LESSON 9

More Bayes and Simulating Probabilities

CSCI 3022

Content Credit: Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain

Warm-Up: Serendipity

- There population of CU undergraduates is n=31,000 students
- Suppose you are friends with r=100 people.
- You walk into a classroom and you see k=160 random people.
- Assume each group of k CU undergrads is equally likely to be in the room.

What is the probability that you see at least one friend in the room?

A). 10%   B) 20%   C) 40%   D) 60%   E) None of these

Course Logistics: Your Fourth Week At A Glance

Mon 2/5	Tues 2/6	Wed 2/7	Thurs 2/8	Fri 2/9	Sat 2/10	
	(Optional): Attend Notebook Discussion with our TA (5-6pm Zoom)	Attend & Participate in Class	HW 4 Due: 11:59pm via Gradescope	Quiz 3: Scope - Lessons: L1-L5 (including HW 3, nb3) Attend & Participate in Class		
			Graded HW 3 posted	HW 5 Released		

HW 4 Tip:

Grader checks for probability questions (e.g. 3a and 4a) require your answer to be in the form   integer/integer   (so you will need to simplify first before putting your answer into the grader check).

Solutions to Extra Practice Problems in Lesson 7 added to Appendix of Lesson 7
https://docs.google.com/presentation/d/136JGtCOvI1BOLnlGAZm6fylpwJXz21UKAFE2kTFk_aE/edit#slide=id.g267840ea360_0_8

Roadmap

Lesson 9, CSCI 3022

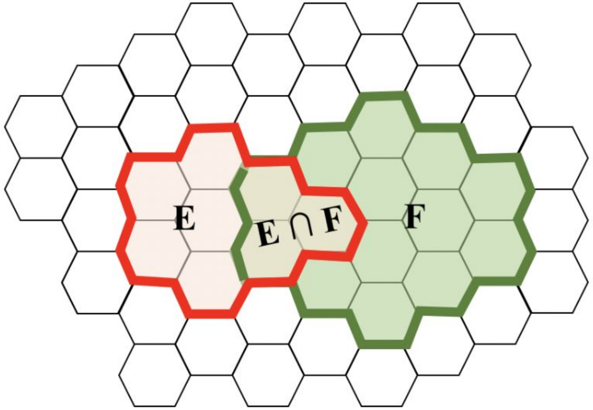
- Bayes’ Theorem
- Simulating Probabilities

Bayes’ Theorem

Lesson 9, CSCI 3022

- Bayes’ Theorem
- Simulating Probabilities

Recall: Law of Total Probability



$$P(E) = P(E \text{ and } F) + P(E \text{ and } F^C)$$

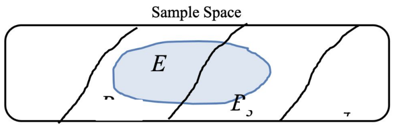
Law of Total Probability

Thm Let F be an event where $P(F) > 0$. For any event E,

$$P(E) = P(E|F)P(F) + P(E|F^C)P(F^C)$$

General Law of Total Probability

Thm For mutually exclusive events $F_1, F_2, \dots, F_n$ such that $F_1 \cup F_2 \cup \cdots \cup F_n = S$,



$$P(E) = \sum_{i=1}^{n} P(E|F_i)P(F_i)$$

Recall from last time: Example

Suppose there's a rare disease with prevalence of 1/1000 in the population.
A test for the disease has a false positive rate of 5% and a false negative rate of 1%.

a). What's the probability you test positive?

b). You test positive. What's the probability you have the disease?

Practice

$P(E) = P(E|F)P(F) + P(E|F^C)P(F^C)$ Law of Total Probability

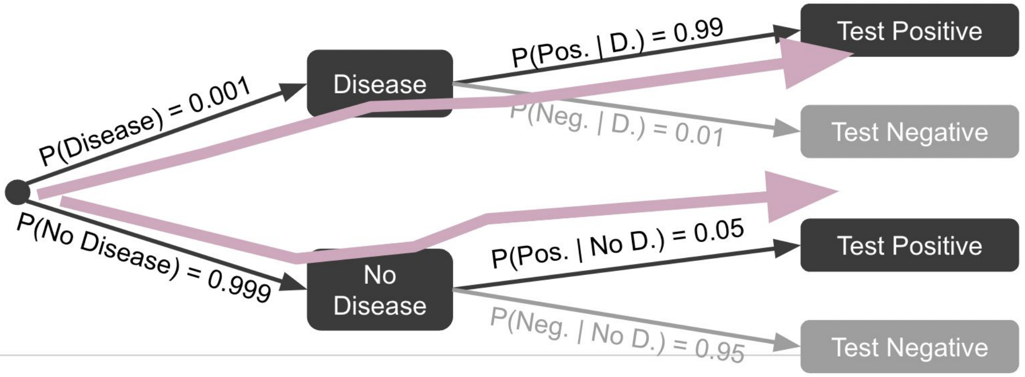
Example:
Suppose there's a rare disease with prevalence of 1/1000 in the population.
A test for the disease has a false positive rate of 5% and a false negative rate of 1%.

a). What's the probability you test positive?

