STATISTICAL FOUNDATIONS

Heike Hofmann

FROM ZERO TO PREDICTIONS

in eight hours

OUTLINE

Day 1: Probability

- Probabilities
- Random Variables
- Relationships between (two) Random Variables
- Distributions: Bernoulli & Normal Distribution
- Central Limit Theorem

Day 2: Statistics

- Estimation of means and proportions (MLE)
- Confidence Intervals
- (Normal) Linear Models
- Predictions

OBJECTIVE

- Later elements (confidence intervals & predictions) relevant to project work
- Review/discuss some of the theoretical foundations
- Connect intuitive understanding to mathematical notation

OBJECTIVE

- Later elements (confidence intervals & predictions) relevant to project work
- Review/discuss some of the theoretical foundations
- Connect intuitive understanding to mathematical notation

Among the people in this class, what is the probability that at least two of us share the same birthday (day, not year)?

FERPA doesn't want us to collect/publish private information.

Instead, let's collect our favorite number between 1 and 366

```
lubridate::yday(lubridate::ymd("2020/mm/dd"))
```

Go to menti.com and use code 95 61 49 & submit your favorite number

THE PLAN

- Material chunked into
  - conceptual overview
  - example
  - activity

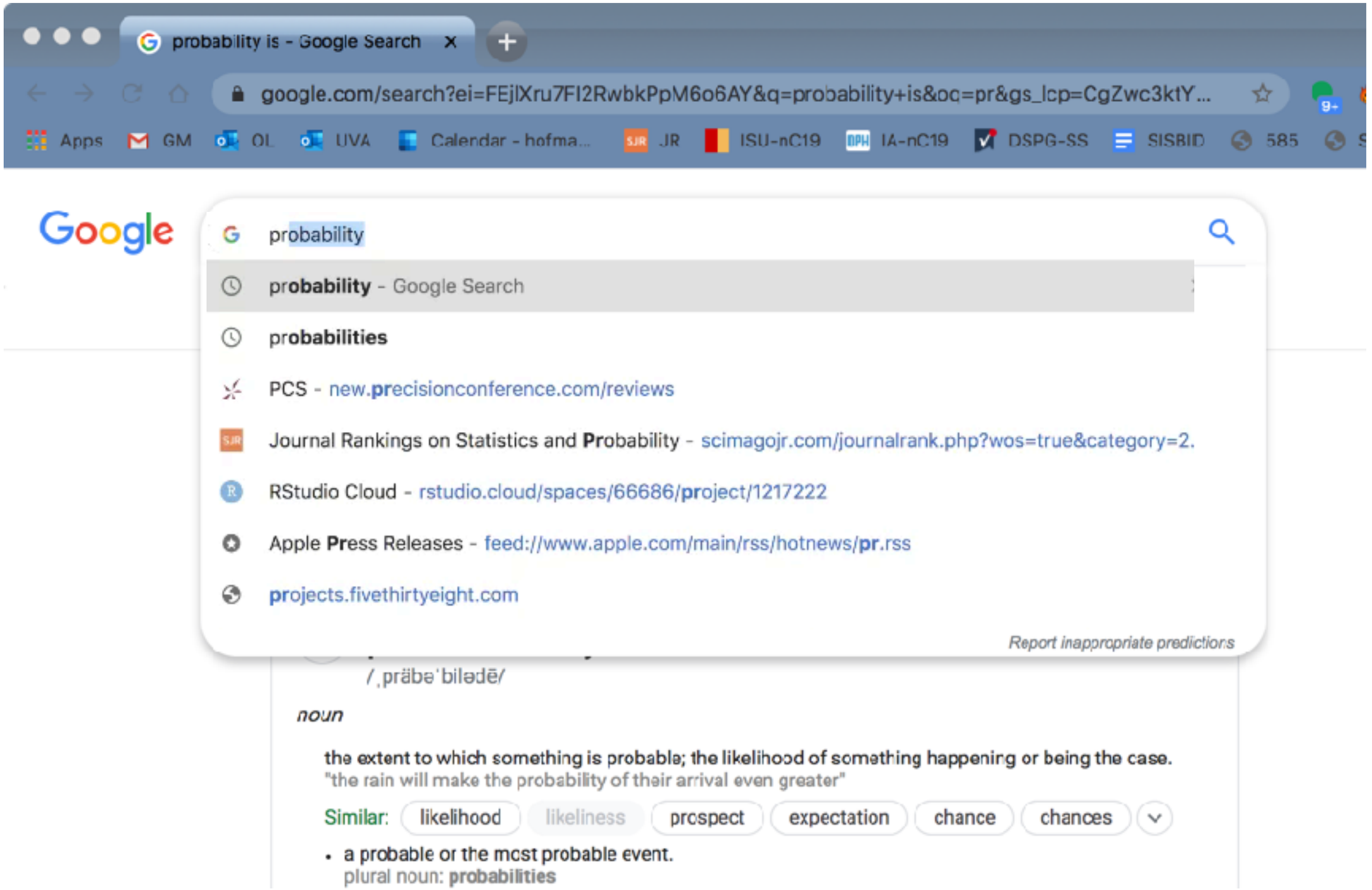
- All materials (pdf, code) are available at https://github.com/DSPG-ISU/Training

DAY 1

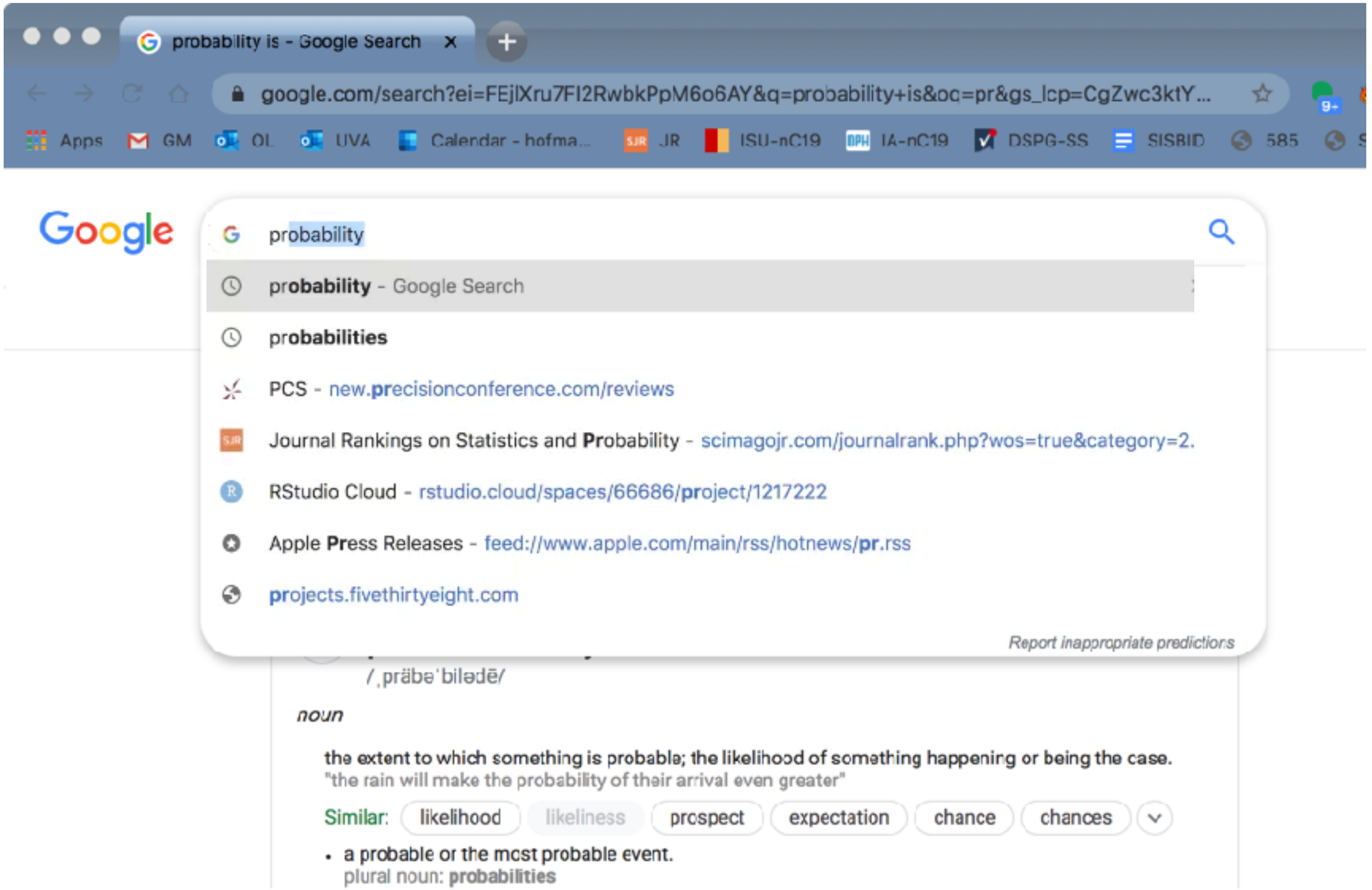
- Probabilities
- Random Variables
- Bernoulli Distribution
- Relationships between (two) Random Variables
- Normal Distribution
- Central Limit Theorem

PROBABILITIES

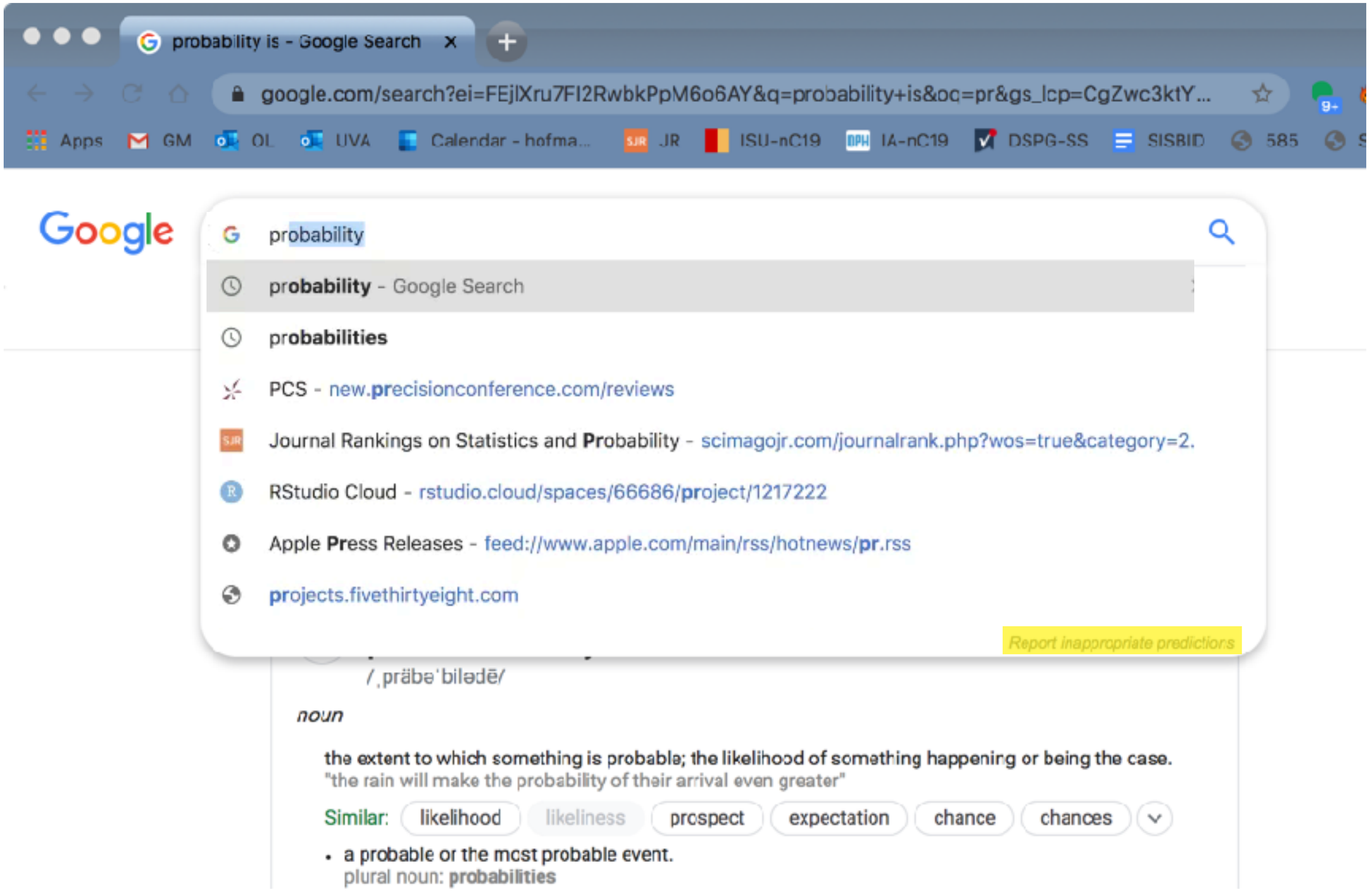
PROBABILITY IS . . .



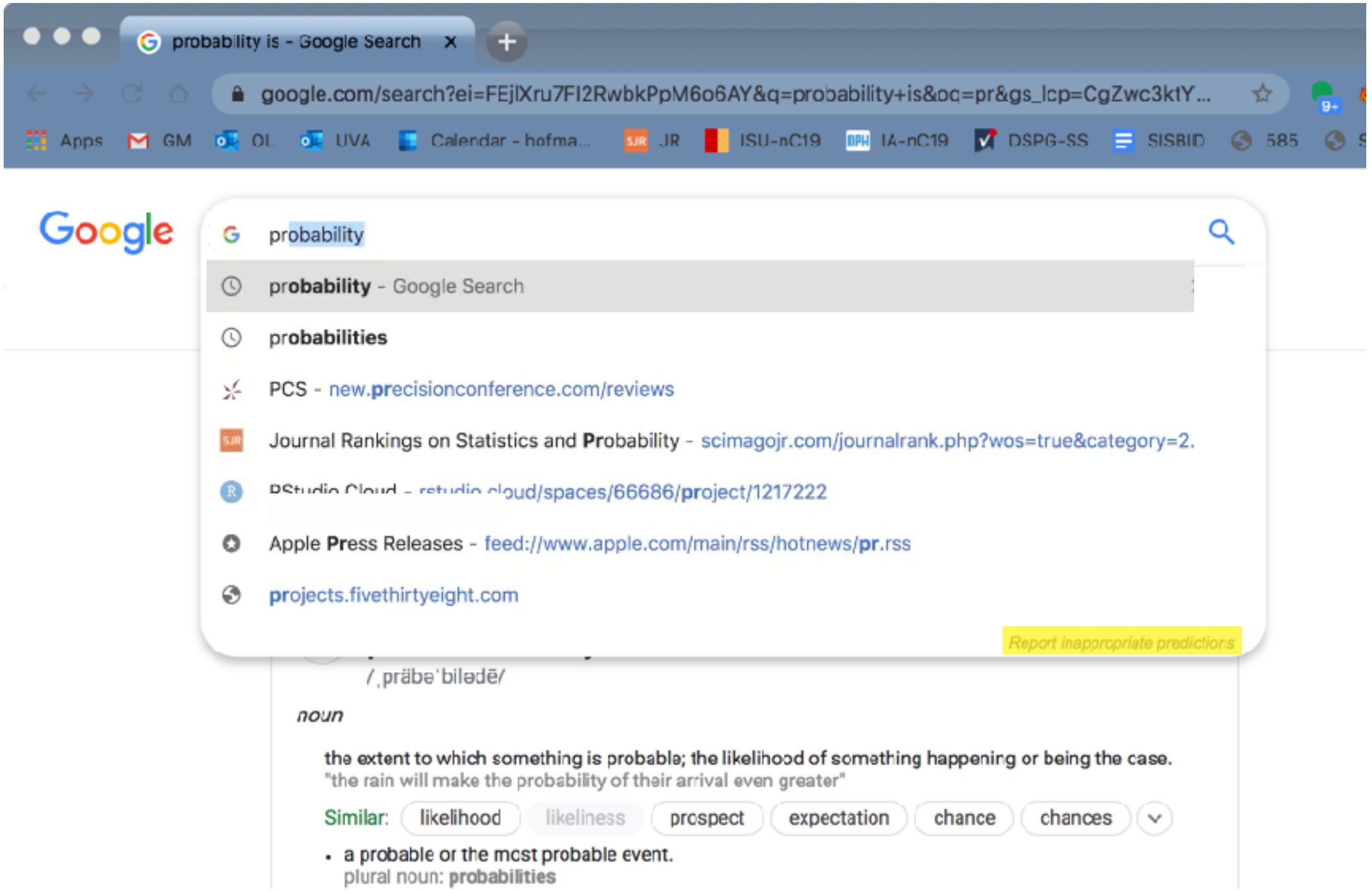
PROBABILITY IS . . .



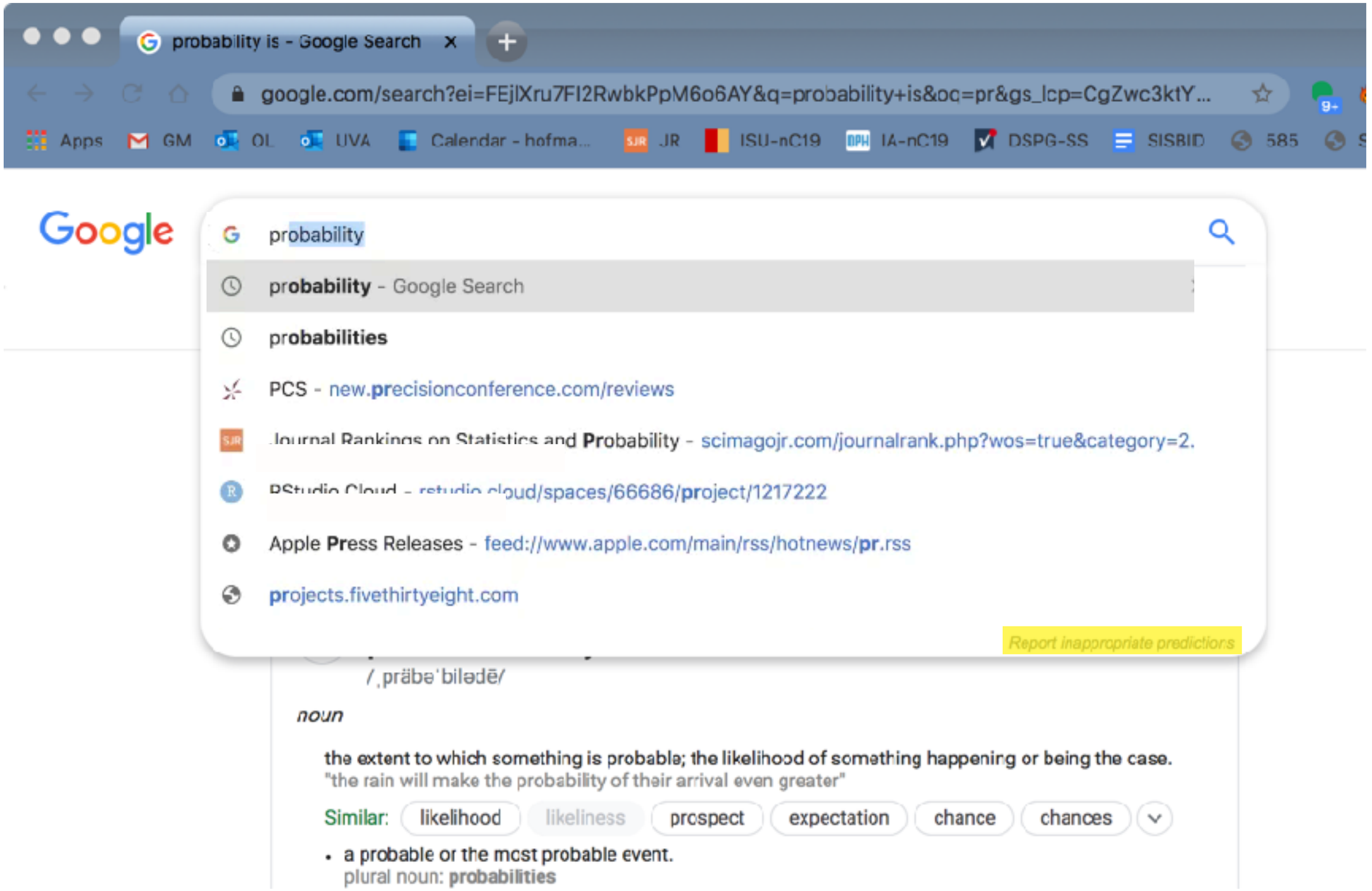
PROBABILITY IS . . .



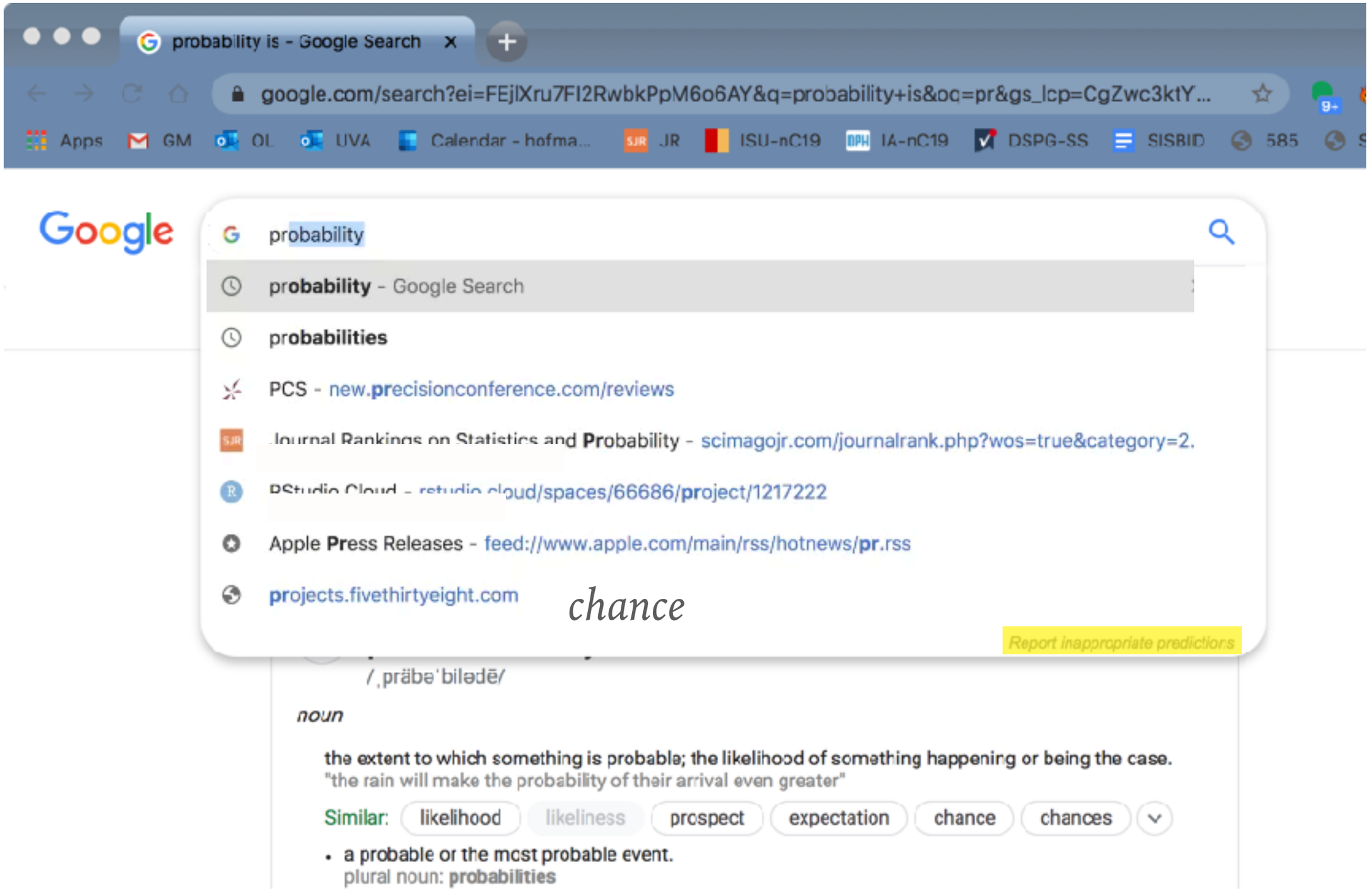
PROBABILITY IS . . .



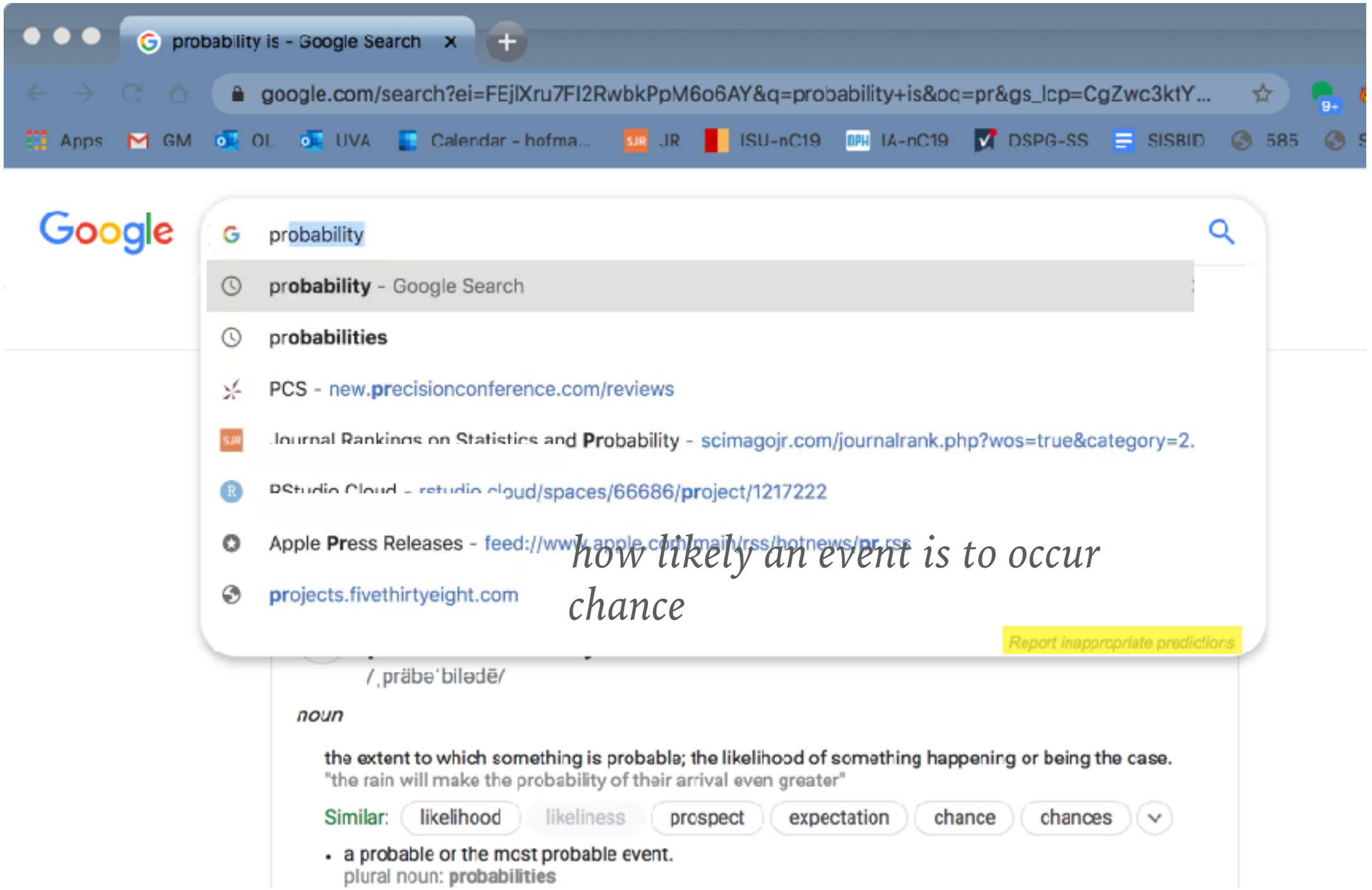
PROBABILITY IS . . .



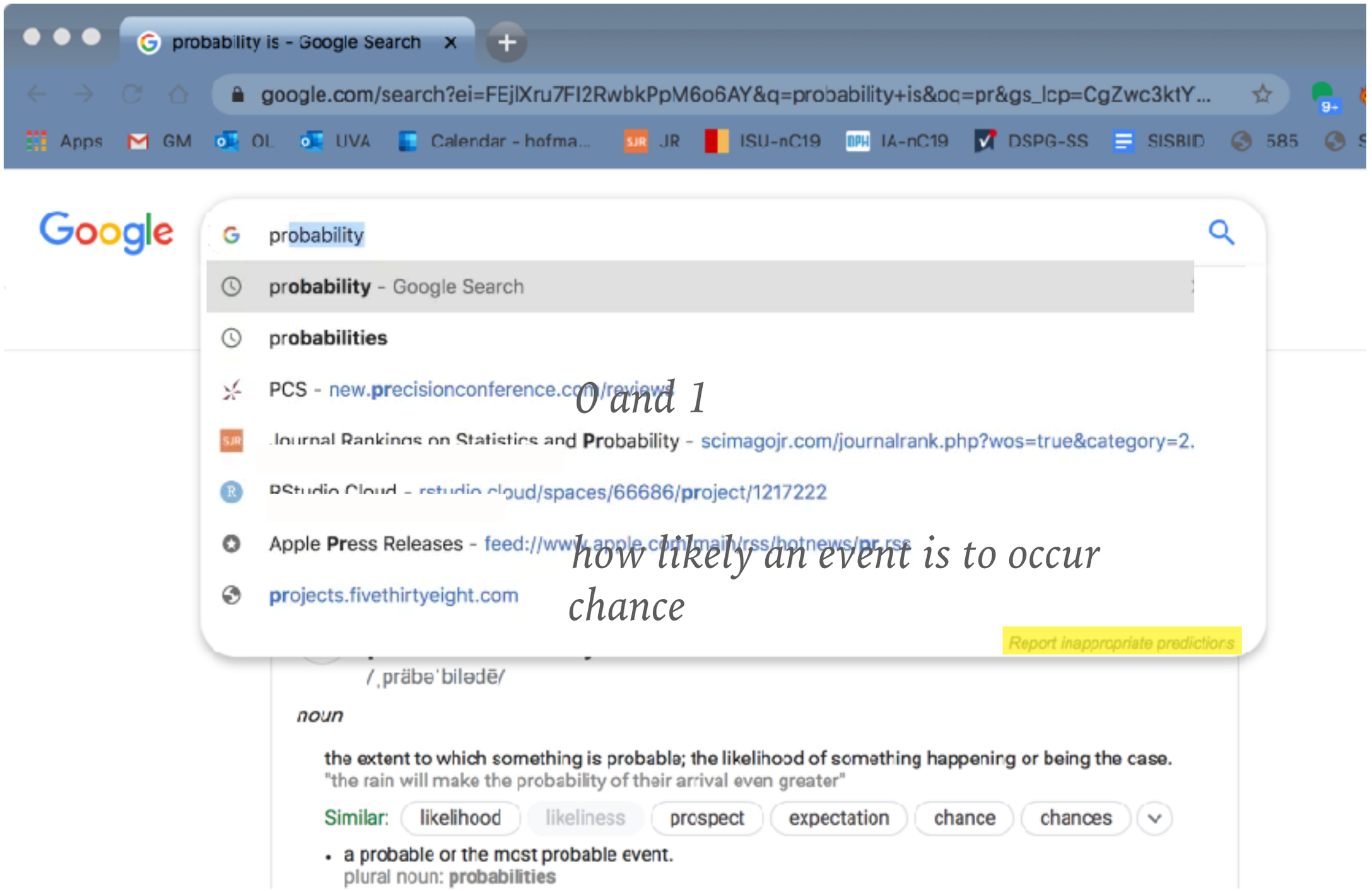
PROBABILITY IS . . .



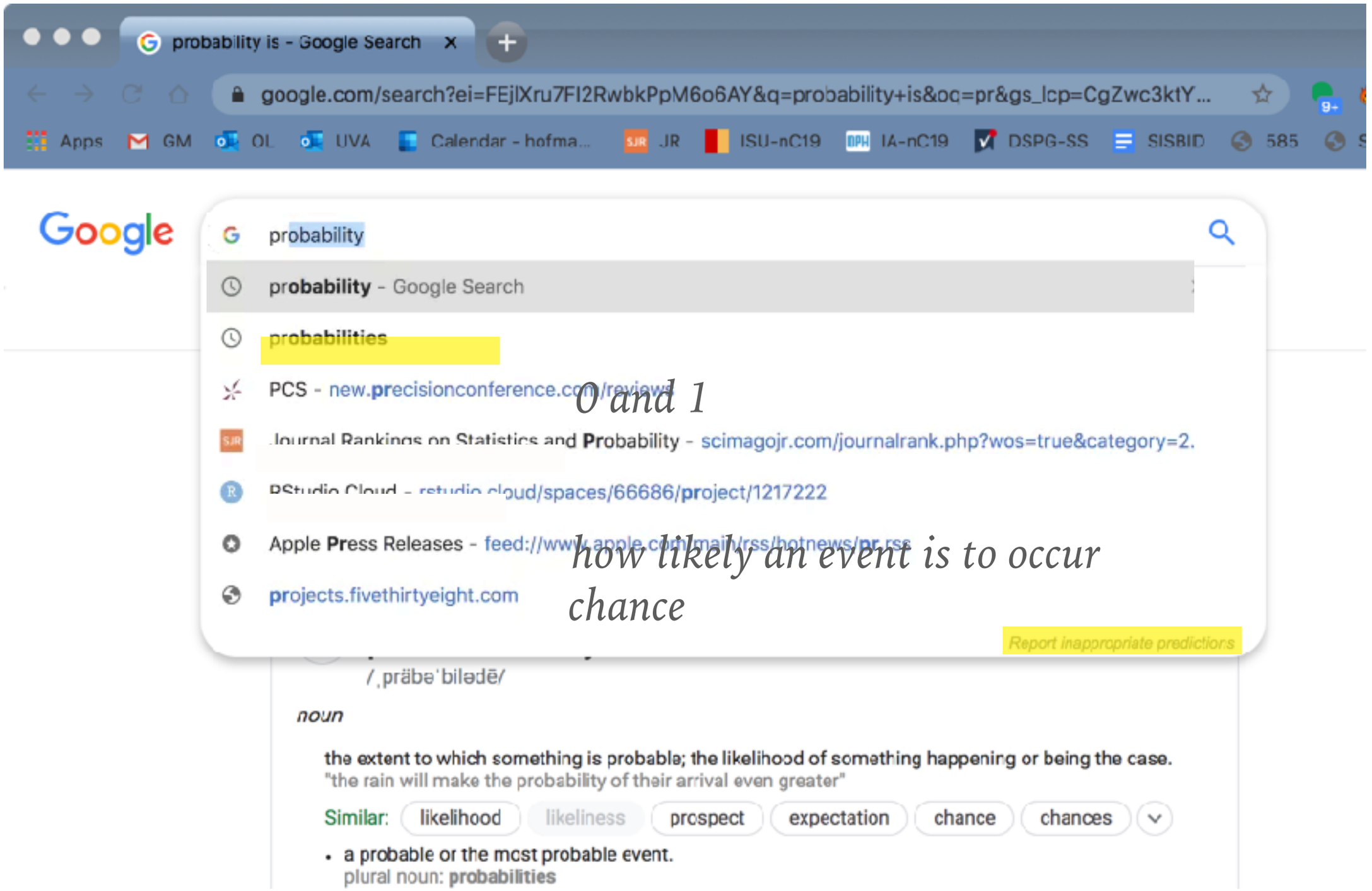
PROBABILITY IS . . .



PROBABILITY IS . . .



PROBABILITY IS . . .



DEFINING PROBABILITY: SAMPLE SPACE & EVENT

DEFINING PROBABILITY: SAMPLE SPACE & EVENT

- A **Random Experiment**
  is a situation/phenomenon without deterministic outcome, but the *set of all possible outcomes* is known.

DEFINING PROBABILITY: SAMPLE SPACE & EVENT

- A **Random Experiment**
  is a situation/phenomenon without deterministic outcome, but the *set of all possible outcomes* is known.

- This set of all possible outcomes is known as the **sample space** of the experiment and is denoted by **Ω**.

DEFINING PROBABILITY: SAMPLE SPACE & EVENT

- A **Random Experiment**
  is a situation/phenomenon without deterministic outcome, but the *set of all possible outcomes* is known.

- This set of all possible outcomes is known as the **sample space** of the experiment and is denoted by **Ω**.

- An **event E**
  is a set of possible outcomes of the experiment (E is a subset of Ω).
  If the outcome of the random experiment is contained in E, we say that E has occurred.

