Essential Statistics

with Python and R

S. Rao Jammalamadaka

University of California, Santa Barbara

Preface

Statistical ideas have become indispensable not only for doing and understanding scientific research but just for being well-informed citizens. We deal with uncertainties around us in everyday life from weather forecasting to stock-market gyrations and are bombarded with polls and advertisements containing claims and counter-claims. So a certain level of "Statistical literacy" is desirable for all of us. The need for such statistical literacy in this modern age was foreseen by H.G. Wells, who said "Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write" and that day, we think, is already upon us!

This book attempts to cover basic ideas of statistics in a direct and succinct way. The goal is to help develop familiarity with statistical concepts among students and others with varied backgrounds. Minimal mathematical background is assumed and the emphasis is on understanding concepts and how they apply to data. Statistical ideas are explained and then illustrated with one or two simple examples. We use practical and realistic examples in many places, and believe a statistical idea is more easily explained through a simple illustrative example.

Besides teaching basic statistics, this book can also serve students and novices as an introduction to modern computational packages that are being commonly used nowadays for statistical analysis, namely **Python** and **R**. Two Appendices at the end of the book provide basic introduction to these two packages, and demonstrate their use by working out sample exercises taken from the book, in a follow-up Appendix. Further introduction to these packages comes in the form of worked out Examples in various chapters throughout the book. Clearly many more complex data analyses can be done using **Python** and **R**, and we restrict ourselves to their use in connection with the basic topics covered in this book on *"Essential Statistics"*.

Contents

List of Figures

Chapter 1

What is Statistics?

1.1 Introduction

The word "statistics" is used primarily in two different contexts. First, to refer to a collection of numbers or data, as in "football statistics" and "crime statistics" and secondly, when we refer to the subject of statistics, which we are about to study.

> The subject of "statistics" may be called the "science of data" and broadly described as "a body of methods for collecting, summarizing, analyzing, and interpreting data".

Each of these aspects of statistics, simple as it may seem, can be discussed in considerable detail and at a highly sophisticated mathematical level. We will avoid both the detail and the mathematics and try to provide a brief glimpse at statistics, focusing on some essential basic techniques.

We now amplify a bit on each of these aspects, namely collecting, summarizing, analyzing and interpreting data, thus providing a bird's eye-view of the subject.

I. Collecting data

Collecting accurate and representative data in the most economical way is the first preliminary step in statistics. How to collect data is the subject of two specialized branches in statistics, called "Sampling Techniques" and "Design of Experiments". Sampling techniques or methods deal with collecting data in the real world as it exists, namely through opinion polls, cross-sectional studies, surveys dealing with political and social issues, etc. Sampling is an essential part of everyday life and we return to it in the next section. The object of Design of Experiments, on the other hand is to design data collection in a more controlled setting in order to answer specific scientific questions, as in agricultural or clinical trials. The goal in either case is to maximize the "information" obtained for a given amount of money or conversely, to minimize the cost for attaining a given level of precision. To illustrate this point through a very simple example, consider two objects whose unknown weights say θ_1 and θ_2 we need to determine, using a simple balance. Since a measurement always involves error and can only be considered as an "estimate", one very simple approach is to take one measurement on each one of these objects, thus getting two separate "estimates". An alternate procedure (which may not make much sense unless you think more about it), is to take the same two measurements, costing the same amount of time and money, but this time on the the "total weight" and the "difference of weights" of the two objects. From these, it is easy to figure out the individual weights θ_1 and θ_2. It can be shown (and we will see later) that the second approach, which costs the same time or effort, gives estimates of θ_1 and θ_2, which are 100% more accurate than the first! This is a consequence of the fact that averages tend to have higher precision, a fact we will learn soon. Thus, designing a proper experiment is very crucial and different methods of collecting information can be quite different in terms of their efficiency.

II. Summarizing

Summarizing data is the next essential step before we can make sense out of any large set of numbers i.e., to understand what the data is about. Suppose we are given the birthweights of say, 1000 babies born in a hospital in a particular year. Such a large set of numbers by themselves would not make much sense even to an expert eye. If they are appropriately summarized, either graphically or numerically, we can figure out some essential features of this data, like where is the **center** and how much is the **spread** around this center. You have no doubt heard the saying "A picture is worth a thousand numbers"! Providing

graphical and numerical summaries of data (often sample data from a larger population), is called **Descriptive statistics**. Surprisingly, in some cases we find that under appropriate assumptions, one or two numerical summaries is all that we need to make inferences about the larger population in place of the original set of 1000 numbers. In the next chapter, we talk about "graphical summaries" such as the bar charts, pie charts, histograms etc. as well as "numerical summaries" such as the mean, median, range, standard deviation to measure certain features of data sets, like the center and spread.

III. Analyzing and interpreting data

It can be argued that analyzing and interpreting data is the heart of statistics. This is also sometimes called **inferential statistics** or **statistical inference**, i.e., drawing conclusions or inferences about the "population" after observing only a subset – a "sample" from it.

1.2 Population, sample and inference

Most often the data statisticians collect is a **sample**, i.e., a randomly selected (and hence representative) subset of the **population**. We use the word **population** to refer to the "entirety of data one can collect on a topic of interest". For instance, if one is interested in the "average family income" in the US, the population here consists of incomes of every family in the US – over 100 million numbers (after one tackles such non-trivial questions as to what we mean by a "family" and what we mean by "family income" etc.). If suppose we are interested in "television rating" for a particular TV show in a given week, the word population then refers to all the data on time spent by each of the potential viewers watching this show. As a third example, consider the situation where one is interested in figuring out the "chance of heads" for a given coin. The entirety of data here refers to the results of all possible tosses with this coin. Since the coin can be flipped forever to obtain more and more data about this coin, the population is theoretically infinite.

In all these cases, a sensible thing to do is to take a representative sample from such a population and try to interpret what this sample has to say, regarding the population. This type of generalizations i.e., drawing conclusions about the population from the observed sample, is called "statistical inference". Statistical inference is a fundamental ingredient in most scientific advances because we can not wait in most cases for "all the data to be in". In

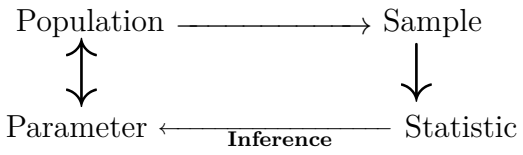
some cases, it is patently unwise or impossible to collect all the data - like in the coin-tossing example. Think what would happen if a manufacturer wanted to determine the "lifetime" of a certain brand of electric bulb and started to burn each of the bulbs manufactured in order to get the "totality of data" on the lifetimes. Not good for business since there would be nothing left to sell!

To recap, we call a "properly" selected subset of population, a **sample** , whose representativeness is guaranteed by selecting the sample units randomly. In most cases, we select a **simple random sample** of size n by giving each unit in the population the same chance of being selected into the sample and also by giving every sample of size n from this population the same chance of being selected. The advantages of random sampling include:
i) reduced cost,
ii) possibility to measure precision in the sample estimates,
iii) greater accuracy and scope are possible, as opposed to a complete population count,
iv) sometimes, sampling is the only way, as when the population is infinite (e.g., coin tossing) or when measurements involve destructive testing (as when studying the life-times of light bulbs).

Sometimes it is known in advance that the characteristic under study is heterogeneous among the various subgroups of the population. For instance, in an agricultural experiment say comparing the yields of different varieties of wheat, we may know that certain plots of land are more fertile than others and will give higher yields. Or in an electoral poll, we may know that Blacks, Hispanics and Caucasians tend to vote quite differently. In such cases, it is best to separate out the population into these subgroups, called "blocks" in an experimental design context or "strata" in a sampling context and then randomly select a proportionate number of units from each subgroup. This idea termed **blocking** avoids the possibility of all fertile plots of land being assigned to a particular variety of wheat or of our poll sample being entirely Caucasian and thus ensures the representativeness of the overall sample. The choice of blocks or strata should be such that all the units within a block are as homogeneous as possible and there is considerable difference between blocks. A simple example of such blocking occurs later in Section 9.1, in what is referred to as paired data.

Population characteristics (like its center or spread) are called **parameters** and are denoted typically by Greek letters such as μ and σ. On the other hand, a sample characteristic, which is computed based on the sample values, is called a **statistic** (yet another meaning

for the word "statistics" — as the plural of the word "statistic") and are generally denoted by English alphabet, using symbols like $\overline{x}$ and s. Statistical inference typically proceeds through "estimating" or sometimes "testing hypotheses" i.e., statements about the unknown parameters, using sample based values, i.e., the statistics. Sometimes, the populations may not be characterized through a few parameters and the inference for such populations, called "nonparametric inference", is not discussed in this book.



Of course, when one tries to generalize from an observed sample to the (incompletely observed) population, there are bound to be pitfalls. One can never be sure that the conclusions are quite right and such generalizations may involve errors. The beauty of statistics is that it allows us to quantify these errors and control them as desired. Since typically a larger sample size provides more information about the population, choosing an appropriately large enough sample size is one way to reduce the error. We will come back to this idea of choosing an appropriate sample size, in later chapters.

Consider the following simple example to illustrate the process of inference. In estimating the "chance of heads", recall that the coin can be tossed indefinitely. We are forced to stop at some finite point, say after 100 tosses. This is a sample of size 100 and suppose the data collected looks like this (with H and T standing for Heads and Tails respectively):

H, H, T, H,....., T.

Suppose there are altogether 46 heads (and the other 54 are tails) in this sample of 100 tosses. It appears reasonable to declare the observed proportion of heads in our sample, namely $(46/100) = 0.46$ as our best guess or estimate of the "chance of heads". How certain are we? Of course there is no guarantee that if we tossed the same coin another 100 times, we will get again 46 heads and 54 tails. This kind of variation from sample to sample, is referred later on as **sampling variability**. How different can that other estimate be, from 0.46? If we can declare something like, "Well, we have 0.46 for our estimate but if we repeat this again, 90% of the time it will be within 0.08 of 0.46", i.e., the error in our estimate is no more than 0.08, with 90% chance –that would be somewhat reassuring. This is the type

of errors we are talking about measuring. This type of measurement of errors, is based on ideas of probability, which we explore in Chapter 3.

EXERCISES

1.1 Critique the following statements:

(a) The average height of the family is 5.5 feet whereas the average depth of the stream is only 4 feet. So the family can safely walk across the stream.

(b) Statistics is like accounting since both deal with numbers.

(c) Most of the people who wrote the company had nothing but praise for its product. So the product must be good.

(d) Almost a 1,000 people died of a deadly disease in spite of being vaccinated for it. So the vaccination must be useless.

1.2 Define what the population is in each of the following cases and how you might draw a sample:

(a) A limnologist wants to study the amount of dissolved phosphates in a lake.

(b) A medical doctor wishes to study if lung cancer is caused by smoking.

(c) A biologist wants to study the average length of the tail of a certain species of adult monkeys.

(d) You wish to find what proportion of Californians are in favor of a certain ballot initiative.

1.3 How does descriptive statistics differ from inferential statistics?

1.4 Find a topic of statistical inference in today's paper and report on it critically. Does the "stock table" giving price changes of stocks represent statistics? If so, in what sense?

1.5 Market research is often done by telephoning people randomly selected from a telephone directory. This clearly eliminates people who have no phones or have unlisted numbers. Can you think of any ways to include one or both these groups in a survey?

1.6 How representative is the sample and how valid is the inductive reasoning in the following examples?

(a) A young lady has been jilted by 3 boyfriends, all of them above 5 feet 6 inches and she vows never to date guys above 5 feet 6 inches.

(b) A doctor takes a small amount of blood from you and concludes your blood sugar is normal.

(c) You buy a sack of potatoes after closely inspecting a few that you can actually see.

1.7 Explain why the following poll which was actually advertised in a major newspaper, might not be unbiased: "Should handgun control be tougher? You call the shots in a special call-in poll tonight. If yes call 1-900-720-6181. If no, call 1-900-720-6182. Charge is $1.00 for the first minute."

1.8 It is known that 1 in 12 industrial workers have a drinking problem. So the supervisors are warned "If you have 12 people working for you, one of them has a drinking problem. If you do not believe it, you are kidding yourself!". Criticize the conclusion.

Chapter 2

Descriptive Statistics: Graphical and Numerical Summaries

As the British physicist and mathematician Lord Kelvin said in 1883: ".... when you can not measure it, when you can not express it in numbers, your knowledge is of a meager and unsatisfactory kind ... scarcely advanced to the stage of science". And since repeated measurements even under identical conditions vary, Galileo (1564-1642) emphasized this further by saying: "Measure, measure, measure. Measure again and again to find out the difference and the difference of the difference". This is what statisticians really do — measure things and repeat these measurements, in order to figure out the magnitude of the characteristic of interest, as well as its variability or the measurement error.

2.1 Scales of measurement

We saw that a "population" refers to all possible data on a given topic of interest. A particular member of this population will be referred to as an "**individual**", on whom we make the measurement(s). Such measurements can be "**qualitative**" (or attributes) like the eye-color, marital status etc. or be "**quantitative**" — taking on numerical values, like the height, the family income etc. More systematically, one can define any measurement as being on 4 **levels or scales**, each successively more precise than the preceding. The first two levels of measurement (**Nominal** and **Ordinal**) lead to "Qualitative" variables whereas

the last two (**Interval** and **Ratio**) correspond to "Quantitative" variables.

(i) **Nominal scale**: This is the most basic level of measurement where an individual is classified into a group depending on whether or not they possess certain attributes or characteristics. For example, "gender" may be measured as "male" or "female", while "marital status" is measured as "single", "married", "divorced", etc. There is no natural order or ranking to these nominal classes, i.e., we can not say whether male or female is a better or higher category.

(ii) **Ordinal scale**: Here, not only do we have a label or name for each category as in the nominal scale, but also an order attached to them. For example, we may rate a product as "excellent", "very good", "good", "poor" or when a student is given a grade A, B, C, D and F. In the latter example, supposedly, an A is better than a B and so on!

(iii) **Interval scale**: Here the measurements can be quantified with numerical values which bear an order relationship. In addition, arithmetic differences are meaningful and intervals of equal width signify equal differences in the characteristic as e.g., temperature measured in degrees or the scores on a test. However, a zero on this scale does not signify the absence of the characteristic i.e., there is no "absolute zero". For instance, when the temperature is zero, it does not mean there is no temperature.

(iv) **Ratio scale**: This is a level of measurement which has all the properties of the interval scale and in addition, possesses an "absolute zero", signifying the absence of the characteristic. For instance things like height and weight belong to the ratio scale and taking ratios makes sense here, unlike things measured on interval scale. One who is 6 feet tall is twice as tall as one who is only 3 feet tall. On the other hand, we could not say that a city where the temperature is 60 °F on a given day is twice as hot as one where the temperature is 30 °F!

2.2 Graphical Summaries

Qualitative variables can be displayed graphically by bar charts, pie charts etc.

A **Bar Chart** is a simple graphical display in which the length of each bar represents the frequency in each category. To draw one:

(i) On the horizontal axis, write the names of the categories,

(ii) On the vertical axis, mark the frequency of each of these categories using an appropriate scale ,

(iii) Place a bar (or rectangle) above each category label with the height corresponding to the frequency. Base widths of these bars should be equal.

We now give instructions needed to generate the same bar chart using **R** or **Python**. But before using either of these software packages, the reader is referred to the Appendices B and C and get familiar with the general introduction given there, to these two popular packages.

R code instruction:
A Bar Chart can be generated in **R** by the built-in function using the Syntax:
(Note that "height" is the name of the data set, "main= " defines the title for the chart, and "names.arg= " generates names for each bar).

```
barplot(height , names.arg = , main = )
```

Python code instruction:
In **Python** programming, "matplotlib" package is used to generate different kinds of plots using the Syntax:

```
import matplotlib.pyplot as plt
plt.show()
```

The specific **Python** syntax for plotting Bar Chart is:

```
plt.bar( )
```

Example 2.1: For purposes of national mortality statistics, every death is attributed to one underlying condition. According to the National Center for Health Statistics (U.S. Dept. of Health and Human Services), the 6 leading causes of death in 1995 are as shown in the table below, along with a bar chart in Figure 2.1 :

Deaths in 1995 along with the cause (in thousands)	
Heart diseases	738
Cancer	538
Stroke	158
Pulmonary diseases	103
Accidents	93
Others	682
All causes	2,312

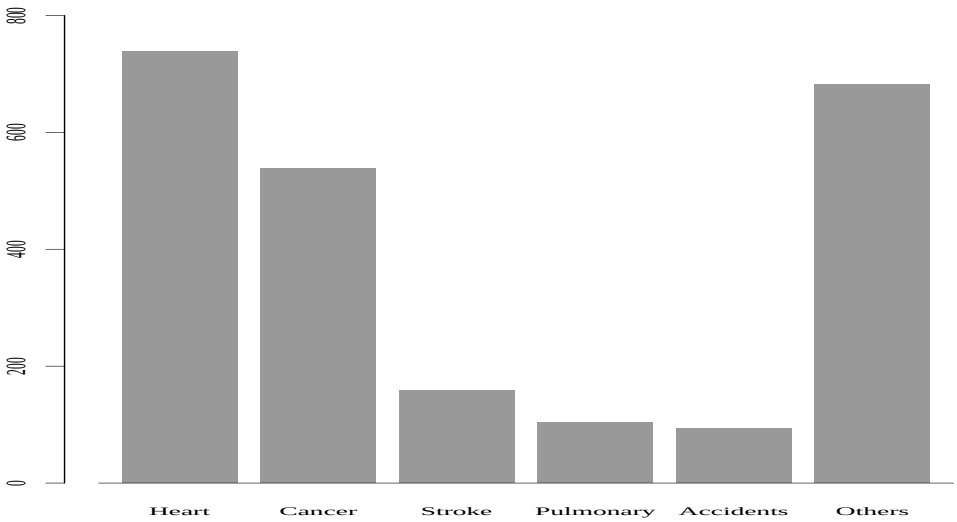


Figure 2.1: Bar Chart of causes of death

R code:

```
dataset=c(738,538,158,103,93,682)
barplot(height=dataset,main="Bar Chart of causes of death", names.arg=c("Heart
    ","Cancer","Stroke","Pulmonary","Accidents","Others"))
```

Python code:

```
dataset=(738,538,158,103,93,682)
name=('Heart','Cancer','Stroke','Pulmonary','Accidents','Others')
plt.bar(name,dataset)
plt.title('Bar Chart of causes of death')
plt.show()
```

◇

A **Pie Chart** on the other hand is one in which the total "pie" is divided into slices proportional to the number of observations (frequency) in each category. It's main advantage is in letting the eye see how the total count is divided up into the different categories. To draw a pie chart:

(i) Find the relative frequency (or percentage) corresponding to each category (e.g., heart disease has $738/2312 = 0.3192$ of the total),

(ii) Find the angle corresponding to each of the categories, keeping in mind that there are 360 degrees in all, at the center. If f is the frequency of a particular category and n is the total number of the observations, then the corresponding degrees in the pie graph is $\frac{f}{n} \times 360$ (e.g., heart disease should correspond to an angle of $0.3192 \times 360 = 115$ degrees). A pie chart for this data is given in Figure 2.2.

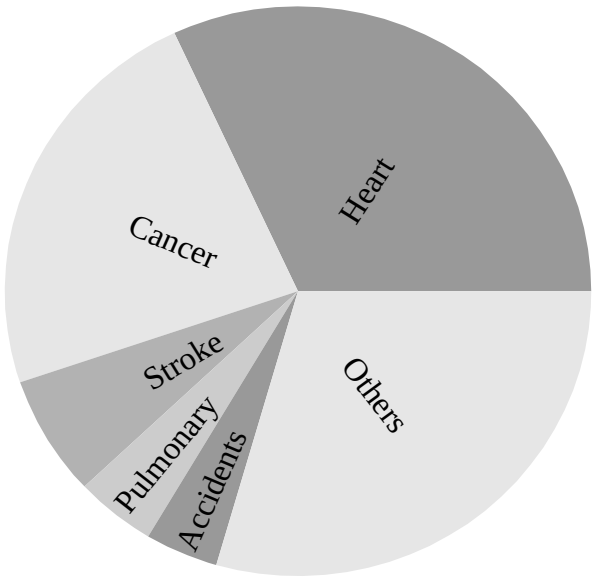


Figure 2.2: Pie Chart of causes of death

R code instruction:
Similarly, a pie chart can also be plotted in **R** by using the built-in function with the command:
Note that x represents the data set, "radius = k" defines the size of the pie chart where k is a constant, labels are the names of each category.

```
pie(x, labels = names(x), radius = k, main = )
```

R code:

```
pie(dataset,radius=1.0,labels=c("Heart","Cancer","Stroke","Pulmonary","
    Accidents","Others"),main="Pie Chart of causes of death")
```

Python code instruction:
Color can be defined for each part using "colors= " inside of the *plt.plot*(). The Syntax for plotting a Pie Chart is:

```
plt.pie( )
```

Python code:

```
colors=('r','y','g','b','grey','purple')
plt.pie(dataset,labels=name,colors=colors,radius=1.0)
plt.title('Pie Chart of causes of death')
plt.show()
```

◇

Quantitative variables on the other hand, can be represented in many ways. We will describe just two basic graphical methods — **histograms** and **stem-and-leaf plots**.

A good way to summarize and make sense of a large set of numbers is to form a **frequency distribution** which tells us where the values are and how frequently they occur. To form a frequency table (or frequency distribution table), we proceed as follows:

(i) Locate the minimum and maximum values among the data.

(ii) Break this range of values into a small number of groups — called "class intervals" or "bins".

(iii) Find the frequency in each bin i.e., how many data points fall into each of these class intervals.

Remark 2.1 An important question here is "how many bins should one use?" Any frequency table, while it provides a more easily understandable summary, involves loss of information since we can not reconstruct the original values of the data from the frequencies in the bins. Smaller the number of bins (with large bin-widths), more such loss of information. On the other hand, it is not good to have too many bins either since this will mean not enough summarization has taken place. Typically 5-15 bins are reasonable depending on the size of the data. In terms of easy interpretation, it is preferable to have equal bin-widths although it is not always practical, especially when the distribution has long tails with very few observations there.

We can now construct what is known as a **histogram** from such a frequency table. A histogram is a just a bar chart with classes (bins) on the x-axis, with the corresponding frequencies represented on the y-axis. Note that by looking at a histogram, one can judge where the center is and how much spread there is around that center. Also one can observe the "**shape**" of the histogram i.e., whether it is **symmetric or asymmetric**. If it is asymmetric and the longer tail is on the right hand side, the distribution is called "**positively skewed**" or "right skewed", whereas if the longer tail is on the left hand side, it is called "**negatively skewed**" or "left skewed".

Sometimes, we are interested in the *proportions* of values that fall into each bin, (called "relative frequencies") instead of the frequencies. This is called the relative frequency distribution, from which we can construct a relative frequency histogram. The relative frequencies, of course, add up to 1. Note that the shape of a relative frequency histogram is exactly the same as the corresponding histogram, only the scale on the y-axis is different. For instance, the relative frequency histogram for the website data (Example 2.2 below) will look exactly the same as Figure 2.3 except on the y-axis, the values 5, 10, 15, 20 are replaced by 0.1, 0.2, 0.3 and 0.4 respectively.

R code instruction:
In general, a frequency table is necessary as preparation for plotting the histogram. However, in **R** programming, a histogram can be generated without having a frequency table first by using a built-in function:
Note: x is still the data set, xlab is the name of category, and ylab is frequency by default.

```
hist(x, xlab = , ylab= , main= )
```

Python code instruction:
In order to define the shape of the histogram, "histtype=shape" is used inside of this function. However, it takes the shape of bars by default if it is not specified. The function for plotting histogram is:

```
plt.hist( )
```

Example 2.2: Consider the following data on the number of "hits" per day for the website of a statistics course, over a 50 day period is:

20, 14, 21, 29, 43, 17, 15, 26, 8, 14, 39, 23, 16, 46, 28, 11, 26, 35, 26, 28,

22, 30, 17, 23, 9, 27, 18, 22, 19, 25, 31, 55, 63, 52, 16, 13, 23, 33, 43, 49,

25, 32, 26, 51, 39, 42, 55, 41, 36, 32.

Construct a frequency table and draw a histogram.

Solution: By selecting class-intervals of width 10 units each, we get the following frequency table. The last column in the table represents "relative frequencies" i.e., frequencies in each class-interval divided by the total frequency which is 50 in this case.

Frequency Table of website hits

Class interval	frequency	Relative frequency
0-9	2	0.04
10-19	11	0.22
20-29	17	0.34
30-39	9	0.18
40-49	6	0.12
50-59	4	0.08
60-69	1	0.02
Total	50	1.00

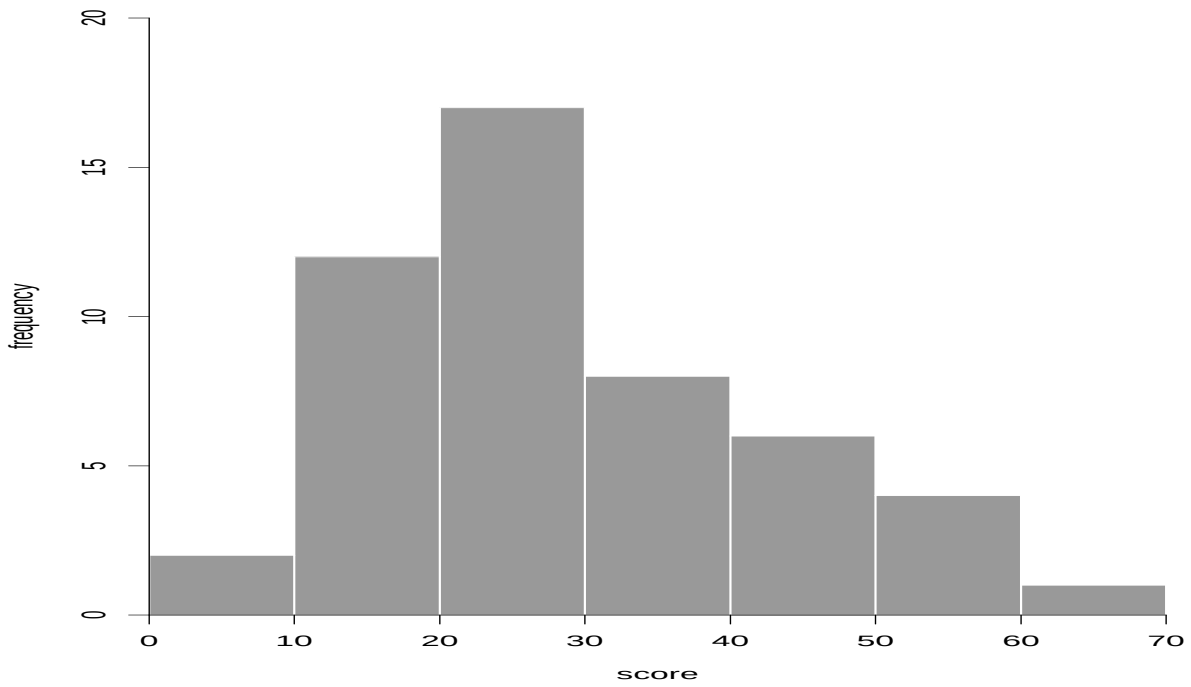


Figure 2.3: Histogram of website hits

R code:

```
dataset1=c(20,14,21,29,43,17,15,26,8,14,39,23,16,
           46,28,11,26,35,26,28,22,30,17,23,9,27,
           18,22,19,25,31,55,63,52,16,13,23,33,43,
           49,25,32,26,51,39,42,55,41,36,32)
hist(dataset1,main="Histogram of website hits",xlab="score")
```

Python code:

```
dataset=(20,14,21,29,43,17,15,26,8,14,39,23,16,46,28,11,26,35,
         26,28,22,30,17,23,9,27,18,22,19,25,31,55,63,52,16,13,23,
         33,43,49,25,32,26,51,39,42,55,41,36,32)
plt.hist(dataset,color='grey')
plt.title('Histogram of website hits')
plt.xlabel('score')
plt.show()
```

◇

A **stem-and-leaf diagram** is essentially a bar chart drawn sideways. The stem-and-leaf plot works well with small data sets and this is one of the few graphical displays where the original data is not lost. To draw a stem-and-leaf diagram, pick one or more leading digits (generally all but the last significant digit) for the **stem** and attach the **leaves** representing the trailing digit(s). Stem values are listed in a vertical column, with a leaf for each observation, besides the corresponding stem value. We illustrate this with an example.

R code instruction:
There is also a built-in function for stem and leaf plot:
Note: Unlike other graphs, stem-and-leaf diagrams do not contain names since "main=" is not a part of this function. x is still the data-set, "scale=n" controls the plot length, and "width=t" is the desired width of plot.

```
stem(x, scale = n, width = t)
```

Python code instruction:
A stem and leaf plot uses the function:

```
plt.stem( )
```

Example 2.3: The midterm scores in a statistics class of 25 students are:

92, 88, 74, 83, 86, 64, 82, 85, 80, 66, 83, 98, 77, 69, 61, 57, 78, 86, 90, 81, 87, 79, 62, 89, 72.

A stem-and-leaf plot of this data is given by

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

◇

For example, the observation 74 is represented here as 7 | 4 . The number of stems corresponds to choosing the number of classes in a frequency table. If there are too many leaves on a single stem, one can use a **split stem-and-leaf diagram**, splitting the same

stem into two, once with the low leaves (the lower half) and once with the high leaves (the upper half). For Example 2.3, this gives:

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

R code:

```
dataset2=c(92,88,74,83,86,64,82,85,80,66,83,98,77,
           69,61,57,78,86,90,81,87,79,62,89,72)
stem(dataset2,scale=1,width=100)
```

Python code:

```
dataset=(92,88,74,83,86,64,82,85,80,66,83,98,77,69,61,
         57,78,86,90,81,87,79,62,89,72)
plt.stem(dataset)
plt.show()
```

2.3 Numerical Summaries

2.3.1 Measuring the center of a data set

It will be useful to have a summary measure — a single number that represents the **center** of a set of values. There are several possible measures of the center and the most important among these are: the **Mean**, the **Median** and the **Mode**.

We will now employ some symbols to represent our data, so we do not have to write sets of specific numbers each time. This will simplify presentation of the ideas. The values observed in a sample of size n, maybe represented as $x_1, x_2, ..., x_n$ where

x_1 = the value on the 1^{st} individual,

x_2 = the value on the 2^{nd} individual

$\vdots$

x_n = the value on the nth individual.

Here n represents the size of the sample. Of course, in any given problem in practice, these $(x_1, x_2, ..., x_n)$ are specific numbers. To measure the center for such a data set, we may use:

(i) **Mean:** The arithmetic mean, sometimes loosely called the **average** of a set of data, is obtained by summing all the values in the data and dividing by the number of values. In a physical sense, this corresponds to the "center of gravity" of a system where unit weights are attached at each of the data points. We use $\overline{x}$ as the standard notation for the mean of x values, so that:

$$\begin{aligned} \text{Mean, } \overline{x} &= \text{(sum of all the observations) / (the total number of observations)} \\ &= \frac{x_1 + x_2 + \cdots + x_n}{n} \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i, \end{aligned}$$

where $\sum_{i=1}^{n} x_i$ = sum over all the observations x_i, with $i = 1$ to n,

and n = total number of observations.

(ii) **Median:** A median is the "the middle value" when the data set is arranged in increasing (or decreasing) order of magnitude. An equal number of observations are smaller than this median as are larger than this.

If the number of observations in the sample n, is odd, the median is the $\left(\frac{n+1}{2}\right)^{th}$ largest value.
On the other hand, if n is even, the average of the two middle values, namely the $\left(\frac{n}{2}\right)^{th}$ and the $\left(\frac{n}{2}+1\right)^{th}$ is taken as the median.

For example, if $n = 21$, the median is the 11^{th} largest value, whereas if $n = 20$, the median is the average of the 10^{th} and 11^{th} largest values. In either case, note that 10 observations are smaller than the median value and the same number of observations are larger than the median value.

(iii) **Mode:** Mode is the most frequent value in a given data set.
The mode may not always be uniquely defined since no single value may be most frequent.

R code instruction:
Built-in functions are constructed in **R** to find mean and median with the Syntax:

```
mean( )
```

```
median( )
```

Even though there is no built-in function in **R** for finding mode, another function in ***statip*** package can give the value of mode. The Syntax is:

```
install.packages("statip")
library(statip)
mfv( )
```

Python code instruction:
Numpy library is used to find the mean and median for a given data set with the Syntax:

```
import numpy as np
np.mean( )
```

```
import numpy as np
np.median( )
```

Stats package from ***Scipy*** library is used to find the mode for a given data set with Syntax:

```
from scipy import stats
stats.mode( )
```

Example 2.4: Suppose we have a data set with $n = 5$ observations, say the scores obtained by a student in the 5 quizzes given to her. Suppose her scores are
8, 5, 7, 3, 7.

Then, the mean $= \frac{8+5+7+3+7}{5} = 6$.

To find the median, we first order the data, resulting in

$$3, 5, 7, 7, 8.$$

Since we have an odd number of observation, the median is the $\left(\frac{5+1}{2}\right)^{th}$ or the 3^{rd} largest observation, which in this case is 7.

Finally the mode here is 7 (most frequent value) since it occurs twice compared to other values which occur only once.

If however n were 4, say with the data being, 3, 5, 7, 7 then the median would be the average of the middle two values, namely 5 and 7, which is 6.

R code:

```
dataset=c(8,5,7,3,7)
mean(dataset)
median(dataset)
library(statip)
mfv(dataset)
```

The output of this **R** code is:

```
> dataset=c(8,5,7,3,7)
> mean(dataset)
[1] 6
> median(dataset)
[1] 7
> library(statip)
> mfv(dataset)
```

```
[1] 7
```

◇

An **outlier** is an observation, which is far outside the regular pattern of data. Outliers can be due to errors in measurement and/or in recording, in which case people tend to disregard them. However, often outliers may point towards something very important and the cause of outliers should be investigated. The median and the mode are resistant to outliers and are referred to as **resistant** or **robust** measures of center, whereas the arithmetic mean is not robust, as the following simple example illustrates.

Example 2.5: Suppose in Example 2.4 with $n = 5$, the first observation is changed from 8 to 38. Clearly 38 is a very large or an "extreme" value compared to the rest of the numbers — an outlier. Then for the new data set with the values ordered,

$$3,\ 5,\ 7,\ 7,\ 38,$$

we get

$$\overline{x} = \frac{3 + 5 + 7 + 7 + 38}{5} = 12,$$

$$\text{Median} = 7,$$

$$\text{Mode} = 7.$$

Python code:

```
import numpy as np
dataset=(3,5,7,7,38)
np.mean(dataset)
np.median(dataset)
from scipy import stats
stats.mode(dataset)
```

The output of this **Python** code is:

```
12.0 7.0
ModeResult(mode=array([7]), count=array([2]))
```

◇

Note that by changing a single observation, the mean $\overline{x}$ jumped from 6 to 12 whereas the median and mode did not change — illustrating a single or a few outliers do not affect the median or mode, while they can change the mean substantially.

Remark 2.2 Now some general guidelines in choosing an appropriate measure of center: If the data is symmetrically distributed and approximately follows a Normal or bell-shaped curve (about which we shall talk in Chapter 5 and in later chapters), the mean $\overline{x}$ is the preferred measure. The mean is also unambiguously defined and has some nice mathematical properties. In fact for nice symmetrical distributions with one peak, it is easy to see that the mean, the median and the mode coincide. However if the data comes from a decidedly skewed distribution as can be judged from the plots of a histogram or a stem-and-leaf diagram, or when outliers are present, we would prefer a more robust measure of the center like the median or the mode. For instance, when talking about house-prices (or incomes), since we would not want a few very expensive homes (or a few rich individuals) to distort the center, one would prefer to use the "median selling price" or "median income" rather than the mean, to represent the center.

2.3.2 Measuring the spread of data

Variability or diversity is natural part of life and the theory of statistics is based on such inherent variability which is all around us. Without this, imagine how monotonous and dull life would be! Every individual would be the exact replica of every other and from a statistical point of view, we would not need to take any more than a sample of size one to reveal the mysteries of the population. Fortunately such is not the case!

To measure this variation, (also called the **spread**, **scatter** or **dispersion**) in a given data set, there are several possible measures that one could use. We will discuss three of the most common and important measures of spread. These are the **Range**, the **Inter-Quartile Range** and the **Standard Deviation**.

(i) **Range** is the difference between the largest and the smallest values.

Range = (Maximum value - Minimum value)

Clearly, the more spread out the data is, the larger the range and vice versa. Thus range is a simple-to-calculate measure of spread.

(ii) **Inter-quartile range (IQR)**

If the number of observations is large (say 20 or more), it is sometimes useful to extend the idea of median and divide the ordered data into quarters. The points for division into quarters are called **quartiles** just like the point of division into halves is called the median. With even larger samples, one can divide the data into even smaller fractions of $\frac{1}{100}^{ths}$, called the **percentiles** . We are familiar with statements, with respect to standardized tests like the SAT, GRE etc., where we say one's score is in the 85^{th} percentile, meaning that 85 percent of the students taking this test, got scores lower than this and the other 15 percent got higher scores. In particular, we note that:

- The 50^{th} percentile is the **median** that we discussed before.
- The 25^{th} percentile is also called the **first quartile** and is denoted by $\mathbf{Q_1}$. It is also the median of the lower half of the observations, i.e., those which are to the left of the overall median in the ordered list.
- The 75^{th} percentile is also called the **third quartile** and is denoted by $\mathbf{Q_3}$. It is also the median of the top half of the observations, i.e., those which are to the right of the overall median in the ordered list.

Note that the median is also the second quartile. The range between the $\mathbf{Q_1}$ and $\mathbf{Q_3}$, which contains the middle 50 percent of the data, is another measure of spread for a given data set and is called the **Interquartile range**. This is clearly a resistant or robust measure since it ignores very large or very small values at either end.

Interquartile Range, $\mathbf{IQR} = (\mathbf{Q_3} - \mathbf{Q_1})$.

R code instruction:
Built-in functions are constructed in **R** to find IQR and range:

```
IQR( )
```

```
range( )
```

Note: The range( ) can only output the Maximum and Minimum values of a given data set. Therefore, following command is used to find the range:

```
max( )-min( )
```

Python code instruction:
Stats is also used to find IQR:

```
from scipy import stats
stats.iqr( )
```

Then one finds the range using the same method:

```
max( )-min( )
```

Example 2.3 (contd.): (i)Find the range and the interquartile range (**IQR**).
(ii)Suppose the score 79 is entered as 19 by mistake. Find the new range and **IQR**.
(iii)Suppose the student with the midterm score of 57 dropped out of the course, later. What is the new range and the **IQR**?

Solution:

(i) First, arrange the 25 data points in increasing order:

57 61 62 64 66 69 72 74 77 78 79 80 81 82 83 83 85 86 86 87 88 89 90 92 98

The **range** = 98 - 57 = 41. The median is the 13^{th} largest observation, which is 81.

$\mathbf{Q}_1$ is the median of the 12 observations to the left of the median = $\frac{69+72}{2}$ = 70.5,

$\mathbf{Q}_3$ is the median of the 12 observations to the right of the median = $\frac{86+87}{2}$ = 86.5.

Therefore, **IQR** = 86.5 - 70.5 = 16

R code:

```
dataset=c(57,61,62,64,66,69,72,74,77,78,79,80,81,82,
          83,83,85,86,86,87,88,89,90,92,98)
max(dataset)-min(dataset)
IQR(dataset)
```

The output of this **R** code is:

```
> dataset=c(57,61,62,64,66,69,72,74,77,78,79,80,81,82,83,83,85,86,
86,87,88,89,90,92,98)
> max(dataset)-min(dataset)
[1] 41
> IQR(dataset)
[1] 16
```

(ii) Now, the ordered data is:

19 57 61 62 64 66 69 72 74 77 78 80 81 82 83 83 85 86 86 87 88 89 90 92 98

Now the **range** = 98 - 19 = 79. The median is the 13^{th} largest observation, which is 81,

$\mathbf{Q}_1$ is the median of the 12 observation to the left of the median = $\frac{66+69}{2}$ = 67.5,

$\mathbf{Q}_3$ is the median of the 12 observation to the right of the median = $\frac{86+87}{2}$ = 86.5.

Thus the **IQR** = 86.5 - 67.5 = 19. The IQR changed very little whereas the range is considerably different.

(iii) Ordering the 24 students in the class, the data is

61 64 69 74 77 78 79 80 81 82 82 82 83 83 85 86 86 86 87 88 89 90 92 98.

The **range** = 98 - 61 = 37. The median is $\frac{82+83}{2}$ = 82.5.

$\mathbf{Q}_1$ is the median of the 12 observation to the left of the median = $\frac{78+79}{2}$ = 78.5,

$\mathbf{Q}_3$ is the median of the 12 observation to the right of the median = $\frac{86+87}{2}$ = 86.5,

so that **IQR** = 86.5 - 78.5 = 8.0

Python code:

```
data=(61,64,69,74,77,78,79,80,81,82,82,82,83,83,85,86,86,86,87,88,
89,90,92,98)
max(data)-min(data)
from scipy import stats
stats.iqr(data)
```

The output of this **Python** code gives the range and the IQR:

```
37.0 8.0
```

◇

The five values, namely, (the minimum value, $\mathbf{Q}_1$, the median, $\mathbf{Q}_3$, the maximum value) provide a useful description of where the data is and how much the spread is and is called the **Five Point Summary**.

Five point summary = (minimum, $\mathbf{Q}_1$, median, $\mathbf{Q}_3$, maximum)

R code instruction:
In order to find the five values at once, the function is:

```
summary( )
```

Boxplots

The information contained in the five-point summary, can also be represented graphically by what is called a **boxplot**, which draws a box of a desired width from $\mathbf{Q}_1$ through $\mathbf{Q}_3$, noting the center at the median. The box has **whiskers** which connect up to the minimum value at one end and the maximum value at the other end.

R code instruction:

```
boxplot(X, horizontal=TRUE)
#horizontal=TRUE sets the boxplot in the horizontal display. If it is not
    specified, the boxplot will be displayed vertically by default.
```

Python code instruction:

```
import matplotlib.pyplot as plt
plt.boxplot( )
plt.show()
```

Example 2.3 (contd.): For the midterm scores of the 25 students:

Minimum = 57, $\mathbf{Q}_1$ = 70.5, median = 81, $\mathbf{Q}_3$ = 86.5 and maximum = 98.

Thus the five point summary consists of these 5 numbers and is (57, 70.5, 81, 86.5, 98). The boxplot is given in Figure 2.4.

ch

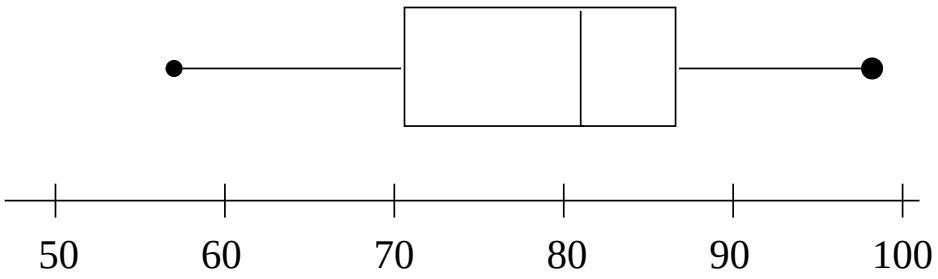


Figure 2.4: Boxplot of 25 scores

R code:

```
dataset=c(57,70.5,81,86.5,98)
boxplot(dataset, horizontal=TRUE)
```

◇

Remark 2.3 Although there is not a formal definition of an **outlier**, we may consider a data point an outlier, if it is 1.5× **IQR** above $\mathbf{Q}_3$ or 1.5×**IQR** below $\mathbf{Q}_1$. If such outliers are present in a given data set, then the whiskers are drawn only to points within this range ($\mathbf{Q}_1$ - 1.5×**IQR**, $\mathbf{Q}_3$ + 1.5×**IQR**) with outliers being identified by small open circles –isolated points unconnected to the main part of the data.

(iii) **Standard deviation**

Using the mean, $\overline{x}$ as the reference point or center, we find the distance of each point x_i from the mean. These values $(x_i - \overline{x})$ are called the **deviations from the mean** and can be negative as well as positive. These deviations always mathematically add up to a zero since the negative ones can be shown to exactly cancel out the positive ones.

Fact 2.1: $\sum(x_i - \overline{x}) = 0$ for any data set.

Thus, averaging these deviations does not tell us anything useful about the data spread. On the other hand, squaring these deviations makes them all positive and if we then average these, it will result in a useful measure of spread. It turns out that it is better to divide the sum of these squared deviations by $(n-1)$, rather than n, to give us what we will later call

an "unbiased" measure of the true variation. This division by $(n-1)$ may also be justified by the fact that only $(n-1)$ of these deviations are actually independently determined, since their total is always fixed at zero by Fact 2.1. Thus we define a very useful alternate measure of spread namely the

$$\boxed{\textbf{Sample Variance, } s^2 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}}$$

An algebraically equivalent computational formula for variance is

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

If the data is more spread out from the center, then the deviations are larger and so is the resulting value of s^2 and s. Thus, s is a useful measure of variation or spread in a given data set.

Fact 2.2: Since we are adding the squares of deviations, it is clear that

$s^2 \geq 0$ for all data sets

and

$s^2 = 0$, if and only if there is no variation in the data

i.e., all the values are the same.

The positive square root of the variance is called the

$$\boxed{\textbf{Sample Standard Deviation (SD), } s = \sqrt{\text{variance}} = \sqrt{s^2}.}$$

Notice that taking the square root gets us back the original units. For instance, if the

weights are measured in lbs, s^2 is in (lbs.^{2}) while s brings it back to the original units, viz., the lbs.

R code instruction:

```
sd( )
#function for calculating the sample standard deviation
var( )
#function for calculating the sample variance
```

Python code instruction:
Statistics package is used to find the sample standard deviation and the sample variance.

```
import statistics
#For the sample standard deviation
statistics.stdev( )
#For the sample variance
statistics.variance( )
```

Example 2.4 (contd.):

In this example, the "deviations from the mean", i.e. $(x_i - \overline{x})$, are

$$(3\text{-}6) = \text{-}3, \qquad (5 - 6) = \text{-}1, \qquad (7 - 6) = 1, \qquad (7 - 6) = 1, \qquad (8 - 6) = 2$$

and their sum is zero as claimed in Fact 2.1. However the sample variance

$$s^2 = \frac{9+1+1+1+4}{5-1} = \frac{16}{4} = 4$$

Thus the standard deviation, which is the square root of the variance, is 2. In this example it is easy to see that the SD is zero if and only if all the 5 observations are the same, namely 6, 6, 6, 6, 6 i.e., when data has no variation.

Python code:

```
data=(3,5,7,7,8)
import statistics
statistics.variance(data)
statistics.stdev(data)
```

The output of this **Python** code gives:
4.0 2.0 ◇

Remark 2.4 If we ever needed to reconstruct the original data from a frequency table, for instance to find the average or the sample variance it is a reasonable approximation to assume that all the observations falling inside a bin, have their values at the "**midpoint**" of that class-interval. Thus, in the above formula for mean and standard deviation, replace $(\sum x_i)$ by $(\sum f_i m_i)$ and $(\sum x_i^2)$ by $(\sum f_i m_i^2)$, where m_i is the mid-point of the class interval i and f_i is the frequency of that interval.

Remark 2.5 Most calculators have a button for σ_{n-1}. This is the s with the denominator $(n-1)$, that we want.

In general, a large part of statistical inference is based on the so-called **Normal Distribution** (which we discuss in Chapter 5) and then the preference is to use the mean $\overline{x}$ as the measure of center and s^2 (or s) as the measure of spread. Although it is beyond the scope of this book, these statistics $(\overline{x}, s)$ are called "sufficient statistics" for such data. i.e., they contain "all" the information about the model that the entire data possesses. Also $(\overline{x}, s)$ are clearly and unambiguously defined algebraically. On the other hand, in dealing with distributions which have longer or heavier tails and when outliers are present, more robust measures like the median and the **IQR** are preferred to $\overline{x}$ and s.

EXERCISES

2.1 Determine the level/scale of measurement for each of the following variables:

(a) the number of hours you study

(b) the weight of a newborn baby

(c) the blood type

(d) price of IBM stock

(e) the score a figure-skater receives from a judge.

2.2 A Statistics class consisted of the following majors (with the corresponding number of students):

Communication	36
Biology	15
Economics	22
Psychology	32
Undeclared	40

What is the variable being measured and what scale is it measured in? Plot a bar graph and a pie chart representing this data.

2.3 (a) What distinguishes measurements that can be made in "Ratio scale" versus those that can only be made in "Interval scale"?

(b) Can a measurement be in both interval and ratio scale? Give an example.

(c) Can qualitative data be in interval scale?

(d) What scale is "the price of IBM computer stock" measured in?

2.4 Of the total of 14,526 domestic freshman admissions at the University of California, Santa Barbara in 1997, 140 were American Indian/Alaskan, 438 were African American, 2,215 were Chicano/Latino, 3,075 were Asian/Pacific Islander, 7,933 were White/Other, while 725 declined to state their race. Draw a bar chart and a pie chart representing this data.

2.5 Suppose a radar records the speeds of several automobiles, based on which the mean, median, mode, range, IQR and s^2 were computed. But on closer scrutiny, it was found that the radar had been calibrated wrongly and all the recorded speeds are 5 miles too fast. How does it affect each of these summary measures?

2.6 The following data represent the number of hours that 30 students devote to homework and study in a typical week:

5, 5, 16, 7, 18, 8, 3, 5, 10, 16, 18, 26, 25, 30, 35, 5, 30, 7, 6, 11, 14, 12, 10, 17, 10, 9, 12, 13, 10, 7.

Make a split stem-and-leaf plot. From the plot, what can you say about the center, spread and shape (symmetry) of this data.

2.7 The following data gives the weight gain (in ozs.) of 10 rabbits in a nutrition experiment.

10, 8, 12, -2, 8, 8, 9, 11, 5, 4

(a) Construct a boxplot of the data.

(b) Give a measure of center and a measure of spread based on this boxplot.

2.8 The following data gives the waiting times in minutes of 15 calls to a company's customer service telephone line.

10, 1, 2, 9, 5, 3, 4, 1, 0, 5, 6, 5, 3, 9, 3

(a) Find the "five-point" summary and draw a box-plot for this data.

(b) Calculate the sample mean and sample standard deviation for this data.

2.9 When you drive on a certain highway at 55 miles per hour, you pass nearly the same number of cars as the number that pass you in the same direction. Is 55 the mean, median or the mode of speed on that highway?

2.10 Albert Michelson (1852-1931) won a Nobel Prize in physics in 1907 for development and use of precision optical instruments. In 1882, he obtained the following measurements on the speed of light (in kms. per second):

299,883	299,796	299,611	299,781	299,774	299,696	299,748	299,809
299,816	299,682	299,599	299,578	299,820	299,573	299,797	299,723
299,778	299,711	300,051	299,796	299,772	299,748	299,851	

(a) Find the mean and variance of this data.

(b) Suppose that new measurements are obtained by subtracting the same constant from each of the measurements, i.e., say

$$y_i = x_i - c,$$

Show that the sample variance of the y-values is exactly the same as the sample variance of the x-values, but the mean, $\overline{y} = \overline{x} - c$.

(c) Subtract 299,000 from each of the measurements on the speed of light and recalculate the sample mean and variance and relate to part (b).

2.11 A set of 10 observations have an $\overline{x}$ of 40 and s of 9 while a different set of 10 have a mean of 48 and s of 7. Find the mean and s for the combined set of 20 observations.

2.12 (a) Prove Fact 2.1.

(b) Show that the two formula for the sample variance are equal, i.e.,

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

2.13 The ratings of several football players is summarized in the following stem and leaf plot (with colon representing the decimal):

```
79 : 3 7
80 : 2 3 5 5
81 : 6 6 7 8
82 : 6 6
83 : 4
84 : 8
85 :
86 : 6
```

(a) What is the average rating of a football player in the sample?

(b) What is the median rating of a football player?

(c) Which of these measures is preferable in this context and why?

(d) Construct a boxplot of the data.

2.14 A mathematics achievement test consisting of 100 questions was given to 25 sixth grade students at Maple Elementary school. The following data shows the number of questions answered correctly by each student:

49	61	51	57	49	55	69	88	89	55	77	51	61
68	54	67	57	63	71	77	84	79	75	65	50	

(a) Prepare frequency table with class width 10, starting from 45.

(b) Plot a relative frequency density histogram for it.

(c) Use the frequency table tocalculate $\overline{x}$ and standard deviation.

2.15 A TA is looking at a student's scores for 6 quizzes. The average of the 6 quiz scores is 9 with a standard deviation of 1. Next day the student takes Quiz # 7 and receives a score of 9.

(a) What is the average and the standard deviation of the 7 quizzes?

(b) How well must the student do on Quiz # 8 in order to keep 9 as his average score?

