Statistics and Data Analysis

A. Abebe, J. Daniels, J. W. McKean
Department of Statistics
Western Michigan University

and

J. A. Kapenga
Department of Computer Science
Western Michigan University

Statistical Computation Lab (SCL)
Western Michigan University
Kalamazoo, MI.

Contents

Preface

This is the text book for Stat 160, *Statistics and Data Analysis*, which is a course offered by the Department of Statistics at Western Michigan University. Stat 160 is a first course in Statistics. It is an online course; hence, for a full idea of what the course is about, the reader is invited to go to the web page,

www.stat.wmich.edu/s160

which is produced and maintained by the Statistical Computation Lab (SCL) at Western. The main purpose of Stat 160 is to present some of the main concepts in Statistics and Probability and to show students how useful statistics is in their world. This course will give the students a basic understanding of statistics, which will enable them to see how statistics can help them to make better decisions as consumers, students, parents or professionals. The course is not designed to be a statistical methods course.

This text book is the online text for Stat 160. It covers basic descriptive statistical and graphical procedures for analyzing data sets. Some simple inferential procedures, parametric and nonparametric, will also be taught. These procedures will enable students to explore data sets. This book is not an introductory statistical methods book, but knowledge of its content will help prepare students for a course in methods.

The quantitative prerequisite for this book is essentially high school algebra, Math 110 at Western Michigan University. It does not make use of any higher math and it is not formula driven. In the course, the computer does the heavy computation, not the student.

Chapter 1 covers basic descriptive statistical and graphical procedures for analyzing data sets. This is a long chapter and an important one. Most people at some time will find themselves either trying to explain a data set to others or it will be important for them to understand a description of a data set. The plots discussed in this chapter, such as comparison boxplots and dotplots, are used throughout the book when discussing data sets. Ample discussion is presented on outliers and the concept of robust statistical procedures is introduced. Robustness is used throughout the text.

Chapters 2 through 5 discuss probability and population models. For the most part, our discussion of probability is based on resampling not on formulas. Resampling has become a very powerful tool in statistics where it is often called the bootstrap. Using resampling to solve probability problems, requires the successful modeling of the problem, which in itself requires the understanding

of the problem. Such exercises will serve the student well in his life. Resampling, though, requires coding. We have short circuited this problem by developing general software for many of the resampling situations called for in the book.

Chapter 6 is a short discussion of the Central Limit Theorem which leads directly into Chapter 7 on confidence intervals. We use the percentile bootstrap confidence intervals for many of the confidence intervals discussed in the book.

A discussion of hypotheses testing is presented in Chapter 8 for two sample location problems. The basic method discussed is the two sample Wilcoxon with observed significance levels determined by resampling. This is followed in Chapter 9 on estimation problems for two sample problems.

Chapter 10 presents experimental designs for two sample situations, both completely randomized and paired designs. Tests on hypotheses and estimation (confidence intervals) are discussed for these designs. Chapter 11 discusses regression designs. Both least squares and a robust procedure are presented.

This text could not have been possible without the help of many individuals, too numerous to thank. Certainly students of previous sections of Stat 160 deserve our thanks. Neither the book nor the course would have been possible without the help provided by the Statistical Computation Lab (SCL) at Western Michigan University. Not only have they provided the statistical and computational expertise to develop the statistical software which accompanies the text but they have supported the entire web page development of the course. Thanks also goes to TLT Presidential funding program at Western Michigan University. Their grant to Professors Kapenga and McKean provided Summer (2000) support for the development of the online part of this course. A grant from Sun Microsystems to Professors Kapenga and McKean was also fundamental to the online development as well as the robust content of the course.

The Authors

December 2000
Kalamazoo, MI

Chapter 1

Descriptive Statistics

1.1 Introduction

In this chapter, we discuss describing data sets. Data sets can be thought of as a bunch of numbers or a list of things. For instance, suppose we ask twenty students their weights and then record them as:

```
122   146    65   162   148   155   136   151   151   153
201   156   235   157   160   171   178   197   142   131
```

This is a data set of 20 observations. The number of items in a sample is called the sample size . We often denote the sample size by n. For this data set $n = 20$.

Next suppose we ask the students their hair color and get the responses:

```
Red   Blond Blond Brown Brown Red   Blond Blond Brown
Black Blond Red   Red   Brown Black Brown Red   Black
Brown Blond
```

This is another data set of 20 observations.

Often our data set is a **sample** of observations from some reference . For example, the 20 weights might be sample of the weights of 20 students from a university. We might want to **infer** something about the weights of the population based on this sample. These are problems of statistical inference which we will take up in later chapters. In this chapter, though, we just want to discuss ways for describing data sets.

To begin with, basically data come in two types: **discrete** and **continuous** . Discrete data have natural categories while continuous data do not. The hair color data set is discrete while the weights are continuous. We will treat discrete data first and then continuous data.

1.2 Describing Discrete Data

As we said, discrete data have natural categories. Hence to describe a discrete data set, simply classify the data into their categories. For example, suppose we ask our 20 students their stronger hand; i.e., whether they are left (L) or right (R) handed. The responses are:

```
Hand L  R  R  R  R  L  R  R  R  R
     R  R  R  R  L  R  R  R  R  R
```

Hence this is discrete data with two categories R or L. Classifying the data, we obtain

```
R    L
17   3
```

This is the **distribution** of the data. It is indeed **the** distribution, there is no other.

A picture of the sample distribution is given in Figure 1.1:

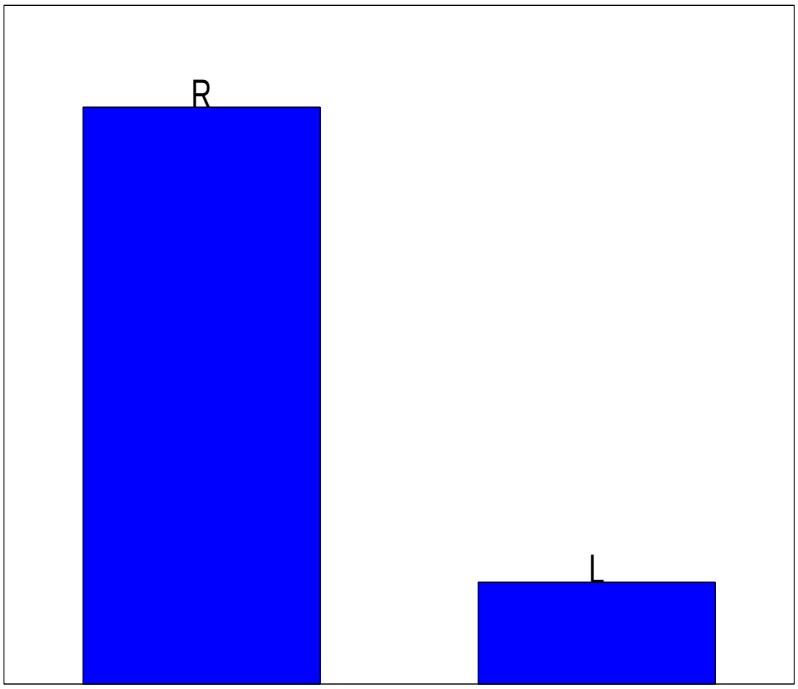
Note how informative this picture is. It tells you immediately that there are many more right-handed people in the sample than left-handed. More than 5 times as many. This picture is much more informative than the 20 L's and R's listed above.

One of interest here is the **sample proportion** of left-handers in the sample which is 3/16 or .1875 (19%). Later in the course, we will discuss how to use the sample proportion to **estimate** the true proportion of left-handed people in the university (population).

Exercise 1.2.1

1. *Obtain the distribution of hair color for the above 20 students. Then draw a histogram of it by hand. Obtain the sample proportion of blonds.*
2. *Obtain the distribution of hair color for the above 20 students using the summary module.*
3. *Sometime ago, Carrie had a deck of 59 baseball cards. The data recorded from this deck is given in Appendix A. The fourth column of this data gives the stronger hand of the baseball player, 0 for right-handed and 1 for left-handed. Obtain the distribution of the strong hand of a baseball player and obtain a histogram of it by hand.*
4. *Repeat the exercise using the summary module to draw the histogram.*
5. *Note that about 11% of the males in America are left-handed. Obtain the sample proportion of left-handed baseball players. Does it seem high compared to 11%? If so, can you think of a reason why it would be high?*

Figure 1.1: Historam of Strong Hand



6. Obtain the distribution and obtain the proportion of ones in the following sample.

```
Data
1 1 1 1 0 0 0 1 1 0
1 0 0 0 1 0 0 0 1 0
0 0 1 1 1 0 0 1 1 1
0 0 0 1 0 1 0 0 1 0
```

1.3 Sample Distributions for Continuous Data

Continuous data are data without natural categories. These are usually measurements such as height, weight, age, temperature, or cholesterol. For weight one might think that 200 is a natural category, but in kilograms 200 pounds is 90.8 KG which is not even an integer. Because we can not measure infinitely precise, measurements are approximations.

Example: Here is a sample of head sizes (maximum measurement across the top of the skull in mm) of 25 Etruscans. This data was taken form the data set Etruscan-Italian head sizes data set given in Appendix A.

```
141   148   132   138   154   142   150   146   155   158   150
140   147   148   144   150   149   145   149   158   143   141
144   144   126
```

So what do we need? A picture, of course. The above picture for the discrete data is a nice visual summary. So we need a sampling distribution of these numbers. Since continuous data have no natural categories we have to create some categories. This results in **a** sample distribution. If we create other categories we will get a different picture. We need a way of creating these pictures fast so that if we don't like a picture we make another one. We will do this with a **stem leaf plot** . The categories are the stems. For instance, suppose for the Etruscan data we choose the interval 120-129 as our first category. Every measurement that falls into this interval has the same first two digits, namely, 12. This is called the stem for the class. The remaining digits of a measurement is called the leaf. For example, the skull size 126 falls into this class; so 126's stem is 12 and its leaf is 6. For a stem-leaf plot we simply put the leaves on the stem. All that is lost (except for possible rounding) is the order of recording of the data which may be important in some applications but it is not in this case. A stem leaf plot of the above data set is:

```
12 6
13 28
14 182607849593144
15 4058008
```

Do you like the picture? No, neither do I. The numbers are too bunched up. We need more categories (stems). But this is easy to do with stem-leaf plots. Lets split each stem into two. In this case the leaves 0 through 4 go on the lower stem while the leaves 5 through 9 go on the upper stem. The picture is

```
12 6
13 2
13 8
14 12043144
14 8678959
15 4000
15 588
```

This picture is better than the first. I wouldn't split the stems again. Although we only have 25 numbers here, certainly the picture is much more informative than looking at the above string of numbers. We can see immediately that in this sample most Etruscans have head sizes between 140-150 mm and there are a few with smaller head sizes.
Note that a stem-leaf plot is also a **histogram** . Technically the histogram is just the picture (not the leaves). We will often use histograms in this class.

Exercise 1.3.1

1. *Consider again Carrie's baseball data given in Appendix A. Glance through the weights (second column) the baseball players. What does a typical baseball player weigh? Do more baseball players weigh over 200 pounds than under 170?*

2. *Obtain a stem-leaf plot of the weights of the baseball players. Now answer the questions in the last problem. For your stem-leaf plot, should the stems be split or grouped together?*

3. *The typical American male weighs about 170-175 pounds. Based on your stem-leaf plot, how would you compare the weights of baseball players to typical American males?*

4. *The typical American male height is 70 inches. What about the heights of baseball players? Base your answer on a stem-leaf plot of the baseball players' heights.*

5. *Obtain a stem-leaf plot of the following using the summary module.*

```
Data
14   117    77   81   205   21   22   157   134   69
193    8   162    0   156  194   17   100    50   53
235   29   191   81   167   29  158   105   171    2
8     89    82   11   247  149  106    61    18  172
```

Try the same example data (given below). Choose stemleaf in the **summary** module after entering the data.

```
12 18 25 15 9 14 21 25 28 125
```

We need a little on **shapes of distributions** so that we can discuss them. We will classify distributions as **symmetric** or **asymmetric** . Symmetric distributions are (approximately if it's a sample distribution) symmetric about a point on the data axis. An example of a symmetric sample distribution is given by:

```
Low:   49

    6 : 4
    6 : 78
    7 : 14
    7 : 556788
    8 : 0122334
    8 : 67799
    9 : 01122223334
    9 : 5555666677788889999
   10 : 000000001122223444
   10 : 568889
   11 : 000001134
   11 : 599
   12 : 123
   12 : 89
   13 : 2
   13 : 56
   14 : 0

High:  161
```

To avoid many empty stems on the ends of stem-leaf plots sometimes, as in the above plot, the low and the high points are just indicated, as 49 and 161 are here. The point of symmetry in this plot is close to 95.
The above plot is unimodal , a single **mode** or peak. Around 95. Here's the stem-leaf plot of a data set which is **bi-modal** , two peaks, and which is symmetric:

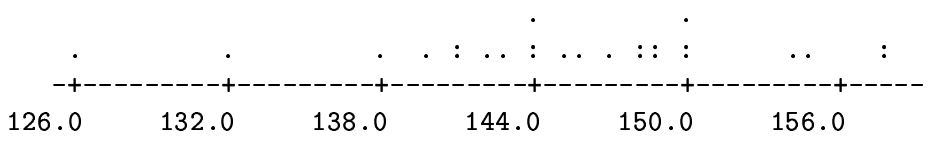
```
  -1 : 2
  -0 :
   0 : 2
   1 : 5
   2 : 5669
   3 : 1125677
   4 : 223445556699
   5 : 011123334444566667788888999
   6 : 00112224444444555555567788899
   7 : 014445566778899
   8 : 0122334455677799
   9 : 011122223334455556666777788889999
  10 : 000000001122223444568889
  11 : 0000011123334599
  12 : 0122389
  13 : 256
  14 : 0
  15 :
  16 : 1
```

A distribution is a **asymmetric** if it is not symmetric. One class of asymmetric distributions of interest is the class of **skewed** distributions. These either have a long tail to the right or to the left. For example, this is a right skewed sample distribution.

```
  0 : 1223444444
  0 : 556666667777777888999
  1 : 0011111123333344
  1 : 5555566788889999
  2 : 011222333334
  2 : 56666789999
  3 : 0114
  3 : 668
  4 : 02
  4 : 58
  5 : 02

High:  617
```

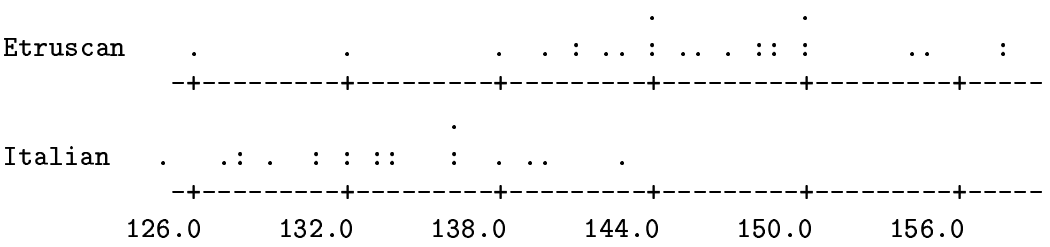
Another plot for continuous data that we will frequently use is the **dotplot** , For a dotplot, simply draw a number line, mark off the range of the data, and then record a dot at the value of each observation. For observations with the same value put the second dot above the first. Here is a dotplot for the Etruscan data.



An interesting application of dotplots concerns **comparison dotplots** of several data sets. Suppose we have several data sets that we want to compare. Simply draw one number line. Then for each sample put a row of dots corresponding to the measurements. For example here are skull measurements of 20 modern Italians taken from the data set Etruscan-Italian head sizes data set.

133	128	136	140	127	136	131	131	128	132	125
133	134	136	134	129	132	139	143	138		

Here is the comparison dotplot between the Italian skull sizes and the above Etruscan skull sizes.



Any conclusions about the Etruscan and Italian skull sizes? It appears that the Etruscans have larger heads than the Italians. As the exercise below shows, this difference occurs when all the data are used. We will discuss inference based on this data set in Chapter 8.

Exercise 1.3.2

1. *Obtain a dot plot of the weights of the twenty students discussed above and listed again as :*

   Weights
   122 146 65 162 148 155 136 151 151 153
   201 156 235 157 160 171 178 197 142 131

2. *For Carrie's baseball data, obtain comparison dotplots of the batting averages (6th column of the data for the hitters only, (signified by a 1 in the 5th column)) by the side of the plate they hit from R, L or Switch, signified by a 1, 2 or 3 in column 3.*

3. *Obtain comparison dotplots of the the Etruscan and Italian data given in Appendix A. Note that the Etruscans formed an ancient civilization in Truscany, Northern Italy, that predated the Romans. There is some question as to where the Etruscans came from. Were they native to Italy or not? Draw conclusions about this mystery based on the comparison dotplots.*

4. *Obtain stem-leaf plots and comparison dotplots for the following 3 samples. Comment on the shape of each.*

```
Sample 1

      76 183 125  24  8 59  25 179  29 101
      55 108  68 128  5 12  35  25 122  39
      59  91  90  81 66 20 178 111 186  26
      5  123 124  45 13 79 158  20  92  23

  Sample 2

      66   9 62 21 11 39 21  24  21  19
      67  71 67  0  4 82 32  91 152 124
      20 108  5 63  1 10 23 125  59  25

  Sample 3

      59 54 19 79 22 81 18 67 61 53
      71 14 10 87 76 49 21 16 35 11
       7 77 90  6 79 55 83 28 11 60
      55 43  9 65 25
```

1.4 The 5 Basic Descriptive Statistics for Continuous Data

Stem-leaf plots and histograms are useful descriptions of a sample but often we want to describe samples or compare samples with a few descriptive statistics. The statistics we discuss next are commonly called the **Five Basic Descriptive Statistics** . First, alas, we need a little notation. We will be discussing samples throughout this course and we need to often call the items something. So for a generic sample of size n lets use

$$x_1, x_2, ..., x_i, ..., x_n$$

where

x_1 denotes the first item (measurement) in the sample,
x_2 denotes the second item (measurement) in the sample,
$\vdots$
x_i denotes the ith item (measurement) in the sample,
$\vdots$
x_n denotes the nth item (measurement) in the sample,

Using this notation we can now define the 5 basic descriptive statistics. We will illustrate these statistics with the sample of *n=25* Etruscan skull sizes, given above, but repeated for convenience:

```
126   132   138   140   141   141   142   143   144   144   144
145   146   147   148   148   149   149   150   150   150   154
155   158   158
```

Note that we have ordered the data from low to high. If you had to choose some numbers to describe this data set, probably the first two you would pick are the **Minimum** and the **Maximum** . The **minimum** of a sample is the smallest measurement, i.e. the first ordered data point. We will denote the minimum by min. For the Etruscan data set min = 126mm. The **maximum** of a sample is the largest measurement, i.e. the nth ordered data point. We will denote the maximum by max. For the Etruscan data set max = 158mm. The min to the max is the range of the data. In fact we call their difference the **range** . For the Etruscan data the range is 158 - 126 = 32 mm. The range is a **measure of scale** , or **dispersion** or **noise**. The range is extremely sensitive to **s**. Outliers are points that are far from the rest of the data. We will formally define "outlier" in the next section. A statistic is said to be **robust** if it is not sensitive to outliers. So the minimum, maximum, and range are not robust statistics.

Now that we have the range of the data, the next statistic is a measure of the center of the sample. We will use the **median**. The median is the middle ordered data point if the sample size

is an odd number and the average of the middle ordered data points if the the sample size is even. 50% of the data is less than or equal to the median and 50% of the data is greater than or equal to the median. For the Etruscan data, upon ordering the data we get,

```
126   132   138   140   141   141   142   143   144   144   144
145   146   147   148   148   149   149   150   150   150   154
155   158   158
```

Since n is 25, $(n+1)/2$ is 13 and, hence, the median is the $13th$ order data point or 146 mm. We shall use Q_2 to denote the median. So half of the Etruscans in the sample had a skull size less than or equal to 146mm and half of the Etruscans had a skull size greater than or equal to 146mm. The median is a **measure of center**. The median is very robust. Half the data would have to change for the median to change.

We now have the range of the data and a measure of the center. How about the middle 50%? This goes from the **First Quartile** to **Third Quartile**. The **first quartile** is the median of the first half of the data. We will denote it by . 25% of the data is less than or equal to the first quartile and 75% of the data is greater than or equal to the first quartile. There are many rules for finding Q_1. In this class we will be using the computer for large data sets and the computer (the statistical software) will compute Q_1. For class and tests lets use a very simple rule. To find the ordered data point, divide n by 4. If the result is an integer use that integer to pick out the ordered data point corresponding to that integer. If the result is a fraction round up to the nearest integer. Pick out the ordered data point corresponding to this integer. For the Etruscan data, 25/4 is 6.25; hence, we round up to 7. The 7th ordered data point is 142, so $Q_1 = 142$mm for the Etruscan data set. The first quartile is a robust statistic.

The **third quartile** is the median of the second half of the data. We will denote it by Q_3. 75% of the data is less than or equal to the third quartile and 25% of the data is greater than or equal to the third quartile. There are many rules for finding Q_3. To find Q_3 by hand, just use the integer we found for the first quartile, but this time count through the data from the high measurements to the low measurements. Hence $Q_3 = 150mm$ (several are tied at 150 but the 8th point from the top in my counting was 149). The third quartile is robust.

The difference between the quartiles is called the **interquartile range** of the sample. It is denoted by **IQR**, so for the Etruscan data $IQR = 150 - 142 = 8mm$. IQR is also a **measure of scale**. It is not sensitive to outliers (25% of the data have to be outliers to affect IQR); hence, IQR is robust.

In summary, for the Etruscan data, the five basic descriptive statistics are: 126, 142, 146, 150 and 158mm. We want to put these summary statistics in a picture but first we need the concept of an outlier, which we will do in the next section.

Lets do one more example which shows how we can get **very quickly** the 5 basic descriptive statistics from a stem leaf plot. Consider the subsample of Italian skull sizes given by,

```
133    128    136    140    127    136    131    131    128    132    125
133    134    136    134    129    132    139    143    138
```

The stem leaf plot is

```
 Stem  Leaves     f   F  FTB
  12   87859      5   5
  13   31123442   8  13
  13   66698      5  18   7
  14   03         2  20   2
```

We have added three columns on the right side of the stem-leaf plot. The column labeled f is the frequency of the class, the column labeled F is the cumulative frequency of the class (the number of data points down through the end of the class), and the column labeled FTB is the cumulative frequency of the class from large numbers to small (the number of data points down through the beginning of the class). Based on this plot and those columns the 5 basic descriptive statistics are a cinch. The minimum is 125 and the maximum is 143.

The sample size is 20 (last number in column F) which is even. So the median is the average of $10th$ $(n/2)$, and the $11th$ ordered data points. To get these look at column F. There are 5 data points down through the end of the first class and there are 13 data points down through the end of the second class; hence, the median must occur in the second class. In the second class the $6th$ through $11th$ ordered data points are 131, 131, 132, 132, 133, 133. Thus the median is $.5(133 + 133) = 133$.

Since 20/4 is 5, the first quartile is the $5th$ ordered data point which is 129 (The largest data point in the first class as dictated by column F). The third quartile is the $5th$ ordered data point from the top to the bottom. By the FTB column its in the second class from the top. The $7th$ ordered (from top to bottom) is 136, the $6th$ is 136, and the $5th$ is 136. So $Q_3 = 136$.

For small data sets we can get these statistics by hand. But for large data sets it is best if the computer gets them for us.

To run the class code for the descriptive statistics choose the summary module of the and choose **Numerical Summaries** after entering the data.

```
12 18 25 15 9 14 21 25 28 125
```

TRY IT!

1.5 Outliers and Box Plots

Frequently the most interesting points of a data set are the points that do not seem to belong; i.e., they seem to differ by a substantial amount from the rest of the data. We call these points **outliers** . These are often points worthy of investigation in order to understand why they differ. Such points can lead to significant discoveries.

For example, each year satellites measure the ozone level over Antarctica. In the early 1980s, however, scientists were so astounded in detecting a dramatical seasonal drop in ozone levels over Antarctica by a fly over that they spent two years rechecking their satellite data. They discovered that satellites had dutifully been recording the ozone collapse but **the computers had not raised an alert because they were programmed to reject such extreme data as anomalies**; see R. Benedick, *Scientific American*, April 1992. This discovery of the drop in ozone levels has had profound influences on manufacturing and society. If the computer had been programmed correctly it would have flagged the outliers and, hence, alerted the scientists to investigate the outliers on the first occasion. Changes in manufacturing could have been made much sooner.

We have chosen the following simple rule for determining when a point is labeled an outlier: First determine the quartiles Q_1 and Q_3. Recall that the interquartile range, $IQR = Q_3 - Q_1$, is a measure of noise or scale for the data set. Points that are beyond the quartiles by one-and-a-half IQR's will be deemed potential **outliers**. I know what you are asking (you are so inquisitive), why this rule? Stay tuned for Chapter 5 when an answer will be provided.

In order to set up a formal mechanism, denote the above distance by h; i.e,

$$h = 1.5IQR = 1.5(Q_3 - Q_1)$$

Next, denote the **lower** and **upper inner fences** by

$$LIF = Q_1 - h,$$

$$UIF = Q_3 + h$$

Hence points beyond these fences are potential **outliers**. Those points of the data set which are closest to the fences but still inside the fences are called the **adjacent points**. There are two adjacent points in a data set, the lower adjacent point (the point inside the fences but closest to LIF) and the higher adjacent point (the point inside the fences but closest to UIF).

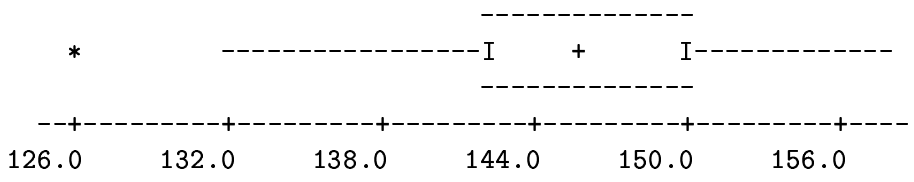
We now have the ingredients to draw a **boxplot** of a data set. This is an easily drawn schematic of the data set which displays the five basic descriptive statistics and the outliers, if there are any. Simply draw a number line, as you did in the dotplot. Find the quartiles on the number line and draw a rectangle above the number line which encloses the quartiles; i.e, this **box** encloses the

middle 50% of the data. Find the median on the number line and place a + in the box above the median. Next find the fences and adjacent points. Draw lines from the ends of the box to the adjacent points. Finally, indicate the outliers by $*$'s.

Sounds horrible, right! Now wait a minute. Consider the sample of $n = 25$ Etruscan skull sizes, given above:

```
126    132    138    140    141    141    142    143    144    144    144
145    146    147    148    148    149    149    150    150    150    154
155    158    158
```

Recall that the quartiles are 142 and 150 and the median is 146. Hence, $h = 1.5(150 - 142) = 12$. Thus the fences are: $LIF = 142 - 12 = 130$ and $UIF = 150 + 12 = 162$. Therefore the adjacent points are 132 and 158 and the point 126 is an outlier. So the boxplot is:

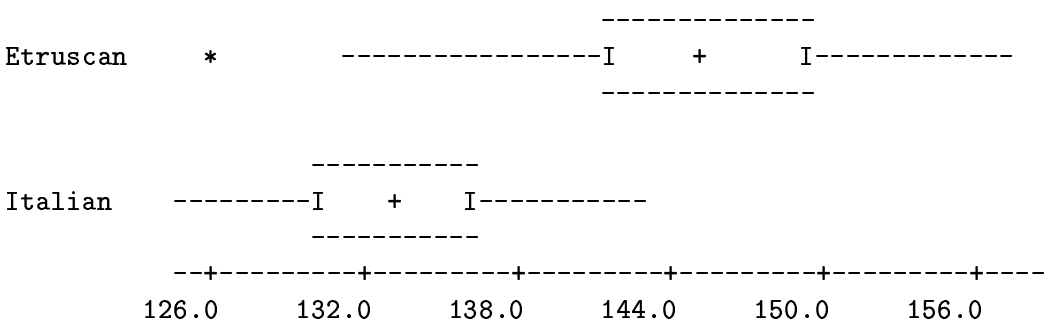


1.6 Comparing Data Sets

As with dotplots, boxplots lend themselves to comparisons. Just make sure that the same number scale is used for each boxplot. Then simply draw the boxplots in rows (or in columns). As an example, reconsider he subsample of Italian skull sizes given by,

```
133    128    136    140    127    136    131    131    128    132    125
133    134    136    134    129    132    139    143    138
```

Recall that the 5 basic descriptive statistics are: 123, 129, 133, 136, and 143. Hence, $h = .5(136 - 129) = 10.5$ and the fences are $LIF = 129 - 10.5 = 118.5$ and $UIF = 136 + 10.5 = 146.5$. The adjacent points are 125 and 143. Based on these statistics, the comparison boxplots are:



Use the summary module to obtain:

1. a boxplot for the example data given here. Enter the data in the first "DATA SETS" window.

   ```
   126    132    138    140
   141    141    142    143
   144    144    144
   ```

2. a comparison boxplot for the above and the following data sets.

   ```
   123   324   145   156   265
   143   221   322   133   233
   142   144   244
   ```

A final remark on this example is in order. Notice that the scales (noise levels) in the data sets are about same; i.e., the interquartiles ranges are about the same, 8 and 7, and ignoring the outlier the ranges are about the same. We do not have much data here to comment on the shapes of the distributions but based on the comparison dotplots above symmetry cannot be discounted.

In light of this, what catches your eye as you look at the box plots? There is a **shift**; that is, the Etruscan data is shifted up from the Italian data. If you draw lines connecting the Etruscan and Italian lower quartiles and then a line connecting their upper quartiles the lines will be almost parallel. The line connecting the medians will also be almost parallel with these lines. In fact, it is tempting to summarize the data with one number which is the difference in the medians. In this case the difference is $146 - 133 = 13$. This is called a **location problem**. These problems are characterized by the samples having similar shapes and scales (noise levels). In such cases, a convenient summary is a difference in locations or centers. Here, that difference is 13; so the Etruscan head sizes are shifted up 13mm from the Italian head sizes. Be very careful, though. This number 13 is based on just two samples. We also need a measure of sample error. If this measure turns out to be greater than 13 then our estimate of shift loses a lot of meaning. In later chapters we will say it is insignificant. If sampling error is small (less than 13 here) then our estimate of shift is meaningful. In later chapters we will say it is significant.

Exercise 1.6.1

1. *A standardized exam was given to two groups of people. The first group took the exam under adverse conditions, (room was too cold, room was dirty, proctor swore at them) while the second group took it under normal conditions. The data are given below. Determine the five basic statistics for the two groups, find the fences, and determine if there are any outliers. Then draw comparison boxplots for the two data sets. Are there any location differences? Scale differences?*

Group 1:	153	150	132	123	148	146	140	154
	137	112						
Group 2:	148	113	69	129	150	129	157	184
	143	167	141	179	124	130	166	

2. *Consider Carrie's baseball data. Obtain back-to-back stem-leaf plots of the height of the hitters and pitchers. Discuss the plots.*

3. *In the last problem, obtain the 5 basic descriptive statistics for the heights of the hitters and pitchers. Obtain the fences, and determine if there are any outliers. Then draw comparison boxplots for the two data sets. Are there any location differences? Scale differences?*

4. *Ten batteries from each of three brands (A, B, and C) were put on test to determine their lifetimes (in hours). Obtain comparison dotplots. Use these dotplots to obtain the 5 basic descriptive statistics for each brand. Bigger means better here. Which brand seems best, if any?*

```
A:    41    289    214    102     38
      94    179     87    116    155

B:    39     65     22     64     22
     191     99     32    142    317

C:    24     95    139    122     41
     360    318     34     43     18
```

1.7 Other Statistics

Although the five basic descriptive statistics go a long way in describing data sets, we will make use of other statistics throughout this course. We can roughly classify them into three broad categories: Measures of Center, Measures of Scale or Noise , Measures of Relationships. In this section we will consider two classes and discuss measures of relationships in the next section.

Denote our generic sample by

$$x_1, x_2, ..., x_n$$

1.7.1 Measures of Center

The **sample median**, Q_2, is one measure of center which we have already discussed. Recall that 50% of the data is less than or equal to Q_2 and 50% of the data is greater than or equal to Q_2. Another measure of center that we will frequently use is the **sample mean**, which is just the arithmetic average of the sample; i.e., add up all the data and divide by the sample size n. In terms of notation we will use $\bar{x}$ to denote the average of $x_1, x_2, ..., x_n$. For example, consider the data (Set 1):

```
Set 1:    11    18     6     4     8    15    22
```

The median is 11. The data add up to 84 and their are 7 data points; hence, the sample mean is $84/7 = 12$. You can use the summary module to obtain the sample mean.

What does the mean mean? The mean is the center of gravity of a histogram of the sample along its horizontal axis. Consider, yet again, the 25 Etruscan skull sizes:

```
126    132    138    140    141    141    142    143    144    144    144
145    146    147    148    148    149    149    150    150    150    154
155    158    158
```

Again enter these data into the data box and choose summary from the analysis menu. The sample average is 145.68 while the median is 146. To get a histogram of the data just click on the histogram button before submitting. The histogram is approximately symmetric so it is not surprising that the mean and the median are similar. But for data sets which are asymmetric these statistics can be quite different.

Furthermore the mean is quite sensitive to outliers. Consider again the simple data set: 11, 18, 6, 4, 8, 15, 22. The median and mean are 11 and 12, respectively. Both statistics are in the center of the data which is where they should be since they are measures of center. Now suppose instead of 22 the last data point is 72. The median of course does not change but the mean is now 19.14. That is, the median is still in the center of the data but the mean has moved beyond the sixth data point, 18. The mean is no longer measuring the center of the data. If the last data point is

220 instead of 22 the mean changes to 40.3, well beyond the center of the data. Below is a table of data sets. The first row is the original set and the subsequent rows are with changed data points for the data point 22. Another statistic given in the table is s which we will discuss later.

	Data						median	mean	IQ	s
Set 1:	11	18	6	4	8	15	22	11	12	12	6.61
Set 2:	11	18	6	4	8	15	72	11	19.1	12	23.8
Set 3:	11	18	6	4	8	15	720	11	112	12	268
Set 4:	11	18	6	4	8	15	2200	11	323	12	828
Set 5:	11	18	6	4	8	15	7200	11	1037	12	2717
Set 6:	11	18	6	4	8	15	72000	11	10295	12	27210

Thus the mean is very sensitive to outliers while the median is not. Hence the mean is not a robust statistic.

Another measure of center which we use occasionally is the median of all the pairwise averages of the data. For the simple data set 11, 18, 6, 4, 8, 15, 22, just order the data and make a table with rows and columns labeled by these data points. Then just compute the average of the pairs associated with row and column elements. This is shown in the table below. These pairwise averages are called **Walsh Averages**. For a pair of data points we only compute the average once; hence, we only need the top half as shown.

	4	6	8	11	15	18	22
4	4	5	6.5	7.5	9.5	11	13
6		6	7	8.5	10.5	12	14
8			8	9.5	11.5	13	15
11				11	13	14.5	16.5
15					15	16.5	18.8
18						18	20
22							22

Ignore the row and column labels and compute median of the other entries in the table. There are 28 entries in the table so the median is the average of the $14th$ and $15th$ entries; i.e, the average of 11.5 and 12 which is 11.75. This estimate is often called the Hodges-Lehmann estimate so we will denote it by HL. Okay. I realize it is not fun to compute this table so you can also do it the easy way. Just enter these data into the data box, choose **summary** from the analysis menu and check the button for **numerical summaries** after submitting. As you see $HL = 11.75$.

The Hodges-Lehmann estimate is robust. If you change the last data point to 72000, the Hodges-Lehmann estimate remains at 11.75.

1.7.2 Measures of Scale or Noise

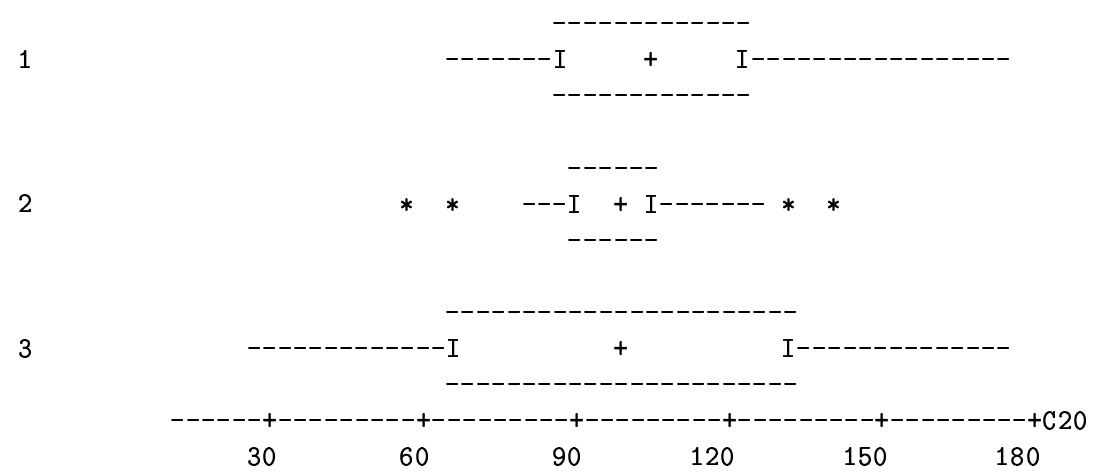
Motivation

It is easy to think of many data sets which have the same center but are quite different otherwise. For example, consider the following three data sets placed in columns 2 through 4 of the table:

	Samp.1	Samp.2	Samp.3
1	88	119	91
2	166	116	98
3	143	92	117
4	110	94	62
5	86	86	51
6	108	81	40
7	133	133	57
8	105	65	74
9	114	82	65
10	126	90	60
11	87	86	26
12	99	98	81
13	72	58	133
14	98	106	174
15	73	99	134
16	137	102	120
17	109	93	119
18	82	101	171
19	122	100	132
20	174	101	88
21	65	126	154
22	99	103	154
23	109	142	94
24	105	103	121
25	79	105	131
Median	105	100	98
Mean	108	99	102

Based on the sample medians and means (last two rows of the table), the center estimates are fairly similar for the data sets, considering the noise level. So if we would only estimate center it would

be hard to tell these data sets apart. But in this class **PLOT DATA** is a must! Comparison boxplots yield:



By the length of the boxes (i.e. interquartile ranges), we see that the noise levels are quite different in the data sets. Sample 3 seems to be twice as noisy as Sample 2 and Sample 2 seems to be twice as noisy as Sample 1. So along with measures of center we need measures of noise. For the third data set the boxplot misses something very important. From the stem leaf plot the data appears to be bimodal. The other two data sets appear to be unimodal.

```
Stem-and-leaf of Sample 1       N  = 25
Leaf Unit = 1.0


    1    6 5
    4    7 239
    8    8 2678
   11    9 899
   (5)  10 55899
    9   11 04
    7   12 26
    5   13 37
    3   14 3
    2   15
    2   16 6
    1   17 4
```

```
Stem-and-leaf of Sample 2       N  = 25
Leaf Unit = 1.0


    1    5 8
    2    6 5
    2    7
    6    8 1266
   12    9 023489
  (8)   10 01123356
    5   11 69
    3   12 6
    2   13 3
    1   14 2


Stem-and-leaf of Sample 3       N  = 25
Leaf Unit = 10


    1    0 3
    3    0 45
    8    0 66677
   12    0 8999
  (1)    1 0
   12    1 22223333
    4    1 55
    2    1 77
```

Again: you must **PLOT the data** and it is best to **use several different different types of plots**. What do the comparison boxplots tell you (5 extra brownie points)?

Measures of Scale

The **range** and the **interquartile range**, IQR, are measures of scale. The range is of course not robust but the interquartile range is. For our three data sets the interquartile ranges are: 37.5, 17.5, and 68, respectively for Sample 1, 2 and 3. The ratios agree with our quick glance at the boxplots above.

We need to discuss an estimate of scale that we use in conjunction with the mean. It is a measure of deviation from the mean. For instance, the value $x_1 - \bar{x}$ is the deviation of the first point from the mean. Hence, we have the n deviations:

$$x_1 - \bar{x}, x_2 - \bar{x}, ..., x_n - \bar{x}$$

It does not matter here whether the deviation is negative or positive. One way to get rid of the sign is to square the deviation. But we still have n squared deviations. So we will take the average of these squared deviations, except we will divide by n - *1* and not n. The resulting statistic is called the **sample variance** and we usually use the symbol s^2 to represent it. However, the units of s^2 are squared units. For example if we are data consists of the weights in pounds of individuals then s^2 will be in pounds squared. We rectify this by taking the square root and we call the resulting statistic the **sample standard deviation**, s. In notation we have

$$s = \sqrt{\frac{\text{Sum}(x - \bar{x})^2}{(n-1)}}$$

Lets use the simple data set 11, 18, 6, 4, 8, 15, 22, for an example. The sample mean is 12, hence the deviations are -1, 6, -6, -8 3, and 10. The squared deviations are 1, 36, 36, 64, 9 and 100. Thus $s^2 = 246/6 = 41$. So that the sample standard deviation is $s = \sqrt{41} = 6.4$. Of course the easy way to compute is to just enter these data into the data box and choosing summary from the analysis menu. Then check the variable name and the **covariance** button.

The sample standard deviation is not robust, as the table below, on the simple example with changes to the last data point, dramatically shows,

	Data						median	mean	IQ	s
Set 1:	11	18	6	4	8	15	22	11	12	12	6.61
Set 2:	11	18	6	4	8	15	72	11	19.1	12	23.8
Set 3:	11	18	6	4	8	15	720	11	112	12	268
Set 4:	11	18	6	4	8	15	2200	11	323	12	828
Set 5:	11	18	6	4	8	15	7200	11	1037	12	2717
Set 6:	11	18	6	4	8	15	72000	11	10295	12	27210

Even the first change (22 to 72) brings almost a 4 fold increase in noise as measured by s. The interquartile range is robust.
What does s mean? We will answer that later in Chapter 5.

Exercise 1.7.1

1. *Use the summary module to obtain these statistics for the two data sets in #1, Exercise 1.4. Using these statistics, obtain comparison boxplots of the two samples.*

2. *Check the robustness of the statistics in the descriptive statistics command on the following two data sets using the summary module.*

```
Data set 1

     102 131 137 63 42 12 23 49 63 21
      56  68  35 63 62 19 85 38 76 29
      31  16   0  8 47 40  2 44  8 16
       7  43   2 50 22  1 51 34  4 78


Data set 2

     1020 131 137 63 42 12 23 49 63 21
       56  68  35 63 62 19 85 38 76 29
       31  16   0  8 47 40  2 44  8 16
        7  43   2 50 22  1 51 34  4 78
```

   Notice that in the second data set,the 102 was changed to 1020. Which statistics were robust to this change? Which weren't?

3. *Same as the last exercise but change the 1020 to 10200.*

4. *Same as the last exercise but change the 10200 to 102000.*

5. *Did Manuel I shortchange the people by having less silver in in later days mintings? Try to answer this question by comparing the following two data sets (use comparison boxplots). The first data set is the amount of silver (percentage)in Manuel's first minting while the second data set is the amount of silver (percentage) in Manuel's fourth minting.*

```
First:     5.9   6.8    6.4   7.0   6.6   7.7   7.2   6.9   6.2
Fourth     5.3   5.6    5.5   5.1   6.2   5.8   5.8
```

6. *Using the LDL levels of quail a drug compound (call it A) was put on test. In the experiment, 30 quail were randomly chosen and 20 were assigned to a placebo and the other 10 to the treatment using Drug A. The drug was mixed in their food. Other than this, though, the quail were treated the same. At the end of the treament period, the Low Density Lipid levels of the quail were measured and are given below. Here smaller is definitely better. The data are real.*

```
Placebo:  64  49  54  64  97  66  76  44  71  89
          70  72  71  55  60  62  46  77  86  71

 Drug A:  40  31  50  48 152  44  74  38  81  64
```

(a) Obtain comparison dot plots of the data and try to decide if the drug A was effective.

(b) Obtain the descriptive statistics for each data sets. Which (difference in means, difference in medians, difference in HL) seem more appropriate here? Why?

1.8 Relationships Between Variables, Part 1: Linear Models

Often we collect observations from several different variables on a subject. A simple example is a form, such as an application form, which are collected from a group of people. Each item on the form corresponds to a variable. For example, suppose it is a form that upper classmen are filling out at an university. Items might include the college GPA, major, ACT score, high school GPA, high school percentile, weight, height, gender, family income, major, etc. We may want to describe each variable separately using the descriptive statistics and plots that we have discussed, but often we also want to investigate the relationship between the variables. For this example, we might be interested in the relationship between college GPA and high school GPA. In particular, we may want to predict college GPA in terms of high school GPA, high school percentile, ACT score, IQ, etc.

In this section, we shall consider a pair of variables. Other examples, besides those above, are:

1. For example X is the height of a person and Y is his/her weight.
2. For example X is the grade of a student on first test and Y is his/her grade on second test.
3. For example X is the points for of a NFL team and Y is the team's win-loss percentage.
4. For example X is the points against of a NFL team and Y is the team's win-loss percentage.

We are interested in the relationship between X and Y. We may further be interested in predicting one variable in terms of the other. For this prediction problem we will always label the variables so that we are interested in predicting Y in terms of X.

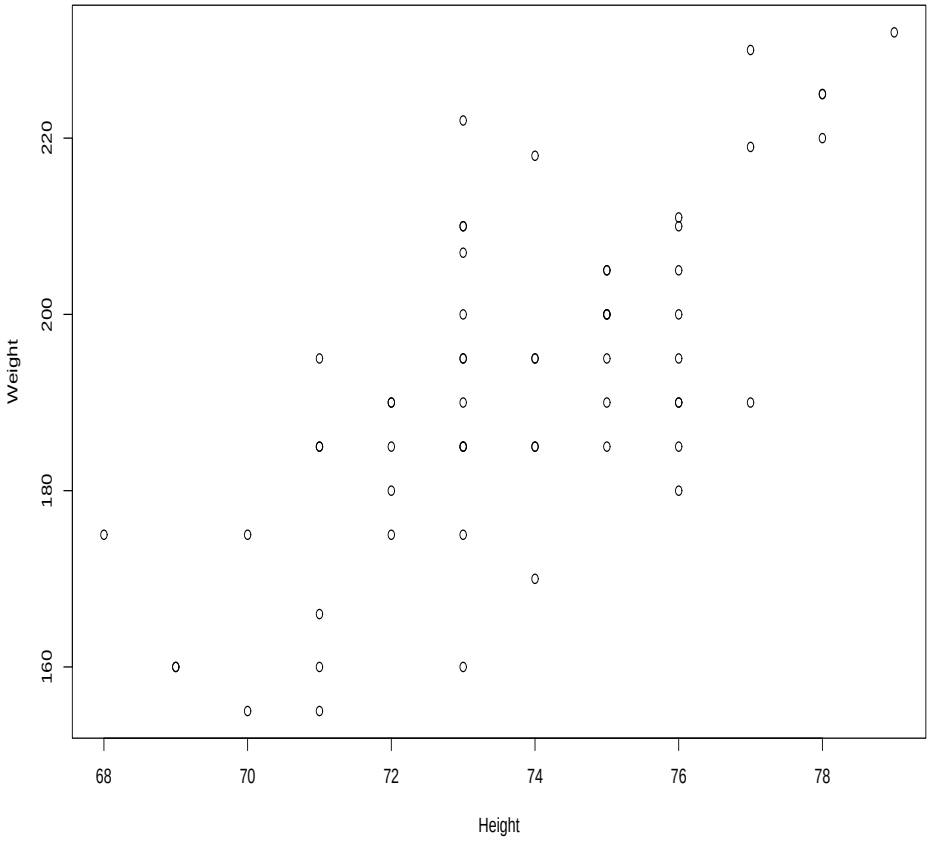
As an example, we will consider the Carrie's baseball data, which is found in Appendix A. The first and second columns of this data contain the heights and weights of the baseball players. Let X denote the height of a ball player and Y denote his weight. Certainly as a first step we plot the data with Y on the vertical axis and X on the horizontal axis. This is called a **scatterplot** of the data. For this data set the scatterplot is given in Figure 1.2

Use the summary module to reproduce the above plot using Carrie's baseball data set.

In order to see if you understand the plot, find the point on the plot corresponding to the ball player who is 5'8" and weighs 175 pounds? Or to the ball player who is 6'3" and weighs about 220 pounds? The relationship between weight and height is increasing, as height increases weight tends to increase too.

Suppose we try to model the data. On the basis of the plot, a linear model is certainly worthy of a first try. Note that the model cannot be deterministic for ballplayers who have the same height have many different weights (in fact a sample of weights). For example, there are at least (some

Figure 1.2: Baseball data : height vs. weight



points overlap) 7 ball players who are 76" tall with weights varying from about 175 pounds to 210 pounds. So the model has to allow for error; i.e, a model of the form:

$$Y = a + bX + e$$

where X denotes the height of a ball player and Y denotes his weight. What are the other parts in the model statement? The variable e denotes **random error**, that is, if there were no error Y would be a deterministic linear function of X. There are two parameters in the model. The most

important parameter is the **slope**, b. It gives the expected change in weight for an increase in 1 inch of height. The **intercept**, a, is the expected weight of a person who is 0" tall. This is absurd. So the intercept in this model has no practical meaning, but we need it to set the line, (there are infinite number of lines with the same slope). We want a model that fits the data well over the range of the X's which in this case is between 68 and 78 inches. The **model is only good where we have data**.

When is a model good? We will discuss this important question in Part 2 of this section. Now we just want to **fit the model**; that is, obtain estimates of a and b. We will first consider a simple **eyeball fit**and then discuss more formal fits.

1. **Eyeball Fit**

   We have selected an easy eyeball method of fit. Pick two points on the plot so that the line passing through them gives a "fairly" good fit. Say the two points are (X_1, Y_1) and (X_2, Y_2). Then an estimate of the slope is

   $$\hat{b} = \frac{Y_2 - Y_1}{X_2 - X_1}$$

   For the baseball data, I chose the points $(69, 160)$ and $(78, 225)$. Hence my estimate of slope is

   $$\hat{b} = \frac{225 - 160}{78 - 69} = \frac{65}{9} = 7.2$$

   Thus I estimate *7.2* more pounds in weight for every inch in height.

   To estimate the intercept, simply take one of the points, say, (X_1, Y_1). Then estimate the intercept by solving the linear equation for a; that is $\hat{a} = Y_1 - \hat{b}X_1$. Based on my first point my estimate is $\hat{a} = 160 - 7.2(69) = -336.800$. Thus we estimate a ball player of 0 height to weigh -336.8 pounds. Actually the newspapers often do this to make fun of scientists. But in this class you know the correct answer to such a farce. Right! **The model is only good where we have data!**. We have no data around *X=0*, so we cannot predict there.

   We thus have our prediction equation,

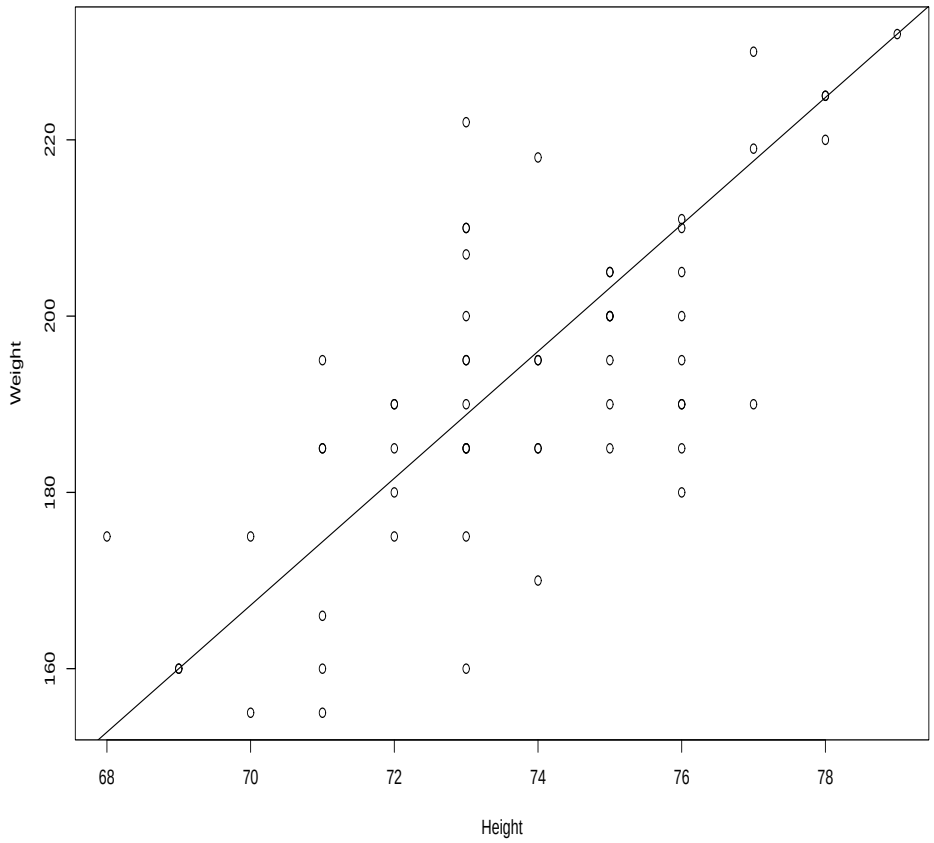
   $$\hat{Y} = -336.8 + 7.2X$$

   Suppose we want to predict the weight of a ball player who is 75" tall. Our prediction is $\hat{Y} = -336.8 + 7.2(75) = 203$ pounds.

