HW5_Markdown

2024-02-29

Manay Divatia

md46245

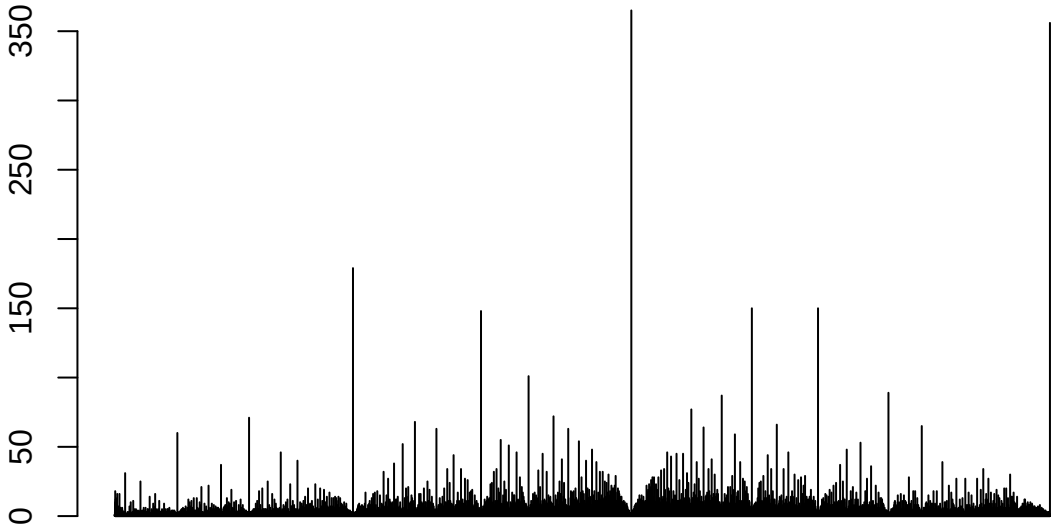
1

Figure 4 is showing how individual income relates to voting in the US. Figure 4 shows an example of Simpson's Paradox. Each individual line is one of the three states of Mississippi, Ohio, and Connecticut. Each open circle is the point on the line where individual income is a whole number from -2 to 2.First looking at the open circles, we see a positive relationship. This means that we see that as income increases, a person is more likely to vote republican. On the other hand, if we focus on the dark circles we see a different trend. The dark circles are the average incomes in those 3 states. We actually see a negative relationship which, in context, means that as individual income increases, the probability of voting republican decreases. This contrasts from the trend we saw with the open circles. Ultimately, we see that when we generalize (the dark circles) we see one trend, but after subsetting the data, we see the opposite trend which is what Simpson's paradox is. My hypothesis for this question is that as individual income, the probability of voting republican increases. I think to test this, I would look at income data for all 50 states and see if the general trend matches that of these 3 states. Based on that, I can either prove or disprove my hypothesis.

2

#a

```
load("~/Downloads/fraud.RData")
P <- sum(russia2011$votes) / sum(russia2011$turnout)
russia2011$rate<- russia2011$votes / russia2011$turnout
tail(names(sort(table(russia2011$rate))), 10)
table(russia2011$rate)
barplot(table(russia2011$rate), name = "barplot")
```

