

CS 440/ECE 448 Lecture 12: Probability

Slides by Svetlana Lazebnik, 9/2016

Modified by Mark Hasegawa-Johnson, 10/2017



Outline

- Motivation: Why use probability?
 - Laziness, Ignorance, and Randomness
 - Rational Bettor Theorem
- Review of Key Concepts
 - Outcomes, Events
 - Random Variables; probability mass function (pmf)
 - Jointly random variables: Joint, Marginal, and Conditional pmf
 - Independent vs. Conditionally Independent events

Outline

- Motivation: Why use probability?
 - Laziness, Ignorance, and Randomness
 - Rational Bettor Theorem
- Review of Key Concepts
 - Outcomes, Events, and Random Variables
 - Joint, Marginal, and Conditional
 - Independence and Conditional Independence

Motivation: Planning under uncertainty

- Recall: representation for planning
- **States** are specified as conjunctions of predicates
 - Start state: $\text{At}(\text{P1}, \text{CMI}) \wedge \text{Plane}(\text{P1}) \wedge \text{Airport}(\text{CMI}) \wedge \text{Airport}(\text{ORD})$
 - Goal state: $\text{At}(\text{P1}, \text{ORD})$
- **Actions** are described in terms of preconditions and effects:
 - $\text{Fly}(\text{p}, \text{source}, \text{dest})$
 - **Precond:** $\text{At}(\text{p}, \text{source}) \wedge \text{Plane}(\text{p}) \wedge \text{Airport}(\text{source}) \wedge \text{Airport}(\text{dest})$
 - **Effect:** $\neg \text{At}(\text{p}, \text{source}) \wedge \text{At}(\text{p}, \text{dest})$

Motivation: Planning under uncertainty

- Let action $A_t = \text{leave for airport } t \text{ minutes before flight}$
 - Will A_t succeed, i.e., get me to the airport in time for the flight?
- Problems:
 - Partial observability (road state, other drivers' plans, etc.)
 - Noisy sensors (traffic reports)
 - Uncertainty in action outcomes (flat tire, etc.)
 - Complexity of modeling and predicting traffic
- Hence a purely logical approach either
 - Risks falsehood: “ A_{25} will get me there on time,” or
 - Leads to conclusions that are too weak for decision making:
 - A_{25} will get me there on time if there's no accident on the bridge and it doesn't rain and my tires remain intact, etc., etc.
 - A_{1440} will get me there on time but I'll have to stay overnight in the airport

Probability

Probabilistic assertions summarize effects of

- Laziness: reluctance to enumerate exceptions, qualifications, etc. --- possibly a deterministic and known environment, but with **computational complexity limitations**
- Ignorance: lack of explicit theories, relevant facts, initial conditions, etc. --- environment that is **unknown** (we don't know the transition function) or **partially observable** (we can't measure the current state)
- Intrinsically random phenomena – environment is **stochastic**, i.e., given a particular (action,current state), the (next state) is drawn at random with a particular probability distribution

Outline

- Motivation: Why use probability?
 - Laziness, Ignorance, and Randomness
 - Rational Bettor Theorem
- Review of Key Concepts
 - Outcomes, Events, and Random Variables
 - Joint, Marginal, and Conditional
 - Independence and Conditional Independence

Making decisions under uncertainty

- Suppose the agent believes the following:
 - $P(A_{25} \text{ gets me there on time}) = 0.04$
 - $P(A_{90} \text{ gets me there on time}) = 0.70$
 - $P(A_{120} \text{ gets me there on time}) = 0.95$
 - $P(A_{1440} \text{ gets me there on time}) = 0.9999$
- Which action should the agent choose?
 - Depends on preferences for missing flight vs. time spent waiting
 - Encapsulated by a *utility function*
- The agent should choose the action that maximizes the *expected utility*:
 - $P(A_t \text{ succeeds}) * U(A_t \text{ succeeds}) + P(A_t \text{ fails}) * U(A_t \text{ fails})$

Making decisions under uncertainty

- More generally: the expected utility of an action is defined as:

$$EU(\text{action}) = \sum_{\text{outcomes of action}} P(\text{outcome} \mid \text{action}) U(\text{outcome})$$

- **Utility theory** is used to represent and infer preferences
- **Decision theory** = probability theory + utility theory

Where do probabilities come from?