2.16 A city administrator is investigating the efficiency of its Fire Department. The time, in minutes, taken by the Fire Department to respond to 15 alarms, is given below:

1, 1, 2, 2, 6, 9, 4, 1, 7, 10, 2, 4, 5, 10, 5

(a) Find the "five-number summary" and draw a boxplot for this data.

(b) Calculate the sample mean and sample standard deviation for this data.

2.17 Find (a) mean (b) median (c) mode and (d) SD of the following 10 observations:

6, 8, 22, 3, 18, 11, 14, 12, 11, 16

2.18 Draw a stem and leaf plot for the following enrollment data for a statistics class:

36	28	33	41	39	36	31	35	33	35	39	41	36
30	38	37	41	44	46	38	33	35	47	27	36	

2.19 Create a stem-and-leaf display, five-number summary and boxplot for these data on reaction times in a psychology experiment.

20.5	18.7	17.2	18.5	19.2	16.7	19.0	21.6	28.3	21.2
18.3	25.3	17.9	19.9	18.9	19.0	15.7	20.3	22.6	18.1
23.0	18.4	20.4	19.3	22.5	21.9	19.2	18.0	19.0	19.5

2.20 What are the mean, median and mode of the following distributions? Which measure of central tendency best reflects each distribution and why?

(a) 1, 1, 1, 100, 100, 100

(b) 2, 3, 4, 6, 100

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

(d) 1, 1, 1, 1, 1, 1, 5, 5, 6, 8, 9

2.21 If $\sum x^2 = 151,\ n = 5$ and $s^2 = 6.5$, what is the mean of the sample?

2.22 The mean advanced mathematics achievement for male and female students in 16 countries having taken advanced mathematics in their final year of secondary school (1994–1995) are listed in the table below. (*Source:* Mulis et al. (1998), Mathematics and Science Literacy in the Final Year of Secondary School, Chestnut Hill, MA.)

Country	Male	Female
Greece	516	505
Australia	531	517
Cyprus	524	509
Demark	529	510
Slovenia	484	464
Sweden	519	496
France	567	543
Canada	529	489
Lithuania	542	490
Russia	569	516
Switzerland	559	503
Italy	484	460
Germany	484	452
Austria	486	406
Czech Republic	524	432
United States	457	426

Draw separate side-by-side boxplots for males and for females and comment if you think there is any ”gender gap” in mathematics.

Chapter 3

Probability ideas

3.1 Introduction

Any discussion of probability starts with what we call a **random experiment**. A **random experiment** is one for which we know in advance, the set of all possibilities or **outcomes** - but we can not predict which of these **outcomes** would occur in any specific trial. The set of all possible **outcomes** is called the **sample space**. We give below a few simple examples of random experiments and their sample spaces:

Example 3.1: Toss a coin and see if it turns up heads or tails.

Possible outcomes = {Head, Tail}.

◇

Example 3.2: Roll a die and observe the score on the top face.

Possible outcomes = {1, 2, 3, 4, 5, 6}.

◇

Example 3.3: Throw a basketball and keep track of how many attempts it takes for the 1st successful basket.

Possible outcomes = {1, 2, 3, 4, . . . }.

◇

Example 3.4: Wait for a taxi in a new town and record the time it takes for the taxi to arrive. The outcomes are possible times t, say in seconds and can be represented by the set

$$\text{Possible outcomes} = \{t : t \geq 0\}.$$

$\diamond$

An **event** is a combination of one or more outcomes. It is typical to denote events by capital letters, $A,\ B,\ \ldots$ etc. . In Example 3.2, we may be interested in the event that the "score is odd". Since the score is odd when we have 1, 3 or 5, we may represent this event as consisting of these three outcomes

A= Score is odd = {1, 3, 5}.

Event B that the score is greater than or equal to 5 is represented by

B = {5, 6 }.

We would like to define probabilities for various events. To be consistently defined, such probability should satisfy the following 3 basic rules, called the axioms:

Rule 1.	Probability of an event A, denoted by, P(A), is a number between zero and one.
Rule 2.	In any random experiment, probability of all the outcomes added together, is one.
Rule 3.	Probability of (A or B) is the sum of the individual probabilities if these events $A,\ B$ have no common outcomes.

Rule 3 can be restated as:

$$\boxed{\text{P}(A \text{ or } B) = \text{P}(A) + \text{P}(B), \text{ if they do not overlap.}}$$

The third axiom or rule can be generalized for any collection of events -not just two, but

we shall keep out of such trouble, for the sake of simplification. A pictorial representation of sets and outcomes is through the so called Venn Diagram:

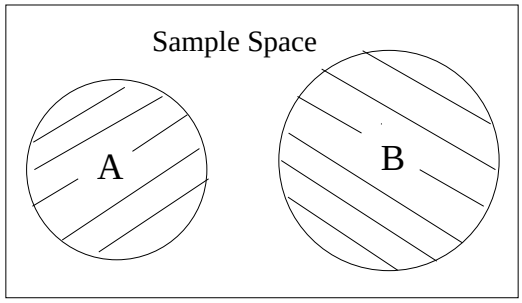


Figure 3.1: Venn Diagram for two events A and B with no common outcomes

Recall the die rolling Example 3.2, where there are 6 outcomes. Since we assumed that the die is **fair**, all scores are **equally likely** i.e., any score has the same probability as any other. However, Rule 2 stipulates that all the probabilities have to add to one. Therefore, it follows that

$$\mathrm{P}(1) = \mathrm{P}(2) = \cdots = \mathrm{P}(6) = \tfrac{1}{6}.$$

From this and by following Rule 3, we can now compute the probability of any event of interest. In particular,

$$\mathrm{P}(A) = \mathrm{P}(\text{ score is odd }) = \mathrm{P}(\{1,3,5\}) = \mathrm{P}(1) + \mathrm{P}(3) + \mathrm{P}(5) = \tfrac{1}{6} + \tfrac{1}{6} + \tfrac{1}{6} = \tfrac{3}{6},$$

while

$$\mathrm{P}(B) = \mathrm{P}(5) + \mathrm{P}(6) = \tfrac{2}{6}.$$

Not always will two events be non-overlapping or **disjoint**. When two events A and B have common outcomes, we can represent them as follows:

Some consequences of the three basic rules or axioms provide us with further "extended rules", 4 and 5:

Rule 4.	$\mathrm{P}(\text{not } A) = 1 - \mathrm{P}(A)$
Rule 5.	$\mathrm{P}(A \text{ or } B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \text{ and } B)$

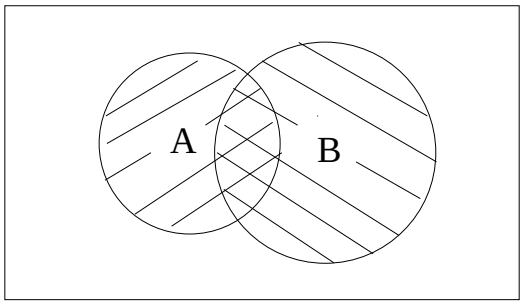


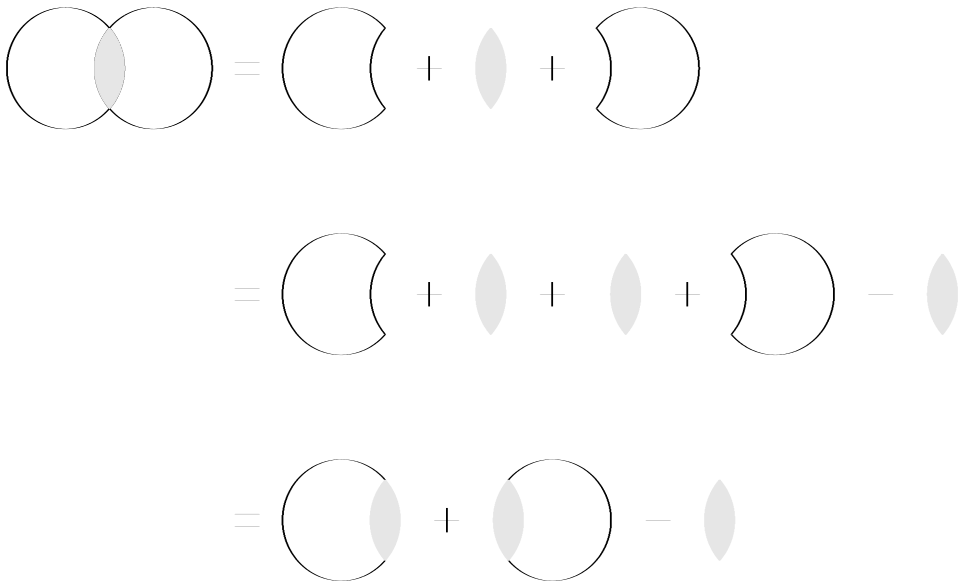
Figure 3.2: Venn Diagram when A and B overlap

Observe that A and "not A" do not have any common outcomes and they make up the whole space. So by Rules 2 and 3,

$$\mathrm{P}(A) + \mathrm{P}(\text{ not } A) = \mathrm{P}(\text{all outcomes}) = 1,$$

from which the Rule 4 follows. A "proof" of Rule 5 is as follows:

Note that (A or B) is



Therefore,

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

In Example 3.2, since A and B refer to the outcome 5 which has probability $\frac{1}{6}$, we get

$\mathrm{P}(A \text{ or } B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \text{ and } B) = \frac{3}{6} + \frac{2}{6} - \frac{1}{6} = \frac{4}{6}$

Example 3.5: Suppose we select a 2 digit random number. Find the probability it is divisible by either 3 or 4.

Solution: There are altogether 100 outcomes, all of them equally likely, namely

outcomes = {00, 01, 02, 03, . . . , 99}.

Define the events

A = selected number is divisible by 3 = {00, 03, 06, 09, 12, . . ., 99}.

$\mathrm{P}(A) = \frac{34}{100}$

B = selected number is divisible by 4.

$\mathrm{P}(B) = \frac{25}{100}$

$\mathrm{P}(\text{not } A) = 1 - \mathrm{P}(A) = 1 - \frac{34}{100} = \frac{66}{100}$

$\mathrm{P}(A \text{ and } B) = \mathrm{P}(\text{number is divisible by 3 and 4})$
$= \mathrm{P}(\text{number divisible by 12}) = \frac{9}{100}$. Therefore

$\mathrm{P}(A \text{ or } B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \text{ and } B) = \frac{34}{100} + \frac{25}{100} - \frac{9}{100} = \frac{50}{100} = 0.5$.

◇

Remark 3.1 In such cases as in Example 3.2 and Example 3.5, where the total number of outcomes is finite and they are all equally likely, then probability is obtained as a ratio of the number of favorable outcomes to the total number of outcomes, by using the formula

$$\boxed{\mathrm{P}(A) = \frac{\text{number of outcomes that make up the event } A}{\text{total number of outcomes in the experiment}}.}$$

Example 3.6: If three people are selected at random, the chance that they all have the same birthday is $\frac{365}{365\times365\times365} = \frac{1}{(365)^2}$, observing there are 365 days in a year (ignoring leap years) and these are equally likely.

◇

Remark 3.2 None of these rules tell us exactly how we assign probabilities in a given experiment. Such probability assignment is done using known physical or natural laws that govern the experiment (e.g., a coin from a US mint is typically fair, certain combination of atmospheric conditions lead to rain, etc.) or empirical evidence, i.e., past experience. Sometimes probabilities may be merely **subjective**, i.e., expert or educated guesses.

3.2 Conditional probability and independence

Sometimes, in a random experiment, there may be partial information available on the outcome, so that the set of possible outcomes is no longer the original larger space that we started with.

Example 3.2 (contd.): Suppose we define the 2 events:

A= Score is odd = { 1, 3, 5 },

B= Score is 2

and C= Score is 3.

P(B) is 1/6 since there are six possible outcomes to start with. But if we are given the event A has occurred, i.e., we are told an odd score is the outcome, then the **conditional probability** of B, given A has happened, denoted by P($B|A$) (the vertical bar inside the probability statement is read "given"), is 0 since the score can not be 2 any longer. In fact, the occurrence of A makes our new sample space of possible outcomes to be { 1, 3, 5 }. For instance, P($C|A$) = 1/3 since it is one of the possible 3 outcomes that can happen. Thus the conditional probability of B given A is the chance of the outcomes in B relative to those in A or

$$\boxed{\text{P}(B|A) = \text{P}(A \text{ and } B)/\text{P}(A).}$$

◇

Referring to the Venn Diagram 3.2 is helpful. Originally the rectangular box represented the set of all outcomes but once we are given A, the set of outcomes are restricted to those

inside the circle A. In the above example, since the numerator $P(A \text{ and } B) = 0$, so is $\mathrm{P}(B|A)$.

Example 3.6: Suppose we throw a pair of fair dice and observe the 2 scores. Let

A=first die comes up with score 2

B=second die comes up with score 3

C=sum of the two scores is 5

D=sum of the two scores is less than or equal to 5.

Find $\mathrm{P}(A)$, $\mathrm{P}(A|B)$, $\mathrm{P}(A|C)$ and $\mathrm{P}(C|D)$.

Solution: There are 36 equally likely outcomes. (See Example 4.2 later for the complete list of these 36 outcomes i.e., the sample space). From there, it can be found that

$\mathrm{P}(A) = 6/36 = 1/6,$

$\mathrm{P}(A|B) = \mathrm{P}(A \text{ and } B)/\mathrm{P}(B) = (1/36)/(1/6) = 1/6,$

$\mathrm{P}(A|C) = \mathrm{P}(A \text{ and } C)/\mathrm{P}(C) = (1/36)/(4/36) = 1/4,$

$\mathrm{P}(C|D) = \mathrm{P}(C \text{ and } D)/\mathrm{P}(D) = (4/36)/(10/36) = 4/10 = 2/5.$

$\diamond$

Which conditional probabilities are of interest depends on what we already know. In an interesting recent court case, a lawyer for the defense stated that the odds that a wife-batterer actually kills his wife is quite small— something like 1 in a 1000. This is indeed a small fraction if one considers the number of cases where the husband is guilty of murdering his wife as a proportion of all wife-battering cases. If on the other hand, we are given the additional information that she is not only a battered wife, but was also killed by somebody, then the chance that the husband did it, is quite high —something of the order of 1 in 3. With this additional information that she is also killed, we restrict our consideration only to those wife-battering cases where the wife is also killed, to see how frequently the husbands are guilty of the crime. And that, according to some statistics, is not small at all !

If the chance of an event A is the same as the conditional chance of A given another event B, it means that the occurrence (or non-occurrence) of B does not affect the chance of A.

Then we say A is **independent** of B. It can be seen that this is a symmetric relationship and we have

A and B are independent if $\mathrm{P}(B|A) = \mathrm{P}(B)$
or, $\mathrm{P}(A|B) = \mathrm{P}(A)$
or, $\mathrm{P}(A \text{ and } B) = \mathrm{P}(A)\mathrm{P}(B)$.

All the three statements are equivalent, i.e., if any of one of them is true, then all three statements are true. For example,

$$\mathrm{P}(B|A) = \mathrm{P}(B)$$

$$\Leftrightarrow \quad \frac{\mathrm{P}(A \text{ and } B)}{\mathrm{P}(A)} = \mathrm{P}(B)$$

$$\Leftrightarrow \quad \mathrm{P}(A \text{ and } B) = \mathrm{P}(A) \times \mathrm{P}(B).$$

This leads to another extended rule, the so called **multiplication rule** for independent events:

Rule 6. $\mathrm{P}(A \text{ and } B) = \mathrm{P}(A) \times P(B)$, if A and B are independent.

Example 3.6 (contd.): In this example, it is seen that A and B are independent since

$\mathrm{P}(A|B) = \mathrm{P}(A) = 1/6$.

Knowing that the score on the second die is 3 does not change the chance of getting score 2 (or any other event) relating to the first die. The multiplication rule for such independent

events allows us to say that

$$\begin{aligned} & \text{P(score 4 on first die and score 6 on the second die)} \\ = \ & \text{P(score 4 on the first die)} \times \text{P(score 6 on the second die)} \\ = \ & 1/6 \times 1/6 \\ = \ & 1/36. \end{aligned}$$

◇

Independence of events is what allows us to say, for instance, that if the chance of making a "free throw" is 0.2 each time for a particular basketball player, then in 3 attempts,

$$\begin{aligned} \text{P(basket, no basket, basket)} = \ & \text{P(basket)} \times \text{P(no basket)} \times \text{P(basket)} \\ = \ & (0.2)(0.8)(0.2) \\ = \ & 0.032. \end{aligned}$$

3.3 Bayes theorem

A collection of events ($B_1,\ B_2,\ \ldots,\ B_k$) is said to be **mutually exclusive** if no two of them overlap. If, at the same time, they also cover the entire set of outcomes i.e., together add up to the whole sample space, they are called **collectively exhaustive** events. Such events **partition** the whole space just like walls partition the living area of a typical home. In such a case, we can write the probability of any event A as the sum of the probabilities of (A and B_1), (A and B_2), $\ldots$ (A and B_k), since when A occurs, it has to occur in conjunction with either B_1 or with B_2 or ... with B_k. Using the definition of conditional probability, $\text{P}(A \text{ and } B_i) = \text{P}(A|B_i)\text{P}(B_i)$, we thus have the formula for **total probability**

$$\begin{aligned} \text{P}(A) &= \text{P}(A \text{ and } B_1) + \text{P}(A \text{ and } B_2) + \cdots + \text{P}(A \text{ and } B_k) \\ &= \text{P}(A|B_1)\text{P}(B_1) + \text{P}(A|B_2)\text{P}(B_2) + \cdots + \text{P}(A|B_k)\text{P}(B_k). \end{aligned}$$

In other words, if we know the probability of an event A conditional on other events and we

also know what proportion of the time these conditions hold, we can put these together to find $P(A)$.

Example 3.7: A microchip manufacturing plant has 3 machines that produce the chips. Machine 1 produces 30% of the output and of these, 2% are defective; Machine 2 produces 45% of the output and of these, 1% are defective; Machine 3 produces the remaining 25% of the chips and of these, 3% are defective. Find the probability that a randomly selected chip from this company is defective.

Solution: Define the events

A= chip is defective

B_1= chip is made by Machine 1

B_2= chip is made by Machine 2

B_3= chip is made by Machine 3.

Then, using the formula for total probability given above,

$$\begin{aligned} & \mathrm{P}(\text{a randomly selected chip is defective}) \\ = \ & \mathrm{P}(A) \\ = \ & \mathrm{P}(A|B_1)\mathrm{P}(B_1) + \mathrm{P}(A|B_2)\mathrm{P}(B_2) + \mathrm{P}(A|B_3)\mathrm{P}(B_3) \\ = \ & (0.02)(0.30) + (0.01)(0.45) + (0.03)(0.25) \\ = \ & 0.018. \end{aligned}$$

◇

The so called **Bayes formula** (due to Reverend Thomas Bayes (1702-1761)), allows us now to calculate some "inverse" probabilities. Since, by the definition of conditional proability,

$$\mathrm{P}(B_i|A) = \mathrm{P}(A \text{ and } B_i)/\mathrm{P}(A)$$

and we can write the numerator as the product of $\mathrm{P}(A|B_i)\mathrm{P}(B_i)$ while the denominator has the total probability formula, we obtain

$$\mathrm{P}(B_i|A) = \frac{\mathrm{P}(A|B_i)\mathrm{P}(B_i)}{\mathrm{P}(A|B_1)\mathrm{P}(B_1) + \mathrm{P}(A|B_2)\mathrm{P}(B_2) + \ldots + \mathrm{P}(A|B_k)\mathrm{P}(B_k)}.$$

This formula allows us to find the conditional probability of B_i given A using the other type of conditional probabilities namely, $\mathrm{P}(A|B_i)$.

Example 3.7 (contd.): If a chip selected from this factory's output turns out to be defective, find the probability that it was made by Machine 3.

Solution: In our notations, we need $\mathrm{P}(B_3|A)$, which, by the Bayes formula is

$$\mathrm{P}(B_3|A) = \frac{\mathrm{P}(A|B_3)\mathrm{P}(B_3)}{\mathrm{P}(A|B_1)\mathrm{P}(B_1) + \mathrm{P}(A|B_2)\mathrm{P}(B_2) + \mathrm{P}(A|B_3)\mathrm{P}(B_3)} = \frac{0.0075}{0.018} = 0.42.$$

◇

EXERCISES

3.1 California's lotto involves picking six numbers without repetition from 1 to 51. If you buy a ticket and pick a set of six numbers, what is the probability that you select the six winning numbers?

3.2 A student's chance of passing a statistics test is 0.8 if he studies and only 0.2 if he does not study before the test. If the chance that this student studies is 0.6, find the probability

(a) that he will study and pass the course;

(b) that he will pass this course;

(c) that he studied given that he passes the course.

3.3 On a day in December, the probability that it will snow in Boston is 0.4 and the probability that it will snow in Moscow is 0.7. Assuming independence, find the probability that it will snow:

(a) in both cities.

(b) in neither city.

(c) only in Moscow.

(d) in exactly one city.

3.4 $A,\ B$ are two events such that $\mathrm{P}(A) = 0.5,\ \mathrm{P}(B) = 0.2$. Find $\mathrm{P}(A \text{ or } B)$ if A and B are (a) disjoint (b) independent.

3.5 Suppose a weather bureau reports that the chance of precipitation tomorrow is 70%, while the chance that the temperature will be less than 32^oF is 25%. If these two events are independent, what is the probability that tomorrow it will

(a) snow? (assume that the precipitation turns into snows at a temperature less than 32°F.)

(b) rain?

(c) be less than 32°F but not snow?

(d) be over 32°F but not rain?

3.6 A die is loaded so that any even score is twice as likely as any odd score. Find

(a) $\mathrm{P}(\text{score} \geq 4)$

(b) P(odd score).

3.7 A committee of four people is selected from a group of five men and six women.

(a) How many ways can this be done?

(b) What is the probability that a randomly selected committee has members all of the same sex?

3.8 Assume that all the 12 months in a year are equally likely to be the month of birth. If we selected 3 people at random, what is the probability that

(a) all 3 have the same birth month

(b) all of them are born in July

(c) all 3 have distinct birth months

(d) at least 2 of them have the same month of birth

3.9 Three students A, B, C can solve problems in a statistics book with probabilities 0.6, 0.8 and 0.3 respectively. Assuming that they all work independently, for a randomly selected problem, find the probability that

(a) all three can solve it

(b) A and B can solve it but not C

(c) either A or B can solve it

(d) at least one of the 3 can solve it

3.10 Let A and B be two events such that $\mathrm{P}(A) = 0.5$, $\mathrm{P}(A \text{ or } B) = 0.7$. Find $\mathrm{P}(B)$ if

(a) A, B are independent events.

(b) A, B are disjoint events.

(c) $\mathrm{P}(A|B) = 0.5$.

3.11 Assume that the probability is 0.95 that a jury selected to try a criminal case will arrive at the correct verdict, i.e., find one guilty if actually guilty and find innocent if actually innocent. Suppose the police in a certain town are quite careful in their investigations and 99% of the people brought for trial before a jury are actually guilty. Find the probability that

(a) the jury finds a defendant guilty

(b) a defendant is actually innocent given the jury finds him guilty.

3.12 Two different suppliers A and B provide a manufacturer with the same part. All the supplies of this part are kept in a large bin. In the past, 5% of the parts supplied by A and 9% of the parts supplied by B are defective. A supplies four times as many parts as B. Suppose you reach into the bin and select a part and find it is nondefective. What is the probability that it was supplied by A?

3.13 In the parking lot of a high tech company, 35% of the cars parked are US-made and of these, 20% are luxury cars; 40% are European-made and of these, 40% are luxury cars; the other 25% are Japanese-made and of these, 15% luxury cars. Find the probability that (a) if a car is picked at random in this lot, it is a luxury car (b) if a car is luxury model, find the probability that it is European-made.

3.14 A fair coin is flipped three times. Let A denote the event "head occurs on the first flip" and B, the event "the same face does not occur on all three tosses".

(a) Write down the sample space for this experiment as well as the sample points that correspond to the two events A and B.

(b) Are A and B independent events? Are they mutually exclusive?

(c) Find P(A and (not B)) and P(not $A|B$).

3.15 A student has 4 pairs of socks in his dresser, each pair a different color. He has to dress before dawn without waking his roommate and so he grabs a pair of socks without seeing them and puts them on. What is the probability that they are of matching color?

3.16 A deck of playing cards consists of 52 cards = 4 suits x 13 ranks. A poker hand consists of five different cards, chosen so that any five cards are equally likely. Clubs is one of the suits with 13 of them in the deck. What is the probability that a poker hand will consist of all clubs?

3.17 Five people get into an elevator on the ground floor of a building which has 5 upper floors. What is the probability that they all get off on different floors?

Chapter 4

Random variables

A **random variable** is a numerical aspect or measure of the outcome in a random experiment. Random variables (r.v.'s) are generally denoted by capital letters, X, Y, etc. Notice that the outcomes of a random experiment themselves need not be numerical values. For instance, the random experiment may be to "select 13 cards from a a deck of 52 cards" and the sample space consists of all possible subsets of the 13 cards. One numerical aspect of interest in this case may be

X= the number of aces among the 13 selected cards.

Or the random experiment may be to "select a student from a class of 100 students" and we may be interested in

Y = height of the selected student.

Notice that the random variable X defined above can take only 5 discrete values: 0, 1, 2, 3 and 4, whereas Y can take all values continuously in a given of range of height. This property distinguishes how we deal with the calculation of probabilities corresponding to these random variables. We now discuss these two basic types of random variables:

4.1 **Discrete** random variables

4.2 **Continuous** random variables

4.1 Discrete random variables

A **discrete** random variable is one which takes a finite (or a countable) number of values. A list of such values and the corresponding probabilities is called the **probability distribution** of the r.v.

Values of x	x_1	x_2	. . .	x_n
Probabilities	p_1	p_2	. . .	p_n

These $p_i = \mathrm{P}(X = x_i)$, being probabilities, satisfy the conditions $p_i \geq 0$ and $\sum_{i=1}^{n} p_i = 1$.

Example 4.1: Toss a coin 3 times. There are 8 ($= 2 \times 2 \times 2$) equally likely outcomes (with probability $\frac{1}{8}$ each) namely,

outcomes = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}.

Let

X= # of heads in the 3 tosses.

Then we can see that

$\mathrm{P}(X = 0) = \mathrm{P}(\#\text{ of heads} = 0) = \mathrm{P}(\mathrm{TTT}) = \frac{1}{8}$.

$\mathrm{P}(X = 1) = \mathrm{P}(\#\text{ of heads} = 1) = \mathrm{P}(\mathrm{HTT},\ \mathrm{THT},\ \mathrm{TTH}) = \frac{3}{8}$.

$\mathrm{P}(X = 2) = \mathrm{P}(\#\text{ of heads} = 2) = \mathrm{P}(\mathrm{HHT},\ \mathrm{HTH},\ \mathrm{THH}) = \frac{3}{8}$.

$\mathrm{P}(X = 3) = \mathrm{P}(\#\text{ of heads} = 3) = \mathrm{P}(\mathrm{HHH}) = \frac{1}{8}$.

All this can be summarized in the following table which lists possible values of X and the corresponding probabilities:

x	0	1	2	3
$\mathrm{P}(X = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

$\diamond$

Example 4.2: Consider a slightly more complex random experiment which is to "roll a fair die twice". Then

$$\text{outcomes} = \left\{ \begin{array}{llllll} (1,1), & (1,2), & (1,3), & (1,4), & (1,5), & (1,6), \\ (2,1), & (2,2), & (2,3), & (2,4), & (2,5), & (2,6), \\ (3,1), & (3,2), & (3,3), & (3,4), & (3,5), & (3,6), \\ (4,1), & (4,2), & (4,3), & (4,4), & (4,5), & (4,6), \\ (5,1), & (5,2), & (5,3), & (5,4), & (5,5), & (5,6), \\ (6,1), & (6,2), & (6,3), & (6,4), & (6,5), & (6,6) \end{array} \right\} .$$

Since this is a fair die, all these 36 outcomes are equally likely and the chance of any particular outcome is $\frac{1}{36}$. Define the random variable:

X= total of the two scores.

This random variable can take values from 2 through 12 and the probabilities corresponding to each of these values can be computed. For instance, the total score is 5 corresponds to the outcomes {(1, 4), (2, 3), (3, 2), (4, 1)} and hence,

P(total score is 5) = P (X= 5) = $\frac{4}{36}$.

It can be checked that the **Probability distribution of X** is given by the following table.

x	2	3	4	5	6	7	8	9	10	11	12
P(X = x)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Note that a total score of 7 has the largest probability, with the probabilities decreasing symmetrically for values on either side of 7.

◇

4.2 Continuous random variables

A **continuous random variable** takes all the values in a given interval. For instance, we may select an individual at random and measure his or her height, say X or the weight, Y.

It is clear that X can take all the values within the range of say 0 to 8 feet, not just integer values like 0, 1, 2, . . . as in Examples 4.1 and 4.2.

Continuous random variables take too many (an uncountable number of) values and this makes it impossible for us to make a list of values and assign probabilities to each individual value. Instead, in the continuous case, we define probabilities through what is known as a **density curve**. This is a curve above the x-axis and with a total area of 1 under it. Probabilities for intervals are defined as areas under such a density curve, i.e.,

$$\mathrm{P}(a < X \leq b) = \text{Area under the density curve between } a \text{ and } b.$$

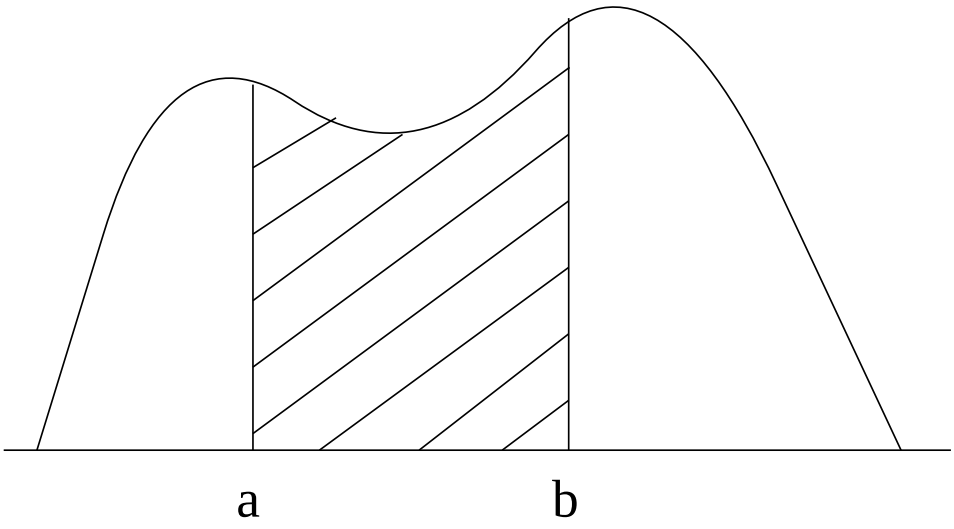


Figure 4.1: Continuous density curve

Example 4.3: Let X be the time (in minutes) taken for a fire truck to arrive after one dials 911. Suppose this time can equally well be anywhere between (2, 15) minutes. Such a situation where all the values within a range are equally likely for a continuous random variable, is called a continuous **uniform distribution** and the random variable is called a continuous **uniform random variable**. The density curve has the same height and has a rectangular shape. In this example, since the base is of length 13 units, the height of the density curve is $\frac{1}{13}$, so that we have

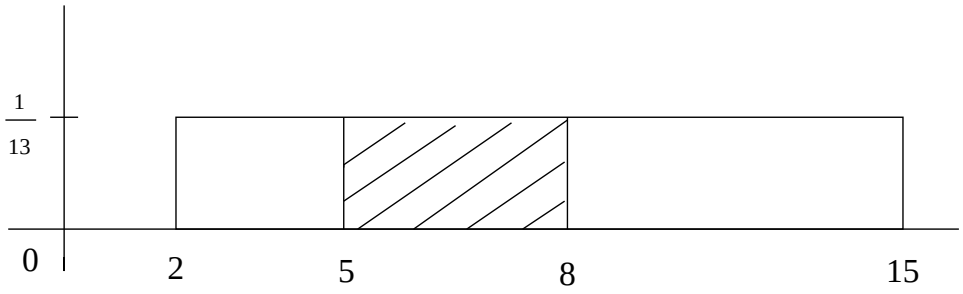


Figure 4.2: Uniform distribution

$$\mathrm{P}(5 < X \leq 8) = (\text{Area between 5 and 8}) = 3 \times \tfrac{1}{13} = \tfrac{3}{13}$$

whereas,

$$\mathrm{P}(X \leq 6) = \mathrm{P}(2 < X \leq 6) = 4 \times \tfrac{1}{13} = \tfrac{4}{13} .$$

◇

4.3 Mean and Variance of a random variable

Suppose X is a discrete random variable with a given probability distribution. We know it can take different values with different probabilities. One can define the center of such a probability distribution called the average, the mean or the expected value of the random variable. It is denoted by μ for the **mean** or $\mathrm{E}(X)$ for the **expected value** of the random variable X and is given by

$$\boxed{\mu = (x_1 \times p_1) + (x_2 \times p_2) + \cdots + (x_n \times p_n) = \sum_{i=1}^{n} x_i p_i.}$$

The **mean** represents the "long-run average" value of the random variable X. That is, if we keep a record of the actual values the random variable takes and average these over a large number of times, it will be very close to μ. Intuitively, this makes sense as a "weighted average" with the weights being the probabilities or the proportion of times the random variable takes a particular value.

One can also define a measure of **spread** around this mean by calculating the deviations from the mean $(x_i - \mu)$, squaring them and averaging these with the same weights, p_i. This process of taking deviations from the mean and averaging their squares is similar to calculating the sample variance s^2. This gives us the **variance** of the r.v. X, denoted by σ^2,

$$\begin{aligned} \mathrm{Var}(X), \sigma^2 &= (x_1 - \mu)^2 \times p_1 + (x_2 - \mu)^2 \times p_2 + \cdots + (x_n - \mu)^2 \times p_n \\ &= \sum_{i=1}^{n} (x_i - \mu)^2 \times p_i. \end{aligned}$$

It can be shown that

$$\sigma^2 = \sum x_i^2 p_i - \mu^2,$$

so that once we have μ, it suffices to calculate $\sum x_i^2 p_i$. The positive square root of this variance, i.e., the σ, is called the **standard deviation (SD)** of the random variable X.

Example 4.1 (contd.): Recall X= number of heads in 3 tosses of a fair coin. From the probability distribution given before, we have mean

$$\mu = 0 \times \tfrac{1}{8} + 1 \times \tfrac{3}{8} + 2 \times \tfrac{3}{8} + 3 \times \tfrac{1}{8} = \tfrac{12}{8} = \tfrac{3}{2} = 1.5.$$

Thus, if we repeat the "experiment of tossing a fair coin 3 times" a very large number of times and record the number of heads each time, the average of the number of heads will be very close to 1.5.

The variance

$$\begin{aligned} \sigma^2 &= (0 - 1.5)^2 \times \frac{1}{8} + (1 - 1.5)^2 \times \frac{3}{8} + (2 - 1.5)^2 \times \frac{3}{8} + (3 - 1.5)^2 \times \frac{1}{8} \\ &= \frac{6}{8} = \frac{3}{4} = 0.75 \end{aligned}$$

Alternatively, we can also calculate σ^2 by noting

$$\sum x_i^2 p_i = 0^2 \times \frac{1}{8} + 1^2 \times \frac{3}{8} + 2^2 \times \frac{3}{8} + 3^2 \times \frac{1}{8} = \frac{24}{8} = 3,$$

so that $\sigma^2 = 3$ - $(1.5)^2 = 0.75$.

The standard deviation, $\sigma = \sqrt{0.75} = 0.866$.

$\diamond$

One can define the mean μ (as a measure of center or the center of gravity) and the variance σ^2 (as a measure of spread) for continuous r.v.s also, but this involves calculus and is beyond the scope of this book. However, one may see for instance, in Example 4.3, that the center of gravity is at the middle of the range, so that

$\mu = \frac{2+15}{2} = 8.5$ minutes.

EXERCISES

4.1 The number of ice cubes put into a soft drink glass in a restaurant is a random variable X with the following probability distribution:

x	1	2	3	4	5
$\mathrm{P}(X = x)$	0.1	0.2	0.4	0.2	0.1

Find

(a) $\mathrm{P}(X \geq 2)$,

(b) the mean and the standard deviation of the random variable X.

4.2 The number of passengers in a car has the following probability distribution:

x	1	2	3	4	5
$\mathrm{P}(X = x)$	0.4	0.3	0.15	?	0.05

(a) Find the probability that there are exactly 4 passengers in a car.

(b) What is the probability of at most 4 passengers in a car?

(c) What is μ? How do you interpret this number?

(d) Find the standard deviation of the number of passengers in a car.

4.3 The number of defects (denoted by X) in a new car from a certain manufacturer has the following probability distribution:

x	0	1	2	3	4
$P(X = x)$	0.50	0.30	0.10	?	0.05

(a) Find the probability that there are exactly 3 defects in a new car.

(b) What is the probability of at most 2 defects in a new car?

(c) What is the mean and standard deviation of the number of defects?

4.4 If X is the difference in scores when a pair of fair dice are rolled, find its probability distribution. What is the mean μ and standard deviation σ for X?

4.5 A statistics professor has determined from past experience the following probability distribution for X, the GPA (with A = 4, ... , F = 0):

x	0	1	2	3	4
$P(X = x)$	0.10	0.15	0.25	?	0.15

(a) Find P($X = 3$) and P($X \leq 1$).

(b) Find the mean and the standard deviation for the GPA in this course.

4.6 In the Goren point-count system for bidding in contract bridge, cards are assigned points as follows (there are 52 cards in a deck):

Ace	King	Queen	Jack	Other
4	3	2	1	0

(a) Give the probabilities that a card drawn at random will have 0, 1, 2, 3, and 4 points, respectively.

(b) Compute the mean value of the number of points assigned to a card selected at random.

(c) Find the average number of points in a hand of 13 cards dealt at random. (*Hint*: Determine the total number of points in the deck and decide how they would be distributed among the four players, on the average.)

4.7 In a lottery your winnings/return, say X, is a random variable. Suppose it has the following probability distribution:

x	0	100	1,000
$\mathrm{P}(X = x)$	0.998	0.001	0.001

(a) What is your expected winnings, $\mathrm{E}(X)$, in this situation?

(b) Find also the $\mathrm{Var}(X)$, i.e., σ^2.

(c) If a ticket for this lottery is priced at \$2.00, would it make sense to buy a ticket? How does the variance affect your decision?

4.8 The following table gives the probability distribution for the rating (the number of stars) X of hotels in a large town.

Book rating (x)	1	2	3	4	5
$\mathrm{P}(X = x)$	.125	.632	.093	.099	.051

(a) Find the probability that a randomly picked hotel has 4 or 5 stars.

(b) Find the mean rating and interpret this result.

(c) Find the standard deviation of hotel ratings.

4.9 In a certain state, historical records indicate that it takes anywhere from 1 to 5 attempts for a person to acquire his or her first driver's license. The following table represents the number of attempts X, needed in order to acquire the license.

Number of attempts (x)	1	2	3	4	5
$P(X = x)$	0.42	0.33	?	0.04	0.01

(a) Find the probability that one does not get a drivers licence in the first two attempts.

(b) Find the mean and the standard deviation for X, the number of attempts one needs.

4.10 An outdoor rock concert draws a "full-house" and the organizers make $30,000 profit if the day of the event is sunny, they make only a $15,000 profit if it is cloudy and they expect to lose $10,000 if it rains on the day of the event. Suppose the weather bureau predicts that there is a 20% chance of rain and 30% chance for a cloudy day and 50% chance for the day to be sunny.

(a) Write down the probability distribution for X, the profit the organizers make. What is the expected profit and its standard deviation?

(b) Would it be worthwhile for the organizers to buy an insurance policy costing $5,000 from an insurance company, if they cover the $10,000 loss in the event of rain? (Find the expected profit again, in this case)

(c) If $5,000 is too much, what is a reasonable premium to pay, to cover the loss in the event of rain?

4.11 A Pizza shop sells pizzas in four different sizes. If X denotes the size (diameter) of the pizza ordered, based on last years' sales, the following information was obtained:

x	12"	14"	16"	18"
$P(X = x)$	.15	.25	.40	.20

Find the mean and s.d. of the pizza size.

4.12 The number of calls X, to a 911 number in a small city has the following probability distribution:

x	0	1	2	3	4	5	6
$\mathrm{P}(X = x)$	.1	.2	.2	.3	.1	.05	.05

(a) Find the probability that there will be three or more calls on a given day.

(b) Find the expected number of calls μ and the s.d. σ.

4.13 A particular statistics professor always takes a few minutes extra before she dismisses the class. Suppose X denotes the extra time (in minutes) and it can equally well be anywhere between 0 and 8 minutes i.e., the density curve has a constant height between 0 and 8 minutes.

(a) What is the height of this density curve?

(b) What is the probability that she dismisses a class (i) before 3 minutes (ii) between 5 and 6 minutes, on a given day?

(c) What is the average length of extra time, μ ? What can you say about the standard deviation?

4.14 Suppose the amount of gasoline customers buy from a gas station (in number of gallons) is denoted by the random variable X. From past records, it was found that the density curve is a triangle, starting at height zero when x=0 and going with constant slope to a height of (1/10) at x=20.

(a) Draw a picture of this density curve and verify the total area is 1. (Recall the area of a triangle = (height x base)/2).

(b) Find the probability that a customer who pulls in buys, (i) less than 10 gallons (ii) between 12 and 15 gallons (iii) more than 16 gallons.

(c) Try and figure out the average number of gallons bought by a customer, remembering that the average μ corresponds to the center of gravity of this triangular shape. What can you say about the standard deviation of X?

Chapter 5

Binomial and Normal distributions

In this chapter, we consider two very basic probability distributions that are omnipresent in statistics. One of them is a discrete probability distribution — the so-called **Binomial distribution**, which arises in connection with counts and/or proportions. The other one we will discuss is a continuous distribution — called the **Normal distribution**, which is a typical model assumed in connection with continuous measurements like heights, weights, scores, etc.

5.1 Binomial distribution

The **Binomial distribution** is a very important and basic discrete probability distribution, which arises under the following three conditions:

(i) A random experiment has just 2 possible outcomes, which we label, for convenience as

Success, S and **Failure**, F.

(ii) The probability of success, denoted by p, $(0 < p < 1)$ is fixed from trial to trial. It is clear that

P(Failure) = 1 - P(Success) = 1 - p.

(iii) n independent trials of this experiment are conducted.

Then, we define

$X =$ **Number of successes** in these n trials.

The number of trials n and the probability of success on any single trial p, completely determine the situation and the probabilities of various events. We refer to this as Binomial$(n,\ p)$ or Bin$(n,\ p)$ distribution. Note that an outcome of these n trials can be written as a string of n symbols consisting of S or F. If k is any number, i.e., an integer between 0 and n, then the probability of any specific outcome with k successes and $(n-k)$ failures such as

$$\underbrace{\text{S S} \cdots \text{S}}_{k \text{ times}} \underbrace{\text{F F} \cdots \text{F}}_{n-k \text{ times}}$$

is

$$\underbrace{(p \times \cdots \times p)}_{k \text{ times}} \underbrace{((1-p) \times \cdots \times (1-p))}_{n-k \text{ times}} = p^k (1-p)^{n-k}.$$

The outcome given above has the k successes in the first k places and all the failures in the last $(n-k)$ places. If, on the other hand we ask for the probability of k successes out of n trials, no matter which k places they occur, we have to add the probabilities of all such outcomes (n-tuples) which have exactly k successes and $(n-k)$ failures in them. It turns out that there are $\binom{n}{k}$ of them each with the same probability $p^k(1-p)^{n-k}$, where

$$\binom{n}{k} = \text{number of ways of selecting } k \text{ places out of } n \text{ places} = \frac{n!}{k!(n-k)!},$$

$$\text{and} \qquad n! = n \cdot (n-1) \cdot \cdots \cdot 3 \cdot 2 \cdot 1 \ .$$

Thus, for instance, $\quad 3! = 3 \cdot 2 \cdot 1 = 6 \quad$ while $\quad 5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120.$

Also, $\binom{4}{2} = \dfrac{4 \cdot 3 \cdot 2 \cdot 1}{(2 \cdot 1)(2 \cdot 1)} = 6$, while $\binom{5}{2} = \dfrac{5!}{2!3!} = \dfrac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(2 \cdot 1)(3 \cdot 2 \cdot 1)} = 10.$

Thus, the Binomial probabilities are given by

$$\text{P}(k \text{ successes out of } n \text{ trials}) = \text{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k},$$
$$k = 0, 1, 2, \cdots, n,$$
$$0 < p < 1.$$

Table B in the Appendix lists these probabilities for $n = 0, 1, . \ . \ ., 20$ and $p = .1, .2, .25, .3, .4, .5, .6, .75, .7, .8, .9$.

Python Code Instruction:
Random variables following binomial distribution can be simulated in **Python** by the code below:

Syntax:

```
np.random.binomial(n, p, size)
```

- n (integer) is the total number of trials
- p (floating value in (0, 1) is the probability of success
- size (integer) represents the particular number of successes

R code instruction:
In **R** the computation becomes more succinct and accurate: probability of a certain point $P(X = a)$ for example can be computed by *dbinom*() function whose syntax is represented as following:

Syntax:

```
dbinom(x, size= , prob= )
```

- x is the particular number of successes

- *size* is total number of trials
- *prob* is the probability of success
  The output is the probability $P(X = x)$

Example 5.1: What is the probability that in a family with 5 children 2 are girls?
Solution: Here the basic trial or random experiment occurs when the parents have a child. Since the sex of the child is of interest, there are 2 possibilities. Let us label "female child" as a "Success". Then assuming boy/girl are equally likely,

$$p = \text{P(Success)} = 0.5\ , \qquad (1 - p) = 0.5$$

and n, the number of trials = 5. Then $X = \#$ of female children among the 5, is conceptually the same as the number of successes in the $n = 5$ trials. This has a Bin (5, 0.5) distribution. So, using the formula, we get

$$\text{P}(X = 2) = \binom{5}{2}(0.5)^2(0.5)^3 = \frac{10}{2^5} = \frac{10}{32} = \frac{5}{16}.$$

Also, the probability of having all girls in this family is

$$\text{P}(X = 5) = \binom{5}{5}(0.5)^5(0.5)^0 = \frac{1}{32},$$

which is also the chance of all boys, i.e., no girls since

$$\text{P}(X = 0) = \binom{5}{0}(0.5)^0(0.5)^5 = \frac{1}{32}.$$

◇

Python code:

```
import numpy as np  #import numpy package
girl_number=np.random.binomial(5,0.5,10000) #simulate bin(5,0.5) for 10000
    times
```

```
count=0 # number of 2 occurs in array girl_number
for i in range (10000):
    if girl_number[i]==2:  # check every element in girl_array if it's 2
        count=count+1
print('The simulated probability is', count/10000)  #probability that 2 are
    girls among the 5 kids
```

The output for this **Python** code gives::

```
The simulated probability is 0.3061
```

As we can clearly see, the simulated value is pretty close to the theoretical value since the simulation was done using a very large sample size (10,000) and it accurately reflects the $P(X = 2)$ for a population which follows a $Bin(5, 0.5)$.

R code:

```
# P(X=2)
dbinom(2,size=5, prob=0.5)
# P(X=5)
dbinom(5,size=5, prob=0.5)
# P(X=0)
dbinom(0,size=5, prob=0.5)
```

The output for this **R** code gives::

```
> dbinom(2,size=5, prob=0.5)
[1] 0.3125
> dbinom(5,size=5, prob=0.5)
[1] 0.03125
> dbinom(0,size=5, prob=0.5)
[1] 0.03125
```

The **R** code, with a one-line built-in function, will directly calculate the probability that we want. Though the outputs of **Python** and **R** for this question agree from a numerical perspective, **R** code would have the more accurate answer compared to **Python**, as in this Example 5.1, since it is obtained by a direct calculation and not by simulation.

Example 5.2: A commuter plane has 10 seats. However, the airline books 12 people on each flight since not everyone who books a seat will actually show up for the flight. Suppose the chance of a person, who makes a reservation, actually showing up is 0.8.

Let $X = \#$ of who actually show up for the flight among the 12 who reserve.

The possible values for X are $x = 0, 1, 2, \ldots, 11, 12$ and the probabilities are given by the Binomial formula with $n = 12$, $p = 0.8$. For instance,

$$\mathrm{P}(X = 10) = \binom{12}{10}(.8)^{10}(.2)^2 = 0.2835$$

by direct calculation or from Table B. Again, using Table B or by direct calculation,

$$\begin{aligned}
\mathrm{P}(\text{someone is bumped}) &= \mathrm{P}(X = 11 \text{ or } 12) \\
&= \mathrm{P}(X = 11) + P(X = 12) \\
&= \binom{12}{11}(0.8)^{11}(0.2)^1 + \binom{12}{12}(0.8)^{12}(0.2)^0 \\
&= .2062 + .0687 \\
&= .2749
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{P}(\text{at least one seat is empty}) &= \mathrm{P}(X = 0) + \mathrm{P}(X = 1) + \cdots + \mathrm{P}(X = 9) \\
&= 1 - [\mathrm{P}(X = 10) + \mathrm{P}(X = 11) + \mathrm{P}(X = 12)] \\
&= 1 - (0.2062 + 0.0687 + 0.2835) \\
&= 0.4416.
\end{aligned}$$

$\diamond$

Remark 5.1 Do the 12 passengers correspond to 12 independent trials i.e., does the chance of success on one trial (i.e., someone showing up) depend on the chance of success on other trials (i.e., others showing up)? Since some passengers may be friends or members of the

same family and may make decision as a group, the trials may not strictly speaking, be independent and therefore the binomial distribution may not be appropriate. However, for simplicity, we assumed they are independent.

Fact 5.1: If X has Bin$(n,\ p)$ distribution, then the mean of X,

$$\mu_X = np$$

and the standard deviation of X,

$$\sigma_X = \sqrt{np(1-p)}.$$

Example 4.1 (contd.): Fair coin tossed 3 times. Let X = # of heads in the 3 tosses. X has a Bin$(3,\ \frac{1}{2})$. Therefore, it has

$$\text{mean, } \mu = np = 3 \cdot \frac{1}{2} = 1.5$$

and

$$\text{standard deviation, } \sigma = \sqrt{np(1-p)} = \sqrt{3 \cdot \frac{1}{2} \cdot \frac{1}{2}} = \sqrt{\frac{3}{4}} = 0.8660.$$

Note that we obtained the same values for μ and σ in Chapter 4 by a direct computation that uses their definitions.

$\diamond$

Example 5.2 (contd.): 12 passengers book seats. $p = 0.8$ is the chance for an individual to show up. Here

$\mu = np = 12(0.8) = 9.6$

$\sigma = \sqrt{np(1-p)} = \sqrt{12(0.8)(0.2)} = \sqrt{1.92} = 1.38$. Thus, on the average 9.6 people show up at the plane and the standard deviation is 1.38.

$\diamond$

Python/R Instruction

One common limitation that **Python** and **R** share in calculating the mean or the sd of a binomial distribution is that both involve a large amount of calculation or simulation. So the theoretical derivations and formula given above are useful.