P(Pos) = P(Pos, D) + P(Pos, No D)

= $\boxed{\text{P(D.) P(Pos. | D.) + P(No D.) P(Pos. | No D.)}}$

= (.001)*(0.99) + (0.999) * (0.05) = **0.05094**



Example

Suppose there's a rare disease with prevalence of 1/1000 in the population.
A test for the disease has a false positive rate of 5% and a false negative rate of 1%.

a). What's the probability you test positive?

P(Pos) = P(Pos, D) + P(Pos, No D)

$= \boxed{\text{P(D.) P(Pos. | D.) + P(No D.) P(Pos. | No D.)}}$

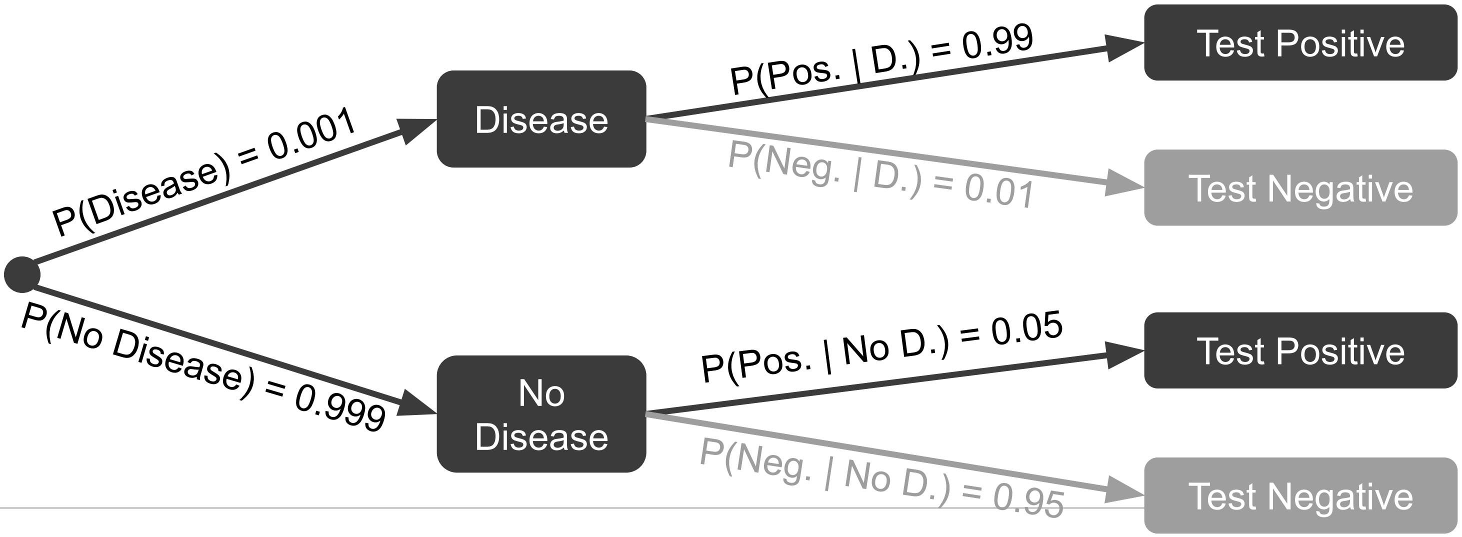
= (.001)*(0.99) + (0.999) * (0.05) = **0.05094**

b). You test positive. What's the probability you have the disease?

There’s a rare disease with prevalence of 1/1000 in the population.
A test for the disease has a false positive rate of 5% and a false negative rate of 1%.
b). You test positive. What’s the prob you have the disease?

$$P(D. \mid Pos.) = \frac{P(D., Pos.)}{P(Pos.)} = \frac{P(D.)\,P(Pos. \mid D.)}{P(Pos.)}$$

