

Session 5: Probability 2

Stats 60/Psych 10
Ismael Lemhadri
Summer 2020

News

- Probability Review - Tuesday 14th, 1:30PM PDT
- Problems already available on the course website
- Try to solve them before the review!

News

- Probability Review - Tuesday 14th, 1:30PM PDT
- Problems already available on the course website
- Try to solve them before the review!

Last time

- What is a probability?
- Classical probability
- Empirical probability

This time

- Conditional probability
- Bayes' rule

Conditional probability

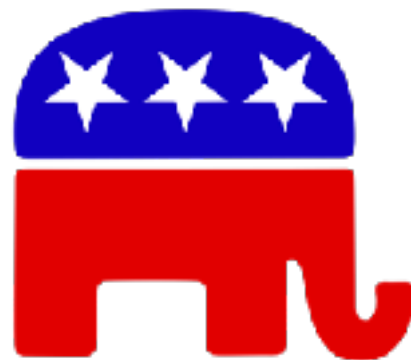
- Simple probabilities:
 - What is the likelihood that a US voter was a Republican in 2016?
 - $p(\text{Republican}) = 0.44$
 - What is the likelihood that a US voter voted for Donald Trump in the 2016 Presidential Election?
 - $P(\text{TrumpVoter}) = 0.46$

Conditional probability

- Simple probabilities:
 - What is the likelihood that a US voter was a Republican in 2016?
 - $p(\text{Republican}) = 0.44$
 - What is the likelihood that a US voter voted for Donald Trump in the 2016 Presidential Election?
 - $P(\text{TrumpVoter}) = 0.46$
- Conditional probability: Probability of one event, given that some other has occurred
 - $P(\text{TrumpVoter}|\text{Republican}) = ?$

Tree
diagram

$p(R)$



$p(DJT|R)$



$p(HRC|R)$



$p(D)$



$p(DJT|D)$



$p(HRC|D)$



Population
(registered
Democrats or
Republicans
who voted for
either DJT
or HRC)

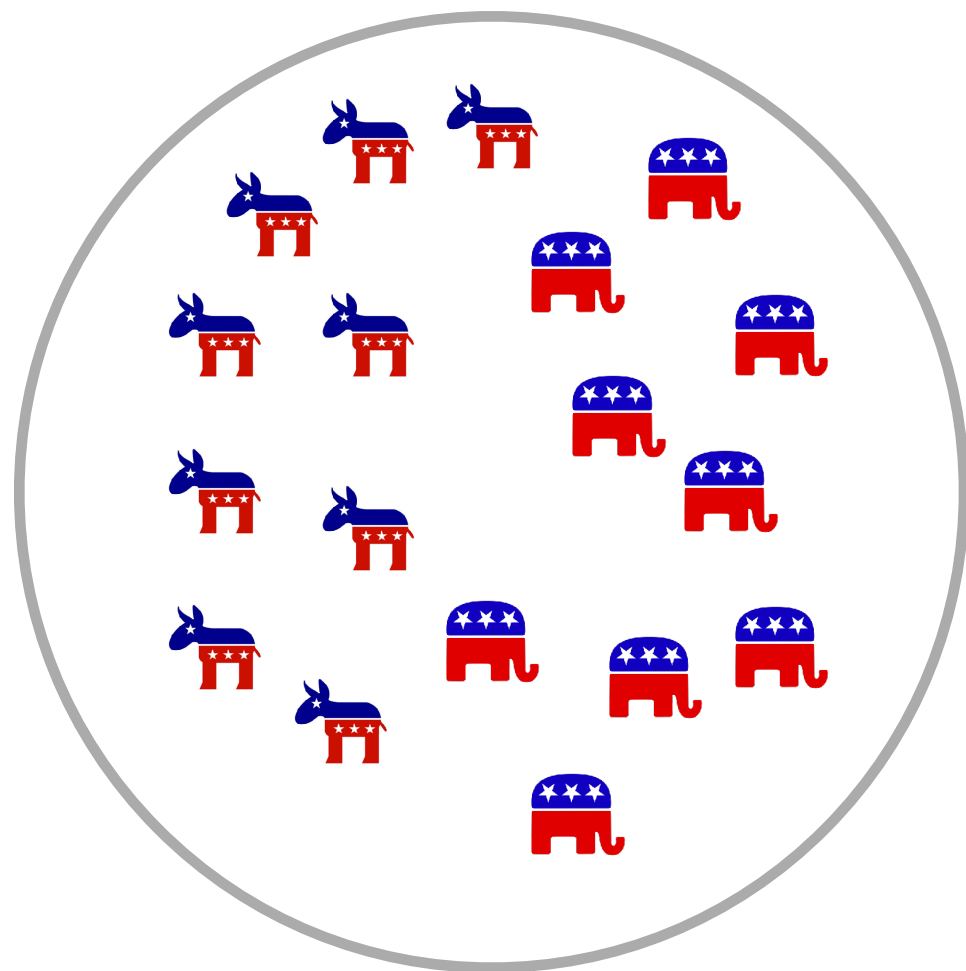
Computing conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

