COMS BC1016
Introduction to Computational Thinking and Data Science

Lecture 11: Sampling, Distributions, and Models

Oct 8, 2025

Upcoming Schedule

Today

Bring a pencil/pen

Date	Topic	Lab	Assignment
10/6	10 - Simulation and Probability Slides lec10-demo.ipynb		HW4 - Probability, Simulation, Estimation (Due 10/15) Courseworks
10/8	11 - Sampling, Distributions and Models Slides lec11-demo.ipynb	Lab 5 - Simulations (Due 10/10) Courseworks	HW3 Due
10/13	Programming/Python Review		
10/15	Midterm Review	*No Lab*	HW4 Due
10/20	**Midterm Exam**		
10/22	Special Topics - Bias in AI	*No Lab*	

Lecture Outline

- Sampling
  - Distributions
- Assessing Models
  - *Swain v Alabama*

Sampling

Sample

A **sample** is a subset of a population you choose to utilize in your analysis

- Picking samples is a fundamental part of Data Science
  - Did you sample enough / collect enough data?
  - Is the data representative?

Deterministic vs Random Samples

Deterministic Sample: Sampling scheme doesn't involve chance, results are always the same

- Examples:
  - First 100 students when listed in alphabetical order
  - `cat_tbl.where('Coloring', 'tuxedo')`

Name	Age	Weight	Coloring	Sex	Owner
Ruby	14	8	tuxedo	F	Alice
Gertrude	15	12	tuxedo	F	Alice
Hamby	8	16	tabby	M	Bob
Fig	3	7	tabby	F	Bob
Corina	6	10	tortie	F	Carol
Frito	2	8.5	tabby	M	Carol

Deterministic vs Random Samples

Random Sample: Each element has a probability of being chosen

- Selection probabilities for each element are known *before the sample is drawn*
- Not all individuals or groups have to have equal chance of being selected
  - Example: drawing a face card vs a numbered card
- Example: `np.random.choice(np.arange(10))`

Randomly Selecting from Arrays

To select uniformly at random from array `some_array`

- `np.random.choice(some_array)`

To select `n` number of random elements from array `some_array`

- `np.random.choice(some_array, n)`

Randomly Sampling Tables

Returns a table with `n` rows sampled *with replacement* from Table `tbl`

```
- tbl.sample(n)
```

Returns a new table with `n` rows sampled *without replacement* from `tbl`

```
- tbl.sample(n, with_replacement=False)
```

Convenience Sampling

Random sampling requires knowing the probability of selection *ahead of time*

- Not fully controlling selection doesn't necessarily make it 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 it is a **sample of convenience** and not a random sample!

Notebook Demo - Sampling

Motivating Example

Suppose a pharmaceutical company is conducting a clinical trial for a new diabetes drug. They want to draw a random sample of patients from the general population to test how effective the drug is.

- Goal of random sampling is to make sure results can be generalized to the whole population
    - To do this accurately, we need to know the selection probability of every subgroup (e.g., age, gender, race, socioeconomic status, …) in the population

What can go wrong?

Suppose the company only recruits participants from wealthier urban neighborhoods. This may unintentionally:

- Exclude low-income individuals
- Underrepresent certain racial or ethnic groups
- Miss patients in rural areas
- Oversample people with access to private healthcare

Overall sample is *biased* because selection probability was effectively zero for certain groups (they didn't have the chance to be included)

- Claims about the drug apply to the people sampled but not necessarily the population as a whole

