Election Fraud in Russia

Nicholas Archambault

18 April 2021

Introduction

Benford's law, or the law of anomalous numbers, is among the most fascinating of statistical phenomena. The principle maintains that the leading digits of naturally occurring collections of numbers are more likely to be small than large, a truth that has been demonstrated in a wide variety of randomized data sets, including bills, street addresses, stock prices, population totals, river lengths, and physical constants.

The law is named after physicist Frank Benford, who accrued for analysis over 20,000 numbers from magazine articles, atomic weights, lake drainage rates, and a welter of other sources. The naturally-arising patterning underlying Benford's law was first noticed in 1881 by astronomer Simon Newcomb when he realized that tables found in logarithm reference books with smaller leading digits values were more worn and handled than tables with larger leading digits values.

A set of numbers is said to satisfy Benford's law if leading digit $d \in \{1, ..., 9\}$ occurs with the probability

$$P(d) = log_{10}\left(1 + \frac{1}{d}\right)$$

.

This leads to the following probability proportions:

d	$P(d)$
1	30.1%
2	17.6%
3	12.5%
4	9.7%
5	7.9%
6	6.7%
7	5.8%
8	5.1%
9	4.6%

The probability $P(d)$ is proportional to the space between d and $d + 1$ on a logarithmic scale; Benford's distribution is thus expected when the *logarithms* of the numbers which comprise the data — not the numbers themselves — are uniformly and randomly distributed. The distribution of the data should span several orders of magnitude in order to allow Benfordian behavior to emerge. Collections of data such as national areas or populations provide suitable distributional widths. Vatican City, for example, is 0.4 km^2 in area while Russia is 1.7×10^7 km^2; the Pitcairn Islands have 50 inhabitants while China has 1.3×10^9.

Benford's law and the analysis of irregular distributions of leading, trailing, or other digits within vote totals has been used as a simple method for identifying election fraud in political processes across the globe. This project will examine voting patterns in the 2011 Russian election of the State Duma, or the federal legislature

of Russia. The ruling United Russia party ultimately won the election, which was tainted by accusations of fraud. Demonstrators assembled to protest the electoral results on the basis of irregular digit distribution that violated Benford's law. The Kremlin denied all accusations of election interference.

Variables and their descriptions can be found in the following table.

Variable	Description
N	total number of voters in a precinct
turnout	total turnout in a precinct
votes	total number of votes for the winner in a precinct

1. Conspicuous Fractions

In disputing the official results of the 2011 Russian election, protesters cited as their main argument the conspicuously high frequency of common fractions such as 1/4, 1/3, and 1/2 in the vote share total obtained by United Russia in precincts across the nation.

We begin to analyze the election results by computing United Russia's vote share — votes divided by turnout — and identifying the 10 most frequently occurring fractions the party earned across Russian precincts. Plotting a histogram of the proportion of each fraction with percentile breaks will allow us to differentiate between similar fractions, such as 1/2 and 51/100, in order to determine whether the protesters' claims are valid.