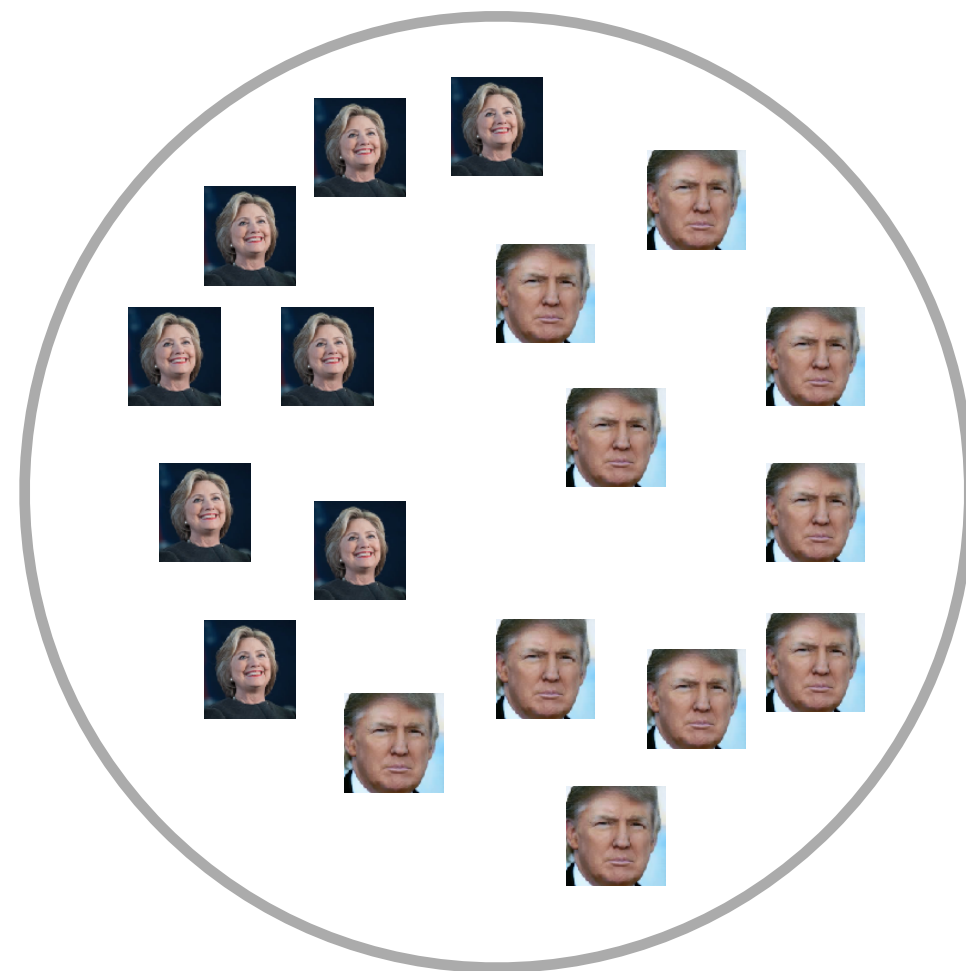
$$P(\textit{TrumpVoter}|\textit{Republican}) = \frac{P(\textit{TrumpVoter} \cap \textit{Republican})}{P(\textit{Republican})}$$

