Refresher on Discrete Probability

STAT 27725/CMSC 25400: Machine Learning

Shubhendu Trivedi

University of Chicago

October 2015

Background

- Things you should have seen before
  - Events, Event Spaces
  - Probability as limit of frequency
  - Compound Events
  - Joint and Conditional Probability
  - Random Variables
  - Expectation, variance and covariance
  - Independence and Conditional Independence
  - Estimation

Background

- Things you should have seen before
  - Events, Event Spaces
  - Probability as limit of frequency
  - Compound Events
  - Joint and Conditional Probability
  - Random Variables
  - Expectation, variance and covariance
  - Independence and Conditional Independence
  - Estimation
- This refresher WILL revise these topics.

Three types of Probability

- **Frequency of repeated trials:** if an experiment is repeated infinitely many times, $0 \leq p(A) \leq 1$ is the fraction of times that the outcome will be A.

Three types of Probability

- **Frequency of repeated trials:** if an experiment is repeated infinitely many times, $0 \leq p(A) \leq 1$ is the fraction of times that the outcome will be A. Typical example: number of times that a coin comes up heads.

Three types of Probability

- **Frequency of repeated trials:** if an experiment is repeated infinitely many times, $0 \leq p(A) \leq 1$ is the fraction of times that the outcome will be A. Typical example: number of times that a coin comes up heads. Frequentist probability.

Three types of Probability

- **Frequency of repeated trials:** if an experiment is repeated infinitely many times, $0 \leq p(A) \leq 1$ is the fraction of times that the outcome will be A. Typical example: number of times that a coin comes up heads. Frequentist probability.
- **Degree of belief:** A quantity obeying the same laws as the above, describing how likely we think a (possibly deterministic) event is.

Three types of Probability

- **Frequency of repeated trials:** if an experiment is repeated infinitely many times, $0 \leq p(A) \leq 1$ is the fraction of times that the outcome will be A. Typical example: number of times that a coin comes up heads. Frequentist probability.
- **Degree of belief:** A quantity obeying the same laws as the above, describing how likely we think a (possibly deterministic) event is. Typical example: the probability that the Earth will warmer by more than $5°F$ by 2100.

Three types of Probability

- **Frequency of repeated trials:** if an experiment is repeated infinitely many times, $0 \leq p(A) \leq 1$ is the fraction of times that the outcome will be A. Typical example: number of times that a coin comes up heads. Frequentist probability.
- **Degree of belief:** A quantity obeying the same laws as the above, describing how likely we think a (possibly deterministic) event is. Typical example: the probability that the Earth will warmer by more than $5°F$ by 2100. Bayesian probability.

Three types of Probability

- **Frequency of repeated trials:** if an experiment is repeated infinitely many times, $0 \leq p(A) \leq 1$ is the fraction of times that the outcome will be A. Typical example: number of times that a coin comes up heads. Frequentist probability.
- **Degree of belief:** A quantity obeying the same laws as the above, describing how likely we think a (possibly deterministic) event is. Typical example: the probability that the Earth will warmer by more than $5°F$ by 2100. Bayesian probability.
- **Subjective probability:** "I'm 110% sure that I'll go out to dinner with you tonight."

Three types of Probability

- **Frequency of repeated trials:** if an experiment is repeated infinitely many times, $0 \leq p(A) \leq 1$ is the fraction of times that the outcome will be A. Typical example: number of times that a coin comes up heads. Frequentist probability.
- **Degree of belief:** A quantity obeying the same laws as the above, describing how likely we think a (possibly deterministic) event is. Typical example: the probability that the Earth will warmer by more than $5^\circ F$ by 2100. Bayesian probability.
- **Subjective probability:** "I'm 110% sure that I'll go out to dinner with you tonight."

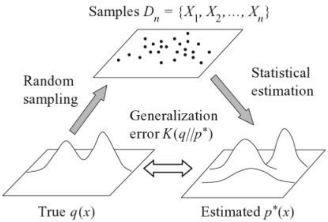
Mixing these three notions is a source of lots of trouble.

Three types of Probability

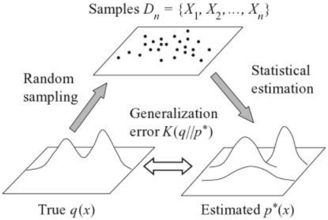
- **Frequency of repeated trials:** if an experiment is repeated infinitely many times, $0 \leq p(A) \leq 1$ is the fraction of times that the outcome will be A. Typical example: number of times that a coin comes up heads. Frequentist probability.
- **Degree of belief:** A quantity obeying the same laws as the above, describing how likely we think a (possibly deterministic) event is. Typical example: the probability that the Earth will warmer by more than $5^\circ F$ by 2100. Bayesian probability.
- **Subjective probability:** "I'm 110% sure that I'll go out to dinner with you tonight."

Mixing these three notions is a source of lots of trouble. We will start with the frequentist interpretation and then discuss the Bayesian one.

