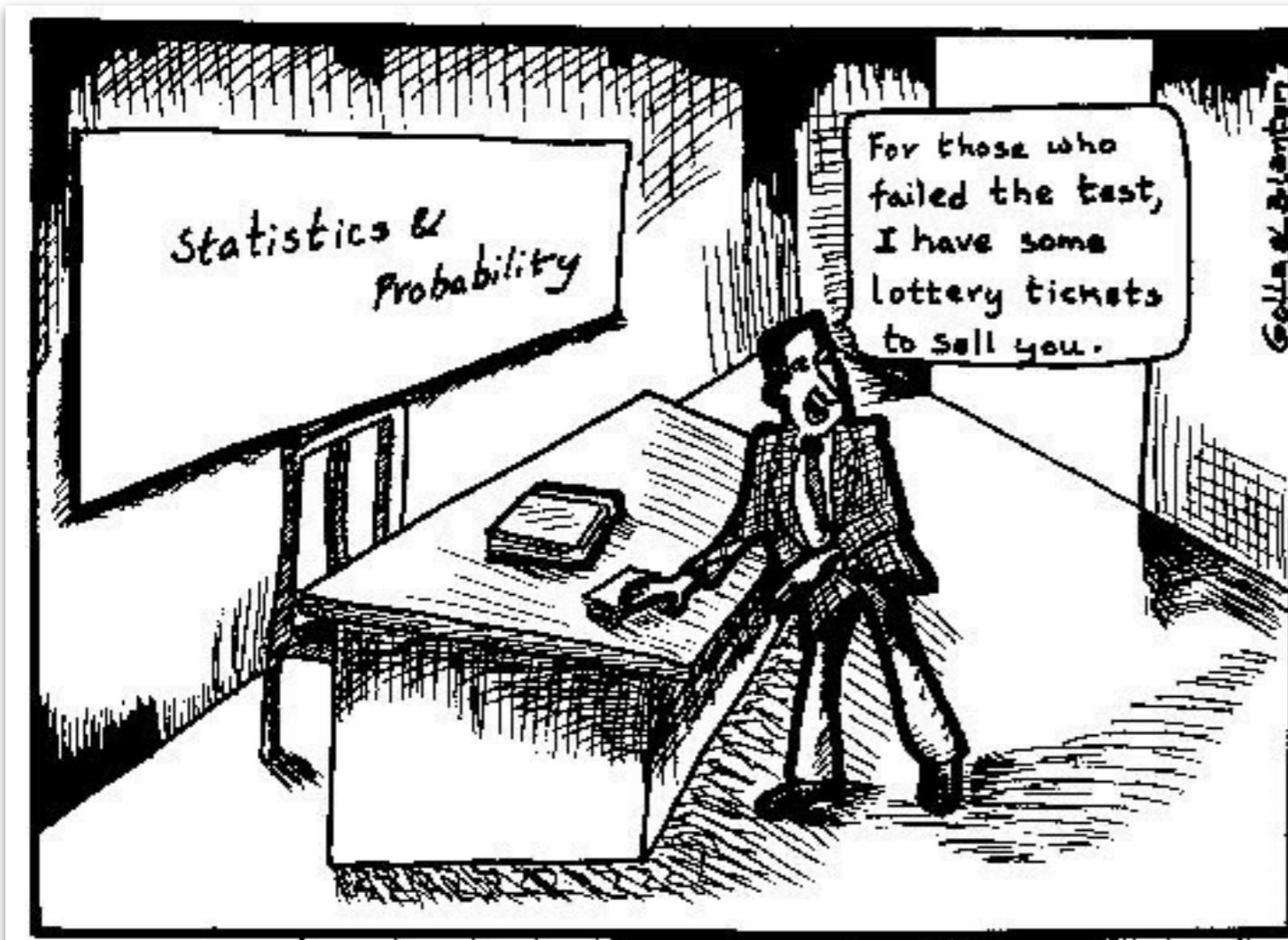


Probability



Chat

What's the present you most wanted as a kid?

To: Everyone More ▾

Type message here...

O COLLABORATIVE PLAYLIST
psych252

<https://tinyurl.com/psych252spotify22>

PLAY ...

We're listening to
"Video" by "India.Arie"
submitted by j.av1997

01/14/2022

Logistics

Solutions to practice problems uploaded

W22-PSYCH-252-01 > Files > slides > 05_data_wrangling2 > R

Search for files  0 items selected

Name ▲ Date Created Date Modified Modified By Size

- Statistical Methods for Behavioral Sciences
- final_project
- homework
- sections
- slides
 - 01_introduction
 - 02_visualization1
 - 03_visualization2
 - 04_data_wrangling1
 - 05_data_wrangling2
 - R
 - data
 - figures

Name	Date Created	Date Modified	Modified By	Size
05_data_wrangling2.Rproj	Tuesday	Tuesday		205 bytes
data	Tuesday			--
data_wrangling2_solutions.html	3:49pm	3:49pm	Tobias Gerstenberg	837 KB
data_wrangling2_solutions.Rmd	3:49pm	3:49pm	Tobias Gerstenberg	4 KB
data_wrangling2.html	Tuesday	Tuesday		909 KB
data_wrangling2.Rmd	Tuesday	Tuesday		28 KB
figures	Tuesday			--

Class 5

Tobias Gerstenberg

- 5 Data wrangling 2: Exercise solutions
 - 5.1 Load packages and data set
 - 5.2 Settings
 - 5.3 Practice 1
 - 5.4 Practice 2
 - 5.5 Practice 3

5 Data wrangling 2: Exercise solutions

5.1 Load packages and data set

Let's first load the packages that we need for this chapter.

```
library("knitr") # for rendering the RMarkdown file
library("tidyverse") # for data wrangling
```

5.2 Settings

```
opts_chunk$set(comment = "")
options(dplyr.summarise.inform = F)
```

And let's load the data set into the environment

```
df.starwars = starwars
```

5.3 Practice 1

Find out what the average height and mass (as well as the standard deviation) is from different species in different homeworlds. Why is the standard deviation NA for many groups?

```
df.starwars %>%
  group_by(species, homeworld) %>%
  summarise(mean_height = mean(height, na.rm = T),
            mean_mass = mean(mass, na.rm = T),
            sd_height = sd(height, na.rm = T),
            sd_mass = sd(mass, na.rm = T),
            n = n()) %>%
  ungroup()
```

```
# A tibble: 58 x 7
  species homeworld  mean_height mean_mass sd_height sd_mass     n
  <chr>   <chr>        <dbl>      <dbl>    <dbl>      <dbl> <int>
1 Aleena  Aleen Minor      79       15       NA       NA     1
2 Besalisk Ojom          198      102       NA       NA     1
3 Cerean   Cerea          198       82       NA       NA     1
4 Chagrian Champala      196      NaN       NA       NA     1
5 Clawdite Zolan          168       55       NA       NA     1
6 Droid    Naboo           96        32       NA       NA     1
7 Droid    Tatooine        132      53.5     49.5    30.4    2
8 Droid    <NA>            148      140      73.5    NA      3
9 Dug     Malastare         112       40       NA       NA     1
10 Ewok    Endor            88        20       NA       NA     1
# ... with 48 more rows
```

Final project

W22-PSYCH-252-01 > Files > final_project > final_report > 2021

Search for files Q 0 items selected

▼ Statistical Methods for Behavioral ...

- ▼ final_project
 - ▼ final_report
 - 2019
 - ▼ 2021
 - homework
 - sections
 - slides

Name ▲	Date Created	Date Modified	Modified By	Size	⋮
 anjie_cao.pdf	Tuesday	Tuesday	Tobias Gerste...	857 KB	⋮
 danyang_fan.pdf	Tuesday	Tuesday	Tobias Gerste...	15.6 MB	⋮
 hannah_marshall.pdf	Tuesday	Tuesday	Tobias Gerste...	2.6 MB	⋮
 madi_jamie_catie.pdf	Tuesday	Tuesday	Tobias Gerste...	4.3 MB	⋮

Effects of similarity and complexity on adult looking time

Anjie Cao
2021-03-18 12:47:12

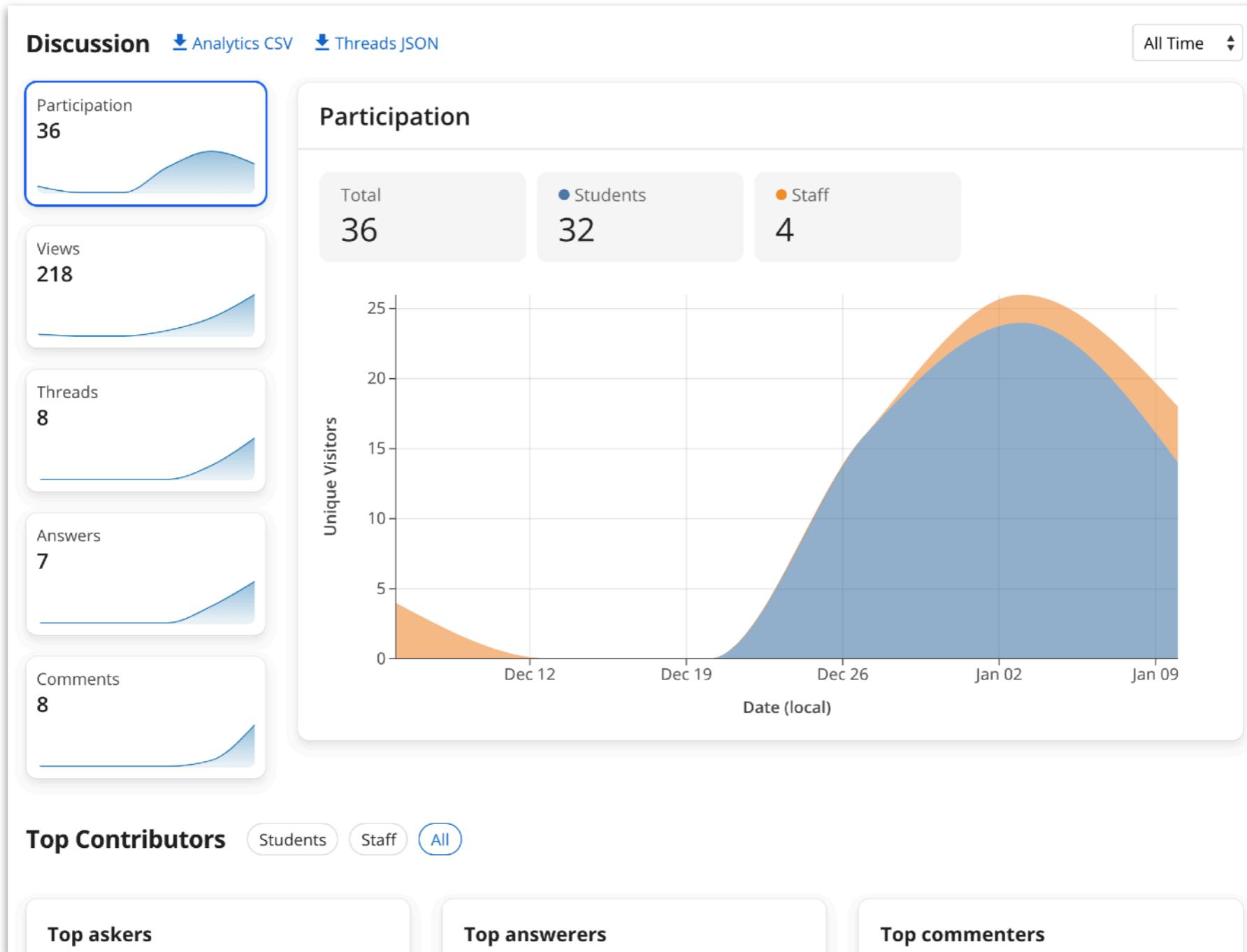
Contents

1	Introduction	1
1.1	Research questions	2
1.2	Hypotheses	2
2	Methods	3
2.1	Participants	3
2.2	Stimuli	3
2.3	Procedure	3
3	Results	3
3.1	Confirmatory analysis	4
3.1.1	Manipulation check	4
3.1.2	Effect of trial number, complexity and familiarity	4
3.2	Exploratory analysis	5
3.2.1	Exponential fit	5
3.2.2	Block-level complexity effect	6
3.2.3	Relationships between complexity ratings and dishabituation effect	7
4	Discussion	7
	References	8

1 Introduction

Looking time paradigms are widely used in infant research. In a typical looking time study, infants would be presented with some stimuli over and over again as their looking time decreases. Afterward, new stimuli would show up and evoke a longer looking time. As simple as this testing procedure is, it has been employed by researchers to answer questions across many aspects of early cognitive development, including categorization (Quinn, Eimas, and Rosenkrantz 1993), causal perception (Leslie and Keeble 1987), object concept

Ed discussion bonus points

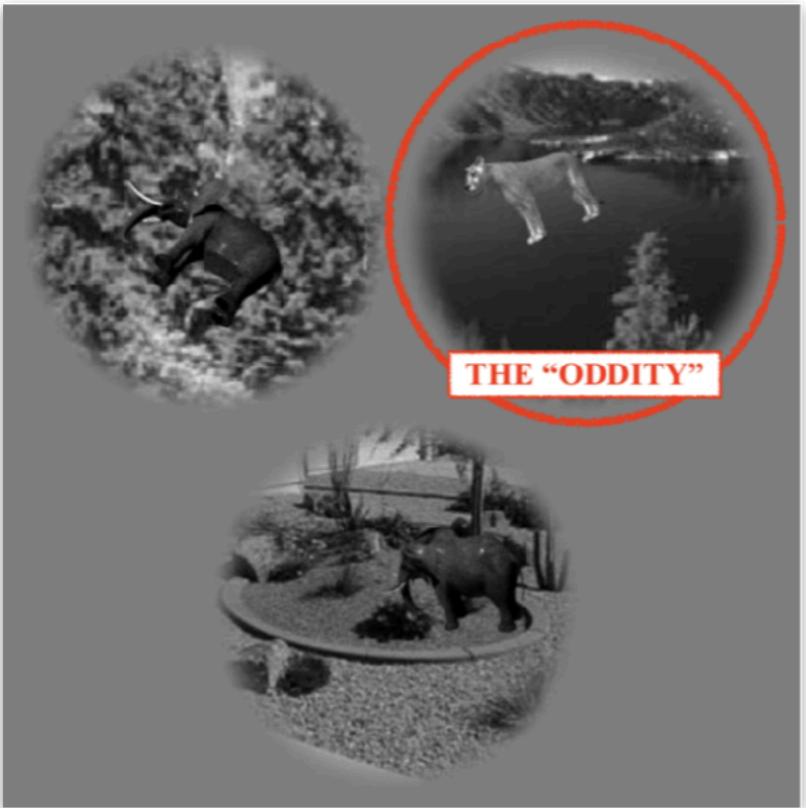


up to 2% bonus points to the final grade for active participation on Ed discussion

Plan for next week

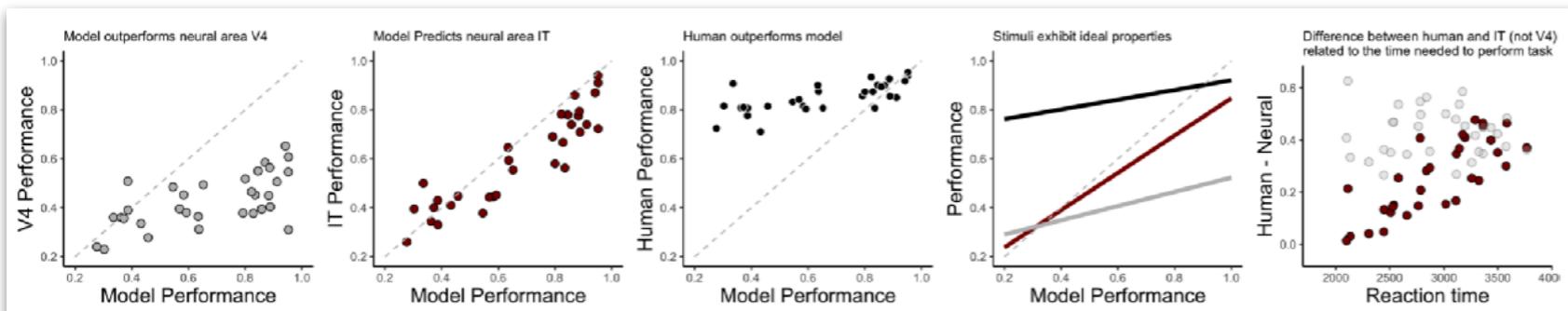
- no class on Monday (Martin Luther King, Jr. Day)
- class on Wednesday will be offered in Hybrid format
- sections next week will be offered both in person and via zoom (one TA each)

Homework 2



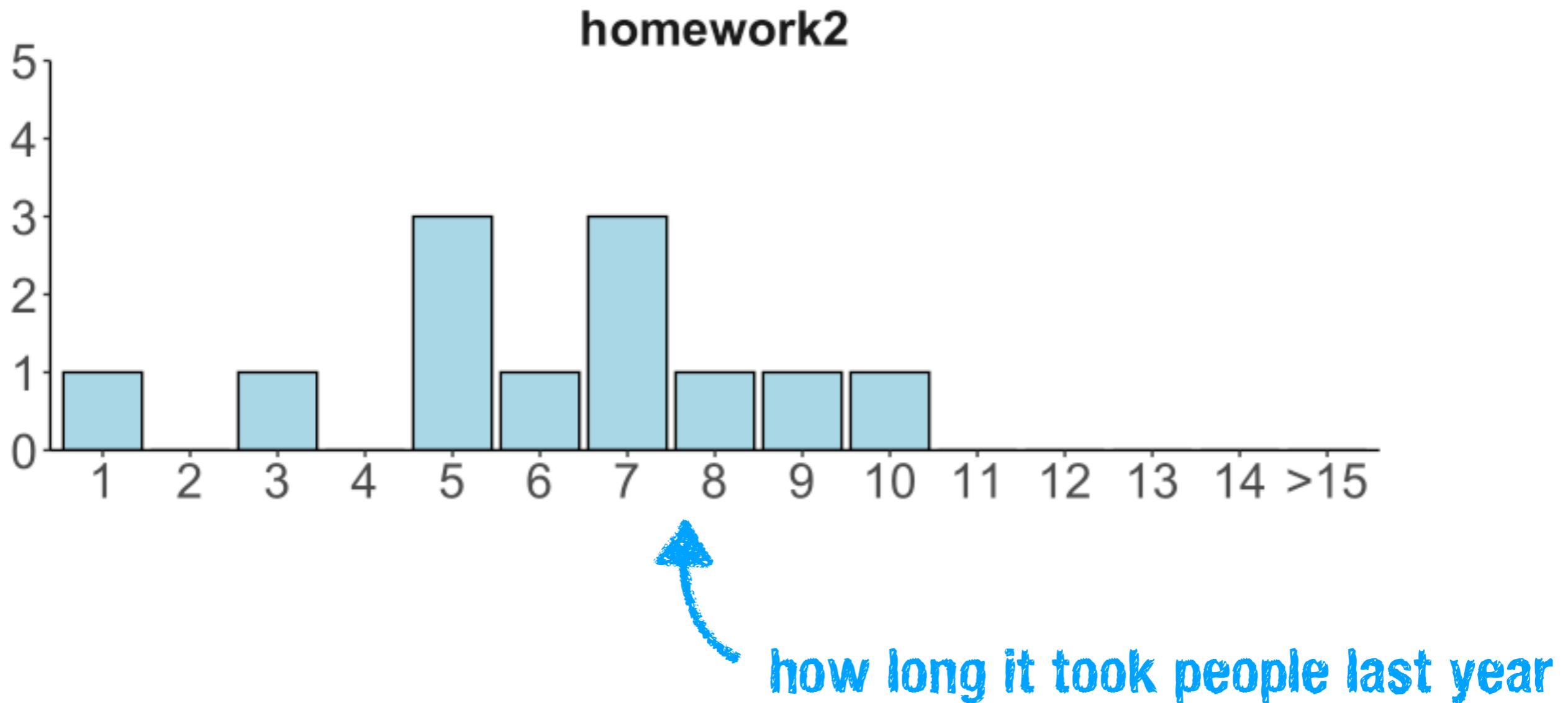
cool experiment

messy dataset



neat plot!

Homework 2



Homework 2

- Due **Thursday 20th, at 8pm**
- Don't wait until the very last moment to knit your RMarkdown file into a pdf. It may not compile and debugging takes time ...
- You can upload earlier versions of your homework on Canvas and still update until the deadline.
- Get and give help via Ed Discussion!

