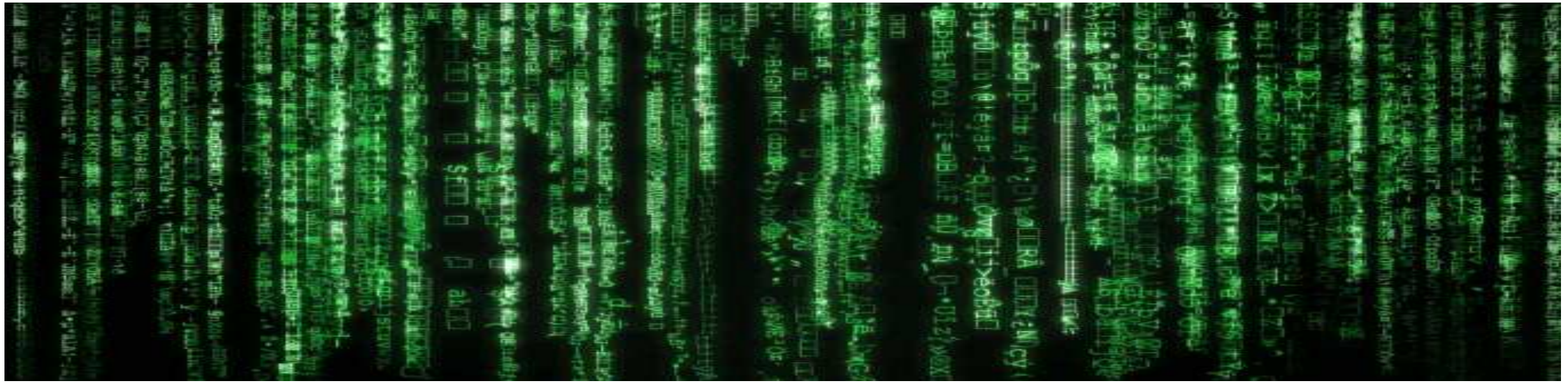

INF 110 Discovering Informatics

More Sampling, and Model Testing

NAU
NORTHERN
ARIZONA
UNIVERSITY

Introduction

- Data scientists often must make conclusions based on random samples.

- Let’s understand what random samples actually are.

Definitions

- ***Deterministic Samples***
  - The sampled elements of a set are specified by ***you.***
  - ***These are not random.***

Live Code Movies Dataset

- Practice deterministic sampling with a table.
  - Each row represents an individual.
  - Each individual is a movie.
  - Sampling individuals can be achieved by sampling the rows of the table.
  - Contents of each row are the values of different variables.
  - These variables are measured on the same individual.
  - The contents of the sampled rows form samples of values of each variable.
- Load a table
- Use .take and .where to choose elements of a set.

Definitions

- ***Deterministic Samples***
  - The sampled elements of a set are specified by ***you.***
  - ***These are not random.***

- ***Important terminology for random samples:***
  - ***Population***
    - the set of all elements from whom a sample will be drawn
  - ***Probability Sample***
    - A sample for which it is possible to calculate the chance with which any subset of elements may enter the sample.
    - Elements do not have to have the same chance of being chosen.

A Random Sampling Scheme

- Choose two people from a population that consists of three people.
- Person A is chosen with probability 1
- One of persons B or C is chosen according to a coin toss:
  - heads choose B, tails choose C
- The probability sample size is 2.
- These are the chances for all non empty subsets:

```
A: 1
B: 1/2
C: 1/2
AB: 1/2
AC: 1/2
BC: 0
ABC: 0
```

Live Code Systematic Sampling

- Imagine the elements of the population as a list.
- Choose a random position early in the list, and then choose evenly spaced positions after that.

Live Code Systematic Sampling

- This is a ***probability sample.***
- In our scheme, all rows have a 1/10 chance of being chosen.
  - Row 23 can only be chosen if Row 3 is chosen first = 1/10.
- Not all subsets have an equal chance of being chosen.
  - selected rows are evenly spaced (every 10)
  - therefore, most subsets of rows have zero chance of being chosen.

Two Types of Random Sampling

1. Sampling ***with*** replacement.
   - This is the default behavior of `np.random.choice`

2. Sampling ***without*** replacement.
   - AKA simple random sampling
   - Sampled individuals are not replaced before the next individual is drawn.
   - Think about when a deal a deck of cards.
   - To use `np.random.choice` for simple random sampling, use `replace=False`

Convenience Sampling

- Drawing random samples is not easy!
- If you stand on a street corner and sample the first ten people that walk by, this is not random!
- You don’t know ahead of time the probability of each person entering the sample.

Live Code Empirical Distributions

- “empirical” = “observed”
- Empirical distributions are distributions of observed data
  - i.e., data in random samples.

Discrete Variables

- When the successive values are separated by a fixed amount.
  - For a die, this separation is 1.
- The histogram showing the uniform distribution of the probabilities of each die roll outcome is a ***discrete histogram***.
  - the array die_bins specifies the bins and ensures each bar is centered over the corresponding integer value.

Live Code Empirical Distributions

- The discrete histogram shows the theoretical probability of each result. This is called a ***probability distribution***.
  - Not based on observed data, you don't need to roll any dice to study it.

- Let's visualize some empirical distributions with empirical histograms.
  - whereas before we have used np.random.choice, we will use a different method, .sample, that makes it a little easier to sample from a table.

Law of Averages

- ***If*** a chance experiment is repeated independently and under identical conditions, ***then*** the proportion of times an event occurs will get closer to the theoretical probability of the event.

- When the experiment is repeated a large number of times, the results will meet the predictions of the theoretical probability distribution.

Live Code Sampling from a Population

- We will study a population of flight delay times.

Lessons from the Empirical Histograms

- For a large random sample, the empirical histogram resembles the histogram of the population.

- USE LARGE RANDOM SAMPLES!

Live Code Empirical Distribution of a Statistic

- We are often interested in numerical quantities associated with a population.
  - In a population of:
    - voters, what percent will vote for Candidate A?
    - Facebook users, what is the largest number of friends someone has?
    - United flights, what is the median delay?
- The quantities are called ***parameters.***

Live Code Empirical Distribution of a Statistic

Simulating a Statistic

Step 1: decide which stat to simulate

Step 2: define a function that returns one simulated value of the stat

Step 3: Decide how many simulated values to generate

Step 4: Use a **for** loop to generate an array of simulated values.

Live Code Assessing a Model

- ***Swain vs Alabama (1965)***
  - Robert Swain, a black man, was convicted in Alabama in 1962.
    - And sentenced to death by an ***all white jury***.
  - He appealed all the way to the Supreme Court
    - main argument: the jury was not impartial, it was unrepresentative.
      - 8% of the 100-person jury panel were black
      - Yet at that time, 26% of men in the county were Black!
  - SCOTUS ruled against Swain 6-3. But should they have?

  - Can we reject a model where the panel was selected at random and ended up with a small number of Black panelists by chance?

Live Code Assessing a Model

1. Simulate based on a model of a randomly selected panel.
2. Generate an empirical distribution of a large number of simulations of randomly selected panels.
3. Visualize this distribution using a histogram.
4. Determine if the observed data fits within this empirical distribution.