GR5065 Final Exam Answer Key

Ben Goodrich

April 26, 2021

```
set.seed(20210426L)
options(mc.cores = parallel::detectCores())
```

1 Bayes' Theorem in the Real World

Based on

https://www.theguardian.com/world/2021/apr/18/obscure-maths-bayes-theorem-reliability-covid-lateral-flow-tests-probability

1.1 Are You Positive?

Under the assumptions of the article, the prior probability of being infected is $\frac{1}{340}$, the probability that an infected person tests positive is (essentially) 1, and the probability that an uninfected person falsely tests positive is (at most) $\frac{1}{1000}$. Thus, the conditional probability of being infected given a positive test can be expressed using Bayes' Rule as:

$$\Pr\left(\text{infected} \mid \text{positive test}\right) = \frac{\Pr\left(\text{infected}\right) \times \Pr\left(\text{positive test} \mid \text{infected}\right)}{\Pr\left(\text{positive test}\right)} =$$

$$\frac{\Pr\left(\text{infected}\right) \times \Pr\left(\text{positive test} \mid \text{infected}\right)}{\Pr\left(\text{infected}\right) \times \Pr\left(\text{positive test} \mid \text{infected}\right) + \Pr\left(\text{uninfected}\right) \times \Pr\left(\text{positive test} \mid \text{uninfected}\right)} =$$

$$\frac{1/340 \times 1}{1/340 \times 1 + 339/340 \times 1/1000} = \frac{1/340}{1/340 + 0.339/340} = \frac{1}{1.339} \approx 0.75$$

1.2 An Obscure Question

Peuplier's prior beliefs could be represented by a Beta distribution with a median of (at most) 0.02 and a 99-th percentile of 0.1. Unfortunately, the Beta distribution distribution does not have an explicit inverse CDF function, so we would have to arrive at the shape parameters that are consistent with those beliefs by trial and error. However, according to Wikipedia,

https://en.wikipedia.org/wiki/Beta_distribution#Median

the median of a Beta distribution is approximately

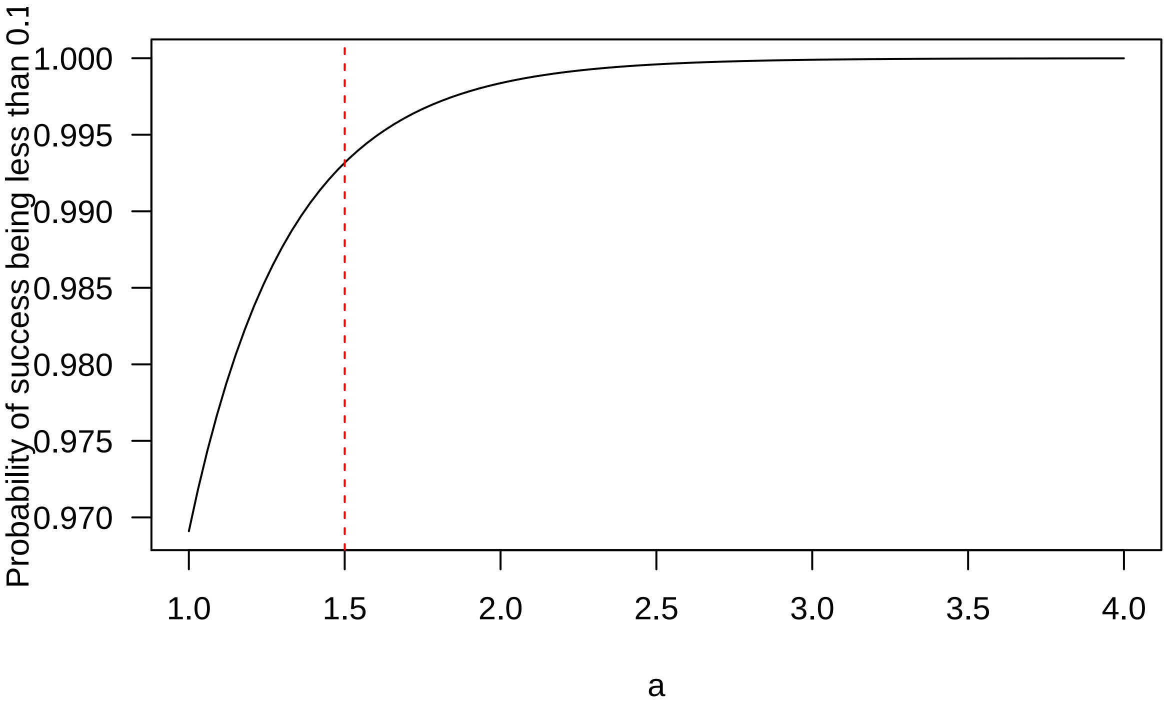
$$F^{-1}\left(0.5 \mid a, b\right) = \frac{a - 1/3}{a + b - 2/3}$$