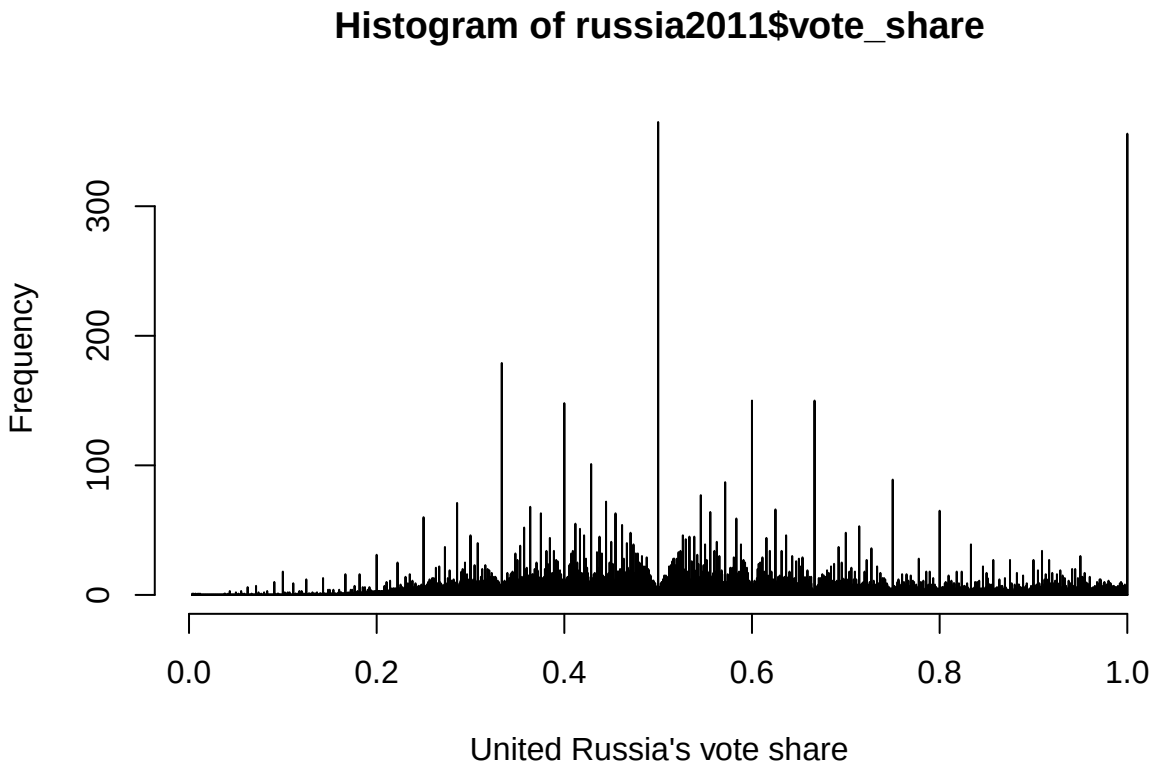
```
# Set seed
set.seed(12)

# Load packages and data
library(MASS)
library(beepr)
load("fraud.RData")

# Calculate United Russia's vote share
russia2011$vote_share <- russia2011$votes / russia2011$turnout
sum(russia2011$votes) / sum(russia2011$turnout)
```

```
## [1] 0.5009214
```

```
# Get vote share count
vote_share_count <- table(russia2011$vote_share)

# Print 10 most common fraction
sort(vote_share_count, decreasing = TRUE)[1:10]
```

```
##
##               0.5                 1 0.333333333333333               0.6
##               365               356               179               150
## 0.666666666666667               0.4 0.428571428571429              0.75
##               150               148               101                89
## 0.571428571428571 0.545454545454545
##                87                77
```

```
# Plot
hist(russia2011$vote_share, breaks = length(vote_share_count),
     xlab = "United Russia's vote share")
```



United Russia's total vote share earned was just barely over the threshold for a majority, at 50.09%. The ten most common vote share fractions include many common fractions with small numerators and denominators, such as 1/2, 1, 1/3, 2/3, 2/5, and 3/4.

The histogram of most common vote shares won by United Russia across precincts confirms that fractions with smaller numerators and denominators are conspicuously frequent. This is a red flag, and we should be suspicious of election fraud since a human fraudulently inputting electoral results would be likely to unwittingly over-represent common fractions. A free and fair election would result in more randomized proportions among precincts; proportions would not be so neat and tidy, clustering specifically around convenient and common values.

2. Monte Carlo Simulation

Though the previous analysis raises questions about the cleanness of the 2011 Russian election, the mere existence of high frequencies at low fractions does not imply election fraud. We can conduct a Monte Carlo simulation under the alternative assumption that vote share is assumed to follow a binomial distribution in order to investigate the possibility that the high frequencies of low fractions arose purely by chance.

The binomial distribution is a generalization of the Bernoulli distribution, which maintains that a random variable can take on one of two distinct values. As opposed to the Bernoulli distribution, where the two-valued random variable is tested only once — think of flipping a coin — the binomial distribution represents the number of times the random variable attains each value across multiple independent trials. In other words, the binomial distribution models the number of times a coin lands on heads across a fixed number of coin flips.