ROLLING A DIE

- Roll a six-sided die

  Define outcome as number on the face on the top once the die has come to a stop

  This is a Random Experiment

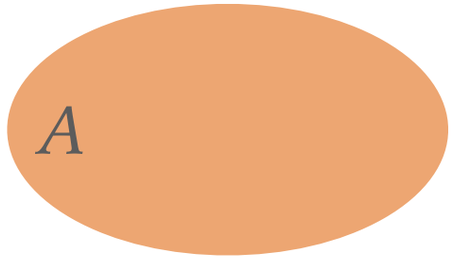
- Sample Space
  $\Omega = \{1, 2, 3, 4, 5, 6\}$

- Possible events?
  Roll a six: $E = \{6\}$
  Roll an odd number:
  $E = \{1,3,5\}$
  Roll a number higher than 2:
  $E = \{3,4,5,6\}$

SET NOTATION

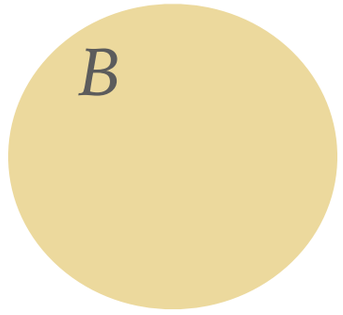
SET NOTATION

- Given are events A and B



$A \subset \Omega$

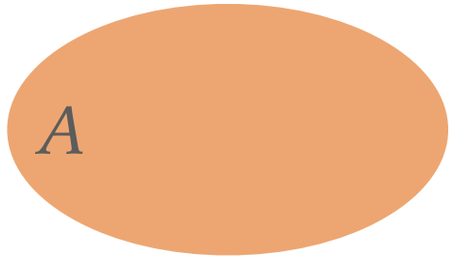
A "is a subset of" Ω



$B \subset \Omega$

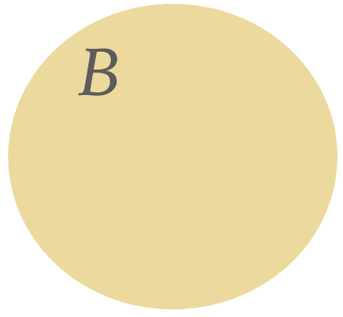
SET NOTATION

- Given are events A and B



$A \subset \Omega$

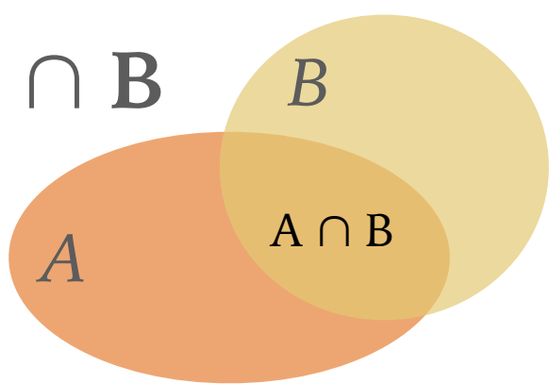
A "is a subset of" Ω



$B \subset \Omega$

- Relationship between A and B:

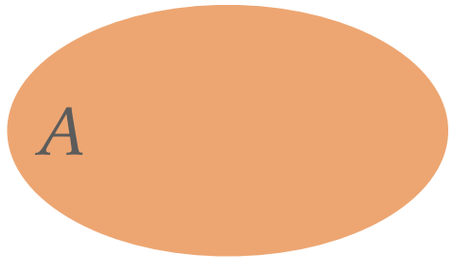
- **Intersection** of **A and B** is **A ∩ B**
  set of all elements that are
  both in A and in B



logical "and"

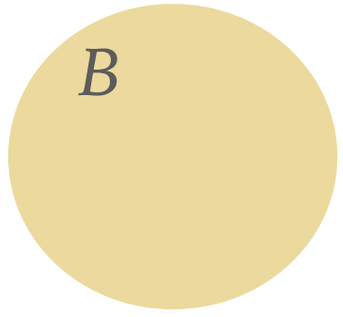
SET NOTATION

- Given are events A and B



$A \subset \Omega$

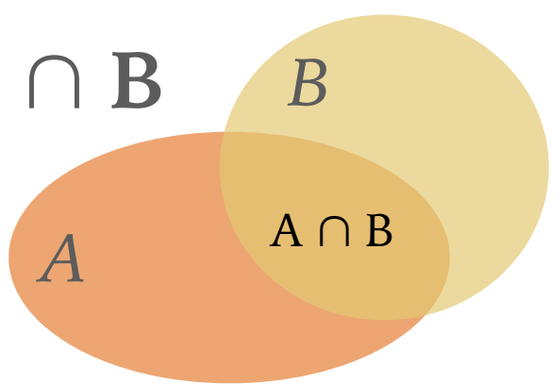
A "is a subset of" Ω



$B \subset \Omega$

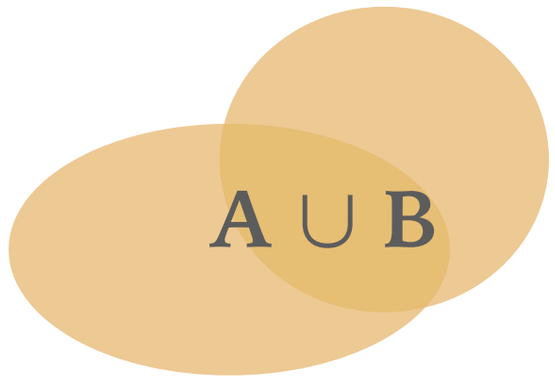
- Relationship between A and B:

- **Intersection** of **A and B** is **A ∩ B**
  set of all elements that are
  both in A and in B



logical "and"

- **Union** of **A or B** is **A ∪ B**
  set of elements that are
  in A, in B or in both

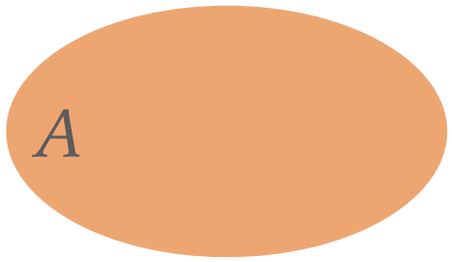


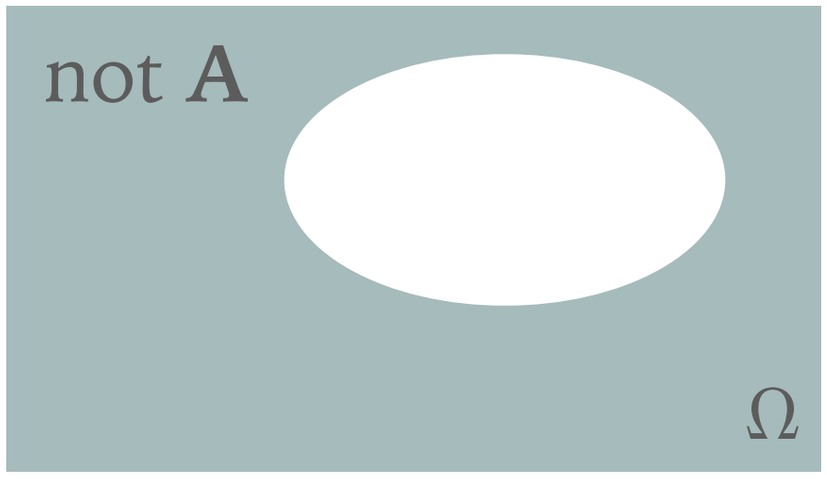
logical "or"

SET NOTATION

SET NOTATION

- A and not A:



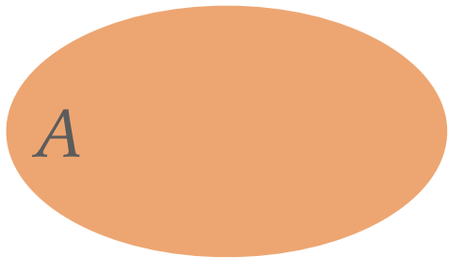


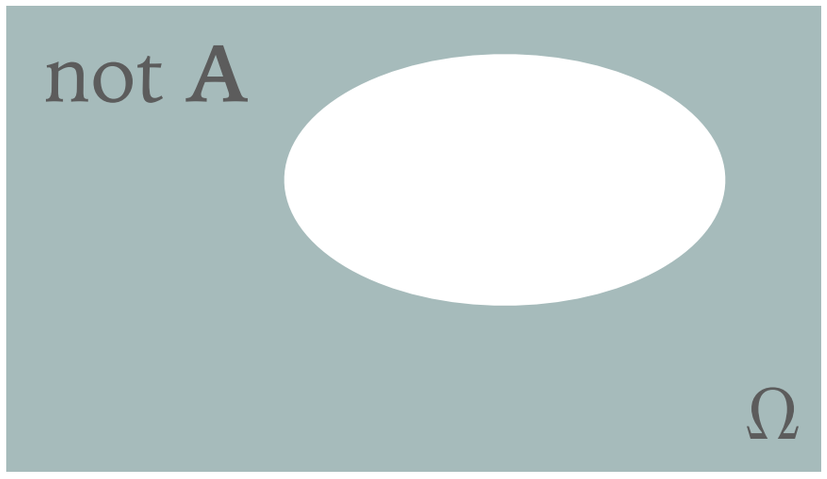
not A is called the **complement of A**

Notation: !A, ¬A

SET NOTATION

- A and not A:



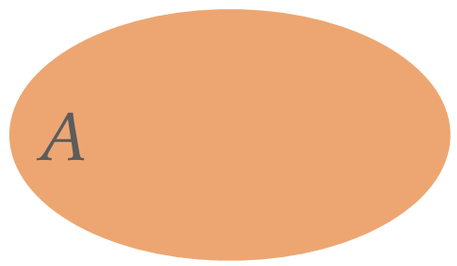


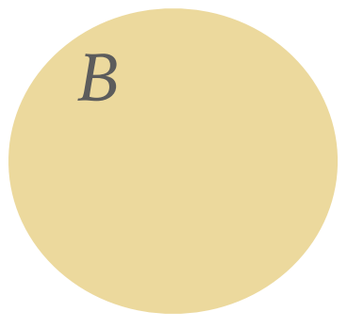
not A is called the **complement of A**

Notation: !A, ¬A

- A and B are disjoint if they do not have any elements in common

  A ∩ B = ø

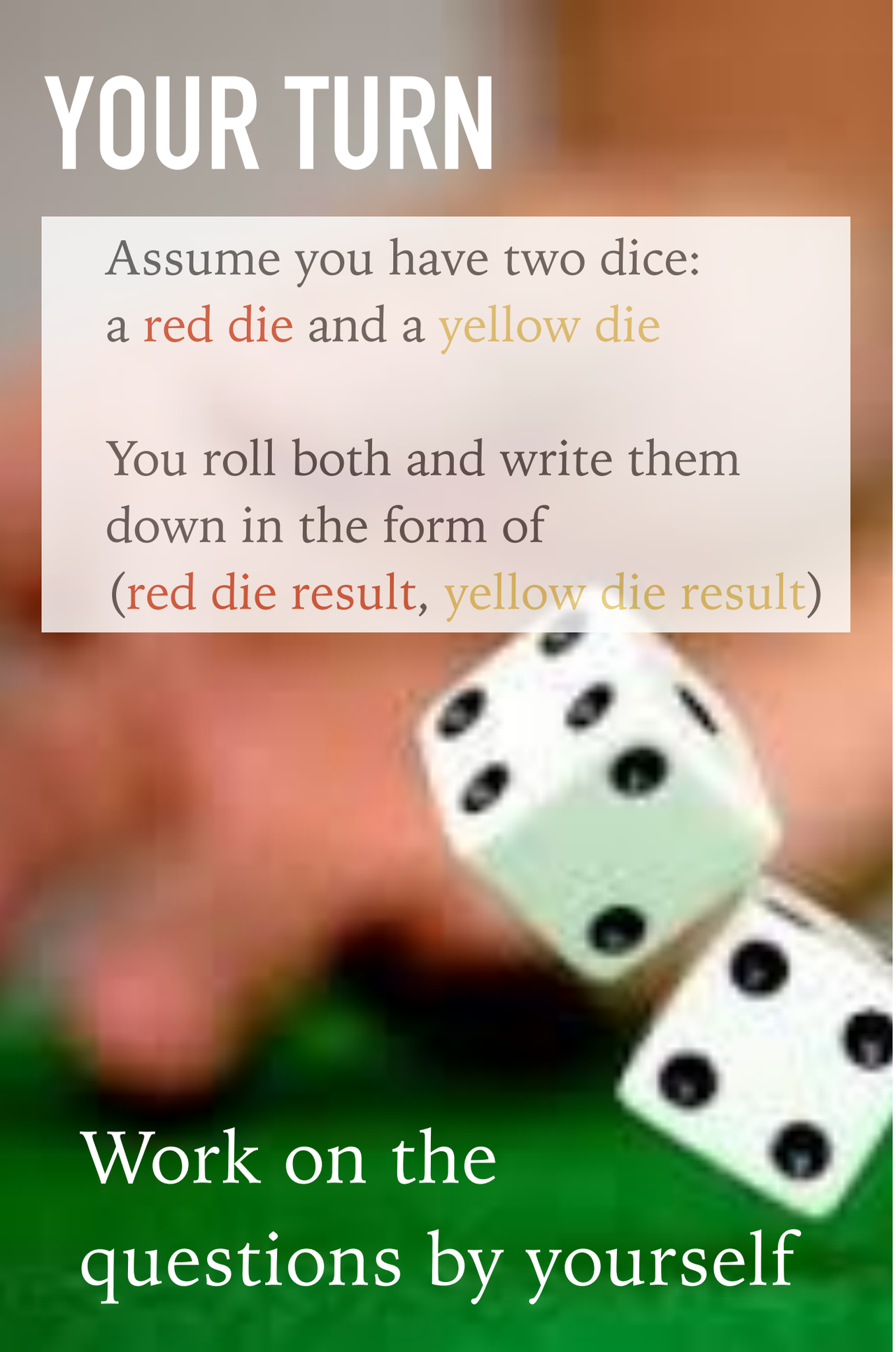




YOUR TURN

Assume you have two dice:
a red die and a yellow die

You roll both and write them down in the form of
(red die result, yellow die result)



Work on the questions by yourself

ROLLING TWO DICE

- Write out the Sample Space Ω
- A is the event that the sum of the two dice is 7
  B is the event that both dice show the same face
  C is the event that the red die is 2.
  Write the events as subsets of Ω
- What is the intersection of B and C?
  Are A and C disjoint?
- What can you say about the intersection of A, B, and C?

YOUR TURN: SOLUTION

- Sample space of rolling two (distinguishable dice):

$$\Omega = \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6),$$
$$(2,1), (2,2), (2,3), (2,4), (2,5), (2,6),$$
$$(3,1), (3,2), (3,3), (3,4), (3,5), (3,6),$$
$$(4,1), (4,2), (4,3), (4,4), (4,5), (4,6),$$
$$(5,1), (5,2), (5,3), (5,4), (5,5), (5,6),$$
$$(6,1), (6,2), (6,3), (6,4), (6,5), (6,6) \} =$$

$$= \{(i, j) \mid 1 \leq i \leq 6, 1 \leq j \leq 6\}$$

YOUR TURN: SOLUTION

- A is the event that the sum of the two dice is 7
  A = {(1,6), (2,5),(3,4),(4,3), (5,2), (6,1)}
- B is the event that both dice show the same face
  B = {(1,1), (2,2), (3,3), (4,4), (5,5), (6,6) }
- C is the event that the red die shows a 2.
  C = { (2,1), (2,2), (2,3), (2,4), (2,5), (2,6)}

- $B \cap C = \{(2,2)\}$
  $A \cap C = \{(2,5)\}$; A and C are not disjoint
  $A \cap B \cap C = \{\} = \emptyset$ disjoint

DEFINING PROBABILITY

- (1) $0 \leq P(A) \leq 1$
  for all events A
- (2) $P(\Omega) = 1$.
- (3) For disjoint events
  A and B:
  $P(A \cup B) = P(A) + P(B)$

Kolmogorov (1933)

SOME PROPERTIES

- $P(\varnothing) = 0$

  proof: $1 = P(\Omega) = P(\Omega \cup \varnothing) = P(\Omega) + P(\varnothing) = 1 + P(\varnothing)$

  (Ω and $\varnothing$ are disjoint)

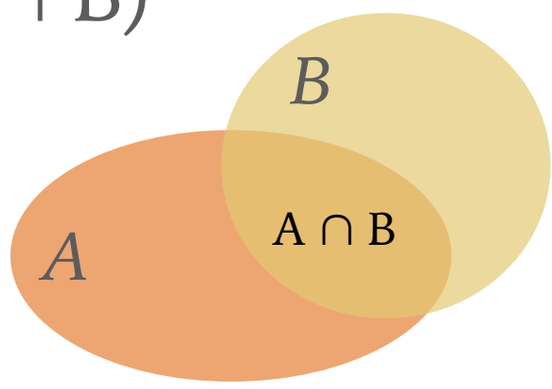
- $P(\neg A) = 1 - P(A)$

  proof: $1 = P(\Omega) = P(A \cup \neg A) = P(A) + P(\neg A)$

  (A and $\neg$A are disjoint)

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

  subtract $P(A \cap B)$
  to only count it once



HOW DO WE ASSIGN PROBABILITIES?

HOW DO WE ASSIGN PROBABILITIES?

- Assign a probability to each element in Ω
  Sometimes we know (from theoretical model),
  sometimes we estimate (from observed data)

HOW DO WE ASSIGN PROBABILITIES?

- Assign a probability to each element in Ω
  Sometimes we know (from theoretical model),
  sometimes we estimate (from observed data)

- Roll of a Fair Die
  $\Omega = \{1,2,3,4,5,6\}$
  Die is fair: all sides come up equally
  $P(\{1\}) = P(\{2\}) = \ldots = P(\{6\}) = 1/6$

HOW DO WE ASSIGN PROBABILITIES?

- Assign a probability to each element in Ω
  Sometimes we know (from theoretical model),
  sometimes we estimate (from observed data)

- Roll of a Fair Die
  $\Omega = \{1,2,3,4,5,6\}$
  Die is fair: all sides come up equally
  $P(\{1\}) = P(\{2\}) = \ldots = P(\{6\}) = 1/6$

- National Pet Owner Survey
  Number of U.S. Households that Own a Pet (in millions)

Bird	5.7
Cat	42.7
Dog	63.4
Horse	1.6

Freshwater	11.5
Saltwater Fish	1.6
Reptile	4.5
Small Animal	5.4

YOUR TURN

A box contains 8 identical looking AA batteries.
One of the batteries is dead.

Work on the questions by yourself

ENERGY?

- The energizer bunny needs a new battery.
- Assuming that you randomly pick a battery and insert it. What is the probability that the bunny starts up?
- How does this probability change if the bunny needs two batteries?

YOUR TURN: SOLUTION

- Let's define $\Omega = \{B1, B2, \ldots, B8\}$ and assume that B8 is defect We have a 1/8 chance to pick B8 at random, i.e. the bunny works with probability 0.875.

YOUR TURN: SOLUTION

- Two picks:
  There's various ways to define a sample space for two picks, maybe the easiest one to work with is when we enumerate the full list of possibilities:
  Ω = { (B1, B2), (B1, B3), … (B2, B1), (B2, B3), … (B8, B7)}
- All elements in Ω have the same probability: 1/56
- P(Bunny works) = P(B8 not picked) = (7*6)/56 = 6/7

YOUR TURN: SOLUTION

- Two picks:
  There's various ways to define a sample space for two picks, maybe the easiest one to work with is when we enumerate the full list of possibilities:
  Ω = { (B1, B2), (B1, B3), … (B2, B1), (B2, B3), … (B8, B7)}

- All elements in Ω have the same probability: 1/56

- P(Bunny works) = P(B8 not picked) = (7*6)/56 = 6/7

Generally:
if all elements in the sample space are equally likely to be selected,

the probability for an event E is proportional to the frequency of the elements in E

$$P(E) = |E|/|\Omega|$$

SAMPLING - CODE

- R function
  `sample(x, size, replace = FALSE, prob = NULL)`
- from `x` number of elements
  draw `size` many

Roll six sided die twice:

```
> sample(6, 2, replace = TRUE)
[1] 1 5
```

Roll six sided die seven times:

```
> sample(6, 7, replace = TRUE)
[1] 5 6 5 1 4 4 3
```

YOUR TURN

A lot of board games need several dice for playing. Analyzing best strategies theoretically is complicated

- Risk is a board game for world domination played with dice.
- When a player tries to take over a region from another player, she will use a number of dice (usually 3) and roll.
- The defender also rolls a number of dice (also usually 3).
- The two sets of dice are lined up from highest to lowest.
- The defender loses a die if it is lower than the attacker's, otherwise the attacker loses

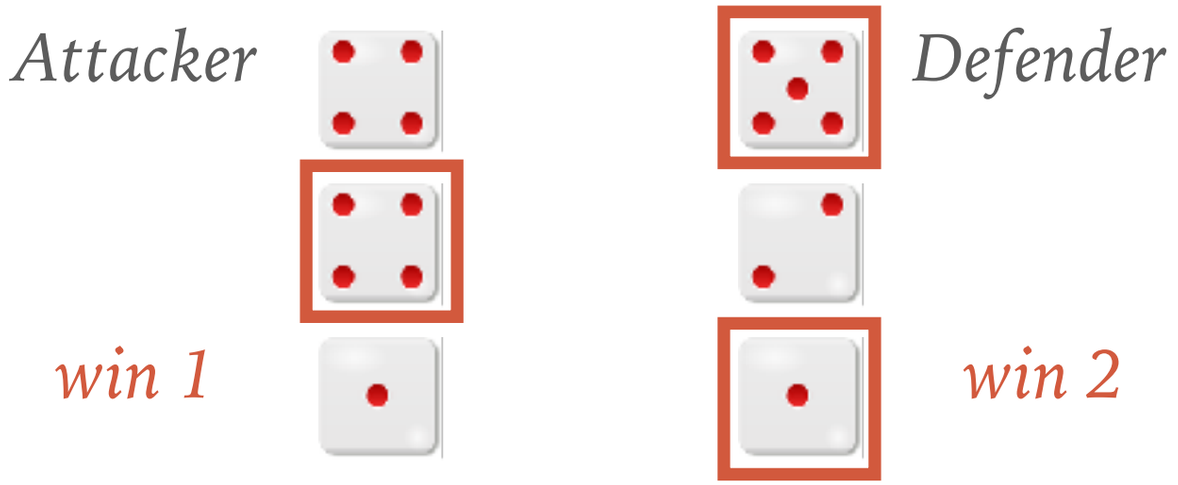
Attacker

Defender

YOUR TURN

A lot of board games need several dice for playing. Analyzing best strategies theoretically is complicated

- Risk is a board game for world domination played with dice.
- When a player tries to take over a region from another player, she will use a number of dice (usually 3) and roll.
- The defender also rolls a number of dice (also usually 3).
- The two sets of dice are lined up from highest to lowest.
- The defender loses a die if it is lower than the attacker's, otherwise the attacker loses

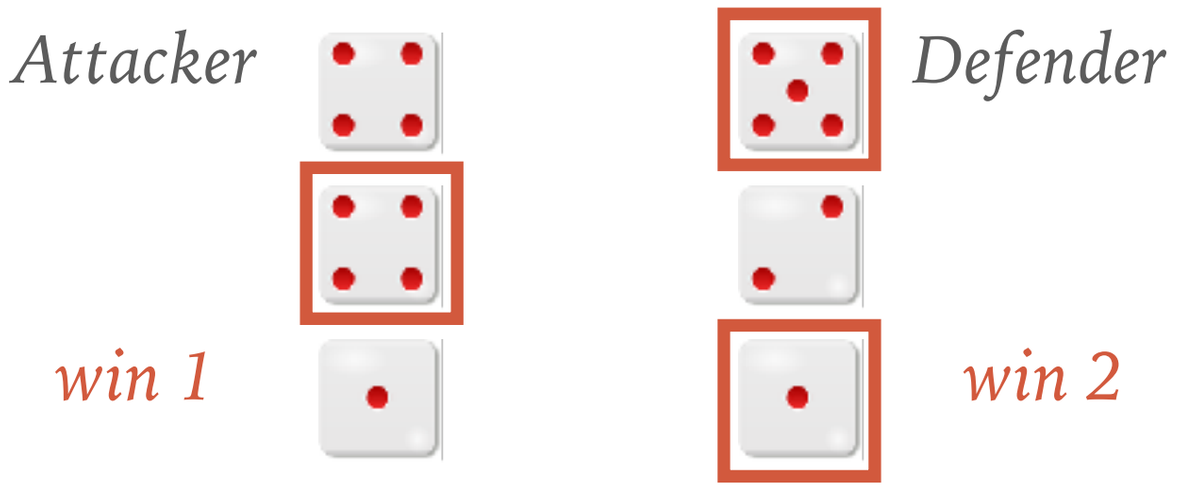


YOUR TURN

A lot of board games need several dice for playing. Analyzing best strategies theoretically is complicated

Work on the questions by yourself

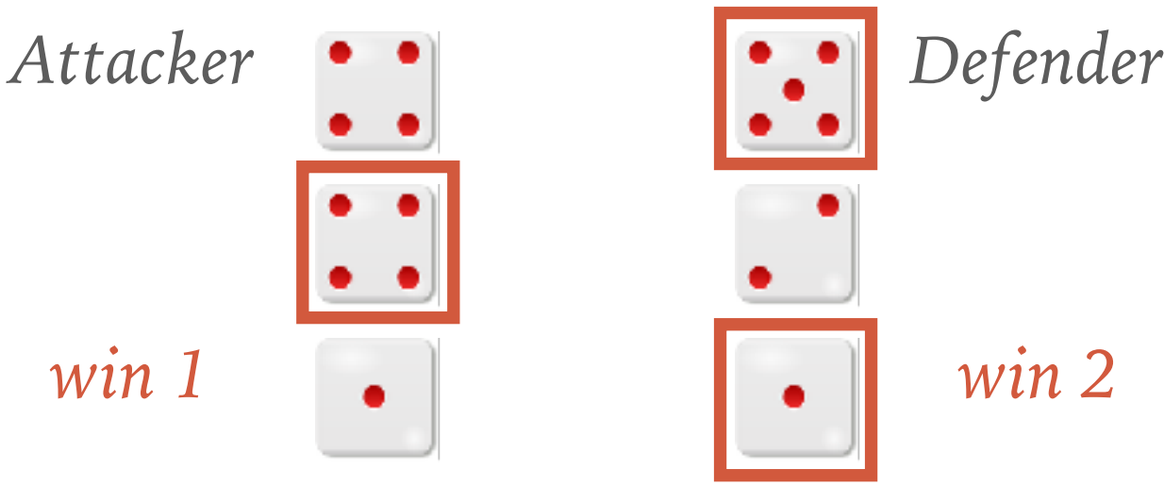
- Risk is a board game for world domination played with dice.
- When a player tries to take over a region from another player, she will use a number of dice (usually 3) and roll.
- The defender also rolls a number of dice (also usually 3).
- The two sets of dice are lined up from highest to lowest.
- The defender loses a die if it is lower than the attacker's, otherwise the attacker loses



YOUR TURN

- Risk is a board game for world domination played with dice.
- When a player tries to take over a region from another player, she will use a number of dice (usually 3) and roll.
- The defender also rolls a number of dice (also usually 3).
- The two sets of dice are lined up from highest to lowest.
- The defender loses a die if it is lower than the attacker's, otherwise the attacker loses

Work on the questions by yourself



- Based on the R function `sample` write a function `attack3()` that simulates an attacks and returns the difference in wins and losses for the attacker (-1 in the above example)
- Use `replicate` or `rerun` (`purrr`) to simulate 100 attacks. What is the probability that the attacker loses more dice than the defender?

YOUR TURN: SOLUTION

```
attack3 <- function() {
  attacker <- sort(sample(6, 3, replace = TRUE), decreasing = TRUE)
  defender <- sort(sample(6, 3, replace = TRUE), decreasing = TRUE)
  wins <- sum(attacker > defender)
  losses <- sum(attacker <= defender)
  wins - losses
}
```

To make random results reproducible, we can set the seed

Here, `set.seed(20200610)` *was used*

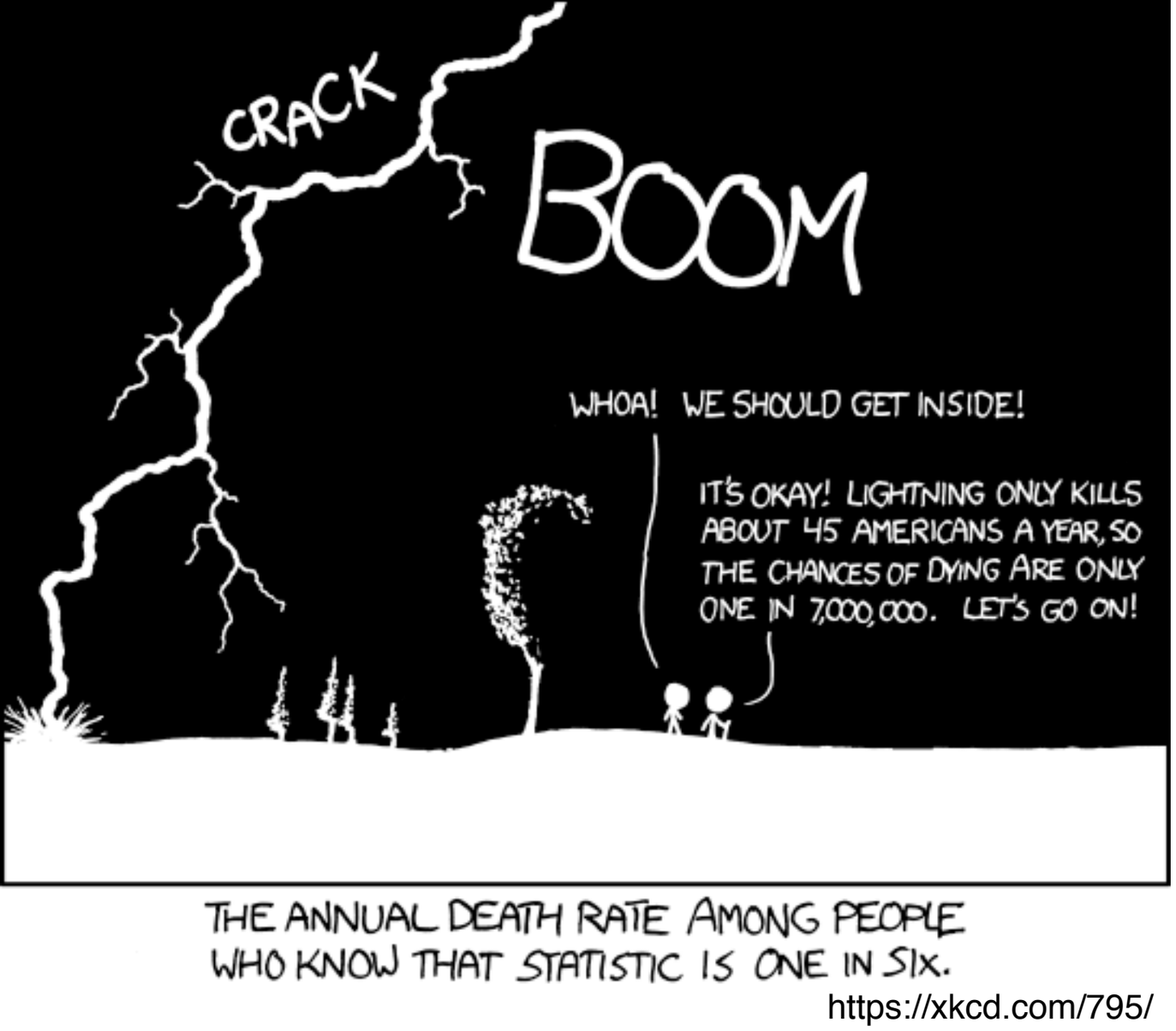
```
n <- 100

attacks <- replicate(n, attack3())
table(attacks)/n
attacks
  -3   -1    1    3
0.34 0.34 0.21 0.11
```

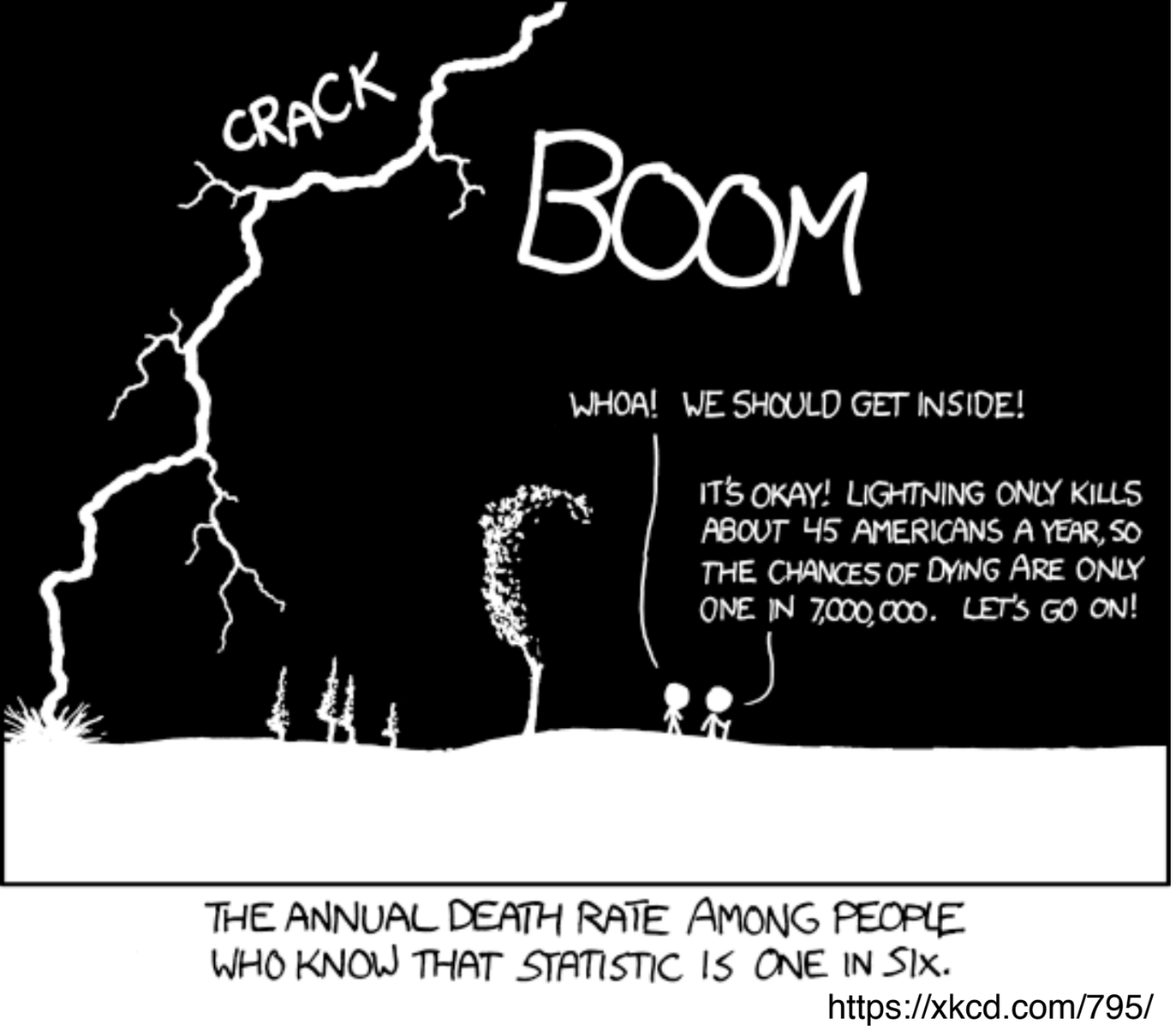
```
attacks <- rerun(n, attack3())
table(unlist(attacks))/n
 -3   -1    1    3
0.34 0.34 0.21 0.11
```