- **Frequentism**

- Probabilities are relative frequencies
- For example, if we toss a coin many times, $P(\text{heads})$ is the proportion of the time the coin will come up heads
- But what if we're dealing with events that only happen once?
 - E.g., what is the probability that Team X will win the Superbowl this year?
 - "Reference class" problem

- **Subjectivism**

- Probabilities are degrees of belief
- But then, how do we assign belief values to statements?
- What would constrain agents to hold consistent beliefs?

The Rational Bettor Theorem

- Why should a rational agent hold beliefs that are consistent with axioms of probability?
 - For example, $P(A) + P(\neg A) = 1$
- Suppose an agent believes that $P(A)=0.7$, and $P(\neg A)=0.7$
- Offer the following bet: if A occurs, agent wins \$100. If A doesn't occur, agent loses \$105. Agent believes $P(A) > 100/(100+105)$, so agent accepts the bet.
- Offer another bet: if $\neg A$ occurs, agent wins \$100. If $\neg A$ doesn't occur, agent loses \$105. Agent believes $P(\neg A) > 100/(100+105)$, so agent accepts the bet. **Oops...**
- **Theorem**: An agent who holds beliefs inconsistent with axioms of probability can be convinced to accept a combination of bets that is guaranteed to lose them money

Are humans “rational bettors”?

- Humans are pretty good at estimating some probabilities, and pretty bad at estimating others. What might cause humans to mis-estimate the probability of an event?
 - Large-scale nonlinearities upset weather forecasts
 - System might change over time
 - Large number of options
 - People over-estimate probability of very rare events
 - Under-estimate (or over-estimate?) probability of very bad events
 - Over-estimate how much we know about the situation
 - Emotional attachment to any particular outcome
 - Over-estimate probability of highly visible/much talked-of events
- What are some of the ways in which a “rational bettor” might take advantage of humans who mis-estimate probabilities?
 - Insurance sales: over-estimate probability of bad events
 - War & politics
 - Stock market

Outline

- Motivation: Why use probability?
 - Laziness, Ignorance, and Randomness
 - Rational Bettor Theorem
- Review of Key Concepts
 - Outcomes, Events, and Random Variables
 - Joint, Marginal, and Conditional
 - Independence and Conditional Independence

Outcomes of an Experiment

The SET OF POSSIBLE OUTCOMES (a.k.a. the “sample space”) is a listing of all of the things that might happen:

1. Mutually exclusive. It’s not possible that two different outcomes might both happen.
2. Collectively exhaustive. Every outcome that could possibly happen is one of the items in the list.
3. Finest grain. After the experiment occurs, somebody tells you the outcome, and there is nothing else you need to know.

Example experiment: Alice, Bob, Carol and Duane run a 10km race to decide who will buy pizza tonight.

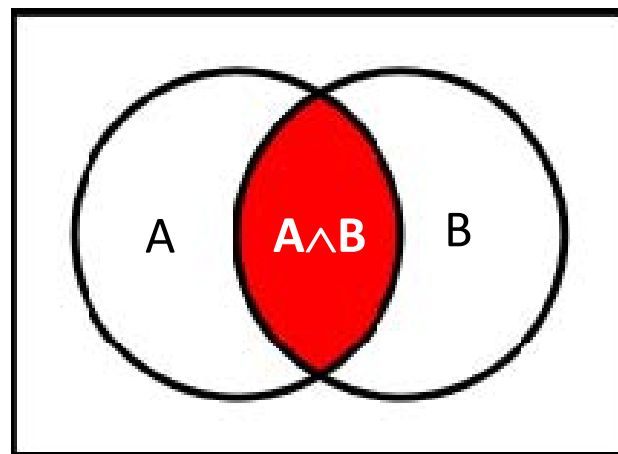
Outcome = a listing of the exact finishing times of each participant.

