David J. Olive

Data Science Without Much Math

January 15, 2019

Preface

This book covers Data Science for students who meet the college entry requirements for Mathematics, but who have not had College Algebra. Hence the student should have had high school algebra and some computer experience. At Southern Illinois University (SIU), the course Math 102 meets the core curriculum. The texts Verzani (2014) and Wilcox (2017) seemed to be close to the appropriate level for such a Data Science course.

A good introductory Statistics text with a College Algebra prerequisite is Moore (2007). This text, and later editions, is used in Math 282 at SIU. This text tries to teach students how to read computer output and to give problems that many Math 282 students would find easy. The *R* software is used. See R Core Team(2016).

Some highlights of this text follow.

- The free software *R* is used.

Downloading the book's R functions *dspack.txt* and data files *dsdata.txt* into *R*: The commands

```
source("http://lagrange.math.siu.edu/Olive/dspack.txt")
source("http://lagrange.math.siu.edu/Olive/dsdata.txt")
```

The *R* software is used in this text. See R Core Team (2016).

Acknowledgements

Contents

Chapter 1
Introduction

This chapter gives a brief introduction to Statistics and Data Science, and describes data sets.

1.1 Introduction

Definition 1.1. Statistics is the science of extracting useful information from data.

There are at least three definitions for Data Science. First, for some researchers, Statistics = Data Science. Hence Data Science is the science of extracting useful information from data. Second, many researchers consider Data Science to be a new interdisciplinary field or discipline that is an extension of Statistics. See Cleveland (2001) and Figure 1 in Cook and Forzani (2018). Third, some researchers consider Data Science to be Statistics applied to big data sets. We ignore the third definition.

This book gives an introduction to the Statistics portion of Data Science for students who have had high school algebra and some computer experience. Hence the course is the lowest level Statistics (or Data Science) course that a college student should be able to take. Students good at math should take a first course in Statistics that has a calculus prerequisite. There are also Statistics courses that have College Algebra as a prerequisite.

This text uses the free statistical software *R*. See R Core Team (2016).

Chapter Two considers graphs for summarizing data such as bar graphs, boxplots, dot plots, histograms, and stem plots. Chapter Three considers numerical summaries that are Statistics, such as the sample mean and the sample median. Chapter 4 considers the normal distribution while Chapter 5 covers scatterplots and correlation. The remaining chapters cover regression, sampling, probability, confidence intervals, hypothesis tests, and classification and regression trees.

1.2 The Data Set

A data set or dataset is a collection of data. *Individuals* are the objects described by a data set. A *random variable* or *variable* is a characteristic recorded about an individual.

Definition 1.2. A **case** or **observation** consists of p random variables measured for one person or thing. The ith case $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T$. A data set consists of n cases, and n is the sample size.

Example 1.1.

```
crancap     hdlen   hdht
 1485       175     132
 1450       191     117
 1460       186     122
 1425       191     125
 1430       178     120
 1290       180     117
   90        75      51
```

The above data set has $p = 3$ random variables and $n = 7$ cases. The random variables are the head measurements *cranial capacity, head length,* and *head height.* The ith case $\boldsymbol{x}_i$ is usually written as a column so $\boldsymbol{x}_i^T$ is written as a row (the T is called the transpose). Hence the first case $\boldsymbol{x}_1^T = (1485, 175, 132)$. The third random variable is $\boldsymbol{v}_3$ written as the third column. Hence $\boldsymbol{v}_3^T = (132, 117, 122, 125, 120, 117, 51)$. The first row in the data set is a header giving abbreviations for the random variable names.

Assume that the data $\boldsymbol{x}_i$ has been observed and stored in an $n \times p$ matrix

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_p \end{bmatrix}$$

where the ith row of $\boldsymbol{W}$ is the ith case $\boldsymbol{x}_i^T$ and the jth column $\boldsymbol{v}_j$ of $\boldsymbol{W}$ corresponds to n measurements of the jth random variable x_j for $j = 1, ..., p$.

In the statistical software R, the data set will be denoted by a symbol, such as x, or a name, such as "major."

In this text, often the data set will consist of one random variable, for example, height. Then the R code below puts the data into x.

```
x <- c(132,117,122,125,120,117,51) #<- means ``gets"
#hdht <- x  #hdht has the same data as x
> x
[1] 132 117 122 125 120 117  51
```

```
#> is the prompt on the computer screen
#the pound sign, #, is used to insert comments
```

Here the random variable is *head height* and there are $n = 7$ cases.

Definition 1.3. A *categorical variable* takes on several categories.

Tips: i) Often count the number in each category or find the percentage. ii) Adding or averaging the categories does not make sense.

Definition 1.4. A *quantitative variable* takes on numerical values.

Tip: Adding or averaging a quantitative variable makes sense.

Example 1.2. Consider a) *race*, b) *hair color*, c) *gender*, d) *height*, and e) *number of emails received* on a specified day. The first three variables are categorical while the last 2 variables are quantitative.

1.3 Summary

1) Statistics = Data Science is the science of extracting information from data.

2) A *categorical variable* takes on several categories.
Tips: i) Often count the number in each category or find the percentage. ii) Adding or averaging the categories does not make sense.

3) A *quantitative variable* takes on numerical values.
Tip: Adding or averaging a quantitative variable makes sense.

4) From a story problem, you should be able to determine the individuals and the variables. You should know whether the variable is categorical or quantitative.

1.4 Complements

There are many Statistics and Data Science texts at a higher level than this one. For students with College Algebra, Moore (2007) is a good text.

1.5 Problems

1.1. Four of the following five variables are categorical. Which variable is quantitative? race, hair color, gender, major, age

1.2. Four of the following five variables are quantitative. Which variable is categorical? height, weight, gender, GPA, age

1.3. A student can receive a grade point average (GPA) of any number between 0.0 and 4.0. What type of variable is "GPA"?

1.4. A student can receive a grade of A, B, C, D or F. What type of variable is "grade"?

Chapter 2
Summarizing Data With Graphs

This chapter considers bar graphs, box plots, dot plots, histograms, and stem plots. The distribution of a categorical variable lists counts (frequencies) or percents.

Definition 2.1. The *distribution* of a variable tells what values it takes and how often.

2.1 The Bar Graph for Categorical Data

Example 2.1. 2017 SIU student gender F: 47.3%, M: 52.7%. For decades, SIU has been one of the few universities where the female percentage is lower than the male percentage. In *R*, the bar graph is made with the barplot command. See Figure 2.1 and the *R* code below.

```
barplot(c(47.3,52.7),names.arg=c("F","M"))
```

Definition 2.2. A *bar graph* (or barplot or barchart or bar plot) is used to display categorical data. The vertical axis height = percent or count and the horizontal axis has categories. Separate the bars with a space. Bar widths are equal.

Example 2.2. The *R* data set `VADeaths` gives the death rates measured per 1000 population per year. They are cross-classified by age group (rows) and population group (columns). The age groups are: 50-54, 55-59, 60-64, 65-69, 70-74 and the population groups are Rural/Male, Rural/Female, Urban/Male and Urban/Female. Try the following command.

```
barplot(VADeaths)
```

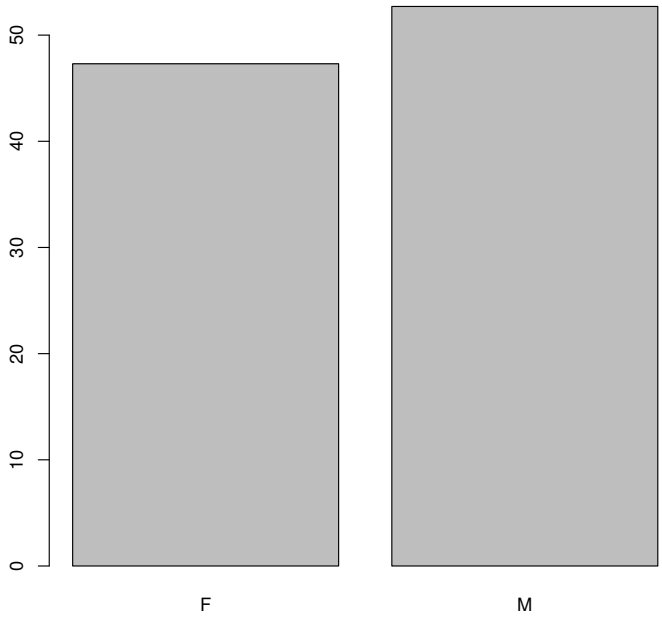


Fig. 2.1 Bar Graph for 2017 SIU Student Gender

2.2 Graphs for Quantitative Variables

We will use two data sets. The *R islands* data set gives the areas in thousands of square miles of the land masses which exceed 10,000 square miles. The Buxton (1920) data set has several variables that are measurements taken on 87 people. We will be interested in the *height* in mm in the variable `buxy`. Five heights were recorded near 19mm (about 0.7 inches) high. These five cases are outliers.

Definition 2.3. An *outlier* is a case that lies far away from the bulk of the data.

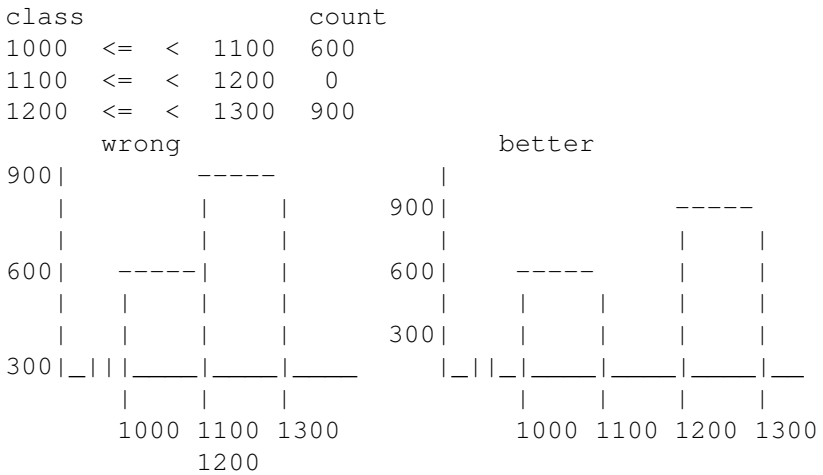
Definition 2.4. A *(frequency) histogram* is a graph that summarizes the distribution of a quantitative variable. Divide the range of the data into *classes* of *equal* width. Each observation should fall in *exactly one* class. Find the count of observations for each class (often use a tally) if a distribution table of classes and counts is not given. On the horizontal axis, mark the scale

of the variable. Put the count (frequency) on the vertical axis. Bars have the same width. Labeling the top of each bar with the count can be useful.

Since the bars have equal width, the area of each bar is proportional to the percentage of observations in each class. Bar graphs have gaps between bars, histograms have no gaps unless a class has a count of 0.

Warning. Do not make breaks in the vertical axis for a bar graph or a histogram. If a break is made, then the area of the bars is no longer proportional to the percentage of observations in each class. Distances should be equally spaced. There can be one break on the horizontal axis for a histogram, but distances should be equally spaced. **This error is so common**, that the axes should be given for exams and quizzes for this class.

Example 2.3. The set of axes to the lower left has a break in the vertical axis and the distance from 0 to 300 is not equal to the distance between 300 and 600 or 600 and 900. The horizontal axis is wrong since the interval from 1100 to 1200 corresponding to a zero count has been omitted. The set of axes to the lower right is better although the distance from 0 to 300 appears to be less than the distance of 300 to 600. There is one break in the horizontal axis just before 1000, denoted by the two small vertical bars. The distance from the right edge of the graph to 1000 is not the same as the distances of 100 that separate the four numbers. The horizontal break can be useful to show detail that would be obscured if no break was used: 1000 to 1300 would be a small part of the graph if the horizontal axis used 0, 100, 200, ..., 1300, 1400. Note that an observation of 1200 goes with the 3rd class, not the 2nd class.



To interpret a histogram, look for an overall pattern and deviations from the pattern. Shape, center, and spread are important. The center is where the histogram is located (a typical or center value on the horizontal axis). There are three common shapes: left skewed, right skewed, and approximately

symmetric. A graph is symmetric if graph is a mirror image about some midpoint. A graph is right skewed if it has a long right tail, e.g. income data, where most incomes are 50000 or less but a few incomes are very large. A graph is left skewed if it has a long left tail. The left tail is the leftmost part of the graph while the right tail is the rightmost part of the graph. If x is right skewed then $-x$ is left skewed. If x is left skewed then $-x$ is right skewed.

Example 2.4. Figure 2.2 has four histograms: the one in the upper left is approximately symmetric and the two tails are about the same length. The one in the upper right is right skewed with a long right tail. The one in the lower left is left skewed with a long left tail. The last histogram of $\log(x)$ is left skewed, but is less skewed than the histogram of x. Each artificial data set has 1000 cases. See the following R code.

```
par(mfrow=c(2,2)) #four graphs
hist(rnorm(1000))
x <- rexp(1000)
hist(x)
hist(-x)
hist(log(x))
par(mfrow=c(1,1))
```

Using $y = \log(x)$ can cause y to have less skew than x if x is skewed. This function is especially useful for reducing skew if $x > 0$ and $\max(x)/\min(x) \geq 10$ where $\max(x)$ is the largest (maximum) value in the data set, and $\min(x)$ is the smallest (minimum) value in the data set.

Definition 2.6. The logarithm function is the inverse function of exponentiation: $\log_b(b^y) = y$ where b is the base of the logarithm. Hence $\log_{10}(100) = \log_{10}(10^2) = 2$. We will usually use the natural logarithm with base $b = e \approx 2.72$, denoted by $y = \log(x)$. We need $x > 0$.

