University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAC32 (K. Butler), Midterm Exam
October 24, 2016

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
- 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 11 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.

When giving SAS code, you can provide code that runs either 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	7	
2	9	
3	9	
4	7	
5	6	
6	8	
7	9	
8	6	
9	8	
10	12	
Total:	81	

1. Coffee travel mugs are designed to keep coffee warm. A student decided to compare four different brands of travel mug. Each time, the student heated water to 90° C, poured it into the mug and sealed it. After 30 minutes, the student opened up the mug, measured the temperature of the water, and recorded the difference between the original temperature and the temperature after 30 minutes. Some of the data are shown in Figure 1 in the booklet of code and output.

   These data have been stored in a file `coffee.txt` in your current working folder in R Studio and also in SAS Studio.

   In later parts of this question, you do not need to repeat the code that you used in earlier parts. (For example, in doing computations on the data set, you may assume that the data set was already successfully read in.)

   (a) (2 marks) Give code to read the data into a data frame in R, and to display the whole data frame.

   (b) (3 marks) Give code to read the data into a SAS data set, and to display the whole data set.

   (c) (2 marks) Give SAS code to calculate the mean temperature difference for each brand of cup.

(d) (3 marks) Give R code to calculate the mean temperature difference for each brand of cup.

(e) (3 marks) Give R code to make `ggplot`-style side-by-side boxplots of temperature difference for each cup. (You may assume that `ggplot2` has already been loaded with `library(ggplot2)`.)

(f) (3 marks) Give SAS code to make side-by-side boxplots of temperature difference for each cup.

2. A car magazine obtained data on a large number of models of American and Japanese cars. The magazine was interested in testing whether the gas mileage, measured in miles per US gallon of gasoline, was on average better for Japanese cars than American ones. The structure of the data is shown in Figure 2 in the booklet of code and output.

   (a) (3 marks) Give R code to carry out a suitable t-test that will address the magazine's interest. Briefly justify your choice of alternative hypothesis. You may assume that the data have already been read into a data frame called `carmpg`.

   (b) (2 marks) The results of my t-test are shown in Figure 3. What do you conclude, in the context of the data?

   (c) (3 marks) Figure 4 shows side-by-side boxplots of the gas mileages for the cars from each country. Explain briefly why we might initially have concerns about the validity of the t-test, but explain why these concerns are not serious here.

   (d) (1 mark) How would you tell *how much* better the gas mileages are for Japanese cars than American ones?

3. Police trainees sit in a darkened room facing a projector screen. Ten different licence plates are projected on the screen, one at a time, for 5 seconds each, separated by 15-second intervals. After the last 15-second interval, the lights are turned on, and the trainees are asked to write down as many of the 10 licence plate numbers as they can (in any order at all). The number of licence plates correctly recalled is recorded.

   Fifteen of the trainees (randomly chosen) went through the above procedure, and were then given a week-long memory training course. After that, they were re-tested. The number of licence plates correctly recalled before and after training is shown in Figure 5.

   (a) (2 marks) Are these data matched pairs or two separate samples? Explain briefly.

   (b) (3 marks) The data have been read into a SAS data set with two variables, called `before` and `after`. Give code to run a suitable *t*-test to determine whether the training program is *helpful*.

   (c) (2 marks) I ran your code from above and obtained the results shown in Figure 6. What do you conclude from this analysis, in terms that would be helpful to the people running the police training program?

(d) (1 mark) The confidence interval shown in Figure 6 makes no sense. How would you change your code to obtain a sensible 95% confidence interval?

(e) (3 marks) Look at Figures 7, 8, and 9. For *each* Figure, choose *one* of these two things: (i) if it is relevant to the trustworthiness of the test, describe what it tells you; (ii) if it is not relevant, explain (very) briefly why not.

(f) (2 marks) A police officer who had some training in statistics a long time ago says "but these are numbers of licence plates identified out of 10. They must be binomial, and cannot be normal. Therefore you have no business doing a *t*-test." How would you respond to this?

4. The IRS (Internal Revenue Service) is an American government organization whose purpose is to make sure people pay the tax they owe. The IRS says it should take the average taxpayer 150 minutes to fill out their tax forms. A journalist has been hearing about people taking much longer to complete their tax forms, and so arranges a study in which 30 randomly-chosen taxpayers record the number of minutes it takes them to complete their tax forms. The journalist's aim is to show that it actually takes longer than 150 minutes on average to complete the forms. The data are shown in Figure 10.

(a) (2 marks) A histogram of the data is shown in Figure 11. Note that the `bins=10` in the code means "choose the bin width so that are ten (possibly empty) classes of equal width that the values are classified into". Why is the median a the better measure of centre (location) than the mean for these data? Explain briefly.

(b) (2 marks) Explain briefly why a sign test would be better than a t-test to assess the journalist's claim. (There may be some overlap with the previous part. If you have made the relevant points somewhere within these two parts, I am happy.)

