

Basic Probability and Statistics

Yingyu Liang

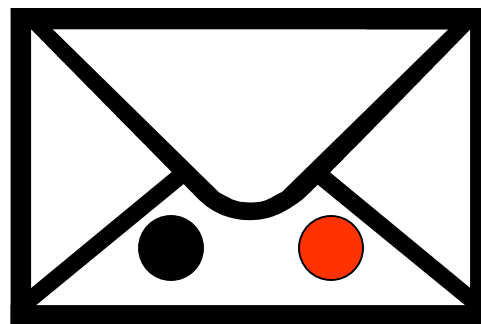
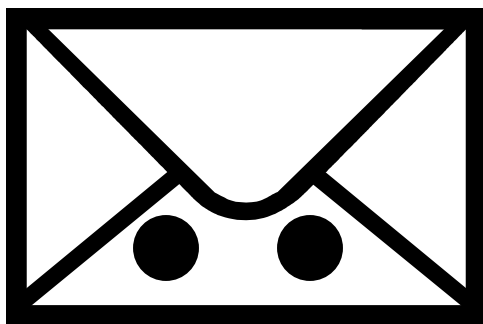
`yliang@cs.wisc.edu`

**Computer Sciences Department
University of Wisconsin, Madison**

[based on slides from Jerry Zhu, Mark Craven]

Reasoning with Uncertainty

- There are two identical-looking envelopes
 - one has a red ball (worth \$100) and a black ball
 - one has two black balls. Black balls worth nothing



- You randomly grabbed an envelope, randomly took out one ball – it's black.
- At this point you're given the option to switch the envelope. **To switch or not to switch?**

Outline

- Probability
 - random variable
 - Axioms of probability
 - Conditional probability
 - Probabilistic inference: Bayes rule
 - Independence
 - Conditional independence

Uncertainty

- Randomness
 - Is our world random?
- Uncertainty
 - Ignorance (practical and theoretical)
 - Will my coin flip ends in head?
 - Will bird flu strike tomorrow?
- Probability is the language of uncertainty
 - Central pillar of modern day artificial intelligence

Sample space

- A space of outcomes that we assign probabilities to
- Outcomes can be binary, multi-valued, or continuous
- Outcomes are mutually exclusive
- Examples
 - Coin flip: {head, tail}
 - Die roll: {1,2,3,4,5,6}
 - English words: a dictionary
 - Temperature tomorrow: R_+ (kelvin)

Random variable

- A variable, x , whose domain is the sample space, and whose value is somewhat uncertain
- Examples:
 - x = coin flip outcome
 - x = first word in tomorrow's headline news
 - x = tomorrow's temperature
- Kind of like $x = \text{rand}()$

Probability for discrete events

- Probability $P(x=a)$ is the fraction of times x takes value a
- Often we write it as $P(a)$
- There are other definitions of probability, and philosophical debates... but we'll not go there
- Examples
 - $P(\text{head})=P(\text{tail})=0.5$ fair coin
 - $P(\text{head})=0.51$, $P(\text{tail})=0.49$ slightly biased coin
 - $P(\text{head})=1$, $P(\text{tail})=0$ Jerry's coin
 - $P(\text{first word} = \text{"the"} \text{ when flipping to a random page in NYT})=?$
- Demo: Search **"The Book of Odds"**

Probability table

- Weather

Sunny	Cloudy	Rainy
200/365	100/365	65/365

- $P(\text{Weather} = \text{sunny}) = P(\text{sunny}) = 200/365$
- $P(\text{Weather}) = \{200/365, 100/365, 65/365\}$
- For now we'll be satisfied with obtaining the probabilities by counting frequency from data...

