University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAC33 (K. Butler), Final Exam
April 18, 2022

Aids allowed: 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 is open book, as above.
This exam has **??** numbered pages of questions.
In addition, you have an additional booklet of Figures to refer to during the exam. Contact an invigilator if you do not have this.
The maximum marks available for each part of each question are shown next to the question part.
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.

Question 1 (**??** marks)

In a chemical procedure called differential pulse polarography, a chemist measured the maximum `Current` generated (in millionths of an ampere) in a solution containing a given amount of nickel (`Ni`), measured in parts per billion. The machine outputs the data in the format shown in Figure 2. This is in a file called `nickel.txt` in the same folder where you are currently running R Studio.

(a) (3 marks) What R code would read the data shown in Figure 2 into a dataframe called `nickel`?

(b) (3 marks) What would be an appropriate graph for these data? Explain briefly. What `ggplot` R code would draw this graph?

Question 2 (**??** marks)

Acquired immunodeficiency syndrome (AIDS) is a chronic, potentially life-threatening condition caused by the human immunodeficiency virus (HIV). By damaging your immune system, HIV interferes with your body's ability to fight infection and disease. There's no cure for HIV/AIDS, but medications can dramatically slow the progression of the disease. These drugs have reduced AIDS deaths in many developed nations.

Data was collected on 2843 patients diagnosed with AIDS in Australia before July 1, 1991. At that time, there were no good medications for AIDS.

The variables measured were, in order:

- the `state` of Australia that the patient lives in. Some of the states are combined.
- the `sex` of the patient
- the date of `diag`nosis with AIDS
- the date of `death` or of last observation (if the patient was still alive at last observation)
- the `status` of the patient at the last observation (A = alive, D = dead)
- `T.categ`: the transmission category (how the patient got AIDS in the first place)
- the `age` at diagnosis in years.

Some of the dataframe, called `aids`, is shown in Figure 3.

For each of the parts below, give (Tidyverse) code that would accomplish the task shown.

(a) (2 marks) Display the median and first and third quartiles of age at diagnosis for patients from each state.

(b) (2 marks) Display the columns of the dataframe whose names begin with "d", without explicitly naming any columns.

(c) (3 marks) Display the patients who are either from the state of Queensland (QLD) or are aged over 50 (or both).

(d) (3 marks) Find the earliest and latest last-observation dates for the patients who were still alive at last observation. I don't want to know about patients who died.

(e) (3 marks) Display only the age and how the patient first got AIDS for only the patients from the state of Victoria (VIC).

(f) (3 marks) Display the age and status of each of the 12 oldest patients who come from New South Wales (NSW).

Question 3 (**??** marks)

An environmentalist heard reports that a community was discharging untreated sewage into a river, and feared that this might have an impact on the ability of the river to support aquatic life. To investigate, the environmentalist chose at random five locations upstream from the community and five locations downstream from the community. At each one, they took a sample of the river water and measured the dissolved oxygen content, in parts per million. The higher the dissolved oxygen content is, the better the water is for supporting aquatic life, and dumping sewage into a river will decrease the dissolved oxygen content.

(Hint: a river flows downhill, so that water flows from upstream of a point to downstream of that point, and not the other way around.)

The data are shown in Figure 4, in the form that the environmentalist collected it. The data frame is called `river_water`.

(a) (3 marks) Describe briefly why the data as shown in Figure 4 are not suitable for an appropriate analysis, as we did it in lecture. (If you find it helpful, as part of your answer you can describe what an appropriate format would be.)

(b) (2 marks) What code would transform the data in Figure 4 into a suitable layout? Save your results into a dataframe `river2`.

(c) (4 marks) The environmentalist knows from previous experience that dissolved oxygen content values have very close to a normal distribution, though not necessarily with equal spreads at different locations. Some possible analyses are shown in Figure 5 through Figure 8. One of these analyses is the most appropriate, and the other three are less appropriate. For each Figure in turn, briefly discuss whether or not it is the most appropriate and why.

(d) (3 marks) From your best analysis (of the previous part), what conclusion do you draw, in the context of the data and bearing in mind what the environmentalist would like to know?

Question 4 (**??** marks)

This question uses the dataframes shown in Figure 9 through Figure 12. In each case, I show you some code that starts from one of these dataframes, and I ask you to write down what the output is. No explanation is necessary, but if you are wrong, an explanation might get you some partial credit.

(a) (3 marks) Suppose the following code is run:

```
d1 %>% pivot_longer(y:z, names_to = "name", values_to = "value")
```

What would be the output from this code?

(b) (3 marks) Suppose the following code is run:

```
d2 %>%
  pivot_longer(starts_with("y"), names_to = c(".value", "when"), names_sep = "_")
```

What would be the output from this code?

(c) (3 marks) Suppose the following code is run:

```
d3 %>% pivot_wider(names_from = x, values_from = y)
```

What would be the output from this code?

(d) (4 marks) Suppose the following code is run:

```
d4 %>% pivot_wider(names_from = y, values_from = z)
```

What would be the output from this code?

