University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAC32 (K. Butler), Final Exam
December 14, 2018 19:00-22:00

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

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.

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)
library(broom)
library(ggrepel)
```

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	11	
2	12	
3	9	
4	12	
5	10	
6	11	
7	9	
8	11	
9	9	
10	11	
11	9	
Total:	114	

1. The data in Figure 1 (in the booklet of code and output) show grain yields in modern and historic times, measured in kilograms per hectare, for 16 grain-producing regions of the world. The data have been uploaded to your SAS Studio account, in a file called `grain-yields.txt`.

   (a) (3 marks) Give SAS code for reading the data into a SAS data set called `grain`.

   (b) (2 marks) Give SAS code to display only the two columns of your data set called `Modern` and `Historic`.

   (c) (3 marks) Give SAS code to produce an appropriate graph of this data set, bearing in mind that the column `Region` simply labels the regions and is not a variable as such.

   (d) (3 marks) Give SAS code to calculate the median and interquartile range of the modern and the historic grain production.

2. Blood pressure is measured with two numbers: the "systolic" (larger) number, and the "diastolic" (smaller) number. The systolic blood pressure is thought to depend on age and weight. Data are available for a sample of 11 men. The data are shown in Figure 2.

   (a) (2 marks) Give code to read the data *into an R data frame* called `bp`, as we did it in class. Assume that the data are in a file `bp.txt` in the same folder as your current project in R Studio.

   (b) (3 marks) Give R code to create a new column in your data frame called `weight_class` which has the values `heavy` if each man's weight is above the mean weight, and `light` otherwise. Use `tidyverse` ideas. (You do not need to save the new data frame.)

   (c) (3 marks) Suppose that the data frame you created in the previous part is saved in data frame `bp2`. Give R code that uses this data frame to plot systolic blood pressure against age, distinguishing the points by whether each man is lighter or heavier than the mean.

   (d) (4 marks) Give R code to run a multiple regression predicting systolic blood pressure from age and weight, and to display the results, including output like Figure 3. If output like what you see in the Figure would be produced by your code, I don't mind what other output would be produced. You should use the (quantitative) weight as read in from the file, not the categorical `weight_class` that you created earlier.

(e) (3 marks) Part of the output from your regression is shown in Figure 3. Imagine that you have two men of the same (unknown) age, but the second man is 10 pounds heavier than the first man. What precisely does the regression output tell you about how their systolic blood pressures will compare on average? Explain briefly.

3. A child psychologist reads in the literature that children typically start walking at 12.5 months. The child psychologist believes this is incorrect (as an average for all children), so she takes a sample of 18 children and measures the age (in months) at which each child starts walking. The data are shown in Figure 4. This is a SAS data set called `walking` with the column names as shown. Assume that it is the most recently created data set.

(a) (2 marks) Give SAS code to produce the graph in Figure 5.

(b) (4 marks) Use Figure 5 along with the output in Figure 6 to assess the evidence for the child psychologist's belief. What do you conclude, in the context of the data? Make it clear which parts of Figure 6 you are using.

4. Rabies is a deadly virus that spreads to people from the saliva of infected animals. The rabies virus is usually transmitted through a bite. Animals most likely to transmit rabies include bats, coyotes, foxes, raccoons and skunks.

   A study was made of rabies cases in foxes in southern Germany. 31 locations were sampled altogether in two regions, labelled A and B here. The data, in a SAS data set called `foxrabies`, are shown in Figure 7. The values in `cases` are counts of the number of cases of rabies at that location. Some analysis is shown in Figure 8.

   (a) (2 marks) Suppose that one interest of the researchers who collected the data was to see whether the mean number of cases per location exceeded 3, regardless of region. Explain briefly why Figure 8 *does not* enable you to assess the evidence for this.

   (b) (3 marks) Give SAS code that will produce output to allow the researchers to assess their interest of the previous part. If your code will produce output that almost addresses the researchers' interest but does not quite do so, explain briefly what you will need to do with the output from your code. You may assume that the distribution of the number of cases is approximately normal.

   (c) (4 marks) What is the main conclusion that you *can* draw from Figure 8? Explain briefly. Assume that anything needing to be approximately normal *is* approximately normal; if you need to make any other assumptions, make them and state what additional assumptions you make.

   (d) (3 marks) Look at Figure 9. What do you conclude about the appropriateness of a t procedure here? Explain briefly.

