University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAC33 (K. Butler), Midterm Exam
March 4, 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 8 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.

Your code should use `tidyverse` and `ggplot` ideas, as used in this course.

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)

Drivers in the Vancouver area suspect that the price of gas depends on the community they are in. One driver collected data on the price of a litre of regular gas at three randomly-chosen Chevron, Esso and Shell stations in each of three communities. The data are shown in Figure 2 in the booklet of Figures, as laid out in the file, and the file is stored as `vancouver_gas.txt` in your current folder in R Studio.

(a) (3 marks) Give R code to read in the file to a data frame `gas` and to display (some of) that data frame.

(b) (2 marks) We want to make a plot showing how gas prices compare among the different communities, ignoring which station each price comes from. Explain briefly why a boxplot is suitable for this.

(c) (2 marks) Give R code to make a boxplot as described in the previous part.

(d) (2 marks) The boxplot drawn by your code is shown in Figure 3. Is there a community where gas is noticeably cheaper or more expensive on average than the other communities? Explain briefly.

(e) (2 marks) Why is it that the Langley boxplot doesn't have a box? Explain briefly.

(f) (3 marks) Give R code to compute the mean and standard deviation of gas prices, along with the number of stations, for each community.

Question 2 (9 marks)

Is it really true that adult males in North America have a mean weight of more than 160 pounds? To assess this, a random sample of 16 males was taken, with the weights shown in Figure 4. The data frame is called `weights`, and the column of weights in it is called `weight`.

(a) (2 marks) A normal quantile plot is shown in Figure 5. What do you conclude from this? Explain briefly.

(b) (3 marks) Give R code to run a suitable t-test.

(c) (2 marks) The output from your t-test is shown in Figure 6. What do you conclude, in the context of the data?

(d) (2 marks) Figure 7 shows a summary of the data frame. Someone says to you "but the mean is greater than 160: you should definitely reject the null hypothesis." How do you respond to them? Explain briefly.

Question 3 (8 marks)

In each of the scenarios below, state whether we have one sample or two independent samples (or some other kind of sampling), and whether we should use a one-sided or two-sided procedure. In each case, justify your choices briefly. (One point for an answer about samples or sidedness *with a good reason*. No credit if you have no reason.)

(a) (2 marks) Many students have complained that the soft-drink vending machine in the student recreation room dispenses a smaller amount of drink than the vending machine in the faculty lounge. A student randomly samples servings from each machine and records the size of each serving in millilitres.

(b) (2 marks) An automotive company wants to compare the wearing quality of two brands A and B of tire. To do this, six test cars are used and one tire of each brand is placed on one randomly chosen wheel. (A standard brand of tires is placed on the other wheels.) After the test, each tire of brands A and B is assessed for wear (in thousandths of an inch).

(c) (2 marks) The Robertson square-drive screw has several advantages over a slotted or Phillips-head screw. A catalog reported that Robertson #8 wood screws fail only after an average of 48 inch-pounds of torque is applied (a much larger torque than for other types of screw). An independent testing laboratory randomly samples 22 Robertson #8 wood screws and records the inch-pounds of torque at which each of them fails, to see whether the catalog is correct or whether the average torque is less than 48 inch-pounds.

(d) (2 marks) The pulse rates of 13 randomly-chosen women were recorded, and a 95% confidence interval for the mean pulse rate of all women was calculated using `t.test`.

Question 4 (7 marks)

Encyclopedia Britannica defines latent heat as "energy absorbed or released by a substance during a change in its physical state (phase) that occurs without changing its temperature". For example, melting ice turns it from a solid from a liquid, and requires more heat than would simply heating ice without melting it.

In an experiment, two different methods were used to study the latent heat of ice fusion (melting). Water was cooled to -0.72 degrees Celsius (so that it froze). The water specimens were then heated back up to 0 degrees Celsius, and the heat required to do so was measured by one of two methods, an electrical method or a method of mixtures. The method used for a particular specimen was chosen at random. The required heat is called `heat_change` in the data. The data, in data frame `fusion`, is shown in Figure 8.

(a) (4 marks) A boxplot of the data is shown in Figure 9. Three different analyses are shown in Figures 10, 11, and 12. Based on this information, carry out what you think is the most reasonable analysis, justifying your decision, and obtain a conclusion in the context of the data.

(b) (3 marks) The people who gave you the data say to you "the boxplot is nice, but what we would really like to see is a suitable normal quantile plot". Give R code to produce a normal quantile plot that is suitable for this analysis.

Question 5 (12 marks)

The exponential distribution has density function $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ (0 otherwise). For example, the exponential density function for $\lambda = 0.5$ is shown in Figure 13. It has the same shape for any value of λ. The distribution has mean $1/\lambda$ and median $\ln(2)/\lambda$, with $\ln(2) = 0.6931$ approximately. The R function `rexp` will draw a random sample from an exponential distribution; it has two inputs, the sample size and the value of λ, in that order.

(a) (2 marks) If we believe that our data comes from an exponential distribution, explain briefly why we would prefer to do a sign test for the median rather than a t-test for the mean.

(b) (2 marks) How might you have guessed at the shape of the exponential distribution (in general) even if I had not shown you Figure 13? Explain briefly.

(c) (2 marks) What value of λ would produce an exponential distribution with a median of 10? Show your calculation.

(d) (4 marks) The function `pval_sign0` in `smmr` takes as input a null median and a vector, and returns the two-sided P-value for the sign test of that null median. Figure 14 shows how it works. Give code that uses `pval_sign0` to estimate the power of the sign test to reject a median of 5 when the median is actually 10, against a two-sided alternative, for data from an exponential distribution and a sample size of 50. You may assume that `smmr` has been loaded with `library(smmr)`. For full marks, do this without a loop.

(e) (2 marks) I got an answer of about 0.8 for my estimated power. Describe, to someone who doesn't know about power, what your result means.

Question 6 (12 marks)

Are there more arrests made for violations of the narcotics drug laws in larger cities than in smaller communities? A study was made of 24 communities that were classified by size: "large cities" (greater than 250,000 people), "small cities" (under 250,000 people), "suburbs", "rural". For each community, the rates of arrest (for these violations) were recorded per 10,000 inhabitants. The data are shown in Figure 15, in data frame `narc`.

(a) (2 marks) What makes analysis of variance an appropriate method to consider for these data? Explain briefly.

(b) (2 marks) On studying the boxplot in Figure 16, the statistician involved with the study decided that the arrest rate values were sufficiently close to normally distributed, given the small sample sizes (six observations per group). The statistician therefore proceeded with the analysis in Figure 17. What should the statistician conclude from the analysis, in the context of the data?

(c) (3 marks) Is it useful to study Figure 18? Explain briefly why or why not. If it is useful, summarize as concisely as possible what you conclude, in the context of the data. For this, you might like to think about how the city sizes rank in terms of arrest rates.

(d) (2 marks) An alternative analysis is shown in Figure 19. Under what kind of circumstances might such an analysis be appropriate? Explain briefly.

(e) (3 marks) There are two important differences between the results of the analysis of Figures 17–18, and those of the analysis of Figure 19. What are they? What feature of the data do you think might have caused this to happen? Explain briefly.

Use the area below 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.