BE540 - Introduction to Biostatistics
Computer Illustration

Topic 1 – Summarizing Data
Software: R

A Visit to Yellowstone National Park, USA

Source:
Chatterjee, S; Handcock MS and Simonoff JS *A Casebook for a First Course in Statistics and Data Analysis.* New York, John Wiley, 1995.

Setting:
Upon completion of BE540, you decide to take a vacation to the United States. Of particular interest is seeing an eruption of the famous "Old Faithful" geyser at Yellowstone National Park. Unfortunately, your time is limited and you do not wish to miss seeing an eruption.

This worked example illustrates descriptive analysis of a data set of 222 interval times between eruptions of the Old Faithful Geyser, measured during August 1978 and 1979.

Data File:
GEYSER1.DAT - This is a data set in ASCII format.

Description of Data:
There are three variables, in the following order:

INDEX - An index of the date of the eruption. We will not be using this variable.

DURATION - The duration of the eruption in minutes.

INTERVAL - The length of the interval between the current eruption and the next eruption.

Objective:
Describe the pattern of eruptions and predict the interval of time to the next eruption.

Before you begin:

About R software:

IR is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.

R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.

You can download R for free at **http://www.r-project.org/**

Preliminaries:

(1) To do this illustration yourself, launch R in a new window. Then, as you proceed, use "copy and paste" from this illustration into the R command line. Press "Enter" to execute.

(2) R commands, and instructions to press the "enter" key are in blue

(3) Note: Where you see green, you need to replace what is written in this illustration with your own input.

(4) R commands that begin with the number symbol, #, are comments.

1. Read in the data.

Download and save the data file into a folder.

```
# To read in the data file use the following command. Note that the path to the
# actual file has forward slashes (/) and not (\)

geyser.data <- read.table('C:/Users/YourPathToDataFile/geyser1.dat',
          sep='', col.names=c('index', 'duration', 'interval'))
```

