Introduction to Statistics

Benjamin Yakir, The Hebrew University

January 2, 2010

Chapter 2

Descriptive Statistics

(Based on the CS book, Chapter 2. Descriptive Statistics)

2.1 Student Learning Objectives

Once you have collected data, what will you do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample often is overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that you will learn to describe data. Your agent might also provide you with a graph of the data.

In this chapter, you will study numerical and graphical ways to describe and display your data. This area of statistics is called "Descriptive Statistics". You will learn to calculate, and even more importantly, to interpret these measurements and graphs.

By the end of this chapter, the student should be able to:

- Display data graphically and interpret graphs: histograms and box plots.
- Recognize, describe, and calculate the measures of location of data: quartiles and percentiles.
- Recognize, describe, and calculate the measures of the center of data: mean, median, and mode.
- Recognize, describe, and calculate the measures of the spread of data: variance, standard deviation, and range.

2.2 Displaying Data

A statistical graph is a tool that helps you learn about the shape or distribution of a sample. The graph can be a more effective way of presenting data than a mass of numbers because we can see where data clusters and where there are

only a few data values. Newspapers and the Internet use graphs to show trends and to enable readers to compare facts and figures quickly.

Statisticians often graph data first in order to get a picture of the data. Then, more formal tools may be applied.

Some of the types of graphs that are used to summarize and organize data are the dot plot, the bar chart, the histogram, the stem-and-leaf plot, the frequency polygon (a type of broken line graph), pie charts, and the box plot. Our emphasis will be on histograms and box plots. All these plots are tightly linked to the notion of *frequency* of the data. Hence, we initiate our discussion by considering this notion.

2.3 Frequency

Twenty students were asked how many hours they worked per day. Their responses, in hours, are listed below:

5; 6; 3; 3; 2; 4; 7; 5; 2; 3; 5; 6; 5; 4; 4; 3; 5; 2; 5; 3.

Let us create an "R" object of the name "work.hours" that contain these data:

```
> work.hours <- c(5,6,3,3,2,4,7,5,2,3,5,6,5,4,4,3,5,2,5,3)
```

Next, let us create a table that summarizes the different values of working hours and the frequency in which these values appear in the data:

```
> table(work.hours)
work.hours
2 3 4 5 6 7
3 5 3 6 2 1
```

Recall that function "table" takes as input a vector of data and produces as output the frequencies of the different values.

We may have a clearer understanding of the meaning of the output of the function "table" if we presented outcome as a frequency listing the different data values in ascending order and their frequencies. For that end we may apply the function "as.data.frame" to the outcome of the given function an obtain:

```
> as.data.frame(table(work.hours))
  work.hours Freq
2          2    3
3          3    5
4          4    3
5          5    6
6          6    2
7          7    1
```

A frequency is the number of times a given datum occurs in a data set. According to the table above, there are three students who work 2 hours, five students who work 3 hours, etc. The total of the frequency column, 20, represents the total number of students included in the sample.

The function "as.data.frame" transforms its input into a data frame, which is the standard way of storing statistical data. We will introduce data frames in the next section.

A relative frequency is the fraction of times an answer occurs. To find the relative frequencies, divide each frequency by the total number of students in the sample - in this case, 20. Relative frequencies can be written as fractions, percents, or decimals.

As an illustration let us compute the relative frequencies in our data:

```
> freq <- table(work.hours)
> freq
work.hours
2 3 4 5 6 7
3 5 3 6 2 1
> sum(freq)
[1] 20
> freq/sum(freq)
work.hours
   2    3    4    5    6    7
0.15 0.25 0.15 0.30 0.10 0.05
```

We stored the stored the frequencies in an object called "`freq`". The content of the object are the frequencies 3, 5, 3, 6, 2 and 1. The function "`sum`" sums the components of its input. Since the sum of the frequencies is the total number of students that responded to the survey, which is 20. Hence, when we apply the function "`sum`" to the object "`freq`" we get 20 as an output.

The outcome of dividing an object by a number is a division by the number of each element in the object. Therefore, when we divide "`freq`" by "`sum(freq)`" (the number 20) we get a vector of relative frequencies. The first entry to this vector is $3/20 = 0.15$, the second entry is $5/20 = 0.25$, and the last entry is $1/20 = 0.05$. The sum of the relative frequencies is equal to 1:

```
> sum(freq/sum(freq))
[1] 1
```

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current value.

```
> cumsum(freq/sum(freq))
   2    3    4    5    6    7
0.15 0.40 0.55 0.85 0.95 1.00
```

Observe that the cumulative relative frequency of the smallest value 2 is the frequency of that value (0.15). The cumulative relative frequency of the second value 3 is the sum of the relative frequency of the smaller value (0.15) and the relative frequency of the current value (0.25), which produces a total of $0.15 + 0.25 = 0.40$. Likewise, for the third value 4 we get a cumulative relative frequency of $0.15 + 0.25 + 0.5 = 0.55$ The last entry of the cumulative relative frequency column is one, indicating that one hundred percent of the data has been accumulated.

