University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAC32 (K. Butler), Midterm Exam
October 13, 2018

READ THE BOX BELOW, AND FOLLOW THE INSTRUCTIONS IN IT.

Aids allowed:
- My lecture slides
- Any notes that you have taken in this course
- Your assignments and feedback on them
- My assignment solutions
- The course R text
- Non-programmable, non-communicating calculator

Past exams are *not* allowed.

Before you begin, complete the signature sheet, but sign it only when the invigilator collects it. The signature sheet shows that you were present at the exam.

This exam has 35 numbered pages of questions. Please check to see that you have all the pages.

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). If you need more space, use the backs of the pages, but be sure to draw the marker's attention to where the rest of the answer may be found.

The maximum marks available for each part of each question are shown next to the question part. In addition, the total marks available for each *page* are shown at the bottom of the page, and also in the table on the next page.

Code for R graphs should be in `ggplot` style, as in lecture. There may be partial credit for "base" graphs.

For any questions below involving R code, you may assume that this code has already been run:

```
library(tidyverse)
```

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.

Last name: _____

First name: _____

Student number: _____

For marker's use only:

Page	Points	Score
1	3	
3	3	
5	3	
10	3	
11	2	
12	2	
13	6	
14	4	
17	2	
20	8	
21	2	
22	3	
24	3	
25	4	
26	3	
27	7	
29	6	
31	3	
33	5	
34	3	
35	3	
Total:	78	

1. Children in elementary school are supposed to master 50 "basic concepts" by the time they reach the end of the second grade. There is a test called the Boehm Test of Basic Concepts that can be given to elementary-school children to see how well they have mastered these concepts.

   For children with hearing problems, it may be harder to master these concepts (for example, because they cannot always hear the teacher). The data in Figure 1, stored in a file `hoh.txt` in your current folder, are the grade levels and Boehm test scores for a sample of hard-of-hearing children.

   (a) (2 marks) Give R code to read the data shown in Figure 1 into a data frame called `test_scores`, using a `tidyverse` function shown in class.

   Solution: Examining the data, we see that the data values are separated by exactly one space, so `read_delim` is the thing:

```
test_scores=read_delim("hoh.txt"," ")

## Parsed with column specification:
## cols(
##   grade = col_character(),
##   boehm = col_double()
## )
```

   Two points for that, one point for `read_delim` with an error (eg. forgetting to specify the delimiter), or using the wrong data frame name, no points for anything else. Note that `read.table` will work, but is not an answer to this question because I did not use it in class. (R veterans beware!)

   The tidyverse function `read_table` *will not work, because the columns are not aligned.* Likewise, neither will `read_tsv`, because there are no tabs anywhere in the file. If there had been, the file would have looked like this:

```
grade boehm
kindergarten 17
kindergarten 20
kindergarten 24
kindergarten 34
kindergarten 34
kindergarten 38
first 23
```

   with apparently more than one space but with the columns aligned depending on the length of the previous thing.

   (b) (1 mark) Give R code to display your data frame. (Code that displays the first ten rows of your data frame in R Studio is acceptable.) Don't overthink this; it's only one point!

   Solution: Just this, the simplest of giveaways:

```
test_scores
```

```
## # A tibble: 24 x 2
##    grade        boehm
##    <chr>        <dbl>
##  1 kindergarten    17
##  2 kindergarten    20
##  3 kindergarten    24
##  4 kindergarten    34
##  5 kindergarten    34
##  6 kindergarten    38
##  7 first           23
##  8 first           25
##  9 first           27
## 10 first           34
## # ... with 14 more rows
```

You don't need to make it any harder than this. (The clue is that this is only one point, so it's got to be something pretty simple.) If you *must*, something like this also gets the point:

```
test_scores %>% head(10)
```

```
## # A tibble: 10 x 2
##    grade        boehm
##    <chr>        <dbl>
##  1 kindergarten    17
##  2 kindergarten    20
##  3 kindergarten    24
##  4 kindergarten    34
##  5 kindergarten    34
##  6 kindergarten    38
##  7 first           23
##  8 first           25
##  9 first           27
## 10 first           34
```

or this:

```
print(test_scores)
```

```
## # A tibble: 24 x 2
##    grade        boehm
##    <chr>        <dbl>
##  1 kindergarten    17
##  2 kindergarten    20
##  3 kindergarten    24
##  4 kindergarten    34
##  5 kindergarten    34
##  6 kindergarten    38
##  7 first           23
##  8 first           25
##  9 first           27
## 10 first           34
## # ... with 14 more rows
```

or this:

```
View(test_scores)
```

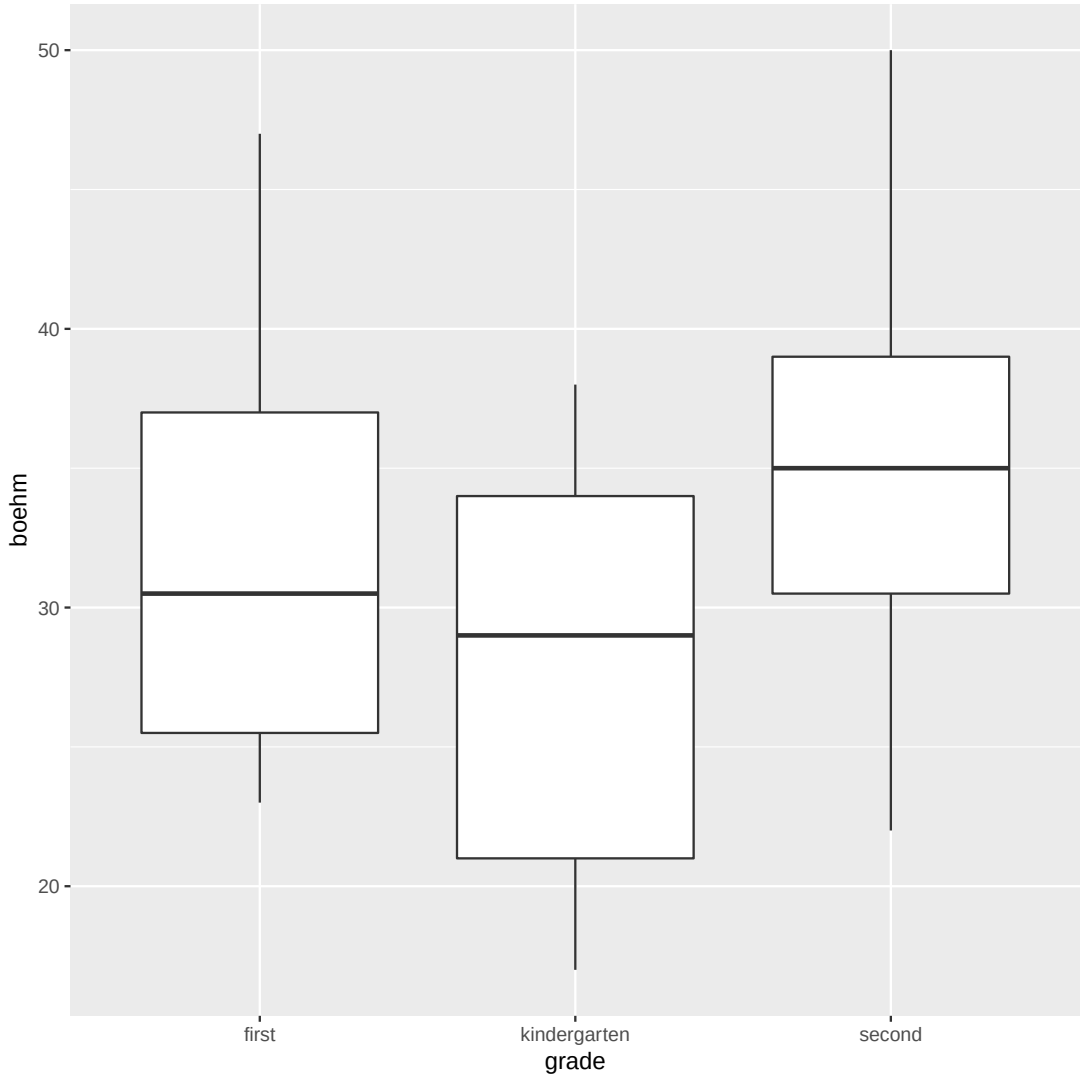
but you're giving yourself extra work for no gain. `View` *must* have a capital V if you want a point for it.

If you mistakenly used a different data frame name in the previous part, use here the same name as you used there. You need to be consistent.

(c) (3 marks) Give R code to make a suitable graph of these data.

Solution: One quantitative and one categorical variable, so a boxplot is the simplest way to go:

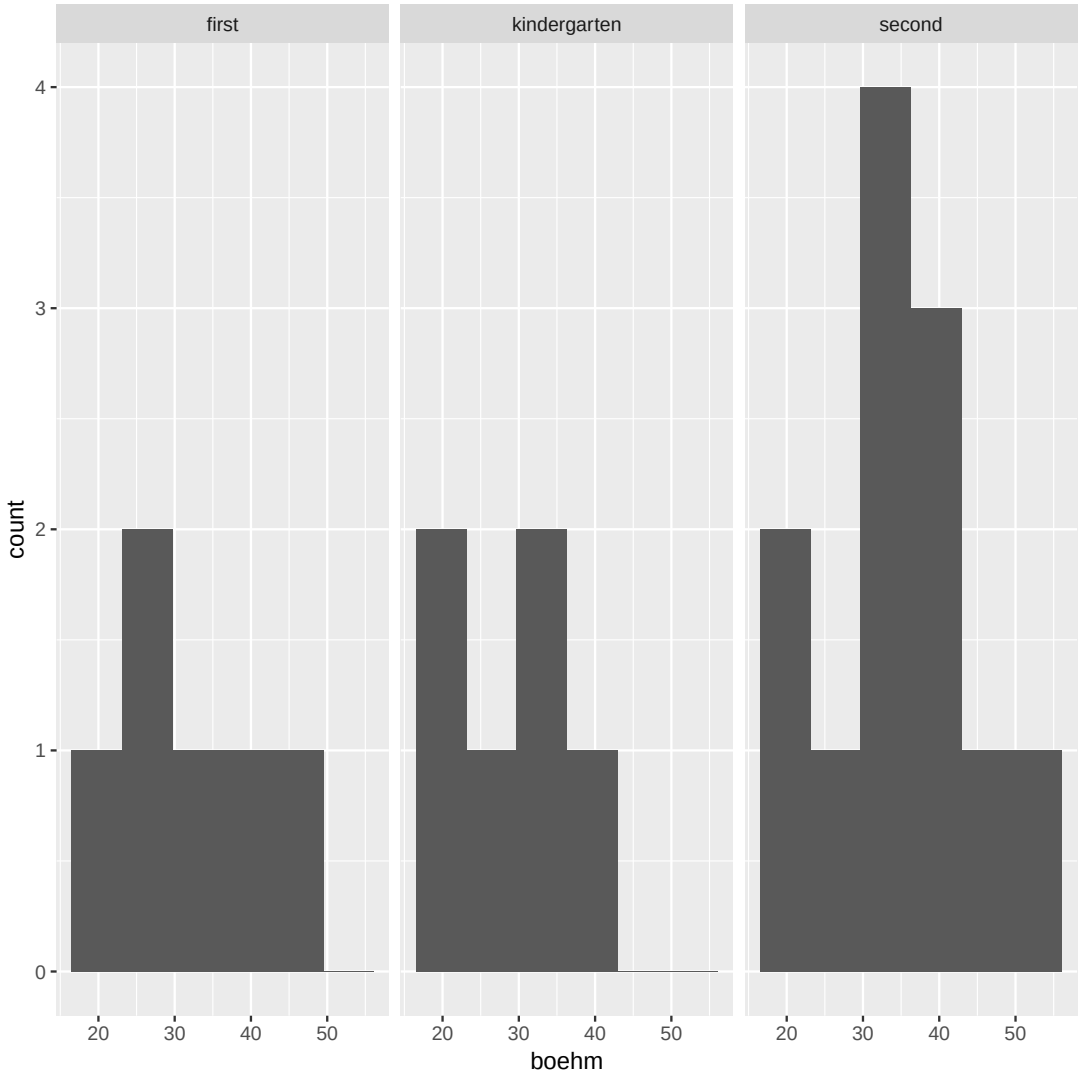
```
ggplot(test_scores, aes(x=grade, y=boehm))+geom_boxplot()
```



Three points for this, two points for a boxplot with errors (eg. getting the `x=` and `y=` the wrong way around), one point for any other plot correctly done (but see below), nothing for any other plot with errors.