   The scatterplot of the data superimposed with our eyeball fit is given in Figure 1.3. Note that you can obtain the predicted value for a given height, say 75", by drawing a vertical line

starting at 75" on the horizontal axis and ending when it intersects our fitted line. Do this to determine the predicted weight of a ball player who is 72" tall.

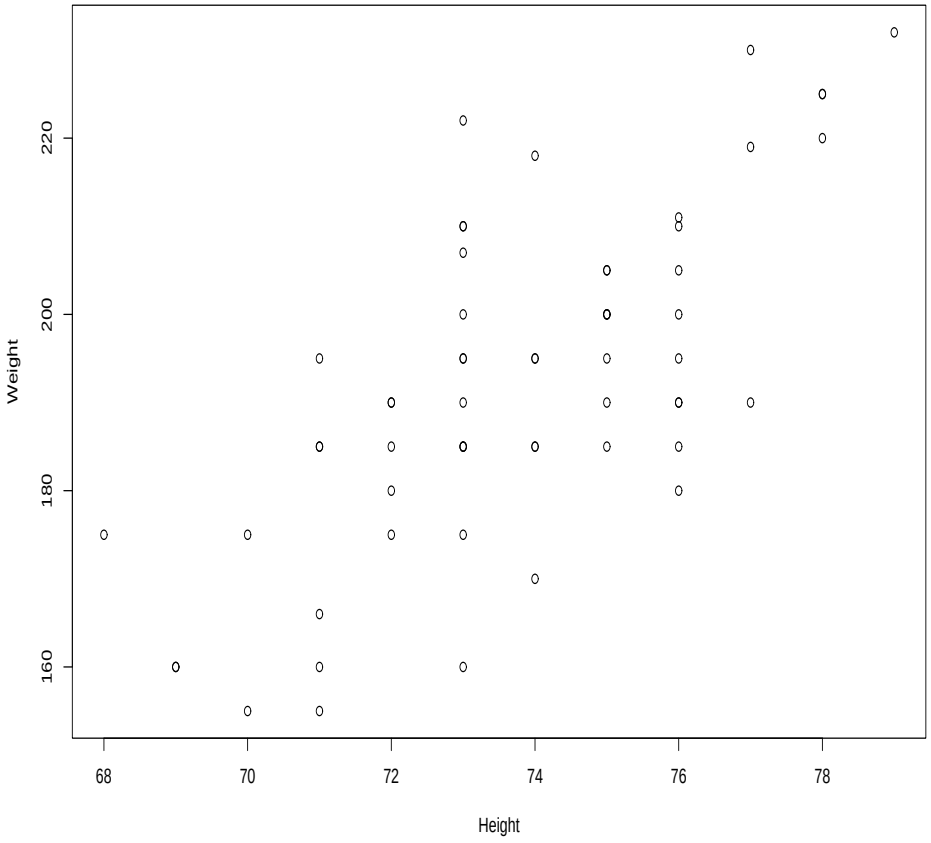
Figure 1.3: Baseball data : eyeball fit



2. **Least Squares Fit**

We will present two methods. The first is the method of **least squares**, which we will often denote by **LS**. Consider again the data set consisting of the weights and heights of baseball players. For convenience the scatter plot of the data is given in Figure 1.4:

Figure 1.4: Baseball data : height vs. weight



Try eyeballing a fit of a straight line on this plot, say, $Y = a + bX$. Consider the point $(77, 190)$ the lowest point at height 77". It probably will not be on your fitted line, so in choosing your line you missed the point by the deviation

$$190 - (a + bX)$$

This deviation is an error determined by the fit. Since two points determine a line, in choosing your fit you will have committed many errors, at least 57 because there are 59 data points).

As a goal in determining the fit, choose the line which minimizes these deviations or errors. It does not matter whether the deviation is positive or negative. The method of least squares minimizes the average of the squared deviations. It does result in equations for estimates of a and b, which we will give below. But at the moment lets just use it. The LS fit is:

```
The regression equation is
 Weight = - 213 + 5.49 Height
```

Hence LS estimates an increase of 5.49 pounds for every inch of increase in height. As an example in terms of prediction, the LS predicted weight of a ball player who is 75" tall is $\hat{Y} = -213 + 5.49(75) = 198.75$ pounds. Of a ball player who is 70" tall is $\hat{Y} = -213 + 5.49(70) = 171.30$ pounds. Locate the points $(75, 198.75)$ and $(70, 171.30)$ on the above plot. Then draw the line determined by those two points. This is the LS fit. It should look like Figure 1.5:

Reproduce the above results using Carrie's baseball data set. Choose regression from the analysis menu after entering the data. Choose weight as a response variable and height as a predictor.

We will make frequent use of the LS fit in later chapters but there is one problem with it. It is not robust. The LS fit is easily distorted by outliers. Lets look at this using the baseball data. Note at height 68" there is one player whose weight is at 175 pounds. Suppose the weight was recorded as 275 pounds. Although high, this weight is not inconceivable for a ball player. The LS fit of this changed data is:

```
The regression equation is
 Weight = - 88.2 + 3.82 Height
```

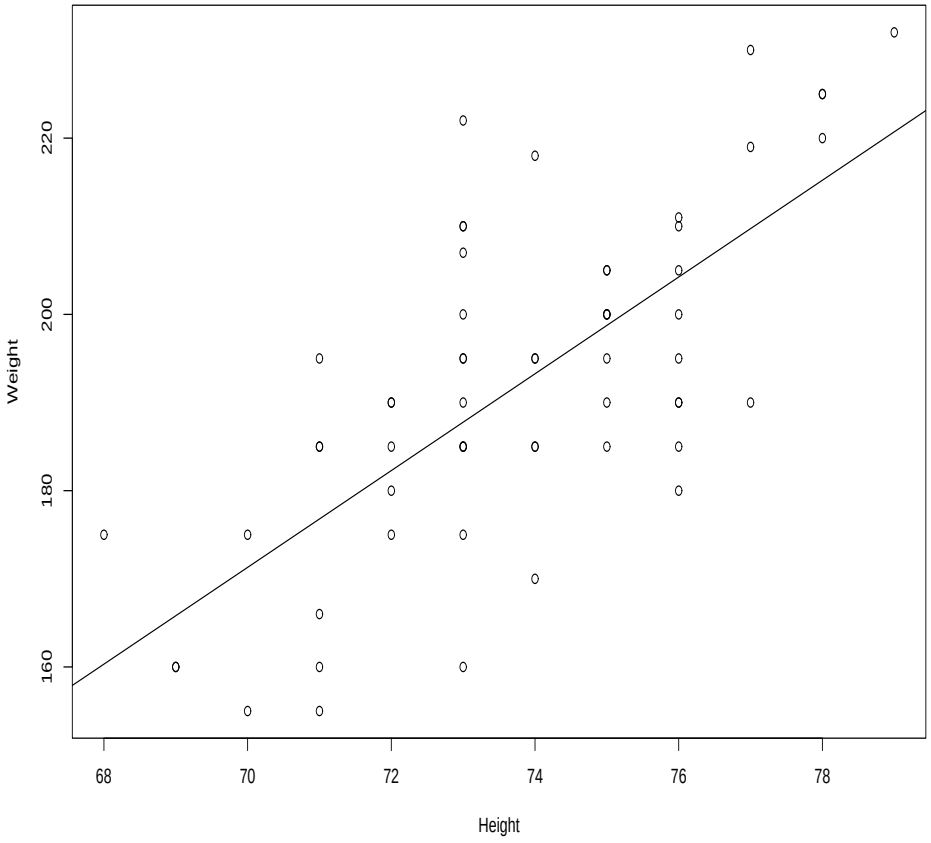
This is quite a change from the previous fit. In particular, the slope estimate has changed from 5.49 to 3.82, a difference of 1.67 pounds. That is, because of one data point we now predict weight to increase 1.67 pounds less for each one inch in height. We can also see the effect on the plot. See Figure 1.6.

Notice how the outlier pulled up the LS fit, resulting in a very poor to the bulk of the data. One data point drove the fit!

3. **Wilcoxon Fit**

As an alternative to LS, we present the Wilcoxon fit. Recall that the LS fit minimizes the averaged squared deviation from the chosen line. An outlier will have a large deviation and under the LS procedure its influence is made much greater by the squaring of this deviation. Because of the square, deviation times deviation, LS is weighing the large deviation by a

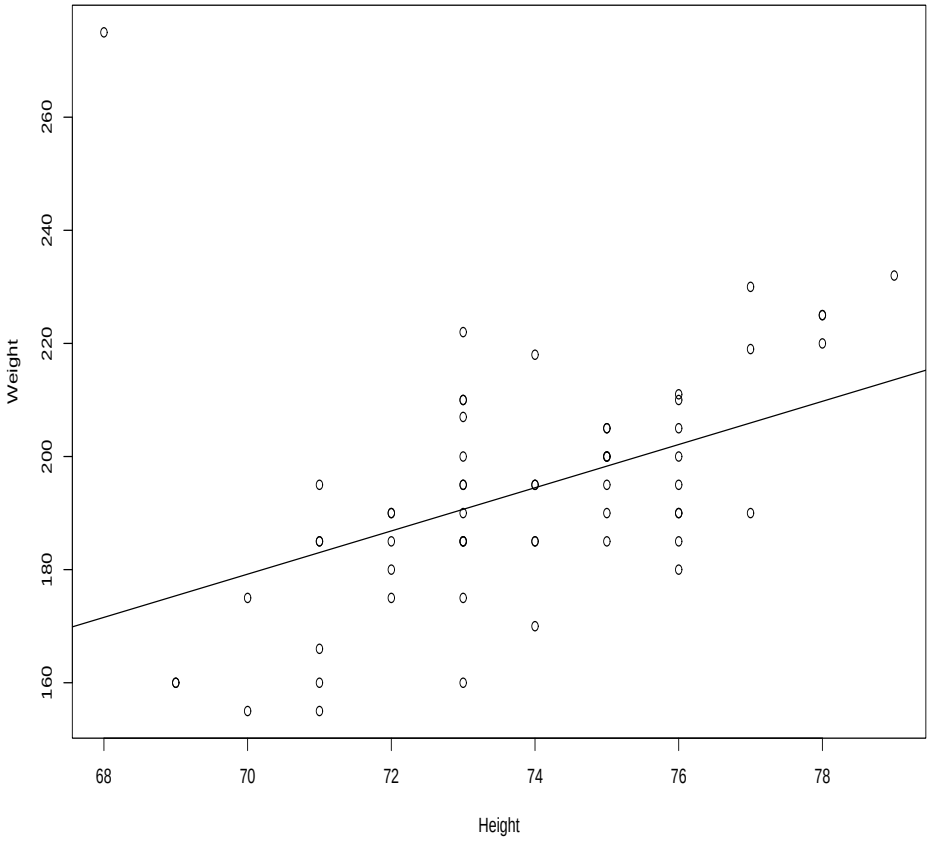
Figure 1.5: Baseball data : LS fit



large weight. The Wilcoxon, though, uses a much smaller weight in determining the chosen line. The Wilcoxon fit is less sensitive than the LS fit at least for outliers in the Y-direction. For good data, no outliers, the Wilcoxon fit is in close agreement with the LS fit. This the Wilcoxon fit is **robust fit**.

The regression module gives the option of a LS fit or a Wilcoxon fit. The Wilcoxon fit of the good data results in:

Figure 1.6: Baseball data : Changed data LS fit




```
Weight = -228 + 5.71 Height
```

Recall that the LS estimate of slope is 5.72 whereas the Wilcoxon estimate is 5.71, quite close. The Wilcoxon fit can be used like the other fits for prediction. For instance, if a ball player is 75" tall then the Wilcoxon fit predicts a weight of $-228 + 5.71(75) = 200.25$ pounds.

On the changed data, the Wilcoxon fit is

```
Weight = -225 + 5.67 Height
```

The change in slope estimates is very slight (.05 pounds). Unlike the LS fit, the Wilcoxon fit is not sensitive to the outlier.

Exercise 1.8.1

1. *Let X be the length (cm) of a laboratory mouse and let Y be its weight (gm). Consider the data for X and Y given below. Obtain a scatterplot of the data and comment on the plot.*

X	Y
16	32
15	26
20	40
13	27
15	30
17	38
16	34
21	43
22	64
23	45
24	46
18	39

2. *For the data set in Problem #1, eyeball a linear fit obtaining an estimate of the slope and the intercept.*
   (a) *Plot your fit.*
   (b) *Use your plotted fit, to predict the weight of a mouse that is 20 cm long.*
   (c) *Use your predicition equation to predict the weight of a mouse that is 25 cm long.*
   (d) *What does the estimate of slope mean in terms of the problem?*
   (e) *What does the estimate of intercept mean in terms of the problem?*
3. *Use the formulas given in class to determine the LS fit for the data given in Problem #1. (ANS: LS slope is: 2.405).*
4. *Plot your fit.*
5. *Compare the LS fit with your eyeball fit? Which is a better fit? Why?*

6. *Use the LS predicition equation to predict the weight of a mouse that is 25 cm long.*

7. *What does the estimate of slope mean in terms of the problem?*

8. *Use the regression module to scatterplot the data and obtain the LS and Wilcoxon fits. Write the Wilcoxon fit down.*

   (a) *Plot the Wilcoxon fit on your plot in #1.*

   (b) *Compare the Wilcoxn and the LS. Which is a better fit? Why?*

   (c) *Use the Wilcoxon predicition equation to predict the weight of a mouse that is 25 cm long.*

   (d) *What does the estimate of slope mean in terms of the problem?*

9. *Consider the height weight of the baseball players in Carrie's baseball data (Appendix A). Obtain the scatterplot of height versus weight, the LS fit, and the Wilcoxon fit.*

1.9 Relationships Between Variables, Part 2: Residual Analysis

Picking a model for a problem is a major undertaking. If the model fits well then it can be used to increase understanding of the problem and/or for prediction. For instance, by fitting the linear model to the heights and weights of the baseball data, we could see that weight seems to increase 5 pounds for each additional gain of one inch in height. This is not a great discovery but it is easy to think of situations where this rate of change is quite important. For example, consider a cancer drug that is supposed to reduce the size of a tumor and the experiment is the shrinkage (Y) of tumor size for a given dose (X) of the drug. If a linear model seems appropriate then the slope is expected reduction in tumor size when the dose is increased by one unit. In this section, we deal with the question of model adequacy.

We will only discuss the simple linear model. So we are considering two variables X and Y and we want to examine the adequacy of the model

$$Y = a + bX + e$$

The variable e denotes **random error**, that is, if there were no error Y would be a deterministic linear function of X.

When is a model good? At first, one might say when there is no error. But for all the data that we consider in this class there will always be error. For the baseball data above, there is a distribution of weight for each height. Actually we will say a model is good if there is no connection between e and $a + bX$; that is, the random error is free of X. Hence, for predicting Y, we have found the model that contains all the information based on X. Now there may be other variables which help in predicting Y. These will be contained in e. So the assumption we want to verify on a model is:

Model Assumption: The random error component is independent of the X component.

How would we check this assumption? If we knew the random errors, e, we could just plot them against $a + bX$. A random scatter would indicate that the errors do not depend on $a + bX$; i.e., the errors are free of $a + bX$. Thus the model is good. However, we don't know the errors, we only know Y and X. But using Y and X we estimate a and b. This leads to an estimate of $a + bX$, the predicted value of Y, which we label as $\hat{Y}$. Our estimate of the error is $Y - \hat{Y}$. This is called the **residual**, literally, what's left. We will denote the residual by $\hat{e}$, that is

$$\hat{e} = Y - (\hat{a} + \hat{b}X)$$

Then we can check our model assumption by plotting $\hat{e}$ versus $\hat{Y}$. This is called the residual plot. A random scatter indicates a good model. If it is not a random scatter then we need to rethink the model.

For example, consider the LS fit of the original baseball data, (no outlier). The prediction equation is $\hat{Y} = -213 + 5.49X$.

For each data point, we can find the predicted value and then the residual. We can then plot the residuals versus the fitted values to check our model assumption. For example the first data point is $(74, 218)$. The predicted value is $\hat{Y} = -213 + 5.49(74) = 193.26$. Hence the residual is $\hat{e} = Y - \hat{Y} = 218 - 193.26 = 24.74$ pounds. So we under predicted the weight of the first individual by 24.74 pounds. Hence one point on the residual plot is $(193.26, 24.74)$. Figure 1.7 contains the complete residual plot. Locate the point $(193.26, 24.74)$ on the plot. Determine the residual for the data point $(76, 200)$ and find it on the plot.

The residual plot is given by the **regression** module. Check the "Plot residuals vs predicted value" button if you wish the residual plot to be returned.

As a final example, consider the changed data set. The LS residual plot is given in Figure 1.8. Notice how the outlier stands out.

The Wilcoxon residual plot for the changed data set is given in Figure 1.9. Notice that the outlier stands out even further at 120 compared to 100 on the LS plot. Again the outlier draws the fit, thus shortening the distance (residual) between the outlier and the fit.

The LS estimates of slope and intercept are given by

$$\hat{b} = \frac{\text{Sum}(X \times Y) - n\bar{X}\bar{Y}}{\text{Sum}(X^2) - n\bar{X}^2}$$

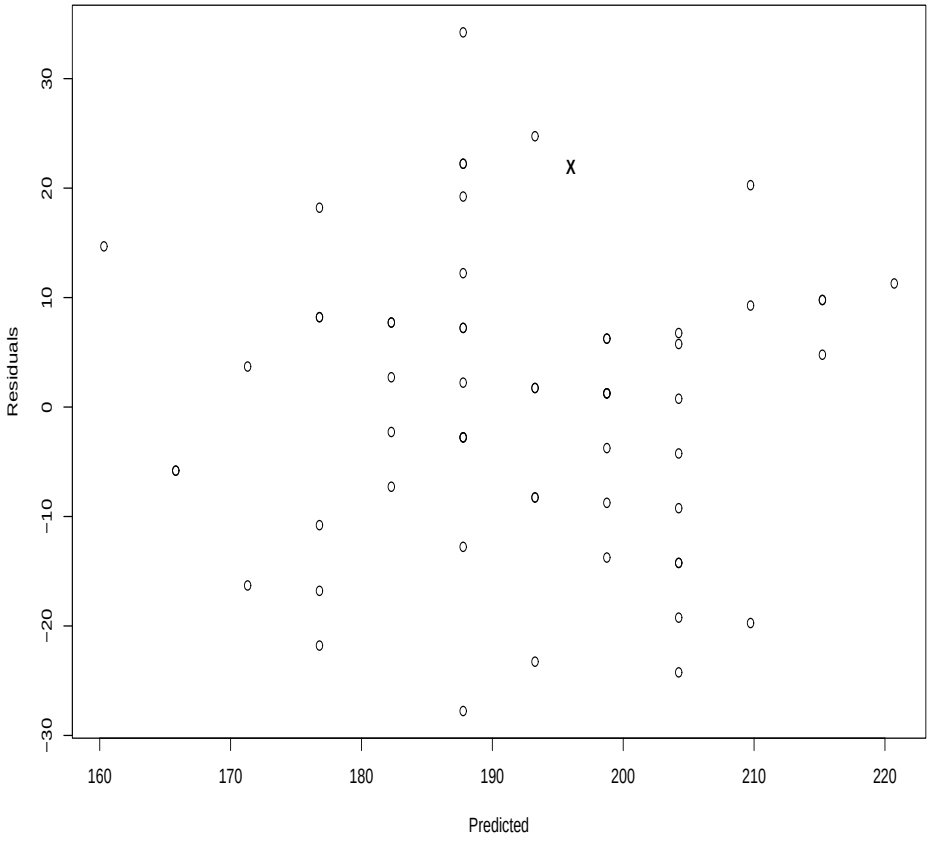
$$\hat{a} = \bar{Y} - \hat{b}\bar{X}$$

Exercise 1.9.1

1. *Let X be the length (cm) of a laboratory mouse and let Y be its weight (gm). Consider the data for X and Y given below.*

X	Y
16	32
15	26
20	40
13	27
15	30
17	38
16	34
21	43

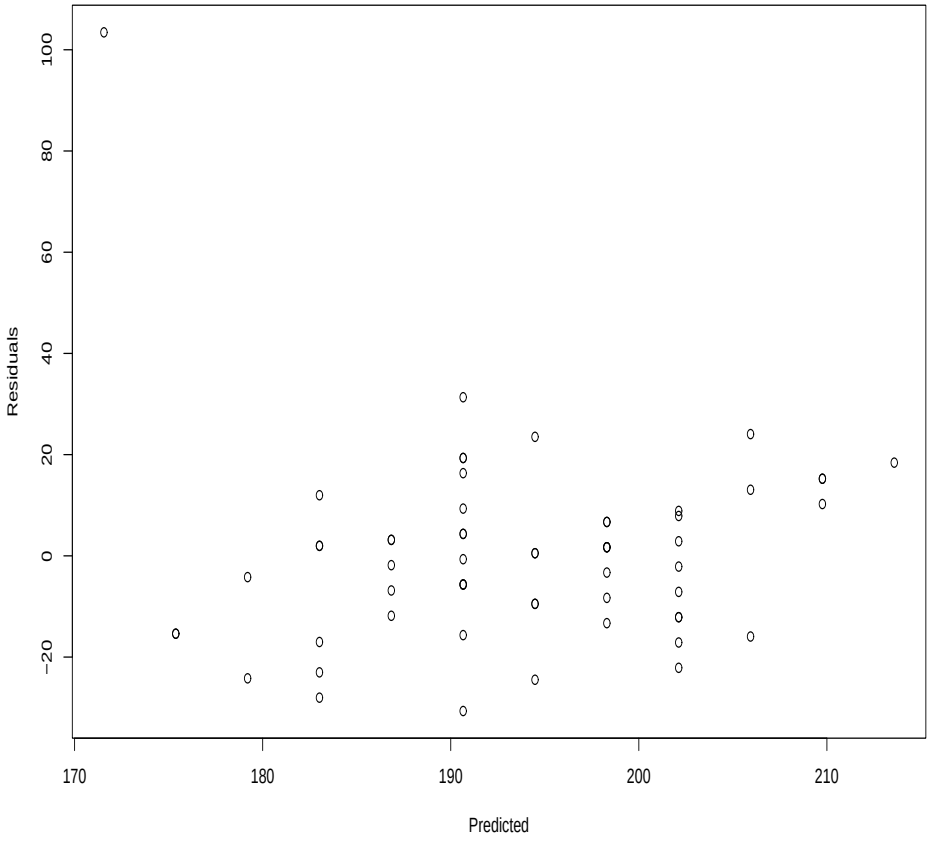
Figure 1.7: Baseball data : LS residual plot



22	64
23	45
24	46
18	39

Recall that you obtained an eyeball fit of this data in the last exercise. Use your fitted line (don't calculate!) to obtain the predicted value for each value of x*. Then by subraction* $(Y - \hat{Y})$ *obtain the residuals.*

Figure 1.8: Baseball data : LS residual plot for the changed data



(a) *Plot the residuals versus the fitted values. Comment on the plot.*

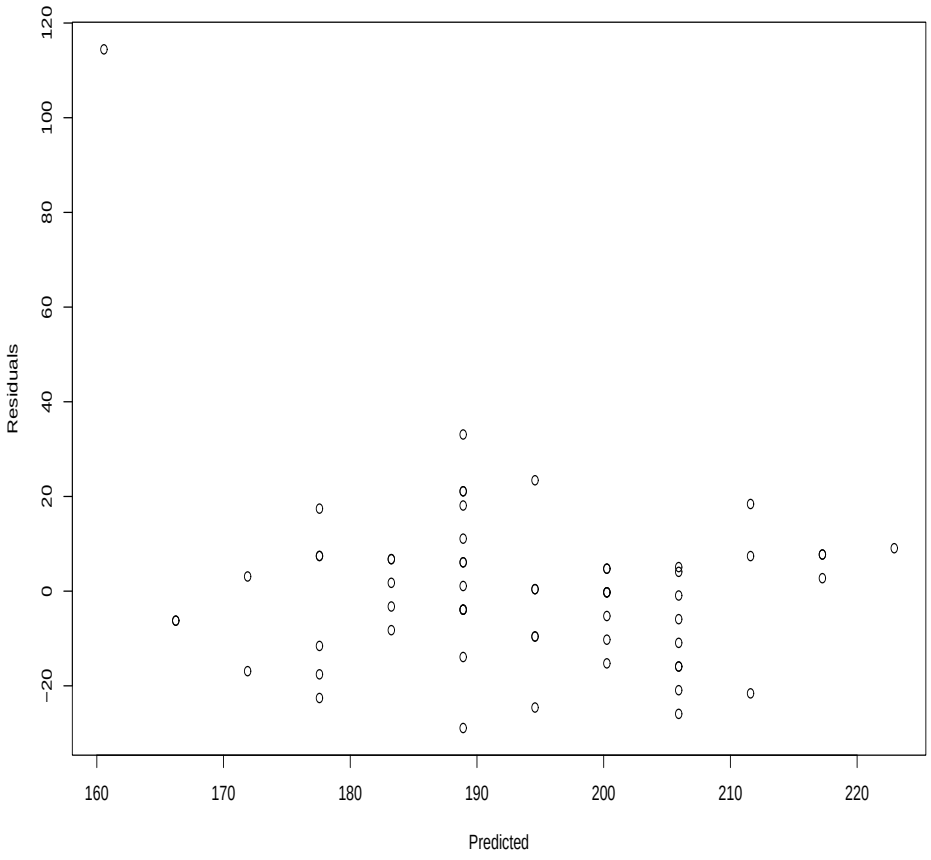
(b) *Obtain a stem leaf plot of the residuals.*

(c) *Obtain the 5 basic descriptive statistics for the residuals. Are there any outliers?*

2. *Recall that you obtained the LS fit for the above data in the last problem set. Calculate the LS residual for Case 9 (* $x = 22, y = 64$ *).*

3. *Recall that you obtained the Wilcoxon fit for the above data in the last problem set. Calculate*

Figure 1.9: Baseball data : Wilcoxon residual plot for the changed data



the Wilcoxon residual for Case 9 ($x = 22, y = 64$). From which of the 3 fits, LS, Wilcoxon or eyeball, would you spot the outlier more readily? Why?

1.10 Relationships Between Variables, Part 3: Measures of Relationships

In this section, we discuss measures of relationships between two variables X and Y. It is easiest to start with no relationship. What do we mean by no relationship? Suppose we had a lot of data on *(X, Y)* and obtained a scatterplot of Y versus X. If the plot was a random scatter then we would conclude that the variables X and Y are not related. What if they are related? Look at the six plots in Figure 1.10. In the first, we would probably conclude that X and Y are not related. Plot 2, we would characterize as probably a linear relationship, certainly exhibiting random error. Plot 3 is similar to Plot 2, although the pattern is not quite as tight. Plot 4 shows some negative drift. Plots 5 and 6 show the strongest relationships (tightest patterns) among the plots. Plot 5 shows a very strong circular relationship while Plot 6 a very strong quadratic pattern. It seems that a measure of a relationship should depend on what type of relationship it is. In this section, we will only be concerned for the most part about linear relationships and we will consider measures of such a relationship. It should not be surprising that this measure will indicate no (linear) relationship for the two strongest relationships in the plots.

Consider Plot 2 again. We want to measure the linear relationship exhibited in this plot. Two simple lines will help a lot. On the x-axis locate the sample mean of the X's ($\bar{X} = 0.6176199$) and draw a vertical line through this point. On the y-axis locate the sample mean of the Y's ($\bar{Y} = 0.6032577$) and draw a horizontal line through this point. Figure 1.11 shows these lines.

The lines intersect at $(\bar{X}, \bar{Y})$, (locate it). This is our new center. Next Label the quadrants I, II , III and IV, beginning at the upper right quadrant and continuing counter-clockwise. The coordinates of (X, Y) relative to the new center are $(X - \bar{X}, Y - \bar{Y})$. The signs on the coordinates are $(+, +)$, $(-, +)$, $(-, -)$, and $(+, -)$ as we go around the quadrants I, II , III and IV, respectively. Then it's easy to come up with many measures of linear relationships. A simple one is to count the number of points with the same sign (those in quadrants I and III) and subtract the number of points with different signs (those in quadrants II and IV). High values of this measure indicate a positive linear relationship while low values indicate a negative linear relationship.

Instead of counting like and unlike signs, we consider a measure which takes the product of these new coordinates. Thus we have n products, one for each point in the plot. Consider as a measure their average:

$$s_{XY} = \frac{\text{Sum}((X - \bar{X})(Y - \bar{Y}))}{n}$$

which is called the **sample covariance**. Positive values of this measure indicate a positive linear relationship while negative values indicate a negative linear relationship. Is this measure robust? No, you are catching on.

For a given data set, we can always make this measure larger (or smaller) by changing the units. Suppose we have a positive linear relationship and X is measured in feet. If we change the X's to

inches then s_{XY} increases by the factor 12. If we change the X's to mm's then s_{XY} increases by the factor 304.8. Thus we need to standardize our measure. In this chapter (we revisit this problem in Chapter 11), we will insist on an absolute measure which in absolute value cannot exceed 1. This is called the **sample correlation coefficient** and it is simply s_{XY} divided by the product of the standard deviations of the X's and the Y's, (except we divide by n and not $n-1$; i.e,

$$r = \frac{s_{XY}}{\sqrt{\frac{\text{Sum}(X-\bar{X})^2\text{Sum}(Y-\bar{Y})^2}{n^2}}}$$

This is our measure of linear relationship. As we said, for all data sets, $-1 \leq r \leq 1$. The extreme values are interesting:

$$r = 1 \text{ means a perfect positive relationship;}$$
$$r = -1 \text{ means a perfect negative relationship.}$$

Values of r close to zero indicate little or no linear relationship.

The values of r for each of the plots in Figure 1.10 is indicated in Figure 1.12.

As we thought, the strongest relationships score 0 with our measure because they are both nonlinear. The best linear pattern is Plot 2, although Plot 3 is close. The negative drift, Plot 3, registers $r = -.43$ and the first plot shows little linearity as initially thought.

We can do a bit more with the sample correlation coefficient. It is associated with the LS fit. It can be shown that $r = \left(\frac{s_Y}{s_X}\right)\hat{b}$ where $\hat{b}$ is the LS estimate of slope. So r contains information on the fit.

We can be more precise. Consider the variation (or noise) in the Y data. A measure of this variation is the sample variance s_Y^2 of the Y's. When we fit the linear model $Y = a + bX + e$ we should account be able to account for some of this variability (X should be of help in predicting Y. In fact, $r^2 100\%$ is the percentage of variation accounted for in the LS fit of Y versus X. We call this the **coefficient of determination** and we often use capital R^2 to denote it. Consider the values of R^2 for Plots 1-6. $R^2 = .007$ for Plot 1; hence we have accounted for .7% of the variation in Y. $R^2 = .66$ for Plot 2; hence we have accounted for 66% of the variation in Y. $R^2 = .59$ for Plot 3; hence we have accounted for 59% of the variation in Y. $R^2 = .18$ for Plot 4; hence we have accounted for 18% of the variation in Y. Of course for the last two plots, $R^2 = 0$. The value of R^2 can be obtained using the regression module.

The measures r and R^2 are not robust. We will consider alternative measures of r later, but for now we do offer an alternative to R^2 , labeled as R_W^2. This is the measure that corresponds to the robust Wilcoxon fit. This is not as sensitive as R^2 to outliers. We show this for the baseball height

and weight data. Recall that we changed the original data by inserting an outlier. The plots in Figure 1.13 show the original and changed data along with their R^2's and R^2_W's.

For the LS fit, notice that due to one outlier, the percentage of variation accounted for dropped from 50% to 19%. The measure corresponding to the robust Wilcoxon fit only changed from .44 to .39.

Exercise 1.10.1

1. *Scatterplot the following data and guess the correlation coefficient. Then compute it, (ans: .161).*

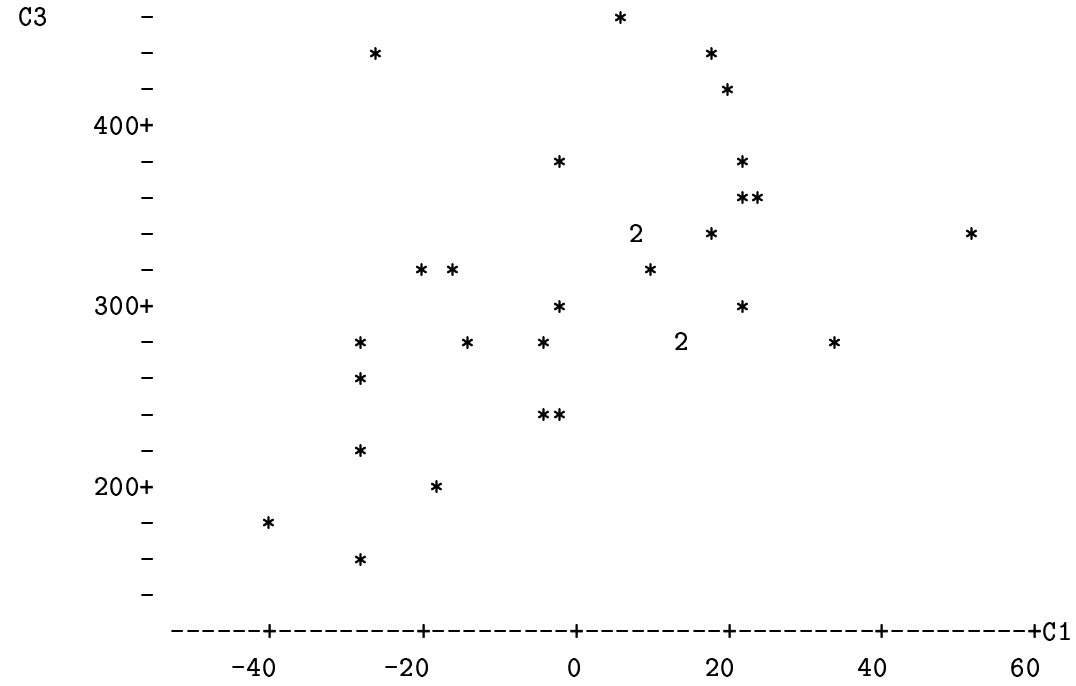
1	2
2	4
3	4
6	3

2. *Reconsider exercise #1 of Exercise 1.6. The data are given below. Scatterplot the data and guess the correlation coefficient. Recall that the LS estimate of slope was 2.405. Suppose the sample standard deviations of x and y are given by 3.58 and 10.42. Compute the correlation coefficient. (Ans: .825)*

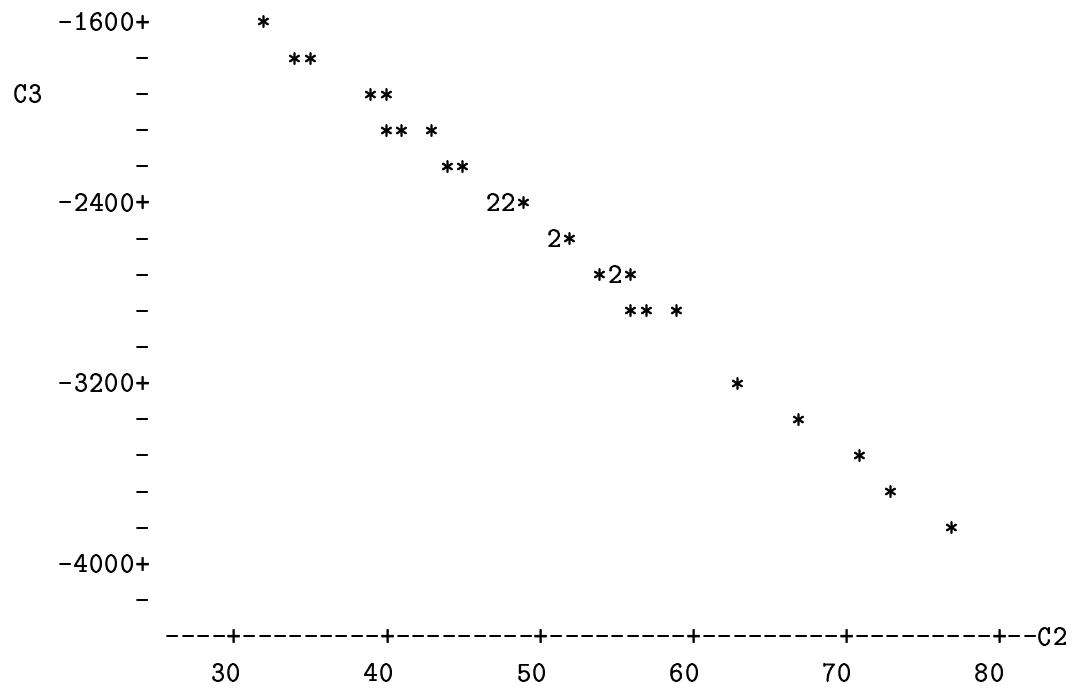
x	y
16	32
15	26
20	40
13	27
15	30
17	38
16	34
21	43
22	64
23	45
24	46
18	39

3. *In the last problem, what percent of variation of y is accounted for by x?*

4. *Which correlation seems appropriate for the following plot: -.678, .956, .892, .483, .045 ?*



5. Same as last problem for the following plot: .999, 0.0, .002,-.999, .500, .764



6. Same as last problem for the following plot: 0.0,.999, .500, -.18, .008, -.500

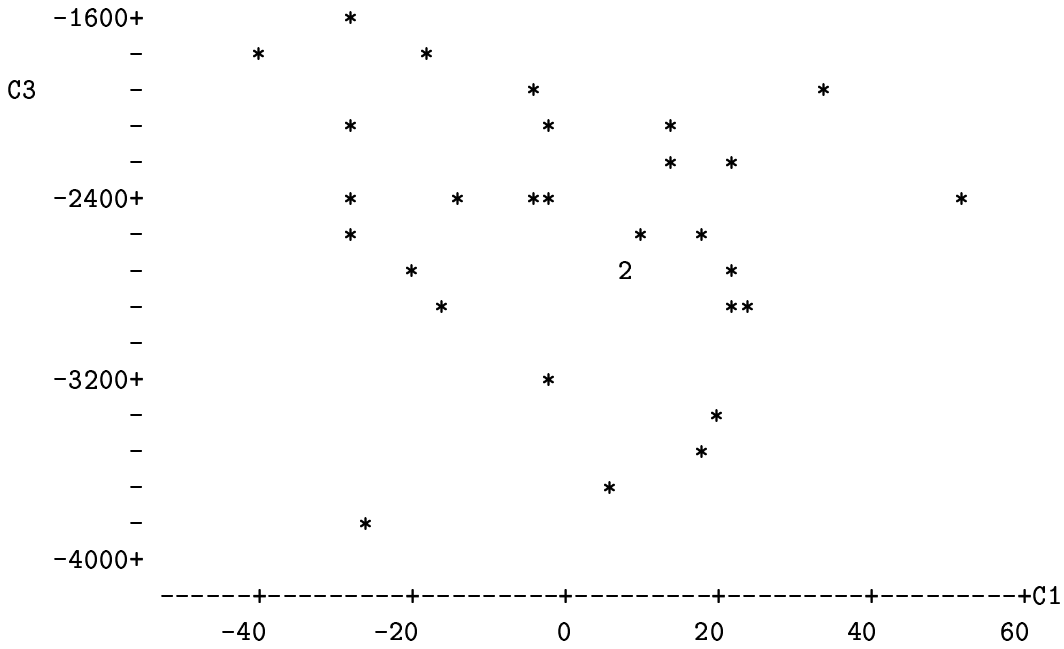


Figure 1.10: Scatter plots

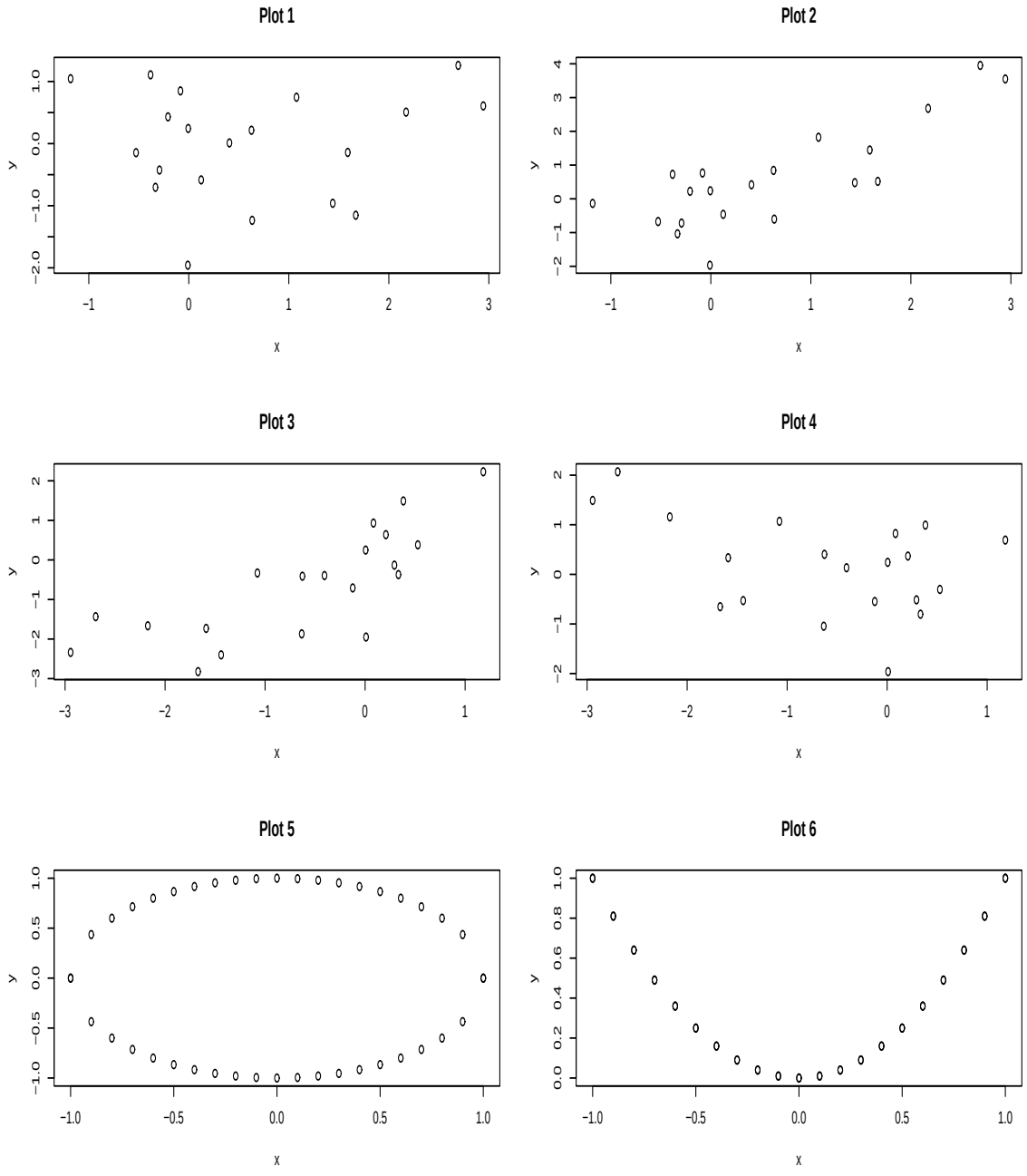


Figure 1.11: Plot 2 with sample means

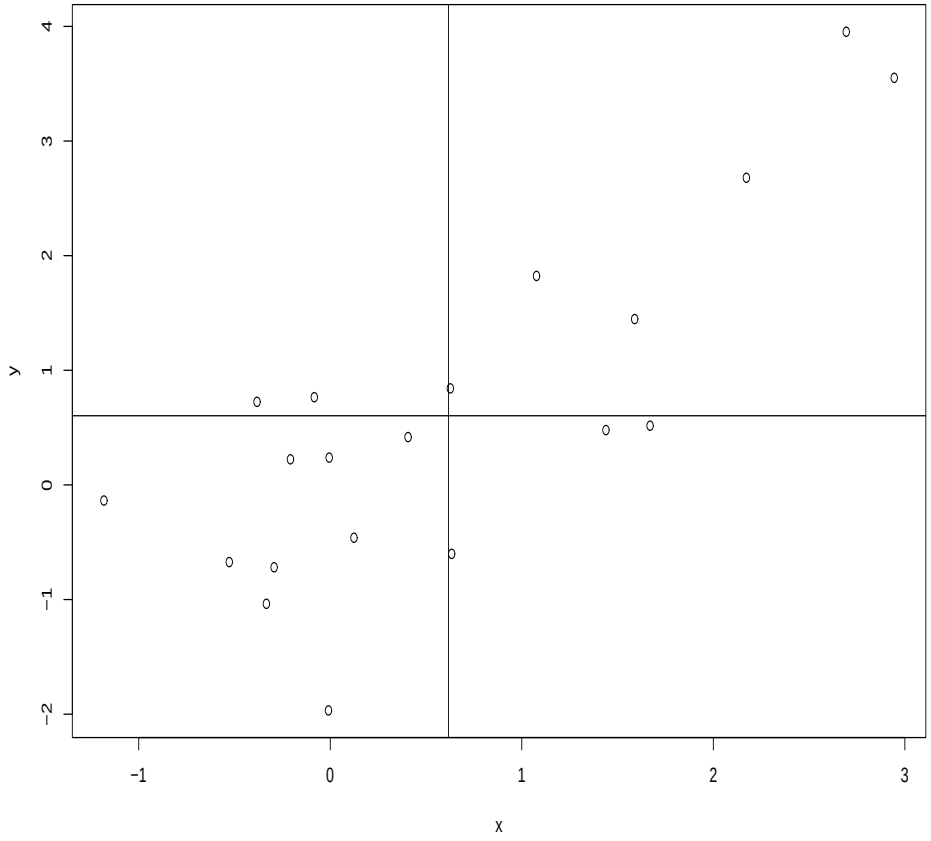


Figure 1.12: Scatter plots with values of r

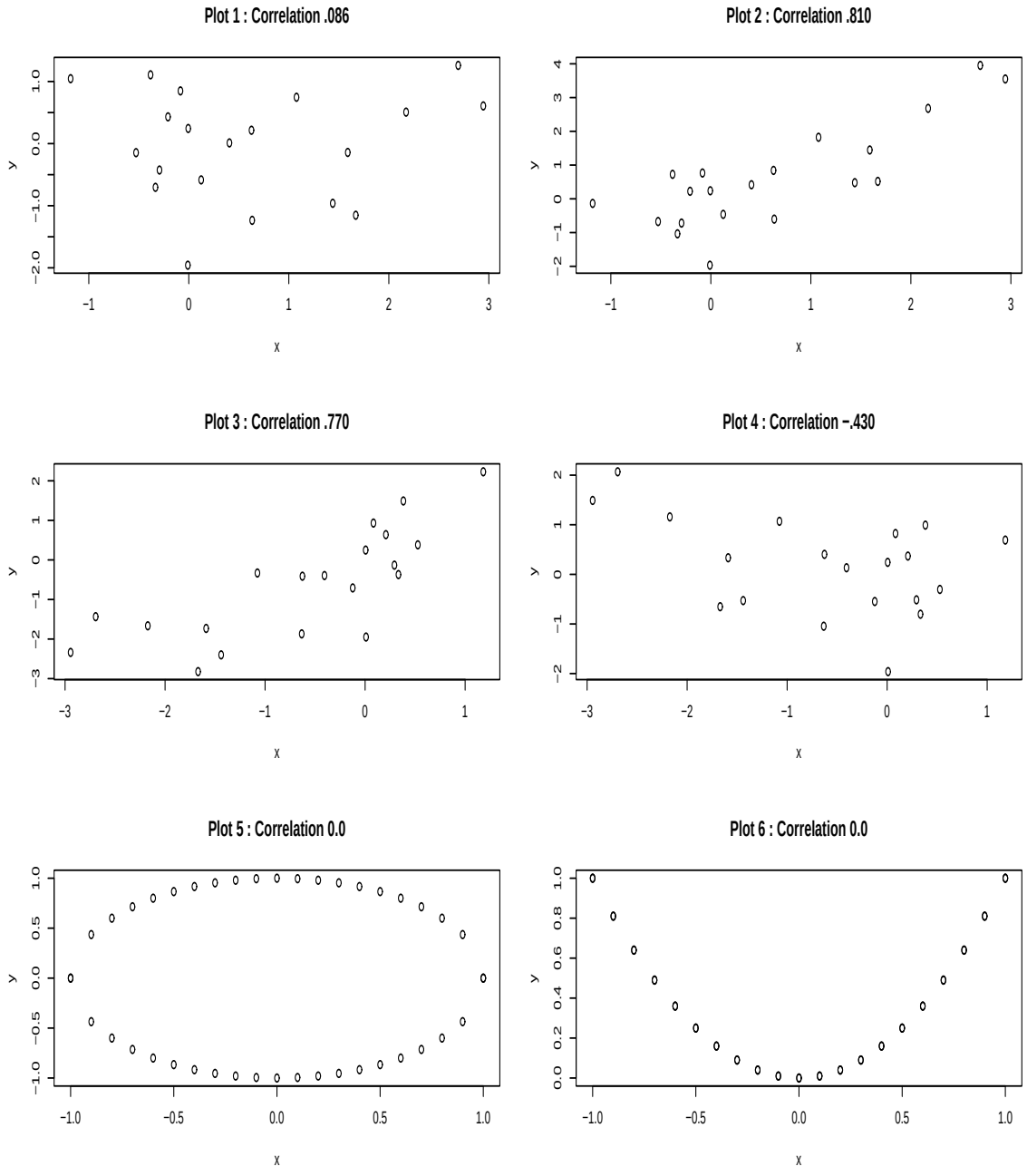
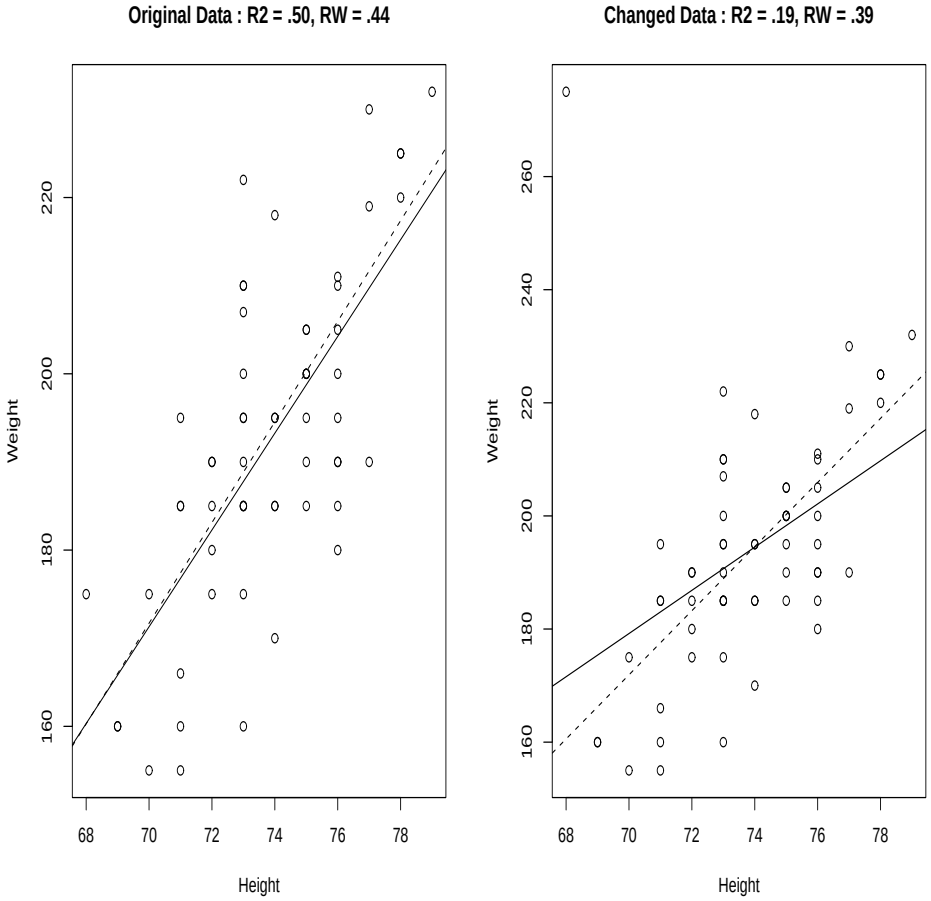


Figure 1.13: Coefficients of determination for LS and Wilcoxon fits



Chapter 2

Probability

2.1 Introduction

We need some, not much, **probability** for this class. It will help with assessing noise in samples. But, also, we can solve some very interesting problems in a simple fashion. In order to look at such problems several years ago, we would have had to stop and develop some mathematics. With **resampling** we no longer have to do this.

Consider first some simple examples:

1. Flip a *fair* coin. What's the probability of a head?
2. Roll a *fair* 6-sided die. You win the game if a 1 or 2 is the upface. What's the probability that you win?
3. Roll a pair of *fair* 6-sided dice. You win the game (on the first roll) if the sum of the upfaces is 7 or 11. What's the probability that you win? We will refer to this as the game craps in subsequent text.
4. Five cards are dealt from a standard *well shuffled* deck of 52 cards. What's the probability that the hand contains a pair. That is, what's the probability that in five card poker you open with a pair?
5. In a simple lotto you pick a number from 1 to 50. Later, to determine the winner, one number is selected at random. Find the probability that you win.
6. In Lotto 2, you select 4 numbers from the numbers 1 through 50. Find the probability that you win.

