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

Discussion #1 Solutions

Name:

Technical Forum Etiquette

You will undoubtedly meet countless computing errors in your data science career. This is normal, and one of the reasons why technical forums like StackOverflow (https://stackoverflow.com/) are so popular.

We are more than happy to help you resolve issues that arise during this course. At the same time, we need to help prepare you for what happens beyond this course. As such, we will be enforcing good forum etiquette on Piazza.

Before Posting

Before posting on Piazza do the following try to find an answer by:

- Reading the docs.
- Searching the Web.
- Searching Piazza for existing posts. Use the folder/filter system.
- Inspection or experimentation.

MCVEs

Sometimes you'll come across a coding problem that will just completely stump you, and this is fine—you're here to learn after all! In order to maximize benefit to the class, we ask that you take the time to craft coding questions in a way that doesn't give away direct solutions to homework problems. The best way to do this is by abstracting the problem to minimal, complete, and verifiable examples (MCVEs).

By providing the code, you enable your would-be helper to quickly go straight to helping you troubleshoot rather than placing the onus of tediously transcribing say a screenshot into code. Moreover, by removing the homework context from the question, you can post publicly to increase the chances of someone answering you and to help others who are in a similar situation.

For example, you find yourself wanting to reshape a dataset in a homework. Instead of posting a screenshot of the specific dataset and all of your code, you should make a small abstract data frame and pose your question around it. "I'm trying to change df into df2 (provided below) but don't know what this operation is called."

```
import pandas as pd
df = pd.DataFrame({
    'foo': ['one', 'one', 'one', 'two', 'two', 'two'],
    'bar': ['A', 'B', 'C', 'A', 'B', 'C'],
    'baz': [1, 2, 3, 4, 5, 6],
    'zoo': ['x', 'y', 'z', 'q', 'w', 't']
})
df2 = pd.DataFrame({
    'foo': ['one', 'two'],
    'A': [1, 4],
    'B': [2, 5],
    'C': [3, 6]
})
```

Reporting Errors

Another problem that arises is when you think you have a solution, but your code acts in a way that you did not anticipate. Worse, Python then returns an error message that you don't quite understand how to read. The best course of action then would be as follows. Try to reproduce the same problem using a MCVE. If you can do this, then now you can attempt debugging on the easier-to-reason-about data instead. If you can't, then your job is to document exactly how to reproduce the error in the more complicated setting:

1. Provide the copy-pastable MCVE. Format this nicely using Piazza's code environment.
2. State what you expected to see and a copy/screenshot of what you saw instead. Often this will be the entire error message.
3. State how you've tried to resolve the problem

When you ask your question, display the fact that you have done these things first and comment on what you have learned through doing them. Some examples of good effort:

- I've tried Googling the terms X, but the most relevant page I could find only talks about Y. Any hints on better search terms or resources I can read up on?
- I've tried reading the docs found at this ⟨ link ⟩, but I don't quite understand what they are expecting for argument X.
- I found this post on StackExchange ⟨ link ⟩. This is almost what I want except that X, Y, Z.

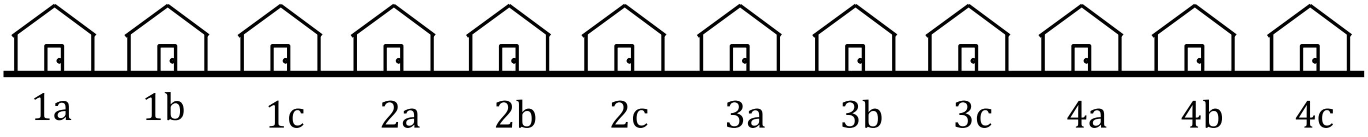
Finally, after you've written your post,

1. Write informative subject lines. Avoid titles like "HELP!" or "pandas error"
2. Use the folder/filter system in Piazza to tag your posts
3. Summarize your findings after you resolve your problem.

Resources

- **Posing Good Questions** `https://stackoverflow.com/help/how-to-ask`
- **Constructing Examples** `https://stackoverflow.com/help/mcve`

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

   > **Solution:** A cluster sample (more specifically, a "multi-stage cluster 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:

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

   (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:

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

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?

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.