University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAC32 (K. Butler), Midterm Exam
October 19, 2019

Aids allowed (printed or handwritten): My lecture overheads (slides); Any notes that you have taken in this course; Your marked assignments; My assignment solution; Non-programmable, non-communicating calculator.

This exam has 51 numbered pages of questions. Check to see that you have all the pages. There is an additional empty page that you can use if you need more space for any answers.

In addition, you should have an additional booklet of output to refer to during the exam. Contact an invigilator if you do not have this.

Answer each question in the space provided (under the question).

The maximum marks available for each part of each question are shown next to the question part.

You may assume throughout this exam that the code shown in Figure 1 of the booklet of code and output has already been run.

The University of Toronto's Code of Behaviour on Academic Matters applies to all University of Toronto Scarborough students. The Code prohibits all forms of academic dishonesty including, but not limited to, cheating, plagiarism, and the use of unauthorized aids. Students violating the Code may be subject to penalties up to and including suspension or expulsion from the University.

Question 1 (11 marks)

When you drive along the highway, you might notice a lot of vegetation (grass, plants etc.) along the side of the highway, in the middle of on-ramps, etc. This vegetation has to be attended to: in particular, if it gets more than 30 cm tall, it is dangerous because car drivers cannot see what is on the other side. Highway authorities need to devise a system to keep the vegetation less than 30 cm tall without spending a lot of money on maintenance. In Texas, an experiment was run. The vegetation next to a typical highway was divided into 36 sections. The equipment used to mow the vegetation has an adjustable blade which can be adjusted to a height of 5, 10, and 20 centimetres (0.05, 0.10, and 0.20 metres). In addition, the vegetation can be mowed 1, 2, or 3 times a year. Each section of the highway was randomly assigned one of these blade heights and mowing frequencies (in such a way that each combination was used on four sections of the highway). One year after the vegetation was last mowed, the maintenance crew came back and measured the maximum height of vegetation on that section of the highway.

The data file is shown in Figure 2 (in the Booklet of Code and Output), and this data file is also stored in `mowing.txt` in your current project in R Studio Cloud (or, if you prefer, in the folder of your current project in R Studio on your computer).

(a) (3 marks) Using something from the `tidyverse`, give R code to read this data into a data frame called `mowing`.

My answer:

Take a look at the data file. The data values are not separated by a single space, but are aligned in columns (like the data on three different drugs in the file-reading section of the notes). This is therefore `read_table`:

```
mowing <- read_table("mowing.txt")
```

```
## Parsed with column specification:
## cols(
##  height = col_character(),
##  frequency = col_character(),
##  vegetation = col_double()
## )
```

Using an = instead of the left-arrow (for assignment to a variable, anywhere) also works, and is also therefore good (anywhere the left-arrow is also good).

If you have learned R somewhere other than from me, you might have learned this:

```
mowing2 <- read.table("mowing.txt", header=T)
mowing2
```

```
##    height  frequency vegetation
## 1   0.05m 1-per-year       17.3
## 2   0.05m 1-per-year       19.3
## 3   0.05m 1-per-year       15.0
## 4   0.05m 1-per-year       16.7
## 5   0.10m 1-per-year       16.0
## 6   0.10m 1-per-year       15.6
## 7   0.10m 1-per-year       16.9
## 8   0.10m 1-per-year       15.0
## 9   0.20m 1-per-year       16.7
## 10  0.20m 1-per-year       17.9
```

Question 1 continues...

```
## 11  0.20m 1-per-year       15.9
## 12  0.20m 1-per-year       13.7
## 13  0.05m 2-per-year       22.4
## 14  0.05m 2-per-year       20.8
## 15  0.05m 2-per-year       24.5
## 16  0.05m 2-per-year       21.7
## 17  0.10m 2-per-year       23.9
## 18  0.10m 2-per-year       23.6
## 19  0.10m 2-per-year       21.7
## 20  0.10m 2-per-year       23.8
## 21  0.20m 2-per-year       24.7
## 22  0.20m 2-per-year       26.3
## 23  0.20m 2-per-year       27.2
## 24  0.20m 2-per-year       26.4
## 25  0.05m 3-per-year       18.6
## 26  0.05m 3-per-year       17.9
## 27  0.05m 3-per-year       16.1
## 28  0.05m 3-per-year       19.4
## 29  0.10m 3-per-year       22.2
## 30  0.10m 3-per-year       25.6
## 31  0.10m 3-per-year       21.8
## 32  0.10m 3-per-year       23.6
## 33  0.20m 3-per-year       27.0
## 34  0.20m 3-per-year       25.3
## 35  0.20m 3-per-year       23.8
## 36  0.20m 3-per-year       28.0
```

This, as you see, works, but *it is from base R, not the Tidyverse*, so it is only one mark.

You might be tempted to think that the values are separated by tabs. This is not what happened here. You can tell because tab-separated stuff tends to have columns aligned on the *left*:

```
first   second  third
a       x       10
bb      y       9
ccc     zz      22
```

Take a look at the athletes data, which really was separated by tabs. In here, the columns are aligned because all the text was about the same length, but the values appear to be left-justified in their columns.

If the things in different columns are of different widths, the columns might not be aligned at all:

```
first    second third
short    1      22
a lot longer    4      23
```

The 4 in the last row is actually a value for `second`, not `third`; it only ended up where it did because the text `a lot longer` spilled into the second column.

The usual place that tab-separated data comes from is copying and pasting from a spreadsheet.

This is not something I would recommend doing (you can always save the spreadsheet as `.csv` and read that in). If you want to try it: open up Excel or another spreadsheet program, and make a sheet like my last one just above, with the text in the first column and two more columns of numbers after that. Copy and paste the values into Notepad or some other text editor, and see what they look like. (You could also try creating a New Text Document in R Studio and paste into that.)

Grading: three marks is a bit generous for this kind of thing, but it *is* the first part of the first question:

- Full marks for using `read_table` correctly;
- two marks for using it but making an error, for example failing to save the result of the `read_table` or calling the data frame something else (read the question!);
- one mark for using `read_tsv` or `read.table` correctly, or for making some kind of effort at getting `read_delim` to work (eg. using multiple spaces for the delimiter, which won't actually work but is a plausible idea).
- Nothing for anything else; for example, if you use `read.table` and forget `header=T`, don't expect to get anything here.

Extra: you might have caught on to the "right-assignment" idea of a pipeline; something funky like this also works, in that spirit:

```
my_url="mowing.txt"
my_url %>% read_table() -> mowing3
```

```
## Parsed with column specification:
## cols(
##  height = col_character(),
##  frequency = col_character(),
##  vegetation = col_double()
## )
```

and to demonstrate:

```
mowing3
```

```
## # A tibble: 36 x 3
##    height frequency  vegetation
##    <chr>  <chr>           <dbl>
##  1 0.05m  1-per-year       17.3
##  2 0.05m  1-per-year       19.3
##  3 0.05m  1-per-year       15
##  4 0.05m  1-per-year       16.7
##  5 0.10m  1-per-year       16
##  6 0.10m  1-per-year       15.6
##  7 0.10m  1-per-year       16.9
##  8 0.10m  1-per-year       15
##  9 0.20m  1-per-year       16.7
## 10 0.20m  1-per-year       17.9
## # ... with 26 more rows
```

Of course, if you want to define the file name into a variable first and then do something rather more conventional with it, that's fine too.

(b) (1 mark) What R code would display (at least some of) the values in your data frame?

My answer:

This is only one point, so it has to be something simple: just the name of the data frame:

```
mowing
```

```
## # A tibble: 36 x 3
##    height frequency  vegetation
##    <chr>  <chr>           <dbl>
##  1 0.05m  1-per-year       17.3
##  2 0.05m  1-per-year       19.3
##  3 0.05m  1-per-year       15
##  4 0.05m  1-per-year       16.7
##  5 0.10m  1-per-year       16
##  6 0.10m  1-per-year       15.6
##  7 0.10m  1-per-year       16.9
##  8 0.10m  1-per-year       15
##  9 0.20m  1-per-year       16.7
## 10 0.20m  1-per-year       17.9
## # ... with 26 more rows
```

Anything equivalent that works is also good, say:

```
glimpse(mowing)
```

```
## Observations: 36
## Variables: 3
## $ height     <chr> "0.05m", "0.05m", "0.05m", "0.05m", "0.10m", "0.10m...
## $ frequency  <chr> "1-per-year", "1-per-year", "1-per-year", "1-per-ye...
## $ vegetation <dbl> 17.3, 19.3, 15.0, 16.7, 16.0, 15.6, 16.9, 15.0, 16....
```

or something like `View(mowing)`, which will work in R Studio but not here. It's still good as an answer to this question, though.

This is a good place to remind you not to overthink my questions! (The usual thing to do after reading in a data frame is to take a look at it, which is all we are doing here.)

(c) (2 marks) The column `height` is really a number, but ought to be treated as a categorical variable. Why is that? Explain briefly.

My answer: In a designed experiment like this one, the "treatments" are generally treated as categorical (and something like an analysis of variance is run). The reason is that we want to compare the effect of each treatment (on the response variable) with each other one. In this case, we want to compare the amount of vegetation at each `height`, and (looking ahead) we might want to use something like boxplots to do it with, which would make sense if `height` is categorical.

Another way to go is that there are only three possible values; the heights in between don't

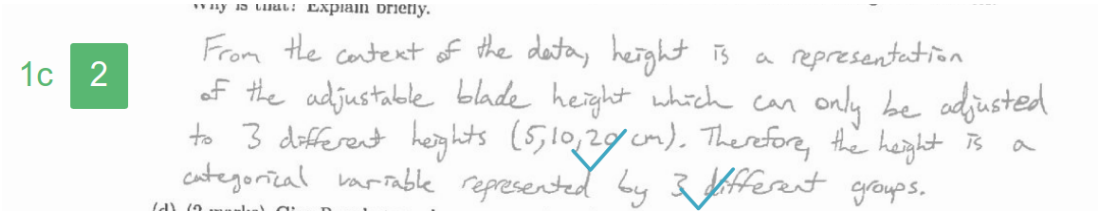
make any sense (in the context of this question anyway).

There are different things you could say, but I think the most obvious is that you want to do a comparison of vegetation between the different heights, so treating the height as categorical is the way to do this. I am writing this before people write the exam, so I will say that I am planning to be fairly relaxed about grading this one; if you say something sensible I am likely to be good with it.

There is one point for saying something about how the values get read in as text (and will thus be treated as categorical. This is not full marks, though, because it doesn't get at why *we* should be treating it as categorical (even had it been read in as a number).

Yes, there will be judgement on the grader's part here.

An actual answer:

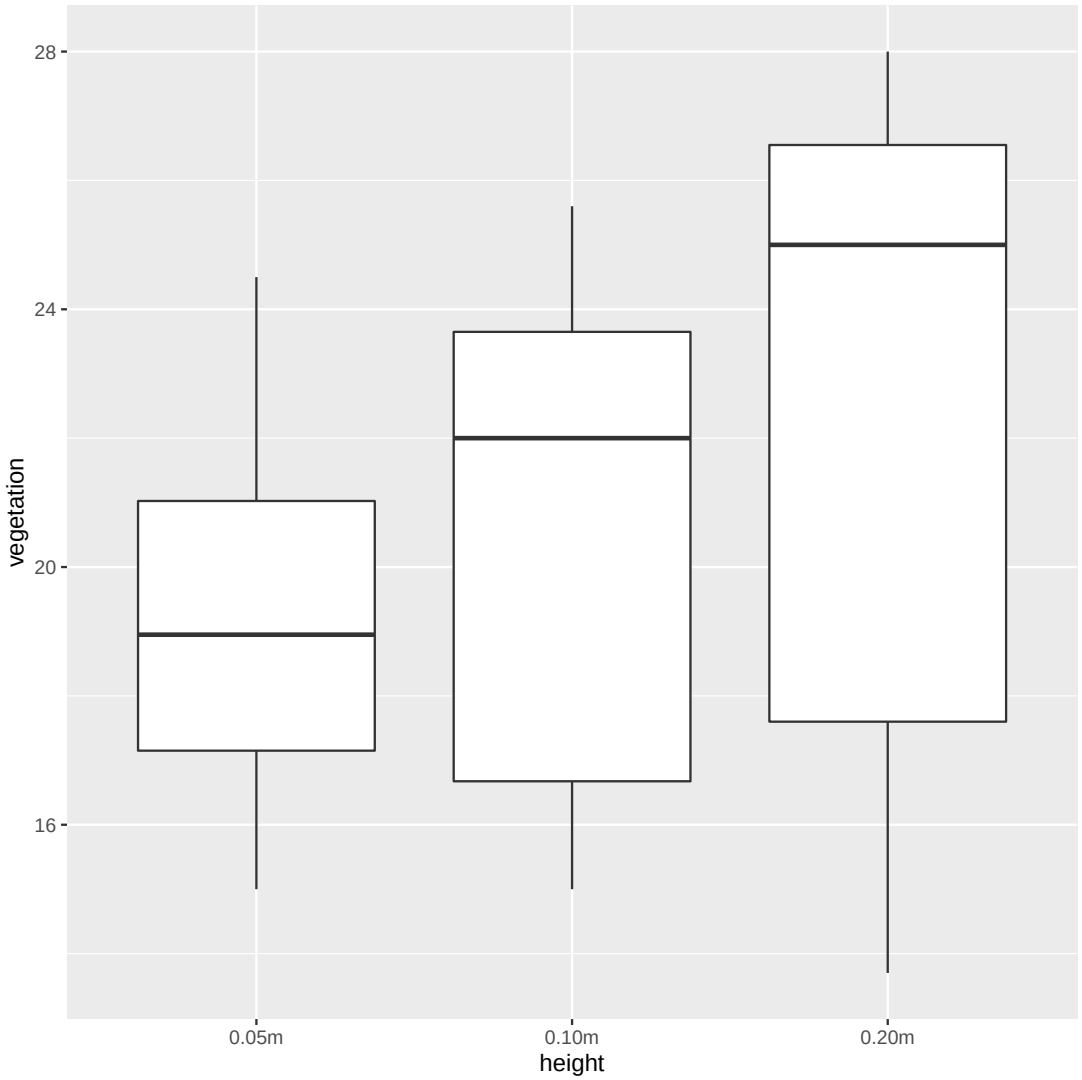


(d) (2 marks) Give R code to make an appropriate plot of `height` and `vegetation`, ignoring (for this part) `frequency`.

