Chapter 3

Descriptive Statistics

3.1 Student Learning Objectives

This chapter deals with numerical and graphical ways to describe and display data. This area of statistics is called *descriptive statistics*. You will learn to calculate and interpret these measures and graphs. By the end of this chapter, you should be able to:

- Use histograms and box plots in order to display data graphically.
- Calculate measures of central location: mean and median.
- Calculate measures of the spread: variance, standard deviation, and interquartile range.
- Identify outliers, which are values that do not fit the rest of the distribution.

3.2 Displaying Data

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 may ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample is often overwhelming. A better way may 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.

A statistical graph is a tool that helps you learn about the shape of the 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 start the analysis by graphing the data in order to get an overall picture of it. Afterwards, more formal tools may be applied.

In the previous chapters we used the bar plot, where bars that indicate the frequencies in the data of values are placed over these values. In this chapter

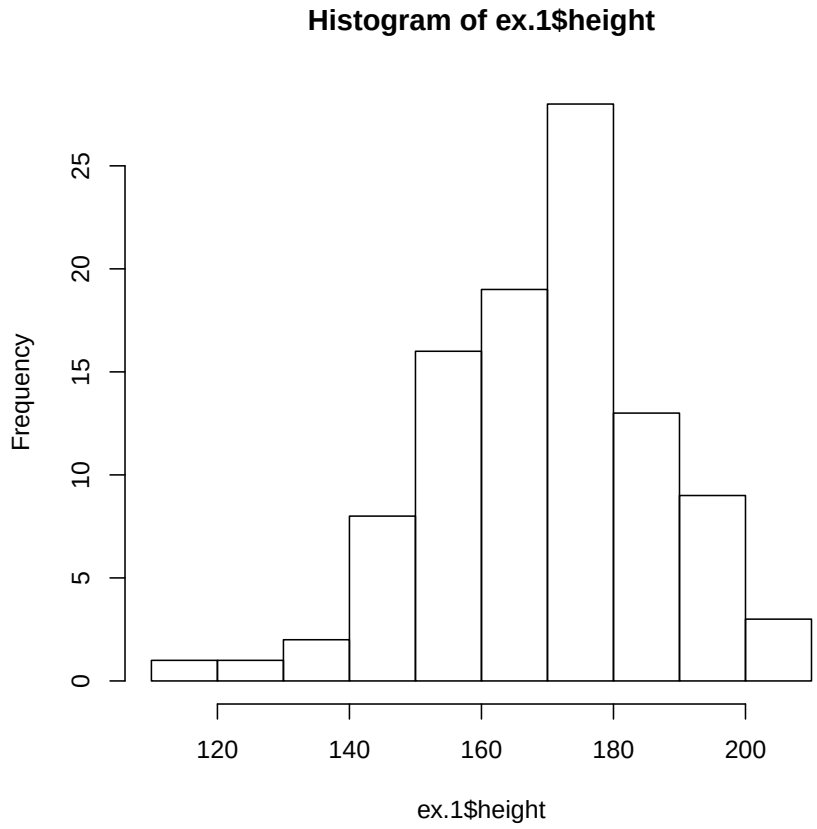

Figure 3.1: Histogram of Height

our emphasis will be on histograms and box plots, which are other types of plots. Some of the other types of graphs that are frequently used, but will not be discussed in this book, are the stem-and-leaf plot, the frequency polygon (a type of broken line graph) and the pie charts. The types of plots that will be discussed and the types that will not are all tightly linked to the notion of *frequency* of the data that was introduced in Chapter 2 and intend to give a graphical representation of this notion.

3.2.1 Histograms

The *histogram* is a frequently used method for displaying the distribution of continuous numerical data. An 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 produce a histogram in `R` by the application of the function "`hist`" to a sequence of numerical data. Let us read into `R` the data frame "`ex.1`" that contains data on the sex and height and create a histogram of the heights:

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

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

The outcome of the function is a plot that apears in the graphical window and is presented in Figure 3.1.

The data set, which is the content of the CSV file "ex1.csv", was used in Chapter 2 in order to demonstrate the reading of data that is stored in a external file into R. The first line of the above script reads in the data from "ex1.csv" into a data frame object named "ex.1" that maintains the data internally in R. The second line of the script produces the histogram. We will discuss below the code associated with this second line.

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 presents frequencies and is labeled "Frequency". By the examination of the histogram one can appreciate the shape of the data, the center, and the spread of the data.

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 (that is, $160 < \texttt{height} \leq 170$). Consequently, the height of that box[1] is 19.

The input to the function "hist" should be a sequence of numerical values. In principle, one may use the function "c" to produce a sequence of data and apply the histogram plotting function to the output of the sequence producing function. However, in the current case we have already the data stored in the data frame "ex.1", all we need to learn is how to extract that data so it can be used as input to the function "hist" that plots the histogram.

Notice the structure of the input that we have used in order to construct the histogram of the variable "height" in the "ex.1" data frame. 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
```

This is a numeric sequence and can serve as the input to a function that expects a numeric sequence as input, a function such as "hist". (But also other functions, for example, "sum" and "cumsum".)

[1] In some books an histogram is introduced as a form of a density. In densities the *area* of the box represents the frequency or the relative frequency. In the current example the height would have been $19/10 = 1.9$ if the area of the box would have represented the frequency and it would have been $(19/100)/10 = 0.019$ if the area of the box would have represented the relative frequency. However, in this book we follow the default of R in which the height represents the frequency.

There are 100 observations in the variable "`ex.1$height`". So many observations cannot be displayed on the screen on one line. Consequently, the sequence of the data is wrapped and displayed over several lines. Notice that the square brackets on the left hand side of each line indicate the position in the sequence of the first value on that line. Hence, the number on the first line is "`[1]`". The number on the second line is "`[16]`", since the second line starts with the 16th observation in the display given in the book. Notice, that numbers in the square brackets on your `R Console` window may be different, depending on the setting of the display on your computer.

3.2.2 Box Plots

The *box plot*, or box-whisker plot, gives a good graphical overall impression of the concentration of the data. It also shows 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 once more in the next section.

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 size or smaller than the median and half the values are the same size or larger than it. For example, consider the following data that contains 14 values:

$$1,\ 11.5,\ 6,\ 7.2,\ 4,\ 8,\ 9,\ 10,\ 6.8,\ 8.3,\ 2,\ 2,\ 10,\ 1\ .$$

Ordered, from smallest to largest, we get:

$$1,\ 1,\ 2,\ 2,\ 4,\ 6,\ 6.8,\ 7.2,\ 8,\ 8.3,\ 9,\ 10,\ 10,\ 11.5\ .$$

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 data and the third quartile is the middle value of the upper half of the data. For illustration consider the same data set from 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. The number 2, which is part of the data in this case, is the first quartile which is denoted Q1. One-fourth of the values are the same or less than 2 and three-fourths of the values are more than 2.

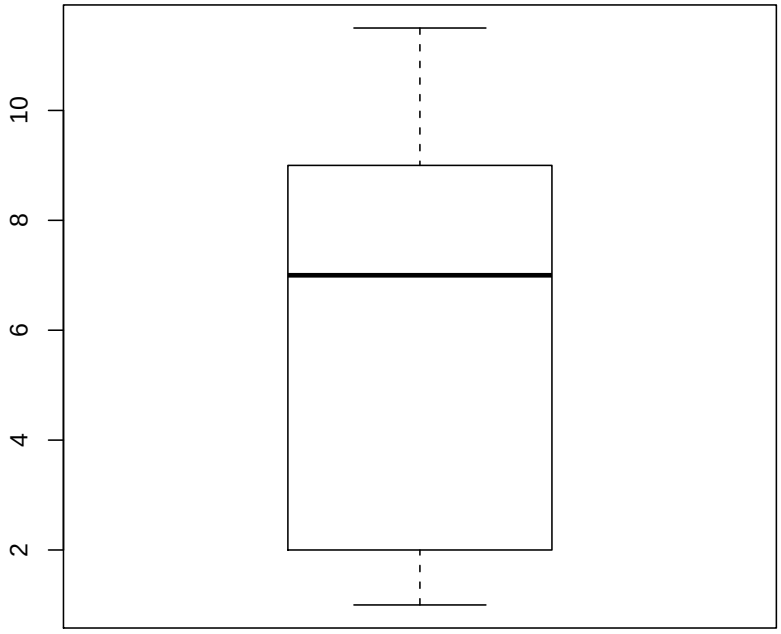

Figure 3.2: Box Plot of the Example

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. The number 9 is the third quartile which is denoted Q3. Three-fourths of the values are less than 9 and one-fourth of the values[2] are more than 9.

Outliers are values that do not fit with the rest of the data and lie outside of the normal range. Data points with values that are much too 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-quartile range* (IQR). The inter-quartile range is the distance between the third quartile (Q3) and the first quartile (Q1), i.e., $\text{IQR} = \text{Q3} - \text{Q1}$. A data point that is larger than the third quartile plus 1.5 times the inter-quartile range will be marked as a potential outlier. Likewise, a data point smaller than the first quartile minus 1.5 times the inter-quartile

[2]The actual computation in `R` of the first quartile and the third quartile may vary slightly from the description given here, depending on the exact structure of the data.

range will also be so marked. Outliers may have a substantial effect on the outcome of statistical analysis, therefore it is important that one is alerted to the presence of outliers.

In the running example we obtained an inter-quartile 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.

In the construction of a box plot one uses 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 central 50% of the data fall within the box.

One may produce a box plot with the aid of the function "`boxplot`". The input to the function is a sequence 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:

```
> boxplot(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 3.2. 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 quartile and 2 for the first quartile. The median 7 is marked inside the box.