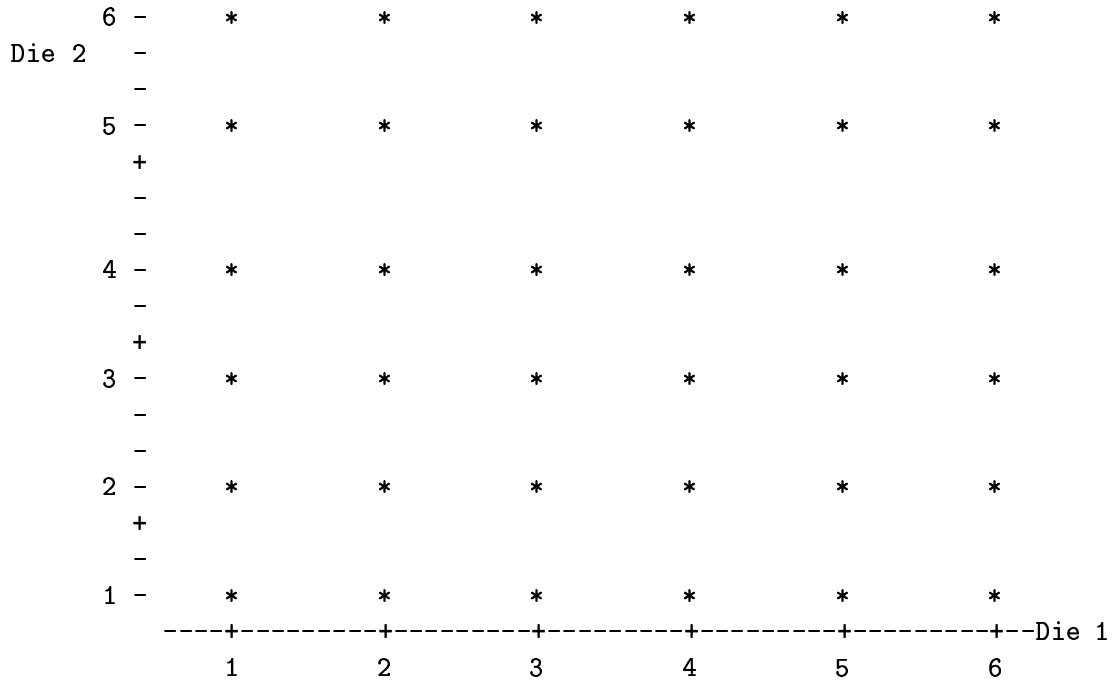
If you don't know the answers to the questions in the first two examples, the answers are given at the end of this section. But the answers for examples (3),(4) and (6) are not that easy to get.

We need a little nomenclature here which easily leads to the solution for (3) and will help contemplate the solution for (4) and (6).

- An **experiment** results in an outcome.
- The collection of all outcomes is the **sample space**. We shall denote sample spaces with the letter S.

Examples:

1. Flip a coin: $S = \{H, T\}$.
2. Roll a six sided die: S=*{1,2,3,4,5,6}*.
3. Roll a pair of 6-sided dice: S = *{(1,1),(1,2), (1,3), ..., (6,6)}*. That is, S consists of 36 pairs of integers. Here's a picture of S: (Read the points as (Die 1, Die 2).)



4. Five cards are dealt from a standard deck of 52 cards. I don't think I'll list S in this case since it contains over 2 and half million elements.

5. Play simple lotto : $S = \{1,2, \ldots, 50\}$.

6. Play Lotto 2 : $S = \{$ *all subsets of 4 numbers drawn from 1 through 50* $\}$.

But in terms of probability it is not the sample space that is of basic interest, but subsets of it.

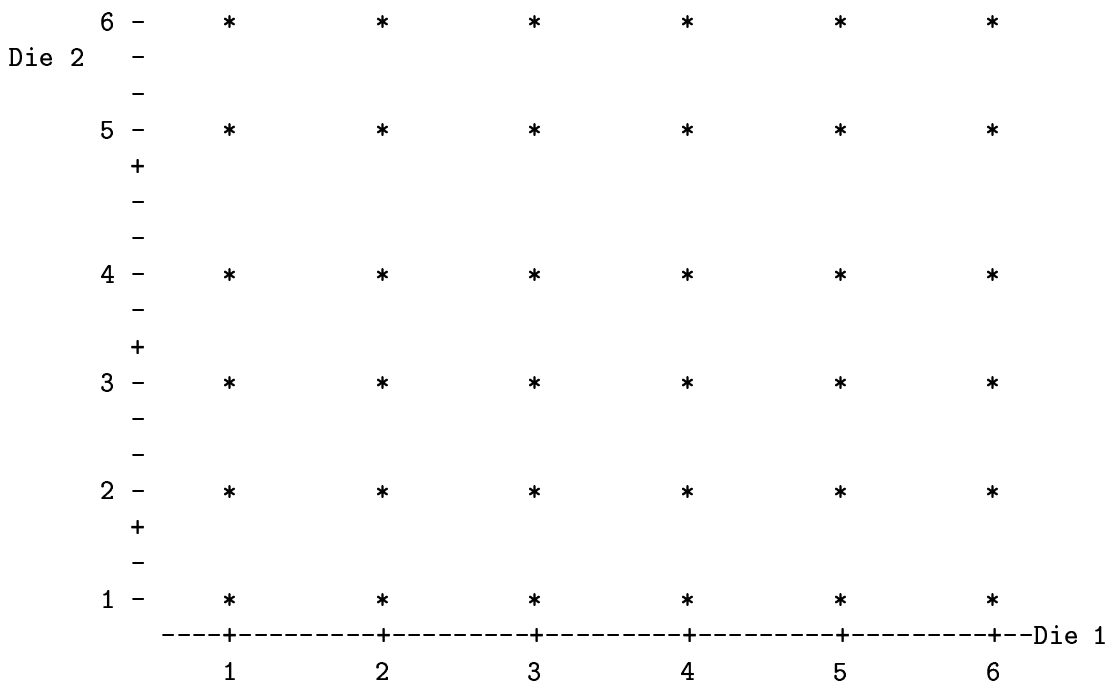
An **event** is a subset of S. Denote events by A, B, C, etc. We say the event A **occurs** if the experiment results in an outcome in A; i.e., A comes up. The **complement** of the event A occurs if A does not occur. We will sometimes write the complement of A by A^c.

Examples

1. Flip a coin: $A=\{H\}$.

2. Roll a six sided die: $B=\{1,2\}$.

3. Roll a pair of 6-sided dice: *A= sum of upfaces 7 or 11.* Find A on the picture:



Note that the 7's (sum of upfaces 7) fall along the main diagonal starting with the point (1,6) and ending with the point (6,1).

4. Five cards are dealt from a standard deck of 52 cards : *A= just a pair.* Alas, A is too big to list, also, because it has over a million elements. But here is one element: {*Jack of hearts, jack of clubs, 7 of diamonds, 9 of spades, 2 of clubs*}. What's another such hand?

5. A is the event that you picked the winning number in the simple lotto.

6. (a) A is the event that you picked the 4 winning numbers in Lotto 2.
   (b) You buy 100 Lotto 2 tickets. B is the event that one of your tickets is the winner.

Answers :

1. The probability of a head on the flip of a *fair* coin is 1/2.

2. The probability of getting a 1 or 2 on a roll of a fair 6-sided die is $2/6 = 1/3$.

Exercise 2.1.1

1. *List the sample space, list the event of interest, and its complement for the experiment: Spin a spinner with the numbers 1 through 10 on it. Suppose we are interested in the event an odd number spun.*

2. *List the sample space, list the event of interest, and its complement for the experiment: Roll a pair of 6-sided dice. We are interested in the event that both dice are the same.*

3. *List the sample space, list the event of interest, and its complement for the experiment: A pizza can have none, one, two or three of the toppings onions, extra cheese, or peppers. We are interested in a pizza with only two toppings.*

4. *List the sample space, list the event of interest, and its complement for the experiment: From a standard deck of 52 cards, three cards are dealt (without replacement) and their color is observed. We are interested in getting 3 red cards.*

2.2 Probabilities

We want the probability of event. Probabilities have to satisfy only the following three requirements:

A **probability** is an assignment of numbers to events so that

1. The probability of an event A is a number between 0 and 1.
2. The probability of the sample space is 1.
3. If two events cannot occur at the same time the probability that one or the other occurs is the sum of the probabilities of the individual events.

We will denote the probability of A by $P(A)$. Notice that the first two requirements are not special. They simply state that probabilities are numbers between 0 and 1 and if the experiment is performed the sample space occurs with probability 1. The third requirement makes sense intuitively. For example, in the game of craps, the events A = *sum of upfaces is 7* and B = *sum of upfaces is 11* cannot occur at the same time, so the probability of a *7* or *11* is $P(A) + P(B)$.

For a discrete (finite) sample space there is an easy way to obtain many examples of probabilities. Consider a subspace with m elements, say, $S = \{x_1, ..., x_m\}$. Let $p_1, ..., p_m$ be m fractions between 0 and 1 which sum to 1. For an event A, consider the assignment

$$P(A) = \text{sum of all } p_i\text{'s for which the element } x_i \text{ is in } A.$$

Then the assignment P is a probability.

Example: Suppose a spinner with the numbers 1 through 6 on it is spun and the number spun is observed. The sample space is $S = \{1,\ 2,3,4,5,6\}$. Let A be the event $A = \{1,2\}$. The following are four *different probabilities* on S and the resulting probability of $A = \{1,2\}$.

1. $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6$
   $P(A) = 2/6 = 1/3$.
2. $p_1 = p_2 = .25$, $p_3 = p_4 = .15$, $p_5 = p_6 = .1$.
   $P(A) = .50$.
3. $p_1 = .10$, $p_2 = .25$, $p_3 = .1$, $p_4 = .15$, $p_5 = .30$, $p_6 = .1$.
   $P(A) = .35$.
4. $p_1 = 1/21$, $p_2 = 2/21$, $p_3 = 3/21$, $p_4 = 4/21$, $p_5 = 5/21$, $p_6 = 6/21$.
   $P(A) = 3/21 = 1/7$.

5. $p_1 = .80$, $p_2 = .04$, $p_3 = .10$, $p_4 = .02$, $p_5 = .03$, $p_6 = .01$
   $P(A) = .84$.

6. $p_1 = 3/12$, $p_2 = 2/12$, $p_3 = 2/12$, $p_4 = 1/12$, $p_5 = 3/12$, $p_6 = 1/12$
   $P(A) = 5/12$.

Exercise 2.2.1

1. *In the last example, obtain 5 more different probabilities on S.*

2. *In the last example, how many probabilities are there on S?*

3. *In Exercise 2.1, #1, assume the spinner is fair. What is the assignment of probabilities? What is the probability of A, the event of interest? What is the probability of A^c, the event of interest?*

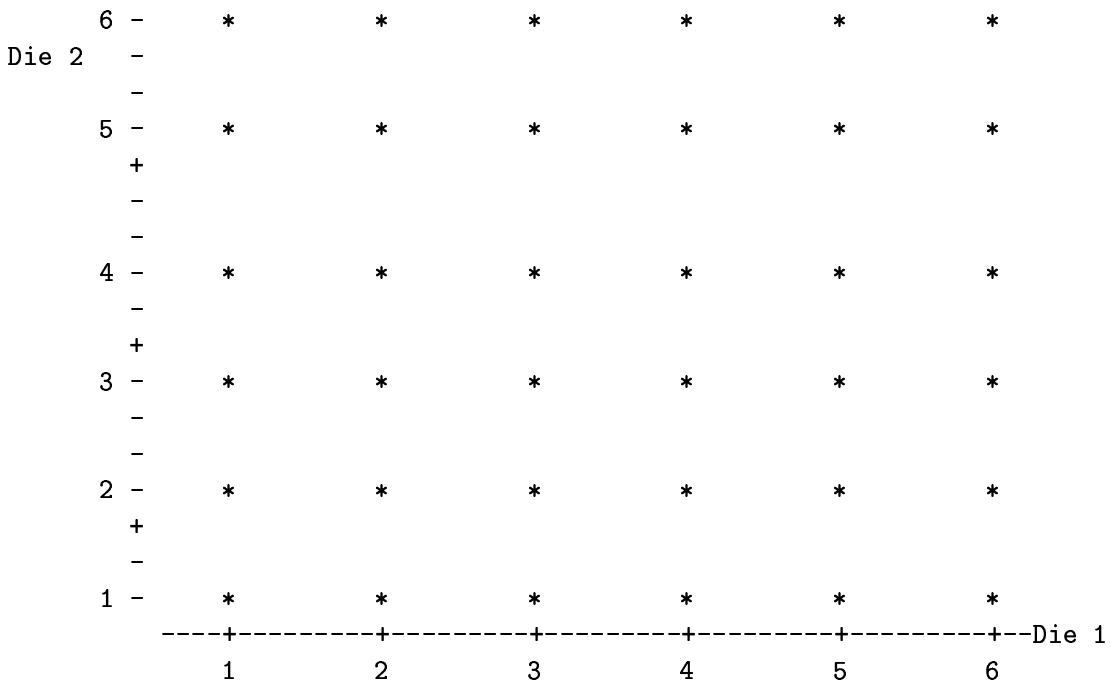
4. *In Exercise 2.1, #3, assume that the probability of any topping on a pizza is 1/2. What is the assignment of probabilities? What is the probability of A, the event of interest? What is the probability of A^c, the event of interest?*

2.3 More on Probability

The first probability in the last example, i.e., the fair 6-sided die, is a special case. It is called the **equilikely case**. For this example, it is the assignment of probabilities under the **assumption that the die is fair**. In real life, this assumption is a **statistical hypothesis,** which we may want to test. For example, you are playing craps for high stakes; hence, you may want to test to see if the die is fair. The die has only to be shaved slightly (loaded die) to change the probabilities. Of course the fourth example above is a die loaded to high numbers.

How would you test to see if the die is fair?

Lets answer one of the questions posed above. Suppose we roll two fair dice. What's the probability that the sum of the upfaces is 7 or 11? Here's the sample space again:



Since the dice are fair, it seems that each of the points is equilikely. Since there are 8 (6 as "7" and 2 as "11") elements in the event of interest, the probability of a "7" or "11" is 8/36.

Note if we assume the equilikely case for assigning probabilities, then the probability of any

event is just the number of elements in that event divided by the number of elements in the sample space.

Exercise 2.3.1

1. *Six cards with the numbers 1 through 6 on them are well shuffled and two cards are dealt (without replacment). Find the probability that the sum of the numbers on the two cards is 7. Note order is not important here. For example, the hand with cards 1,2 in it is the same as the hand with cards 2,1. In the sample space there are 15 elements. List them. Then find the probability that the sum of the numbers on the two cards is 7. Why is your answer different from the craps game answer?*

2.4 Relative Frequency

We need a few notes on the relative frequency idea of probability. Suppose we want to determine the probability of some event A. Suppose that we can repeat the experiment over and over again, such that the trials are:

1. **Independent** of one another.
2. **Identical**, in that conditions do not change from one trial to another.

Let N be the number of trials and let $\#(A)$ denote the number of times A occurred. Then

$$\text{The probability of } A \text{ is approximately } \frac{\#(A)}{N}$$

and the approximation gets better as N gets larger. Note that the relative frequency idea of probability obeys the three axioms of probability.

2.5 Determination of Probabilities 1: Tree Diagrams

We will discuss three ways of determining probabilities: enumeration (listing of the sample space), tree diagrams, and resampling. Resampling will be discussed in the next chapter and enumeration is what we have been doing. For instance, we solved the probability of a "7" or "11" in the game of craps by listing the sample space and just observing the number of times we get "7" or "11". This gets tedious very quickly and for many problems is quite unrealistic. For example consider the following problem:

Urn Problem: Suppose we have an urn with 30 blue balls and 50 red balls in it and that these balls are identical except for color. Suppose further the balls are well mixed and that we draw 3 balls, without replacement. Determine the probability that the balls are all of the same color.

Even for this simple problem there are 82160 elements in the sample space. But note that this problem is sequential in nature; i.e, there are 3 steps (draw Ball 1, draw Ball 2, and draw Ball 3). In such cases a simple **tree diagram** can often solve the problem or even if unfinished lead to the answer. We will draw our trees horizontally. So for this problem, begin by putting a dot on the center left side of your paper. Then for the first ball, it is either blue or red. Beginning at the dot, trace a branch up for blue, putting the probability of the first ball being blue, 30/80, on it and a "B" at the end. Likewise, trace a branch down for red with 50/80 on it and an R at the end. Hence at the end of the first step your tree looks like Figure 2.1.

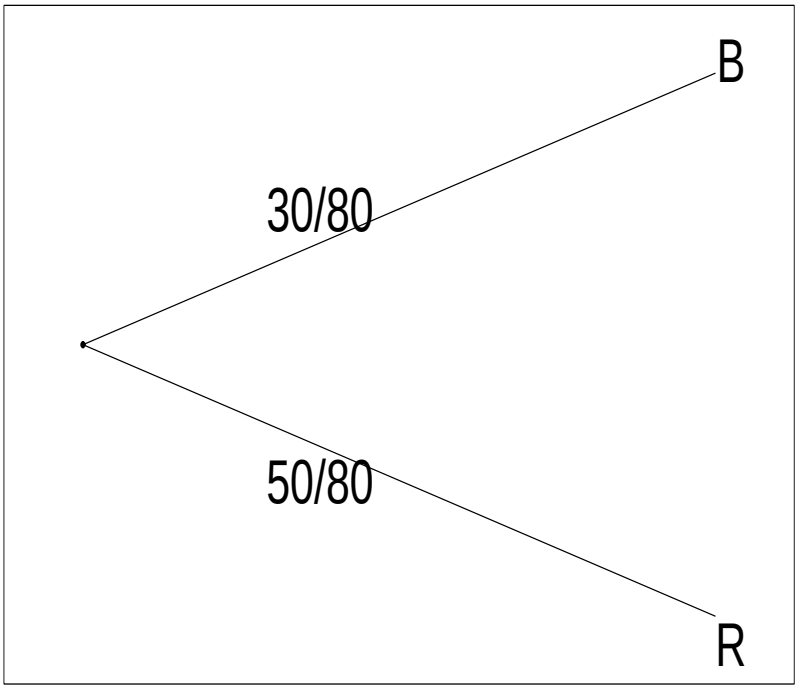
Next do the second step at each of the ends of the first step. The second ball is either blue or red. That is, at the "B", draw one branch up for second ball blue with the probability of 29/79 on it (on this branch, you have already drawn one of the blue balls, so there are 29 blue balls left out of 79 balls), and end it with a "B". Next draw one branch down for second ball red with the probability 50/79 (a blue ball was drawn on the first step so there are 50 red balls left out of 79 balls) and end it with an "R".

Now you try the second step at the "R" of the first step. If you have done it right, your tree diagram at the end of the second step should look like Figure 2.2.

Hey this is easy stuff! Now you try the third step. The ball can be blue or red so there will be two branches at the end of the four second step branches. If you have done it right, your tree diagram at the end of the third and last step should look like:

Look at the node of the final "B", "B", "B" branch. This means blue ball on first step, blue ball on second step, and blue ball on third step. What's the probability of this? It's easy. 30 out of 80 times you go up to the first "B", and *of those times* 29 out of 79 times you go up to the second "B", and *of those times* 28 out of 78 times you go up to the third "B". The key word, here, is **of**. That is 28/78 of 29/79 of 30/80 times you get to the "B", "B", "B" node. Hence the probability

Figure 2.1: Tree diagram : Step 1



of three blue balls is:

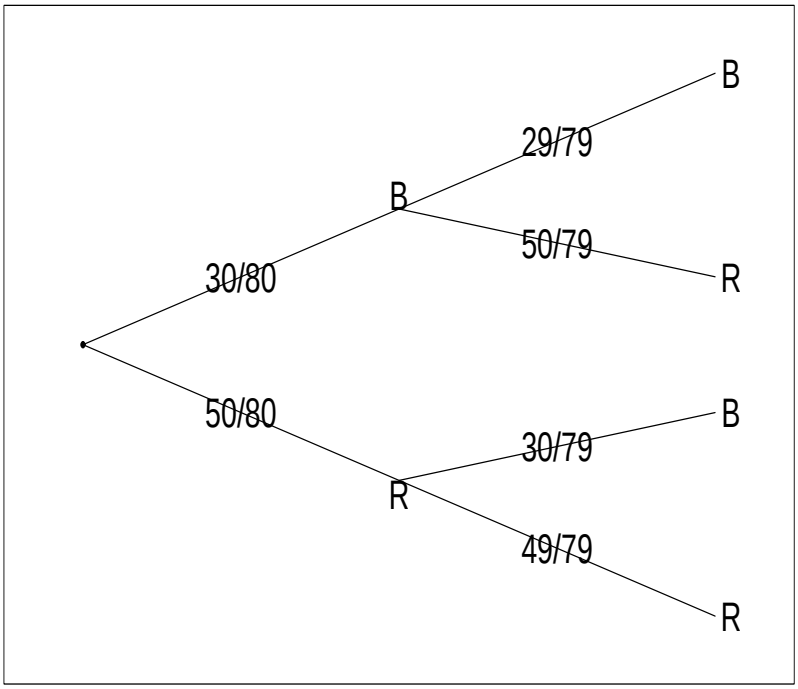
$$\frac{28}{78} \times \frac{29}{79} \times \frac{30}{80} = .0494.$$

Likewise the probability of three reds is (follow the bottom most branch to its final node) is

$$\frac{48}{78} \times \frac{49}{79} \times \frac{40}{80} = .2386.$$

Finally, the probability that the balls are the same color is the probability of either 3 reds or 3 blues. These events cannot happen at the same time (in fact all last 8 nodes are disjoint events);

Figure 2.2: Tree diagram : Step 2

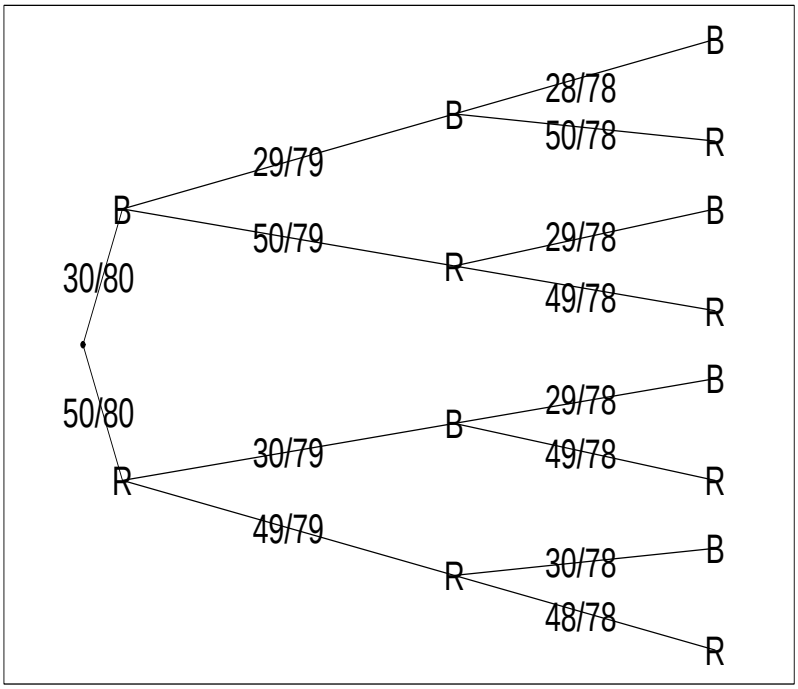


hence, the probability that the balls are the same color is *.0494 + .2386 = .2880.* That is, almost 30% of the time you will draw three balls out of the urn which are the same color.

Exercise 2.5.1

1. *In the urn example, find the probability of getting 2 redballs and one blue ball.*

Figure 2.3: Tree diagram : Step 3



2. *In the urn example, find the probability of getting the first two balls red.*

3. *In the urn example, suppose we darw 4 balls without replacement. Now find the probability that all 4 balls are of the same color.*

4. *Use a tree diagram to determine the probability of getting three red cards, when three cards are dealt (without replacement) from a well shuffled standard deck of 52 cards.*

5. *Six cards with the numbers 1 through 6 on them are well shuffled and two cards are dealt*

(without replacment). Use a tree diagram to determine the probability of that the sum of the numbers on the two cards is 7.

6. *In the urn example, find the probability of getting 2 redballs and one blue ball.*

2.6 Independence

In solving the urn problem, we stumbled across the concept of independence. This will be important to us and we need to spend a few minutes on it. Consider again the urn problem:

Urn Problem: Suppose we have an urn with 30 blue balls and 50 red balls in it and that these balls are identical except for color. Suppose further the balls are well mixed and that we draw 3 balls, without replacement. Determine the probability that the balls are all of the same color.

Recall the final tree diagram given in Figure 2.3. The probabilities on the branches are called **conditional probabilities**. For example:

1. Let B_2 denote the event that the second ball is blue and
2. let A_1 denote the event that the first ball is blue.

Then the probability on the first step upward branch is the probability that B_2 occurs given that A_1 has occurred; i.e, 29/79. This is called the conditional probability of B_2 given A_1 and we will denote it by $P(B_2|A_1)$. The bar is pronounced "given".

In general for two events A and B, if $P(B|A) = P(B)$, i.e, knowledge of A did not change the predicition of B, then we say A and B are **independent events**.

Note for the urn problem, if, as above, B_2 denotes the event that the second ball is blue and A_1 denotes the event that the first ball is blue then $P(B_2|A_1) = 29/79$. What is the $P(B_2)$? All B_2 says is that the second ball drawn is blue. So to determine the $P(B_2)$ go the the final tree diagram and look at all the end nodes for which the second ball is blue. Then add up all the probabilities associated with these end nodes. Try it. Now add up the probabilities. You get 30/80, the same as the $P(A_1)$. Surprised? This may be counterintuitive, but what is so special about the first ball over the second ball? Nothing. Okay, what is the probability that the third ball is blue? It's 30/80. If you don't believe me go to the tree diagram and add up the probabilities on the final nodes associated with a blue third ball. If we continued with this urn game and drew all 80 balls without replacement, what is the probability that the eightieth ball is blue? YOU GUESSED IT, 30/80.

Before we forget it, for the urn problem we showed that $P(B_2|A_1) = 29/79$ and $P(B_2) = 30/80$; hence, A_1 and B_2 are not independent events. We say that they are dependent events.

Lets return to the urn problem once again. Suppose we do the **sampling with replacement**. That is we remove a ball, record its color, put it back in the urn, mix the balls well, and then remove the next ball. We do this until we get 3 balls. Now what's $P(B_2|A_1)$, i.e, the probability that the second ball is blue given the first ball is blue. In this case, **sampling with replacement**,

it's 30/80, because the first ball is replaced and, hence, the contents of the urn are the same on the second draw as they were on the first draw. Thus the events A_1 and B_2 are independent events in the case of sampling with replacement.

Actually we can get a neat formula out of conditional probability that will be useful from time to time and it will also give us a way to use independence.

Let A and B be arbitrary events. We want to determine $P(B|A)$. Suppose that the tree diagram is too complicated. But we can use relative frequency. So we repeat the experiment many times, say, N. Now of these times we only want the times that A occurs $\#(A)$, (because we want the probability of B given that A **has occurred**).

Now of the times that A has occurred, count those times that B has occurred. What have you counted in this last count? Wait! Whatever this count is, lets denote it by *"Last Count"*. I claim that

$$P(B|A), \text{ is approximately } \frac{\textit{"Last Count"}}{\#(A)}.$$

But what have we counted in getting *"Last Count"*.? Say it! That's right! We have counted the number of times both A and B occurred simultaneously, i.e, *#(A and B)*. Hence

$$P(B|A), \text{ is approximately } \frac{\#(A \text{ and } B)}{\#(A)}$$

But I won't change anything by dividing both the numerator and denominator of this fraction by N, the number of times that we repeated the experiment. Hence

$$P(B|A), \text{ is approximately } \frac{\#(A \text{ and } B)/N}{\#(A)/N}.$$

As N gets large this last fraction gets close to $\frac{P(A \text{ and } B)}{P(A)}$. Thus we define the conditional probability of B given A as

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

Lets rewrite it as a formula for $P(A \text{ and } B)$, which is

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

This is called the **multiplicative law**. Finally, if A and B are independent events we get

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

If we can **recognize independence** then we can use this formula to compute *P(A and B)*. As an example, consider the following.

Jet Example: A jet airplane has 3 engines which function independently of one another. The probability that an engine fails in flight is .0001. Furthermore, the plane can fly if at least one engine is functioning. Determine the probability that the airplane has a successful flight.

The event we want to consider is A = at least one engine operates throughout the flight. Consider the complement of A, A^c which is the event all three engines fail.

1. Let B_1 be the event that engine one fails.
2. Let B_2 be the event that engine two fails.
3. Let B_3 be the event that engine three fails.

Hence, A^c is the event B_1 and B_2 and B_3 occurs. Thus

$$P(A) = 1 - P(A^c) = 1 - P(B_1 \text{ and } B_2 \text{ and } B_3)$$

It seems that the engines function independently of one another; hence, B_1, B_2, and B_3 are independent events. So

$$P(B_1 \text{ and } B_2 \text{ and } B_3) = .0001 \times .0001 \times .0001 = .00000000001.$$

Hence $P(A) = .999999999999$.

Exercise 2.6.1

1. *Suppose we flip a fair coin 4 times. What's the probability of 4 heads?*
2. *A basketball player free throw percentage is .70. If he shoots 6 free throws, find the probability that he makes all 6. What are we assuming here that may not be true?*
3. *Suppose in the Jet airplane example, that one engine is broken before takeoff, but the plane takes off anyway. Determine the probability that the plane arrives safely.*
4. *Suppose A and B cannot occur at the same time. Are they independent?*
5. *In a call in poll, are the calls independent of one another?*
6. *Suppose the subjects in a poll are selected by random phone calling. Are the calls independent of one another?*
7. *A newspaper reporter goes out to the mall and asks people a question of local interest. Are these respondees independent of one another?*

Chapter 3

Resampling

3.1 Introduction

We have discussed determination of probabilities of events by enumeration and tree diagrams. These are useful for some small problems but are very limited. For example, the probability of opening with a pair in 5 card poker is impossible to obtain by these methods. We could turn to the theory of probability but that would involve higher mathematics. Fortunately, with ever increasing speed of computers we have another way, **resampling**. Using resampling we can estimate the probability of the event and, further, we can increase the accuracy of the estimation by simply increasing the number of resamples.

Another advantage of resampling is that you have to build a **model** to accomplish it and you can only build a *correct* model if you understand the problem. There are basically 4 steps to resampling. We outline the steps in general and then give several examples.

Let A be the event of interest.

1. **Choose a model and define a trial**. In class, this often means portraying the sample space and event accurately using a table of random digits. The **trial (repetition of the experiment)** must be done explicitly.

2. **Define the event of interest** in terms of Step 1. We must be able to compute the *P(A)* in terms of the trial.

3. Obtain N trials of the experiment. Count the occurrences of the event A. Denote this count by *#(A)*. It is extremely important that:

   (a) The trials are **independent** of one another.

   (b) The trials are performed under **identical** conditions.

   If one or both of these conditions fail then there is **NO** guarantee whatsoever that the result in Step 4 is an estimate of the *P(A)*! Furthermore there is **generally NO WAY** to estimate the error of the estimate! It is indeed usually **GIGO** *Garbage In, Garbage Out.*

4. Estimate the $P(A)$ by $\frac{\#(A)}{N}$.

Lets do a simple problem. On the roll of a **fair** 6 sided die, determine the probability that a 1 or 2 is the upface. Tough problem, right? The answer is $2/6 = 1/3 = .333$. But this is a simple problem with which to demonstrate resampling. Here's the first 3 steps of the resampling experiment:

1. Use random single digit random numbers 0 through 9. Discard (actually skip) digits 0, 7, 8, and 9.

2. The event A is a 1 or 2.

3. Pick at random a starting point in the 10 digit random number table given in Appendix B. This is the first outcome. Read the succeeding outcomes **one after another** going down that column to the end. Then move to the top of the next column and continue until we have N trials.

Notice how explicit we were in describing how to do the N trials (Step 3). Notice that it ensures independent and identical trials (the digits in the table are random). This is a MUST! Failure to do so results in **GIGO**.

Lets do 30 repetitions of this experiment. We will use the table of random numbers. To make sure we are all on the same wavelength, I will use numbers in the first column, starting at the top. Remember to skip the digits 0, 7, 8, and 9.

Here are 30 trials:

```
5 5 1 4 3 2 6 2 2 2 4 1 6 1 6 3 6 3 6 5 6 3 4 1 4 2 6 2 1 4
```

Notice that a 1 or 2 came up 11 times. Hence our estimate of the probability of a 1 or 2 is $11/30 = .3667$. Close to the true value.

Hey, we are on a roll! Lets try the urn problem of the previous chapter. Tough problem, but here is a resampling model:

1. Choose two digit random numbers, 00 through 99. Discard 00 and 81 through 99. The numbers 01 through 30 represent a blue ball while the numbers 31 through 80 represent a red ball. Select 3 numbers and discard ties (Here the problem is sampling without replacement).

2. If we get 3 numbers from 01 through 30 then 3 blue ball were obtained and if we get 3 numbers 31 through 80 then 3 red balls were obtained. In either case, 3 of the same color occurred. Count these up.

3. Pick at random a starting point in the 10 digit random number table. Use 2 columns. This is the first outcome. Read the succeeding outcomes **one after another** going down that those 2 columns to the end. Then move to the top of the next 2 columns and continue until we have N trials.

Lets obtain 30 repetitions of this experiment. We will use the table of random numbers. To make sure we are all on the same wavelength, I will use numbers in the first 2 columns, starting at the top.

```
59, 58, 12;   02, 41, 30;   29, 60, 20;  01, 21, 04;  07, 24, 06;  42, 15, 65;
19, 09, 06 ;  66, 38, 63;   31, 61, 55;  63, 73, 30;  47, 15, 49;  25, 62, 29;
75, 18, 48;   60, 53, 25;   29, 53, 21.
```

Lets turn them into colored balls:

```
R, R, B;      B, R, B;      B, R, B;      B, B, B;      B, B, B;      R, B, R;
B, B, B ;     R, R, R;      R, R, R;      R, R, B;      R, B, R;      B, R, B;
R, B, R;      R, R, B;      B, R, B.
```

So our estimate of the probability that all the balls are of the same color is: 5/15.

What's the error here? In the next two chapters, we will consider this in some detail. But for now, lets just state the error as follows. Denote our estimate of the probability of interest by $\hat{p}$. It is read "p hat". Then our error of estimation is

$$2\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

Notice that the error decreases proportionally by $\sqrt{N}$; hence, the more repetitions the smaller the error. For the urn problem, $\hat{p} = .3333$ and the error is 0.2434262. Notice that the interval $(\hat{p} - \text{error}, \hat{p} + \text{error})$ traps the true probability of .2880. This error is huge, because N is so small. Alas, I got very bored doing 15 repetitions of this experiment. But guess what? Yep, you got it. The computer will not get bored doing 10,000 reps. In which case the error is about 0.0091. (I used the correct value .2880 for this calculation. In practice, use the estimate $\hat{p}$).

Exercise 3.1.1

1. *Paula has 6 pairs of earrings in a box. She grabs two of the earrings in the box (sampling without replacement). Find the probability that she has a matched set of earrings.*

   Using the random number table, model this problem. (Hint: Use 0,1 for first pair; 2,3 for second pair; etc. Now the length of the trial is 2 (that's all she grabs and remember it's sampling without replacement).

   Next resample 10 trials of your model. For each trial record success (got a matched pair) or failure (did not get a matched pair). Obtain $\hat{p}$ your estimate of the desired probability. Calculate the error of estimation.

2. *When his alarm goes off, John hits the snooze button on it 80% of the time. If he fails to hit it, he gets up. The snooze alarm only works for 6 hits. Find the probability that John sleeps at least an extra 20 minutes.*

   Using the random number table, model this problem. (Hint: Let 1-8 denote John hitting the button and 0,9 denote he doesn't. Note that the length of the trial is either 6 or when the first 0 or 9 occurs before 6.)

Next resample 10 trials of your model. For each trial record the extra sleep John got (for example, suppose the trial is 4, 6 ,9. Then John slept for an extra 20 minutes which is a success for the event we want). Obtain $\hat{p}$ your estimate of the desired probability. Calculate the error of estimation.

3. *20 passengers are on a bus that enters a foreign country. 12 of these passengers are women. At the gate to the foreign country, a guard gets on the bus and selects 6 people at random for an extensive visa check. Find the probability that (a) all 6 are males. Find the probability that (b) all 6 are females. Find the probability that (c) 4 are females.*

   Using the random number table, model this problem.

   Next resample 10 trials of your model. For each trial record the success or failure for each of (a), (b), and (c). Obtain $\hat{p}$ your estimate of the desired probability for each event. Calculate the error of estimation.

4. *Betty is playing 5 card draw poker. She holds 3 hearts and 2 clubs. In the draw, she decides to discard her 2 clubs and get two more cards. Find the probability that she will get a flush in hearts, i.e., her 2 cards in the draw are hearts.*

   Using the random number table, model this problem.

   Next resample 10 trials of your model. For each trial record the success or failure for the desired event. Obtain $\hat{p}$ your estimate of the desired probability. Calculate the error of estimation.

5. *Jack pays \$10 to play a dice game in which 5 fair dice are rolled. If the dice result in:*

   (a) All dice come up 6, Jack wins \$500.

   (b) All dice are the same, Jack wins \$100.

   (c) All dice are even, Jack wins \$20.

   (d) Else Jack wins nothing.

   Find the probability that Jack wins some money.

   Using the random number table, model this problem.

   Next resample 10 trials of your model. For each trial record the success or failure for the desired event. Obtain $\hat{p}$ your estimate of the desired probability. Calculate the error of estimation.

3.2 Class Code for Resampling

To effectively use resampling to estimate probabilities of desired events requires many trials, say 1000 to 10,000. No sane person is going to do this with a random number table, but again the computer will not get bored doing 10,000 trials. Further, by setting up the model as we have been doing we do have a correct algorithm for the resampling. We could try to learn to code these algorithms into a computer program. This would necessitate the learning of a computer program which is not the purpose of this class.

So what can we do? Setting up the model as we have been doing is the most important thing here. If you can set up the model correctly then you **understand** the problem. This is the **important idea**. Thinking back on the problems we have solved, the tedious thing here is using the random number table to do the trials. So we have constructed class code that will do all this work for us. It requires input but if have modeled the problem correctly this will be easy.

Lets go back to our simple example in the last section. Here's the problem and our resampling model:

On the roll of a **fair** 6 sided die, determine the probability that a 1 or 2 is the upface. Here's the first 3 steps of the resampling experiment:

1. Use single digit random numbers 0 through 9. Discard (actually skip) digits 0, 7, 8, and 9.
2. The event A is a 1 or 2.
3. Pick at random a starting point in the 10 digit random number table. This is the first outcome. Read the succeeding outcomes **one after another** going down that column to the end. Then move to the top of the next column and continue until we have N trials.

Now lets obtain 20 trials of our resampling experiment, using the class code. We need the following input:

1. Number of trials: lets just do 20 the first time.
2. Minimum value of desired random numbers: 1.
3. Maximum value of desired random numbers: 6.
4. Number to be drawn (length of the trial): 1.
5. With or Without Replacement: With Replacement (although, since the length is one it doesn't matter).

Simply click on the class code (Random number generation for resampling trials) and input these items.

What did you get? Here's what I got. Note that the output is simple: the trial number followed by the outcome of the trial (in this case the upface of a fair die). Our results will differ, since the class code starts at a new place (based on the time of day) for each run.

```
Trial 1
3
Trial 2
6
Trial 3
1
Trial 4
1
Trial 5
6
Trial 6
3
Trial 7
4
Trial 8
5
Trial 9
4
Trial 10
4
Trial 11
1
Trial 12
1
Trial 13
3
Trial 14
6
Trial 15
6
Trial 16
2
Trial 17
```

```
1
Trial 18
1
Trial 19
4
Trial 20
3
```

I got 7 successes (a 1 or a 2) out of 20 trials. Hence my estimate of the probability of a 1 or a 2 is $\hat{p} = 7/20 = .35$ and my estimate of error is *.21* .

Note that we still have to examine the trials to see if the desired event came up or didn't. Hence, it is hard to see us doing 1000 trials to get a good estimate. But again, the main point is **SET UP A CORRECT MODEL** and if you input the right numbers and understand when the event occurs or doesn't occur on a trial then **YOU DO UNDERSTAND THE PROBLEM!**

Lets do the urn problem with the class code. Recall the problem and our resampling solution of the last section. Here is a resampling model:

1. Choose two digit random numbers, 00 through 99. Discard 00 and 81 through 99. The numbers 01 through 30 represent a blue ball while the numbers 31 through 80 represent a red ball. Select 3 numbers and discard ties (sampling without replacement).

2. If we get 3 numbers from 01 through 30 then 3 blue ball were obtained and if we get 3 numbers 31 through 80 then 3 red balls were obtained. In either case, 3 of the same color occurred. Count these up.

3. Pick at random a starting point in the 10 digit random number table. Use 2 columns. This is the first outcome. Read the succeeding outcomes **one after another** going down those 2 columns to the end. Then move to the top of the next 2 columns and continue until we have N trials.

The input for 20 trials via class code is:

1. Number of trials: lets just do 20 the first time.

2. Minimum value of desired random numbers: 1.

3. Maximum value of desired random numbers: 80.

4. Number to be drawn (length of the trial): 3.

5. With or Without Replacement: Without Replacement.

Click on the class code and input these items.

What did you get? Here's what I got. Note that the output is simply the trial number followed by the outcome of the trial (in this case the three balls drawn).

```
Trial 1
31      73      79
Trial 2
1       30      80
Trial 3
15      42      65
Trial 4
52      53      61
Trial 5
30      46      54
Trial 6
17      24      76
Trial 7
10      34      52
Trial 8
69      74      77
Trial 9
2       18      47
Trial 10
4       32      59
Trial 11
24      26      80
Trial 12
1       22      42
Trial 13
33      48      65
Trial 14
42      48      70
Trial 15
30      65      77
Trial 16
30      67      71
Trial 17
2       24      48
Trial 18
```

```
9       32      77
Trial 19
42      65      79
Trial 20
18      59      70
```

Note that all red came up in trials 1, 4, 8, 13 ,14 and 19. All blue never came up. So the estimate of the desired probability is $\hat{p} = 6/20 = .3$ and the standard error of estimation is *.204*.

Exercise 3.2.1
This exercise uses the class code (Random number generation for resampling trials).

1. *Use the class code to obtain 20 trials of your resampling experiment for Problem #1 in the last set of exercises.*
2. *Use the class code to obtain 20 trials of your resampling experiment for Problem #2 in the last set of exercises.*
3. *Use the class code to obtain 20 trials of your resampling experiment for Problem #3 in the last set of exercises.*
4. *Use the class code to obtain 20 trials of your resampling experiment for Problem #4 in the last set of exercises.*
5. *Use the class code to obtain 20 trials of your resampling experiment for Problem #5 in the last set of exercises.*
6. *1000 parts are shipped into a factory. Your job is to obtain a random sample of 20 (without replacement) of these parts for inspection. If the parts are tagged 1001 through 2000, use the class code to obtain your sample.*
7. *For the last problem, suppose your quality control plan rejects the shipment, if 5 or more of the sampled parts are defective. Suppose that really 20% of the shipped parts are defective. Determine the probability of returning the lot using the quality control plan.*

   Estimate the desired probability by doing 30 resamplings.
8. *Same as the last problem but now only 10% are defective.*
9. *We can solve a problem we have been discussing (opening with a pair, in 5 card poker) but the counting is a bit tedious, (need to count by 13's fast). But if enough of you do, say, 5 poker hands we can combine the results. Use the numbers 1 through 52 to denote the cards. Let*

   - *1, 14, 27, 30 denote Ace.*

- *2, 15, 28, 31 denote a two.*
- *Etc.*

Now sample 5 numbers (length of trial) 1 through 52 without repacement. Do this 5 times (5 trials). Count as a success a pair (not 3 nor 4 of a kind).

Chapter 4

Discrete Populations (Probability Models)

4.1 Random Variables

Recall that a probability assigns numbers to events. But in many problems there are only a few events of interest and, furthermore, they can often be characterized in terms of a variable.

For example, in the first roll in the game of craps (roll a pair of dice) the events of interest are: the sum of upfaces is 2, or 3, or 4, ... , or 12. Hence, there are only 11 events of interest. If we let

$$X = \text{the sum of the upfaces}$$

then the events of interest can be expressed as: *X=2*, *X=3*, ..., or *X=12*. Hence, X characterizes the events of interest. We call X a **random variable**.

As another problem, reconsider the urn problem where the urn contained 30 blue balls and 50 red balls and that 3 of these balls were selected at random without replacement. Recall we wanted to determine the probability that all the balls were of the same color. Just let

$$X = \text{the number of blue balls in the sample of size 3}$$

then the event of interest is *X =0 or X = 3*. Hence, X characterizes the events of interest.

The **range** of a random variable is the set of values it can assume. For example in the game of craps, the range of X is {2, 3, ... ,12} while in the urn problem, the range of X is {0, 1, 2, 3}. As another example, let X be the height of an adult male in inches. It is hard, even, impossible to come up with minimum or maximum of X ; hence, a convenient range is the interval $(0, \infty)$. This seems odd at first, but keep in mind we are trying to *model* height. Actually the best model of height employs a range of $(-\infty, \infty)$. We will discuss this later.

Essentially, random variables come in two types: **discrete** and **continuous** random variables. A discrete random variable has a finite (or listable) range. The range of a continuous random variable is an interval of numbers. In the first two examples, the random variables are discrete while in the last example on height, the random variable is continuous.

Exercise 4.1.1

1. *Let X denote the number of aces in a 2 card hand drawn without replacement fro a standard deck 0f 52 cards. What is the range of X? Is it discrete or continuous?*
2. *In the last problem, let Y denote the average area of the two dealt cards. Assume that we can measure area infinitely precise. What is the range of Y? Is it discrete or continuous?*

3. *In the urn problem discussed above, let Z denote the number of red balls in the sample (without replacement) of size 3. What is the range of Z? Is it discrete or continuous?*

4. *Let X denote the temperature at noon in Kalamazoo in centigrade. What is the range of X? Is it discrete or continuous?*

5. *Let X denote the number of people in a queue at a bank teller's window. What is the range of X? Is it discrete or continuous?*

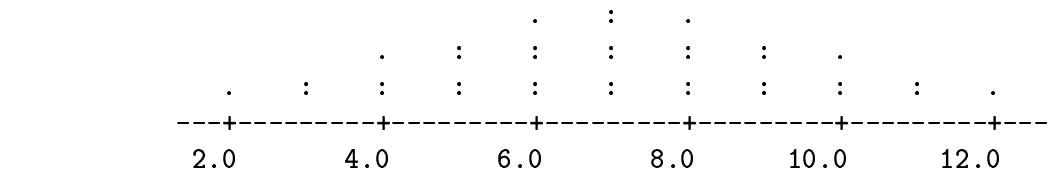
4.2 Discrete Populations (Probability Models)

As discussed above **discrete random variables** have a finite (or listable) range. What are their probability models? It's easy. In fact, you knew this before taking this course. Right! The **probability model of a discrete random variable** is its range and associated probabilities.

For example, in the first roll in the game of craps (roll a pair of dice) let X = the sum of the upfaces. Then the range of X is *2, 3, 4, ... ,12*. Now **ASSUME** that the dice are fair. Upon recalling the picture of the sample space, it is easy to determine the probability model of X. For example, the probability that $X = 3$ means the probability that a *(1,2)* or a *(2,1)* comes up which is *(1/36) + (1/36) = 2/36*. Using the same reasoning for the other range items, we obtain the probability model for X:

Range	2	3	4	5	6	7	8	9	10	11	12
Probabilities	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Instead of a table, how about a picture of the probability model? Just plot the probabilities (vertical) versus the range (horizontal). Here is a crude plot:



Alas, two more items. A little notation here is useful. For a discrete random variable X, let *p(x)* denote the probability that X assumes the value x. Often *p(x)* is called the **probability mass function.** For example, in the dice problem above. We will denote the probability that X is 7 by *p(7)*; hence, *p(7) = 6/36*.

Note that, in general for any discrete random variable, *p(x)* is a fraction and the sum of all the *p(x)* (over the range of X) is 1.

Although the term probability model makes sense here, it is often not used in practice. Usually we call the probability model of X, the **distribution** of X. It is confusing since we used the term distribution with sampling distributions of Chapter 1. We will sort this out later.

Exercise 4.2.1

1. *Let X denote the number spun on a fair spinner with the numbers 1, 2, and 3 on it. Determine the probability model of X.*

2. *In the last problem, suppose we spin the the spinner twice. Let S be the sum of the numbers spun. The range of S is 2, 3, 4, 5, 6. Use a tree diagram to determine the probability model of S.*

3. *Repeat the last problem if the spinner is spun 3 times.*

4. *Let X denote the number of aces in a 2 card hand drawn without replacement from a well shuffled standard deck 0f 52 cards. Then the range of X is $\{0, 1, 2\}$. Use a tree diagram to determine the probability model of X.*

5. *Repeat the last problem under sampling with replacement.*

6. *Let X denote the number of hearts in a 2 card hand drawn without replacement from a well shuffled standard deck 0f 52 cards. Then the range of X is $\{0, 1, 2\}$. Use a tree diagram to determine the probability model of X.*

7. *In the urn problem (with the balls well mixed) discussed above, let Z denote the number of red balls in the sample (without replacement) of size 3. Use a tree diagram to determine the probability model of Z. What is the range of Z? Is it discrete or continuous?*

4.3 Parameters

It is important to see the distinction between a probability model and a sample drawn from it. In this course, for the most part, we will be dealing with a sample. But this sample is **generated** by a probability model. In this section, we will discuss this for a very simple model. This also motivates **parameters** which are characteristics of the probability model.

Here is the probability model. Consider the fair spinner with the numbers 1, 2, and 3 on it. Let X denote the number spun. Then the probability model for X is

Range	1	2	3
Probabilities	1/3	1/3	1/3

In practice we won't know the probability model for X. In this case, we won't know if the spinner is fair or not. But in practice we can take a sample from the probability model. Based on the sample, perhaps we can say something about the probability model. Now it is very important that the sample is a **random sample**.

A sample is a **random sample** if:

1. The items in the sample are drawn independently of one another.
2. Conditions do not change as the sample is drawn.

In this case, the spins of the spinner are independent of one another and we are not changing the chances of a 1, 2 or 3 from spin to spin.

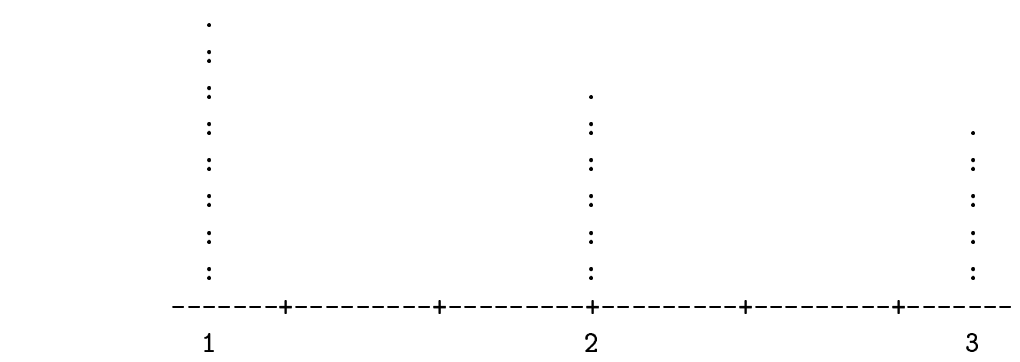
Here is a Sample drawn from the probability model. Suppose we decide on a sample size of 100. The following is a sample of the probability model; i.e., I spun a fair spinner 100 times: So here are the results of **my** random sample:

2	2	1	1	2	2	1	2	1	1	1	1	1	1	1
1	2	3	1	3	1	3	1	1	2	2	3	3	2	1
2	3	2	3	3	1	2	1	2	1	3	2	1	3	3
1	2	2	2	1	3	3	1	1	1	1	3	1	1	1
1	2	1	3	2	2	1	2	2	2	3	2	3	2	3
3	3	3	1	1	1	1	2	2	1	1	3	2	3	2
1	1	3	2	1	3	3	1	1	2					

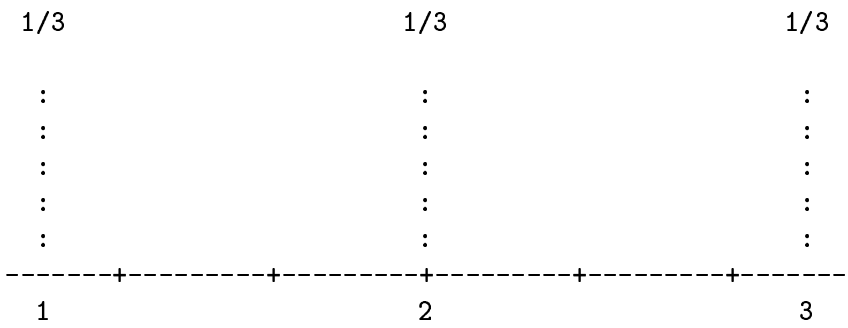
We of course tally up the sample and get the sample distribution:

Range	1	2	3
Frequency	43	31	26
Relative Frequency	.43	.31	.26

This sample distribution is an **estimate** of the probability model for X. That is, .43 is our estimate of *p(1)*, .31 is our estimate of *p(2)*, and .26 is our estimate of *p(3)*. The histogram of the sample



is our estimate of the graph of the probability model



What's that you're thinking? It seems a little off if the spinner is *fair*? Be careful, you are starting to think statistically. You may be even thinking of a formal test statistic to see if such a sample could be generated by the fair spinner probability model. Hey, we'll get there soon!