For the median and quartiles, see Chapter 3. The box plot is roughly symmetric if the line corresponding to the median is close to the middle of the plot, and the whiskers have about the same length. If the right whisker is longer than the left (or the circles extend further to the right), then the data is likely right skewed. If the left whisker is longer than the right (or the circles extend further to the left), then the data is likely left skewed.

Definition 2.7. The *five number summary* is the minimum, Q_1, the median, Q_3 and the maximum. The *box plot* or boxplot is a box from Q_1 to Q_3 with a line at the median $= Q_2$. If sketched by hand, whiskers extend from Q_1 to the minimum and from Q_3 to the maximum. R uses another rule to make the whiskers, say Q_1 to low and Q_3 to high, and puts circles past the whiskers to indicate possible outliers.

Warning. Computer output often gives several numbers besides the five number summary, such as the (sample) mean.

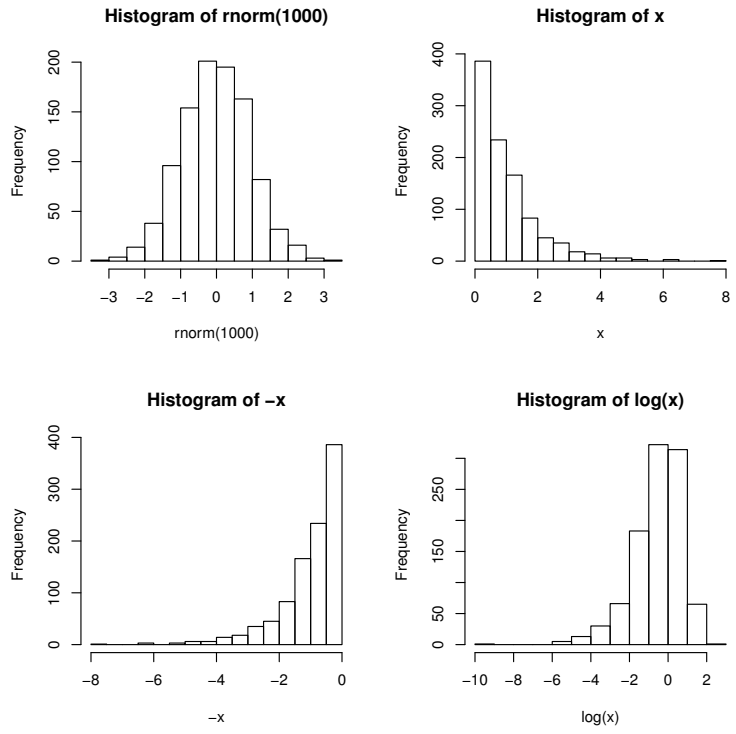


Fig. 2.2 Histograms for Example 2.4.

Example 2.5. Figure 2.3 shows histograms and boxplots for the islands data and the Buxton (1920) heights data. Note that the islands data is right skewed and the heights data has outliers (gaps in the plot). The box plot for the heights data looks roughly symmetric if the outliers are ignored.

Stem plots and dot plots are for small data sets.

Definition 2.8. To make a *stem plot*, divide the data into groups = stems that contain all but the final digit = leaf. Place the stems in order in a vertical column (e.g. smallest on top, largest on bottom). A vertical line separates the stems from the leaves. Each leaf is written to the right of its stem in increasing order. Write the stem and leaf units on the plot or tell where the decimal goes.

R tells where the decimal goes and truncates the data rather than rounding the data. Suppose the stem is 4 and the leaf is 5. If the decimal is j digits to the right of the | (stem), then the value is $4.5(10^0) = 4.5$ if $j = 0$ (the decimal is at the |), $4.5(10) = 4.5(10^1) = 45$ if $j = 1$, $4.5(100) = 4.5(10^2) = 450$ if

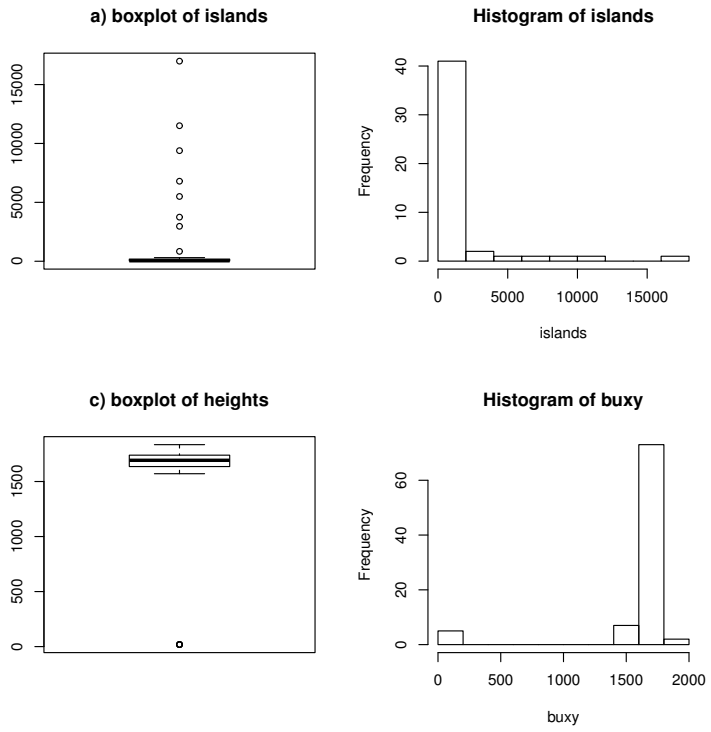


Fig. 2.3 Histograms and Box Plots for Example 2.5.

$j = 2$, and $4.5(10^3) = 4.5(1000) = 4500$ if $j = 3$. If the decimal was 1 digit to the left of the | (stem), then the value is 0.45. If the stem unit is ones and the leaf unit is tenths, then the value is $4(1) + 5(0.1) = 4.5$. If the stem unit is tens and the leaf unit is ones (very common), then the value is $4(10) + 5(1) = 45$.

Example 2.6. For the island data, the 4 and 5 correspond to $4500 = 4.5(1000)$. The 16|0 corresponds to $16000 = 16.0(1000)$. For log(islands), stem 9 with leaf 7 corresponds to 9.7 and $\log(16988) = 9.740$.

```
max(islands)  #largest value
[1] 16988
16000  #17 | 0 would have been better
> stem(islands)
```

```
  The decimal point is 3 digit(s) to the right of the |

   0 | 000000000000000000000000000000111111222338
   2 | 07
   4 | 5
   6 | 8
   8 | 4
  10 | 5
  12 |
  14 |
  16 | 0

> stem(log(islands))

  The decimal point is at the |

  2 | 566666778889
  3 | 01234444556778889
  4 | 134445
  5 | 22467
  6 | 7
  7 |
  8 | 0268
  9 | 147
par(mfrow=c(2,2)) #four graphs
boxplot(islands)
title("a) boxplot of islands")
hist(islands)
boxplot(buxy)
title("c) boxplot of heights")
hist(buxy)

stripchart(islands) #dot plot
title("a) dot plot for islands")
stripchart(log(islands))
title("b) dot plot for log(islands)")
stripchart(buxy)
title("c) dot plot for heights")
stripchart(buxy,method="jitter")
title("d) jittered dot plot for heights")

boxplot(log(islands))
?stripchart #same as help(stripchart)
hist(log(islands))
par(mfrow=c(1,1))
```

Definition 2.9. A *dot plot* consists of an axis and plotted points for each value of the data set.

Example 2.7. Figure 2.4 shows dot plots for the islands data and the Buxton (1920) heights data. Note that the islands data is right skewed and the heights data has outliers (gaps in the plot). Jitter adds some noise to the plotted points so it is easier to see the plotted points. The heights data has 5 outliers, and it is easier to see that there is more than one outlier in the jittered dot plot of Figure 2.4d) than in the dot plot of Figure 2.4c). See the above *R* commands.

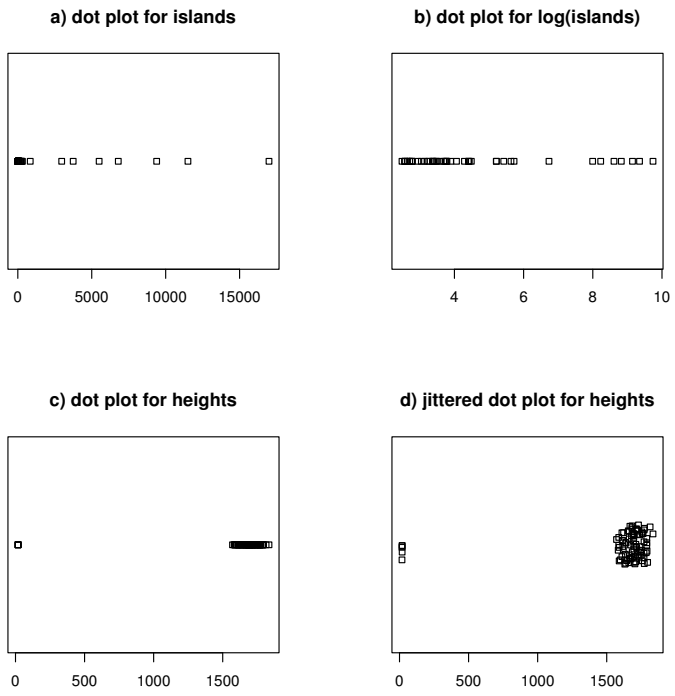


Fig. 2.4 Dot Plots for Example 2.7.

Example 2.8: Gender Intelligence. The following data is from the Nov. 16, 2011 *Chicago Tribune.* The data is the percentage of boys and girls meeting or exceeding the state standards at each grade level. Note that the math and science results are almost the same for grades 3 to 8, suggesting that gender intelligence is about the same. In 2001, more males than females took science and math in high school. According to a 2008 fall ABC news

report, female 11th graders took as much science as males, and from about 2008, females have tended to score at least as high as males on standardized tests in science. Since each result has two categories: meeting or exceeding standards and failing to meet standards, the results could be displayed with a bar graph for each gender with two bars instead of four: just display the percent meeting or exceeding standards for each gender and for each grade discipline combination. (For grades 3 or 4 and 8 or 7, the first 3 categories were for the 1st grade, 3 or 8, while the last two categories science and social science were for grades 4 or 7.)

Table 2.1 Comparing Boys and Girls

grade	reading		writing		math		science		social science	
	B	G	B	G	B	G	B	G	B	G
3 or 4	60	65	53	63	74	74	66	65	62	60
5	57	60	63	77	61	62				
8 or 7	64	67	52	71	51	50	72	72	60	60
11	54	61	53	65	56	52	54	47	62	53

In some countries, gender differences are large (significant) for science and math. See Beaton et al. (1996). The differences were not likely due to gender differences in intelligence, but rather to the amount of classes taken by each gender, the percentage of female teachers in the sciences and mathematics, whether the teachers or students think that boys are better in science and math, gender opportunities for higher education, et cetera. For example, in the 1996 study, eighth grade girls disliked the sciences much more than boys, and 20% of the science teachers were female. In the US, 54% of the science teachers were female.

Example 2.9: Sexual Activity. If we use a Big Bang Theory term for sex, such as coitus, each time a man has sex with a woman, a woman has sex with a man, and vice versa. Hence the total number of times all women have sex in a year is equal to the total number of times all men have sex in a year. Men tend to claim to have more partners while women tend to claim to have fewer partners than the actual number. Table 2.2, taken from Student (1998), shows the estimated annual occasions of sex by age and gender in the USA. The numbers can be roughly explained by the numbers of men and women in each group. Also the male prison population is much higher than that of women. There are more young men than young women, so young women have the most sexual activity. Old men who are willing and able are greatly outnumbered by old women who are willing and able, and hence old men have much more sexual activity than old women.

Table 2.2 Estimated Annual Occasions of Sexual Activity

gender/age	18-24	25-34	35-40	40-54	55-64	65-74	75up
M	83	85	73	55	52	23	13
W	86	84	65	50	25	10	2

2.3 Summary

1) You should know how to make a bar graph for categorical data. Bar graph bars should have bases that have the **same length**.

2) **Do not make breaks in the vertical axis of bar graphs and histograms** because the area of the bars is proportional to the percentage of cases in each class.

3) Given a distribution table, make a histogram. Given a histogram and a rule for the endpoints (e.g. bar includes right endpoint but not the left endpoint), you should be able to make a distribution table.

4) Given a list of numbers, make a stemplot. **Include the stem units and leaf units on the plot.** For example the number 205 will have stem 20 and leaf 5 with stem units = tens and leaf units = ones. Note that $20(10)+1(5) = 205$.

5) Sometimes a list of numbers is presented as a stemplot, then you are asked to find the mean, median, and SD of the list. The stem and leaf units are used to determine what the list of numbers is.

6) Given a list of numbers or *R* output, find the five number summary: min, Q1, median, Q3, and max. Recall that the data is **sorted from smallest to largest.** The median is the "middle number", Q1 is the median of the sorted numbers to the left of the median, and Q3 is the median of the numbers to the right of the median. Use the five number summary to make a box plot.

7) Given a box plot, bar graph, stemplot, histogram, or dot plot, be able to give a short summary of what the plot tells you. For example are outliers present, is the histogram symmetric, right skewed or left skewed? How does the proportion of one category compare to the proportion of another category? *R* commands are `boxplot, barplot, stem, hist`, and `stripchart`.

8) From a story problem, you should be able to determine the individuals and the variables. You should know whether the variable is categorical or quantitative.