The binomial random variable X records the number of successes out of n independent and identical trials, with success probability p. This can be interpreted as the number of times X takes on the specific value x and modeled with the equation

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

.

The expression $\binom{n}{x}$ is read as "of n, choose x," and its full form is $\frac{n!}{(n-x)!x!}$.

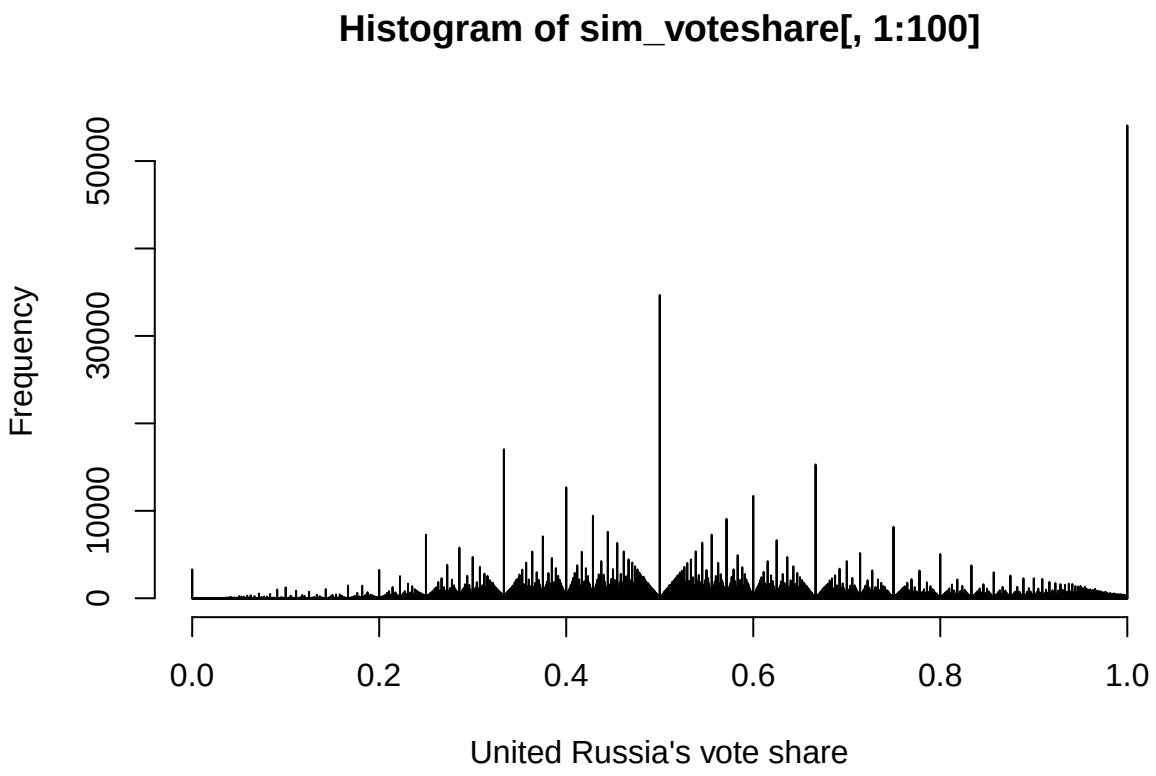
To conduct the simulation, we assume turnout for a precinct follows a binomial distribution where size n equals the total number of voters and probability p equals the turnout rate for the precinct. The vote share for United Russia in the precinct is also assumed to follow a binomial distribution where size n equals the number of voters who turned out and probability p equals the observed United Russia vote share.

