Assignment 5

Due Tuesday February 25 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. You will be using something from `smmr` here, so load that as well. Install it first (see the lecture notes if you need help).

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

Hand in question 2.

1. Work through Chapter 10 of PASIAS (on matched pairs and the matched pairs sign test).
2. Some people believe that the full moon can cause changes in behaviour. In one study, aggressive behaviour in dementia patients (in a hospital) was observed. The experimenters suspected that there might be more aggressive behaviour on days close to a full moon. The number of incidents of aggressive behaviour for each patient on each day was recorded. After that, each day was classified as a "moon day" (within 3 days of a full moon) or an "other day" (not close to a full moon), and, for each patient, the mean number of aggressive incidents on moon days and other days was recorded. These data are in `http://ritsokiguess.site/STAC33/moon.csv` as a `.csv` file. (Note: we only have averages for moon days and other days for each patient. If we had numbers of aggressive incidents for each day and each patient, we would be dealing with repeated measures, multiple observations per individual under different conditions. But we don't; just analyzing the means is simpler.)

   (a) (2 marks) Read in and display (some of) the data.

   Solution: Everything is familiar here:

```
my_url="http://ritsokiguess.site/STAC33/moon.csv"
aggression <- read_csv(my_url)
```

```
## Parsed with column specification:
## cols(
##  patient = col_double(),
##  moonday = col_double(),
##  otherday = col_double()
## )
```

```
aggression
```

```
## # A tibble: 15 x 3
##    patient moonday otherday
##      <dbl>   <dbl>    <dbl>
##  1       1    3.33     0.27
##  2       2    3.67     0.59
##  3       3    2.67     0.32
##  4       4    3.33     0.19
##  5       5    3.33     1.26
##  6       6    3.67     0.11
##  7       7    4.67     0.3
##  8       8    2.67     0.4
##  9       9    6        1.59
## 10      10    4.33     0.6
## 11      11    3.33     0.65
## 12      12    0.67     0.69
## 13      13    1.33     1.26
## 14      14    0.33     0.23
## 15      15    2        0.38
```

Extra: there are, as expected, three columns: an identifier for each patient, and two measurements for each one, the mean number of aggressive incidents on moon days and other days. This is, as it turns out, exactly the format we need; there is no need to reshape the data using `pivot_longer` or any of the other things we see later.

(b) (2 marks) Explain briefly, in the context of this data set, why this is a matched pairs experiment.

Solution: Each person has two measurements, or the two columns of measurements `moondays` and `otherdays` are related because they come from the same person, or something else like that.

The point is that these are *not* two separate samples of people; they are the same people, measured twice. You need to be clear on the difference, because the appropriate analysis depends on the way the experiment was designed.