2.4 Complements

2.5 Problems

2.1. The stem-and-leaf display above is for 71 Stat 3011 final exam scores from around 1998. The lowest score was a 30 while the highest was a 92. The mean score was 69.2, the median score was 72, and the standard deviation of the scores was 15.8.

```
3 | 045678
4 | 266
5 | 01247899
6 | 01444555678889
7 | 011222334566888889        Stem: tens
8 | 1112222344456777888       Leaf: ones
9 | 122
```

If score of 59 or lower was a failing grade, what proportion of students failed this final?

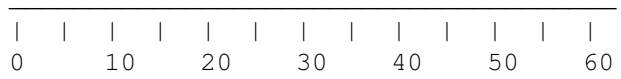
a) 0.17 b) 0.24 c) 0.76 d) 0.024 e) 0.048

2.2. Twelve students took Math 580. From their quiz scores listed below, make a stemplot. Put the stem and leaf units to the right of the plot.

80 89 90 97 72 91 81 88 83 87 87 88

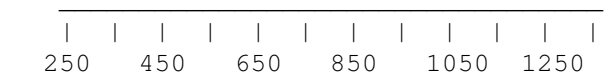
2.3. The weights, in ounces, of malignant tumors removed from 51 patients is displayed in the table below. Data is from Mendenhall and Beaver (1991, p. 19). Make a histogram or bar graph of the data and **state which plot you used**. Use labels like those shown below.

Class	Count
$10 - < 20$	5
$20 - < 30$	19
$30 - < 40$	10
$40 - < 50$	13
$50 - < 60$	4



2.4. Data for federal aid per capita for the 50 states in 1986 is summarized below. Data is from Mendenhall and Beaver (1991, p. 19). Make a histogram or bar graph of the data and **state which plot you used**Use labels like those shown below.

```
class             count
250  <=  <  450     26
450  <=  <  650     20
650  <=  <  850      1
850  <=  < 1050      1
1050 <=  < 1250      1
```



2.5. Data is taken from the following newspaper article: Herndobler, K. (Aug. 27, 2002), "Illinois ACT Scores Bring Down the National Average," *Daily Egyptian*. The ACT scores during the 2001-2002 academic year were 22.8 for Carbondale, 20.8 for the USA, 20.1 for Illinois, and 16.5 for Chicago. (This was the first year that all Illinois high school juniors had to take the ACT, and the number taking the ACT jumped from 89000 in the previous year to 129000.) Display the ACT data with either a histogram or bar graph. Which plot did you use?

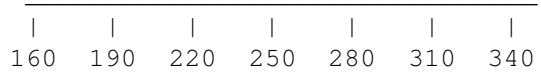
2.4. The lengths of reign of 13 rulers of England and Great Britain are listed below. Data is from Rossman and Chance (2011, p. 147).

```
21  13  35  19  35  10  17  56  35  20  50  22  13
```

Make a stem plot for the data. Do not forget to include the stem and leaf units.

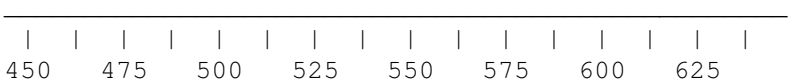
2.5. The above data below is from Gould and Ryan p. 117. The numbers above are revenue (in millions) from the top ten Pixar animated movies as of June 2010. $Q_1 = 206, M = 245$, and $Q_3 = 261$. Draw a box plot for the movie data. Use labels like those shown below.

```
192  163  246  256  340  261  244  206  224  293
```



2.6. Suppose that the mean GRE math scores for the 50 states and the District of Columbia were entered into a computer. The computer gave the following descriptive statistics. From these statistics, draw a boxplot of the 51 GRE scores. Use labels like those shown below.

```
 N  MEAN   MEDIAN  STDEV   MIN    MAX    Q1      Q3
51  529.30 521.00  34.83   473.00 600.0 500.0  557.0
```



2.7. The number of violent crimes (per 100000 people) in 2002 are summarized for the eight states below. Data is from Sullivan (2006, pp. 50, 87). From these statistics, draw a boxplot.

2.8. In the 1996 presidential election, about 50% of the voters voted Democrat and 40% voted Republican and about 10% of the voters voted "Other". Display this information with an appropriate plot and say what type of plot you used. Use labels like those shown below.

```
N  MEAN   MEDIAN  STDEV   MIN  MAX   Q1   Q3
8  479.5   497    112.60  311  599   388  578
```

```
 |    |    |    |    |    |    |    |    |
300  340  380  420  460  500  540  580  620
```

2.9. Fourteen students took Math 484. a) From their scores listed below, make a stemplot. Put the stem and leaf units to the right of the plot.

```
78 100 85 100 87 90 91 100 98 98 95 100 91 99
```

2.10. The proportion of US adults over 25 whose highest educational attainment (no high school degree, a high school degree, some college but no degree, BA or BS degree, or advanced degree) are given below. Data is from Sullivan (2014, p. 59). Make a histogram or bar graph of the data and **state which plot you used**. Use labels like those shown below.

```
noHS         0.1544
HS           0.3202
some college 0.1715
BA or BS     0.2612
adv. degree  0.0927
```

```
noHS      HS     somecollege   BAorBS   AdvDeg
```

2.11. A major newspaper reported the Budweiser's share of the beer market. Make a bar graph or histogram from the table below.

```
brand of beer    percent of market share
Budweiser           25%
Bud Light           22%
Other               53%
```

2.12. The percentage of intercity passengers travelling by bus, air, subway and Amtrack is listed below. Make a bar graph or histogram from the table below.

```
type       percent
Air         40.6 %
```

Amtrak	1.6%
Bus	29.2%
Subway	28.6%

2.13. Suppose that the grade distribution for Math 282 in 2010 was A - 12, B - 27, C - 8, D - 2, F - 1. (So 12 students get A's, 27 get B's, etc.) Make a histogram or bar chart of this data (one of these is inappropriate).

2.14. The November 16, 2001 *Chicago Tribune* reported the percentage of boys and girls meeting or exceeding state standards from achievement tests at several grade levels in Math, Science and Social Science. Make a histogram or bar graph of the data and **state which plot you used**. Use labels like those shown below. Does the plot suggest that males are
a) smarter, b) less smart **or** c) about as smart as females? (**Circle one.**)

group	label	% passing
Math 8th grade Female	(MF)	50
Math 8th grade Male	(MM)	51
Science 7th grade Female	(SF)	72
Science 7th grade Male	(SM)	72
Soc. Sci. 7th grade Female	(SSF)	60
Soc. Sci. 7th grade Male	(SSM)	60

MF MM SF SM SSF SSM

Chapter 3
Summarizing Data With Statistics

simulation an

3.1 Summary

1) Given a list of n numbers, find the mean $\overline{x}$, median, and standard deviation SD $= S$. Recall that $\overline{x} = \sum x/n$, $S = \sqrt{\sum (x - \overline{x})^2/(n-1)}$ and the median is found by finding the middle number(s) of the data ordered from smallest to largest.

3.2 Complements

3.3 Problems

3.1. The plot below shows 9 IQ scores for the 9 top Stat 5021 students.

```
10 | 1
11 | 2  6
12 | 3  7  8
13 | 4  9              stem: tens
14 | 5                 leaf: ones
```

Find the mean $\bar{x}$ and the median of the 9 scores.

a) $\bar{x} = 125$, median $= 127$
b) $\bar{x} = 12.5$, median $= 12.7$
c) $\bar{x} = 125$, median $= 125$
d) $\bar{x} = 12.5$, median $= 12.5$
e) $\bar{x} = 127$, median $= 125$

3.2. Which statement is false, or are none of the statements false?

a) A population is the entire collection of individuals about which information is desired.
b) A sample is a subset of the population actually examined.
c) For a stem plot, the units for the stem and leaf should appear someplace in the display.
d) If a histogram is right skewed, then the mean is greater than the median.
e) None of the four statements above is false.

3.3.

x^2	$(x-\bar{x})^2$
1	4
4	1
9	0
16	1
25	4

i) Find $\sum(x-\bar{x})^2$ and the standard deviation S.

a) $\sum(x-\bar{x})^2 = 10$, S = 2.5 b) $\sum(x-\bar{x})^2 = 5$, S = 1.12
c) $\sum(x-\bar{x})^2 = 10$, S = 1.58 d) $\sum(x-\bar{x})^2 = 2$, S = 0.25
e) $\sum(x-\bar{x})^2 = 10$, S = 3.16

ii) If $\sum x = 15$, find $\sum x^2$ and $\bar{x}$.

a) $\sum x^2 = 225$, $\bar{x} = 3.0$ b) $\sum x^2 = 225$, $\bar{x} = 15$ c) $\sum x^2 = 55$, $\bar{x} = 15$ d) $\sum x^2 = 55$, $\bar{x} = 3$ e) not possible

3.4. Consider the following list.

40, 36, 36, 38, 50, 64

The mean and median of the list are

a) mean = 44, med = 36, b) mean = 36, med = 40 c) mean = 44, med = 39, d) mean = 264, med = 36 e) mean = 40, med = 40

3.5. Below is a list of the state income tax paid by a TA from 1993 to 1997. Find the sample mean and median of these numbers.

```
41          296          303          276          345
```

3.6. The time in seconds for five rats to complete a maze were 24, 37, 38, 43, and 33.

a) Find the sample mean and median of these numbers.

b) Find the standard deviation s of these numbers.

3.7. Suppose that the number of Math 282 students who failed to turn homework was counted for the last five homeworks. The results were 7, 11, 9, 7, and 6.

a) Find the sample mean and median of these numbers.

b) Find the standard deviation s of these numbers.

3.8. Following Daniel, suppose that the fasting blood glucose levels of 5 children were 56, 61, 57, 77, and 62.

a) Find the sample mean and median of these numbers.

b) Find the standard deviation s of these numbers.

3.9. Suppose that $\sum(x_i - \overline{x})^2 = 616$ and $n = 6$. Find the standard deviation s.

Chapter 4
The Normal Distribution

Example 4.1. From texts on abnormal psychology and a PBS documentary based on Hicks and Wattenberg (2000), individuals with IQ scores 2 standard deviations below average (2.5% of the population) are retarded, mildly retarded, mentally deficient, or morons. Low IQ would be better. Individuals in the upper 2.5% are high IQ or gifted.

4.1 Summary

9) Know how to do a **forwards calculation using table A.** In the story problem you will be told that X is approximately normal with some mean and SD. You will be given one or two X^* values and asked **to find a proportion** or probability or chance. Draw a line and mark down the mean and the X^* values. Standardize each X value by taking the z-score $Z^* = (X^* - \mu)/\sigma$. If you want the chance that X is **less than** X^*, then table A gives the correct value. If you want the chance that X is **greater than** X^* take 1 - table A value. If you want the chance that X **is between** X_1^* and X_2^* subtract the smaller value from the larger value of table A. Given a z-score, to use table A you use the leftmost column and top row of table A. Intersect this row and column to get a 4 digit decimal which is equal to the area to the left of Z^*.

2) Know how to do a **backwards calculation using table A.** Here you are **given a proportion and asked to find** one or two X^* values. Table A gives areas to the **left** of Z^*. So if you are asked to find the top 5%, that is the same as finding the bottom 95%. If you are asked to find the bottom 25%, table A gives the correct value. If you are asked to find the two values containing the middle 95%, then 5% of the area is outside of the middle. Hence .025 area is to the left of $X^*(lo)$ and $.025 + .95 = .975$ area is to the left of $X^*(hi)$. Once you know the area to the left of X^*, find the largest 4 digit number smaller than the desired area and the smallest 4 digit number

larger than the desired area. These two numbers will be found in the middle of table A. Take the number closest to the desired area, and to find the corresponding Z^*, examine the row and column containing the number. Go along the row to the entry in the leftmost column of table A and go along the column to the top row of table A. For example, if your 4 digit number is .9750, $Z^* = 1.96$. To get the corresponding X^*, use $X^* = \mu + \sigma Z^*$.

3) Given a density that is box shaped with base from a to b, know that the height of the density is $1/(b-a)$ and that the chance that X is between c and d where $a \leq c < d \leq b$ is given by (base)(height) $= (d-c)/(b-a)$.

4.2 Complements

4.3 Problems

4.1.

Chapter 5
Scatterplots and Correlation

simulation an

5.1 Summary

5.2 Complements

5.3 Problems

5.1. The correlation coefficient

a) can be any positive or negative number.
b) measures the variability of the x and y values.
c) is a measure of linear association.
d) only takes on values between 0 and 1.
e) shows whether x causes y.

5.2. In the table below, note that $\overline{x} = 0.5333$, $s_x = 0.3266$, $\overline{y} = 14.0$, and $s_y = 0.6132$. Use the table below to find the correlation r between x and y.

x	y	$z_x = (x - \bar{x})/s_x$	$z_y = (y - \bar{y})/s_y$	product $z_x z_y$
0.1	14.9	−1.3267	1.4677	−1.9472
0.2	14.5	−1.0205	0.8154	−0.8321
0.5	13.4	−0.1020	−0.9785	0.0998
0.7	14.1	0.510	0.1631	0.0832
0.8	13.4	0.8166	−0.9785	−0.7990
0.9	13.7	1.1228	−0.4892	−0.5493

Chapter 6
Regression

simulation an

6.1 Summary

1) Be able to find the least squares line $\hat{y} = a + bx$ from R output (a is under *Coef* and to the right of *Constant* while b is below a). A typical table and a table with numbers are shown below.

```
predictor        coef      stdev      T       Pvalue
Constant          a
x                 b       unimportant numbers for exam 2

predictor        coef      stdev      T       Pvalue
constant         272.819   63.4963   4.297    0.0000
sternal height   1.01482   0.04537   22.370   0.0000
```

2) Be able to find the least squares line $\hat{y} = a + bx$ given 2 means, 2 SD's and the correlation r. Recall that $b = rs_y/s_x$ and $a = \bar{y} - b\bar{x}$. Remember that **the response** y is what you want to predict. The explanatory variable x is used to help predict y.