Homework 2

- **We encourage you to work in groups!**
- It's more fun
- You learn more (learning through explaining)

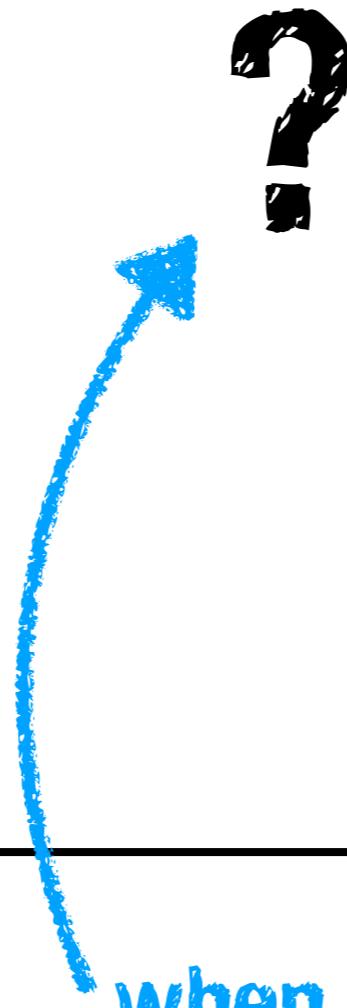
Homework 2

- You can adjust the figure size in your output by using the code chunk options



<https://yihui.org/knitr/options/>

It's a mystery



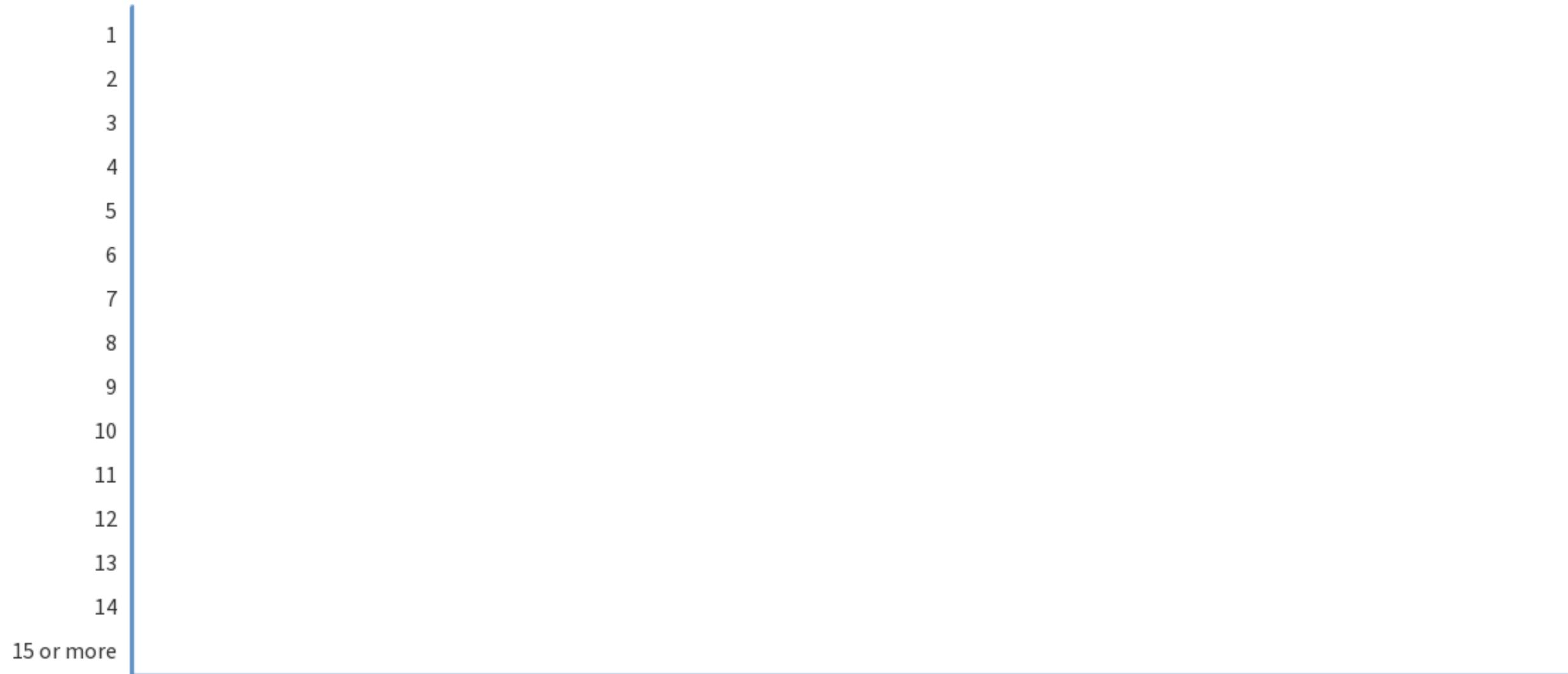
when I want to hide stuff from you

Your feedback

🌐 When poll is active, respond at **pollev.com/psych252**

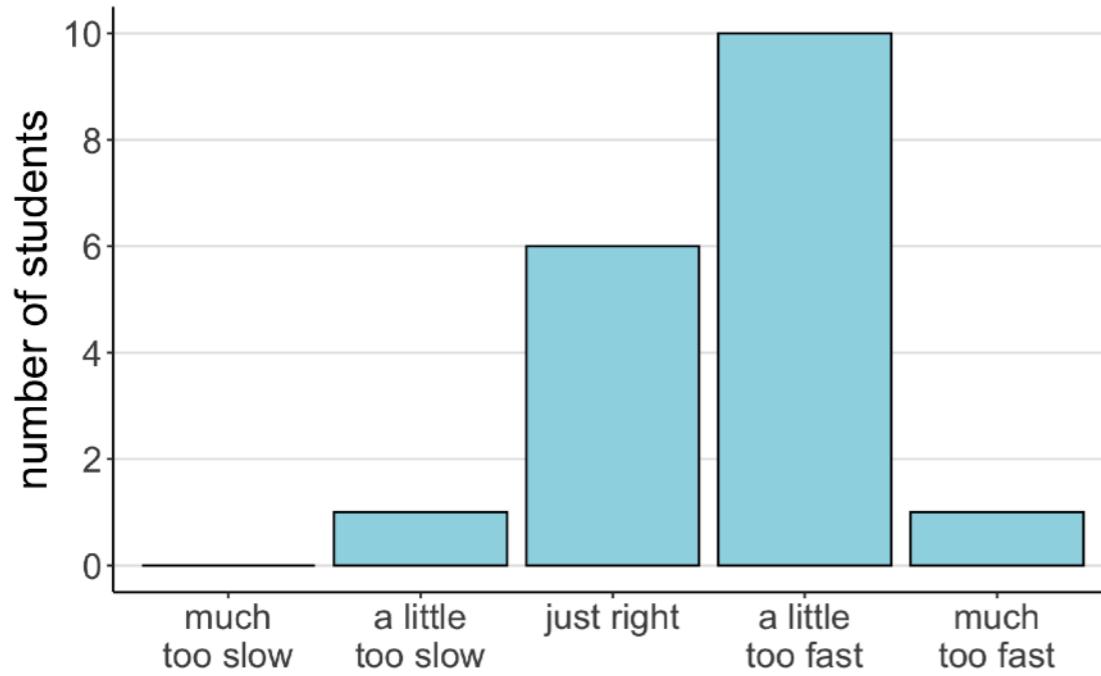


How many hours did it take you to complete Homework 1?

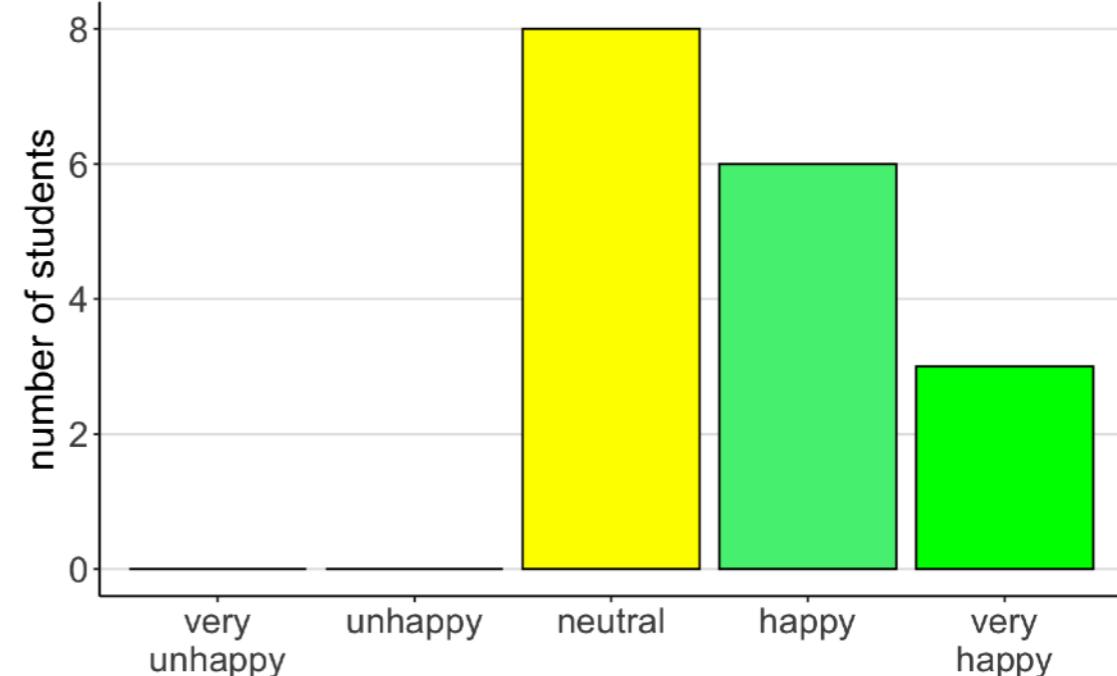


Your feedback

How was the pace of today's class?



How happy were you with today's class overall?



wish we had spent more time on the more complex stuff later in class like pivoting, etc. that stuff is generally a lot less intuitive than simple stuff like adding new columns. I mean, hopefully we won't have to waste our time tidying data and we can learn actually statistics in this class so it won't matter... but -_(ゞ)_/-

Really would like another wrangling class just focusing on making data tidy/combining dataframes.

breakout rooms where we work on code chunks are great!

I finally got the hang of a practice problem and was able to contribute and reproduce the code on my own!! Woohoo!!!!

Most things are improved by chinchillas. If an opportunity arises, they would be much appreciated.

I'm nervous about how much I actually have learned and worried that I'm in worse shape than I think I am. I understand the material in the class but not sure how well I can replicate the code from scratch, or come up with what I need to do myself.

Outline

- Introduction to probability / Recap
 - Motivation
 - Counting possibilities
 - **Clue** guide to probability
 - Understanding Bayes' Rule
 - Getting Bayes' right matters!
 - Building a Bayesis

Motivation

What does statistics have to do with probability?



Theory

Our goal is to develop theories. In psychology, theories of how the mind works.



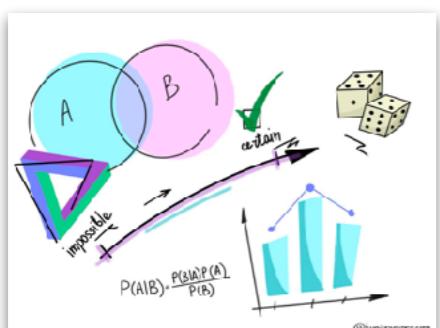
Prediction

Our theories need to make testable/falsifiable predictions.



Uncertainty

Because the domains that we are interested in are fundamentally uncertain (e.g. we want to say something about people generally but can only test a sample), we formulate and test these predictions using statistical models.

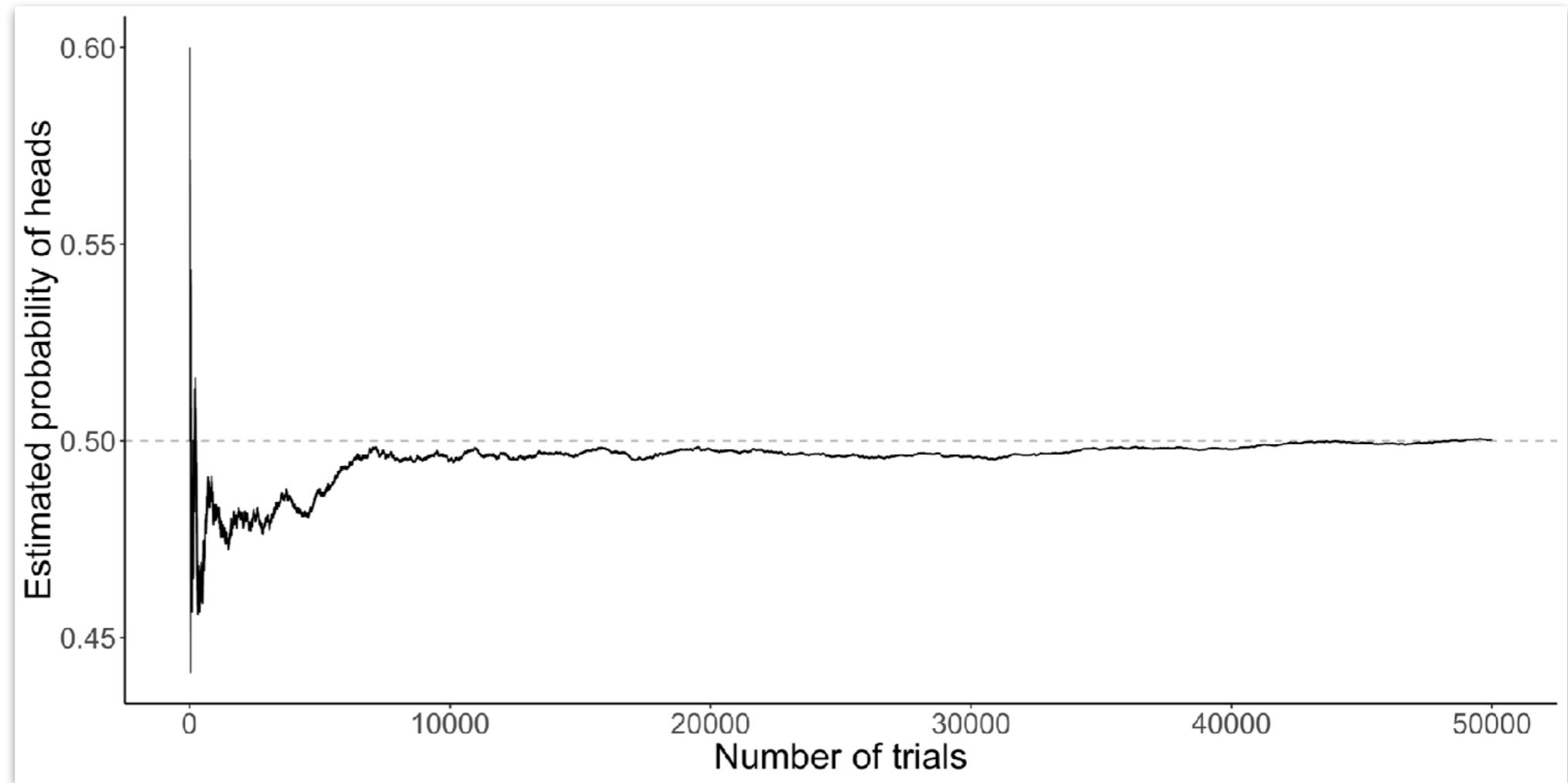


Probability

Probability theory is the formal language for dealing with uncertainty.

Frequentist interpretation

Probabilities = **long-range frequencies**



law of large numbers = empirical probability will approximate the true probability as the sample size increases

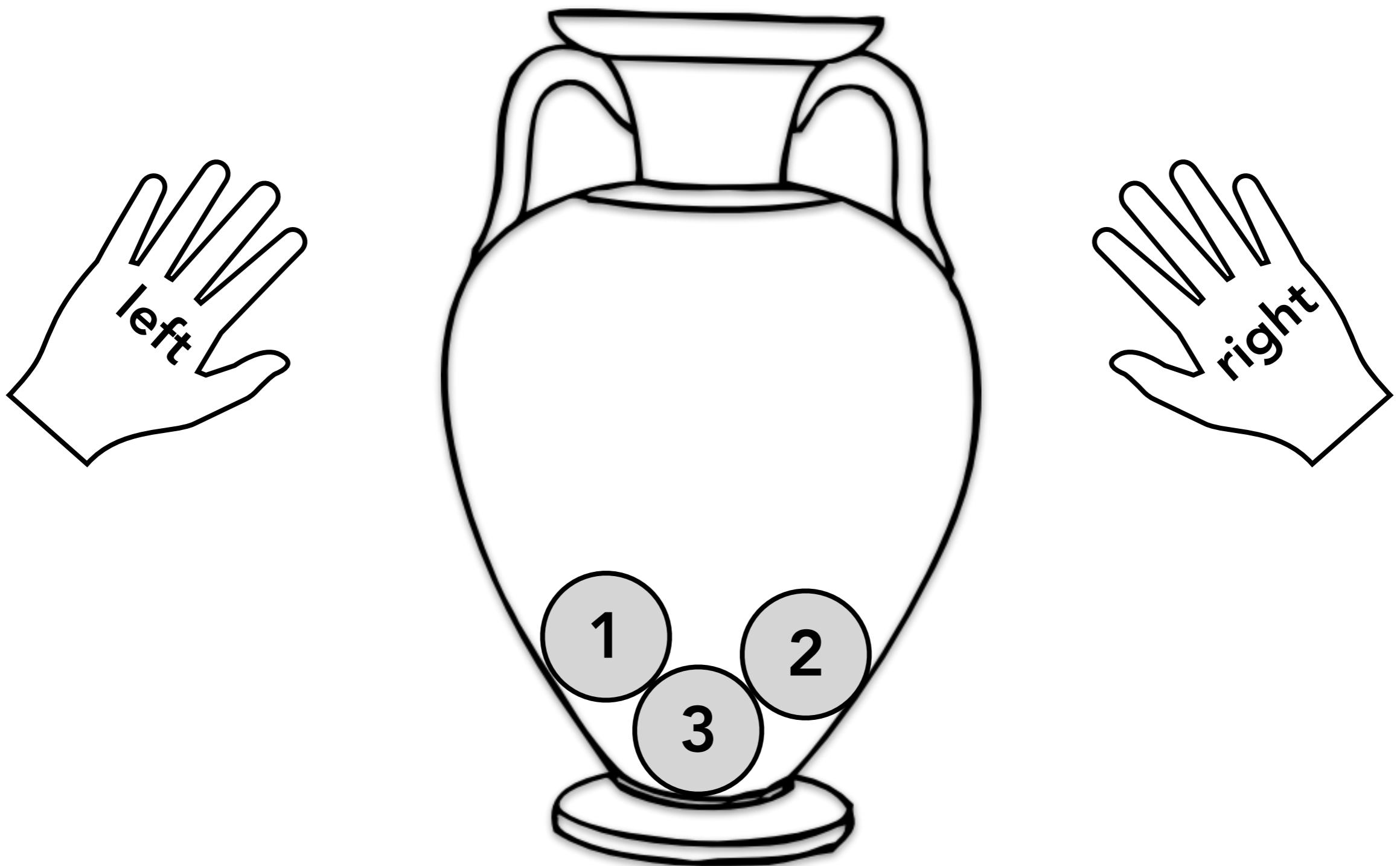
Subjective/Bayesian interpretation

Probabilities = **subjective degrees of belief**

- applies to events which may only happen once
- "**What's the probability that humans will land on Mars someday?**"
- probabilities are not a property of the world, but of a person's beliefs about the world
- at the heart of Bayesian data analysis

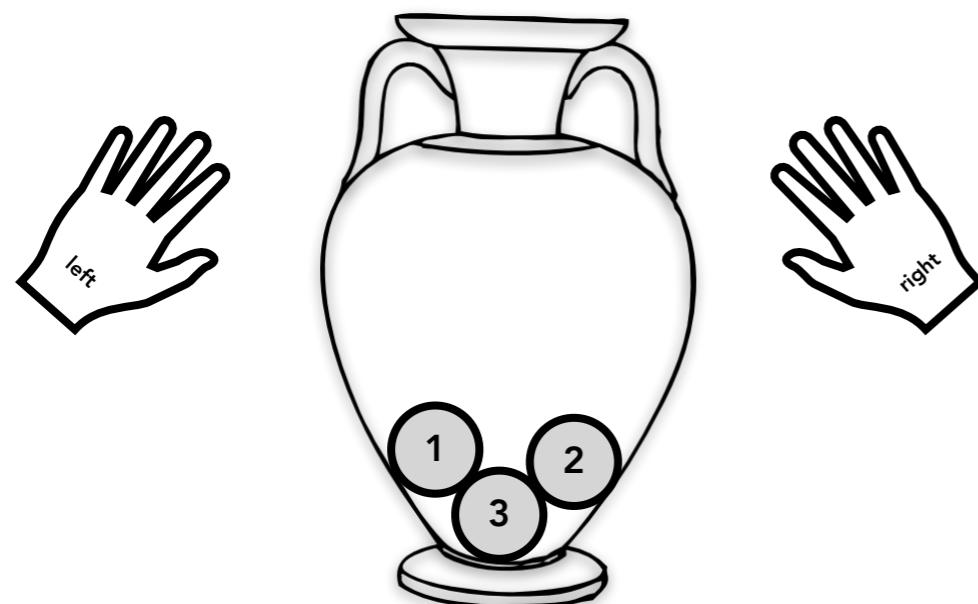
Counting possibilities

no stats class without urns!