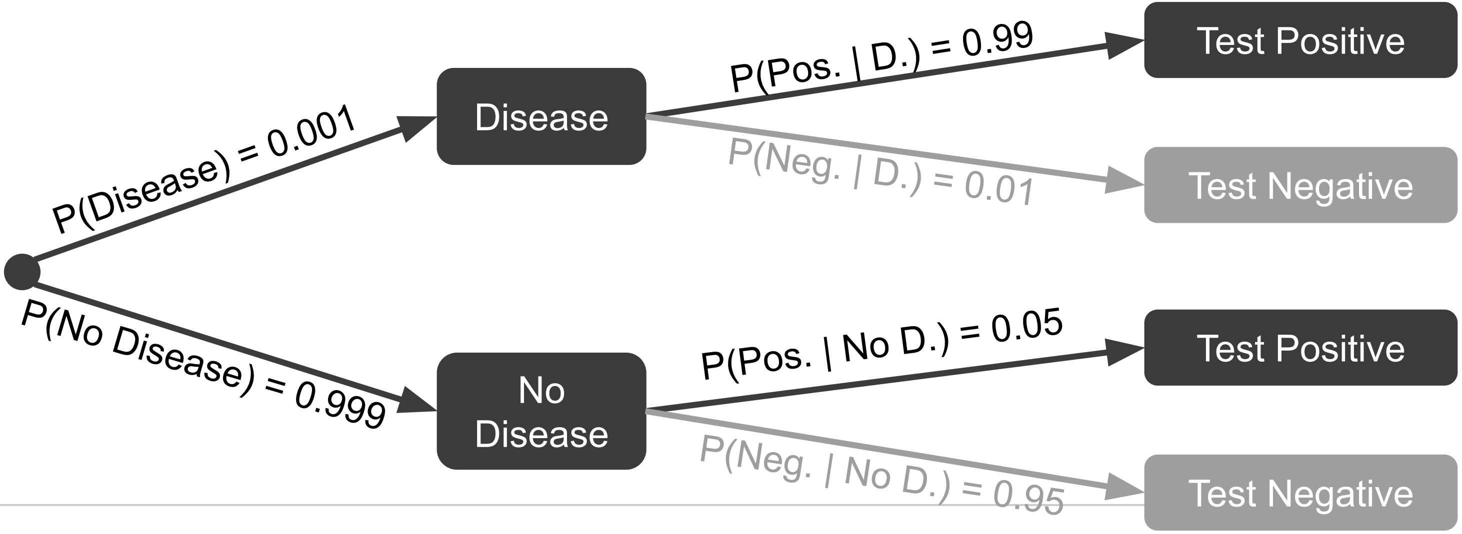
Famous Equation Alert!
Bayes’ Thm



There's a rare disease with prevalence of 1/1000 in the population.
A test for the disease has a false positive rate of 5% and a false negative rate of 1%.
b). You test positive. What's the prob you have the disease?

$$P(D., Pos.) = P(D.)\, P(Pos. \mid D.) = 0.00099$$

$$P(D. \mid Pos.) = \frac{0.00099}{0.00099 + 0.04995} = 0.0194 \approx 2\%$$



Bayes' Theorem

$P(E|F) \Rightarrow P(F|E)$

<u>Thm</u> For any events E and F where $P(E) > 0$ and $P(F) > 0$,

$$\boxed{P(F|E) = \frac{P(E|F)P(F)}{P(E)}}$$

<u>Proof</u>

2 steps!

<u>Expanded form</u>:

$$\boxed{P(F|E) = \frac{P(E|F)P(F)}{P(E|F)P(F) + P(E|F^C)P(F^C)}}$$

<u>Proof</u>

1 more step!

Famous Equation: Bayes' Theorem

Bayes' Thm allows us to **update probabilities** by incorporating observations:

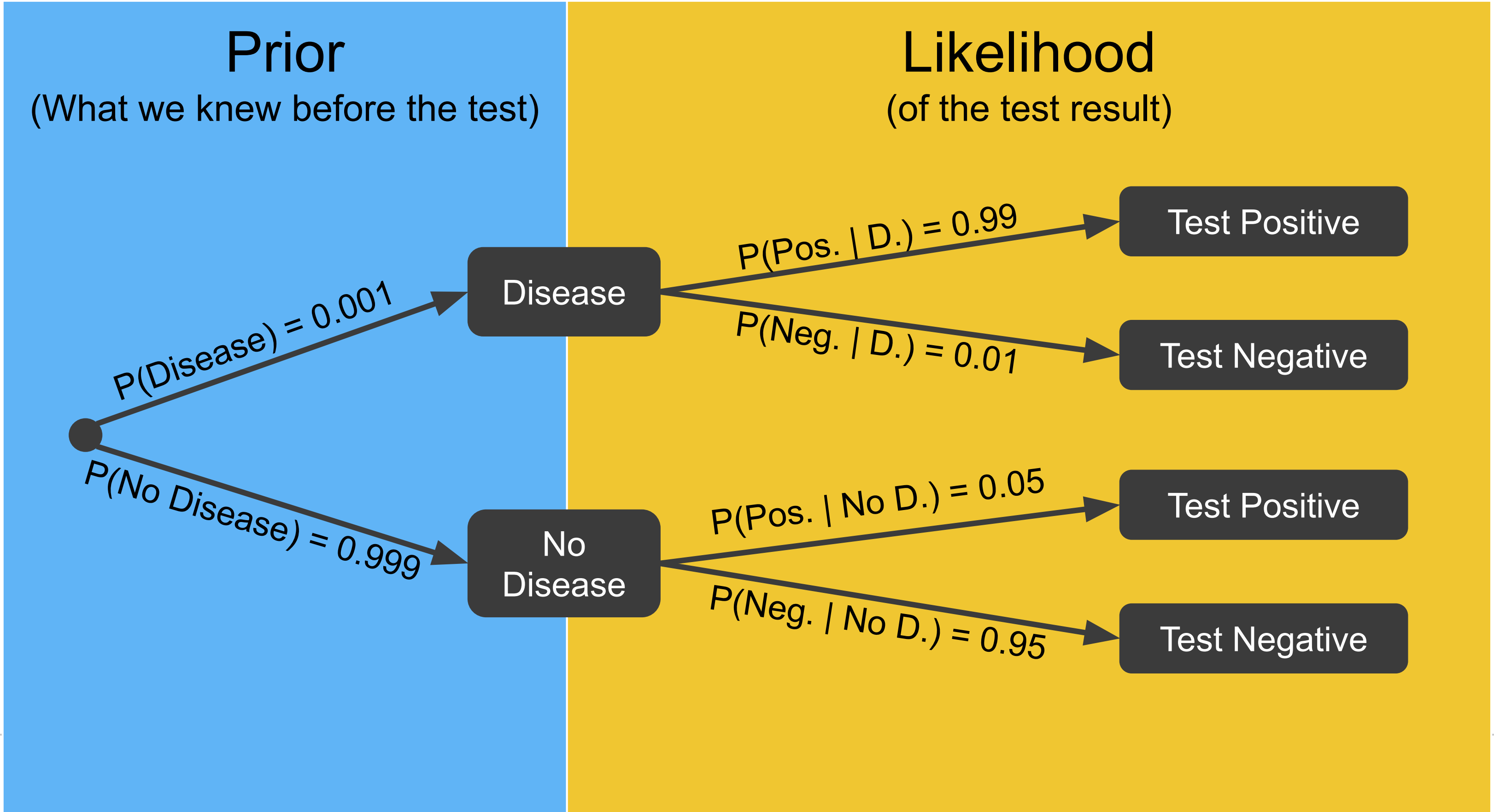
Posterior

Prior

Likelihood (Observation)

$$P(\text{Disease} \mid \text{Pos. Test}) = \frac{P(\text{Disease})\, P(\text{Pos. Test} \mid \text{Disease})}{P(\text{Pos. Test})}$$

Tree Diagrams and Terminology



Bayes' Theorem

Review

likelihood prior

posterior

$$P(F|E) = \frac{P(E|F)P(F)}{P(E)}$$

normalization constant

Mathematically:

$$P(E|F) \to P(F|E)$$

Real-life application:

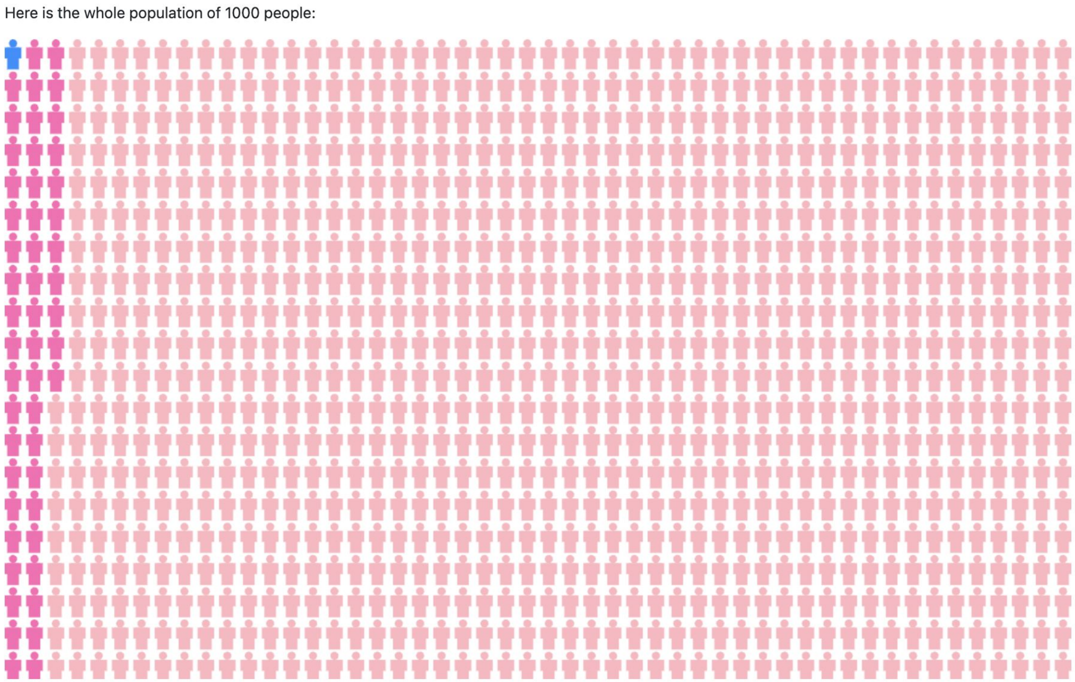
Given new evidence E, update belief of fact F
Prior belief → Posterior belief
$P(F) \to P(F|E)$

A Closer Look at the Answer

Assume a patient is picked at random.

