University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 / STA 1007 (K. Butler), Final Exam
April 24, 2018, 2:00–5:00pm (3 hours)

Aids allowed:

- My lecture overheads (slides)
- Any notes that you have taken in this course
- Your marked assignments
- Non-programmable, non-communicating calculator

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 34 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 in the table on the next page.

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	
2	4	
3	2	
4	5	
5	7	
7	5	
8	2	
9	6	
11	6	
13	8	
15	5	
16	2	
17	2	
18	7	
19	4	
21	8	
23	4	
27	5	

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

1. An experiment is done to test the effect of a toxic substance on insects. At each of six dose levels, 250 insects are exposed to the substance and the number of insects that die is counted. We are interested in whether there is an association between dose and how likely an insect is to die, and if so, what kind of relationship it is.

   The data are shown in Figure 2 (in the booklet of code and output). The values are separated by spaces, not tabs.

   (a) (2 marks) Explain briefly why logistic regression is a sensible way to assess our problem of interest.

   My answer: We are looking for a relationship between dose and whether or not an insect survives. The response variable, surviving or not, is categorical (with two categories), and so to predict it we need logistic regression.

   For two points, I am looking for: that the response variable is what we need to look at, that it is categorical, and how you know this (eg. by saying that the two categories are lived and died). I could have insisted on (but didn't) you saying that the response variable has *exactly two* categories, not more than two.

   The reason I gave you this question is that it requires (below) a two-column response, which you hadn't seen on an exam before. This means that the data are laid out with many (250) insects on each line of the data file. The crucial thing is what happens to *each individual insect*: it either lives or dies.

   (b) (1 mark) Give R code to read in the data, as shown in Figure 2, from a file called `toxicity.txt` in the current folder.

   My answer: This is easy, so only one point.

   Separated by more than one space and aligned by columns, so `read_table` is the thing:

```
insects=read_table("toxicity.txt")
```

```
## Parsed with column specification:
## cols(
##  Dose = col_double(),
##  SampSize = col_double(),
##  Deaths = col_double()
## )
```

   That's all you need.

   `read_delim` *will not work*, because the data values are separated by more than one space. I think `read_table2` will also work, and so will the old-fashioned `read.table` with a dot. (I didn't teach you that, but I didn't exclude it, so if that's what you said, that was OK.)

   Extra: with only six lines of data, this will display the whole thing:

```
insects
```

```
## # A tibble: 6 x 3
##    Dose SampSize Deaths
##   <dbl>    <dbl>  <dbl>
## 1     1      250     28
## 2     2      250     53
## 3     3      250     93
## 4     4      250    126
## 5     5      250    172
## 6     6      250    197
```

(c) (4 marks) Give R code to fit a suitable logistic regression for these data, predicting the probability of death, and to display the results.

My answer: The first thing to note is that each line of the data file represents more than one insect (actually, 250 of them), so that we will need to make a two-column response. Another way to see this is that the "response" column `Deaths` is not a categorical "lived" or "died", but a *number* of insects who died.

The second thing to note is that we don't actually *have* the number of insects who lived, only the total and the number who died, and our two-column response requires the number who died and the number who lived (in that order, since we are predicting the probability of dying).

Having created the two-column response, the last two parts are to run the logistic regression and to display the output (with `summary`). So, the whole thing looks something like this:

```
insects = insects %>% mutate(Survived=SampSize-Deaths)
response = with(insects,cbind(Deaths,Survived))
insects.1 = glm(response~Dose, family="binomial", data=insects)
summary(insects.1)
```

```
##
## Call:
## glm(formula = response ~ Dose, family = "binomial", data = insects)
##
## Deviance Residuals:
##       1        2        3        4        5        6
## -0.5092  -0.1115   0.7461  -0.2869   0.4744  -0.5599
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.64367    0.15610  -16.93   <2e-16 ***
## Dose         0.67399    0.03911   17.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 383.0695  on 5  degrees of freedom
## Residual deviance:   1.4491  on 4  degrees of freedom
## AIC: 39.358
##
## Number of Fisher Scoring iterations: 3
```

There are different ways to create the `response`, but `glm` requires it to be a `matrix`, so as long as you get it there, you will be good, eg:

```
response = insects %>% mutate(Survived=SampSize-Deaths) %>%
    select(Deaths, Survived) %>% as.matrix()
response
```

```
##      Deaths Survived
## [1,]     28      222
## [2,]     53      197
## [3,]     93      157
## [4,]    126      124
## [5,]    172       78
## [6,]    197       53
```

or you can avoid creating a named column like my `Survived` at all:

```
response=with(insects,cbind(Deaths,SampSize-Deaths))
```

You should really use the `SampSize` column rather than hard-coding 250 in any of your code. What if the sample size for one of the doses changes? (I took off a point for this, because I think the best answer uses `SampSize`.)

Marking notes: one point for creating a column of survivors, one point for creating the response matrix with two columns (deaths first, since that's what we model the probability of), one point for fitting a logistic regression with `family="binomial"` in it, and one rather gimme point for displaying the `summary` of the model you fitted. (If you fitted a completely wrong model, but you displayed the summary of it, you got one point. That's what I mean by "gimme".) If you combined the calculation of survivors with the making of the response matrix, two points for both together.

I realized I confused the issue by using the word "predict" without accompanying it with "in the model". This use of the word "predict" was meant to guide you towards what should be in the first column of the two-column response (deaths). "Modelling the probability of death" would have been clearer, maybe. If you did something like a prediction, I just ignored it. "Display the results" meant "display the results of the logistic regression".

Another way to ask this question would have been to have you do (give code for) actual predicted probabilities of death for some different doses (which I would have specified somehow), and later to ask you to check the predictions for consistency with what you would expect from the positive slope (below). But this would have been rather similar to the midterm, so I didn't.

(d) (2 marks) Part of the output from your logistic regression is shown in Figure 3. Is there evidence that the probability of death depends on dose, and does the probability of death go up or down as dose increases? Explain briefly.

My answer: The P-value on `Dose` is very small, so there is definitely an effect of dose on the probability of death. The coefficient, 0.67, is positive, so as dose goes up, the probability of death goes up also.

This is as you would expect for a poison. This is also what you would have guessed looking at the data, where the number of insects at each dose is the same, and the number of deaths goes sharply up: this is way more of a trend than you would expect by chance.

I wanted you to pick out two things: the small P-value and the positiveness of the slope, and tell me what they mean about the data. I wanted you to use Figure 3 for the bulk of your conclusions: that is, to make the connection between the positive slope and the probability of death increasing as the dose increases. If you want to look back at Figure 2, you can use that to *add* to your answer (something like "in the data, as dose goes up, more of the insects die") or to convince yourself that the answer you have here is correct, but the reason I was asking for here came from Figure 3.

Be careful about trying to interpret the *number* 0.67; if you do that, you have to do it correctly ("as the dose increases by one, the *log-odds of death* increases by 0.67"). If the probability itself increased by that much for each unit increase in dose, the probability would soon get bigger than 1, which makes no sense. The safe way to tackle this is to recognize the limits of your own knowledge; if you remember the stuff about the log-odds, you can say that, but it's enough to say what the positive slope means. I tried to guide you towards the latter by the way I phrased the question (the bit before "explain briefly"): "does it go up or down" rather than "at what rate does it go up or down".

The output is actually from `tidy` from package `broom`, because the usual `summary` output contains, in `Call`, the code that was used to fit the model, and I wanted you to come up with that yourself.

2. When children are learning to solve a puzzle, does it make any difference how much encouragement they get? In a study, 20 children were randomly assigned to one of four treatment groups:

   - Constant reward
   - Frequent reward
   - Infrequent reward
   - No reward

   according to how often they received a reward or encouragement from the experimenter as they were trying to solve a puzzle. For each child, the number of attempts required to solve the puzzle was recorded (a smaller number of attempts means that the child solved the puzzle more quickly).

   The researchers have some specific research hypotheses in mind:

   - Any reward (average of) will produce faster solution than no reward
   - Constant reward will produce faster solution than (the average of) frequent and infrequent reward.
   - Frequent reward will produce faster solution than infrequent reward.

   A "faster solution" is one that takes *fewer* attempts. The data are shown in Figure 4.

   (a) (2 marks) Explain briefly why the researchers decided to use contrasts rather than the standard ANOVA F-test followed by Tukey.

   My answer: The researchers have three specific comparisons in mind, the ones detailed above, rather than all (six) comparisons of each reward level with each other reward level.

   The idea is that focusing on the specific comparisons we care about should enable us to get a more powerful test, compared to one that compares things we do care about *and* things we don't care about. You can also take the view that the comparisons above were specified *ahead of time* (a *priori* is the fancy name for this).