5. As the world population increases, it is important to grow enough grain to feed everyone. Sixteen large grain-producing regions were sampled. The production of grain (in kilograms per hectare) was recorded for each of these regions, both modern production and historical production. The historical production figures come from the mid-20th century. Do we have evidence that modern production is higher than historical production? The data are shown in Figure 1. We previously used this data set for another question.

   You may assume that you have a SAS data set called `grain` that contains the data from Figure 1, and that this is the most recently-created data set.

   (a) (2 marks) To compare modern and historic grain production, why would a matched pairs test be better than a two-sample test? Explain briefly.

   (b) (2 marks) Give SAS code to carry out a suitable matched-pairs t-test for these data.

   (c) (2 marks) What key assumption is made in order to be able to trust the paired t-test?

   (d) (4 marks) Give SAS code to produce a graph that will enable you to assess the assumption you named in the previous part. If you need to make any design decisions for your graph, it is up to you to decide on something sensible.

6. A personnel officer developed four different aptitude tests and administered them to 25 new entry-level clerical employees. Each of these employees went through a "probationary period" in which they could get accustomed to the work involved. At the end of each employee's probationary period, they were assessed for proficiency on the job. The personnel officer would like to know whether any of the aptitude tests were able to predict job proficiency, and, if so, which aptitude tests. The data are shown in Figure 10. This display is of a SAS data set called `jobs`, which is the most recently created one.

   All the test and proficiency scores are numbers; they have no units.

   (a) (2 marks) Give the SAS code to run the regression whose output is shown in Figures 11 through 13.

   (b) (3 marks) In Figure 11, what is the parameter estimate for the aptitude test labelled `test2`? What does the value mean, in the context of the data?

   (c) (4 marks) Look at Figures 12 and 13. We are looking for any problems with the regression. Assess: (i) plot of residuals against the fitted values, (ii) normal quantile plot of residuals, (iii) plots of residuals against explanatory variables. For each of these, indicate which plot(s) on the Figures you are looking at, and what you conclude. In addition, make an overall comment about the appropriateness of the regression.

   (d) (2 marks) What *one* thing would you do next to improve this regression? Explain briefly.

7. Figure 14 shows an R data frame called dd, with a column called row of row numbers, and two data columns called xx and yy. A scatterplot of yy against xx is shown in Figure 15, with the data points labelled by which row of the data frame they come from.

   (a) (3 marks) Give the R code that was used to produce Figure 15.

   (b) (4 marks) Figure 16 shows a function rsq that takes as input any data frame d that has columns xx and yy, runs a regression predicting yy from xx, and outputs the R-squared from that regression. Figure 17 shows a function omit1 that takes as input two things: (i) any data frame d that has columns xx and yy, and (ii) a row number i. The function *omits row i*, and then runs the regression of (the remaining) yy on (the remaining) xx, outputting the R-squared from that regression.

   I will not be asking how these functions work. You may take it for granted that they do what I say they do, even if you don't see how they do it.

   Your task is to add a new column called rsquared to data frame dd. In that column should be the value of R-squared for the regression of yy on xx using all the observations *except* the one in that row. For example, the first entry in the column rsquared should be the value of R-squared for the regression of yy on xx using all the xx and yy values except for the first (xx, yy) pair.

   Give R code that will create this new column rsquared, using tidyverse ideas, and using the function omit1 defined in Figure 17. You do not have to save your new data frame.

   (c) (2 marks) The output from running your code of the previous part is shown in Figure 18. Describe how the outlier influences the results and explain briefly why it has the effect it does.

8. There is evidence that smiling can affect judgments of possible wrongdoing. This phenomenon, termed the "smile-leniency effect", was the focus of a study in 1995.

   136 subjects were asked to judge a case of possible academic misconduct. The subjects each received a file of evidence that a student had cheated on an exam. This evidence was the same for all the subjects except for one thing: a photo of the student (who allegedly cheated) with one of four facial expressions: a "felt smile" (that is, a genuine smile), a false smile, a miserable smile and a neutral face. The facial expression was randomized to subjects, with each type of smile appearing the same number of times. (Yes, there is such a thing as a "miserable smile", apparently.)

   After reviewing the evidence and looking at the photo they received, each subject had to respond to two questions on a 9-point scale. First, how they felt about the case as presented (the student should be given the benefit of the doubt, strongly disagree to strongly agree). Second, what they think an appropriate punishment should be (scale: cleared of all charges, up to given the maximum academic punishment). Each subject's answers were combined into a "leniency score", with a higher value meaning more lenient (the subject felt that the student should be punished less and/or that the student should be given the benefit of the doubt).

