University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAC33 (K. Butler), Midterm Exam
February 18, 2022

This exam is open book, open Internet. Anything taken from outside the course materials must be cited: that is, if you want any credit for it, you must say where it came from. "Course materials" means my lecture slides, lecture videos, solutions to this semester's assignments, and PASIAS. There is no credit for uncited outside materials, and there may not be full credit even if you cite (if there is a way to solve the problem using materials from this course).
This exam has 20 numbered pages of questions.
In addition, you have an additional booklet of Figures to refer to during the exam. You should keep this open in another window.
This exam is online. You should hand in the output from a previewed R Notebook (or knitted R Markdown document) in HTML format that can be read by the grader. There is no credit for a file that cannot be read. It is your responsibility to check that the file you hand in can be read.
The exam closes at 9:15 pm. If you have an Accessability extension, your contact there will inform you when the exam closes for you. The exam may not be handed in after it closes. It is your responsibility to make sure that you can knit/preview your exam so that you can submit it on time. (It is recommended that you knit/preview your exam frequently so that you receive the earliest warning of any problems.)
You will be allowed multiple attempts to hand in your work. The latest readable attempt will be the one graded.
The maximum marks available for each part of each question are shown next to the question part.
You may assume throughout this exam that the code shown in Figure 1 of the booklet of Figures has already been run.
The University of Toronto's Code of Behaviour on Academic Matters applies to all University of Toronto Scarborough students. The Code prohibits all forms of academic dishonesty including, but not limited to, cheating, plagiarism, and the use of unauthorized aids. Students violating the Code may be subject to penalties up to and including suspension or expulsion from the University.

Question 1 (10 marks)

A 1992 study at a hospital in Chicago compared two different ways to take care of geriatric patients who had been transferred from the emergency room to a geriatric ward. The two methods are here labelled Treatment and Control. For each patient, the total hospitalization cost was recorded (in US dollars). Some of the dataset is shown in Figure 2 of the booklet of Figures, and the entire dataset is here:[1]
`https://ritsokiguess.site/datafiles/hospitalization.txt`

(a) (3 marks) Read in and display (some of) the data, and briefly justify why your method is the best for these data.

My answer: The data display suggests that the data values are separated by single spaces, and thus that they can be read in with `read_delim` with second input a single space:

```
my_url <- "https://ritsokiguess.site/datafiles/hospitalization.txt"
hosp <- read_delim(my_url, " ")
```

```
## Rows: 60 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: " "
## chr (1): method
## dbl (1): cost
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
hosp
```

```
## # A tibble: 60 x 2
##    method     cost
##    <chr>     <dbl>
##  1 control     478
##  2 control    2969
##  3 control    7062
##  4 treatment   528
##  5 treatment  2391
##  6 treatment  5928
##  7 control     605
##  8 control    3151
##  9 control    7284
## 10 treatment   650
## # i 50 more rows
```

You could also use `read_delim` with no second input:

```
hosp <- read_delim(my_url)
```

```
## Rows: 60 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: " "
## chr (1): method
## dbl (1): cost
```

[1] Copy and paste the URL. I put it all on one line so that this would work. This applies to other questions on this exam as well.

Question 1 continues...

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
  hosp
## # A tibble: 60 x 2
##    method     cost
##    <chr>     <dbl>
##  1 control     478
##  2 control    2969
##  3 control    7062
##  4 treatment   528
##  5 treatment  2391
##  6 treatment  5928
##  7 control     605
##  8 control    3151
##  9 control    7284
## 10 treatment   650
## # i 50 more rows
```

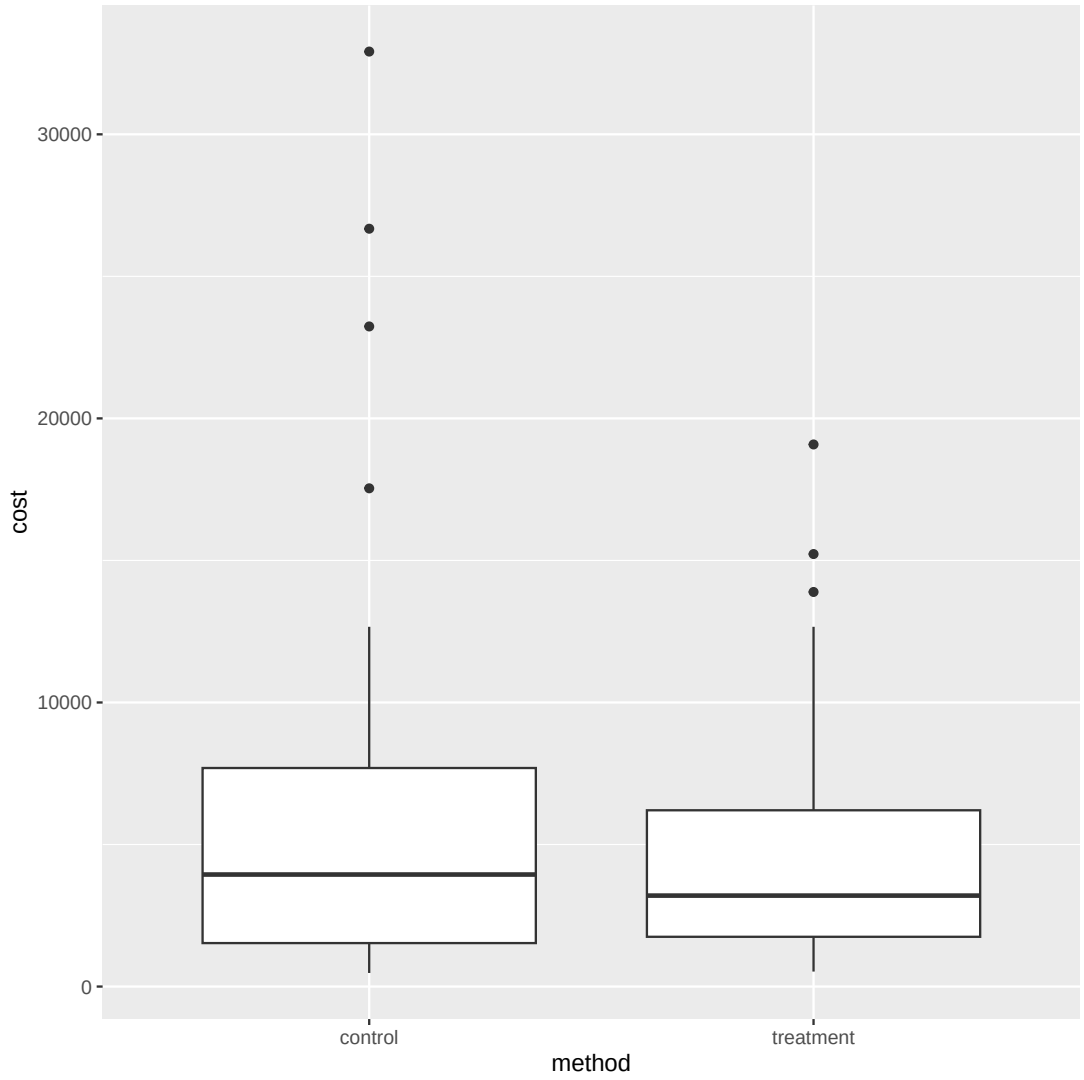
but then you must draw your reader's attention to the fact that `read_delim` successfully *guessed that the data values were separated by a single space.*

(b) (3 marks) Make a suitable graph of the two columns of data, and explain (very) briefly why this was a good choice of graph.

My answer:

There is one quantitative variable `cost` and one categorical variable `method`, so a boxplot is called for:

