University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAC32 (K. Butler), Final Exam
December 19, 2016 2:00–5:00pm

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 12 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.

Figure 1 of the booklet of code and output shows the R packages that I loaded in preparing this exam. You may assume that these packages have been loaded, and your code can use anything within them without further comment.

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	15	
2	10	
3	13	
4	7	
5	8	
6	9	
7	7	
8	8	
9	9	
10	11	
11	7	
12	7	
Total:	111	

1. The data shown in Figure 2 of the booklet of code and output come from a survey. Ten individuals are listed here; the columns respectively indicate a numerical ID, gender, age, income (on some scale), and the scores on three tests, which will be labelled R1, R2 and R3.

   Each of the question parts below can be answered with three lines of code or less (2017: except for (a), which might be longer).

   (a) (3 marks) The data have been stored in a file `survey.txt` in your folder on SAS Studio. Give SAS code to read the data in with suitable variable names. (You may lose marks for any unnecessary code.) Extra for 2017: why can I not use last year's data file?

   (b) (2 marks) Give SAS code to calculate the mean scores on each of the tests R1, R2 and R3 for each gender.

   (c) (2 marks) Give SAS code to obtain side-by-side boxplots of age for each gender.

   (d) (2 marks) The same data set, now in a file `survey.txt` (2017: `survey2.txt`) in your current R working folder, can be read into a data frame called `survey2` (2016: `survey`) in R, using `read.table` with a suitable value for `header`. What value? (2017: using `read_delim`, somehow. Show how.) And what names will the columns have?

   (e) (3 marks) What R code will give the columns names that describe what the columns contain?

   (f) (3 marks) Use something from `dplyr/tidyverse` to find the mean income for each gender. Give the code you would use. (You may assume that `dplyr` has already been loaded.)

2. A clinic provides a program which is supposed to help their clients lose weight. They take a sample of 15 clients in the program. They weigh each client before the program begins, and again three months later. The data are shown in Figure 3. The clinic has just hired a Statistics graduate student to help them determine whether the program is effective or not.

   (a) (2 marks) Why are these data matched pairs rather than two independent samples? Explain briefly.

   (b) (3 marks) The graduate student has just learned about the spaghetti plot, and wants to draw one for the clinic. Figure 4 shows a line of code that the graduate student used in preparation for making the spaghetti plot. Describe briefly what columns the data frame `wtloss2` contains, and what values are in each column. (If you want to list the values, list only a few, enough to show that you know what they are.)

   (c) (3 marks) What do you learn from the spaghetti plot, shown in Figure 36? Explain briefly. (Note that this plot is shown at the *end* of the booklet of code and output.)

   (d) (2 marks) Why is the spaghetti plot more informative for these data than side-by-side boxplots of the before and after weights would be? Explain briefly.

3. Some psychologists study "stereotype threat". Consider people who identify as members of a group characterized by the ability to perform some task at a high level. According to the theory, those people can feel threatened if they are told that some other group can perform that task better. This threat has the effect of lowering their performance at the task they are good at.

   To assess whether stereotype threat occurs in practice, 23 white male students, who were known to be good at math, were were recruited for a study. These students were randomly divided into two groups. The first group, a control group labelled `exam`, were asked to write a difficult math exam. The second group, before writing the same exam, were told "Asian students typically do better than other students in math exams". This group was labelled `threat`. For each student, the score on the math exam was recorded, a higher score being better. Our task is to investigate whether the threat did indeed lower the scores of the students in the second group.

   The data are shown in Figure 5.

   (a) (2 marks) Side-by-side boxplots of exam scores by group (exam only or exam plus threat) are shown in Figure 6. Do the boxplots appear to support the researchers' hypothesis? Explain briefly.

   (b) (2 marks) By looking at the boxplots in Figure 6, give *two* reasons why you might be unwilling to run a two-sample t-test on these data. (The recruited white male students were divided into two groups of approximately equal size; a reason that applies to both groups counts as two reasons.)

   (c) (2 marks) Look at the computations in Figure 7. Explain briefly what `obs` and `omd` contain.

   (d) (4 marks) Look at the function in Figure 8. Describe in words (i) what kind of input the function requires, (ii) what the function does (without using the word "sample"), (iii) what the function returns to the outside world.

   (e) (3 marks) In Figure 9, a randomization test is carried out. What do you conclude from it, in the context of the data? Explain briefly. (2017 on: don't answer this question. Think about how you might analyze these data instead.)

