MA5701 - Final Project

With your proposal approved and data collected, you will be performing a *minimum* of two distinct hypothesis testing procedures and summarizing your findings in a report or presentation.

- If your group chooses to give a presentation (via powerpoint slides or on a poster), the presentation session will be during the scheduled final exam period, at 12:45pm on Monday, December 16 in Fisher 139
  - If your group chooses to take this route (which should hopefully be more fun than a written report), please contact me by **5pm on Wednesday, December 11** so that scheduling may be fit
- If your group chooses to produce a final written report, this should be submitted to the Canvas assignment page by **5pm on Wednesday, December 18**

Taking either the report option or the presentation option, you will be submitting...

- your report, poster files, or presentation files, saved as either a .pdf or .docx file
- your set of data, saved as a .csv file
- a single R program that you used to perform your analyses, saved as a .R file

Requirements of the report/presentation and other submissions follow:

Format

Your project should have

- a descriptive title
- three sections titled "Introduction", "Methods", and "Results" (a "Conclusions" section is nice, but not formally required)
- within the "Results" section, two subsections which have a brief but descriptive title describing the specific analyses that are investigated there

All requested graphics should...

- be labeled using the format "Figure 1: Some Short but Descriptive Title"
  - this title can be done in R, or if you find it more convenient you can add the title label in your word processing document (or LaTeXif such typesetting is in your repertoire)
- have short but descriptive axes labels
- if appropriate, have a legend (getting a title on the legend can be messy in R, let me know if you need help with that)
- have axes which cover the entire range of the plotted points/bars/lines
  - recall that the default in R doesn't always cover the range of your data, so you may have to use the `xlim` or `ylim` arguments in your plotting functions
- **if** you feel more descriptive information is needed with your graphic, feel free to add a caption to it

All requested tables should...

- be labeled using the format "Table 1: Some Short but Descriptive Title"

R program

Your R program should...

- be able to be run without producing any errors
- perform all of the analyses found in your report

Introduction Section

The Introduction section should be used to draw the reader into your report, as well as give a basic description of the problems that are being investigated. This section should typically be about 2 or 3 paragraphs long (there is no formal length requirement) and for full credit should address include each of the following points:

- an *explicit* description of the **population** that is being targeted
- an *explicit* description of all of the **variables** that will be observed
- a description of the **purpose/motivation** of the experiment
  - why have you chosen this population and set of variables for investigation? (**Do not** just say "Because we have to do a project for the course")
  - what sorts of relationships do you intend to investigate with the data?

Methods Section

The Methods section should tell the reader what sort of work had to go into the planning and execution of the experiment. This should include the following:

- an *explicit* description of the **sampling unit**
- an *explicit* description of whether this was a controlled experiment or an observational study
- for all numerical variables in your data set (that you end up using in your analyses), make sure to state the units; for all categorical variables that you use, make sure to explicitly list out all of the categories this variable takes (perhaps in a table, if there are quite a few categories)

For the handful of you who choose to perform a controlled experiment, you should...

- specifically describe the independent variable(s) that you were able to control (what values of this variable did you choose, and why?)
- specifically describe the thought process you used in conjecturing about other (unobserved) variables which you think may affect the response variable, and the steps you took to try to keep these variables constant
- discuss any potential problems that you may have run into while trying to control your independent variable or other variables

For the majority of you who choose to analyze data from an observational study, you should...

- provide hyperlinks to all of the online sources that you used in collecting your data
- discuss any extra work you may have had to do go "clean" your data (creating extra variables, removing unnecessary variables, etc.)
- if you created a survey for collecting data, provide a link to the survey that was used (or place all of the questions, with the *exact wording that was used*, in an appendix) as well as a discussion of the dates that the survey was administered and how you tried to recruit responders to take the survey

Results Section

Your Results section will take up the majority of the length of the report/presentation, mostly due to graphics and tables. It is here where you will present the results of your statistical analysis of your data.

Before jumping into the analyses, remind the reader of the essential problems you wish to address with your two analyses. This should only take a few sentences (you don't need to go into specifics with the statistical analyses that will be used at this point).

Also state the total number of observations that you have in your data set here.