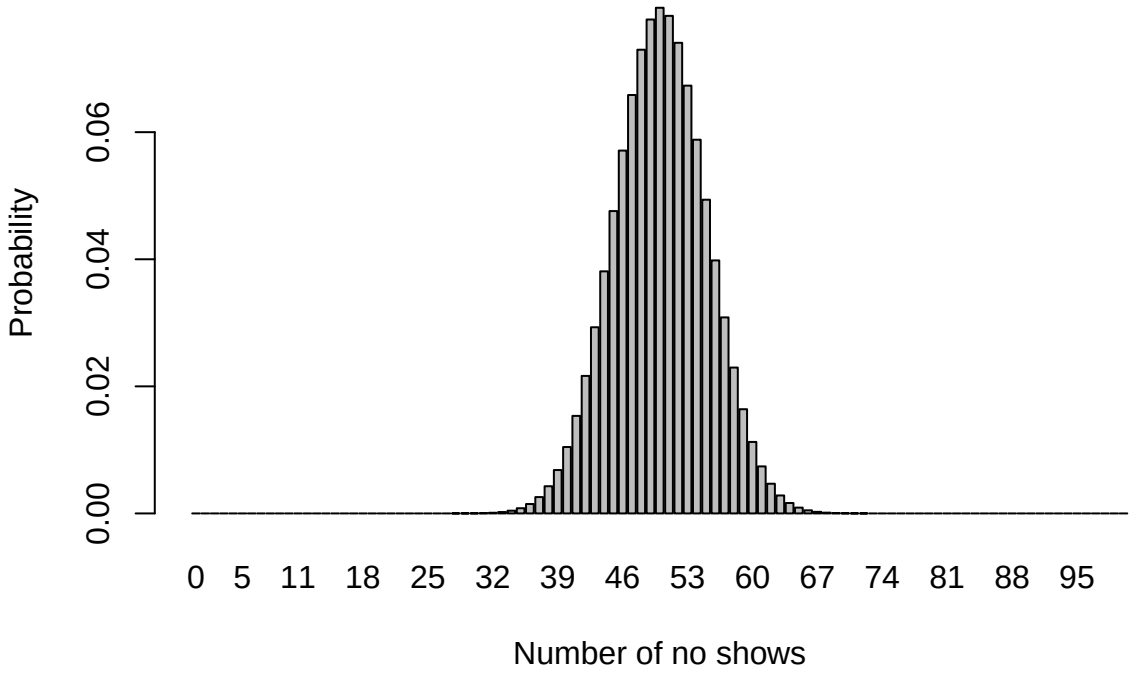


The points around 1/2 and 2/3 are low but for those exact fractions, the instances skyrocket. #b

```
NMC = 100
k_grid = seq(0, 100, by=1)
dbinom(k_grid, NMC, sum(russia2011$turnout) / sum(russia2011$N))
```

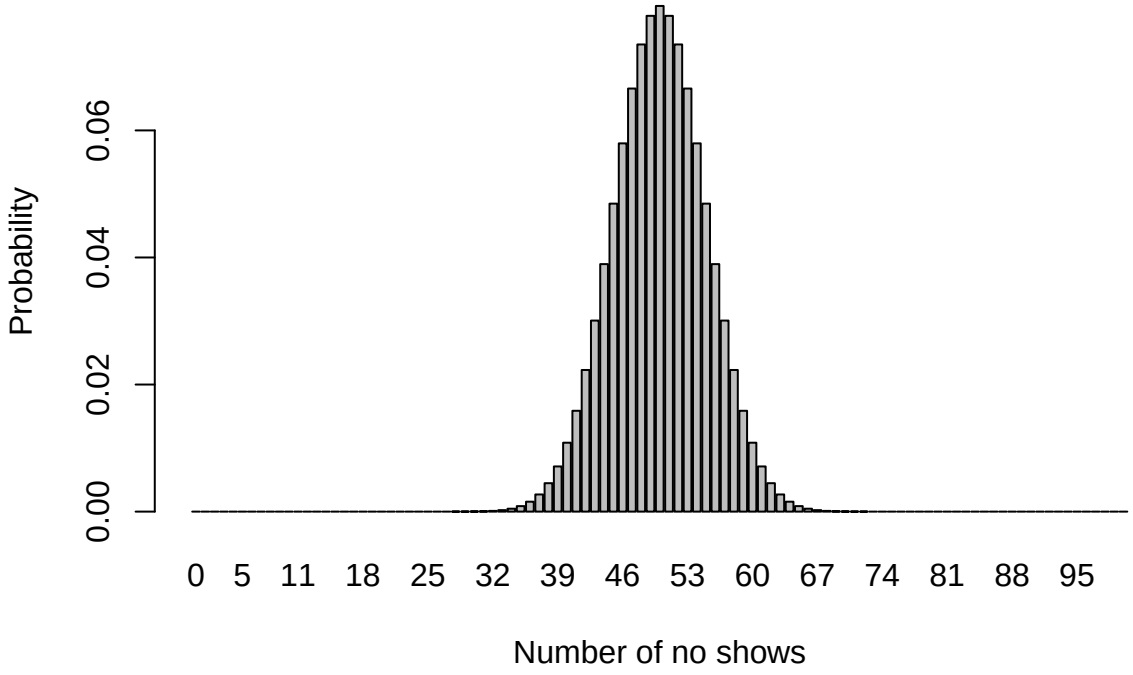
```
##   [1] 1.275334e-39 1.847635e-37 1.324993e-35 6.270624e-34 2.203002e-32
##   [6] 6.127853e-31 1.405636e-29 2.734608e-28 4.605537e-27 6.820527e-26
##  [11] 8.991903e-25 1.065844e-23 1.145235e-22 1.123121e-21 1.011137e-20
##  [16] 8.398647e-20 6.463989e-19 4.627253e-18 3.091157e-17 1.932741e-16
##  [21] 1.134021e-15 6.258697e-15 3.255972e-14 1.599704e-13 7.435522e-13
##  [26] 3.274744e-12 1.368540e-11 5.433976e-11 2.052461e-10 7.382469e-10
##  [31] 2.531227e-09 8.280559e-09 2.586730e-08 7.722152e-08 2.204583e-07
##  [36] 6.022748e-07 1.575425e-06 3.947920e-06 9.482385e-06 2.183921e-05
##  [41] 4.825020e-05 1.022960e-04 2.081871e-04 4.068231e-04 7.635194e-04