```
ggplot(hosp, aes(x = method, y = cost)) + geom_boxplot()
```



An alternative angle for the explanation is that we have groups defined by the categorical variable `method`, and we want to compare the (quantitative) `cost` for each of them.

Either way is good, but it is important to say which variable is categorical and which is quantitative, or else it looks as if you don't know.

(c) (2 marks) Briefly describe the *shapes* of each of the distributions of costs.

My answer:
Both distributions are skewed to the right, or have upper (high) outliers. Either is good.

That's all you need. I was not asking about centre or location (so you don't need to compare the medians) and I was not asking about spread (so you don't need to compare the heights of the boxes). Expect to lose something if you write too much. The point is to develop judgment about what your reader is interested in, and saying only that.

(d) (2 marks) Why does it make practical sense that your cost distributions should have the shapes that they do? Explain briefly.

My answer:
The cost of anything cannot be less than zero, but it could be very high (it has a lower limit, but may not have an upper limit). This is particularly true for something like the cost of hospital care, which could be very high indeed if the patient needs some sort of specialized care. Even more so given that these are geriatric patients (old people).

Question 2 (9 marks)

It is known that smoking makes the lungs work less effectively. One measurement method is the carbon monoxide diffusing capacity (DL) of the lungs. A higher value of DL indicates that the lungs are working better. Non-smokers are known to have a mean DL value of 100. A sample of 20 smokers is taken, and the DL is recorded for each one. Some of the data are shown in Figure 3, and the entire dataset is here: `https://ritsokiguess.site/datafiles/lung-capacity.csv`

(a) (1 mark) Using `read_csv`, read in the data and display some of the values.

My answer:
I gave this one away. All you need to do is to read the question carefully enough:

```
my_url <- "https://ritsokiguess.site/datafiles/lung-capacity.csv"
lungs <- read_csv(my_url)
```

```
## Rows: 20 Columns: 1
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (1): DL
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
lungs
```

```
## # A tibble: 20 x 1
##       DL
##    <dbl>
##  1 104.
##  2  92.3
##  3 101.
##  4 103.
##  5  88.6
##  6  61.7
```

