

SEIS 631

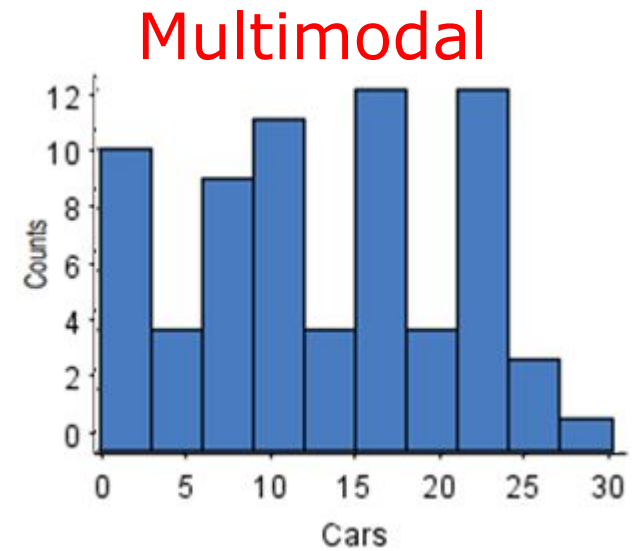
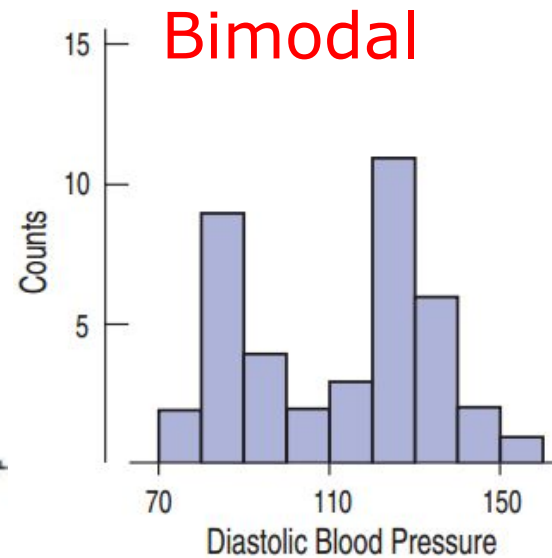
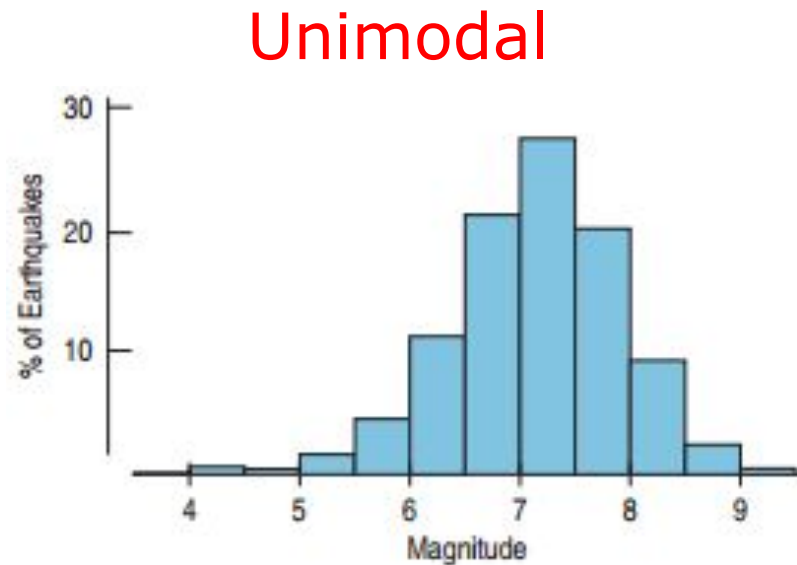
Foundations of Data Analysis



Shapes of Distributions

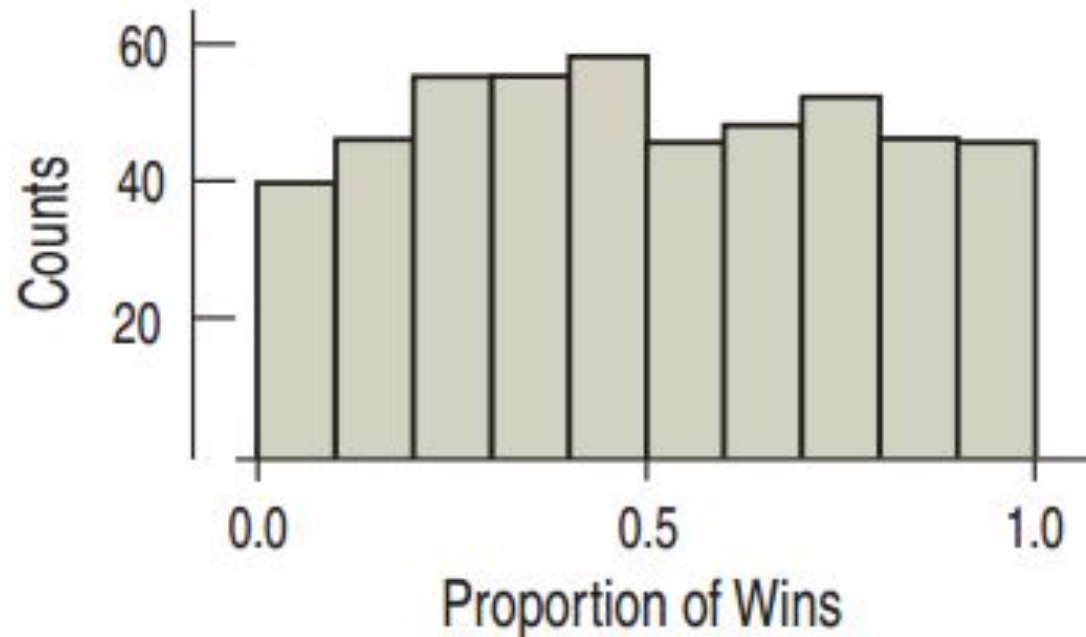
Modes

- A Mode of a histogram is a hump or high-frequency bin.
 - One mode → Unimodal
 - Two modes → Bimodal
 - 3 or more → Multimodal



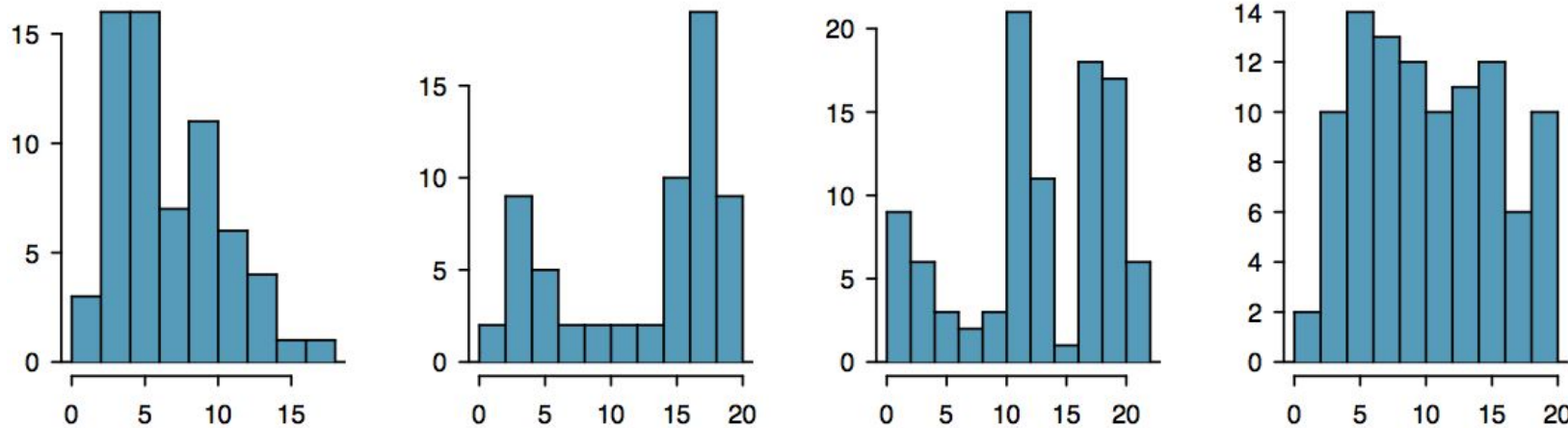
Uniform Distributions

- **Uniform Distribution:** All the bins have the same frequency, or at least close to the same frequency.
- The histogram for a uniform distribution will be **flat**.



Shape of a Distribution: Modality

Does the histogram have a single prominent peak (unimodal), several prominent peaks (bimodal/multimodal), or no apparent peaks (uniform)?

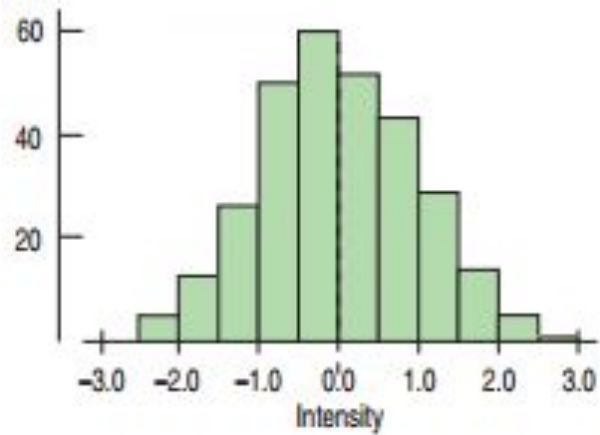


Note: In order to determine modality, step back and imagine a smooth curve over the histogram

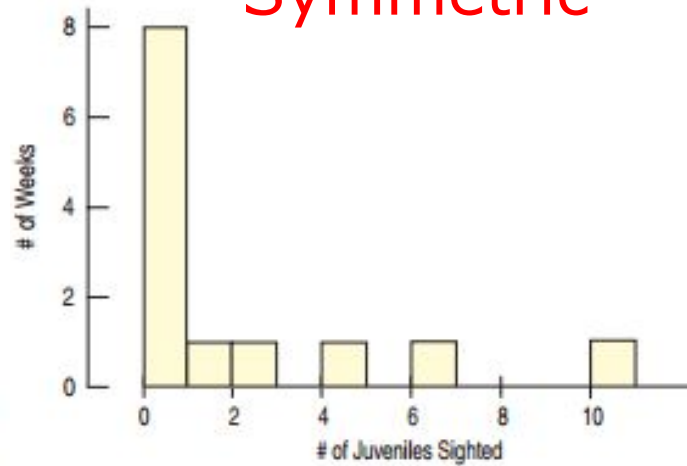
Symmetry

- The histogram for a **symmetric** distribution will look the same on the left and the right of its center.

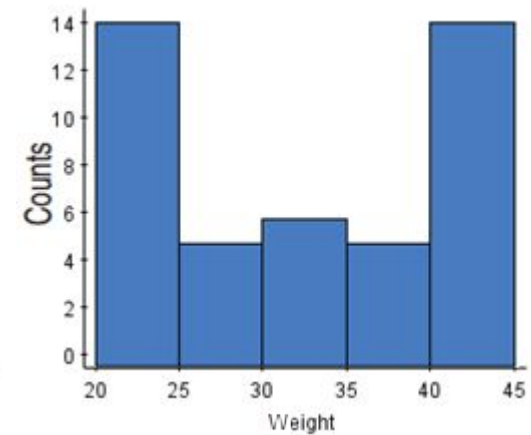
Symmetric



Not
Symmetric

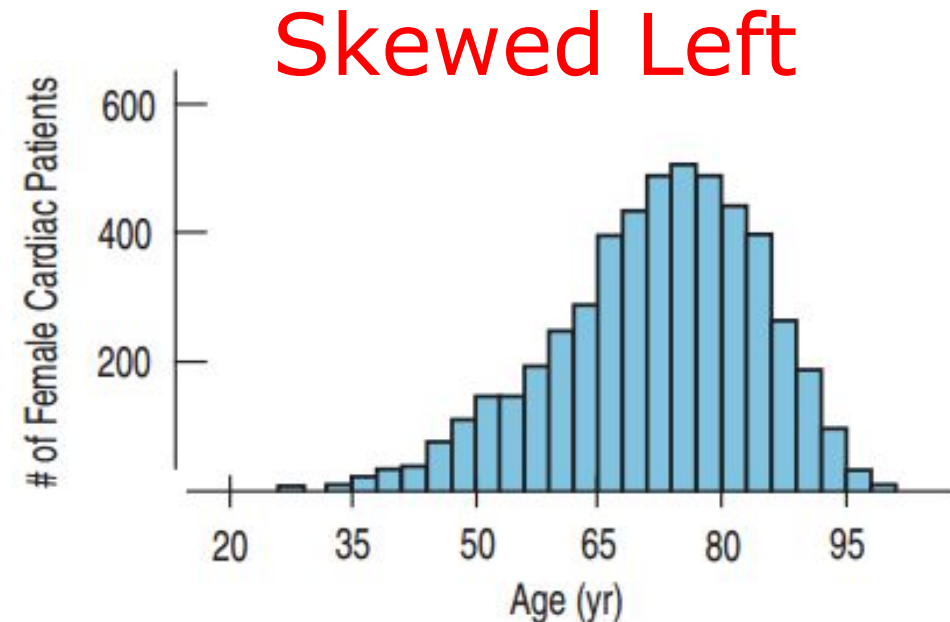
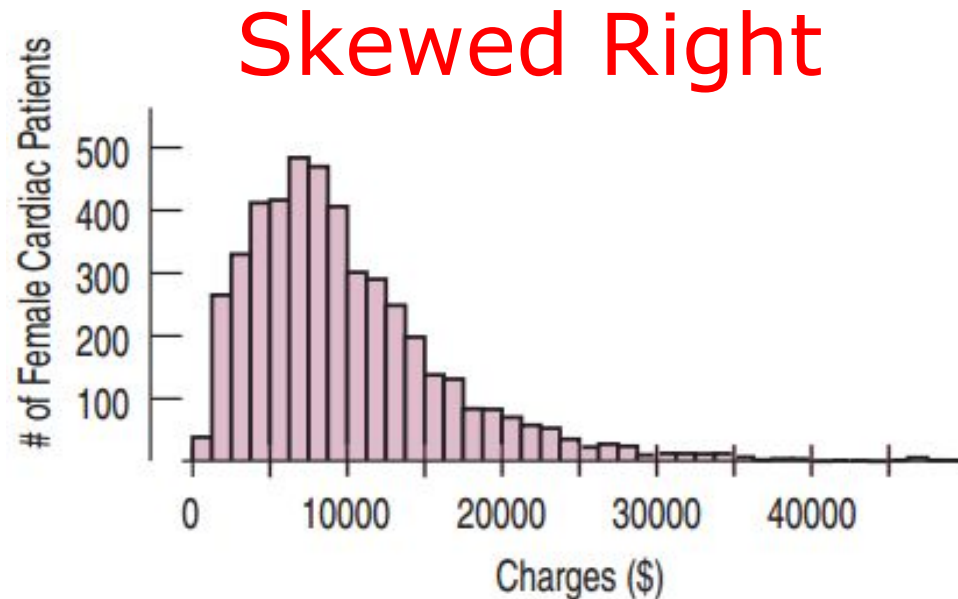


Symmetric



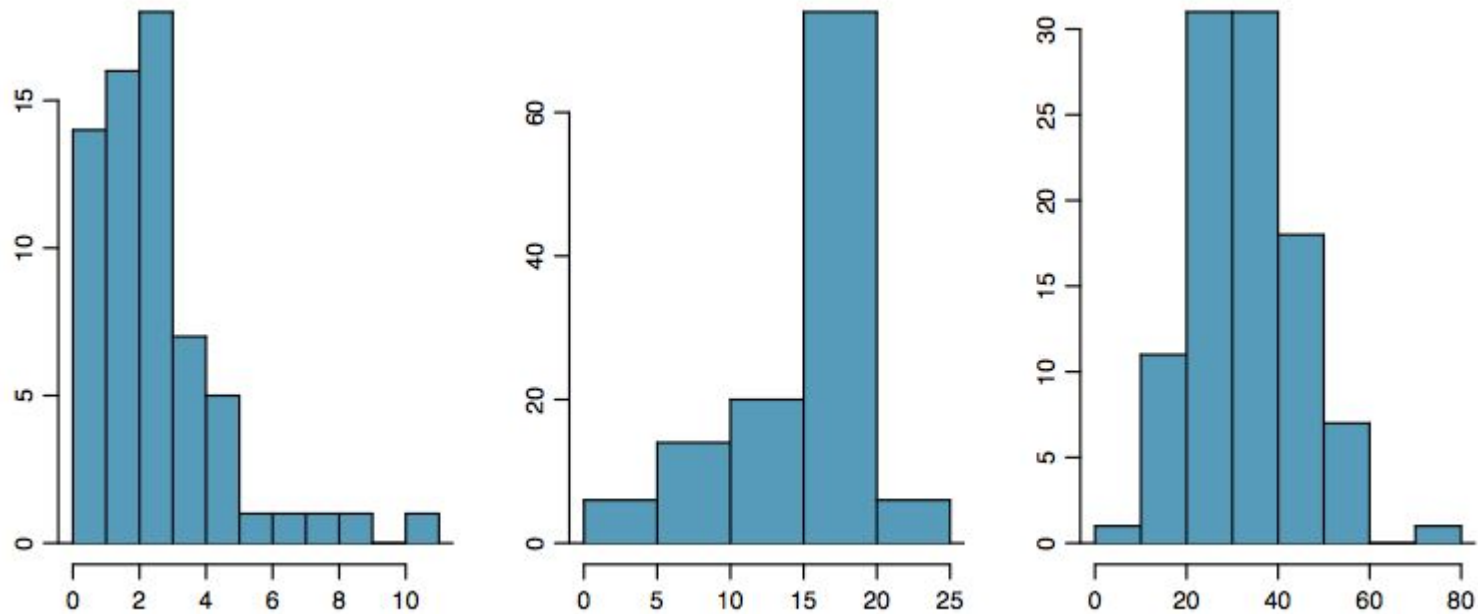
Skew

- A histogram is **skewed right** if the longer tail is on the right side of the mode.
- A histogram is **skewed left** if the longer tail is on the left side of the mode.



Shape of a Distribution: Skewness

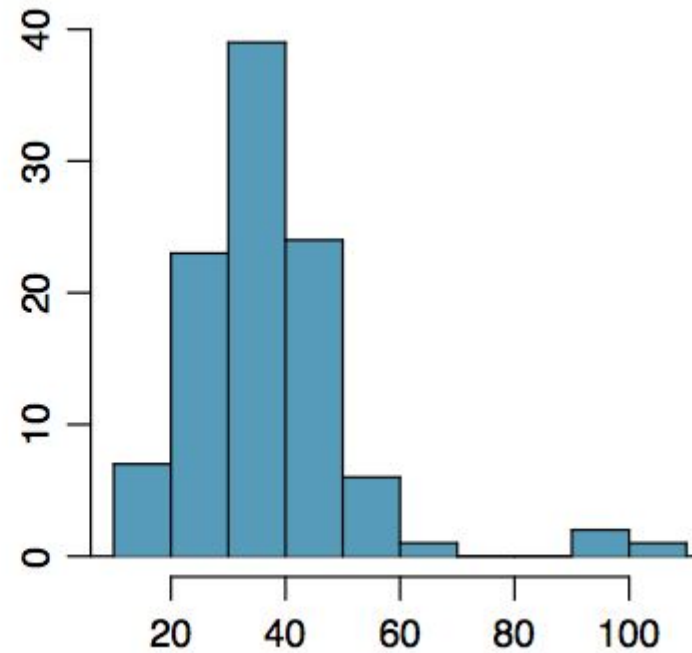
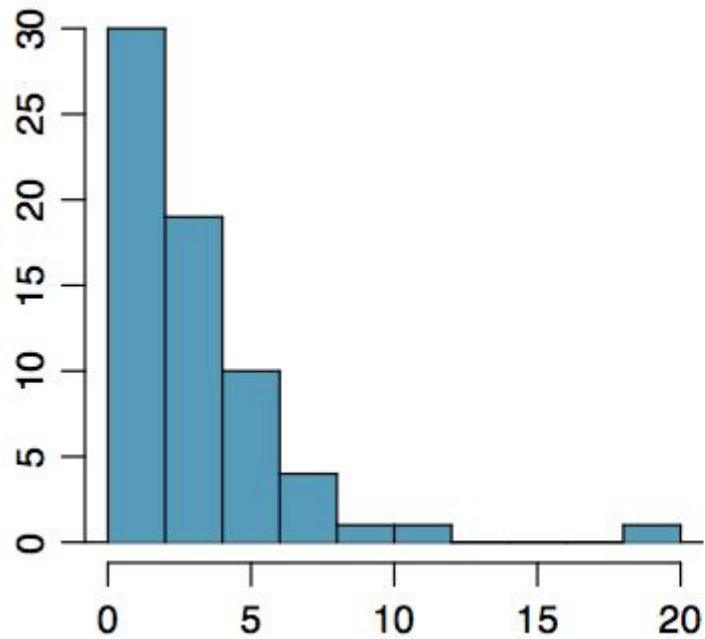
Is the histogram right skewed, left skewed, or symmetric?



Histograms are said to be skewed to the side of the long tail.

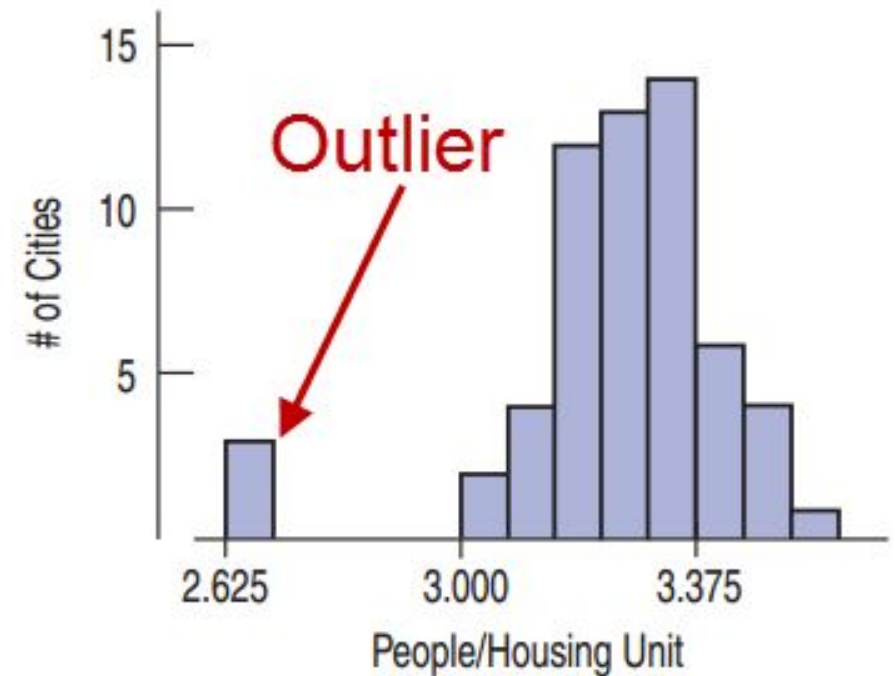
Shape of a Distribution: Unusual Observations

Are there any unusual observations or potential outliers?



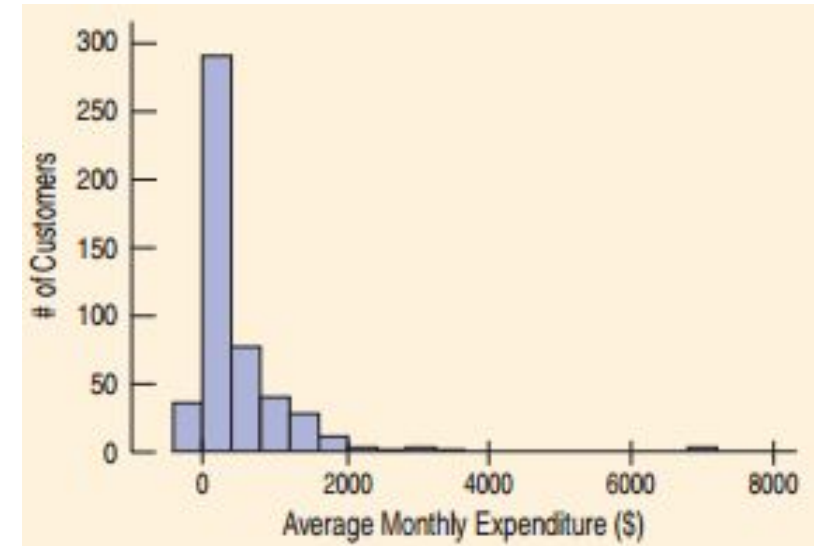
Outliers/ Extreme Values

- An **Potential Outlier** is a data value that is far above or far below the rest of the data values.
- An potential outlier is sometimes just an error in the data collection.
- An potential outlier can also be the most important data value.
 - Income of a CEO
 - Pinocchio's nose length after lying
 - Elevation at Death Valley



Example

- The histogram shows the amount of money spent by a credit card company's customers. Describe and interpret the distribution.
 - The distribution is **unimodal**. Customers most commonly spent a small amount of money.
 - The distribution is **skewed right**. Many customers spent only a small amount and a few were spread out at the high end.
 - There is an **extreme value** at around **\$7000**. One customer spent much more than the rest of the customers.



Commonly observed shapes of distributions

Modality

unimodal



bimodal



multimodal



uniform



Skewness

right skew



left skew



symmetric



A photograph of Bill Maher on a talk show set. He is wearing a dark suit, a white shirt, and a green tie. He is looking down and slightly to his left. The background is dark with some blurred lights. The text "Real Time with Bill Maher" is overlaid in white. Below it, "July, 2014 Episode 324" and "Bill Maher & Reihan Salam" are also overlaid in white. In the bottom right corner, the word "DAYS" is visible in a stylized font.

Real Time with Bill Maher
July, 2014 Episode 324
Bill Maher & Reihan Salam

DAYS

Checking the Numbers

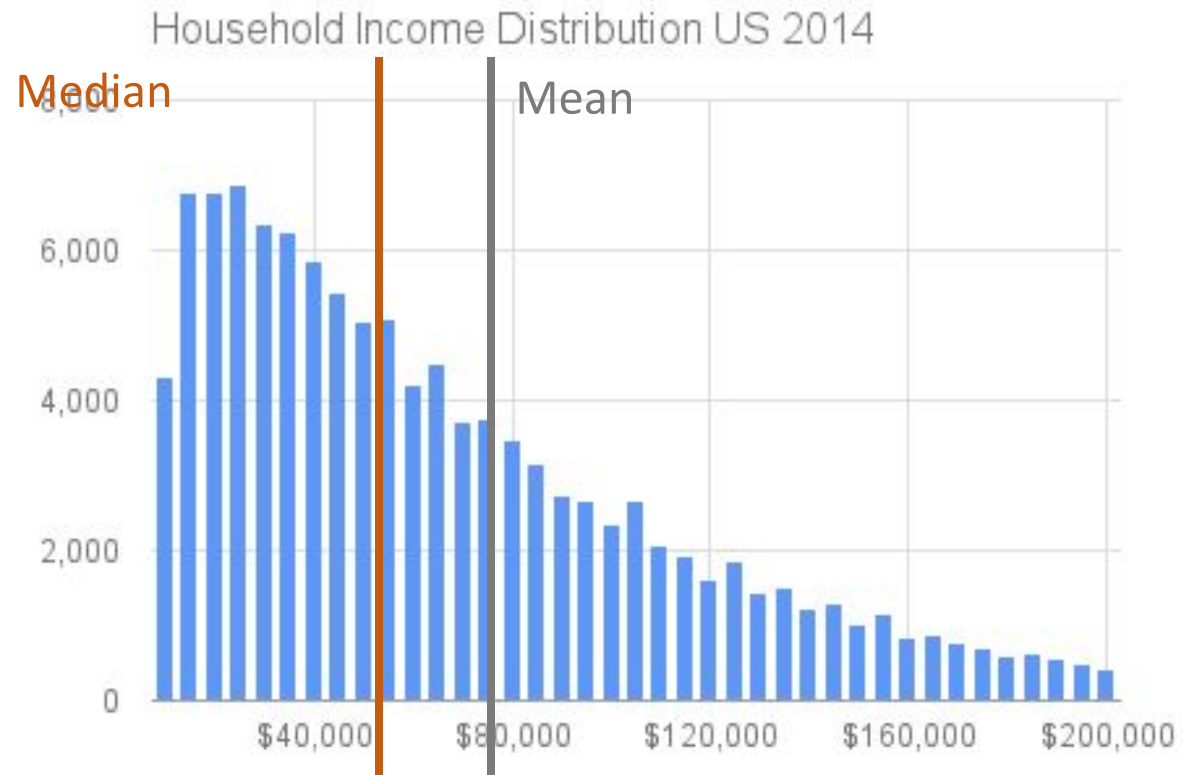
Bill Maher

- Median Household Income
 - \$51,000
 - Actual: \$53,000

Reihan Salam

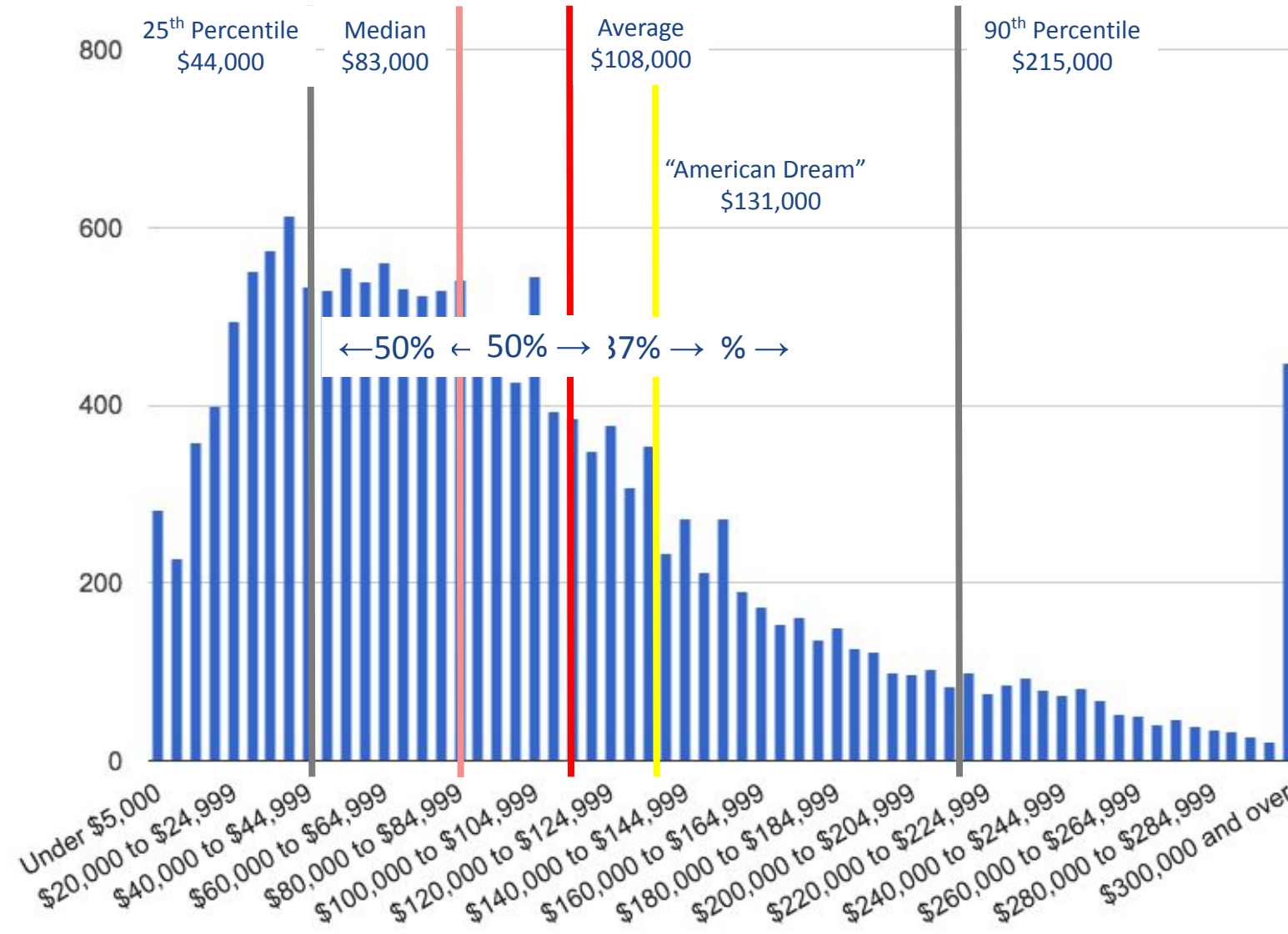
- Median Income: Family of 4
 - \$80,000
 - Actual: \$83,000
- Mean Income: Family of 4
 - \$100,000
 - Actual: \$108,000