Probability for discrete events

- Probability for more complex events A
 - $P(A=\text{"head or tail"})=?$ fair coin
 - $P(A=\text{"even number"})=?$ fair 6-sided die
 - $P(A=\text{"two dice rolls sum to 2"})=?$

Probability for discrete events

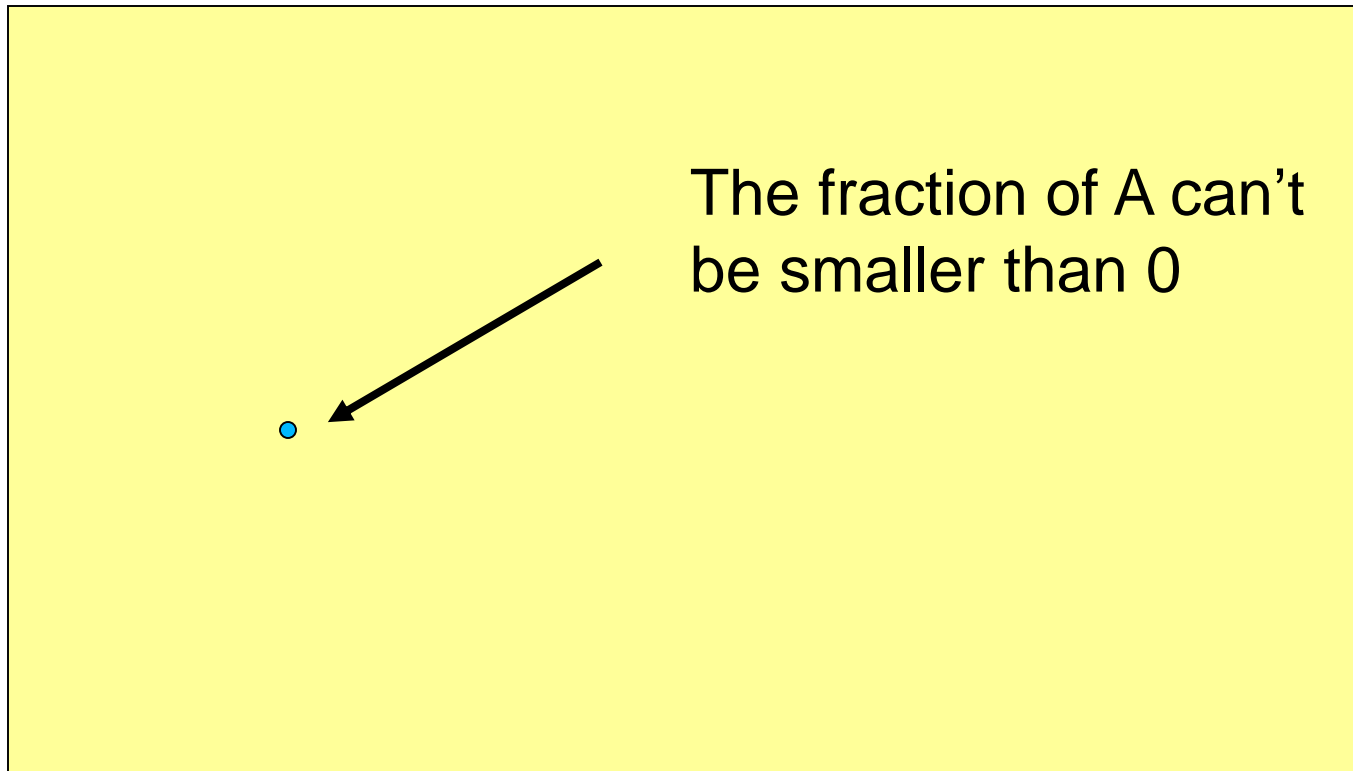
- Probability for more complex events A
 - $P(A=\text{"head or tail"})=0.5 + 0.5 = 1$ fair coin
 - $P(A=\text{"even number"})=1/6 + 1/6 + 1/6 = 0.5$ fair 6-sided die
 - $P(A=\text{"two dice rolls sum to 2"})=1/6 * 1/6 = 1/36$

The axioms of probability

- $P(A) \in [0,1]$
- $P(\text{true})=1, P(\text{false})=0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

The axioms of probability

- $P(A) \in [0,1]$
- $P(\text{true})=1, P(\text{false})=0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