Limits the calculation to the set of B events

Another view on conditional probability

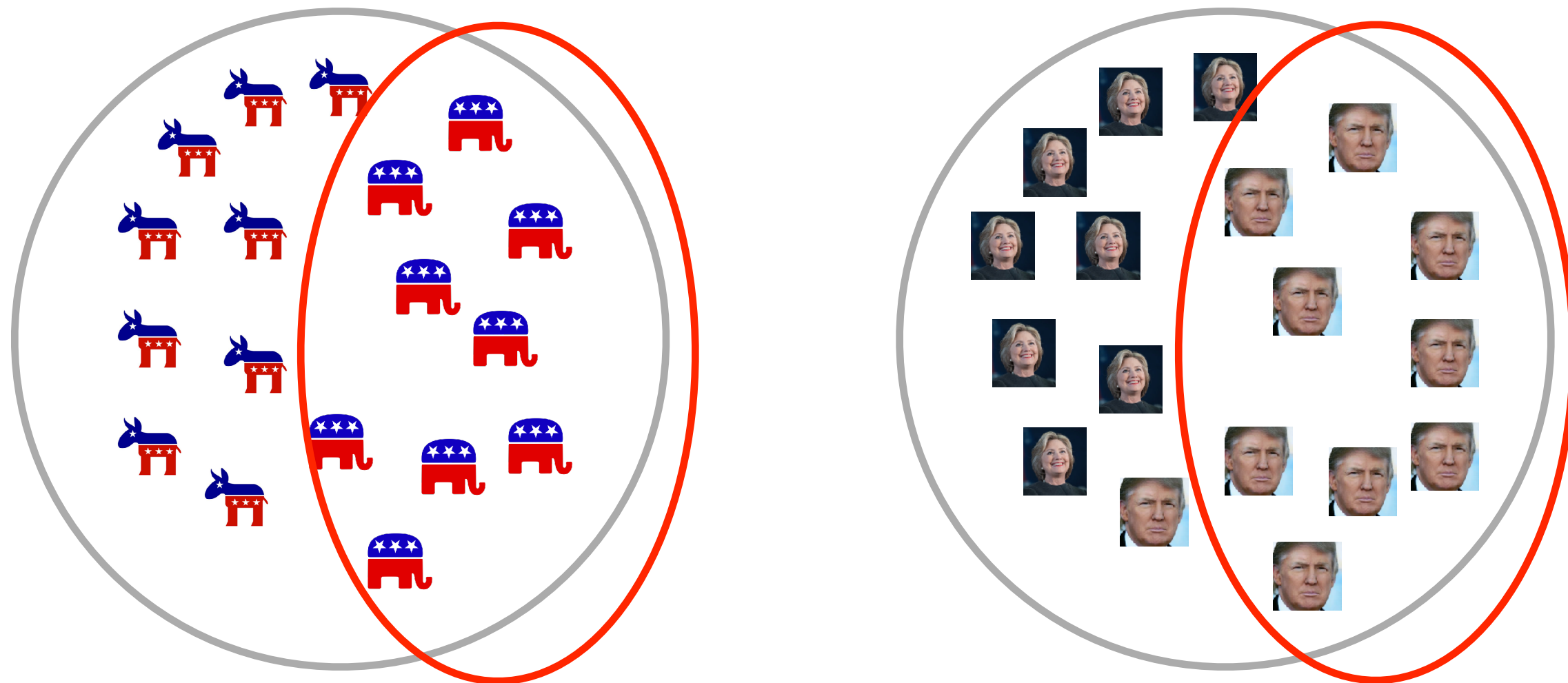


$$P(D) = 9/18 = 0.5$$
$$P(R) = 1 - P(D) = 0.5$$



$$P(DJT) = 10/18 = 0.55$$
$$P(HRC) = 1 - P(DJT) = 0.45$$

Another view on conditional probability



$$P(\text{DJT}) = 10/18 = 0.55$$

$$P(\text{DJT}|\text{R}) = ?$$

$$P(\text{DJT}|\text{R}) = 9/9 = 1.0$$

What does “independent” mean to you?

