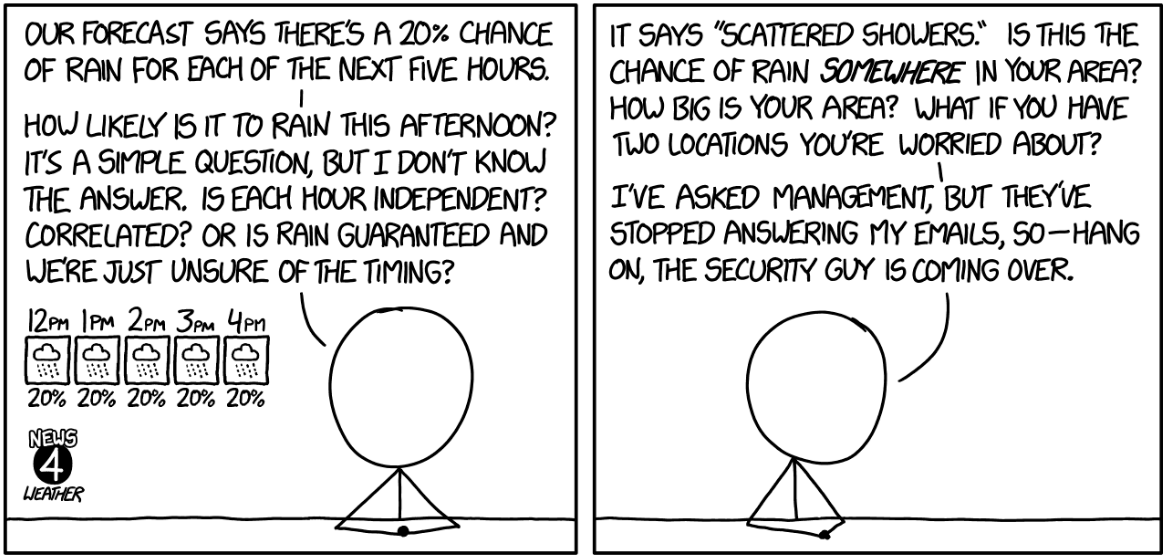
Stat 88: Probability and Mathematical Statistics in Data Science



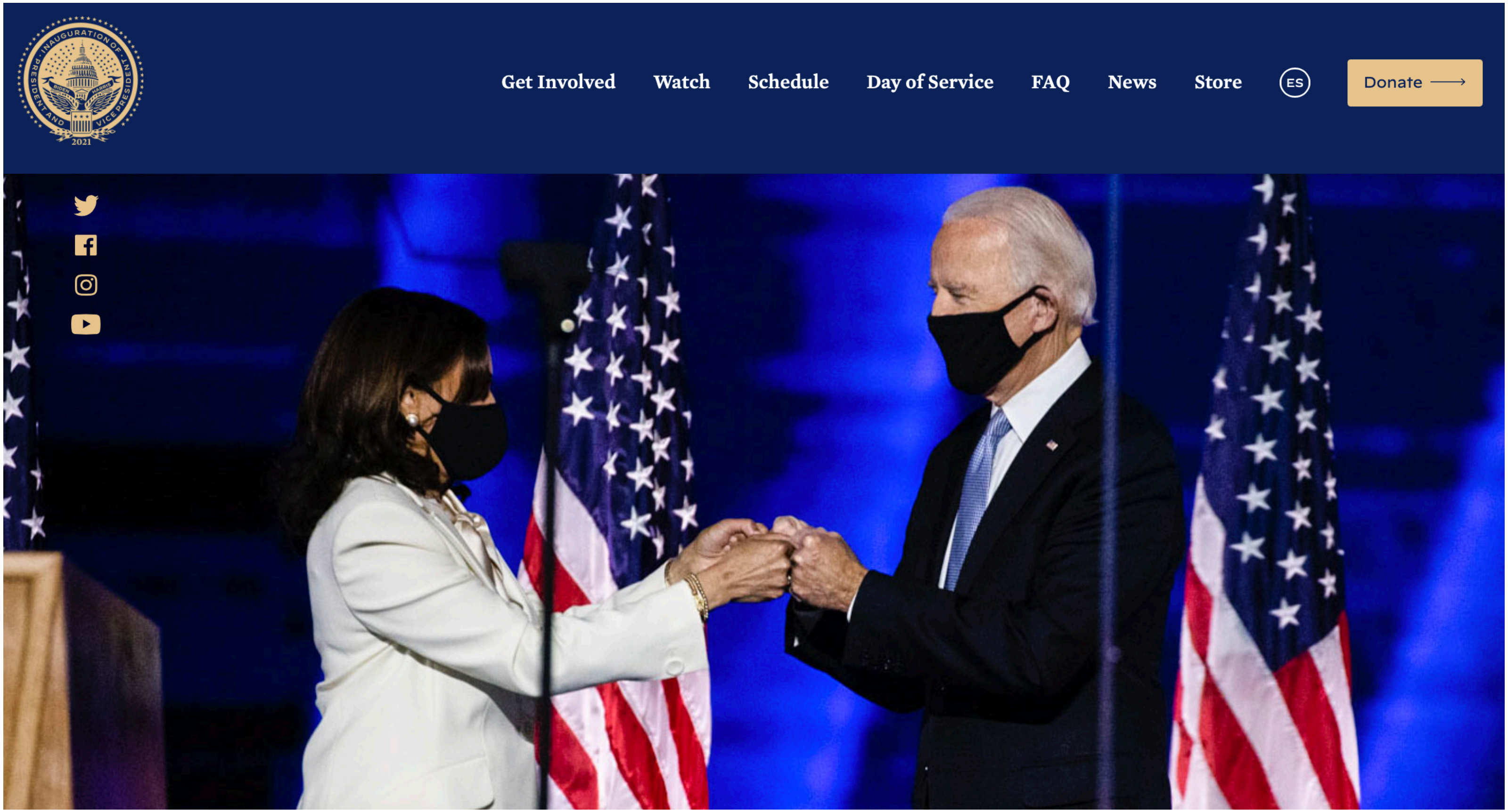
https://imgs.xkcd.com/comics/meteorologist.png

