University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAC32 (K. Butler), Midterm Exam
October 26, 2015

Aids allowed:

- My lecture overheads
- Any notes that you have taken in this course
- Your assignments and feedback on them
- My assignment solutions
- The course R text
- The course SAS text
- 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 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 the total marks for each question are shown in the table on the next page.

When giving SAS code, it is acceptable to use code that runs on the online version of SAS Studio, or on the version that runs on a virtual machine. Either version is acceptable.

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	2	
1	2	
1	6	
1	2	
2	6	
3	4	
3	2	
3	6	
4	7	
5	10	
5	2	
6	8	
6	2	
6	6	
7	2	
7	3	
7	3	
8	11	
Total:	84	

1. Professional baseball in North America is played in two leagues, the American League (in which the Blue Jays play) and the National League (in which the Montreal Expos used to play). Figure 1 in the booklet of code and output shows, for each team, the league they play in and the number of home runs hit in the team's stadium. (The number of home runs is a total for the 2011 season; each team plays 81 home games.) The data have been saved in a file called `hr.txt`, both in the working folder of R Studio and in your home folder on SAS Studio.

   Give suitable code that will accomplish the tasks described below. You should need only between one and three lines of code in each case.

   (a) (2 marks) Read the data into an R data frame.

   (b) (2 marks) Draw side-by-side boxplots of home runs for each league, in R.

   (c) (2 marks) Calculate the mean number of home runs for each of the two leagues, in R.

   (d) (2 marks) Read the data into a SAS data set. (For this part, assume that the top line of the data file shown in Figure 1 has been omitted. This is to make your job easier. The solution where that top line has been included will also be considered correct.)

   (e) (2 marks) Draw side-by-side boxplots of home runs for each league, in SAS.

   (f) (2 marks) Calculate the mean number of home runs for each of the two leagues, in SAS.

2. The Finger Lakes is a district of New York State famous for producing wines. There are several different vineyards along three of the lakes. For each vineyard, the selling price of a case of wine is recorded, and each vineyard is classified by which lake it is on (labelled `location` in the data set). SAS boxplots are shown in Figure 2 in the booklet of code and output.

   (a) (2 marks) Describe how the locations differ in terms of average case price, if at all. (You will have to decide what you mean by "average".)

   (b) (2 marks) Describe how the locations differ in terms of inter-quartile range, if at all.

   (c) (2 marks) For the location that has the smallest inter-quartile range, how might the *standard deviation* be a misleading measure of spread? Explain briefly. (If you found more than one location that has the smallest IQR, pick any one of them, say which one you picked, and then answer the question.)

3. A study was carried out on boys with attention deficit hyperactivity disorder (ADHD). The researchers rated boys' performance on a number of tasks. One task was "has difficulty organizing work", and each boy was rated on a 0–4 scale, with 0 meaning "has no difficulty" and 4 meaning "has a lot of difficulty". 282 boys with ADHD were rated. *Only* the integer values $0, 1, 2, 3, 4$ were used for ratings; there were no decimal ratings. Some output is shown in Figures 3 and 4. You may refer to the task as "this task" as you answer the questions below, rather than writing out "has difficulty organizing work" every time.

   (a) (2 marks) Figure 3 shows some of the data, and a table of how many boys received each rating. Explain briefly what is assumed about the data in order to use a t-test (or t confidence interval) and why that assumption cannot be satisfied exactly here.

   (b) (2 marks) Despite what you said in part (a), explain briefly why you would have no real hesitations about using a t-test (or t confidence interval) here.

   (c) (2 marks) The output in Figure 4 contains a confidence interval. Explain *precisely* what that confidence interval means.

   (d) (3 marks) Figure 4 also contains a hypothesis test. Give the null and alternative hypotheses for this test, defining any symbols that you use. Also, what do you conclude in the context of the data?

   (e) (3 marks) R obtained a P-value for the test in Figure 4. Describe the process by which you would obtain a P-value for this test using either statistical tables or an R function *other* than `t.test`. I don't want a numerical answer, just a description of the process. Use the fact that the t-distribution with large degrees of freedom can be closely approximated by a standard normal distribution. Your normal table, if you go that way, gives the probability of observing a value *less* than the given z.

4. I have a spreadsheet that contains names and information about different types of coffee drinks served in a certain coffee shop. Some of the spreadsheet is shown in Figure 5 of the booklet of code and output. I want to read this information into SAS.

   (a) (4 marks) Tell me how to get the data in the spreadsheet into SAS Studio so that (in the next part) it can be read into a SAS data set. You need to provide enough detail so that I can reproduce your process myself and end up with some kind of file in SAS Studio. (You can assume that I know how to do basic spreadsheet operations.)

   (b) (3 marks) Give code that will read the file that you saved on SAS Studio into a SAS data set, preserving the names of the coffee drinks as well as possible. Depending on what you did in the previous part, your code might be quite simple.

