Assignment 3

Due Thursday September 26 at 11:59pm on Quercus

As before, the questions without solutions (here, the last two) 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.

As ever, you'll want to begin with:

```
library(tidyverse)
```

1. Work through problems 7.1 through 7.4 in Chapter 7 of PASIAS. This will prepare you for the fertilizer question below.
2. Work through problem 8.1 of PASIAS. This will prepare you for the question about the exponential distribution below.
3. Corn plants are treated with one of two fertilizers, called A and B. The amount of corn produced by each plant is measured; this is called the "yield". We are interested in seeing whether the mean yield differs between the two fertilizers. The available plants (which you can think of as a sample of "all possible plants") were randomly assigned to fertilizers. The data are in `http://www.utsc.utoronto.ca/~butler/assgt_data/ferto.txt`, the values separated by a single space.

   (a) (3 marks) Read in and display (at least some of) the data. How many plants got each fertilizer? (You should get R to figure this out rather than counting them by hand.)

   (b) (3 marks) Run a two-sample t-test to compare the mean yields of plants given the two different fertilizers. What do you conclude, in the context of the data?

   (c) (3 marks) Obtain the confidence interval from your output of the previous part. What precisely is this a confidence interval for? How is it consistent with the result of your test? Explain briefly.

   (d) (2 marks) Do you think a 99% confidence interval would contain zero for these data? Explain briefly why or why not.

   (e) (4 marks) Run a pooled t-test for these data, and, by looking at a suitable graph, explain briefly whether you prefer it to the original (Welch) test.

   (f) (2 marks) The data as I originally received it looked like this:

   ```
   ferto_wide=read_delim("ferto_wide.txt", " ")
   ```

```
## Parsed with column specification:
## cols(
##   A = col_double(),
##   B = col_double()
## )
```

```
ferto_wide
```

```
## # A tibble: 16 x 2
##        A     B
##    <dbl> <dbl>
##  1   452   546
##  2   874   547
##  3   554   774
##  4   447   465
##  5   356   459
##  6   754   665
##  7   558   467
##  8   574   365
##  9   664   589
## 10   682   534
## 11   547   456
## 12   435   651
## 13   245   654
## 14    NA   665
## 15    NA   546
## 16    NA   537
```

Explain briefly how this is not an appropriate data layout for us to use to do a two-sample *t*-test with.

4. The exponential is a continuous, positive-valued probability distribution. We will investigate its shape shortly. The R function `rexp` draws random samples from an exponential distribution. It needs two inputs: a sample size, and a parameter called the "rate" that is one over the mean.

   (a) (2 marks) Draw a random sample of 100 values from an exponential distribution with mean 10. Save the values in a data frame, and display at least some of the data frame.

   (b) (3 marks) Make a histogram of your random sample. Why might you have doubts about using a *t*-test on data from this exponential distribution? Explain briefly.

   (c) (4 marks) Generate a large number of random samples of size 100 from an exponential distribution with mean 10. For each one, run a (two-sided) *t*-test that the mean is 10. For each of those *t*-tests, extract the P-value. Count up and display how many of those P-values are less than 0.05. (This is the same idea as power by simulation.)

   (d) (3 marks) In the previous part, was the null hypothesis actually true or false? Would rejecting it be correct or some kind of error? Out of your "large number" of random samples, what percentage of them would you expect to result in rejection of the null hypothesis? What percentage actually did? Explain briefly.

   (e) (3 marks) Use simulation to estimate the power of the *t*-test to (correctly) reject a null mean of 13 when the true mean is actually 10 and the data is a sample of size 100 from an exponential distribution with rate 0.1.