SEIS 631
Foundations of Data Analysis

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…

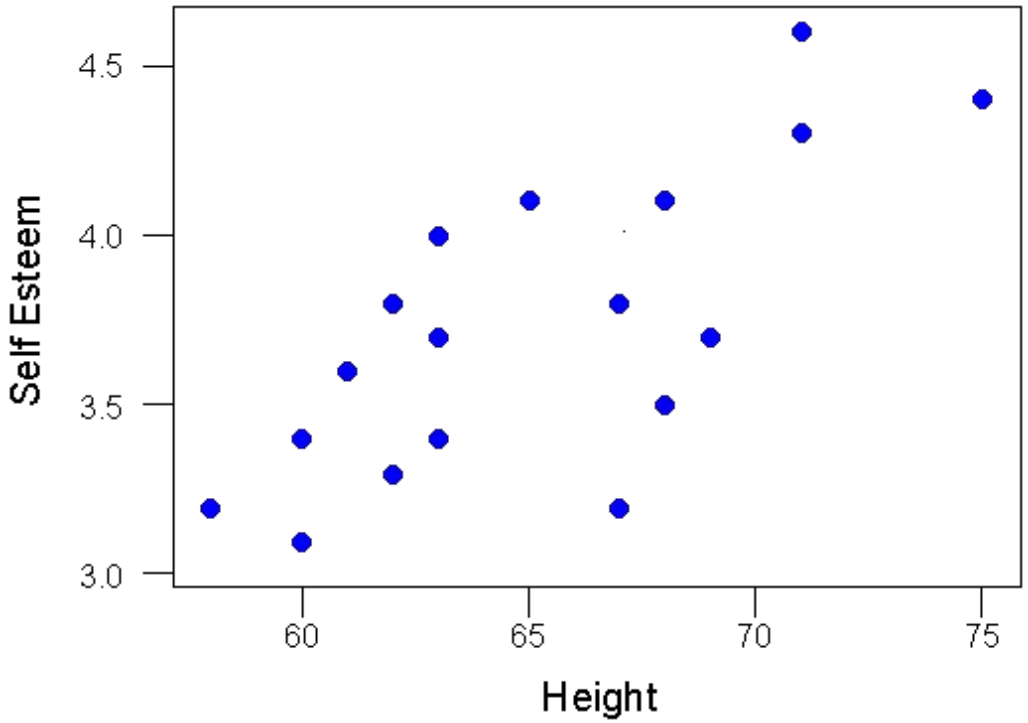
Associated vs. independent

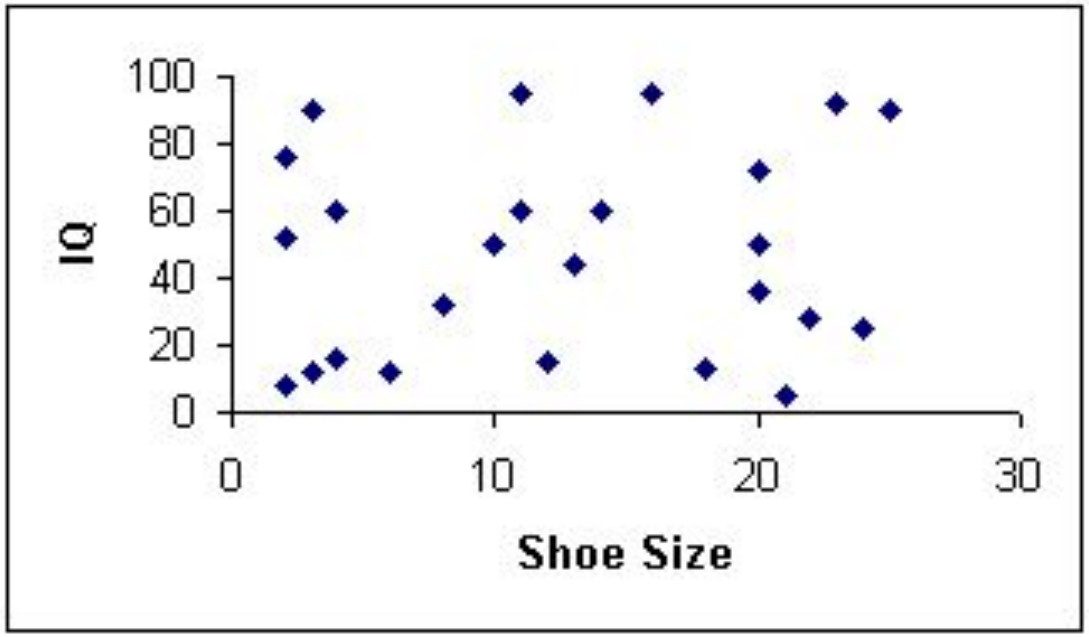
- When two variables show some connection with one another, they are called associated variables.
  - Associated variables can also be called dependent variables and vice-versa.
- If two variables are not associated, i.e. there is no evident connection between the two, then they are said to be independent.

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





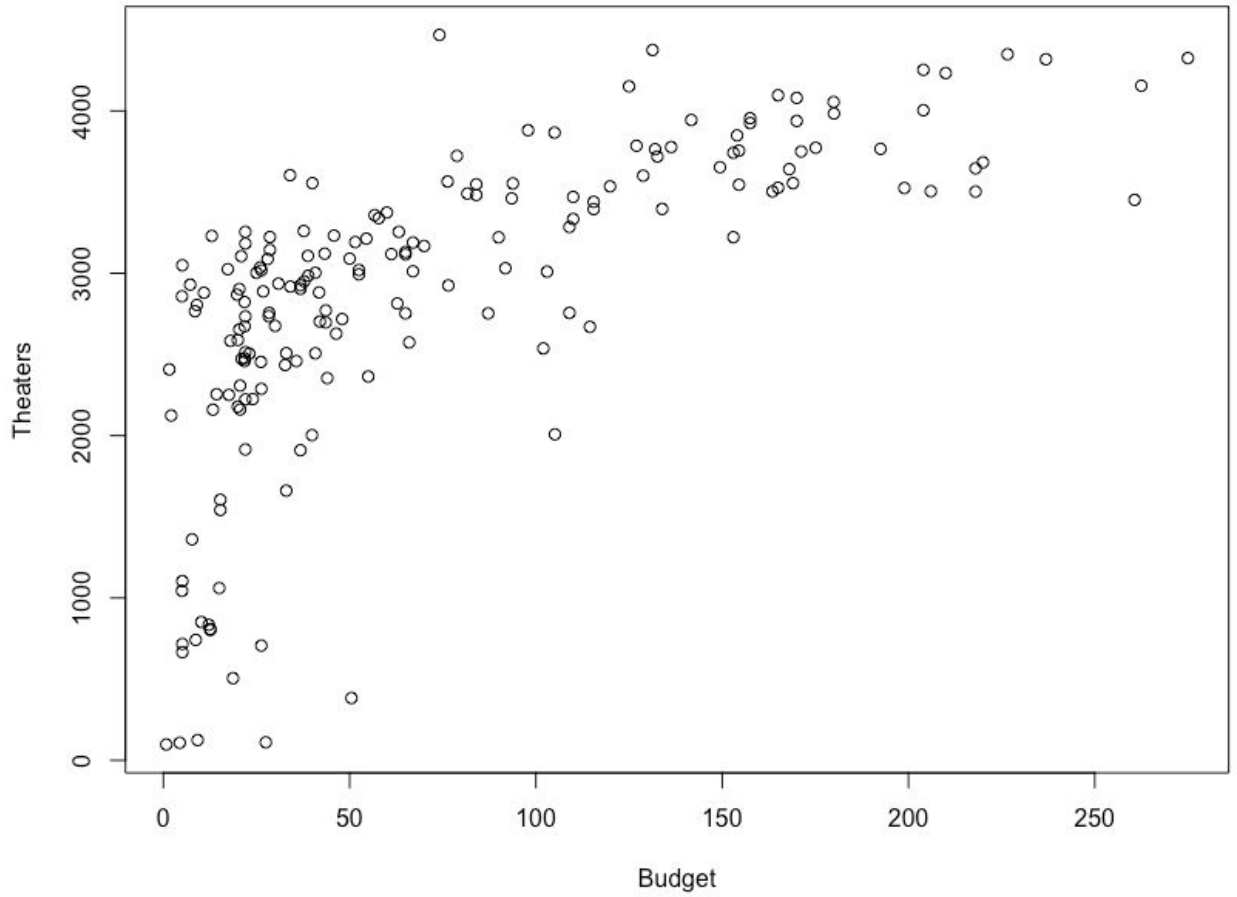
Data Analysis Step 1: Explore & Describe
Two Variables: Both Numeric

- Scatterplot

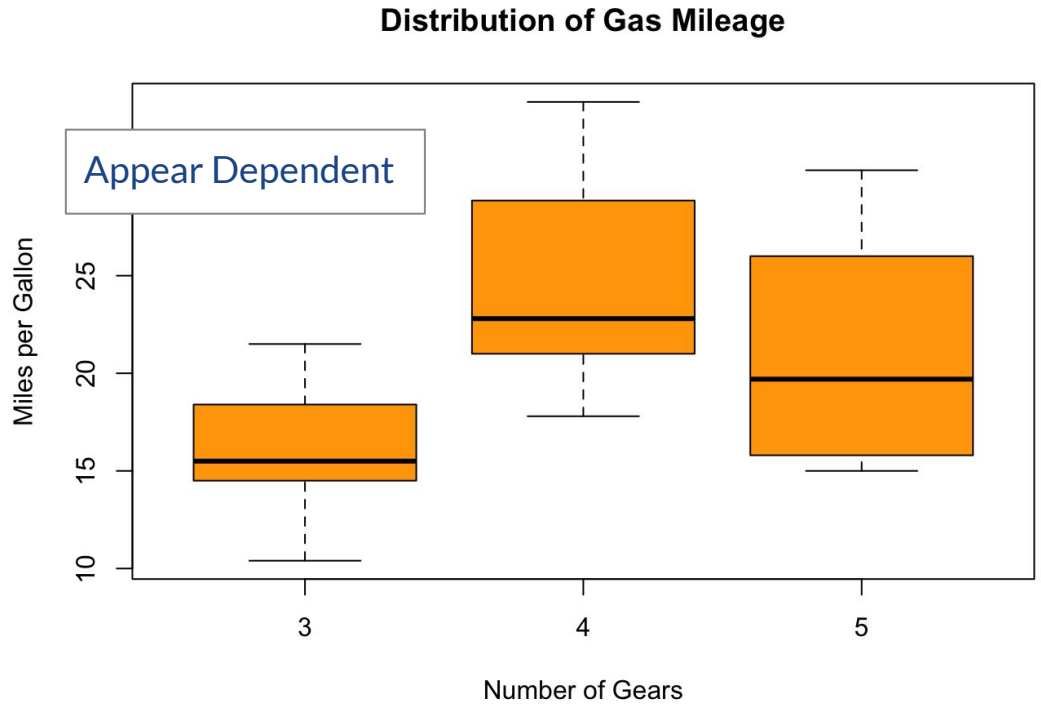
Are they independent?

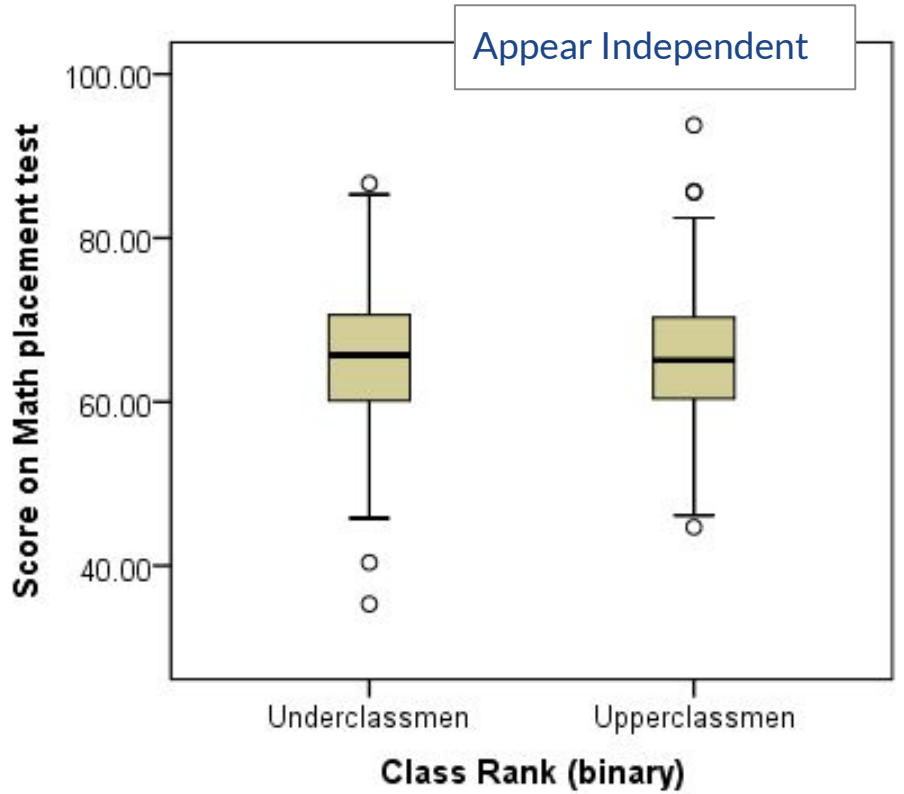
How do we describe the relationship?