if and only if both a and b are greater than 1. If we set the right-hand side equal to $\frac{20}{1000}$,

$$\frac{20}{1000} = \frac{a - 1/3}{a + b - 2/3} \implies b = 49a - 16$$

We then need to choose $a > 1$, such that the probability of exceeding $\frac{100}{1000}$ is very small.

```r
curve(pbeta(0.1, a, 49 * a - 16), from = 1, to = 4, xname = "a", las = 1,
      ylab = "Probability of success being less than 0.1")
abline(v = 1.5, col = "red", lty = "dashed")
```



So, we could set $a = 1.5$ and $b = 57.5$ or thereabouts.

The posterior probability could then be computed by exploiting the natural conjugacy between a Beta prior and a binomial likelihood with y observed successes out of $N = 1000$ independent trials that have a common success probability θ.

$$\Pr\left(\theta \mid a, b, y, N\right) \propto \theta^{a-1}\left(1-\theta\right)^{b-1}\theta^{y}\left(1-\theta\right)^{1000-y} = \theta^{a+y-1}\left(1-\theta\right)^{b+1000-y-1} = \theta^{a^{*}-1}\left(1-\theta\right)^{b^{*}-1}$$

where $a^* = a + y$ and $b^* = b + 1000 - y$, which is the kernel of a Beta PDF with shapes a^* and b^*.

1.3 mRNA Vaccines

Bayesian decision theory says to make the decision that maximizes expected utility, or equivalently minimizes expected loss, where the probabilities governing the expectation are given by Bayes rule. In this case, the loss function is assumed to be "excess deaths", which depends on various unknowns that we need a posterior distribution over, such as the effectiveness of the mRNA vaccines after one or two doses, how long the immunity lasts, etc.

BionTech / Pfizer and Moderna only conducted clinical trials of a two-dose regime, although they observed few cases of covid19 among participants in the clinical trials who were two weeks past their first dose and awaiting their second dose. In addition, there is observational data from countries like Israel that people who have only received their first dose tend not to get covid19. You would want to condition on this information when obtaining the posterior distribution of vaccine effectiveness.

In retrospect, it appears as if the British approach was better than the U.S. approach, in part because there are a lot of people in the United States who do not want a vaccine and in part because more mRNA doses became available than were anticipated as of the end of 2020. Thus, almost any adult in the United States who wanted a vaccine could have gotten both doses by now even if the policy were like the British one where few people received second doses until everyone who wanted one received their first dose. However, whether the British and American *decisions* were prudent *at the time* is subject to debate.

1.4 Strategic Lawsuits Against Public Participation (SLAPP)

The Frequentist perspective on the process of “demonstrating a reasonable probability of succeeding in their case” would be something like “among all cases that have essentially the same claims, what proportion of them are found to be in favor of the plaintiff or defendant?” This is difficult to apply because cases are not filed randomly and no two cases have identical claims, which raises the questions of which cases are sufficiently similar and are there enough of them to accurately estimate the probability.