Lecture 1: 1/20/2021

Course introduction and the basics

Shobhana M. Stoyanov

were
I know many of you ~~would rather be~~ watching…



Agenda

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

- Icebreaker padlet

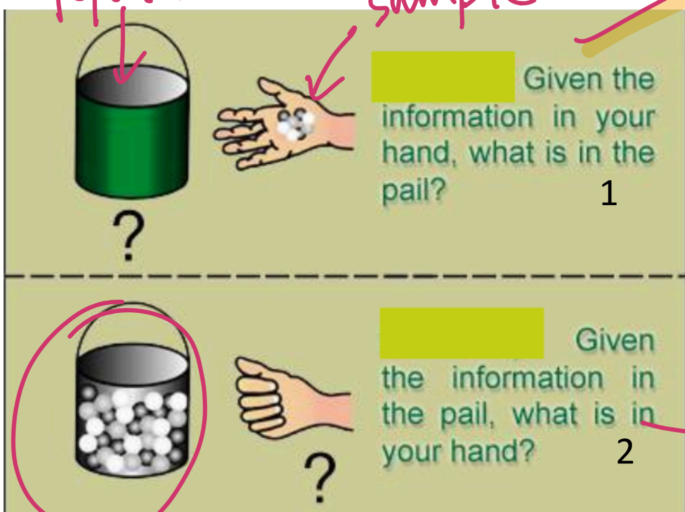
- The Basics:
  - Section 1.1: Probabilities as Proportions
  - Section 1.2: Exact Calculation or Bound

Icebreaker

- In the breakout room, introduce yourself to your classmates. How are you feeling today? You can write your thoughts on the padlet.

