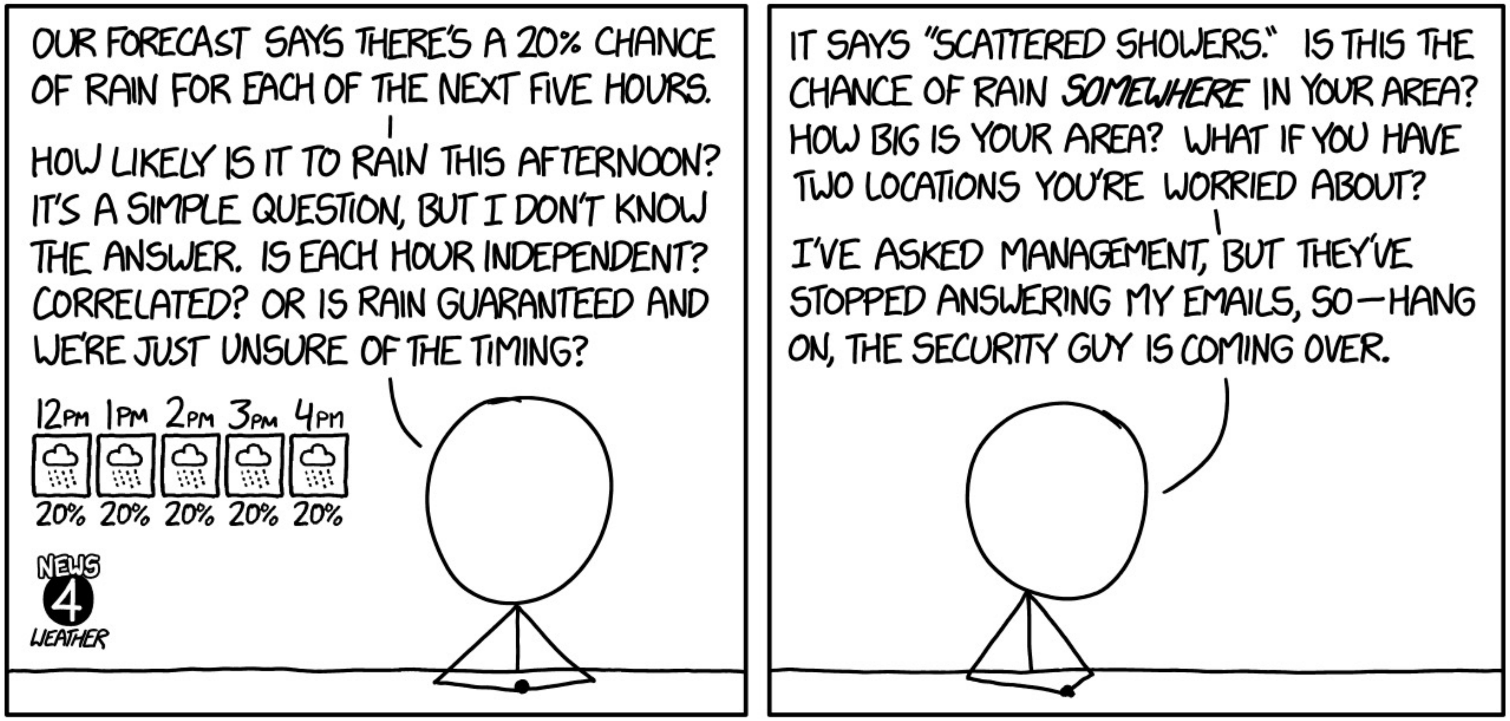
Stat 88: Probability and Mathematical Statistics in Data Science


OUR FORECAST SAYS THERE'S A 20% CHANCE OF RAIN FOR EACH OF THE NEXT FIVE HOURS.
HOW LIKELY IS IT TO RAIN THIS AFTERNOON? IT'S A SIMPLE QUESTION, BUT I DON'T KNOW THE ANSWER. IS EACH HOUR INDEPENDENT? CORRELATED? OR IS RAIN GUARANTEED AND WE'RE JUST UNSURE OF THE TIMING?
12PM 1PM 2PM 3PM 4PM
20% 20% 20% 20% 20%
NEWS 4 WEATHER
IT SAYS "SCATTERED SHOWERS." IS THIS THE CHANCE OF RAIN SOMEWHERE IN YOUR AREA? HOW BIG IS YOUR AREA? WHAT IF YOU HAVE TWO LOCATIONS YOU'RE WORRIED ABOUT?
I'VE ASKED MANAGEMENT, BUT THEY'VE STOPPED ANSWERING MY EMAILS, SO—HANG ON, THE SECURITY GUY IS COMING OVER.

