Artificial Intelligence

CS 165A

Jan 15 2019

Today

→ Uncertainty (Ch 13)

Announcements

- Course website:

https://www.cs.ucsb.edu/~yuxiangw/classes/CS165A-2019winter/

- Homework 1 will be posted in the Assignment subdirectory midnight **Jan 17 (Thursday).**
  - Homework submission in **hard copies**
  - Exact location will be announced on Piazza.
  - Print-outs of latex created pdf are prefered!
  - **Due date Jan 29**. Start early!

Quick Review of Probability

From here on we will assume that you know this…

containing anonymous slides (slides 4-13) from the Web

Deterministic vs. Random Processes

- In deterministic processes, the outcome can be predicted exactly in advance
  - Eg. Force = Mass x Acceleration. If we are given values for mass and acceleration, we exactly know the value of force

- In random processes, the outcome is not known exactly, but we can still describe the *probability distribution* of possible outcomes
  - Eg. 10 coin tosses: we don't know exactly how many heads we will get, but we can calculate the probability of getting a certain number of heads

Events

- An **event** is an outcome or a set of outcomes of a random process

Example: Tossing a coin three times

Event A = getting exactly two heads = {HTH, HHT, THH}

Example: Picking real number X between 1 and 20

Event A = chosen number is at most 8.23 = $\{X \leq 8.23\}$

Example: Tossing a fair dice

Event A = result is an even number = {2, 4, 6}

- Notation: P(A) = Probability of event A
- **Probability Rule 1:**

$$\mathbf{0 \leq P(A) \leq 1 \text{ for any event } A}$$

Sample Space

- The **sample space** S of a random process is the set of all possible outcomes

  Example: one coin toss

  S = {H,T}

  Example: three coin tosses

  S = {HHH, HTH, HHT, TTT, HTT, THT, TTH, THH}

  Example: roll a six-sided dice

  S = {1, 2, 3, 4, 5, 6}

  Example: Pick a real number X between 1 and 20

  S = all real numbers between 1 and 20

- **Probability Rule 2: The probability of the whole sample space is 1**

$$P(S) = 1$$

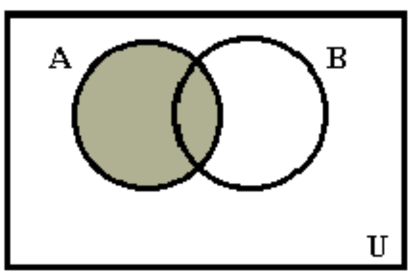
Combinations of Events

- The **complement** A^c of an event A is the event that A does not occur
- **Probability Rule 3:**

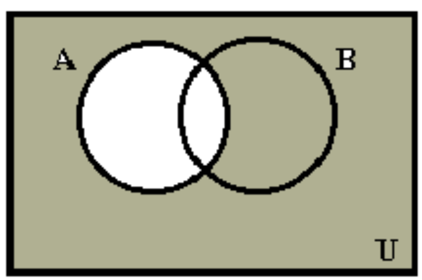
$$P(A^c) = 1 - P(A)$$

- The **union** of two events A and B is the event that either A or B or both occurs
- The **intersection** of two events A and B is the event that both A and B occur

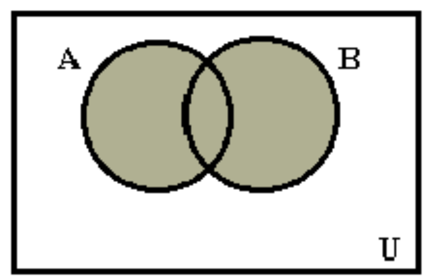
Event A



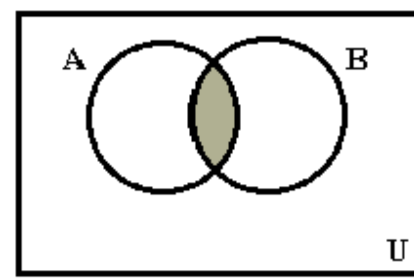
Complement of A



Union of A and B



Intersection of A and B



Disjoint Events

- Two events are called **disjoint** if they can not happen at the same time
  - Events A and B are disjoint means that the intersection of A and B is zero
- Example: coin is tossed twice
  - S = {HH,TH,HT,TT}
  - Events A={HH} and B={TT} are disjoint
  - Events A={HH,HT} and B = {HH} are not disjoint

- **Probability Rule 4: If A and B are disjoint events then**

$$\mathbf{P(A \text{ or } B) = P(A) + P(B)}$$

Independent events

- Events A and B are **independent** if knowing that A occurs does not affect the probability that B occurs

