DS 100/200: Principles and Techniques of Data Science Date: January 30, 2020

Discussion #2

Name: Raguvir Kunani

This discussion consists of a quick recap of sampling methods and biases, some tips and examples of how to caclulate probabilities, and some SQL.

As stated in Lecture 1, some time in every discussion will be spent on selected homework problems.

Sampling and Bias

1. A campus organization wants to take a sample of Berkeley students who are registered for classes this semester. To do this, the organization takes a simple random sample of 20 classes from among all classes offered this semester, and then takes all students in those classes. You can assume that the organization has access to complete enrollment information all classes.

   (a) Is this a simple random sample of students? Explain.

   no; people in more classes have a higher chance

   (b) Is this a probability sample of students? Explain.

   yes; chose 20 classes randomly

2. The Current Population Survey is a national survey run by the Census Bureau. It is thorough and reliable, and thus is sometimes used as a benchmark to assess the accuracy of other surveys. As part of an assessment of its own phone surveys, the Pew Research Center found that the response rates have been dropping over the years. Still, on most measures, its estimates were comparable to those of the Current Population Survey. But for example 55% of respondents in the most recent Pew Survey said they did some type of volunteer work for or through an organization in the past year, compared with 27% in the Current Population Survey.

   How do you think this difference might have arisen?

Response Bias : people don't respond truthfully

Non-response Bias : people don't respond ← people who do respond to a PEW survey are more likely to volunteer

Selection Bias : systematic exclusion of certain groups

See my "Probability and Sampling" Resources for a more in-depth explanation of this problem

Finding Chances

Golden rules for finding the chance of an event:

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

3. Let n be a positive integer. Consider a sample of size n drawn at random with replacement from a population in which a proportion p of the individuals are called successes.

   (a) For an integer k such that $0 \le k \le n$, which of the following are equal to the chance of getting exactly k successes in the sample?

   (i) $p^k(1-p)^{n-k}$
   (ii) $\binom{n}{k}p^k(1-p)^{n-k}$
   (iii) $\binom{n}{n-k}p^k(1-p)^{n-k}$
   (iv) $\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$

$\binom{n}{k}$ = "n choose k" = $\frac{n!}{k!(n-k)!}$

   (b) Which of the following are equal to the chance of getting at least one success in the sample?

   (i) $np(1-p)^{n-1}$
   (ii) $\sum_{k=2}^{n}\binom{n}{k}p^k(1-p)^{n-k}$
   (iii) $\sum_{k=1}^{n}\binom{n}{k}p^k(1-p)^{n-k}$
   (iv) $1-p^n$
   (v) $1-(1-p)^n$

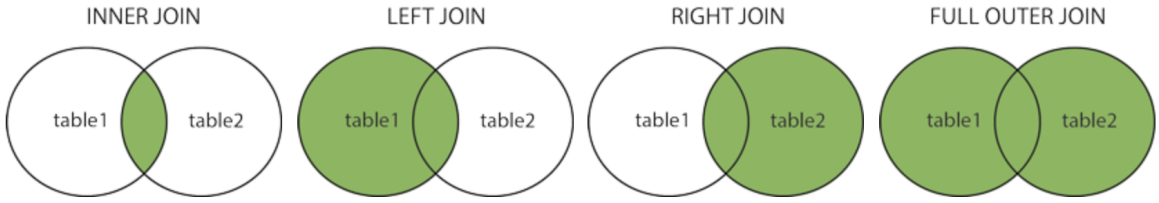
p^n: all successes

$1-p^n$: not all successes ➡ at least 1 failure

$(1-p)^n$: all failures

$1-(1-p)^n$: not all failures ➡ at least 1 success

SQL



Note: You do not always have to use the JOIN keyword to join sql tables. The following are equivalent:

```
SELECT column1, column2
FROM table1, table2
WHERE table1.id = table2.id;

SELECT column1, column2
FROM table1 JOIN table2
ON table1.id = table2.id;
```

4. Describe which records are returned from each type of join in the figure above. How does a cross join relate to these types of joins?

See the joins example from my Discussion 3 slides, it shows the difference between the types of joins with an example.

5. Consider the following real estate schema:

```
Homes(home_id int, city text, bedrooms int, bathrooms int,
area int)
Transactions(home_id int, buyer_id int, seller_id int,
transaction_date date, sale_price int)
Buyers(buyer_id int, name text)
Sellers(seller_id int, name text)
```

Fill in the blanks in the SQL query to find the id and selling price for each home in Berkeley. If the home has not been sold yet, **the price should be NULL**.

```
SELECT h.home_id, t.sale_price
FROM Homes AS h
   LEFT          JOIN  Transactions As t
ON h.home_id = t.home_id
WHERE h.city = "Berkeley";
```

Approach

1) FROM: need to know the selling price (in Transactions)
need to know the location (in Homes)

Both of these tables have a common column home_id, so we can join these 2 tables to get the info we need.

2) JOIN: since home_id is unique, if the home_id of a home in Homes is equal to the home_id of a home in Transactions, these 2 homes must be the same home. We want a LEFT JOIN to make sure the output has a row for each home, regardless of whether the home has been sold. If a home has not been sold, the LEFT JOIN will automatically assign it a NULL price.

3) WHERE: we only want the sale prices of Berkeley homes, so we only consider the homes that have "Berkeley" as its city.

4) SELECT: we only care about the id and price, but we have to specify which tables these columns come from.