Solutions to Assignment #2

1. Read the data contained in the MS Excel file Westvaco.xls, which may be downloaded from`http://pages.pomona.edu/~ajr04747`, into a dataframe that you may name `WestvacoD`. Define variables `Age` and `RIF` as described in Appendix A in the class notes at `http://pages.pomona.edu/~ajr04747`. Obtain vectors `laidoff` and `kept` containing ages of laid off workers and kept workers, respectively. Print a box plot showing the distributions of the two variables and write a paragraph comparing the two distributions. Does the picture show enough evidence to conclude that the company discriminated against older workers?

   ***Solution**:* Source the codes

   ```
   laidoff <- array(dim = 0)
   L <- length(Age)
   for (i in 1:L) if (RIF[i]>=1) laidoff <- c(laidoff,Age[i])
   ```

   and

   ```
   kept <- array(dim = 0)
   L <- length(Age)
   for (i in 1:L) if (RIF[i]==0) kept <- c(kept,Age[i])
   ```

   to generate vectors `laidoff` and `kept` containing ages of laid off workers and kept workers, respectively. Note that these codes assume that you have extracted the vectors `Age` and `RIF` from the Westvaco data set.

   Typing

   ```
   boxplot(laidoff,kept, names=c("Laid off", "Kept"), ylab="Age")
   ```

   in R yields the picture of the box plots for `laidoff` and `kept` in Figure 1. The figure gives a strong graphical indication of an age bias against older workers by Westvaco. However, this statement needs to be corroborated by a test of significance. □

2. Write an R script which picks out those workers that get paid hourly from the `WestvacoD` dataframe and puts their ages into a vector called `hourly`. Test the script and explain the procedure you used. Do the same for the salaried workers putting their ages into a vector named `salaried`. Plot box plots of the distributions of the two variables in the same graph and compare. Is there any significant difference between the two distributions?

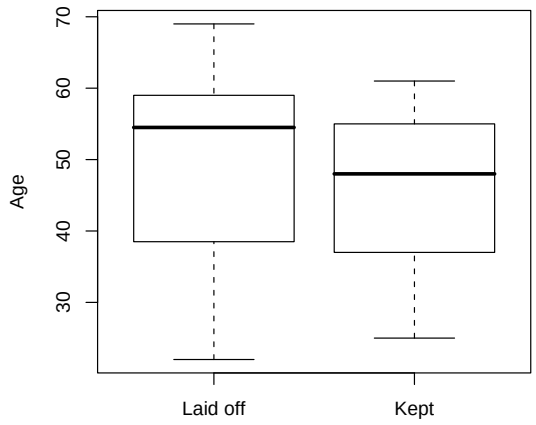

Figure 1: Comparing distribution of ages of laid off versus kept workers

***Solution**:* Begin by picking out the `Pay` column form the Westvaco data set by typing

```
Pay <- WestvacoD$Pay
```

Then, source the code

```
hourly <- array(dim = 0)
L <- length(Age)
for (i in 1:L) if (Pay[i]=="H") hourly <- c(hourly,Age[i])
```

The code will loop through the `Pay` vector and check if it has an "H" in each cell. If so, the age associated to that worker gets appended to the `hourly` vector. The corresponding code for the salaried workers is

```
salaried <- array(dim = 0)
L <- length(Age)
for (i in 1:L) if (Pay[i]=="S") salaried <- c(salaried,Age[i])
```

The graph comparing the boxplots for the two variables is shown in Figure 2. Even thought he median age for the horly workers is

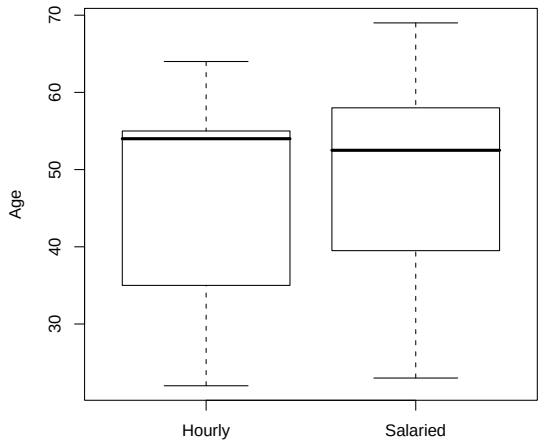

Figure 2: Comparing distribution of ages of hourly versus salaried workers

higher than that of the salaried ones, salaried workers tend to be older than the hourly ones. There is a wider range of ages for the salaried workers; however, the inter–quartile range for the hourly workers is wider than that of the salaried ones. □

3. Write R scripts that pick out the workers that get laid off in a rounds 1, 2, 3, 3, 4 and 5, and puts them into vectors `round1`, `round2`, ..., `round5`, respectively. Plot box plots of the distributions of all the variables in the same graph. Discuss the picture.

Solution: To pick out the workers that were laid off in the first round, source the code

```
round1 <- array(dim = 0)
L <- length(Age)
for (i in 1:L) if (RIF[i]==1) round1 <- c(round1,Age[i])
```

This generates the vector `round1` containing the ages of the those workers. In the same way we can generate vectors for the other rounds by changing the condition `(RIF[i]==1)` accordingly.