   There was one mark to be had for getting somewhere towards these ideas, or an answer that I felt was a bit light on the "explain briefly". I was looking for something that talked about how contrasts were useful *here*, rather than copying something from your notes (which is almost never going to get you full marks).

   (b) (3 marks) I created an ordered factor from `reward` in Figure 5. Write contrasts C1, C2, C3 that will enable the researchers to test their research hypotheses, in the order that they are given above. (This is most easily done as three lines of R code.)

   My answer: This is what I used in my code in Figure 6:

```
C1=c(1/3,1/3,1/3,-1)
C2=c(1,-1/2,-1/2,0)
C3=c(0,1,-1,0)
```

   Anything equivalent, like switching plus and minus signs through a contrast, or multiplying a contrast through by a constant, is equally good. Thus `C1` could be written as `(-1,-1,-1,3)`, for example, or even `(-0.5,-0.5,-0.5,1.5)`, which is weird, but it works, so it gets full marks.

   The reason for creating the ordered factor and showing you what I had done was so that you would get the reward levels in the right order. The way I usually ask this on an assignment is to ask *you* to figure out what the right order is, but I couldn't do that on an exam, hence this way. It turns out that the four reward levels were actually in alphabetical order *and* logical order, so I probably didn't need to do this, but I thought it was better to be clear.

   Marking: two points for getting two of the contrasts correct or one correct and two nearly right. *Signs matter*: if you get any signs wrong, the contrasts won't be orthogonal (mine are,

Question 2 continues… This page: _____ of possible 5 points.

and so are yours if you got them right). A common problem was to fail to get the second and third things in C2 to have the same sign. One point for anything down from that to "three things that look like contrasts", that is, three sets of four numbers that seem to add to zero but contain the wrong numbers. There are no words in this answer; I set it up so that you know which order the rewards are in, and you know which numbered contrast is which.

(c) (3 marks) What R code (one or two lines, depending how you write it) would arrange it so that the `lm` in Figure 6 does the right tests (instead of the default use of `Constant` as a baseline and comparing everything with that)?

My answer: Note that *I made this out of 3 points instead of 2*, so that this page adds up to 12 and the exam adds up to 104. I explain why below.

Make a matrix out of your contrasts and set this as `contrasts` of the right thing. I used these two lines:

```
m=cbind(C1,C2,C3)
contrasts(puzzle$rewardf)=m
```

If you prefer, do it in one line by not defining `m` and putting the `cbind` on the `contrasts` line.

Note that the thing you set `contrasts` of *must* be a `factor` (of some kind). If you say `puzzle$reward`, that is *wrong* because `reward` is text, as the data display in Figure 4 shows.

I realized as I was marking this that I wanted to distinguish between someone who wrote just

```
m=cbind(C1,C2,C3)
```

and then stopped (1 point), and someone who wrote

```
m=cbind(C1,C2,C3)
contrasts(puzzle$reward)=m
```

using the *text* `reward` rather than the *factor* `rewardf`, which is an important error but shows a lot more knowledge than someone who didn't write down the `contrasts` line at all. So this is two points out of (now) three, and the correct answer is three points. I went through the pile twice to check that I got the 2s and 3s correct.

Some people will appear to get 12 out of 11 on this page! (It's really 12 out of 12.)

(d) (4 marks) What do you conclude from the analysis in Figure 6, in the context of the data? Note: there is a subtle feature of the analysis which, for full marks, you should address.

My answer: This analysis tests the three research hypothesis, in the order that they were given above. That was the reason for all the setup.

The subtlety is that the tests in this analysis are *two-sided* (as tests for regression slopes always are), but the research hypotheses given above are *one-sided*, with a "faster solution" (implying a "less than" in the alternative hypothesis). Now, you don't know which way around I did my contrasts, since I didn't show you, so the strictly correct way is to say that the tests in Figure 6 only enable us to conclude whether there is *a difference* in means for the appropriate contrast, and if there is a difference, it is not clear which way that difference goes. It could go the *opposite way around* to the research hypothesis. However, the numbers in the `Estimate` column are all negative, which would suggest that the means are the right way around for "quicker" in the research hypotheses. (That is actually how it worked out, as you can check from my contrasts in these solutions.)

So, for the final point, you need to say that the research hypotheses are one-sided and that these tests are two-sided, or mention that the Estimates are negative which would put you on the "correct side" for a one-sided test (and you would then be able to halve the P-values, but that would not affect any conclusions about rejecting nulls at $\alpha = 0.05$; halving the P-values here is

optional, since the main thing I wanted you to see was the one-sided vs. two-sided thing).

I realized, the way I worded the question, if you drew a two-sided conclusion, that was enough. I ought to have asked you to relate the output back to the research hypothesis, but since I didn't, a two-sided solution along the lines of "there is a significant difference between any reward and no reward" gets the fourth point.

The first three points are:

1. Any reward does produce faster solutions than no reward (P-value 0.0012 or half that)
2. Constant reward *does not* produce faster solutions than frequent or infrequent reward (P-value 0.183 or half that)
3. Frequent reward does produce faster solutions than infrequent reward (P-value 0.0010, rounded, or half that).

You needed to show me that you knew what the contrasts were testing, in terms of what the data are. Don't expect to get much for saying that C1 and C3 are significant and C2 is not and quitting there, since that shows almost no insight. (This is probably 1 point, though you might get a second if you demonstrated some additional understanding.)

3. Diabetes is a disease in which a person has high blood sugar over a long period. Two ways of treating diabetes are a special diet, and regular injections of insulin. The aim of these treatments is to keep blood sugar at an acceptable level.

   A study was carried out in which 20 adults wth diabetes were randomly assigned to one of two treatment groups, `diet` and `insulin`. The study lasted six months. Each subject had their fasting blood sugar measured at the beginning and at the end, and the change in fasting blood sugar level, `FBS_change`, was recorded. It is also suspected that the fasting blood sugar levels, and therefore their change, will depend on the age of the subject. (Blood sugar levels also depend on what the subject has eaten, and so are measured when the subject has *not* eaten for a specified number of hours.)

   The dataset is shown in Figure 7.

   (a) (3 marks) A plot is shown in Figure 31. (This is at the end of the Code and Output because it is in colour.) What does this plot tell you about any possible relationships between change in fasting blood sugar and: (i) treatment, (ii) age of subject, (iii) the treatment-age combination? Explain briefly.

   My answer: You need to address all three of those issues somehow. I don't mind (at all) what you conclude, as long as you look at the right thing on the plot each time.

   (i) compare the red points (or line) with the blue one(s). You can say that the points are all mixed up, thus no treatment effect. Or the blue line is higher than the red one, so (`FBS_change` is (slightly) higher for the insulin group than for the diet group (adjusting for age, or comparing people of same age and different treatments).

   (ii) both lines go (slightly) uphill, or there is a weak upward trend in the points, indicating a (small) age effect: older people have a (slightly) higher `FBS_change`. Or conclude that there isn't really any trend, so there isn't really any age effect.

   (iii) compare the slopes of the two lines. You could say that they are (slightly) different, indicating a (slight) dependence on the combination of age and treatment. Or you could say that they are "not substantially different", indicating no dependence on the combination. (Try to save the word "significant" for later when you do a test.) If you think the slopes are not really different, this would indicate that `FBS_change` depends possibly on age, and (separately) possibly on treatment, but the value of one has no bearing on the effect of the other.

   You can say a lot less than I did, but I want *an* assessment of the right thing on the graph in each case. (I really don't mind what you conclude; looking at the right thing is what's important here.)

   When you compare the *slopes* of the lines, you are assessing *interaction*. If you want to look for a treatment effect, you need to consider whether one of the lines (or sets of points) is consistently higher than the other. (If you're comparing the points, it's easier to do this if you fix an age and compare, then note that it didn't matter what you fixed the age at. Comparing the points, you are likely to conclude that there isn't really any effect of treatment.)

   What you conclude here will be either consistent or inconsistent with what you conclude from the ANCOVA (coming up). I don't mind whether you are consistent or inconsistent, but the final part will invite you to think about that again. If you just look at the lines, you can be rather easily convinced that something is going on, because the lines here, like the lines on an interaction plot, can hide a lot of variability.

   (b) (2 marks) Why is an analysis of covariance suitable here? Explain briefly.

   My answer: We have a quantitative response `FBS_change` and, most important, a categorical explanatory variable `Treatment` *and* a quantitative one `Age`. The crucial thing is to note that your two explanatory variables are of different types, and which is which (this is the "explain briefly" part).

I want to be sure that you understand the difference between a categorical variable and its levels. Here, we have one categorical explanatory variable `Treatment`, but it has two levels `diet` and `insulin`.