Why do we need Probability in Machine Learning

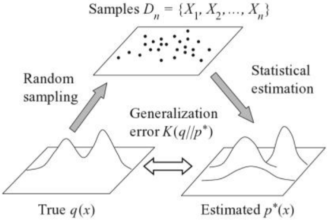


Why do we need Probability in Machine Learning



- To analyze, understand and predict the performance of learning algorithms (Vapnik Chervonenkis Theory, PAC model, etc.)

Why do we need Probability in Machine Learning



- To analyze, understand and predict the performance of learning algorithms (Vapnik Chervonenkis Theory, PAC model, etc.)
- To build flexible and intuitive **probabilistic models**.

Basic Notions

Sample space

- Random Experiment: An experiment whose outcome cannot be determined in advance, but is nonetheless subject to analysis
  1. Tossing a coin
  2. Selecting a group of 100 people and observing the number of left handers

Sample space

- Random Experiment: An experiment whose outcome cannot be determined in advance, but is nonetheless subject to analysis
  1. Tossing a coin
  2. Selecting a group of 100 people and observing the number of left handers
- There are three main ingredients in the model of a random experiment

Sample space

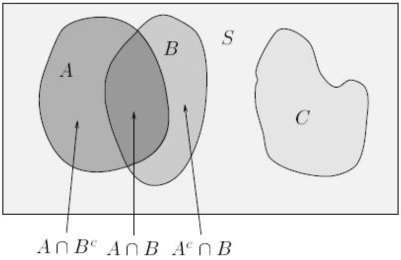
- Random Experiment: An experiment whose outcome cannot be determined in advance, but is nonetheless subject to analysis
  1. Tossing a coin
  2. Selecting a group of 100 people and observing the number of left handers
- There are three main ingredients in the model of a random experiment
- We can't predict the outcome of a random experiment with certainty, but can specify a set of possible outcomes

Sample space

- Random Experiment: An experiment whose outcome cannot be determined in advance, but is nonetheless subject to analysis
  1. Tossing a coin
  2. Selecting a group of 100 people and observing the number of left handers
- There are three main ingredients in the model of a random experiment
- We can't predict the outcome of a random experiment with certainty, but can specify a set of possible outcomes
- **Sample Space:** The sample space Ω of a random experiment is the set of all possible outcomes of the experiment
  1. {H, T}
  2. {1, 2, ..., 100 }

Events

- We are often not interested in a single outcome, but in whether or not one of a *group* of outcomes occurs.
- Such subsets of the sample space are called **events**
- Events are sets, can apply the usual set operations to them:
  1. $A \cup B$: Event that A or B or both occur
  2. $A \cap B$: Event that A and B both occur
  3. A^c: Event that A does not occur
  4. $A \subset B$: event A will imply event B
  5. $A \cap B = \emptyset$: Disjoint events.



Axioms of Probability

- The third ingredient in the model for a random experiment is the specification of the probability of events

Axioms of Probability

- The third ingredient in the model for a random experiment is the specification of the probability of events
- The probability of some event A, denoted by $\mathbb{P}(A)$, is defined such that $\mathbb{P}(A)$ satisfies the following axioms
  1. $\mathbb{P}(A) \geq 0$
  2. $\mathbb{P}(\Omega) = 1$
  3. For any sequence $A_1, A_2, \ldots$ of disjoint events we have:

$$\mathbb{P}\big(\cup_i A_i\big) = \sum_i \mathbb{P}(A_i)$$

Axioms of Probability

- The third ingredient in the model for a random experiment is the specification of the probability of events
- The probability of some event A, denoted by $\mathbb{P}(A)$, is defined such that $\mathbb{P}(A)$ satisfies the following axioms
  1. $\mathbb{P}(A) \geq 0$
  2. $\mathbb{P}(\Omega) = 1$
  3. For any sequence $A_1, A_2, \dots$ of disjoint events we have:

$$\mathbb{P}\big(\cup_i A_i\big) = \sum_i \mathbb{P}(A_i)$$

- Kolmogorov showed that these three axioms lead to the rules of probability theory
- de Finetti, Cox and Carnap have also provided compelling reasons for these axioms

Some Consequences

- Probability of the Empty set: $\mathbb{P}(\emptyset) = 0$

Some Consequences

- Probability of the Empty set: $\mathbb{P}(\emptyset) = 0$
- Monotonicity: if $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$

Some Consequences

- Probability of the Empty set: $\mathbb{P}(\emptyset) = 0$
- Monotonicity: if $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$
- Numeric Bound: $0 \leq \mathbb{P}(A) \leq 1 \; \forall A \in S$

Some Consequences

- Probability of the Empty set: $\mathbb{P}(\emptyset) = 0$
- Monotonicity: if $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$
- Numeric Bound: $0 \leq \mathbb{P}(A) \leq 1 \; \forall A \in S$
- Addition Law: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
- $\mathbb{P}(A^c) = \mathbb{P}(S \setminus A) = 1 - \mathbb{P}(A)$