The Bayesian perspective is that judges should update their prior beliefs in light of the claims that are made. This also raises difficult questions as to what priors a judge should have and how much credibility should they give to evidence that may subsequently be challenged on legal and / or factual grounds if the case proceeds and discovery is conducted (although in this particular situation, the allegations pertain to statements on Twitter, YouTube, etc. so I do not know what else needs to be discovered). Despite these problems, I think judges would prefer the Bayesian perspective to the Frequentist one, although they do not explicitly calculate posterior distributions, because the question is essentially one of “degree of belief”, rather than one of “What is the proportion of times that something happens in the limit as the number of randomizations approaches infinity?”

For what it is worth, Mike Postle lost his first anti-SLAPP case against one of the other people that he sued and seems likely to lose the remaining ones.

2 Beta Regression

```r
library(readxl)
dataset <- read_excel("Supplementary-File-Select-county-level-factors-04-06.xlsx",
                      sheet = "County Factors", skip = 1, # to not use the original column names
                      col_names = c("FIPS", "County", "State", "Hesitant", "StronglyHesitant",
                                    "Age18_24", "Age25_39", "Age40_54", "Age55_64", "Age65_",
                                    "Male", "Hispanic", "White", "Black", "Asian", "OtherRace",
                                    "LessHS", "HighSchool", "SomeCollege", "CollegeDegree",
                                    "Married", "Widowed", "Divorced", "NeverMarried",
                                    "SVI", "CVAC", "Vaccinated18_", "Vaccinated65_"),
                      col_types = c("skip", "text", "text", "skip", "skip",
                                    "skip", "numeric", "numeric", "numeric", "numeric",
                                    "numeric", "numeric", "skip", "numeric", "numeric", "numeric",
                                    "skip", "numeric", "numeric", "numeric",
                                    "skip", "numeric", "numeric", "numeric",
                                    "skip", "skip", "skip", "numeric"))
dataset$Vaccinated65_ <- dataset$Vaccinated65_ / 100 # convert from percentage to proportion
dataset <- dataset[!is.na(dataset$Vaccinated65_), ]  # drop a few observations with missing values
```