Expect one point if you had the rough idea but lacked clarity, for example you didn't distinguish between response and explanatory variables, or didn't distinguish between their types, or didn't tell me which variable was which *here*. It's one thing to know that ANCOVA requires one quantitative and one categorical explanatory variable, but it's another thing to look at a data set and say which variable is of which type. As a statistician, you have to be able to do both.

(c) (2 marks) An analysis of covariance is shown in Figure 8. What do you conclude from it, in the context of the data?

My answer: Look *only* at the interaction term. This is not significant, and so should be removed; it means that `FBS_change` does *not* depend on the combination of age and treatment. That is, there is only (possibly) an age effect and a treatment effect.

This is one of those cases where writing too much will cost you a point. The two people who got away with talking about the main effects used the word "suggest": the other two P-values are big, so we might *guess* that taking out the interaction term won't be enough to make them significant (which is a good intuition).

A couple of people said "No other conclusions may be drawn", which I *love*.

(d) (2 marks) A further analysis is shown in Figure 9. Is it necessary to look at this? If it is necessary, what do you conclude from it, in the context of the data? If it is not necessary, explain briefly why not.

My answer: The interaction was not significant, so we need to re-fit the model with the interaction removed, which is being done here.

This output tells us that neither of the main effects is significant either: that is, there is no significant effect of age on `FBS_change`, and there is no significant effect of treatment on `FBS_change` either. That is to say, any trends that we observed on the plot in Figure 31 are just chance.

If the interaction *had* been significant, that would have been our conclusion, and we would not have needed to look here. (It is possible to have a significant interaction with a significant, say, treatment effect; this could happen if the insulin treatment was consistently better, but the *amount* by which it is better depends on `Age`. This would show up on the plot by the blue line being consistently (and significantly) higher than the red one, but the lines not being parallel.)

Marking note: if you (erroneously) interpreted the main effects in (c) (for which you lost a point there), it seemed logical that you would then conclude that looking at Figure 9 was a waste of time, for the reason that you didn't (by that logic) need to look at the main effects again. Thus, I checked to see whether you had lost a point in (c), and if you said this in (e) you got full credit for (e).

(e) (2 marks) More output from the analysis of the previous part is shown in Figure 10. Is it helpful to look at this? If it is helpful, what do you conclude from it, in the context of the data? If it is not, explain briefly why not.

My answer: This is a tidy version of the `summary` output from the previous analysis, with the P-values removed (since we already looked at those).

We concluded in the previous part that there are no significant main effects, so these coefficients are not significantly different from zero, and there is therefore nothing to say about them. So it is not helpful to look at these.

(If, say, `Age` had had a significant effect, this output would be interesting because it would tell us how much fasting blood sugar changes per one-year increase of age; a positive slope means that it increases as age increases.)

Marking: two points for saying that it is not helpful to look at these because of non-significance, zero for almost anything else. The only reason I can see for looking at these coefficients at all is if you are thinking about what might happen with a larger sample size, but even then, the non-significance means that if you were to collect more data, what would happen then is pretty much impossible to guess given what happened here. There is just too much scatter. (I didn't give any points for discussion of what the coefficients mean, because I think *here* there was no value in trying to interpret those numbers. If they had been significantly nonzero, this would have been exactly the *right* thing to do.)

(f) (2 marks) Compare your conclusions from Figure 31 and from parts (c) through (e). Describe briefly how they are consistent or inconsistent.

My answer: This depends, of course, on what you concluded before.

If you found something on the plot, your conclusions will be inconsistent with the analysis. If you like, you can walk your conclusion from the graphs back a bit and say something like "but there is a lot of random variability", or that the apparent trends on the graph are just chance.

If you had the insight to realize that those apparent trends on the plot were likely just random

chance, your conclusions will be consistent, and then you can mention that this is the kind of conclusion you expected to draw, or something like that.

There is a lot of scatter (random variability) on Figure 31, which supports the idea that there is no real upward trend and neither line is really different from being flat. The apparent upward trends and different slopes are just chance.

With this much random variability, either the trends would have to be a lot more upward to be significant, or we would have to have a lot more data to have any hope of finding a significant result.

What I am looking for here is some insight connecting the graph with the conclusions you drew. This is one of my "are you surprised?" questions. I wanted to bring you back to the graph now that you have some actual inferential answers, and have you think about everything together. (This is probably going to be hard to mark, but it's the kind of thing I think you as a statistician should be considering.)

One student described how Figure 31 "gave us an illusion", which I thought was a really nice way to say it.

I realize now that I could have made the question out of 3, and asked you to speculate on *why* you thought the graph and (c)–(e) were inconsistent, if that's what you found. Some people did this anyway, and I wanted to give them an extra mark, but that would be unfair on the people who answered the question as asked (highlighting consistency or inconsistency between graph and (c)–(e)), so I didn't.

4. What might influence a person's ability to remember a list of words? One possibility is that if the words on the list are related in some way, it would be more difficult to remember which specific words were on the list. For example, if the list contained "baseball", "soccer" and "hockey", you might have trouble remembering whether the list also contained "football" or "basketball", but if there was only one sport, you might more easily remember which one it was.

   A study was conducted in which subjects were presented with three different lists of words at three different times. The three word lists contained the same number of words and the same mixture of common and rare words, but varied according to whether, and how, the words were related to one another. Specifically, subjects had to learn word lists of these different types:

   - `unrelated`: the words are unrelated to each other.
   - `semantic`: the words have similar meanings (to other words on the list).
   - `phonological`: the words have similar sounds (to other words on the list).

   The order in which a subject was given the word lists was randomized.

   The response variable was the number of words on each list that the subject successfully remembered. The data are shown in Figure 11. (There are sixteen subjects; the subjects with IDs 10 and 11 withdrew from the study.)

   (a) (2 marks) Why do you think that each subject was given the word lists in a random order?

   > **My answer:** There might be an association with order, for example, the subjects might learn words better on the later lists because of practice (or they might learn the later lists worse because of being tired). Any plausible association with order is good here. The best answers mentioned something specific that could cause bias; if it seemed plausible, I was good with it.
   >
   > The words *within* each list were probably also given in a random order, but I wasn't asking about those; I was asking why the three *lists* were given in a random order.

   (b) (2 marks) What is it about the design of this experiment that makes a repeated measures analysis necessary?

   > **My answer:** Each subject did all three word lists, and thus produced three different numbers of words remembered (one for each list). Thus we have repeated measures for each subject. Another way to say this is that the three measurements for each subject are likely to be correlated, because they are on the same subject.
   >
   > If you got at one of these, I was good. I think the easiest one was "there was more than one measurement for each subject". Copying a list of properties of a repeated measures design from your notes without any critical analysis of the crucial issues *for these data* most likely got you only one mark.
   >
   > The contrast is with independent observations, where each subject only gets a randomly-chosen *one* of the word lists. Because subjects are likely to differ from each other, this would be an inferior way to design the study; the variability between subjects gets mixed up with the random variability (whereas in the actual study design we are controlling for differences between subjects and only the random variability is left).

   (c) (2 marks) Some analysis is shown in Figure 12. In the second line of code (the one with `lm` in it), why did I need the number `1` after the squiggle? Explain briefly.

   > **My answer:** There are no "between-subjects factors" here, because all the subjects did the exact same thing (they were all given three word lists to remember). This line is where the

Question 4 continues… This page: _____ of possible 6 points.

between-subjects stuff goes. So I have to supply the `1` to say "just an intercept" or "there is nothing here".

An example of a between-subjects factor in this kind of context could be that some of the subjects were given some training in how to remember word lists, and the others were not. Our data frame would then have a column `training` with values `yes` and `no`, one value for each subject, and the `lm` line would be `response`-squiggle-`training` instead.

"There are no other explanatory variables" was just about enough to get the two points, but the word "other" is important; an answer of "there are no explanatory variables" needs something to say how you know that. (By itself it's one point.) For example, you can refer back to the data to say that the only other variable is `id`, and that does not belong in the model because all the subjects do the same thing. (If we had used a mixed model here, I would have arranged the data in long format, with each `id` appearing three times, once for each word list, and I would have used `id` as a random effect.)

There were some really good answers getting at the within-subjects and between-subjects thing.

(d) (3 marks) What do you conclude from the analysis in Figure 12, in the context of the data?

My answer: The P-value is (slightly) greater than 0.05, so there is not quite a significant effect of the within-subject factor. This is normally time, but in this case it is type of word list: it is saying that there is not (quite) a significant difference between the number of words recalled in word lists of the three different types.

This was probably a disappointment to the researchers; they would probably have liked to have been able to show that the relatedness of the words in the word list made a difference to the number of words recalled.