3) After being given or finding the slope and intercept or the line $\hat{y} = a+bx$, be able to predict y for a given value of x.

4) Given the least squares line $\hat{y} = a+bx$, be able to put it on a scatterplot.

5) Know that least squares should only be used if the scatterplot is linear (football shaped). A residual is $y - \hat{y}$ and a residual plot should be football shaped with zero slope. Know that extrapolation is risky.

6.2 Complements

6.3 Problems

6.1. It is desired to predict treadmill *speed* from an animal's body *mass* (in kilograms). Data is from Ross and Chance, 4th ed., p. 626. The least squares equation is $\hat{y} = 0.9894 + 0.005484x$.

a) What is the response variable?

b) Predict the speed if mass = 140.

6.2. Use the following information for problems 10) and 11). It is desired to predict *price* in dollars of a textbook from the number of *pages* of the textbook. Data is from Rossman and Chance, 4th ed., p. 630. The least squares equation is $\hat{y} = -3.42 + 0.147x$.

a) What is the response variable?

b) Predict the price if pages = 800.

Coefficient Estimates for Problem 6.3

Label	Estimate	Std. Error	t-value	p-value
Constant	-11.2262	6.74092	-1.665	0.1024
m750	1.00369	0.0140630	71.371	0.0000

6.3. Consider studying extremely high SAT verbal scores by state. Suppose that it is desired to predict the number of female students *f750* who get a 750 or higher on the SAT verbal score from the number of male students who got a 750 or higher on the SAT verbal score = x = *m750*. The data was obtained from $n = 50$ states. Predict *f750* if *m750* = 1240.

6.4. It is desired to predict the number of women (who got a 750 or higher on the SAT Verbal) from the number of men (who got a 750 or higher on the SAT Verbal). The least squares equation is $\hat{y} = -11.2262 + 1.00369x$.

a) What is the response variable?

b) Predict the number of women from a state where the number of men was 950.

6.5. It is desired to predict *brain weight* (in grams) from *head size*. The least squares equation is $\hat{y} = -1192.65 + 161.846x$.

a) What is the response variable?

b) Predict the brain weight if head size = 15.

6.6. It is desired to predict a fifth grade student's *reading score* from the student's *IQ score*. Data is from Moore (2007, p. 118). The least squares equation is $\hat{y} = -33.4 + 0.882x$.

a) What is the response variable?

b) Predict the reading score for a student who has an IQ score of 125.

6.7. It is desired to predict the weight of the brain (in grams) (y) from a measurement of the size of the head (x). Suppose that the least squares

line is $\hat{y} = a + bx = -1192.82 + 161.855x$. Predict the weight of the brain if $x = 13.28$.

6.8. Suppose it is desired to predict the yearly return from the stock market from the return in January. Assume that the correlation $r = 0.605$.

variable	mean	standard deviation
Y= yearly return	9.07	15.35
x = January return	1.75	5.36

a) Find the slope of the least squares line.

b) Find the intercept of the least squares line.

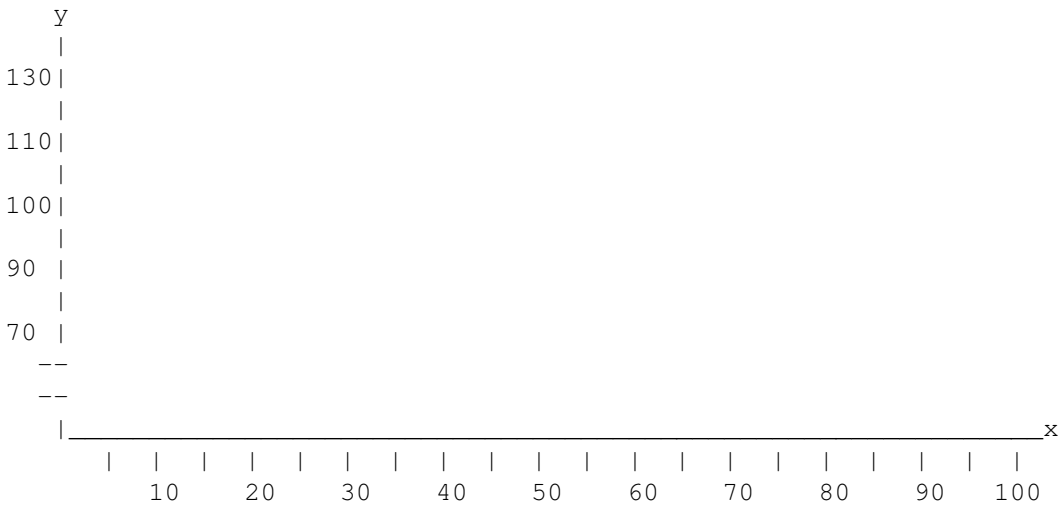
Label	Estimate	Std. Error	t-value	p-value
Constant	4.42337	0.856933	5.162	0.0000
shell mass	0.132712	0.00576835	23.007	0.0000

6.9. Predict a mussel's muscle mass if its shell mass = 128 using the output above.

6.10. Cement gives off heat as it hardens and the amount of heat is related to the amounts of certain chemicals in the cement. Suppose that the least squares line for predicting y = heat from x = chemical A is $\hat{y} = 57.4237 + 0.78912x$.

a) Predict the heat if the amount of chemical A is 40.

b) Suppose that x ranges from 20 to 100. Draw the least squares line. Show work.



predictor	coef	SE	T	P
constant	64.93	8.49058	6.76	0.0000
chem2	0.635	0.16839	4.69	0.0007

6.11. The table above represents output for 12 children. The explanatory variable x is age in months while the response variable y is the child's height in cm. Predict the height of a child who is 24 months old with the least squares regression line. Show work.

Chapter 7
Sampling

7.1 Nonscientific Surveys

Example 7.1. Following Von Hoffman (2000), the magazine *Time* had a web site that asked who was the most important person of the 20th century. Before a software crash, the leading vote getter was Ronnie O'Brien, an Irish soccer player. Then *Time* restricted the survey to 100 people, and Elvis won.

7.2 Scientific Surveys

Example 7.1. According to an April 2011 ABC news, 50% of sexually active young adults contract an STD and 1/8 couples of child bearing age can not have children.

Example 7.2. The December 2010, p. 3 NEA Higher Education Advocate observed that students of inexperienced teachers tend to do better on common finals than those of seasoned teachers. See Carrell and West (2010). An interpretation is that inexperienced teachers teach towards the final and seasoned teachers prepare students better. The evidence for this claim was that Calc I students from the experienced teachers do better in Calc II. Suppose 60% of inexperienced teachers and 50% of seasoned teachers students pass the Calc I final. A likely more accurate interpretation is that seasoned teachers are worse teachers so their students who pass are smarter on average. So in later classes, the smaller smarter group does better than the larger less smart group. Similarly, a larger proportion of women take the SAT and ACT than men. This fact drags down the average female score as compared to the average male score.

Example 7.3: People Massively Overrate Themselves. Price (2006) reported that a survey of professors at the University of Nebraska showed 94

percent thought they were better than average teachers at their own institution.

7.3 Sampling Distribution and the CLT

Example 7.1. Suppose that the population consists of the IQ scores of four children. The four scores are 110, 122, 125, and 132. The mean μ of the scores is 122.25. A random sample of size 2 will be selected without replacement. then the sampling distribution of $\bar{x}$ is shown in the table below. What is the probability that $\bar{x} > 118$?

sample	110,122	110,125	110,132	122,125	122,132	125,132
value of $\bar{x}$	116	117.5	121	123.5	127	128.5
probability	1/6		1/6	1/6	1/6	1/6

Solution: $P(\bar{x} > 118) = P(\bar{x} \geq 121) = 4/6 = 0.6667$.

7.4 Summary

1) Know the difference between an individual and a population.

2) Know that association does not imply causation.

3) Know what a lurking variable is.
2) Know how to get a SRS using table B. See Q4 4, HW4 D, E.

4) Know that voluntary response samples and samples of convenience are bad **regardless of the sample size** while probability samples are good.

5) Know that SRS's are too expensive so multistage samples are used when interviewers are sent out. Random digit dialing is used for many opinion polls.

6) Know that the accuracy of the probability sample depends on the size of the sample. Two SRS's of the same size have the same accuracy (if the sample size is small compared to both population sizes). **Bigger samples have greater accuracy.** Some times you will be given the sample size, sometimes a percentage of the population sampled (then you need to figure out which sample is larger).

7) Know that the normal approx for $\overline{x}$ holds for $n \geq 5$ if the population of x is approximately normal. Know that the CLT **does not apply** if $n \leq 30$ and x comes from a highly skewed population. Unless told otherwise, assume CLT holds for $n \geq 100$ even for a highly skewed population.

8) Know that for the CLT to apply, the data needs to be a SRS or observations from a randomized experiment (eg coin tossing). If the data comes from a sample of convenience or a voluntary response sample, you can not find probabilities such as $P(\overline{x} < a)$.

9) Know that $\mu_{\overline{x}} = \mu$ and $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

10) Know how to do a forwards calculation involving $\overline{x}$.

11) Law of large numbers. Figure out the mean μ. If μ is favorable (eg stock market, number of questions likely to get right if you are a good student) larger sample sizes n are better than smaller. If μ is not favorable (eg casino gambling or guessing on a multiple choice exam) smaller sample sizes are better. See Q5 4?, HW6 D, G.

7.5 Complements

7.6 Problems

7.1. Which statement about samples is false?

a) A simple random sample of 500 Carbondale voters is as accurate as a SRS of 500 Chicago voters.
b) To determine uneployment figures, the US government uses a multistage cluster sample because a simple random sample would be too expensive.
c) Response bias is an important source of bias in sample surveys.
d) Cosmopolitan asked its married female readers whether they had committed adultery or not. 39% of the respondants said yes. A probability sample said that 26% of married women have committed adultery. The value of 26% is less likely to be near the true percentage of married women who have committed adultery than 39%.
e) A probability method uses the impartial use of chance.

7.2. Suppose x is from a highly skewed distribution with mean $\mu = 12$ and standard deviation $\sigma = 1.6$. Assume that the sample mean $\bar{x}$ is computed from a sample of size $n = 16$. Find $P(\bar{x} > 13)$.

a) 0.994 b) 0.894 c) 0.006 d) 0.106
e) can't solve, n is not large enough

7.3. Suppose x is from a normal distribution with mean $\mu = 12$ and standard deviation $\sigma = 1.6$. Assume that the sample mean $\bar{x}$ is computed from a sample of size $n = 16$. Find $P(\bar{x} > 13)$.

a) 0.994 b) 0.894 c) 0.006 d) 0.106

7.4. The population consists of the IQ scores of four children. The four scores are 110, 122, 125, and 132. The mean μ of the scores is 122.25. A random sample of size 2 will be selected without replacement. Find the sampling distribution of the sample mean $\bar{x}$.

a)

value of $\bar{x}$	110	122	125	132
probability	1/4	1/4	1/4	1/4

b)

value of $\bar{x}$	116	117.5	121	122	123.5	125
probability	1/6	1/6	1/6	1/6	1/6	1/6

c)

value of $\bar{x}$	116	117.5	121	123.5	127	128.5
probability	1/6	1/6	1/6	1/6	1/6	1/6

d)

value of $\bar{x}$	110	115	120	125	130	135
probability	1/6	1/6	1/6	1/6	1/6	1/6

e) none of the above

7.5. In the television show *Dancing with the Stars*, viewers dial a telephone number corresponding to their favorite performer. The performer with the least amount of votes is removed from the contest. What type of sample is being used?

7.6. The magazine *Cosmopolitan* asked its married female readers whether they had committed adultery or not. 39% of the respondents said yes. A probability sample said that 26% of married women have committed adultery. Is the true percentage of married women who have committed adultery more likely to be near 39% or 26%? Explain briefly.

7.7. According to a 1992 survey in the magazine *Esquire*, out of 1000 students surveyed, 10% had committed a lewd act (such as had sex in public) in their school's library. What kind of sample was used to collect this data?

7.8. Following Moore (2003, p. 192) advice columnist Ann Landers once asked her female readers if they would be content with affectionate treatment from men with no sex ever. Over 90000 women responded with 72% saying "yes." Is 72% too high, too low or about right? What type of sample is being used?

7.9. Following Moore (2000, p. xxviii), Ann Landers once asked her readers, "if you had it to do over again, would you have children?" 70% of the nearly 10,000 parents who wrote in said "No". What type of sample is being used?

7.10. According to the May 7, 2002 Daily Egyptian, a random sample of 100 students at the **Student Center** was asked about the proposed 18% tuition hike. 88 of the 100 students were opposed. Do you think that this sample was actually a simple random sample, or was it some other type of sample? Explain briefly. (Possibly Useless Hint: how would you obtain a simple random sample of students at the Student Center?)

7.11. The magazine *Time* had a web site that asked who was the most important person of the 20th century. Before a software crash, the leading vote getter was Ronnie O'Brien, an Irish soccer player. Do you think that a probability sample such as a simple random sample (SRS) was used to obtain this result? If not, what kind of sample was used. Explain briefly.

7.12. According to a February 2000 *U. Magazine* college sex survey, "Cheating happens outside of class, too. 28% of us admit that we've cheated on a past or current significant other, but 41% of us have been cheated on. Ouch!" What type of sample was used to collect this data?

7.13. Pete Rose was banned from baseball for gambling. A sports website asked its viewers whether Rose should be reinstated or not. More than 90% of the respondants said that Rose should be reinstated. What kind of sample was used and are the results likely to be accurate?

