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 3 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.

(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.

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

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

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.

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

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

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

[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 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.

(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.

(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?

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.

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

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

(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.

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?

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

(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.

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