- Example: tossing two coins

  Event A = first coin is a head

  Event B = second coin is a head

  Independent

- Disjoint events cannot be independent!
  - If A and B can not occur together (disjoint), then knowing that A occurs does change probability that B occurs

- **Probability Rule 5: If A and B are independent**

$$\mathbf{P(A \text{ and } B) = P(A) \times P(B)}$$

multiplication rule for independent events

Equally Likely Outcomes Rule

- If all possible outcomes from a random process have the same probability, then
- P(A) = (# of outcomes in A)/(# of outcomes in S)
- Example: One Dice Tossed

P(even number) = |2,4,6| / |1,2,3,4,5,6|

- Note: equal outcomes rule only works if the number of outcomes is "countable"
  - Eg. of an uncountable process is sampling any fraction between 0 and 1. Impossible to count all possible fractions !

Combining Probability Rules Together

- Initial screening for HIV in the blood first uses an enzyme immunoassay test (EIA)
- Even if an individual is HIV-negative, EIA has probability of 0.006 of giving a positive result
- Suppose 100 people are tested who are all HIV-negative. What is probability that at least one will show positive on the test?
- First, use complement rule:

P(at least one positive) = 1 - P(all negative)

Combining Probability Rules Together

- Now, we assume that each individual is independent and use the multiplication rule for independent events:

$$\mathbf{P(all\ negative) = P(test\ 1\ negative) \times \ldots \times P(test\ 100\ negative)}$$

- P(test negative) = 1 - P(test positive) = 0.994

$$\mathbf{P(all\ negative) = 0.994 \times \ldots \times 0.994 = (0.994)^{100}}$$

- So, we finally we have

$$\mathbf{P(at\ least\ one\ positive) = 1 - (0.994)^{100} = 0.452}$$

Random variables (R.V.)

- A random variable is a variable whose possible values are outcomes of a random process or random event.

- Example: three tosses of a coin
  - S = {HHH,THH,HTH,HHT,HTT,THT,TTH,TTT}
  - Random variable X = number of observed tails
  - Possible values for X = {0,1, 2, 3}

- Why do we need random variables?
  - We use them as a model for our observed data

Probability notation and notes

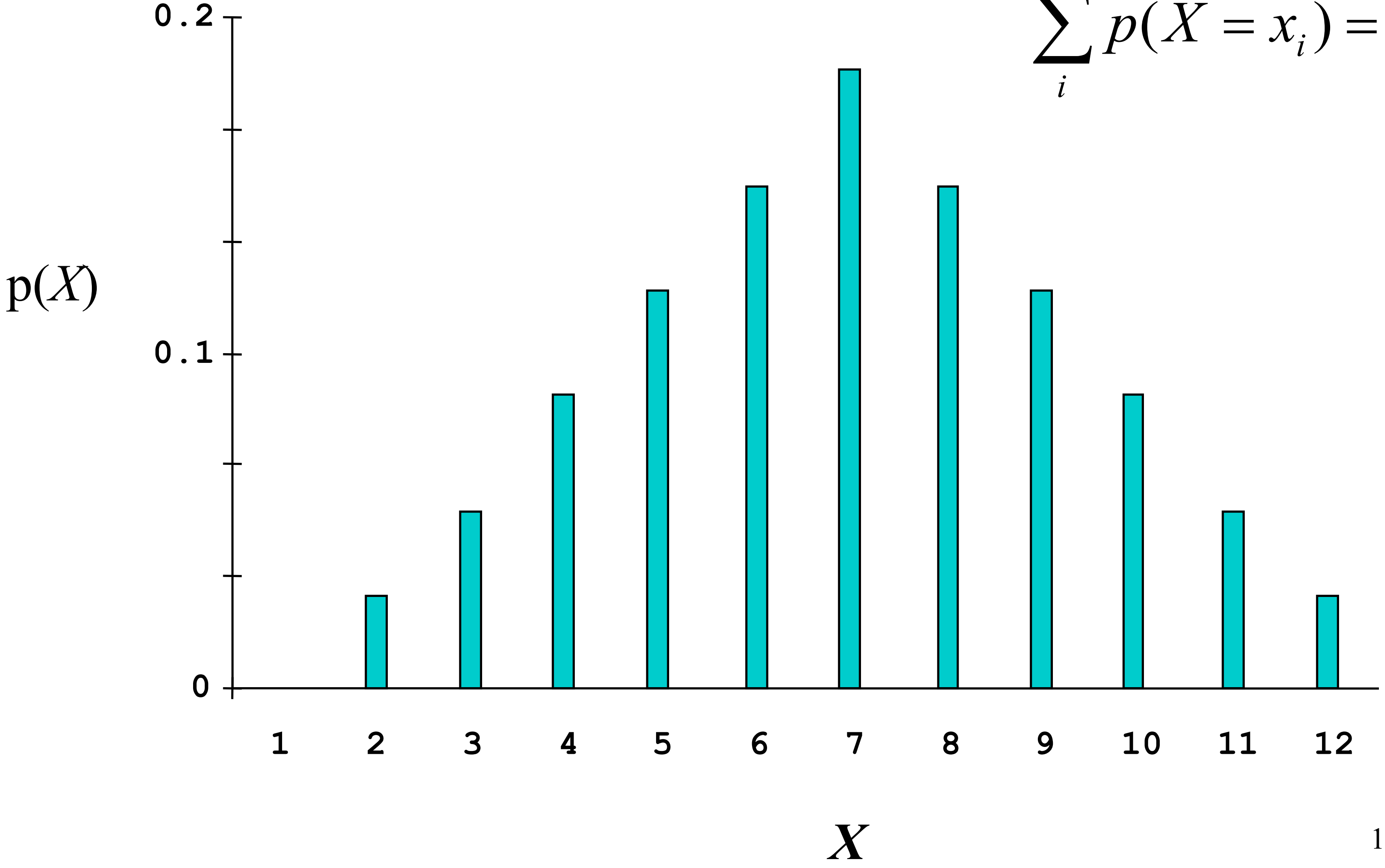
- Probabilities of *propositions*
  - P(A), P(the sun is shining)
- Probabilities of *random variables*
  - $P(X = x_1)$, $P(Y = y_1)$, $P(x_1 < X < x_2)$
- P(A) usually means P(A = True) (A is a proposition, not a variable)
  - This is a probability **value**
  - Technically, P(A) is a probability *function*
- $P(X = x_1)$
  - This is a probability **value** (P(X) is a probability *function*)
- P(X)
  - This is a probability **mass function** or a **probability density function**
- Technically, if X is a variable, we should not write $P(X) = 0.5$
  - But rather $P(X = x_1) = 0.5$