Looking at the Distribution



Households of 4 Total Income

2014



Another Example

- Two people have a total height of 11ft. What's the most likely guess for the heights of the two individuals?
 - 5 ft 6 in.
- Two people have a total income of \$36 million dollars. What's the most likely guess for the incomes of the two individuals?
 - \$0 and \$36,000,000

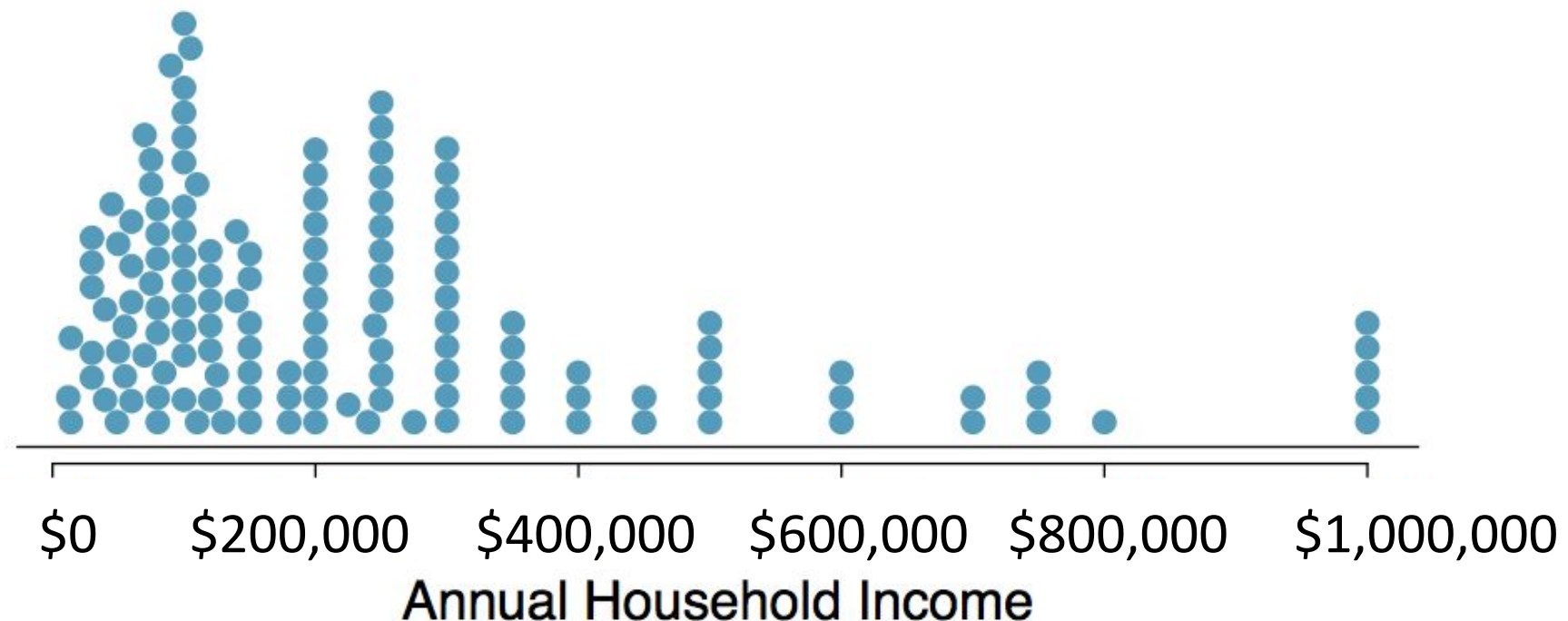
Outliers/Extreme Values?

Why is it important to look for extreme values?

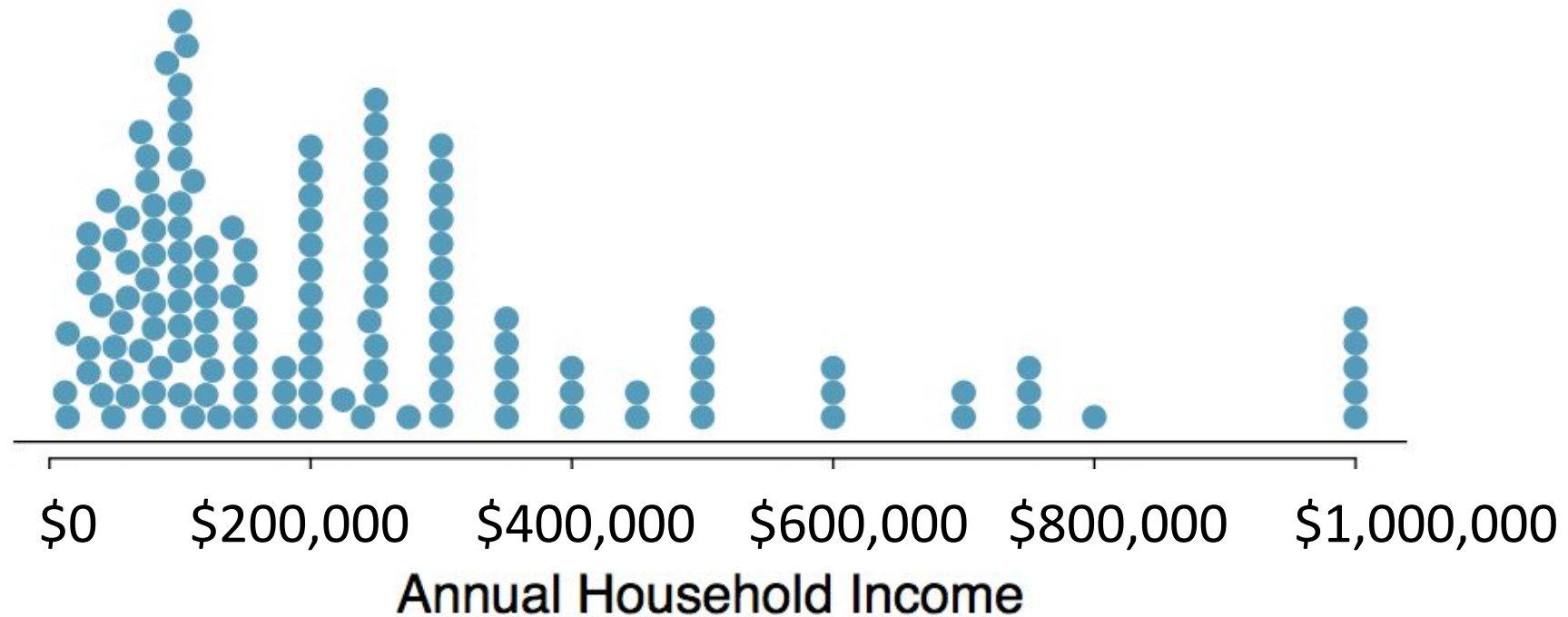
- Identify extreme skew in the distribution.
- Identify data collection and entry errors.
- Provide insight into interesting features of the data.

Extreme Observations & Summary Statistics

How would sample statistics such as mean, median, SD, and IQR of household income be affected if the largest value was replaced with \$10 million? What if the smallest value was replaced with \$10 million?



Robust Statistics



scenario	robust		not robust	
	median	IQR	\bar{x}	s
original data	190K	200K	245K	226K
move largest to \$10 million	190K	200K	309K	853K
move smallest to \$10 million	200K	200K	316K	854K

Robust Statistics

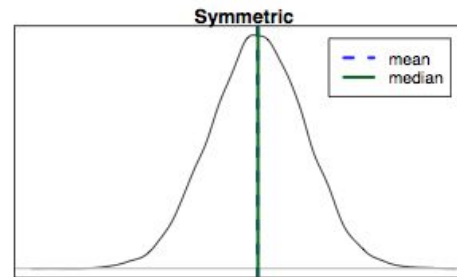
Median and IQR are more robust to skewness and outliers than mean and SD. Therefore,

- for skewed distributions it is often more helpful to use median and IQR to describe the center and spread
- for symmetric distributions it is often more helpful to use the mean and SD to describe the center and spread

Mean vs. Median

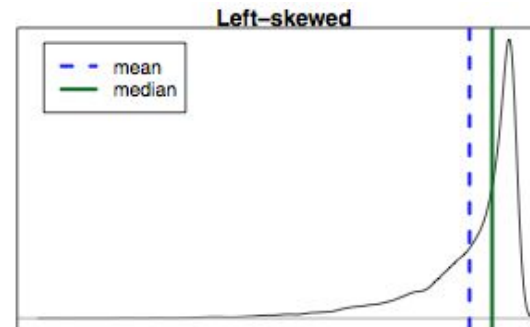
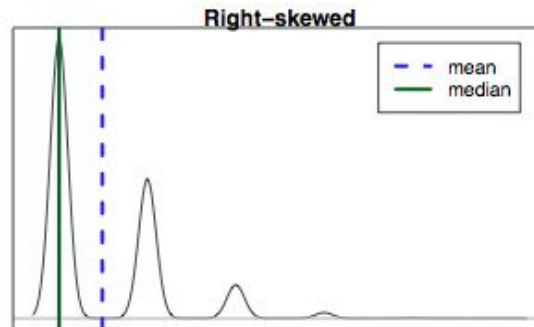
If the distribution is symmetric, center is often defined as the mean:

mean \sim median



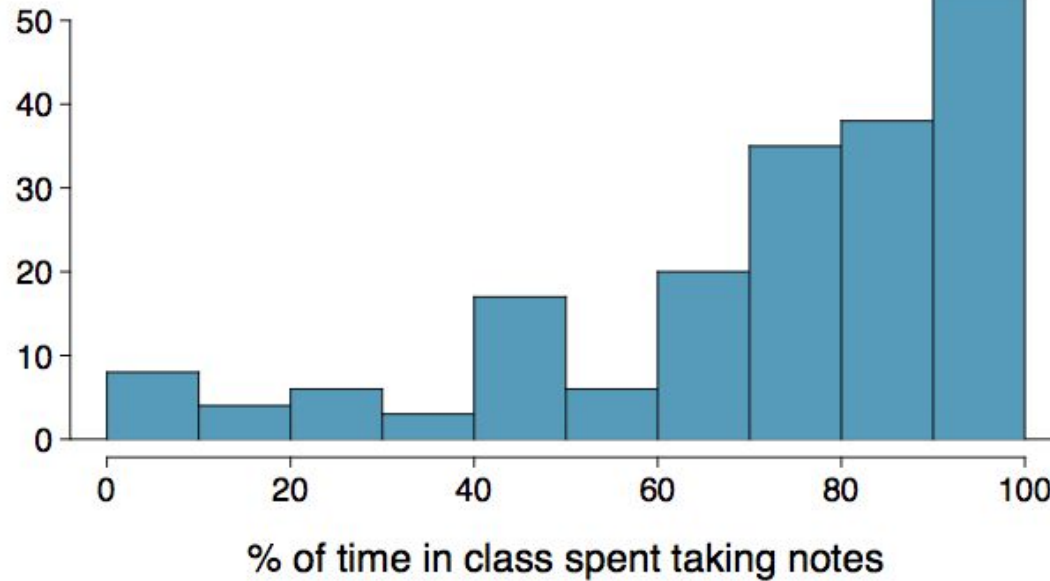
If the distribution is skewed or has extreme outliers, center is often defined as the median

- Right-skewed: mean $>$ median
- Left-skewed: mean $<$ median



Question

Which is most likely true for the distribution of percentage of time actually spent taking notes in class versus on Facebook, Twitter, etc.?



(a) mean > median

(b) mean ~ median

(c) mean < median

(d) impossible to tell

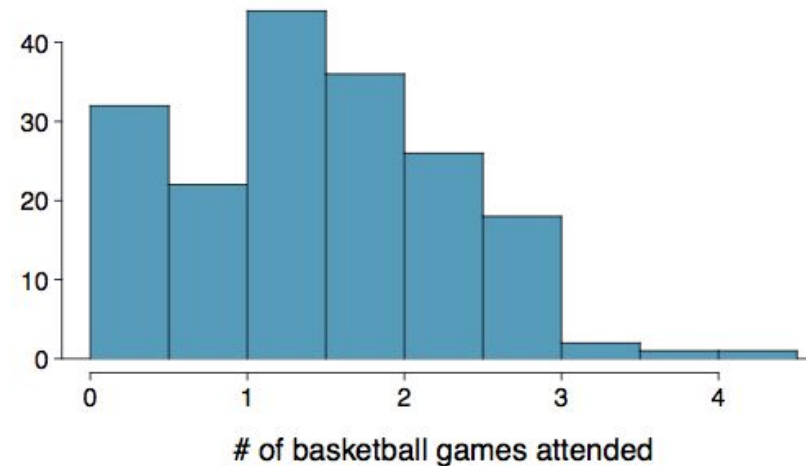
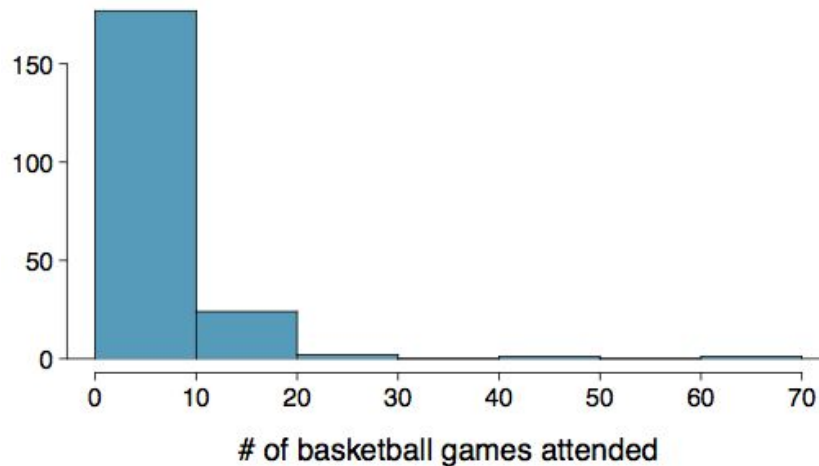
Data Transformations

- Transformation is a rescaling of the data using a function.
- When data are very strongly skewed, we sometimes transform them so they are easier to model.

Extremely Skewed Data

A common transformation is the log transformation.

The histograms on the left shows the distribution of number of basketball games attended by students. The histogram on the right shows the distribution of log of number of games attended.



Pros and Cons of Transformations

Skewed data are easier to model with when they are transformed because outliers tend to become far less prominent after an appropriate transformation.

However, results of an analysis might be difficult to interpret because the log of a measured variable is usually meaningless.

Summary: Goals of Transformations

- To see the structure of the data differently.
- To reduce the skew to help modeling.
- To straighten a nonlinear relationship in a scatterplot.

Summarizing or Exploring a Numerical Dataset

- Focus on 5 aspects (at least) for Numerical Data
 - Center
 - Spread
 - Skew
 - Clusters/Modality
 - Extreme Values
- Always look at multiple data representations

Try It

- Find a numeric variable in one of the Fivethirtyeight or Openintro Data sets and explore and summarize it.
 - Make sure to try to describe each of the 5 aspects on the previous slide.
 - Generate boxplots and histograms
- Focus on 5 aspects (at least) for Numerical Data
 - Center
 - Spread
 - Skew
 - Clusters/Modality
 - Extreme Values
- Always look at multiple data representations

Anticipating Shape, Symmetry & Skew

- What modality and symmetry would you expect from the following data sets.
 - Heights of women in the U.S.
 - Income of individuals in California
 - Outcomes from rolling a 20 sided die 10,000 times
 - Magnitudes of earthquakes that occurred between 2010 and 2018
 - Weights of elephants worldwide

Anticipating Shape, Symmetry & Skew

- If data values have a hard limit on one side but not the other, they are more likely to be skewed (especially if the center of the distribution is near the limit)
 - For example:
 - You can't earn less than \$0
 - You can't score more than 100% on a test
- If the center is far from any limits, the distribution is likely to be symmetric.
 - For example:
 - You can't be less than 0 inches tall, but average heights are not at all close to 0. Heights are usually distributed symmetrically

Exploring Nominal/Categorical Variables

General Social Survey

Year	Marital Status	Age	Race	Income	Party ID	Religion	Denomination
2004	Married	61	White	Not applicable	Strong democrat	Catholic	Not applicable
2008	Married	84	White	Not applicable	Strong democrat	Protestant	United methodist
2010	Married	61	White	Refused	Not str republican	Catholic	Not applicable
2008	Never married	26	Other	\$25000 or more	Strong democrat	Catholic	Not applicable
2014	Never married	26	White	\$25000 or more	Ind,near rep	Other	Not applicable
2006	Married	51	White	\$25000 or more	Independent	None	Not applicable
2014	Married	65	White	\$25000 or more	Not str democrat	Buddhism	Not applicable
2002	Widowed	84	White	Not applicable	Not str republican	Protestant	No denomination
2002	Divorced	37	White	Lt \$1000	Independent	Catholic	Not applicable

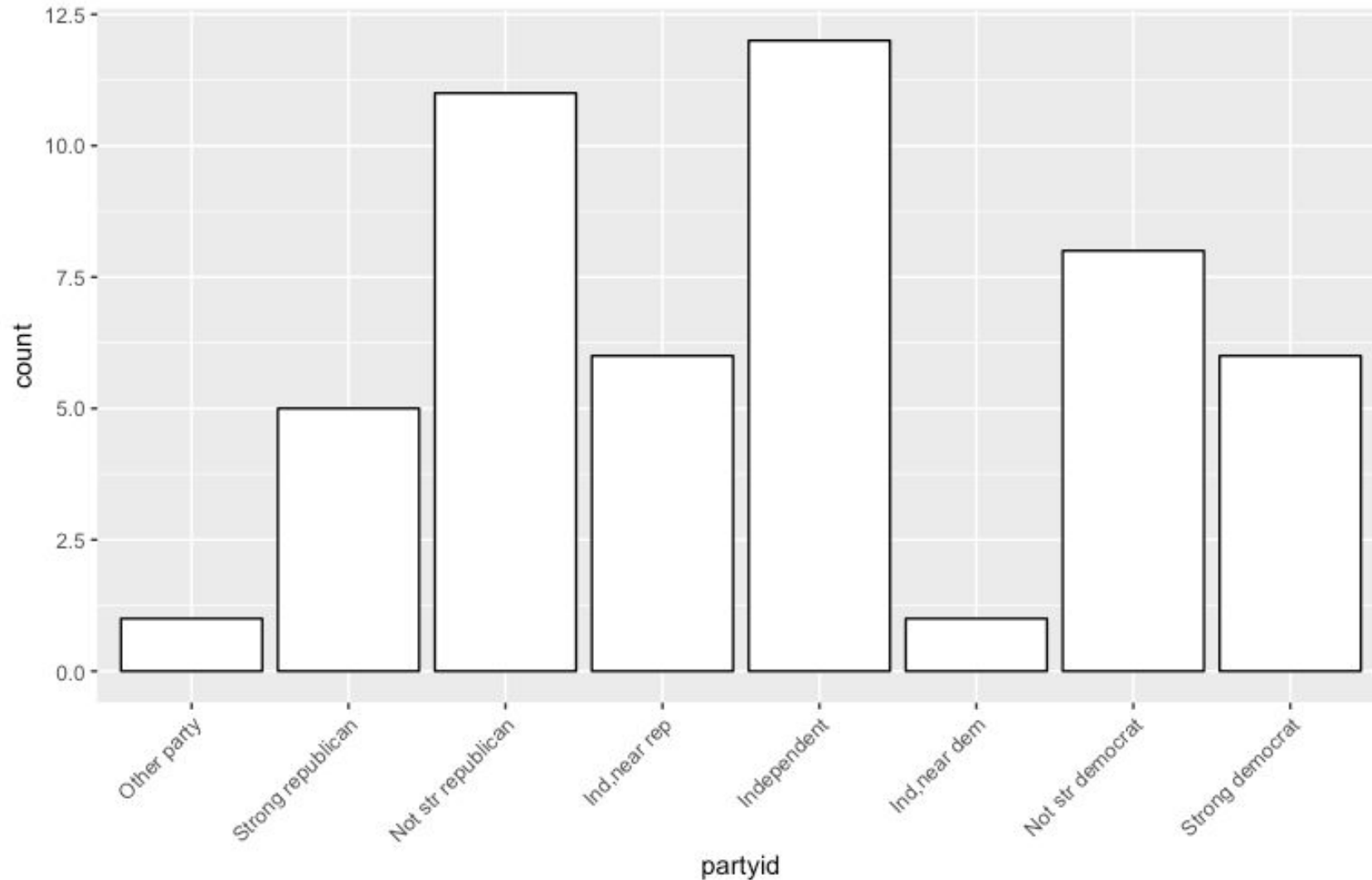
Categorical Data

Means, Medians, sd, etc. are not applicable

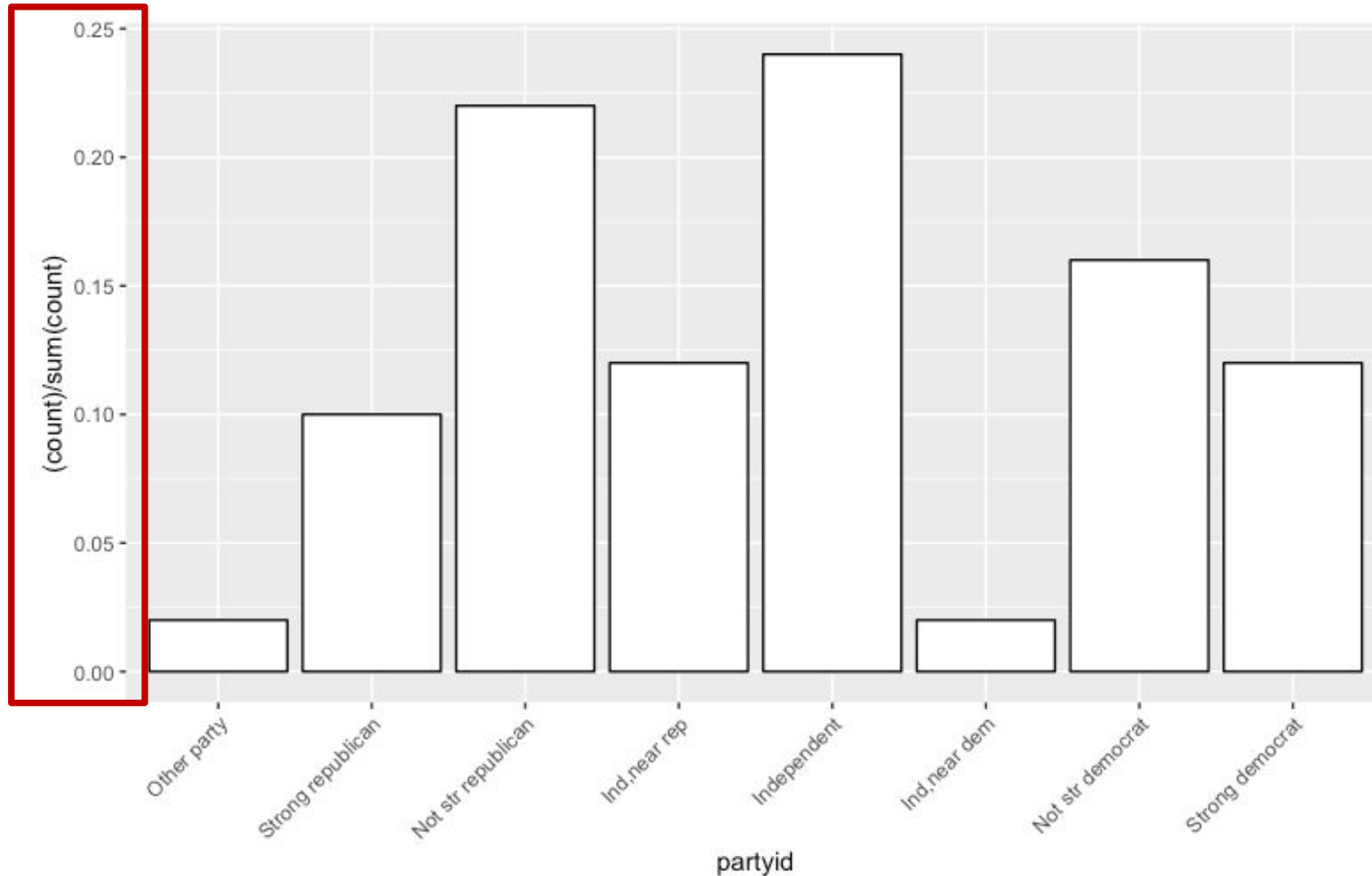
Instead:

- **Counts**: How many are in each group?
- **Proportions**: What % of the total are in each group?