   Some of the data set is shown in Figure 19, and a table of sample sizes, means, and medians for each group (for all the data) is shown in Figure 20. The data frame is called `smile_leniency`.

   (a) (4 marks) Give the R code that was used to make the plots in **Figure 22**.

   (b) (2 marks) Data of this kind are often analyzed using analysis of variance. What are the two major assumptions required for an analysis of variance to be appropriate?

   (c) (3 marks) Use any or all of Figures 20 through 22 to assess the assumptions you named in part (b). Do you conclude that the assumptions are satisfied, or not? Explain briefly.

   (d) (2 marks) Figures 23, 24 and 25 show three possible analyses of these data. Which analysis do you think is most appropriate? Explain briefly.

(e) (2 marks) Based on what you said in part (d), what do you conclude from your most appropriate analysis, in the context of these data?

(f) (3 marks) Figures 26, 27 and 28 show some further analysis. Which, if any, of these Figures should you look at? What do you conclude, in the context of the data? (If you are not entitled to look at any of these Figures, explain briefly why.)

9. A manufacturer of felt-tip markers wanted to know whether a new advertising display, featuring a picture of a physician, was more effective at selling the markers than the current display, featuring a picture of an athlete. The current display was located in the stationery area (of a drugstore chain), while the new display could be located in either the stationery area or at the checkouts. Fifteen stores took part in the study. In all of them, the current display (the athlete picture, located in the stationery area) was used for three weeks, and the sales recorded. Then, for the next three weeks, each store was randomly allocated one new display: either (i) the athlete display in the stationery area, as before (control), (ii) the physician display in the stationery area, or (iii) the physician display at the checkouts. Sales were then recorded for those next three weeks, in suitable units. The data are shown in Figure 29. `sales1` and `sales2` refer to the sales in the first and second three-week periods.

(a) (2 marks) Why do you think the sales for the first three-week period were recorded, even though the new display was not being used?

(b) (2 marks) A regression model was fitted, predicting `sales2` from `sales1` and `Treatment`. Output from the fit is shown in Figures 30 and 31. What do you conclude from Figure 30, in the context of the data? Explain briefly.

(c) (3 marks) In Figure 31, what precisely does the number 19.12547 tell you, in the context of the data?

(d) (2 marks) The previous regression assumed that the rate at which sales in the second three-week period increased with sales in the first three-week period was the same for each treatment. Is that supported by the graph in Figure 34? Explain briefly. (Note that the graph is at the end of the Booklet of Code and Output.)

10. You might recall the 68–95–99.7 rule from a previous course. It is sometimes called the "empirical rule". It says that 68% of the values in a normal distribution with mean μ and standard deviation σ lie within one standard deviation of the mean, 95% of them lie within two standard deviations of the mean, and 99.7% of them (that is, almost all) lie within three standard deviations of the mean.

For those who prefer this mathematically: if X has a normal distribution with mean μ and SD σ, then $P(\mu - \sigma < X \le \mu + \sigma) = 0.68$, $P(\mu - 2\sigma < X \le \mu + 2\sigma) = 0.95$, and $P(\mu - 3\sigma < X \le \mu + 3\sigma) = 0.997$.

This idea gives us an alternative way to estimate `sigma` for use in a SAS normal quantile plot, as we explore. You do not need any normal tables for this question.

(a) (4 marks) Using this rule, explain briefly why $\mu - \sigma$ and $\mu + \sigma$ are the 16th and 84th percentiles (respectively) of a normal distribution.

(b) (2 marks) Consider the difference d between the 84th percentile and the 16th percentile. In a normal distribution, what is d in terms of μ and σ? (d might depend on one or both of these.)

(c) (2 marks) How, therefore, could you use the 16th and 84th percentiles of sample data to estimate σ? Explain briefly.

(d) (2 marks) Consider the data in Figure 32, whose 16th and 84th percentiles and median are shown in Figure 33, labelled `x16`, `x84` and `x50` respectively. Use these values to estimate `mu` and `sigma` for a SAS normal quantile plot.

(e) (3 marks) Give SAS code to draw a normal quantile plot of the data `x` that are shown in Figure 32, using your estimated μ and σ.

(f) (2 marks) What would be the advantage to estimating μ and σ as in this question, rather than the way SAS does it using `mu=est sigma=est`? Explain briefly.

This page: ______________ of possible 9 points.