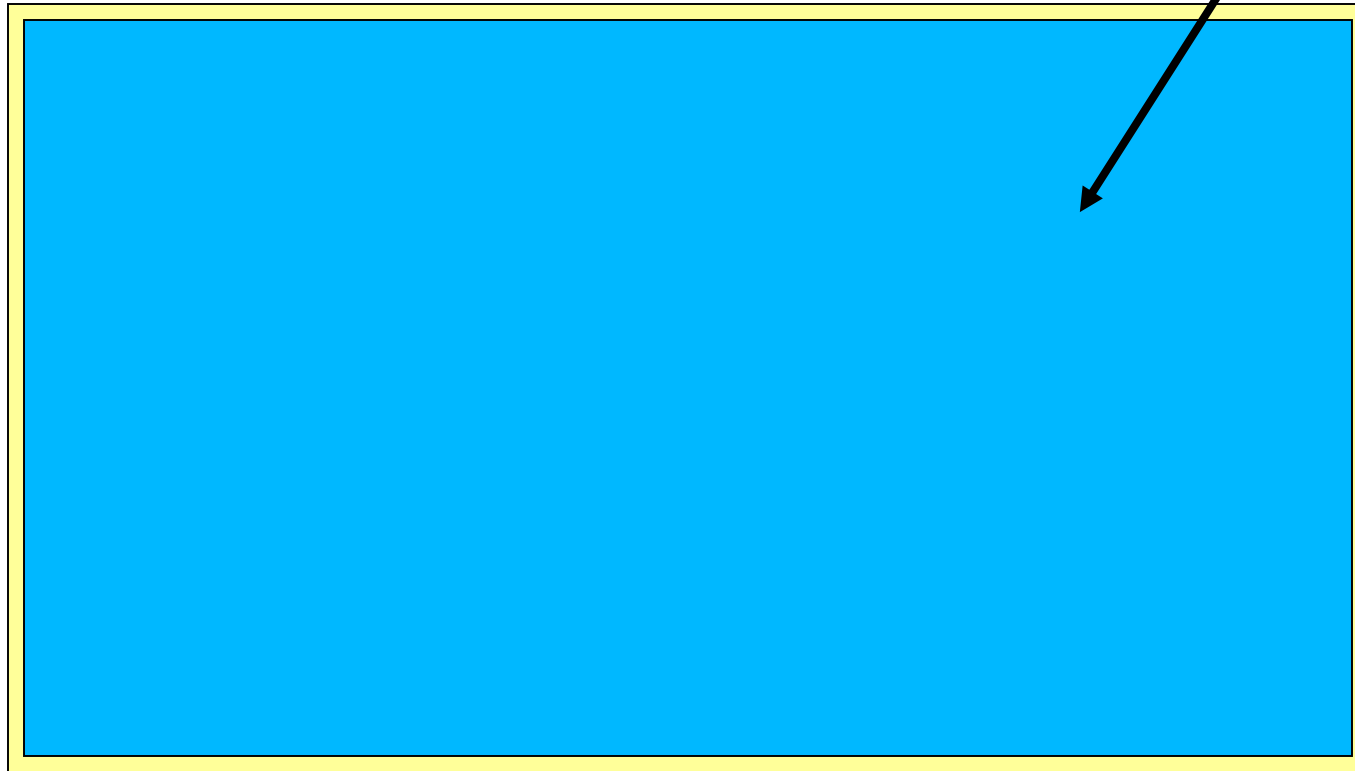


Sample space

The axioms of probability

- $P(A) \in [0,1]$
- $P(\text{true})=1, P(\text{false})=0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

The fraction of A can't be bigger than 1



Sample space

The axioms of probability

- $P(A) \in [0,1]$
- $P(\text{true})=1$, $P(\text{false})=0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

Valid sentence: e.g. “x=head or x=tail”

