QML - Summative Assessment 1

YOUR EXAM NUMBER

2023-10-18

1 Instructions

PLEASE READ CAREFULLY

DUE Week 8 - Thu 9 November 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 render this Rmd file now, 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.
- If you are having issues with this, feel free to use any other sans-serif font you have installed on your machine.

This assessment covers Weeks 1 to 7.

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 plot and patterns you notice in each plot you make. You should also mutate the data if necessary and add informative labels.

Note that for each exercise we expect you to create a single plot, i.e. data should be split according to all the variables listed in the exercise instructions. Practically, this means you should use one single call of `ggplot()` per exercise.

2.1.1 Exercise 1.1

- Read `data_e1_1.csv`, from https://lingbuzz.net/lingbuzz/006708.
- Make a plot that shows, for each age group, what proportion of responses correspond to each value (i.e., each score on the acceptability Likert scale), and split this plot by restrictor condition (0 and 1, i.e., absent and present).
- Describe the plot and the patterns you see.

2.1.2 Exercise 1.2

- Read `data_e1_2.csv`, from https://doi.org/10.1016/j.cognition.2008.12.007. The files contains data from the Japanese participants only.
- "H" is "homophone", "LR" is the /l~r/ contrast and "PB" the /p-b/ contrast. "F" are the fillers.
- Filter the data so that the only Procedure it contains is trial data (`TrialProc`) and all contrasts except `F`.
- Create a plot that shows the proportion of correct/incorrect responses in each condition for each contrast.
- Describe the plot and the patterns you see.

2.1.3 Exercise 1.3

- Use the same data frame as for Ex 1.2 (`data_e1_2`).
- Plot *logged* reaction times in each condition, also dividing the plot by accuracy and contrast (using any method or combination of methods that effectively distinguishes these variables: different colours, different facets, etc.). Also include points that represent the median values for each combination of condition, accuracy, and contrast.
- Include median values as points in each combination of condition, accuracy and contrast.
- Describe the plot and the patterns 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 will help.

