Assignment 5

Due Thursday October 19 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. Athletes are concerned with measuring their body fat percentage. Two different methods are available: one using ultrasound, and the other using X-ray technology. We are interested in whether there is a difference in the mean body fat percentage as measured by these two methods, and if so, how big that difference is. Data on 16 athletes are at `http://www.utsc.utoronto.ca/~butler/c32/bodyfat.txt`.

   We saw this data set before.

   (a) Read in the data again.

   > **Solution:** This kind of thing. Since you looked at the data (didn't you?), you'll know that the values are separated by single spaces:

```
myurl="http://www.utsc.utoronto.ca/~butler/c32/bodyfat.txt"
bodyfat=read_delim(myurl," ")
```

```
## Parsed with column specification:
## cols(
##   athlete = col_integer(),
##   xray = col_double(),
##   ultrasound = col_double()
## )
```

```
bodyfat
```

```
## # A tibble: 16 x 3
##    athlete  xray ultrasound
##      <int> <dbl>      <dbl>
##  1       1  5          4.75
##  2       2  7          3.75
##  3       3  9.25       9
##  4       4 12         11.8
##  5       5 17.2       17
##  6       6 29.5       27.5
##  7       7  5.5        6.5
##  8       8  6          6.75
##  9       9  8          8.75
## 10      10  8.5        9.5
## 11      11  9.25       9.5
## 12      12 11         12
## 13      13 12         12.2
## 14      14 14         15.5
## 15      15 17         18
## 16      16 18         18.2
```

(b) Calculate the differences, and make a normal quantile plot of them. Is there any evidence that normality of differences fails? Explain briefly.

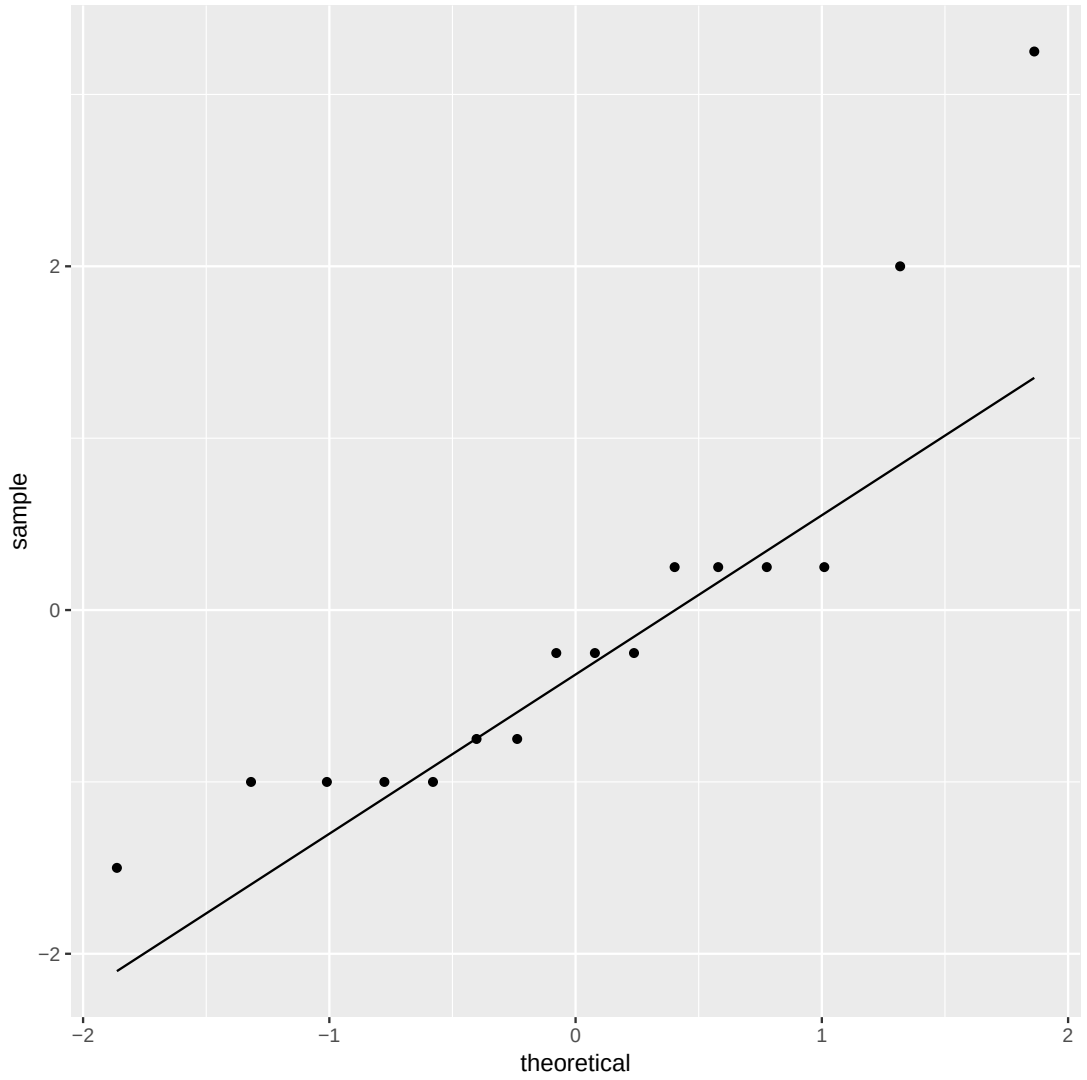
Solution: This is a good place to look ahead. We'll need the differences in two places, most likely: first for the normal quantile plot, and second for the matched-pairs sign test. So we should calculate and save them first:

```
bodyfat %>% mutate(diff=xray-ultrasound) -> bodyfat2
```

I seem to be using a 2 on the end to name my dataframe-with-differences, but you can use whatever name you like.

Then, not forgetting to use the data frame that we just made:

```
ggplot(bodyfat2,aes(sample=diff))+stat_qq()+stat_qq_line()
```



This is showing a little evidence of skewness or outliers (depending on your point of view: either is good). The lowest and highest values are both too high, and the pattern of points on the plot is kind of curved (which would be evidence of skewness). Or you could say that the two highest values are too high, with the other values being more or less in line (that would be evidence of outliers at the upper end). I like outliers better than skewness, since those bottom-end points are not far off the line. I would also accept "no substantial problems", if you can make the case that those two highest points are not too far off the line. With only 16 observations as we have here, even truly normal data would stray off the line a bit.

As ever, your explanation is more important than your conclusion. Can you justify what you think?

If you took your differences the other way around, as `ultrasound` minus `xray`, your plot will also be the other way around, with the "outliers" at the bottom. That's good too.

(c) Previously, we did a matched-pairs t-test for these data. In the light of your normal quantile plot, do you think that was a good idea? Explain briefly.

Solution: We are looking for the differences to be approximately normal, bearing in mind that we have a sample of size 16, which is not that large. Say what you think here; the points, if I were giving any here, would be for the way in which you support it.

The comment I made before when we did a matched-pairs t-test was that the P-value was so large and non-significant that it was hard to imagine any other test giving a significant result. Another way of saying that is that I considered these differences to be "normal enough", given the circumstances.

You might very well take a different view. You could say that these differences are clearly not normal, and that the sample size of 16 is not large enough to get any substantial help from the Central Limit Theorem. From that point of view, running the t-test is clearly not advisable.

(d) Use the sign test appropriately to compare the two methods for measuring body fat. (Use `smmr` if you wish.) What do you conclude, as ever in the context of the data?

Solution: That means using a sign test to test the null hypothesis that the median difference is zero, against the alternative that it is not zero. (I don't see anything here to indicate that we are looking only for positive or only for negative differences, so I think two-sided is right. You need some reason to do a one-sided test, and there isn't one here.)

Remembering again to use the data frame that has the differences in it:

```
sign_test(bodyfat2, diff, 0)
```

```
## $above_below
## below above
##    10     6
##
## $p_values
##   alternative   p_value
## 1       lower 0.2272491
## 2       upper 0.8949432
## 3   two-sided 0.4544983
```

The two-sided P-value is 0.4545, so we are nowhere near rejecting the null hypothesis that the median difference is zero. There is no evidence that the two methods for measuring body fat show any difference on average.

The table of aboves and belows says that there were 6 positive differences and 10 negative ones. This is not far from an even split, so the lack of significance is entirely what we would expect.

Extra: this is the same conclusion that we drew the last time we looked at these data (with a matched-pairs t-test). That supports what I said then, which is that the t-test was so far from being significant, that it could be very wrong without changing the conclusion. That is what seems to have happened.

2. The data for this question are in `http://www.utsc.utoronto.ca/~butler/c32/cereal-sugar.txt`. The story here is whether breakfast cereals marketed to children have a lot of sugar in them; in particular, whether they have more sugar on average than cereals marketed to adults.

(a) Read in the data (to R) and display the data set. Do you have a variable that distinguishes the children's cereals from the adults' cereals, and another that contains the amount of sugar?

Solution:

```
my_url="http://www.utsc.utoronto.ca/~butler/c32/cereal-sugar.txt"
cereals=read_delim(my_url," ")
```

```
## Parsed with column specification:
## cols(
##  who = col_character(),
##  sugar = col_double()
## )
```

```
cereals
```

```
## # A tibble: 40 x 2
##    who      sugar
##    <chr>    <dbl>
##  1 children  40.3
##  2 children  55
##  3 children  45.7
##  4 children  43.3
##  5 children  50.3
##  6 children  45.9
##  7 children  53.5
##  8 children  43
##  9 children  44.2
## 10 children  44
## # ... with 30 more rows
```

The variable `who` is a categorical variable saying who the cereal is intended for, and the variable `sugar` says how much sugar each cereal has.

(b) Calculate the mean sugar content for each group of cereals (the adults' ones and the children's ones). Do they look similar or different?

