COMS BC1016
Introduction to Computational Thinking and Data Science

Lecture 12: Probability and Sampling

March 2, 2026

Upcoming Labs and HW4

- **There are no labs this week or next week**
  - Labs resume after spring recess (March 25/26)
- **HW 4 is due next week Monday**
  - It was released last week Monday, but we're giving you all a little extra time to work on it
  - HW 5 will not be released until after spring recess

Midterm Logistics

- Midterm is **next week** on Wednesday, March 11
- You are permitted to bring a single formula sheet (5”x8” index card, double-sided) that will be submitted along with the exam
  - Exam is otherwise closed-note and no computers
- **Exam will only cover material up until this lecture**
  - Any new material introduced after this lecture will not be on the exam
- There will be a review session during class on Monday, March 9

Probability

Probability

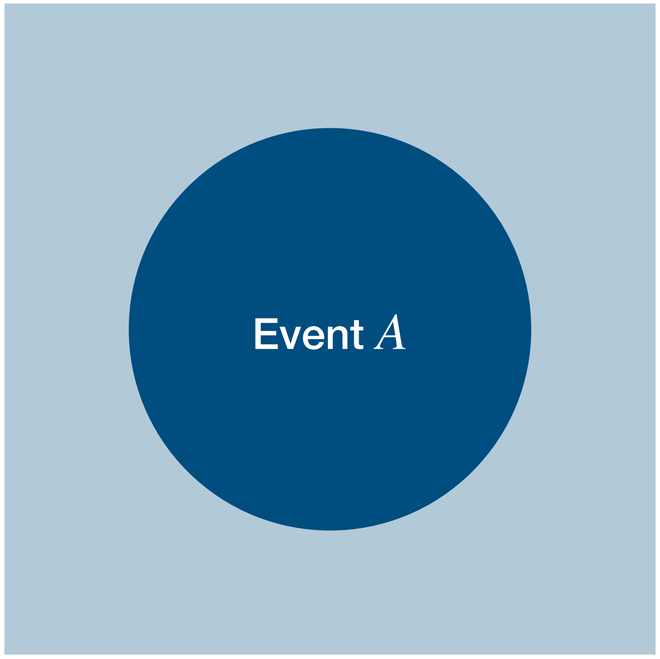
$\mathrm{P}(A) =$ Probability of event A happening

- $\mathrm{P}(A) = 0$
  - 0% chance of event A happening
  - Event A is impossible
- $\mathrm{P}(A) = 1$
  - 100% chance of event A happening
  - Event A is certain

Complements

If an event has a chance of happening N, then the chance it *doesn't* happen is 1-N

- e.g., if chance of happening is 70%, chance of not happening is 30%



Equally Likely Outcomes

Assuming all outcomes are equally likely:

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

Exercise A

- 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's the chance that I get the Queen followed by the King?

Exercise A

What's the chance that I get the Queen followed by the King?

- Let A be event "Queen then King"

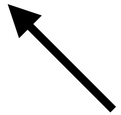
$$\mathrm{P}(A) = \frac{\text{number of outcomes that make } A \text{ happen}}{\text{total number of outcomes}}$$

Exercise A

What's the chance that I get the Queen followed by the King?

- Let A be event "Queen then King"

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



6

Draw 1	Draw 2
Ace	Queen
Ace	King
Queen	King
Queen	Ace
King	Ace
King	Queen

Exercise A

What's the chance that I get the Queen followed by the King?

- Let A be event "Queen then King"

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

1

6

Draw 1	Draw 2
Ace	Queen
Ace	King
Queen	King
Queen	Ace
King	Ace
King	Queen

Exercise A

What's the chance that I get the Queen followed by the King?

- Let A be event "Queen then King"

$$\mathrm{P}(A) = \frac{\text{number of outcomes that make } A \text{ happen}}{\text{total number of outcomes}} \quad (1 \text{ over the numerator}, 6 \text{ over the denominator})$$

$$= \frac{1}{6}$$

Draw 1	Draw 2
Ace	Queen
Ace	King
Queen	King
Queen	Ace
King	Ace
King	Queen

Multiplication Rule

The chance that two events A and B both happen:

$$\text{P}(A \text{ and } B) = \text{P}(A) \times \text{P}(B \text{ happens given } A \text{ happens})$$

Exercise A (another way)

What's the chance that I get the Queen followed by the King?