https://imgs.xkcd.com/comics/meteorologist.png

Lecture 1: 1/18/2022

Course introduction and the basics, 1.1, 1.2

Shobhana M. Stoyanov

Agenda

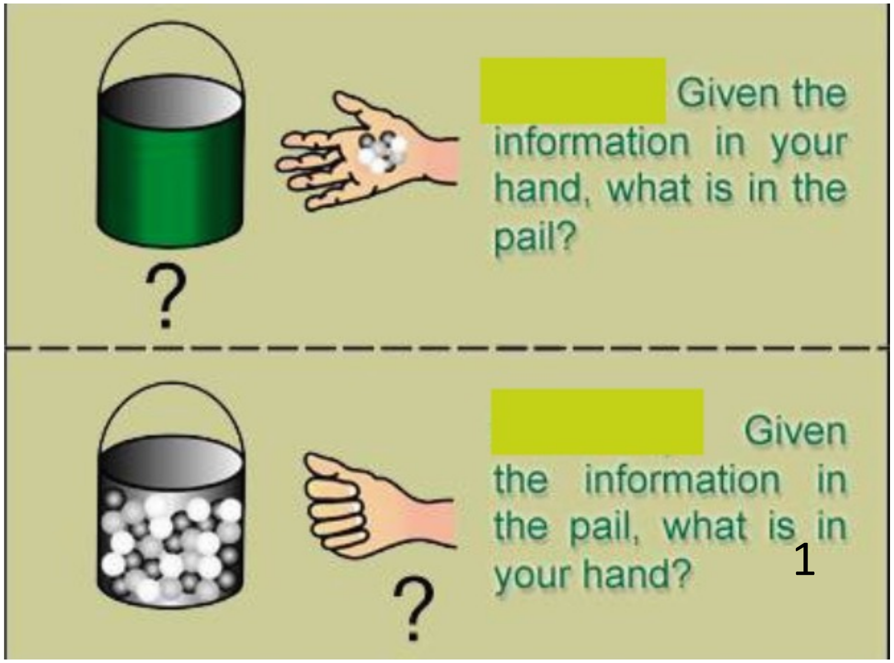
- Course resources:
  - Course site: http://stat88.org
  - Announcements and discussions: Piazza
  - Assignments and grades: Gradescope

- Write your questions on the google doc (we will have one for each lecture): https://tinyurl.com/2p9d58t6

- The Basics:
  - Section 1.1: Probabilities as Proportions
  - Section 1.2: Exact Calculation or Bound

Probability vs Statistics

- Discuss which is probability and which is statistics:



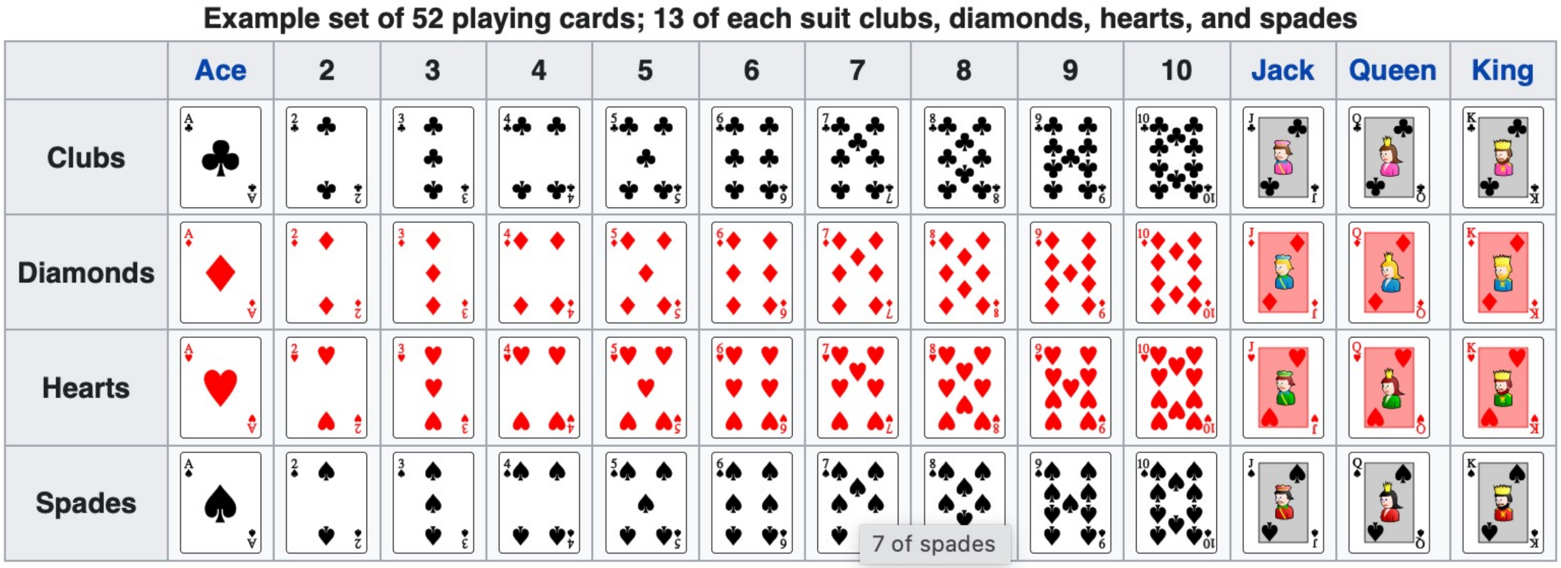
[redacted] Given the information in your hand, what is in the pail?

[redacted] Given the information in the pail, what is in your hand?

1

2

Cards



Example set of 52 playing cards; 13 of each suit clubs, diamonds, hearts, and spades

	Ace	2	3	4	5	6	7	8	9	10	Jack	Queen	King
Clubs													
Diamonds													
Hearts													
Spades													

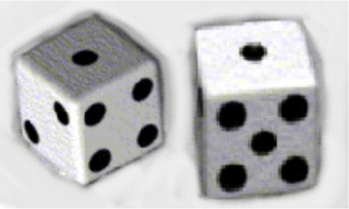
If you have a well-shuffled deck of cards, and deal 1 card from the top, what is the chance of it being the queen of hearts? What is the chance that it is a queen (any suit)? What assumptions are you making?

If you deal 2 cards, what is the chance that at least *one* of them is a queen? How do these relate to populations and samples?

Section 1.1: Probabilities as proportions

- We can think about probability as a numerical measure of uncertainty, and we will define some basic principles for computing these numbers.
- These basic computational principles have been known for a long time, and in fact, gamblers thought about these ideas a lot. Then mathematicians investigated the principles.

- Famous problem: will the probability of **at least one six** in **four** throws of a die be equal to prob of **at least a double six** in 24 throws of a pair of dice.

- Note: single = die, plural = dice: vs 

Origins of probability: de Méré's paradox

Questions that arose from gambling with dice.

