Chapter 2

Descriptive Statistics: Summary Statistics

In this chapter we will discuss the following topics:

- How to find the mean of a set of observations with the R-function **mean()**.
- How to find the median of a set of observations with the R-function **median()**.
- How to find the standard deviation of a set of observations with the R-function **sd()**.
- How to find the variance of a set of observations with the R-function **var()**.
- How to find the quartiles of a set of observations with the R-function **quantile()**.
- How to find outliers in the data.
- How to find the five-number summary of a set of observations with the R-function **summary()**.
- How to compute a boxplot with the R-function **boxplot()**.

Mean, Median, Standard deviation, Variance, and the Quartiles

Given a set of n observations x_1, x_2,..., x_n. The *mean* of the observations is given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The mean is a measure of central tendency of the distribution. Another measure of central tendency is the *median* which is defined to be the value such that half of the observations are greater than and half are smaller than that value when the data is sorted in ascending order. In an exact symmetric distribution, the mean and the median are equal.

The *variance* of the observations is given by

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

and is a measure of spread of the distribution from the mean. Notice that we divided by $n-1$ instead of by n. The explanation of this is as follows: Notice that we have

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0. \tag{0.1}$$

This means that we have $n-1$ variables that can vary freely since the value of the last one can be determined from the equation in (0.1). The estimation of the mean $\bar{x}$ in the expression for the variance resulted in the loss of one degrees of freedom leading to $n-1$ degrees of

freedom for the variance s^2. We will return to the concept of degrees of freedom in a later chapter. The *standard deviation* s of the observations is the square root of the variance,

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

and has the same measurement unit as the mean.

The *first quartile* is the value such that 25% of the observations are lower than that value when the data is sorted in ascending order. The *third quartile* is the value such that 75% of the observations are lower than that value when the data is sorted in ascending order. Notice that R uses a different algorithm for computing quartiles than the textbook so the results might be slightly different.

Problem. Given the following set of observations: 1, 2, 3, 4, 2, 3, 1, 2, 5, 9, 10. Find the mean, median, standard deviation, variance, the first- and third quartile of the observations.

Solution. We obtain:

```
> x=c(1,2,3,4,2,3,1,2,5,9,10)
> mean(x)
[1] 3.818182
> median(x)
[1] 3
> var(x)
[1] 9.363636
> sd(x)
[1] 3.060006
> quantile(x,0.25)
25%
  2
> quantile(x,0.75)
75%
4.5
```

Notice that R computes the quartiles differently than the textbook.

Outliers

Outliers are extreme observations in the data set. We give here a rough guide for how to identify them: Define the *interquartile range* as $IQR = Q3 - Q1$, where Q1 and Q3 are the first and third quartiles, respectively. The *suspected outliers* are the observations that are smaller than

$$Q1 - 1.5 \cdot IQR$$

and greater than

$$Q3 + 1.5 \cdot IQR.$$

Problem. Determine the suspected outliers in the following data: 1, 2, 3, 4, 2, 3, 1, 2, 5, 9, 10.

Solution. We have

```
> x=c(1,2,3,4,2,3,1,2,5,9,10)
> Q3=quantile(x,0.75)
> Q1=quantile(x,0.25)
> IQR=Q3-Q1
> I1=Q1-1.5*IQR
> I1
  25%
-1.75
> I2=Q3+1.5*IQR
> I2
 75%
8.25
> which(x<I1)
integer(0)
> which(x>I2)
[1] 10 11
> x[which(x<I1)]
numeric(0)
> x[which(x>I2)]
[1]  9 10
```

The suspected outliers have values greater than 8.25 and smaller than -1.75. It follows that the suspected outliers are observations number 10 and 11 which have values 9 and 10, respectively.

Explanation. The code can be explained as follows:

- The entry $x > I2$ returns a logical vector with entry **TRUE** if **x** is greater than the value stored in **I2** and **FALSE** if x is smaller. The command **which**$(x > I2)$ returns the locations in the vector x for which the entries are TRUE.
- The code **x**[**which**$(x > I2)$] returns the values of the observations stored in the vector x for which $x > I2$.

The five-number Summary

The *five-number summary* of a distribution shows in the following order: the smallest observation, the first quartile, the median, the third quartile, and the largest observation.

Problem. Find the five-number summary of the following set of observations: 1, 2, 3, 4, 2, 3, 1, 2, 5, 9, 10.

