STA 303H1S / 1002 HS -Winter 2018 Assignment # 1
"In or Out"

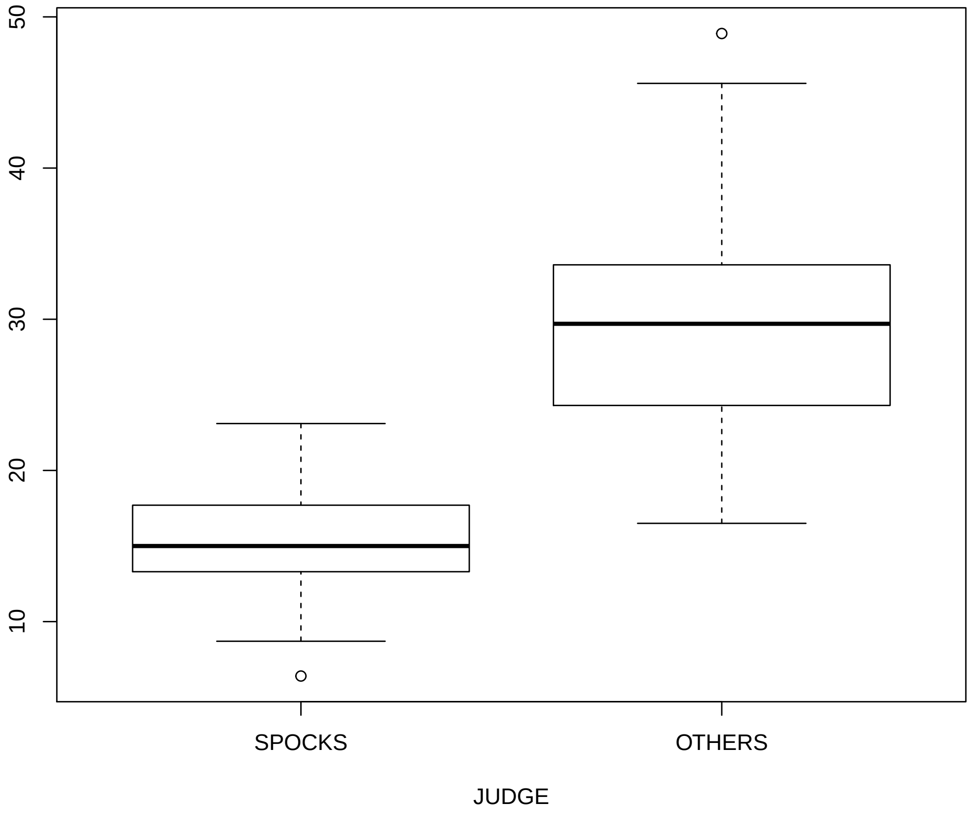
Due: In Crowdmark via Blackboard by 10pm on Thursday, January 25, 2018.
Late assignments will be subjected to a penalty of 5% per hour late.

Grading: There are 2 main questions. The grand total for this assignment is 100 marks.

Instructions:

- Use R (or R Studio) to do the analysis for the following questions.
- Use a benchmark significant level of 5%.
- Compile your solution as a PDF document (Word, LaTeX or Rmarkdown can be your base).
- Presentation of solutions is very important. Your assignment should have two main sections- Solutions and Appendix. Include relevant plots and quote relevant numbers from your R output for your solutions. In the Appendix, include your R code and other output. Marks will be awarded for excellent presentation.
- Write and submit **your own work**. For instance, personalized your code as much as possible, using your first name. **All plots produced must be given a title with the last 4 digits of your student number**.
- Where appropriate, your answers are expected to be written in plain English.

1. (30 marks) Consider the box plot below, drawn in R, based on the data in the file "juries.csv".



(a) (10 marks) Recall the 1.5IQR Rule which is used to identify potential outliers. Show, using this rule, how the two points identify as outliers.

(b) (15 marks) Recreate the side-by-side box plots of percent of women on venires for Spock's judge and the other judges **without identifying outliers**.

(c) (5 marks) Comment on the difference between the schematic box plot (which does not identify outliers) and the modified box plot (which identifies outliers).

2. (70 marks) Consider the data, "assign1data.csv" based on the heights of 166 students in our class and answer the questions that follow. The variables in the dataset are:

- `id`- an identification number from 1 to 166
- `height`- height in inches
- `sex`- sex of student

(a) (5 marks) Was the data based on an experiment or an observational study? Briefly discuss the limitations on the statistical inference we can draw from this data.

(b) (5 marks) Which variables are categorical? Name the levels of each categorical variable.

(c) (20 marks) Conduct an appropriate hypothesis test to determine whether there is a difference between the heights of *Males* and *Females*. Include the following:

i. Side-by-side boxplots
ii. Null and Alternative Hypotheses
iii. A test statistic and it's distribution
iv. Test assumptions
v. Test diagnostics (checking model assumptions)
vi. P-value
vii. Results (brief discussion and conclusion)

(d) (5 marks) Name two(2) statistical methods which are equivalent to your method used in part (c) above.

(e) (25 marks) *Create a subset of the data by removing the row of observations whose 'id' matches the last 2 digits of your student number*. For instance, this can be done in R by `shivon.subset` $<-$`shivon.data`[$-$`100`,] if my student number ends with '00'.
Then redo the analyses of part (c) above with your data subset.

(f) (10 marks) Compare your results of part (c) and part (e). Do you think that the observation removed was influential?