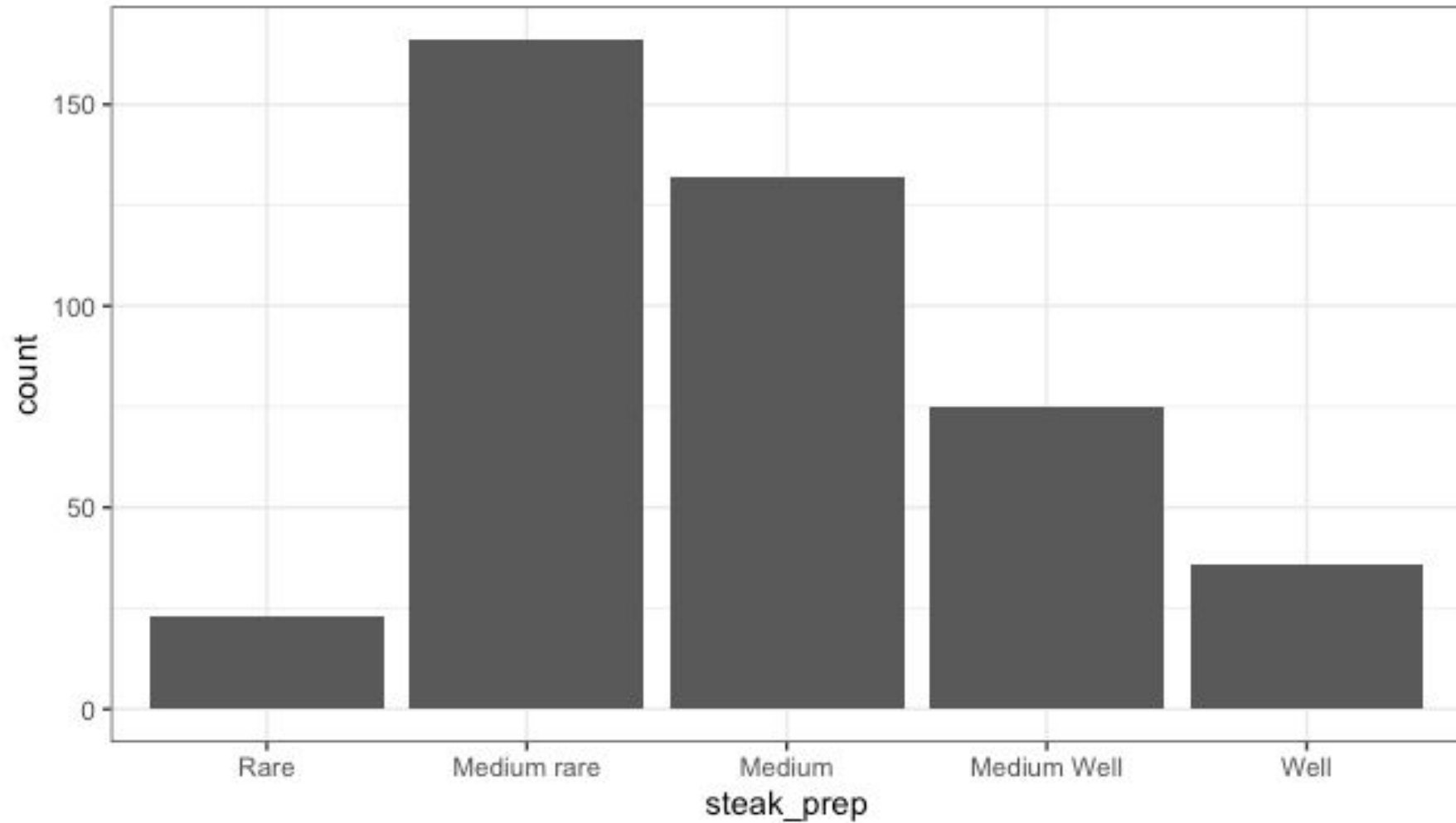
Visualizing Counts: Bar Charts



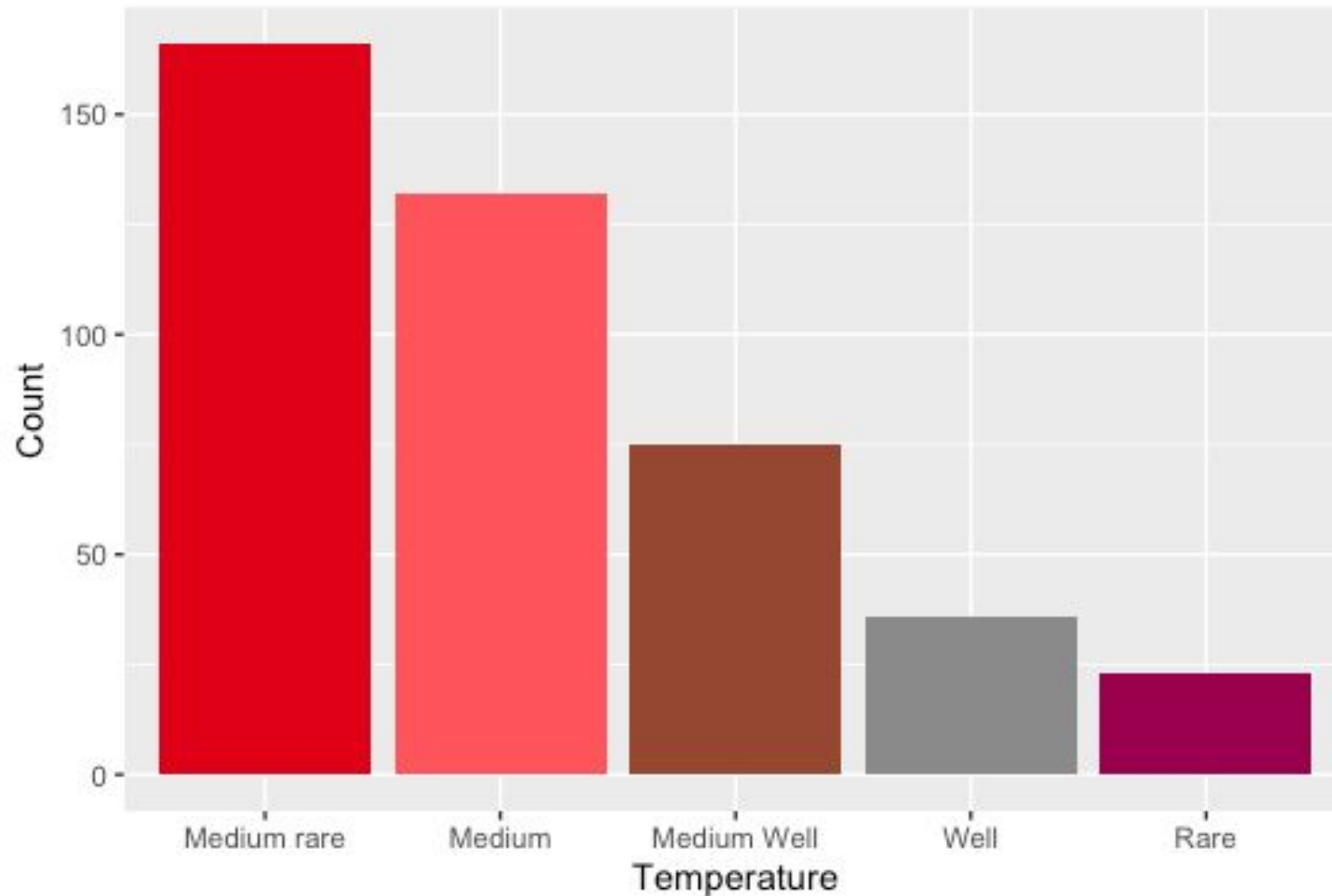
Visualizing Proportions: Bar Charts



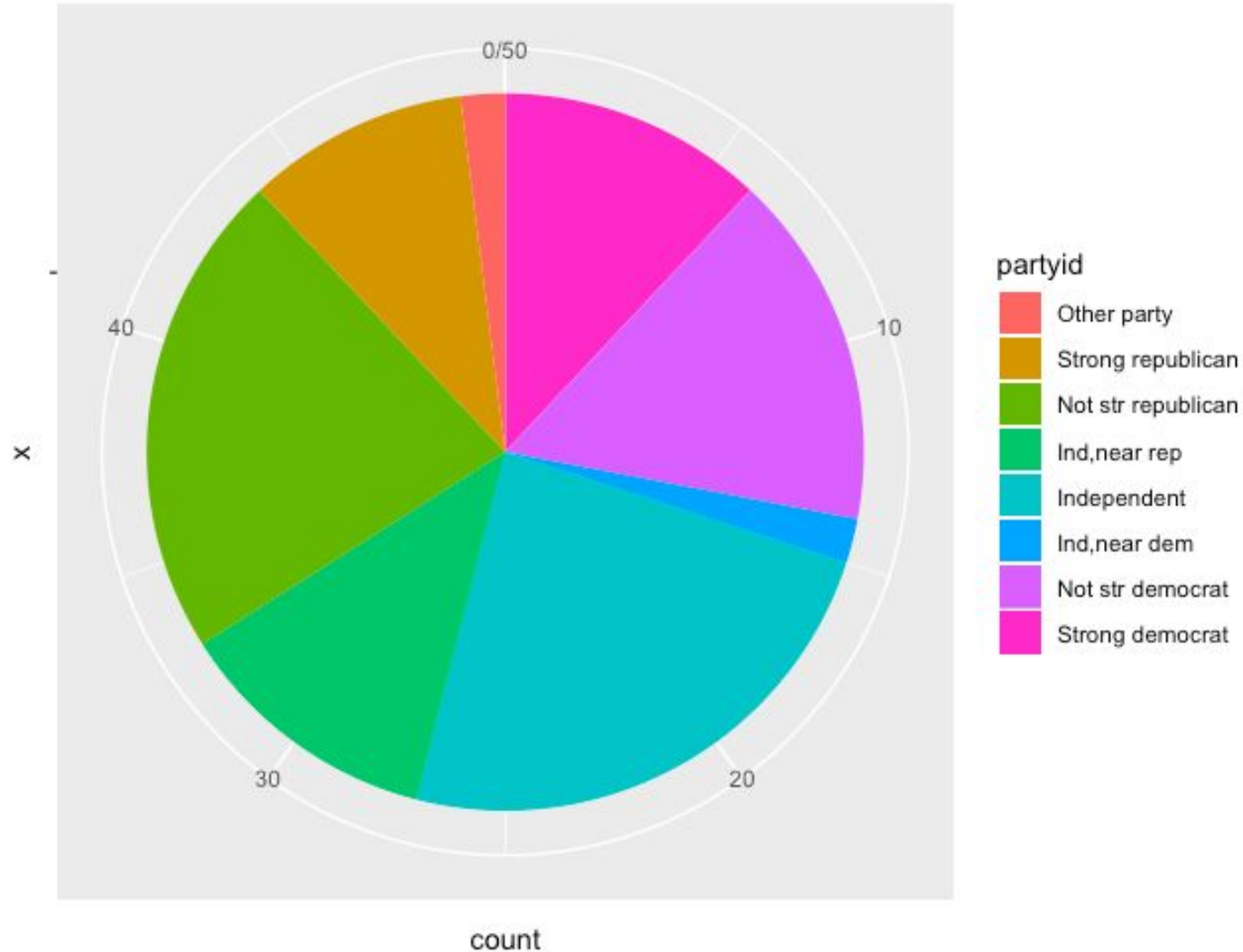
Bar Charts: Another Example



Bar Charts: Another Example



Visualizing Proportions: Pie Charts



Relationships Between Variables: Contingency Tables

A table that summarizes data for two categorical variables is called a **contingency table**.

	Rare	Medium Rare	Medium	Medium Well	Well	Total
No Smoke	16	136	111	"Marginal Values"- row or column totals		356
Smoke	6	30	20			72
Total	22	166	131	74	35	428

Overall Total

Relationships Between Variables: Contingency Tables

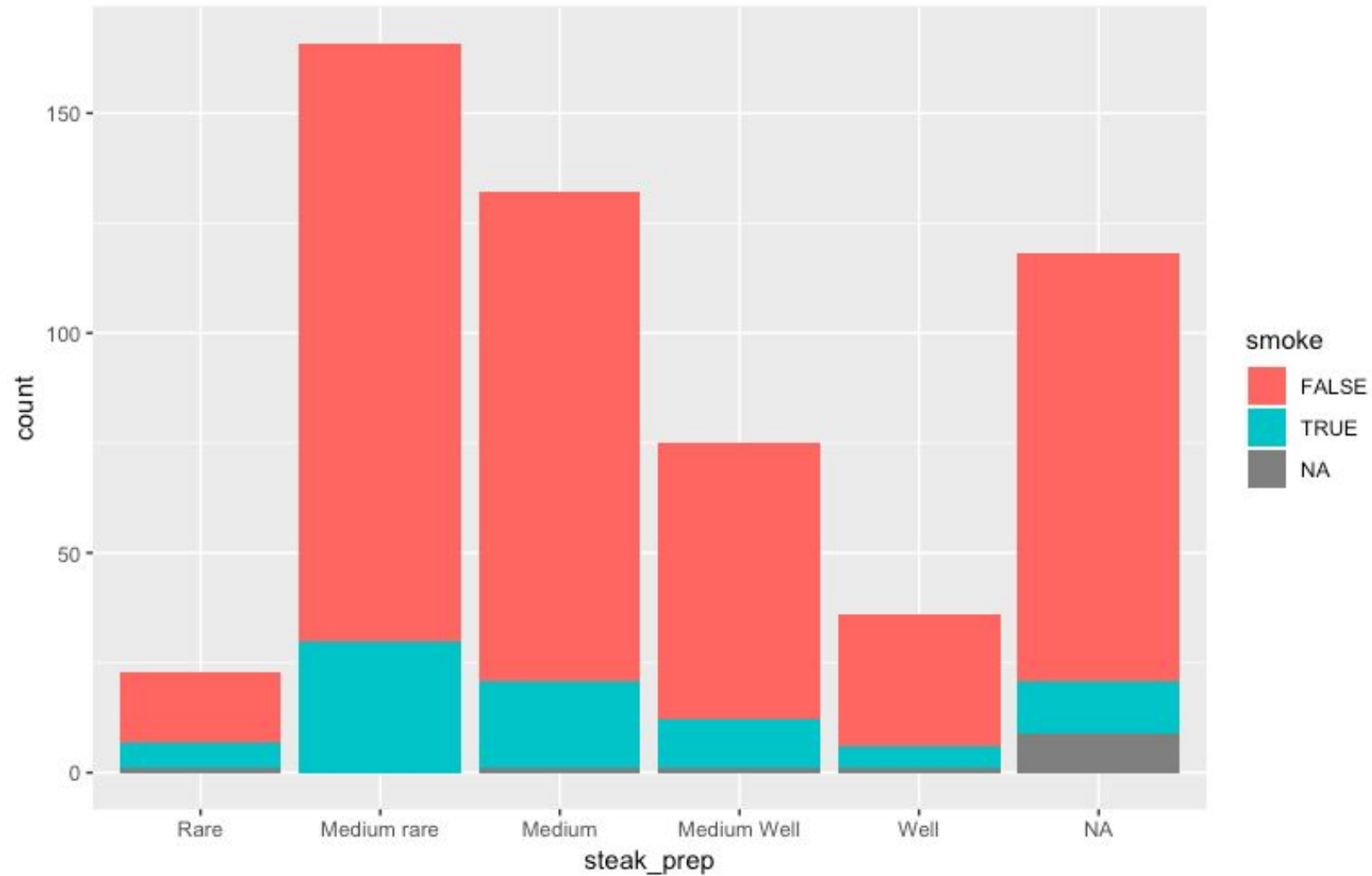
Does there appear to be a relationship between smoking and preference in steak preparation?

	Rare	Medium Rare	Medium	Medium Well	Well	Total
No Smoke	16	136	111	63	30	356
Smoke	6	30	20	11	5	72
Total	22	166	131	74	35	428

Contingency Table Proportions

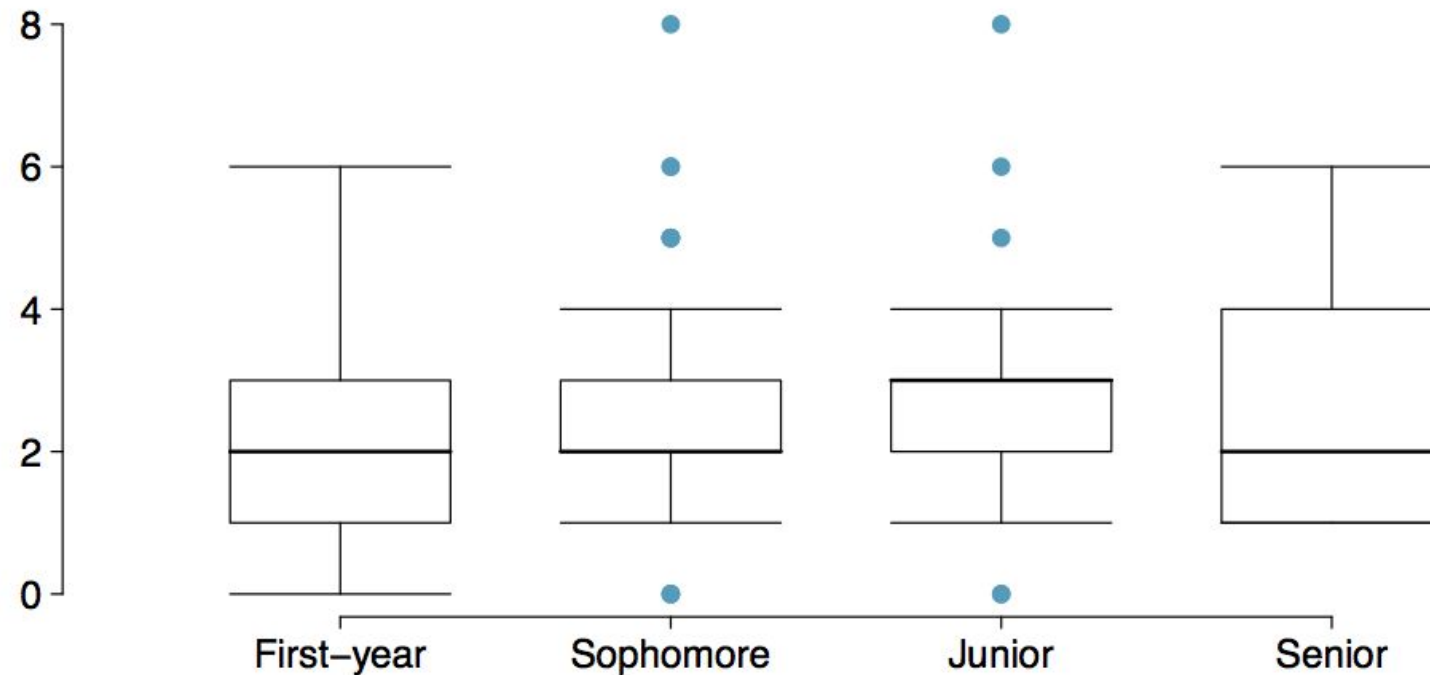
- Proportions can be calculated in three ways:
 - Based on overall totals (What % of the whole?)
 - Based on row totals (What % of smokers?)
 - Based on column totals (What % of Rare Steak Lovers?)
- Be very careful of the denominator!!

Segmented Bar Charts



Side by Side Box Plot

Does there appear to be a relationship between class year and number of clubs students are in?



box plot of number of clubs college students are involved with and their class year

Data Analysis Step 1: Explore & Describe A Single Variable

Quantitative

- Center, Spread, Shape, Symmetry, Extreme Values
- Histogram
- Boxplot
- other (e.g. dotplot, lineplot...)

Qualitative

- Mode
- Proportions for each level
- Frequency Table
- Bar Chart
- Pie chart (last resort)

Data Analysis Step 1: Explore & Describe

Two Variables: Both Categorical

- Contingency Table
- Double Bar Chart
- Segmented Bar Chart
- Mosaic Plot

Are they independent?

Independence

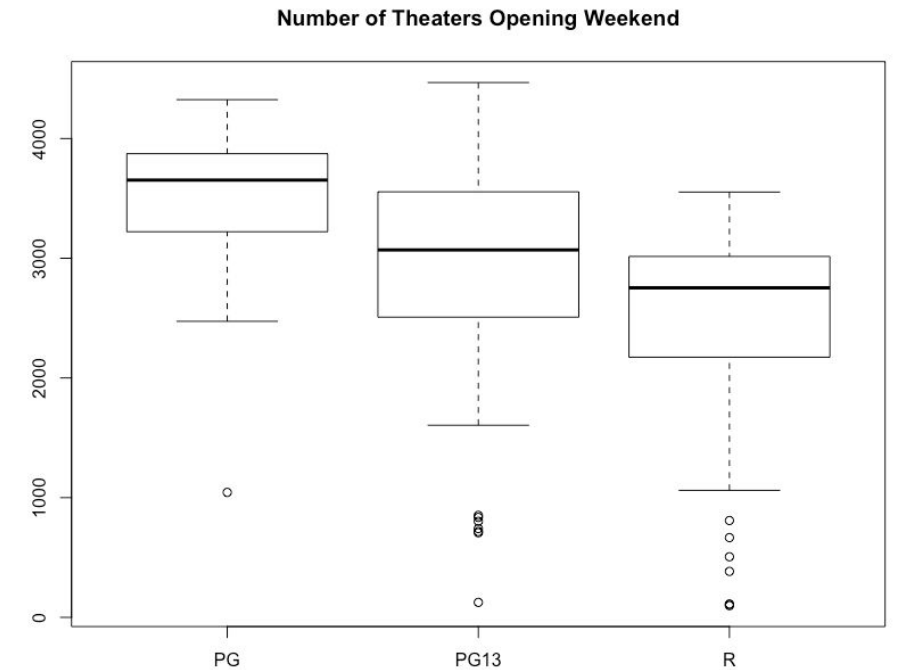
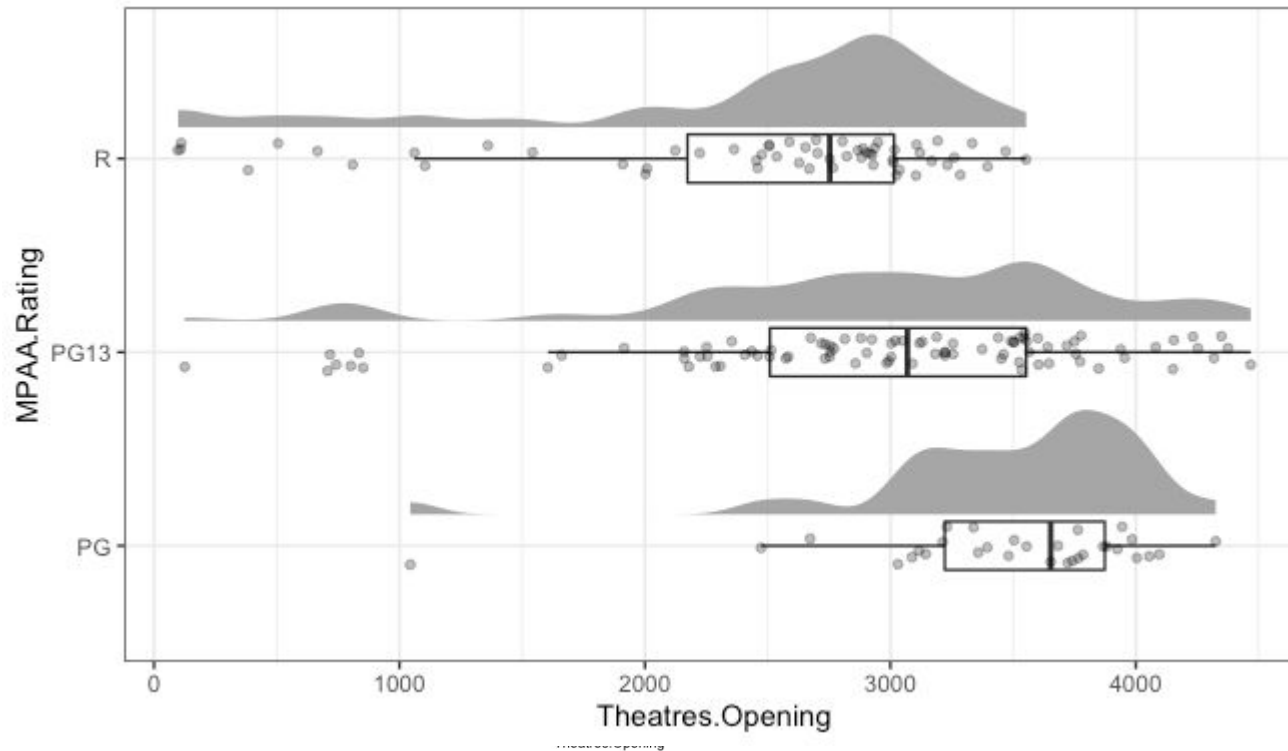
Two variables are

- **Independent** if they do not affect each other.
 - A change in one variable doesn't imply a change in another.
 - Patterns across levels of a variable are all similar
- **Dependent** if they do affect each other.
 - The variables change together in some way
 - Patterns for one level are different from patterns in another level.

Data Analysis Step 1: Explore & Describe

Two Variables: One Categorical, One Numeric

- Side by side boxplots
- Transparent Histograms
- Raincloud plots



Are they independent?

Data Analysis Step 1: Explore & Describe

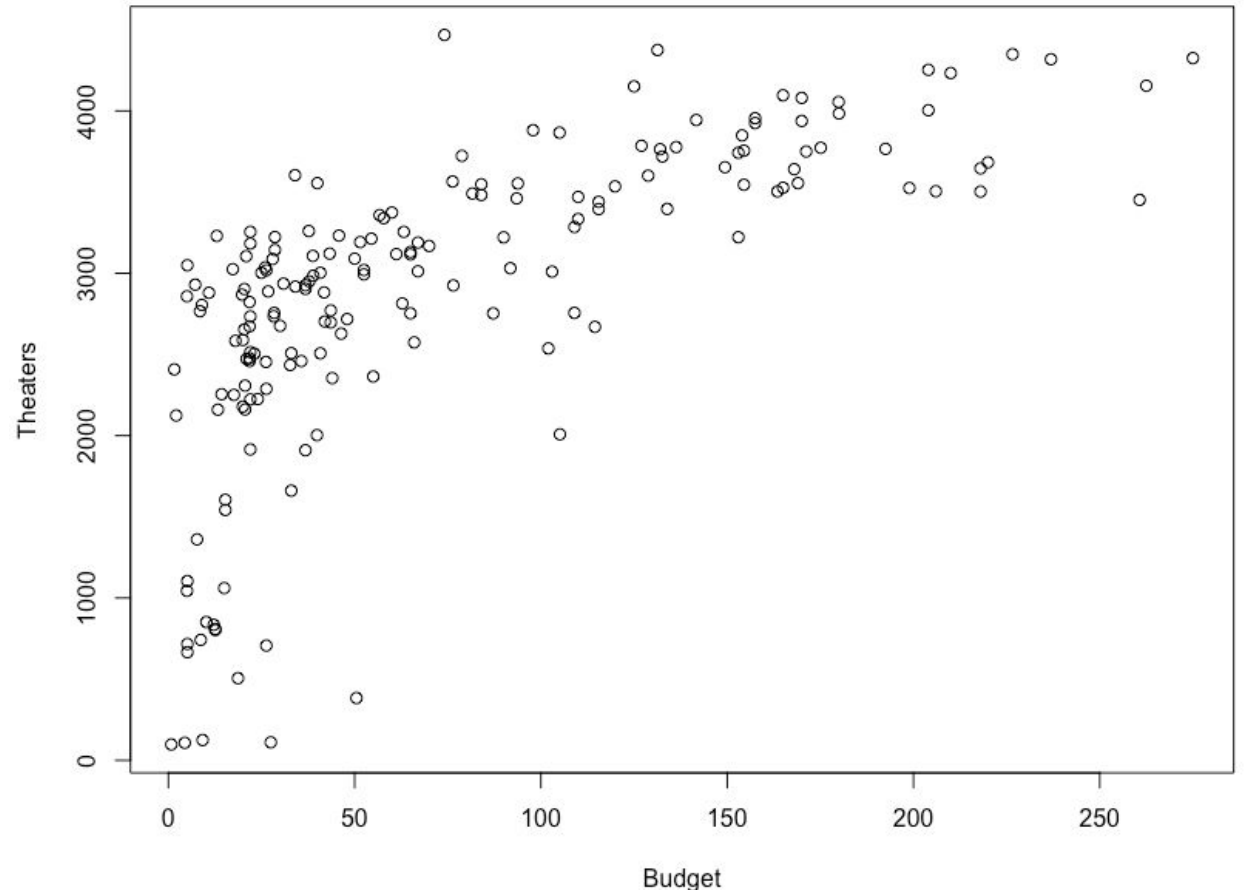
Two Variables: Both Numeric

- Scatterplot

Are they independent?

How do we describe the relationship?

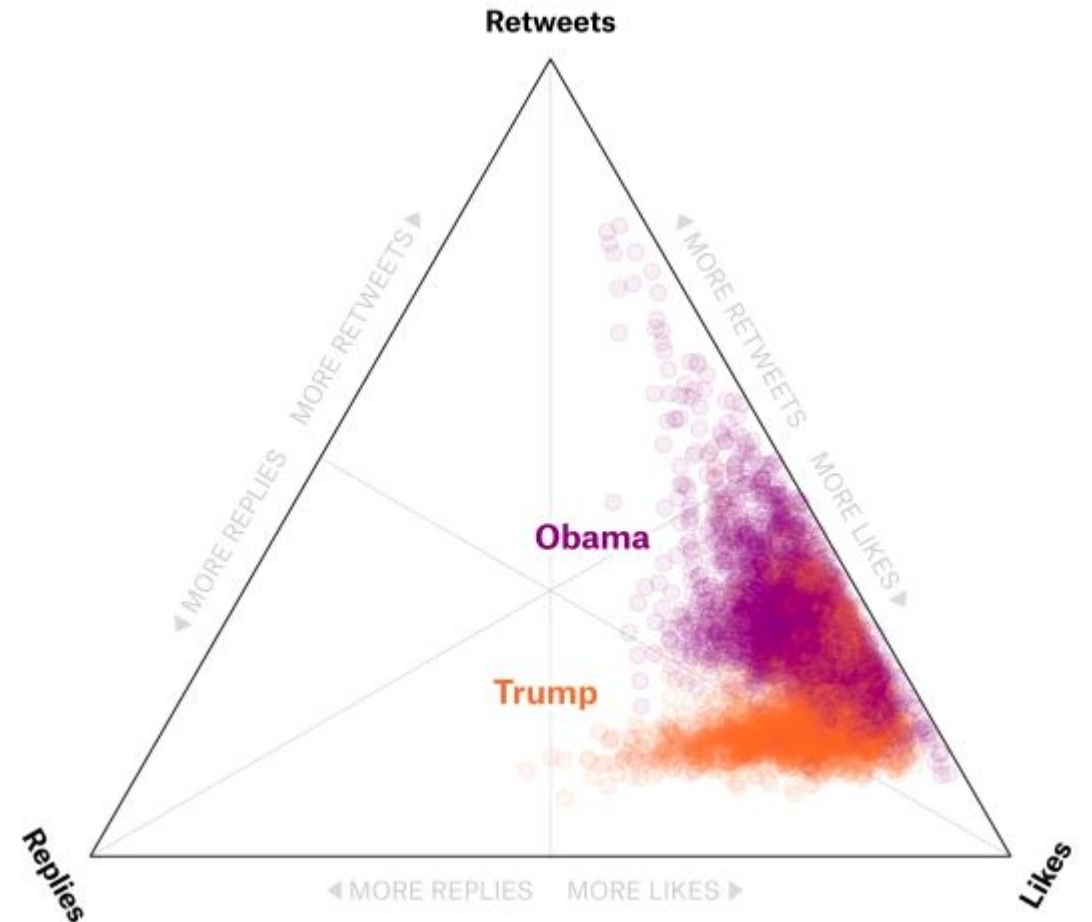
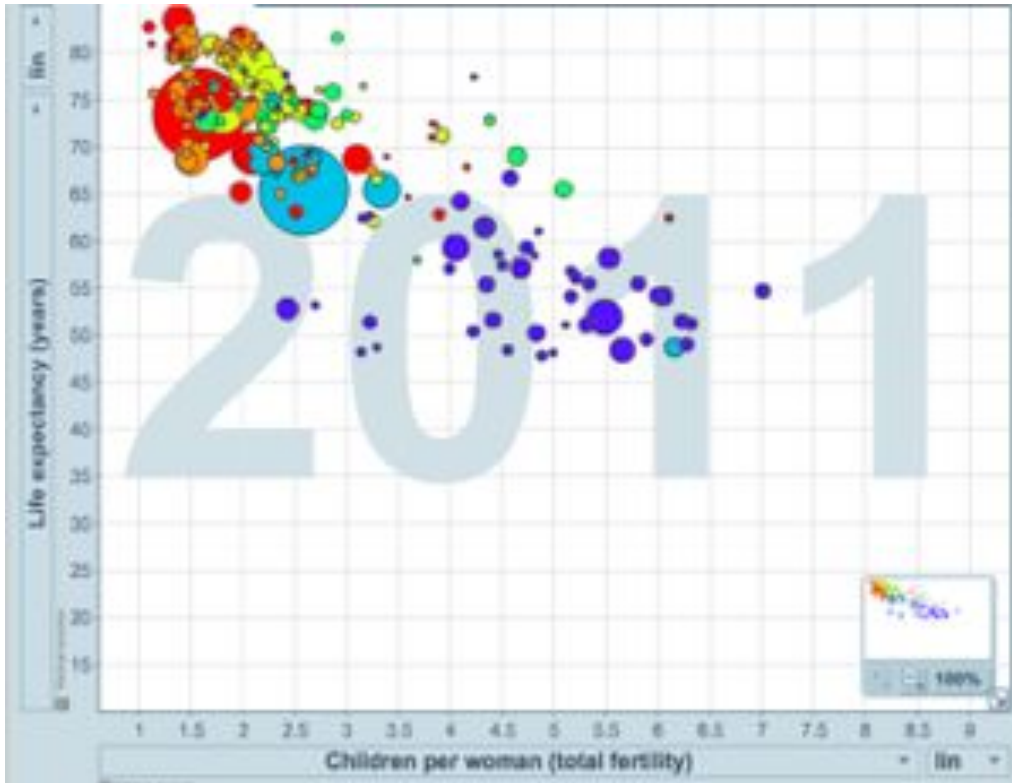
- Fairly strong
- Positive
- Non-linear



Data Analysis Step 1: Explore & Describe

More than Two Variables

Be creative!