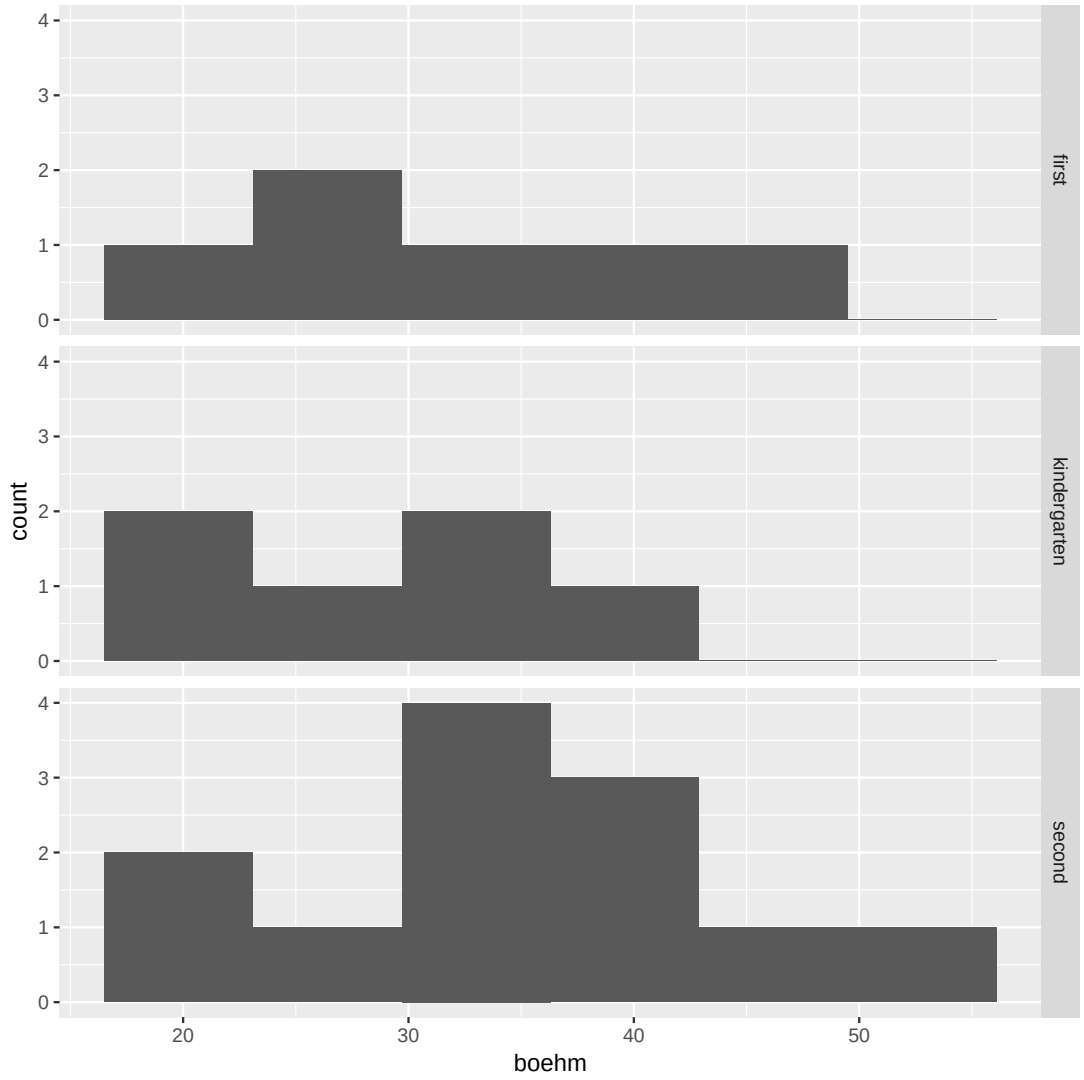
The only other plot that would be acceptable here is faceted histograms, which would go this way:

```
ggplot(test_scores, aes(x=boehm))+
    geom_histogram(bins=6)+facet_wrap(~grade)
```



or this way:

```
ggplot(test_scores, aes(x=boehm))+
    geom_histogram(bins=6)+facet_grid(grade~.)
```



If you go with histograms, failing to supply a number of bins is here an error and will cost you a point (I don't mind how many bins you use). I would prefer the histograms to be one above another for ease of comparison, which you can accomplish with this `facet_grid` or by using `facet_wrap` with `ncol=1`, but I will also accept the histograms in a row (including `facet_grid` with `grade` as the "x").

Three points if your code is one of the two variations of facetted histograms above, with any number of bins other than the default (30 bins, way too many).

Missing out the dot in `facet_grid` will cost you one point (since that is an actual error).

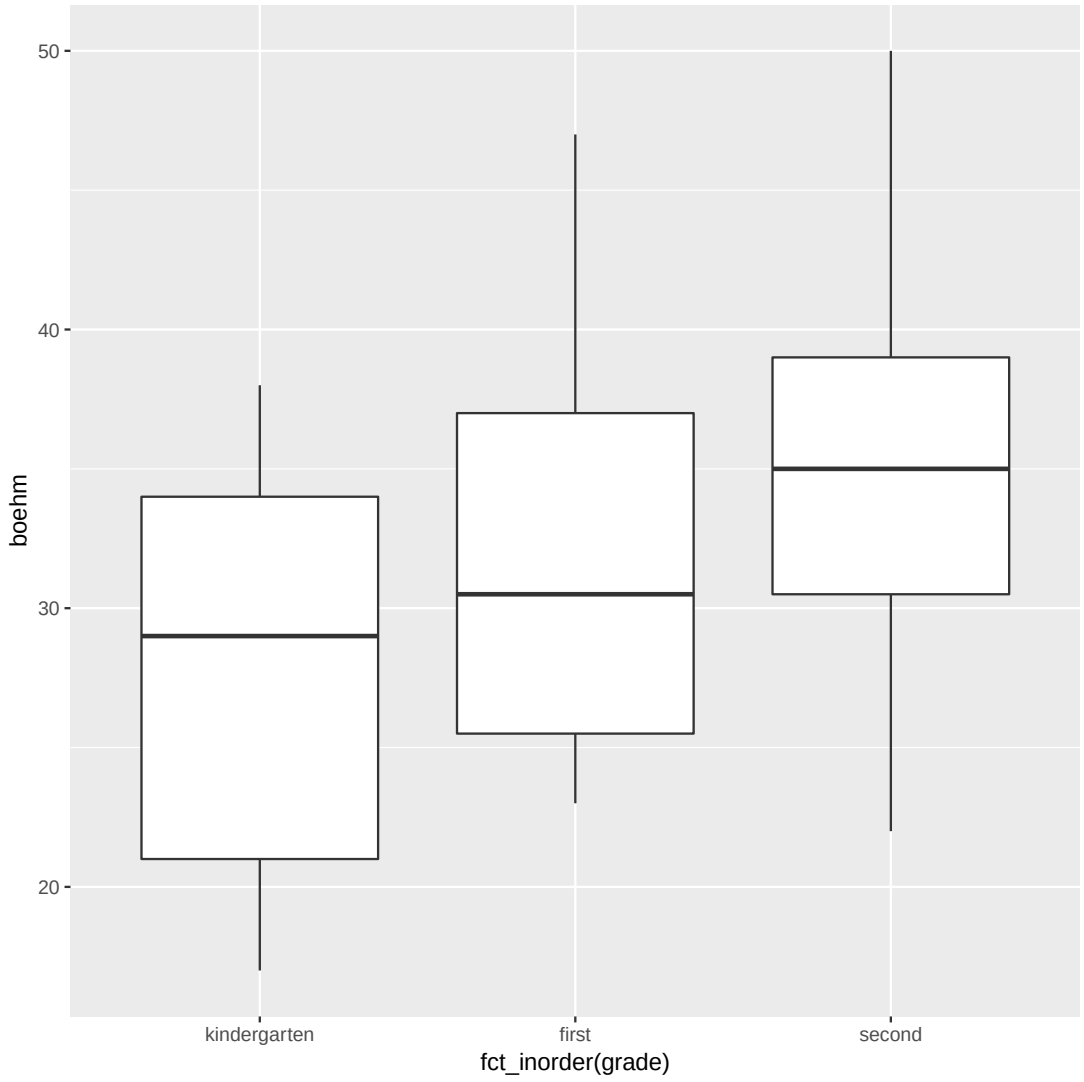
Making a boxplot is way easier, though.

(d) (3 marks) On the plot you made in the previous part, how will the grade levels display? What would be a better way for them to display? How could you modify your code of the previous part to create this better display?

Solution: In alphabetical order (one point for this). This is also true for facetted histograms (check the facet labels). A better display would be youngest to oldest in grade order, which would be kindergarten, first, second (one point, either for giving the grades in the right order or describing what the order should be, such as "in grade order, youngest to oldest" or "the same order that the grades appear in the data file"). This is the `fct_inorder` thing from the assignment 1 solutions, since the rows of data *are* in that order in Figure 1. (I put the rows in that order on purpose.) The best answer here is "replace `x=grade` with `x=fct_inorder(grade)`"; you don't need to write the whole code out again (the final one point).

To show you that works:

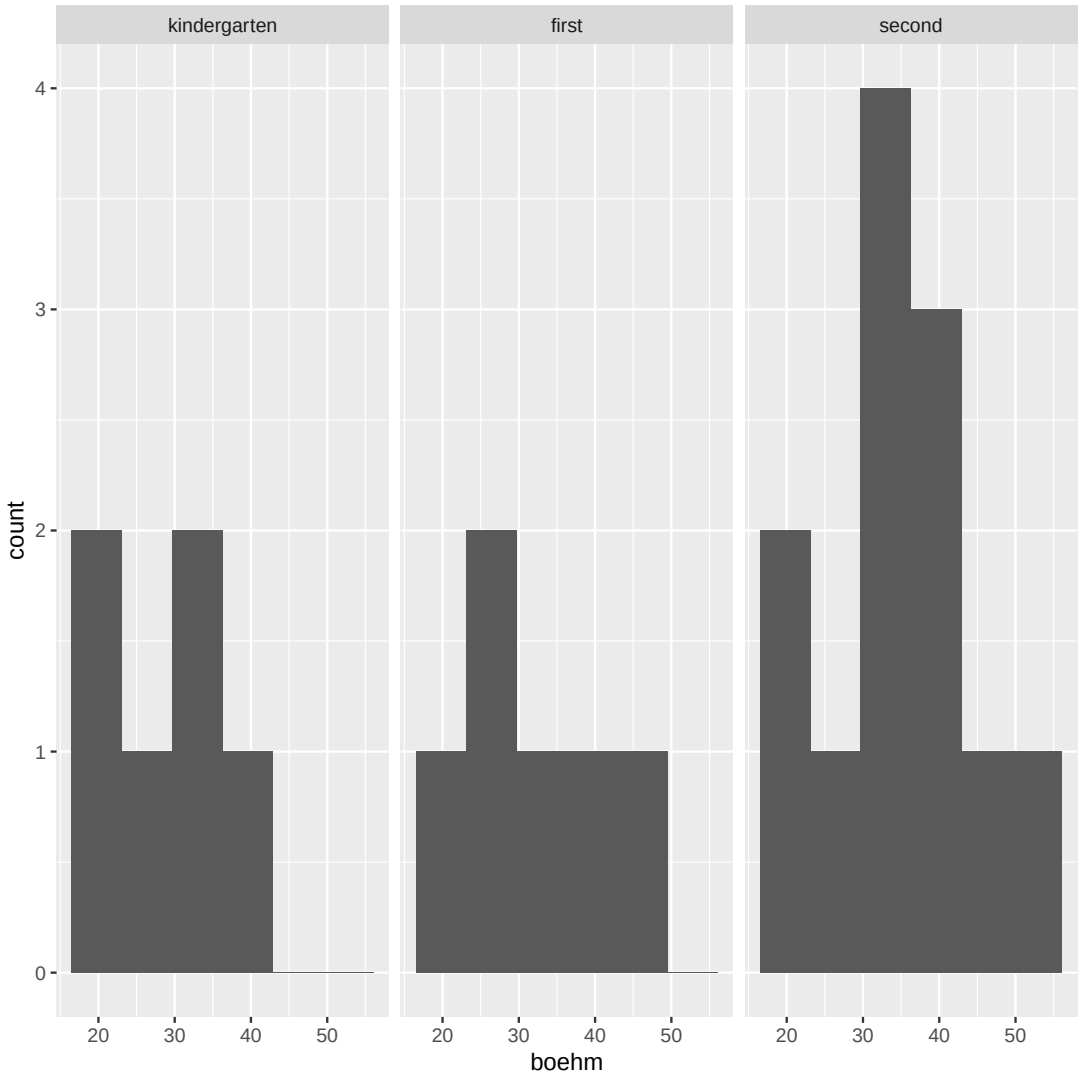
```
ggplot(test_scores, aes(x=fct_inorder(grade), y=boehm))+geom_boxplot()
```



This nicely shows the average scores increasing by grade level.

