STAC32
Assignment 1

Due Thursday Sep 25

What I am looking for here is the code you used to answer the questions, plus the output the code produces (you can edit this if it is long), plus (of course) your answers to the questions. You might find it convenient to have a Word (or WordPad or similar) document and copy-paste your code and output into there, then type whatever else you need.

For R, copying the stuff in the Console (bottom left) window is most of what you need. In your document, you should use a fixed-width font like Courier to make the tables of results look as they should. For graphs click on the Export button. What may work for you is a direct copy-paste via the clipboard; if that doesn't work, save the plot as an Image (on your Desktop, say) and then Insert Image From File in your document.

For SAS, you can copy and paste everything into your document. You will probably need to Paste Special the output in tables (so that the HTML stays as HTML). When I tried it, the graphs pasted successfully as images. Everything *should* be good. Things sometimes go awry, though, so check (especially) that all your tables of output came across as they should have done. (Rows of tables sometimes get mixed up.) For the code, copy-paste it from the Code tab, and in your document, change the font to something like Courier so that it looks appropriate. If copying and pasting does not work, you'll need to take a screenshot (Alt-PrintScreen on my computer) and paste that into your document as an image.

The links to the data files (the URLs) should be clickable. Let me know if you have any problems with this. As I see them, they have pale blue boxes around them.

1. Here are the annual numbers of deaths from tornados in the United States from 1990 through 2000:

```
53 39 39 33 69 30 25 67 130 94 40
```

   (a) Start up R Studio, and enter these data into a variable using `c()`. Verify (by typing the name of your variable) that the values were read in correctly.

(b) Find the mean and standard deviation of the number of deaths from tornados, quoting what you think is an appropriate number of decimal places.

(c) Find the median and quartiles of the number of deaths from tornados. (It is easier to obtain the five-number summary and get them all in one go.)

(d) Obtain a histogram of the number of deaths from tornados. Comment briefly on its *shape* (is it skewed, or roughly symmetric?)

(e) Obtain a boxplot of the number of deaths from tornados. Does this tell the same story as the histogram? Explain briefly.

(f) Which do you think would be the better numerical summary of the data: the mean and standard deviation, or the median and quartiles? Explain briefly. (No computation needed for this part.)

2. A student wanted to compare the efficiency of various coffee travel mugs. She decided to try four different mugs. Each time, she heated water to 80 degrees C, poured it into a mug, and sealed the lid. After 30 minutes, she measured the temperature again, and recorded the temperature difference. (In the data set, the temperature differences are never negative, because in each case the water cooled down if anything.) The data are at `http://www.utsc.utoronto.ca/%7ebutler/c32/coffeecups.txt`. Open this file in your web browser. (The URL above should be clickable.)

   (a) Start up SAS. Make a new SAS file if you want to. Click on the Code tab. Write a `data` step that will read in the two variables, mug type (which is text) and temperature difference (a number). In your `data` step, include a `cards` (or `datalines`) line, and copy-paste the data in below that. Then add a a line with a `;` by itself, and then a `proc print;` line, to list the data, and below that, a `run;` line. *Check to see that you have semicolons in all the right places*, that is to say, at the end of every line of code, but *not* the individual lines of data. You can mimic what you saw in Lab 1.

   (b) Use `proc means` to obtain the means and SDs of temperature difference for each type of mug. Which mug appears to be best? Which appears worst? (You will have to define what you mean by “best”).

   (c) Obtain side-by-side boxplots of temperature difference for each type of mug. Does this support your conclusions from the means? Is there any evidence of skewness? Explain briefly.

3. The 50 states of the US were each classified (geographically) as “east or midwest” or “southern or western”. For each state, the population growth (in percent) between 1990 and 2000 was recorded. We are interested in whether the population growth is different in a southern-or-western state than it is in an eastern-or-midwestern state.

(a) Use SAS for this problem. The data file is at `http://www.utsc.utoronto.ca/%7ebutler/c32/pop.txt`. Open it in your web browser. Create a new SAS file, click on the Code tab, and copy-paste the data into it. In the Program Editor, select File and Save As, and save the data in a file of your choice. `pop.txt` will do if you can't think of a better name. SAS will add a `.sas` on the end of the file name; find the file on the left, right-click it and Rename to get rid of the `.sas`. Create another new SAS file. Then enter some code that will read the data in from the file and print it out (using `infile` in your `data` step).

(b) To assess the evidence that the regions differ, obtain side-by-side boxplots and a table of means for the two regions. You can do this by adding some code onto what you wrote to read in the data (above). Explain briefly how, if at all, the two groups differ. (Pay attention to differences in centre, spread and possibly shape.)