$P(\text{A loses more dice than D}) = 0.68 = 68\%$

CONDITIONAL PROBABILITY



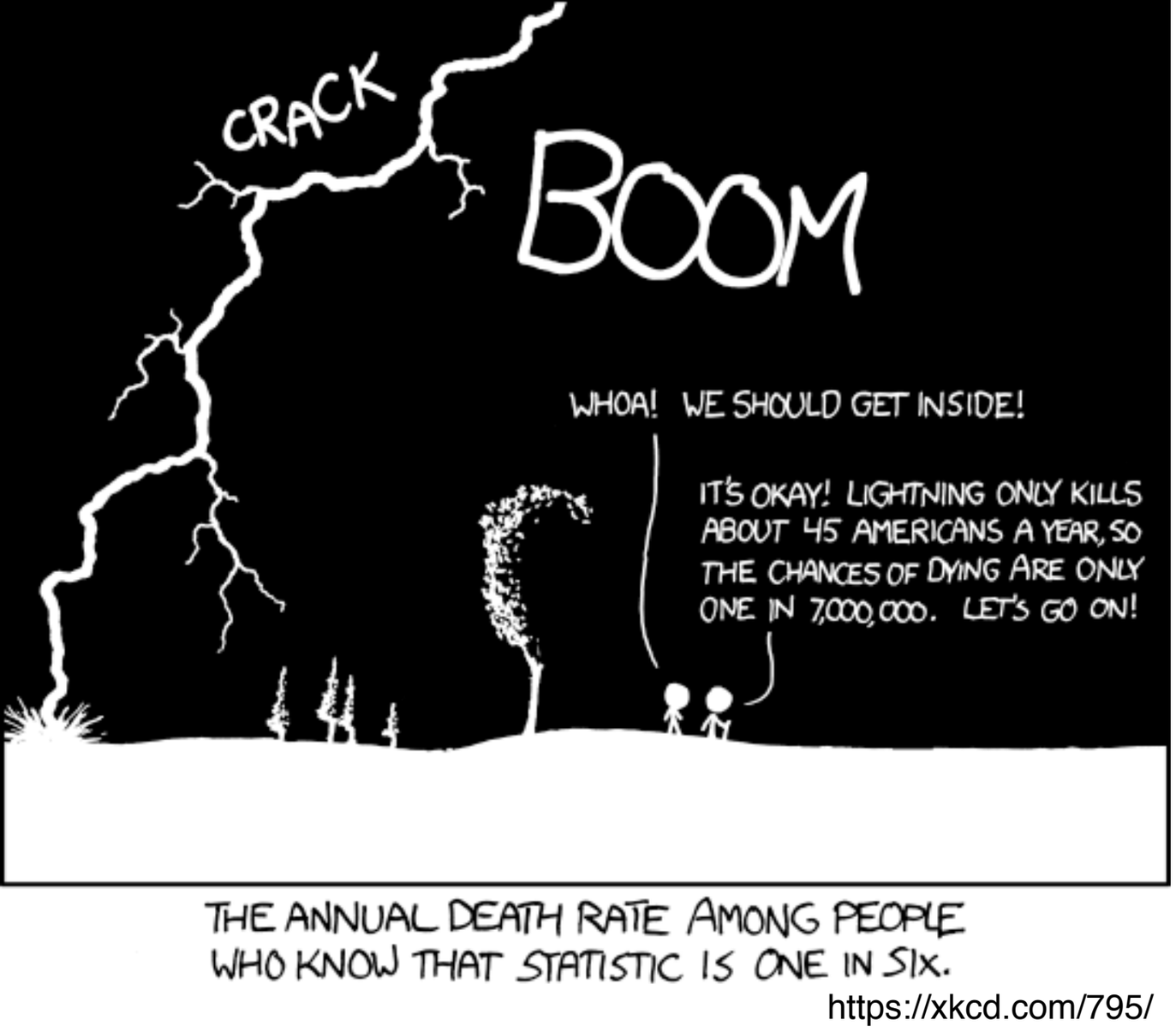
https://xkcd.com/795/



https://xkcd.com/795/

CONDITIONAL PROBABILITY

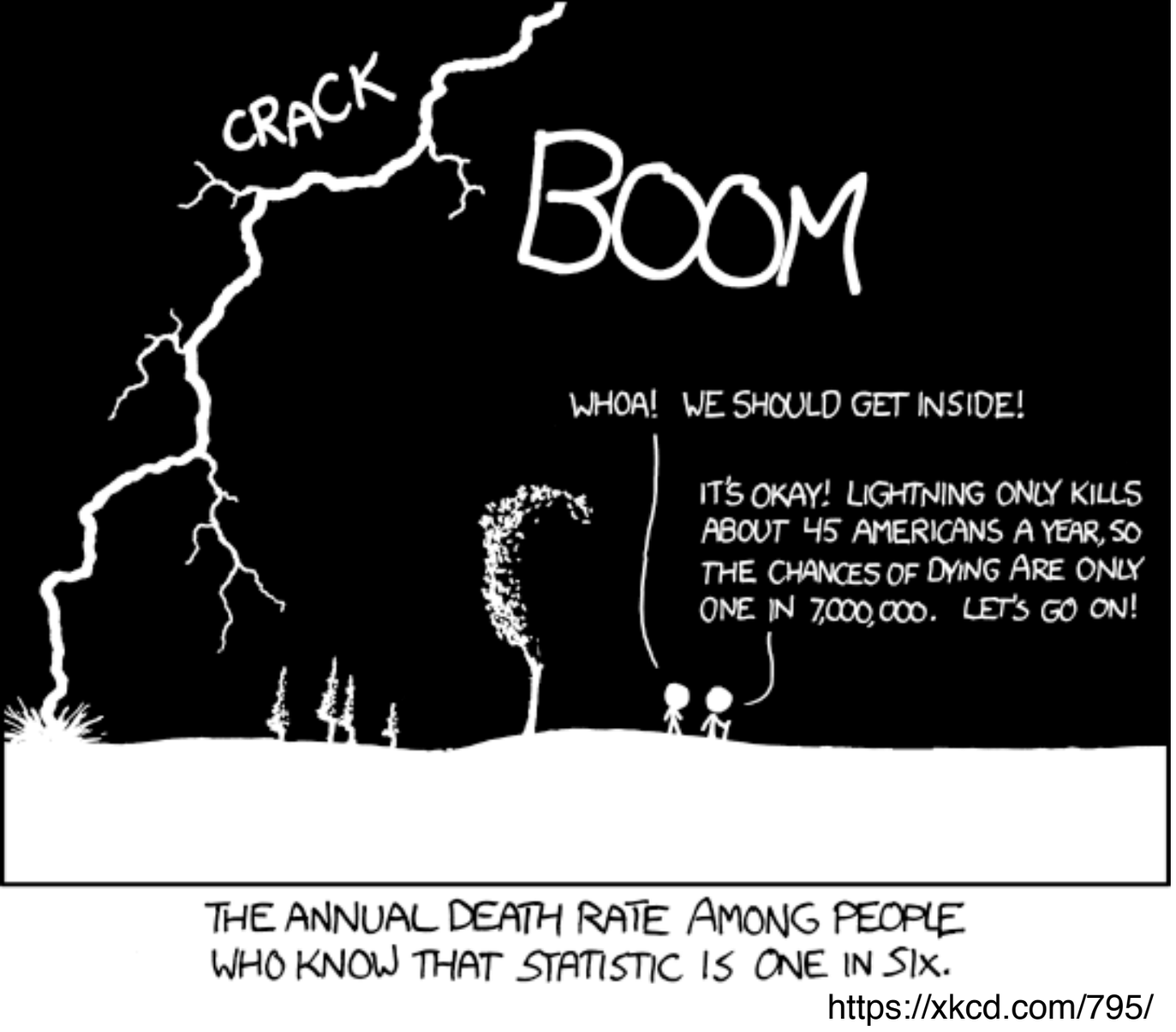
- In the presence of additional information the assessment of event A's probability might change



https://xkcd.com/795/

CONDITIONAL PROBABILITY

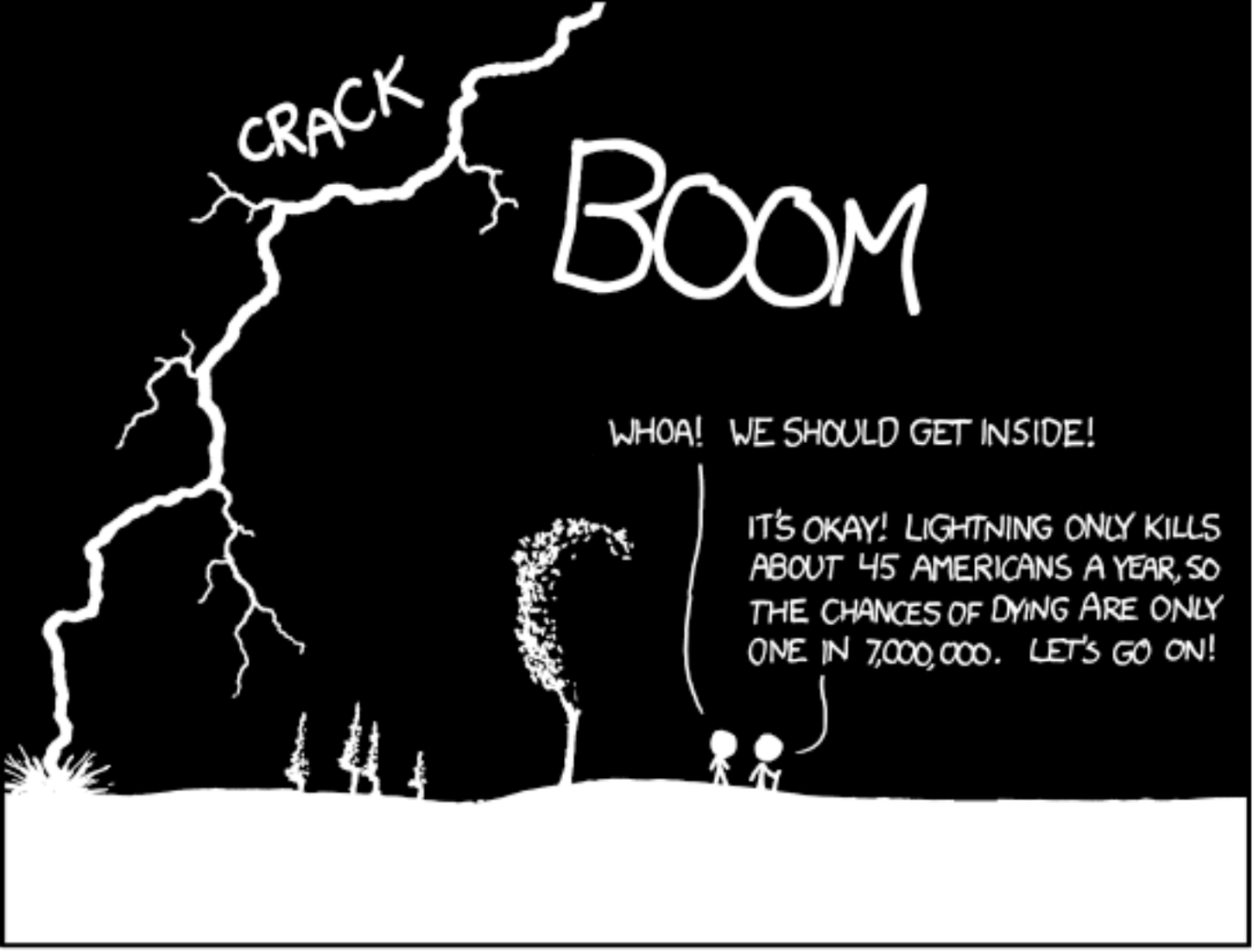
- In the presence of additional information the assessment of event A's probability might change
- P(A | B) is conditional probability of A given B



https://xkcd.com/795/

CONDITIONAL PROBABILITY

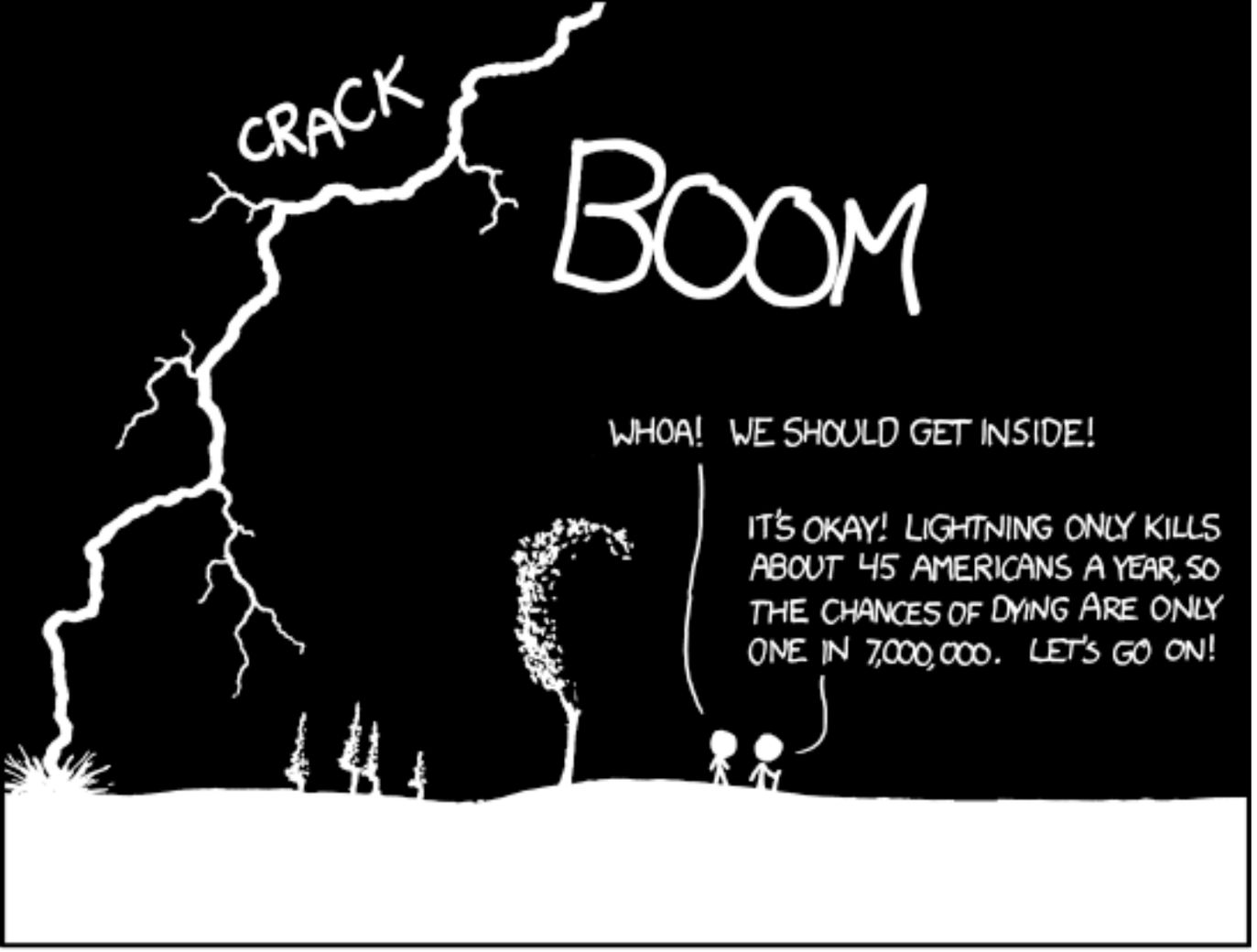
- In the presence of additional information the assessment of event A's probability might change
- P(A | B) is conditional probability of A given B



https://xkcd.com/795/

CONDITIONAL PROBABILITY

- In the presence of additional information the assessment of event A's probability might change
- P(A | B) is conditional probability of A given B
- $$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

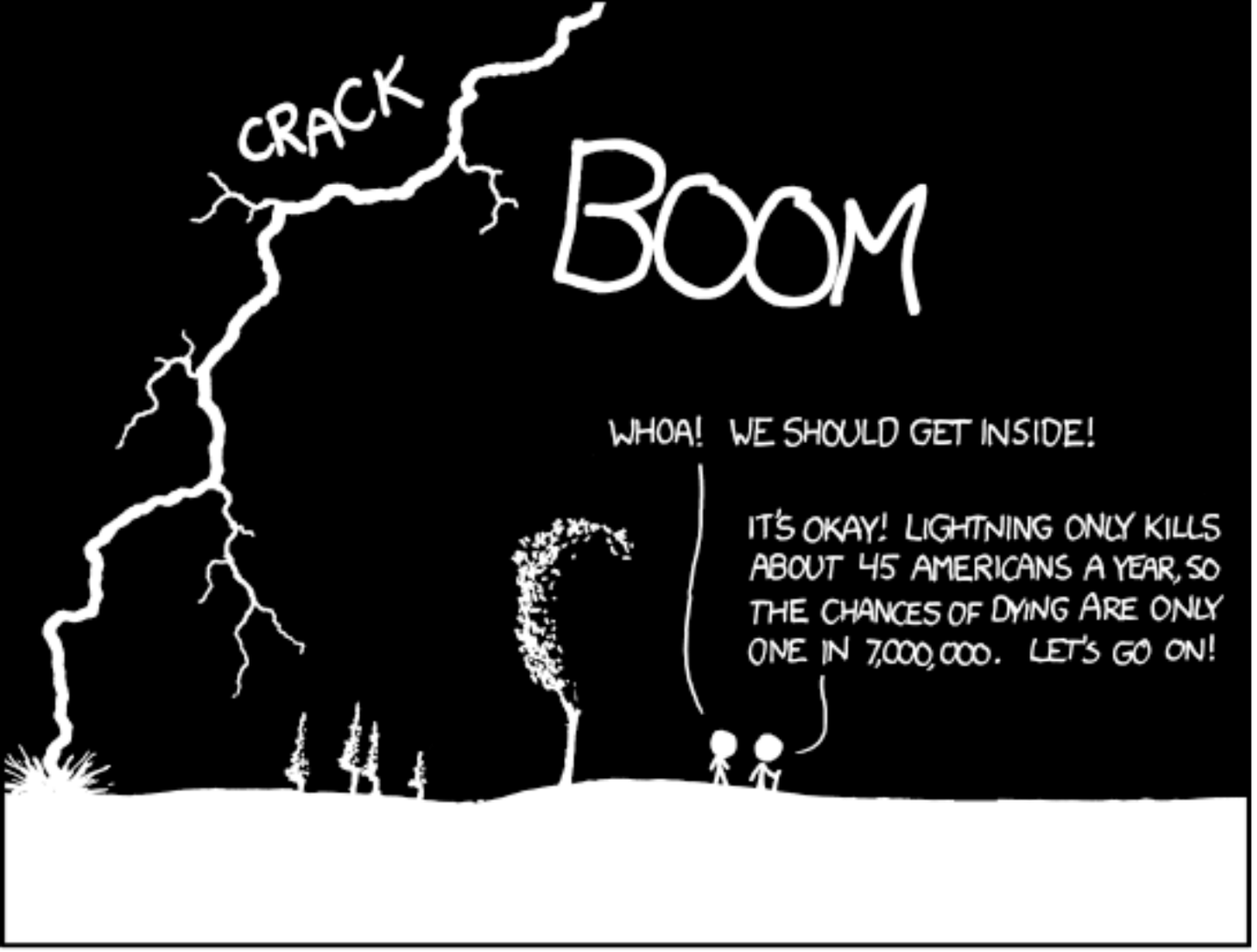


https://xkcd.com/795/

CONDITIONAL PROBABILITY

- In the presence of additional information the assessment of event A's probability might change
- P(A | B) is conditional probability of A given B
- $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$

- P(killed by lightning) = 1/7,000,000

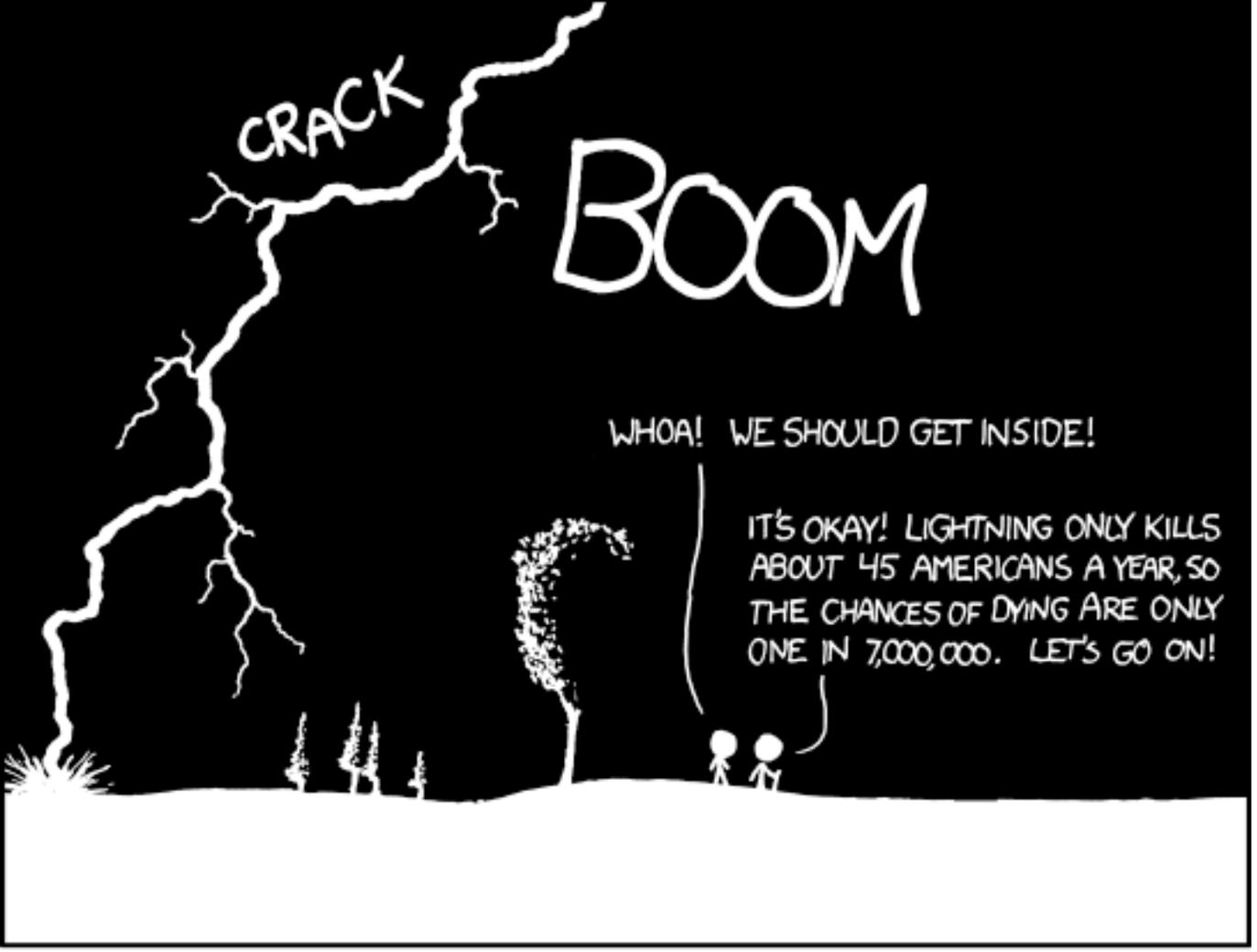


https://xkcd.com/795/

CONDITIONAL PROBABILITY

- In the presence of additional information the assessment of event A's probability might change
- P(A | B) is conditional probability of A given B
- $$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

- P(killed by lightning) = 1/7,000,000
- P(killed by lightning | relying blindly on this statistic) = 1/6

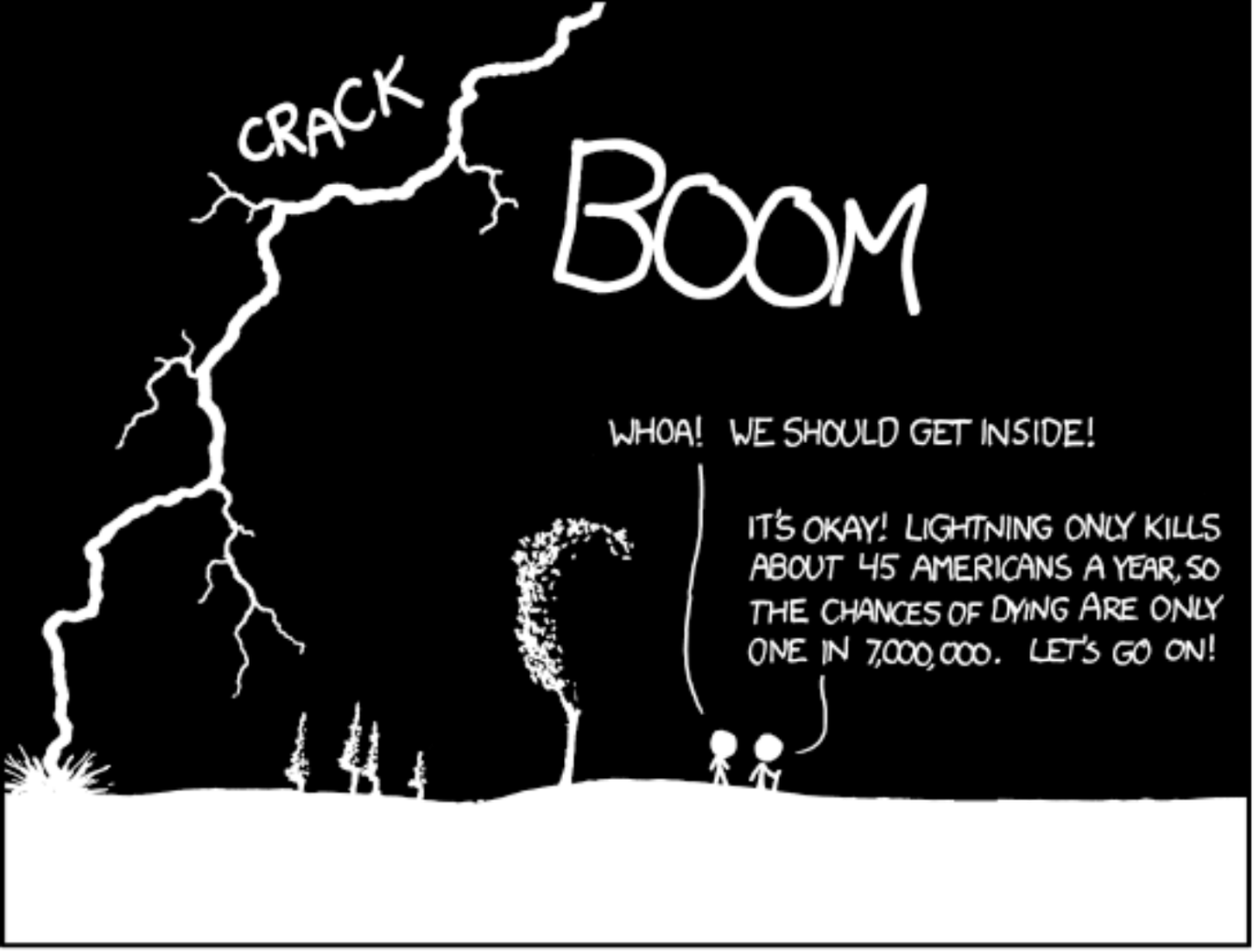


https://xkcd.com/795/

CONDITIONAL PROBABILITY

- In the presence of additional information the assessment of event A's probability might change
- P(A | B) is conditional probability of A given B
- $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$

- P(killed by lightning) = 1/7,000,000
- P(killed by lightning | relying blindly on this statistic) = 1/6
- 2nd line is a joke! Just don't blindly rely on some statistic …



https://xkcd.com/795/

CONDITIONAL PROBABILITY

- In the presence of additional information the assessment of event A's probability might change
- P(A | B) is conditional probability of A given B
- $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$

- P(killed by lightning) = 1/7,000,000
- P(killed by lightning | relying blindly on this statistic) = 1/6
- 2nd line is a joke! Just don't blindly rely on some statistic …
- … especially not when you have additional info (just LOOK!)

YOUR TURN

Out of a group of 40 students all play at least one of badminton, volleyball or table tennis.

8 students play all three games, 10 students play badminton and table tennis
20 students play table tennis and volleyball, 12 students play badminton and volleyball 30 students play table tennis, 25 students play volleyball.

Work on the questions by yourself

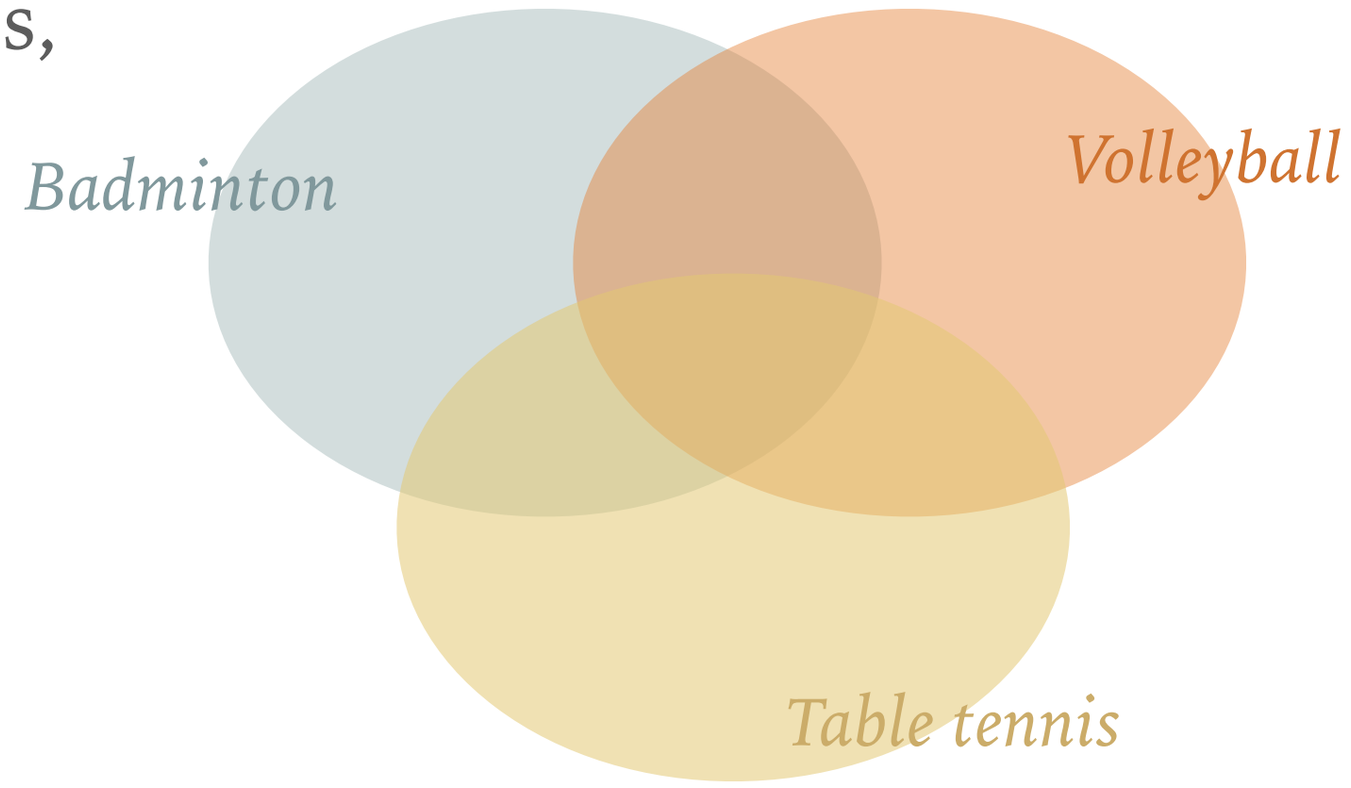
SPORTS?

- Assume one student is picked at random out of this group. What is the probability that he/she
  - plays badminton?
  - plays at least two sports?
- Assume the student you've picked is a volleyball player. What is now the probability that he/she plays
  - badminton?
  - at least two sports?

YOUR TURN: SOLUTION

- Out of a group of 40 students, all play at least one of badminton, volleyball or table tennis.

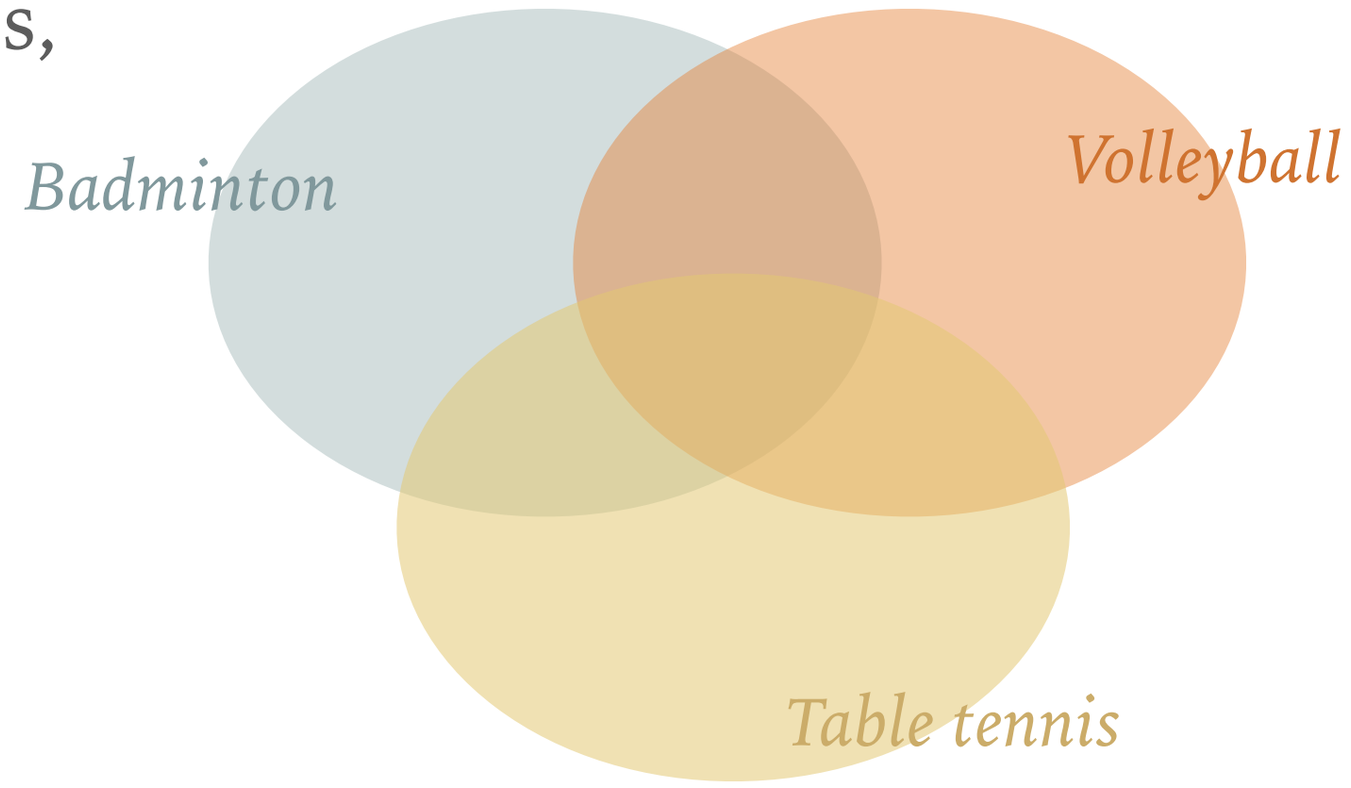
- 8 students play all three games,
  10 students play badminton and table tennis
  20 students play table tennis and volleyball,
  12 students play badminton and volleyball
  30 students play table tennis,
  25 students play volleyball.



YOUR TURN: SOLUTION

- Out of a group of 40 students, all play at least one of badminton, volleyball or table tennis.

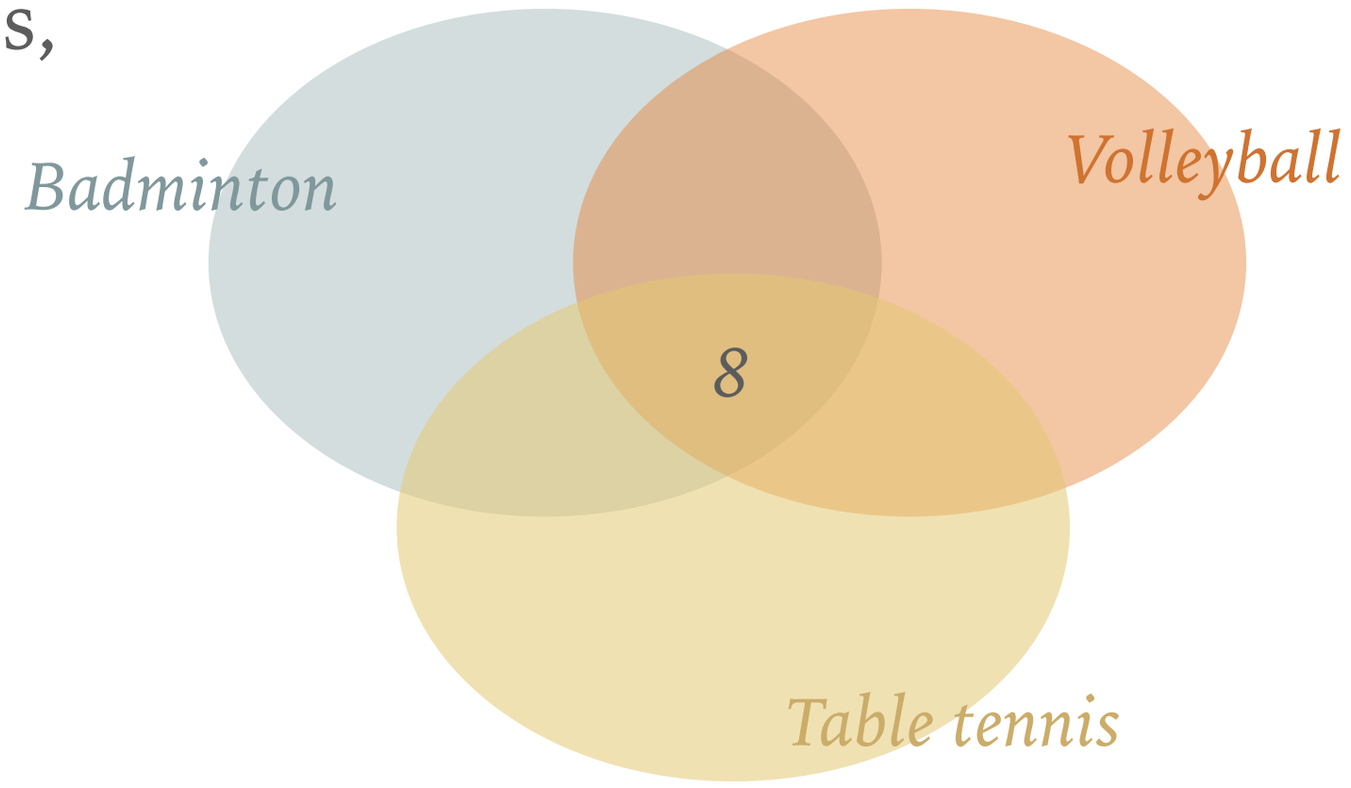
- 8 students play all three games,
  10 students play badminton and table tennis
  20 students play table tennis and volleyball,
  12 students play badminton and volleyball
  30 students play table tennis,
  25 students play volleyball.



YOUR TURN: SOLUTION

- Out of a group of 40 students, all play at least one of badminton, volleyball or table tennis.

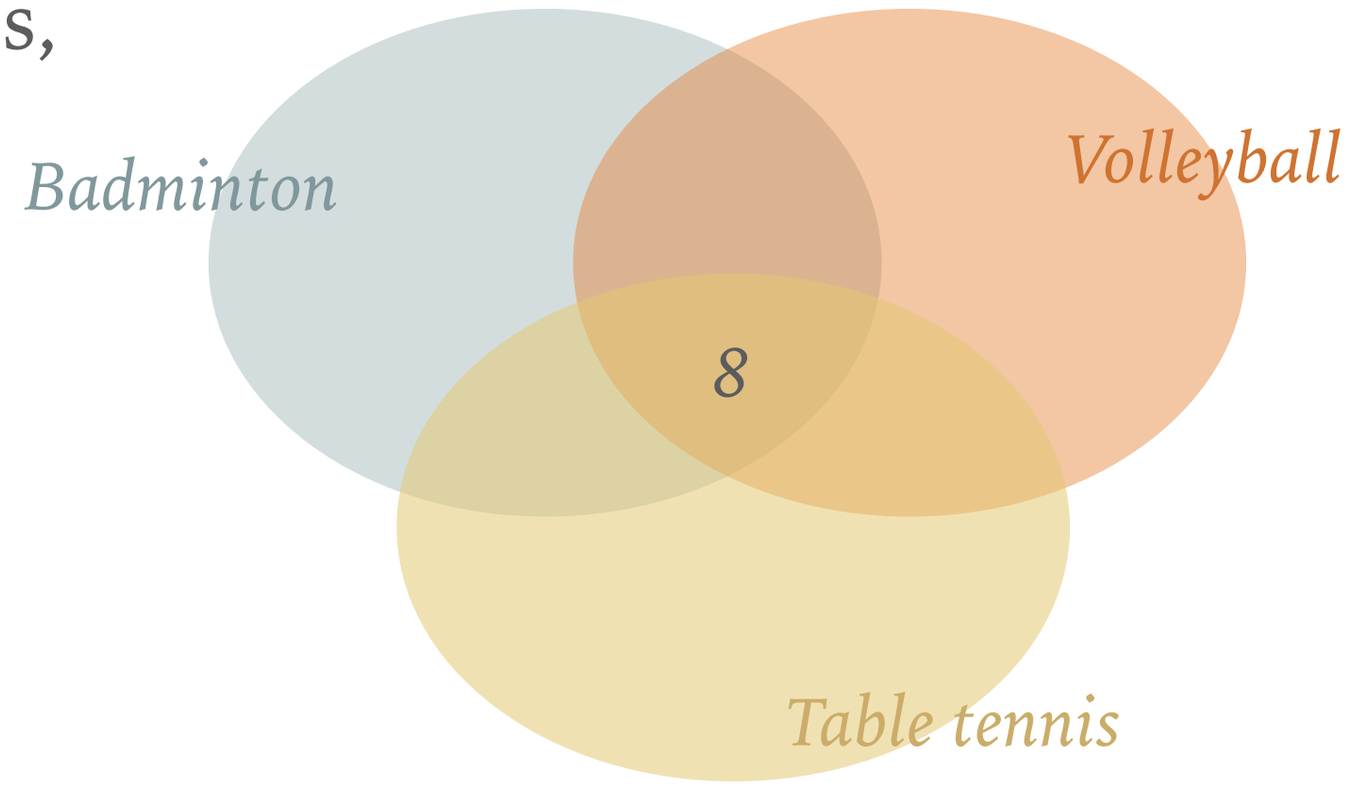
- 8 students play all three games,
  10 students play badminton and table tennis
  20 students play table tennis and volleyball,
  12 students play badminton and volleyball
  30 students play table tennis,
  25 students play volleyball.



YOUR TURN: SOLUTION

- Out of a group of 40 students, all play at least one of badminton, volleyball or table tennis.