(c) (2 marks) Write down the null and alternative hypotheses for your sign test, defining any symbols you use.

(d) (2 marks) What does the table in Figure 12 tell us? Explain briefly.

(e) (2 marks) Figure 13 is a table of the binomial distribution with $n = 30$ and $p = 0.5$. The number in the `prob` column is the probability of the number of successes in the `success` column *or fewer*. Thus, for example, the probability of 5 successes or fewer is 0.00016 (1.6×10^{-4}).

Use this table to obtain a P-value for your sign test. *Do not* express your answer in scientific notation; you should give an ordinary decimal number.

(f) (2 marks) Does the journalist "have a story": that is, is there evidence in favour of his claim? Explain briefly.

(g) (3 marks) Figure 14 shows a function to obtain a two-sided P-value for a sign test with null median `med` and data `z`. (This previously appeared in one of the assignment solutions). This function is used in Figure 15. Use the results shown there to obtain a 95% confidence interval for the median time spent filling out the IRS forms, to the accuracy that the output permits.

(h) (2 marks) The journalist also did a t-test that the mean was 150 minutes, against the alternative that it was greater, and obtained a P-value of 0.06. What do you think it is about the data that would make the t-test come out almost significant, while the sign test is nowhere near significant? Explain briefly.

5. Suppose you have data

$$16, 17, 18, 20, 20, 23, 25, 29, 34$$

and you want to test whether the population *third quartile* could be 21 (as the null hypothesis) against the alternative that the population third quartile is *less* than 21. You want to use the same idea as the sign test for the median.

(a) (2 marks) What would you use as the *test statistic* for your test (that is, the value you would look up in a table to get a P-value for)?

(b) (4 marks) How would you obtain a P-value for testing that the third quartile of the population from which these data came is 21? Explain in words what you would do. Give R code if you think it makes your explanation clearer. (I should be able to reproduce what you say and get the right answer for the right reason.)

6. A psychologist is studying pattern-recognition skills in grade 4 children. Each child is given a test with 10 patterns to identify. Each child does the test under one of four different conditions:

Praise: child given praise for right answers and no comment about wrong answers

Criticism: child given criticism for wrong answers and no comment for right answers

Interest: child given no praise or criticism, but observer expresses interest in what the child is doing

Silence: observer remains silent while watching the child

The response variable for the study is the number of patterns out of 10 that each child correctly identifies. The research question is whether the different conditions have any kind of impact on pattern-recognition ability. The data are shown in Figure 16.

(a) (3 marks) The mean and SD of the number of patterns identified for each setting, along with boxplots, are shown in Figures 17 and 18. What *two* assumptions do we need to assess for analysis of variance? Explain briefly why those assumptions are difficult to assess here. You do not need to go further in assessing the assumptions. (We will assume in the rest of the question that those assumptions are sufficiently satisfied.)

(b) (2 marks) An analysis of variance is shown in Figure 19. What precisely do you conclude from this, in the context of the data?

(c) (3 marks) The results from Tukey's method are shown in Figure 20. Was it a good idea to obtain these results? Explain briefly, in the context of the data. If you think it was a good idea, what do you conclude from the results?

7. Suppose we are trying to organize a study that estimates the number of hours of TV people watch per week. In particular, a study carried out ten years ago found that people watched a mean of 20 hours of TV per week, with a standard deviation of 8 hours per week. We are trying to see if the mean has changed since then (for example, it may have gone down because people are spending more time online, for example on Facebook, rather than watching TV as they used to do).

   (a) (3 marks) We have decided that a change in mean of 2 hours per week is of interest to us. We have a budget to survey 40 people. We want to know the power of our test to detect this change. Write SAS code to find this out (under the assumption that the number of hours of TV watched has approximately a normal distribution).

   (b) (2 marks) Our supervisor has said that she wants to achieve power 0.80, and will provide resources to collect a sample that is large enough. To find out how large a sample will be needed, what changes would you make to the code of the previous part?

   (c) (3 marks) Another way to obtain the power in this situation is by simulation. The assumption used by SAS in the previous parts is that the population distribution is normal. Recall that the null mean is 20 and that the true mean differs from the null mean by 2 (the direction does not matter). Write an R function, with no input, that draws a random sample of size 40 from the true normal distribution, tests whether its mean is significantly different from the null mean, against a two-sided alternative, and returns the P-value of the test.

   (d) (2 marks) Give R code to run your function 1000 times and to count how many P-values are less than 0.05. (This number, divided by 1000, is the simulated power.)

   (e) (2 marks) What would be an advantage to using the simulation approach to assessing power over the `proc power` approach, in general? Explain briefly, giving an example if you wish.

Notes

[1] "There are differences among the means" also works as an answer here, because it says that at least one pair of means differs without implying that they all do. In mathematical terms, it's the difference between $\exists$ and $\forall$.

[2] The function f is *literally* a black box!

[3] There are some caveats about how you can't make things easier for yourself by getting an earlier part wrong, but this is the basic idea.