fec16 Exploratory Data Analysis

Richard Robbins

March 24, 2022

Contents

1 Introduction

The w203 Unit 10 homework assignment is based on Federal Election Commission 2016 data available in the `fec16` R package. As I worked on the assignment I noticed several things that piqued my curiosity. As a result, after completing the assignment I reviewed the data more closely. This report describes my most meaningful findings.

I focus only on select data elements from the `fec16::results_house` dataset that pertain to the general election, as opposed to primary or runoff elections. I also consider data from the `fec16::campaigns` dataset, but only to a limited degree.

The Federal Election Commission 2016 Federal Election report (the “Report”) appears at https://www.fec.gov/resources/cms-content/documents/federalelections2016.pdf. The Report includes explanatory notes that inform this analysis.

Table 1: California 1st District Winner

candidate	party	votes	percent
H2CA02142	REP	185,448	0.59

Table 2: New York 1st District Winner

candidate	party	votes	percent
H8NY01148	REP	158,409	0.49
H8NY01148	CRV	23,327	0.07
H8NY01148	REF	843	0.00
H8NY01148	IDP	5,920	0.02

```r
campaigns <- fec16::campaigns
results_house <- fec16::results_house
```

2 Splintered Vote Reporting and Party Affiliation

2.1 Splintered Vote Reporting

In some cases, the `results_house` data includes several rows that represent the votes received by a candidate. I refer to this phenomenon as "splintered vote reporting". When the `results_house` dataset is used to analyze the total votes received by candidates, splintered vote reporting must be taken into account. The FEC touches on this topic in the explanatory notes to the Report.

> "Combined Parties" represents all the valid votes cast for one candidate, regardless of party. (This method is used where a candidate may be listed on the ballot more than once, with different party designations, i.e., in Connecticut, New York and South Carolina.) These votes are then broken down and listed by party.

In the simplest case, a candidate appears on the ballot once and the relevant data is reflected in a single row. Consider, for example, Republican Doug LaMalfa. He won the election to represent California's 1st district. He was elected with 185,448 votes and 59% of the vote. Table 1 shows how his vote data is reflected in `results_house`. On the other hand, where a candidate is listed on the ballot more than once, the `results_house` dataset includes one row for each party designated for the candidate. For example, consider Lee Zeldin, a Republican, who won the race to represent the New York's 1st district. He was elected with 188,499 votes and 58% of the vote. Table 2 shows how his vote data is reflected in `results_house`

The `results_house` dataset includes 1,291 rows with candidate vote data. If we take splintered vote reporting into account, we have 1,190 rows. If we count the number of winning candidates in `results_house` without considering splintered vote reporting, we get 493 which is clearly wrong as the House of Representatives in 2016 (and now) how 435 members plus six additional non-voting members.

2.2 Party Affiliation

As described above, a single candidate can appear in voting reports with multiple party affiliations. The `results_house` dataset reflects 84 separate parties. If we focus only on candidates participating in the general election, we get 76 parties. Finally, if we look only at the winners, we get 15 parties.

Table 3: Party Results Cross Check

party	derived	true
Republican	237	242
Democrat	196	197
Other	6	2
Total	439	441

The 435 voting members of the 2016 House of Representatives consisted of 241 Republicans and 194 Democrats. The six non-voting delegates included three Democrats (District of Columbia, Guam and the United States Virgin Islands), one Republican (American Samoa), one Independent (Northern Mariana Islands) and one member of the New Progressive Party (Puerto Rico). See, the Report.

The difficulty is that the `results_house` dataset does not make it easy to identify with certainty the principal party affiliation of the candidates. Other `fec16` datasets are well suited to this purpose. The `campaigns` dataset does not contain duplicate entries for candidates and it includes two relevant variables, `cand_pty_affiliation` and `pty_cd`.

The `campaigns` dataset includes 1,898 candidates and no duplicate candidate identification numbers. The `cand_pty_affiliation` variable includes 924 Republicans, 743 Democrats and 231 people who are neither Republicans nor Democrats. The `pty_cd` numeric variable is consistent with `cand_pty_affiliation`, where 1 denotes a Democrat, 2 a Republican, and 3 people who are neither Republicans or Democrats.

2.3 Correcting For Splintered Vote Reporting and Identifying Principal Party Affiliation

We can address the splintered vote reporting problem by grouping `results_house` information by candidate identification ID and deriving the total general election votes and the total general election percentages from the rows in the group. We can take the resulting dataset and add party affiliation by joining the `campaigns` dataset on the `cand_id` variable and utilizing either `cand_ptuy_affiliation` or `pty_cd`. In the following example, we remove rows that do not contain general election voting information and retain other variables of interest, including `state`, `district_id`, `incumbent` and `won`.

```
df <- results_house %>%
  drop_na(general_votes) %>%
  group_by(cand_id, state, district_id, incumbent, won) %>%
  summarize(total_general_vote = sum(general_votes),
            total_general_percent = sum(general_percent),
            .groups = "keep") %>%
  ungroup()

df <- inner_join(df, campaigns %>% select (cand_id, pty_cd), by="cand_id")

df <- df %>%
  mutate(party = case_when(pty_cd == 1 ~ "Democrat",
                           pty_cd == 2 ~ "Republican",
                           pty_cd == 3 ~ "Other"))
```

Table 3 compares the breakdown of winning candidates and their party affiliations as reflected in the data frame assembled with the immediately preceding code with the true results. The numbers are getting close but it appears that there are still some flaws. We explore those in greater detail below.

2.4 Voting Percentage Distributions

Before we fine tune our work, let's take a step back and consider the impact of properly accounting for splintered vote recording. When a candidates vote total is splintered, we end up with too many data points and the nunber of votes and voting percentage with each is too low. The primary implication of this is that we would expect to see an incorrect concentration of lower percentage votes in a distribution of voting percentages and a corresponding aggregate descrease in voting percentages that should be reflected at higher levels. This shift should also be reflected in the mean.