Solution. We have:

```
> x=c(1,2,3,4,2,3,1,2,5,9,10)
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   3.000   3.818   4.500  10.000
```

Notice here that R also computes the mean.

Boxplots

The graph of the five-number summary is called a *boxplot*. The horizontal line in the box marks the median and the box spans the first and third quartile. In most textbooks, the smallest and largest observations are marked by lines or whiskers extended out from the box. In R however, there are different ways to draw a boxplot. The entry **range=** is an option to add to the arguments in **boxplot()**. The option **range** specifies how far the lines are extended out from the box:

I. If the vale of **range** is positive, the lines extend to the most extreme observation that are within or on **range** times the interquartile range from the box. The default in R, is **range=1.5**. All observations that fall outside these range of values will be marked as outliers indicated by small circles (unless we turn that option off in R). Notice that the default value of range=1.5 is in accordance with the definition of suspected outliers given in the section about outliers.

II. If we put **range=0** then the lines extend to the extreme observations in accordance with the textbook definition provided above. In this case, no outliers will be marked in the plot.

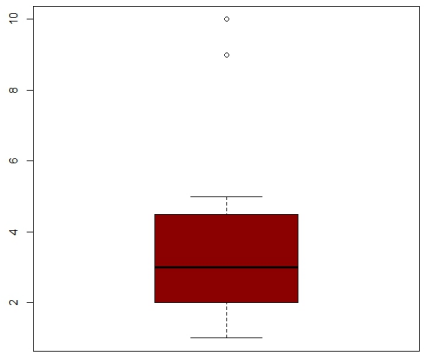
We illustrate these two ideas in the next two examples.

Problem. Plot a simple **Boxplot** of the observations 1, 2, 3, 4, 2, 3, 1, 2, 5, 9, 10 that shows the suspected outliers using the type of boxplot described in part I above.

Solution. We have

```
> x=c(1,2,3,4,2,3,1,2,5,9,10)
> boxplot(x,col="darkred")
```

We see that there are two suspected outliers in the data marked as circles. The suspected outliers have values 9 and 10. The lines are extended out from the box to the observations with values 1 and 5. The middle horizontal line is the median which has value 3. The box spans the first quartile which has value 2 and the third quartile which has value 4.5.

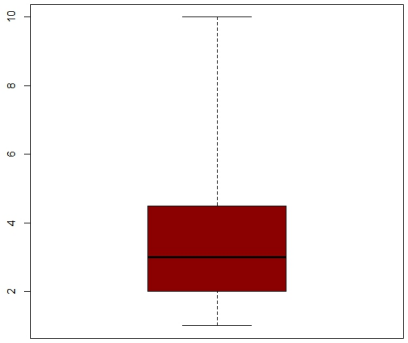


Problem. Plot a simple **Boxplot** of the observations 1, 2, 3, 4, 2, 3, 1, 2, 5, 9, 10 using the type of boxplot described in part II above. (This is the textbook definition of a boxplot).

Solution. We have

```
> x=c(1,2,3,4,2,3,1,2,5,9,10)
> boxplot(x,range=0,col="darkred")
```

Here the lines extend out from the box to the observations with minimum value 1 and maximum value 10.

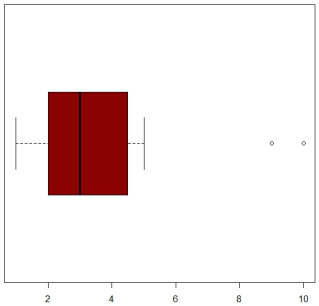


Notice here that R returns a vertical boxplot. To plot a horizontal boxplot, add the command **horizontal=TRUE**.

Problem. Plot a simple horizontal **Boxplot** of the observations 1, 2, 3, 4, 2, 3, 1, 2, 5, 9, 10.