Example 5.3: Roll a fair die 36 times and let X = number of times face 6 shows up. What is the average and standard deviation of this random variable?
Solution: Here, each time one rolls a die, there are six possible scores and one might think a Binomial is not appropriate. But since we are just interested in whether "the score is 6" (a success) or "not 6" (a failure), we have a Binomial situation. We have 36 independent trials of a Binomial experiment where we think of getting a score 6 as a "success" and thus, X has a Binomial distribution, with $n = 36$ and $p = \frac{1}{6}$. We write X has Bin(36, $\frac{1}{6}$) distribution (in symbols, $X \sim$ Bin(36, $\frac{1}{6}$)). Hence, for instance,

$$\mathrm{P}(X = 5) = \binom{36}{5}\left(\frac{1}{6}\right)^5\left(\frac{5}{6}\right)^{31},$$

$$\mathrm{P}(X = 4 \text{ or } 5) = \mathrm{P}(X = 4) + \mathrm{P}(X = 5) = \binom{36}{4}\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^{32} + \binom{36}{5}\left(\frac{1}{6}\right)^5\left(\frac{5}{6}\right)^{31}.$$

From the formulae for binomial mean and standard deviation,

$$\text{Mean of } X,\ \mu = np = 36 \times \frac{1}{6} = 6$$

and

$$\text{standard deviation of } X,\ \sigma = \sqrt{np(1-p)} = \sqrt{36 \cdot \frac{1}{6} \cdot \frac{5}{6}} = \sqrt{5} = 2.2361.$$

◇

Python code:

```
import numpy as np
dice=np.random.binomial(36,1/6,10000)  #simulate the expert 10000 times
```

```
count_5=0  # number of 5s in array dice
count_4_and_5=0  #number of 4s and 5s in array dice
for i in range(10000):
    if dice[i]==5:
        count_5=count_5+1
    if dice[i]==5 or dice[i]==4:
        count_4_and_5=count_4_and_5+1
print('The probability of P(X=5) is ', count_5/10000)
print('The probability of P(X=4 or 5) is', count_4_and_5/10000)

dice_mean=np.mean(dice)  #find mean of array
dice_std=np.std(dice)  # find standard deviation of array
print('The binomial mean and standard deviation are ',dice_mean, 'and',
    dice_std)
```

The output for this **Python** code gives::

```
The probability of P(X=5) is  0.1721
The probability of P(X=4 or 5) is 0.3006
The binomial mean and standard deviation are  6.0395 and 2.2326
```

R code:

```
#p(x=5)
dbinom(5, size=36, prob=1/6)
#p(x=4 or 5)
dbinom(5, size = 36, prob = 1/6)+dbinom(4, size = 36, prob = 1/6)
sum(dbinom(4:5,size=36, prob = 1/6))
# find mean: compare sample result with theoretical result
bin_sample <- rbinom(10000, size = 36, p=1/6)
mean(bin_sample)
sd(bin_sample)

new_bin_sample <- rbinom(10000000, size = 36, p=1/6)
mean(new_bin_sample)# more accurate due to increased sample size
sd(new_bin_sample)#more accurate
```

The output for this **R** code gives::

```
> dbinom(5, size=36, prob=1/6)
```

```
[1] 0.1701991
> dbinom(5, size = 36, prob = 1/6)+dbinom(4, size = 36, prob = 1/6)
[1] 0.3031671
> sum(dbinom(4:5,size=36, prob = 1/6))
[1] 0.3031671
> # find mean: compare sample result with theoretical result
> mean(bin_sample)
[1] 6.0068
> sd(bin_sample)
[1] 2.260501
> mean(new_bin_sample)
[1] 6.00079
> sd(new_bin_sample)
[1] 2.236377
```

We used a new built-in function in **R**, namely "*rbinom*()" for working with Example 5.3: This function randomly generates data from $bin(size, p)$.

```
bin_sample<- rbinom(10000, size=36, p=1/6)
```

meaning that there are 10000 trials from a Bin(36,1/6) randomly generated and stored in an array called "bin sample". Then the mean and standard deviation of array "bin sample" is used to approximate the population mean and standard deviation. When we increase the sample size from 10,000 to 10,000,000 as we did in the succeeding lines, the corresponding mean and standard deviation become closer to the theoretical value, as one would expect.This phenomenon, whereby the sample mean approaches the true mean with increased sample sizes, is more generally true and is called the "Law of Large Numbers" (See Remark 6.3 later).

5.2 Normal distribution

This is a very useful continuous distribution used to model many data sets whose frequency curves are approximately **bell-shaped** i.e., the frequency curve is symmetric around a single

peak at the center and the frequencies decline rather rapidly, as we move away from this center. For instance, the distribution of heights, weights, SAT scores etc. tend to have such bell-shaped curves. A Normal distribution is completely described by its central value or point of symmetry, μ and the spread around this center, denoted by the standard deviation, σ. We write N(μ, σ) for such a distribution. There is a different Normal curve for each pair

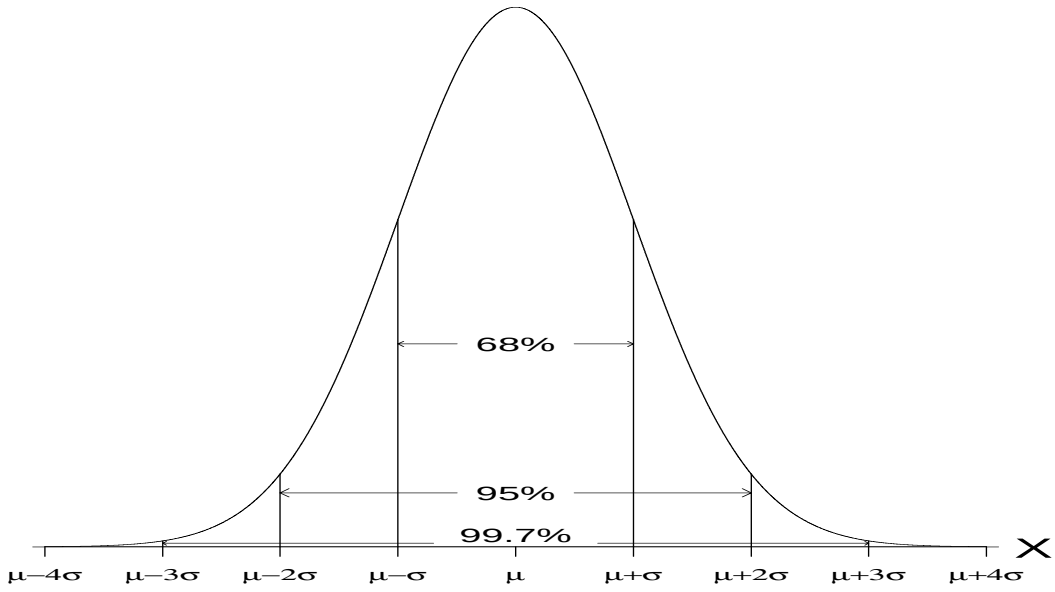


Figure 5.1: Normal curve

of values of μ and σ. These μ and σ are called parameters of the Normal distribution and their roles in describing the center and variability are illustrated in Figures 5.2 and 5.3. Note that the mean μ tells us where the curve is centered. On the other hand, the parameter σ is a measure of spread. Smaller the σ, the taller and more concentrated the curve is at the center, so that there is a greater chance of observing a value near the mean.

Some properties of a Normal curve

5.1 The curve is symmetric and centered around the mean, μ.

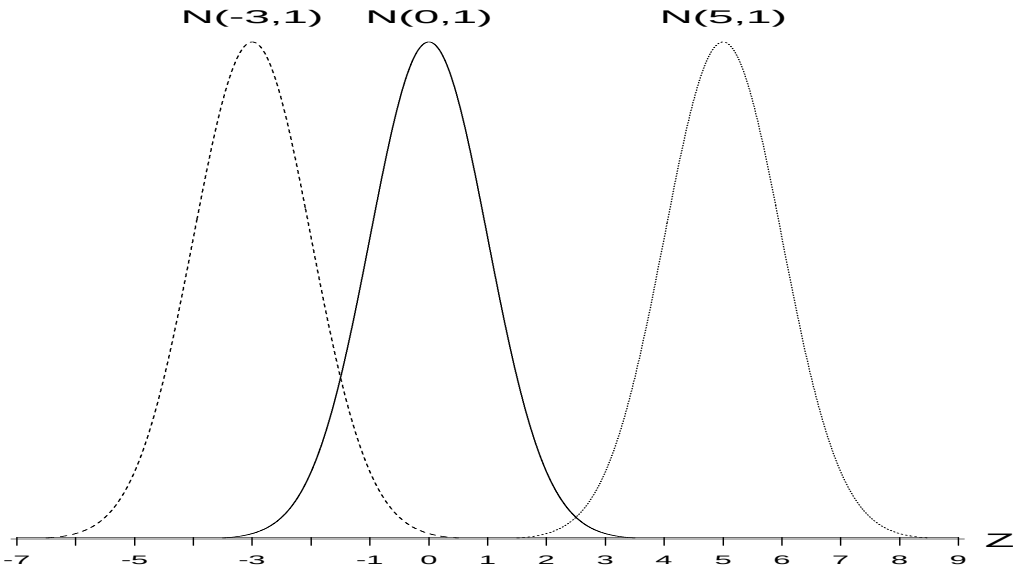


Figure 5.2: Normal curves with different μ's and same σ

5.2 Although theoretically the curve lies between -∞ and ∞,

$$\boxed{(\mu - 3\sigma,\ \mu + 3\sigma) \text{ has } 99.7\% \text{ probability.}}$$

Hence, $(\mu - 3\sigma,\ \mu + 3\sigma)$ can actually be considered as the "practical range" for a N($\mu,\ \sigma$) distribution. Similarly, if we take a symmetric interval around the mean going one standard deviation on each side, it has approximately 68% probability, while an interval two standard deviations from the mean has nearly 95% probability, i.e.,

$$\boxed{(\mu - \sigma,\ \mu + \sigma) \text{ has } 68\% \text{ probability and } (\mu - 2\sigma,\ \mu + 2\sigma) \text{ has } 95\% \text{ probability.}}$$

This rule is sometimes called the $1\sigma - 2\sigma - 3\sigma$ rule or 68- 95- 99.7 rule.

5.3 In particular, the Normal curve centered at $\mu = 0$ and with $\sigma = 1$, is called the **Standard Normal** curve and a random variable having such a standard Normal distribution is usually denoted by Z. See Figure 5.4 for the curve of a standard Normal random variable. If X is any Normal random variable with a N($\mu,\ \sigma$) distribution, then it can be **standardized** by subtracting its mean μ and dividing by σ to obtain a standard Normal or N(0, 1) distribution.

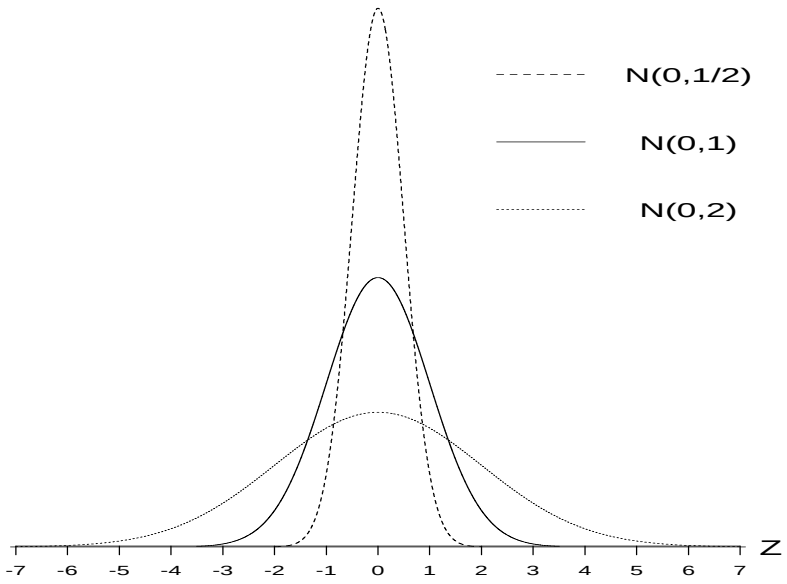


Figure 5.3: Normal curves with same μ and different σ's

That is, if X has a N($\mu,\ \sigma$) distribution, $Z = \dfrac{X-\mu}{\sigma}$ has a N(0, 1) distribution.

Areas under a Normal curve

In some simple cases, it is possible to find the probabilities corresponding to a N($\mu,\ \sigma$) distribution, by simply using Property 5.2 above, namely the 68-95-99.7 rule.

Example 5.4: Suppose it is given that the height of adult women has a Normal distribution, with $\mu = 64.5$ inches and $\sigma = 2.5$ inches. The probability that a randomly selected woman has height between 59.5 and 69.5 inches, i.e., between $(\mu - 2\sigma,\ \mu + 2\sigma)$ is 95% while the probability that her height exceeds 72, which is $(\mu + 3\sigma)$ inches, is $\frac{1}{2} \times (1 - 0.997)$ or 0.0015 because of symmetry and the 99.7% rule.

$\diamond$

However, in most cases, we need to use the Property 5.3, which allows us to translate

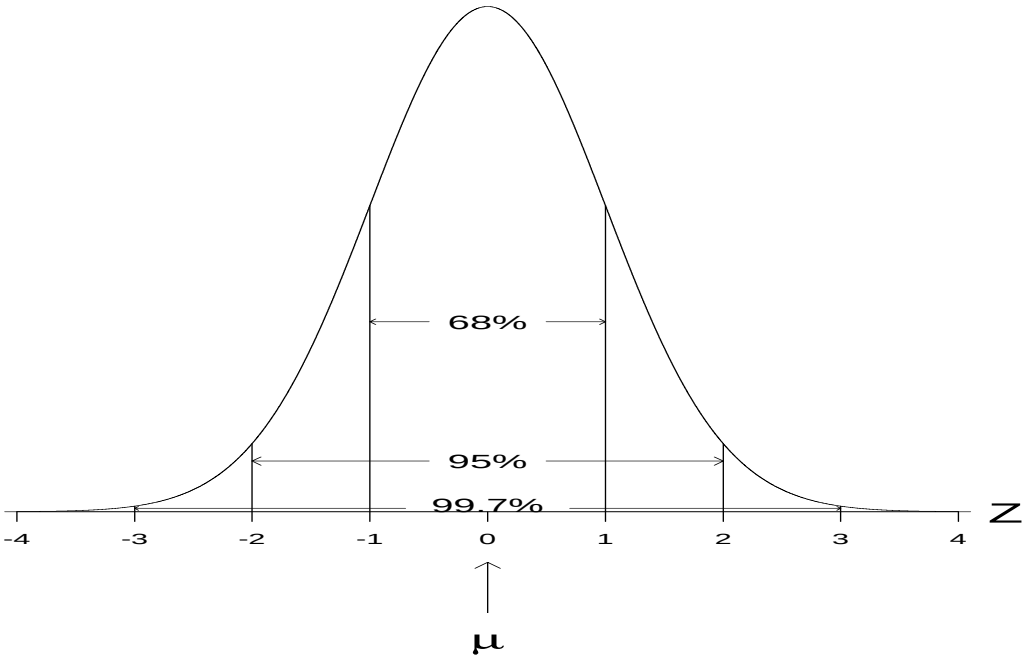


Figure 5.4: Standard Normal curve

the problem into one on Z and then look up the areas for the standard Normal or the Z variable, which are tabulated in Table A of the book. This table given on the inside front cover as well as in the Appendix, provides the area to the left of any given value z.

For instance, from this table, one can read that the area to the left of 1.20 is 0.8849, that to the left of -1.32 is 0.0934 and that to the left of 0 is 0.5000.

The last probability reflects the fact that the N(0, 1) curve is symmetric about zero with a probability of 0.5 on either side. If one wants the probability to the right of a given number, i.e., on the right tail, recalling that the total probability is one,

$$\mathrm{P}(Z > 1.2) = 1 - \mathrm{P}(Z \leq 1.2) = 1 - 0.8849 = 0.1151.$$

Or by symmetry of the tail areas,

$$\mathrm{P}(Z > 1.2) = P(Z < -1.2) = 0.1151.$$

If we want the area between any two points, we take the cumulative area till the larger

value and subtract the cumulative area till the smaller value, which leaves us the with area in between these two points. Thus

$$\mathrm{P}(-1.32 < Z \leq 1.2) = \mathrm{P}(Z \leq 1.2) - \mathrm{P}(Z \leq -1.32) = 0.8849 - .0934 = 0.7915.$$

The last three probability calculations can be stated more generally as:

$$\begin{aligned} &\mathrm{P}(Z > a) = 1 - \mathrm{P}(Z \leq a) \\ &\mathrm{P}(Z > a) = \mathrm{P}(Z < -a) \\ &\mathrm{P}(a < Z \leq b) = \mathrm{P}(Z \leq b) - \mathrm{P}(Z \leq a) \end{aligned}$$

Python Code Instruction:

When simulating normal distribution, **Numpy** package in **Python** can do the job by using the Syntax:

```
np.random.normal(mean, sd, size)
```

- mean is the mean of the Normal distribution
- sd is the standard deviation of the population
- size represents the sample size that will be generated

R Code Instruction:

In **R**, computing probabilities of normal distribution can be done using *pnorm*() function using the Syntax:

```
pnorm(x, mean= , sd= , lower.tail= )
```

- x is the value of interest
- *mean* is the mean of the normal distribution

- *sd* is the standard deviation
- *lower.tail* decides whether the area under the curve is greater than x or less than x The output would be the probability $P(X > x)$ or $P(X \leq x)$

Example 5.5: Suppose X, the weight in lbs. of a college student, has a N(150, 20) distribution. Find the probability that a randomly selected student weighs more than 160lbs.

Then, translating any question on X to one on $Z = \frac{X-\mu}{\sigma} = \frac{X-150}{20}$, we get:

$$\mathrm{P}(X \leq 160) = \mathrm{P}(X-150 \leq 160-150) = \mathrm{P}\left(\frac{X-150}{20} \leq \frac{160-150}{20}\right) = \mathrm{P}(Z \leq 0.5) = 0.6915$$

from Table A while

$$\mathrm{P}(X > 160) = 1 - \mathrm{P}(X \leq 160) = 1 - 0.6915 = 0.3085.$$

◇

Python code:

```
import numpy as np
count=0
x=np.random.normal(150, 20, 10000)
for i in range (10000):
    if x[i]<=160:
        count=count+1
print('The probability that a randomly selected student weighs more than 160
   lbs is ', count/10000)
print('The probability that a randomly selected student weighs less than 160
   lbs is ', 1-count/10000)
```

The output for this Python code gives::

```
The probability that a randomly selected student weighs more than 160lbs
is  0.6934
The probability that a randomly selected student weighs less
than 160lbs
is  0.3098
```

R code:

```
#p(x<=160)
pnorm(160, mean=150, sd=20, lower.tail = T)
#p(x>=160)
pnorm(160, mean=150, sd=20, lower.tail = F)
p1-norm(160, mean=150, sd=20, lower.tail = T) #same as line above
```

The output for this R code gives:

```
> pnorm(160, mean=150, sd=20, lower.tail = T)
[1] 0.6914625
> pnorm(160, mean=150, sd=20, lower.tail = F)
[1] 0.3085375
> 1-pnorm(160, mean=150, sd=20, lower.tail = T)
[1] 0.3085375
{
```

Example 5.4 (contd.): Recall X, the height of women has a N(64.5, 2.5) distribution. Hence,

$$\mathrm{P}(X \leq 70) = \mathrm{P}(X - 64.5 \leq 70 - 64.5) = \mathrm{P}\left(\frac{X - 64.5}{2.5} \leq \frac{70 - 64.5}{2.5}\right) = \mathrm{P}(Z \leq 2.2) = 0.9861,$$

$$\begin{aligned}\mathrm{P}(60 < X \leq 70) &= \mathrm{P}\left(\frac{60 - 64.5}{2.5} < \frac{X - 64.5}{2.5} \leq \frac{70 - 64.5}{2.5}\right) = \mathrm{P}(-1.8 < Z \leq 2.2) \\ &= \mathrm{P}(Z \leq 2.2) - \mathrm{P}(Z \leq -1.8) = 0.9861 - 0.0359 = 0.9502.\end{aligned}$$

◇

Table A can be used not only to read the areas for a given value of z but also to read off the z value corresponding to a given area. For instance from Table A, the value z which has a 20% probability below it is -0.84 while the value z with a 95% probability to its left, falls between 1.64 and 1.65 and is indeed 1.645.

Example 5.6: Suppose the score, X on a certain test has a N(430, 100) distribution. Find the 90th percentile score on this test i.e., x such that $\mathrm{P}(X \leq x) = 0.9$.

$$0.9 = \mathrm{P}(X \leq x) = \mathrm{P}\left(\frac{X - 430}{100} \leq \frac{x - 430}{100}\right) = \mathrm{P}\left(Z \leq \frac{x - 430}{100}\right).$$

From Table A, the z-value corresponding to an area of 0.90 is 1.28, so that we have the equation

$$\begin{aligned} \frac{x-430}{100} &= 1.28 \\ (x-430) &= 128 \\ x &= 430+128=558. \end{aligned}$$

i.e., if the scores follow this Normal distribution, then there is a 90% chance that a randomly selected student will have a score of 558 or less, or to put it differently, 90% of the students have a score of 558 or less.

◇

Summary of Python/R codes

As the code discussed in this chapter will be frequently called upon in the next few chapters, we provide a brief review of the ideas here:

- Python

  5.1 the general procedure for calculating probabilities using Numpy:

  - **set**:
    Import Numpy and set initial values for counters
  - **simulate**:
    Using *numpy.random*() simulate desired number of samples that follow certain distribution in an array
  - **count**:
    Count the number of values of interest in the array using "For loop"
  - **calculate**:
    Calculate the probability as the proportion of values of interest to the total number of simulated values

  5.2 Binomial distribution

  - Syntax:
    $numpy.random.binomial(n, p, size)$
    - n (integer) is the number of trials;
    - p (float is between 0 and 1) is the probability of success;

    * size (integer) represents the sample size.

5.3 Normal distribution

- Syntax:
  $numpy.random.normal(mean, sd, size)$
  * mean is the mean of the population following N(mean, sd)
  * sd is standard deviation of the population;
  * size (integer) is the sample size that will be generated.

5.4 Other functions

- $np.means()$
  calculates the mean of an array
- $np.sd()$
  calculates the standard deviation of an array

- R

5.1 Binomial distribution

- $dbinom(x, size, prob)$
  * probability function; probability of a certain value "x"
  * x: number of successes
  * size: total number of trials
  * prob: probability of success in a single trial
  * output P(X=x).
- $pbinom(q, size, prob, lower.tail = TRUE)$
  * Distribution function; cumulative probability before, or after a certain critical value
  * q: critical number of successes
  * size: total number of trials
  * prob: probability of success in a single trial
  * Lower.tail: if TRUE then calculate the area to the left of critical value, and if FALSE then calculate the area to the right of critical value
  * e.g. P(X greater than 160): $pbinom(160, size, prob, lower.tail = FALSE)$
- $qbinom(p, size, prob, lower.tail = TRUE)$

    - Quantile function; Exact value of a certain percentile
    - p: specified percentile of interest
    - size: total number of trials
    - prob: probability of success in a single trial
    - Lower.tail: if TRUE then calculate the value x, area to the left of which is p, if FALSE then calculate x, the area to the right of which is p
    - e.g. the 25th percentile of bin(n, p): $qbinom(0.25, size, prob, lower.tail = TRUE)$
  - $rbinom(n, size, prob)$
    - Random Generation function
    - n: number of generated data
    - size: total number of trials
    - prob: probability of success in a single trial
    - e.g. randomly generate 1000 data following a bin(n, prob) $rbinom(1000, size, prob)$

5.2 Normal distribution

  - $dnorm(x, mean =, sd =)$
    - Density function at a given point
    - x: critical value
    - mean: mean of normal distribution
    - sd: standard deviation of normal distribution
    - i.e. $f_X(x) = dnorm(x, mean =, sd =)$
  - $pnorm(q, mean =, sd =, lower.tail =)$
    - Distribution function; probability after or before a certain value of r.v. X
    - q: critical value
    - mean: mean of normal distribution
    - sd: standard deviation of normal distribution
    - Lower.tail: if TRUE then calculate the area to the left of the critical value, and if FALSE then calculate the area on the right of critical value
  - $qnorm(p, mean =, sd =, lower.tail =)$
    - Quantile function; Exact value of a certain percentile
    - p: critical value

    - mean: mean of normal distribution
    - sd: standard deviation of normal distribution
    - Lower.tail: if true then calculate the area on the left of critical value if false then calculate the area on the right of critical value
- $rnorm(n, mean =, sd =)$
    - Random variables generating function
    - n: number of samples generated
    - mean: mean of normal distribution
    - sd: standard deviation of normal distribution

EXERCISES

5.1 In a medical study it is found that there is 5% chance of a false positive in a mammogram test i.e., suggests presence of breast cancer but biopsy reveals that cancer is not present. If a woman has 30 mammograms during her lifetime (say, once a year, between the ages of 40 and 70), what is the chance of at least one false positive during her lifetime?

5.2 Suppose a multiple choice test has 20 questions with each question having 4 choices, only one of which is correct. Suppose an unprepared student writes the answers, each time randomly picking one of the 4 choices. What is the probability that the student will get

(a) exactly 4 answers correct?

(b) 8 or more correct?

5.3 A girl-scout makes visits to 15 households over a weekend, trying to sell cookies. She feels that there is a 30% chance of selling cookies at any one household.

(a) What probability model would be appropriate for describing the number of households that buy cookies?

(b) Compute the probability that 10 or more households buy cookies, among the 15 she visits.

5.4 If I throw a peanut up in the air, there is a probability $\frac{4}{5}$ that I can catch it in my mouth. I throw up in the air four peanuts. Assume that the throws are independent.

(a) What is the probability that I catch in my mouth

i. all four ?

ii. exactly two ?

iii. at least one?

(b) Find the mean (μ) and variance (σ^2) for the number I catch in my mouth.

5.5 Since 2 out of 3 is the same proportion as 4 out of 6, can you conclude that the probability of getting 2 heads in 3 tosses of a fair coin is the same as that of getting 4 heads in 6 tosses?

5.6 The Intelligence Quotient (IQ) of a person is normally distributed with mean 100 and standard deviation 16.

(a) What percentage of the population possess an IQ in the interval (84, 116)? What rule, if any, did you use to arrive at your answer?

(b) MENSA is an organization for people with IQ in the top 2% of the population. What IQ should a person possess to get admitted to MENSA?

5.7 Suppose the weight of newborn babies is normally distributed with mean 7 pounds and standard deviation 1 pound.

(a) Find the probability that a given newborn weighs more than 5 pounds.

(b) What percentage of the newborns has weight greater than 8 pounds and less than 9 pounds?

(c) Find the weight such that 10% of newborns have weight less than or equal to that value.

5.8 Suppose that the daily high temperature in January in New York City's Central Park has a Normal distribution with mean $\mu = 33$ degrees and $\sigma = 4$ degrees.

(a) What is the probability that a day in January will have a high temperature of more than 30 degrees?

(b) What is the probability that a day in January will have a high temperature between 34 and 40 degrees?

(c) What is the temperature such that 9% of days in January will have daily highs less than that temperature?

5.9 Suppose the lengths of frogs bred for human consumption follow a Normal distribution with mean length $\mu = 20$ cms and $\sigma = 5$ cm. They are classified as "small" if the length is less than 15 cm, "medium" if between 15 and 25 cms, "large" if between 25 and 30 and "jumbo" if longer than 30 cms.

(a) What is the probability that a frog you select at random is "jumbo"?

(b) What is the probability that a randomly selected frog is "medium"?

(c) If you select two frogs at random for dinner, what is the probability that one is "large" while the other one is "jumbo"?

5.10 When customers call a certain 800 number, the time for which they are put on hold is normally distributed with a mean of 3 minutes and standard deviation of 1 minute.

(a) What is the probability that a randomly chosen caller has to wait on hold for more than 5 minutes?

(b) If I have to call the number five times a week, what is the probability that I wait for more than 5 minutes exactly once during the week?

5.11 Suppose that the scores, X, on a college entrance examination are normally distributed with a mean of 550 and a standard deviation of 90. A certain university will consider for admission only those applicants, whose scores fall in the top 10%. Find the minimum score an applicant must achieve in order to be considered for admission to the university.

5.12 Suppose the "waiting time at a bank teller's window" is normally distributed with mean $\mu = 5$ minutes and standard deviation $\sigma = 1.2$ minutes. If a customer visits this bank on two different days, what is the probability that she has to wait more than 7 minutes on both these occasions?

5.13 A beer distributor believes that the actual amount of beer in a 12 ounce can of beer has a Normal distribution with a mean of 12 ounces and a standard deviation of 1 ounce. If a 12 ounce beer can is randomly selected, find the probability that

(a) the 12 ounce can of beer will actually contain less than 11 ounces of beer.

(b) 12 ounce can of beer will actually contain more than 12.5 ounces of beer.

(c) the 12 ounce can of beer will actually contain between 10.5 and 11.5 ounces of beer.

5.14 A tire manufacturer claims that the steel-belted radial tire manufactured by their company has tread life that is normally distributed with a mean life of 22,000 miles and a standard deviation of 18,000 miles.

(a) Find the probability that a tire selected at random will last

i. less than 18,000 miles.

ii. more than 18,000 miles.

(b) If two tires are selected at random, what is the probability that both will last less than 18,000 miles?

5.15 A car battery of a certain brand has a mean life, $\mu = 5$ years with a standard deviation, $\sigma = 1.5$ years. Assume a Normal distribution for the lifetime.

(a) What is the probability that a battery of this brand will last longer than 8 years?

(b) If this battery is guaranteed for 3 years, what percent of these can be expected to need replacement before the warranty period?

(c) If this company wishes to replace no more than 10% of the battery sold under its guarantee program, how long should the guarantee period be?

5.16 A machine produces nails whose length X is a random variable with a Normal distribution with $\mu = 2$" and $\sigma = 0.02$".

(a) Find the probability that a nail produced by this machine is

i. shorter than 1.95".

ii. "defective" where a defective nail is one whose length is either < 1.95" or > 2.05".

(b) If three nails are picked from this machine's production, what is the probability that all three are good (i.e., there are no defectives)?

5.17 Suppose the IQ scores for American children have a Normal distribution with mean $\mu = 100$ and variance $\sigma^2 = 25$.

(a) Find the probability that a child's IQ is between 88 and 110.

(b) Find the probability that a child will have IQ higher than 130.

(c) Find the IQ score which separates the lowest 20% from the higher 80% of the scores.

5.18 A children's bicycle says the "safe maximum load" for any rider is 150 lbs. If a UCSB student gets on this bike, what is the probability that she will have a "safe ride", given that the weights of students at UCSB are normally distributed with $\mu = 120$ lbs. and $\sigma = 10$ lbs.?

5.19 A cafeteria coke machine dispenses coke into 6-ounce cups in such a way that the actual amount dispensed into any particular cup is normally distributed with standard deviation of 0.1 ounce. The machine can be set so that the mean of the amount dispensed is any desired level. At what level should the mean be set so that the 6-ounce cup will overflow about 2% of the time?

5.20 Suppose at a given hospital, newborn birth weights are normally distributed with an average of 120 ounces and a standard deviation of 10 ounces. A doctor knows that infants with hypothyroidism have high birth weights and are are always above the 90th percentile in birth weight. At what birth weight should the doctor be concerned about the baby having hypothyroidism?

5.21 In a study of long distance runners, the average weight was found to be 140 lbs. with a standard deviation of 10 lbs. Assuming the distribution of runners' weights to be normal,

(a) Find the proportion of runners who weigh

i. more than 155 lbs.?

ii. between 122 and 158 lbs.?

(b) If 90% of the runners weigh less than Peter, find Peter's weight.

5.22 If Z has a N(0, 1) distribution, find the area

(a) below 1.00.

(b) below -1.5.

(c) above 2.4.

(d) under the single value 2.

(e) between 1 and 2.

(f) between -1 and 2.2.

5.23 Suppose the heights of adult males follow a Normal distribution with mean $\mu = 68$ inches and $\sigma = 3$ inches. Find

(a) the proportion of males whose heights are between 65 and 71 inches.

(b) the proportion of males who are taller than 6 feet.

(c) the chance that if two males are selected at random from this population, both are taller than 6 feet.

(d) the proportion of males that is exempt, if a certain police department does not recruit anyone who is shorter than 5 feet.

(e) the first quartile i.e., a value below which 25% of the male heights lie.

5.24 The composite scores in a statistics course have a Normal distribution with a mean μ of 68 and a standard deviation σ of 8. If the teacher decides to give the top 15% of the students an A grade, what is the lowest score needed for an A ? If the bottom 10% are to be given an F, below what score is it an F?

5.25 Consider a student who is taking a multiple choice exam where there are 5 possible answers for each question. Since the student has not studied or attended any of the classes, the student decides to randomly guess each question. Suppose there are 10 questions on this exam.

(a) Find the probability the student gets the first 2 guesses right and the last 8 guesses wrong.

(b) Find the probability that the student gets 2 questions out of the 10 correct (give an exact expression and also use table).

(c) If it takes 6 or more correct answers to pass this test, what is the chance that this student will pass?

(d) What is the expected number of correct guesses and its variance?

Chapter 6

Sampling distributions

6.1 Sampling variability

Recall that population characteristics, like its center and spread, are called **parameters** and are denoted by μ and σ respectively. Suppose we are interested in the average height, μ, of all the adult males in the United States. To estimate it, we may select a representative random sample of 100 men. Suppose in this sample of 100 men, we get

$$\overline{X} = \text{ sample mean } = 5'5''.$$

If we repeat the same process again, i.e., draw another random sample of 100 and calculate the sample mean, then we may (and almost surely will) get a different value for the sample mean, say 5'2" — the third time possibly 4'6" . The values of the sample means differ because the samples, on which they are based, differ. The important question here is, if we repeat this many times, what possible distribution of values of these sample means would we get? This is called the **sampling distribution** of $\overline{X}$, i.e., the distribution of values of $\overline{X}$ over all possible samples. If we know features of this sampling distribution, like the location of its center and how spread out it is about this center, it will help us in interpreting and using the sample mean for inference.

The idea applies to the sampling distribution of any other **statistic** (recall a statistic is just a quantity based on the sample). For example, in tossing a coin consider the "probability of getting heads", say p, which typically we do not know. Suppose a sample of size n=20

is obtained by tossing the coin 20 times and say, we obtain 12 heads in this sample of 20 tosses. Then the **observed** proportion of heads i.e.,

$$\text{proportion of heads in the sample } = \frac{12}{20} = 0.6$$

can be used as an **estimate** of the unknown p. If we toss the same coin again and obtain another sample of 20, we may end up getting 11 heads — so that, this time, proportion of heads in the sample $= \frac{11}{20} = 0.55$. The question again is: "What is the **sampling distribution** of the observed proportion, i.e., the distribution of values of this observed proportion of heads over repeated sampling?" In some important cases, we are able to obtain such sampling distributions and use the center and spread of such distributions for purposes of inference.

Remark 6.1 One simple way to generate the sampling distribution of the observed proportion when $n = 20$ and the true proportion $p = \frac{1}{2}$, is to do the following experiment:

I. Toss a fair coin 20 times, observe the number of heads, note down the observed proportion of heads.
II. Repeat step I again and again.
III. Draw a histogram of the different values of the observed proportion, which ranges from 0 to 1. This process can be simulated easily on a computer.

Definition: A statistic that is used to estimate a parameter is said to be **unbiased** for that parameter if the sampling distribution of this statistic centers at the true parameter value.

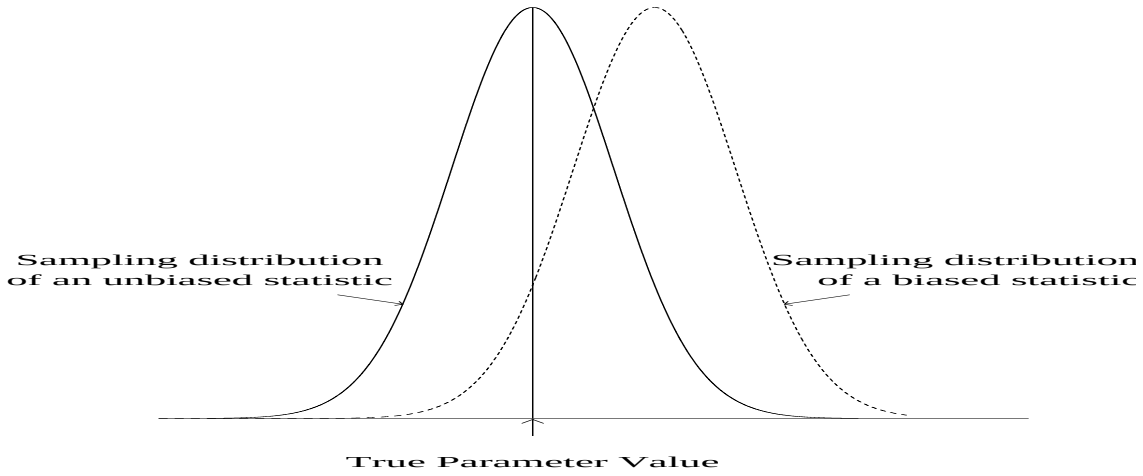


Figure 6.1: Biased and unbiased statistics

That is, even though the statistic will vary from sample to sample, on the average, the statistic centers at the correct or true value which we are trying to estimate. As we proceed, we will be able to check that the sample mean $\overline{X}$ is unbiased for the true population mean μ and and observed sample proportion in the Binomial situation is unbiased for the true probability p. The sample variance s^2 with a denominator of $(n-1)$ that we defined earlier, is also an unbiased estimate of the true variance σ^2, although, this is a bit harder to check.

When two different statistics have sampling distributions, both of which center around the correct or true value (i.e., both are unbiased), we would prefer the one which has a smaller spread around this center. Such a statistic is more likely to be closer to the true value which we are trying to estimate and hence preferable. See Figure 6.2.

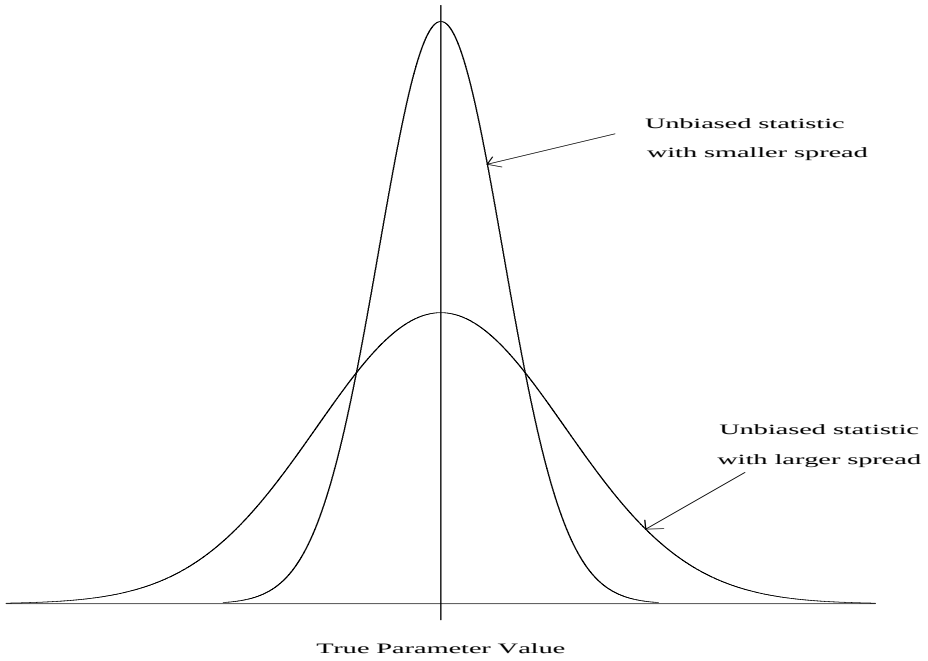


Figure 6.2: Among two unbiased statistics, the one with the smaller spread is preferred.

Coding Instruction:

In this chapter, we will mainly use the **R** functions that were previously discussed, to get a deeper understanding of sampling distributions. While **Python** can also be used to consider such questions, it is somewhat more complicated and not so straightforward at the introductory level; thus our main focus in this chapter will be on **R**.
One such illustration is to look at the distribution of the sample mean from a binomial and demonstrate the normal approximation to the binomial distribution.

6.2 Distribution of the sample mean

Let $(X_1\ ,\ X_2\ ,.\ .\ .\ ,\ X_n)$ be a sample from a distribution with center at μ and standard deviation σ. We observed that $\overline{X} = \frac{1}{n}\sum X_i$ can take different values for different samples. We now state some results on the sampling distribution of $\overline{X}$ i.e., the distribution of values of $\overline{X}$ when we repeatedly draw such samples.

Fact 1: The mean and standard deviation for the $\overline{X}$ distribution are given by

$$\mu_{\overline{X}} = \mu$$

and

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\,.$$

i.e., the sampling distribution of $\overline{X}$ centers at the same value μ as the center of any individual observation. On the other hand, the spread of the $\overline{X}$ distribution (as described by its standard deviation) *reduces* by a factor of $\sqrt{n}$ and is $\frac{1}{\sqrt{n}}$ of the standard deviation of any individual value. Thus the $\overline{X}$ values, i.e., averages, tend to be much less variable or more stable than single observations.

Remark 6.2 The fact that the sampling distribution of $\overline{X}$ tends to be centered at μ can be rephrased to say that the sample mean $\overline{X}$ is **unbiased** for μ.

Remark 6.3 Note that since $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$, this standard deviation of $\overline{X}$ tends to zero as $n \to \infty$.

Recall that a variable with variance or standard deviation of zero is a constant (in this case, μ), so that we may conclude that $\overline{X}$ comes arbitrarily close to μ for large enough n. This fact is referred to as the **Law of Large Numbers**, which says that the sample mean $\overline{X}$ approaches the true mean, μ when the sample size is sufficiently large.

For Fact 1, we assumed that the individual values have mean μ and standard deviation σ and nothing about the what distribution they have. Suppose we assume further that the individual values are *from a* N(μ, σ) *curve*. Then $\overline{X}$ also follows a normal curve, with the center and standard deviation indicated in Fact 1 above, i.e., $\mu_{\overline{X}} = \mu$, $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$.

Fact 2:

If X_1 , X_2 ,. . . , X_n are from $N(\mu,\ \sigma)$, $\overline{X}$ has a N(μ, $\frac{\sigma}{\sqrt{n}}$) distribution.

Since $\overline{X}$ is normally distributed with $\mu_{\overline{X}} = \mu$, $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$, we may **standardize** it to get

$$Z = \frac{(\overline{X} - \mu)}{(\sigma/\sqrt{n})} = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}$$

which has a N(0, 1) distribution.

Example 6.1: Students in a certain university have a weight distribution that is known to be N(150, 20). Let X_1 , X_2 ,. . . , X_{16} represent the weights of 16 randomly selected students from this university. If $\overline{X}$ is the average weight for this sample of the 16 students, find P($\overline{X} > 160$).

Solution: Recall from Fact 2 that $\overline{X}$ has N(μ, $\frac{\sigma}{\sqrt{n}}$) $\equiv$ N(150, $\frac{20}{4}$) = N(150, 5) distribution.

Therefore,

$$\begin{aligned} \mathrm{P}(\overline{X} > 160) &= \mathrm{P}\left(\frac{\overline{X} - 150}{5} > \frac{160 - 150}{5}\right) \\ &= \mathrm{P}\left(Z > \frac{10}{5}\right) \\ &= \mathrm{P}(Z > 2) \\ &= 1 - \mathrm{P}(Z \leq 2) \\ &= 1 - 0.9772 \\ &= 0.0228. \end{aligned}$$

◇

R code:

```
# given that X~N(150,20); 16 students are randomly chosen
# calculate sample mean and standard deviation
miu_bar=150
sigma_bar=20/sqrt(16)
# calculate P(X_bar>160)
pnorm(160,mean = miu_bar,sd=sigma_bar,lower.tail = FALSE)
```

The output for this **R** code gives::

```
> miu_bar=150
> sigma_bar=20/sqrt(16)
> pnorm(160,mean = miu_bar,sd=sigma_bar,lower.tail = FALSE)
[1] 0.02275
```

Example 6.1 (contd.): An elevator at this university has a carrying capacity of 1500 pounds. What is the probability that 9 students who enter the elevator will have a safe ride, i.e., their total weight is less than 1500 pounds?

Solution: By Fact 2, $\overline{X}$ has $\mathrm{N}(\mu, \frac{\sigma}{\sqrt{n}}) \equiv \mathrm{N}(150, \frac{20}{3}) = \mathrm{N}(150, \frac{20}{3})$. Therefore,

$$\mathrm{P}(\text{Total weight of 9 people is} < 1500) = \mathrm{P}(\overline{X} < \frac{1500}{9}) = \mathrm{P}(\overline{X} < 166.67)$$

$$= \mathrm{P}\left(\frac{\overline{X} - 150}{\left(\frac{20}{3}\right)} < \frac{166.67 - 150}{\left(\frac{20}{3}\right)}\right) = \mathrm{P}(Z < 2.5) = 0.9938$$

i.e., more than 99% chance that they will have a safe ride (unless all these 9 guys are from the university's football team!— in which case we may have a very unusual and not a representative sample!)

◇

R code:

```
miu_bar=150
sigma_bar=20/sqrt(16)
# total weight of 9 people less than 1500: P(X total<1500=P(X_bar<1500/9))
pnorm(1500/9,mean = miu_bar,sd=sigma_bar,lower.tail = TRUE)
```

The output for this **R** code gives:

```
miu_bar=150
sigma_bar=20/sqrt(16)
> pnorm(1500/9,mean = miu_bar,sd=sigma_bar,lower.tail = TRUE)
[1] 0.9995709
```

We note that for calculating the mean and standard deviation of $\overline{X}$ with a distribution of $\mathrm{N}(\mu,\sigma)$, we merely switch σ_X that we used in the calculation of X with a distribution of $\mathrm{N}(\mu, \sigma)$ for a single observation, to $\sigma_X/\sqrt{n}$

Example 6.2: Suppose X, the score on a certain test has $\mathrm{N}(500, 100)$. Let $X_1, \ldots, X_{16}$ be a sample of scores for 16 individuals and let $\overline{X}$ be the average score for these 16 people. Find $\mathrm{P}(550 < \overline{X} \leq 600)$.

Solution: $\overline{X}$ has normal distribution with $\mu_{\overline{X}} = \mu = 500$, and with $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{16}} = \frac{100}{4} = 25$.

Hence,

$$\begin{aligned} \mathrm{P}(550 < \overline{X} \leq 600) &= \mathrm{P}\left(\frac{550-500}{25} < \frac{\overline{X}-500}{25} \leq \frac{600-500}{25}\right) \\ &= \mathrm{P}(2 < Z \leq 4) \\ &= \mathrm{P}(Z \leq 4) - \mathrm{P}(Z \leq 2) \\ &= 1 - 0.9772 = 0.0228. \end{aligned}$$

◇

It is a somewhat surprising fact that even if the sample is not from a normal distribution as we assumed in Fact 2, the sampling distribution of $\overline{X}$ *is well approximated by the normal curve, provided the sample size is large.* This is called the **Central Limit Theorem**.

Fact 3 (Central Limit Theorem): If $X_1,\ X_2,\ \ldots,\ X_n$ are any set of observations with mean μ and standard deviation σ, their sample mean, $\overline{X}$, has approximately a $\mathrm{N}(\mu,\ \frac{\sigma}{\sqrt{n}})$ distribution, if n is *sufficiently large.*

Remark 6.4 An important question is how large should n be before we can use this normal approximation. Unfortunately, there is no simple answer, as it depends on a number of factors. However, the larger the n, the better the normal curve approximation — at least a sample of size of 20 is desirable in most cases, before we use this approximation.

Example 6.3: Let $X_1,\ X_2,\ \ldots,\ X_{25}$ be the lifetimes of electronic components, with $\mu = 700$ hours, $\sigma = 10$ hours. Find $\mathrm{P}(\overline{X} \leq 702)$, where $\overline{X}$ is the sample mean of the 25 lifetimes.
Solution: Typically such data as lifetimes, have a right-skewed distribution and not a normal curve. However since $n = 25$ is reasonably large, by Fact 3,

$$\overline{X} \text{ has approximately a } \mathrm{N}\left(\mu,\ \frac{\sigma}{\sqrt{n}}\right) = \mathrm{N}\left(700,\ \frac{10}{\sqrt{25}}\right) = \mathrm{N}(700,\ 2) \text{ distribution.}$$

Thus,

$$\mathrm{P}(\overline{X} \leq 702) = \mathrm{P}\left(\frac{\overline{X} - 700}{2} \leq \frac{702 - 700}{2}\right) = \mathrm{P}(Z \leq 1) = 0.8413.$$

$\diamond$

Remark 6.5 The main difference between Facts 2 and 3 is that Fact 2 holds true for *any* sample size as long as the original data is assumed to be from a normal distribution, while Fact 3 (or the Central Limit Theorem) is true even when the observations are not from a normal curve, but the approximation is good only for sufficiently *large* n.

6.3 Normal approximation to the Binomial

Recall the Bin(n, p) random variable X, which counts the number of successes in n trials with probability of success p on each trial. When the number of trials n is sufficiently large, the Binomial probabilities can be approximated by the Normal probabilities. This result can be stated in either of the two following equivalent ways — for the distribution of the observed *number* of successes, X (Fact 4) or for the distribution of the the observed *proportion* of successes, $\widehat{p} = \frac{X}{n}$, which is a fraction (Fact 4′).

$*$ We can also use **Python** to test the accuracy of Fact 4. (See Appendix)

Fact 4: For n large enough, the Bin(n, p) random variable X can be approximated by a normal distribution with

$$\mu_X = np,$$

and

$$\sigma_X = \sqrt{np(1-p)}.$$

Fact 4′: For n large enough, the distribution of observed Binomial proportion $\widehat{p} = \frac{X}{n}$, is approximately Normal with

$$\mu_{\widehat{p}} = p, \text{ the true proportion}$$

and with

$$\sigma_{\widehat{p}} = \sqrt{\frac{p(1-p)}{n}}.$$

Remark 6.6 It can be demonstrated that Facts 4 and 4′ are indeed consequences of the Central Limit Theorem, stated in Fact 3.

Remark 6.7 For this approximation to be reasonable, a general rule of thumb is that n and p should be such that $np \geq 5$ and $n(1-p) \geq 5$. The approximation becomes quite good when $np \geq 10$ and $n(1-p) \geq 10$.

For instance, following the stricter rule, if $p = \frac{1}{2}$, n should be greater than or equal to 20, whereas if p $= \frac{1}{4}$, n should be greater than or equal to 40 for the Normal approximation to be quite good. We plot in Figure 6.3 Binomial distributions for different n and p to demonstrate the normal approximation. (Code for creating this plot is given in the Appendix).

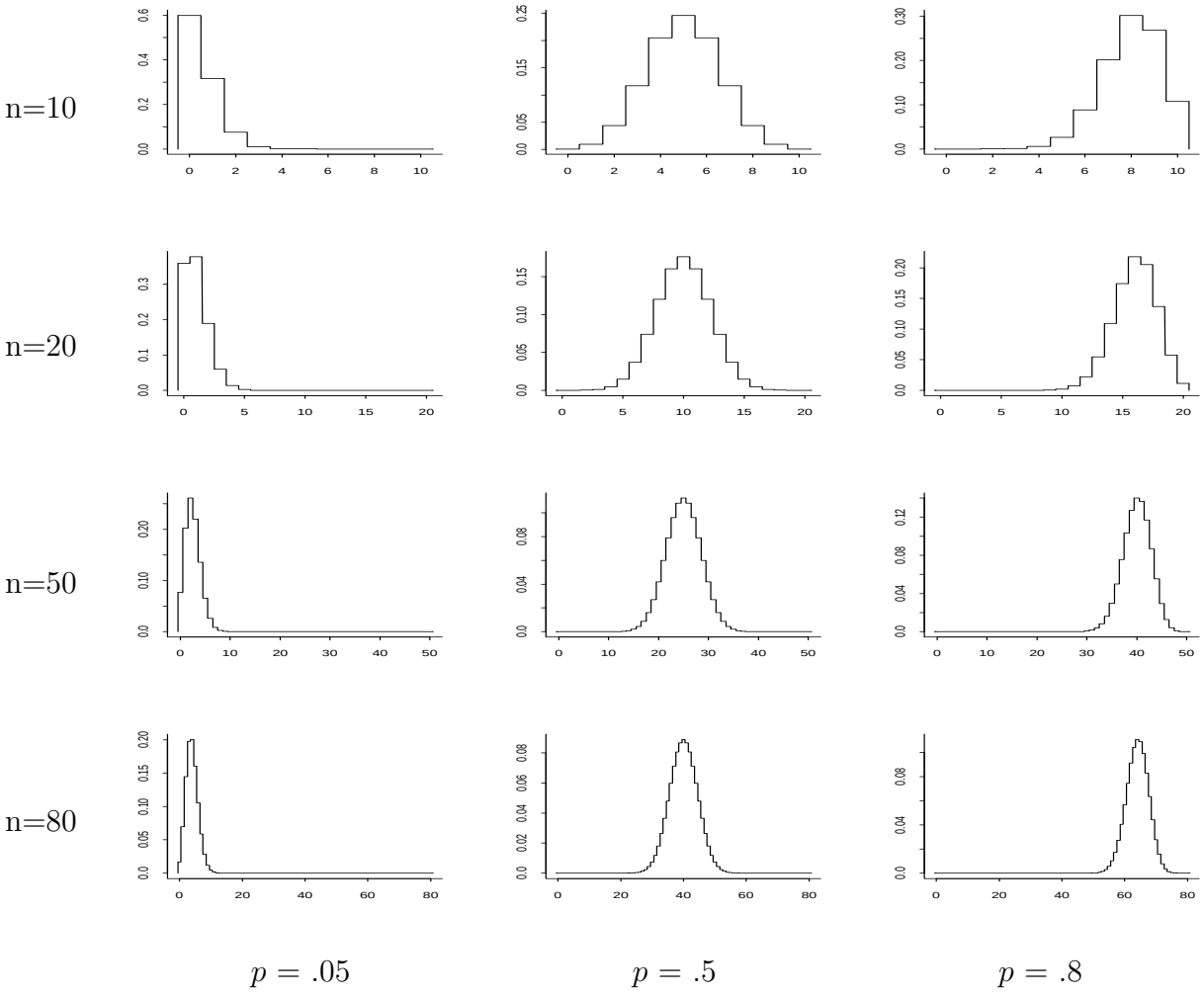


Figure 6.3: Normal approximation to the Binomial

Remark 6.8 We mentioned that Facts 4 and 4′ are equivalent since both these lead to the same **standardization** or Z-scores i.e., If X is a Binomial and n is sufficiently large, we can use the Normal approximation with

$$\mu_X = np \text{ and } \sigma_X = \sqrt{np(1-p)}$$

leading to

$$Z = \frac{(X - np)}{\sqrt{np(1-p)}} = \frac{\frac{1}{n}(X - np)}{\frac{1}{n}\sqrt{np(1-p)}} = \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

which is the standardization one would use for $\widehat{p} = \frac{X}{n}$, from Fact 4′.

