Homework #2: Summary Statistics

This written homework assignment should be *typed* and written in the format of a *short report*.

- You **do not** need to fuss with the visual aspects of your report (such as creating two columns, or wrapping text around figures, or making tables which are "pretty", or trying to typeset mathematical notation, . . .)
- For full credit, you **do** need to. . .
  - write in complete sentences
  - use paragraph breaks judiciously
  - use the section headings that are provided ("Hand Spans", "Age of Students", . . .)
  - address all of the requested points in the problems that follow

Also, make sure that *every* requested graphic has a short but descriptive title, labeled axes (or axis) with units (when appropriate), and a legend (if appropriate).

If any of these expectations are unclear, make sure to *ask questions* in a timely fashion (i.e. not four hours before the deadline).

You will be performing several hypothesis tests on the data that you investigated previously in the first report. Recall that you should use the command

```
> survey <- read.csv('stats.survey.csv', na.strings='')
```

in order to read the data file properly.

DO NOT use any function like `na.omit` on your data, which removes more observations than are necessary. The various test functions in R will automatically remove missing data when performing the hypothesis test, and if you need calculations of a sample mean or sample standard deviation you can use the `na.rm=TRUE` argument in the appropriate functions.

Hand Spans

Two of the variables in the data set are

- `WrHnd`, the span of the student's writing hand (measured in cm); the span of someone's hand is the distance from the tip of their thumb to the tip of their pinky finger when their hand is fully stretched
- `NWHnd`, the span of the student's non-writing hand

Someone has the *a priori* hypothesis that the span of someone's writing hand will be *greater* than the span of their non-writing hand. Test this hypothesis with an appropriate t-test as well as all appropriate nonparametric tests that were investigated in class: the sign test, the Wilcoxon signed rank test, or the Wilcoxon rank sum test (which one(s)?)

This section of your report should include the following components:

1. A brief discussion as to whether this design consists of two *independent* samples or two *paired* samples. **Explain the reasoning** that you use in making this determination.
2. A list of the different hypothesis tests that can be applied to this problem, given your determination from the previous discussion.
3. For the appropriate t-test (Welch t-test or paired t-test?). . .

- give *verbal descriptions* of both null and alternative hypotheses
  - if you are using a one-sided alternative (should you be?) make sure to specify the direction of the alternative hypothesis in relation to the two populations
- in a table, include the sample means of both the writing hand span and the non-writing hand span, the sample mean of the differences between these two spans, the standard deviation of the sample of differences, the degrees of freedom for the *t*-test, the value of the test statistic (t_{obs}), and the *p*-value
- report a 95% **two-sided** confidence interval on the mean difference in hand spans
  - this can be either in a table or reported in the text
  - give a **verbal interpretation** to this interval which includes units; specifically, what can you infer from this interval with 95% confidence?
- create a normal q-q plot of the differences in writing hand spans
  - you can use the default axes labels from R, but make sure to change the title to be more descriptive
  - comment on whether or not the normal q-q plot suggests the pouplation is normal (there is one "right" answer here!)
  - comment on whether or not the the *t*-test could be used with this data
- decide whether you should reject or retain the null hypothesis using a significance level of $\alpha = 0.05$, and **interpret** your decision (do not just simply state "reject H_0" or "retain H_0")

4. For the appropriate nonparametric test(s)...

   - write down the single R function and arguments that you use
   - state the *p*-value of the test
   - state the decision of the test, based off of a significance level of $\alpha = 0.05$

5. Provide a brief discussion on the conclusions that can be drawn from your analyses. Specifically, all of the tests yield the same decision when $\alpha = 0.05$? If not, which of the tests do you have the most confidence in? **Explain your reasoning** by referring to the assumptions.

Age of Students Consider two other variables in this data set:

- `Age`, a student's age (in years)

- `MI` (or `M.I` if you have loaded the data frame from the `MASS` package in R directly), which is a categorical variable indicating how the student reported their height
  - the levels are `Metric` if they reported their height in cm (these are students in Australia) and `Imperial` if they reported their height in feet and inches

We are interested in knowing if there is any difference in the mean age between students who report their height in imperial vs. metric units; there is no *a priori* suspicion as to which population might have a greater mean age.

This section of your report should include the following components:

1. A brief discussion as to whether this design consists of two *independent* samples or two *paired* samples. **Explain the reasoning** that you use in making this determination.

2. A list of the different hypothesis tests (*t*-test and nonparametric tests) that can be applied to this problem, given your determination from the previous discussion.

3. For the appropriate *t*-test (Welch *t*-test or paired *t*-test?)...

- give *verbal descriptions* of both null and alternative hypotheses
  - if you are using a one-sided alternative (should you be?) make sure to specify the direction of the alternative hypothesis in relation to the two populations
- in a table, include the sample means of ages of students from either population (those who answer in metric units and those who answer in imperial units), the sample standard deviations of the ages from these two populations, the degrees of freedom for the *t*-test, the value of the test statistic (t_{obs}), and the *p*-value
- report a 95% **two-sided** confidence interval on the difference in mean ages between the two populations
  - this can be either in a table or reported in the text
  - give a **verbal interpretation** to this interval which includes units; specifically, what can you infer from this interval with 95% confidence?
- create normal q-q plots of the ages from both populations
  - you can use the default axes labels from R, but make sure to change the title to be more descriptive
  - comment on whether or not the normal q-q plots suggests the populations are normal (there is one "right" answer here!)
- state the decision of the test when using a significance level of $\alpha = 0.05$

4. For the appropriate nonparametric test(s)...
   - write down the single R function and arguments that you use
   - state the *p*-value of the test
   - state the decision of the test, based off of a significance level of $\alpha = 0.05$
5. Provide a brief discussion on the conclusions that can be drawn from your analyses. Specifically, **interpret the conclusions** of the decision of either test. Which test yields a significant result: the *t*-test or the appropriate nonparametric test?

Exercise Frequency Consider two other variables in this data set:

- `Sex`, taking the values `Female` and `Male`
- `Exer`, which is a categorical variable indicating how often the student reported that they exercise
  - the levels of this variable are `None`, `Some`, and `Freq` (frequently)

Suppose that we are interested in determining if there is any difference in the proportions of students who report that they *frequently* exercise between the populations of female and male students. We have no *a priori* suspicion that one group over the other will have a greater proportion.

This section of your report should include the following components:

1. A verbal statement of the null and alternative hypotheses
2. A table which gives the total number of females (n_1), total number of females who reported exercising frequently (x_1), the total number of males (n_2), and the total number of males who reported exercising frequently (x_1)
3. Another table which reports the two sample proportions $\hat{p}_1$ and $\hat{p}_2$, the value of the χ^2-statistic, and associated *p*-value from `prop.test`; you can use the default continuity correction when using `prop.test`
4. The decision of the test when using a significance level of $\alpha = 0.05$, and a **verbal interpretation of the conclusions** of this test