- Fairly strong
- Positive
- Non-linear

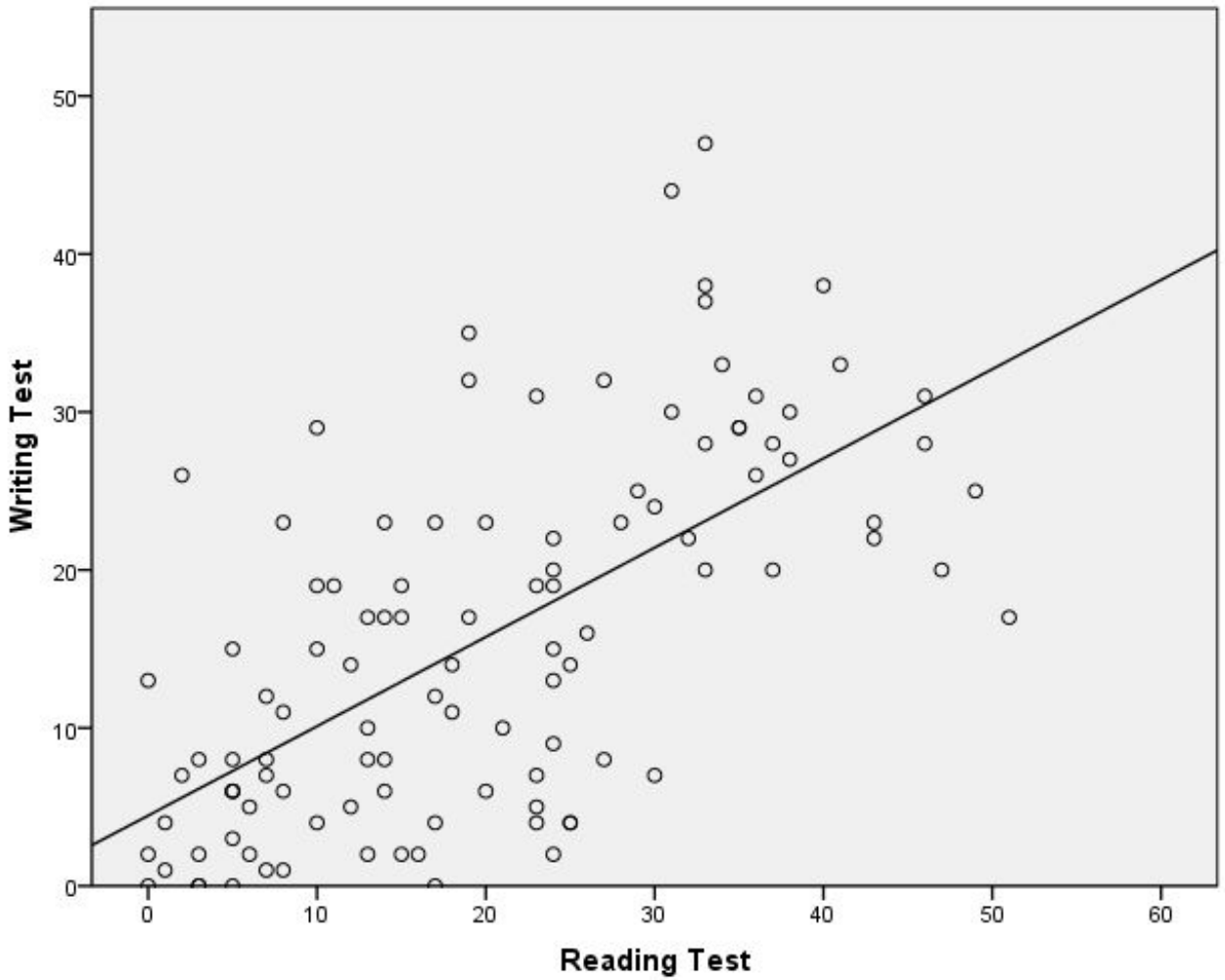


Relationships between Variables Example
Numeric vs. Categorical





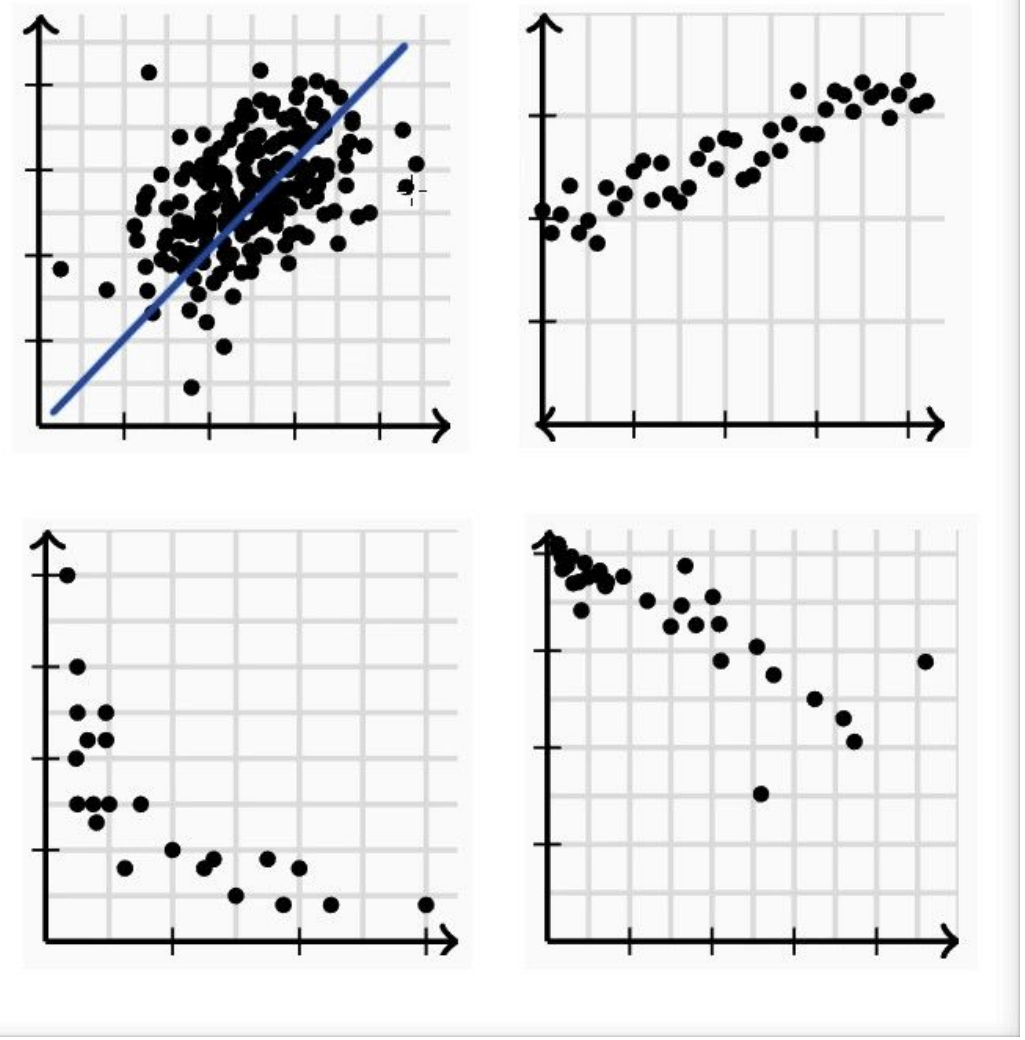
Relationships between Variables Example
Numeric vs. Numeric



Linear
Weak
Positive
Few extreme values

Relationships between Variables More Examples: Numeric vs. Numeric

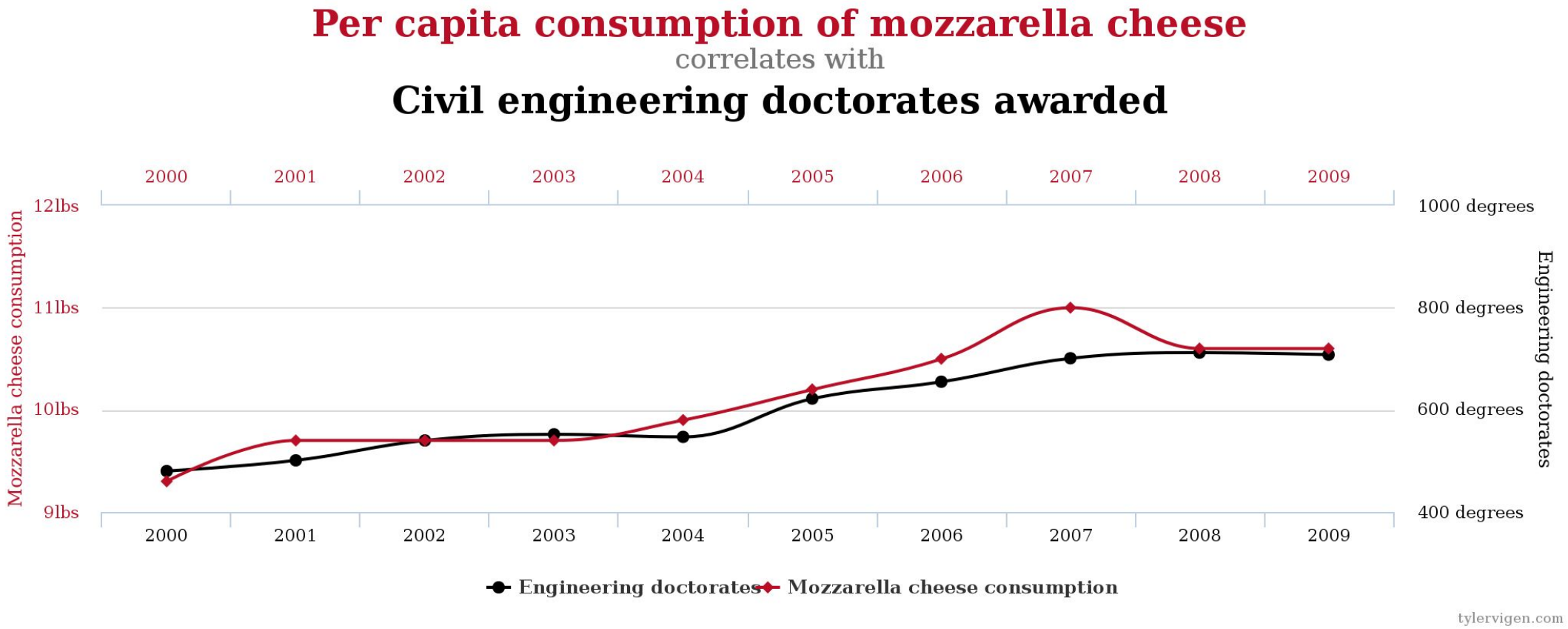
- Which ones are strong? Weak?
- Which ones are positive? negative?
- Which ones are linear? Curved?
- Which ones have outliers?



Identifying Relationships

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

Correlation ≠ Causation



- Source: http://www.tylervigen.com/spurious-correlations

Descriptive Statistics Tell a Story

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

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.

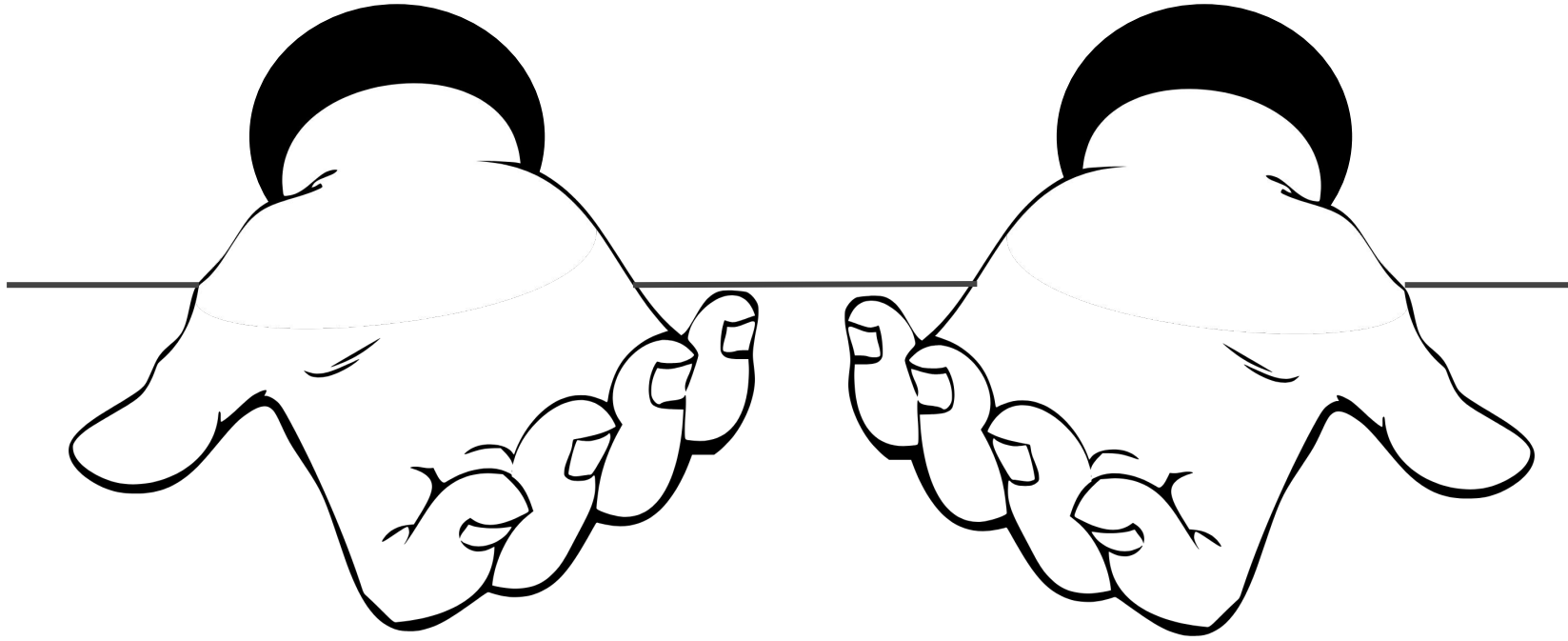
Can psychics sense your aura?

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

Can we put this claim to the test?

