Election Prediction

Bob, Becca, Rena, Grace

10/16/2020

1. Introduction

The outcome of the 2016 election not only stunned the nation, but also sent shockwaves through the statistical and polling communities. Over the course of the election year in 2016 up until the week of the election, poll predictions of Hillary Clinton's likelihood of beating Donald Trump ranged from 71%- 99% probability [1]. So when Trump beat these odds, the polling industry lost a lot of trust from the general public [2].

This small, specialized industry that is the political polling business, has a great deal of influence on how the election is portrayed in the news media, voter decisions, and candidate partisan policy initiatives. [1]. Million dollar decisions on advertising and campaign strategy are dictated by polls.

Polls conducted early in the election year are a weak predictor of election outcomes because the general public has paid less attention to the race and is less knowledgeable of the candidates' platforms. Voters that tend to sway between parties often report that they are undecided, however these are arguably the most important opinions in predicting the election outcomes. [3] As the election nears, polls become more accurate, but that is where historical prediction models have been lacking – they fail to take into account the differential accuracy in poll results. Historical models, regression based models that rely on outcomes from past elections and structural factors, predict election outcomes at a single point in time which renders high levels of uncertainty [4].

Drew A. Linzer developed a dynamic bayesian forecasting model to predict the U.S. presidential election at the national and state level that combines these historical models with everchanging poll updates. Linzer's model uses hierarchical specification to handle states being polled on different days and takes into account sampling errors of the polls and national campaign effects [4].

In this report we aim to:

- Predict the outcome of the presidential election and the electoral college vote using the Linzer model to predict swing state outcomes in combination
- Predict whether the US Senate remains in Republican control by using an adaptation of the Linzer model and FiveThirtyEight Senate poll data for each state
- Predict the outcomes of all 13 NC Congressional elections using Linzer model with input from FiveThirtyEight Senate poll data for North Carolina along with our model that predicts who will vote in North Carolina
- Predict the outcome of the NC Senate election and the associated uncertainty using the Linzer model.

Taking into account the anomaly that was the 2016 election, we chose to use the Linzer model to best account for slight and unexpected changes leading up to election day. In Section 2 of this report we will discuss datasets used for all tasks, including a brief exploratory data analysis. In Section 3, we formulate models and methodologies that answer all the research questions. Section 4 will present model diagnostic and sanity check validation of prediction. Then, in section 5 we present the major results of our analysis. Section 6 will be focused on conducting sensitivity analysis to test data imputation hypothesis and prior choices.

2. Data Source and EDA

2.1 Description of Data

In order to answer these questions, we used a total of four datasets: senate polls, house polls, 2020 US presidential election polls, and North Carolina voter registration history snapshot dataset. Both the senate and house polls dataset comes from the fivethirtyeight website [5], while the presidential election polls dataset comes from the Economist website [6].Senate and house polls have 38 variables and 4061 and 2655 observations respectively. The presidential election polls have 1447 observations and 17 variables which are: poll state, pollster, sponsor, start and end date, entry time, number of observations, population, method, Biden, Trump, Biden margin, other, undecided, URL, include, and notes.

The North Carolina voter history dataset contains information on the 2016 elections and about how voters voted (method, election, county, party affiliation, precinct). This will help us model and identify potential voters in 2020 (Interim report). On top of the created model, we used the 2020 voter registration snapshot dataset to identify potential voters, and use this information as additional input to model house polls results.

2.2 Exploratory Data Analysis

To get a basic understanding of the four datasets we were working with, we went through the data to understand what we were working with. In the senate and the house polls, we explored the variables of: methodology of the poll, the state in which the poll was conducted, which party the candidate was, and the percentage of the poll. In the presidential election polls, we explored similar variables of: method of poll and state, in addition to the margin between Biden and Trump. Because all three of these datasets had similar variables, we are able to compare them to each other. In Figure A, we can see how the methods of polling were distributed between the house, senate, and presidential polls. In Figure B, we can see how each poll's distribution of which state was polled, and in Figure C, we can see how the house and senate polls' candidate party are different from each other. Finally, in Figure D, we can see the distribution of the margin between Biden and Trump.

To explore the North Carolina voter history dataset, we looked at variables such as method, county, and party affiliation of the voter. We can see in Figure E that the majority of North Carolina voters are from two counties, Wake and Mecklenberg. Party affiliation and the method that the voters used are included in our appendix.