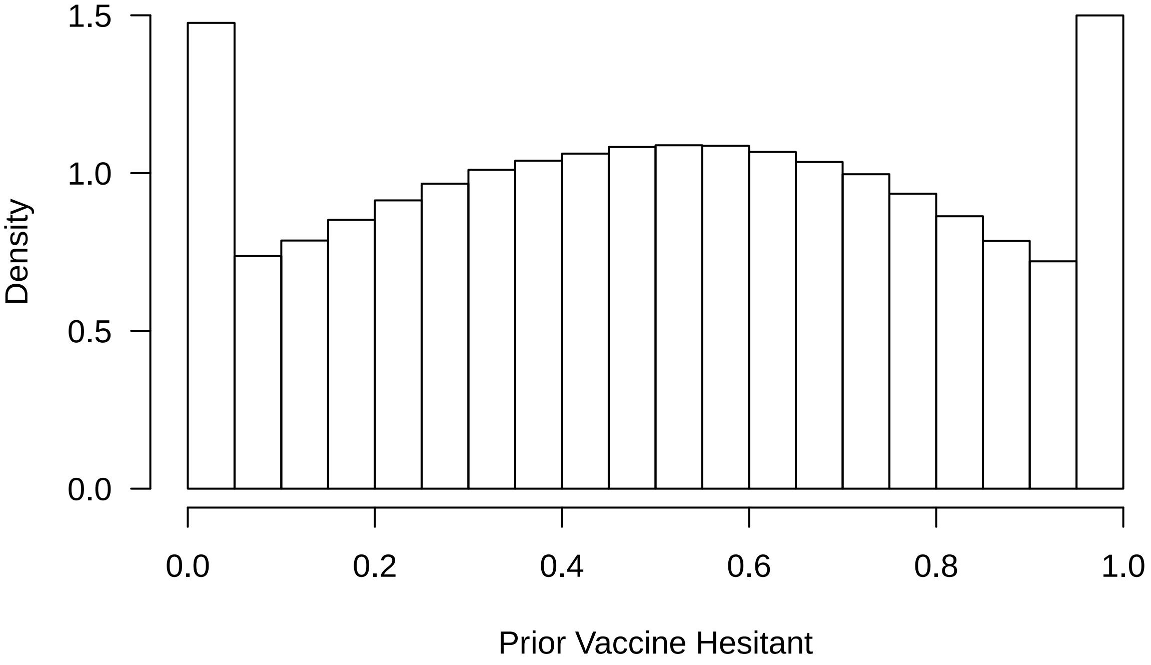
2.1 Prior Predictive Distribution

```r
X <- model.matrix(Vaccinated65_ ~ . - County - State, data = dataset) # design matrix
X <- sweep(X, MARGIN = 2, STATS = colMeans(X), FUN = `-`)            # now centered
draws <- t(replicate(1000, {
  alpha_ <- rnorm(1, mean = 0, sd = 1)
  beta_  <- rnorm(n = ncol(X), mean = 0, sd = 1)
  eta_ <- alpha_ + X %*% beta_
  mu_ <- plogis(eta_)
  phi_ <- rexp(1, rate = 1 / 10) # prior expectation of phi is 10
  a_ <- mu_ * phi_
  b_ <- (1 - mu_) * phi_
  y_  <- rbeta(n = nrow(X), shape1 = a_, shape2 = b_)
  return(y_)
}))
```

Too many people drew the outcome from a normal distribution or if they drew from a beta distribution, drew once rather than `nrow(X)` times (once for each observation).