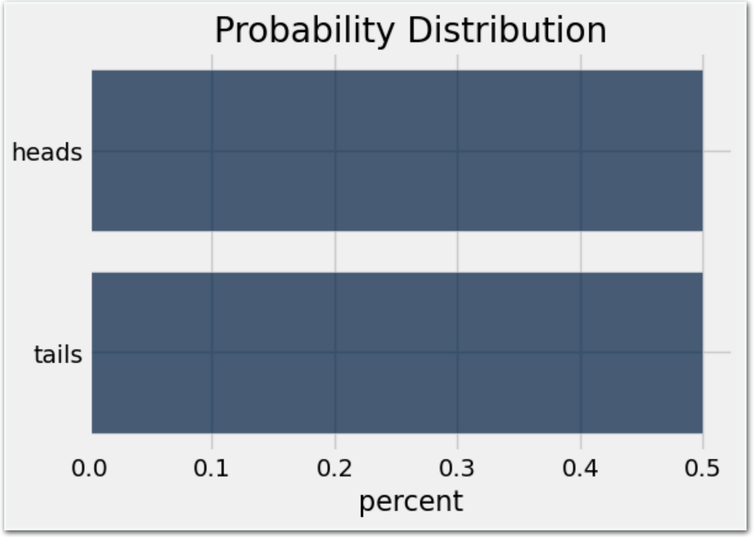
Distributions

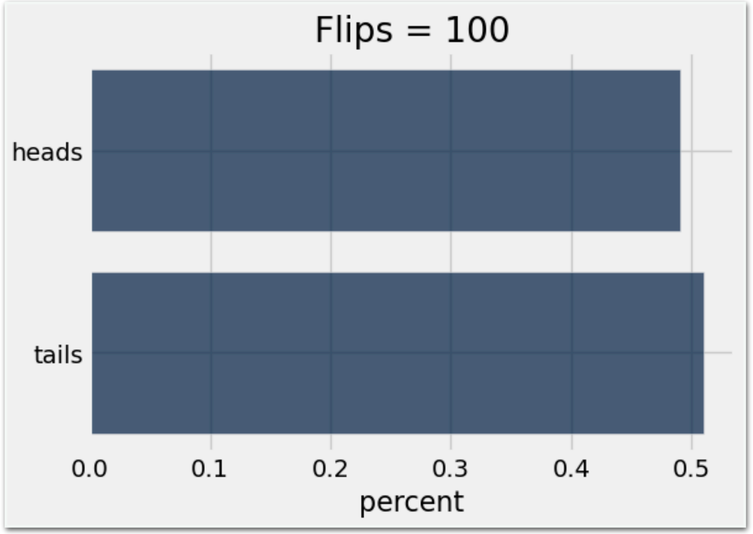
Distributions

- Recall we did categorical distributions (how often each unique value appears)
    - If we have a population, we can get a list of unique values and how often they appear
    - If we take a sample, the list of unique values may change based on the sample
        - There can be differences in what’s in the population vs what we see based on taking samples

Distributions

- **Probability Distribution**: All possible values & the probability of each value
  - Example: 50% heads, 50% tails
- **Empirical Distribution**: The observed results (values and outcomes) of an experiment
  - Example: I flipped 100 coins and N of them were heads and rest were tails