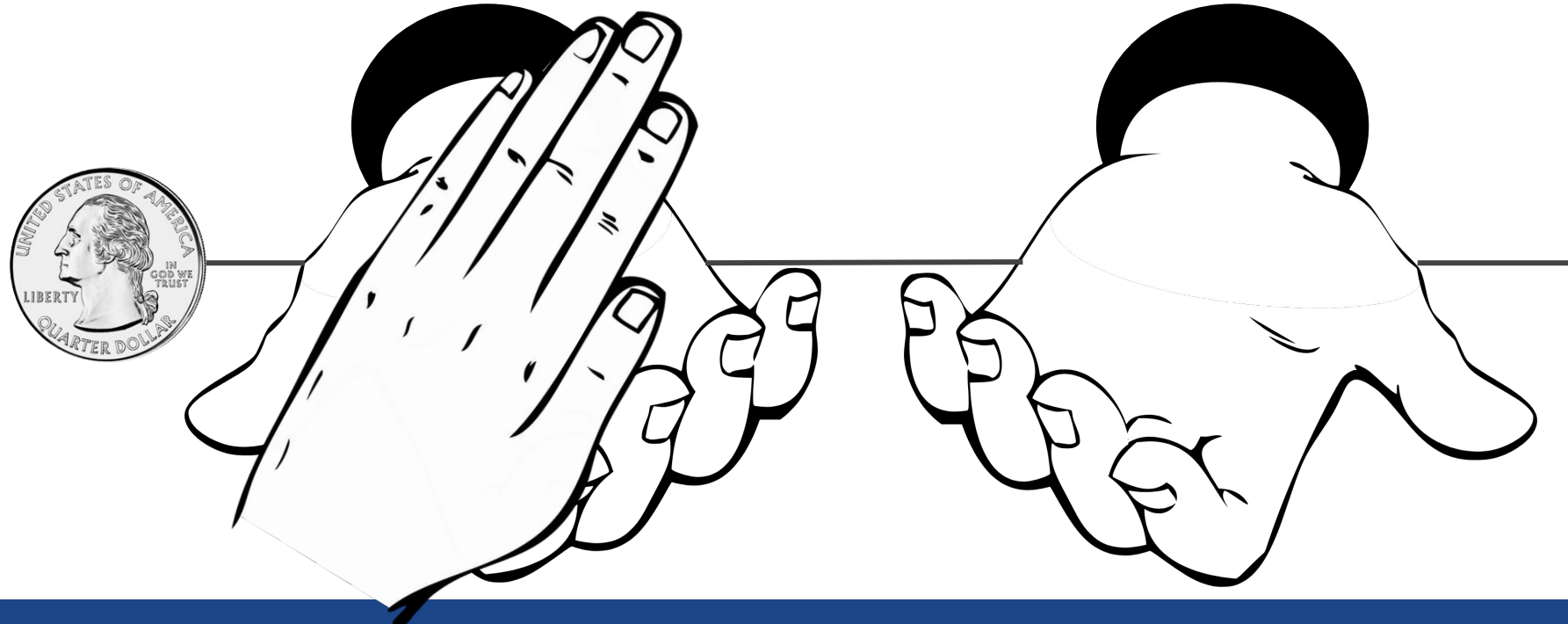
The Experiment

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



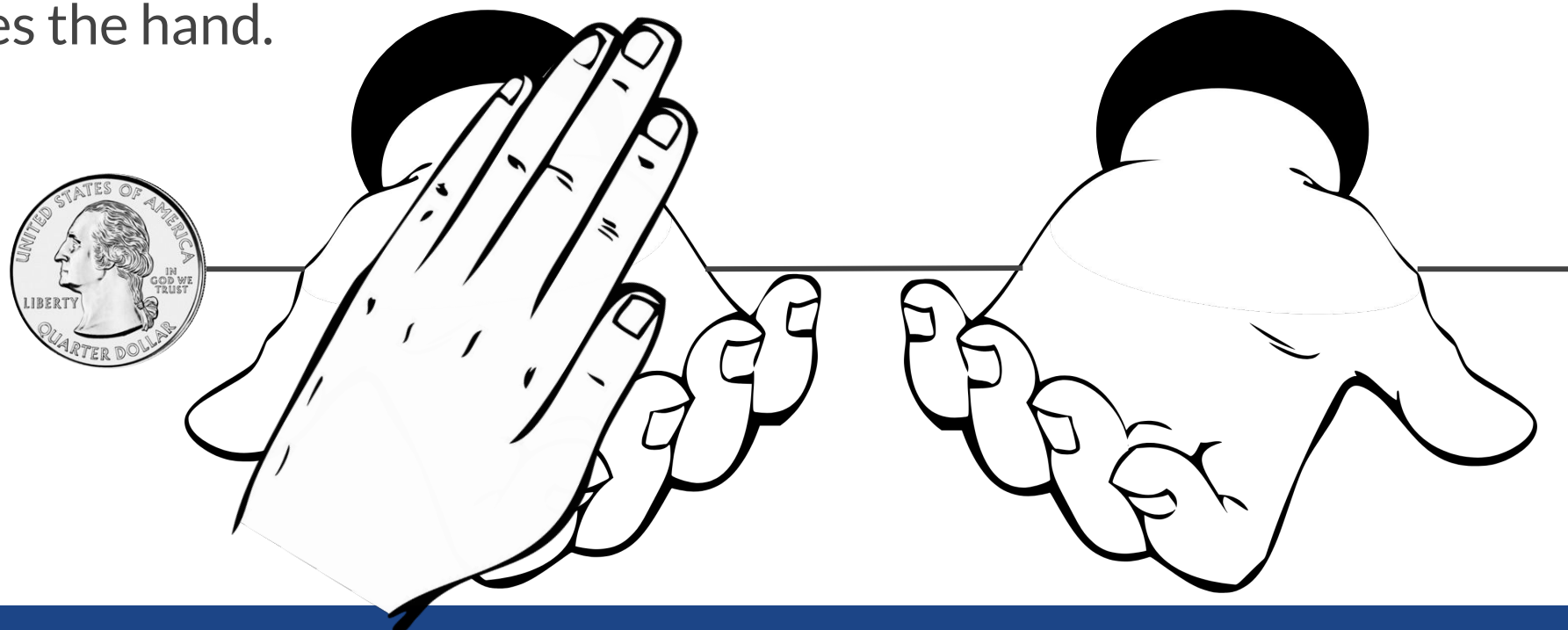
The Experiment

- 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.
    5. Do the patterns match your assumption of randomness?

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?

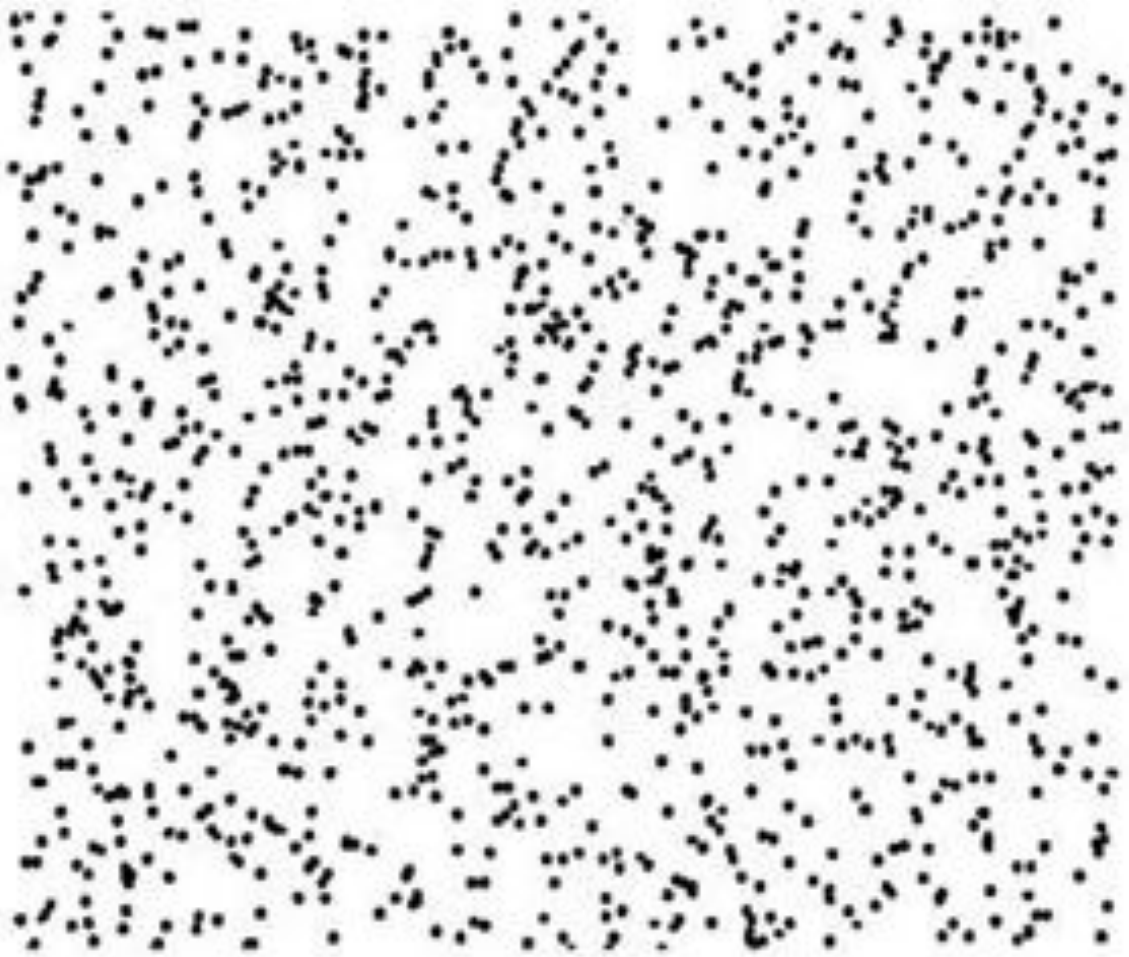
Let's test my claim with a simulation.

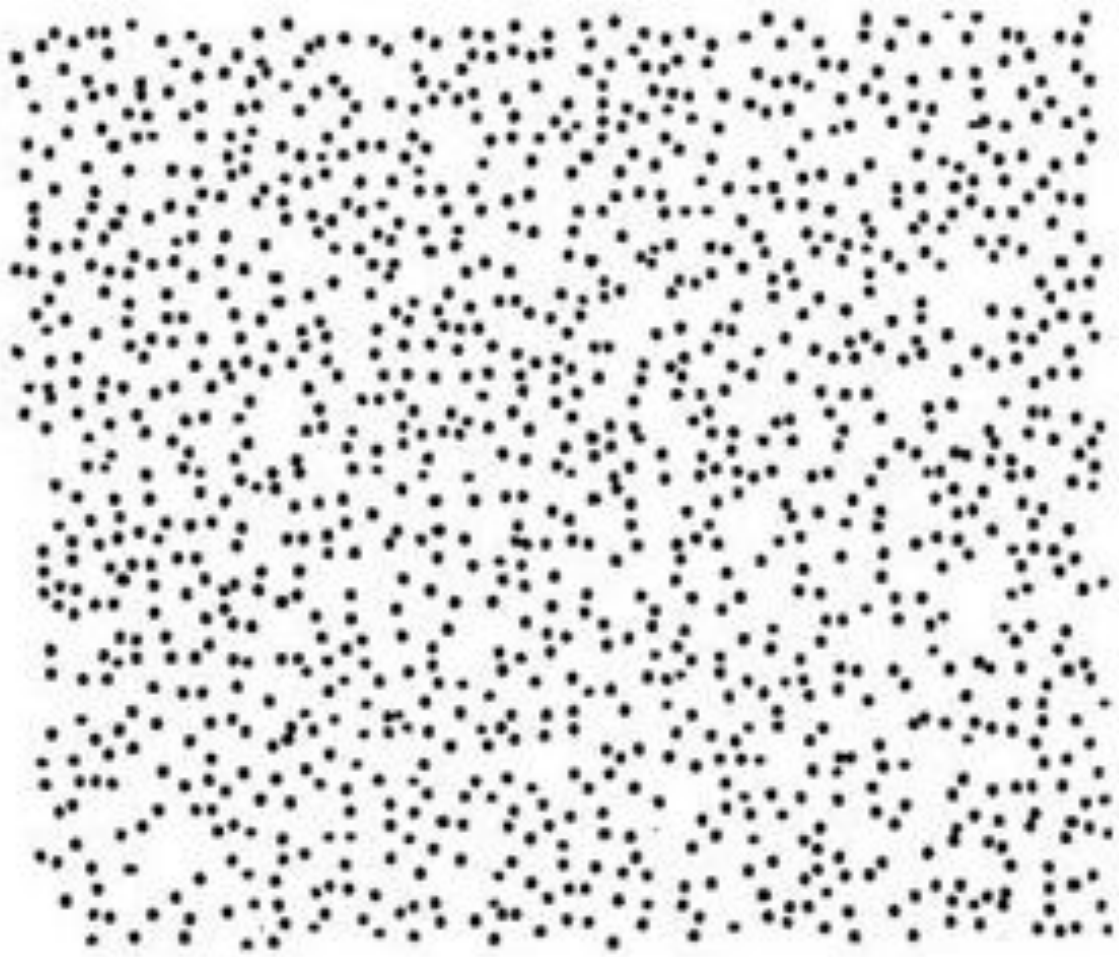
https://justflipacoin.com/

Simulations in R

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
  - Streaming Music 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, Steph Curry has made 43.3% of his three pointers. What's the probability that he makes his next three pointer?*
  - Answer: 0.433
  - **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 Tomorrow?*
  - Answer: 85% (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:

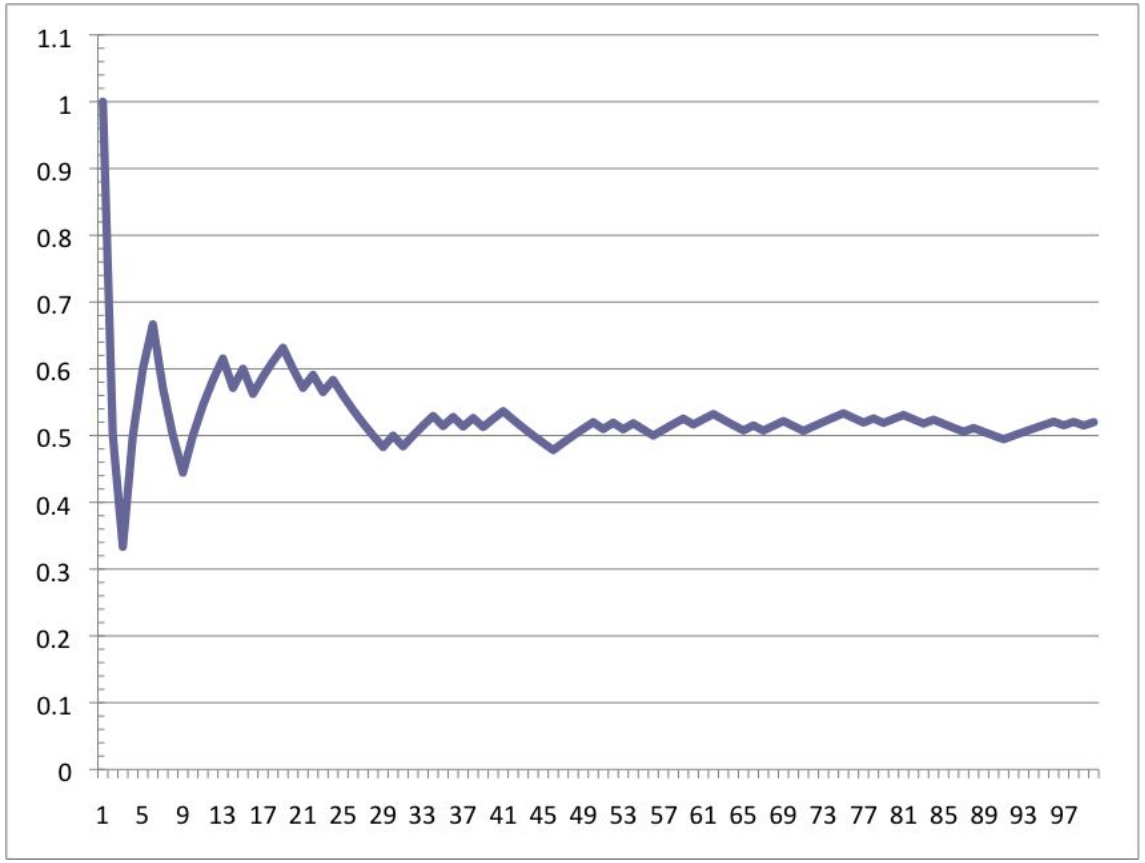
- A Bayesian interprets probability as a subjective degree of belief: For the same event, two separate people could have different viewpoints and so assign different probabilities.

- Largely popularized by revolutionary advance in computational technology and methods during the last twenty years.

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?