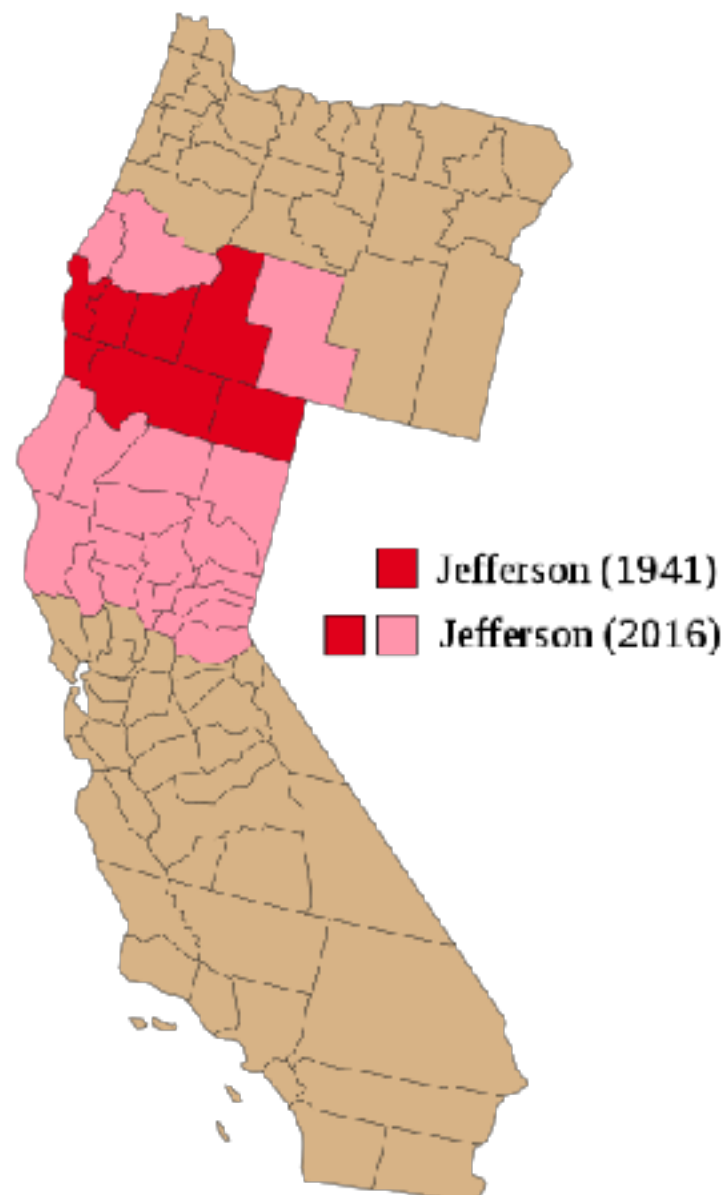
Statistical Independence

- Knowing about one thing does not tell us anything about the other

$$P(A|B) = P(A)$$

- Knowing the value of B doesn't give us any additional information about the value of A
- They are statistically unrelated
- This has a very different meaning from the common language meaning of “independence”

Example: The proposed “independent” state of Jefferson



Let's suppose they succeeded
For a current resident of CA:

$$P(\text{CA})=0.986$$

$$P(\text{JF})=0.014$$

$$P(\text{CA}|\text{JF})=0$$

political independence =
statistical dependence!

In general, mutually independent
events will be statistically dependent
(assuming $p>0$)



National Health and Nutrition Examination Survey

- NHANES is a program of studies by the CDC designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations.
- The survey examines a nationally representative sample of about 5,000 persons each year.
- The NHANES interview includes demographic, socioeconomic, dietary, and health-related questions. The examination component consists of medical, dental, and physiological measurements, as well as laboratory tests administered by highly trained medical personnel.
- Available in R:
 - `library(NHANES)`

An example: Are physical activity and mental health independent in NHANES?

PhysActive

Participant does moderate or vigorous-intensity sports, fitness or recreational activities (Yes or No).

DaysMentHlthBad

Self-reported number of days participant's mental health was not good out of the past 30 days.

```
NHANES_adult = NHANES_adult %>%  
  mutate(badMentalHealth=DaysMentHlthBad>7)
```

Are two draws from a single deck of cards (without replacing the first card) independent?

yes **A**

no **B**

An example: Are physical activity and mental health independent in NHANES?

```
NHANES_adult %>%  
  summarize(badMentalHealth=mean(badMentalHealth))
```

P(badMentalHealth)
0.164

```
NHANES_adult %>%  
  group_by(PhysActive) %>%  
  summarize(badMentalHealth=mean(badMentalHealth))
```

P(badMentalHealth ~Active)	0.200
P(badMentalHealth Active)	0.132

Physical activity is good - let's do some!

Why independence matters



https://www.ted.com/talks/peter_donnelly_shows_how_stats_fool_juries

Reversing a conditional probability

- We know $P(A|B)$
- How do we find out what $P(B|A)$ is?
 - Why would this ever be useful?

Airport screening




we know: $P(\text{positive test} \mid \text{explosives})$
we want to know: $P(\text{explosives} \mid \text{positive test})$

Medical testing

- Prostate specific antigen (PSA)
- Tests can be characterized by two factors:
 - Sensitivity:
 - $P(\text{positive test} \mid \text{disease})$
 - ~80%
 - Specificity:
 - $1 - P(\text{positive test} \mid \text{no disease})$
 - ~70%

Prostate Cancer Facts


Prostate cancer
is the most common
cancer to affect Canadian men


During his lifetime,
1 in 7
men will be diagnosed with the disease¹

EARLY DETECTION SAVES LIVES.
When detected early, the survival rate
of prostate cancer can
BE OVER 90%²

For more information on prostate cancer,
visit prostatecancer.ca

Prostate Cancer Information Service
1-855-PCC-INFO (1-855-722-4636)

 Prostate Cancer
Canada

Charitable Registration Number: BN 69127 0944 RR0001 Source: Canadian Cancer Society, 2017; American Cancer Society, 2012

© 2018 Prostate Cancer Canada

Table of possible outcomes

	Has disease	Does not have disease
Positive test	“hit” $P(D \cap T)$	“false alarm” $P(\sim D \cap T)$
Negative test	“miss” $P(D \cap \sim T)$	“true negative” $P(\sim D \cap \sim T)$

Sensitivity: $P(\text{positive test} \mid \text{has disease})$

How do we compute it?

$$\text{Sensitivity} = \text{hits} / (\text{hits} + \text{misses})$$

Table of possible outcomes

	Has disease	Does not have disease
Positive test	“hit” $P(D \cap T)$	“false alarm” $P(\sim D \cap T)$
Negative test	“miss” $P(D \cap \sim T)$	“true negative” $P(\sim D \cap \sim T)$

Specificity: $P(\text{negative test} \mid \text{no disease})$

How do we compute it?