The computation of the cumulative relative frequency was carried out with the aid of the function "`cumsum`". This function take as an input a numerical vector and produces as output a numerical vector of the same length with the cumulative sums of the components of the input vector.

Glossary

Frequency: The number of times a value of the data occurs.

Relative Frequency: The ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes.

Cumulative Relative Frequency: The term applies to an ordered set of observations from smallest to largest. The Cumulative Relative Frequency is the sum of the relative frequencies for all values that are less than or equal to the given value.

2.4 Reading Data into R

In the examples so far the size of the data set was very small and we were able to input directly the data into R. In more practical settings the data sets that will be required to be analyzed will be much larger and it will be very inefficient to try and enter them manually. In this section we will learn how to upload data from a file in the Comma Separated Values (CSV) format.

The file "ex1.csv" contains data on the sex and height of 100 individuals. This file is given in the CSV format. We will use the R function "read.csv" in order to upload the content of this file to R. Run the following line of code in the Rweb window:

```
> ex.1 <- read.csv(file="ex1.csv")
```

Observe that this code will probably not work on a local copy of R since the file "ex1.csv" is not recognized by your local copy. Below we give explanation how to read CSV files into local copies of R. For the time being, however, let us work on the internet version of the program Rweb.

Consider the content of R object "ex.1":

```
> ex.1
         id    sex height
1   5696379 FEMALE   182
2   3019088   MALE   168
3   2038883   MALE   172
4   1920587 FEMALE   154
5   6006813   MALE   174
6   4055945 FEMALE   176
.       .      .      .
.       .      .      .
.       .      .      .
98  9383288   MALE   195
99  1582961 FEMALE   129
100 9805356   MALE   172
>
```

(Noticed that we erased the middle rows. In the R window you should obtain the full table.) The object "ex.1" is a data frame. Data frames are the standard tabular format of storing statistical data. The columns of the table are called *variables* and corresponds to measurements. In this example the three variables are:

id: The ID of each subject. A unique identifying number with 7 digits.

sex: The sex of each subject. The values are either `Male` or `FEMALE`.

height: The height (in centimeter) of each subject. A numerical value.

When the values of the variable are numerical we say that it is a quantitative variable. On the other hand, if the variable has qualitative or level values we say that it is a factor. In the given example, `sex` is a factor and `height` is a quantitative variable.

The rows of the table are called *observations* and correspond to the subjects. In this data set there are 100 subjects, with subject number 1, for example, being a female of height 182 cm and ID number 5696379. Subject number 98, on the other hand, is a male of height 195 cm and ID number 9383288.

The function "`read.csv`" takes as input the address of a CSV file and produces as output a data frame object with the content of the file.

In the example above the file "`ex1.csv`" is located in the working directory of `R` that runs on the university's servers. In such a case the address of the file is its name. However, if the file is located in a different directory or a remote location then the full address of the file should be supplied. For example, a copy of the file `ex1.csv` can be found on the internet in the URL `http://temp/temp/ex1.csv` (**FIX**). Running the function "`read.csv`" as before, but with the argument `file="http://temp/temp/ex1.csv"` will read into `R` the content of that copy.

Alternatively, one may save a copy of a file locally in some directory and change the working directory of the local installation of `R` to the directory where the file was stored. This can be done (in the Windows version) by selecting the option "`File/Change Dir...`" in the upper toolbar of the `R` window and browse to the target directory. After setting the directory that contains the CSV file as the working directory one use the argument `file="ex1.csv"` as in the original demonstration.

One may browse and edit CSV files and create new files with the aid of standard electronic spreadsheet programs such as Microsoft's Excel or OpenOffice's Calc. Opening a CSV file by the spreadsheet program will produce a spreadsheet with the content of the file. Values in the cells of the spreadsheet may be modified directly. When saving, one should pay attention to save the file in the CVS format. Similarly, new CSV files may be created entering the data in an empty sheet. The first row should include the name of the variable, preferably as a single character string with no empty spaces. The following rows contain the observations. When saving, the spreadsheet should be saved in the CSV format (use the "Save by name" option).

