ST 314 Data Analysis 06

Topics:

- Single Factor ANOVA and Multiple Comparisons Procedures
- Describing Visual Displays
- Lessons Covered: 31- 33
- Textbook Chapter (Optional): 10

Grading:

- Points are listed next to each question and should total 25 points overall.
- Grading will be based on the content of the data analysis as well as the overall appearance of the document.
- Late assignments will not be graded.

Deadlines:

- Final Submission: **Monday, December 1st**. All submissions must be PDF files.

Instructions:

- Clearly label and **type answers** to the questions on the proceeding pages in Word, Google Docs, or other word processing software.
- Insert **diagrams or plots as a picture** in an appropriate location.
- Math Formulas need to be typed with Math Type, LaTeX, or clearly using key board symbols such as +, -. *, /, sqrt() and ^
- Submit assignment to the Gradescope link as a PDF. Indicate the pages to the individual questions and also verify the correct document has been uploaded. Failing to follow this direction may result in point deductions.

Allowances:

- You may use any resources listed or posted on the Canvas page for the course.
- You are encouraged to discuss the problems with other students, the instructor and TAs, however, all work must be your own words. Duplicate wording will be considered plagiarism.
- Outside resources need to be cited. Websites such as Chegg, CourseHero, Koofers, etc. are discouraged, but if used need to be cited and used within the boundaries of academic honesty.

ST 314 Data Analysis 06

Part 1. (25 points)
Single Factor ANOVA is a method we use when we want to compare a quantitative variable among more than two categories. It evaluates whether the means of different treatment groups, or populations, are equivalent. When we only have two populations then we can perform a two-sample t procedure, but when we have more populations we need to examine the data with Single Factor ANOVA.

In the R script `DA6_Single_Factor_ANOVA.R`, follow along with the analysis that compares average number of roommates between majors for the ST314 online students. You will need to upload the student information dataset from Canvas.

Once you have reviewed the example analysis, conduct your own by choosing one of the following three options:

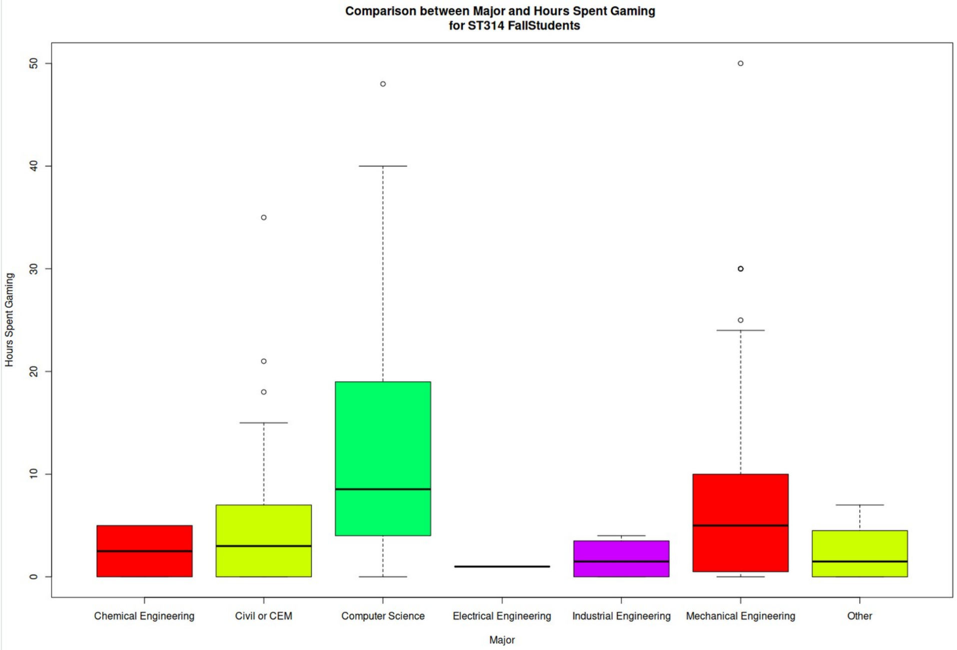
Option 1: Is there evidence average number of terms a student has been studying at OSU differs between majors for ST314 students?

Option 2: Is there evidence average weekly gaming hours per week differs between majors for ST314 students?

Option 3: Is there evidence average anticipated salary differs between majors for ST314 students?

For the option you selected, answer the questions below. Use a significance level of 0.10.

a. **(3 point)** Create side-by-side boxplot of the data and add color and a title to your plot. Paste your plot.



b. **(2 point)** From the side-by-side box plot, does there look to be a difference between the averages? Explain your reasoning.