- Prior probability of disease
  - P(Disease) = 0.001 = 0.1%
- Posterior probability of disease given positive test
  - P(Disease | Test positive) = 0.0194... ≅ 2%
- Bigger than the prior, but still pretty small
- Should we approve such a test?
  - The test has **low error rates** compared to most tests
- How can this be?

There's a rare disease with prevalence of 1/1000 in the population.
A test for the disease has a false positive rate of 5% and a false negative rate of 1%.
b). You test positive. What's the probability you have the disease?



- Out of 1000 people:
  - # of people with disease: 1000*(1/1000) = 1
  - # of people who test positive and have the disease:

    1000*(.001*.99) ≅ 1
  - # of people who test positive and don't have the disease:

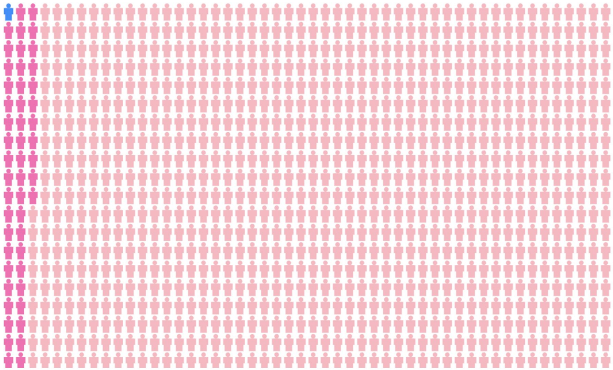
    1000*(.999*.05) ≅ 50
  - # of people who test positive: 1+50 = 51

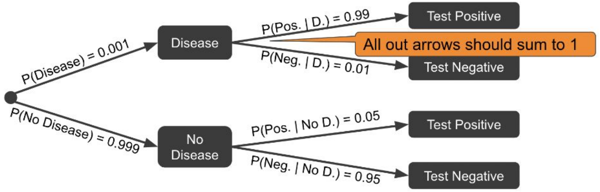
P(D | Pos) ≅ 1/51 ≅ 2%

There's a rare disease with prevalence of 1/1000 in the population.
A test for the disease has a false positive rate of 5% and a false negative rate of 1%.

c). You test NEGATIVE. What's the prob you have the disease?

Here is the whole population of 1000 people:





- Out of 1000 people:
  - # of people with disease: 1000*(1/1000) = 1
  - # of people who test positive and have the disease: 1000*(.001*.99) ≅ 1
  - # of people who test positive and don't have the disease: 1000*(.999*.05) ≅ 50
  - # of people who test positive: 1+50 = 51

P(D | Pos) ≅ 1/51 ≅ 2%

- # of people who test negative and DO have the disease: 1000*(.001*.01) =0.01 ≅ 0
- # of people who test negative and don't have the disease: 1000*(.999*.95) ≅ 949
- # of people who test negative: 0+949= 949

P(D | Neg)=0.01/949 =0.00001≅0%

Suppose there's a rare disease with prevalence of 1/1000 in the population.
A test for the disease has a false positive rate of 5% and a false negative rate of 1%.

Why it's still good to get tested

E = you test positive
F = you actually have the disease
E^C = you test negative

I have a 0.001 chance of having disease.

$P(F)$

Take test, results positive

With these test results, I now have a 0.02 chance of having disease

$P(F|E)$

Take test, results negative

With these test results, I now have a 0.00001 chance of having disease!!!