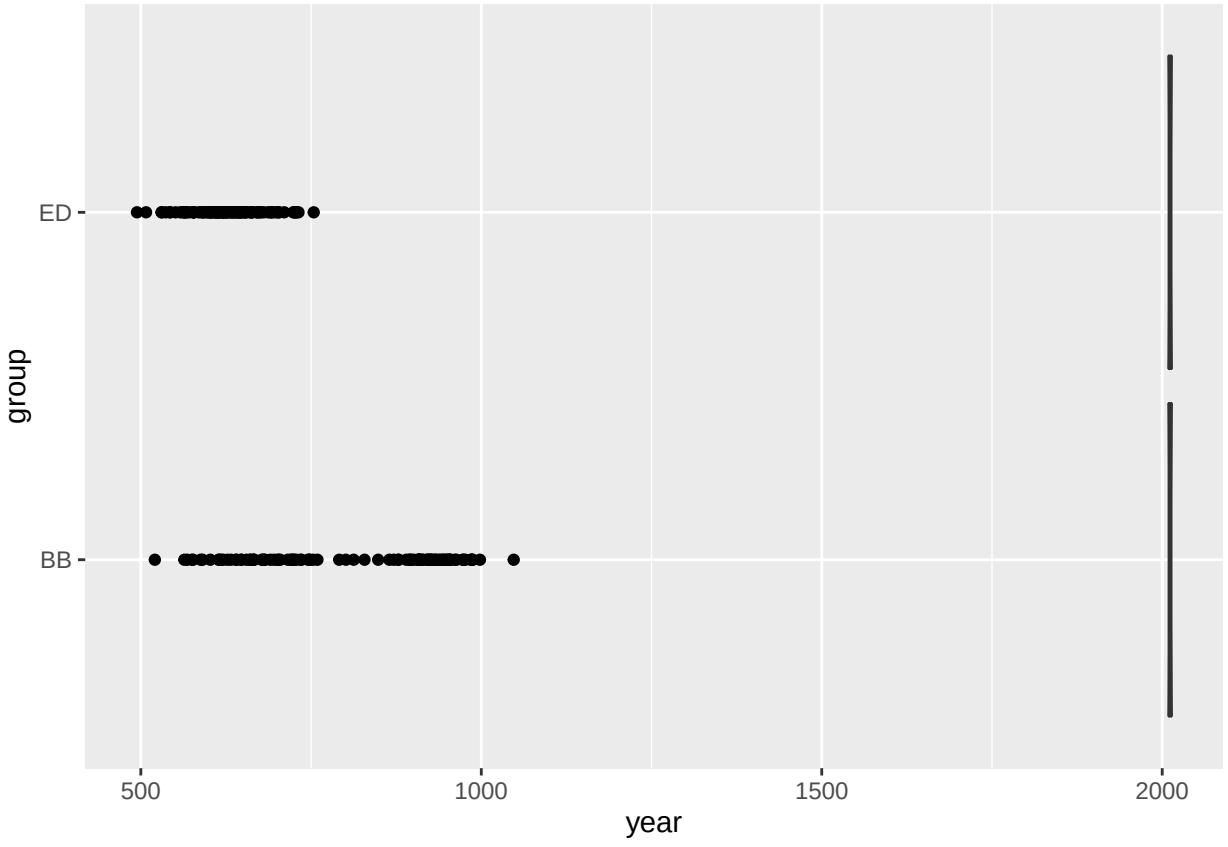
2.2.1 Exercise 2.1

- Data from Vocal Learning in the Functionally Referential Food Grunts of Chimpanzees. The original data was not available, so the current data frame contains simulated data based on the values reported in the paper.

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

```
## Rows: 200 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): group
## dbl (2): year, peak_freq
##
## 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(year, group)) +
  geom_violin() +
  geom_point(aes(x = peak_freq))
```