- Let A be event "Queen in the first draw"

- Let B be event "King in the second draw"

P(A and B) = P(A) × P(B happens given A happens)

Exercise A (another way)

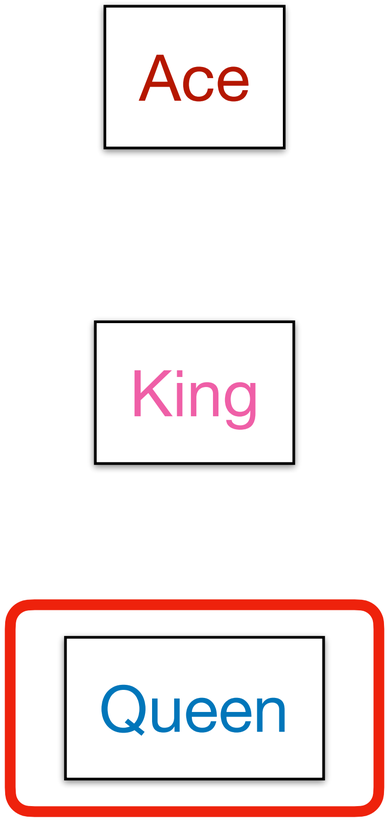
What's the chance that I get the Queen followed by the King?

- Let A be event "Queen in the first draw"

- Let B be event "King in the second draw"

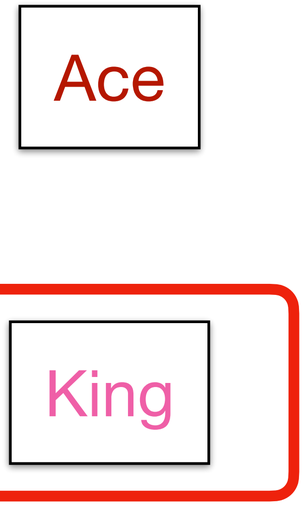
$\mathrm{P}(A \text{ and } B) = \mathrm{P}(A) \times \mathrm{P}(B \text{ happens given } A \text{ happens})$

$\frac{1}{3}$



Exercise A (another way)

What's the chance that I get the Queen followed by the King?



- Let A be event "Queen in the first draw"

- Let B be event "King in the second draw"

$$\mathrm{P}(A \text{ and } B) = \mathrm{P}(A) \times \mathrm{P}(B \text{ happens given } A \text{ happens})$$

$$\frac{1}{3} \qquad \frac{1}{2}$$

Multiplication Rule

The chance that two events A and B both happen:

$$\text{P}(A \text{ and } B) = \text{P}(A) \times \text{P}(B \text{ happens given } A \text{ happens})$$

- 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

Addition Rule

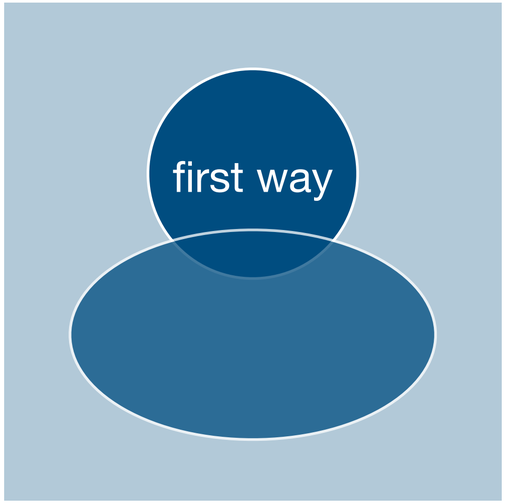
- If event A can happen in exactly one of two ways, then

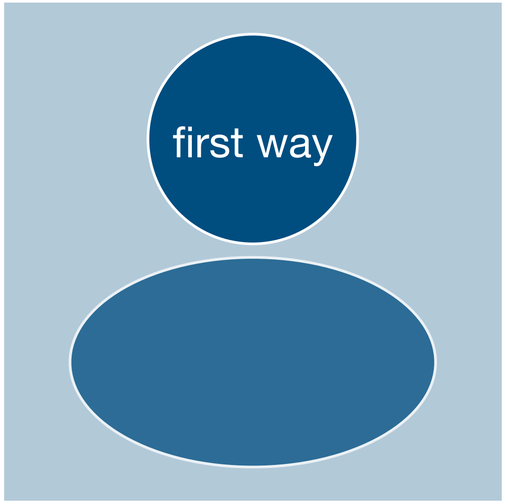
$$\mathrm{P}(A) \leq \mathrm{P}(\text{first way}) + \mathrm{P}(\text{second way})$$