Alternatively, something like:

```
ggplot(test_scores, aes(x=boehm))+
    geom_histogram(bins=6)+facet_wrap(~fct_inorder(grade))
```

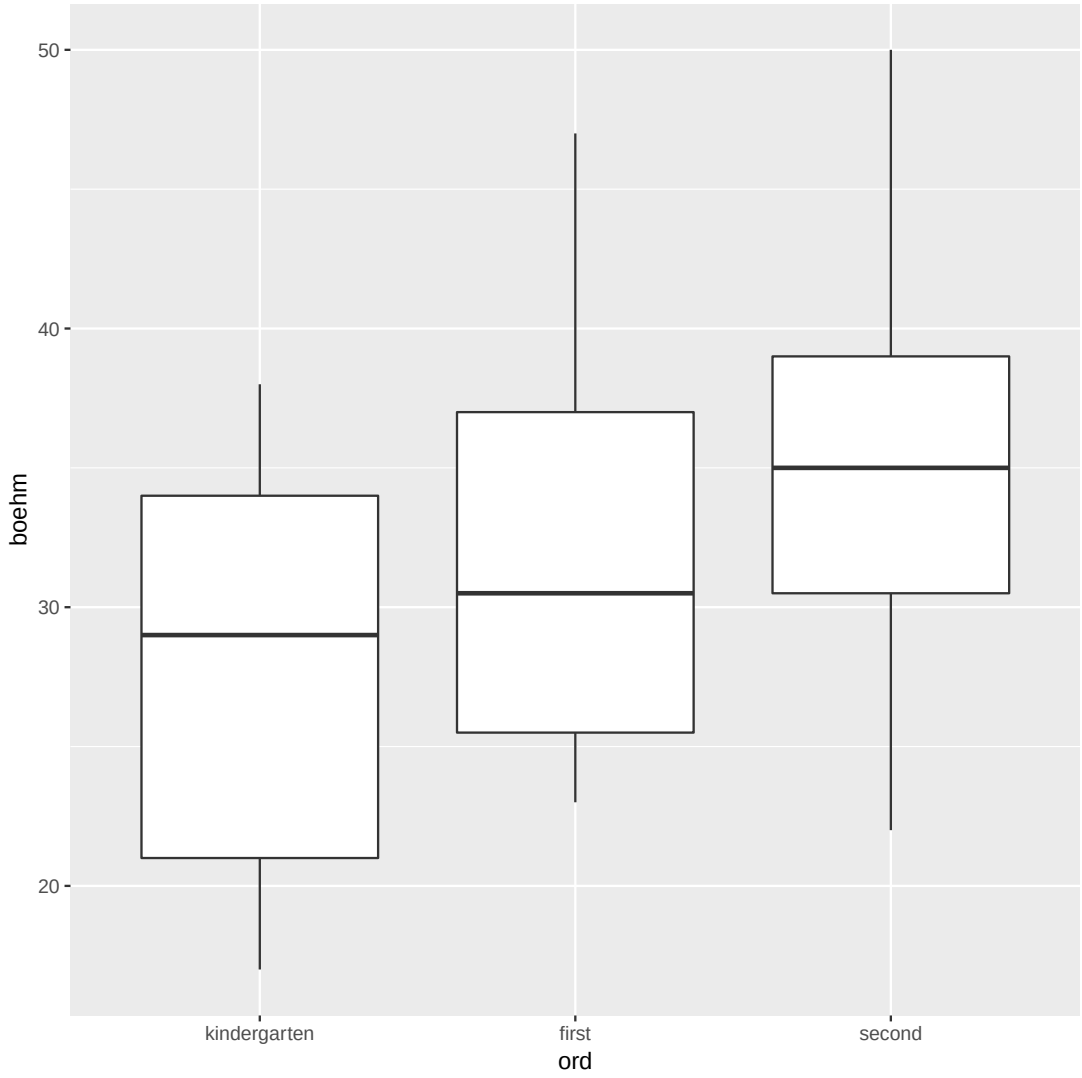


This one is hard to compare horizontally since there were more students in grade 2. But if you have this kind of code, I'm fine with it.

If you use `fct_inorder` when you *first* create your plot (in the previous part), and at some point you show some understanding of why that is a good idea (for example, if you say that failing to do so will put the boxes in alphabetical order), then you can get full marks for these two parts. If you use `fct_inorder` in the previous part *without comment*, then we will probably ignore that you have done so. In that case, you can rescue yourself here by saying something like "the grade levels will display in the previous part in the order that they appear in the data. I did that so that they did not appear in alphabetical order, which makes no sense." This shows the proper understanding of what I was trying to get at.

It's also good to define the ordered factor first like this, and then use the (named) ordered factor in your plot:

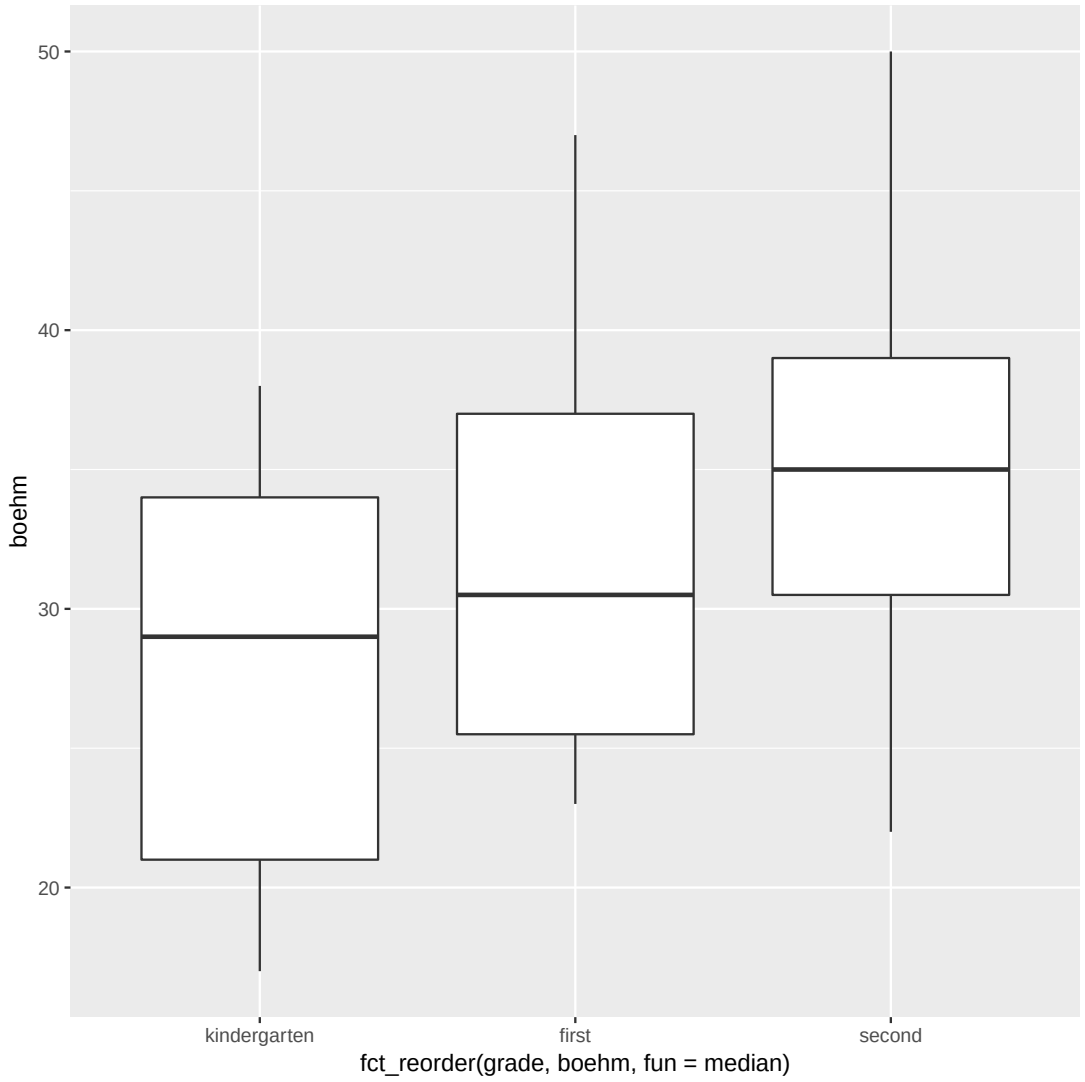
```
test_scores %>%
    mutate(ord=fct_inorder(grade)) -> ts2
ggplot(ts2, aes(x=ord, y=boehm))+geom_boxplot()
```



This is actually the best answer to the question, but for this exam, the simpler answer with the "inorder" stuff on the axis is fine.

Another answer I would support is something like "the graph will look better if the grades are in order of Boehm scores", and then you can use `fct_reorder` thus:

```
ggplot(test_scores, aes(x=fct_reorder(grade, boehm, fun=median), y=boehm))+
    geom_boxplot()
```



If you use `fct_reorder`, I would like to have some sense that you are doing so for a good reason, one that fits with your answer to "what would be a better display". If you have done this, I would accept this kind of answer, even though I don't think it's as good an approach as putting the grades in grade order. I know it comes out the same, but I care about the *method* over the *result*. However, if it looks as if you have written `fct_reorder` as a guess, without any idea of why you would prefer it over `fct_inorder`, don't expect so much sympathy.

(e) (3 marks) Give R code to calculate the mean Boehm test score for each grade level, using a technique that we have seen in class.

Solution: What I am after is this:

```
test_scores %>% group_by(grade) %>%
    summarize(xbar=mean(boehm))
```

```
## # A tibble: 3 x 2
##   grade         xbar
##   <chr>        <dbl>
## 1 first         32.3
## 2 kindergarten  27.8
## 3 second        35.2
```

Choose your own name for the column containing the group means. Optionally, do the `fct_inorder` thing again (full marks with or without it):

```
test_scores %>% group_by(fct_inorder(grade)) %>%
    summarize(xbar=mean(boehm))
```

```
## # A tibble: 3 x 2
##   `fct_inorder(grade)`  xbar
##   <fct>                <dbl>
## 1 kindergarten          27.8
## 2 first                 32.3
## 3 second                35.2
```

Three points for this, two for something containing both `group_by` and `summarize` but with an error, one point for something with only one of these two.

Rather unwillingly, I'll give you two points for this:

```
aggregate(boehm~grade, data=test_scores, mean)
```

```
##          grade    boehm
## 1        first 32.33333
## 2 kindergarten 27.83333
## 3       second 35.25000
```

since it works, but this is not as we have done it in class.

If you try this, expecting to get means for the groups, it doesn't work:

```
t.test(boehm~grade, data=test_scores)
```

```
## Error in t.test.formula(boehm ~ grade, data = test_scores):  grouping factor
must have exactly 2 levels
```

because there are more than two grade levels. It turns out that `aov` doesn't help you if you try that:

```
boehm.1=aov(boehm~grade, data=test_scores)
summary(boehm.1)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## grade        2  220.9  110.46   1.399  0.269
## Residuals   21 1658.4   78.97
```

because this doesn't give the group means either. This kind of does:

```
boehm.2=lm(boehm~grade, data=test_scores)
summary(boehm.2)
```

```
##
## Call:
## lm(formula = boehm ~ grade, data = test_scores)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.250  -7.458  -0.250   6.167  14.750
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         32.333      3.628   8.912  1.4e-08 ***
## gradekindergarten   -4.500      5.131  -0.877    0.390
## gradesecond          2.917      4.443   0.656    0.519
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.887 on 21 degrees of freedom
## Multiple R-squared:  0.1176,Adjusted R-squared:  0.03351
## F-statistic: 1.399 on 2 and 21 DF,  p-value: 0.269
```

When we do regression with categorical variables later, we find that the intercept is the mean of the first group (the one not named, so `first`), and the "slopes" for the other groups are the *differences in mean* between the "baseline" first group and the one you're looking at. Thus, for example, the kindergarten mean is $32.333 - 4.5 = 27.833$. But this is a way harder way of answering the question than `group_by` and `summarize`. (I typically have one kind of thing in mind; there may be other ways to answer the question, but figuring out what I had in mind will lead you to the easiest way.)

2. Between 1954 and 2017, 58 people, a mixture of males and females, successfully swam across Lake Ontario from south to north. The time was recorded, in minutes, that each swimmer took. It typically takes about 1200 minutes (20 hours) for a swimmer to cross Lake Ontario. This is an endurance event.

   It typically takes longer to swim from north to south, so the data used in this question are only for the swims from south to north (for example, from Niagara-on-the-Lake to Leslie Street Spit).

   (a) (2 marks) Boxplots of the swimming times for males and females are shown in Figure 2. What do you notice about the *shapes* of the distributions of times? Comment briefly on the shapes of both distributions.

Solution: The female times have a long upper whisker, so they are right-skewed. The male times have a longer upper whisker *and* upper-end outliers, so they are right-skewed as well.

This is only two marks, so "they are both right-skewed" is a minimal acceptable answer.

Extra: a few people did swim across the lake from north to south, but since it typically takes longer in this direction, I omitted these swimmers from the data set for this question. By doing so, I made the patterns appear more clearly. (It is not generally a good idea to mix up data that you know is going to be different.) Also, some people have swum across the lake more

than once, so the number of distinct swimmers is less than 58. This is something we ignore here.