```
##  7  88.0
##  8 109.
##  9  73.0
## 10  90.7
## 11  71.2
## 12  73.2
## 13 123.
## 14  84.0
## 15  82.1
## 16 107.
## 17  91.1
## 18  76.0
## 19  89.2
## 20  90.5
```

Twenty observations of one variable called `DL`.

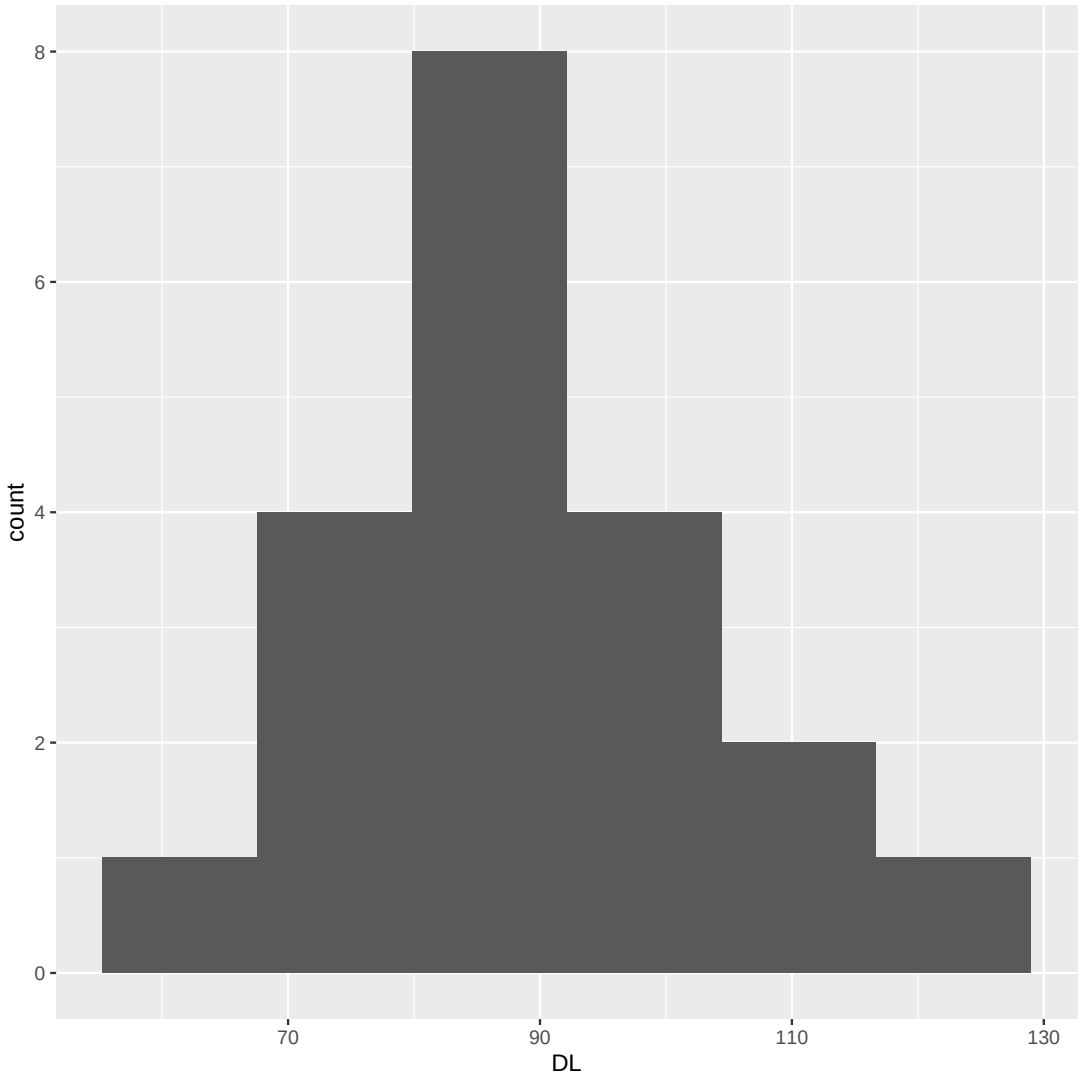
This was meant to be an absolute gimme. If you didn't read the question, for example you used `read.csv` instead of `read_csv`, or you failed to display the data you read in, expect it to cost you.

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

My answer:

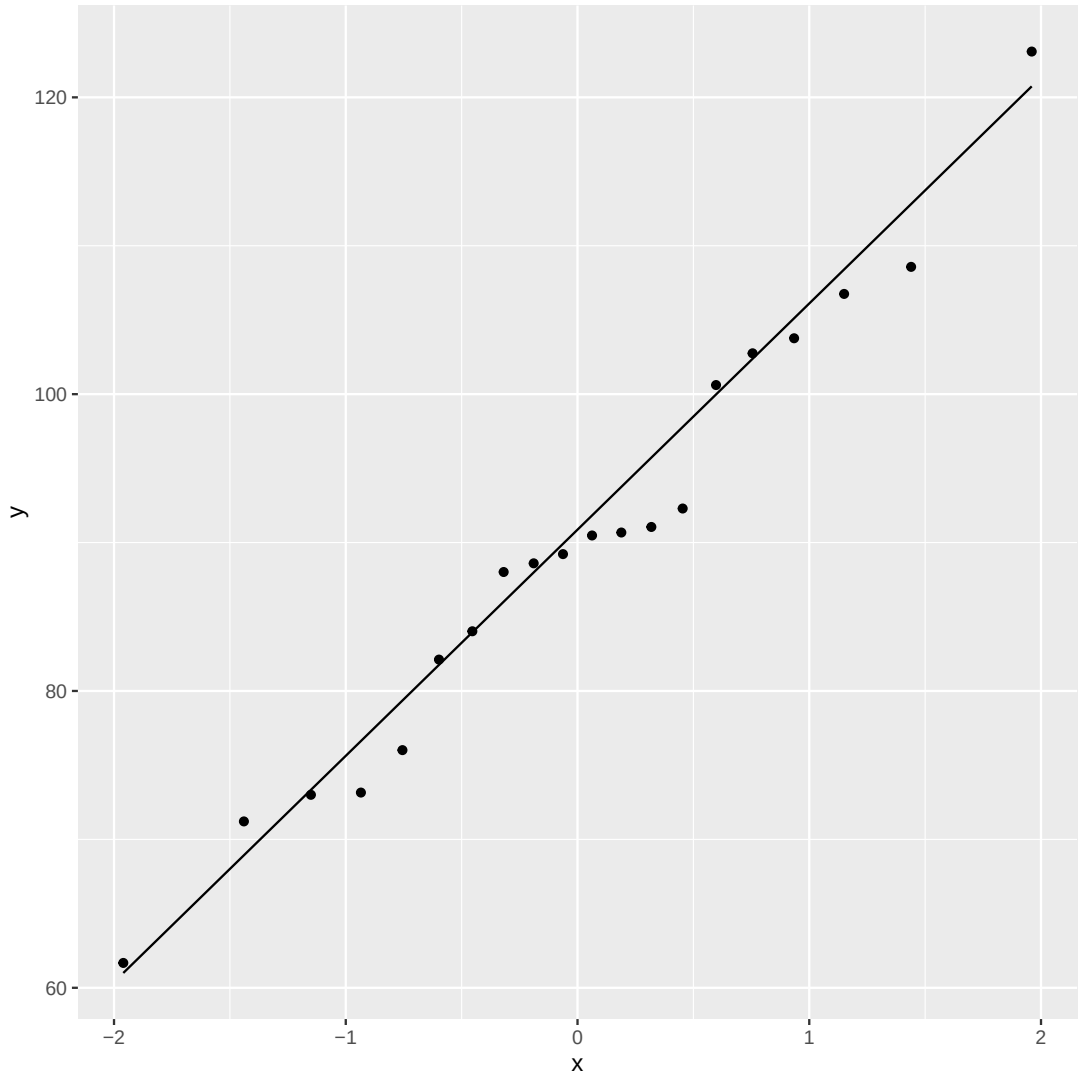
One quantitative variable, so a histogram. There are 20 observations, so the maximum number of bins to consider is about 7. You should certainly show evidence of having thought about the number of bins. I used six bins:

```
ggplot(lungs, aes(x = DL)) + geom_histogram(bins = 6)
```



If you have read ahead (in this question and in the course), you might also think of a normal quantile plot:

```
ggplot(lungs, aes(sample = DL)) + stat_qq() + stat_qq_line()
```



The question asked for *a* plot, meaning exactly one plot, so you will lose points if you provide two. If you think two plots would be better, you need to choose the one you think is best (and, as a general piece of strategy, providing some reasoning will help you if I happen to disagree with your choice).

In this case, either plot is acceptable, since the later parts will ask you to think about normality.

Make sure you use a suitable number of bins on the histogram, based on how much data you have. The grader's guideline was between about 4 and 7. More or less than that won't show the shape, which is what you are after.

In this course, we are using `ggplot` for drawing graphs. I expect you to do the same. You did

not learn (for example) `hist` from me, so you need to say where you *did* learn it from to get any credit.

(c) (2 marks) Make the argument, based on your graph, that a t-procedure (test or confidence interval) will be appropriate here.

My answer:

There are two things to consider:

- the distribution of the `DL` values is itself close to normal
- the sample size is moderate (or use the adjective of your choice), so that the Central Limit Theorem will deal with any remaining non-normality.

There should be a consideration of normality, and there should also be something about the sample size (which could also be that the `DL` values are close enough to being normal that we don't need to worry about the sample size at all). You need to say something about sample size in your answer.

My take is that even if you had doubts about the normality, for example that the distribution was too "peaky" or that it was slightly skewed to the right, the sample size is easily big enough to take care of those. (The normal quantile plot says more clearly that there are no problems with the normality.)

(d) (4 marks) Is there evidence from this data that smokers' lungs are less efficient than non-smokers' lungs? Justify your answer.

My answer:

When we are looking for evidence (and in particular the strength of evidence), this indicates doing a hypothesis test (rather than a confidence interval). The null hypothesis is that the mean DL for smokers is the same as for non-smokers, ie. 100, and the alternative is that the mean DL for smokers is *less* than 100. (The hypotheses need to be statements about populations, in this case the population of "all smokers" of which these 20 observations are a random sample.)

Hence the code is this:

```
with(lungs, t.test(DL, mu = 100, alternative = "less"))
```

```
##
##  One Sample t-test
##
## data:  DL
## t = -3.0443, df = 19, p-value = 0.003336
## alternative hypothesis: true mean is less than 100
## 95 percent confidence interval:
##      -Inf 95.61714
## sample estimates:
## mean of x
##  89.85475
```

The P-value of 0.003 is much less than 0.05, and therefore we reject the null hypothesis in favour of the alternative; that is, we have (strong) evidence that the mean DL for smokers is less than 100, and therefore that their lungs work less effectively than non-smokers' lungs.

You might think of tackling this sort of problem with a confidence interval, like this:

