University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAC32 (K. Butler), Final Exam
December 19, 2019

Aids allowed (printed or handwritten): My lecture overheads (slides); Any notes that you have taken in this course; Your marked assignments; My assignment solution; Non-programmable, non-communicating calculator.

This exam has 13 numbered pages of questions. Check to see that you have all the pages. There is an additional empty page that you can use if you need more space for any answers.

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).

The maximum marks available for each part of each question are shown next to the question part.

You may assume throughout this exam that the code shown in Figure 1 of the booklet of code and output has already been run.

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.

Question 1 (14 marks)

"Time-of-day pricing" is a plan by which electricity customers are charged at a higher rate for using electricity at peak hours (hours at which there is a large demand for electricity), and a lower rate for use at off-peak hours. A study was carried out by a large electricity company to measure customer satisfaction with various pricing schemes. The study consisted of two factors: price ratio (how many times more expensive peak-hours electricity is than off-peak electricity; ratios of 2:1, 4:1 and 8:1) and peak period length (6, 9, or 12 hours).

For each combination of price ratio and peak period length, known as a "plan", four customers were randomly selected and charged for electricity use according to that plan for a certain time. At the end of this time period, they were given a questionnaire that assessed satisfaction with the plan they had been on. The questionnaire results were summarized into an overall level of satisfaction, with a minimum (worst) value of 10 and a maximum (best) value of 38. The data are shown in Figure 2.

(a) (4 marks) The data are stored in a file on your SAS Studio called `timeofday.txt`. Give SAS code to read in and display the data set.

(b) (3 marks) What SAS code would display the mean satisfaction score for each combination of peak period length and price ratio? (It is fine if your output will include other things as well as the mean.)

(c) (2 marks) Do you have a `var` line in your code for the previous part? Do you need one? Explain briefly.

(d) (3 marks) The electricity company wanted to see how customer satisfaction depended on the *combination* of peak period length and price ratio. To do this, they made the graph shown in Figure 23. (This is at the end of the Figures because it is in colour.) Give the SAS code that was used to make this graph.

(e) (2 marks) Look again at Figure 23. The peak period lengths are labelled "06", "09" and "12" so that they come out in a sensible order on the graph. For all the price ratios, the "average" cost of electricity is the same; this means that when the ratio is 8–1, the peak price is highest and the off-peak price is lowest, compared to all the other ratios. On the boxplot boxes, SAS uses the symbol O, X or + (one symbol for each colour of box) to denote the mean of the data in that box. Describe one thing you can conclude from Figure 23 about average (mean or median) satisfaction as it depends on peak period length and/or price ratio, and explain briefly why your conclusion makes sense, based on what you know or can guess about electricity prices.

Question 2 (15 marks)

Bulimia is an eating disorder. People who suffer from bulimia have an unrealistic body image, and will consume a lot of food at one time, followed by feelings of guilt or shame. Such people often have a fear of being evaluated negatively by others.

25 female students took part in a study. 11 of them suffered from bulimia and the other 14 had "normal" eating habits. Each of the students also completed a questionnaire called FNE which evaluates the "fear of negative evaluation", a higher score on the FNE indicating a greater fear. We are interested in finding out whether people suffering from bulimia tend to have a greater fear of negative evaluation than people who do not.

The data are shown in Figure 3.

(a) (4 marks) A suitable t-test is run, with output shown in Figure 4 and Figure 5. Give the SAS code that produced all of this output.

(b) (2 marks) The statistician involved with this study decided to run a t-test. Why do you think she decided to do this, rather than using some other test, for these data? Explain briefly.

(c) (3 marks) What do you conclude from Figure 4, in the context of the data? Explain briefly. In your answer, you should give a P-value and justify why that P-value is appropriate.

(d) (2 marks) Describe a population to which it would make sense to generalize these results.

(e) (4 marks) The data as it originally came to me is shown in Figure 6. This is a SAS data set, and so `Obs` labels the rows; it is not a real column. Give SAS code that will create a new data set that looks like Figure 3. (It is OK to have some extra missing values in the new data set, and it is OK to have the observations in a different order as long as they are all there.) The data set shown in Figure 6 is called `negevalwide`; your new data set should be called `negevallong`.