Suppose you compute the sample mean of sample distribution. You get the value $\bar{x} = 183/100 = 1.83$. Now since the histogram is an estimate of the probability model, what is 1.83 an estimate of?

It's not hard to see. There are two ways to calculate $\bar{x}$, here. One way is to add up the 100 numbers and divide by the sample size 100. However, in adding up these numbers you have added 3 to itself 26 times. Hence a much easier way is to use the tallying and add as follows:

$$\bar{x} = \frac{(1 \times 43) + (2 \times 31) + (3 \times 26)}{100}$$

$$\bar{x} = \left(1 \times \frac{43}{100}\right) + \left(2 \times \frac{31}{100}\right) + \left(3 \times \frac{26}{100}\right)$$

By the last line it is easy to see what $\bar{x}$ is estimating. Again, .43 is our estimate of *p(1)*, .31 is our estimate of *p(2)*, and .26 is our estimate of *p(3)*. Hence, $\bar{x}$ is estimating

$$\mu = \big(1 \times p(1)\big) + \big(2 \times p(2)\big) + \big(3 \times p(3)\big)$$

That's the Greek letter *mu*. μ is called the **mean** of the probability model. It is the center of gravity along the horizontal axis of the graph of the probability model. So in this example 1.83 is an estimate of μ.

This estimate is the result of one sample! If I spin the spinner another 100 times, I am going to get a different estimate of μ. If you spin it 100 times you are going to get a different estimate, too. In fact, if everyone in class spins the spinner 100 times we are going to get different estimates, (there may be a few ties because the probability model is discrete).

So the important thing to determine is: "How much does $\bar{x}$ miss μ by?" That's the way to think. Keep it up!

In the spinner example, if the spinner is *fair* then $p(1) = p(2) = p(3) = 1/3$ and $\mu = 2$. So $\bar{x} = 1.83$ is an estimate of $\mu = 2$, in this case. We missed by .17.

In practice, we will not know the population mean. But, hopefully, we will have a random sample. We will calculate $\bar{x}$. We will estimate with a degree of confidence, "How much does $\bar{x}$ miss μ by?"

In general, the probability model mean, μ, is called a **parameter** of the probability model. For a discrete random variable X, to determine the mean, as in the spinner problem, we simply cross multiply the range values by the associate probabilities and total it up; that is,

$$\mu = \text{Sum}\{x \times p(x)\}, \qquad \text{over } x \text{ in the range of } X.$$

Exercise 4.3.1

1. *Let S denote the sum of two numbers spun on a fair spinner with the numbers 1, 2, and 3 on it. The range of S is 2, 3, 4, 5, 6. Determine the probability model* **mean** *of S.*
2. *Repeat the last problem if the spinner is spun 3 times.*
3. *Let X denote the number of aces in a 2 card hand drawn without replacement from a well shuffled standard deck 0f 52 cards. Then the range of X is {0, 1, 2}. Determine the probability model* **mean** *of X.*
4. *Repeat the last problem under sampling with replacement.*
5. *Let X denote the number of hearts in a 2 card hand drawn without replacement from a well shuffled standard deck 0f 52 cards. Then the range of X is {0, 1, 2}. Determine the probability model* **mean** *of X.*
6. *In the urn problem (with the balls well mixed) discussed above, let Z denote the number of red balls in the sample (without replacement) of size 3. Determine the probability model* **mean** *of Z.*

4.4 More Parameters

There are several other parameters that we need to mention. The first is the probability model **variance.** To understand this parameter, return to spinner example of the last section. Recall we had spun a fair spinner 100 times. This resulted in our sample. Recall that we computed the sample mean and this motivated the probability model mean μ. Suppose next we calculate the sample variance s^2. But we can do this easy using the tallied sample results as follows:

$$s^2 = \left((1-1.83)^2 \times \frac{43}{100}\right) + \left((2-1.83)^2 \times \frac{31}{100}\right) + \left((3-1.83)^2 \times \frac{26}{100}\right) = .6611$$

Actually I should have divided by 99 instead of 100, but with n so large it won't matter much. By the last line it is easy to see what s^2 is estimating. Again, .43 is our estimate of *p(1)*, .31 is our estimate of *p(2)*, .26 is our estimate of *p(3)*, and 1.83 estimates μ. Hence, s^2 is estimating

$$\sigma^2 = \big((1-2)^2 \times p(1)\big) + \big((2-2)^2 \times p(2)\big) + \big((3-2)^2 \times p(3)\big)$$

That's the Greek letter *sigma*. σ^2 is called the **probability model variance** and its square root σ is called the **probability model standard deviation**. It is the center of gravity along the horizontal axis of the graph of the probability model. So in this example, assuming the spinner is fair, .6611 is an estimate of $\sigma^2 = 2/3$ and its square root, .8131, is an estimate of $\sigma = \sqrt{2/3} = .8165$.

Three other parameters of interest are the median and quartiles of the probability model. These are used more for the continuous probability models, so we will present them later.

Exercise 4.4.1

1. *Let S denote the sum of two numbers spun on a fair spinner with the numbers 1, 2, and 3 on it. The range of S is 2, 3, 4, 5, 6. Determine the probability model* **variance** *of S.*
2. *Let X denote the number of aces in a 2 card hand drawn without replacement from a well shuffled standard deck 0f 52 cards. Then the range of X is {0, 1, 2}. Determine the probability model* **variance** *of X.*
3. *Repeat the last problem under sampling with replacement.*
4. *In the urn problem (with the balls well mixed) discussed above, let Z denote the number of red balls in the sample (without replacement) of size 3. Determine the probability model* **variance** *of Z.*

4.5 Binomial Probability Model

The binomial probability model offers a simple but very useful model. An example of this model is the number of heads on 20 flips of a coin. This is certainly simple, but how about: The number of people answering yes to the question, "Do you like the way the president is doing his job?" Both of these are binomial. The second is of interest monthly (if not daily) in the US.

A binomial model is characterized by trials which either end in success (heads) or failure (tails). These are sometimes called **Bernoulli** trials.

Suppose we have n Bernoulli trials and p is the probability of success on a trial. Then this is a **binomial model** if

1. The Bernoulli trials are independent of one another.

2. The probability of success, p, remains the same from trial to trial.

Don't those two assumptions look familiar? They should! This is nothing more than the rules for a random sample applied to a particular case; i.e., the sample items are independent of one another and conditions don't change from sample item to sample item.

The **binomial random variable**, X, is just the number of successes in the n trials. Over the n trials, there could be one success, two successes, etc., up to n successes. So the range of X is the set $\{0, 1, 2, \ldots, n\}$. We will often write X is *bin(n,p),* which is read "X is binomial n, p". We can determine (obtain an explicit formula) for the probability model of X.

For this class, we have written some class code to obtain these probabilities. An example will demonstrate it. Suppose we want the probability of getting 7 heads in ten flips of a fair coin. That is, X is *bin(10,.5)* and we want $P(X = 7)$. In class code the input is:

1. $k = 7$

2. $p = .5$

3. $n = 10$

Choose **probability** from the analysis menu, select the appropriate probability distribution and enter these values. You should get the results:

```
P( X = k)      P( X <= k)
0.1171875      0.9453125
```

Hence, the probability of getting 7 heads in 10 flips of a fair coin is .1171875. Also, the probability of getting at most 7 heads in 10 flips of a fair coin is .9453125. At most 7 heads in 10 flips is the same event as 7 or less heads. These later probabilities will be useful.

As another example, suppose we have a fair spinner with the numbers 1 through 10 on it. Suppose success is a 1 or 2 or 3, while 4 through 10 are failures; i.e., we win if 3 or less is spun. Now suppose the spinner is spun six times. Let X be the number of times we win. Then X is *bin(6,.3)*. Lets determine the distribution of X by using the probability module. It's easy. Our input is *n=6*, $p = .3$ and we let k vary from 0 through 6. Try it! You'll get (rounded to 4 places):

range: x	0	1	2	3	4	5	6
p(x)	0.1176	.3025	.3241	.1852	.0595	.0102	.0072

Notice how the distribution peaks around 2 and then decreases. This will always happen for a binomial.

In general, for a *bin(n,p)* the probabilities of a binomial will increase until np and then decrease. The probability distribution will be symmetric if $p = 1/2$, skewed right if $p < 1/2$ and skewed left if $p > 1/2$. See the exercises for examples.

Further, the mean of a *bin(n,p)* is $\mu = np$ and the variance is $\sigma^2 = np(1-p)$.

Exercise 4.5.1

1. *Find the probability of getting 10 heads on 20 flips of a fair coin.*
2. *Find the probability of 5 aces (1's) on 5 rolls of a die.*
3. *Jack reports he has ESP. To prove it, he states the color (red or black) of a card drawn at random from a deck of 52 cards. He does this for 30 cards (with replacement). Suppose he is correct on 18 of the cards. He states, "See, I got more than half correct? " What do you think? One way of reasoning here is: If Jack is just guessing how odd is it that he gets 18 or more correct out of 30?; i.e., obtain the probability that $X \geq 18$ when X is bin(30,.5).*
4. *Same as last problem but this time JACK gets 24 out of 30 correct.*
5. *Clyde hits 70% of his free throws in basketball. Determine the probability that Clyde makes 8 out of 14 free throws.*
6. *In the last problem, Clyde plays a game in which he sinks only 4 out of 17 free throws. The coach benches him the next game, saying "Clyde, you're slipping." Clyde says, "Hey, coach it's just a bad night." To which the coach says, "A pretty rare night Clyde." Who is right? Consider the probability that $X \leq 4$ when X is bin(17,.7).*

7. *Obtain and comment on the distribution of a binomial probability model with* $n = 6$ *and* $p = .5$.

8. *Obtain and comment on the distribution of a binomial probability model with* $n = 6$ *and* $p = .7$.

4.6 Poisson Probability Model

Another discrete probability model of interest is the **Poisson**. Actually an understanding of the Poisson will help you get through the frustration of queues. You know, "When I came into this store there was no one in the line (queue) at the fast checkout, but when I went to check out there were 6 people ahead of me."

Poisson random variables deal with counts of events over time.

1. The number of customers entering a deli over, say, noon hour.
2. The number of calls entering a switch board from 9 to 10 in the morning.
3. The number of tornadoes which touchdown in Kalamazoo County in July.
4. The number of plane crashes in one year.

These are all examples of random variables which are counts of events over time.

There are two axioms for a Poisson process.

1. In a small interval of time, the probability of an event occurring is λ times the length of the interval of time. Further, the probability that two or more events occur is practically 0.
2. If two intervals of time do not overlap, the occurrence or non occurrence of events in these intervals are independent of one another.

Consider a bank. It seems reasonable, that two customers will generally not enter the bank simultaneously, (i.e., in a short interval of time at most one). And the longer the short interval of time, the more probable it becomes that some customer will enter. Further the customers that enter between 9 and 10 are independent of those entering between 10 and 11.

These are assumptions and for a given situation they may or may not be true. For example, if a motorist is driving to a drive-in bank window and he sees a long queue then he may decide to do his banking later; i.e., dependence broke down. Actually these assumptions never truly hold but often the approximation is close to reality and predictions based on the model are often fairly accurate.

Let X denote the number of events in one unit of time. Then under these assumptions we can obtain the probability model for X. The range of X is $\{0, 1, 2, 3, \ldots\}$ and its probability distribution can be obtained. The mean of X is λ and its variance is also λ.

As with the binomial, we use the probability module to obtain the probabilities. To obtain the probability that $X = k$, the input consists of k and λ.

Chapter 5

Continuous Probability Models

5.1 Uniform Probability Model

The probability model of a discrete random variable was evident to you before you took this course. Sure, the notation is new but in playing games like craps you knew the probabilities of interest; eg, probability of a "7". Recall that in general, the probability model of a discrete random variable with range $\{1, 2, ..., k\}$ consists of probabilities $P(X = i)$ for $i = 1, 2, ..., k$. The probability of a continuous random variable is not as evident. Lets begin with an example where we know the answers. This will motivate the continuous probability model.

Suppose we choose a real number at random between 0 and 1. Let X be the number chosen. Then X is a continuous random variable with the interval *(0,1)* as its range. Certain probabilities are obvious here:

1. The probability that X is between 0 and 1/2 is 1/2.
2. The probability that X is between 1/2 and 1 is 1/2.
3. The probability that X is between 0 and 1/4 is 1/4.
4. The probability that X is between 1/2 and 3/4 is 1/4.
5. The probability that X is between 1/8 and 2/8 is 1/8.

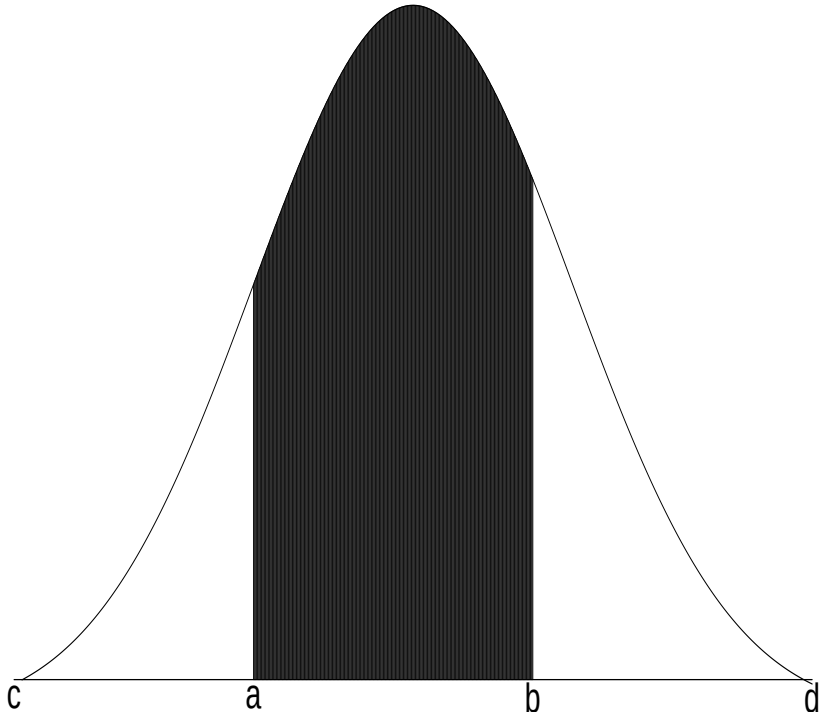
Are you ready for the big jump? What's the probability that X is between a and b when a and b are real numbers between 0 and 1? It's b - a, the length of the interval. Go back to the list above and check if this isn't so for those cases. This leads to the following, though:

1. The probability that X is between 1/4 and 3/4 is 1/2.
2. The probability that X is between 3/8 and 5/8 is 1/4.
3. The probability that X is between 7/16 and 9/16 is 1/8.
4. The probability that X is between 15/32 and 17/32 is 1/16.

Note that we could continue this list forever. Each of the above intervals contains the number 1/2 and that further the length of each succeeding interval is getting smaller. Hence, we must have *P(X = 1/2) = 0.* But this is true for any real number a between 0 and 1; i.e., *P(X = a) = 0.* In general, for continuous random variables the discrete probability model will not work. But the probabilities of intervals are the probabilities of interest and this is how we define the model.

For a continuous random variable X whose range is the interval *(c,d)* the **probability model** of X is a curve *f(x)* such that the probability that X is between a and b is the area under the curve

Figure 5.1: Area under the curve



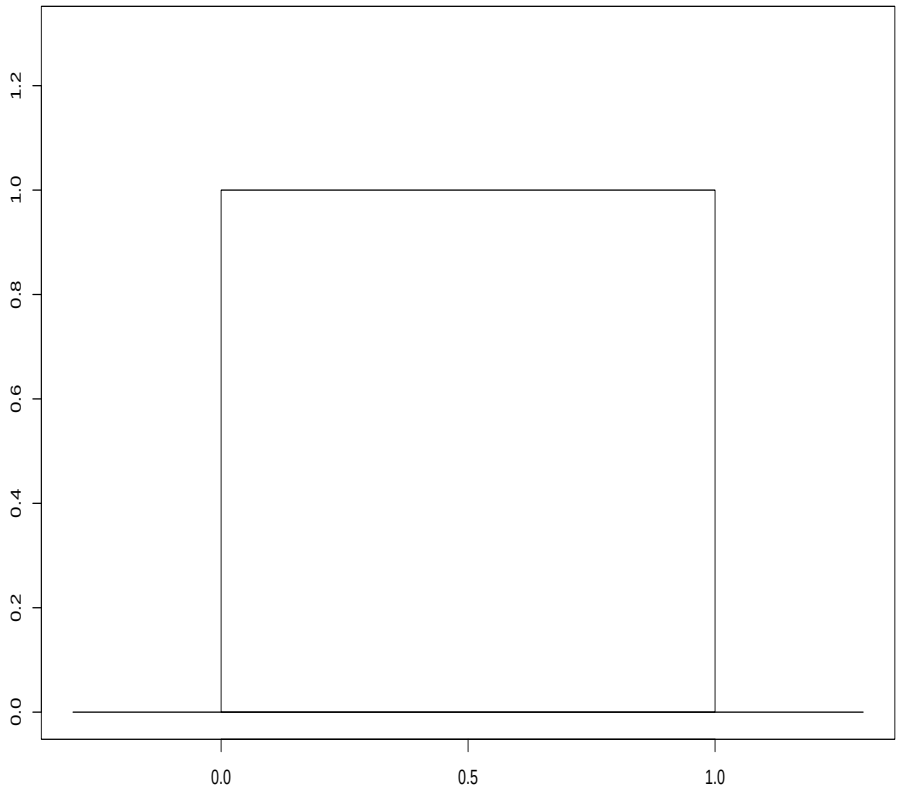
between a and b, that is, for some function *f(x)* the $P(a < X < b)$ is the area under the curve as shown in Figure 5.1

Notice that *f(x)* cannot be negative and that the total area under the curve must be one.

Consider the above example where X is a number chosen at random between 0 and 1. If we draw a straight line with slope 0 and height 1 above the interval *(0,1)*, then the area under this line over the interval *(a,b)* is $(b - a) \times 1 = b - a$, which is our desired probability. The curve is

given in Figure 5.2. This is called the **uniform probability model**.

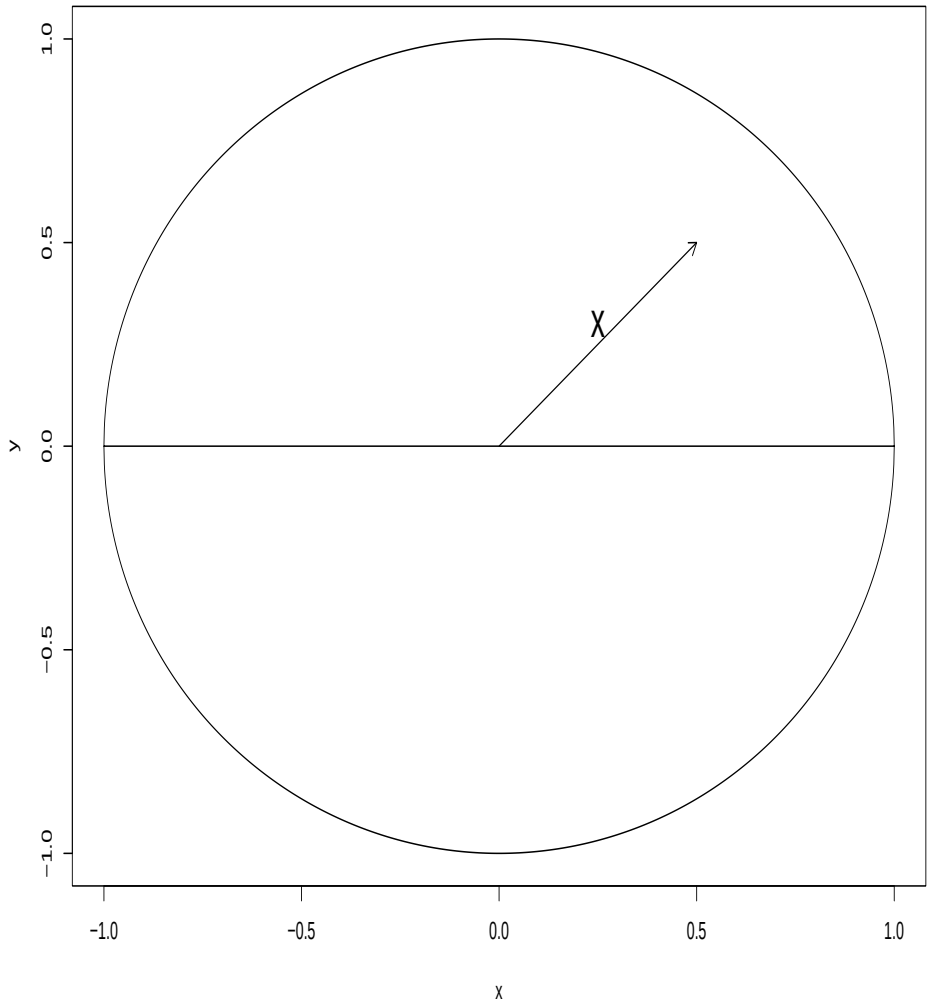
Figure 5.2: Uniform(0,1)



Here's a second example. Suppose we choose a point at random inside the unit circle, a circle with radius 1 and center at the origin, (sketch it!).

Let X be the distance between the chosen point and the origin, (sketch it!). See my sketch in Figure 5.3. Then the range of X is between 0 and 1, just like the uniform. But the probabilities are unlike the uniform. For example, it is much more likely that X is between 3/4 and 1 than between 0 and 1/4, (Why? Sketch it!). In fact, you can show that the probability model for X is a line over

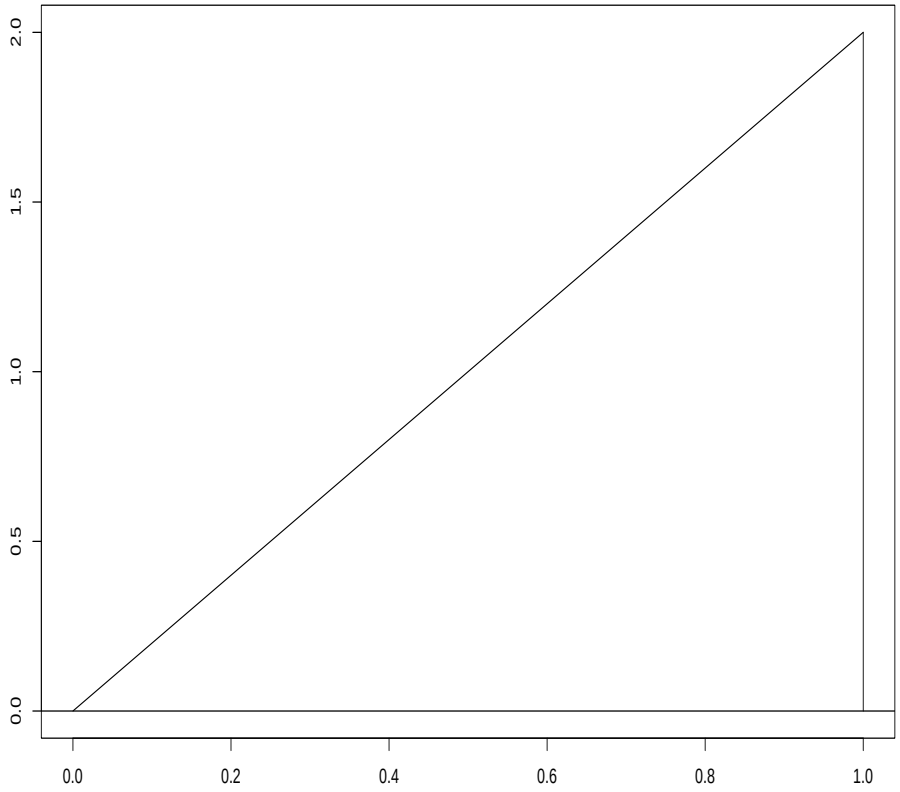
Figure 5.3: Circle with radius 1



the interval *(0,1)* with intercept at the origin and slope 2 (a triangle!, sketch it!). The graph is

Recall that the area of a triangle is $\frac{1}{2}\times$ base $\times$ height. Show that the area of the triangle is 1.

Figure 5.4: Triangular distribution



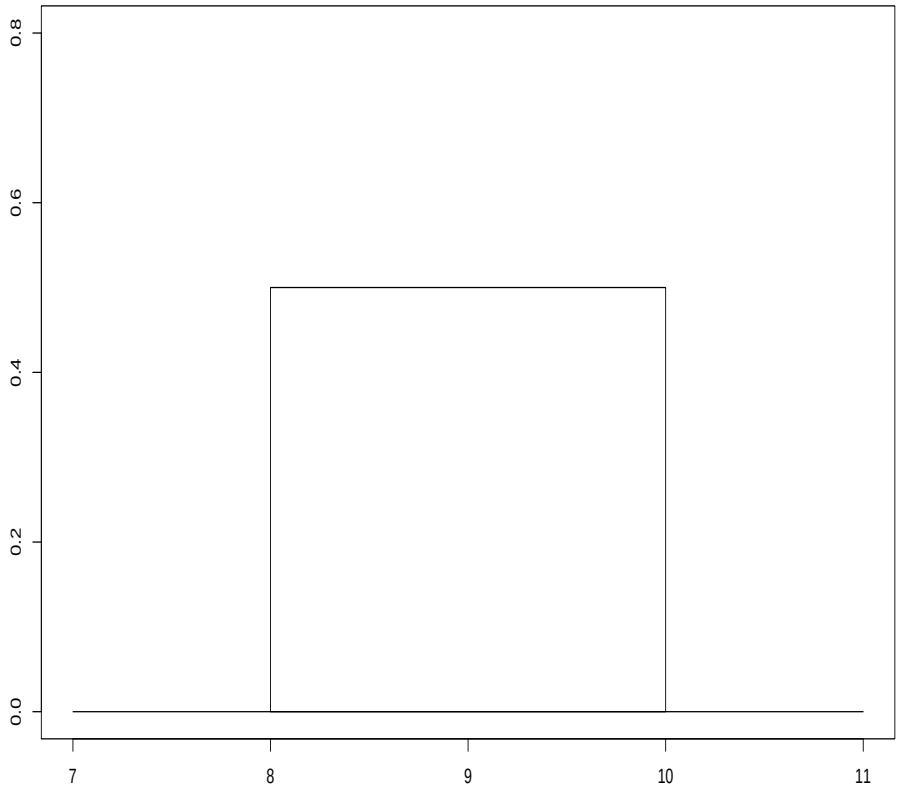
Next shade in the area under the curve from 1/4 to 3/4. Determine this area. You have just found $P(1/4 < X < 3/4)$.

Exercise 5.1.1

1. *Let X be a number chosen at random between 8 and 10. Determine the probabilities that X is between 8.5 and 9.5 and X is between 8.5 and 8.7. Determine the probability that X is*

between a and b. Verify that the probability model for X has the graph given in Figure 5.5.

Figure 5.5: Uniform(8,10)

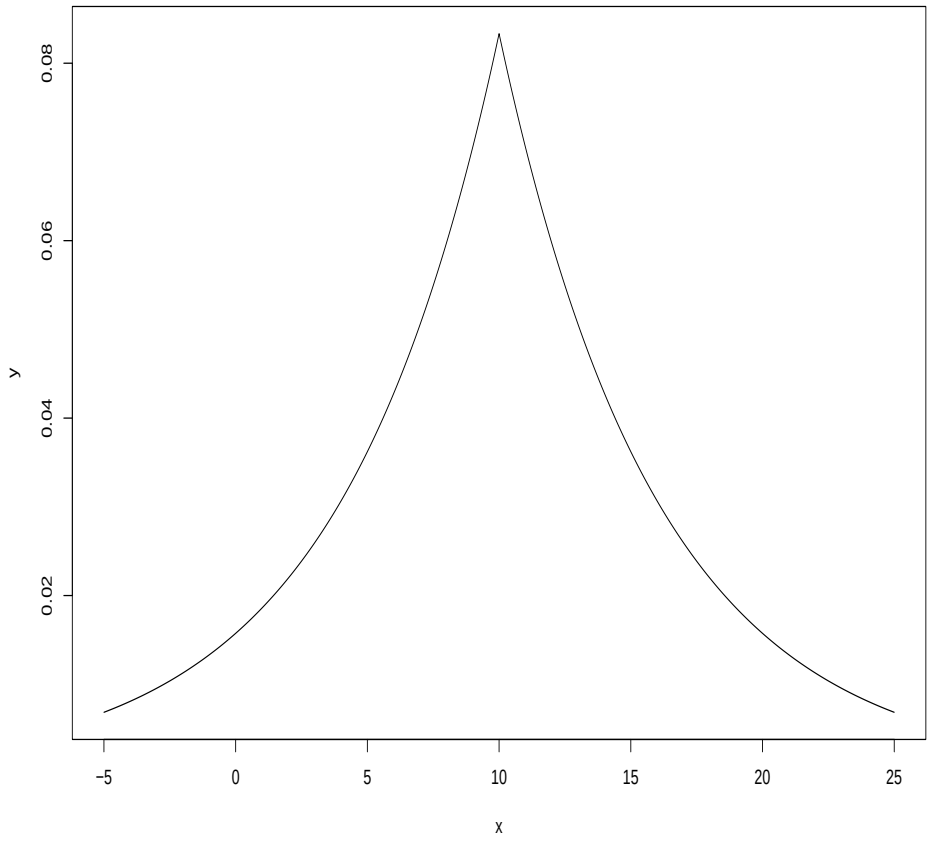


2. *For the second example above, find the probability that X is between 0 and 1/2.*
3. *For the second example above, find the probability that X is between 3/4 and 1.*
4. *For the second example above, find the probability that X is between 0 and 1/4.*
5. *For the second example above, (choose a point at random inside the unit circle), find c so that the probability that X is between 0 and c is 1/2. (Hint Sketch it!).*

5.2 Parameters

There are several parameters associated with a continuous probability model that will prove useful. These are measures for probability models, so we will classify them into location and scale. To visualize these parameters, consider the probability model in Figure 5.6.

Figure 5.6: A continuous probability model



While reading the material below try to locate the parameters on it. Some of these will be "guesses". The answers will be given at the end.

Also consider the following random sample from this model, (I obtained this sample).

24.0720	7.4773	18.4161	10.1440	10.4192	7.0660	3.9749
0.5231	7.9999	28.7453	6.0621	12.3729	5.3767	23.1134
7.9949	16.0966	10.4424	4.1915	8.4586	9.9292	

As we proceed, calculate the statistics based on this sample which are estimates of the parameters. In fact, the stem leaf plot or the histogram are estimates of the probability curve. These estimates (stem leaf, histogram) are poor because the sample is so small.

1. **Location Parameters**

   The **median**, θ, is the point located along the horizontal axis of the probability model which divides the probability mass in two. That is, half the time the variable is less than θ and half the time the variable is greater than θ. Okay! Locate the median on the probability model above.

   Calculate the sample median, Q_2. This is the estimate of the median you located on the probability model above.

   The **mean**, μ, is the center of gravity along the horizontal axis of the probability model. This is the point where the probability mass would balance. Obviously if the probability curve is symmetric, the mean and median would agree . Locate the mean on the probability model above.

2. **Scale or Noise Parameters**

   The **first quartile**, q_1, is the point located along the horizontal axis of the probability model which divides the probability mass into 1/4 (to the left of q_1) and 3/4 (to the right of q_1). Similarly, the **third quartile**, q_3, is the point located along the horizontal axis of the probability model which divides the probability mass into 3/4 (to the left of q_3) and 1/4 (to the right of q_3). Their difference, $iqr = q_3 - q_1$, is called the **interquartile range** of the probability model. The interquartile range is a parameter. The interquartile range is a scale or noise parameter. Locate the quartiles on your probability model above.

   A second scale parameter is the **population standard deviation**, σ . Recall that we gave a formula for it in the discrete case. We would have to use Calculus for the continuous case. But we can discuss it. The sample standard deviation, s, is an estimate of σ. Calculate your estimate.

Answers for the continuous probability model

Parameter	Parameter Value	Estimate Based on Sample
median	$\theta = 10$	9.19
mean	$\mu = 10$	11.14
First quartile	$q_1 = 5.84$	6.31
Third quartile	$q_3 = 14.16$	15.17
Interquartile range	$q_3 - q_1 = 8.32$	8.86
Standard deviation	$\sigma = 6$	7.37

5.3 Normal Distribution

One of the most important continuous probability models is the **normal probability model**. It is important because of the **Central Limit Theorem**, which we will discuss in the next chapter. Briefly, the Central Limit Theorem says that if you add up a bunch of random errors (independent errors under identical conditions) the distribution of this sum is approximately **bell shaped**.

In this section, we will discuss the model. For example, let the variable X be the height of an adult American male. Then X is approximately normally distributed. It is centered at 70" (i.e., the mean height is $\mu = 70$") and its standard deviation is 4", (i.e. $\sigma = 4$"). For this example, we say that X is normal with mean 70 and variance 16 and we will write it as *N(70,16)*. A picture of the distribution is given in Figure 5.7.

Suppose we want to determine the probability that a man is over 6 feet tall. That's easy. Just find the area under this curve between 72 and infinity. What's that, you didn't take calculus? Well no worries. We will often be finding probabilities like this so we have, of course, a class code to do so. There are two steps: make a *z-score* and then choose probability from the analysis menu.

A fact that we need here is that a random variable X is $N(\mu, \sigma^2)$ if and only if the random variable $Z = \frac{X-\mu}{\sigma}$ is $N(0, 1)$. Note that the distribution of Z does not depend on μ and σ.

To solve our problem, we only need to make a *z-score* for $X = 72$, which is *z = (72-70)/4 = .5*. Hence, the probability that $X > 72$ is the same as the probability that $Z > .5$. So we only need to compute the area under the distribution of Z from .5 to infinity. In terms of the distributions, we want the shaded area in Figure 5.8.

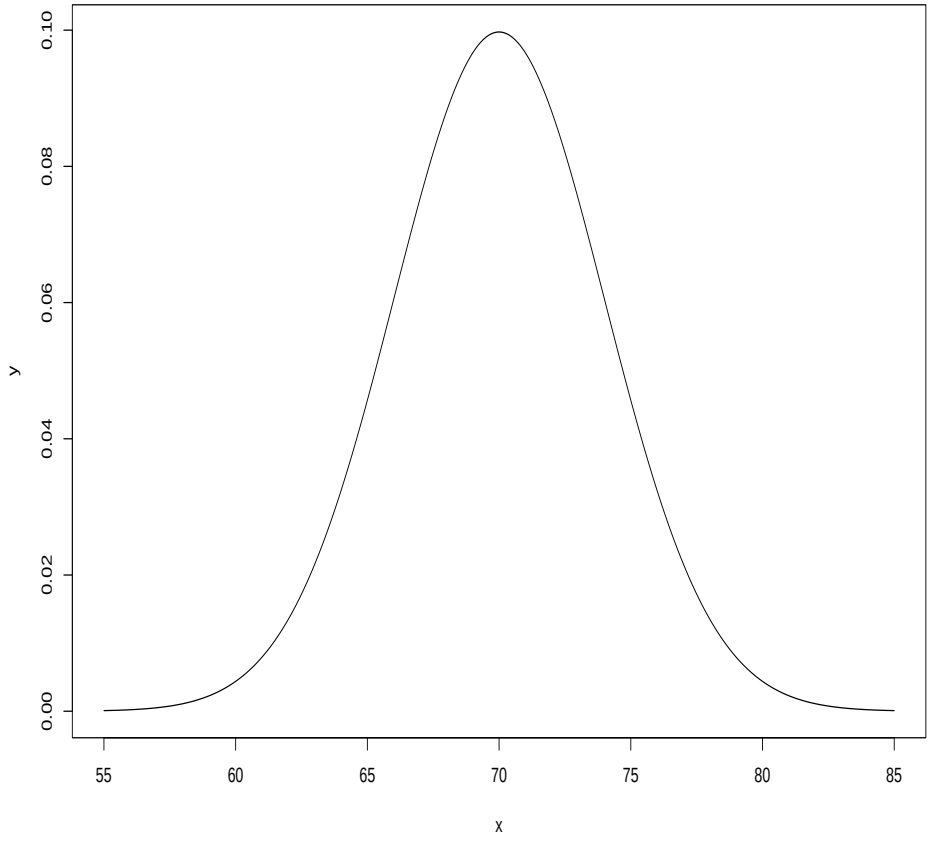
Actually the class code will give us the area to the left of .5. But recall that the total area under the curve is 1; hence, our answer is *1 - (the area to the left of .5)*.

To solve this problem with CC, just choose probability from the analysis modules. You want Cumulative Normal Probabilities and enter .5 at the k-window. Try it. Remember to get the answer subtract what you see from 1; i.e. the probability that a man is over 6 foot tall is *1-.6914624612740131 = .3086*.

Lost? Lets try another one. What's the probability that a man is between 66 and 77 inches tall? Compute the *z-scores* of 66 and 77. You will get *-1* and *1.75*. We want the shaded area in Figure 5.9.

Again go to CC and put in 1.75, record the answer. Then put in -1, record the answer. Subtract the second from the first and you have the probability that a man is be between 66 and 77 inches tall. Try it. Hence the answer is *.9599 - .1586 = .8013*.

Figure 5.7: A N(70,16) distribution

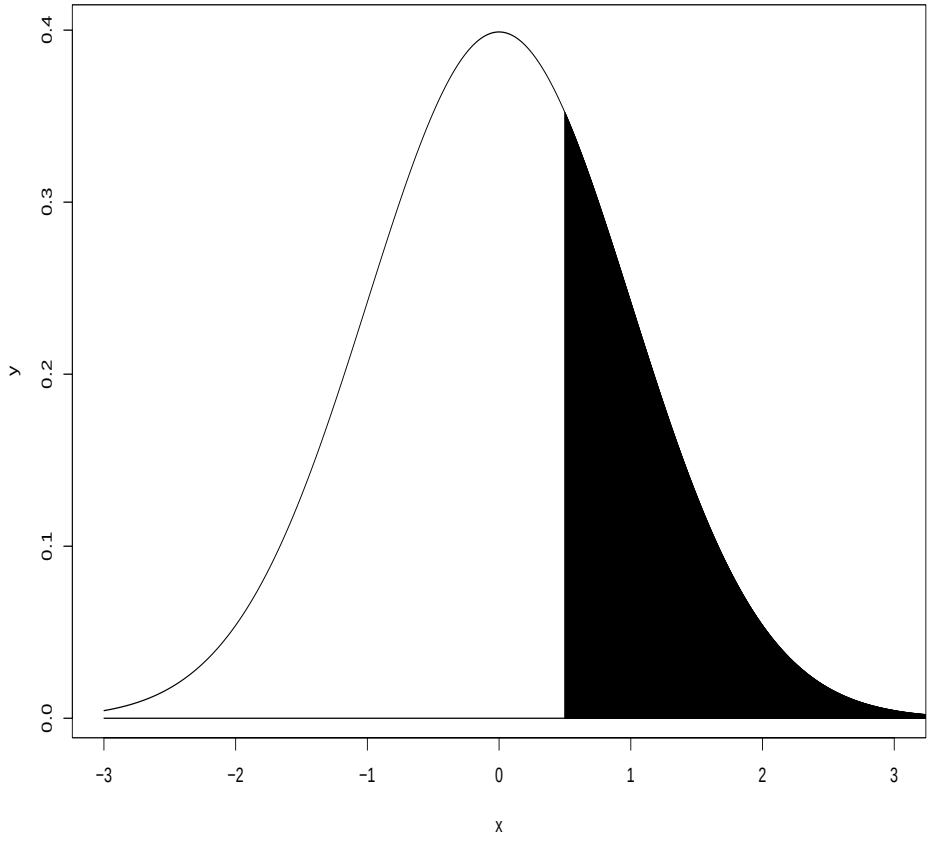


Exercise 5.3.1

1. *Consider the height example given above. Determine the probability that a man is between 63 and 71 inches tall. (Ans: 0.5586).*

2. *Suppose the passing grade on a standardized exam is 450. Suppose we know that scores on this exam are approximately normally distributed with mean 430 and standard deviation 50.*

Figure 5.8: $P(Z > .5)$



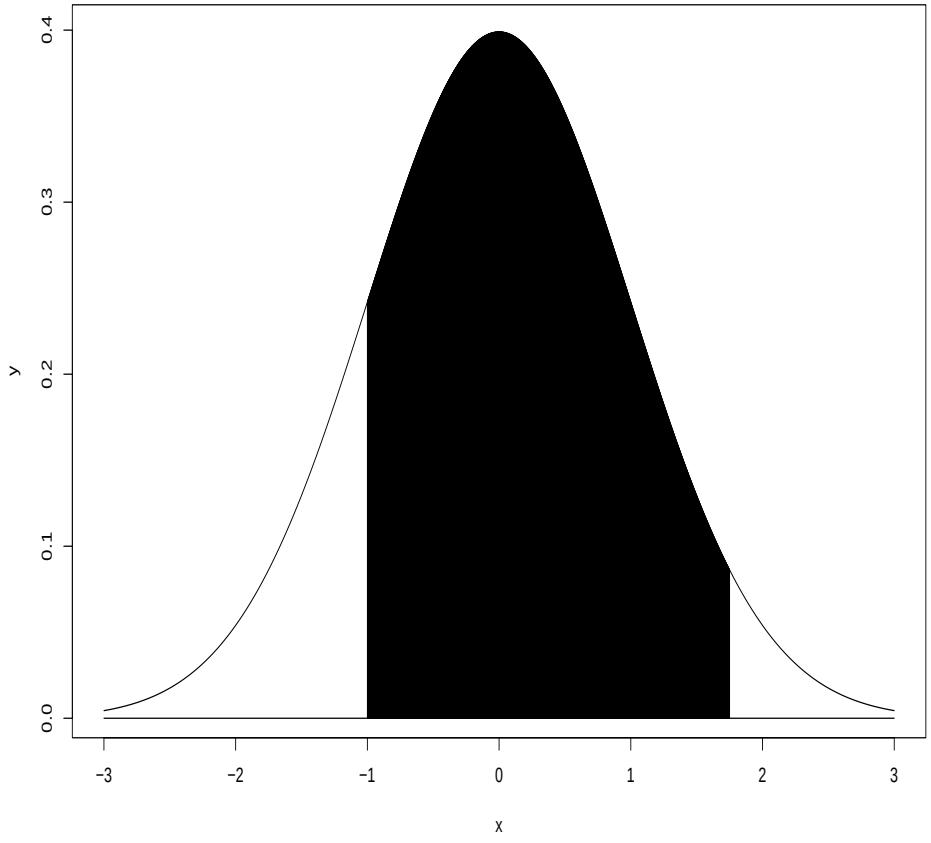
3. *Find the probability that a person passes the exam. (Ans:.3445).*

4. *Find the probability that a person scores over 500. (Ans: .0807).*

5. *Find the probability that a person scores between 400 and 480. (Ans: .5670) .*

6. *Suppose a part is acceptable only if it is less that .1" long. Suppose the lengths of these parts are approximately normally distributed with mean .09" and standard deviation .018". Find the probability that a part is acceptable. (Ans: .7107).*

Figure 5.9: $P(-1 < Z < 1.75)$



7. *In the last problem, suppose we make 20 of these parts. Find the probability that at least 10 are acceptable. (Ans: .9870).*

5.4 Normal Quantiles

Consider the above problem on height of a male. Assume that height is normally distributed with mean 70" and standard deviation 4". Suppose we want to determine the $90th$ percentile in height; i.e, the height which is surpassed by only 10% of men. Call this point c.

Well what can we do it? What's that? A *z-score* of course. Get back to z. Easy. $z = \frac{c-70}{4}$. We can't determine z, but the area to the left of z on the standard normal distribution curve is .90. See Figure 5.10

Now we need to find the value of z which corresponds to .90. It's easy. Just ask CC. This time when you click go to the Normal Percentage Point part of the page and enter .90 in the p-window. Then click submit and read the answer, which is 1.28. Try it.

Hence, by the above equation, $1.28 = (c - 70)/4$; i.e., $c = 75.12$. Therefore only 10% of men are taller than 75", (a very short basketball player).

How about finding the quartiles of the general normal distribution with mean μ and standard deviation σ? For the first quartile, say, q_1, we have $z = \frac{q_1-\mu}{\sigma}$ and the area to the left is .25. From the CC we get $z = -.67$, so that $q_1 = (-.67 \times \sigma) + \mu$. Now you try it for the third quartile, q_3. Another way is to use symmetry: the quartiles have got to be symmetric with respect to μ; hence $q_3 = (+.67 \times \sigma) + \mu$.

Thus the interquartile range for a normal distribution is

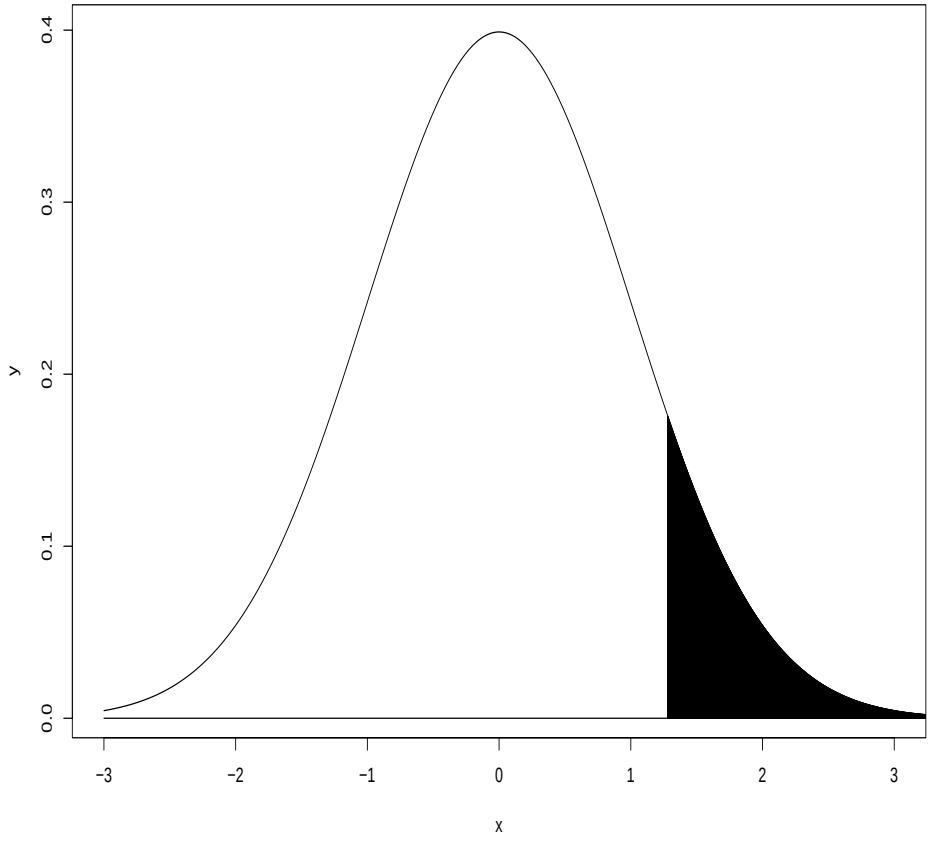
$$iqr = q_3 - q_1 = 2 \times .67 \times \sigma = 1.34 \times \sigma$$

i.e, the interquartile range for any normal distribution is $1.34 \times \sigma$.

Recall this is the **population** interquartile range. Hence, suppose I tell you the population is normal, but μ and σ are unknown. Suppose it is extremely important that you obtain an estimate of σ. And I give you 1000 data points to use in your estimation of σ. But then the batteries in your calculator died. And then your friend the computer whiz went to the movies. So what do you do? You don't want to compute the sample standard deviation. This is 1000 numbers! Alas! But then you notice that on the last sheet of the pages containing the data, someone has sketched a boxplot. You obtain the length of the box and divide it by 1.34. This is an estimate of σ. Why? Will it be a good estimate? What's that? How much did the estimate miss by? Hey, you catching on.

Exercise 5.4.1

Figure 5.10: $P(Z > 1.28) = .10$



1. *Consider the height example given above. Determine the first and third population quartiles. (Ans: 67.30204, 72.69796).*

2. *In the last problem, a basketball coach remarks that the even shortest professional basketball player, exceeds the 98th percentile in height. Does this make sense? (Help with answer: 98th percentile is 78.215).*

3. *Suppose we know that scores on this exam are approximately normally distributed with mean*

430 and standard deviation 50. Determine the first and third population quartiles. (Ans: 396.2755, 463.724 5).

4. *Smith College only accepts the upper 20% of people taking the exam in #3. What is the lowest score a person can make on the exam and still be acceptable to Smith College? (Ans. 472.0811).*

5. *Verify that for any normal population, the probability that a measurement falls in the interval* $\mu - 2\sigma$ *to* $\mu + 2\sigma$ *is .9544997.*

6. *Verify that for any normal population, the probability that a measurement falls in the interval* $\mu - 3\sigma$ *to* $\mu + 3\sigma$ *is .9973002.*

5.5 Empirical Rule

Empirical Rule : If the histogram of the data is approximately mound shaped then

1. About 68% of the data fall in the interval $\bar{X} - s$ to $\bar{X} + s$.
2. About 95% of the data fall in the interval $\bar{X} - 2s$ to $\bar{X} + 2s$.
3. About 99.5% of the data fall in the interval $\bar{X} - 3s$ to $\bar{X} + 3s$.

This is based on the normal distribution. Suppose X has a normal distribution with mean μ and standard deviation σ. To determine the $P(\mu - \sigma < X < \mu + \sigma)$, form the z-scores

$$\begin{aligned} Z_1 &= \frac{\mu - \sigma - \mu}{\sigma} = -1, \text{ and;} \\ Z_2 &= \frac{\mu + \sigma - \mu}{\sigma} = 1. \end{aligned}$$

Now call up the class code. The area under the curve to the left of $Z_1 = -1$ is .1586. The area to the left of $Z_2 = 1$ is .8413. Hence

$$P(\mu - \sigma < X < \mu + \sigma) = .8413 - .1586 = .6827$$

or about 68%. The other two parts of the rule are obtained in the problems.

Chapter 6

Central Limit Theorem

6.1 Some Probability Examples

Why are you taking this class? There are several reasons but **the main reason** is the **Central Limit Theorem**. It's the reason why statistics works. In this chapter we consider this theorem. In the rest of book, we will make use of it. We begin with a few probability examples which illustrate it.

Consider our favorite example: a fair spinner with the numbers 1, 2 and 3 on it. Let X be the number spun. Then the probability model for X is:

Range	1	2	3
Prob.	1/3	1/3	1/3

Better yet, a picture of it is found in Figure 6.1.

Note from the picture that the mean is 2; i.e., $\mu = 2$. Show that the standard deviation, σ, is $\sqrt{2/3}$. If you cannot show it, ask your instructor in class to show you.

Now suppose we spin it twice. Let S_2 be the sum of the numbers spun and let $\bar{X}_2$ be the average of the numbers spun. For example, if the numbers 1 and 3 are spun then $S_2 = 4$ and $\bar{X}_2 = 2$. What is the probability model for S_2? Just use a tree diagram (i.e., two spins, top branch is 1 and 1, etc.: SKETCH IT!). My sketch is given in Figure 6.2.

How about $\bar{X}_2$? Well that's just $S_2/2$. So the probabilities don't change, just the range values. If you have done your tree diagram correctly, the probability model for $\bar{X}_2$ is given by:

Range	1	3/2	2	5/2	3
Prob.	1/9	2/9	3/9	2/9	1/9

See Figure 6.3.

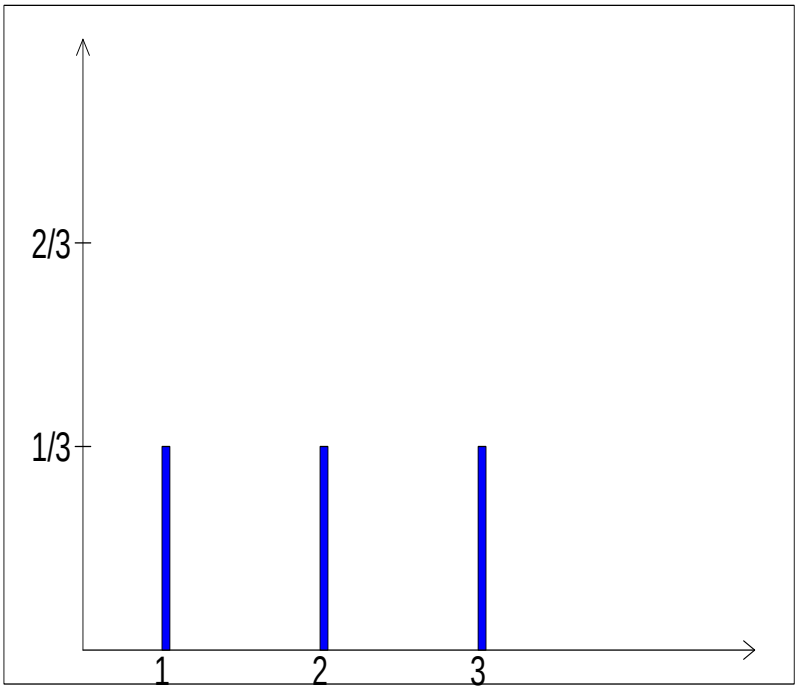
Note that the range of $\bar{X}_2$ is from 1 to 3, which is the same range as X. The probability distribution of $\bar{X}_2$ is quite different than that of X. The distribution of X is flat and uniform. On the other hand, the distribution of $\bar{X}_2$ is unimodal and mound shaped.