3. Model Formulation

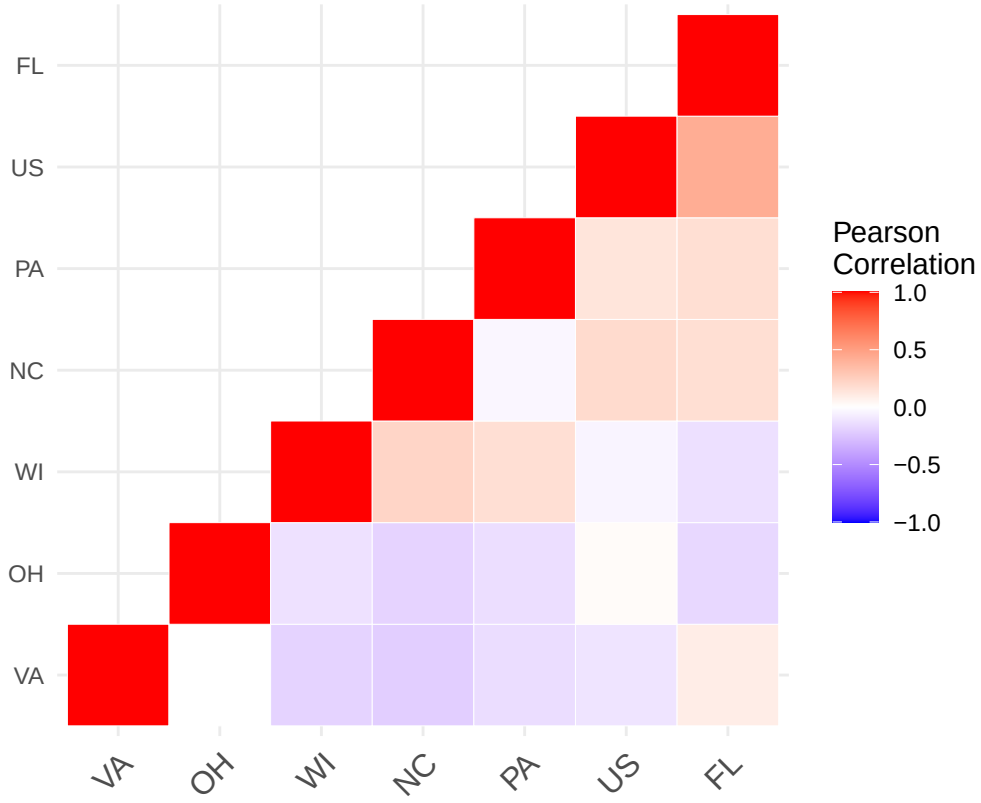
3.1 Model for Question 1

In this section, we build a model to answer our first research question: Predict the outcome of the presidential election and the electoral college vote using the Linzer model to predict swing state outcomes in combination. We follow the route of Linzer and build our model. Specifically, the motivation is that we want to model Joe Biden's percentage of EC vote in the state specific level of detail. We believe that throughout time, each state's preference for Joe Biden is randomly varying and follows a random walk process. Besides, each state's uncertainty of their attitude towards Joe Biden is also state-widely different. We make the assumption that such uncertainty is constant throughout time. Therefore, we create the following extension of Linzer Model.

$$
\begin{aligned}
y_k &\sim N\left(\beta_{i[k]t[k]}, (\sigma_y^2)_{i[k]}\right) \\
\text{for } t > 1 : \beta_{it} &\sim N\left(\beta_{i,t-1}, (\sigma_\beta^2)\right) \\
\text{for } t = 1 : \beta_{i1} &\sim N\left(\mu_0, \sigma_0^2\right) \\
(\sigma_y^2)_{i[k]} &\sim \text{InvGamma}(\nu_y, \tau_y) \\
(\sigma_\beta^2) &\sim \text{InvGamma}(\nu_\beta, \tau_\beta)
\end{aligned}
$$

$$
\begin{aligned}
\mu_0 &\sim N(50, 17) \\
\sigma_0^2 &\sim \text{InvGamma}(\frac{1}{2}, \frac{1}{2}) \\
\nu_y &\sim \text{Uni}(0, 100) \\
\tau_y &\sim \text{Uni}(0, 100) \\
\nu_\beta &\sim \text{Uni}(0, 100) \\
\tau_\beta &\sim \text{Uni}(0, 100)
\end{aligned}
$$

Where k index the polls, i index the states, and t index the date. $i[k]$ represents the k^{th} poll's corresponding state index, and $t[k]$ represents the k^{th} poll's corresponding date index. We model samples from the economist poll data of 2020 and therefore are able to sample the posterior distribution of 1) Joe Biden's chance of winning, and 2) total electoral vote. The sampled posterior distribution y (percentage of supportance within each state) will be calculated by comparing to rounding by 50% to simulate the "winner takes all" procedure. After such adjustment, we can simply multiply the number of electoral college votes of each state to obtain a posterior distribution of electoral college votes of the entire nation. Thus, 2) can be obtained. By comparing to 270 again provides the probability of Joe Biden being elected.



3.2 Model for Question 2, 4

In this section, we're answering question 2 and 4 jointly: Predict whether the US Senate remains in republican control and predict the outcome of the NC Senate election. Similarly, using the above model with exactly the same notation, and by switching Economist dataset to FiveThirtyEight senate polls data, the exact inference procedure can be produced. After modeling and predicting the Republican's support percentage, we can predict whether the Republican or Democrat party wins the state majority. This procedure includes the state of North Carolina. Furthermore, by summing the posterior samples of Republican senators in each state on election day, we can obtain the posterior distribution of US Senate remains in Republican Party's control.

```
senator_poll_plot_data$p <- jags_sims_senator$BUGSoutput$mean$p
senator_poll_plot_data$sigma2y <- jags_sims_senator$BUGSoutput$mean$sigma2_y[senator_jags_data$r]
senator_poll_plot_data$binom_v <- (senator_poll_plot_data$p)*(100-senator_poll_plot_data$p)/ senator$sa


senator_var <- senator_poll_plot_data %>%
  ggplot(aes(x=binom_v)) +
  geom_density() +
  geom_vline(aes(xintercept = sigma2y),color = "red") +
  facet_wrap(~ state)
```

3.3 Model for Question 3

In this section, we answer research question 3: Predict the outcomes of all 13 NC Congressional elections. The dataset we've used is the FiveThirtyEight house poll data. However, the dataset's sample sizes on each NC congressional district are small. This is due to the fact house polls happen less frequently compared to either senate polls or presidential polls. Besides, not all congressional districts are competitive because there is no term limit for representatives. FiveThirtyEight has only recorded poll history data for competitive districts including districts 2,3,7,8,9,11,13. Thus, correlation between non-competitive districts cannot be captured without inputting extra datasets. Thus, we concatenated this section's modeling procedure with the output of the interim reports' "who vote" results.