Antoine Gombaud,
Chevalier de Méré

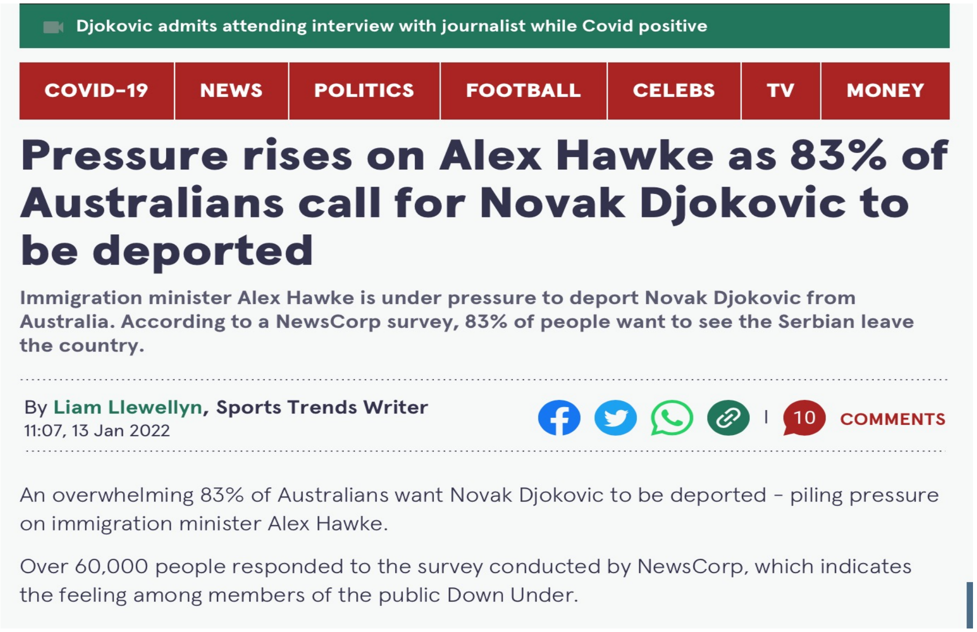
Blaise Pascal

Pierre de Fermat

The dice players
Georges de La Tour
(17th century)

Terminology

- Suppose we have an action that results in exactly one of several possible *outcomes* or results, and chance or randomness is involved - that is, each time we perform the action, the outcome will be different, and we don't know exactly which outcome will occur.
- Such an action is called an *experiment* or a *random experiment*.
- Examples: toss a coin, roll a die, take a random sample of people and see how many agree with Australian government's decision to deport Novak Djokovic.


Djokovic admits attending interview with journalist while Covid positive
COVID-19
NEWS
POLITICS
FOOTBALL
CELEBS
TV
MONEY
10
COMMENTS

Pressure rises on Alex Hawke as 83% of Australians call for Novak Djokovic to be deported

Immigration minister Alex Hawke is under pressure to deport Novak Djokovic from Australia. According to a NewsCorp survey, 83% of people want to see the Serbian leave the country.

By **Liam Llewellyn**, **Sports Trends Writer**
11:07, 13 Jan 2022

An overwhelming 83% of Australians want Novak Djokovic to be deported – piling pressure on immigration minister Alex Hawke.

Over 60,000 people responded to the survey conducted by NewsCorp, which indicates the feeling among members of the public Down Under.

Terminology

- Suppose we have an action that results in exactly one of several possible *outcomes* or results, and chance or randomness is involved - that is, each time we perform the action, the outcome will be different, and we don't know exactly which outcome will occur.
- Such an action is called an *experiment* or a *random experiment*.

- A collection of all possible outcomes of an action is called a *sample space* or an *outcome space* . Usually denoted by Ω (sometimes also by *S*).
- An *event* is a collection of outcomes, so a subset of Ω.

Computing probabilities

- If you have a well-shuffled deck of cards, and deal 1 card from the top, what is the chance of it being the queen of hearts? What is the chance that it is a queen (any suit)?

- How did you do this? What were your assumptions?

- Say we roll a die. What is Ω?

- What is the chance that the die shows a multiple of 3? What were your assumptions?

Chance of a particular outcome

- We usually think of the chance of a particular outcome (roll a 6, coin lands heads etc) as the number of ways to get that outcome divided by the total possible number of outcomes.

$$\frac{\# \; of \; particular \; outcomes \; of \; interest}{total \; \# \; of \; outcomes \; possible}$$

- So if A is an event (subset of Ω), then $P(A) = \frac{\#(A)}{\#(\Omega)}, A \subseteq \Omega$