Question 3 (10 marks)

Some people think that a mild amount of stress actually improves performance, but greater amounts of stress can be disruptive. In a study, 28 subjects were each given a task in which they had to keep a pointer on a moving disk. A clock records the time (in seconds) that the pointer is in contact with the disk. A larger value is better. Each subject is randomly assigned to one of four treatments, that relate to who was watching while they were trying to keep the pointer on the disk:

- `none`: no audience
- `experimenter`: experimenter as audience
- `peers`: peers (fellow students) as audience
- `faculty`: senior faculty members as audience

The data are shown in Figure 7. (This is the output from `proc print`.) The data set is called `stress`, which you may assume is the most recently created one.

(a) (3 marks) A boxplot is shown in Figure 8. Give the SAS code that was used to make the boxplot. (It is acceptable if your boxes will be in a different order from mine.)

(b) (2 marks) Based on what you see in Figure 8, what would be an appropriate test to compare the times for the four groups, testing the null hypothesis that they all have the same mean or median? Explain briefly.

(c) (3 marks) Give SAS code to run your preferred test from the previous part. You do not need to give code for any followup tests you might run if your preferred test gives a significant result.

(d) (2 marks) At the beginning of the question there was a suggestion that moderate stress levels might be associated with best performance. Is that supported by the data, from the output you have? Explain briefly. (If you cannot tell, describe what you would need to see and why.)

Question 4 (23 marks)

A university professor in California cycles to work. He has three routes that he uses, labelled by the name of the street that most of the route is on. He randomly chooses a route each day, and records the time in seconds that it takes to get from his house to the bike parking at his university. The data, as the professor recorded it, is shown in Figure 9. Note that he did not cycle each route the same number of times. The data has been read into a data frame called `biking`, as shown in the Figure. Give R, that is, Tidyverse code to accomplish the tasks below.

(a) (4 marks) Reorganize the data so that it has one row for each of the 52 completed rides, and columns called `street` and `time` that contain respectively the name of the street on which he cycled and the time in seconds that it took to complete the ride.

(b) (2 marks) Using the data in Figure 9, display the columns called `Oxnard` and `Rice` (and *not* `Rose`).

(c) (3 marks) Using the data in Figure 9, display the columns whose names begin with R, *without naming any columns or referring to them by number.*

(d) (3 marks) Some of the biking data as you rearranged it in part (a) is shown in Figure 10, as a data frame called `biking_long`. Use this data frame for the rest of the question.

Display the number of times each route was ridden, together with the mean time for each route.

(e) (2 marks) Display the rows numbered 2 through 6 of `biking_long`.

(f) (3 marks) Display the rows of `biking_long` where the street is Rice.

(g) (3 marks) Display the times of the rides in `biking_long` which took 600 seconds or less, along with the street that each one was on.

(h) (3 marks) Display the times of the five slowest rides, along with the streets they were on.

Question 5 (6 marks)

A study was carried out to assess the effects of smoking on exercise. Twenty-seven people were classified into three groups by smoking history as non-smokers (`non`), moderate smokers (`mod`) and heavy smokers (`heavy`). Each person was randomly assigned to one of three types of exercise: a stationary bicycle (`bike`), a treadmill (`tread`), or stair-climbing (`step`). There were three people in each combination of smoking history and exercise type. Each person was asked to begin exercising, and their time, in minutes, until "maximal oxygen uptake" was measured. The longer this time is, the fitter the person is (it means that they can exercise at a steady rate for a longer time).

The data as recorded are shown in Figure 11. The column `id` labels subject *within each group*.

(a) (4 marks) Give R code to arrange these data into a column of smoking-history group called `smoke`, a column of exercise types called `exercise`, and the time to maximal oxygen uptake, called `oxygen`. Bear in mind that there are 27 people in the data set, so your code will need to produce a data frame that has 27 rows. Your answer can contain a column `id` or not. Either is good.

(b) (2 marks) Part of the data set, after you have finished tidying it, is shown in Figure 12. Describe *in words* a suitable graph for this data set, and justify your choice briefly. Think about whether there is any value in including `id` in your graph.