When two variables are related

- Two variables are related when they change together.
 - Different values of variable A correspond with different values of variable B.
- When variables are related, we can do two things:
 - Describe the relationship
 - How strong is the relationship?
 - What pattern does the relationship show?
 - Make inferences
 - Is this relationship predictive of the larger population?
 - Does this relationship indicate the possibility of a cause-effect relationship.

When two variables are related

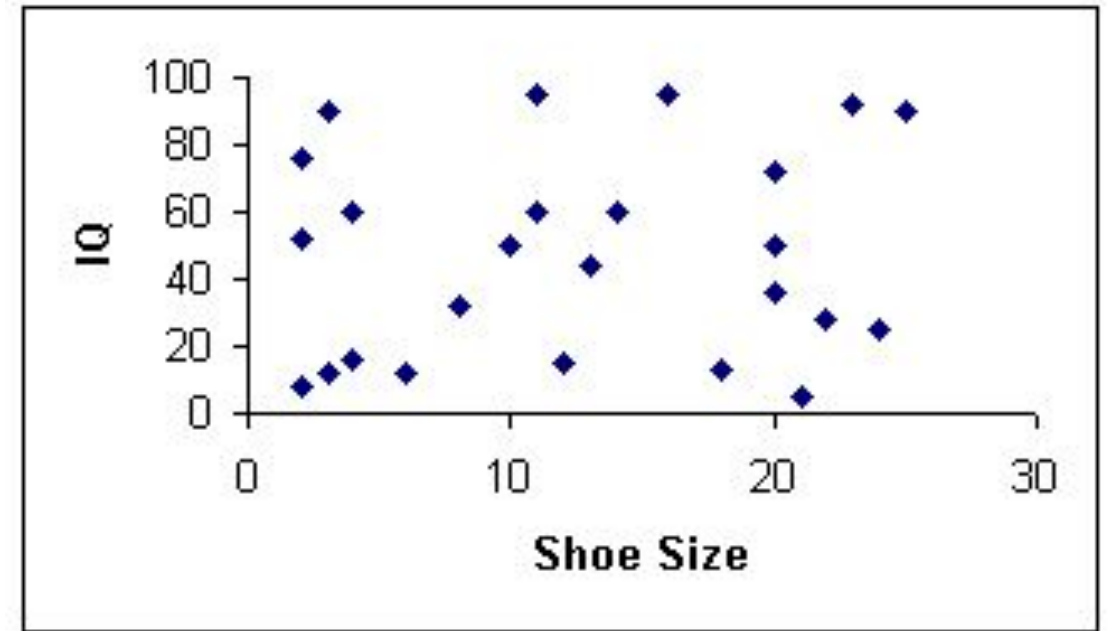
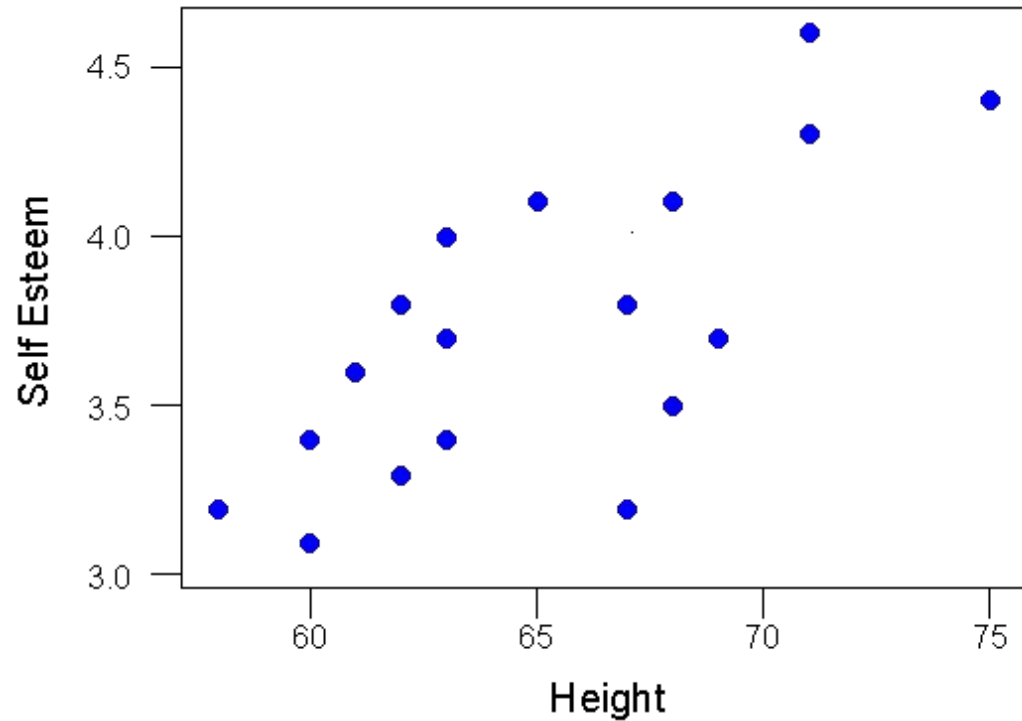
- Two variables are related when they change together.
 - Different values of variable A correspond with different values of variable B.
- When variables are related, we can do two things:
 - Describe the relationship
 - How strong is the relationship?
 - What pattern does the relationship show?
 - Make inferences
 - Is this relationship predictive of the larger population?
 - Does this relationship indicate the possibility of a cause-effect relationship.

LATER...

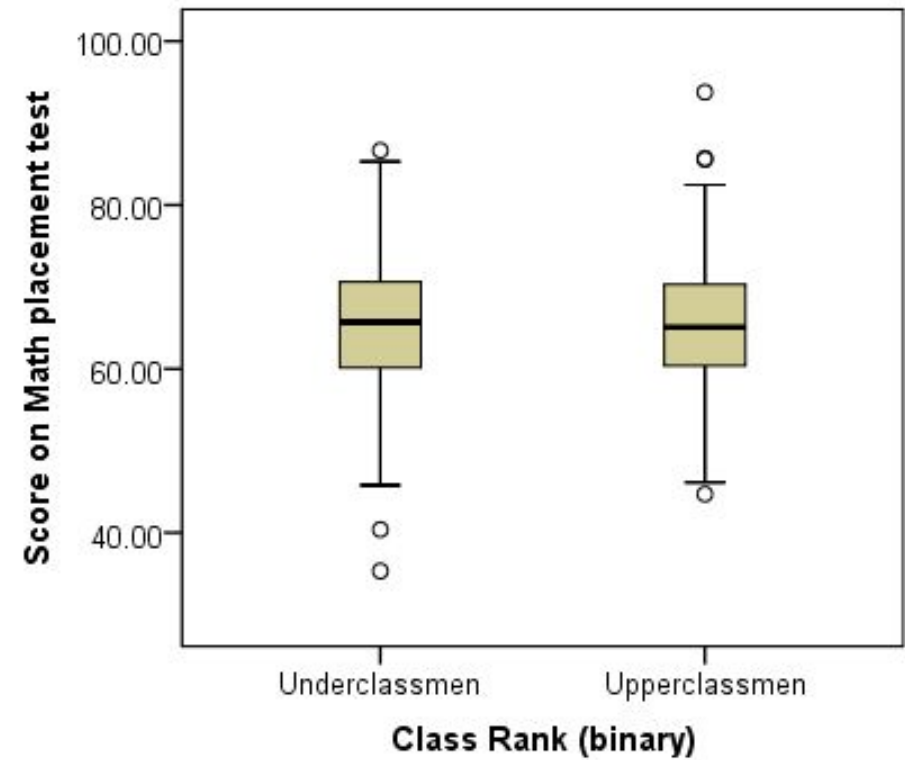
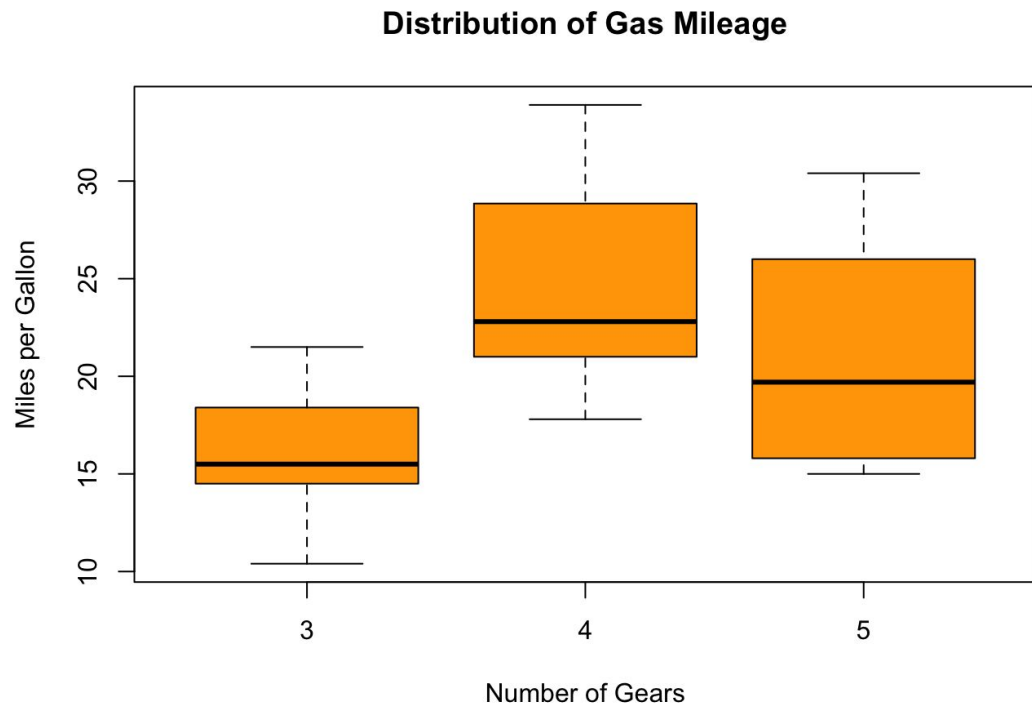
Relationships between Two Quantitative Variables

- Strength: Strong or Weak?
 - A strong relationship is indicated by how close to the trend the data fall
- Direction: Positive, Negative
 - Do the variables change in the same direction or opposite directions?
- Shape: Linear, Curved
- Outliers

Relationships in Variables



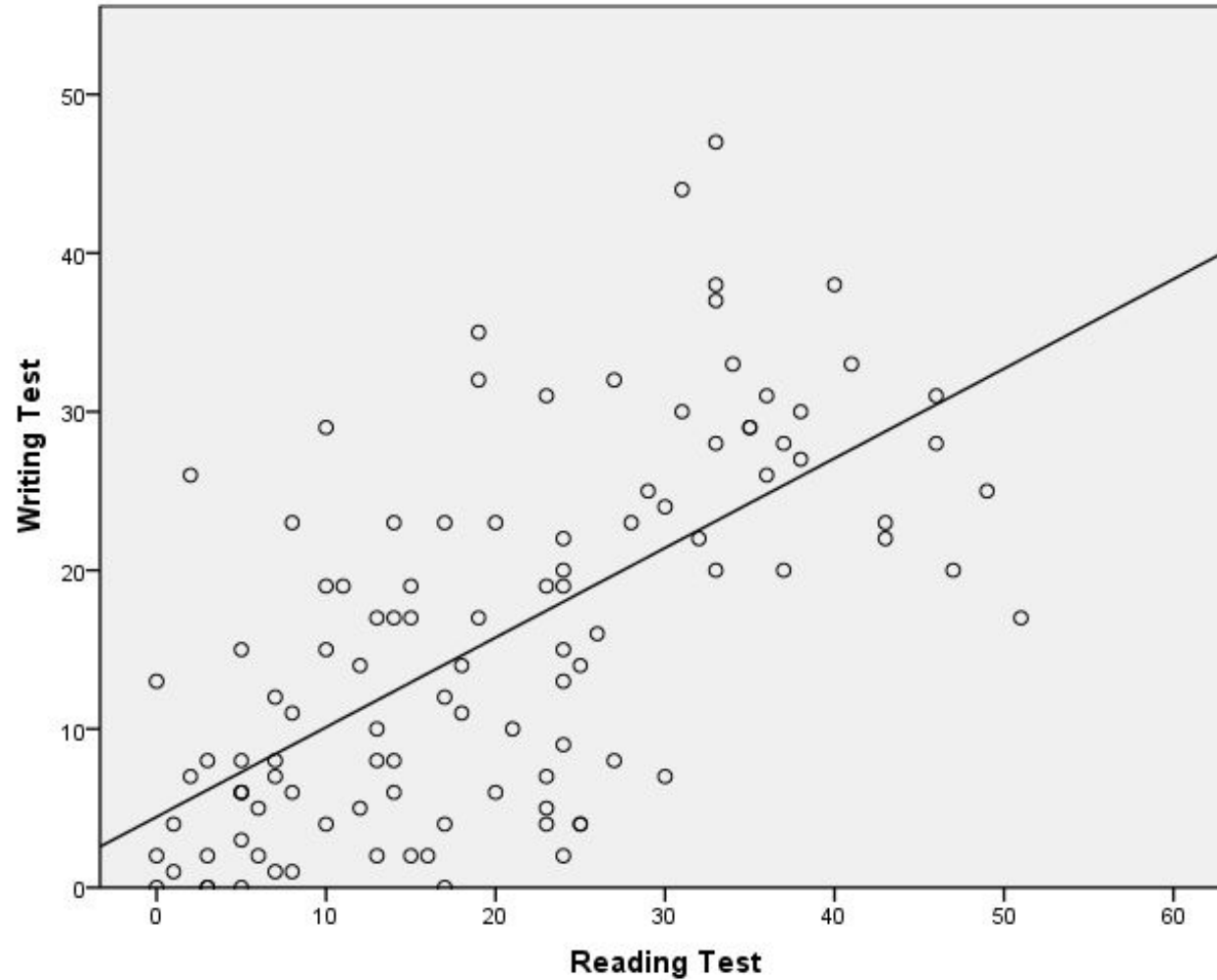
Relationships between Variables



Associated vs. independent

- When two variables show some connection with one another, they are called associated variables.
 - Associated variables can also be called dependent variables and vice-versa.
- If two variables are not associated, i.e. there is no evident connection between the two, then they are said to be independent.

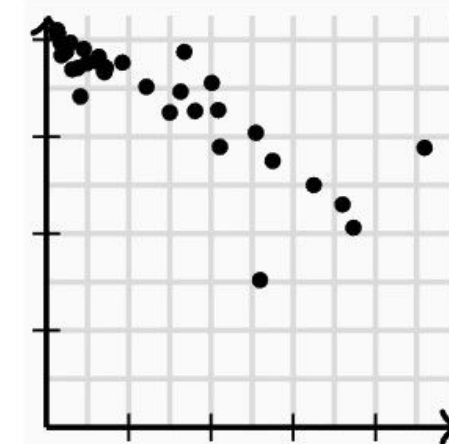
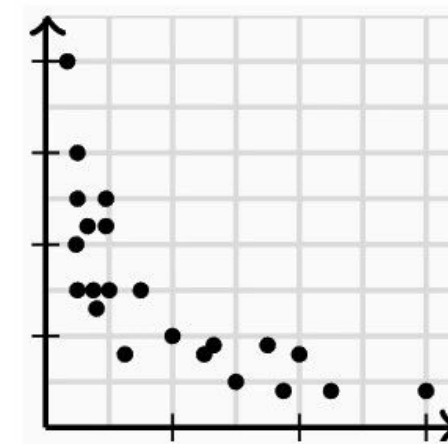
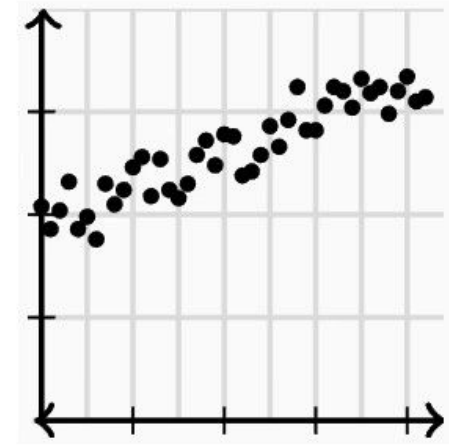
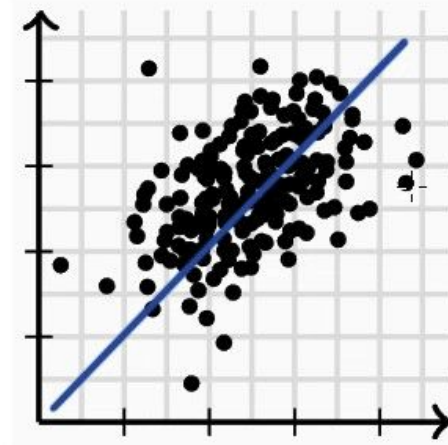
Example



Linear
Weak
Positive
Few extreme
values

More Examples

- Which ones are strong? Weak?
- Which ones are positive? negative?
- Which ones are linear? Curved?
- Which ones have outliers?

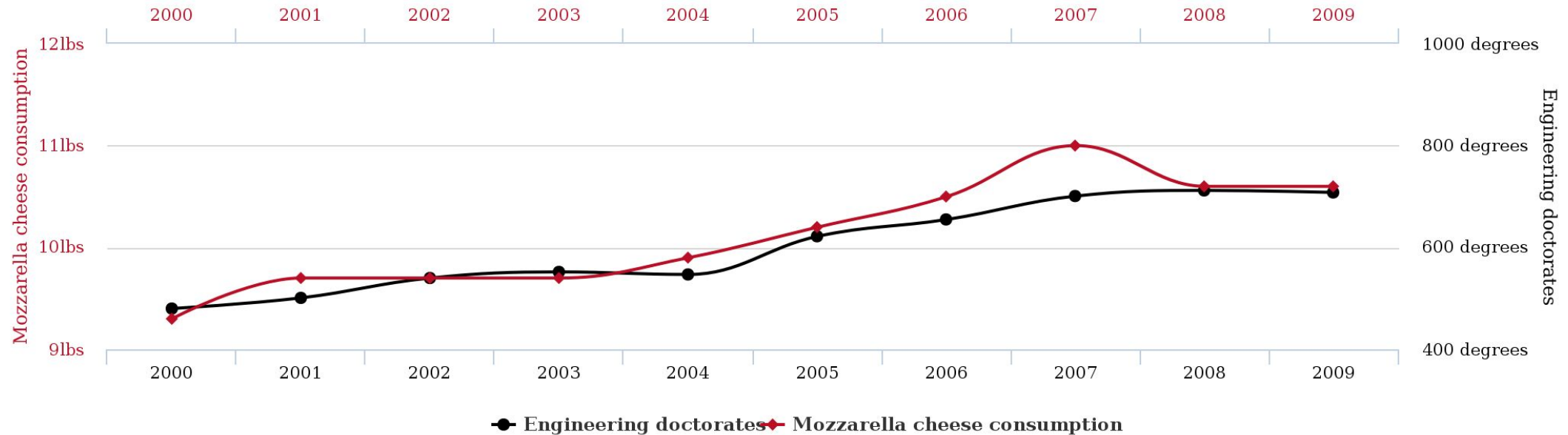


Identifying Relationships

- To identify relationships we create studies
 - Correlations are relatively straightforward to find
 - Causation is extremely difficult

Correlation \neq Causation

Per capita consumption of mozzarella cheese
correlates with
Civil engineering doctorates awarded



Descriptive Statistics in the Wild

The Washington Post

Democracy Dies in Darkness

North America has lost 29% of its birds in 50 years

A sweeping new study says a steep decline in bird abundance, including among common species, amounts to "an overlooked biodiversity crisis."

By Karin Brulliard

The New York Times

3 Billion North American Birds Have Vanished: 'It's Just Staggering'

The number of birds in the United States and Canada has declined by 3 billion, or 29 percent, over the past half-century, scientists find.

7m ago [466 comments](#)

Simulations and the Need for Probability

Can psychics sense your aura?

- Practitioners of Therapeutic Touch claim to be able sense people's auras.



Can psychics sense your aura?

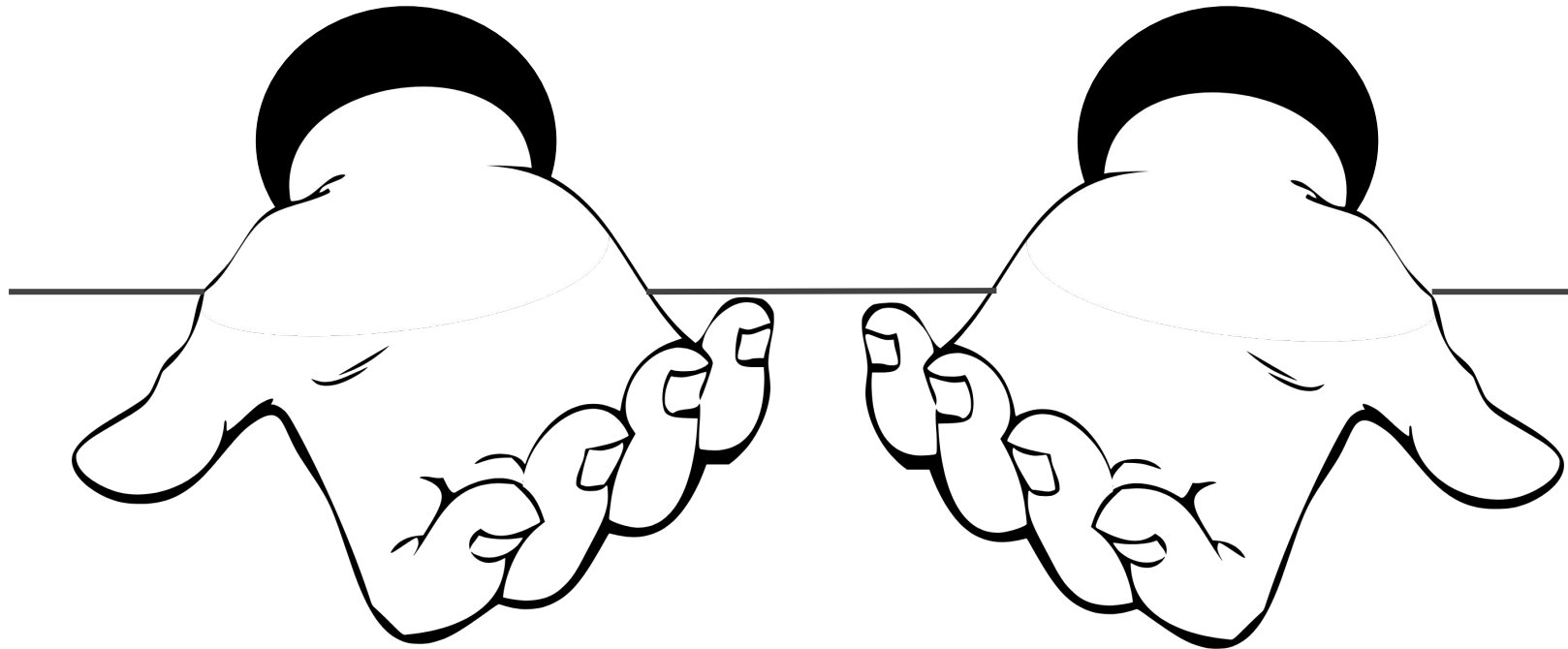
- Practitioners of Therapeutic Touch claim to be able sense people's auras.

Can we put this claim to the test?



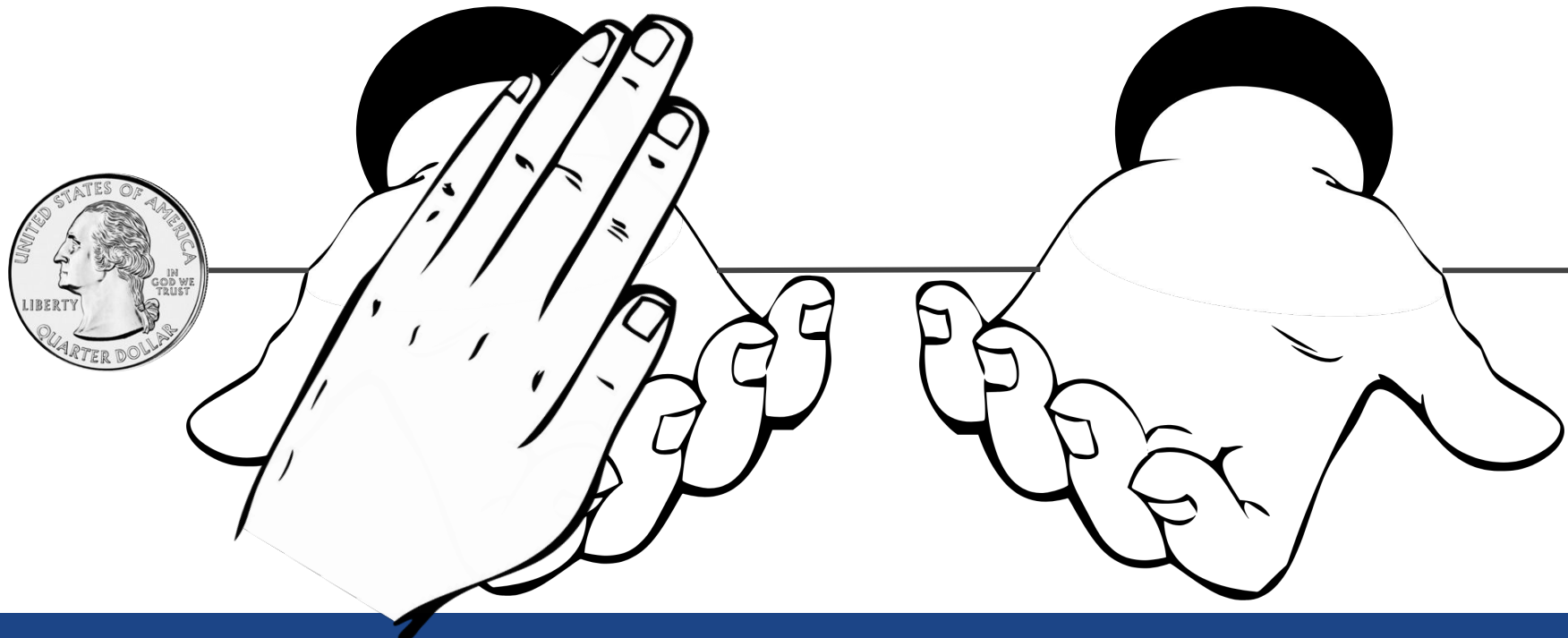
The Experiment

- 21 Therapeutic Touch Healers participate in a test
- The healers place their hands through an opaque screen



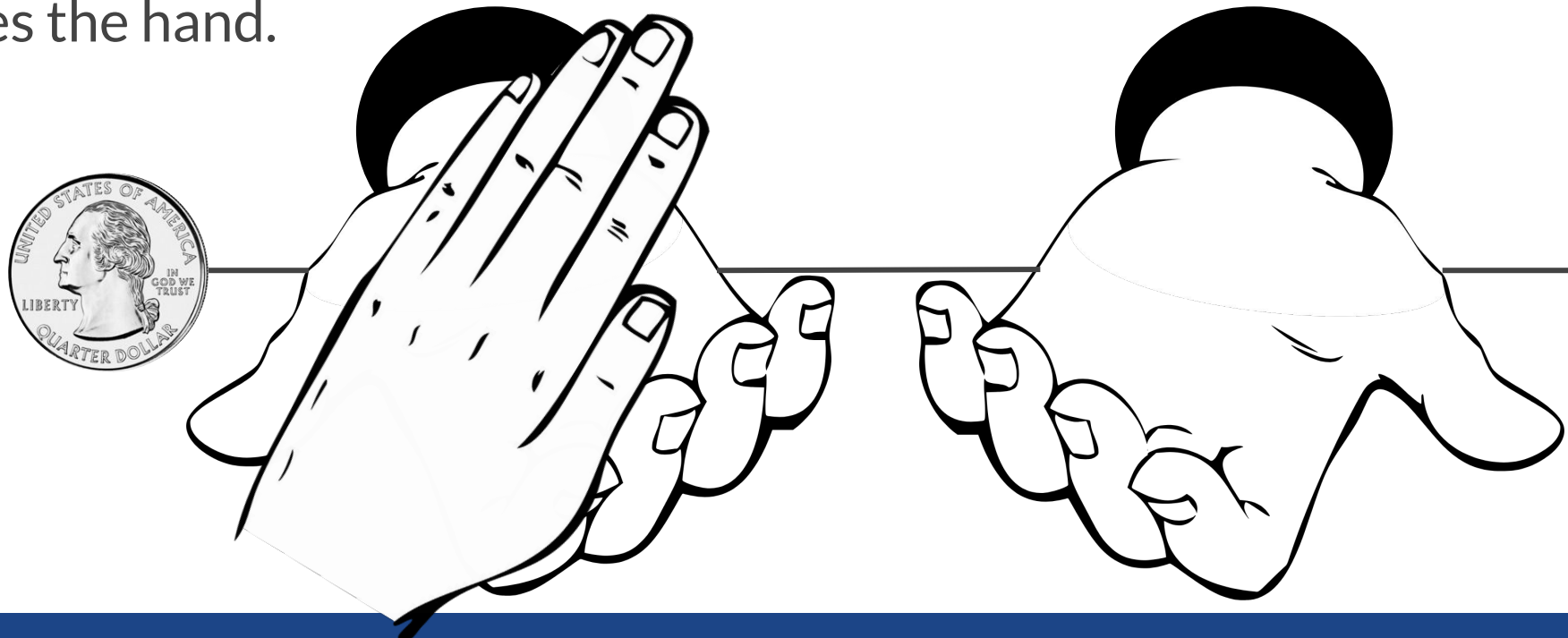
The Experiment

- 21 Therapeutic Touch Healers participate in a test
- The healers place their hands through an opaque screen
- The researcher flips a coin. If it lands heads, she places her hand over (but not touching) the healer's right hand, if tails over the healer's left hand.



The Experiment

- 21 Therapeutic Touch Healers participate in a test
- The healers place their hands through an opaque screen
- The researcher flips a coin. If it lands heads, she places her hand over (but not touching) the healer's right hand, if tails over the healer's left hand.
- The healer identifies the hand.



The Results

- 280 trials in total
- Of those, the healers correctly identified which hand 171 times
 - $171/280 = 61\%$ of the time.

