DS 100: Principles and Techniques of Data Science	Date: August 29, 2018

Discussion #1

Name:

Probability & Sampling

1a 1b 1c 2a 2b 2c 3a 3b 3c 4a 4b 4c
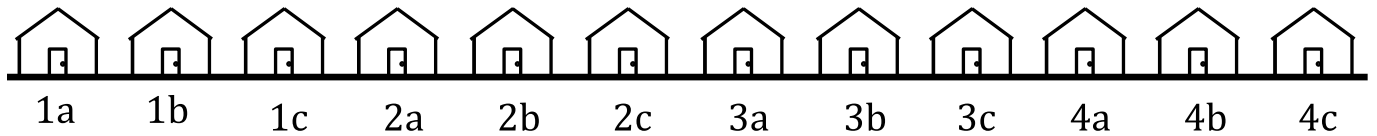
1. Kalie wants to measure interest for a party on her street. She assigns numbers and letters to each house on her street as illustrated above. She picks a letter "a", "b", or "c" at random and then surveys every household on the street ending in that letter.

   (a) What kind of sample has Kalie collected?

   (b) What is the chance that two houses next door to each other are both in the sample?

   (c) Now suppose Kalie instead picks one house beginning with '1' at random, one house beginning with '2' at random, and so on, so she surveys four houses, one of each number. What kind of sample has Kalie collected?

   (d) Kalie randomly selects 4 houses on the street. In each house, she randomly selects one household member to interview. What kind of sample has Kalie collected?

2. There are 32 participants in a randomized clinical trial: 8 are male and 24 are female. 16 are assigned to treatment and the others are put into the control group. What is the probability that none of the men are in the treatment group if:

   (a) the treatment was assigned using stratified random sampling, grouping by gender?

   (b) the treatment was assigned using simple random sampling?

   (c) the treatment was assigned using cluster random sampling of 2 groups of 8 using clusters as described below?

Cluster	Male	Female
A	0	8
B	3	5
C	5	3
D	0	8

A Big Data Fail

Consider the 1936 federal presidential election of FDR vs Al Landon. The magazine Literary Digest's straw poll had correctly predicted the outcome of the previous five presidential elections. Running up to the election, they polled over 10 million individuals including

```
- magazine subscribers
- registered automobile owners
- telephone owners
```
and received responses from about 2.4 million of those polled. The Literary Digest predicted Landon would win in a landslide. By contrast, George Gallup's quota sample consisted of bi-weekly surveys of 2000 individuals, and correctly predicted a landslide for FDR.

3. What are some potential sources of bias in each of these polling schemes?