Assignment 6

Due Thursday October 31 at 11:59pm on Quercus

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.

Begin with the usual:

```
library(tidyverse)
```

Hand in question 2 below.

1. Work through chapter 13 of PASIAS.

2. A cosmetic company created a small trial of a new cream for treating skin blemishes. It measured the effectiveness of the new cream (`New`) compared to the leading cream on the market (`Old`) and a placebo (`Control`). Thirty people were put into three groups of 10 at random, although just before the trial began 2 people from the control group and 1 person from the test group for the existing cream dropped out. The data as originally recorded are in `http://www.utsc.utoronto.ca/~butler/assgt_data/cosmetic-wide.txt`. The values in each column are the number of blemishes removed from each person during the trial (so a higher number is better).

   For this question, write a report, with a title, an Introduction, an Analysis and a Conclusion, that would make sense to someone who didn't know about this data set or what you were trying to do with it. Your report needs to read like something *you* wrote, not like some questions that I gave you. Parts (a), (b), etc, below, are to guide you about what to include in the report; *do not* include the part names in your report. (The grader will check that you have discussed the appropriate things somewhere in your report.) The grader also has 3 discretionary marks to award for the quality of writing, on the following scale: 3: excellent writing with few to no spelling or grammatical errors; 2: writing that is easy to read but contains grammatical errors; 1: writing that is hard to understand; 0: no attempt was made to turn this assignment into a report.

   (a) (4 marks) Begin your report with an Introduction that describes, *in your own words*, what the data set is, what is being measured, and what you hope to find out.

   > **Solution:** You need to say something about the data set being how effective a new cream is at removing skin blemishes, and that it is a comparison between a new cream, the leading cream on the market, and a placebo. The variable being measured is the number of (skin) blemishes removed. The aim of the study is to compare the new cream with the other creams in terms

of blemish removal (or, if you like, to compare the creams with each other). Or, say that you hope to find out whether the new cream is better than the other creams at blemish removal.

Extra: the way we'll do the analysis, we are really comparing all the creams with each other, but it's certainly perfectly good to say that you are trying to find out whether the new cream is better (at removing blemishes).

(b) (2 marks) Begin your Analysis section. Read in and display the data. Why are there some `NA` values?

> **Solution:** The usual, data values being separated by single spaces:
>
> ```
> my_url="http://www.utsc.utoronto.ca/~butler/assgt_data/cosmetic-wide.txt"
> cosmetic=read_delim(my_url, " ")
> ```
>
> ```
> ## Parsed with column specification:
> ## cols(
> ##   New = col_double(),
> ##   Old = col_double(),
> ##   Control = col_double()
> ## )
> ```
>
> ```
> cosmetic
> ```
>
> ```
> ## # A tibble: 10 x 3
> ##       New   Old Control
> ##     <dbl> <dbl>   <dbl>
> ##  1     81    48      18
> ##  2     32    31      49
> ##  3     42    25      33
> ##  4     62    22      19
> ##  5     37    30      24
> ##  6     44    30      17
> ##  7     38    32      48
> ##  8     47    15      22
> ##  9     49    40      NA
> ## 10     41    NA      NA
> ```
>
> The `NA` values represent the people who dropped out, and thus their measurements are missing. There are actually 27 ($= 10 + 9 + 8 = 30 - 3$) people who completed the study.
>
> Extra: the same thing would have to be done any time the three treatment groups had uneven numbers of subjects (for whatever reason), because the columns of a data frame all have to be the same length (and if they are not, you have to add NA values at the end until they are). If you recorded the data in a spreadsheet, reading it in with `read_csv` or using `read_excel`, blank cells would be translated to `NA` by this process.

(c) (2 marks) These data are not in a suitable layout for making plots or for running any kind of analysis of variance. Why not? Explain briefly.

> **Solution:** This is not tidy format. The column names are *levels* of a categorical variable (that might be called Treatment), and the numbers are all numbers of blemishes (recorded under the different treatments). We need to organize it so that there is one column called `blemishes` and

another column called `treatment` that says which treatment group that number of blemishes belongs to.