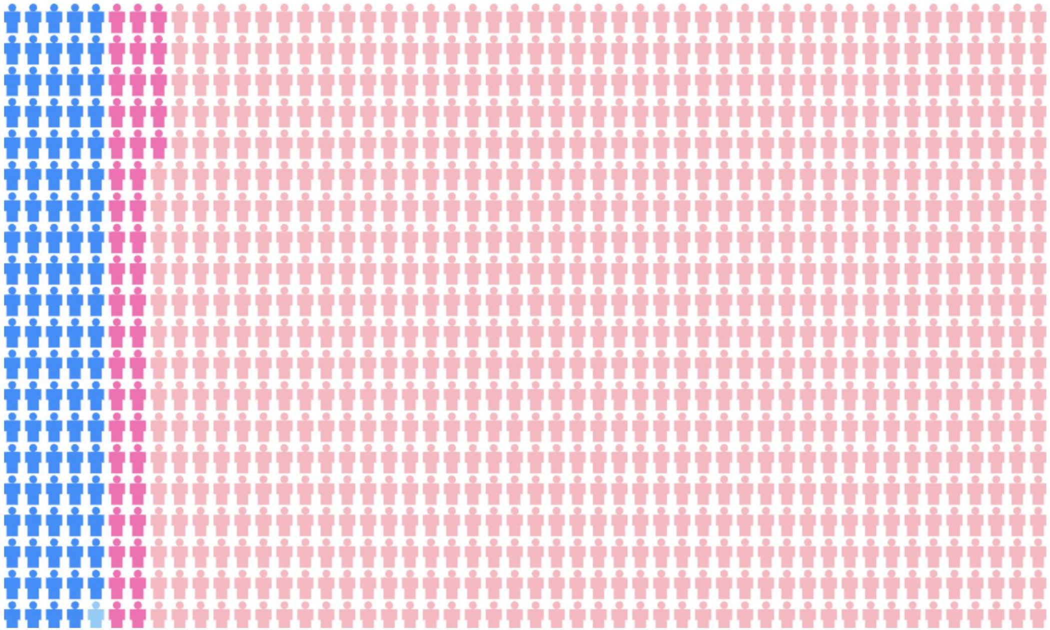
$P(F|E^C)$

Assumptions Matter

- **“Assume a patient is picked at random.”**
  - But usually, people aren’t picked at random for medical tests
  - So our intuition about randomly picked patients may not be great
- For a ***randomly picked patient, the result does make sense***, because the disease is very rare.
- What if the doctor believes there is a 10% chance the patient has the disease?

There’s a rare disease with prevalence of 1/1000 in the population.
A test for the disease has a false positive rate of 5%
and a false negative rate of 1%.
b). You test positive. What’s the prob you have the disease?

- What if, based on additional information your doctor believes you are in a subpopulation where there is a 10% chance you have the disease?



- Out of 1000 people:

- # of people with disease:
1000*(1/10) = 100

- # of people who test positive and have the disease:

  1000*(0.1*0.99) ≅ 99

- # of people who test positive and don’t have the disease:

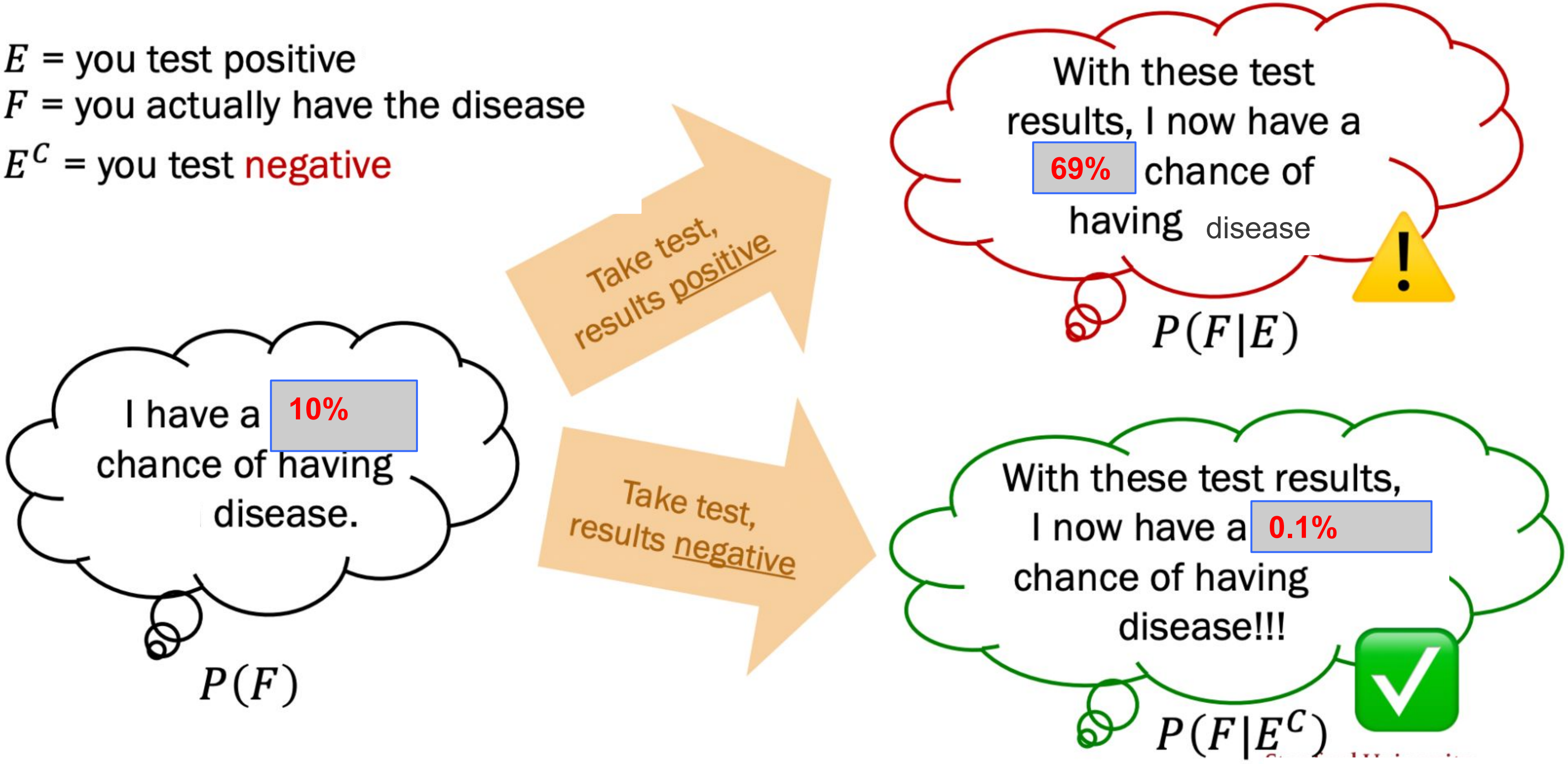
  1000*(.90*.05) ≅ 45

- # of people who test positive:
99+45 = 144

P(D | Pos) ≅ 99/144 ≅ 69%

Assumptions Matter

E = you test positive
F = you actually have the disease
E^C = you test negative



Subjective Probabilities

likelihood prior

posterior

$$P(F|E) = \frac{P(E|F)P(F)}{P(E)}$$

normalization constant

Recall: the probability of an outcome can be defined as:

[Frequentist] The **frequency** with which it will occur in repeated trials:

$$\mathrm{P(Event)} = \lim_{n \to \infty} \frac{\mathrm{count(Event)}}{n}$$

[Bayesian] Someone's **subjective degree** of belief that it will occur

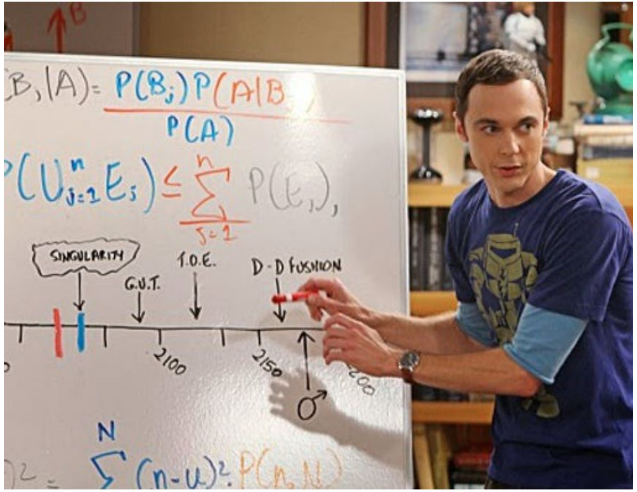
Why use subjective priors?

- To quantify your degree of uncertainty about an outcome, even when there is no physical randomization
    - i.e. chance of CU football team getting into a Bowl game next year
    - i.e. chance of the "Big One" in the next 30 years

Purpose of Bayes' Theorem

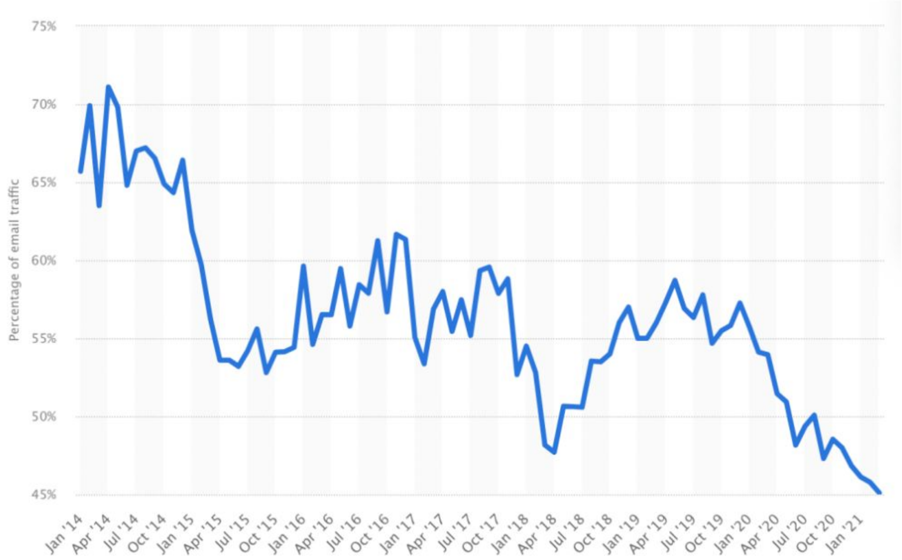
- Update your prediction based on new information