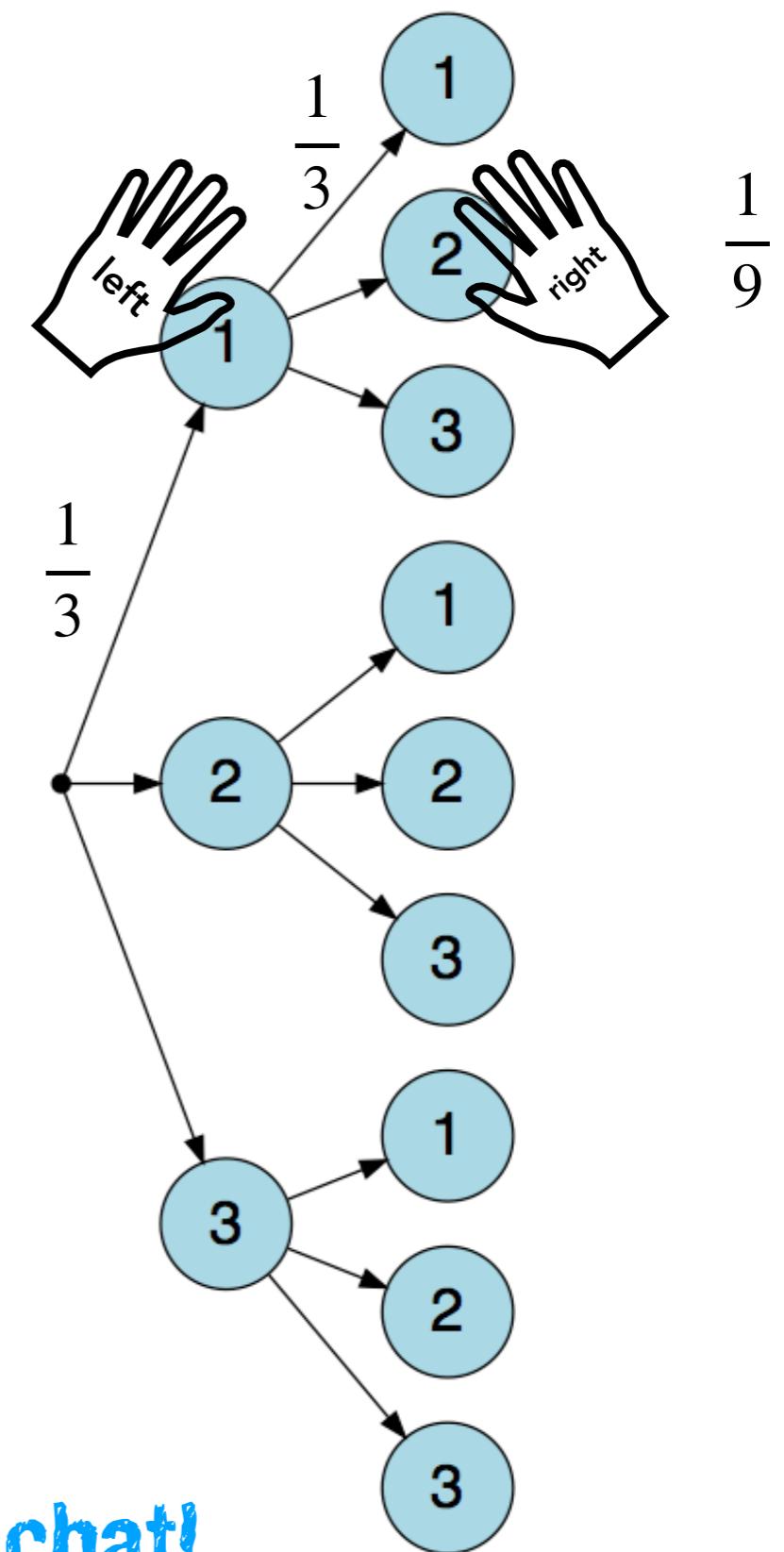
Sampling with replacement

$$p(\text{left} = 1, \text{right} = 2) = ?$$



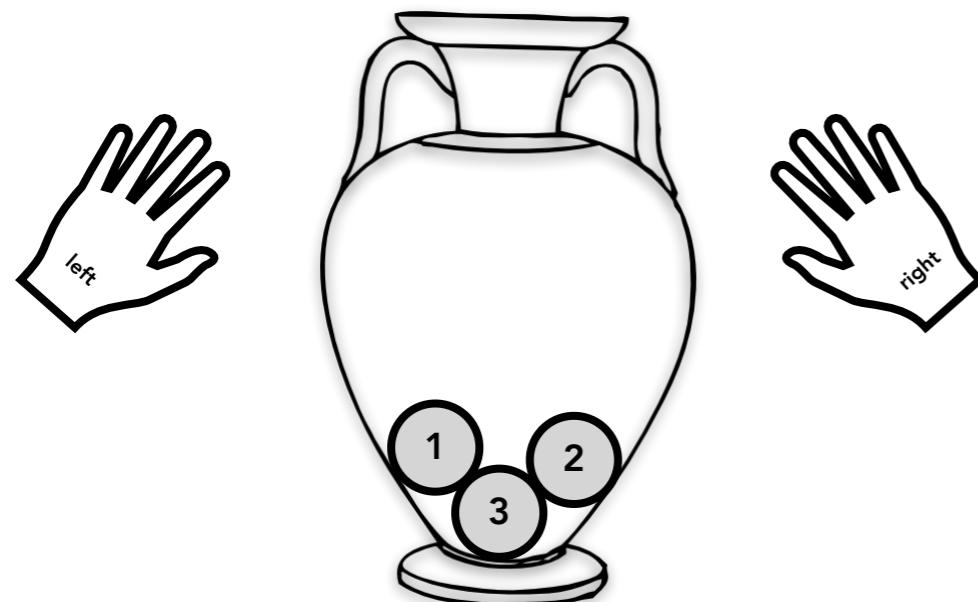
What is the probability that I first draw the 1 with my left hand, and then, after putting the 1 back into the urn again, draw the 2 with my right hand?

→ write in the chat!



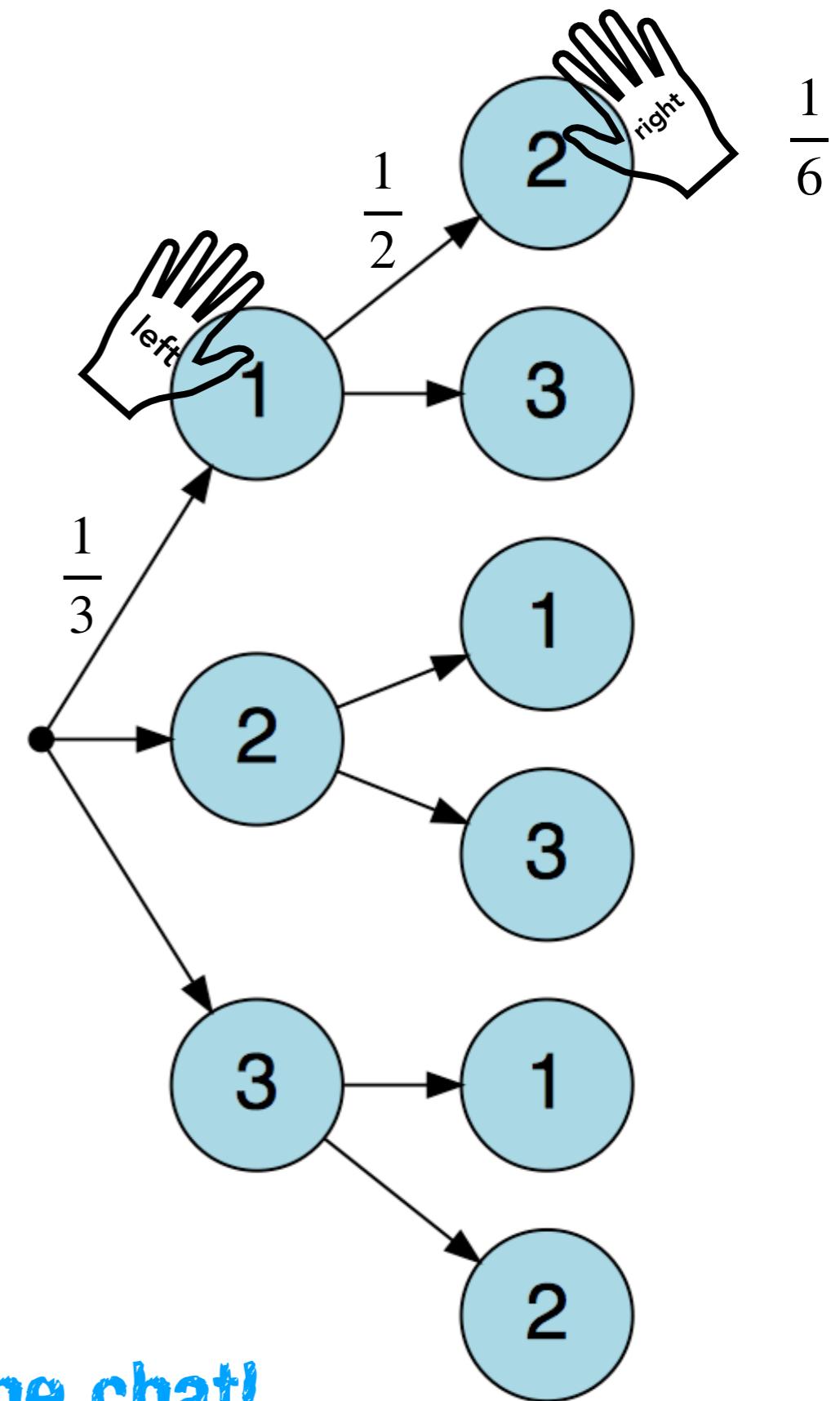
Sampling without replacement

$$p(\text{left} = 1, \text{right} = 2) = ?$$



What is the probability that I first draw the 1 with my left hand, and then, without putting the 1 back into the urn, draw the 2 with my right hand?

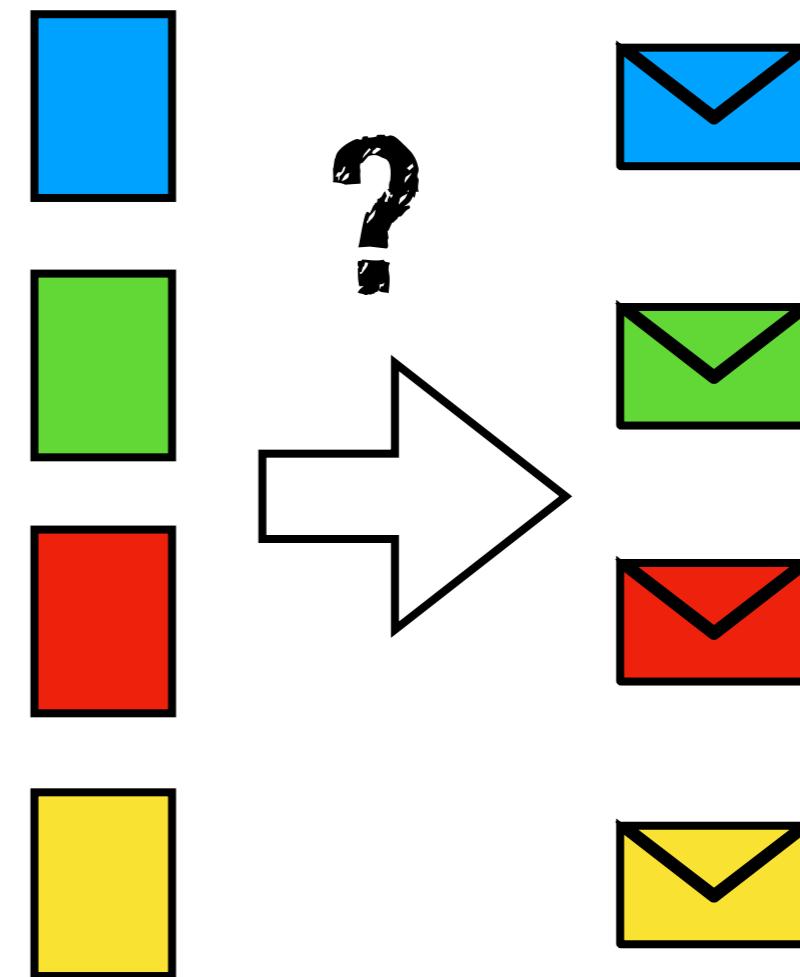
→ write in the chat!



Random secretary



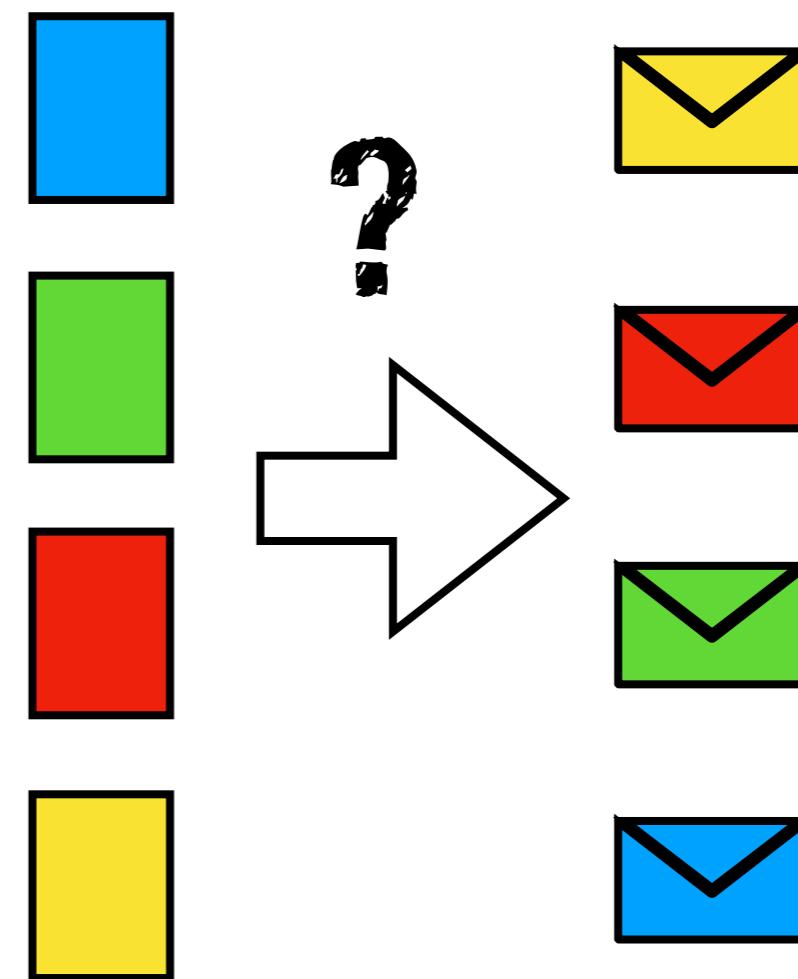
A secretary types four letters to four people and addresses the four envelopes. If he inserts the letters at random, each in a different envelope, what is the probability that exactly three letters will go into the right envelope?



Random secretary



A secretary types four letters to four people and addresses the four envelopes. If he inserts the letters at random, each in a different envelope, what is the probability that exactly three letters will go into the right envelope?

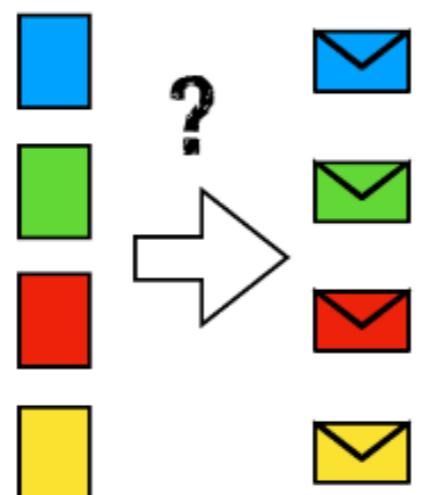


What is the probability that exactly three letters will go into the right envelope?

Random secretary



A secretary types four letters to four people and addresses the four envelopes. If he inserts the letters at random, each in a different envelope, what is the probability that exactly three letters will go into the right envelope?

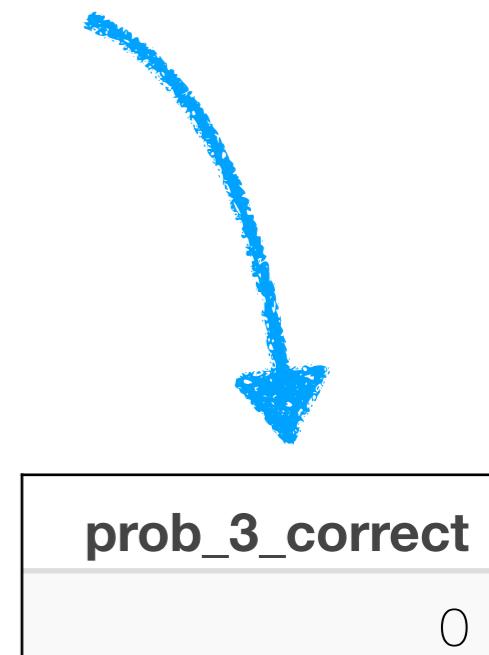


0% 25% 50% 75% 100%

Random secretary

```
1 df.letters = permutations(x = 1:4, k = 4) %>%
2   as_tibble(.name_repair = ~ str_c("person_", 1:4)) %>%
3   mutate(n_correct = (person_1 == 1) +
4         (person_2 == 2) +
5         (person_3 == 3) +
6         (person_4 == 4))
7
8 df.letters %>%
9   summarize(prob_3_correct = sum(n_correct == 3) / n())
```

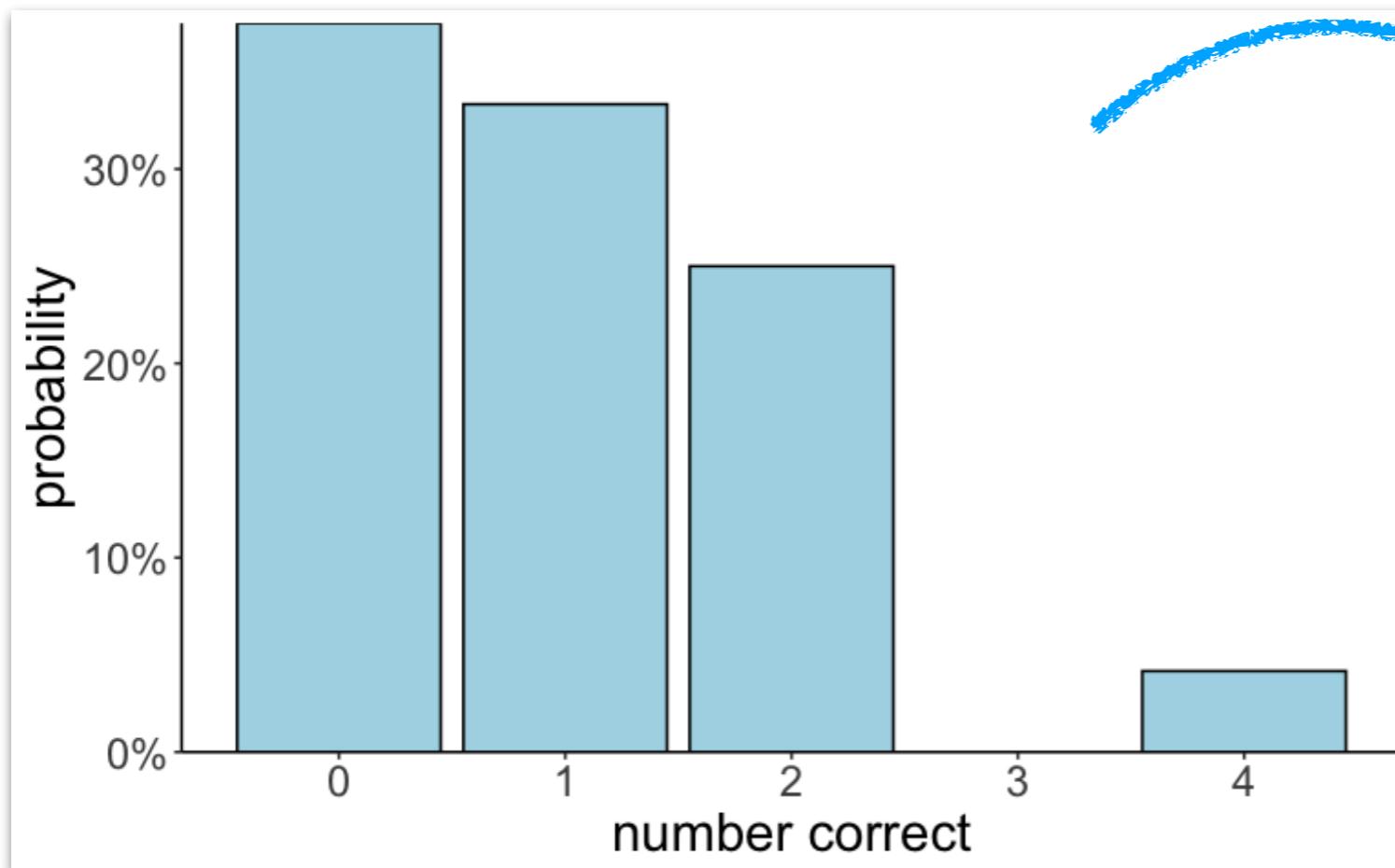
person_1	person_2	person_3	person_4
1	2	3	4
1	2	4	3
1	3	2	4
1	3	4	2
1	4	2	3
1	4	3	2
2	1	3	4
2	1	4	3
2	3	1	4



⋮ 24 rows total

Random secretary

```
1 ggplot(data = df.letters,
2         mapping = aes(x = n_correct)) +
3   geom_bar(aes(y = stat(count)/sum(count)),
4           color = "black",
5           fill = "lightblue") +
6   scale_y_continuous(labels = scales::percent,
7                      expand = c(0, 0)) +
8   labs(x = "number correct",
9        y = "probability")
```



probability of getting
0, 1, 2, 3, or 4
envelopes to the
correct person

Naive definition of probability

$$P_{\text{naive}}(A) = \frac{\text{number of outcomes favorable to } A}{\text{number of outcomes}}$$

if all outcomes are equally likely!

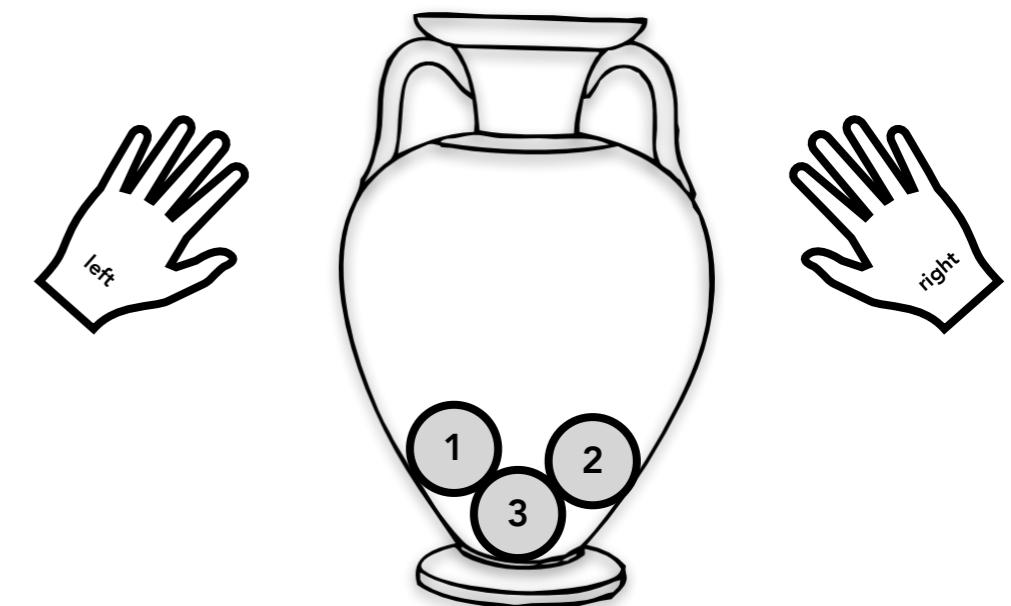
Definitions

Experiment: Activity that produces or observes an outcome.

Drawing 2 marbles from the urn with replacement, and noting the order.

