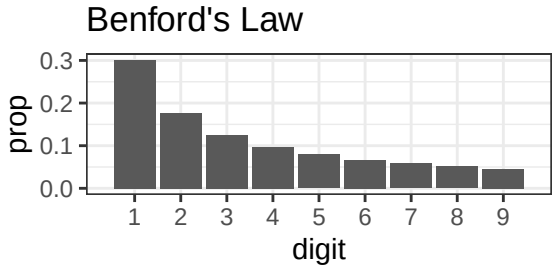
Lab 5: Informal Inference



Part I: 2009 Iran Election

On June 12 2009, the Republic of Iran held an election where President Mahmoud Ahmadinejad sought re-election against three challengers. One of the challengers, Mir-Hossein Mousavi. When it was announced that Ahmadinejad had won handily, there were widespread allegations of election fraud. There are many methods, both quantitative and qualitative, to detect election fraud. In this lab we will explore just one proposed method.

```
library(tidyverse)
library(stat20data)
data(iran)
```

Exploratory Data Analysis

1. What is the unit of observation in the `iran` data frame? What are the dimensions?
2. Which cities had the highest proportion of total votes cast for Ahmadinejad? Please return the top several city names along with the province name and the proportions.
3. Which cities had the highest proportion of total votes cast for Mousavi? Please return the top several city names along with the province name and the proportions.
4. How many cities did Mousavi win?
5. *How does the proportion of total votes that were voided compare between cities won by Mousavi and cities won by Ahmadinejad? This can be answered either with a plot or with summary statistics. Describe in words how they compare.
6. *What proportion of the total votes cast were won by Ahmadinejad and Mousavi, respectively?

First Digit Distribution

7. What proportion of vote counts for Ahmadinejad have "1" as a first digit?
8. *Create a plot showing the distribution of first digits in the Ahmadinejad's vote counts. Does this plot appear to match the ideal Benford's distribution? Where does it deviate?
9. *Would you consider this meaningless, weak, moderate, or strong evidence of election fraud? Why or why not?

Part II: 2020 US Election

The OpenElections project obtains and standardizes precinct-level results from US elections, including the 2020 US Presidential Election. To access the data, search OpenElections' GitHub page (https://github.com/openelections) and click on a link to a data repository for the state of your choosing. Select the `2020` folder and find a file that ends in `.csv`. Some notes:

- Each state uses a different format, so click through a couple states' repositories until you find one that will allow you to study voting patterns at the precinct-level.
- To read the csv file into R, you will need to point R to the raw version of the data set. To view the raw csv you will either click the button that says "Raw" at the top right of the data frame on GitHub or click the link that says "View Raw Data". When you are looking at the raw csv file, the url in your browser is the one you can use to access the file from within R.
- There may be strange extra rows in your data, such as a row tallying total overall votes. Visually inspect the data to see if anything jumps out and be sure to take this into consideration when doing your analysis.

Exploratory Data Analysis

10. What state did you choose to study? What is the unit of observation in your state's data frame? What are the dimensions?

11. Create a table of the total votes won by each candidate for President, arranged in descending order. Who won in your state? Compare your table to a student that is studying a different state. Who won in their state? Which margin was larger?

12. Which 5 counties had the greatest number of total votes for Joseph Biden? Which 5 counties had the greatest number of total votes for Donald J. Trump? If there is a strong disagreement between those two lists, i.e., a county that cast lots of votes for Biden and few for Trump or vice-versa, do a bit of research into that countie(s) and describe a bit about it (its principle cities, its economy, etc).

13. *Take a look through the data set and note the other variables that are recorded - each state records slightly different things. Ask and answer one additional question of your choosing about the election in your state using this data set. The question should be able to be answered either with a plot or a data frame.

First Digit Distribution

14. Use this data to create a plot of that state's first digit distribution by precinct. Use the number of votes cast for Joseph Biden in each precinct.

15. Does the election you chose appear to fit Benford's distribution better or worse than the Iran election?

Part III: Creating a Distance Statistic

16. We've used plots to assess the difference between the observed distribution and the distribution expected by Benford's Law, but it can be useful to quantify that difference with a statistic. We can start by building a dataframe that directly compares the proportions for each of the digits.

    Using the data wrangling techniques from the last lab, reformulate the `iran` data into the following dataframe and call it `iran_props`.

```
## # A tibble: 9 x 3
##   first_digit obs_prop ben_prop
##         <dbl>    <dbl>    <dbl>
## 1           1   0.260    0.301
## 2           2   0.232    0.176
```

```
## 3           3   0.148    0.125
## 4           4   0.0820   0.0969
## 5           5   0.0874   0.0792
## 6           6   0.0628   0.0669
## 7           7   0.0464   0.0580
## 8           8   0.0546   0.0512
## 9           9   0.0273   0.0458
```

17. *Formulate your own statistic to measure the distance between the observed proportions (`obs_prop`) and those expected by Benford's Law (`ben_prop`). There are many many possible choices, but some are more useful than others. Describe this statistic in words (or write out the formula for it if you are comfortable using LaTeX), then calculate it for this data.

18. *Repeat exercises 16 and 17 for the US state data. Which distribution is judged to be farther from the ideal Benford's distribution using your distance statistic? What do you conclude about the presence of fraud in these elections?