Data 100, Discussion 2 – Sampling, Probability, SQL

Suraj Rampure

Friday, January 31st, 2020

Reminders

All materials will be posted at

surajrampure.com/teaching/ds100.html

Please fill out

tinyurl.com/feedbacksuraj

Lastly, Homework 2 and Lab 1 are both due on Monday.

Agenda

- Sampling
- Probability
- SQL

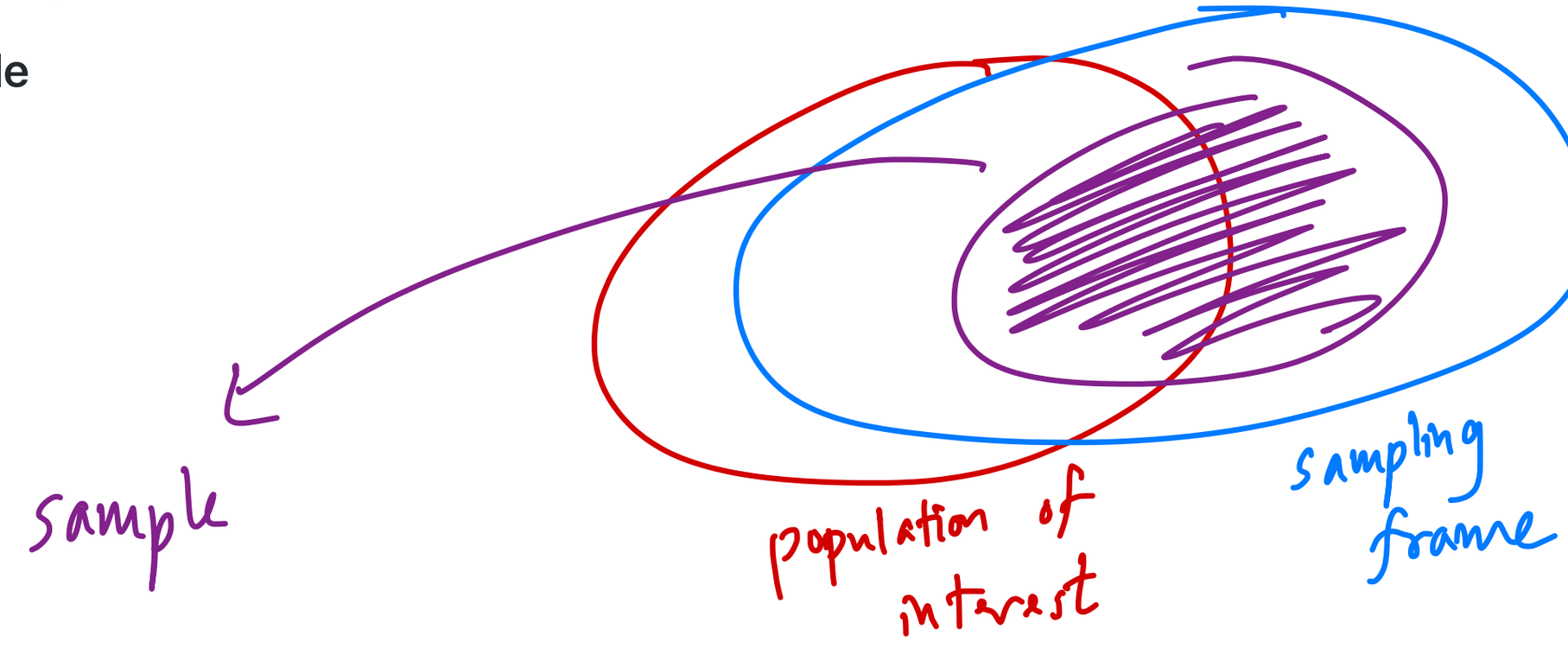
Sampling Terminology

Let's discuss and define the following terms.

Population of Interest → trying to learn about

Sampling Frame → the set of all individuals you can sample from

Sample



Types of Samples

Probability Samples

- Involves randomness.
    - If you were to repeat your sampling process again, you'd likely end up with a different sample.
- Must be able to compute the probability that any individual is part of the sample.
- All individuals in the population need not have the same chance of being selected.

Non-Probability Samples

- Convenience samples.
- Quota samples.

Simple random samples (SRS)

type of probability sample!

Sampling uniformly at random from the population without replacement.

- All equal-sized subsets must have the same chance of being in the sample.
  - **Each element of the population must have the same chance of being in the sample.**
    Note: You should keep in mind what exactly your population is.

e.g. all pairs, all triplets, etc.

- For example, if we have five students, A, B, C, D, and E, and want to select two of them, our sample could 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 $\frac{5-1}{\binom{5}{2}} = \frac{4}{10} = \frac{2}{5}$.

Question: Why use SRS, and probability samples in general? When do we need to?

"Golden Rules" of Probability

Golden rules for finding the chance of an event (taken from the worksheet):

- **List the ways:** list all the distinct ways the event can happen, and add the chances of all the ways. "addition rule"
- If the list above looks long and complicated, make the list of ways in which **the event doesn't happen**; it might be simpler. "complement"
- **If an event involves multiple trials**, like a number of random draws, imagine yourself conducting the experiment one trial at a time.

Binomial Basics

n trials, each has probability p of success
each independent

$$P(k \text{ successes}) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\binom{n}{k} \rightarrow \frac{n!}{k!(n-k)!}$$

"n choose k"

HTHT
HHTT
HTTH
THHT
TTHH
THTH

} 6 ways, each with same probability

Homework 2, Question 2d

Suppose each of five survey organizations takes a sample of voters at random with replacement from the population of voters in Part c*, independently of the samples drawn by the other organizations.

- Three of the organizations use a sample size of 200
- One organization uses a sample size of 300
- One organization uses a sample size of 400

Write an expression that evaluates to the chance that in at least one of the five samples the majority of voters favor Candidate C.

$$P(\text{at least 1 majority}) = 1 - P(\text{no majority})$$

$$= 1 - P(\text{no maj, sample 1}) \cdot P(\text{no maj, sample 2}) \cdot \ldots \cdot P(\text{no maj, sample 5})$$

SQL

SQL is a declarative language, as opposed to the imperative languages you may be used to. You *declare* what you want, not how to find it.

The general format of a SQL query:

```
SELECT <column expression list>
FROM <list of tables>
[WHERE <predicate>]
[GROUP BY <column list>
    [HAVING <predicate>] ]
[ORDER BY <column list>]
[LIMIT <number of rows>];
```