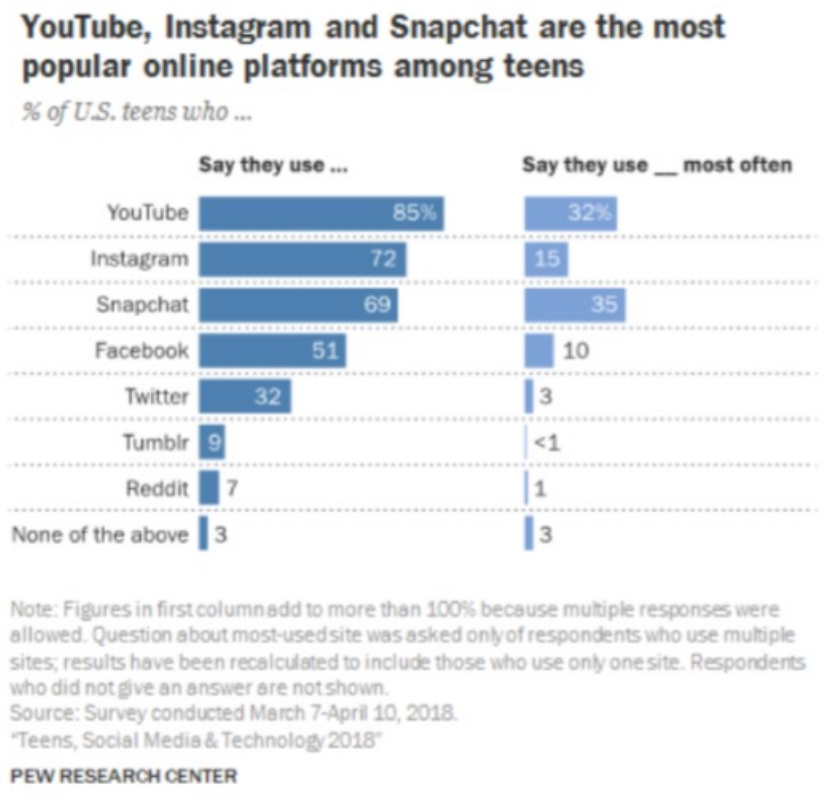
Cards

- If you have a well-shuffled deck of cards, and deal 1 card from the top, what is the chance of it being the queen of hearts? What is the chance that it is a queen (any suit)?

- If you deal 2 cards, what is the chance that at least *one* of them is a queen?

Not equally likely outcomes

- What if our assumptions of equally likely outcomes don't hold (as is often true in life, data are messier than nice examples).
- Here is a graphic from Pew Research displaying the results of a 2018 survey of social media use by US teens.



YouTube, Instagram and Snapchat are the most popular online platforms among teens

% of U.S. teens who ...

Note: Figures in first column add to more than 100% because multiple responses were allowed. Question about most-used site was asked only of respondents who use multiple sites; results have been recalculated to include those who use only one site. Respondents who did not give an answer are not shown.
Source: Survey conducted March 7-April 10, 2018.
"Teens, Social Media & Technology 2018"