Some Consequences

- Probability of the Empty set: $\mathbb{P}(\emptyset) = 0$
- Monotonicity: if $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$
- Numeric Bound: $0 \leq \mathbb{P}(A) \leq 1 \ \forall A \in S$
- Addition Law: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
- $\mathbb{P}(A^c) = \mathbb{P}(S \setminus A) = 1 - \mathbb{P}(A)$
- Axioms of probability are the only system with this property: If you gamble using them you can't be be unfairly exploited by an opponent using some other system (di Finetti, 1931)

Discrete Sample Spaces

- For now, we focus on the case when the sample space is countable $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$

Discrete Sample Spaces

- For now, we focus on the case when the sample space is countable $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$
- The probability $\mathbb{P}$ on a discrete sample space can be specified by first specifying the probability p_i of each **elementary event** ω_i and then defining:

Discrete Sample Spaces

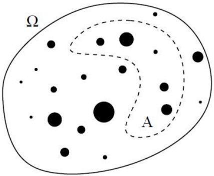
- For now, we focus on the case when the sample space is countable $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$
- The probability $\mathbb{P}$ on a discrete sample space can be specified by first specifying the probability p_i of each **elementary event** ω_i and then defining:

$$\mathbb{P}(A) = \sum_{i:\omega_i \in A} p_i \ \forall A \subset \Omega$$

Discrete Sample Spaces

- For now, we focus on the case when the sample space is countable $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$
- The probability $\mathbb{P}$ on a discrete sample space can be specified by first specifying the probability p_i of each **elementary event** ω_i and then defining:

$$\mathbb{P}(A) = \sum_{i:\omega_i \in A} p_i \ \forall A \subset \Omega$$



Discrete Sample Spaces

$$\mathbb{P}(A) = \sum_{i:\omega_i \in A} p_i \ \forall A \subset \Omega$$

Discrete Sample Spaces

$$\mathbb{P}(A) = \sum_{i:\omega_i \in A} p_i \ \forall A \subset \Omega$$

- In many applications, each elementary event is equally likely.
- Probability of an elementary event: 1 divided by total number of elements in Ω
- **Equally likely principle:**If Ω has a finite number of outcomes, and all ar equally likely, then the possibility of each event A is defined as

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}$$

Discrete Sample Spaces

$$\mathbb{P}(A) = \sum_{i:\omega_i \in A} p_i \ \forall A \subset \Omega$$

- In many applications, each elementary event is equally likely.
- Probability of an elementary event: 1 divided by total number of elements in Ω
- **Equally likely principle:** If Ω has a finite number of outcomes, and all ar equally likely, then the possibility of each event A is defined as

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}$$

- Finding $\mathbb{P}(A)$ reduces to counting
- What is the probability of getting a full house in poker?

Discrete Sample Spaces

$$\mathbb{P}(A) = \sum_{i:\omega_i \in A} p_i \ \forall A \subset \Omega$$

- In many applications, each elementary event is equally likely.
- Probability of an elementary event: 1 divided by total number of elements in Ω
- **Equally likely principle:**If Ω has a finite number of outcomes, and all ar equally likely, then the possibility of each event A is defined as

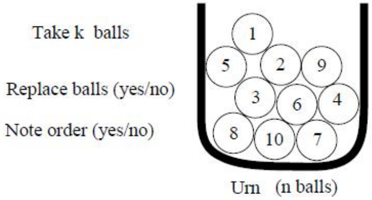
$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}$$

- Finding $\mathbb{P}(A)$ reduces to counting
- What is the probability of getting a full house in poker?

$$\frac{13\binom{4}{3} \cdot 12\binom{4}{2}}{\binom{52}{5}} \approx 0.14$$

Counting

- Counting is not easy! Fortunately, many counting problems can be cast into the framework of drawing balls from an urn



	with replacement	without replacement
ordered		
not ordered		

Choosing k of n distinguishable objects

	with replacement	without replacement
ordered	n^k	$n(n-1)\ldots(n-k+1)$
not ordered	$\binom{n+k-1}{n-1}$	$\binom{n}{k}$

Choosing k of n distinguishable objects

	with replacement	without replacement
ordered	n^k	$n(n-1)\ldots(n-k+1)$
not ordered	$\binom{n+k-1}{n-1}$	$\binom{n}{k}$

$\longrightarrow$ usually goes in the denominator

Indistinguishable Objects

If we choose k balls from an urn with n_1 red balls and n_2 green balls, what is the probability of getting a particular sequence of x red balls and $k - x$ green ones?
What is the probability of any such sequence? How many ways can this happen? (this goes in the numerator)

Indistinguishable Objects

If we choose k balls from an urn with n_1 red balls and n_2 green balls, what is the probability of getting a particular sequence of x red balls and $k - x$ green ones?
What is the probability of any such sequence? How many ways can this happen? (this goes in the numerator)

	with replacement	without replacement
ordered	$n_1^x n_2^{k-x}$	$n_1 \ldots (n_1 - x + 1) \cdot n_2 \ldots (n_2 - k + x$
not ordered	$\binom{k}{x} n_1^x n_2^{k-x}$	$k! \binom{n_1}{x} \binom{n_2}{k-x}$

Joint and conditional probability

Joint:

$$\mathbb{P}(A, B) = \mathbb{P}(A \cap B)$$

Conditional:

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

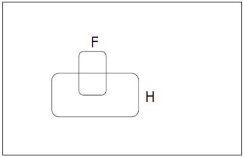
AI is all about conditional probabilities.