Sample Space: Set of possible outcomes for an experiment.

$$\Omega = \{(1, 1), (1, 2), \dots, (2, 1), \dots, (3, 3)\}$$



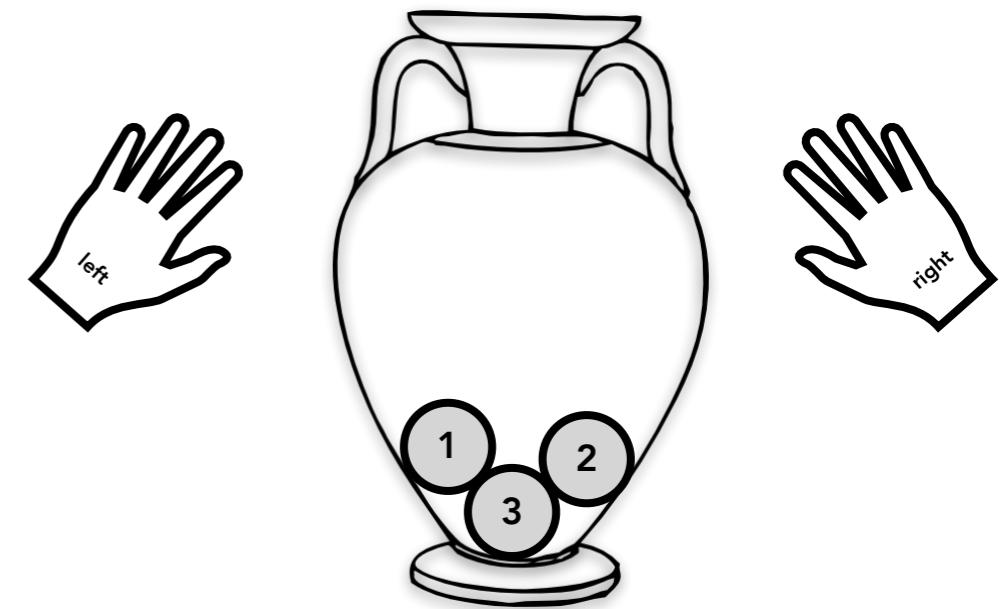
Event: Subset of the sample space. $(1, 1)$

Definitions

If $P(X_i)$ is the probability of event X_i

1. Probability cannot be negative.

$$P(X_i) \geq 0$$



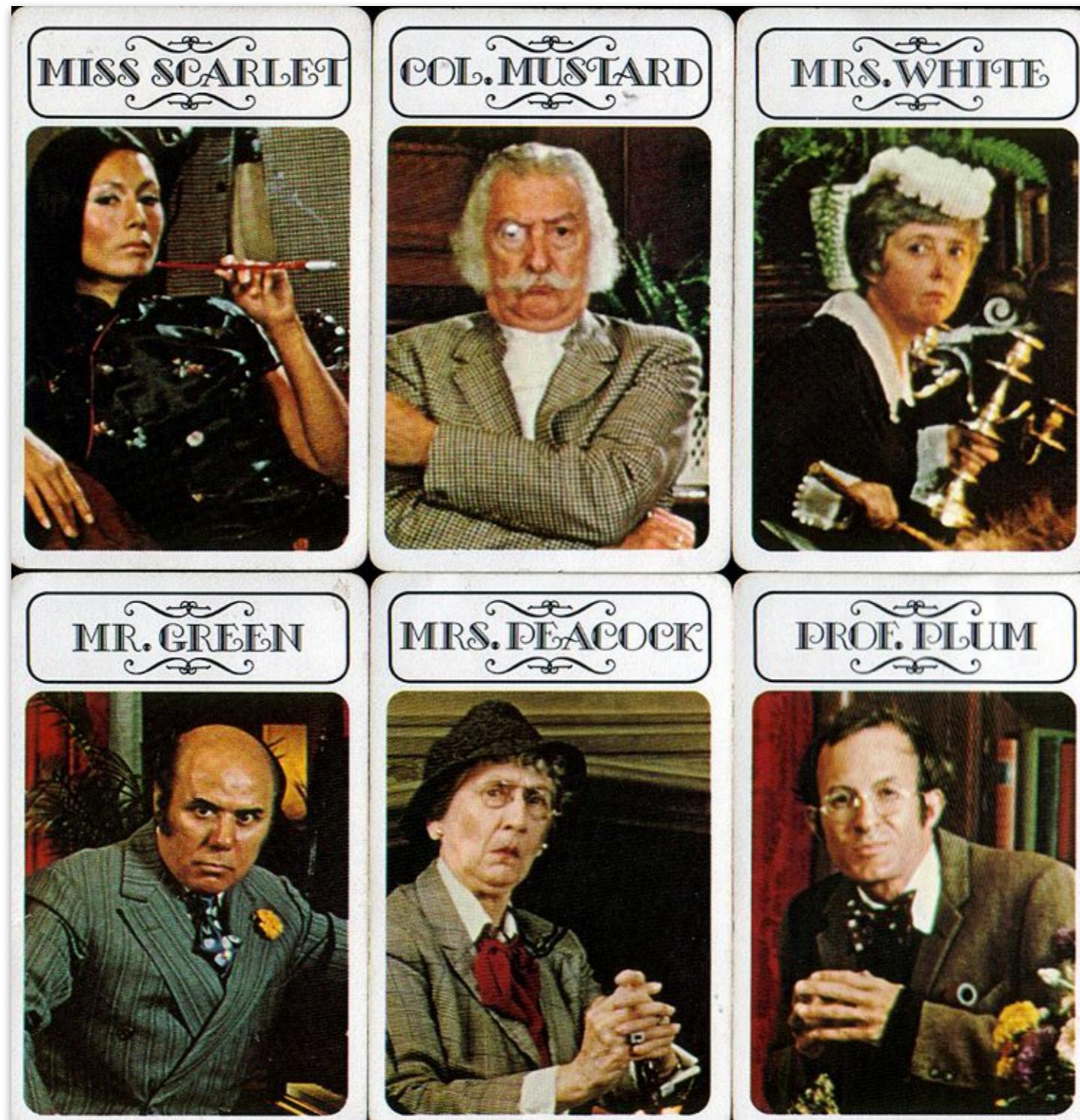
2. Total probability of all outcomes in the sample space is 1.

$$\sum_{i=1}^N P(X_i) = P(X_1) + P(X_2) + \dots + P(X_N) = 1$$

clue guide to probability

Clue guide to probability

Who killed Mr Boddy?



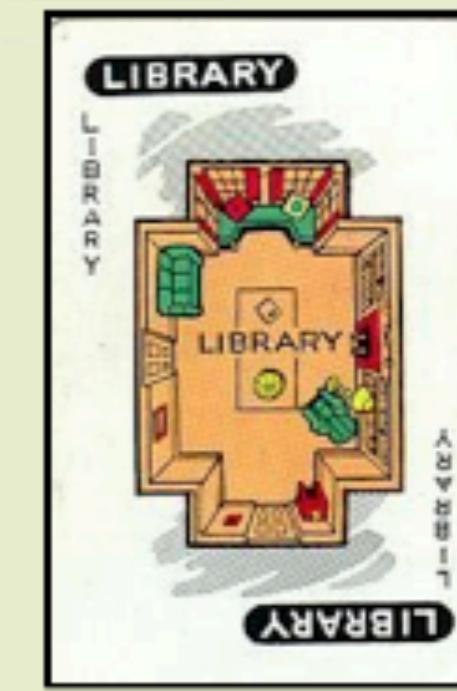
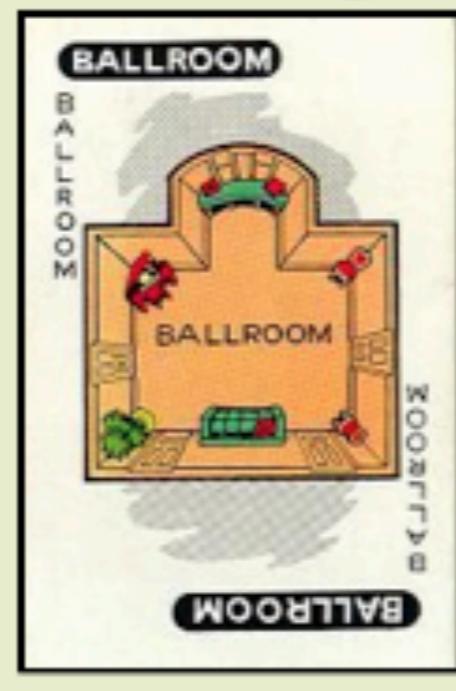
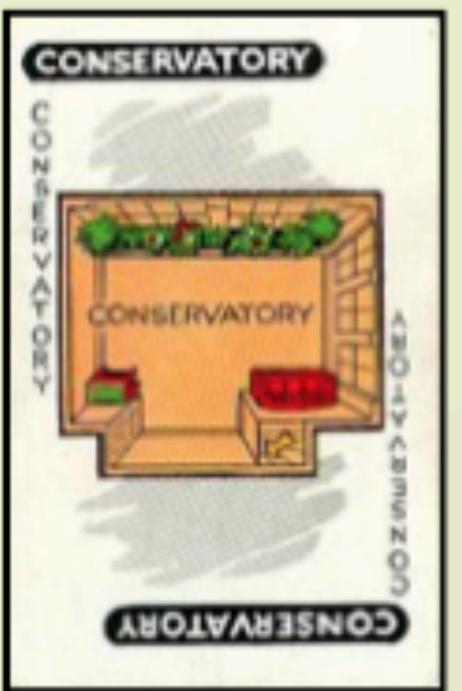
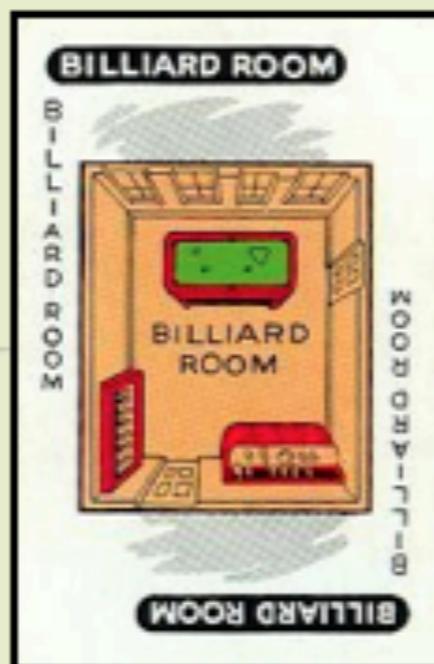
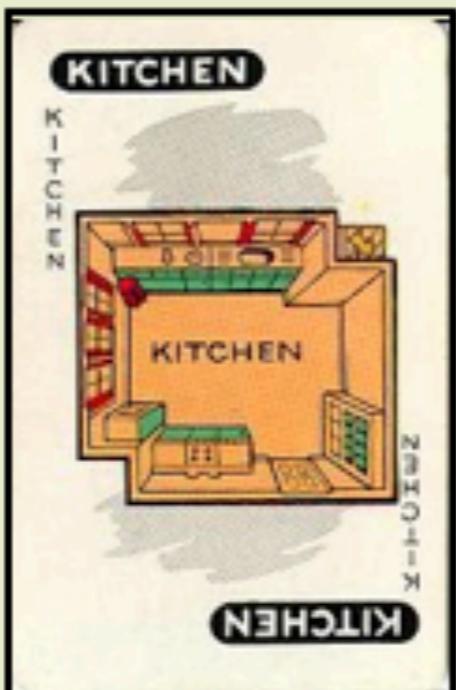
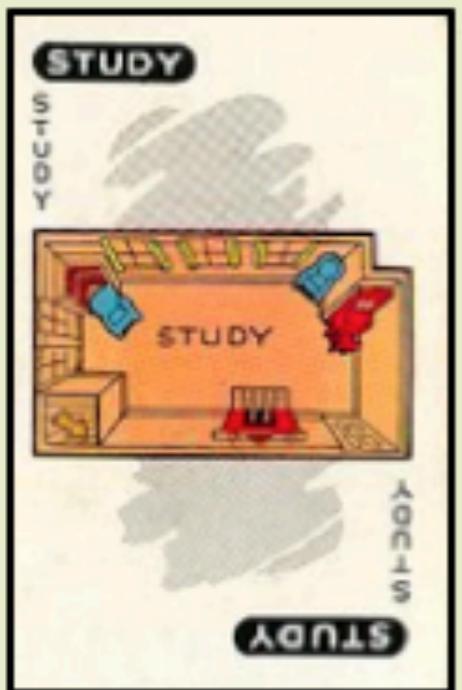
Clue guide to probability

Who killed Mr Boddy, **with what?**

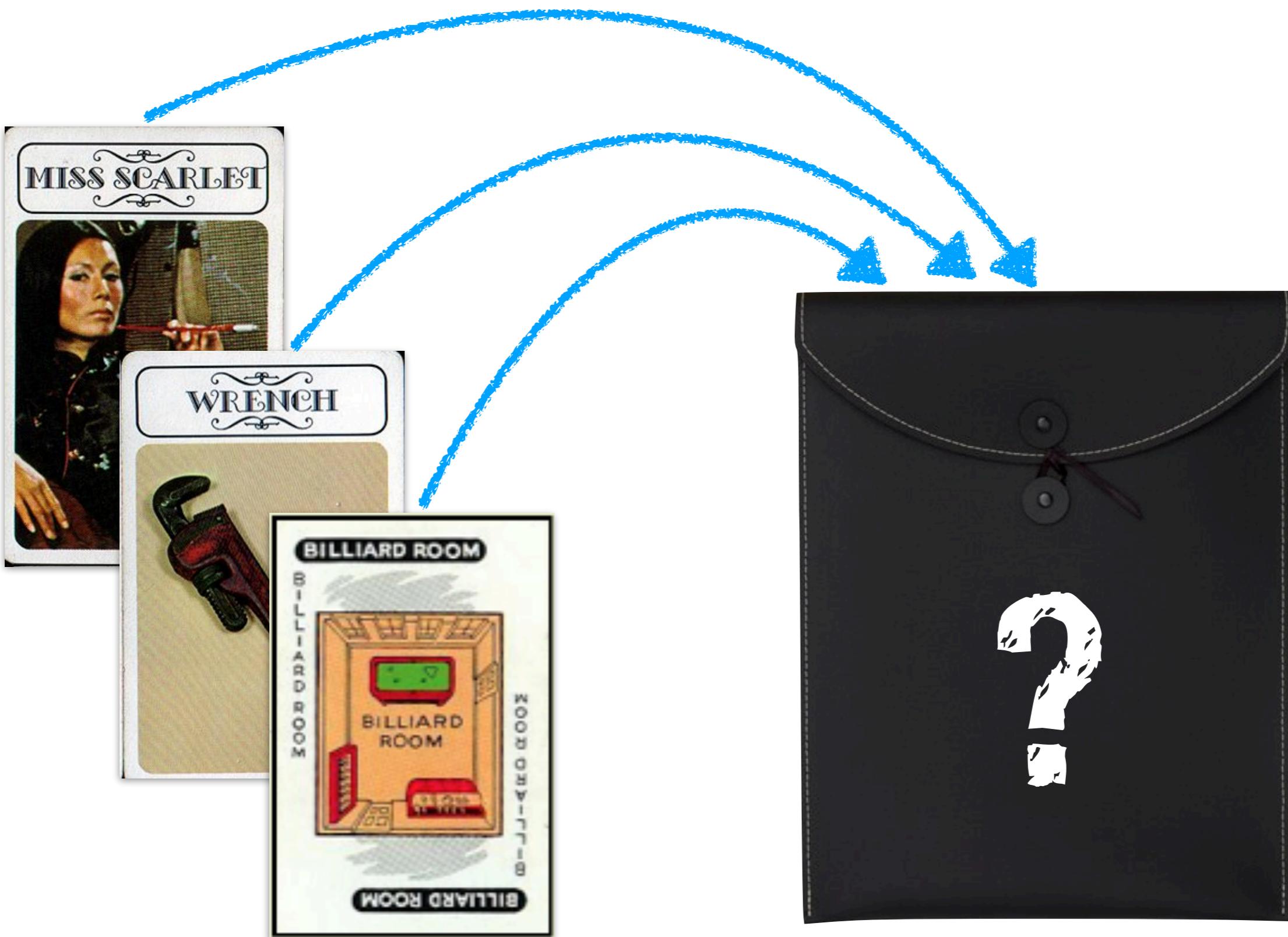


Clue guide to probability

Who killed Mr Boddy, with what, and where?



Clue guide to probability



Clue guide to probability

```
1 who = c("ms_scarlet", "col_mustard", "mrs_white",
2       "mr_green", "mrs_peacock", "prof_plum")
3 what = c("candlestick", "knife", "lead_pipe",
4         "revolver", "rope", "wrench")
5 where = c("study", "kitchen", "conservatory",
6           "lounge", "billiard_room", "hall",
7           "dining_room", "ballroom", "library")
8
9 df.clue = expand_grid(who = who,
10                      what = what,
11                      where = where)
```

all combinations

Ω

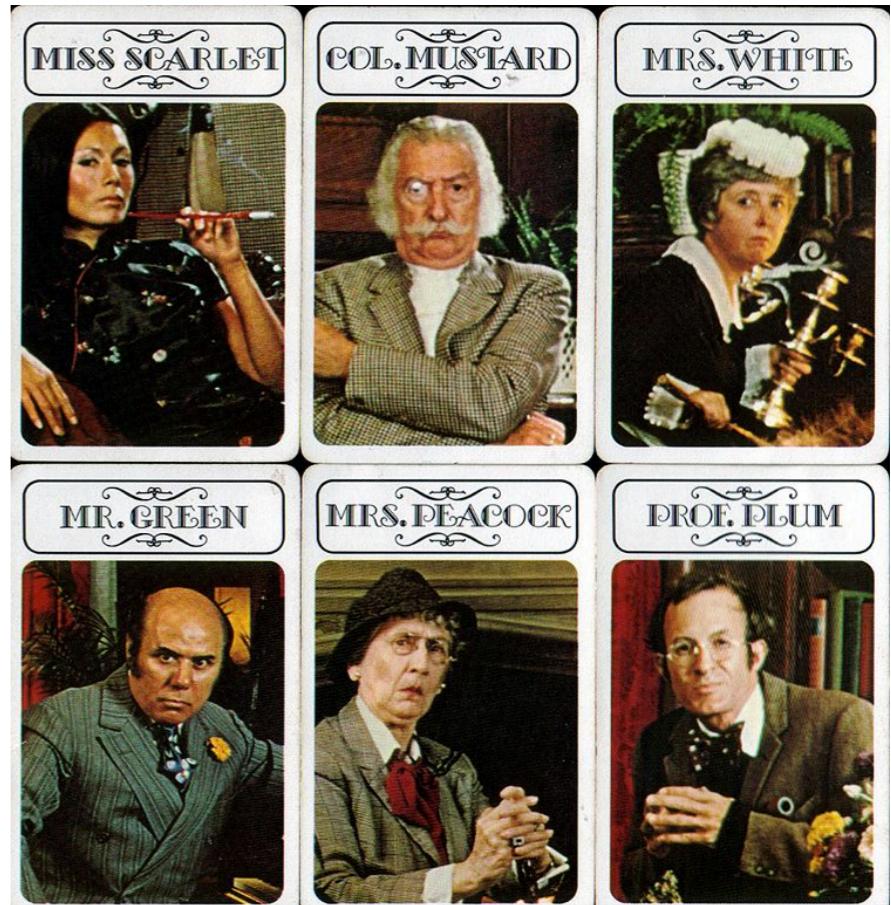
who	what	where
ms_scarlet	candlestick	study
ms_scarlet	candlestick	kitchen
ms_scarlet	candlestick	conservatory
ms_scarlet	candlestick	lounge
	⋮	

nrow(df.clue) = 324

Clue guide to probability

Who?

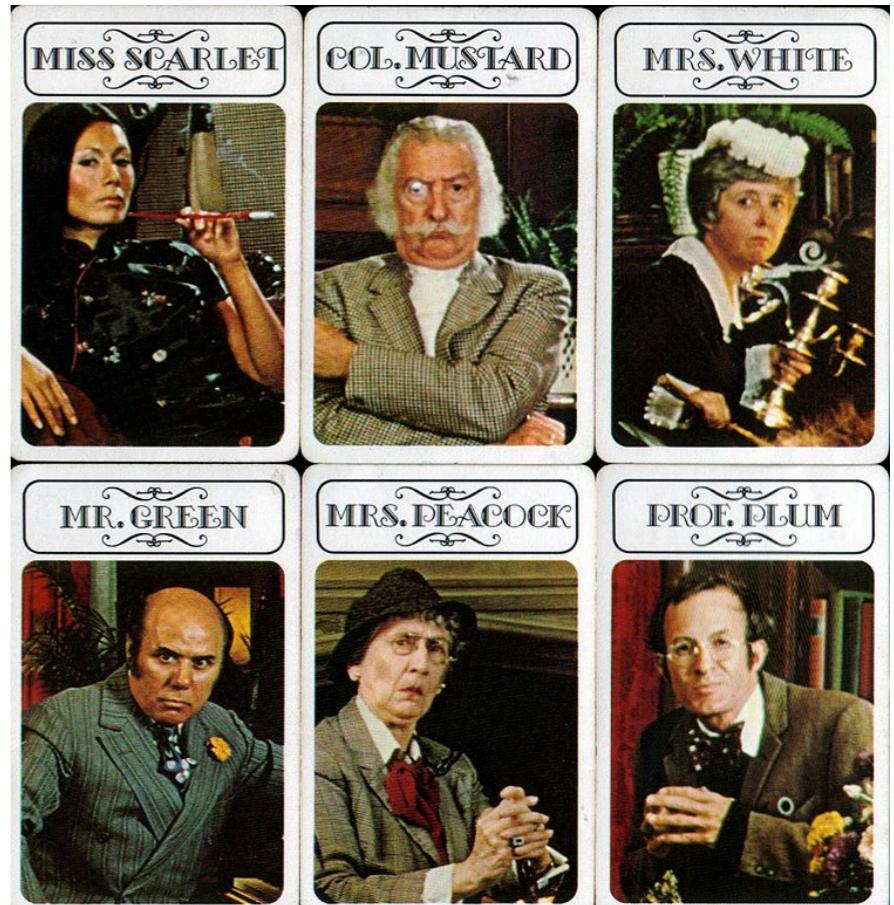
- 6 suspects
- mutually exclusive and exhaustive
- $p(\text{who} = \text{one of the six}) = 1$
- each equally likely a priori
- $p(\text{who} = \text{Prof. Plum}) = \frac{1}{6}$



Clue guide to probability

Who?