7.14. Suppose a simple random sample of 2000 of Springfield residents and a simple random sample of 1000 of Carbondale residents are taken. Both samples are to be used to estimate the proportion of unemployed. Which sample is more accurate, or is the accuracy about the same? (Hint: Springfield is more than 3 times as large as Carbondale.) Explain.

7.15. Suppose a simple random sample of 1% of Springfield residents and a simple random sample of 1% of Carbondale residents are taken. Both samples are to be used to estimate the proportion of unemployed. Which sample is more accurate, or is the accuracy about the same? (Hint: Springfield is more than 3 times as large as Carbondale. Are the sample sizes the same or is one larger?) Explain.

7.16. Television ratings are important to advertisers. Suppose that a simple random sample of 400 Carbondale residents and a simple random sample of 400 Chicago residents are taken to estimate to determine the percentage of television viewers that watch "Desperate Housewives." Chicago has roughly 100 times as many people as Carbondale. Which sample is more accurate, or are the accuracies about the same? Explain briefly.

7.17. Suppose a simple random sample of 200 University of Illinois students and a simple random sample of 400 SIU students. SIU has about 2/3 of the number of students that the University of Illinois has. Which sample is more accurate (for determining the percentage of female students) or are the accuracies about the same? Explain.

7.18. Television ratings are important to advertisers. Suppose that a simple random sample of 400 Carbondale residents and a simple random sample

of 200 Chicago residents are taken to estimate to determine the percentage of television viewers that watch the "American Idol." Chicago has roughly 100 times as many people as Carbondale. Which sample is more accurate, or are the accuracies about the same? Explain briefly.

7.19. Suppose that a student taking a multiple choice quiz has an 80% chance of answering a question correctly. Each question has options a) b) c) d) and e). Suppose that the student needs to get 75% or more of the questions correct to get a B. Would the student want the quiz to have 4 questions or 20 questions? Explain briefly.

7.20. Suppose that a gambler needs \$4000 by Friday but only has \$2000. He goes to a casino and decides to play "red and black" on the roulette wheel. The probability of winning "red and black" is about 0.47. To maximize his chance of doubling his money, should the gambler make one \$2000 bet or should he make many \$100 bets? Explain.

7.21. Suppose that the population consists of the IQ scores of four children. The four scores are 110, 122, 125, and 132. The mean μ of the scores is 122.25. A random sample of size 2 will be selected without replacement. then the sampling distribution of $\bar{x}$ can be found by completing the table below. What is the probability that $\bar{x} > 118$?

sample	110,122	110,125	110,132	122,125	122,132	125,132
value of $\bar{x}$						
probability	1/6		1/6	1/6	1/6	1/6

7.22. Suppose X is from a highly skewed distribution with mean $\mu = 12$ and standard deviation $\sigma = 1.6$. Assume that the sample mean $\overline{X}$ is computed from a sample of size $n = 16$. If possible, find $P(\overline{X} > 13)$.

Chapter 8
Probability

simulation an

8.1 Summary

1) Know that for any event A, $0 \leq P(A) \leq 1$.

2) **Probability rules:** i) $P(S) = 1$
ii) **Complement rule:** P(not A) = 1− P(A).
iii) A and B are **disjoint events** if A and B have no outcomes in common: P(A and B) = 0. Hence if A occurs, B did not occur and vice verca. If A and B are disjoint, then the **addition rule for 2 disjoint events** is P(A or B) = P(A) + P(B).

iv) Finite S. If $S = \{e_1, ..., e_k\}$ then $0 \leq P(e_i) \leq 1$, $\sum_{i=1}^{k} P(e_i) = 1$. If e_i is a sample point, then $P(A) = \sum_{i:e_i \in A} P(e_i)$. That is, $P(A)$ is the sum of the probabilities of the sample points in A. If all of the outcomes e_i are *equally likely,* then $P(e_i) = 1/k$ and
$P(A) =$ (number of outcomes in $A)/k$ if S contains k outcomes.

v) Two events A and B are **independent** if knowing that one occurs does not change the probability that the other occurs. If events are not independent, then they are dependent. Two events A and B are independent if P(A and B) = P(A) P(B). The events $A_1, ..., A_n$ are independent if knowing any subset of one to n - 1 events occurred does not change the probabilities of the other events.

vi) **Multiplication rule for 2 independent events:** If A and B are independent, then P(A and B) = P(A) P(B).

vii) **General Addition Rule:** P(A or B) = P(A) + P(B) − P(A and B). Notice that if A and B are disjoint, then P(A or B) = P(A) + P(B).
Notice that if A and B are independent, then P(A or B) = P(A) + P(B) − P(A) P(B).

viii) **Addition rule for n disjoint events:** If $A_1, ..., A_n$ are disjoint, then
$P(A_1 \text{ or } A_2 \text{ or } \cdots \text{ or } A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$. This is the probability that at least one of the n events occurs.

ix) **Multiplication rule for n independent events:** If $A_1, ..., A_n$ are independent, then $P(A_1 \text{ and } A_2 \text{ and } \cdots \text{ and } A_n) = P(A_1)P(A_2)\cdots P(A_n)$. This is the probability that all n events occur.

3) Table of probabilities with some outcomes blank. Use the fact that all of the probabilities add to 1. See Q5 4?, HW 5 C, and HW6 C.

4) Given a story problem, list the outcomes that make up an event (especially for die problems). Often you can use order to find S. Using a table to find S if two die are tossed or if a die is tossed twice and to find S if a coin is flipped 2, 3, or 4 times are typical examples. After listing all outcomes in S, use these outcomes to find $P(A)$.
5) **Toss two die** (eg red or green) (or toss a die twice with a 1st die, 2nd die). Find the probability that the sum of the two die = k. Solution: fill a

table with 36 entries and find the number of entries where the sum is equal to k. These entries lie on a diagonal. Let E_k = "sum of the dice is k". Then $P(e_k)$ = P(sum of the dice is equal to k) = (number of table entries where the sum is k)/(number of table entries). Frequently a 4, 5, or 6-sided die will be used. For a 6–sided die the number of table entries is (6)(6) = 36 and

k	2	3	4	5	6	7	8	9	10	11	12
P(sum)	1/36	2/36	3/36	4/26	5/36	6/36	5/36	4/36	3/36	2/36	1/36

where P(sum) is P(sum of two dice = k).

6) Given P(A), find P(not A). Given P(not A), find P(A). Use the complement rule: P(not A) = 1 − P(A).

7) **General Addition Rule:** P(A or B) = P(A) + P(B) − P(A and B). Notice that if A and B are disjoint, then P(A or B) = P(A) + P(B).
Notice that if A and B are independent, then P(A or B) = P(A) + P(B) − P(A) P(B). Given any three of the above probabilities, use the general additive rule to find the fourth probability. USING A PROBABILITY VENN DIAGRAM CAN BE USEFUL. The probabilities in 4 regions are P(A and not B), P(A and B), P(not A and B) and P(not A and not B). The four regions are disjoint.

8) Given P(A), P(B), and that A and B are disjoint, find P(A and B) or find
P(A and B). If A and B are disjoint, P(A and B) = 0 while P(A or B) = $P(A) + P(B)$.

9) Given P(A), P(B), and that A and B are independent, find P(A and B) or find P(A or B). If A and B are independent, P(A and B) = P(A) P(B) while
P(A or B) = $P(A) + P(B) - P(A)P(B)$.

10) Know P(x was at least k) = $P(x \geq k)$ and P(x at most k) = $P(x \leq k)$.

11) Suppose there are n independent identical trials and x counts the number of successes (outcome D) and the p = probability of success for any given trial. Then
i) P(x=0) = P(**none** of the n trials were successes) = $(1-p)^n$.
ii) $P(x \geq 1)$ = P(**at least one** of the trials was a success) = $1 - P(x = 0) = 1 - (1-p)^n$.
iii) P(x=n) = P(**all** n trials were successes) = p^n.
iv) $P(x < n)$ = P(**not all** n trials were successes) = $1 - P(x = n) = 1 - p^n$.

8.2 Complements

8.3 Problems

8.1. Let $P(A) = 0.4$ and $P(B) = 0.2$ and suppose that A and B are disjoint. Which statement is true? (Choose one answer. Write a very brief explanation by each statement.)

a) P(A and B) = 0.08.
b) P(A or B) = 0.6
c) P(A or B) = 0.08.
d) A and B are independent.
e) Need more information in order to decide whether A and B are independent.

8.2. Which of the 1st 4 statements is false, or are none of the statements false?

a) The sampling distribution of $\bar{x}$ is approximately normal if the sample size n is sufficiently large (assume the population size is at least 10 times as large as the sample size).
b) The sample proportion from a voluntary response sample is usually a good estimator of the population proportion.
c) If the probability of getting an A in Math 282 = 0.20 and the probability of getting a B is 0.25, then the probability of getting an A or B is 0.45.
d) The probability of an outcome is interpreted as the long run proportion of the time that the outcome will occur.
e) None of the four statements above is false.

8.3. Suppose events A and B are independent, P(A) = 0.5 and P(B) = 0.3. Then

a) P(A or B) = 0.15.
b) P(A or B) = 0.80.
c) P(A and B) = 0.80
d) P(A and B) = 0.15.
e) Events A and B are disjoint.

8.4. Suppose that 23% of Math 282 students receive a grade of an A. What is the probability that a randomly selected Math 282 student does not get an A in the class?

8.5. Suppose that the grade distribution x for Math 282 is given below. Find the probability that a randomly selected Math 282 student receives a D or an F?

x	A	B	C	D	F
probability	0.30	0.35	0.20	?	??

8.6. The probabilities for animals killed in steel traps meant for coyotes is shown below. Data was modified from McClave and Sincich (1991, pp. 90-92, 223). What is the probability that the steel trap killed a coyote?

filed by	Skunk	Raccoon	Opossum	Other	Coyote
probability	0.139	0.073	0.061	0.167	?

8.7. According to *The USA Today*, in 2001 Illinois, California New York and Pennsylvania are the biggest pumpkin producers. The pumpkin distribution is shown below where Other stands for the remaining states. What is the probability that a randomly selected pumpkin was grown in Illinois (IL)?

grade	IL	CA	NY	PA	Other
probability	?	0.201	0.127	0.122	0.143

8.8. The probabilities based on a 1994 survey asked females about their opinion on holiday shopping. Data is modified slightly from Johnson and Kuby (1999, p. 176). See the table given below. Find the probability that a randomly selected female would say that holiday shopping is a chore.

	"pleasure"	"chore"	"no big deal"	"nightmare"
probability	0.49	?	0.19	0.10

a) Find the probability that a randomly selected female would say that holiday shopping is a chore.

b) Find the probability that a randomly selected female would say that holiday shopping is a nightmare.

8.9. When Germany bombed London in World War II, a rumor was spread that the bombs were guided. To test this claim, intelligence divided London into 576 regions of 0.25 square kilometers each. If the bombs were hitting at random, then the number of hits would have the probabilities given below. (It turned out that the bombs were indeed falling at random.)

X=# of hits in a region	0	1	2	3	4 or more
probability	0.40	0.37	0.16	0.06	

a) What was the probability of a region getting 4 or more hits $P(X \geq 4)$?

b) What was the probability of a region getting at least two hits?

8.10. The probability of the age and gender of college students in 2003 is given below. For example, the probability that a randomly selected college student was female and 25 or older was 0.2142. Find the probability that a randomly selected college student was male. (Hint: three of the categories below are for males. Common sense says between 0.4 and 0.6.)

grade	F15-17	M15-17	F18-24	M18-24	F25up	M25up
probability	0.0053	?	0.3406	0.2823	0.2142	0.1538

8.11. The above table below represents the USA age distribution according to the 1990 Census.

age	0-19	20-39	40-59	60-79	80 and up
probability	0.29	0.33	0.21	0.14	

a) Find the probability that a randomly selected person in the United States will be 60 or over.

b) Suppose that two persons are randomly selected. Find the probability that they are both between the ages of 20 and 39.

8.12. According to a UCLA study, in a recent year steel traps meant for coyotes killed 25026 coyotes, 6348 skunks, 3345 raccoons, 2698 oppossums, 1367 porcupines, 682 beavers, and 273 dogs. An additional 5243 other types of animals were killed. A total of 44982 animals were killed by the traps. Suppose that an investigator randomly selects an animal killed by the trap. What is the probability that the animal was a skunk?

8.13. Suppose P(A) = 0.8 and P(B) = 0.1 What is P(A and B) if A and B are independent?

8.14. Suppose P(A) = 0.8, P(B) = 0.1 and that A and B are disjoint. Find
P(A or B).

8.15. Suppose P(A) = 0.5, P(B) = 0.2 and that A and B are independent. Find P(A or B).

8.16. Suppose P(A) = 0.5 and P(B) = 0.2. What is P(A or B) if A and B are disjoint?

8.17. A bowl contains ten marbles. Three are red, 2 are white, and 5 are blue. Suppose that a marble is selected at random from the bowl. What is the probability that the marble obtained is blue?

8.18. According to the 2000 Census, there are 13.8 million men and 13.2 million women in the 18 to 24 age group. If a person is randomly selected from the 18 to 24 age group, what is the probability that the selected person will be a man?

8.19. Suppose that the Math 282 class has 25 women and 11 men. What is the probability that a randomly selected person from this class will be a man?

8.20. From Brase and Brase, p. 89, according to the *Statistical Abstract of the United States*, 44% of people who received a bachelor's degree had a high school GPA higher than 3.75 (on a 4 point scale), and 27% had a high school GPA between 3.25 and 3.75 (inclusive). What is the probability that a randomly selected person from who has received a bachelor's degree had a high school GPA less than 3.25?