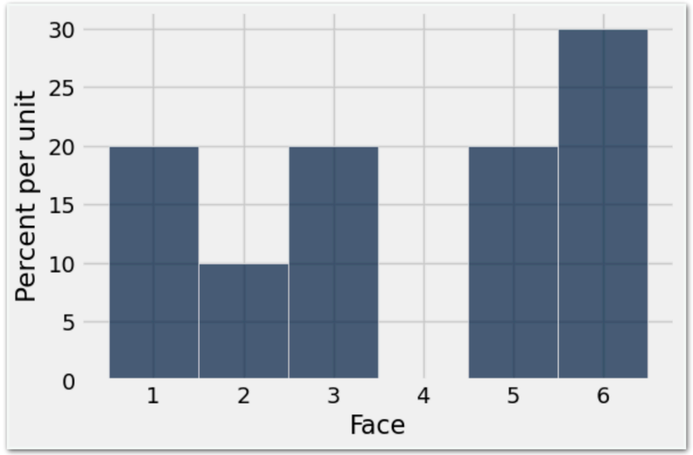




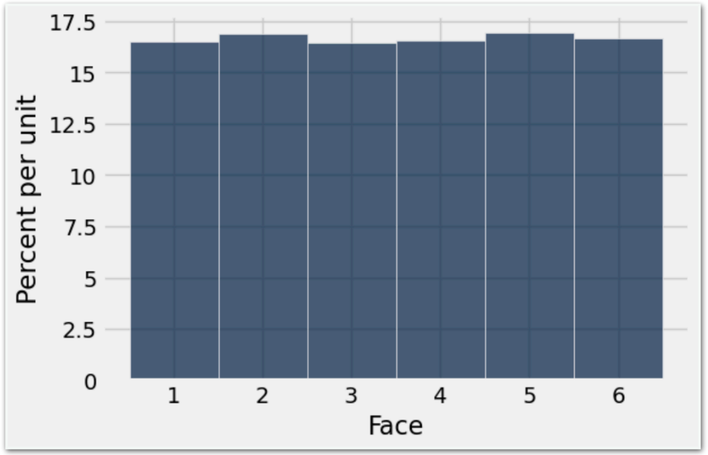

```
def emperical_prob(cnt):
    return cnt/flips

toss = make_array('heads', 'tails')
outcomes = np.random.choice(toss, flips)
results = Table().with_column('Coin Flip', outcomes).group('Coin Flip')
results = results.with_column('count', results.apply(emperical_prob, 'count'))
results = results.relabel('count', 'percent')
results.barh('Coin Flip')
plots.title('Flips = ' + str(flips))
```

Law of Averages / Large Numbers

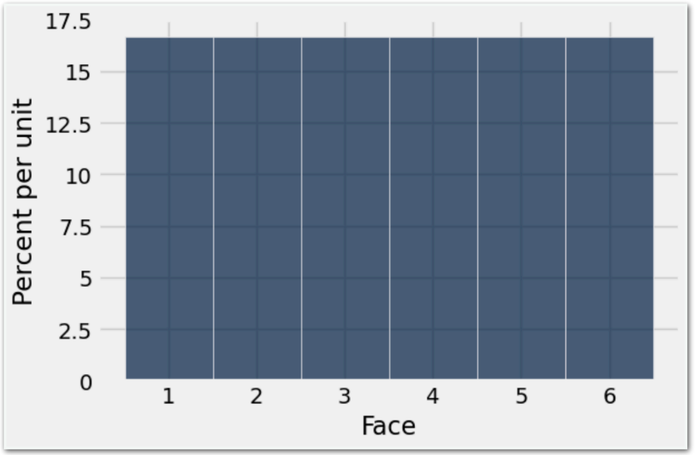
If a chance experiment is repeated many times, independently and under the same conditions, then the proportion of times that an event occurs gets closer to the true probability of the event



Empirical Distribution of 10 rolls



Empirical Distribution of 10,000 rolls



Probability Distribution for Rolling Dice

Terminology

Parameter: Number associated with the population

- Example: average, max, min, mean

Statistic: A number calculated from the sample, can be used to describe the distribution

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

- Example: sample mean, sample max, sample min

Statistical Inference

Statistical Inference: drawing conclusions based on data in random samples

- Create an estimate of an unknown value using sample data and statistics
- Inference occurs from not being able to know an entire population
  - Estimates change based on the sample you draw
  - Statistics help you measure how much you expect those differences to vary

Mean vs Median

- Mean is the average
  - Sum of all the elements divided by the number of elements
- Median is the “middle value”
  - Value that separates the lower half and higher half of a sample