We will conduct 1,000 simulated trials and construct a histogram of the fractions observed in the simulated results.

```
# Calculate turnout rate
russia2011$turnout_rate <- russia2011$turnout / russia2011$N

# Assign number of simulations
sims <- 1000

# Initialize results matrix
results <- matrix(NA, ncol = sims, nrow = nrow(russia2011))

mc_simulation <- function(dataset) {
  sim_turnout <- rbinom(nrow(dataset),
                        size = dataset$N,
                        prob = dataset$turnout_rate)
  sim_votes <- rbinom(nrow(dataset),
                      size = sim_turnout,
                      prob = dataset$vote_share)
  sim_votes / ifelse(sim_turnout == 0, 1, sim_turnout)
}

# Reset seed
set.seed(12)

# Generate simulation (without 'for' loop)
sim_voteshare <- replicate(sims, mc_simulation(russia2011))
sim_vote_share_count <- table(sim_voteshare)

# Sort by count; print 10 most common fractions
sort(sim_vote_share_count, decreasing = TRUE)[1:10]
```

```
## sim_voteshare
##                1              0.5 0.333333333333333 0.666666666666667
##           538366           348609           171917           153763
##              0.4              0.6 0.428571428571429 0.571428571428571
##           126505           117894            96288            90394
##             0.75 0.444444444444444
##            80807            76134
```

```
# Plot
hist(sim_voteshare[, 1:100], breaks = length(vote_share_count),
     xlab = "United Russia's vote share")
```



Common fractions still constitute the top ten most frequently occurring vote share values. Just as before, we find high frequencies of fractions with low numerators and denominators, such as 1, 1/2, 1/3, 2/3, 2/5 and 3/5. Though they don't follow the same order of frequency, nine of the ten most common values in this simulated analysis also appear in the previous section's analysis of the official election data.

Based on the frequency of the fraction 1, it seems that the Monte Carlo simulation expects a 100% vote share for United Russia far more often than it actually occurred in the official data. We also note that other fractions appear at different frequencies between the two sets of results: 2/5 and 3/5, for example, seem to be less common in the Monte Carlo results than in the official.

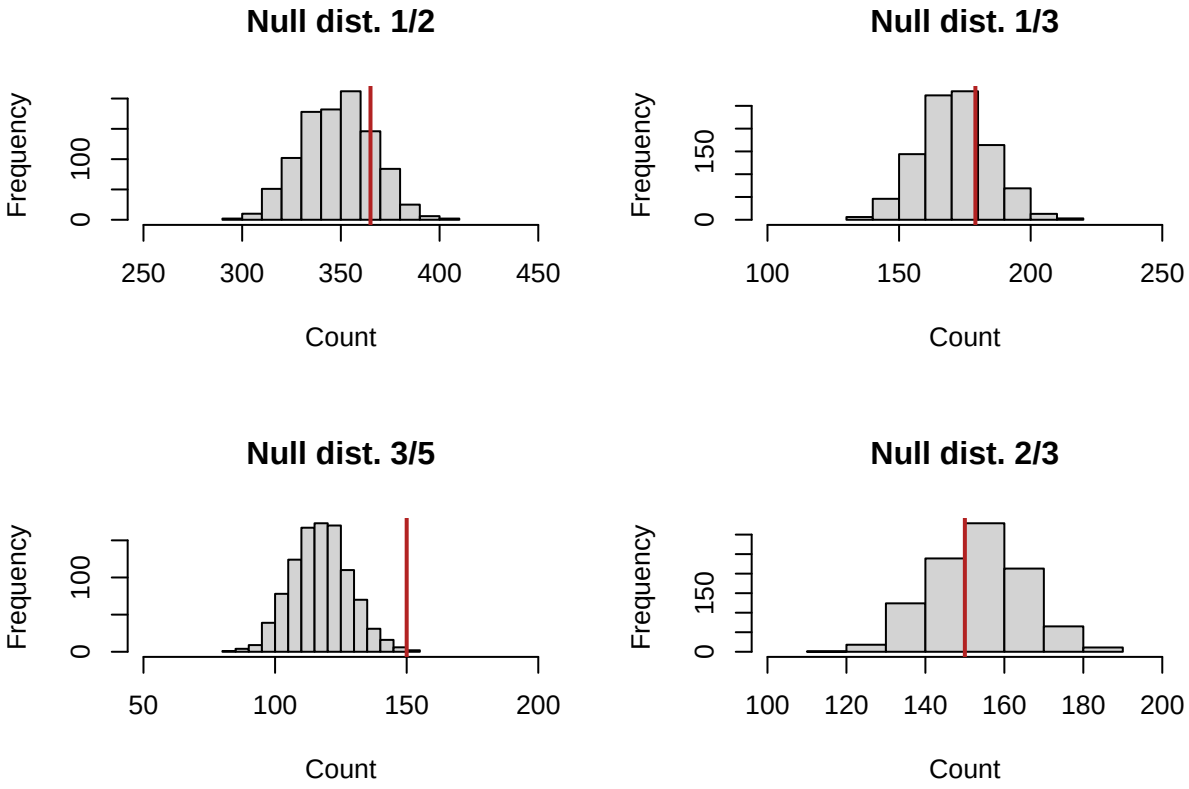
This analysis does not conclusively establish that electoral fraud was committed during the 2011 Russian election; in fact, it pushes back on that notion. The following sections will provide more insight into whether fraud actually occurred.