8.21. Suppose that a large automobile insurance company has 10,000 car insurance policies from clients who qualified as "safe drivers". From past

records, the mean annual claim from such clients is known to be $\mu = \$350$. The insurance company sells such a policy for \$375 and expects to make, on average, \$25 profit per client for a total profit of \$250,000. What fact, law, or theorem is the insurance company using in order to forecast its profit accurately?

8.22. The probabilities based on a 1994 survey asked females about their opinion on holiday shopping. The data is modified slightly from Johnson and Kuby (1999, p. 176). See the table given below. Find the probability that a randomly selected female would say that holiday shopping is a chore.

	"pleasure"	"chore"	"no big deal"	"nightmare"
probability	0.49	?	0.19	0.10

8.23. The probabilities for filing a petition for divorce in 1986 are shown below. Data is from Mendenhall and Beaver (1991, p. 90). What is the probability that the petition for divorce was filed jointly?

filed by	Wife	Husband	Jointly
probability	0.615	0.326	?

8.24. In Spring 1999 the grade distribution for Math 282 is recorded below. Other stands for withdrawals, incompletes, etc. What is the probability that a randomly selected Math 282 student received a grade of an A?

grade	A	B	C	D	F	Other
probability	?	0.282	0.113	0.070	0.014	0.085

8.25. Suppose P(A) = 0.5 and P(B) = 0.4. What is P(A and B) if A and B are independent?

8.26. Suppose P(A) = 0.5 and P(B) = 0.4. What is P(A and B) if A and B are disjoint?

8.27. Data is from Moore (2010, p. 204). In a recent year, let A be the event vehicle sold was domestic (made in North America), B be the event that the vehicle was a light truck and C = "A and B" be the event that the vehicle was a domestic light truck. Then P(A) = 0.77, P(B) = 0.52, and P(A and B) = 0.44. Find P(A or B), the probability that the vehicle was domestic or a light truck.

8.28. Suppose P(A) = 0.4, P(B) = 0.2 and P(A and B) = 0.1. Use the general addition rule to find P(A or B).

8.29. Suppose two **4 sided die** are rolled. What is the probability that the **sum** of the two die is equal to 5?

Chapter 9
Confidence Intervals and Hypothesis Testing

simulation an

9.1 The t Test and CI

9.2 Matched Pairs

9.3 Two Sample t

9.4 One Sample z for a Proportion

Example 9.1. *Cosmopolitan* magazine printed a survey inviting the reader to complete a survey about her sexual habits. 2673 women responded and 170 of the women stated that they had more than one sexual partner in the past year. Can a confidence interval be made for the proportion p of women who had more than one sexual partner in the past year? Explain.

Solution: No, a voluntary response sample was used.

9.5 Two Sample z for 2 Proportions

9.6 Inference for Regression

9.7 Chi-Squared Tests

9.8 Summary

1) The following problem is **very important.** Recognizing which of the 5 tests or CI's to use. i) one sample t,
ii) matched pairs,
iii) two sample t,
iv) one sample z for a proportion,
v) two sample z for 2 proportions.

2) A forwards calculation for $\hat{p}$ **will be on the exam**.

Step 0) $\mu_{\hat{p}} = p$ and $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$. Step i) draw line with $\hat{p}$ value(s) and p. Step ii) get zscore: z $= \frac{value - p}{\sigma_{\hat{p}}}$. Step iii) draw z–curve. Step 4 use table A to get the appropriate probability.

3) **Know how to choose the sample size** for a desired margin of error for a CI for p: use $n = \left(\frac{z*}{m}\right)^2 p^*(1-p^*)$ where p^* is a good guess of p. Use $p^* = 0.5$ if no good guess is given. Round n up. Keywords: to within 0.05 or $\pm.05$ means $m = 0.05$.

4) Know the interpretation of the CI as an interval of reasonable values. If 100 95% CI's are generated, about 95 will contain the parameter and about 5 will not.

5) Know the interpretation of $\alpha = \delta$ for tests of hypotheses. If $\alpha = \delta = 0.05 =$ P(type I error) and if 100 tests of hypotheses are performed where Ho is true for all 100 of the tests, then about 95 will fail to reject Ho, but about 5 will wrongly reject Ho.

6) Know how to compute $\hat{p} = X/n =$ (count of successes)/(sample size).

7) You may need to compute $\overline{x}$ and s in order to make a t–interval.

8) The degrees of freedom df tells you what table to use for tests, not what procedure should be used, eg σ unknown and s given make a t–procedure regardless of n.

9) Using output is important. Output makes CI's easy and steps ii) and iii) of hypothesis tests easy. Also check whether the test statistic is given.

10) You need to know when the methods can not be used, eg when the CLT does not hold and when the sample sizes for proportions are too small. Can not use the methods if the data is from a voluntary response sample or a sample of convenience.

11) For the one and two sample z-intervals for proportions and for t–intervals if df > 30, use the third line from the bottom of table C to get the cutoff z* or t* (1.645 for 90%, 1.96 for 95% or 2.576 for 99%). For t–intervals with $df \leq 30$, find the column with the level, eg 95% and the row with the df. Intersect to get t*. Eg if df $= 5$ then for a 95% CI, t* $= 2.571$.

12) Given a test statistic, **know how to find the p–value**. Know that $0 \leq$
p–value ≤ 1. **Making a sketch of the normal or t curve is a useful book keeping technique.**

A) Always use z-table A for z-tests and for t-tests (including matched pairs) if df > 30. If a t-test is used, let $to = z_o$.

i) For a right tail test (Ha $>$), pval $= 1 - P(z < z_o)$. If $z_o > 3.49$, then pval $= 0.0$, if $z_0 < -3.49$, then pval $= 1.0$.
ii) For a left tail test (Ha $<$), pval $= P(z < z_o)$. If $z_o > 3.49$, then pval $= 1.0$, if $z_0 < -3.49$, then pval $= 0.0$.
iii) For two tail (Ha $\neq$), pval $= 2(P(z < -|z_o|))$. If $z_o > 3.49$, then pval $= 0.0$, if $z_0 < -3.49$, then pval $= 0.0$.

B) Use table C to approximate p-values for t tests (including matched pairs) if df ≤ 30. Tip: if df > 5 then the pval from table A should be within 0.1 of the pval from table C.

i) For right tail, if to falls between two t* values, then the pval is between 2 "One-sided P" pvalues (eg if df $= 5$ and to $= 3.05$, then $0.01 <$ pval < 0.02). If to < 0 or if to $<$ smallest t* value (eg df $= 5$ and to $= 0.555$), then pval $> .25$ (the "One-sided P" pvalue furthest to the left). If to $>$ largest t* value (eg df $= 5$ and to $= 17.75$), then pval $= 0.0$ (less than 0.0005, the "One-sided P" pvalue furthest to the right).

ii) For left tail, if to > 0, then pval > 0.25. If to < 0 then compute $|to|$ and use (symmetry and) the rules for the right tail test: that is, if to < 0 and $|to|$ is between two t* values, then the p-value is between two "One-sided P" pvalues (eg if df $= 5$ and to $= -1.57$, then $0.05 <$ pval < 0.10). If to < 0 and $|to|$ is bigger than the largest t* value (eg df $= 5$ and to $= -44.67$), then pval $= 0.0$. If to < 0 and $|to|$ is less than the smallest t* value (eg df $= 5$ and to $= -0.17$), then pval > 0.25.

iii) For two tail, use the last line of Table C: Two-sided P. If $|to|$ is between two t* values, then the pval is between two "Two-sided P" pvalues. If $|to|$ is bigger than the largest t* value (eg df $= 5$ and $|to| = 33.79$), then pval $= 0.0$ (less than 0.001, the "Two-sided P" pvalue furthest to the right). If $|to|$ is smaller than the smallest t* value (eg df $= 5$ and $|to| = 0.37$), then pval > 0.5 (the "Two-sided P" value furthest to the left).

13) **Confidence intervals:** A confidence interval is an interval of reasonable values for the parameter and has the form estimator ± cutoff SE(estimator) (= estimator ± margin of error) where SE(estimator) is the standard error of the estimator. The cutoff t^* is obtained from table C. Use third line from the bottom of table C if the cutoff is z^* (or if df > 30). If the cutoff is z^*, then 1.645 is used for the 90% CI, 1.96 for the 95% CI and 2.576 for the 99% CI.

14) **tests of hypotheses:** All tests of hypotheses have the same 4 steps:
i) State Ho and Ha.
ii) Obtain the test statistic (possibly from output).
iii) Find the p–value (possibly from output).
iv) If the p–value $\leq \delta$, reject Ho, otherwise fail to reject Ho. Write a nontechnical sentence explaining the decision. (Use $\delta = \alpha = 0.05$ if $\delta = \alpha$ is not given.)

15) P–values are obtained from table A if they are z-tests or if they are t-tests with
df > 30. Use table C to get p–values if the test is a t-test with df < 30. Often p-values and the test statistic are given in computer output.

Use the notes given in class for more details of the following five procedures.

i) **t test and interval** for μ: ALWAYS ON EXAMS. Sample SD s is given and pop. SD σ is unknown.

The test statistic for Ho: $\mu = \mu_o$ is $t_o = \dfrac{\overline{x} - \mu_o}{s/\sqrt{n}}$

Get p–value from table C if $df = n - 1 \leq 30$ otherwise use table A.

The CI for μ is $\overline{x} \pm t^* \dfrac{s}{\sqrt{n}}$

Get t^* from table C with $df = n - 1$ if $df \leq 30$. Otherwise use $t^* = z^*$ line near the bottom of table C. The central limit theorem (CLT) should hold for $\overline{x}$. The data should be from a SRS or measurements from an experiment.

ii) **matched pairs t test and interval** for μ: The n pairs (x_i, y_i) are independent, but x_i and y_i are dependent, eg x and y are **two measurements on the same person or thing (taken at the same time or "before and after"), or on twins, or on litter mates**. Often 2 "treatments" are randomly assigned to the same individual or thing. Suppose x has mean μ_1 and y has mean μ_2. The matched pairs procedures are simply the one sample t procedures applied to the differences $d_i = x_i - y_i$. Let $\mu_d = \mu_1 - \mu_2$. Let $\overline{x}_d$ be the sample mean and let s_d be the sample standard deviation of the differences d_i. The subscript d is often not used.

The test statistic for Ho: $\mu_d = 0$ is $t_o = \dfrac{\overline{x}_d - 0}{s_d/\sqrt{n}}$.

Get p–value from table C if $df = n - 1 \leq 30$ otherwise use table A.

The CI for $\mu_d = \mu_1 - \mu_2$ is $\overline{x}_d \pm t^* \dfrac{s_d}{\sqrt{n}}$.

Get t^* from table C with $df = n-1$ if $df \leq 30$. Otherwise use $t^* = z^*$ line near the bottom of table C. The central limit theorem (CLT) should hold for $\overline{x}_d$. The data should be from a SRS of pairs or pairs of measurements from an experiment.

iii) **2 sample t test and interval** for $\mu_1 - \mu_2$:

The test statistic for Ho: $\mu_1 = \mu_2$ is $t_o = \dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$.

The CI for $\mu_1 - \mu_2$ is $(\overline{x}_1 - \overline{x}_2) \pm t^* \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$.

Use df from output or use df = smaller of $n_1 - 1$ and $n_2 - 1$. Get p–values and cutoff t^* from table C if df ≤ 30 otherwise get the p-value from table A and cutoff $t^* = z^*$ near the bottom of table C. The data should be from two independent SRS's or from measurements from an experiment on two groups (where the individuals were randomly assigned to each group). If $n_1 = n_2 \equiv n$ then the procedure can be used for $n \geq 5$ if both populations have the **same shape**. Otherwise, the CLT should hold for both $\overline{x}_1$ and $\overline{x}_2$.

iv) **z test and interval for** p:

The test statistic for Ho: $p = p_o$ is $z_o = \dfrac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}$.

The CI for p is $\hat{p} \pm z^* \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$.

Sample size $n \approx \left(\dfrac{z^*}{m}\right)^2 p^*(1-p^*)$, round up. The value p^* is a good guess for p. If no good guess is available, use $p^* = 0.5$.

Get the p-value from table A and cutoff z^* near the bottom of table C.

The data should be a proportion obtained from a SRS or an experiment. The population size should be at least ten times the sample size. For a test, need both $np_o \geq 10$ and $n(1-p_o) \geq 10$. For a CI, need both the number of successes $X = n\hat{p} \geq 15$ and the number of failures $n - X = n(1-\hat{p}) \geq 15$.

v) **z test and interval for** $p_1 - p_2$**:**

The test statistic for Ho:$p_1 = p_2$ is: $z_o = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$ where $\hat{p} = \dfrac{X_1 + X_2}{n_1 + n_2}$. Here X_i is the count of successes in sample i, $i = 1, 2$.

The CI for $p_1 - p_2$ is: $(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$.

Get the p-value from table A and cutoff z^* near the bottom of table C. The data should be proportions obtained from two ind. SRS's (or a randomized controlled experiment). The population size should be at least ten times the sample size. For a test, need both the number of successes $X_i = n_i\hat{p}_i \geq 5$

and the number of failures $n_i - X_i = n_i(1 - \hat{p}_i) \geq 5$. For a CI, need both $X_i = n_i\hat{p}_i \geq 10$ and $n_i - X_i = n_i(1 - \hat{p}_i) \geq 10$ for $i = 1, 2$.

```
Response      = Y
Coefficient Estimates
Label    Estimate Std. Error    t-value    p-value
Constant   a       not       important   for final
x          b       SE(b)    to = b/SE(b)  pvalue for
                                          Ho beta = 0
-----------------------------------------------------
Response      = brnweight
Terms         = (size)
Coefficient Estimates
Label     Coef     St. Dev      T       P
Constant  305.945   35.1814    8.696   0.0000
size      0.271373  0.009866  27.505   0.0000
```

16) Be able to find the least squares line $\hat{y} = a + bx$ from Minitab output. A typical table and a table with numbers are shown above. Then predict y for a given x.