I expect you to be able to say not only that something does not have a significant effect, but also what that "something" is.

I am not picky about α; if you want to compare your P-value with 0.10 and declare this significant, go ahead; or if you stuck with $\alpha = 0.05$ you could use words like "almost significant" or "marginally significant" to describe this one.

(e) (1 mark) A spaghetti plot is shown in Figure 13. Some of the points on the plot are joined by lines. Which ones?

My answer: The ones that belong to the same *subject*.

There's a couple of ways to get at this: either you can remember that this is what normally goes on a spaghetti plot, or you can look at the code above the plot and see `group=id`, which means that the `geom_line` will draw lines between observations with the same `id`, that is, that belong to the same subject.

This was a pretty easy one to mark because most people either got the idea or didn't.

In case you are wondering, I didn't add any colour to this spaghetti plot because there was no between-subject factor like a treatment that some of the subjects received. That is normally what is distinguished by colour on a plot like this. (I could also distinguish the subjects by colour, which would require 16 different colours, and then we get into the issue of distinguishing that many colours. So I didn't do that. If you thought I should have done, that would be fair comment.)

(f) (2 marks) In the code above Figure 13, I used `gather`. Why did I need to do that? I am looking for an answer specific to this task, not a general description of what `gather` does.

My answer: For the plot, I needed all the numbers of words recalled in *one* column, and I needed them labelled by what kind of word list the number recalled came from. This is exactly what this *gather* does. That's a two-point answer.

A one-point answer would be "you have wide format and you need long format", or "to make the data set into long format". This gets at what I am doing, but not why I need to do it. I circled the "why" in the question if I didn't feel you talked about the "why I need to do it" enough.

I find thinking in terms of "long format" and "wide format" helpful. If you say that we have wide format but we need long format *to make the plot*, that's two points becaues it gets at everything.

(g) (2 marks) Compare what you conclude from Figure 12 and Figure 13. Do the conclusions seem to be consistent or inconsistent? Explain briefly.

My answer: You can come to almost any conclusion here, as long as you are able to support it.

The analysis said that there is no difference in (mean) recall between the three types of word list. On the spaghetti plot, that would be consistent with the lines going (on average) straight across. Decide for yourself whether that's what you see; if it is, the conclusions are consistent, and if not, they are inconsistent.

My feeling is that the lines between `phonological` and `semantic` go sometimes up and sometimes down, so there is not really any difference in recall between those. With `unrelated`, though, compared to `semantic`, the lines seem to go more up than down, and so I wouldn't have been surprised to see a significant effect of type of word list. That would be an argument in favour of the results being inconsistent from the two Figures.

I would resolve that inconsistency one of two ways: (i) the lines that go down on the spaghetti plot, even though there are only three of them, go down more sharply than the upward-sloping lines go up, so that in terms of mean, there is not much difference; (ii) the P-value, though not significant, is close to 0.05, and so the suggestion of mostly upward lines here is indicative of a result that is almost significant. (If the pattern had been a little clearer, it would have *been* significant.)

If you said something logically sound, whatever it was, you probably got two points here. There were, as you might guess, a lot of ways to get full marks.

There is one other oddity here, which is that the variable on the x-axis of a spaghetti plot (usually time) makes the most sense if it is ordered (and then the ones neighbouring in time are joined). Here, what I called `relatedness` is nominal rather than ordered; the plot would have looked different if the three different types of word list had been arranged in a different order, which they could reasonably have been.

5. A rootstock is part of a plant, often a root, from which new above-ground growth can be produced. This can be done by "grafting" one or more cuttings from another plant onto the rootstock in such a way that they will grow into healthy plants. In the data shown in Figure 14, six apple tree rootstocks are used. On each of these, eight apple trees are grown. The individual trees are labelled in the column called `row`; there are 48 of them. The idea is that trees grown on the same rootstock should have consistent properties, and different properties than those grown on other rootstocks. To see whether that is true for these apple trees, four variables are measured for each tree:

   - `girth4`: the girth (circumference) of the tree trunk after 4 years (mm, multiplied by 100)
   - `extension`: extension growth at 4 years (metres)
   - `girth15`: the girth of the tree trunk after 15 years (same units as the girth after 4 years)
   - `weight` of above-ground portion of tree after 15 years (in pounds, multiplied by 1000).

   The purpose of multiplying by the different numbers was to get values that were comparable in size.

   Our aim is to see whether the trees from the different rootstocks differ on any of the measured variables, and, if so, which ones.

   (a) (2 marks) Explain briefly how MANOVA will help us address at least part of our aim (and which part it will help with).

   My answer: MANOVA will enable us to see whether the rootstocks are more different on any of our variables (or combinations of them) than you would expect by chance. That is, it will enable us to decide whether there are any differences worth trying to find later.

   It won't tell us how the rootstocks differ from each other, but we will take that up later (that's what discriminant analysis does).

   One point for something like "MANOVA is used when we have multiple response variables", which is true, but says nothing about what it will tell us that will be useful here. I'm trying to get you to distinguish between "are the rootstocks different on any of the variables?", which is MANOVA, and "if they are different, *how* are they different", which is discriminant analysis. It's just like ANOVA followed by Tukey ("are the groups different on the one response" vs. "how are they different".)

   (b) (3 marks) Give code to run the MANOVA and to display the results. Give some thought to whether the `rootstock` variable is the right kind of thing. (You may assume that you already have a data frame `rootstocks` as in Figure 14.)

   My answer:

   There are three parts:

   1. Create a response variable out of our four measured variables.
   2. Run the MANOVA, predicting the response from the rootstock *as a categorical variable* (since the numbers don't mean anything as numbers; they are just labels).
   3. Display the `summary` of the results.

   That means you need something like this:

```
response=with(rootstocks,cbind(girth4,extension,girth15,weight))
rootstocks.1=manova(response~factor(rootstock),data=rootstocks)
summary(rootstocks.1)
```

Question 5 continues...

```
##                   Df Pillai approx F num Df den Df    Pr(>F)
## factor(rootstock)  5 1.3055   4.0697     20    168 1.983e-07 ***
## Residuals         42
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

You need to turn `rootstock` into a categorical variable somehow; this seems to be the cleanest way. Or you could define the `factor` version of `rootstock` first, and then use that in the MANOVA. Up to you.

This was the code that produced Figure 15.

If you want to do it the `Manova` way, you need this:

```
library(car)
## Warning: package 'car' was built under R version 3.5.1
## Loading required package: carData
## Warning: package 'carData' was built under R version 3.5.1
##
## Attaching package: 'car'
## The following object is masked from 'package:dplyr':
##
##     recode
## The following object is masked from 'package:purrr':
##
##     some
rootstocks.2a=lm(response~factor(rootstock),data=rootstocks)
rootstocks.2b=Manova(rootstocks.2a)
rootstocks.2b
##
## Type II MANOVA Tests: Pillai test statistic
##                   Df test stat approx F num Df den Df    Pr(>F)
## factor(rootstock)  5    1.3055   4.0697     20    168 1.983e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

having created `response` first, again.

Some notes:

- I'm not a fan of `attach`, but if you want to use it here instead of `with`, it will work, so that's OK.
- This is not repeated measures, so you *do not* need the `idata` and `idesign` stuff. (These are not even four different but related variables, as the word lists were.)
- Some people got the brackets messed up with the `factor`. This I forgave, since the point of the question was to see whether you knew to turn `rootstock` into a factor and had made a reasonable attempt at doing so. (The same thing in STAC32 would be a different story.) There was a clue about the `factor` thing in the next Figure, where I did the discriminant analysis. I did `factor(rootstock)` there, so the suggestion was that I needed it here as well.

(c) (2 marks) How do you know that the discriminant analysis in Figure 16 was worth doing? Explain briefly.

My answer: This is a brief "briefly": because the MANOVA in Figure 15 was significant (had a small P-value). This ought to be consistent with what you said in (a). That is to say, there are some differences between rootstocks for us to find out about.

I wasn't asking a general question about what discriminant analysis does, and I wasn't asking for an interpretation of any LDs (that's coming up later). There was one point available for an answer that hinted sufficiently clearly about the value of a discriminant analysis *here* without mentioning the significant MANOVA.

(d) (2 marks) Looking at Figure 16, how many linear discriminants do you think you should pay attention to? Explain briefly.

My answer:

This is the `svd` part of the output.

The first two singular values are much larger than the other two, which suggests paying attention to the first two LDs, but not the third and fourth ones.

Make a different call if you can justify it; for example, you might say that fourth singular value is much smaller than the other three, and therefore we should look at LD3 as well. Basically, if you make it sufficiently clear that you were looking at the SVD values (by naming them or giving the values), and you make *some* call about which ones were big and which ones were small, I was happy.