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

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

The resulting box plot is presented in Figure 3.3. In order to assess the plot let us compute quartiles 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 sequence, produce the minimal and maximal entries, as well the first, second and third quartiles (the second is the Median). It also computes the average of the numbers (the Mean), which will be discussed in the next section.

Let us compare the results with the plot in Figure 3.3. 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 quartile 158.0 and the upper end with 180.2, which is the third quartile. 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 than this threshold. Hence it is an outlier and it is marked 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.

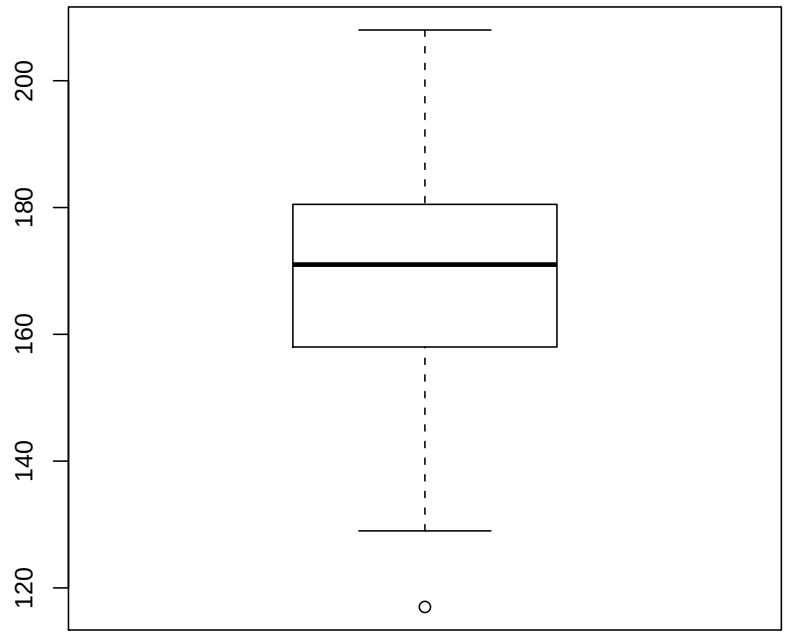

Figure 3.3: Box Plot of Height

3.3 Measures of the Center of Data

The two most widely used measures of the central location of the data are the *mean* (average) and the *median*. To calculate the average weight of 50 people one should add together the 50 weights and divide the result by 50. To find the median weight of the same 50 people, one may order the data and locate a number that splits the data into two equal parts. 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. Nonetheless, the mean is the most commonly used measure of the center.

We shall use small Latin letters such as x to mark the sequence of data. In such a case we may mark the sample mean by placing a bar over the x: $\bar{x}$ (pronounced "x bar").

The mean can be calculated by averaging the data points or it also can be calculated with the relative frequencies of the values that are present in the data. In the latter case one multiplies each distinct value by its relative frequency and then sum the products across all values. To see that both ways of calculating

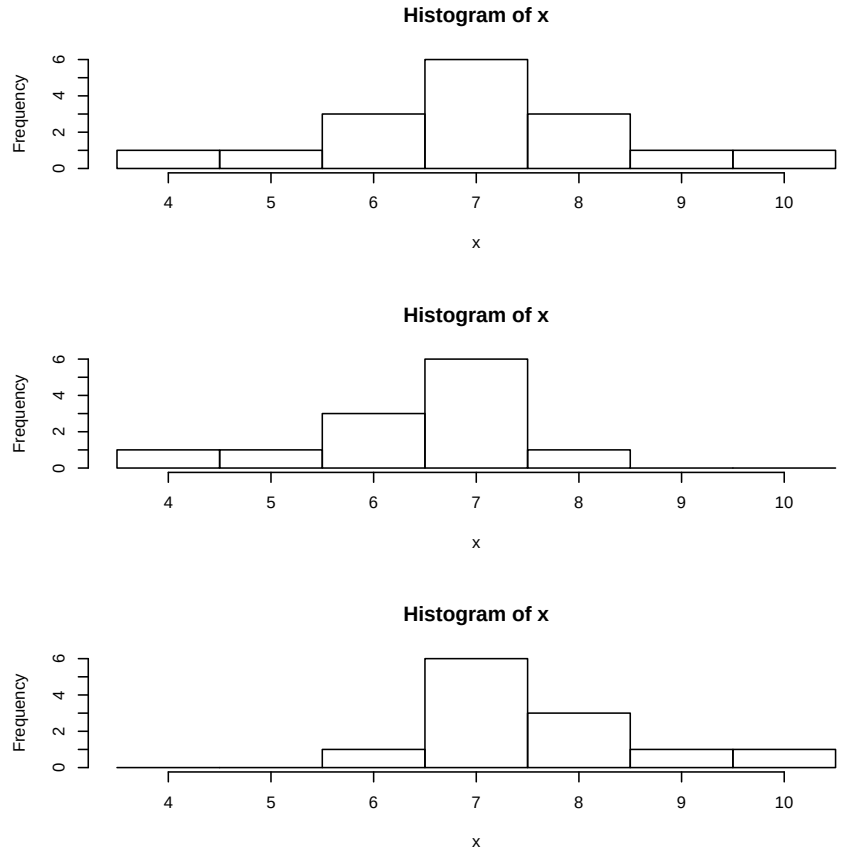

Figure 3.4: Three Histograms

the mean are the same, consider the data:

$$1,\ 1,\ 1,\ 2,\ 2,\ 3,\ 4,\ 4,\ 4,\ 4,\ 4\ .$$

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:

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

3.3.1 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 3.4. Each interval has width one and each value is located at 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 to 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
```

The mean and the median are each 7 for these data. In a perfectly symmetrical distribution, the mean and the median are the same[3].

The functions "`mean`" and "`median`" were used in order to compute the mean and median. Both functions expect a numeric sequence as an input and produce the appropriate measure of centrality of the sequence as an output.

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 3.4. 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 towards 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
```

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

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 3.4. 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
```

[3]In the case of a symmetric distribution the vertical line of symmetry is located at the mean, which is also equal to the median.

The median is yet again equal to 7, but this time the mean is greater than 7. Again, the mean reflects the skewing.

In summary, if the distribution of data is skewed to the left then the mean is less than the median. If the distribution of data is skewed to the right then the median is less than the mean.

Examine the data on the height in "`ex.1`":

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

Observe that the histogram of the height (Figure 3.1) is skewed to the left. This is consistent with the fact that the mean is less than the median.

3.4 Measures of the Spread of Data

One measure of the spread of the data is the inter-quartile range that was introduced in the context of the box plot. However, the most important measure of spread is the standard deviation.

Before dealing with the standard deviation let us discuss the calculation of the variance. If x_i is a data value for subject i and $\bar{x}$ is the sample mean, then $x_i - \bar{x}$ is called the deviation of subject i from the mean, or simply the deviation. In a data set, there are as many deviations as there are data values. The variance is in principle the average of the squares of the deviations.

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. Here are the ages of her students to the nearest half a 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 .

In order to explain the computation of the variance of these data 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
```

Pay attention to the fact that we did not write the "`+`" at the beginning of the second line. That symbol was produced by `R` when moving to the next line to indicate that the expression is not complete yet and will not be executed. Only after inputting the right bracket and the hitting of the Return key does `R` carry out the command and creates the object "`x`". When you execute this example yourself on your own computer make sure not to copy the "`+`" sign. Instead, if you hit the return key after the last comma on the first line, the plus sign will be produced by `R` as a new prompt and you can go on typing in the rest of the numbers.

The function "`length`" returns the length of the input sequence. 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 sequence `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 number that is presented in the output.

From a more technical point of view observe that the expression that computed the deviations, "`x - x.bar`", involved the deletion of a single value (`x.bar`) from a sequence with 20 values (`x`). The expression resulted in the deletion of the value from each component of the sequence. This is an example of the general way by which `R` operates on sequences. The typical behavior of `R` is to apply an operation to each component of the sequence.