Note from the picture that the mean is 2; i.e., $\mu = 2$. You can show that $\sigma = \frac{\sqrt{2/3}}{\sqrt{2}}$.

Suppose we spin it three times. Let S_3 be the sum of the numbers spun and let $\bar{X}_3$ be the average of the numbers spun. For example, if the numbers 1, 3 and 3 are spun then $S_3 = 7$ and $\bar{X}_3 = 7/3$. What is the probability model for S_3? Just put the set of third branches on the above tree diagram. How about $\bar{X}_3$? Well that's just $S_3/3$. So the probabilities don't change, just the range values. If you have done your tree diagram correctly, the probability model for $\bar{X}_3$ is given by:

Range	1	4/3	5/3	6/3	7/3	8/3	3
Prob.	1/27	3/27	6/27	7/27	6/27	3/27	1/27

Figure 6.1: Probability model : spinner

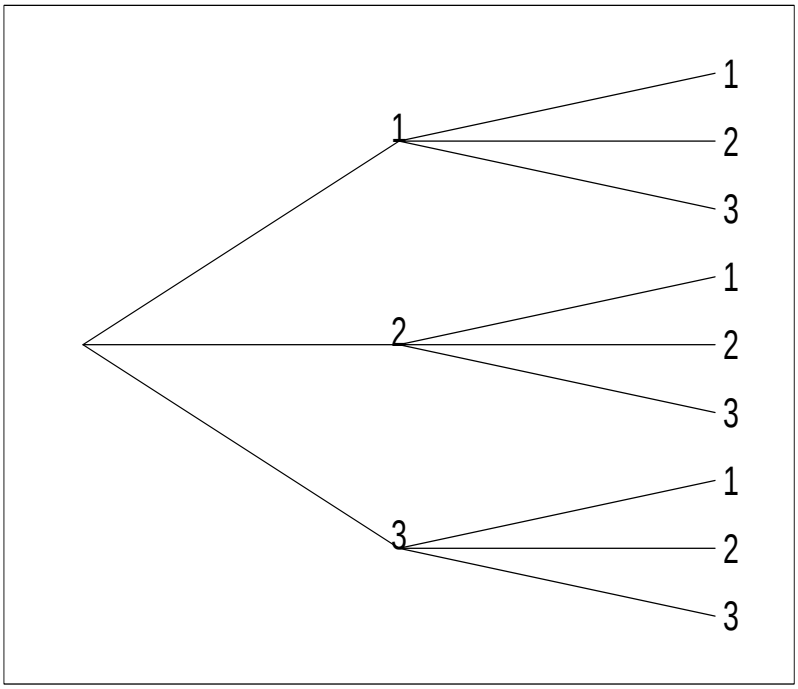


See Figure 6.4.

Note from the picture that the mean is 2; i.e., $\mu = 2$. You can show that $\sigma = \frac{\sqrt{2/3}}{\sqrt{3}}$.

Ready to jump? What's happening here? As the number of spins n increases, the mass is moving towards the center. It seems that $\bar{X}_n$ is more "likely" to be in the middle. The mean of $\bar{X}_n$ is *2*, the same mean as in the original model. The variance is, however, getting smaller. Again the mass is moving towards the center, the distribution of $\bar{X}_n$ is becoming less dispersed.

Figure 6.2: Probability model : S_2



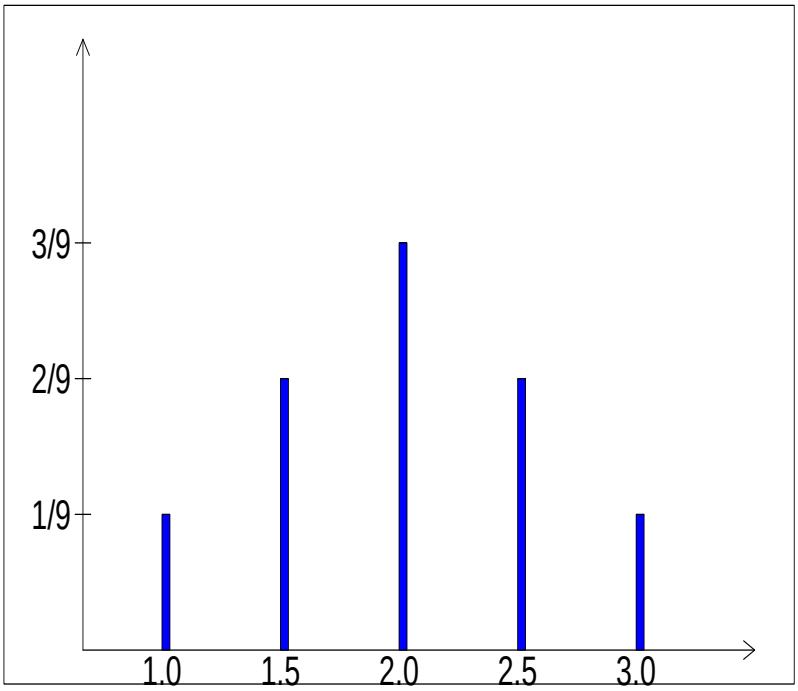
Lets try one more example.

Suppose 2 can never be spun. That is the distribution of X, the number spun, is

Range	1	2	3
Prob.	1/2	0	1/2

See Figure 6.5.

Figure 6.3: Probability model : $\bar{x}_2$

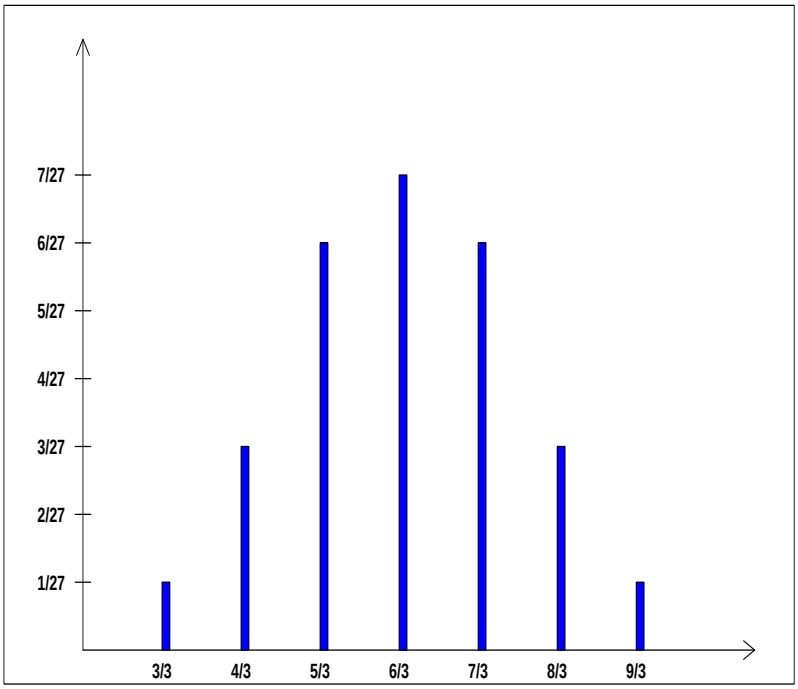


Note from the picture that the mean is 2; i.e., $\mu = 2$. You can show that $\sigma = \sqrt{5}$.

Now suppose we spin it twice. Let S_2 be the sum of the numbers spun and let $\bar{X}_2$ be the average of the numbers spun. What is the probability model for S_2? Just use a tree diagram (i.e., two spins, top branch is 1 and 1, etc.: SKETCH IT!). How about $\bar{X}_2$? Well that's just $S_2/2$. The distribution of $\bar{X}_2$ is given by:

Range	1	2	3

Figure 6.4: Probability model : $\bar{x}_3$



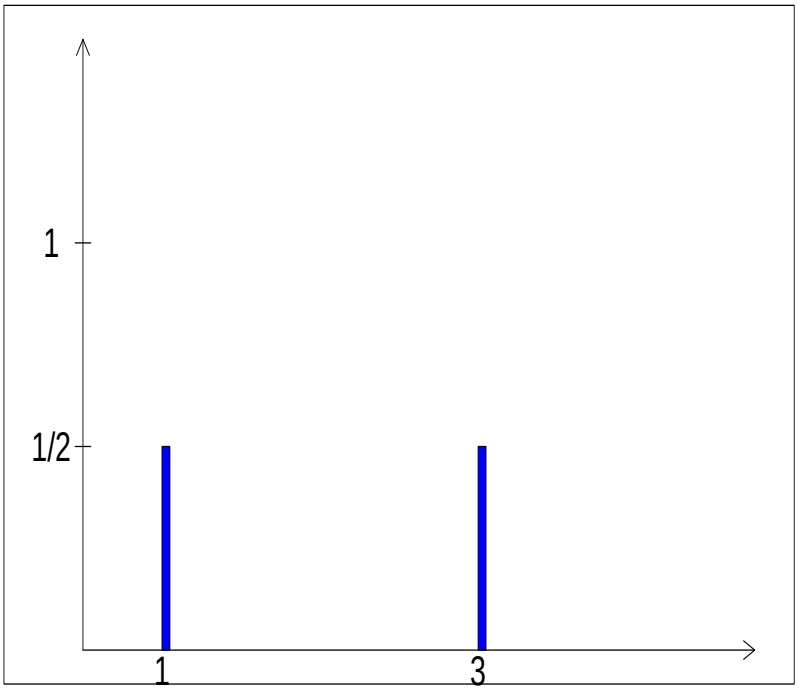

```
Prob.   1/4   2/4   1/4
```

Better yet, see Figure 6.6.

Even for this probability model, the mass for $\bar{X}_2$ is moving to the middle. The mean of the distribution of $\bar{X}_2$ and its standard deviation is $\sqrt{4.5}$.

Exercise 6.1.1

Figure 6.5: Probability model : Spinner with no 2



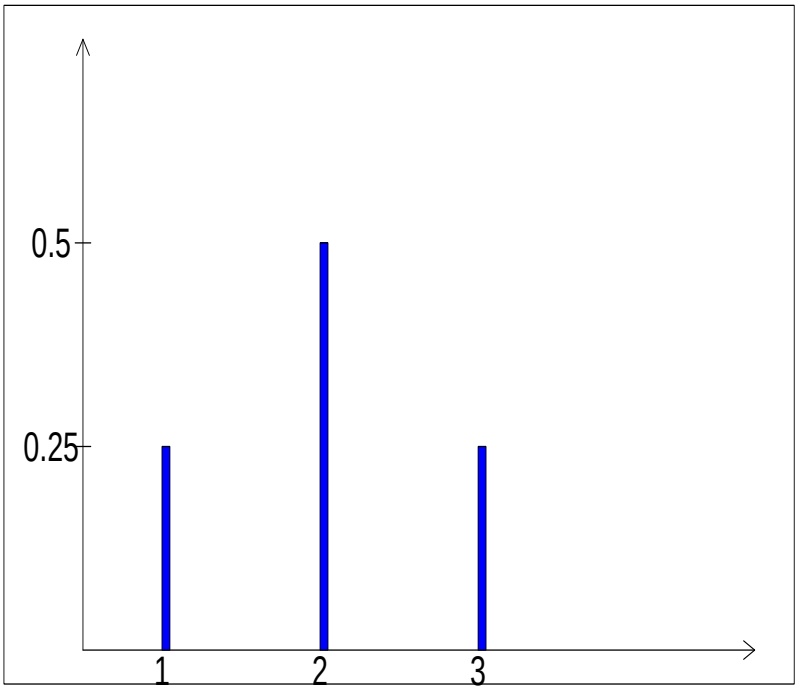
1. *Suppose the population has the distribution*

Range	0	1
Prob.	.3	.7

Let $\bar{X}_3$ be the sample average of a sample of size 3 from this population. Using a tree diagram, show that the distribution of $\bar{X}_3$ is

Range	0	1/3	2/3	1

Figure 6.6: Probability model : $\bar{x}_2$




```
Probs  0.027 0.189 0.441 0.343
```

2. *Show that the means of the population and the distribution of $\bar{X}_3$ have the same value .7.*
3. *Show that the variance of the population in #1 is .21.*
4. *Show that the variance of $\bar{X}_3$ in #1 is .07.*

6.2 Central Limit Theorem

The spins on the spinner were independent of one another and conditions (the probabilities) did not change from spin to spin. These are the two important ingredients to get the central limit effect.

In this class, we will be dealing with samples from a population. If the sample items are independent of one another and conditions remain the same while the sampling is being conducted then the distribution of the sample sum and sample average will be mound shaped as in our simple spinner example above. We state this next, with a little bit of notation:

Central Limit Theorem: Let $X_1, X_2, ..., X_n$ be a random sample from a population with mean μ and standard deviation σ. Let $\bar{X}$ be the sample average of $X_1, X_2, ..., X_n$. Then the distribution of $\bar{X}$ is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$.

Remarks on the Central Limit Theorem, (CLT).

1. The distribution of $\bar{X}$ approaches a normal distribution as n gets large. In this way, the approximation improves as n increases.
2. There are probability models for which the CLT does not hold but we will not encounter these situations in this course.
3. The sum of independent (or nearly independent) random variables is key to the CLT. This is how we will use it in later chapters.

Consider the height of an adult male. It's approximately normally distributed with mean 70 inches and standard deviation 4. A picture of the distribution (population) is given in Figure 6.7:

Based on the empirical rule, 95% of adult males will have heights in the interval $(62, 78)$.

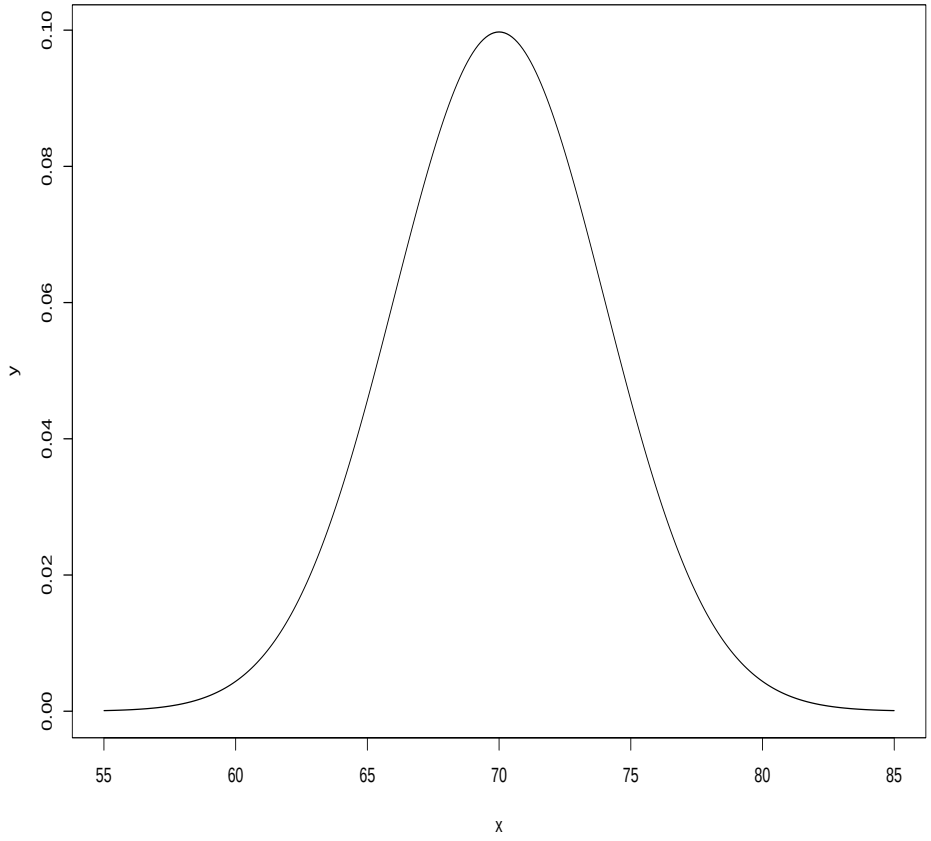
Next suppose we take a sample of 16 adult males. By the Central Limit Theorem $\bar{X}_16$ will be approximately normal with mean 70 and standard deviation $4/\sqrt{16} = 1$. A picture of this distribution is found in Figure 6.8.

Notice that this distribution is less variable than the original population. Based on the empirical rule, 95% of the time the average height of 16 adult males will fall in the interval $(68, 72)$.

Next suppose we take a sample of 64 adult males. By the Central Limit Theorem $\bar{X}_{64}$ will be approximately normal with mean 70 and standard deviation $4/\sqrt{64} = .5$. A picture of this distribution is found in Figure 6.9.

Notice that this distribution is less variable than the original population and the distribution of the average height of a sample of 16 adult males. Based on the empirical rule, 95% of the time the average height of 64 adult males will fall in the interval $(69, 71)$.

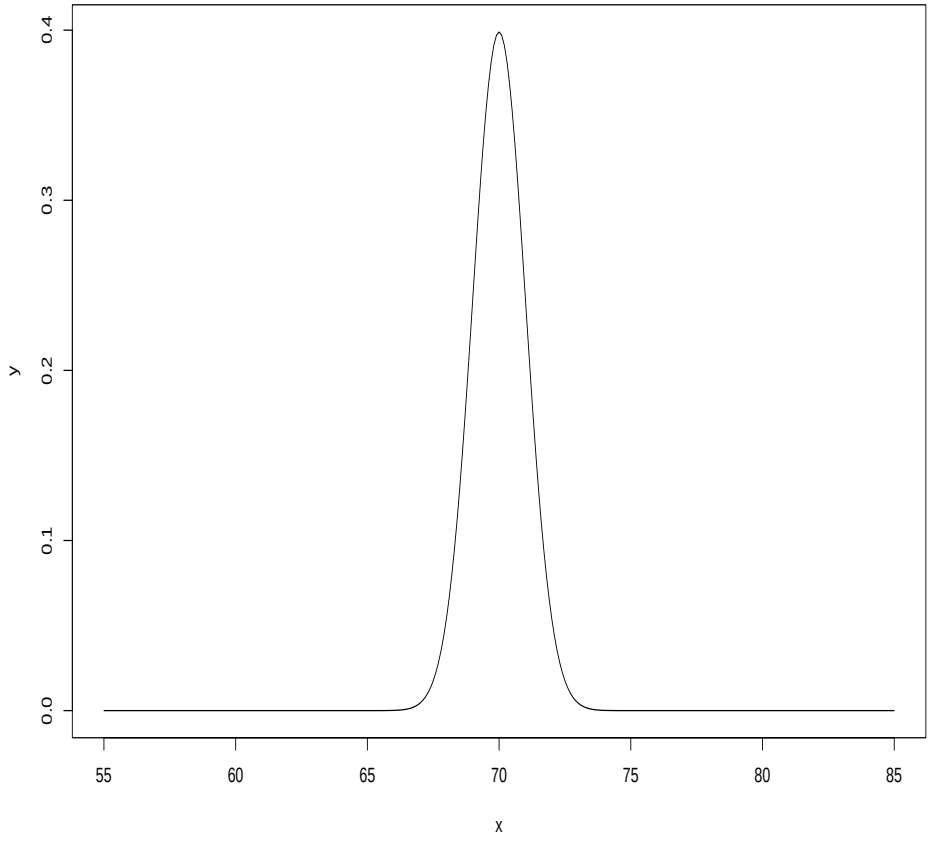
Figure 6.7: N(70, 16)



As the Central Limit Theorem dictates, as the sample gets large the distribution of the sample average becomes less and less variable (the noise is being cut by the $\sqrt{n}$, i.e. the standard deviation of the sample average is $\sigma/\sqrt{n}$.) Hence the sample average is getting closer to the population mean μ.

We will use the sample average to estimate μ. What's that you say? Speak louder. **It's not the estimate but how much it misses by!** Hey, you are right again. How much did it miss

Figure 6.8: N(70,1)



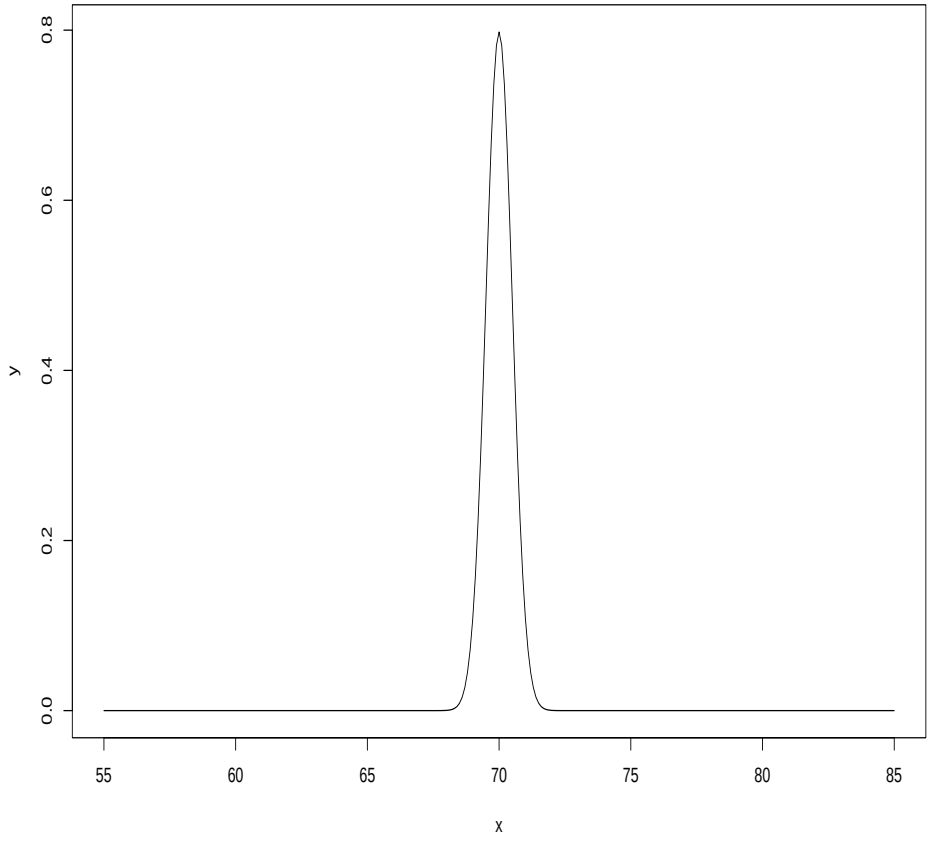
by? Hey, that's the topic of the next chapter. For now, lets solve a few interesting problems with the CLT.

Elevator Problem. Sixteen adult males approach an elevator on the 100th floor of Everett Tower. The elevator has the sign:

```
Maximum Weight 2900 lbs
```

They enter the elevator. Find the probability that the cable snaps and they plunge to their death;

Figure 6.9: N(70,.25)

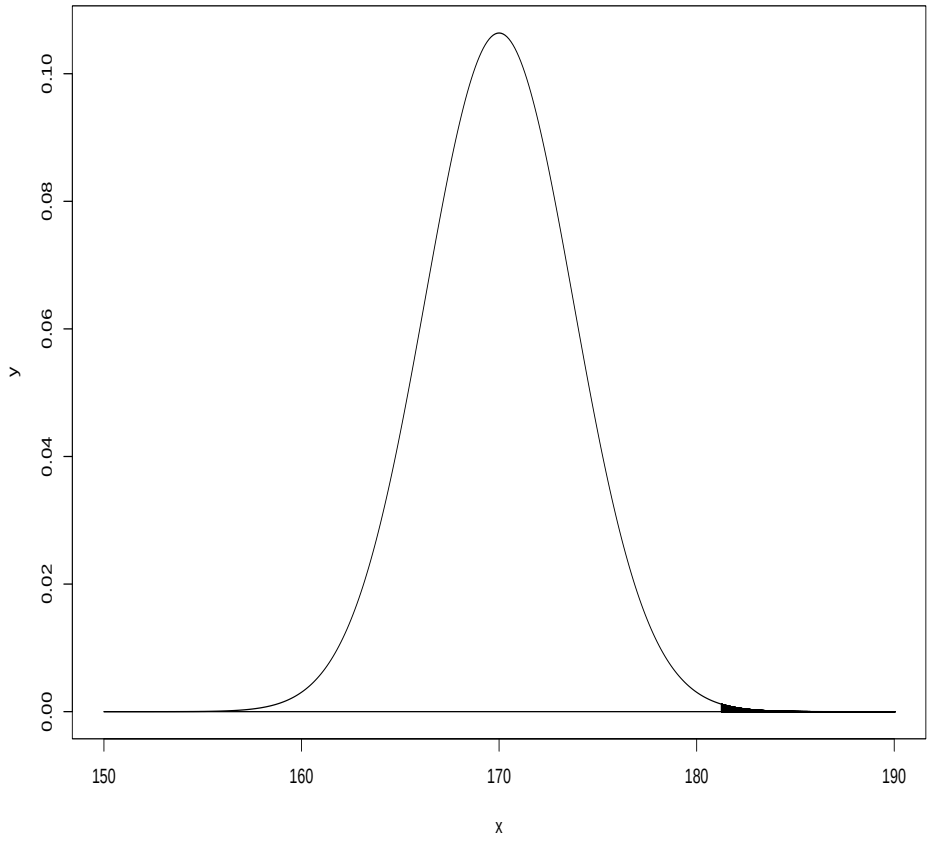


i.e., their combined weight exceeds 2900 pounds.

Looks hard but we can do it with the help of the CLT and Doctor Population. From Doctor Population we need the average weight and standard deviation of an adult male. There is plenty of information on this. So Doctor Population consults his blue book and tells us that the mean weight of an adult male is 170 pounds and the standard deviation is 15 pounds. This is all we need.

Since we have expressed the CLT for sample averages, first express the problem in terms of the sample average; i.e., find the probability that the sample average of 16 adult males exceeds $2900/16 = 181.25$. By the CLT, the distribution of the sample average is approximately normal with mean 170 and standard deviation $15/\sqrt{16} = 3.75$. The probability that we want is the shaded area in Figure 6.10.

Figure 6.10: $P(\bar{X} > 181.25)$



The z-score is $(181.25 - 170)/3.75 = 3$. Using the probability module, the probability that the the sample average of 16 adult males exceeds $2900/16 = 181.25$ is $1 - .9986 = .0014$. Hence *only*

once out of a 1000 times will the cable snap if 16 males enter the elevator at one time.

Note, we made certain assumptions to solve this problem. The 16 males should be independent of one another, no kin, friends, etc. They must be a random sample form the general population, no football team, etc.

As a final note. It is not odd for the weight of one adult male to exceed 181.25 pounds, (z-score is $z = \frac{181.25-170}{15} = .75$); hence (assuming an approximate normal distribution), the probability that the weight of one adult male exceeds 181.25 pounds is $1 - .7733 = .2367$. This happens 24% of the time. But it is odd that the average (based on a random sample) weight of 16 adult males exceeds 181.25 pounds.

Exercise 6.2.1

1. *The scores on a general test have mean 450 and standard deviation 50. It is highly desirable to score over 480 on this exam. A person can get into Smith's College prestigious MBA program if he/she scores over 480. In one location 25 people sign up to take the exam. The average score of these 25 people exceeds 490. Is this odd? Should the test center investigate? Answer on the basis of the CLT.*

2. *A machine fills cereal boxes at a factory. Due to an accumulation of small errors (different flakes sizes, etc.) it is thought that the amount of cereal in a box is normally distributed with mean 22 oz. for a supposedly 20 oz. box. Suppose the standard deviation of the amount filled is 1.3 oz. A federal regulatory selects four of these boxes at random and finds that the average content of these boxes is less than 18 oz. This official knows that the company claims the mean content to be 22 oz. He promptly fines the company. Who is right? Use the CLT in your answer.*

3. *Sixteen adult males are in a pit which is 98 feet deep. They decide to stand on one another (feet to head), hoping that the person on top can grip the top of the pit and get out, and, hence go for help. What's the probability that their plan succeeds? (Ans: .0003).*

Chapter 7

Confidence Intervals

7.1 Introduction

The beginning of a study is formed by a **question**. This question usually defines the **population** (or populations) of interest. If we knew the population then we could answer the question.

For example, a question that is of current interest on campus is, "Are students willing to buy a computer which will be paid for by an increase in tuition?" The population here is all students at WMU. Now if we knew the population we would know how many students (what proportion) are willing to have an increase in tuition to offset the purchase of a computer. And this could be done, but it will take time (read as MONEY) to gather this information.

In an attempt to answer this question we obtain a random sample from the population. By random sample (here we go again) we mean

1. **The items in the sample are independent of one another.**
2. **Conditions do not change as the sample is gathered.**

With regards to the question on the cost of a computer, we would select n students at random and ask them the above question. The number who answer yes divided by n would be our estimate of the true proportion. As usual, the question we must answer is: **How much did our estimate miss by?** That's the topic of this chapter.

Another question along these lines is: "What's the family income of a student (or more specifically, the income paying the student's tuition)?" We mean of course the population distribution of the family incomes of all the students. We could use our sample to estimate this. Actually the histogram of the family incomes of the sample students is our estimate of the population distribution of the family incomes of all the students. From working with histograms, though, you know that we need quite a large sample to estimate this population distribution accurately. We could rephrase the question as: "What's the **average** family income of a student (or more specifically, the **average** income paying the student's tuition) ?" Then our sample average would be an estimate of the population average. Again, the question we must answer is: **How much did our estimate miss by?** We need an estimate of error of estimation.

7.2 Confidence Intervals for Means

Lets pick on the mean, μ. That is, we have a population with **unknown** mean μ. So we take a random sample of size n from this distribution, say, $X_1, X_2, ..., X_n$. Then our estimate of μ is the sample average $\bar{X}$.

Income Example: Suppose we take a sample of 25 students from Smith University and record their family incomes. Suppose the incomes (in thousands of dollars) are:

```
28    29    35    42    42    44    50    52    54    56    59
78    84    90    95   101   108   116   121   122   133   150
158   167   235
```

The data have been sorted. So the lowest income is \$28,000 and the highest income is \$235,000. The average is (Either add up all the numbers or use the summary module) 89.96, i.e, about \$90,000. So now we need to determine how much our estimate missed by.

In general, our estimate of μ is $\bar{X}$. And we know something about the distribution of $\bar{X}$. The Central Limit Theorem tells us that the distribution of $\bar{X}$ is approximately normal with mean μ (the population mean) and standard deviation $\sigma/\sqrt{n}$, (σ is the population standard deviation). By the empirical rule, 95% of the time $\bar{X}$ falls in the interval $\mu - 1.96\frac{\sigma}{\sqrt{n}}$ to $\mu + 1.96\frac{\sigma}{\sqrt{n}}$, (1.96 is more accurate than 2 which we have been using). A picture of it is seen in Figure 7.1.

We need an interval which we are fairly confident contains μ. The interval in the above plot ($\mu - 1.96\frac{\sigma}{\sqrt{n}}, \mu + 1.96\frac{\sigma}{\sqrt{n}}$) occurs 95% of the time. **It's endpoints are the 2.5 and 97.5 percentiles of the distribution of $\bar{X}$.** But we can't use it because we don't know μ. Well if you don't know it, estimate it. Ignoring σ, consider the interval

$$(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}})$$

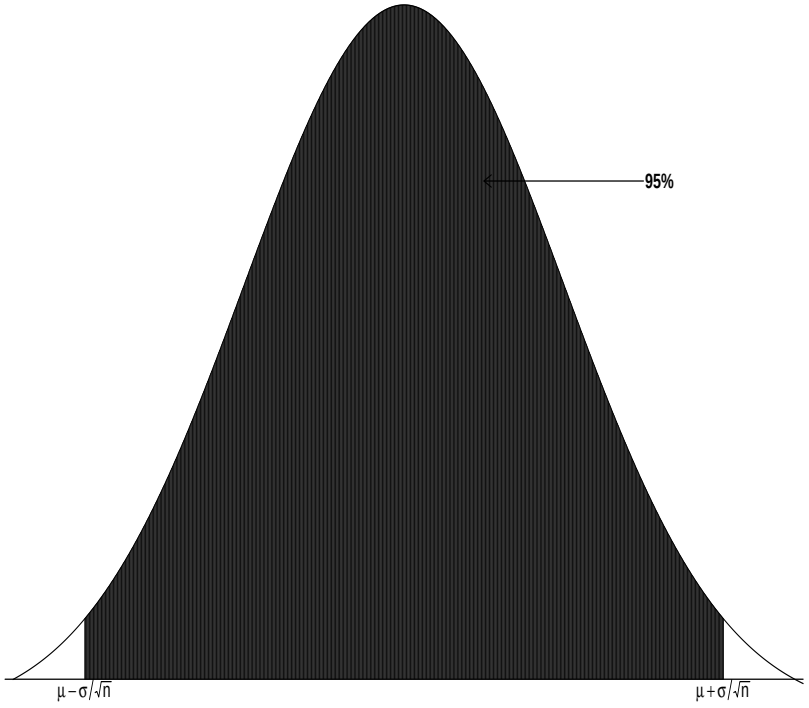
Oddly enough, this interval works. When will this interval not cover μ? If $\bar{X} < \mu - 1.96\frac{\sigma}{\sqrt{n}}$ then the right side of the interval $\bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$ will be less than μ. This will happen 2.5% of the time. If $\bar{X} > \mu + 1.96\frac{\sigma}{\sqrt{n}}$ then the left side of the interval $\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}$ will be greater than μ. This will happen 2.5% of the time. If these two things don't occur then the interval $(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}})$ will contain μ. That is, this interval will contain μ 95% of the time.

What's that? We don't know σ so we can't use the interval! That's right. We will replace σ by the sample standard deviation s. Thus the interval we will use is:

$$(\bar{X} - 1.96\frac{s}{\sqrt{n}}, \bar{X} + 1.96\frac{s}{\sqrt{n}})$$

Income Example: Lets apply to the income example. Recall that the data are:

Figure 7.1: A 95% confidence interval




```
28     29     35     42     42     44     50     52     54     56     59
78     84     90     95    101    108    116    121    122    133    150
158    167    235
```

Recall the average income is 89.96. The sample standard deviation is (Either do it by hand or check the **numerical summaries** button in the summary module):

```
Rweb:> # STANDARD DEVIATION of x
Rweb:> var(1)^.5
```

```
[1]    51.68
```

Hence $s = 51.68$. Note for the interval we actually need $s/\sqrt{n}$ which is called **Standard Error of the Mean** : $s/\sqrt{n} = 10.33$. So the interval we want is:

```
(89.96 - 1.96*10.33, 89.96 + 1.96*10.33)
             (69.71, 110.21)
```

So we estimate the mean family income of a Smith University student to be between \$69,710 to \$110,210. Our error of estimation is $1.96 \times 10.33 = 20.25$; i.e., \$20,250. That seems like a lot. How can we reduce the error of estimation? A larger sample size; i.e, as n gets larger, $s/\sqrt{n}$ gets smaller.

Interpretation. What is this interval? One way of thinking about it is: the probability that the random interval $(\bar{X} - 1.96\frac{s}{\sqrt{n}}, \bar{X} + 1.96\frac{s}{\sqrt{n}})$ traps μ is .95. What the heck does this mean? Think of it this way. This interval is a result of a Bernoulli trial with probability of success .95. In practice, we have only one sample and one interval. It will either catch μ or not. But it is the outcome of a Bernoulli trial with probability of success .95. Hence, we are fairly confident of success. So we call it a 95% confidence interval.

Other Remarks. There are two approximations in our confidence interval:

1. It is based on the Central Limit Theorem which says the distribution of $\bar{X}$ is approximately normal, and we used it as exactly normal.
2. We estimated σ by s.

So our confidence interval is really an approximate confidence interval. It's close enough in most applications.

A final remark of considerable importance: The end points of our confidence interval are estimates of the 2.5 and 97.5 percentiles of the distribution of $\bar{X}$, the estimator. This will be very important in the section after next.

Exercise 7.2.1

1. *To set ideas, obtain a 95% confidence interval for μ if the data are:*

```
10   12  16  18  24
```

Do this one by hand. The sample mean and standard deviation are easy to get and $\sqrt{5} = 2.24$.

2. *Obtain a 95% confidence interval for μ if the data are:*

```
76  87  98 102 111 114 115 115 120 126
```

First boxplot the data. Next mark the sample average and the endpoints of the confidence interval on the plot. Here's some output from the summary module to do the confidence interval:

```
Rweb:> summary(variables)
        x
 Min.   : 76.0
 1st Qu.: 99.0
 Median :112.5
 Mean   :106.4
 3rd Qu.:115.0
 Max.   :126.0

Rweb:> # STANDARD DEVIATION of x
Rweb:> var(x)^.5
[1] 15.5863
```

3. *Obtain a 95% confidence interval for μ if the data are:*

```
6   8  14  30  31  32  51  57  87  87 109 145 156 171 342
```

First boxplot the data. Next mark the sample average and the endpoints of the confidence interval on the plot. Here's the output from the summary module to do the confidence interval:

```
Rweb:> summary(variables)
        x
 Min.   :  6.0
 1st Qu.: 30.5
 Median : 57.0
 Mean   : 88.4
 3rd Qu.:127.0
 Max.   :342.0

Rweb:> # STANDARD DEVIATION of x
Rweb:> var(x)^.5
[1] 88.8005
```

4. *Consider the following sample of Etruscan skull sizes: Obtain a 95% confidence interval for* μ.

   ```
   141  145  145  146  142  126  144  146  154  149  143  131
   ```

   (Ans: 142.6 ± 4.25*).*

5. *Same as the last question for a sample of size 10 of Italian skull sizes:*

   ```
   134  132  126  134  131  130  130  125  132  126
   ```

   (Ans: 130 ± 2.04*).*

6. *Plot the confidence intervals from the last two problems on a line. What do conclude about the true mean skull sizes of Etruscans and Italians based on this comparison?*

7. *Now use all the Etruscan and Italian data (Appendix A) to do the last three exercises.*

7.3 Confidence Intervals for Proportions

Suppose we have a population proportion of interest. There are many examples:

1. The proportion of left-handed professional baseball players.
2. President Clinton's rating.
3. The proportion of patients with a specific disease who are under a new drug.
4. The proportion of graduating high school students who can read at the eighth grade level.
5. The proportion of Republicans who will vote for Bush.
6. The proportion of Democrats who will vote for Bush.
7. The proportion of Republicans who will vote for Gore.
8. The proportion of Democrats who will vote for Gore.
9. The proportion of citizens who will not vote.

Let p denote the population proportion. To estimate p, we sample the population and form the sample proportion which we will call $\hat{p}$.

Baseball Example. Consider the first example above: *The proportion of left-handed professional baseball players.* We have a sample of size 59 from this population.

There are 15 left-handed baseball players so the sample proportion is $\hat{p} = 15/59 = .2542$. Thus .2542 is our estimate of the proportion of left-handed professional baseball players. How much did it miss by?

In general, $\hat{p}$ is a sample average, (Record Success as 1 and Failure as 0, then the sum of these 0's and 1's is the number of successes and the average (divide sum by n) is $\hat{p}$). Hence we can invoke the Central Limit Theorem to determine a confidence interval for p. We use a slightly different standard error, though. The **Confidence Interval for** p is:

$$(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$$

You have already used this. The error here is the error (except 1.96 replaces 2) that we used for our estimates of probability based on resampling. This confidence interval has the same interpretation as the one in the last section; i.e., we are fairly confident that the true population proportion is contained in the interval.

Baseball Example. For the baseball data,

$$1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96\sqrt{\frac{.2542(1-.2542)}{59}} = .1111$$

Hence, the confidence interval is *(.2542 - .1111, .2542 + .1111) = (.143, .365).* So provided this data came from a random sample, we are fairly confident that the true percentage of left-handed professional baseball players is between 14 and 37%.

Exercise 7.3.1

1. *The cure rate for a the standard treatment of a disease is 45%. Dr. Snyder has perfected a primitive treatment which he claims is much better. As evidence, he says that he has used his new treatment on 50 patients with the disease and cured 25 of them. What do you think? Is this new treatment better. Use a 95% confidence interval to answer the question. (Ans.: (.36,.64)).*

2. *Experimenters injected a growth hormone gene into thousands of carp eggs. Of the 400 carp that grew from these eggs, 20 incorporated the gene into their DNA (Science News, May 20, 1989). Calculate a 95% confidence interval for the proportion of carp that would incorporate the gene into their DNA. From Statistics, S. Rasmussen, CA: Brooks/Cole, 1992. (Ans.: (.03,.07)).*

3. *Using Carrie's baseball data, estimate the proportion of professional baseball players who weigh 200 or more pounds. Find a 95% confidence interval for this proportion and interpret it. (Ans: 21 out of 59 weigh 200 or more pounds. So the CI is: (.23, .48)).*

7.4 Confidence Intervals Based on Resampling

In this section we discuss a way of obtaining confidence intervals in a variety of situations. These are based on **resampling** and are often referred to as **bootstrap percentile confidence intervals**.

Recall from Section 7.2 that the endpoints of the confidence interval for the population mean, μ, are actually estimates of the approximate (Central Limit Theorem) $2.5th$ and $97.5th$ percentiles of the distribution of $\bar{X}$. We will use this for our bootstrap confidence intervals.

Consider the population median, θ. We will obtain a bootstrap confidence interval for the median. This is often a parameter of interest because it divides the population in half. That is, half the time a population item is less than θ and half the time a population item is greater than θ. For instance, suppose the population of interest is the income of an American family. If you knew θ then you would know if your family is in the bottom half or top half of American families when it came to income.

As another example, suppose you were doing research on a new battery to power an electrical automobile. Suppose the median lifetime in miles of the current battery is 300 miles. You feel that the new battery is a vast improvement. Let θ be the median lifetime in miles of your new battery. You don't know θ, but you would like to show it is an improvement; i.e., it is over 300 miles. How would you investigate this?

It's easy, right? You don't know the population (the lifetime in miles of a typical new battery), so you will have to use a sample. So you select 20 new batteries and put them on test. **WAIT!!!!!!**. It is extremely important that the batteries are selected:

1. **Independent of one another.**
2. **They were manufactured under the same conditions.**

If these conditions are not meant get ready for **GIGO**, **Garbage In, Garbage Out**.

These assumptions are very important and they have to be followed. You can see why we are often dealing with small samples. In this case, we are destroying the battery when we sample it, (you can recharge it, but a recharged battery is not in the population of interest!). How long a recharged battery lasts is of interest but in the present experiment, we re not measuring the effect of recharging. This may be a later experiment. Also, we are doing research on the battery so you may be tempted to make modifications to the battery as we sample. **Nope,** not allowed for this violates assumption 2. (In certain situations this can be done but it is a much different experimental design ; see the section on regression design.)

Continuing with our example, suppose you do select 20 new batteries at random and put them on test, (20 cars of the same type are selected, one of the new batteries are installed in each car, and they are driven over the same route). Suppose the (sorted) lifetimes of the batteries in miles are:

```
196    204    233    256    258    313    315    322    403     408
483    510    538    559    586    722    806    875    930    1192
```

A dotplot of the sample is:

```
        .
   : .:  :   :   . ....      .   .   .  .            .
---+---------+---------+---------+---------+---------+---C1
 200       400       600       800      1000      1200
```

Dotplot is skewed right and shows a lot of noise. The sample median is $Q_2 = .5 \times (408+483) = 445.5$. This is above 300 (the lifetime in miles of the old battery). But wait! Five of the new batteries lasted less than 300 miles. So are we sure??? Of course not, $Q_2 = 445.5$ is just an estimate of the population median θ. We need to know **how much it missed by**.

We of course need to know the distribution of Q_2, but we don't. Next, how about a Central Limit Theorem from which we could obtain the approximate distribution of Q_2. Such theorems exist, but the approximate standard error of Q_2 is not easy to estimate. How about estimating the $2.5th$ and $97.5th$ percentiles of the distribution of Q_2? Hey, now you are cooking!

Okay, we need the distribution of Q_2. But we don't know it. We could do it this way, though. Simply do the experiment over and over, say, 1000 times. For each of those times, calculate Q_2. Form a histogram of these 1000 Q_2's and pick off the $2.5th$ and $97.5th$ percentiles. (This is the same as sorting the 1000 Q_2's and selecting the $25th$ and $976th$ sorted Q_2.

Back to the battery experiment! We just have to do 1000 experiments. **WHAT!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!**. **That's** 1000×20 **batteries** that we need. Ridiculous. Of course, we can't do this.

So what's to do. Louder, I can't hear you. You got it!!!!!!!!!!!!!!!!!! **Resample the sample 1000 times**. For each resample calculate the sample median. Form a histogram of these 1000 Q_2 and pick off the $2.5th$ and $97.5th$ percentiles. (This is the same as sorting the 1000 Q_2's and selecting the $25th$ and $976th$ sorted Q_2's. These percentiles do indeed estimate the percentiles of the true distribution of Q_2. It's such a simple idea and it works. You simply resample the sample. To insure independence, you resample **with replacement** and you use the same sample size.

For example, here is a resample of the sample of lifetimes of the batteries (I have sorted them):

```
196   196   204   204   204   204   256   258   315   315
315   322   483   538   559   559   806   875   875   930
```

The sample median is 315. Here is another resample:

```
196   204   256   256   256   313   313   403   483   510
538   538   538   538   538   586   722   806   875  1192
```

The sample median is 524. We only have 998 more resamples to go. Of course, the computer will do this for us.

We will have two levels of CC for this.

1. Input the data and select the number of resamples (trials) that you want. For class use, the concept (what is a confidence interval) is more important than real use so we will often just do 100 resamples. As you will see, the 100 resampled medians will be returned in order. Avoiding fractions, we will choose the $3rd$ and the $98th$ items from these sorted 100 resample medians as our 95% confidence interval for the median.

2. For the second level, 1000 resamples will be done. They will not be printed out. Just the $2.5th$ and $97.5th$ percentiles will be printed out.

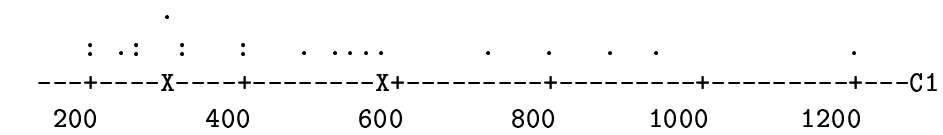
I did 100 resamples by getting the class code (One-Sample bootstrap means and medians), dropping the data in the input box, entering my id, putting in for 100 resamples (bootstraps), and clicking submit. This gave me:

```
286.5   313.0   314.0   314.0   315.0   315.0   317.5   317.5   318.5
318.5   318.5   318.5   318.5   318.5   322.0   358.0   359.0   361.5
361.5   362.5   362.5   362.5   362.5   362.5   362.5   365.0   365.0
365.0   402.5   403.0   403.0   403.0   403.0   403.0   405.5   405.5
405.5   405.5   405.5   405.5   408.0   408.0   408.0   408.0   408.0
408.0   408.0   408.0   412.5   443.0   443.0   445.5   445.5   445.5
445.5   445.5   459.0   459.0   459.0   470.5   483.0   483.0   483.0
483.0   483.0   483.0   483.0   483.0   483.0   483.5   496.5   496.5
496.5   496.5   496.5   496.5   510.0   510.0   510.0   510.0   510.0
510.0   510.5   521.0   524.0   524.0   524.0   524.0   524.0   534.5
534.5   538.0   538.0   538.0   548.5   548.5   559.0   559.0   572.5
572.5
```

Hence my confidence interval is *(314, 559)*. This is only based on 100 resamples, so lets use the terminology: *we are* **fairly** *confident that the true population median is between 314 and 559.* Note that the interval did not include 300.

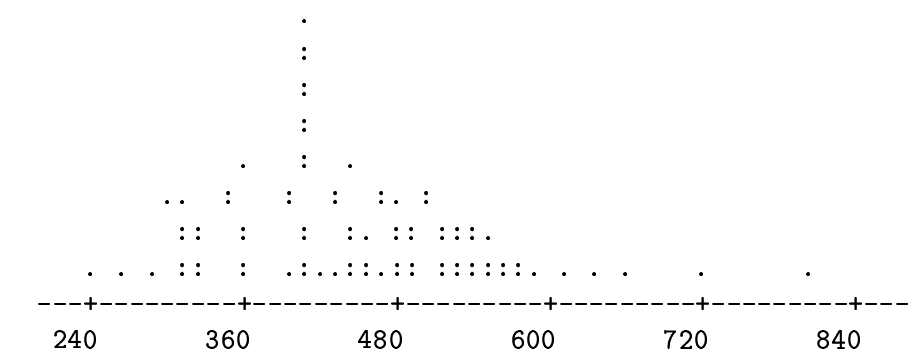
For the practical confidence interval based on 1000 resamples, I got the interval *(314, 572.5)* This is based on 1000 resamples, so we will use the terminology: *we are* **95%** *confident that the true population median is between 314 and 559.* Note that the interval did not include 300. Because this interval did not contain 300 and all values in the interval exceeded 300, we are confident that the new battery is an improvement. Is it a practical improvement? This is a question for the engineers to determine.

Next is a dotplot of the sample showing the location (X's) of the confidence interval:



A dotplot of the 1000 resample medians is given by

```
Each dot represents 14 points
```



It is fairly symmetric in the middle with an obvious tail to the right. It shows a central limit effect as we have seen with the sample mean.

Exercise 7.4.1

1. *Consider the following simple data set.*

77	79	81	91	106	114	126	132

Obtain 5 resamples of this data set using the previous resampling code which we used for probability. (Just sample with replacement the numbers: min = 1, max = 8, trials = 5, numbers to be drawn 8. These are the sample item numbers for the resample. For example if the numbers you draw are:

```
6       5       4       4       4       5       8       6
```

Then your resample is: 114, 106, 91, 91 91, 106, 132, 114

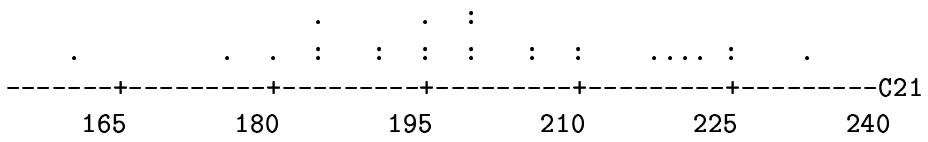
Obtain 4 more resamples. Calculate the median of each. Compare your resampled medians with the sample median.

2. *For the last problem use the class code (One-Sample bootstrap means and medians (Sorted)) to obtain a 95% confidence interval for the true population median based on 100 resamples.*

3. *For problem #1, use the class code (One-Sample CI's for the mean and median) to obtain a 95% confidence interval for the true population median based on 1000 resamples. Dotplot your data set and locate the confidence interval and sample median on your plot.*

4. *Below are the weights of the pitchers in Carrie's baseball data set. Obtain the sample median. Use the class code (One-Sample CI's for the mean and median) to obtain a 95% confidence interval for the true population median weight of a professional baseball pitcher based on 1000 resamples. Locate your interval on the dotplot plot below and interpret your interval.*

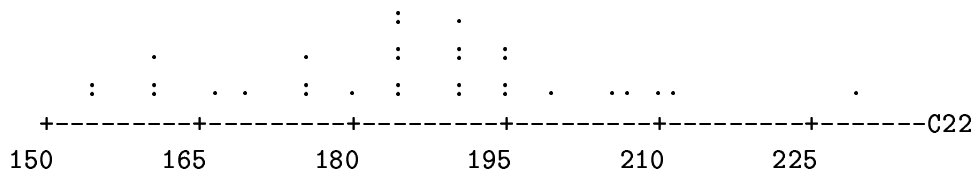
```
160     175     180     185     185     185     190     190     195     195     195
200     200     200     200     205     205     210     210     218     219     220
222     225     225     232
```



5. *Do the last problem for the weights of the hitters:*

```
155     155     160     160     160     166     170     175     175     175     180
185     185     185     185     185     185     185     190     190     190     190
190     195     195     195     195     200     205     207     210     211     230
```

.



6. *Plot your CI's for the last two problems on the same real number line. What do you conclude about the true median weights of hitters and pitchers based on this plot?*

7. *Select one of your textbooks or a novel that you are reading. Select a passage at random (not dialogue). Then count up the number of words in the first sentence of the passage. Record this number. Repeat this for 30 sentences. This your sample of size 30. Dotplot your data and describe the shape. Determine the sample median. Next obtain a 95% confidence interval for the true median sentence length. Locate the interval on your dotplot. What does it mean?*

Chapter 8

Tests of Hypotheses

8.1 Introduction

The beginning of a study is formed by a **question**. This question usually defines the **population** (or populations) of interest. If we knew the population then we could answer the question. Often we rephrase the question in terms of hypotheses. The one hypothesis often reflects the current state (standard, no change) while a second hypothesis reflects change. The first hypothesis is refereed to as the null hypothesis and is denoted by H_0, while the second hypothesis is designated as the alternative hypothesis and is denoted by H_A.

In this chapter we are concerned with two populations. For these problems, there is a natural null hypothesis, namely, that the two populations are the same. Consider the following questions for two population problems along with associated hypotheses.

1. At a pharmaceutical company, a new drug has been developed which should reduce cholesterol much more than their current drug on the market. Is this true? Hypotheses:

   - H_0: New drug has the same effect on cholesterol as the current drug.
   - H_A: New drug reduces cholesterol more than the current drug.