```
with(lungs, t.test(DL))
##
##  One Sample t-test
##
## data:  DL
## t = 26.963, df = 19, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  82.87969 96.82981
## sample estimates:
## mean of x
##  89.85475
```

and then you say that the mean DL for smokers is, with 95% confidence, between 82.9 and 96.8. This interval is completely less than 100, so we have evidence that smokers' mean DL is less than 100. This much is true, but what we lose by doing it this way is that we don't know *how* strong the evidence is. All we know is that the P-value is less than $0.05/2 = 0.025$,[2] but we don't know *how much* less it is. The one-sided CI[3] in my first output suffers from the same problem; it says that the P-value is less than 0.05, but not how much less. The reason this matters is that the person reading your report might have a different standard of "strong enough evidence" than you: they might be using an alpha of 0.001 or some other value that you don't know, and if you only give a 95% confidence interval, they won't be able to make their decision about whether the mean DL for smokers is "really" less than 100.

Some notes of things the grader saw, in ascending order of severity:

- make sure you state your hypotheses clearly, particularly the alternative.
- use a one-sided test, since we only care about whether the smokers' lungs are *less* efficient than those of the population as a whole (or, the population of non-smokers)
- we *never know* whether the alternative hypothesis is true or not: that is the *whole reason* for doing a test, to draw a conclusion or make an inference about the hypotheses, without being able to observe directly what is true and what is false (the only way we could observe it is by observing the entire population).
- not stating *any* hypotheses makes it impossible to tell what your test is doing.
- you cannot draw a conclusion about a population just by looking at the histogram.

Question 3 (11 marks)

A company manufactures light bulbs. One particular model has a mean lifetime of 810 hours with a standard deviation of 200 hours. The company is investigating whether to put a new manufacturing process into production. 50 light bulbs will be produced by the new process and their lifetimes measured. Suppose the mean lifetime of bulbs produced by the new process has mean 850 hours, and the same standard deviation as before. The company will be doing a suitable hypothesis test to see whether the mean lifetime of the light bulbs has increased, compared to 810 hours. Assume that the light bulb lifetimes have a gamma distribution.

In R, random values from a gamma distribution are generated using `rgamma`. In R, this has three inputs: the number of random values to generate, and the gamma distribution parameters `shape` and `rate` (in that order). The mean of the gamma distribution is the shape divided by the rate, and the variance is the shape divided by the rate-squared. (Only the rate is squared in the variance.)

(a) (2 marks) Demonstrate that a gamma distribution with shape 18.0625 and rate 0.02125 has a mean of 850 and a standard deviation of 200.

My answer:

Use R as a calculator:

```
shape <- 18.0625
rate <- 0.02125
shape / rate
## [1] 850
sqrt(shape / rate^2)
## [1] 200
```

Check. The last one is the square root of the variance. As simple as this.

There is partial credit if you draw a random sample from this gamma distribution and show that the mean and SD of the sample come out close to the desired values. You need to say that the values are close for 1 point, otherwise only 0.5. A few people used a Law of Large Numbers idea by showing that if you draw a larger sample, the mean and SD are closer to the desired values. 1.5 for this. There is no credit for a histogram of sampled values without further explanation.

(b) (5 marks) Estimate by simulation the power of a test to reject the previous mean of 810 hours in favour of a larger alternative, if the light bulbs were in fact manufactured by the new process which yields lifetimes with a gamma distribution that has a mean of 850 hours, using a sample size of 50 light bulbs. Do the test with a t-test. Do you think the company would be happy with the results? Explain briefly. Suggestion: begin by setting the random number seed.

My answer:

Using a null mean of 810, a true mean of 850, a standard deviation of 200, a t-test, and generating the random gammas with the scale and rate you checked above (that give the right mean and SD):

```
set.seed(457299)
my_shape <- 18.0625
my_rate <- 0.02125
```

```
tibble(sim = 1:1000) %>%
  rowwise() %>%
  mutate(my_sample = list(rgamma(50, my_shape, my_rate))) %>%
  mutate(test = list(t.test(my_sample, mu = 810, alternative = "greater"))) %>%
  mutate(p_value = test$p.value) %>%
  count(p_value <= 0.05)
```

```
## # A tibble: 2 x 2
## # Rowwise:
##   `p_value <= 0.05`     n
##   <lgl>             <int>
## 1 FALSE               620
## 2 TRUE                380
```

My estimated power is 0.380. Yours will probably be a bit different, because of the randomness, most likely somewhere between 35% and 40%. This is rather small for a power, so the company will probably not be very happy. (To raise the power to something acceptable, the company will need to manufacture more than 50 light bulbs with the new process. How many? That's the subject of the next part.)

Make sure to end with something that says whether you would expect the company to be happy with the result you got.

Comments:

- You don't have to set the random number seed. I did because I didn't want my results to change in between doing it and writing about it. Some people said the power was estimated as a different value from what their simulation said. If it looked as if the knitting caused this, I let it go.
- Use the values of shape and rate you checked in the previous part because you want data generated from a gamma distribution with mean 850 and SD 200, and that shape and rate will get you that. If you thought the mean and/or SD should be something else, you'll need to calculate the shape and rate that go with those.
- Run the right t-test with the right null mean and alternative. (The null is 810; we know the truth is 850, and we want to see whether the known-wrong null is rejected in favour of a mean that is *greater*, as we think it should be.)
- Get the P-value from each one.
- Count how many of those P-values are 0.05 or less; the fraction of TRUE, out of 1000 (for me), is the estimated power. (The larger fraction of FALSE estimates the probability of a type II error; this is distressingly large.)

Using `power.t.test` is wrong because that assumes a normal distribution for the lifetimes (also, the hint in the question is "estimate" the power, rather than "calculate"). Likewise, using `rnorm` is wrong to generate the random samples. (Doing so suggests that you were copying some code of mine without understanding what you were trying to achieve here.) You might have drawn a histogram in the previous part and concluded that the lifetimes are "approximately normal",

but they are *exactly* gamma, and you know how to generate random values from a gamma distribution because you just did it.

I didn't actually teach the `rerun` way, but since that did appear in the first version of the recorded lecture (and you may have seen that elsewhere in the course), I'm good with that if you can do it right:

```
rerun(1000, rgamma(50, my_shape, my_rate)) %>%
  map( ~t.test(., mu = 810, alternative = "greater")) %>%
  map_dbl("p.value") %>%
  enframe(value="pvals") %>%
  count(pvals <= 0.05)
```

```
## Warning:  `rerun()` was deprecated in purrr 1.0.0.
## i Please use `map()` instead.
##   # Previously
##   rerun(1000, rgamma(50, my_shape, my_rate))
##
##   # Now
##   map(1:1000, ~ rgamma(50, my_shape, my_rate))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
## # A tibble: 2 x 2
##   `pvals <= 0.05`     n
##   <lgl>           <int>
## 1 FALSE             612
## 2 TRUE              388
```

The answers either way will be the same, to within randomness.

(Don't smoosh your code together all on one line. Future-you will thank current-you for the readable code, like when you come to study for the final exam.)

(c) (4 marks) By running two more simulations, what can you say about the sample size required to obtain a power of 0.80 in this situation?

My answer:

In my case, the power I got from a sample size of 50 was too small, so I need to rerun the simulation with a bigger sample size. Copy and paste your simulation code, and change the 50 to something bigger. I chose 100:

```
tibble(sim = 1:1000) %>%
  rowwise() %>%
  mutate(my_sample = list(rgamma(100, my_shape, my_rate))) %>%
  mutate(test = list(t.test(my_sample, mu = 810, alternative = "greater"))) %>%
  mutate(p_value = test$p.value) %>%
  count(p_value <= 0.05)
```

```
## # A tibble: 2 x 2
## # Rowwise:
```

```
##   `p_value <= 0.05`     n
##   <lgl>             <int>
## 1 FALSE               363
## 2 TRUE                637
```

My power is now 0.646, still not big enough. I need to increase the sample size some more. My guess is that 200 is too much, so I'll try 150. Copy and paste and edit again:

```
tibble(sim = 1:1000) %>%
  rowwise() %>%
  mutate(my_sample = list(rgamma(150, my_shape, my_rate))) %>%
  mutate(test = list(t.test(my_sample, mu = 810, alternative = "greater"))) %>%
  mutate(p_value = test$p.value) %>%
  count(p_value <= 0.05)
```

```
## # A tibble: 2 x 2
## # Rowwise:
##   `p_value <= 0.05`     n
##   <lgl>             <int>
## 1 FALSE               209
## 2 TRUE                791
```

That was a lucky break: 0.793, very close to 0.8. So the desired sample size is close to, a little bigger than, 150.

For you, *I don't mind* what sample sizes you choose, as long as your choices make sense. Assuming that the power you found in the previous part was less than 0.8, for your first attempt you need to make your sample size bigger than 50, but anything at all bigger than 50 is fine. Then, for your second simulation in this part, increase the sample size again if the power came out too small and *decrease* it if it came out too big.

One point for a simulation with a (presumably) larger sample size, one point for making the case for the next simulation, one point for doing that, and the fourth for saying something about the sample size needed to get the desired power.

If you thought `power.t.test` was the way to go, do something sensible here using it again, either trying some bigger sample sizes, or using it to find the sample size for the target power 0.80. It was wrong before, but if you used it before, you are not making any additional errors using it again, so you can get full marks for this part using it (as long as you used it before). About the only way you can use `power.t.test` in this part and get away with it is to use it to give you a starting guess at the sample size, bearing in mind that it won't be quite right because the population distribution is gamma, rather than normal. (If you do that, it helps to say that this is what you are doing, to make sure I realize that you understand.)

End your work with the best statement you can make about the desired sample size, for example:

- if the sample sizes you tried were 100 and 200, the desired sample size is between 100 and 200 (because the power is too big with $n = 200$)
- if the sample sizes you tried were 60 and 70, the desired sample size is greater than 70 (because the power is too small with both of those).

"The sample size must be greater than 70" doesn't seem like a very illuminating answer, but for the purposes of this question, it is perfectly good if the sample sizes you tried were something

like 60 and 70. I limited you to two more simulations to discourage you from spending too much time on this question.

If you didn't make very much progress for (b), I tried to give you credit for your (c) based on how logically it followed from what you did in (b). This means that you can still get full credit for (c), but you cannot make (c) easier than it should have been had you done (b) correctly. (If you did, expect to lose something in (c) as well.)

Some people reran their simulation with the *same* sample size of 50. It wasn't clear to me what you were hoping to learn from this, since the answers came out about the same as they did before. If you're trying to figure out a sample size, what you need to do is to learn something about the relationship between sample size and power, and you will need to *change* the sample size to have any prospect of success.

Question 4 (10 marks)

Twenty students wrote a test that was marked out of 100 points. The test scores are here: `https://ritsokiguess.site/datafiles/test-scores.csv`

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

My answer:

The file is a `.csv`, so again `read_csv` is again the thing:

```
my_url <- "https://ritsokiguess.site/datafiles/test-scores.csv"
test_scores <- read_csv(my_url)
```