Discrete and continuous probabilities

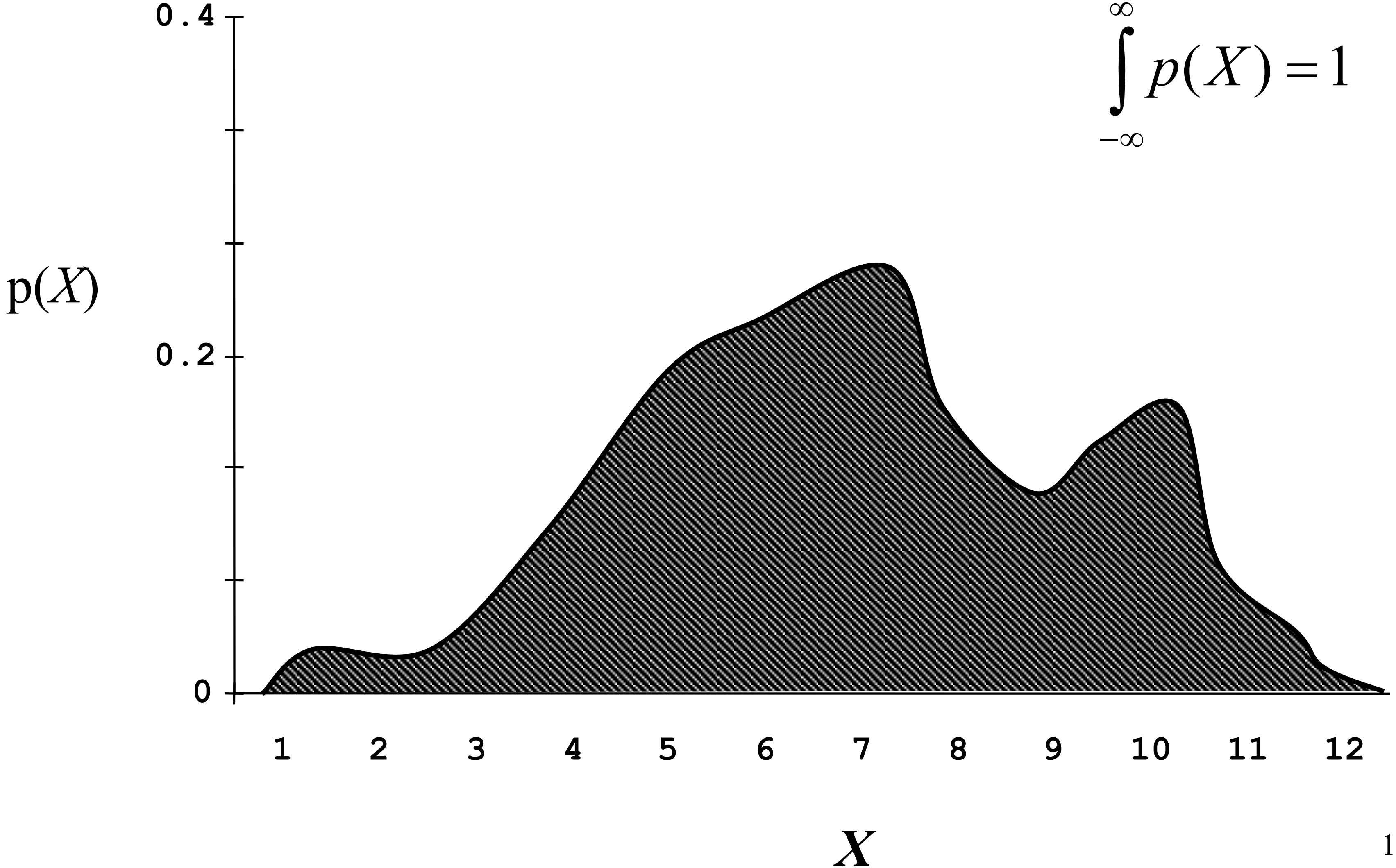
- Discrete: Probability Mass Function P(X, Y) is described by an MxN matrix of probabilities
  - Possible values of each: $P(X=x_1, Y=y_1) = p_1$
  - $\sum P(X=x_i, Y=y_j) = 1$
  - P(X, Y, Z) is an MxNxP matrix
- Continuous: Probability density function (**pdf**) P(X, Y) is described by a 2D function
  - $P(x_1 < X < x_2, y_1 < Y < y_2) = p_1$
  - $\int P(X, Y)\, dX\, dY = 1$