My answer:

I think I have rather given the game away with this one: `height` is (from the previous part) categorical, and `vegetation` is quantitative, so a boxplot is the way to go:

```
ggplot(mowing, aes(x=height, y=vegetation)) + geom_boxplot()
```

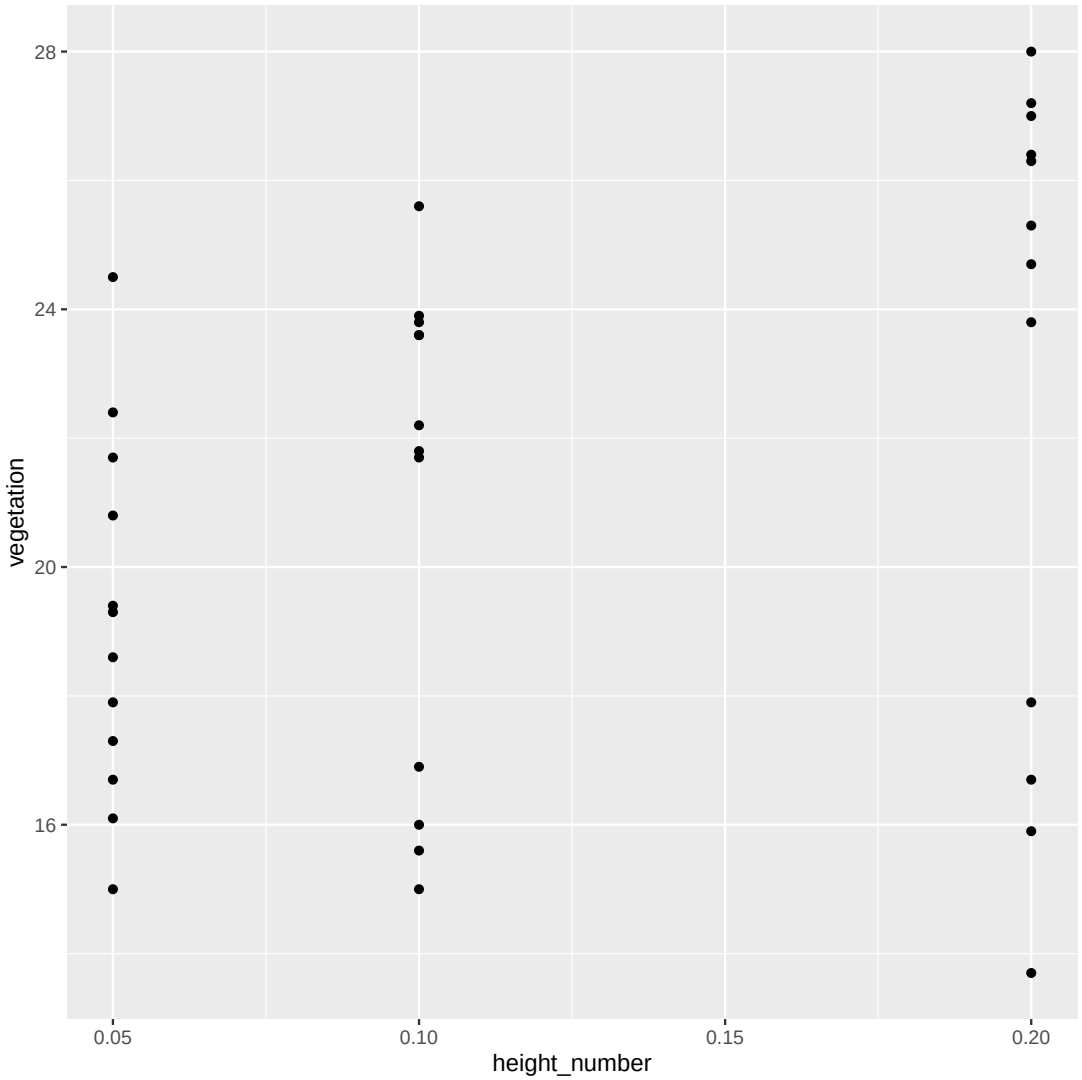


This part was meant to be as easy as that.

The boxplots themselves are not very informative, but that's OK because you won't be seeing the results in the exam.

Extra: if we had treated `height` as quantitative, a scatterplot would have been the thing. This one looks rather odd. We have to do a bit of work first to get the numbers out of `height`:

```
mowing %>% mutate(height_number = parse_number(height)) %>%
    ggplot(aes(x=height_number, y=vegetation)) + geom_point()
```



You might know this kind of thing as a "dotplot". The reason this one looks rather odd is that there were only a few different values of `height` (only three different heights in the entire data set). If the height could have been *anything* between 5cm and 30cm (randomly chosen, say), so that there were a whole bunch of different heights, we would have been looking for a (maybe linear) relationship between the now-quantitative height and the amount of vegetation, and then a scatterplot would definitely have been the thing.

I tried to give you a clue here by arranging the height and frequency values so that they would be read in as text rather than numbers. There is (for 2019 students) an echo of the first assignment here, in that the kind of *values* you read in from the file may or may not correspond to the kind of *variables* you want to use. There, the problem one was the dates (do you treat them

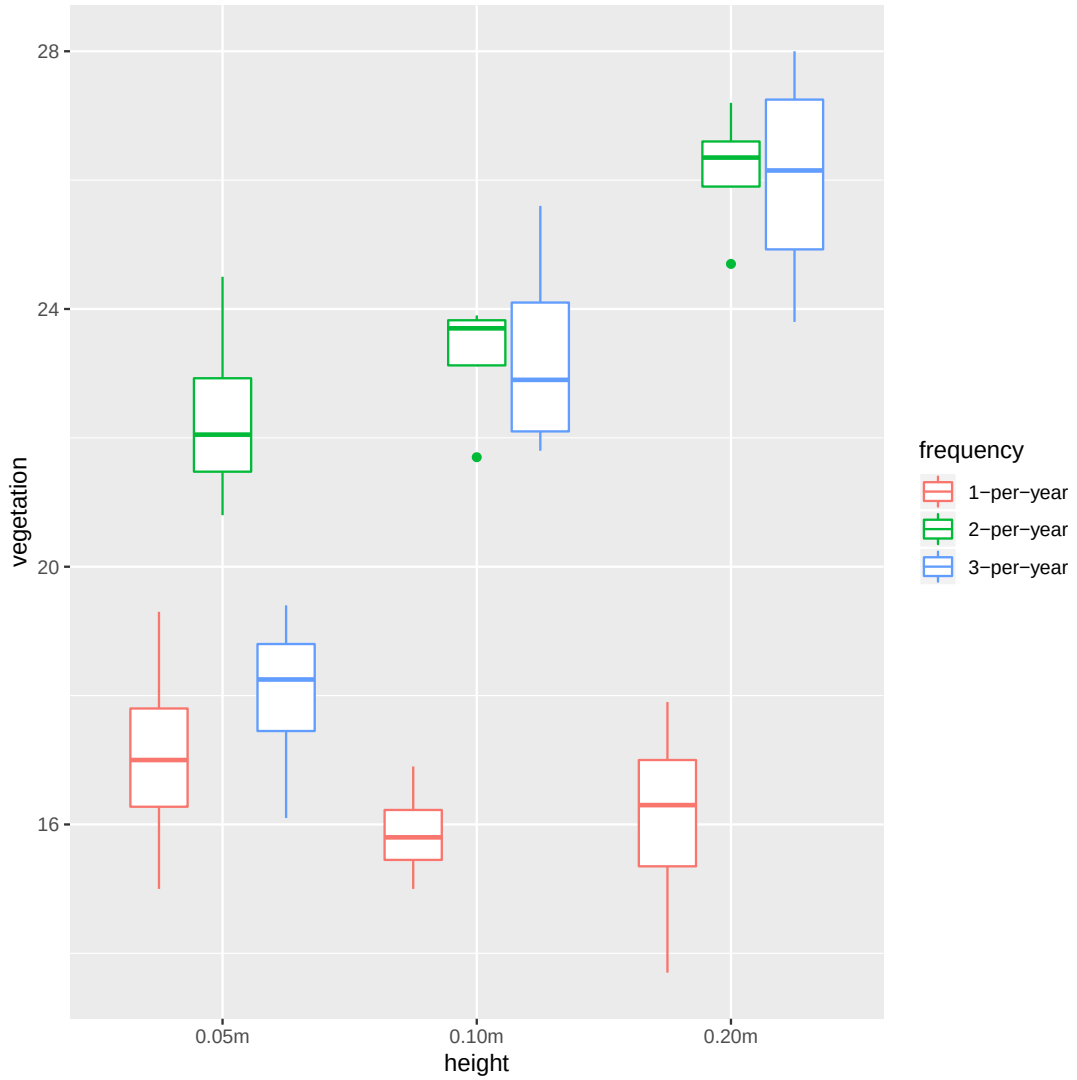
as quantitative or categorical?); here, the problem is `height` and also `frequency`, which were actually numbers but ought to be treated as categorical so that we can compare the vegetation for the different ones.

(e) (3 marks) We now want to make a suitable graph that includes all three variables. Describe the kind of graph that you would draw, and give R code to produce it.

My answer: Look back at the table of graphs in my notes. We have one quantitative variable and two categorical variables, so a grouped boxplot is the thing. One point. (Or facetted boxplots. See later.)

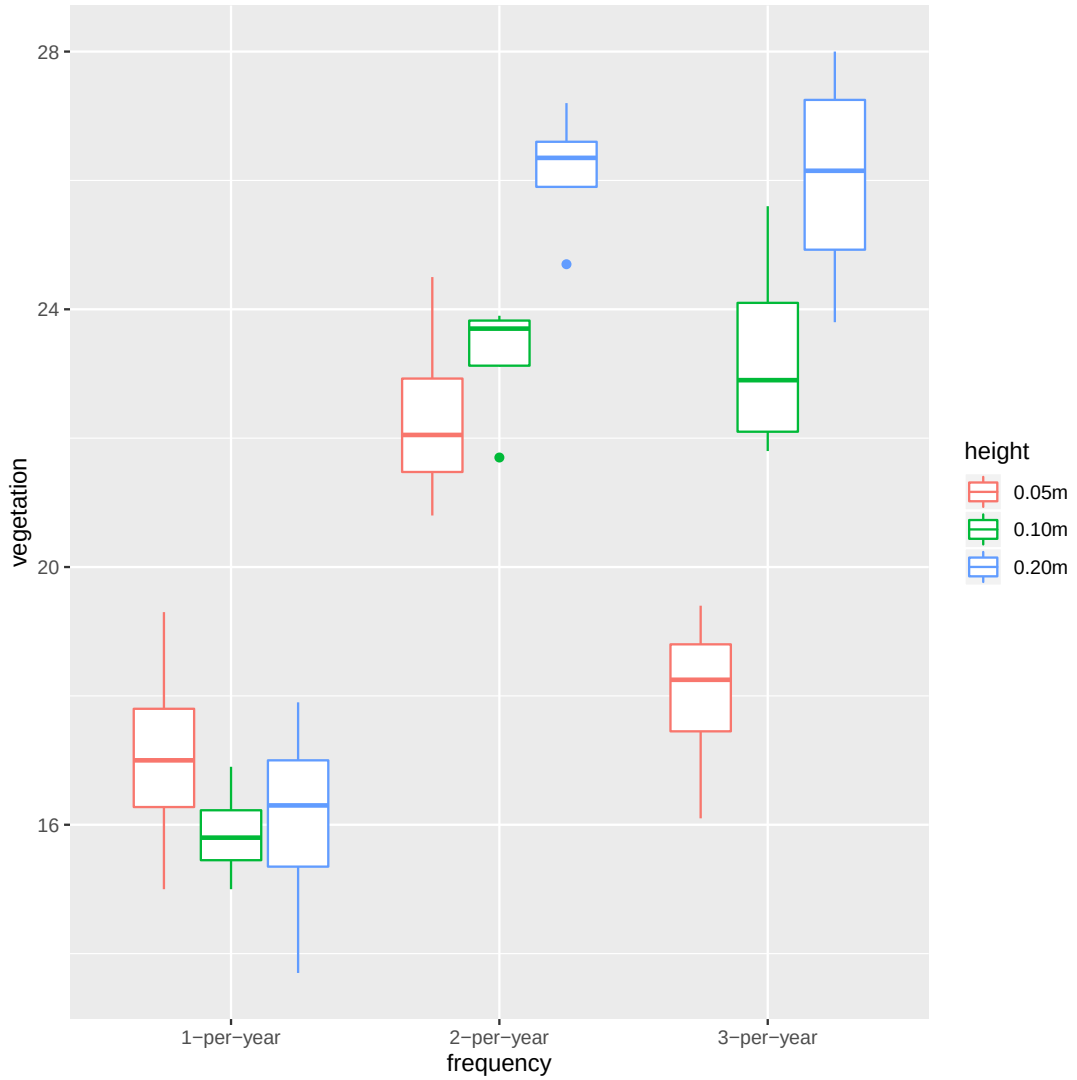
This kind of code will produce one:

```
ggplot(mowing, aes(x=height, y=vegetation, colour=frequency)) + geom_boxplot()
```



`x` and `colour` can be the other way around:

```
ggplot(mowing, aes(colour=height, y=vegetation, x=frequency)) + geom_boxplot()
```

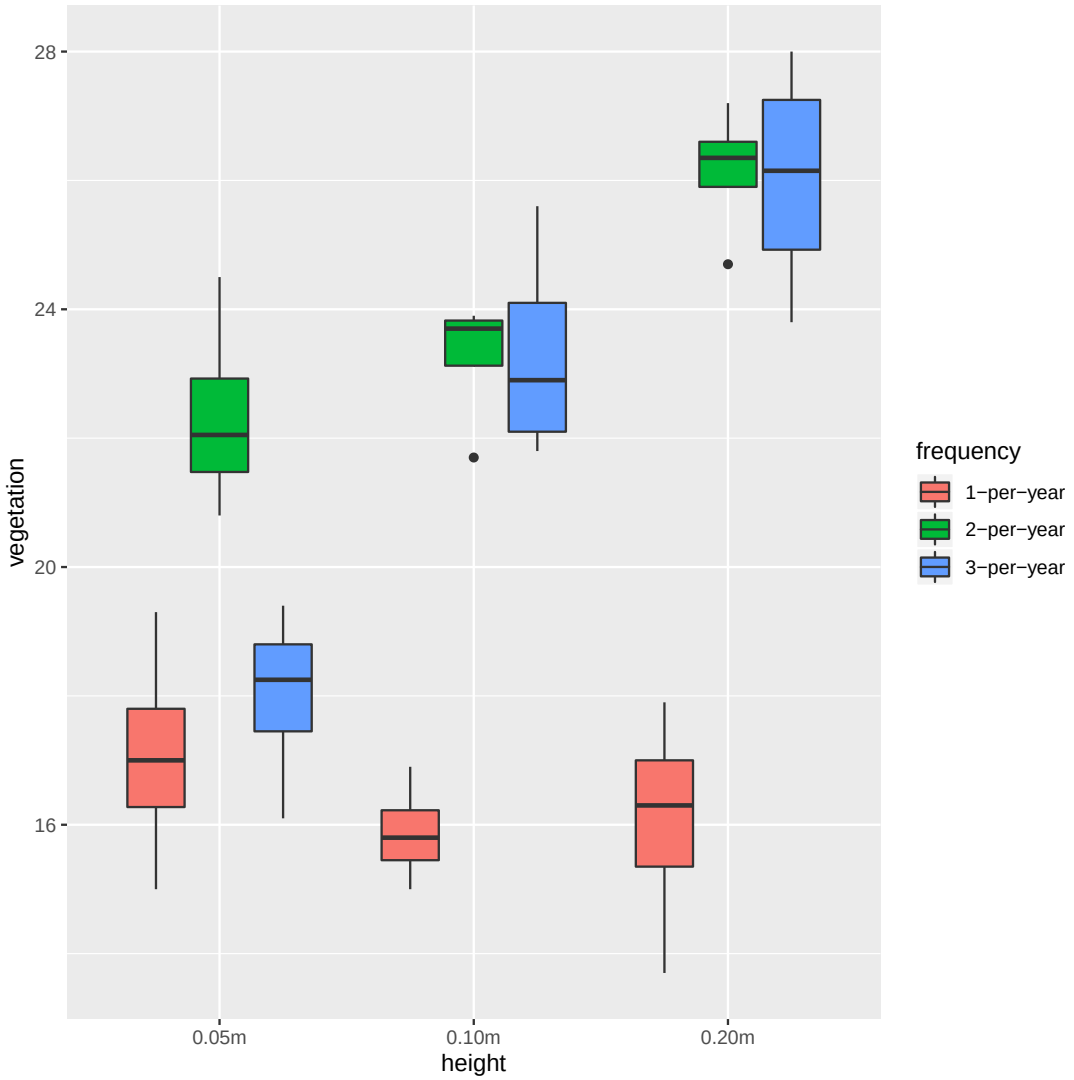


The look is different, but the plot is equally good. The order of the inputs to the `aes` is immaterial, so as long as you have an `x`, a `y` and a `colour` somewhere, it's good.