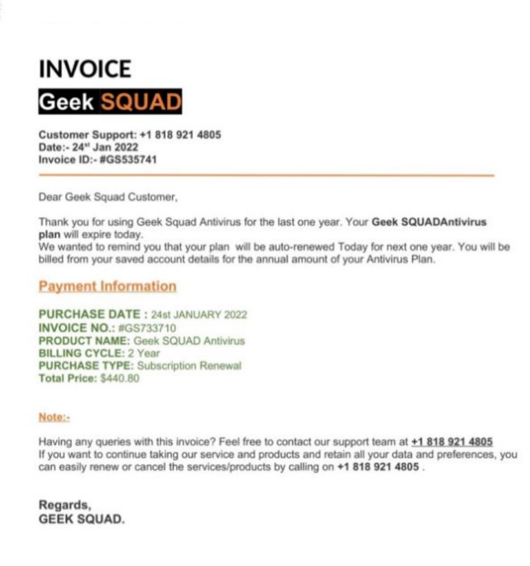
- In a multi-stage experiment,
  find the chance of an event at an earlier stage,
  an earlier stage,
  given the result of a later stage



More Practice

Detecting spam email





INVOICE

Geek SQUAD

Customer Support: +1 818 921 4805
Date:- 24th Jan 2022
Invoice ID:- #GS535741

Dear Geek Squad Customer,

Thank you for using Geek Squad Antivirus for the last one year. Your **Geek SQUADAntivirus plan** will expire today.
We wanted to remind you that your plan will be auto-renewed Today for next one year. You will be billed from your saved account details for the annual amount of your Antivirus Plan.

Payment Information

PURCHASE DATE : 24st JANUARY 2022
INVOICE NO.: #GS733710
PRODUCT NAME: Geek SQUAD Antivirus
BILLING CYCLE: 2 Year
PURCHASE TYPE: Subscription Renewal
Total Price: $440.80

Note:-

Having any queries with this invoice? Feel free to contact our support team at +1 818 921 4805
If you want to continue taking our service and products and retain all your data and preferences, you can easily renew or cancel the services/products by calling on **+1 818 921 4805** .

Regards,
GEEK SQUAD.

We can easily calculate how many existing spam emails contain "Dear":

$$P(E|F) = P\left(\text{"Dear"} \;\middle|\; \begin{matrix}\text{Spam}\\ \text{email}\end{matrix}\right)$$

But what is the probability that a mystery email containing "Dear" is spam?

$$P(F|E) = P\left(\begin{matrix}\text{Spam}\\ \text{email}\end{matrix} \;\middle|\; \text{"Dear"}\right)$$

More Practice

Detecting spam email

$$P(F|E) = \frac{P(E|F)P(F)}{P(E|F)P(F) + P(E|F^C)P(F^C)}$$ Bayes' Theorem

- 60% of all email in 2016 is spam.
- 20% of spam has the word "Dear"
- 1% of non-spam (aka ham) has the word "Dear"

You get an email with the word "Dear" in it.

What is the probability that the email is spam?

A). ~13%
B). ~52%
C). ~83%
D). ~97%
E). none of these

1. Define events & state goal
2. Identify known probabilities
3. Solve

More Practice

Detecting spam email

$$P(F|E) = \frac{P(E|F)P(F)}{P(E|F)P(F) + P(E|F^C)P(F^C)}$$ Bayes' Theorem

- 60% of all email in 2016 is spam.
- 20% of spam has the word "Dear"
- 1% of non-spam (aka ham) has the word "Dear"

You get an email with the word "Dear" in it.

What is the probability that the email is spam?

A). ~13%
B). ~52%
C). ~83%
D). ~97%
E). none of these

1. Define events & state goal
2. Identify known probabilities
3. Solve

Let: E: "Dear", F: spam

Want: P(spam | "Dear")

$= P(F|E)$

More Practice

- 60% of all email in 2016 is spam.
- 20% of spam has the word "Dear"
- 1% of non-spam (aka ham) has the word "Dear"

You get an email with the word "Dear" in it.
What is the probability that the email is spam?

1. Define events & state goal

Let: E: "Dear", F: spam
Want: $P(\text{spam} \mid \text{"Dear"})$
$= P(F|E)$

Note: You should still know how to use Bayes/ Law of Total Prob., but drawing a probability tree can help you identify which probabilities you have. The branches are determined using the problem setup.

A). ~13%
B). ~52%
C). ~83%
D). ~97%
E). none of these

Simulating Probabilities

Lesson 9, CSCI 3022

- DEMO!