Discrete probability distribution

$$\sum_i p(X = x_i) = 1$$



Continuous probability distribution



$$\int_{-\infty}^{\infty} p(X) = 1$$

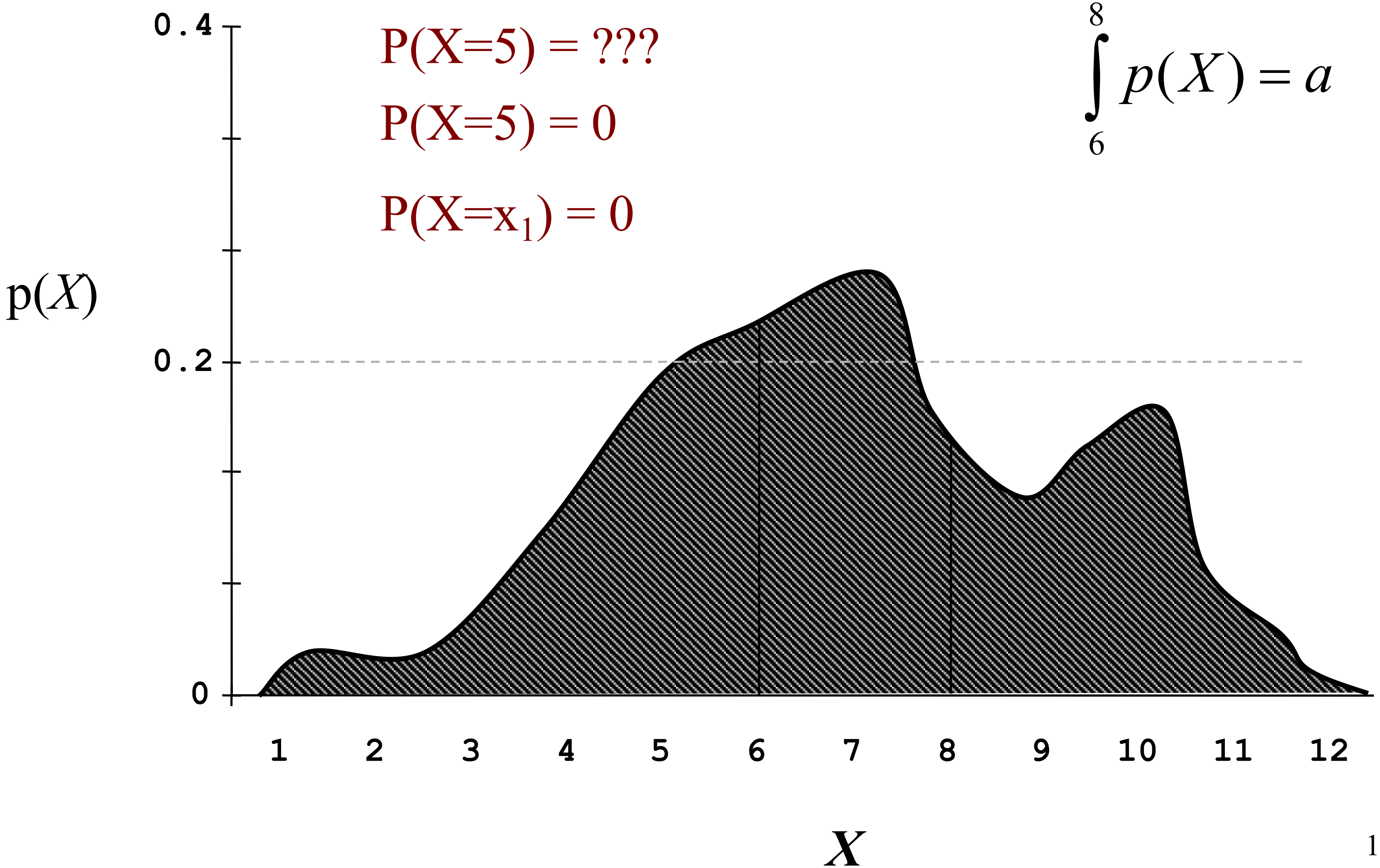
Continuous probability distribution

P(X=5) = ???

P(X=5) = 0

$P(X=x_1) = 0$

$$\int_6^8 p(X) = a$$



Three Axioms of Probability