What does this mean?

The Results

- 280 trials in total
- Of those, the healers correctly identified which hand 171 times
 - $171/280 = 61\%$ of the time.

*There are (at least) two plausible explanations for these results.
What are they?*

The Results

- If they just guess, the healers have a 50% chance of guessing correctly.
- 61% is better than that

...but

- It's only a little better. Maybe they just got lucky?
- What are the chances of getting 171 out 280 guesses correct just by chance?

To answer this we need to learn some probability

The Actual Experiment

Rosa, E., et al. (1998). *A Close Look at Therapeutic Touch*. JAMA, 279(13).

- Emily Rosa was 11 years old when she conducted her study!
- 21 TT healers did actually participate
- 280 trials
- *But her results were different. I'll share later...*

Simulations

- Without working out the mathematics, we can often use simulations to answer probabilistic questions.
 1. Assume whatever you are investigating is a random event.
 2. Assign probabilities based on reasonable assumptions.
 3. Simulate the event in question repeatedly.
 4. Look for patterns in the results.

Simulations

My claim: *I am a very skilled coin-flipper. I can't do it every time, but most times I can flip the coin in such a way that it comes up heads.*

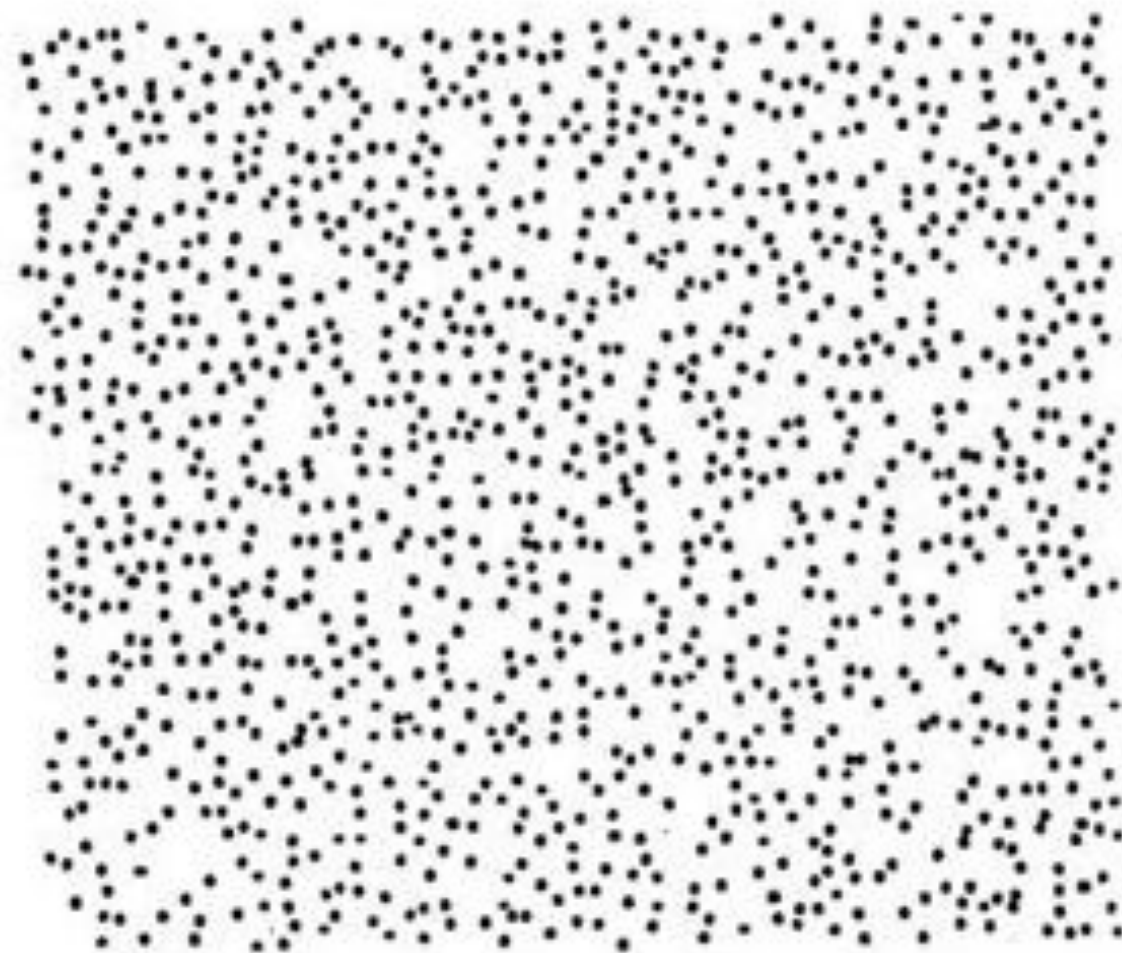
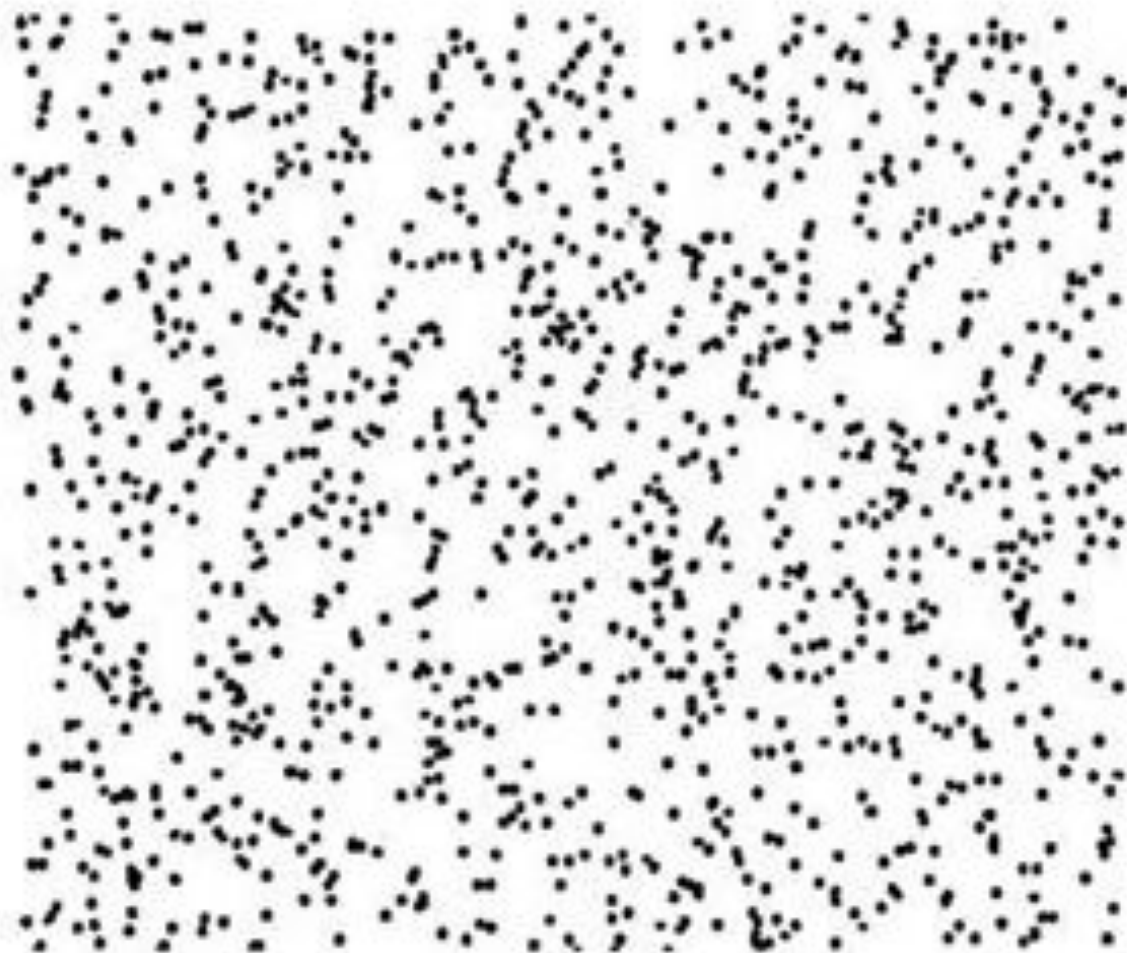
My evidence: I flipped a coin 17 times and it came up heads 11.

How credible is my claim?

What are the two explanations for these results?

Probability and Distributions

Random processes



Random processes

- A **random process** is a situation in which we know what outcomes could happen, but we don't know which particular outcome will happen.
- Examples:
 - coin tosses
 - dice rolls
 - Weather
 - Spotify Playlists
 - Whether the stock market goes up or down tomorrow, etc.

Probability

- *What's the probability of rolling a 3 with a fair die?*
 - Answer: $1/6$
 - **CLASSICAL APPROACH:** Theoretically speaking, all six sides of the die are equally likely (implied by the word "fair"), thus we expect 1 out of 6 rolls to be a 3.
- *In his career, Derrick Rose has made 82.4% of his free throws. What's the probability that he makes his next free throw?*
 - Answer: 0.824
 - **EMPIRICAL APPROACH:** Probabilities are based on past performance, and are found by looking at the proportion of "successes" out of all trials.
- *What's the probability that it rains on Saturday?*
 - Answer: 40% (according to weather.gov)
 - **SUBJECTIVE APPROACH:** Probabilities measure degree of belief. Can be different for different people. Often the result of some predictive model.

Probability

There are several possible interpretations of probability but they (almost) completely agree on the mathematical rules probability must follow:

- $P(A)$ = Probability of event A
- $0 \leq P(A) \leq 1$ (fractions, decimals, percents)
- $P(A)$ measures the likelihood of event A happening

Probability

There are several possible interpretations of probability but they (almost) completely agree on the mathematical rules probability must follow:

- $P(A)$ = Probability of event A
- $0 \leq P(A) \leq 1$ (fractions, decimals, percents)
- $P(A)$ measures the likelihood of event A happening

Frequentist interpretation:

- The probability of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times.

Probability

There are several possible interpretations of probability but they (almost) completely agree on the mathematical rules probability must follow:

- $P(A)$ = Probability of event A
- $0 \leq P(A) \leq 1$ (fractions, decimals, percents)
- $P(A)$ measures the likelihood of event A happening

Frequentist interpretation:

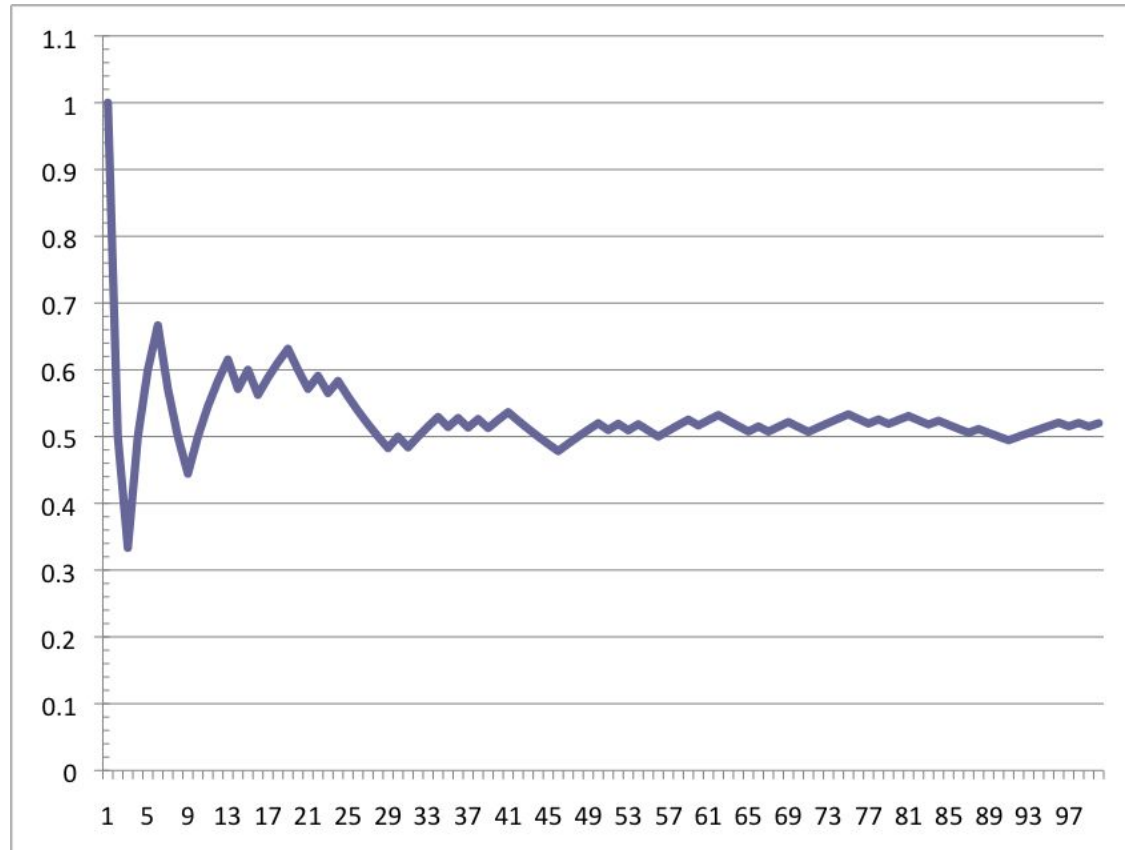
- The probability of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times.

Bayesian interpretation:

Law of large numbers

Law of large numbers states that as more observations are collected, the proportion of occurrences with a particular outcome converges to the probability of that outcome.

Prop of head by
of coin tosses



Law of large numbers and Cancer Rates

Highest Brain Cancer Rates	Lowest Brain Cancer Rates
South Dakota	Wyoming
Nebraska	Vermont
Alaska	North Dakota
Delaware	Hawaii
Maine	DC

What is common among these states?

Law of large numbers and Cancer Rates

Highest Brain Cancer Rates	Lowest Brain Cancer Rates
South Dakota	Wyoming
Nebraska	Vermont
Alaska	North Dakota
Delaware	Hawaii
Maine	DC

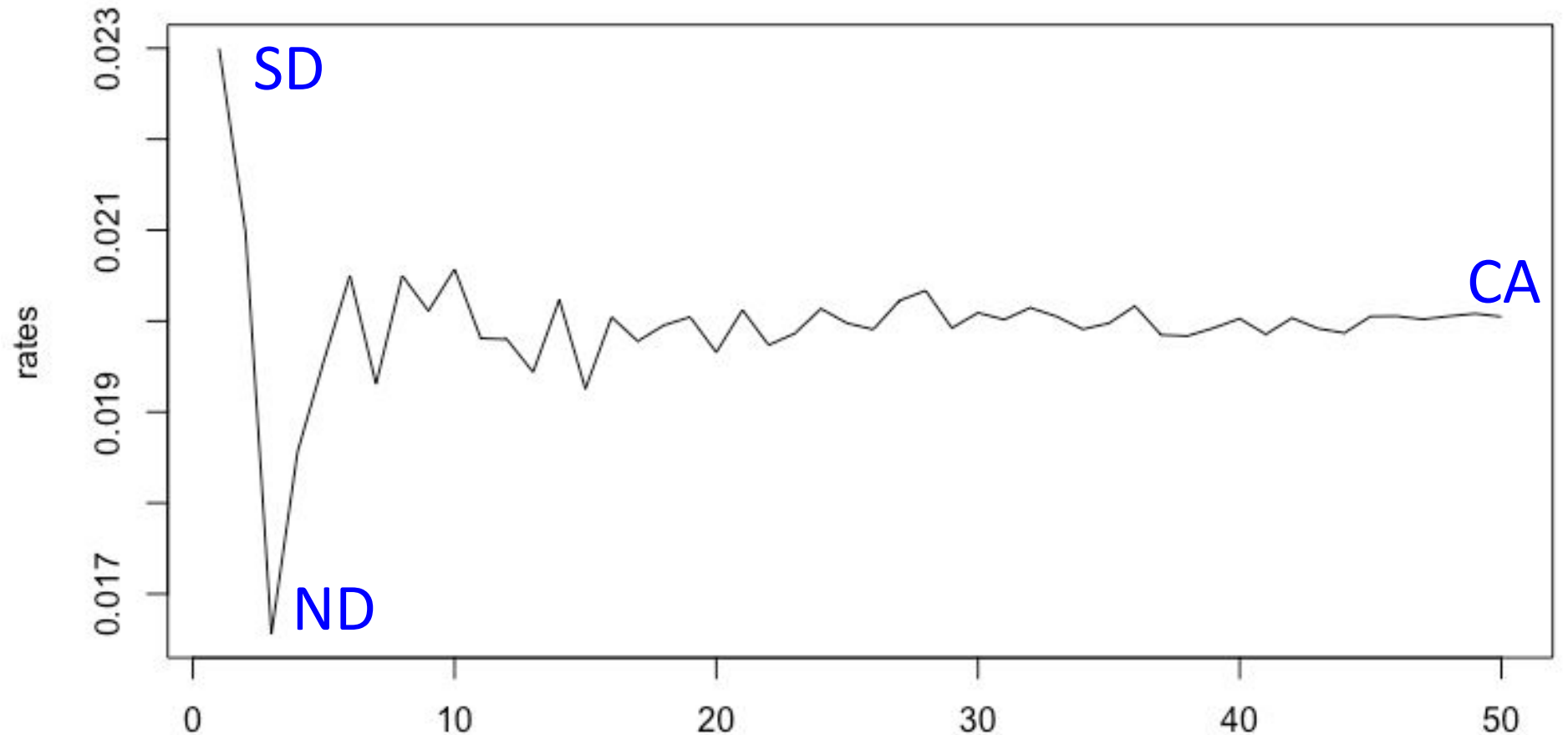
These are all states with low populations.

State with smaller populations that are more prone to swings in one direction of the other.

Law of large numbers and Cancer Rates

Hypothetical Brain
Cancer Rates by
State

(US Brain cancer
rate is typically ~1%
to 2% in a year)



Calculating Probabilities

- Good Things / All Things
 - When all possible outcomes are equally likely, we calculate a probability as the number of possible “good” outcomes, divided by the total number of possible outcomes.
- What’s the probability of rolling a 4 on a six sided die (assuming all sides are equally likely)?

$$P(\text{Rolling a 4}) = 1 \text{ good thing (the 4)} / 6 \text{ total things} = \frac{1}{6}$$

Calculating Probabilities

- If I randomly choose a letter, what is the probability it is a vowel?

$$P(\text{Vowel}) = 5 \text{ vowels (a, e, i, o, u)} / 26 \text{ letters} = 5/26$$

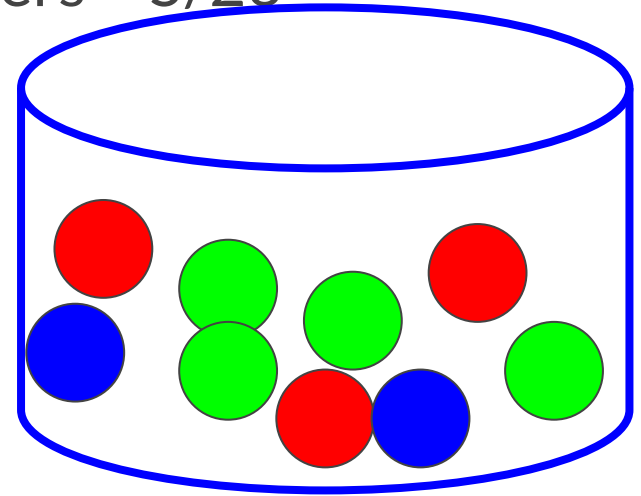
- I randomly choose a marble from the jar.

Calculating Probabilities

- If I randomly choose a letter, what is the probability it is a vowel?

$$P(\text{Vowel}) = 5 \text{ vowels (a, e, i, o, u)} / 26 \text{ letters} = 5/26$$

- I randomly choose a marble from the jar.

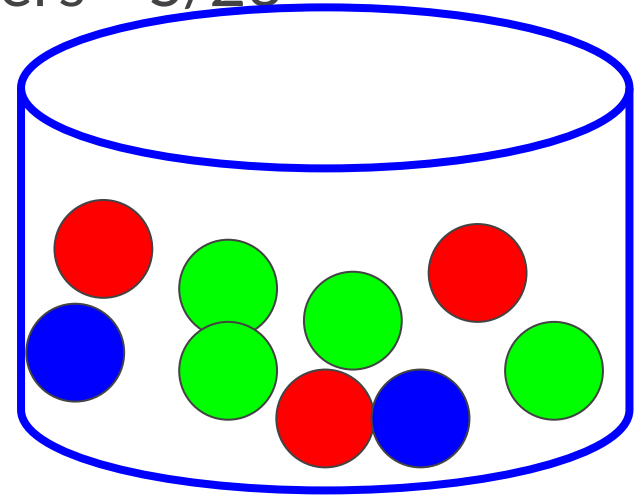


Calculating Probabilities

- If I randomly choose a letter, what is the probability it is a vowel?

$$P(\text{Vowel}) = 5 \text{ vowels (a, e, i, o, u)} / 26 \text{ letters} = 5/26$$

- I randomly choose a marble from the jar.
 - What is the probability I choose red?

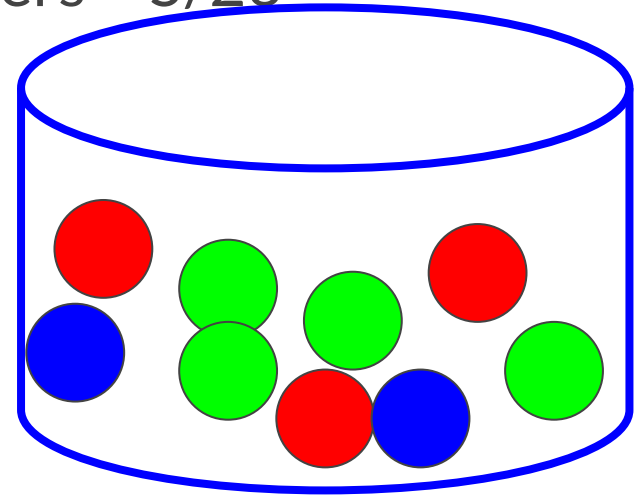


Calculating Probabilities

- If I randomly choose a letter, what is the probability it is a vowel?

$$P(\text{Vowel}) = 5 \text{ vowels (a, e, i, o, u)} / 26 \text{ letters} = 5/26$$

- I randomly choose a marble from the jar.
 - What is the probability I choose red?
 - $P(\text{Red}) = 3 \text{ red} / 9 \text{ marbles} = 3/9 = 1/3$

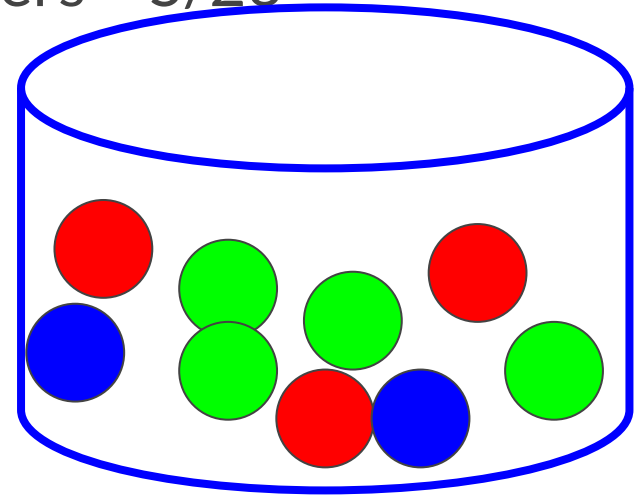


Calculating Probabilities

- If I randomly choose a letter, what is the probability it is a vowel?

$$P(\text{Vowel}) = 5 \text{ vowels (a, e, i, o, u)} / 26 \text{ letters} = 5/26$$

- I randomly choose a marble from the jar.
 - What is the probability I choose red?
 - $P(\text{Red}) = 3 \text{ red} / 9 \text{ marbles} = 3/9 = 1/3$
 - What is the probability I don't choose blue?

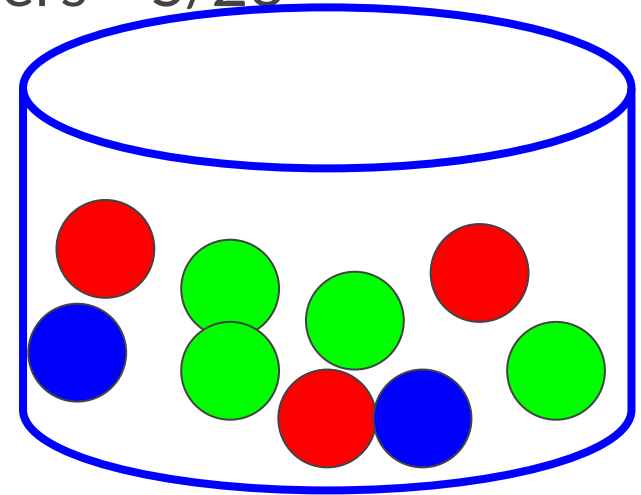


Calculating Probabilities

- If I randomly choose a letter, what is the probability it is a vowel?

$$P(\text{Vowel}) = 5 \text{ vowels (a, e, i, o, u)} / 26 \text{ letters} = 5/26$$

- I randomly choose a marble from the jar.
 - What is the probability I choose red?
 - $P(\text{Red}) = 3 \text{ red} / 9 \text{ marbles} = 3/9 = 1/3$
 - What is the probability I don't choose blue?
 - $P(\text{Not Blue}) = 7/9$

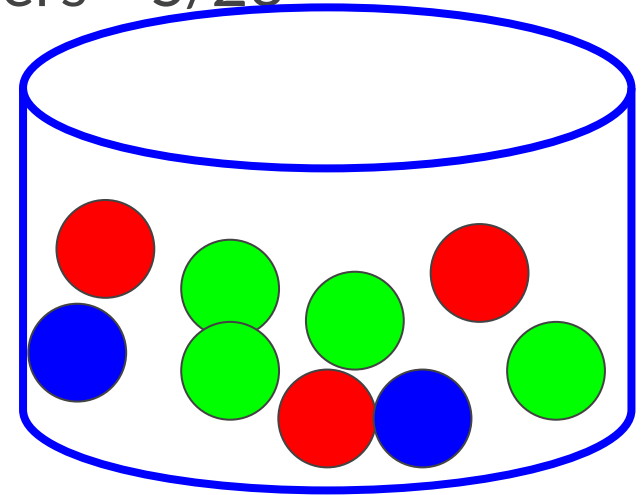


Calculating Probabilities

- If I randomly choose a letter, what is the probability it is a vowel?

$$P(\text{Vowel}) = 5 \text{ vowels (a, e, i, o, u)} / 26 \text{ letters} = 5/26$$

- I randomly choose a marble from the jar.
 - What is the probability I choose red?
 - $P(\text{Red}) = 3 \text{ red} / 9 \text{ marbles} = 3/9 = 1/3$
 - What is the probability I don't choose blue?
 - $P(\text{Not Blue}) = 7/9$

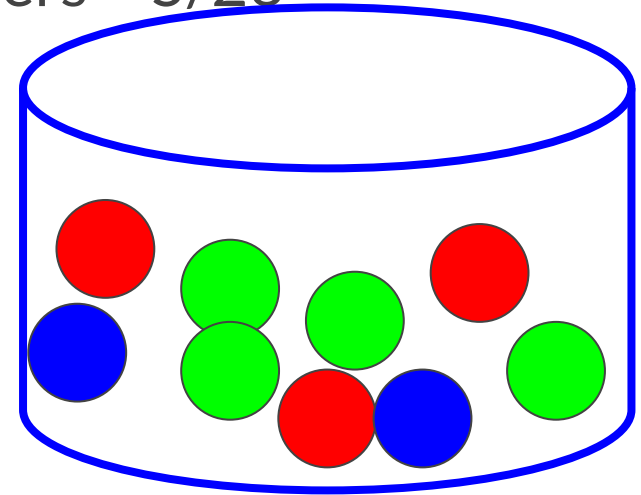


Calculating Probabilities

- If I randomly choose a letter, what is the probability it is a vowel?