Say a sensible fraction of that: what makes the data untidy, what it would look like if it were tidy, or both.

(d) (3 marks) Use something from the `tidyverse` to get the data into the right format, and show that this is what you have done. Find out how to get rid of the missing values, so that your final data frame has 27 rows for the 27 non-missing observations.

> **Solution:** This is mostly a textbook application of `pivot_longer` or `gather`. (Either is good.) I am going to save my result, which is probably a good idea because I'll be using it again later.
>
> `pivot_longer`, including the removal of the missing values, goes like this:
>
> ```
> cosmetic %>%
>     pivot_longer(everything(), names_to="treatment",
>                  values_to="blemishes", values_drop_na=T) -> cosmetic_tidy
> cosmetic_tidy
>
> ## # A tibble: 27 x 2
> ##    treatment blemishes
> ##    <chr>         <dbl>
> ##  1 New              81
> ##  2 Old              48
> ##  3 Control          18
> ##  4 New              32
> ##  5 Old              31
> ##  6 Control          49
> ##  7 New              42
> ##  8 Old              25
> ##  9 Control          33
> ## 10 New              62
> ## # ... with 17 more rows
> ```
>
> This is now tidy.
>
> Extra: for you (not for the report), the inputs to `pivot_longer` are:
>
> - the columns to make longer. Here, that is all of them; you could also use something like `New:Control` to select them.
> - the current column names will go to a new column called `treatment`.
> - the current column values will go to a new column called `blemishes`.
> - drop any values that are missing. (This you will probably have to look up.)
>
> For the `gather` approach, you have to remember which order the inputs go. The inputs are: what makes the columns different (different treatments), what makes them the same (all numbers of blemishes), and the columns to gather up. For the first two, you get to choose names for the new columns; for the last one, you can either say something like `New:Control` or use the handy select-helper `everything()` which gathers up all the columns. To get rid of the missing values, you'll have to explore the options on `gather` (eg. at https://tidyr.tidyverse.org/reference/gather.html); the one you want is `na.rm`, `rm` being shorthand for "remove":

```
cosmetic %>% gather(treatment, blemishes,
                    everything(), na.rm=T) -> cosmetic_tidy2
cosmetic_tidy2

## # A tibble: 27 x 2
##    treatment blemishes
##    <chr>         <dbl>
##  1 New              81
##  2 New              32
##  3 New              42
##  4 New              62
##  5 New              37
##  6 New              44
##  7 New              38
##  8 New              47
##  9 New              49
## 10 New              41
## # ... with 17 more rows
```

27 rows, and no missing values.

Extra: it doesn't actually hurt if you leave the missing values in (you'll get a warning when you draw a graph, but the actual real data will still be plotted). But for this assignment, I wanted you to find out how to get rid of them.

(e) (3 marks) Make a suitable plot that will enable you to assess the distributions of values in each treatment group. What do you see?