4. A small New England college has historically admitted students with a mean SAT score of 520, with a standard deviation of 80. SAT scores are calibrated to have an approximately normal distribution. This year, the college admissions department has started a campaign with the aim of increasing the quality of the students admitted to the college. They are hoping that the mean SAT of students admitted next year will be 530. The plan is to admit 100 students next year.

   (a) (2 marks) The college administration plans to run a suitable t-test next year to determine whether the mean SAT score has increased. Why would the college admissions department be interested in the *power* of this test? Explain briefly.

   (b) (3 marks) If the college achieves what it hopes, how likely is it that they will be able to demonstrate a statistically significant increase in mean SAT score? Use the appropriate one of Figures 10 through 13 to obtain your answer, and explain briefly why your choice is correct. If none of the choices is appropriate, state this and explain briefly why.

   (c) (2 marks) In each of Figures 10 through 13, `power.t.test` is used. The admissions department is curious about how many students it would have to admit to make the power of its test equal to 0.80. How would you change the code used in the appropriate one of these Figures to find this out?

5. Plants are much more sensitive to environmental light than humans. Experiments on plants can therefore be affected by using regular light. However, if the experimenter works in darkness, they cannot see what they are doing! Experiments on plants are usually performed in a dark room under "safelight", under which the experimenter can see, but where the plants are minimally affected.

   In an experiment, two different kinds of safelight (labelled A and B) were tested, each at high and low intensities (labelled H and L, so that for example BH means safelight B at high intensity). A control group, Darkness (labelled D) was also used. 40 seedlings were randomly allocated to one of the five groups (8 seedlings to each). The seedlings were allowed to grow for 20 days, and at the end of that time, their height was measured. The question of interest is "do either of the safelights, at either of the intensities, have an effect on plant height?"

   The structure of the data is shown in Figure 14.

   (a) (3 marks) The researcher intends to compare the heights for the different groups using analysis of variance. What two assumptions about ANOVA do the boxplots in Figure 15 enable you to assess? Do those assumptions appear to be satisfied? Explain briefly.

   (b) (2 marks) Figure 16 and 17 show (respectively) an analysis of variance and Mood's median test for these data. Which test do you think is better? Explain briefly.

   (c) (3 marks) Looking at the appropriate one of Figures 16 and 17, what P-value do you obtain? Express your P-value in scientific notation or as a decimal number. What, precisely, do you conclude? Explain briefly. (Mood's median test has a "warning: chi-squared approximation may be incorrect". You may ignore this.)

(d) (2 marks) Look again at Figure 17. What is it about the table `tab` that would make Mood's median test give the kind of P-value that it did? Explain briefly. (Answer this even if you think the ANOVA is the better test to use.)

(e) (2 marks) What would be your recommendation for the kind of safelight to use, bearing in mind the purpose of a safelight? Explain briefly. (You might like to look back at the boxplots.)

6. The salaries of a number of employees at a company were recorded. Along with the employee's salary, three other variables were recorded: the level of education attained by the employee (1 is Bachelor's, 2 is Master's, 3 is PhD), the number of years of experience after completing their degree, and the number of employees supervised in the employee's current position. The aim is to see which, if any, of these three variables helps to predict salary.

   The data are read in and some of the values are shown in Figure 18.

   Note that throughout this question, the variable `degree` is treated as being numeric (so that each extra level of education is worth one extra "point", so to speak). You may assume that this is a reasonable way to handle `degree` for these data.

   (a) (3 marks) A regression is run for predicting salary from the other three variables. The output and graphs are shown in Figures 19 and 20. Do you think this regression is appropriate to describe the relationship between salary and the other variables? Explain briefly, referring to items in the output as necessary. You should assess all the relevant items in the output.

   (b) (2 marks) Some output from `proc transreg` is shown in Figure 21. Looking at this output, what do you conclude?

(c) (2 marks) The analyst decides to predict log of salary from the other variables. (This may or may not have been suggested by any of the previous output.) The output is shown in Figure 22 and Figure 23. Is there anything that would cause you to doubt the appropriateness of this regression? Explain briefly.

(d) (3 marks) Think about what you know or can guess about the type of relationship being investigated here. For the regression in Figure 22, would you expect the slopes to be positive or negative? Are they? Explain briefly.

(e) (2 marks) What, if anything, would you do next, after the regression in Figure 22? Your choices are (i) to remove one or more explanatory variables, (ii) to use a different transformation than the analyst did, (iii) declare yourself satisfied with the regression. Explain briefly.