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 sequence of length 20. We apply the square function to this sequence. This function is applied to each of the components of the sequence. 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 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
```

If the variance is produced as a result of dividing the sum of squares by the number of observations minus one then the variance is called the *sample variance*.

The function "`var`" computes the sample variance and the function "`sd`" computes the standard deviations. The input to both functions is the sequence 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
```

In the computation of the variance we divide the sum of squared deviations by the number of deviations minus one and not by the number of deviations. The reason for that stems from the theory of statistical inference that will be discussed in Part II of this book. Unless the size of the data is small, dividing by n or by $n-1$ does not introduce much of a difference.

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.

The sample standard deviation, s, is either zero or is larger than zero. When $s = 0$, there is no spread and the data values are equal to each other. 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 is a number that measures how far data values are from their mean. For example, if the data contains the value 7 and if the mean of the data is 5 and the standard deviation is 2, then the value 7 is one standard deviation from its mean because $5 + 1 \times 2 = 7$. We say, then, that 7 is one standard deviation larger than the mean 5 (or also say "to the right of 5"). If the value 1 was also part of the data set, then 1 is two standard deviations smaller than the mean (or two standard deviations to the left of 5) because $5 - 2 \times 2 = 1$.

The standard deviation, when first presented, may not be too simple to interpret. 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 is less so. 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.

3.5 Solved Exercises

Question 3.1. Three sequences of data were saved in 3 R objects named "x1", "x2" and "x3", respectively. The application of the function "summary" to each of these objects is presented below:

```
> summary(x1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   2.498   3.218   3.081   3.840   4.871
> summary(x2)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0001083 0.5772000 1.5070000 1.8420000 2.9050000 4.9880000
> summary(x3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.200   3.391   4.020   4.077   4.690   6.414
```

In Figure 3.5 one may find the histograms of these three data sequences, given in a random order. In Figure 3.6 one may find the box plots of the same data, given in yet a different order.

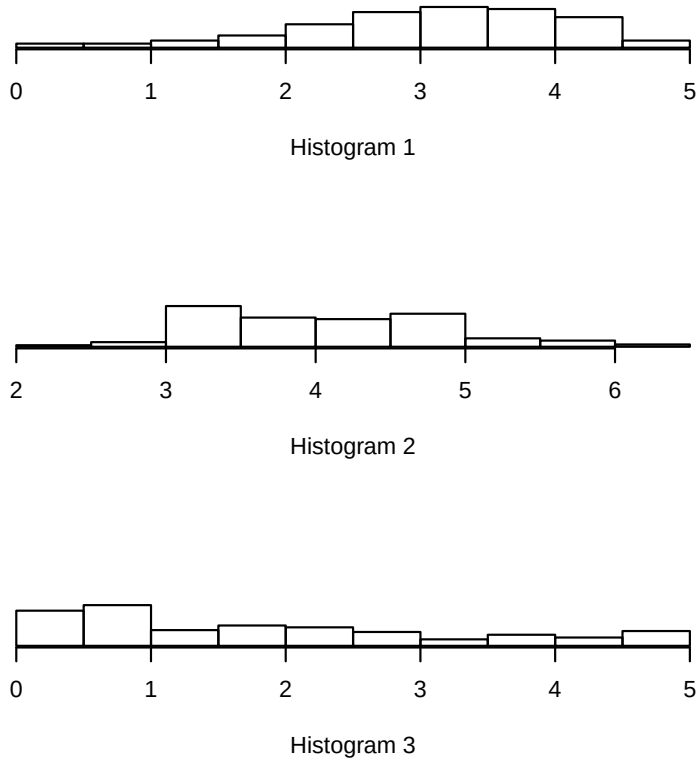

Figure 3.5: Three Histograms

1. Match the summary result with the appropriate histogram and the appropriate box plot.

2. Is the value 0.000 in the sequence "`x1`" an outlier?

3. Is the value 6.414 in the sequence "`x3`" an outlier?

Solution (to Question 3.1.1): Consider the data "`x1`". From the summary we see that it is distributed in the range between 0 and slightly below 5. The central 50% of the distribution are located between 2.5 and 3.8. The mean and median are approximately equal to each other, which suggests an approximately symmetric distribution. Consider the histograms in Figure 3.5. Histograms 1 and 3 correspond to a distributions in the appropriate range. However, the distribution in Histogram 3 is concentrated in lower values than suggested by the given first and third quartiles. Consequently, we match the summary of "`x1`" with Histograms 1.

Consider the data "`x2`". Again, the distribution is in the range between 0 and slightly below 5. The central 50% of the distribution are located between 0.6 and 1.8. The mean is larger than the median, which suggests a distribution skewed

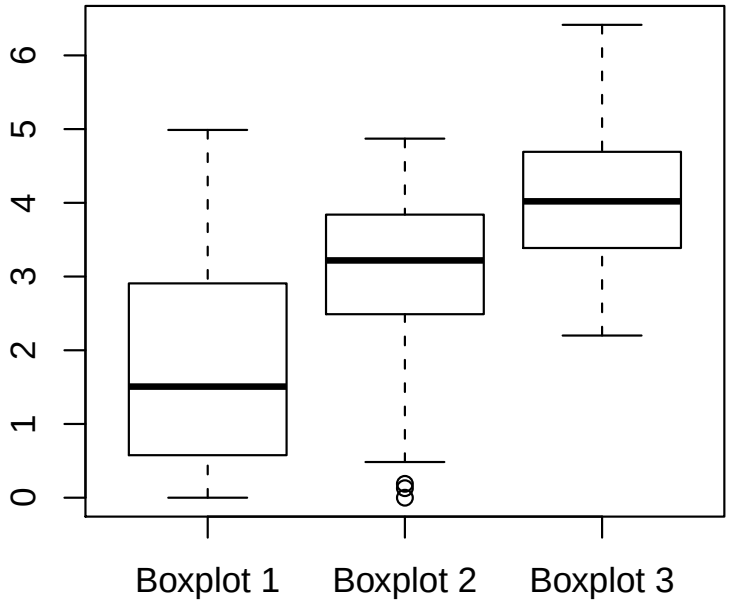

Figure 3.6: Three Box Plots

to the right. Therefore, we match the summary of "`x2`" with Histograms 3.

For the data in "`x3`" we may note that the distribution is in the range between 2 and 6. The histogram that fits this description is Histograms 2.

The box plot is essentially a graphical representation of the information presented by the function "`summary`". Following the rational of matching the summary with the histograms we may obtain that Histogram 1 should be matched with Box-plot 2 in Figure 3.6, Histogram 2 matches Box-plot 3, and Histogram 3 matches Box-plot 1. Indeed, it is easier to match the box plots with the summaries. However, it is a good idea to practice the direct matching of histograms with box plots.

Solution (to Question 3.1.2): The data in "`x1`" fits Box-plot 2 in Figure 3.6. The value 0.000 is the smallest value in the data and it corresponds to the smallest point in the box plot. Since this point is below the bottom whisker it follows that it is an outlier. More directly, we may note that the inter-quartile range is equal to $IQR = 3.840 - 2.498 = 1.342$. The lower threshold is equal to $2.498 - 1.5 \times 1.342 = 0.485$, which is larger that the given value. Consequently, the given value 0.000 is an outlier.

Solution (to Question 3.1.3): Observe that the data in "x3" fits Box-plot 3 in Figure 3.6. The vale 6.414 is the largest value in the data and it corresponds to the endpoint of the upper whisker in the box plot and is not an outlier. Alternatively, we may note that the inter-quartile range is equal to $IQR = 4.690 - 3.391 = 1.299$. The upper threshold is equal to $4.690 + 1.5.299 = 6.6385$, which is larger that the given value. Consequently, the given value 6.414 is not an outlier.

Question 3.2. The number of toilet facilities in 30 buildings were counted. The results are recorded in an R object by the name "x". The frequency table of the data "x" is:

```
> table(x)
x
 2  4  6  8 10
10  6 10  2  2
```

1. What is the mean ($\bar{x}$) of the data?
2. What is the sample standard deviation of the data?
3. What is the median of the data?
4. What is the inter-quartile range (IQR) of the data?
5. How many standard deviations away from the mean is the value 10?

Solution (to Question 3.2.1): In order to compute the mean of the data we may write the following simple R code:

```
> x.val <- c(2,4,6,8,10)
> freq <- c(10,6,10,2,2)
> rel.freq <- freq/sum(freq)
> x.bar <- sum(x.val*rel.freq)
> x.bar
[1] 4.666667
```

We created an object "x.val" that contains the unique values of the data and an object "freq" that contains the frequencies of the values. The object "rel.freq" contains the relative frequencies, the ratios between the frequencies and the number of observations. The average is computed as the sum of the products of the values with their relative frequencies. It is stored in the objects "x.bar" and obtains the value 4.666667.