Sample
space

The axioms of probability

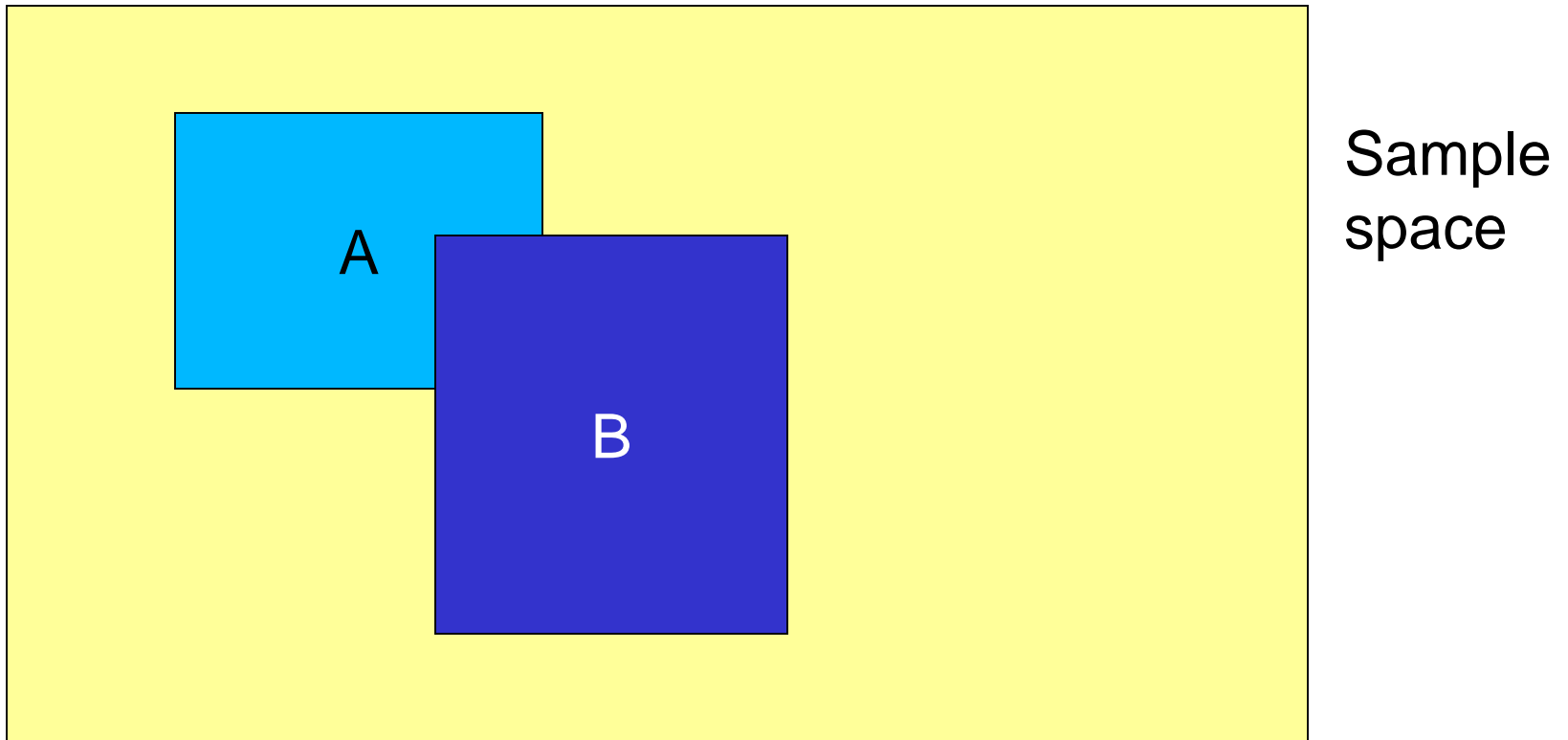
- $P(A) \in [0,1]$
- $P(\text{true})=1$, $P(\text{false})=0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

Sample
space

Invalid sentence:
e.g. “x=head AND x=tail”

The axioms of probability

- $P(A) \in [0,1]$
- $P(\text{true})=1, P(\text{false})=0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



Some theorems derived from the axioms

- $P(\neg A) = 1 - P(A)$ picture?
- If A can take k different values $a_1 \dots a_k$:
$$P(A=a_1) + \dots P(A=a_k) = 1$$
- $P(B) = P(B \wedge \neg A) + P(B \wedge A)$, if A is a binary event
- $P(B) = \sum_{i=1 \dots k} P(B \wedge A=a_i)$, if A can take k values

Joint probability

- The **joint** probability $P(A=a, B=b)$ is a shorthand for $P(A=a \wedge B=b)$, the probability of both $A=a$ and $B=b$ happen

$P(A=a)$, e.g. $P(1^{\text{st}} \text{ word on a random page} = \text{"San"}) = 0.001$
(possibly: San Francisco, San Diego, ...)

$P(B=b)$, e.g. $P(2^{\text{nd}} \text{ word} = \text{"Francisco"}) = 0.0008$
(possibly: San Francisco, Don Francisco, Pablo Francisco ...)