Glossary

Data Frame: A tabular format for storing statistical data. Columns correspond to variables and rows correspond to observations.

Variable: A measurement that may be carried out over a collection of subjects. The outcome of the measurement may be numerical, which produces a quantitative variable; or it may be qualitative, in which case a factor is produced.

Observation: The evaluation of a variable (or variables) for a given subject.

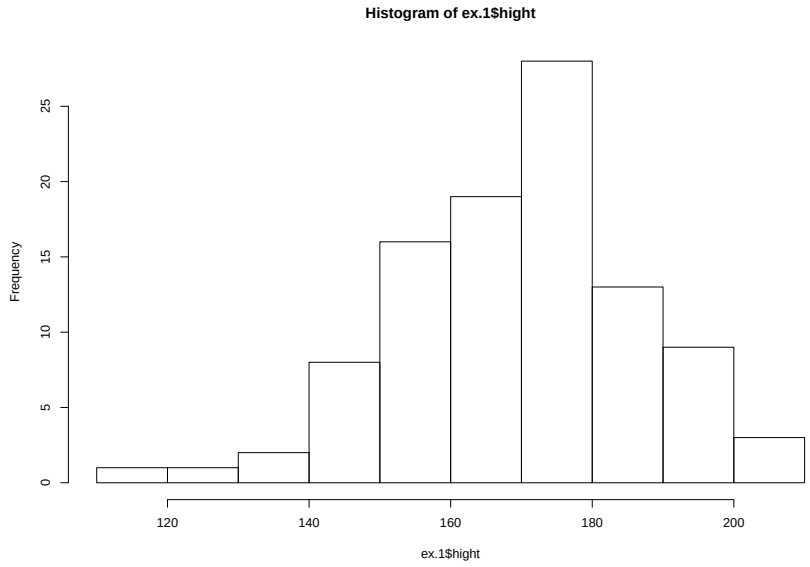

Figure 2.1: Histogram of Height

CSV Files: A digital format for storing data frames.

2.5 Histograms

For most of the work you do in this book, you will use a histogram to display the data. One advantage of a histogram is that it can readily display large data sets. A rule of thumb is to use a histogram when the data set consists of 100 values or more.

One may obtain an histogram in `R` by the application of the function "`hist`" to a vector of numerical data. Let us create an histogram of the height in `ex.1` data frame:

```
> hist(ex.1$height)
```

The outcome of the function is a plot that apears in the graphical window and is presented in Figure 2.1.

A histogram consists of contiguous boxes. It has both a horizontal axis and a vertical axis. The horizontal axis is labeled with what the data represents (the height, in this example). The vertical axis is labeled either "Frequency". The histogram can give you the shape of the data, the center, and the spread of the data. (The next section tells you how to calculate the center and the spread.)

The histogram is constructed by dividing the range of the data (the x-axis) into equal intervals, which are the bases for the boxes. The height of each box represents the count of the number of observations that fall within the interval. For example, consider the box with the base between 160 and 170. There is a total of 19 subjects with height larger that 160 but no more than 170 (i.e., $160 <$ `height` ≤ 170). Consequently, the height of that box is 19.

The input to the function "`hist`" should be a vector of numerical values. Notice the structure of the input that we used in order to construct the histogram of the variable `height` in the `ex.1` data frame. In general, one may address the variable `variable.name` in the data frame `dataframe.name` using the format: "`dataframe.name$variable.name`". Indeed, when we type the expression "`ex.1$height`" we get as an output the values of the variable `height` from the given data frame:

```
> ex.1$height
  [1] 182 168 172 154 174 176 193 156 157 186 143 182 194 187 171
 [16] 178 157 156 172 157 171 164 142 140 202 176 165 176 175 170
 [31] 169 153 169 158 208 185 157 147 160 173 164 182 175 165 194
 [46] 178 178 186 165 180 174 169 173 199 163 160 172 177 165 205
 [61] 193 158 180 167 165 183 171 191 191 152 148 176 155 156 177
 [76] 180 186 167 174 171 148 153 136 199 161 150 181 166 147 168
 [91] 188 170 189 117 174 187 141 195 129 172
```

2.6 Box Plots

Box plots or box-whisker plots give a good graphical image of the concentration of the data. They also show how far from most of the data the extreme values are. In principle, the box plot is constructed from five values: the smallest value, the first quartile, the median, the third quartile, and the largest value. The median, the first quartile, and the third quartile will be discussed here, and then again in the section on measuring data in this chapter. We use these values to compare how close other data values are to them.

