SQMB - Summative Assessment 1

YOUR EXAM NUMBER

2023-03-03

1 Instructions

PLEASE READ CAREFULLY

DUE THURSDAY, 30 MARCH 2023 AT NOON

You must include your **exam number as the author** in the document preamble above.

You'll notice that line 10 of the preamble says `mainfont: DejaVu Sans`. We would appreciate it if you'd use this font (or at least some other sans-serif font). **Try to knit this Rmd file immediately, before making any changes, to see if you have this font installed.**

- If you get an error message saying that DejaVu Sans could not be found, you can download it here: http://sourceforge.net/projects/dejavu/files/dejavu/2.37/dejavu-fonts-ttf-2.37.zip.
- Then, install it appropriately for your operating system.
  - Here's a guide for Windows: https://www.digitaltrends.com/computing/how-to-install-fonts-in-windows-10/.
  - Here's a guide for Mac: https://support.apple.com/en-gb/HT201749.

Do each exercise by completing tasks, answering questions and/or providing code if required. Please **keep your written answers as concise as possible.**

Feel free to **add as many code chunks as you want** throughout.

When you are ready to submit:

1. Render the Rmd file to **PDF**.
2. **Rename** the PDF to your exam number only.
3. **Upload** the PDF file to Learn.

2 Exercises

2.1 Exercise 1: Creating plots

The next three exercises below require you to read in a particular file, filter/transform the data as needed, and create one or more plots that appropriately illustrate the described aspects of the data. Please also include a concise written description of the patterns you notice in each plot you make.

2.1.1 Exercise 1.1

Based on `data_e1_1.csv`: Plot centered speech rate against f0 midpoint, colouring by condition and faceting by vowel. Describe what you see.

2.1.2 Exercise 1.2

Based on `data_e1_2.csv`: Plot logged reaction times by language, by environment, and by age. Describe what you see.

2.1.3 Exercise 1.3

Based on `data_e1_3.csv`: Plot the proportion of correct responses for each trial in the easy and difficult conditions, faceting by priming setting. Describe what you see.

2.2 Exercise 2: Critiquing and correcting plots

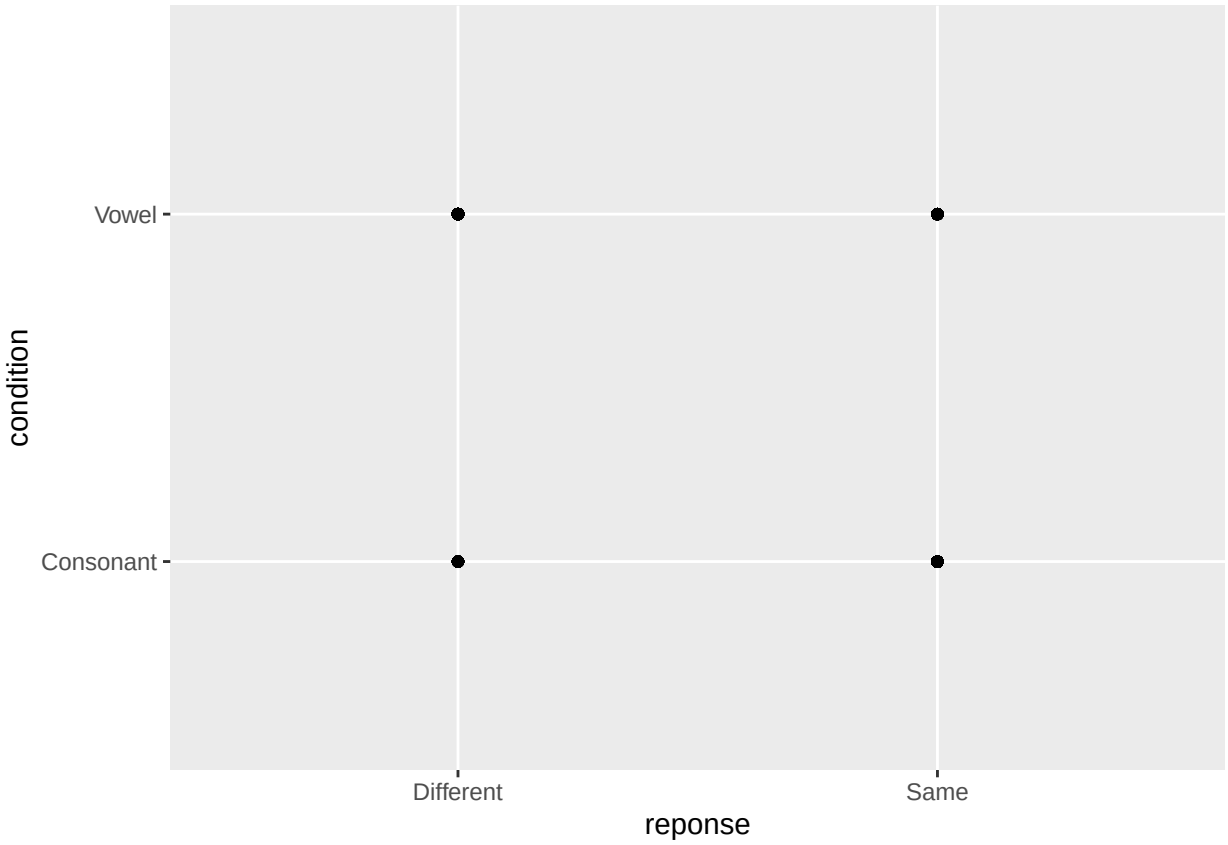
The following two plots are not appropriate for the type of data they show. Briefly describe what is wrong with each plot, try to figure out what the plots might be aiming to visualise, and write your own code to create a more appropriate plot for the exact same data. (If you're unsure about the kind of data you're dealing with, having a look at the data frame might help.)