This is not asking about how many LDs there *are* (though I often do ask about that: $\min(6-1, 4) = 4$). That kind of question would only be one mark. I want to know how many of them we should be interested in, which is not necessarily all the ones there are. (Since I was being generous, a correct answer of this kind got one mark.)

(e) (2 marks) What would make a tree have a very *negative* score on `LD1`? Explain briefly.

My answer: Looking at `scaling`, only the two girth measurements have any substantial influence over LD1, and a tree will have a negative score on LD1 if these are both *small*.

I was fairly relaxed about this one. If you mentioned all four variables with sizes according to the sign of their loadings, that was fine; if you want to mention extension being large as well as the two girth measurements being small, that was fine. You really ought to pick out the variables whose coefficients on LD1 are *far from zero*, be they positive or negative, rather than only looking for negative coefficients (since a small value on the girth measurements would do it too). The one thing I was less relaxed about is that *you need to talk about at least one of the girth measurements*, or else you only get one point. (Taking the point of view that only `girth15` matters, since its coefficient is "much" bigger in size than the others, was defensible.)

Somebody said that the trees needed to be "ungirthy", which I didn't think was actually a word, but I think it conveys the impression very well!

The Wikipedia article on tree measurement tells you exactly what tree girth is and how it should be measured.

(f) (2 marks) What would make a tree have a large positive score on `LD2`? Explain briefly.

My answer: `girth15` has a positive coefficient on LD2, and `weight` a negative one, so LD2 will be large and positive for a tree that has a large `girth15` and a small `weight`.

This is the same idea as the previous part. If you have picked out something between "large `girth15`" and that plus small on weight, or small on all three of the others, I was happy. I probably ought to have directed you a bit more (like I do on the principal components question below, where I ask for the most important *two* variables on each component). So this ended up being easier (for you) than it might have been otherwise, since pretty much anything not actually wrong would get you two points.

My feeling is that the coefficients on LD2 for `girth4` and `extension` are close to zero, so that LD2 only depends (negatively) on weight and (positively) on `girth15`. But you're entitled to make a different call.

(g) (3 marks) Look at Figure 17. Find an observation (a tree) that is misclassified. Which row number is it in the data frame? Was this tree's rootstock close to being gotten correct? Explain briefly, showing that you know what the columns of the data frame represent.

My answer: An example answer: the tree in row 3 is misclassified, since it is actually from rootstock 1 but was predicted to be from rootstock 6. The posterior probability of it being rootstock 6 is 0.319, and for (correct) rootstock 1 is 0.228, only a little lower. So it was close to being gotten correct.

More discussion: the columns are these:

- `row`: the row number, for ease of identification
- `obs`: the actual rootstock that tree was grown from
- `pred`: the predicted rootstock, based on the measured variables
- `1` through `6`: the posterior probabilities of a tree being from each of those rootstocks, based on the measured variables.

A tree is misclassified if `obs` and `pred` are different. For example, the tree in row 3 was actually from rootstock 1 but was predicted to be from rootstock 6. (I don't mind which tree you pick,

but tell me what row it comes from, and make sure the observed and predicted rootstock are *different*.) Looking at the posterior probabilities for the tree in row 3, it has probability 0.285 of being from rootstock 1, 0.228 of being from rootstock 4, and probability 0.319 of being from rootstock 6 (and much lower probabilities of being from any of the other rootstocks). Since these highest posterior probabilities are very close to each other, there is considerable doubt about which rootstock it came from, and if we had been luckier, we would have gotten this one correct. (Or, the posterior probability of being from rootstock 1 was not that small.)

The answer you get will depend on which row you pick. Row 7, for example, is actually a 1, but has posterior probability 0.826 of being a 4. This was not close to being gotten correct, since the posterior probability of being a 1 was much lower.

I want you to state or (strongly) imply that you know that `obs` is the actual rootstock, `pred` is the predicted rootstock, and the numbered columns are the (posterior) probabilities of that tree having come from the rootstock with that number. There's no need to go to the detail I did, but I want to be confident that you know what you're looking at.

If I had made a `data.frame` out of the posterior probabilities, the columns would have come out labelled `X1` through `X6`, which I thought was less clear, so I went this way. I would have liked not to have scientific notation in the column for rootstock 6, but we get what we get.

The "backticks" in the names of the numbered columns are there because a plain number is not a legal column name, so we have to "quote" them to use them as column names. `data.frame` *makes* them legal column names by putting an `X` in front of them, but the Tidyverse doesn't change anything without your permission.

(h) (2 marks) Look at Figure 33. This figure is in colour, at the end of the Code and Output. The overall misclassification rate is around 50%. Is that about what you would expect, looking at this plot? Explain briefly.

My answer: The plot shows that the rootstocks are somewhat mixed up, but there are clusters that are predominantly one rootstock, which will be gotten right. So it makes sense that some of the trees will be predicted correctly and some incorrectly, and a misclassification rate around 50% seems about right.

The MANOVA (by the way) shows that this plot is a lot more clustered than you would expect by chance, but whether it's clustered enough to give useful predictions is another question. If there were no clusters at all, the misclassification rate would be what you would get by guessing, $5/6 = 83\%$. Here, it is a lot better than that, so there must be *some* clustering.

I am looking for something that suggests the misclassification rate won't be very low (points mixed up) and something suggesting that it won't be very high (some clustering). You should say both of these things for two points.

(i) (2 marks) In Figure 33, find a tree that scores *low* (ie., very negatively) on LD1. Find that tree in Figure 14. Are its variable values high and/or low as you predicted for such a tree in part (e)? Explain briefly.

My answer: Again, the answer you get will depend on the tree you pick. I pick tree 46, way over on the left of Figure 33. In part (e), I said that a tree with a very negative score on LD1 will have small values for both `girth4` and `girth15`. Look at the data in Figure 14 and try to judge where tree 46 (or tree 29, if you picked that one) stands relative to the others on these two variables. (If you said something about `weight` and `extension` in (e), you ought to assess those as well, but you may find them not so remarkable.)

It looks to me as if tree 46 is among the smallest on `girth4` and the smallest of all on `girth15`. So it is not at all surprising that it scores very negatively on LD1. (This tree's extension is on

the low side and its weight looks like the lowest of all, so if you considered these variables, you should be surprised by these.)

Tree 29 is done the same way: it is one of only a handful of trees less than 1 on `girth4`, and its `girth15` value is about the third or fourth smallest. The story is again consistent with what we said before. (The extension is lowish and the weight is on the low side, again showing that what matters are the loadings far away from zero.)

I don't need detail; I'm looking for a qualitative impression. Words like "among the smallest" or "one of the smallest" are precise enough for me.

I realize that it would have made your life easier to ask you, for example, to look only at the two girth measurements and ask you to comment on their smallness. But I didn't do that.

If you asserted that a variable's value for your chosen tree was large or small and *it wasn't*, expect to lose a point. (I checked.) If you don't name a tree, don't expect to get any points. (And make sure you pick a tree on the *left*, not at the bottom, which would be low `LD2` instead.)

6. A chemical analysis was carried out of 178 wines from a certain region in Italy. Thirteen different quantitative variables were measured for each wine. These variables are listed in Figure 18. Some of the data is shown in Figure 19. (Each wine has a text ID beginning with V, since the Italian for "wine" is "vino".)

   (a) (2 marks) Why is it not possible to run discriminant analysis on these data? Explain briefly.

   My answer: There are no known groups. Or, we have to find the groups ourselves (via a cluster analysis).

   Saying what a discriminant analysis does is only a starting point. You need to go beyond "predicting groups" to say that there are no groups here, and then you have an answer.

   I said "not possible" rather than "not desirable"; for example, in another situation, we might have groups but a non-significant MANOVA, and then it would be kind of dumb to run a discriminant analysis, but it would still be possible. This is not what we have here, though, so a discussion of this kind does not answer the question.

   (b) (2 marks) Why would a K-means cluster analysis be more appropriate for these data in their current form than a hierarchical cluster analysis?

   My answer: We have measurements on variables, which a K-means analysis can use, and not dissimilarities between wines, which we would require for something like Ward's method.

   You can (if you must) copy from your notes what K-means and hierarchical clustering do, but if you want the second point you need to display some kind of understanding of what data we have *here*: no dissimilarities, so no hierarchical analysis (or, we have measurements on variables, so we can go ahead and do K-means).

   (c) (2 marks) A calculation is shown in Figure 20. Explain briefly what the calculation does and why it is necessary.

   My answer: It standardizes all the numeric columns. This is necessary because they are measured on (very) different scales, and we want to treat each variable equally in the cluster analysis. Note that you need to know precisely what `scale` does.

