University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAC32 (K. Butler), Midterm Exam
October 24, 2017

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 xx 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.

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.

Downloading files

A computer science student wants to know whether it takes longer to download files at certain times of the day. To assess this, she placed a file on a remote server, and repeatedly downloaded it from there at different times of day: "early" (7:00am), "evening" (5:00pm) and "late-night" (midnight). The student downloaded the file 16 times at each of the three times of day, recording the download time in seconds, for a total of 48 observations. The data are shown in Figure 2 of the booklet of code and output.

(1) (3 points) Is the data file formatted as a CSV, space-delimited, or something else? Make a reasonable assumption about the format of the data, and give code to read the data file into R. The data is in the file `download.txt` in the current folder (the folder in which R is working). Use the name `downloads` for your input data frame.

(2) (2 points) Give R code to draw a boxplot of download time for each time of day. (You may assume that you successfully read the data into a data frame called `downloads`.)

(3) (2 points) Explain briefly why a boxplot is a sensible plot to draw for these data.

(4) (2 points) The boxplot drawn by your code is shown in Figure 3. Does there seem to be a good time of day to download the file? Explain briefly.

High jump

The high jump has been an Olympic event since 1896 for men and 1928 for women. The winning height (in metres) in the high jump at each Olympic Games for men and women was recorded. The data set is shown in Figure 4. We are interested in making a graph that shows how the winning height changes over time. The data has been read into a data frame `highjump`.

(5) (2 points) What would be a suitable graph to display all the variables *on one plot*?

(6) (3 points) Give R code to draw your graph of the previous question.

(7) (3 points) Give R code for one plot *command* that produces two appropriate subplots, one for men and one for women. (By "subplots" I mean graphs like ones you made above, but one containing only men and one containing only women, contained within one larger plot.)

The Stroop test

In psychology, the “Stroop test” is a test in which colour names are printed in a different colour. For example, the word “red” might be printed in blue ink. The idea of the test is that you name the colour *of the ink* as fast as you can (for a whole collection of colour names printed in different colour ink). There is a mental conflict involved: when you see “red” printed in blue ink, you have to stop yourself saying “red” and say “blue” instead.

A researcher thinks that hypnotizing people helps them to do better on the Stroop test, and so conducts an experiment on twelve subjects. Six of these subjects, randomly chosen, do the Stroop test as normal (without being hypnotized), while the other six are hypnotized first. The score on the test is a combination of the time taken and the number of mistakes made; a lower score is better. The data are shown in Figure 5}. I saved the data as `stroop.txt` in your current folder in R Studio.

(8) (4 points) Give code to read in and display the data.

(9) (2 points) What kind of plot would be most suitable to display these data? We do not want to display the subject numbers on the graph. Justify your choice briefly.

(10) (2 points) Give code to draw the plot that you proposed in the previous part.

(11) (3 points) Give code to carry out a suitable t-test to determine whether the hypnotized subjects do better at the Stroop test on average than the unhypnotized ones. The researchers decided that the hypnotized and unhypnotized subjects have about the same spread of scores. If you think your test should be one-sided, justify your choice of side.

(12) (3 points) The plot you drew earlier shows approximate normality and approximately equal spread for the test scores within each group, of people who were hypnotized and people who were not. The output for your t-test is shown in Figure 6. What do you conclude from it, in the context of the data?

Australian athletes again

In lecture, we looked at the Australian athletes data set. This contains information on 202 athletes who play various different sports. Thirteen variables are measured for each athlete. Some of the data set is shown in Figure 7.

(13) (2 points) The variable `BMI` is the "body mass index". It is the ratio of height to weight-squared, and is often used as a measure of whether a person is over-weight or under-weight for their height. Give R code to obtain a 90% confidence interval for the mean BMI of all athletes (of which these are a sample). Your code can obtain other things as well as the confidence interval, but must obtain the appropriate confidence interval among its output. (The data frame is called `athletes`, and has already been read into R.)

(14) (2 points) The 90% confidence interval found by your code in the previous part goes from 22.6 to 23.3. One of the three interpretations below is the best one. Which one? Explain briefly:

(i) the procedure by which this confidence interval was produced would give you an interval containing the population mean BMI for 90% of all possible samples; (ii) 90% of all athletes have BMI values between 22.6 and 23.3; (iii) the interval from 22.6 to 23.3 has probability 0.90 of containing the population mean BMI.