Solution: `group_by` and `summarize`:

```
cereals %>% group_by(who) %>%
  summarize(sugar_mean=mean(sugar))
```

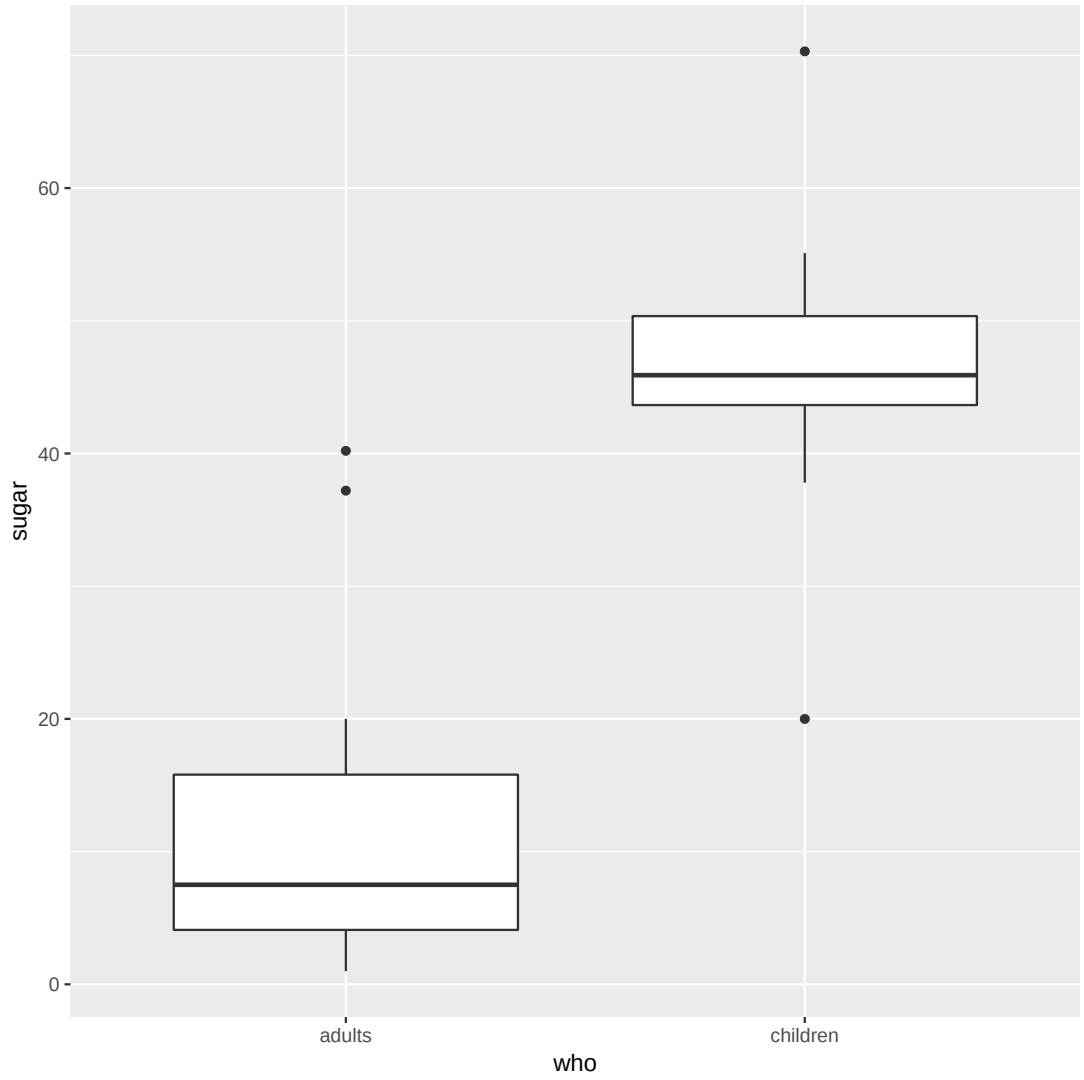
```
## # A tibble: 2 x 2
##   who      sugar_mean
##   <chr>         <dbl>
## 1 adults         10.9
## 2 children       46.6
```

These means look very different, though it would be better to look at a boxplot (coming up in a moment).

(c) Make side-by-side boxplots of the sugar contents of the two types of cereal. What do you see that is out of the ordinary?

Solution: The usual:

```
ggplot(cereals,aes(x=who,y=sugar))+geom_boxplot()
```



I see outliers: two high ones on the adults' cereals, and one high and one low on the children's cereals.

My thought above about the means being very different is definitely supported by the medians being very different on the boxplots. We should have no trouble declaring that the "typical" amounts of sugar in the adults' and children's cereals are different.

(d) Explain briefly why you would not trust a two-sample t-test with these data. (That is, say what the problem is, and why it's a problem.)

Solution: The problem is the outliers (which is rather a giveaway), but the reason it's a problem is that the two-sample t-test assumes (approximately) normal data, and a normal distribution doesn't have outliers.

Not only do you need to note the outliers, but you also need to say why the outliers cause a problem *in this case*. Anything less than that is not a complete answer.

(e) Run a suitable test to see whether the "typical" amount of sugar differs between adult's and children's cereals. Justify the test that you run. (You can use the version of your test that lives in a package, if that is easier for you.) What do you conclude, in the context of the data?

Solution: Having ruled out the two-sample t-test, we are left with Mood's median test. I didn't need you to build it yourself, so you can use package `smmr` to run it with:

```
library(smmr)
median_test(cereals,sugar,who)
```

```
## $table
##           above
## group      above below
##   adults       2    19
##   children    18     1
##
## $test
##        what        value
## 1 statistic 2.897243e+01
## 2        df 1.000000e+00
## 3   P-value 7.341573e-08
```

We conclude that there *is* a difference between the median amounts of sugar between the two groups of cereals, the P-value of 0.00000007 being extremely small.

Why did it come out so small? Because the amount of sugar was smaller than the overall median for almost all the adult cereals, and larger than the overall median for almost all the children's ones. That is, the children's cereals really do have more sugar.

Mood's median test doesn't come with a confidence interval (for the difference in population medians), because whether or not a certain difference in medians is rejected depends on what those medians actually are, and the idea of the duality of the test and CI doesn't carry over as we would like.

My daughter likes chocolate Cheerios, but she also likes Shredded Wheat and *Bran Flakes*. Go figure. (Her current favourite is Raisin Bran, even though she doesn't like raisins by themselves.)

Mood's median test is the test we should trust, but you might be curious about how the t-test stacks up here:

```
t.test(sugar~who,data=cereals)
```

```
##
##  Welch Two Sample t-test
##
## data:  sugar by who
## t = -11.002, df = 37.968, p-value = 2.278e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -42.28180 -29.13925
## sample estimates:
##   mean in group adults mean in group children
##               10.90000               46.61053
```

The P-value is *even smaller*, and we have the advantage of getting a confidence interval for the difference in means: from about 30 to about 40 units less sugar in the adult cereals. Whatever the units were.

3. Two new short courses have been proposed for helping students who suffer from severe math phobia. The courses are labelled A and B. Ten students were randomly allocated to one of these two courses, and each student's score on a math phobia test was recorded after they completed their course. The math phobia test produces whole-number scores between 0 and 10, with a higher score indicating a greater fear of mathematics. The data can be found in `http://www.utsc.utoronto.ca/~butler/c32/mathphobia.txt`. We start with R for this question.

   (a) Read in the data and check, however you like, that you have 10 observations, 5 from each course.

Solution: This doesn't need much comment:

```
my_url="http://www.utsc.utoronto.ca/~butler/c32/mathphobia.txt"
math=read_delim(my_url," ")
```

```
## Parsed with column specification:
## cols(
##  course = col_character(),
##  phobia = col_integer()
## )
```

```
math
```

```
## # A tibble: 10 x 2
##    course phobia
##    <chr>   <int>
##  1 a           8
##  2 a           7
##  3 a           7
##  4 a           6
##  5 a           6
##  6 b           9
##  7 b           8
##  8 b           7
##  9 b           2
## 10 b           1
```

This will do, counting the `a` and `b`. Or, to save yourself that trouble:

```
math %>% count(course)
```

```
## # A tibble: 2 x 2
##   course     n
##   <chr>  <int>
## 1 a          5
## 2 b          5
```

Five each. The story is to get the computer to do the grunt work for you, if you can make it do so. Other ways:

```
math %>% group_by(course) %>% summarize(count=n())
```

```
## # A tibble: 2 x 2
##   course count
##   <chr>  <int>
## 1 a          5
## 2 b          5
```

and this:

```
with(math,table(course))
```

```
## course
## a b
## 5 5
```

giving the same answer. Lots of ways.

Extra: there is an experimental design issue here. You might have noticed that each student did only *one* of the courses. Couldn't students do both, in a matched-pairs kind of way? Well, it's a bit like the kids learning to read in that if the first of the courses reduces a student's anxiety, the second course won't appear to do much good (even if it actually would have been helpful had the student done that one first). This is the same idea as the kids learning to read: once you've learned to read, you've learned to read, and learning to read a second way won't help much. The place where matched pairs scores is when you can "wipe out" the effect of one treatment before a subject gets the other one. We have an example of kids throwing baseballs and softballs that is like that: if you throw one kind of ball, that won't affect how far you can throw the other kind.

(b) Do a two-sample t-test to assess whether there is a difference in mean phobia scores after the students have taken the two courses. What do you conclude? (You have no *a priori*[1] reason to suppose that a particular one of the tests will produce a higher mean than the other, so do a two-sided test.)

Solution: A two-sided test is the default, so there is not much to do here:

```
t.test(phobia~course,data=math)
```

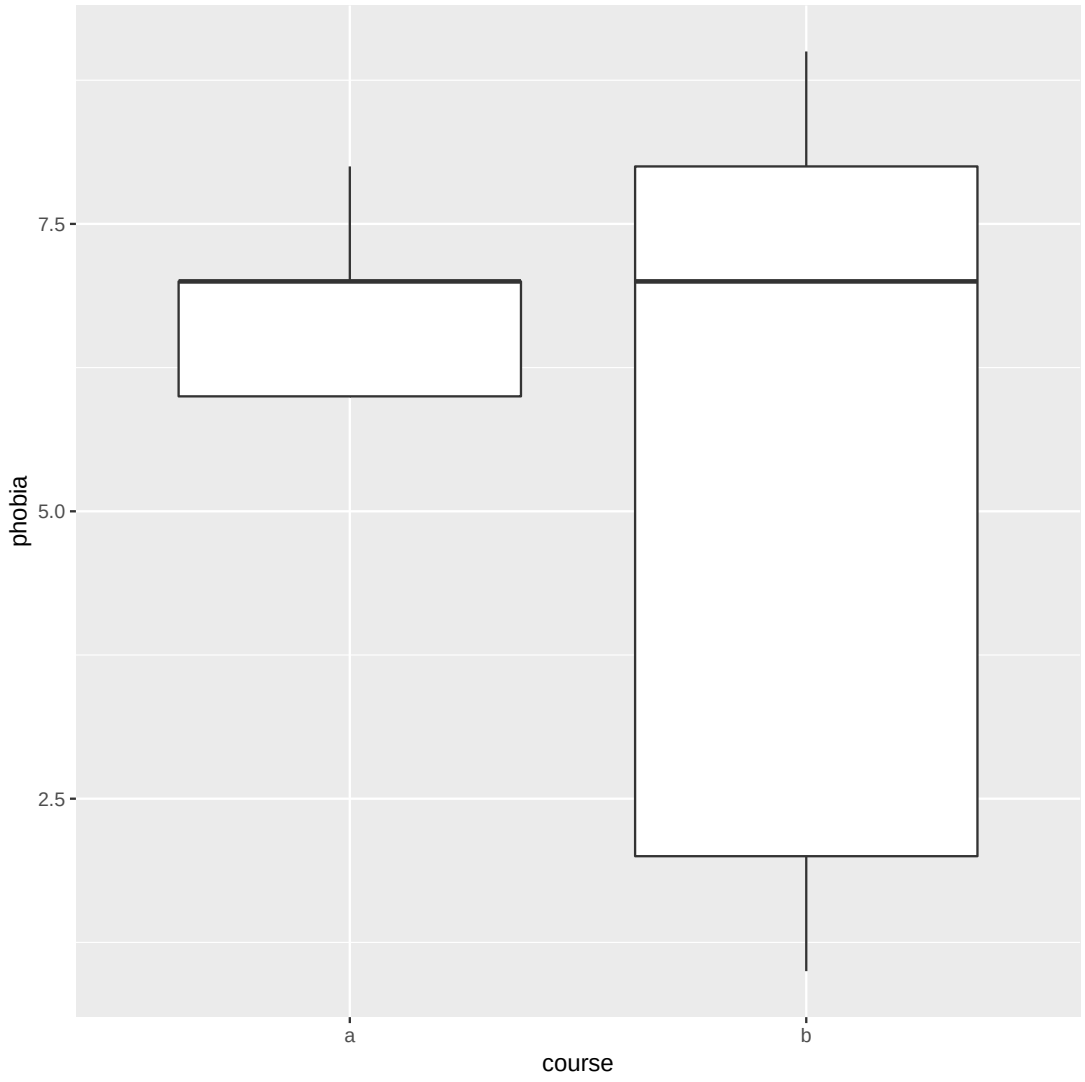
```
##
##  Welch Two Sample t-test
##
## data:  phobia by course
## t = 0.83666, df = 4.4199, p-value = 0.4456
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.076889  5.876889
## sample estimates:
## mean in group a mean in group b
##             6.8             5.4
```