1, 3, 3, 6, 7, 8, 9

$$\text{mean} = \frac{1+3+3+6+7+8+9}{7}$$

$$= \frac{37}{7}$$

$$= 5.28$$

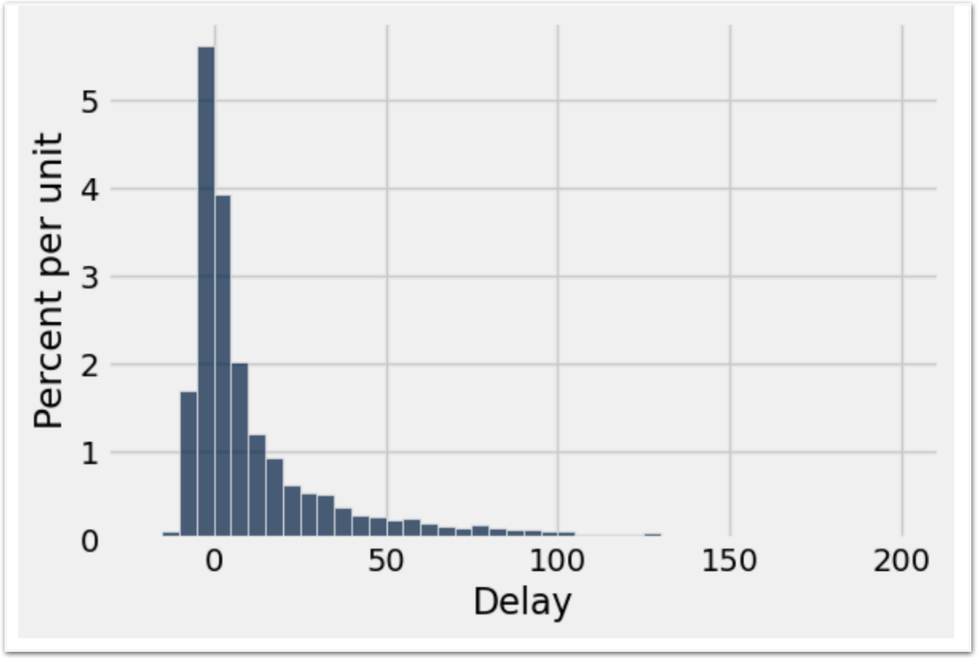
$$\text{median} = 6$$

United Flights Example

United Flight Delays

- If the distribution isn't centered, we expect the median to either fall to the left / right of the mean based on the skew

- Since delays skews to the left, we expect the median to be less than the mean



Probability vs Empirical Distribution of a Statistic

For every sample size, a statistic has both a probability and empirical distribution

- **Probability Distribution**: All possible values and their associated probabilities

- **Empirical Distribution**: Based on simulated values of a statistic and include all observed values along with the proportion of times the value appeared

    - Can approximate probability distribution if number of repetition is high

Assessing Models

Models

A model is a set of assumptions about data

- In data science, many models involve assumptions about the processes that involve randomness

- Question: Does the model fit the data?

Assessing Models

- Suppose we have a statistical model that describes how data should behave (based on certain assumptions)
    - If we can use that model to **generate (simulate)** fake data, then we can see what the model “thinks” the data should look like
    - By simulating data, we can see the kinds of outcomes or patterns the model expects, i.e., its **predictions**
- We can then compare the predictions to the data that were observed (irl data)
- If the data and the model’s predictions are not consistent, that is evidence against the model.

1960s Supreme Court Case: *Swain v Alabama*

Amendment VI of the U.S. Constitution:

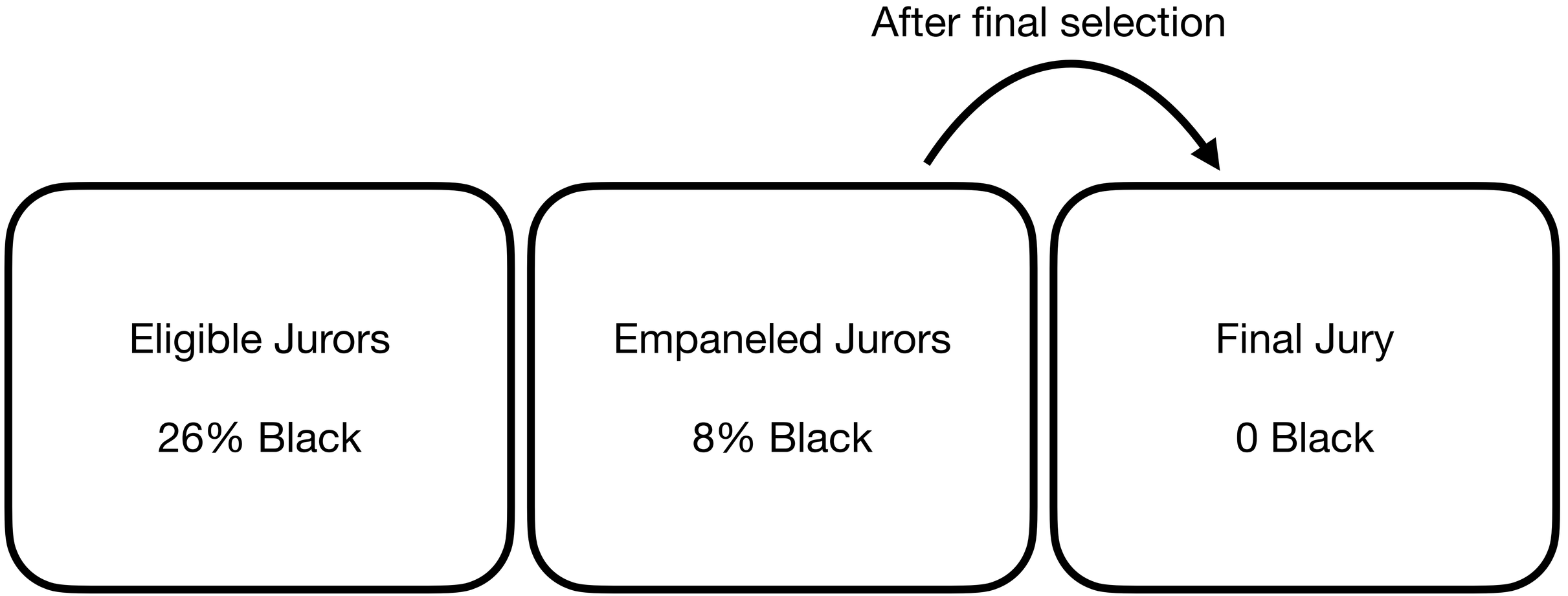
"In all criminal prosecutions, the accused shall enjoy the right to a speedy and public trial, by an impartial jury of the State and district wherein the crime shall have been committed"