Example 6.4: If a fair coin is tossed 100 times, what is the chance that the observed proportion of heads exceeds 0.6?
Solution: If X denotes the number of heads in these 100 tosses, it has a Binomial distribution with $n = 100$ and $p = \frac{1}{2}$. In terms of X, we want $\mathrm{P}(X > 60)$. The exact answer can be obtained by using the binomial probabilities i.e.,

$$\begin{aligned} \mathrm{P}(X > 60) &= \mathrm{P}(X = 61) + \mathrm{P}(X = 62) + \cdots\cdots \\ &= \binom{100}{61}(.5)^{61}(.5)^{39} + \binom{100}{62}(.5)^{62}(.5)^{38} + \cdots\cdots . \end{aligned}$$

Clearly, this is a very difficult expression to evaluate numerically partly because of the large factorials involved. On the other hand, we can use Fact 4′ (or Fact 4), since n is large enough. The observed proportion $\widehat{p}$ is approximately normal with

$$\mu_{\widehat{p}} = p = 0.5$$

and

$$\sigma_{\widehat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(.5)(.5)}{100}} = \frac{.5}{10} = 0.05.$$

Thus, we may compute the probabilities of interest by using the much simpler Normal approximation to the Binomial distribution, namely

$$\mathrm{P}(X > 60) = \mathrm{P}\left(\frac{X}{n} > \frac{60}{100}\right) = \mathrm{P}(\widehat{p} > 0.6) = \mathrm{P}\left(\frac{\widehat{p} - 0.5}{.05} > \frac{0.6 - 0.5}{.05}\right) = \mathrm{P}(Z > 2) = 0.0228.$$

◇

R code:

```
#Example 6.4
# probability function of the binomial distribution can be used to get P(X>60)
n=100
p=1/2
```

```
pbinom(60,size = n,prob = p,lower.tail = FALSE)
# bin(100,1/2) is also approximately N(50,5) since:
norm_mean=n*p
norm_std=sqrt(100*1/2*(1-1/2))
pnorm(60,mean = norm_mean,sd=norm_std,lower.tail = FALSE)
# the Normal approximation to the binomial proportion p_hat, will get an easy
   and close enough answer.
```

The output for this **R** code gives::

```
> n=100
> p=1/2
> pbinom(60,size = n,prob = p,lower.tail = FALSE)
[1] 0.0176001
> norm_mean=n*p
> norm_std=sqrt(100*1/2*(1-1/2))
> pnorm(60,mean = norm_mean,sd=norm_std,lower.tail = FALSE)
[1] 0.02275013
```

Example 6.5: Suppose the true proportion of foreign cars in California is known to be $p = 0.4$. Suppose we observe a sample of $n = 100$ cars. What is the chance that (a) more than half the cars observed are foreign-made? (b) the observed proportion of foreign cars is between 0.35 and 0.45?

Solution: From Fact 4′, $\widehat{p}$ is approximately Normally distributed with

$$\mu_{\widehat{p}} = 0.4, \ \sigma_{\widehat{p}} = \sqrt{\frac{(0.4)(0.6)}{100}} = 0.05.$$

Therefore,

(a) P(50 or more foreign cars) = P($\widehat{p} \geq 0.5$) = P($\frac{\widehat{p}-0.4}{.05} \geq \frac{0.5-0.4}{.05}$) = P($Z \geq 2$) = .025 .

(b) Similarly,

$$\mathrm{P}(.35 < \widehat{p} \leq .45) = \mathrm{P}(-1 < Z \leq 1) = 0.68.$$

◇

Example 6.6: Suppose the admissions office at a certain university sends out 1000 admission letters to prospective students. Suppose it is known from past statistics, that the probability that an admitted student will actually enroll at that university is 0.40. Find the probability that fewer than 420 of the admitted students students will enroll at this university.

Solution: Let X = Number of students who actually enroll among the 1000 offered admission. This is a Binomial distribution with $n = 1000$ and $p = 0.40$. Hence,

$$\mathrm{P}(X \le 420) = \binom{1000}{0}(.4)^0(.6)^{1000} + \binom{1000}{1}(.4)^1(.6)^{999} + \cdots + \binom{1000}{420}(.4)^{420}(.6)^{580}$$

which is considerably messy.

On the other hand, from Fact 4, we know a $\mathrm{Bin}(n,\ p)$ can be approximated by a $\mathrm{N}(np,\ \sqrt{np(1-p)})$. Here

$$\mu_X = np = 1000(.40) = 400,$$
$$\sigma_X = \sqrt{np(1-p)} = \sqrt{1000(.4)(.6)} = \sqrt{240} = 15.5.$$

Since n is large, using this Normal approximation, X has an approximate $\mathrm{N}(400,\ 15.5)$ distribution in this case. Thus

$$\mathrm{P}(X \le 420) = \mathrm{P}\left(\frac{X - 400}{15.5} \le \frac{420 - 400}{15.5}\right) \approx \mathrm{P}(Z \le 1.29) = 0.9015.$$

An equivalent way to handle this problem is to use Fact 4′, which says that $\widehat{p}$ has approximately a $\mathrm{N}\left(p,\ \sqrt{\frac{p(1-p)}{n}}\right)$ distribution — in this example, $\mathrm{N}(0.4,\ 0.0155)$ since

$$\mu_{\widehat{p}} = p = 0.4, \qquad \sigma_{\widehat{p}} = \sqrt{\frac{p(1-p)}{n}} = \frac{(.4)(.6)}{1000} = \sqrt{\frac{.24}{1000}} = 0.0155.$$

Therefore,

$$\mathrm{P}(X \le 420) = \mathrm{P}(\widehat{p} \le .42) = \mathrm{P}(Z \le 1.29) = 0.9015.$$

◇

EXERCISES

6.1 An insurance company is studying the age at death of its clients. If the standard deviation in the population is 6.2 years, how likely is the sample mean based on 64 observations to differ from the unknown population mean by 1 year or less?

6.2 The distribution of scores for persons over 16 years of age on the Wechsler Adult Intelligence Scale (WAIS) is approximately normal with mean 100 and standard deviation 15. The WAIS is one of the most common "IQ tests" for adults.

(a) What is the probability that a randomly chosen individual has a WAIS score of 105 or higher?

(b) What are the mean and standard deviation of the average WAIS score $\overline{x}$ for an independent sample of 60 people?

(c) What is the probability that the average WAIS score of the sample of 60 people is 105 or higher?

(d) Would your answers to (a), (b) or (c) be affected if the distribution of WAIS scores in the adult population were not normal?

6.3 If the SAT verbal scores are normally distributed with mean 420 and a standard deviation of 100 and SAT math scores are distributed normally with mean 480 and standard deviation of 100, answer the following questions:

(a) Assuming independence, find the probability one gets more than 600 on verbal AND more than 700 on math. Is this the same as asking the chance of the total score (verbal + Math) being more than 1300? Without calculating, which of these outcomes do you think has a higher probability?

(b) If 10 tenth-graders attempt this exam in a particular high school, what is the chance their average math score is more than 550?

6.4 Suppose the lifetime of light bulbs follow a distribution with mean 700 hours and standard deviation of 25.

(a) For a random sample of 100 light bulbs, give an approximate probability distribution for the sample average, and give justification through a fact or theorem.

(b) For a random sample of 100 light bulbs, find an approximate probability that the sample average of lifetimes is greater than 750.

6.5 A soda fountain dispenses to any customer an amount X at random anywhere between 6.5 and 7.5 ounces, i.e., X is uniformly distributed between 6.5 and 7.5 ounces.

(a) Is the amount dispensed by the beverage machine a discrete or a continuous random variable? Explain.

(b) Find the probability that the amount of beverage dispensed to a randomly chosen customer is more than 7 ounces.

(c) Suppose 49 customers used the particular machine during the last hour. Find the (approximate) probability that the total amount of beverage dispensed in the last hour is more than 350 ounces. (Assume that the mean and variance of X are 7 and $\frac{1}{12}$ respectively.)

6.6 The length of a TV commercial (in seconds) is a random variable with the following probability distribution:

Length (in secs.)	10	15	30	45	60
Probability	.11	.27	$2x$	x	.02

(a) What is the value of x?

(b) What is the average length of a TV commercial?

(c) Suppose we randomly choose 36 TV commercials and calculate their mean length. What is the standard deviation of this mean?

(d) What is the approximate probability that the total length of 36 randomly chosen TV commercials is between 15 and 20 minutes?

6.7 A buyer for a lumber company must determine whether to buy a piece of land containing 5,000 pine trees. If at least 1,000 of the trees are 32 feet or taller, the buyer will purchase the land; otherwise, she will not. The owner of the land reports that the distribution of the heights of the trees has mean of 30 feet, and standard deviation of 3 feet. Based on this information, what should the buyer decide? State any assumptions you made.

6.8 The probability that a door-to-door salesperson makes a sale when she visits a house is 0.2. There are 10 houses on a particular street.

(a) What is the probability that she makes at least one sale on this street?

(b) What is the probability that the number of sales is less than or equal to 3?

(c) On a typical day, the salesperson visits 100 houses. What is the approximate probability that she makes a sale in at least 40% them?

6.9 The probability that a person with a certain disease is cured after taking a new medicine is 0.40. Suppose there are 10 patients and the probability of a patient being cured is independent of what happens to the other patients.

(a) What is the exact probability that at least three patients out of the ten will be cured?

(b) What is the mean and the standard deviation of the number of patients that will be cured out of the 10?

(c) Suppose that new medicine has now been tested out on a different group of 500 patients. What is the approximate probability that at least 190 of the patients will be cured?

6.10 A restaurant offers its patrons free dinner on the 10th of each month, if their birthday falls on that month. What is the approximate probability that out of 200 customers that they plan to serve on the coming May 10th, more than 25 will be entitled to a free meal? (You may assume that the birthday of a person is equally likely to fall in any one of the 12 months.)

6.11 A radio station in Santa Barbara claims 30% of the people in town listen to it on a regular basis i.e., p = 0.3. In a random sample of 200 radio listeners, find the probability that the number who tune to this station is 50 or less.

6.12 A new drug aimed at alleviating mental depression seems to be 80% effective, according to an intensive prelicensing testing program.

(a) If the new drug is prescribed for 15 patients, with what probability can we expect fewer than 11 to benefit from it?

(b) If the drug is prescribed for 150 patients, how likely is it that fewer than 110 will benefit?

(c) Of 1500 patients taking this drug, what are the chances that fewer than 1100 will benefit from it?

(d) With 15,000 patients being given the new drug, what is the probability that fewer than 11,000 will benefit from it?

6.13 A statistics professor decides to give a 20-question true-false quiz to determine who has read the weekly assignment. She wants to choose the passing mark such that the probability of passing a student who guesses on every question is less than .05. What score should she set as the lowest passing grade?

6.14 Suppose that 20% of all adults jog regularly.

(a) If we selected 6 people randomly, what is the chance that 2 of these 6 jog regularly?

(b) Suppose we take a random sample of 200 adults and observe the proportion of adults who jog in our sample. Find the probability that this is between 15% and 22%.

6.15 Suppose there is a 80% chance that an insect egg hatches and becomes an adult. If a "mommy insect" lays 100 eggs, find the chance that 75 or more of these eggs hatch and grow into adults.

6.16 Suppose 55% of self-employed workers in the United States do not have health insurance coverage (i.e. they are uninsured). 100 self-employed workers are randomly surveyed. What is the probability that at least one half of those surveyed will be uninsured?

Chapter 7

Estimation and Confidence Intervals

Thus far, we have been describing how the number of successes follows a Binomial distribution and how variables like heights or scores or their means follow a Normal curve. We have assumed all this time that we knew the so-called **parameters** — the true proportion or probability p in a Binomial distribution or the μ and σ for a Normal curve. However, in practice, these parameters are typically unknown and one of the main purposes of statistics is (i) to estimate the parameters or (ii) to test hypotheses about these unknown parameters using sample data. This is, as we said before, part of **statistical inference** and in this chapter, we will try to cover some basic ideas in estimation.

7.1 Point estimates

The parameters p, μ or σ are generally unknown. For instance, in the Binomial context, the true proportion or "probability of success" p is unknown. However, if we conducted the Binomial experiment n times and observed say, X successes among these n trials, it is intuitively clear that the observed proportion of successes in n trials, $\widehat{p} = \frac{X}{n}$ is a reasonable estimate of p. Indeed, in Fact $4'$ in Chapter 6, we observed

$$\mu_{\widehat{p}} = p,$$

so that such $\widehat{p}$ values, which can vary from one sample to the other, still center at the true (but unknown) value p. This means that $\widehat{p}$ is an **unbiased** estimator of p. Similarly, in the

case of data $(X_1,\ X_2,\ \cdots,\ X_n)$ from a Normal curve with unknown center μ, the sample mean

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n},$$

is a natural heuristic first guess at this μ. Again Fact 1 of Chapter 6 tells us that

$$\mu_{\overline{X}} = \mu,$$

so that $\overline{X}$ is indeed an **unbiased** estimator of μ. Similarly

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

turns out to be a good estimate of the variance, σ^2.

Such estimates are called **point estimates** since they give a single value or point as an estimate of the corresponding unknown quantity. In fact, it can be shown that these are all very good point estimates in the sense that they are not only **unbiased** (i.e., centered at the value we are trying to estimate) but also have the smallest possible spread around that center among all unbiased estimators (refer Section 6.1).

Remark 7.1 We use "$\widehat{\ }$" on top of an unknown parameter to indicate that it is an estimate of that parameter. For instance, we have already used $\widehat{p}$ for an estimate of p . Thus, we can summarize our point estimates by writing an estimate of p as

$$\widehat{p} = \frac{X}{n},\ \text{if } X \text{ successes are observed in } n \text{ trials.}$$

If $(X_1,\ X_2,\ \ldots,\ X_n)$ is a random sample from $\mathrm{N}(\mu,\ \sigma)$, then

$$\widehat{\mu} = \overline{X}$$

and

$$\widehat{\sigma^2} = S^2.$$

These are are known to be the *best* (unbiased as well as having minimum variance) point estimates of the corresponding parameters.

R Code Instruction:

In **R**, if the unbiased estimator is a mean of the sample data, it can be obtained using the built-in function *mean*( ) that was mentioned before. Similarly for the estimator of population proportion p, to calculate $\widehat{p}$, the equation $\widehat{p} = \frac{X}{n}$ is used. As an example, the following data set contains different genders and we calculate the proportion of females in the data set.

Example :

```
dataset1=c("female","male","male","female","male","female","female")
X=sum(dataset1=="female")
n=length(dataset1)
phat=X/n
print(phat)
```

The output for this **R** code gives::

```
> dataset1=c("female","male","male","female","male","female","female")
> X=sum(dataset1=="female")
> n=length(dataset1)
> phat=X/n
> print(phat)
[1] 0.5714286
```

Note: For estimating the variance, the built-in function *var*() is used as mentioned in Chapter 2.

7.2 Confidence interval for mean with known σ

An alternative to finding point estimates, is to seek a possible range of values for the unknown parameter and to state along with it the confidence that we have in such an interval of values. The latter approach is called **interval estimation** and the intervals are called **confidence intervals**.

If $(X_1, \ldots, , X_n)$ is a sample of n observations, each independently drawn from a Normal

curve with mean μ and standard deviation σ, we saw (Fact 2) that

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

has a Normal distribution with $\mu_{\overline{X}} = \mu$, $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$. Hence,

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

has a N(0, 1) distribution. Thus, from the 68-95-99.7 rule, or more accurately from Table A, we can claim that the Z-value lies between $(-1.96,\ 1.96)$ with a 95% probability. In other words, the chance is 95% that

$$-1.96 \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96$$

or

$$(-1.96)\frac{\sigma}{\sqrt{n}} \leq (\overline{X} - \mu) \leq (1.96)\frac{\sigma}{\sqrt{n}}$$

or

$$\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}.$$

Thus,

The interval $(\overline{X} \pm\ 1.96\frac{\sigma}{\sqrt{n}})$ contains μ with 95% chance.

This interval, $(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}},\ \overline{X} + 1.96\frac{\sigma}{\sqrt{n}})$, is entirely known once we have the data and σ is known and can hence be used as an interval estimate of μ. The interval is called a **95% confidence interval for μ when σ is known** and the corresponding probability, 95%, is called the **confidence level** or **confidence coefficient**.

We can, of course, arrange for any other confidence level C (usually 0.90, 0.95, 0.99, etc.) by looking up Table A for the appropriate Z-value. If C is the desired confidence level, we need a Z-value to correspond to a *right tail area* of $\frac{1-C}{2}$. It is more common to use $(1 - \alpha)$ in place of C, where α is some small given value (usually 0.1, 0.05, 0.01, etc.) Then, we seek

values $-z$ and z such that there is a probability $(1-\alpha)$ in between these two values and $\frac{\alpha}{2}$ in each tail. We denote such a z by $z_{\alpha/2}$. This makes the area between $-z_{\alpha/2}$ and $z_{\alpha/2}$ the probability $C = (1-\alpha)$ that we desire.

For instance, 1.96 corresponds to a 95% confidence level, while 2.576 corresponds to a 99% confidence level. Or, in other words, $z_{.025} = 1.96$ and $z_{.005} = 2.576$.

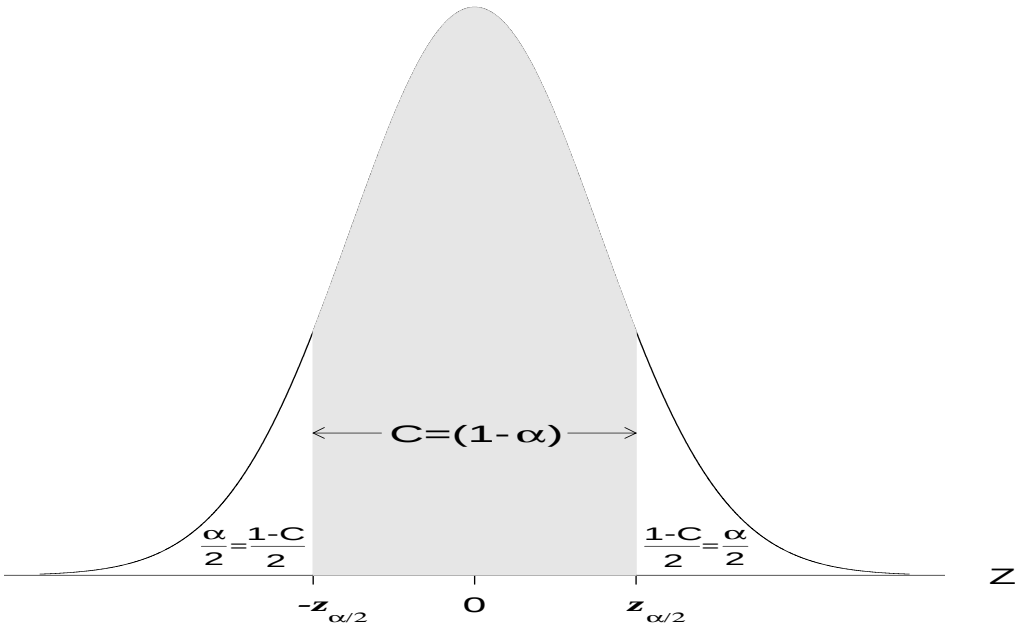


Figure 7.1: Confidence level, $C = (1-\alpha)$

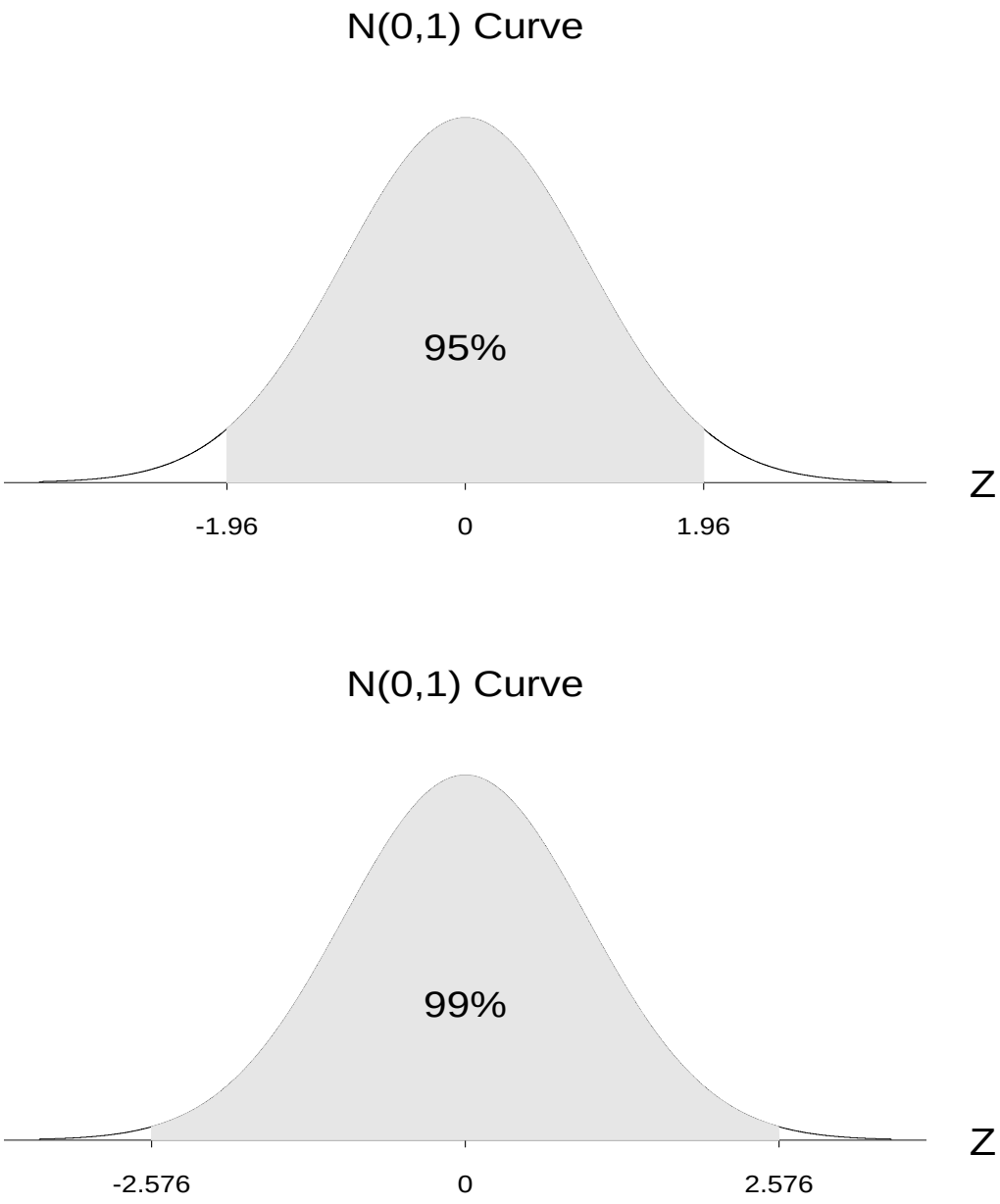


Figure 7.2: Confidence interval for Normal mean

The $z_{\alpha/2}$ values corresponding to different confidence levels are obtained readily from the last row of Table C of this book.

We may now restate the more general confidence interval for the unknown mean μ, as follows:

A $C = (1 - \alpha)$ confidence interval for μ is given by

$$\overline{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

if the data is from a Normal curve with σ *known*.

R Code Instruction:

From what has already been discussed, we know that $\overline{X}$ follows a normal distribution with $\mu_{\overline{X}}=\mu$, and $\sigma_{\overline{X}}=\frac{\sigma}{\sqrt{n}}$, so that the interval estimator with known σ is obtained as $\overline{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$ with 95% confidence. Recall that the interval (-1.96, 1.96) contains 95% probability under the standard normal distribution. In **R**, we can use the built-in function *qnorm*() to find such values or z-scores corresponding to any specified probability. For Example:

```
qnorm(0.95)
qnorm(0.975)
```

The output for this **R** code gives:

```
> qnorm(0.95)
[1] 1.644854
> qnorm(0.975)
[1] 1.959964
```

A confidence interval in this case can be found through R using *qnorm*() function. We will use Example 7.1 to illustrate this.

Example 7.1: Suppose X, the scores of a test on numerical ability, have a Normal distribution with unknown μ and known σ= 60. Suppose the sample values ($X_1, \ldots, X_{900}$) give a sample mean $\overline{X} = 272$. Find a 95% as well as a 98% confidence interval for the (unknown) true mean μ.

Solution: We have

$$\widehat{\mu} = \overline{X} = 272,$$

as a point estimate of μ. A 95% confidence interval for μ is given by:

$$\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}},\ \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

$$= \left(272 - 1.96\frac{60}{\sqrt{900}},\ 272 + 1.96\frac{60}{\sqrt{900}}\right)$$

$$= (272 - 1.96 \times 2,\ 272 + 1.96 \times 2)$$
$$= (272 - 3.92,\ 272 + 3.92)$$
$$= (268.08,\ 275.92).$$

For a 98% confidence interval for μ, we replace 1.96 by 2.326 (see Table C) and get

$$272 \pm (2.326) \times 2 = 272 \pm 4.652 = (267.348,\ 276.652).$$

◇

R code:

```
xbar=272
sigma=60
n=900
ci=qnorm(0.95)*(sigma/sqrt(n))
confidenceInterval=xbar+c(ci,-ci)
print(confidenceInterval)
```

The output for this **R** code gives:

```
> xbar=272
> sigma=60
> n=900
> ci=qnorm(0.95)*(sigma/sqrt(n))
> confidenceInterval=xbar+c(ci,-ci)
> print(confidenceInterval)
[1] 275.2897 268.7103
```

Remark 7.2 From Fact 3, we know that if the sample size n is large,

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

is approximately N(0, 1), no matter what the parent population is. We can therefore use the same confidence intervals and confidence statements for μ for *any data set, provided* n *is large.*

7.3 Choice of sample size

The margin of error (or half-width of the confidence interval), $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, depends on the following three factors:

- the confidence level desired (through $z_{\alpha/2}$)
- the variation in the data, σ
- the sample size, n.

Given the first two, we can ask "What is the sample size n, needed for the margin of error to be a specified value, say m?"

In other words, find n so that

$$m = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$\text{or,} \qquad \sqrt{n} = z_{\alpha/2}\frac{\sigma}{m}$$

$$\text{or,} \qquad n = \frac{z^2_{\alpha/2}\sigma^2}{m^2}.$$

Thus,

For constructing a $C = (1 - \alpha)$ level confidence interval for μ when σ is known,

$$n = \left(z_{\alpha/2} \frac{\sigma}{m} \right)^2 = \frac{z_{\alpha/2}^2 \sigma^2}{m^2}$$

is the sample size required for the margin of error to be m.

Example 7.2: Suppose $(X_1, \ \ldots \ , \ X_n)$ are heights from N$(\mu, \ \sigma)$, with σ given to be 10". If in a sample of size 16, we obtain $\overline{X} = 60$" , find a 95% confidence interval for μ. Find the n needed in order that the margin of error $m = 3$ inches.

Solution: Since we want a confidence level of 95%, C = 0.95. From the last row of Table C, the appropriate $z_{\alpha/2} = 1.96$. Thus, the required confidence interval is:

$$\begin{aligned} & \overline{x} \pm \ z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ = \ & 60 \pm 1.96 \frac{10}{\sqrt{16}} \\ = \ & 60 \pm 4.9 \\ = \ & (55.1, \ 64.9) \end{aligned}$$

However, if we wish to keep the margin of error to $m = 3$ inches, then the required sample size is obtained by substituting in the equation above and we obtain

$$n = \left(z_{\alpha/2} \frac{\sigma}{m} \right)^2 = \left(1.96 \times \frac{10}{3} \right)^2 = 42.68,$$

which we *round up* to 43. Thus if we increase the sample size to 43 instead of the original 16, we can reduce the margin of error from 4.9 inches to 3 inches.

$\diamond$

7.4 Confidence interval for mean with unknown σ

If σ is not known, estimate this by the sample standard deviation S, where

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

Then, in the Z-statistic, we replace σ by S, getting

$$t \equiv \frac{\overline{X} - \mu}{S/\sqrt{n}}.$$

Notice that when we know σ we use it to compute the Z-statistic whereas when it is unknown, we replace σ by the sample estimate S. The result is now called a **t-statistic** and it does not follow the N(0,1) distribution any longer.

This t-statistic however follows a $t(n-1)$ distribution where $(n-1)$ is called the **degrees of freedom (df)**. A t-distribution is very similar in shape to the Standard Normal or Z distribution except that it has somewhat *heavier (or fatter) tails* (See Figure 7.3). In other words, there is less concentration near the center compared to the Standard Normal curve. As the degrees of freedom of a t distribution increases, it becomes indistinguishable from a N(0, 1) distribution. The t-values corresponding to the commonly used upper-tail areas are given in Table C. Each row corresponds to a different df with the last row (∞ df) giving us the Z-values of the N(0, 1) distribution.

Recall when σ is known, we used the fact that

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

has a N(0, 1) distribution to obtain the confidence interval for μ, given by

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

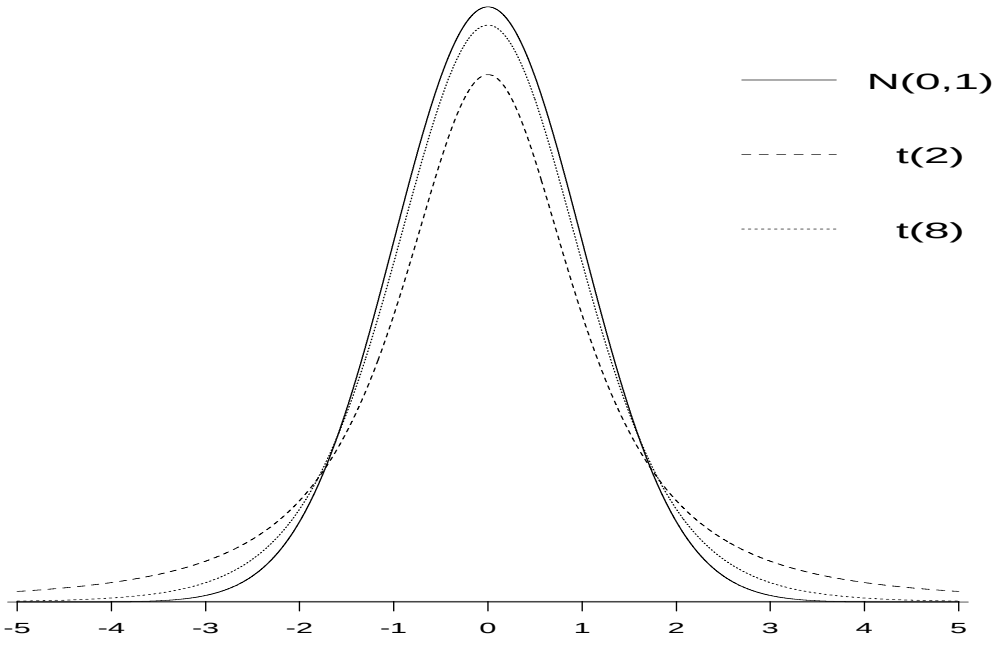


Figure 7.3: t-distributions and the Standard Normal distribution

Similarly when σ is unknown, we can replace σ by S and use the corresponding fact that

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has a $t(n-1)$ distribution. This results in the confidence interval

$$\overline{X} \pm \ t_{\alpha/2} \ \times \frac{S}{\sqrt{n}},$$

with confidence level $(1-\alpha)$. Note that $t_{\alpha/2}$ coming from Table C, corresponding to $(n-1)$ df, satisfies:

$$\mathrm{P}\left(t(n-1) > t_{\alpha/2}\right) = \frac{\alpha}{2}.$$

We thus get:

A $C = (1 - \alpha)$ confidence interval for μ is given by

$$\overline{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}}$$

when data is from a Normal curve with σ *unknown*.

R Code Instruction:

For finding a confidence interval for the mean with unknown variance, t-score is used instead of the z-score. The **R** program can find the t-score for specified probability, using the built-in function

```
qt(p, df)
```

where *df* represents the degree of freedom.

Example 7.3: A scientist who is studying the brain-weights of tigers took a random sample 16 animals and measured their brain-weights in ounces. Suppose this data gave a sample mean $\overline{x}$ =10 and standard deviation $s = 3.2$. Assuming that these weights follow a Normal distribution, find a 95% confidence interval for the true mean weight μ.
Solution: Here, $n = 16$, $\overline{X} = 10$, $S = 3.2$. Since we want a 95% confidence, $\alpha = 0.05$ and $\alpha/2$= 0.025. Looking up Table C, we have $t_{\alpha/2} = 2.131$ corresponding to 15 df, so that the required confidence interval is

$$10 \pm \ 2.131 \times \frac{3.2}{\sqrt{16}} = 10 \pm 1.705 = (8.295, \ 11.705).$$

R code:

```
xbar1=10
s=3.2
n=16
ci1=qt(0.95,df=n-1)*(s/sqrt(n))
ConfidenceInterval=xbar1+c(-ci1,ci1)
print(ConfidenceInterval)
```

The output for this **R** code gives:

```
> xbar1=10
> s=3.2
> n=16
> ci1=qt(0.95,df=n-1)*(s/sqrt(n))
> ConfidenceInterval=xbar1+c(-ci1,ci1)
> print(ConfidenceInterval)
[1]  8.29556 11.70544
```

Thus we are 95% confident that the true mean weight of tiger's brain is somewhere between 8.295 and 11.705 ozs. (Just for comparison, a human brain weighs on the average 46 ozs while an African elephant has a brain weighing 158 ozs, on the average). ◇

7.5 Confidence interval for σ

We mentioned in Section 7.1 that the sample variance, S^2, is a good estimate of the true (population) variance, σ^2. To obtain a confidence interval for σ (or σ^2), we need to introduce yet another distribution called the **chi-square** (written as χ^2) **distribution**.

Fact: If $(X_1,\ X_2,\ \ldots,\ X_n)$ is a sample from a N$(\mu,\ \sigma)$ population, then

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

has a chi-square distribution with $(n-1)$ degrees of freedom (df), written as $\chi^2(n-1)$.

The critical values corresponding to chi-square distributions with different degrees of freedom are tabulated in Table D. Unlike the Normal or t distributions, the χ^2 distribution is not quite symmetric around any value. See Figure 7.4 for some examples of χ^2 distribution.

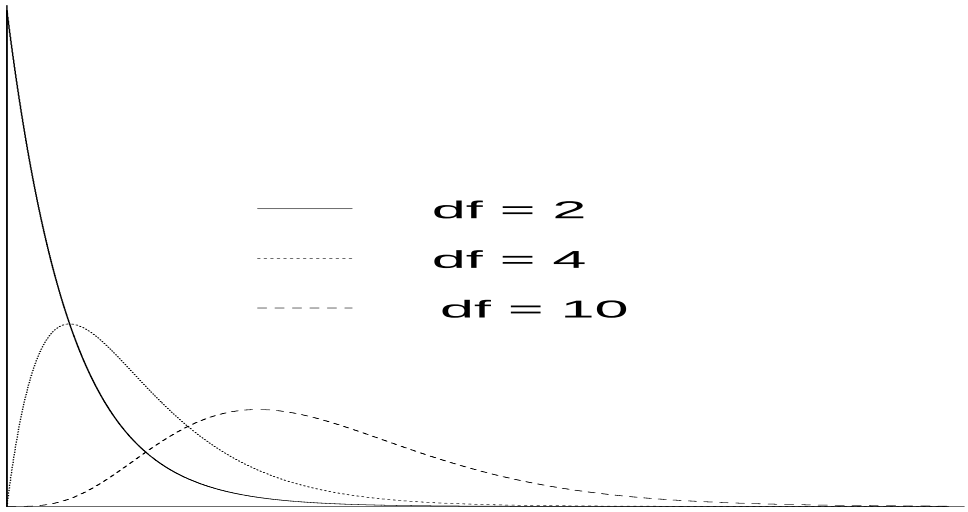


Figure 7.4: Density curves of χ^2 distributions with 2, 4 and 10 df

The procedure we are about to describe (called the equal-tails method) works well for reasonably large n, although choosing equal tails is not the "Optimal" thing to do. If we wish to set a $C = (1 - \alpha)$ level confidence interval for σ for a given sample size n, consult a $\chi^2(n-1)$ distribution and find the low and high values $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$ respectively so that there is a probability of $\frac{\alpha}{2}$ on each tail (and hence a probability of C = $(1 - \alpha)$ in between $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$). Such values can be found from Table D.

Then the probability is $C = (1 - \alpha)$ that

$$\chi^2_{1-\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\alpha/2}$$

which can be equivalently rewritten to say

$$\sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2}}} \leq \sigma \leq \sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}}.$$

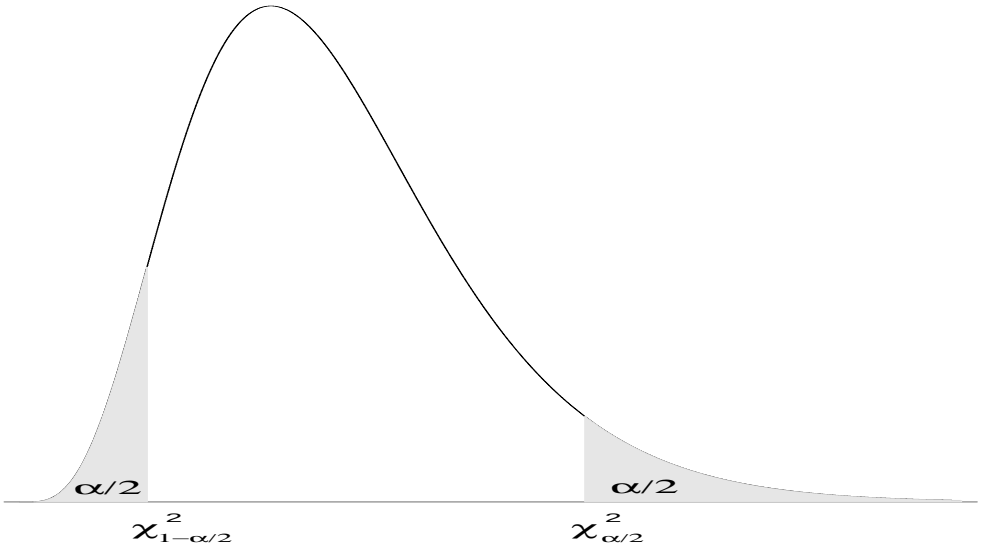


Figure 7.5: Confidence region for a χ^2 distribution

A $C = (1 - \alpha)$ level confidence interval for the standard deviation σ based on a sample from a N(μ, σ), is given by

$$\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}} \right)$$

where $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$ are values from a $\chi^2_{(n-1)}$ distribution with areas $\frac{\alpha}{2}$ below $\chi^2_{1-\alpha/2}$ and above $\chi^2_{\alpha/2}$, respectively.

R Code Instruction:

For finding a confidence interval for σ, **R** programming can find χ_α, using the built-in function:

```
qchisq(p, df)
```

where *df* represents the degree of freedom.

Example 7.3 (contd.): Recall here $n = 16$, $\overline{X} = 10$ and $S = 3.2$. Find a 95% confidence interval for σ.

Solution: We consult a $\chi^2(15)$ distribution to find $\chi^2_{0.95} = 7.26$ as the value with 5% probability below it and $\chi^2_{0.05} = 25.00$ as the value with 5% probability above. Therefore, the 90% confidence interval is

$$\left(\sqrt{\frac{15 \times (3.2)^2}{(25.00)^2}}, \sqrt{\frac{15 \times (3.2)^2}{(7.26)^2}}\right)$$

or

$$(2.48,\ 4.60).$$

That is, we have 90% confidence that the true σ is in between 2.48 and 4.60.

R code:

```
s=3.2
n=16
chi1=qchisq(0.95,df=n-1)
chi2=qchisq(0.05,df=n-1)
ci1=sqrt((n-1)*(s^2)/(chi1^2))
ci2=sqrt((n-1)*(s^2)/(chi2^2))
Ci=s+c(-ci1,ci2)
print(Ci)
```

The output for the **R** code gives:

```
> print(Ci)
[1] 2.704175 4.906878
```

◇

7.6 Confidence interval for proportion

In the Binomial context, suppose p = probability of success is unknown (p is also called the population proportion or the true proportion). To estimate p, we take a sample of n

observations and suppose X of these result in successes. Then we have seen that

$$\widehat{p} = \frac{X}{n} = \text{ observed proportion}$$

is a good point estimate of p. For instance, if 10 tosses result in 6 heads, then $\widehat{p} = \frac{6}{10} = 0.6$ is a point estimate of "the probability of getting heads".

For large n, $\widehat{p}$ is again approximately Normally distributed (Fact 4′) with

$$\mu_{\widehat{p}} = p \quad \text{ and } \sigma_{\widehat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

so that

$$Z = \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

has an approximate N(0, 1) distribution. Therefore, using the $z_{\alpha/2}$ corresponding to a specified confidence level $C = (1 - \alpha)$, there is a chance C = $(1 - \alpha)$ that

$$-z_{\alpha/2} \leq \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2}$$

or

$$\widehat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \widehat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}.$$

In large samples, it is justified to plug in $\widehat{p}$ in place of p inside the square root on either side of the above inequality. This results in

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

as the confidence interval for p with confidence level C.

A $C = (1 - \alpha)$ confidence interval for the population proportion p is given by

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},$$

where $\widehat{p} = \frac{X}{n}$ is the observed sample proportion and the sample size n is large.

R code Example:

```
dataset1=c("female","male","male","female","male","female","female")
X=sum(dataset1=="female")
n=length(dataset1)
phat=X/n
qhat=1-phat
ci2=qnorm(0.90)*sqrt(phat*qhat/n)
confidenceint=phat+c(-ci2,ci2)
print(confidenceint)
```

The Output for the **R** code gives:

```
 > print(confidenceint)
[1] 0.3317222 0.8111350
```

As before, we can ask how large a sample we should draw to attain a given precision. Suppose we wish to find n such that the margin of error $z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$ is some pre-selected value m. Solving an equation like the one in the last section, we have

$$n = \frac{z^2_{\alpha/2}p(1-p)}{m^2}.$$

Since p is unknown, we have to plug in our best guess of p, say p_0, in the above equation. If no guess is available, a conservative approach is to take $p = \frac{1}{2}$. This choice gives us the maximum sample size needed, no matter what the true value of p is. This results in

$$n = \frac{z^2_{\alpha/2}(\frac{1}{2})^2}{m^2} = \frac{z^2_{\alpha/2}}{4m^2}.$$

For constructing a $C = (1 - \alpha)$ level confidence interval for p

$$n = \frac{z^2_{\alpha/2}\, p_0(1 - p_0)}{m^2},$$

is the sample size required for the margin of error to be m, where p_0 is the best prior guess about p.
If no prior information about p is available, use the conservative value $p_0 = \frac{1}{2}$.

Example 7.4: In obtaining a 90% confidence interval for p, if it is desired to have an error of no more than 3 percentage points i.e., $m = 0.03$, how large an n do we need?
Solution: Since $(1 - \alpha) = 0.90$, $\frac{\alpha}{2} = .05$ and $z_{\alpha/2} = 1.645$. We want $m = 0.03$. Therefore,

$$n = \frac{(1.645)^2}{4(.03)^2} = 751.67$$

or, rounding upwards, we need a sample of $n = 752$.

◇

R code:

```
z=qnorm((1-0.9)/2,lower.tail = FALSE)
m=0.03
n=z^2/(4*m^2)
n
```

The output for this **R** code gives:

```
> n
[1] 751.5398
```

EXERCISES

7.1 A psychologist is studying "learning" in rats and wants to determine the overall time required for rats to learn to traverse a maze. She randomly selected 12 rats and found

the times they took to be:

5.1, 1.6, 6.8, 2.3, 5.4, 1.8, 2.6, 6.2, 3.6, 3.4, 4.4, 4.8

(a) Assuming this data is from a N(μ, σ) distribution where σ is known to be 3.7, estimate the average time required for rats to learn to traverse this maze with a 90% confidence interval.

(b) If we wished to estimate this with 90% confidence and with a margin of error, m = 1 minute, how large a sample will be needed?

7.2 The League for Moral Rectitude felt it was necessary to verify if the average hemline of skirts worn by girls at a local college is more than 4" above the knee. Suppose μ denotes the true mean (inches above the knee). Careful measurement of a random sample of 15 girls yielded the following data (in terms of inches above the knee):

2, 5, 7, 4, 5, 7, 6, 8, 0, 9, 4, 5, 3, 2, 4

Find a 95% confidence interval for μ if

(a) the true standard deviation σ is known to be 1.5".

(b) σ is not known and is estimated from the data.

7.3 If a sample of 40 men have a mean weight of 178 pounds with a sample standard deviation s of 8.5, find a 99% confidence interval for the population mean weight.

7.4 A company is interested in estimating the mean number of days of sick leave taken by its employees.

(a) The firm's statistician selects 25 personnel files at random and notes the number of sick days taken by each employee. The following sample statistics are computed: $\overline{x}$ = 12.2 days and s = 10 days. Find a 90% confidence interval for the mean number of sick days taken by each employee.

(b) How many personnel files would the statistician have to go through to estimate the mean number of sick days within a margin of error of 2 days with a 99% confidence interval? (Assume σ = 10 days)

7.5 A cigarette manufacturer measures the "nicotine content" in a particular brand of cigarettes he produces. A sample of 6 cigarettes give the following nicotine contents (in mgs.):

21, 19, 15, 18, 12, 17

Find a 95% confidence interval for the mean nicotine content,μ for this brand of cigarettes. (Assume the nicotine content is normally distributed with both μ and σ unknown.)

7.6 A random sample of the birth weights of 20 babies born in a Santa Barbara hospital are recorded. This sample has $\overline{x} = 6.87$ lbs. and $s = 1.76$. Find a 99% confidence interval for the mean weight of babies born in this hospital.

7.7 Nine measurements of the percentage of sugar in 9 boxes of cereal A yielded the following results:

20.0, 20.2, 20.1, 19.9, 20.0, 19.8, 19.9, 20.2, 19.9

Construct a 90% confidence interval for the mean percentage of sugar in cereal A.

7.8 A physical model suggests that the mean temperature increase in the water used as coolant in a compressor chamber should not be more than 5°C. Temperature increases in the coolant measured on 8 independent runs of the compressing unit revealed the following data:

6.4, 4.3, 5.7, 4.9, 6.5, 5.9, 6.4, 5.1

Give a 95% confidence interval for the mean increase of the temperature in the coolant.

7.9 The amount of dissolved oxygen in water is an important measure of the quality of water and its ability to support aquatic life. The following readings (mg/l) were obtained when 12 random samples downstream from an industrial region were tested.

7.13 6.68 6.14 6.39 5.14 5.28 4.47 5.75 7.05 4.78 6.79 5.67

Estimate the true mean content μ using a 95% confidence interval.

7.10 A manager of a large production facility wants to determine the average time required to assemble a widget. A random sample of the times to produce 15 assembled widgets gave $\overline{x} = 15.2$ minutes and $s = 2.2$ minutes.

(a) Assuming that the assembly times are normally distributed, estimate the mean assembly time with 95% confidence.

(b) How would your answer to part (a) change if the population standard deviation were known to be 2.0 minutes?

7.11 If 12 out of a sample of 50 Halloween-revelers at Isla Vista are from out of town, find a 95% confidence interval for the true proportion of people coming from out of town for the Halloween.

7.12 In a telephone survey on the subject of the death penalty, 250 out of the 400 people contacted said they are in favor. Find a point estimate for the "true" proportion of people who support death penalty, as well as a 95% confidence interval for it.

7.13 A random sample of 200 persons from the labor force of a large city are interviewed, and 22 of them are found to be unemployed. Give a 95% confidence interval for the unemployment rate in the city.

7.14 A news reporter would like to predict the percentage of members of Congress who favor a certain controversial change in the banking laws. If a poll of 120 members yields 72 who support the change, find a 90% confidence interval for the true percentage of members favoring the change.

7.15 A news organization is interested in finding out the proportion of people who believe the explanation of a high-ranking government official concerning an alleged incident.

(a) Find the smallest sample size needed to attain a desired margin of error of 0.05 for a 95% confidence interval for the true proportion.

(b) Suppose now that out of 100 people surveyed randomly, 55 people were found to believe the explanation. Give a 90% confidence interval for the true proportion of people who believe the official.

(c) If 10 people are sampled randomly and the true proportion is 0.5, what is the exact probability that the number of people who believe the official is less than or equal to 8?

7.16 A survey is to be conducted to estimate the true proportion of faculty in the UC system who favor "Affirmative Action". How large a sample should be used if we want that with 90% confidence, the sample proportion will not differ from the true proportion by more than 0.03?

7.17 An employee of the CHP desires to estimate the true proportion p of California drivers who wear seat belts.

(a) How many drivers should be sampled in order that the sample proportion will not differ from the true proportion by more than 0.03 with 90% confidence.

(b) Suppose 100 cars were randomly stopped and 85 of these drivers wear seat belts on a regular basis, find a 95% confidence interval for p, the true proportion.

7.18 You have been asked to conduct a survey to determine the proportion of students at UCSB who favor a fee increase to support a UCEN expansion. If you would like your $\widehat{p}$ to be in error of no more than ± 0.05 when constructing a 90% confidence interval, find how large a sample you will need for your survey.

7.19 (a) Of the 500 cars observed on a California freeway, 160 were foreign imports. Find a 90% confidence interval for the true proportion of foreign imports, assuming that this is a representative sample of cars in California.

(b) A survey is to be conducted to estimate the proportion of citizens who favor trade restriction on imports. How large should the sample be so that with 98% confidence, the sample proportion will not differ from the true proportion by more than 0.04?

7.20 A personal computer manufacturer's marketing division is interested in determining the proportion of potential customers that would be willing to buy a computer with a non-Intel microprocessor. After a random sample of 1000 potential customers is surveyed, the sample proportion is found to be 0.45. Give a 90% confidence interval for the true proportion.

Chapter 8

Testing Hypotheses for a single sample

In the last chapter, we were concerned with estimating unknown parameters by using point estimates as well as through confidence intervals. Sometimes, we are more concerned with verifying if the observed data fits a certain hypothesis about these parameters or contradicts it. A **hypothesis is a statement about the unknown parameter**, say p or μ. For instance, if p is the probability of observing a heads after flipping a coin, we may wish to test the hypothesis that the coin is fair, i.e.,

$$\mathrm{H} : p = \frac{1}{2}.$$

Or, if μ refers to the true or population average height of individuals, we may wish to verify if the observed sample data is consistent with the hypothesis

$$\mathrm{H} : \mu = 64 \text{ inches}.$$

Based on the sample data, we wish to test if the hypothesis H is true or false and accordingly accept or reject it. If it is consistent with the data, we have no reason to reject H (i.e., we accept H) and otherwise, we reject H. Before we accept or reject a hypothesis, we should ask what the alternatives are and set up an alternative hypothesis, relative to which we judge the original hypothesis.

8.1 Introduction

A **null hypothesis**, written H_0, is a statement denoting "no effect", "no change" or "status quo". An **alternative hypothesis**, written H_a , reflects the "expected change" or what is called the "research hypothesis".

In hypotheses testing context, we take the attitude that we will hold on to the null hypothesis as true and reject it only if there is sufficient evidence against it — much the same attitude as expressed in statements like "innocent until proven guilty" or the conservative philosophy, "do not fix things if they are not broken". For instance,

Example 8.1: A coin when tossed 100 times, gives 62 heads. Is this a fair coin? Set up H_0 and H_a.
Solution: We assume the coin is fair until proven otherwise, so that we set up the null hypothesis

$$H_0 : p = \frac{1}{2}.$$

The alternative hypothesis in this general context is that the coin is *not fair* or that p is different from $\frac{1}{2}$. In other words,

$$H_a : p \neq \frac{1}{2}.$$

Such an alternative is called a **two-sided alternative**, since values on either side of the hypothesized value $\frac{1}{2}$ are allowed. In some cases, we may have more specific information or even hunches about p. In such cases we may set up the alternative to be either

$$H_a : p > \frac{1}{2}$$

or

$$H_a : p < \frac{1}{2}.$$

These two latter type of alternatives are called **one-sided alternatives**, since they allow for alternative values to be only on one side of the null-hypothesis value, $\frac{1}{2}$ in this case.