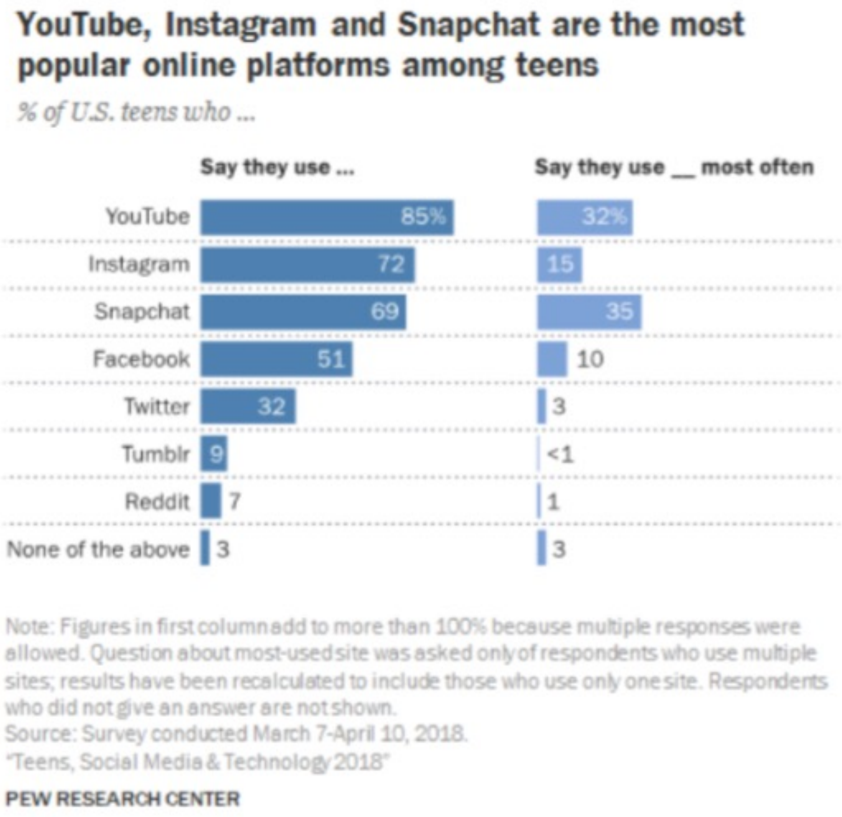
PEW RESEARCH CENTER

- What is the difference b/w 2 charts?

- Why do the % add up to more than 100 in the first graph?

- Second graph gives us a *distribution* of teens over the different categories

Not equally likely outcomes



YouTube, Instagram and Snapchat are the most popular online platforms among teens

% of U.S. teens who ...

Note: Figures in first column add to more than 100% because multiple responses were allowed. Question about most-used site was asked only of respondents who use multiple sites; results have been recalculated to include those who use only one site. Respondents who did not give an answer are not shown.
Source: Survey conducted March 7-April 10, 2018.
"Teens, Social Media & Technology 2018"

PEW RESEARCH CENTER

1. What is the chance that a randomly picked teen uses FB most often?

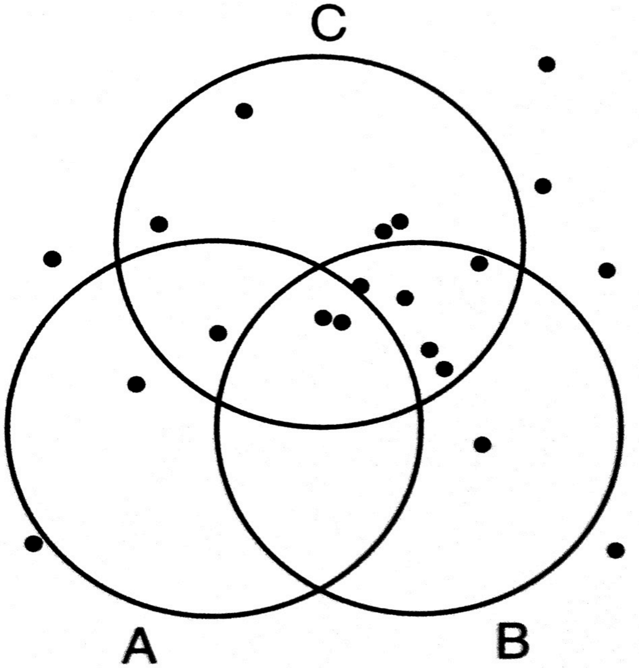
2. What is the chance that a randomly picked teen did *not* use FB most often?

3. What is the chance that FB *or* Twitter was their favorite?

4. What is the chance that the teen used FB, just not most often?

5. Given that the teen used FB, what is the chance that they used it most often?

Venn Diagrams



Consider the Venn diagram above. (The sample space consists of all the dots.) What is the probability of A? What about A or B? A or B or C?

So far:

- If all the possible outcomes are *equally likely*, then each outcome has probability $1/n$, where $n = \#(\Omega)$
- Let $\mathrm{A} \subseteq \Omega$, $P(A) = \frac{\#(A)}{\#(\Omega)}$
- Probabilities as proportions
- Sum of the probabilities of all the distinct outcomes should add to 1
- $0 \leq P(A) \leq 1, \mathrm{A} \subseteq \Omega$
- A *distribution* of the outcomes over different categories is when each outcome appears in one and only one category.
- Venn diagrams

- When we get some information about the outcome or event whose probability we want to figure out, ***our outcome space reduces***, incorporating that information.

Conditional probability

- In the last question, we used the information that the teen used FB. We were told the teen used FB, and *then* asked to compute the chance that FB was their favorite.
- This is called the *conditional probability that the teen used Facebook most often, given that they used Facebook* and denoted by:

Conditional probability

- This probability we computed is called a ***conditional probability***. It puts a condition on the teen, and *changes* (restricts) the universe (the sample space) of the next outcome, a teen who likes FB best.
- To compute a conditional probability:
  - First restrict the set of all outcomes as well as the event to *only* the outcomes that *satisfy* the given **condition**
  - Then calculate proportions accordingly
- How do the probabilities in #1 and #5 compare?