The usual choice of which categorical variable to call `x` and which to call `colour` is to have the categorical variable with more categories be the `x`. Here, though, both categorical variables have three categories, so there is no reason to prefer one choice over the other. (I was kind of glad this one came out this way, because it would have been rather picky to deduct something for *not* having the variable with more categories on the x-axis.)

You can also replace the `colour` with `fill` to get another look, but an equally good one:

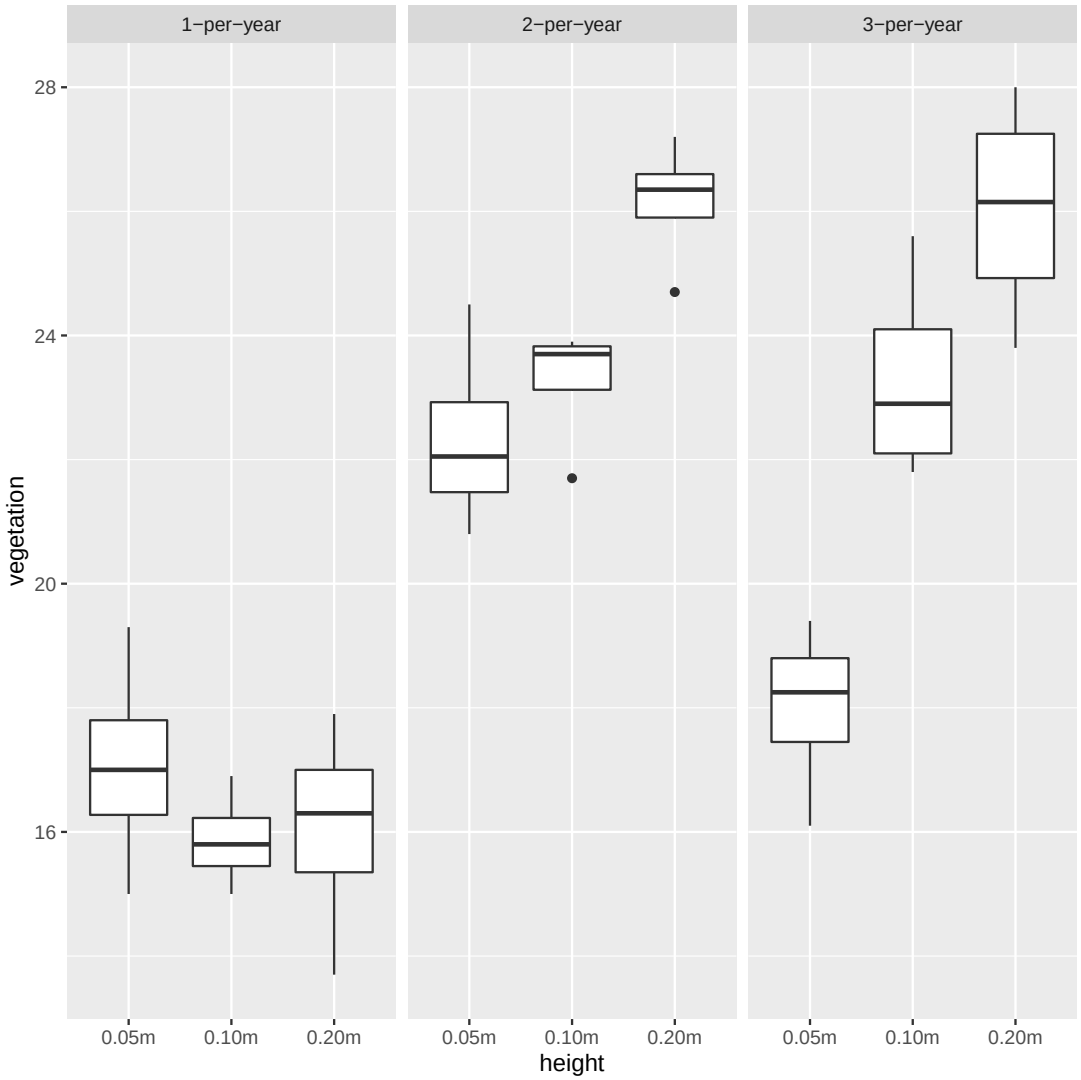
```
ggplot(mowing, aes(x=height, y=vegetation, fill=frequency)) + geom_boxplot()
```



The logic is that `colour` colours the *outside* of the boxes, but `fill` uses a different colour for each *whole* box.

Yet another possibility is facets. This might have come to you if you thought there were "too many variables", and you knew what to do if there was one categorical variable fewer. This is in some ways easier than a grouped boxplot because you take your boxplot from above and "add the facets to it":

```
ggplot(mowing, aes(x=height, y=vegetation)) + geom_boxplot() +
    facet_wrap(~frequency)
```



Or put `height` in the `facet_wrap` and `frequency` as the `x`, but I think the way I did it is a more logical consequence of what you did before, so is more likely to be what you'd think of. (I actually like this as an answer from you, because it shows that you get the concept of how facets work. We didn't do an example like this, that I recall.)

As I look at this, the facetted boxplots are quite similar in concept to the grouped boxplots (with the facets corresponding to the x-axis groups), so the finishing point is similar even though the code is different.

If you thought there were two quantitative variables: well, on the face of it, you were wrong, but if you made some sensible effort to create a second one (eg. with `parse_number` or `separate`), then you would be correct. In the more likely case that you asserted two quantitative variables

but didn't make them, you get nothing for naming the graph, but you get two if you drew what you said you wanted.

Lots of possibilities makes this a pain to grade, but that's the nature of this course. One point for naming the graph you are going to draw, plus two points for drawing it properly. For the drawing part, one point if you had one or more errors in your code (but still had something sensible). Another way to get one point for the drawing part (and thus one overall) is if you said to draw a *simpler* graph than one of these (in the grader's estimation) and drew it correctly. (Normally, if you make an error earlier and then what follows would have been correct had you not made an error, you get full marks for the later thing, but if your earlier error makes your later work *simpler*, that does not apply.)

Extra: I wanted to ask you to interpret (one of) the above graphs, but I couldn't figure out a way to ask that without giving away the kind of graph I wanted you to draw (which was the actual point of the question: did you know that a grouped boxplot or something equivalent was the way to show these data?).

The story, looking at one of the graphs, is actually a rather interesting one: the most consistent way to keep the vegetation down is to mow it only once per year (which is also the cheapest, because the maintenance crews will be least busy). If you mow more often than that, having a higher blade generally also results in more vegetation.

This might strike you as odd, because you would think there would be *less* vegetation if you mow it more often, and *less* vegetation if you use a larger blade height (so that you are cutting more of it down each time). But the story seems to be opposite: the more often or more vigorously you cut the vegetation back, the *faster* it grows back. I don't know whether this makes any kind of sense to you, but the closest thing I know of is that when you're growing flowers, and you want lots of flowers, you cut off the old flowers as soon as they've finished blooming, because that will encourage the growth of new ones. I guess this is the same idea. It's one of those cases where you take your results back to an expert and ask "Here, does this make any sense?".

Question 2 (13 marks)

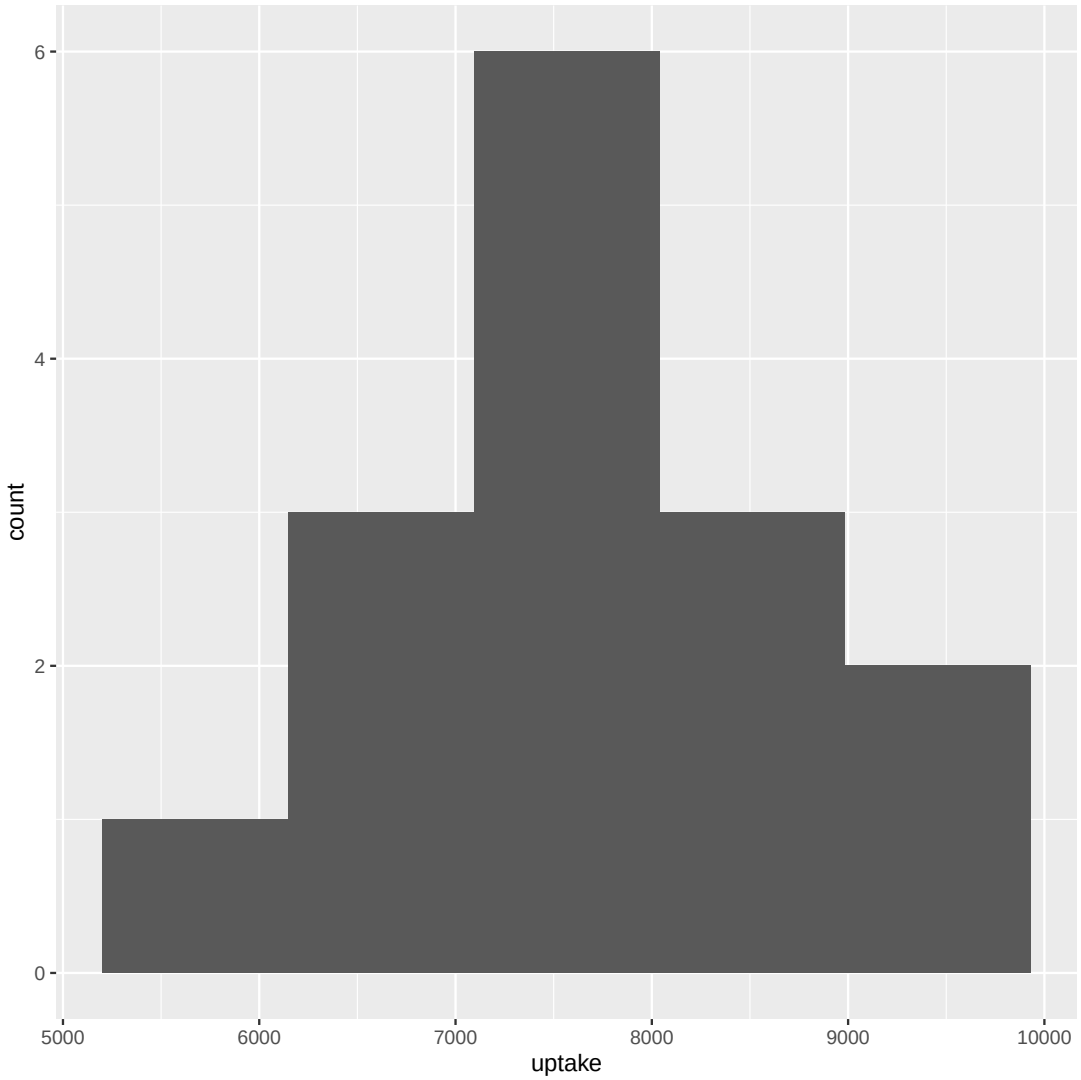
People are concerned about the use of nitrates as meat preservatives. One study looked at possible effects of these chemicals. Bacteria cultures were grown in a medium containing nitrates, and the rate of uptake of radio-labelled amino acids was then determined for each culture. The data are shown in Figure 3, in units of disintegrations per minute. The data are shown in Figure 3.

It is known that the mean rate of uptake for cultures grown without nitrates is 8000 in these units. We will be investigating whether the addition of nitrates results in a decrease in mean uptake rates.

(a) (2 marks) A histogram is shown in Figure 4. Give the code that was used to produce this histogram. (The data frame is called `nitrates`.)

My answer: Copying and pasting from my code and output (the code I didn't show you there):

```
ggplot(nitrates, aes(x=uptake)) + geom_histogram(bins=5)
```



I don't usually ask you these this way around, but your thinking ought to be "well, this is a histogram" and you piece together the `ggplot` from the name of the data frame and the name of the (one) column in it. You know the column is called `uptake`, because that appears in Figure 3, and also on the x-axis of the histogram. Since you are used to thinking about a number of bins for histograms you draw, I expect you to write a `bins=` inside the `geom_histogram` and then go to the actual histogram and count the number of bins it has. (I made this part of the exercise nice and easy for you, although with 15 observations, five bins is a very sensible kind of number, so I wasn't really faking it up.)

Two points for the above. I can't think of any variations on it that would also be correct. (Actually, I can, but I haven't talked about `binwidth` with you folks.) One point if you forget

the `bins` or mess up the name of either the data frame or the column in it. If you manage to make more than one mistake, don't expect to get much.

(b) (2 marks) The researchers wanted to assess the effect of nitrates on "typical" uptake rate. They ran a t-test rather than a sign test. Why do you think they decided to do this? Explain briefly.

My answer: The histogram in Figure 4 looks very much normal, with a symmetric shape for the distribution of uptake rates, and no outliers.

"The distribution is normal" is not enough; which distribution, and how do you know it's normal? That would be a one-point answer. Give me just "normal" as an answer, and I'm wondering whether you deserve anything at all.

(c) (3 marks) Give code to obtain a suitable t-test.

My answer: We want to prove that the mean is *less* than 8000, so that's the alternative. The null mean is thus 8000. So, one of these (either one is good):

```
with(nitrates, t.test(uptake, mu=8000, alternative="less"))
```

```
##
##  One Sample t-test
##
## data:  uptake
## t = -0.81599, df = 14, p-value = 0.2141
## alternative hypothesis: true mean is less than 8000
## 95 percent confidence interval:
##      -Inf 8244.674
## sample estimates:
## mean of x
##    7788.8
```

or

```
t.test(nitrates$uptake, mu=8000, alternative="less")
```

```
##
##  One Sample t-test
##
## data:  nitrates$uptake
## t = -0.81599, df = 14, p-value = 0.2141
## alternative hypothesis: true mean is less than 8000
## 95 percent confidence interval:
##      -Inf 8244.674
## sample estimates:
## mean of x
##    7788.8
```

The one-sample `t.test` doesn't take a `data=`, so you can't do it that way.

The usual three-pointer mark scheme: three if all correct (in one of the variations), two if one error, one if more than one error but something substantial correct.

(d) (2 marks) The output from your t-test is shown in Figure 5. What do you conclude from it, in the context of the data?

My answer: The P-value of 0.2141 is not less than 0.05, so we do not reject the null hypothesis, and thus there is no evidence that the mean uptake rate is less than 8000. We have not shown that adding nitrates to the culture decreases the uptake rate.

This is not a proof that the mean uptake rate has stayed the same (since that is not the way hypothesis tests work), but it is suggestive of that, unless we can find another explanation.

Two for that, one for getting as far as "fail to reject the null", or for getting all the way but messing something up.

(e) (2 marks) What *two* changes to the code for your hypothesis test would produce output containing a 95% confidence interval for the true mean uptake rate? (If you only think you need one change, explain briefly why your one change is sufficient.)

My answer: Take out the `mu=8000` and the `alternative="less"`. The first is what makes `t.test` test that null hypothesis, and the second is what makes the test one-sided, and you don't need either of those for a confidence interval.

You might be wondering whether you have to specify the 95% confidence level with `conf.level`, and the answer is that you don't because it's the default (it'll be the confidence level if you don't specify one).

Another possibility is to leave the `mu=8000` in. This will do a test, which you then ignore. But you *must* remove the `alternative="less"`, because doing a one-sided test will also produce a *one*-sided confidence interval, which is not how we've done them. Notice how in Figure 5, the confidence interval given goes all the way down to minus infinity? You can optionally replace it with `alternative="two.sided"`, but that is also the default, so there is no need. Since I asked for two changes and we only have one so far (taking `alternative` out and maybe replacing it with something else), you can supply the addition of `conf.level=0.95` for the other change. This isn't necessary, but it doesn't hurt either, and since I asked for *two* changes, you should either supply a second one or explain why it is not needed (such an explanation might be "remove only `alternative="less"` and ignore the test and P-value that come out").

If you got the `t.test` code wrong above, do your best to give two changes from what you wrote that will get the CI. We will try to be sympathetic.