   There are these things to say for a complete answer:

   - Only work on the numeric columns
   - Standardize all of those
   - This is because the variables are measured on different scales and we want to treat them all equally (or equivalent).

   I gave two points for a complete or almost-complete answer. If you mentioned one or two of the three things, you got one point. I definitely wanted to see the "why" part if I was going to give you two. (This is another one of those questions that could have been out of three, but I don't have a good reason to change it, so two it stays.)

   (d) (2 marks) Explain briefly why it is necessary to look at a scree plot before running the K-means analysis.

   My answer: The input to `kmeans` includes the number of clusters, and the scree plot will tell us how many clusters we should use.

   If you're going to be complete about this, we need to know the number of clusters in advance

Question 6 continues… This page: _____ of possible 8 points.

because it needs to be input to `kmeans`; we don't have another way to figure out how many clusters to use. But I will certainly take "because it tells us how many clusters to use" or similar. (Once again, if this were a three-pointer, I might be more picky, but it isn't, so I'm not.)

Don't get confused between this and factor analysis (or principal components). They all use scree plots, but they do so for different purposes. You might say that cluster analysis is "grouping individuals" and factor analysis is "grouping variables". (The latter is used to summarize a potentially large number of variables with many fewer.) But this is not factor analysis, so don't talk about that here.

If you want to talk about elbows here, go ahead. In fact, you can talk about pretty much whatever you like. I didn't punish you for saying something incorrect like "elbow minus 1". It's all OK, as long as you use the words "number of clusters" or "how many clusters" somewhere. That's the most important thing to say.

(e) (4 marks) Suppose you have a function called `ss` that accepts as input a number of clusters and a data frame, and returns as output the total within-cluster sum of squares for a K-means cluster analysis with that many clusters on the input data frame.

Give code to obtain a data frame containing the total within-cluster sum of squares for each number of clusters from 2 through 20 inclusive, using the appropriate data set, and use that data frame to make a scree plot.

My answer: This is the `map` way (it is what I used to produce the scree plot in Figure 21):

```
d = tibble(clusters=2:20) %>%
    mutate(wss=map_dbl(clusters,ss,wines2))
d
```

```
## # A tibble: 19 x 2
##    clusters   wss
##       <int> <dbl>
##  1        2 1649.
##  2        3 1271.
##  3        4 1169.
##  4        5 1102.
##  5        6 1038.
##  6        7  974.
##  7        8  920.
##  8        9  875.
##  9       10  839.
## 10       11  800.
## 11       12  773.
## 12       13  745.
## 13       14  718.
## 14       15  694.
## 15       16  681.
## 16       17  651.
## 17       18  634.
## 18       19  620.
## 19       20  603.
```

Don't forget to use the standardized `wines2`. That was kind of the reason we standardized it in the first place.

I "gave" you the function `ss` so that you could use this technique (which you could lift from the lecture notes and change some names). You could also do it the "functionless" way, replacing the function with a squiggle and the definition of a function. This requires a little care, though, since my function also got rid of that column of IDs, so I have to do that here as well:

```
wines3 = wines2 %>% select(-id)
dx = tibble(clusters=2:20) %>%
    mutate(wss=map_dbl(clusters,~kmeans(wines3,.,nstart=20)$tot.withinss))
dx
```

```
## # A tibble: 19 x 2
##    clusters   wss
##       <int> <dbl>
##  1        2 1649.
##  2        3 1271.
##  3        4 1169.
##  4        5 1099.
##  5        6 1033.
```

```
##  6        7  973.
##  7        8  927.
##  8        9  877.
##  9       10  837.
## 10       11  805.
## 11       12  771.
## 12       13  744.
## 13       14  722.
## 14       15  698.
## 15       16  670.
## 16       17  653.
## 17       18  639.
## 18       19  617.
## 19       20  602.
```

`kmeans` takes a data frame, then a number of clusters, then any additional stuff like `nstart`. However, I'd rather do this in two steps, thus:

```
wines3 = wines2 %>% select(-id)
dy = tibble(clusters=2:20) %>%
    mutate(km=map(clusters,~kmeans(wines3,.,nstart=20))) %>%
    mutate(wss=map_dbl(km,"tot.withinss"))
dy
```

```
## # A tibble: 19 x 3
##    clusters km             wss
##       <int> <list>       <dbl>
##  1        2 <S3: kmeans> 1649.
##  2        3 <S3: kmeans> 1271.
##  3        4 <S3: kmeans> 1169.
##  4        5 <S3: kmeans> 1095.
##  5        6 <S3: kmeans> 1033.
##  6        7 <S3: kmeans>  972.
##  7        8 <S3: kmeans>  927.
##  8        9 <S3: kmeans>  879.
##  9       10 <S3: kmeans>  837.
## 10       11 <S3: kmeans>  797.
## 11       12 <S3: kmeans>  773.
## 12       13 <S3: kmeans>  742.
## 13       14 <S3: kmeans>  716.
## 14       15 <S3: kmeans>  695.
## 15       16 <S3: kmeans>  675.
## 16       17 <S3: kmeans>  656.
## 17       18 <S3: kmeans>  635.
## 18       19 <S3: kmeans>  617.
## 19       20 <S3: kmeans>  602.
```

The column `km` contains the whole `kmeans` fit for each number of clusters, and then from that I extract the total within-group SS. I think this makes my process clearer, as long as you understand my process: the first `map` obtains all the K-means objects, and then the second `map_dbl` says "for all the things in `km`, give me the piece called `tot.withinss` from each one". Tricky, but very powerful.

If you want to do it using a loop, you can, but there's some extra housekeeping to do, something like this:

```
clusters=2:20
```

```
wss=numeric(0)
for (i in clusters) {
    wss[i]=ss(i,wines2)
}
d2=tibble(clusters=1:20,wss)
d2
```

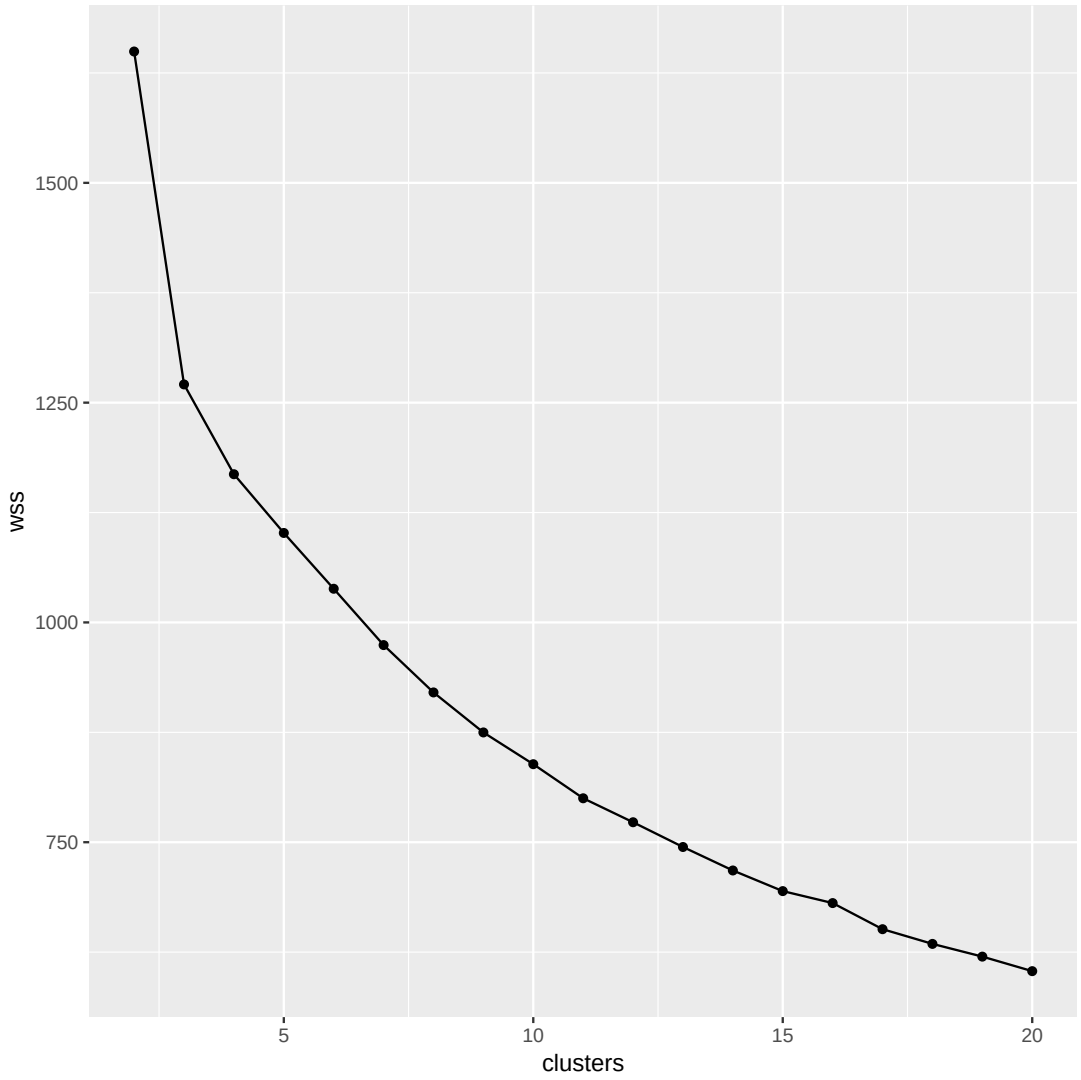
```
## # A tibble: 20 x 2
##    clusters   wss
##       <int> <dbl>
##  1        1   NA
##  2        2 1649.
##  3        3 1271.
##  4        4 1169.
##  5        5 1095.
##  6        6 1033.
##  7        7  973.
##  8        8  919.
##  9        9  877.
## 10       10  835.
## 11       11  800.
## 12       12  771.
## 13       13  739.
## 14       14  722.
## 15       15  695.
## 16       16  674.
## 17       17  655.
## 18       18  639.
## 19       19  626.
## 20       20  609.
```