An alternative approach is to reconstruct the original data from the frequency table. A simple trick that will do the job is to use the function "rep". The first argument to this function is a sequence of values. If the second argument is a sequence of the same length that contains integers then the output will be composed of a sequence that contains the values of the first sequence, each repeated a number of times indicated by the second argument. Specifically, if we enter to this function the unique value "x.val" and the frequency of the values "freq" then the output will be the sequence of values of the original sequence "x":

```
> x <- rep(x.val,freq)
> x
 [1]  2  2  2  2  2  2  2  2  2  2  4  4  4  4  4  4  6  6  6
 [20] 6  6  6  6  6  6  6  8  8 10 10
> mean(x)
[1] 4.666667
```

Observe that when we apply the function "`mean`" to "`x`" we get again the value 4.666667.

Solution (to Question 3.2.2): In order to compute the sample standard deviation we may compute first the sample variance and then take the square root of the result:

```
> var.x <- sum((x.val-x.bar)^2*freq)/(sum(freq)-1)
> sqrt(var.x)
[1] 2.425914
```

Notice that the expression "`sum((x.val-x.bar)^2*freq)`" compute the sum of square deviations. The expression "`(sum(freq)-1)`" produces the number of observations minus 1 ($n-1$). The ratio of the two gives the sample variance.

Alternatively, had we produced the object "`x`" that contains the data, we may apply the function "`sd`" to get the sample standard deviation:

```
> sd(x)
[1] 2.425914
```

Observe that in both forms of computation we obtain the same result: 2.425914.

Solution (to Question 3.2.3): In order to compute the median one may produce the table of cumulative relative frequencies of "`x`":

```
> data.frame(x.val,cumsum(rel.freq))
  x.val cumsum.rel.freq.
1     2        0.3333333
2     4        0.5333333
3     6        0.8666667
4     8        0.9333333
5    10        1.0000000
```

Recall that the object "`x.val`" contains the unique values of the data. The expression "`cumsum(rel.freq)`" produces the cumulative relative frequencies. The function "`data.frame`" puts these two variables into a single data frame and provides a clearer representation of the results.

Notice that more that 50% of the observations have value 4 or less. However, strictly less than 50% of the observations have value 2 or less. Consequently, the median is 4. (If the value of the cumulative relative frequency at 4 would have been exactly 50% then the median would have been the average between 4 and the value larger than 4.)

In the case that we produce the values of the data "`x`" then we may apply the function "`summary`" to it and obtain the median this way

```
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   4.000   4.667   6.000  10.000
```

Solution (to Question 3.2.4): As for the inter-quartile range (IQR) notice that the first quartile is 2 and the third quartile is 6. Hence, the inter-quartile range is equal to 6 - 2 = 4. The quartiles can be read directly from the output of the function "`summary`" or can be obtained from the data frame of the cumulative relative frequencies. For the later observe that more than 25% of the data are less or equal to 2 and more 75% of the data are less or equal to 6 (with strictly less than 75% less or equal to 4).

Solution (to Question 3.2.5): In order to answer the last question we conduct the computation: $(10 - 4.666667)/2.425914 = 2.198484$. We conclude that the value 10 is approximately 2.1985 standard deviations above the mean.

3.6 Summary

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.

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

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

(Sample) 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:

$$s^2 = \frac{\text{Sum of the squares of the deviations}}{\text{Number of values in the sample} - 1} .$$

(Sample) Standard Deviation: A number that is equal to the square root of the variance and measures how far data values are from their mean. $s = \sqrt{s^2}$.

Discuss in the forum

An important practice is to check the validity of any data set that you are supposed to analyze in order to detect errors in the data and outlier observations. Recall that outliers are observations with values outside the normal range of values of the rest of the observations.

It is said by some that outliers can help us understand our data better. What is your opinion?

When forming your answer to this question you may give an example of how outliers may provide insight or, else, how they may abstract our understanding. For example, consider the price of a stock that tend to go up or go down at most 2% within each trading day. A sudden 5% drop in the price of the stock may be an indication to reconsidering our position with respect to this stock.

Commonly Used Symbols

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

Commonly Used Expressions

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

Formulas:

- Mean: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \sum_x \big(x \times (f_x/n)\big)$
- Variance: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{n}{n-1} \sum_x \big((x - \bar{x})^2 \times (f_x/n)\big)$
- Standard Deviation: $s = \sqrt{s^2}$

Chapter 4

Probability

4.1 Student Learning Objective

This section extends the notion of variability that was introduced in the context of data to other situations. The variability of the entire *population* and the concept of a *random variable* is discussed. These concepts are central for the development and interpretation of statistical inference. By the end of the chapter the student should:

- Consider the distribution of a variable in a population and compute parameters of this distribution, such as the mean and the standard deviation.
- Become familiar with the concept of a random variable.
- Understand the relation between the distribution of the population and the distribution of a random variable produced by sampling a random subject from the population.
- Identify the distribution of the random variable in simple settings and compute its expectation and variance.

4.2 Different Forms of Variability

In the previous chapters we examined the variability in data. In the statistical context, data is obtained by selecting a sample from the target population and measuring the quantities of interest for the subjects that belong to the sample. Different subjects in the sample may obtain different values for the measurement, leading to variability in the data.

This variability may be summarized with the aid of a *frequency table*, a table of *relative frequency*, or via the *cumulative relative frequency*. A graphical display of the variability in the data may be obtained with the aid of the *bar plot*, the *histogram*, or the *box plot*.

Numerical summaries may be computed in order to characterize the main features of the variability. We used the *mean* and the *median* in order to identify the location of the distribution. The *sample variance*, or better yet the *sample standard deviation*, as well as the *inter-quartile range* were all described as tools to quantify the overall spread of the data.

The aim of all these graphical representations and numerical summaries is to investigate the variability of the data.

The subject of this chapter is to introduce two other forms of variability, variability that is not associated, at least not directly, with the data that we observe. The first type of variability is the *population variability*. The other type of variability is the variability of a *random variable*.

The notions of variability that will be presented are abstract, they are not given in terms of the data that we observe, and they have a mathematical-theoretical flavor to them. At first, these abstract notions may look to you as a waste of your time and may seem to be unrelated to the subject matter of the course. The opposite is true. The very core of statistical thinking is relating observed data to theoretical and abstract models of a phenomena. Via this comparison, and using the tools of statistical inference that are presented in the second half of the book, statisticians can extrapolate insights or make statements regarding the phenomena on the basis of the observed data. Thereby, the abstract notions of variability that are introduced in this chapter, and are extended in the subsequent chapters up to the end of this part of the book, are the essential foundations for the practice of statistics.

The first notion of variability is the variability that is associated with the population. It is similar in its nature to the variability of the data. The difference between these two types of variability is that the former corresponds to the variability of the quantity of interest across all members of the population and not only for those that were selected to the sample.

In Chapters 2 and 3 we examined the data set "`ex.1`" which contained data on the sex and height of a sample of 100 observations. In this chapter we will consider the sex and height of *all* the members of the population from which the sample was selected. The size of the relevant population is 100,000, including the 100 subjects that composed the sample. When we examine the values of the height across the entire population we can see that different people may have different heights. This variability of the heights is the population variability.

The other abstract type of variability, the variability of a random variable, is a mathematical concept. The aim of this concept is to model the notion of randomness in measurements or the uncertainty regarding the outcome of a measurement. In particular we will initially consider the variability of a random variable in the context of selecting one subject at random from the population.

Imagine we have a population of size 100,000 and we are about to select at random one subject from this population. We intend to measure the height of the subject that will be selected. Prior to the selection and measurement we are not certain what value of height will be obtained. One may associate the notion of variability with uncertainty — different subjects to be selected may obtain different evaluations of the measurement and we do not know before hand which subject will be selected. The resulting variability is the variability of a random variable.

Random variables can be defined for more abstract settings. Their aim is to provide models for randomness and uncertainty in measurements. Simple examples of such abstract random variables will be provided in this chapter. More examples will be introduced in the subsequent chapters. The more abstract examples of random variables need not be associated with a specific population. Still, the same definitions that are used for the example of a random variable that emerges as a result of sampling a single subject from a population will

apply to the more abstract constructions.

All types of variability, the variability of the data we dealt with before as well as the other two types of variability, can be displayed using graphical tools and characterized with numerical summaries. Essentially the same type of plots and numerical summaries, possibly with some modifications, may and will be applied.