2.2 Reasonableness

```
hist(c(draws), prob = TRUE, xlab = "Prior Vaccine Hesitant", las = 1, main = "")
```

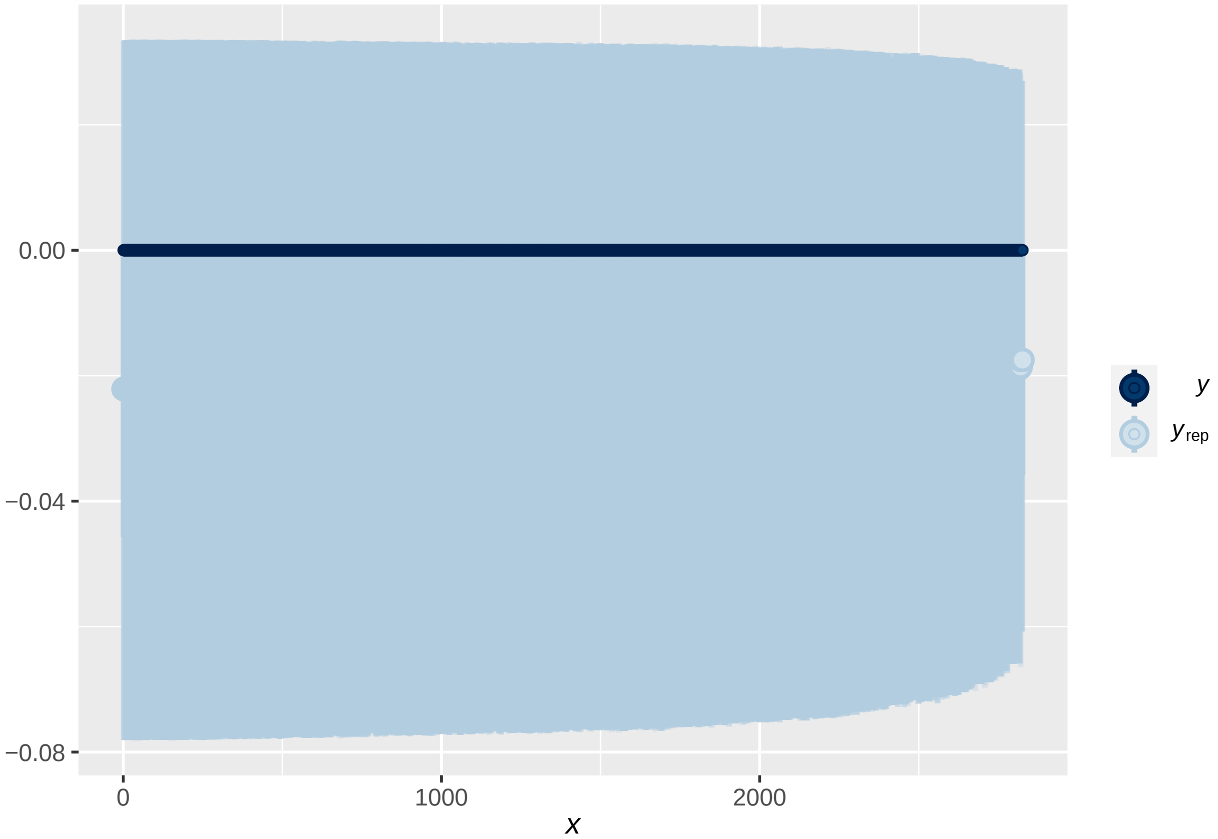


2.3 Posterior Distribution

```
library(rstanarm)
post <- stan_glm(Vaccinated65_ ~ . - County - State, data = dataset,
                 family = mgcv::betar, prior_intercept = normal(0, 1),
                 prior = normal(0, 1), prior_aux = exponential(rate = 1 / 10))
```

2.4 Interpretation

```
high <- dataset
high$Male <- 0.6
low <- dataset
low$Male <- 0.4
mu_diff <- posterior_epred(post, newdata = high) - posterior_epred(post, newdata = low)
mu_diff <- mu_diff[ , order(apply(mu_diff, MARGIN = 2, FUN = median))]
bayesplot::ppc_intervals(y = rep(0, length = ncol(mu_diff)), yrep = mu_diff)
```



In any one county, the expected difference in the outcome between having 60% men vs. 40% men is very uncertain and could be positive or negative. However, in aggregate the difference is more likely to be negative than positive.

2.5 Posterior Predictive Checks

One would think that this

```
PPD <- posterior_predict(post)
```

or something like `pp_check` that calls the above internally would allow you to investigate how well the model fits. However, due to a bug in rstanarm with `family = mgcv::betar`, it throws an error. So, I just gave everyone the points, but you could have persisted by drawing from the posterior predictive distribution yourself with something like

```
mu <- posterior_epred(post)
phi <- as.data.frame(post)$`(phi)`
a <- mu * phi
b <- (1 - mu) * phi
y <- matrix(rbeta(n = prod(dim(a)), shape1 = a, shape2 = b),
            nrow = nrow(a), ncol = ncol(a))
```

```
bayesplot::ppc_intervals_grouped(y = dataset$Vaccinated65_, yrep = y,
                                 x = dataset$CollegeDegree, group = dataset$State,
                                 facet_args = list(ncol = 4))
```