Solution: The obvious thing, I think, is a boxplot:

```
ggplot(cosmetic_tidy, aes(x=treatment, y=blemishes)) + geom_boxplot()
```
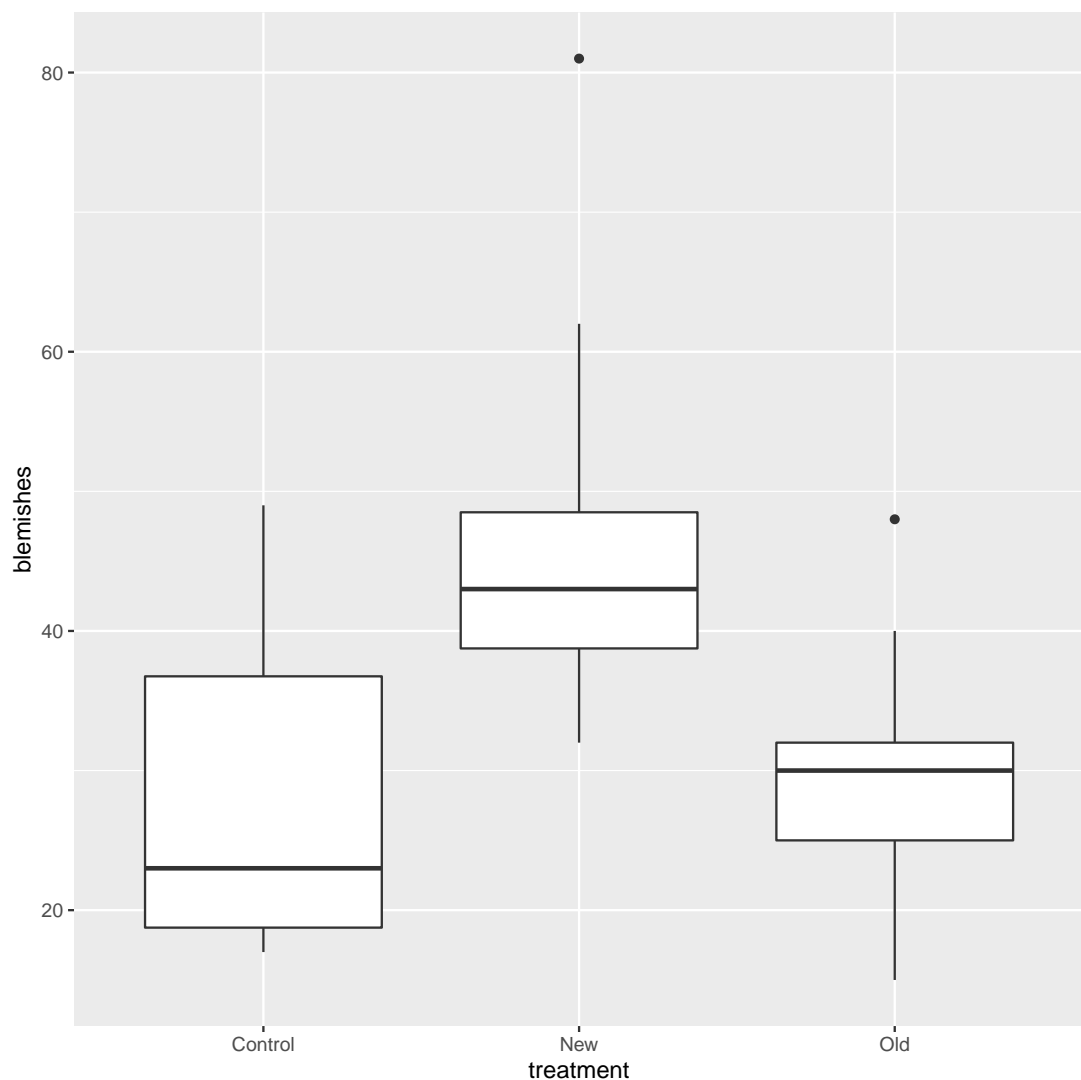
I see: outliers in the New and Old groups, right-skewness in the Control group (so overall non-normality), and the spread for the Control group looks bigger (although it is difficult to be sure with the outliers).

I'd expect you to note the outliers and the skewness, at least.

