University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAC32 (K. Butler), Final Exam
December 7, 2017 9:00-12:00

IT IS ASSUMED THAT YOU HAVE READ THE BOX BELOW.

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

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 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.

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.

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:

```r
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	7	
2	10	
3	11	
4	6	
5	10	
6	8	
7	8	
8	10	
9	8	
10	8	
11	8	
12	10	
Total:	104	

1. Four weight loss programs are being compared. These are: (i) a low calorie diet, (ii) a low fat diet, (iii) a low carbohydrate diet, (iv) a placebo diet (the participants in this group are told that they are "participating in a study of healthy behaviour"). Twenty participants are randomly allocated to one of these four programs. Each participant's weight is measured at the start, and again after 8 weeks. The response variable is the weight loss, measured in pounds (the weight at the start minus the weight at the end). A larger weight loss indicates that the diet is more helpful. A negative weight loss indicates that the diet actually resulted in a *gain* of weight.

   The data are shown in Figure 1. We will be analyzing the data in R. The diet names are written as single words.

   (a) (2 marks) Explain briefly what feature the data has that will allow it to be read in with `read_delim`.

   (b) (2 marks) Why would analysis of variance be a plausible technique to use to analyze these data? Explain briefly.

   (c) (3 marks) Explain briefly how the data in Figure 1 are not currently in the correct format for an analysis of variance. In answering this question, you can describe what the correct format would be, and how the format shown in the Figure differs from that.

(d) (3 marks) The data in Figure 1 have been read into a data frame called `diets`. Using this data frame, give R code that will create a new data frame called `diets2` that contains the data in the appropriate format for the analysis of variance.

(e) (3 marks) Using your new data frame `diets2`, give R code to run an analysis of variance comparing the diets, and to display the F-test from the analysis of variance.

(f) (4 marks) The ANOVA and Tukey analyses for these data are shown in Figures 2 and 3. Give a complete conclusion based on these two Figures, in the context of the data. In your conclusion, explain briefly whether or not you need to refer to Figure 3, and if you do, why.

2. In a packing plant, a machine packs cartons with jars. A new packing machine is being tested. Will it pack faster on average than the machine currently used? To test that hypothesis, the times it takes each machine to pack ten cartons are recorded. The data are shown in Figure 4, and have been uploaded as a file `packing.txt` to my account `ken` on SAS Studio.

   (a) (3 marks) Give code to read the data into a SAS data set called `packing`. Look carefully at the format of the data before you write any code.

   (b) (2 marks) Give SAS code to find the mean packing times classified by whether the machine is old or new. (Output that contains the mean and other things is fine.)

   (c) (4 marks) The analyst at the packing company wants all the packing times in one column of a SAS data set, with a second column saying which machine the packing time came from. The new columns should be called `packtime` and `machine` respectively. The columns from the original data set should be removed. Give SAS code to create a data set in this format.

   (d) (2 marks) Using the new data set created in the previous part, give code to find the mean packing time classified by whether the machine is old or new. (Output that contains the mean and other things is fine.)

3. A researcher is planning a study to look at a new treatment. The mean of the standard procedure (in suitable units) is 12, but the researcher hopes that the new treatment will increase the mean to 16. The standard deviation of the measurements is believed from previous studies to be about 8.5. The researcher will use a one-sample t-test. The researcher's primary question is: how large a sample size is needed to obtain probability at least 0.75 of correctly rejecting the null hypothesis that the mean is 12, in favour of an alternative that the mean is greater than 12?

   (a) (3 marks) Give SAS code to find the necessary sample size.

   (b) (3 marks) The output of your code from part (a) is shown in Figure 5. Figure 6 shows the output from almost the same code, except that I added `alpha=0.01` to the code, to carry out the test at $\alpha = 0.01$. Compare the sample sizes from Figures 5 and 6. Are the results surprising, or not? Explain (reasonably) briefly.