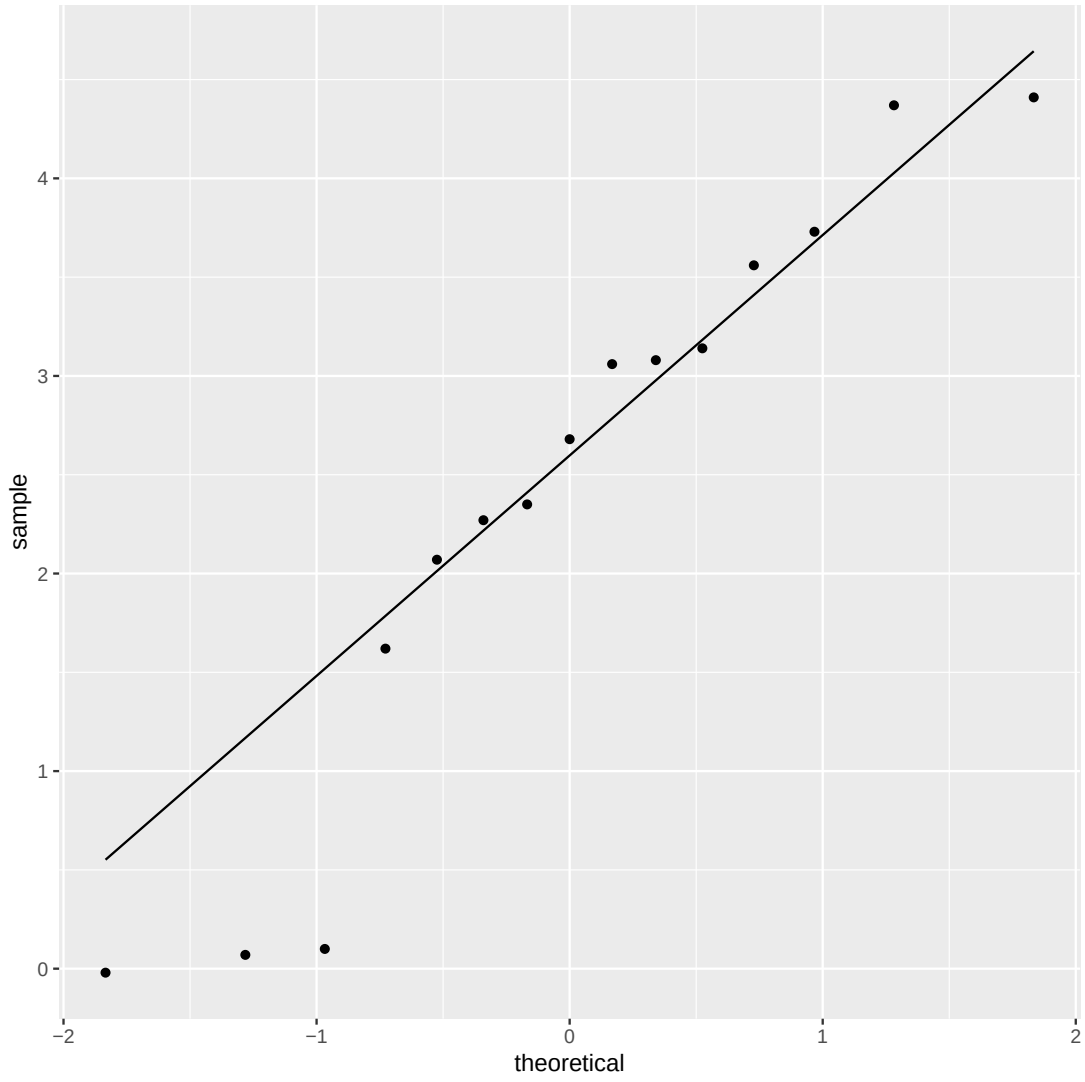
(c) (3 marks) What is the principal assumption behind the matched pairs t-test? Obtain a graph that would enable you to assess this assumption. (You don't need to interpret the graph until later.)

Solution: The principal assumption is that the *differences* between the two measurements for each individual have (approximately) a normal distribution. Thus the first thing you need to

do, whatever plot you plan to draw, is to calculate the differences. It is best to create a column of differences in the data frame and save it, since we'll be using it again later.

The plot I like best is a normal quantile plot of the differences:

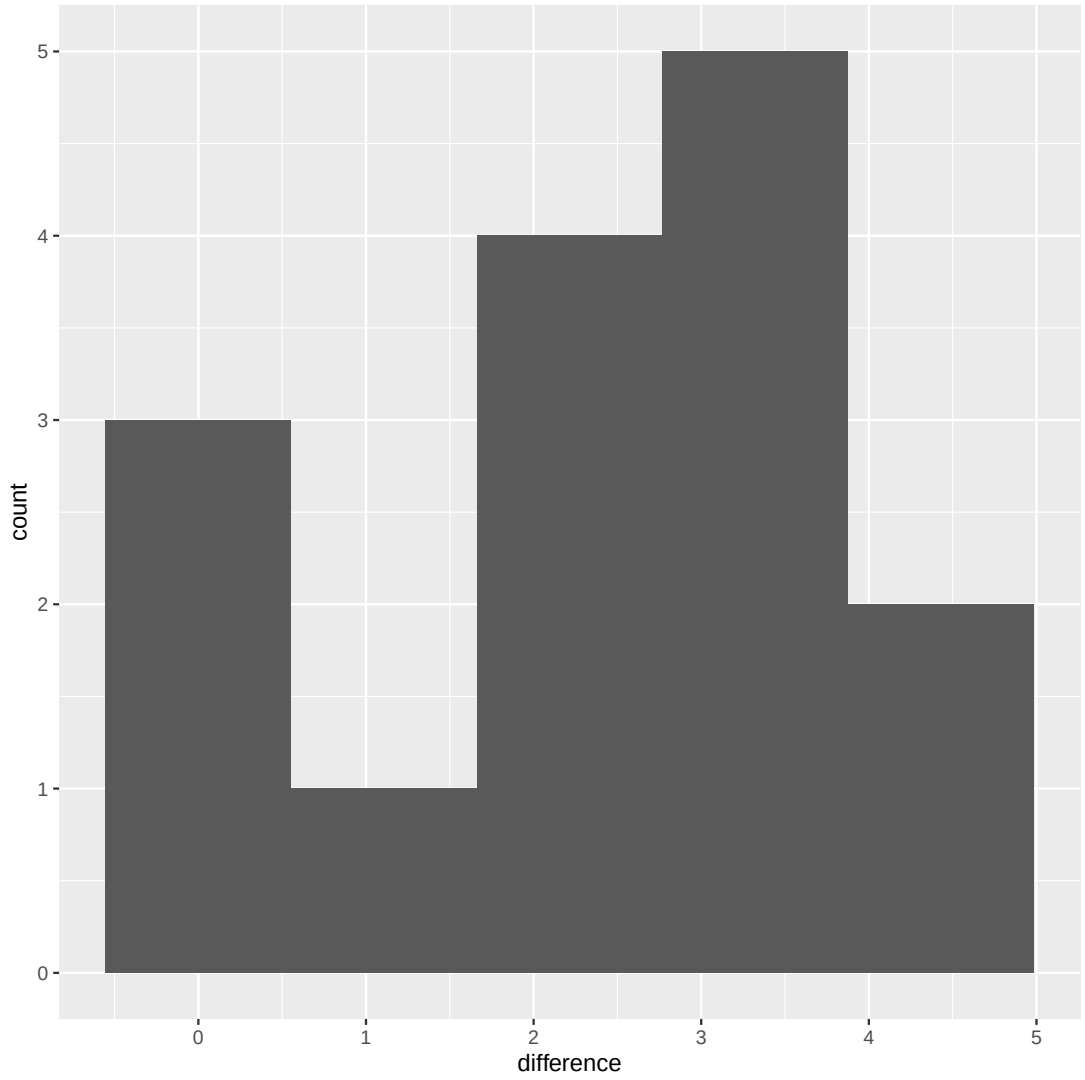
```
aggression %>% mutate(difference=moonday-otherday) -> aggression2
ggplot(aggression2, aes(sample=difference)) +
    stat_qq() + stat_qq_line()
```



If you calculate the differences and then pipe the resulting data frame directly into `ggplot`, you are making extra work for yourself later (but that is also OK for this part). Calculating the differences *outside* of a data frame will work, but will be more difficult. Try to stick to the `tidyverse` way of doing things.

An alternative here would be the standard one-quantitative-variable plot: a histogram of the differences. There are only 15 observations, so you can't really justify more than about 5 bins (which will give you a fairly crude picture of the shape):

```
ggplot(aggression2, aes(x=difference)) + geom_histogram(bins=5)
```



The interpretation of the histogram, however, is similar to the interpretation of the normal quantile plot. See later.

(d) (3 marks) Carry out a suitable matched-pairs t-test and interpret the results. (Do this even if you think some other test is better. The critical analysis comes later.)

Solution: Remember that the experimenters suspected, if anything, that the full moon would *increase* the number of aggressive incidents, so you need an appropriate *one-sided* test.