Solution. We have

```
> x=c(1,2,3,4,2,3,1,2,5,9,10)
> boxplot(x,horizontal=TRUE,col="darkred")
```

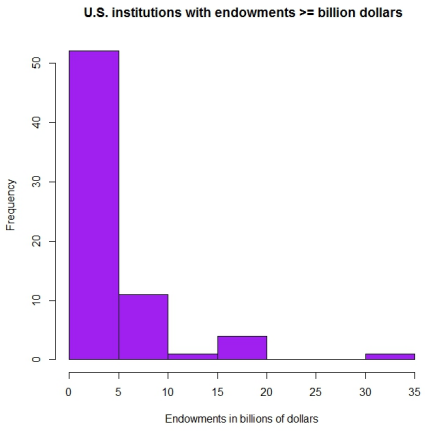


Some Practical problems

Problem. The histogram below shows the U.S. institutions of higher education with endowments greater than or equal to one billion dollars [2]. The data can be found in an external file called *Endowment.txt*. Here is part of the data:

```
   Endowment
1     30.435
2     19.345
3     18.264
4     17.036
..     ....
66     1.088
67     1.054
68     1.036
69     1.000
```

(A) Find the mean, median, variance, and standard deviation of the endowments.
(B) Determine if there are any suspected outliers in the data.
(C) Draw a boxplot of the distribution of endowments.



Solution to part (a). We have

```
> data=read.table("Endowment.txt",header=T)
> attach(data)
> mean(Endowment)
[1] 4.088522
> median(Endowment)
[1] 1.799
> var(Endowment)
[1] 27.50901
> sd(Endowment)
[1] 5.244903
```

We notice that the mean=4.09 is greater than the median=1.80. This can be explained by the right skewness in the distribution of endowments. The median is here a better measure of center of the distribution than the mean since the distribution is strongly skewed.

Solution to part (b). We will check for outliers:

```
> Q1=quantile(Endowment,0.25)
> Q3=quantile(Endowment,0.75)
> IQR=Q3-Q1
> I1=Q1-1.5*IQR
> I2=Q3+1.5*IQR
> I2
   75%
10.341
> outliers1=which(Endowment<I1)
> outliers2=which(Endowment>I2)
> outliers1
integer(0)
> outliers2
[1] 1 2 3 4 5
```

We see that the first five observations are suspected outliers.

Solution to part (c). We have

```
> boxplot(Endowment,ylab="Endowments",
+ main=" U.S. institutions with endowments >= billion dollars",col="pink")
```

We see from the boxplot that the data is strongly right-skewed since the third quartile is farther above the median than the first quartile is below the median.

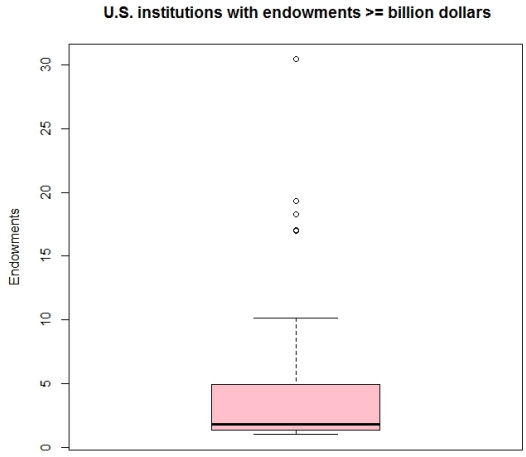


Figure 1: Boxplot showing the suspected outliers

We include here a boxplot defined as in part II above. (This is the textbook definition of a boxplot).

```
> boxplot(Endowment,range=0,ylab="Endowments",
+ main=" U.S. institutions with endowments >= billion dollars",col="pink")
```

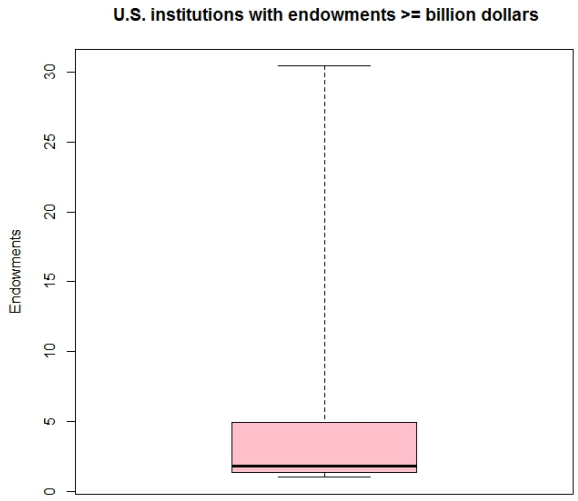


Figure 2: Boxplot with lines extending out from the box to the minimum and maximum value of the observations

Explanation. The code can be explained as follows:

- The command

  ```
  data=read.table("C:Chapter2/Endowment.txt",header=T)
  ```

  imports the data set from an external text file named **Endowment.txt** and store the data in the data frame **data**.

- The command **attach(data)** puts the data frame **data** into the search path such that we have access to the variables in the data frame.