17) Using a t-table, be able to find the 100 $(1-\delta)$ % CI for β is $b \pm t^* \, SE(b)$. If $df = n - 2 \leq 30$, get t^* from table C. If $df > 30$ use $t^* = z^*$ in table C (as usual).

18) Be able to perform the 4 step t–test of hypotheses:
i) State the hypotheses Ho: $\beta = 0$ Ha: $\beta \neq 0$.
ii) Find the test statistic $to = b/SE(b)$ from output (usually).
iii) Find the p–value from output (usually).
iv) If p–value $< \delta$, reject Ho and conclude that x is a useful linear predictor of y. If p–value $\geq \delta$, fail to reject Ho and conclude that x is not a useful linear predictor of y. Get x and y from the story problem and use $\delta = 0.05$ if δ is not given.

The p–value can also be obtained from table C (with the "Two Sided P" line) if $df = n - 2 \leq 30$: p–value $= 2P(t_{n-2} > |to|)$. Use table A if $df = n - 2 > 30$: p–value $= 2P(Z < -|to|)$. Note that "linear" is crucial. It could be that x is a very useful nonlinear predictor for y, but not a good linear predictor.

19) A 100 $(1-\delta)$% confidence interval (CI) for $\mu_y = \alpha + \beta x^*$ when $x = x^*$ is for the parameter (mean) μ_y while a 100 $(1 - \delta)$% prediction interval (PI) for a new observation y_{new} when $x = x^*$ is for the random variable y_{new}. If both intervals are given by output, know which is which. See HW13 C.

20) Suppose that there are two categorical variables: the row variable with r categories and the column variable with c categories. Know how to perform

the 4 step test:
i) Ho: there is no relationship between the two categorical variables
Ha: there is a relationship.
ii) test statistic = X^2
iii) p-value = $P(\chi^2_{(r-1)(c-1)} > X^2)$.
iv) Reject Ho if the p–value $\leq \delta$, and conclude that there is a relationship between the two categorical variables. If the p-value $> \delta$, fail to reject Ho and conclude that there is no relationship between the two variables.

Sometimes X^2 is given by output but sometimes you need to compute the expected count and the chisquare contribution. Recall that the expected cell count = (row total)(column total)/(table total). The chisquare cell contribution = $(O - E)^2/E$ where O and E are the observed and expected cell counts. The expected cell count and the cell chisquare contribution need to be computed for each of the rc cells. Finally, X^2 is the sum of all rc cell chisquare contributions.

Sometimes the p–value is given by output but sometimes it needs to be obtained from table E. The df = (r-1)(c-1). Since this test is always a right tail test, find the two values in the df row of table E that are closest to X^2. Then the p–value is between the values on the top row of the table. For example, if df = 5 and $X^2 = 13.00$ then 12.83 and 13.39 bracket X^2 and $0.02 < pvalue < 0.025$. If X^2 is big and way off table E, then p-value < 0.0005. For example, if df = 5 and $X^2 = 57$, then p–value = 0. If X^2 is small and way off table E, then p-value > 0.25. For example, if df = 5 and $X^2 = 4.33$, then p–value > 0.25.

21) Know the difference between an observational study and an experiment.

22) In this class, double blinded completely randomized controlled (comparative) experiments are best. Next best are single blinded completely randomized controlled (comparative) experiments and completely randomized controlled (comparative) experiments are still very good. Observational studies are ok.

Experiments that are controlled (comparative) but not randomized and experiments that (are not comparative) have a treatment group but no control group are bad (analogous to voluntary response samples and samples of convenience).

Know that randomization is the most important step in an experiment. Randomization washes out the effects of lurking variables, makes the treatment group like the control group except for the treatment, and allows one to find valid confidence intervals and two sample tests of hypotheses.

23) If you are given the results of observational studies and completely randomized experiments and the two results differ, then conclude that the results from the completely randomized experiment are correct.

25) Know how to use the random numbers to divide n individuals into a treatment group and a control group.

9.9 Complements

9.10 Problems

9.1. 4 of the studies listed below were observational studies. Which study was a randomized controlled experiment? Circle an answer and explain your choice.

a) a study on a vaccine for AIDS
b) a study of the effect of cocaine on the heart.
c) a study of the effect of alcohol on the liver
d) a study on factors that cause people to murder
e) a study on sexual behavior of Americans

9.2. The Sept. 26, 1998 Star Tribune reported that the first experimental vaccine for infants against pneumococcus bacteria (causes miningitis) was highly effective. If the design was good this study was

a) an observation study. b) a controlled but not randomized controlled experiment.
c) a double blinded placebo study. d) a randomized controlled experiment.
e) a double blinded observational study.

9.3. In a double blinded randomized controlled experiment that uses a placebo, the most important step is

a) double blinding
b) using a placebo
c) using randomization to assign subjects to treatment or control
d) eliminating bias by controlling for all confounding factors
e) none of the above.

9.4. Suppose that 40% of business majors at US universities are women and that a simple random sample of 100 business majors is selected. Let $\hat{p}$ be the proportion of women in the sample. Find the mean and standard deviation of the sampling distribution of $\hat{p}$. (Hint: $\mu_{\hat{p}} = p$, $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$.)

9.5. A corn soy blend was developed in 1996 to be a highly nutritious cheap food. Data is from Moore and McCabe (1999, p. 507). Suppose that the amounts of vitamin C were obtained for a simple random sample of size $n = 8$ taken from a production run. Assume that the sample mean $\overline{x} = 22.50$ and the sample standard deviation $s = 7.19$, and that the distribution of x is approximately normal. Find a 95% confidence interval for the mean vitamin C content during the production run, if possible.

9.6. A corn soy blend was developed in 1996 to be a highly nutritious cheap food. Data is from Moore and McCabe (1999, p. 507). Suppose that the amounts of vitamin C were obtained for a simple random sample of size $n = 8$ taken from a production run. Assume that the sample mean $\overline{x} = 22.50$ and the sample standard deviation $s = 7.19$, and that the distribution of x is highly skewed. Find a 95% confidence interval for the mean vitamin C content during the production run, if possible.

9.7. Suppose that 80 90% confidence intervals are made for the proportion of SIU students that will graduate in the Spring. About how many confidence intervals will contain the true proportion p?

9.8. Following Mendenhall and Beaver (p. 329), a manufacturer of gunpowder has developed a new powder that is designed to produce a muzzle velocity of 3000 feet per second. Eight shells are loaded with charge and their muzzle velocities are measured. Assume that the central limit theorem holds, that $\overline{x} = 2958.75$ and that the sample SD $s = 39.26$. Find a 95% confidence interval for the mean muzzle velocity μ.

9.9. Suppose that it is desired is desired to predict the total income of a bank from its total assets (in billions of dollars). The output below uses data from the 20 largest banks in 1973.

```
Coefficient Estimates
Label      Estimate  Std. Error  t-value  p-value
constant   -0.59     14.48        -0.04    0.968
assets      5.840     1.363        4.285   0.001
```

a) Predict the *income* if *assets* = 7.5.

b) Find a 99% confidence interval for β.

c) Do a 4 step test for $\beta \neq 0$.

9.10. Suppose that it is desired to predict a person's *height* in mm from their *height while kneeling* $= x$. The data was obtained from $n = 112$ people.

```
Coefficient Estimates
Label     Estimate  Std. Error  t-value  p-value
Constant  4.14159   53.2161      0.078   0.9381
x         1.34740    0.04282    31.470   0.0000
```

a) Predict *height* if *height while kneeling* = 1240.

b) Find a 90% confidence interval for β.

c) Do a 4 step test for $\beta \neq 0$.

9.11. The Sept. 26, 1998 *Star Tribune* reported that the first vaccine for young children against the pneumococcus bacteria (causes blood poisoning, meningitis) was tested. 19000 children got the vaccine (treatment group) and 19000 other children got a placebo (a shot of a saline solution). None of the

children who got the vaccine developed pneumococcal diseases while 22 of the other children did. **How** did researchers **decide** which of the 38000 children were placed in the treatment group and which were placed in the control group? **Did the vaccine help** (explain briefly)?

9.12. According to an April 2002 ABC nightly news report, six observational studies suggested that taking Estrogen after menopause decreased the risk for heart disease in women. Two randomized controlled experiments suggested that taking Estrogen after menopause increased the risk for heart disease in women. Should women take Estrogen after menopause? Explain briefly.

9.13. Knee surgery is frequently performed to relieve chronic knee pain from osteoarthritis. In a single blinded randomized controlled experiment, the knee surgery was performed on half of the patients. The other half of the patients were put to sleep and received incisions in their knees, but no surgery (this was the placebo). In this study, both groups received about the same amount of pain relief. In four other studies all of the patients received the knee surgery and the investigators concluded that the surgery does provide relief from pain. Should people suffering from chronic knee pain undergo the surgery? Explain briefly.

9.14. Following Freedman (2005, p. 4), in the 1960s there were several studies on whether mammography (screening women for breast cancer by x–rays) could speed up detection of breast cancer early enough to matter. Some studies that were not randomized controlled experiments were negative: they said mammography did not help. Then the first large scale randomized controlled study offered 31000 women the treatment: 4 rounds of annual screening (clinical exam and mammography) and another 31000 women were offered the usual health care (control group). Death rates after 5 years were were 1.3 per 1000 for the treatment group and 2.0 per 1000 for the control group. **How** did researchers **decide** which of the 62000 women were placed in the treatment group and which were placed in the control group? **Did mammography help** (explain briefly)?

9.15. DES was given to pregnant woman to prevent miscarriage. Eight controlled experiments were performed to study the effectiveness of DES. The women in the treatment group received DES while the women in the control group were given a placebo. Three studies were randomized controlled experiments while five studies did not use randomization. The rate of miscarriage was about the same for all eight of the treatment groups. For the randomized controlled studies the rate of miscarriage in the control groups was about the same as that of the treatment groups. But for the five studies that did not use randomization, the control group rate was much higher than the treatment group rate. Does the drug DES help prevent miscarriage? Explain.

9.16. An investigator believes that a low calorie high nutrition diet will increase the lifetime of primates. The investigator has 10 baby rhesus monkeys. She will give five monkeys the low calorie high nutrition diet and five

of the monkeys the standard diet. Explain or demonstrate how to assign the monkeys to the two types of diet.

9.17. (Modifying Moore, p. 219 slightly.) Sellers of the supplement gingko extract gave a group of 400 senior citizens a daily pill of gingko extract, and the study reported improved memory and concentration. Another study used 230 healthy people over 60 years old. Half of the subjects were in the treatment group that received a daily pill of gingko extract and half of the subjects were in the control group that received a placebo pill. This study found no evidence of improved memory and concentration after six weeks. **How** did researchers **decide** which of the 230 people were placed in the treatment group and which were placed in the control group? **Does gingko extract improve memory and concentration** (explain briefly)?

9.18. A physiology student from a 2000 Math 282 class believed that caffeine would increase the amount of potassium in the urine. Five students measured the amount of potassium in their urine after taking a caffeine pill (μ_1) and before taking the pill (μ_2). Assume that the appropriate procedure can be used. The output above is the data collected from the five students. What type of test should be used (one sample t, matched pairs, or 2 sample t)? Explain briefly.

```
mu = mud = mu1 - mu2
test            alternative    T-value p-value
matched pairs:       mu < 0    0.35    0.635
```

9.19. 21 subjects worked a paper maze wearing a scented or unscented mask. Each subject worked the maze twice, one time with each mask (in random order). Let μ_1 be the mean time to work the maze wearing the unscented mask and let μ_2 be the mean time wearing the scented mask. Using the output above, test whether using the unscented mask decreased the time to do the maze.

9.20. Roughly 30% of hunters issued a deer permit in Illinois harvest a deer. It is desired to estimate the proportion of hunters in Southern Illinois with a deer permit who harvest a deer. How large a sample must you test in order to estimate the proportion of Southern Illinois hunters who harvest a deer within ± 0.05 with 95% confidence?

9.21. A corn soy blend was developed in 1996 to be a highly nutritious cheap food. Suppose that the amounts of vitamin C were obtained for a simple random sample of size $n = 8$ taken from a production run. If the sample mean $\bar{x} = 22.50$ and the sample standard deviation $s = 7.19$, find a 95% confidence interval for the mean vitamin C content during the production run. You may assume that the data comes from a normal distribution.

9.22. At Northern Illinois University, several hundred students take the statistics course. Assume that 28 students were selected at random and their scores from exam 1 and exam 2 were entered into the computer. Below is some computer output from the data. Perform a 4 step test for whether the

mean score of the first midterm was lower than the mean score of the second midterm. Assume that the appropriate test can be used.

two sample t-test for $H_o : \mu_1 = \mu_2$ vs $H_a : \mu_1 < \mu_2$, T = -1.15, p-value = 0.13

two sample t-test for $H_o : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$, T = -1.15, p-value = 0.26

matched pairs test for $H_o : \mu_1 = \mu_2$ vs $H_a : \mu_1 < \mu_2$, T = -1.28, p-value = 0.04

matched pairs test for $H_o : \mu_1 = \mu_2$ vs $H_a : \mu_1 > \mu_2$, T = -1.28, p-value = 0.96

9.23. A SRS of male college students and a SRS of female college students were taken. The students were asked questions. From these questions it was determined that the number of frequent binge drinkers was 1630 for men and 1684 for women. It is desired to know whether the proportion for males is greater than the proportion for females.

a) Find the sample proportion of frequent binge drinkers for men.

b) If $SE(\hat{p_1} - \hat{p}_2) = 0.00622$, find a 95% confidence interval for $p_1 - p_2$.

c) What is the value of the pooled proportion $\hat{p}$.

d) What are the hypotheses for the test?

e) Suppose the test statistic is equal to 9.34. What is the p-value for the test?

f) What is the conclusion for the test?