4. A study is undertaken to explore how age is related to sense of smell. Altogether, 180 subjects, aged between 20 and 89, are exposed to 40 different odours: for each odour, subjects are asked to choose which one of four words best describes the odour. A "smell score" is calculated for each subject, based on the number of odours correctly identified (and therefore a higher smell score is better). A subject's age is recorded as an age group as follows: group 1: age ≤ 25; group 2: $25 <$ age ≤ 40; group 3: $40 <$ age ≤ 55; group 4: $55 <$ age ≤ 70; group 5: age > 70. A higher-numbered age group thus contains older people.

   It is believed that smell score will decrease with age on average, but that in older people, some will have an excellent sense of smell, but some will have a much worse sense of smell.

   Some of the data set is shown in Figure 7. (There are too many observations to show them all.)

   (a) (2 marks) Figure 8 shows side-by-side boxplots of smell score by age group. Because there is a lot of data, the statistician on the study is not concerned about either the skewness or the outliers. What other assumption is made in order to do analysis of variance, and how is that assumption not satisfied here? Explain briefly.

   (b) (2 marks) Explain briefly why Welch's method for analysis of variance is preferable to the standard ANOVA here. (There are two points to make.)

   (c) (3 marks) Give SAS code to run the Welch ANOVA for these data.

   (d) (3 marks) Give the F value and P-value for the Welch ANOVA, as shown in Figure 9. What do you conclude, in the context of the data?

5. This question is about R Markdown.

   (a) (2 marks) Suppose you have just opened a new R Markdown document. What code would insert a new section called "Introduction"?

   (b) (2 marks) How would you add a code chunk to your R Markdown document? Give the code you would use, or describe the procedure you would follow.

   (c) (2 marks) Describe briefly in your own words why it is a good idea to use an R Markdown document (at least if you are coding in R), compared to copying and pasting code and output into a Word document.

   (d) (2 marks) Your boss looks at the output from your R Markdown document and says "this would be better if you had a histogram *right there*" (and points at the spot). When you get back to your computer, what steps would you take to produce a new output for your boss with the histogram in the right place? Describe those steps. (There are two or three steps depending on how you count them.)

6. In Chile in 1988, there was a "plebiscite" or special election, to determine whether army general Augustin Pinochet should continue as President for another 8 years. Before the plebiscite, a survey was taken of 2700 respondents. For each respondent, the following variables were recorded. They appear in the data frame in this order:

   - `region` of Chile where the respondent lives: SA, in Santiago; M, in the metropolitan Santiago area; N, in northern Chile; S, in southern Chile; C, in central Chile.
   - `population`: the population of the city where the respondent lives.
   - `sex`: M (Male) or F (Female).
   - `age` in years.
   - `education`: P: primary (elementary) only; S: secondary (high school): PS: post-secondary (university or college).
   - `income`: monthly income in pesos.
   - `statusquo`: result of a questionnaire about General Pinochet. A positive number means the respondent likes Gen. Pinochet, a negative number means they dislike him.
   - `vote`: how the respondent intends to vote: Y (yes, for Pinochet), N (no, against Pinochet), A (will abstain, that is, not vote at all), U (undecided).

   The data frame is called `chile`. Give R code, using ideas from the `tidyverse`, to accomplish the following tasks:

   (a) (2 marks) Display only the columns containing sex and education.

   (b) (2 marks) Display the columns `age`, `income` and `vote` *without naming* the columns.

   (c) (2 marks) Display the first 15 rows of the data frame.

   (d) (2 marks) Display a random sample of 5 respondents.

(e) (2 marks) Display the respondents who earn more than 70,000 pesos per month.

(f) (3 marks) Display the twelve lowest `statusquo` scores, along with an indication of how each of these respondents intends to vote.

7. A marketing researcher studied the sales of a product that was introduced ten years ago. The data are shown in Figure 11. The `year` is the number of years since the product was introduced, and the `sales` are annual sales, so that for example the sales for year 0 are the sales for the first year after the product was introduced (between 0 years and 1 year after introduction). The researcher is interested in modelling the relationship between sales and time.