```
## Rows: 20 Columns: 1
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (1): score
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
test_scores
```

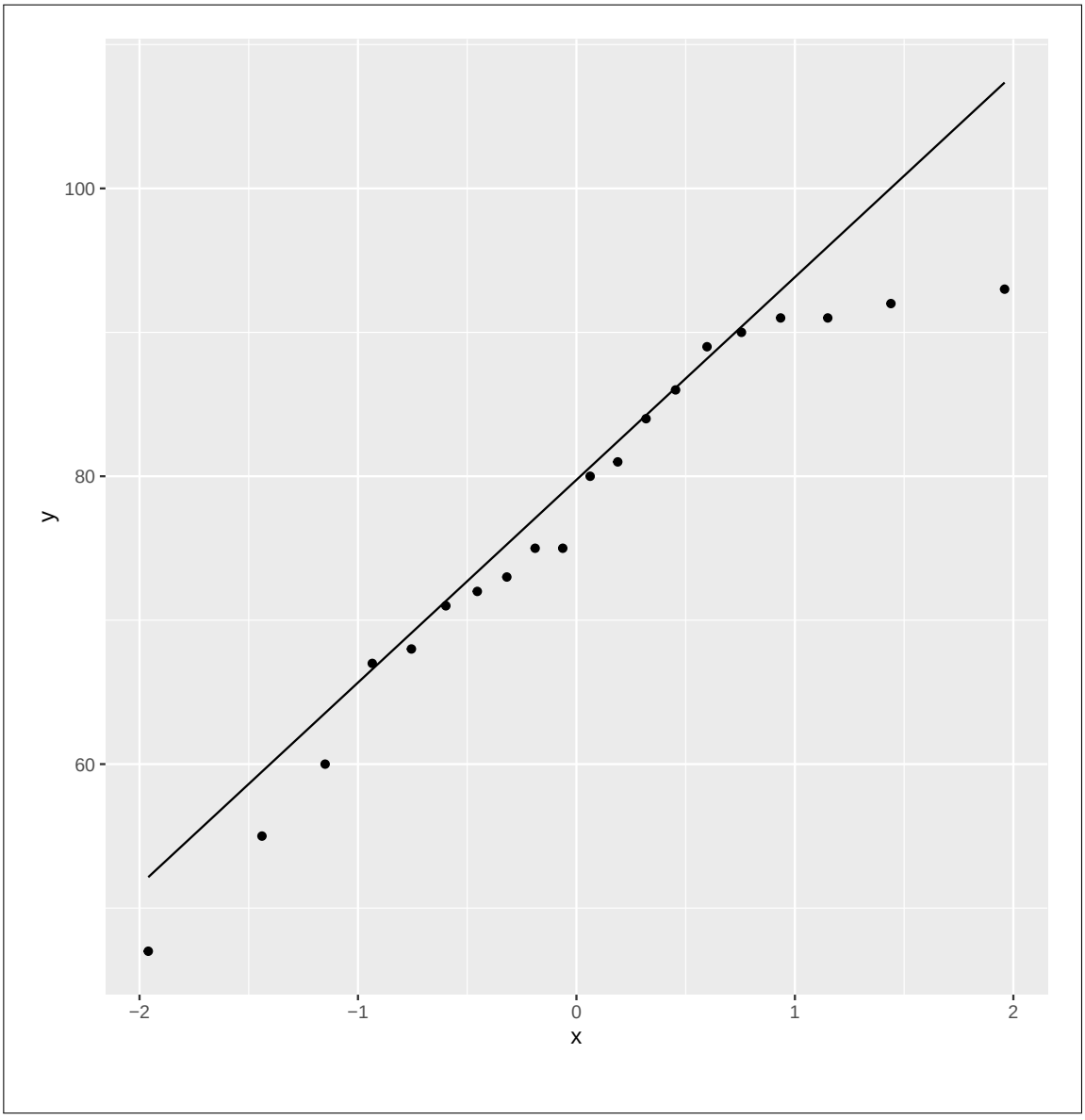
```
## # A tibble: 20 x 1
##    score
##    <dbl>
##  1    71
##  2    73
##  3    90
##  4    81
##  5    93
##  6    86
##  7    89
##  8    68
##  9    91
## 10    92
## 11    67
```

```
## 12    60
## 13    84
## 14    80
## 15    55
## 16    75
## 17    75
## 18    47
## 19    72
## 20    91
```

(b) (3 marks) Draw a normal quantile plot of the (one column of) data.

My answer:
Thus:

```
ggplot(test_scores, aes(sample = score)) + stat_qq() + stat_qq_line()
```



(c) (3 marks) What do you conclude from your plot? Explain briefly.

My answer: Compared to a normal distribution, the scores go down too low at the bottom (have a long lower tail) and are too bunched up at the top (have a short upper tail). Thus the distribution of scores is skewed to the left.

It is better to reason it out this way rather than to try to memorize which kind of skew goes with an upward- or downward-opening curve. This one, if you want to call it that, is a downward-opening curve. Some other software switches the roles of the observed data (here on the y-axis)

and the expected data if the normal is correct (x-axis here), and if all you do is memorize curve shapes, you will have the direction of skewness *wrong* the moment you switch software.

There are three test scores at the bottom that appear to be consistently below the line, so I think it is better to call this left-skewed (a feature of the whole distribution) rather than saying that there are outliers at the bottom (implying three isolated unusual values) and a short tail at the top.

(d) (2 marks) Based on what you know or can guess about test scores, does it make sense that the distribution of scores for this class would have this kind of shape? Explain briefly.

My answer:

A left-skewed distribution implies some values quite a bit lower than the others. This makes sense for a distribution of test scores; most students did well, but a few (the under-prepared ones) did worse than the others. You could also talk about the top end: several students got above 90 but nobody got close to 100, suggesting that there was one difficult question that nobody did that well on. (This is different from an easy exam, in which a number of students might get close to 100, which serves as an upper limit that the scores are otherwise skewed away from. If you like, you could think of this one as having an upper limit of about 95 and the scores are skewed away from that.)

Question 5 (10 marks)

A property tax office received complaints that a particular tax assessor (labelled A in the data) was biased. An experiment was conducted to compare assessor A with another tax assessor B from the same office. There had been no complaints about assessor B. Eight properties were selected, and each one was assessed by both assessors. The assessments, in thousands of dollars, are shown in Figure 5.

There is no coding in this question.

(a) (3 marks) Two tests are shown in Figure 6 and Figure 7. Based on what you know about the data so far, which of these tests is more appropriate and why?

My answer:

The first test is for two independent samples, and the second one is for matched pairs. We know in this case that each property was assessed by *both* assessors A and B, and so this is matched pairs rather than two independent samples. Hence, the test in Figure 7 is more appropriate.

Extra: we haven't looked at any graphs, so we don't know whether some kind of sign test or Mood's median test would be better than some kind of t-test. But, from what we know about the design of the experiment, we certainly do know that a matched-pairs t-test is better than a two-sample t-test. The right thing to do, therefore, would be to assess the normality of the differences (the assumption that is required for the matched-pairs t) and to do a sign test on the differences if that is not normal enough given the small sample size (only 8 pairs).

(b) (2 marks) What do you conclude from the more appropriate test, in the context of the data?

My answer:

Looking at Figure 7, the P-value is 0.026, which is less than 0.05, so we reject the null hypothesis that the two assessors have an equal mean assessment, and conclude that assessor A is biased (the mean assessments are different). The P-value doesn't tell us which way assessor A is biased, but the positive numbers in the confidence interval say that the bias is that A's assessment is actually higher. (Thinking in terms of taxes, people will not complain if the tax assessment is too low, but they will complain if it is too high. This is a flipped-around version of the fact that students will not complain if their exam mark is too high, but they will if they think it is too low.)

Extra: if you mistakenly concluded that the two-sample test was the one to do, then you need to get the P-value of 0.6956 from Figure 6 and conclude that there is no difference in mean assessment between the two assessors, and thus that Assessor A is not biased. You need to be consistent; if you conclude that the two-sample test is better and then draw a conclusion from the matched-pairs test, you are making *two* mistakes, but if you conclude that the two-sample test is better and then draw a conclusion from the same test, you are only making one.

(c) (3 marks) Looking at the numbers in Figure 5, explain briefly how it is that the P-values for the two tests in Figures 6 and 7 are so different.

My answer:

This requires you to think about how the tests are computed differently.

The two-sample test treats the A values and the B values independently ("two independent

samples") and each set of values looks rather variable, because the properties are different one from another. If you think about how the Welch t-statistic is calculated (from Lecture 3a), it has $s_1^2/n_1 + s_2^2/n_2$ on the bottom, and the two s_i^2 are both large, so the test statistic will be small in size (close to 0) and the P-value will be large.

The paired test looks at the differences between A and B for each property, and these are not only small but also *consistent* from one property to the next: even if the assessments are both large for a property, they tend to differ by about the same amount than the two assessments would if they were both small. So the $s/\sqrt{n}$ in the (one-sample) formula for the test statistic will tend to be small, the test statistic will be larger in size, and its P-value will be smaller. This is exactly what happened here.

Another way to come at this is to say that the variability in assessments comes from two sources: one, because the assessors are different, and two, because the *properties* are different. The matched pairs test eliminates the second source (by taking differences) and so the s in the denominator focuses on whether the assessors are different. The two-sample test doesn't eliminate anything, and so the variability is a mixture of both sources; the test is then less sensitive (as well as being wrong, of course).

Saying only that the two-sample test is wrong and the matched-pairs test is right is true, but not very insightful. One out of three.

Extra: if you were comparing say four assessors instead of two, you wouldn't be able to eliminate the second source of variability because you wouldn't know which differences to take. In that case, you would have to properly accommodate both sources of variability, and you would have to allow for the assessors' valuations to be correlated overall because the properties differ (if one assessor's valuation is higher than average, the others will tend to be as well, because it's a high-valued property). The right way to handle this is called "repeated measures ANOVA", which you might see in your multivariate course. In fact, you can use repeated measures to analyze matched pairs (and you will get the same result), but if you are only comparing two "treatments", as here, you have the get-out of being able to take differences.

(d) (2 marks) What other piece of output would be helpful to you in assessing the appropriateness of your preferred analysis? Explain briefly.

My answer:

The word "appropriateness" suggests assessing assumptions. The key assumption here is that the differences from the matched pairs test are sufficiently close to normal. So the thing to ask for is a graph, a normal quantile plot of the differences (best) or a histogram of the differences (second best). Then you will know whether the matched pairs can be trusted, or whether you have to look at a sign test of the differences (which is very unlikely to be unbalanced enough to be significant with such small samples).

You could ask for the output for the matched pairs sign test, but the problem with that is that without the graph you have no idea whether you should prefer the matched pairs t or sign tests. And if the P-values come out different enough to lead to different conclusions (as they might), you don't know which test to believe. So this is not as good as asking for a graph.

Again, consistency is key. If you thought this was a two-sample situation, you need to ask for a graph that would assess the assumptions for that (normal quantile plots for *each* assessor's

assessments). That would be better than asking for the results from the other test, which would be Mood's median test in this case. Expect the grader to check back and make sure you were consistent.

End of Exam This page: 0 marks.