- If the two ways are independent (i.e., no overlap), then

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

In this class, we'll mostly deal with independent events





Exercise B

- 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's the chance that one is a Queen *and* one is a King?

Exercise B: Equal Probabilities

What's the chance that one is the Queen and one is the King?

- Let A be drawing a Queen and King

Draw 1	Draw 2
Ace	Queen
Ace	King
Queen	King
Queen	Ace
King	Ace
King	Queen

Exercise B: Equal Probabilities

What's the chance that one is the Queen and one is the King?

- Let A be drawing a Queen and King

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

Draw 1	Draw 2
Ace	Queen
Ace	King
Queen	King
Queen	Ace
King	Ace
King	Queen

Exercise B: Equal Probabilities

What's the chance that one is the Queen and one is the King?

- Let A be drawing a Queen and King

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

$$= \frac{2}{6}$$

Draw 1	Draw 2
Ace	Queen
Ace	King
Queen	King
Queen	Ace
King	Ace
King	Queen

Exercise B: Equal Probabilities

What's the chance that one is the Queen and one is the King?

- Let A be drawing a Queen and King

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

$$= \frac{2}{6}$$

$$= \frac{1}{3}$$

Draw 1	Draw 2
Ace	Queen
Ace	King
Queen	King
Queen	Ace
King	Ace
King	Queen

Exercise B: Addition Rule

What's the chance that one is the Queen and one is the King?

- Let A be drawing a Queen and King

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

Draw 1	Draw 2
Ace	Queen
Ace	King
Queen	King
Queen	Ace
King	Ace
King	Queen

Exercise B: Addition Rule

What's the chance that one is the Queen and one is the King?

- Let A be drawing a Queen and King

$$\begin{aligned} \mathrm{P}(A) &= \mathrm{P}(\text{first way}) + \mathrm{P}(\text{second way}) \\ &= \mathrm{P}(\text{Queen then King}) + \mathrm{P}(\text{King then Queen}) \end{aligned}$$

Draw 1	Draw 2
Ace	Queen
Ace	King
Queen	King
Queen	Ace
King	Ace
King	Queen

Exercise B: Addition Rule

What's the chance that one is the Queen and one is the King?

- Let A be drawing a Queen and King

$$\begin{aligned} \mathrm{P}(A) &= \mathrm{P}(\text{first way}) + \mathrm{P}(\text{second way}) \\ &= \mathrm{P}(\text{Queen then King}) + \mathrm{P}(\text{King then Queen}) \\ &= \frac{1}{6} + \frac{1}{6} \end{aligned}$$

Draw 1	Draw 2
Ace	Queen
Ace	King
Queen	King
Queen	Ace
King	Ace
King	Queen

Exercise B: Addition Rule

What's the chance that one is the Queen and one is the King?

- Let A be drawing a Queen and King

$$
\begin{aligned}
\mathrm{P}(A) &= \mathrm{P}(\text{first way}) + \mathrm{P}(\text{second way}) \\
&= \mathrm{P}(\text{Queen then King}) + \mathrm{P}(\text{King then Queen}) \\
&= \frac{1}{6} + \frac{1}{6} \\
&= \frac{1}{3}
\end{aligned}
$$

Draw 1	Draw 2
Ace	Queen
Ace	King
Queen	King
Queen	Ace
King	Ace
King	Queen

Exercise

A population has 100 cats including Ruby and Gertrude.
We sample 2 cats at random without replacement.

What are the following probabilities?

1. P(neither Ruby nor Gertrude are in the sample)

2. P(both Ruby and Gertrude are in the sample)

Pictured: Ruby and Gertrude

Exercise

A population has 100 cats including Ruby and Gertrude.
We sample 2 cats at random without replacement.