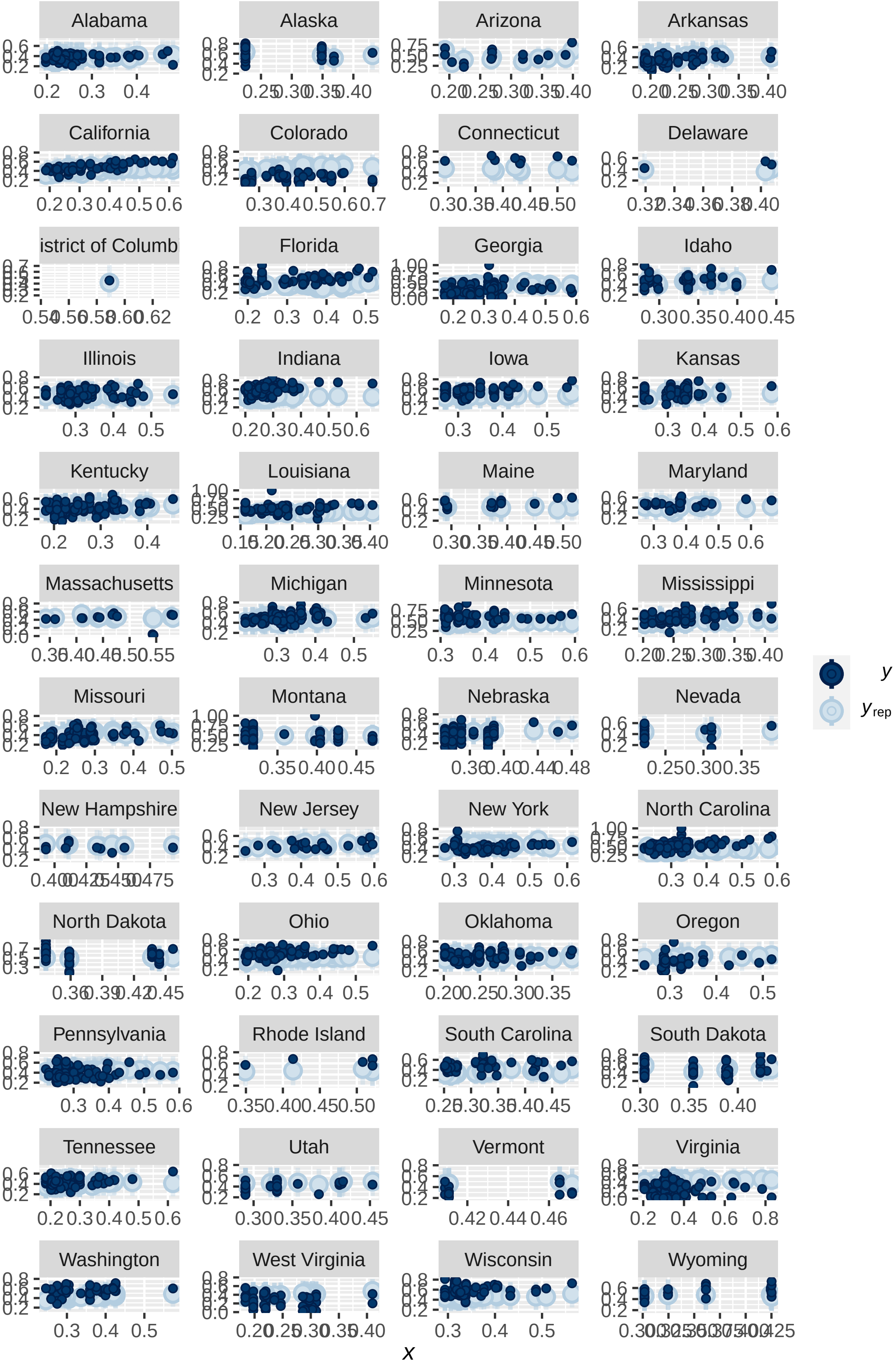


Figure 1: Predictive Accuracy by State

3 US Presidential Election

```
library(readr)
polls <- read_csv("2020 US presidential election polls - all_polls.csv",
                  col_types = cols(start.date = col_date(format = "%m/%d/%Y"),
                                   end.date = col_date(format = "%m/%d/%Y"),
                                   entry.date.time..et. =
                                     col_datetime(format = "%m/%d/%Y %H:%M:%S")))
polls$end.date[is.na(polls$end.date)] <- "2020/10/08" # fix misformatted dates
polls$mode[112] <- "Online" # fix typo
polls$state[polls$state == "--"] <- "USA" # these were national polls
polls$days_to_election <- as.integer(as.Date("11/03/2020", format = "%m/%d/%Y") -
                                     polls$end.date)
```

3.1 Frequentist Modeling

This dataset is not a cluster random sample. Thus, it would not be appropriate to have random intercepts by state because (essentially) all states are included, rather than having a small sample of states from a much larger population. However, it could be appropriate to have a random intercept for each poll.

