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

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

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

Past exams are *not* allowed.

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

This exam has 8 numbered pages of questions. Please check to see that you have all the pages.

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

Answer each question in the space provided (under the question). If you need more space, use the backs of the pages, but be sure to draw the marker's attention to where the rest of the answer may be found.

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

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

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

```
library(tidyverse)
```

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

Last name: _____

First name: _____

Student number: _____

For marker's use only:

Page	Points	Score
1	9	
2	7	
3	12	
4	10	
5	13	
6	7	
7	9	
8	11	
Total:	78	

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

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

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

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

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

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

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

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

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

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

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

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

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

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

   The data are shown in Figure 4.

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

   (b) (4 marks) For this part and the following parts, you may assume that the data have been read into a data frame called `crackers`.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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