The P-value of 0.4456 is nowhere near less than 0.05, so there is no evidence at all that the mean math phobia scores are different between the two courses.

(c) Draw boxplots of the math phobia scores for each group (one line of code). What is the most striking thing that you notice?

Solution:

```
ggplot(math,aes(x=course,y=phobia))+geom_boxplot()
```



Boxplot `a` is just weird. The bar across the middle is actually at the top, and it has no bottom. (Noting something sensible like this is enough.) Boxplot `b` is hugely spread out.[2]

By way of explanation: the course `a` scores have a number of values equal so that the 3rd quartile and the median are the name, and also that the first quartile and the minimum value are the same:

```
tmp=math %>% filter(course=="a")
tmp %>% count(phobia)
```

```
## # A tibble: 3 x 2
##   phobia     n
##    <int> <int>
## 1      6     2
## 2      7     2
## 3      8     1
```

```
summary(tmp)
```

```
##     course              phobia
##  Length:5           Min.   :6.0
##  Class :character   1st Qu.:6.0
##  Mode  :character   Median :7.0
##                     Mean   :6.8
##                     3rd Qu.:7.0
##                     Max.   :8.0
```

The phobia scores from course A are two 6's, two 7's and an 8. The median and third quartile are both 7, and the first quartile is the same as the lowest value, 6.

Technique note: I wanted to do two things with the phobia scores from course A: count up how many of each score, and show you what the five-number summary looks like. One pipe won't do this (the pipe "branches"), so I saved what I needed to use, before it branched, into a data frame `tmp` and then used `tmp` twice. Pipes are powerful, but not *all*-powerful.

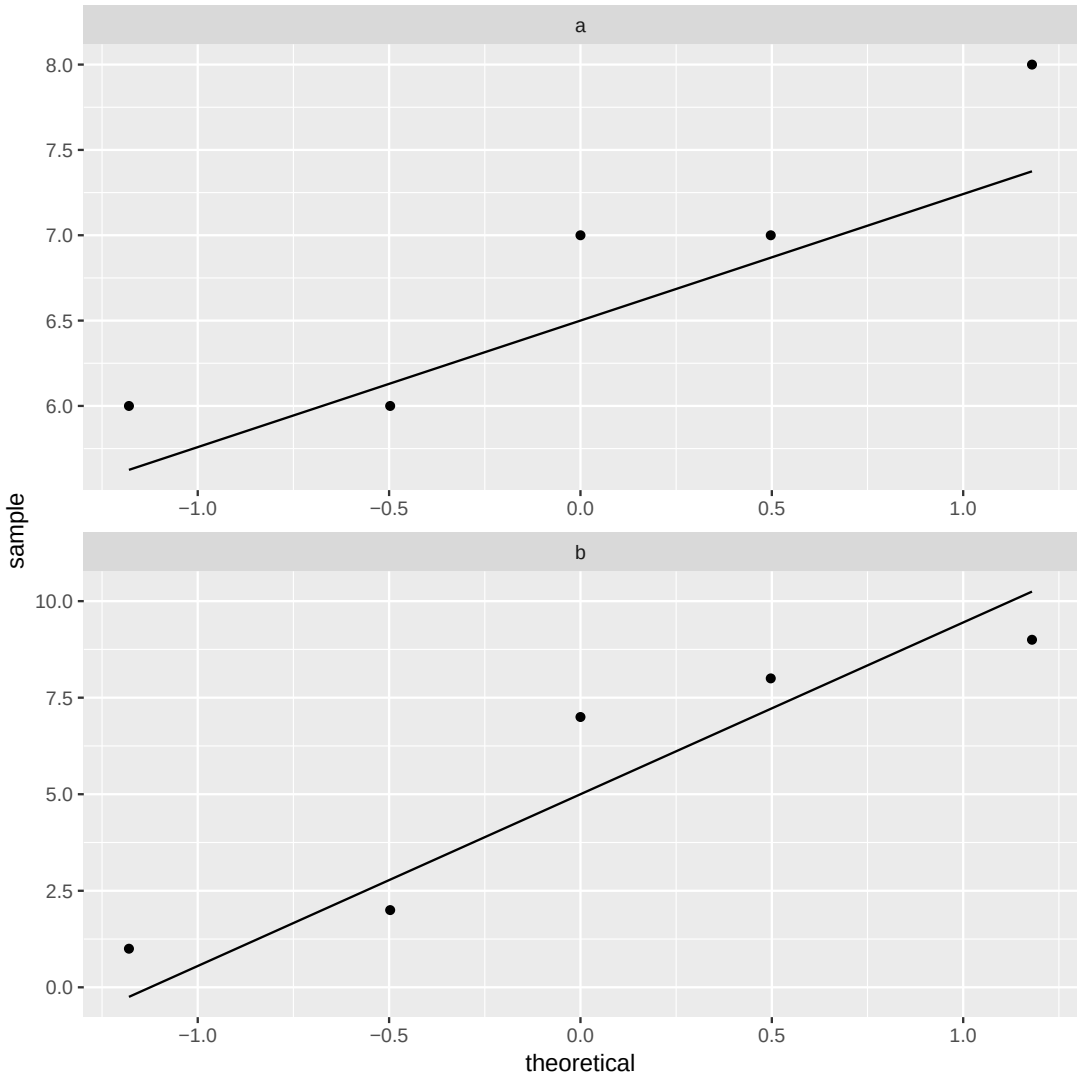
(d) Explain briefly why a t-test would not be good for these data. (There are two things that you need to say.)

Solution: The easiest way to structure this is to ask yourself first what the t-test needs, and second whether you have it.

The t-test assumes (approximately) normal data. The boxplot for group `a` doesn't even look symmetric, and the one for group `b` has an oddly asymmetric box. So I think the normality is in question here, and therefore another test would be better. (This is perhaps a bit glib of an answer, since there are only 5 values in each group, and so they can certainly look non-normal even if they actually are normal, but these values are all integers, so it is perhaps wise to be cautious.)

We have the machinery to assess the normality for these, in one shot:

```
ggplot(math,aes(sample=phobia))+
    stat_qq()+stat_qq_line()+
    facet_wrap(~course,ncol=1, scales="free")
```



I don't know what *you* make of those, but they both look pretty straight to me (and there are only five observations, so it's hard to judge). Course `b` maybe has a "hole" in it (three large values and two small ones). Maybe. I dunno. What I would *really* be worried about is outliers, and at least we don't have those.

I mentioned in class that the t-tests are robust to non-normality. I ought to have expanded on that a bit: what really makes the t-test still behave itself with non-normality is when you have *large* samples, that is, when the Central Limit Theorem has had a chance to take hold. (That's what drives the normality not really being necessary in most cases.) But, even with small samples, exact normality doesn't matter so much. Here, we have two tiny samples, and so we have to insist a bit more, but only a bit more, on a more-or-less normal shape in each group. (It's kind of a double jeopardy in that the situation where normality matters most, namely with small samples, is where it's the hardest to judge, because samples of size 5 even from a normal distribution can look very non-normal.)

But, the biggest threats to the t-test are big-time skewness and outliers, and we are not suffering too badly from those.

(e) Run a suitable test to compare the "typical" scores for the two courses. (You can use the version from a package rather than building your own.) What do you conclude?

Solution: This is an invite to use `smmr`:

```
library(smmr)
median_test(math,phobia,course)
```

```
## $table
##      above
## group above below
##     a     1     2
##     b     2     2
##
## $test
##        what     value
## 1 statistic 0.1944444
## 2        df 1.0000000
## 3   P-value 0.6592430
```

We are nowhere near rejecting equal medians; in fact, both courses are very close to 50–50 above and below the overall median.

If you look at the frequency table, you might be confused by something: there were 10 observations, but there are only $1+2+2+2 = 7$ in the table. This is because three of the observations were equal to the overall median, and had to be thrown away:

```
math %>% summarize(med=median(phobia))
```

```
## # A tibble: 1 x 1
##     med
##   <dbl>
## 1     7
```

```
math %>% count(phobia)
```

```
## # A tibble: 6 x 2
##   phobia     n
##    <int> <int>
## 1      1     1
## 2      2     1
## 3      6     2
## 4      7     3
## 5      8     2
## 6      9     1
```

The overall median was 7. Because the actual data were really discrete (the phobia scores could only be whole numbers), we risked losing a lot of our data when we did this test (and we didn't have much to begin with). The other thing to say is that with small sample sizes, the frequencies in the table have to be *very* lopsided for you to have a chance of rejecting the null. Something like this is what you'd need:

```
x=c(1,1,2,6,6,6,7,8,9,10)
g=c(1,1,1,1,1,2,2,2,2,2)
d=tibble(x,g)
median_test(d,x,g)
```

```
## $table
##      above
## group above below
##     1     0     3
##     2     4     0
##
## $test
##        what       value
## 1 statistic 7.000000000
## 2        df 1.000000000
## 3   P-value 0.008150972
```

I faked it up so that we had 10 observations, three of which were equal to the overall median. Of the rest, all the small ones were in group 1 and all the large ones were in group 2. This is lopsided enough to reject with, though, because of the small frequencies, there actually was a warning about "chi-squared approximation may be inaccurate".[3]

4. Do people understand medical instructions better at certain times of the day? In a study, students in a grade 12 class are randomly divided into two groups, A and B. All students see a video describing how to use an infant forehead thermometer. The students in Group A see the video at 8:30 am, while the students in Group B see the same video at 3:00 pm (on the same day). The next day, all the students are given a test on the material in the video (graded out of 100). The observed scores are in `http://www.utsc.utoronto.ca/~butler/c32/forehead.txt` (values separated by spaces).