- Discuss which is probability and which is statistics:

population sample Statistics



sample → population

inference.

coin: $p = P(H)$ → toss 100 times 55 H. $\hat{p} = 0.55$

Probability

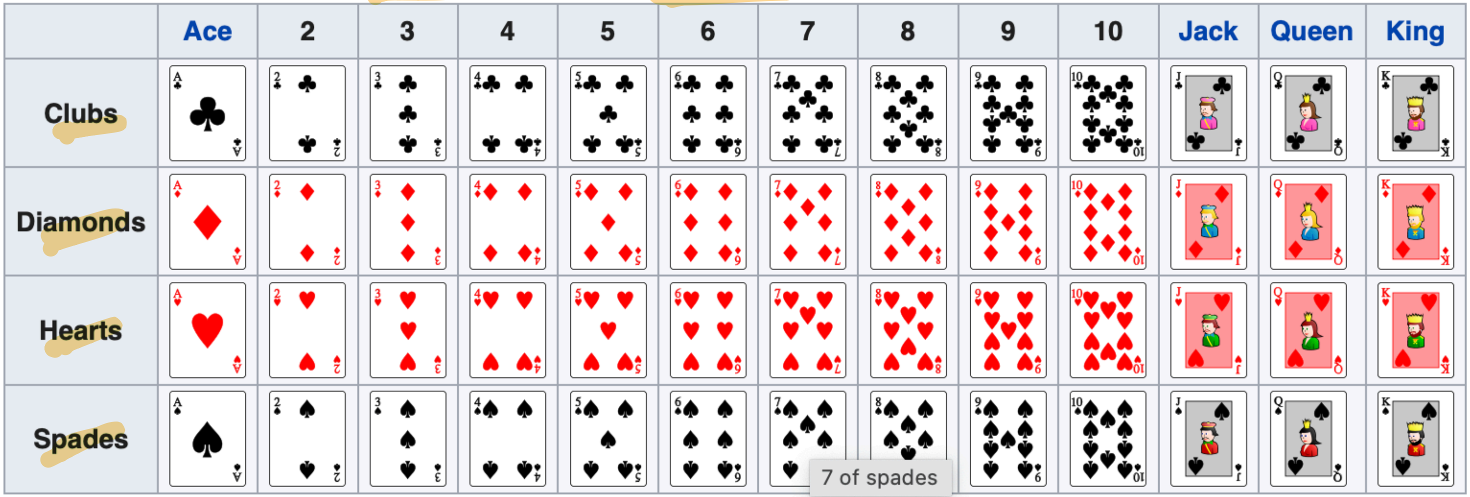
Fair coin $P(H) = 0.5$

What is the chance of 55 H in 100 tosses

100 tosses

Cards

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

1/52

Assumption: All cards are equally likely.

4/52

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

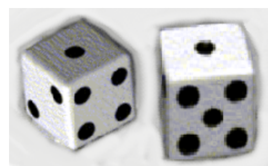
popn = deck of cards

sample = hand that is dealt.

Section 1.1: Probabilities as proportions

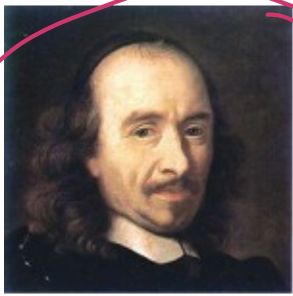
- We can think about probability as a numerical measure of uncertainty, and we will define some basic principles for computing these numbers.
- These basic computational principles have been known for a long time, and in fact, gamblers thought about these ideas a lot. Then mathematicians investigated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

De Méré's Paradox

- 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)

Examples Toss a coin once: outcomes : H, T

expt.

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*.

$\Omega = \{H, T\}$

- 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.

Expt: roll a die once $\Omega = \{1, 2, 3, 4, 5, 6\}$

Exercise ① Write down n outcomes of rolling up a pair of dice: Ω

6

$6 \times 6 = 36$

(1,1) (1,2) (1,3) - - .

② Write down Ω for expt of tossing a coin twice, three times.

Computing probabilities

$\#(\Omega) = 4$ $\#(\Omega) = 8$

- 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)?

prob of Q♡ $= \dfrac{1}{52}$ ← Q♡, ← 52 possible

prob Q $= \dfrac{4}{52}$ ← $\dfrac{\#(Q♡, Q♢, Q♣, Q♤)}{\#(\Omega)}$

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

Equally likely outcomes

- Say we roll a die. What is Ω?

$\{1, 2, 3, 4, 5, 6\}$

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

$\dfrac{\#(\{3,6\})}{\#(\Omega)} = \dfrac{2}{6}$.

Prob as a proportion: Prob is a numerical measure.

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.

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

subset

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

$A \subseteq \Omega$
contained.

