University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAC32 (K. Butler), Final Exam
December 11, 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 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.

Hawks

A "blind" is a place where people can watch wild animals or birds without being noticed by them. A college maintains a hawk blind at Lake MacBride, Iowa, where students can observe hawk and other birds. Birds of two species of hawk, Cooper's Hawk (`CH` in the data file) and Red-tailed Hawk (`RT` in the data file) were captured, weighed and then released; the weight was measured in grams. Some of the data file is shown in Figure 2 (the remaining lines are laid out in the same way). The data file is called `hawks.txt`, and is in the same folder in which you are currently running R Studio.

(1) (3 points) What code would read the data from the file into a dataframe called `hawks`, and display at least some of the dataframe that was read in?

(2) (3 points) What code will calculate and display the mean and standard deviation of weight, and the number of hawks observed, for each species of hawk?

(3) (2 points) What code will draw a suitable graph of the two variables in this dataframe?

(4) (2 points) What code will display the species and weight of all the hawks that weigh less than 300 grams?

(5) (3 points) What code will display only the weights of only the five heaviest Cooper's Hawks?

(6) (2 points) A dataframe `hawk_names` is shown in Figure 3. What code would use this dataframe, as well as dataframe `hawks`, to create a dataframe that contains the full names of the species of each hawk, along with each hawk's weight? The dataframe produced by your code may contain other columns, and you do not need to save it.

(7) (1 point) In the dataframe `hawk_names`, suppose the column called `species` had instead been called `abb` (short for "abbreviation"). How would your code from the previous question need to change in order to achieve the same result?

(8) (3 points) Some code and output is shown in Figure 4. What do you conclude from it, in the context of the data? Explain, showing your logic clearly.

Energy expenditure

Do obese people use more energy on average than lean people? To find out, the total energy expenditure (in megajoules) was recorded over 24 hours for 22 people. 13 of these people were lean and 9 were obese. The data are shown in Figure 5, in dataframe `energy`. The energy expenditure is shown in column `expend`, and the classification of each person as lean or obese is shown in column `stature`.

(9) (2 points) A graph is shown in Figure 6. On the basis of this graph, and anything else you have learned about the data so far, explain briefly why any kind of two-sample t-test would not be appropriate.

(10) (2 points) What code would run a suitable test to see whether obese people use more energy on average than lean people? ("Average" can mean either mean or median, as appropriate.)

(11) (2 points) The output from the code you gave in the previous part is shown in Figure 7. What is an appropriate P-value for testing the alternative hypothesis of interest to us? Explain briefly.

(12) (2 points) Hence, what do you conclude, in the context of the data?

Income and education

These data are annual incomes in 2005 of a random sample of 2,584 Americans who were selected for the National Longitudinal Survey of Youth in 1979 and who had paying jobs in 2005. The columns are:

- `Subject`: a subject ID
- `Educ`: the amount of education (see below)
- `Income2005`: annual income in 2005
- `tr_income`: transformed 2005 income (which we will use in our analysis).

The categories in `Educ` are:

- `012` (less than high school, 0–12 years)
- `12` (completed high school, 12 years)
- `13-15` (some college or university, 13–15 years)
- `16` (completed university, 16 years)
- `16+` (started or completed graduate school, more than 16 years).

The people who collected the data expected to see that more education went with a higher income on average (or a higher transformed income).

Some of the data, in dataframe `incomes`, is shown in Figure 8. Summary data of transformed income for each education category are shown in Figure 9.

(13) (3 points) Plots of the transformed incomes are shown in Figure 10. What code was used to make this Figure?

(14) (2 points) According to Figure 10, what do the distributions of transformed incomes appear to have in common (that is, what is the same about them)? Why, nonetheless, is it *not* necessary to run Mood's median test here to compare the transformed incomes, based on the information you have?

(15) (3 points) Two analyses are shown in Figure 11 and Figure 12. Which analysis is better? Explain briefly. Your answer should make it clear what each analysis is.

(16) (3 points) What conclusions do you draw from your chosen analysis, bearing in mind what the people who collected the data would like to know?

Tidying

The following questions are about tidying data.