   (a) Read the data into R and display the (first ten) values.

Solution: Separated by spaces, so `read_delim`:

```
my_url="http://www.utsc.utoronto.ca/~butler/c32/forehead.txt"
instr=read_delim(my_url," ")
```

```
## Parsed with column specification:
## cols(
##  group = col_character(),
##  score = col_integer()
## )
```

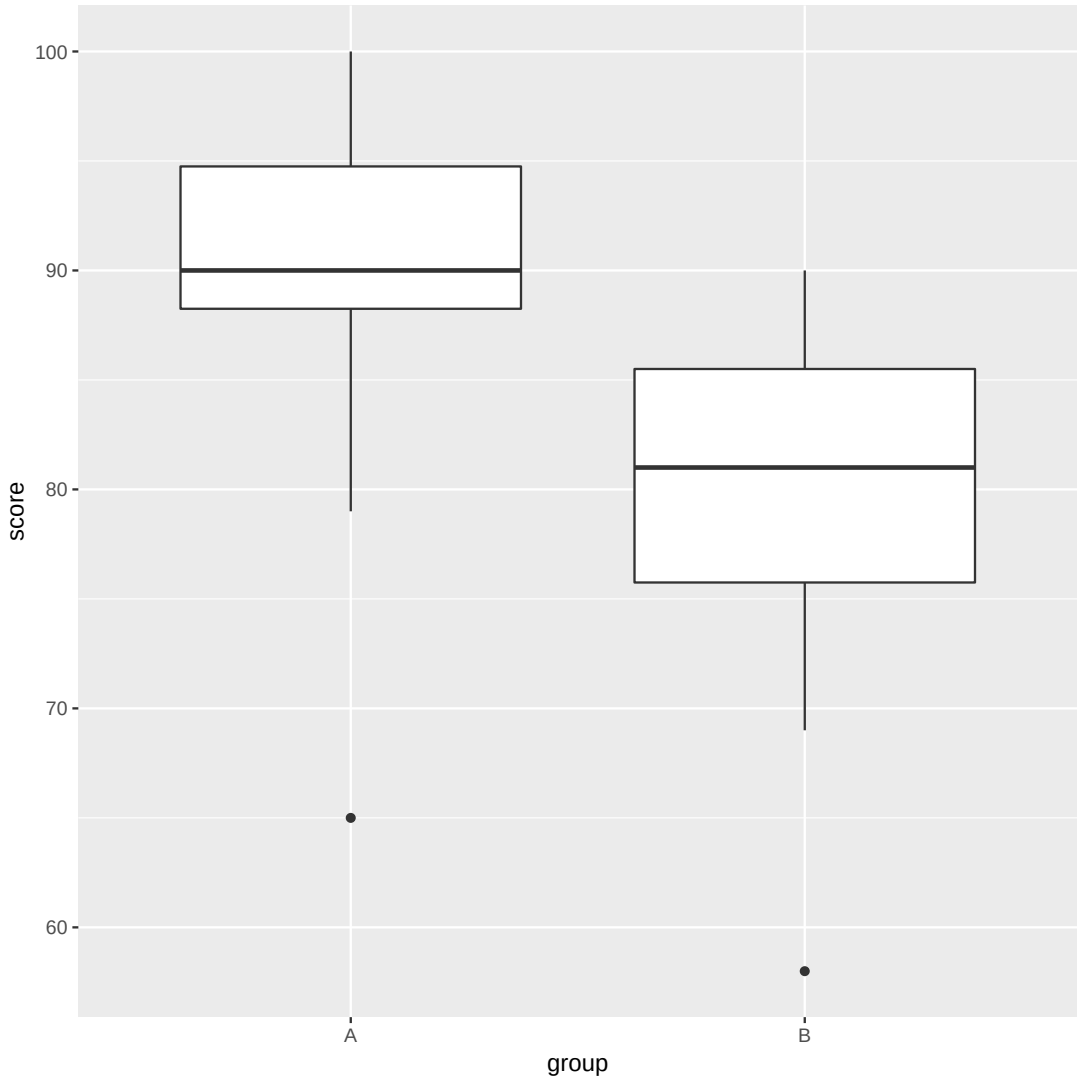
```
instr
```

```
## # A tibble: 18 x 2
##    group score
##    <chr> <int>
##  1 A        88
##  2 A        89
##  3 A        79
##  4 A       100
##  5 A        98
##  6 A        89
##  7 A        65
##  8 A        94
##  9 A        95
## 10 A        91
## 11 B        87
## 12 B        69
## 13 B        78
## 14 B        79
## 15 B        83
## 16 B        90
## 17 B        85
## 18 B        58
```

(b) Obtain a suitable plot that will enable you to assess the assumptions for a two-sample t-test.

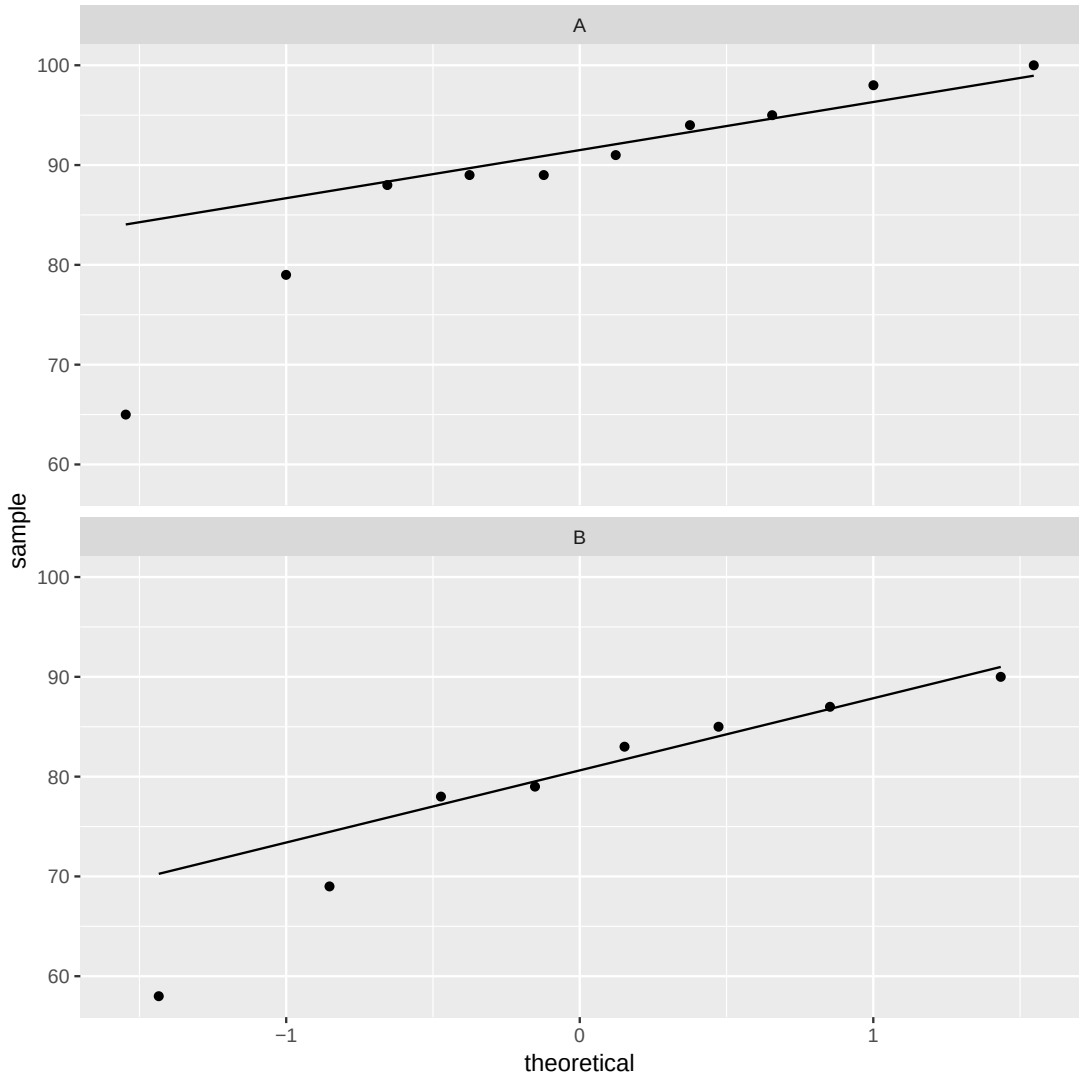
Solution: We need the values in each group to be approximately normally distributed. Side-by-side boxplots will do it:

```
ggplot(instr,aes(x=group,y=score))+geom_boxplot()
```



or, if you like, separate (facetted) normal quantile plots, which I would do this way:

```
ggplot(instr,aes(sample=score))+
    stat_qq()+stat_qq_line()+
    facet_wrap(~group,ncol=1)
```



(c) Why might you have doubts about using a two-sample t-test here?

Solution: We are looking for non-normality in at least one of the groups. Here, both groups have an outlier at the low end that would be expected to pull the mean downward. I don't think there is left-skewness here, since there is no particular evidence of the high-end values being bunched up: the problem in both cases with normality is at the low end.

One way or another, I'm expecting you to have noticed the outliers.

Extra: last year, when I first drew the normal quantile plots, there was no `stat_qq_line`, so you had to imagine where the line went if you did it this way. Without the line, these plots look somewhat curved, which would have pointed to left-skewness, but now we see that the lowest observation is too low, and maybe the second-lowest one as well, while the other observations are just fine.

(d) Run Mood's median test as in class (*without* using `smmr`). What do you conclude, in the context

of the data? What recommendation would you make about the time of day to see the video? (You might get a warning about "chisquared approximation being incorrect", which you can ignore here.)

Solution: The overall median first:

```
instr %>% summarize(med=median(score))
```

```
## # A tibble: 1 x 1
##     med
##   <dbl>
## 1  87.5
```

87.5, which is not equal to any of the data values (they are all integers). This will avoid any issues with values-equal-to-median later.

Then, create and save a table of the value by group and above/below median. You can count either above or below (it comes out equivalently either way):

```
tab=with(instr,table(group,score>87.5))
tab
```

```
##
## group FALSE TRUE
##     A     2    8
##     B     7    1
```

Then, chi-squared test for independence (the null) or association of some kind (the alternative). The `correct=F` is saying not to do Yates's correction, so that it would come out the same if you were doing it by hand ("observed minus expected, squared, divided by expected" and all that stuff).

```
chisq.test(tab,correct=F)
```

```
## Warning in chisq.test(tab, correct = F): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 8.1, df = 1, p-value = 0.004427
```

The P-value is 0.0044, which is (much) smaller than 0.05, and therefore you can reject independence and conclude association: that is, whether a student scores above or below the median depends on which group they are in, or, that the median scores are different for the two groups.