The median, a number, is a way of measuring the "center" of the data. You can think of the median as the "middle value," although it does not actually have to be one of the observed values. It is a number that separates ordered data into halves. Half the values are the same number or smaller than the median and half the values are the same number or larger. For example, consider the following data:

1; 11.5; 6; 7.2; 4; 8; 9; 10; 6.8; 8.3; 2; 2; 10; 1

Ordered from smallest to largest:

1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

The median is between the 7th value, 6.8, and the 8th value 7.2. To find the median, add the two values together and divide by 2:

$$\frac{6.8 + 7.2}{2} = 7$$

The median is 7. Half of the values are smaller than 7 and half of the values are larger than 7.

Quartiles are numbers that separate the data into quarters. Quartiles may or may not be part of the data. To find the quartiles, first find the median or second quartile. The first quartile is the middle value of the lower half of the

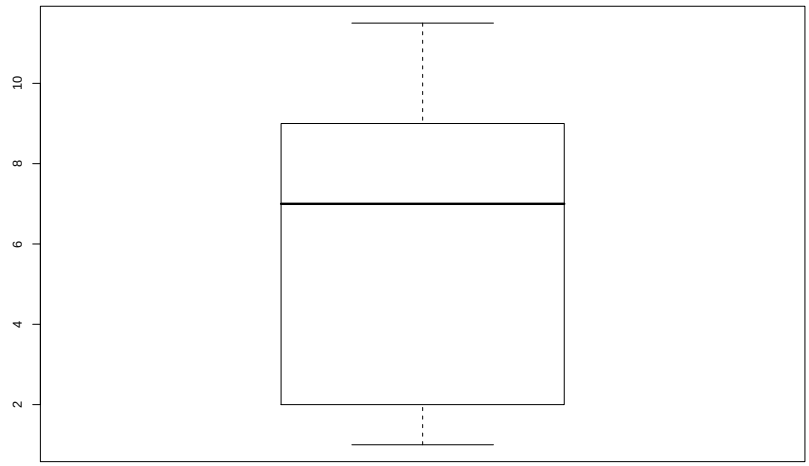

Figure 2.2: Box Plot of the Example

data and the third quartile is the middle value of the upper half of the data. To get the idea, consider the same data set shown above:

1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

The median or second quartile is 7. The lower half of the data is 1, 1, 2, 2, 4, 6, 6.8. The middle value of the lower half is 2.

1; 1; 2; 2; 4; 6; 6.8

The number 2, which is part of the data, is the first quartile. One-fourth of the values are the same or less than 2 and three-fourths of the values are more than 2.

The upper half of the data is 7.2, 8, 8.3, 9, 10, 10, 11.5. The middle value of the upper half is 9.

7.2; 8; 8.3; 9; 10; 10; 11.5

The number 9, which is part of the data, is the third quartile. Three-fourths of the values are less than 9 and one-fourth of the values are more than 9.

Outliers are values that do not fit with the rest of the data. Data points with values that are much to large or much too small in comparison to the vast majority of the observations will be identified as outliers. In the context of the construction of a box plot we identify potential outliers with the help of the inter-quantile range (IQR). The inter-quantile range is the distance between the third quantile (Q3) and the first quantile (Q1), i.e., IQR = Q3 - Q1. A data point that is larger than the third quantile plus 1.5 times the inter-quantile range will be marked as a potential outlier. Likewise, a data point smaller than the first quantile minus 1.5 times the inter-quantile range will also be so marked. Potential outliers always need further investigation.

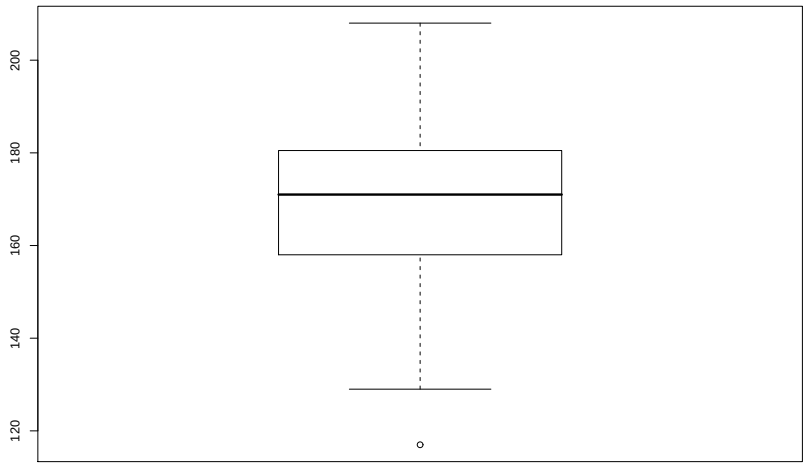

Figure 2.3: Box Plot of Height

In the running example we obtained an inter-quantile range of size 9-2=7. The upper threshold for defining an outlier is $9 + 1.5 \times 7 = 19.5$ and the lower threshold is $2 - 1.5 \times 7 = -8.5$. All data points are within the two thresholds, hence there are no outliers in this data.

