Individual Homework 09

```
library(bayesrules) # R package for our textbook
library(tidyverse) # Collection of packages for tidying and plotting data
library(janitor) # Helper functions like tabyl
library(rstan) # for fitting models
library(rstanarm) # for fitting standard regression models
library(broom.mixed) # for tidy() function
library(bayesplot) # helpful plotting functions
library(tidybayes) # helpful for wrangling Bayesian model output
library(patchwork)
```

> **Important!**
>
> This HW/GW/Quiz will be due on Monday of Week 10, instead of the usual Friday. Sorry for any confusion!

1 BR 16.6 (Big words: getting to know the data)

Recall from Section 16.7 the Abdul Latif Jameel Poverty Action Lab (J-PAL) study into the effectiveness of a digital vocabulary learning program, the Big Word Club (BWC) (Kalil, Mayer, and Oreopoulos 2020). In our analysis of this program, we'll utilize weakly informative priors with a baseline understanding that the average student saw 0 change in their vocabulary scores throughout the program. We'll balance these priors by the big_word_club data in the bayesrules package. For each student participant, big_word_club includes a school_id and the percentage change in vocabulary scores over the course of the study period (score_pct_change). We keep only the students that participated in the BWC program (treat == 1), and thus eliminate the control group.

```
data("big_word_club")
big_word_club <- big_word_club %>%
  filter(treat == 1) %>%
  select(school_id, score_pct_change) %>%
  na.omit()
```

(a) How many schools participated in the Big Word Club?
(b) What's the range in the number of student participants per school?
(c) On average, at which school did students exhibit the greatest improvement in vocabulary? The least?
(d) Construct and discuss a plot which illustrates the variability in score_pct_change within and between schools.

2 BR Exercise 16.7 (Big words: setting up the model)

In the next exercises you will explore a hierarchical one-way ANOVA model (16.12) of Y_{ij}, the percentage change in vocabulary scores, for student i in school j.

(a) Why is a hierarchical model, vs a complete or no pooled model, appropriate in our analysis of the BWC program?
(b) Compare and contrast the meanings of model parameters μ and μ_j in the context of this vocabulary study.
(c) Compare and contrast the meanings of model parameters σ_y and σ_μ in the context of this vocabulary study.

3 BR 16.8

Exercise 16.8 (Big words: simulating the model)

(a) Simulate the hierarchical posterior model of parameters $(\mu_j, \mu, \sigma_y, \sigma_\mu)$ using 4 chains, each of length 10000.
(b) Construct and discuss Markov chain trace, density, and autocorrelation plots.
(c) Construct and discuss a pp_check() of the chain output.

4 BR 16.11

Suppose we continue the vocabulary study at each of Schools 6 and 17 (which participated in the current study) and Bayes Prep, a school which is new to the study. In this exercise you'll make predictions about $Y_{new,j}$, the vocabulary performance of a student that's new to the study from each of these three schools j.

(a) *Without* using the `posterior_predict()` shortcut function, simulate posterior predictive models of$Y_{new,j}$ for School 6 and Bayes Prep. Display the first 6 posterior predictions for both schools.
(b) Using your simulations from part (a), construct, interpret, and compare the 80% posterior predictive intervals of $Y_{new,j}$ for School 6 and Bayes Prep.

(c) Using `posterior_predict()` this time, simulate posterior predictive models of $Y_{new,j}$ for each of School 6, School 17, and Bayes Prep. Illustrate your simulation results using `mcmc_areas()` and discuss your findings.
(d) Finally, construct, plot, and discuss the 80% posterior prediction intervals for all schools in the original study.

5 BR 17.7 (adapted)

(a) Simulate the posteriors from model (1) and (2) from BR exercise 17.6
(b) Report a posterior predictive check from each model. Which do you prefer?
(c) Using model (2), identify the person whose reaction time changes the *most* with sleep deprivation. Report their posterior median regression model
(d) Using model (2), identify the person who has the *slowest* baseline reaction time. Report their posterior median regression model
(e) Simulate, plot, and discuss the posterior predictive model (under both 1 and 2) of reaction time after 5 days of sleep deprivation for two subjects: you and Subject 308.

