Assignment 5

Due Thursday October 10 at 11:59pm on Blackboard

As before, the questions without solutions are an assignment: you need to do these questions yourself and hand them in (instructions below).

The assignment is due on the date shown above. An assignment handed in after the deadline is late, and may or may not be accepted (see course outline). My solutions to the assignment questions will be available when everyone has handed in their assignment.

You are reminded that work handed in with your name on it must be *entirely your own work*.

Assignments are to be handed in on Quercus. See `https://www.utsc.utoronto.ca/~butler/c32/quercus1.nb.html` for instructions on handing in assignments in Quercus. Markers' comments and grades will be available there as well.

Start with this. I think it's likely we'll be using something from `smmr` here, so I'm loading that as well. Install it first (see the lecture notes if you need help).

```
library(tidyverse)
library(smmr)
```

1. Work through Chapter 10 of PASIAS (on matched pairs and the matched pairs sign test). This will help you with the catfood question below.
2. Work through problems 9.4 through 9.6 in PASIAS (on Mood's median test). This will help you with the Yukon wildfires problem below.
3. Which cat food do cats prefer? A pet food company was comparing two recipes, A and B. A random sample of 10 cats was taken. Two bowls of food were placed in front of each cat, with each bowl containing food prepared from one of the recipes. Trained observers gave each a score for each food, depending on the cat's behaviour and the amount of food eaten from each bowl. The data are in `http://www.utsc.utoronto.ca/~butler/assgt_data/catfood.txt`. The pet food company is interested in whether there is any evidence that cats prefer one of the recipes over the other.
   (a) (2 marks) Read in and display the data.
   (b) (2 marks) What kind of experimental design is this: matched pairs, two independent samples, or something else? Explain briefly.
   (c) (3 marks) Run a suitable t-test. What do you conclude, in the context of the data? (If you need to do any data manipulation to do the test, do that first.)
   (d) (4 marks) Make a suitable graph to assess the assumptions of your t-test. What do you conclude about the validity of your t-test? Explain briefly.
   (e) (3 marks) Use `smmr` to run a suitable sign test. How does the result compare to that of your t-test?
4. In the Yukon Territory, is more forested area being destroyed by wildfires than in the past? The data are in `http://www.utsc.utoronto.ca/~butler/assgt_data/Yukon_Wildfires.csv` as a `.csv` file. The data file contains five columns: the year, the number of wildfires caused by lightning strikes, the number

caused by humans, the total of the last two, and the total number of hectares of forest destroyed by wildfires in that year.

(a) (2 marks) Read in and display (some of) the data. Note that the variable names have Capital Letters.

(b) (2 marks) One way of comparing any of these variables in the past to the same variable more recently is to divide the time period into two parts. For example, we can compare up to (and including) 1980 with 1981 to the present (the data set goes from 1950 to 2004). In your data frame, make a new column called `recent` that is `TRUE` if the year is 1981 or greater and `FALSE` otherwise. Save the resulting data frame, and display at least some of it. (Hint: set your new variable equal to the appropriate logical condition, or, if you must, use `ifelse`.)

(c) (2 marks) Use your data frame with `recent` in it to make a suitable plot of the `Hectares` values, one that could be used to assess the assumptions for a two-sample t-test.

(d) (3 marks) Use something from `smmr` to compare the median hectares destroyed in the two time periods? What precisely do you conclude, in the context of the data? Explain briefly.

This is the end of what you need to hand in, but I originally had a couple of other things I was going to ask you to do (before realizing that the assignment was too long with them in). I've added those back below. You do **not** need to hand these in, but working through them might give you some more insight:

(e) Find the median of all the `Hectares` values.

(f) Count (using `count`) the number of `Hectares` values above (and below) the overall median for each of your two time periods (`recent` being `TRUE` or `FALSE`). Hints: (i) in `count` you can have either a column or a logical condition based on a column being greater than some value, or both; (ii) you can just type the number you obtained in the previous part.

(g) Does it look as if more recent years have a greater number of hectares of forest destroyed, on average? Explain briefly. (Of course, you know the answer now, but pretend for the moment you don't. I originally had this part in before you did the test and got the P-value.)