Homework 7

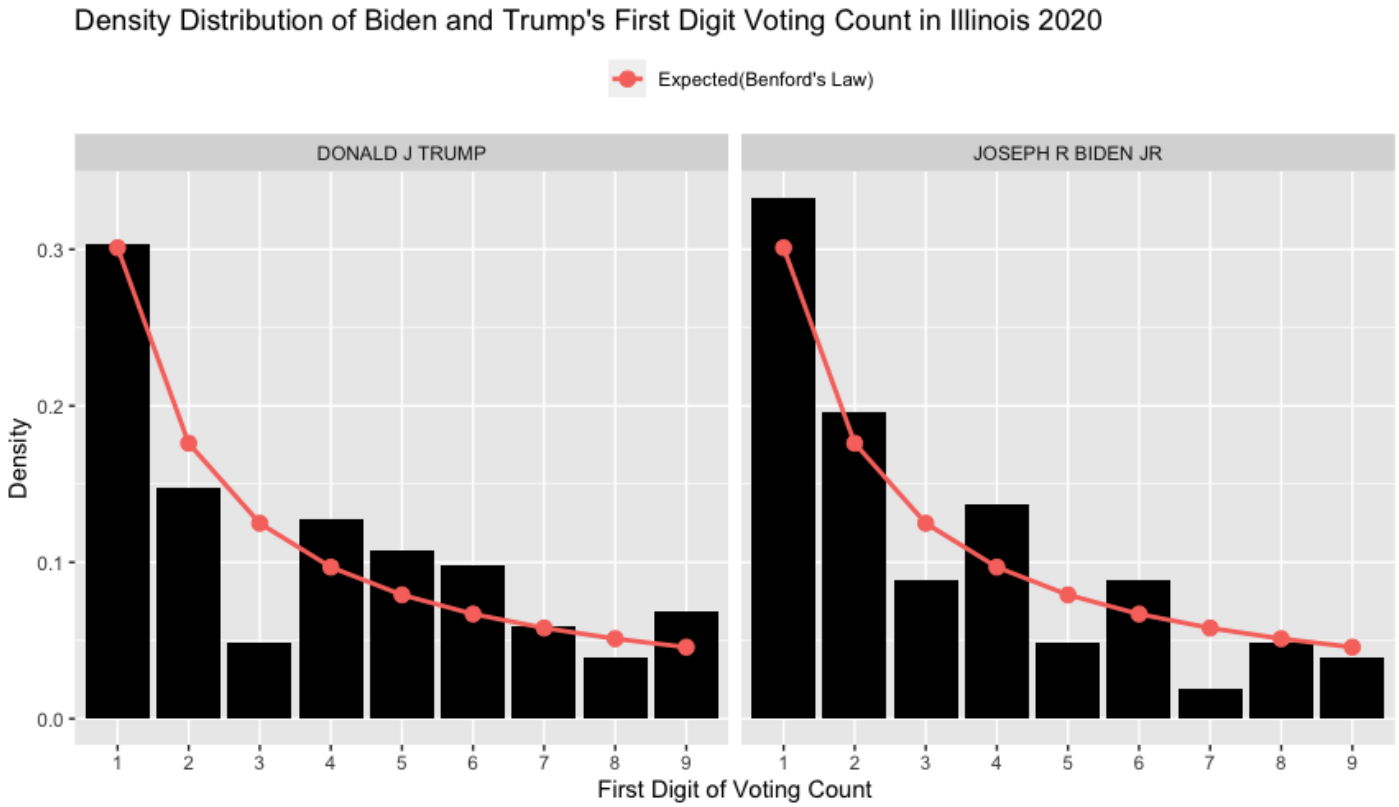
Yiwei He (yh9vhg), Da Lin (dl2de), Ziyue Jin (zj5qj)

Problem 1

Based on Trump and Biden's election record in Illinois 2020, we find each candidate's first-digit voting count distribution to be deviating from the expected distribution, therefore violating Benford's Law.

For Trump, densities of first digits equal to 4,5,6,9 are higher than Benford's distribution, whereas digits equal to 2 and 8 are lower. Density of 3 is substantially lower, less than half of the theoretical distribution. For Biden, densities of first digits equal to 1,2,4,6 are higher than Benford's distribution, whereas digits equal to 3,5,7,9 are lower.