3.2 Bayesian Modeling

```
library(brms)
my_prior <- prior(normal(0, 1), class = "b") + prior(normal(0, 2), class = "Intercept") +
  prior(exponential(10), class = "sd") + prior(exponential(5), class = "sds") +
  prior(exponential(10), class = "sigma")
post <- brm(biden_margin ~ s(days_to_election) + population +
              (1 | pollster) + (1 | state), data = polls, family = gaussian,
            prior = my_prior)
```

3.3 Prediction

```
predict_df <- data.frame(state = sort(unique(polls$state)), pollster = NA_character_,
                         population = "lv", days_to_election = 0L)
PPD <- posterior_predict(post, newdata = predict_df, re_formula = ~(1 | state))
colnames(PPD) <- predict_df$state
as.matrix(sort(colMeans(PPD > 0)))
```

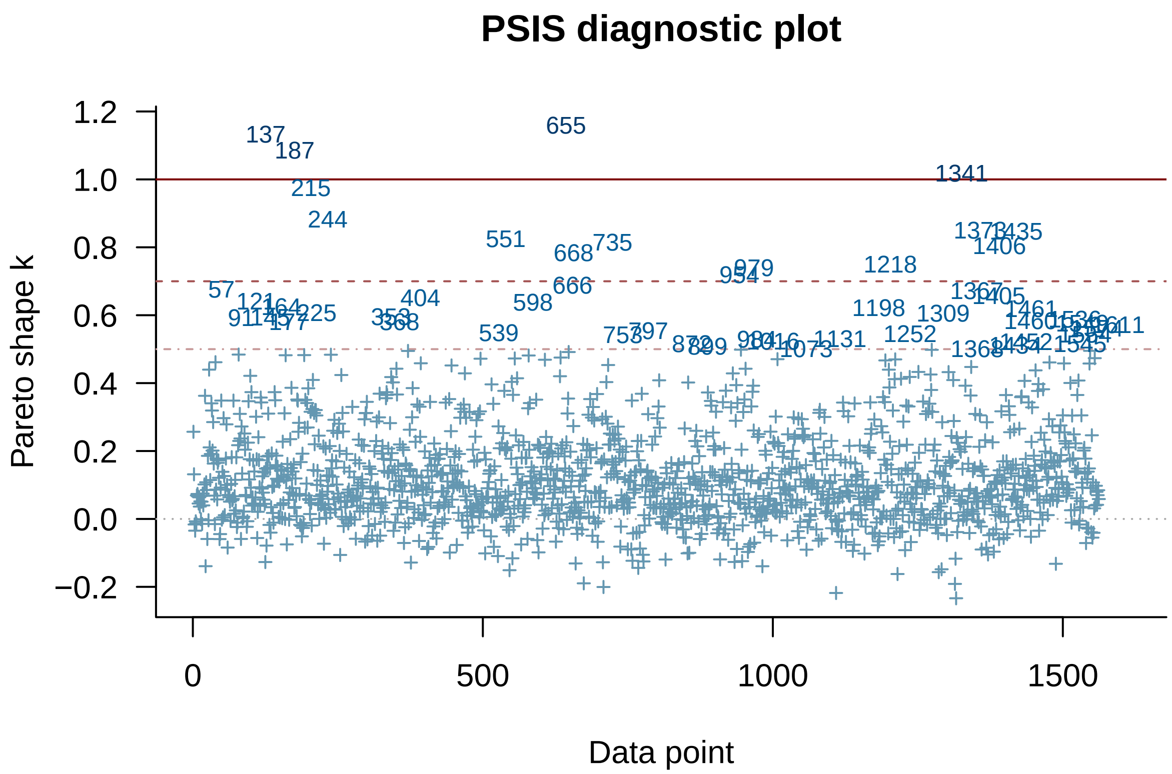
```
##        [,1]
## AL  0.00000
## AR  0.00000
## IN  0.00000
## KY  0.00000
## MS  0.00000
## ND  0.00000
## OK  0.00000
## UT  0.00000
## WV  0.00000
## WY  0.00000
## LA  0.00025
## TN  0.00025
## KS  0.00125
## SD  0.01225
## MT  0.01250
## SC  0.01475
## MO  0.03425
## AK  0.04275
## TX  0.44425
## IA  0.48550
## OH  0.48900
## GA  0.69375
## NC  0.85575
## AZ  0.94225
## FL  0.94550
## NV  0.98600
## PA  0.99125
## NH  0.99675
## WI  0.99675
## MI  0.99925
## USA 0.99925
## MN  0.99950
## CA  1.00000
## CO  1.00000
## CT  1.00000
## DE  1.00000
## HI  1.00000
## IL  1.00000
## MA  1.00000
## MD  1.00000
## ME  1.00000
## NJ  1.00000
## NM  1.00000
## NY  1.00000
## OR  1.00000
## VA  1.00000
## VT  1.00000
## WA  1.00000
```