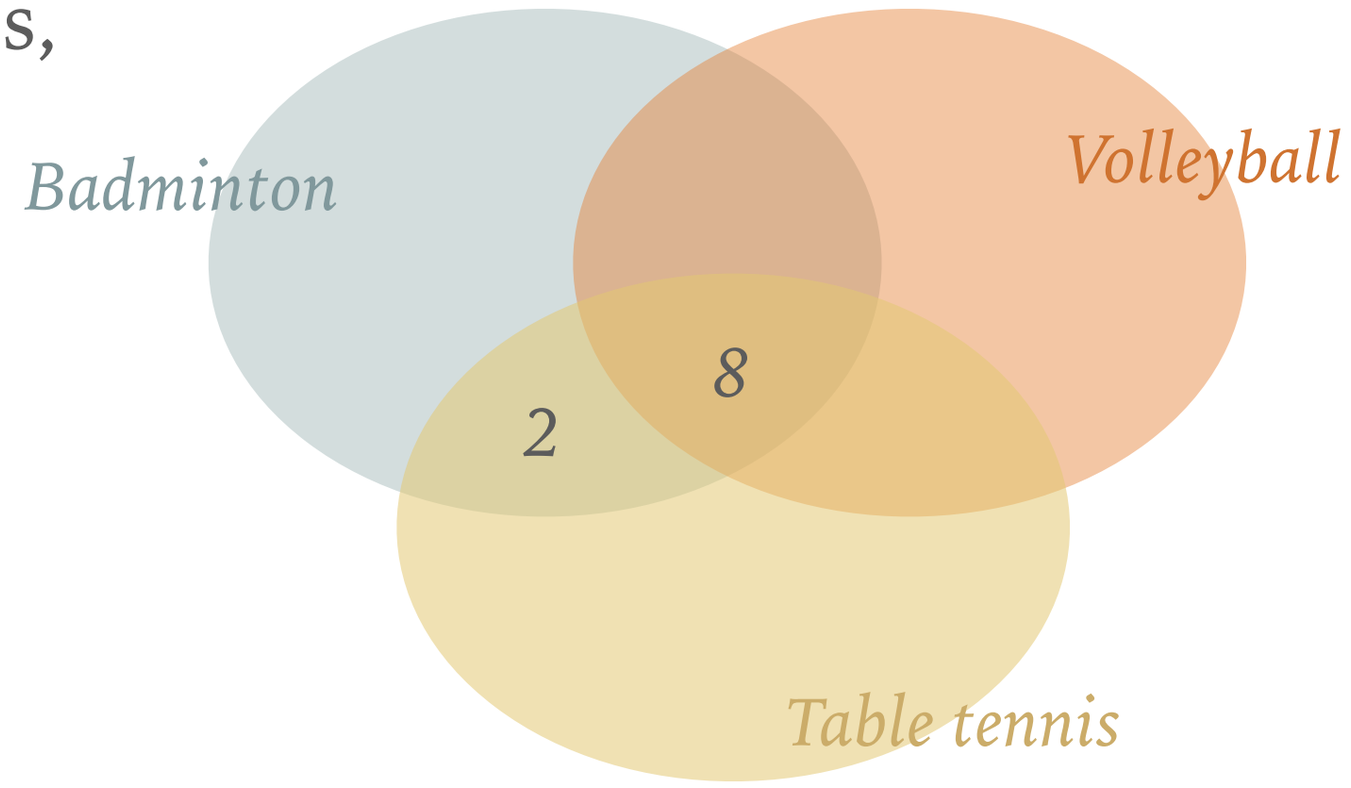
- 8 students play all three games,
  10 students play badminton and table tennis
  20 students play table tennis and volleyball,
  12 students play badminton and volleyball
  30 students play table tennis,
  25 students play volleyball.



YOUR TURN: SOLUTION

- Out of a group of 40 students, all play at least one of badminton, volleyball or table tennis.

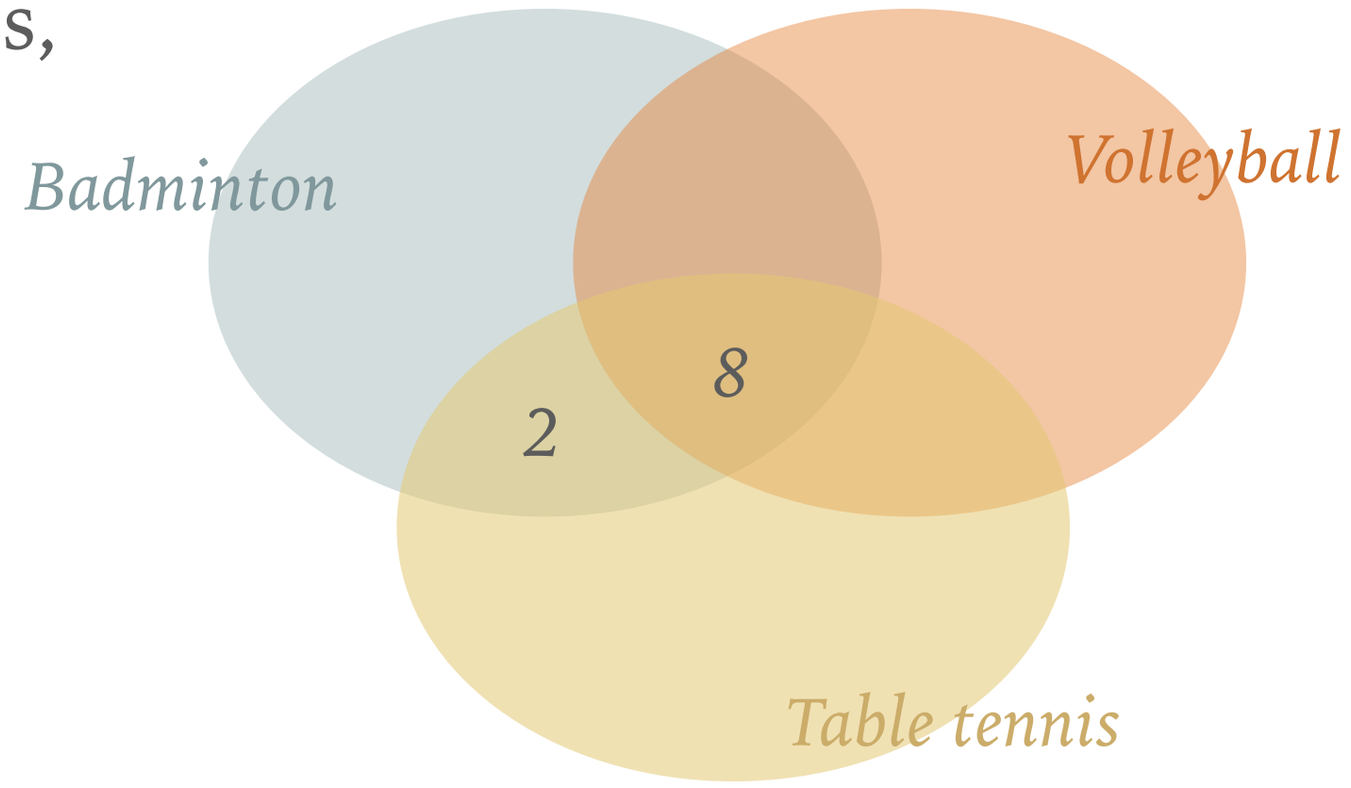
- 8 students play all three games,
  10 students play badminton and table tennis
  20 students play table tennis and volleyball,
  12 students play badminton and volleyball
  30 students play table tennis,
  25 students play volleyball.



YOUR TURN: SOLUTION

- Out of a group of 40 students, all play at least one of badminton, volleyball or table tennis.

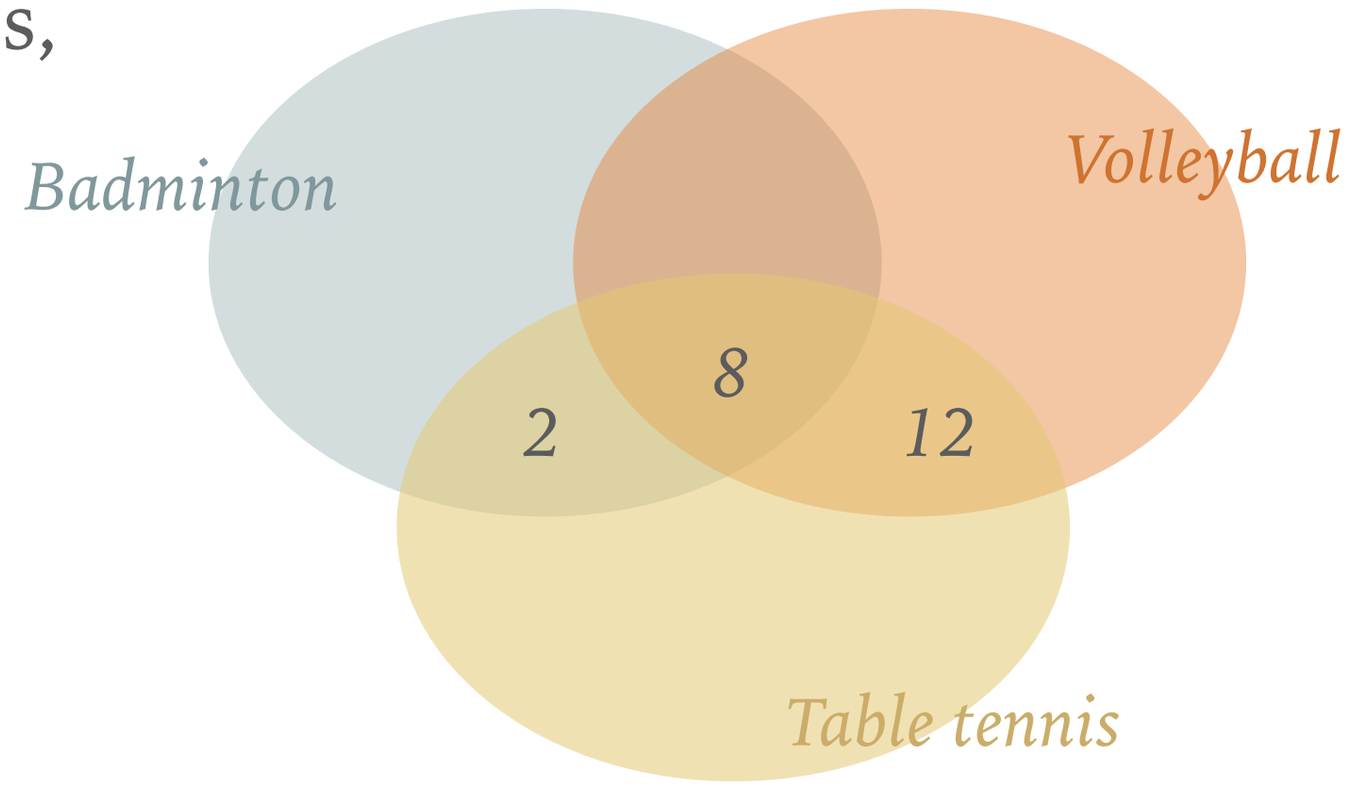
- 8 students play all three games,
  10 students play badminton and table tennis
  20 students play table tennis and volleyball,
  12 students play badminton and volleyball
  30 students play table tennis,
  25 students play volleyball.



YOUR TURN: SOLUTION

- Out of a group of 40 students, all play at least one of badminton, volleyball or table tennis.

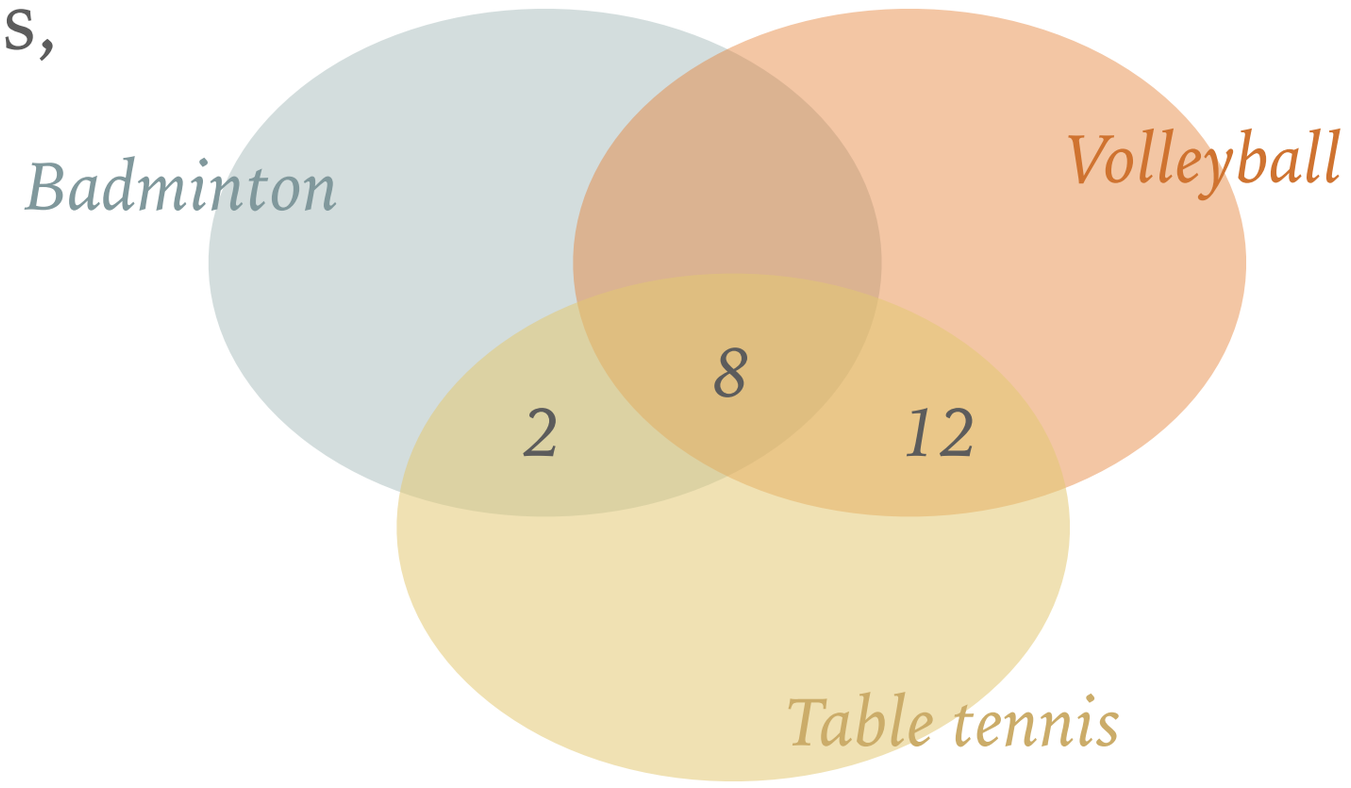
- 8 students play all three games,
  10 students play badminton and table tennis
  20 students play table tennis and volleyball,
  12 students play badminton and volleyball
  30 students play table tennis,
  25 students play volleyball.



YOUR TURN: SOLUTION

- Out of a group of 40 students, all play at least one of badminton, volleyball or table tennis.

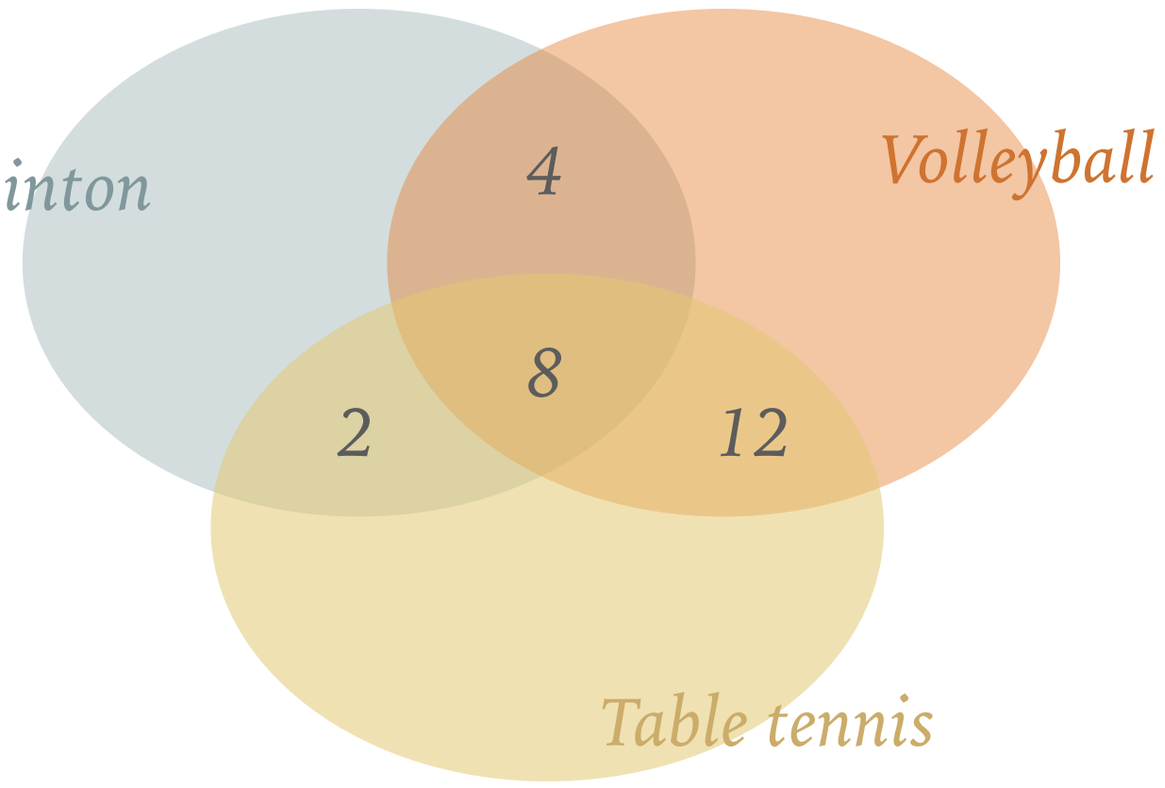
- 8 students play all three games,
  10 students play badminton and table tennis
  20 students play table tennis and volleyball,
  12 students play badminton and volleyball
  30 students play table tennis,
  25 students play volleyball.



YOUR TURN: SOLUTION

- Out of a group of 40 students, all play at least one of badminton, volleyball or table tennis.

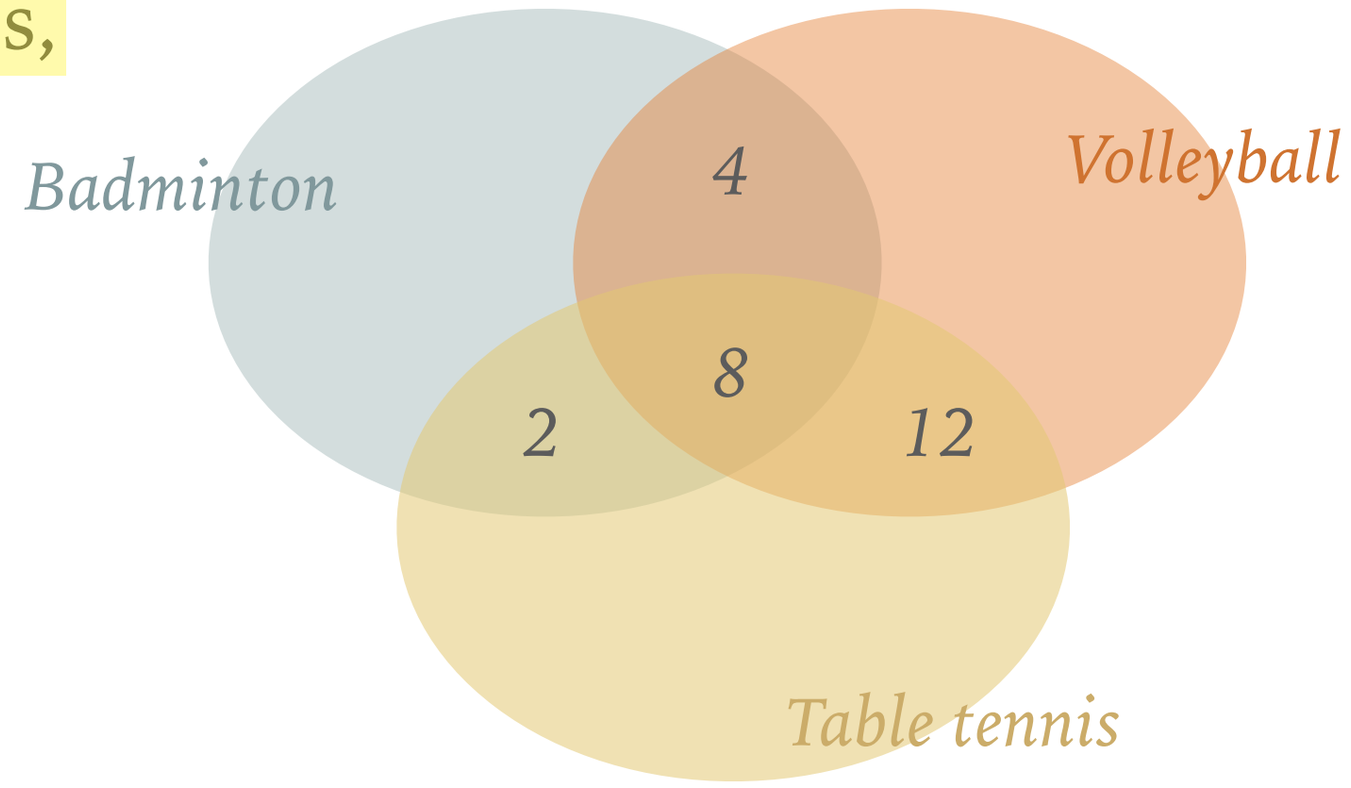
- 8 students play all three games,
  10 students play badminton and table tennis
  20 students play table tennis and volleyball,
  12 students play badminton and volleyball
  30 students play table tennis,
  25 students play volleyball.



YOUR TURN: SOLUTION

- Out of a group of 40 students, all play at least one of badminton, volleyball or table tennis.

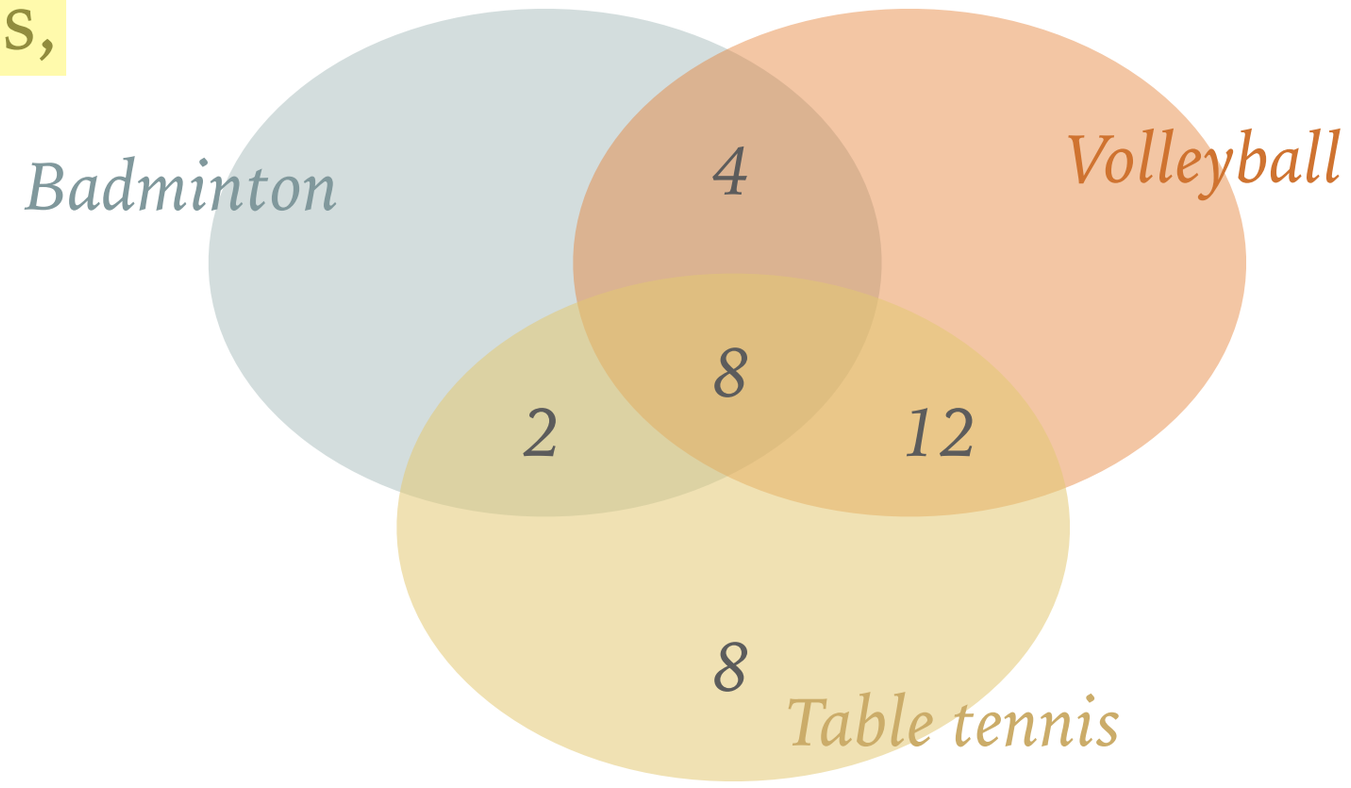
- 8 students play all three games,
  10 students play badminton and table tennis
  20 students play table tennis and volleyball,
  12 students play badminton and volleyball
  30 students play table tennis,
  25 students play volleyball.



YOUR TURN: SOLUTION

- Out of a group of 40 students, all play at least one of badminton, volleyball or table tennis.

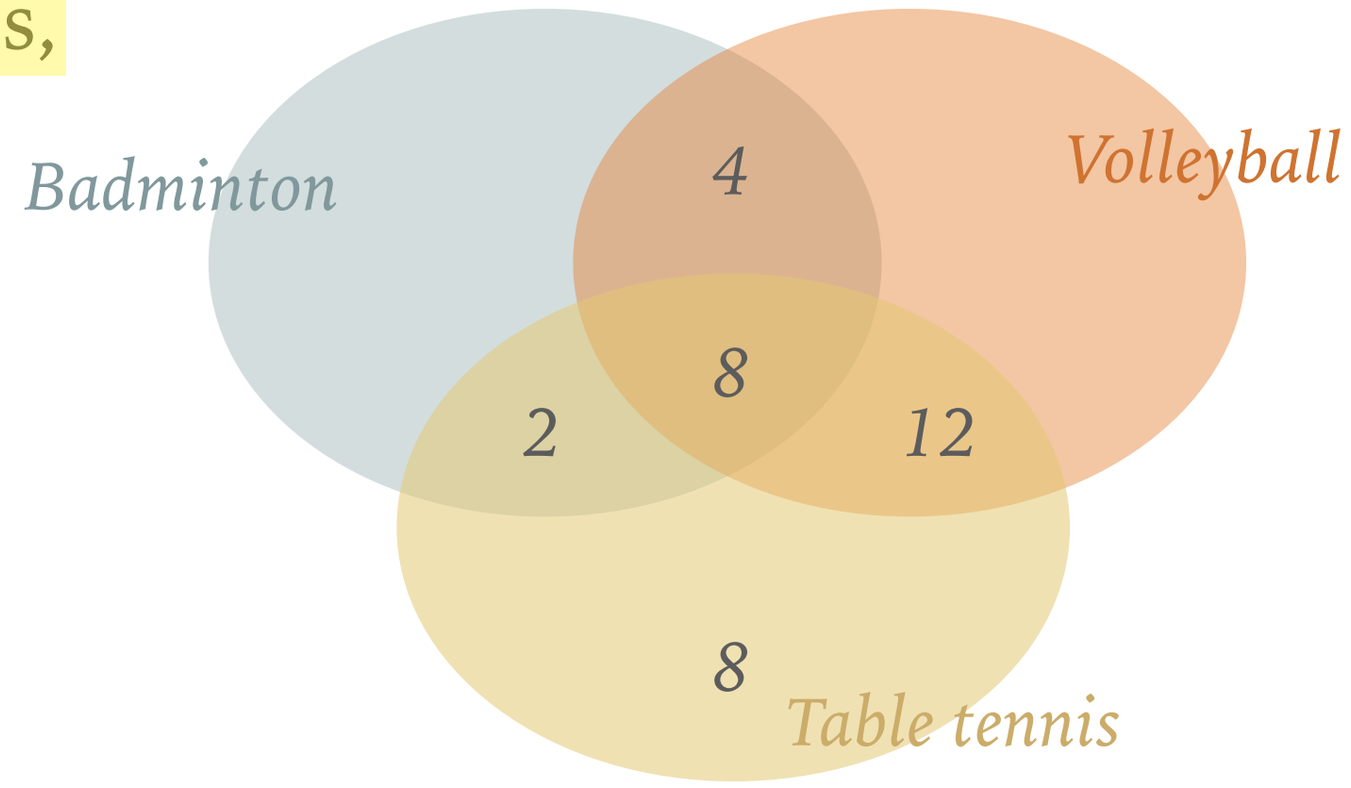
- 8 students play all three games,
  10 students play badminton and table tennis
  20 students play table tennis and volleyball,
  12 students play badminton and volleyball
  30 students play table tennis,
  25 students play volleyball.



YOUR TURN: SOLUTION

- Out of a group of 40 students, all play at least one of badminton, volleyball or table tennis.

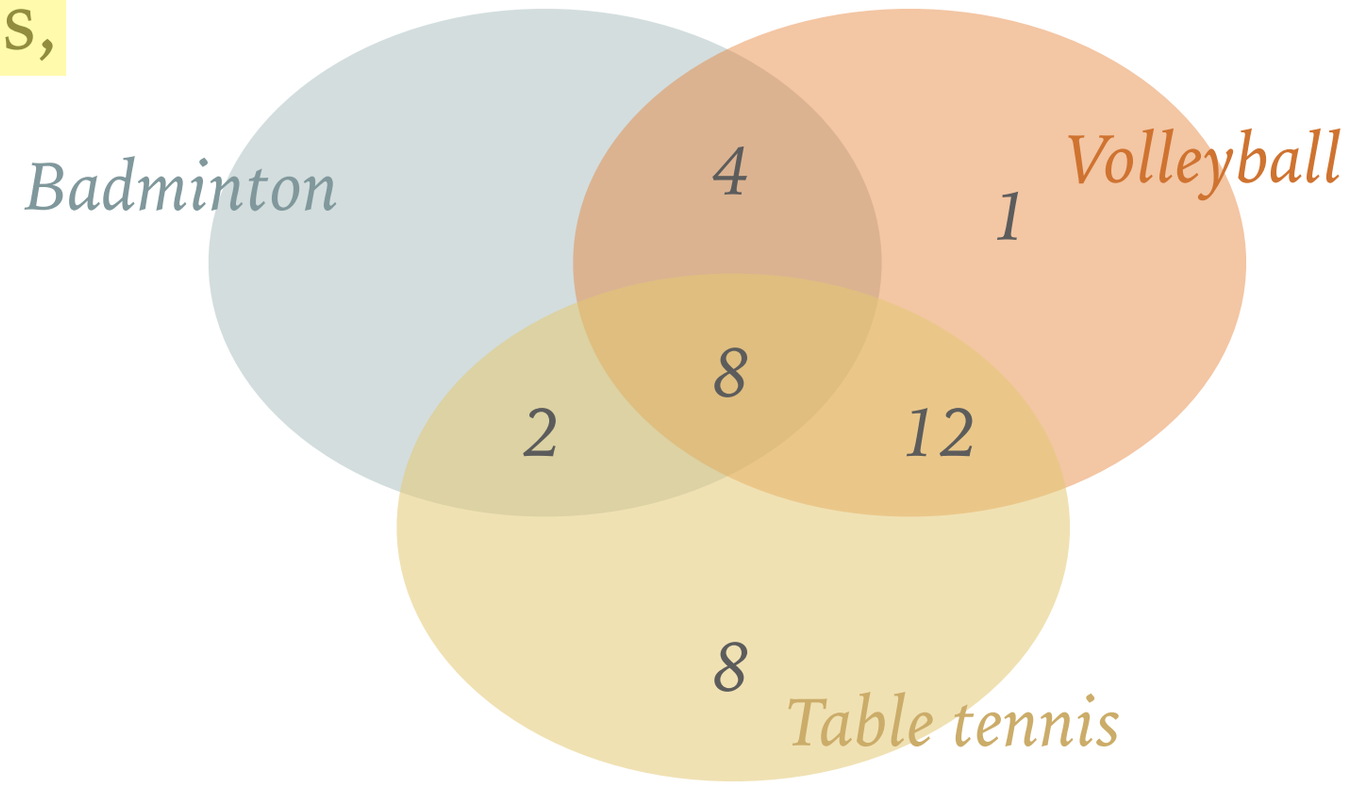
- 8 students play all three games,
  10 students play badminton and table tennis
  20 students play table tennis and volleyball,
  12 students play badminton and volleyball
  30 students play table tennis,
  25 students play volleyball.



YOUR TURN: SOLUTION

- Out of a group of 40 students, all play at least one of badminton, volleyball or table tennis.

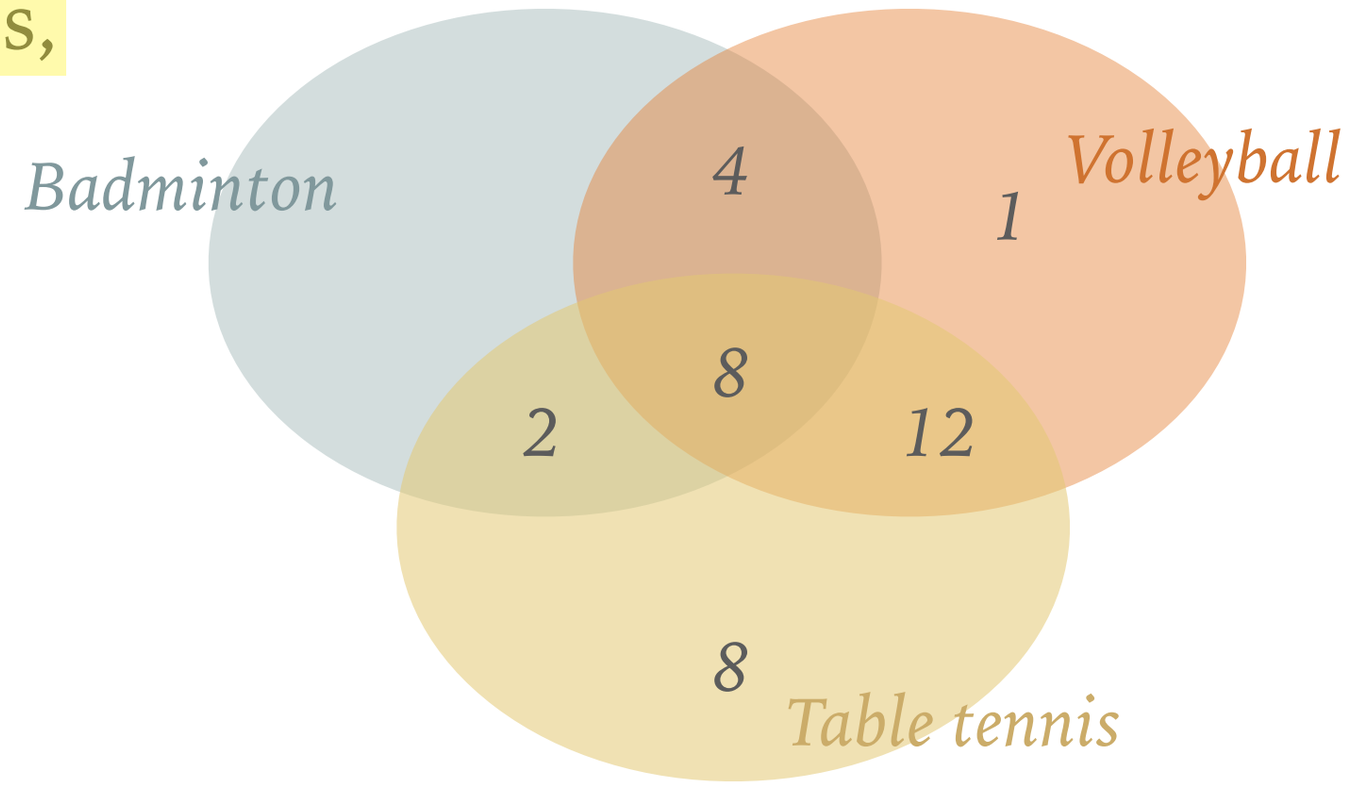
- 8 students play all three games,
  10 students play badminton and table tennis
  20 students play table tennis and volleyball,
  12 students play badminton and volleyball
  30 students play table tennis,
  25 students play volleyball.



YOUR TURN: SOLUTION

- Out of a group of 40 students, all play at least one of badminton, volleyball or table tennis.

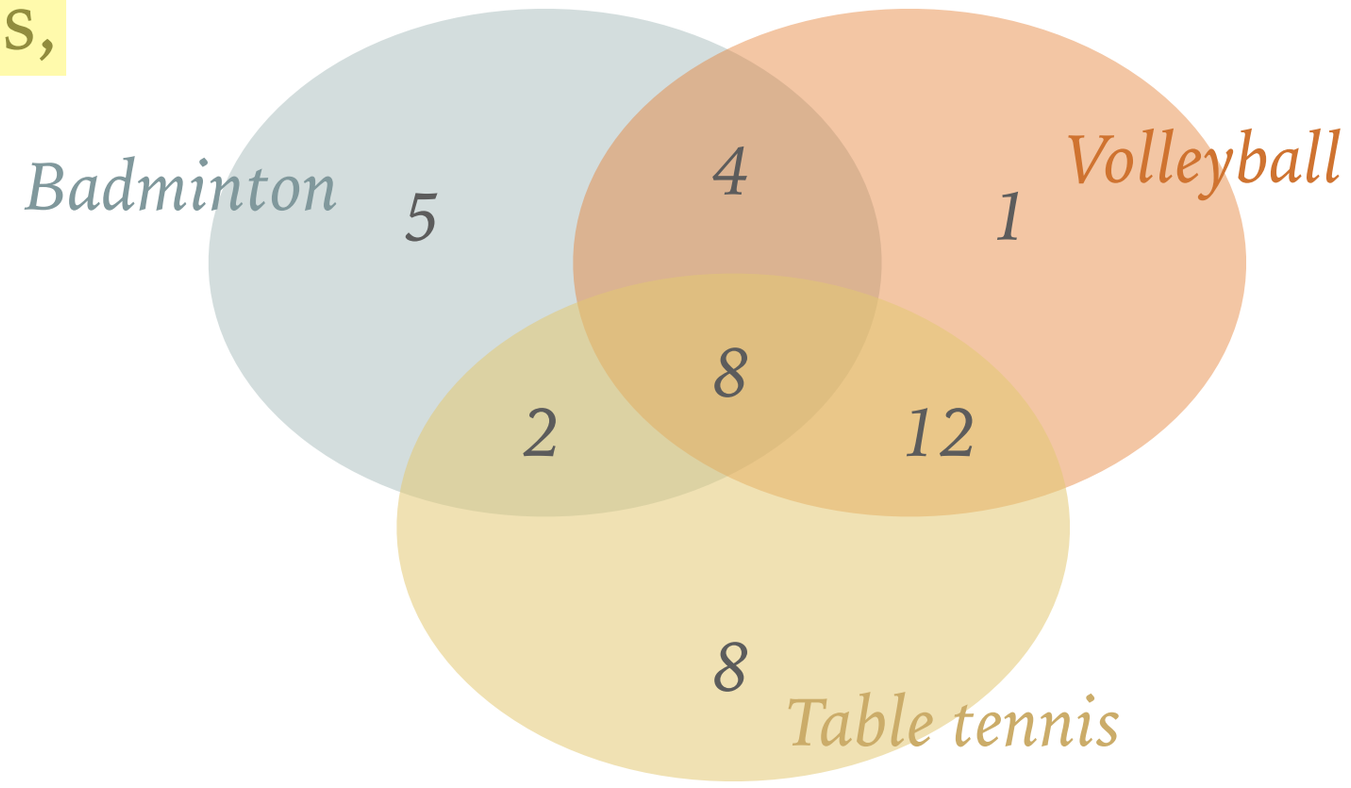
- 8 students play all three games,
  10 students play badminton and table tennis
  20 students play table tennis and volleyball,
  12 students play badminton and volleyball
  30 students play table tennis,
  25 students play volleyball.