Conditional Probability

- $\mathbb{P}(A|B) =$ fraction of worlds in which B is true that also have A true

Conditional Probability

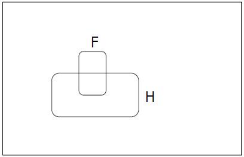
- $\mathbb{P}(A|B) =$ fraction of worlds in which B is true that also have A true



- $H =$ "Have a headache", $F =$ "Have flu".

Conditional Probability

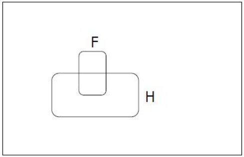
- $\mathbb{P}(A|B)$ = fraction of worlds in which B is true that also have A true



- H = "Have a headache", F = "Have flu".
- $\mathbb{P}(H) = \frac{1}{10}, \mathbb{P}(F) = \frac{1}{40}, \mathbb{P}(H|F) = \frac{1}{2}$

Conditional Probability

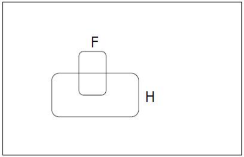
- $\mathbb{P}(A|B) =$ fraction of worlds in which B is true that also have A true



- $H =$ "Have a headache", $F =$ "Have flu".
- $\mathbb{P}(H) = \frac{1}{10}, \mathbb{P}(F) = \frac{1}{40}, \mathbb{P}(H|F) = \frac{1}{2}$
- "Headaches are rare and flu is rarer, but if you are coming down wih flu, there is a 50-50 chance you'll have a headache."

Conditional Probability

- $\mathbb{P}(A|B)$ = fraction of worlds in which B is true that also have A true



- H = "Have a headache", F = "Have flu".
- $\mathbb{P}(H) = \frac{1}{10}, \mathbb{P}(F) = \frac{1}{40}, \mathbb{P}(H|F) = \frac{1}{2}$
- "Headaches are rare and flu is rarer, but if you are coming down wih flu, there is a 50-50 chance you'll have a headache."

Conditional Probability

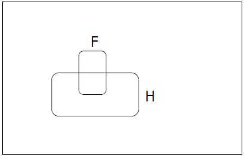
- $\mathbb{P}(H|F)$: Fraction of flu-inflicted worlds in which you have a headache

Conditional Probability

- $\mathbb{P}(H|F)$: Fraction of flu-inflicted worlds in which you have a headache
- $\mathbb{P}(H|F) = \frac{\text{Number of worlds with flu and headache}}{\text{Number of worlds with flu}}$

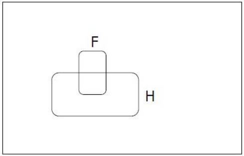
Conditional Probability

- $\mathbb{P}(H|F)$: Fraction of flu-inflicted worlds in which you have a headache
- $\mathbb{P}(H|F) = \frac{\text{Number of worlds with flu and headache}}{\text{Number of worlds with flu}}$



Conditional Probability

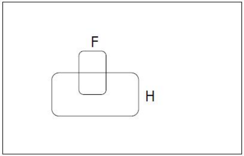
- $\mathbb{P}(H|F)$: Fraction of flu-inflicted worlds in which you have a headache
- $\mathbb{P}(H|F) = \frac{\text{Number of worlds with flu and headache}}{\text{Number of worlds with flu}}$



- $\mathbb{P}(H|F) = \frac{\text{Area of H and F region}}{\text{Area of F region}} = \frac{\mathbb{P}(H \cap F)}{\mathbb{P}(F)}$

Conditional Probability

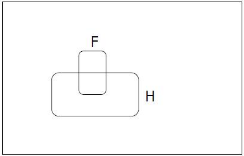
- $\mathbb{P}(H|F)$: Fraction of flu-inflicted worlds in which you have a headache
- $\mathbb{P}(H|F) = \frac{\text{Number of worlds with flu and headache}}{\text{Number of worlds with flu}}$



- $\mathbb{P}(H|F) = \frac{\text{Area of H and F region}}{\text{Area of F region}} = \frac{\mathbb{P}(H \cap F)}{\mathbb{P}(F)}$
- Conditional Probability: $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$

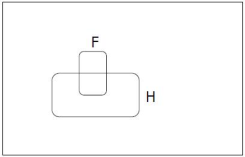
Conditional Probability

- $\mathbb{P}(H|F)$: Fraction of flu-inflicted worlds in which you have a headache
- $\mathbb{P}(H|F) = \frac{\text{Number of worlds with flu and headache}}{\text{Number of worlds with flu}}$