5. A student suspected that files might take longer to download at certain times of the day. To examine this, he placed a file on a server, and then, at certain times of the day on randomly chosen days, he downloaded the file and noted how many seconds it took. The times of day were 7:00am (denoted "early" in the data set), 5:00pm ("evening") and midnight ("late"). The data are shown in the booklet of code and output as Figure 6. Some analysis is shown in Figure 7 and Figure 8.

   (a) (1 mark) In Figure 7, what *null hypothesis* is being tested?

   (b) (1 mark) In Figure 7, what *alternative hypothesis* is being tested?

   (c) (2 marks) What do you conclude from the analysis of Figure 7?

   (d) (2 marks) Why is the analysis in Figure 8 worth doing? Explain briefly.

   (e) (2 marks) What do you conclude from Figure 8?

   (f) (2 marks) Which time of day has quickest downloads, on average? Is it significantly quicker than all the other times of day? (Note that a *small* time goes with a quick download.)

   (g) (2 marks) Look at Figure 9 of the booklet of code and output, which shows the distributions of download times at each different time of day. Concerning your conclusions above, do you have (i) no doubts, (ii) moderate doubts or (iii) severe doubts about their validity? Explain briefly.

6. In Question 5, we looked at an analysis of download times for a file downloaded at different times of day. Figures 7 and 8 showed the output from the analyses, but the code was not shown. Your job in this question is to give the code. You may assume that the data in Figure 6 have already been read into a data frame `dl`.

   (a) (4 marks) What two or three lines of R code will produce the output in Figure 7?

   (b) (2 marks) What one or two lines of R code will produce the output in Figure 8?

7. Arsenic is toxic to humans. People can be exposed to it through contaminated drinking water, food, dust and soil. A new way of examining a person's exposure to arsenic is to examine their toenail clippings. Levels of arsenic, in parts per million, measured from the toenail clippings of 19 people in New Hampshire, are shown in Figure 10 in the booklet of code and output. A boxplot of the arsenic levels is shown in Figure 11.

   (a) (2 marks) Why would a sign test be better here than a t-test here as a test of "centre" or "location"? Explain briefly but precisely. (I am looking for *two* points.)

   (b) (2 marks) I read the data into a SAS data set containing a variable named `arsenic`. Give SAS code that will obtain a sign test that the median is 0.400 (against a two-sided alternative).

   (c) (1 mark) The output is quite lengthy. Tell me something that would be next to the sign test in the output so that I can find it more easily. (No explanation needed.)

   (d) (2 marks) Figure 12 shows an R function I wrote that will take a hypothesized median and run the sign test for that hypothesized median on the input data. It returns the two-sided P-value for the sign test in question. Figure 13 shows the P-value of the sign test on the arsenic data for various different null medians. What is the P-value for the test that you gave code for in (b)? What do you conclude about the median from this?

   (e) (3 marks) Use the output in Figure 13 to obtain a 90% confidence interval for the population median. Explain briefly how you obtained your interval.

8. The cardiovascular system in humans consists of the heart and blood vessels, which circulate blood around the body. It is important that this system operates properly, and there are a number of "risk factors" that might prevent it from doing so. One of these risk factors is a sedentary lifestyle.

   One study compared a group of runners (who averaged at least 15 miles per week of running) with a control group of "generally sedentary" people. The gender of each person was also noted. The outcome variable was the number of heart beats per minute after running on a treadmill for 6 minutes. The dataset we will use for this question is shown in Figure 14. (This is an excerpt of a much larger dataset.)

   In this question, we will be using the R package `dplyr`. You can assume for this question that the package has already been installed (with `install.packages`) and loaded (with `library`). The data have been read into a data frame called `runners`. Give code (using `dplyr` tools) to accomplish the following tasks. You should need no more than three lines of code in each case, and each part *can* be done with two lines or fewer:

   (a) (2 marks) Display the runners, but not the "generally sedentary" people.

   (b) (3 marks) Display just the genders of the people that have `beats` less than 100.

   (c) (3 marks) Obtain the mean and standard deviation of `beats` for both the runners and the sedentary people.

9. Does logging (cutting down trees) in an area have an effect on the number of tree species present some years later? A study of logging in Borneo looked at 12 forest plots that had never been logged, and 9 (otherwise similar) plots that had been logged 8 years earlier. Some code and output is shown in Figures 15 through 17.

   (a) (2 marks) Why, looking at the Figures, would you have doubts about doing a two-sample t-test to compare the tree species in the two types of plots? How would a randomization test be better? Explain briefly.

   (b) (3 marks) Describe briefly how the function in Figure 16 does something different from the first two lines of that Figure. (You can do this by describing what they each do, and then explaining how they are different.)

   (c) (2 marks) What does the top line of Figure 17 calculate, and how does it do it (in general terms, not in detail?)

   (d) (2 marks) The research hypothesis was that a logged plot would have *fewer* tree species present 8 years later than an unlogged plot. Explain how the histogram in Figure 18 supports this hypothesis.

   (e) (2 marks) Obtain a suitable P-value for the randomization test from Figure 17. What do you conclude in terms of the effect of logging?