University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAC32 (K. Butler), Midterm Exam
October 21, 2022

Aids allowed (on paper, no computers):

- My lecture overheads (slides)
- Any notes that you have taken in this course
- Your marked assignments
- My assignment solutions
- Non-programmable, non-communicating calculator

This exam has 8 numbered pages of questions plus this cover page.
In addition, you have an additional booklet of Figures to refer to during the exam.
The maximum marks available for each part of each question are shown next to the question part.
If you need more space, use the last page of the exam. Anything written on the back of the page will not be graded.
You may assume throughout this exam that the code shown in Figure 1 of the booklet of Figures has already been run.

1. The questions here are all about reading in data from files.

   (a) [3] A data file is laid out as shown in Figure 2 (in the booklet of Figures). This file is stored in a file `biscuit1.txt` in the same folder as your R Studio project. What code will read these data values into a dataframe called `test1` and display that dataframe?

   (b) [3] A second data file is shown in Figure 3. This file is stored in `biscuit2.txt` in the same folder as your R Studio project. Explain briefly why the code below will read in the data from the file. (You may treat it as given that the data read in will be stored in a dataframe called `test2`.)

```
test2 <- read_delim("biscuit2.txt", " ")
```

   (c) [2] Explain briefly why the code shown in Figure 4 also works.

   (d) [4] A collaborator gives you a spreadsheet, stored in `precious.xlsx`, as shown in Figure 5, containing data that you would like to read into R. The spreadsheet is a "workbook" containing two sheets, `Sheet1` containing some notes, and `Sheet2` containing the data you want. You have saved the file in the same folder as your current R Studio project.

   What are *two* distinct ways you might read the data in? Describe what you would do, or give code for doing it, as appropriate.

2. A chicken farmer keeps track of the weights of eggs (in grams) produced by their chickens for each of 24 days (labelled days A through X in the data). On each day, 10 eggs were randomly selected and weighed. Some of the data are shown in Figure 6.

   (a) [2] A graph is shown in Figure 7. What code was used to draw this graph?

   (b) [4] Considering Figure 7, what code would compute a suitable measure of centre and spread of egg weight for each day? Justify your choices briefly.

   (c) [2] The column `day_number` contains the same information as `day`, except that the days are now numbered rather than indicated by letters. Would the plot of Figure 7 still work using `day_number`, instead of `day`? Explain briefly, and, if necessary, indicate how you would change your code of (a) to make the plot work using `day_number`.

3. Many cities require certain pet animals, such as cats and dogs, to be licensed. Seattle is one of those cities. The dataframe `seattlepets`, some of which is shown in Figure 8, contains information about all the pets licensed in Seattle over a certain time period, including the licence number, the date it was issued, what kind of animal it is (in `species`), what breed it is within `species` (`primary_breed`), the animal's name, and so on.

   In the questions below, what code would display the appropriate part of the dataframe, or summarize it in an appropriate way?

   (a) [3] Display only the animal names and kind of animal they are, and display the first ten rows of those.

   (b) [2] How many animals are there of each species?

   (c) [2] Display only the pet pigs (you can display all the variables).

   (d) [3] For dogs (only), find the three most common names.

(e) [4] For cats (only), find the percentage of all cats that have a name that is included in the ten most popular cat names.

(f) [2] It turns out that 6.6% of all the cats in Seattle have a name that is one of the most popular 10 cat names. (This is the answer your code of the previous part would have given if run on the actual data). Do you think that there are many different names people give their cats in Seattle, or only a few different names? Explain briefly. (There is no code required in this part.)

4. A researcher was interested in comparing ages of hockey defencemen and forwards, and looked at data from the 2019 Ottawa Senators hockey team. Some of the dataset is shown in Figure 9. The column `Forward` takes the value `yes` if the player in question is a forward, and `no` if the player is a defenceman. (This dataset does not include goaltenders.)

(a) [2] The researcher decided to use a t-test to compare the ages of forwards and defencemen. Looking at the information in Figure 10, how would you argue *in favour of* that decision?

(b) [3] A t-test is run in Figure 11. What code was used to run this t-test?

(c) [2] What do you conclude from the t-test in Figure 11, in the context of the data?

(d) [2] There are two kinds of t-test that might have been used here. Out of these two tests, was the one run here the better choice? Explain briefly.

5. Figure 12 shows a power analysis. Use this Figure to answer the questions below, except where reference to another Figure is made.

   Explanations are not required, except where explicitly asked for, but may be worth partial credit if given and your answer is not correct. (An incorrect answer with no explanation is zero.) This is a long question, but answers should come quickly if you know what you are doing.

   (a) [2] What is the assumed true population mean?

   (b) [2] What is the assumed population distribution? How do you know? (Your answer should contain the name of a probability distribution.)

   (c) [2] What is the null hypothesis being tested?

   (d) [2] What is the alternative hypothesis being tested? Explain briefly.

(e) [2] What is the sample size in Figure 12?

(f) [2] Why do two of the code lines in Figure 12 have `list` in them? Explain briefly.

(g) [2] How do you know that this simulation is estimating power, and not the probability of making a Type I error? Explain briefly.

(h) [2] What is your estimate of the power of the test?

(i) [2] A second power analysis is shown in Figure 13. How does this differ from the first power analysis, Figure 12, in the context of the data?

(j) [2] Suppose that you are aiming for a power of 0.80. What do Figures 12 and 13, taken together, tell you about the sample size you should use? Explain briefly.

6. Thirty-one determinations of nickel content, in parts per million, were made from a Canadian syenite rock. The data, in column `nickel` of dataframe `abbey`, are shown in Figure 14. Our aim is to estimate the "average" nickel content of rock of this type, where "average" could be mean or median.

   (a) [2] What feature of Figure 15 would lead you to have doubts about using a t procedure (test or confidence interval) here? Explain briefly.

   (b) [1] What other feature of the data should also be considered in assessing the use of a t procedure? (No explanation needed.)

   (c) [4] What is the graph in Figure 16 a graph of? How does this graph *add* to your previous conclusion? What, therefore, is your overall conclusion about the validity of a t-procedure for these data? Explain briefly.

   (d) [2] Two confidence intervals are shown in Figures 17 and 18. Which one do you prefer, and why? Your explanation should make it clear why one of them is bad and the other one is good.

(e) [3] A hypothesis test is shown in Figure 19. What is the null hypothesis for this test? Suppose you are trying to detect any change from the null hypothesis. What do you conclude, in the context of the data?

(f) [2] How might you have guessed from Figure 19, even without seeing the P-value, that the P-value would be small? Explain briefly.

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