There does appear to be a significant difference in the averages of time spent playing video games depending on major. Computer Science majors appear to have the highest average, followed by Mechanical Engineering majors. The remaining majors differ by less, but Civil or CEM Engineering appears to hold third place.

c. **(2 point)** State the appropriate null and alternative hypothesis for the Single Factor ANOVA F test.

$$H_0 : \mu_{major\,1} = \mu_{major\,2} = \cdots = \mu_I$$
$$H_a : At\,least\,two\,population\,means\,differ$$

d. **(3 points)** State the conditions for the Single Factor ANOVA F Test. Is it reasonable to seem that these conditions are satisfied? Explain your reasoning. *If not, still proceed.*

- Samples are Random and representative of population.
  - True. These samples are random and representative of the population of students at OSU.
- Based on CLT, $n'_I s$ s are sufficiently large for $\overline{X'_I s}$ to be approx. normal.
  - False. Only Computer Science, Civil or CEM and Mechanical Engineering are sufficiently large to be representative of the population.
- The I populations are independent.
  - True. The populations are independent.
- The I population variances are equal: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_I^2$
  - False
    - $\sigma_{Chemical\,Engineering} = 3.535534$
    - $\sigma_{Civil\,o\,r\,CEM} = 6.855991$
    - $\sigma_{Computer\,Science} = 11.074066$
    - $\sigma_{Industrial\,Engineering} = 2.061553$
    - $\sigma_{Mechanical\,Engineering} = 9.938138$
    - $\sigma_{Other} = 2.939671$

e. Perform the Single Factor ANOVA F test in R.

1. **(2 point)** Paste the ANOVA table.

	DF	Sum Sq	Mean Sq	F Value	PR (>F)
Majors	6	2100	350.0	3.746	0.00145
Residuals	312	19897	93.4		

2. **(2 points)** From the ANOVA table, what is the average between group variability and the average within group variability, respectively the MSTr and MSE?

$$MSTr = 350.0$$
$$MSE = 93.4$$

f. Use the F statistic and p-value from the ANOVA table to state whether there is a significant difference between at least two of the group means.

1. **(2 points)** State whether to reject the null. State the test statistic and p-value.

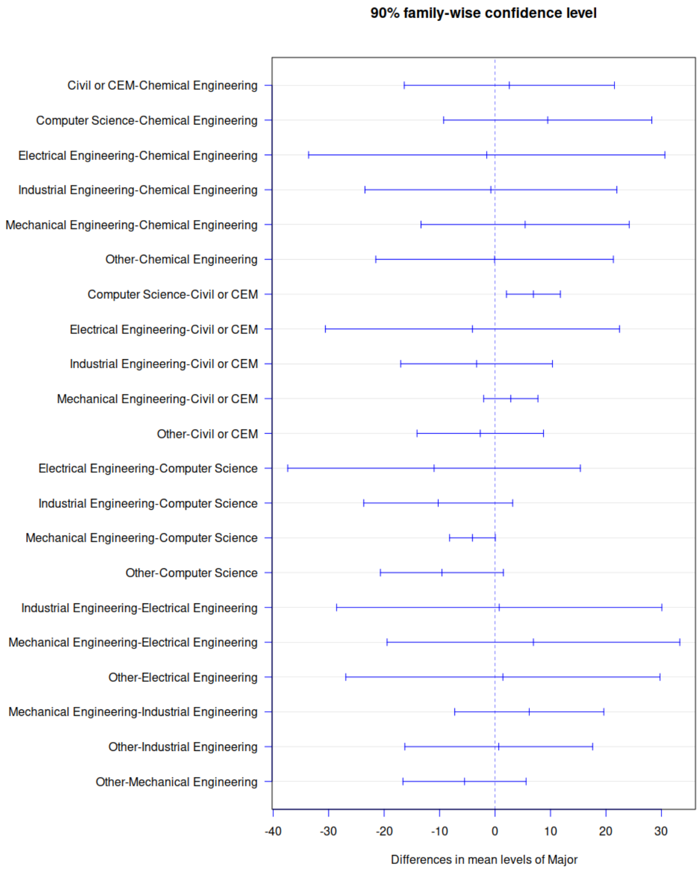
   We very strongly reject the null hypothesis, as we are using a significance level of 0.10 and our P value is 0.00145 (F Value = 3.746).