To view your data set that you read into R type

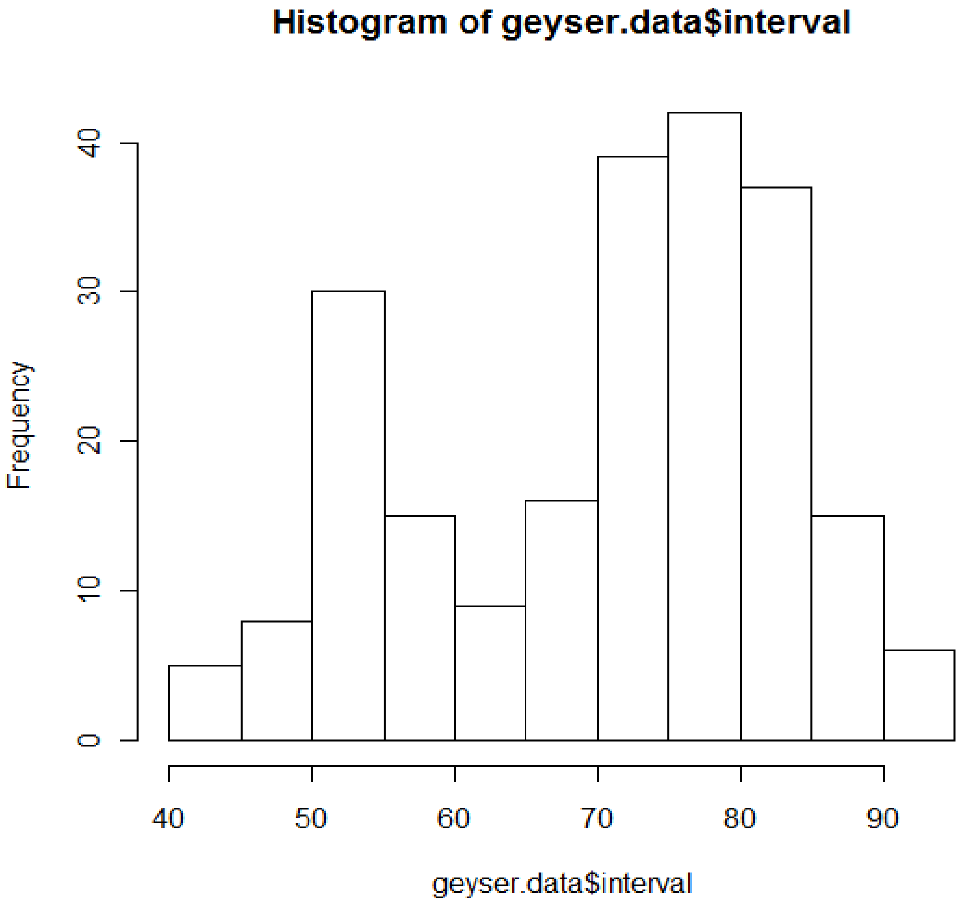
```
geyser.data
```

then press **"Enter"**

2. Obtain a Histogram of Interval Times.

```
# Obtain a histogram of  interval times
hist(geyser.data$interval)
```

You should see



Remarks

The interval times are in the range of 40 to 100 minutes, approximately.
There appears to be two groupings of interval times.
They are centered at 55 and 80 minutes, approximately.
Interestingly, there is a gap in the middle.

3. Save this histogram as a picture that you can print directly or that you can insert into a document such as this one.

TIP: **To save a copy of the histogram right-click on the histogram and press "cntl+C". Then paste it into any word processing document**.

4. Instead of a histogram, we might have constructed a stem-leaf diagram.

```
stem(geyser.data$interval)
```

You should see:

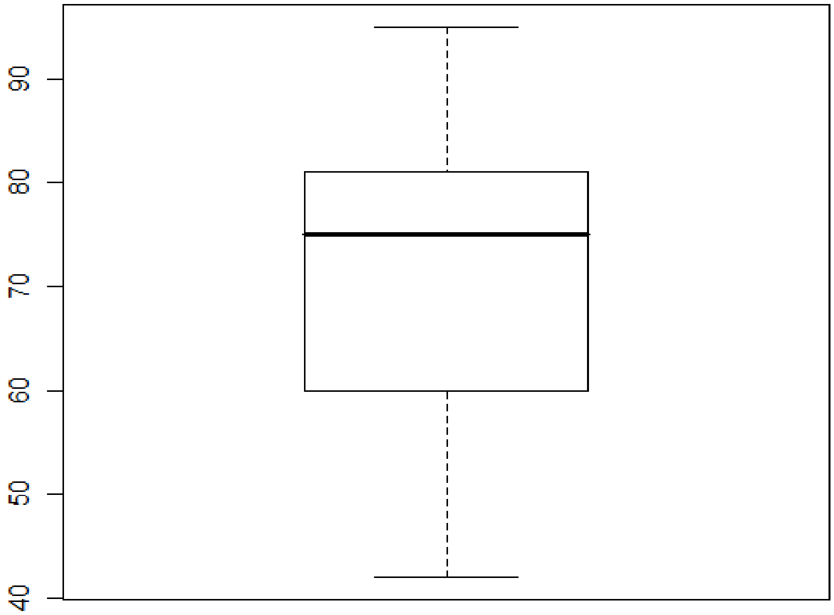
```
  The decimal point is 1 digit(s) to the right of the |

  4 | 234
  4 | 55788999
  5 | 0011111111111111222333334444
  5 | 555566677778889
  6 | 0000111112223
  6 | 66677788999
  7 | 00000111112222233333333344444
  7 | 55555555555555566666666667777777788888889999
  8 | 0000000000000111111111222222222233333333444444444
  8 | 5666666788899
  9 | 00011134
  9 | 5
```

Remarks.
You can see that a stem and leaf diagram is very similar to a histogram. However, we can also see that the minimum and maximum interval times are 42 and 95 minutes, respectively, and that the median time is 75 minutes.

