3.4) Data Design

Vitor Kamada

December 2019

Reference

Tables, Graphics, and Figures from

Principles and Techniques of Data Science

Lau et al. (2019): Ch 2 Data Design

https://www.textbook.ds100.org/ch/02/design_intro.html

Presidential Election in 1936

Companies	Sample Size	Landon	Roosevelt
Literary Digest	10M	57%	43%
Gallup	50K	44%	56%
Actual Result		37%	62%

Polled a sample based on telephone and car registrations

It was not a **Probability Sampling**

1948: The Gallup Poll

Quota Sampling: each interviewer polled a set number of people from each demographic class

Thomas Dewey would earn 5% more of the popular vote

But, Truman won with 5% more votes

Gallup Poll predicted 2-6% more Republican votes than the actual results for the 3 elections prior

Are events C and D independent?

$$P(A \cap B) = P(A) \times P(B)$$

$$P(\heartsuit \cap ace) = P(A\heartsuit) = \frac{1}{52}$$

$$P(\heartsuit) \times P(ace)$$

$$\frac{1}{4} \times \frac{1}{13} = \frac{1}{52}$$

C = a randomly selected US citizen is over 90 years old

D = the citizen is male

Simple Random Sample (SRS) of size 2

Population: 6 individuals [A,F]

AB	BC	CD	DE	EF
AC	BD	CE	DF	
AD	BE	CF		
AE	BF			
AF				

$$P(AB) = P(BC) = ... + P(AF) = \frac{1}{15}$$

Cluster vs Stratified Sampling

3 Clusters: (A,B), (C,D), (E,F)

$$P(A\_in\_sample) = P(AB) = \frac{1}{3}$$

$$P(AC) = 0$$

Strata 1: A, B, C, D **Strata 2**: E, F

SRS of one individual from each strata

(A,E) (A,F) (B,E) (B,F) (C,E) (C,F) (D,E) (D,F)

$$P(A\_in\_sample) = \frac{2}{8}$$

$$P(AB) = 0$$

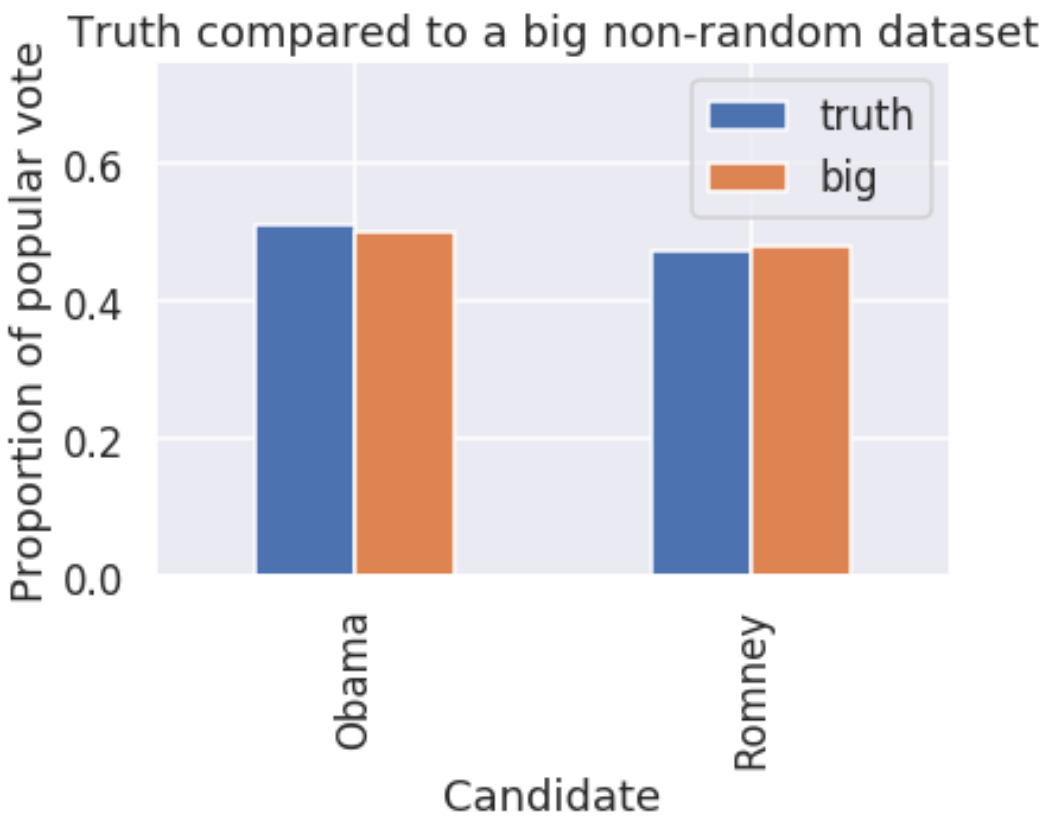
2012: Obama vs Mitt Romney

```
total = 129085410
obama_true_count = 65915795
romney_true_count = 60933504
obama_true = obama_true_count / total
romney_true = romney_true_count / total

# 1 percent off
obama_big = obama_true - 0.01
romney_big = romney_true + 0.01
```

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
sns.set()
sns.set_context('talk')
```

```
pd.DataFrame({'truth': [obama_true, romney_true],
    'big': [obama_big, romney_big],}, index=['Obama', 'Romney'],
     columns=['truth', 'big']).plot.bar()
plt.title('Truth compared to a big non-random dataset')
plt.xlabel('Candidate')
plt.ylabel('Proportion of popular vote')
plt.ylim(0, 0.75);
```



Simple Random Samples vs "Big Data"

120M voters in 2012

```
srs_size = 400
big_size = 60000000
replications = 10000

def resample(size, prop, replications):
    return np.random.binomial(n=size, p=prop,
                              size=replications) / size

srs_simulations = resample(srs_size, obama_true, replications)
big_simulations = resample(big_size, obama_big, replications)
```

```
bins = bins=np.arange(0.47, 0.55, 0.005)
plt.hist(srs_simulations, bins=bins, alpha=0.7, normed=True, label='srs')
plt.hist(big_simulations, bins=bins, alpha=0.7, normed=True, label='big')
plt.title('Proportion of Obama Voters for SRS and Big Data')
plt.xlabel('Proportion')
plt.ylabel('Percent per unit')
plt.xlim(0.47, 0.55)
plt.ylim(0, 50)
plt.axvline(x=obama_true, color='r', label='truth')
plt.legend();
```