Homework #1: Summary Statistics

The written homework assignments 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 as prompted in the following problems
  - 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).

Statistics Survey

The data file `stats.survey.csv` contains responses from 237 students enrolled in Statistics I at the University of Adelaide. (This data actually comes with another R package, called `MASS`.)

In order to import this data file *correctly*, you will have to add one argument to `read.csv`:

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

Make sure after importing this data file that, after running `str` on the data frame, you see that the variable `Sex` has 2 levels (not 3) and `Smoke` has 4 levels (not 5).

Perform the requested analyses, making sure to split your report into sections with titles as are used below. You will not be submitting your R program with this report, but it is **strongly recommended** that you save your program in a convenient location. You should be able to run this program from a fresh R session and get the same exact results each time.

Student Ages

In this section of your report, you will provide a summary of the single variable `Age`. Include. . .

- a histogram of the variable `Age` (including a title, labels on both axes, units of years for the age, and the use of the `ylim` argument to ensure the range of the axis covers all bar heights)
- a table which reports the five-number summary (min, three quartiles, and max), along with the sample mean, sample IQR, and sample standard deviation (of the variable `Age`)
- comments on the graph and summary statistics
  - comment on the center, spread (via discussing both the range and the middle half), and skewness (in what direction is the skew, if there is any? how strong is it?)
  - explain which of the mean or median would be a better measure of center in this sample, and why you think so
  - explain which of the IQR or standard deviation would be a better measure of spread in this sample, and why you think so

  - discuss the presence of any outliers in the sample and potential reasons for their being present

Student Heights

In this section of your report, you will summarize the heights of students in a variety of ways. Note there is a variable `Height` in the data frame, which is measured in centimeters. As we do not reside in Australia, create a variable `Height_Inch` in your R program which is defined as `Height/2.54`.

Under the *Student Heights* section heading, write a short report which includes the following:

- a set of parallel boxplots of `Height_Inch` split by the levels of `Sex` (including title and labels on *both* axes, with units as appropriate)
- a table containing the mean, median, and standard deviation of `Height_Inch` split by the levels of `Sex`
- a set of brief comments on these summaries
  - state any *a priori* hypotheses you have about the difference in mean height between the two groups; do you expect to see any significant differences in spread between the two groups?
  - for each group (of female and male students), provide brief descriptions of center, spread, (perceived) skew, and the presence of any outliers
  - provide subjective comparisons of both center and spread between the two groups
  - you may use some of the comments made in the various lecture worksheets as a template for how to structure your comments (but feel free to add your own personal style)

Exercise and Smoking Trends

The `Exer` variable is split into three levels, `None`, `Some`, and `Frequent`, and the `Smoke` variable is split into four levels. Before performing an analysis of these variables, first run the commands

```
> Exer <- factor(Exer, levels=c('None', 'Some', 'Frequent'))
> Smoke <- factor(Smoke, levels=c('Never', 'Occasional', 'Regular', 'Heavy'))
```

(this will only work if you've used `attach(survey)` at the beginning of your program).

In this section of your report (titled "*Exercise and Smoking Trends*"), include...

- a brief description of any *a priori* hypotheses you might have about the relationship between exercise frequency and smoking frequency (if you have no expectations, it is fine to say so)
- a side-by-side bar chart of the variables `Exer` and `Smoke`, making sure that this plot has...
  - a short but descriptive title
  - labels on both axes
  - a y-axis which adequately covers the entire range of the bar heights
  - a legend with a *title* (use the code in the Categorical Variables handout as a template)
- a table of sample *percentages* of the levels of `Smoke`
- a table of sample *percentages* of the levels of `Exer`
- a discussion of whether or not there appears to be a relationship between these two variables (frequency of smoking and frequency of exercise)
  - explain what you should expect to *roughly* see in the bar chart if the two variables were independent of one another

Hand Spans

The two variables `WrHnd` and `NWHnd` are the hand spans, measured in centimeters, of a student's writing hand and non-writing hand, respectively. Create a new variable called `DiffHnd` which is the difference `WrHnd - NWHnd`.

Include in this portion of your report...

- a scatterplot of `Wr.Hnd` (on the *y*-axis) and `NW.Hnd` (on the *x*-axis), including a descriptive title and labels (with units) on both axes
- a brief discussion on this plot which includes
  - a statement of the sample correlation coefficient between the two variables
  - a subjective description of the nature of the relationship between the two variables (linear with a positive slope, linear with a negative slope, nonlinear, no relationship?)
- also create a histogram of the single variable `DiffHnd`, making sure it has a title and axis labels
- comment on the histogram of differences...
  - discuss the center by stating the mean and median of the differences
  - discuss the spread by stating the standard deviation, min, and max of the differences
  - discuss any perceived skew or outliers