DATA 8

Spring 2020

Lecture 15

Sampling

Weekly Goals

- Monday
  - Simulation
  - Chances
- **Today**
  - Methods of sampling
  - Distributions of large random samples
- Friday
  - Models that involve chance
  - Assessing the consistency of the data and the model

Announcements

- HW 5 due tomorrow
- Project 1 due this Friday (tomorrow for a bonus point)
  - If you flagged a submission for checkpoint, reflag it to be graded for the final submission
- Midterm on March 13th, 7PM
  - Scope: up to and including **A/B testing**
  - Review material on Piazza
  - **Review on 03/11, 6-8PM in 10 Evans & 2050 VLSB**
  - Fill out conflict form by **this Friday**

Probability Review

Equally Likely Outcomes

Assuming all outcomes are equally likely, the chance of an event A is:

$$P(A) = \frac{\text{number of outcomes that make A happen}}{\text{total number of outcomes}}$$

Multiplication Rule

Chance that two events *A* and *B* both happen

= P(*A* happens) x P(*B* happens given that *A* has happened)

- The answer is *less than or equal to* each of the two chances being multiplied
- The more conditions you have to satisfy, the less likely you are to satisfy them all

Another Question

- I have three cards: **ace of hearts**, **king of diamonds**, and **queen of spades**.
- I shuffle them and draw two cards *at random without replacement.*

- What is the chance that one of the cards I draw is a King and the other is Queen?

AK AQ KA KQ QA QK

Addition Rule

If event *A* can happen in *exactly one* of two ways, then

$$P(A) = P(\text{first way}) + P(\text{second way})$$

- The answer is *greater than or equal to* the chance of each individual way

Complement: At Least One Head

- Recall: $P(A) = 1 - P(\text{not } A)$
- In 3 tosses:
  - Any outcome *except* TTT
  - $P(TTT) = (1/2) \times (1/2) \times (1/2) = 1/8$
  - $P(\text{at least one head}) = 1 - P(TTT) = 1 - (1/8) = 87.5\%$
- In 10 tosses:
  - 1 - (1/2)**10 ≅ 99.9%

Discussion Question

A population has 100 people, including Rick and Morty.
We sample two people at random without replacement.

(a) P(both Rick and Morty are in the sample)
= P(first Rick, then Morty) + P(first Morty, then Rick)
= (1/100) * (1/99) + (1/100) * (1/99) = 0.0002

(b) P(neither Rick nor Morty is in the sample)
= (98/100) * (97/99) = 0.9602

Sampling

Random Samples

- Deterministic sample:
  - Sampling scheme doesn't involve chance

- Random sample:
  - Before the sample is drawn, you have to know the selection probability of every group of people in the population
  - NOTE: Not all individuals / groups have to have equal chance of being selected

(Demo)

Sample of Convenience

- Example: sample consists of whoever walks by
- Just because you think you're sampling "randomly", doesn't mean you have a random sample.
- If you can't figure out ahead of time
  - what's the population
  - what's the chance of selection, for each group in the population

then you don't have a random sample

Distributions

Probability Distribution

- Random quantity with various possible values
- “Probability distribution”:
  - All the possible values of the quantity
  - The probability of each of those values
- If you can do the math, you can work out the probability distribution without ever simulating it
- But... simulation is often easier!

Empirical Distribution

- “Empirical”: based on observations
- Observations can be from repetitions of an experiment
- “Empirical Distribution”
  - All observed values
  - The proportion of times each value appears

(Demo)

Large Random Samples

Law of Averages / Law of Large Numbers

If a chance experiment is repeated many times,
independently and under the same conditions,
then the proportion of times that an event occurs
gets closer to the theoretical probability of the event

As you increase the number of rolls of a die, the proportion of times you see the face with five spots gets closer to 1/6

Empirical Distribution of a Sample

If the sample size is large,

then the empirical distribution of a uniform random sample

resembles the distribution of the population,

with high probability

(Demo)

A Statistic

Inference

- **Statistical Inference:**
  Making conclusions based on data in random samples
- **Example**:
  fixed
  Use the data to guess the value of an unknown number

  depends on the random sample

  Create an **estimate** of the unknown quantity

Terminology

- **Parameter**
    - A number associated with the population
- **Statistic**
    - A number calculated from the sample

A statistic can be used as an **estimate** of a parameter

(Demo)

Probability Distribution of a Statistic

- Values of a statistic vary because random samples vary
- “Sampling distribution” or “probability distribution” of the statistic:
    - All possible values of the statistic,
    - and all the corresponding probabilities
- Can be hard to calculate
    - Either have to do the math
    - Or have to generate all possible samples and calculate the statistic based on each sample

Empirical Distribution of a Statistic

- Empirical distribution of the statistic:
  - Based on simulated values of the statistic
  - Consists of all the observed values of the statistic,
  - and the proportion of times each value appeared

- Good approximation to the probability distribution of the statistic
  - if the number of repetitions in the simulation is large