Using the model to predict who will vote in 2020, we ran this binary output model on all the registered voters in NC to predict whether they'll vote or not. In the NC voter registration profile 2020 snapshot dataset, there is a column indicating party affiliation of each potential voter. Thus, we can predict the percentage of Republican voters for each congressional district and use this percentage as an imputed value for polls percentage. In this way, we're essentially imputing polling results for non-competitive districts, which can be also fed into the Linzer Model defined below. Such imputation is reasonable but bold. Thus, sensitivity analysis of imputed percentage will be further explored later in section ********

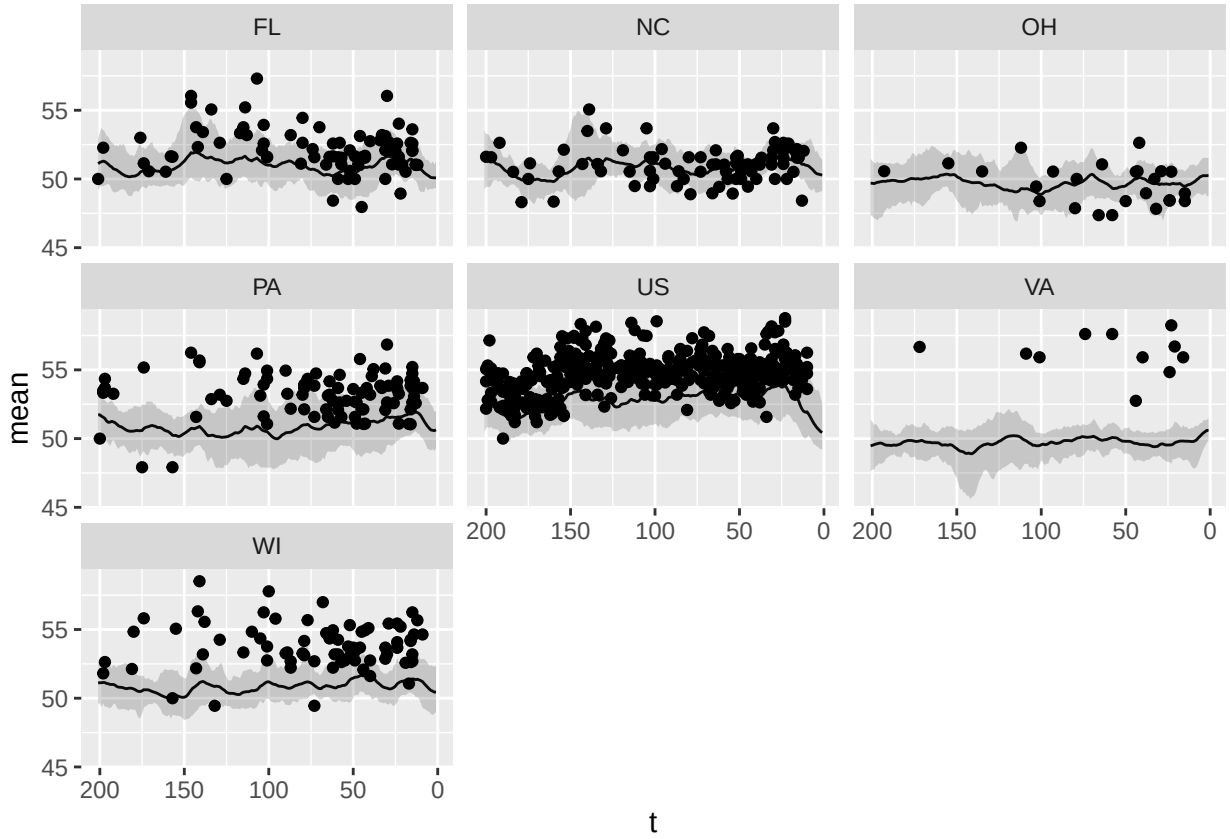
Below is the model we're using for this task:

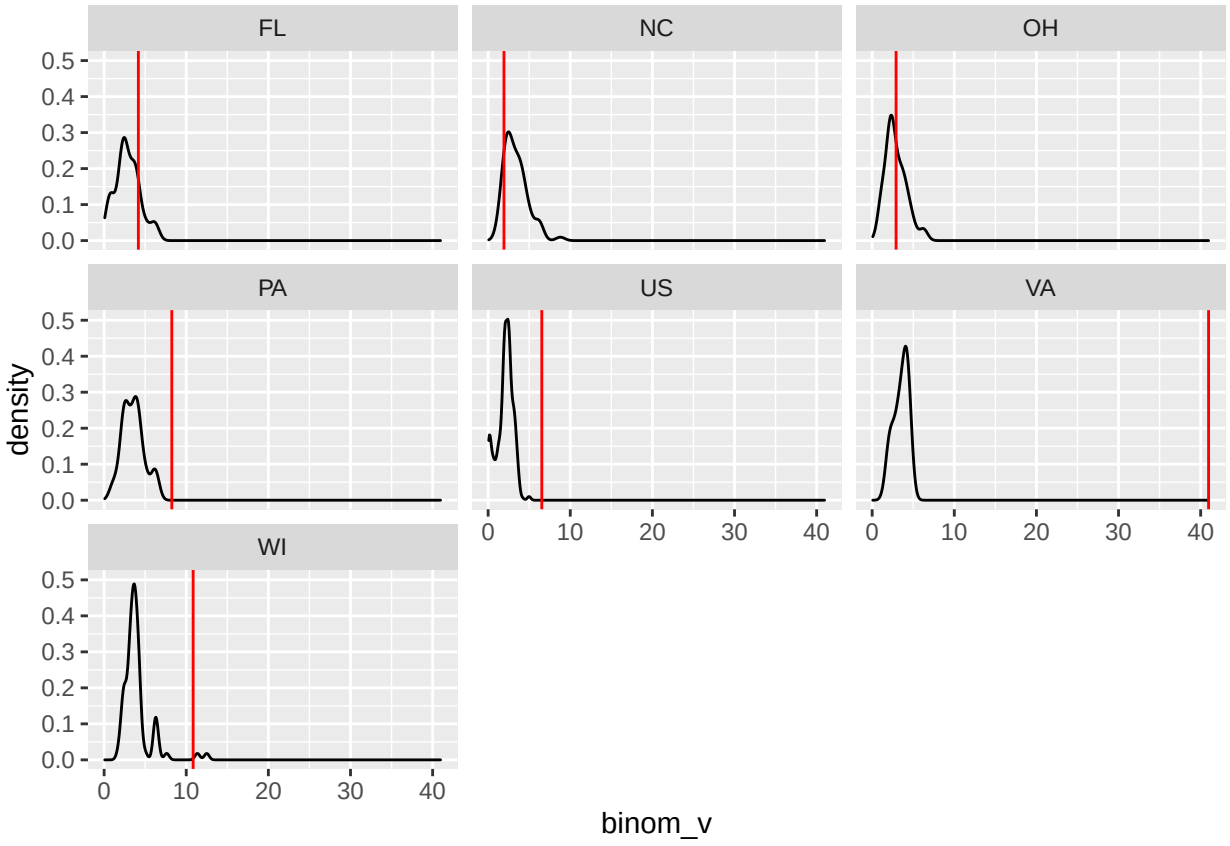
$$
\begin{aligned}
y_k &\sim N\left(\beta_{i[k]t[k]}, (\sigma_y^2)_{i[k]}\right) \\
\text{for } t > 1 : \beta_{it} &\sim N\left(\beta_{i,t-1}, (\sigma_\beta^2)\right) \\
\text{for } t = 1 : \beta_{i1} &\sim N\left(\mu_0, \sigma_0^2\right) \\
(\sigma_y^2)_{i[k]} &\sim \text{InvGamma}(\nu_y, \tau_y) \\
(\sigma_\beta^2) &\sim \text{InvGamma}(\nu_\beta, \tau_\beta)
\end{aligned}
$$

$$
\begin{aligned}
\mu_0 &\sim N(50, 17) \\
\sigma_0^2 &\sim \text{InvGamma}(\frac{1}{2}, \frac{1}{2}) \\
\nu_y &\sim \text{Uni}(0, 100) \\
\tau_y &\sim \text{Uni}(0, 100) \\
\nu_\beta &\sim \text{Uni}(0, 100) \\
\tau_\beta &\sim \text{Uni}(0, 100)
\end{aligned}
$$

Where k index the polls or the imputed polls, i index the congressional district, and t index the polls date. $i[k]$ represents the k^{th} poll's corresponding congressional district index, and $t[k]$ represents the k^{th} poll's corresponding date index.

```
## Warning: Removed 985 rows containing missing values (geom_point).
```






```
## [1] 0.92
```