$\diamond$

Example 8.2: Suppose there are two treatments or drugs (or a treatment and a control) that we wish to compare . Often we need to test if one treatment or drug is better than the

other. In such cases the null hypothesis states

$$H_0: \text{ The two drugs are equally effective}$$

versus the two-sided alternative hypothesis,

$$H_a: \text{ The effects of the two drugs are different}$$

or, versus the one-sided alternative

$$H_a: \text{ Drug 1 is better than drug 2}$$

or

$$H_a: \text{ Drug 1 is worse than drug 2.}$$

◇

Example 8.3: Suppose the average score on midterms over the past few years has been equal to 70 out of a possible 100 points. This year, suppose with a sample of $n = 100$, we obtain $\overline{X} = 73$. Are this year's students any smarter? Assume that the midterm scores follow a Normal distribution with $\sigma = 10$.

Solution: Suppose we use μ to represent this year's *true* (unknown) mean score on the midterm. Remember that $\overline{X} = 73$ is just the mean of one sample and it can vary if we took another sample. So we set up the null hypothesis that the students this year have the same *true* mean as before (indicating status-quo), i.e.,

$$H_0 : \mu = 70$$

versus the alternative hypothesis suggested in the question, viz., that they are smarter,

$$H_a : \mu > 70.$$

◇

To decide whether to accept or to reject the null hypothesis, we ask "If the hypothesis H_0 is true and the true mean μ is still only 70, how likely are we to see a sample mean of 73 or even larger values than 73?" If the observed sample mean of 73 is quite possible, i.e.,

consistent with a true mean of 70, there is no reason for us to change our mind about μ still being 70.

Remark 8.1 While a null hypothesis as we said before, represents status-quo or no change, it is not always clear-cut as to how to set up an alternative hypothesis. If no specific possible alternative values are being indicated by the question (corresponding to a research hypothesis), the default is a two-sided test.

8.2 P-value approach

A **P-value** is the probability of observing a value of the test statistic *at least as contradictory* to the null hypothesis (and favoring the alternative hypothesis) as the observed value, when the null hypothesis is indeed true i.e., the probability of finding the observed value or *more extreme or contradictory* values than the observed one, if the null hypothesis is true.

Thus, the P-value (also called the **observed significance level**) is a measure of credibility of the null hypothesis, given the data. If the P-value of the observed data is very small, we reject H_0. As will be seen in the following examples, if the alternative is one-tailed, the P-value is the tail area beyond the observed value, *in the same direction as the alternative hypothesis*. If the alternative is two-tailed, the P-value is the probability of observing a test statistic value at least as different from the hypothesized value, *on either side*, i.e., twice the observed tail area.

How small should the P-value be before we reject the null hypothesis? There is no simple answer but the smaller the P-value, the more convincing the disagreement between the data and H_0. Typically a P-value of less than 1% or 5% might be convincing enough for most purposes to decide to reject the null hypothesis.

In particular, let's consider testing $H_0 : \mu = \mu_0$ when σ is known for a Normal sample. Recall,

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

has a N(0, 1) distribution under the null hypothesis. The computation of the P-value, i.e., the tail area depends on the alternative.

If the alternative is $\mathrm{H}_a : \mu > \mu_0$, and z is the calculated value, the P-value is the tail area to the right of z, i.e., $\mathrm{P}(Z > z)$. If the alternative is $\mathrm{H}_a : \mu < \mu_0$, and z is the calculated value, the P-value is the tail area to the left of z, i.e., $\mathrm{P}(Z < z)$. Finally, if the alternative is $\mathrm{H}_a : \mu \neq \mu_0$, and z is the calculated value, the P-value is $2 \times \mathrm{P}(Z > |z|)$. See Figure 8.1 for illustration.

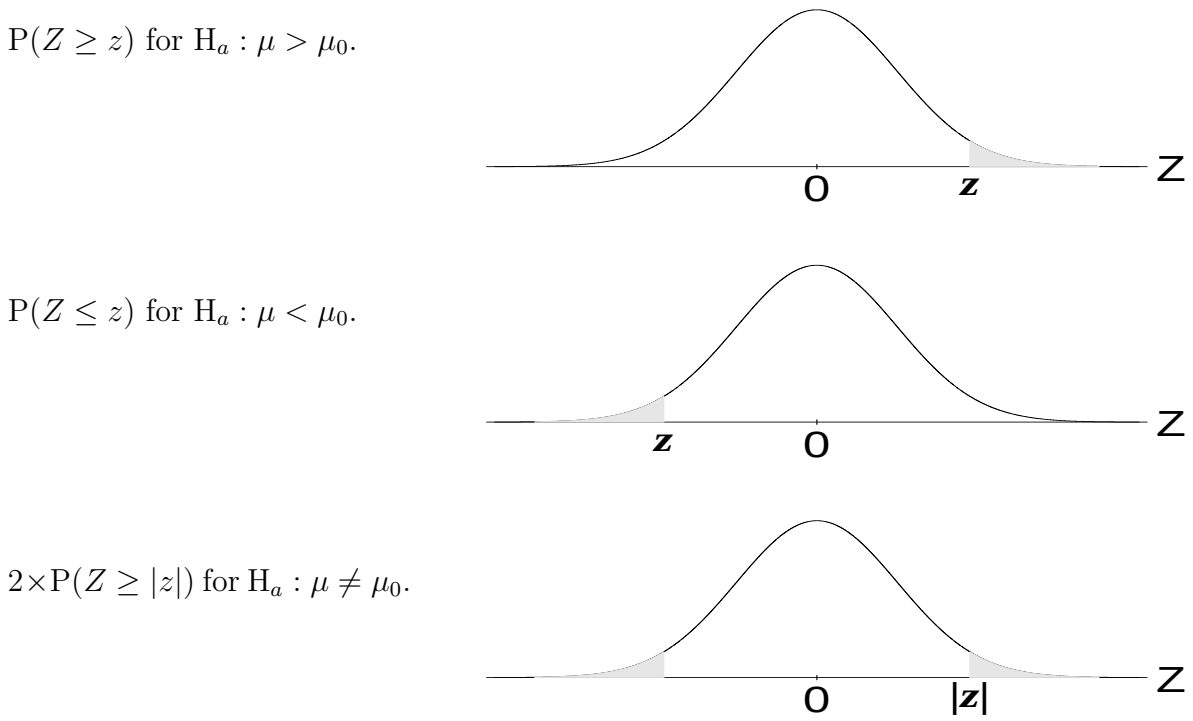


Figure 8.1: P-values for testing $\mathrm{H}_0 : \mu = \mu_0$ against various alternatives.

R Code Instruction:

By implementing **R** in the context of hypothesis testing, we would mostly use it to calculate the critical values and the p-value; Essentially no brand new functions are needed. We now take the example below to demonstrate the implementation of **R**:

Example 8.3 (contd.): If H_0 is true and $\mu = 70$, from Fact 2, the sample mean $\overline{X}$ follows

a normal distribution with

$$\mu_{\overline{X}} = \mu = 70, \qquad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1.$$

The P-value in this example is the probability that the sample mean can be as large as 73 or larger when the null hypothesis is true. Thus

$$P\text{-value } = \mathrm{P}(\overline{X} \geq 73) = \mathrm{P}\left(\frac{\overline{X} - 70}{1} \geq \frac{73 - 70}{1}\right) = \mathrm{P}(Z \geq 3) = 0.0013.$$

R code:

```
# H0: miu=70  vs  Ha: miu>70
n <- 100
sd_x <- 10
miu_xbar <- 70
sd_xbar <- sd_x/sqrt(n)
# The P-value in this example is the probability that the sample mean can be
    as large as 73 or larger when the null hypothesis is true, thus p_val=P(x_
    bar >= 73)=P(Z >= z) with
z = (73-mu_xbar)/sd_xbar
p_val=1-pnorm(z) #where pnorm(abs(z))=P(Z<z)
p_val
```

The output for this **R** code gives::

```
> n <- 100
> sd_x <- 10
> mu_xbar <- 70
> sd_xbar <- sd_x/sqrt(n)
> z <- (73-mu_xbar)/sd_xbar
> p_val=1-pnorm(z) #where pnorm(abs(z))=P(Z<z)
> p_val
[1] 0.001349898
```

Since this is considerably small, for instance less than 0.01, the null hypothesis does not appear credible and we reject H_0. We may thus conclude that the true mean for this year's students is more than 70, that is, they are indeed smarter.

◇

Example 8.3 (contd.): If on the other hand a sample mean of $\overline{X} = 73$ is based not on 100 observations, but only say on $n = 25$, do we still reject H_0?
Solution: In this case, $\mu_{\overline{X}}$ is still equal to 70 under H_0 , but note that now

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = \frac{10}{5} = 2.$$

Thus,

$$P\text{-value } = \mathrm{P}(\overline{X} \geq 73) = \mathrm{P}\left(Z \geq \frac{73-70}{2}\right) = \mathrm{P}(Z \geq 1.50) = 1 - 0.9332 = 0.0668.$$

R code:

```
# if p_val=P(x_bar>=73)=P(Z>=z) and n=25
n <- 25
sd_x <- 10
mu_xbar <- 70
sd_xbar <- sd_x/sqrt(n)
z <- (73-mu_xbar)/sd_xbar
p_val=1-pnorm(z)
p_val
```

The output for this **R** code gives:

```
> n <- 25
> sd_x <- 10
> mu_xbar <- 70
> sd_xbar <- sd_x/sqrt(n)
> z <- (73-mu_xbar)/sd_xbar
> p_val=1-pnorm(z)
> p_val
[1] 0.0668072
```

Since this is not sufficiently small (say smaller than 1% or 5%), we have no reason to reject H_0 that $\mu = 70$. Thus the same difference of 3 extra points in the sample average is convincing enough for us to reject $H_0 : \mu = 70$ if the sample mean is based on $n = 100$, but not if it is based only on a sample of size $n = 25$.

In other words, sample means based on $n = 25$ have more variability and we are not able to rule out chance variability as the cause for those extra 3 points.

◇

Example 8.3 (contd.): Suppose $n = 25$, $\overline{X} = 68$ and we want to test $H_0 : \mu = 70$ versus the other one-sided alternative, $H_a : \mu < 70$.
Solution: As before, given the null hypothesis, $\mu_{\overline{X}} = 70$ and $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = \frac{10}{5} = 2$.

Here, the observed sample mean is 68. To compute the P-value, we need to interpret what we mean by more "extreme" values. Under the alternative that $\mu < 70$, we expect *smaller* $\overline{X}$ values. Values of $\overline{X}$ equal to 68 or smaller are considered more extreme relative to the hypothesized mean of 70 and in favor of the present H_a. Thus, the P-value now is the chance of seeing the observed sample mean value of 68 or values even smaller than this (the left tail). We get

$$P\text{-value } = \mathrm{P}(\overline{X} \leq 68) = \mathrm{P}\left(\frac{\overline{X} - 70}{2} \leq \frac{68 - 70}{2}\right) = \mathrm{P}(Z \leq -1) = 0.1587.$$

R code:

```
# if p_val=P(x_bar<=68)=P(Z<=z) while n=25
n <- 25
sd_x <- 10
mu_xbar <- 70
sd_xbar <- sd_x/sqrt(n)
z <- (68-mu_xbar)/sd_xbar
p_val=pnorm(z)
p_val
```

The output for this **R** code gives:

```
> n <- 25
```

```
> sd_x <- 10
> mu_xbar <- 70
> sd_xbar <- sd_x/sqrt(n)
> z <- (68-mu_xbar)/sd_xbar
> p_val=pnorm(z)
> p_val
[1] 0.1586553
```

This is not too small; for instance, it is not less than 0.01 (or even 0.05), and so we do not reject H_0 .

◇

So far, we were computing P-values for one-sided alternatives like $H_a : \mu > 70$ or $H_a : \mu < 70$.

How does one compute the P-value for a 2-sided alternative, say $H_a : \mu \neq 70$?

Example 8.3 (contd.): Suppose $n = 25$ and we observe $\overline{X} = 68$. We wish to test

$$H_0 : \mu = 70 \text{ vs. } H_a : \mu \neq 70.$$

If the alternative allows values on either side of the hypothesized values, "extreme" values of $\overline{X}$ are those that are either too small or too large, relative to 70. Hence, in terms of the calculation of the P-value for this example, this refers to $\overline{X}$ values at least 2 units smaller than 70 as well as those at least 2 units larger than 70. But by the symmetry of Normal curves,

$$P\text{-value } = \mathrm{P}(\overline{X} \leq 68) + \mathrm{P}(\overline{X} \geq 72) = 2 \times \mathrm{P}(\overline{X} \leq 68) = 2 \times 0.1587 = 0.3174.$$

R code:

```
# if 2-sided test H0: mu=70 vs Ha: mu != 70
# then p_val=P(x_bar<=68)+P(x_bar>=72)
n <- 25
sd_x <- 10
mu_xbar <- 70
sd_xbar <- sd_x/sqrt(n)
z <- (68-mu_xbar)/sd_xbar
```

```
z_1 <- (72-mu_xbar)/sd_xbar
p_val=pnorm(z) + (1-pnorm(z_1))
p_val
# it could also be calculated as 2*P(Z<=z)
p_val_same=2*pnorm(-abs(z))
p_val
```

The output for this **R** code gives:

```
> n <- 25
> sd_x <- 10
> mu_xbar <- 70
> sd_xbar <- sd_x/sqrt(n)
> z <- (68-mu_xbar)/sd_xbar
> z_1 <- (72-mu_xbar)/sd_xbar
> p_val=pnorm(z) + (1-pnorm(z_1))
> p_val
[1] 0.3173105
> # it could also be calculated as 2*P(Z<=z)
> p_val_same=2*pnorm(-abs(z))
> p_val
[1] 0.3173105
```

This P-value is not small and we have no grounds to reject H_0.

◇

As we can see, the calculation of p-value depends on the alternative hypothesis. But in essence, the p-value calculation is just to find the tail-area –either one tail or both tails depending on the alternative being one-sided or two-sided; if the p-value i.e. the area, we calculate is small, say less than some α, then we would reject H0.

Example 8.4: Suppose the mean height for men is known to be 66", and on the basis of a sample of size $n = 36$ women, we get $\overline{X} = 62$". Can we determine if on the average, women are shorter than men? (Use $\sigma = 10$.)

Solution: If μ represents the true mean height for women, we wish to test

$$H_0 : \mu = 66 \text{ versus } H_a : \mu < 66.$$

If H_0 is true, $\overline{X}$ is Normal with

$$\mu_{\overline{X}} = \mu = 66 \text{ and } \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{36}} = \frac{10}{6} = 1.67.$$

Hence,

$$P\text{-value } = \mathrm{P}(\overline{X} \leq 62) = \mathrm{P}\left(\frac{\overline{X} - 66}{1.67} \leq \frac{62 - 66}{1.67}\right) = \mathrm{P}(Z \leq -2.4) = 0.0082.$$

Since this P-value is small (less than 0.05 or 0.01), we reject H_0 and conclude that the women are shorter on the average.

$\diamond$

Example 8.5: Suppose the verbal SAT score for $n = 100$ students gives $\overline{X} = 500$, and it is known that σ =100. Test the hypothesis $H_0 : \mu = 475$ versus the 2-sided alternative, $H_a : \mu \neq 475$.
Solution: Here

$$z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{500 - 475}{100/\sqrt{100}} = \frac{25}{10} = 2.5$$

with

$$\mathrm{P}(Z > 2.5) = 0.0062.$$

Recall that the P-value for a 2-sided alternative takes into account the possibility that extreme values can be on either tail. We therefore double this to get

$$P\text{-value } = 2(.0062) = .0124.$$

Since this is not smaller than 0.01, we do not have a strong enough reason to reject H_0. However, this P-value is smaller than 0.05, and someone using that as a yard-stick would reject H_0. Thus, this P-value may be considered a borderline result in terms of providing conclusive enough evidence in favor of the alternative.

$\diamond$

8.3 Fixed Level of significance

An alternate way to test hypotheses is by using a fixed level of significance, denoted by α. This **level of significance** represents the **rate of false rejection** of H_0 that we are willing to commit, typically a small number like, $\alpha = 0.05$ or 0.01. While we can not altogether avoid this false rejection of the null hypothesis (also called **Type I error**), we can control it by selecting this level α to be reasonably small. There is also another kind of error called the **Type II** error, which corresponds to the **rate of false acceptance** of H_0, i.e., how frequently we say H_0 is true even though it is actually false. All tests we discussed are "optimal" in the sense that they minimize this Type II error for a given Type I error or level of significance. This procedure, which is equivalent to the P-value approach, can be summarized in the following 3 steps:

Step 1 Find the observed value of $z = \left(\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\right)$

Step 2 For given level α, find z_α in cases A and B below and $z_{\alpha/2}$ in case C, using Table A or the last row of Table C.

Step 3

Case A : If we are testing $H_0 : \mu = \mu_0$ vs. $H_a : \mu > \mu_0$, reject H_0 if $z > z_\alpha$.

Case B : If we are testing $H_0 : \mu = \mu_0$ vs. $H_a : \mu < \mu_0$, reject H_0 if $z < -z_\alpha$.

Case C : If we are testing $H_0 : \mu = \mu_0$ vs. $H_a : \mu \neq \mu_0$, reject H_0 if $|z| > z_{\alpha/2}$.

See Figure 8.2 for illustration.

$Z > z_\alpha$ for $H_a : \mu > \mu_0$.

$Z < -z_\alpha$ for $H_a : \mu < \mu_0$.

$|Z| > z_{\alpha/2}$ for $H_a : \mu \neq \mu_0$.

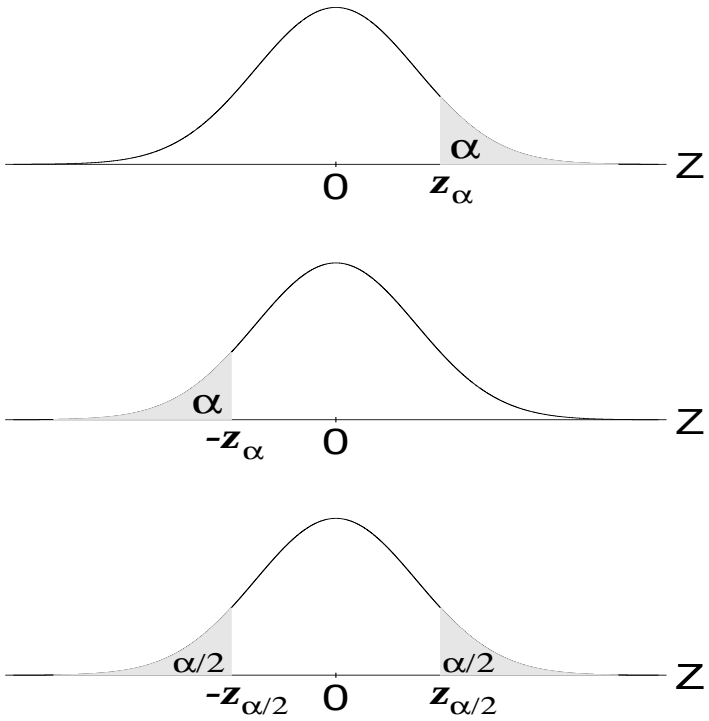


Figure 8.2: Rejection regions for testing $H_0 : \mu = \mu_0$ with fixed significance level α.

Example 8.4 (contd.): If, based on a sample of size $n = 36$ we get a sample mean of $\overline{x} = 64$", test $H_0 : \mu = 66$ versus $H_a : \mu < 66$. Assume $\sigma = 6$" and use $\alpha = 0.05$.
Solution:

Step 1 First we compute

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{64 - 66}{6/\sqrt{36}} = -2.$$

Step 2 From Table C, corresponding to $\alpha = .05$, the tabulated value of z_α is 1.645.

Step 3 Since the alternative is as in Case B, we reject H_0 if $z < -1.645$.

R code:

```
# one-sided test H0: mu=66 vs Ha: mu<66
n <- 36
mu_xbar <- 64
```

```
sd_x <- 6
alpha <- 0.05
sd_xbar <- sd_x/sqrt(n)
z=(mu_xbar-66)/(sd_xbar)
#find critical value Z_c
Z_c <- -qnorm(1-alpha)
print(z)
print(Z_c)   # comparing z with critical value Z_c
```

The output for this **R** code gives:

```
> n <- 36
> mu_xbar <- 64
> sd_x <- 6
> alpha <- 0.05
> sd_xbar <- sd_x/sqrt(n)
> z=(mu_xbar-66)/(sd_xbar)
> #find critical value Z_c
> Z_c <- -qnorm(1-alpha)
> print(z)
[1] -2
> print(Z_c)   # comparing z with critical value Z_c
[1] -1.644854
```

In this case, since $-2 < -1.645$, we do reject H_0 and conclude that the true mean is less than 66.

Very similar to the p-value approach, a fixed level of significance approach can also be used to do the same job. Depending on the alternative hypothesis, H0 would be rejected when the calculated vale of z falls into the shaded rejection region,

◇

Python Code Instruction:

As for **Python**, a new package called *Scipy*, which is as widely used as *Numpy*, will be introduced in the next few sections. While *Scipy* is quite an extensive package, we would just use few of its basic functions.

Example 8.5 (contd.): Recall here $n = 100$, $\overline{x} = 500$, and $\sigma = 100$. Let $\alpha = .01$. Also recall that we already tested $H_0 : \mu = 475$ versus $H_a : \mu \neq 475$ earlier using the P-value approach. To use this alternate approach,

Step 1 Compute

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{500 - 475}{100/\sqrt{100}} = \frac{25}{10} = 2.5.$$

Step 2 Since this is a 2-sided alternative, we find $z_{\alpha/2}$ corresponding to $\frac{\alpha}{2} = \frac{.01}{2} = .005$, which, from Table C, is $= 2.576$.

Step 3 Reject H_0 if $|z| > 2.576$. Since 2.5 is not larger than the critical value, 2.576, we do not reject H_0 — the same conclusion we reached before by using the alternate P-value approach.

Python code:

```
import numpy as np
from scipy.stats import norm
n=100
mu_xbar=500
sd=100
alpha=0.01
sd_xbar=sd/10
z=(mu_xbar-475)/sd_xbar #compute z
Z_c=norm.ppf(1-(alpha/2)) # looking for critical value when 2-sided testing
p_val=2*norm.cdf(-abs(z)) # p value when 2-sided testing
print('z value is ',z,'. critical value is',Z_c,'. p value is ',p_val)
```

The output for this **Python** code gives:

```
z value is  2.5 . critical value is 2.5758 with a p-value of
0.012419
```

Compared with the codes in **R**, the structure of how *scipy.stats* calculates the critical value z and the p-value, is quite similar. We will now take a look at how **R** would approach the testing problems.

R Solution:

```
# two-sided test H0: mu=475 vs Ha: mu != 475
mu_xbar <- 500
n <- 100
sd_x <- 100
sd_xbar <- sd_x/sqrt(n)
alpha <- 0.01
z <- (mu_xbar-475)/sd_xbar
# p-value approach
p_val=2*pnorm(-abs(z))
print(p_val)  #
# Fixed level of significance approach
Z_c <- qnorm(1-alpha/2)
print(Z_c)
```

The output for this **R** code gives:

```
> mu_xbar <- 500
> n <- 100
> sd_x <- 100
> sd_xbar <- sd_x/sqrt(n)
> alpha <- 0.01
> z <- (mu_xbar-475)/sd_xbar
> # p-value approach
> p_val=2*pnorm(-abs(z))
> print(p_val)
[1] 0.01241933
> Z_c <- qnorm(1-alpha/2)
> print(Z_c)
[1] 2.575829
```

◇

8.4 One sample t-test

In Sections 8.2 and 8.3 we illustrated testing ideas for the unknown μ after assuming that the population standard deviation σ is known. In practice, σ also is typically unknown. In

this section, we consider testing $H_0 : \mu = \mu_0$ when σ is not known. Recall that we mentioned in Section 7.4 that when σ is unknown, in the z-statistic

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

we merely replace the unknown σ by its estimate s, resulting in the t-statistic

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

which has a $t(n-1)$ distribution. Apart from this substitution and using a t-distribution in place of a z-distribution, the mechanisms for hypothesis testing remain the same as in last two sections. We now recite the three steps in fixed α testing, for this situation.

Step 1 Calculate

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}.$$

Step 2 Find t_α or $t_{\alpha/2}$ from Table C using $(n-1)$ df.

Step 3

Case A: If the alternative is $H_a : \mu > \mu_0$, reject H_0 if $t > t_\alpha$.

Case B: If the alternative is $H_a : \mu < \mu_0$, reject H_0 if $t < -t_\alpha$.

Case C: If the alternative is $H_a : \mu \neq \mu_0$, reject H_0 if $|t| > t_{\alpha/2}$.

R Code Instruction: When dealing with a t distribution, the entire structure of our **R** code remains the same except for substituting $qnorm()$ by $qt()$

Python Code Instruction: Everything else in **Python** also remains the same except for adding a new built-in function $t.ppf()$ which would display the probability density function of the t distribution using the Syntax:

```
t.ppf (x, df)
```

which provides the probability density of the t distribution with df degrees of freedom, at the value x.

Example 8.6: Suppose a sample of $n = 36$ results in $\overline{x} = 64$" and $s^2 = 25$. Test $H_0 : \mu = 66$ versus $H_a : \mu < 66$ (Use $\alpha = .05$).
Solution:

Step 1

$$\text{Here } t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{64 - 66}{5/\sqrt{36}} = \frac{-12}{5} = -2.4.$$

Step 2 The df here is $(n - 1) = 36 - 1 = 35$ and $t_{.05} \approx 1.690$ from Table C.

Step 3 Since $-2.4 < -1.690$, we reject H_0 and conclude that the true mean is less than 66.

◇

R code:

```
# one-sided test H0: mu=66 vs Ha: mu<66
alpha=0.05
n=36
mu_xbar=64
samp_var=25
t <- (mu_xbar-66)/sqrt(samp_var/n)
t
# Fixed level of significance approach
t_c <- qt(alpha,n-1)
print(t_c)
# p-value approach
p_val <- 1-pt(abs(t),df=n-1)
print(p_val)
```

The output for this **R** code gives:

```
> alpha=0.05
> n=36
> mu_xbar=64
> samp_var=25
> t <- (mu_xbar-66)/sqrt(samp_var/n)
```

```
> t
[1] -2.4
> # Fixed level of significance approach
> t_c <- qt(alpha,n-1)
> print(t_c)
[1] -1.689572
> # p-value approach
> p_val <- 1-pt(abs(t),df=n-1)
> print(p_val)
[1] 0.01092493
```

Python code:

```
import numpy as np
from scipy.stats import t
alpha=0.05
n=36
mu_xbar=64
samp_var=25
test_result=(64-66)/(5/6)
print('the test result is',test_result)
t_val=-t.ppf(1-alpha,n-1)  # critical value of t
print('the critical value of t is',t_val)
```

The output for this **Python** code gives:

```
the test statistic is -2.4 whereas
the critical value of t is -1.6896
```

Example 8.7: Suppose we have a sample of size $n = 16$, from which we obtain $\overline{x} = 10$ and $s = 3.2$. Test $\mathrm{H}_0 : \mu = 8$ vs. $\mathrm{H}_a : \mu > 8$ using $\alpha = .05$.

Solution:

Step 1

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{10 - 8}{3.2/\sqrt{16}} = \frac{8}{3.2} = 2.5.$$

Step 2 Since $n = 16$, df $= (n-1) = 15$. Corresponding to this row and for $\alpha = .05$, we get $t_\alpha = 1.753$ from Table C.

Step 3 Reject H_0 if $t > t_\alpha$.

Since, 2.5>1.753, we do indeed reject $H_0 : \mu = 8$ in favor of the alternative $H_a : \mu > 8$. ◇

Example 8.7 (contd.): For the same problem, test $H_0 : \mu = 8$ vs. $H_a : \mu \neq 8$.

Step 1 It is the same as above.

Step 2 We still have df = 15, but since this is a 2-tailed alternative, we have $\alpha/2 = .025$, so that $t_{\alpha/2} = 2.131$.

Step 3 Reject H_0 if $|t| > t_{\alpha/2}$. Since 2.5>2.131, we do reject H_0 .

Remark 8.2 The P-values are a bit harder to compute for the t-tests since Table C for the t-distribution is not as extensive as Table A for the z-statistic. We can still use this table to find the approximate range for the P-value. For instance, in Example 8.7, with df = 15 and observed (or calculated) value of t = 2.5, we can look up Table C to conclude that the P-value is between 0.01 and 0.02. This follows since the tail probability for 2.249 is 0.02 and the tail probability for 2.602 is 0.01 and our value 2.5 is right in between. Most statistical computer software packages provide the P-values such as for t-distributions.

8.5 Tests on σ

Often, we are interested in testing and controlling the variability in given measurements. For instance, a bottler filling soda cans wants the mean to be 12 ozs. (say) and neither too much more nor too much less. She may wish to test a hypothesis about the true σ, say $H_0 : \sigma = \sigma_0$. Such a test can be conducted as follows:

Step 1 Calculate

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}.$$

Step 2 Find χ^2_α (or, for a two-sided test, two numbers $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$) from Table D for a given df and a given level of significance α.

Step 3

Case A: If the alternative is $H_a : \sigma > \sigma_0$, reject H_0 if $\chi^2 > \chi^2_\alpha$, where χ^2_α is the value with area α above it.

Case B: If the alternative is $H_a : \sigma < \sigma_0$, reject H_0 if $\chi^2 < \chi^2_{1-\alpha}$ where $\chi^2_{1-\alpha}$ is the value with area α below it.

Case C: If the alternative is $H_a : \sigma \neq \sigma_0$, reject H_0 if $\chi^2 < \chi^2_{1-\alpha/2}$ (with area $\frac{\alpha}{2}$ to the left) or if $\chi^2 > \chi^2_{\alpha/2}$ (with area $\frac{\alpha}{2}$ to the right).

R Code Instruction:

qchisq() is the function to calculate probabilities under a chi-square distribution; We use the example below to demonstrate its implementation.

Python Code Instruction:

In *scipy.stats* package, we would use *chi*2 as our tool to solve problems involving the chi-square distribution. $chi2.isf()$ is just similar to *qchisq*() in **R**, and calculates the probabilities for a chi-square distribution.

Example 7.3 (contd.): Suppose on the basis of $n = 16$ observations, we find $s = 3.2$. Test the hypothesis $H_0 : \sigma = 4.2$ versus the alternative $H_a : \sigma < 4.2$, using $\alpha = .05$.
Solution:

Step 1 We compute

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{15(3.2)^2}{(4.2)^2} = 8.71.$$

Step 2 and 3 Here the alternative $H_a : \sigma < 4.2$ is one-sided (see Case B above). From a χ^2 distribution with $(n-1) = 15$ df, we find the lower tail critical value $\chi^2_{1-\alpha} = \chi^2_{0.95} = 7.26$. Since the computed value of χ^2, namely 8.71, is not below 7.26, we have do not have strong enough reason to reject the hypothesis that the actual value of σ is 4.2.

R code:

```
std=3.2
sigma=4.2
n=16
alpha=0.05
chi=(n-1)*std*std/(sigma*sigma) #observed value of chi-square
print(chi)
chi_critical=qchisq(alpha, df=n-1) #critical value of chi-square
print(chi_critical)
```

The output for this **R** code gives:

```
> std=3.2
> sigma=4.2
> n=16
> alpha=0.05
> chi=(n-1)*std*std/(sigma*sigma) #observed chi-square
> print(chi)
[1] 8.707483
> chi_critical=qchisq(alpha, df=n-1) #critical value of chi-square
> print(chi_critical)
[1] 7.260944
```

Python code:

```
from scipy.stats import chi2
sd=3.2
n=16
sigma=4.2
chi=(n-1)*sd*sd/(sigma*sigma)
chi_critical= chi2.isf(q=0.95, df=n-1)#critical value of chi square
print('calculated value of chi-square is', chi)
print('critical value of chi-square is', chi_critical)
```

The output for this **Python** code gives:

```
calculated value of chi-square is 8.70748
while critical value of chi-square is 7.26094
```

◇

8.6 Large sample tests on proportions

In dealing with counts or proportions, we are interested in testing hypotheses about the true proportion p. Suppose in a Binomial experiment with n trials, we observed x successes with resulting sample proportion

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

Based on this, we wish to test

$$\mathrm{H}_0 : p = p_0,$$

where p_0 is a specified value. Recall from Fact $4'$ that if we have a large enough sample,

$$z = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

has an approximate N(0,1) distribution under the hypothesis that the true p is p_0. This fact can be used to test hypotheses on p.

Step 1 Calculate

$$z = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

Step 2 Find z_α from table.

Step 3

If the alternative hypothesis is $\mathrm{H}_a : p > p_0$, reject H_0 if $z > z_\alpha$.

If the alternative hypothesis is $H_a : p < p_0$, reject H_0 if $z < -z_\alpha$.

If the alternative hypothesis is $H_a : p \neq p_0$, reject H_0 if $|z| > z_{\alpha/2}$.

Example 8.8: When a coin is tossed 100 times, suppose we get 60 heads and 40 tails. Test if the coin is fair versus the alternative it is loaded in favor of heads, using significance level $\alpha = 0.05$.

Solution: We want to test

$$H_0 : p = \frac{1}{2} \text{ (fair coin) versus } H_a : p > \frac{1}{2} \text{ (here } p_0 = \frac{1}{2} = 0.5).$$

Since

$$\widehat{p} = \text{ observed proportion } = \frac{60}{100} = 0.6,$$

Step 1

$$z = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{(0.6) - (0.5)}{\sqrt{\frac{(0.5)(0.5)}{100}}} = \frac{(0.1)(10)}{0.5} = 2.$$

Steps 2 and 3 For this one sided alternative, we reject H_0 if $z > z_\alpha$. From Table A, $z_\alpha = 1.645$. Since $z = 2 > 1.645$, we reject the null hypothesis H_0 that the coin is fair.

R code:

```
p_hat=60/100
p0=1/2
n=100
alpha=0.05
std=sqrt(p0*(1-p0)/n)
z=(p_hat-p0)/std #observed z
print(z)
Z_c <- qnorm(1-alpha) # critical z for Ha:p > 1/2
print(Z_c)
```

The output for this **R** code gives:

```
> p_hat=60/100
> p0=1/2
```

```
> n=100
> alpha=0.05
> std=sqrt(p0*(1-p0)/n)
> z=(p_hat-p0)/std #observed z
> print(z)
[1] 2
> Z_c <- qnorm(1-alpha) # critical z for Ha:p > 1/2
> print(Z_c)
[1] 1.644854
```

Python code:

```
import numpy as np
from scipy.stats import norm
p_hat=40/100 #observed p
p0=1/2 # p in H0
n=100
sd=np.sqrt(p0*(1-p0)/n)
alpha=0.05
z=(p_hat-p0)/sd #compute z
Z_c=-norm.ppf(1-(alpha)) # looking for critical value when one-sided testing
print('z value is ',z,'. critical value is',Z_c)
```

The output for this **Python** code gives:

```
[1] z value is  -1.9999. critical value is -1.6448
```

◇

For quick and easy reference, we summarize all these tests as well as those that appear in the next chapter, in the form of a Table in Appendix F.

EXERCISES

8.1 A government agency is planning to send a spaceship to the moon with you in it. Clearly, you are quite concerned about returning safely. The trip will take place only

if a test of the hypothesis **H**: "the ship will return safely", is accepted, instead of alternative **A**: "the ship will not return safely". What do the Type I and Type II errors represent here? Which will be of more concern to you?

8.2 A label on a certain cereal package states that the true mean weight of the packages is 16 ounces (denote this by $\mu = 16$). A consumer group insists that the true mean weight is less than the stated weight. Suppose that the total weight of a sample of 100 boxes is 1550 ounces, and that the standard deviation of the packages is known to be 1.0 ounce.

(a) Test the hypothesis that $\mu = 16$ versus the suggested alternative. Let $\alpha = 0.05$.

(b) Give a 98% confidence interval for μ.

(c) Suppose that μ actually is equal to 16. What is the probability that the sample mean of 100 boxes is between 15.8 and 16.1?

8.3 Suppose the amount of soda in a can is normally distributed with a mean μ of 12 ounces. On the other hand a random sample of 20 cans from a bottling plant gave $\overline{x} =$ 11.2 ounces and $s^2 = 0.4$. Test if there is any validity to a consumer complaint that they are being short-changed. State a null and alternative hypothesis, and test them using $\alpha = 0.05$.

8.4 A company named Acme Semiconductors has developed a new microprocessor. It wants to test how fast one of these new chips can conduct a certain benchmark calculation. Suppose that the time it takes to complete the calculation is normally distributed. After 10 runs, the sample average time to completion is 32.7 seconds, and the sample variance is 16. Letting $\alpha = 0.05$, conduct a two-sided test of the hypothesis that the true average time to completion is 30 seconds.

8.5 For the data in Exercise 7.2, test the hypothesis $\mathbf{H}_0 : \mu = 4$ against the alternative $\mathbf{H}_\text{a} : \mu > 4$ using $\alpha = 0.05$ under both situations of (a) and (b).

8.6 The manufacturer in Exercise 7.5 claims in his advertisements that the mean nicotine content is no more than 16 mgs. Test their claim at $\alpha = 0.10$ level of significance.

8.7 Use the data in Exercise 7.7. The manufacturer of cereal A claims that the percentage of sugar does not exceed 19.8%. Test the hypothesis that they are right. Use $\alpha = .05$.

8.8 Use the data in Exercise 7.8. Do the data contradict the assertion of the physical model? (Test at the level 5%.)

8.9 A manufacturer of automobile tires claims that the average number of trouble-free miles given by one line of tires made by his company is more than 40,000. When sixteen randomly picked tires were tested, the mean number of miles was 39,200, with the sample standard deviation s equal to 8,200 miles. At the 5 percent level of significance, is the manufacturer's claim justified?

8.10 Suppose the weight of a fish caught off of the Santa Monica Pier is normally distributed. From a random sample of 20 recently caught fish, the sample mean is found to be 22.0 ounces and the sample variance is found to be 25.0 ounces. The Fish and Wildlife Department states that the true mean weight is 24.0 ounces. Fishermen claim that due to increased pollution in the bay, the true mean weight is less than 24.0 ounces. Test the claim of the Fish and Wildlife Department at $\alpha = 0.05$.

8.11 The average IQ of entering high school students is known to be 113. A random sample of 49 students who were firstborn children was found to have a mean IQ of 117 with a standard deviation of 15. Is it reasonable to conclude that firstborns have a different IQ from that of the general population of students? Give your conclusions using $\alpha = 0.01$. What is the P-value of the test?

8.12 (a) A soft-drink dispensing machine is supposed to dispense 8 ozs. per cup. The machine needs to be adjusted if the true standard deviation, σ, of the amount per cup is greater than 1.2 ozs. Does the machine need adjustment if the 6 sample cups gave

6.8, 7.8, 8.2, 8.1, 7.6, 8.4

respectively? (Test the appropriate hypothesis using $\alpha = .10$.)

(b) In part (a), set up a 99% confidence interval for σ^2.

8.13 Ten third-graders took an achievement test and scored as follows:

34, 28, 31, 26, 36, 23, 29, 27, 37, 26

(a) Past achievement test scores have been normally distributed with a mean of 25. Is there evidence to suggest that the performance of this year's class is significantly above the past average? Set-up and perform an appropriate test using $\alpha = .10$.

(b) Construct a 99% confidence interval for the true mean score, μ.

(c) Find a 95% confidence interval for the true standard deviation, σ.

8.14 It has been conjectured that the Dow Jones Index for stock prices is equally likely to rise as not rise on a given day relative to the previous day's index value. Over a period of 1000 days, suppose that for 550 of those days the Dow Jones index value rose from the previous day's value. Test this hypothesis at $\alpha = 0.02$.

8.15 If in 80 tosses of a coin, one observes 48 heads, test the hypothesis that the coin is fair. Use $\alpha = .05$.

8.16 A real estate agent in Santa Barbara claims that during summer, 70% of the days are "sunny". In other words, one can see the sun for more than 4 hours without interruption. A client spends 10 days in Santa Barbara and observes only 5 "sunny" days. Test the hypothesis that the agent is telling the truth. (Use $\alpha = .05$.)

8.17 A politician takes a random sample of 50 voters and finds that 28 of these 50 are pro-choice. Use a 5% level of significance to test if the actual proportion of pro-choice voters is significantly more than 0.5. What is the P-value of the test?

8.18 Two girls who share an apartment take turns at washing dishes. Out of a total of 10 broken dishes that occurred over a quarter, 8 were caused by the younger girl. Do you think she is clumsier (or can the event be attributed to chance?) [Use a binomial distribution with n = 10 and $p = \frac{1}{2}$ to find the P-value.]

8.19 It is believed that a certain die is loaded, and specifically that the tendency for a "six" to appear is greater than if the die were fair. Suppose that after rolling the die n = 96 times, "six" appeared 24 times. Test the hypothesis that the die is fair at $\alpha = 0.01$.

8.20 A sample of 900 cars passing through a busy intersection in Santa Barbara was taken on a particular morning. It was found that 13% of the cars observed were SUVs. The purpose of the study was to determine whether the proportion of SUVs in Santa Barbara has increased from the last year's value of 10%.

(a) State H_0 and H_a.

(b) What is the P-value of your test?

(c) What would you conclude at the 5% significance level?

8.21 In a large city, 30% of the households had a particular newspaper delivered to their doors. After the newspaper conducted an aggressive marketing campaign to increase that figure, a random sample of 200 households was surveyed. Of the sample group, 85 households now have the paper delivered.

(a) Can we conclude at 5% level of significance that the campaign was a success?

(b) Find the P-value of the test.

Chapter 9

Comparing two samples

Quite frequently, we wish to compare the means of two groups or populations based on samples from each of them. For example, we may have available a sample of m observations on men's heights $(x_1,\ x_2,\ \ldots,\ x_m)$ and another independently drawn sample of n observations on women's heights $(y_1,\ y_2,\ \ldots,\ y_n)$. It could be of interest to test if the mean heights are the same for these two groups. As another example, $(x_1,\ x_2,\ \ldots,\ x_m)$ may represent observations corresponding to a "treatment" group, whereas $(y_1,\ y_2,\ \ldots,\ y_n)$ represent observations for the "control" group. We would want to test the hypothesis that the treatment made no difference or, that the treatment and control groups have the same mean i.e.,

$$\mathrm{H}_0 : \mu_x = \mu_y$$

versus either of the one-sided alternatives

$$\mathrm{H}_a : \mu_x > \mu_y, \quad \mathrm{H}_a : \mu_x < \mu_y \text{ or } \quad \mathrm{H}_a : \mu_x \neq \mu_y.$$

9.1 Paired t-test

Before we proceed to deal with genuine two-sample questions such as when comparing 2 **independent samples**, we will discuss a special situation where the observations $(X,\ Y)$ actually come as matched pairs and hence are not independent. For instance X could be the score on a test for an individual before (s)he undergoes certain training and Y the score for the same individual after such training. Or, $(X,\ Y)$ could be the heights of twins or siblings.

In such cases, the independence assumption about X and Y is not valid. In Chapter 1, we mentioned briefly the advantages of "blocking" and the paired situation is a simple yet very useful and powerful example of such blocking. Here a block consists of the two twins or siblings or sometimes even the same person being measured before and after. By pairing the data and differencing, we eliminate the so-called block-effect and capture what we are actually trying to measure.

In this case, it is easy to see that the problem of testing the equality of the means can be translated into one which says that the **differences** between X and Y have mean zero. Let μ_x be the true mean for X's (say, before training) and μ_y the true mean for the Y's (say, after training). We wish to test the null hypothesis that training makes no difference, i.e.,

$$\mathrm{H}_0 : \mu_x = \mu_y$$

versus say, the alternative that the mean is bigger after training than before training

$$\mathrm{H}_a : \mu_x < \mu_y.$$

Let d stand for the difference $(X - Y)$. Then it is easy to see that the mean of the difference, μ_d, is equal to the difference in the means, $(\mu_x - \mu_y)$. We can now rephrase H_0 and H_a in terms of μ_d and formulate the hypothesis as :

$$\mathrm{H}_0 : \mu_d = 0 \text{ versus } \mathrm{H}_a : \mu_d < 0.$$

If the data consists of the pairs of observations (x_i, y_i), we can then proceed as follows. First, compute the differences:

observation	before	after	difference
1	x_1	y_1	$d_1 = (x_1 - y_1)$
2	x_2	y_2	$d_2 = (x_2 - y_2)$
	$\vdots$		
n	x_n	y_n	$d_n = (x_n - y_n)$

Using these differences $(d_1,\ d_2,\ \ldots,\ d_n)$, we can now do a one-sample t-test as described

in Section 7.4. We must find

$$\overline{d} = \frac{1}{n}\sum_{1}^{n} d_i$$

and

$$s_d^2 = \frac{\sum_1^n (d_i - \overline{d})^2}{n-1}.$$

We now can compute the usual one-sample t-statistic,

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

with x replaced by d and hypothesized mean μ_0 here being zero (or it can be any other specified value for the mean difference that might be postulated) , so that we have:

$$t = \frac{\overline{d} - 0}{s_d/\sqrt{n}} = \frac{\overline{d}}{s_d/\sqrt{n}}$$

which has the $t(n-1)$ distribution. Against the one-sided alternative we have stated, we reject H_0 if $t < -t_\alpha$. This is called the **paired t-test** or **matched pairs t-test**.

Example 9.1: The amount of lactic acid in the blood was examined for 10 men, before and after a strenuous exercise, with the following results:

Before:	15	16	13	13	17	20	13	16	14	18
After:	33	20	30	35	40	37	18	26	21	19

(a) Test if exercising changes the level of lactic acid in blood. Use $\alpha = .005$.

(b) Find a 95% confidence interval for the mean change in the blood lactose level.

Solution:

(a) Let d = (After level - Before level). The hypotheses are $H_0 : \mu_d = 0$ versus $H_a : \mu_d \neq 0$ since we are just checking if there is a change. The 10 values for d are 18, 4, 17, 22, 23, 17, 5, 10, 7 and 1 with $\overline{d}$ =12.4, $s_d^2 = 63.1556$. Thus,

$$t = \frac{\overline{d}}{s_d/\sqrt{n}} = 4.9342.$$

With df $= 10 - 1 = 9$ and $\alpha = .005$, $t_{\alpha/2} = 3.69$. We reject H_0 if $|t| > t_{\alpha/2}$. Since $4.9342 > 3.690$, we do indeed reject H_0 .

(b) By Table C, $t_{\alpha/2} = 2.262$ for 95% confidence level when df $= 9$. Since $s_d^2 = 63.1556$, s_d $= 7.9470$. Hence, the confidence interval for the mean change in lactic acid level μ_d, is

$$12.4 \pm (2.262)\frac{7.9470}{\sqrt{10}} = 12.4 \pm 5.6845 = (6.7155,\ 18.0845)$$

◇

R code:

We are given the number of men is 10, and use alpha = 0.005. We use the *mean* function to find the mean of the differences of x and y.

```
dbar<-mean(diff)
dbar
s_d_sq<-63.1556
t_acid<-dbar/(sqrt(s_d_sq/10))
t_acid
t_stat_acid=qt(.9975,df=9)
round(t_stat_acid, digits=3)
acid_lower<-dbar - qt(0.975,df=9)*sqrt(s_d_sq/10)
acid_upper<-dbar + qt(0.975,df=9)*sqrt(s_d_sq/10)
print(paste(round(acid_lower,digits=4),round(acid_upper,digits=4)))
```

The output for this **R** code gives:

```
> dbar<-mean(diff)
> dbar
[1] 12.4
```

```
> s_d_sq<-63.1556
> t_acid<-dbar/(sqrt(s_d_sq/10))
> t_acid
[1] 4.934189
> t_stat_acid=qt(.9975,df=9)
> round(t_stat_acid, digits=3)
[1] 3.69
> acid_lower<-dbar - qt(0.975,df=9)*sqrt(s_d_sq/10)
> acid_upper<-dbar + qt(0.975,df=9)*sqrt(s_d_sq/10)
> print(paste(round(acid_lower,digits=4),round(acid_upper,digits=4)))
[1] "6.715 18.085"
```

9.2 Comparing the means for two independent samples

We now return to the problem of two independent samples.

1^{st} sample: $x_1,\ x_2,\ \ldots,\ x_m$, with mean μ_x (say).

2^{nd} sample: $y_1,\ y_2,\ \ldots,\ y_n$ say with mean μ_y (say).

How we test the hypothesis $\mathrm{H}_0 : \mu_x = \mu_y$ depends on what we know or are prepared to assume about the other unknowns, namely the standard deviations for the two groups, σ_x and σ_y. This leads to several cases:

Case (1) Assume that the standard deviations for X and Y observations, σ_x and σ_y are known.

Step 1 Calculate

$$z = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}}.$$

Steps 2 and 3 If the alternative hypothesis is $\mathrm{H}_a : \mu_x > \mu_y$, reject H_0 if $z > z_\alpha$.

If the alternative hypothesis is $\mathrm{H}_a : \mu_x < \mu_y$, reject H_0 if $z < -z_\alpha$.

If the alternative hypothesis is $H_a : \mu_x \neq \mu_y$, reject H_0 if $|z| > z_{\alpha/2}$.

Case (2) $\sigma_x,\ \sigma_y$ are unknown but may be assumed equal.

In this case, we need to calculate the sample variance of the x's, s_x^2, and the sample variance of the y's, s_y^2, by the usual formulae:

$$s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2,$$

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2.$$

If it is reasonable to assume that the variances in both these groups are nearly equal (say from past experience) then we can combine these two separate estimates of variances to obtain what is called the **pooled variance**

$$\begin{aligned} s_p^2 &= \frac{1}{m+n-2}\left\{\sum_{i=1}^{m}(x_i - \overline{x})^2 + \sum_{i=1}^{n}(y_i - \overline{y})^2\right\} \\ &= \frac{1}{m+n-2}\left\{(m-1)s_x^2 + (n-1)s_y^2\right\}. \end{aligned}$$

Then, we compute the t-statistic

$$t = \frac{\overline{x} - \overline{y}}{s_p\sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

which follows a $t(m+n-2)$-distribution, i.e., a t-distribution with $(m+n-2)$ degrees of freedom.

Example 9.2: A medication for blood pressure was administered to a group of 13 randomly selected patients with elevated blood pressures while a group of 15 was given a placebo. At the end of three months, the following data was obtained on their Systolic Blood Pressure.

	n	sample mean	s
Control group, x	15	180	50
Treated group, y	13	150	30

Test if the treatment has been effective. (Assume the variances are the same in both the groups and use $\alpha = .01$.)

Solution: Let μ_x denote the true mean blood pressure for the control group and μ_y denote the true mean blood pressure for the treated group. The hypothesis of interest is $H_0 : \mu_x = \mu_y$ versus $H_a : \mu_x > \mu_y$. Here

$$s_p^2 = \frac{1}{m+n-2}\left\{(m-1)s_x^2 + (n-1)s_y^2\right\} = \frac{(15-1)\,50^2 + (13-1)\,30^2}{15+13-2} = 1761.54.$$

Thus,

$$t = \frac{(\overline{x} - \overline{y}) - 0}{s_p\sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{180 - 150}{\sqrt{1761.54}\sqrt{\frac{1}{15} + \frac{1}{13}}} = 1.8863.$$

Corresponding to $15 + 13 - 2 = 26$ degrees of freedom, $t_{0.01} = 2.479$. Since $t = 1.8863$ is not greater than $t_{0.01} = 2.479$, we have no reason to reject H_0. In other words, there is not enough evidence to conclude that the medicine is effective.

◇

R code:

We use the $qt()$ command in **R** to get $t_{0.01}$ and compare this with our observed value of t, which is 1.8863.

```
t_stat <- 1.8863
t_0.01 <- qt(.99, df=26)
t_0.01
```

The output for this **R** code gives the critical value to be:

```
> t_stat <- 1.8863
```

```
> t_0.01 <- qt(.99, df=26)
> t_0.01
[1] 2.47863
```

Case (3) σ_x, σ_y are unknown and possibly unequal.

Step 1 In this case, we compute the statistic

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}}},$$

whose distribution can be **approximated** by a $t(k)$-distribution where k is given either by the simpler option B or the more complicated option A. Option A provides a reasonably close approximation to the true distribution, when the m and n are 10 or more each. The simpler option B provides what is called a conservative procedure, meaning *less than* the value calculated from the $t(k)$-distribution, i.e., the true Type I error is often less than the assumed or nominal level of significance.

Option A: The degrees of freedom is obtained from the two sample variances given above by using the formula

$$k = \frac{\left(\frac{s_x^2}{m} + \frac{s_y^2}{n}\right)^2}{\sqrt{\left(\frac{s_x^2}{m}\right)^2 \frac{1}{m-1} + \left(\frac{s_y^2}{n}\right)^2 \frac{1}{n-1}}}.$$

If the resulting value of k is not a integer, use the integer part of it. Thus, if k turns out to be 15.2, use $k = 15$ degrees of freedom in looking up the tables. This procedure provides a more accurate approximation to the degrees of freedom, although it is slightly complicated.