Specificity = $\text{true negatives} / (\text{false alarms} + \text{true negatives})$

A person selected at random receives a test for a disease and the result is positive. What do we need to know in order to determine the likelihood that the person actually has the disease? (select all that apply)

The specificity of the test

The sensitivity of the test

The probability of getting the test

The probability that the person has the disease

Interpreting test results

- A person receives a positive test result
- We know the likelihood of a positive test given the disease
 - Sensitivity of the test: $P(\text{positive test} | \text{disease})$
- But what we really want to know is: is the likelihood that the person actually has the disease?
 - $P(\text{disease} | \text{positive test})$
- How do we compute this “inverse probability”?

Bayes' rule

- A way to invert a conditional probability

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

- In the context of science:

$$P(hypothesis|data) = \frac{P(data|hypothesis)P(hypothesis)}{P(data)}$$

Deriving Bayes' rule

- Remember the definition of conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- Rearrange to get the rule for computing joint probability of A and B:

$$P(A \cap B) = P(A|B)P(B)$$

- So if we want to compute $P(B|A)$:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)}$$

Bayes' rule

- For two outcomes, we can express it in a slightly clearer way using the sum rule for probabilities:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

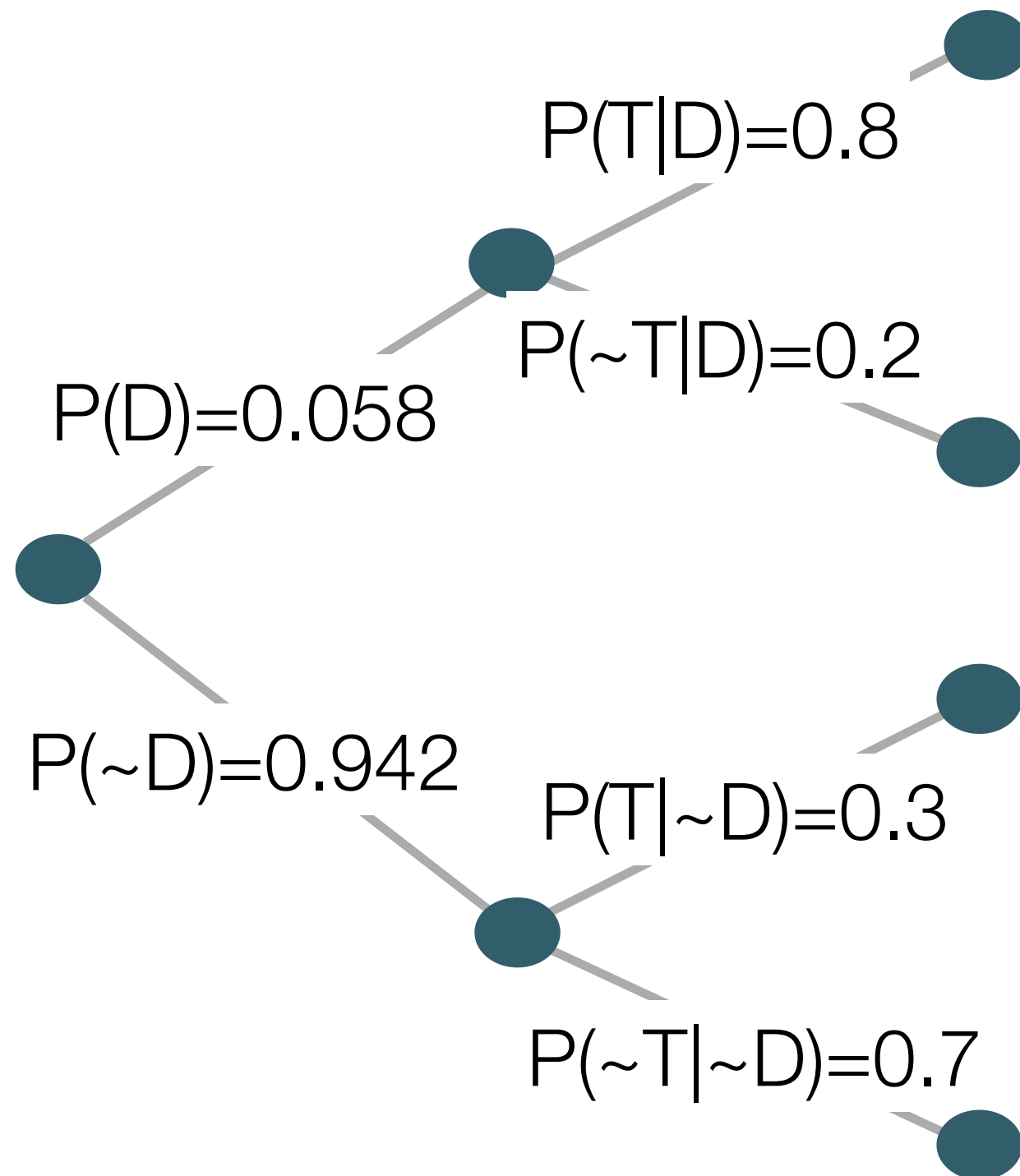
$$P(B) = P(B|A) * P(A) + P(B| \sim A) * P(\sim A)$$