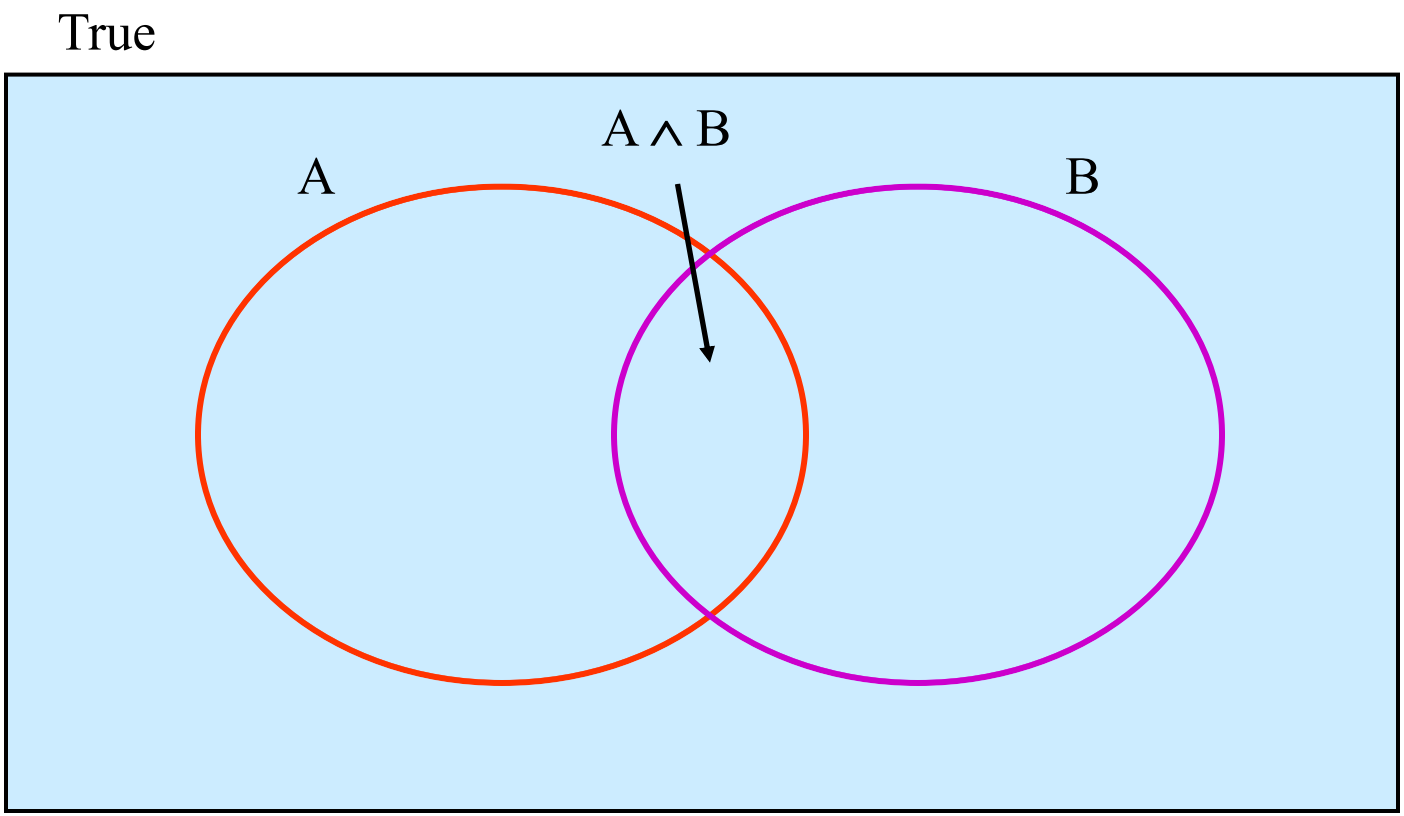
1. The probability of every event must be nonnegative
   - For any event A, $P(A) \geq 0$
2. Valid propositions have probability 1
   - $P(True) = 1$
   - $P(A \vee \neg A) = 1$
3. For disjoint events $A_1, A_2, \ldots$
   - $P(A_1 \vee A_2 \vee \ldots) = P(A_1) + P(A_2) + \ldots$