3. Comparing Real and Simulated Distributions

To more comprehensively judge the Monte Carlo simulation results against the official results, we compare the observed fraction of vote shares within a bin of a certain size with its simulated counterpart. We will generate histograms of simulated frequency distributions for four of the most commonly occurring fractions — 1/2, 1/3, 3/5, and 2/3 — and overlay the frequencies of those fractions in the official election data to determine how conspicuous those official frequencies were.

```
# Function to assess the number of observed fractions
fraction_dist <- function(fraction, range) {
  number <- sum(russia2011$vote_share == fraction)
  simulated_number <- apply(sim_voteshare == fraction, 2, sum)
  hist(simulated_number, xlim = range, xlab = "Count", breaks = 10,
       main = paste0("Null dist. ", as.character(fractions(fraction))))
  abline(v = number, col = "firebrick", lwd = 2)
}
```

```r
par(mfrow = c(2, 2))
fraction_dist(1/2, c(250, 450))
fraction_dist(1/3, c(100, 250))
fraction_dist(3/5, c(50, 200))
fraction_dist(2/3, c(100, 200))
```



The above plots show the counts of precincts for which top four most common fractions appear in the 1,000-trial Monte Carlo simulated data. These simulations constitute the *null distribution* for each fraction, the distribution we assume appearance frequencies of that fraction would follow if there were no election fraud. The true mean — the frequency with which those fractions were *actually* observed in the official election data — is shown with the solid red line.

If the 2011 Russian election were completely free and fair, we would expect to see nearly identical means of these four fractions in simulated and official data. The observed number of precincts reporting fractions 1/3 and 2/3 are indeed similar to their respective null distributions. The number of precincts reporting fraction 3/5, however, is at the tail of its null distribution, raising suspicions of fraud. We would not expect to observe so many precincts reporting 3/5 if the null hypothesis of no election fraud were true.