2.2.1 Exercise 2.1

```
data_e2_1 <- read_csv("data/data_e2_1.csv")
```

```
## Rows: 600 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): reponse, condition
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
data_e2_1 %>%
  ggplot(aes(reponse, condition)) +
  geom_point()
```

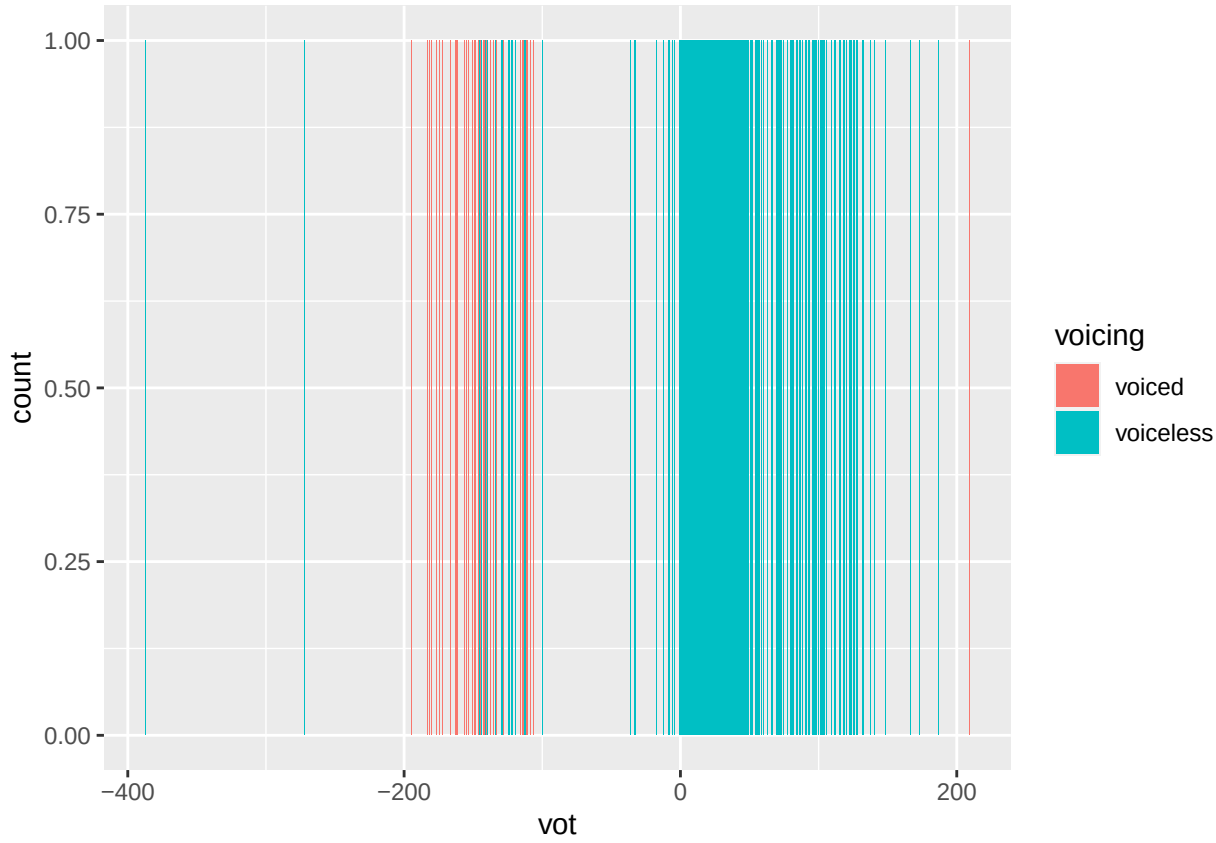


2.2.2 Exercise 2.2

```
data_e2_2 <- read_csv("data/data_e2_2.csv")
```

```
## Rows: 1035 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): voicing
## dbl (1): vot
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
data_e2_2 %>%
  ggplot(aes(vot, fill = voicing)) +
  geom_bar(width=0.5)
```



2.3 Exercise 3: Choosing appropriate summary measures

Read in `data_e3.csv` and obtain summary measures (central tendency: mean, median, or mode; dispersion: standard deviation or range) for each variable in the data. Make sure to pick the correct measure(s) for the respective variable type.

Then, for each variable, briefly state the type of variable, the chosen measure(s), and report the value of the measure(s).