To construct a box plot, use a vertical rectangular box and two vertical "whiskers" that extend from the ends of the box to the smallest and largest data values that are not outliers. Outlier values, if any exist, are marked as points above or blow the endpoints of the whiskers. The smallest and largest non-outlier data values label the endpoints of the axis. The first quartile marks one end of the box and the third quartile marks the other end of the box. The middle fifty percent of the data fall inside the box. The box plot gives a good quick picture of the data.

One may produce a box plot with the aid of the function "`box plot`". The input to the function is a vector of numerical values and the output is a plot. As an example, let us produce the box plot of the 14 data points that were used as an illustration:

```
> box plot(c(1,11.5,6,7.2,4,8,9,10,6.8,8.3,2,2,10,1))
```

The resulting box plot is presented in Figure 2.2. Observe that the end points of the whiskers are 1, for the minimal value, and 11.5 for the largest value. The end values of the box are 9 for the third quantile and 2 for the first quantile. The median 6.5 is marked inside the box.

Next, let examine the box plot for the height data:

```
> box plot(ex.1$height)
```

The resulting box plot is presented in Figure 2.3. In order to assess the plot let us compute quantiles of the variable:

```
> summary(ex.1$height)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  117.0   158.0   171.0   170.1   180.2   208.0
```

The function `summary`, when applied to a numerical vector, produce the minimal and maximal entries, as well the first, third and second quantiles (the latter is the Median). It also computes the average of the numbers (the Mean), which will be discussed later.

Let us compare the results with the plot in Figure 2.3. Observe that the median 171 coincides with the thick horizontal line inside the box and that the lower end of the box coincides with first quantile 158 and the upper end with 180.2, which is the third quantile. The inter-quartile range is $180.2 - 158.0 = 22.2$. The upper threshold is $180.2 + 1.5 \times 22.2 = 213.5$. This threshold is larger than the largest observation (208.0). Hence, the largest observation is not an outlier and it marks the end of the upper whisker. The lower threshold is $158.0 - 1.5 \times 22.2 = 124.7$. The minimal observation (117.0) is less then this threshold. Hence it is an outlier and it is mark as a point below the end of the lower whisker. The second smallest observation is 129. It lies above the lower threshold and it marks the end point of the lower whisker.

Glossary

Median: A number that separates ordered data into halves: half the values are the same number or smaller than the median and half the values are the same number or larger than the median. The median may or may not be part of the data.

Quartiles: The numbers that separate the data into quarters. Quartiles may or may not be part of the data. The second quartile is the median of the data.

Outlier: An observation that does not fit the rest of the data.

Interquartile Range (IQR) : The distance between the third quartile (Q3) and the first quartile (Q1). IQR = Q3 - Q1.

2.7 Measures of the Center of the Data

The "center" of a data set is a way of describing location. The two most widely used measures of the "center" of the data are the mean (average) and the median. To calculate the mean weight of 50 people, add the 50 weights together and divide by 50. To find the median weight of the 50 people, order the data and find the number that splits the data into two equal parts (previously discussed under box plots in this chapter). The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the center.

The mean can also be calculated by multiplying each distinct value by its relative frequency and then summing across all values. The letter used to represent the sample mean is an x with a bar over it (pronounced "x bar"): $\bar{x}$.

The Greek letter μ (pronounced "mew") represents the population mean. If you take a truly random sample, the sample mean is a good estimate of the population mean.

To see that both ways of calculating the mean are the same, consider the sample:

$$1; 1; 1; 2; 2; 3; 4; 4; 4; 4; 4$$

In the first way of calculating the mean we get:

$$\bar{x} = \frac{1+1+1+2+2+3+4+4+4+4+4}{11} = 2.7 \, .$$

Alternatively, we may note that the distinct values in the sample are 1, 2, 3, and 4 with relative frequencies of 3/11, 2/11, 1/11 and 5/11, respectively. The alternative method of computation produces:

$$\bar{x} = 1 \times \frac{3}{11} + 2 \times \frac{2}{11} + 3 \times \frac{1}{11} + 4 \times \frac{5}{11} = 2.7 \, .$$

You can quickly find the location of the median by using the expression $(n+1)/2$. The letter n is the total number of data values in the sample. If n is an odd number, the median is the middle value of the ordered data (ordered smallest to largest). If n is an even number, the median is equal to the two middle values added together and divided by 2 after the data has been ordered.