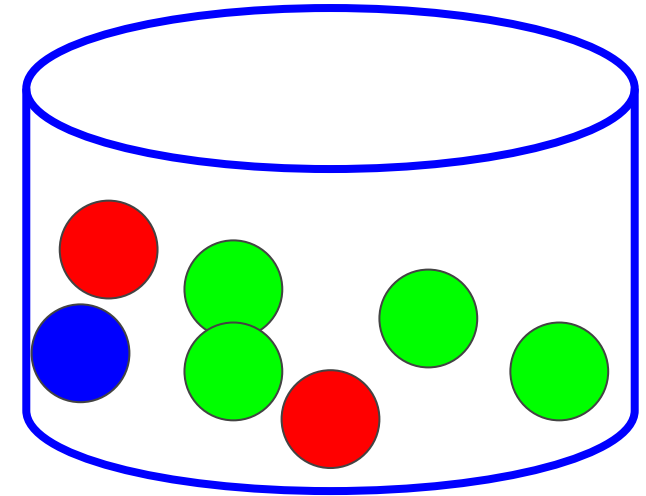
$$P(\text{Vowel}) = 5 \text{ vowels (a, e, i, o, u)} / 26 \text{ letters} = 5/26$$

- I randomly choose a marble from the jar.
 - What is the probability I choose red?
 - $P(\text{Red}) = 3 \text{ red} / 9 \text{ marbles} = 3/9 = 1/3$
 - What is the probability I don't choose blue?
 - $P(\text{Not Blue}) = 7/9$



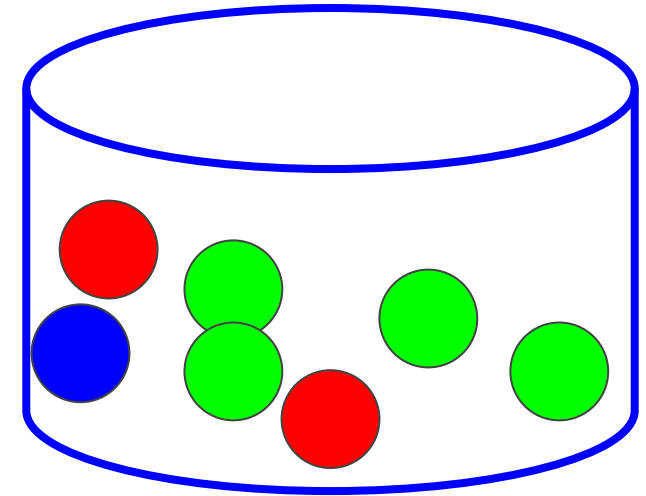
Sample Space

- We call the list of all possible outcomes, the *SAMPLE SPACE* and we usually label it S .
 - Sample space for flipping a single coin:
 - $S = \{\text{Head}, \text{Tail}\}$
 - Sample space for drawing a single marble:



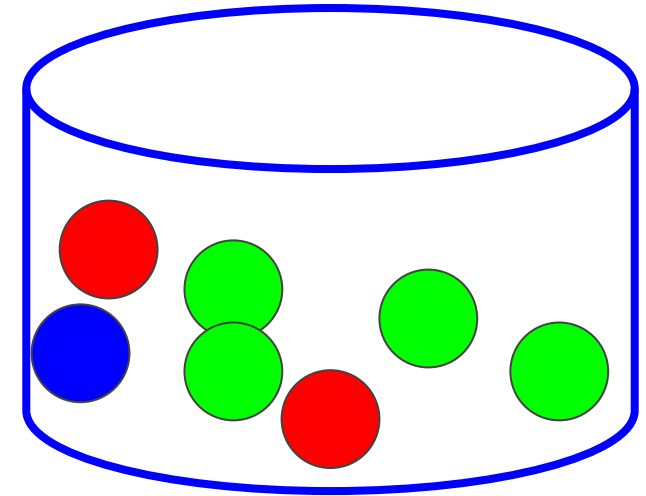
Sample Space

- We call the list of all possible outcomes, the *SAMPLE SPACE* and we usually label it S .
 - Sample space for flipping a single coin:
 - $S = \{\text{Head}, \text{Tail}\}$
 - Sample space for drawing a single marble:
 - $S = \{\text{Red}, \text{Green}, \text{Blue}\}$



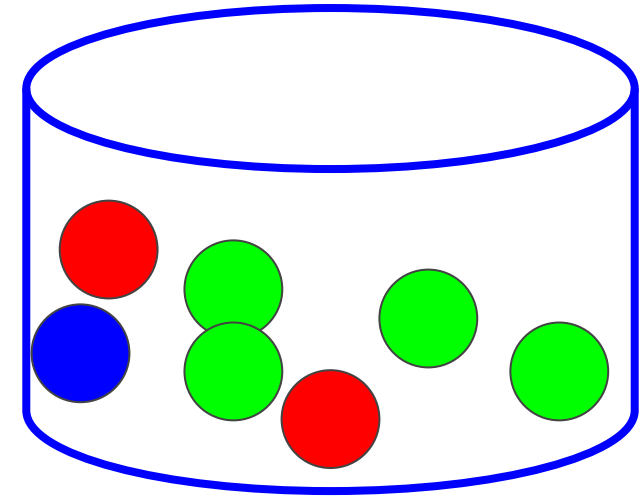
Sample Space

- We call the list of all possible outcomes, the *SAMPLE SPACE* and we usually label it S .
 - Sample space for flipping a single coin:
 - $S = \{\text{Head}, \text{Tail}\}$
 - Sample space for drawing a single marble:
 - $S = \{\text{Red}, \text{Green}, \text{Blue}\}$
 - Sample space for flipping two coins:



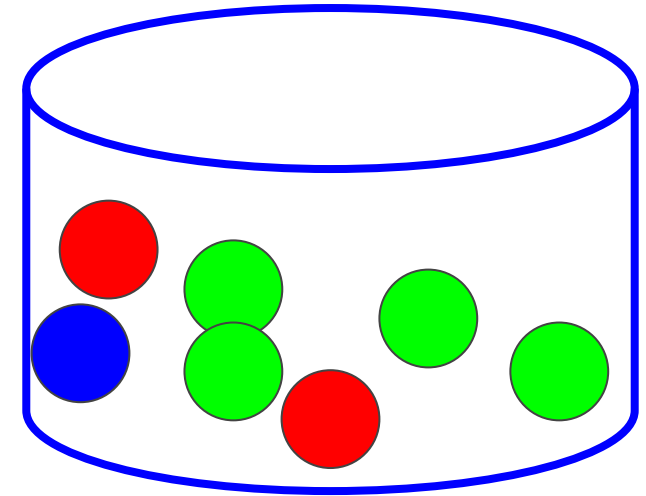
Sample Space

- We call the list of all possible outcomes, the *SAMPLE SPACE* and we usually label it S .
 - Sample space for flipping a single coin:
 - $S = \{\text{Head}, \text{Tail}\}$
 - Sample space for drawing a single marble:
 - $S = \{\text{Red}, \text{Green}, \text{Blue}\}$
 - Sample space for flipping two coins:
 - $S = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$ (why 4 elements?)



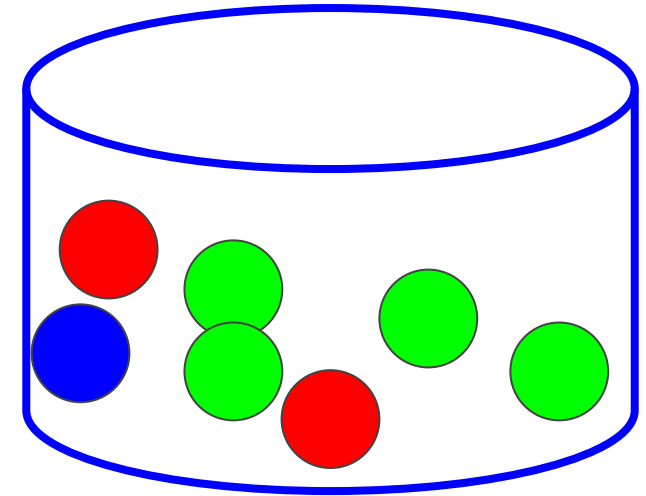
Sample Space

- We call the list of all possible outcomes, the *SAMPLE SPACE* and we usually label it S .
 - Sample space for flipping a single coin:
 - $S = \{\text{Head}, \text{Tail}\}$
 - Sample space for drawing a single marble:
 - $S = \{\text{Red}, \text{Green}, \text{Blue}\}$
 - Sample space for flipping two coins:
 - $S = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$ (why 4 elements?)
 - Sample space for drawing two marbles:



Sample Space

- We call the list of all possible outcomes, the *SAMPLE SPACE* and we usually label it S .
 - Sample space for flipping a single coin:
 - $S = \{\text{Head}, \text{Tail}\}$
 - Sample space for drawing a single marble:
 - $S = \{\text{Red}, \text{Green}, \text{Blue}\}$
 - Sample space for flipping two coins:
 - $S = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$ (why 4 elements?)
 - Sample space for drawing two marbles:
 - $S = \{\text{RR}, \text{RG}, \text{RB}, \text{GG}, \text{GR}, \text{GB}, \text{BR}, \text{BG}\}$



Complementary Events

Complementary events are *two* mutually exclusive events whose probabilities add up to 1. Together they span the entire sample space.

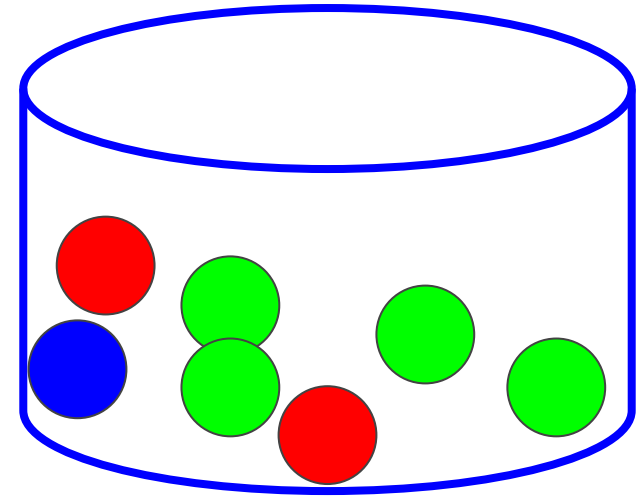
- You flip coin. If we know that it does not come up heads, what is the result?
 - { ~~H~~, **T** } Head and Tail are **complementary** outcomes.
- You flip two coins, if we know that they are not both tails, what are the possible results?

$$S = \{ \text{HH}, \text{HT}, \text{TH}, \text{TT} \}$$

{ **HH**, **HT**, **TH** } and { **TT** } are **complementary**

Applying Complementary Events

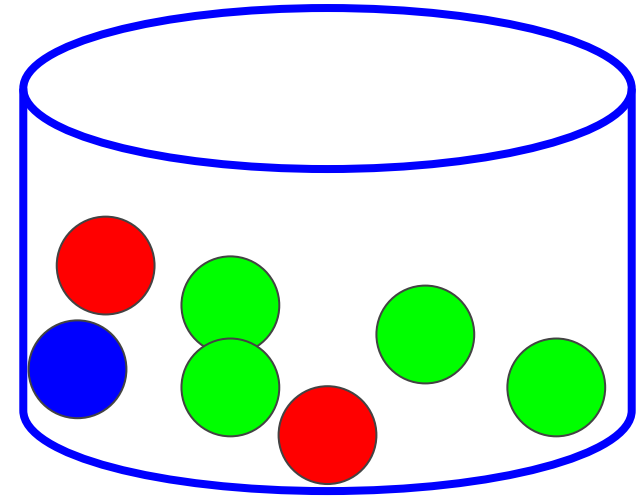
*What's the probability of drawing a **red** or a **green** marble?*



Applying Complementary Events

*What's the probability of drawing a **red** or a **green** marble?*

- $S = \{\text{Red}, \text{Green}, \text{Blue}\}$
- $\{\text{Red}, \text{Green}\}$ and $\{\text{Blue}\}$ are complementary

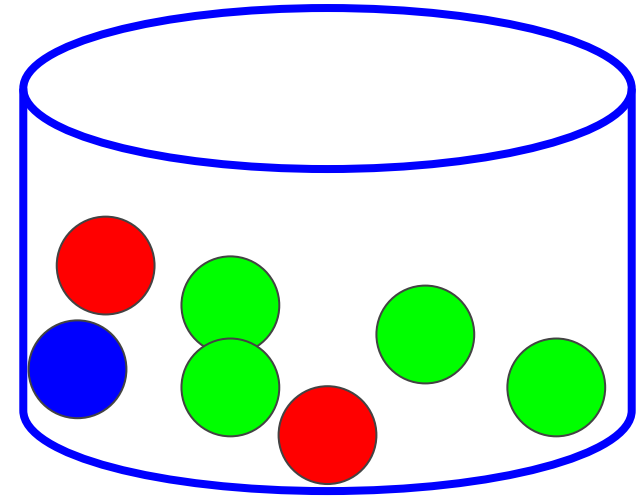


Applying Complementary Events

What's the probability of drawing a *red* or a *green* marble?

- $S = \{\text{Red}, \text{Green}, \text{Blue}\}$
- $\{\text{Red}, \text{Green}\}$ and $\{\text{Blue}\}$ are complementary

$$P(\text{Red or Green}) = 1 - P(\text{Blue}) = 1 - 1/7 = 6/7$$



Disjoint vs. Complementary

- Do the sum of probabilities of two disjoint outcomes always add up to 1?
- Do the sum of probabilities of two complementary outcomes always add up to 1?
- Complementary events are always disjoint
- However, disjoint events are not necessarily always complementary.

Question

In a survey, 52% of respondents said they are Democrats. What is the probability that a randomly selected respondent from this sample is a Republican?

- (a) 0.48
- (b) more than 0.48
- (c) less than 0.48
- (d) cannot calculate using only the information given

Combining Probabilities

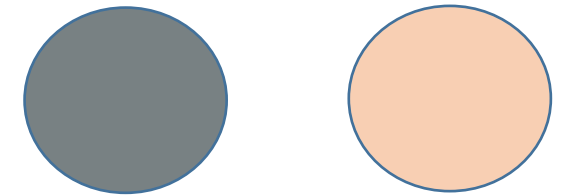
- Probabilities are, unfortunately, not usually that simple.
- Often we calculate probabilities for complex events by combining the probabilities for simpler events.
- How we combine depends on how constituent events are related.
 - Do both events happen at the same time?
 - Is one event contingent on the other?

**How can two (or more) events
happen?**

Combining Probabilities: Terminology

Intersection of two events: A **AND** B

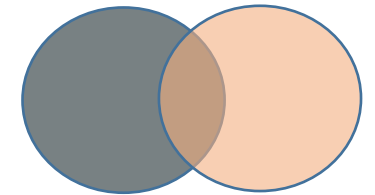
- The set of all outcomes where **both A and B** happen.



Disjoint (mutually exclusive) outcomes: Cannot happen at the same time.

$$P(A \text{ and } B) = 0$$

- The outcome of a single coin toss cannot be a head and a tail.
- A student both cannot fail and pass a class.
- A single card drawn from a deck cannot be an ace and a queen.



$$P(A \text{ and } B) \neq 0$$

Combining Probabilities: Terminology

Union of two events: A **OR** B

- The set of all outcomes where either A or B (or both) happen.

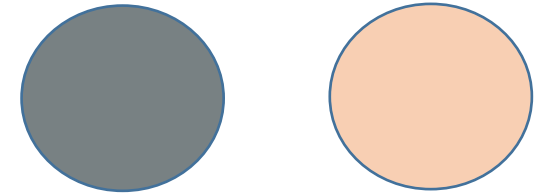
Combining Probabilities: Terminology

Union of two events: A **OR** B

- The set of all outcomes where either A or B (or both) happen.

Disjoint (mutually exclusive) outcomes:

- Cannot happen at the same time.



$$P(\text{A or B}) = P(A) + P(B)$$

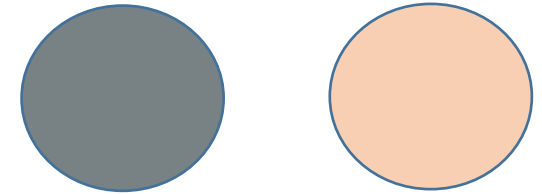
Combining Probabilities: Terminology

Union of two events: A **OR** B

- The set of all outcomes where either A or B (or both) happen.

Disjoint (mutually exclusive) outcomes:

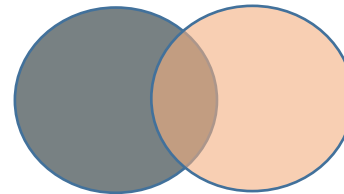
- Cannot happen at the same time.



$$P(\text{A or B}) = P(A) + P(B)$$

Non-disjoint outcomes:

- Can happen at the same time.



$$P(\text{A or B}) = P(A) + P(B) - P(\text{A and B})$$

Example: Intersection of Events

What is the probability of drawing a jack from a well shuffled full deck?

$$P(\text{Jack}) = 4/52 = 1/13$$

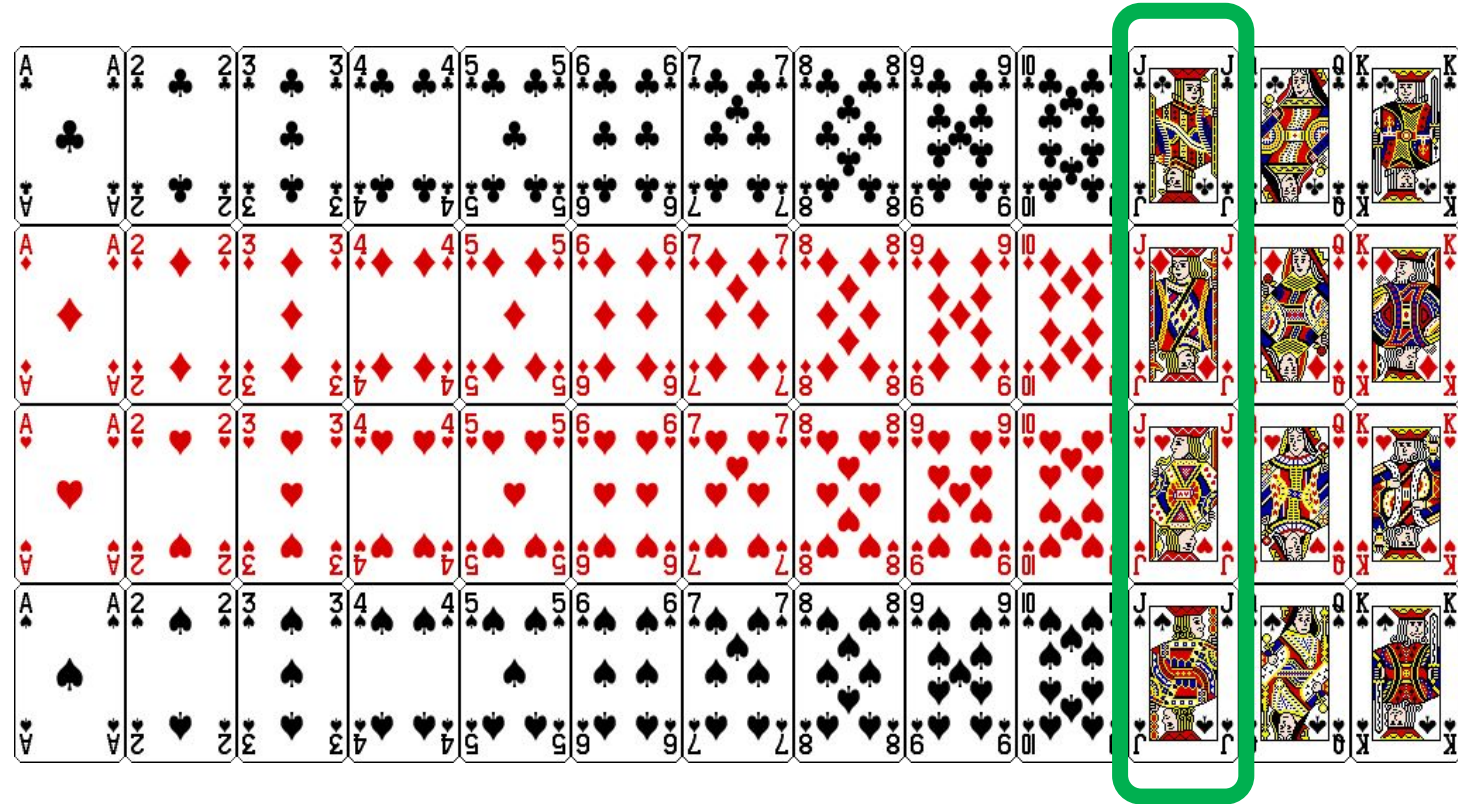


Figure from <http://www.milefoot.com/math/discrete/counting/cardfreq.htm>

Example: Intersection of Events

What is the probability of drawing a red card from a well shuffled full deck?

$$P(\text{red}) = 26/52 = 1/2$$

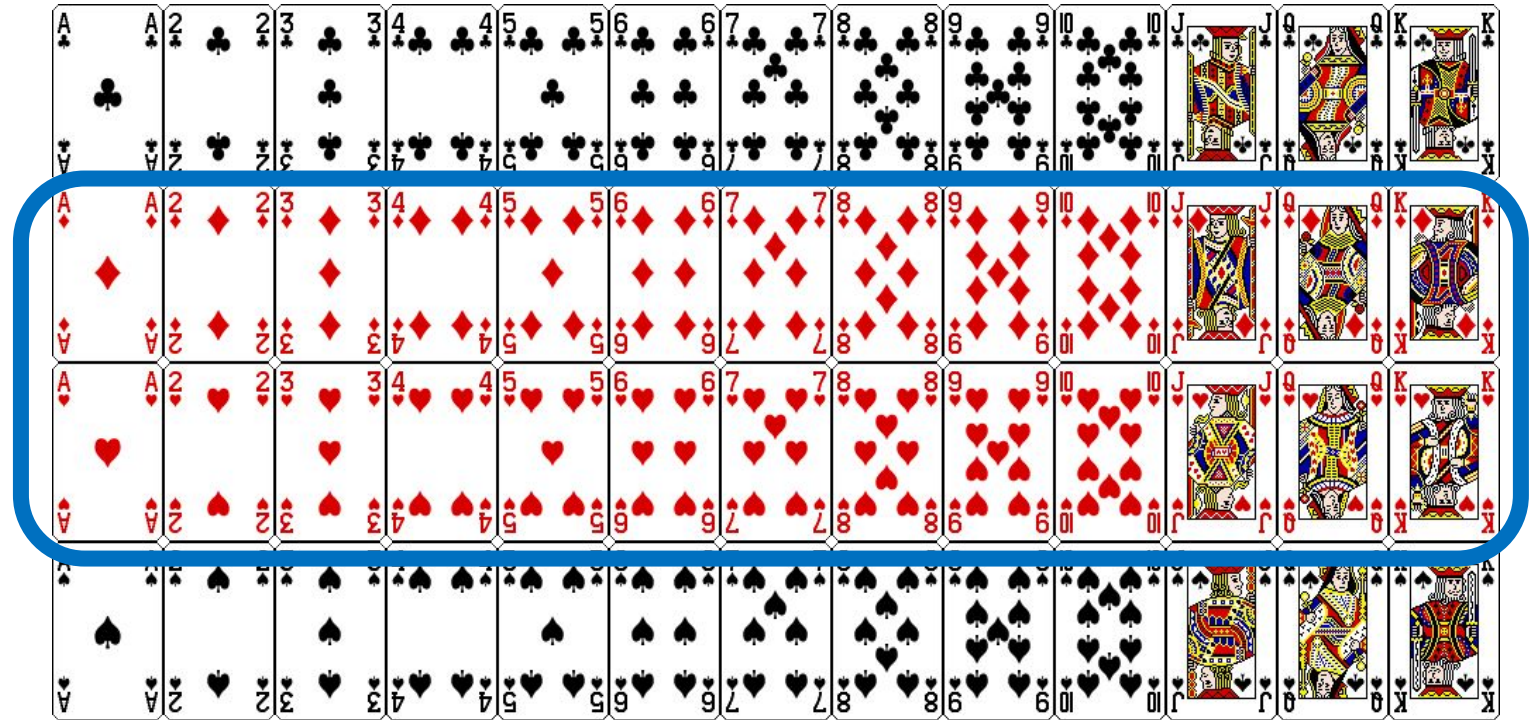


Figure from <http://www.milefoot.com/math/discrete/counting/cardfreq.htm>

Example: Intersection of Events

What is the probability of drawing a card that is **red** and a **jack** from a well shuffled full deck?

$$P(\text{red and jack}) \\ = 2/52 = 1/26$$

$P(\text{red and jack}) \neq 0$
so “being red” and
“being a jack” are
non-disjoint

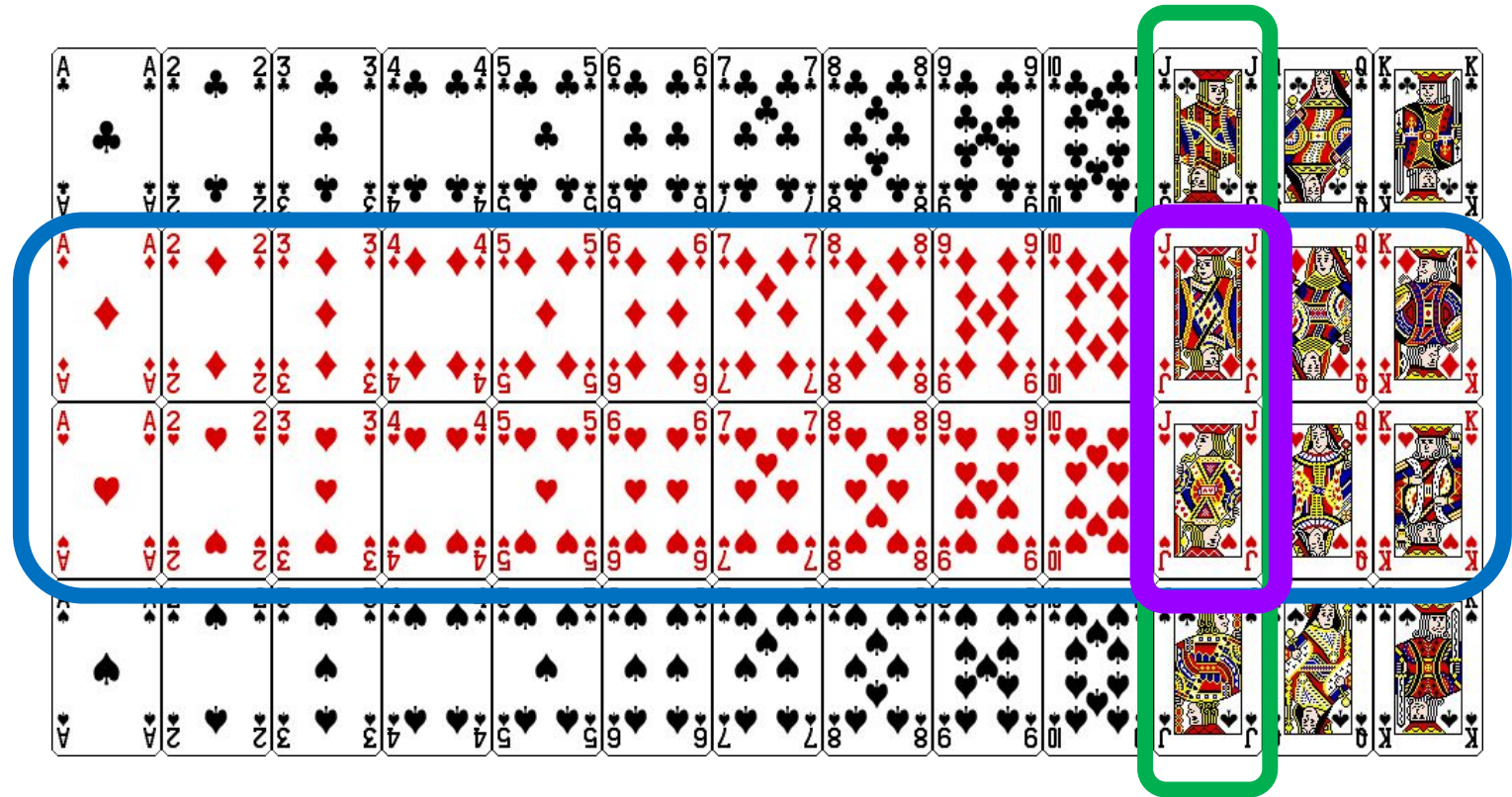


Figure from <http://www.milefoot.com/math/discrete/counting/cardfreq.htm>

Example: Intersection of Events

What is the probability of drawing a king from a well shuffled full deck?

$$P(\text{King}) = 4/52 = 1/13$$

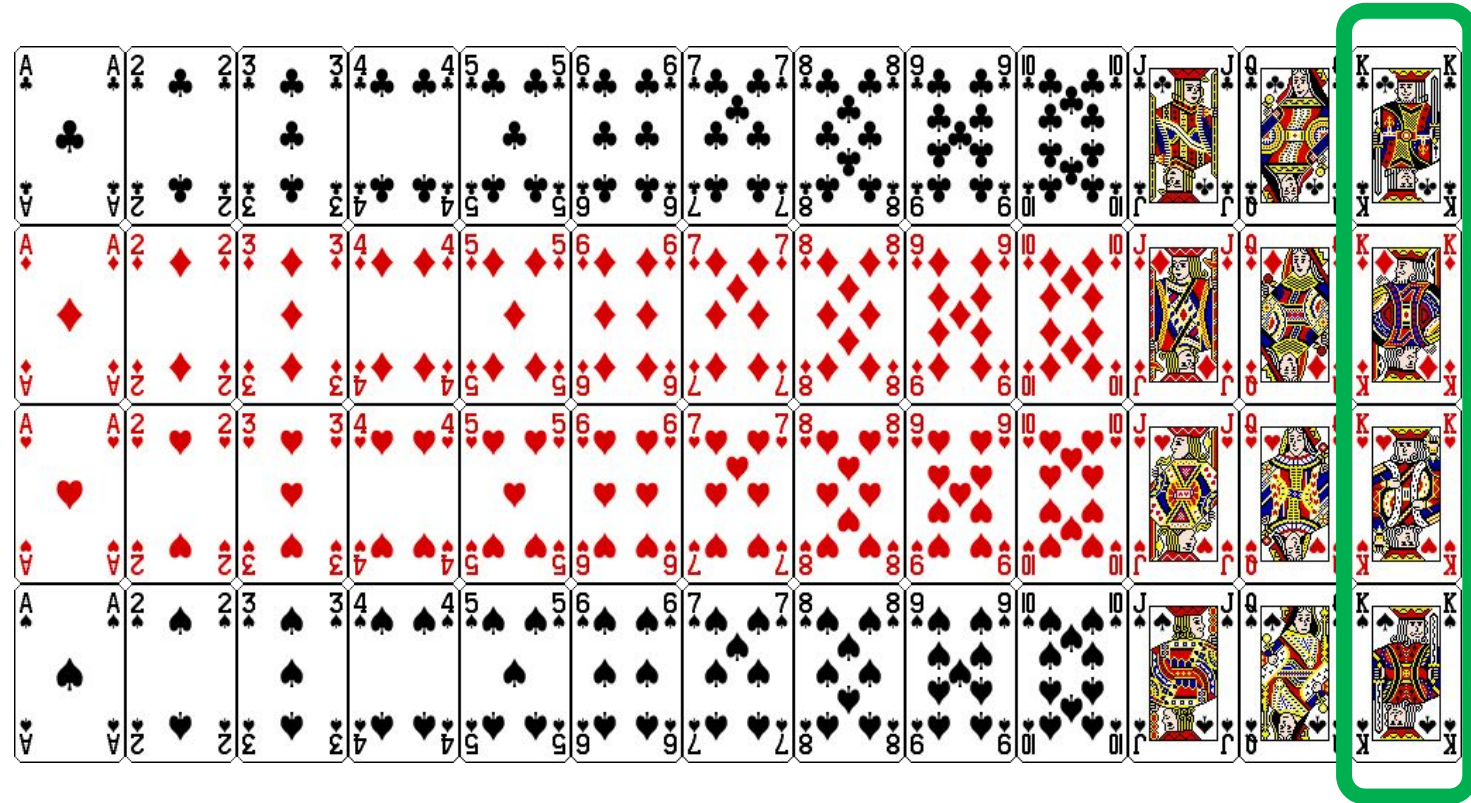


Figure from <http://www.milefoot.com/math/discrete/counting/cardfreq.htm>

Example: Intersection of Events

What is the probability of drawing a 3 from a well shuffled full deck?