As a last thing: if I ask for *changes* from one piece of code to do another job, you need to tell me what changes. Giving me code that will do the second job is not telling me how the first code changes, so it is not answering the question. Expect to get one out of 2 if you do this. (Imagine your boss asking "how do I change this code to get a confidence interval?", and you just give her the code to get the confidence interval. She will look at you funny, because you haven't answered the question she asked.)

(f) (2 marks) A 95% confidence interval for the population mean uptake rate is shown in Figure 6. The researchers thought that maybe adding nitrates to the bacterial culture might reduce the mean uptake rate by 500. By looking at the confidence interval, do you think that the researchers would have had a reasonable chance of being able to prove that the mean uptake rate was less than 8000, using the sample size that they had? Explain briefly.

My answer: A 500-unit reduction from 8000 would be 7500. You see that *both* 7500 and 8000 are inside the confidence interval, so either one is plausible in the light of the data: the data don't allow us to distinguish between them. This is because the confidence interval is too wide; we would need a larger sample size to be able to reject a mean of 8000 when the population mean is actually 7500, because the data values are too variable.

This sounds an awful lot like power, and it is, but I didn't give you the machinery to see what the power would have been. I wanted to test your intuition: if the confidence interval is wide, the reason we didn't reject the null might be that things are too uncertain: that is to say, given the small sample size *and* the large amount of variability in the data, the test didn't have enough power. (I want you to mention both of those things for the two points.) Of course, we

might have failed to reject the null of 8000 because it is actually *true*, but the point is that with large variability and small sample size we cannot distinguish the two possibilities.

Extra: with a little work, we can figure out the power of a test with this sample size to reject 8000 when the mean is actually 7500, thus:

```
nitrates <- read_csv("nitrates.csv")
```

```
## Parsed with column specification:
## cols(
##  uptake = col_double()
## )
```

```
nitrates %>% summarize(s=sd(uptake))
```

```
## # A tibble: 1 x 1
##       s
##   <dbl>
## 1 1002.
```

and then, using the sample SD as an estimate of the population SD:

```
power.t.test(n=15, delta=8000-7500, sd=sd(nitrates$uptake),
             type="one.sample", alternative="one.sided")
```

```
##
##      One-sample t test power calculation
##
##               n = 15
##           delta = 500
##              sd = 1002.431
##       sig.level = 0.05
##           power = 0.5763105
##     alternative = one.sided
```

This is actually not as bad as I was fearing, but it's still not very good, and with this much variability, a larger sample size would definitely be a good idea.

Another way to go at this is to think about the length of the confidence interval. This is twice the margin of error, since we have to go both up and down from the sample mean. Thus the length is:

$$2t^* s/\sqrt{n}$$

where t^* is the number that comes out of the t-table, which depends on the degrees of freedom, but for a 95% CI it's about 2. s is the sample SD, which is going to be about 1000 (on the evidence of the data that we have). If we are going to get one of 8000 and 7500 inside the interval and one outside, we want the length to be less than 500, which means solving:

$$2(2)(1000)/\sqrt{n} = 500$$

or

$$\sqrt{n} = 4000/500 = 8.$$

Using our very rough numbers, the sample size would have to be something like $8^2 = 64$ or more to get the CI to be this short.

Another way to get to the same place is to note that the CI is currently just over 1000 long; we want it to be a little less than half that long, and to do that will mean multiplying the sample size that we have by a bit more than $2^2 = 4$. The sample size is now 15, so that once again means taking a sample size of something over 60.

I'm not quite sure what kinds of answers I'm going to get to this one, because you haven't seen one like this before. My aim with this kind of question is to see if you can produce some sensible thinking that shows some attention to the relevant issues.

Question 3 (14 marks)

In a power plant, water is used for cooling, and the water is then discharged into a nearby river. It has been determined that as long as the mean temperature of the discharged water is no more than 65 degrees Celsius, there will be no negative effects on the river's ecosystem. To find out whether the power plant is discharging water that is too warm, a scientist will take 50 water specimens (at randomly selected times) and record the temperature of each one. If μ denotes the mean temperature of all the discharged water, the scientist will then test $H_0 : \mu = 65$ against $H_a : \mu > 65$.

(a) (2 marks) Describe a type I error in this context.

My answer: A type I error is rejecting the null hypothesis when it is true: that is to say, declaring the mean temperature of the discharged water to be greater than 65 degrees when it actually is 65 (or less).

One point if you get as far as "rejecting the null when true". You need to talk about water discharge temperatures to have a shot at the second point. That applies to the next part as well.

(b) (2 marks) Describe a type II error in this context.

My answer: Failing to reject the null when it is actually false: that is, saying that the mean temperature is 65 when it is actually higher than that.

(c) (4 marks) The population standard deviation of water temperature measurements is believed to be about 15 degrees. Give R code to **estimate** the power of the test when the mean water temperature is actually 68 degrees. Assume that water temperature readings have at least approximately a normal distribution.

My answer: I bolded the word "estimate" to alert you that this is by simulation. Steps:

- generate lots of samples (of size 50) from a normal distribution (with mean 68 and SD 15)
- for each one, run a t-test with a null mean of 65 and a one-sided alternative (don't forget that)
- for each of those t-tests, pull out the P-value and save it
- make a table of how many of those P-values are less than 0.05.

This is how it goes:

```
rerun(1000, rnorm(50, 68, 15)) %>%
    map(~t.test(., mu=65, alternative="greater")) %>%
    map_dbl("p.value") -> pvals
tibble(pvals) %>% count(pvals<0.05)
```

```
## # A tibble: 2 x 2
##   `pvals < 0.05`     n
##   <lgl>          <int>
## 1 FALSE            583
## 2 TRUE             417
```

The guideline is one mark for each line correct. The grader will decide how picky to be in case of mistakes (and they will be consistent from student to student). You can replace 1000 on the first line by any number about 100 or bigger (this is the "lots of samples"). Or you can look ahead at Figure 7 to see that I actually used 1000 samples as my "lots of them".

(d) (2 marks) The output from your code is shown in Figure 7. What is your estimated power for this test?

My answer: 413 out of 1000, or 0.413. A gimme two points if ever there was one. (One point for giving me the probability of a type II error, or of something else that is obviously "almost" 0.413.)

(e) (2 marks) The scientist's manager would prefer to design the test to have a power of 0.7. Using Figure 8, approximately how many water specimens would the scientist need to take to achieve this?

My answer: Looking at the graph, getting power 0.7 would require a sample size of about 120. (I would accept anything between about 110 and 130, although a sample size of 125 gives more power than 0.7: the point just above 0.7 on the y-axis).

I didn't ask for an explanation. However, if you don't give one, the grading choices are full marks (correct), *nothing* (wrong). If you have the right idea but get the answer wrong, you may get a point if you explain what you did. A number of people gave an answer of 150 or so; I thought this was obviously too high, but you get a point for being "close". The printed graph came out less clear than I would have liked. Apologies. This might have had something to do with the high estimates.

Extra: I made this graph with `power.t.test` so that it is slightly inconsistent with my simulation. The exact sample size required is thus:

```
power.t.test(power=0.7, delta=68-65, sd=15,
             type="one.sample", alternative="one.sided")
##
##      One-sample t test power calculation
##
##               n = 119.0051
##           delta = 3
##              sd = 15
##       sig.level = 0.05
##           power = 0.7
##     alternative = one.sided
```

119 fails by a minuscule amount to give enough power:

```
power.t.test(n=119, delta=68-65, sd=15,
             type="one.sample", alternative="one.sided")
##
##      One-sample t test power calculation
##
##               n = 119
##           delta = 3
##              sd = 15
##       sig.level = 0.05
##           power = 0.6999836
##     alternative = one.sided
```

and so the required sample size actually *is* 120. But I didn't ask *you* for code; I asked you to use the Figure.

Extra: in `power.t.test` the appropriate alternative is written `one.sided`, but in the simulation above, you are using the ordinary `t.test`, and so `alternative="greater"` as usual there.

(f) (2 marks) Explain briefly why your answer to the previous part compares with the original number of water specimens in the way that it does.

My answer: The original 50 water specimens gave a power of only about 0.4, which was too low, and so to increase the power we have to take more water specimens. Hence the answer to the previous part is quite a bit bigger than 50.

This is true from the power curve (it has a positive slope everywhere), but it is also a general principle: all else equal (as it is here), larger sample goes with larger power. I would like to be

convinced that you know how it is supposed to work, and that it *does* work that way in this example.

An actual answer:

3f 2

As sample size increases, so does power. Therefore at a smaller sample size of 50, power is very weak (0.413) increases to 0.7 when sample size goes up to 125.

Question 4 (8 marks)

In each of the situations below, say whether you would use a *one-sided* or *two-sided* test, and also whether you would use a *one-sample* or *two-sample* test. In each case, therefore, you need to say something about both the number of sides and the number of samples, and give a brief justification of your choices.

(a) (2 marks) A study of wait times in coffee shops was carried out in Boston. The researchers were concerned about whether females had to wait longer than males on average. They observed a number of customers, and for each one recorded the wait time in seconds and whether each customer was male or female.

My answer: Two-sample, one-sided.

We are comparing males' wait times and females' wait times *with each other*, so this is a two-sample situation. We are trying to show that females wait longer than males (rather than just whether average wait times are different), so this is one-sided.

Extra: I tried to phrase this in a way to help you imagine a column of wait times, with a second column of genders, which is exactly what we've been using for our two-sample tests. I thought about saying "a sample of males had their wait time recorded, and also a sample of females", which would sound more like that "wide data" at the end of Assignment 3 (for 2019 students), and I think that would have been unnecessarily confusing. So I didn't do it that way.

Marking guide for all of these: one point for properly justifying the number of samples, and one point for properly justifying the number of sides. Thus, getting the answer right but the explanation wrong is worth nothing. In this course, you have to get a good answer *for a good reason*, and there is no credit for an answer you can't support.

(b) (2 marks) A machine part has a hole in it that is supposed to be exactly 5 centimetres in diameter. The machine part is produced by a process that can be adjusted to produce parts with holes of different sizes. The process supervisor takes a sample of 10 parts (from the process at its current settings) and finds the mean and standard deviation of hole sizes of these parts. If the mean hole size is too far away from 5 cm, the process supervisor will need to adjust the process settings.

My answer: One-sample, two-sided.

There is one sample of parts (of size $n = 10$), the holes in which are compared to some external standard (a diameter of 5 cm). Any departure from 5 cm is a concern, either too large or too small (the last sentence), so the test needs to be two-sided. (In general, if something is supposed to be a certain size, then either too big or too small is a problem.)

(c) (2 marks) The instructor of a large introductory psychology class believes that students need to spend 10 hours studying for the final exam to master the material, and is concerned that students are studying less than they should. The student newspaper reported that, for a random sample of 411 students in the course, the mean time spent studying for the final exam was 7.74 hours with a standard deviation of 3.40 hours.

My answer: One-sample, one-sided.

We are comparing one sample of students with some external value (the instructor's belief of 10 hours). The 10 hours did not come from a second sample, so there is only one sample here.

Question 4 continues… This page: 6 marks.

The instructor is trying to demonstrate that students study *less* than she thinks they need to, which calls for a one-sided alternative hypothesis "mean is less than 10".

If the comparison had been between students now and students five years ago (say), with a mean study time from a sample of students from five years ago, that would have required a two-sample test. But it was not that: the comparison was of one sample of students with some external value.

Extra: I gave you some numbers, so there actually is enough information to do the test here, not that I needed you to. The data are summary statistics, so you have to do it "by hand" rather than with `t.test`:

```
t_stat <- (7.74-10)/(3.40/sqrt(411))
t_stat
## [1] -13.47567
p_value <- pt(t_stat, 411-1)
p_value
## [1] 8.020945e-35
```

`pt` gives the probability of getting a value less or equal to the first input, on a t distribution with degrees of freedom equal to the second input. The P-value is very small, so there is indeed evidence that (all) students in this course study fewer hours for the final exam than the instructor thinks they should. This is a one-sided lower-tail test (the alternative is that the students study *less* than 10 hours), so looking at the lower tail of the t-distribution is correct. (If you were doing this with tables, you'd forget about the minus sign on on -13.47, and look up 13.47 on the appropriate degrees of freedom line, probably "infinity", noting that 13.47 is way off the end.)

(d) (2 marks) Forty men were recruited from a dating site. Each man has a profile in which he reports his height. The researchers also recorded each man's actual height, and compared it with the height reported in that man's profile. The researchers were trying to find out whether men on this dating site systematically reported themselves as being taller than they actually were.

(If you think this one will be neither a one-sample nor a two-sample test, describe what you think it is instead.)

My answer: This is matched pairs (or one-sample), one-sided.

There are two measurements for each man: their actual height and the height they reported. This is matched pairs, or you could think of it as one sample of differences between the two heights. It is not two independent samples (which is what "two-sample" requires), because the first reported height and first actual height will be dependent on each other (they are the same man). The researchers were trying to find out whether the reported heights were on average bigger than the actual heights, which requires a one-sided test (rather than just whether they were different, which would have been two-sided).

Extra: the source I got this scenario from had some data as well, for women as well as men. The men did overstate their heights, by an average of half an inch (significantly higher than zero), while the women clearly did not (not significantly higher than zero).

If you wanted to compare males and females here, you would note that the males and fe-

males were independently selected, and do a two-sample test on the two sets of matched-pair differences!

My thanks to Aasha for the idea for this question.

Question 5 (7 marks)

Medical research has shown that repeated wrist extension beyond 24 degrees increases the risk of hand and wrist injuries. Some students at Cornell University were given a proposed new mouse design. While using the mouse, each student's wrist extension was measured. Our interest is in whether the average wrist extension is greater than 24 degrees, where the average could be the mean or median.

(a) (4 marks) A histogram of the wrist extension values is shown in Figure 9. Two possible analyses of these data are shown in Figures 10 and 11. Which of these analyses do you prefer and why, and what do you therefore conclude in the context of the data?

My answer: On the histogram, I see two low outliers, or two wrist extension values that are much lower than the others. I think this is two low outliers, but if you want to call it left-skewed, that works as well.

With the outliers, we don't trust the t-test (or, indeed, the mean), so we look at the sign test in Figure 11. The P-value is 0.0013 (one-sided), so we reject the null median of 24 and conclude that the median wrist extension is greater than 24 degrees.

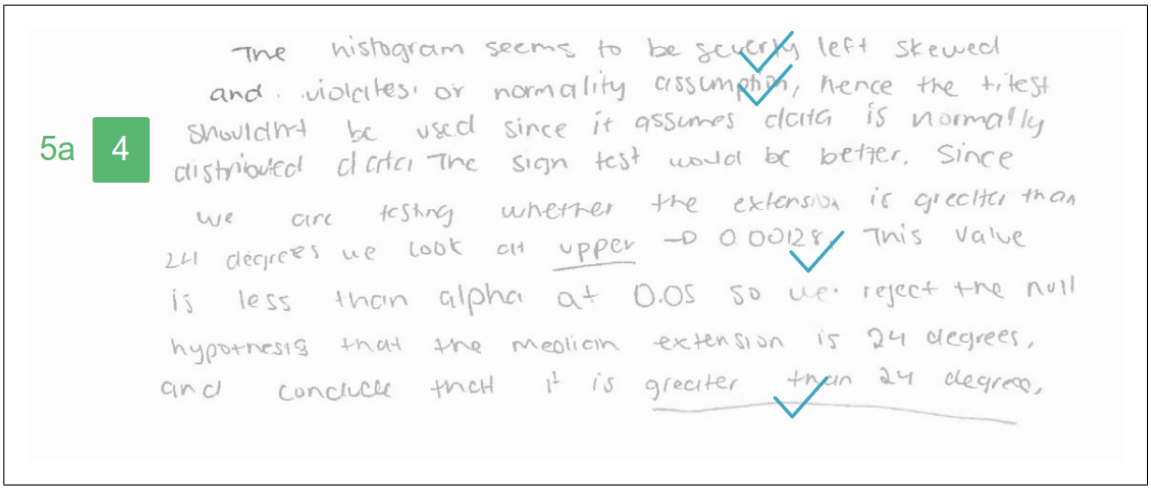
Two points for choosing the sign test with a good reason (no points if you don't have a reason). A good reason would be "there are two low outliers", a one-point answer would be choosing the sign test "because there are outliers" without saying specifically where they are. I was usually persuaded to give 1 out of 2 by an assertion of "not normal" in an otherwise good argument, but you need to tell me *how* it's not normal.