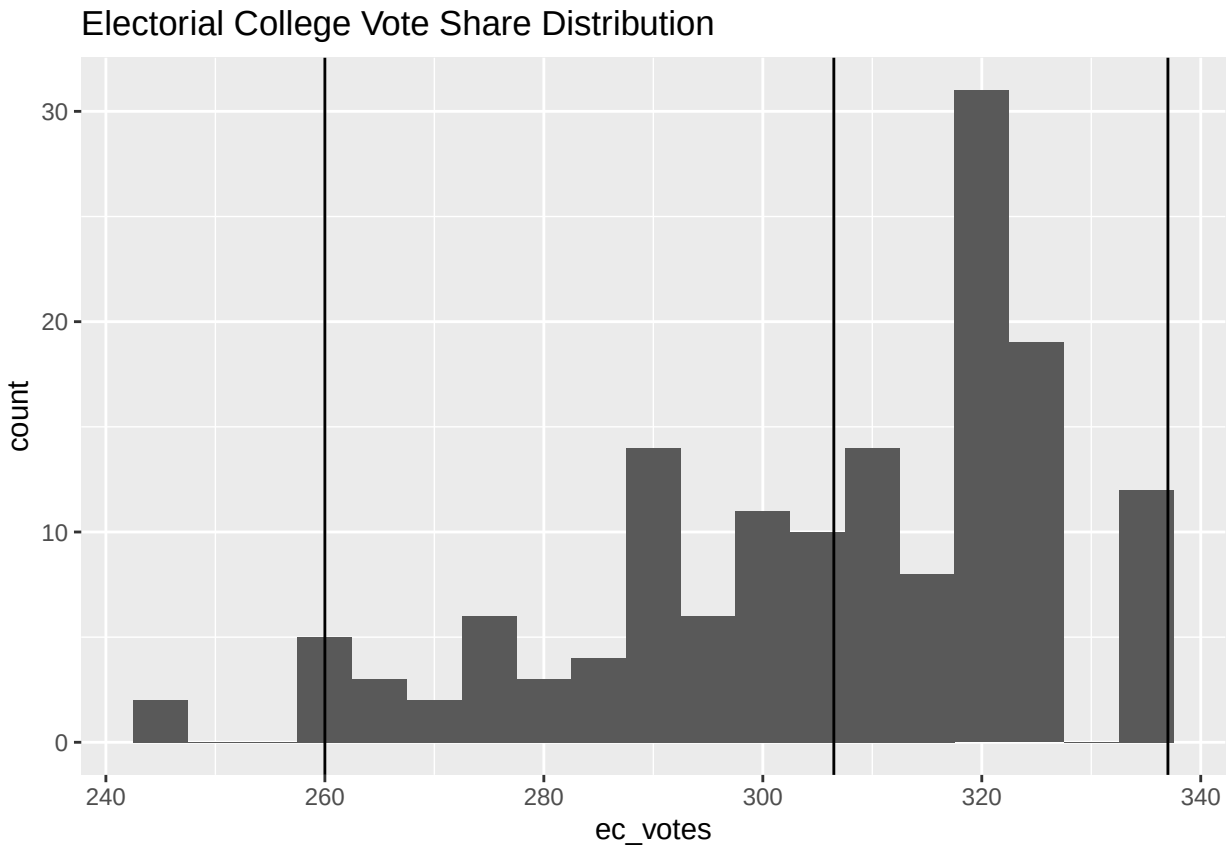


Table 1: EC Vote Total

	x
2.5%	260
97.5%	337

Table 2: Swing State Winning Probability

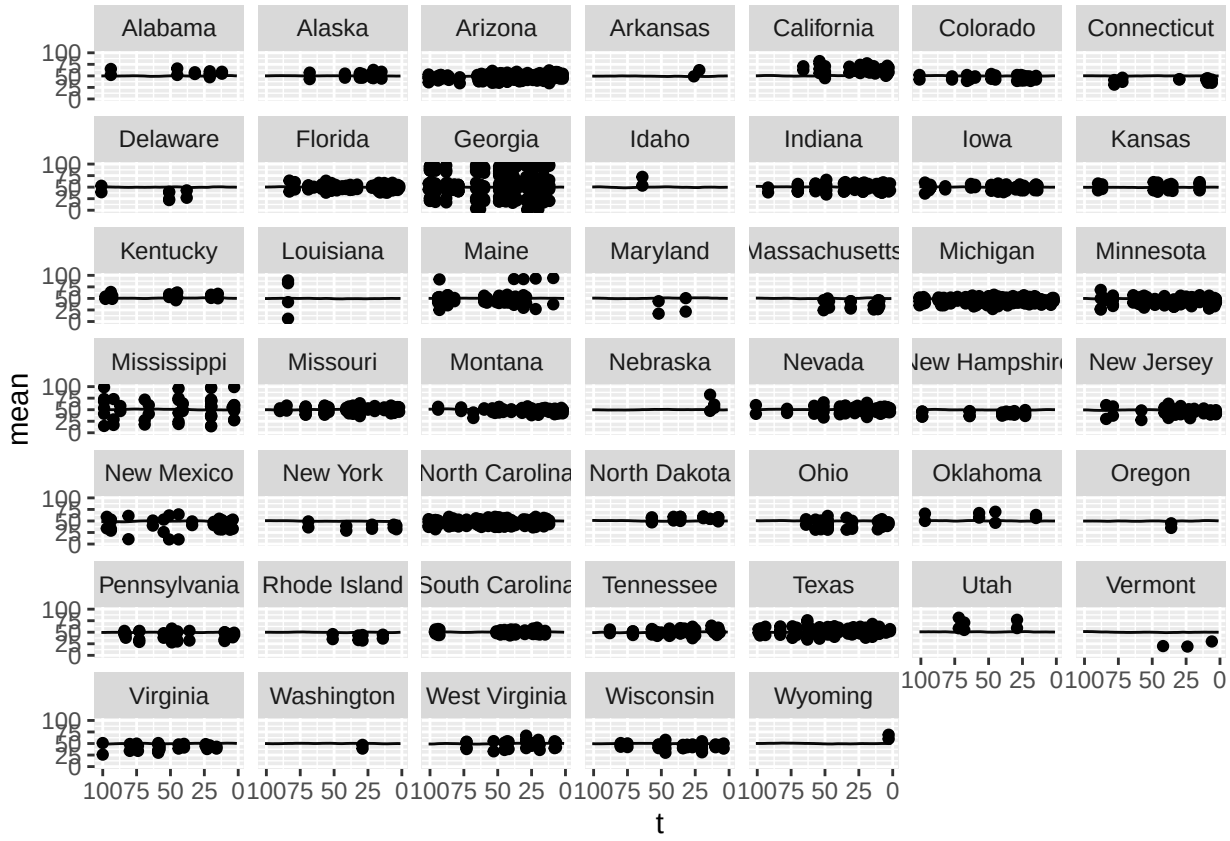
	Winning Probability
WI	0.5733333
PA	0.7800000
US	0.6933333
FL	0.5200000
NC	0.6866667
VA	0.9466667
OH	0.6066667

Table 3: Senator All State Vote Share Percentage Interval Estimate

	2.5%	50%	97.5%
Texas	49.48865	49.93430	50.47043
Michigan	48.70744	49.71781	50.28788

