University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAC33 (K. Butler), Midterm Exam
March 3, 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. A study was made of three species of crab. In particular, interest was in the `Force` exerted by the crab's claw as it closed, and how that depended on the `Height` of the next joint of the claw. Also noted in the dataset was the `Species` of crab (text, possibly including a space). Some of the data file is shown in Figure 2. `Force` is measured in Newtons, and `Height` is measured in millimetres.

   (a) [4] What R code would read the data from the data file, called `crab-claws.txt`, into a dataframe called `crabs` and display the dataframe? The file is in the same folder that you are running R Studio in. (Here and elsewhere, if I ask for the code, I want the code and *not* the output. If I want the output, I will ask for it specifically.)

   (b) [3] What code would make a suitable graph of the two variables `Force` and `Height`? (It is a good idea to name the graph you are trying to draw in case you go wrong.)

   (c) [2] Describe in words how you would add `Species` to your graph of the previous part, and describe how you would *change* your code of the previous part to make this happen. (If you prefer, you can write out the complete code to draw this graph.)

   (d) [3] What code would show the number of crabs and the mean and standard deviation of `Force`, for each species separately?

   (e) [2] A graph is shown in Figure 3. This may or may not resemble the graph you gave code for earlier. From this graph, which species has the smallest standard deviation of `Force`? Explain briefly.

2. In question 1, we looked at some data on crab claws. The data, in dataframe `crabs`, are shown in Figure 4. There are three variables, `Species` (categorical), `Height` (quantitative), and `Force` (also quantitative). In this question, you are asked to give R (Tidyverse) code to accomplish the task described.

   (a) [2] Display all the variables *except* `Species`, without naming any of the other variables.

   (b) [2] Display all the categorical (text) columns, without using or even having to know their names.

   (c) [2] Display all the crabs for which `Force` is greater than 20.

   (d) [3] Display only the `Species` of the crabs that have `Force` less than 5.

   (e) [3] Find how many crabs of each species have `Force` less than 5.

3. A well-known cereal manufacturer has been adjusting the recipe for its frosted flakes so that a serving of the cereal should contain less sugar. One hundred servings of the latest batch of frosted flakes have been obtained, and some of the data, in grams of sugar per serving, is shown in Figure 5. The dataframe is called `cereal_sugar`.

   (a) [3] A graph is shown in Figure 6. What code was used to draw the graph?

   (b) [3] Last year's recipe for frosted flakes contained a mean of 38.5 grams of sugar per serving. What code would carry out a suitable hypothesis test, bearing in mind what the cereal manufacturer would like to have evidence for?

   (c) [3] The test you gave code for in the previous part has P-value 0.0002. Bearing in mind what the cereal manufacturer would like to know, what do you conclude from the test and why, in the context of the data?

   (d) [4] The cereal manufacturer's statistician is not happy with Figure 6, so decides to obtain a bootstrap sampling distribution of the sample mean. Some code is shown at the top of Figure 7. Unfortunately, the statistician spilled his coffee on the printout of his code, and only the four lines of code shown in Figure 7 could be read. What code was missing, and which line of the code shown should it come after?

(e) [2] Why did you need a `list` around something in your code in the previous part? Explain briefly.

(f) [3] What do you conclude from the graph at the bottom of Figure 7? Explain briefly. (You may assume that the graph came from the code at the top of the Figure, but including the missing code that you supplied earlier.)

4. A company believes that it has a "cleaner" manufacturing process than other companies making the same product. To assess this belief, the company measures the carbon monoxide emissions from smoke stacks at its factory, and from smoke stacks at a competitor's factory. The data are shown in Figure 8. The dataframe is called `Monoxide`. A lower carbon monoxide emission indicates a cleaner manufacturing process.

(a) [3] What code would make an appropriate graph of this dataset?

(b) [3] A test is shown in Figure 9. What do you conclude from the results shown in the Figure, and why, in the context of the data?

(c) [2] Why is there no `mu` in the code in Figure 9?

(d) [3] The company wants to obtain a 90% confidence interval for the difference in mean emission between the two groups. What code would get this interval?

(e) [2] A graph is shown in Figure 10. This may or may not be the same as the graph you gave code for in an earlier part. Does this graph give you any doubts about the appropriateness of your t-test? Explain briefly.

5. The beta distribution is a continuous distribution on $[0, 1]$. It has two parameters, called `a` and `b`. The R function `rbeta` obtains random samples from a beta distribution. It has three inputs. In order, these are the sample size, `a`, and `b`.

   (a) [2] What R code would draw a random sample from a beta distribution with `a` of 3, `b` of 6, of size 20?

   (b) [4] Suppose you are testing the null hypothesis that the population mean is 0.5, against the alternative hypothesis that it is less than 0.5, and the data is assumed to come from a beta distribution with `a` of 4 and `b` of 5. What code would enable you to estimate the power of a t-test under these assumptions, with a sample size of 30? (Note: the mean of this beta population is $4/(4 + 5) \simeq 0.44$.)

   (c) [2] The output from your code in the previous part is shown in Figure 11. Your boss asks you what sample size you would need to obtain power 0.75. What **change** would you make to your code of the previous part in an attempt to answer this question? Explain briefly.

6. Heights of 100 randomly-sampled adult males were measured, in whole numbers of inches. Some of the dataset is shown in Figure 12, and a graph is shown in Figure 13. Our aim is to test whether the population median height is 71.5 inches, or whether there is evidence that it is different from this.

   (a) [2] A sign test is run with the output shown in Figure 14. What code was used to obtain these results?

(b) [2] Why might you guess that the P-value in Figure 14 will be small, even before you look at the P-value? (You can base your answer on any of the output, *except* for the P-values.)

(c) [3] What do you conclude from the sign test, in the context of the data? Explain briefly.

(d) [3] Part of a binomial table for $n = 100, p = 0.5$ is shown in Figure 15. The column `x` denotes the number of successes, and the column `prob` is the probability of obtaining exactly that many successes. The probabilities have been rounded to six decimals. (If only five decimals are shown, the last one is zero.) You may assume that the probability of 20 successes or less is 0 to the accuracy shown. Show how you can use this table to obtain the P-value for your sign test. (If you do not have a calculator, show the calculation you would do.)

(e) [2] Based on what you have seen in this question, and assuming that the mean height is also of interest, do you think it would have been better to do a t-test here? Justify your answer 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.