Source: ***How to Not Be Wrong*** – Jordan Elllenberg

Law of large numbers and Cancer Rates

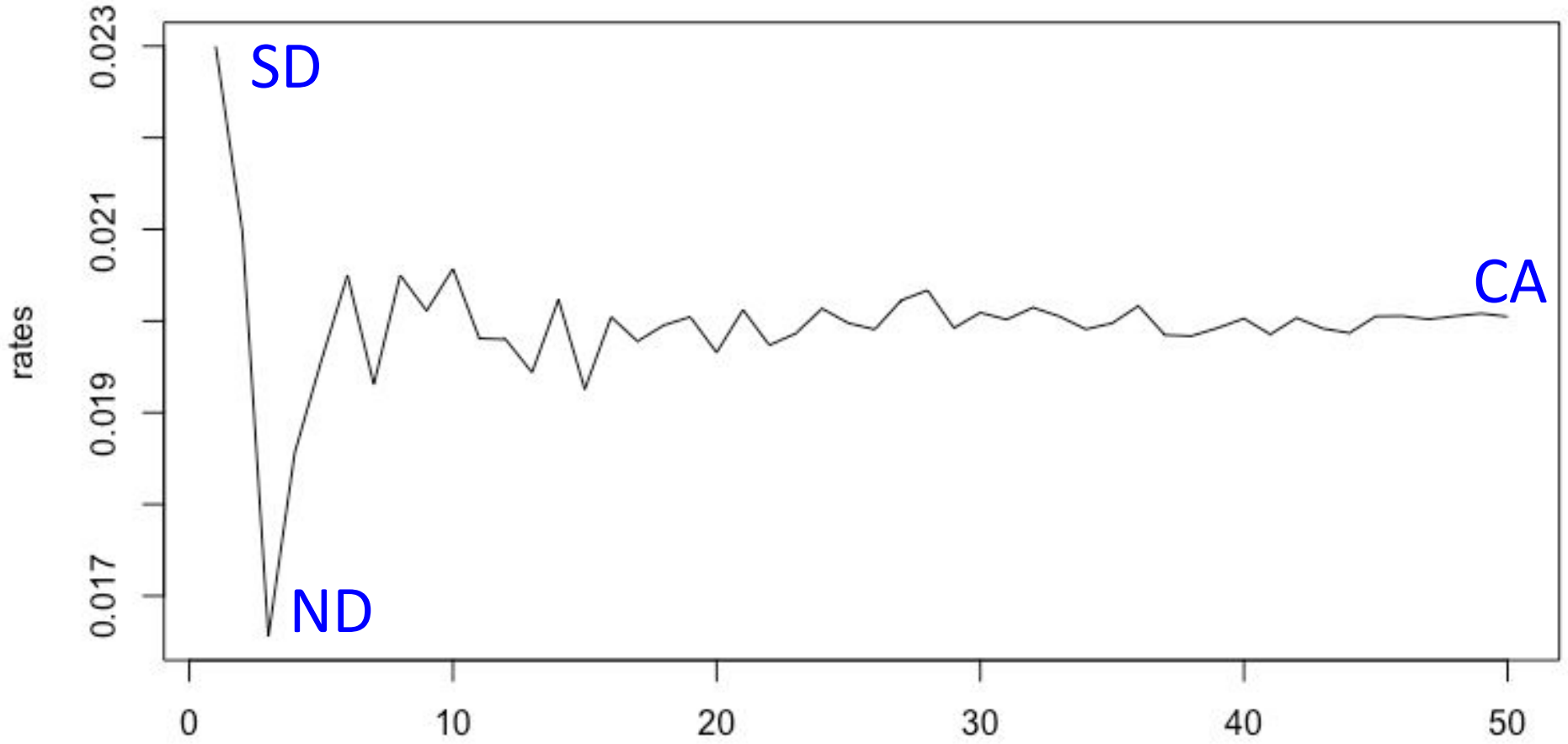
Highest Brain Cancer Rates	Lowest Brain Cancer Rates
South Dakota	Wyoming
Nebraska	Vermont
Alaska	North Dakota
Delaware	Hawaii
Maine	DC

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(Rolling a 4) = 1 good thing (the 4) / 6 total things = ⅙

Notation: P(EVENT) is how we write probabilities and you should read this as "The probability of this event" or "The probability that EVENT happens"

Calculating Probabilities

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

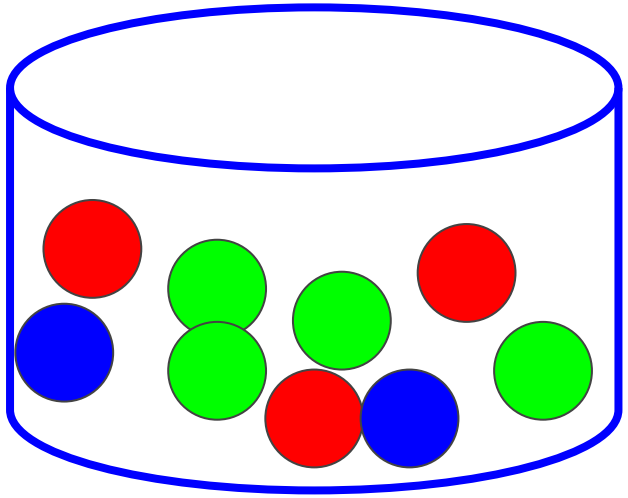
P(Vowel) = 5 vowels (a, e, i, o, u) / 26 letters = 5/26

Calculating Probabilities

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

P(Vowel) = 5 vowels (a, e, i, o, u) / 26 letters = 5/26

- I randomly choose a marble from the jar.

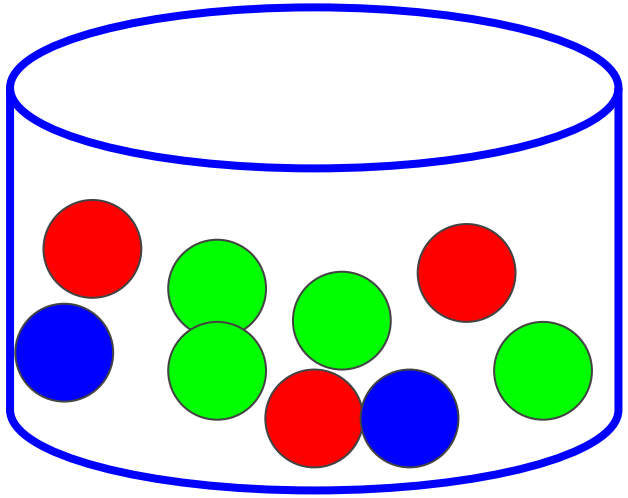


Calculating Probabilities

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

P(Vowel) = 5 vowels (a, e, i, o, u) / 26 letters = 5/26

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

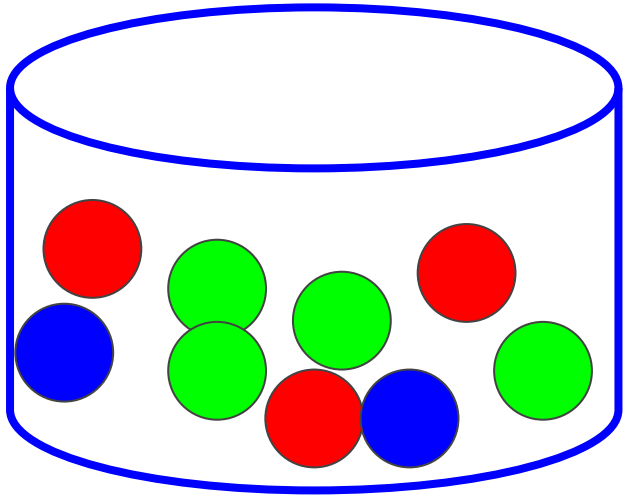


Calculating Probabilities

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

P(Vowel) = 5 vowels (a, e, i, o, u) / 26 letters = 5/26

- I randomly choose a marble from the jar.
  - What is the probability I choose red?
    - P(Red) = 3 red / 9 marbles = 3/9 = ⅓

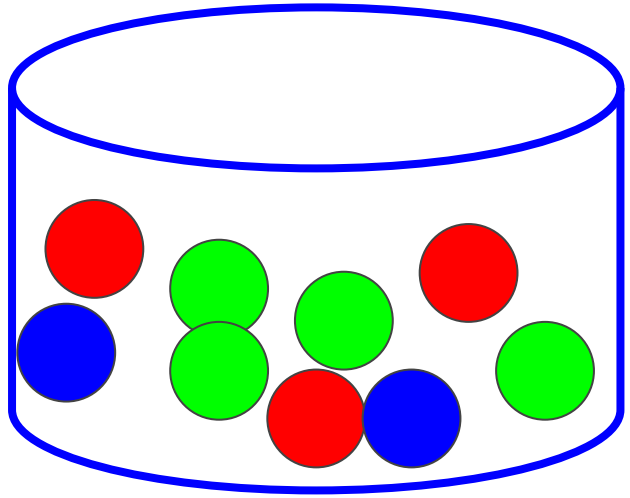


Calculating Probabilities

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

P(Vowel) = 5 vowels (a, e, i, o, u) / 26 letters = 5/26

- I randomly choose a marble from the jar.
  - What is the probability I choose red?
    - P(Red) = 3 red / 9 marbles = 3/9 = ⅓
  - What is the probability I don't choose blue?

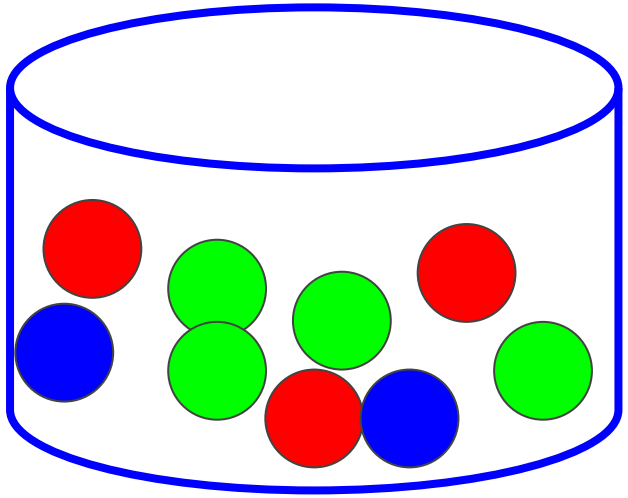


Calculating Probabilities

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

P(Vowel) = 5 vowels (a, e, i, o, u) / 26 letters = 5/26

- I randomly choose a marble from the jar.
  - What is the probability I choose red?
    - P(Red) = 3 red / 9 marbles = 3/9 = ⅓
  - What is the probability I don't choose blue?
    - P(Not Blue) = 7/9

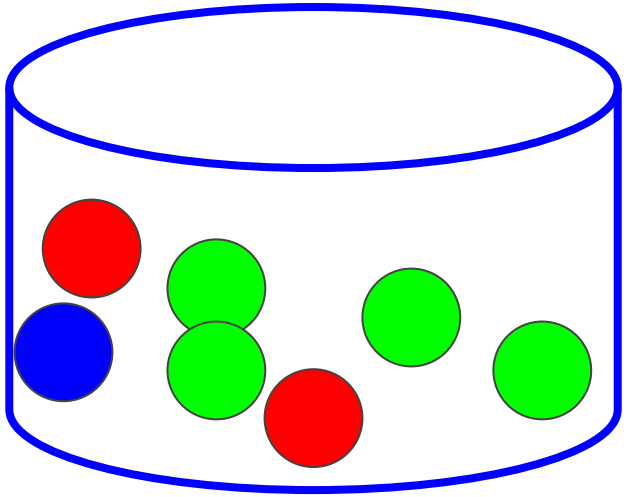


Calculating Probabilities

- There are currently 27 teams in Major League Soccer here in the United States. What is the probably that Minnesota United (Go Loons!) win the league championship this year?
  - P(MNUFC Champions) = UNKNOWN!! Outcomes are not equally likely.

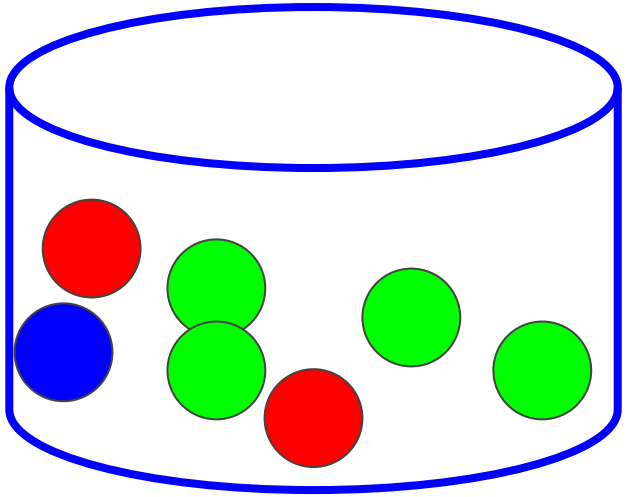
Sample Space

- 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 = {Head, Tail}
  - Sample space for drawing a single marble:



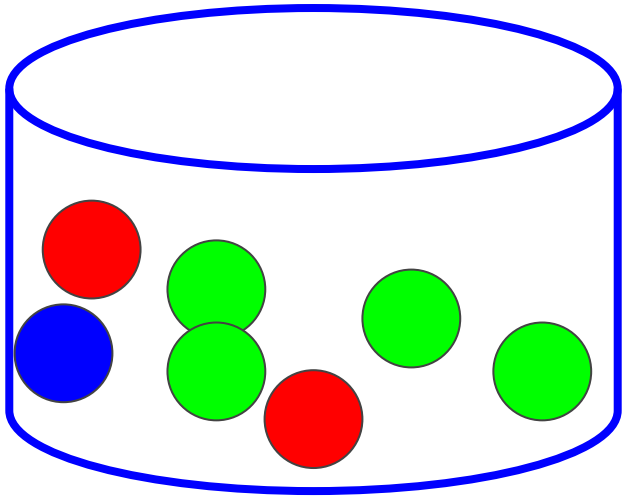
Sample Space

- 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 = {Head, Tail}
  - Sample space for drawing a single marble:
    - S = {Red, Green, Blue}



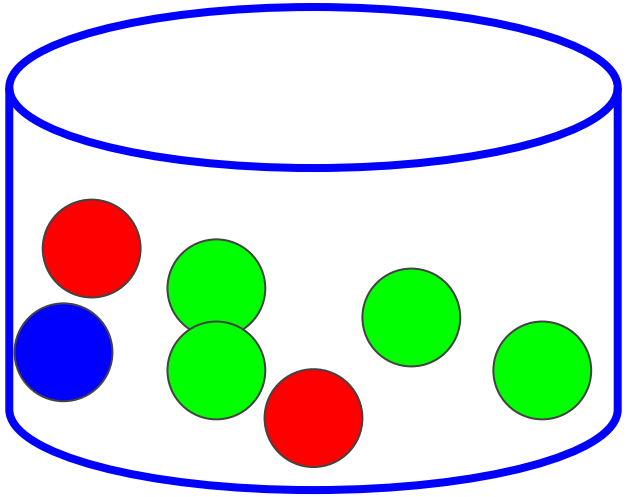
Sample Space

- 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 = {Head, Tail}
  - Sample space for drawing a single marble:
    - S = {Red, Green, Blue}
  - Sample space for flipping two coins:



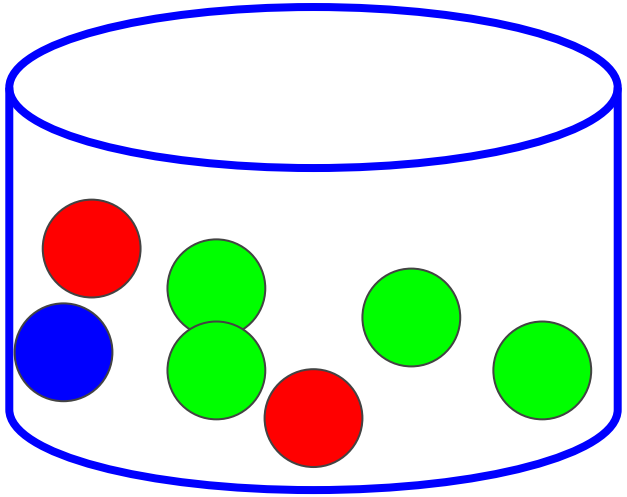
Sample Space

- 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 = {Head, Tail}
  - Sample space for drawing a single marble:
    - S = {Red, Green, Blue}
  - Sample space for flipping two coins:
    - S = {HH, HT, TH, TT} (why 4 elements?)