That goes this way, most simply:

```
with(aggression, t.test(moonday, otherday, paired=T, alternative="greater"))
```

```
##
##  Paired t-test
##
## data:  moonday and otherday
## t = 6.4518, df = 14, p-value = 7.591e-06
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.768559      Inf
## sample estimates:
## mean of the differences
##                2.432667
```

Alternatives: use the dollar sign twice to say which data frame you are getting the columns from; use the columns the other way around and then have `alternative="less"` (in which case, your test statistic will have the opposite sign from mine but the P-value will be the same). Another alternative is to directly use the column of differences that you computed and saved (if you did):

```
with(aggression2, t.test(difference, mu=0, alternative="greater"))
```

```
##
##  One Sample t-test
##
## data:  difference
## t = 6.4518, df = 14, p-value = 7.591e-06
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  1.768559      Inf
## sample estimates:
## mean of x
##  2.432667
```

In this case, you should specify a null mean. The final possibility is if you calculated the differences the other way around (so that most of them will be negative), and then you will need `alternative="less"`.

Whichever one of those variants you used (and it doesn't matter which one you did, as long as you got the same P-value), you come to the same conclusion. The null hypothesis is that there are the same average number of aggressive incidents on moon days and other days, and the alternative is that there are *more* incidents on moon days. Here, the P-value 0.0000076 is very small, so we reject the null in favour of the alternative, and conclude that there are more aggressive incidents on moon days than other days (on average), for all dementia patients.

You need to get all the way to the end to get the points for the conclusion. "Reject the null hypothesis" is only about halfway towards the final answer, because you as an applied statistician have to tell the world what rejecting the null hypothesis *means* in terms of a decision or an action. In this example, the conclusion might imply an action of having more staff in the dementia ward of a hospital on days around a full moon as a safeguard against the expected increase in aggressive incidents.

(e) (3 marks) Carry out a suitable matched-pairs sign test on these data, and interpret the results. (Do this even if you think some other test is better.)

Solution: For this one, you need the differences (whether or not you used them in the previous part). If you didn't save them before, you'll need to calculate them again. If that's the case, the coder in you should be reminded that saving them before would have been a good idea (so you go back and do it there).

I'll use my data frame `aggression2` with the differences in it. This is where you need to have `smmr` ready to go:

```
sign_test(aggression2, difference, 0)
```

```
## $above_below
## below above
##     1    14
##
## $p_values
##   alternative      p_value
## 1       lower 0.9999694824
## 2       upper 0.0004882812
## 3   two-sided 0.0009765625
```

(data frame, column, null median). This gives you three P-values, of which you want the one labelled `upper`. This is 0.00049, which is also very small, so we again reject the null hypothesis that the median difference is 0, in favour of the alternative that the median is greater than zero: that is, that is, that the median number of aggressive incidents is greater on moon days than it is on other days.

If you took your differences the other way around, you need the P-value marked `lower` because, according to the experimenters' beliefs, most of the other-day values should be lower than the moon-day values.

Extra: you'll note that the conclusion is the same both times, but the P-value is (even) smaller for the matched-pairs t-test.

Extra extra: the top table in the output explains *why* the P-value for the matched pairs sign test is so small: there are 14 positive differences and only 1 negative one. This is so far from an even split that it would be very unlikely if the differences were equally likely to be positive or negative.

(f) (2 marks) Which of the two tests you did is better? Explain briefly.

Solution: Go back to the graph you drew, and ask yourself whether it looks sufficiently close to normal, given the sample size. If it does, you should prefer the t-test; if not, the sign test.

My take is that, on the normal quantile plot, the three lowest values are a bit too low. Given that, I have doubts about the normal distribution, and therefore I prefer the sign test.

This is also the same impression I would get from the histogram (if that's what you drew): if the bin on the left had one or zero observations instead of three, I would be happy, but those three observations could be low outliers, or the distribution could be bimodal, either of which would argue against the t-test. Thus I would prefer the matched-pairs sign test.

However, I would also entertain the other point of view, properly justified. For example, you could say that those three low observations on the normal quantile plot are not so far from the line, particularly not the lowest one. We have a sample size of 15, which, while not large, is also not tiny, so we get a little help from the Central Limit Theorem. Also, from this size sample, you

might get this kind of deviation from normality just by chance, if the population of differences actually is normal. Thus, by (at least some of) this argument, I prefer the matched-pairs t-test.