##  [46] 1.376537e-03 2.384432e-03 3.968928e-03 6.348924e-03 9.761120e-03
##  [51] 1.442421e-02 2.048728e-02 2.796850e-02 3.669669e-02 4.627253e-02
##  [56] 5.606744e-02 6.527209e-02 7.299575e-02 7.840258e-02 8.085743e-02
##  [61] 8.004695e-02 7.604440e-02 6.929987e-02 6.055748e-02 5.072030e-02
##  [66] 4.069708e-02 3.126653e-02 2.298667e-02 1.616119e-02 1.085842e-02
##  [71] 6.966626e-03 4.264594e-03 2.488487e-03 1.382811e-03 7.309493e-04
##  [76] 3.671061e-04 1.749485e-04 7.899928e-05 3.374805e-05 1.361559e-05
##  [81] 5.177953e-06 1.852232e-06 6.217666e-07 1.953503e-07 5.727644e-08
##  [86] 1.561958e-08 3.946885e-09 9.201434e-10 1.969285e-10 3.846733e-11
##  [91] 6.811370e-12 1.084390e-12 1.536853e-13 1.915277e-14 2.066304e-15
##  [96] 1.890663e-16 1.426610e-17 8.522866e-19 3.779837e-20 1.106268e-21
## [101] 1.602701e-23
```

```
barplot(dbinom(k_grid, NMC, P), names.arg = k_grid,
        xlab='Number of no shows',
        ylab='Probability')
```