Events

- Probabilistic statements are defined over *events*, or sets of world states
 - $A = \text{"It is raining"}$
 - $B = \text{"The weather is either cloudy or snowy"}$
 - $C = \text{"The sum of the two dice rolls is 11"}$
 - $D = \text{"My car is going between 30 and 50 miles per hour"}$
- An EVENT is a SET of OUTCOMES
 - $B = \{ \text{outcomes : cloudy OR snowy} \}$
 - $C = \{ \text{outcomes : } d1+d2 = 11 \}$
- Notation: $p(A)$ or $P(A)$ is the probability of the set of world states (outcomes) in which proposition A holds

Kolmogorov's axioms of probability

- For any propositions (events) A, B
 - $0 \leq P(A) \leq 1$
 - $P(\text{True}) = 1$ and $P(\text{False}) = 0$
 - $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
 - Subtraction accounts for double-counting
- Based on these axioms, what is $P(\neg A)$?
- These axioms are sufficient to completely specify probability theory for *discrete* random variables
 - For continuous variables, need *density functions*



Outcomes = Atomic events

- **OUTCOME or ATOMIC EVENT:** is a complete specification of the state of the world, or a complete assignment of domain values to all random variables
 - Atomic events are mutually exclusive and exhaustive
- E.g., if the world consists of only two Boolean variables *Cavity* and *Toothache*, then there are four outcomes:
 - $\neg Cavity \wedge \neg Toothache$
 - $\neg Cavity \wedge Toothache$
 - $Cavity \wedge \neg Toothache$
 - $Cavity \wedge Toothache$

Random variables

- We describe the (uncertain) state of the world using **random variables**
 - Denoted by capital letters
 - **R**: *Is it raining?*
 - **W**: *What's the weather?*
 - **D**: *What is the outcome of rolling two dice?*
 - **S**: *What is the speed of my car (in MPH)?*
- Just like variables in CSPs, random variables take on values in a *domain*
 - Domain values must be *mutually exclusive* and *exhaustive*
 - **R** in {True, False}
 - **W** in {Sunny, Cloudy, Rainy, Snow}
 - **D** in {(1,1), (1,2), ... (6,6)}
 - **S** in [0, 200]

Random variables

- A random variable can be viewed as a function that maps from outcomes to real numbers (or integers, or strings)
- For example: the event “Speed=45mph” is the set of all outcomes for which the speed of my car is 45mph

Probability Mass Function (pmf)

- We use a capital letter for a random variables (RV=the function that maps from outcomes to values), and a small letters for the actual value that it takes after any particular experiment.
- $X_1 = x_1$ is the event “random variable X_1 takes the value x_1 ”
- $p(X_1 = x_1)$ is a **number**: the probability that this event occurs.
 - We call this number the “probability mass” of the event $X_1 = x_1$
 - The function is called the “probability mass function” or pmf
 - Shorthand: $p(x_1)$ using a small letter x_1
 - Subscript notation, which we won’t use in this class: $p_{X_1}(x_1)$
- $p(X_1)$ using a capital letter X_1 is a **function**: the entire table of the probabilities $X_1 = x_1$ for every possible x_1

Events and Outcomes

- An OUTCOME (ATOMIC EVENT) is a particular setting of all of the random variables
 - *Outcome = (die 1 shows 5 dots, die 2 shows 6 dots)*
- An EVENT is a SET of OUTCOMES
 - *“The sum of the two dice rolls is 11” = { set of all outcomes such that $D1+D2 = 11$ }*
 - *“D1=5” = {set of all outcomes such that D1=5, regardless of what D2 is }*
- $P(EVENT) = \sum_{outcomes \in EVENT} P(outcome)$

Functions of Random Variables

- Suppose we are not really interested in any given random variable, instead we're only interested in a function of the random variables
- Example: the game of craps. We're only interested in the sum of the two dice, e.g., what is the probability that the sum of the two dice is greater than 10.
- Define $S=D1+D2$. How can we calculate the pmf for S ?

Outline

- Motivation: Why use probability?
 - Laziness, Ignorance, and Randomness
 - Rational Bettor Theorem
- Review of Key Concepts
 - Outcomes, Events, and Random Variables
 - Joint, Marginal, and Conditional
 - Independence and Conditional Independence

Joint probability distributions

- A **joint distribution** is an assignment of probabilities to every possible atomic event

Atomic event	P
$\neg \text{Cavity} \wedge \neg \text{Toothache}$	0.8
$\neg \text{Cavity} \wedge \text{Toothache}$	0.1
$\text{Cavity} \wedge \neg \text{Toothache}$	0.05
$\text{Cavity} \wedge \text{Toothache}$	0.05

- Why does it follow from the axioms of probability that the probabilities of all possible atomic events must sum to 1?