For a die roll, $\Omega = \{1, 2, 3, 4, 5, 6\}$
A is the event that we roll a multiple of 3.
$A = \{3, 6\}$

$$P(A) = \frac{\#(A)}{\#(\Omega)} = \frac{2}{6}$$

Rules for P(A)

① $0 \leq P(A) \leq 1$
② add up all the probs of single outcomes, should get 1.

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)?

$\#(\Omega) = 52$ $A = \{Q♡\}$ $P(A) = \frac{1}{52}$

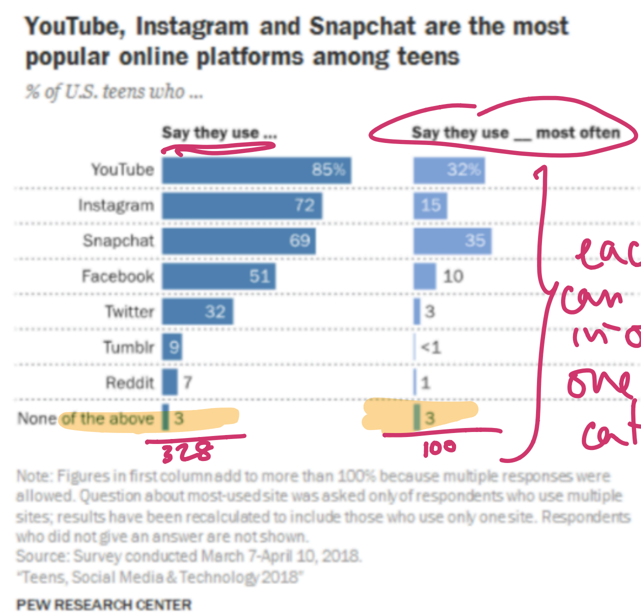
- If you deal 2 cards, what is the chance that at least *one* of them is a queen?

$A = \{(Q,Q), (Q, \text{not } Q), (\text{not } Q, Q)\}$ ← at least 1 Q.

$$P(A) = \frac{4 \cdot 3 + 4 \cdot 48 + 48 \cdot 4}{52 \cdot 51} \approx 0.149$$

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.



- What is the difference b/w 2 charts?

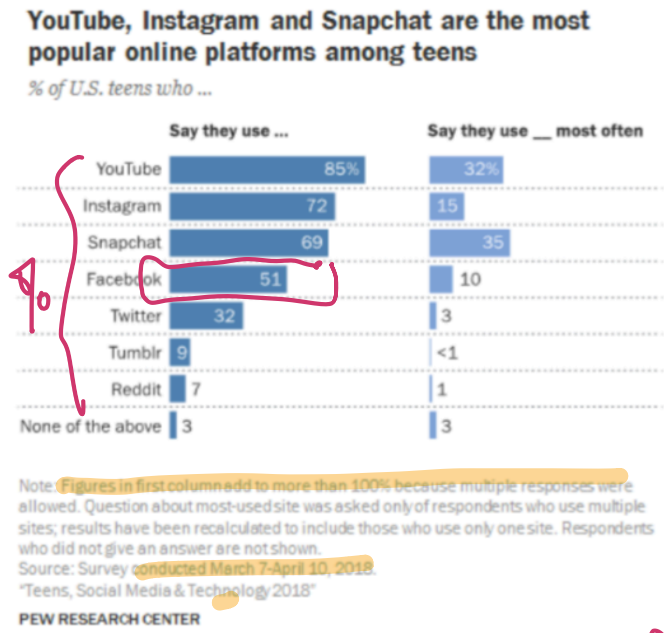
- Why do the % add up to more than 100 in the first graph?

each teen can be in only one category.

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

This is called a dsn

Not equally likely outcomes



1. What is the chance that a randomly picked teen uses FB most often? ~ 10%

2. What is the chance that a randomly picked teen did *not* use FB most often? ~ 90% = 100% − 10%

3. What is the chance that FB *or* Twitter was their favorite? 13% = (10 + 3)%

P(FB) = 51%

4. What is the chance that the teen used FB, just not most often? 51 − 10 = 41%

New outcome space = FB users

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

Venn diagram for use

Ω = all teens

Favorite