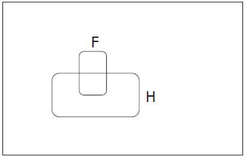
- $\mathbb{P}(H|F) = \frac{\text{Area of H and F region}}{\text{Area of F region}} = \frac{\mathbb{P}(H \cap F)}{\mathbb{P}(F)}$
- Conditional Probability: $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$
- **Corollary:** The Chain Rule $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$

Probabilistic Inference



- H = "Have a headache", F = "Have flu".
- $\mathbb{P}(H) = \frac{1}{10}$, $\mathbb{P}(F)\frac{1}{40}$, $\mathbb{P}(H|F) = \frac{1}{2}$

Probabilistic Inference



- H = "Have a headache", F = "Have flu".
- $\mathbb{P}(H) = \frac{1}{10}$, $\mathbb{P}(F)\frac{1}{40}$, $\mathbb{P}(H|F) = \frac{1}{2}$
- Suppose you wake up one day with a headache and think: "50 % of flus are associated with headaches so I must have a 50-50 chance of coming down with flu"
- Is this reasoning good?

Bayes Rule: Relates $\mathbb{P}(A|B)$ to $\mathbb{P}(A|B)$

Sensitivity and Specificity

	TRUE	FALSE
predict $+$	true $+$	false $+$
predict $-$	false $-$	true $-$

- Sensitivity $= \mathbb{P}(+|\text{disease})$
- FNR $= \mathbb{P}(-|T) = 1 -$ sensitivity
- Specificity $= \mathbb{P}(-|\text{healthy})$
- FPR $= \mathbb{P}(+|F) = 1 -$ specificity

Mammography

- Sensitivity of screening mammogram $\mathbb{P}(+|\text{cancer}) \approx 90\%$
- Specificity of screening mammogram $\mathbb{P}(-|\text{no cancer}) \approx 91\%$
- Probability that a woman age 40 has breast cancer $\approx 1\%$ If a previously unscreened 40 year old woman's mammogram is positive, what is the probability that she has breast cancer?

Mammography

- Sensitivity of screening mammogram $\mathbb{P}(+|\text{cancer}) \approx 90\%$
- Specificity of screening mammogram $\mathbb{P}(-|\text{no cancer}) \approx 91\%$
- Probability that a woman age 40 has breast cancer $\approx 1\%$ If a previously unscreened 40 year old woman's mammogram is positive, what is the probability that she has breast cancer?

$\mathbb{P}(\text{cancer}|+) =$

Mammography

- Sensitivity of screening mammogram $\mathbb{P}(+|\text{cancer}) \approx 90\%$
- Specificity of screening mammogram $\mathbb{P}(-|\text{no cancer}) \approx 91\%$
- Probability that a woman age 40 has breast cancer $\approx 1\%$ If a previously unscreened 40 year old woman's mammogram is positive, what is the probability that she has breast cancer?

$$\mathbb{P}(\text{cancer}|+) = \frac{\mathbb{P}(\text{cancer}, +)}{\mathbb{P}(+)} =$$

Mammography

- Sensitivity of screening mammogram $\mathbb{P}(+|\text{cancer}) \approx 90\%$
- Specificity of screening mammogram $\mathbb{P}(-|\text{no cancer}) \approx 91\%$
- Probability that a woman age 40 has breast cancer $\approx 1\%$ If a previously unscreened 40 year old woman's mammogram is positive, what is the probability that she has breast cancer?

$$\mathbb{P}(\text{cancer}|+) = \frac{\mathbb{P}(\text{cancer}, +)}{\mathbb{P}(+)} = \frac{\mathbb{P}(+|\text{cancer})\,\mathbb{P}(\text{cancer})}{\mathbb{P}(+)} =$$

Mammography

- Sensitivity of screening mammogram $\mathbb{P}(+|\text{cancer}) \approx 90\%$
- Specificity of screening mammogram $\mathbb{P}(-|\text{no cancer}) \approx 91\%$
- Probability that a woman age 40 has breast cancer $\approx 1\%$ If a previously unscreened 40 year old woman's mammogram is positive, what is the probability that she has breast cancer?

$$\mathbb{P}(\text{cancer}|+) = \frac{\mathbb{P}(\text{cancer}, +)}{\mathbb{P}(+)} = \frac{\mathbb{P}(+|\text{cancer})\,\mathbb{P}(\text{cancer})}{\mathbb{P}(+)} =$$

$$\frac{0.01 \times .9}{0.01 \times .9 + 0.99 \times 0.09} \approx$$

Mammography

- Sensitivity of screening mammogram $\mathbb{P}(+|\text{cancer}) \approx 90\%$
- Specificity of screening mammogram $\mathbb{P}(-|\text{no cancer}) \approx 91\%$
- Probability that a woman age 40 has breast cancer $\approx 1\%$ If a previously unscreened 40 year old woman's mammogram is positive, what is the probability that she has breast cancer?

$$\mathbb{P}(\text{cancer}|+) = \frac{\mathbb{P}(\text{cancer}, +)}{\mathbb{P}(+)} = \frac{\mathbb{P}(+|\text{cancer})\,\mathbb{P}(\text{cancer})}{\mathbb{P}(+)} =$$

$$\frac{0.01 \times .9}{0.01 \times .9 + 0.99 \times 0.09} \approx \frac{0.009}{0.009 + 0.09} \approx \frac{0.009}{0.1} \approx 9\%$$

Message: $\mathbb{P}(A|B) \neq \mathbb{P}(B|A)$.