Joint probability distributions

- A **joint distribution** is an assignment of probabilities to every possible atomic event
- Suppose we have a joint distribution of N random variables, each of which takes values from a domain of size D
 - What is the size of the probability table?
 - Impossible to write out completely for all but the smallest distributions

Notation

- $p(X_1 = x_1, X_2 = x_2, \dots, X_N = x_N)$ refers to a single entry (atomic event) in the joint probability distribution table
 - Shorthand: $p(x_1, x_2, \dots, x_N)$
 - Subscript notation, which we won't use in this class:
 $p_{X_1, X_2, \dots, X_N}(x_1, x_2, \dots, x_N)$
- $p(X_1, X_2, \dots, X_N)$ refers to the entire joint probability distribution table
- $P(A)$ can also refer to the probability of an event
 - E.g., $X_1 = x_1$ is an event

Marginal probability distributions

- From the joint distribution $p(X,Y)$ we can find the **marginal distributions** $p(X)$ and $p(Y)$

P(Cavity, Toothache)	
$\neg Cavity \wedge \neg Toothache$	0.8
$\neg Cavity \wedge Toothache$	0.1
$Cavity \wedge \neg Toothache$	0.05
$Cavity \wedge Toothache$	0.05

P(Cavity)	
$\neg Cavity$?
$Cavity$?

P(Toothache)	
$\neg Toothache$?
$Toothache$?

Joint -> Marginal by adding the outcomes

- From the joint distribution $p(X,Y)$ we can find the **marginal distributions** $p(X)$ and $p(Y)$
- To find $p(X = x)$, sum the probabilities of all atomic events where $X = x$:

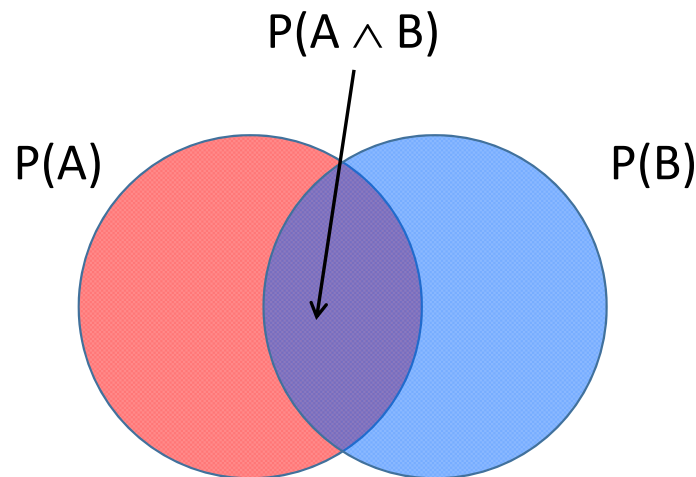
$$P(X = 1) = P(X = 1, Y = 1) + P(X = 1, Y = 2) + P(X = 1, Y = 3) + \dots$$

- This is called **marginalization** (we are *marginalizing out* all the variables except X)

Conditional Probability: renormalize (divide)

- Probability of cavity given toothache:
 $P(\text{Cavity} = \text{true} \mid \text{Toothache} = \text{true})$

- For any two events A and B,
$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)} = \frac{P(A, B)}{P(B)}$$



Conditional probability

P(Cavity, Toothache)	
$\neg\text{Cavity} \wedge \neg\text{Toothache}$	0.8
$\neg\text{Cavity} \wedge \text{Toothache}$	0.1
$\text{Cavity} \wedge \neg\text{Toothache}$	0.05
$\text{Cavity} \wedge \text{Toothache}$	0.05

P(Cavity)	
$\neg\text{Cavity}$	0.9
Cavity	0.1

P(Toothache)	
$\neg\text{Toothache}$	0.85
Toothache	0.15

- What is $p(\text{Cavity} = \text{true} \mid \text{Toothache} = \text{false})$?
 $p(\text{Cavity} \mid \neg\text{Toothache}) = ?$
- What is $p(\text{Cavity} = \text{false} \mid \text{Toothache} = \text{true})$?
 $p(\neg\text{Cavity} \mid \text{Toothache}) = ?$