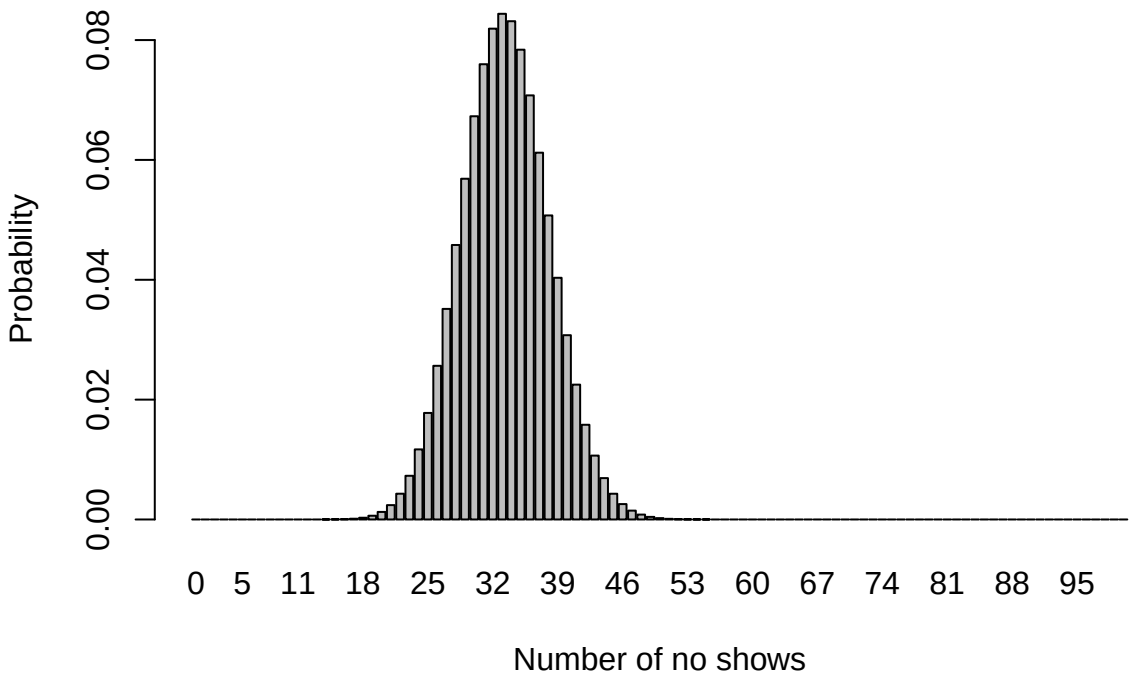


After plotting, we see that the distribution should be a lot more. This does suggest the idea of voter fraud. However, this itself isn't enough evidence to make that claim. #c

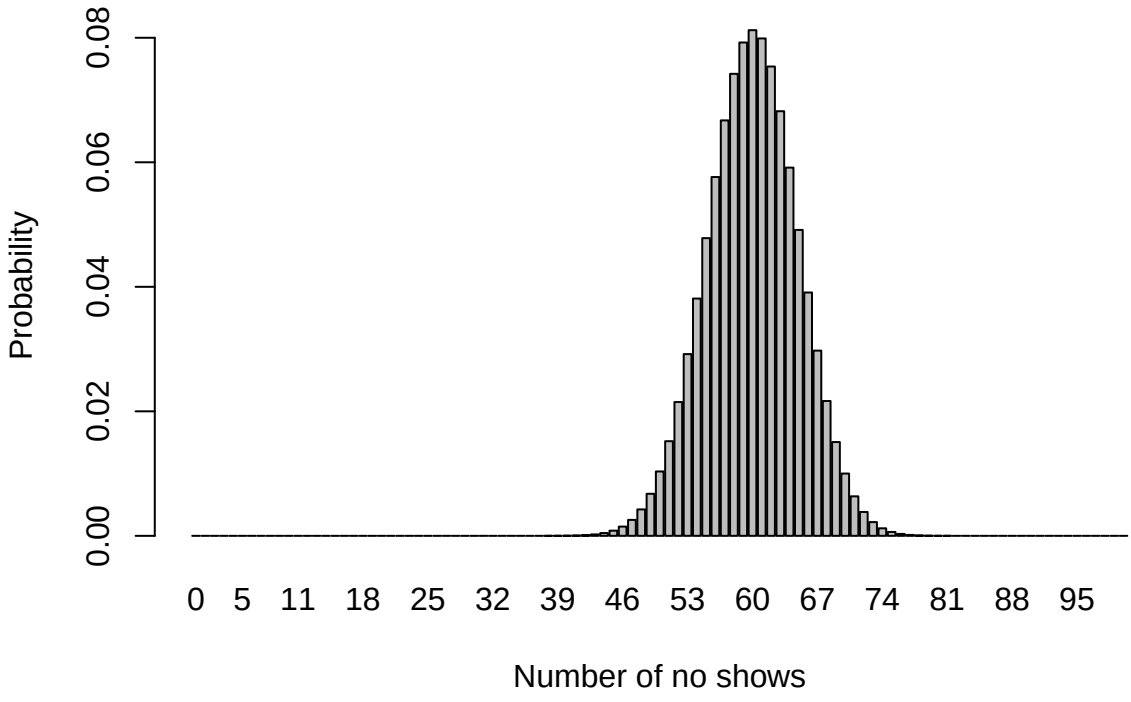
```
c_plots <- function(P) {
  barplot(dbinom(k_grid, NMC, P), names.arg = k_grid,
          xlab='Number of no shows',
          ylab='Probability')
}
c_plots(0.5)
```