1. P(neither Ruby nor Gertrude are in the sample)

P(A and B) = P(A) $\times$ P(B happens given A happens)

Exercise

A population has 100 cats including Ruby and Gertrude.
We sample 2 cats at random without replacement.

1. P(neither Ruby nor Gertrude are in the sample)

First pick:
100 cats
98 are not Ruby or Gertrude
↳ $P(A) = \frac{98}{100}$

$$P(A \text{ and } B) = P(A) \times P(B \text{ happens given } A \text{ happens})$$

A: Gertrude and Ruby are not chosen in the first pick

Exercise

A population has 100 cats including Ruby and Gertrude.
We sample 2 cats at random without replacement.

1. P(neither Ruby nor Gertrude are in the sample)

$$\text{P}(A \text{ and } B) = \text{P}(A) \times \text{P}(B \text{ happens given } A \text{ happens})$$

A: Gertrude and Ruby are not chosen in the first pick

B: Gertrude and Ruby are not chosen in the second pick

First pick:
100 cats
98 are not Ruby or Gertrude
↳ $P(A) = \frac{98}{100}$

Second pick:
99 cats after A (removed 1)
97 are not Ruby or Gertrude
↳ $P(B \text{ given } A) = \frac{97}{99}$

Exercise

A population has 100 cats including Ruby and Gertrude.
We sample 2 cats at random without replacement.

1. P(neither Ruby nor Gertrude are in the sample)

First pick:
100 cats
98 are not Ruby or Gertrude
↳ $P(A) = \frac{98}{100}$

$$P(A \text{ and } B) = P(A) \times P(B \text{ happens given } A \text{ happens})$$
$$= \frac{98}{100} \times \frac{97}{99}$$
$$= 0.96$$

Second pick:
99 cats after A (removed 1)
97 are not Ruby or Gertrude
↳ $P(B \text{ given } A) = \frac{97}{99}$

Exercise

A population has 100 cats including Ruby and Gertrude.
We sample 2 cats at random without replacement.

2. P(both Ruby and Gertrude are in the sample)

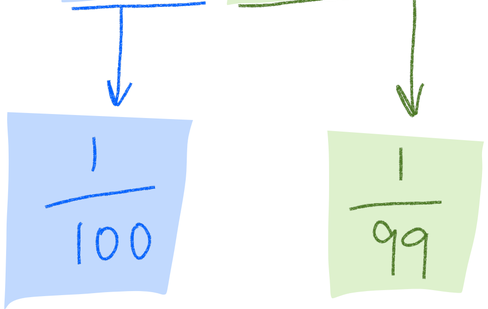
$\mathrm{P}(A) = \mathrm{P}(\text{first way}) + \mathrm{P}(\text{second way})$

Exercise

A population has 100 cats including Ruby and Gertrude.
We sample 2 cats at random without replacement.

2. P(both Ruby and Gertrude are in the sample)

$$\begin{aligned} \mathrm{P}(A) &= \mathrm{P}(\text{first way}) + \mathrm{P}(\text{second way}) \\ &= \mathrm{P}(\text{Ruby then Gertrude}) + \mathrm{P}(\text{Gertrude then Ruby}) \end{aligned}$$