Bayes' rule

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

(Bayes, Thomas (1763) An Essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*)

Rev. Thomas Bayes (1701–1761)

Prosecutor's fallacy: Sally Clark

Sally Clark (1964–2007)

- Two kids died with no explanation.
- Sir Roy Meadow testified that chance of this happening due to SIDS is $(1/8500)^2 \approx (73 \times 10^6)^{-1}$.
- Sally Clark found guilty and imprisoned.
- Later verdict overturned and Meadow struck off medical register.

Fallacy: $\mathbb{P}(\text{SIDS}|2\,\text{deaths}) \neq \mathbb{P}(\text{SIDS}, 2\,\text{deaths})$

$\mathbb{P}(\text{guilty}|+) = 1 - \mathbb{P}(\text{not guilty}|+) \neq 1 - \mathbb{P}(+|\text{not guilty})$

Independence

Two events A and B are **independent**, denoted $A \perp B$if

$$\mathbb{P}(A, B) = \mathbb{P}(A)\, \mathbb{P}(B).$$

Independence

Two events A and B are **independent**, denoted $A \perp B$if

$$\mathbb{P}(A, B) = \mathbb{P}(A)\,\mathbb{P}(B).$$

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

Independence

Two events A and B are **independent**, denoted $A \perp B$if

$$\mathbb{P}(A, B) = \mathbb{P}(A)\,\mathbb{P}(B).$$

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

$$\mathbb{P}(A^c|B) = \frac{\mathbb{P}(B) - \mathbb{P}(A, B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)(1 - \mathbb{P}(A))}{\mathbb{P}(B)} = \mathbb{P}(A^c)$$

Independence

A collection of events $\mathcal{A}$ are **mutually independent** if for any $\{i_1, i_2, \ldots, i_n\} \subseteq \mathcal{A}$

$$\mathbb{P}(\bigcap_{i=1}^{n} A_i) = \prod_{i=1}^{n} \mathbb{P}(A_i)$$

If A is independent of B and C, that does not necessarily mean that it is independent of (B, C) (example).

Conditional independence

A is conditionally independent of B given C, denoted

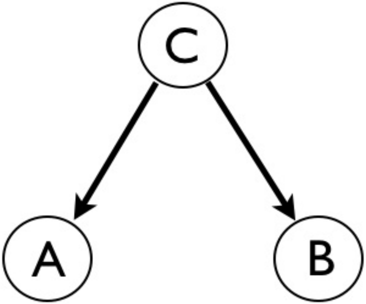
$$A \perp B \,|\, C$$

if

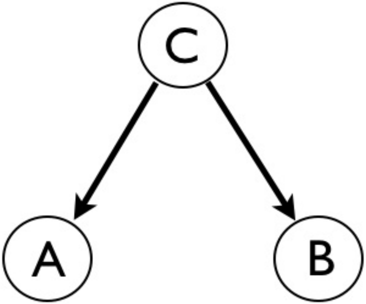
$$\mathbb{P}(A, B|C) = \mathbb{P}(A|C)\,\mathbb{P}(B|C).$$

$A \perp B \,|\, C$ does not imply and is not implied by $A \perp B$.

Common cause

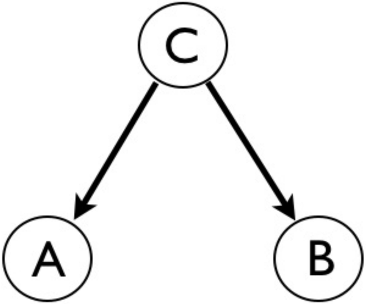


Common cause



$$p(x_A, x_B, x_C) = p(x_C)\, p(x_A|x_C)\, p(x_B|x_C)$$

Common cause

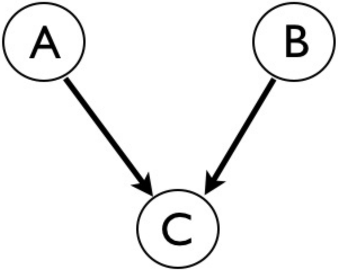


$$p(x_A, x_B, x_C) = p(x_C)\, p(x_A|x_C)\, p(x_B|x_C)$$

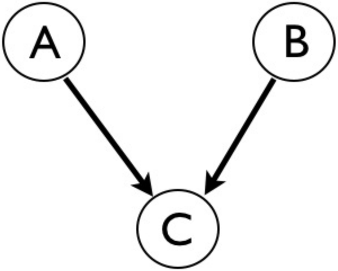
$X_A \not\perp X_B$ but $X_A \perp X_B \,|\, X_C$