(b) (2 marks) Numerical summaries of the swimming times for males and females are shown in Figure 3. How do these summaries support your answers to the previous part about distributional shape? Explain briefly.

Solution: For both males and females, the mean is bigger than the median. This makes sense, because the mean will be made larger by the few very large values, while the median will not be.

You can't compare the SD and IQR in the same way, because they are not comparable numbers to begin with. For example, in a normal distribution, the IQR is bigger than the SD (1.35 times bigger, in fact), but a distribution with a lot of outliers could have a much bigger SD compared with the IQR. So comparing SD and IQR is a lot less clear than comparing mean and median.

3. A company sells crackers. The company wants to see what kind of promotion will best improve sales of crackers, so it runs a study. Three promotion types were considered:

   - `sampling`: allowing customers to taste the crackers by giving out free samples
   - `shelf_regular`: additional shelf space in the regular location where the crackers are sold
   - `shelf_display`: Special display shelves at the end of the aisle (in addition to regular shelf space).

   Fifteen stores were used for the study. Five stores were randomly chosen to receive each promotion type. Each store used the same price and advertising for the crackers. The outcome variable was the number of cases of the crackers sold during the study period, `prom_sales` in the data set. The stores varied somewhat in size, so the company also recorded the number of cases of the crackers sold in each store during the previous time period, `prev_sales`. The data set also includes a variable `store` that identifies each store. This variable plays no further part in this question.

   The data are shown in Figure 4.

   (a) (2 marks) Why did I need to use `read_table` to read in the data?

Solution: Because the data values are aligned in columns with more than one space in between. ("Separated by more than one space" is only one point.)

Extra: this will read in the data:

```
crackers=read_table("crackers.txt")
```

```
## Parsed with column specification:
## cols(
##  store = col_double(),
##  promotion = col_character(),
##  prom_sales = col_double(),
##  prev_sales = col_double()
## )
```

```
crackers
```

```
## # A tibble: 15 x 4
##    store promotion     prom_sales prev_sales
##    <dbl> <chr>              <dbl>      <dbl>
##  1     1 sampling              38         21
##  2     2 sampling              39         26
##  3     3 sampling              36         22
##  4     4 sampling              45         28
##  5     5 sampling              33         19
##  6     6 shelf_regular         43         34
##  7     7 shelf_regular         38         26
##  8     8 shelf_regular         38         29
##  9     9 shelf_regular         27         18
## 10    10 shelf_regular         34         25
## 11    11 shelf_display         24         23
## 12    12 shelf_display         32         29
## 13    13 shelf_display         31         30
## 14    14 shelf_display         21         16
## 15    15 shelf_display         28         29
```

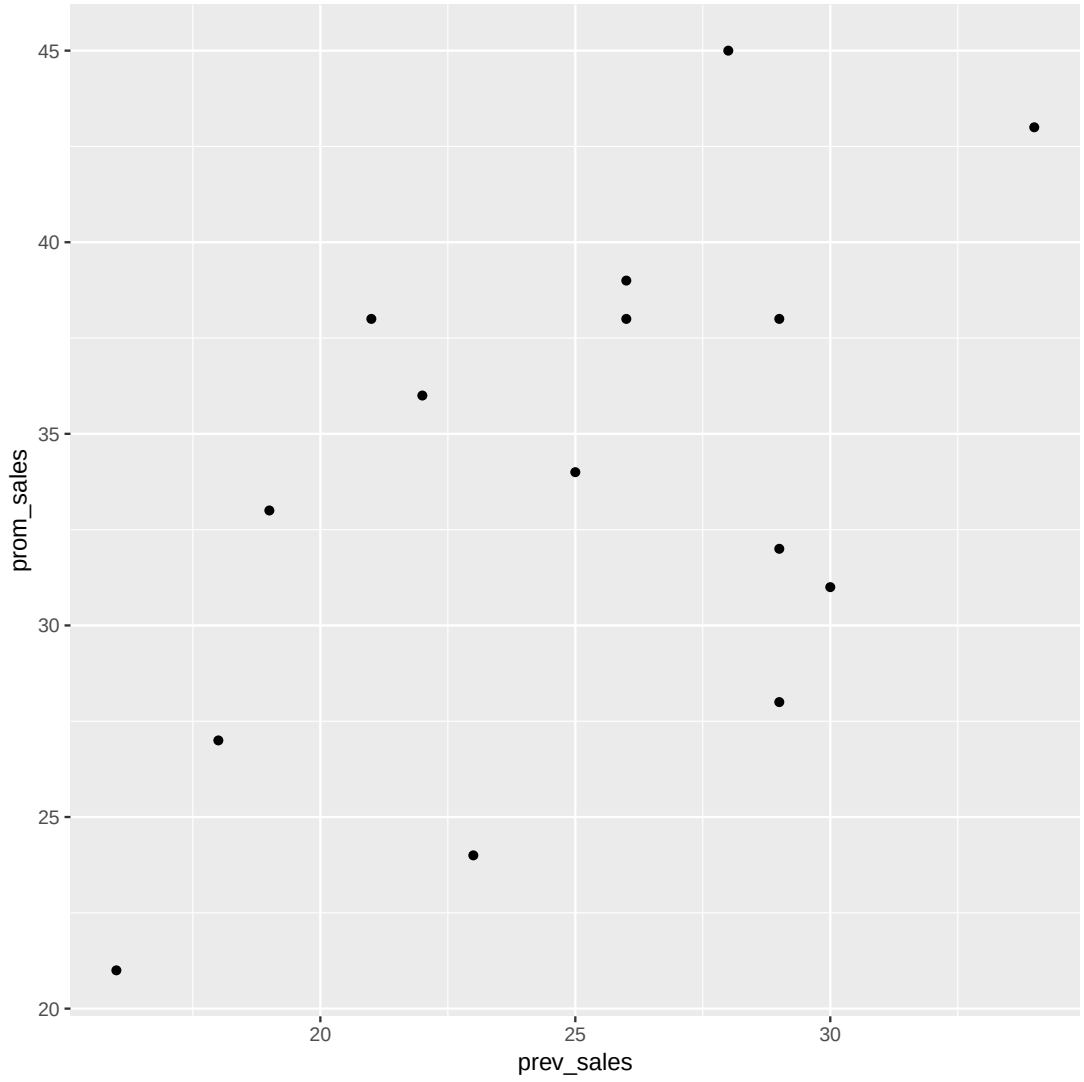
   (b) (4 marks) For this part and the following parts, you may assume that the data have been read into

a data frame called `crackers`.

What would be a good graph to show the two columns of sales (and not the type of promotion)? Give R code to draw this graph.

Solution: Two quantitative variables, so a scatterplot (one point). Sales in the promotion period is the outcome variable, so this goes on the y-axis:

```
ggplot(crackers, aes(x=prev_sales, y=prom_sales))+geom_point()
```



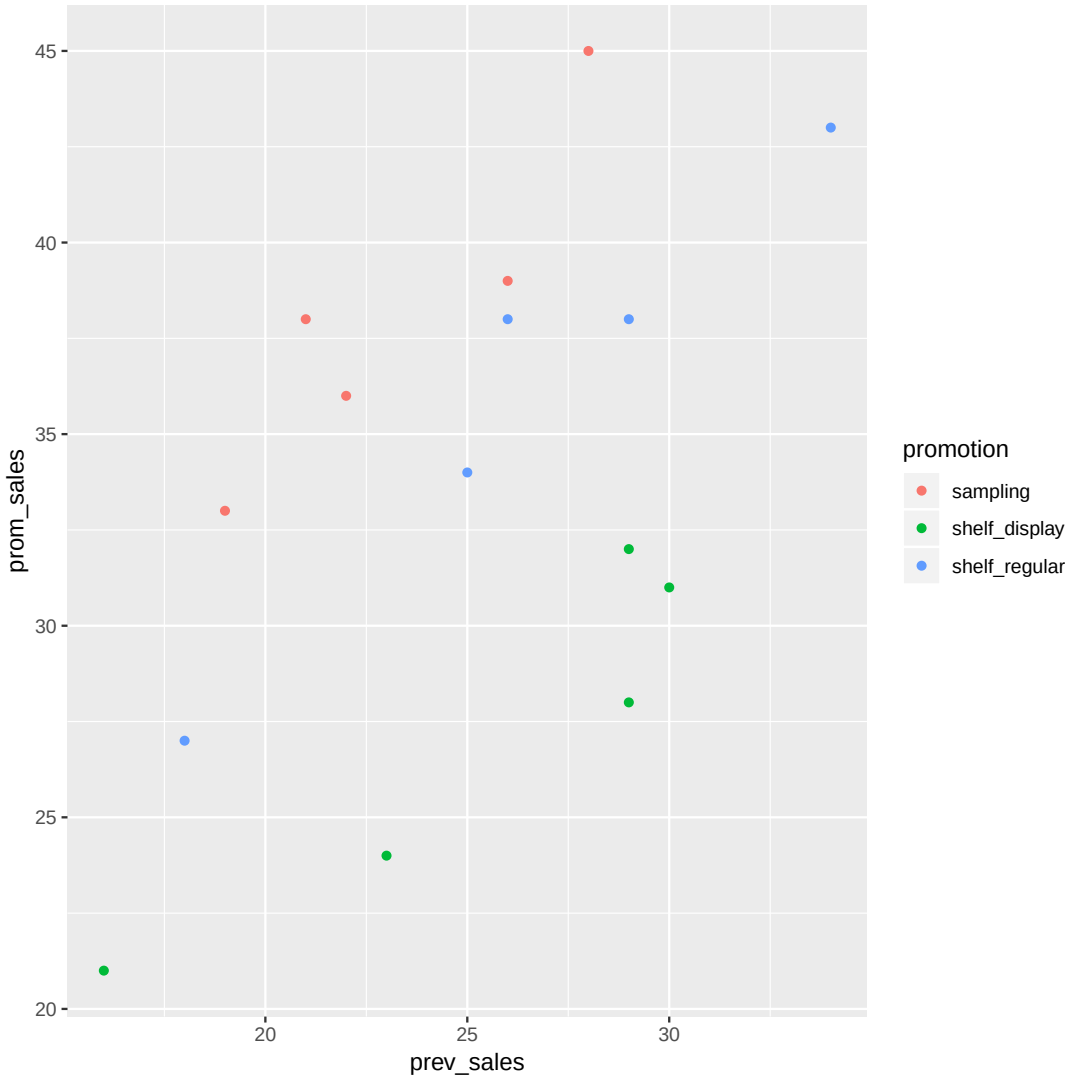
Three points for a correctly-drawn graph, minus one per error, one point if you have something non-trivial correct.

This graph shows the weakest of upward trends.

(c) (4 marks) Describe a graph that would show all three variables (that is, everything except `store`), and give R code to produce it.

Solution: A scatter plot with the points distinguished by promotion type, most easily by making them different colours (one point):

```
ggplot(crackers, aes(x=prev_sales, y=prom_sales, colour=promotion))+
    geom_point()
```

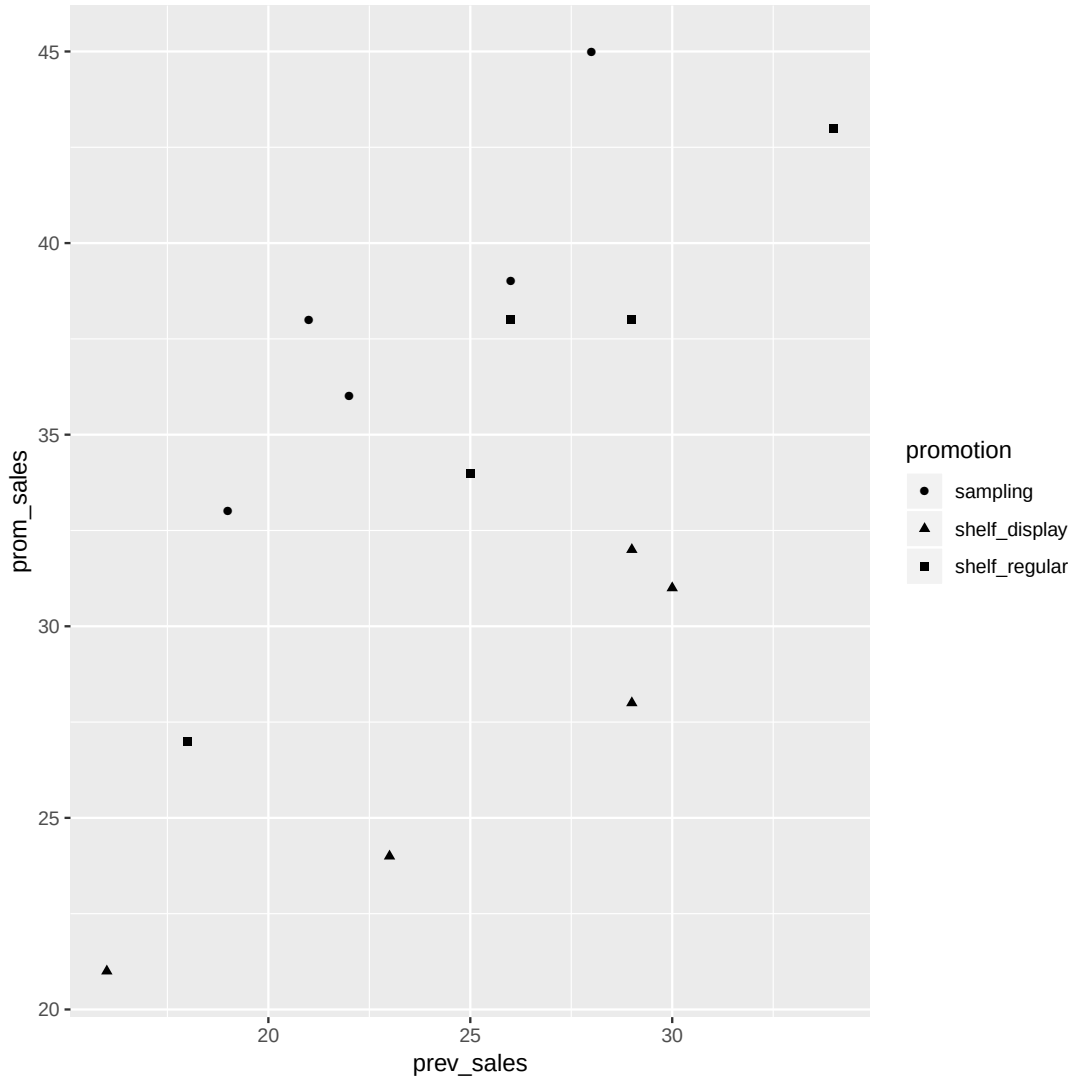


Minus one per error, one if you have something relevant correct.