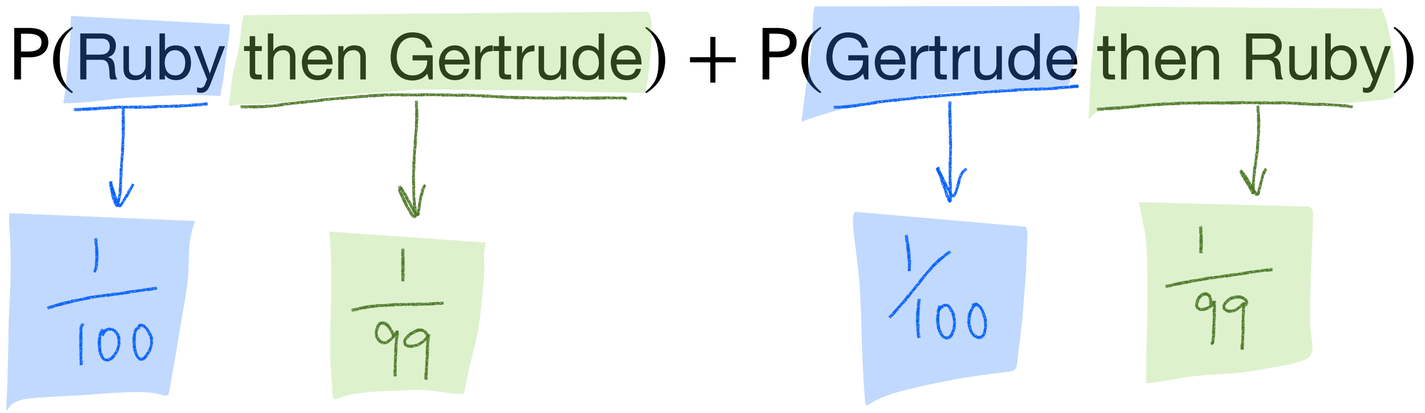


Exercise

A population has 100 cats including Ruby and Gertrude.
We sample 2 cats at random without replacement.

2. P(both Ruby and Gertrude are in the sample)

$$\begin{aligned} \mathrm{P}(A) &= \mathrm{P}(\text{first way}) + \mathrm{P}(\text{second way}) \\ &= \mathrm{P}(\text{Ruby then Gertrude}) + \mathrm{P}(\text{Gertrude then Ruby}) \end{aligned}$$



Exercise

A population has 100 cats including Ruby and Gertrude.
We sample 2 cats at random without replacement.

2. P(both Ruby and Gertrude are in the sample)

$$
\begin{aligned}
\mathrm{P}(A) &= \mathrm{P}(\text{first way}) + \mathrm{P}(\text{second way}) \\
&= \mathrm{P}(\text{Ruby then Gertrude}) + \mathrm{P}(\text{Gertrude then Ruby}) \\
&= \frac{1}{100} \times \frac{1}{99} + \frac{1}{100} \times \frac{1}{99} \\
&= 0.0002
\end{aligned}
$$

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!

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 categorical distributions (how often each unique value appears)

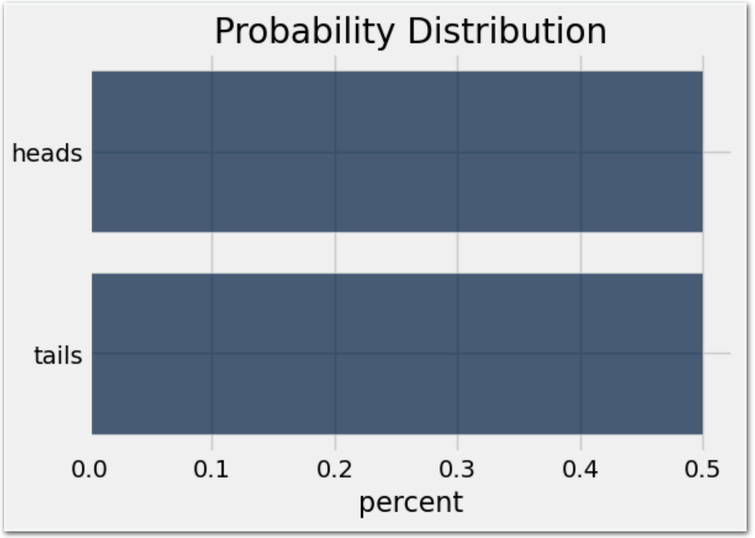
- If we have a population, we can get a list of unique values and how often they appear

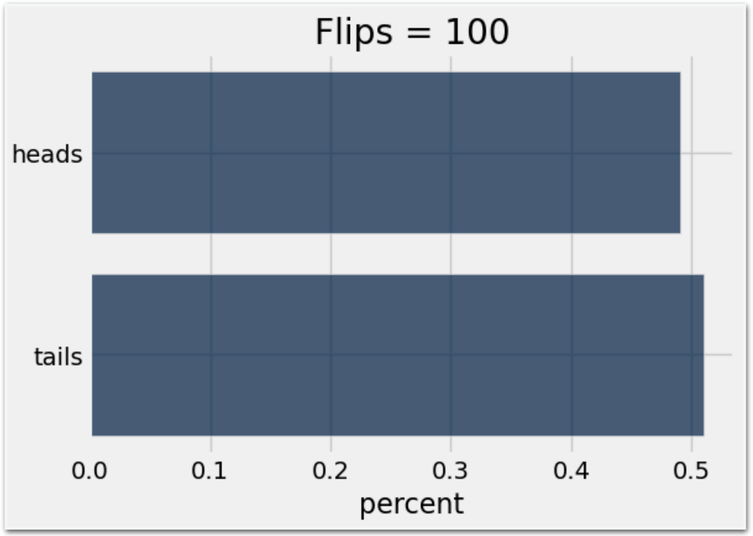
- If we take a sample, the list of unique values may change based on the sample