A point to remember is that the variability of the data relates to a concrete list of data values that is presented to us. In contrary to the case of the variability of the data, the other types of variability are not associated with quantities we actually get to observe. The data for the sample we get to see but not the data for the rest of the population. Yet, we can still discuss the variability of a population that is out there, even though we do not observe the list of measurements for the entire population. (The example that we give in this chapter of a population was artificially constructed and serves for illustration only. In the actual statistical context one does not obtain measurements from the entire population, only from the subjects that went into the sample.) The discussion of the variability in this context is theoretical in its nature. Still, this theoretical discussion is instrumental for understanding statistics.

4.3 A Population

In this section we introduce the variability of a population and present some numerical summaries that characterizes this variability. Before doing so, let us review with the aid of an example some of the numerical summaries that were used for the characterization of the variability of data.

Recall the file "`ex1.csv`" that contains data on the height and sex of 100 subjects. (The data file can be obtained from `http://pluto.huji.ac.il/~msby/StatThink/Datasets/ex1.csv`.) We read the content of the file into a data frame by the name "`ex.1`" and apply the function "`summary`" to the data frame:

```
> ex.1 <- read.csv("ex1.csv")
> summary(ex.1)
       id              sex         height
 Min.   :1538611   FEMALE:54   Min.   :117.0
 1st Qu.:3339583   MALE  :46   1st Qu.:158.0
 Median :5105620               Median :171.0
 Mean   :5412367               Mean   :170.1
 3rd Qu.:7622236               3rd Qu.:180.2
 Max.   :9878130               Max.   :208.0
```

We saw in the previous chapter that, when applied to a numeric sequence, the function "`summary`" produces the smallest and largest values in the sequence, the three quartiles (including the median) and the mean. If the input of the same function is a factor then the outcome is the frequency in the data of each of the levels of the factor. Here "`sex`" is a factor with two levels. From the summary we can see that 54 of the subjects in the sample are female and 46 are male.

Notice that when the input to the function "`summary`" is a data frame, as is the case in this example, then the output is a summary of each of the variables

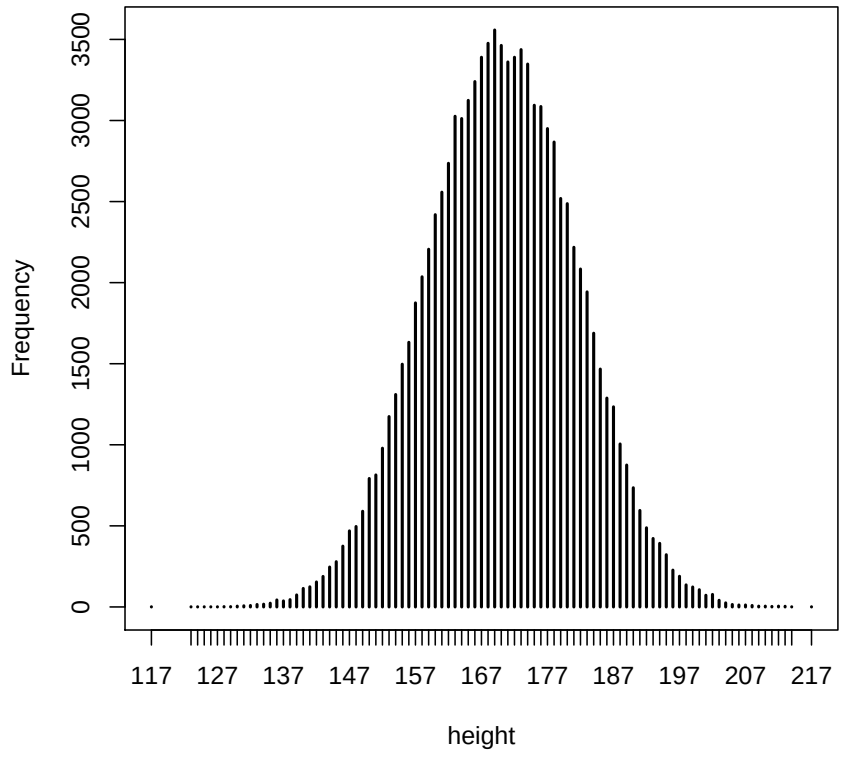

Figure 4.1: Bar Plot of Height

of the data frame. In this example two of the variables are numeric ("`id`" and "`height`") and one variable is a factor ("`sex`").

Recall that the mean is the arithmetic average of the data which is computed by summing all the values of the variable and dividing the result by the number of observations. Hence, if n is the number of observations ($n = 100$ in this example) and x_i is the value of the variable for subject i, then one may write the mean in a formula form as

$$\bar{x} = \frac{\text{Sum of all values in the data}}{\text{Number of values in the data}} = \frac{\sum_{i=1}^{n} x_i}{n} \ ,$$

where $\bar{x}$ corresponds to the mean of the data and the symbol "$\sum_{i=1}^{n} x_i$" corresponds to the sum of all values in the data.

The median is computed by ordering the data values and selecting a value that splits the ordered data into two equal parts. The first and third quartile are obtained by further splitting each of the halves into two quarters.

Let us discuss the variability associated with an entire target population. The file "`pop1.csv`" that contains the population data can be found on the internet (`http://pluto.huji.ac.il/~msby/StatThink/Datasets/pop1.csv`). It

is a CSV file that contains the information on sex and height of an entire adult population of some imaginary city. (The data in "ex.1" corresponds to a sample from this city.) Read the population data into R and examine it:

```
> pop.1 <- read.csv(file="pop1.csv")
> summary(pop.1)
       id               sex            height
 Min.   : 1000082   FEMALE:48888   Min.   :117.0
 1st Qu.: 3254220   MALE  :51112   1st Qu.:162.0
 Median : 5502618                  Median :170.0
 Mean   : 5502428                  Mean   :170.0
 3rd Qu.: 7757518                  3rd Qu.:178.0
 Max.   : 9999937                  Max.   :217.0
```

The object "pop.1" is a data frame of the same structure as the data frame "ex.1". It contains three variables: a unique identifier of each subject (id), the sex of the subject (sex), and its height (height). Applying the function "summary" to the data frame produces the summary of the variables that it contains. In particular, for the variable "sex", which is a factor, it produces the frequency of its two categories – 48,888 female and 51,112 – a total of 100,000 subjects. For the variable "height", which is a numeric variable, it produces the extreme values, the quartiles, and the mean.

Let us concentrate on the variable "height". A bar plot of the distribution of the heights in the entire population is given in Figure 4.1 [1]. Recall that a vertical bar is placed above each value of height that appears in the population, with the height of the bar representing the frequency of the value in the population. One may read out of the graph or obtain from the numerical summaries that the variable takes integer values in the range between 117 and 217 (heights are rounded to the nearest centimeter). The distribution is centered at 170 centimeter, with the central 50% of the values spreading between 162 and 178 centimeters.

The mean of the height in the entire population is equal to 170 centimeter. This mean, just like the mean for the distribution of data, is obtained by the summation of all the heights in the population divided by the population size. Let us denote the size of the entire population by N. In this example $N = 100,000$. (The size of the sample for the data was called n and was equal to $n = 100$ in the parallel example that deals with the data of a sample.) The mean of an entire population is denoted by the Greek letter μ and is read "*mew*". (The average for the data was denoted $\bar{x}$). The formula of the population mean is:

$$\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}} = \frac{\sum_{i=1}^{N} x_i}{N} \ .$$

Observe the similarity between the definition of the mean for the data and the definition of the mean for the population. In both cases the arithmetic average is computed. The only difference is that in the case of the mean of the data the computation is with respect to the values that appear in the sample whereas for the population all the values in the population participate in the computation.

[1] Such a bar plot can be produced with the expression "plot(table(pop.1$height))".

In actual life, we will not have all the values of a variable in the entire population. Hence, we will not be able to compute the actual value of the population mean. However, it is still meaningful to talk about the population mean because this number exists, even though we do not know what its value is. As a matter of fact, one of the issues in statistics is to try to estimate this unknown quantity on the basis of the data we do have in the sample.

A characteristic of the distribution of an entire population is called a *parameter*. Hence, μ, the population average, is a parameter. Other examples of parameters are the population median and the population quartiles. These parameters are defined exactly like their data counterparts, but with respect to the values of the entire population instead of the observations in the sample alone.

Another example of a parameter is the population variance. Recall that the sample variance was defined with the aid of the deviations $x_i - \bar{x}$, where x_i is the value of the measurement for the ith subject and $\bar{x}$ is the mean for the data. In order to compute the sample variance these deviations were squared to produce the squared deviations. The squares were summed up and then divided by the sample size minus one $(n-1)$. The sample variance, computed from the data, was denoted s^2.