$P(A=a, B=b)$, e.g. $P(1^{\text{st}} = \text{"San"}, 2^{\text{nd}} = \text{"Francisco"}) = 0.0007$

Joint probability table

		weather		
temp		Sunny	Cloudy	Rainy
	hot	150/365	40/365	5/365
	cold	50/365	60/365	60/365

- $P(\text{temp}=\text{hot}, \text{weather}=\text{rainy}) = P(\text{hot}, \text{rainy}) = 5/365$
- The full joint probability table between N variables, each taking k values, has k^N entries (that's a lot!)

Marginal probability

- Sum over other variables

		weather		
		Sunny	Cloudy	Rainy
temp	hot	150/365	40/365	5/365
	cold	50/365	60/365	60/365
Σ		200/365	100/365	65/365

$$P(\text{Weather}) = \{200/365, 100/365, 65/365\}$$

- The name comes from the old days when the sums are written on the margin of a page

Marginal probability

- Sum over other variables

		weather			
temp		Sunny	Cloudy	Rainy	Σ
	hot	150/365	40/365	5/365	195/365
	cold	50/365	60/365	60/365	170/365

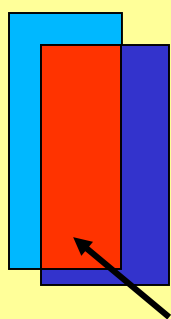
$$P(\text{temp}) = \{195/365, 170/365\}$$

- This is nothing but $P(B) = \sum_{i=1 \dots k} P(B \wedge A=a_i)$, if A can take k values

Conditional probability

- The **conditional** probability $P(A=a \mid B=b)$ is the fraction of times $A=a$, **within the region that** $B=b$

$P(A=a)$, e.g. $P(1^{\text{st}} \text{ word on a random page} = \text{"San"}) = 0.001$



$P(B=b)$, e.g. $P(2^{\text{nd}} \text{ word} = \text{"Francisco"}) = 0.0008$

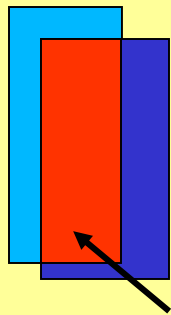
$P(A=a \mid B=b)$, e.g. $P(1^{\text{st}} = \text{"San"} \mid 2^{\text{nd}} = \text{"Francisco"}) = \mathbf{0.875}$
(possibly: San, Don, Pablo ...)

Although “San” is rare and “Francisco” is rare,
given “Francisco” then “San” is quite likely!

Conditional probability

- $P(\text{San} \mid \text{Francisco})$
= $\#(1^{\text{st}}=\text{S and } 2^{\text{nd}}=\text{F}) / \#(2^{\text{nd}}=\text{F})$
= $P(\text{San} \wedge \text{Francisco}) / P(\text{Francisco})$
= $0.0007 / 0.0008$
= 0.875

$P(S)=0.001$
$P(F)=0.0008$
$P(S,F)=0.0007$



$P(B=b)$, e.g. $P(2^{\text{nd}} \text{ word} = \text{"Francisco"}) = 0.0008$

$P(A=a \mid B=b)$, e.g. $P(1^{\text{st}}=\text{"San"} \mid 2^{\text{nd}}=\text{"Francisco"}) = \mathbf{0.875}$
(possibly: San, Don, Pablo ...)

Conditional probability

- In general, the conditional probability is

$$P(A = a | B) = \frac{P(A = a, B)}{P(B)} = \frac{P(A = a, B)}{\sum_{\text{all } a_i} P(A = a_i, B)}$$

- We can have everything conditioned on some other events C, to get a conditional version of conditional probability

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

‘|’ has low precedence.
This should read $P(A | (B, C))$

The chain rule

- From the definition of conditional probability we have the chain rule

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

- It works the other way around

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

- It works with more than 2 events too

$$P(A_1, A_2, \dots, A_n) =$$

$$P(A_1) * P(A_2 | A_1) * P(A_3 | A_1, A_2) * \dots * P(A_n | A_1, A_2, \dots, A_{n-1})$$

Reasoning

How do we use probabilities in AI?

