Data 100, Discussion 1 – Probability and Sampling

Suraj Rampure

Friday, January 28th, 2019

Suraj Rampure

I'm a junior EECS major from Windsor, Ontario, Canada. This is my second time TAing for DS 100, but I've TA'd for CS 61A and Data 8 previously. I'm also a part of CSM, and I'm the instructor for Introduction to Mathematical Thinking.

- **Email**: suraj.rampure@berkeley.edu – feel free to email me about anything!
- **Lab**: Tuesdays, 12-1 and 1-2PM, Evans B6
- **Discussion**: Fridays, 12-1 and 1-2PM, Etcheverry 3119
- **Office Hours**: 12-2PM, Soda 411

All discussion/lab slides and other resources will be posted at surajrampure.com/teaching.

tinyurl.com/goatjames

Before leaving, please fill out this form. It contains some introductory questions, just for me to learn who is in the class.

Sampling Techniques

Discuss the following types of samples:

- **Convenience sample**
- **Simple random sample**
- **Cluster sample**
- **Stratified sample**
- **Multi-stage sample**

Convenience sample

Quite literally, sampling whoever is convenient

Simple random sample (SRS)

Sampling uniformly at random from the population

- For example, if we have five students, A, B, C, D, and E, and want to select two of them, our sample will look like AB, AC, AD, AE, BC, BD, BE, CD, CE, or DE.
- There are $\binom{5}{2}$ total possible samples, and each is equally likely (with probability $\frac{1}{\binom{5}{2}} = \frac{1}{10}$).
- Each student appears in exactly 4 of the samples, so the probability that any one specific student appears in our sample is $\frac{5-1}{\binom{5}{2}} = \frac{4}{10} = \frac{2}{5}$.
    - (Extra): In general, if we have n students and want to choose k of them, the probability that one specific student is chosen in our sample is $\frac{\binom{n-1}{k-1}}{\binom{n}{k}}$.

Cluster sample

In cluster sampling, we first split our population into clusters. Then, we use SRS to select clusters themselves, and sample everyone within a cluster.

- For example, if we wanted to survey some number of CS and EECS majors at Berkeley, we could split our population into 1st years, 2nd years, 3rd years and 4th years. We could choose two of these clusters at random (i.e. 1/2, 1/3, 1/4, 2/3, 2/4, 3/4) and then **sample all students** in the two clusters.
- We could also choose just one cluster.

Stratified sample

In stratified sampling, we also split our population into groups (now called strata). However, we now use SRS to sample *within every strata*.

- Continuing with the above example, if we were to proceed with stratified sampling, we would use SRS to select **some** number of 1st, 2nd, 3rd, and 4th years

Multi-stage sample

Multi-stage samples can be thought of as a combination of cluster samples and stratified samples. In a cluster sample, we:

1. Divide our population into clusters
2. Use SRS to select clusters
3. Use SRS within each cluster

- In our example, this would mean first using SRS to select some number of class ranks (e.g. 1st years and 3rd years)
- Then, we would use SRS to sample **some** 1st years, and **some** 3rd years

Recap of the three "grouped" sampling techniques:

- Cluster: Use SRS to select groups, sample everyone within selected groups
- Stratified: Use SRS to sample within every group
- Multi-stage: Use SRS to select groups, use SRS to sample within selected groups