One more point for getting the right P-value and using it to reject the null, and one last point for saying something sensible about median wrist extension.

Choosing a test on the basis of its P-value, as some people tried to do, is *bad science*. This is the same problem that doing multiple t-tests has: you're giving yourself multiple chances to reject a null, and thus the α for your testing procedure is not 0.05 even if you think it is. Look at the data first (with a graph), and use *that* to decide what to do.

Extra: if I were organizing this question differently, I would now ask you about generalizing the results. These were students at Cornell, and the extent to which you can generalize these results depends on how representative students at Cornell (an Ivy League school) are of students generally. Alternatively, the question is whether you think these Cornell students look like a random sample of "all possible students", in terms of their wrist extension when operating a mouse.

An actual answer:



(b) (3 marks) If you don't remember `filter`, look at Figure 12 to see how it works.
Some further analysis is shown in Figure 13. By comparing this Figure with Figures 10 and 11, what do you learn about the behaviour of the t-test and sign test, and what general principle does it illustrate? Explain briefly.

My answer: `wrist2` differs from `wrist` in that the outliers have been removed (the outlying values, from the boxplot, were less than 10, and they were the only values that were, and we selected the values greater than 10, so we omitted the outliers and nothing else).

The extra analysis in Figure 13 is based on the data set `wrist2` without the outliers. The issue is how the behaviour of the t-test and the sign test change, when you remove the outliers. I think the obvious thing to look at is the P-values. We now have four tests, t and sign with and without the outliers, and the P-values look like this:

Outliers	t-test	Sign test
With	0.4212	0.0026
Without	0.00008	0.00007

The P-value for the t-test has changed dramatically when you take out the outliers, going from clearly nonsignificant to clearly significant. The P-value for the sign test has also changed (becoming a lot smaller), but it has not changed as much and (perhaps more important) its level of significance has not changed (it's "strongly significant", or whatever words you like to use, both times).

The picture we're getting from here is that the outliers have an enormous impact on the t-test, but not so much on the sign test. (This is the "general principle" I was fishing for.) This is another way of seeing that when you have outliers, neither the mean nor the t-test will be very reliable, whereas the sign test will still be, because it is based on the median which is not affected (so much) by outliers.

Another way of looking is along the rows of my table: with the outliers in, the P-values are completely different, so it then matters a great deal which test you do. But with the outliers removed, the two tests are giving almost identical results. The moral again is that the t-test can be greatly affected by outliers, while the sign test isn't.

Extra: we said, looking at the histogram in Figure 9, that the distribution of the values that were not outliers looked pretty normal. Hence, when you take them out, the t-test and sign test are telling pretty much the same story.

Grading: one mark for explaining how `wrist` and `wrist2` are different. One for comparing the t-tests between Figures 10 and 13, and also for comparing the sign tests between Figures 11 and 13. You need both for the second mark. (I figure you'll be able to do this even if you have no idea what else is going on, but it's really only about a third of the whole thing.) One mark for making an overall comparison of the results in the context of how the data sets are different. An answer like "the t-test changes a lot but the sign test only a little" by itself is thus only one mark (the second one), because you haven't said how the two data sets are different, and you haven't said why the results are interesting, or are what you would guess (because of how the data sets are different). Saying how `wrist` and `wrist2` are different is (on the face of it) only one mark, but it is crucial to getting the third mark as well.

I was also sympathetic to a discussion like "if you remove the outliers, the data are approximately normal, and then it makes sense to look at the t-test".

I meant this part to be fairly difficult, since I was leaving you to do some detective work, and I didn't give you many clues.

Extra extra: these were one-sided tests, so it didn't make much sense to look at confidence intervals, but if we do that instead, a similar picture emerges. First, with the outliers:

```
t.test(wrist$extension)
```

```
##
##  One Sample t-test
##
## data:  wrist$extension
## t = 20.311, df = 24, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  21.77689 26.70311
## sample estimates:
## mean of x
##     24.24
```

```
ci_median(wrist, extension)
```

```
## [1] 24.00098 26.99609
```

and then without the outliers:

```
t.test(wrist2$extension)
```

```
##
##  One Sample t-test
##
## data:  wrist2$extension
## t = 62.529, df = 22, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  25.01156 26.72757
## sample estimates:
```

```
## mean of x
##  25.86957
ci_median(wrist2, extension)
## [1] 25.00000 26.99609
```

The confidence intervals without the outliers are shorter, but the one for the median changes less than the one for the mean. This is consistent with what happened to the P-values.

An actual answer:

(b) (3 marks) If you don't remember `filter`, look at Figure 12 to see how it works. Some further analysis is shown in Figure 13. By comparing this Figure with Figures 10 and 11, what do you learn about the behaviour of the t-test and sign test, and what general principle does it illustrate? Explain briefly.

5b 3

By filtering out the outlier, the t-test's p-value is lower, and becomes significant. On the other hand, the sign test ended up with a lower p-value, but not by much, and it was already significant. This shows us that the t-test is more susceptible to outliers and cannot be trusted with non-normal data, whereas the sign test does not significantly affect the sign test.

Another:

5b 3

The code on Figure 13 get rid of the 2 outliers. With the outliers, the t-test cannot be performed and it's, it will cause type II error. After we get rid of the 2 outliers, the t-test can be used, the p-value is very significant. For the sign test, the outliers don't effect the sign test significantly. In short, the t-test is significant influenced by outlier, but sign test isn't.

Exam continues... This page: 7 marks.

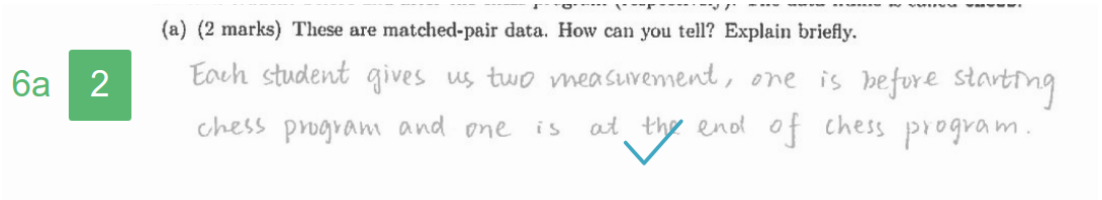
Question 6 (11 marks)

Is it true that learning to play chess can improve your memory? In a study, sixth-grade students who had not previously played chess took weekly chess lessons and played chess daily for 9 months. Each student took a memory test called the "Test of Cognitive Skills" before starting the chess program and again at the end. The data are shown in Figure 14, with `pre_test` and `post_test` denoting the scores for each student before and after the chess program (respectively). The data frame is called `chess`.

(a) (2 marks) These are matched-pair data. How can you tell? Explain briefly.

My answer: Each of the 12 students has two measurements (one before and one after the chess program). Or, if this were going to be two-sample, we would have two independent groups of 12 students, one group measured before and one after, and thus 24 students altogether. Or, the measurements in (say) the first row are all on the 1st student, so they will not be independent. One of those, or something like it. Say what makes it matched pairs, or what makes it *not* two (independent) samples. To be precise, say what makes *this* data set matched pairs, rather than giving me a definition of what matched pairs is. Your job is to show how that definition applies to *this* data set.

An actual answer:



(b) (3 marks) Give code to run a suitable t-test on these data.

My answer:

This is a matched pairs t-test, so your code should reflect that one way or another. You can use the actual measurements, thus:

```
with(chess, t.test(pre_test, post_test, alternative="less", paired=T))
```

```
##
##  Paired t-test
##
## data:  pre_test and post_test
## t = -4.564, df = 11, p-value = 0.0004057
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -87.69083
## sample estimates:
## mean of the differences
##                -144.5833
```

or switch pre and post around and make the alternative `"greater"`, or use the differences (since they are already in the data frame):

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

Question 6 continues... This page: 5 marks.

```
##
##  One Sample t-test
##
## data:  difference
## t = 4.564, df = 11, p-value = 0.0004057
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  87.69083      Inf
## sample estimates:
## mean of x
##  144.5833
```

since the differences were after minus before. (I took the differences this way around, so you have to use them this way around — no `less` for the alternative here. Of course, if this were your data frame and you were sitting in front of R Studio, you could take the differences whichever way around you like.)

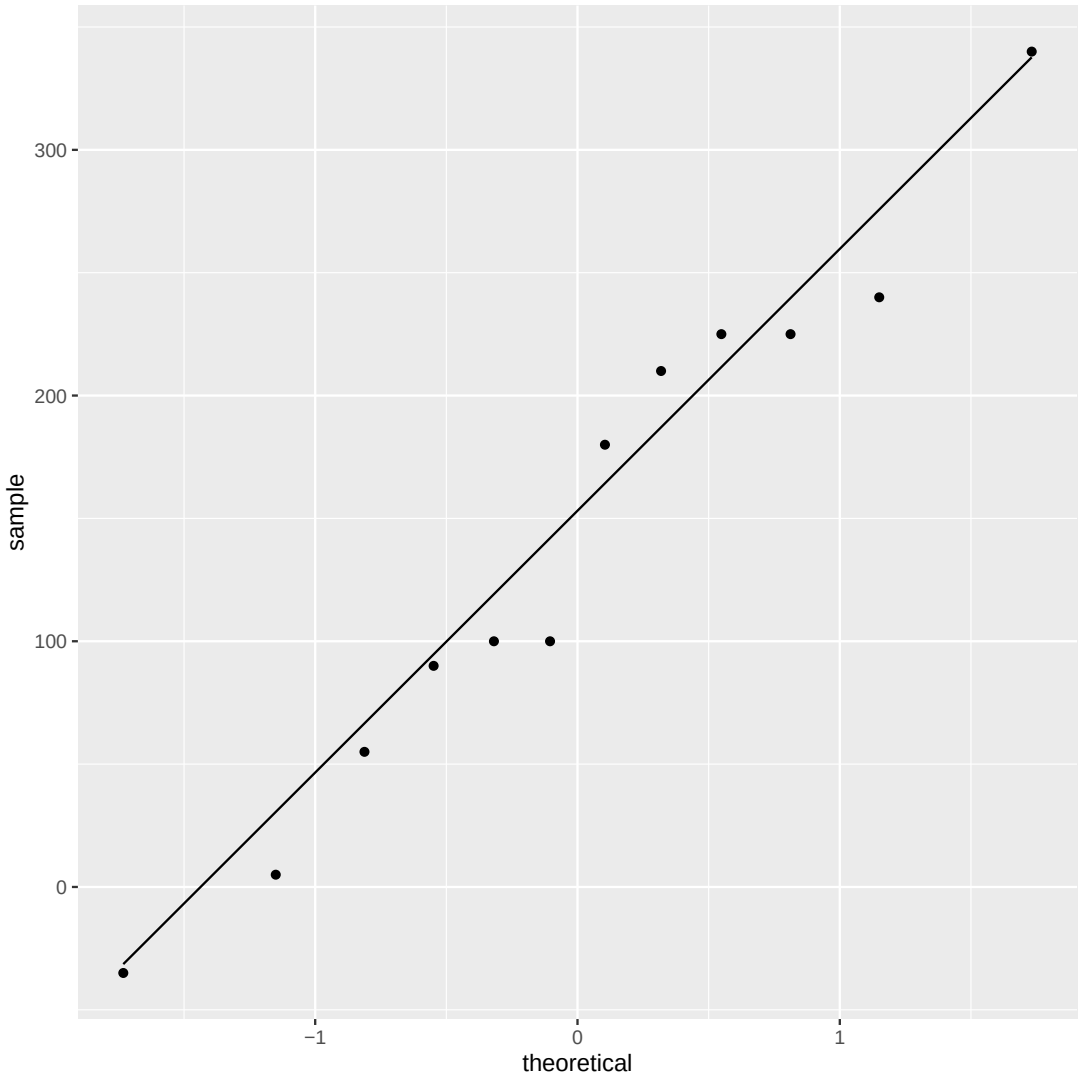
A lot of people (75 of them!) forgot the `alternative`-equals-something, and a few got the wrong one-sided alternative. If you do it the `paired=T` way, the alternative says how the column you put first compares to the one you have second, *in that order*, so that if you have `pre_test` first, it's got to be "less". (Also, the differences are already there, so you're wasting effort by computing them again, if you did that. If you did that, you rather betrayed that you were adapting something from your notes without thinking carefully about whether you needed it.)

(c) (3 marks) What code would produce a suitable graph for assessing whether the t-test you just did was appropriate?

My answer: The key assumption is that the *differences* are normal enough. It doesn't matter whether the pre-test or post-test scores by themselves are normal or not. Thus, a plot of the differences is needed. (There is no need for `pivot_longer` or anything like that, because we don't care about the test scores themselves.)

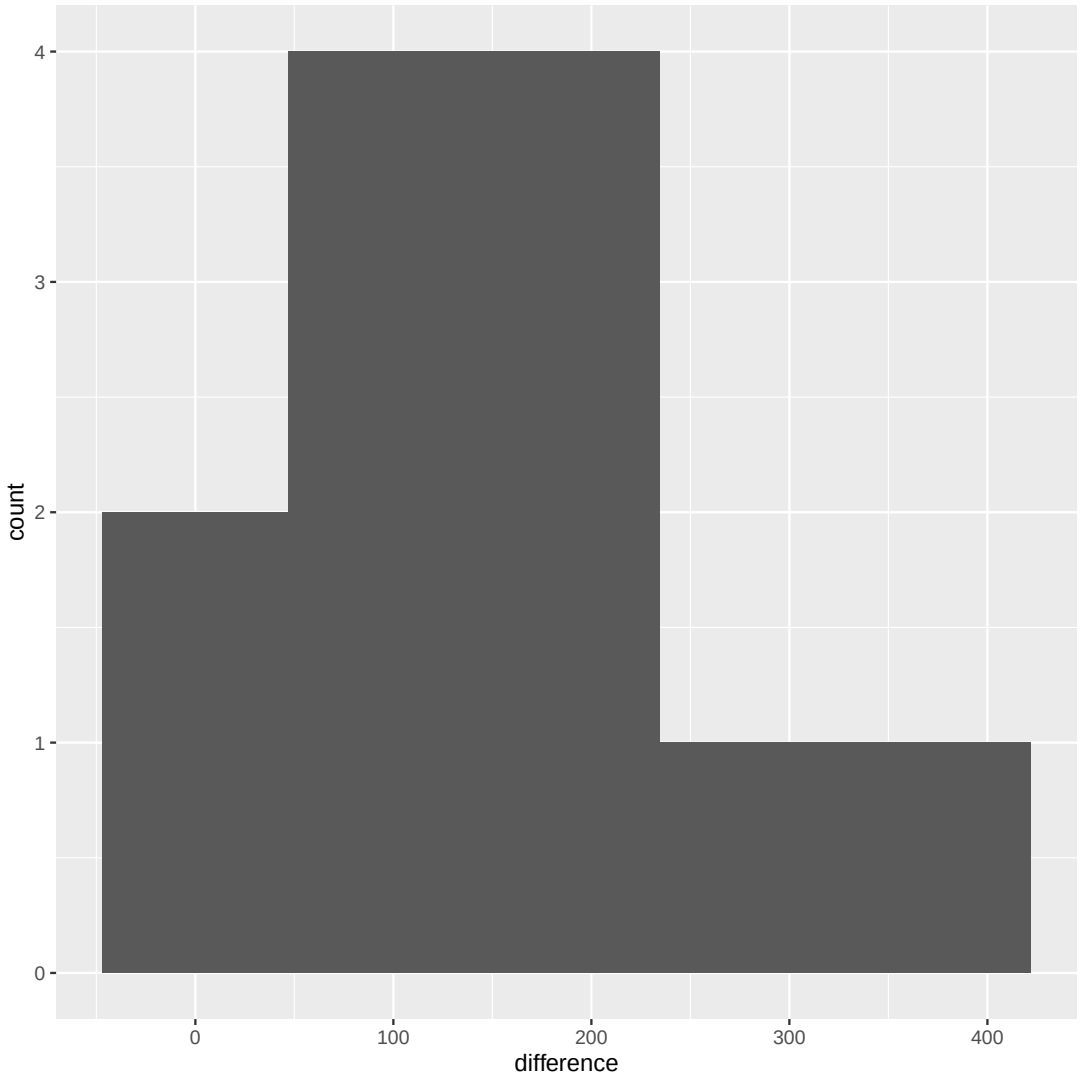
Perhaps the best plot is a normal quantile plot:

```
ggplot(chess, aes(sample=difference)) + stat_qq() + stat_qq_line()
```



but a histogram (or even a one-sample boxplot) would be fine also:

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



You should use a relatively small number of bins since there are only twelve differences.