Example: Lung cancer $\perp$ Yellow teeth | Smoking

Explaining away

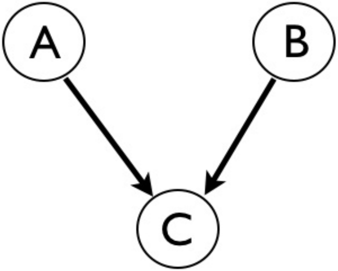


Explaining away



$$p(x_A, x_B, x_C) = p(x_A)\, p(x_B)\, p(x_C | x_A, x_B)$$

Explaining away

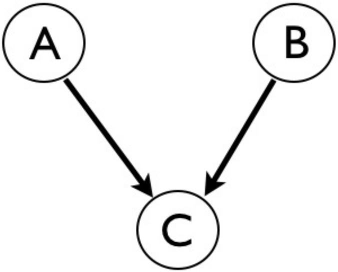


$$p(x_A, x_B, x_C) = p(x_A)\, p(x_B)\, p(x_C | x_A, x_B)$$

$$X_A \perp X_B \quad \text{but} \quad X_A \not\perp X_B \,|\, X_C$$

Example: Burglary $\not\perp$ Earthquake | Alarm

Explaining away

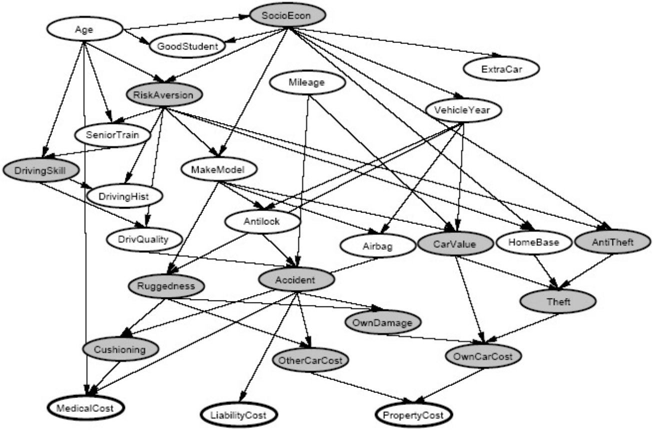


$$p(x_A, x_B, x_C) = p(x_A)\, p(x_B)\, p(x_C | x_A, x_B)$$

$$X_A \perp X_B \qquad \text{but} \qquad X_A \not\perp X_B \,|\, X_C$$

Example: Burglary $\not\perp$ Earthquake | Alarm Even if two variables are independent, they can become dependent when we observe an effect that they can both influence

Bayesian Networks



Simple case: POS Tagging. Want to predict an output vector $\mathbf{y} = \{y_0, y_1, \ldots, y_T\}$ of random variables given an observed feature vector $\mathbf{x}$ (Hidden Markov Model)

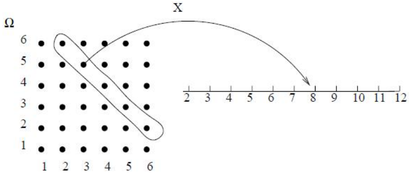
Random Variables

Random Variables

- A Random Variable is a function $X : \Omega \mapsto \mathbb{R}$

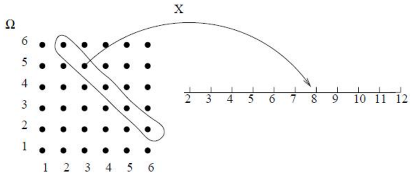
Random Variables

- A Random Variable is a function $X : \Omega \mapsto \mathbb{R}$
- Example: Sum of two fair dice



Random Variables

- A Random Variable is a function $X : \Omega \mapsto \mathbb{R}$
- Example: Sum of two fair dice



- The set of all possible values a random variable X can take is called its **range**
- **Discrete** random variables can only take isolated values (probability of a random variable taking a particular value reduces to counting)

Discrete Distributions

- Assume X is a discrete random variable. We would like to specify probabilities of events $\{X = x\}$

Discrete Distributions

- Assume X is a discrete random variable. We would like to specify probabilities of events $\{X = x\}$
- If we can specify the probabilities involving X, we can say that we have specified the probability distribution of X

Discrete Distributions

- Assume X is a discrete random variable. We would like to specify probabilities of events $\{X = x\}$
- If we can specify the probabilities involving X, we can say that we have specified the probability distribution of X
- For a countable set of values $x_1, x_2, \ldots x_n$, we have $\mathbb{P}(X = x_i) > 0, i = 1, 2, \ldots, n$ and $\sum_i \mathbb{P}(X = x_i) = 1$