The pair of histograms in Figure 1 demonstrates the anticipated effects. The upper histogram in the panel shows the distribution of voting percentages without correction for splintered voting. The bottom histogram in the panel shows the distribution of voting percentages after the correction for splintered voting. The most prominent feature of the upper histogram is a pronounced density mass on the left hand side that grows as the percentages on the x axis drop. After the correction is applied, the pronounced mass is dramatically diminished and the acceleration as voting percentages drop is disippated. The mean of the percentages reflected in the upper histogram is .342 whereas the mean of the percentages increases to .509 after the correction is applied.

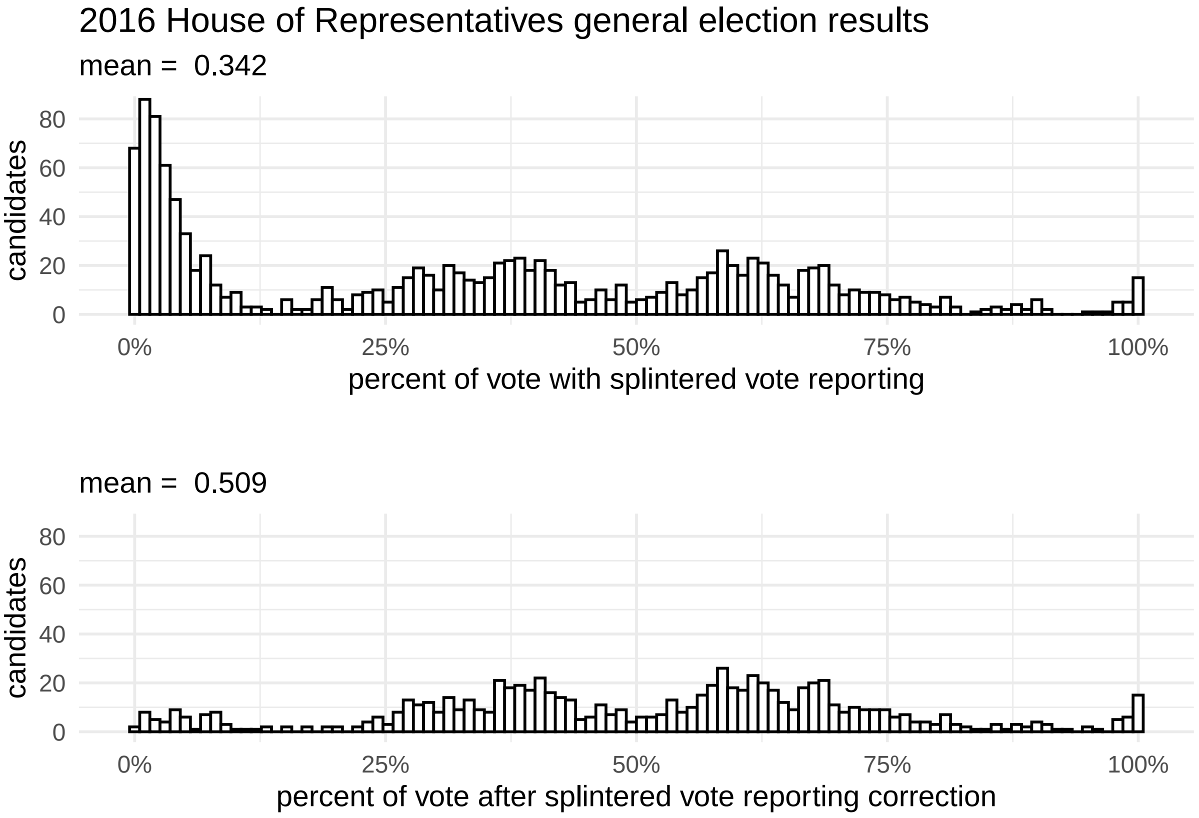

Figure 1: Voting Percentage Distribution

3 Elections for Unexpired Terms

In the ordinary course, an election to the House of Representatives is for a full term that commences when the term of the incumbent seat holder expires. However, frome time to time a vacancy occurs before the expiration of the incumbent's term. Some states have separate elections for unexpired terms. In some cases,

Table 4: Unexpired Term Contests

state	district_id	count
HI	01 - UNEXPIRED TERM	10
KY	01 - UNEXPIRED TERM	1
KY	1 - UNEXPIRED TERM	1
PA	02 - UNEXPIRED TERM	2

results for unexpired term elections are commingled in the `results_house` data together with full term elections. The FEC references this in the first paragraph of the body of the Report (with emphasis added):

> This publication has been prepared by the Federal Election Commission to provide the public with the results of elections held in the fifty states during 2016 for the offices of United States President, United States Senator and United States Representative. Also included are the results for Delegate to Congress from American Samoa, the District of Columbia, Guam, the Northern Mariana Islands, the U.S. Virgin Islands and Resident Commissioner for Puerto Rico. **Additionally, there are results for the special elections to fill the unexpired terms in Hawaii's 1st Congressional District, Kentucky's 1st Congressional District and Pennsylvania's 2nd Congressional District.** The Commission undertakes this project on a biennial basis in order to respond to public inquiries.

Table 4 lists the unexpired term races in `results_house` together with the number of observations for each. Note that there are inconsistent references to the contest relating to the unexpired term for the 1st District of Kentucky. For flexibility, I will add a new variable to the dataframe to indicate whether a row in the dataset pertains to an election for an unexpired term.

4 Uncontested Elections

5 Non-Voting Delegates

6 Write In Ballots

7 Small Defects