6 IRT for Forensic Fingerprint Examiners

In forensic fingerprint analysis, individual examiners look at a pair of fingerprint images and try to decide if they came from the same source or not. They can also make "inconclusive" determinations if there is not enough information in the images to decide either way. In large(ish) research studies, some examiners report conclusive decisions far more than others, and some questions (pairs of images) receive far more inconclusives than others. Some researchers have hypothesized that, since examiners knew they were participating in a study to determine their error rate, they may have been "strategic" and reported more inconclusives than they typically do in casework. Since not every examiner answers every question, figuring out which inconclusives are due to the questions, and which inconclusives are due to the answers is harder than it initially appears.

In this problem, you'll work with a subset of the 2011 FBI "Black Box" study of fingerprint examiners, called `blackbox_responses`. The distribution of `pct_conclusive` by examiner and question are shown in the histograms below:

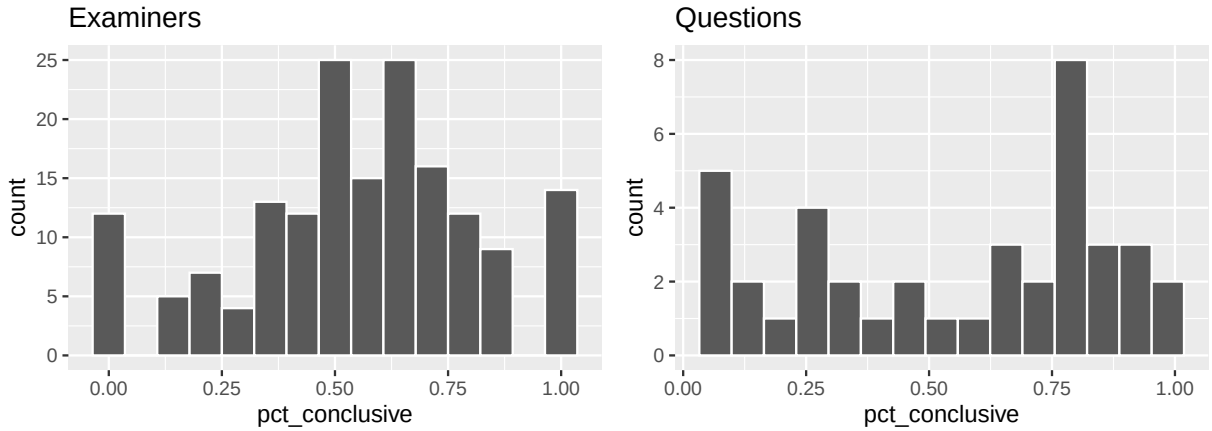
```
responses <- read_csv("https://aluby.github.io/stat340-f25/data/blackbox_responses.csv")

p1 <- responses |>
  group_by(ex_id) |>
  summarize(
    pct_conclusive = mean(conclusive)
  ) |>
  ggplot(aes(x = pct_conclusive)) +
```

```
  geom_histogram(bins = 15, col = "white") +
  labs(title = "Examiners")

p2 <- responses |>
  group_by(q_id) |>
  summarize(
    pct_conclusive = mean(conclusive)
  ) |>
  ggplot(aes(x = pct_conclusive)) +
  geom_histogram(bins = 15, col = "white") +
  labs(title = "Questions")

p1 + p2
```



(a) Compute the percent conclusive for each participant and for each question. Identify the 5 participants who were the *most* and *least* conclusive; and the questions that were the *most* and *least* conclusive

(b) Take a peek at the data and explain why looking at the percent correct for each participant or each question won't give us ideal estimates for tendencies to be conclusive (*Hint:* look at how many questions there are, and how many questions each participant answered)

(c) Fit an IRT model to this data using either {rstanarm} or Stan (I recommend *not* fitting a discrimination parameter). Report the fixed effects table and interpret any fixed effect parameters.

(d) Identify the 5 participants who were the *most* and *least* conclusive; and the questions that were the *most* and *least* conclusive according to the IRT-like model.

(e) Explain any differences between your answers to (a) and (d)