2. A new method for teaching statistics utilizing technology has been developed. Is it more successful than the usual lecture approach? Hypotheses:

   - H_0: The teaching methods are equally as effective.
   - H_A: The new teaching method is more successful than the usual approach.

3. Based on head sizes (maximum head breadths), are the ancient Etruscans different from modern Italians? Hypotheses:

   - H_0: Typical head sizes of Etruscans and Italians are the same.
   - H_A: Head sizes of Etruscans and Italians differ.

4. A new variety of wheat is developed which should yield more wheat per acre than a current popular variety.

   - H_o: Yields of the two wheat varieties are about the same.
   - H_A: The yields of the new variety of wheat are larger than the current popular variety.

It is easy to think of many such examples because we make many comparisons daily . The only new stuff is the labeling of the hypotheses. Each of the alternative hypotheses are of the form: (a) one population is better (bigger, larger) than the other. There are two other classes of alternatives:

(b) one population is worse than the other (actually this is the same as the other is better) and (c) the populations are different. We will just consider (a) for a while and discuss the others later.

In life, we must often decide between conflicting claims and usually **we must decide in the face of uncertainty**. We will never be sure which hypothesis is correct but perhaps we can have some confidence, never 100%, that our decision is correct.

8.2 A Testing Procedure

After casting our questions into hypotheses, we need a formal way to test H_0 versus H_A. We need to decide whether to accept H_0 and, hence, reject H_A or whether to reject H_0 and, hence, accept H_A. We must decide one way or another, there is no fence sitting here.

Alas, there are two types of errors we can make:

1. **Type I Error**: We reject H_0, when H_0 is true.
2. **Type II Error**: We accept H_0, when H_A is true.

For example, recall the first example above: *At a pharmaceutical company, a new drug has been developed which should reduce cholesterol much more than their current drug on the market. Is this true? The hypotheses are: H_0: New drug has the same effect on cholesterol as the current drug; and H_A: New drug reduces cholesterol more than the current drug.* The errors in words of this problem are:

1. **Type I Error**: We declare the new drug is more effective than the current drug on the market, when really it is not more effective.
2. **Type II Error**: We declare the new drug is not more effective when it really is more effective.

We need information to decide which hypothesis is true. So we take a random sample from each population and base our decision on these samples. We will use the samples to from a decision rule to make a decision on which hypothesis H_0 or H_A is true. Because we must decide, we may make either a Type I or Type II error. Usually Type I error is regarded as the more serious error. For instance, in the two population problem suppose the first population represents the *standard* while the second population represents the *new*. In rejecting H_0 we are claiming the *new* is better than the *standard*. Hence, a Type I error here means we are claiming the *new* is better when it really isn't. In real life, this often means shelling out dollars (buying the new, retooling the assembly line, installing a new expensive teaching method) for something that is not better. Of course, a Type II error is serious, also, because *you have missed something which is better.*

Getting back to our **decision rule**: We have two samples and we must make a decision in the face of uncertainty. So we choose a **test statistic**, say T, and a **decision rule** say, "We reject H_0 and accept H_A if T is too large." How large is too large? We pick a probability for Type I error, say α, usually .05 or smaller and then determine how large is too large. **IT'S EASY**. Yuck, how about an example which leads into our first test statistic?

8.3 The Wilcoxon

Simple Example. Suppose our company makes batteries, and, in particular, we make an expensive battery, called the XX, that is used in the space station. Suppose you have a bright idea of how to increase the life time of the battery by changing one of the resources used in its manufacture. You call your new battery the YY. Your hypotheses are:

- H_0: Battery YY's lifetime is the same as Battery XX's.
- H_A: Battery YY's lifetime is longer than Battery XX's.

So you take a sample of XX batteries, say 6 of them, and a sample of 5 YY batteries. These 11 batteries are made under identical conditions, (except for the new resource that goes into the YY's). Also they must have been made independent of one another. (**good sampling is expensive, but you avoid GIGO**).

The test statistic that we have chosen, T, is simple: just count up the number of times a YY battery beats (lasts longer than) a XX battery. This is called the two sample Wilcoxon test statistic, which we will refer to as the Wilcoxon test statistic. Note that there are $30 = 5 \times 6$ match ups between the samples. Under the null hypothesis, H_0, you expect T to be $(1/2) \times 30 = 15$; i.e., under H_0 in the 30 match ups, you expect half the time that the YY battery will last longer than the XX battery and half the time that the XX battery will last longer than the YY battery. You reject H_0 in favor of H_A if T is too large.

Suppose the data are:

```
XX    49     53     74    111    113    335
YY    62    101    167    174    190
```

To compute T just go use each YY data point:

```
62 beats 2 XX's, namely 49, 53
101 beats 3 XX's, namely 49, 53, 74
167 beats 5 XX's, namely 49, 53, 74, 111, 113
174 beats 5 XX's, namely 49, 53, 74, 111, 113
190 beats 5 XX's, namely 49, 53, 74, 111, 113
```

So $T = 20$.

So T is 20, this is more than 15. The question is: Is this enough more? We will answer that after a few remarks and exercises.

Exercise 8.3.1

1. *Obtain a comparison dotplot of the two samples (X and Y) below. Let T be the number of time a Y beats a X. Under the null hypothesis, what do you expect T to be? Next compute T.*

X	78	108	121	123	127	140	141
Y	104	107	119	124	135	136	

   (Ans: T= 17).

2. *Below are the batting averages of the switch hitters and the left-handed hitters from the baseball data set. Obtain a comparison dotplot. Dotplot the two samples. Let T be the number of time an average of a left-handed hitter is bigger than the average of a switch-hitter. Under the null hypothesis, what do you expect T to be? Next compute T .*

Switch	.212	.218	.236	.242	.251	.251	.254	.261	.270	.282
Left	.238	.271	.279	.283	.284	.290	.300	.303		

   (Ans: T = 71).

3. *Consider the following samples of Italian and Etruscan skull sizes. Let T be the number of time an Etruscan skull size is bigger than an Italian skull size. Under the null hypothesis, what do you expect T to be? Next compute T. It's easier if you sort the samples first!*

Ital.	134	132	126	134	131	130	125	132	126			
Etru.	141	145	145	146	142	126	144	146	154	149	143	131

4. *Below are the batting averages of the right-handed hitters and the left-handed hitters from the baseball data set. Dotplot the two samples. Let T be the number of times an average of a left-handed hitter is bigger than the average of a right-handed hitter. Under the null hypothesis, what do you expect T to be? Next compute T .*

Right	.225	.238	.239	.243	.244	.245	.262	.271	.271
	.274	.274	.276	.282	.286	.286			
Left	.238	.271	.279	.283	.284	.290	.300	.303	.240

5. *Did Manuel I shortchange the people by having less silver in in later days mintings? Try to answer this question by comparing the following two data sets (use comparison boxplots). Let T be the number of times a First minting has a higher percentage than a Fourth minting. Under the null hypothesis, what do you expect T to be? Next compute T .*

```
First:    5.9  6.8   6.4  7.0  6.6  7.7  7.2  6.9  6.2
Fourth    5.3  5.6   5.5  5.1  6.2  5.8  5.8
```

General Case.

We need a little notation. In general (not just the battery example), let $X_1, X_2, ..., X_m$ denote the random sample from the first population and let $Y_1, Y_2, ..., Y_n$ denote the random sample from the second population. Denote the Wilcoxon test statistic by

$$T = \#\{Y_j > X_i\}$$

Read : T is the number of matches between Y and X in which Y is larger than X.

There are $m \times n$ matches. If H_0 is true, we expect T to be $\frac{m \times n}{2}$.

Actually a table that proves useful here and in the next chapter is the table of differences. Sort each sample. Then columns of the table are sorted Y's and the rows are sorted X's. The entries in the table are the differences $Y_j - X_i$. The statistic T is just the number of positive differences. Here is the table for the battery data.

	62	101	167	174	190
49	13	52	118	125	141
53	9	48	114	121	137
74	-12	27	93	100	116
111	-49	-10	56	63	79
113	-51	-12	54	61	77
335	-273	-234	-168	-161	-145

In general, we need to know how large T should be to reject H_0 in favor of H_A. The key is very large values of T should be rare if H_0 is true. So we calculate the probability that T is greater than or equal to the observed value of T assuming that H_0 is true. This is called the **p-value** or the **observed significance level** of the test. Oh, oh! We need the distribution of T assuming that H_0 is true. How do we get that? What's that? Resampling! That's right. We approximate this distribution by resampling.

We need to resample assuming H_0 is true. We can do this by combining the samples into one large sample of size $N = m + n$. Then sample with replacement from this combined sample, randomly assigning m of these values to be the *new* X's and the remaining n of these values to be the *new* Y's. Note that the null hypothesis is true for these new samples, they are from the big combined sample.

Battery Example

Let's try it on the battery data. Recall that the samples are:

```
XX    49     53     74    111    113    335
YY    62    101    167    174    190
```

Now combine it into one data set:

```
Null Population:   49     53     74    111    113    335  62
                  101    167    174    190
```

Resample with replacement from this data set and assign the first 6 to be a X and the last 5 to be a Y. (I did this by mixing the numbers together in a hat, drawing one out, recording it, putting it back in, mixing them up, ETC!!! Here's the results:

```
New X's:   335    167     53    335     74     62
New Y's:    62     49     53    174    190
```

Now compute T. (Here, we are going to get some $Y = X$, so we will count such a match as 1/2). Hence starting with 62, $T = 1.5 + 0 + .5 + 4 + 4 = 10$. Recall that the value of T on the original sample was 20. So the event $T \geq 20$ did not occur.

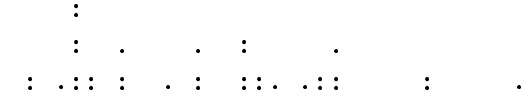
Now do this 1000 times and count the times the event $T \geq 20$ occurs. Divide this number by 1000 and we have the *p-value* of the test.

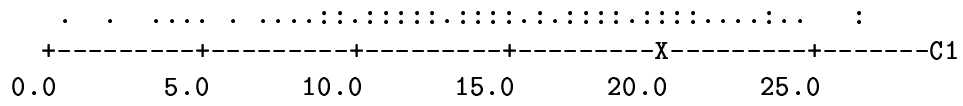
The class code discussed below will do this. But for now, here are the results of doing it 100 times. These are the 100 sorted resampled test statistics:

```
  0.5  2.0  3.5  4.0  4.5  5.0  6.0  7.0  7.5  8.0  8.5  9.0  9.0  9.5  9.5
*10.0 10.5 10.5 10.5 10.5 11.0 11.0 11.5 11.5 11.5 12.0 12.0 12.0 12.0 12.0
 12.0 12.0 12.0 12.5 12.5 12.5 12.5 13.0 13.5 13.5 13.5 13.5 13.5 14.0 14.0
 14.5 14.5 15.0 15.0 15.0 15.5 16.0 16.0 16.0 16.0 16.0 16.5 17.0 17.0 17.5
 17.5 17.5 17.5 17.5 17.5 18.0 18.0 18.0 18.0 18.5 18.5 18.5 19.0 19.5 19.5
 19.5 20.0 20.0 20.0 20.0 20.5 20.5 20.5 20.5 20.5 21.0 21.0 21.5 22.0 22.5
 23.0 23.5 23.5 23.5 23.5 24.0 24.5 26.5 26.5 26.5
```

I put a * at the resample we just did (i.e., resampled $T = 10$). How many times did the resampled T exceed 20? Well just count them up: 24 times. So the *p-value* of the test was .24. That's not too rare! One-out-of-four times. Hence, we would probably not reject H_0. We would conclude: There is insufficient evidence to conclude that Battery YY lasts longer than Battery XX.

There is nothing like a picture of a *p-value* . Here's a dotplot of the 100 resampled T's.





I put an X on 20. If you count the dots from 20 on you will get 24. If this were a histogram, .24 would be the shaded area to the right of 20.

Now you try it with the class code (Two-Sample bootstrap Wilcoxon statistic). Drop the XX and YY samples into the boxes (they are printed below), enter 100 for the number of trials, and click submit. You will get back 100 sorted Wilcoxon's. Determine the *p-value*; i.e., the number of resampled T's which exceed 20. The data are:

XX	49	53	74	111	113	335
YY	62	101	167	174	190	

Exercise 8.3.2

1. *In the last set of exercises, you obtained T be the number of time a Y beats a X. Now use the class code (Two-Sample bootstrap Wilcoxon statistic (Sorted)) to compute the p-value based on 100 trials.*

X	78	108	121	123	127	140	141
Y	104	107	119	124	135	136	

2. *Below are the batting averages of the switch hitters and the left-handed hitters from the baseball data set. Let T be the number of time an average of a left-handed hitter is bigger than the average of a switch-hitter. Recall T = 71. Now use the class code (Two-Sample bootstrap Wilcoxon statistic (Sorted)) to compute the p-value based on 100 trials.*

Switch	.212	.218	.236	.242	.251	.251	.254	.261	.270	.282
Left	.238	.271	.279	.283	.284	.290	.300	.303		

3. *Consider the following samples of Italian and Etruscan skull sizes. Let T be the number of time an Etruscan skull size is bigger than an Italian skull size. You computed T in the last set of exercises. Now use the class code (Two-Sample bootstrap Wilcoxon statistic (Sorted)) to compute the p-value based on 100 trials.*

Ital.	134	132	126	134	131	130	130	125	132	126		
Etru.	141	145	145	146	142	126	144	146	154	149	143	131

4. *Below are the batting averages of the right-handed hitters and the left-handed hitters from the baseball data set. Let T be the number of times an average of a left-handed hitter is bigger than the average of a right-handed hitter. You computed T in the last set of exercises. Now use the class code (Two-Sample bootstrap Wilcoxon statistic (Sorted)) to compute the p-value based on 100 trials.*

Right	.225	.238	.239	.243	.244	.245	.262	.271	.271
	.274	.274	.276	.282	.286	.286			
Left	.238	.271	.279	.283	.284	.290	.300	.303	.240

5. *Did Manuel I shortchange the people by having less silver in in later days mintings? Try to answer this question by comparing the following two data sets (use comparison boxplots). Let T be the number of times a first minting has a higher percentage than a Fourth minting. You computed T in the last set of exercises. Now use the class code (Two-Sample bootstrap Wilcoxon statistic (Sorted)) to compute the p-value based on 100 trials.*

First:	5.9	6.8	6.4	7.0	6.6	7.7	7.2	6.9	6.2
Fourth	5.3	5.6	5.5	5.1	6.2	5.8	5.8		

8.4 Wilcoxon: Other Alternatives

The problem we have been looking at can be summarized as follows: We have two populations, say X and Y. We think that a measurement from population Y is typically larger than a measurement from population X. This is our alternative hypothesis H_A. The null hypothesis is that the populations are the same. Again:

- H_0: Populations are the same.
- H_A: A measurement from population Y is typically larger than a measurement from population X.

Our procedure is to draw a random sample from Population X and a random sample from Population Y. We then calculate T the number of times a Y beats an X. If T is too large we reject H_0 in favor of H_A, where large is measured by the p-value.

Another set of alternatives is:

- H_0: Populations are the same.
- H_A: A measurement from population Y is typically smaller than a measurement from population X.

For example, suppose a person takes golf lessons. Then his score should improve; i.e., after the lesson scores should be smaller than before lessons scores. Certainly, a test procedure is to use the Wilcoxon, but now reject if T is too small. Here's an example.

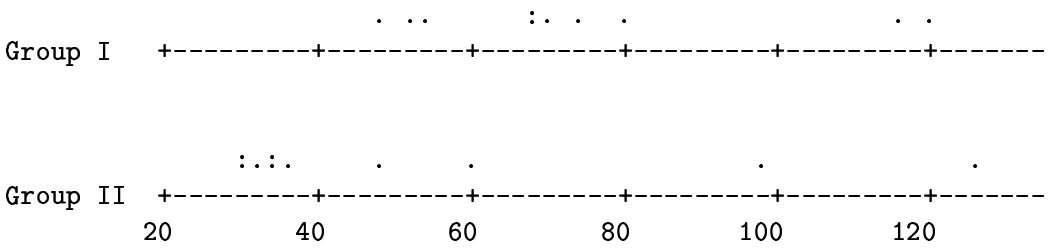
Twenty quail were randomly assigned to two groups, 10 to each. The quail in Group I were given a diet without a drug compound while the quail in Group II were given a diet with a drug compound inserted, which hopefully reduces LDL (low-density-lipid) cholesterol. Except for the difference in diet the quail were treated the same. At the end of the study their LDL levels were measured. The hypotheses are:

- H_0: The LDL levels of both groups are about the same.
- H_A: LDL levels of quail in Group II are typically smaller than LDL levels in Group I.

Here is the sorted data:

```
Group I:    47  52  54  67  68  69  73  79 116 120
Group II:   30  30  31  33  34  36  47  59  98 125
```

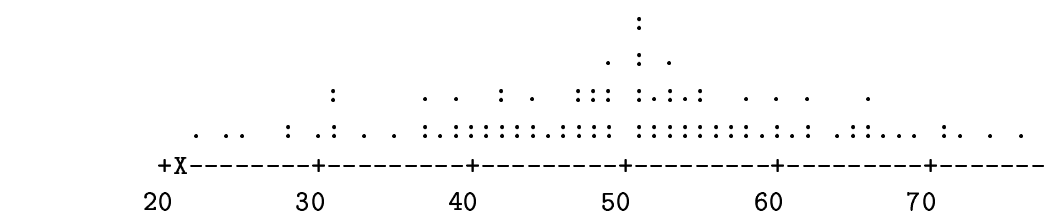
and a dotplot



It appears that the drug compound had an effect. The value of the Wilcoxon is $T = 21.5$ which is certainly smaller than what we would expect if H_0 is true, i.e., $.5 \times 100 = 50$. Is this small enough? We need the p-value which is the probability that $T \leq 21.5$ under H_0. To estimate this, we obtained 100 resampled T 's:

```
22.0 23.5 24.5 28.0 28.0 30.0 30.5 30.5 31.0 31.0 32.5 35.0 36.5 37.0 37.0
37.5 38.5 39.0 39.0 39.5 40.0 40.5 41.0 41.5 41.5 41.5 42.0 43.0 43.0 43.5
44.0 44.0 44.5 45.5 46.0 46.5 46.5 46.5 46.5 47.5 48.0 48.0 48.0 48.5 48.5
49.0 49.0 49.0 50.5 50.5 50.5 50.5 51.0 51.0 51.0 51.0 51.5 51.5 52.0 52.5
52.5 52.5 53.0 53.0 53.5 54.0 54.0 54.5 55.0 55.0 55.0 55.5 56.0 56.5 56.5
57.5 57.5 57.5 58.5 59.5 59.5 59.5 61.0 61.5 61.5 62.0 64.0 64.5 64.5 66.0
66.0 66.0 66.5 68.0 69.0 71.0 71.0 72.0 74.0 75.5
```

The estimated p-value is $0/100 = 0$. Here is a picture of the p-value



I also ran 1000 resampled T's which resulted in 13 resampled T's being less than or equal to 21.5. Hence the p-value is $13/1000 = .013$. Based on this evidence, we reject H_0 in favor of H_A and conclude that the drug is effective in reducing LDL cholesterol.

The third alternative is the alternative of ignorance; i.e., the populations differ. Formally,

- H_0: Populations are the same.
- H_A: A measurement from population Y is either typically smaller than a measurement from population X or typically larger than a measurement from population X.

Using the Wilcoxon, we would reject H_0 in favor of H_A if T is too small or too large. In this case, to determine the p-value we first determine if T is on the down side (T smaller than $mn/2$) or on the up side (T greater than $mn/2$). If it is on the down side we double the estimated probability that T is less than or equal to the observed value of T , while its on the up side we double the estimated probability that T is greater than or equal to the observed value of T . Hey, lets cut the chatter and do an example.

From *Statistical Concepts and Methods,* Page 321, Bhattacharya and Johnson (1977), New York: Wiley. The peak oxygen intake per unit of body weight, called the aerobic capacity of an individual performing a strenuous activity is a measure of work capacity. For a comparative study, measurements are recorded for a group of 12 Peruvian Highland natives and 10 U.S. lowlanders who have spent considerable time in high altitudes. Do these groups seem to differ in peak oxygen intake? The hypotheses are:

- H_0: Peruvian Highlanders tend to have the same peak oxygen intake as the U.S. acclimatized Lowlanders.
- H_A: Peruvian Highlanders tend to differ with respect to peak oxygen intake from the U.S. acclimatized Lowlanders.

Here's the sorted data:

```
Peru     34 35 36 38 38 42 43 46 48 50 52 55
US       30 32 32 33 36 38 41 43 44 46
```

Let T be the number of times a Peruvian has a higher peak oxygen intake than a US person. The value of the Wilcoxon is $T = 77.5$ which exceeds $120/2 = 60$. So T is on the upside. Hence the p-value is twice the probability that T is greater than or equal to 87.5. To estimate the p-value here are 100 resampled T's under H_0:

```
24.5  24.5  31.5  32.5  34.0  35.0  38.0  38.5  39.5  41.5  42.0  42.5
44.0  44.5  45.0  45.0  45.5  46.5  47.0  47.5  48.5  48.5  49.5  51.0
52.0  52.0  52.0  52.5  52.5  52.5  52.5  53.0  53.0  53.5  54.0  54.5
54.5  54.5  54.5  55.0  55.5  56.0  57.0  57.0  57.5  58.5  60.5  60.5
61.5  61.5  62.0  62.5  62.5  63.0  63.5  63.5  63.5  64.0  64.5  65.0
65.0  65.5  65.5  65.5  65.5  66.0  67.0  68.0  68.0  68.5  68.5  69.0
69.5  69.5  69.5  70.0  70.0  70.5  70.5  71.0  73.0  74.0  75.0  76.5
77.0  77.0  77.5  78.5  79.0  80.5  83.5  85.0  85.0  86.5  89.0  89.0
89.5  90.5  93.5 101.0
```

Based on these resampled T's, we estimate the p-value to be $2 * 6/100 = .12$. Assuming this pattern holds for 1000 resampled T's, we would not reject H_0 in favor of H_A. Our conclusion would be, that

there is insufficient evidence that Peruvian Highlanders differ from U.S. acclimatized Lowlanders with reference to peak oxygen intake. You are asked in the problems to estimate the p-value based on 1000 samples.

Exercise 8.4.1

1. *In the last example, use class code (Two-Sample Hypothesis and CI (Wilcoxon)) to determine the p-value based on 1000 resampled T's. What is your conclusion in terms of the data? Obtain comparison dotplots of the data.*

2. *Consider two data sets which are labeled as A and B and are given below. Suppose we want to test that the B's tend to be smaller than the A's. Determine the Wilcoxon test statistic and the p-value based on 1000 resamples using class code (Two-Sample Hypothesis and CI (Wilcoxon)).*

```
A:      12 16 18 25 30
B:      8  10 19 22 28
```

3. *Is the Wilcoxon robust? As a verification consider the following two samples. We want to test to see if the B's tend to be smaller than the A's. Determine the Wilcoxon test statistic and the p-value based on 1000 resamples using class code (Two-Sample Hypothesis and CI (Wilcoxon)).*

```
A:       70  72  87  88 102 112
B:       41 43 54 67 74 78 87 91
```

   Next change, the 70 to 7, the first A. Determine the Wilcoxon test statistic and the p-value based on 1000 resamples using class code (Two-Sample Hypothesis and CI (Wilcoxon)). Did your conclusion change?

   Next change it to -7000. Determine the Wilcoxon test statistic and the p-value based on 1000 resamples using class code (Two-Sample Hypothesis and CI (Wilcoxon)). Did your conclusion change?

4. *Recall the following problem: Select one of your textbooks or a novel that you are reading. Select a passage at random, Not dialogue. Then count up the number of words in the first sentence of the passage. Record this number. Repeat this for 15 sentences. This is your sample of size 15. Do this and then select a second book of the same type but by a different author and repeat the procedure for this second author.*

   State H_0 and H_A. Use the Wilcoxon to test these hypotheses. Use 1000 resampled T's. Obtain comparison dotplots. Conclude in terms of the problem.

Chapter 9

Estimation of Effect : Two Independent Samples

9.1 Introduction

This chapter is a continuation of the last chapter. As in the last chapter, we have two populations X and Y but we now want to estimate the difference between the populations. We do need to make one important assumption:

The populations differ by at most a shift in locations (centers).

Fortunately we have at least visual checks for this assumption in comparison dotplots, boxplots and back-to-back stem leaf plots based on the samples we obtain. For example, if the lengths of the boxes in the comparison boxplots are much different then this is an indication that scale (or noise) level is also different between the populations. Or, if, provided the sample sizes are large enough, the shapes of the back-to-back stem leaf plots are quite different then this would indicate that the populations differ by more than a shift in locations.

Under this assumption, the problem is easily parameterized. Let Δ be the **difference in locations** of the populations. In many problems, we think of Δ as the **effect** between the populations. If μ_1 is the mean of the first population and μ_2 is the mean of the second population then $\Delta = \mu_2 - \mu_1$. But Δ is also the difference in population medians, **shift is shift**. Hence, if θ_1 is the median of the first population and θ_2 is the median of the second population then $\Delta = \theta_2 - \theta_1$. So we want to estimate Δ and we will be done. What's that? Louder, I can't hear you. Right! We must also estimate the error of estimation. **We want a confidence interval for Δ, too.** How much did our estimate of Δ miss Δ by?

One final word. The value to check for in the confidence interval is **0**. For if 0 is in the confidence interval then there may be no difference between the populations. Note this is another way of testing for a difference between populations. In particular, consider the hypotheses:

- H_0: Populations X and Y are the same.
- H_A: A measurement from population Y is either typically smaller than a measurement from population X or typically larger than a measurement from population X.

We are now dealing with a location problem, so we recast these hypotheses as:

- H_0: $\Delta = 0$.
- H_A: $\Delta \neq 0$.

We reject H_0 in favor of H_A, if 0 is not in the confidence interval.

9.2 Estimation and Confidence Interval Based on the Wilcoxon

Suppose we have two populations which we assume differ by at most a shift in locations. Let Δ be the difference in locations: Population Y - Population X. We draw random samples from each population. Let $X_1, X_2, ..., X_m$ denote the sample of size m from the first population. Let $Y_1, Y_2, ..., Y_n$ denote the sample of size n from the second population. Think of Δ as positive for a moment. Then typical Y's are shifted up from a typical X's by Δ. If we knew Δ, we could **unshift** the Y's by subtracting Δ from each Y. This leads to a point estimate. Confused? Here's an example for the Wilcoxon point estimate:

Suppose the samples are:

- X: 8 12 16
- Y: 14 19 22

Lets estimate Δ by the median of the differences $Y_j - X_i$. Here are the 9 differences.

```
14-8=6,  14-12=2, 14-16=-2, 19-8=11, 19-12=7,
19-16=6, 22-8=14, 22-12=10, 22-16=6.
```

Here are the sorted differences:

```
-2  2   3   6   6  7   10   11   14
```

As our point estimate we will take the median of the differences, i.e., 6. Here are the X's and the unshifted Y's; i.e., $Y - 6$:

```
X:       8   12  16
Y - 6:   8   13  16
```

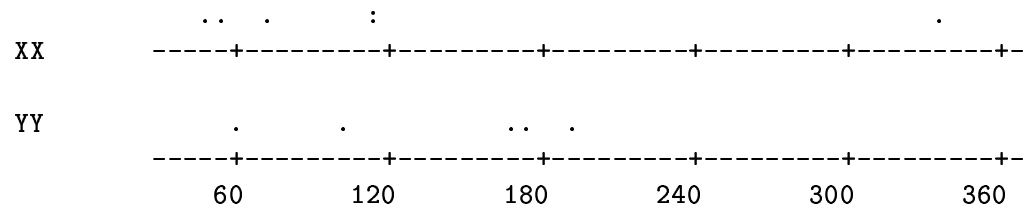
Now compute the Wilcoxon test statistic on the X's and the unshifted Y's. You will get $T = 4.5$ which is $\frac{m \times n}{2} = 9/2$. This is what you expect T to be if there are no differences. Hence, the median of the differences has unshifted the Y's.

In general, the estimate of the shift in locations based on the Wilcoxon is the median of the differences $Y_j - X_i$.

Consider the battery example of the last chapter. Recall that we had two types of batteries XX and YY and we wanted to see if a typical YY lasts longer that a typical XX. Lets estimate the difference in lifetimes of typical YY and XX batteries. Here are the samples (lifetime in hours):

```
XX    49     53     74    111    113    335
YY    62    101    167    174    190
```

Here is the comparison dotplot:



It seems though YY's are beating XX's. To estimate the shift we need to get all 30 differences of the form $YY_j - XX_i$. When we get this estimate by hand calculation, the table of differences discussed in the last chapter really helps. Sort the samples. Then the columns of the table are the sorted Y's and the rows of the table are the sorted X's. Then obtain the differences $Y_j - X_i$. As you will see the median is easy to get.

	62	101	167	174	190
49	13	52	118	125	141
53	9	48	114	121	137
74	-12	27	93	100	116
111	-49	-10	56	63	79
113	-51	-12	54	61	77
335	-273	-234	-168	-161	-145

Our point estimate is the median which is 53 (do a quick stem-leaf then compute the median). Could you guess it from the plot? (Take the YY's shift them back 53 units. Do these "aligned" samples seem about the same?). So a typical YY battery lasts 53 hours longer than a typical XX battery. Takes care of that problem. What's that? Oh right, **it could just be sampling error. We need a confidence interval!**

We will use percentile confidence intervals based on resampling. So its old stuff! The steps for a general situation are:

1. Resample m X's with replacement.
2. Resample n Y's with replacement.
3. Obtain the median of the differences of the resampled Y's minus the resampled X's.
4. Record this median.
5. Repeat steps (1) through (4) 1000 times.
6. Sort the 1000 medians,

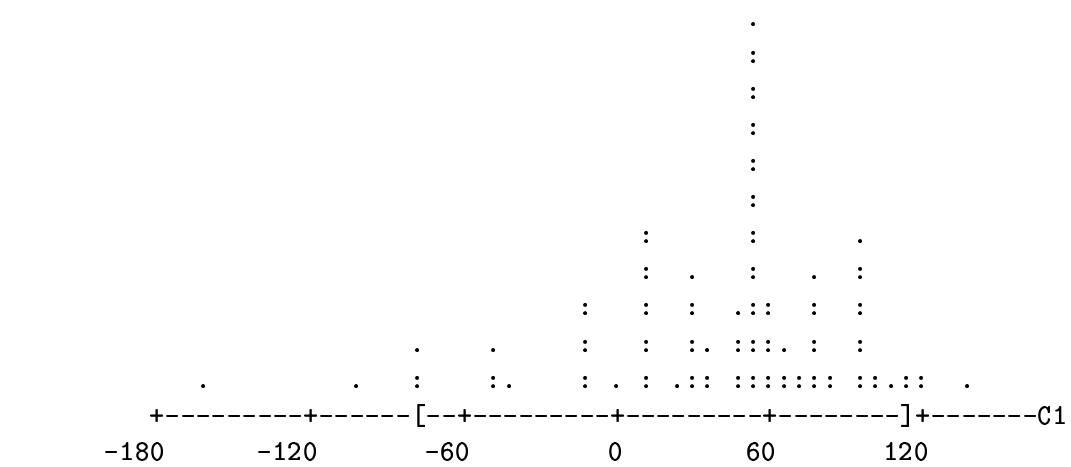
7. Pick off the *25th* and *976th* sorted medians. This is our 95% confidence interval.

It looks great until you see step (5). To get an idea of what's going on, I did step (5) 100 times. Here are the resorted 100 resampled medians of the differences:

-161.0	-105.0	-78.5	-78.5	-76.0	-49.0	-49.0	-45.5	-44.5	-12.0
-11.0	-11.0	-11.0	-10.0	-10.0	-0.5	9.0	9.0	9.0	9.0
11.0	11.0	13.0	13.0	13.0	13.0	22.0	27.0	27.0	27.0
27.0	27.0	27.0	30.5	34.5	37.5	37.5	48.0	48.0	48.0
48.0	48.0	51.0	51.0	51.0	52.0	52.0	52.0	52.0	53.0
53.0	53.0	54.0	54.0	54.5	55.0	56.0	56.0	56.0	56.0
56.0	56.0	56.5	57.5	58.5	61.0	61.0	61.0	61.0	63.0
63.0	64.5	70.0	72.5	77.0	77.0	77.0	77.0	77.0	79.0
79.0	85.0	86.0	93.0	93.0	93.0	93.0	93.0	93.0	93.0
93.0	96.5	100.0	100.0	107.0	114.0	116.0	117.0	121.0	137.0

The confidence interval is $(-78.5, 117)$. It contains 0, hence, the results are inconclusive. Remember we took differences of the form YY minus XX, so positive values in the CI means YY beats XX, but negative values mean XX beats YY. Our conclusion would be: a typical YY battery has a shorter lifetime than a typical XX by 78.5 hours to a typical YY battery has a longer lifetime than a typical XX by 117 hours. Though right, this sounds a bit odd. It is better to say the results were inconclusive. The value of 53 did not overcome the noise level. Note that on this data set, this is the same conclusion which we came to in Chapter 8.

A picture is worth a 1000 words, so here is a histogram of the 100 resampled medians of the differences. I have located the CI on it with *[ ]*'s.



Using 1000 resamples, I got the confidence interval $(-105, 115)$. So the conclusion remains the same.

Using the class code (Two-Sample hypothesis test and confidence interval for the location parameter based on the Wilcoxon) you try it. Simply bring up class code in a second window, drop the XX sample in the first box (data set 1), drop the YY sample in the second box (data set 2), and submit.

Exercise 9.2.1

1. *To set ideas work on this simple data set.*

   ```
   X   12 15 18
   Y   16 19 25 28
   ```

   (a) *Obtain all 12 differences (Y minus X).*

   (b) *Next obtain the point estimate, the median of the differences.*

   (c) *Subtract this estimate from the Y's and obtain the value of the Wilcoxon test statistic.*

2. *For the last problem, use the following list of random numbers to obtain 2 resampled median of differences.*

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

3. *Consider the batting averages of the switch hitters and the left-handed hitters from the baseball data set. Using the class code (Two-Sample Hypothesis and CI (Wilcoxon)), obtain the estimate of the difference (Left minus switch) of batting averages and determine a 95% confidence interval for the difference. What does the interval mean in terms of the problem?*

   ```
   Switch  .212  .218  .236  .242  .251  .251  .254  .261  .270  .282
    Left   .238  .271  .279  .283  .284  .290  .300  .303
   ```

4. *Consider the following samples of Italian and Etruscan skull sizes. Use the class code (Two-Sample Hypothesis and CI (Wilcoxon)) to obtain the estimate of difference between a typical Etruscan skull and an Italian skull. Obtain a 95% confidence interval and interpret it in terms of the problem.*

   ```
   Ital.  134  132  126  134  131  130  130  125  132  126
   Etru.  141  145  145  146  142  126  144  146  154  149  143  131
   ```

5. *Let Δ be the difference in weight between a typical pitcher and hitter, professional baseball players. Using the class code (Two-Sample Hypothesis and CI (Wilcoxon)) estimate Δ and determine a 95% confidence interval for it based on the following data. What does the interval mean in terms of the problem?*

```
Hitters:
 155   155   160   160   160   166   170   175   175   175   180
 185   185   185   185   185   185   185   190   190   190   190
 190   195   195   195   195   200   205   207   210   211   230

Pitchers:
 160   175   180   185   185   185   190   190   195   195   195
 200   200   200   200   205   205   210   210   218   219   220
 222   225   225   232
```

9.3 Estimation and Confidence Intervals Based on Means and Medians

There are estimation schemes other than the one based on the Wilcoxon. One that is commonly used is the difference of the means. It is similar to the above discussion except instead of the median of the differences, we consider the difference of the sample means.

Again, suppose we have two populations which we assume differ by at most a shift in locations. Let Δ be the difference in locations: Population Y - Population X. Remember, **shift is shift**. So for here write $\Delta = \mu_2 - \mu_1$, where μ_1 and μ_2 are the true population means of Populations X and Y, respectively. We draw random samples from each population. Let $X_1, X_2, ..., X_m$ denote the sample of size m from the first population. Let $Y_1, Y_2, ..., Y_n$ denote the sample of size n from the second population. Our estimate of Δ is $\bar{Y} - \bar{X}$.

Suppose the samples are:

- X: 8 12 16
- Y : 14 19 24

Then $\bar{X} = 12$ and $\bar{Y} = 19$. Hence the estimate of Δ is $19 - 12 = 7$.

This is only an estimate, so once again we need to get a confidence interval. But the algorithm discussed in the last section will still work. Simply replace median of differences with difference in means; i.e.,

1. Resample m X's with replacement.
2. Resample n Y 's with replacement.
3. Obtain the difference in sample means of these resamples.
4. Record this difference.
5. Repeat steps (1) through (4) 1000 times.
6. Sort the 1000 difference in means,
7. Pick off the *25th* and *976th* sorted differences in means. This is our 95% confidence interval.

This becomes very tedious, so again we have a class code, Two-Sample hypothesis test and confidence interval for the location parameter based on the mean, to obtain the point estimate and the confidence interval. It works just like one in the last section.

In the same way, we could use medians instead of means. Although this seems similar to the procedure using the Wilcoxon, it is much different.

Which procedure should we use in practice? That's a hard question to answer. The interval based on the means is not robust. So if there are outliers present, we avoid using this interval. The other two intervals are robust. Of these two, I would choose the Wilcoxon. It offers protection but it is also more powerful in most cases, giving shorter confidence intervals. The exercises will be helpful here.

Exercise 9.3.1

1. *To investigate the robustness of the three point estimates, consider the following data set:*

```
X   12 15 18
Y   16 19 25 28
```

   (a) *Obtain the three estimates: median of differences, difference in means, difference in medians. (Answers: 7, 7, 7).*

   (b) *Next replace the Y observation 28 by 2800. Obtain the three estimates: median of differences, difference in means, difference in medians. (Answers: 7, 700, 7).*

2. *We will use the next two problems to investigate the robustness of the confidence intervals.*

   (a) *Obtain comparison dotplots of the following data:*

```
X:
 31   32   33   37   37   44   44   45   45   46   50   50   50
 57   57   58   59   59   67   67

Y:
 40   45   45   47   50   52   53   53   54   54   55   61    63
 66   67   68   73   73   76   83
```

   (b) *Using the class code (Two-Sample Hypothesis and CI (Wilcoxon)) obtain the estimate of Δ and the confidence interval for it using the Wilcoxon.*

   (c) *Using the class code (Two-Sample Hypothesis and CI (mean)) obtain the estimate of Δ and the confidence interval for it using the difference in means.*

   (d) *Using the class code (Two-Sample Hypothesis and CI (median)), obtain the estimate of Δ and the confidence interval for it using the difference in medians.*

(e) Compare the intervals.

3. *Consider the samples (same as last problem but the typo of 67 on the last data point of the X's was discovered and its true value of 670 has been put in):*

```
X:
 31   32   33   37   37   44   44   45   45   46   50   50   50
 57   57   58   59   59   67   670

Y:
 40   45   45   47   50   52   53   53   54   54   55   61   63
 66   67   68   73   73   76   83
```

(a) Using the class code (Two-Sample Hypothesis and CI (Wilcoxon)) obtain the estimate of Δ and the confidence interval for it using the Wilcoxon.

(b) Using the class code (Two-Sample Hypothesis and CI (mean)) obtain the estimate of Δ and the confidence interval for it using the difference in means.

(c) Using the class code (Two-Sample Hypothesis and CI (median)) obtain the estimate of Δ and the confidence interval for it using the difference in medians.

(d) Compare the intervals.

9.4 Difference Between Proportions

There are many other two sample problems which time does not allow us to consider. You can always sign up for another stat class and I'll be happy to recommend some to you. But we would be remiss if we didn't discuss the difference in proportions problem. Also, it's a cinch with resampling.

So consider two population proportions. Examples are far too numerous to list here. One that occurs all too often is: the president's rating this month versus his rating last month. Another one that is the sign of the times is: Candidate A is worried about his financial backing. He needs to show that his popularity (i.e. proportion that will vote for him) is rising in order to attract more money. So he decides that he will come out strongly for (or against) an issue that is very popular with certain segments of the population. He then talks about this nonstop on morning, afternoon, and evening talk shows. Population I is the proportion of voters who favor him before this takes place and Population II is the proportion of voters who favor him after this change and the subsequent push on the talk shows. He must convince his backers that there has been an increase in the proportion of voters who favor him.

Ah, notation, but it's simple here. Just let p_1 and p_2 be the population proportions of Populations I and II, respectively. We are interested in estimating and determining a confidence interval for $p_2 - p_1$. So we draw random samples from Populations I and II of size m and n, respectively. Our estimate of $p_2 - p_1$ is just the difference in sample proportions. We will do the CI next, but first lets look at an example:

There are two different treatments (Drug I and Drug II) for a certain disease. Which is better? A scientist comes up with the following plan: He selects 100 patients who have the disease. He randomly assigns them to Drug I or II by a preassigned random scheme (in particular he does not decide!). The patients are treated by doctors who do not know which drug the patient is getting. At the end of the treatment period the proportion cured by each drug is tabulated. This is called a **double blind study**. Suppose the results are:

	Cured	Not Cured
Drug I	39	13
Drug II	26	22

The estimate of $p_2 - p_1$ is $26/48 - 39/52 = .54 - .75 = -.21$. So it looks like Drug I is better. What's that? Oh yes. How could I forget? Small samples, sampling error, etc. We need a confidence interval.

Resampling to the rescue. Recall that sample proportions are sample means. So we can use the algorithm of the last section, but we do need the samples. These are not the tabled values above,

but what produced the tabled results. The first sample consists of 39 *1*'s and 13 *0*'s. The second sample consists of 26 *1*'s and 22 *0*'s. Here they are:

```
X Drug I:
     1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
     1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0
     0 0 0 0 0 0 0 0 0 0 0 0

Y Drug II:
     1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
     1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0
     0 0 0 0 0 0 0 0
```

Now just call up class code for difference in means, drop these samples in the X and Y boxes and submit. To set ideas, here are the results of 100 resampled difference in proportions:

```
-0.47115385 -0.45352564 -0.43269231 -0.42948718 -0.41346154 -0.39583333
-0.34455128 -0.33012821 -0.33012821 -0.32852564 -0.32532051 -0.31089744
-0.31089744 -0.30929487 -0.30769231 -0.30769231 -0.30608974 -0.30608974
-0.30608974 -0.30608974 -0.30608974 -0.29647436 -0.29326923 -0.29006410
-0.29006410 -0.29006410 -0.28846154 -0.28525641 -0.27243590 -0.27083333
-0.26923077 -0.26923077 -0.26762821 -0.26602564 -0.26442308 -0.25480769
-0.25160256 -0.25160256 -0.25000000 -0.24839744 -0.24679487 -0.24679487
-0.24519231 -0.23717949 -0.23237179 -0.23237179 -0.23237179 -0.23237179
-0.23237179 -0.22756410 -0.22756410 -0.22756410 -0.22115385 -0.21153846
-0.21153846 -0.21153846 -0.21153846 -0.20993590 -0.20833333 -0.20512821
-0.20352564 -0.20192308 -0.20032051 -0.19230769 -0.19070513 -0.19070513
-0.18910256 -0.18910256 -0.18910256 -0.18910256 -0.18750000 -0.18108974
-0.17628205 -0.17628205 -0.17467949 -0.16506410 -0.16506410 -0.15544872
-0.15064103 -0.15064103 -0.14903846 -0.14903846 -0.14743590 -0.14583333
-0.12500000 -0.11378205 -0.10897436 -0.10737179 -0.10576923 -0.10576923
-0.09935897 -0.09455128 -0.09134615 -0.08814103 -0.08493590 -0.06410256
-0.04807692 -0.03365385 -0.01442308  0.05608974
```

Hence our estimate is $-.21$ and our confidence interval is $(-.43, -.03)$. The interval does not include 0, so we would conclude that Drug I is better. Here we will want the CI based on at least 1000 bootstraps. I did this and got the interval $(-.39, -.02)$. Hence, I get the same conclusion. Now you try it.

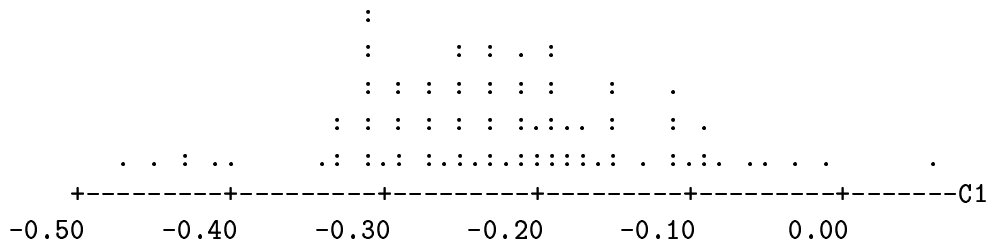
Exercise 9.4.1

1. *If you didn't do it, use class code to obtain a 95% confidence interval for the true difference in proportions for the above example of the two different drugs.*

2. *Should the president be worried? Polls of wealthy financial backers before and after she made a controversial decision were tabulated and given to her. What do you think? Base your answer on a 95% confidence interval.*

	Will Contribute	Will Not
Before Decision	68	13
After Decision	38	20

3. *In the example of the two different drugs found above, the sorted resampled differences in proportions were given. Here's a dot plot of them. Locate the estimate and the confidence interval on it.*



Chapter 10

Design of Experiments

10.1 Introduction

After reading the title of the last chapter "Estimation of Effect" you may have said, "Here's **effect**, where is **cause**?" It is one thing to observe a relationship between variables but it is another to establish cause and effect. Controlled experiments are the best way to try to establish cause and effect. In this chapter we offer two types of controlled experiments. Here's an example of an uncontrolled experiment.

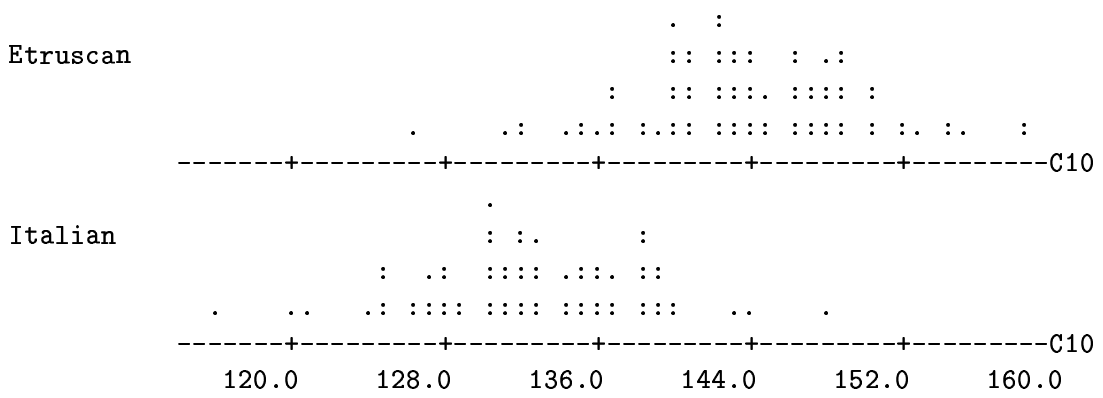
Consider again the Etruscan and Italian skull size data set. We have analyzed this from time to time. A complete two sample analysis is given at the end of this section. Based on this analysis there is a difference between Etruscan and Italian skull sizes. Recall that scientists were trying to establish a link between ancient Etruscans and modern Italians, (the Italian skulls were recent); that is, the Etruscans were native to Italy. Our statistical analysis is not supportive of that link but does it really show that the Etruscans were not native to Italy? The problem here is that there are many other variables that could cause the change in skull size : diet, environmental changes, etc. There is no way to control these variables.

This is an observational study. These studies are important. This certainly is evidence against the link, but other evidence needs to be gathered. We'll come back to these discussions later, but first lets talk about controlled experimental designs.

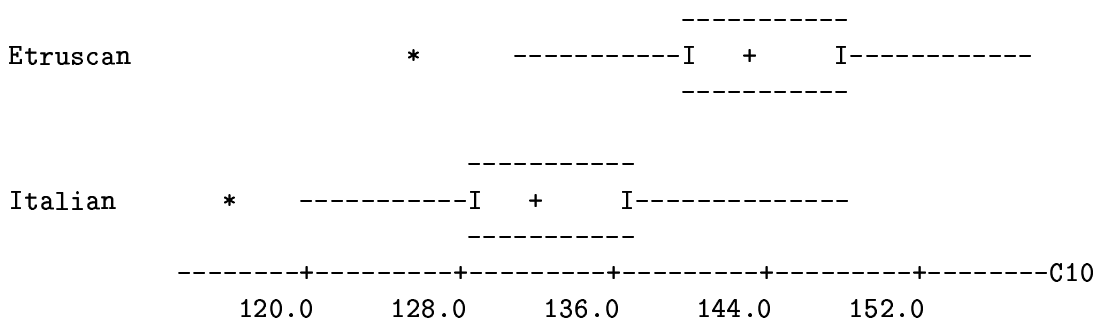
Two sample analysis of Etruscan Italian example

Were the ancient Etruscan native to Italy? To help answer this question we have two samples of skull sizes. The first sample consists of the maximum head breadths of 84 Etruscan skulls while the second sample consists of the maximum head breadths of 70 modern Italian skulls. The data is given in Appendix A.

Comparison dotplot of the data sets:



Comparison boxplot of the data sets:



The plots indicate that Etruscan skull sizes are larger than the Italian skull sizes. Furthermore, they indicate that it is a location problem.

For formal inference, let Delta be the shift in location from typical Italian skull sizes to typical Etruscan skull sizes. We will first test the hypotheses:

$$H_0 : \Delta = 0$$

versus

$$H_A : \Delta \neq 0$$

We will use the two sample Wilcoxon to test these hypotheses. (Just click on the class code: two sample wilcoxon). The Test statistics is

$$T = \#\{\text{Etruscan} > \text{Italian}\} = 5401.$$

This far exceeds the expected value of T under H_0, which is $84(70)/2 = 2940$. The p-value is .000. So we reject H_0 with high confidence. The Etruscan skull sizes are larger.

The Wilcoxon estimate of shift in location is 11mm and the confidence interval for Δ is (10, 13). So typical Etruscan head sizes are from 10 to 13mm larger than typical Italian skull sizes.

10.2 Completely Randomized Designs

We will consider two populations, but here we will call them responses due to two different treatments. So suppose we have two treatments, say, T_1 and T_2. Let X be the response under T_1 and Y be the response under T_2. T_1 may be a placebo (standard, control, old, etc.). It is easy to think of examples. For instance, consider a new diet drink. Let X be the reduction in weight following a low fat diet and Let Y be the reduction in weight following a low fat diet and which uses the diet drink. Or, let X be the durability of house paint XX and let Y be the durability of house paint YY. You only have to paint a house once to realize the importance of this experiment.

Now in a controlled experiment the treatments have an effect on the response variables, but all other variables are kept in control (at the same level), as much as possible. When investigators get ready to do a controlled experiment they often sit and discuss all variables which could have a bearing on the response. This is a very important part of the experiment. For example consider the diet example. What else has a bearing on weight reduction? Exercise, life style, age, heredity, sex, physical condition, etc. There are many, many variables. These will have to be controlled as well as possible. In certain cases, you may not be able to control a variable. Such variables are called covariates and there are certain designs where their effects are taken into account, but we will not consider these in this course. But needless to say, uncontrolled variables may jeopardize the experiment.

We will assume the location assumption of the last chapter, which we rewrite as,

The distributions of Y and X differ by at most a shift in locations (centers), say Δ.

Again, we at least have visual checks, comparison boxplots, dotplots with which to assess this assumption.

Our target parameter is Δ, this is the effect. There is a natural null hypothesis, i.e.,

- H_0: $\Delta = 0$.

Alternatives may be one or two sided. For convenience, lets assume the alternative is

- H_A: $\Delta \neq 0$.

So we want to test hypotheses, estimate the effect, and find a confidence interval for it.

In this section, we consider a **completely randomized design, (CRD)**.

- We randomly select N experimental units at random from our reference population. We randomly assign m of these units to Treatment 1 and the other n of these units to Treatment 2. The experiment (study) is run for a pre assigned time and during this time all other

variables are kept under control. At the end of the assigned time, we measure the responses for the m units which were assigned to Treatment 1, call them $X_1, ..., X_m$. And we measure the responses for the n units which were assigned to Treatment 2, call them $Y_1, ..., Y_n$. It is assumed that these responses are independent of one another.

Too wordy? Lets look at the experiment which produced the quail data discussed in Chapter 1. Recall it was an experiment involving a drug which hopefully reduces LDL cholesterol levels. There were two treatments: Placebo (Treatment 1) and an active drug compound which hopefully reduces LDL cholesterol (Treatment 2). Let Δ denote the effect (i.e. typical quail's LDL level on placebo minus typical quail's LDL level on treatment or the true mean level of a quail on placebo minus the mean level of a quail on the treatment). Our hypothesis of interest is H_0: $\Delta = 0$ versus H_A: $\Delta > 0$. We will also estimate Δ and find a confidence interval for it using the Wilcoxon analysis of the last chapter.

The Experiment: 30 quail were randomly selected (these are the experimental units) from a reference population. 20 were randomly assign to Treatment 1 (a placebo) and the other 10 to Treatment 2. For those on Treatment 2 an active drug compound was mixed with their diet. Those on Treatment 1 had the same diet without the drug compound. Over the course of the experiment, the quail were treated the same. Same amount of exercise, same types of pens, etc. At the end of the time period their LDL cholesterol levels were measured.

