2019 Statistics Graduate Bootcamp
University of California, Irvine

TA: Derenik Haghverdian
Department of Statistics

Updated - September 11, 2019

1. (a) Load the `PlantGrowth` data set in `R` with the command

   ```
   data("PlantGrowth")
   ```

   You can read about the data set with the command

   ```
   ?PlantGrowth
   ```

   (b) Find the mean of the three treatment groups using the `aggregate` function.

   (c) Similarly, find the sample standard deviation of the three treatment groups.

   (d) Use side-by-side boxplots to compare the distributions of the three treatment groups.

   (e) Write a function in `R` to perform a two-sample t-test to test whether the means from group 1 and group 2 are significantly different under the assumption of equal variance. Use the following function definition:

   ```
   my.ttest <- function(grp1, grp2, alpha = 0.05) {
   # Code to perform calculations goes here...
   ...
   # Return these values
   return(data.frame(t.diff, P, df)
   }
   ```

   (f) Use your t-test function to test whether the plants in treatment group 1 had significantly different growth on average compared to the plants in the control group.

   (g) Compare the results from the previous part to the results given by the `t.test` function with the argument $var.equal = TRUE$.

   (h) Plot the reference distribution and indicate the value of the test statistic on the plot.

   (i) State the conclusion of the hypothesis test in context.

   (j) Fit a linear regression model to test for a significant difference in mean growth across the treatment 1 and control groups using `fit <- lm(weight ~ group, data=PlantGrowth)`

2. (a) Load in the baseball players data set that is in the ISLR library and call it *dat*. You can install the ISLR library using the `install.packages("ISLR")` command.

   (b) Find out how many observations are in the data set by `nrow(dat)`. How many variables are in the data set?

   (c) We are interested in understanding the relationship between player salary and different measures of player performance. Plot a histogram of the salary variable. Use the `main` option in the plot function to set a title for the plot. What do you notice about the distribution of the salary variable?

   (d) Generate a histogram for log(`Salary`). What do you notice about the distribution of log(`Salary`)?

(e) Create a new variable `logSalary` = log(`Salary`) and add it to `dat`.

(f) Use the `scatterplotMatrix` function in the `car` package to generate a scatterplot matrix for the variables `logSalary, HmRun, Hits, RBI, Errors`. What variables appear to have a significant linear association with `logSalary`? Do any variables have a significant non-linear association with `logSalary`?

(g) Generate a scatter plot of `logSalary` against `Hits`. Add a title and change the axis labels. Run `?plot` for help to modify plot elements.

(h) Fit a linear regression model to determine if there is a significant linear association between `logSalary` and `Hits`. Run the following code.

```
fit <- lm(logSalary ~ Hits, data=dat)
summary(fit)
plot(dat$logSalary ~ dat$Hits) # Add in code for title and labels
abline(fit)
```

(i) Compute a 95% confidence interval for the `Hits` coefficient. Use the degrees of freedom and standard error given by `summary(fit)`. Verify your calculations using `confint(fit)`.

(j) Diagnostic plots for a linear model fit can be generated easily by `plot(fit)`. Run this and examine the first two plots for signs of violation of the linear regression assumptions.