Looking at each promotion type separately, the upward trends are a bit stronger.

Any way of distinguishing the promotion types is good, for example this:

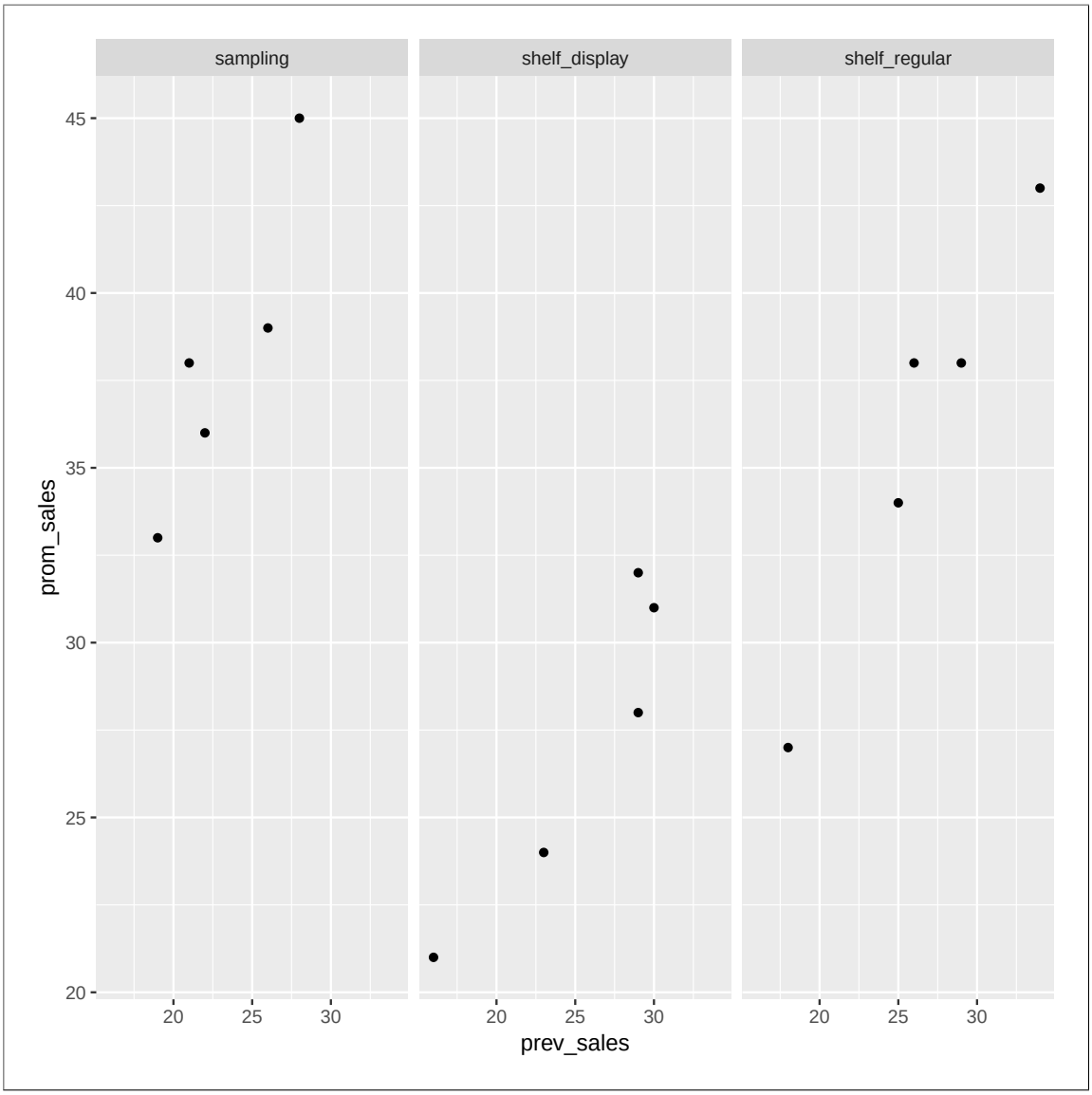
```
ggplot(crackers, aes(x=prev_sales, y=prom_sales, shape=promotion))+
    geom_point()
```



I happen not to like this as much, but it gets the job done, so full marks.

You might instead think of facetting on the categorical variable:

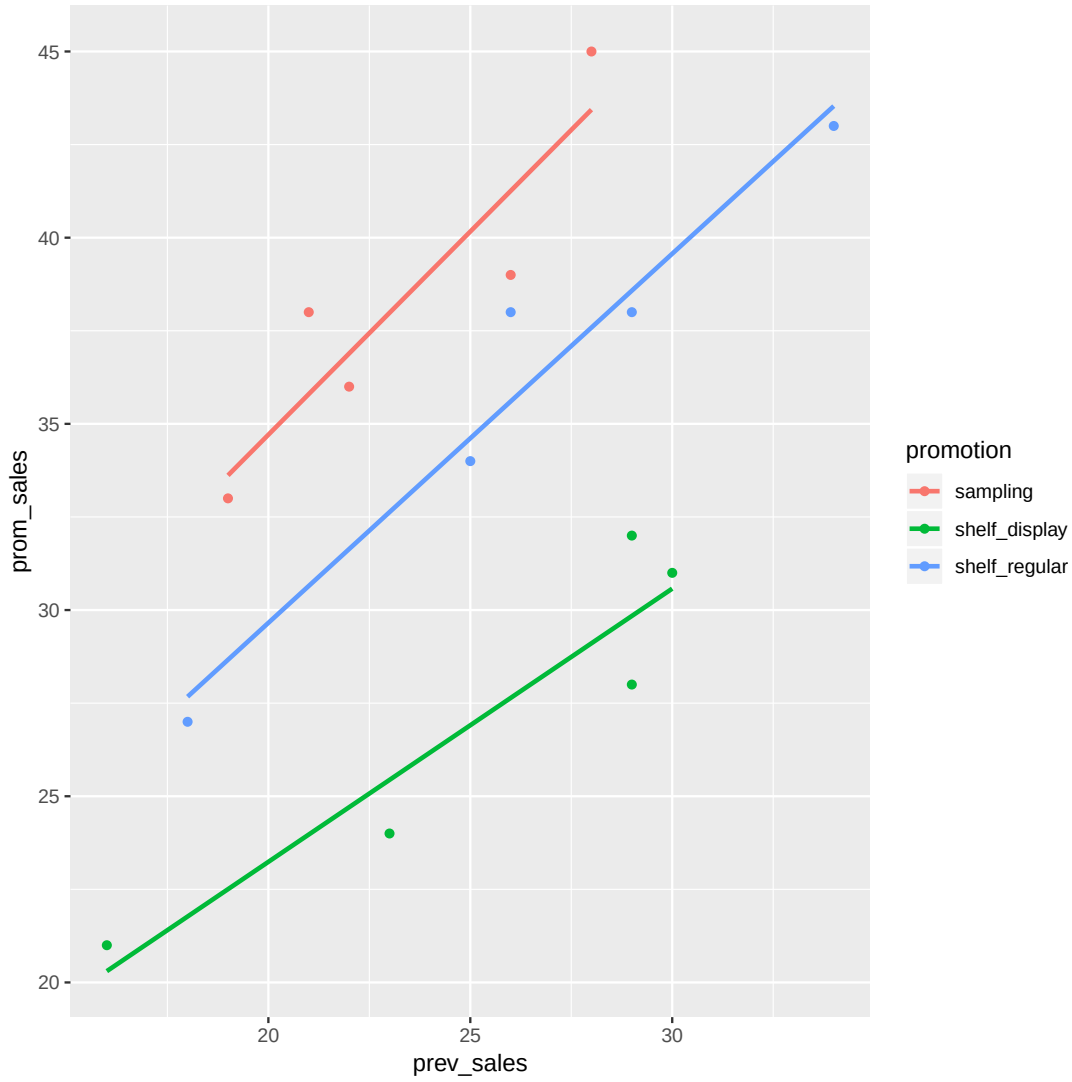
```
ggplot(crackers, aes(x=prev_sales, y=prom_sales))+geom_point()+
    facet_wrap(~promotion)
```



(d) (2 marks) What R code would you *add* to your previous plot to add three regression lines, one for each promotion type?

Solution: This code would do it:

```
ggplot(crackers, aes(x=prev_sales, y=prom_sales, colour=promotion))+
    geom_point()+geom_smooth(method="lm",se=F)
```



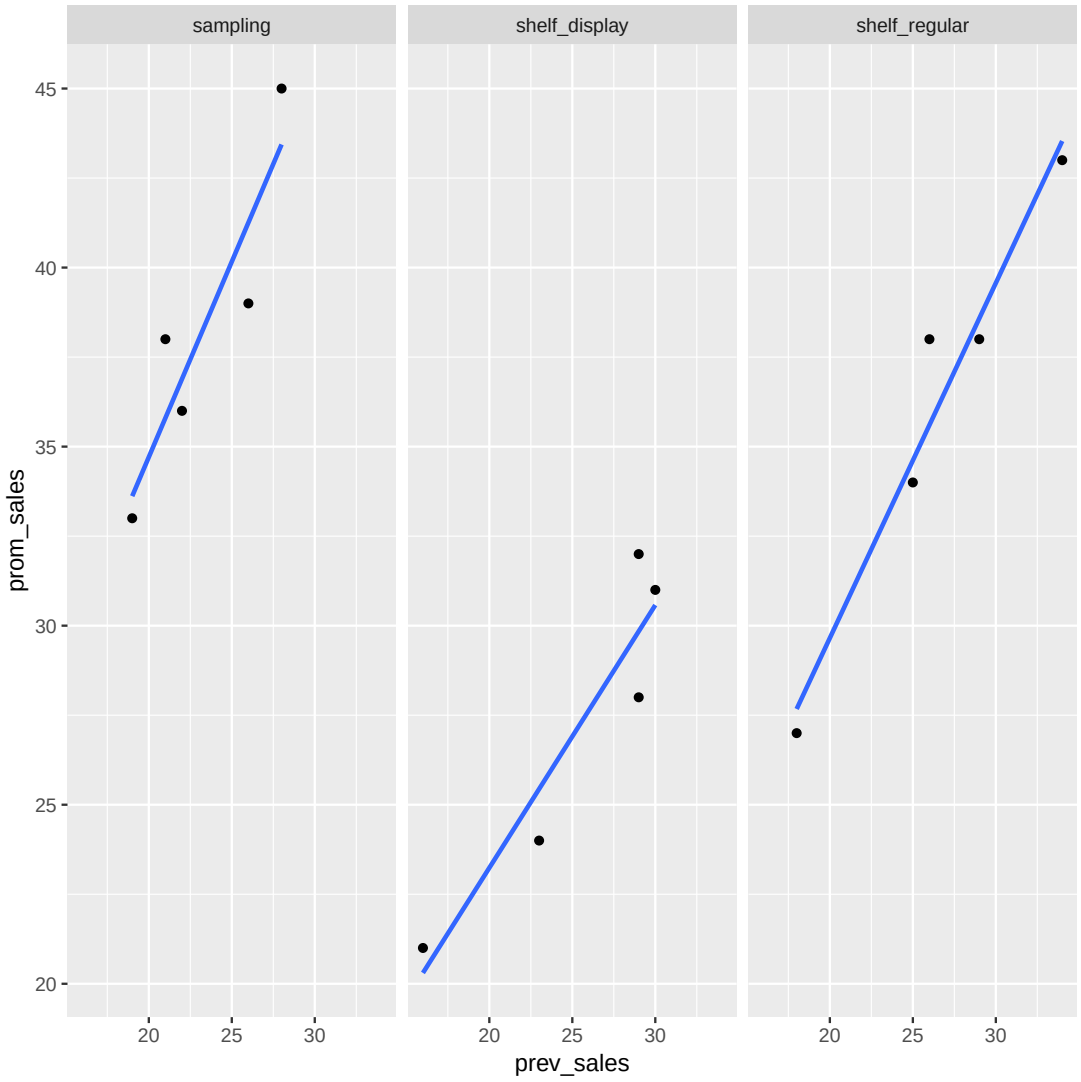
so you need to give me just the `geom_smooth()` piece. You don't need to write out the whole thing again (although, if you do, the only thing you've wasted is your time, since it won't cost you any points).