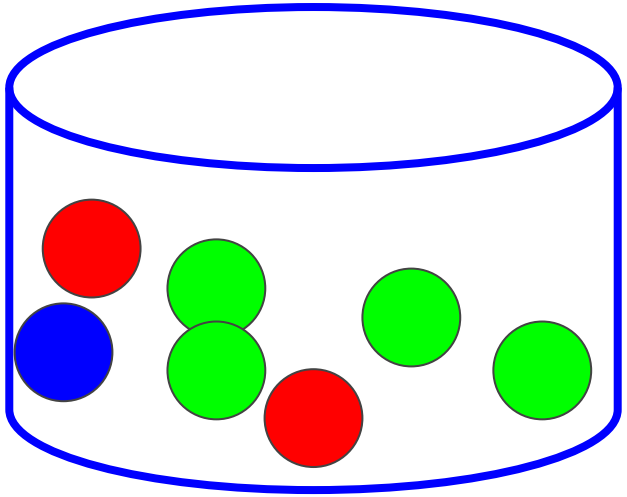
Sample Space

- 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 = {Head, Tail}
  - Sample space for drawing a single marble:
    - S = {Red, Green, Blue}
  - Sample space for flipping two coins:
    - S = {HH, HT, TH, 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 = {Head, Tail}
  - Sample space for drawing a single marble:
    - S = {Red, Green, Blue}
  - Sample space for flipping two coins:
    - S = {HH, HT, TH, TT} (why 4 elements?)
  - Sample space for drawing two marbles:
    - S = {RR, RG, RB, GG, GR, GB, BR, 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.

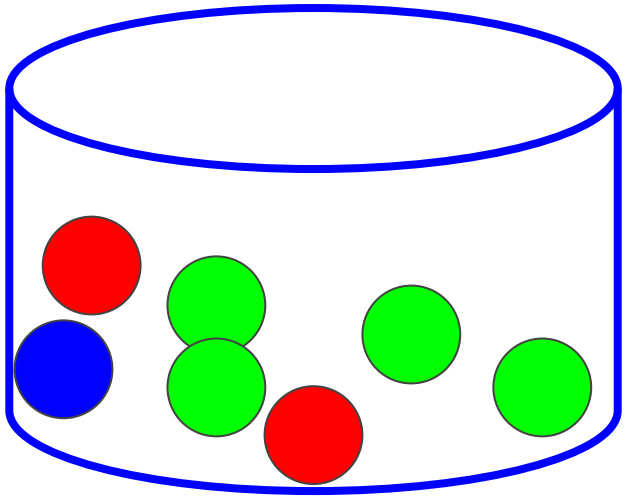
- You flip two coins, if we know that they are not both tails, what are the possible results?

S = { HH, ~~TT~~, HT, TH }

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

Applying Complementary Events

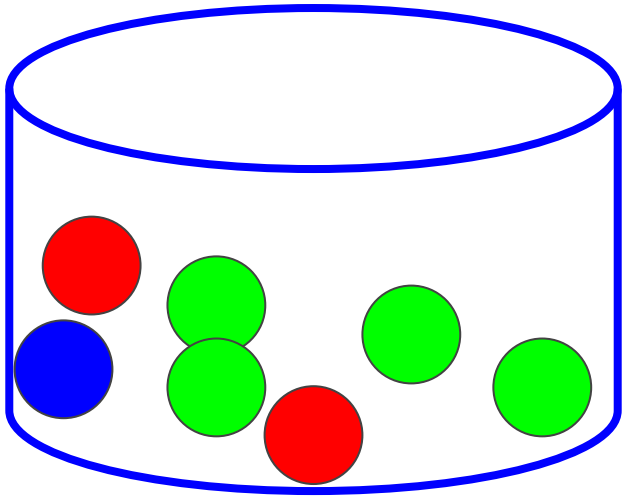
What's the probability of drawing a red or a green marble?



Applying Complementary Events

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

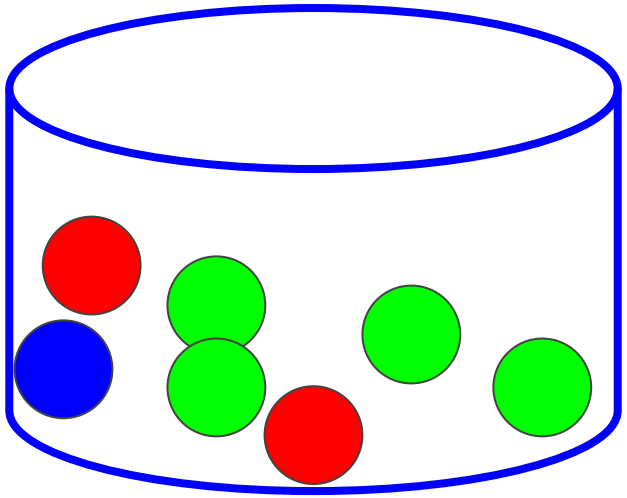
- S = {Red, Green, Blue}
- {Red, Green} and {Blue} are complementary



Applying Complementary Events

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

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



P(Red or Green) = 1 - P(Blue) = 1 - 1/7 = 6/7

Disjoint vs. Complementary

Disjoint events are two events that can't both happen.
They are *mutually exclusive*

- Do the sum of probabilities of two complementary outcomes always add up to 1?
- Do the sum of probabilities of two disjoint 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.

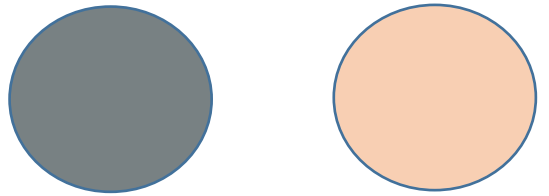
- How we combine depends on how constituent events are related.
  - Do both events happen at the same time?
  - Is one event contingent on the other?

How can two (or more) events happen?

Combining Probabilities: Terminology

Intersection of two events: A *AND* B

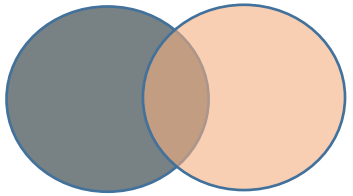
- The set of all outcomes where both A and B happen.



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

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

- 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.



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

Non-disjoint outcomes: Can happen at the same time.

- A student can get an A in Stats and A in Econ in the same semester.
- We will see how to calculate these soon.

Combining Probabilities: Terminology

Union of two events: A *OR* B

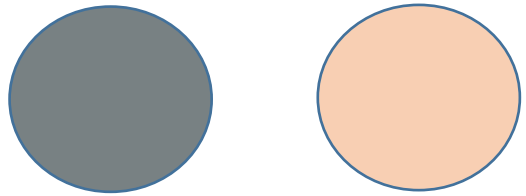
- The set of all outcomes where either A *or* B (or both) happen.

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.

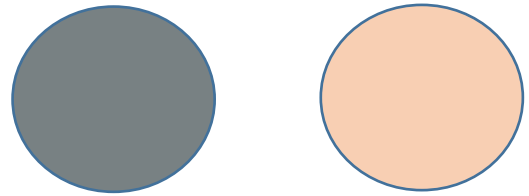
$$P(\text{A or B}) = P(A) + P(B)$$

Combining Probabilities: Terminology

Union of two events: A *OR* B

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

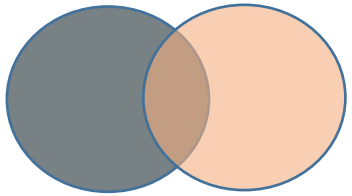
Disjoint (mutually exclusive) outcomes:



- 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(Jack) = 4/52 = 1/13

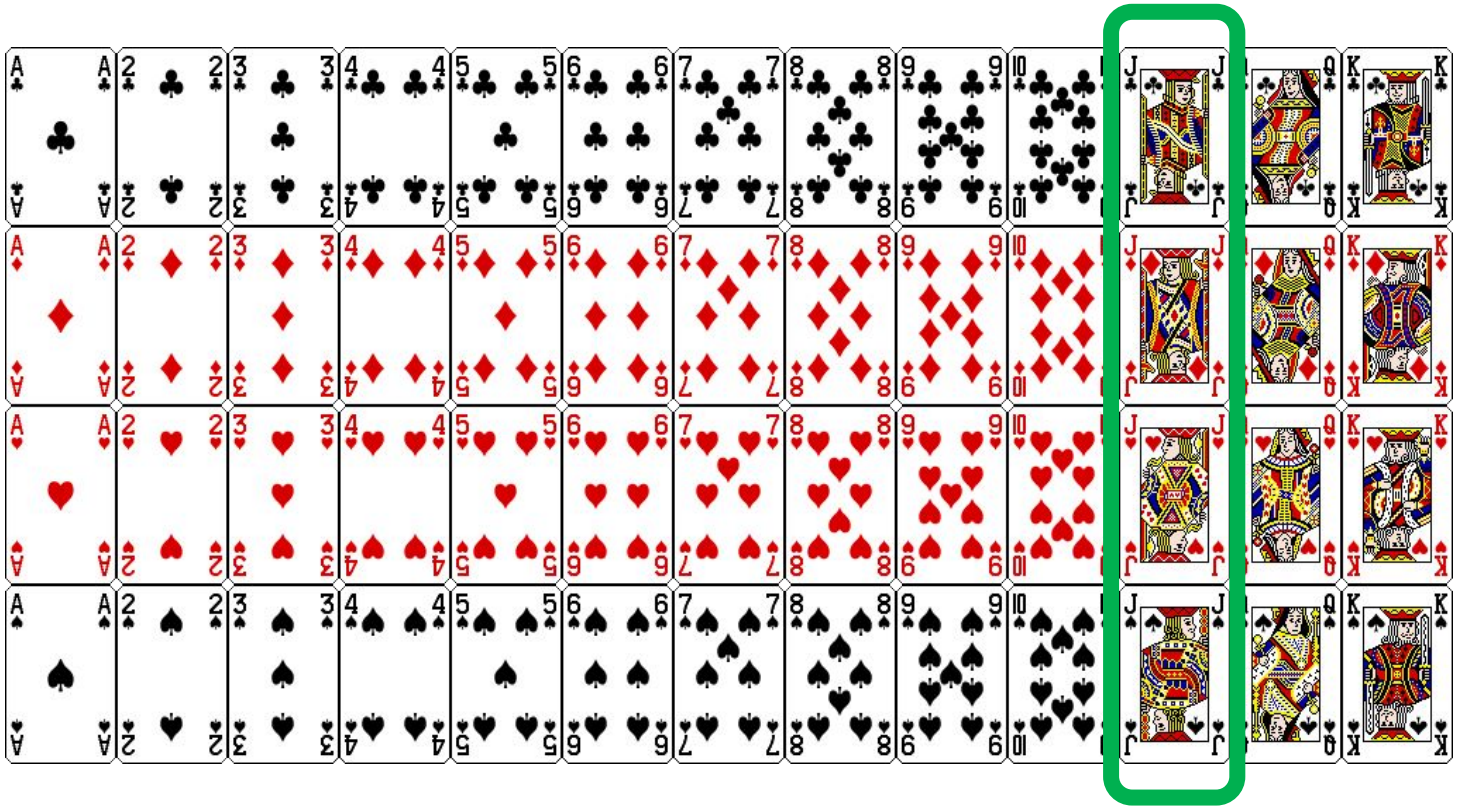


Figure from http://www.milefoot.com/math/discrete/counting/cardfreq.htm

Example: Intersection of Events

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

P(red) = 26/52 = 1/2

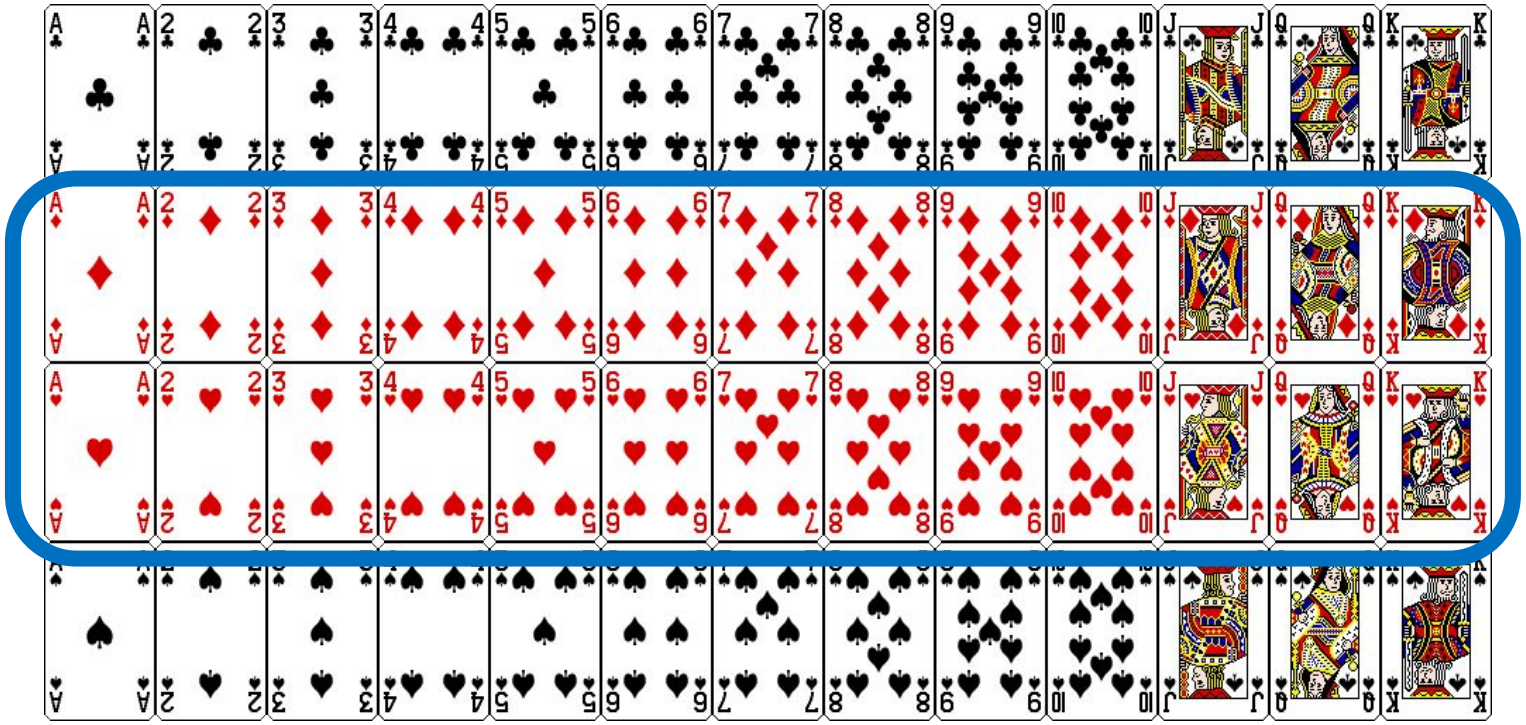


Figure from http://www.milefoot.com/math/discrete/counting/cardfreq.htm

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(red and jack)
= 2/52 = 1/26