Figure 3 is a graph of the box plots for all the rounds.

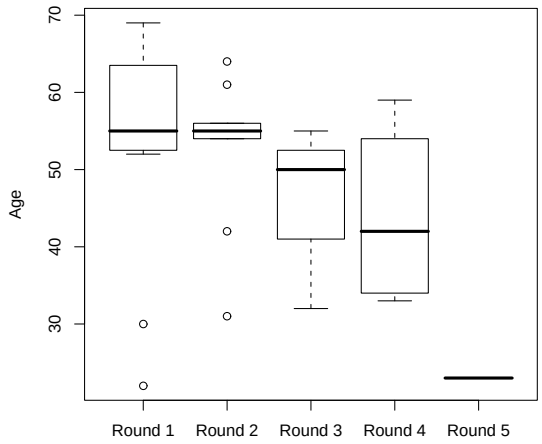

Figure 3: Comparing distribution of ages of laid off workers at various rounds

The box plots in Figure 3 show that the median ages of the laid off workers in the first two rounds is quite high (around 55). There is a preponderance of high ages in the first round of layoffs. By the time of the fourth round, the median age of laid off workers has diminished (primarily because many of the older workers have been laid off in previous rounds) and the distribution slightly more even. □

4. In a study reported by Roseann M. Lyle et al. in the *Journal of the American Medical Association,* Vol. 257 (1987), pp. 1772-1776, titled *Blood Pressure and metabolic effects of calcium supplementation in normotensive white and black man*, results of an experiment testing the effects of calcium supplementation on the decrease of systolic blood pressure were reported. The experiment involved 21 black men who were randomly divided into a group of 10 and a group of 11. The first group of men was given calcium supplement pills for 12 weeks, and the second group (the *control* group) received a placebo pill identical to the calcium supplement also for 12 weeks (this is an example of a *randomized comparative experiment*). The men were all instructed to take one pill a day. The men involved in the experiment did not know which pill they were taking. Also, the people administering the pills and those measuring the subjects' blood pressure did not know which pill the men were taking (this is known as a *double blind* experiment). Systolic blood pressures were measured at the beginning and end of the experiment. The data from the experiment are recorded in the MS

Excel file, `CalciumBloodPressureData.xls`, which may be downloaded from `http://pages.pomona.edu/~ajr04747` (follow the link to Data Files in the Math 58 section). Decease in the systolic blood pressure are also recorded in the last column of the data set (the one labeled `dec`). Note that a negative value in the decrease columns indicates that the blood pressure actually went up.

Import the table in the spreadsheet to a dataframe in R which you may name `CalciumBloodPressureD`. Operate on the dataframe to obtain variables `group` and `dec`. Pick out the `dec` values for those subjects in the `calcium` group, and those for the `placebo` group. Put them into vectors called `calcium` and `placebo`, respectively.

Compute number summaries for each of the variable `calcium` and `placebo`. Is there any difference in the decrease in blood pressure in the two groups?

***Solution**:* The codes

```
calcium <- array(dim = 0)
L <- length(dec)
for (i in 1:L) if (group[i]=="Calcium") calcium <- c(calcium,dec[i])
```

and

```
placebo <- array(dim = 0)
L <- length(dec) for (i in 1:L)
if (group[i]=="Placebo") placebo <- c(placebo,dec[i])
```

will do the trick. The numerical summaries for the two variables are:
for the Calcium treatment group

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-5.00   -2.75    4.00    5.00   10.75   18.00
```

and for the control group

```
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-11.00 -3.00     -1.00     -0.64   1.00       12.00
```

Both the mean and the median for the treatment group are considerably higher than those for the placebo croup. So the data seems to

suggest that calcium supplementation does have some effect on the decrease of blood pressure. □

5. Plot box plots of `calcium` and `placebo` in the same graph. Print the plot and discuss your observations. Would you say that calcium supplementation has an effect in the decrease of blood pressure?

   Solution: The picture of the box plots is in Figure 4.

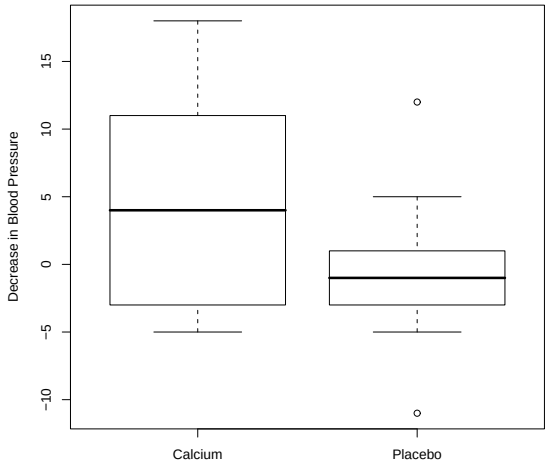

Figure 4: Comparing decrease in blood pressure between calcium treatment versus placebo groups

The box plots in Figure 4 suggest strongly that calcium supplementation does have some effect on the decrease of blood pressure. □