- From these axioms, all other properties of probabilities can be derived.
  - E.g., derive $P(A) + P(\neg A) = 1$

Some consequences of the axioms

- Unsatisfiable propositions have probability 0
  - $P(False) = 0$
  - $P(A \wedge \neg A) = 0$
- For any two events A and B
  - $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
- For the complement A^c of event A
  - $P(A^c) = 1 - P(A)$
- For any event A
  - $0 \leq P(A) \leq 1$
- For independent events A and B
  - $P(A \wedge B) = P(A)\, P(B)$

Venn Diagram



Visualize: P(True), P(False), P(A), P(B), P(¬A), P(¬B), P(A ∨ B), P(A ∧ B), P(A ∧ ¬B), …

Joint Probabilities

- A **complete probability model** is a single joint probability distribution over all propositions/variables in the domain
  - $P(X_1, X_2, \ldots, X_i, \ldots)$
- A particular instance of the world has the probability
  - $P(X_1=x_1 \wedge X_2=x_2 \wedge \ \ldots \wedge X_i=x_i \wedge \ \ldots) = p$
- Rather than stating knowledge as
  - Raining $\Rightarrow$ WetGrass
- We can state it as
  - P(Raining, WetGrass) = 0.15
  - P(Raining, ¬WetGrass) = 0.01
  - P(¬Raining, WetGrass) = 0.04
  - P(¬Raining, ¬WetGrass) = 0.8

	¬WetGrass	WetGrass
¬Raining	0.8	0.04
Raining	0.01	0.15

Marginal and Conditional Probability

- Marginal, or Prior, Probability
  - **Probabilities** associated with a proposition or variable, **prior to any evidence**
  - E.g., P(WetGrass), P(¬Raining)
- Conditional, or Posterior, Probability
  - **Probabilities after evidence is gathered**
  - P(A | B) – "The probability of A given that we know B"
  - After (posterior to) procuring evidence
  - E.g., P(WetGrass | Raining)

$$P(X \mid Y) = \frac{P(X,Y)}{P(Y)} \qquad \text{or} \qquad P(X \mid Y)\,P(Y) = P(X,Y)$$

Assumes P(Y) nonzero

Where does the word “marginal” come from?

- A joke (by Larry Lesser and Dennis Pearl):
  - Teacher: To get the marginal of X from the joint pdf of X and Y, you should integrate.....
  - Student: Can you go over why?
  - Teacher: Correct!

- “Statistics is the only field where you can be marginalized while being integrated at the same time.” – unknown quote

- Historical reason: actuary practice.

The chain rule

$$P(X,Y) = P(X \mid Y)\,P(Y)$$

By the Chain Rule

$$P(X,Y,Z) = P(X \mid Y,Z)P(Y,Z)$$

$$= P(X \mid Y,Z)\,P(Y \mid Z)\,P(Z)$$

$$or, equivalently$$

$$= P(X)P(Y \mid X)P(Z \mid X,Y)$$

Notes:
- Precedence: '|' is lowest
- E.g., P(X | Y, Z) means which?
  P((X | Y), Z)
  P(X | (Y, Z)) ⇦

Joint probability distribution

From $P(X,Y)$, we can always calculate:

$P(X)$	$P(X=x_1)$
$P(Y)$	$P(Y=y_2)$
$P(X\|Y)$	$P(X\|Y=y_1)$
$P(Y\|X)$	$P(Y\|X=x_1)$
	$P(X=x_1\|Y)$
	etc.