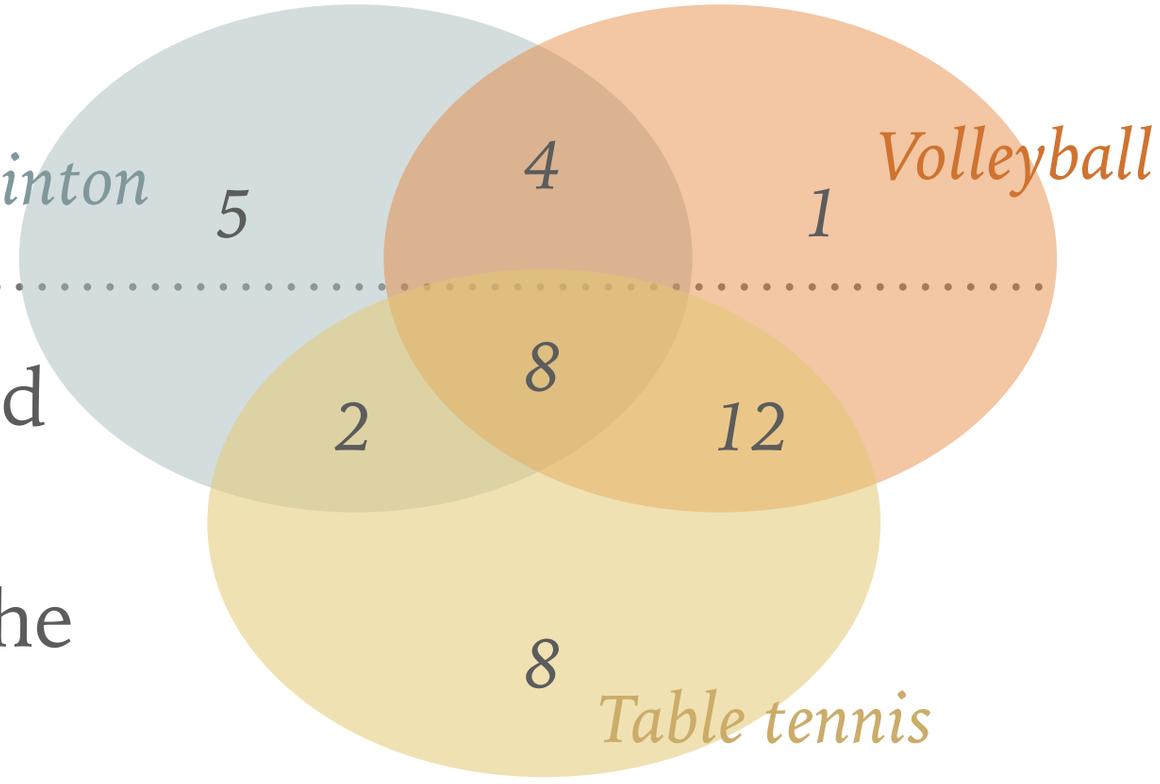
YOUR TURN: SOLUTION

- Out of a group of 40 students, all play at least one of badminton, volleyball or table tennis.

- 8 students play all three games,
  10 students play badminton and table tennis
  20 students play table tennis and volleyball,
  12 students play badminton and volleyball
  30 students play table tennis,
  25 students play volleyball.

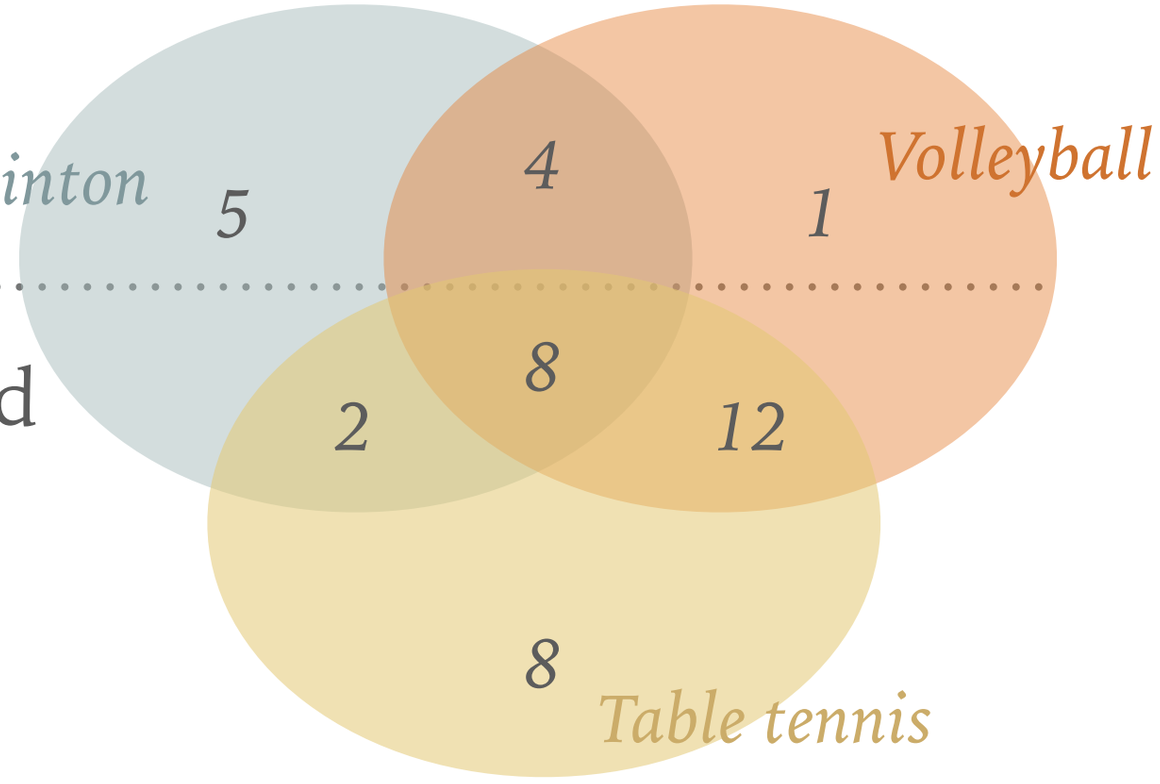


YOUR TURN: SOLUTION



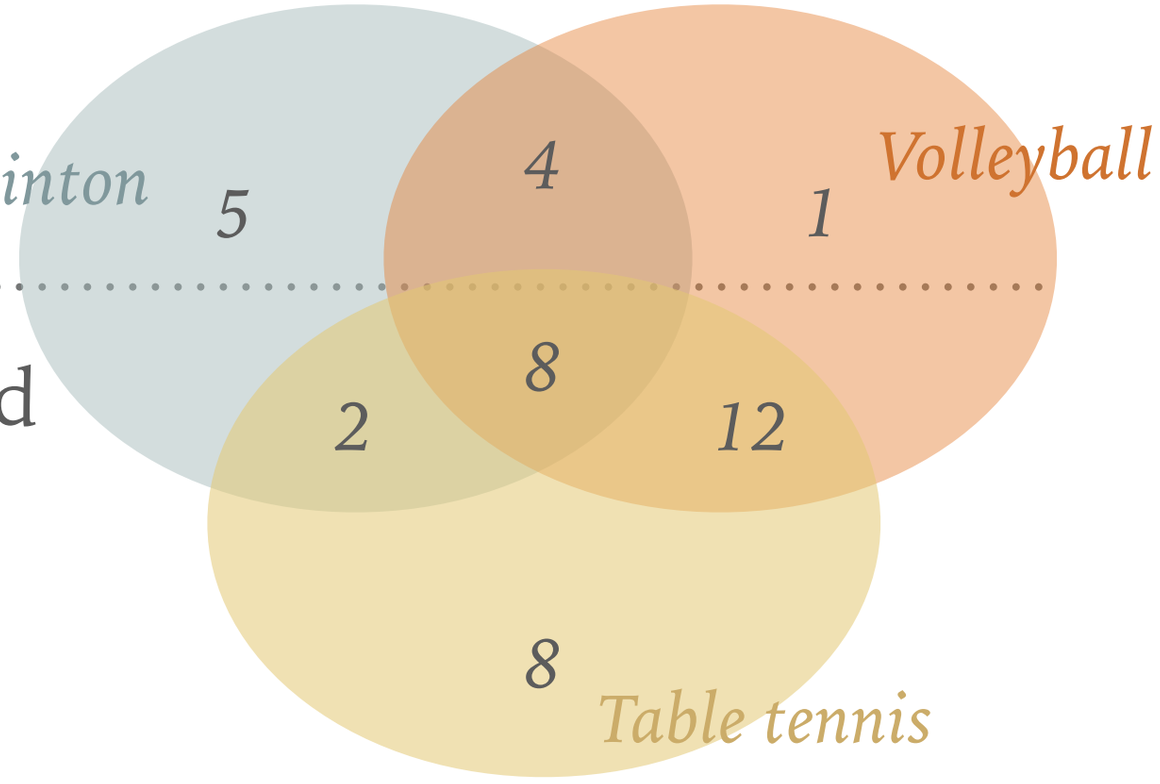
- Assume that one student is picked at random out of this group. What is the probability that he/she
  - plays Badminton?
    P(Badminton) = (5 + 2 + 8 + 4)/40 = 19/40
  - plays at least two sports?
    P(at least two sports) = (2+8+4+12)/40 = 26/40

YOUR TURN: SOLUTION



- Assume the student you've picked is a volleyball player.
  What is now the probability that he/she
  - plays Badminton?
    P(Badminton | Volleyball) = (8 + 4)/25 = 12/25
  - plays at least two sports?
    P(at least two sports | Vb) = (8+4+12)/25 = 24/25

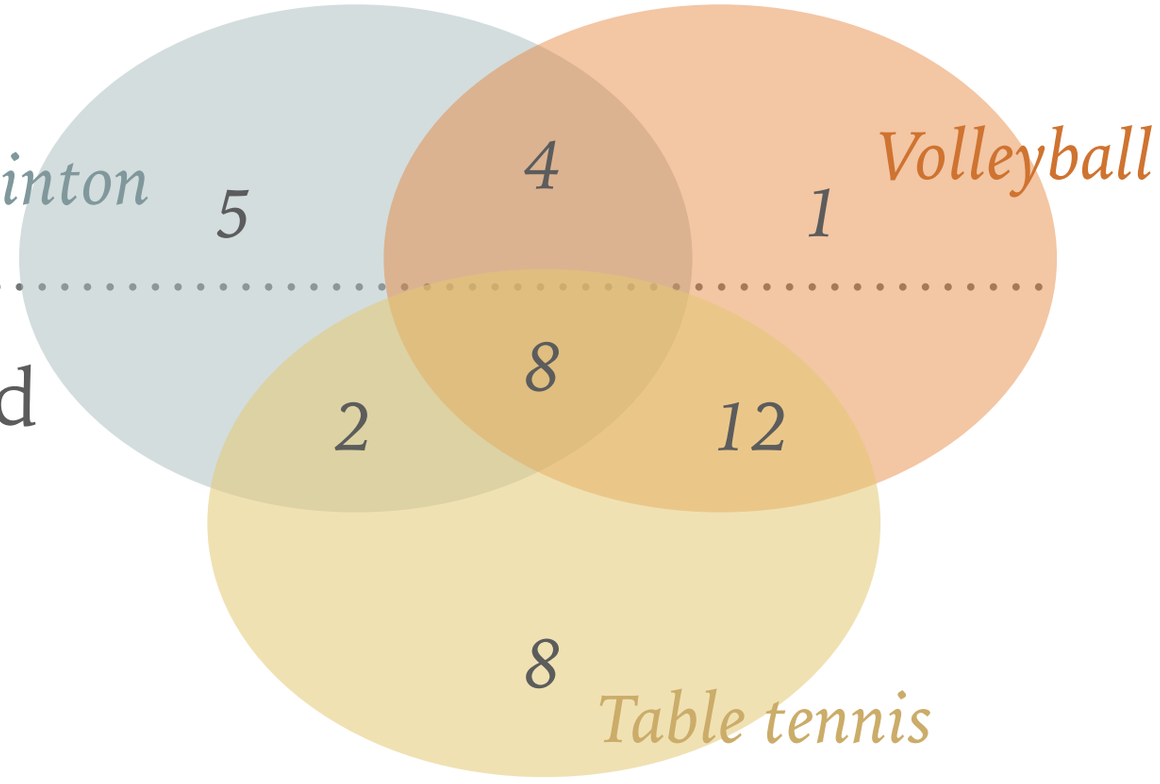
YOUR TURN: SOLUTION



- Assume the student you've picked is a volleyball player.
  What is now the probability that he/she
  - plays Badminton?
    P(Badminton | Volleyball) = (8 + 4)/25 = 12/25
  - plays at least two sports?
    P(at least two sports | Vb) = (8+4+12)/25 = 24/25

Compare:

YOUR TURN: SOLUTION



- Assume the student you've picked is a volleyball player.
  What is now the probability that he/she
  - plays Badminton?
    P(Badminton | Volleyball) = (8 + 4)/25 = 12/25
  - plays at least two sports?
    P(at least two sports | Vb) = (8+4+12)/25 = 24/25

Compare: P(Badminton) = 19/40

P(at least two sports) = 26/40

(STATISTICAL) INDEPENDENCE

- Events A and B are **independent**, if
  $P(A \cap B) = P(A) \cdot P(B)$

- equivalently
  events A and B are **independent**, if knowing the outcome of B does not impact the probability of A:
  $P(A \mid B) = P(A)$

We will get back to these concepts …

QUESTIONS?

We'll take a 5-minute break now

5 MINUTE BREAK

4

3

2

1

We'll take a 5-minute break now

. . . READY?

RANDOM VARIABLES

RANDOM VARIABLE

- A function X that assigns each element of the sample space a real valued number is called a **random variable**:

$$X : \Omega \rightarrow R$$

- We distinguish between discrete and continuous random variables,

  e.g. #faulty devices found in factory line per day (countable)
  amount of milk per day from goat Dottie (measurable)
  height of NBA players (measurable)
  #wins in five games (countable)

DISCRETE RANDOM VARIABLES

- **Probability Mass Function**, PMF
  The function $p_X(x) = P(X = x)$ is called the probability mass function of random variable X.

- **Properties of a PMF**
  - (i) $0 \leq p_X(x) \leq 1$ for all $x \in \{x_1, x_2, x_3, ...\}$
  - (ii) $\sum p(x_i) = 1$ (probabilities over all possible outcomes sum to 1)

EXAMPLE: 3 COIN FLIPS

EXAMPLE: 3 COIN FLIPS

- A fair coin is flipped 3 times

EXAMPLE: 3 COIN FLIPS

- A fair coin is flipped 3 times
- Sample space
  Ω = {HHH, HHT, HTH, HTT,
  THH, THT, TTH, TTT}
  all elements in Ω occur equally likely

EXAMPLE: 3 COIN FLIPS

- A fair coin is flipped 3 times
- Sample space
  Ω = {HHH, HHT, HTH, HTT,
  THH, THT, TTH, TTT}
  all elements in Ω occur equally likely
- Define $X : \Omega \to R$ as number of heads:
  $X(HHH) = 3$, $X(HTH) = 2$, …
  $X(TTT) = 0$

EXAMPLE: 3 COIN FLIPS

- A fair coin is flipped 3 times
- Sample space
  Ω = {HHH, HHT, HTH, HTT,
  THH, THT, TTH, TTT}
  all elements in Ω occur equally likely
- Define $X : \Omega \to R$ as number of heads:
  $X(HHH) = 3$, $X(HTH) = 2$, …
  $X(TTT) = 0$
- Probability Mass Function: $P(X = x)$
  $x \in \{0, 1, 2, 3\}$:
  $P(X = 0) = 1/8$ $P(X = 1) = 3/8$
  $P(X = 2) = 3/8$ $P(X = 3) = 1/8$

EXAMPLE: 3 COIN FLIPS

- A fair coin is flipped 3 times
- Sample space
  Ω = {HHH, HHT, HTH, HTT,
  THH, THT, TTH, TTT}
  all elements in Ω occur equally likely
- Define $X : \Omega \rightarrow R$ as number of heads:
  $X(HHH) = 3$, $X(HTH) = 2$, …
  $X(TTT) = 0$
- Probability Mass Function: $P(X = x)$
  $x \in \{0, 1, 2, 3\}$:
  $P(X = 0) = 1/8$ $P(X = 1) = 3/8$
  $P(X = 2) = 3/8$ $P(X = 3) = 1/8$
- Two or more heads?
  $P(X \geq 2) = 3/8+1/8 = 0.5$
  At least one head?
  $P(X \geq 1) = 1 - P(X = 0) = 7/8$

R TO THE RESCUE

R TO THE RESCUE

- Package discreteRV

```
install.packages("discreteRV")
library(discreteRV)
```

R TO THE RESCUE

- Package discreteRV

```
install.packages("discreteRV")
library(discreteRV)
```

- Define a random variable

```
~~fair.coin <- RV(c("H", "T"), probs = 0.5)~~
fair.coin <- RV(c("1", "0"), probs = 0.5)
```

R TO THE RESCUE

- Package discreteRV
```
install.packages("discreteRV")
library(discreteRV)
```
- Define a random variable
```
~~fair.coin <- RV(c("H", "T"), probs = 0.5)~~
fair.coin <- RV(c("1", "0"), probs = 0.5)
```
- Multiple flips of the coin:

R TO THE RESCUE

- Package discreteRV

```
install.packages("discreteRV")
library(discreteRV)
```

- Define a random variable

```
~~fair.coin <- RV(c("H", "T"), probs = 0.5)~~
fair.coin <- RV(c("1", "0"), probs = 0.5)
```

- Multiple flips of the coin:

```
➤ iid(fair.coin, 2)
  Random variable with 4 outcomes
  Outcomes 0,0 0,1 1,0 1,1
  Probs    1/4 1/4 1/4 1/4
```

R TO THE RESCUE

- Package discreteRV
  ```
  install.packages("discreteRV")
  library(discreteRV)
  ```
- Define a random variable
  ~~`fair.coin <- RV(c("H", "T"), probs = 0.5)`~~
  ```
  fair.coin <- RV(c("1", "0"), probs = 0.5)
  ```
- Multiple flips of the coin:
  - `iid(fair.coin, 2)`
    ```
    Random variable with 4 outcomes
    Outcomes 0,0 0,1 1,0 1,1
    Probs    1/4 1/4 1/4 1/4
    ```
  - `iid(fair.coin, 2, fractions = FALSE)`
    ```
    Random variable with 4 outcomes
    Outcomes  0,0  0,1  1,0  1,1
    Probs    0.25 0.25 0.25 0.25
    ```

R TO THE RESCUE

- Package discreteRV
```
install.packages("discreteRV")
library(discreteRV)
```
- Define a random variable
```
~~fair.coin <- RV(c("H", "T"), probs = 0.5)~~
fair.coin <- RV(c("1", "0"), probs = 0.5)
```
- Multiple flips of the coin:
  - `iid(fair.coin, 2)`
```
Random variable with 4 outcomes
Outcomes 0,0 0,1 1,0 1,1
Probs    1/4 1/4 1/4 1/4
```
  - `iid(fair.coin, 2, fractions = FALSE)`
```
Random variable with 4 outcomes
Outcomes  0,0  0,1  1,0  1,1
Probs    0.25 0.25 0.25 0.25
```
  - `iid(fair.coin, 3, fractions = FALSE)`
```
Random variable with 8 outcomes
Outcomes 0,0,0 0,0,1 0,1,0 0,1,1 1,0,0 1,0,1 1,1,0 1,1,1
Probs    0.125 0.125 0.125 0.125 0.125 0.125 0.125 0.125
```

R TO THE RESCUE

- Package discreteRV
```
install.packages("discreteRV")
library(discreteRV)
```
- Define a random variable
```
~~fair.coin <- RV(c("H", "T"), probs = 0.5)~~
fair.coin <- RV(c("1", "0"), probs = 0.5)
```
- Multiple flips of the coin:

R TO THE RESCUE

- Package discreteRV

```
install.packages("discreteRV")
library(discreteRV)
```

- Define a random variable

```
~~fair.coin <- RV(c("H", "T"), probs = 0.5)~~
fair.coin <- RV(c("1", "0"), probs = 0.5)
```

- Multiple flips of the coin:
- Sum of heads in n flips:

```
(heads10 <- SofIID(fair.coin, 10, fractions = FALSE))
Random variable with 11 outcomes
Outcomes     0     1     2     3     4     5     6     7     8     9    10
Probs    0.001 0.010 0.044 0.117 0.205 0.246 0.205 0.117 0.044 0.010 0.001
```

R TO THE RESCUE

- Package discreteRV

```
install.packages("discreteRV")
library(discreteRV)
```

- Define a random variable

```
~~fair.coin <- RV(c("H", "T"), probs = 0.5)~~
fair.coin <- RV(c("1", "0"), probs = 0.5)
```

- Multiple flips of the coin:
- Sum of heads in n flips:

```
(heads10 <- SofIID(fair.coin, 10, fractions = FALSE))
Random variable with 11 outcomes
Outcomes     0     1     2     3     4     5     6     7     8     9    10
Probs    0.001 0.010 0.044 0.117 0.205 0.246 0.205 0.117 0.044 0.010 0.001
```

- Evaluate probabilities

```
P(heads10 > 3)
[1] 0.828125
```

R TO THE RESCUE

- Package discreteRV

```
install.packages("discreteRV")
library(discreteRV)
```

- Define a random variable

```
~~fair.coin <- RV(c("H", "T"), probs = 0.5)~~
fair.coin <- RV(c("1", "0"), probs = 0.5)
```

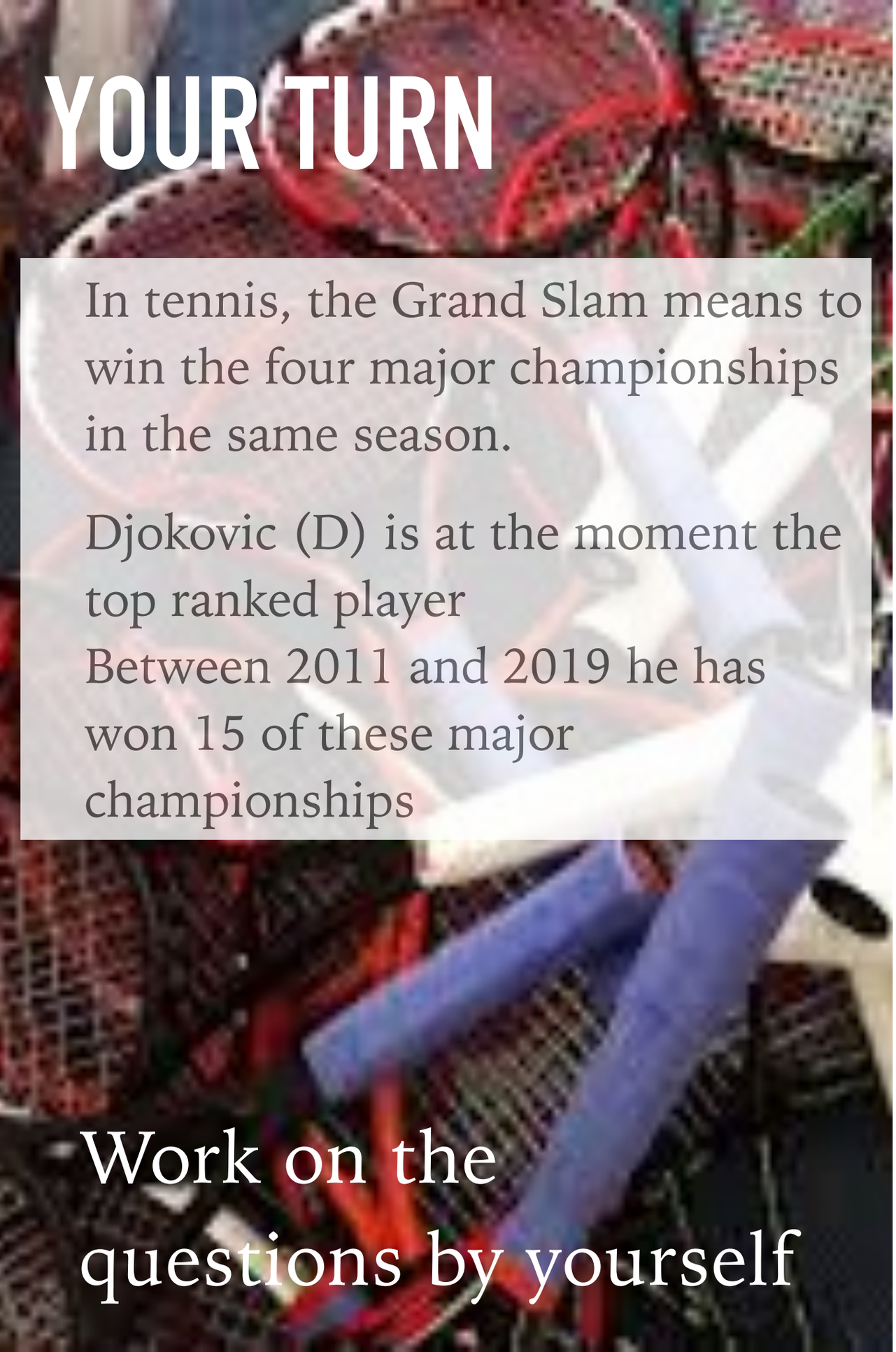
- Multiple flips of the coin:
- Sum of heads in n flips:

```
(heads10 <- SofIID(fair.coin, 10, fractions = FALSE))
Random variable with 11 outcomes
Outcomes     0     1     2     3     4     5     6     7     8     9    10
Probs    0.001 0.010 0.044 0.117 0.205 0.246 0.205 0.117 0.044 0.010 0.001
```

- Evaluate probabilities

```
P(heads10 > 3)
[1] 0.828125
```

Now we can also easily evaluate situations where not all elements in the sample space have the same probability



YOUR TURN

In tennis, the Grand Slam means to win the four major championships in the same season.

Djokovic (D) is at the moment the top ranked player
Between 2011 and 2019 he has won 15 of these major championships

Work on the questions by yourself

GRAND SLAM

- Assume Djokovic has a chance of 15/36 to win a major championship
- Set up a sample space for Djokovic's outcomes in a season (of four major championships)
- Set up a random variable W for the number of wins in a season.
- What is the probability for Djokovic to win the Grand Slam?
- Use discreteRV functionality, if you can/want to.

YOUR TURN: SOLUTION – GRAND SLAM

YOUR TURN: SOLUTION – GRAND SLAM

- Let D denote a win by Djokovic, and L a loss

YOUR TURN: SOLUTION – GRAND SLAM

- Let D denote a win by Djokovic, and L a loss
- Ω = {DDDD, DDDL, DLDD, …, LLLL}

YOUR TURN: SOLUTION – GRAND SLAM

- Let D denote a win by Djokovic, and L a loss
- Ω = {DDDD, DDDL, DLDD, …, LLLL}
- Code and results:

```
D <- RV(c(1,0), probs = c(15/36, 21/36))
```

```
(wins <- SofIID(D, n=4, fractions = FALSE))
Random variable with 5 outcomes
Outcomes     0     1     2     3     4
Probs    0.116 0.331 0.354 0.169 0.030
```

YOUR TURN: SOLUTION – GRAND SLAM

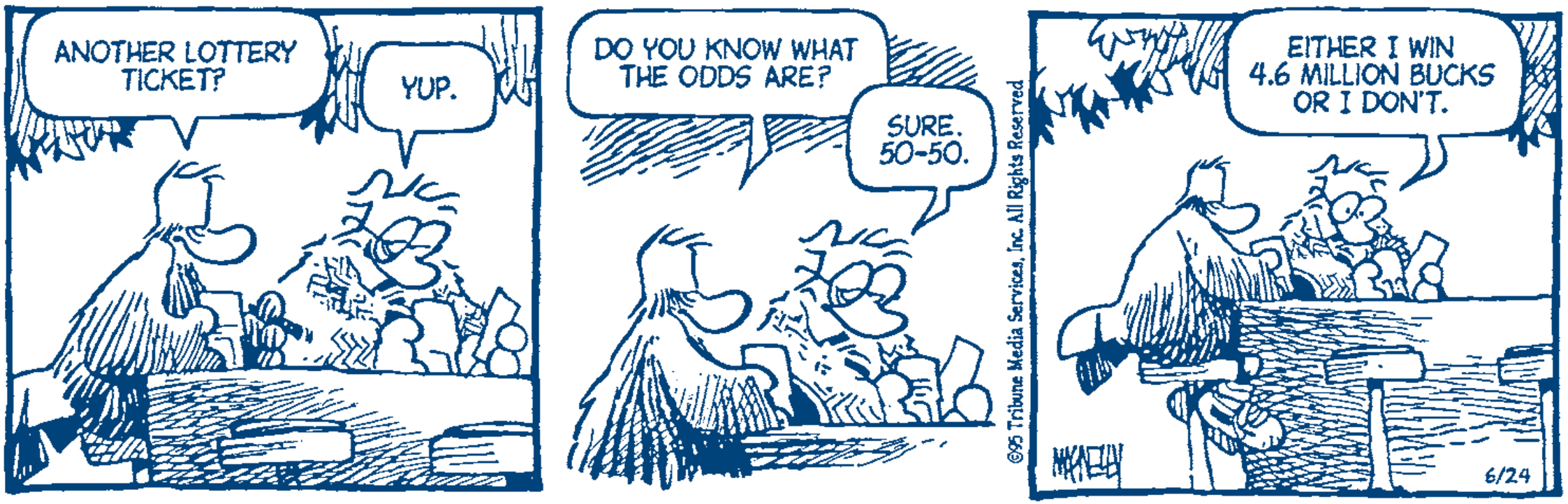
- Let D denote a win by Djokovic, and L a loss
- Ω = {DDDD, DDDL, DLDD, …, LLLL}
- Code and results:

```
D <- RV(c(1,0), probs = c(15/36, 21/36))

(wins <- SofIID(D, n=4, fractions = FALSE))
Random variable with 5 outcomes
Outcomes     0     1     2     3     4
Probs    0.116 0.331 0.354 0.169 0.030
```

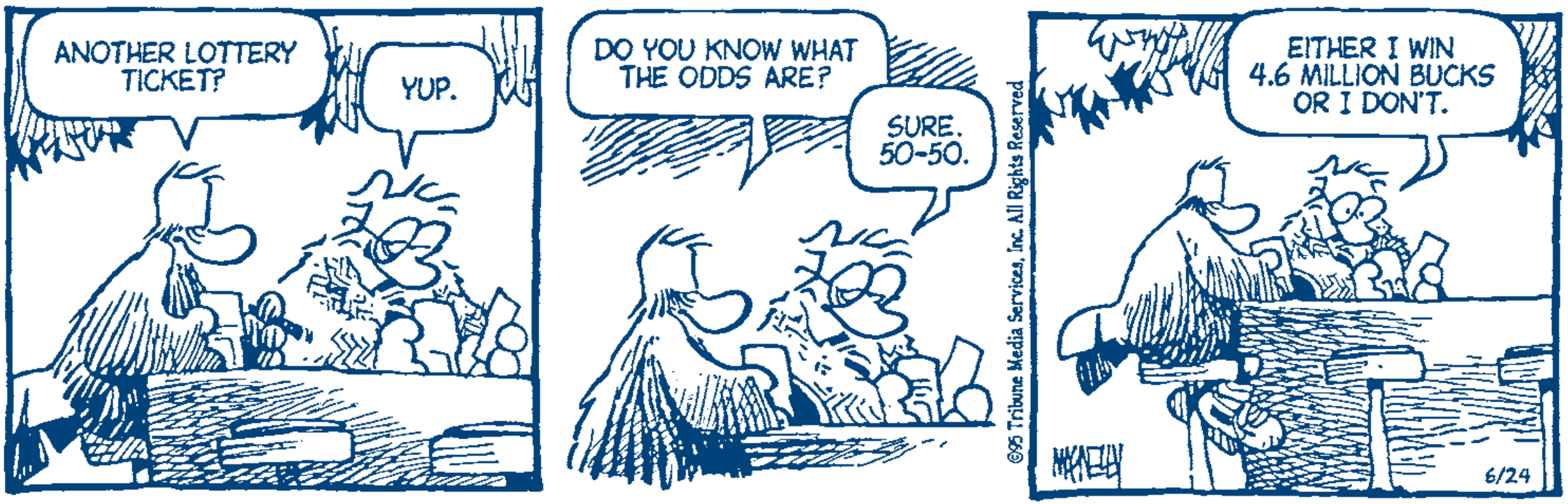
- The probability for the Grand Slam is therefore about 3% in a season.

EXPECTED VALUE



i

EXPECTED VALUE

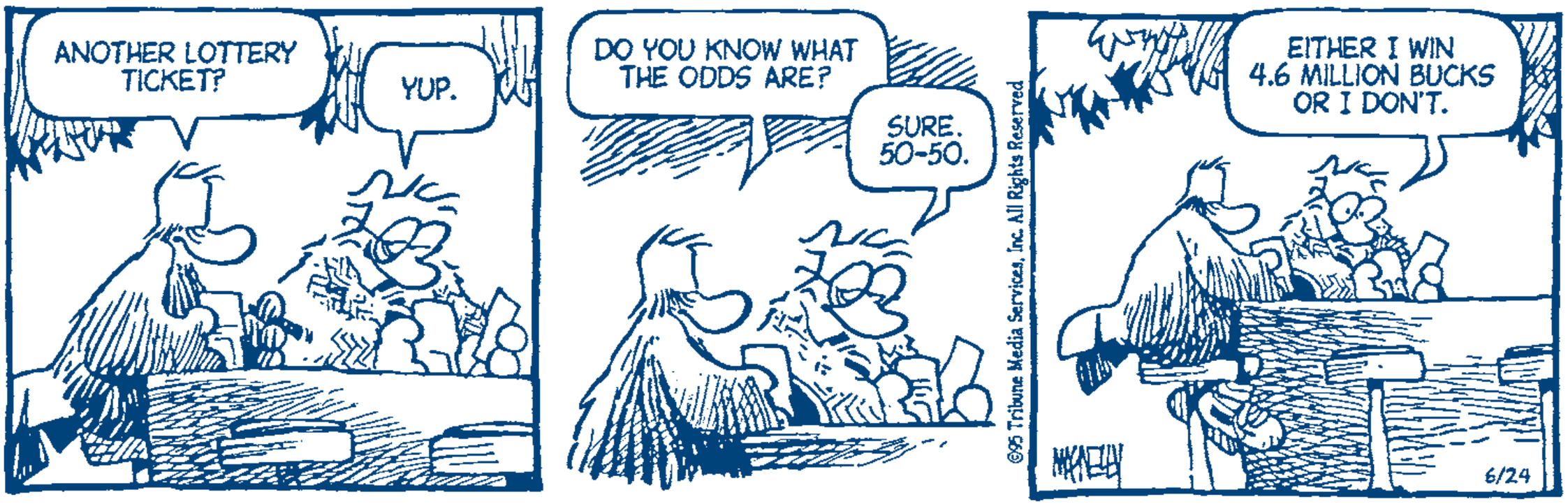


- The **expected value of a function h(X)** is defined as

$$E[h(X)] = \sum_i h(x_i) \cdot p_X(x_i)$$

i

EXPECTED VALUE



- The **expected value of a function h(X)** is defined as

$$E[h(X)] = \sum_i h(x_i) \cdot p_X(x_i)$$

- The most important version of this is $h(x) = x$, which gives **expected value of random variable X:**

$$E[X] = \sum_i x_i \cdot p_X(x_i) = \mu$$

EXAMPLES OF EXPECTED VALUES

i

EXAMPLES OF EXPECTED VALUES

- The most important version of this is $h(x) = x$, which gives **expected value of random variable X:**

$$E[X] = \sum_i x_i \cdot p_X(x_i) = \mu$$

EXAMPLES OF EXPECTED VALUES

- The most important version of this is $h(x) = x$, which gives **expected value of random variable X:**

$$E[X] = \sum_i x_i \cdot p_X(x_i) = \mu$$

- Expected value of roll of a fair die:

$$E[\text{fair die}] = 1/6 \cdot 1 + 1/6 \cdot 2 + 1/6 \cdot 3 + 1/6 \cdot 4 + 1/6 \cdot 5 + 1/6 \cdot 6 = \\ = 3.5$$

EXAMPLES OF EXPECTED VALUES

- The most important version of this is $h(x) = x$, which gives **expected value of random variable X:**

$$E[X] = \sum_i x_i \cdot p_X(x_i) = \mu$$

- Expected value of roll of a fair die:

$$E[\text{fair die}] = 1/6 \cdot 1 + 1/6 \cdot 2 + 1/6 \cdot 3 + 1/6 \cdot 4 + 1/6 \cdot 5 + 1/6 \cdot 6 = \\ = 3.5$$

- Expected value of number of wins of major tournaments

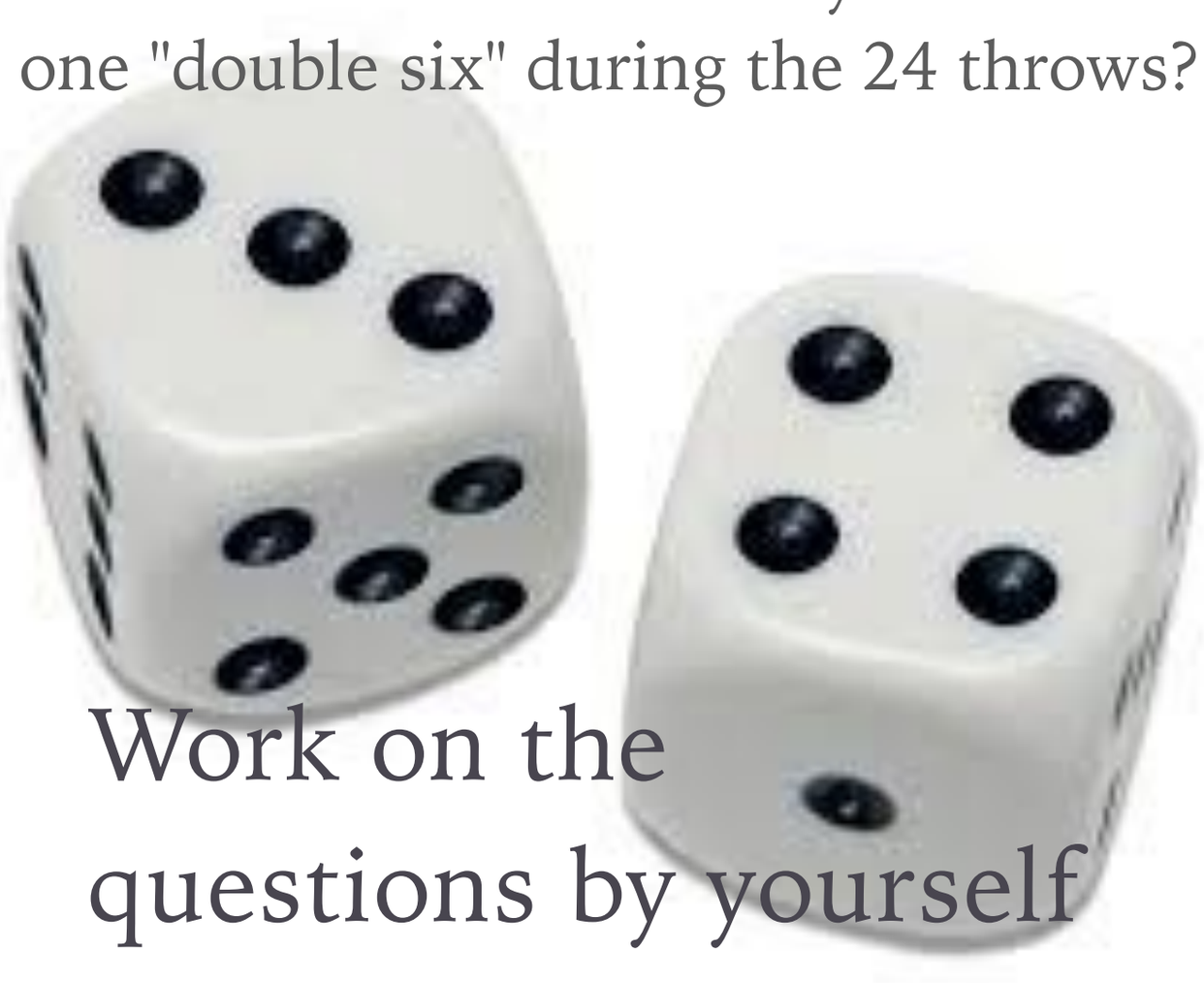
```
E(wins)
```
```
[1] 1.666667
```

YOUR TURN

Antoine Gombaud, Chevalier de Méré, French nobleman with an interest in gambling, posed the question in 1654:

Throw a pair of dice 24 times

Should one bet even money on at least one "double six" during the 24 throws?