You might also be looking ahead to ANOVA assumptions and realize that normality is going to be of some importance; that would tell you to make normal quantile plots:

```
ggplot(cosmetic_tidy, aes(sample=blemishes)) + stat_qq() + stat_qq_line() +
  facet_wrap(~treatment)
```
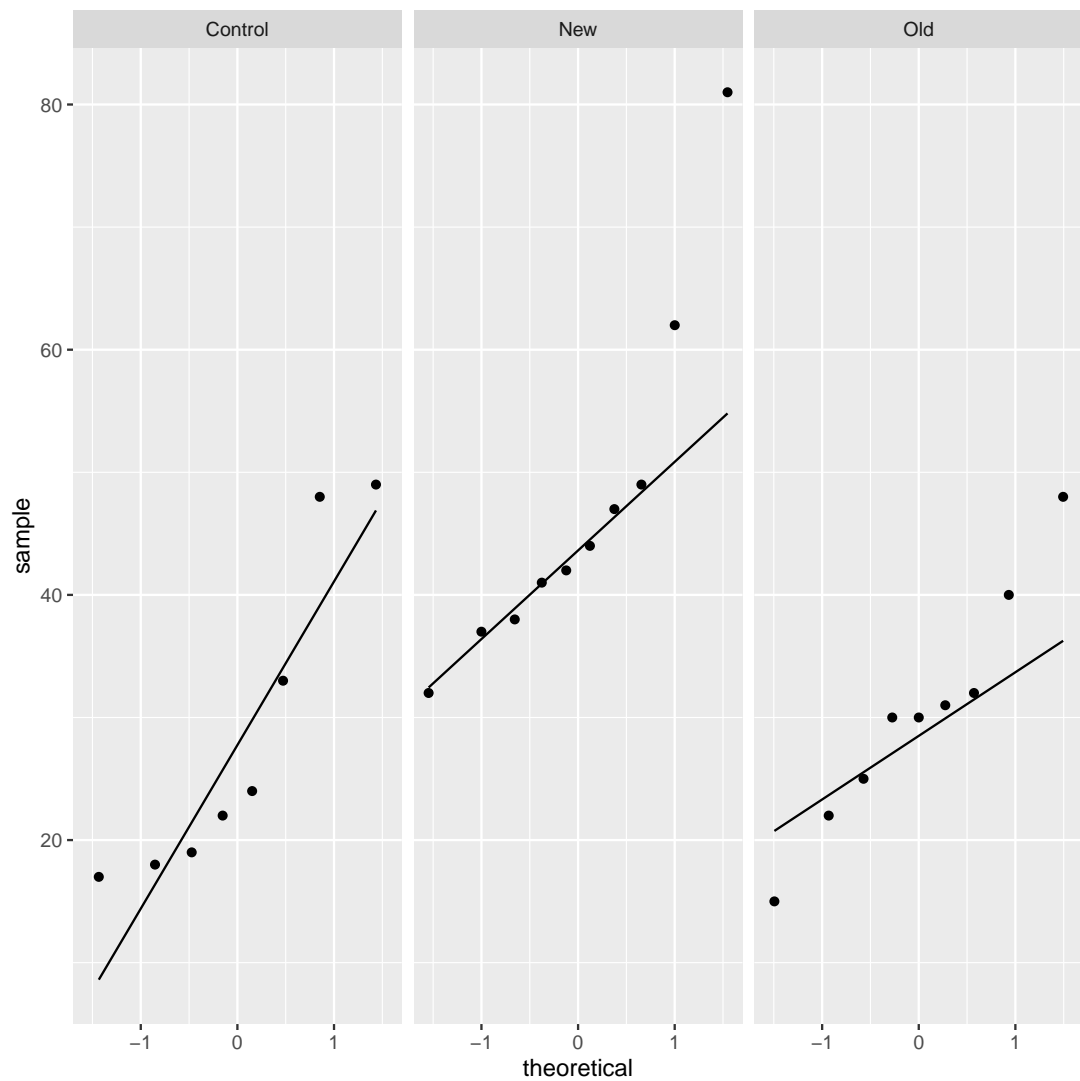
The Control group has a shape that you might read as curved (skewness), or you might read it as an S-bend (bimodal), and the New and Old groups might even have two outliers at the top, not just one. If you go this way, say something about failure of normality in each (or at least one) of the groups and you're good.

(f) (4 marks) Carry out a suitable analysis to compare the number of blemishes among the treatment groups. Do an overall test, and follow up with comparisons between the groups if it is a sensible thing to do. If your comparisons between groups seem to be inconsistent with your overall test, discuss briefly what seems to have happened.