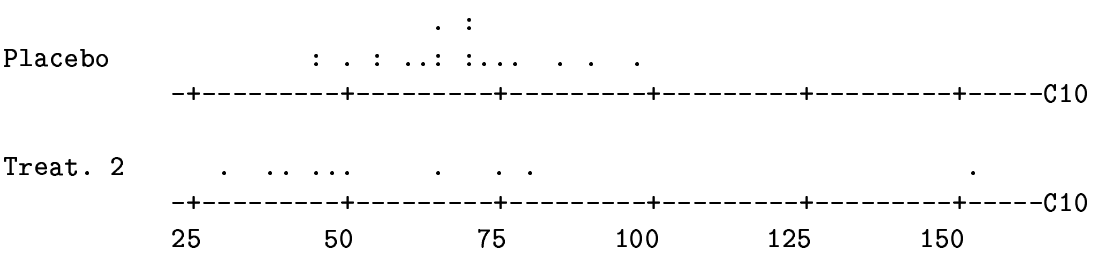
The CRD of course produces two samples. The statistical analyses described in the last Chapter would be appropriate. Remember to do comparison dotplots or boxplots to check on the location assumption. Lets look at the quail experiment.

The data are:

```
Placebo:      64  49  54  64  97  66  76  44  71  89
              70  72  71  55  60  62  46  77  86  71

Treatment2:   40  31  50  48 152  44  74  38  81  64
```

A comparison dotplot is:



There is not much data here. Ignoring the outlier, the scales do not seem to be that different. The LDL levels of the treated quail seem to be shifted lower.

The value of the Wilcoxon is $T = \#\{\text{Placebo } > T_2\} = 134.5$. Using class code (Two-Sample hypothesis test and confidence interval for the location parameter based on the Wilcoxon), the *p-value* is .055. Hence, there is evidence at the .05 level of significance that the treated quail have lower LDL levels than the placebo group. So the drug would be earmarked for further study. The estimate of the effect Δ is 14 and a confidence interval is $(-8.5, 25.5)$. Notice that the confidence interval contains 0. This does not contradict the test because it was a one-sided test.

Exercise 10.2.1

1. *From Rasmussen, Statistics with Data Analyses, CA: Brooks-Cole. Investigators wanted to compare the drugs morphine and nalbuphine on their effect in changing pupil size. So they selected 11 volunteers and randomly assigned 6 of them to several does of morphine and the other 5 to several doses of nalbuphine. Before receiving the drug their pupil sizes were measured. After waiting a prescribe amount of time after the dosages of the drugs they measured the change in diameters of the subjects pupils. The data are:*

Treatment	Change in pupil diameter					
Morphine	.08	.8	1.0	1.9	2.0	2.4
Nalbuphine	-.3	.0	.2	.4	.8	

   (a) *Obtain comparison dotplots of the data. Comment.*

   (b) *Let Δ be the effect of the different drugs on pupil size (morphine minus nalbuphine). We want to test*

   - *H_0: $\Delta = 0$ versus*
   - *H_A: $\Delta \neq 0$*

   Obtain the Wilcoxon test statistic and compare it to what we would expect under Ho. Use the class code (Two-Sample Hypothesis and CI (Wilcoxon)) to determine the p-value. Conclude in terms of the problem.

   (c) *Obtain the estimate of and a confidence interval for the effect, using the Wilcoxon. Conclude in terms of the problem.*

2. *Suppose we wanted to investigate the difference in the thicknesses of a pages in two books, but all we had was a ruler with eighths-of-inches. Set up an experimental design to do this. (Note you can measure the thickness of a bunch of pages even though you cannot measure a page). What plots could you use here? What are the parameters of interest? What are the hypotheses? What analysis would you use?*

3. *From Rasmussen, Statistics with Data Analyses, CA: Brooks-Cole. Researchers wanted to study the effect of regular alcohol assumption on plasma estrogen. The participants in the experiment were 20 adult male squirrel monkeys, of similar age and good health (What variables are being controlled here?). They randomly divided the monkeys into two equal sized groups. Monkeys in the alcohol group consumed a steady diet of 12% ethyl alcohol while those in the control group had the same diet with no alcohol. The results are:*

Alcohol:	3.17	2.52	2.59	4.25	3.27	4.92
	5.46	2.83	4.80	2.26		
Control	6.57	5.81	5.63	5.75	4.54	5.35
	4.16	5.12	4.69	4.52		

   (a) *Obtain comparison dotplots of the data. Comment.*

   (b) *Let Δ be the effect of the different drugs on pupil size (morphine minus nalbuphine). We want to test*

   - H_0: $\Delta = 0$ *versus*
   - H_A: $\Delta \neq 0$

   Obtain the Wilcoxon test statistic and compare it to what we would expect under H_0. Use the class code (Two-Sample Hypothesis and CI (Wilcoxon)) to determine the p-value. Conclude in terms of the problem.

   (c) *These sort of experiments produced what warning level?*

10.3 Randomized Paired Design

Noise is often the villain in the analysis of an experimental design. There is just too much noise to see the target. The design we introduce next is an effective noise reducer. The price is a loss of information (nothing comes free). Also, as you will see, often it is not possible to do.

The setup is the same as the completely randomized design. We have two treatments, T_1 and T_2, applied to a response. We still want to test an estimate the effect, Δ. The difference is that we can select a pair of experimental units. For example, identical twins on a study involving humans, the same house for a study on two house paints (halve the North wall), the same field for a study on two varieties of wheat, etc. As I said, "this may be impossible to do."

- **Randomized Paired Design:** Randomly select n paired experimental units from the reference population. Within a pair, randomly assign one of the pair to Treatment 1 and the other to Treatment 2. The experiment (study) is run for a pre assigned time and during this time all other variables are kept under control. At the end of the assigned time, we measure the responses for the n paired experimental units. Letting X and Y denote the responses for Treatments 1 and 2, respectively the data are in the paired form: $(X_1, Y_1), ..., (X_n, Y_n)$.

Although, the pairs are independent, within a pair there is dependency. In fact, the more dependency within a pair, the more the noise reduction. Hence, the two-independent-sample analysis of Chapter 9 is out. The key is that Δ is still a typical Y minus a typical X, i.e., read that as Y_i minus a typical X_i where the subscript i refers to the ith pair. Thus the sample of interest **IS THE DIFFERENCES**. That is,

$$D_1 = Y_1 - X_1, D_2 = Y_2 - X_2, ..., D_n = Y_n - X_n$$

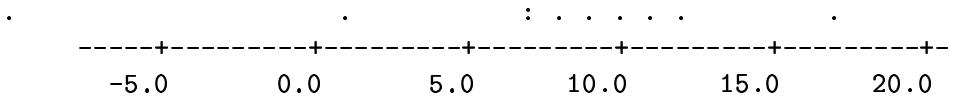
Too wordy! Lets have an example. This is taken from Siegel, *Nonparametric Statistics*. Ten pairs of identical twins, age 4, were randomly selected for an experiment to investigate how nursery school affects the the social awareness of a 4 year old. For each pair, one twin was randomly assigned to go to nursery school while the other stayed home. At the end of the time period, all 20 took the same test and their scores were recorded (bigger means more socially aware). The data are, pair number in column 1, nursery school twin in column 2, response of stay-at-home twin in column 3, difference in responses in column 4:

pair	N	H	D
1	74	63	11
2	43	33	10
3	61	41	20
4	79	67	12
5	80	65	15

6	73	80	-7
7	56	43	13
8	98	84	14
9	84	74	10
10	52	48	4

There are two immediate observations form this data set:

1. The twin who went to nursery school seems to be more socially aware. A dotplot on the differences is given next and a formal analysis is discussed below.



2. The pairing has really cut the noise. The range of the nursery school scores is $98 - 43 = 55$, the range of the stay-at-home scores is $84 - 33 = 51$, but the range of the differences is $20 - (-7) = 27$. Hence the noise level has been cut by about 1/2. The reason this reduction in noise takes place here is that four year olds are all over the map on social relationships. Some are ready for school, some are far from ready, some talk continuously while others are very shy, etc. And the scores reflect this, (note the scores 98 and 43 in column 1). But identical twins are alike in social awareness (before the experiment). So if one twin scores high then so does the other while if one twin scores low so does the other. This certainly makes sense for these our identical twins. Within a pair the scores are much more similar and, hence, the differences are smaller.

Alright! I hear you clamoring. This is *ad hoc*. We want *p-values*. We WANT estimates and confidence intervals. Put up or shut up.

We can't use Chapter 9 but since we have a single sample, the D's, we can use Chapter 7 for estimates and confidence intervals. For example, we can estimate Δ by the median of the paired-differences which is *11.5*. A confidence interval for the median is *(10, 14.5)* which can be obtained using the class code (One sample bootstrap confidence intervals for the population mean and median) and typing in the paired differences in the big data box. Selecting median and submiting produces the bootstrap confidence interval for the median.

Exercise 10.3.1

1. *Finish the example for the twin data. Recall the paired differences were:*

pair	N	H	D
1	74	63	11
2	43	33	10
3	61	41	20
4	79	67	12
5	80	65	15
6	73	80	-7
7	56	43	13
8	98	84	14
9	84	74	10
10	52	48	4

   (a) *Obtain the value of the Wilcoxon test statistic. (Actually determine the number of negative averages (2) and subtract it for 10(11)/2.*

   (b) *Obtain the p-value for a two sided-test. Use the class code of course (Wilcoxon for paired designs). Conclude in terms of the problem.*

   (c) *Obtain (from class code) the estimated effect and the associated confidence interval. Conclude in terms of the problem.*

2. *From Cushney and Peebles (1905)a, J. of Phisiology: Ten patients were selected for a study. The average number of hours that they slept was deterimed. There were two parts to the study. In Part 1, they were given by a flip of the coin one of two drugs, Laevo and Dextro, and the average (over a week) number of excess hours (over their usual average) was recorded. In Part 2 (after a wash out period), they were given the other drug, and the average (over a week) number of excess hours (over their usual average) was recorded. The data are:*

Patient	Dextro	Laevo
1	0.7	1.9
2	-1.6	0.8
3	-0.2	1.1
4	-1.2	0.1
5	-0.1	-0.1
6	3.4	4.4
7	3.7	5.5
8	0.8	1.6
9	0.0	4.6
10	2.0	3.4

(a) *Obtain the value of the Wilcoxon test statistic, (diff = D - L).*

(b) *Compare it what you would expect under* H_0.

(c) *Obtain the p-value for a two sided-test. Use the class code (Wilcoxon for paired designs) of course. Conclude in terms of the problem.*

(d) *Obtain (from class code) the estimated effect and the associated confidence interval. Conclude in terms of the problem.*

3. *The data below are some measurements recorded by Charles Darwin in 1878. They consist of 15 pairs of heights in inches of cross-fertilized plants and self-fertilized plants, Zea mays, each pair grown in the same pot.*

POT	CROSS	SELF
1	23.500	17.375
2	12.000	20.375
3	21.000	20.000
4	22.000	20.000
5	19.125	18.375
6	21.550	18.625
7	22.125	18.625
8	20.375	15.250
9	18.250	16.500
10	21.625	18.000
11	23.250	16.250
12	21.000	18.000
13	22.125	12.750
14	23.000	15.500
15	12.000	18.000

(a) *Obtain the value of the Wilcoxon test statistic, (diff = C - S).*

(b) *Compare it what you would expect under* H_0.

(c) *Obtain the p-value for a two sided-test. Use the class code (Wilcoxon for paired designs). Conclude in terms of the problem.*

(d) *Obtain (from class code) the estimated effect and the associated confidence interval. Conclude in terms of the problem.*

10.4 Signed-Rank Wilcoxon

Another analysis though is based on the one-sample Wilcoxon. Recall the hypotheses we want to test are:

- H_0 : $\Delta = 0$ versus
- H_A: $\Delta \neq 0$.

where Δ is the true location of the paired differences.

Consider our **simple example**:

Y	X	D
6	10	-4
13	15	-2
30	25	5
40	31	9

Thus the 4 differences are: -4, -2, 5, 9. Notice that the positive numbers are slightly larger. So the edge is to the positive side; although the test should be far from significant.

Our analysis is based on the one sample Wilcoxon test statistic. This is often referred to as the **Signed-Rank Wilcoxon**. So, let us label it the **SRW** procedure.

The SRW statistic is

$$W = \#\Big\{\frac{D_i + D_j}{2} > 0\Big\}$$

We will often refer to these paired-averages, $\frac{D_i+D_j}{2}$, by the name **Walsh averages**. For each pair (D_i, D_j) of differences we only count the corresponding Walsh average once. An easy way to calculate these averages is by the table given below. Sort the D's. The columns are the D's and the rows are the D'd too. The entries are the Walsh averages. Since we only need one, just the top half of the table is formed as shown.

ave. with	-4	-2	5	9
	**********	**********	**********	**********
-4	* -4	-3	.5	2.5
-2	*	-2	1.5	3.5
5	*		5	7
9	*			9

Our SRW statistic is the number of positive averages in the table. Hence the test statistic is $W = 7$.

How many positive averages would you expect if H_0 is true. Well half should be positive and half should be negative. Since there are in general $\frac{n(n+1)}{2}$, where n is the number of differences, in this case we expect W to be $.5 * (4(5)/2)$ or 5. At 7, W is not far from what you expect it to if the null hypothesis is true. Again we need a *p-value* which we can get by resampling. (How would you do this resampling? It must be under H_0. We come back to this in a moment. Just assume we can do it). This can be obtained using the class paired sample analysis code. Drop the differences *-4, -2, 5, 9* into the data box. The test statistic and *p-value* are returned to you. If you do this, you will get a *p-value* of about .37. You certainly cannot reject.

The point estimate of the effect, Δ, is the median of $(D_i + D_j)/2$. Looking back up at the table, you see the median is $.5(1.5 + 2.5) = 2$. A confidence interval is based on resampling the paired differences $-4, -2, 5, 9$. The class paired sample analysis code will also return the point estimate and the confidence interval. Try it. You should get 2 as the point estimate. My confidence interval (based on 1000 resamples is $(-4, 9)$, which contains 0 (hardly a surprise here, right?).

How do we do the resampling for the *p-value*? H_0 must be true; i.e., the true Δ must be 0. Just take the differences (in this case *-4, -2, 5, 9*) and subtract off the point-estimate (in this case 2). This will center the differences for the Wilcoxon around 0. Our table for these "centered" differences is:

```
ave. with   -6    -4    3   7
          ************************
     -6   *  -6     -5   -1.5   .5
     -4   *          -4    -.5   1.5
      3   *                3    5
      7   *                     7
```

The Wilcoxon test statistic here is 5, just what you expect under H_0. The class code does this type of resampling for its *p-value*.

10.5 Difference Between Proportions : Dependent Samples

In Chapter 9 we talked about a difference in proportions. This involved two independent samples. Another difference in proportions which involves one sample occurs quite frequently and we would be remiss if we didn't discuss it.

As an example suppose a national poll was conducted in which voter were asked who they are going to vote for if the election was held today : Bush, Gore, Nader, etc., I don't know, I never vote, etc. Let p_B and p_G be the two proportions of citizens who will vote for Bush and Gore, respectively. Then the difference of interest is $p_B - p_G$. The estimate is, of course, $\hat{p}_B - \hat{p}_G$ the sample proportions in our poll.

We need a confidence interval for it. We could get a confidence interval by resampling but in this case we will use the CLT and give a formula. The main reason for doing this is that these types of polls occur all the time, so in the future you may want to impress your friends and obtain the error margin of the poll. Suppose in the poll n votes were sampled (at random!!). The error in the poll is

$$\text{Error } = 1.96\sqrt{\frac{\hat{p}_B + \hat{p}_G - (\hat{p}_B - \hat{p}_G)^2}{n}}$$

and the 95% confidence interval for $p_B - p_G$ is

$$\hat{p}_B - \hat{p}_G \pm 1.96\sqrt{\frac{\hat{p}_B + \hat{p}_G - (\hat{p}_B - \hat{p}_G)^2}{n}}$$

Conclusions based on the confidence interval:

1. If 0 is in the confidence interval then the results are inconclusive. The paper might use the term "too close to call".

2. If the confidence interval consists entirely of negative values then the result is significant and the poll is predicting that Gore will win. Remember the poll begins with, "If the election was held today, ... ". The poll is only good for "this time". Things can change, but at the moment the poll is predicting that Gore will win, with 95% confidence in that prediction.

3. If the confidence interval consists entirely of positive values then the poll is predicting that Bush will win, with 95% confidence in that prediction.

Example : Poll was over 1500 voters. The results are

Bush	Gore	All Others
580	595	325

Then $\hat{p}_B = 580/1500 = 0.3867$ and $\hat{p}_G = 595/1500 = 0.3967$; hence the error is

$$\text{Error } = 1.96\sqrt{\frac{\hat{p}_B + \hat{p}_G - (\hat{p}_B - \hat{p}_G)^2}{n}} = \sqrt{\frac{.7834 - .0001}{1500}} = .0448$$

So the 95% confidence interval is

$$(.3867 - .3967) \pm .0448$$

$$(-.0348\,,\ .0548)$$

Based on this confidence interval (0 is in it), we would say the election is too close to call.

Chapter 11

Regression : Second Pass

11.1 Introduction

"Regression is better, the second time around...". Nothing like those ancient oldies.

We discussed regression in Chapter 1. It is one of the most widely used techniques in statistics for various reasons. In this chapter, we want to tie in some inference and discuss it from both experimental designs and observational studies point-of-views. Since we have been talking about experimental designs, we shall begin with it.

11.2 Regression Experimental Designs: A Beginning Example

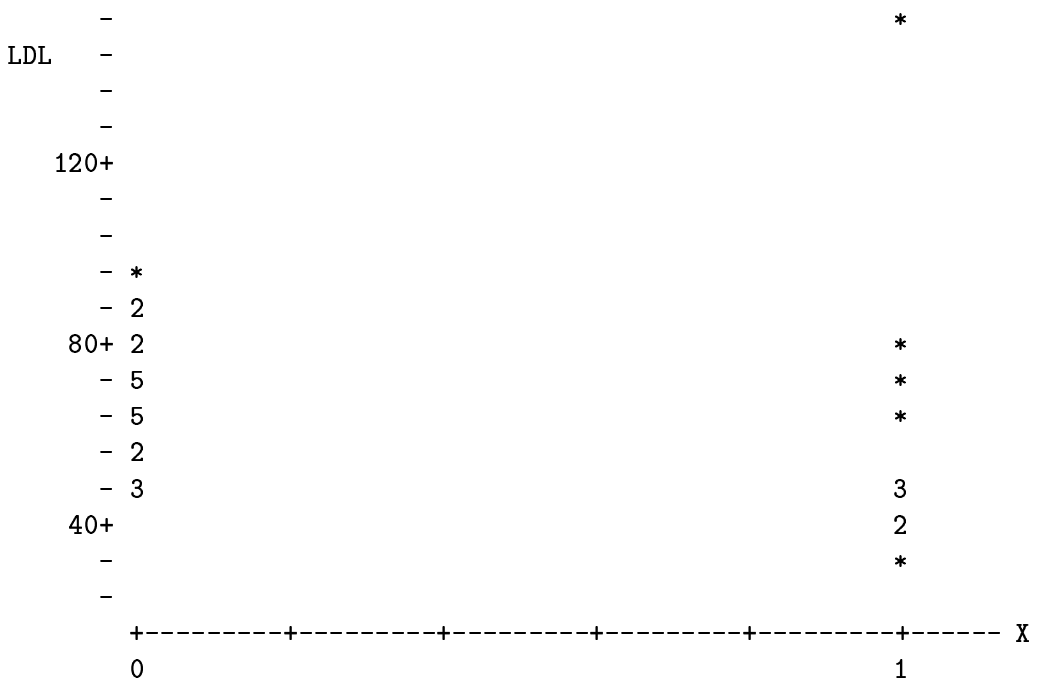
Lets begin with an example of a completely randomized design and set it up as a regression study. Then the generalization is easy.

Recall the cholesterol study on quail discussed in the last chapter: The Experiment: 30 quail were randomly selected (these are the experimental units) from a reference population. 20 were randomly assign to Treatment 1 (a placebo) and the other 10 to Treatment 2. For those on Treatment 2 a active drug compound was mixed with their diet. Those on Treatment 1 had the same diet without the drug compound. Over the course of the experiment, the quail were treated the same. Same amount of exercise, same types of pens, etc. At the end of the time period their LDL cholesterol levels were measured. The data are:

```
Placebo:      64  49  54  64  97  66  76  44  71  89
              70  72  71  55  60  62  46  77  86  71

Treatment2:   40  31  50  48 152  44  74  38  81  64
```

This doesn't look like a regression problem but it is. Set the independent variable to $x = 0$ if the response (LDL) level is from a quail in the placebo group, and set the independent variable to $x = 1$ if the response (LDL) level is from a quail in the active drug group. Thus we have 20 x's set at 0 and 10 x's set at 1. Our scatter plot would be 64 versus 0, ... , 71 versus 0, 40 versus 1, ..., 64 versus 1. Hence, the plot is



The numbers stand for how many points are at that location; i.e., the 5 means that there are 5 points at that location. The * means that there is one point at that location. So count them to see that indeed there are 20 points over $x = 0$ and 10 points over $x = 1$. Note the huge outlier in the treated group that we talked about in the last chapter.

Now eyeball a line through the points, ignoring the outlier (a robust eyeball fit). Here's what I did: I chose the line that goes through $(0, 77)$ (that's between the 2 and the top 5 over $x = 0$) and the point $(64, 1)$ (that's the * above the 3 over $x = 1$). NOW TRACE THIS LINE IN!!!!!!!!!!!!!

What's the slope of my eyeball fit? That's easy. The change in Y over the change in x is: $(77 - 64)/(0 - 1) = -13$. Now more importantly, what does this slope mean? If you think about it, it is an estimate of the change in centers of the two groups. That is, it is an estimate of the effect between the two treatment groups. Recall from the last chapter that the Wilcoxon estimate of the effect was -14. Hence all completely randomized designs can be put into a regression context. This is true of paired designs too but we will not go into it, (you can always take additional statistics courses).

11.3 Regression Experimental Designs

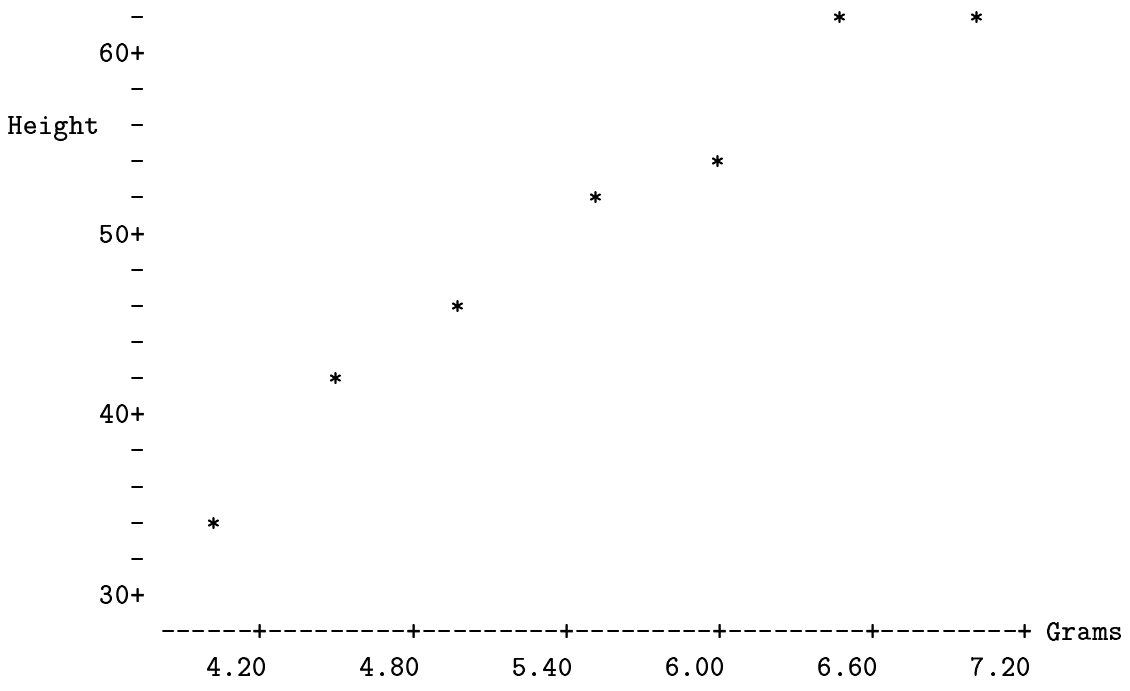
Consider a response which is related to an independent variable. For the last example, suppose we measure the LDL level of a quail given a specific dose level of the drug. Now we vary the dose level for different quail. This would be an example of a regression experimental design. Instead of introducing a lot of notation, here's a simple definition of such a design. Read through it and then read the examples that follow.

Controlled Regression Design. We want to investigate a response over several different levels of an independent variable. Randomly select n experimental units and randomly assign a preassigned number to each level of the independent variable. Keep all other variables which could influence the response at a predetermined fixed level. At the end of the experiment time period measure the responses.

Suds Example. Here is another simple example (From, Draper and Smith (1966), *Applied Regression Analysis*, New York: Wiley): For a manufacturer of dishwasher detergent, the height of soap suds in the dishpan is important, even though it is a psychological factor. The suds height should depend on the amount of detergent used. So 7 pans of water were prepared. To each (by random assignment) an amount of dishwasher detergent was added. Then the dishpan was agitated for a set amount of time and the height of the suds was measured. Some of the variables controlled here were: temperature of water, time of agitation, type of dishpan, and measurement of the height conducted in the same way. The data are:

Grams of Product (X):	4	4.5	5.0	5.5	6.0	6.5	7.0
Height of Suds mm(Y):	33	42	45	51	53	61	62

The plot of interest is a scatter plot of Height versus Grams:



There is an increasing relationship between height of suds and grams of detergent. It looks fairly linear except it seems to taper off for the high suds levels. Using the regression module, we fit the linear model:

```
Height of Suds = a + b*(Grams of detergent) + error
```

We used the Wilcoxon option. The prediction equation is

```
Predict Height = -3.33 + 9.67*(Grams of detergent)
```

The estimate of slope is 9.67, that is we estimate the height of suds to increase 9.67 mm for each additional gram of detergent. We could also use the equation to predict the height of the suds level for values of grams of detergent. For instance, for 6 gm of detergent we predict the suds level to be

```
Predicted height = -3.33 + 9.67*6 = 54.69
```

Inference. The only inference we will consider is a confidence interval for the slope parameter. The estimation of slope is just that, an estimate. We need to estimate how much it missed the true slope by. We will also use this confidence interval to test the hypotheses:

- H_0: $b = 0$ versus H_A: $b \neq 0$.

Our decision rule is simple, we reject H_0 in favor of H_A if *0* is not in the confidence interval for b.

We will use a Central Limit Theorem confidence interval for b. Besides the estimation class code prints out the standard errors of the estimates. These are in the table which follows the regression equation. The first numerical column gives the estimate and the second column gives the estimated standard deviation of the estimate (i.e., the standard error). Our confidence interval is then of the form:

- $\hat{b} \pm 1.96 \times$ Std. Err.

Suds Example, continued. From the class code, the estimated slope was 9.67 with *Stdev = 1.21*. Hence the confidence interval is:

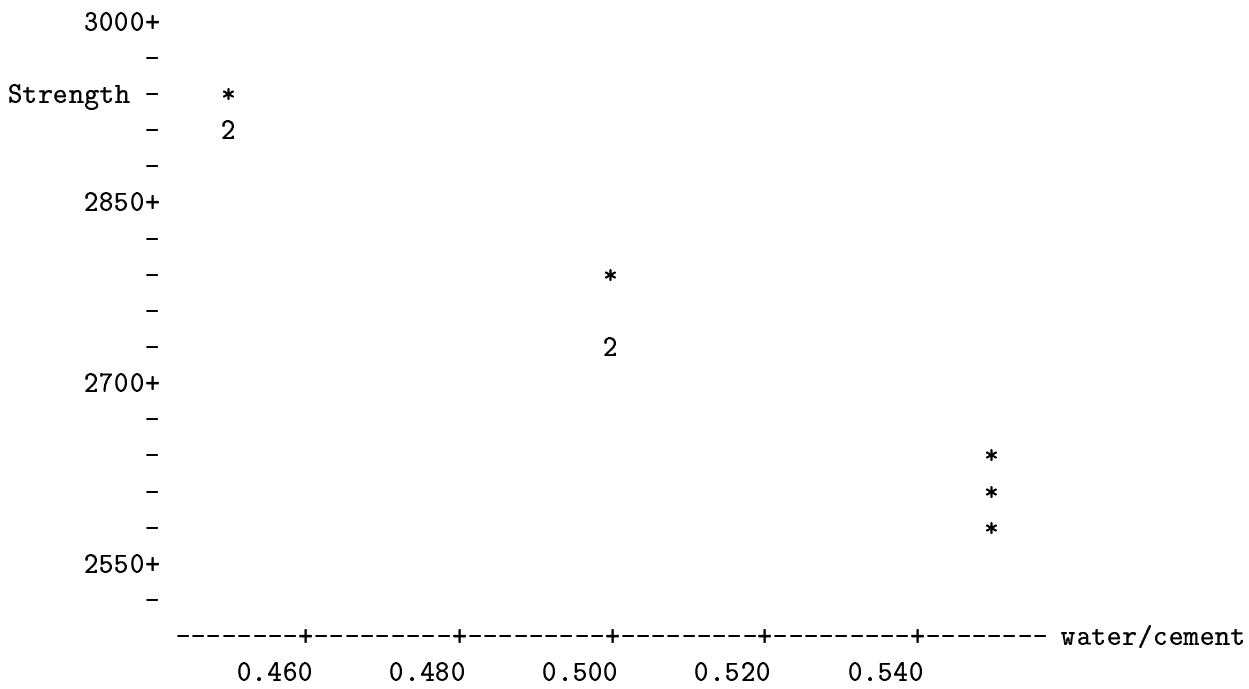
- $9.67 \pm 1.96 \times 1.21$ or $(7.30, 12.04)$.

Hence we estimate the height of the suds to increase 7 to 12 mm in height for every gram of additional detergent. The confidence interval does not include 0 so we reject H_0 in favor of H_A and we conclude that there is a positive linear relationship between the height of suds and the amount of detergent.

Concrete Example. (From Vardeman (1994), *Statistics for Engineering Problem Solving*, Boston: PWS.) A study was performed to investigate the relationship between the strength (psi) of concrete and water/cement ratio. Three settings of water to cement were chosen (.45, .50, .55). For each setting 3 batches of concrete were made. Each batch was measured for strength 14 days later. All other variables were kept constant (mix time, quantity of batch, same mixer used (which was cleaned after every use), etc.). Here's the data:

Water/cement	0.45	0.45	0.45	0.50	0.50	0.50	0.55	0.55	0.55
Strength	2954	2913	2923	2743	2779	2739	2652	2607	2583

Here's a scatter plot:



The plot indicates a decreasing relationship between strength of concrete and water to cement ratio,; i.e., the more water one uses, the weaker the cement. Clicking on regression module, and using the Wilcoxon estimate, we obtain the prediction equation

- $Strength = 4345 - 3160 * (w/c)$.

What does the estimate of the slope mean?

Keeping the range of x in mind (.1), it is best to phrase this as for each additional tenth of water to cement, we estimate the strength of the concrete to drop by 316 psi. From the class code, we form a confidence interval for slope by:

- $-3160 \pm 1.96 \times 277.4$ or $(-3703.7, -2616.3)$.

Since 0 is not in the confidence interval we reject H_0. One way of concluding would be: for each *additional tenth* of water to cement, we estimate the strength of the concrete to drop from 262 to 370 psi.

There is a lot more to experimental designs than we have covered in this chapter. The effects of more than one variable at a time changing on the response can be analyzed. These variables are set at certain values (the design of the experiment) and other variables are controlled. If they

cannot be controlled then they are recorded. These will be used as covariates to adjust the analysis. These items are beyond this course. In fact there are several courses you can take at Western on experimental design.

There are many situations, though, where we can not design an experiment, (set the levels of the independent variables). These are basically observational studies which we discuss in the next section.

Exercise 11.3.1

1. *(From Bhattacharyya and Johnson (1977), Statistical Concepts and Methods, New York: Wiley). A study was performed to investigate the relationship between speed and stopping distance for an automobile. 10 cars were selected (same year, model, etc.). Each was driven at preassigned speed and when the driver attained that speed the he applied the brakes. The distance to a complete stop was then measured. The data are:*

Speed (X)	:	20	20	30	30	30	40	40	50	50	60
Distance (Y):		16.3	26.7	39.2	63.5	51.3	98.4	65.7	104.1	155.6	217.2

   (a) *Assuming this was a designed experiment what other variables besides car were controlled?*

   (b) *Scatter plot this data (Y versus X). Comment on the plot. Does it look linear?*

   (c) *Regardless of your discussion in the last part, use the regression module to fit the model. Predict the stopping distance for an initial speed of 35. Predict the stopping distance for an initial speed of 55.*

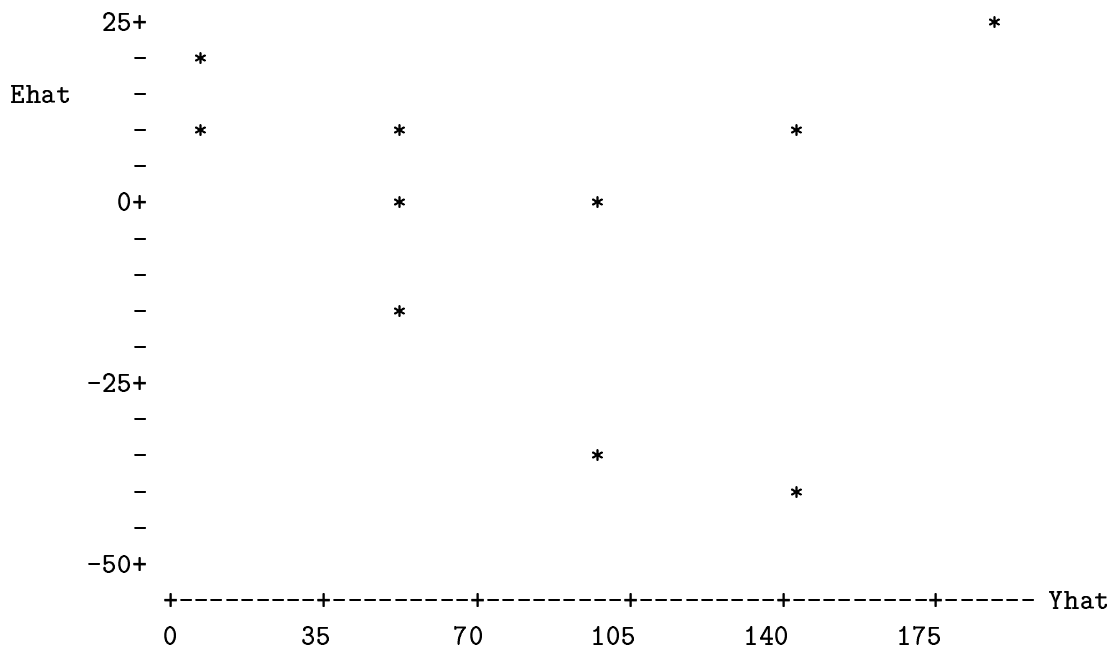
   (d) *Use your predictions in the last part to plot your fit on the scatter plot. Comment? Interpret the estimate of slope.*

   (e) *Obtain a confidence interval for the slope parameter. What does it mean in terms of the problem? Use it to test H_0 . Conclude in terms of the problem.*

   (f) *Determine the fit and the residual for the response 98.4 at $x = 40$.*

   (g) *Next obtain the residual plot. Does the observation (40, 98.4) seem to be an outlier? Is the scatter random? See the next problem for the answer.*

2. *Here is the residual plot for the last problem:*



It is not a random scatter. Sometimes a simple transformation will help. Consider the square root of the stopping distances. These are given by:

Speed (X) :	20	20	30	30	30	40	40	50	50	60
SqrtDistance	4.03	5.16	6.26	7.96	7.16	9.91	8.10	10.20	12.47	14.73

Repeat the last problem using these responses. Notice interpretation changes. As you will see, the residual plot improves considerably but there are still problems with it.

3. *(From Vardeman (1994), Statistics for Engineering Problem Solving, Boston: PWS.) A study was performed to investigate the relationship between the carburetor jetting size and the time of a Camaro for a quarter-mile run. The data are:*

Jet Size	76	68	70	72	74	76
Time	15.08	14.60	14.50	14.53	14.79	15.02

   (a) *Assuming this was a designed experiment what other variables besides car model were controlled?*

   (b) *Scatter plot this data. Comment on the plot. Does it look linear?*

   (c) *Regardless of your discussion in the last part, use the regression module to fit the model. Predict the time for a jet size of 76. Predict the time for a jet size of 68.*

(d) Use your predictions in the last part to plot your fit on the scatter plot. Comment? Interpret the estimate of slope.

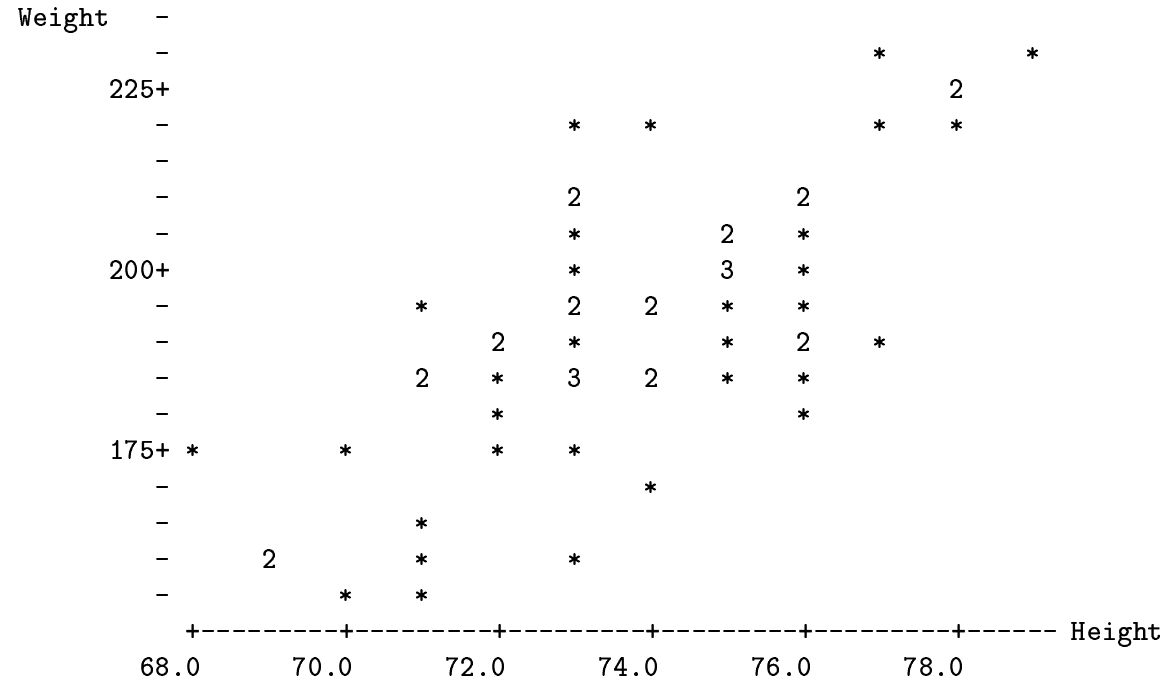
(e) Obtain a confidence interval for the slope parameter. What does it mean in terms of the problem? Use it to test H_0*. Conclude in terms of the problem.*

11.4 Observational Studies

In order to study the relationships among variables, observational studies are performed. Unlike controlled experimental designs where only certain variables are allowed to vary (at prespecified levels), in observational studies the variables are observed and recorded. Often some of the variables are controlled as much as possible. Consider a long term study on a drug involving humans where a variable that needs to be controlled is diet. The diet guidelines are set but these will probably be broken from time to time (or maybe often) by some of the human subjects. Contrast this with a lab setting, where the diet of animals can be controlled.

In observational studies, cause and effect are hard (often impossible) to establish. But associations and predictabilities among variables can be investigated. Such associations and predictabilities may be further studied in a lab setting.

Here's a simple example. Let Y be the weight of a baseball player and let X be the height of a baseball player. Recall the scatter plot which is given by:



Here is the data set:

Height (X)

74	75	77	73	69	73	78	76	77	78	76	72	73
73	74	75	72	75	76	76	72	76	68	73	69	76
77	74	75	73	79	72	75	70	75	78	73	75	74
71	73	76	73	75	73	73	74	72	73	71	71	71
73	74	76	71	76	71	70						

Weight (Y) This paired data, e.g., the 74 goes with 218, etc.

218	185	219	185	160	222	225	205	230	225	190
180	185	200	195	195	185	190	200	180	175	195
175	185	160	211	190	195	200	207	232	190	200
175	200	220	195	205	185	185	210	210	195	205
175	190	185	190	210	195	166	185	160	170	185
155	190	160	155							

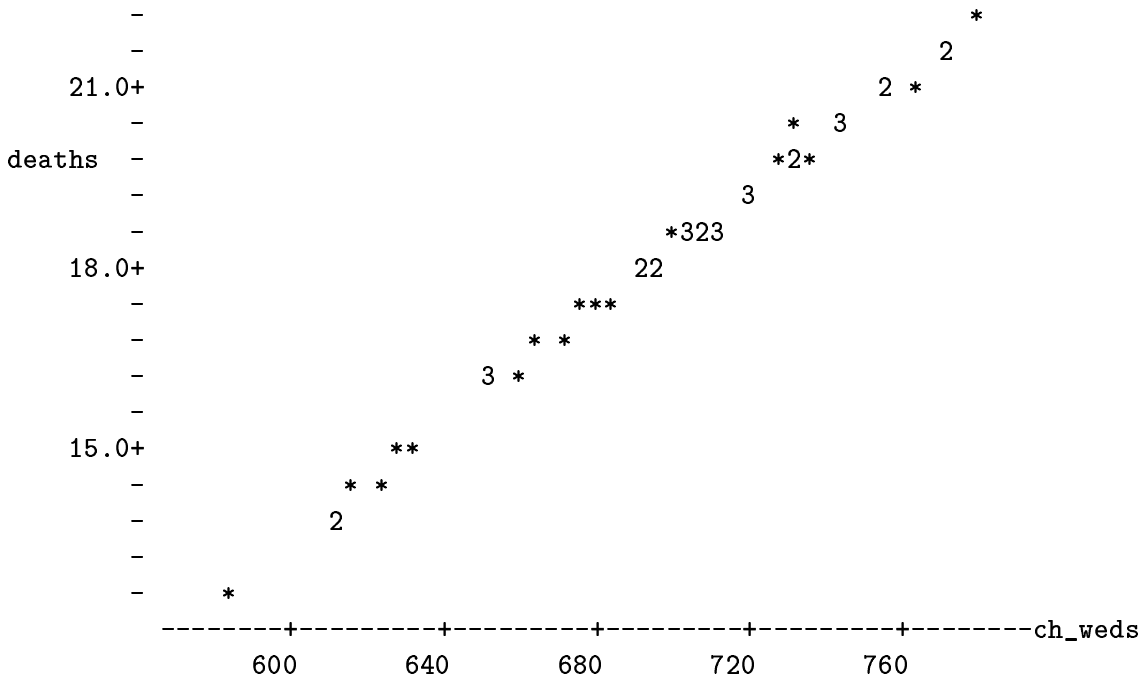
If we enter this data set into the data box and choose regression, we get the prediction equation (Wilcoxon):

- *Predict Wt = -228.57 + 5.71*Height* .

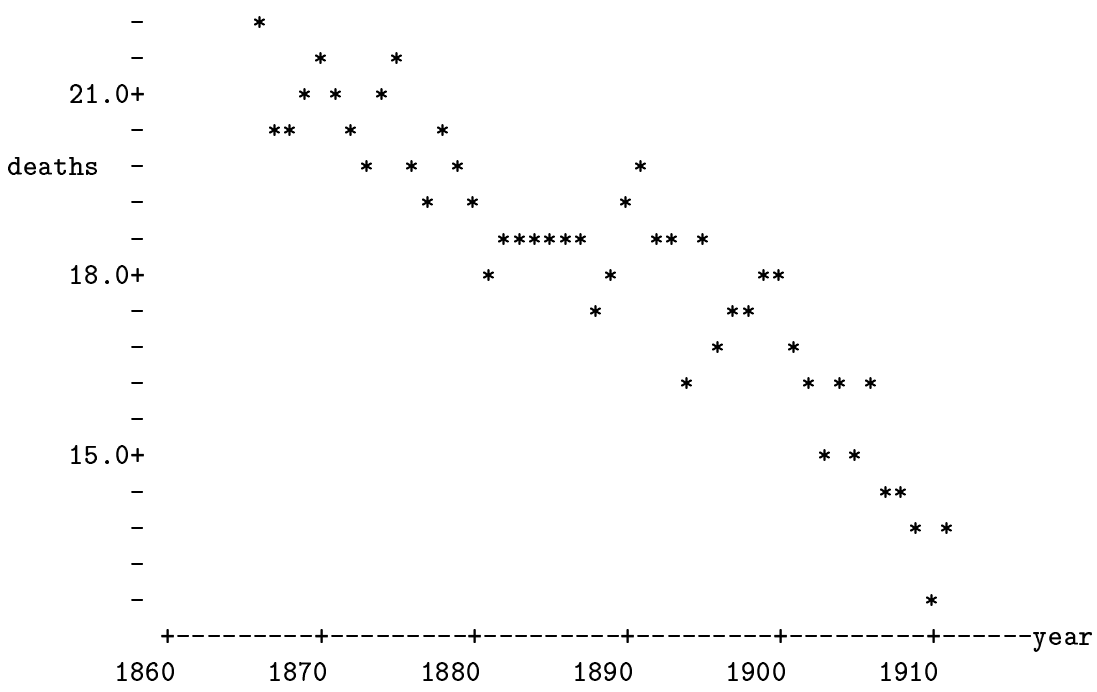
There is an association between height and weight, An increasing relationship. We predict a baseball player's weight in terms of his height. A confidence interval for the slope parameter is *(4.2, 7.2)*; hence, we predict the weight of a ball player to increase between 4 to 7 pounds for each additional increase in 1 inch of height, (4 to 7 pounds per inch). We are not saying taller causes heavier, this is absurd. But we are observing an association between height and weight. We are saying that if a ball player is taller then he is more likely to be heavier.

To make better predictions, there may be other variables to consider. In the height-weight data, a measure of body build would be useful. In a more advance class, we would discuss these issues.

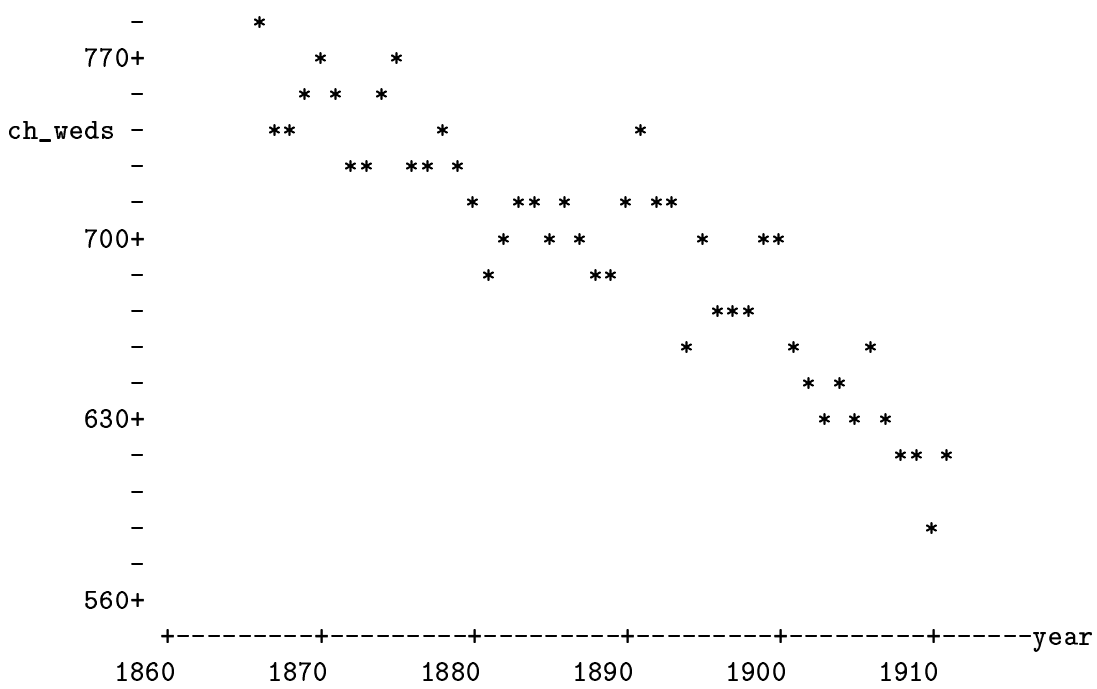
We do need to emphasize one thing concerning observational studies. **There must be a reason to explore associations and predictions.** An example here is worth thousands of words. Let Y be the number of deaths per 100,000 in England for a year in the late 1800's and let X be the number of church weddings (in thousands) in England for that year. There is no reason to seek an association between these variables. But suppose we do. The data is given in Appendix A. The scatter plot of the data is:



The relationship is linear. In fact the pattern is quite tight. It is clear from the plot: to reduce deaths, reduce church weddings! There is a variable here causing this pattern. It is time! These data are recorded over the years. Here is a plot of the death rate versus year:



Great strides in science were made in these years (Louis Pasteur, etc.) that helped the death rate to plummet. Here is a plot of church weddings versus year:



Church attendance dropped over these years. Hence both variables decrease with respect to year and thus have an increasing relationship when plotted with each other. So that solves the puzzle. Time is called a **lurking** variable here.

In an observational study, make sure you are including variables for which a relationship between them makes sense. If a paradox occurs (such as death rate and church wedding rate) look for a lurking variable.

Exercise 11.4.1

1. *(From Bhattacharyya and Johnson (1977), Statistical Concepts and Methods, New York: Wiley). Below are used-car prices (in thousands of dollars) for a foreign compact (1970's data) with their ages in years.*

Age	1	2	2	3	3	4	6	7	8	10
Price	2.45	1.80	2.00	2.00	1.70	1.20	1.15	.69	.60	.47

(a) *Plot the data, Price versus Age. Comment on the car buyer's lament (depreciation).*

(b) *Use the regression module to obtain the Wilcoxon fit of a linear model to the data.*

(c) *Obtain a 95% confidence interval for slope and interpret it in terms of the problem.*

(d) *Predict the price of an 11 year-old compact.*

(e) *What are some other X variables that would help predict price?*

(f) *If we had much older cars, would you expect to see a continual down hill trend? Why?*

2. *(From Hettmansperger and McKean (1998), Robust Nonparametric Statistical Methods, London: Arnold). Below are the number of telephone calls (tens of millions) made in Belgium for the years 1950-1973:*

Year	50	51	52	53	54	55	56	57	58	59	60	61
Calls	0.44	0.47	0.47	0.59	0.66	0.73	0.81	0.88	1.06	1.20	1.35	1.49

Year	62	63	64	65	66	67	68	69	70	71	72	73
Calls	1.61	2.12	11.90	12.40	14.20	15.90	18.20	21.20	4.30	2.40	2.70	2.90

(a) *Plot the data and comment on the plot (There were a few years where a recording error was made. Find those years).*

(b) *Use the regression module to obtain both the least squares and Wilcoxon fits of the data set.*

(c) *Plot these fits. Which would you use for prediction for the number of calls in 1974.*

11.5 How Regression Got Its Name

In the mid to late 1800's, a scientist called Galton was working with large observational studies on humans. One of these data sets consisted of the heights of fathers and first sons. In our terminology let

- X = the height of the father
- Y = the height of the first, fully grown, son.

Assume we want to predict Y in terms of X . When Galton plotted Y versus X the scatter filled in a large oval. The trend was linear and increasing. The least squares fit went through the center of the data $(\bar{X}, \bar{Y})$, as it always does, with positive slope. For this data $\bar{X}$ is about the same as $\bar{Y}$; i.e., the average heights were about the same over these two adjacent generations, (this was certainly true in the 1800's). This was true of the scale (standard deviations) also.

Suppose the slope of the least squares fit was 1. Then since the line goes through $(\bar{X}, \bar{Y})$ which are about the same then you would predict the height of the first son to be same as the height of the father. Galton noticed, though, that the slope of the line was definitely (significantly) less than 1. Hence for father's whose heights were taller than the average, the line predicts the son to be shorter than the father. Likewise, for father's whose heights were shorter than the average, the line predicts the son to be taller than the father. That is, taller fathers tend to have shorter sons and shorter fathers tend to have taller sons. There is a **regression towards the mean effect**. That's how regression got its name. Actually it is a good thing that this phenomenon occurs. Why?