Work on the questions by yourself

A GAMBLER'S DISPUTE

- Should the Chevalier make this bet, if each game costs 1 Franc and he will win 2 Francs if at least one double six occurs, but will not get anything otherwise?
- De Méré asked his friends Pascal and Fermat for help, you can ask R

YOUR TURN SOLUTION – A GAMBLER'S DISPUTE

YOUR TURN SOLUTION – A GAMBLER'S DISPUTE

- Rolling a double six in a pair of fair dice happens with probability 1/36:

```
doublesix <- RV(c(0,1), probs=c(35/36, 1/36))
twentyfour <- SofIID(doublesix, n = 24, fractions =
FALSE)
P(twentyfour >= 1)
[1] 0.4914039
```

YOUR TURN SOLUTION – A GAMBLER'S DISPUTE

- Rolling a double six in a pair of fair dice happens with probability 1/36:

```
doublesix <- RV(c(0,1), probs=c(35/36, 1/36))
twentyfour <- SofIID(doublesix, n = 24, fractions =
FALSE)
P(twentyfour >= 1)
[1] 0.4914039
```

- We now bring the money into the game:

```
chevalier <- RV(c(-1,rep(1, length(probs(twentyfour))-1),
                  probs = probs(twentyfour))
E(chevalier)
[1] -0.01719226
```

YOUR TURN SOLUTION – A GAMBLER'S DISPUTE

- Rolling a double six in a pair of fair dice happens with probability 1/36:

```
doublesix <- RV(c(0,1), probs=c(35/36, 1/36))
twentyfour <- SofIID(doublesix, n = 24, fractions =
FALSE)
P(twentyfour >= 1)
[1] 0.4914039
```

- We now bring the money into the game:

```
chevalier <- RV(c(-1,rep(1, length(probs(twentyfour))-1),
                  probs = probs(twentyfour))
E(chevalier)
[1] -0.01719226
```

- The overall expected value is negative, i.e. the Chevalier is losing 0.017 Francs on average in each game

A DILEMMA

- In a particular game the objective is to get as many chips as possible.
- With a special card the chance to get a chip is 10%. We can add the chances from multiple cards.
- You have 2 of these cards.
- Strategy I: play the two cards sequentially
- Strategy II: play the two cards together

INVESTIGATING A DILEMMA

INVESTIGATING A DILEMMA

- Using discreteRV we define random variables for playing cards sequentially and together

INVESTIGATING A DILEMMA

- Using discreteRV we define random variables for playing cards sequentially and together

  - ```
    one.card <- RV(c(0,1), probs=c(0.9,.1))
    (sequentially <- SofIID(one.card, n=2))
    Random variable with 3 outcomes
    Outcomes       0      1      2
    Probs    81/100   9/50  1/100
    ```
    ```

INVESTIGATING A DILEMMA

- Using discreteRV we define random variables for playing cards sequentially and together

  - ```
    one.card <- RV(c(0,1), probs=c(0.9,.1))
    (sequentially <- SofIID(one.card, n=2))
    Random variable with 3 outcomes
    Outcomes       0      1      2
    Probs    81/100   9/50  1/100
    ```

  - ```
    (together<- RV(c(0,1), probs = c(0.8, 0.2)))
    Random variable with 2 outcomes
    Outcomes   0   1
    Probs    4/5 1/5
    ```

INVESTIGATING A DILEMMA

- Using discreteRV we define random variables for playing cards sequentially and together

  - ```
    one.card <- RV(c(0,1), probs=c(0.9,.1))
    (sequentially <- SofIID(one.card, n=2))
    Random variable with 3 outcomes
    Outcomes       0      1      2
    Probs    81/100   9/50  1/100
    ```

  - ```
    (together<- RV(c(0,1), probs = c(0.8, 0.2)))
    Random variable with 2 outcomes
    Outcomes   0   1
    Probs    4/5 1/5
    ```

- Expected values

  ```
  E(sequentially)
  [1] 0.2
  E(together)
  [1] 0.2
  ```

INVESTIGATING A DILEMMA

- Using discreteRV we define random variables for playing cards sequentially and together

  - ```
    one.card <- RV(c(0,1), probs=c(0.9,.1))
    (sequentially <- SofIID(one.card, n=2))
    Random variable with 3 outcomes
    Outcomes       0      1      2
    Probs     81/100   9/50  1/100
    ```

  - ```
    (together<- RV(c(0,1), probs = c(0.8, 0.2)))
    Random variable with 2 outcomes
    Outcomes   0   1
    Probs    4/5 1/5
    ```

- Expected values

  ```
  E(sequentially)
  [1] 0.2
  E(together)
  [1] 0.2
  ```

What should we do?

VARIANCE OF A RANDOM VARIABLE

- The variance of a random variable is measuring its expected squared deviation from the mean
- The variance of a random variable X is defined as:
  $\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2] = \sum_i (x_i - \mathrm{E}[X])^2 \cdot p_X(x_i)$

  The variance is measured in squared units of X.

- $\sigma = \mathrm{Var}[X]^{1/2}$ is called the standard deviation of X, its units are the original units of X.

In the previous example:

- `V(sequentially)`
  `[1] 0.18`
- `V(together)`
  `[1] 0.16`

YOUR TURN

Work on the
questions by yourself

A FAIR COIN

- The expected value of a fair six-sided die is 3.5
- What is its variance?

YOUR TURN

Work on the
questions by yourself

A FAIR COIN

- The expected value of a fair six-sided die is 3.5
- What is its variance?

The variance is given as

$$Var[X] = \sum(x_i - E[X])^2 \cdot p_X(x_i)$$

$$Var[X] = (1 - 3.5)^2 \cdot 1/6 + (2 - 3.5)^2 \cdot 1/6 + (3 - 3.5)^2 \cdot 1/6 + (4 - 3.5)^2 \cdot 1/6 + (5 - 3.5)^2 \cdot 1/6 + (6 - 3.5)^2 \cdot 1/6 =$$

$$2.9167$$

$$sd(X) = 1.71$$

CHEBYCHEV'S INEQUALITY

- How can we use the variance?

Chebychev's Theorem

For any positive real number k, and random variable X with variance σ^2:

$$(*) \quad P(|X - E[X]| \leq k\sigma) \geq 1 - \frac{1}{k^2}$$

	k = 1	k = 2	k = 3	k=4	k = 5
(*) ≥	0	0.75	0.8889	0.9375	0.96

read as: at least 1-1/k² of a random variable's probability is within kσ of its expected value

CONTINUOUS RANDOM VARIABLE

- For a continuous random variable $X : \Omega \to R$ we define
- **Distribution function** of X

$$F_X(t) = P(X \leq t)$$

- Properties:
  - $0 \leq F_X(t) \leq 1$ for all $t \in R$
  - F_X is monotone increasing,
    i.e. for $x_1 \leq x_2$ we know $F_X(x_1) \leq F_X(x_2)$
  - $\lim_{t \to -\infty} F_X(t) = 0$ and $\lim_{t \to +\infty} F_X(t) = 1.$

- **Density function** of X is the derivative of the distribution function:

$$f_X(t) = F_X'(t)$$

  - Properties: $f_X(t) \geq 0$ for all t and $1 = \int f_X(t)\, dt$

CONTINUOUS RANDOM VARIABLE

- For a continuous random variable X : $\Omega \rightarrow$ R we define
- Expected value of h(X)

$$E[h(X)] = \int_x h(x) \cdot f_X(x)$$

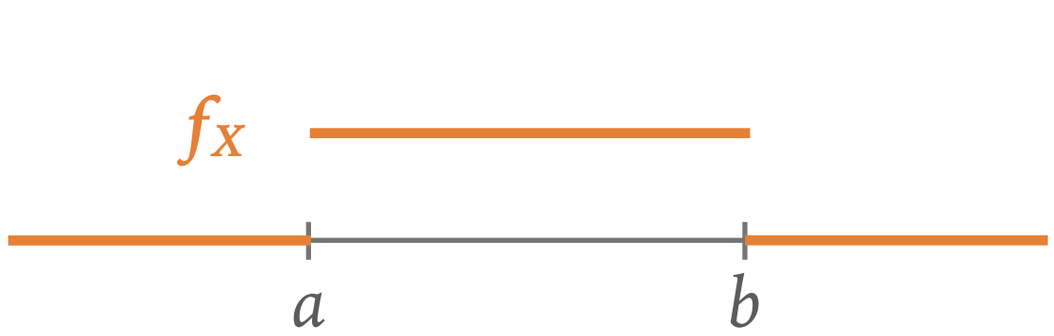
- Variance of X

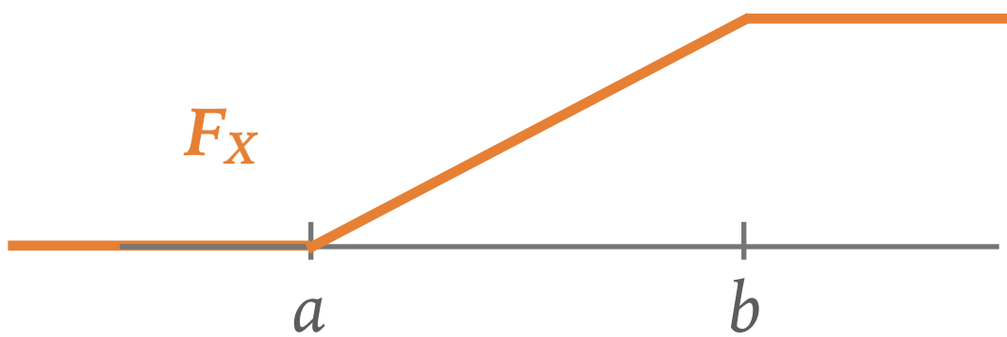
$$\begin{aligned} Var[X] &= E[(X - E[X])^2] = \\ &= \int_{-\infty}^{\infty} (x - E[X])^2 f_X(x) dx \end{aligned}$$

with that: Var[X] =E[X²] -E[X]² (mixed term gives -2E[X²])

A SPECIAL DISTRIBUTION: UNIFORM

- All values in interval have the same chance of being selected
  e.g. random number generation, sampling, …
- Define the uniform density on interval [a, b] as:
  $f(x) = 1/(b-a)$ for $x \in [a, b]$ and $f(x) = 0$ everywhere else





- What is E[X]? What is Var[X]?

$$E[X] = \int x\, f(x)dx = \frac{(a-b)^2}{2(b-a)} = \frac{a+b}{2} = \mu$$

$$Var[X] = \int_a^b (x-\mu)^2 f(x)dx = \int_{\mu-b}^{b-\mu} u^2 \frac{du}{b-a} = \frac{(b-a)^2}{12}$$

YOUR TURN

For all major distributions R provides set of functions of the form rXXX, dXXX, pXXX, and qXXX

Where rXXX creates random numbers from distribution XXX, dXXX is the density, pXXX is the distribution, and qXXX is the quantile function.

Work on the questions by yourself

RANDOM NUMBERS

- The function runif in R is parameterized as `runif(n, min = 0, max = 1)` where `n` is the number of random numbers between `min` and `max`.
- Create a vector of 1000 uniform random numbers between 0 and 10 and save the result as `x`.
- Give a summary of the vector (with the function `summary`) and check how far the values are away from what they are supposed to be theoretically (with `qunif`)

YOUR TURN SOLUTION - RANDOM NUMBERS

```
set.seed(10111213)


y <- runif(1000, min = 0, max = 10)
summary(y)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
0.001642 2.765500 5.304320 5.190713 7.688299 9.999371 
```

```
qunif(c(0, 0.25, 0.5, 0.75, 1), min = 0, max = 10)
```

```
[1]  0.0  2.5  5.0  7.5 10.0
```

```
# mean and median are the same values in a symmetric
# density function
```

QUESTIONS?

RELATIONSHIPS BETWEEN MULTIPLE RANDOM VARIABLES

JOINT DISTRIBUTION OF X AND Y

JOINT DISTRIBUTION OF X AND Y

- The joint distribution of X and Y of two discrete RVs is defined as

$$p_{XY}(x, y) = P(X = x \cap Y = y)$$

JOINT DISTRIBUTION OF X AND Y

- The joint distribution of X and Y of two discrete RVs is defined as

$$p_{XY}(x, y) = P(X = x \cap Y = y)$$

- For two continuous RVs X and Y the joint distribution is

$$F_{XY}(x, y) = P(X \leq x \cap Y \leq y)$$

JOINT DISTRIBUTION OF X AND Y

- The joint distribution of X and Y of two discrete RVs is defined as

$$p_{XY}(x, y) = P(X = x \cap Y = y)$$

- For two continuous RVs X and Y the joint distribution is

$$F_{XY}(x, y) = P(X \leq x \cap Y \leq y)$$

- **Example:**
  The General Social Survey is a bi-annual survey by NORC of about 2500 Americans asking about current issues.
  In 2016 Americans were asked about their happiness level (not too happy, pretty happy, very happy).

HOW HAPPY ARE YOU?

HOW HAPPY ARE YOU?

- Let G be gender, and H happiness level

	Female	Male
not too happy	261	191
pretty happy	883	718
very happy	442	364

HOW HAPPY ARE YOU?

- Let G be gender, and H happiness level

	Female	Male
not too happy	261	191
pretty happy	883	718
very happy	442	364

- Joint PMF of G and H

p_{GH}	Female	Male	p_H
not too happy	0.09	0.07	0.16
pretty happy	0.31	0.25	0.56
very happy	0.15	0.13	0.28
p_G	0.55	0.45	1.00

HOW HAPPY ARE YOU?

- Let G be gender, and H happiness level

	Female	Male
not too happy	261	191
pretty happy	883	718
very happy	442	364

- Joint PMF of G and H

p_{GH}	Female	Male	p_H
not too happy	0.09	0.07	0.16
pretty happy	0.31	0.25	0.56
very happy	0.15	0.13	0.28
p_G	0.55	0.45	1.00

marginal distributions

INDEPENDENCE

INDEPENDENCE

- Random variables X and Y are independent, if:

$$P(X = x \cap Y = y) = P(X = x) \cdot P(Y = y)$$

  for all x and y

INDEPENDENCE

- Random variables X and Y are independent, if:

$$P(X = x \cap Y = y) = P(X = x) \cdot P(Y = y)$$

  for all x and y

INDEPENDENCE

- Random variables X and Y are independent, if:

$$P(X = x \cap Y = y) = P(X = x) \cdot P(Y = y)$$

for all x and y

- For random variables X and Y and a, b $\in$ R:

$$Var[aX + bY] = a^2\,Var[X] + b^2\,Var[Y] + 2ab\,Cov(X,Y)$$

where $Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$ is the **covariance between X and Y**

IS HAPPINESS INDEPENDENT OF GENDER?

p_{GH}	Female	Male	p_H
not too happy	$0.09 =?= 0.16 \cdot 0.55$	$0.07 =?= 0.16 \cdot 0.45$	0.16
pretty happy	$0.31 =?= 0.56 \cdot 0.55$	$0.25 =?= 0.56 \cdot 0.45$	0.56
very happy	$0.15 =?= 0.28 \cdot 0.55$	$0.13 =?= 0.28 \cdot 0.45$	0.28
p_G	0.55	0.45	1.00

IS HAPPINESS INDEPENDENT OF GENDER?

- Is $P(G = g \cap H = h) = P(G = g) \cdot P(H = h)$ for all h, g?

p_{GH}	Female	Male	p_H
not too happy	$0.09 =?= 0.16 \cdot 0.55$	$0.07 =?= 0.16 \cdot 0.45$	0.16
pretty happy	$0.31 =?= 0.56 \cdot 0.55$	$0.25 =?= 0.56 \cdot 0.45$	0.56
very happy	$0.15 =?= 0.28 \cdot 0.55$	$0.13 =?= 0.28 \cdot 0.45$	0.28
p_G	0.55	0.45	1.00

IS HAPPINESS INDEPENDENT OF GENDER?

- Is $P(G = g \cap H = h) = P(G = g) \cdot P(H = h)$ for all h, g?
- compare by calculating differences

p_{GH}	Female	Male	p_H
not too happy	$0.09 =?= 0.16 \cdot 0.55$	$0.07 =?= 0.16 \cdot 0.45$	0.16
pretty happy	$0.31 =?= 0.56 \cdot 0.55$	$0.25 =?= 0.56 \cdot 0.45$	0.56
very happy	$0.15 =?= 0.28 \cdot 0.55$	$0.13 =?= 0.28 \cdot 0.45$	0.28
p_G	0.55	0.45	1.00

IS HAPPINESS INDEPENDENT OF GENDER?

- Is $P(G = g \cap H = h) = P(G = g) \cdot P(H = h)$ for all h, g?
- compare by calculating differences

p_{GH}	Female	Male	p_H
not too happy	0.09 =?= 0.087	0.07 =?= 0.07	0.16
pretty happy	0.31 =?= 0.309	0.25 =?= 0.248	0.56
very happy	0.15 =?= 0.156	0.13 =?= 0.125	0.28
p_G	0.55	0.45	1.00

IS HAPPINESS INDEPENDENT OF GENDER?

- Is $P(G = g \cap H = h) = P(G = g) \cdot P(H = h)$ for all h, g?
- compare by calculating differences
- are different results **'similar enough'**?

p_{GH}	Female	Male	p_H
not too happy	0.09 =?= 0.087	0.07 =?= 0.07	0.16
pretty happy	0.31 =?= 0.309	0.25 =?= 0.248	0.56
very happy	0.15 =?= 0.156	0.13 =?= 0.125	0.28
p_G	0.55	0.45	1.00

CHISQUARE TABLE TEST OF INDEPENDENCE

CHISQUARE TABLE TEST OF INDEPENDENCE

- Let X and Y be two discrete random variables with I and J levels.
- **Null hypothesis**
  H_o: X and Y are independent.

  Alternative hypothesis
  H_1: X and Y are not independent
- The test statistic $C = \sum_{ij} (o_{ij} - e_{ij})^2 / e_{ij}$ is used to evaluate the difference between observed values o_{ij} and expected values: $e_{ij} = p_X(i) \cdot p_Y(j)$
- C follows a χ_c^2 distribution with $c = (I-1)\cdot(J-1)$
- in R: `chisq.test`

YOUR TURN

The critical values for a chi square distribution with 2 degrees of freedom are:

90%	*4.61*
95%	*5.99*
99%	*9.21*

Work on the questions by yourself

HAPPINESS

- Based on the numbers above calculate a chi-square statistic.
- Compare your chi-square statistic to the critical values on the left and interpret your finding.
- Characterize deviation from independence

YOUR TURN

The critical values for a chi square distribution with 2 degrees of freedom are:

90%	*4.61*
95%	*5.99*
99%	*9.21*

Work on the questions by yourself

HAPPINESS

	Age < 60	60+
not too happy	315	135
pretty happy	1145	452
very happy	533	269

- Based on the numbers above calculate a chi-square statistic.
- Compare your chi-square statistic to the critical values on the left and interpret your finding.
- Characterize deviation from independence

YOUR TURN SOLUTION - HAPPINESS

```
if (!require(classdata)) {
  devtools::install_github("heike/classdata")
}
library(classdata)
library(tidyverse)

happy16 <- happy %>% dplyr::filter(year==2016)
dframe <- happy16 %>%
  mutate(age60 = age < 60) %>%
  select(age60, happy) %>% na.omit()

with(dframe, chisq.test(age60, happy))
Pearson's Chi-squared test

data:  age60 and happy
X-squared = 6.97, df = 2, p-value = 0.03065
```

PROPERTIES OF EXPECTED VALUES AND VARIANCES

PROPERTIES OF EXPECTED VALUES AND VARIANCES

- For random variables X and Y and a, b $\in$ R:

$$E[aX + bY] = a\,E[X] + b\,E[Y]$$

  i.e. expected values are linear functions

PROPERTIES OF EXPECTED VALUES AND VARIANCES

- For random variables X and Y and a, b $\in$ R:

$$E[aX + bY] = a\,E[X] + b\,E[Y]$$

  i.e. expected values are linear functions

- For random variable X and a $\in$ R:

$$\mathrm{Var}(aX) = a^2\,\mathrm{Var}(X)$$

$$\mathrm{sd}(X) = a\,\mathrm{sd}(X)$$

PROPERTIES OF EXPECTED VALUES AND VARIANCES

- For random variables X and Y and a, b $\in$ R:

$$E[aX + bY] = a\,E[X] + b\,E[Y]$$

  i.e. expected values are linear functions

- For random variable X and a $\in$ R:

$$Var(aX) = a^2\,Var(X)$$

$$sd(X) = a\,sd(X)$$

- For random variables X and Y and a, b $\in$ R:

$$Var[aX + bY] = a^2\,Var[X] + b^2\,Var[Y] + 2ab\,Cov(X,Y)$$

  where $Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$ is the **covariance between X and Y**

CORRELATION COEFFICIENT

CORRELATION COEFFICIENT

- **Correlation between X and Y** is defined as
  $r(X,Y) = \mathrm{Corr}(X,Y) = \mathrm{Cov}(X, Y)/(\mathrm{Var}(X)\cdot\mathrm{Var}(Y))$

CORRELATION COEFFICIENT

- **Correlation between X and Y** is defined as
  $r(X,Y) = Corr(X,Y) = Cov(X, Y)/(Var(X) \cdot Var(Y))$

- Properties
  - $-1 \leq r(X,Y) \leq 1$ for all X, Y

CORRELATION COEFFICIENT

- **Correlation between X and Y** is defined as
  $r(X,Y) = Corr(X,Y) = Cov(X, Y)/(Var(X) \cdot Var(Y))$

- Properties
  - $-1 \leq r(X,Y) \leq 1$ for all X, Y
  - If X, Y are independent, then $r(X,Y) = 0$
    ! The reverse is not true !
    i.e. if $r(X,Y) = 0$, X and Y are NOT always independent!

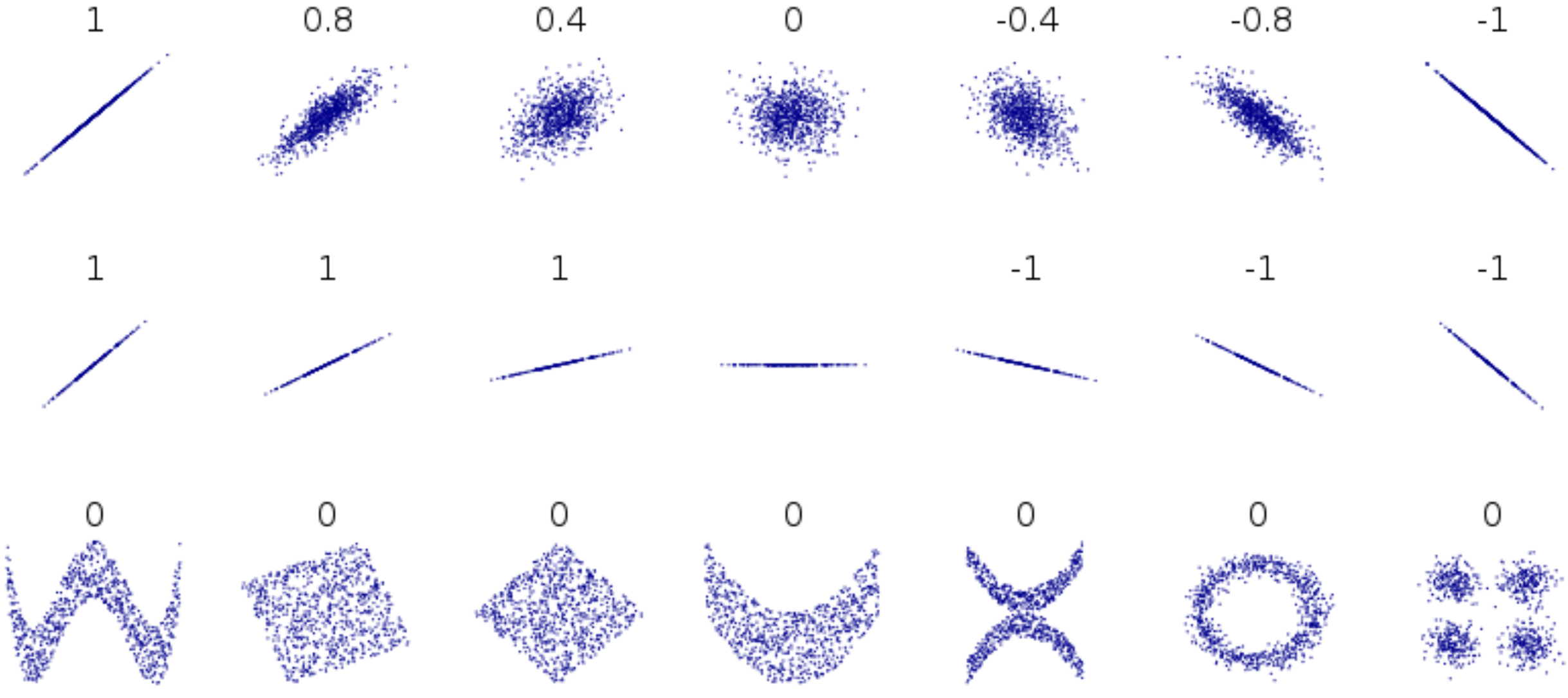
CORRELATION COEFFICIENT

- **Correlation between X and Y** is defined as
  $r(X,Y) = Corr(X,Y) = Cov(X, Y)/(Var(X) \cdot Var(Y))$

- Properties
  - $-1 \leq r(X,Y) \leq 1$ for all X, Y
  - If X, Y are independent, then $r(X,Y) = 0$
    ! The reverse is not true !
    i.e. if $r(X,Y) = 0$, X and Y are NOT always independent!

    *The correlation coefficient is a measure for the amount of **linear dependence** between X and Y*

Correlation



BEWARE CAUSALITY!

BEWARE CAUSALITY!

- Statistics only rarely allows us to say something about causal relationships.

BEWARE CAUSALITY!

- Statistics only rarely allows us to say something about causal relationships.
- Instead, statistical models usually only allow us to say something about correlation or co-occurrence, i.e.: with an increase in X we also see an increase in Y.

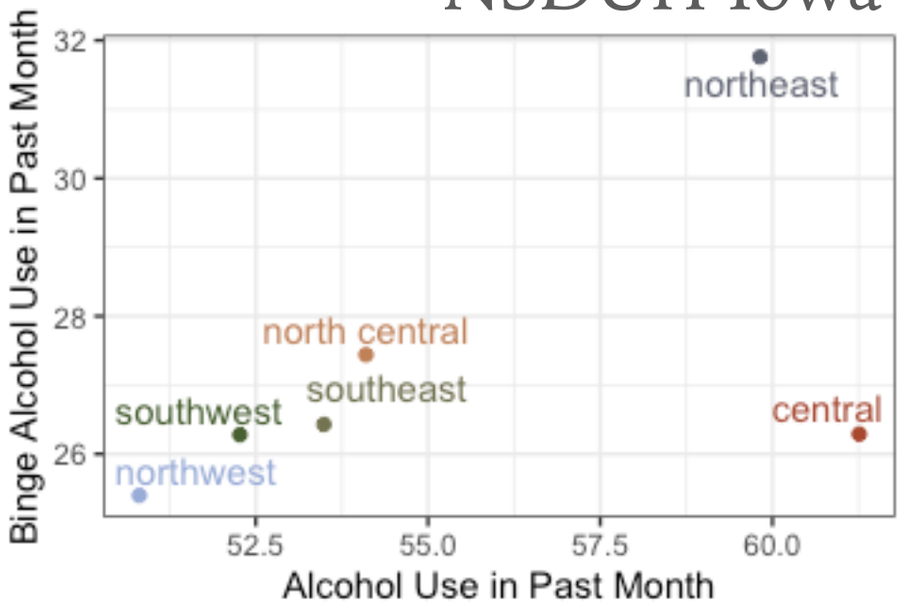
BEWARE CAUSALITY!

- Statistics only rarely allows us to say something about causal relationships.
- Instead, statistical models usually only allow us to say something about correlation or co-occurrence, i.e.: with an increase in X we also see an increase in Y.
- That is different from saying that Y increases when we increase X.

BEWARE CAUSALITY!

- Statistics only rarely allows us to say something about causal relationships.
- Instead, statistical models usually only allow us to say something about correlation or co-occurrence, i.e.: with an increase in X we also see an increase in Y.
- That is different from saying that Y increases when we increase X.

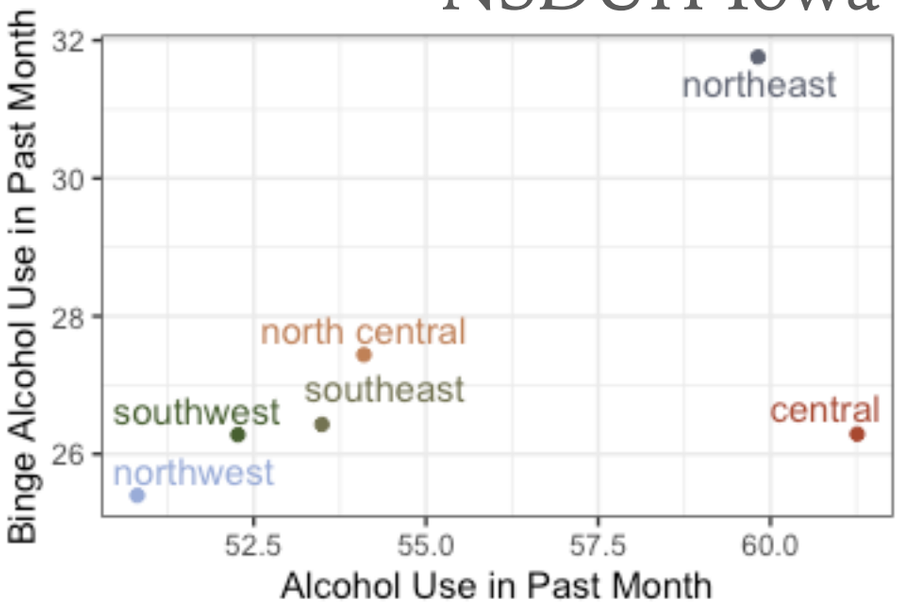
NSDUH Iowa

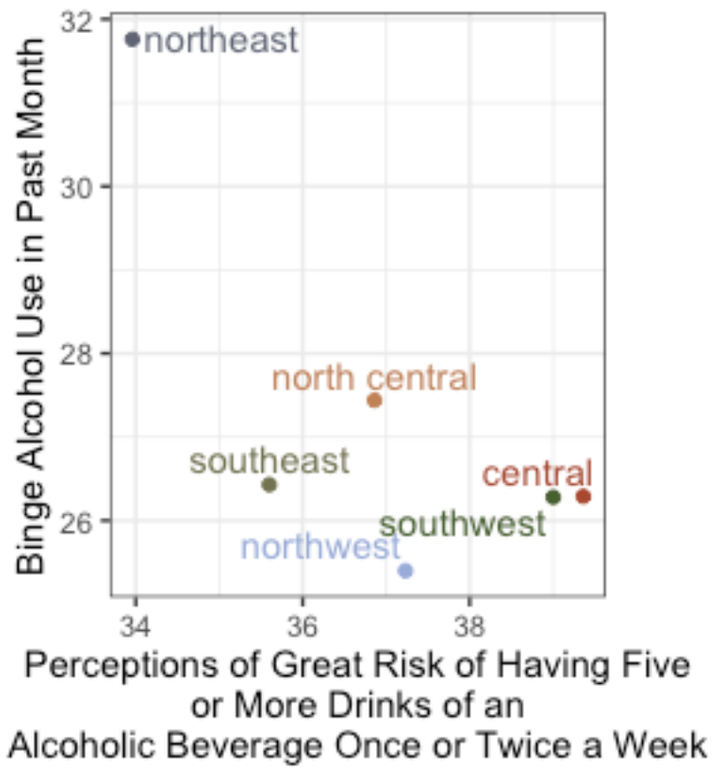


BEWARE CAUSALITY!

- Statistics only rarely allows us to say something about causal relationships.
- Instead, statistical models usually only allow us to say something about correlation or co-occurrence, i.e.: with an increase in X we also see an increase in Y.
- That is different from saying that Y increases when we increase X.

NSDUH Iowa





QUESTIONS?

We'll take a 5-minute break now

5 MINUTE BREAK

4

3

2

1

We'll take a 5-minute break now

. . . READY?

BINOMIAL DISTRIBUTION

BERNOULLI EXPERIMENT

BERNOULLI EXPERIMENT

- **Bernoulli Experiment:**
  Random experiment with only two outcomes: Success/Failure
  P(success) = p

BERNOULLI EXPERIMENT

- **Bernoulli Experiment:**
  Random experiment with only two outcomes: Success/Failure
  $P(\text{ success }) = p$

- **Random Variable X** defined as

$$X(\text{Success}) = 1 \quad \text{and} \quad X(\text{Failure}) = 0$$

- **PMF:** $p_X(0) = 1 - p$ and $p_X(1) = p$

- $E[X] = 0 \cdot (1 - p) + 1 \cdot p = p$
  $\text{Var}[X] = (0-p)^2 \cdot (1-p) + (1-p)^2 \cdot p = p(1-p)$

YOUR TURN

A target consists of a small yellow area in the middle and a ring of other colors around it.
Assume that the yellow area in the middle is 1/10 of the area around it

Work on the questions by yourself

BULLSEYE

- A novice shooter aims for the bullseye but only manages to hit the target area at random, i.e. her chances of hitting the yellow bull's eye is 1/10.
- Assume there are three independent attempts (she also does not learn that quickly) in which the target is hit.
- Let X be the random variable counting the number of hits of the yellow area.
- What is the pmf? What is E[X]? What is Var[X]?

YOUR TURN SOLUTION – BULLSEYE

YOUR TURN SOLUTION – BULLSEYE

- We could cheat and use R:

```
shot <- RV(c(1,0), probs = c(1/10, 9/10))
(three.shots <- SofIID(shot, n=3))
Random variable with 4 outcomes
Outcomes          0          1          2          3
Probs    729/1000 243/1000  27/1000   1/1000
```