Conditional distributions

- A conditional distribution is a distribution over the values of one variable given fixed values of other variables

P(Cavity, Toothache)	
$\neg \text{Cavity} \wedge \neg \text{Toothache}$	0.8
$\neg \text{Cavity} \wedge \text{Toothache}$	0.1
$\text{Cavity} \wedge \neg \text{Toothache}$	0.05
$\text{Cavity} \wedge \text{Toothache}$	0.05

P(Cavity Toothache = true)	
$\neg \text{Cavity}$	0.667
Cavity	0.333

P(Cavity Toothache = false)	
$\neg \text{Cavity}$	0.941
Cavity	0.059

P(Toothache Cavity = true)	
$\neg \text{Toothache}$	0.5
Toothache	0.5

P(Toothache Cavity = false)	
$\neg \text{Toothache}$	0.889
Toothache	0.111

Normalization trick

- To get the whole conditional distribution $p(X \mid Y = y)$ at once, select all entries in the joint distribution table matching $Y = y$ and renormalize them to sum to one

P(Cavity, Toothache)	
$\neg \text{Cavity} \wedge \neg \text{Toothache}$	0.8
$\neg \text{Cavity} \wedge \text{Toothache}$	0.1
$\text{Cavity} \wedge \neg \text{Toothache}$	0.05
$\text{Cavity} \wedge \text{Toothache}$	0.05



Select

Toothache, Cavity = false	
$\neg \text{Toothache}$	0.8
Toothache	0.1



Renormalize

P(Toothache Cavity = false)	
$\neg \text{Toothache}$	0.889
Toothache	0.111

Normalization trick

- To get the whole conditional distribution $p(X \mid Y = y)$ at once, select all entries in the joint distribution table matching $Y = y$ and renormalize them to sum to one
- Why does it work?

$$\frac{P(x, y)}{\sum_{x'} P(x', y)} = \frac{P(x, y)}{P(y)} \quad \text{by marginalization}$$

Product rule

- Definition of conditional probability: $P(A | B) = \frac{P(A, B)}{P(B)}$
- Sometimes we have the conditional probability and want to obtain the joint:

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

Product rule

- Definition of conditional probability: $P(A | B) = \frac{P(A, B)}{P(B)}$
- Sometimes we have the conditional probability and want to obtain the joint:

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

- The chain rule:

$$\begin{aligned} P(A_1, K, A_n) &= P(A_1)P(A_2 | A_1)P(A_3 | A_1, A_2)K P(A_n | A_1, K, A_{n-1}) \\ &= \prod_{i=1}^n P(A_i | A_1, K, A_{i-1}) \end{aligned}$$

Product Rule Example: The Birthday problem

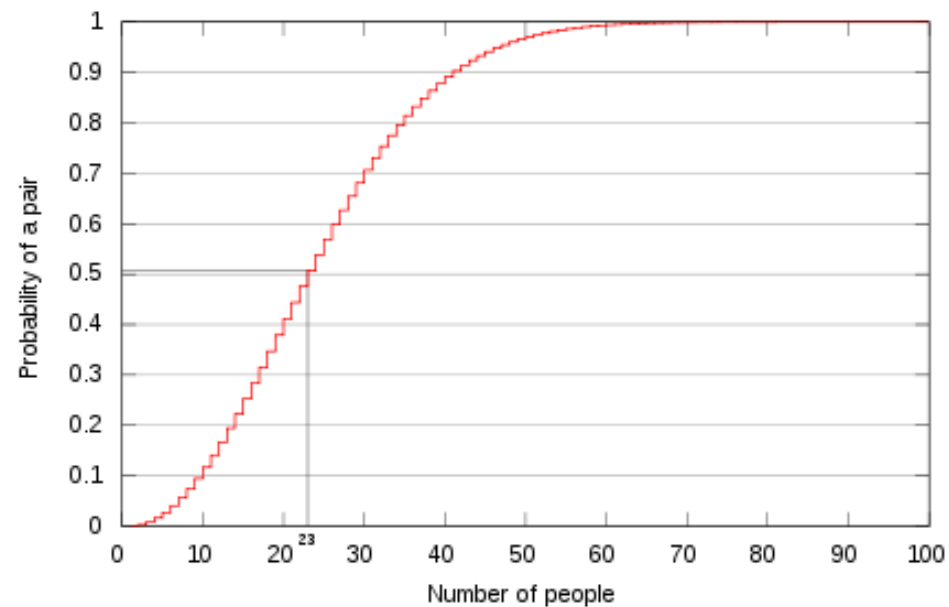
- We have a set of n people. What is the probability that two of them share the same birthday?
- Easier to calculate the probability that n people *do not* share the same birthday