The population variance is defined in a similar way. First, the deviations from the population mean $x_i - \mu$ are considered for each of the members of the population. These deviations are squared and the average of the squares is computed. We denote this parameter by σ^2 (read "*sigma square*"). A minor difference between the sample variance and the population variance is that for the latter we should divide the sum of squared deviations by the population size (N) and not by the population size minus one $(N-1)$:

$$\begin{aligned}\sigma^2 =& \text{The average square deviation in the population}\\ =& \frac{\text{Sum of the squares of the deviations in the population}}{\text{Number of values in the population}}\\ =& \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}\ .\end{aligned}$$

The standard deviation of the population, yet another parameter, is denoted by σ and is equal to the square root of the variance. The standard deviation summarizes the overall variability of the measurement across the population. Again, the typical situation is that we do not know what the actual value of the standard deviation of the population is. Yet, we may refer to it as a quantity and we may try to estimate its value based on the data we do have from the sample.

For the height of the subjects in our imaginary city we get that the variance is equal to $\sigma^2 = 126.1576$. The standard deviation is equal to $\sigma = \sqrt{126.1576} = 11.23199$. These quantities can be computed in this example from the data frame "`pop.1`" with the aid of the functions "`var`" and "`sd`", respectively[2].

[2] Observe that the function "`var`" computes the sample variance. Consequently, the sum of squares is divided by $N-1$. We can correct that when computing the population variance by multiplying the result by $N-1$ and dividing by N. Notice that the difference between the two quantities is negligible for a large population. Henceforth we will use the functions "`var`" and "`sd`" to compute the variance and standard deviations of populations without the application of the correction.

4.4 Random Variables

In the previous section we dealt with the variability of the population. Next we consider the variability of a random variable. As an example, consider taking a sample of size $n = 1$ from the population (a single person) and measuring his/her height.

The object `pop.1$height` is a sequence with 100,000 entries. Think of it as a population. We will apply the function "`sample`" to this sequence:

```
> sample(pop.1$height,1)
[1] 162
```

The first entry to the function is the given sequence of heights. When we set the second argument to 1 then the function selects one of the entries of the sequence at random, with each entry having the same likelihood of being selected. Specifically, in this example an entry that contains the value 162 was selected. Let us run the function again:

```
> sample(pop.1$height,1)
[1] 192
```

In this instance an entry with a different value was selected. Try to run the command several times yourself and see what you get. Would you necessarily obtain a different value in each run?

Now let us enter the same command without pressing the return key:

```
> sample(pop.1$height,1)
```

Can you tell, before pressing the key, what value will you get?

The answer to this question is of course *"No"*. There are 100,000 entries with a total of 94 distinct values. In principle, any of the values may be selected and there is no way of telling in advance which of the values will turn out as an outcome.

A random variable is the future outcome of a measurement, **before** the measurement is taken. It does not have a specific value, but rather a collection of potential values with a distribution over these values. After the measurement is taken and the specific value is revealed then the random variable ceases to be a random variable! Instead, it becomes data.

Although one is not able to say what the outcome of a random variable will turn out to be. Still, one may identify patterns in this potential outcome. For example, knowing that the distribution of heights in the population ranges between 117 and 217 centimeter one may say in advance that the outcome of the measurement must also be in that interval. Moreover, since there is a total of 3,476 subjects with height equal to 168 centimeter and since the likelihood of each subject to be selected is equal then the likelihood of selecting a subject of this height is 3,476/100,000 = 0.03476. In the context of random variables we call this likelihood *probability*. In the same vain, the frequency of subjects with hight 192 centimeter is 488, and therefore the probability of measuring such a height is 0.00488. The frequency of subjects with height 200 centimeter or above is 393, hence the probability of obtaining a measurement in the range between 200 and 217 centimeter is 0.00393.

4.4.1 Sample Space and Distribution

Let us turn to the formal definition of a random variable: A random variable refer to numerical values, typically the outcome of an observation, a measurement, or a function thereof.

A random variable is characterized via the collection of potential values it may obtain, known as the *sample space* and the likelihood of obtaining each of the values in the sample space (namely, the probability of the value). In the given example, the sample space contains the 94 integer values that are marked in Figure 4.1. The probability of each value is the height of the bar above the value, divided by the total frequency of 100,000 (namely, the relative frequency in the population).

We will denote random variables with capital Latin letters such as X, Y, and Z. Values they may obtain will be marked by small Latin letters such as x, y, z. For the probability of values we will use the letter "P". Hence, if we denote by X the measurement of height of a random individual that is sampled from the given population then:

$$\mathrm{P}(X = 168) = 0.03476$$

and

$$\mathrm{P}(X \geq 200) = 0.00393\ .$$

Consider, as yet another example, the probability that the height of a random person sampled from the population differs from 170 centimeter by no more than 10 centimeters. (In other words, that the height is between 160 and 180 centimeters.) Denote by X the height of that random person. We are interested in the probability $\mathrm{P}(|X - 170| \leq 10)$.[3]

The random person can be any of the subjects of the population with equal probability. Thus, the sequence of the heights of the 100,000 subjects represents the distribution of the random variable X:

```
> pop.1 <- read.csv(file="pop1.csv")
> X <- pop.1$height
```

Notice that the object "`X`" is a sequence of lenght 100,000 that stores all the heights of the population. The probability we seek is the relative frequency in this sequency of values between 160 and 180. First we compute the probability and then explain the method of computation:

```
> mean(abs(X-170) <= 10)
[1] 0.64541
```

We get that the height of a person randomly sampled from the population is between 160 and 180 centimeters with probability 0.64541.

Let us produce a small example that will help us explain the computation of the probability. We start by forming a sequence with 10 numbers:

```
> Y <- c(6.3, 6.9, 6.6, 3.4, 5.5, 4.3, 6.5, 4.7, 6.1, 5.3)
```

[3] The expression $\{|X - 170| \leq 10\}$ reads as "the absolute value of the difference between X and 170 is no more that 10". In other words, $\{-10 \leq X - 170 \leq 10\}$, which is equivalent to the statement that $\{160 \leq X \leq 180\}$. It follows that $\mathrm{P}(|X - 170| \leq 10) = \mathrm{P}(160 \leq X \leq 180)$.

The goal is to compute the proportion of numbers that are in the range $[4, 6]$ (or, equivalently, $\{|Y - 5| \leq 1\}$).

The function "`abs`" computes the absolute number of its input argument. When the function is applied to the sequence "`Y-5`" it produces a sequence of the same length with the distances between the components of "`Y`" and the number 5:

```
> abs(Y-5)
 [1] 1.3 1.9 1.6 1.6 0.5 0.7 1.5 0.3 1.1 0.3
```

Compare the resulting output to the original sequence. The first value in the input sequence is 6.3. Its distance from 5 is indeed 1.3. The fourth value in the input sequence is 3.4. The difference 3.4 - 5 is equal to -1.6, and when the absolute value is taken we get a distance of 1.6.

The function "`<=`" expects an argument to the right and an argument to the left. It compares each component to the left with the parallel component to the right and returns a logical value, "`TRUE`" or "`FALSE`", depending on whether the relation that is tested holds or not:

```
> abs(Y - 5) <= 1
 [1] FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE
```

Observe the in this example the function "`<=`" produced 10 logical values, one for each of the elements of the sequence to the left of it. The first input in the sequence "`Y`" is 6.3, which is more than one unit away from 5. Hence, the first output of the logical expression is "`FALSE`". On the other hand, the last input in the sequence "`Y`" is 5.3, which is within the range. Therefore, the last output of the logical expression is "`TRUE`".

Next, we compute the proportion of "`TRUE`" values in the sequence:

```
> mean(abs(Y - 5) <= 1)
 [1] 0.4
```

When a sequence with logical values is entered into the function "`mean`" then the function replaces the `TRUE`'s by 1 and the `FALSE`'s by 0. The average produces then the relative frequency of `TRUE`'s in the sequence as required. Specifically, in this example there are 4 `TRUE`'s and 6 `FALSE`'s. Consequently, the output of the final expression is $4/10 = 0.4$.

The computation of the probability that the sampled height falls within 10 centimeter of 170 is based on the same code. The only differences are that the input sequence "`Y`" is replaced by the sequence of population heights "`X`" as input. the number "`5`" is replaced by the number "`170`" and the number "`1`" is replaced by the number "`10`". In both cases the result of the computation is the relative proportion of the times that the values of the input sequence fall within a given range of the indicated number.

The probability function of a random variable is defined for any value that the random variable may obtain and produces the *distribution* of the random variable. The probability function may emerge as a relative frequency as in the given example or it may be a result of theoretical modeling. Examples of theoretical random variables are presented mainly in the next two chapters.

Consider an example of a random variable. The sample space and the probability function specify the distribution of the random variable. For example,

assume it is known that a random variable X may obtain the values 0, 1, 2, or 3. Moreover, imagine that it is known that $\mathrm{P}(X = 1) = 0.25$, $\mathrm{P}(X = 2) = 0.15$, and $\mathrm{P}(X = 3) = 0.10$. What is $\mathrm{P}(X = 0)$, the probability that X is equal to 0?