Option B: A simpler option is to take $k = \min(m - 1,\ n - 1)$. For instance, if we are dealing with samples of sizes $m = 10$, $n = 15$, the df $= k = \min(9,\ 14) = 9$.

Step 2 This step is the same for Case 2 and 3 (options A and B) except for the different degrees of freedom.

If the alternative hypothesis is $H_a : \mu_x > \mu_y$, reject H_0 when $t > t_\alpha$.

If the alternative hypothesis is $H_a : \mu_x < \mu_y$, reject H_0 when $t < -t_\alpha$.

If the alternative hypothesis is $H_a : \mu_x \neq \mu_y$, reject H_0 when $|t| > t_{\alpha/2}$.

9.3 Confidence interval for the difference of two means

In Case (1) of last section, when the σ_x and σ_y are known, we may use the z-statistic to obtain a $(1-\alpha) = C$ level confidence interval for $(\mu_x - \mu_y)$ as

$$(\overline{x} - \overline{y}) \pm \ z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}.$$

If $\sigma_x,\ \sigma_y$ are not given and we are prepared to assume they are equal (Case (2) of last section), we use the t-statistic to get

$$(\overline{x} - \overline{y}) \pm \ t_{\alpha/2}\, s_p \sqrt{\frac{1}{n} + \frac{1}{m}}.$$

Here, $t_{\alpha/2}$ has degrees of freedom $(m + n - 2)$. Finally, in Case (3) of the last section, we have the confidence interval

$$(\overline{x} - \overline{y}) \pm \ t_{\alpha/2}\sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}},$$

where we use a t-distribution with k df (see Option A or Option B in Case (3) for the appropriate k) to get $t_{\alpha/2}$.

Example 9.2 (contd.): Construct a 95% confidence interval for the difference in the means of blood pressures for the two groups , i.e., $(\mu_x - \mu_y)$.
Solution: As we have done before in the testing context, we assume that there is a common unknown variance, which we estimate by the pooled variance. Recall $\overline{x} = 180$, $\overline{y} = 150$ and

s_p is $\sqrt{1761.54} = 41.9707$. Corresponding to 26 df and $C = (1 - \alpha) = .95$, $t_{\alpha/2} = 2.056$. Therefore, the 95% confidence interval for $\mu_x - \mu_y$ is

$$(180 - 150) \pm (2.056)(41.9707)\sqrt{\frac{1}{13} + \frac{1}{15}}$$

$$= 30 \pm 32.7 = (-2.7, 62.7).$$

◇

R code:

We construct a 95% confidence interval where our difference in means is 30, standard deviation is 41.97, and sample sizes are 13 and 15. We use $qt()$ to find $t_{0.025}$.

```
sd.pool<-41.97
t_.025<-qt(.975,df=26)
lower<-30 - t_.025*sd.pool*sqrt(1/13+1/15)
upper<-30 + t_.025*sd.pool*sqrt(1/13+1/15)
print(paste(round(lower,digits=4),round(upper,digits=4)))
```

The output for this **R** code gives:

```
> sd.pool<-41.97
> t_.025<-qt(.975,df=26)
> lower<-30 - t_.025*sd.pool*sqrt(1/13+1/15)
> upper<-30 + t_.025*sd.pool*sqrt(1/13+1/15)
> print(paste(round(lower,digits=4),round(upper,digits=4)))
[1] "-2.6912 62.6912"
```

9.4 Comparing two proportions in large samples

Suppose there are x successes in one Binomial experiment with m independent trials and y successes in a second Binomial experiment with n independent trials. Let p_1 and p_2 denote the true probabilities for these two populations. We wish to test $H_0 : p_1 = p_2$.

Step 1 We compute

$$\widehat{p}_1 = \frac{x}{m}$$

$$\widehat{p}_2 = \frac{y}{n}$$

as well as

$$\widehat{p} = \text{ combined or pooled proportion } = \frac{x+y}{m+n}.$$

The test statistic to use in this case is

$$z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})(\frac{1}{m}+\frac{1}{n})}}.$$

This follows a Standard Normal distribution under the null hypothesis when both m and n are sufficiently large (say, at least 25 or 30 each).

Steps 2 and 3

If the alternative hypothesis is $H_a : p_1 > p_2$, reject H_0 when $z > z_\alpha$.

If the alternative hypothesis is $H_a : p_1 < p_2$, reject H_0 when $z < -z_\alpha$.

If the alternative hypothesis is $H_a : p_1 \neq p_2$, reject H_0 when $|z| > z_{\alpha/2}$.

Example 9.3: A sample of 180 college graduates was surveyed, 100 of them men and 80 women, and each was asked if they make more or make less than \$40,000 dollars a year. The following data was obtained

	$\geq$ \$40,000	$<$ \$40,000	Total
Men	60	40	100
Women	30	50	80
	90	90	180

Are men more likely to make more than \$40,000 than women?

Solution:

Let p_1 = true proportion of men making more than \$40,000 and
p_2 = true proportion of women making more than \$40,000.

We want to test

$$\mathrm{H}_0 : p_1 = p_2 \text{ (there is no difference between men and women)}$$

versus

$\mathrm{H}_a : p_1 > p_2$ (a higher proportion of men make more than \$40,000 compared to women).

Here,

$$m = 100,\ \widehat{p}_1 = \frac{60}{100} = 0.6,\ n = 80,\ \widehat{p}_2 = \frac{30}{80} = 0.375.$$

Also,

$$\widehat{p} = \text{ combined/pooled proportion } = \frac{60+30}{100+80} = \frac{90}{180} = 0.5.$$

Thus, the test statistic is:

$$z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}\,(1-\widehat{p})\left(\frac{1}{m}+\frac{1}{n}\right)}} = \frac{0.6 - 0.375}{\sqrt{(0.5)(0.5)\left(\frac{1}{100}+\frac{1}{80}\right)}} = 3.$$

We reject H_0 since $z > z_\alpha$ which is 1.645 corresponding to $\alpha = 0.05$.

◇

R code:
For a z-statistic, we can again use $qt()$ to determine $z_{0.05}$ by setting $df = Inf$ since a t-distribution with $df = Inf$ corresponds to a z distribution.

```
z<-3
z_0.05<-qt(.95,df=Inf)
z_0.05
```

The output for this **R** code gives::

```
> z<-3
> z_0.05<-qt(.95,df=Inf)
> z_0.05
[1] 1.644854
```

As mentioned in Chapter 8, we summarize all the tests that appear in this and the last chapter in the form of a Table in Appendix F.

EXERCISES

9.1 A new drug treatment for patients with elevated cholesterol levels is administered to six patients for a four-month test period. Results of the experiment are given in the following table:

Patient		1	2	3	4	5	6
Cholesterol count	Before	252	311	280	293	312	327
	After	211	251	241	248	256	268

(a) Treating the subjects as a random sample of patients with a high cholesterol count, test whether the drug decreases the mean cholesterol level at least by 50 units. Use 5% level of significance.

(b) What is the P-value of this test?
(*Hint*: Draw the picture and think!)

9.2 A psychologist suspects that the first-born twin of a pair of twins is smarter, on average. She measures the IQs of 8 pairs of twins and obtains the following data:

twin	1	2	3	4	5	6	7	8
IQ of first-born	103	98	107	127	116	134	152	110
IQ of second-born	101	96	102	125	117	131	140	112

Does this confirm the psychologist's theory? Do a one-sided test using $\alpha = 0.05$.

9.3 In a "watch-your-weight" program, 9 participants were randomly selected and interviewed. Their weights as they entered the program and their weights after 10 weeks into the program were recorded.

Entering weight	208	215	170	140	182	177	150	191	190
Weight after 10 weeks	200	200	172	143	165	159	150	175	172

(a) Using the information on "weight losses" for the above 9 people, construct a 90% confidence interval for the "true mean" weight loss. What assumptions are you making?

(b) Is the claim that the program enables one to lose 10 lbs. on average substantiated? Test using level $\alpha = .05$.

9.4 A placement exam in mathematics was given to 10 students who had new-math and to 20 students who had a traditional math training. The mean score of the modern math students was 88 points and that of the traditional math students was 82 points. Suppose it may be assumed that the variances of the score for modern math and traditional math are known to be $\sigma_1^2 = 20$ and $\sigma_2^2 = 12$ respectively. At the 5 percent level of significance, do the true mean scores differ significantly? Assume that the scores in the populations are normally distributed.

9.5 An psychologist wishes to test the hypothesis that "blondes have more fun" based on the "Subjective Fun Scale" (higher scores indicating more fun). 13 brunettes are selected at random and their SFS measured. He then bleaches the hair of all 13 subjects blonde, waits two weeks and administers the SFS again. The scores are given in the following table.

Person	1	2	3	4	5	6	7	8	9	10	11	12	13
Brunettes	70	47	57	61	65	57	58	58	62	56	52	60	61
Blondes	50	56	63	73	51	65	60	63	64	66	61	83	71

(a) Use $\alpha = 0.05$ to test whether blondes have more fun.

(b) If in the above experiment, the investigator selects 13 brunettes at random and independently another 13 blondes at random and aministers SFS to them, resulting in the same table. Use $\alpha = 0.05$ to test if blondes have more fun.

(c) If your conclusions are different in (a) and (b), how do you explain it?

9.6 To compare two programs for training industrial workers to perform a skilled job, 20 workers are included in an experiment. Of these 10 are selected at random to be trained by method 1; the remainder to be trained by method 2. After completion of training, all the workers are subjected to a time-and-motion test that records the speed pf performance of a skilled job. The following data are obtained:

Method 1	15	20	11	23	16	21	18	16	27	24
Method 2	23	31	13	19	23	17	28	26	25	28

(a) Test the hypothesis $\mathbf{H}_0$: mean job time after training with method 1 is the same as after training with method 2 against alternative **A**: the first mean time is less than the second. (Assume normality and use level $\alpha = .05$).

(b) At what level **H** is to be accepted?

(c) Suppose that the data above are obtained from paired samples (e.g., each column of the table represents observations of two workers of the same age, mental and physical abilities, etc.). Answer now the questions (a) and (b).

9.7 In one discussion section of a statistics course, there were 15 students and for the midterm exam, their mean was 70 points and the sample variance equal to 36. In another section, there were 20 students whose mean was 78 with a sample variance of 64. Is there reason to believe the second section is significantly better than the first? (Use $\alpha = 0.01$).

9.8 The manufacturer of a new type of baseball claims that their product has the same playing characteristics as the one currently in use. To verify the manufacturer's claim, 20 baseballs of the currently used type and 15 of the new type are placed in an automatic ball-hitting machine, and the distance the machine hits each baseball is recorded.

The average distance for the baseball in current use is 352 ft with a standard deviation of 20 ft, while that for the new ball is 375 ft with a standard deviation of 25 ft. Do you think that the manufacturer's claim is valid? (Use $\alpha = 0.10$).

9.9 In a study of cereal beetle damage on oats, researchers measured the number of beetle larvae per stem in small plots of oats after randomly applying one of the two treatments: no pesticide (the control), or malathion at the rate of 0.25 lb/acre. The data appear nearly normal. Here are the summary statistics:

Group	Treatment	n	$\overline{x}$	s
1	Control	12	3.5	1.2
2	Malathion	15	1.4	0.5

(a) State $\mathbf{H}_0$ and $\mathbf{H}_\mathrm{a}$.

(b) Test $\mathbf{H}_0$ versus $\mathbf{H}_\mathrm{a}$ assuming that the 2 population variances are equal. (Use level = .05).

(c) Test $\mathbf{H}_0$ versus $\mathbf{H}_\mathrm{a}$, this time without assuming that the variances are equal. Use level = .05 and the "conservative" degrees of freedom.

9.10 Electrical measurements on the nerve activity of 6 rats poisoned with DDT gave the following observations:

12.2, 16.8, 25.0, 22.4, 8.4, 20.6

while six other unpoisoned (control) rats gave the measurements

11, 9.7, 12.0, 9.4, 8.2, 6.6

Test the hypothesis that the mean level of electrical activity is the same in both groups, versus the alternative hypothesis that poisoning increases the electrical activity. (Use $\alpha = .05$). Assume that the two samples come from normal distributions with the same variance.

9.11 The percentage fat-content for samples of ice cream of two different brands, say A and B, are given below:

Brand A	7.2	8.5	7.4	3.2	8.9	6.7	9.4
Brand B	9.1	8.5	7.9	5.6	8.4		

(a) Assuming both these are from normal populations with common unknown σ^2, test if these brands have the same percentage fat on the average. (Hint: Test $\mathbf{H}_0$: $\mu_A = \mu_B$ versus $\mathbf{H}_\text{a}$: $\mu_A \neq \mu_B$, where μ_A, μ_B are the true means for the two brands respectively. Use α= .05)

(b) Do the test without the assumption of equal σ.

9.12 The Reproductive Biology Research Foundation conducted an experiment to determine the effect of marijuana on male sexuality (See Newsweek, April 27, 1974). In the experiment 20 young men were picked at random from among those who had smoked marijuana at least 4 days a week for a minimum of 6 weeks, without using any other drugs during that period. A control group of 20 young men who had never smoked marijuana was used for comparison. The measurement of sexuality used was the level of the male sex hormone, testosterone in the blood. The results are tabulated in the chart below.

	Mean level of testosterone	Standard deviation
Marijuana Group	416	150
Control Group	742	130

Test **H** : the two true means are the same. Use $\alpha = .05$.

9.13 In order to determine driving habits, an auto insurance company surveyed 30 male and 25 female drivers. Each was asked how many miles he or she had driven in the past year. The means and the standard deviations are shown in the accompanying table.

	Male	Female
Mean	9,117	10,014
Standard deviation	3,249	3,960

Can we conclude at the 5% significance level that the male and female drivers differ in the number of miles driven per year?

9.14 On the issue of an "Oil Initiative" in Santa Barbara county, an opinion poll was conducted which gave the following results: Among 80 registered Democrats sampled, 65 said they are in favor of the Initiative while among 75 registered Republicans, 50 said they are in favor of it. Test if there is significant difference between Democrats and Republicans, using 5% level of significance.

9.15 The percentages of adults who are right-handed, left-handed, and ambidextrous are well documented. What is not so well known is that a similar phenomenon can be found in animals. Dogs, for example, can be either right-pawed or left-pawed. Suppose that in a random sample of 200 beagles it is found that 55 are left-pawed and that in a random sample of 200 collies 40 are left-pawed. Can we conclude that the true proportion of collies that are left-pawed is significantly different from the true proportion of beagles that are left-pawed? Let $\alpha = 0.05$.

Chapter 10

Bivariate Data: Correlation and Regression

So far, we have dealt with situations where we measured only a single characteristic or variable on each individual — so-called univariate data. Often, it is of interest to measure two or more variables on the same individual. For instance, we may measure (height and weight) or (SAT score and GPA) of each individual; (price of a commodity and demand), (the amount of fertilizer and yield of a crop), etc. If we just measure two variables on each individual, the resulting data is called bivariate data and takes the form

$$(x_1,\ y_1),\ (x_2,\ y_2),\ \ldots\ ,\ (x_n,\ y_n),$$

where $(x_i,\ y_i)$ represent measurements on the "same" i^{th} individual. The primary goals in studying two variables together, are
(i) **correlation** studies: to study the type and amount of association between x and y and/or
(ii) **regression** studies: to predict say y from x by setting up a simple equation,if possible, that relates y to x.

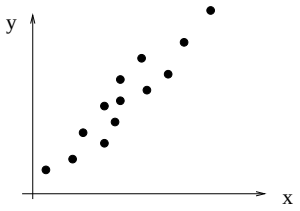
10.1 Correlation

The first step in trying to figure out what if any kind of association there is between x and y is to make a **Scatter-plot**. After looking through the x and y values, find an appropriate

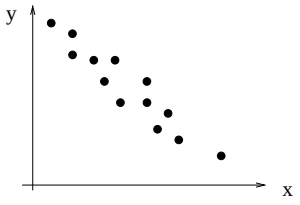
scale for the horizontal and vertical axes. A scatter plot consists of just plotting each of the observations (x_i, y_i) as a point on the x-y plane.

The type of relationship or association between the two variables can be broadly categorized as:

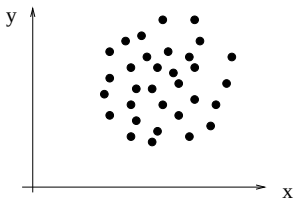
(i)Positive association: Two variables x and y are said to be **positively associated** if the larger x-values tend to be associated with the larger y-values and the smaller x-values tend to be associated with smaller y-values, i.e., x and y tend to be either both large or both small relative to the respective means. For instance, one may expect measurements on height (x) and weight (y) to behave this way because taller people tend also to be heavier.



(ii) Negative association: Two variables are said to be **negatively associated** if larger x-values are generally associated with smaller y-values and vice versa, i.e., the two variables tend to move in opposite directions. For instance, the relationship between the price (x) and demand (y) for a commodity would be of this kind since the higher the price, typically the less the demand for that commodity.



(iii) No association If there is no particular association, the points in a scatter-plot tend to be all over the place, with no discernible patterns.



A descriptive as well as a quantitative measure of such relationship is given by the **correlation coefficient** between x and y, denoted by r_{xy} (often written simply as r). It is obtained as an "average" of the standardized distance of x_i from $\overline{x}$ multiplied with the standardized distance of y_i from $\overline{y}$, i.e.,

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right).$$

Since each of the standardized deviations can be positive or negative, this measure itself can be positive or negative. However, if x and y have a positive association, recall that large x's tend to be observed with large y's . Hence, positive deviations of x, i.e. $(\frac{x_i-\overline{x}}{s_x})$, get multiplied with positive deviations of y, i.e. $(\frac{y_i-\overline{y}}{s_y})$, and negative deviations of x get multiplied with negative deviations of y. In either case, the product of such deviations is positive and when added up, this results in a positive value for r_{xy}. Similarly it can be seen that r takes negative values when there is a negative association between x and y. Thus the sign of the correlation coefficient represents the **type** of association, i.e., positive values of r indicating positive association, while negative r indicating negative association between x and y. Further, it can be shown that the correlation coefficient r takes values between -1 and +1, i.e., $-1 \leq r \leq 1$. Its magnitude indicates the **strength** of such association. For instance, $r = 0.2$ represents a weak positive association while an r value of -0.8 represents a stronger negative relationship between the variables.

Computational formula for r :

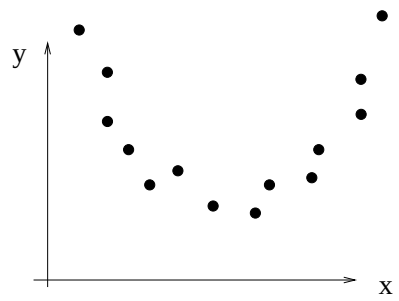
It is somewhat cumbersome to calculate the standardized deviations for the x and y values for each individual observation and multiply them and add them together as required by the definition of r. An easier formula to compute this correlation coefficient is based on computing first

$\sum x_i$, $\sum y_i$, $\sum x_i^2$, $\sum y_i^2$ and $(\sum x_i y_i) = (x_1y_1 + x_2y_2 + \ldots + x_ny_n)$.

Then

$$r = \frac{[n\,(\sum x_i y_i) - (\sum x_i)(\sum y_i)]}{\sqrt{[n(\sum x_i^2) - (\sum x_i)^2]\,[n(\sum y_i^2) - (\sum y_i)^2]}}$$

Remark 10.1 r measures **linear** relationship. If the relation between x and y is of a curvilinear type, the correlation is an inadequate measure of such association. Even though there is a strong (of a non-linear or quadratic type) relationship between x and y as indicated in the following graph, the correlation coefficient may turn out to be zero.



Remark 10.2 Watch out for **spurious** correlation, i.e. false correlation that may exist because of other so-called **lurking** or hidden variables.

Example 10.1: In a study on home fires in a town it was found that $r_{xy} = 0.85$, where x = amount of damage and y = number of firemen on the scene. In spite of this strong positive correlation, it is inappropriate to conclude that the more firefighters are used, the more damage occurs. In this case, z (= the size of the fire) is driving both x and y, and is called the lurking/hidden variable.

$\diamond$

Remark 10.3 Correlation does not imply **causation**. Even if there is a strong correlation, we can not say which is the cause and which is the effect between x and y, without further controlled studies.

For instance, in the case of smoking and incidence of lung cancer, just observing a high correlation between the two factors does not imply smoking causes lung cancer. As the tobacco companies would have us believe, it could be the other way around!

r^2 is called **coefficient of determination**. It can be shown that r^2 is the ratio of variance in y explained through its dependence on x, to the total variance in y.

10.2 Regression

The second important goal in studying two variables together (i.e. bivariate data) is to "predict" one variable say y, using the other variable x. This may be accomplished by setting up a simple equation like

$$y = \alpha + \beta x.$$

In this straight-line relation between y and x, α represents the "intercept", i.e. where the line meets the vertical or y-axis, and β represents the "slope" , i.e. the change in y for each unit increase in x.

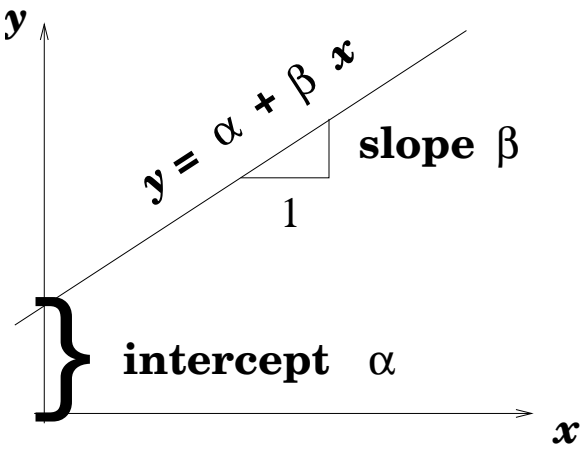


Figure 10.1: Intercept and slope of a regression line

This is called a **simple linear regression** equation which connects y to x. Here y is called the **dependent or response** variable and x the **independent or explanatory** variable.

How do we set up such an equation? In other words, how do we estimate the intercept α and the slope β? We need sample data on such x and y , from which we can base finding the best fitting regression line. This data, (x_i, y_i), is sometimes called the **training data**. We use the so-called **Least squares principle** to estimate these unknown α and β. If we use the equation $y = \alpha + \beta\ x$, then corresponding to the actual value of x_i, our prediction of the value of y would be $(\alpha + \beta\ x_i)$, while the observed value corresponding to this x_i is

actually y_i. Thus, corresponding to this point, we have an error given by

$$e_i = \text{ observed } y_i - \text{ predicted } y_i = y_i - (\alpha + \beta x_i)$$

The "Least Squares Principle" is to find α and β which minimize these sum of squared

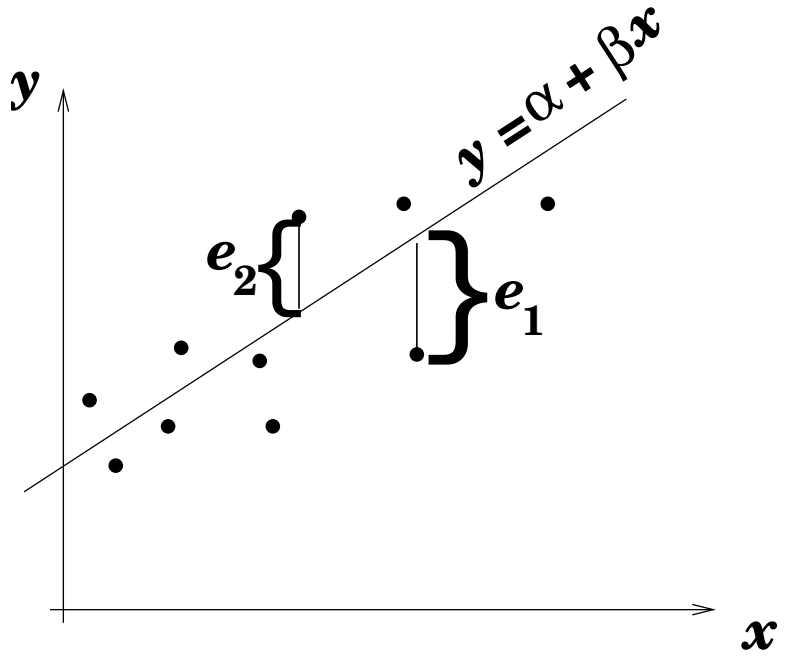


Figure 10.2: Least squares method

errors, namely, $\sum(y_i - \alpha - \beta x_i)^2$. The solution, called the least squares estimate of β, is given by

$$b = \widehat{\beta} = r_{xy}(\frac{s_y}{s_x}) = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}.$$

Observe that this estimate of slope has the same sign as r_{xy} since positive association between x and y gives a positive slope and vice versa. Also, the sample estimate of the intercept is given by

$$a = \widehat{\alpha} = \overline{y} - b\overline{x}\,.$$

.

Thus the best fitting or the least-squares regression line of y on x is given by

$$\widehat{y} = a + bx$$

where

$$b = \widehat{\beta} = r_{xy} \cdot \frac{s_y}{s_x}.$$

and

$$a = \overline{y} - b\overline{x}\,.$$

10.3 Inference for regression

Before we start using a regression equation for prediction, we should ask if doing so has any value. In other words, does y depend on x significantly or is what we have found just random noise? Notice that in the latter situation the regression line becomes horizontal, indicating y does not change with x. However, the sample estimate b may never be exactly equal zero even if the true slope $\beta = 0$. Thus we want to test the null hypothesis

H_0: $\beta = 0$ (i.e. y does not depend on x).

For testing this, we need to calculate the residuals

$$e_i = (\text{observed } y_i - \text{ predicted } y_i) = (y_i - \widehat{y}_i) = y_i - (a + bx_i).$$

and the so called **Residual Sum of Squares, RSS**

$\sum(\text{residual})^2 = \sum {e_i}^2$.

Based on this we calculate an estimate of the standard deviation, namely

$$s_e = \sqrt{\frac{\sum(\text{residual})^2}{n-2}} = \sqrt{\frac{\sum(y_i - \widehat{y}_i)^2}{n-2}} = s_y\sqrt{\frac{n-1}{n-2}(1 - r_{xy}^2)}.$$

Step 1 Calculate

$$\text{standard error of } b = SE_b = \frac{s_e}{\sqrt{\sum(x_i - \overline{x})^2}} = \frac{s_e}{s_x\sqrt{n-1}}$$

and finally the test statistic

$$t = \frac{b}{\text{SE}_b}.$$

This has a t-distribution with $(n-2)$ df under H_0.

Step 2 For a two-sided alternative, H_a: $\beta \neq 0$, reject H_0if $|t| > t_{\alpha/2}$.

(You know by now what to do for the one-sided alternatives).

Prediction:

Here we have to make a distinction between two different scenarios. First, we may be interested in predicting the "mean response" in y corresponding to a given value $x = x^*$. This is done using our regression equation given by $\alpha + \beta x^*$, and it is estimated by

$$\widehat{\mu}_y = a + bx^*.$$

It can be shown that it has standard error

$$SE_{\widehat{\mu}_y} = s_e\sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum(x_i - \overline{x})^2}}$$

To obtain a confidence interval for this predicted mean, we use

$$\widehat{\mu}_y \pm t_{\alpha/2}\left(SE_{\widehat{\mu}}\right)$$

where $t_{\alpha/2}$ is the $\text{C} = (1-\alpha)$ level value for the t distribution with $(n-2)$ df from Table C.

As opposed to this mean response, sometimes we may wish to predict an "individual" response of y corresponding to x= x^*. Since the difference between the mean and an individual value y is the error, which is best estimated by zero, we can still use the equation we

already have set up. The best prediction of y is also given by

$$\widehat{y} = a + bx^*.$$

One can also obtain a C-level confidence interval for this predicted value, called the **prediction interval**, which is given by

$$\widehat{y} \pm \ t_{\alpha/2} \left(SE_{\widehat{y}} \right)$$

where

$$SE_{\widehat{y}} = s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum (x_i - \overline{x})^2}}.$$

Again, $t_{\alpha/2}$ is the $(1 - \alpha)$ =C level value corresponding to a t distribution with $(n - 2)$ df given in Table C. Observe the extra term "1+" under the square root in the formula for SE, which accounts for the additional variability in an individual response as opposed to that of the mean response.

Remark 10.4 Notice from the formula for the standard error, $\text{SE}_{\widehat{y}}$, that the error in the predicted value increases as x^* is farther away from $\overline{x}$. This implies that extrapolating for values x^* far away from the sample mean is not desirable.

Example 10.2: An economist studying the relationship between income and savings (measured in thousands of dollars) collected the following data on 10 households. We illustrate through this example, many of the concepts discussed in this chapter starting with a scatter plot.

Family	Income, x	Savings, y
1	25	0.5
2	28	0.3
3	35	0.8
4	39	1.6
5	44	1.8
6	48	3.1
7	52	4.3
8	55	4.4
9	65	5.6
10	72	7.2

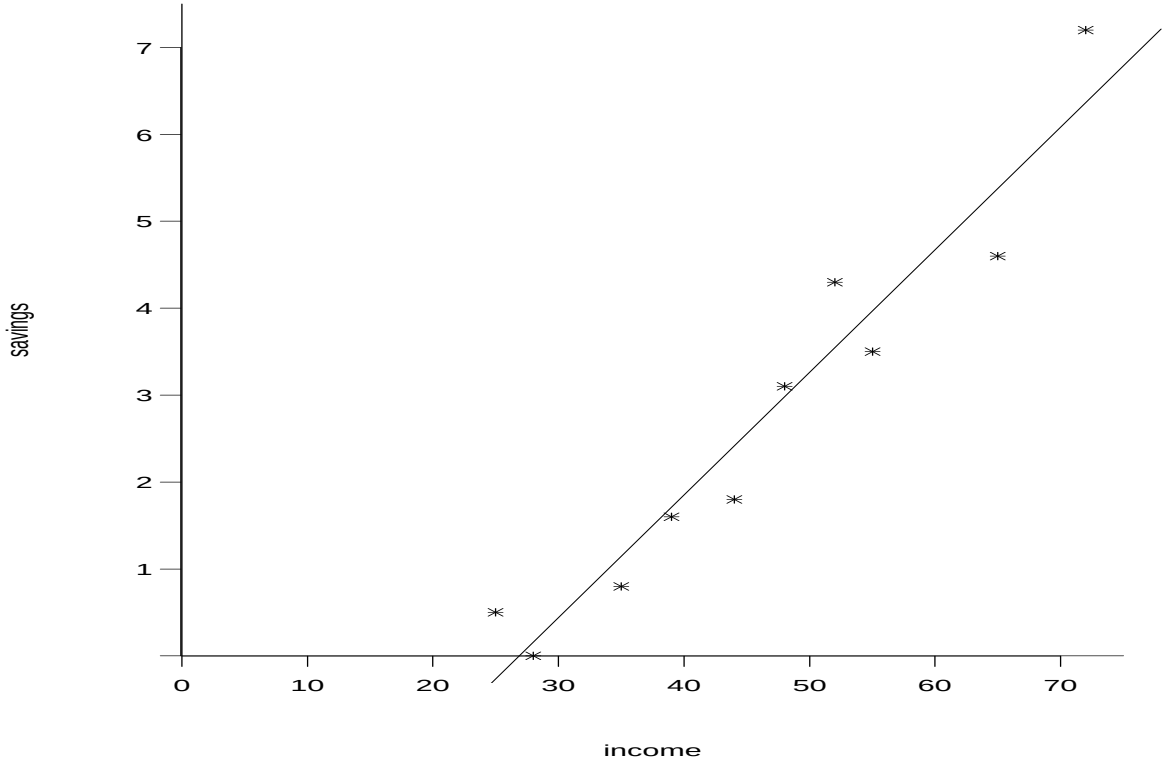


Figure 10.3: Scatter plot of savings vs. income and the best fitting line

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
25	0.5	625	0.25	12.5
28	0.3	784	0.09	8.4
35	0.8	1225	0.64	28.0
39	1.6	1521	2.56	62.4
44	1.8	1936	3.24	79.2
48	3.1	2304	9.61	148.8
52	4.3	2704	18.49	223.6
55	4.4	3025	19.36	242.0
65	5.6	4225	31.36	364.0
72	7.2	5184	51.84	518.4
463	29.6	23533	137.44	1687.3
$\sum x_i$	$\sum y_i$	$\sum x_i^2$	$\sum y_i^2$	$\sum x_i y_i$

$$r = \frac{[10(1687.3) - (463)(29.6)]}{\sqrt{[10(23533) - 463^2][10(137.44) - 29.6^2]}} = .9804$$

$$b = \frac{[10(1687.3) - (463)(29.6)]}{10(23533) - 463^2} = 0.1511$$

$$a = \overline{y} - b\overline{x} = 2.96 - .1511 \times 46.3 = -4.0359.$$

y_i	$\widehat{y}_i$	$e_i = y_i - \widehat{y}_i$	e_i^2
0.5	-0.2594	0.7594	0.5767
0.3	0.1940	0.1060	0.0112
0.8	1.2520	-0.4520	0.2043
1.6	1.8566	-0.2566	0.0659
1.8	2.6124	-0.8124	0.6599
3.1	3.2170	-0.1170	0.0137
4.3	3.8215	0.4785	0.2289
4.4	4.2750	0.1250	0.0156
5.6	5.7865	-0.1865	0.0348
7.2	6.8445	0.3555	0.1264
		Total	1.9375

$$s_e = \sqrt{\frac{1.9375}{10-2}} = .4921$$

$$SE_b = \frac{s_e}{\sqrt{\sum(x_i-\overline{x})^2}} = \frac{s_e}{\sqrt{(\sum x_i^2) - \frac{(\sum x_i)^2}{n}}} = \frac{.4921}{\sqrt{23533 - \frac{463^2}{10}}} = .0107$$

$$t = \frac{b}{SE_b} = \frac{.1511}{.0107} = 14.1215$$

From Table C, P-value $= 2\mathrm{P}(T > |t|) = 2\mathrm{P}(T > 14.1215) = 0$. Hence, reject H_0: $\beta = 0$. A 99% confidence interval for β is given by:

$$b \pm t_{\alpha/2} SE_b = .1511 \pm 3.355 \times .0107 = (0.1152, 0.1870).$$

A 99% confidence interval for μ_y when $x = 50$ is given by:

$$(a + bx^*) \pm t_{\alpha/2} s_e \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum(x_i - \overline{x})^2}}$$

$$= -4.0381 + .1511 * 50 \pm 3.355 \times .4921 \sqrt{\frac{1}{10} + \frac{(50-46.3)^2}{23533 - \frac{463^2}{10}}}$$

$$= 3.5169 \pm .5389$$

$$= (2.9780, 4.0558)$$

Finally a 99% prediction interval for y when $x = 50$ is provided by:

$$(a + bx^*) \pm t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum (x_i - \overline{x})^2}}$$

$$= -4.0381 + .1511 * 50 \pm 3.355 \times .4921 \sqrt{1 + \frac{1}{10} + \frac{(50 - 46.3)^2}{23533 - \frac{463^2}{10}}}$$

$$= 3.5169 \pm 1.7367$$

$$= (1.7802, 5.2536)$$

◇

R code instructions:

We can find the sample correlation coefficient r using the function *cor*, other regression estimates using *summary*(), and plot a scatter plot and best fitting line. As seen in the output for the **R**-code, this gives an r value of 0.9804, slope estimate b of 0.15115, intercept a of -4.03812, residual standard error s_e of 0.4921, standard error SE_b of 0.01075, and $t-$value of 14.062.

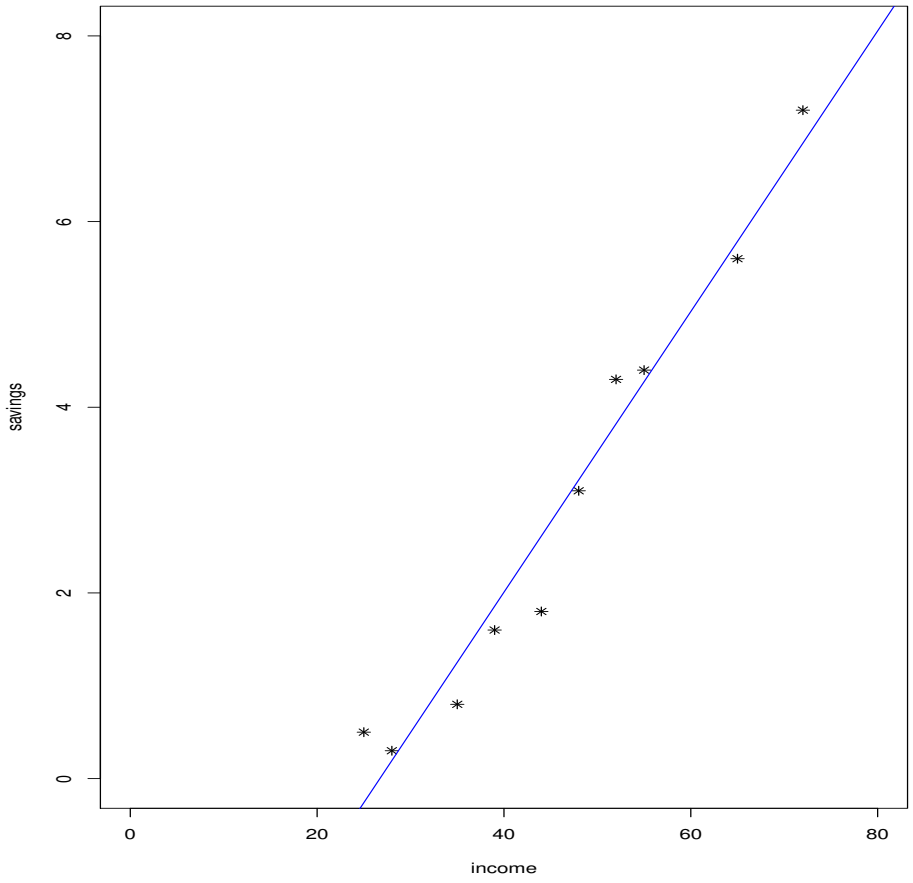
R code:

```
x_i<-c(25,28,35,39,44,48,52,55,65,72)
y_i<-c(0.5,0.3,0.8,1.6,1.8,3.1,4.3,4.4,5.6,7.2)
r<-cor(x_i,y_i)
round(r,digits=4)
plot(y_i~x_i,xlab="income",ylab="savings",
     main="Figure 10.3",
     xlim=c(0,80),ylim=c(0,8),pch=8)
abline(s,col="blue")
s<-lm(y_i~x_i)
summary(s)
# 99% Confidence Interval for Beta
t_stat=qt(p=.995,df=8)
round(t_stat,digits=3)
print(paste(round(0.15115-t_stat*0.01075,digits=4),
            round(0.15115+t_stat*SE_b,digits=4)))
```

The output for this **R** code gives:

```
> x_i<-c(25,28,35,39,44,48,52,55,65,72)
> y_i<-c(0.5,0.3,0.8,1.6,1.8,3.1,4.3,4.4,5.6,7.2)
> r<-cor(x_i,y_i)
> round(r,digits=4)
[1] 0.9804
> plot(y_i~x_i,xlab="income",ylab="savings",
     main="Figure 10.3",
     xlim=c(0,80),ylim=c(0,8),pch=8)
> abline(s,col="blue")
```

Figure 10.3




```
> s<-lm(y_i~x_i)
```

```
> summary(s)
Call:
lm(formula = y_i ~ x_i)

Residuals:
     Min       1Q   Median       3Q      Max
-0.81236 -0.23908 -0.00548  0.29789  0.75944

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.03812    0.52144  -7.744 5.51e-05 ***
x_i          0.15115    0.01075  14.062 6.35e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4921 on 8 degrees of freedom
Multiple R-squared:  0.9611,Adjusted R-squared:  0.9563
F-statistic: 197.7 on 1 and 8 DF,  p-value: 6.352e-07

> # 99% Confidence Interval for Beta
> t_stat=qt(p=.995,df=8)
> round(t_stat,digits=3)
[1] 3.355
> print(paste(round(0.15115-t_stat*0.01075,digits=4),
            round(0.15115+t_stat*0.01075,digits=4)))
[1] "0.1151 0.1872"
```

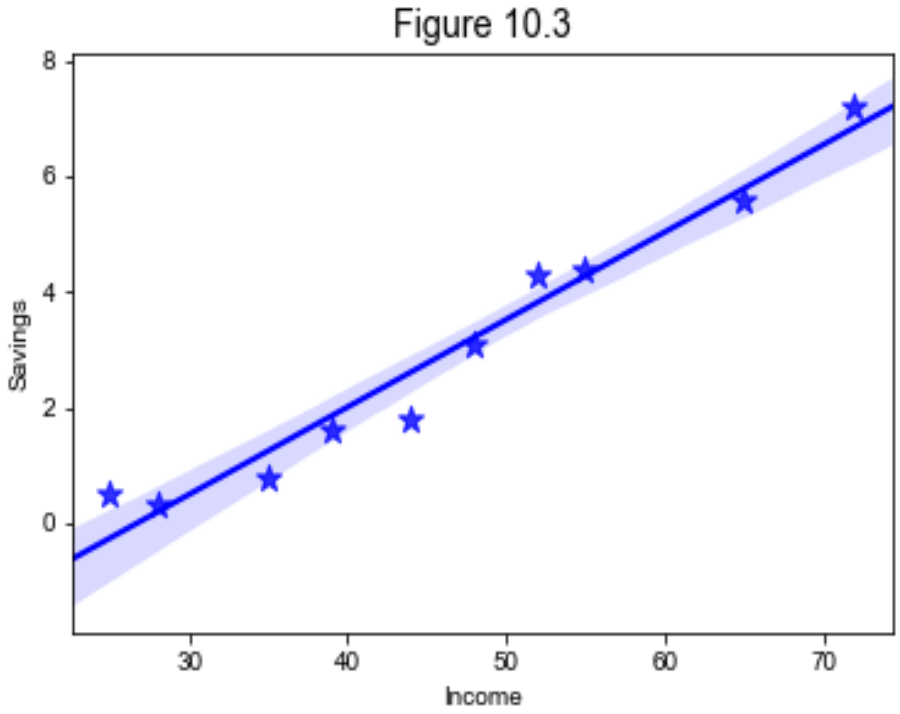
Python instruction:
We use the function *sns.regplot*() to plot our scatter plot and the best fitting line. We begin by importing the necessary libraries to run our **Python** code.

Python code:

```
# Import libraries
import pandas as pd
import matplotlib.pyplot as plt
```

```
import seaborn as sns
# Code for scatter plot and best fitting line
x=pd.Series([25,28,35,39,44,48,52,55,65,72], name='Income')
y=pd.Series([0.5,0.3,0.8,1.6,1.8,3.1,4.3,4.4,5.6,7.2], name='Savings')
sns.regplot(x=x, y=y, color="blue", marker="*", scatter_kws={"s": 110})
sns.set(font_scale=1.2)
plt.title("Figure 10.3")
```

The output for this **Python** code gives:



Residual Analysis

Even the best-fitting linear regression does not completely account for the variability in the y values. It still leaves the so-called "residuals" or the unexplained part, namely

$$e_i = (\text{observed } y_i - \text{ predicted } y_i) = y_i - (a + bx_i).$$

The sum of squares of these residuals, appropriately called the residual sum of squares RSS, represents the left-over variation in y that the dependence on x, has failed to account for. Or one may consider the residual sum of squares as that part of variation in the data not explained by the linear regression model. We may then ask what proportion of the total variation in y values has the linear regression accounted for? Recall that the total variation in the y's is represented by $\sum(y_i - \overline{y})^2$ and is called the $Total\ SS$. Thus the answer to our question is given by

$$(Total\ SS - RSS)/Total\ SS = 1 - (RSS/Total\ SS).$$

This quantity is called the **Coefficient of Determination** and is a measure of how well the linear regression model fits the data. It can be shown that in this case of simple linear regression, this coefficient of determination is exactly r^2, the square of the correlation coefficient that we discussed earlier. The larger it is, the better the fit of the linear regression model.

Since the residuals reflect what is "left over" in the observed y, after the assumed linear dependence on x is accounted for, one might ask if these are purely vestiges of noise, centered at zero and with constant variance or if there are any significant and noticeable patterns still left in them. Various plots of residuals are possible including (i) a time-ordered plot of the residuals r_i versus i, if the data is collected over time, to see if one can detect dependence on time, (ii) a plot of the residuals r_i versus the x_i, to see if there are other detectable patterns which may indicate that a linear regression does not adequately describe the dependence of y on x. This plot can also reveal if the variation in r_i changes with the magnitude of x_i, invalidating the assumption of constant variance that is usually made in the linear regression model. (iii)Also of concern is the assumption of normality of the residuals, which can be checked either by using some rather complex statistical tests or graphically by making what is known as a "normal probability plot". A normal probability plot is a scatter plot of the "ordered residuals" against a set of expected values of an ordered sample of the same size from a standard normal distribution. Such expected values are found from statistical theory and are available in many computer packages and tables. For instance, if $n = 10$, such values are

-1.539, -1.001, -0.656, -0.376, -0.123, 0.123, 0.376, 0.656, 1.001 and 1.539.

Here -1.539 has the interpretation that it is the mean value (long-run average) of the smallest value if a sample of size 10 were to be drawn from a N(0,1), -1.001 is the mean value of the second smallest out of 10 observations from N(0,1), etc. When a normal probability plot is approximately a straight line, then we have reason to believe that the data, in this case the residuals, are from a normal curve.

Example 10.2 (contd.): Examine the residuals in Example 10.2, to see if the normality assumption is valid.

Solution: We order the 10 residuals obtained earlier and plot these against the 10 ordered scores given above and obtain the plot given below. A visual inspection reveals that this plot is approximately linear and therefore we can conclude that normality is a reasonable assumption.

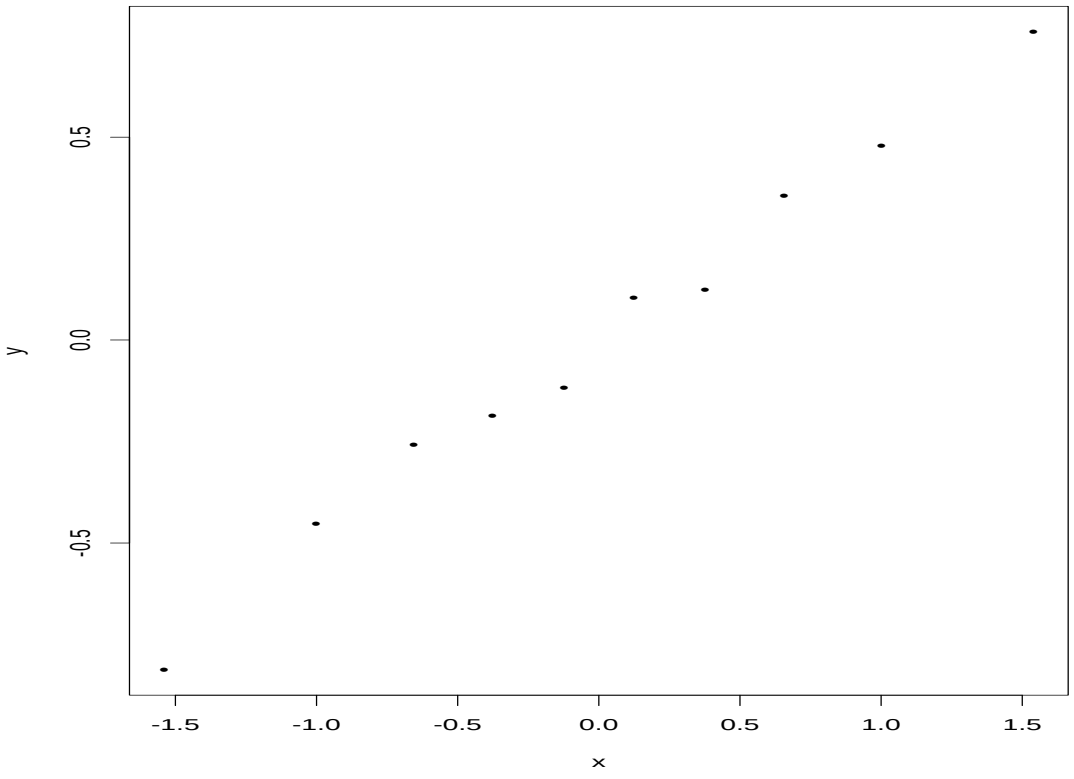


Figure 10.4: Residual plot to check normality

R code:

The function *qqplot*() is used to check normality.

```
x1<-c(-1.539,-1.001,-0.656,-0.376,-0.123,
    0.123,0.376,0.656,1.001,1.539)
y_i<-c(0.5,0.3,0.8,1.6,1.8,3.1,4.3,4.4,5.6,7.2)
y_ihat<-c(-0.2594,0.1940,1.2520,1.8566,2.6124,
    3.2170,3.8215,4.2750,5.7865,6.8445)
ei<-y_i-y_ihat
y.lm = lm(ei ~ x1)
y.res = resid(y.lm)
qqplot(x1,y.res,ylim=c(-1,1),xlab="x",ylab="Residuals",
    xlim=c(-1.7,1.7),pch=8,cex=0.8,main="Figure 10.4")
```

The output for this **R** code is presented in Figure 10.4.

Python code:

The function *sm.qqplot*() is used to check for normality. We begin by importing the necessary libraries to run our **Python** code.

```
# Import libraries
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
import numpy as np
# Code to check normality
y_i=[0.5,0.3,0.8,1.6,1.8,3.1,4.3,4.4,5.6,7.2]
y_ihat=[-.2594,.194,1.252,1.8566,2.6124,3.2170,3.8215,4.275,5.7865,6.8445]
x_1=[-1.539,-1.001,-0.656,-0.376,-.123,.123,.376,.656,1.001,1.539]
e_i=np.subtract(y_i,y_ihat)
model=sm.OLS(e_i,x_1)
results=model.fit()
res=results.resid
fig=sm.qqplot(res,marker='*',markersize=9,color="black")
plt.title("Figure 10.4")
plt.xlabel("x")
plt.ylabel("Residuals")
plt.show()
```

The output for this **Python** code is as given in Figure 10.4.

◇

In many cases, the dependent variable y may depend on more than one independent variable x, as is assumed in simple linear regression. For example. the yield y of a crop may depend on the amount of fertilizer applied, x_1, as well as the amount of rainfall, x_2. Or, one's blood pressure y may depend on one's weight, x_1, one's age, x_2, as well as several other factors. If we believe that the dependence is linear, one can use a **Multiple Linear Regression** equation of the form

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

Again the least squares principle provides a set of simultaneous equations in the unknown parameters from which it is possible to estimate these coefficients and find the best-fitting line

$$y = a + b_1 x_1 + b_2 x_2 + ... + b_k x_k.$$

Confidence statements on the regression coefficients as well as tests of hypotheses are available and the theory and ideas can be viewed essentially as extensions of those that we already discussed for simple linear regression. However for such models, the formulas and the computations can get quite messy and are best handled by using computer packages.

EXERCISES

10.1 The gas mileage of an automobile first increases and then decreases as the speed increases. Suppose the following data on speed x (miles per hour) and mileage y (miles per gallon) is obtained.

Speed	20	30	40	50	60
Mileage	24	28	30	28	24

(a) Make a scatter-plot of mileage versus speed.

(b) Show that the correlation coefficient $r_{xy} = 0$ in this example. Do you really believe there is no association between the 2 variables? Explain the zero correlation in spite of possible association between speed and mileage.

10.2 A car company has been advertising on television. The number of commercials broadcasts in a month (x) and the corresponding number of car sales in the same month (y, in thousands) are given in a table below.

x	15	10	9	11	8	7	12	13
y	33	21	20	25	17	17	26	27

(a) Plot the data.

(b) Find the least squares regression line $\hat{y} = a + bx$. Test at a 5% level whether b is equal to 0.

(c) Predict the number of cars sold in a month when 20 commercials are broadcast during that month and give a 90% prediction interval.

10.3 Suppose it is reasonable to assume that the weight of newborn babies, y in lbs, is a linear function of its age, x in months, during the first 6 months. The following values were obtained for $n = 10$ babies:

$\overline{x} = 3.7$, $\overline{y} = 15.6$, $s_x^2 = 23.1$ and $s_{xy} = \frac{1}{n-1}\sum(x_i - \overline{x})(y_i - \overline{y}) = 47.8$.

(a) Find the best fitting line and interpret what the intercept and the slope denote.

(b) Use this equation to predict the weight of a baby who is (i) two-months old and (ii) 5 months old.

10.4 The following table represents x, the weight in pounds and y, the miles per gallon in the city, for 16 car models:

x	2705	3470	3935	3195	4025	2270	1845	3735
y	25	19	16	21	15	29	31	15
x	1695	2285	3695	2970	2295	3240	2045	2985
y	46	26	17	26	29	19	33	21

(a) Plot a scatter diagram

(b) Calculate the correlation coefficient and interpret it.

(c) Find the best fitting regression line which predicts the mileage y for a given weight x.

(d) Test if the slope of the regression line is zero.

(e) Predict the value of y when x= 2500 and find a 95% confidence interval for this predicted value.

