University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAC32 (K. Butler), Midterm Exam
November 4, 2024

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 10 numbered pages of questions.

In addition, you have an additional booklet of Figures to refer to during the exam.

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

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.

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.

Reading data files

The following questions are about reading data files into R. Here, and elsewhere on this exam where I ask for code, you *do not* need to show the output that the code will produce; full credit can be obtained by giving correct code.

(1) (3 points) Read Scenario A in Figure 2. What R code would read these data into an R dataframe, and display at least some of that dataframe?

(2) (2 points) Read Scenario B in Figure 4. What R code would read these data into an R dataframe, and display at least some of that dataframe?

(3) (3 points) Read Scenario C in Figure 6. Describe the process you would use to get the data into a dataframe called `animals` in R Studio on `r.datatools.utoronto.ca`. Give enough detail to allow your reader to reproduce your process and get the same results you did. (You may assume that R Studio on `r.datatools` is currently open in a tab in your web browser.) Your answer will probably need to contain words *and* code.

Diabetes

The dataframe shown in Figure 7 contains five numerical measurements made on 145 non-obese adult patients classified into three groups. The dataframe is called `diabetes`.

The three primary variables are glucose intolerance (`glucose`), insulin response to oral glucose (`insulin`) and insulin resistance (`sspg`). Two additional variables, the relative weight (`rw`) and fasting plasma glucose (`fpg`), are also included.

The column `group` contains the classifications of the subjects into three groups, obtained by current medical criteria: "normal", "chemical diabetic", and "overt diabetic".

(4) (2 points) What would be an appropriate graph of the insulin resistance and group variables (one graph showing both variables)? Explain briefly why your graph is an appropriate choice.

(5) (3 points) What code would draw the graph that you named in the previous part?

(6) (3 points) A graph is shown in Figure 8. What code was used to draw this graph?

(7) (2 points) Sketch a one-sample boxplot of the data shown in Figure 8.

(8) (3 points) Another graph is shown in Figure 9. What kind of graph is it? What are two distinct things that you learn from this graph?

Diabetes revisited

The following questions also use the dataframe `diabetes`, some rows of which are shown in Figure 7. This contains five numerical measurements made on 145 non-obese adult patients classified into three groups.

(9) (3 points) What code would find the number of observations, the median, and the interquartile range of values of `fpg` for each `group`?

(10) (2 points) What code would display (all the rows of) all the columns that have the letter "g" in their names somewhere, either uppercase or lowercase, without naming or numbering any columns?

(11) (2 points) What code did I use to make the display in Figure 7?

(12) (2 points) For only the patients whose `glucose` value is more than 1000, what code will display all their information?

(13) (3 points) What code will display the `sspg` values for patients in the `overt` group that have the largest 8 values of `insulin` (but not display these for patients in any other group, or display any other variables)?

Heights of males

Figure 10 shows the heights of a sample of 100 males, measured in inches. Use these data to answer the following questions.

(14) (2 points) A histogram of the heights is shown in Figure 11. Based on this histogram and anything else you have learned about these data so far, explain briefly why it reasonable to use t procedures (test or confidence interval) to make inferences about the population mean height of males.

(15) (4 points) In 2016, the mean height for American males was 69.1 inches. Using Figure 12, what do you conclude? (You may assume that the data in Figure 10 were collected this year.) Make sure you show your thought process clearly.

(16) (3 points) What do you conclude from the output of Figure 13? (The code that produced the output is shown above the output.) Your answer should make it clear that you know what the output is.

(17) (2 points) Appropriately state a 95% confidence interval for the mean height of all males this year.

Lower back pain

In many industrial settings involving physical labour, lower back pain (LBP) is a serious health problem. A study measured the lateral range of motion (measured in degrees) of workers in a steel factory, some of whom had a history of lower back pain and some of whom did not. The researchers suspected that the workers with no history of lower back pain would have a larger range of motion. Some of the data are shown in Figure 14. There were 28 workers in the no-LBP group and 31 in the LBP group.

(18) (3 points) A boxplot is shown in Figure 15. On the basis of the boxplot, the researchers decided to run a t-test to compare the mean lateral range of motion between workers with a history of low back pain and those without. What do you think the researchers' reasoning was?

(19) (3 points) What code would carry out your preferred t-test in this situation? Justify any choices you make.

(20) (2 points) The P-value for your test is 0.0019. What do you conclude, in the context of the data?

Mystery code

(21) (4 points) Some code and output is shown in Figure 16. What precisely do you conclude from the output?

Planning a comparative experiment

(22) (3 points) Two different heat treatments will be compared in their effects on the strength of steel ingots. (An ingot is defined as "a mass of metal cast into a convenient shape for storage or transportation.") Eight ingots will be cast using each treatment. In the units used, strength has a standard deviation of 0.5 units for each treatment. We want to know how likely it is that a null hypothesis of equal strengths will be rejected if one of the treatments has a mean strength 0.8 units greater than the other, using an appropriate two-sided t-test. What code will calculate this? (You may assume that the distributions of strengths are normal.)

(23) (2 points) What *changes* would you make to your code of the previous question to find the required sample size to obtain a power of 0.90?

(24) (2 points) The output from your changes in the previous question is shown in Figure 17. What does this output tell you about the total number of observations you need to make?

Wisconsin Card Sorting Test

The Wisconsin Card Sorting Test is widely used by psychiatrists with patients who have a brain injury or other mental illness. Fifty patients were given this test, with some of the data shown in Figure 18. The dataframe is called `wcst`. A higher `score` is better.

(25) (2 points) A histogram of the scores is shown in Figure 19. The distribution is skewed, but the sample size is fairly large. What would you calculate and make a graph of in order to determine whether the sample size is large enough to enable you to use a t-test? (No explanation needed.)

(26) (3 points) Based on what the researchers saw in the graph of what you named in the previous part, they decided to run a sign test. The median score in the general population is 25. What code would run a suitable test to see whether the median score in this population is less than 25?

(27) (2 points) The output from your code in the previous question contains more than one P-value. Which is the correct one to use here? Explain briefly.

(28) (3 points) A 99% confidence interval for the population median `score` is shown in Figure 20. What code was used to obtain this confidence interval?

(29) (2 points) Let M denote the population median score. Suppose we test $H_0 : M = 20$ against $H_a : M \neq 20$. What does the confidence interval in Figure 20 tell you about the P-value of this test? Explain briefly.

Comparing lecture sections

A large class has 164 students divided into three lecture sections denoted `a`, `b` and `c`. For each student, their mark `m1` on the midterm exam is recorded, along with their lecture section (`lecture`). Some of the data are shown in Figure 21. The course instructor is interested whether there is any difference in performance among the lecture sections.

(30) (2 points) What two key assumptions are required for the analysis in Figure 22?

(31) (2 points) What do you conclude from the analysis in Figure 22, assuming that it is an appropriate choice of analysis?

(32) (1 point) What is the analysis in Figure 23? Explain (very) briefly why it is appropriate to use it here.

(33) (3 points) What do you conclude from Figure 23? Your answer should consider which lecture sections have the highest and lowest means.

If you need any more space, use this page, labelling each answer with the question number it belongs to.