$$P(3) = 4/52 = 1/13$$

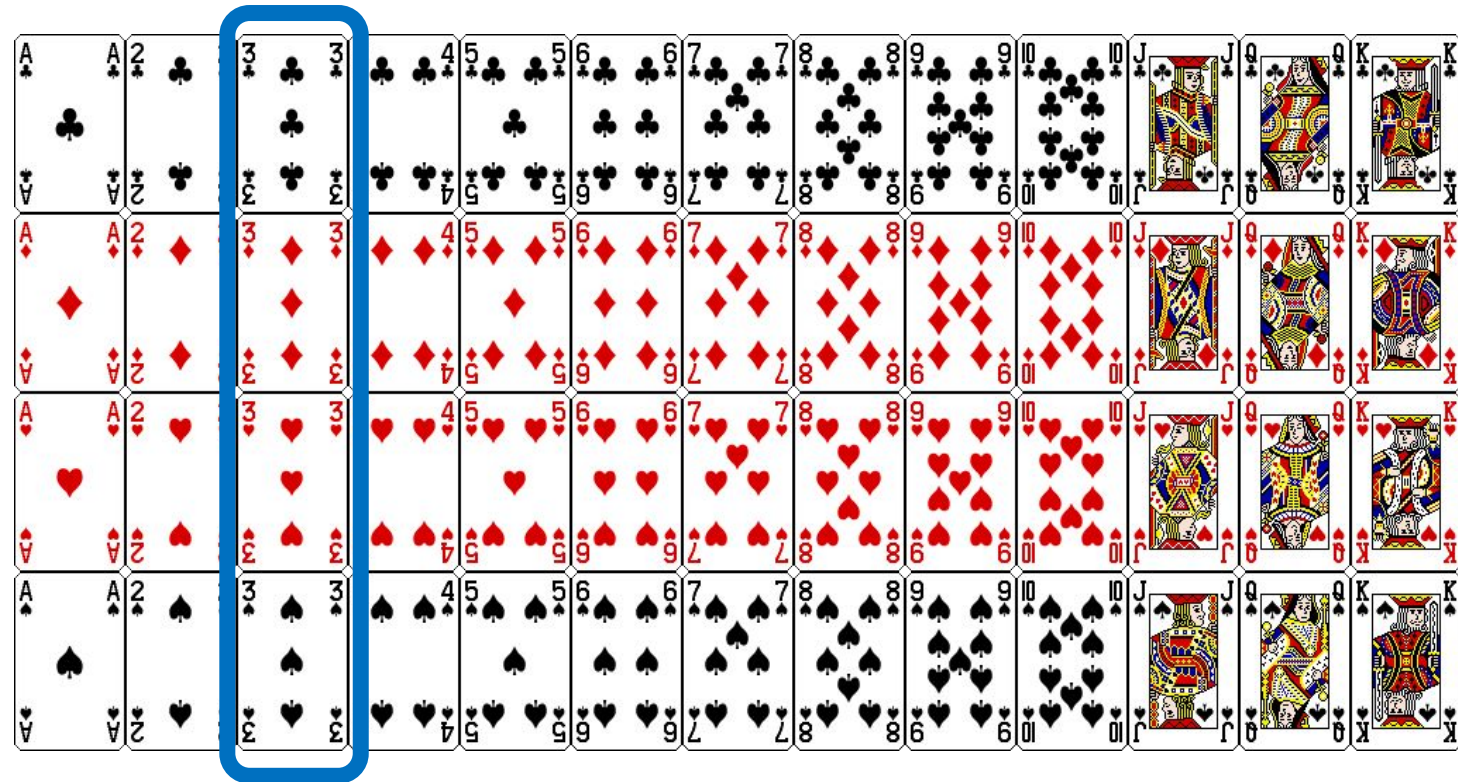


Figure from <http://www.milefoot.com/math/discrete/counting/cardfreq.htm>

Example: Intersection of Events

What is the probability of drawing a card that is a **king** and a **3** from a well shuffled full deck?

$$P(\text{king and } 3) \\ = 0/52 = 0$$

$P(\text{king and } 3) = 0$
so “being a king” and
“being a 3” are *disjoint*

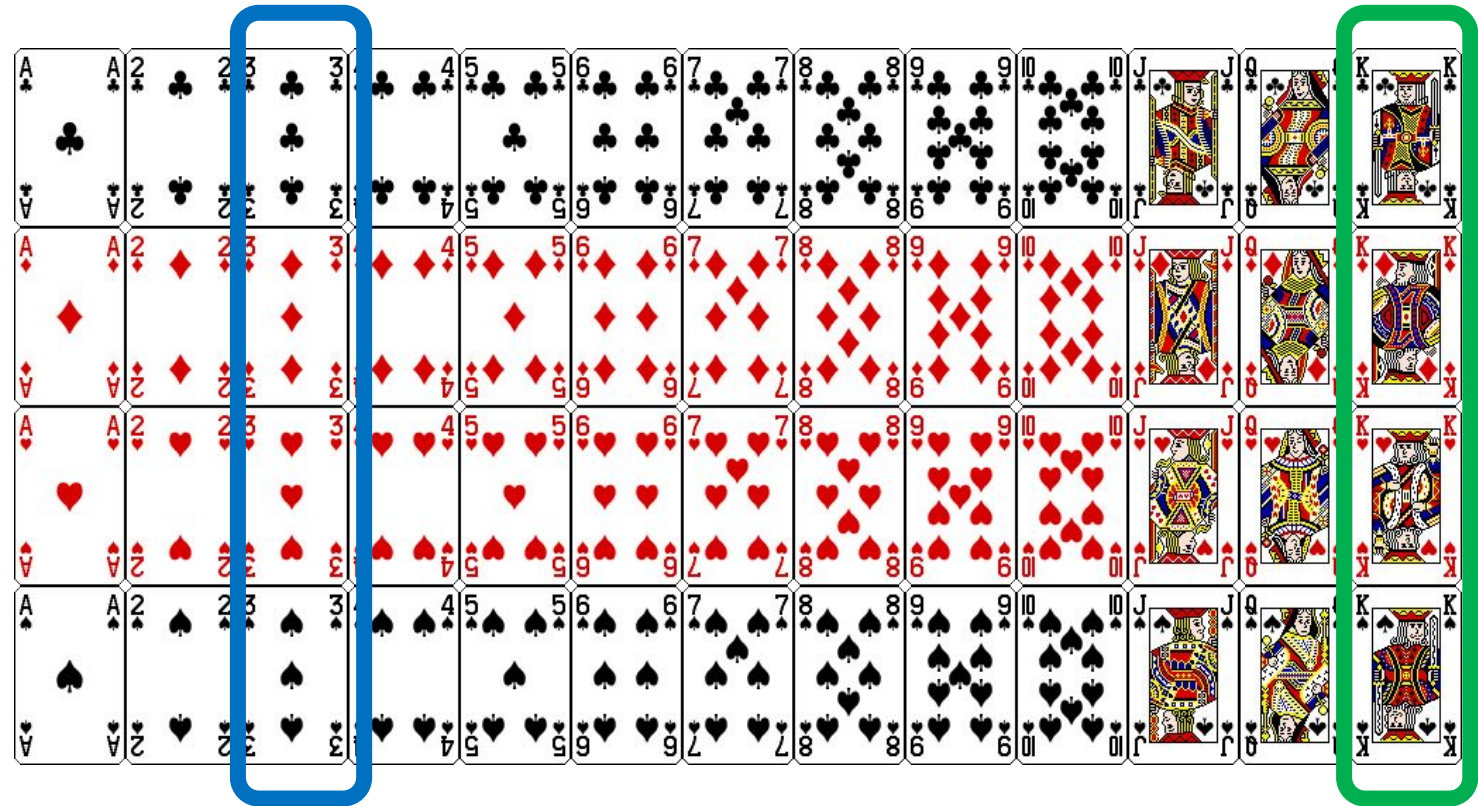


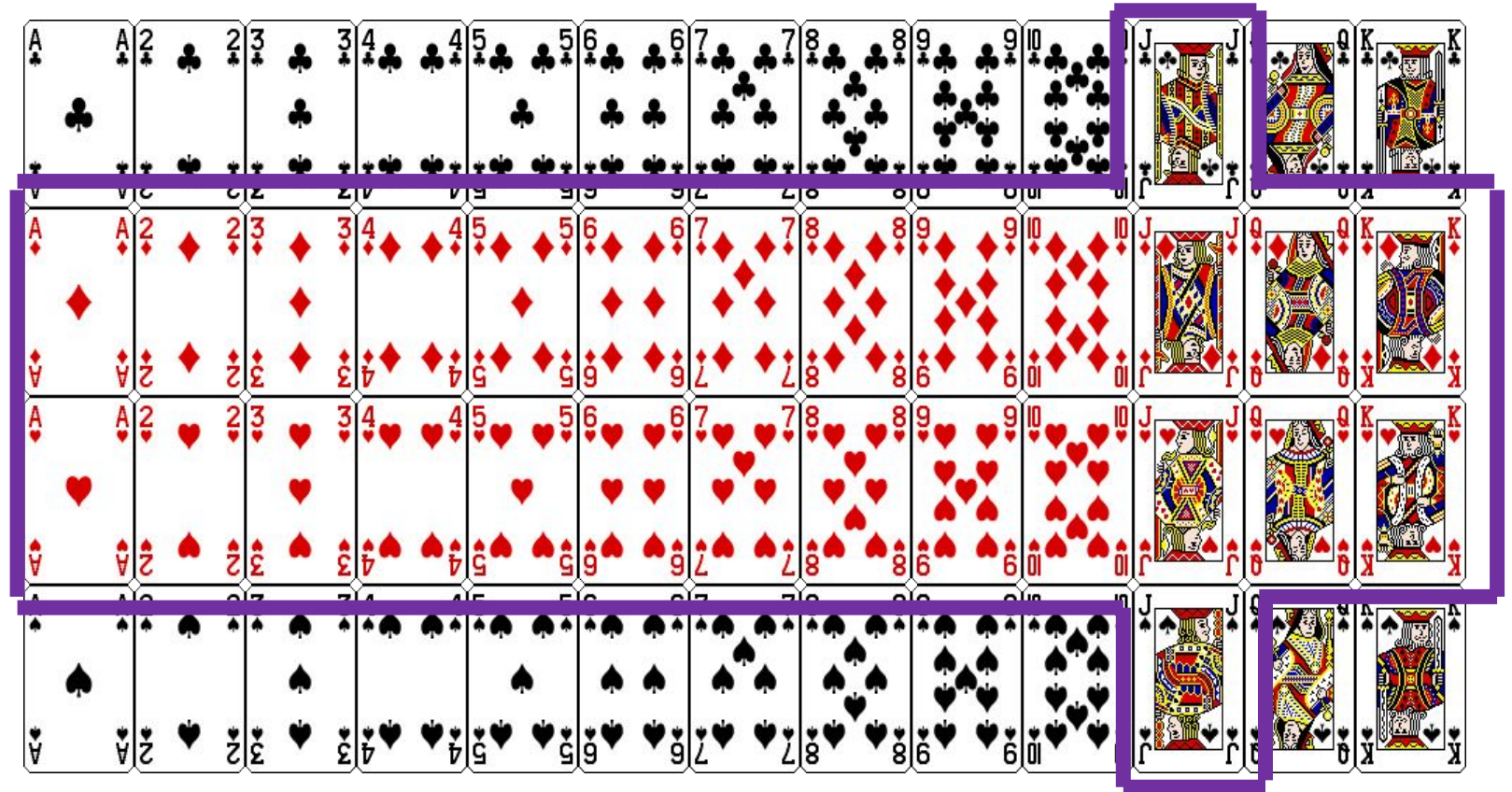
Figure from <http://www.milefoot.com/math/discrete/counting/cardfreq.htm>

Example: Union of Events

What is the probability of drawing a **red card** **OR** a **jack** from a well shuffled full deck?

Approach 1: Count the
good outcomes

$$P(\text{Red OR Jack}) = 28/52 \\ = 7/13$$



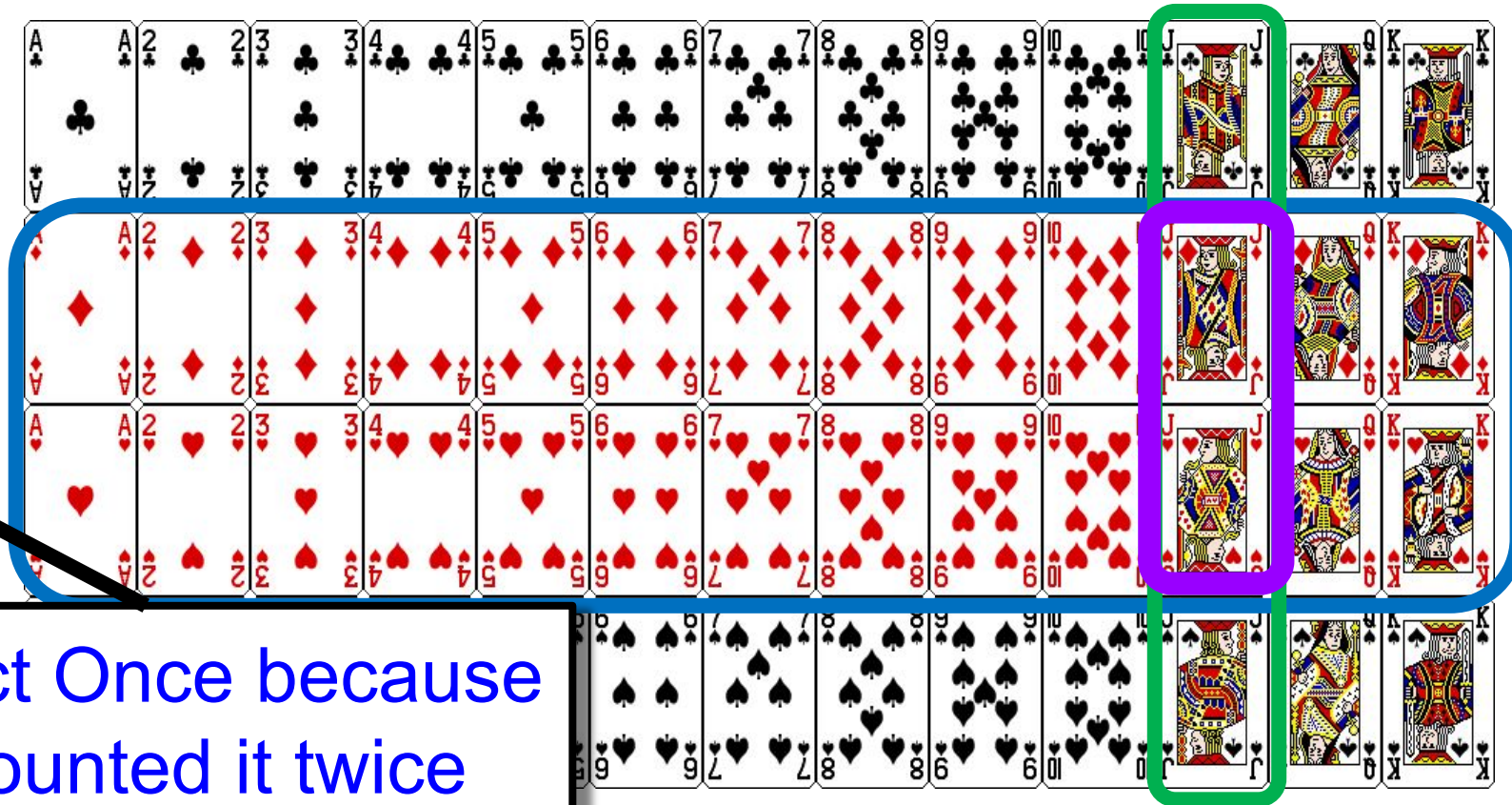
Example: Union of Events

What is the probability of drawing a **red card** **OR** a **jack** from a well shuffled full deck?

Approach 2: Consider the events separately

$$\begin{aligned} P(\text{Red OR Jack}) &= P(\text{Red}) + P(\text{Jack}) - P(\text{Red and Jack}) \\ &= 26/52 + 4/52 - 2/52 \\ &= 28/52 = 7/13 \end{aligned}$$

Subtract Once because
we counted it twice

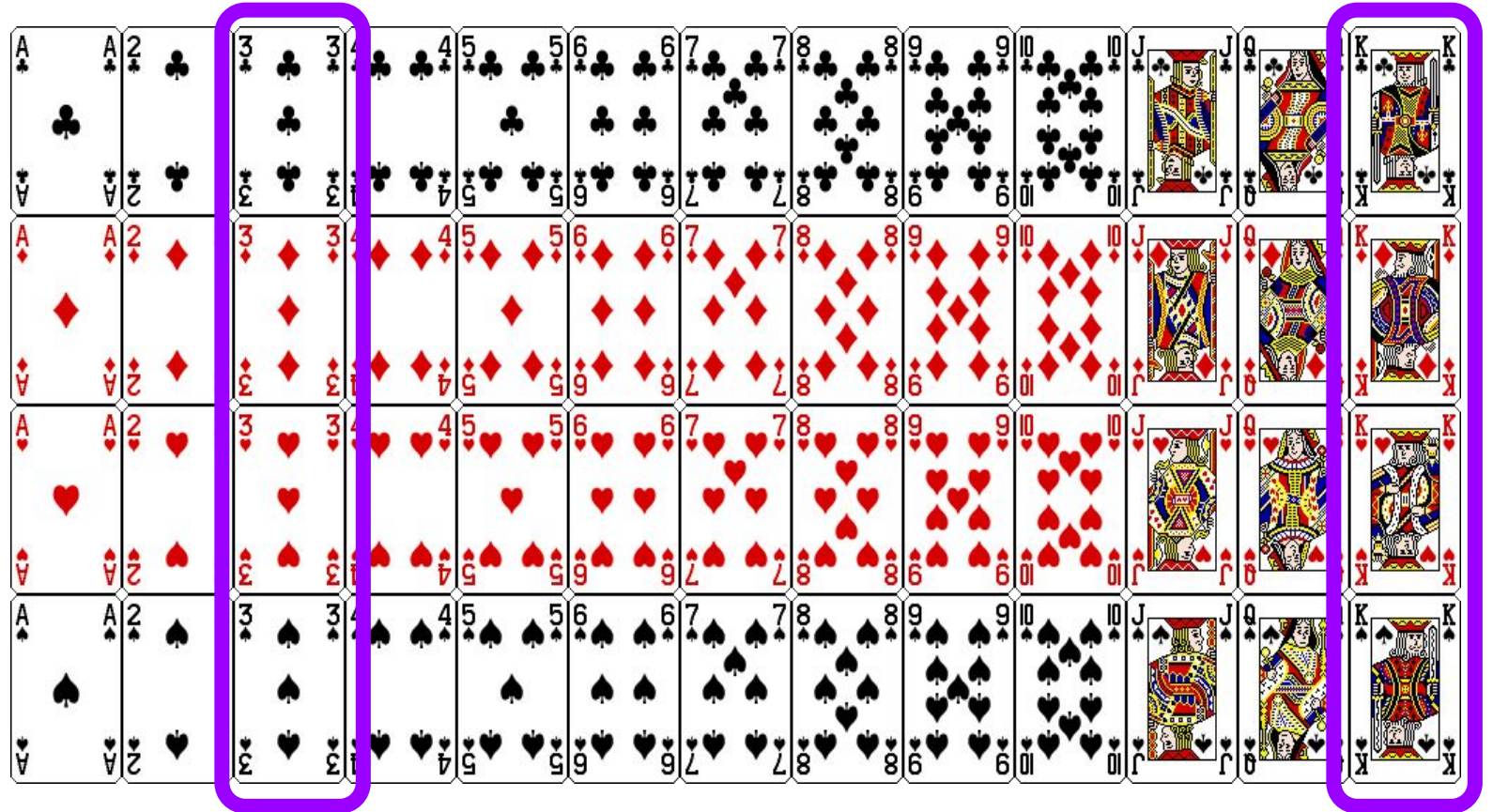


Example: Union of Events

What is the probability of drawing a **king** OR a **3** from a well shuffled full deck?

Approach 1: Count the
good outcomes

$$\begin{aligned} P(\text{king OR } 3) &= 8/52 \\ &= 2/13 \end{aligned}$$



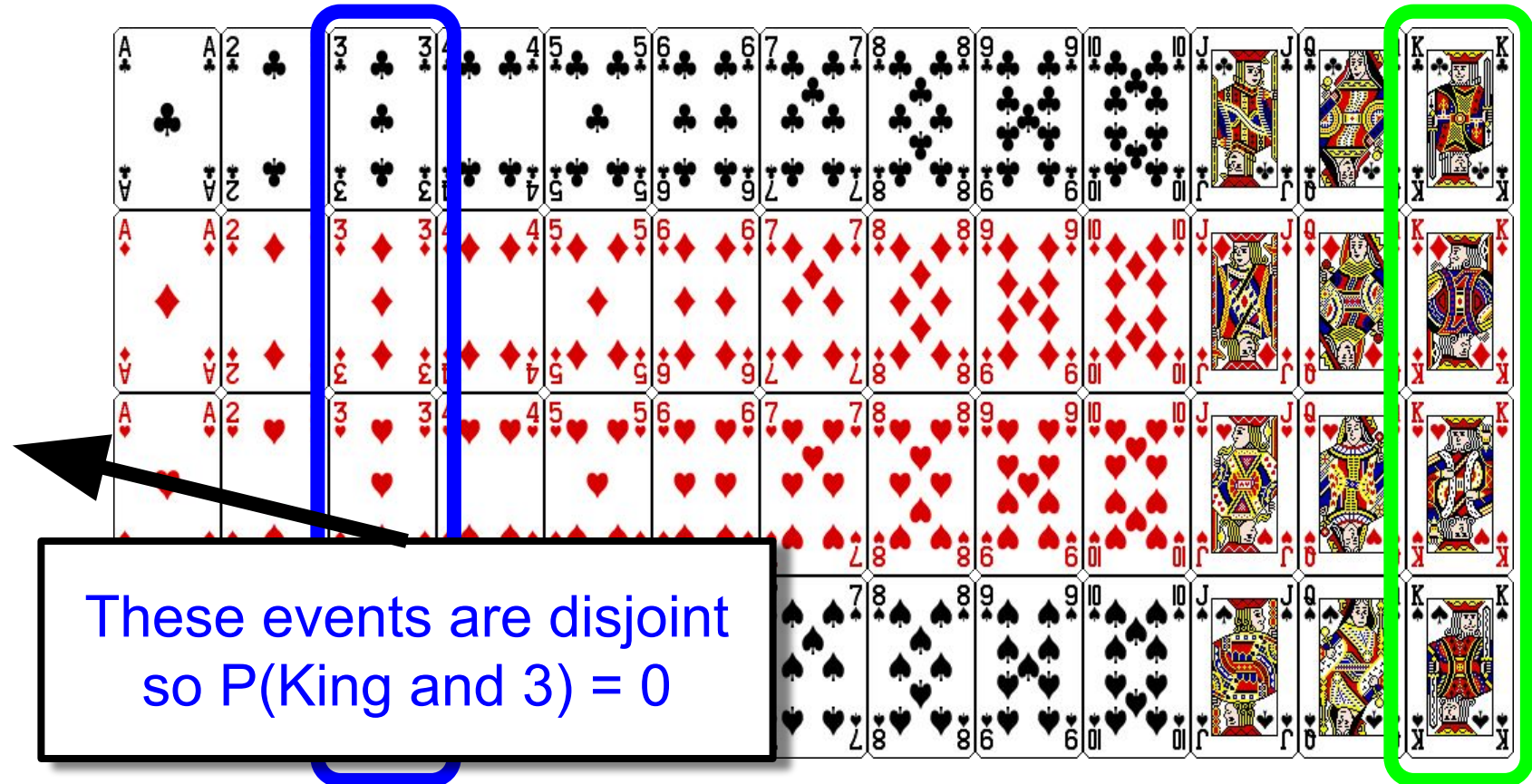
Example: Union of Events

What is the probability of drawing a **king** OR a **3** from a well shuffled full deck?

Approach 2: Consider the events separately

$$\begin{aligned} P(\text{King OR } 3) \\ = P(\text{King}) + P(3) - P(\text{King AND } 3) \end{aligned}$$

These events are disjoint
so $P(\text{King and } 3) = 0$



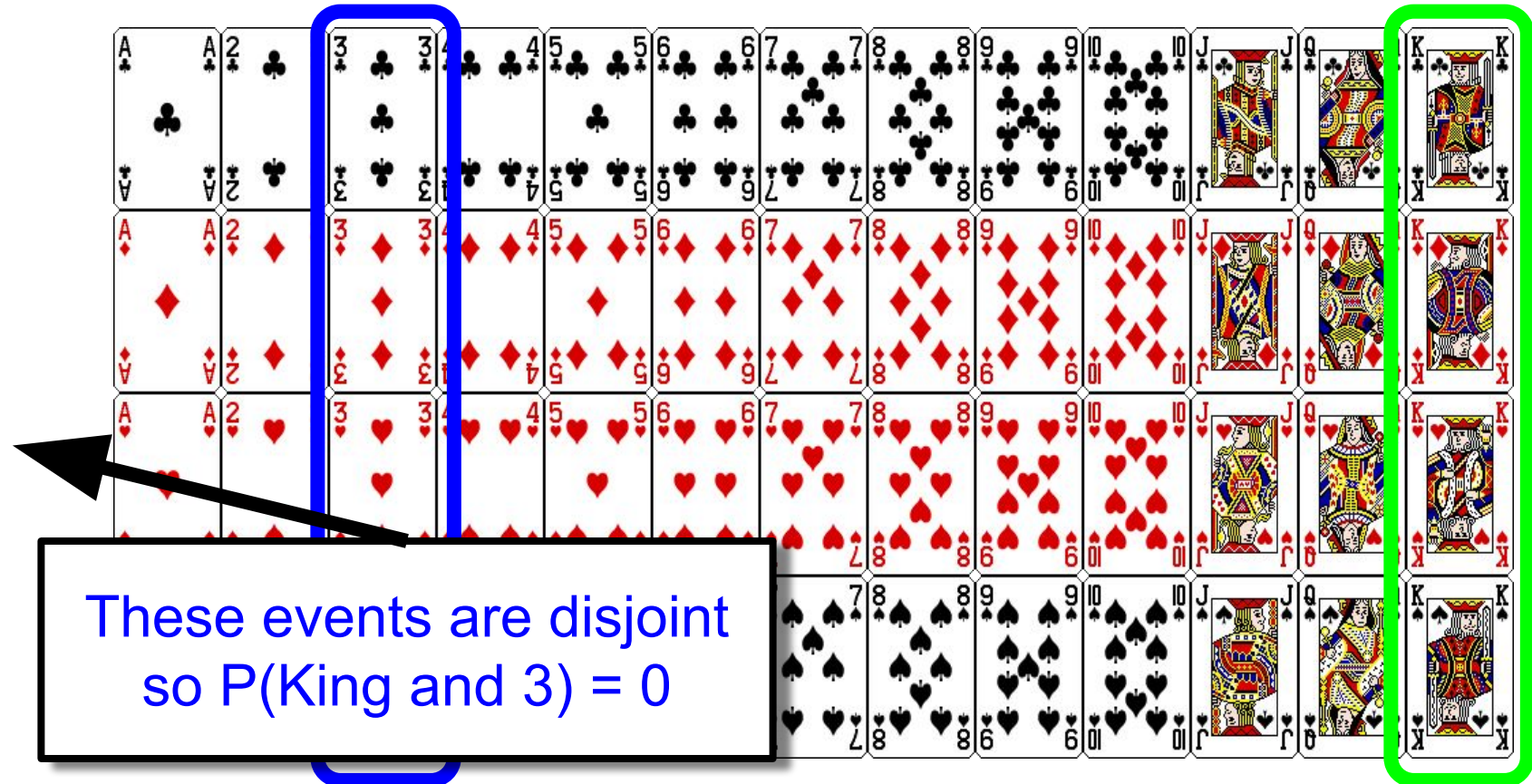
Example: Union of Events

What is the probability of drawing a **red card** **OR** a **jack** from a well shuffled full deck?

Approach 2: Consider the events separately

$$\begin{aligned} P(\text{King OR } 3) &= P(\text{King}) + P(3) - P(\text{King AND } 3) \\ &= P(\text{King}) + P(3) - 0 \end{aligned}$$

These events are disjoint
so $P(\text{King and } 3) = 0$



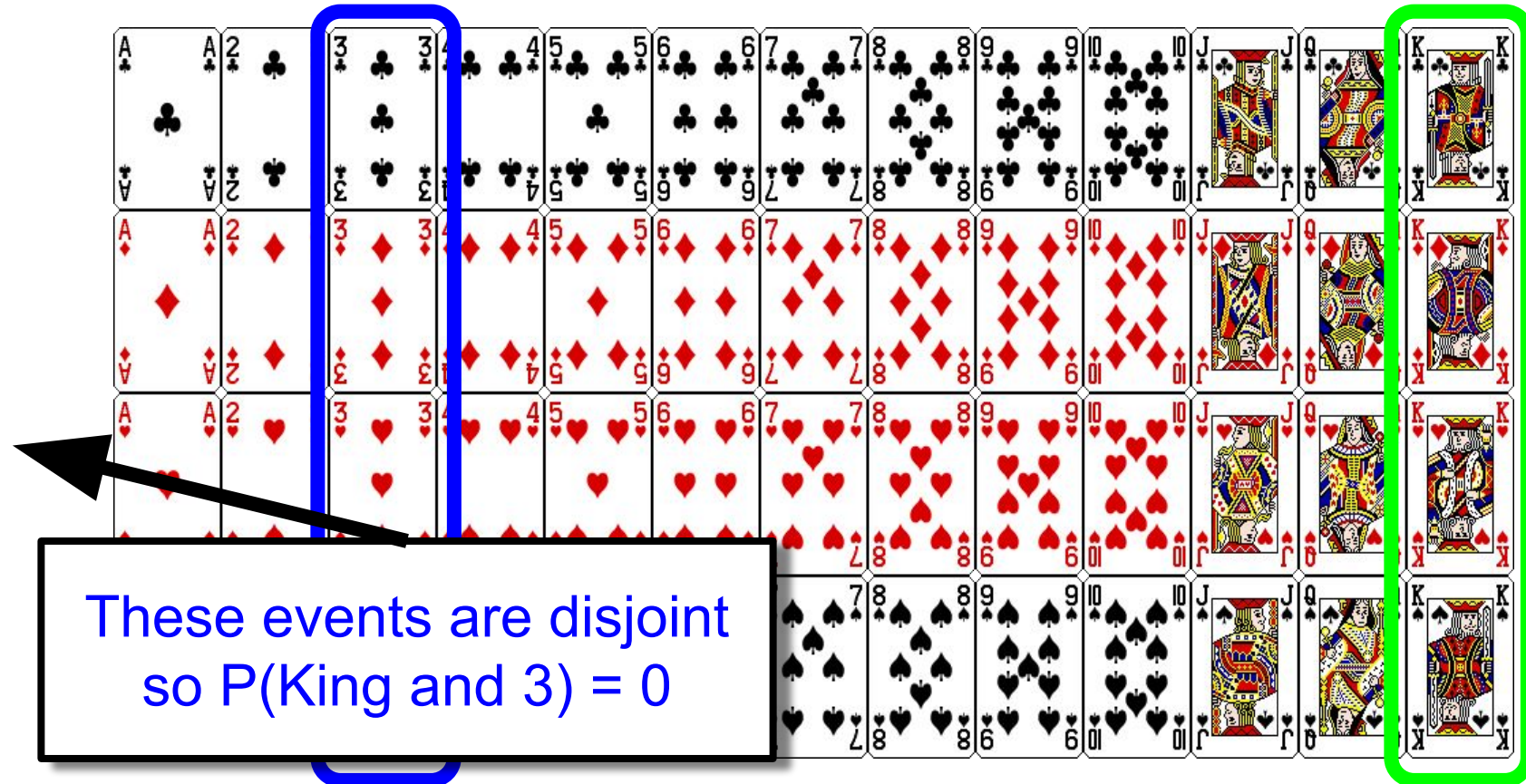
Example: Union of Events

What is the probability of drawing a **red card** OR a **jack** from a well shuffled full deck?