Does regression towards the mean occur for other data sets? It does. Suppose we have observational data (X's and Y's both random). Suppose the data follow the linear model

- $Y = a + b*X + Error$,

where the errors are independent of the X's. Now suppose that **the variance of Y is the same as the variance of X.** (This is the key assumption; i.e., the variances are the same). Then we can show that the absolute value of b is less than 1. Hence if $b > 0$ then the model exhibits **regression towards the mean.**

Here's an example with real data. The data consist of the scores 36 students made on two tests in there statistics course. These were hour exams (over 20 questions). Test 1 was the first test and Test 2 was taken about a month later. So we want to predict Test 2 scores in terms of Test 1. Here's the data:

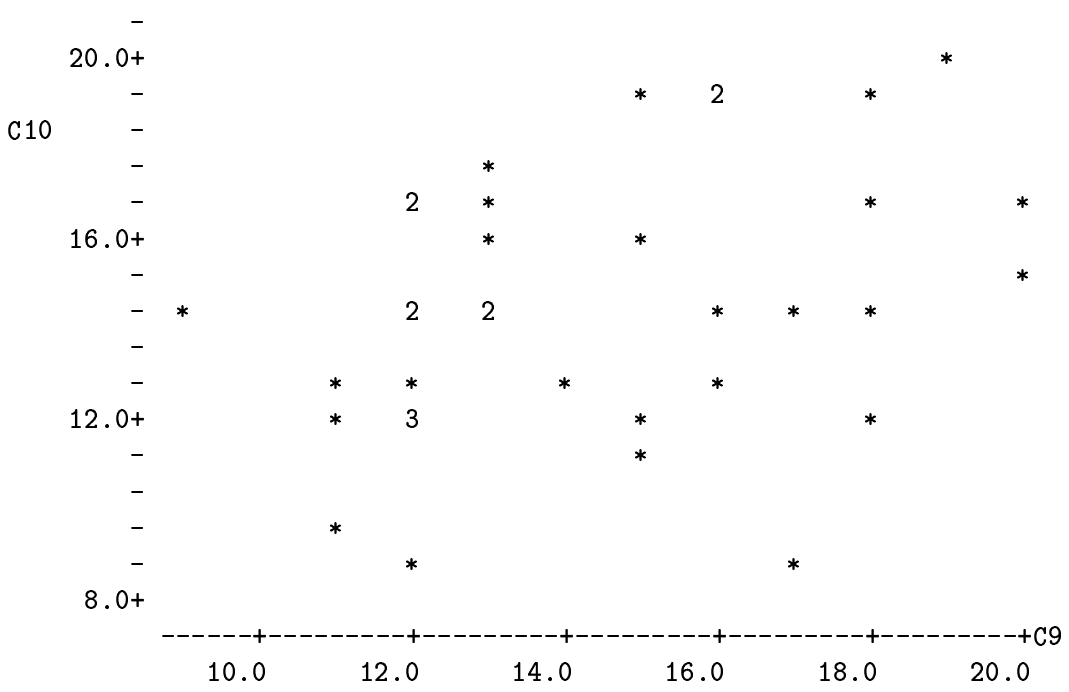
```
Test 1
```

```
    12    17    16    18    12    12    12    20    18    18    11    13    15
    16    20    13    15    11     9    12    17    12    16    15    19    13
    13    16    18    12    12    15    11    14    12    13

Test 2 (these data are paired (same order) with Test 1

    14    14    19    17    12    14    13    17    14    19    12    16    16
    19    15    14    11    13    14    17     9    12    13    12    20    18
    17    14    12     9    12    19    10    13    17    14
```

As I noted this is paired data. The first student scored 12 on his first test and 14 on his second test. Here is a scatter plot of the data:



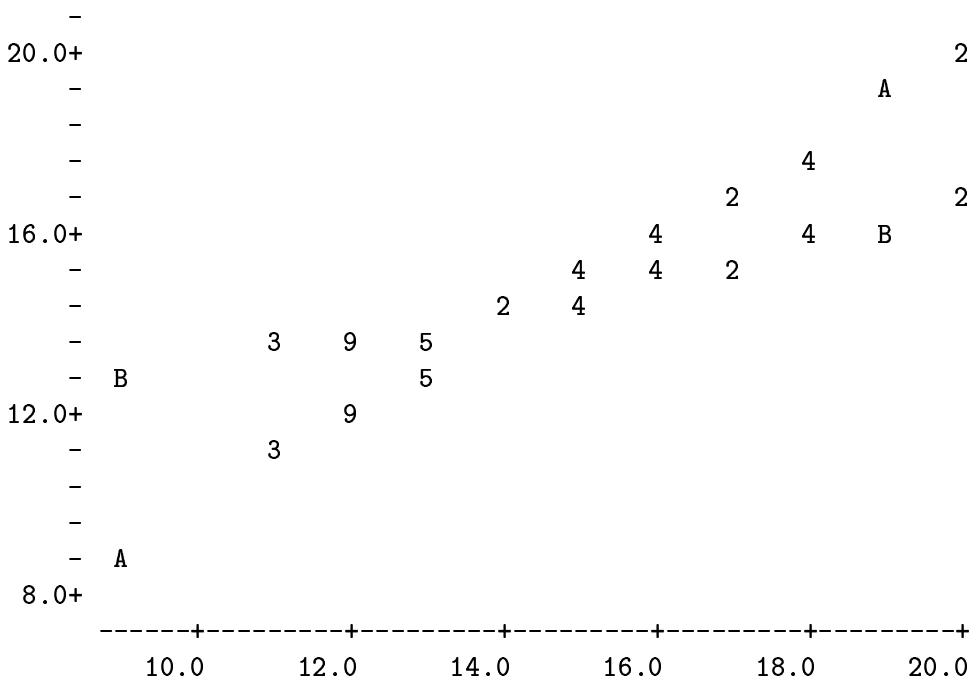
The averages are 14.4 and 14.5 for Tests 1 and 2, respectively. The standard deviations are 2.86 and 2.93 for Tests 1 and 2, respectively. The least squares fit is

- $\hat{\text{Test}2} = 9.4 + .35 * Test1$

The slope is less than 1 which is not surprising since the standard deviations are about the same. Hence this data set exhibits regression towards the mean.

You can see it in the data too. Note that two students scored 20 on the first test. They scored less than 20 on the second test. Note the four students who scored 18 on the first test. Three of these scored less than 18 on the second test while 1 scored higher. Likewise, notice the 5 students who scored 13 on Test 1. They all scored higher on the second test.

As a final thought on regression towards the mean, the plot below shows the least squares fit contrasted with the line through points where the second coordinate is the same as the first coordinate (i.e., scores on second test exactly the same as on first).



Finally to contrast the regression towards the mean effect to The least squares estimate of slope is .35 which is much less than 1. Sketch the fit (show it goes through $(14.4, 14.5)$).

Exercise 11.5.1 *1. Use the regression module to obtain the Wilcoxon fit for the above data.*

2. The scores below are test scores for students in a Stat class over two tests, Test 1 and Test2.

```
Test 1
  37  17  23  40  37  39  35  29  32  40  26  39  34
  29  38  21  36  38  14  27  34  38  25  18  39  37
  36  12  34  26

Test 2  (Paired data, first student scored 37, 28 on tests 1 and 2 respectively)

  28  24  20  32  39  36  40  33  23  36  21  30  30
  21  22  24  27  20   8  31  28  30  25  16  31  18
  25   6  36  20
```

(a) Plot the data, Test 2 versus Test 1.

(b) Use the summary module to find the standard deviations of the two data sets. Do you think these data will exhibit the regression towards the mean effect?

(c) Use regression module to obtain the Wilcoxon fit. Do the data exhibit the regression towards the mean effect?

Appendix A

Data Sets

A.1 Carrie's Baseball Data

These data come from the back-side of 59 baseball cards that Carrie had. There are 6 columns: c1 contains the heights; c2 contains the weights; c3 is 1 if the player hits from the right side, 2 if from the left side and 3 if the player is a switch hitter; c4 is 0 if the player throws right-handed and it is 1 if left-handed; c5 is 0 if the player is a pitcher and 1 if he is a fielder; c6 is the ERA if the player is a pitcher and his batting average if the player is a fielder.

```
74 218 1 1 0 3.330
75 185 1 0 1 0.286
77 219 2 1 0 3.040
73 185 1 0 1 0.271
69 160 3 0 1 0.242
73 222 1 0 0 3.920
78 225 1 0 0 3.460
76 205 1 0 0 3.420
77 230 2 0 1 0.303
78 225 1 0 0 3.460
76 190 1 0 0 3.750
72 180 3 0 1 0.236
73 185 1 0 1 0.245
73 200 2 1 0 4.800
74 195 1 0 1 0.276
75 195 1 0 0 3.660
72 185 2 1 1 0.300
75 190 1 0 1 0.239
76 200 1 0 0 3.380
```

```
76 180 2 1 0 3.290
72 175 2 1 1 0.290
76 195 2 1 0 4.990
68 175 2 0 1 0.283
73 185 1 0 1 0.271
69 160 1 0 1 0.225
76 211 3 0 1 0.282
77 190 3 0 1 0.212
74 195 1 0 1 0.262
75 200 1 0 0 3.940
73 207 3 0 1 0.251
79 232 2 1 0 3.100
72 190 1 0 1 0.238
75 200 2 0 0 3.180
70 175 2 0 1 0.279
75 200 1 0 1 0.274
78 220 1 0 0 3.880
73 195 1 0 0 4.570
75 205 2 1 1 0.284
74 185 1 0 1 0.286
71 185 3 0 1 0.218
73 210 1 0 1 0.282
76 210 2 1 0 3.280
73 195 1 0 1 0.243
75 205 1 0 0 3.700
73 175 1 1 0 4.650
73 190 2 1 1 0.238
74 185 3 1 0 4.070
72 190 3 0 1 0.254
73 210 1 0 0 3.290
71 195 1 0 1 0.244
71 166 1 0 1 0.274
71 185 1 1 0 3.730
73 160 1 0 0 4.760
74 170 2 1 1 0.271
76 185 1 0 0 2.840
71 155 3 0 1 0.251
76 190 1 0 0 3.280
71 160 3 0 1 0.270
70 155 3 0 1 0.261
```

A.2 Etruscan-Italian Data

These are maximal head measurements (across the top of the skull) in mm of 84 ancient Etruscans and 70 modern Italians.

Head Sizes of Etruscans

```
141 148 132 138 154 142 150 146 155 158 150 140 147 148 144 150
149 145 149 158 143 141 144 144 126 140 144 142 141 140 145 135
147 146 141 136 140 146 142 137 148 154 137 139 143 140 131 143
141 149 148 135 148 152 143 144 141 143 147 146 150 132 142 142
143 153 149 146 149 138 142 149 142 137 134 144 146 147 140 142
140 137 152 145
```

Head Sizes of Italians

```
133 138 130 138 134 127 128 138 136 131 126 120 124 132 132 125
139 127 133 136 121 131 125 130 129 125 136 131 132 127 129 132
116 134 125 128 139 132 130 132 128  139 135 133 128 130 130 143
144 137 140 136 135 126 139 131 133 138 133 137 140 130 137 134
130 148 135 138 135 138
```

A.3 Mortality-Church Wedding Data

Year	Deaths	Church weddings
1866	22.0	780
1867	20.5	745
1868	20.4	743
1869	20.8	755
1870	21.5	770
1871	21.3	762
1872	20.2	732
1873	20.0	728
1874	21.0	755
1875	21.5	770
1876	19.8	730
1877	19.4	721
1878	20.4	744
1879	19.8	730
1880	19.5	720
1881	18.0	692
1882	18.5	705
1883	18.8	710
1884	18.8	710
1885	18.6	702
1886	18.6	708
1887	18.4	700
1888	17.5	682
1889	17.8	690
1890	19.2	720
1891	20.0	735
1892	18.5	708
1893	18.8	712
1894	16.2	658
1895	18.4	704
1896	16.6	670
1897	17.2	675
1898	17.4	678
1899	18.2	694
1900	18.2	694
1901	16.8	664

1902	16.2	650
1903	15.2	630
1904	16.2	650
1905	15.0	628
1906	16.4	652
1907	14.5	624
1908	14.2	614
1909	14.0	610
1910	12.8	582
1911	14.0	610

Appendix B

Table of 10-Digit Random Numbers

5965 2913 5612 6361 7075 5490 9626 4307 0840 7945
5801 9383 6173 8358 9236 5543 5811 5520 5814 7864
1223 5344 3649 6397 1678 4400 7715 7614 1209 7729
0220 2108 0784 8837 3916 0282 4490 3442 6471 6593
4131 9772 7594 8863 0874 1864 8117 6411 7012 2682
3074 5746 2723 5681 0989 8015 0818 5380 9981 3758
2939 6585 6658 7756 7916 9770 2868 2128 2665 2386
6003 5982 8829 2833 8160 2101 3365 4121 4522 8216
2039 2993 4362 6363 2914 4955 6364 5237 6456 5561
0176 2425 2968 3834 6077 4302 3499 9938 7231 2136
2161 1365 2764 7836 1584 2421 4247 2930 0783 9989
0407 1760 7048 1929 9034 0242 0753 4851 9465 0791
0055 7981 7760 2215 3323 4727 8884 8066 7965 3939
0726 2104 9164 6275 5464 4073 1715 3215 7883 8087
2475 9583 8713 1445 2702 4952 4307 5796 2913 0589
0686 1266 4341 9760 9608 5773 7394 9333 4752 8395
4223 4033 3734 8221 2055 5131 0065 1626 7742 5806
9596 5241 3230 3269 4836 9776 2894 5740 1557 2515
1581 5007 6906 8933 9981 3175 4979 4525 5334 6038
6558 6350 1273 6164 7125 1481 3084 1517 4748 0956
1974 7635 1129 0593 7963 3817 0148 1377 5165 6568
8671 4147 7231 3509 9032 4233 9087 3328 9044 3152
0979 6984 8428 7697 8859 5363 2984 2649 9244 7035
0635 0334 7219 7422 9571 1053 5954 4040 5777 2440
6686 8703 3451 1548 9797 0816 9342 0240 5814 9593
3878 6600 8703 9512 5588 2446 1842 0882 2024 7736

9869 8361 8090 8666 7540 6516 3343 7379 1140 5565
8969 4225 6202 8102 5691 8499 6466 7775 0721 9345
6339 8671 8023 3701 8250 0274 9339 5135 4475 7960
3187 5353 9213 1705 5580 1432 5962 8191 1676 5861
6142 5175 6497 9478 6278 8939 3902 0076 2004 9201
8286 5570 4400 3640 9650 5709 6855 3454 5397 9991
5531 0150 6376 0494 8239 1639 5611 5803 5645 0851
6357 6828 4497 2508 9084 7544 5964 3718 1007 9333
7376 2940 3503 3317 0465 2912 6500 3883 2539 6516
3060 1836 3740 7183 2965 3246 4028 5528 8607 5611
4767 1322 7035 6171 9065 2024 2318 5460 5571 2092
1550 2362 4356 9447 4196 1101 6479 3928 3321 3684
4956 5537 9056 3006 2066 7296 3018 3878 2927 9268
2504 8074 7591 9689 2755 3226 1726 9222 3633 9816
8328 3942 7243 1717 3592 9307 2738 3856 0684 9873
6227 3172 3764 9551 0426 6061 8384 5473 7418 8053
2946 2893 4927 2197 3452 6104 2255 2268 7063 1443
7574 3933 8021 2711 6276 7146 2391 1984 2962 3634
9042 6919 4140 4545 6873 3748 5053 8284 4120 1819
1839 7794 6640 0492 6833 0485 6422 5213 0394 2643
4861 2514 5827 7994 4041 9929 8055 3514 7126 4064
6051 9425 6381 7204 3938 3430 5952 2753 3471 5992
5306 1578 1198 6256 1865 5631 2852 1416 6313 4460
2521 8837 4158 5485 7726 4380 7901 6142 6385 6755

Appendix C

Notation and Abbreviations

- n denotes the size of a sample.
- Q_1 denotes the sample first quartile.
- Q_2 denotes the sample median (or the second quartile).
- Q_3 denotes the sample third quartile.
- **IQR** denotes the sample interquartile range.
- $\bar{x}$ denotes the sample mean.
- s denotes the sample standard deviation.
- s^2 denotes the sample variance.
- μ denotes the population mean.
- σ denotes the population standard deviation.
- σ^2 denotes the population variance.
- q_1 denotes the first population quartile.
- q_3 denotes the third population quartile.
- iqr denotes the population interquartile range.
- θ denotes the population median.
- A^c denotes the complement of the set A.

- H_0 denotes the null hypothesis.
- H_A denotes the alternative hypothesis.
- Δ denotes the true difference in shift (effect). Estimate denoted by $\hat{\Delta}$.
- p denotes the population proportion. Estimate denoted by $\hat{p}$.
- $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2.
- $bin(n, p)$ denotes a binomial distribution with parameters n and p.
- T denotes the Wilcoxon test statistic.

- **CC** abbreviates "Class Code".
- **LIF** abbreviates "Lower Inner Fence".
- **UIF** abbreviates "Upper Inner Fence".
- **HL** abbreviates "Hodges-Lehmann".
- **LS** abbreviates "Least Squares".
- **CI** abbreviates "Confidence Interval".
- **SRW** abbreviates "Signed-Rank Wilcoxon".
- **CRD** abbreviates "Completely Randomized Design".

Appendix D

Practice Final Examination

Attempt all problems.

1. Suppose the population of incomes of people working in industry and who have a masters degree in Statistics is positively skewed with mean of \$55 (in thousands of dollars) and a standard deviation of 3. Suppose we take a sample of size 100 from this population and form the arithmetic average $\overline{X}$. If we did this repeatedly, what would be the shape of the histograms of $\overline{X}$'s and in what interval would the middle 68% of $\overline{X}$'s lie?

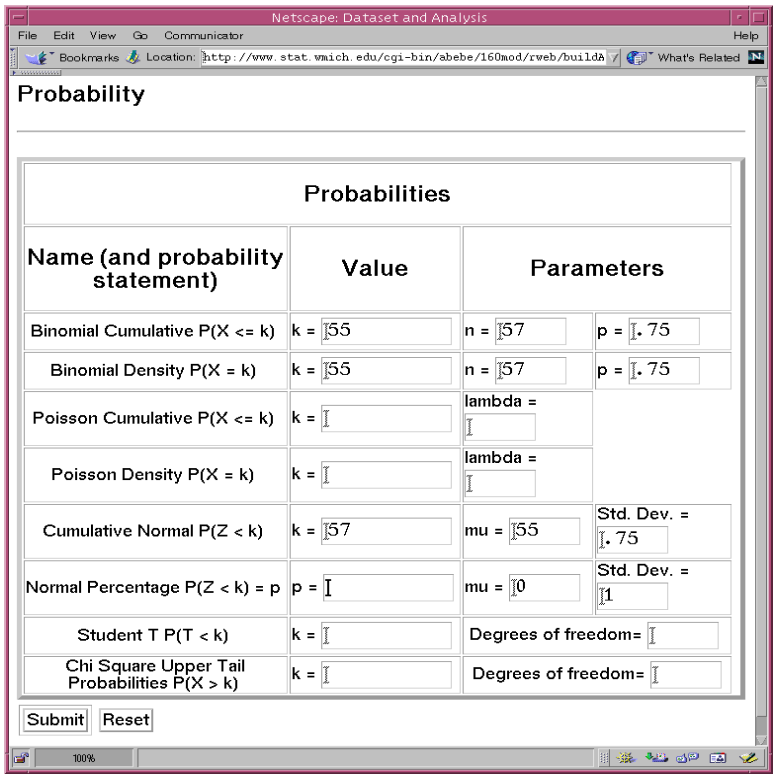
   (a) Positively skewed and $(52, 57)$.

   (b) Positively skewed and $(54.7, 55.3)$.

   (c) Mound shaped and $(54.7, 55.3)$.

   (d) Mound shaped and $(52, 57)$.

2. In the last problem, from the print out below find the probability that the average income of 16 such people exceeds 57.




```
Rweb:> # CUMULATIVE BINOMIAL DISTRIBUTION
Rweb:> pbinom(55, 57, .75)
[1] 0.9999985
Rweb:> # BINOMIAL PROBABILITY
Rweb:> dbinom(55, 57, .75)
[1] 1.340548e-05
Rweb:> # CUMULATIVE NORMAL DISTRIBUTION
Rweb:> pnorm(57, 55, .75)
[1] 0.9961696
```

(a) .9962

(b) .0038

(c) .9999

(d) .0001

3. To be accepted into Stanford's Graduate Business School, a candidate must pass a four hour written entrance exam, the SBSE. Ajax Prep Company offers an expensive study course to prepare a person for this exam. Ajax makes the claim that 80% of the students who finish their course pass the SBSE. Pam works for a government agency that thinks Ajax's claim is dubious and that their true percentage of alumni who pass the SBSE is somewhat lower than 80%.

   So Pam collects a random sample of 64 students who finished Ajax's course and determines that 45 of them passed the SBSE. Based on this information Pam forms a 95% confidence interval for the true percentage and makes a decision. What was Pam's interval and what was her decision?

   (a) $(.65, .76)$, Ajax's claim is false!

   (b) $(.59, .81)$, Ajax's claim is false!

   (c) $(.59, .81)$, no evidence against Ajax.

   (d) $(.65, .76)$, no evidence against Ajax.

4. Jane works for Dick's Real Estate Agency, Spot Reality. Dick wants to determine the median owner's asking price in an exclusive neighborhood. So Jane obtains the following random sample of owner's asking prices: (In thousands of dollars):

```
580 552 928 757  84 394 528 373 859 460 258 998
```

   What is Jane's estimate?

   (a) 461.

   (b) 564.25.

   (c) 540.

   (d) 277.50.

5. In the last problem, Jane was not satisfied with just an estimate, so she used resampling code to obtain the 100 resampled medians:

```
373.0 383.5 383.5 383.5 383.5 383.5 394.0 394.0 427.0 427.0
427.0 427.0 427.0 460.0 460.0 460.0 460.0 460.0 460.0 461.0
473.0 494.0 494.0 494.0 506.0 506.0 506.0 506.0 506.0 506.0
506.0 506.0 520.0 520.0 528.0 528.0 528.0 540.0 540.0 540.0
540.0 540.0 540.0 540.0 540.0 552.0 552.0 552.0 552.0 552.0
552.0 552.0 552.0 552.0 552.0 552.0 554.0 554.0 554.0 554.0
554.0 554.0 554.0 554.0 566.0 566.0 566.0 566.0 566.0 580.0
580.0 580.0 580.0 580.0 580.0 580.0 642.5 654.5 654.5 668.5
668.5 668.5 668.5 668.5 668.5 693.5 705.5 719.5 719.5 719.5
754.0 757.0 757.0 808.0 808.0 808.0 808.0 859.0 859.0 859.0
```

From this she obtained a 95% confidence interval. What was Jane's 95% confidence interval and what does it mean?

(a) $(383.5, 859.0)$, Jane is fairly confident that this interval contains the true median owner's asking price.

(b) $(383.5, 859.0)$, Jane is fairly confident that this interval contains the true range of the owner's asking price.

(c) $(496, 632)$, Jane is fairly confident that this interval contains the true median owner's asking price.

(d) $(496, 632)$, Jane is fairly confident that this interval contains the true range of the owner's asking price.

6. Consider the last two problems. Suppose Jane took a larger random sample say of size 36 and use it to obtain a new estimate of the true median and a new 95% confidence interval. What is true, in general, about the length of the new confidence interval?

   (a) The new interval would have about the same length as the old interval.

   (b) The new interval would have a shorter length than the old interval.

   (c) Can't say because its another sample.

   (d) Since the new sample size is larger the new confidence interval would also be larger.

7. Four pea plants of a certain variety are grown without fertilizer, while five of the same variety are grown with fertilizer. Other than the presence or absence of fertilizer the plants received the same treatment. Let Δ be the true mean (or median) increase in plant height due to fertilizer. We want to test the hypotheses

$$H_0 : \ \Delta = 0 \text{ versus } H_A : \ \Delta > 0 \ .$$

   The experiment resulted in the following data:

```
                                Height (in.)
Without Fertilizer(Control):    19  8 16 17
With    Fertilizer(Treated):    20 13 25 18 15
```

Determine the value of the Wilcoxon test statistic for this data and determine what we would expect it to be if H_0 is true.

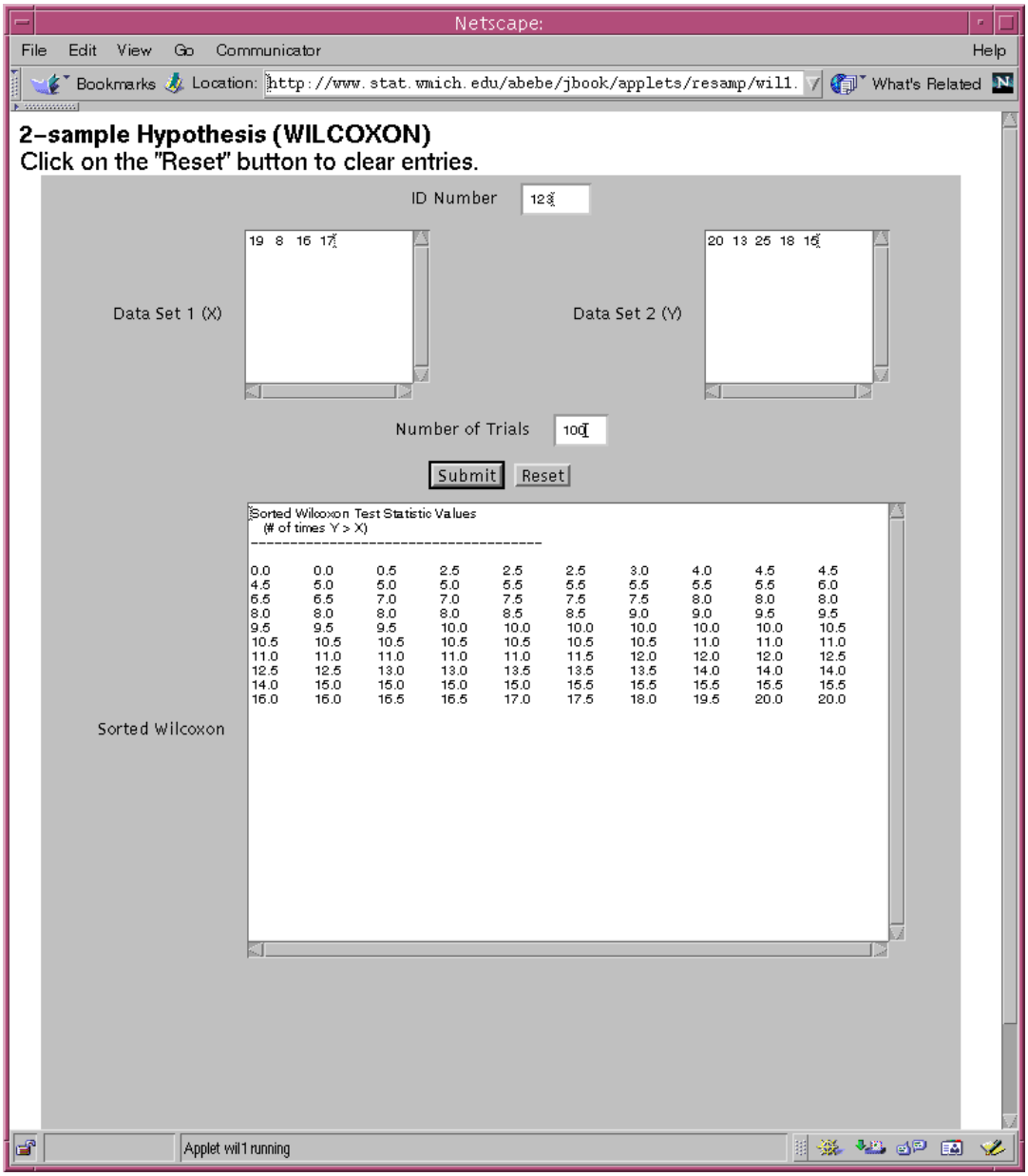
(a) 13 (expect it to be 0).

(b) 3.2 (expect it to be 0).

(c) 13 (expect it to be 10).

(d) 3.2 (expect it to be 10).

8. The data were combined into one big sample which was resampled 100 times. In each resampling, 4 were allocated to be new control items and 5 were allocated to be new treated items. For each resampling, the Wilcoxon test statistic was obtained and is given below. Obtain the observed significance level and make the proper decision if your maximum Type I error is at most 5%.


Netscape:
File Edit View Go Communicator Help
Bookmarks Location: http://www.stat.wmich.edu/abebe/jbook/applets/resamp/wil1.
What's Related
2-sample Hypothesis (WILCOXON)
Click on the "Reset" button to clear entries.
ID Number
123
Data Set 1 (X)
19 8 16 17
Data Set 2 (Y)
20 13 25 18 15
Number of Trials
100
Submit
Reset
Sorted Wilcoxon
Sorted Wilcoxon Test Statistic Values
(# of times Y > X)
0.0 0.0 0.5 2.5 2.5 2.5 3.0 4.0 4.5 4.5
4.5 5.0 5.0 5.0 5.5 5.5 5.5 5.5 5.5 6.0
6.5 6.5 7.0 7.0 7.5 7.5 7.5 8.0 8.0 8.0
8.0 8.0 8.0 8.0 8.5 8.5 9.0 9.0 9.5 9.5
9.5 9.5 9.5 10.0 10.0 10.0 10.0 10.0 10.0 10.5
10.5 10.5 10.5 10.5 10.5 10.5 10.5 11.0 11.0 11.0
11.0 11.0 11.0 11.0 11.0 11.5 12.0 12.0 12.0 12.5
12.5 12.5 13.0 13.0 13.5 13.5 13.5 14.0 14.0 14.0
14.0 15.0 15.0 15.0 15.0 15.5 15.5 15.5 15.5 15.5
16.0 16.0 16.5 16.5 17.0 17.5 18.0 19.5 20.0 20.0
Applet wil1 running

(a) .28, conclude that typical fertilized peas are taller than unfertilized peas.

(b) .28, no evidence to conclude that typical fertilized peas are taller than unfertilized peas.

(c) .56, no evidence to conclude that typical fertilized peas are taller than unfertilized peas.

(d) .56, conclude that typical fertilized peas are taller than unfertilized peas.

9. Using the data of Problem 7, Obtain the Wilcoxon estimate of Δ.

   (a) 1.5.

   (b) 3.2.

   (c) 13.

   (d) 2.5.

10. Besides an estimate of the effect Δ, suppose we also want a confidence interval for Δ. Which resampling plan below would we use.

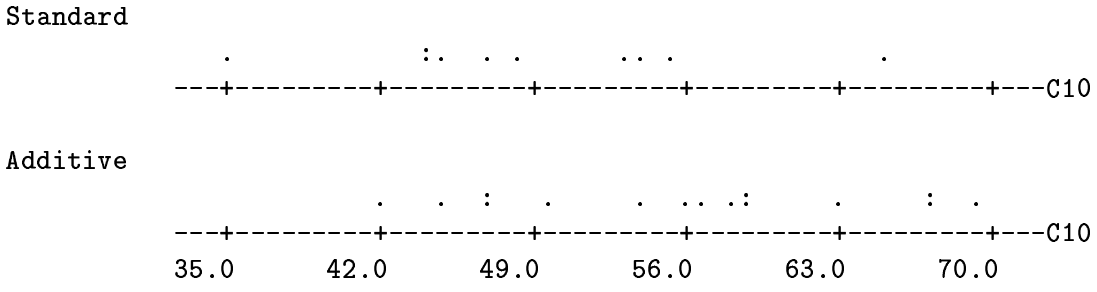
    (a) Combine the original samples into one sample and then resample with replacement from the big sample allocating items to new samples.

    (b) Resample from each sample without replacement.

    (c) Resample from each sample with replacement.

    (d) Combine the original samples into one sample and then resample without replacement from the big sample allocating items to new samples.

11. 25 cars were put on test. The first 10 used a standard fuel while the others used a fuel designed (hopefully) to increase miles per gallon. The same amount of fuel was used in each car. Below are the comparison dotplots of the cars' miles per gallon.



What else if anything needs to be done to "correctly" infer about the new additive?

(a) Obtain a point estimate and confidence interval for the effect (difference in means or medians).

(b) It is clear from the boxplots that the new fuel additive is effective, so no further statistics are needed.

(c) It is clear from the boxplots that the new fuel additive gives similar miles per gallon as the standard, so no further statistics are needed.

(d) Obtain a point estimate for the effect (difference in means or medians).

12. A new type of surgery for a certain heart disease has been developed. In order to test it, 50 patients who have the disease were selected. Half of them got the new surgery while the others received the standard surgery. After the surgery, each patient's surgery was rated a success, a failure or no change by a team of doctors who did not know what surgery the patient had received.

    Suppose we decide to rate the surgeries on their success rates. Let p_N be the number of successful surgeries for the new operation and let p_S be the number of successful surgeries for the standard operation. Based on the data below estimate $p_N - p_S$.

	Success	Failure	No Change
New Surgery	16	7	2
Standard Surgery	10	10	5

(a) .64

(b) .24

(c) .40

(d) .195

13. For the last problem, suppose we want to test

$$H_0 : \ p_N = p_S \text{ versus } H_A : \ p_N \neq p_S \ .$$

    Using 2000 resamples, we obtain the 95% confidence interval $(-.03, .50)$. Which of the following statements is the proper conclusion for testing H_0 versus H_A.

(a) The sample sizes are far too small to conclude anything.

(b) There is sufficient evidence at the .05 level to conclude that the new surgery is better than the standard.

(c) There is insufficient evidence at the .05 level to conclude that the new surgery is better than the standard.

(d) It is clear from the data that the new surgery is better than the standard, so confidence intervals are not needed.

14. The following data are the monthly rental prices for a random sample of 10 unfurnished studio apartments in the center of a large city.

    `955, 1000, 985, 980, 940, 975, 965, 999, 1247, 1119`

    List the 5-number summary (min, Q1, median, Q3, max).

    (a) 940, 965, 982.5, 1000, 1247

    (b) 955, 985, 960, 999, 1119

    (c) 940, 955, 982.5, 1119, 1247

    (d) 955, 985, 957.5, 999, 1119

15. In order to estimate how much water will be needed to supply the community of Falling Rock in the next decade, the town council asked the city manager to find out how much water typical family uses. A random sample of 15 Falling Rock families used the following amount of water (in thousands of gallons) in the past year.

    `4.1, 13.1, 14.0, 14.6, 15.5, 16.4, 16.9, 18.2, 18.3,`
    `18.8, 19.7, 21.5, 22.7, 23.8, 32.2`

    Identify the outliers (if any) in this data set.

    (a) 4.1, 32.2

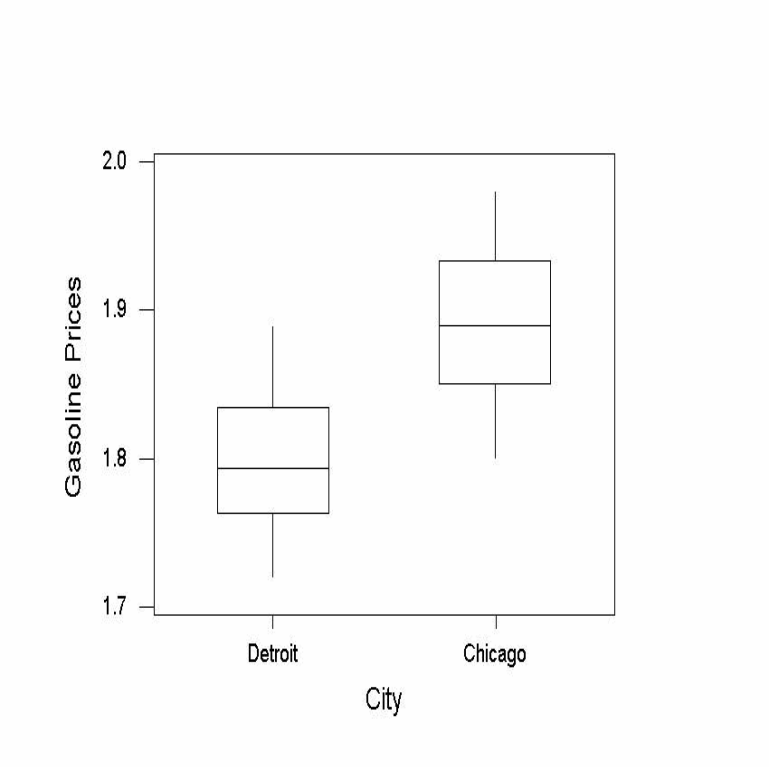
    (b) 4.1, 13.1, 23.8, 32.2

    (c) There aren't any outliers

    (d) 18.2

    The next two questions refer to the following situation:

    The following side-by-side boxplots represent the prices of gasoline (per gallon) based on a random sample of gasoline stations in Detroit and Chicago.



16. What can be said about the scale (variation) between the Detroit and Chicago gasoline prices?

    (a) Chicago has less variation
    (b) Chicago has more variation
    (c) Detroit has more variation
    (d) Approximately Equal

17. What can be said about the shift between Detroit and Chicago gasoline prices?

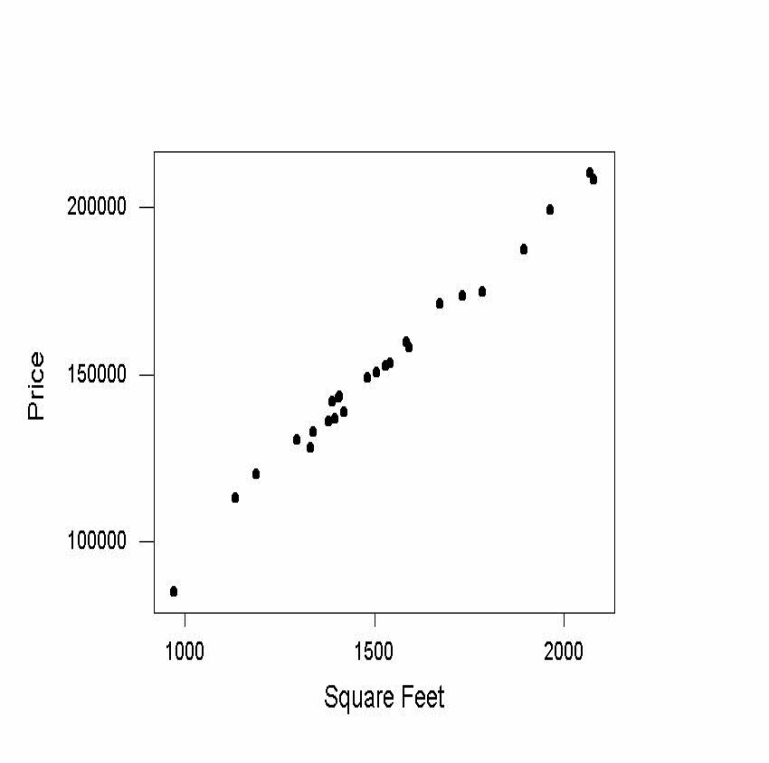
    (a) Detroit has higher prices
    (b) The prices are approximately equal

(c) Chicago has higher prices

(d) Since the variances are different, it is impossible to tell.

The next two questions refer to the following situation:

An agent for a residential real estate company would like to predict selling prices for homes based upon the size (amount of square footage). A sample of 25 homes in a particular neighborhood was selected. A regression analysis revealed the following scatterplot and regression equation.



The regression equation is

```
Price = - 7598 + 105 Square Feet
```

18. If a home has 1800 square feet of living area, what selling price would this regression model predict?

    (a) $196,698

    (b) Unable to predict, we would be extrapolating

    (c) $181,402

    (d) $189,000

19. According to this model, for every additional square foot of living area, by how much will the price of a home change?

    (a) Decrease $7598

    (b) Increase $7598

    (c) Decrease $105

    (d) Increase $105

20. Your job is to assemble computers. A local company sends you fuses used in the construction of the computers. Your company estimates that 20% of these fuses are defective. You have just received a shipment of 100 fuses from the local company (80 fuses good, 20 fuses defective). You pick 3 fuses at random from this shipment. If your job is to assemble 3 computers, what is the probability you will have 0 defective fuses in your 3 computers. (Hint: Use a tree diagram to calculate)

    (a) .0071

    (b) .5081

    (c) .4919

    (d) .8

21. A survey of 100 people was taken. The question was: "Please check the appropriate response regarding if you have used the following products over the past month:" The answers from these 100 people are as follows:

Event	Response
Taken Tylenol	60
Taken Pepto-Bismol	25
Taken Both Tylenol and Pepto-Bismol	15

```
Taken Neither Tylenol nor Pepto-Bismol  30

Total                                  100
```

Are the events "Taken Tylenol" and "Taken Pepto-Bismol" independent events?

(a) Yes

(b) No

(c) More information is required

(d) Only if this is a random sample of 100 people

The next two questions refer to the following situation:

I wish to estimate the probability of getting a three or more of a kind "3-ones, 3-twos, 3-threes, 3-fours, 3-fives, or 3-sixes" on the first roll of 5 fair dice (Like in the game Yahtzee). I perform 15 resampling trials with the following results:

```
Trial 1
5 6 2 5 2

Trial 2
4 3 6 6 5

Trial 3
5 6 5 5 1

Trial 4
2 1 4 6 1

Trial 5
5 5 3 2 4

Trial 6
3 4 6 1 4

Trial 7
6 4 1 6 4

Trial 8
1 5 5 6 6
```

```
Trial 9
5 3 1 6 1

Trial 10
6 6 6 3 6

Trial 11
4 3 3 2 3

Trial 12
3 5 2 6 4

Trial 13
6 6 3 6 1

Trial 14
3 2 6 5 6

Trial 15
6 1 2 5 6
```

22. What is the estimate on the probability of getting three or more of a kind?

    (a) .733

    (b) 0

    (c) .267

    (d) .6

23. What is the error of estimation for this probability?

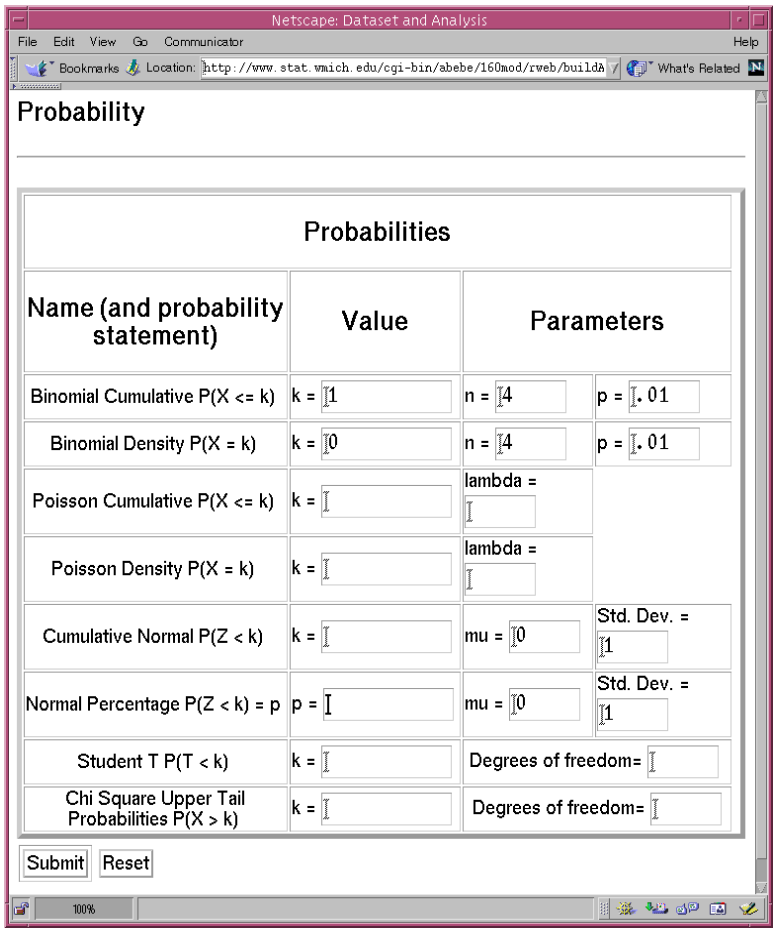
    (a) 0

    (b) .2668

    (c) .2529

    (d) .2285

24. Consider a metabolic defect that occurs in one of every 100 births. If four infants are born in a particular hospital on a given day, what is the probability that at least one has the defect? Use the following output from the probability module.




```
Rweb:> # CUMULATIVE BINOMIAL DISTRIBUTION
Rweb:> pbinom(1, 4, .01)
[1] 0.999408
Rweb:> # BINOMIAL PROBABILITY
Rweb:> dbinom(0, 4, .01)
[1] 0.960596
```

(a) 0.96

(b) 0.04

(c) 0.999

(d) 0.001

25. Suppose that buses arrive at a bus stop every 15 minutes and that the waiting time for the next bus to arrive has a uniform distribution on the interval from 0 to 15 minutes. Find the probability that a person's waiting time will exceed 10 minutes.

(a) 5/15

(b) 4/15

(c) 1/10

(d) 1/2

26. If X has a normal distribution with mean 30 and standard deviation 5, which of the following has the greatest probability?

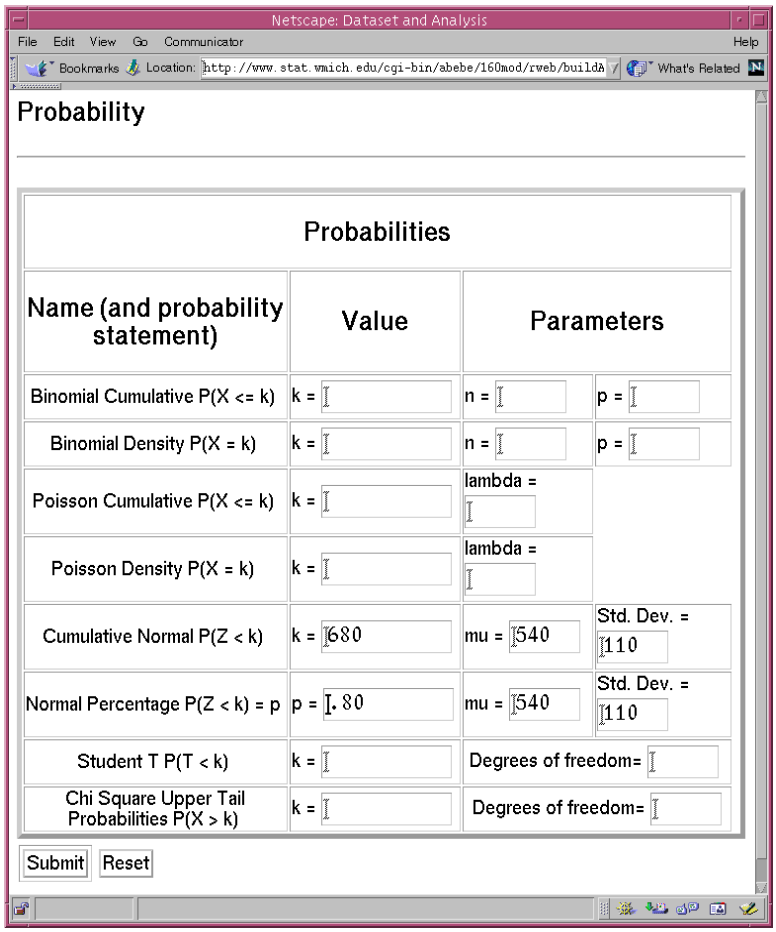
(a) $X < 35$

(b) $X > 30$

(c) $X > 20$

(d) $X < 37.5$

27. The scores of a national achievement test were approximately normally distributed with a mean of 540 and a standard deviation of 110. If you achieve a score of 680, what percentage of those who took the examination score lower than you? Use the following probability module output.




```
Rweb:> # CUMULATIVE NORMAL DISTRIBUTION
Rweb:> pnorm(680, 540, 110)
[1] 0.8984426
Rweb:> # NORMAL PERCENTAGE POINT
Rweb:> qnorm(.80, 540, 110)
[1] 632.5783
```

(a) 0.800

(b) 0.102

(c) 0.898

(d) 0.975

28. Refer to the situation and output given in the previous question. **C** College admits students whose score on the test is among the top 20%. What is the lowest score that would guarantee admission?

(a) 680.0

(b) 632.6

(c) 540.0

(d) 650.0

29. A study was performed to investigate the relationship between the carburetor jetting size and the time of a Camaro for a quarter-mile run. The data are:

Jet Size	76	68	70	72	74	76
Time	15.08	14.60	14.50	14.53	14.79	15.02

The Wilcoxon analysis output (from the regression module) is given below :

	Coef	Std. Err	t-ratio
intercept	9.4675300	2.3165700	4.08687
Jet	0.0724995	0.0318255	2.27803

Use the results above to predict the time for a jet size of 76. What is the predicted time?

(a) 15.05

(b) 15.02

(c) 15.08

(d) 14.98

30. Consider the situation in the previous question. Use the Wilcoxon fit to obtain a 95% confidence interval for the slope parameter and use it to test:

- H_0 : Slope is 0

- H_1 : Slope is not 0.

The interval and conclusion are:

(a) (0.01 , 0.14) ; Slope is not 0.

(b) (0.04 , 0.10) ; Slope is not 0.

(c) (-1.89 , 2.03) ; Slope is 0.

(d) (0.04 , 0.10) ; Slope is 0.

31. To decide whether a newly developed gasoline additive increases gas mileage you will compare the gas mileage for cars with and without the additive. A recent study randomly selected a single group of 5 cars and had each of the 5 cars driven both with and without the additive.

```
With(Y)          :       25.7    20.0    28.4    13.5    18.4
Without(X)       :       24.9    18.8    27.7    13.0    18.8
Diff(Y-X)        :
```

What is the value of the signed rank Wilcoxon *test statistic*?

(a) 14

(b) 9

(c) 0.65

(d) 15

32. Consider the data and context of the above question. What is the value of the centered signed rank Wilcoxon test statistic, i.e. the expected under H_0?

(a) 0.65

(b) 7.5

(c) 14

(d) 10

33. Consider the data and context of the previous 2 questions. Suppose our interest is in determining whether there is a difference between the two additives. So, the differences were computed and put in the class code for paired Wilcoxon. The class code reported a 95% confidence interval for the difference. The interval is (0.05 , 1.0). What conclusion do you draw based on the interval?

(a) There is a difference.

(b) There is no difference.

(c) Inconclusive.

(d) Not enough information.

34. A group of 9 students were randomly assigned to be taught by two different teaching techniques. They were tested at the end of a specified period of time. The following are the data.

```
Technique 1 : 65 87 73 79
Technique 2 : 75 69 83 81 72
```

What type of design is this?

(a) Randomized paired design

(b) Completely randomized design

(c) Controlled regression design

(d) Latin square design

35. Regression was performed using a response variable (Y) and a predictor (X). The regression equation obtained is $\hat{Y} = 4 - 0.23X$. Does the data show regression towards the mean?

(a) Yes

(b) No

(c) Insufficient information.

Appendix E

Bibliography

Below are references to books cited in the text plus some additional references. There are many good introductory statistical methods books in the literature. These include the books by Bhattacharya and Johnson, Kitchens, and Rasmussen cited below. The books by Stout et al. and Simon also discuss resampling for statistical inference. The books of Siegel and Castellan and of Hollander and Wolfe offer an elementary treatment of nonparametrics statistics (which includes the Wilcoxon procedures discussed in the text). The book of Hettmanspeger and McKean offers a more theoretical treatment, but it also includes the baseball data. The article by McKean, Vidmar and Sievers discusses the random drug screen which produced the quail data used frequently in the text. The particular quail data set used is presented in the article by McKean and Vidmar. The accompanying software of the text made use of the software cited in the references to Kapenga et al. and R.

Banfield, J. (1999), Rweb:Web-based Statistical Analysis, *Journal of Statistical Software*, 41, 01.

Bhattacharya, G. K. and Johnson, R. A. (1977), *Statistical Concepts and Methods*, New York: John Wiley & Sons.

Hettmansperger, T. P. and McKean, J. W. (1998), *Robust Nonparametric Statistical Methods*, London: Arnold.

Hollander, M. and Wolfe, D., A. (1999), *Nonparametric Statistical Methods, 2nd Edition*, New York: Wiley.

Kapenga, J. A., McKean, J. W. and Vidmar, T. J. (1988), *RGLM: Users Manual*, Amer. Statist. Assoc. Short Course on Robust Statistical Procedures for the Analysis of Linear and Nonlinear Models, New Orleans.

Kapenga, J. A., McKean, J. W. and Vidmar, T. J. (1995), *RGLM: Users Manual, Version 2*, SCL Technical Report, Dept. of Statistics, Western Michigan University.

Kitchens, L. (1998), *Exploring Statistics*, 2nd Edition, Pacific Grove, CA: Duxberry Express.

McKean, J. W. and Vidmar, T. J. (1994), A comparison of two rank-based methods for the analysis of linear models, *The American Statistician*, 48, 220-229.

McKean, J. W., Vidmar, T.J., and Sievers, G. L. (1989), A robust two stage multiple comparison procedure with application to a random drug screen, *Biometrics 45*, 1281-1297.

R (2000), R Development Core Team, R-core@r-project.org.

Rasmussen, S. (1992), *An Introduction to Statistics with Data Analysis*, Pacific Grove, CA: Brooks/Cole Publishing Company.

Siegel, S. and Castellan, N. J. (1988), *Nonparametric Statistics for the Behavior Sciences*, 2nd Edition, New York: McGraw Hill.

Simon, J. (1995), *Resampling: The New Statistics*, Arlington : Resampling Stats, Inc.

Stout, W. F., Travers, K. J. and Marden, J. (1998), *Statistics Making Sense of Data*, Rantoul, IL: Mobius.

Vardeman. S. B. (1994), *Statistics for Engineering Problem Solving*, Boston: PWS Publishing Company.

Wilcoxon, F. (1945), Individual comparisons by ranking methods. *Biometrics*, 1, 80-83.

Index