In this case, I see no real problems with normality (with only 12 observations, you probably won't get a nice bell curve), and therefore I have no problem with using a t-test. But of course I couldn't ask you that because I didn't show you the graphs.

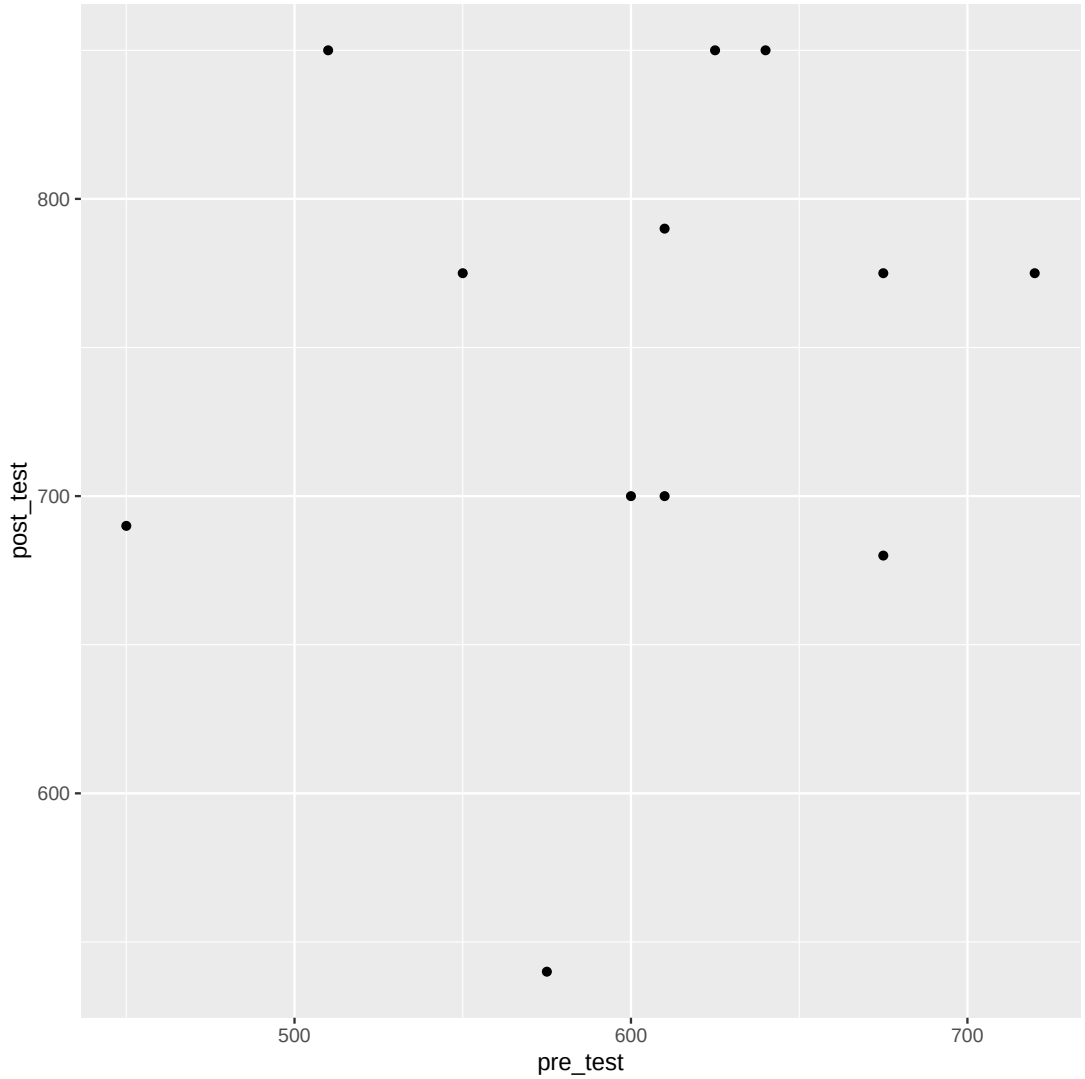
I don't know why so many people *calculated* the differences when they were already in the data frame. I lived with something like `mutate(diff=post_test-pre_test)`, but not something like `mutate(diff=difference)` which made me scratch my head rather. That cost you a point, because why do that extra work? Or using `diff` instead of `difference`, perhaps because you had `diff` in your notes. In *this* data frame, the differences are called `difference`.

You might have thought that a facetted normal quantile plot was the thing, or side-by-side

boxplots. But that would work for two independent samples, not for matched pairs. For *that*, it doesn't matter what the distributions of the before and after values were; they could be anything, as long as the differences are sufficiently normal.

I was willing to give one point for a graph showing the pre-test and post-test scores in some sensible fashion, even though it wouldn't assess the assumptions for a matched-pairs test. The obvious thing is a scatterplot:

```
ggplot(chess, aes(x=pre_test, y=post_test)) + geom_point()
```

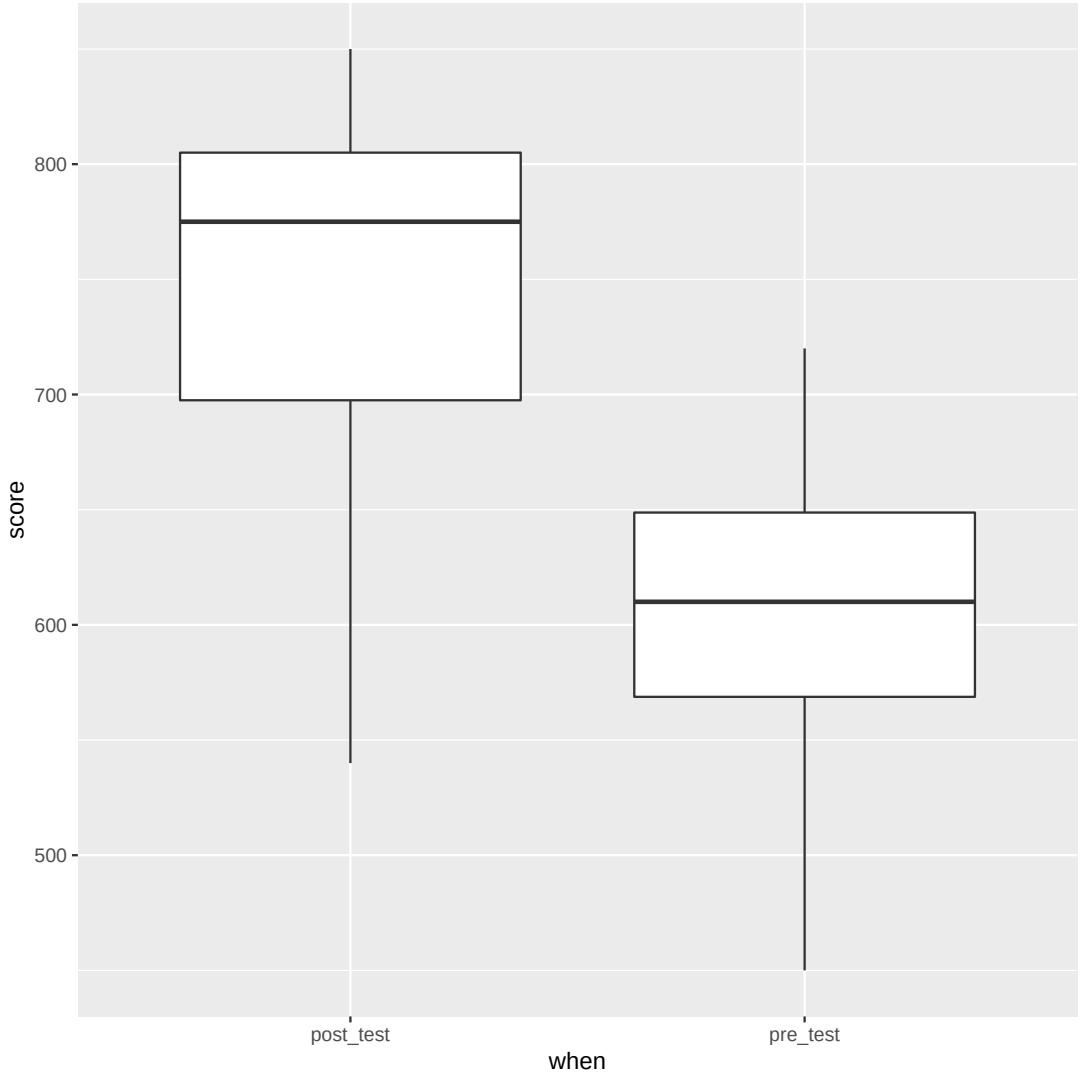


But you have to draw this graph correctly to get the point. (This is actually kind of interesting because typically in matched pairs you'll have some subjects that score high both before and after, and some that score low, and you'll see a trend. Not so much here, but quite often with

matched pairs.)

This idea, however, does not help:

```
chess %>%
    pivot_longer(pre_test:post_test, names_to="when", values_to="score") %>%
    ggplot(aes(x=when, y=score)) + geom_boxplot()
```

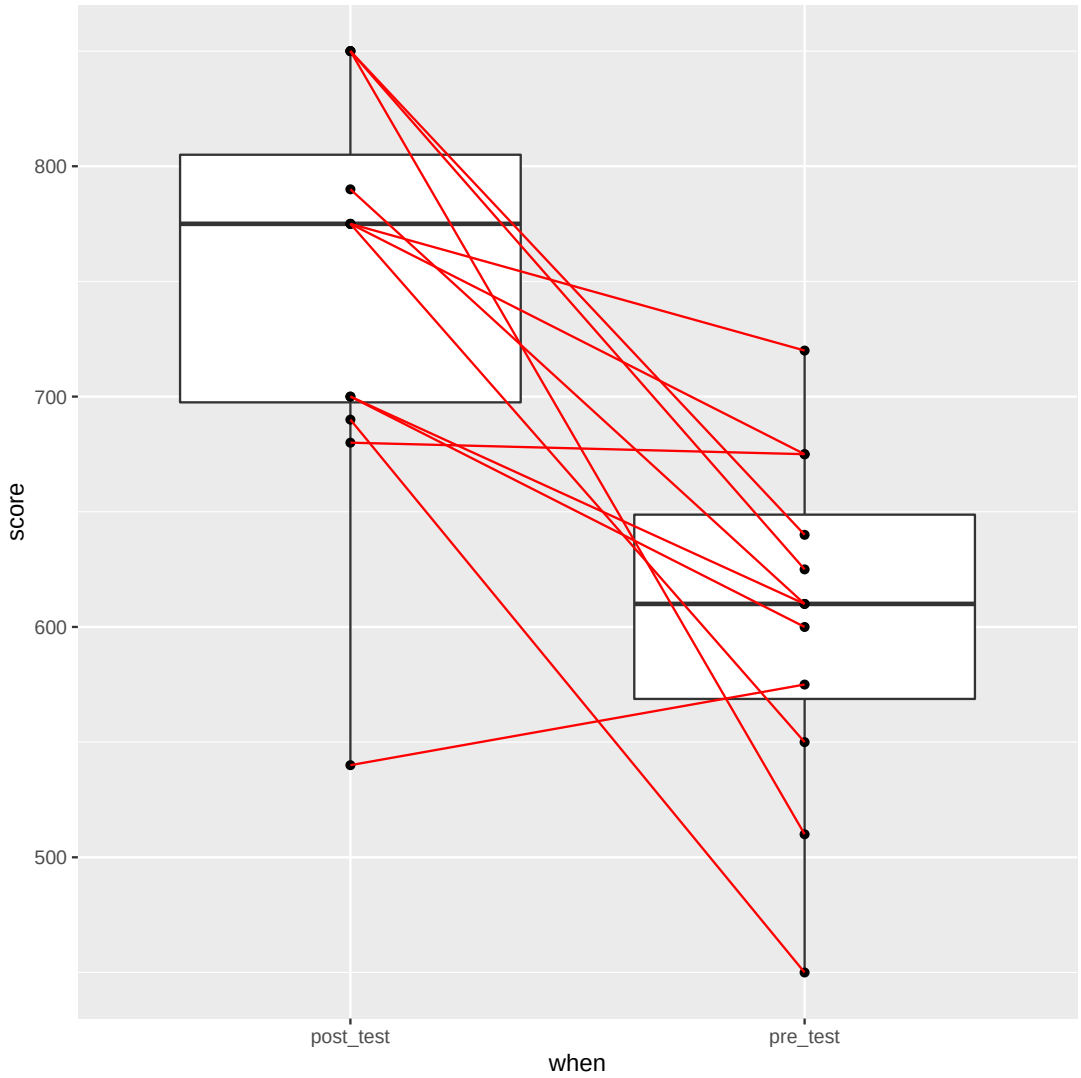


It's no help because it *breaks* the connection between the two test scores for each student. It would work if you had two independent samples (that were presented to you in wide format), but with matched pairs the connection between the two test scores is rather the point.

Here is a way of making the above graph useful:

```
chess %>%
```

```
pivot_longer(pre_test:post_test, names_to="when", values_to="score") %>%
ggplot(aes(x=when, y=score)) + geom_boxplot() +
geom_point() + geom_line(aes(group=student), colour="red")
```



This is a truly hideous graph, and wins no points for style. However, it does do two things: (i) it shows that post-test scores are higher than pre-test scores on average (the boxplots), and (ii) it shows that all students except one had a higher post-test score than pre-test one (the points, and the lines that join them go downhill in all but one case).

A couple of coding things: using `group` inside `geom_line` controls which points are joined by lines (the ones belonging to the same student, here), and putting the `colour` *outside* the `aes` makes all the lines literally red, rather than something like choosing a colour based on what

group they're in.

(d) (3 marks) Why might you have guessed, looking at the data in Figure 14, that a suitable sign test for comparing memory test scores from before and after the program would have produced a significant result? Explain briefly.

My answer: The sign test here is looking at whether or not the differences have median zero: that is to say, it's based on how many of them are above zero and how many below.

Only one of the twelve differences is negative, meaning that only one of the students had a memory test score that was higher before the chess program than after. (All the others had a higher score afterwards, which is what you'd expect to see if the chess program really helps.) So, we observe (i) an uneven split of the differences above and below zero, and (ii) it is uneven in the direction that we would expect if the chess program helps with memory.

This is a one-sided test, because we want to see whether memory is *better* afterwards, so it is important to think about whether we are on the "right side": that is, whether the data are pointing in the direction of the alternative hypothesis. (If they are pointing in the opposite direction, we stop right there and say that the P-value is "large", and say that there is no evidence that the chess program has a positive effect; in a case like that, the chess program would be *harmful* to memory rather than helpful.)