Question 6 (12 marks)

Earlier, we worked with the data in Figure 2. This concerned an analysis of electricity pricing, and the satisfaction of customers with different pricing plans. In each of the parts below, give SAS code to accomplish the task described. This data is stored in the SAS data set `tod`.

(a) (3 marks) Create a new SAS data set from `tod` that contains only the column `satisfaction`.

(b) (3 marks) Create a new SAS data set that contains the first 8 rows (inclusive) of the data set `tod`.

(c) (3 marks) Create a new SAS data set that contains only those observations where the ratio is `4-1`.

(d) (3 marks) Obtain *only* the mean `satisfaction` for only those observations whose `ratio` is `8-1`. That is, your code must not calculate any other summary statistics, and must not calculate any means for other ratios. Do this without creating any new datasets.

Question 7 (15 marks)

Dorothy sells life insurance. She does this by visiting her clients' homes. We want to find out whether it is true that the more home visits she makes, the more people buy insurance from her. She collects data on the number of home visits she makes each week, and the number of life insurance policies she sells. The data are shown in Figure 13.

(a) (3 marks) A scatterplot is shown in Figure 14. Describe any relationship you see. Hint: linear or curved? Up or down? Strong, moderate or weak? Briefly justify your choice about the strength of the relationship.

(b) (2 marks) A regression was fitted to predict sales from visits, with the results shown in Figure 15. Which numbers from the output support your conclusions in (a) about (i) the direction (up/down) of the trend, (ii) the strength of the trend? Explain briefly (one number and a brief explanation for each of (i) and (ii)).

(c) (2 marks) What do you conclude from the P-value on the Intercept line in Figure 15? Does this make sense in the context of the data? Explain briefly.

Question 7 continues... This page: 7 marks.

(d) (2 marks) Does Figure 15 support a hypothesis that there really is a relationship between the number of visits in a week and the number of sales in that week? Explain briefly.

(e) (1 mark) Using Figure 15, what would you predict Dorothy's sales to be in a week where she makes 12 visits? (Use your calculator if you need to.)

(f) (2 marks) How accurate would you expect your prediction of the previous part to be: highly accurate, moderately accurate, not accurate at all, or something else? Cite something from the output to support your answer, and explain briefly.

(g) (3 marks) Two more plots, ones that normally go with a regression, are shown in Figures 16 and 17. What are these plots, and what do you conclude from them? Explain briefly. (You need to say three things: a conclusion from each plot (including a statement of what it is), and an overall conclusion.)

Question 8 (16 marks)

The life expectancy is measured as the number of years a baby born today can expect to live. This typically depends on the country a baby is born in, or on variables that say something about that society.

The data in Figure 18 show, for a number of countries, the life expectancy in years (male and female averaged), the number of televisions per person, and the number of doctors per 1000 people.

(a) (3 marks) A regression model is shown in Figure 19. Would you have expected each of the three numbers shown in the Estimate column to be positive, given what the data represent? Explain briefly, for each one.

(b) (4 marks) The plots of residuals against the explanatory variables are shown in Figure 20. Give the R code that was used to produce this plot, using the data frame shown in Figure 18, which is called `life`, and the regression model object `life.1`.

(c) (3 marks) What do you conclude from Figure 20, and thus how would you proceed with model-building? You may assume that the normal quantile plot of residuals and the plot of residuals against fitted values are satisfactory.

(d) (2 marks) I fitted another model (not shown) that was supposed to improve things. The residuals against fitted values and the normal quantile plot of the residuals are both again satisfactory. The residuals against the explanatory variables are shown in Figure 21. Do you think, on the basis of this Figure, that the model I fitted is now satisfactory? Explain briefly.

(e) (4 marks) An alternative model is fitted, with output shown in Figure 22. This may or may not be a model that you would recommend, based on your earlier answers. Look at the two additional estimates, compared to the regression output shown in Figure 19. What do their signs (positive or negative) tell you about the form of this relationship? In addition, does this form of relationship with these estimates make sense in the context of these data? Explain briefly. Hint: a quadratic $y = ax^2 + bx + c$ has a maximum or minimum at $x = -b/2a$.

Use this page if you need more space to write your answers. Be sure to label any answers here with the question and part that they belong to.

End of Exam This page: 0 marks.