$$\begin{aligned} &P(B_1, \dots, B_n \text{ distinct}) \\ &= P(B_1, B_2 \text{ distinct})P(B_1, B_2, B_3 \text{ distinct} | B_1, B_2 \text{ distinct}) \dots \\ &\quad P(B_1, B_2, \dots, B_n \text{ distinct} | B_1, \dots, B_{n-1} \text{ distinct}) \end{aligned}$$

$$P(B_1, \dots, B_n \text{ distinct}) = \left(\frac{364}{365}\right) \left(\frac{363}{364}\right) \dots \left(\frac{365-n+2}{365-n+1}\right)$$

The Birthday problem

- For 23 people, the probability of sharing a birthday is above 0.5!



http://en.wikipedia.org/wiki/Birthday_problem

Outline

- Motivation: Why use probability?
 - Laziness, Ignorance, and Randomness
 - Rational Bettor Theorem
- Review of Key Concepts
 - Outcomes, Events, and Random Variables
 - Joint, Marginal, and Conditional
 - Independence and Conditional Independence

Independence \neq Mutually Exclusive

- Two events A and B are *independent* if and only if
$$p(A \wedge B) = p(A, B) = p(A) p(B)$$
 - In other words, $p(A | B) = p(A)$ and $p(B | A) = p(B)$
 - This is an important simplifying assumption for modeling, e.g., *Toothache* and *Weather* can be assumed to be independent?
- Are two *mutually exclusive* events independent?
 - No! Quite the opposite! If you know A happened, then you know that B _didn't_ happen!!
$$p(A \vee B) = p(A) + p(B)$$

Independence \neq Conditional Independence

- Two events A and B are *independent* if and only if
$$p(A \wedge B) = p(A) p(B)$$
 - In other words, $p(A | B) = p(A)$ and $p(B | A) = p(B)$
 - This is an important simplifying assumption for modeling, e.g., *Toothache* and *Weather* can be assumed to be independent
- **Conditional independence:** A and B are *conditionally independent* given C iff
$$p(A \wedge B | C) = p(A | C) p(B | C)$$
 - Equivalent:
$$p(A | B, C) = p(A | C)$$
 - Equivalent:
$$p(B | A, C) = p(B | C)$$

Conditional independence: Example

- *Toothache*: boolean variable indicating whether the patient has a toothache
- *Cavity*: boolean variable indicating whether the patient has a cavity
- *Catch*: whether the dentist's probe catches in the cavity
- If the patient has a cavity, the probability that the probe catches in it doesn't depend on whether he/she has a toothache
$$p(\text{Catch} | \text{Toothache}, \text{Cavity}) = p(\text{Catch} | \text{Cavity})$$
- Therefore, *Catch* is conditionally independent of *Toothache* given *Cavity*
- Likewise, *Toothache* is conditionally independent of *Catch* given *Cavity*
$$p(\text{Toothache} | \text{Catch}, \text{Cavity}) = p(\text{Toothache} | \text{Cavity})$$
- Equivalent statement:
$$p(\text{Toothache}, \text{Catch} | \text{Cavity}) = p(\text{Toothache} | \text{Cavity}) p(\text{Catch} | \text{Cavity})$$

Random Audience Participation Slide

- List some pairs of events that are independent
 - ... here is a pair of events
- List some pairs of events that are mutually exclusive
 - here is some different pair of events
- List some pairs of events that are conditionally independent given knowledge of some third event
 - ... whoa, now we need event triples. ...