Assignment 2

Due Thursday October 9, 16:00 or earlier

1. Major league baseball teams in North America play in one of two leagues: the American League, in which the Blue Jays play, and the National League, in which the Montreal Expos used to play. In baseball, pitching and batting are very different skills, so that pitchers tend not to be very good at batting. In the National League, the pitcher has to bat, but in the American League, the pitcher does not bat (in his place, a special player called the Designated Hitter bats).

   The aim of the American League's Designated Hitter rule is to make the game more exciting, that is, there should be more runs on average in games involving American League teams than in games involving National League teams. The file `http://www.utsc.utoronto.ca/%7ebutler/c32/baseball.txt` contains the number of runs per game for each team in major league baseball. The top line is the names of the variables.

   (a) Using R, read in the data file and check that you have three columns: the name of the team, the number of runs per game, and the league in which that team plays.

   (b) Make side-by-side boxplots of runs for each league. How do the averages (medians) compare?

   (c) Do a (two-sample) t-test to compare the mean runs per game for the two leagues. Does the American league have a significantly higher mean number of runs per game?

2. Re-do the previous question using SAS. That is:

   (a) read in the data file and check that the values look reasonable.

   (b) make side-by-side boxplots of runs for the two different leagues

   (c) do a two-sample t-test to assess the evidence that the mean runs in the American league is higher than the mean runs in the National league. Which two-sample t-test did R give you? How can you tell?

3. Some students bought six randomly chosen bags of Dorito snacks (that is, from different stores at different times). They measured the net weight of

the Doritos in each bag (that is, just the snacks and not the bag containing them). The results are in `http://www.utsc.utoronto.ca/%7ebutler/c32/doritos.txt`. (Again, the link should be clickable as is.)

(a) Using R, read in the data and check that you have the right values.

(b) The bags of Doritos are supposed to weigh 29.2 grams. Test whether the data are consistent with this. (That means, you are looking for evidence that the population mean is *different from* 29.2; if you fail to reject the correct null hypothesis, the data are consistent with a mean of 29.2.) Doing the test will get you a 95% confidence interval for the mean as well. Do you reject the null hypothesis that the mean weight is 29.2 grams, at the $\alpha = 0.05$ level? Is the value 29.2 inside your confidence interval? Are these conclusions consistent with each other? Explain briefly.

(c) An 80% confidence interval is shorter than a 95% one. From the output that you have so far, can you tell whether 29.2 should be inside or outside an 80% confidence interval for the mean? Explain briefly. Then obtain the 80% confidence interval and show that you were right.

4. OK, so you guessed this one was coming: repeat the previous question using SAS.... This time, just check that your results agree with R's.

(a) read in the data and check that everything looks all right.

(b) do a t-test and obtain a 95% confidence interval

(c) obtain an 80% confidence interval. SAS likes you to specify `alpha`, which is one minus the confidence level.

5. An experimenter is planning a study. It is to be a one-sample study, to test the null hypothesis that a mean is 125, against an alternative that it is different from 125.

(a) The experimenter plans to use a sample of size $n = 20$. The population standard deviation is believed to be about 10. What power does the experimenter have for (correctly) rejecting the null hypothesis if the mean is actually 130? Use R for this. You'll need to specify values for `n`, `delta`, `sd`, `type` and `alternative`, where `delta` is the difference between the true mean and the null mean. If you can't figure out what to put for these, look at the help via `?power.t.test`.

(b) What sample size does the experimenter need to obtain power 0.75 under the same conditions as above (without `n=20`, obviously)? Use R again.

(c) Compare the sample sizes and powers in (a) and (b). Do the results make sense?

(d) Repeat parts (a) and (b) using SAS. Do you get consistent answers? How do the R and SAS answers to the sample size calculation differ? (You can mimic the example in the lecture notes, but you'll need to find something to replace `twosamplemeans` with, and you'll need to figure out what goes after `test=`. The help page at `http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_power_sect007.htm` might offer some hints (that link ought to be clickable, even though it seems to split over multiple lines). The lab session on this might also give you a clue.

6. R has various add-ons that people have written that extend what R does. These add-ons are called *packages*. To use a package, there are two steps: you first have to install it on your machine, and then you have to make it available for use. To install it, you type `install.packages("packagename")` (with the quotes), and when that has finished, signalled by the word `DONE`, you can use ("load") the package by typing `library(packagename)` without the quotes. You only have to install a package once, but you have to do the `library` thing in every R Studio session. That is, while R Studio is still open, all your packages are good, but as soon as you close it, the next time you open it you have to do the `library` thing again for any packages you want to use.