5. In this example, a Box and Whisker plot is not very informative.

```
# Obtain a box and whisker plot of interval times
boxplot(geyser.data$interval)
```



Remarks:

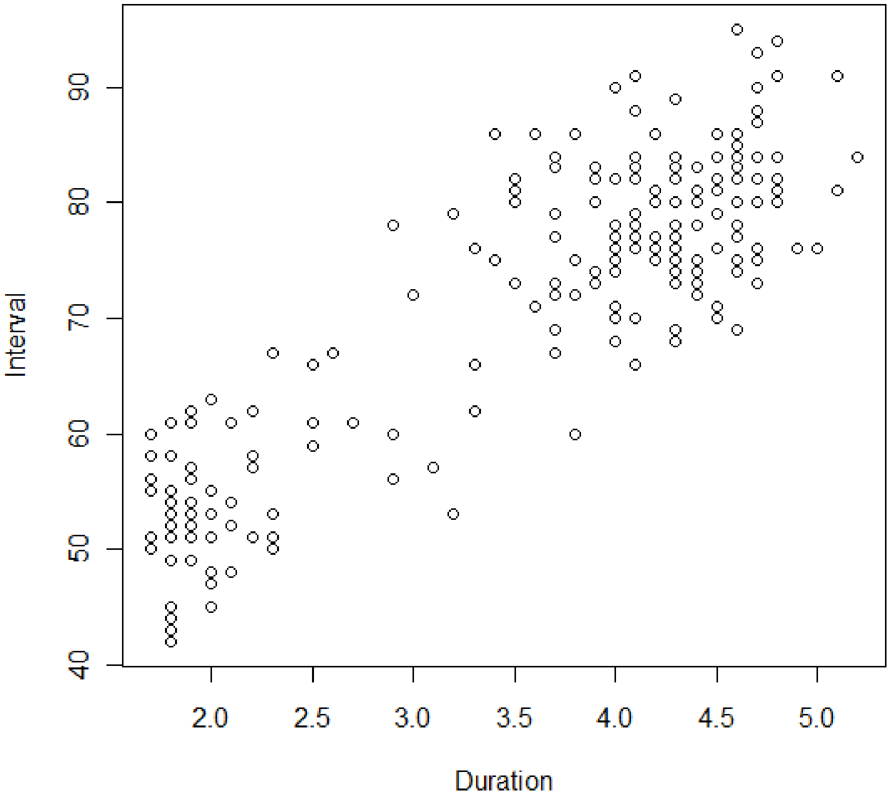
Both the histogram and stem and leaf summaries suggested that there are two groups of interval times. This cannot be seen in a Box and Whisker plot.

Box and Whisker plots are excellent for summarizing the distribution of ONE population. They are not informative when the sample being summarized actually represents MORE THAN ONE population.

6. We have information on duration of eruption also. One possibility is that the duration of the current eruption is a predictor of the interval time to the next eruption. To investigate this possibility, construct a scatter plot of interval time versus duration. Plot the predictor DURATION on the horizontal axis (X) and the outcome INTERVAL time to the next eruption on the vertical axis (Y).

```
# Obtain a scatter plot of duration vs. interval
plot (geyser.data$duration, geyser.data$interval, xlab='Duration', ylab='Interval')
```

You should see:



Remarks

The scatter plot confirms a suspected positive association. Longer duration times appear to predict longer intervals to the next eruption. Interestingly, the scatter plot still suggests that there are two distinct subgroups, distinguished by durations of less than versus greater than three minutes.

7. Create two subgroups based on duration and construct separate box and whisker plots of interval times for the interval times that follow eruptions less than 3 minutes in duration and the interval times that follow eruptions greater than 3 minutes in duration.

