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”?