The `se=F` is optional; full marks with or without it, if everything else is correct. If you omit it, you get grey "envelopes" around the lines; I think it is clearer without them, but that's just my opinion.

Note that the *one* `geom_smooth` will plot one line for *each* colour, so you get all three lines at once. It's not more difficult than that.

If you facetted in the previous part, the same addition will get you the appropriate regression line on each facet:

```
ggplot(crackers, aes(x=prev_sales, y=prom_sales))+
    geom_point()+geom_smooth(method="lm",se=F)+
    facet_wrap(~promotion)
```



The interpretation of these graphs (which I kind of wanted to ask you for: maybe I'll put one of those on another assessment) is that promotional sales are higher if previous sales are higher (for all promotion types), and, if all else is equal (that is, for any fixed value of previous sales), sales are highest for `sampling`, that is, when customers get to sample the crackers. So this is the most effective promotional strategy.

4. A sample of 52 people have their pulse rates measured. The aim to estimate the mean pulse rate of "all people" (that is, all the people of which these data are a sample). The data frame is called `pulserates` and the column of interest is called `Pulse`. A histogram of the data is shown in Figure 5.

   (a) (3 marks) What, precisely, do you conclude about the data from Figure 6? You should discuss only inferences that make sense.

   Solution: There is no test being done here, or at least, if there is, it's testing that the mean is *zero* against a two-sided alternative. That doesn't make much sense.

   So, what this is doing is making a 90% confidence interval for the mean pulse rate of "all people". With 90% confidence, the mean pulse rate of all people is between 71.2 and 74.2.

   You should:

   - say that this is a 90% confidence interval
   - say that it is an interval for an appropriate population mean
   - round off the ends of the interval to one or two decimal places, since you would typically be communicating a confidence interval to other people, and *they* don't want to see lots of decimals.

   Expect to lose one point for each of each of those that you forget, or each extra one that you talk about (eg. a test), with the proviso that if you say *something* sensible you should get one point.

   (b) (2 marks) Do you think that a t procedure can be trusted here? Explain briefly, using one of the Figures to support your opinion.

   Solution: This is thinking about the assumptions behind the t-test, so look at the histogram in Figure 5. We need this to be approximately normal (possibly with consideration of the sample size). I think there are two ways you can go:

   - this distribution is itself approximately normal, so there are no issues with the t interval and it can be trusted
   - this distribution is not normal (eg. bimodal), but it has no outliers or skewness, so the sample size of $n = 52$ is large enough for the Central Limit Theorem to work well, and thus the interval can be trusted. Another way to say this is to describe the distribution as "not badly non-normal", and then talk about the Central Limit Theorem as it applies here.

   I don't think any other answers will work. I really don't think this picture is non-normal enough to cause any problems at all, and I don't see how you could make a case for that. Two points for one of the above stated clearly, one for one of those not stated clearly enough, zero for anything else.

   (c) (3 marks) Give code that will produce the output in Figure 7.

   Solution: This is doing a hypothesis test whose alternative is that the mean is greater than 71, and so the null is that the mean is equal to 71. You need to say three things:

   - which data frame the column `Pulse` is coming from
   - what the null mean is

- which direction the alternative hypothesis is

This is the code I used:

```
with(pulserates, t.test(Pulse, mu=71, alternative="greater"))
```

This is also good:

```
t.test(pulserates$Pulse, mu=71, alternative="greater")
```

Minus one per error, down to one point if you have the `t.test` *plus* one other thing correct. (No credit for `t.test` by itself, since that was above Figure 6.) Grading guide: two points for the alternative, one point for the null mean.

If you leave the `conf.level` in, you must change it to `conf.level=0.95`, but you don't need to have it, since 95% is the default confidence level.

(d) (2 marks) Look again at the output shown in Figure 7. What do you conclude from it, in the context of the data?

Solution: The P-value of 0.03274 is less than 0.05, so we reject the null hypothesis (that the mean is 71) in favour of the alternative that the mean is greater than 71. We conclude that the mean pulse rate is greater than 71. Or, there is evidence that the mean pulse rate is greater than 71.

If you talk about the confidence interval here at all, expect to lose a mark. This confidence interval is *one*-sided, and ours are two-sided in this course. Part of the issue is knowing what *not* to say.

Extra: there is a clue here from the (correct) confidence interval in Figure 6. The 90% confidence interval *did not* contain 71. Thus the P-value for the two-sided test is less than 0.10. Our sample mean is greater than 71, so we are on the correct side, so the one-sided P-value is less than $0.10/2 = 0.05$. This is consistent with the test done here. The proper reasoning gets the same conclusion both ways. (Taking another step: 71 is just outside the 90% CI, and the P-value here is a little less than 0.05, which is what you'd expect.)

5. The chest circumference of healthy newborn baby girls is normally distributed with mean 13.0 inches and standard deviation 0.7 inches. (These figures come from observing a very large number of newborn baby girls.) A population group from a remote region has a different genetic makeup, and therefore possibly a different mean chest circumference in its newborn baby girls.

   (a) (3 marks) Suppose in fact that newborn baby girls in the remote region have a mean chest circumference of 12.8 inches. A sample of 25 such girls will be taken in the remote region. Give code to *calculate* the power of a t-test for the mean, testing that the girls from the remote region have the same mean chest circumference as the general population, against a two-sided alternative.

Solution: The true mean is 12.8, the null mean is 13, the population SD is 0.7 (use the value given in the question) and the sample size is 25. Taking the hint that "calculate" means "obtain an exact answer":

```
power.t.test(n=25, delta=12.8-13, sd=0.7, type="one.sample",
             alternative="two.sided")
```

```
##
##      One-sample t test power calculation
##
##               n = 25
##           delta = 0.2
##              sd = 0.7
##       sig.level = 0.05
##           power = 0.2781731
##     alternative = two.sided
```

It is best to include the two-sidedness in `alternative`, to make it clear what you are doing, but it works without, since that is the default:

```
power.t.test(n=25, delta=12.8-13, sd=0.7, type="one.sample")
```

```
##
##      One-sample t test power calculation
##
##               n = 25
##           delta = 0.2
##              sd = 0.7
##       sig.level = 0.05
##           power = 0.2781731
##     alternative = two.sided
```

Omitting the `alternative` will work, and is therefore full marks.

The `one.sample` *must* be included, since the default is a *two*-sample t-test:

```
power.t.test(n=25, delta=12.8-13, sd=0.7)
```

```
##
##      Two-sample t test power calculation
##
##               n = 25
##           delta = 0.2
##              sd = 0.7
##       sig.level = 0.05
##           power = 0.1660423
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

(the power is less than we got for our one-sample test since *both* of the two samples have uncertainty attached to them, so there is more uncertainty present overall.)

Marks: you need all of `n`, `delta`, `sd` and `type`. Expect to lose one for each wrong or missing one, with the proviso that you get one if you have `power.t.test` and even *one* correct input. My thinking is that if there are any errors you do not deserve full marks, and if you have anything substantial right, you deserve at least one.

Extra: I learned something just now about `power.t.test`. If you work out the power of a two-sided test, as we just did, it *only* gets the probability from one tail, the one in the same direction as the truth. In this case, the null mean is 13 and the true mean is 12.8, so the 0.278 is the probability of getting a sample mean far enough *less* than 13 to reject with. Since this is a two-sided test, we could have observed a sample mean like 13.8 that is *bigger* than 13 and so far away from 13 that we reject in the upper tail. This ought to be pretty unlikely if the population mean is really 12.8, but it is not impossible. The way to include this in the calculation of the power is to include `strict=T`, thus:

```
power.t.test(n=25, delta=12.8-13, sd=0.7, type="one.sample",
             alternative="two.sided", strict=T)
```

```
##
##      One-sample t test power calculation
##
##               n = 25
##           delta = 0.2
##              sd = 0.7
##       sig.level = 0.05
##           power = 0.2786014
##     alternative = two.sided
```

You see now that the power is a teeny bit bigger, since it includes both tails. This one is strictly speaking the right answer, but since I didn't even know about it myself until just now, I have no reason to expect you to know it.

Extra extra: this explains something I had in one of my assignment solutions, which I now have to go back and correct. If the null hypothesis is *correct* in a two-sided test (which translates here to `delta` being zero), the probability of rejecting it should be 0.05, since rejecting it is now a type I error:

```
power.t.test(n=25, delta=0, sd=0.7, type="one.sample",
             alternative="two.sided")

##
##      One-sample t test power calculation
##
##               n = 25
##           delta = 0
##              sd = 0.7
##       sig.level = 0.05
##           power = 0.025
##     alternative = two.sided
```

Except that it isn't, because `power.t.test` is only counting one tail, so it is getting half the right answer. Adding `strict=T` fixes everything:

```
power.t.test(n=25, delta=0, sd=0.7, type="one.sample",
             alternative="two.sided", strict=T)

##
##      One-sample t test power calculation
##
##               n = 25
##           delta = 0
##              sd = 0.7
##       sig.level = 0.05
##           power = 0.05
##     alternative = two.sided
```

As it should be. Sometimes reading the help file teaches you something new!

(b) (3 marks) The power value from running your code from the previous part is shown in Figure 8. Explain precisely but briefly what this number tells you.

Solution: The power is the probability of correctly rejecting the null hypothesis when it is wrong. For full marks, you need to correctly say what that means for these data, such as:

> The null hypothesis is that the mean chest circumference is 13 inches. 0.28 is the probability that this null hypothesis will be correctly rejected, when the mean is actually 12.8 inches.

(This is my second attempt at wording this, so you might need to have a couple of tries to come up with something clear.)

If you prefer, do it with symbols, as long as you define them first:

> Let μ be the mean chest circumference of all newborn baby girls in the remote region. The value 0.28 is the probability of correctly rejecting $H_0 : \mu = 13$ when in fact $\mu = 12.8$.

(It took me two goes to word that one properly as well.)

Somewhere in your answer, you need to say what the null hypothesis is, what the truth is, and that the value given in the Figure is the probability that the null hypothesis is rejected (I like to add the word "correctly" to emphasize that we *should* be rejecting the null here because it

is wrong).

Three marks for getting all three of those. Two marks if you get as far as "the probability of the null hypothesis being rejected when it should be rejected" but getting the true and null means the wrong way around, or not mentioning them. I particularly want to see the idea of "the null hypothesis being rejected when $\mu = 12.8$", since the chance of rejecting the null depends in part on how wrong it is. (If you have "when ..." in my grading, this is the bit you missed.) One mark for messing up the definition of power but getting the means right, or for giving the definition of power *in general* without applying it to this problem. (Knowing the theory is good, but you will not be very useful to the world if you don't know how to apply the theory!)

(c) (4 marks) Give R code to *simulate* the power of this test, but now with a sample of size 100. Use the technique given in lecture.

Solution: I want you to take one of the examples you've seen and adapt it for this case. There is not too much adaptation required:

```
rerun(1000, rnorm(100,12.8,0.7)) %>%
    map(~t.test(.,mu=13)) %>%
    map_dbl("p.value") -> pvals
tibble(pvals) %>% count(pvals<=0.05)
```

```
## # A tibble: 2 x 2
##   `pvals <= 0.05`     n
##   <lgl>           <int>
## 1 FALSE             175
## 2 TRUE              825
```

This is the code I ran in Figure 9. (There is a surreptitious `set.seed` to make the answers come out the same here and there. Otherwise, different random samples would give a different result. I think the correct answer is a bit less than this.)

The first number in `rerun` is the number of simulations; anything 1000 or bigger is fine. The three numbers in `rnorm` are the size of each sample, the true mean and the true SD (the latter taken from the question). The value of `mu` in the `t.test` is the null mean. Expect the grader to be checking these carefully. Detail matters.

Minus one point per error, down to a minimum of 1 if you got one of my four lines correct. I call something like switching 12.8 and 13 one error, becasuse at least you're recognizing that there should be different values in the two places.

Common errors are switching the true and null means, and getting the 100 in the wrong place. The first number in `rerun` should be some large number of simulations (at least 1000), and the sample size is the first number in `rnorm`.

Extra: how accurate is my simulated power? As in the assignment:

```
prop.test(825,1000)

##
##  1-sample proportions test with continuity correction
##
## data:  825 out of 1000, null probability 0.5
## X-squared = 421.2, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.799694 0.847766
## sample estimates:
##     p
## 0.825
```

The "true power" is between 0.80 and 0.85, with 95% confidence. I could run 10,000 simulations to nail down the answer more precisely.

In the textbook from which I got this scenario, they were doing power for a z-test assuming that the population standard deviation was known. In our case, we are doing a one-sample t, using the sample standard deviation, so my guess is that the power ought to be a little less, though with a sample of size 100 it shouldn't make much difference. The textbook computes the exact power to be 0.805, which is very much consistent with what we have here (inside the confidence interval, for example).

(d) (3 marks) The result of your simulation is shown in Figure 9. I did this the same way as in lecture. What is the estimated power of this test for a sample of size 100? Compare your result to the one in Figure 8. Does the result of your comparison make sense? Explain briefly.

Solution: The estimated power from the simulation is $825/1000 = 0.825$. (An easy one point.)

This is bigger than the power from a sample size of 25, which makes sense because with everything else the same (as it is here), we should have a greater chance of (correctly) rejecting a null hypothesis that is wrong if we have a larger sample size. (This is what those power curves we looked at in class were demonstrating.)

Say "larger sample size should go with larger power, and it does here" for the remaining two points. One point for each thing. Once again, it's one thing to know what *should* happen, but the important thing for us is to verify that it actually *did* happen.

I need some kind of reason that makes sense to give you the second and third point (even if you compared the powers and said that the one here was much bigger).

Note that the sample size in (c) was 100, the first number in `rnorm`. The first number in `rerun`, 1000, was the number of *simulations*, so that the simulated power was based on 1000 simulated samples, each sample of size 100. It's important to keep these straight. (For a simulation, the 1000 is something "large", at least 1000, but the sample size is the exact sample size you want the power for.)

The idea between Figures 8 and 9 is that we changed the scenario, but the true mean did not change; the only thing that changed was the sample size, so it is the effect of increasing the sample size that we saw. (There is going to be *some* randomness attached to the simulation, but not nearly as much as the difference in powers between the two Figures, so randomness is not an explanation of the change in power.)

6. 25 NHL (hockey) players are randomly sampled and their annual salaries recorded, in millions of dollars. Some of the data frame is shown in Figure 10. The first column is a number to identify the player whose salary is in the second column. Salaries are often very right-skewed, and so a sign test is preferable to a t-test when making inferences for the "typical" salary.

   (a) (2 marks) What null hypothesis is being tested in Figure 11? (If you use any symbols, define what those symbols mean.)

   Solution: That the median annual salary (1 point) is equal to 1.0 million dollars (the second point).

   If you like it with symbols, for example let M be the median salary of all NHL players in millions of dollars, and then we are testing $H_0 : M = 1.0$. Don't use the symbol μ, because that by convention is the population *mean*, which is not what we're testing here. (I suppose you could define μ to be the *median* salary of all NHL players, but it would look weird.)

   We know the hypothesized median is 1.0 because of the inputs to `sign_test`: these are the data frame, the column and the hypothesized median.

   (b) (1 mark) In Figure 11, I use `sign_test` from `smmr`. What are the numbers 7 and 18 in the output?

   Solution: There are no points for saying "above" or "below", since that is in the output! The issue is *what* they are above and below; these are how many salaries are above (18) and below (7) the hypothesized median salary of 1.0 million dollars.

   Use "hypothesized median" or "null median" or whatever numerical answer you got from (a). I originally had these two parts together, but I thought it was easier to separate them out. Having done that, I couldn't really offer more than one point for this.

   (c) (2 marks) Is there evidence that the "typical" annual salary is *greater* than the value you had in your null hypothesis in (a)? Explain briefly.

   Solution: This means pulling out the appropriate P-value from Figure 11, *telling me which one it is*, and drawing the appropriate conclusion. The upper-tailed P-value is 0.022, which is smaller than 0.05, so there *is* evidence that the median salary exceeds 1.0 million dollars.

   Give me the P-value (one point) and a conclusion with the word "median" in it (since that's what the sign test uses; the second point).

   When I started watching hockey in the late 1980s, you had to be a superstar like Gretzky or Mario Lemieux to make a million a year, but I guess things have moved on since then.

   (d) (2 marks) Describe how the P-value labelled `upper` on Figure 11 was obtained. (You may or may not have used this P-value elsewhere in this question.) Give a description in words or with code, as you prefer.

   Solution: If the null hypothesis is true and the median really is 1.0, the number of values either above or below 1.0 will have a binomial distribution with $n = 25$ (25 players in the sample) and $p = 0.5$. The P-value is then the probability of obtaining 18 or more salaries above 1.0, or, alternatively, of obtaining 7 or fewer salaries below 1.0, from the null distribution.

   That's the words way. If you want to do it with code, one of these is what you need:

```
sum(dbinom(18:25,25,0.5))
## [1] 0.02164263
sum(dbinom(0:7,25,0.5))
## [1] 0.02164263
```

or even (if you know about `pbinom`, which is the probability of that many successes *or less*):

```
pbinom(7,25,0.5)
## [1] 0.02164263
1-pbinom(17,25,0.5)
## [1] 0.02164263
```

The second of those requires clear thinking: the binomial distribution is *discrete*, taking values only on the integers, so the complement of "18 or more" is "*17* or less". I used to show `pbinom` in class, but summing up the individual probabilities seems easier to understand.

As you see, all of those give the right answer.

"Half of the two-sided P-value" doesn't offer any insight, so it gets no points. Also, if I had wanted code to produce *that* output, I would have asked it that way. I wanted to know how that number was calculated.

Extra: the nomenclature in R is that `d` followed by the name of a distribution gives the probability density (for a continuous random variable) or the probability mass function (for a discrete one such as this). "Probability mass function" is just a fancy way to say "the probability of exactly one value". `p` followed by the name of a distribution is the cumulative distribution function: the probability of that value *or less*. This works for both discrete and continuous distributions. The last one is `q`, which is the "inverse cumulative distribution function", for example:

```
qnorm(0.975)
## [1] 1.959964
```

On a standard normal distribution, the value 1.96 has 97.5% of the probability less than it (and 2.5% more), hence this is also true:

```
pnorm(1.96)
## [1] 0.9750021
```

Close enough for government work, as they say.

7. One program to help people stop smoking cigarettes is "posthypnotic suggestion". Each subject is hypnotized, and while under hypnosis is told to avoid cigarettes. Subjects for whom the program works will subconsciously choose to smoke fewer cigarettes after the hypnosis than before.

   Eighteen subjects agreed to test the program. Each subject recorded the number of cigarettes they smoked the day before the program, and also the number of cigarettes they smoked the day after the program. The data are shown in Figure 12. The three columns are an identifier for the subject, the number of cigarettes smoked the day before the program, and the number smoked the day after.

   (a) (2 marks) Why is this a matched-pairs study rather than two independent samples? Explain briefly.

   Solution: Each smoker is measured twice (both before and after the program), so there are two observations for each subject. (Or, there are 36 measurements but only 18 subjects, or anything equivalent.)

   "Because this is before and after data" is only one point. The important thing is that they are before and after measurements *on the same people*. (It is true that before and after measurements usually are matched pairs, but I could have had one group of people measured before and a second, different, group measured after, and then it would have been two independent samples. I was pleased to see that some students recognized this.)

   I don't want to know here about why matched-pairs studies are better in general. I want to know why *this* one is matched pairs. (Applied statistics is all about seeing how the theory you learned applies, or does not apply, to *the data in front of you*.)

   Extra: you might also have quibbled with this study design (not as an answer to this question, but in general) because there is no control group, just 18 subjects who agreed to participate. I don't know what kind of placebo you would devise in this situation; maybe you invite the control group to listen to a doctor explain the hazards of smoking, or an inspirational speaker persuading all the smokers that they have it within them to stop smoking, or something like that. This kind of control group would be emphasizing the subconscious aspect of the hypnotic suggestion; if the smokers smoked less after the post-hypnotic suggestion than after one of these placebo treatments, it must have been the subconscious nature of the hypnosis that made a difference.

   Extra extra: if you have a control group like the above as well, you have two sets of before-and-after measurements, one for the treatment group and one for the control group. Thinking about a suitable analysis is likely to make your head hurt, since you have a "within-subjects factor" of time (before and after) and also a "between-subjects factor" of treatment (hypnosis and placebo). The easiest way to think about this is to take the differences before minus after (since the amount of smoking is expected to go *down*) for each subject, and compare the differences between treatment and control using a two-sample test. The trick of working out the differences gets rid of the within-subject variation. From such a test, you might be able to conclude that the hypnosis reduces smoking by more than the placebo treatment does. The test below tests whether smoking has been reduced *at all*.

   (b) (4 marks) Three normal quantile plots are shown in Figures 13, 14, and 15. Which of these plots do you need to consider in order to assess the assumptions for a matched-pairs t-test? Explain briefly. For your chosen plot or plots, what do you conclude about the appropriateness of a matched-pairs t-test? Explain briefly.

   Solution: The assumption we need for a matched-pairs t-test is that the *differences* are approximately normal. The normal quantile plot for these is shown in Figure 15. The normality of the before values and the after values separately is not required, and so Figures 13 and 14

are not relevant to this.

Having chosen a plot, you now need to assess it for normality. On Figure 15, the points do not follow the line very well, so the data are not even approximately normal. The way these data fail to be normal is that they are too bunched up both at the bottom and the top: that is, the tails are too *short* compared to the normal. Often, this happens with something like a uniform flat-top distribution, or sometimes with a bimodal distribution (two humps). (I explore this further below.) I would also take the interpretation that the left tail is OK and the right tail is too short (too bunched up), so the distribution overall is left-skewed. Say what you see. If it looks like something you could reasonably infer, I'm good with it. It's real data, so there isn't one absolutely right answer. Those extreme points are not outliers, though, because they are *not extreme enough* for a normal distribution. I had to deduct a point from an otherwise good answer if you thought they were outliers. (You might call them something like "inliers".) "Far away from the line" is a safe way to describe them, or, as somebody wrote, "the tails are both quite far from expressing overall normality", which I think gets it nicely.

You can reasonably say that this is a smallish sample of differences, so they may not *look* very normal. Having said that, though, I think the departures away from the line at the ends for these data are too systematic (a clear S-bend shape) to be normal-plus-chance, which would look more "wiggly" without a clear shape to the wiggles. (You can explore that yourself: generate a few random normal samples of size 15 using `rnorm`, and draw normal quantile plots of them. I'd be very surprised if you saw a shape like this among them.)

Finally, you need a recommendation for doing or not doing a matched-pairs t-test that is in line with your conclusion about normality. (The alternative test here, which you don't need to say, is a *sign* test on the differences. A Mood's median test would be for a two-sample test.) A (subtle) point in favour of doing a t-test anyway is that the tails are *short*, so that the mean is still a good measure of centre even though the distribution is not very normal. You can make your own call on this based on what you see.

Choosing among Figures 13–15 on the basis of which one looks most normal is *wrong* (and very possibly dishonest data analysis).

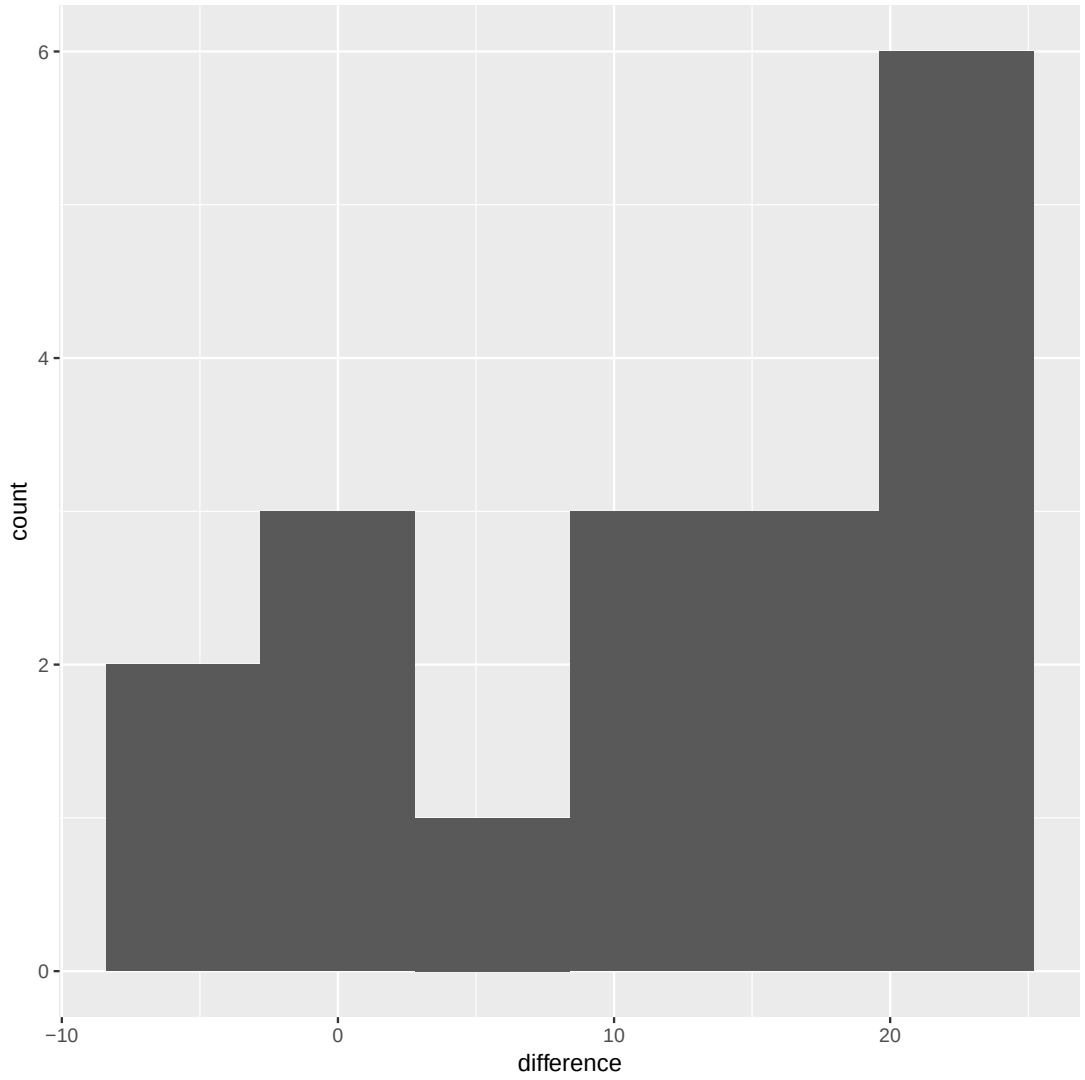
Marking: two points for choice and explanation, two points for conclusion from chosen plot(s). Nothing (out of the first two) for making a choice without also explaining *why* you are making that choice.

Extra: I'm curious about what's going on. I have to read the data in again:

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

## Parsed with column specification:
## cols(
##   subject = col_double(),
##   before = col_double(),
##   after = col_double()
## )

smoking %>%
    mutate(difference=before-after) -> smoking2
ggplot(smoking2, aes(x=difference))+geom_histogram(bins=6)
```