For example, if the total number of data values is 97, then $(n+1)/2 = (97+1)/2 = 49$. The median is the 49th value in the ordered data. If the total number of data values is 100, then $(n+1)/2 = (100+1)/2 = 50.5$ The median occurs midway between the 50th and 51st values.

2.7.1 The Law of Large Numbers and the Mean

The Law of Large Numbers says that if you take samples of larger and larger size from any population, then the mean $\bar{x}$ of the sample gets closer and closer to the population mean μ. This is discussed in more detail in the section (**REF**) The Central Limit Theorem of this course.

2.7.2 Skewness, the Mean and the Median

Consider the following data set:

$$4,\ 5,\ 6,\ 6,\ 6,\ 7,\ 7,\ 7,\ 7,\ 7,\ 7,\ 8,\ 8,\ 8,\ 9,\ 10$$

This data produces the upper most histogram in Figure 2.4. Each interval has width one and each value is located in the middle of an interval. The histogram displays a symmetrical distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other.

Let us compute the mean and the median of this data:

```
> x <- c(4,5,6,6,6,7,7,7,7,7,7,8,8,8,9,10)
> mean(x)
[1] 7
> median(x)
[1] 7
```

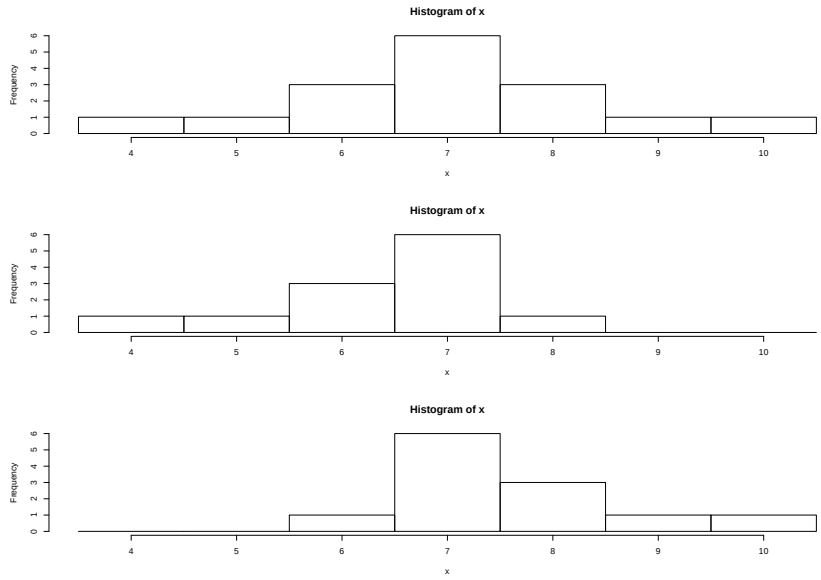

Figure 2.4: Three Histograms

The mean and the median are each 7 for these data. In a perfectly symmetrical distribution, the mean and the median are the same.

The histogram for the data:

$$4,\ 5,\ 6,\ 6,\ 6,\ 7,\ 7,\ 7,\ 7,\ 7,\ 7,\ 8$$

is not symmetrical and is displayed in the middle of Figure 2.4. The right-hand side seems "chopped off" compared to the left side. The shape of the distribution is called skewed to the left because it is pulled out to the left.

Let us compute the mean and the median for this data:

```
> x <- c(4,5,6,6,6,7,7,7,7,7,7,8)
> mean(x)
[1] 6.416667
> median(x)
[1] 7
```

A histogram that is skewed to the left. The median is still 7, but the mean is less than 7. The relation between the mean and the median reflects the skewing. (Notice that the original data is replaced by the new data when object `x` is reassigned.)

Consider yet another set of data:

$$6,\ 7,\ 7,\ 7,\ 7,\ 7,\ 7,\ 8,\ 8,\ 8,\ 9,\ 10$$

The histogram for the data is also not symmetrical and is displayed at the bottom of Figure 2.4. Notice that it is skewed to the right. Compute the mean and the median:

```
> x <- c(6,7,7,7,7,7,7,8,8,8,9,10)
```

```
> mean(x)
[1] 7.583333
> median(x)
[1] 7
```

The median is still 7, but this time the mean is greater than 7. Again, the mean reflects the skewing.