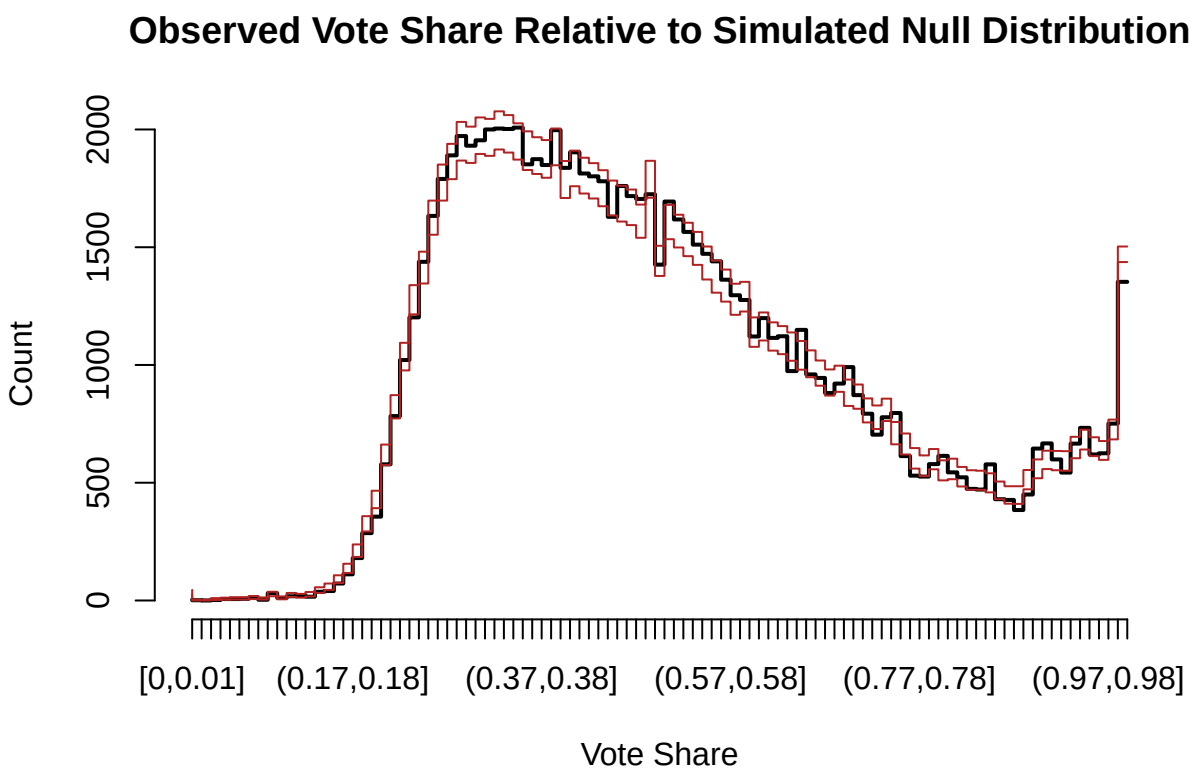
We are, however, testing only a handful fractions, and one of them is bound to look different from the null distribution by pure chance, even if the null *is* true. The more disciplined approach is to test all fractions at the same time.

4. Identifying Suspicious Percentile Bins

To conduct more rigorous analysis, we will compare the relative frequencies of observed fractions with their simulated frequencies beyond the four examined in the previous section. By choosing a bin size of 0.01, we divide the distribution into 100 constituent fractions and compute the proportion of all vote share observations for which each fraction accounts. We can then examine whether or not the proportion for a fraction observed in the official election data falls within a 95% confidence interval of simulated proportions of that same fraction.

If a high number of observed proportions for the 100 different fractions falls outside the corresponding range of simulated proportions, we can be reasonably confident that fraud occurred.

```
# Sequence of percentile bins
bins <- seq(0, 1, 0.01)

# Function to get bin counts
bin_counter <- function(votes) {table(cut(votes, bins, include.lowest = TRUE))}

# Get observed bin count
obs_bin_count <- bin_counter(russia2011$vote_share)

# Simulated bin count
# Rows are bins, columns are simulation round
sim_bin_count <- apply(sim_voteshare, 2, bin_counter)

# Get 2.5 and 97.5 percentiles for each bin over simulation rounds
sim025 <- apply(sim_bin_count, 1, quantile, probs = 0.025)
sim975 <- apply(sim_bin_count, 1, quantile, probs = 0.975)

# Plot
plot(obs_bin_count, xlab = "Vote Share", ylab = "Count", type = "S",
     main = "Observed Vote Share Relative to Simulated Null Distribution")
lines(sim025, col = "firebrick", type = "S")
lines(sim975, col = "firebrick", type = "S")
```




```
# Count number of times observed count is more extreme than 2.5 and 97.5 percentiles
sum(obs_bin_count < sim025 | obs_bin_count > sim975)
```

```
## [1] 30
```

The observed counts seem to follow the simulated null distribution reasonably well, but we identify some outliers. Notably, in 34 of the 100 bins the proportion of observed vote shares falls outside the 2.5 and 97.5 percentiles of the null distribution for that bin. That number is conspicuously high — just by chance, we would expect around five bins to fall outside their counterparts' middle 95% range. A nearly seven-fold increase in the number of outlying bins over our expectation indicates that the null hypothesis is not true, perhaps signalling fraud.

It is possible that the middle 95% of each bin does not truly capture 95% of the data. In the following section, we will apply these same analysis techniques to different election data to see whether the high number of outlying bins persists.