The warning is because the expected frequencies are on the small side (if you have done this kind of problem by hand, you might remember something about "expected frequencies less than 5". This is that.) Here, the P-value is so small that we can afford to have it be inaccurate by a bit and still not affect the conclusion, so I think we are safe.

As for which group is better, well, the easiest way is to go back to your boxplots and see that the median for group A (8:30 am) is substantially higher than for group B (3:00pm). But you can also see it from your frequency table, if you displayed it:

```
tab
```

```
##
## group FALSE TRUE
##     A     2    8
##     B     7    1
```

Most of the people in the 8:30 am group scored above the median, and most of the people in the 3:00 pm group scored below the median. So the scores at 8:30 am were better overall.

As I write this, it is just after 3:00 pm and I am about to make myself a pot of tea!

Extra: about that `correct=F` thing. There was a point of view for a long time that when you are dealing with a 2×2 table, you can get better P-values by, before squaring "observed minus expected", taking 0.5 away from the absolute value of the difference. This is called Yates's correction. It is in the same spirit as the "continuity correction" that you might have encountered in the normal approximation to the binomial, where in the binomial you have to have a whole number of successes, but the normal allows fractional values as well. In about the 1960s, the usefulness of Yates's correction was shot down, for general contingency tables. There is, however, one case where it *is* useful, and that is the case where the row totals and column totals are *fixed*.

What do I mean by that? Well, first let's look at a case where the totals are *not* all fixed. Consider a survey in which you want to see whether males and females agree or disagree on some burning issue of the day. You collect random samples of, say, 500 males and 500 females, and you count how many of them say Yes or No to your statement.[4] You might get results like this:

```
         Yes  No  Total
Males    197 303   500
Females  343 157   500
Total    540 460  1000
```

In this table, the row totals must be 500, because you asked this many males and this many females, and each one must have answered something. The column totals, however, are not fixed: you didn't know, ahead of time, that 540 people would answer "yes". That was just the way the data turned out, and if you did another survey with the same design, you'd probably get a different number of people saying "yes".

For another example, let's go back to Fisher (yes, *that* Fisher). A "lady" of his acquaintance claimed to be able, by drinking a cup of tea with milk and sugar in it, whether the milk or the sugar had been added first. Fisher, or, more likely, his housekeeper, prepared 8 cups of tea, 4 with milk first and 4 with sugar first. The lady knew that four of the cups had milk first, and her job was to say which four. The results might have been like this:

```
                      Actual
                  Milk first  sugar first  Total
Lady   Milk first          3            1      4
says   sugar first         1            3      4
       Total               4            4      8
```

This time, all of the row totals and all of the column totals must be 4, regardless of what the lady thinks. Even if she thinks 5 of the cups of tea actually had milk first, she is going to pick 4 of them to say that they have milk first, since she knows there are only 4. In this case, all of the row and column totals are fixed at 4, and the right analysis is called Fisher's Exact Test, based on the hypergeometric distribution. In a 2×2 table like this one, there is only one "degree of

freedom", since as soon as you specify one of the frequencies, say the number of cups where the lady said milk first and they actually were milk first, you can work out the others. But, leaving that aside, the usual chi-squared analysis is a perfectly good approximation, especially if the frequencies are large, and especially if you use Yates's correction.

It is clear that Fisher must have been English, since he was able to get a publication out of drinking tea.

How does that apply to Mood's median test? Well, let's remind ourselves of the table we had:

```
tab
```

```
##
## group FALSE TRUE
##     A     2    8
##     B     7    1
```

We know how many students were in each group: 10 in group A and 8 in B. So the row totals are fixed. What about the columns? These are whether each observation was above or below the overall median. There were 18 observations altogether, so there *must* be 9 above and 9 below.[5] So the column totals are fixed as well. All totals fixed, so we should be using Yates's correction. I didn't, because I wanted to keep things simple, but I should have done.

R's `chisq.test` by default *always* uses Yates's correction, and if you don't want it, you have to say `correct=F`. Which is why I have been doing so all through.

(e) Run Mood's median test on these data using my `smmr` package, and verify that you get the same answer.

Solution: Not much to it, since the data is already read in:

```
library(smmr)
median_test(instr, score, group)
```

```
## $table
##      above
## group above below
##     A     8     2
##     B     1     7
##
## $test
##        what       value
## 1 statistic 8.100000000
## 2        df 1.000000000
## 3   P-value 0.004426526
```

Identical, test statistic, degrees of freedom and P-value. The table of frequencies is also the same, just with columns rearranged. (In `smmr` I counted the number of values below the overall median, whereas in my build-it-yourself I counted the number of values above.)

5. Before a movie is shown in theatres, it receives a "rating" that says what kind of material it contains. `https://en.wikipedia.org/wiki/Motion_Picture_Association_of_America_film_rating_system` explains the categories, from G (suitable for children) to R (anyone under 17 must be accompanied by parent/guardian). In 2011, two students collected data on the length (in minutes) and the rating category, for 15 movies of each rating category, randomly chosen from all the movies released that

year. The data are at `http://www.utsc.utoronto.ca/~butler/c32/movie-lengths.csv`.

(a) Read the data into R, and display (some of) what you read in.

Solution: `read_csv`:

```
my_url="http://www.utsc.utoronto.ca/~butler/c32/movie-lengths.csv"
movies=read_csv(my_url)
```

```
## Parsed with column specification:
## cols(
##  length = col_integer(),
##  rating = col_character()
## )
```

```
movies
```

```
## # A tibble: 60 x 2
##    length rating
##     <int> <chr>
##  1     25 G
##  2     75 G
##  3     88 G
##  4     63 G
##  5     76 G
##  6     97 G
##  7     68 G
##  8     82 G
##  9     98 G
## 10     74 G
## # ... with 50 more rows
```

Something that looks like a length in minutes, and a rating.

(b) Count how many movies there are of each rating.

Solution:

```
movies %>% count(rating)
```

```
## # A tibble: 4 x 2
##   rating     n
##   <chr>  <int>
## 1 G         15
## 2 PG        15
## 3 PG-13     15
## 4 R         15
```

Fifteen of each rating. (It's common to have the same number of observations in each group, but not necessary for a one-way ANOVA.)

(c) Carry out an ANOVA and a Tukey analysis (if warranted).

Solution: ANOVA first:

```
length.1=aov(length~rating,data=movies)
summary(length.1)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## rating       3  14624    4875   11.72 4.59e-06 ***
## Residuals   56  23295     416
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This P-value is 0.00000459, which is way less than 0.05.

Having rejected the null (which said "all means equal"), we now need to do Tukey, thus:

```
TukeyHSD(length.1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = length ~ rating, data = movies)
##
## $rating
##                 diff        lwr       upr     p adj
## PG-G       26.333333   6.613562 46.053104 0.0044541
## PG-13-G    42.800000  23.080229 62.519771 0.0000023
## R-G        30.600000  10.880229 50.319771 0.0007379
## PG-13-PG   16.466667  -3.253104 36.186438 0.1327466
## R-PG        4.266667 -15.453104 23.986438 0.9397550
## R-PG-13   -12.200000 -31.919771  7.519771 0.3660019
```

Cast your eye down the `p adj` column and look for the ones that are significant, here the first three. These are all comparisons with the G ("general") movies, which are shorter on average than the others (which are not significantly different from each other).

If you like, you can make a table of means to verify that:

```
movies %>% group_by(rating) %>%
    summarize(mean=mean(length))
```

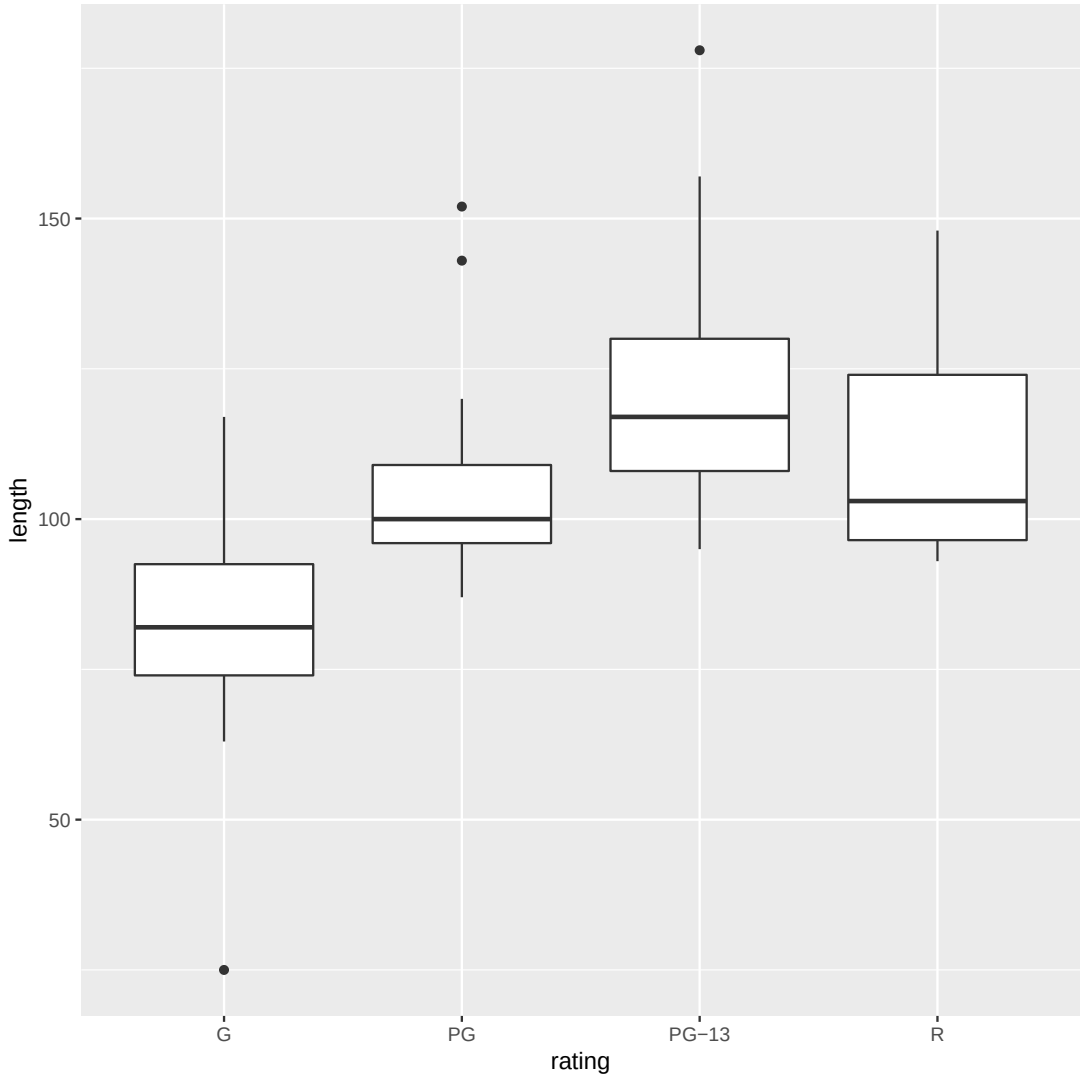
```
## # A tibble: 4 x 2
##   rating  mean
##   <chr>  <dbl>
## 1 G       80.6
## 2 PG     107.
## 3 PG-13  123.
## 4 R      111.
```

When we do this problem in SAS, you'll see the Tukey get handled a different way, one that you might find more appealing.

(d) Make a graph to assess whether this ANOVA is trustworthy. Discuss your graph and its implications briefly.

Solution: The obvious graph is a boxplot:

```
ggplot(movies, aes(x=rating, y=length))+geom_boxplot()
```



For ANOVA, we are looking for approximately normal distributions within each group and approximately equal spreads. Without the outliers, I would be more or less happy with that, but the G movies have a low outlier that would pull the mean down and the PG and PG-13 movies have outliers that would pull the mean up. So a comparison of means might make the differences look more significant than they should. Having said that, you could also say that the ANOVA is *very* significant, so even considering the effect of the outliers, the differences between G and the others are still likely to be significant.

Extra: the way to go if you don't trust the ANOVA is (as for the two-sample t) the Mood's median test. This applies to any number of groups, and works in the same way as before:

```
library(smmr)
median_test(movies,length,rating)
```

```
## $table
##        above
## group   above below
##   G         2    13
##   PG        7     7
##   PG-13    12     3
##   R         8     6
##
## $test
##        what        value
## 1 statistic 13.752380952
## 2        df  3.000000000
## 3   P-value  0.003262334
```

Still significant, though not quite as small a P-value as before (which echoes our thoughts about what the outliers might do to the means). If you look at the table above the test results, you see that the G movies are mostly shorter than the overall median, but now the PG-13 movies are mostly *longer*. So the picture is a little different.

Mood's median test does not naturally come with something like Tukey. What you can do is to do all the pairwise Mood's median tests, between each pair of groups, and then adjust to allow for your having done several tests at once. I thought this was generally useful enough that I put it into `smmr` under the name `pairwise_median_test`:

```
pairwise_median_test(movies,length,rating)
```

```
## # A tibble: 6 x 4
##   g1    g2      p_value adj_p_value
##   <chr> <chr>     <dbl>       <dbl>
## 1 G     PG    0.00799      0.0479
## 2 G     PG-13 0.0000590    0.000354
## 3 G     R     0.0106       0.0635
## 4 PG    PG-13 0.0106       0.0635
## 5 PG    R     0.715        4.29
## 6 PG-13 R     0.273        1.64
```

You can ignore those (adjusted) P-values rather stupidly bigger than 1. These are not significant.

There are two significant differences in median length: between G movies and the two flavours of PG movies. The G movies are significantly shorter (as you can tell from the boxplot), but the difference between G and R movies is no longer significant (a change from the regular ANOVA).

You may be puzzled by something in the boxplot: how is it that the G movies are significantly shorter than the PG movies, but not significantly shorter than the R movies, *when the difference in medians between G and R movies is bigger*? In Tukey, if the difference in means is bigger, the P-value is smaller.[6] The resolution to this puzzle, such as it is, is that Mood's median test is not directly comparing the medians of the groups (despite its name); it's counting values above and below a *joint* median, which might be a different story.

Hand the next two questions in:

6. The author of a popular book called *Mindless Eating* studies factors that make people eat without realizing how much they are eating. (Have you ever opened a bag of chips, and then, a little later, wondered where they all went?) One study was of 10 males and 10 females, who were allowed to take as many M&M candies as they wanted from a bowl during a study session. Do males or females tend to take more candies?

   The data in `https://www.utsc.utoronto.ca/~butler/c32/mm.txt` show each subject's gender and how many candies they took.

   (a) (3 marks) Read the data into R and count up how many male and female subjects there are.

   Solution:

   Separated by exactly one space, so `read_delim`:

```
my_url="https://www.utsc.utoronto.ca/~butler/c32/mm.txt"
mm=read_delim(my_url," ")
```

```
## Parsed with column specification:
## cols(
##  gender = col_character(),
##  candies = col_integer()
## )
```

   If you display this, you'll see only the males (who were listed first in the file). Hence I changed up the question to get you to count how many there were of each gender:

```
mm %>% count(gender)
```

```
## # A tibble: 2 x 2
##   gender     n
##   <chr>  <int>
## 1 Female    10
## 2 Male      10
```

   or if you prefer being more long-winded:

```
mm %>% group_by(gender) %>% summarize(how_many=n())
```

```
## # A tibble: 2 x 2
##   gender how_many
##   <chr>     <int>
## 1 Female       10
## 2 Male         10
```

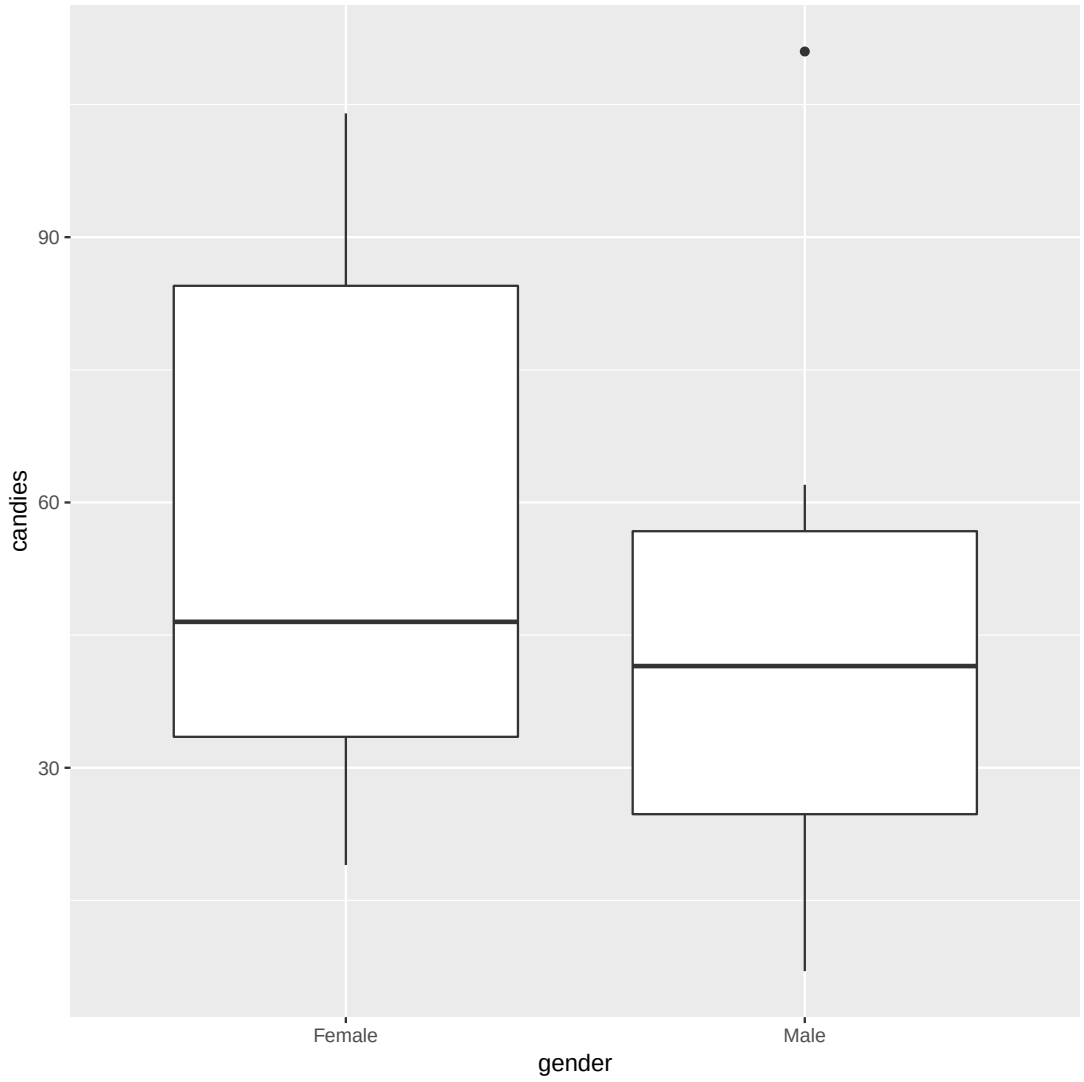
   Ten males and ten females (which you need to say), as in the preamble of the question.

   (b) (2 marks) Make a suitable plot to display the distributions of M&Ms taken by males and females.

   Solution:

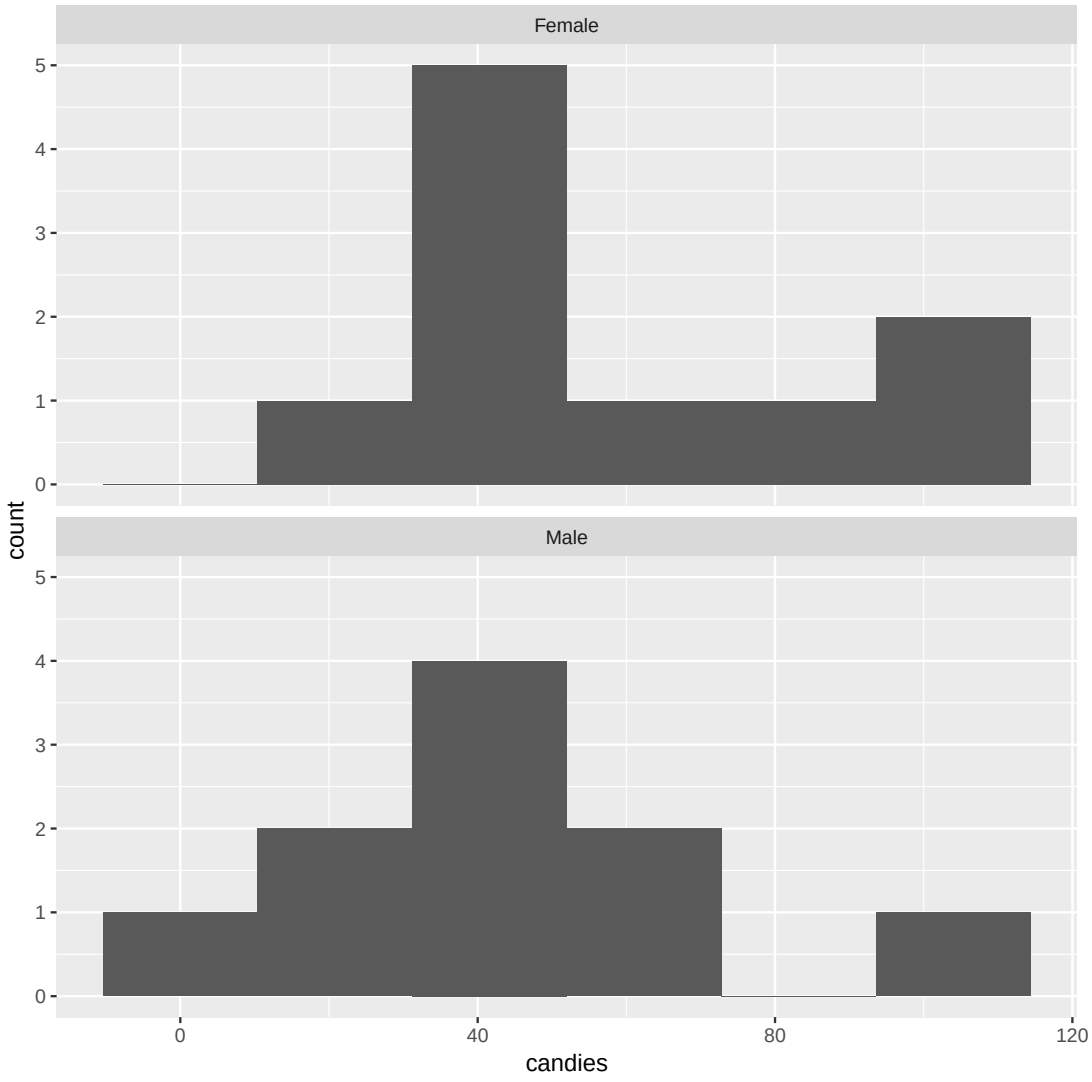
   One categorical variable (gender) and one quantitative (number of M&Ms), so a boxplot is my first port of call:

```
ggplot(mm, aes(x=gender, y=candies))+geom_boxplot()
```



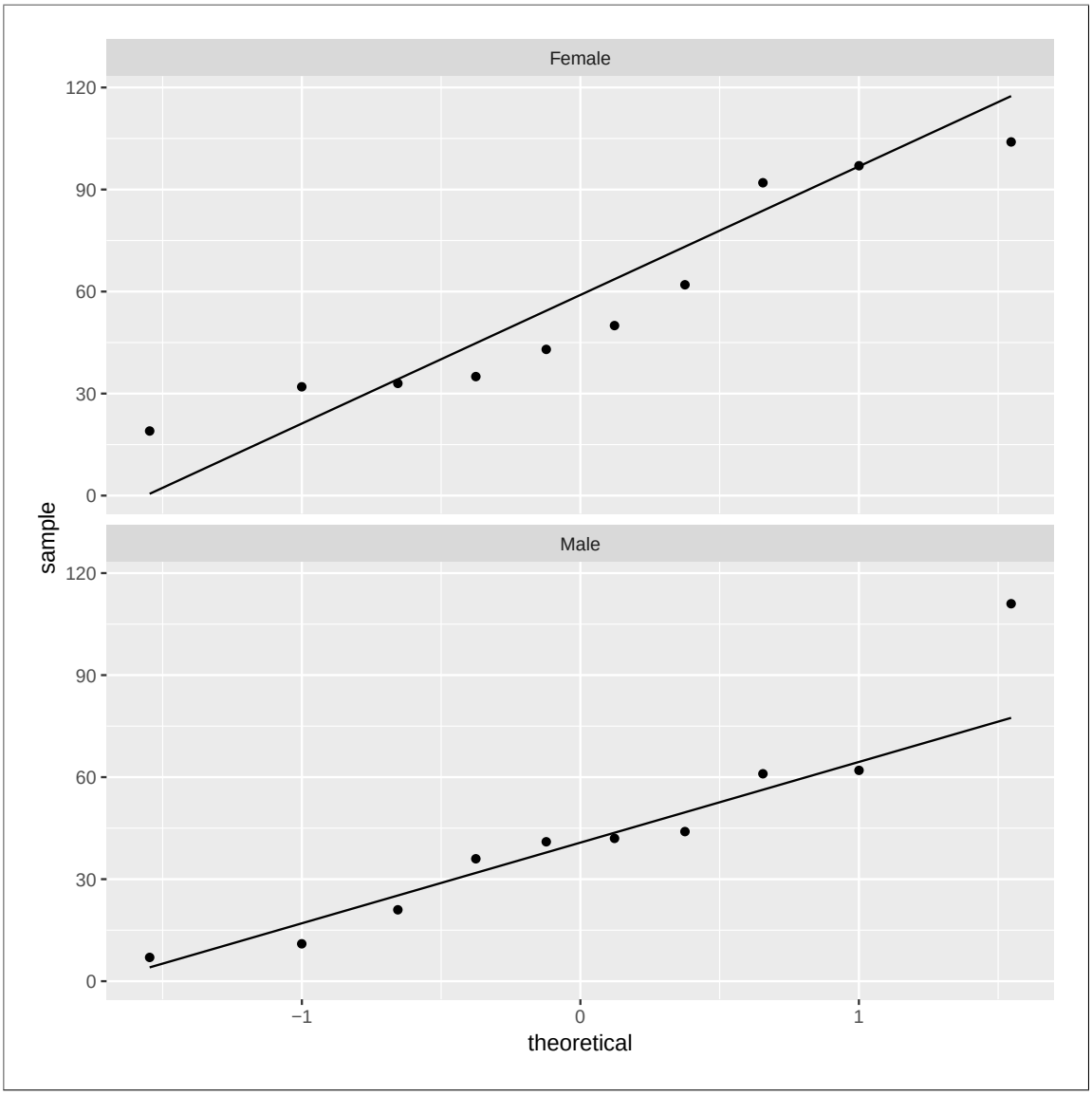
If you like, facetted above-and-below histograms will also work:

```
ggplot(mm, aes(x=candies))+geom_histogram(bins=6)+facet_wrap(~gender,ncol=1)
```



I like the boxplots better because they provide a clearer comparison of the two groups. You could even (almost) make a case for above-and-below facetted normal quantile plots, but that would be specifically be for assessing normality, as opposed to a general comparison of the distributions:

```
ggplot(mm, aes(sample=candies))+stat_qq()+stat_qq_line()+facet_wrap(~gender,ncol=1)
```



(c) (2 marks) What does your plot tell you about the typical numbers of candies taken by males and females, and how they compare with each other? Explain briefly.

Solution:

Looking at the boxplot, the median number of candies taken by females is slightly higher than by males: the median is 45 or 50 for females and about 40 for males. (You should say that the boxplot gives you the median, something about what each median is, at least approximately, and which one is higher.) If you prefer, you can say that the medians are very similar, and "they are both close to 45" or something like that. If you drew histograms, make some kind of sensible call about whether the centres of them are. If you draw normal quantile plots, you won't be able to answer this one, which will be a hint that some other kind of plot would be better.

(d) (2 marks) Why would a two-sample t-test not be best for comparing the typical number of candies taken by people of each gender?

Solution: A t-test would require both distributions to be approximately normal, given that there are only ten observations in each. The obvious thing is the high outlier on the male boxplot; there was one male who took a lot of candies. The female distribution looks fairly symmetric. If you drew histograms, neither distribution looks especially normal (but that might be the small samples). The normal quantile plots are designed to answer this kind of question; the female one looks acceptably normal (given the small sample size), and most of the male one does too, except for the high value that is too high.

Some kind of sensible comment given your graph. I would also accept drawing a normal quantile plot here and commenting on it.

(e) (3 marks) Run Mood's median test on these data, using the `smmr` package. What do you conclude, in the context of the data? (Note that the test result as given is two-sided.)

Solution: I already loaded the `smmr` package. You may need to do that first, but I don't:

```
median_test(mm,candies,gender)
```

```
## $table
##         above
## group    above below
##   Female     6     4
##   Male       4     6
##
## $test
##        what     value
## 1 statistic 0.8000000
## 2        df 1.0000000
## 3   P-value 0.3710934
```

There is no evidence of a difference in median numbers of candies taken by males and females. (This is not a great surprise to me, looking at the graphs, or indeed, looking at the very near 50–50 split in the table of aboves and belows.) The test is two-sided, so it tells us whether there is *any* difference. This makes sense, given that we had no reason ahead of time to expect a particular one of the genders to take more candies.

If you really wanted a one-sided test, you would do the usual two-part thing: ask whether you are on the correct side, and, if you are, halve the two-sided P-value. This test is set up this way because it's based on the chi-squared test, which is looking for *any* association between group and being above/below the overall median. If you have more than two groups (your categorical variable has more than two levels), the only sensible kind of alternative is a multi-sided one, since you are testing only that "not all the groups have the same median". SAS (later) does distinguish the two-group Mood median test from the one with more than two groups; when you have two groups, it will also give you a one-sided P-value.

7. Earlier in the course, we looked at some data on travel coffee mugs. 32 mugs were used, 8 each of four different brands. Each mug was filled with hot coffee and closed. 30 minutes later, each mug was opened and the temperature of the coffee in the mug measured. The experimenter recorded the temperature decrease (in degrees Fahrenheit) for each mug. Thus, a small temperature decrease is good (coffee in such a mug will remain warm for longer).

   The data are in `https://www.utsc.utoronto.ca/~butler/c32/coffee.txt`.

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

Solution: The usual `read_delim`:

```
my_url="https://www.utsc.utoronto.ca/~butler/c32/coffee.txt"
coffeecups=read_delim(my_url, " ")
```

