SQMB - Summative Assessment 1

YOUR EXAM NUMBER

2023-02-22

1 Instructions

PLEASE READ CAREFULLY

DUE THURSDAY 8 DECEMBER AT NOON

Remember to include your **exam number as the author** in the document preamble above.

Please, **use a sans-serif font**. I recommend the DejaVu Sans font. You can download the DejaVu fonts here: http://sourceforge.net/projects/dejavu/files/dejavu/2.37/dejavu-fonts-ttf-2.37.zip.

Complete each exercise by completing tasks, answering questions and/or providing code if required.

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 Excercise 1

- You are given three data files and you have to create one plot for each.
- Each data file requires you to read in the file, filter/transform the data and create a plot that shows a particular aspect of the data.
- Add a brief description of the patterns you notice in the plot.
- Feel free to add more code chunks as needed.

2.1.1 Exercise 1.1

`data_e1_1.csv`: Plot centred speech rate against f0 at vowel mid-point, by condition and group, faceting by vowel.

2.1.2 Exercise 1.2

`data_e1_2.csv`: Plot logged reaction times by language, environment, and age.

2.1.3 Exercise 1.3

`data_e1_3.csv`: Plot proportions of incorrect vs correct responses across trial number in the easy and difficult condition, faceting by priming setting.

2.2 Exercise 2

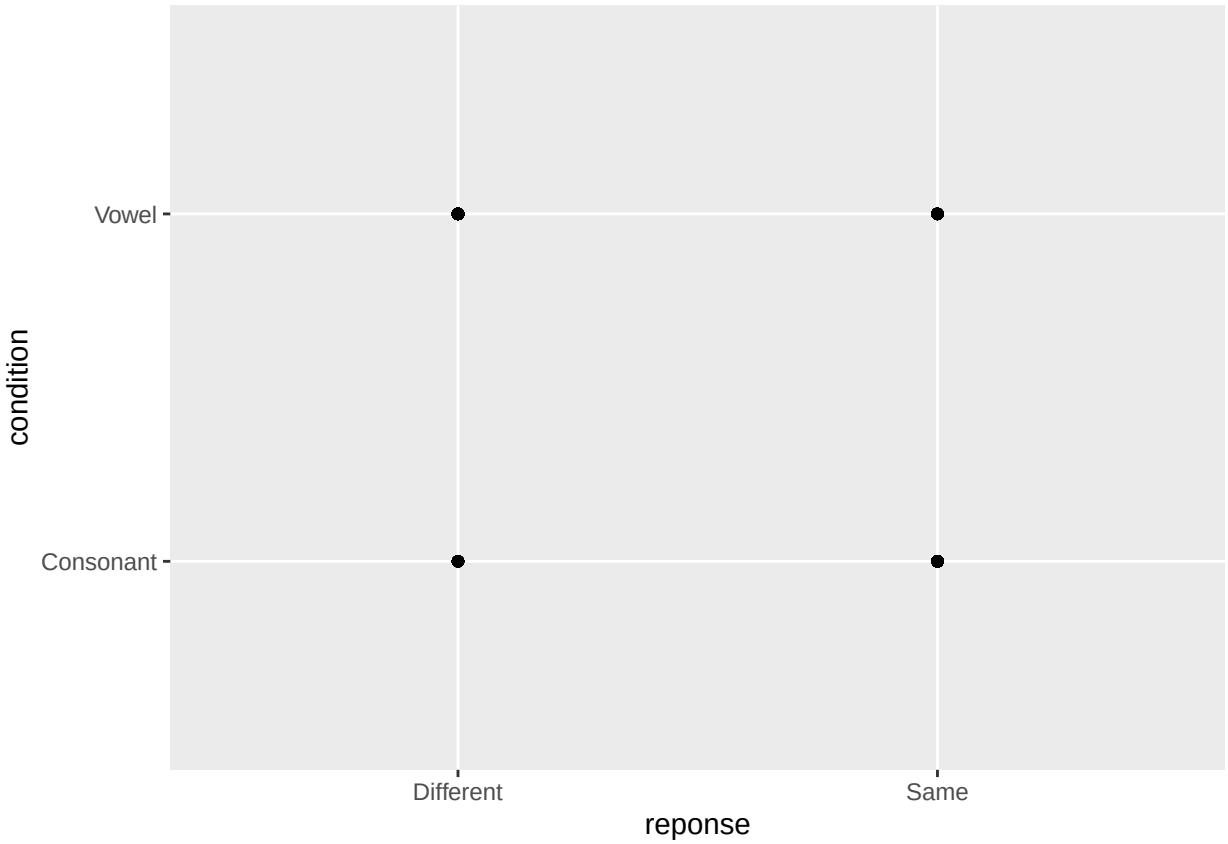
- The following plots are not appropriate for the type of data they show.
- Briefly describe what is wrong with the plot and write code to create a more appropriate plot (looking at the data frame might help you).
- Feel free to add more code chunks as needed.