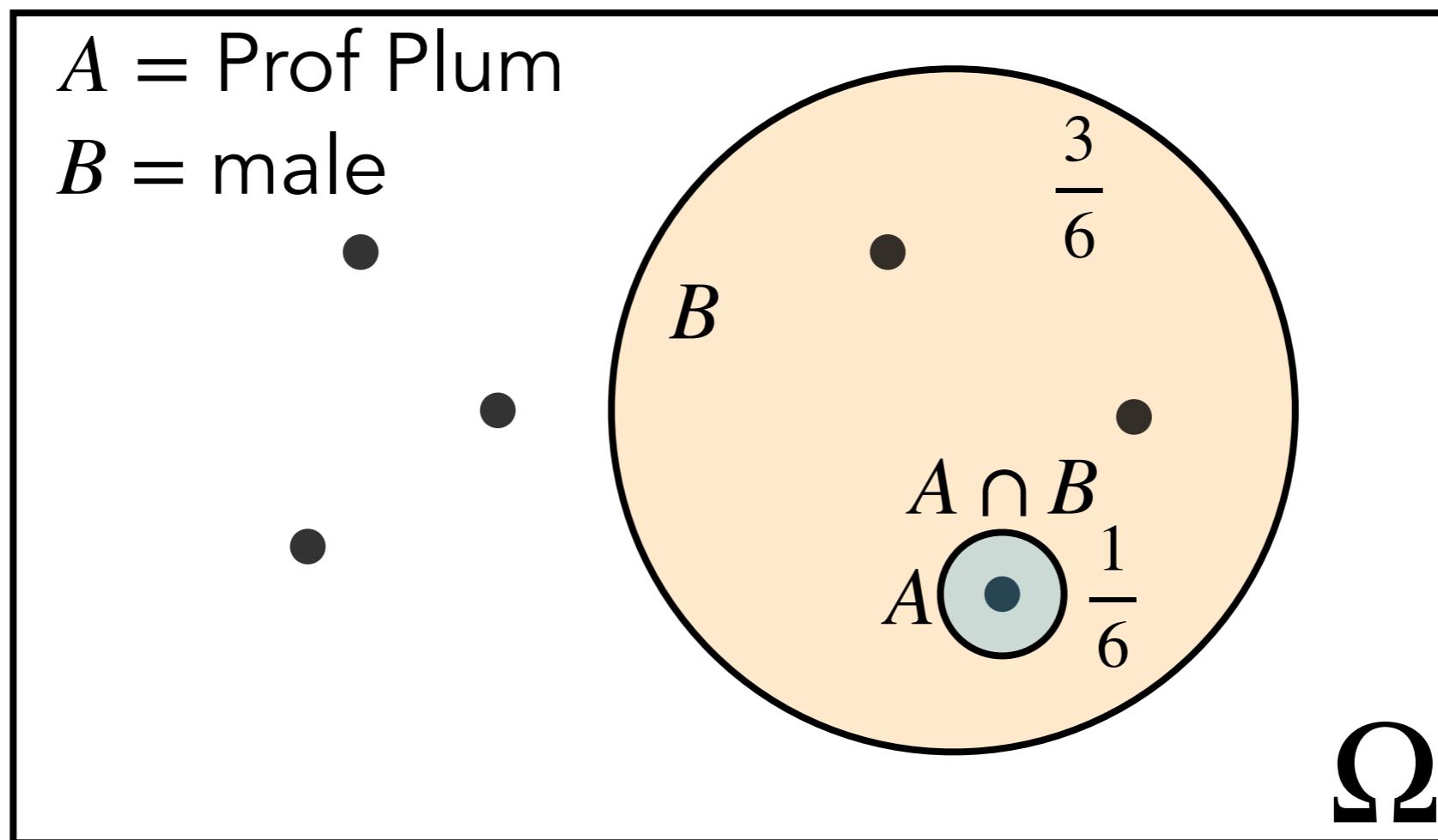
- *conditional probability:*
- $p(A | B)$ (probability of A given B)
- **Definition:** $p(A | B) = \frac{p(A, B)}{p(B)}$



Probability that it was Prof Plum, given that the murderer was male?

$$p(\text{Prof. Plum} | \text{male}) = ?$$

Clue guide to probability



Probability that it was Prof Plum, given that the murderer was male?

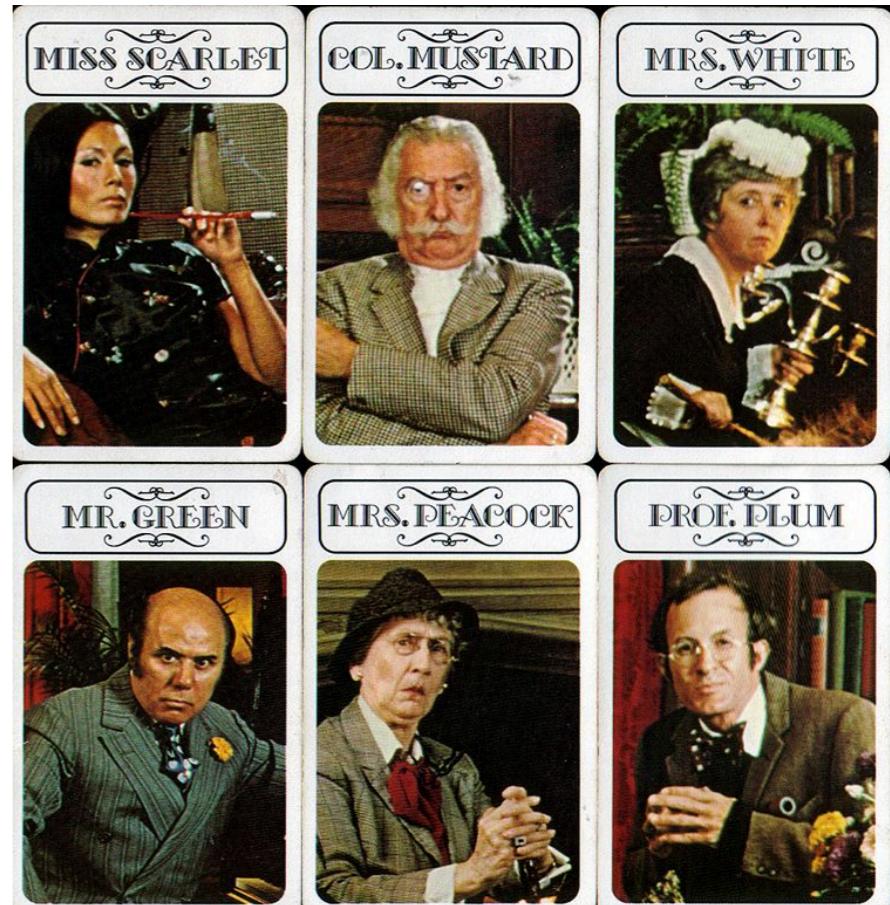
Definition: $p(A | B) = \frac{p(A, B)}{p(B)} = \frac{1}{3}$

$$p(A) = \frac{1}{6} \quad p(A, B) = \frac{1}{6} \quad p(B) = \frac{3}{6}$$

Clue guide to probability

Who?

- conditional probability:
- $p(A | B)$ (probability of A given B)
- **Definition:** $p(A | B) = \frac{p(A, B)}{p(B)}$
- $p(\text{Prof. Plum} | \text{male}) = \frac{1/6}{1/2} = 1/3$



who	gender
col_mustard	male
mr_green	male
prof_plum	male
ms_scarlet	female
mrs_white	female
mrs_peacock	female

```

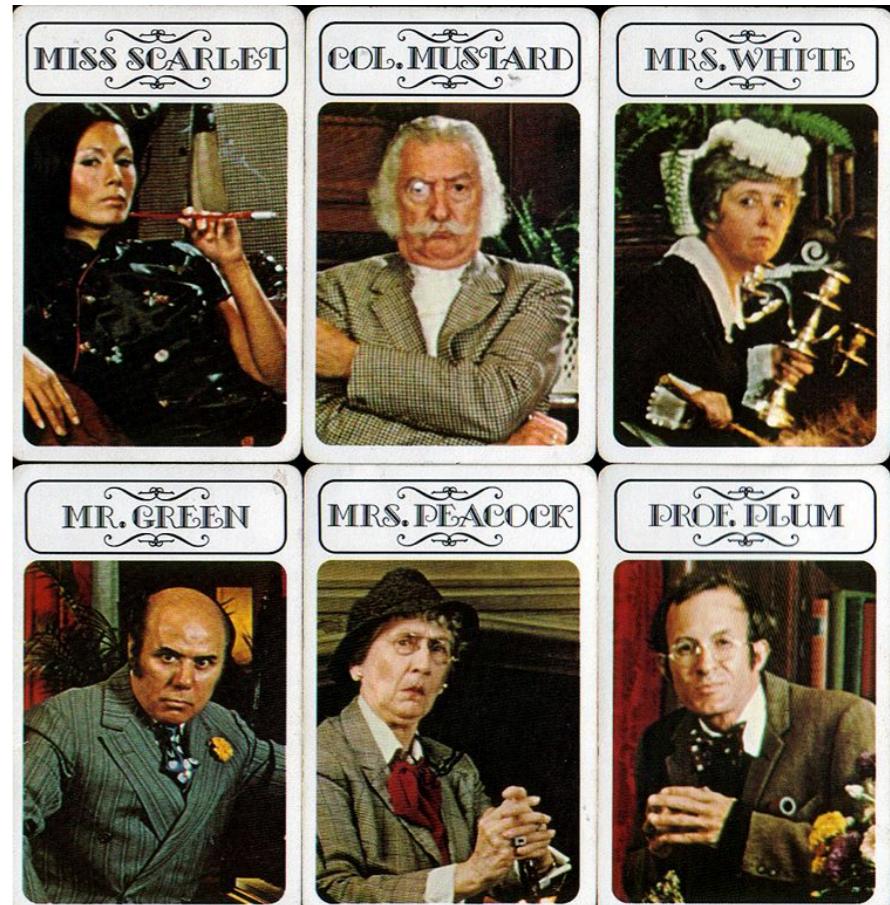
1 df.suspects = df.clue %>%
2   distinct(who) %>%
3   mutate(gender = ifelse(
4     test = who %in% c("ms_scarlet",
5                           "mrs_white",
5                           "mrs_peacock"),
6     yes = "female",
6     no = "male"))

```

Clue guide to probability

Who?

- conditional probability:
- $p(A | B)$ (probability of A given B)
- **Definition:** $p(A | B) = \frac{p(A, B)}{p(B)}$
- $p(\text{Prof. Plum} | \text{male}) = \frac{1/6}{1/2} = 1/3$



who	gender
col_mustard	male
mr_green	male
prof_plum	male
ms_scarlet	female
mrs_white	female
mrs_peacock	female

```

1 df.suspects %>%
2   summarize(p_prof_plum_given_male =
3     sum(gender == "male" &
4       who == "prof_plum") /
5     sum(gender == "male"))

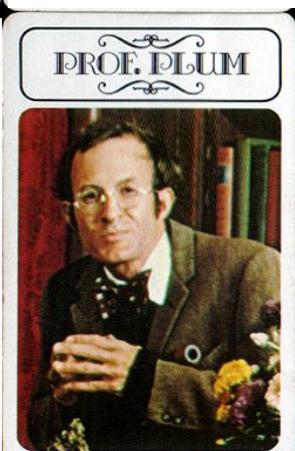
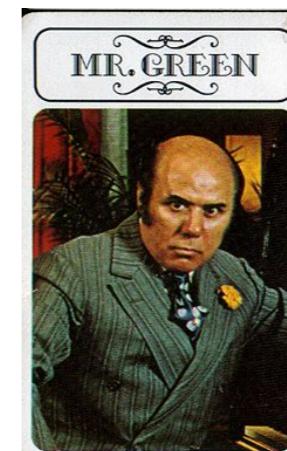
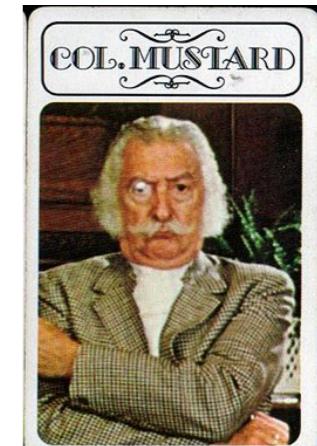
```

use naive definition of probability

Clue guide to probability

Who?

- *conditional probability:*
- $p(A | B)$ (probability of A given B)
- **Definition:** $p(A | B) = \frac{p(A, B)}{p(B)}$
- $p(\text{Prof. Plum} | \text{male}) = \frac{1/6}{1/2} = 1/3$

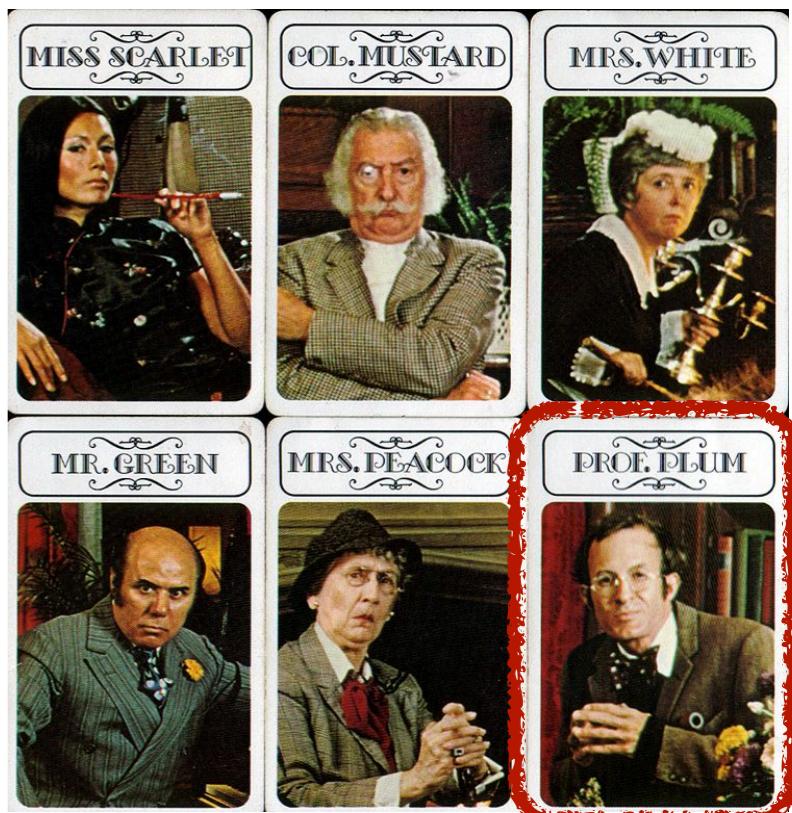


who	gender
col_mustard	male
mr_green	male
prof_plum	male

```
1 df.suspects %>%
2   filter(gender == "male") %>%
3   summarize(p_prof_plum_given_male =
4             sum(who == "prof_plum") /
5             n())
```

Clue guide to probability

Who?



- *independence:*
- A and B are independent if
- **Definition:** $p(A | B) = p(A)$
- (probability of A does not change if you know B)

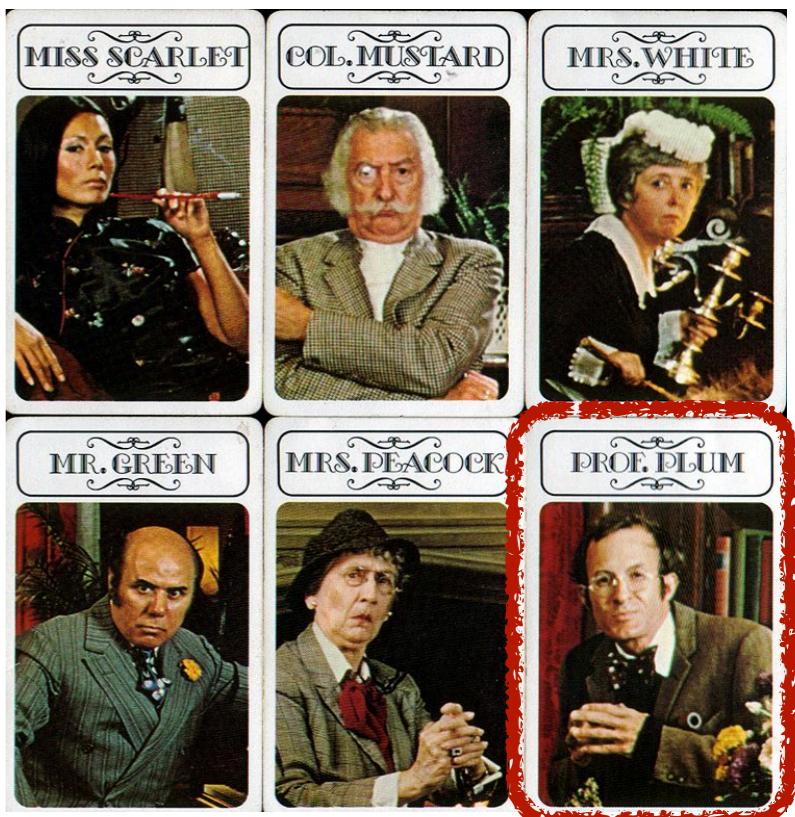
What?



- $p(\text{Prof Plum} | \text{candle stick}) = p(\text{Prof Plum})$
- each card (who and what) is drawn from a separate pack of cards

Clue guide to probability

Who?

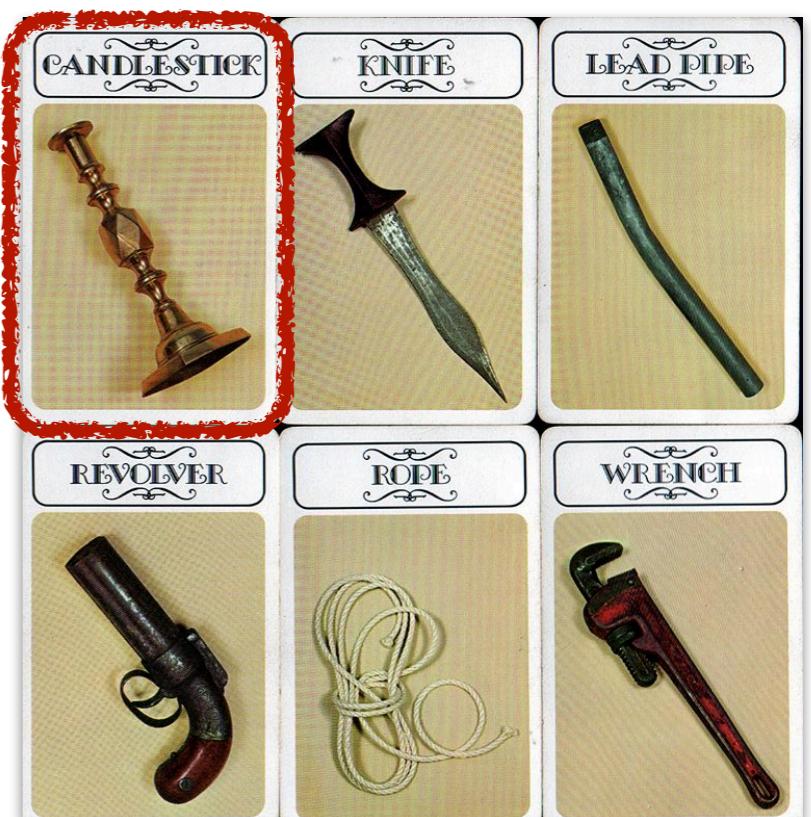


- joint probability:
- if A and B are independent then
- **Definition:** $p(A, B) = p(A) \cdot p(B)$

- $p(\text{Prof Plum, candle stick}) =$
 $p(\text{Prof Plum}) \cdot p(\text{candle stick}) =$

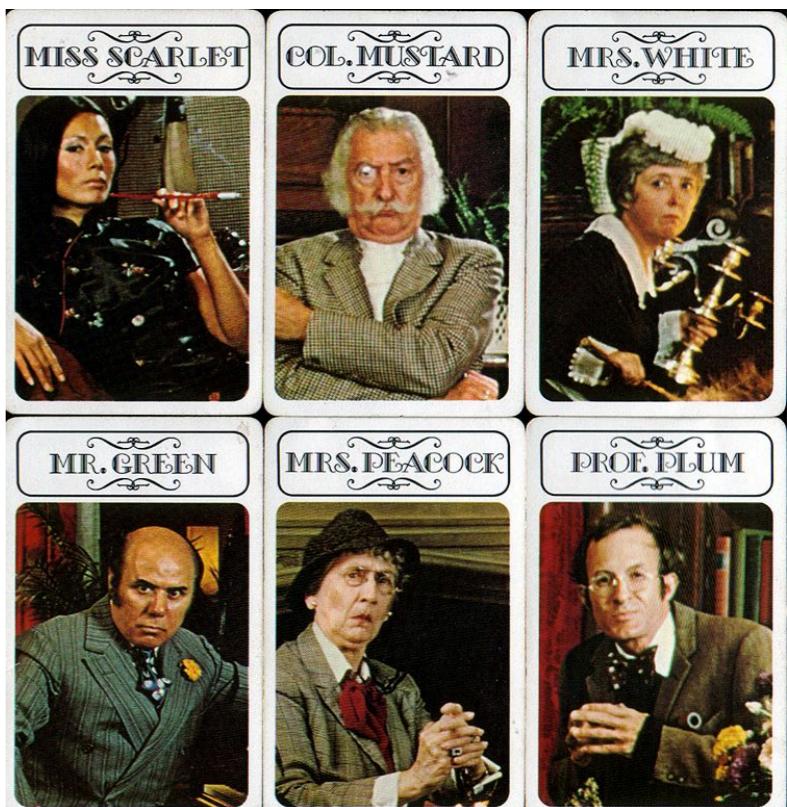
$$\frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$$

What?



Clue guide to probability

Who?