(15) (3 points) A histogram is shown in Figure 8. Based on this histogram, do you have any doubts about the confidence interval calculated above? Explain briefly. What, if anything, would you do instead?

Power

The next four questions are about power of hypothesis tests. The first two questions refer to the same situation (described in the first question below), while the third and fourth questions are independent of those and of each other.

(16) (3 points) A measurement varies according to a normal distribution with unknown mean μ but known SD 2. Suppose we are interested in testing the null hypothesis $H_0 : \mu = 8$ against the alternative $H_a : \mu > 8$. If in fact $\mu = 9$ and we have a sample size of 50, how likely are we to correctly reject H_0, at $\alpha = 0.05$? Give code to do the calculation.

(17) (2 points) How would you change your code of the previous part to find the sample size needed to achieve probability at least 0.7 of correctly rejecting the null hypothesis?

(18) (3 points) We are now comparing a treatment against a control. The standard deviation of both sets of measures is 10. Suppose we take a sample of 25 observations from each group and test the null hypothesis that the treatment and control have the sample mean, against an alternative that the treatment mean is greater. If in actual fact the treatment mean is 5 units higher than the control mean, what R code would *calculate* the power here?

(19) (3 points) Look at Figure 9. Explain briefly what the results at the bottom indicate. Do *not* explain what each line of the function does.

Diabetes

The typical age of onset of diabetes (that is, the age at which someone with diabetes is diagnosed as having diabetes) is 37 years. A sample of people with diabetes is taken from a certain population. The data are shown in Figure 10. We are interested in whether the typical age of onset for these people is different from that of other people with diabetes.

(20) (2 points) A histogram of the ages of onset is shown in Figure 11. Why do you think the statistician in this study chose to use a sign test rather than a *t*-test?

(21) (2 points) Write appropriate null and alternative hypotheses for the sign test.

(22) (2 points) A sign test was carried out, as shown in Figure 12. Obtain a P-value from the output, and state your conclusion in the context of the data.

(23) (3 points) Some other output for the same data is shown in Figure 13. Use this output to describe how `sign_test` obtained its P-value.

Zinc in drinking water

Trace metals in drinking water affect the flavour, and an unusually high concentration of them can pose a health hazard. Zinc concentration in water at the bottom of a lake and in water at the surface was measured at each of 10 different locations. We are interested in whether zinc concentration is higher on average at the bottom of the lake. If it is, drinking water should be drawn from near the surface of the lake. The data are shown in Figure 14.

(24) (2 points) Explain briefly why a matched-pairs analysis would be more suitable here than a two-sample analysis.

(25) (3 points) Figure 15 and Figure 16 show two possible analyses for these data. Which of the two analyses is more suitable? What, therefore, do you conclude in the context of the data?

(26) (3 points) What specific assumption are we making in order to trust the results of our *t*-test? Which one or ones of Figure 17, Figure 18, and Figure 19 enable us to assess this assumption? What do you conclude? Explain briefly.

Prenatal care

A new approach to prenatal care is proposed for pregnant women living in a rural community. The new program involves in-home visits during the course of pregnancy. A trial with 15 pregnant women is designed to evaluate whether women who participate in the new program deliver healthier babies than women receiving the usual care. The outcome is the APGAR score indicator measured 5 minutes after birth. APGAR scores range from 0 to 10 with scores of 7 or higher considered normal (healthy), 4–6 low and 0–3 critically low.

The data are shown in Figure 20. The column `care` shows whether each pregnant woman received the usual care (`usual`) or extra prenatal home visits (`visits`).

(27) (3 points) Figure 21 shows the results of a hypothesis test run on these data. The first part of the output from `median_test` shows a table. What do the numbers 5 and 1 on the second row of that table represent? Explain briefly.

(28) (2 points) In Figure 21, look again at the first part of the output from `median_test`, specifically the items labelled `$grand_median` and `$table`. Do these suggest that the new program is helpful? How can you tell? Explain briefly. (This part is not asking about P-values.)

(29) (3 points) What do you conclude from the result of the test at the bottom of Figure 21, in the context of the data? In writing your conclusion, bear in mind what the researchers running this study are trying to prove.

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