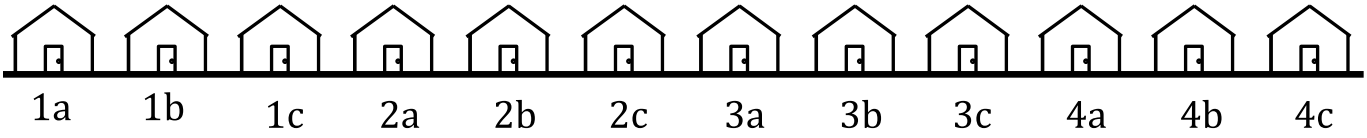
DS 100: Principles and Techniques of Data Science Date: January 25, 2019

Discussion #1 Solutions

Name:

Probability & Sampling



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?

   Solution: A cluster sample (each group of houses ending in a certain letter is a cluster).

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

   Solution: None of the adjacent houses end in the same letter, so the chance is zero.

   (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?

   Solution: A stratified sample.

   (d) Kalie randomly selects 4 houses without replacement on the street. In each house, she randomly selects one household member to interview. What kind of sample has Kalie collected?

   Solution: A multi-stage sample.

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?

   Solution: 0

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

   Solution:

   $$\frac{24 \times 23... \times 10 \times 9}{32 \times 31 \times ... \times 18 \times 17}$$

   or

   $$\frac{\binom{24}{16}}{\binom{32}{16}}$$

   (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

   Solution:

   $$\frac{1}{\binom{4}{2}} = \frac{1}{6}$$

   Solution: Alternative Analytical solution: Another way to think about how to solve this problem is to first list out all the possible 2-cluster groups that can be chosen. The 6 2-cluster groups would be (A, B), (A, C), (A, D), (B, C), (B, D), and (C, D). Out of these 6 groups, only one of the groups would result in 0 men chosen, namely the (A, D) group. Therefore, the probability that none of the men are in the treatment group would be $\frac{1}{6}$.

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 $2,000$ individuals, and correctly predicted a landslide for FDR.

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

Solution: Possible answers: The Literary Digest poll was more likely to get responses from wealthier families (car-owners, telephone owners, and those with disposable incomes to subscribe to a magazine). Those with strong enough opinions to respond to the poll are likely to be different from those who did not. The quota sample, while better than the Literary Digest poll, suffers from its own implicit biases. Interviewers were required to search for subgroups like "seven white males under 40 living in a rural area" but beyond that were given their own discretion in choosing who to interview. In fact, in the 1948 election, Gallup's method incorrectly predicted that Thomas Dewey would win over Harry Truman.

Data-Driven Study Design: COMPAS Algorithm for Predicting Recidivism

Recidivism is the tendency of a convicted criminal to reoffend. The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm, developed by the company Northpointe (now equivant), predicts recidivism risk based on variables related to criminal history, drug involvement, and juvenile delinquency. It is used by US courts for the purpose of case management, to predict a defendant's risk of committing more crimes.

4. We will examine the COMPAS algorithm and, in particular, a ProPublica study pointing to its racial bias (https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm). We will discuss general issues raised by the application of such algorithms, e.g., in terms of ethics, privacy, security, and governance? We will also walk through steps you might take to address questions related to the accuracy and potential racial bias of the COMPAS algorithm.

   The questions are meant be discussed with the people around you as a group and there is no right or wrong answer.

   (a) What would be the population of interest if we were conducting an analysis of the COMPAS algorithm?

   Solution: Some ideas, but not limited to are:

   Ideally: All convicted people who were assessed by the COMPAS algorithm.

   Reality: Many studies have used subsets of people assessed by COMPAS. For example, ProPublica (nonprofit org that does investigative journalism) looked at 10,000 criminal defendants in Broward County, Florida. This is because Florida has strong open-records laws and Broward Country largely used COMPAS.

   (b) What are some features or attributes that were used by COMPAS to design the algorithm? Are there features or attributes that you think should've been included or taken out?

   Solution: Although we don't know exactly what features were used by the COMPAS algorithm, some types of variables COMPAS used included criminal history, drug involvement, juvenile delinquency.

   Some ideas of features that could be included (not limited to) are housing status, such as if a person has a home that he/she lives in, income, weight, etc.

(c) How do you define "accuracy" and "racial bias"?

Solution: Some ideas or avenues of exploration for defining "accuracy" might be:

Accuracy: captures overall performance predicting risk of recidivism. It is usually defined as % of instances accurately classified.
Note: there are many other metrics that capture performance.

Bias: Bias measures how close the average of an estimator is to the parameter.

One way to define **racial bias** in the context of an algorithm is an algorithm that exploits systematic prejudices in decision-making. In reality, there is no single way to measure racial bias.What are some metrics we could use? Maybe we could answer this question: Controlling for all other factors (ie. holding them constant), does race make a difference in the prediction?

(d) How should data be collected or obtained to assess the accuracy of predictors like COMPAS? Would you sample at random from the population of interest?

Solution: To wholistically evaluate the COMPAS algorithm, you may want to ensure that your sample is representative of different racial, ethnic groups and income levels.

Ideally: You would like to take a random sample from the population of interest.

Reality: Many counties do not make COMPAS data available (strong record laws).

(e) What are some ways we can assess the accuracy of COMPAS?

Solution: One way to measure accuracy (certainly not the only way and it is up for debate) of COMPAS is to assess certain metrics, such as the ones below:

False Positive: Incorrectly predicting that an individual is high risk.

False Negative: Incorrectly predicting that an individual is not high risk.

Another set of metrics one could use, as will be discussed later in the class when we talk about logistic regression is:

Precision: Out of all the individuals I predicted as high risk, what proportion of the individuals actually recidivated?

Recall: Out of all the individuals who actually recidivated, what proportion of individuals did I predict as high risk?

(f) Think about the concepts of false positives and false negatives in this scenario. What are the ramifications or costs of a false positive and/or false negative?

Solution: One way to view ramifications of false positives and false negatives is below:

Ramification of False Positive: Incorrectly predicting that an individual is high risk. A consequence would be falsely jailing an individual that wouldnt have recidivated.

Ramification of False Negative: Incorrectly predicting that an individual is not high risk. A consequence would be not jailing an individual that would have recidivated.

Note: As data scientists and algorithm designers, it is up to you how you come up with metrics to evaluate your algorithms and models and how you assess the ramifications of those metrics. All of the above questions are ones you should be thinking of when designing and analyzing an algorithm to make sure it truly answers the questions you are trying to solve.