P(red and jack) ≠ 0
so "being red" and "being a jack" are *non-disjoint*

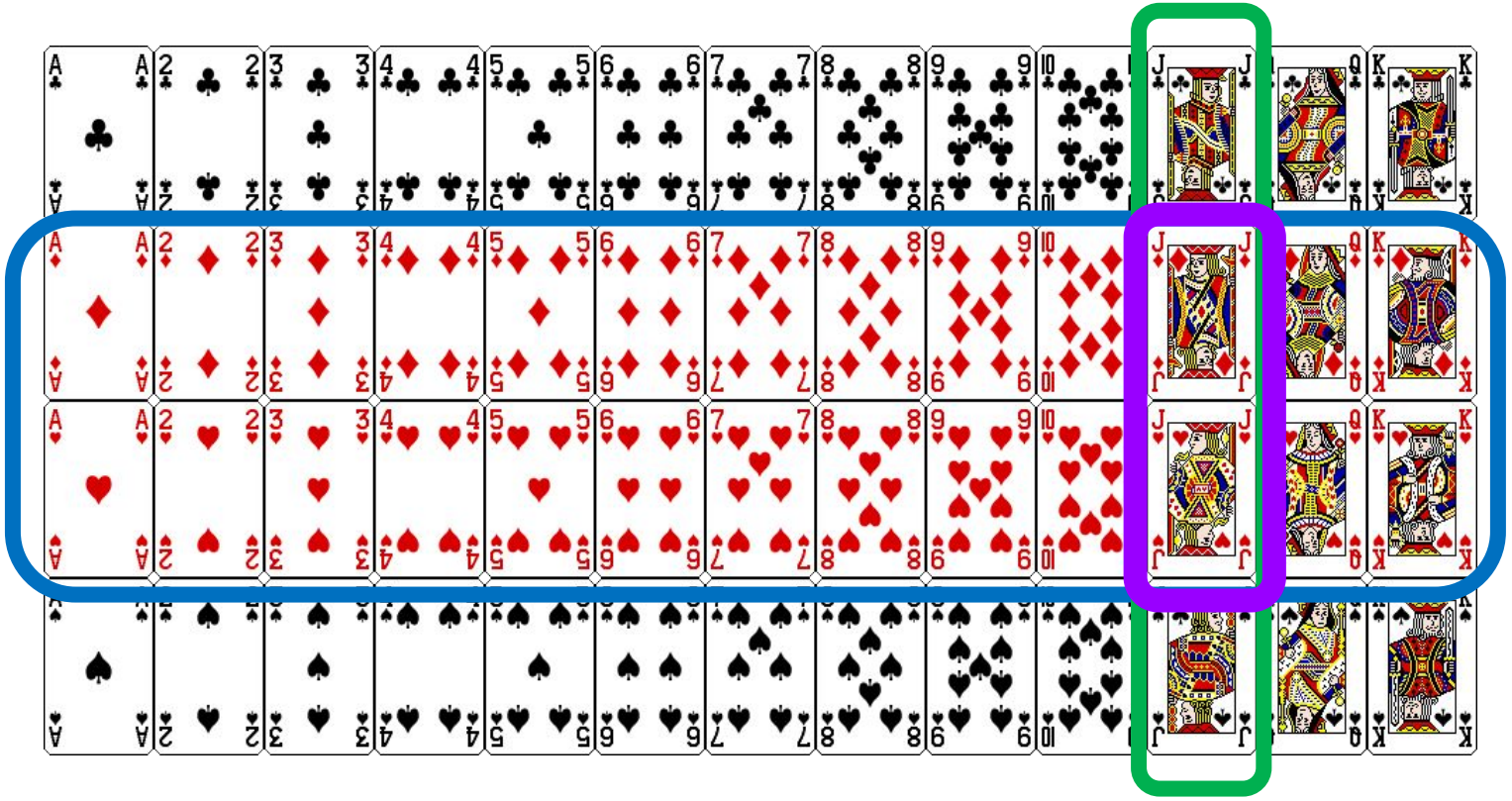


Figure from http://www.milefoot.com/math/discrete/counting/cardfreq.htm

Example: Intersection of Events

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

P(King) = 4/52 = 1/13

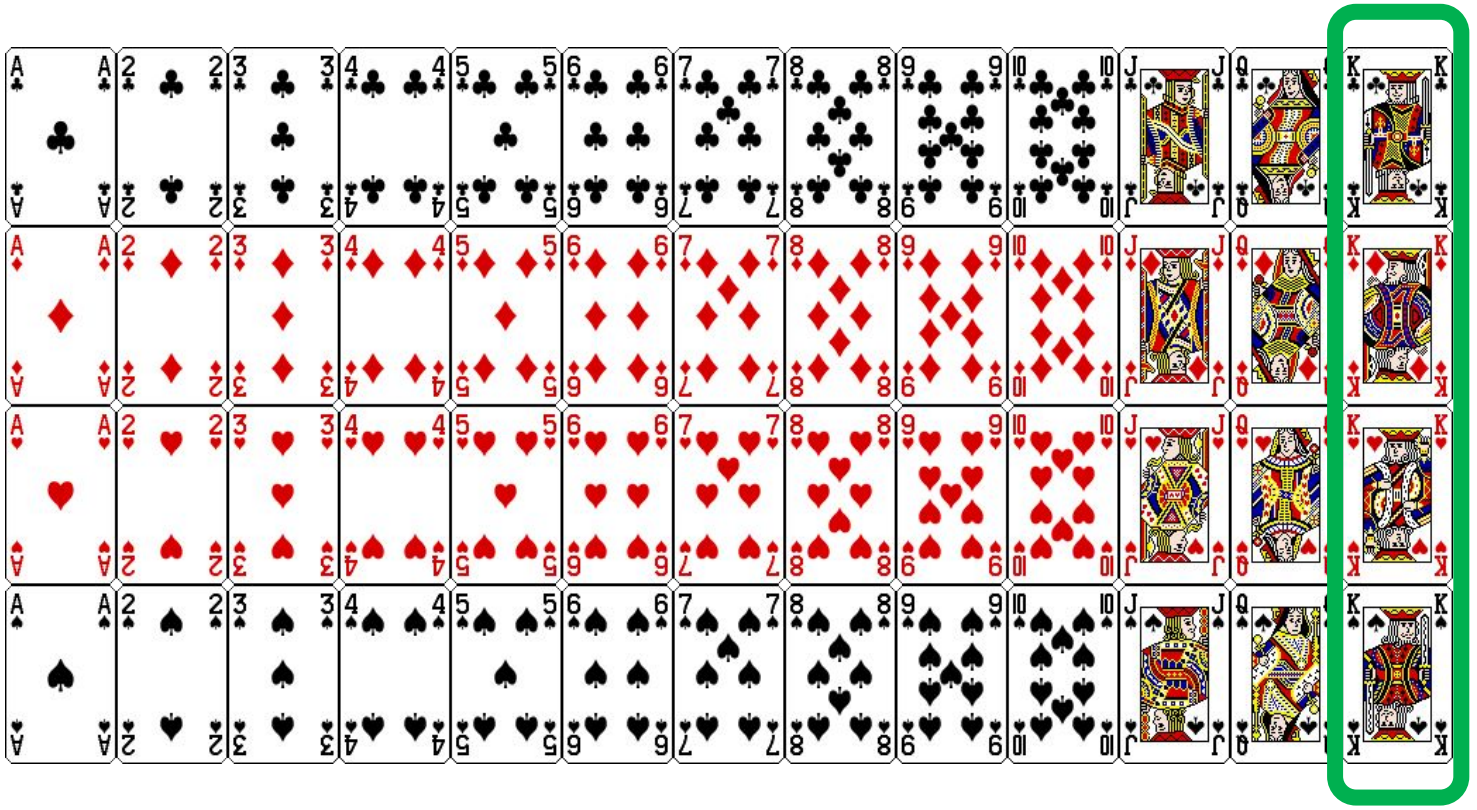


Figure from http://www.milefoot.com/math/discrete/counting/cardfreq.htm

Example: Intersection of Events

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

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

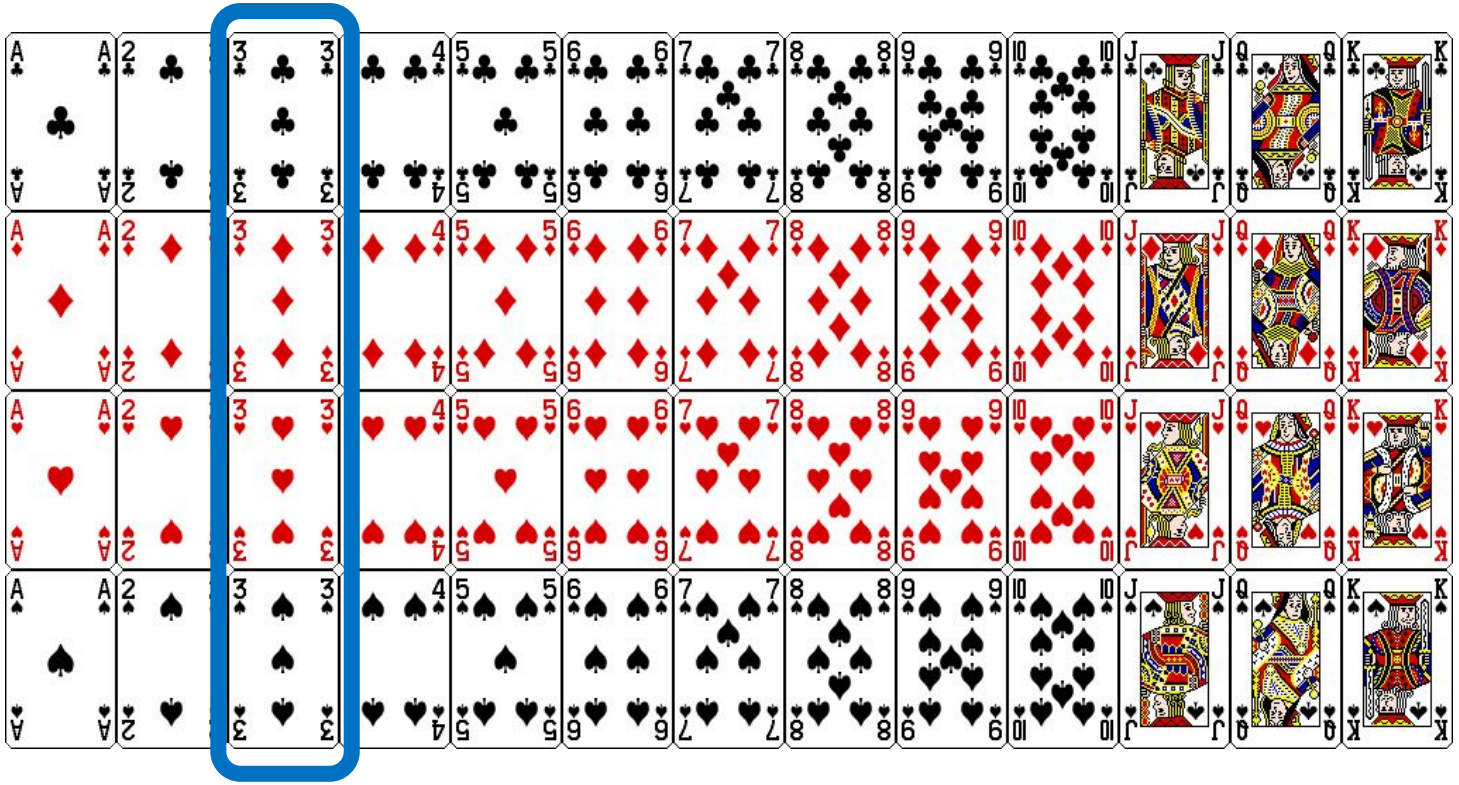


Figure from http://www.milefoot.com/math/discrete/counting/cardfreq.htm

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(king and 3)
= 0/52 = 0

P(king and 3) = 0
so “being a king” and
“being a 3” are *disjoint*

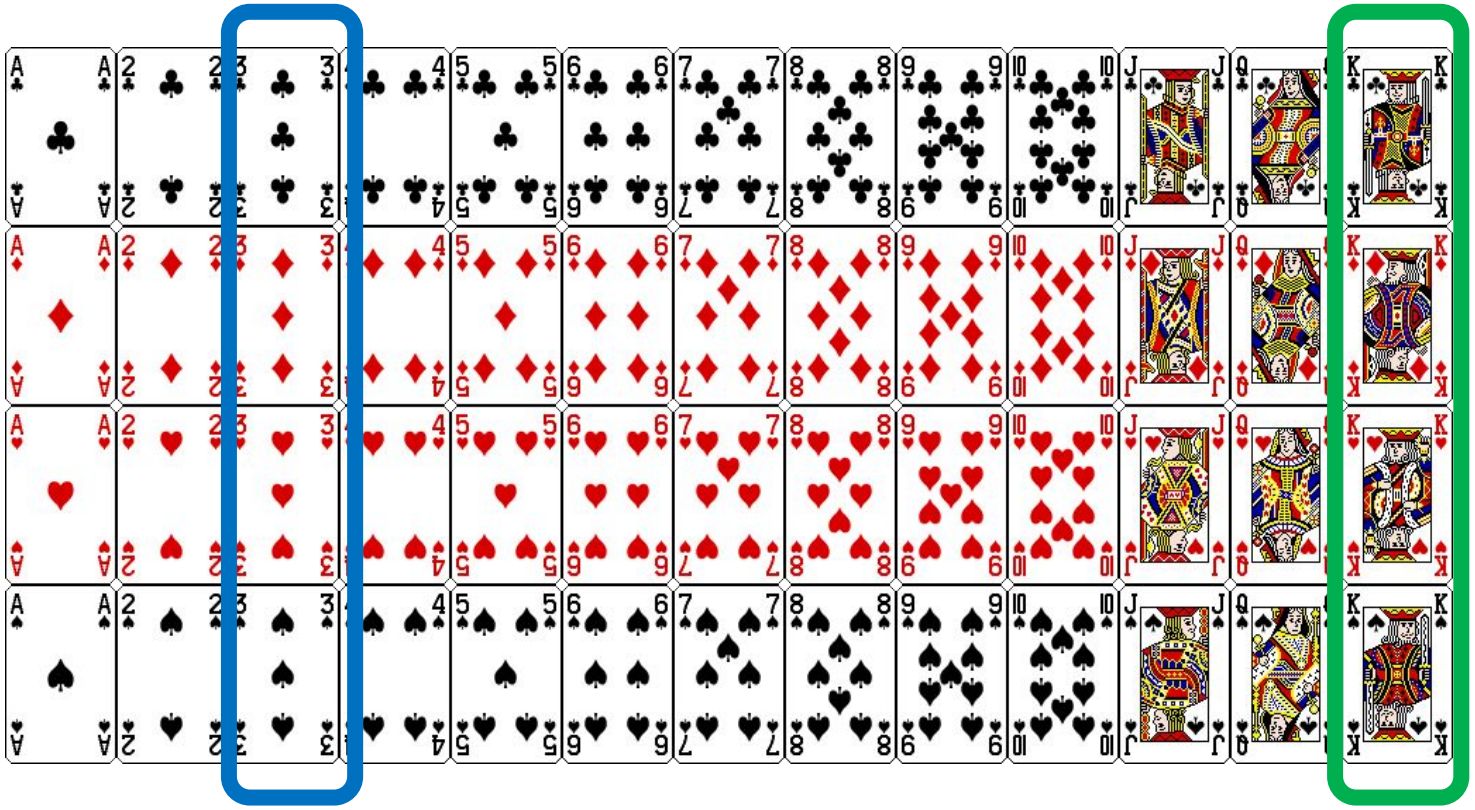


Figure from http://www.milefoot.com/math/discrete/counting/cardfreq.htm

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(Red OR Jack) = 28/52
= 7/13