To summarize, generally if the distribution of data is skewed to the left the mean is less than the median. If the distribution of data is skewed to the right than the median is less than the mean. Finally, let us examine the data on the height:

```
> mean(ex.1$height)
[1] 170.11
> median(ex.1$height)
[1] 171
```

Observe that the histogram of the height (Figure 2.1) is skewed to the left. This is consistent with the fact that the mean is less than the median.

Skewness and symmetry are important in the context of the Central Limit Theorem that will be discussed in later chapters.

Glossary

Mean: A number that measures the central tendency. A common name for mean is 'average.' The term 'mean' is a shortened form of 'arithmetic mean.' By definition, the mean for a sample (denoted by $\bar{x}$) is

$$\bar{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$$

and the mean for a population (denoted by μ) is

$$\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}} .$$

Median: A number that separates ordered data into halves. Half the values are the same number or smaller than the median and half the values are the same number or larger than the median. The median may or may not be part of the data.

2.8 Measures of the Spread of the Data

The most common measure of spread is the standard deviation. The standard deviation is a number that measures how far data values are from their mean. For example, if the mean of a set of data containing 7 is 5 and the standard deviation is 2, then the value 7 is one (1) standard deviation from its mean because $5 + 1 \times 2 = 7$.

The number line may help you understand standard deviation. If we were to put 5 and 7 on a number line, 7 is to the right of 5. We say, then, that 7 is one standard deviation to the right of 5. If 1 were also part of the data set, then 1 is two standard deviations to the left of 5 because $5 - 2 \times 2 = 1$.

In general, for a mean $\bar{x}$ and a standard deviation s we may use the formula:

$$\text{value} = \bar{x} + (\#\text{ofSTDEVs}) \times s \ ,$$

where #ofSTDEVs = the number of standard deviations.

If x is a value and $\bar{x}$ is the sample mean, then $x - \bar{x}$ is called a deviation. In a data set, there are as many deviations as there are data values. Deviations are used to calculate the sample standard deviation.

Calculation of the Sample Standard Deviation

To calculate the standard deviation, calculate the variance first. The variance is the average of the squares of the deviations. The standard deviation is the square root of the variance. You can think of the standard deviation as a special average of the deviations (the $x - \bar{x}$ values). The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma) represents the population standard deviation. We use s^2 to represent the sample variance and σ^2 to represent the population variance. If the sample has the same characteristics as the population, then s should be a good estimate of σ.

Consider the following example: In a fifth grade class, the teacher was interested in the average age and the standard deviation of the ages of her students. What follows are the ages of her students to the nearest half year:

9, 9.5, 9.5, 10, 10, 10, 10, 10.5, 10.5, 10.5, 10.5, 11, 11, 11, 11, 11, 11, 11.5, 11.5, 11.5

We are interested in the computation of the standard deviation for these data. First, let us create an object `x` that contains the data:

```
> x <- c(9,9.5,9.5,10,10,10,10,10.5,10.5,10.5,10.5,11,11,11,11,11,
+ 11,11.5,11.5,11.5)
> length(x)
[1] 20
```

The function "`length`" returns the length of the input vector. Notice that we have a total of 20 data points.

The next step involves the computation of the deviations:

```
> x.bar <- mean(x)
> x.bar
[1] 10.525
> x - x.bar
 [1] -1.525 -1.025 -1.025 -0.525 -0.525 -0.525 -0.525 -0.025
 [9] -0.025 -0.025 -0.025  0.475  0.475  0.475  0.475  0.475
[17]  0.475  0.975  0.975  0.975
```

The average of the observations is equal to 10.525 and when we delete this number from each of the components of the vector `x` we obtain the deviations. For example, the first deviation is obtained as 9 - 10.525 = -1.525, the second deviation is 9.5 - 10.525 = -1.025, and so forth. The 20th deviation is 11.5 - 10.525 = 0.975, and this is the last one.