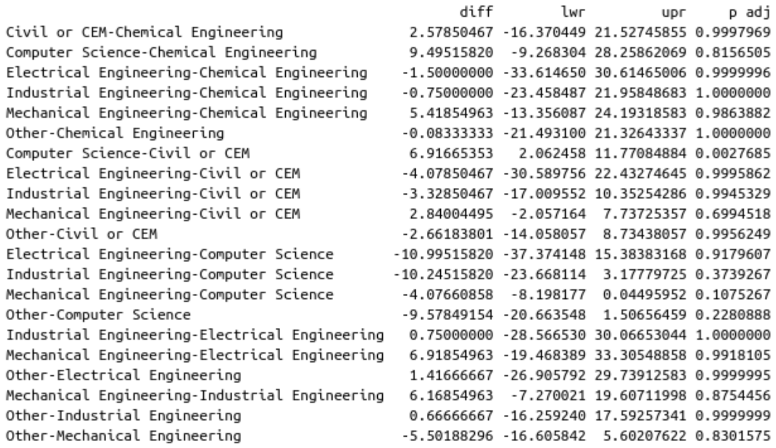
2. **(2 points)** Include a statement in terms of the strength of evidence in terms of the alternative. Include context.

   There is very significant evidence provided that different majors spend different amounts of time on average playing video games.

g. Using the Tukey's Multiple Comparison procedure output. Are there any individual comparisons that are significant at the 0.10 significance level?

1. **(2 point)** Paste R output for the multiple comparisons procedure.





```
                                                      diff        lwr         upr     p adj
Civil or CEM-Chemical Engineering               2.57850467 -16.370449 21.52745855 0.9997969
Computer Science-Chemical Engineering           9.49515820  -9.268304 28.25862069 0.8156505
Electrical Engineering-Chemical Engineering    -1.50000000 -33.614650 30.61465006 0.9999996
Industrial Engineering-Chemical Engineering    -0.75000000 -23.458487 21.95848683 1.0000000
Mechanical Engineering-Chemical Engineering     5.41854963 -13.356087 24.19318583 0.9863882
Other-Chemical Engineering                     -0.08333333 -21.493100 21.32643337 1.0000000
Computer Science-Civil or CEM                   6.91665353   2.062458 11.77084884 0.0027685
Electrical Engineering-Civil or CEM            -4.07850467 -30.589756 22.43274645 0.9995862
Industrial Engineering-Civil or CEM            -3.32850467 -17.009552 10.35254286 0.9945329
Mechanical Engineering-Civil or CEM             2.84004495  -2.057164  7.73725357 0.6994518
Other-Civil or CEM                             -2.66183801 -14.058057  8.73438057 0.9956249
Electrical Engineering-Computer Science       -10.99515820 -37.374148 15.38383168 0.9179607
Industrial Engineering-Computer Science       -10.24515820 -23.668114  3.17779725 0.3739267
Mechanical Engineering-Computer Science        -4.07660858  -8.198177  0.04495952 0.1075267
Other-Computer Science                         -9.57849154 -20.663548  1.50656459 0.2280888
Industrial Engineering-Electrical Engineering   0.75000000 -28.566530 30.06653044 1.0000000
Mechanical Engineering-Electrical Engineering   6.91854963 -19.468389 33.30548858 0.9918105
Other-Electrical Engineering                    1.41666667 -26.905792 29.73912583 0.9999995
Mechanical Engineering-Industrial Engineering   6.16854963  -7.270021 19.60711998 0.8754456
Other-Industrial Engineering                    0.66666667 -16.259240 17.59257341 0.9999999
Other-Mechanical Engineering                   -5.50188296 -16.605842  5.60207622 0.8301575
```

2. **(2 point)** List all comparisons that are significant (or state those that are not).

The following comparisons showed the greatest difference with relatively low P values, indicating that they are somewhat significant and different:

- Computer Science-Chemical Engineering
- Electrical Engineering-Computer Science
- Industrial Engineering-Computer Science
- Other-Computer Science

Only one comparison had a P value that was lower than our significance level of 0.1, and this was the comparison between Computer Science and Civil or CEM.

3. **(3 points)** Interpret the 90% F-W confidence interval for the difference with the smallest p-value (even if it is not significant).

The interval with the smallest P value was that between Computer Science and Civil or CEM. This has a difference of 6.92 hours difference in average gaming (CS averaging 6.92 hours more than CEM), with a 90% confidence that ranges from CS majors gaming 2.06 hours more than CEM majors to CS majors gaming 11.77 hours more than CEM majors.