- An *impartial* jury should be selected from a panel that is representative from the population of the relevant region

Robert Swain was a Black man indicted and convicted by an all white jury in Talladega County, Alabama

- He appealed to the Supreme Court due to the lack of representation on juries in Talladega County

Swain vs. Alabama Jury



The Appeal

- Swain’s argument: the juries in Talladega County were not representative of the population and were therefore unfair

- Supreme Court Decision:

  - “The overall percentage disparity has been small” (between 26% and 8%) - deemed not different enough to indicate that Black panelists were systematically excluded

Our Question

- Would an 8% black jury be a realistic outcome if jury section were truly unbiased?

- In this case - the distribution we care about is categorical

  - % black vs % non-black

Sampling from a Categorical Distribution

- Sample at random from a categorical distribution

  – `sample_proportions(sample_size, pop_distribution)`

- `pop_distribution` is a list or array that adds up to 1

- Function returns an array containing the empirical distribution of the categories in the sample

Steps in Assessing a Model

1. Choose a statistic that will help you decide whether the data supports the model or an alternative view of the world

2. Simulate the statistic under the assumptions of the model

3. Draw a histogram of the simulated values
   - This is the model's prediction for how the statistic should come out

4. Compute the statistic from the sample in the study
   - If the two are not consistent: evidence against the model
   - If the two are consistent: data supports the model *so far*

Assessing *Swain v Alabama*

1. Choose a statistic that will help you decide whether the data supports the model or an alternative view of the world

2. Simulate the statistic under the assumptions of the model

3. Draw a histogram of the simulated values

4. Compute the statistic from the sample in the study

Model: Panelists were selected at random and the small number of Black panelists is by chance

Alternative view: too few Black panelists for it to have been a random sample

Statistic: Number (count) of Black panelists

Notebook Demo: *Swain vs. Alabama*

Our Conclusion

- Percent of Black panelists (8%) was **highly unlikely** under random sampling
  - It's *possible*, but extremely unlikely
  - Suggests the assumptions about the model are wrong
- We observed **statistical bias**, when differences between the parameters and the statistics are systematically in one direction
- The verdict was eventually overruled in *Baton v. Kentucky* (1986)
  - "The dismissal of jurors without stating a valid cause for doing so—may not be used to exclude jurors based solely on their race."

Upcoming Schedule

Next Week

Date	Topic	Lab	Assignment
10/6	10 - Simulation and Probability Slides lec10-demo.ipynb		HW4 - Probability, Simulation, Estimation (Due 10/15) Courseworks
10/8	11 - Sampling, Distributions and Models Slides lec11-demo.ipynb	Lab 5 - Simulations (Due 10/10) Courseworks	HW3 Due
10/13	Programming/Python Review		
10/15	Midterm Review	*No Lab*	HW4 Due
10/20	**Midterm Exam**		
10/22	Special Topics - Bias in AI	*No Lab*	