From a more technical point of view observe the expression that computed the deviations, "`x - x.bar`", involved the deletion of a single value (`bar.x`) from a vector with 20 values (`x`). The expression resulted in the deletion of the value from each component of the vector. This is an example of the general way by which `R` operates on vector. The typical behavior of `R` is to apply an operation to each component of the vector. As yet another illustration of this property consider the computation of the squares of the deviations:

```
> (x - x.bar)^2
 [1] 2.325625 1.050625 1.050625 0.275625 0.275625 0.275625
 [7] 0.275625 0.000625 0.000625 0.000625 0.000625 0.225625
[13] 0.225625 0.225625 0.225625 0.225625 0.225625 0.950625
[19] 0.950625 0.950625
```

Recall that "`x - x.bar`" is a vector of length 20. We apply the square function to this vector. This function is applied to each of the components of the vector. Indeed, for the first component we have that $(-1.525)^2 = 2.325625$, for the second component $(-1.025)^2 = 1.050625$, and for the last component $(0.975)^2 = 0.950625$.

For the sample variance we sum the square of the deviations and divide by the total number of data values minus one $(n - 1)$. The standard deviation is obtained by taking the square root of the variance:

```
> sum((x - x.bar)^2)/(length(x)-1)
[1] 0.5125
> sqrt(sum((x - x.bar)^2)/(length(x)-1))
[1] 0.715891
```

The function "`var`" computes the sample variance and the function "`sd`" computes the standard deviations. The input to both functions is the vector of data values and the outputs are the sample variance and the standard deviation, respectively:

```
> var(x)
[1] 0.5125
> sd(x)
[1] 0.715891
```

Explanation of the computation of the variance and the standard deviation

The deviations show how spread out the data are about the mean. The value 9 is farther from the mean than 9.5. The deviations -1.525 and -1.025 indicate that. If you add the deviations, the sum is always zero. (For this example, there are 20 deviations.) So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers. The variance, then, is the average squared deviation. It is small if the values are close to the mean and large if the values are far from the mean.

The variance is a squared measure and does not have the same units as the data. Taking the square root solves the problem. The standard deviation measures the spread in the same units as the data.

For the sample variance, we divide by the total number of data values minus one $(n-1)$. Why not divide by n? The answer has to do with the population variance. The sample variance is an estimate of the population variance. By dividing by $(n-1)$, we get a better estimate of the population variance.

Your concentration should be on what the standard deviation does, not on the arithmetic. The standard deviation is a number which measures how far the data are spread from the mean. Let a calculator or computer do the arithmetic.

The sample standard deviation, s, is either zero or larger than zero. When $s = 0$, there is no spread. When s is a lot larger than zero, the data values are very spread out about the mean. Outliers can make s very large.

The standard deviation, when first presented, can seem unclear. By graphing your data, you can get a better "feel" for the deviations and the standard deviation. You will find that in symmetrical distributions, the standard deviation can be very helpful but in skewed distributions, the standard deviation may not be much help. The reason is that the two sides of a skewed distribution have different spreads. In a skewed distribution, it is better to look at the first quartile, the median, the third quartile, the smallest value, and the largest value. Because numbers can be confusing, always graph your data.

Glossary

Standard Deviation: A number that is equal to the square root of the variance and measures how far data values are from their mean. Notation: s for sample standard deviation and ??for population standard deviation.

Variance: Mean of the squared deviations from the mean. Square of the standard deviation. For a set of data, a deviation can be represented as $x - \bar{x}$ where x is a value of the data and $\bar{x}$ is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and 1.

2.9 Summary of Formulas

2.9.1 Commonly Used Symbols

- The symbol $\sum$ means to add or to find the sum.
- n = the number of data values in a sample.
- N = the number of people, things, etc. in the population.
- $\bar{x}$ = the sample mean.
- s = the sample standard deviation.
- μ = the population mean.
- σ = the population standard deviation.
- f = frequency.
- f/n = relative frequency.
- x = numerical value.

2.9.2 Commonly Used Expressions

- $x \times (f_x/n)$ = A value multiplied by its respective relative frequency.
- $\sum_{i=1}^{n} x_i$ or $\sum_{i=1}^{N} x_i$ = The sum of the values
- $\sum_x (x \times f_x/n)$= The sum of values multiplied by their respective relative frequencies.
- $(x - \bar{x})$ or $(x - \mu)$ = Deviations from the mean (how far a value is from the mean).
- $(x - \bar{x})^2$ or $(x - \mu)^2$ = Deviations squared.

2.9.3 Mean Formulas:

- $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \sum_x \left(x \times (f_x/n)\right)$
- $\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$

2.9.4 Standard Deviation Formulas:

- $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$
- $\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2}$

2.9.5 Formulas Relating a Value, the Mean, and the Standard Deviation:

- value = mean + (#ofSTDEVs)(standard deviation), where #ofSTDEVs = the number of standard deviations
- $x = \bar{x}$ + (#ofSTDEVs)(s).
- $x = \mu$ + (#ofSTDEVs)(σ).