2.4 Exercise 4: Identifying probability distributions

For each variable in the table below, specify in the “Probability distribution” column whether it’s (in principle) distributed according to a Gaussian, a log-normal, or a Bernoulli distribution, or according to some different one (put “other” in this case).

If you have doubts about any of the variables, you can write about it briefly below the table.

	Variable	Probability distribution
1	Vowel duration (ms)	
2	Formant values (hz)	
3	Accuracy (binary)	
4	Readability (0-100)	
5	Reaction times (ms)	
6	Number of relative clauses	
7	Scots vs English	
8	Counts of infant gestures	
9	Logged reaction times	
10	Ratio of 1st vs non-1st pronouns	

2.5 Exercise 5: Contrast coding

Look at the following table that represents the coding of a categorical predictor `background`.

	Bangladeshi	Mandarin	English
Bangladeshi	1	0	0
Mandarin	0	2	0
English	0	0	3

Assume that the intention was to use treatment coding, and to order the levels according to R's default alphabetical order. Briefly explain what is wrong with this table, and create a new table with the correct coding scheme (you can use https://tablesgenerator.com/markdown_tables# to format the table in markdown).

2.6 Exercise 6: Running a Bayesian linear model

Imagine you've just run the following study:

> Participants are asked to listen to nonce words and choose whether they think the word refers to a small object or a big object. Half of the words are of the `kiki` type (back consonants and high front vowels), while half are of the `baba` type (front consonant and low vowels). The expectation is that `kiki` words should elicit more `small` responses and `baba` words more `big` responses. The data also contains reaction time, and the expectation there is that shorter reaction times would correlate with a greater effect of hearing a `kiki` word vs a `baba` word.

Read the `data_e7.csv` file. It contains the following columns:

- `subject`: the subject's ID.
- `response`: whether the subject has chosen `small` or `big`.
- `condition`: `kiki` vs `baba` word.
- `RT`: reaction time in ms.

As needed, change columns to factors and specify the order of levels.

Run a linear model that helps you answer the following questions:

1. Do `baba` words elicit more `big` responses than `kiki` words do?
2. Is the effect of `baba` words on response greater with shorter reaction times?

(Optional: If it helps, you can try creating plots or summary data frames that show how the data is distributed across these variables. This could offer a sense check for whether your model is producing plausible output.)

In code *and* in a brief written description, report the model specification and the results. Also visualise the model's estimates.

2.7 Exercise 7: Critiquing and correcting a linear model

Imagine you are asked to review a paper. Below you can find the description of a mock study, including the details of the linear model the researchers have run. The results are not included.

> We recorded 50 subjects while they read 100 sentences on a screen. For each subject, half of the sentences were presented together with pictures of natural landscapes. The other half was presented together with pictures of urban landscapes. Background sounds were delivered via headphones to the subject in each trial of the natural and urban setting condition. For each setting (natural vs urban), half of the trials had natural sounds (birds, waterfalls, wind, waves, thunders) and half had urban sounds (traffic noise, sirens, people walking).
>
> For each trial we measured speech rate as number of syllables per second (syl/s). The hypothesis is that speech rate will be faster in the urban setting trials relative to the natural setting trials. Moreover, the effect of background noise (natural vs urban sounds) will decrease speech rate in the natural setting but not in the urban setting. In other words, we expect the visual setting (natural vs urban) to prime speakers to slow down their speech rate, but we expect the auditory setting (natural vs urban) to make a difference only in the visual natural setting.
>
> To assess these expectations, we ran a linear model using a Gaussian distribution with visual setting (natural vs urban) as the outcome variable. We included the following predictors: speech rate (centered) and auditory setting (natural vs urban). In R syntax: `brm(visual ~ speech_rate_c + auditory)`.

Now, critique the analysis (i.e., explain what is wrong with it) and run a more appropriate linear model to assess the research hypotheses of the study based on the provided data (`data_e8.csv`). Report the model specification and the results of your linear model with respect to the hypotheses.

2.8 Exercise 8: True or false

Read the statements below. For each, indicate whether the statement is true or false by adding an "x" in the relevant column of the table below.

Statements

1. Frequentist CIs and Bayesian CrIs have the same interpretation.
2. The types of predictor variables in a linear model decide which distribution family to use in the model.
3. Strip charts are among the plots that can be used to visualise the individual observations of a continuous variable.
4. All population-level effects returned by a model summary are conditional posterior probabilities.
5. 95% CrIs are more informative than 60% CrIs.
6. Quantitative methods are objective way to work with data.
7. The most appropriate distribution family for a binary variable is the Bernoulli family.
8. About 68% of the data in a Gaussian distribution is contained within the range marked by "mean – 1 SD" and "mean + 1 SD".
9. `plogis(0.3 + 0.2)` is equivalent to `plogis(0.3) + plogis(0.2)`.
10. The goal of statistics is to definitively accept or reject hypotheses.
11. When creating a density plot, the *y*-axis can be interpreted as the absolute probability of obtaining a specific value.
12. The number of parameters to estimate in a Bayesian model is equivalent to the number of predictors in the model.

TRUE or FALSE

	True	False
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		

2.9 Exercise 9: Using interactions

Explain when it is important to include interactions between predictors in your model.