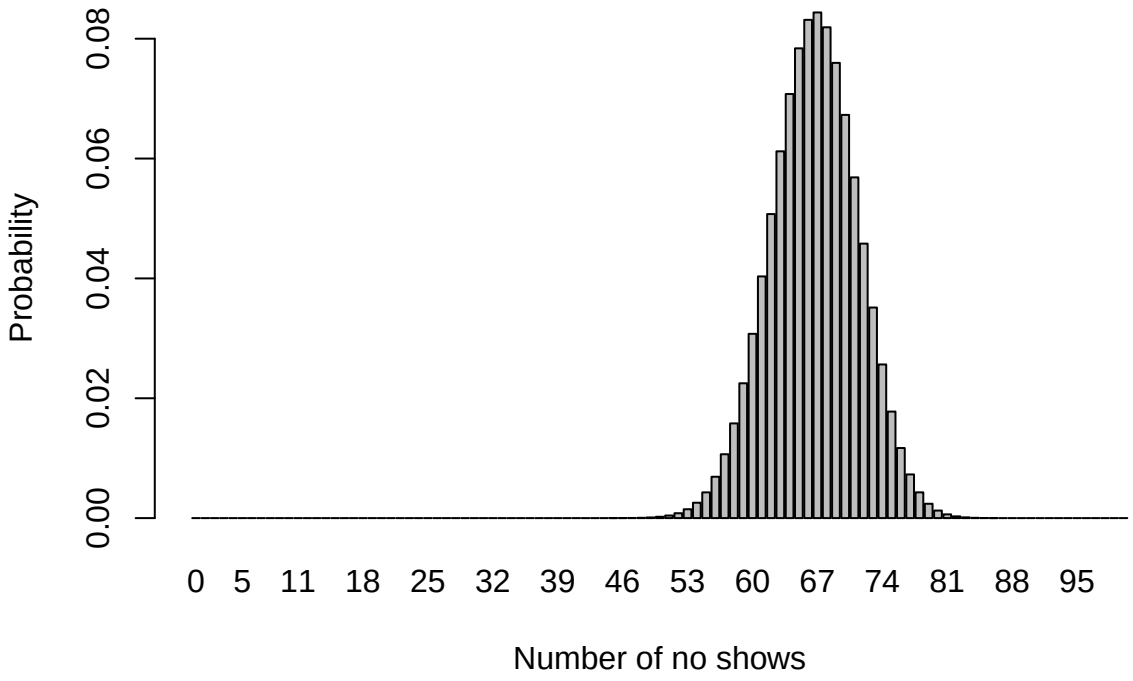
```
c_plots(1/3)
```




```
c_plots(3/5)
```




```
c_plots(2/3)
```

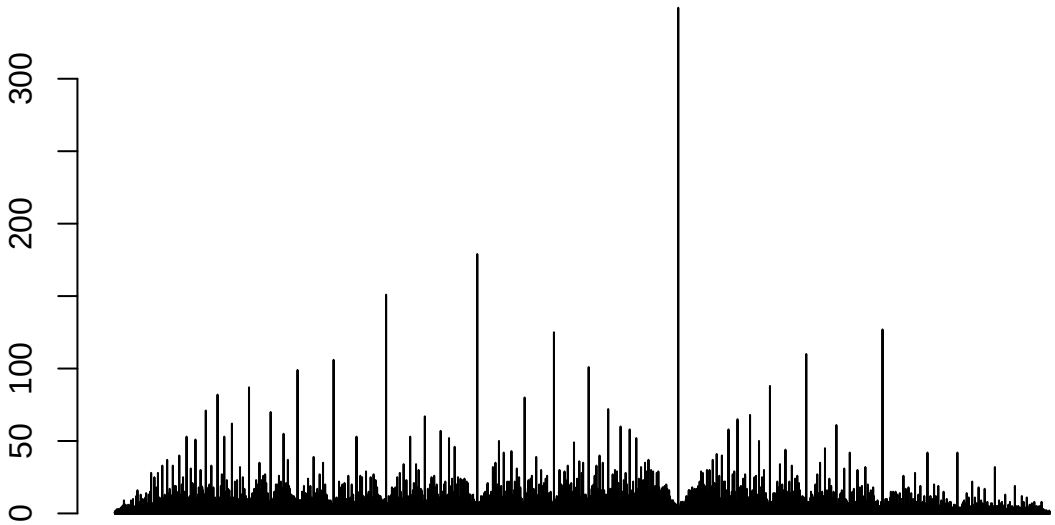


After plotting, we see how the graphs should be when distributed with the different commonly occuring fractions of 1/2, 1/3, 3/5, and 2/3. Again, the distribution does suggest election fraud. #d

```
d_plots <- function(P) {
barplot(table(russia2011$rate) - dbinom(k_grid, NMC, P), names.arg = k_grid,
        xlab='Number of no shows',
        ylab='Probability')
}
#d_plots(0.5)
#d_plots(1/3)
#d_plots(3/5)
#d_plots(2/3)
```