Pick two analyses from the following list (*at least one* of these must be a **one-way ANOVA** or **simple linear regression**); repeats are acceptable (you could perform two one-way ANOVA analyses, or two simple linear regressions):

- a one-way ANOVA (or Welch's ANOVA, or Kruskal-Wallis test)

  - some of you may have experimental designs which involve a **repeated measures ANOVA**; ask me for further guidance if you wish to perform such an analysis
  - some of you may have experimental designs which would involve a **two-way ANOVA**; ask me for further guidance if you wish to perform such an analysis
- a simple linear regression
  - if you wish to perform a multiple linear regression, ask your instructor for further guidance
- a comparison of two group means (via a pooled *t*-test, Welch's *t*-test, or the Wilcoxon rank sum test)
- a comparison of two group means with a paired sampling scheme (via a paired *t*-test, the Wilcoxon signed rank test, or the sign test)
- a comparison of two population proportions (via the z-test on the difference in two population proportions or Fisher's exact test)
- a comparison of two categorical variables arranged in a contingency table (via Pearson's χ^2 test or Fisher's exact test)
  - ask your instructor for further guidance on this test if you wish to take this route

You will then have two *subsections* (with descriptive but short titles) which contain analyses of your data from the above choices (please feel free to include more if you have a fun and rich set of data).

- if you were performing a controlled experiment, you only need one subsection and analysis, which is either a one-way ANOVA or a simple linear regression

Deviations from the above requirements require prior instructor approval.

Requirements on what are needed for each potential analysis you might perform follow:

One-way ANOVA

To perform the one-way ANOVA (or related procedures, such as Welch's ANOVA or the Kruskal-Wallis test), you need **at least three** categories for your independent variable. Your analysis and subsection should include all of the following components:

- a set of parallel boxplots of your response variable split by the levels of the independent variable
- a table which includes the sample size, sample mean, sample median, and sample standard deviation of your response variable for each of the *k* samples defined by your independent variable
- an ANOVA table (including the columns "Source of Variability", "df", "SS", "MS", "F", and "*p*-value")
  - you don't need to include the "Total" row (which isn't included in the R output anyhow)
- a discussion on the validity of the assumptions of the one-way ANOVA which refers to…
  - a normal q-q plot of the residuals
  - the *p*-value from the Shapiro-Wilk test applied to the residuals
  - a plot of the residuals against the fitted values
  - the *p*-value of Levene's test applied to the squared residuals
- if the assumptions are violated, report also the *p*-value from either Welch's ANOVA or the Kruskal-Wallis test
  - provide a *brief* discussion as to why you make the choice of either Welch's ANOVA or the Kruskal-Wallis test (or the choice to stick with the one-way ANOVA)
- a discussion of the **accurate** conclusions you may make based off of the one-way ANOVA, Welch's ANOVA, or Kruskal-Wallis test
  - state the choice of significance level α that you use in coming to this decision (which need not be the "typical" choice of $\alpha = 0.05$)
  - you may feel free to adopt Fisher's philosophy to significance testing, and not use a "hard cutoff" of α in coming to a decision of the test, but may talk about "strong", or "moderate" or "marginal" significance as you see fit

If you come to the decision that there are significant differences between the group means/medians, you should follow up with *post hoc* multiple comparisons

- report the output of the `TukeyHSD` function in a table if you used the one-way ANOVA

- report the output of the `pairwise.t.test` function with the `pool.sd=FALSE` argument if you use Welch's ANOVA
- report the output of Dunn's test (you must install the `dunn.test` package in R if you haven't done so yet) if you used the Kruskal-Wallis test

The output for any of these multiple comparison procedures should be reported in a table with the pair of groups that are being compared in one column, and the corresponding *p*-value in a second column; if you have much too many comparsions to make, you can include in the table the rows corresponding to comparisons with *p*-values which are <0.15. Provide a brief discussion as to whih differences you find to be significant by referring to this table and the figure containing the parallel boxplots.

Simple Linear Regression

To perform a simple linear regression, your independent variable and dependent variable should both be *roughly* continuous (ask your instructor if you're unsure about this requirement). Your analysis should include the following:

- a scatter plot of the response variable (on the *y*-axis) and independent variable (on the *x*-axis)
  - this plot should have the regression line overlaid on top of it
  - this plot should also have a set of 95% confidence bands with a dashed plotting style
  - this plot should also have a set of 95% prediction bands with a solid and thick plotting style (refer to the "Regression in R" handout for example code on how to plot these)
- a table of summary statistics of the response and independent variables, including sample means (for x and y), sample standard deviations (for x and y), and the sample correlation
- a table similar to that output from the `summary` function in R, which has rows corresponding to slope and intercept and columns corresponding to...
  - the point estimate of the parameter
  - the standard error
  - the *t*-value
  - the corresponding *p*-value
- include in the caption of this table the coefficient of determination (R^2), the residual standard error $\hat{\sigma}$, and the residual (error) degrees of freedom
- specifically discuss the validity of each of the assumptions of normality, homoscedasticity, and an appropriate linear model by referring to...
  - a normal q-q plot of the residuals
  - the *p*-value of the Shapiro-Wilk test applied to the residuals
  - a plot of the residuals against the fitted values
- if you feel that a transformation of the response or independent variable is warranted, fit this model instead (we really only investigated the log transform, but ask your instructor about others if need be)
  - if you perform a transformation of the data, make sure to explicitly state you have done so in your discussion; your tables and plots requested above should then be associated with the transformed data
- if you feel that a polynomial model is warranted, include a couple of extra variables (don't go beyond a cubic polynomial) in the model
  - your table of point estimates, standard errors, *t*-values and *p*-values will then have another row or two corresponding to a quadratic or cubic term
- discuss the significance of your model by referring to the *p*-values; what **inferences** can you make from your analysis?