X

Y

	x_1	x_2	x_3
y_1	0.2	0.1	0.1
y_2	0.1	0.2	0.3

P(X,Y)

	x_1	x_2	x_3
y_1	0.2	0.1	0.1
y_2	0.1	0.2	0.3

P(X)

x_1	x_2	x_3
0.3	0.3	0.4

P(Y)

y_1	0.4
y_2	0.6

P(X|Y)

	x_1	x_2	x_3
y_1	0.5	0.25	0.25
y_2	0.167	0.333	0.5

P(Y|X)

	x_1	x_2	x_3
y_1	0.667	0.333	0.25
y_2	0.333	0.667	0.75

$P(X=x_1, Y=y_2) = ?$

$P(X=x_1) = ?$

$P(Y=y_2) = ?$

$P(X|Y=y_1) = ?$

$P(X=x_1|Y) = ?$

Probability Distributions

	Continuous vars	Discrete vars
P(X)	Function (of one variable)	M vector
P(X=x)	Scalar*	Scalar
P(X,Y)	Function of two variables	MxN matrix
P(X|Y)	Function of two variables	MxN matrix
P(X|Y=y)	Function of one variable	M vector
P(X=x|Y)	Function of one variable	N vector
P(X=x|Y=y)	Scalar*	Scalar

* - actually zero. Should be $P(x_1 < X < x_2)$

Bayes' Rule

Thomas Bayes: 1701 - 1761

- Since $P(X,Y) = P(X \mid Y)\,P(Y)$

  and $P(X,Y) = P(Y \mid X)\,P(X)$

- Then $P(X \mid Y)\,P(Y) = P(Y \mid X)\,P(X)$

$$P(X \mid Y) = \frac{P(Y \mid X)\;P(X)}{P(Y)}$$ Bayes' Rule

Funny fact: Thomas Bayes is arguably a frequentist.

Stephen Fienberg. "When did Bayesian inference become 'Bayesian'?." *Bayesian analysis* 1.1 (2006): 1-40.

https://projecteuclid.org/euclid.ba/1340371071

Bayes' Rule

- Similarly, P(X) conditioned on two variables:

$$P(X \mid Y \quad) = \frac{P(Y \mid X \quad) P(X \quad)}{P(Y \quad)}$$

$$P(X \mid \quad Z) = \frac{P(Z \mid X \quad) P(X \quad)}{P(Z \quad)}$$

- Or N variables:

$$P(X_1 \mid X_2 \qquad\qquad) = \frac{P(X_2 \mid X_1 \qquad\qquad) P(X_1 \qquad\qquad)}{P(X_2 \qquad\qquad)}$$

Bayes' rule for Bayesian Inference

- This simple equation is very useful in practice
  - Usually framed in terms of hypotheses (H) and data (D)
    - Which of the hypotheses is best supported by the data?

Likelihood
(causal knowledge)

Prior probability

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

Posterior probability
(diagnostic knowledge)

Normalizing constant

$$\underbrace{P(H_i \mid D)} = k\, P(D \mid H_i)\, \underbrace{P(H_i)}$$

Bayes' rule example: Medical diagnosis

- Meningitis causes a stiff neck 50% of the time
- A patient comes in with a stiff neck – what is the probability that he has meningitis?
- Need to know two things:
  - The prior probability of a patient having meningitis (1/50,000)
  - The prior probability of a patient having a stiff neck (1/20)
- ?

$$P(M \mid S) = \frac{P(S \mid M)\, P(M)}{P(S)}$$

- P(M | S) = (0.5)(0.00002)/(0.05) = 0.0002

Example (cont.)

- Suppose that we also know about whiplash
  - P(W) = 1/1000
  - P(S | W) = 0.8
- What is the relative likelihood of whiplash and meningitis?
  - P(W | S) / P(M | S)

$$P(W \mid S) = \frac{P(S \mid W)\, P(W)}{P(S)} = \frac{(0.8)(0.001)}{0.05} = 0.016$$

So the relative likelihood of whiplash vs. meningitis is (0.016/0.0002) = 80

A useful Bayes rule example

A test for a new, deadly strain of anthrax (that has no symptoms) is known to be 99.9% accurate. Should you get tested? The chances of having this strain are one in a million.

What are the random variables?

A – you have anthrax (boolean)

T – you test positive for anthrax (boolean)

Notation: Instead of P(A=True) and P(A=False), we will write P(A) and P(¬A)

What do we want to compute?

P(A|T)

What else do we need to know or assume?

Priors: P(A) , P(¬A)

Given: P(T|A) , P(T|¬A), P(¬T|A), P(¬T|¬A)

Possibilities

A T	¬A T
A ¬T	¬A ¬T

Example (cont.)