	2.5%	50%	97.5%
Alabama	49.36794	49.94929	50.63500
Georgia	49.41129	49.93638	50.57295
North Carolina	49.38202	49.83676	50.25965
Tennessee	49.60345	50.06526	50.78474
Arizona	49.01161	49.74505	50.35547
Iowa	49.05625	49.92159	50.56213
Nebraska	49.36209	49.83959	50.39243
Massachusetts	49.28333	49.81077	50.31590
Kansas	49.03423	49.79416	50.37000
Alaska	49.19220	49.72012	50.22294
Kentucky	49.52827	50.02476	50.57869
Montana	49.33858	49.82322	50.35748
Minnesota	49.33862	50.04631	50.61206
Oklahoma	49.29089	49.89460	50.36911
Colorado	49.29429	50.05029	50.87112
South Carolina	49.41807	49.88414	50.40613
Virginia	49.67020	50.19278	50.60253
New Mexico	49.33469	49.93196	50.51173
Arkansas	49.30187	49.97594	50.54286
New Jersey	49.29358	49.85576	50.28157
New Hampshire	49.37435	49.97463	50.66964
Maine	49.52431	49.94446	50.46235
West Virginia	49.49968	50.04234	50.75991
Oregon	49.50307	50.09993	50.76482
Delaware	49.11628	49.81549	50.54978
Idaho	49.15113	49.73873	50.29138
Mississippi	49.37968	50.03097	50.66838
Louisiana	49.13613	49.99302	50.54872
Florida	49.29684	49.87618	50.62078
Nevada	49.25424	49.91041	50.82405
Indiana	49.45093	49.98615	50.36335
Missouri	49.48347	50.06575	50.81421
Pennsylvania	49.38802	49.86989	50.30653
Wyoming	49.41795	50.13526	50.84998
California	49.57884	50.03115	50.70710
Ohio	49.38909	50.00994	50.47344
New York	49.23308	49.81312	50.24726
Wisconsin	49.43468	49.83374	50.40279
Connecticut	49.12114	49.79219	50.41237
Vermont	49.41725	49.96039	50.51073
North Dakota	49.46003	50.01059	50.68093
Rhode Island	49.46959	50.00278	50.46348
Washington	49.18311	49.92345	50.50333
Utah	49.69696	50.23810	50.82617
Maryland	49.38380	50.00682	50.61093

```
## Warning: Removed 4104 rows containing missing values (geom_point).
```


Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
Florida
Georgia
Idaho
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming
mean
t
100
75
50
25
0

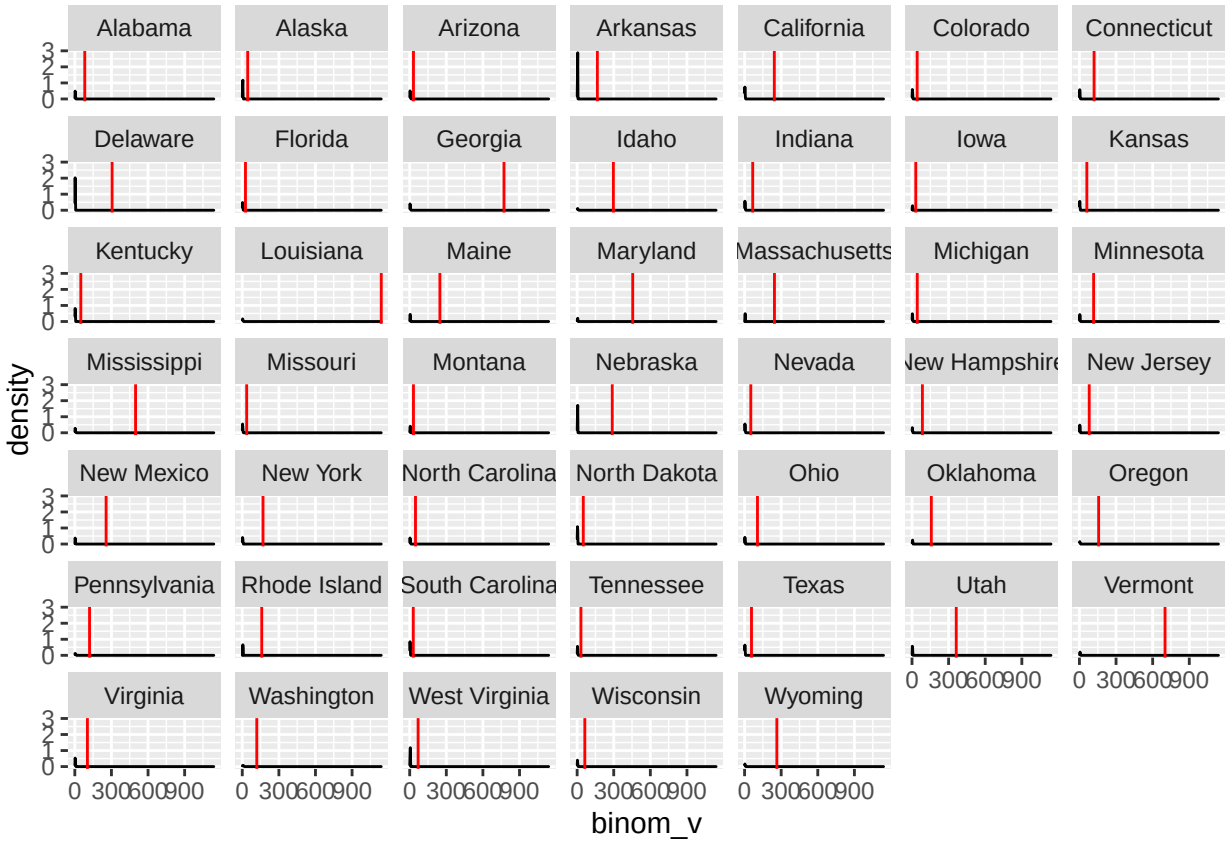


Table 4: Senator All State Winning Probability

	Winning Probability
Texas	0.4133333
Michigan	0.1800000
Alabama	0.4400000
Georgia	0.3800000
North Carolina	0.2000000
Tennessee	0.6333333
Arizona	0.2200000
Iowa	0.4066667
Nebraska	0.3133333
Massachusetts	0.2400000
Kansas	0.2333333
Alaska	0.1733333
Kentucky	0.5666667
Montana	0.2600000
Minnesota	0.5666667
Oklahoma	0.3466667
Colorado	0.5733333
South Carolina	0.3600000
Virginia	0.7266667
New Mexico	0.4266667
Arkansas	0.4733333
New Jersey	0.2800000
New Hampshire	0.4466667