- dependence:
- **Definition:** $p(A | B) \neq p(A)$
- **Definition:** $p(A, B) = p(A) \cdot p(B | A)$
- if women were more likely than men to use the revolver then
- $p(\text{Mrs. White} | \text{Revolver}) > p(\text{Mrs. White})$

What?



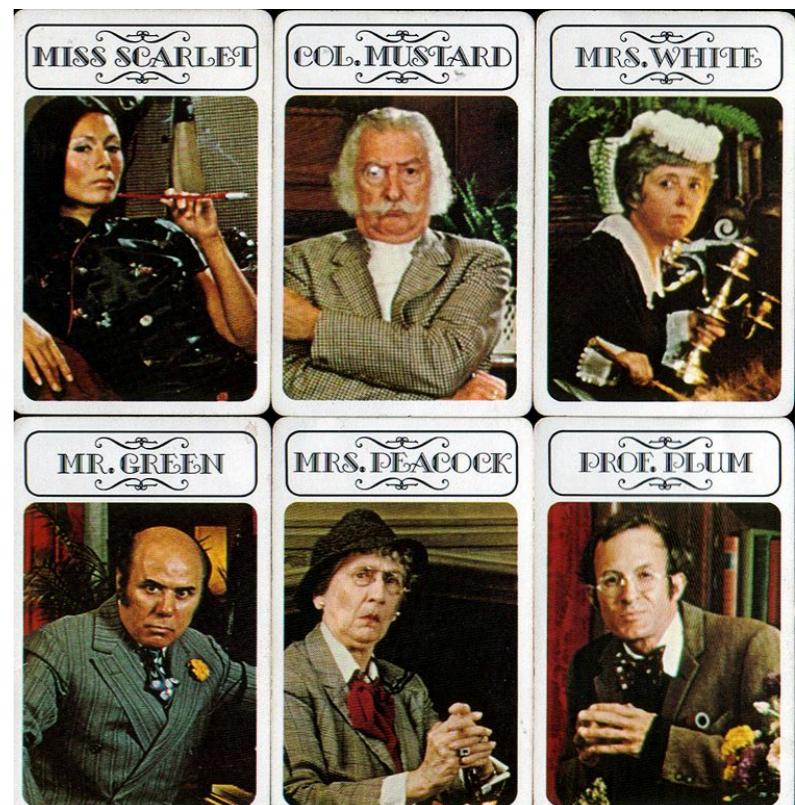
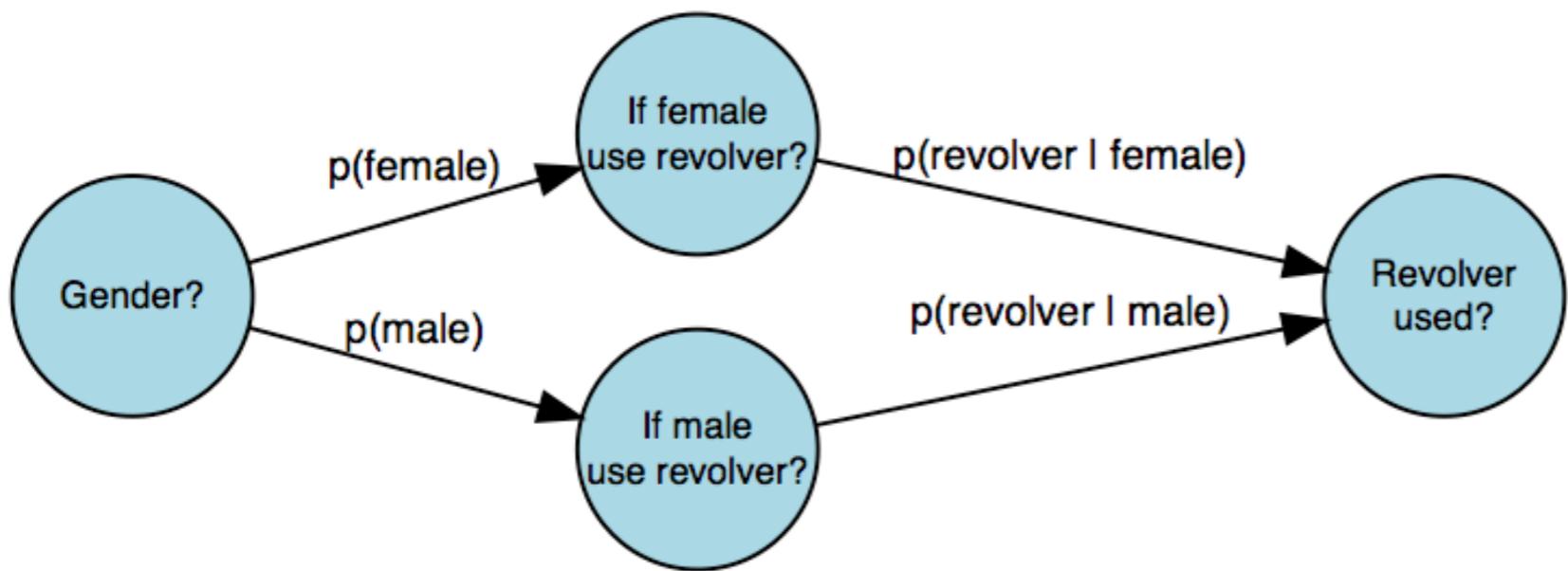
Clue guide to probability

- law of total probability
- Definition:

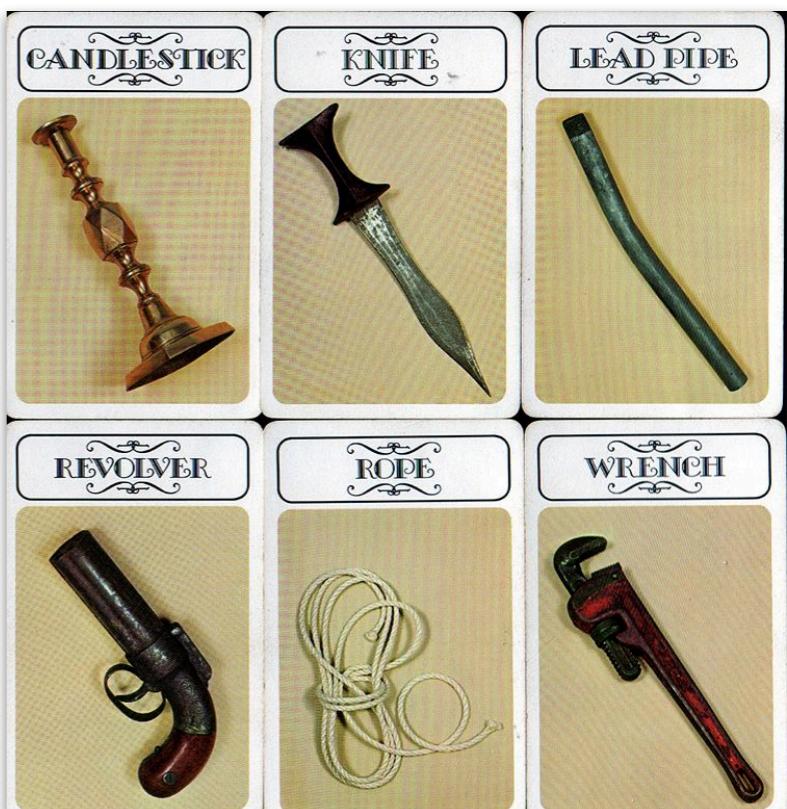
$$p(A) = p(A | B) \cdot p(B) + p(A | \neg B) \cdot p(\neg B)$$

$$p(A) = \sum_{i=1}^n p(A | B_i) \cdot p(B_i)$$

$p(\text{what} = \text{Revolver}) = ?$



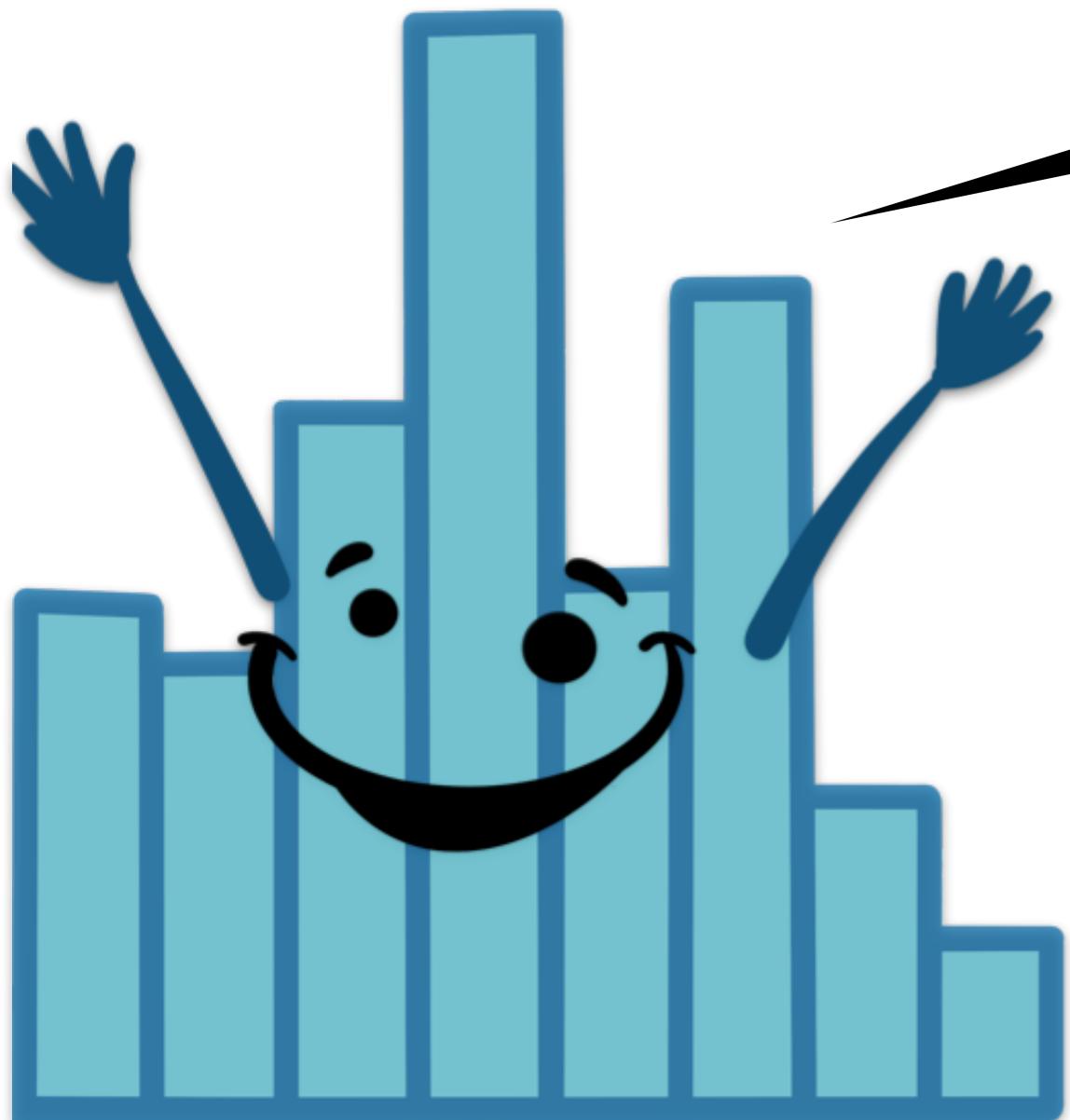
Who?



We're listening to "The Bird"
by "Anderson .Paak"
submitted by Ari Beller

02:00

stretch break!



Understanding Bayes' rule

Clue guide to probability

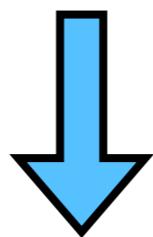
Bayes Theorem in a few steps



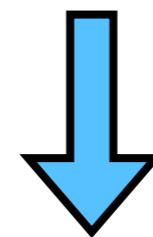
Clue guide to probability

- Bayes' theorem (derivation)

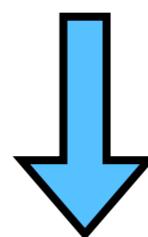
$$p(B | A) = \frac{p(A, B)}{p(A)}$$



$$p(A | B) = \frac{p(A, B)}{p(B)}$$



$$p(A, B) = p(B | A) \cdot p(A) = p(A | B) \cdot p(B)$$



$$p(B | A) = \frac{p(A | B) \cdot p(B)}{p(A)}$$

Clue guide to probability

$$p(B|A) = \frac{p(A|B) \cdot p(B)}{p(A)}$$

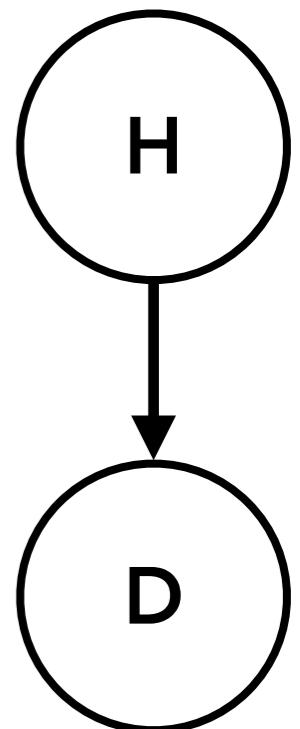
posterior **likelihood** **prior**

$$p(H|D) = \frac{p(D|H) \cdot p(H)}{p(D)}$$

subjective probability
interpretation

H = Hypothesis

D = Data



formal framework for learning from data

updating our prior belief $p(H)$, to a posterior belief $p(H|D)$
given some data

Clue guide to probability

posterior $p(H|D) = \frac{\text{likelihood} \cdot \text{prior}}{p(D)}$ $H = \text{Hypothesis}$
 $D = \text{Data}$

probability of the data?!

law of total probability:

$$p(D) = \sum_{i=1}^n p(D|H_i) \cdot p(H_i)$$

**take into account all the different ways
in which the data could have come about**

Clue guide to probability

A patient named Fred is tested for a disease called *conditionitis*, a medical condition that afflicts **1% of the population**. The test result is positive, i.e., the test claims that Fred has the disease. Let **D** be the event that Fred has the disease and **T** be the event that he tests positive.

Suppose that the test is “95% accurate”; there are different measures of the accuracy of a test, but in this problem it is assumed to mean that **P(T|D) = 0.95** and **P(¬T|¬D) = 0.95**. The quantity $P(T|D)$ is known as the *sensitivity* (= true positive rate) of the test, and $P(\neg T|\neg D)$ is known as the *specificity* (= true negative rate).

Find the conditional probability that Fred has *conditionitis*, given the evidence provided by the positive test result.

What's the probability that Fred has conditionitis?

A patient named Fred is tested for a disease called *conditionitis*, a medical condition that afflicts **1% of the population**. The test result is positive, i.e., the test claims that Fred has the disease. Let **D** be the event that Fred has the disease and **T** be the event that he tests positive.

Suppose that the test is "95% accurate"; there are different measures of the accuracy of a test, but in this problem it is assumed to mean that $P(T|D) = 0.95$ and $P(\neg T|\neg D) = 0.95$. The quantity $P(T|D)$ is known as the *sensitivity* (= true positive rate) of the test, and $P(\neg T|\neg D)$ is known as the *specificity* (= true negative rate).

Find the conditional probability that Fred has *conditionitis*, given the evidence provided by the positive test result.

- 1% 16% 50% 73% 95%

Breakout rooms

Tasks: A patient named Fred is tested for a disease called *conditionitis*, a medical condition that afflicts **1% of the population**. The test result is positive, i.e., the test claims that Fred has the disease. Let **D** be the event that Fred has the disease and **T** be the event that he tests positive.

Suppose that the test is “95% accurate”; there are different measures of the accuracy of a test, but in this problem it is assumed to mean that **$P(T|D) = 0.95$** and **$P(\neg T|\neg D) = 0.95$** . The quantity $P(T|D)$ is known as the *sensitivity* (= true positive rate) of the test, and $P(\neg T|\neg D)$ is known as the *specificity* (= true negative rate).

Find the conditional probability that Fred has *conditionitis*, given the evidence provided by the positive test result.

Size: ~3 people

Time: 5 minutes

Report: We will vote again.



What's the probability that Fred has conditionitis?

A patient named Fred is tested for a disease called *conditionitis*, a medical condition that afflicts **1% of the population**. The test result is positive, i.e., the test claims that Fred has the disease. Let **D** be the event that Fred has the disease and **T** be the event that he tests positive.

Suppose that the test is "95% accurate"; there are different measures of the accuracy of a test, but in this problem it is assumed to mean that $P(T|D) = 0.95$ and $P(\neg T|\neg D) = 0.95$. The quantity $P(T|D)$ is known as the *sensitivity* (= true positive rate) of the test, and $P(\neg T|\neg D)$ is known as the *specificity* (= true negative rate).

Find the conditional probability that Fred has *conditionitis*, given the evidence provided by the positive test result.

1% 16% 50% 73% 95%

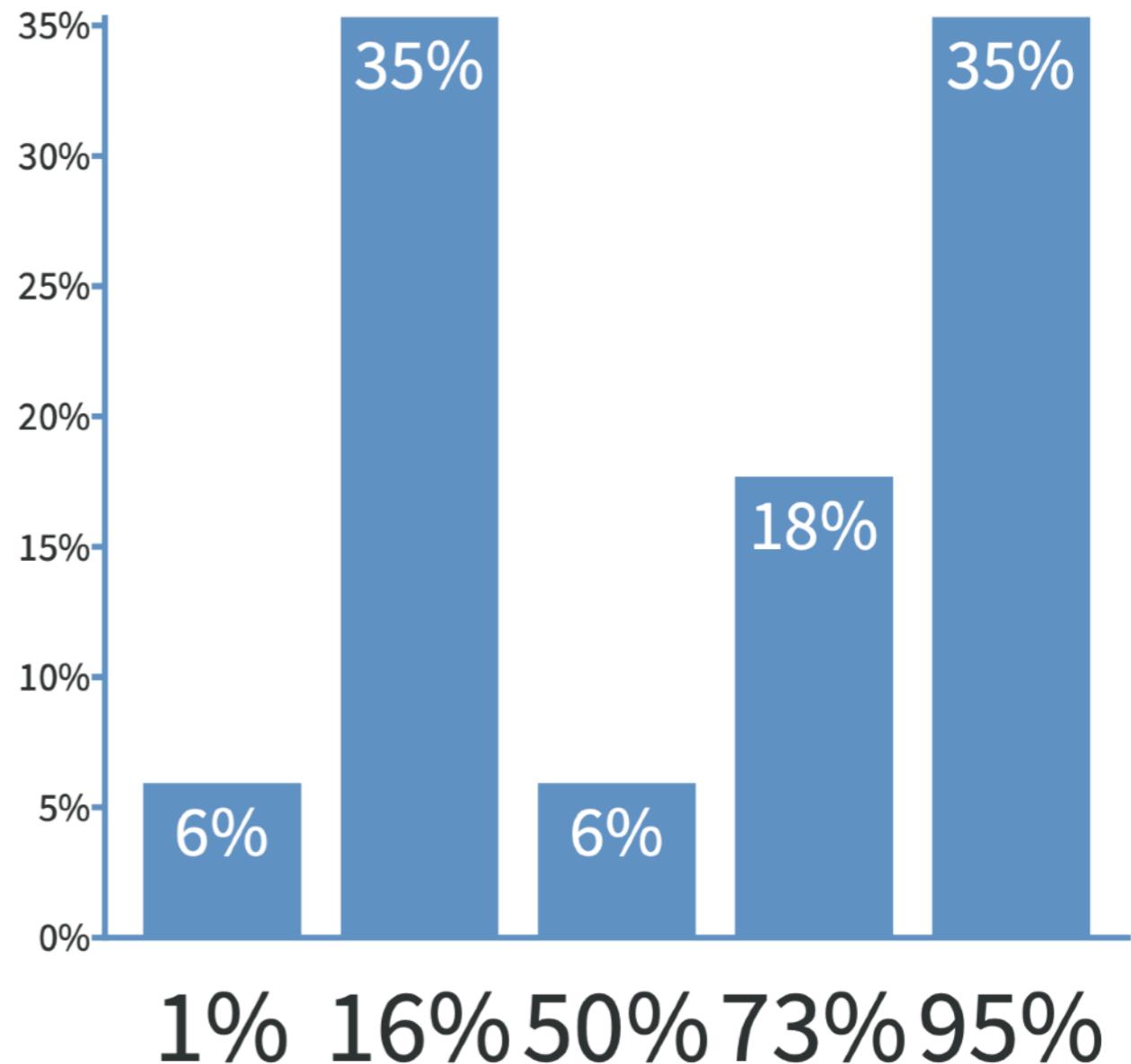


What's the probability that Fred has conditionitis?

A patient named Fred is tested for a disease called *conditionitis*, a medical condition that afflicts **1% of the population**. The test result is positive, i.e., the test claims that Fred has the disease. Let **D** be the event that Fred has the disease and **T** be the event that he tests positive.

Suppose that the test is "95% accurate"; there are different measures of the accuracy of a test, but in this problem it is assumed to mean that $P(T|D) = 0.95$ and $P(\neg T|\neg D) = 0.95$. The quantity $P(T|D)$ is known as the *sensitivity* (= true positive rate) of the test, and $P(\neg T|\neg D)$ is known as the *specificity* (= true negative rate).

Find the conditional probability that Fred has *conditionitis*, given the evidence provided by the positive test result.