A point for looking at whether the differences are above or below *zero*, and a point each for something resembling (i) and (ii). Alternatively, note that 11 of the 12 post-test scores are higher than the corresponding pre-test scores, and then say that this is (i) unbalanced, (ii) in the right direction.

Talking about a sign test being good because there are outliers misses the point here. I am saying that we are *going* to do a sign test (whether you think it's a good idea or not), and I am testing your intuition about what the result of that sign test might be.

You might remember that the example from lecture, with the two drugs, also had 12 observations. There, 9 observations were below zero and 3 were above, and this was not quite significant. On this basis it is reasonable to guess that an 11–1 split is unbalanced enough to be significant:

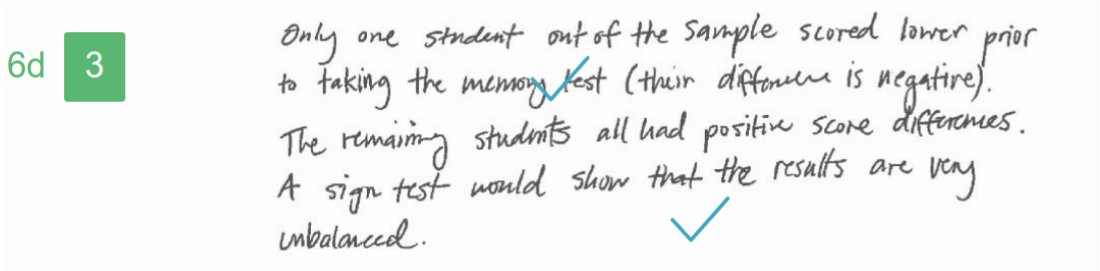
```
sign_test(chess, difference, 0)
```

```
## $above_below
## below above
##     1    11
##
## $p_values
##   alternative     p_value
## 1       lower 0.999755859
## 2       upper 0.003173828
## 3   two-sided 0.006347656
```

It would even be significant two-sided, but this one is a one-sided "greater" test, so the P-value is actually 0.003.

An actual answer:

6d 3



Asserting that pre- and post-test scores are "obviously different" is not an answer. The point of the question is *how you know this*.

Extra: the t-test, which I said I was perfectly happy with, had a smaller P-value of 0.0004 (extra zero). This, I think, is one of those cases where the t-test uses the data more efficiently than the sign test does, and so comes out with a smaller P-value. Here, the one negative difference is one of the smallest differences in size, so in some sense the evidence for the average difference being positive is stronger than the "1 out of 12" that the sign test uses.

Extra extra: you may run into a test called the "signed-rank test" that ranks the differences in order by absolute size. This would have a smaller P-value than the sign test:

```
wilcox.test(chess$difference, mu=0, alternative="greater")
```

```
## Warning in wilcox.test.default(chess$difference, mu = 0, alternative = "greater"):
cannot compute exact p-value with ties
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  chess$difference
## V = 76, p-value = 0.002082
## alternative hypothesis: true location is greater than 0
```

The R name comes from the name Wilcoxon that is often attached to this test. (There actually was also a statistician called Willcox, so the name of the R function is rather confusing.) In general though, I don't like this test as much as the sign test because it comes with an assumption of a *symmetric* distribution: that is to say, a difference above the hypothesized median is worth the same as a difference *of the same size* below. But, to my mind, if you don't trust normality, you probably don't trust symmetry much either. The sign test doesn't make this kind of assumption: what matters is above or below, and that's it.

Question 7 (10 marks)

Some people have the ability to remember accurately vast amounts of information about themselves, without using mnemonic tricks or extra practice. This ability is called "Highly Superior Autobiographical Memory" or HSAM. A study recruited adults with diagnosed HSAM and also control individuals of similar age without HSAM. The aim of the study was to determine what makes HSAM work. All the subjects in the study were given a large number of cognitive and behavioral tests. Some of the results for a visual memory test are shown in Figure 15. A higher score is better.

(a) (3 marks) What code would run a suitable Mood's median test on these data? (You may assume `library(smmr)` has already been run.)

My answer: The data frame is called `hsam` (from Figure 15). `memory` is groups and `test_score` is quantitative, thus:

```
median_test(hsam, test_score, memory)
```

```
## $table
##          above
## group     above below
##   control     5    10
##   hsam        8     0
##
## $test
##        what       value
## 1 statistic 9.435897436
## 2        df 1.000000000
## 3   P-value 0.002127789
```

Get the right inputs, with the right names, in the right order.

You can build it yourself (I said you had `smmr`, so there is no need). To do that, get hold of the overall median, a table showing the values in each group above and below, and a chi-squared test on that table. A mark for each of those, but a lot of work for three marks.

(b) (3 marks) The output from your Mood's median test is shown in Figure 17. What do you conclude from this output, in the context of the data?

My answer: The P-value, 0.0021, is less than 0.05, so we reject the null hypothesis in favour of the alternative (one point).

In this case, the null hypothesis is that the two memory groups have equal median test score, and the alternative is that the medians are *different* (remember that our conception of Mood's median test is always two-sided, because it can also be used for comparing more than two groups, like ANOVA, where a directional alternative makes no sense). Thus, we have evidence that the median test scores for the HSAM and control groups are different. Two more points.

You might reasonably claim that this is a memory test and you would expect the HSAM group to do *better* on it (so that this should have been our alternative). You can't conclude this directly from the output, but you can make it work. First you notice that the HSAM results are all *above* the overall median, and the majority of the control group scored *below*, which is pointing in the direction of the HSAM subjects doing better. (That is, we are on the right side.) Then, you can justifiably halve the P-value to 0.0011, and *then* you can conclude that

HSAM subjects are better at this visual memory test than control subjects.

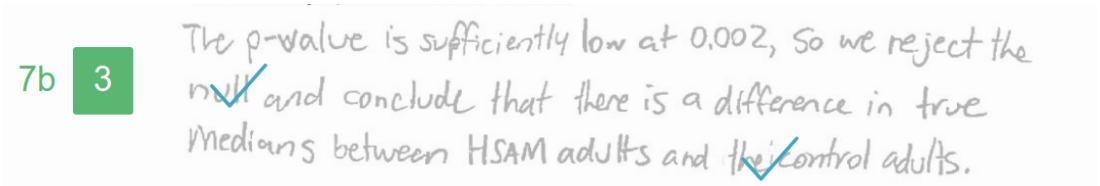
Full marks for a properly-reasoned conclusion that the two group medians are different, or for a properly-reasoned conclusion that the median test score is higher for HSAM people (along the lines of the previous paragraph). I also accept something like "the group medians are different, but the table points to the test scores being higher for HSAM people".

Two out of three for asserting that the HSAM median is significantly higher without properly justifying that conclusion.

If you want to talk about the `table` output of the median test you can, but you need to use it to make some kind of overall conclusion, rather than just citing the numbers, like "test scores are typically higher for the HSAM group".

You do know, don't you, that Mood's median test is for *comparing* medians with each other, like the two-sample t? The fact that the overall median is here 4 is just a stepping-stone on the way to doing the test, not part of its conclusion.

An actual answer:



(c) (1 mark) There are 29 observations in the data set, but only 23 in the table in the Mood's median test in Figure 17. What happened to the others?

My answer: The hint is that the observations in Figure 15 are all fairly small whole numbers, and so quite a lot of them might be equal to each other. What `median_test` does if any observations are equal to the overall median is to throw them away and use only the remaining observations for the test. That must be what happened here.

"They were discarded because they were equal to the overall median" is what I want to see for the point.

Is that actually true? Well, what is the overall median?

```
hsam %>% summarize(m=median(test_score)) %>% pull(m) -> med
med
```

```
## [1] 4
```

It's 4. (I put that in Figure 16 to give you a hint.) Then, how many observations from each group are equal to it?

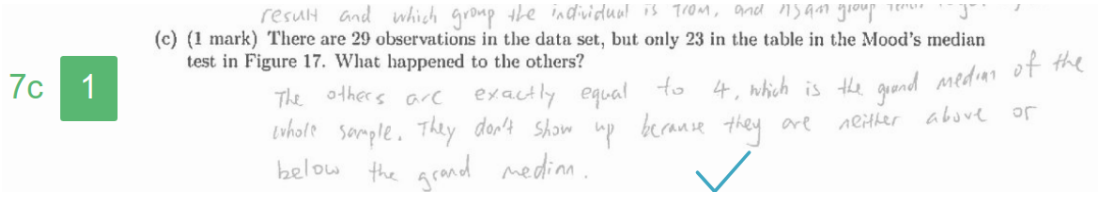
```
hsam %>% count(test_score==med, memory)
```

```
## # A tibble: 4 x 3
##   `test_score == med` memory      n
##   <lgl>               <chr>   <int>
## 1 FALSE               control    15
## 2 FALSE               hsam        8
## 3 TRUE                control     4
## 4 TRUE                hsam        2
```

The last two rows of this tell us that 4 observations in the control group and 2 in the `hsam` group, totalling 6, are equal to the overall median 4. Check. You can eyeball this yourself since the actual data set is in Figure 15.

More people got this than I expected. Some people seemed to work it out from the fact that the table contains "above" and "below", and the data values equal to the overall median were neither of those. Which is exactly right.

An actual answer:



It has nothing to do with outliers. The point of Mood's median test is that you can keep outliers in and it won't bother the test; for example, a high outlier will count as just one above the overall median, no matter how high it is.

(d) (3 marks) A pair of graphs is shown in Figure 18. Explain briefly what you conclude from this, and thus discuss whether we should have run a Mood's median test or whether some other test would have been better instead. If you think some other test would have been better, give the name of the test you would prefer. (If you would prefer to see some other plot to help you decide, describe what you would like to see and why.)

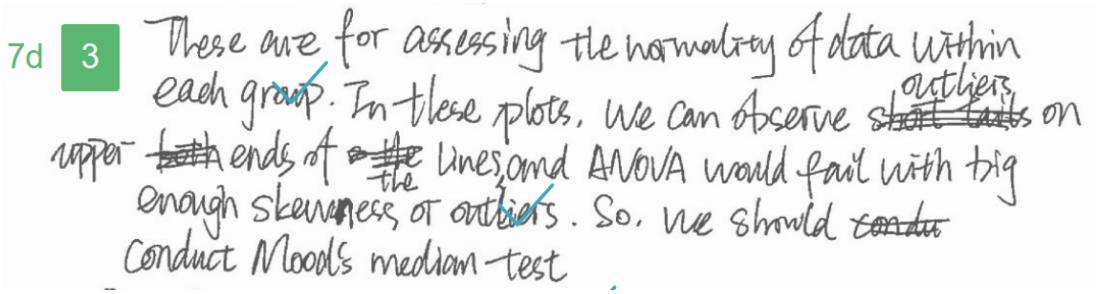
My answer: The competition here is some flavour of two-sample t-test, which you would run if the plots in Figure 18 are *both* acceptably normal. Make a call on that. For the control group, you might say that the highest observation is an outlier, or you might say that there is a little skewness (to the right) because the lowest two observations are a bit too big as well (the lower tail is a bit short and the upper tail is a bit long). For the HSAM observations, it's hard to be sure since there are only 10 of them. You might say that the highest one or two observations are outliers, but they are not extreme.

Whatever you thought of these plots: all you need is to find a problem in one of them, and then you have a problem with whichever two-sample t-test that you were going to run instead. If that's the case, you have a reason for running the Mood's median test that you did. If you found no problems here, then you should run one of the two-sample t-tests (I don't mind which one), because that will make more efficient use of the data.

Grading: one point for saying (or implying that you know) that the plots are assessing the normality of the test scores, one point for making a call of what you see (this is really a free point, since you can reasonably conclude that normality is good enough in both graphs or fails in at least one of them), and one last point for saying what would be a good test (either flavour of two-sample t if you think normality is OK, the Mood's median test that we did if not). You need to mention normality (or non-normality) somewhere, since the point of a normal quantile plot is to assess whether data are approximately normal.

An actual answer:

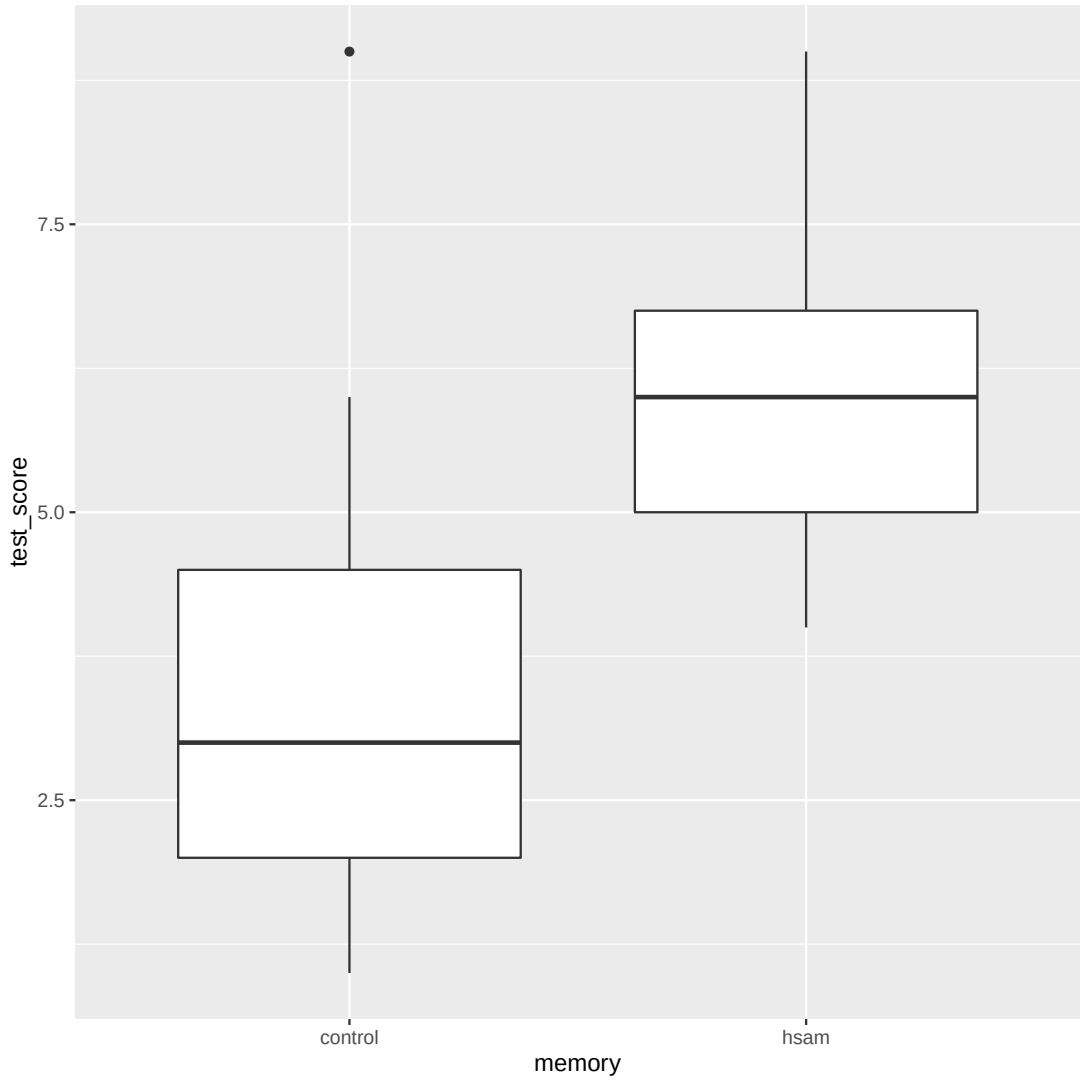
7d 3



The collections of horizontal points are because of the discreteness of the data (the test scores were all smallish whole numbers). These are not really of concern to us when assessing normality: the values cannot *really* be normal exactly, but what is important is whether they are non-normal enough to be a problem.