10.5 Eight randomly selected people were asked to watch a 1-hour television program. In the middle of the show, a commercial advertising a breakfast cereal appeared. Each person was shown a commercial of a different length ranging from 20 to 48 seconds, with the essential content being the same in all cases. After the show, each person was given a test to determine how much he or she remembered about the product. The commercial times and test scores (on a 20-point scale) are given below:

Person	1	2	3	4	5	6	7	8
Length of commercial(x)	20	24	28	32	36	40	44	48
Test score(y)	10	8	10	11	14	16	12	13

Summary statistics: $\sum$x= 272, $\sum$y= 94, $\sum xy = 3,320$, $\sum x^2 = 9,920$,
$\sum y^2 = 1,150$.

(a) Do you think a straight line explains the data? Draw a scatter plot to decide.

(b) Calculate the correlation coefficient of the data. What percentage of the total variation in test scores would be explained by a regression line?

(c) Calculate the regression line for predicting test-score using commercial length.

(d) What is the predicted test score of a person who watches a 30 second commercial?

10.6 It has been conjectured that the age at which a child speaks his/her first word can predict the score on an aptitude test given to children entering first grade. Let x = the first month a word is spoken, y = the score on the aptitude test. Data was collected on 8 kids.

x	15	10	9	11	8	7	12	13
y	76	90	93	87	97	100	84	80

Note that $\sum_{i=1}^{8} x_i y_i = 7358$, $\sum_{i=1}^{8} x_i^2 = 953$, $\sum_{i=1}^{8} y_i^2 = 62959$.

(a) Draw a scatter plot (use a scale of 0-20 on x and 70-110 on y).

(b) Compute the correlation coefficient between x and y.

(c) Briefly, interpret the value you obtain for the correlation coefficient.

(d) Obtain the least squares regression line, $y = a + bx$.

(e) Predict the aptitude test score for a child whose first spoken word came at 9 months.

10.7 The following data has been gathered on the used cars for sale:

x = age of the car (in years)	1.0	0.5	5.0	3.0	0.6
y = asking Price (in dollars)	12,900	15,100	5,900	8,950	15,800

x = age of the car (in years)	1.0	4.0	2.0	1	3.5
y = asking Price (in dollars)	11,200	7,750	9,800	12,920	11,200

(a) Draw a scatter plot to see it and what type of association there is between how old the car is (x) and asking price, y.

(b) Obtain the regression line $y = a + bx$. What do a and b signify in this problem?

(c) Predict the asking price for a car which is 4.5 years old and a 95% prediction interval for this value.

(d) Test if the regression line in statistically significant, i.e., β =0. (Use $\alpha = .05$.)

10.8 A study of soil erosion produced the following data on the rate at which water flows across land (x) and the resulting amount of erosion (y).

Flow rate	31	85	126	247	375
Eroded Soil	82	195	218	301	607

(a) Find the least squares regression line, y = a + bx, for predicting the amount soil eroded as a function of flow rate.

(b) Find the predicted amount of soil erosion when the flow rate is x = 100.

(c) Find the correlation coefficient between the two variables and interpret its value.

10.9 In order to estimate the current year's inventory of tires, a tire company sampled 6 of its dealers, in each case obtaining this year's inventory (Y, in 100's) along with last year's (X, in 100's). The summary statistics are as follows:

$$\overline{X} = 5,\ \overline{Y} = 8,\ \sum_{i=1}^{6} x_i^2 = 450,\ \sum_{i=1}^{6} y_i^2 = 684,\ \sum_{i=1}^{6} x_i y_i = 50.$$

(a) Find the best linear regression line Y = a + bX, showing how this year's inventory Y is related to last year's value X. In what sense is this the "best" line?

(b) At a 5% level, test the hypothesis that last year's inventory X is not useful in predicting Y.

(c) Suppose the actual value of an inventory X for last year turned out to be 8 for a dealer. Corresponding to this value of $x_0 = 8$, find a 95% confidence interval for the predicted value of Y, the number of tires in this year's inventory.

10.10 Eight students obtained the following scores (out of 100) for a psychology test (X) and a statistics test (Y).

Student	A	B	C	D	E	F	G	H
Psych. test score (X)	70	38	90	50	96	65	58	77
Stat. test score (Y)	75	56	60	68	56	90	78	52

(a) Compute the sample correlation coefficient, r_{XY}. Interpret its meaning in this context.

(b) Find the best regression line Y = A + BX for predicting the statistics score from the psychology score. In what sense is this the "best" line? Use it to predict the statistics score of someone whose psychology score is X = 60 points.

10.11 Six students of a statistics course recorded the number of hours per week they spent in studying for the course, x, and their score in the final exam, y. The following table gives this data:

hours studied (x):	5	6	8	4	6	7
Final score (y):	70	75	85	60	72	90

(a) Obtain the correlation coefficient and interpret it.

(b) Obtain the least squares regression line, $y = a + bx$.

(c) Predict the score y for a student who studies $x^* = 5.5$ hours and find a 90% confidence interval for this predicted value.

10.12 Data was collected from n = 6 people about the number of years of schooling, x, and their beginning salary in thousands, y.

# of years in school (x):	12	16	16	18	20	21
beginning salary (y):	18	22	24	31	42	40

(a) Compute the sample correlation coefficient, r, and interpret your result in this context.

(b) Find the least squares regression line, $y = a + bx$.

(c) Test if this dependence of y on x is statistically significant (use $\alpha = .05$).

(d) Predict your salary if you decide to stop with a Bachelor's degree, i.e., $x = 16$ years of schooling. Also, find a 90% confidence interval for this predicted value.

10.13 Some recent studies indicate that <u>moderate</u> consumption of wine is beneficial to your heart. Let x be the annual consumption of wine in liters per capita, and y the annual

death rate due to heart disease per population of 100,000. The following table, (*Source:* Laura Shapiro, "Food to Your Health?", *Newsweek*, January 22, 1996), gives data on x and y for 10 countries.

Country	x	y
France	63.5	61.1
Italy	58.0	94.1
Switzerland	46.0	106.4
Australia	15.7	173.0
Britain	12.2	199.9
United States	8.9	176.0
Russia	2.7	373.6
Czech Republic	1.7	283.7
Japan	1.0	34.7
Mexico	0.2	36.4

(a) Compute the correlation coefficient between x and y, and interpret it. Can you argue a cause and effect relation between x and y?

(b) Set up a regression line of y on x and test if $\beta = 0$. Use significance level = 0.05.

(c) Predict the heart disease rate when $x = 10$ liters and find a 95% prediction interval.

10.14 According to the authoritative "Third International Mathematics and Science Study" done in 1995, American 12th graders rank near the bottom in Math and Science literacy, placing 18th among the 21 countries compared. We look at part of the data to see if these test scores Y, can be explained by different factors such as time (hours per day) students spend on hobbies outside school (e.g. such as watching TV, playing sports, etc.) denoted as x_1, the hours spent on a paid job x_2, the hours spent on homework x_3 and finally the per capita expenditure in US dollars that the government spends on elementary and secondary education x_4. The table below summarizes the mean scores of students of different countries on a test measuring knowledge of math and science given in the 1994-95 school year, as well as the factors mentioned above. (*Source:*

Mulis et al. (1998), Mathematics and Science Literacy in the Final Year of Secondary School, Chestnut Hill, MA, Boston College)

Country	mean score y	hobbies x_1	paid job x_2	homework x_3	per capita public expenditure x_4
Netherlands	559	6.5	1.8	1.7	725
Sweden	555	5.3	0.5	1.9	1,163
Iceland	541	5.7	1.8	2.1	1,173
Norway	536	5.9	1.8	1.9	1,292
Switzerland	531	5.3	0.6	2.0	1,383
Denmark	528	5.2	1.5	2.4	1,349
Canada	526	5.4	2.2	2.7	904
New Zealand	525	5.5	1.7	2.2	415
Austria	519	5.6	0.5	2.0	1,058
Australia	525	4.8	1.4	3.3	663
Slovenia	514	4.7	0.5	2.2	300
France	505	4.5	0.6	3.4	847
Czech Republic	476	7.0	1.2	1.4	120
Russia	476	7.6	0.2	3.5	-
Italy	475	5.4	0.6	4.0	557
United States	471	5.9	3.1	1.7	1,040
Lithuania	465	6.6	0.8	3.2	29
Cyprus	447	4.2	0.6	3.2	374
South Africa	352	4.9	0.9	4.8	164

(a) Compute the correlation coefficients between y and x_1, y and x_2, y and x_3, y and x_4, respectively and interpret them. Can you argue a cause and effect relation between y and the x's?

(b) Set up a regression line of y on each of the x's separately and test if $\beta = 0$ in each case. (Use significance level = .05).

(c) Does the time students spend on homework explain the differences in the test scores? Find a 95% confidence interval for the mean score if U.S. students were to get an extra hour of homework per day (i.e., let $x_3 = 2.7$).

(d) How much better can we expect the U.S. students to do if we increased the expenditure on them, say to $x_4 = \$1{,}500$. Find a 95% prediction interval.

10.15 The following data from the Third International Mathematics and Science Study compare the scores of twelfth graders in advanced mathematics (x) and physics (y).

Country	x	y
Sweden	512	573
Switzerland	533	488
Denmark	522	534
Canada	509	485
Austria	436	435
Australia	525	518
Slovenia	475	523
France	557	466
Czech Republic	469	451
Russia	542	545
United States	442	423
Cyprus	518	494
Greece	513	486
Germany	465	522

(a) Make a scatter plot and describe the type of association between math and physics scores.

(b) Compute the correlation coefficient between x and y, and interpret it.

(c) Set up a regression line of y on x and test if $\beta = 0$. Use significance level = 0.05.

Appendix A

Answers to Odd Numbered Problems

Chapter 1

1.1a. An average of 5.5 feet does not mean that all or even most of the family members are taller than 4 feet. Also, even with an average depth of 4 feet, the stream can be very very deep in parts!

b. Statistics deals with drawing conclusions about the population while accounting deals with actual amounts of money.

c. It is possible that many unsatisfied customers simply do not write.

d. Without the vaccine, the number could have been 100,000 instead of 1,000. It is better to compare the proportions dead, among the vaccinated and unvaccinated groups.

1.3 The former summarizes data, while the latter draws conclusions about the population from data.

1.5 Do a street survey, a door-to-door survey or collect information in front of a supermarket. Unlisted numbers are covered by randomly selecting the digits of a telephone number instead of from a telephone directory

1.7 The poll will be biased if people in one group are more likely to call in than people from another group. Also the $1.00 charge excludes all but those who are highly motivated.

Chapter 2

2.1 a. ratio b. ratio c. nominal d. ratio e. ordinal

2.3 a. existence of "absolute zero" and a meaningful ratio of two data points. b. Yes. Money you have. c. No. d. ratio

2.5 Each of the first three will be 5 miles faster than its corresponding true value, while the rest three will be correct.

2.7 a. (-2,5,8,10,12) b. Center: mean = 8; Spread: IQR = 5.

2.9 Median

2.11 mean = 44, $s = 8.8556$

2.13 a. 818.1333 b. 81.6 c. Median, since data is skewed. d. (79.3, 80.3, 81.6,82.6, 86.6)

2.15 a. average=9, s=.9129 b. 9

2.17 a. 12.1 b. 11.5 c. 11 d. 5.685

2.19

15.	7							
16.	7							
17.	2	9						
18.	0	1	3	4	5	7	9	
19.	0	0	0	2	2	3	5	9
20.	3	4	5					
21.	2	6	9					
22.	5	6						
23.	0							
24.								
25.	3							
26.								
27.								
28.	3							

(15.7, 18.4, 19.2, 21.2, 28.3)

2.21 ± 5

Chapter 3

3.1 $\frac{1}{18009460}$

3.3 a. .28 b. .18 c. .42 d. .54

3.5 a. .175 b. .525 c. .075 d. .225

3.7 a. 330 b. .0606

3.9 a. .144 b. .336 c. .92 d. .944

3.11 a. .9410 b. .000531

3.13 a. .2675 b. .5981

3.15 $\frac{1}{7}$

3.17 $\frac{24}{625}$

Chapter 4

4.1 a. .9 b. $\mu = 3$, $\sigma = 1.0954$

4.3 a. .05 b. .9 c. .85, 1.1079

4.5 a. .35, .25 b. 2.3, 1.1874

4.7 a. 1.10 b. 1008.79 c. No, even though the expected winnings is \$1.10, the variance is very large and 998 out of a 1000 end up losers.

4.9 a. .75 b. 1.89, .9262

4.11 a. 15.3 b. 1.9261

4.13 a. $\frac{1}{8}$ b. $\frac{3}{8}, \frac{1}{8}$ c. 4, 2.3094

Chapter 5

5.1 .7854

5.3 a. Binomial with $n = 15$, $p = .30$ b. .0037

5.5 No. $\binom{3}{2}(0.5)^2(0.5)^1 = .375$, but $\binom{6}{4}(0.5)^4(0.5)^2 = .234375$

5.7 a. P(X > 5) =P(Z > -2) = .9772 b. .1359 c. 5.72

5.9 a. .0228 b. .6826 c. $0.1359 \times 0.0228 = 0.0031$

5.11 665.2

5.13 a. .1587 b. .3085 c. .2405

5.15 a. .0228 b. .0918 c. 8.84 years

5.17 a. .9690 b. 0 c. 95.80

5.19 $P(X > 6) = P\left(Z > \frac{6-\mu}{.1}\right) = .02$, $\mu = 6 - 2.055 \times .1 = 5.7945$

5.21 a. .0668 b. .9282 c. 152.8

5.23 a. .6826 b. .0918 c. .0084 d. .0038 e. 65.99

5.25 a. .0164 b. .7373 c. 0 d. 2, 1.6

Chapter 6

6.1 $P(|\overline{X} - \mu| \leq 1) = P(-1.29 \leq Z \leq 1.29) = .8030$

6.3 a. $.0359 \times .0139 = .0005$. No. The latter, since it covers any combination of scores totalling 1300, including (600 and 700). b. .0136

6.5 a. Continuous, X is anywhere between 6.5 and 7.5. b. .5 c. .0003

6.7 P(X > 32) = P(Z > .67) = .2514;

$$P(p > \tfrac{1000}{5000}) = P(p > .2) = P\left(Z > \frac{.2-.2514}{\sqrt{\frac{(.2514)(.7486)}{5000}}}\right) = P(Z > -8.3780) = 1$$

Assume that the heights are normally distributed.

6.9 a. P(X ≥ 3) = P(X=0)-P(X=1)-P(X=2)=.8327 b. μ=4, σ=1.5492 c. P(X > 190) = P(Z > -.91)=.8186

6.11 $P(\widehat{p} > 2.5)$ = P(Z > -1.54)=.9382

6.13 X is Bin(20, .5) ≈ N(10, $\sqrt{5}$) $P(X > a)=P\left(Z > \frac{a-10}{\sqrt{5}}\right)$ = .05. $\frac{a-10}{\sqrt{5}}$ = 1.645, a = 13.68. So a = 14.

6.15 P(X ≥ 75) = P(Z ≥- 1.25) = .8944

Chapter 7

7.1 a. $4.0 \pm 1.645 \times \frac{3.7}{\sqrt{12}}$ = (2.243, 5.757). b. $n = \left(\frac{1.645 \times 3.7}{1}\right)^2$=37.045. Rounded up to the next integer gives n = 38.

7.3 $178 \pm 2.704 \times \frac{8.5}{\sqrt{40}}$ = (174.3659, 181.6341)

7.5 $17 \pm 2.571 \times \frac{3.162}{\sqrt{6}}$ = (13.681, 20.319)

7.7 $20 \pm 1.860 \times \frac{.141421}{\sqrt{9}}$ = (19.9123, 20.0877)

7.9 $5.9392 \pm 2.201 \times \frac{.8971}{\sqrt{12}}$ = (5.3692, 6.5092)

7.11 $\widehat{p} = \frac{12}{50}$ = .24. $.24 \pm 1.960 \times \sqrt{\frac{(.24)(.96)}{50}}$ = (.1216, .3584)

7.13 $\widehat{p} = \frac{22}{200}$ = .11. $.11 \pm 1.960 \times \sqrt{\frac{(.11)(.89)}{200}}$ = (.0666, .1534)

7.15 a. $n = \left(\frac{1.960}{2 \times 0.05}\right)^2$=384.16. Round up to 385.

b. $\widehat{p} = \frac{55}{100}$ = .55, $.55 \pm 1.645 \times \sqrt{\frac{(.55)(.45)}{100}}$ = (.4682, .6318)

c. P(X ≤ 8) = 1 - P(X=9) - P(X=10) = .9893

7.17 a. 752 b. (.7800, .9200)

7.19 a. (.286, .354) b. 846

Chapter 8

8.1 Type I: error of concluding that the space-ship will not return safely when in fact it will. Type II: error of concluding that the ship will return safely when in fact it will not. Type II is of more concern.

8.3 $H_0 : \mu = 12$ vs. $H_a : \mu < 12$; t = -5.657 < $-t_{.05}(19) = -1.729$. Reject H_0.

8.5 a. $z = 1.8934 > 1.645$. Reject H_0. b. t = 1.1669 $\not>$ $t_{.05}(14) = 1.761$. Do not reject H_0.

8.7 $H_0 : \mu = 19.8$ vs. $H_a : \mu > 19.8$; t = 4.2433 > $t_{.05}(8) = 1.860$. Reject H_0.

8.9 $H_0 : \mu \geq 40000$ vs. $H_a : \mu < 40000$; t = -.3902 $\not<$ $t_{.05}(15) = 1.753$. Do not reject H_0.

8.11 a. $H_0 : \mu = 113$ vs. $H_a : \mu \neq 113$; t = 1.86 $\not>$ $t_{.005}(48) \approx 2.704$. Do not reject H_0.

b. P-value is between .05 and .10.

8.13 a. $H_0 : \mu = 25$ vs. $H_a : \mu > 25$; t = 3.18 > $t_{.10}(9) = 1.383$. Reject H_0.

b. $29.7 \pm 3.250 \times \frac{4.6678}{\sqrt{10}} = (24.9027, 34.4973)$

c. $\chi^2_{1-\alpha/2} = 2.700$, $\chi^2_{\alpha/2} = 19.023$, $\left(\sqrt{\frac{9\,(4.6678)^2}{19.023}}\,,\ \sqrt{\frac{9\,(4.6678)^2}{2.700}}\right) = (3.21, 8.52)$

8.15 $H_0 : p = .5$ vs. $H_a : p \neq .5$; $z = 1.7889$. $|z| \not> z_{.025} = 1.960$. Do not reject H_0.

8.17 $H_0 : p = .5$ vs. $H_a : p > .5$; $z = 0.855 \not> z_{.05} = 1.645$. Do not reject H_0. P-value = .1977

8.19 $H_0 : p = \frac{1}{6}$ vs. $H_a : p > \frac{1}{6}$; $z = 2.1909 \not> z_{.01} = 2.326$. Do not reject H_0.

8.21 a. $H_0 : p = .3$ vs. $H_a : p > .3$; $z = 3.8576 > z_{.05} = 1.645$. Reject H_0. b. P-value = 0.

Chapter 9

9.1 a. $H_0 : \mu_d \leq 50$ vs. $H_a : \mu_d > 50$; t= 0 $\not>$ $t_{.05}(5) = 2.015$. Do not reject H_0.

b. 0.5

9.3 a. $9.677 \pm 1.860 \times \frac{9.042}{9} = (4.061, 15.273)$. Weight loss is normally distributed.

b. $H_0 : \mu_x = \mu_y$ vs. $H_a : \mu_x < \mu_y$; t= -.110 $\not<$ $-t_{.05}(8) = -1.860$. Do not reject H_0.

9.5 a. t = 1.55 $\not>$ $t_{.05}(12) = 1.782$. Do not reject H_0. b. t = 1.63 $\not>$ $t_{.05}(24) = 1.711$. Do not reject H_0.

9.7 t = -3.3806 < $-t_{.01}(33) = -2.43$. Reject H_0.

9.9 a. $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_1 > \mu_2$

b. t= 6.162 > $t_{.05}(25) = 1.708$. Reject H_0.

c. t= 5.68 > $t_{.05}(11) = 1.796$. Reject H_0.

9.11 a. $s_p = 1.8148$. t = .538 $\not>$ $t_{.025}(10) = 2.228$. Do not reject H_0.

b. t = -.5783 $\not>$ $t_{.025}(4) = 2.776$. Do not reject H_0.

9.13 $H_0 : \mu_x = \mu_y$ vs. $H_a : \mu_x \neq \mu_y$; t = .9065. |.9065| = .9065 $\not>$ $t_{.025}(24) = 2.064$. Do not

reject H_0.

9.15 $H_0 : p_1 = p_2$ vs. $H_a : p_1 \neq p_2$; $z = 1.7624$. $|1.7624| = 1.7624 \not> z_{.025} = 1.960$. Do not reject H_0.

Chapter 10

10.1 b. $r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$. In the sum, (20,24) cancels with (60, 24) and (30, 28) with (50, 28), while (40, 30) results a zero summand.

10.3 a. $\widehat{y} = 7.9437 + 2.0693x$. The line intercept with the y axis at (0, 7.9437) and a baby will grow 2.0693 lbs. (on average) every month.

b. (i) 12.0823 lbs. (ii) 18.2902 lbs.

10.5 a. Yes, roughly. b. r = .709; r^2 = .503, 50.3 c. $\widehat{y} = 5.63 + .18x$ d. 11.03

10.7 a. negative association b. $\widehat{y} = 14939.0933 - 1753.2839x$

c. 7049.3158, $s_e = \sqrt{178811608} = 1495.0401$, 95% CI = (3064.4696, 11034.1620)

d. $t = \frac{-1753.2839}{310.3641} = -5.6491$.—t—> $t_{.025} = 2.306$. Reject H_0.

10.9 a. $\widehat{y} = 11.1665 - .6333x$. The sum of squares of residuals will be minimized.

b. t = -1.63. $|t| > t_{.025} = 2.776$. Do not reject H_0.

c. 11.1665 -.6333*8 $\pm 2.776 \times 6.7023\sqrt{1 + \frac{1}{6} + \frac{(8-5)^2}{450-\frac{30^2}{6}}}$ = (-14.2530, 26.4532).

10.11 a. r = .9165. Strong positive linear relationship. b. $\widehat{y} = 33.33 + 7.00x$ c. 71.83, (60.6, 83.1)

10.13 a. r = .3948. No. b. $\widehat{y} = 190.7307 - 1.7552x$. t = -1.2153. $|t| \not> t_{.025} = 2.306$. Do not reject H_0.

c. 173.1787; 173.1787 $\pm 2.306 \times 108.0030\sqrt{1 + \frac{1}{10} + \frac{(10-20.99)^2}{9998.01-\frac{209.9^2}{10}}}$ = (-90.5842, 436.9416)

10.15 a. positive association b. r_{xy} = .4833, a moderate linear relationship

c. $\widehat{y} = 217.9872 + 0.5545x$, t = 1.9122. $|t| \not> t_{.025} = 2.179$. Do not reject H_0.

Appendix B

Introduction to R

B.1 What is **R**?

R is one of the most popular and useful programming languages that has free and open resources. It has many statistical computing and graphing tools, allowing one to analyze data sets, large and small. Many built-in functions and packages are already part of **R**, making it very convenient for users. Users can compile and run **R** on various operating systems including Windows, Mac OS X, and Linux.

B.2 Getting started with **R** and **RStudio**

(a) Download and Install **R**

- Go to `https://www.r-project.org/`
- Click on Download CRAN
- Choose the mirror closest to your location e.g.: `http://cran.stat.ucla.edu/`
- Select your operating system
- Select the most recent version of **R** (R-3.5.1.pkg)
- Follow the step-by-step installation to install **R**

(b) Download and Install **RStudio**

- **RStudio** is an open source and enterprise-ready professional software for **R**, which is a more handy version for **R** programming.
- Go to `http://www.rstudio.com`
- Click on RStudio Download

- Choose the first one, RStudio Desktop Open Source License, click on DOWNLOAD
- Choose the Installers for Supported Platforms and download the **RStudio**
- Follow the installer's simple steps to install **R**

(c) Video Link for tutorial to download **R** and **RStudio**

- For Mac: `https://www.youtube.com/watch?v=d-u`$_7 vdag - 0$
  $ForWindows$: `https://www.youtube.com/watch?v=9-RrkJQQYqY`

B.3 • Basics in **R** Programming

(a) Basic Interaction with Interpreter

Data analysis in **R** typically proceeds as an interactive dialogue with the interpreter. To start with, we type in an **R** command at the `>` prompt, press the Enter key, and the interpreter responds by executing the command and returning a result, producing graphical output, or sending output to a file or device.

The **R** language includes the following usual arithmetic operators:

- `+` Addition
- `-` Subtraction
- `*` Multiplication
- `/` Division
- `^` or `**` Exponentiation
- `sqrt` Square Root
- Here are some simple examples of arithmetic in **R**:

```
89-6 #subtraction
5+8  #addition
6*7  #multiplication
76/12 #division
3^3 #exponentiation
sqrt(81) #square root

```

the output for this **R** code gives:

```
> 89-6 #subtraction
[1] 83
```

```
> 5+8  #addition
[1] 13
> 6*7  #multiplication
[1] 42
> 76/12 #division
[1] 6.333333
> 3^3 #exponentiation
[1] 27
> sqrt(81) #square root
[1] 9
```

In **R**, text to the right of the pound sign # is ignored by the interpreter. Therefore, we use pound sign (#) to add a comment to any line of code, just like we have used above.

(b) Data Frame and Variables

Creating a data frame is an important first step in any statistical analysis since we have to manipulate and analyze data for better interpretation. A data frame is composed of two types of variables in **R** programming. The first type of variables are numeric variables, which consist of numbers such as 19, 28, etc. The second type of variables are non-numerical variables, consisting of characters, other symbols, logical vectors, combination of numbers and symbols such as $TRUE$, $toys$, $\_ex\_$, $route101$, etc. A little different from numeric variables, quotation marks are needed to operate character variables in **R** programming; $cfunction$ is used to operate on more complex data structures than individual numbers. We can combine different types of data with the c function and operators (< -) or (=) to assign variables. Also, we directly type the variable name to show the result. Here are some examples:

```
variable_a <- c(1,2,3,4,5,6)
variable_a
variable_b <- c('monkey', 'cats', 'dogs')
variable_b
variable_c <- c('monkey34', 34, 'cats___', 98, 'dogs')
variable_c

```

the output for this **R** code gives:

```
> variable_a <- c(1,2,3,4,5,6)
> variable_a
[1] 1 2 3 4 5 6
> variable_b <- c('monkey', 'cats', 'dogs')
> variable_b
[1] "monkey" "cats"   "dogs"
> variable_c <- c('monkey34', 34, 'cats___', 98, 'dogs')
> variable_c
[1] "monkey34" "34"       "cats___"  "98"       "dogs"
```

B.4 Frequently used functions

As a tool for statistical analysis, there are many built-in functions in **R** programming that help users to find the information more accurately and conveniently.

- `mean()` is used to find the mean of a dataset
- `sd()` is used to find the standard deviation of a dataset
- `var()` is used to find the variance of a dataset
- `summary()` is used to find the five-number summary of a dataset
- `qnorm()` is used to find the z-score for the corresponding p-value in the z-table
- `pnorm()` is used to find the p-value for the corresponding z-score in the z-table
- `pt()` is used to find the p-value with corresponding to a t-score in the t-table
- `qt()` is used to find the t-score with corresponding p-value in the t-table
- `plot()` is used to display all kinds of graphical analysis of a dataset

However, several powerful functions are not contained in **R** by default, but we can install packages to use these functions. To install a package, we only have to use the function *install.packages*("*nameofthepackage*"). For example, we can install the package " *statip* " in **R** by saying

```
1 install.packages('statip')
2
```

The package has been successfully installed if it shows the following line at the end of the output.

```
The downloaded binary packages are in /var/folders/zs/h8y5w6dn6294t8sf2zg
kwb5h0000gn/T//RtmpFLEJ9t/downloaded_packages
```

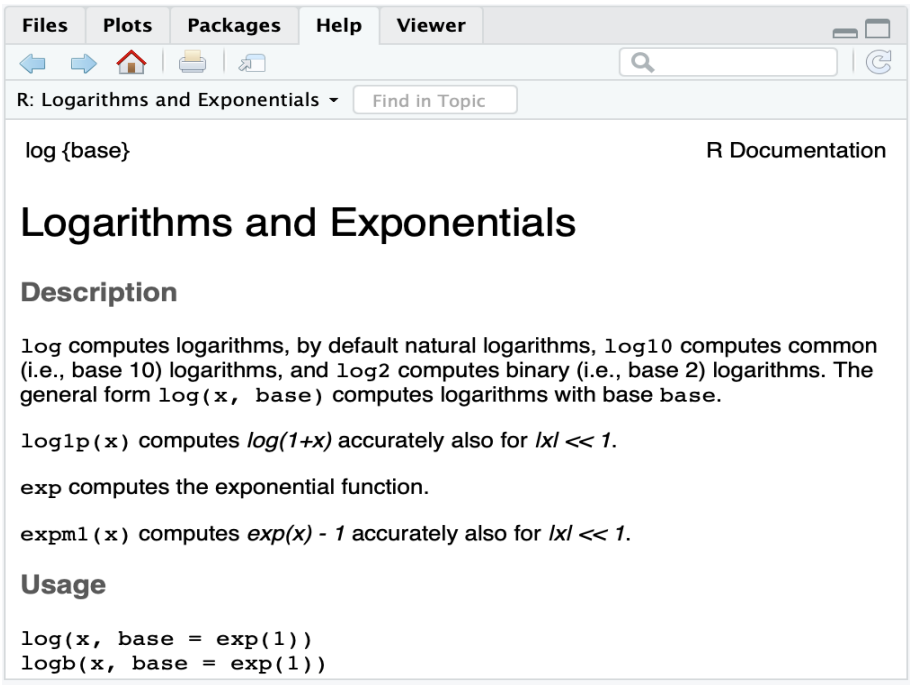
Note that the directory would be different for each user depending on different settings of the computer; but once this line prints out, package installation is complete.

B.5 Getting help with **R** programming

There is a host of other resources in the **R** programming, therefore, we can use `help()` function or ? operator to get information about an **R** function. For example:

```
1 help(log)
```

Here is the result displaying on the lower right window:



? operator offers exactly the same result. Therefore, even if we forgot the usage of a function in **R**, we can use these two functions to search for the information we want.

Appendix C

Introduction to Python

C.1 What is **Python**?

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. It is like a pseudocode in that it allows users to express powerful ideas in a few lines of code while also being very readable. Users can import libraries such as *numpy*, *matplotlib*, *scipy*, *seaborn*, and *pandas* that make **Python** a powerful environment for computing. **Jupyter Notebook** is an open-source web application allowing users to create documents and easily share them with others using email, Dropbox, Github, and the Jupyter Notebook Viewer, and is a good place to start.

C.2 Getting started with **Jupyter Notebook** for **Python**

(a) Download and install **Jupyter Notebook** and **Anaconda** distribution

- Visit the project website at https://www.jupyter.org.
- Click on *Install the Notebook*, and then select *install the Anaconda distribution*. Using the **Anaconda** distribution is recommended for new **Python** users who would like to set up their development environment from scratch.
- Follow the link https://www.anaconda.com/download/ to the **Anaconda** download page and choose between installers for Windows, Mac OS X, and Linux.
- Once installed, use the command *$ jupyter notebook* in your terminal to start **Jupyter Notebook**.

- This will open up the **Jupyter Notebook** application in your browser. To create a new *Python 3* file, click on the button *New* and then *Python 3* located on the right side of your browser (next to Upload).

C.3 Basic data types in **Python**

- Like several other programming languages, **Python** has a number of basic types including integers, floats, booleans, and strings, which behave in ways similar to other programming languages.
- Unlike many languages, **Python** does not have unary increment (x++) or decrement (x−−) operators.

  Numbers: include integers and floats.

  Booleans: **Python** uses English words like *and*, *not*, and *or*, instead of symbols like (&&, ||, *etc.*).

  Strings: **Python** has great support for strings and can be used with single or double quotes.

C.4 Built-in Containers

- **Python** has several built-in containers including lists, loops, and functions.

  List: **Python** equivalent of arrays but are resizable and can contain elements of different types. It stores a series of items in a specific order, which can be accessed using an index or within a loop e.g.

  - nums = [1, 5, 2]
  - print(nums[−1])

  Loops: **Python** allows users to loop over elements of a list e.g.

  - pets = [“cat”, “dog”, “fish”]
  - for pet in pets:
    - print(pet)

  Function: They are named blocks of code to do a specific job.

  - **Argument**: information passed to a function
  - **Parameter**: information received by a function

C.5 ***Numpy*** Library

- ***Numpy*** is the core scientific computing library in **Python**. To get started, type *import numpy as np* in **Jupyter** and run it.
- It can be used to deal with powerful $N-$dimensional arrays and matrices, along with high-level mathematical functions.

  For instance, **numpy.random.binomial(*n, p, size*)**: draws samples from a binomial distribution where n is number of trials, p is probability of success, and *size* is output shape. (see Ch 5)

  Similary **numpy.random.uniform(*low, high, size*)**: draws samples from a uniform distribution where *low* is the lower bound of the interval, *high* is the upper bound, and *size* is output shape. (Ch 5)

  numpy.mean(*a*): calculates the average of the elements of array a.

  numpy.std(*a*): calculates the standard deviation of the elements of array a.

C.6 ***Matplotlib*** Library

- ***Matplotlib*** is a **Python** 2D plotting library which allows one to plot figures and create visualizations such as line graphs and scatterplots. To get started, type *import matplotlib.pyplot as plt* in **Jupyter**.
- Plots can be customized using customizations such as **plt.scatter()**, **plt.title()**, **plt.xlabel()**, **plt.ylabel()**, and **plt.show()** to add titles, labels, and scaling axes.

  plt.savefig() saves a plot whereas **plt.figure()** is used to set a custom figure size.

C.7 ***Scipy*** Library

- ***Scipy*** is a **Python**−based ecosystem that contains a large number of probability distributions and a growing library of statistical functions. Type*from scipy import stats* or *from scipy.stats import norm* in **Jupyter**.

  norm.ppf displays the percent point function or inverse PDF.

  norm.cdf displays the cumulative density function (CDF).

  stats.mode displays the most frequent value in a given data set.

  stats.iqr displays the inter-quartile value in a given data set.

C.8 ***Seaborn*** Library

- ***Seaborn*** is a **Python** data visualization library based on ***matplotlib*** and provides a high-level interface for drawing attractive and informative statistical graphics. Get started by writing *import seaborn as sns* in **Jupyter**.

  sns.regplot() plots data and a linear regression model fit.

  sns.set() sets aesthetic parameters in one step.

Appendix D

Python and R Code for Answering Selected Exercises from the Book

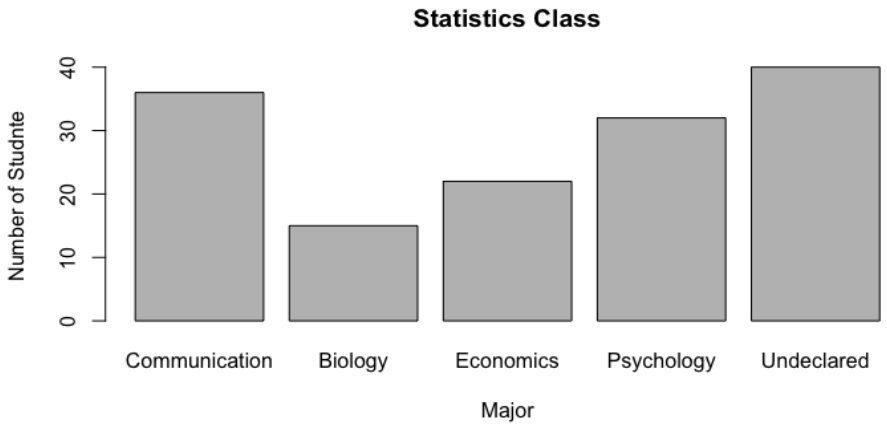
Chapter 2

- Ex. 2.2

  Using **R**:

```
student=c(36,15,22,32,40)
major=c("Communication","Biology","Economics","Psychology","Undeclared")
barplot(student,names.arg = major,xlab="Major",ylab="Number of Student",
    main="Statistics Class")
```

  the output for this **R** code gives:

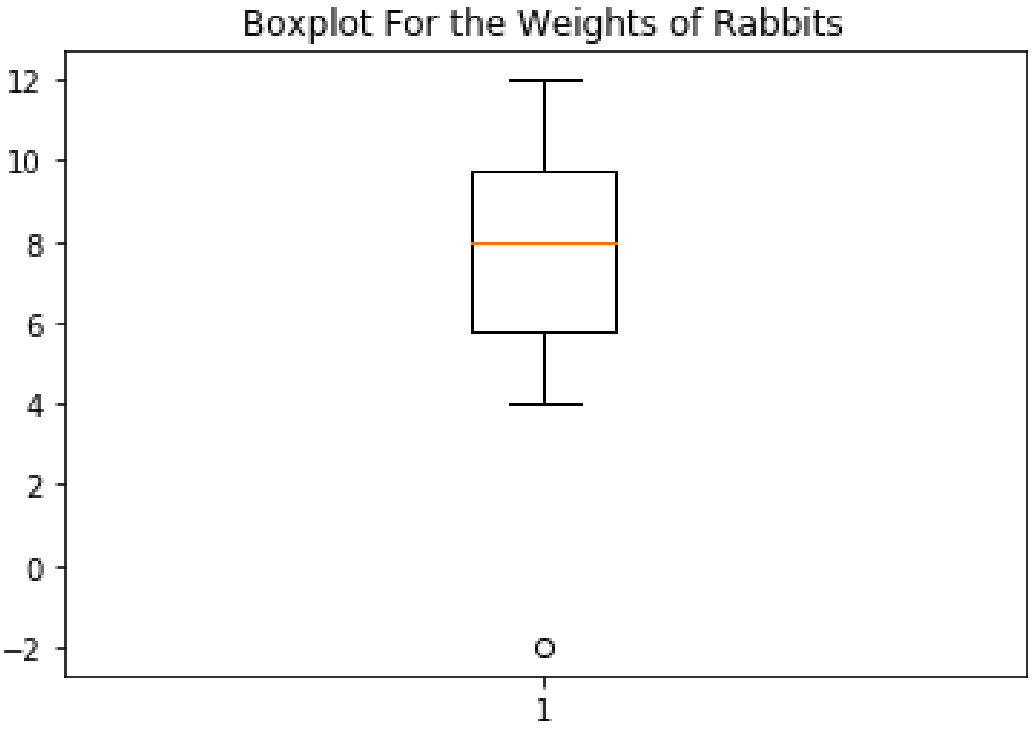


- Ex. 2.7

  Using **Python**:

```
weight=(10,8,12,-2,8,8,9,11,5,4)
import matplotlib.pyplot as plt
plt.boxplot(weight)
plt.title('Boxplot For the Weights of Rabbits')
plt.show()
```

  the output for this **Python** code gives:



- Ex. 2.8

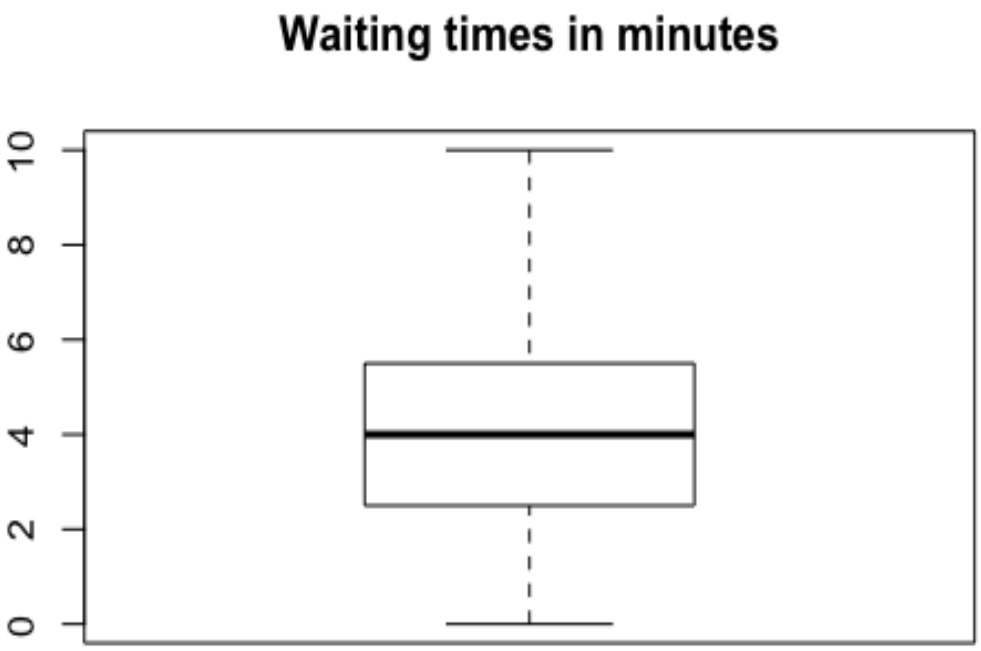
  Using **R**:

  a.)

```
time=c(10,1,2,9,5,3,4,1,0,5,6,5,3,9,3)
summary(time)
boxplot(time,main="Waiting time in minutes")
```

  the output for this **R** code gives:

```
> time=c(10,1,2,9,5,3,4,1,0,5,6,5,3,9,3)
> summary(time)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0     2.5     4.0     4.4     5.5    10.0
```



b.)

```
mean(time)
sd(time)
```

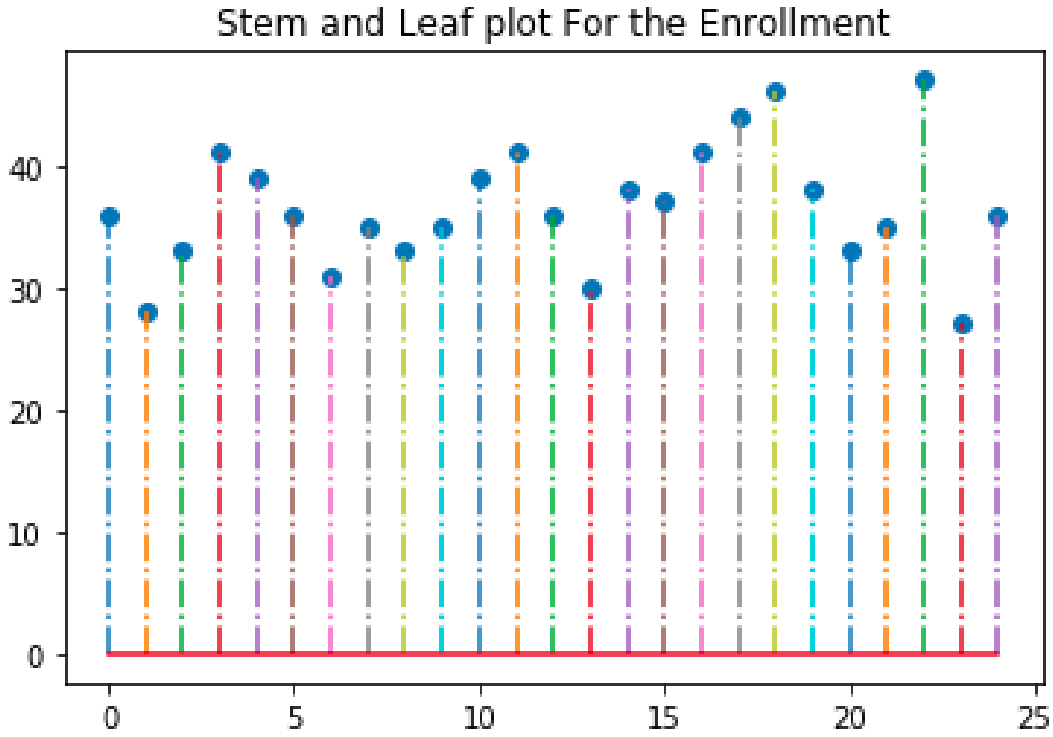
the output for this **R** code gives:

```
> mean(time)
[1] 4.4
> sd(time)
[1] 3.065942
```

- Ex 2.18
  Using **Python**:

```
enrollment=(36,28,33,41,39,36,31,35,33,35,39,41,36,
            30,38,37,41,44,46,38,33,35,47,27,36)
import matplotlib.pyplot as plt
plt.stem(enrollment,linefmt="-.")
plt.title('Stem and Leaf plot For the Enrollment')
plt.show()
```

The output of the **Python** code is:



Chapter 5

- Ex. 5.2

  Using **R**:

```r
# Ex 5.2
#p(x=4)
dbinom(4,size=20, prob=1/4)
#p(x=8 or more)
sum(dbinom(8:20, size=20, prob=1/4))

```

  the output for this **R** code gives:

```
> dbinom(4,size=20, prob=1/4)
[1] 0.1896855
> sum(dbinom(8:20, size=20, prob=1/4))
[1] 0.1018119
```

  Using **Python**:

```python
import numpy as np
n=20
p=1/4
count_4=0
count_8_more=0
x=np.random.binomial(n, p, 10000)
for i in range (10000):
    if x[i]==4:
        count_4=count_4+1
    if x[i]>=8:
        count_8_more=count_8_more+1
print(count_4/10000)
print(count_8_more/10000)

```

the output for this **Python** code gives:

```
0.1862
0.1022
```

- Ex. 5.9 a) b)

  Using **R**:

```r
#prob jumbo
1-pnorm(30, mean=20, sd=5)
pnorm(30, mean=20, sd=5, lower.tail = F) # same as line above
#prob medium
pnorm(30, mean=20, sd=5)-pnorm(25, mean=20, sd=5)

```

The output for this **R** code gives:

```
> 1-pnorm(30, mean=20, sd=5)
[1] 0.02275013
> pnorm(30, mean=20, sd=5, lower.tail = F) # same as line above
[1] 0.02275013
> pnorm(30, mean=20, sd=5)-pnorm(25, mean=20, sd=5)
[1] 0.1359051
```

Using **Python**:

```python
import numpy as np
mu=20
sigma=5
count_jumbo=0
count_medium=0
x=np.random.normal(mu, sigma, 10000)
for i in range (10000):
    if x[i]>30:
        count_jumbo=count_jumbo+1
    if x[i]>=15 and x[i]<=25:
        count_medium=count_medium+1
print(count_jumbo/10000)
```

```
print(count_medium/10000)

```

the output for this **Python** code gives:

```
0.0179
0.6866
```

- Ex. 5.22
  Using **R**:

```
#p(x<1)
pnorm(1.00, mean=0, sd=1)
#p(x<-1.5)
pnorm(1.5, mean=0, sd=1)
#p(x>2.4)
1-pnorm(2.4, mean=0, sd=1)
#p(x=2)
dnorm(2, mean=0, sd=1)
#between 1 and 2
pnorm(2, mean=0, sd=1)-pnorm(1, mean=0, sd=1)
#between -1 and 2.2
pnorm(2.2, mean=0, sd=1)-pnorm(-1, mean=0, sd=1)

```

the output for this **R** code gives:

```
> pnorm(1.00, mean=0, sd=1)
[1] 0.8413447
> #p(x<-1.5)
> pnorm(1.5, mean=0, sd=1)
[1] 0.9331928
> #p(x>2.4)
> 1-pnorm(2.4, mean=0, sd=1)
[1] 0.008197536
> #p(x=2)
> dnorm(2, mean=0, sd=1)
[1] 0.05399097
```

```
> #between 1 and 2
> pnorm(2, mean=0, sd=1)-pnorm(1, mean=0, sd=1)
[1] 0.1359051
> #between -1 and 2.2
> pnorm(2.2, mean=0, sd=1)-pnorm(-1, mean=0, sd=1)
[1] 0.8274413
```

- Ex. 5.23 a) b) d) e)
  Using **R**:

```
#a)
pnorm(71, mean=68, sd=3)-pnorm(65, mean=68, sd=3)
#b)
# 6 feet is 6*12=72 inches
1-pnorm(72, mean=68, sd=3)
#d)
pnorm(60,mean=68, sd=3)
#e)
qnorm(p=0.25, mean=68, sd=3)

```

the output for this **R** code gives:

```
> #a)
> pnorm(71, mean=68, sd=3)-pnorm(65, mean=68, sd=3)
[1] 0.6826895
> #b)
> 1-pnorm(72, mean=68, sd=3)
[1] 0.09121122
> #d)
> pnorm(60,mean=68, sd=3)
[1] 0.003830381
> #e)
> qnorm(p=0.25, mean=68, sd=3)
[1] 65.97653
```

Chapter 6

- Interpreting Fact 4:

  Fact 4 is considered the core principle in the whole chapter. Before we dig into specific problems, we could use **Python** to first test its accurateness:
  Suppose we have a binomial distribution of $Bin(50, 1/6)$; the simulation goes like:

```
import numpy as np
n=50
p=1/6
x=np.random.binomial(n,p,1000000) # simulating random array following
    binomial distribution
sim_mean=np.mean(x)  #mean of the simulated array x
sim_sd=np.std(x)  # sd of simulated array x
theoretical_mean=n*p  # formula in fact 4
theoretical_sd=sqrt(n*p*(1-p)) # formula in fact 4
print("the simulated result mean and standard deviation is", sim_mean,
    sim_sd)
print("the theoretical result mean and standard deviation is",
    theoretical_mean,theoretical_sd)

```

  The output for this **Python** code gives:

```
[1] the simulated result mean and standard deviation is 8.329962 2.63570124987
[2] the theoretical result mean and standard deviation is 8.333333333333332
2.63523138347
```

  As we can see, the simulated values are very close to the result expected from the formula, and the accuracy is very high.