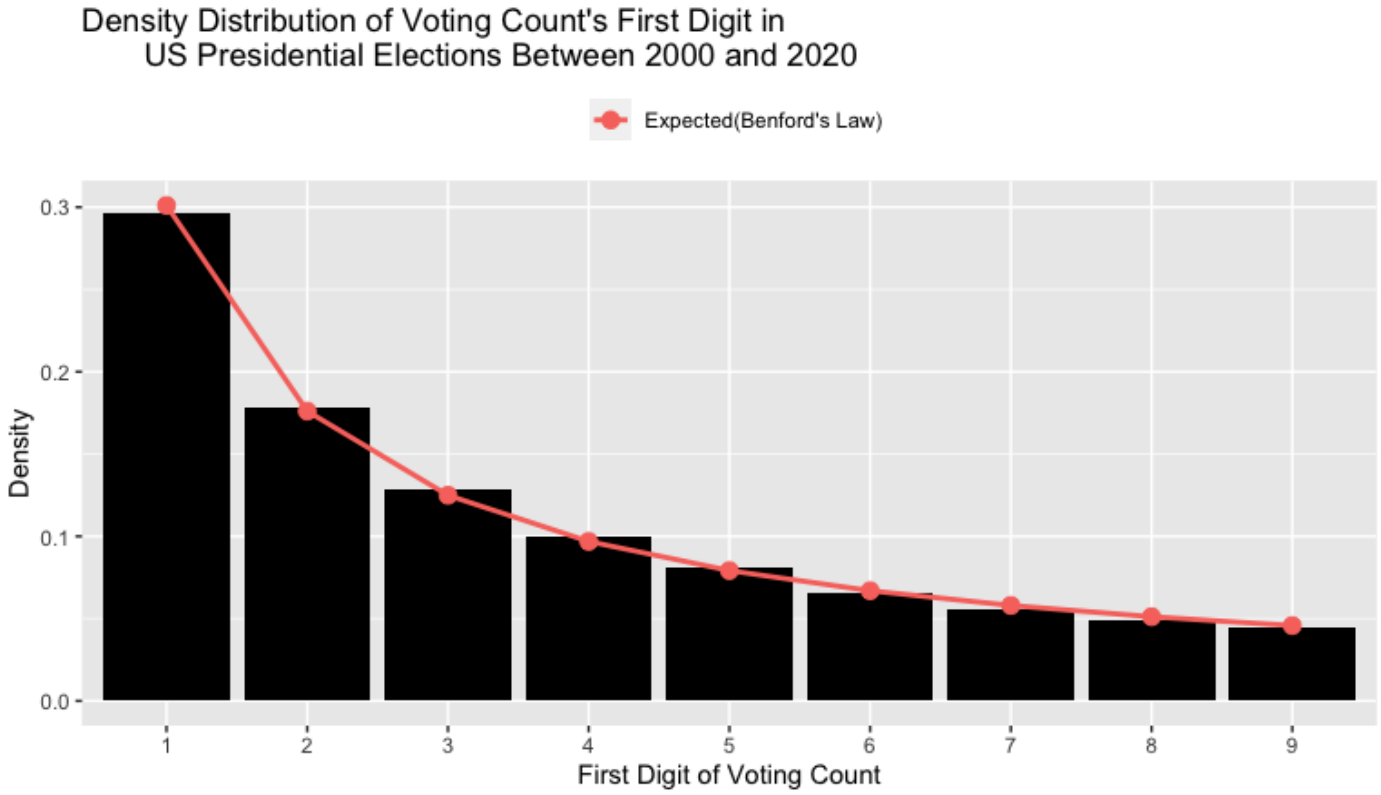
Both Trump and Biden have some digits that don't follow the theoretical distribution. Therefore, we conclude that there might be election fraud associated with their campaign in Illinois in 2020.



Problem 2

The following visualization is based on voting counts associated with each recorded election candidate (whose voting count isn't null or zero) across all stats between 2000 and 2020. We did this to observe whether results found in part (a) can be applied to an overall distribution.

The sample distribution seems to correspond with the Benford curve, with minor digressions (< 1% in density differences). Therefore, even though we observed violations in part (a), there isn't sufficient evidence for us to conclude that election fraud exists in general because we now have a different result based on a much larger sample.



Appendix

```
### Problem 1
library(tidyverse)
library(ggplot2)
dt <- read.csv("~/Desktop/Homework 7/CountyData_2000-2020.csv")
dt$digits <- substr(as.character(dt$candidatevotes),1,1)
illinois <- dt %>% filter(year==2020 & state=="ILLINOIS")
benford <- data.frame(digits=1:9,val=1:9)
for (i in 1:9) {
  benford$val[i] <- log10((i+1)/i)
}

## overall
vote <- illinois %>% filter(candidate=="JOSEPH R BIDEN JR" | candidate=="DONALD J
TRUMP")

ggplot(vote,aes(x=digits)) +
  geom_bar(aes(y=..prop..,group=candidate,fill="black"),fill="black") +
  geom_point(data=benford,aes(x=digits,y=val,color="tomato"),size=3) +
  geom_line(data=benford,aes(x=digits,y=val,color="tomato"),lwd=1) +
  scale_color_discrete(labels = c("Expected(Benford's Law)")) +
  labs(title="Density Distribution of Biden's First Digit Voting Count in Illinois 2020",
       x="First Digit of Voting Count",y="Density") +
  theme(legend.position="top",legend.title=element_blank()) +
  facet_grid(.~candidate)

### Problem 2
overall <- dt %>% filter(digits!="0" & is.na(digits)==F)
ggplot(overall,aes(x=digits)) +
  geom_bar(aes(y=..prop..,group=1,fill="black"),fill="black") +
  geom_point(data=benford,aes(x=digits,y=val,color="tomato"),size=3) +
  geom_line(data=benford,aes(x=digits,y=val,color="tomato"),lwd=1) +
  scale_color_discrete(labels = c("Expected(Benford's Law)")) +
  labs(title="Density Distribution of Voting Count's First Digit in
       US Presidential Elections Between 2000 and 2020",
       x="First Digit of Voting Count",y="Density") +
  theme(legend.position="top",legend.title=element_blank())
```