   (a) The package `BSDA` includes an implementation of the sign test. Install and load this package. Check that everything worked by typing `head(Tablrock)`. This should show you (the top 6 observations of) a data frame that has over 600 observations on 16 atmospheric variables at Mount Mitchell, North Carolina. One of these variables is `dew`, which we will use later.

   (b) `attach` the `Tablrock` data frame, and make a histogram of the `dew` values. What do you think is the most important feature of the histogram? (Why does the horizontal axis stretch so far to the left? Look carefully!)

   (c) We are interested in testing whether the "typical" `dew` value is 15 (against an alternative that it is something other than 15). Do you think a t-test or a sign test will be more appropriate? Explain briefly. (Do both the next two parts regardless of your conclusion here.)

   (d) Do the sign test for testing that the median is 16, against the alternative that it is not 16. In package `BSDA`, the appropriate function is `SIGN.test`, thus spelled, so you need to get the capital letters right. In `SIGN.test`, you need to give the vector of data values, and you specify the null median by `md=`. What do you conclude?

   (e) Do a t-test for testing that the mean is 16, against the alternative that it is not 16. What do you conclude? (This is the one-sample version of `t.test`: feed it a vector of data and the null mean via `mu=`.)

(f) Compare the 95% confidence intervals from the sign test and the t-test. Are they similar or noticeably different? Explain briefly.

7. The data from the last question is also in the file `http://www.utsc.utoronto.ca/%7ebutler/c32/tablrock.txt` The top line of the data file contains the names of all the variables.

   (a) Read the data into SAS[1] and make a histogram of the `dew` values. Do you see the same shape as you did with R?

   (b) Use `proc univariate` to obtain (all at once) the sign test that the median of `dew` is 16, and the t-test that the mean is 16. Are your results consistent with the ones from R? (There's a hint in Lab 3 about how you specify a null mean/median in `proc univariate`.)

8. In the time between 4000 BCE and 2000 BCE,[2] it is believed that there was some interbreeding between Egyptians of that era and other races. One of the few ways we have to determine this is to make measurements from what survives – for example, skull breadths. It is believed that skull breadth changes only slowly over time within a race, but skull breadths can be noticeably different among different races. This makes skull breadth a reasonable measurement to assess interbreeding.

   Two separate samples of 30 Egyptians were obtained dating from 4000BCE and 2000BCE, and their skull breadths were measured. The data are in `http://www.utsc.utoronto.ca/%7ebutler/c32/egyptians.txt`.

   (a) Read the data into R. Convince yourself that the values read in are reasonable, by looking at the data.

   (b) We need all the skull breadths in one column and the dates in a second, rather than having the data in two separate columns. Use `stack`, as in the lecture notes, to put the data in the form that we want, and convince yourself that you actually *do* have the data in that form.

   (c) Make side by side boxplots of skull breadth, one for each date. Do you have any doubts about the suitability of a t-test for comparing the skull breadths? Explain briefly.

   (d) Find the means of the two groups using `aggregate`. Save the results in a variable.

   (e) Find the difference between the two sample means,

   (f) Now we're going to do one randomization, as follows:

      i. `attach` the data frame. Shuffle the years randomly. Store the shuffling in a variable.

[1] You will need rather a long `input` line.

[2] BCE means "before common era", what some of you may know as "BC".

ii. Calculate the means for the shuffled groups, storing the results in a variable.

iii. Calculate the mean difference for the shuffled group. How does it compare to the observed difference that you calculated in (d)?

(g) Obtain 10,000 randomized mean differences, by running the commands of (f) in a loop, and saving the mean difference you get each time. (You can mimic the example in class.) Run `summary` on, and make a histogram of, your results. 10,000 randomizations will take a few seconds to run, depending on how fast your computer is. Everyone's results will be different here (the grader will be checking).

(h) Does your histogram suggest that a difference in means as big as you observed (in (e)) is likely or not? Explain briefly.

(i) Use `table` to count how many randomized mean differences are at least as big as the one you observed in (e). This, divided by 10,000, is the P-value of the randomized test. What do you conclude from this?