5. Testing Other Elections for Fraud

To put the results of the previous section into context and determine whether 34 outlying bins is conspicuously high, we can run the same bin-comparison procedure for other election results and evaluate the number of outlying bins identified.

The 2003 Russian and 2011 Canadian elections featured no major voting irregularities. In contrast, the 2012 Russian presidential election was marred by allegations of fraud. We can collect the previous steps of analysis into a function and obtain outlying percentile results for each of these elections to determine whether such analysis produces meaningful indication of fraud.

```
set.seed(12)

# Combine previous steps of percentile bin simulation into function
test_fraud <- function(dataset) {
  dataset$vote_share <- dataset$votes / dataset$turnout
  dataset$turnout_rate <- dataset$turnout / dataset$N
  sim_voteshare <- replicate(sims, mc_simulation(dataset))
  obs_bin_count <- bin_counter(dataset$vote_share)
  sim_bin_count <- apply(sim_voteshare, 2, bin_counter)
  sim025 <- apply(sim_bin_count, 1, quantile, probs = 0.025)
  sim975 <- apply(sim_bin_count, 1, quantile, probs = 0.975)
  sum(obs_bin_count < sim025 | obs_bin_count > sim975)
}

test_fraud(canada2011)
```

```
## [1] 20
```

```
test_fraud(russia2003)
```

```
## [1] 22
```

```
test_fraud(russia2012)
```

```
## [1] 22
```

We previously found that 34 of the 100 percentile bins for the 2011 Russian election fall outside the middle 95% of their null distributions. There were no reports or allegations of fraud in either the 2003 Russian election or the 2011 Canadian election, so we would assume that roughly 5 of the 100 percentile bins for those elections fall outside the middle 95% of their null distributions. Instead, we find that 20 bins display such behavior. Furthermore, our simulation process yields the same number of outlying bins for the 2003 and 2012 Russian elections — one of which reported no voting irregularities, while the other has been empirically confirmed as fraudulent (see *Rozenas 2017*, the paper on which this exercise is based). This indicates that the 34 outlying bins identified in Section 4 do not singlehandedly condemn the 2011 Russian election as fraudulent.

We sense, however, based on the fact that 34 is still markedly higher than either 20 or 22, that something about the 2011 election was amiss. Though we cannot definitively conclude based on this analysis that election fraud occurred, we should be suspicious enough to probe further into the results.

Conclusion

This project was inspired by Benford's law, a fascinating and somewhat spooky phenomenon wherein the distributions of naturally-occurring collections of digits is skewed: smaller digits occur at a much higher frequency than larger ones.

We applied principles of Benford's law to evaluate allegations of election fraud in Russia's 2011 election for state legislature. The 'official' results of vote shares across Russian precincts featured a conspicuously high proportion of common fractions, such as 1/2 and 1/3. In conducting a 1,000-trial Monte Carlo simulation of vote share fraction proportions, we did not find a marked distributional difference from the proportions found in the official results.

We continued to investigate the potential for fraud by comparing the official proportions of the four most common vote share fractions with their respective bootstrapped distributions from the Monte Carlo simulation. We then expanded this procedure to include 100 different fractions; we found that for the election in question, 34 of the 100 were observed in proportions that fell outside the middle 95% of Monte Carlo-simulated distributions for those fractions. We would expect that just 5 of 100 would exhibit this behavior by chance; as such, this analysis heightened suspicions of election fraud.

Finally, we contextualized the results by conducting the same procedure for two elections without any allegations of voting irregularities, as well as one marred by accusations of fraud. The two fair and free elections returned 20 and 22 outlying fraction distributions, while the known fraudulent election returned 22 as well. Though we continued to suspect that the 2011 Russian election under examination was tainted by fraud, comparison to other global elections weakened this presumption. Even fair and free elections rendered high numbers of outlying fraction distributions relative to the 5 that would naturally be expected to exhibit irregular behavior. We cannot, therefore, conclude through this method of analysis that the 2011 Russian election contained fraud.

This exercise was based in part on the 2017 work of Arturas Rozenas: "Detecting Election Fraud from Irregularities in Vote-Share Distributions," which appeared in *Political Analysis*, vol. 25, no. 1, pp. 41-56.