9.24. Makers of generic drugs must **show that the generic drug does not differ significantly from the reference drug.** Let μ_1 be the mean absorption of the reference drug and let μ_2 be the mean absorption of the generic drug. The output for 3 possible tests is shown below. Suppose that **two** simple random samples of 20 patients were taken. The 1st 20 get the reference drug and the 2nd 20 get the generic drug. The data is from a randomized experiment so the appropriate procedure can be used.

```
matched pairs:    t = -0.14 pvalue=0.89
95% CI for mu ref - mu gen: (-538,472)
2 sample t   :    t = -0.14 pvalue=0.89
95% CI for mu ref - mu gen: (-524,457)
```

a) Which procedure should be used: matched pairs, or 2 sample t. **Explain.**

b) Do a 4 step test of hypotheses.

c) Give a 95% confidence interval for the difference in mean absorptions.

9.25. Makers of generic drugs must **show that the generic drug does not differ significantly from the reference drug.** Let μ_1 be the mean

absorption of the reference drug and let μ_2 be the mean absorption of the generic drug. The output for 3 possible tests is shown below. Suppose 20 subjects were used to test the drugs. Subjects 1 to 10 received the generic drug first and one week later received the reference drug. Subjects 11 to 20 received the reference drug 1st then the generic drug one week later. The data is from a randomized experiment so the appropriate procedure can be used.

```
matched pairs:    t = -0.14 pvalue=0.89
95% CI for mu ref - mu gen: (-538,472)
2 sample t   :    t = -0.14 pvalue=0.89
95% CI for mu ref - mu gen: (-524,457)
```

a) Which procedure should be used: matched pairs, 2 sample t. **Explain.**

b) Do a 4 step test of hypotheses.

c) Give a 95% confidence interval for the difference in mean absorptions.

9.26. ACT college prep course claims to improve ACT scores for students. To test the claim, a simple random sample of 14 identical twins was taken. One twin took the prep course while the other twin did not. Let μ_1 be the mean ACT score of students who take the prep course and let μ_2 be the mean score of students who do not take the test. Output gave $t = 3.012$ and pvalue $= 0.005$.

a) What test procedure should be used. Circle the correct answer and **give a brief explanation.**

i) One sample Z test
ii) One sample t test
iii) two sample t test
iv) matched pairs t test
v) pooled two sample t test

b) Assume that the assumptions for the appropriate test hold. Use a 4 step test to decide if the ACT prep course improves ACT scores.

9.27. A SRS of 15 patients with high blood pressure was taken. Let μ_1 be the mean systolic bood pressure before taking medicine and let μ_2 be the mean systolic bood pressure after taking the medicine. The investigator wanted to know whether the medicine helped reduce the blood pressure: is $\mu_1 > \mu_2$? Let $\mu_d = \mu_1 - \mu_2$.

```
test           alternative     T    p-value 95% CI
                                            for  mu1 - mu2:
matched pairs: mu_d not = 0   8.12  0.0000  (13.93, 23.93)
matched pairs: mu_d     > 0   8.12  0.0000  (13.93, 23.93)
matched pairs: mu_d     < 0   8.12  1.000   (13.93, 23.93)
2 sample t   : mu1 not = mu2  2.56  0.0166  (3.7, 34.1)
2 sample t   :     mu1 > mu2  2.56  0.0083  (3.7, 34.1)
2 sample t   :     mu1 < mu2  2.56  0.9917  (3.7, 34.1)
```

a) Which procedure should be used? **Explain briefly.**

b) Do a 4 step test of hypotheses.

c) Give a 95% confidence interval for the difference in mean blood pressure levels.

test	df	T	p-value
matched pairs	4	12.83	0.00
equal variances	8	0.57	0.58
unequal variances	7	0.57	0.58

Ho: $\mu_1 - \mu_2 = 0$ where μ_1 is mean wear from tire brand A. If appropriate, $\mu_D = \mu_1 - \mu_2$.

9.28. Suppose that a manufacturer desires to compare the wearing qualities of two brands of tires, A and B. For comparison, a tire of type A and a tire of type B are randomly assigned and mounted on the rear wheels of each of five cars. The cars are then operated for a specified number of miles, and the amount of wear is recorded for each tire. Assume that the conditions for the appropriate test hold.

a) Which test should be used? Explain.

b) Test whether the mean amount of wear from brand A is different from the mean amount of wear from brand B.

9.29. Makers of generic drugs must **show that the generic drug does not differ significantly from the reference drug.** Let μ_1 be the mean absorption of the reference drug and let μ_2 be the mean absorption of the generic drug. The output is shown below. Suppose a SRS of 20 subjects received the generic drug and a SRS of 20 subjects received the reference drug.

test	alternative	T	p-value	95% CI for mu1 - mu2:
matched pairs:	mu_d not = 0	-0.14	0.89	(-538,472)
matched pairs:	mu_d > 0	-0.14	0.555	(-538,472)
matched pairs:	mu_d < 0	-0.14	0.445	(-538,472)
2 sample t :	mu1 not = mu2	-0.14	0.89	(-524,457)
2 sample t :	mu1 > mu2	-0.14	0.555	(-524,457)
2 sample t :	mu1 < mu2	-0.14	0.445	(-524,457)

a) Which procedure should be used? **Explain briefly.**

b) Do a 4 step test of hypotheses.

c) Give a 95% confidence interval for the difference in mean absorption.

9.30. 21 subjects worked a paper maze wearing a scented or unscented mask. Each subject worked the maze twice, one time with each mask (in random order). Let μ_1 be the mean time to work the maze wearing the

unscented mask and let μ_2 be the mean time wearing the scented mask. Investigators wanted to know whether using the unscented mask increased the time to do the maze. Let $\mu_d = \mu_1 - \mu_2$ be written as $mud = mu1 - mu2$.

```
test            alternative    T     p-value 95% CI
                                                 for mud:
matched pairs:      not = 0    0.35  0.730  (-4.76,6.67)
matched pairs:          > 0    0.35  0.365  (-4.76,6.67)
matched pairs:          < 0    0.35  0.635  (-4.76,6.67)
2 sample t    : mu1 not = mu2  0.22  0.82   (-7.7, 9.6)
2 sample t    :     mu1 > mu2  0.22  0.41   (-7.7, 9.6)
2 sample t    :     mu1 < mu2  0.22  0.59   (-7.7, 9.6)
```

a) Which procedure should be used: matched pairs or 2 sample t? **Explain briefly.**

b) Do a 4 step test of hypotheses.

c) Give a 95% confidence interval for the difference in mean times.

9.31. A simple random sample of 3420 students attending private schools showed that 917 were smokers. A SRS of 5131 students attending public schools showed that 1503 were smokers. Let p_1 be the proportion of private school students that smoke and let p_2 be the proportion of public school students that smoke. It is desired to test if p_1 is less than p_2. Assume that the test statistic is -2.33 and the pvalue is 0.0099.

a) Find the sample proportion $\hat{p}_1$ of private students that smoke.

b) State the hypotheses and the test statistic.

c) Find the pvalue and give the conclusion.

quality of life		Canada	USA	row total
much better	observed	75	541	616
	expected	(77.37)	(538.63)	
	cell chisq	[0.0726]	[0.0104]	
somewhat better	observed	71	498	569
	expected	(71.47)	( )	
	cell chisq	[0.0031]	[ ]	
about the same	observed	96	779	875
	expected	(109.91)	(765.09)	
	cell chisq	[ ]	[0.2529]	
somewhat worse	observed	50	282	332
	expected	(41.70)	(290.30)	
	cell chisq	[1.6520]	[0.2373]	
much worse	observed	19	65	84
	expected	(10.55)	(73.45)	
	cell chisq	[6.7680]	[0.9721]	
column total		311	2165	2476

9.32. Data is from Moore (p. 557). Suppose that a simple random sample of 311 Canadians and a SRS of 2165 USA citizens who have had a heart attack are taken. One year after the heart attack, each person is asked about the current quality of their life relative to what it had been before the heart attack. The output is shown above.

Perform a 4 step test for whether quality of life and country are related. Be sure to show the degrees of freedom and how table E is used.

9.33. Suppose that a simple random sample of 386 student from 4 majors (accounting, administration, economics, and finance) was taken at a big university. Each student was classified acoording to major and gender. From the output below, give a four step test for an appropriate set of hypotheses.

```
                Female     Male        Total
Acct.              68       56          124
                 72.28    51.72

Admin              91       40          131
                 76.36    54.64

Econ.               5        6           11
                  6.41     4.59

Finance            61       59          120
                 69.95    50.05

Total             225      161          386

ChiSq = 0.253 + 0.354 +
```

```
2.807 + 3.923 +
0.311 + 0.434 +
1.145 + 1.600 = 10.827
```

Chapter 10
Classification and Regression Trees

simulation an

10.1 Summary

10.2 Complements

10.3 Problems

Beaton, A.E., Martin, M.O., Mullis, I.V.S., Gonzales, E.J., Smith, T.A., and Kelly, D.L. (1996), *Science Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study,* TIMSS International Study Center, Chestnut Hill, MA.

Becker, R.A., Chambers, J.M., and Wilks, A.R. (1988), *The New S Language: a Programming Environment for Data Analysis and Graphics,* Wadsworth and Brooks/Cole, Pacific Grove, CA.

Brase and Brase *Understandable Statistics: Concepts and Methods*

Brase and Brase *Understanding Basic Statistics*

Buxton, L.H.D. (1920), "The Anthropology of Cyprus," *The Journal of the Royal Anthropological Institute of Great Britain and Ireland,* 50, 183-235.

Carrell, S.E., and West, J.E. (2010), "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors," *Journal of Political Economy*, University of Chicago Press, 118, 409-432.

Cleveland, W.S. (2001), "Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics," *International Statistical Review / Revue Internationale de Statistique*, 69, 21-26.

Cook, R.D., and Forzani, L. (2018), "Big Data and Partial Least Squares Prediction," *The Canadian Journal of Statistics*, 46, 62-78.

Freedman, D.A. (2005), *Statistical Models Theory and Practice,* Cambridge University Press, New York, NY.

Freedman, D., Pisani, R., and Purves, R. (1978), *Statistics*, W.W. Norton, New York, NY.

Gould and Ryan *Introductory Statistics: Exploring the world Through Data*,

Gould and Ryan *Essential Statistics*,

Hicks, L., and Wattenberg, B.J. (2000), *The First Measured Century: An Illustrated Guide to Trends in America, 1900-2000*, American Enterprise Institute Press, available from (www.pbs.org/fmc/book.htm).

Johnson, R. and Kuby, P. (1999), *Just the Essentials of Elementary Statstics*, Brooks/Cole, Pacific Grove, CA.

Lewis, K. (199?), "Seven, and Wiser, Virgins", from "The Best of Teaching Statistics,"

McKenzie, J.D., and Goldman, R. (1999), *The Student Edition of MINITAB,* Addison Wesley Longman, Reading, MA.

Mendenhall, W., and Beaver, R.J. (1991), *Introduction to Probability and Statistics*.

Moore, D.S. (1995, 2000, 2003, 2007, 2010), *The Basic Practice of Statistics,* 1st, 2nd, 3rd, 4th, and 5th ed., W.H. Freeman, New York, NY.

Moore, D.S. (2010), *Essential Statistics,* W.H. Freeman, New York, NY.

Moore, D.S., and McCabe, G.P. (1999), *Introduction to the Practice of Statistics,* 3rd ed., W.H. Freeman, New York, NY.

Moore, D.S., Notz, W.I., and Fligner, M.A. (2018), *The Basic Practice of Statistics,* 8th ed., W.H. Freeman, New York, NY.

Price, P.C. (2006), "Are You as Good a Teacher as You Think?", *Thought & Action*, fall, 7-14.

R Core Team (2016), "R: a Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

Rossman, A.J., and Chance, B.L. (2011), *Workshop Statistics: Discovery with Data*, 4th ed., Wiley, Hoboken, NJ.

Somers, C.H. (2000), "The War Aginst the Boys," *The Atlantic Monthly*, May, 59-74.

Stavig, V. (2004), "Bread and Peace," *Minnesota,* January-February, 38-40.

Student, J. (1998), "No Sex, Please... We're College Graduates," *American Demographics,* 20, 18.

Sullivan, M. (2006), *Statistics: Informed Decisions Using Data*, 2nd ed., Prentice Hall, Englewood Cliffs, NJ.

Sullivan, M. (2014), *Fundamentals of Statistics: Informed Decisions Using Data*, 4th ed., Pearson, Boston, MA.

Tanur, J.M., Mosteller, F., Kruskal, W.H., Lehmann, E.L., Link, R.F., Pieters, R.S., and Rising, G.R. (eds.) (1989), *Statistics: a Guide to the Unknown*, 3rd ed., Wadsworth & Brooks/Cole, Pacific Grove, CA.

Verzani, J. (2014), *Using R for Introductory Statistics*, 2nd ed., Chapman & Hall/CRC, Boca Raton, FL.

Von Hoffman, C. (2000), "Survey Says ...," *The Standard*.

Wilcox, R.R. (2017), *Understanding and Applying Basic Statistical Methods Using R*, Wiley, Hoboken, NJ.

Index