   (a) (3 marks) A scatterplot is shown in Figure 12. Give the SAS code that was used to create this plot.

   (b) (2 marks) Figures 13 and 14 show the output from fitting a regression predicting sales from year. Would you say that the regression model fits well or badly? Cite a *number* from the output to support your answer, and explain briefly how it does that.

(c) (3 marks) Look at Figure 14. Is there any evidence that the regression is unsatisfactory? You should assess at least two of the graphs in the Figure. If you think the regression is unsatisfactory, suggest a way of improving it.

(d) (3 marks) Figure 15 shows some SAS code and its output. What do you learn from this Figure, and what does it suggest to do next? Explain briefly.

(e) (2 marks) Compare your answers to parts (c) and (d). Are they consistent or inconsistent? Explain briefly.

8. The "performance IQ" or PIQ is one aspect of intelligence. This is understood to depend on brain size (measured from MRI scans) and body size (height, in inches; weight, in pounds). Data on these variables were collected for a sample of 38 individuals. Part of the dataset is shown in Figure 16. **Use $\alpha = 0.01$ for all the hypothesis tests in this question.**

   (a) (2 marks) A regression predicting performance IQ from the three explanatory variables is shown in Figure 17. Even at $\alpha = 0.01$, explain briefly why this output *does not* permit us to remove both `Height` and `Weight` from this regression.

   (b) (2 marks) Give R code to fit the model of Figure 18, *without* using `lm` and without a `data=`. You may assume that the model `piq.1` shown in Figure 17 has already been fitted. (A correct answer can be obtained in one line of code.)

   (c) (4 marks) A second regression is shown in Figure 18, and another test is shown in Figure 19. What do you conclude, at $\alpha = 0.01$, from the test in Figure 19, in the context of the data? What does that tell you about the appropriate model to use to predict performance IQ? Explain briefly.

9. In 2008, a major college in the US was monitoring salaries to see whether there were systematic differences in salary between male and female faculty members. Part of the data set is shown in Figure 20. The variables shown are:

   - `rank`: Assistant Professor (lowest), Associate Professor, Professor (highest)
   - `discipline`: classified as `A` ("theoretical") or `B` ("applied")
   - `yrs.since.phd`: number of years since Ph. D. earned
   - `yrs.service`: number of years since hired (at this college)
   - `sex`: Female or Male
   - `salary`: in (US) dollars (response).

   (a) (2 marks) Figure 21 shows the results of a *t*-test for comparing salaries of male and female faculty members. The P-value is very small. However, despite this, this *t*-test is not the strongest evidence that females are being discriminated against. Explain briefly why not. (You may assume that the distributions of salaries within each gender are sufficiently close to being normal.)

   (b) (3 marks) Look at Figure 22. Some of the explanatory variables here are categorical rather than quantitative. Explain briefly why the two coefficient estimates that start with `rank` have values that make sense, given what the data represents.

   (c) (3 marks) What precisely does Figure 22 tell you about the evidence for gender discrimination in terms of salary? Explain briefly.

10. A staff analyst at a manufacturer of microcomputer components has compiled monthly data for the past 16 months. The variables are as follows:

    - `month` and `year`: the month and year for which the data was collected. The month and the year are both numbers.
    - `comp`: the value of the manufacturer's components sold (in thousands of dollars)
    - `indproc`: the total value of all industry production that uses these components (in millions of dollars).

    The file has been read into a SAS data set called `micro`, with `month` and `year` stored as numbers. The data set as read in is displayed in Figure 23.

    (a) (3 marks) Give code to create a new data set that contains the month and year turned into an actual SAS date called `date`. The dates you create should be the first day of the appropriate month, and the columns `month` and `year` of the read-in data set should be removed.

    (b) (2 marks) If I run simply `proc print` on the new data set, how will the dates be displayed?

    (c) (2 marks) What code would I add to `proc print` to display the dates in UK format, that is, day-month-year with the month as a number and the year as 4 digits?

    (d) (3 marks) What code could I use to plot `indproc` against time, joining the points by lines, and making sure the time axis is appropriately labelled? (*You* have to decide what "appropriately" means here.)