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B|A) * P(A) + P(B| \sim A) * P(\sim A)}$$

60 year old male: $P(\text{disease in next 10 years})=0.058$

Sensitivity: $P(T|D)=0.8$

Specificity: $P(\sim T|\sim D)=0.7$



$$P(D|T) = \frac{0.8 \cdot 0.058}{0.8 \cdot 0.058 + 0.3 \cdot 0.942}$$
$$= 0.14$$

What do these probabilities mean?

- The person either has a disease or doesn't
- How should we interpret this probability?
- Objective probability
 - long-run relative frequency that the hypothesis is true
- Subjective probability
 - our degree of belief in the hypothesis
 - how plausible is the hypothesis?

What do these probabilities mean?

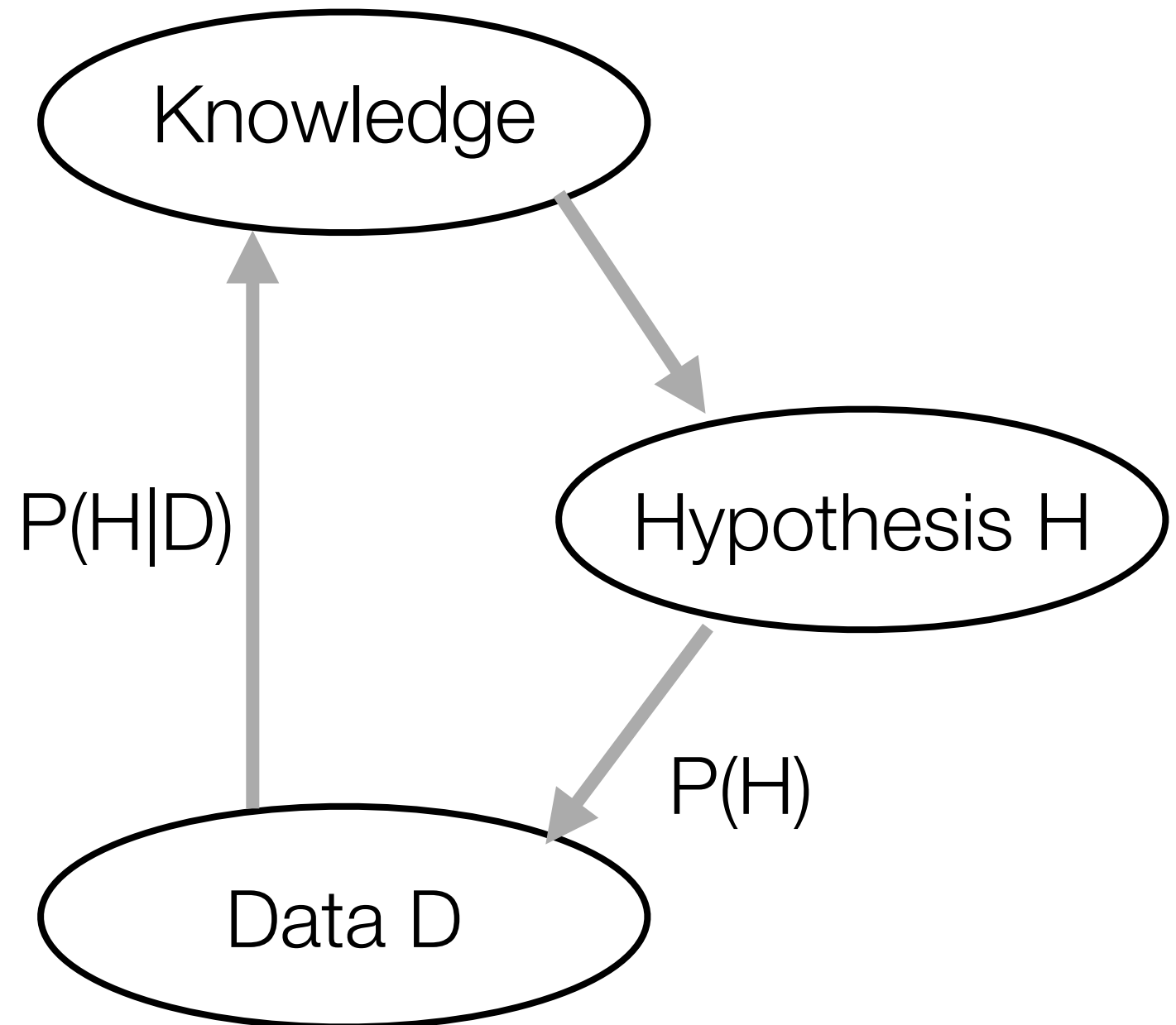
- The person either has a disease or doesn't
- How should we interpret this probability?
- Objective probability
 - long-run relative frequency that the hypothesis is true
- Subjective probability
 - our degree of belief in the hypothesis
 - how plausible is the hypothesis?