We know:

- Given: $P(T|A) = 0.999$, $P(T|\neg A) = 0.001$, $P(\neg T|A) = 0.001$, $P(\neg T|\neg A) = 0.999$
- Prior knowledge: $P(A) = 10^{-6}$, $P(\neg A) = 1 - 10^{-6}$

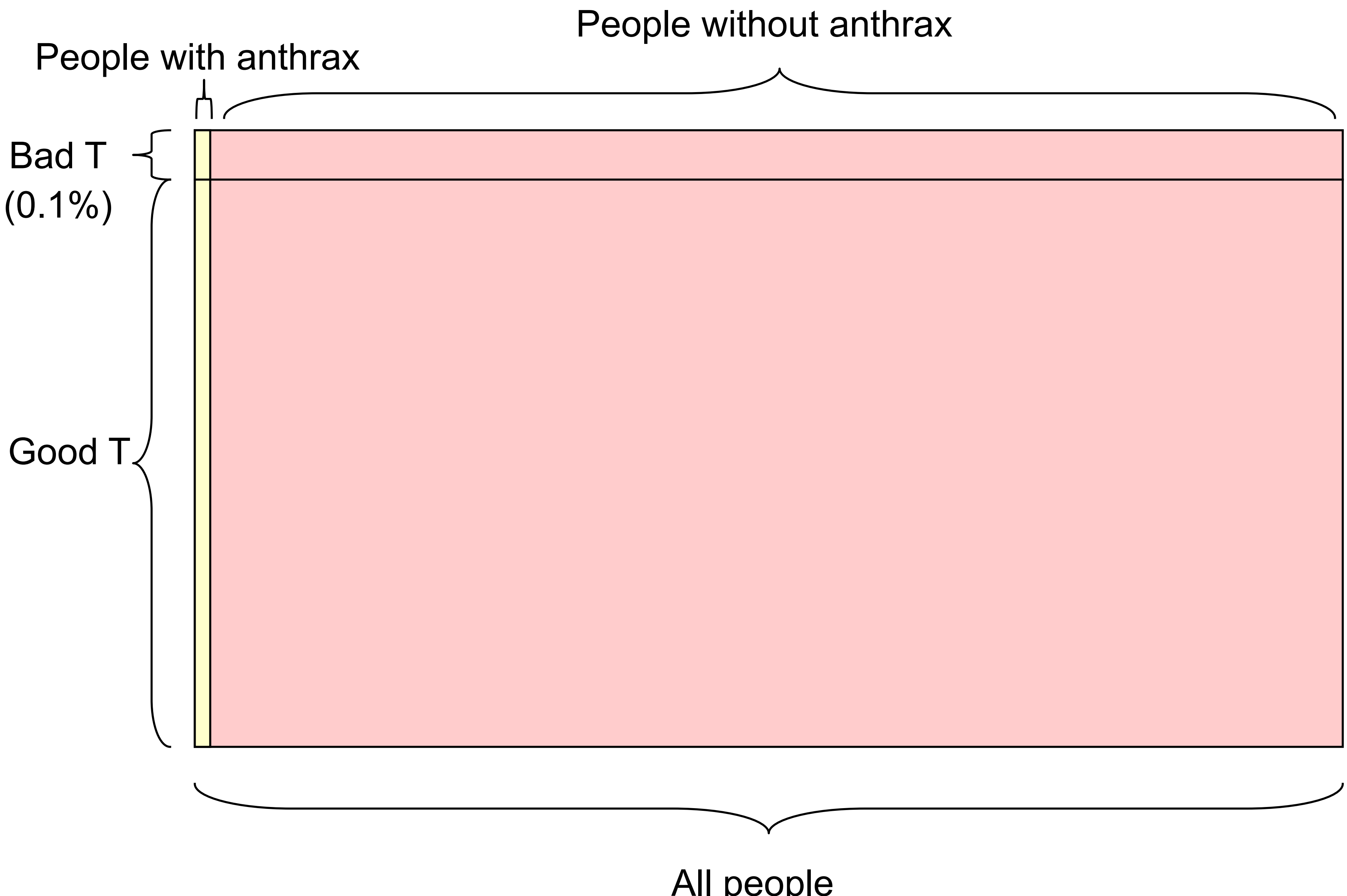
Want to know $P(A|T)$

- $P(A|T) = P(T|A)\ P(A) / P(T)$

Calculate $P(T)$ by marginalization

- $P(T) = P(T|A)\ P(A) + P(T|\neg A)\ P(\neg A) = (0.999)(10^{-6}) + (0.001)(1 - 10^{-6}) \approx 0.001$
- So $P(A|T) = (0.999)(10^{-6}) / 0.001 \approx 0.001$
- Therefore $P(\neg A|T) \approx 0.999$

What if you work at a Post Office?


People without anthrax
People with anthrax
Bad T
(0.1%)
Good T
All people

For you to think about / discuss on Piazza

1. Space complexity of representing a joint distribution of n discrete variables.

2. Time complexity of calculating the marginals / conditionals

3. Bayesian vs. frequentist definition of probabilities.

- Bonus points for class participation!