Two Independent Samples

If you have already performed a one-way ANOVA or simple linear regression, you may perform a two-sample *t*-test (if you have two independent samples). Your analysis should include...

- a set of parallel boxplots of your response variable split by the levels of the independent variable

- a table which includes the sample size, sample mean, sample median, and sample standard deviation of your response variable for the two samples defined by your independent variable
- a discussion on the validity of the...
  - normality assumption, which refers to two normal q-q plots of the two samples (if you like, you could produce just a single normal q-q plot of the *residuals* by treating this like a one-way ANOVA)
  - homoscedasticity assumption, which refers to your parallel boxplots
- if you are comfortable with the normality assumption, report in a table the following results associated with both the pooled *t*-test and Welch's *t*-test (on two separate rows):
  - the difference in sample means (this will be the same for both rows)
  - the degrees of freedom
  - the value of the *t*-statistic
  - the *p*-value
  - a 95% confidence interval on the difference in means
- if you are not comfortable with the normality assumption, create a table with results from the Wilcoxon rank sum test (use the `conf.int=TRUE` argument in the `wilcox.test` function)
  - the difference in location (as reported by R)
  - the value of the test statistic W (as reported by R) (don't worry about trying to interpret this quantity)
  - the *p*-value
  - the 95% confidence interval on the difference in location
- state the conclusions that you will draw from the test (pooled *t*-test, Welch's *t*-test, or the rank sum test)
  - was this a one-tailed or two-tailed test? make sure you explicitly state this in your discusssion, and that your *p*-value reflects this fact

Two Dependent Samples

If you have already performed a one-way ANOVA or simple linear regression, you may perform a paired *t*-test (**if** you have two dependent samples). Your analysis should include...

- a scatter plot of the two variables
  - use the command `abline(a=0, b=1)` after your plotting function to add a line to the plot for making easy comparisons
- a table of summary statistics including sample size, sample mean, sample median, and sample standard deviation with rows corresponding to the first sample, the second sample, and the sample of differences
- a normal q-q plot of the sample of *differences*, along with the *p*-value of the Shapiro-Wilk test applied to these differences
- a discussion as to whether or not the assumption of normally distributed differences is appropriate
- if a *t*-test is appropriate, then include a table which contains the mean of the differences, the standard deviation of the differences, the degrees of freedom, value of the *t*-statistic, the *p*-value, and a 95% confidence interval on the population mean of the differences
- if a *t*-test is not appropriate, include a table with rows corresponding to the sign test and Wilcoxon signed rank test, and columns corresponding to the value of the test statistic (S as reported by `SignTest` in R, or `V` as reported by the signed rank test), the *p*-value, and the 95% confidence interval that is reported by either function (use the `conf.int=TRUE` argument in the `wilcox.test`) function
- state the conclusions that you will draw from the test(s) you performed
  - was this a one-tailed or two-tailed test? make you explicitly state this in your discussion, and that your *p*-value reflects this fact

Difference in Two Proportions

If you have already performed a one-way ANOVA or simple linear regression, you may perform a test on the difference between two population proportions. Your analysis should include...

- a contingency table of the counts of "successes" and "failures" in each of your two populations (use more appropriate labels than "success" or "failure" or "population 1" and "population 2")
  - include marginal row and column totals in your contingency table as well
- a table containing the difference in sample proportions, its standard error (sorry, you'll have to go back to the notes and implement the formula "by hand"), the value of the χ^2-statistic (called `X-squared` in `prop.test`), the p-value, and a 95% confidence interval on the difference between the proportions
- if you have "small" sample sizes (according to the rule of thumb used in class), report the p-value of Fisher's exact test
- state the conclusion that you will draw from the test you performed
  - was this a one-tailed or two-tailed test? make you explicitly state this in your discussion, and that your p-value reflects this fact