7. Back in 2000, an instructor at the University of California Davis collected some information about the students in her class. There were a lot of variables, but we will focus on these:

   - `Height`: the student's height (inches)
   - `GPA`: the student's GPA (4-point scale)
   - `Sex`: whether the student identified as `Male` or `Female`
   - `Alchol`: the number of alcoholic drinks consumed in the past week
   - `momheight`: the height of the student's mother (inches)
   - `dadheight`: the height of the student's father (inches)

   The variable name `Alchol` is mis-spelled. That's how it was in the original data. I wonder if the instructor had been drinking too much alchol.

   Some of the relevant data are shown in Figure 24.

   (a) (2 marks) What does the code in Figure 25 do? Explain briefly.

   (b) (2 marks) Figure 26 shows a regression predicting students' height from the other variables. Based on this output, what regression would you fit next? Explain briefly. (If you are happy to stop at this regression, say so, but you should also offer an explanation.)

   (c) (2 marks) Look at the regression in Figure 27, which may or may not be your preferred regression from part (b), and the output in Figure 28 below it. What do you conclude from Figure 28? Explain briefly.

   (d) (2 marks) Figure 29 shows a residual plot for the regression in Figure 27. Do you see (i) no problems, (ii) one problem, (iii) two or more problems? Explain briefly.

8. A manager has data on sales of asphalt roofing shingles from data in 26 sales districts. For each district (numbered 1–26), the recorded variables are:

   - the annual sales (in thousands of dollars),
   - the promotional expenditures (thousands of dollars),
   - number of active accounts,
   - number of competing brands,
   - district potential (a coded value, but a higher value indicates more sales potential).

   The data are shown in Figure 30. The manager is interested in seeing whether any of the data values recorded are possible errors. 2018: skip this question.

   (a) (1 mark) What is the name of the quantity calculated in the first three lines of Figure 32?

   (b) (2 marks) What is the *purpose* of the calculation in the last line of Figure 32?

   (c) (2 marks) Which sales districts are possible errors, according to the criterion given here?

   (d) (4 marks) For the observation with the most extreme value from the output in Figure 32, describe what made it come out that way, referring to Figure 31 as necessary, and using what you know or can guess about sales of a product.

9. The data in Figure 33 are (some of the) results from the English Premier soccer league for the 2016-2017 season, which started in August 2016. The columns are:

   - The date of the match, year first, then month, then day.
   - The name of the home team (text, 27 characters long including the trailing spaces)
   - The name of the away (road) team (likewise, text, 27 characters long including the trailing spaces)
   - The number of goals scored by the home team
   - The number of goals scored by the away team

   There are 130 lines of data altogether.

   The columns are separated by commas. Imagine that the (entire) data set has been uploaded as `england.csv` to your file storage on SAS Studio, and also to your working folder on R Studio.

   (a) (4 marks) 2016: Give a SAS data step to read the data into a SAS data set, so that the variables are read in properly and are of the appropriate types. (By "data step", I mean SAS code starting with `data` and the name of a data set.) 2017: skip this; see the next part.

   (b) (3 marks) Write a `proc import` statement to read in the same data.

   (c) (2 marks) Starting from the data set you read in with `proc import`, what code would save it as a permanent data set (on disk)?

   (d) (2 marks) 2017: we might not have done this either. 2016: Imagine you have closed down SAS and started it up again. What code would display the first 8 rows of the permanent data set that you created in the last part, *without* using a `data` step or `proc import`?

(e) (3 marks) Now the same data in R. Give code that will read in the data of Figure 33 *as text rather than factors*, and will also create a data frame where the dates are stored as R Dates.

(f) (2 marks) What R `dplyr` code would display all the games that the team Liverpool played (both home games and away games)?

(g) (2 marks) What R `dplyr` code would show, for each different date:

- the total number of goals scored,
- and the number of games played,

on that date?

10. A study was conducted to evaluate the performance of a diesel engine run on three different types of fuel. The response variable was called the Mass Burning Rate, and it was thought to depend on both the Fuel type and on the Brake Power. Three observations were (intended to be) taken at each brake power; each observation is labelled with a unique one-letter ID. You may assume that the data have been read into a SAS data set called `synfuels`, as shown in Figure 34.

    (a) (4 marks) A plot is shown in Figure 35. Your task is to give the SAS code that produced this plot, bearing in mind all the features of the plot.

    (b) (3 marks) How would you label the points with the `id` *only* for the `AdvancedTiming` fuel, and not for the other fuels? Give code to accomplish this. In the part of your code that draws the revised plot, you only need to describe any *changes* to your code from (a) (that is, you don't need to write out the whole thing again).