Solution:

Your choices here are:

- regular ANOVA

- Welch ANOVA

- Mood's median test

I think the clear outliers rule out sufficient normality (with 10 or fewer observations in each group), so to my mind the only sensible choice is the Mood's median test:

```
library(smmr)
median_test(cosmetic_tidy, blemishes, treatment)

## $table
##          above
## group      above below
##   Control     2     5
##   New         9     1
##   Old         2     7
##
## $test
##        what         value
## 1 statistic 10.463492063
## 2        df  2.000000000
## 3   P-value  0.005344186
```

The P-value of 0.0053 is definitely less than 0.05 (0.01, whatever), so the median numbers of blemishes in the three treatment groups are not all the same. (This is, strictly speaking, an inference to the populations of "all possible people" who might have ended up in one of the treatment groups.)

To find out which differ from which, we need the pairwise median tests:

```
pairwise_median_test(cosmetic_tidy, blemishes, treatment)

## # A tibble: 3 x 4
##   g1      g2     p_value adj_p_value
##   <chr>   <chr>    <dbl>       <dbl>
## 1 Control New     0.0578       0.173
## 2 Control Old     0.447        1
## 3 New     Old     0.0184       0.0553
```

Curiously, *none* of the groups are significantly different! (This kind of thing is unusual but possible.) I think what happened here is that New vs. Old is close to significance, and only failed to be significant because of the (Bonferroni) adjustment for doing three tests at once. (There are other, better, adjustments than Bonferroni, but that's what I coded in smmr.) Also, we are not helped by the groups being small: 8 or 9 or 10 people in each treatment. Because each observation gives so little information (it's above or below the overall median), it helps to have plenty of data for Mood's median test.

Say that none of the groups are significantly different and why that surprises you, or say that New vs. Old is almost significant, or something like that.

As I said, I don't think there's any real alternative to Mood's median test for data like this. If you did one of the other analyses correctly, expect to get about half the points. I include them here so you can check what you did (if you went this way), not because I think they are in any way valid:

```
blemish.1=aov(blemishes~treatment, data=cosmetic_tidy)
summary(blemish.1)

##            Df Sum Sq Mean Sq F value  Pr(>F)
## treatment   2   1986   992.9   6.242 0.00657 **
## Residuals  24   3818   159.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

TukeyHSD(blemish.1)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = blemishes ~ treatment, data = cosmetic_tidy)
##
## $treatment
##                   diff        lwr       upr     p adj
## New-Control  18.550000   3.610044 33.489956 0.0130468
## Old-Control   1.583333 -13.721066 16.887732 0.9639276
## Old-New     -16.966667 -31.438169 -2.495165 0.0194245
```

On this, New is better than both the others.

Or:

```
oneway.test(blemishes~treatment, data=cosmetic_tidy)

##
##  One-way analysis of means (not assuming equal variances)
##
## data:  blemishes and treatment
## F = 5.3282, num df = 2.000, denom df = 15.251, p-value = 0.01759

library(PMCMRplus)
gamesHowellTest(blemishes~factor(treatment), data=cosmetic_tidy)

##
## Pairwise comparisons using Games-Howell test

## data:  blemishes by factor(treatment)

##     Control New
## New 0.03    -
## Old 0.96    0.02

##
## P value adjustment method:  none

## alternative hypothesis:  two.sided
```

Same again.

(g) (4 marks) End your report with a Conclusion that summarizes what you found out in the analysis, and how it answers your questions from the Introduction (or fails to do so). If you were able to do this study a second time, would you change anything? Why?

Solution: The real question of interest for this report is how the New cream compares with the others. It is, according to the pairwise median tests, almost significantly better than the Old cream (the leading cream on the market) but not significantly better than the Control cream (placebo). (If you thought one of the other tests was better, you are able to say that the New cream is better than both the others.)

I think you are definitely entitled to say that the study should have used a larger sample size! That would have made the conclusions clearer. This is especially so given that we used Mood's median test, which doesn't use the data very efficiently (only whether each data value was above or below the overall median), so you tend to need a larger sample size for Mood's median test than for the others.