Too few people recognized that the columns of the matrix of draws from the posterior predictive distribution correspond to each state. Thus, you could compute the probability of Biden winning each state or enough states to win the Electoral College. Too many people calculated something based on the matrix of posterior predictive draws as a whole, which is not politically relevant.

3.4 Pareto k Estimates

```
plot(loo(post), label_points = TRUE)
```

```
## Warning: Found 15 observations with a pareto_k > 0.7 in model 'post'. It is
## recommended to set 'moment_match = TRUE' in order to perform moment matching for
## problematic observations.
```



```
polls[c(137, 187, 215, 244, 353, 363, 450, 655, 668, 1341, 1435, 1574),
      c("state", "pollster", "end.date", "population", "mode")]
```

```
## # A tibble: 12 x 5
##    state pollster                        end.date   population mode
##    <chr> <chr>                           <date>     <chr>      <chr>
##  1 WY    University of Wyoming           2020-10-28 lv         Live Phone
##  2 AR    University of Arkansas          2020-10-21 lv         Live Phone
##  3 LA    University of New Orleans       2020-10-22 lv         Online
##  4 SD    Mason-Dixon Polling & Strategy  2020-10-21 lv         Live Phone
##  5 HI    MRG Research                    2020-10-07 rv         Live Phone
##  6 CO    Keating Research                2020-10-13 lv         Live Phone
##  7 NJ    Fairleigh Dickinson University  2020-10-05 lv         Online
##  8 VT    Braun Research                  2020-09-15 lv         Live Phone
##  9 AL    The Tyson Group                 2020-08-18 lv         Live Phone
## 10 TN    East Tennessee State University 2020-05-01 lv         Live Phone
## 11 WA    EMC Research                    2020-04-06 rv         Online
## 12 MT    University of Montana           2020-02-22 lv         Online
```

The observations with high Pareto k values tended to be polls of states that are rarely polled (because no one cares about those states) and / or by pollsters that rarely conduct polls (often because they are done by a university in that state). Thus, if any one of those observations were omitted, it would have a non-negligible effect on the posterior distribution, particularly for the state and / or pollster deviations in the intercept. There is nothing wrong with the posterior distribution, but it is not possible to rely on the quickest estimator of the ELPD, which assumes that any observation could be dropped without having a major effect on the posterior distribution.