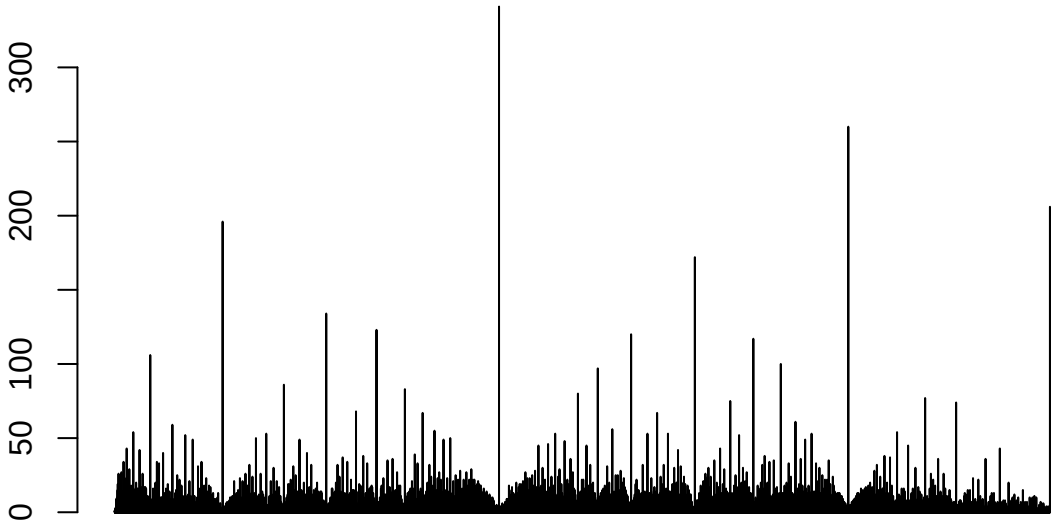
The difference we see here shows that proportion do lie in the <2.5 range and the 97.5> range. This leads to the greater evidence and supports the claim of election fraud in Russia. #e

```
barplot(table(canada2011$votes / canada2011$turnout), name = "barplot")
```



barplot

```
barplot(table(russia2003$votes / russia2003$turnout), name = "barplot")
```