Clue guide to probability

what we know

$$P(D) = 0.01$$

$$P(T|D) = 0.95$$

$$P(T|\neg D) = 0.05$$

what we want to know

$$P(D|T) = ?$$

$$p(D|T) = \frac{p(T|D) \cdot p(D)}{p(T)} \text{ Bayes' rule}$$

$$p(D|T) = \frac{p(T|D) \cdot p(D)}{p(T|D) \cdot p(D) + p(T|\neg D) \cdot p(\neg D)}$$



law of total
probability

$$p(D|T) = \frac{0.95 \cdot 0.01}{0.95 \cdot 0.01 + 0.05 \cdot 0.99} \approx 0.16$$

Clue guide to probability

what we know

$$P(D) = 0.01$$

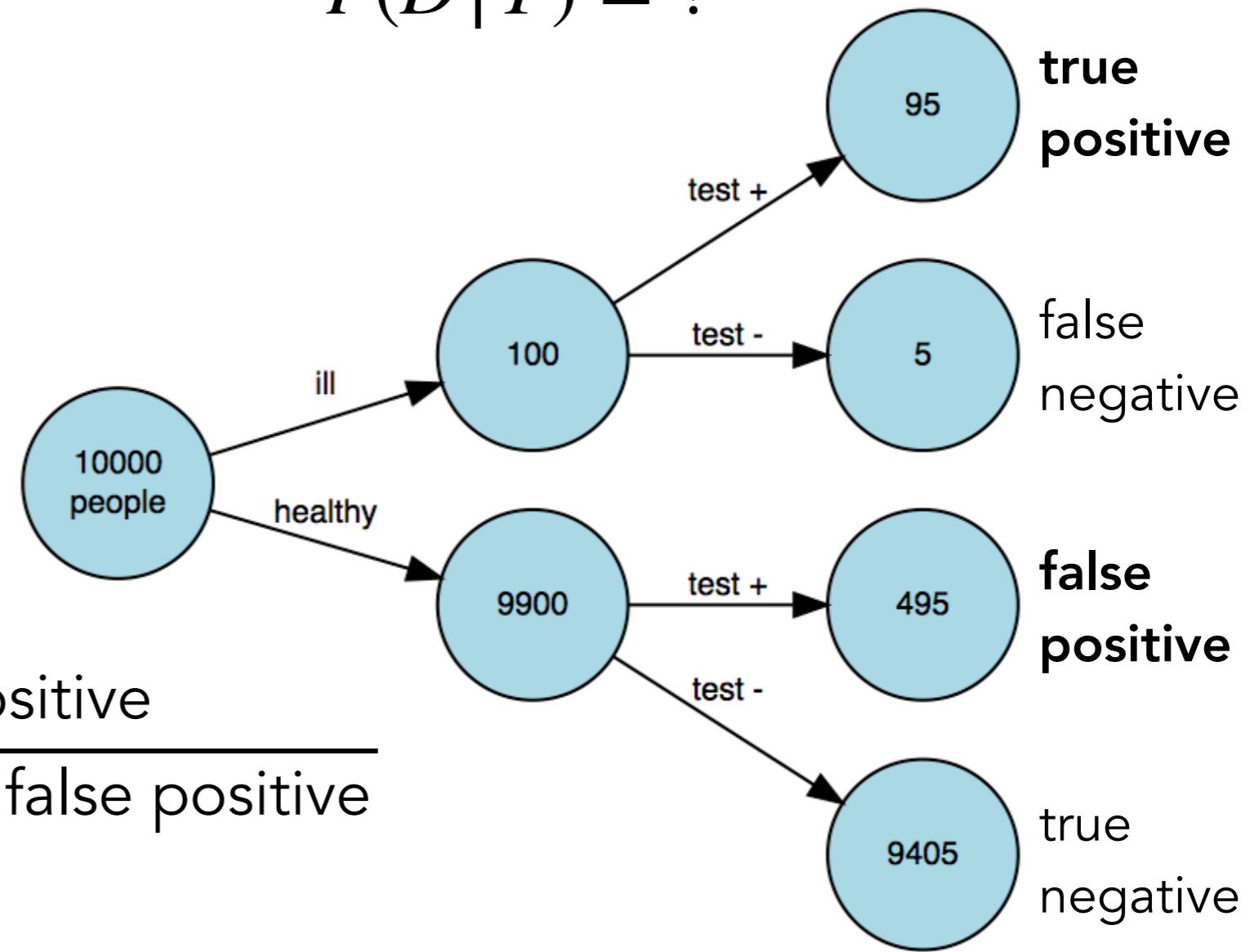
$$P(T|D) = 0.95$$

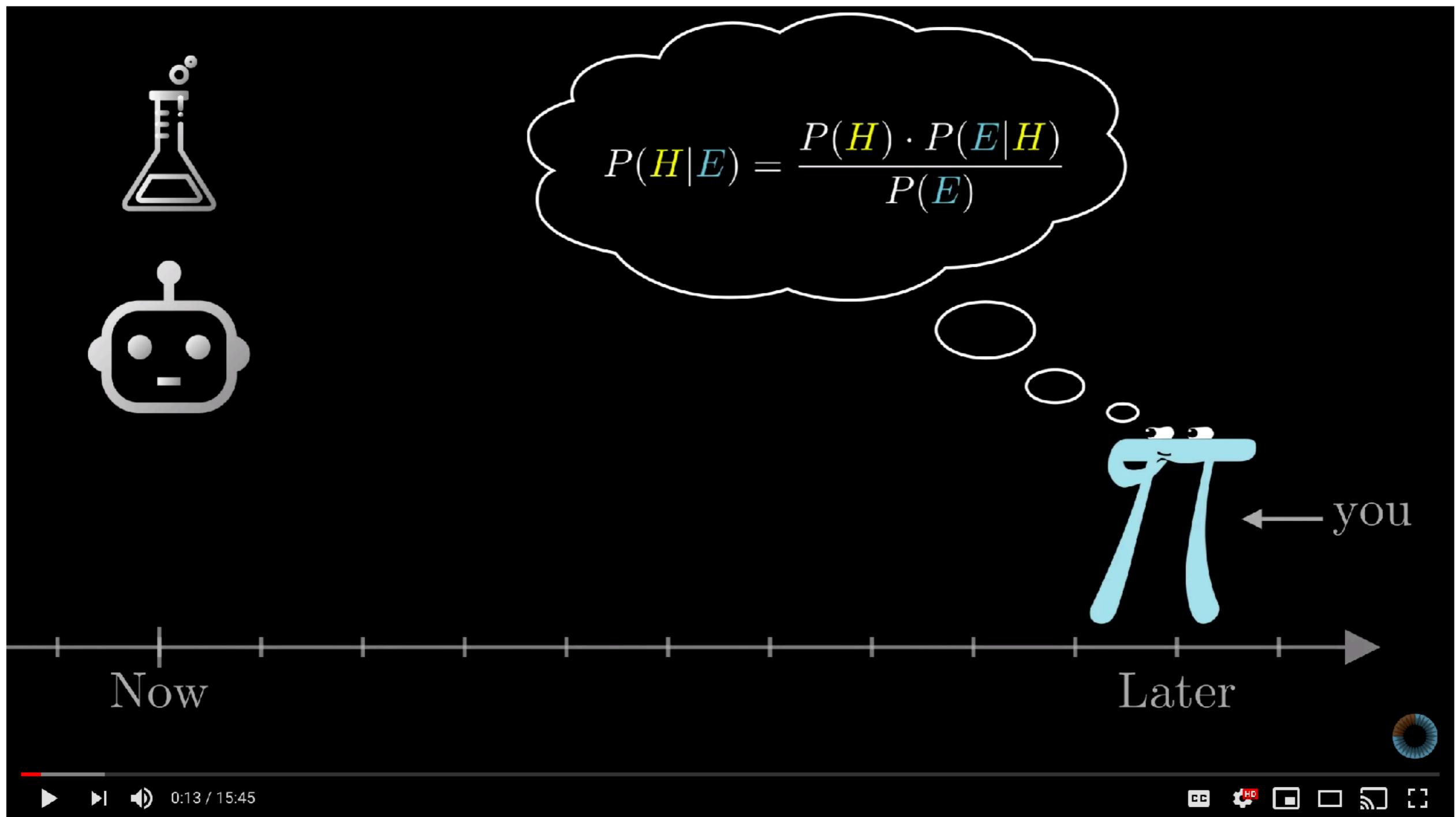
$$P(T|\neg D) = 0.05$$

$$\begin{aligned} P(D|T) &= \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \\ &= \frac{95}{95 + 495} \\ &\approx 0.16 \end{aligned}$$

what we want to know

$$P(D|T) = ?$$





Bayes theorem, and making probability intuitive

461,105 views • Dec 22, 2019

26K

228

SHARE

SAVE

...

<https://www.youtube.com/watch?v=HZGCoVF3YvM&feature=youtu.be>

Bayes' theorem in three panels

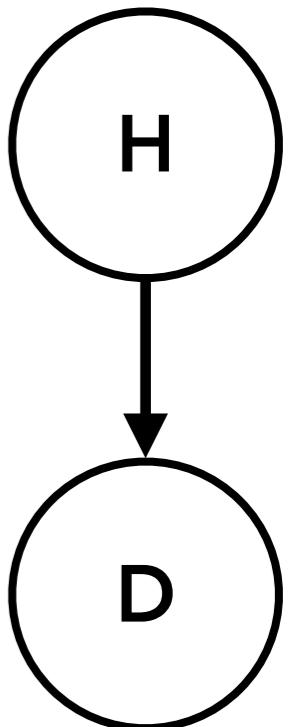
posterior

$$p(H|D) = \frac{\text{likelihood} \cdot \text{prior}}{p(D)}$$

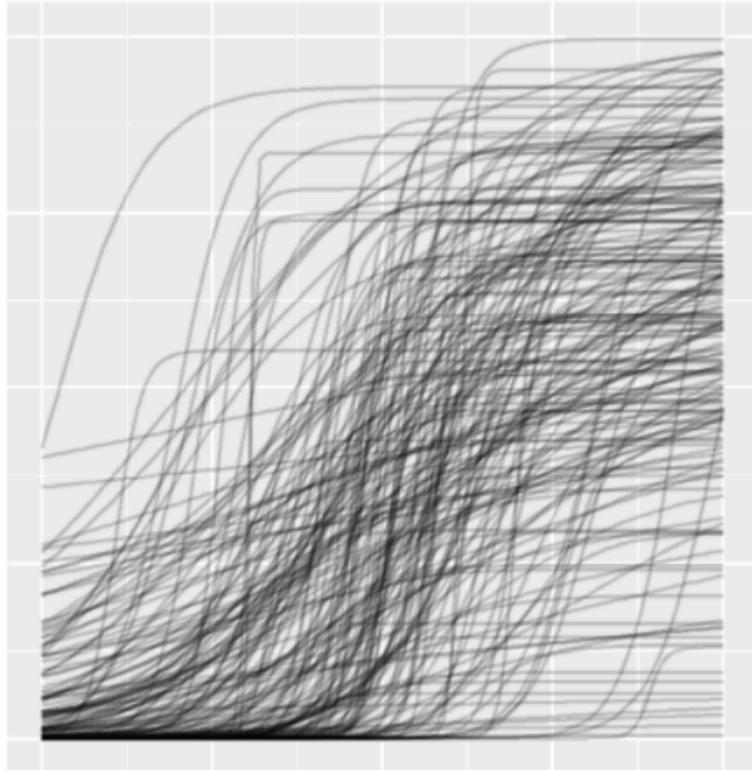
subjective probability
interpretation

H = Hypothesis

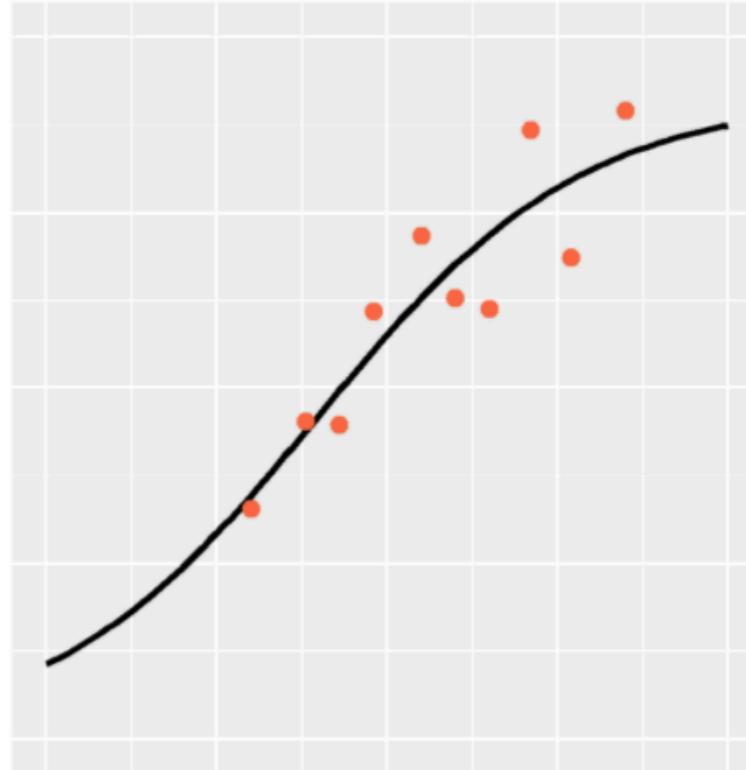
D = Data



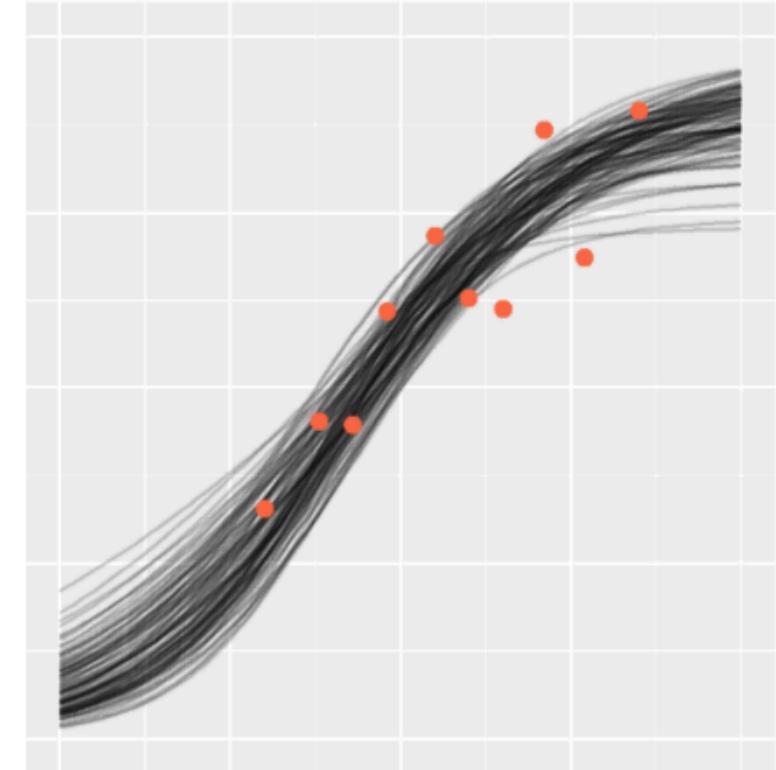
Plausible lines before seeing data



How well does the data fit the lines



Plausible lines after seeing data



Getting Bayes' right matters

Getting Bayes right matters!

Officer characteristics and racial disparities in fatal officer-involved shootings

David J. Johnson^{a,b,1}, Trevor Tress^b, Nicole Burkell^b, Carley Taylor^b, and Joseph Cesario^b

^aDepartment of Psychology, University of Maryland at College Park, College Park, MD 20742; and ^bDepartment of Psychology, Michigan State University, East Lansing, MI 48824

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved June 24, 2019 (received for review March 5, 2019)

Despite extensive attention to racial disparities in police shootings, two problems have hindered progress on this issue. First, databases of fatal officer-involved shootings (FOIS) lack details about officers, making it difficult to test whether racial disparities vary by officer characteristics. Second, there are conflicting views on which benchmark should be used to determine racial disparities when the outcome is the rate at which members from racial groups are fatally shot. We address these issues by creating a database of FOIS that includes detailed officer information. We test racial disparities using an approach that sidesteps the benchmark debate by directly predicting the race of civilians fatally shot rather than comparing the rate at which racial groups are shot to some benchmark. We report three main findings: 1) As the proportion of Black or Hispanic officers in a FOIS increases, a person shot is more likely to be Black or Hispanic than White, a disparity explained by county demographics; 2) race-specific county-level violent crime strongly predicts the race of the civilian shot; and 3) although we find no overall evidence of anti-Black or anti-Hispanic disparities in fatal shootings, when focusing on different subtypes of shootings (e.g., unarmed shootings or "suicide by cop"), data are too uncertain to draw firm conclusions. We highlight the need to enforce federal policies that record both officer and civilian information in FOIS.

officer-involved shootings | racial disparity | racial bias | police use of force | benchmarks



Claim:

"White officers are not more likely to shoot minority civilians than non-White officers"

$$\Pr(\text{shot}|\text{minority civilian, white officer}, X)$$

$$- \Pr(\text{shot}|\text{minority civilian, minority officer}, X) \leq 0,$$

[1]

What the statistic says:

"whether a person fatally shot was more likely to be Black (or Hispanic) than White"

Original claim:

Requires Bayes' rule

$$\begin{aligned} & \Pr(\text{shot}|\text{minority civilian, white officer}, X) \\ & - \Pr(\text{shot}|\text{minority civilian, minority officer}, X) \\ & \quad \Pr(\text{min. civ. } |\text{shot, white off.}, X) \\ & \quad \times \Pr(\text{shot}|\text{white off.}, X) \\ & = \frac{\Pr(\text{min. civ. } |\text{shot, white off.}, X) \times \Pr(\text{shot}|\text{white off.}, X)}{\Pr(\text{minority civilian}|\text{white officer}, X)} \\ & \quad - \frac{\Pr(\text{min. civ. } |\text{shot, min. off.}, X) \times \Pr(\text{shot}|\text{min. off.}, X)}{\Pr(\text{minority civilian}|\text{minority officer}, X)}. \end{aligned}$$

[2]

authors didn't have the relevant data to support their claim!

paper was retracted

Johnson, D. J., Tress, T., Burkell, N., Taylor, C., & Cesario, J. (2019). Officer characteristics and racial disparities in fatal officer-involved shootings. *Proceedings of the National Academy of Sciences*, 116(32), 15877–15882.

Knox, D., & Mummolo, J. (2020). Making inferences about racial disparities in police violence. *Proceedings of the National Academy of Sciences*, 117(3), 1261–1262.

Getting Bayes right matters!

The screenshot shows the Zotero application interface. On the left is a library list with various items, including a highlighted entry: "Officer characteristics and racial disparities in f..." by Johnson et al., which has been retracted. On the right, a detailed view of this item is shown. A red banner at the top states "This work has been retracted." Below it, the retraction date is listed as "Retracted on 7/10/2020". Under "Concerns/Issues About Results", it says "Any question, controversy or dispute over the validity of the results". Under "Upgrade/Update of Prior Notice", it says "Either a change to or affirmation of a prior notice". A link to "Retraction Notice" is provided. At the bottom, there is a "More details:" section with a link to "https://retractionwatch.com/2020/07/08/retraction-of-paper-on-police-killings-and-race-not-due-to-mob-pressure-or-distaste-for-the-political-views-of-people-citing-the-work-approvingly-say-authors/". A note at the bottom right indicates "Data from Retraction Watch".

Title	Creator	Date Added
The R Inferno	Burns	... 9/29/2020, 1:4...
Officer characteristics and racial disparities in f...	Johnson et al.	... 6/15/2020, 4:5...
Reply to Knox and Mummolo and Schimmack and ...	Johnson and C...	... 6/15/2020, 4:5...
Young unarmed nonsuicidal male victims of fatal ...	Schimmack an...	... 6/15/2020, 4:5...
Making inferences about racial disparities in polic...	Knox and Mum...	... 6/15/2020, 4:5...
Quasi-experimental causality in neuroscience and ...	Marinescu et al.	... 5/29/2020, 10:...
Equivalence Testing for Psychological Research: A ...	Isager et al.	... 5/29/2020, 10:...
TreeBUGS: An R package for hierarchical multinom...	Heck et al.	... 5/27/2020, 1:4...
Analysis of variance with unbalanced data: an upd...	Hector et al.	... 4/15/2020, 1:4...
Bayesian analysis of factorial designs.	Rouder et al.	... 4/2/2020, 12:5...
Model comparison in ANOVA	Rouder et al.	... 4/2/2020, 12:4...
Bayesian Data Analysis	Gelman et al.	... 4/2/2020, 12:2...
d-separation		... 3/22/2020, 4:0...
To Bayes or Not To Bayes? That's no longer the qu...	Fokoue	... 3/22/2020, 4:0...
Science is not a signal detection problem	Wilson et al.	... 3/22/2020, 4:0...
Credible Confidence: A Pragmatic View on the Fre...	Albers et al.	... 3/22/2020, 4:0...
Effective Analysis of Reaction Time Data	Whelan	... 3/16/2020, 11:...
Are Reaction Time Transformations Really Benefici...	Schramm and 3/16/2020, 11:...
To transform or not to transform: using generaliz...	Lo and Andrews	... 3/16/2020, 11:...

Tip: Use Zotero as a reference manager!

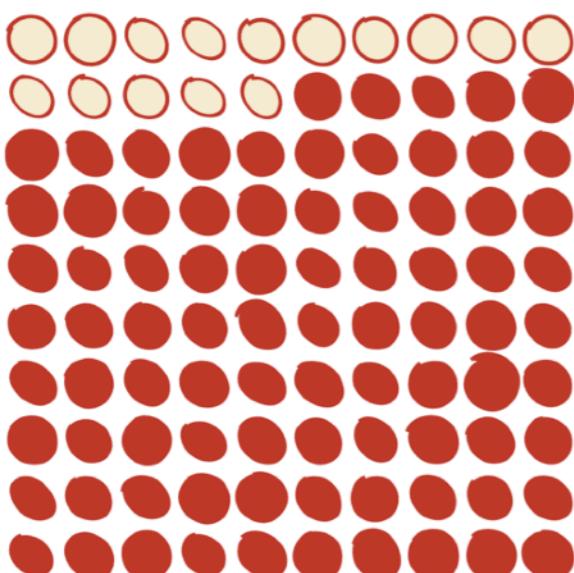
Getting Bayes right matters!