Question 5 (**??** marks)

A realtor in Taiwan collected some information about houses for sale in the city of Taipei, as follows:

- `sale_date`: the date on which the house was sold (in fractional years; for example 2013.5 means June of 2015)
- `age` of the house when it was sold (in years)
- `mrt`: distance from the house to the nearest MRT (rapid transit) station (in metres)
- `conv`: number of convenience stores within walking distance of the house (count)
- `price`: selling price per unit area, in suitable units (tens of thousands of Taiwan dollars per "ping", where a ping is 3.3 square metres)

The realtor's aim was to predict selling price from the other variables, so as to be able to set a good price for houses in Taipei that come on the market in the future. Some of the data is shown in Figure 13.

(a) (2 marks) A Box-Cox analysis is shown in Figure 14. Why do you think the realtor decided to predict the log of the selling price (per unit area), rather than some other function of the selling price?

(b) (2 marks) A regression analysis, predicting log price from the other variables, is shown in Figure 15. Which, if any, explanatory variables should the realtor remove from the regression? Explain briefly.

(c) (2 marks) Suppose the realtor decided to remove the explanatory variable `mrt` from the regression. What would happen to R-squared?

(d) (3 marks) Residual plots for the regression of log-price on the other explanatory variables are shown in Figures 16 through 18. Do you see any problems with these residual plots? If so, describe the problems you see; if not, explain how you know that the plots are satisfactory.

Question 6 (**??** marks)

In this question, we will be writing and using functions.

(a) (3 marks) Write an R function called `f1` that has one input called `n` which returns the sum of the first `n` integers. For example, `f1(3)` should evaluate to 6 because $1 + 2 + 3 = 6$. For maximum points, your function should be as concise and clear as possible. You may assume that the input to your function is a positive integer.

(b) (3 marks) How would you rewrite your function `f1` to calculate the sum of the integers between a second input `lo` and the first input `n`, inclusive? `lo` should be an optional input that is given the value 1 if it is not specified by the user. Your new function should be called `f2`.

(c) (2 marks) How would you best use your function `f2` to (i) sum the integers between 3 and 6, (ii) sum the integers between 1 and 10? ("Between" means "inclusive" here.)

(d) (3 marks) Suppose you were to run your function `f2` with inputs `lo = 3` and `n = 2`. What do you think should happen in this case? Rewrite your function `f2` to handle cases like this.

(e) (3 marks) For the remainder of the question, we go back to using the function `f1` that you defined earlier.

You are given three values of `n` in a vector, like this:

```
ns <- c(4, 6, 10)
```

How would you best run `f1` on all three of these values of `n` at once, arranging the answers back into a vector? Give the code that would do it.

(f) (4 marks) Suppose the values of `n` we want to run `f1` on are now in a column of a dataframe like this:

```
d <- tibble(n = ns)
d
```

```
## # A tibble: 3 x 1
##       n
##   <dbl>
## 1     4
## 2     6
## 3    10
```

What are two *different* ways in which we might create a column `first` that contains the results of running `f1` on those three values of `n`?

Question 7 (**??** marks)

You observe a process that produces values Y from an exponential distribution with rate β. That is to say, the density function of Y is $f(y) = \beta \exp(-\beta y)$ for $y \geq 0$ and 0 otherwise, where $\beta > 0$. The mean of this distribution is $1/\beta$. You want to estimate β using Bayesian methods.

You believe before looking at any data that β is most likely between 0.1 and 0.5, a fact that is summarized by a gamma prior distribution with shape 4.6 and rate 17.2. The prior density is shown in Figure 19. (95% of this distribution is between 0.1 and 0.5.)

You observe the data shown as `my_y` in Figure 20. The maximum likelihood estimate of β is $\hat{\beta} = 1/\bar{y}$, as shown in Figure 21.

In Stan, the exponential distribution is written `exponential()`, and the gamma distribution as `gamma()`. Inside the brackets go the name or value of the one parameter for the exponential distribution and the names or values for the two parameters for the gamma distribution.

(a) (3 marks) Write the `model` section of a Stan program that will sample from the posterior distribution of β.

(b) (2 marks) Write the `parameters` section of your Stan program.

(c) (2 marks) Write the `data` section of your Stan program.

(d) (3 marks) Assuming that your Stan program compiles correctly into an object called `expo`, what R code would run it for the data in Figure 20, to obtain a sampled posterior distribution? You may assume that you already have the data stored as in Figure 20. Save the posterior distribution into `expo_fit`.

(e) (2 marks) A summary of the posterior distribution of `beta` is shown in Figure 22. Why does most of the posterior distribution appear to be less than the maximum likelihood estimate shown in Figure 21?

(f) (3 marks) Suppose you wanted to modify your Stan program to allow estimation of an exponential rate parameter from a sample of any size, and to use a gamma distribution with any parameters as the prior for `beta`. (Call those parameters `theta` and `lambda`.) What *changes* would you make to the Stan code that you wrote earlier?