```
## Parsed with column specification:
## cols(
##   cup = col_character(),
##   tempdiff = col_double()
## )
```

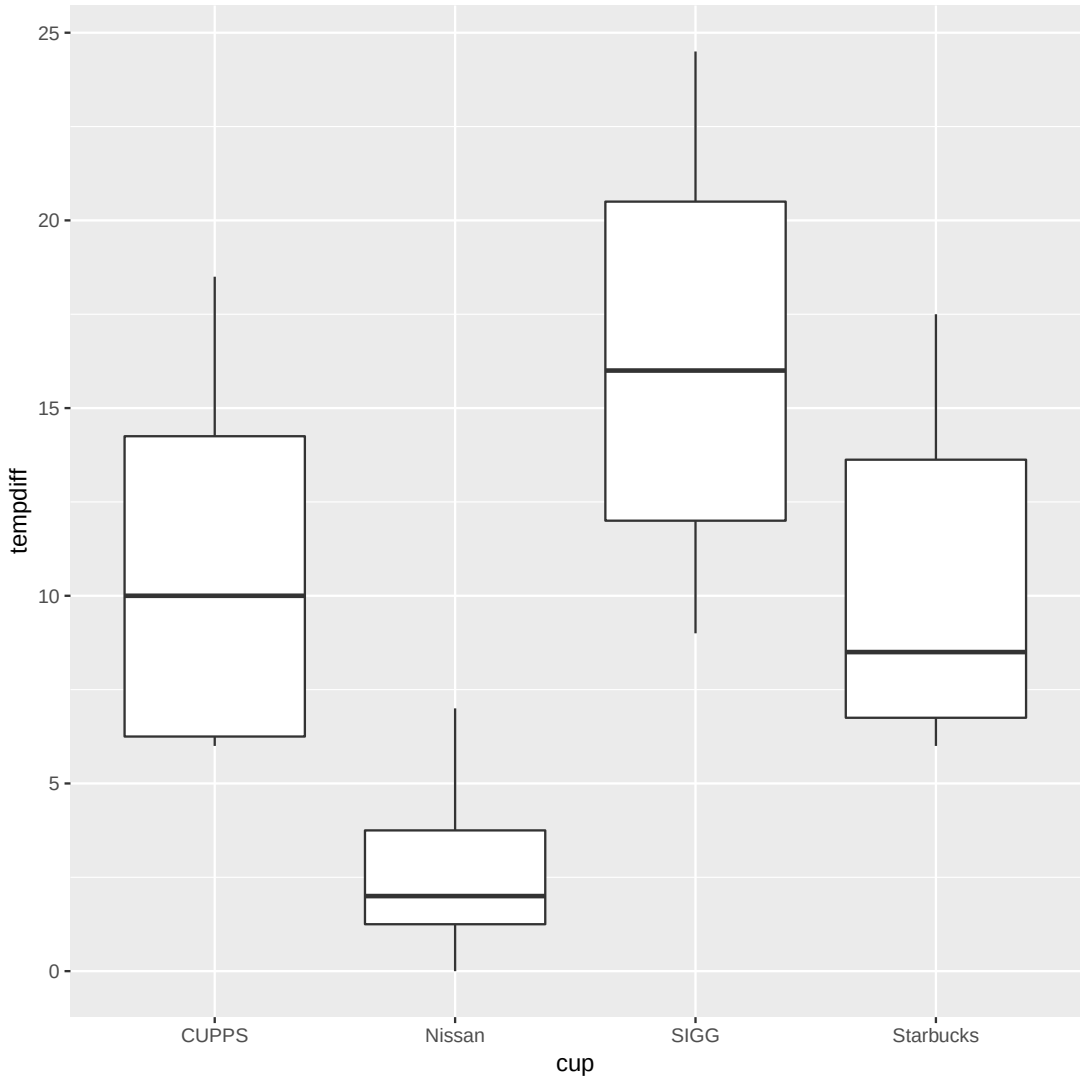
```
coffeecups
```

```
## # A tibble: 32 x 2
##    cup       tempdiff
##    <chr>        <dbl>
##  1 Starbucks     13
##  2 Starbucks      7
##  3 Starbucks      7
##  4 Starbucks     17.5
##  5 Starbucks     10
##  6 Starbucks     15.5
##  7 Starbucks      6
##  8 Starbucks      6
##  9 SIGG          12
## 10 SIGG          16
## # ... with 22 more rows
```

(b) (2 marks) Make a suitable boxplot of the data.

Solution: Thus:

```
ggplot(coffeecups, aes(x=cup, y=tempdiff))+geom_boxplot()
```



You're probably thinking that the ANOVA that is coming up is pretty shaky (small spread on the Nissan cups, some right-skewness in the distributions). We'll come back to that later.

(c) (3 marks) Run an analysis of variance. What do you conclude, in the context of the data?

Solution: `tempdiff` is the response and `cup` the categorical explanatory variable:

```
cups.1=aov(tempdiff~cup, data=coffeecups)
summary(cups.1)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## cup          3  809.8  269.95    12.2 2.77e-05 ***
## Residuals   28  619.7   22.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis is that the mean temperature difference in all four brands of cup is the same. This is clearly rejected, so the means of the temperature differences for the four brands of mug are not all the same. *This is where you stop.*

You might (reasonably) have said that you should not run an ANOVA here because the assumptions are not met. If that's what you thought, you need to run an appropriate analysis here (either Welch or Mood's median test) and then follow it up with the appropriate multiple comparisons if warranted (that is, Games-Howell or pairwise median tests respectively). That is to say, if you go this way, you need one of these here:

```
oneway.test(tempdiff~cup, data=coffeecups)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  tempdiff and cup
## F = 16.97, num df = 3.000, denom df = 14.369, p-value =
## 5.407e-05
```

or

```
median_test(coffeecups, tempdiff, cup)
```

```
## $table
##            above
## group       above below
##   CUPPS         4     3
##   Nissan        0     8
##   SIGG          8     1
##   Starbucks     4     4
##
## $test
##        what        value
## 1 statistic 13.587301587
## 2        df  3.000000000
## 3   P-value  0.003524286
```

In both of those cases, the P-values are very small, so the conclusion you get is the same: the "typical" (mean or median) temperature differences are not the same for the four cups.

(d) (2 marks) If reasonable, run Tukey's method on these data. If it is not reasonable to do so, explain briefly why this is.

Solution: The ANOVA is significant, so not all the brands of mug have the same mean temperature difference, and we want to find out which ones differ from which. (This is motivation for doing the Tukey; you don't need to say this.)

```
TukeyHSD(cups.1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = tempdiff ~ cup, data = coffeecups)
##
## $cup
##                         diff         lwr         upr     p adj
## Nissan-CUPPS     -8.0357143 -14.6832138 -1.38821474 0.0132542
## SIGG-CUPPS        5.7698413  -0.7030218 12.24270429 0.0936942
## Starbucks-CUPPS  -0.5357143  -7.1832138  6.11178526 0.9961503
## SIGG-Nissan      13.8055556   7.5644014 20.04670975 0.0000094
## Starbucks-Nissan  7.5000000   1.0779053 13.92209469 0.0173795
## Starbucks-SIGG   -6.3055556 -12.5467098 -0.06440136 0.0469861
```

Extra: the time you would *not* run Tukey is if the ANOVA is not significant. That would tell you that all the groups have the same mean, and there is nothing more to find out.

You would also be justified in not running Tukey if you thought that `aov` was not the right thing to do. My intention was that you would run `aov` and Tukey and *then* critique it (which was why I arranged the question as I did). I *guess* you deserve full credit (for this part) if you refuse to run Tukey for this reason, though it would really be better, for the purposes of analyzing the data, if you do the appropriate follow-up for the test you thought was best in the previous part. That is, if you did Mood's median test, then do:

```
pairwise_median_test(coffeecups, tempdiff, cup)
```

```
## # A tibble: 6 x 4
##   g1     g2          p_value adj_p_value
##   <chr>  <chr>         <dbl>       <dbl>
## 1 CUPPS  Nissan    0.00341      0.0205
## 2 CUPPS  SIGG      0.131        0.783
## 3 CUPPS  Starbucks 0.797        4.78
## 4 Nissan SIGG      0.0000633    0.000380
## 5 Nissan Starbucks 0.00200      0.0120
## 6 SIGG   Starbucks 0.317        1.90
```

or if you did Welch's ANOVA, then this:

```
library(PMCMRplus)
gamesHowellTest(tempdiff~factor(cup),data=coffeecups)
```

```
##
## Pairwise comparisons using Games-Howell test
```

```
## data:  tempdiff by factor(cup)
```

```
##           CUPPS   Nissan  SIGG
## Nissan    0.02509 -       -
## SIGG      0.20730 0.00018 -
## Starbucks 0.99663 0.00852 0.09593
```

```
##
## P value adjustment method:  none
```

```
## alternative hypothesis:  two.sided
```

The Games-Howell test requires `cup` to be a genuine `factor`. It won't work otherwise.

(e) (2 marks) Are any of the mug brands significantly different from *all* of the other mugs? Which one(s)?

Solution: This is a slightly unusual Tukey, in that I would normally ask you to say something about which differences are significant. Here, most but not all of the differences are significant, so I asked it differently.

One way to answer this is to look for the *non*-significant differences. Any mug that features in one of these is *not* significantly different from all of the others.

The non-significant differences are SIGG-Cupps and Starbucks-Cupps. That includes all the mugs except for Nissan, and you can verify that all three of the comparisons involving Nissan are significant. If you look back at the boxplot, this is saying that the Nissan mugs are significantly *better* at keeping coffee warm than all of the others. (So that would be the mug to recommend.)

If you did pairwise median tests or Games-Howell, the conclusion will be the same, though for these, *only* the differences involving Nissan are significant, so it's a bit easier.

You might have run the `aov` (because I asked you to) and then *not* run the Tukey (because you were worried about the assumptions). That's all right so far, if a little inconsistent, but then you don't have a good answer to this part. In that case, you need to say "because I didn't run Tukey I don't have an answer to this part". (That would be full marks.) The hint is that I used the words "significantly different" in the question, which implies, in the statistical world, that you need to look at a test and a P-value. Trying to do something else (like comparing medians from the boxplot) won't work here because there is no P-value involved, and so you're not getting at "*significantly* different". (This is not what I originally intended, but that's not the point; it's a matter of whether you have answered the question, or stated that you cannot.)

(f) (2 marks) The person who originally analyzed these data said "I am somewhat cautious about these conclusions". Why do you think she said that? Explain briefly.

Solution: This is an invitation to check the assumptions of ANOVA, namely normal distributions with equal spreads within each group. You can assess this by looking back at the boxplots. The Nissan boxplot looks to have a lot less spread than the others, and you might

also have some quibbles about normality, eg. of the CUPPS and Starbucks mugs, which both look right-skewed.

Find one reason (or two reasons) why you don't believe the assumptions for these data, and say something about the analyst's caution coming from her not being completely happy with the assumptions.

(This was the first place I intended you to assess the assumptions, but if you have done that before, and proceeded in an appropriate way given what you concluded, I am happy with that.)

Notes

[1]That is, before looking at the data. This is Latin. It's also the place that the Bayesian "prior distribution" comes from. The "posterior distribution" comes from the Latin *a posteriori*, which means "afterwards", that is, after you have looked at the data.

[2]The two groups have very different spreads, but that is not a problem as long as we remember to do the Welch-Satterthwaite test that does not assume equal spreads. This is the default in R, so we are good, at least with that.

[3]There *was*, in the `chisq.test` inside `median_test`, but in `smmr` I didn't pass that warning back to the outside world.

[4]To simplify things, we'll assume that everyone gave a Yes or a No answer, though you could add a column like "No answer" if you wanted to make it more realistic.

[5]Except in the case of the previous problem, where there were multiple observations equal to the overall median. Which we ignore for the moment.

[6]Actually, this doesn't always work if the sample sizes in each group are different. If you're comparing two small groups, it takes a *very large* difference in means to get a small P-value. But in this case the sample sizes are all the same.