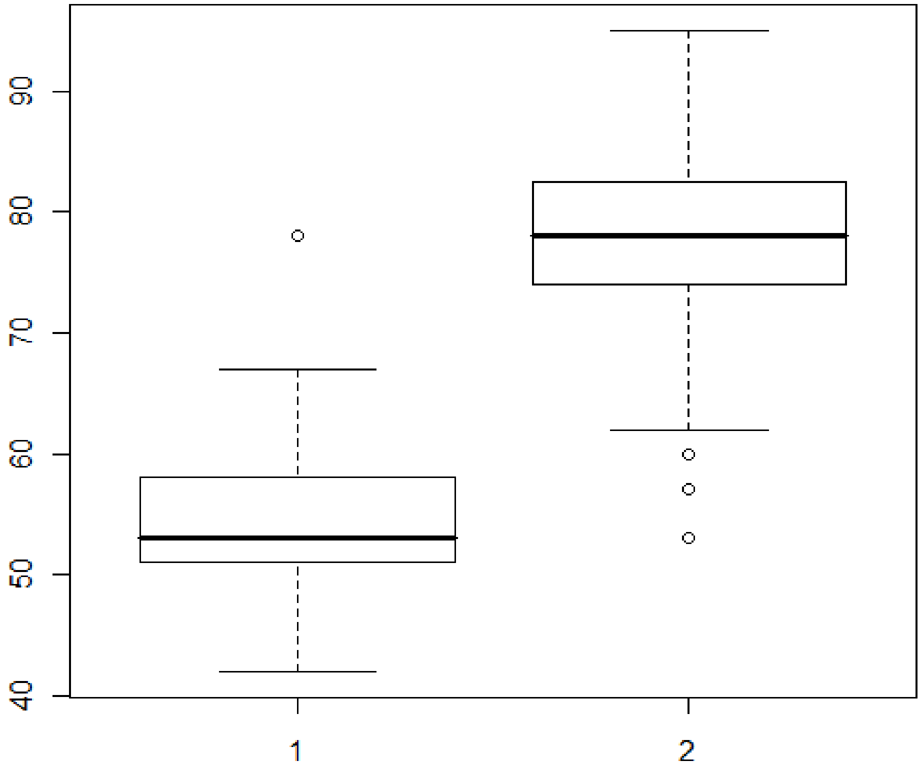
```
# Create two subgoups of geyser data based on duration: 1)-Duration less than 3 minutes
                                                       2)-Duration over 3 minutes


geyser.data1 <- subset(geyser.data,duration<3)
geyser.data2 <- subset(geyser.data,duration>=3)

# Plot box and whisker plots of the two groups side by side

boxplot(geyser.data1$interval, geyser.data2$interval)
```

You should see:



10. Finally, let's look at some numerical summaries, separately for the two groups.

```
# Obtain summary statistics on both groups

summary(geyser.data1$interval)
```

You should see:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  42.00   51.00   53.00   54.46   58.00   78.00
```

```
summary(geyser.data2$interval)
```

You should see:

```
   Min. 1st Qu.  Median    Mean  3rd Qu.    Max.
  53.00   74.00   78.00   78.16   82.50    95.00
```

```
# Alternatively, more descriptive summary can be obtained using
# "describe" command in library "psych".  First, install the R package "psych"

install.packages('psych')

# then invoke the psych library

library(psych)

# Now you can use the "descibe" command to obtain statistics

describe(geyser.data1$interval)
```

You should see:

```
  var  n  mean  sd median trimmed  mad min max range skew kurtosis   se
   1   67 54.46 6.3     53   54.18    4.45  42   78    36     0.81     1.55     0.77
```

```
describe(geyser.data2$interval)
```

You should see:

```
  var   n   mean   sd   median trimmed  mad min max range  skew kurtosis   se
  1    155 78.16  6.89     78   78.22      5.93  53   95    42     -0.33     0.92    0.55
```

So, what should you do? If you arrive to Old Faithful just after an eruption of less than 3 minutes, with 95% confidence, your waiting time to the next eruption will be between 53 and 56 minutes. Alternatively, if you arrive just after an eruption of greater than 3 minutes, with 95% confidence, your waiting time to the next eruption will be between 77 and 79 minutes.

Here is the entire program:

```
geyser.data <- read.table('C:/Users/YourPathToDataFile/geyser1.dat',
          sep='', col.names=c('index', 'duration', 'interval'))


hist(geyser.data$interval)


stem(geyser.data$interval)


boxplot(geyser.data$interval)


plot (geyser.data$duration, geyser.data$interval, xlab='Duration', ylab='Interval')


geyser.data1 <- subset(geyser.data,duration<3)
geyser.data2 <- subset(geyser.data,duration>=3)


boxplot(geyser.data1$interval, geyser.data2$interval)



summary(geyser.data1$interval)
summary(geyser.data2$interval)


install.packages('psych')

library(psych)


describe(geyser.data1$interval)
describe(geyser.data2$interval)
```