`wss` gets an entry for 1 cluster (that never gets filled), so you have to allow for this when making the data frame. The `map` way is much easier since you don't have to worry about this.

Three points for getting this far, somehow.

Finally, to get a scree plot, plot `wss` against `clusters`, joining the points by lines. Use the data frame you got above, either one:

```
ggplot(d,aes(x=clusters,y=wss))+geom_point()+geom_line()
```



This last part is a giveaway; even if you don't know how to do the rest of it, say "if I had the data frame, I would do" and then give the code to make the graph, and get one point. (It actually doesn't matter if your data frame has a missing value in it, as it will if you do it my loop way; you'll get a warning when you make the plot, but the plot itself will be fine.)

I had a slightly different arrangement of random number seed here, so I came out with a slightly different plot. (The clearest elbow, on either of them, is at 3 clusters; where there are other elbows depends on the random number seed.)

Extra: my function `ss` was this:

```
ss
## function(i,d) {
##   km = d %>% select_if(is.numeric) %>%
##     kmeans(i,nstart=20)
##   km$tot.withinss
## }
## <bytecode: 0x562d70e7e130>
```

It grabs only the numeric columns before running `kmeans` (since `kmeans` only accepts numeric

columns).

(f) (2 marks) A scree plot is shown in Figure 21. What do you conclude from this plot? Explain briefly.

My answer: This tells us how many clusters we should use, by taking the number of clusters at an "elbow". I see a clear elbow at 3 clusters, and maybe little elbows at 17 and maybe 15 clusters (perhaps 11 too). So I would recommend 3 clusters, but I would also accept 11 or 15 or 17 clusters (or any other place where you could justify there being an elbow). The point here is not your answer, but your justification of it; there has to be a noticeable elbow at the number of clusters that you assert, and you have to say that.

Beware of "backwards elbows", ones that point up rather than down, like the one at 10. This was about the only choice of number of clusters that I disagreed with; if it could conceivably be seen as a (downward-facing) elbow, I was good with it.

Couple of other things:

- This is not principal components, so the number of clusters is where the elbow is (not minus 1). Expect to lose a point if you subtracted one. (The unlucky souls who picked an elbow at 10 and then subtracted one thus lost *both* points.)
- Don't give me two answers. If one of them is wrong, I have to mark your answer wrong. At the very least, pick one as your "final answer". (You are fighting the odds by giving me two answers; if they're *both* defensible, you get the points; otherwise, you don't.)

You might reasonably say that 3 clusters is "too far up the mountain", since we want the total within-cluster sum of squares to be small. That would be a good argument for preferring 15 or especially 17 clusters. (With 178 wines, we could justify using this many clusters, but the picture later would be harder to interpret. This is why I went for 3.)

(g) (3 marks) I obtained a 3-cluster solution from K-means. (This may or may not be a good number of clusters.) To make a graph of this solution, I ran a discriminant analysis, using the three clusters as known groups. I also drew a biplot. This, and the code to produce it, are shown in Figure 32.

According to the biplot, what do you think is the most important way in which my clusters 1 and 3 differ? Explain briefly.

My answer: Cluster 1 is the red one on the left (with a typically negative LD1 score) and cluster 3 is the blue one on the right (with a typically positive LD1 score). That's one point if you say that.

To go further and identify which variables distinguish the clusters, look for variables with long arrows left and right (since the clusters are arrayed left and right). `flavonoids` is the most obvious one (cluster 3 is high and cluster 1 is low), but you could also pick `colour` (1 high and 3 low), or `proline` or `od280` (3 high and 1 low like `flavonoids`, since they point in almost the same direction). I don't think anything else points left and right to the same degree as these.

Note that the cluster at the point of the arrow is high on that variable and the cluster that the arrow points away from is low on that variable. I want you to say which cluster is high and which is low on your chosen variable (and I really want you to pick *one* since I asked for the "most important way" in which the clusters differ).

This is not quite the same as picking out which arrows point towards a cluster and saying that the cluster is high on those things (or that these things are most important in the cluster). This is because it doesn't say what the cluster is *low* on, and so is an incomplete description of

what is going on. Thus, expect two points for a discussion like this.

7. Five people were trying on ski boots in a store on a Friday evening in January. They were each asked about factors which might influence which ski resort they would go to. The questions on the questionnaire were designed to assess:

   - `cost`: cost of ski ticket
   - `lift`: speed of ski lift
   - `depth`: depth of snow
   - `powder`: moisture of snow (drier snow, called "powder", is better for skiing on).

   Some people might attach greater importance to some of these variables, and some might attach greater importance to others.

   After the questionnaires were completed, a score on each of the above variables was computed for each person, with a higher score indicating greater importance on that variable for that person. The data are shown in Figure 22. Our aim is to see whether these four variables can be summarized by fewer variables.

   (a) (2 marks) A principal components analysis is done in Figure 23 and a scree plot is obtained in Figure 24. What do you conclude from the scree plot? Explain briefly.

   My answer: We are looking for an elbow, a "pointing-downwards" one. The clearest one is at 3, so we should take $3 - 1 = 2$ components. (The corner at 2 is not an "elbow" as we have defined it, so make sure your justification for 2 components is clear.)

   Two components is also supported by the output in Figure 23: two components explain nearly 99% of the variability, and it's hard to improve usefully on that. But I want the major part of your reasoning to come from the scree plot. At the very least, you should conclude *something* from the scree plot. I was OK with using the variance explained to get a number of clusters and using the scree plot to support that choice, even though my intention was for you to do it the other way around.

   That is *not* an elbow at 2, because it points the wrong way (it points up, and it needs to point down). Thus there is no justification for one component. (You will lose a point if you say there is an elbow there.)

   (b) (2 marks) In Figure 25, which two variables are the most important part of each relevant component? Explain briefly.

   My answer: My take is that component 1 is `depth` and `powder`, and component 2 is `cost` and `lift`, with opposite signs. (I changed the question from how I first wrote it, since you might otherwise say that all the variables belong to both components. It is a little "mushy" as it often is for principal components. Rather like wet snow.)

   You ought to talk about the two most important variables in each of however many components you said in the previous part (neither more nor less), but if you talked correctly about more components, I was OK with that.

   There was one point for basically any relevant comment here; you could make more or less any number of errors, and as long as there was something correct there, I'd give you one point.

   (c) (2 marks) A biplot is shown in Figure 26. How does this support your answer to the previous part? Explain briefly.

   My answer:

   `depth` and `powder` (component 1) point mainly left, and `lift` and `cost` (component 2) point mainly up and down. It is not perfect, since each component partly contains all the variables.

Question 7 continues... This page: _____ of possible 6 points.

I want to see something about which directions things are mostly pointing in: the more left-right arrows are the things in component 1, and the more up-down arrows are the things in component 2. Or something implying that, for example "`depth` and `powder` point in the component 1 direction" and the same for `cost` and `lift`.

I marked this part according to whatever you said in the previous part, so that if you have complete support for whatever you said there, you get two points here (and a note like "given (b)" on your exam).

I wasn't asking about the subjects S1 through S5 in this question (since I had already asked about that kind of thing in the trees and rootstocks question). However, if you said something sensible about one or more of the subjects here, you got one point. In the same way, if you said something correct but incomplete, or you otherwise didn't convince me that you really knew what was going on, you got one point here.

