In this project we investigate the viewing of datasets using the quantiles (min,Q_1, Median,Q_3,max) and the graphical display of the boxplot.

1 Five number summary

The definition the quantiles – the minimum, the first quantile Q_1, the median, the third quantile Q_3 and the maximum – given in the book correspond to the `fivenum()` command which gives the "five-number summary".

To illustrate, we can use the built-in data set `rivers` which list the lengths of 141 "major" rivers in North America. To access a built-in data set we use the `data()` function as follows

```
> data(rivers)
> rivers
  [1]  735  320  325  392  524  450 1459  135  465  600  330  336
  ... many are cut out...
> fivenum(rivers)
[1]  135  310  425  680 3710
```

The output shows the shortest river is 135 miles, the longest 3710. As well half the rivers are less than 425 miles long, and 1/2 are between 310 and 680 miles long.

Exercise 1.1 The number of cell phone minutes used per month by a class of 15 students is

`0, 250, 500, 0, 450, 1000, 300, 150, 500, 250, 400, 450, 0, 0, 800`

To enter these into the computer we use the `c()` command as

```
> minutes = c(0, 250, 500, 0, 450, 1000, 300, 150, 500,
+ 250, 400, 450, 0, 0, 800)
```

Enter in the minutes as shown (If you type all the data on one line the `+` won't appear.) and then answer using the computer these questions:

1. What is the most number of minutes? The least?
2. What is the median amount of cell phone minutes?
3. What is the mean amount of cell phone minutes? (Use `mean()`)

Exercise 1.2 The dataset `discoveries` lists the number of "great" inventions and scientific discoveries in each year from 1860 to 1959. Load it in with the command `data(discoveries)`.

1. What is the fewest number of great discoveries per year? The greatest?
2. What is the median number of great discoveries?
3. Fifty percent of the years have discoveries between what and what?

Exercise 1.3 The built-in data set `state` contains a data vector `state.area` which lists the area of the 50 US states in square miles. Load it with `data(state)`.

1. What is the largest state size? The smallest?
2. What is the median state size?
3. Recall the **percentile rank** of a data point is found by the formula

$$\frac{\text{number less than the data point} + 1/2\text{number equal the data point}}{\text{total number of data points}}.$$

For example, the percentile rank of New Jersey with a state area of 7836 is

```
> top = (sum(state.area < 7836) + 1/2* sum(state.area == 7836))
> bottom = length(state.area)
> top/bottom * 100
[1] 9
```

Find the percentile rank of New York with an area of 56400 square miles.

1.1 Mean and standard deviation vs. median and IQR

Recall, the sample mean and standard deviation as well as the median and IQR are used to describe center and spread. The mean and standard deviation are found with `mean()` and `sd()`, whereas the median and IQR are found with `median()` and `IQR()`. We find both pairs for the `rivers` data set with

```
> mean(rivers)
[1] 591.2
> sd(rivers)
[1] 493.9
> median(rivers)
[1] 425
> IQR(rivers)
[1] 370
```

We see that the center and spread given by the mean and standard deviation are both much larger. This is because this data set is skewed with a heavy right tail.

Exercise 1.4 A sample of size 10 from the 2002 salaries of the top 200 CEO's in America gives this data sets (in 10,000's of dollars)

`312, 316, 175, 200, 92, 201, 428, 51, 289, 1126, 822`

Enter this in as follows

```
> ceo = c(312, 316, 175, 200, 92, 201, 428, 51, 289, 1126, 822)
```

Compare the center and spread given by the mean/standard deviation and the median/IQR. Is there a difference? Why?

2 Boxplots

A boxplot is a very nice graphical display of the five-number summary. It shows the range, the center and spread, and the skew of a data set in a compact manner that allows for multiple data sets to be compared. The command to make a boxplot is `boxplot()`.

For example, a boxplot (figure 1) of the rivers is made with

```
> boxplot(rivers)
```

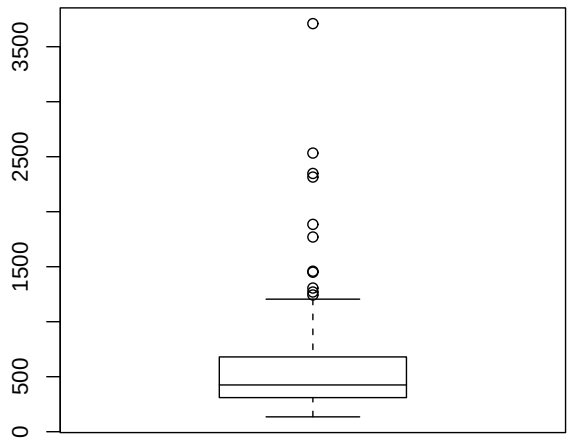

Figure 1: Rivers boxplot `boxplot(rivers)`

Recall the box is drawn to show Q_1, the median and Q_3. The whiskers extend to the min and max *unless* these are too far away (1.5 IQR's) in which case the points are plotted separately.

Notice how you can see the skew of the boxplot with the few long rivers plotted as points. This is why the mean and standard deviation are much bigger than the median and IQR, as a few really large values can throw those off.

Exercise 2.1 Make a boxplot of the `state.area` data set. Does it show any skew? Would you expect the mean to be much different than the median? Check.

Exercise 2.2 Make a boxplot of the CEO data. Does it show much skew?

3 *z*-scores

Recall the definition of a *z*-score for a dataset. For a data set $x_1, x_2, \ldots, x_n$ with mean $\bar{x}$ and standard deviation s the z score of x_i is

$$z_i = \frac{x_i - \bar{x}}{s}$$

These are found easily in R. For example, the CEO dataset has *z*-scores of

```
> (ceo - mean(ceo))/sd(ceo)
 [1] -0.1615 -0.1492 -0.5810 -0.5045 -0.8352 -0.5014  0.1938
 [8] -0.9608 -0.2319  2.3313  1.4004
```

(The same can be done with the `scale()` function, but the output isn't as pretty.)

Exercise 3.1 Find the *z*-scores of the rivers dataset and store them in a variable. Make a boxplot of your result. Does it look like figure 1?

A useful thing with *z*-scores is that they allow one to *compare* two dataset's shape when the scales are different. For example, the a sample of CEO pay in 2000 is given by this data

```
110, 12, 2.5, 98, 1017, 540, 54, 4.3, 150, 432
```

Exercise 3.2 Enter in the CEO data for the year 2000 as follows

```
> ceo.2000 = c(110, 12, 2.5, 98, 1017, 540, 54, 4.3, 150, 432 )
```

Then find the *z* scores of each. For example, for the 2002 data

```
> z.2002 = (ceo - mean(ceo))/sd(ceo)
```

Store the 2000 year *z*-scores in the variable `z.2000` and then compare with side-by-side boxplots using

```
> boxplot(z.2000,z.2002)
```

Are the shapes similar?