You could also argue this from the histogram, perhaps by saying that the sample size is smallish and so we cannot expect the histogram to look too much more normal than this, even if the population of differences *is* normal.

This kind of thing happens a lot in this course. There are often not right answers, but as long as you get an answer that is justifiable and logically sensible, you are good. This is one of those cases that you could argue either way. But if I think the decision is clear one way or the other, I would expect you to agree with me.

Extra: since the conclusion (reject the null; conclude more aggressive incidents on moon days) is the same both ways, it doesn't really matter which test you do. It's a very clear-cut decision either way.

Extra extra: some people assess whether the normal quantile plot is non-normal enough by generating eight more normal quantile plots from data which actually *is* normal (using the same sample size), and plotting all nine plots on a 3×3 array (which you could imagine doing with `facet_wrap`). Then you ask *somebody else* to look at the array of plots and pick out the odd one. If they pick yours, that's evidence that your plot is non-normal; if they pick one of the actually-normal plots, this means that your plot is indistinguishable from normal.

Let me have a go at that.

Let's first generate the eight real normals (and then bash the real data into shape to fit that). The first thing I need is the mean and SD of our data, so I can generate the real normals to be the same:

```
aggression2 %>% summarize(m=mean(difference), s=sd(difference))
```

```
## # A tibble: 1 x 2
##       m     s
##   <dbl> <dbl>
## 1  2.43  1.46
```

Now generate eight actual normal samples with that mean and SD:

```
tibble(i=1:8) %>%
    mutate(norm=map(i, ~rnorm(15, 2.43, 1.46))) %>%
    unnest(norm) -> d8
d8
```

```
## # A tibble: 120 x 2
##        i  norm
##    <int> <dbl>
##  1     1 4.80
##  2     1 1.34
##  3     1 2.04
##  4     1 1.41
##  5     1 2.74
##  6     1 3.47
##  7     1 0.856
##  8     1 3.59
##  9     1 2.44
## 10     1 4.03
## # ... with 110 more rows
```

I did this by creating a list-column each element of which is an entire normal random sample. The `map` idea is the same for-each as we saw when simulating power.

Now we have to glue the original data onto that. This will be easiest if we reconstitute it to have the same two columns with the same two names. The `rep` is "repeat": we need to repeat the sample number 9 15 times because there are 15 differences.[1]

```
tibble(i=rep(9, 15)) %>%
    mutate(norm=aggression2$difference) -> d1
d1
```