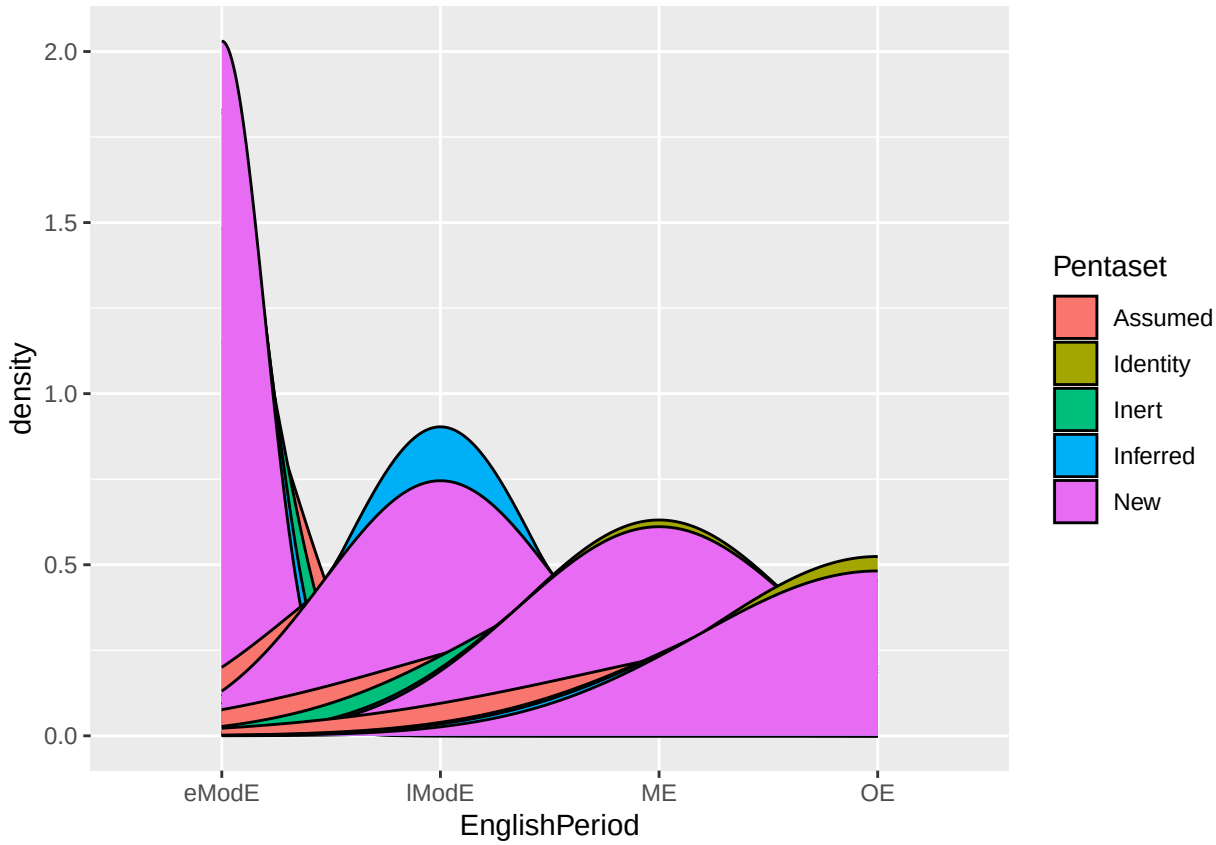
2.2.2 Exercise 2.2

- The data is from The decline of local anchoring: a quantitative investigation and includes occurrences of prepositional phrases in different texts.
- The English period labels are: “OE” = Old English, “ME” = Middle English, “eModE” = Early Modern English, “lModE” = Late Modern English. Make sure the order of the English periods reflects the historical order (as given here).

```
data_e2_2 <- read_csv("data/data_e2_2.csv") %>%
  filter(
    Include == 1
  )
```

```
## New names:
## Rows: 21558 Columns: 34
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (25): TextId, SubType, Title, Author, Genre, Cat, Locs, Locw, ft_clsMain... dbl
## (7): ResId, Date, Size, Words, ft_antdist, Include, TextList lgl (2): ...32,
## ...33
## 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.
## * `` -> `...32`
## * `` -> `...33`
## * `` -> `...34`
```

```
data_e2_2 %>%
  ggplot(aes(EnglishPeriod, fill = Pentaset)) +
  geom_density()
```


2.0
1.5
1.0
0.5
0.0
density
eModE
lModE
ME
OE
EnglishPeriod
Pentaset
Assumed
Identity
Inert
Inferred
New

2.3 Exercise 3: Choosing appropriate summary measures

- Read in `data_e3.csv`. This is simulated data of P300 measurements (Event Related Potential component P300) taken during an auditory odd-ball task by participants with Autistic Spectrum Disorder and a control group. The odd word in each trial differed by a set of features (controlled for as part of the experimental design), the number of which is recorded in `n_feats` (from 1 to 4).
- Obtain summary measures (central tendency: mean, median, or mode; dispersion: standard deviation or range):
  - For each variable on their own.
  - For P300 in each group/response combination.
- Make sure to pick the correct measure(s) for the respective variable type.
- Report all the measures in writing as you would in a paper.

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	Old vs young	
2	Counts of verb occurrences	
3	Hand (left vs right)	
4	Entropy (0 to 1)	
5	Reaction times (seconds)	
6	Non-binary vs female vs male	
7	Politeness (7 point scale from rudest to most polite)	
8	VOT of voiceless stops	
9	Logged word frequency	
10	Number of f0 peaks	

2.5 Exercise 5: Critiquing and correcting a Bayesian linear model

Let's imagine a group of researchers decided to investigate whether speech rate is affected by what kind of landscape you are viewing. You are asked to review the paper where they describe their study and results.

Here you can find the description of this mock study, including the details of the Bayesian 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, some of the sentences were presented together with videos showing one of urban landscapes, natural landscapes, fireplaces, and roller coasters.
>
> 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 roller coaster setting relative to the urban setting, and slower in the natural and fireplace settings relative to the urban setting. Moreover, we expect that the natural and fireplace setting will have the same effect on speech rate.
>
> To assess these expectations, we ran a linear model using a Gaussian distribution with setting as the outcome variable. We included speech rate as the predictor. In R syntax: `brm(setting ~ speech_rate)`. We will use the following treatment contrast coding:

	nature	fireplace	roller coaster	urban
nature	1	0	0	0
fireplace	0	1	0	0
roller coaster	0	0	1	0
urban	0	0	0	1

Now critique the analysis (i.e., explain what is wrong with it in light of the authors' hypotheses) and run a more appropriate linear model to assess the research hypotheses of the study based on the provided data (`data_e5.csv`). Pay attention to the order of the levels in `setting` (if you are unsure, carefully look at how the hypotheses are formulated).

Feel free to also summarise and plot the data, although it is optional and if you do you don't have to add it in the assessment (but you can if you want to).

Report the model specification and the results of your linear model with respect to the hypotheses (i.e. do the results match the hypotheses or not? Can we be certain?).