Approach 2: Consider the events separately

$$\begin{aligned} P(\text{King OR } 3) &= P(\text{King}) + P(3) - P(\text{King AND } 3) \\ &= P(\text{King}) + P(3) - 0 \\ &= P(\text{King}) + P(3) \end{aligned}$$

These events are disjoint
so $P(\text{King and } 3) = 0$



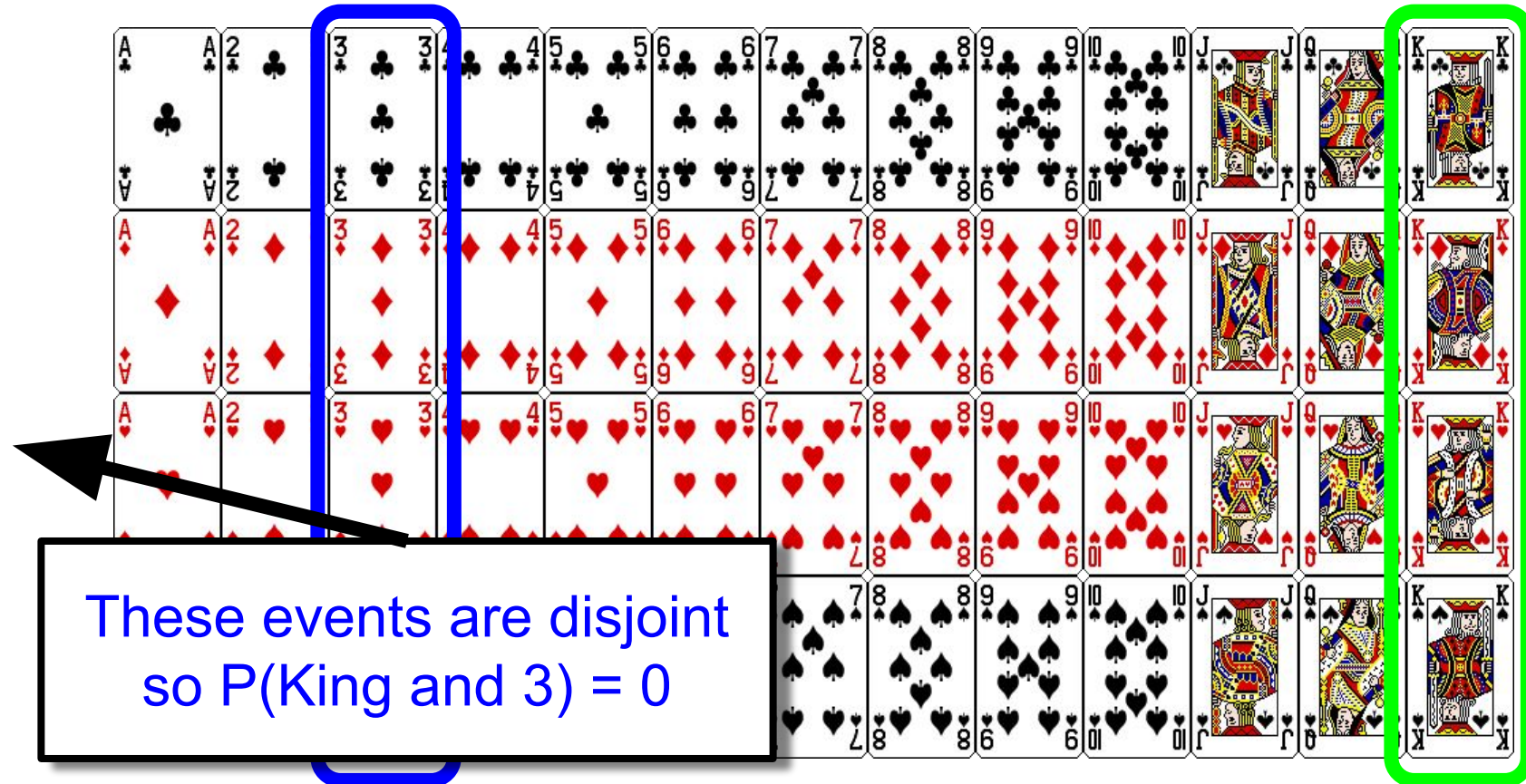
Example: Union of Events

What is the probability of drawing a **red card** OR a **jack** from a well shuffled full deck?

Approach 2: Consider the events separately

$$\begin{aligned} P(\text{King OR } 3) &= P(\text{King}) + P(3) - P(\text{King AND } 3) \\ &= P(\text{King}) + P(3) - 0 \\ &= P(\text{King}) + P(3) \\ &= 4/52 + 4/52 = 8/52 = 2/13 \end{aligned}$$

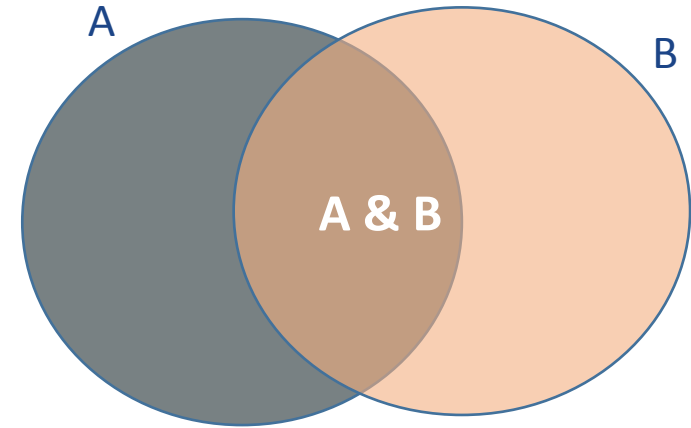
These events are disjoint
so $P(\text{King and } 3) = 0$



Summary

General addition rule

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$



Note: For disjoint events $P(A \text{ and } B) = 0$, so the above formula simplifies to

$$P(A \text{ or } B) = P(A) + P(B)$$

Question

What is the probability that a randomly sampled student thinks marijuana should be legalized or they agree with their parents' political views?

<i>Legalize MJ</i>	<i>Share Parents' Politics</i>		<i>Total</i>
	<i>No</i>	<i>Yes</i>	
No	11	40	51
Yes	36	78	114
Total	47	118	165

- (a) $(40 + 36 - 78) / 165$
- (b) $(114 + 118 - 78) / 165$
- (c) $78 / 165$
- (d) $78 / 188$
- (e) $11 / 47$

Conditional Probability

Sampling with replacement

When sampling with replacement, you put back what you just drew.




- Imagine you have a bag with 5 red, 3 blue and 2 orange chips in it. What is the probability that the first chip you draw is blue?

5 , 3 , 2 

$$Prob(1^{st} \text{ chip } B) = \frac{3}{5 + 3 + 2} = \frac{3}{10} = 0.3$$

- Suppose you did indeed pull a blue chip in the first draw. If drawing with replacement, what is the probability of drawing a blue chip in the second draw?

1st Draw - 5 , 3 , 2 

2nd Draw - 5 , 3 , 2 

$$P(2^{nd} \text{ Chip } B \mid 1^{st} \text{ Chip } B) = 3/10 = 0.3$$

$$P(A | B)$$

Conditional Probability Notation

- When we are examining the probabilities of one event assuming that another event has happened we call it conditional probability.

What's the probability of Event B, given that we know A has happened?

$$P(B | A)$$




$$P(A | B)$$

Assuming Event B, what's the probability of Event A?

Sampling with replacement (cont.)

- Suppose you actually pulled an orange chip in the first draw. If drawing with replacement, what is the probability of drawing a blue chip in the second draw?




1st Draw - 5 , 3 , 2 

2nd Draw - 5 , 3 , 2 

$$P(\text{2nd chip B} \mid \text{1st chip O}) = 3 / 10 = 0.3$$

- If drawing with replacement, what is the probability of drawing two blue chips in a row?

1st Draw - 5 , 3 , 2 

2nd Draw - 5 , 3 , 2 

$$P(\text{1st chip B}) \times P(\text{2nd chip B} \mid \text{1st chip B}) = 0.3 \times 0.3 = 0.09$$

Sampling with replacement (cont.)

When drawing with replacement, probability of the second chip being blue does not depend on the color of the first chip since whatever we draw in the first draw gets put back in the bag.

$$\text{Prob}(B \mid B) = \text{Prob}(B \mid O)$$

In addition, this probability is equal to the probability of drawing a blue chip in the first draw, since the composition of the bag never changes when sampling with replacement.

$$\text{Prob}(B \mid B) = \text{Prob}(B)$$

When drawing with replacement, draws are independent.

Independence

Two processes are independent if knowing the outcome of one **provides no useful information** about the outcome of the other.

- Knowing that the coin landed on a head on the first toss does not provide any useful information for determining what the coin will land on in the second toss.
>> Outcomes of two tosses of a coin are independent.
- Knowing that the first card drawn from a deck is an ace does provide useful information for determining the probability of drawing an ace in the second draw.
>> Outcomes of two draws from a deck of cards (without replacement) are dependent.

Independence

Two events are independent if knowing the outcome of one event **does not change the probability** of the other event happening.




- Knowing that the coin landed on a head on the first toss does not change the probabilities for the second toss.
>> Outcomes of two tosses of a coin are independent.
- Knowing that the first card drawn from a deck is an ace does change the probability for the second draw.
>> Outcomes of two draws from a deck of cards (without replacement) are dependent.

Sampling without replacement

When drawing without replacement you do not put back what you just drew.

- Suppose you pulled a blue chip in the first draw. If drawing without replacement, what is the probability of drawing a blue chip in the second draw?




1st Draw - 5 , 3 , 2 

2nd Draw - 5 , 2 , 2 

$$P(\text{2nd chip B} \mid \text{1st chip B}) = 2 / 9 = 0.22$$

- If drawing without replacement, what is the probability of drawing two blue chips in a row?

1st Draw - 5 , 3 , 2 

2nd Draw - 5 , 2 , 2 

$$P(\text{1st chip B}) \times P(\text{2nd chip B} \mid \text{1st chip B}) = 0.3 \times 0.22 = 0.066$$

Sampling without replacement (cont.)

When drawing without replacement, the probability of the second chip being blue given the first was blue is not equal to the probability of drawing a blue chip in the first draw since the composition of the bag changes with the outcome of the first draw.

$$\text{Prob}(B | B) \neq \text{Prob}(B)$$

When drawing without replacement, draws are not independent.

This is especially important to take note of when the sample sizes are small. If we were dealing with, say, 10,000 chips in a (giant) bag, taking out one chip of any color would not have as big an impact on the probabilities in the second draw.

Question

In most card games cards are dealt without replacement. What is the probability of being dealt an ace and then a 3?

Checking for independence

If $P(A | B) = P(A)$,
then A and B are independent.

Event B does not change the probability of Event A

Question

- In the UK in 90s, doctors estimated the chances of a child dying from SIDS at 1 in 8500.
- Sally Clark had two children die from sids several years apart when both were about the same age.
- Authorities charged her with murder, and at her trial Dr. Roy Meadow testified that the probability of both children dying from sids was
$$\frac{1}{8500}^2 = \frac{1}{73 \text{ mil}}$$
- She was convicted based mostly on this calculation.
- What's wrong with this story?

Sally Clark

- The cause of SIDS was (and still is largely) unknown, so assuming independence was not appropriate.
- In fact, given that unidentified genetic and environmental factors likely contribute to SIDS, it is a better assumption to assume that given one child who dies of SIDS, the second child is more likely to be at risk of the same thing.

There is an additional problem with this story--but we'll save that for another time. Also, this story does not have a happy ending.

Testing for Independence

If $P(A|B) = P(A)$ then A and B are independent.

However, in real life our knowledge of the probabilities is never perfect.

$$P(A|B) \approx P(A)$$

What if $P(A|B) \approx P(A)$?

We need to decide, is it close enough?

Determining dependence based on sample data

- If conditional probabilities calculated based on sample data suggest dependence between two variables, the next step is to conduct a hypothesis test to determine if the observed difference between the probabilities is likely or unlikely to have happened by chance.
- If the observed difference between the conditional probabilities is large, then there is stronger evidence that the difference is real.
- If a sample is large, then even a small difference can provide strong evidence of a real difference.

Combining Events:

The Fundamental Counting Principle

- A pizza parlor has two sizes of pizza, and offers three toppings. How many 1 topping pizzas are there?

2 sizes x 3 toppings = 6 one-topping pizzas.

The fundamental counting principle states that if Event A has N outcomes and Event B has M outcomes, and A and B are independent, then the combined event AB or A and B has $N \times M$ outcomes.

Product rule for independent events

If A and B are independent, $P(A \text{ and } B) = P(A) \times P(B)$

Or more generally, $P(A_1 \text{ and } \dots \text{ and } A_k) = P(A_1) \times \dots \times P(A_k)$

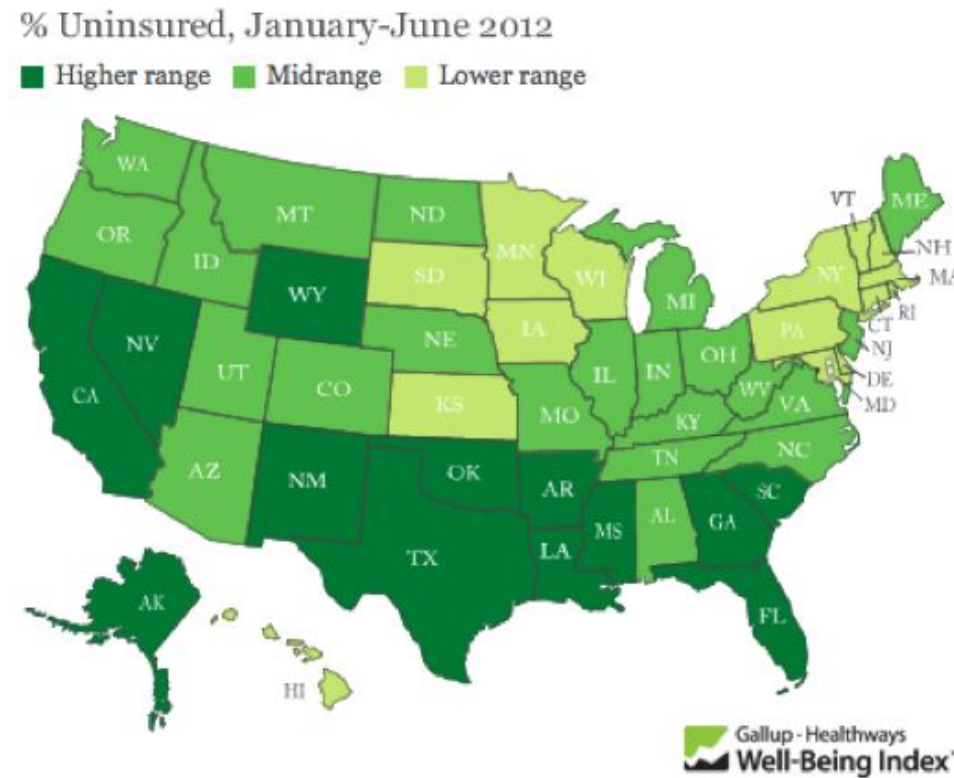
You toss a coin twice, what is the probability of getting two tails in a row?

$$\begin{aligned} &P(\text{T on the first toss}) \times P(\text{T on the second toss}) \\ &= (1 / 2) \times (1 / 2) = 1 / 4 \end{aligned}$$

Question

A Gallup poll suggests that 25.5% of Texans do not have health insurance as of June 2012. Assuming that the uninsured rate stayed constant, what is the probability that two randomly selected Texans are both *insured*?

- (a) 25.5^2
- (b) 0.255^2
- (c) 0.255×2
- (d) $(1 - 0.255)^2$



Putting everything together...

If we were to randomly select 5 Texans, what is the probability that at least one is uninsured?

- If we were to randomly select 5 Texans, the sample space for the number of Texans who are uninsured would be:

$$S = \{0, 1, 2, 3, 4, 5\}$$

- We are interested in instances where at least one person is uninsured:

$$S = \{0, 1, 2, 3, 4, 5\}$$

- So we can divide up the sample space into two categories:

$$S = \{0, \text{at least one}\}$$

Putting everything together...

Since the probability of the sample space must add up to 1:

$$\begin{aligned} &P(\text{at least 1 uninsured}) \\ &= 1 - P(\text{none uninsured}) \\ &= 1 - (1 - 0.255)^5 \\ &= 1 - 0.745^5 \\ &= 1 - 0.23 \\ &= 0.77 \end{aligned}$$

At least 1:

$$P(\text{at least one}) = 1 - P(\text{none})$$

Question

Roughly 20% of undergraduates at a university are vegetarian. What is the probability that, among a random sample of 3 undergraduates, at least one is vegetarian?

(a) $1 - 0.2 \times 3$

(b) $1 - 0.2^3$

(c) 0.8^3

(d) $1 - 0.8 \times 3$

(e) $1 - 0.8^3$

Bayes Theorem

Relating Conditional Probabilities

- Bayes Theorem

$$P(A|B) = \frac{P(A \& B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

Relating Conditional Probabilities

- Bayes Theorem

$$P(A|B) = \frac{P(A \& B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

Example:

- If there fire, then there will be smoke: $P(\text{smoke} | \text{fire}) = 0.9$
- What's the probability that if there is smoke, there is fire? $P(\text{fire} | \text{smoke})$?
 - We need to know the $P(\text{fire}) = 1\%$ (let's say bad fires are rare)
 - And $P(\text{smoke}) = 10\%$ (don't forget about bbqs and fire pits)

$$P(\text{fire}|\text{smoke}) = \frac{P(\text{smoke}|\text{fire}) \cdot P(\text{fire})}{P(\text{smoke})} = \frac{0.9 \cdot 0.01}{0.1} = 0.09$$

General Multiplication Rule

- If A and B are independent

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

- Bayes Theorem

$$P(A|B) = \frac{P(A \& B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

- General Multiplication Rule (Works whether A and B are independent or not)

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

Recap: Disjoint vs. Independent

- Two events that are disjoint (mutually exclusive) if they cannot happen at the same time.
 - $P(A \text{ and } B) = 0$
- Two events are independent if knowing the outcome of one provides no useful information about the outcome of the other.
 - $P(A|B) = P(A)$

Applying Conditional Probability

Relapse

Researchers randomly assigned 72 chronic users of cocaine into three groups: desipramine (antidepressant), lithium (standard treatment for cocaine) and placebo. Results of the study are summarized below.

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

[http://www.oswego.edu/~srp/stats/2 way tbl 1.htm](http://www.oswego.edu/~srp/stats/2_way_tbl_1.htm)

Marginal probability

What is the probability that a patient relapsed?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

Marginal probability

What is the probability that a patient relapsed?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

$$P(\text{relapsed}) = 48 / 72 \sim 0.67$$

Joint probability

What is the probability that a patient received the antidepressant (desipramine) and relapsed?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

Joint probability

What is the probability that a patient received the antidepressant (desipramine) and relapsed?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

$$P(\text{relapsed and desipramine}) = 10 / 72 \sim 0.14$$

Bayes Theorem

Bayes Theorem gives a way to relate conditional probabilities to joint probabilities:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

$$= \frac{10/72}{24/72}$$

$$= \frac{10}{24}$$

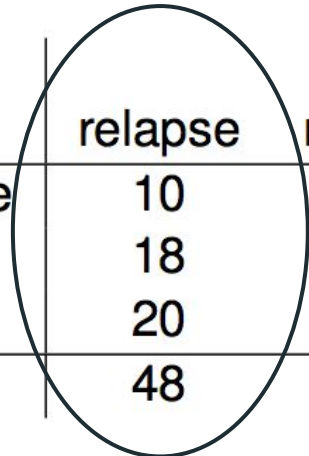
$$= 0.42$$

$$= \frac{P(\text{relapse}|\text{desipramine})}{P(\text{desipramine})}$$

Conditional probability

If we know that a patient relapsed, what is the probability that they received the antidepressant (desipramine)?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72



$$P(\text{desipramine} \mid \text{relapse}) = 10 / 48 \sim 0.21$$

$$P(\text{lithium} \mid \text{relapse}) = 18 / 48 \sim 0.38$$

$$P(\text{placebo} \mid \text{relapse}) = 20 / 48 \sim 0.42$$

Conditional probability

If we know that a patient received the antidepressant (desipramine), what is the probability that they relapsed?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

$$P(\text{relapse} \mid \text{desipramine}) = 10 / 24 \sim 0.42$$

$$P(\text{relapse} \mid \text{lithium}) = 18 / 24 \sim 0.75$$

$$P(\text{relapse} \mid \text{placebo}) = 20 / 24 \sim 0.83$$

Example

Consider the following (hypothetical) distribution of gender and major of students in an introductory statistics class:

	social science	non-social science	total
female	30	20	50
male	30	20	50
total	60	40	100

- The probability that a randomly selected student is a social science major is $P(SS) = 60/100 = 0.6$.
- The probability that a randomly selected student is a social science major given that they are female is $P(SS|F) = 30/50 = 0.6$
- Similarly, $P(SS|M) = 30/50 = 0.6$

Illustration of Conditional Probability

- <http://setosa.io/conditional/>

Quick Recap on Conditional Probabilities

- Use conditional probability when two events are dependent

- Bayes Theorem $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

$$P(\text{desipramine} \mid \text{relapse}) = 10 / 48 \sim 0.21$$

$$P(\text{lithium} \mid \text{relapse}) = 18 / 48 \sim 0.38$$

$$P(\text{placebo} \mid \text{relapse}) = 20 / 48 \sim 0.42$$

Probability Trees

Inverting Probabilities, i.e. $P(A|B) \rightarrow P(B|A)$

Example

Suppose 13% of students earned an A on the midterm. Of those students who earned an A on the midterm, 47% received an A on the final, and 11% of the students who earned lower than an A on the midterm received an A on the final. You randomly pick up a final exam and notice the student received an A. What is the probability that this student earned an A on the midterm?

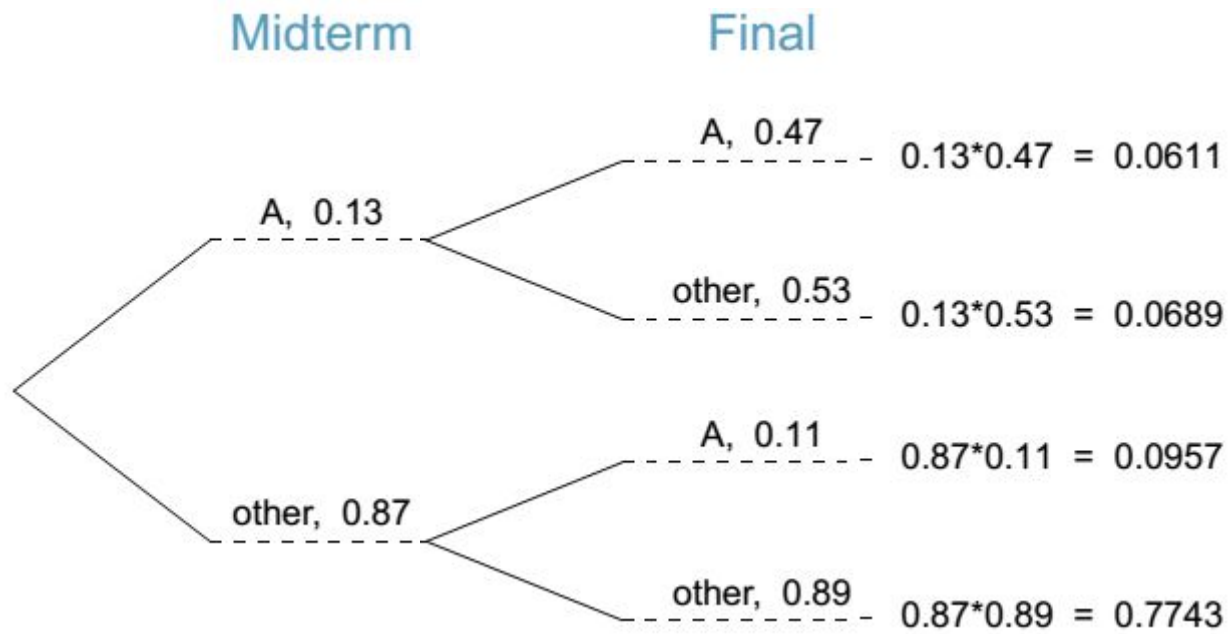
$$P(\text{midterm}=\text{A}) = 0.13$$

$$P(\text{final}=\text{A} \mid \text{midterm}=\text{A}) = 0.47$$

$$P(\text{final}=\text{A} \mid \text{midterm}=\text{other}) = 0.11$$

$$P(\text{midterm}=\text{A} \mid \text{final}=\text{A}) = ?$$

We need to calculate $P(\text{midterm} = \text{A and final} = \text{A})$ and $P(\text{final} = \text{A})$



$$P(\text{midterm} = \text{A and final} = \text{A}) = 0.0611$$

$$P(\underline{\text{final}} = \text{A}) = P(\text{midterm} = \text{other and } \underline{\text{final}} = \text{A}) + P(\text{midterm} = \text{A and } \underline{\text{final}} = \text{A}) = 0.0957 + 0.0611 = 0.1568$$

$$\begin{aligned}
 P(\text{midterm} = \text{A} | \text{final} = \text{A}) &= \frac{P(\text{midterm} = \text{A and final} = \text{A})}{P(\text{final} = \text{A})} \\
 \text{(posterior probability)} & \\
 &= \frac{0.0611}{0.1568} = 0.3897
 \end{aligned}$$

Probability Trees

Inverting Probabilities, i.e. $P(A|B) \rightarrow P(B|A)$

Example

Suppose 13% of students earned an A on the midterm. Of those students who earned an A on the midterm, 47% received an A on the final, and 11% of the students who earned lower than an A on the midterm received an A on the final. You randomly pick up a final exam and notice the student received an A. What is the probability that this student earned an A on the midterm?

$$P(\text{midterm}=\text{A}) = 0.13$$

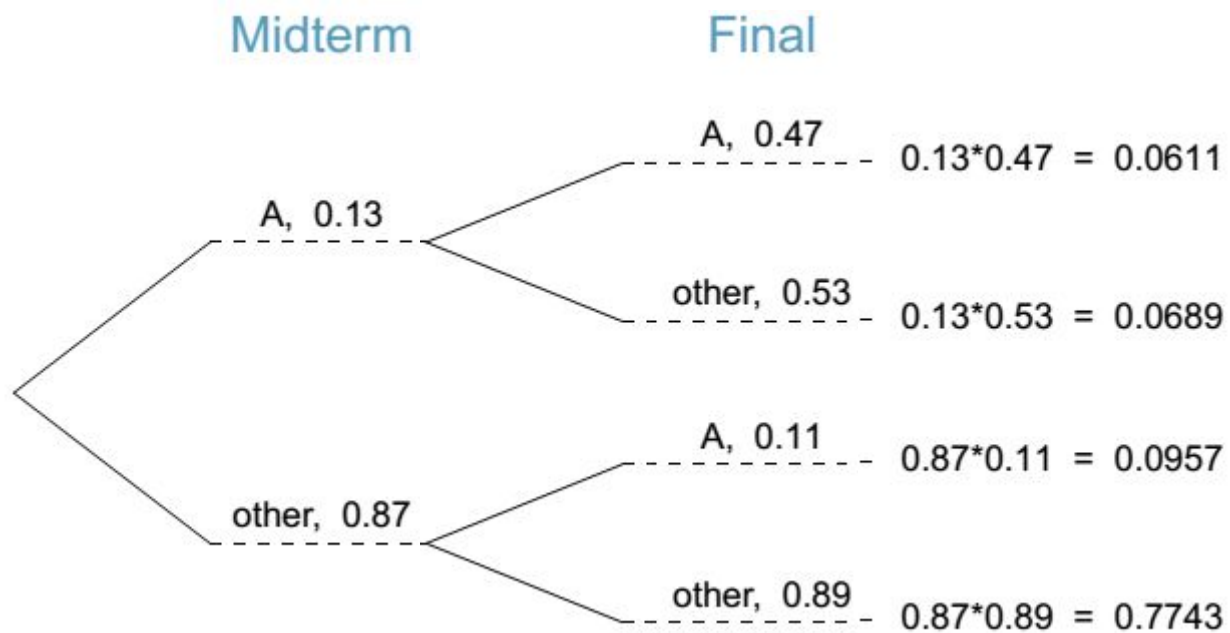
$$P(\text{final}=\text{A} \mid \text{midterm}=\text{A}) = 0.47$$

$$P(\text{final}=\text{A} \mid \text{midterm}=\text{other}) = 0.11$$

$$P(A|B) = P(A \text{ and } B) / P(B) = P(B|A)P(A)/P(B)$$

$$P(\text{midterm}=\text{A} \mid \text{final}=\text{A}) = ?$$

We need to calculate $P(\text{midterm} = \text{A and final} = \text{A})$ and $P(\text{final} = \text{A})$



$$P(\text{midterm} = A \text{ and } \text{final} = A) = 0.0611$$

$$P(\underline{\text{final}} = A) = P(\text{midterm} = \text{other} \text{ and } \underline{\text{final}} = A) + P(\text{midterm} = A \text{ and } \underline{\text{final}} = A) = 0.0957 + 0.0611 = 0.1568$$

$$\begin{aligned}
 P(\text{midterm} = A | \text{final} = A) &= \frac{P(\text{midterm} = A \text{ and } \text{final} = A)}{P(\text{final} = A)} \\
 \text{(posterior probability)} & \\
 &= \frac{0.0611}{0.1568} = 0.3897
 \end{aligned}$$