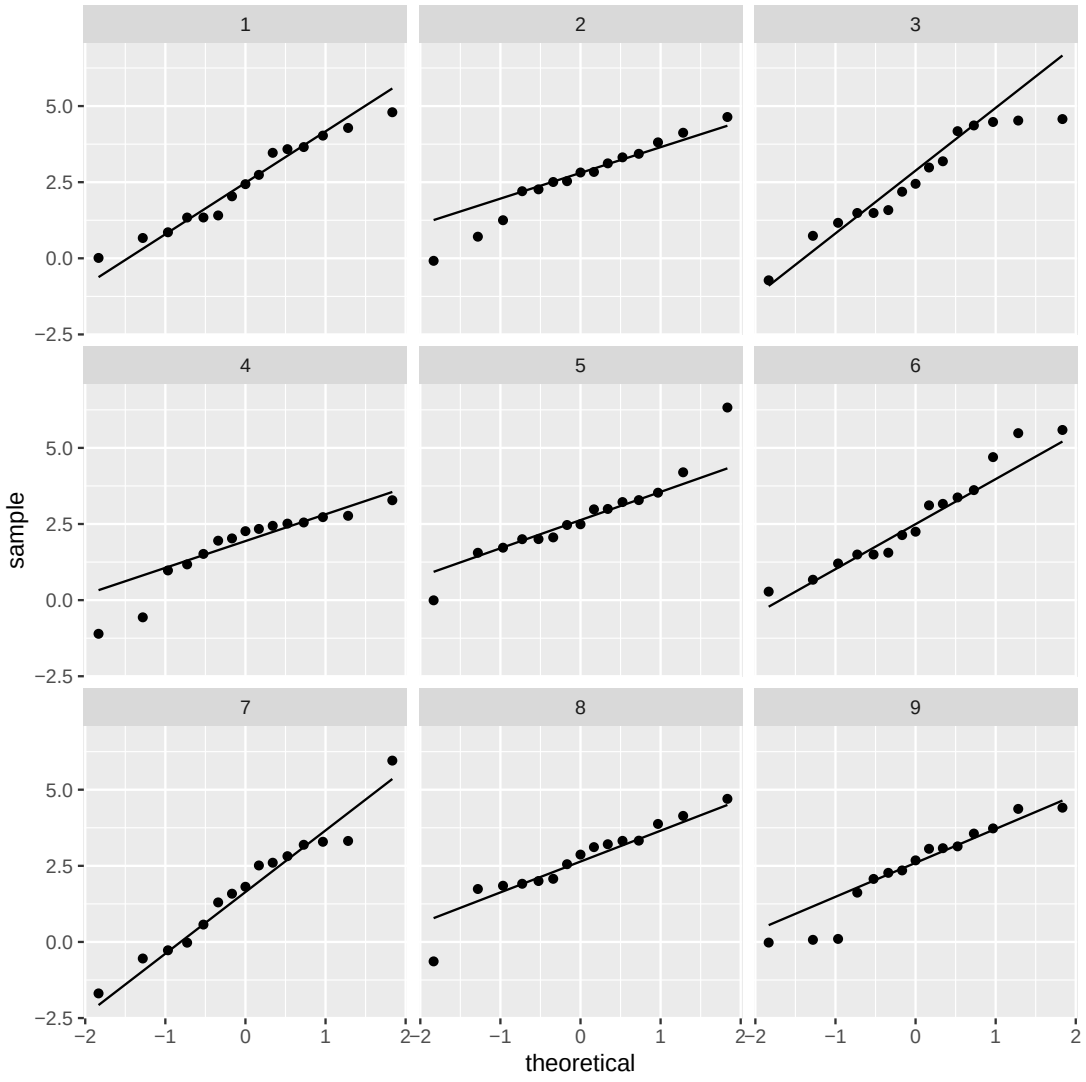
```
## # A tibble: 15 x 2
##        i    norm
##    <dbl>   <dbl>
##  1     9  3.06
##  2     9  3.08
##  3     9  2.35
##  4     9  3.14
##  5     9  2.07
##  6     9  3.56
##  7     9  4.37
##  8     9  2.27
##  9     9  4.41
## 10     9  3.73
## 11     9  2.68
## 12     9 -0.0200
## 13     9  0.07
## 14     9  0.1
## 15     9  1.62
```

then to glue them together we can use `bind_rows` (since the columns have the same names and are in the same order):

```
d <- bind_rows(d8, d1)
```

We should shuffle the sample numbers (since we know the real data is sample number 9), but I am too lazy to do that:

```
ggplot(d, aes(sample=norm)) + stat_qq() +
    stat_qq_line() + facet_wrap(~i)
```



Now, look at these nine normal quantile plots. 1 through 8 are actual normal samples, and 9 is our data. Can you tell the difference between 9 and the rest?

To my eyes, some of the samples look definitely more normal than ours: 1 and 6 definitely, 7 maybe. But some of them look quite similar to our data: 2 with three low values, and 4 with two, but the low values are more extreme than ours. And some of them look non-normal in different ways: 3 has a short upper tail; 5 has outliers at both ends; 8 has a low outlier. The overall picture I see from this is that our data is somewhere in the middle of the non-normality we would see from *actual normal* data, and so there is no reason to call our data non-normal.

This might be a bit of a surprise to you, but it goes to show that data that really is normal can itself look quite non-normal, especially if the sample size (here $n = 15$) is not large.

Notes

[1]If I had obtained the 15 values `norm` first, then I could have created the column `i` as just *one* value 9, and R would have repeated that 15 times to make it the same length as `norm`. But then the columns would have come out in the wrong order.