- The entry **Endowment** $< I1$ returns a logical vector with entry **TRUE** if **Endowment** is smaller than the value stored in **I1** and **FALSE** if **Endowment** is greater. The command **which(Endowment**$< I1$**)** returns the indexes of the observations for which the entries are TRUE.

Problem. R has the built-in data set called **PlantGrowth** [1]. The data in **PlantGrowth** shows the results from an experiment providing the dried weight of plants obtained under the condition of two different treatments and a control. Here is part of the data:

```
   weight group
1    4.17  ctrl
..    ...   ...
10   5.14  ctrl
11   4.81  trt1
..    ...   ...
20   4.69  trt1
21   6.31  trt2
..    ...   ...
30   5.26  trt2
```

Construct a five number summary of the distribution of weight for each of the three groups.

Solution. We have

```
> PlantGrowth
> Group1=PlantGrowth$weight[PlantGrowth$group=="ctrl"]
> Group2=PlantGrowth$weight[PlantGrowth$group=="trt1"]
> Group3=PlantGrowth$weight[PlantGrowth$group=="trt2"]

> summary(Group1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.170   4.550   5.155   5.032   5.292   6.110
> summary(Group2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.590   4.208   4.550   4.661   4.870   6.030
> summary(Group3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.920   5.268   5.435   5.526   5.735   6.310
```

Explanation. The code can be explained as follows:

- The command **PlantGrowth** returns the data stored in the dataframe **PlantGrowth**.
- The command

  ```
  Group1=PlantGrowth$weight[PlantGrowth$group=="ctrl"]
  ```

  extracts the value of the variable **weight** for which **group** is ctrl. We name this vector of observations for **Group1**.

Problem. Construct a boxplot of the distribution of weight for each of the three groups.

Solution. We have:

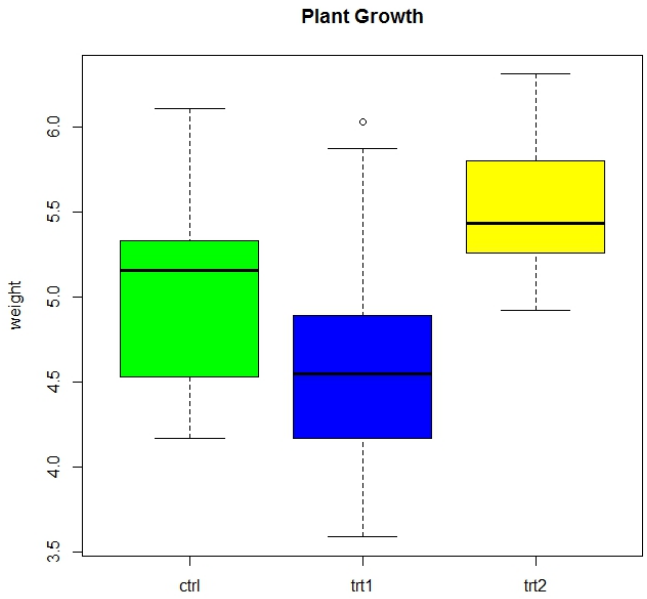
```
> boxplot(weight~group,data=PlantGrowth,ylab="weight",
+ main="Plant Growth",col=c("green","blue","yellow"))
```

Explanation. The code can be explained as follows:

- The commmand

```
boxplot(weight ~ group,data=PlantGrowth)
```

returns a boxplot for the distribution of *weight* versus *group* for data in the dataframe **PlantGrowth**. Thus, we make a boxplot of the distribution of *weight* for each of the three groups in **group**.



The median weight for plants in group trt2 is larger than the median weight for plants in group trt1 and ctrl. The plants in group trt1 have the smallest median weight. The plants in group trt2 also have larger first and third quartiles as well as larger maximum value than the plants in the other two groups. Plants from group trt1 have the smallest first and third quartile as well as the smallest minimum value. The distribution of weight for group ctrl is left-skewed and the distribution of weight for group trt2 is right-skewed.

References

[1] A. J. Dobson, *An Introduction to Statistical Modelling.* London: Chapman and Hall, 1983

[2] National Association of College and University Business Officers and Commonfund Institute (NACUBO), *U.S. and Canadian Institutions Listed by Fiscal Year 2012 Endowment Market Value and Percentage Change* in Endowment Market Value from FY 2011 to FY 2012*, 2013

00