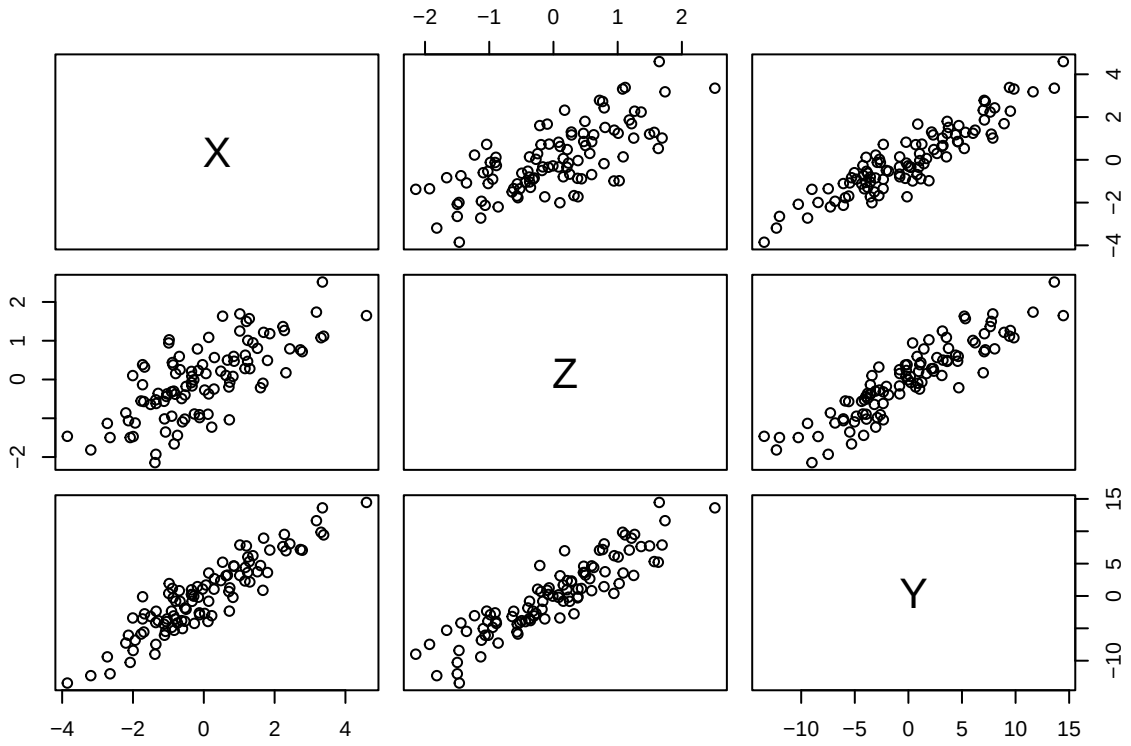


barplot

What is surprising is that we can see a similar outcome in Canada and Russia. Both show points where fractions like 1/2, 1/3 and 1 all appear more often than others. Moreover, they appear significantly more often which gives us evidence that there may have been election fraud during these elections.

3

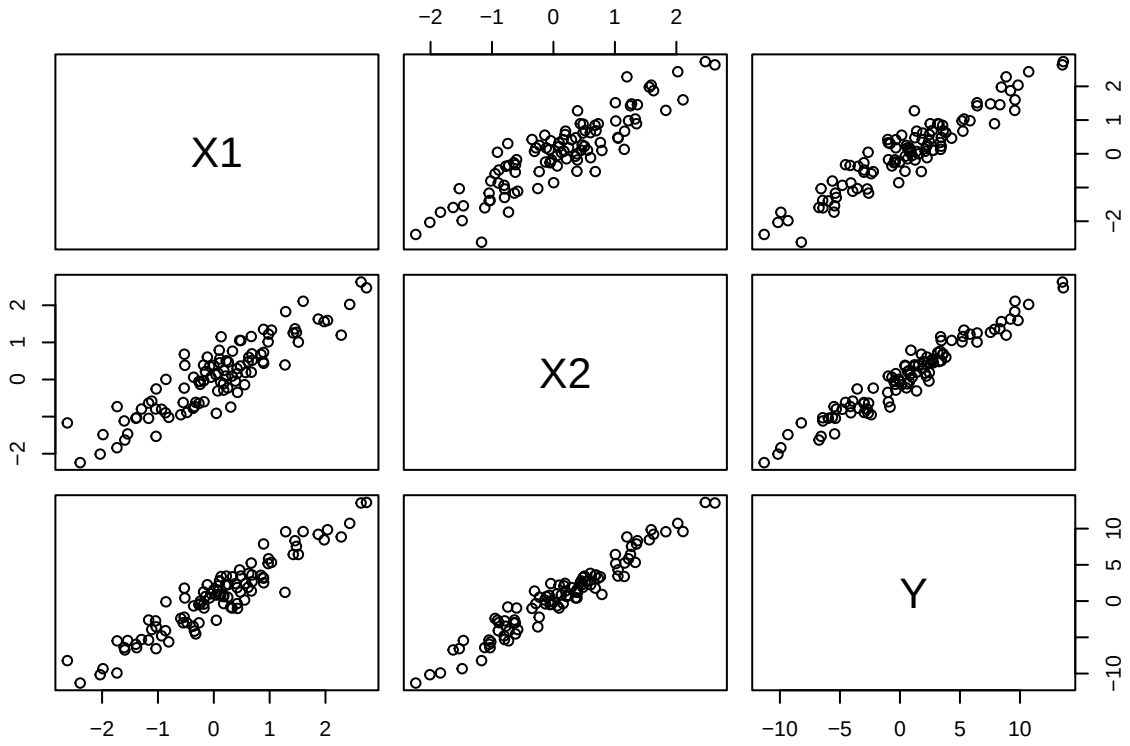
```
n <- 100
Z <- rnorm(n)
X <- Z + rnorm(n)
Y <- 2*X + 3*Z + rnorm(n)
data <- data.frame(X, Z, Y)
pairs(data)
```




```
cor(data)
```

```
##           X         Z         Y
## X 1.0000000 0.7184392 0.9167945
## Z 0.7184392 1.0000000 0.9052867
## Y 0.9167945 0.9052867 1.0000000
```