4. The prices (per case) for 36 different wines from three wine-growing regions of New York state were recorded. The data are available at `http://www.utsc.utoronto.ca/%7ebutler/c32/wines.txt`.

   (a) Open the data file in a web browser. Select and copy all the values, including the header line at the top. Go to R Studio and open a suitable new window top left (File, New and Text File). Paste the values there, and save them as a file `wines.txt` in your current project (File, Save As). Write some code to read in the data from the file (use `read.table`) into a data frame, and verify that the values are correct.

   (b) Draw side-by-side boxplots to compare the wine prices.

   (c) Use `aggregate` to compare the mean values of `CasePrice` for the three `locations`.

   (d) Do you think the mean provides a fair comparison between the wines? What do the boxplots add to the comparison? Explain briefly.

5. The file `http://www.utsc.utoronto.ca/%7ebutler/c32/CrimeRates.xls` contains rates (incidences per 100,000 population) of various types of crime in a selection of US and Canadian cities. Save the file on your computer and open it in your spreadsheet program. (This might be Excel, or it might be something like OpenOffice or LibreOffice, which is what I use. Even though the data are in Excel format, you should be able to open the file in any of those. Let me know if not. (If you don't have a spreadsheet program, download and install a free one such as OpenOffice or Gnumeric). Leave the spreadsheet open, since you'll need it for the next question as well. (You might be able to open the spreadsheet directly, without saving it. That's fine too.)

(a) Select all the cells in the spreadsheet except for the top row (you don't need the row headers). Copy these. Paste the spreadsheet data into an empty Code window in SAS. Save it (File and Save As) into a file, choosing a name like `crime.txt`. (As you recall, you'll have to save it as something like `crime.txt.sas` first, and then rename it.) Next, write code to read in the crime rate data, bearing in mind that those data are separated by tabs and not spaces. Give the columns suitable names. Run `proc print` to verify that everything is OK.

(b) There are two unexpected things about the way this data set has turned out. One has to do with one of the columns, and the other has to do with one of the rows. What are these things?

(c) Make side-by-side boxplots for aggravated assault, for the US cities and the Canadian cities. What differences do you see? Are any of the Canadian cities more like an American one in terms of aggravated assault?

(d) Use `proc means` to compare the US and Canadian cities' homicide rates. What do you see? Do any of the American cities look like a Canadian one when it comes to homicide rates?

6. Go back to the crime rates spreadsheet. Find the missing value in the Boston line, marked with a `*`. Change this to `NA`, which is R's code for "missing value". Save the crime rates spreadsheet as a `.csv` file somewhere in your project folder. We are going to do a similar analysis to the above in R:

   (a) Read the `.csv` file into a data frame using `read.csv`. Verify that the data frame contains the right values.

   (b) Make side-by-side boxplots of aggravated assault rates for the two countries.

   (c) Use `aggregate` to calculate mean homicide rates for the two countries. Note that "homicide" has been misspelled, and you will have to use the misspelled name!

   (d) Use `aggregate` to obtain five-number summaries of homicide rate for each country. (Just do this the way you'd expect it to work. Note how R figures out how to handle the output.)

   (e) Check, from the *numerical* output you have, that R and SAS produce the same answers.

7. The spreadsheet at `http://www.utsc.utoronto.ca/%7ebutler/c32/Cholesterol.xls` contains measurements on blood cholesterol for two groups of people: smokers and ex-smokers.

   (a) Get the data into an R data frame. Describe briefly how the data frame is laid out. (The `NA`s are there to make the two columns the same length, as is required for a data frame.)

(b) Create a new data frame by running a command like this (assuming that your original data frame is called `chol`):

```
R> chol2=stack(chol)
```

Take a look at the new data frame. How does it differ from the original one?

(c) Using either the original or the new data frame (only one will work), produce histograms of the cholesterol levels (i) for the smokers and (ii) for the non-smokers.

(d) Using either the original or the new data frame (only one will work), produce side-by-side boxplots of the cholesterol levels for the two groups (smokers and ex-smokers).

(e) Looking at your boxplots, what notable differences, if any, do you see between the cholesterol levels of the smokers and the ex-smokers?

(f) Use your histograms to compare the smokers and the ex-smokers. Do you see any notable differences here? Are your conclusions consistent with the ones from the boxplots?

(g) Which do you think is easier to use for comparison: the boxplots or the histograms? Explain briefly.