```
E(three.shots)
[1] 0.3

V(three.shots)
[1] 0.27
```

YOUR TURN SOLUTION – BULLSEYE (2)

YOUR TURN SOLUTION – BULLSEYE (2)

- **By hand:**
  we can have a look at the three shots separately as
  $S = S_1 + S_2 + S_3$
  where each S_i is a Bernoulli RV with P(success) $= 1/10$

- $E[S_i] = 1/10$ and $E[S] = E[S_1 + S_2 + S_3] = 3/10$

- $Var[S_i] = 9/100$ and $Var[S] = Var[S_1 + S_2 + S_3] = 27/100$

- PMF of S:
  $p_S(0) = P(S_1 = 0, S_2 = 0, S_3 = 0) = 9^3/10^3 = 0.729$
  $p_S(1) = 3 \cdot 9^2/10^3 = 0.243$
  $p_S(2) = 3 \cdot 9^1/10^3 = 0.027$
  $p_S(3) = 1/1000$

BINOMIAL DISTRIBUTION

- **Situation**: n sequential Bernoulli experiments, with success rate p for a single trial.
  Single trials are independent from each other.
- We are interested in the **number X of successes in n trials**:
- Sample Space $\Omega = \{0,1,2,...,n\}$
- **PMF for X**:
  $p_X(k) = P(X = k)$ for all possible $k = 0,...,n$.
  i.e. want to find probability that in a sequence of n trials there are exactly k successes.

BINOMIAL DISTRIBUTION (2)

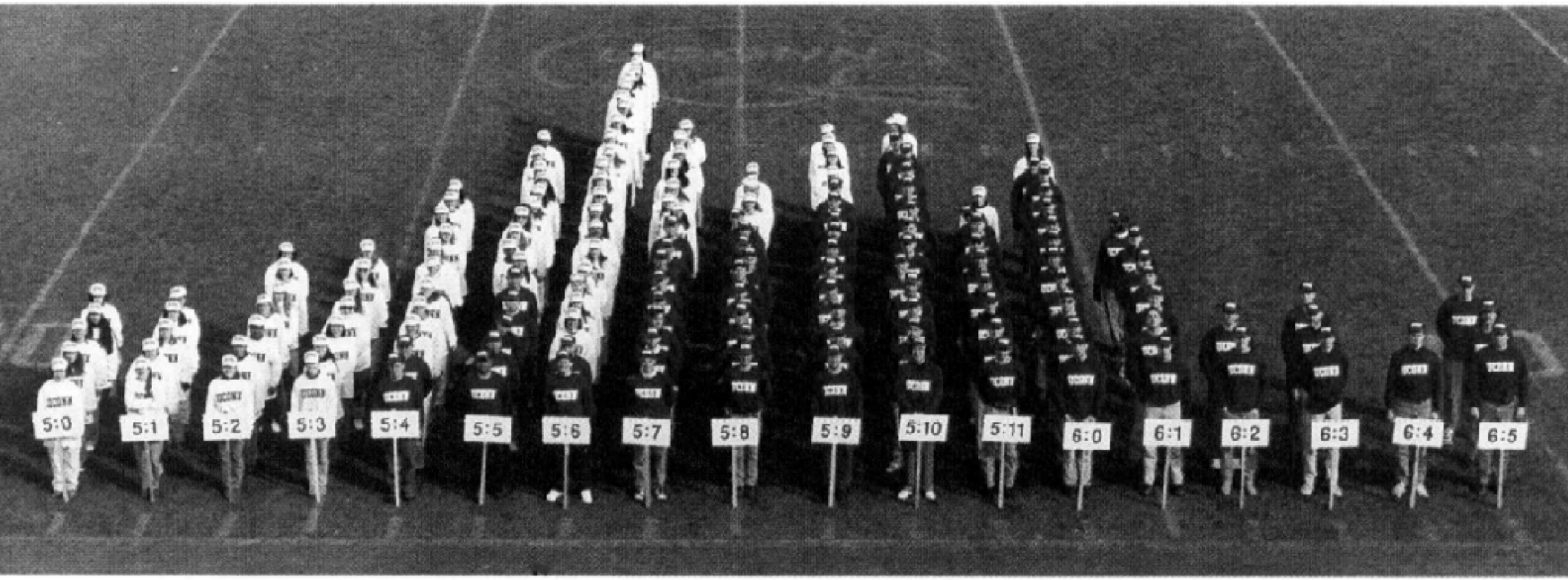
- Probability that in a sequence of n trials there are exactly k successes:
- Probability for sequence s consisting of k successes first, followed by n − k failures $P(s) = p^k(1-p)^{n-k}$.
- How many possibilities are there to have k successes in n trials?
  We can think of this as choosing k spots for successes among n trials.
  The number of options is $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
- This gives a pmf of $p_X(k) = \binom{n}{k} p^k(1-p)^{n-k}$
- E[X] = np Var[X] = np(1-p)

BINOMIAL DISTRIBUTION IN R

- The name of the binomial distribution in R is `binom`
- `rbinom(n, size, prob)`
  `size` is the number of trials, `prob` is the probability of success of a single trial
- `dbinom` is the density, `pbinom` is the distribution, and `qbinom` is the quantile

- Examples using Binomial?
  A box contains 15 components that each have a failure rate of 2%. What is the probability that
- 1. exactly two out of the fifteen components are defective?
  2. at most two components are broken?
  3. more than three components are broken?
  4. more than 1 but less than 4 are broken?

- Let X be the number of broken components. Then X has a B15,0.02 distribution.
- P(exactly two out of the fifteen components are defective)
  $=$ dbinom(2, 15, 0.02) $=$ 0.0323.
- P(at most two components are broken) $=$ P(X $\leq$ 2) $=$ pbinom(2, 15, 0.02) $=$ 0.9670
- P(more than three components are broken)$=$P(X $>$3)$=$1$-$P(X $\leq$3)$=$ 1-pbinom(3, 15, 0.02) $=$ 1$-$0.9998$=$0.0002.
- Alternatively, we could have used pbinom(3, 15, 0.02, lower.tail$=$FALSE)



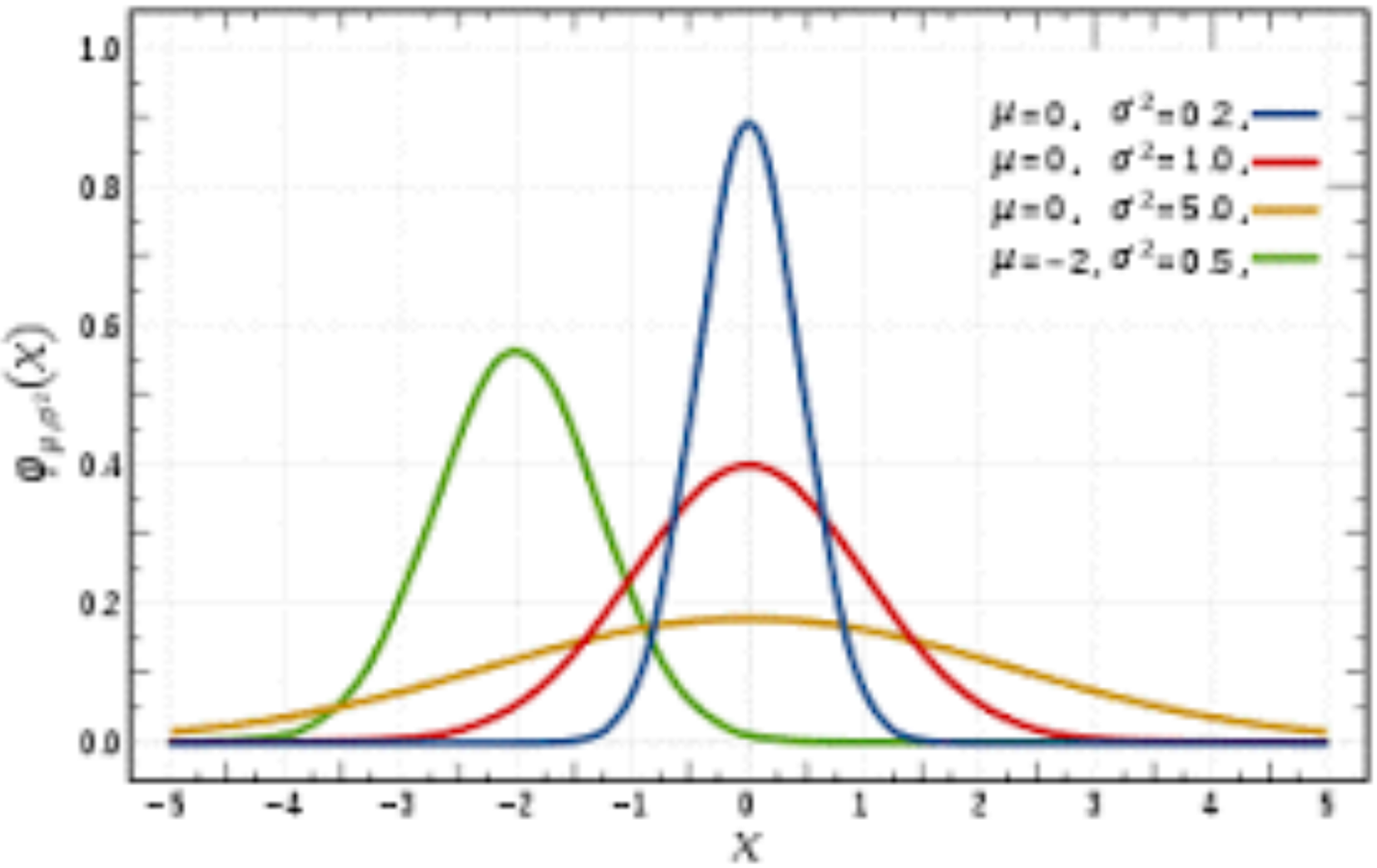
NORMAL DISTRIBUTION

NORMAL DISTRIBUTION

- Archetypical bell-shaped density curve
- Two parameters:
  μ and σ
- μ is the location,
  σ is the scale &
  determines ‘flatness’
- density function

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)}{2\sigma^2}}$$

- E[X] = μ and Var[X] = σ^2



NORMAL DISTRIBUTION

- For a normally distributed random variable X, $X \sim N_{\mu, \sigma 2}$ the distribution function is

$$P(X \leq t) = F_{\mu,\sigma^2}(t) = \int_{-\infty}^{t} f_{\mu,\sigma^2}(x)dx$$

- integral doesn’t have a closed form, i.e. we need lookup tables or software :) to use the normal distribution

NORMAL DISTRIBUTION IN R

- The name of the binomial distribution in R is `norm`
- `rnorm(n, mean = 0, sd = 1)`
  `mean` is the location parameter, `sd` is the scale parameter (standard deviation)
- `dnorm` is the density, `pnorm` is the distribution, and `qnorm` is the quantile

EXAMPLE: CALCULATE PROBABILITIES FROM A NORMAL DISTRIBUTION

- In 2018, total SAT scores had a mean of 1068 and standard deviation of 204 points.
  https://nces.ed.gov/programs/digest/d18/tables/dt18_226.40.asp?current=yes

- Approximate the distribution of SAT scores by a normal distribution with that mean and standard deviation.
  Alternatively, use that $$Z = \frac{\text{SAT} - \mu}{\sigma}$$
  follows a standard normal distribution ($\mu = 0$, $\sigma = 1$).

- What is the probability (for a student in 2018) to have a score
  … between 800 and 1200?
  … greater than 1400?

APPROXIMATELY NORMAL DISTRIBUTION OF SAT SCORES

- We compute the Z-values:
- $\frac{800 - 1068}{204} = \mathbf{-1.31}$ $\frac{1200 - 1068}{204} = \mathbf{0.65}$ $\frac{1400 - 1068}{204} = \mathbf{1.63}$
- Then we use the standard normal distribution function F:
- $P(800<SAT<1200) = F(\mathbf{0.65}) - F(\mathbf{-1.31})$
  $= 0.742 - 0.095 = 0.647.$
- $P(1400 < SAT) = 1 - P(SAT<1400) = 1 - F(\mathbf{1.63}) = 0.052.$

PROPAGATING PROPERTY OF NORMALS

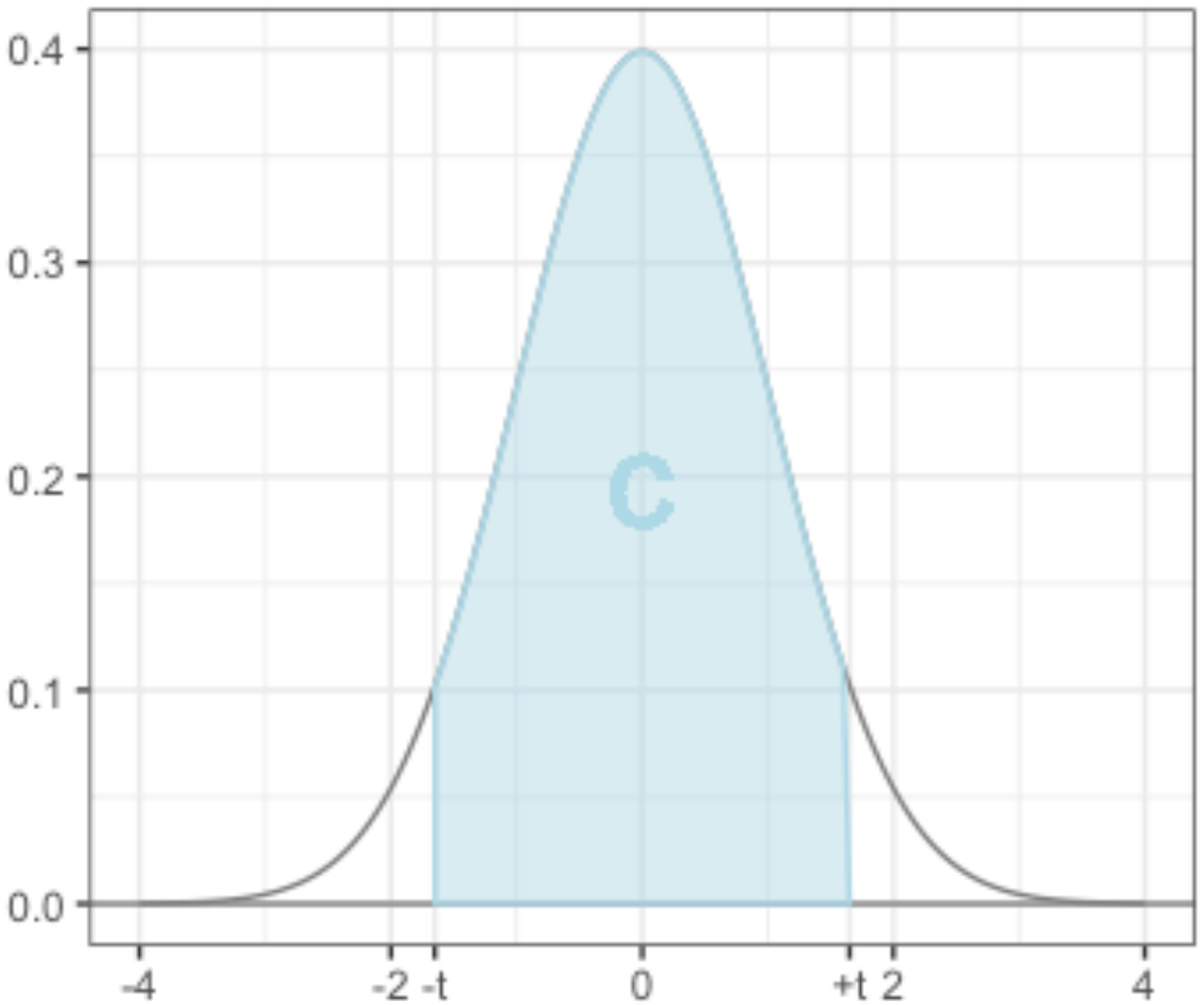
- Let X and Y be normal RVs and a, b $\in$ R, then
  Z = aX + bY is also normal RV with parameters μ_Z and σ_Z^2

  $\mu_Z = E[Z] = E[aX + bY] = a\,E[X] + b\,E[Y]$
  $\sigma_Z^2 = Var[Z] = a^2\,Var[X] + b^2\,Var[Y] + 2ab\,Cov(X,Y)$

- Normalize RV X. For any normal RV $X \sim N_{\mu,\,\sigma^2}$ the normalized RV Z defined as

$$Z = (X - \mu) / \sigma \sim N(0,1)$$