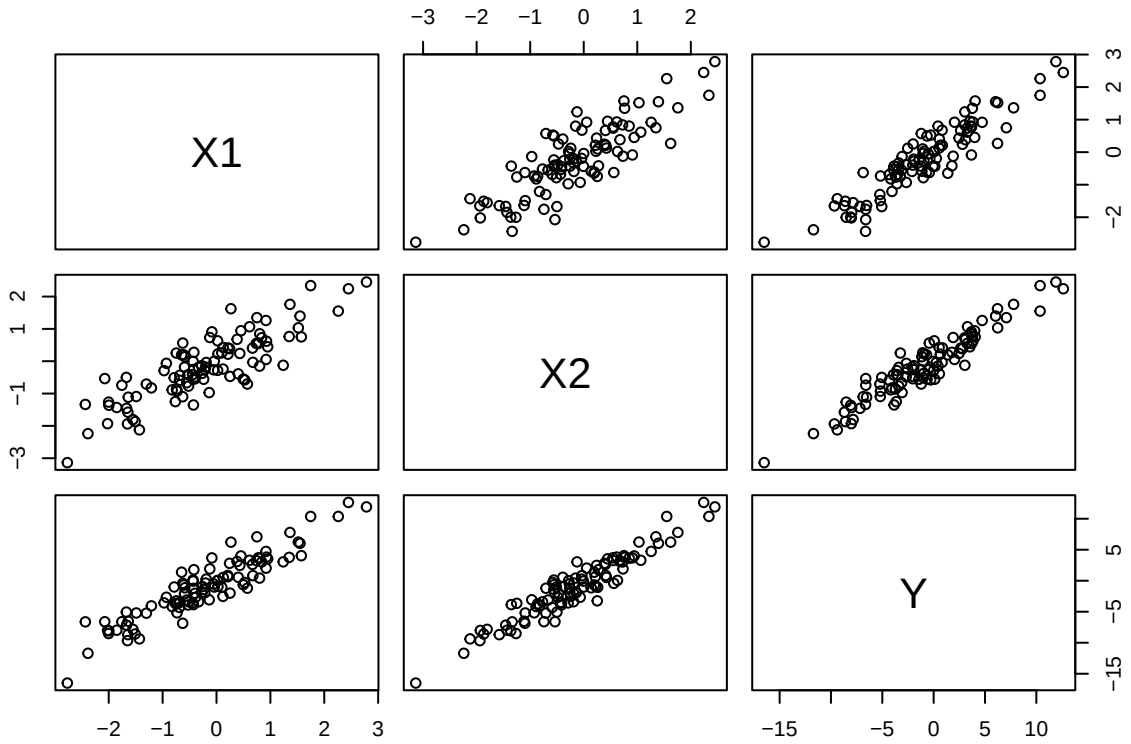
```
n <- 100
X1 <- rnorm(n)
X2 <- 0.8*X1 + rnorm(n, mean = 0, sd = 0.5)
Y <- 2*X1 + 3*X2 + rnorm(n, mean = 0, sd = 1)
confounded_data <- data.frame(X1, X2, Y)
pairs(confounded_data)
```




```
cor(confounded_data)
```

```
##           X1        X2         Y
## X1 1.0000000 0.8944361 0.9427082
## X2 0.8944361 1.0000000 0.9698848
## Y  0.9427082 0.9698848 1.0000000
```

```
n <- 100
X1 <- rnorm(n)
X2 <- 0.8*X1 + rnorm(n, mean = 0, sd = 0.5)
Y <- 2*X1 + 3*X2 + rnorm(n, mean = 0, sd = 1)
confounded_data <- data.frame(X1, X2, Y)
pairs(confounded_data)
```




```
cor(confounded_data)
```

```
##           X1        X2         Y
## X1 1.0000000 0.8431949 0.9266705
## X2 0.8431949 1.0000000 0.9502641
## Y  0.9266705 0.9502641 1.0000000
```

Here are 3 examples of how random variables can display a confounded relationship. What I did for these is that I created random variables but each variable relied on the previous one in some way. Because they relied on the previous one(s) it created a confounding relationship that made both related and taht could be seen through the graphs.