The sample space, the collection of possible values that the random variable may obtain is the collection $\{0, 1, 2, 3\}$. Observe that the sum over the positive values is:

$$\mathrm{P}(X > 0) = \mathrm{P}(X = 1) + \mathrm{P}(X = 2) + \mathrm{P}(X = 3) = 0.25 + 0.15 + 0.10 = 0.50 \; .$$

It follows, since the sum of probabilities over the entire sample space is equal to 1, that $\mathrm{P}(X = 0) = 1 - 0.5 = 0.5$.

Value	Probability	Cum. Prob.
0	0.50	0.50
1	0.25	0.75
2	0.15	0.90
3	0.10	1.00

Table 4.1: The Distribution of X

Table 4.1 summarizes the distribution of the random variable X. Observe the similarity between the probability function and the notion of relative frequency that was discussed in Chapter 2. Both quantities describe distribution. Both are non-negative and sum to 1. Likewise, notice that one may define the cumulative probability the same way cumulative relative frequency is defined: Ordering the values of the random variable from smallest to largest, the cumulative probability at a given value is the sum of probabilities for values less or equal to the given value.

Knowledge of the probabilities of a random variable (or the cumulative probabilities) enables the computation of other probabilities that are associated with the random variable. For example, considering the random variable X of Table 4.1, we may calculate the probability of X falling in the interval $[0.5, 2.3]$. Observe that the given range contains two values from the sample space, 1 and 2, therefore:

$$\mathrm{P}(0.5 \leq X \leq 2.3) = \mathrm{P}(X = 1) + \mathrm{P}(X = 2) = 0.25 + 0.15 = 0.40 \; .$$

Likewise, we may produce the probability of X obtaining an odd value:

$$\mathrm{P}(X = \text{odd}) = \mathrm{P}(X = 1) + \mathrm{P}(X = 3) = 0.25 + 0.10 = 0.35 \; .$$

Observe that both $\{0.5 \leq X \leq 2.3\}$ and $\{X = \text{odd}\}$ refer to subsets of values of the sample space. Such subsets are denoted *events*. In both examples the probability of the event was computed by the summation of the probabilities associated with values that belong to the event.

4.4.2 Expectation and Standard Deviation

We may characterize the center of the distribution of a random variable and the spread of the distribution in ways similar to those used for the characterization of the distribution of data and the distribution of a population.

The *expectation* marks the center of the distribution of a random variable. It is equivalent to the data average $\bar{x}$ and the population average μ, which was used in order to mark the location of the distribution of the data and the population, respectively.

Recall from Chapter 3 that the average of the data can be computed as the weighted average of the values that are present in the data, with weights given by the relative frequency. Specifically, we saw for the data

$$1,\ 1,\ 1,\ 2,\ 2,\ 3,\ 4,\ 4,\ 4,\ 4,\ 4$$

that

$$\frac{1+1+1+2+2+3+4+4+4+4+4}{11} = 1\times\frac{3}{11}+2\times\frac{2}{11}+3\times\frac{1}{11}+4\times\frac{5}{11}\ ,$$

producing the value of $\bar{x} = 2.727$ in both representations. Using a formula, the equality between the two ways of computing the mean is given in terms of the equation:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \sum_x \big(x \times (f_x/n)\big)\ .$$

In the first representation of the arithmetic mean, the average is computed by the summation of all data points and dividing the sum by the sample size. In the second representation, that uses a weighted sum, the sum extends over all the unique values that appear in the data. For each unique value the value is multiplied by the relative frequency of the value in the data. These multiplications are summed up to produce the mean.

The expectation of a random variable is computed in the spirit of the second formulation. The expectation of a random variable is marked with the letter "E" and is define via the equation:

$$\mathrm{E}(X) = \sum_x \big(x \times \mathrm{P}(x)\big)\ .$$

In this definition all the unique values of the sample space are considered. For each value a product of the value and the probability of the value is taken. The expectation is obtained by the summation of all these products. In this definition the probability $\mathrm{P}(x)$ replaces the relative frequency f_x/n but otherwise, the definition of the expectation and the second formulation of the mean are identical to each other.

Consider the random variable X with distribution that is described in Table 4.1. In order to obtain its expectation we multiply each value in the sample space by the probability of the value. Summation of the products produces the expectation (see Table 4.2):

$$\mathrm{E}(X) = 0\times 0.5 + 1\times 0.25 + 2\times 0.15 + 3\times 0.10 = 0.85\ .$$

In the example of height we get that the expectation is equal to 170.035 centimeter. Notice that this expectation is equal to μ, the mean of the population[4]. This is no accident. The expectation of a potential measurement of a randomly selected subject from a population is equal to the average of the measurement across all subjects.

[4]The mean of the population can be computed with the expression "`mean(pop.1$height)`"

Value	Probability	$x \times \mathrm{P}(X = x)$
0	0.50	0.00
1	0.25	0.25
2	0.15	0.30
3	0.10	0.30
		$\mathrm{E}(X) = 0.85$

Table 4.2: The Expectation of X

The sample variance (s^2) is obtained as the sum of the squared deviations from the average, divided by the sample size (n) minus 1:

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

A second formulation for the computation of the same quantity is via the use of relative frequencies. The formula for the sample variance takes the form

$$s^2 = \frac{n}{n-1} \sum_x \left((x - \bar{x})^2 \times (f_x/n)\right) \ .$$

In this formulation one considers each of the unique value that are present in the data. For each value the deviation between the value and the average is computed. These deviations are then squared and multiplied by the relative frequency. The products are summed up. Finally, the sum is multiplied by the ratio between the sample size n and $n-1$ in order to correct for the fact that in the sample variance the sum of squared deviations is divided by the sample size minus 1 and not by the sample size.

In a similar way, the variance of a random variable may be defined via the probability of the values that make the sample space. For each such value one computes the deviation from the expectation. This deviation is then squared and multiplied by the probability of the value. The multiplications are summed up in order to produce the variance:

$$\mathrm{Var}(X) = \sum_x \left((x - \mathrm{E}(X))^2 \times \mathrm{P}(x)\right) \ .$$

Notice that the formula for the computation of the variance of a random variable is very similar to the second formulation for the computation of the sample variance. Essentially, the mean of the data is replaced by the expectation of the random variable and the relative frequency of a value is replaced by the probability of the value. Another difference is that the correction factor is not used for the variance of a random variable.

As an example consider the variance of the random variable X. The computation of the variance of this random variable is carried out in Table 4.3). The sample space, the values that the random variable may obtain, are given in the first column and the probabilities of the values are given in the second column. In the third column the deviation of the value from the expectation $\mathrm{E}(X) = 0.85$ is computed for each value. The 4th column contains the square of these deviations and the 5th and last column involves the product of the square deviations and the probabilities. The variance is obtained by summing up the

Value	Prob.	$x - \mathrm{E}(X)$	$(x - \mathrm{E}(X))^2$	$(x - \mathrm{E}(X))^2 \times \mathrm{P}(X = x)$
0	0.50	-0.85	0.7225	0.361250
1	0.25	0.15	0.0225	0.005625
2	0.15	1.15	1.3225	0.198375
3	0.10	2.15	4.6225	0.462250
				$\mathrm{Var}(X) = 1.027500$

Table 4.3: The Variance of X

products in the last column. In the given example:

$$\begin{aligned}\mathrm{Var}(X) =& (0 - 0.85)^2 \times 0.5 + (1 - 0.85)^2 \times 0.25 \\ &+ (2 - 0.85)^2 \times 0.15 + (3 - 0.85)^2 \times 0.10 = 1.0275 \ .\end{aligned}$$

The standard deviation of a random variable is the square root of the variance. The standard deviation of X is $\sqrt{\mathrm{Var}(X)} = \sqrt{1.0275} = 1.013657$.

In the example that involves the height of a subject selected from the population at random we obtain that the variance is 126.1576, equal to the population variance, and the standard deviation is 11.23199, the square root of the variance.

Other characterization of the distribution that were computed for data, such as the median, the quartiles, etc., may also be defined for random variables.

4.5 Probability and Statistics

Modern science may be characterized by a systematic collection of empirical measurements and the attempt to model laws of nature using mathematical language. The drive to deliver better measurements led to the development of more accurate and more sensitive measurement tools. Nonetheless, at some point it became apparent that measurements may not be perfectly reproducible and any repeated measurement of presumably the exact same phenomena will typically produce variability in the outcomes. On the other hand, scientists also found that there are general laws that govern this variability in repetitions. For example, it was discovered that the average of several independent repeats of the measurement is less variable and more reproducible than each of the single measurements themselves.