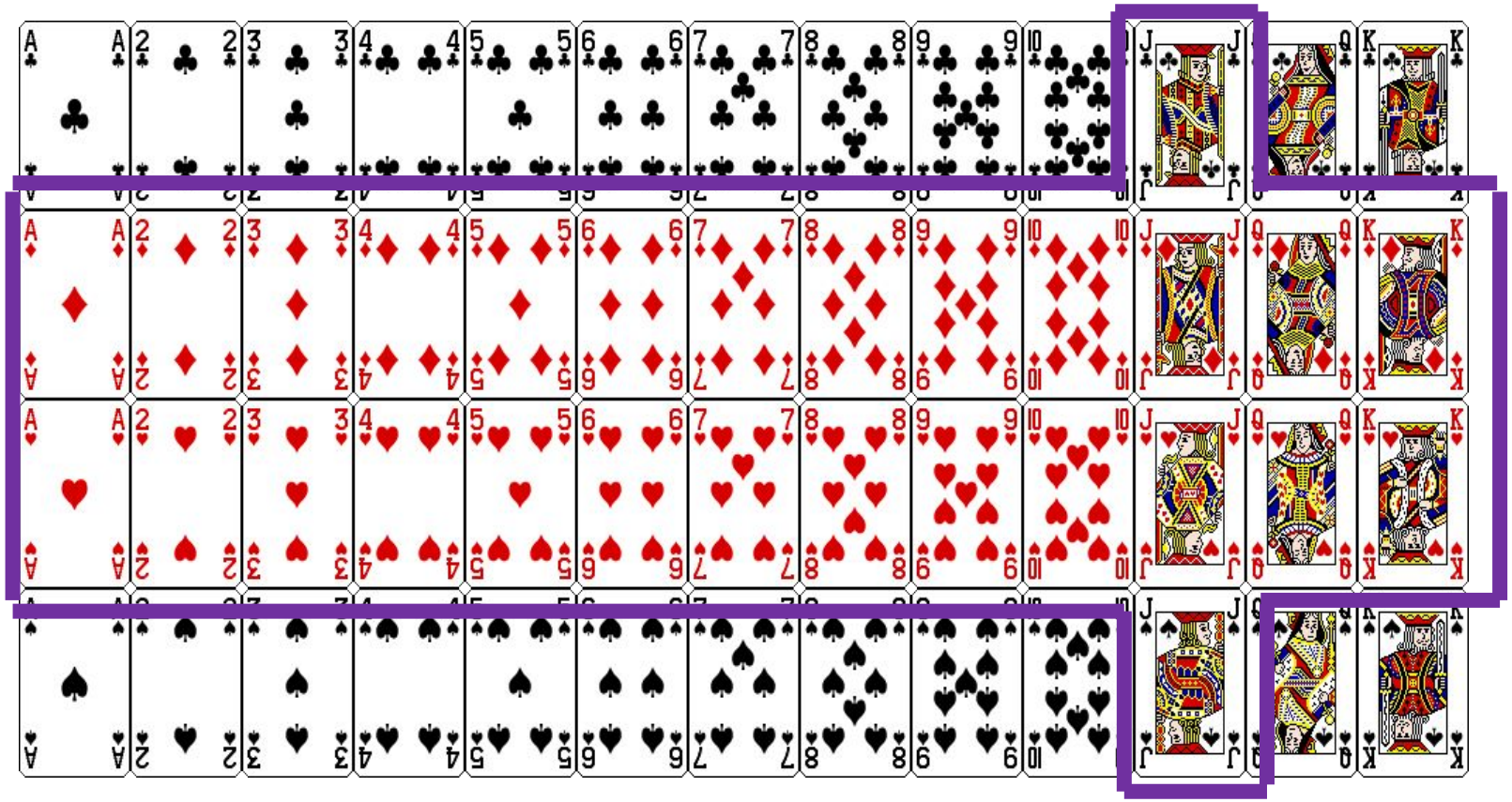


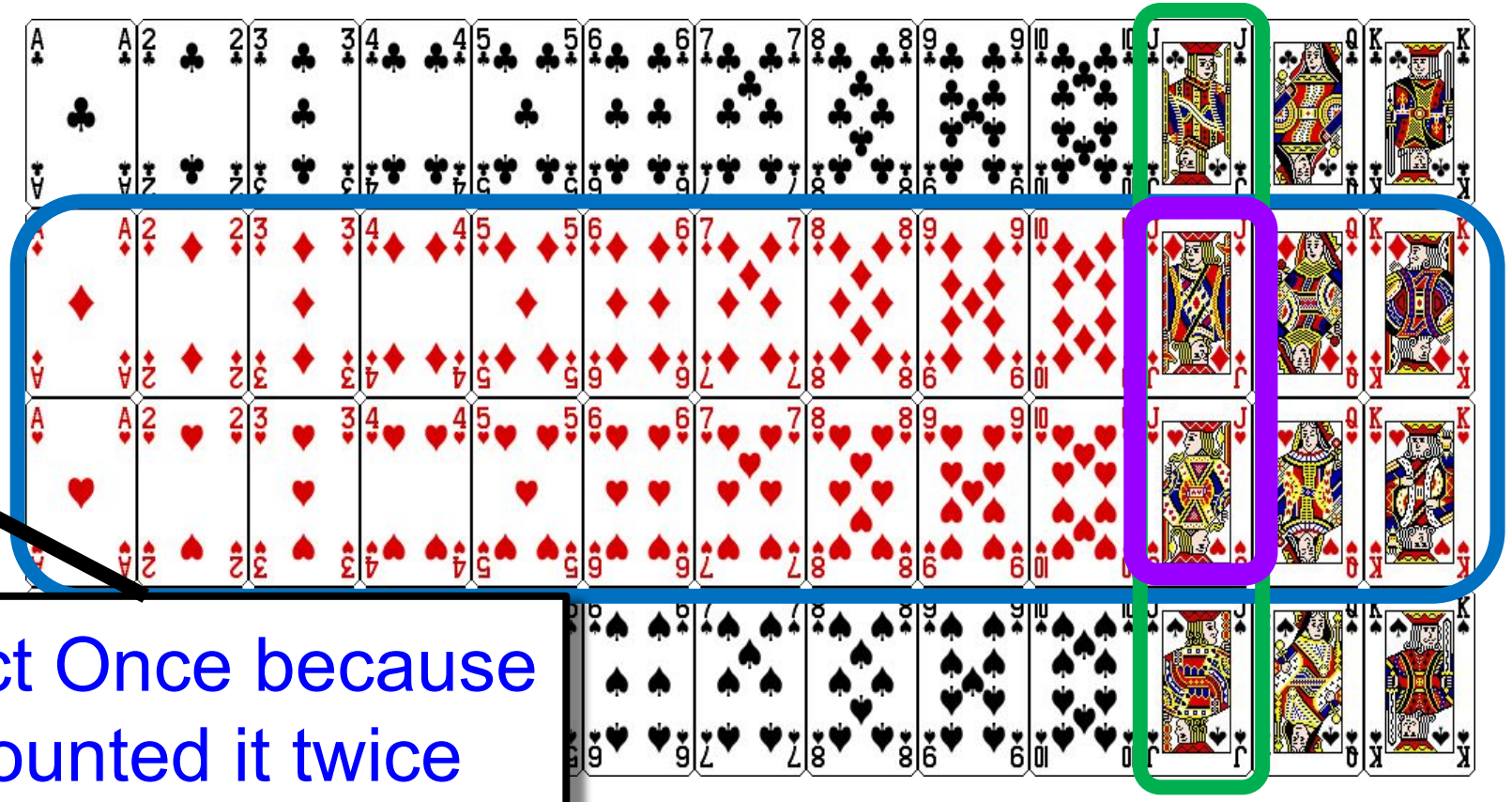
Figure from http://www.milefoot.com/math/discrete/counting/cardfreq.htm

Example: Union of Events

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

Approach 2: Consider the events separately

P(Red OR Jack)
= P(Red) + P(Jack) - P(Red and Jack)
= 26/52 + 4/52 - 2/52
= 28/52 = 7/13



Subtract Once because we counted it twice

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

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

P(king OR 3) = 8/52
= 2/13

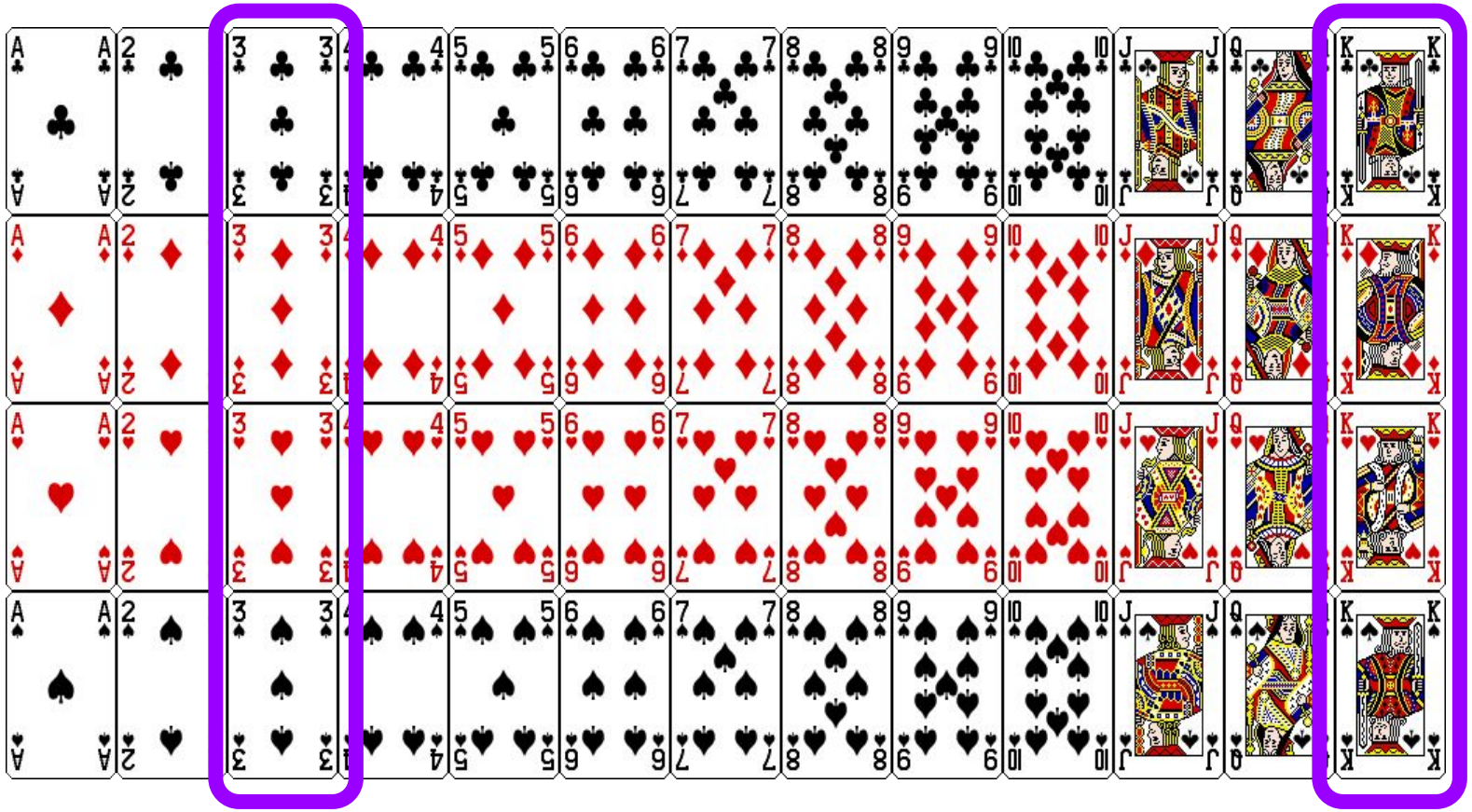


Figure from http://www.milefoot.com/math/discrete/counting/cardfreq.htm

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

P(King OR 3)
= P(King) + P(3) - P(King AND 3)