The New York Times

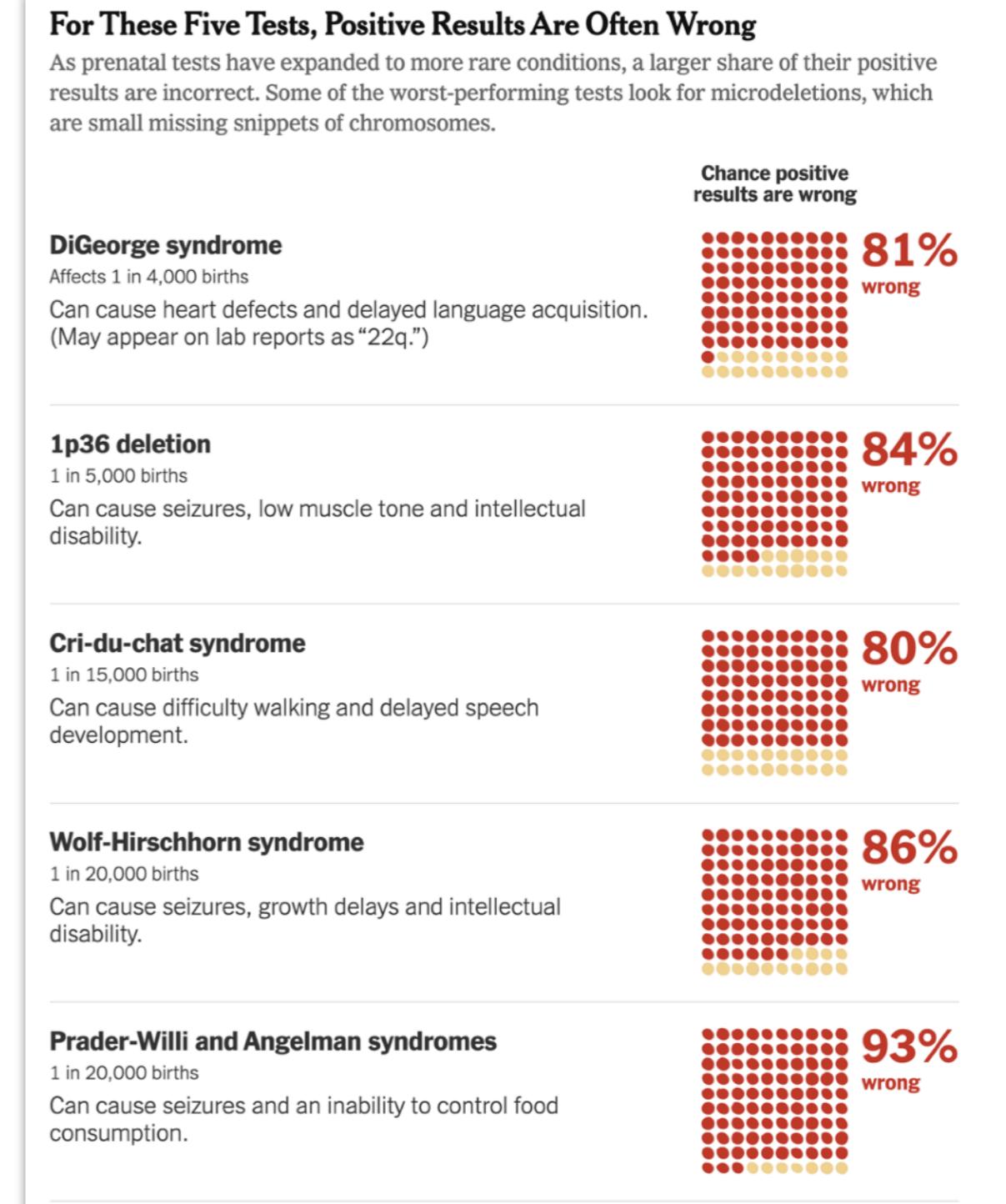
• The Upshot

When They Warn of Rare Disorders, These Prenatal Tests Are Usually Wrong

Some of the tests look for missing snippets of chromosomes. For every 15 times they correctly find a problem ● ...



... they are ● wrong 85 times



Getting Bayes right matters!

sensitivity: $p(T|D) = 0.999$

T = positive test result

specificity: $p(\neg T|\neg D) = 0.999$

$\neg T$ = negative test result

prior: $p(D) = 0.0001$

D = disease

$\neg D$ = no disease

data: T (positive test result)

81% wrong

$$\text{posterior: } p(D|T) = \frac{p(T|D) \cdot p(D)}{p(T|D) \cdot p(D) + p(T|\neg D) \cdot p(\neg D)}$$

$$= \frac{0.999 \cdot 0.0001}{0.999 \cdot 0.0001 + 0.001 \cdot 0.9999} \approx 0.09$$

Getting Bayes right matters!

Most people who are in the hospital
being treated for Covid are vaccinated.

Getting Bayes right matters!

likelihood: $p(H|V) = 0.2$

H = hospitalized

$p(H|\neg V) = 0.5$

$\neg H$ = not hospitalized

prior: $p(V) = 0.8$

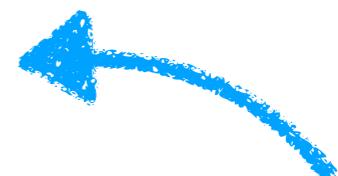
V = vaccinated

$\neg V$ = no vaccinated

data: H (the person is in the hospital)

$$\text{posterior: } p(V|H) = \frac{p(H|V) \cdot p(V)}{p(H|V) \cdot p(V) + p(H|\neg V) \cdot p(\neg V)}$$

$$= \frac{0.2 \cdot 0.8}{0.2 \cdot 0.8 + 0.5 \cdot 0.2} \approx 0.62$$

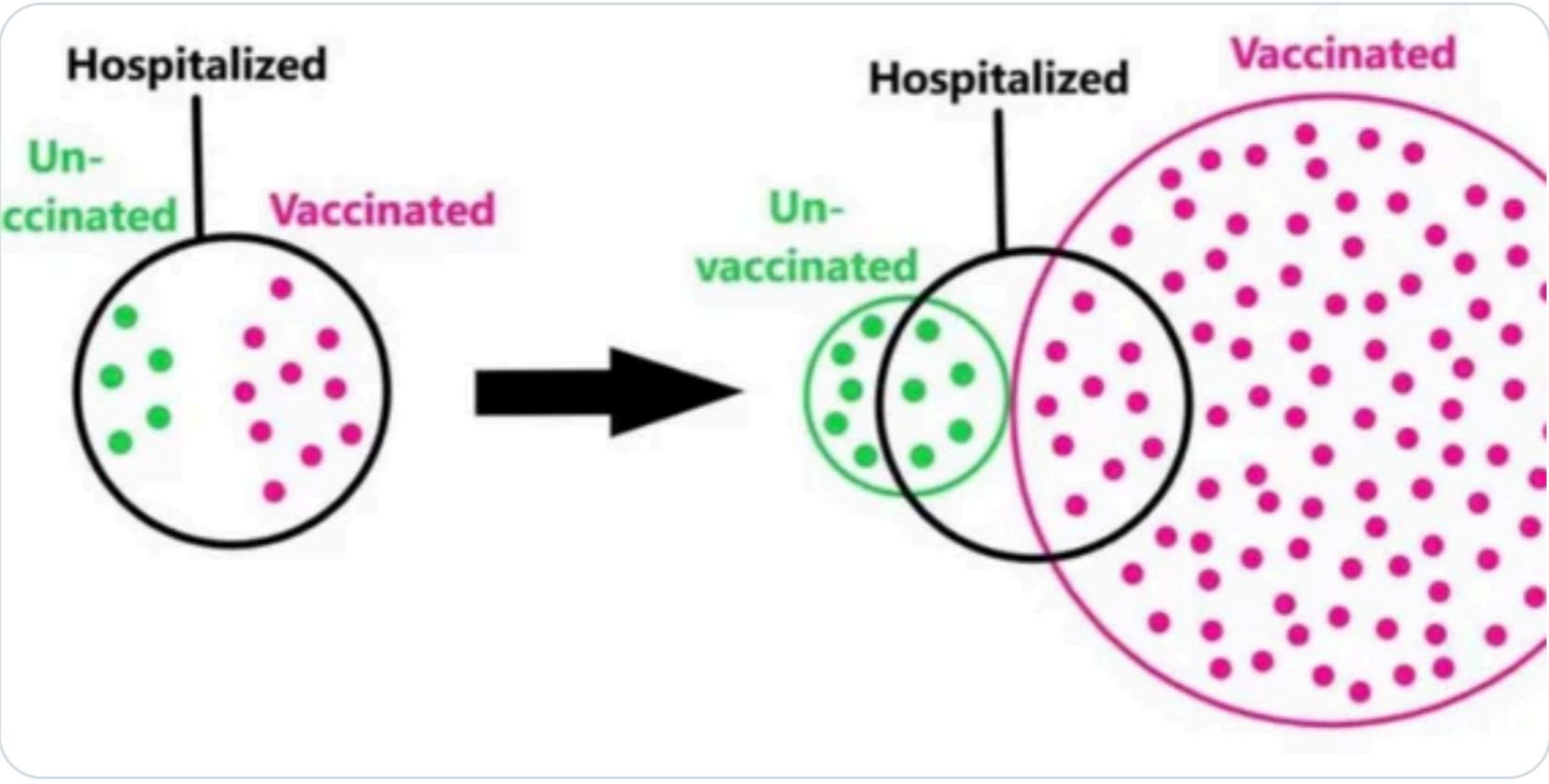


62% of the hospitalized
people are vaccinated

Bayes' rule matters

 **Nick Brown**
@sTeamTraen ...

Stolen from Reddit. May be of some use.



1:22 PM · Nov 20, 2021 · Twitter Web App

565 Retweets **54** Quote Tweets **2,752** Likes

Building a Bayesis



Rolling the dice



Four sided



Six sided

both dice are equally likely to be picked
 $p(\text{4-sided}) = p(\text{6-sided}) = 0.5$

both dice are equal sided
(uniform probability over the different numbers)

Which die do you think was rolled?

$$4 \quad p(\text{4-sided} \mid \text{data}) = ?$$

$$4, 2, 1 \quad p(\text{4-sided} \mid \text{data}) = 0.77$$

$$4, 2, 1, 3, 1 \quad p(\text{4-sided} \mid \text{data}) = 0.88$$

$$4, 2, 1, 3, 1, 5 \quad p(\text{4-sided} \mid \text{data}) = 0$$

Physical reasoning



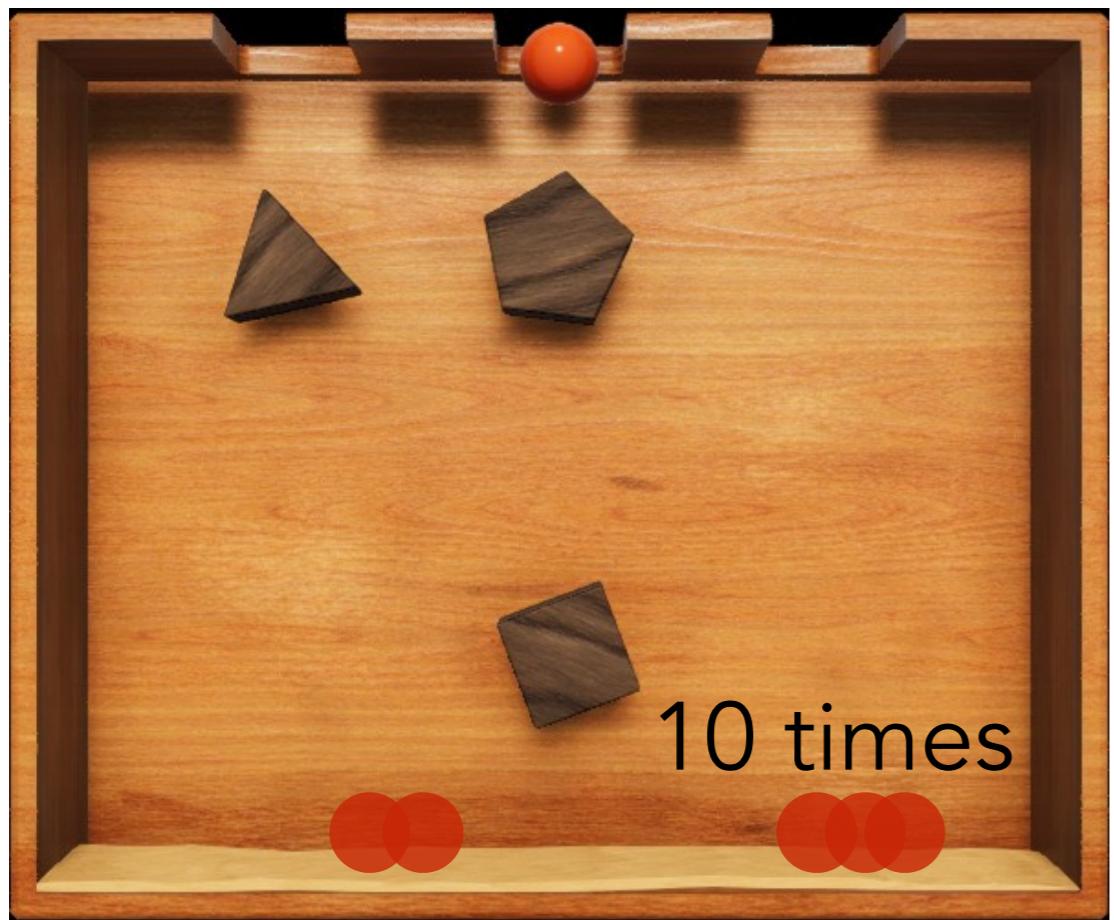
Gerstenberg, T., Siegel, M. H., & Tenenbaum, J. B. (2021). What happened? Reconstructing the past from vision and sound. PsyArXiv. <https://psyarxiv.com/tfjdk>

Physical reasoning



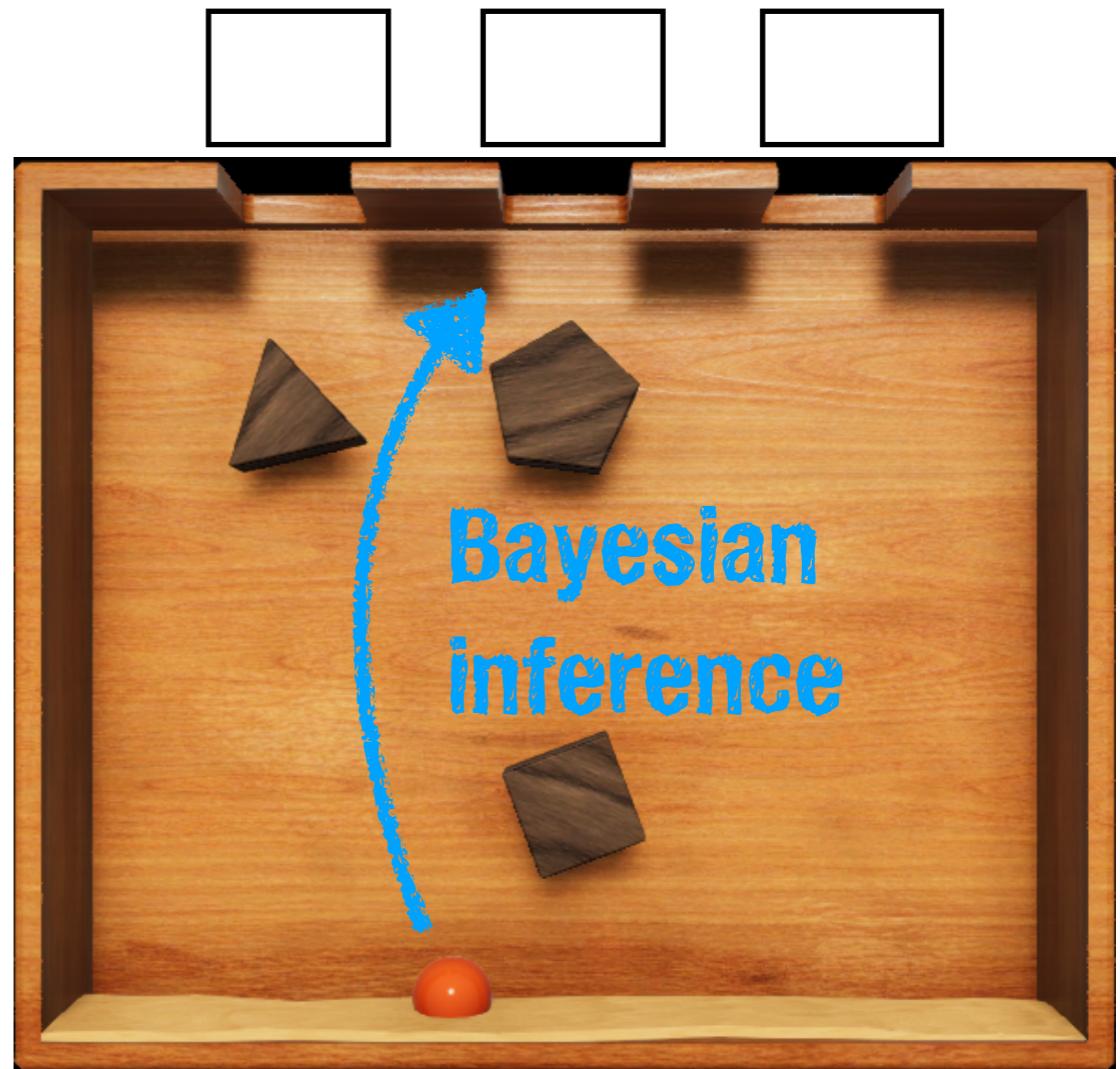
Gerstenberg, T., Siegel, M. H., & Tenenbaum, J. B. (2021). What happened? Reconstructing the past from vision and sound. PsyArXiv. <https://psyarxiv.com/tfjdk>

Prediction



Where will the ball land?

Inference



In which hole was the ball dropped?

Outline

- Introduction to probability / Recap
 - Motivation
 - Counting possibilities
 - **Clue** guide to probability
 - Understanding Bayes' Rule
 - Getting Bayes' right matters!
 - Building a Bayesis

I want more!

Chapter 9 Introduction to probability

[God] has afforded us only the twilight ... of Probability.
– John Locke

Up to this point in the book, we've discussed some of the key ideas in experimental design, and we've talked a little about how you can summarise a data set. To a lot of people, this is all there is to statistics: it's about calculating averages, collecting all the numbers, drawing pictures, and putting them all in a report somewhere. Kind of like stamp collecting, but with numbers. However, statistics covers much more than that. In fact, descriptive statistics is one of the smallest parts of statistics, and one of the least powerful. The bigger and more useful part of statistics is that it provides that let you make inferences about data.

Once you start thinking about statistics in these terms – that statistics is there to help us draw inferences from data – you start seeing examples of it everywhere. For instance, here's a tiny extract from a newspaper article in the Sydney Morning Herald (30 Oct 2010):

"I have a tough job," the Premier said in response to a poll which found her government is now the most unpopular Labor administration in polling history, with a primary vote of just 23 per cent.

This kind of remark is entirely unremarkable in the papers or in everyday life, but let's have a think about what it entails. A polling company has conducted a survey, usually a pretty big one because they can afford it. I'm too lazy to track down the original survey, so let's just imagine that they called 1000 NSW voters at random, and 230 (23%) of those claimed that they intended to vote for the ALP. For the 2010 Federal election, the Australian Electoral Commission reported 4,610,795 enrolled voters in NSW; so the opinions of the remaining 4,609,795 voters (about 99.98% of voters) remain unknown to us. Even assuming that no-one lied to the polling

in the figures/
folder for these
materials on
canvas

Probability Cheatsheet v2.0

Compiled by William Chen (<http://wchen.ca/>) and Joe Blitzstein, with contributions from Sebastian Chin, Yuan Jiang, Yiqi Hou, and Jessie Hwang. Material based on Joe Blitzstein's (<https://stat110.net/>) lectures (<http://www-stat.wharton.upenn.edu/~khs/stat110/>). Licensed under CC BY-NC-SA 4.0. Please share comments, suggestions, and errors at http://github.com/wchen/probability_cheatsheet.

Last Updated September 4, 2016

Counting

Multiplication Rule

Let's say we have a compound experiment (an experiment with multiple components). If the 1st component has n_1 possible outcomes, the 2nd component has n_2 possible outcomes, ..., and the r th component has n_r possible outcomes, then overall there are $n_1 \times n_2 \times \dots \times n_r$ possibilities for the whole experiment.

Sampling Table

The sampling table gives the number of possible samples of size k out of a population of size n , under various assumptions about how the sample is collected.

	Order Matters	Not Matter
With Replacement	n^k	$\binom{n+k-1}{k}$
Without Replacement	$\frac{n!}{(n-k)!}$	$\binom{n}{k}$

Naive Definition of Probability

If all outcomes are equally likely, the probability of an event A happening is:

$$P_{\text{Naive}}(A) = \frac{\text{number of outcomes favorable to } A}{\text{number of outcomes}}$$

Simpson's Paradox

It is possible to have $P(A | B, C) < P(A | B^c, C)$ and $P(A | B, C^c) < P(A | B^c, C^c)$ yet also $P(A | B) > P(A | B^c)$.

Random Variables and their Distributions

PMF, CDF, and Independence

Probability Mass Function (PMF) Gives the probability that a discrete random variable takes on the value x . $p_X(x) = P(X = x)$

The PMF satisfies:

$$p_X(x) \geq 0 \text{ and } \sum_x p_X(x) = 1$$

INTERACTIVE COURSE

Foundations of Probability in R

Start Course **Play Intro Video** **Bookmark**

4 hours 13 Videos 54 Exercises 27,296 Participants 4,350 XP

Course Description

Probability is the study of making predictions about random phenomena. In this course, you'll learn about the concepts of random variables, distributions, and conditioning, using the example of coin flips. You'll also gain intuition for how to solve probability problems through random simulation. These principles will help you understand statistical inference and can be applied to draw conclusions from data.

1 The binomial distribution

One of the simplest and most common examples of a random phenomenon is a coin flip: an event that is either "yes" or "no" with some probability. Here you'll learn about the binomial distribution, which describes the behavior of a combination of yes/no trials and how to predict and simulate its behavior.

VIEW CHAPTER DETAILS **Continue Chapter**

This course is part of these tracks:

- Probability and Distributions with R
- Statistician with R

David Robinson
Principal Data Scientist at Heap

Feedback

How was the pace of today's class?

much a little just a little much
too too right too too
slow slow fast fast

How happy were you with today's class overall?



What did you like about today's class? What could be improved next time?

Thank you!

Have a nice Martin Luther King, Jr. day!

The time is
always right
to do what
is right.

- Martin Luther King, Jr.