There were a lot of two-point questions on this exam that might perhaps have been three points on a different exam. On these, one point out of two covered a wide range between "something relevant" and "not completely correct".

(d) (2 marks) What do the variables in the first principal component have in common with each other? What do the variables in the second principal component have in common with each other?

My answer: I'd say that `depth` and `powder`, which make up component 1, both have to do with the snow at the ski resort, and `cost` and `lift`, which have to do with component 2, have to do with the ski resort itself.

I won't be too picky about this, since I can't expect you to know too much about skiing, but if you can come up with something plausible, I'm happy. I am looking for something descriptive, something *about* the variables.

There were some really good answers here.

Extra: the loadings on `cost` and `lift` have *opposite signs*, which means that component 2 will be high (and positive) for a person if one of those variables is very important and the other one is very unimportant for that person. If you look at the biplot, coming up later, you'll see that people `s1` and `s4` thought that the lift speed was important and the others thought that the cost was (at least somewhat) important.

If you thought we should look at component 3 as well, this depends on `cost` and `lift` with the *same* sign. Someone who scores high on component 3 thinks that `cost` and `lift` are both important or both unimportant. (In the data, one of these tends to be high and the other one low, which is why this component is not important.)

(e) (2 marks) The biplot in Figure 26 suggests that factor analysis will produce a clearer result than principal components. How? Explain briefly.

My answer: Factor analysis is able to rotate the principal components to align them with the axes, for example making `lift` and `cost` point more clearly up and down (and not left and right at all), and `powder` and `depth` more clearly left (or right), and not up or down at all. (Some people said specifically which way to rotate and by how much to accomplish this, which I thought was excellent.)

For two points, I wanted something that *this* Figure told you: that is, some *application* of your knowledge (something that you wouldn't have known had you not looked at this Figure). This is not accomplished by giving me some random theory about why factor analysis is better than principal components *in general*, particularly not if it is blindly copied from your notes with no evidence of understanding. I had some sympathy for issues like "reproducibility" (factor

analysis, unlike principal component, has a hypothesis test for "n factors are sufficient", which gives some kind of assurance that these factors will show up again in another analysis of similar data), so if you said something like that, I may have given you a point, depending on how convincingly it was said.

Extra: according to the biplot, skiers 1 and 2 should have high values for `powder` and `depth`, and skier 4 should have a high value for `lift` and a low one for `cost`.

The data were:

```
skiers
```

```
## # A tibble: 5 x 5
##   skier  cost  lift depth powder
##   <chr> <dbl> <dbl> <dbl>  <dbl>
## 1 s1       32    64    65     67
## 2 s2       61    37    62     65
## 3 s3       59    40    45     43
## 4 s4       36    62    34     35
## 5 s5       62    46    43     40
```

That's exactly how it came out.

(I was going to run the factor analysis, but I was told that 2 factors was "too many for 4 variables".)

8. 1681 residents of twelve areas of Copenhagen (in Denmark) were classified according to four categorical variables:

   - `Sat`: Satisfaction with their housing; low, medium or high
   - `Infl`: Feeling of influence on building management; low, medium or high
   - `Type`: Type of housing; tower, apartment, atrium or terrace
   - `Cont`: Degree of contact with neighbours, low or high

   There is also a column `Freq` showing how many residents fell into that combination of categories.

   The data set (some) is shown in Figure 27. The data set reached me as a `data.frame` so I used `as_tibble` to display it nicely.

   Our aim is to discover which of the other variables are associated with satisfaction, and if so, how they are related.

   (a) (3 marks) Some analysis is shown in Figure 28. There is more than I showed (which is omitted). Describe the general process by which I got from `housing.1` to `housing.5`, and why I stopped at `housing.5`.

   My answer: The procedure is:

   1. Look at the `drop1` table, and remove the least significant term with the highest P-value, if not significant. Otherwise stop.
   2. Fit the model with that term removed (I used `update`).
   3. Produce a new `drop1` table, and go back to step 1.

   Continue until everything in the `drop1` table is significant (which is why I stopped at `housing.5`).

   There are different ways to explain this; anything equivalent is good. There is one point for the "how do I know when to stop", and two points for describing the process in some coherent fashion. For the latter, I want to see something about *why* the terms were removed (the ones that were removed). You can if you want work out which terms must have been removed, but saying that these terms were removed is only part of the story; they were removed *because they had the highest non-significant P-value in the* ***`drop1`***, and that's the important thing. Like a backward elimination in a multiple regression: we're removing something clearly non-significant and then re-evaluating, in the hope of building a model that provides a "parsimonious" description of what's going on. In this kind of modelling, that means (hopefully) coming up with a model that can be understood by looking at a small number of relatively simple tables. You might disagree that we have done that here (especially when you start thinking about (d)), but we've done the best we can and the structure of the data just *is* that complicated.

   (b) (1 mark) Explain very briefly why I am not interested in the `Infl:Cont` and `Type:Cont` terms in Figure 28, even though they are significant.

   My answer: They do not tell me about associations with satisfaction (because they do not contain `Sat`), which is our primary interest.

   I made this only one mark to stop you from thinking there was something complicated here.

   The issue with this is that there *is* an association between, say, influence over management and contact with neighbours. In other circumstances, this might be important to us, in which case we should stop and study it, but our focus here is on Satisfaction, so we care about things associated with that.

Question 8 continues…

This page: _____ of possible 4 points.

The three-way interaction `Sat:Infl:Type` that we study in (d) *does not* contain `Cont` at all, so `Infl:Cont` and `Type:Cont` are separate associations that could be studied. For example, you might expect that people would have more or less contact with neighbours depending on the type of housing they live in. If you live in an apartment, you would see your neighbours all the time (waiting for the elevator, say), whereas if you live in a townhouse, you would only see your neighbours if you happened to be outside at the same time. You would investigate this with the same `xtabs` and possibly `prop.table` idea that you would use for any association.

(c) (2 marks) Use Figure 29 to describe the relationship between the amount of contact with neighbours and a resident's satisfaction with their housing.

My answer: A resident who has high contact with neighbours is slightly less likely to have Low satisfaction with their housing, and slightly more likely to have Medium or High satisfaction.

There is, said differently, a positive correlation between the amount of contact with neighbours and overall satisfaction. (This is about what you'd expect.)

I didn't ask for a reason here, so "high contact with neighbours is associated with higher satisfaction" is enough. As always, though, giving a reason is like an insurance policy: if your answer is wrong, it gives you a chance at part marks. You can also say that there isn't much of an effect, since the proportions in High satisfaction are not that different, but it *is* significant, so you ought to comment on the effect you see, even if you think it is tiny.

I tried hard to give you 2 here if I reasonably could. Otherwise, an appropriate-looking comment would get you 1.

(d) (3 marks) Use Figure 30 to explain what the significant `Sat:Infl:Type` term tells you about the data.

My answer: I'm leaving this one open to see how you handle it. What I'm after is something explaining the significant association between satisfaction and the *influence-housing type combination*: that is to say, the association is not just between influence and satisfaction, but that association is different for each housing type.

Anyway, down to business.

The choice of `margin` means that the three numbers for each housing type and influence level add up to 1. So this says "given that a resident lives in this type of residence and feels this much influence, how likely are they to express low, medium or high satisfaction?" The idea is to look for a different association between influence and satisfaction for different housing types; that would explain why satisfaction is associated with the `Type-Infl` *combination*.

- Tower block residents generally are likely to be highly satisfied regardless of how much influence they have (somewhat higher if high influence, but there is not a lot in it).
- Apartment residents: for them, it makes a big difference how much influence they have on how satisfied they are. If they have low influence, they are likely to have low satisfaction; if high influence, likely also high satisfaction.
- Atrium residents are less satisfied overall than tower-block residents, but the pattern is the same; satisfaction has only a weak association with influence.
- Terrace-house residents show the same pattern as apartment residents: if influence is low, satisfaction is low, and if influence is high, satisfaction is high.

What I'm looking for you to do is to find two types of residence for which the association between influence and satisfaction is *different*. For example, you can say that the pattern is

different for people who live in tower blocks vs. people who live in apartments. That would be enough to explain why we could not remove that three-way interaction. (This is why this is three marks instead of four: I'm after *some* reason why the interaction needed to stay, rather than a complete discussion of the `prop.table` output like I gave.)

Marking guide: a sensible comment, one point; some sensible discussion that does not get at why this interaction is significant, two points; discussion that (for example) shows how the effect of influence is different for different housing types (or otherwise gets at why satisfaction is associated with the influence-type combination), three points.

Yes, this was a difficult one to finish with.