YOUR TURN



CRITICAL VALUES OF N(0,1)

- Find values (-t, t) such that for $X \sim N(0,1)$

$$P(-t \leq X \leq t) = c$$

where $c = 0.9$, 0.95, and 0.99

These values for c are called critical values of the standard normal distribution.

Work on the questions by yourself

YOUR TURN SOLUTION – CRITICAL VALUES OF STANDARD NORMAL

- Find t such that $c = P(-t \leq X \leq t)$ for some values of c
- $P(-t \leq X \leq t) = F_X(t) - F_X(-t) =$ f_X is symmetric around 0

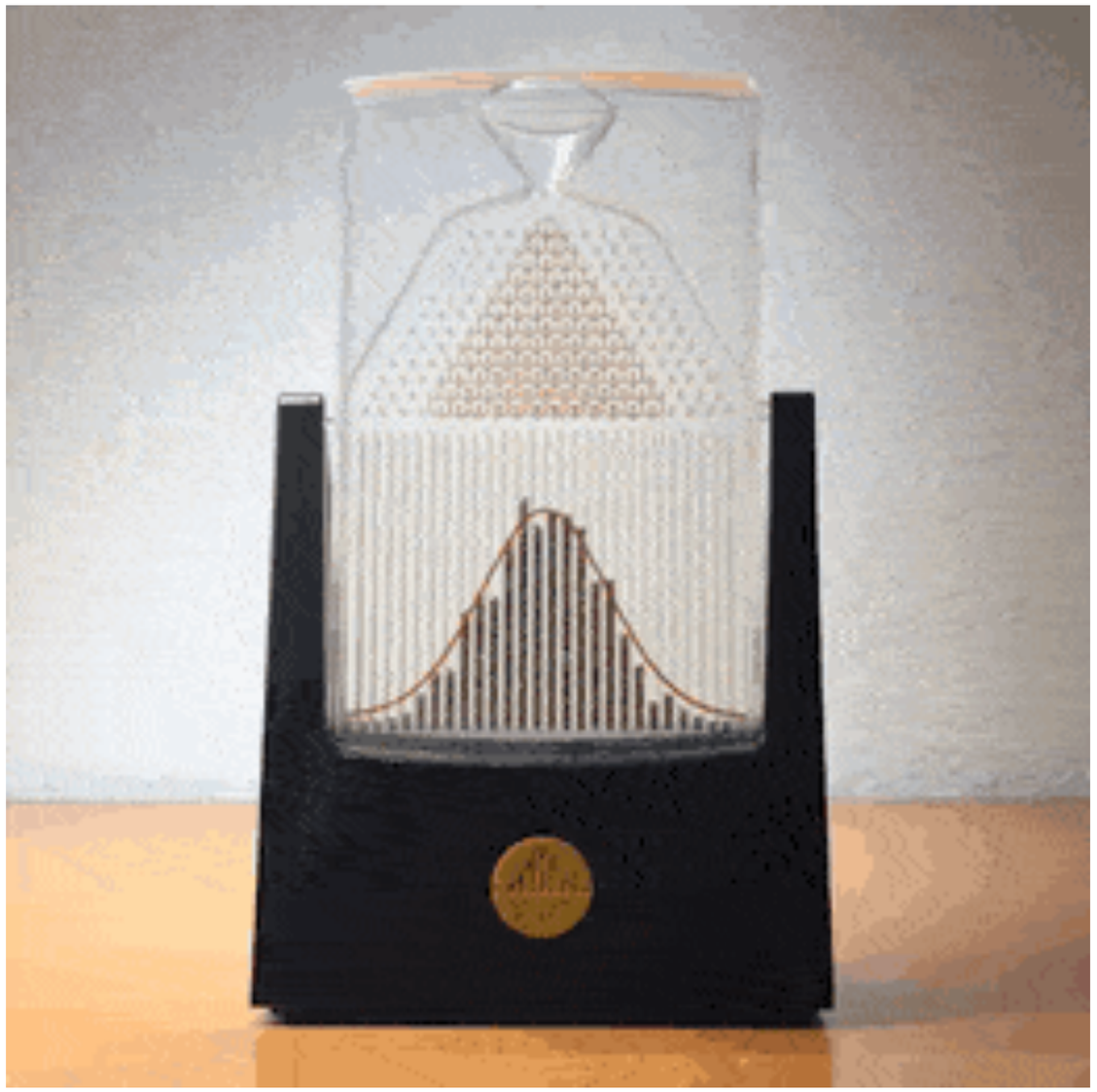
  $= F_X(t) - (1 - F_X(t)) =$

  $= 2\,F_X(t) - 1$
- $c = 2\,F_X(t) - 1$, therefore $F_X(t) = 0.5(c + 1)$

  $t = F_X^{-1}(0.5(c + 1))$ quantile

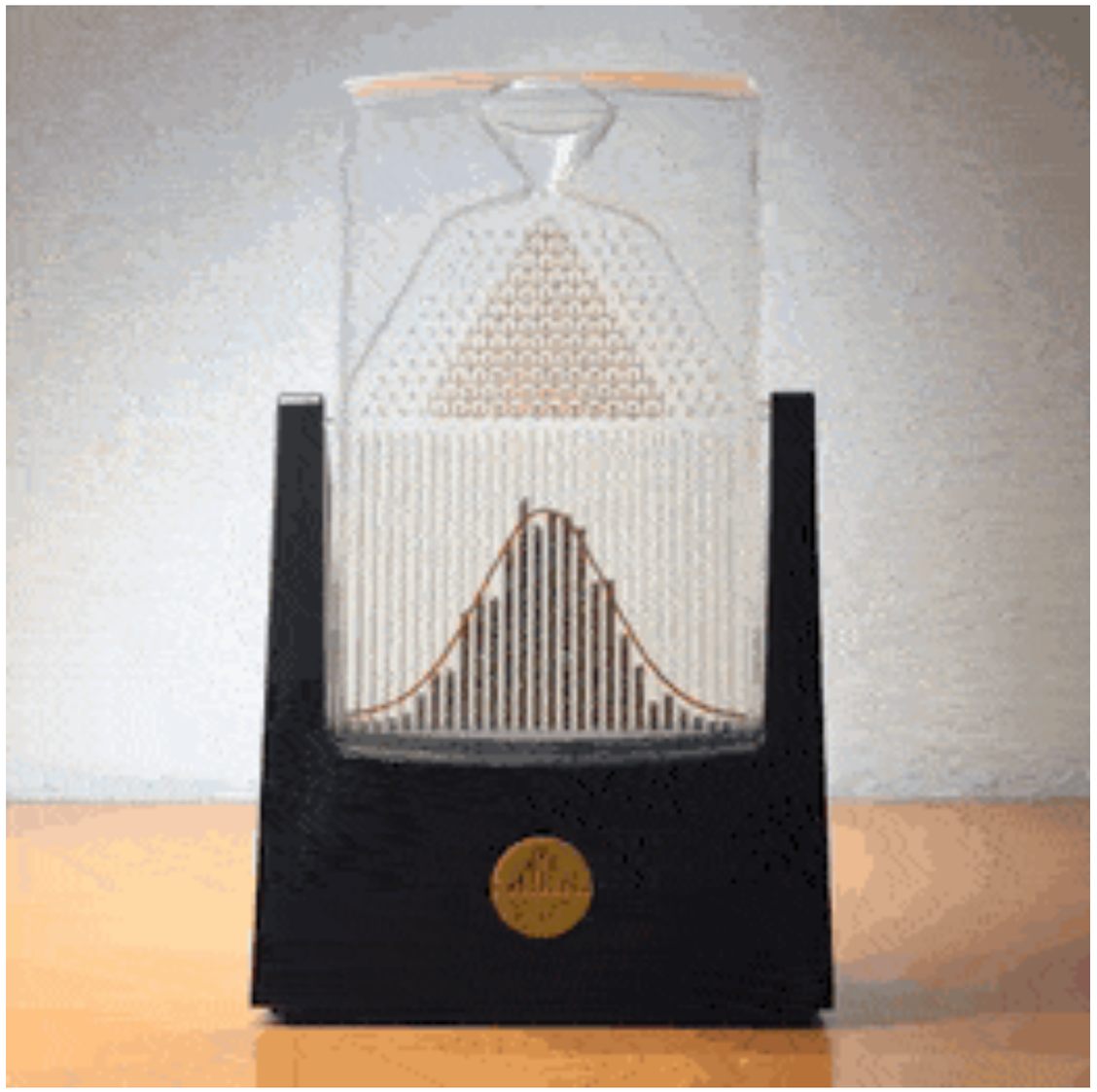
	0.5(c+1)	t
c = 0.9	0.95	1.645
c = 0.95	0.975	1.960
c = 0.99	0.995	2.576

WHAT MAKES THE NORMAL DISTRIBUTION SO SPECIAL?



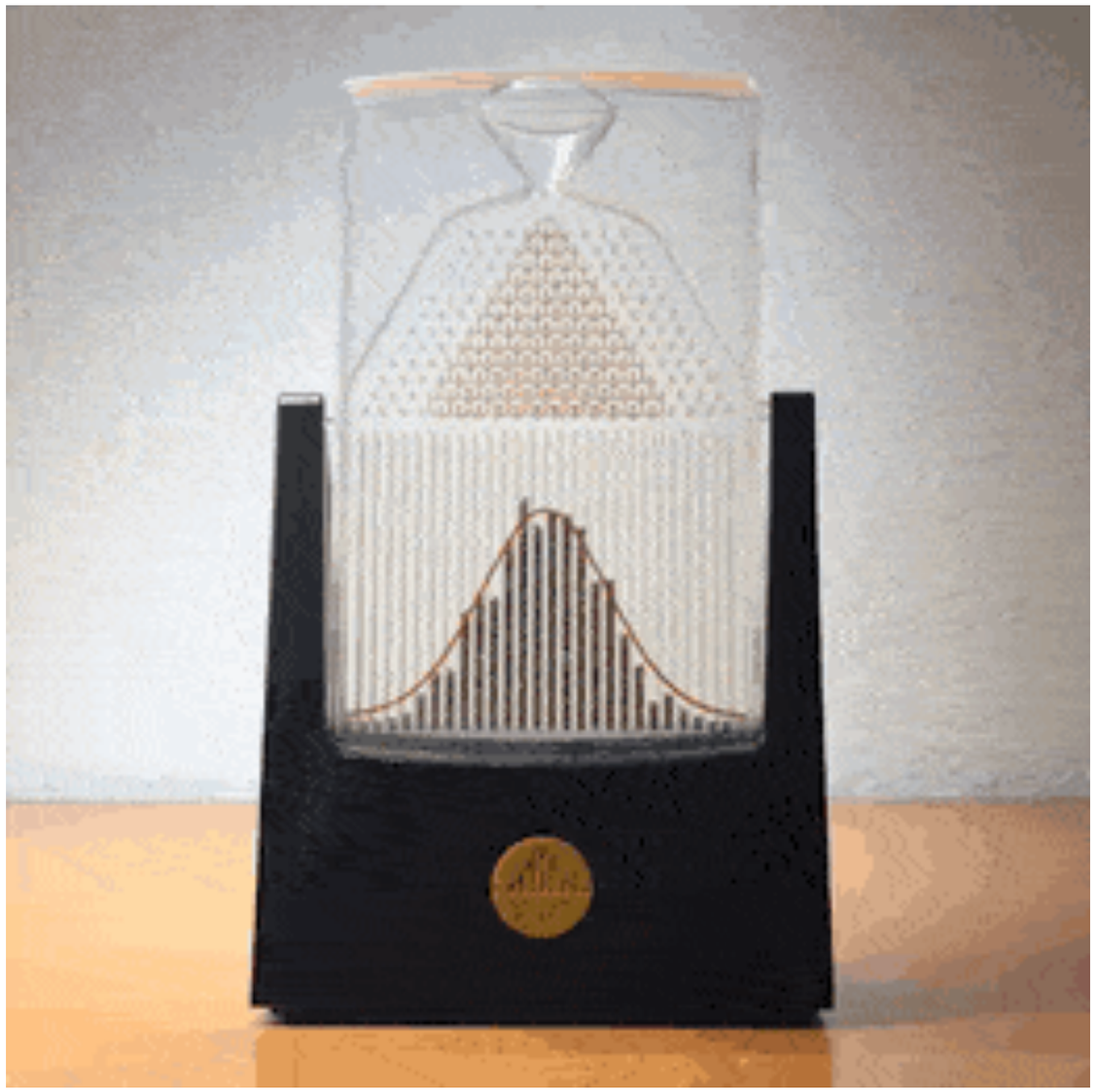
WHAT MAKES THE NORMAL DISTRIBUTION SO SPECIAL?

- **The Galton Board**
  Container of small balls is poured over triangular grid of pegs
  The balls form a bell curve shape



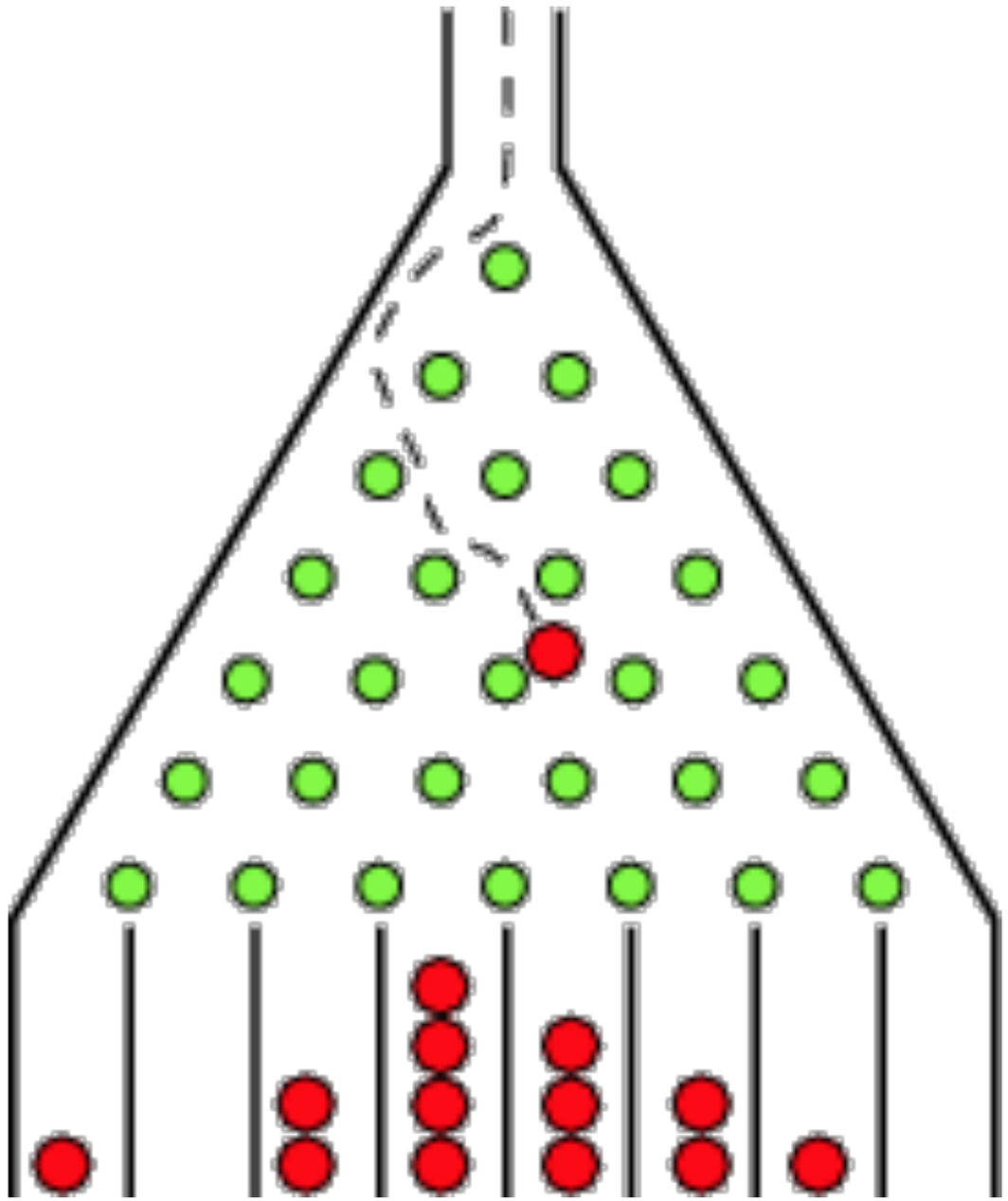
WHAT MAKES THE NORMAL DISTRIBUTION SO SPECIAL?

- **The Galton Board**
  Container of small balls is poured over triangular grid of pegs
  The balls form a bell curve shape

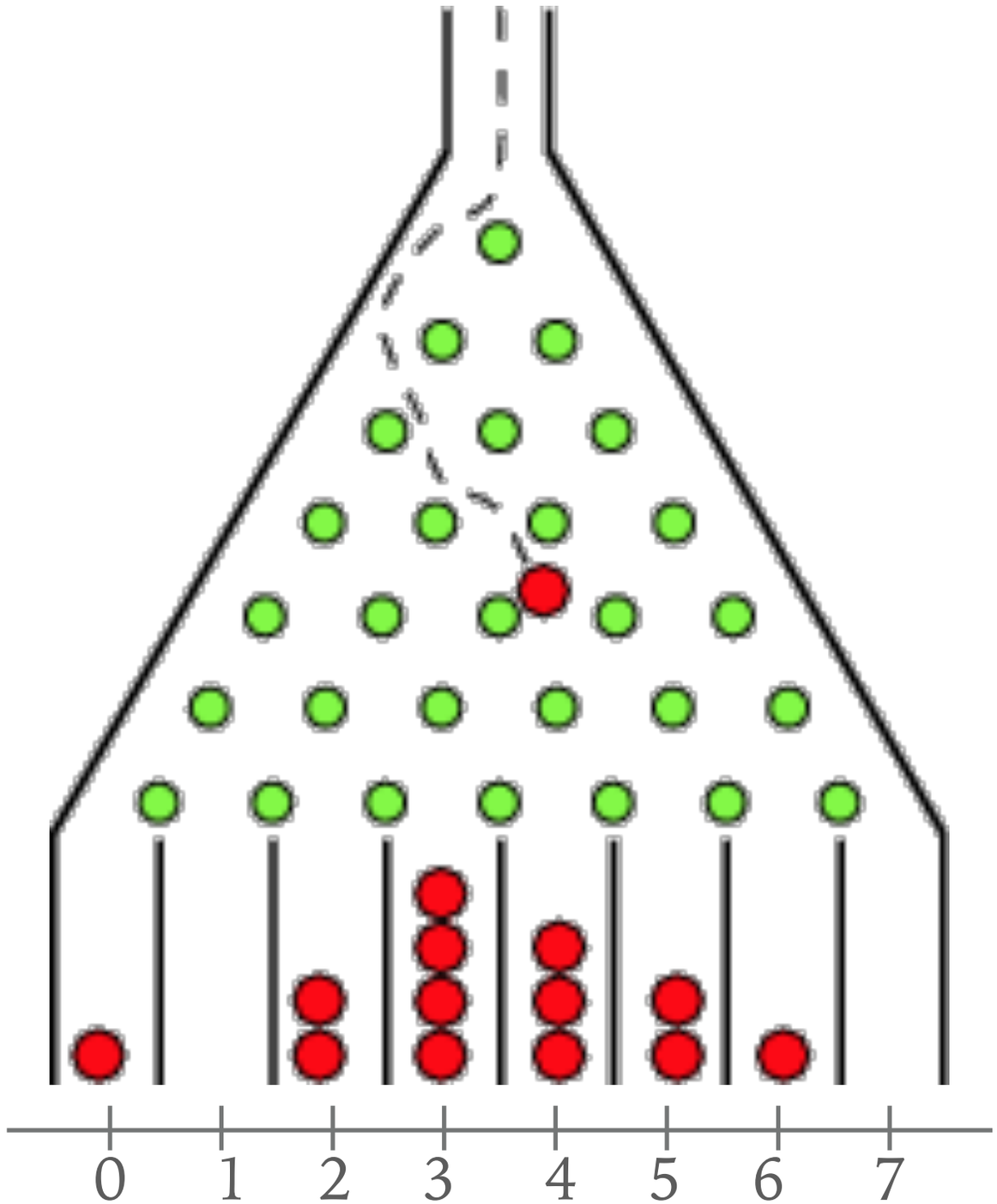


- … final location of each individual ball is random, but the overall shape of the result is same

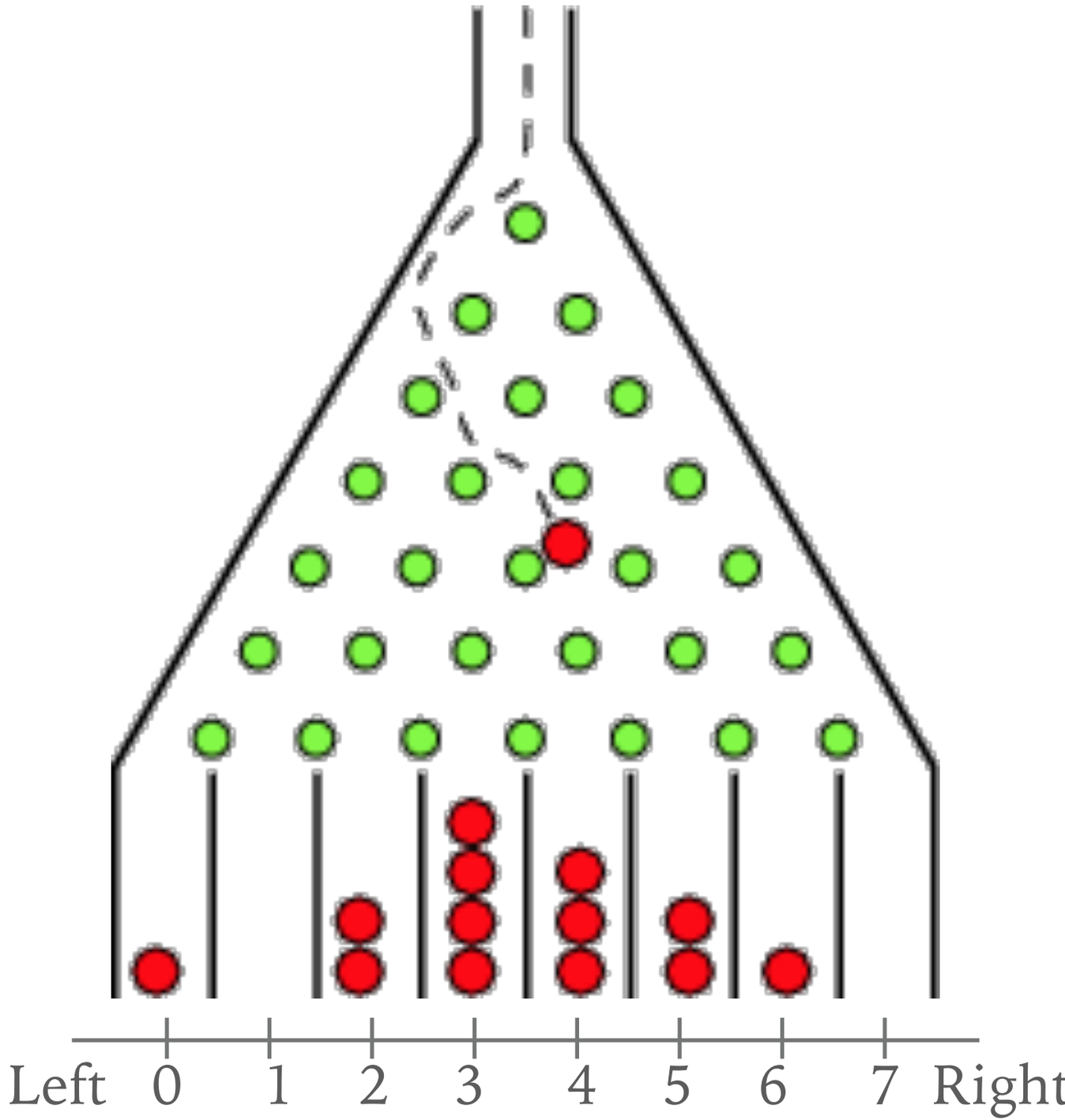
LEFT OR RIGHT?



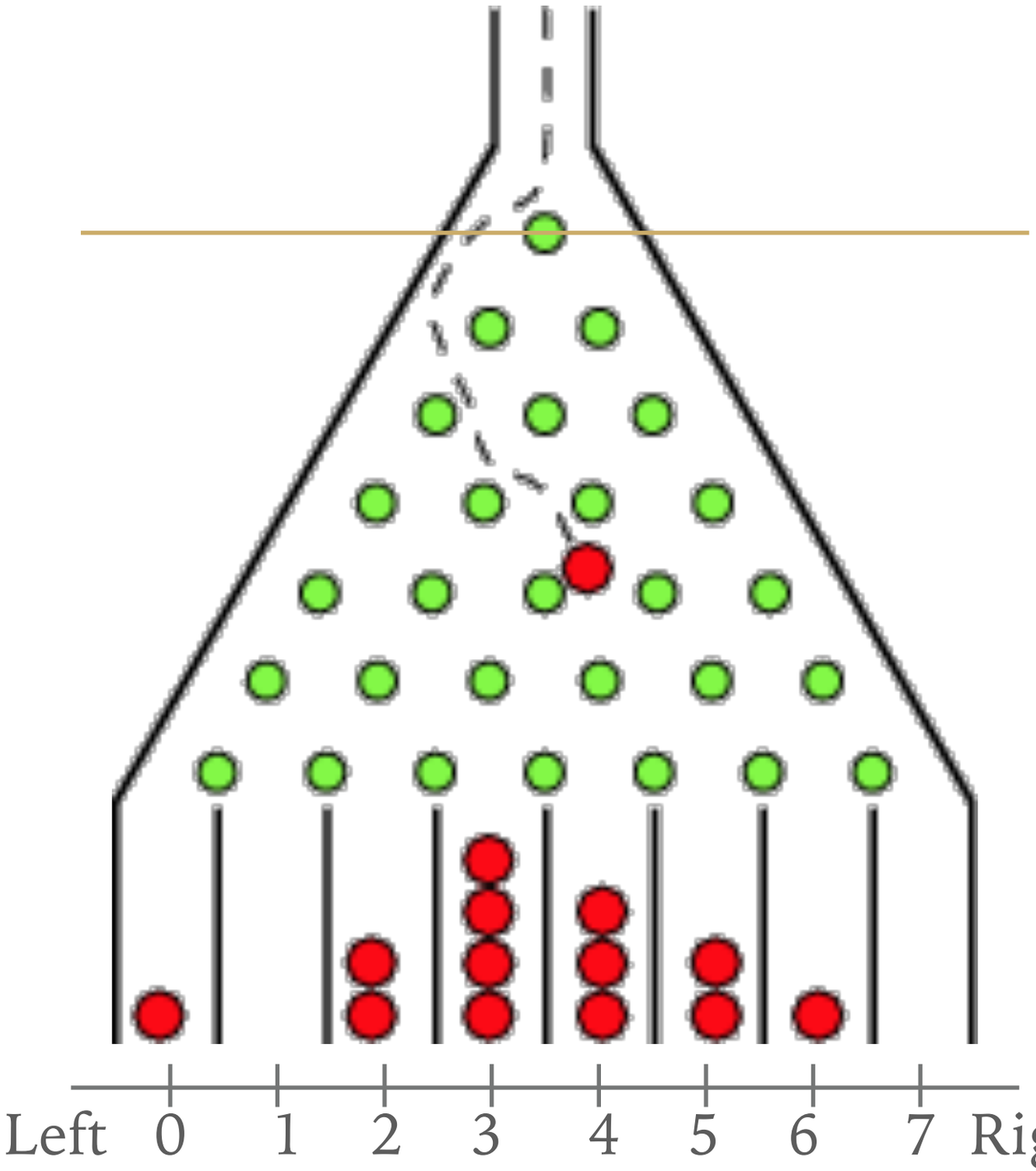
LEFT OR RIGHT?



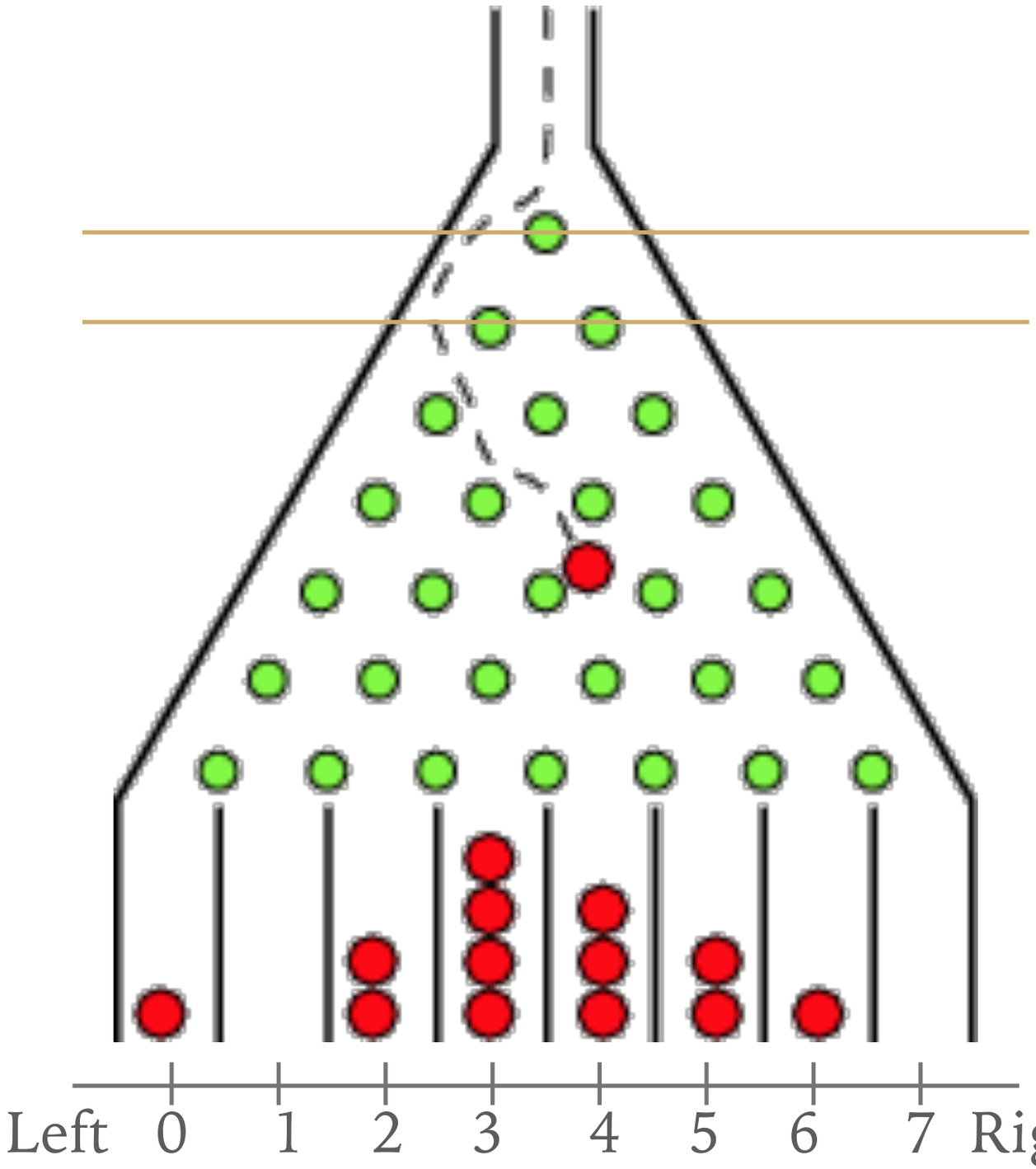
LEFT OR RIGHT?



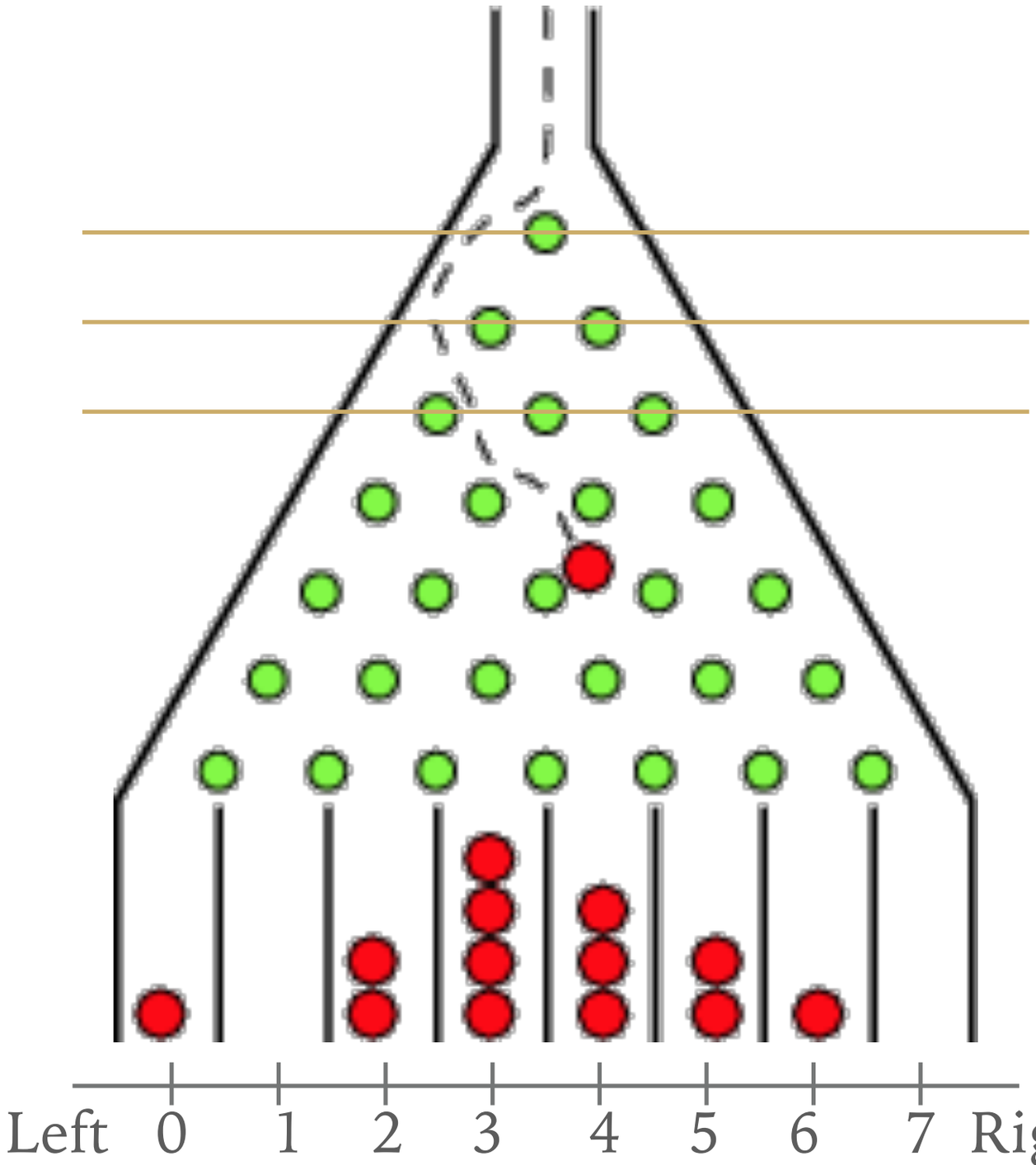
LEFT OR RIGHT?



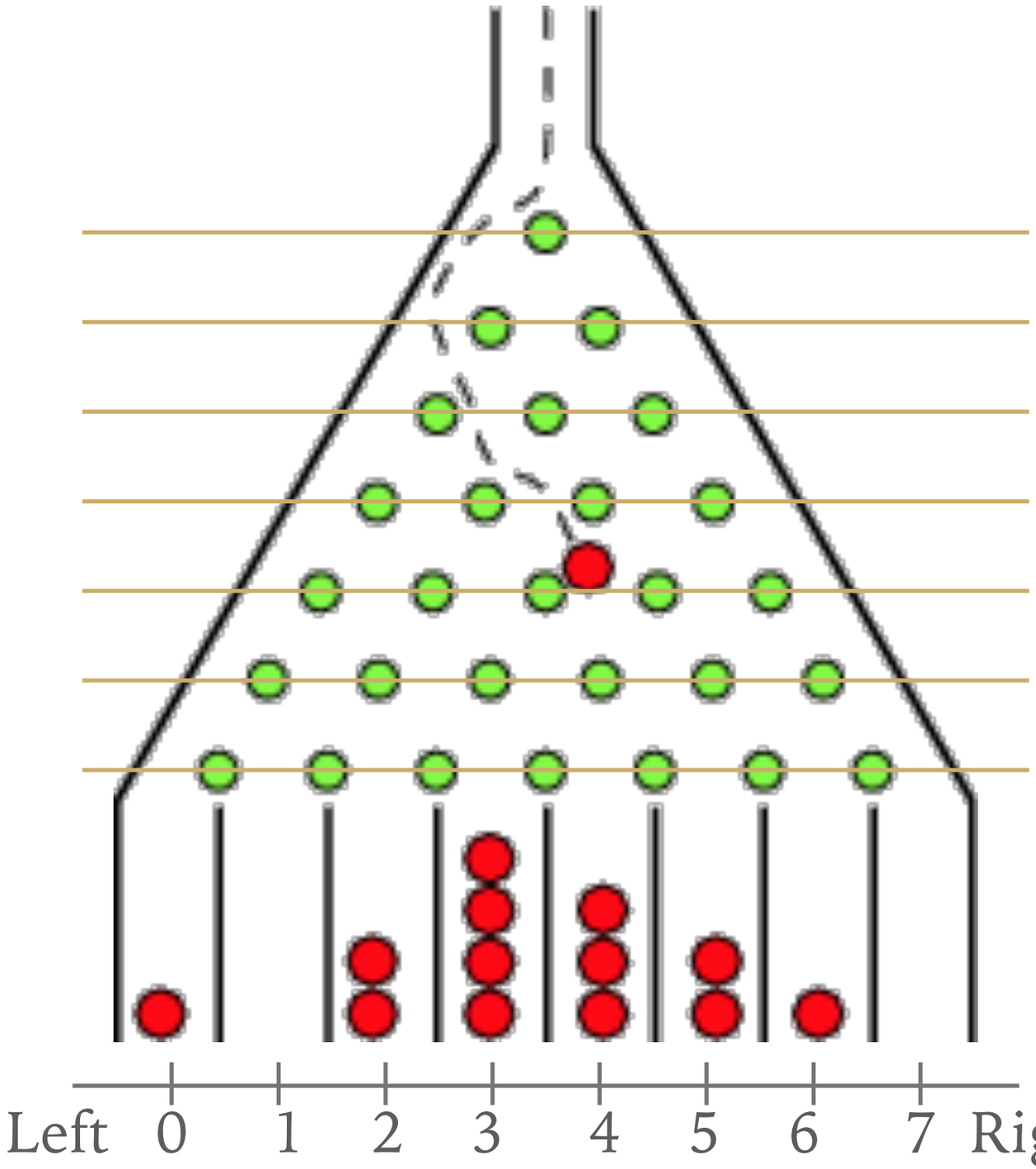
LEFT OR RIGHT?



LEFT OR RIGHT?

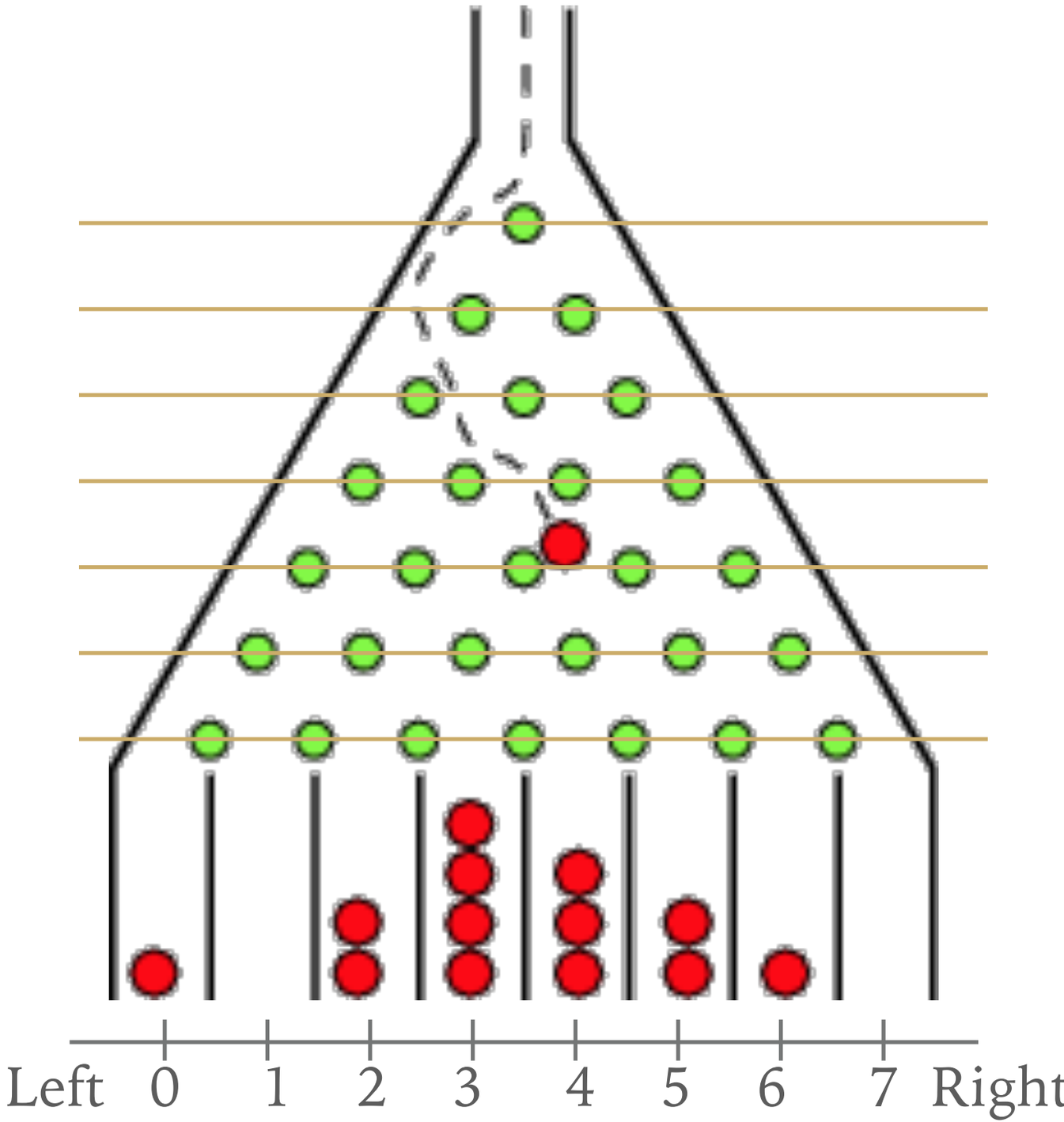


LEFT OR RIGHT?



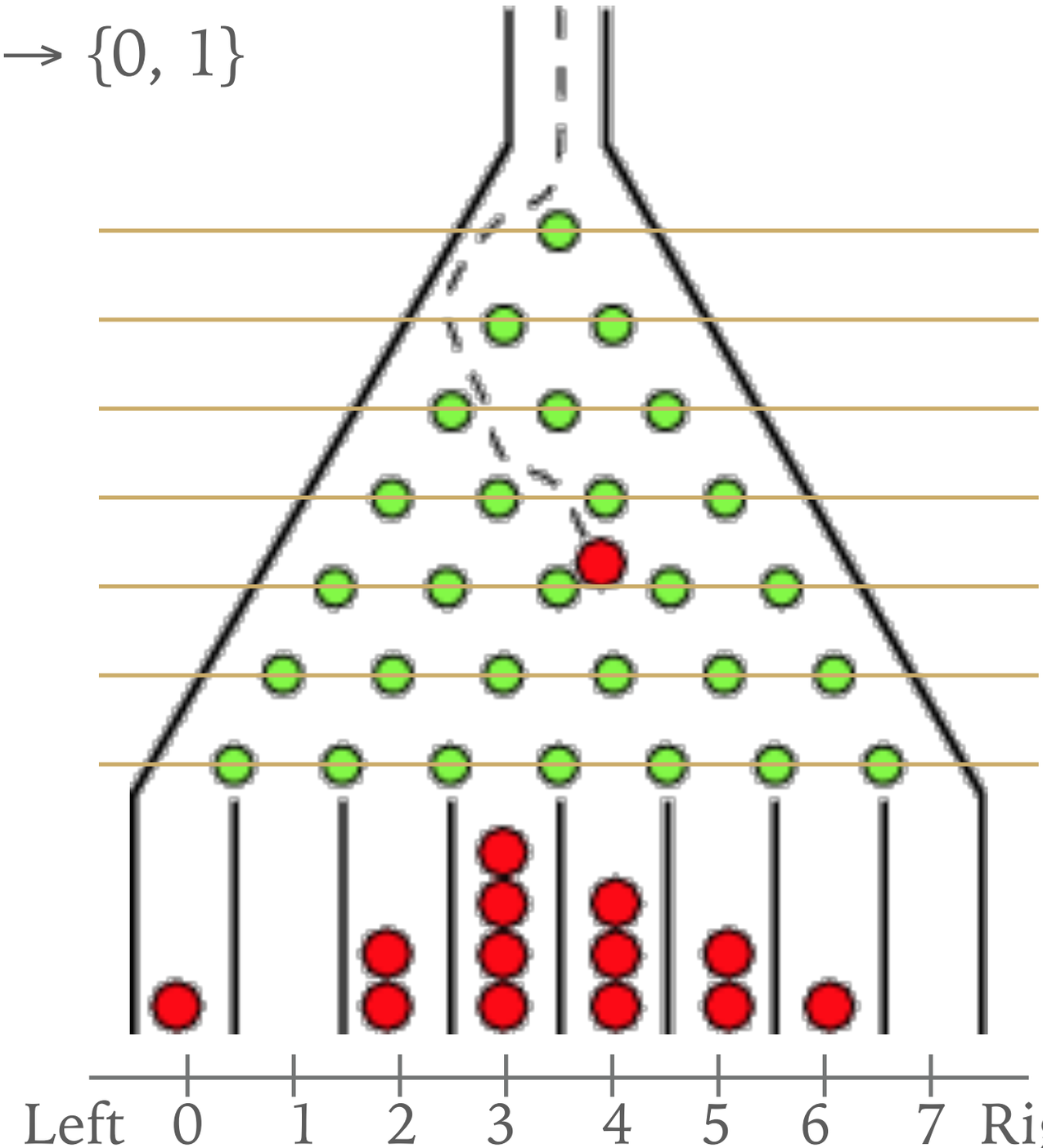
LEFT OR RIGHT?

➤ At each line, the red ball has to make a left/right decision



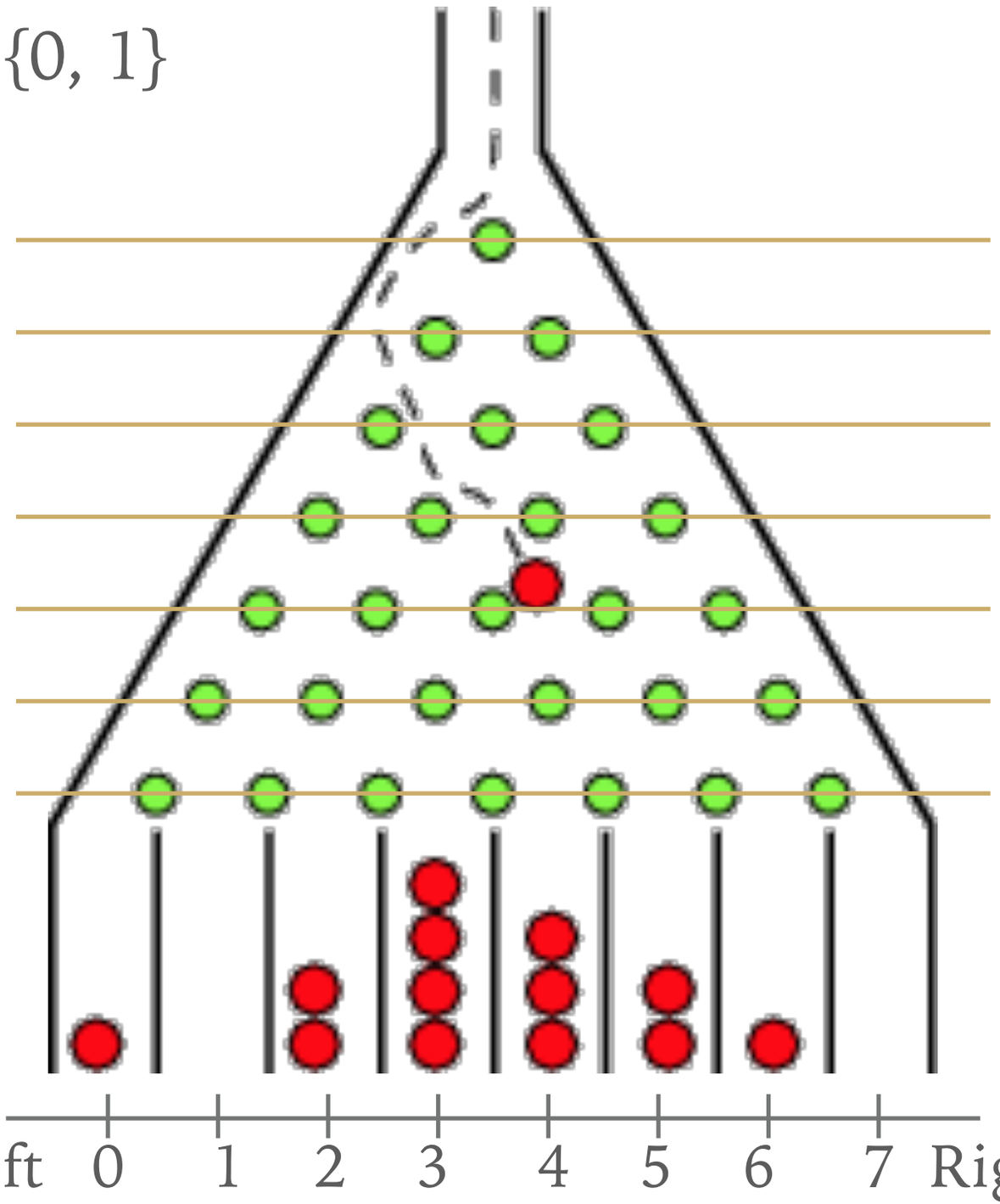
LEFT OR RIGHT?

- At each line, the red ball has to make a left/right decision
- Define RV X_i : {Left, Right} → {0, 1}



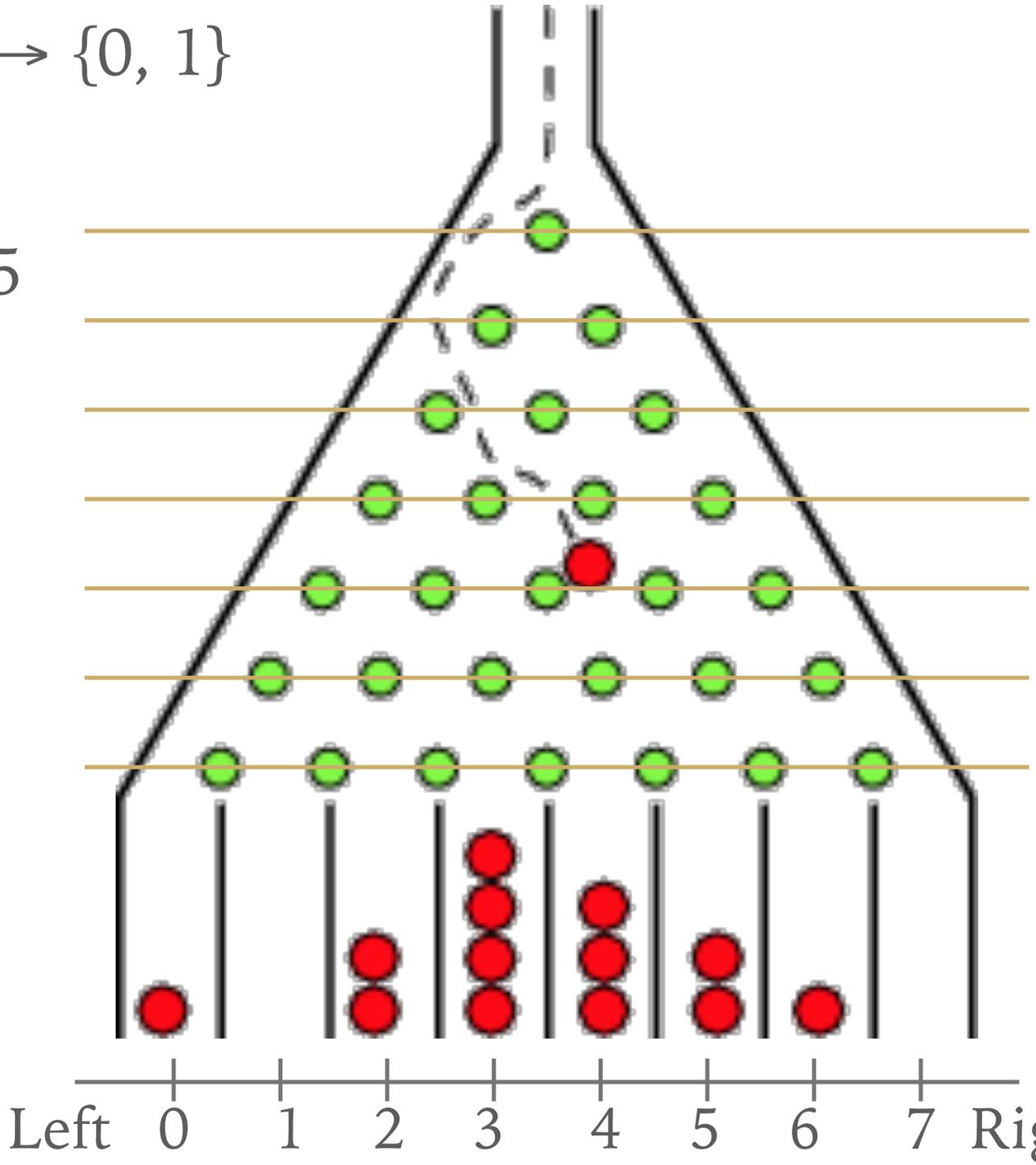
LEFT OR RIGHT?

- At each line, the red ball has to make a left/right decision
- Define RV X_i : {Left, Right} → {0, 1}
- X_i is Bernoulli RV, with
  $P(\{Right\}) = P(\{Left\}) = 0.5$
  $E[X_i] = 0.5$, $Var[X_i] = 0.25$



LEFT OR RIGHT?

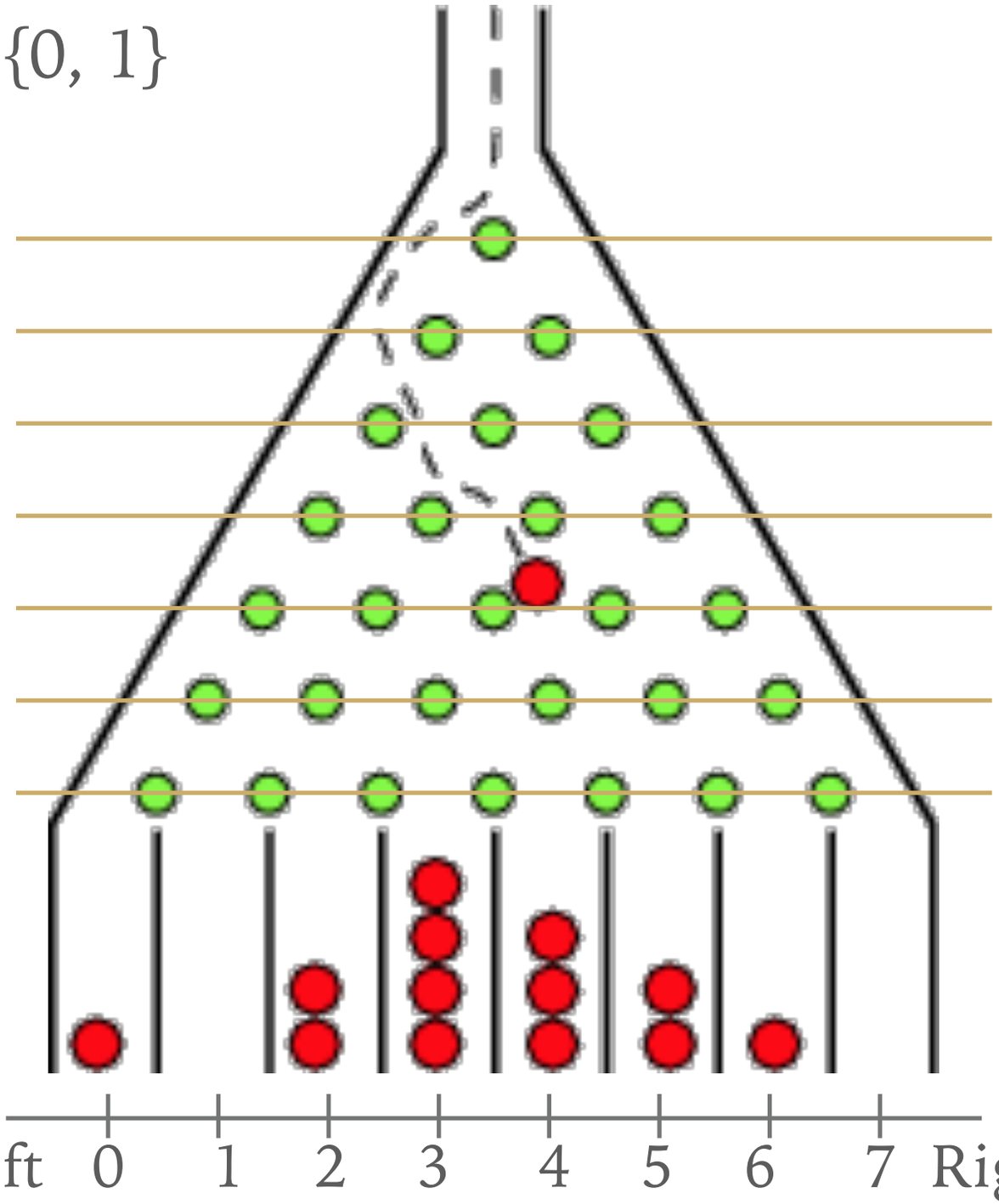
- At each line, the red ball has to make a left/right decision
- Define RV X_i : {Left, Right} → {0, 1}
- X_i is Bernoulli RV, with
  $P(\{Right\}) = P(\{Left\}) = 0.5$
  $E[X_i] = 0.5$, $Var[X_i] = 0.25$
- Final spot is determined as
  $X = X_1 + \ldots + X_7 \sim Bin_{7,\ 0.5}$



LEFT OR RIGHT?

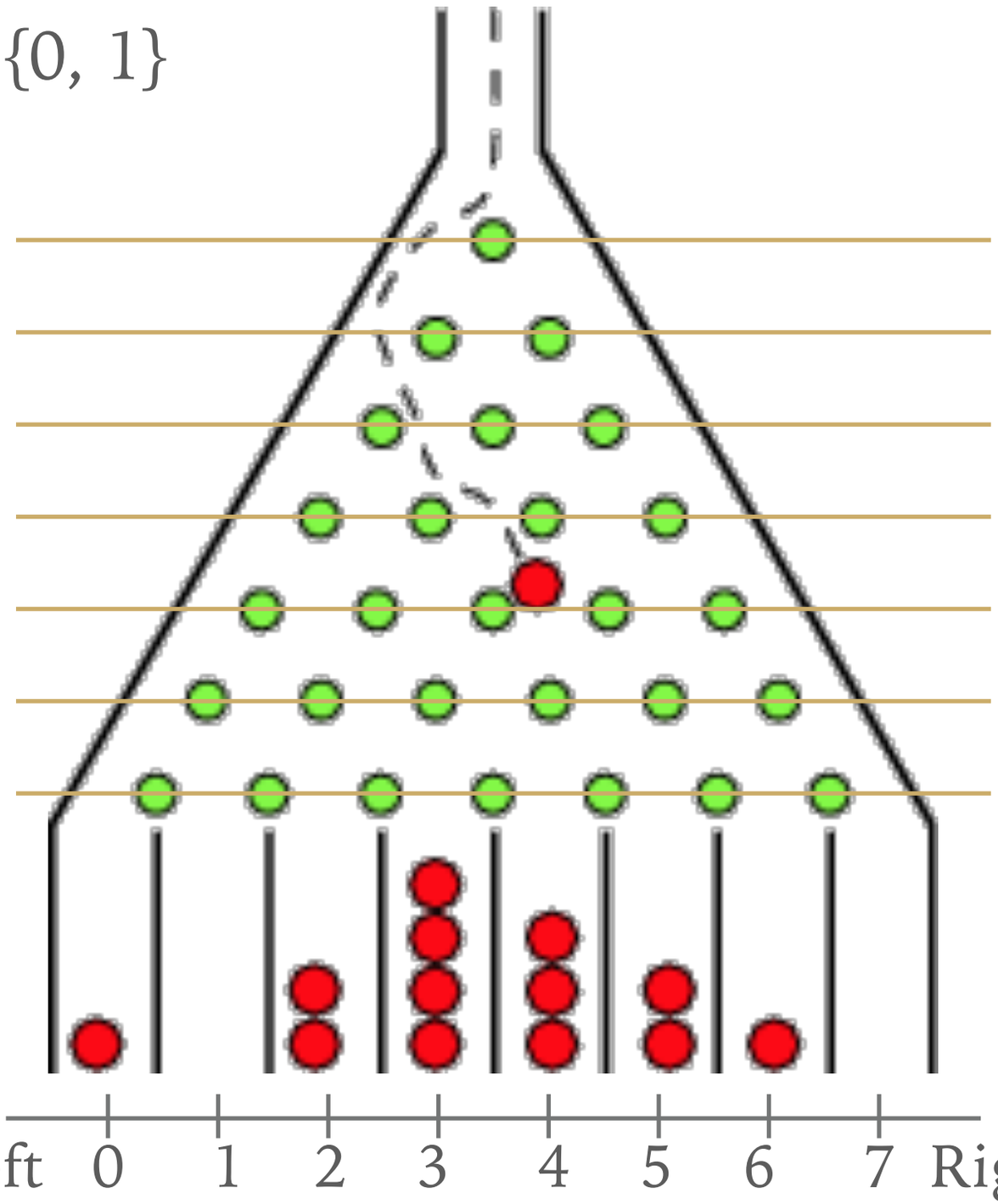
- At each line, the red ball has to make a left/right decision
- Define RV X_i : {Left, Right} → {0, 1}
- X_i is Bernoulli RV, with
  $P(\{Right\}) = P(\{Left\}) = 0.5$
  $E[X_i] = 0.5$, $Var[X_i] = 0.25$
- Final spot is determined as
  $X = X_1 + \ldots + X_7 \sim Bin_{7,\ 0.5}$

Binomial looks like a normal bell curve!



LEFT OR RIGHT?

- At each line, the red ball has to make a left/right decision
- Define RV X_i : {Left, Right} → {0, 1}
- X_i is Bernoulli RV, with
  $P(\{Right\}) = P(\{Left\}) = 0.5$
  $E[X_i] = 0.5$, $Var[X_i] = 0.25$
- Final spot is determined as
  $X = X_1 + \ldots + X_7 \sim Bin_{7,\ 0.5}$



Binomial looks like a normal bell curve!

… this is not a coincidence

CENTRAL LIMIT THEOREM

CENTRAL LIMIT THEOREM

CENTRAL LIMIT THEOREM

- Let X be the sum of n independent, identically distributed RVs $X_1, \ldots, X_n$ with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2$

$$X = X_1 + X_2 \ldots + X_n$$

then X is approximately normally distributed $N(n\mu, n\,\sigma^2)$

CENTRAL LIMIT THEOREM

- Let X be the sum of n independent, identically distributed RVs $X_1, \ldots, X_n$ with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2$

$$X = X_1 + X_2 \ldots + X_n$$

then X is approximately normally distributed $N(n\mu, n\,\sigma^2)$

- CLT sometimes stated as theorem of the average:
Sample mean $\overline{X} = (X_1 + X_2 \ldots + X_n)/n$
has a limiting distribution (i.e. $n \to \infty$) of $N(\mu, \sigma^2/n)$
for n i.i.d. RVs $X_1, \ldots, X_n$ with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2$

NORMAL APPROXIMATION OF BINOMIAL

- The Binomial distribution is defined as a sum of i.i.d Bernoulli variables
- The Binomial with parameters n and p can therefore be approximated by a normal distribution N(np, np(1-p))

- Generally, we assume that n is large enough, if $np > 5$ and $n(1-p) > 5$

NORMAL APPROXIMATION OF BINOMIAL

- A new drug to treat high blood pressure has a side effect of indigestion in 3% of patients. In a study with 120 patients, what is …
- … the probability that 5 or fewer patients experience this side effect?
- … the probability that 10 or more patients experience this side effect?
- Use the normal approximation to the binomial distribution with $n=120$, $p = 0.03$.
- Compare to the exact values from the binomial distribution.

QUESTIONS?