That looks left-skewed. I guess the left tail is not far off what you'd expect for a normal distribution (on the normal quantile plot the points are not so far off the line at the low end), but the points at the upper end are bunched up and that's where the problem lies. Or, indeed, you can look at the "hole" and call this bimodal. All a matter of opinion.

Bear in mind that a histogram is not a precision tool and the sample size is small, so things will not be as clear as you might like.

Whether you look at the normal quantile plot of the differences, or at this histogram, you might note that most of the before-minus-after differences seem to be positive. That gives you some intuition for the test coming up: it might be significant on the correct (upper) side.

(c) (3 marks) Four possible analyses for these data are shown in Figures 16 through 19. Which analysis is the most appropriate one? Explain briefly. What do you conclude from the analysis, in the context of the data?

Solution: This needs to be a matched pairs analysis of some kind, which eliminates Figure 18. That is a two-sample t-test, albeit a slightly non-standard one, since it doesn't have a model formula (with squiggle) in it. (The big clue is the first line of the output.)

I suspect you will have found that the differences did not come from a normal distribution, which would rule out the two matched-pairs t-tests in Figures 16 and 17. The right test is then the sign test on the differences, which is the one in Figure 19. Your explanation should then justify doing a matched-pairs sign test, on the grounds that the differences in Figure 15 did not look normal.

The differences for the sign test were calculated in Figure 15 as before minus after, a number that will be *positive* if the post-hypnotic suggestion has a positive effect on reducing smoking. Thus we should do a one-sided test in the *upper* tail, giving a P-value of 0.0106, and a conclusion that the program *does* reduce smoking (from what it was before).

Another way to come at this is to say that the distribution of differences being short-tailed does not really invalidate using the mean as a measure of centre, and therefore the t-test is OK to use. (What would make the mean not work is outliers or skewness, and we don't have either of those here.) So I suspect that the upper-tailed matched pairs t-test will not be so awful here (Figure 17: the P-value is 0.000059, so there is (stronger) evidence that the program helps). If you came to this conclusion, you get full credit here for looking at this t-test, but not for naming any of the other Figures.

If you found problems with the normality in the previous part but you nonetheless picked Figure 17, you get one point if you interpreted it correctly. Expect to get two points if you picked the right test for a good reason but messed up the interpretation.

If you give no explanation for your choice of output, you are fighting for one point (and that is if you draw an appropriate conclusion from the output you chose).

8. Do trees grow to different heights depending on which side of a building they are planted on? To investigate this, 48 elm tree seedlings were planted, a randomly chosen 12 of them on each of a building's four sides (north, south, east and west). After "several years" of growth (that's what my source says), the heights of the resulting elm trees, in metres, were measured. Some of the data is shown in Figure 20.

   (a) (3 marks) The researchers are planning to run an analysis of variance to compare the tree heights on the different sides of the building. What two major assumptions are required in order to be able to trust an analysis done using `aov`?

   Solution: We need:

   - (approximately) normally distributed data (1 point) within each group (1 more point)
   - with (approximately) equal spreads in each group (1 point)

   `aov` does the F-test-style analysis of variance, so these are the two important assumptions required for it to work properly. (We also need random samples, independent observations etc., but we need those for pretty much everything.)

   Just saying "normally distributed data" is not enough (that's only one point for the first thing); we need (approx) normality *within* each group, not for all the data together. ("Normally distributed residuals" is another way to get at the same thing, since in this kind of model the residual is the observation minus the group mean. If the groups are normal with the same variance, the residuals pooled together, for all four groups, will be normal, but if not, the distribution of the residuals will be a mixture of normals.)

   Make absolutely sure that you know the difference between an *assumption* and a *hypothesis*. A hypothesis is something that we are using a test to decide upon (like, the trees on each side of the building all have the same mean height). An assumption is something that we are not testing, but that ought to be at least approximately true if we are to believe the results of the test (like, the data within each group have normal distributions). As it says in Numerical Recipes, "you will suffer endless agonies if you fail to understand this simple point". (It really says that. Google it if you don't believe me.)

   (b) (2 marks) Normal quantile plots are shown in Figure 21. What do you conclude from these? Explain briefly, bearing in mind that there are only 12 observations per group. (If you need different normal quantile plots, explain briefly what you would need and why you would need it.)

   Solution: These plots are for assessing the normality of data within each group (that is, on each side of the building). In the previous part, you (I hope) decided that this is what you need to check in order to do ANOVA.

   Each of the four plots should be "reasonably" close to the line. I asked you to consider that there are only 12 observations per group in order to dissuade you from asking for perfection. My take is that east, north and south are acceptably normal ("inconsequential wiggles" is how I'd describe their departures from their lines), but west seems definitely non-normal with short tails at both ends. As with residual plots (see later with regression), it doesn't do to stare at these for too long, since then you're bound to see *something* (that probably isn't really there). I would just about accept that the south-side trees have two upper-end outliers, but really I don't think these are far enough above the line to be outliers.

   An additional point here (that applies later) is that ANOVA will fail with big enough skewness or outliers. *Short* tails do not pose a problem. So, looking ahead, I would be happy to do the regular ANOVA here.

Question 8 continues... This page: _____ of possible 5 points.

Extra: these plots do not appear to permit a comparison of spreads (at least, we didn't use them this way in class). This is why I allow you to assume "acceptably equal spreads" later. But actually, you can. The lines join the first and third quartiles for each group, and since the graphs are all on the same scale, if the line is steeper, the group has a greater spread. Thus the west group heights are actually more spread-out than the others, which are all about the same.

You can also get some intuition about means or medians from these plots: if the points are typically higher up in their box (eg. the south-side heights), the mean and median will be bigger.

I am fully expecting this to be hell to mark. What I am looking for is two or three sensible and relevant points, about groups that are not normal, or a number of groups that are normal, in your opinion. If it appears to be a well-founded opinion, I'm happy with it. In the matched-pairs question, I was picky about outliers, so I won't be so much here; something that says the west values are not normal is enough, and here I will take "outliers" as a reason why. (Don't expect me to be as generous again, but my point is that I don't want to penalize you twice for the same misconception.) I am trying hard to give you two points here if I can, so give me something to give you two points for!

In case you were wondering, the numbers of people who thought normality was acceptably good and who thought there was at least one untenable problem were about 50–50. Democracy in action. (This means that from my point of view, the question came out well; if most of you had gone one way, that would have been disappointing.)

(c) (3 marks) Figures 22 and 23 contain two possible analyses for these data. Which analysis do you prefer? Explain briefly. What do you conclude from your chosen analysis? You may assume, if you need to, that the heights on each side of the building have acceptably equal spread.

Solution:

The last sentence of this part is a clue that "equal spread" ought to be one of your assumptions in (a)!

The way you tackle this will depend on what you thought of the normal quantile plots. If you thought they were all acceptably normal, as I did, that plus the (given) assumption about equal spreads is enough to support Figure 22 (the regular ANOVA). If you thought there were unacceptable problems with the normality (that is, at least one of the four distributions was not normal enough), then you should go with Mood's median test in Figure 23. I will accept either of those, *as long as your choice here is consistent with your call in the previous part.*

My call about "all acceptably normal" was based on the West side having *short* tails, and that this wouldn't affect the mean and thus the F. I would also accept "West not normal, therefore use Mood".

The conclusions are basically the same for both tests: for the ANOVA, the P-value 0.007 and you conclude that the mean heights are not all equal for the different sides, and for the Mood's median test, the P-value is 0.011 and you conclude that the *median* heights are not all equal.

One point for choosing a Figure consistent with your call about the normality, one for quoting the P-value from that (or for saying that it is small), and one for a conclusion about heights of trees. Saying "there are differences between groups" is not enough, because it doesn't show that you know what that means *here*. If you did Mood's median test, talking about differences in *means* is an error (which seems kind of obvious when I write it like that, but anyway.)

Extra: had I not thrown in the bit about equal spreads, and had you known how to assess

spreads from a normal quantile plot (a lot to ask, I know), you might have come to the conclusion that the distributions were normal enough, but the spreads differed. That's the kind of situation in which you would run a Welch ANOVA via `oneway.test`. I didn't want to confuse the issue further by providing a third ANOVA and also a Games-Howell that you would have had to choose between, so I allowed you to assume equal spreads. On my first draft of this question, I also gave you boxplots of heights for each side; the West spread there is clearly bigger than the others, and the boxplots appear to show clear non-normality as well, something that is mostly an artefact of the small sample sizes, so I thought that would be too confusing and I took the boxplots out.

As a point of exam technique, I asked you which of the two analyses you *preferred*, so you were supposed to pick one. But saying that a Welch ANOVA would be better is definitely something intelligent to say, so I like that idea as well.

(d) (3 marks) Is it worthwhile to follow up with one of the analyses in Figures 25 or 26? Explain briefly why or why not. If it is worthwhile to follow up with one of those analyses, say which one, and say what conclusions you draw. Use Figure 24 if it is helpful.

Solution: Whichever of the ANOVA or the Mood's median test you looked at in the previous part, the P-value is small (less than 0.05) and so you conclude that the mean/median heights are not all equal (or, there are some differences in mean/median height to find). One point.

That means that you *do* look at one of Figures 25 or 26. Look at the Tukey, Figure 25, if you did the ANOVA, and look at the pairwise median tests, Figure 26, if you did Mood. Make sure you state the number of the Figure you're looking at, so that I can be sure that you have the right one. (Or give me some way to work it out, for example quoting me the small P-value you are looking at. These are different for the Tukey and the pairwise median tests, so I can tell whether you are looking at the right one.)

In both cases, the conclusion is the same: only the east and south sides differ significantly in mean/median height; none of the other differences are significant. Two points. ("Only" is a good word to use here; it saves you a lot of writing.)

A glance at Figure 24 reveals that the south mean and median are both significantly *bigger* than the east ones. Make the appropriate comparison for the followup that you did. As I originally planned it, this was supposed to be the third point, but that seems in retrospect a lot to ask, so "only the east and south sides have different mean/median heights" (as appropriate for the followup test you are doing) is a complete answer.

If you did Mood's median test, make sure you look at the *last* column of Figure 26 with the *adjusted* P-values in it. These are the proper (Bonferroni) adjustment for doing six tests at once; the ones in the `p_value` column are unadjusted, and using those is "dishonest" in the same sense as Tukey's "honestly significant differences". (If you do that, you will find an extra "significant" comparison that is not actually significant when you do it properly.)

Extra: I have to rewrite `pairwise_median_test` to not give P-values greater than 1. A P-value of 2.49 doesn't exactly make much sense! A student discovered a different problem with this function that I also have to fix.