	Winning Probability
Maine	0.4333333
West Virginia	0.5333333
Oregon	0.6400000
Delaware	0.3666667
Idaho	0.2066667
Mississippi	0.5333333
Louisiana	0.5000000
Florida	0.3466667
Nevada	0.3333333
Indiana	0.4533333
Missouri	0.5866667
Pennsylvania	0.2666667
Wyoming	0.5933333
California	0.5200000
Ohio	0.5200000
New York	0.2466667
Wisconsin	0.2733333
Connecticut	0.2800000
Vermont	0.4400000
North Dakota	0.5200000
Rhode Island	0.5000000
Washington	0.3666667
Utah	0.8000000
Maryland	0.5133333

Table 5: Senator All State Vote Share Percentage Interval Estimate

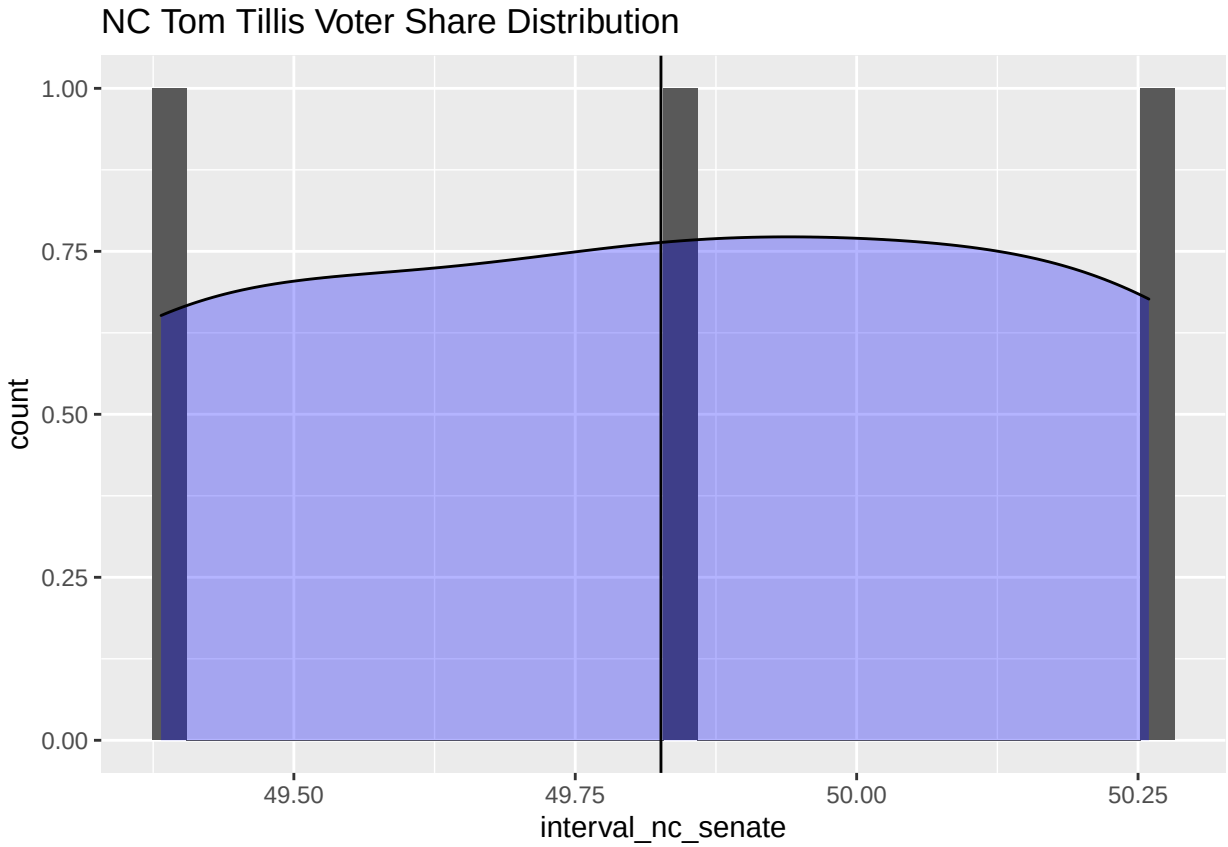
	2.5%	50%	97.5%
Texas	49.48865	49.93430	50.47043
Michigan	48.70744	49.71781	50.28788
Alabama	49.36794	49.94929	50.63500
Georgia	49.41129	49.93638	50.57295
North Carolina	49.38202	49.83676	50.25965
Tennessee	49.60345	50.06526	50.78474
Arizona	49.01161	49.74505	50.35547
Iowa	49.05625	49.92159	50.56213
Nebraska	49.36209	49.83959	50.39243
Massachusetts	49.28333	49.81077	50.31590
Kansas	49.03423	49.79416	50.37000
Alaska	49.19220	49.72012	50.22294
Kentucky	49.52827	50.02476	50.57869
Montana	49.33858	49.82322	50.35748
Minnesota	49.33862	50.04631	50.61206
Oklahoma	49.29089	49.89460	50.36911
Colorado	49.29429	50.05029	50.87112
South Carolina	49.41807	49.88414	50.40613
Virginia	49.67020	50.19278	50.60253
New Mexico	49.33469	49.93196	50.51173
Arkansas	49.30187	49.97594	50.54286
New Jersey	49.29358	49.85576	50.28157
New Hampshire	49.37435	49.97463	50.66964