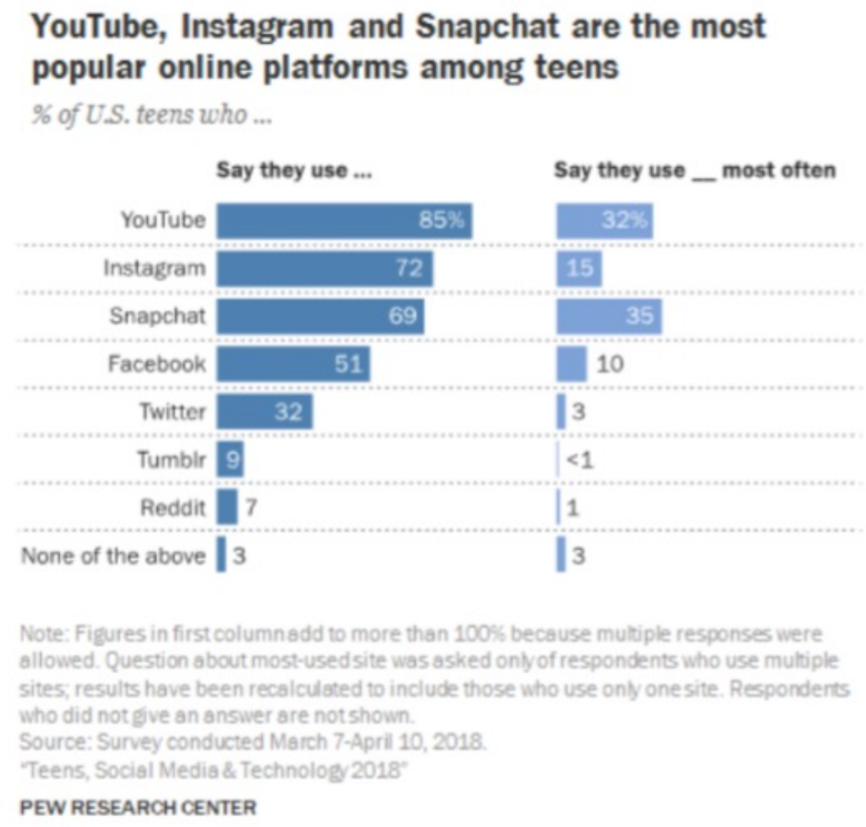
Example

- A ten-sided fair die is rolled twice:

  - If the first roll lands on 1, what is the chance that the second roll lands on a number bigger than 1?

  - Find the probability that the second number is greater than the twice the first number.

Section 1.2: Exact Calculations, or Bound?



Note: Figures in first column add to more than 100% because multiple responses were allowed. Question about most-used site was asked only of respondents who use multiple sites; results have been recalculated to include those who use only one site. Respondents who did not give an answer are not shown.
Source: Survey conducted March 7-April 10, 2018.
"Teens, Social Media & Technology 2018"

PEW RESEARCH CENTER

Recall #3 about FB or Twitter. What was the answer? What can you say about the chance that a randomly selected teen used FB or Twitter (not necessarily most often)?

Example with bounds

- Let A be the event that you catch the bus to class instead of walking, P(A) = 70%
- Let B be the event that it rains, P(B) = 50%
- Let C be the event that you are on time to class, P(C) = 10%
- What is the chance of **at least** one of these three events happening?

- What is the chance of **all three** of them happening?

Rules that we used:

- If all the possible outcomes are *equally likely*, then each outcome has probability *1/n*, where ***n*** = number of possible outcomes.
- If an event A contains k possible outcomes,

  then *P(A)* = *k/n*.

- Probabilities are between 0 and 1

- If two events A and B don't overlap, then the probability of A or B = P(A) + P(B) (since we can just add the number of outcomes in one and the other, and divide by the number of outcomes in Ω)

Rules of probability

- Let’s think about what rules we can lay down, based on what we have seen so far.