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()
```

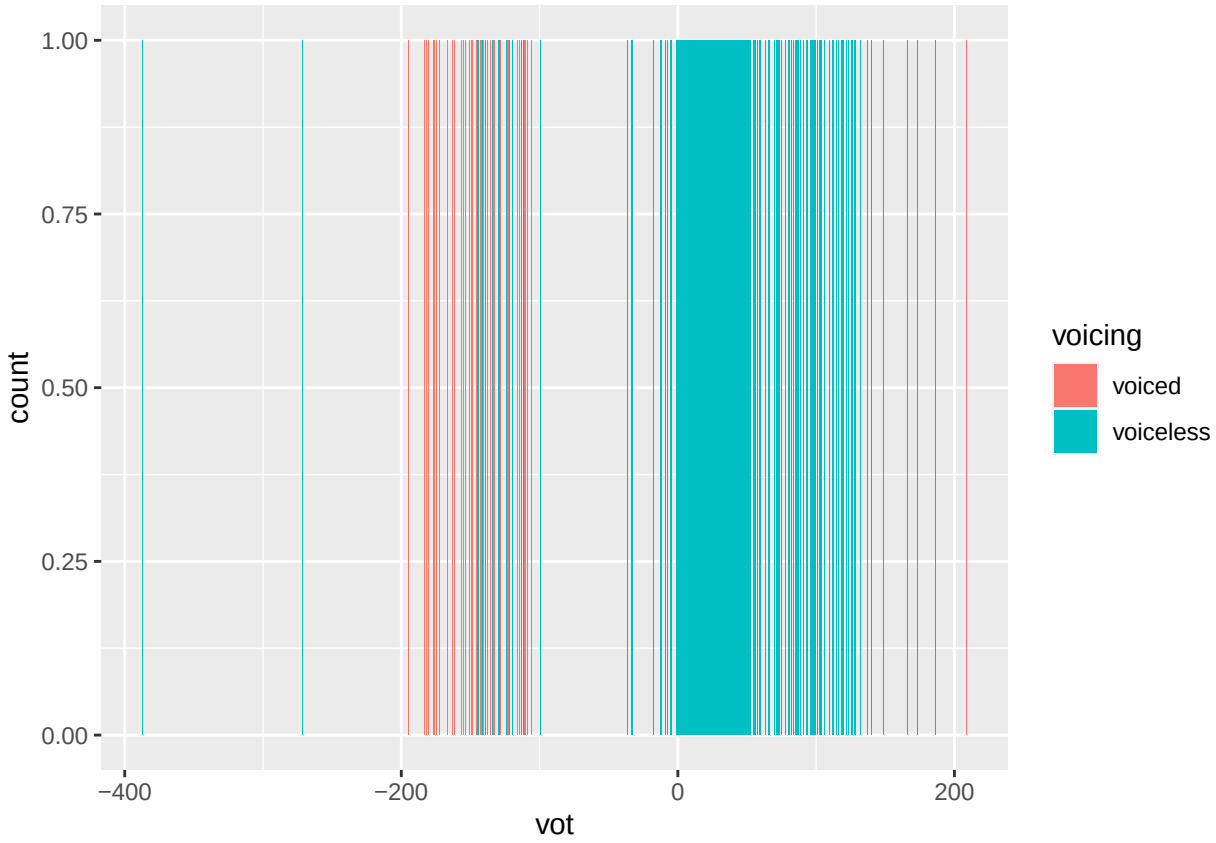


2.2.2 Exercise 2.2

```
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()
```



2.3 Exercise 3

- Read the `data_ex_3.csv` file in R and obtain summary measures (central tendency: mean, median, or mode; and dispersion: standard deviation or range) for each variable in the data.
- Make sure to pick the correct measure depending on the variable type.
- For each variable, state the type of variable and the chosen measures and report the value of the measures.
- Feel free to add more code chunks as needed.

2.4 Exercise 4

- For each of the variables in the table, specify in the `Probability distribution` column whether it's (in principle) distributed according to a Gaussian, log-normal, Bernoulli, Pois-

son distribution or to a different distribution (put "other" in this case).
- If you have doubts about any of the variables, write about it in your words below.

	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

- Look at the following table that represents the coding of a categorical predictor `background`.
- Assume treatment coding has been used and that the default alphabetical order was intended for the order of the levels.
- Explain what is wrong with the table and write a new table with your solution (you can use https://tablesgenerator.com/markdown_tables# to format the markdown table).

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

2.6 Exercise 6

Imagine you run the following study:

Participants are asked to listen to nonce words and choose whether they think the word refers to a small 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. We also recorder reaction time and we expect shorter reaction times to 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.

Make sure to change columns to factors if needed and to 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?
2. Is the effect of `baba` words on response greater with shorter reaction times?

Report the model specification and the results. Also include a plot of the model results.

2.7 Exercise 7

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 (centred) and auditory setting (natural vs urban). In R syntax: `lm(visual ~ speech_rate_c + auditory)`.

Now criticise 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 results of your linear model.

2.8 Exercise 8

- Read the 15 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. A study result is *robust* when a very similar result is obtained with the same data but a different analysis.
5. The population-level effects returned by a model summary are conditional posterior probabilities.
6. Sum-coding a predictor sets the model intercept to the grand mean across the levels of that predictor.
7. 95% CrIs are more important than 60% CrIs.
8. Quantitative methods are an objective enterprise.
9. The most appropriate distribution family for a binary variable is the Bernoulli family.
10. About 68% of the data in a Gaussian distribution is contained within the range marked by "mean - 1 SD" and "mean + 1 SD".
11. `plogis(0.3 + 0.2)` is equivalent to `plogis(0.3) + plogis(0.2)`.
12. The goal of statistics is to test for statistical significance.

13. When creating a density plot, the *y*-axis can be interpreted as the absolute probability of obtaining a specific value.
14. Reproducibility and replicability are the same thing.
15. The number of parameters to estimate in a 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		
13		
14		
15		

2.9 Exercise 9

- Explain why it is important to include interactions between predictors when including multiple predictors in a model.