Extra: there is a small indication here of which kind of two-sample t you should run, if you think that running one is warranted. The *slope* of the line on a normal quantile plot tells you about spread; the HSAM line is less steep than the Control line, suggesting that the HSAM values have less variability, and thus that the Welch t-test is better. I am not penalizing anyone for failing to get this far, since it is a subtle point, and *this* indication would have been much easier to see if I'd given you boxplots:

```
ggplot(hsam, aes(x=memory, y=test_score)) + geom_boxplot()
```



The `control` boxplot is taller (the box) than the `hsam` boxplot, so the control-group distribution of test scores is more spread out. (You might also suggest this as a way of deciding which t-test to run, if you thought we should run one.) On this plot, the control-group outlier really shows up, and the HSAM group appears right-skewed from the whiskers. The Control distribution has, from here, a more or less symmetric distribution apart from the outlier. (If the upper whisker had been longer, that would also have supported skewness, since then the outlier would have been in the same direction as the long tail.) But the normal quantile plot is really a better (more detailed) way to assess normality.

I'm actually suspecting that the t-test is less problematic than it looks, for a few reasons: (i) the control group is not tiny (19 observations), so the outlier may not be so influential; (ii) both

groups appear to be skewed in the same direction, and since the t-test is based on the difference in means, unusually large observations from each group will tend to cancel each other out; (iii) I'm guessing that the result is so significant anyway that it doesn't matter if the distributions are off by a bit (I did this one two-sided to be consistent with the Mood's median test):

```
t.test(test_score ~ memory, data=hsam)
```

```
##
##  Welch Two Sample t-test
##
## data:  test_score by memory
## t = -3.6173, df = 21.556, p-value = 0.001565
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.893598 -1.053771
## sample estimates:
## mean in group control    mean in group hsam
##              3.526316              6.000000
```

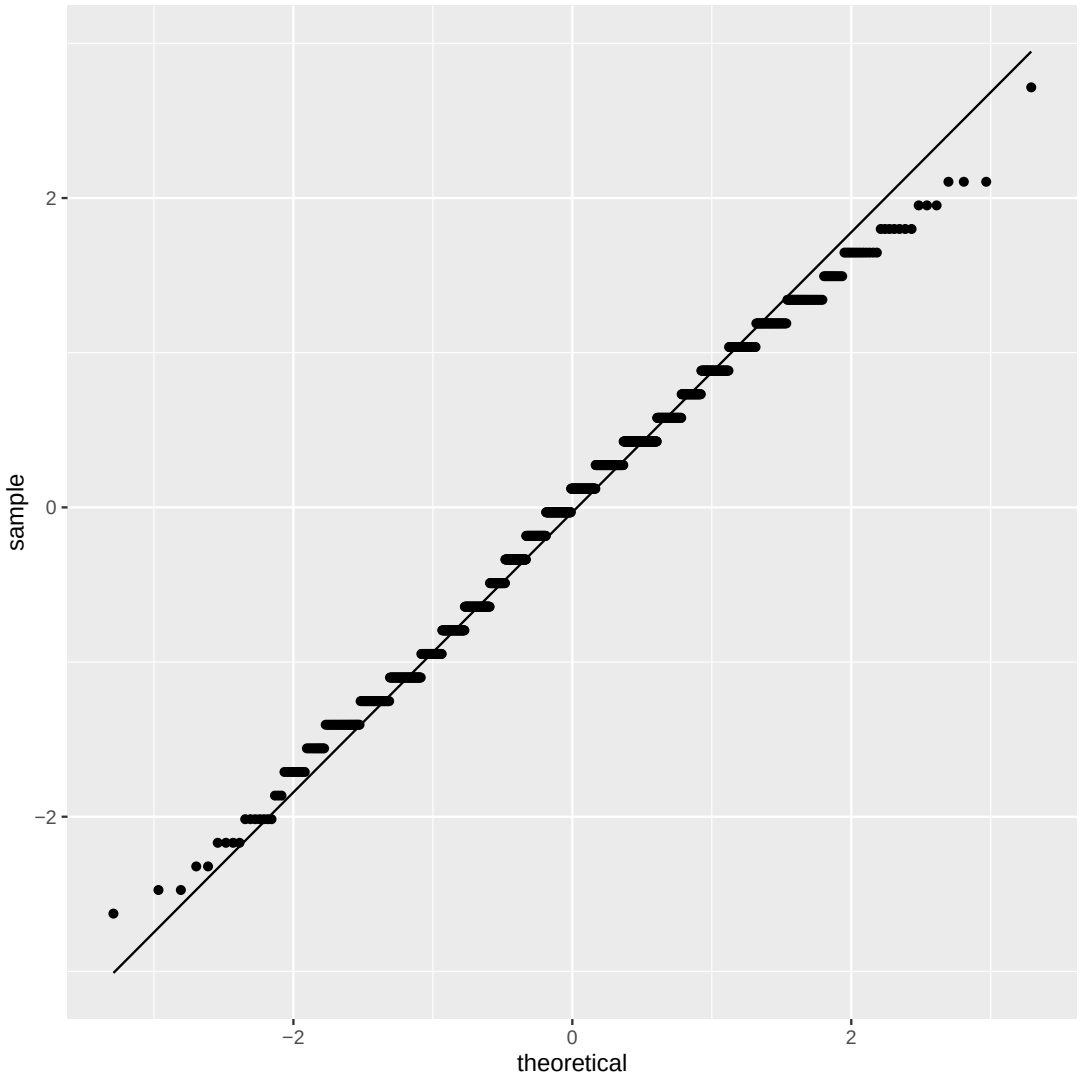
The P-value is a little bit smaller than that of Mood's median test, but, given the foregoing, it's not clear how much we should be trusting it.

A bootstrap would be a way to assess the distribution of the difference in means, since this is the thing that really needs to be approximately normal, and my claim is that this is closer to approximately normal than you might think. Another way to handle this is a "randomization test": under the null hypothesis that the distribution of test scores is the same between the HSAM and control groups, what you do is randomly permute the group labels, and then see how the means of the groups defined by the permuted labels differ. I'm going to define a couple of functions first. The first one creates a new column with the group memberships randomly shuffled (`sample` run like this randomly shuffles whatever is inside it). The second one takes a data frame with a column called `perm_labels` in it (created by the first function), and calculates the mean test score for each group defined by `perm_labels` and returns just the means:

```
make_perm=function(d) {
    d %>% mutate(perm_labels=sample(memory))
}
make_means=function(d) {
    d %>% group_by(perm_labels) %>% summarize(m=mean(test_score)) %>%
        pull(m)
}
```

Now, to business, using these two functions to create a list of resampled and summarized data frames:

```
rerun(1000, make_perm(hsam)) %>%
    map(~make_means(.)) %>%
    map_dbl(~ .[1]-.[2]) -> mean_diffs
ggplot(tibble(mean_diffs), aes(sample=mean_diffs)) + stat_qq() + stat_qq_line()
```



There is only a smallish possible number of sample means, hence the discrete horizontal lines, but the overall pattern is that the normality is actually pretty good.

I should explain my code. It's a similar principle to any simulation (like for example the power by simulation):

- generate 1000 ("a lot") of data frames with the grouping variable `memory` relabelled
- for each of those data frames, find the means of test score by the relabelled groups
- for each of those pairs of means, find the difference between them by taking "the first thing in it minus the second thing in it" and call that `mean_diffs`

- put them into a data frame and make a normal quantile plot of them.

The thing that comes out of `rerun` is an R list, so after we run `make_means` we have a list of pairs of numbers (with a "first thing" and a "second thing" each time). The `map_dbl` takes us from a list of pairs of numbers to a vector of numbers, that I have to make back into a data frame so that I can `ggplot` it.

Question 8 (11 marks)

Farmers know that driving heavy equipment like tractors over the soil, especially if the soil is wet, compresses the soil and makes it more difficult for crops planted in that soil in the future to grow. One way of quantifying this is to measure something called "penetrability", which is a measure of how much resistance plant roots will meet when they try to grow through the soil. On the scale measured, a high penetrability means that plants find it easier to grow.

A study was carried out at a research station. An area of soil was divided into three plots A, B, C. (These are plots of land, not `ggplot` plots). Plot A was driven over by a tractor in wet weather. Plot B was driven over by a tractor in dry weather. Plot C was left as it was. 20 locations were chosen at random within each plot and the soil penetrability measured. Some randomly chosen rows of the data are shown in Figure 19.

(a) (3 marks) Some plots are shown in Figures 20 and 21. Use either or both of these Figures to assess the two major assumptions for analysis of variance (that `aov` would run). In your explanation, make sure to mention which plot you are drawing each conclusion from.

My answer: The two assumptions are that data should be:

- approximately normally distributed within each group
- each group should have about the same spread.

These, especially the first one, should be considered in the light of the sample sizes (here, 20 observations in each group).

I gave you boxplots and normal quantile plots so that you could use whichever one you find easier to interpret. A properly drawn conclusion for each of the two points above drawn from *one* of the graphs is enough.

To business:

For the normality, if you look at the boxplot, you're looking for approximately symmetric with no problematic outliers. We have moderately large samples of 20 in each group, so we can afford to be off from the normality a bit. Plot A I would say is OK, but plot B has an upper outlier (and/or right skewness, because the upper whisker is also longer than the lower one). Plot C is a real problem because it has two outliers a lot higher than the bulk of the data.

If you look at the normal quantile plots, much the same picture emerges about normality: plot B has one outlier and plot C has two, both at the top. You could also read B (and maybe even C) as being curved and therefore right-skewed.

Since the ANOVA requires all three groups to be approximately normal, it's enough to find *one* problem, such as the outliers in plot C.

Now we turn to the equal spreads assumption. This is easier to assess from the boxplots, which is why I included them as well as the normal quantile plots. Look at the heights of the boxes on the boxplots: I would say B and C are similar, but A seems definitely smaller, so that the equal spreads assumption is definitely shaky to my mind. I really don't think you can say the spreads are equal. (This is one of those where I'm not going to let you say *anything*.) Having said that, if you note that A has smaller spread but not enough to damage the equal spreads assumption, *that* I am OK with. (I think you have to say that A having a smaller spread is potentially a problem, but after that you can do what you like.)

Question 8 continues… This page: 3 marks.

If you want to assess this from the normal quantile plots, you can, but it's a bit harder. I would look at the slopes of the lines; B and C are again about the same, but A seems less steep, indicating a smaller spread. Alternatively, you can say that the data values in A are bunched up at the bottom of the box (small spread), while B goes to about halfway up the box and C goes from about halfway up the box to the top. B and C fill more of their boxes, so they have larger spread.

Again, ANOVA requires all three of the spreads to be sufficiently close to equal, so all you need to do is to find one that's different from the others, such as plot A's.)

Feel free to disagree with me on either of these; if you can support your point of view reasonably well, I'm happy with it. (For example, you can state that the outliers are not severe enough, given the largish samples, to invalidate the normality. I'm not sure I agree, but it's a reasonable argument. Also, you can say the group spreads are sufficiently close to equal if you also say that A is a bit less, but you're not worried about it.)

One point for naming the two assumptions, and one point each for using a suitable graph in a suitable way for assessing each assumption. (If you manage to name something that isn't an assumption, you can still get some credit for assessing it with the graphs, unless you name something that is easier to assess, in which case maybe not.) I broke my own rule and gave 1.5 if you had (for example) a complete discussion of normality but no discussion of equal spreads at all.

Make sure you have clear in your mind the difference between a *hypothesis*, like "all the means are equal", something where you want to use the data to see whether it's true or not, and an *assumption*, like "the data within each group are approximately normal", something that needs to be true in order for a test of a hypothesis to be trustworthy. Thus, telling me how the means or medians compare is not what I wanted here. (I might have asked about this, to get your intuition going about what to expect in the analysis you do below, but I didn't: I would have asked something like "would you expect your preferred analysis in (b) to give a significant result? Explain briefly".

(b) (2 marks) Figures 22, 23 and 24 show three possible analyses of these data. Which one of these analyses do you think is the most appropriate? Explain briefly. Your answer should contain a Figure number.

My answer: First decide whether you think the normality assumption is OK. If not, then you should use Mood's median test, Figure 24. If you think it's OK, decide whether you think the equal spreads assumption is OK. If it is, go with the regular ANOVA, Figure 22. If not, go with the Welch ANOVA, Figure 23.

For example, I think the normality fails, so I would say "I choose the Mood's median test in Figure 24 because I think that plots B and C are not sufficiently close to normally distributed". Base your answer on what you concluded in the previous part (I will check for consistency). If you choose the right Figure for a good reason *given what you said in (a)*, I am happy here, even if I completely disagree with your (a). (If your answer to (a) didn't say much, I tried to work out what you seemed to understand.)

Extra: note that if you think normality fails, you actually don't need to worry about equal spreads, because Mood's median test will handle equal or unequal spreads. All that matters for

Mood's test is whether each observation is above or below the overall median, and the spread doesn't have much effect on that. I suppose, if a group has large spread, its observations are more likely to be close to 50–50 above and below the overall median whatever that is, which would make it harder to reject a null of equal medians, but unless the spreads are *really* unequal, the effect of this is likely to be small.

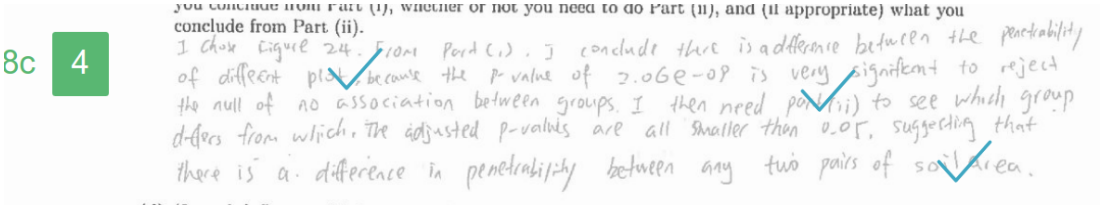
(c) (4 marks) What do you conclude from your chosen analysis, in the context of the data? Note that each analysis has a Part (i) and a Part (ii). Your explanation should include a discussion of what you conclude from Part (i), whether or not you need to do Part (ii), and (if appropriate) what you conclude from Part (ii).

My answer: The conclusions are actually exactly the same whichever of the three analyses you prefer:

- Part (i) has a small P-value, so that you can conclude that the the means/medians (as appropriate) are not all the same. One point.
- Because Part (i) was significant, we look at Part (ii) to decide which of the plots differs from which. One more point.
- Part (ii) shows that all of the plots differ from each other in terms of penetrability. Two more points. (By the usual standards of interpreting the followup test, this is pretty easy this time.)

I was after the results of the hypothesis tests here (because these show whether any differences are real or reproducible). There might be a point for something like a discussion of the number of data values above and below the median, which *suggests* that all the treatments will be different, but doesn't prove it (as the P-values do).

An actual answer:



(d) (2 marks) Do you think your conclusions would make sense to farmers? Explain briefly.

My answer: Go back to the beginning of the question. Plot A was driven over in wet weather, so it should be the most compressed, so the penetrability should be smallest. Plot B was driven over in dry weather, so the penetrability should be a bit larger, while Plot C was not driven over at all, so its penetrability should be largest of all. This is exactly how the data came out (see the boxplot, Figure 20), and not only that, but all the differences are significant, so this points to being something that is always true, not just a happenstance of this data set.

For the two points, I am looking for (i) that the sample means are in the order the farmers

would predict, (ii) that these are significantly different, indicating that the *population* means are in this order too.

Use this page if you need more space to write your answers. Be sure to label any answers here with the question and part that they belong to.