John Maynard
Keynes:

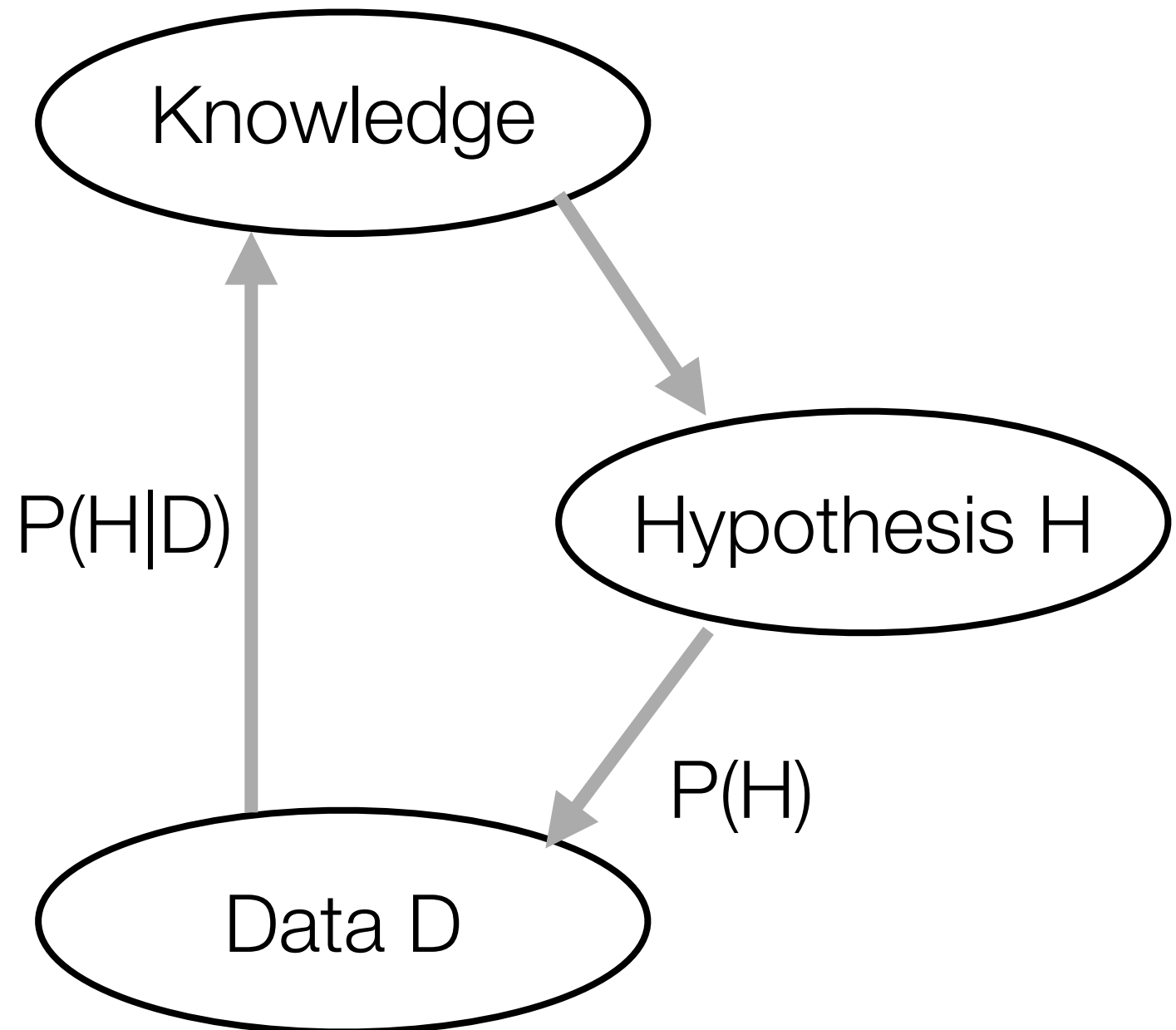
“In the long run,
we are all dead”

Statistics as learning from data



Statistics as learning from data

- We almost always start with some prior knowledge, which leads us to test a hypothesis
 - Perform the PSA test
- We generally have some idea of what to expect
 - e.g. $P(\text{disease in next 10 years})=0.058$
- We update our knowledge based on the data using Bayes' rule
 - $P(\text{disease}|\text{test result})=0.14$



Dissecting Bayes' rule

$$P(A|B) = \frac{P(B|A)}{P(B)} * P(A)$$

Dissecting Bayes' rule


prior: how likely did we think A was before we collected data?

$$P(A|B) = \frac{P(B|A)}{P(B)} * P(A)$$

Dissecting Bayes' rule

posterior: how likely do we think A is after we collected data?