Probability was first introduced as a branch of mathematics in the investigation of uncertainty associated with gambling and games of chance. During the early 19th century probability began to be used in order to model variability in measurements. This application of probability turned out to be very successful. Indeed, one of the major achievements of probability was the development of the mathematical theory that explains the phenomena of reduced variability that is observed when averages are used instead of single measurements. In Chapter **??** we discuss the conclusions of this theory.

Statistics study method for inference based on data. Probability serves as the mathematical foundation for the development of statistical theory. In this chapter we introduced the probabilistic concept of a random variable. This concept is key for understanding statistics. In the rest of Part I of this book we discuss the probability theory that is used for statistical inference. Statistical inference itself is discussed in Part II of the book.

Value	Probability
0	p
1	$2p$
2	$3p$
3	$4p$
4	$5p$
5	$6p$

Table 4.4: The Distribution of Y

4.6 Solved Exercises

Question 4.1. Table 4.6 presents the probabilities of the random variable Y. These probabilities are a function of the number p, the probability of the value "0". Answer the following questions:

1. What is the value of p?
2. $\mathrm{P}(Y < 3) = ?$
3. $\mathrm{P}(Y = \text{odd}) = ?$
4. $\mathrm{P}(1 \leq Y < 4) = ?$
5. $\mathrm{P}(|Y - 3| < 1.5) = ?$
6. $\mathrm{E}(Y) = ?$
7. $\mathrm{Var}(Y) = ?$
8. What is the standard deviation of Y.

Solution (to Question 4.1.1): Consult Table 4.6. The probabilities of the different values of Y are $\{p, 2p, \ldots, 6p\}$. These probabilities sum to 1, consequently

$$p + 2p + 3p + 4p + 5p + 6p = (1 + 2 + 3 + 4 + 5 + 6)p = 21p = 1 \Longrightarrow p = 1/21 \,.$$

Solution (to Question 4.1.2): The event $\{Y < 3\}$ contains the values 0, 1 and 2. Therefore,

$$\mathrm{P}(Y < 3) = \mathrm{P}(Y = 0) + \mathrm{P}(Y = 1) + \mathrm{P}(Y = 2) = \frac{1}{21} + \frac{2}{21} + \frac{3}{21} = \frac{6}{21} = 0.2857 \,.$$

Solution (to Question 4.1.3): The event $\{Y = \text{odd}\}$ contains the values 1, 3 and 5. Therefore,

$$\mathrm{P}(Y = \text{odd}) = \mathrm{P}(Y = 1) + \mathrm{P}(Y = 3) + \mathrm{P}(Y = 5) = \frac{2}{21} + \frac{4}{21} + \frac{6}{21} = \frac{12}{21} = 0.5714 \,.$$

Solution (to Question 4.1.4): The event $\{1 \leq Y < 4\}$ contains the values 1, 2 and 3. Therefore,

$$\mathrm{P}(1 \leq Y < 4) = \mathrm{P}(Y = 1) + \mathrm{P}(Y = 2) + \mathrm{P}(Y = 3) = \frac{2}{21} + \frac{3}{21} + \frac{4}{21} = \frac{9}{21} = 0.4286 \,.$$

Solution (to Question 4.1.5): The event $\{|Y - 3| < 1.5\}$ contains the values 2, 3 and 4. Therefore,

$$\mathrm{P}(|Y-3| < 1.5) = \mathrm{P}(Y = 2) + \mathrm{P}(Y = 3) + \mathrm{P}(Y = 4) = \frac{3}{21} + \frac{4}{21} + \frac{5}{21} = \frac{12}{21} = 0.5714 \,.$$

Solution (to Question 4.1.6): The values that the random variable Y obtains are the numbers 0, 1, 2, ..., 5, with probabilities $\{1/21, 2/21, \ldots, 6/21\}$, respectively. The expectation is obtained by the multiplication of the values by their respective probabilities and the summation of the products. Let us carry out the computation in `R`:

```
> Y.val <- c(0,1,2,3,4,5)
> P.val <- c(1,2,3,4,5,6)/21
> E <- sum(Y.val*P.val)
> E
[1] 3.333333
```

We obtain an expectation $\mathrm{E}(Y) = 3.3333$.

Solution (to Question 4.1.7): The values that the random variable Y obtains are the numbers 0, 1, 2, ..., 5, with probabilities $\{1/21, 2/21, \ldots, 6/21\}$, respectively. The expectation is equal to $\mathrm{E}(Y) = 3.333333$. The variance is obtained by the multiplication of the squared deviation from the expectation of the values by their respective probabilities and the summation of the products. Let us carry out the computation in `R`:

```
> Var <- sum((Y.val-E)^2*P.val)
> Var
[1] 2.222222
```

We obtain a variance $\mathrm{Var}(Y) = 2.2222$.

Solution (to Question 4.1.8): The standard deviation is the square root of the variance: $\sqrt{\mathrm{Var}(Y)} = \sqrt{2.2222} = 1.4907$.

Question 4.2. One invests \$2 to participate in a game of chance. In this game a coin is tossed three times. If all tosses end up "Head" then the player wins \$10. Otherwise, the player losses the investment.

1. What is the probability of winning the game?

2. What is the probability of loosing the game?

3. What is the expected gain for the player that plays this game? (Notice that the expectation can obtain a negative value.)

Solution (to Question 4.2.1): An outcome of the game of chance may be represented by a sequence of length three composed of the letters "H" and "T". For example, the sequence "THH" corresponds to the case where the first toss produced a "Tail", the second a "Head" and the third a "Head".

With this notation we obtain that the possible outcomes of the game are $\{\text{HHH}, \text{THH}, \text{HTH}, \text{TTH}, \text{HHT}, \text{THT}, \text{HTT}, \text{TTT}\}$. All outcomes are equally likely. There are 8 possible outcomes and only one of which corresponds to winning. Consequently, the probability of winning is 1/8.

Solution (to Question 4.2.2): Consider the previous solution. One looses if any other of the outcomes occurs. Hence, the probability of loosing is 7/8.

Solution (to Question 4.2.3): Denote the gain of the player by X. The random variable X may obtain two values: 10-2 = 8 if the player wins and -2 if the player looses. The probabilities of these values are $\{1/8, 7/8\}$, respectively. Therefore, the expected gain, the expectation of X is:

$$\mathrm{E}(X) = 8 \times \frac{1}{8} + (-2) \times \frac{7}{8} = -0.75 \; .$$

4.7 Summary

Glossary

Random Variable: The probabilistic model for the value of a measurement, before the measurement is taken.

Sample Space: The set of all values a random variable may obtain.

Probability: A number between 0 and 1 which is assigned to a subset of the sample space. This number indicates the likelihood of the random variable obtaining a value in that subset.

Expectation: The central value for a random variable. The expectation of the random variable X is marked by $\mathrm{E}(X)$.

Variance: The (squared) spread of a random variable. The variance of the random variable X is marked by $\mathrm{Var}(X)$. The standard deviation is the square root of the variance.

Discussion in the Forum

Random variables are used to model situations in which the outcome, before the fact, is uncertain. One component in the model is the sample space. The sample space is the list of all possible outcomes. It includes the outcome that took place, but also all other outcomes that could have taken place but never did materialize. The rationale behind the consideration of the sample space is

the intention to put the outcome that took place in context. What do you think of this rationale?

When forming your answer to this question you may give an example of a situation from you own field of interest for which a random variable can serve as a model. Identify the sample space for that random variable and discuss the importance (or lack thereof) of the correct identification of the sample space.

For example, consider a factory that produces car parts that are sold to car makers. The role of the QA personnel in the factory is to validate the quality of each batch of parts before the shipment to the client.

To achieve that, a sample of parts may be subject to a battery of quality test. Say that 20 parts are selected to the sample. The number of those among them that will not pass the quality testing may be modeled as a random variable. The sample space for this random variable may be any of the numbers 0, 1, 2, ..., 20.

The number 0 corresponds to the situation where all parts in the sample passed the quality testing. The number 1 corresponds to the case where 1 part did not pass and the other 19 did. The number 2 describes the case where 2 of the 20 did not pass and 18 did pass, etc.

Summary of Formulas

Population Size: N = the number of people, things, etc. in the population.

Population Average: $\mu = (1/N)\sum_{i=1}^{N} x_i$

Expectation of a Random Variable: $\mathrm{E}(X) = \sum_x \big(x \times \mathrm{P}(x)\big)$

Population Variance: $\sigma^2 = (1/N)\sum_{i=1}^{N}(x_i - \mu)^2$

Variance of a Random Variable: $\mathrm{Var}(X) = \sum_x \big((x - \mathrm{E}(X))^2 \times \mathrm{P}(x)\big)$