    - 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

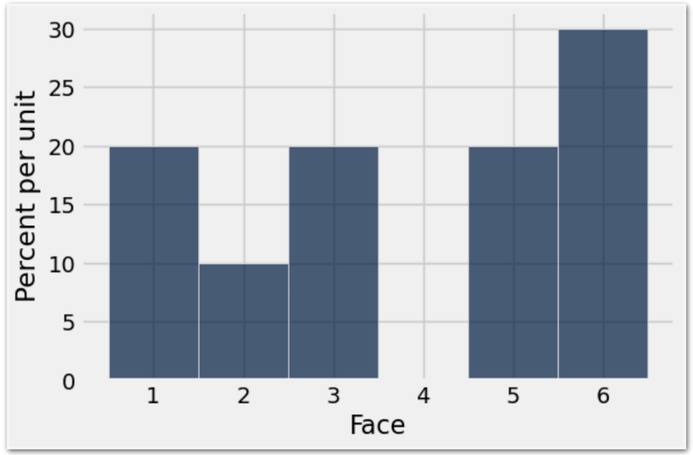




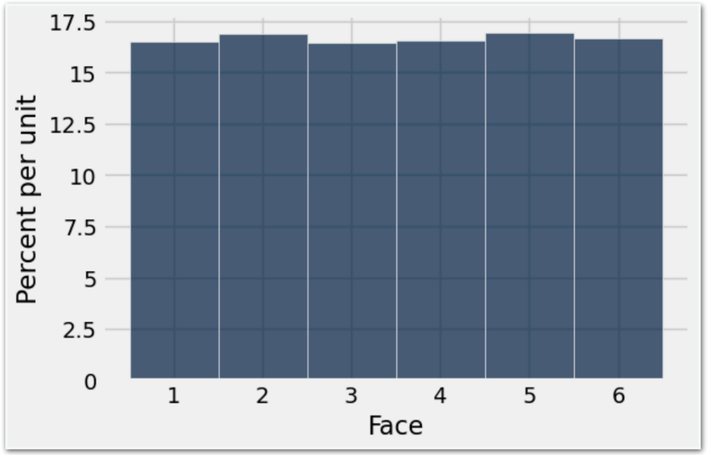

```
def emperical_prob(cnt):
    return cnt/flips

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

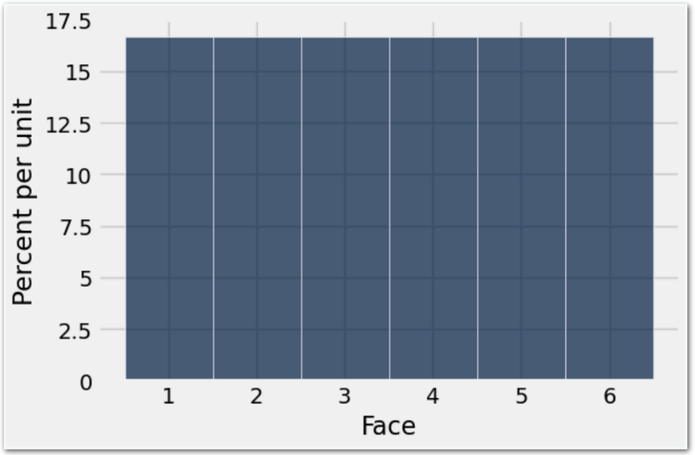
If a chance experiment is repeated many times, independently and under the same conditions, then the proportion of times that an event occurs gets closer to the true probability of the event



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

Next Class

- Today
  - Probability Review
  - Sampling
  - Law of Large Numbers / Law of Averages
- Wednesday
  - Assessing Models