(17) (3 points) Figure 13 shows a dataframe `d1` and, below it, some code run with this dataframe. What will be the *output* of this code?

(18) (4 points) Figure 14 shows a dataframe `d2` and Figure 15 shows a dataframe `d3`. What code will rearrange `d2` into `d3`? For full credit, do this in one step.

(19) (3 points) Dataframe `d4` is shown in Figure 16, and dataframe `d5` is shown in Figure 17. What code will rearrange `d4` into `d5`? Explain briefly how you know your code will work (hint: how do you know that the rows will be correct?).

(20) (3 points) Figure 18 shows a dataframe `d6` and some code that uses `d6`. What will the *output* from this code be?

Growing cattle

As cattle (cows and bulls) grow, the metabolic clearance rate of growth hormone changes. In a study, 14 male cattle were weighed, and their metabolic clearance rate was measured. The data are shown in Figure 19, with the metabolic clearance rates in column `mcr`. The dataframe is called `growth`.

(21) (3 points) A scatterplot of the data is shown in Figure 20. Describe any trend you see (hint: form, direction, strength).

(22) (2 points) Some analysis is shown in Figure 21 and Figure 22. Which part of this analysis tells you that that it might be a good idea to add a squared term in body weight to the regression, and how does it tell you this?

(23) (2 points) A regression that includes body weight squared is shown in Figure 23. What are *two* ways you can tell that adding the squared term was a good idea?

Tadpoles

A tadpole is a small creature that lives in water, and develops into a frog or toad. Can tadpoles adjust the length of their intestines if they are exposed to a fungus called `Bd`? The dataframe `tadpoles`, shown in Figure 24, contains these variables:

- `treatment`: `Bd` if the tadpole was exposed to the fungus, `Control` if not.
- `body`: body length in mm
- `gut_length`: length of the intestine in mm (response)
- `mouthpart_damage`: measure of damage to the mouth (a larger value indicates more damage)

The biological question is whether `gut_length` has an association with `treatment`.

(24) (2 points) A regression was fitted to predict gut length from the other variables, with code shown in Figure 25. `drop1` output from this regression is shown in Figure 26. What do you conclude? Explain briefly.

(25) (1 point) Why is it better to use the `drop1` output to answer the previous question than the `summary` output shown in Figure 27?

(26) (2 points) What would happen if I ran `step` on the model `tadpoles.1`? Explain briefly.

(27) (2 points) Interpret the number 6.442 in the Estimate column of Figure 27.

(28) (2 points) Interpret the number 25.412 in the Estimate column of Figure 27.

(29) (3 points) A plot is shown in Figure 28. What do you conclude from this plot? Your answer should make it clear that you know what this plot is.

(30) (3 points) Another plot is shown in Figure 29. What *code* was used to draw this plot?

(31) (1 point) Figure 30 shows another plot from the regression. Why is this one a different type of plot from the ones in Figure 29?

(32) (2 points) Does Figure 30 indicate any problem with the regression? Explain briefly.

Exponential growth

Exponential growth is defined by the formula

$$y = y_0(1 + r)^t$$

where y_0 is the initial value of some quantity, r is the growth rate per unit time (expressed as a proportion less than 1), t is the amount of time, and y is the final value of the quantity.

(33) (3 points) Write an R function called `expgrowth` that will accept an initial value, growth rate per unit time, and amount of time (in that order), and will return the final value. Use the same notation for things as in my formula above.

(34) (2 points) How would you use your function to find the final value after 60 seconds of a quantity that grows at 5% per second, starting from an initial value of 20?

(35) (2 points) Exponential growth only makes sense if the initial value is positive. How would you change your function to give an error if the input value is not positive?

(36) (3 points) What code would use your function, along with some variation on `map`, to create a dataframe that contains a column `time` with values from 0 through 10 (seconds) and the values of a quantity `y` at each of those times that has initial value 20 and grows by exponential growth at a rate of 5% per second?

(37) (4 points) Write a function called `plot_expgrowth`, using your previous work, that:

- takes as input an initial value, growth rate, and maximum time
- constructs a dataframe that contains times from 0 up to the input maximum time (in steps of 1 time unit) and values of the quantity at each of those times
- makes a plot of the quantity against time.

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