	2.5%	50%	97.5%
Maine	49.52431	49.94446	50.46235
West Virginia	49.49968	50.04234	50.75991
Oregon	49.50307	50.09993	50.76482
Delaware	49.11628	49.81549	50.54978
Idaho	49.15113	49.73873	50.29138
Mississippi	49.37968	50.03097	50.66838
Louisiana	49.13613	49.99302	50.54872
Florida	49.29684	49.87618	50.62078
Nevada	49.25424	49.91041	50.82405
Indiana	49.45093	49.98615	50.36335
Missouri	49.48347	50.06575	50.81421
Pennsylvania	49.38802	49.86989	50.30653
Wyoming	49.41795	50.13526	50.84998
California	49.57884	50.03115	50.70710
Ohio	49.38909	50.00994	50.47344
New York	49.23308	49.81312	50.24726
Wisconsin	49.43468	49.83374	50.40279
Connecticut	49.12114	49.79219	50.41237
Vermont	49.41725	49.96039	50.51073
North Dakota	49.46003	50.01059	50.68093
Rhode Island	49.46959	50.00278	50.46348
Washington	49.18311	49.92345	50.50333
Utah	49.69696	50.23810	50.82617
Maryland	49.38380	50.00682	50.61093

Table 6: North Carolina Vote Share Interval Estimate

	x
2.5%	49.38202
50%	49.83676
97.5%	50.25965

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


NC Tom Tillis Voter Share Distribution
1.00
0.75
0.50
0.25
0.00
count
49.50
49.75
50.00
50.25
interval_nc_senate

```
## [1] 0.2
```

```
## [1] 0.16
```

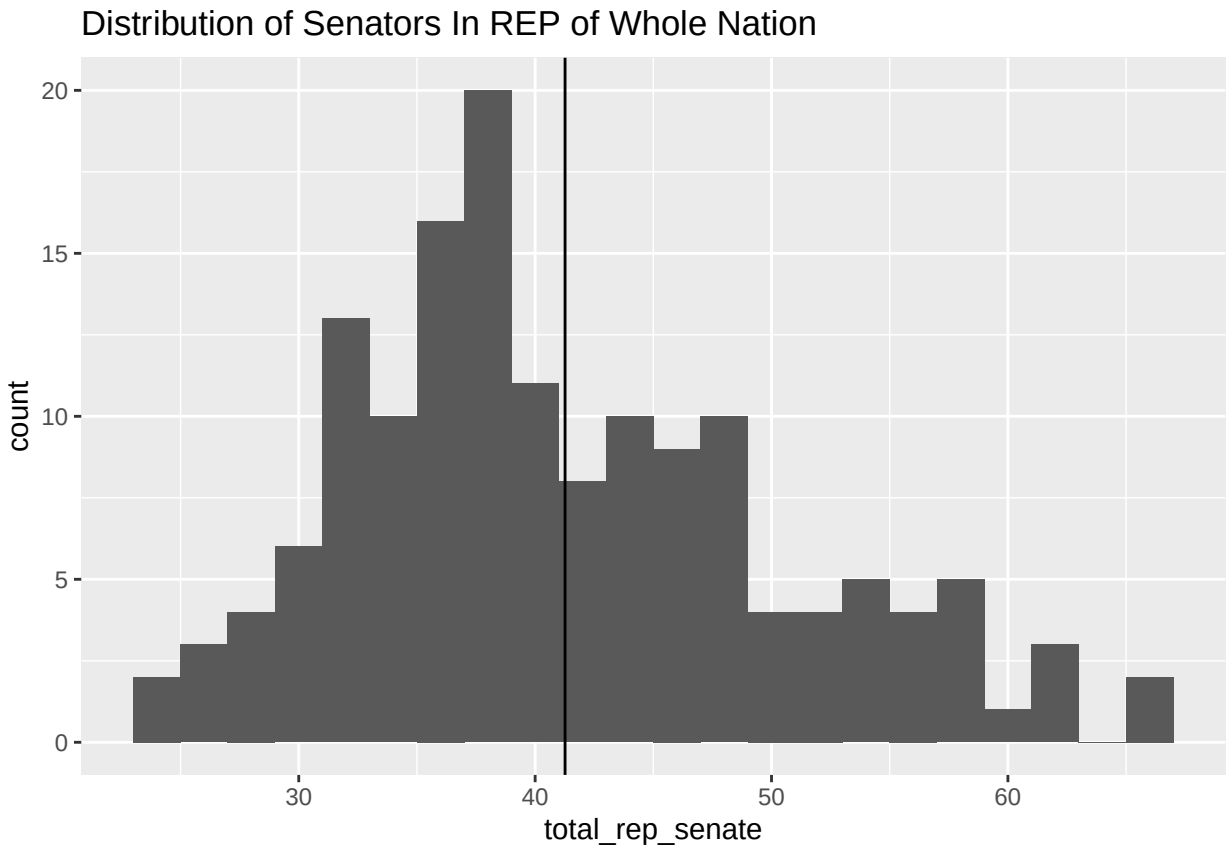


Table 7: House All Congressional District Winning Probability (REP)

	Winning Probability
District 11	0.2400000
District 8	0.6266667
District 9	0.1800000
District 3	0.3400000
District 13	0.4400000
District 2	0.6066667
District 7	0.3600000
district 1	0.5000000
district 4	0.2600000
district 5	0.8066667
district 6	0.0400000
district 10	0.4866667
district 12	0.2733333

Table 8: house All State Vote Share Percentage Interval Estimate (REP)

	2.5%	50%	97.5%
District 11	48.92757	49.72709	50.53072
District 8	49.04345	50.14734	51.21412

	2.5%	50%	97.5%
District 9	47.92971	49.61795	50.39143
District 3	47.99315	49.66380	51.14133
District 13	48.55711	49.89879	50.93699
District 2	49.14999	50.11197	52.17065
District 7	48.31677	49.68463	50.93979
district 1	48.72039	50.00610	51.44350
district 4	48.77145	49.70914	50.41983
district 5	49.36884	50.55082	53.05599
district 6	47.56290	48.81924	50.05846
district 10	48.58524	49.98259	51.04029
district 12	48.83427	49.68021	50.47296