- Interpreting Figure 6.3: Normal approximation to Binomial

  In the Figure 6.3, we increase the sample size n from 10 to 80 as we go from the top to the bottom; and vary the p values 0.05, 0.5 and 0.8 as we go from left to right. These plots demonstrate how the normal approximation gets better with increasing n and for p values of closer to 0.5. The **R** code for plotting these graphs is attached below, for the readers who would like to experiment with other values of n and p.

```r
x1 <- seq(0,10,by=1)
y1 <- dbinom(x1,10,0.05)
plot11 <- plot(x1,y1,col="red",pch=16)
y2 <- dbinom(x1,10,0.5)
plot12 <- plot(x1,y2,col="red",pch=16)
y3 <- dbinom(x1,10,0.8)
plot13 <- plot(x1,y3,col="red",pch=16)

x2 <- seq(0,20,by=1)
y4 <- dbinom(x2,20,0.05)
plot21 <- plot(x2,y4,col="green",pch=16)
y5 <- dbinom(x2,20,0.5)
plot22 <- plot(x2,y5,col="green",pch=16)
y6 <- dbinom(x2,20,0.8)
plot23 <- plot(x2,y6,col="green",pch=16)

x3 <- seq(0,50,by=1)
y7 <- dbinom(x3,50,0.05)
plot31 <- plot(x3,y7,col="blue",pch=16)
y8 <- dbinom(x3,50,0.5)
plot32 <- plot(x3,y8,col="blue",pch=16)
y9 <- dbinom(x3,50,0.8)
plot33 <- plot(x3,y9,col="blue",pch=16)

x4 <- seq(0,80,by=1)
y10 <- dbinom(x4,80,0.05)
plot41 <- plot(x4,y10,col="purple",pch=16)
y11 <- dbinom(x4,80,0.5)
plot42 <- plot(x4,y11,col="purple",pch=16)
y12 <- dbinom(x4,80,0.8)
plot43 <- plot(x4,y12,col="purple",pch=16)

```

- Ex 6.11

```r
# this question meets the prerequisite of using normal distribution as
    approximation
n=200
p=0.3
norm_mean=n*p
norm_std=sqrt(n*p*(1-p))
```

```
#calculate P(X<=50)=1-P(X>50)
1-pnorm(50,mean = norm_mean,sd=norm_std,lower.tail = TRUE)
# if using binomial distribution
1-pbinom(50,size = n,prob = p,lower.tail = TRUE)
# the output are pretty close to each other

```

The output of this **Python** code gives:

```
> n=200
> p=0.3
> norm_mean=n*p
> norm_std=sqrt(n*p*(1-p))
> 1-pnorm(50,mean = norm_mean,sd=norm_std,lower.tail = TRUE)
[1] 0.9385887
> 1-pbinom(50,size = n,prob = p,lower.tail = TRUE)
[1] 0.9304547
```

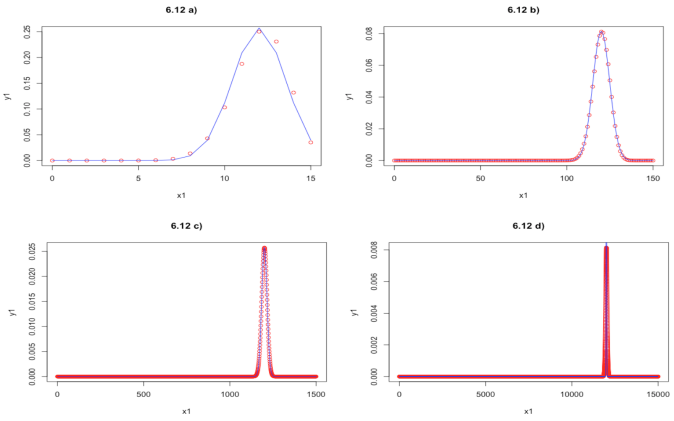
- Ex 6.12

```
p=0.8
# a) exact
pbinom(11,15,prob = p,lower.tail = TRUE)
# a) approximating by normal distribution
pnorm(11,mean = 15*p,sd=sqrt(15*p*(1-p)),lower.tail=TRUE)

# b) exact
pbinom(110,150,prob = p,lower.tail = TRUE)
# b) approximating by normal distribution
pnorm(110,mean = 150*p,sd=sqrt(150*p*(1-p)),lower.tail=TRUE)

# c) exact
pbinom(1100,1500,prob = p,lower.tail = TRUE)
# c) approximating by normal distribution
pnorm(1100,mean = 1500*p,sd=sqrt(1500*p*(1-p)),lower.tail=TRUE)

# d) exact
pbinom(11000,15000,prob = p,lower.tail = TRUE)
# d) approximating by normal distribution
pnorm(11000,mean = 15000*p,sd=sqrt(15000*p*(1-p)),lower.tail=TRUE)

```

Output:

```
> p=0.8
> pbinom(11,15,prob = p,lower.tail = TRUE)
[1] 0.3518379
> pnorm(11,mean = 15*p,sd=sqrt(15*p*(1-p)),lower.tail=TRUE)
[1] 0.2593025

> pbinom(110,150,prob = p,lower.tail = TRUE)
[1] 0.02929984
> pnorm(110,mean = 150*p,sd=sqrt(150*p*(1-p)),lower.tail=TRUE)
[1] 0.02061342
>
> pbinom(1100,1500,prob = p,lower.tail = TRUE)
[1] 2.852455e-10
> pnorm(1100,mean = 1500*p,sd=sqrt(1500*p*(1-p)),lower.tail=TRUE)
[1] 5.411937e-11
>
> pbinom(11000,15000,prob = p,lower.tail = TRUE)
[1] 1.952736e-86
> pnorm(11000,mean = 15000*p,sd=sqrt(15000*p*(1-p)),lower.tail=TRUE)
[1] 6.485784e-93
```

Apart from what is being asked, we can also view the problem from a graphical perspective:



The code that generates these plots is shown below:

```
# a) plot interpretation
x1 <- seq(0,15,by=1)
y1 <- dbinom(x1,15,0.8)
y2 <- dnorm(x1,mean = 15*0.8,sd=sqrt(15*0.8*0.2))
plot11 <- plot(x1,y1,col="red",main = "6.12 a)")
lines(x1,y2,col="blue")
# b) plot interpretation
x1 <- seq(0,150,by=1)
y1 <- dbinom(x1,150,0.8)
y2 <- dnorm(x1,mean = 150*0.8,sd=sqrt(150*0.8*0.2))
plot11 <- plot(x1,y1,col="red",main = "6.12 b)")
lines(x1,y2,col="blue")
# c) plot interpretation
x1 <- seq(0,1500,by=1)
y1 <- dbinom(x1,1500,0.8)
y2 <- dnorm(x1,mean = 1500*0.8,sd=sqrt(1500*0.8*0.2))
plot11 <- plot(x1,y1,col="red",main = "6.12 c)")
lines(x1,y2,col="blue")
# d) plot interpretation
x1 <- seq(0,15000,by=1)
y1 <- dbinom(x1,15000,0.8)
y2 <- dnorm(x1,mean = 15000*0.8,sd=sqrt(1500*0.8*0.2))
plot11 <- plot(x1,y1,col="red",main = "6.12 d)")
lines(x1,y2,col="blue")
```

Chapter 9

- Ex 9.1

  Using **R**:

```
# a)
t_0.05<-qt(0.95,df=5)
round(t_0.05,digits=3)
# b)
acid_lower<-12.4 - qt(0.975,df=9)*sqrt(63.1556/10)
acid_upper<-12.4 + qt(0.975,df=9)*sqrt(63.1556/10)
```

```
print(paste(round(acid_lower,digits=4),
round(acid_upper,digits=4)))
```

the output for this **R** code gives:

```
> # a)
> t_0.05<-qt(0.95,df=5)
> round(t_0.05,digits=3)
[1] 2.015
> # b)
> acid_lower <- 12.4 - qt(0.975,df=9)*sqrt(63.1556/10)
> acid_upper <- 12.4 + qt(0.975,df=9)*sqrt(63.1556/10)
> print(paste(round(acid_lower,digits=4),
    round(acid_upper,digits=4)))
[1] "6.715 18.085"
```

Chapter 10

- Ex 10.4

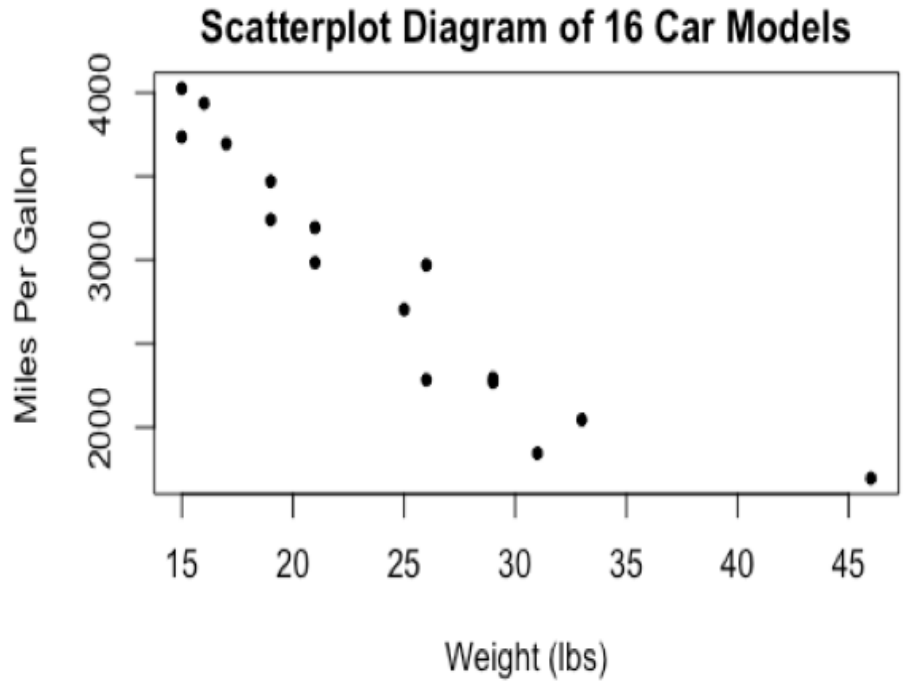
  Using **R**:

```
# a)
x_lb<-c(2705,3470,3935,3195,4025,
2270,1845,3735,1695,2285,3695,
2970,2295,3240,2045,2985)
y_mpg<-c(25,19,16,21,15,
29,31,15,46,26,17,
26,29,19,33,21)
plot(x_lb,y_mpg,main="Exercise 10.4a",
xlab="Weight (lbs)", ylab="Miles Per Gallon",
pch=20,cex=0.8)
# b)
r<-cor(x_lb,y_mpg)
round(r,digits=4)
# c)
c<-lm(y_mpg~x_lb)
plot(c,main="Exercise 10.4c",
      xlab="Weight (lbs)",ylab="Miles Per Gallon",
      pch=20,cex=0.8)
```

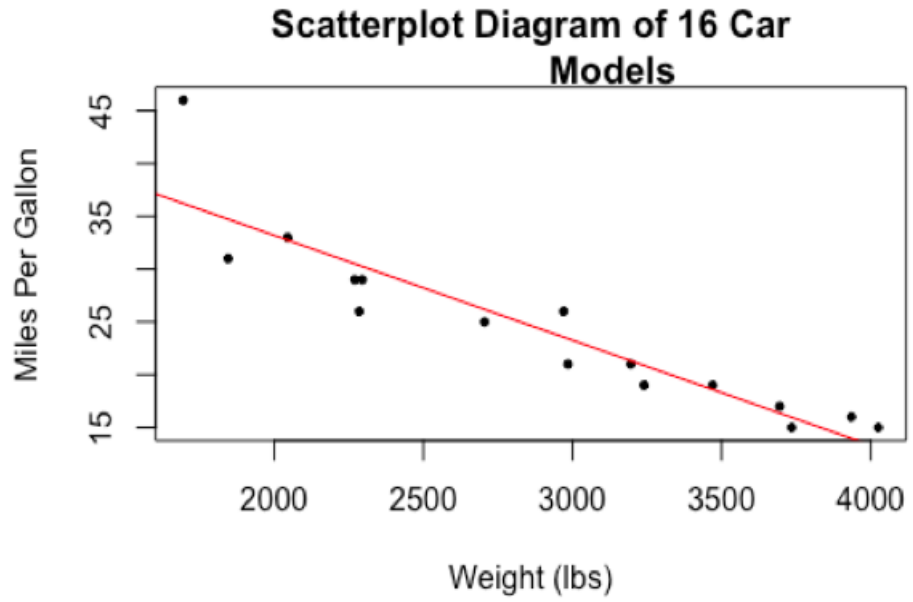
```
abline(c,col="red")
# d)
summary(c)
# e)
x_car <- 2500
y_car <- 53.05-0.01*x_car
y_car
df_car <- data.frame(x_lb=2500)
predict(c, df_car, se.fit=TRUE, level=0.95,
interval = 'confidence')
```

the output for this **R** code gives:

```
> # a)
> x_lb<-c(2705,3470,3935,3195,4025,2270,1845,
3735,1695,2285,3695,2970,2295,3240,2045,2985)
> y_mpg<-c(25,19,16,21,15,29,31,15,46,26,17,26,29,19,33,21)
> plot(x_lb,y_mpg,main="Exercise 10.4a",xlab="Weight (lbs)",
ylab="Miles Per Gallon",pch=20,cex=0.8)
```



```
> # b)
> cor(x_lb,y_mpg)
> round(r,digits=4)
[1] -0.9186
> # c)
> c <- lm(y_mpg~x_lb)
> plot(c,main="Exercise 10.4c",
     xlab="Weight (lbs)",ylab="Miles Per Gallon",
     pch=20,cex=0.8)
> abline(c,col="red")
```




```
> # d)
> summary(c)
Call:
lm(formula = y_mpg ~ x_lb, data = cars)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-4.3522 -1.5926 -0.6320  0.9718  9.7877

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 53.047551   3.417447  15.523 3.23e-10 ***
x_lb        -0.009932   0.001142  -8.694 5.13e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.366 on 14 degrees of freedom
Multiple R-squared:  0.8437,Adjusted R-squared:  0.8326
F-statistic: 75.59 on 1 and 14 DF,  p-value: 5.131e-07
> # e)
> x_car <- 2500
> y_car <- 53.05-0.01*x_car
> y_car
[1] 28.05
> df_car <- data.frame(x_lb=2500)
> predict(c, df_car, se.fit=TRUE, level=0.95, interval = 'confidence')
$fit
       fit      lwr      upr
1 28.21672 26.16373 30.26972

$se.fit
[1] 0.9572027

$df
[1] 14

$residual.scale
[1] 3.365896
```

- Ex 10.11

  Using **R**:

```
# a)
x_hrs<-c(5,6,8,4,6,7)
y_score<-c(70,75,85,60,72,90)
cor(x_hrs,y_score)
# b)
score<-lm(y_score~x_hrs)
summary(score)
```

the output for this **R** code gives:

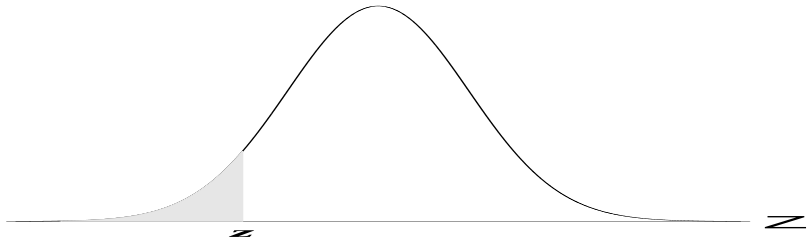
```
> # a)
> x_hrs <- c(5,6,8,4,6,7)
> y_score <- c(70,75,85,60,72,90)
> r <- cor(x_hrs,y_score)
> round(r,digits=4)
[1] 0.9165
> # b)
> score<-lm(y_score~x_hrs)
> summary(score)
Call:
lm(formula = y_score ~ x_hrs, data = exam)

Residuals:
      1       2       3       4       5       6
 1.6667 -0.3333 -4.3333 -1.3333 -3.3333  7.6667

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   33.333      9.375   3.556   0.0237 *
x_hrs          7.000      1.528   4.583   0.0102 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 4.83 on 4 degrees of freedom
Multiple R-squared:   0.84,Adjusted R-squared:    0.8
F-statistic:    21 on 1 and 4 DF,  p-value: 0.01016
```

Appendix E

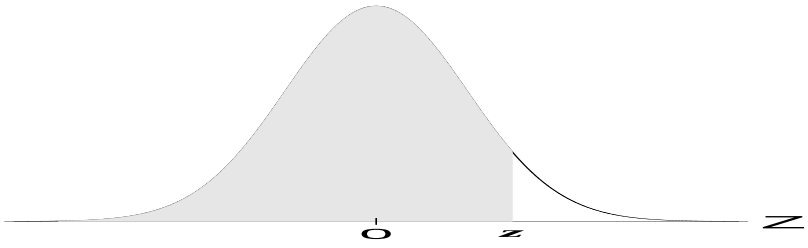
Tables

Table A: Areas for a Standard Normal distribution



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Table A: Areas for a Standard Normal distribution (contd.)



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Table B: Binomial probabilities

		p										
n	k	0.1	0.2	0.25	0.3	0.4	0.5	0.6	0.7	0.75	0.8	0.9
1	0	.9000	.8000	.7500	.7000	.6000	.5000	.4000	.3000	.2500	.2000	.1000
	1	.1000	.2000	.2500	.3000	.4000	.5000	.6000	.7000	.7500	.8000	.9000
2	0	.8100	.6400	.5625	.4900	.3600	.2500	.1600	.0900	.0625	.0400	.0100
	1	.1800	.3200	.3750	.4200	.4800	.5000	.4800	.4200	.3750	.3200	.1800
	2	.0100	.0400	.0625	.0900	.1600	.2500	.3600	.4900	.5625	.6400	.8100
3	0	.7290	.5120	.4219	.3430	.2160	.1250	.0640	.0270	.0156	.0080	.0010
	1	.2430	.3840	.4219	.4410	.4320	.3750	.2880	.1890	.1406	.0960	.0270
	2	.0270	.0960	.1406	.1890	.2880	.3750	.4320	.4410	.4219	.3840	.2430
	3	.0010	.0080	.0156	.0270	.0640	.1250	.2160	.3430	.4219	.5120	.7290
4	0	.6561	.4096	.3164	.2401	.1296	.0625	.0256	.0081	.0039	.0016	.0001
	1	.2916	.4096	.4219	.4116	.3456	.2500	.1536	.0756	.0469	.0256	.0036
	2	.0486	.1536	.2109	.2646	.3456	.3750	.3456	.2646	.2109	.1536	.0486
	3	.0036	.0256	.0469	.0756	.1536	.2500	.3456	.4116	.4219	.4096	.2916
	4	.0001	.0016	.0039	.0081	.0256	.0625	.1296	.2401	.3164	.4096	.6561
5	0	.5905	.3277	.2373	.1681	.0778	.0312	.0102	.0024	.0010	.0003	
	1	.3280	.4096	.3955	.3601	.2592	.1562	.0768	.0284	.0146	.0064	.0004
	2	.0729	.2048	.2637	.3087	.3456	.3125	.2304	.1323	.0879	.0512	.0081
	3	.0081	.0512	.0879	.1323	.2304	.3125	.3456	.3087	.2637	.2048	.0729
	4	.0005	.0064	.0146	.0283	.0768	.1562	.2592	.3601	.3955	.4096	.3280
	5		.0003	.0010	.0024	.0102	.0312	.0778	.1681	.2373	.3277	.5905
6	0	.5314	.2621	.1780	.1176	.0467	.0156	.0041	.0007	.0002	.0001	
	1	.3543	.3932	.3560	.3025	.1866	.0938	.0369	.0102	.0044	.0015	.0001
	2	.0984	.2458	.2966	.3241	.3110	.2344	.1382	.0595	.0330	.0154	.0012
	3	.0146	.0819	.1318	.1852	.2765	.3125	.2765	.1852	.1318	.0819	.0146
	4	.0012	.0154	.0330	.0595	.1382	.2344	.3110	.3241	.2966	.2458	.0984
	5	.0001	.0015	.0044	.0102	.0369	.0938	.1866	.3025	.3560	.3932	.3543
	6		.0001	.0002	.0007	.0041	.0156	.0467	.1176	.1780	.2621	.5314
7	0	.4783	.2097	.1335	.0824	.0280	.0078	.0016	.0002	.0001		
	1	.3720	.3670	.3115	.2471	.1306	.0547	.0172	.0036	.0013	.0004	
	2	.1240	.2753	.3115	.3177	.2613	.1641	.0774	.0250	.0115	.0043	.0002
	3	.0230	.1147	.1730	.2269	.2903	.2734	.1935	.0972	.0577	.0287	.0026
	4	.0026	.0287	.0577	.0972	.1935	.2734	.2903	.2269	.1730	.1147	.0230
	5	.0002	.0043	.0115	.0250	.0774	.1641	.2613	.3177	.3115	.2753	.1240
	6		.0004	.0013	.0036	.0172	.0547	.1306	.2471	.3115	.3670	.3720
	7			.0001	.0002	.0016	.0078	.0280	.0824	.1335	.2097	.4783
8	0	.4305	.1678	.1001	.0576	.0168	.0039	.0007	.0001			
	1	.3826	.3355	.2670	.1977	.0896	.0313	.0079	.0012	.0004	.0001	
	2	.1488	.2936	.3115	.2965	.2090	.1094	.0413	.0100	.0038	.0011	
	3	.0331	.1468	.2076	.2541	.2787	.2188	.1239	.0467	.0231	.0092	.0004
	4	.0046	.0459	.0865	.1361	.2322	.2734	.2322	.1361	.0865	.0459	.0046
	5	.0004	.0092	.0231	.0467	.1239	.2188	.2787	.2541	.2076	.1468	.0331
	6		.0011	.0038	.0100	.0413	.1094	.2090	.2965	.3115	.2936	.1488

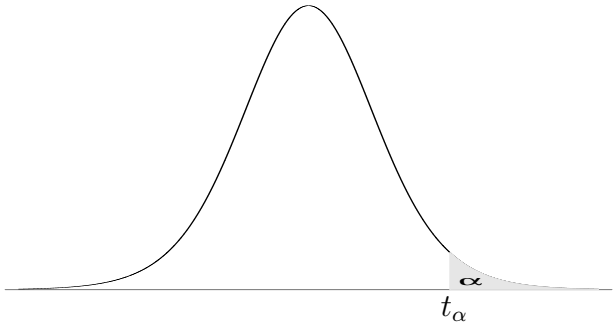
		p										
n	k	0.1	0.2	0.25	0.3	0.4	0.5	0.6	0.7	0.75	0.8	0.9
8	7		.0001	.0004	.0012	.0079	.0313	.0896	.1977	.2670	.3355	.3826
	8				.0001	.0007	.0039	.0168	.0576	.1001	.1678	.4305
9	0	.3874	.1342	.0751	.0404	.0101	.0020	.0003				
	1	.3874	.3020	.2253	.1556	.0605	.0176	.0035	.0004	.0001		
	2	.1722	.3020	.3003	.2668	.1612	.0703	.0212	.0039	.0012	.0003	
	3	.0446	.1762	.2336	.2668	.2508	.1641	.0743	.0210	.0087	.0028	.0001
	4	.0074	.0661	.1168	.1715	.2508	.2461	.1672	.0735	.0389	.0165	.0008
	5	.0008	.0165	.0389	.0735	.1672	.2461	.2508	.1715	.1168	.0661	.0074
	6	.0001	.0028	.0087	.0210	.0743	.1641	.2508	.2668	.2336	.1762	.0446
	7		.0003	.0012	.0039	.0212	.0703	.1612	.2668	.3003	.3020	.1722
	8			.0001	.0004	.0035	.0176	.0605	.1556	.2253	.3020	.3874
	9					.0003	.0020	.0101	.0404	.0751	.1342	.3874
10	0	.3487	.1074	.0563	.0282	.0060	.0010	.0001				
	1	.3874	.2684	.1877	.1211	.0403	.0098	.0016	.0001			
	2	.1937	.3020	.2816	.2335	.1209	.0439	.0106	.0014	.0004	.0001	
	3	.0574	.2013	.2503	.2668	.2150	.1172	.0425	.0090	.0031	.0008	
	4	.0112	.0881	.1460	.2001	.2508	.2051	.1115	.0368	.0162	.0055	.0001
	5	.0015	.0264	.0584	.1029	.2007	.2461	.2007	.1029	.0584	.0264	.0015
	6	.0001	.0055	.0162	.0368	.1115	.2051	.2508	.2001	.1460	.0881	.0112
	7		.0008	.0031	.0090	.0425	.1172	.2150	.2668	.2503	.2013	.0574
	8		.0001	.0004	.0014	.0106	.0439	.1209	.2335	.2816	.3020	.1937
	9				.0001	.0016	.0098	.0403	.1211	.1877	.2684	.3874
	10					.0001	.0010	.0060	.0282	.0563	.1074	.3487
11	0	.3138	.0859	.0422	.0198	.0036	.0005					
	1	.3835	.2362	.1549	.0932	.0266	.0054	.0007				
	2	.2131	.2953	.2581	.1998	.0887	.0269	.0052	.0005	.0001		
	3	.0710	.2215	.2581	.2568	.1774	.0806	.0234	.0037	.0011	.0002	
	4	.0158	.1107	.1721	.2201	.2365	.1611	.0701	.0173	.0064	.0017	
	5	.0025	.0388	.0803	.1321	.2207	.2256	.1471	.0566	.0268	.0097	.0003
	6	.0003	.0097	.0268	.0566	.1471	.2256	.2207	.1321	.0803	.0388	.0025
	7		.0017	.0064	.0173	.0701	.1611	.2365	.2201	.1721	.1107	.0158
	8		.0002	.0011	.0037	.0234	.0806	.1774	.2568	.2581	.2215	.0710
	9			.0001	.0005	.0052	.0269	.0887	.1998	.2581	.2953	.2131
	10					.0007	.0054	.0266	.0932	.1549	.2362	.3835
	11						.0005	.0036	.0198	.0422	.0859	.3138
12	0	.2824	.0687	.0317	.0138	.0022	.0002					
	1	.3766	.2062	.1267	.0712	.0174	.0029	.0003				
	2	.2301	.2835	.2323	.1678	.0639	.0161	.0025	.0002			
	3	.0852	.2362	.2581	.2397	.1419	.0537	.0125	.0015	.0004	.0001	
	4	.0213	.1329	.1936	.2311	.2128	.1208	.0420	.0078	.0024	.0005	
	5	.0038	.0532	.1032	.1585	.2270	.1934	.1009	.0291	.0115	.0033	
	6	.0005	.0155	.0401	.0792	.1766	.2256	.1766	.0792	.0401	.0155	.0005
	7		.0033	.0115	.0291	.1009	.1934	.2270	.1585	.1032	.0532	.0038
	8		.0005	.0024	.0078	.0420	.1208	.2128	.2311	.1936	.1329	.0213
	9		.0001	.0004	.0015	.0125	.0537	.1419	.2397	.2581	.2362	.0852
	10				.0002	.0025	.0161	.0639	.1678	.2323	.2835	.2301
	11					.0003	.0029	.0174	.0712	.1267	.2062	.3766
	12						.0002	.0022	.0138	.0317	.0687	.2824

		p										
n	k	0.1	0.2	0.25	0.3	0.4	0.5	0.6	0.7	0.75	0.8	0.9
13	0	.2542	.0550	.0238	.0097	.0013	.0001					
	1	.3672	.1787	.1029	.0540	.0113	.0016	.0001				
	2	.2448	.2680	.2059	.1388	.0453	.0095	.0012	.0001			
	3	.0997	.2457	.2517	.2181	.1107	.0349	.0065	.0006	.0001		
	4	.0277	.1535	.2097	.2337	.1845	.0873	.0243	.0034	.0009	.0001	
	5	.0055	.0691	.1258	.1803	.2214	.1571	.0656	.0142	.0047	.0011	
	6	.0008	.0230	.0559	.1030	.1968	.2095	.1312	.0442	.0186	.0058	.0001
	7	.0001	.0058	.0186	.0442	.1312	.2095	.1968	.1030	.0559	.0230	.0008
	8		.0011	.0047	.0142	.0656	.1571	.2214	.1803	.1258	.0691	.0055
	9		.0001	.0009	.0034	.0243	.0873	.1845	.2337	.2097	.1535	.0277
	10			.0001	.0006	.0065	.0349	.1107	.2181	.2517	.2457	.0997
	11				.0001	.0012	.0095	.0453	.1388	.2059	.2680	.2448
	12					.0001	.0016	.0113	.0540	.1029	.1787	.3672
	13						.0001	.0013	.0097	.0238	.0550	.2542
14	0	.2288	.0440	.0178	.0068	.0008	.0001					
	1	.3559	.1539	.0832	.0407	.0073	.0009	.0001				
	2	.2570	.2501	.1802	.1134	.0317	.0056	.0005				
	3	.1142	.2501	.2402	.1943	.0845	.0222	.0033	.0002			
	4	.0349	.1720	.2202	.2290	.1549	.0611	.0136	.0014	.0003		
	5	.0078	.0860	.1468	.1963	.2066	.1222	.0408	.0066	.0018	.0003	
	6	.0013	.0322	.0734	.1262	.2066	.1833	.0918	.0232	.0082	.0020	
	7	.0002	.0092	.0280	.0618	.1574	.2095	.1574	.0618	.0280	.0092	.0002
	8		.0020	.0082	.0232	.0918	.1833	.2066	.1262	.0734	.0322	.0013
	9		.0003	.0018	.0066	.0408	.1222	.2066	.1963	.1468	.0860	.0078
	10			.0003	.0014	.0136	.0611	.1549	.2290	.2202	.1720	.0349
	11				.0002	.0033	.0222	.0845	.1943	.2402	.2501	.1142
	12					.0005	.0056	.0317	.1134	.1802	.2501	.2570
	13					.0001	.0009	.0073	.0407	.0832	.1539	.3559
	14						.0001	.0008	.0068	.0178	.0440	.2288
15	0	.2059	.0352	.0134	.0047	.0005						
	1	.3432	.1319	.0668	.0305	.0047	.0005					
	2	.2669	.2309	.1559	.0916	.0219	.0032	.0003				
	3	.1285	.2501	.2252	.1700	.0634	.0139	.0016	.0001			
	4	.0428	.1876	.2252	.2186	.1268	.0417	.0074	.0006	.0001		
	5	.0105	.1032	.1651	.2061	.1859	.0916	.0245	.0030	.0007	.0001	
	6	.0019	.0430	.0917	.1472	.2066	.1527	.0612	.0116	.0034	.0007	
	7	.0003	.0138	.0393	.0811	.1771	.1964	.1181	.0348	.0131	.0035	
	8		.0035	.0131	.0348	.1181	.1964	.1771	.0811	.0393	.0138	.0003
	9		.0007	.0034	.0116	.0612	.1527	.2066	.1472	.0917	.0430	.0019
	10		.0001	.0007	.0030	.0245	.0916	.1859	.2061	.1651	.1032	.0105
	11			.0001	.0006	.0074	.0417	.1268	.2186	.2252	.1876	.0428
	12				.0001	.0016	.0139	.0634	.1700	.2252	.2501	.1285
	13					.0003	.0032	.0219	.0916	.1559	.2309	.2669
	14						.0005	.0047	.0305	.0668	.1319	.3432
	15							.0005	.0047	.0134	.0352	.2059

		p										
n	k	0.1	0.2	0.25	0.3	0.4	0.5	0.6	0.7	0.75	0.8	0.9
16	0	.1853	.0281	.0100	.0033	.0003						
	1	.3294	.1126	.0535	.0228	.0030	.0002					
	2	.2745	.2111	.1336	.0732	.0150	.0018	.0001				
	3	.1423	.2463	.2079	.1465	.0468	.0085	.0008				
	4	.0514	.2001	.2252	.2040	.1014	.0278	.0040	.0002			
	5	.0137	.1201	.1802	.2099	.1623	.0667	.0142	.0013	.0002		
	6	.0028	.0550	.1101	.1649	.1983	.1222	.0392	.0056	.0014	.0002	
	7	.0004	.0197	.0524	.1010	.1889	.1746	.0840	.0185	.0058	.0012	
	8	.0001	.0055	.0197	.0487	.1417	.1964	.1417	.0487	.0197	.0055	.0001
	9		.0012	.0058	.0185	.0840	.1746	.1889	.1010	.0524	.0197	.0004
	10		.0002	.0014	.0056	.0392	.1222	.1983	.1649	.1101	.0550	.0028
	11			.0002	.0013	.0142	.0667	.1623	.2099	.1802	.1201	.0137
	12				.0002	.0040	.0278	.1014	.2040	.2252	.2001	.0514
	13					.0008	.0085	.0468	.1465	.2079	.2463	.1423
	14					.0001	.0018	.0150	.0732	.1336	.2111	.2745
	15						.0002	.0030	.0228	.0535	.1126	.3294
	16							.0003	.0033	.0100	.0281	.1853
17	0	.1668	.0225	.0075	.0023	.0002						
	1	.3150	.0957	.0426	.0169	.0019	.0001					
	2	.2800	.1914	.1136	.0581	.0102	.0010	.0001				
	3	.1556	.2393	.1893	.1245	.0341	.0052	.0004				
	4	.0605	.2093	.2209	.1868	.0796	.0182	.0021	.0001			
	5	.0175	.1361	.1914	.2081	.1379	.0472	.0081	.0006	.0001		
	6	.0039	.0680	.1276	.1784	.1839	.0944	.0242	.0026	.0005	.0001	
	7	.0007	.0267	.0668	.1201	.1927	.1484	.0571	.0095	.0025	.0004	
	8	.0001	.0084	.0279	.0644	.1606	.1855	.1070	.0276	.0093	.0021	
	9		.0021	.0093	.0276	.1070	.1855	.1606	.0644	.0279	.0084	.0001
	10		.0004	.0025	.0095	.0571	.1484	.1927	.1201	.0668	.0267	.0007
	11		.0001	.0005	.0026	.0242	.0944	.1839	.1784	.1276	.0680	.0039
	12			.0001	.0006	.0081	.0472	.1379	.2081	.1914	.1361	.0175
	13				.0001	.0021	.0182	.0796	.1868	.2209	.2093	.0605
	14					.0004	.0052	.0341	.1245	.1893	.2393	.1556
	15					.0001	.0010	.0102	.0581	.1136	.1914	.2800
	16						.0001	.0019	.0169	.0426	.0957	.3150
	17							.0002	.0023	.0075	.0225	.1668
18	0	.1501	.0180	.0056	.0016	.0001						
	1	.3002	.0811	.0338	.0126	.0012	.0001					
	2	.2835	.1723	.0958	.0458	.0069	.0006					
	3	.1680	.2297	.1704	.1046	.0246	.0031	.0002				
	4	.0700	.2153	.2130	.1681	.0614	.0117	.0011				
	5	.0218	.1507	.1988	.2017	.1146	.0327	.0045	.0002			
	6	.0052	.0816	.1436	.1873	.1655	.0708	.0145	.0012	.0002		
	7	.0010	.0350	.0820	.1376	.1892	.1214	.0374	.0046	.0010	.0001	
	8	.0002	.0120	.0376	.0811	.1734	.1669	.0771	.0149	.0042	.0008	
	9		.0033	.0139	.0386	.1284	.1855	.1284	.0386	.0139	.0033	
	10		.0008	.0042	.0149	.0771	.1669	.1734	.0811	.0376	.0120	.0002
	11		.0001	.0010	.0046	.0374	.1214	.1892	.1376	.0820	.0350	.0010
	12			.0002	.0012	.0145	.0708	.1655	.1873	.1436	.0816	.0052
	13				.0002	.0045	.0327	.1146	.2017	.1988	.1507	.0218
	14					.0011	.0117	.0614	.1681	.2130	.2153	.0700

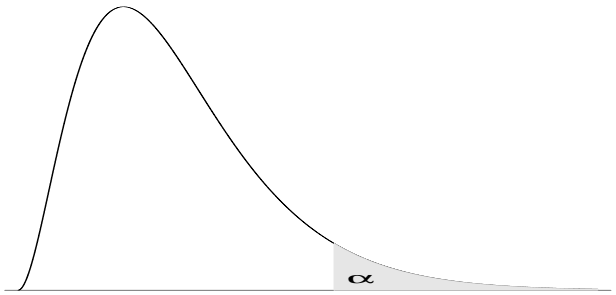
		p										
n	k	0.1	0.2	0.25	0.3	0.4	0.5	0.6	0.7	0.75	0.8	0.9
18	15					.0002	.0031	.0246	.1046	.1704	.2297	.1680
	16						.0006	.0069	.0458	.0958	.1723	.2835
	17						.0001	.0012	.0126	.0338	.0811	.3002
	18							.0001	.0016	.0056	.0180	.1501
19	0	.1351	.0144	.0042	.0011	.0001						
	1	.2852	.0685	.0268	.0093	.0008						
	2	.2852	.1540	.0803	.0358	.0046	.0003					
	3	.1796	.2182	.1517	.0869	.0175	.0018	.0001				
	4	.0798	.2182	.2023	.1491	.0467	.0074	.0005				
	5	.0266	.1636	.2023	.1916	.0933	.0222	.0024	.0001			
	6	.0069	.0955	.1574	.1916	.1451	.0518	.0085	.0005	.0001		
	7	.0014	.0443	.0974	.1525	.1797	.0961	.0237	.0022	.0004		
	8	.0002	.0166	.0487	.0981	.1797	.1442	.0532	.0077	.0018	.0003	
	9		.0051	.0198	.0514	.1464	.1762	.0976	.0220	.0066	.0013	
	10		.0013	.0066	.0220	.0976	.1762	.1464	.0514	.0198	.0051	
	11		.0003	.0018	.0077	.0532	.1442	.1797	.0981	.0487	.0166	.0002
	12			.0004	.0022	.0237	.0961	.1797	.1525	.0974	.0443	.0014
	13			.0001	.0005	.0085	.0518	.1451	.1916	.1574	.0955	.0069
	14				.0001	.0024	.0222	.0933	.1916	.2023	.1636	.0266
	15					.0005	.0074	.0467	.1491	.2023	.2182	.0798
	16					.0001	.0018	.0175	.0869	.1517	.2182	.1796
	17						.0003	.0046	.0358	.0803	.1540	.2852
	18							.0008	.0093	.0268	.0685	.2852
	19							.0001	.0011	.0042	.0144	.1351
20	0	.1216	.0115	.0032	.0008							
	1	.2702	.0576	.0211	.0068	.0005						
	2	.2852	.1369	.0669	.0278	.0031	.0002					
	3	.1901	.2054	.1339	.0716	.0123	.0011					
	4	.0898	.2182	.1897	.1304	.0350	.0046	.0003				
	5	.0319	.1746	.2023	.1789	.0746	.0148	.0013				
	6	.0089	.1091	.1686	.1916	.1244	.0370	.0049	.0002			
	7	.0020	.0545	.1124	.1643	.1659	.0739	.0146	.0010	.0002		
	8	.0004	.0222	.0609	.1144	.1797	.1201	.0355	.0039	.0008	.0001	
	9	.0001	.0074	.0271	.0654	.1597	.1602	.0710	.0120	.0030	.0005	
	10		.0020	.0099	.0308	.1171	.1762	.1171	.0308	.0099	.0020	
	11		.0005	.0030	.0120	.0710	.1602	.1597	.0654	.0271	.0074	.0001
	12		.0001	.0008	.0039	.0355	.1201	.1797	.1144	.0609	.0222	.0004
	13			.0002	.0010	.0146	.0739	.1659	.1643	.1124	.0545	.0020
	14				.0002	.0049	.0370	.1244	.1916	.1686	.1091	.0089
	15					.0013	.0148	.0746	.1789	.2023	.1746	.0319
	16					.0003	.0046	.0350	.1304	.1897	.2182	.0898
	17						.0011	.0123	.0716	.1339	.2054	.1901
	18						.0002	.0031	.0278	.0669	.1369	.2852
	19							.0005	.0068	.0211	.0576	.2702
	20								.0008	.0032	.0115	.1216

Table C: Critical Values of t distribution



df	$t_{.10}$	$t_{.05}$	$t_{.025}$	$t_{.01}$	$t_{.005}$	$t_{.0025}$
1	3.078	6.314	12.71	31.82	63.66	127.3
2	1.886	2.920	4.303	6.965	9.925	14.09
3	1.638	2.353	3.182	4.541	5.841	7.453
4	1.533	2.132	2.776	3.747	4.604	5.59
5	1.476	2.015	2.571	3.365	4.032	4.77
6	1.440	1.943	2.447	3.143	3.707	4.31
7	1.415	1.895	2.365	2.998	3.499	4.029
8	1.397	1.860	2.306	2.896	3.355	3.833
9	1.383	1.833	2.262	2.821	3.250	3.690
10	1.372	1.812	2.228	2.764	3.169	3.581
11	1.363	1.796	2.201	2.718	3.106	3.497
12	1.356	1.782	2.179	2.681	3.055	3.428
13	1.350	1.771	2.160	2.650	3.012	3.372
14	1.345	1.761	2.145	2.624	2.977	3.326
15	1.341	1.753	2.131	2.602	2.947	3.286
16	1.337	1.746	2.120	2.583	2.921	3.252
17	1.333	1.740	2.110	2.567	2.898	3.222
18	1.330	1.734	2.101	2.552	2.878	3.197
19	1.328	1.729	2.093	2.539	2.861	3.174
20	1.325	1.725	2.086	2.528	2.845	3.153
21	1.323	1.721	2.080	2.518	2.831	3.135
22	1.321	1.717	2.074	2.508	2.819	3.119
23	1.319	1.714	2.069	2.500	2.807	3.104
24	1.318	1.711	2.064	2.492	2.797	3.091
25	1.316	1.708	2.060	2.485	2.787	3.078
26	1.315	1.706	2.056	2.479	2.779	3.067
27	1.314	1.703	2.052	2.473	2.771	3.057
28	1.313	1.701	2.048	2.467	2.763	3.047
29	1.311	1.699	2.045	2.462	2.756	3.038
30	1.310	1.697	2.042	2.457	2.750	3.030
40	1.303	1.684	2.021	2.423	2.704	2.971
60	1.296	1.671	2.000	2.390	2.660	2.915
100	1.290	1.660	1.984	2.364	2.626	2.871
500	1.283	1.648	1.965	2.334	2.586	2.820
1000	1.282	1.646	1.962	2.330	2.581	2.813
z-values : ∞	1.282	1.645	1.960	2.326	2.576	2.807
Confidence level C	80%	90%	95%	98%	99%	99.5%

Table D: Critical Values of χ^2 distribution



df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Appendix F

Summary of important formulae and tests

Chapter 2

- Mean

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Sample Variance

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

- Sample Standard Deviation (SD)

$$s = \sqrt{\text{sample variance}} = \sqrt{s^2}$$

- Range = (Maximum value − Minimum value)
- Interquartile Range (IQR) = $Q_3 - Q_1$
- Five point summary consists of : (minimum, Q_1, median, Q_3, maximum)

Chapter 3

- The probability of any event A is a number between 0 and 1.
- In any random experiment, the sum of probabilities of all possible outcomes is 1.
- If A and B are events with no common outcomes,

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

- For any event A,

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

- For any two events A and B,

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

- $$\mathrm{P}(B|A) = \frac{\mathrm{P}(A \text{ and } B)}{\mathrm{P}(A)} \text{ if } \mathrm{P}(B) > 0$$

- $$\mathrm{P}(A \text{ and } B) = \mathrm{P}(A) \times \mathrm{P}(B) \text{ if } A \text{ and } B \text{ are independent}$$

- $$\begin{aligned} \mathrm{P}(A) &= \mathrm{P}(A \text{ and } B_1) + \mathrm{P}(A \text{ and } B_2) + \ldots + \mathrm{P}(A \text{ and } B_k) \\ &= \mathrm{P}(A|B_1)\mathrm{P}(B_1) + \mathrm{P}(A|B_2)\mathrm{P}(B_2) + \cdots + \mathrm{P}(A|B_k)\mathrm{P}(B_k) \end{aligned}$$

- $$\mathrm{P}(B_i|A) = \frac{\mathrm{P}(A|B_i)\mathrm{P}(B_i)}{\mathrm{P}(A|B_1)\mathrm{P}(B_1) + \mathrm{P}(A|B_2)\mathrm{P}(B_2) + \cdots + \mathrm{P}(A|B_k)\mathrm{P}(B_k)}$$

Chapter 4

- Mean of a discrete r.v. X

$$\mu_X = (x_1 \times p_1) + (x_2 \times p_2) + \cdots + (x_n \times p_n) = \sum_{i=1}^{n} (x_i \times p_i)$$

- Variance of a discrete r.v. X

$$\begin{aligned}\sigma_X^2 &= (x_1-\mu)^2\times p_1 + (x_2-\mu)^2\times p_2 + \cdots + (x_n-\mu)^2\times p_n \\ &= \sum_{i=1}^{n}(x_i-\mu)^2\times p_i \\ &= \sum_{i=1}^{n}(x_i^2\times p_i) - \mu^2\end{aligned}$$

Chapter 5

- Number of ways of selecting k items out of n items

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- $\mathrm{P}(k \text{ successes out of } n \text{ trials}) = \binom{n}{k}p^k(1-p)^{n-k}, \qquad k = 0,\ 1,\ \ldots,\ n, \qquad 0 < p < 1.$

- Mean for Binomial Distribution $\mu_X = np$
- Standard deviation for Binomial Distribution $\sigma_X = \sqrt{np(1-p)}$

Chapter 7

- μ lies between $\overline{x}\ \pm\ z_{\alpha/2}\ \frac{\sigma}{\sqrt{n}}$ with confidence $C = (1-\alpha)$ if the data is from a Normal curve with known σ.
- μ lies between $\overline{x} \pm\ t_{\alpha/2}\ \frac{s}{\sqrt{n}}$ with confidence $C = (1-\alpha)$ if the data is from a Normal curve with unknown σ.

- A $C = (1-\alpha)$ level confidence interval for the standard deviation σ based on a sample from a N($\mu,\ \sigma$) curve is given by

$$\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2}}},\ \sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}} \right)$$

  where $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$ are values from a $\chi^2(n-1)$ distribution with areas $\frac{\alpha}{2}$ below $\chi^2_{1-\alpha/2}$ and above $\chi^2_{\alpha/2}$, respectively.

- The true (population) proportion p lies between

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

  with confidence $C = (1-\alpha)$ where $\widehat{p} = \frac{x}{n}$ is the observed proportion and n is large.

- For constructing a $C = (1-\alpha)$ level confidence interval for μ when σ is known,

$$n = \left(z_{\alpha/2}\frac{\sigma}{m} \right)^2 = \frac{z^2_{\alpha/2}\sigma^2}{m^2}$$

  is the sample size required for the margin of error to be m.

- For constructing a $C = (1-\alpha)$ level confidence interval for p,

$$n = \frac{z^2_{\alpha/2}p_0(1-p_0)}{m^2}$$

  is the sample size required for the margin of error to be m, where p_0 is the best prior guess about p.

  If no prior information is available, take $p_0 = \frac{1}{2}$.

Chapter 9

- If σ_x and σ_y are known, use the z-statistic to obtain a $(1-\alpha) = C$ level confidence interval for $\mu_x - \mu_y$

$$(\overline{x} - \overline{y}) \pm \ z_{\alpha/2} \sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}$$

- If σ_x , σ_y are not known but are assumed equal, use the t-statistic to get a $C = (1-\alpha)$ level confidence interval for $\mu_x - \mu_y$

$$(\overline{x} - \overline{y}) \pm \ t_{\alpha/2}\, s_p \sqrt{\frac{1}{n} + \frac{1}{m}}$$

  Here, $t_{\alpha/2}$ has $(m+n-2)$ degrees of freedom and

$$s_p^2 = \frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2}$$

  is the pooled sample variance.

- If nothing is known about σ_x and σ_y, a $C = (1-\alpha)$ level confidence interval for $\mu_x - \mu_y$ is given by

$$(\overline{x} - \overline{y}) \pm \ t_{\alpha/2} \sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}}$$

  where we use a t-distribution with $k = \min\,(m-1, n-1)$ degrees of freedom to get $t_{\alpha/2}$.

Chapter 10

- The sample correlation coefficient

$$\begin{aligned} r_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{s_x}\right)\left(\frac{y_i-\overline{y}}{s_y}\right) \\ &= \frac{n(\sum x_i y_i)-(\sum x_i)(\sum y_i)}{\sqrt{[n(\sum x_i^2)-(\sum x_i)^2]\,[n(\sum y_i^2)-(\sum y_i)^2]}} \end{aligned}$$

- The least squares regression line has the equation:

$$\widehat{y} = a + bx$$

- $$b = \widehat{\beta} = r_{xy}(\frac{s_y}{s_x}) = \frac{n(\sum xy)-(\sum x)(\sum y)}{n(\sum x^2)-(\sum x)^2}$$

- $$a = \widehat{\alpha} = \overline{y} - b\overline{x}$$

- $$\text{residual}_i = e_i = (\text{observed } y_i - \text{ predicted } y_i) = (y_i - \widehat{y}_i) = y_i - (a + bx_i)$$

- $$s_e = \sqrt{\frac{\sum \text{residual}_i^2}{n-2}} = \sqrt{\frac{\sum (y_i-\widehat{y}_i)^2}{n-2}} = s_y\sqrt{\frac{n-1}{n-2}(1-r_{xy}^2)}$$

- $$\text{Standard error of } b = \text{SE}_b = \frac{s_e}{\sqrt{\sum (x_i-\overline{x})^2}} = \frac{s_e}{s_x\sqrt{n-1}}$$

- $$t = \frac{b}{\text{SE}_b}$$

  has a t-distribution with $(n-2)$ degrees of freedom

- Predicted mean corresponding to $x = x^*$ is given by

$$\widehat{\mu}_y = a + bx^*$$

and has standard error

$$\text{SE}_{\widehat{\mu}} = s_e\sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum(x_i - \overline{x})^2}}$$

- A $C = (1 - \alpha)$ level confidence interval for this predicted mean is given by

$$\widehat{\mu} \pm t_{\alpha/2} \times \text{SE}_{\widehat{\mu}}$$

where $t_{\alpha/2}$ comes from a t distribution with $(n - 2)$ df.

- Predicted value corresponding to $x = x^*$ is given by

$$\widehat{y} = a + bx^*$$

and has standard error

$$\text{SE}_{\widehat{y}} = s_e\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum(x_i - \overline{x})^2}}$$

- A $C = (1 - \alpha)$ level prediction interval for this predicted value is given by

$$\widehat{y} \pm \ t_{\alpha/2} \times \text{SE}_{\widehat{y}}$$

where $t_{\alpha/2}$ comes from a t distribution with $(n - 2)$ df.

Assumptions	H_0	H_a	Compute	Reject H_0 if	P-value
§8.3:	$\mu = \mu_0$	$\mu > \mu_0$	$z = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}$	$z > z_\alpha$	P(Z > z)
$x_1, \ldots, x_n$ are N(μ, σ).	$\mu = \mu_0$	$\mu < \mu_0$		$z < -z_\alpha$	P(Z < z)
σ known	$\mu = \mu_0$	$\mu \neq \mu_0$		$\lvert z\rvert > z_{\alpha/2}$	2 P(Z > $\lvert z\rvert$)
§8.4: $x_1, ..., x_n$ are N(μ, σ).	$\mu = \mu_0$	$\mu > \mu_0$	$t = \frac{\overline{x}-\mu_0}{s/\sqrt{n}}$	$t > t_\alpha$(n-1)	P(T > t)
σ unknown.	$\mu = \mu_0$	$\mu < \mu_0$		$t < -t_\alpha$(n-1)	P(T < t)
Find $s^2 = \frac{1}{n-1}\sum_1^n (x_i - \overline{x})^2$	$\mu = \mu_0$	$\mu \neq \mu_0$		$\lvert t\rvert > t_{\alpha/2}$(n-1)	2 P(T > $\lvert t\rvert$)
§8.6: Binomial(n,p)	$p = p_0$	$p > p_0$	$z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$z > z_\alpha$	P(Z > z)
$\hat{p} = \frac{x}{n}$ is the observed	$p = p_0$	$p < p_0$		$z < -z_\alpha$	P(Z < z)
proportion. (n large)	$p = p_0$	$p \neq p_0$		$\lvert z\rvert > z_{\alpha/2}$	2 P(Z > $\lvert z\rvert$)
§9.1: $(x_1, y_1), \ldots, (x_n, y_n)$	$\mu_x = \mu_y$	$\mu_x > \mu_y$	$t = \frac{\overline{d}}{s_d/\sqrt{n}}$	$t > t_\alpha$(n-1)	P(T > t)
are paired data. Find	$\mu_x = \mu_y$	$\mu_x < \mu_y$		$t < -t_\alpha$(n-1)	P(T < t)
$d_i = (x_i - y_i)$ and $\overline{d}$, s_d^2	$\mu_x = \mu_y$	$\mu_x \neq \mu_y$		$\lvert t\rvert > t_{\alpha/2}$(n-1)	2 P(T > $\lvert t\rvert$)
§9.2: σ_x, σ_y known	$\mu_x = \mu_y$	$\mu_x > \mu_y$	$z = \frac{\overline{x}-\overline{y}}{\sqrt{\frac{\sigma_x^2}{m}+\frac{\sigma_y^2}{n}}}$	$z > z_\alpha$	P(Z > z)
$x_1, \ldots, x_m$ are N(μ_x, σ_x).	$\mu_x = \mu_y$	$\mu_x < \mu_y$		$z < -z_\alpha$	P(Z < z)
$y_1, \ldots, y_n$ are N(μ_y, σ_y).	$\mu_x = \mu_y$	$\mu_x \neq \mu_y$		$\lvert z\rvert > z_{\alpha/2}$	2 P(Z > $\lvert z\rvert$)
§9.2: Common σ unknown	$\mu_x = \mu_y$	$\mu_x > \mu_y$	$t = \frac{\overline{x}-\overline{y}}{s_p\sqrt{\frac{1}{m}+\frac{1}{n}}}$	$t > t_\alpha$(m+n-2)	P(T > t)
$x_1, \ldots, x_m$ are N(μ_x, σ).	$\mu_x = \mu_y$	$\mu_x < \mu_y$		$t < -t_\alpha$(m+n-2)	P(T < t)
$y_1, \ldots, y_n$ are N(μ_y, σ). Find $\overline{x}, \overline{y}, s_x^2, s_y^2$ and $s_p = \sqrt{\frac{(m-1)s_x^2+(n-1)s_y^2}{m+n-2}}$	$\mu_x = \mu_y$	$\mu_x \neq \mu_y$		$\lvert t\rvert > t_{\alpha/2}$(m+n-2)	2 P(T > $\lvert t\rvert$)
§9.2: σ_x, σ_y unknown	$\mu_x = \mu_y$	$\mu_x > \mu_y$	$t = \frac{\overline{x}-\overline{y}}{\sqrt{\frac{s_x^2}{m}+\frac{s_y^2}{n}}}$	$t > t_\alpha(k)$ k=min(m-1, n-1)	P(T > t)
$x_1, \ldots, x_m$ are N(μ_x, σ_x).	$\mu_x = \mu_y$	$\mu_x < \mu_y$		$t < -t_\alpha(k)$	P(T < t)
$y_1, \ldots, y_n$ are N(μ_y, σ_y). Find $\overline{x}, \overline{y}, s_x^2, s_y^2$	$\mu_x = \mu_y$	$\mu_x \neq \mu_y$		$\lvert t\rvert > t_{\alpha/2}(k)$	2 P(T > $\lvert t\rvert$)
§9.4: Comparing p_1, p_2:	$p_1 = p_2$	$p_1 > p_2$	$z = \frac{(\hat{p}_1-\hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{m}+\frac{1}{n})}}$	$z > z_\alpha$	P(Z > z)
Find $\hat{p}_1 = \frac{x_1}{m}$, $\hat{p}_2 = \frac{x_2}{n}$ and $\hat{p} = \frac{x_1+x_2}{m+n}$.	$p_1 = p_2$	$p_1 < p_2$		$z < -z_\alpha$	P(Z < z)
m, n large	$p_1 = p_2$	$p_1 \neq p_2$		$\lvert z\rvert > z_{\alpha/2}$	2 P(Z > $\lvert z\rvert$)

Index