- You wake up with a headache (D'oh!).
- Do you have the flu?
- H = headache, F = flu



Logical Inference: if (H) then F . (but the world is often not this clear cut)

Statistical Inference: compute the probability of a query given (conditioned on) evidence, i.e. $P(F|H)$

Inference with Bayes' rule: Example 1

Inference: compute the probability of a query given evidence
(H = headache, F = flu)

You know that

- $P(H) = 0.1$ “one in ten people has headache”
- $P(F) = 0.01$ “one in 100 people has flu”
- $P(H|F) = 0.9$ “90% of people who have flu have headache”
- How likely do you have the flu?
 - 0.9?
 - 0.01?
 - ...?

Bayes rule

Inference with Bayes' rule

Essay Towards Solving a Problem
in the Doctrine of Chances (1764)



$$P(F|H) = \frac{P(F, H)}{P(H)} = \frac{P(H|F)P(F)}{P(H)}$$

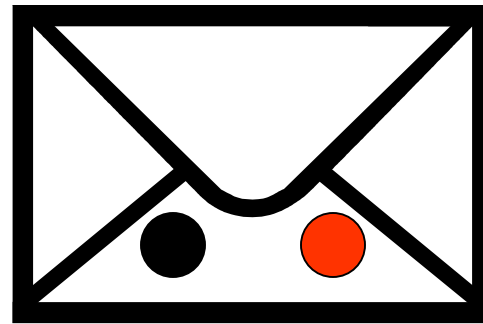
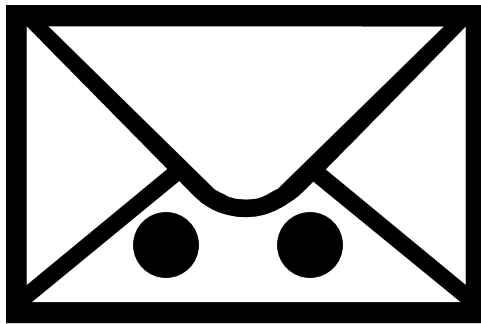
- $P(H) = 0.1$ “one in ten people has headache”
- $P(F) = 0.01$ “one in 100 people has flu”
- $P(H|F) = 0.9$ “90% of people who have flu have headache”
- $P(F|H) = 0.9 * 0.01 / 0.1 = 0.09$
- So there's a 9% chance you have flu – much less than 90%
- But it's higher than $P(F)=1\%$, since you have the headache

Inference with Bayes' rule

- $P(A|B) = P(B|A)P(A) / P(B)$ Bayes' rule
- Why do we make things this complicated?
 - Often $P(B|A)$, $P(A)$, $P(B)$ are easier to get
 - Some names:
 - **Prior $P(A)$** : probability before any evidence
 - **Likelihood $P(B|A)$** : assuming A, how likely is the evidence
 - **Posterior $P(A|B)$** : conditional prob. after knowing evidence
 - **Inference**: deriving unknown probability from known ones
- In general, if we have the full joint probability table, we can simply do $P(A|B)=P(A, B) / P(B)$ – more on this later...

Inference with Bayes' rule: Example 2

- In a bag there are two envelopes
 - one has a red ball (worth \$100) and a black ball
 - one has two black balls. Black balls worth nothing



- You randomly grabbed an envelope, randomly took out one ball – it's black.
- At this point you're given the option to switch the envelope. **To switch or not to switch?**

Inference with Bayes' rule: Example 2

- E: envelope, 1=(R,B), 2=(B,B)
- B: the event of drawing a black ball
- $P(E|B) = P(B|E) * P(E) / P(B)$
- We want to compare $P(E=1|B)$ vs. $P(E=2|B)$

Inference with Bayes' rule: Example 2

- E: envelope, 1=(R,B), 2=(B,B)
- B: the event of drawing a black ball
- $P(E|B) = P(B|E) * P(E) / P(B)$
- We want to compare $P(E=1|B)$ vs. $P(E=2|B)$
- $P(B|E=1) = 0.5$, $P(B|E=2) = 1$
- $P(E=1)=P(E=2)=0.5$
- $P(B)=3/4$ (it in fact doesn't matter for the comparison)

Inference with Bayes' rule: Example 2

- E: envelope, 1=(R,B), 2=(B,B)
- B: the event of drawing a black ball
- $P(E|B) = P(B|E) * P(E) / P(B)$
- We want to compare $P(E=1|B)$ vs. $P(E=2|B)$
- $P(B|E=1) = 0.5$, $P(B|E=2) = 1$
- $P(E=1)=P(E=2)=0.5$
- $P(B)=3/4$ (it in fact doesn't matter for the comparison)
- $P(E=1|B)=1/3$, $P(E=2|B)=2/3$
- After seeing a black ball, the posterior probability of this envelope being 1 (thus worth \$100) is smaller than it being 2
- Thus you should switch