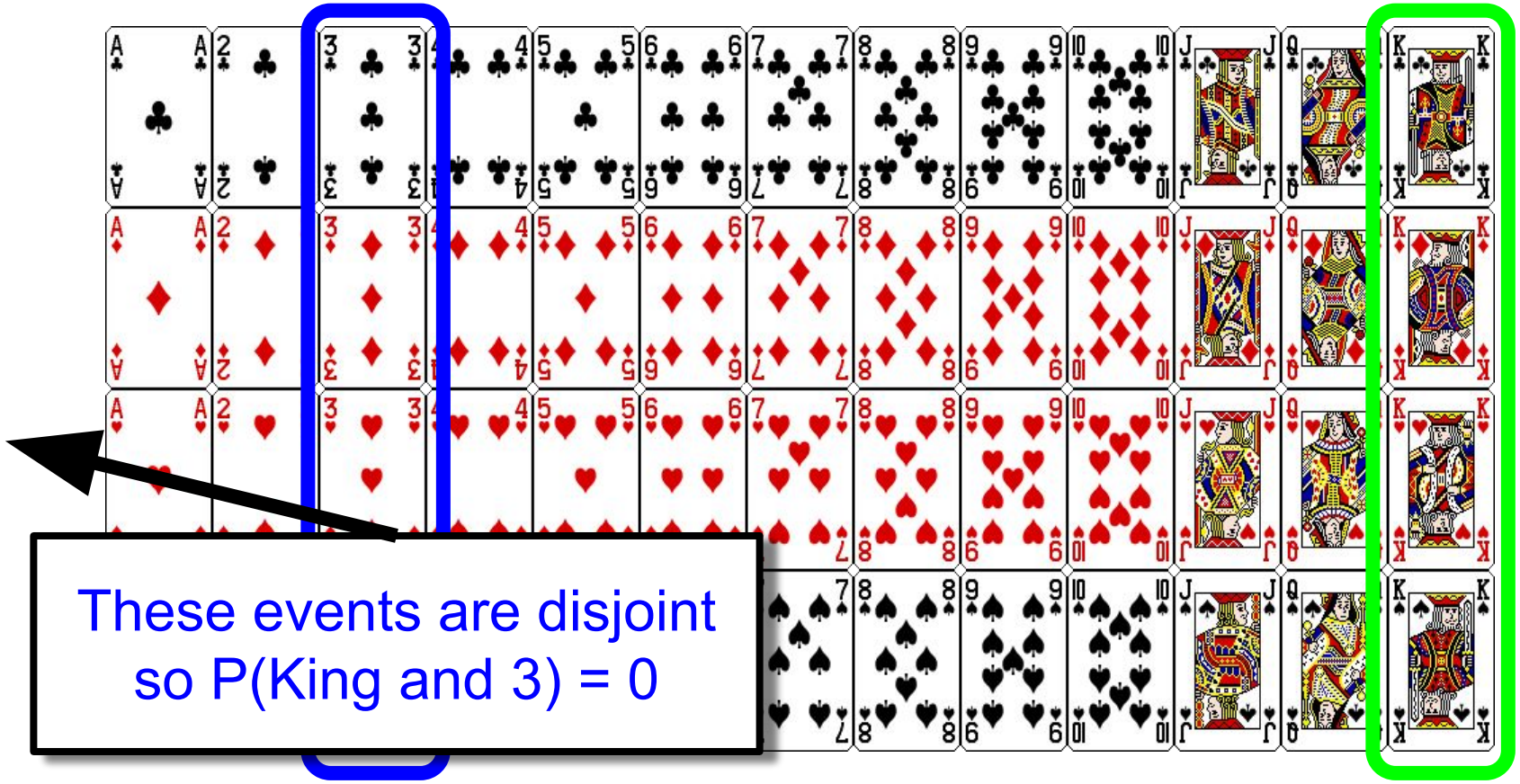


Figure from http://www.milefoot.com/math/discrete/counting/cardfreq.htm

Example: Union of Events

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

Approach 2: Consider the events separately

P(King OR 3)
= P(King) + P(3) - P(King AND 3)
= P(King) + P(3) - 0

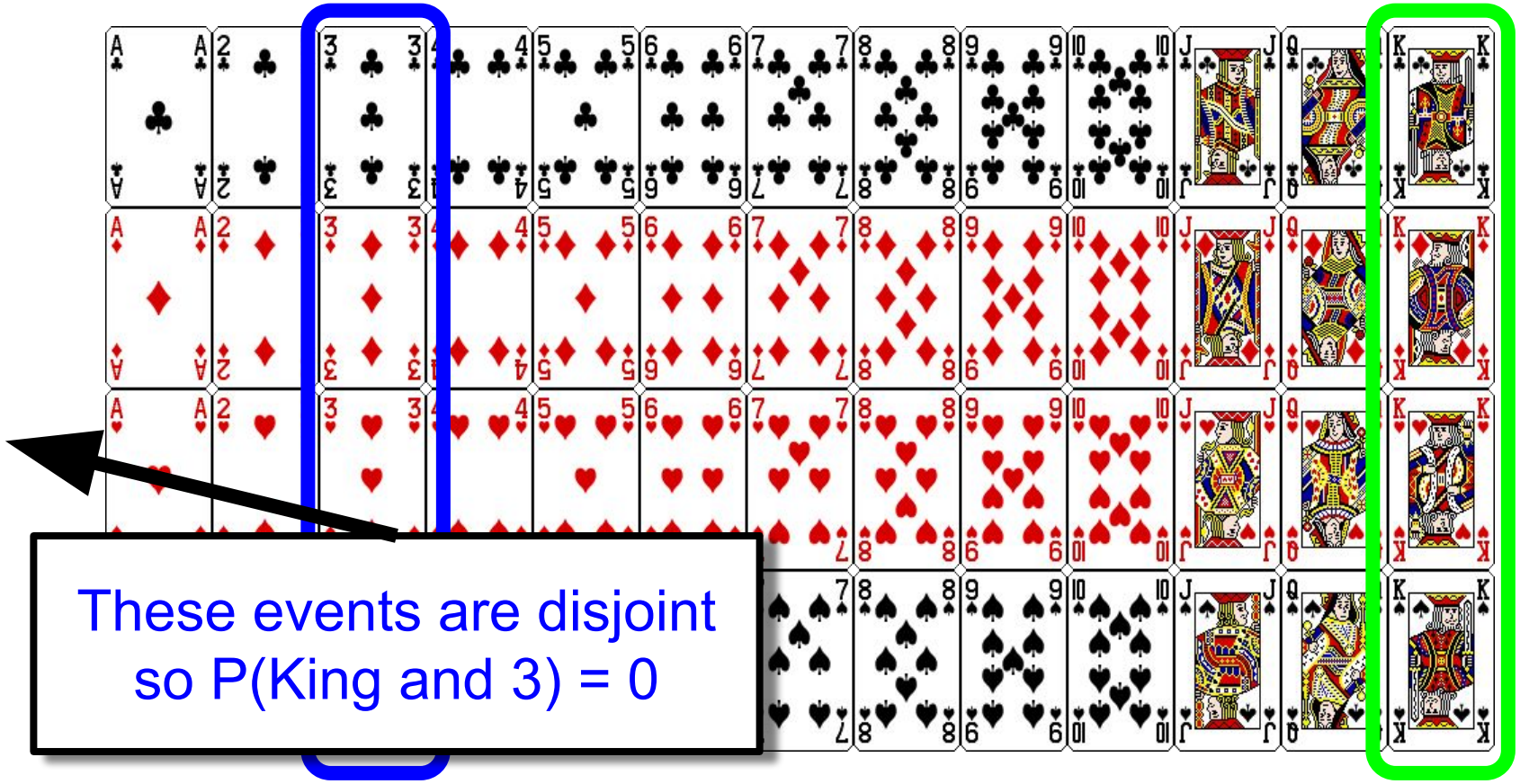


Figure from http://www.milefoot.com/math/discrete/counting/cardfreq.htm

Example: Union of Events

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

Approach 2: Consider the events separately

P(King OR 3)
= P(King) + P(3) - P(King AND 3)
= P(King) + P(3) - 0
= P(King) + P(3)

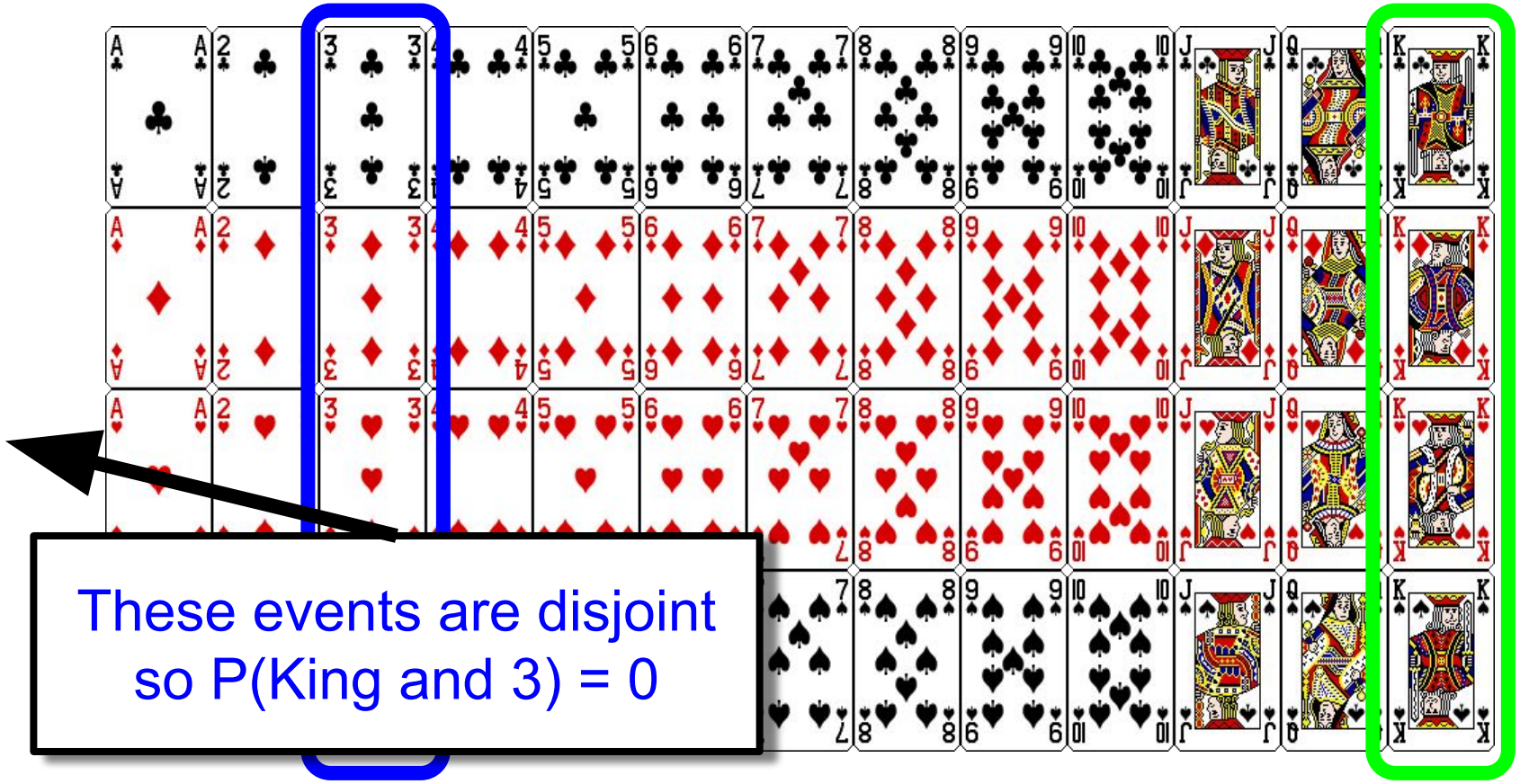


Figure from http://www.milefoot.com/math/discrete/counting/cardfreq.htm

Example: Union of Events

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

Approach 2: Consider the events separately

P(King OR 3)
= P(King) + P(3) - P(King AND 3)
= P(King) + P(3) - 0
= P(King) + P(3)
= 4/52 + 4/52 = 8/52 = 2/13

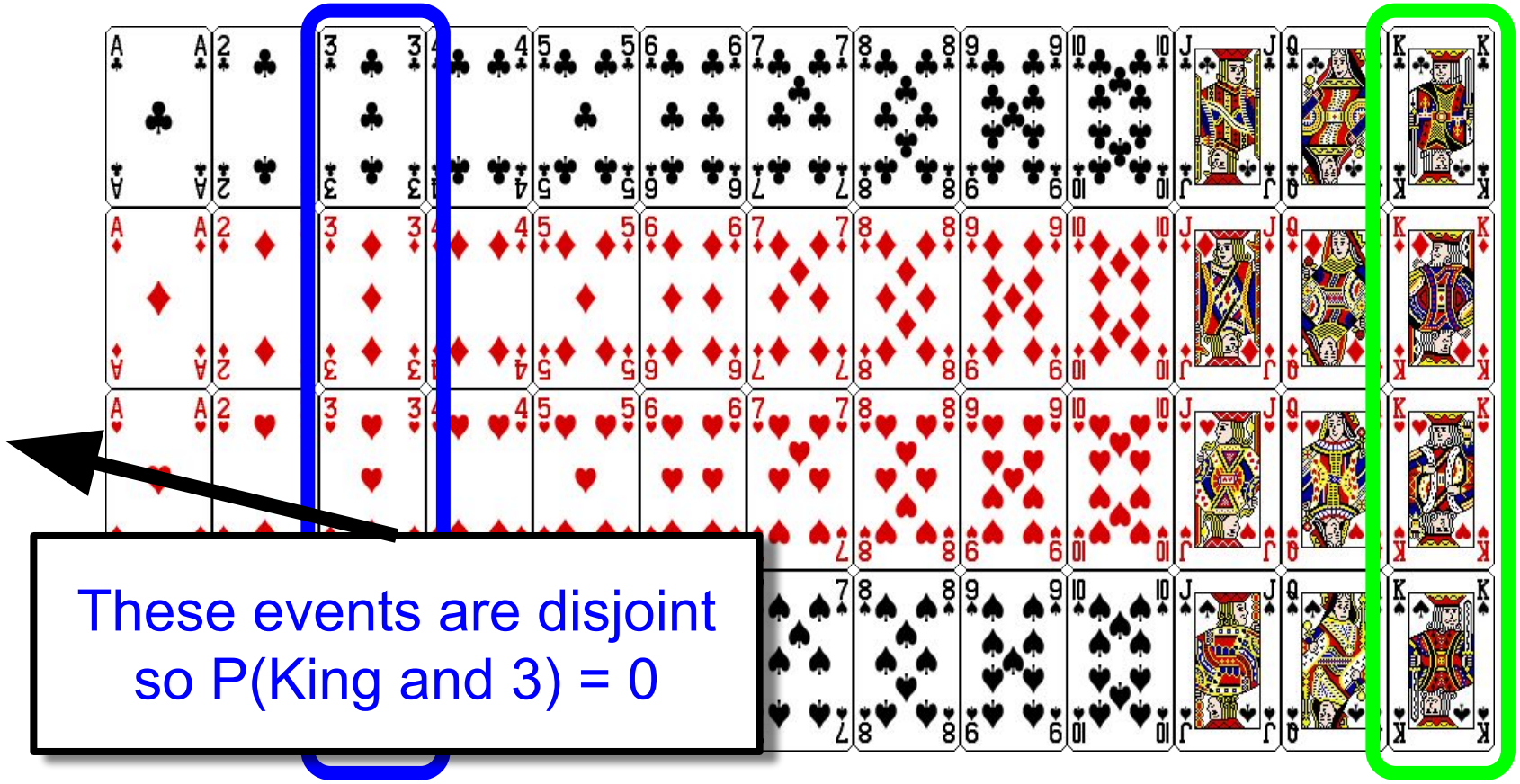
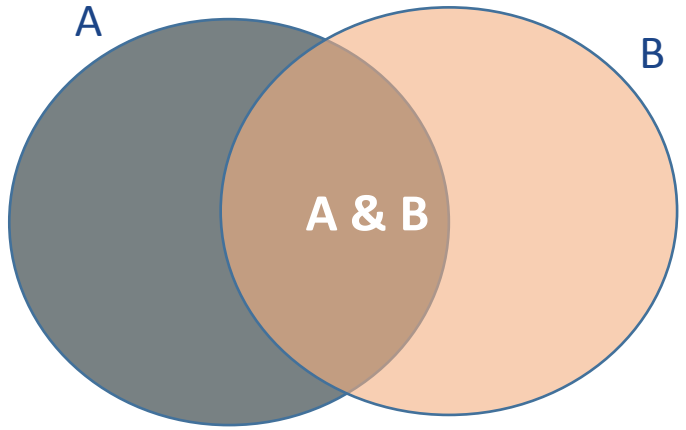


Figure from http://www.milefoot.com/math/discrete/counting/cardfreq.htm

Summary

General addition rule



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

Note: For disjoint events P(A 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?

	Share Parents' Politics		
Legalize MJ	No	Yes	Total
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