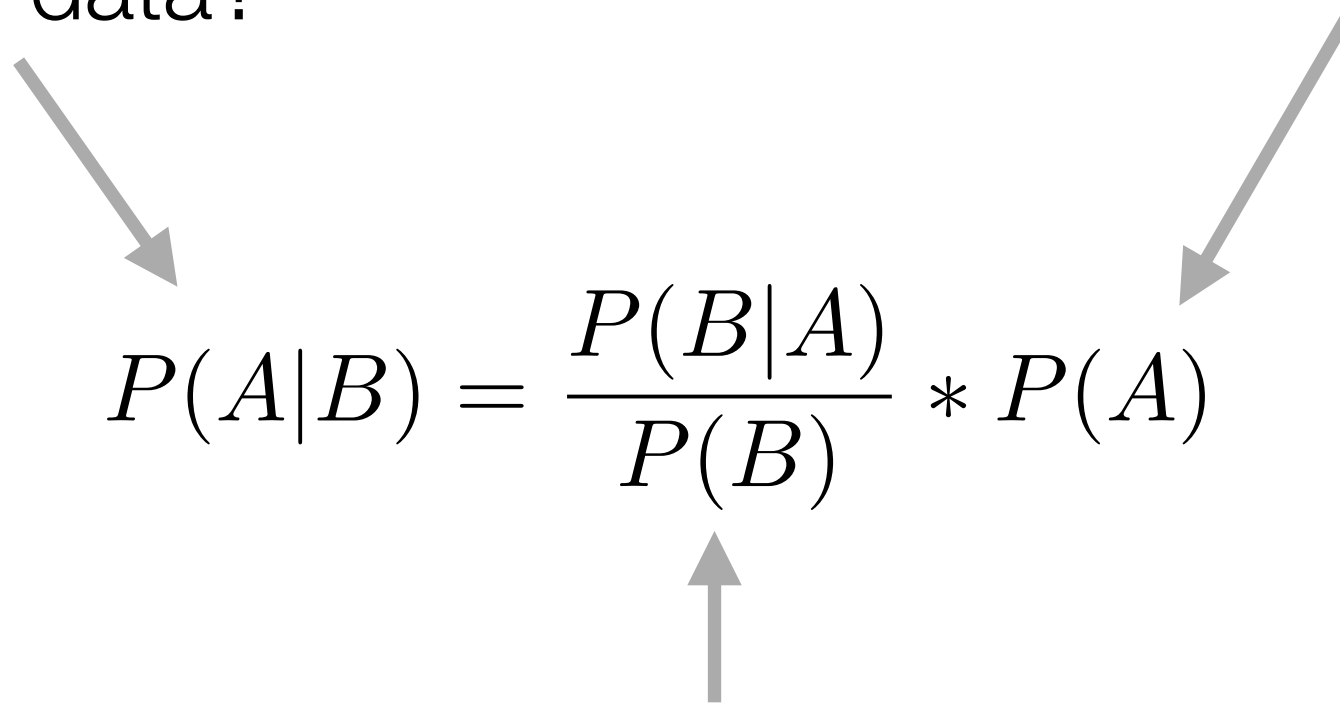
prior: how likely did we think A was before we collected data?


$$P(A|B) = \frac{P(B|A)}{P(B)} * P(A)$$

Dissecting Bayes' rule

posterior: how likely do we think A is after we collected data?

prior: how likely did we think A was before we collected data?


$$P(A|B) = \frac{P(B|A)}{P(B)} * P(A)$$

relative likelihood of the data given A,
versus the overall likelihood
of the data

Odds

- A ratio expressing the likelihood of something happening relative to not happening

$$\text{odds} = \frac{P(A)}{P(\sim A)}$$

- 1/1: “even odds”
- Example: What are the odds of rolling a six using a one-sided die?

$$\text{odds in favor} = \frac{\frac{1}{6}}{\frac{5}{6}} = \frac{1}{5}$$

$$\text{odds against} = \frac{\frac{5}{6}}{\frac{1}{6}} = \frac{5}{1}$$

Bayesian odds

$$\text{prior odds} = \frac{P(A)}{P(\sim A)} \qquad \text{prior odds} = \frac{0.058}{1 - 0.058} = 0.061$$

$$\text{posterior odds} = \frac{P(A|B)}{P(\sim A|B)} \qquad \text{posterior odds} = \frac{0.14}{0.86} = 0.16$$

$$\text{likelihood ratio} = \frac{\text{posterior odds}}{\text{prior odds}} = 2.62$$

Summary

- Conditional probabilities allow to express the likelihood of some event, given some other event
- The statistical concept of independence revolves around whether one variable provides information about the value of another
- Bayes' theorem provides us with the means to invert conditional probabilities