Independence

- Two events A, B are **independent**, if (the following are equivalent)
 - $P(A, B) = P(A) * P(B)$
 - $P(A | B) = P(A)$
 - $P(B | A) = P(B)$
- For a 4-sided die, let
 - A=outcome is small
 - B=outcome is even
 - Are A and B independent?
- How about a 6-sided die?

Independence

- Independence is a domain knowledge
- If A, B are independent, the joint probability table between A, B is simple:
 - it has k^2 cells, but only $2k-2$ parameters. This is good news – more on this later...
- Example: $P(\text{burglary})=0.001$, $P(\text{earthquake})=0.002$. Let's say they are independent. The full joint probability table=?

Conditional independence

- Random variables can be dependent, but **conditionally independent**
- Your house has an alarm
 - Neighbor John will call when he hears the alarm
 - Neighbor Mary will call when she hears the alarm
 - Assume John and Mary don't talk to each other
- JohnCall independent of MaryCall?
 - No – If John called, likely the alarm went off, which increases the probability of Mary calling
 - $P(\text{MaryCall} \mid \text{JohnCall}) \neq P(\text{MaryCall})$

Conditional independence

- If we know the status of the alarm, JohnCall won't affect Mary at all

$$P(\text{MaryCall} \mid \text{Alarm}, \text{JohnCall}) = P(\text{MaryCall} \mid \text{Alarm})$$

- We say JohnCall and MaryCall are **conditionally independent**, given Alarm
- In general A, B are conditionally independent given C
 - if $P(A \mid B, C) = P(A \mid C)$, or
 - $P(B \mid A, C) = P(B \mid C)$, or
 - $P(A, B \mid C) = P(A \mid C) * P(B \mid C)$

Independence example #1

x, y	$P(X = x, Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

x	$P(X = x)$
sun	0.3
rain	0.5
snow	0.2
y	$P(Y = y)$
on-time	0.45
late	0.55

Are X and Y independent here?

Independence example #2

x, y	$P(X = x, Y = y)$
sun, fly-United	0.27
rain, fly-United	0.45
snow, fly-United	0.18
sun, fly-Delta	0.03
rain, fly-Delta	0.05
snow, fly-Delta	0.02

x	$P(X = x)$
sun	0.3
rain	0.5
snow	0.2
y	$P(Y = y)$
fly-United	0.9
fly-Delta	0.1

Are X and Y independent here?

Expected values

- The *expected value* of a random variable that takes on numerical values is defined as:

$$\mathbf{E}[X] = \sum_x xP(x)$$

This is the same thing as the *mean*

- We can also talk about the expected value of a function of a random variable

$$\mathbf{E}[g(X)] = \sum_x g(x)P(x)$$

Expected value examples

- Shoesize

$$\mathbf{E}[Shoesize]$$

$$= 5 \times P(Shoesize = 5) + \cdots + 14 \times P(Shoesize = 14)$$

- Suppose each lottery ticket costs \$1 and the winning ticket pays out \$100. The probability that a particular ticket is the winning ticket is 0.001.

What is the expectation of the gain?

Expected value examples

- Shoesize

$$E[\textit{Shoesize}]$$

$$= 5 \times P(\textit{Shoesize} = 5) + \cdots + 14 \times P(\textit{Shoesize} = 14)$$

- Suppose each lottery ticket costs \$1 and the winning ticket pays out \$100. The probability that a particular ticket is the winning ticket is 0.001.

$$\mathbf{E}[\textit{gain}(\textit{Lottery})]$$

$$= \textit{gain}(\textit{winning})P(\textit{winning}) + \textit{gain}(\textit{losing})P(\textit{losing})$$

$$= (\$100 - \$1) \times 0.001 - \$1 \times 0.999$$

$$= -\$0.9$$

Summary

- Axioms of probability and related properties
- Joint/marginal/conditional probabilities
- Bayes' rule for reasoning
- Independence and conditional independence
- Expectation