Discrete Distributions

- Assume X is a discrete random variable. We would like to specify probabilities of events $\{X = x\}$
- If we can specify the probabilities involving X, we can say that we have specified the probability distribution of X
- For a countable set of values $x_1, x_2, \ldots x_n$, we have $\mathbb{P}(X = x_i) > 0, i = 1, 2, \ldots, n$ and $\sum_i \mathbb{P}(X = x_i) = 1$
- We can then define the **probability mass function** f of X by $f(X) = \mathbb{P}(X = x)$

Discrete Distributions

- Assume X is a discrete random variable. We would like to specify probabilities of events $\{X = x\}$
- If we can specify the probabilities involving X, we can say that we have specified the probability distribution of X
- For a countable set of values $x_1, x_2, \ldots x_n$, we have $\mathbb{P}(X = x_i) > 0, i = 1, 2, \ldots, n$ and $\sum_i \mathbb{P}(X = x_i) = 1$
- We can then define the **probability mass function** f of X by $f(X) = \mathbb{P}(X = x)$
- Sometimes write as f_X

Discrete Distributions

- Example: Toss a die and let X be its face value. X is discrete with range $\{1, 2, 3, 4, 5, 6\}$. The pmf is

x	1	2	3	4	5	6	$\sum$
$f(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

Discrete Distributions

- Example: Toss a die and let X be its face value. X is discrete with range $\{1, 2, 3, 4, 5, 6\}$. The pmf is

x	1	2	3	4	5	6	$\sum$
$f(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

- Another example: Toss two dice and let X be the largest face value. The pmf is

x	1	2	3	4	5	6	$\sum$
$f(x)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$	1

Expectation

- Assume X is a discrete random variable with pmf f.

Expectation

- Assume X is a discrete random variable with pmf f.
- The expectation of X, $\mathbb{E}[X]$ is defined by:

$$\mathbb{E}[X] = \sum_x x\mathbb{P}(X = x) = \sum_x x f(x)$$

Expectation

- Assume X is a discrete random variable with pmf f.
- The expectation of X, $\mathbb{E}[X]$ is defined by:

$$\mathbb{E}[X] = \sum_x x\mathbb{P}(X = x) = \sum_x x f(x)$$

- Sometimes written as μ_X. Is sort of a "weighted average" of the values that X can take (another interpretation is as a center of mass).

Expectation

- Assume X is a discrete random variable with pmf f.
- The expectation of X, $\mathbb{E}[X]$ is defined by:

$$\mathbb{E}[X] = \sum_x x\mathbb{P}(X = x) = \sum_x xf(x)$$

- Sometimes written as μ_X. Is sort of a "weighted average" of the values that X can take (another interpretation is as a center of mass).
- Example: Expected outcome of toss of a fair die - $\frac{7}{2}$

Expectation

If X is a random variable, then a function of X, such as X^2 is also a random variable. The following statement is easy to prove:

Expectation

If X is a random variable, then a function of X, such as X^2 is also a random variable. The following statement is easy to prove:

Theorem

If X is discrete with pmf f, then for any real-valued function g,

$$\mathbb{E}g(X) = \sum_x g(x)f(x)$$

Example: $\mathbb{E}[X^2]$ when X is outcome of the toss of a fair die, is $\frac{91}{6}$

Linearity of Expectation

- A consequence of the obvious theorem from earlier is that Expectation is linear i.e. has the following two properties for $a, b \in \mathbb{R}$ and functions g, h

Linearity of Expectation

- A consequence of the obvious theorem from earlier is that Expectation is linear i.e. has the following two properties for $a, b \in \mathbb{R}$ and functions g, h
- $\mathbb{E}(aX + b) = a\mathbb{E}X + b$

Linearity of Expectation

- A consequence of the obvious theorem from earlier is that Expectation is linear i.e. has the following two properties for $a, b \in \mathbb{R}$ and functions g, h
- $\mathbb{E}(aX + b) = a\mathbb{E}X + b$
  (Proof: Suppose X has pmf f. Then the above follows from $\mathbb{E}(aX + b) = \sum_x (ax + b) f(x) = a \sum_x f(x) + b \sum_x f(x) = a\mathbb{E}X + b$)
- $\mathbb{E}(g(X) + h(X)) = \mathbb{E}g(X) + \mathbb{E}h(X)$

Linearity of Expectation

- A consequence of the obvious theorem from earlier is that Expectation is linear i.e. has the following two properties for $a, b \in \mathbb{R}$ and functions g, h
- $\mathbb{E}(aX + b) = a\mathbb{E}X + b$
  (Proof: Suppose X has pmf f. Then the above follows from $\mathbb{E}(aX + b) = \sum_x (ax + b) f(x) = a \sum_x f(x) + b \sum_x f(x) = a\mathbb{E}X + b$)
- $\mathbb{E}(g(X) + h(X)) = \mathbb{E}g(X) + \mathbb{E}h(X)$
  (Proof: $\mathbb{E}(g(X) + h(X) = \sum_x (g(x) + h(x)) f(x) = \sum_x g(x) f(x) + \sum_x h(x) f(x) = \mathbb{E}g(X) + \mathbb{E}h(X)$)

Variance

- Variance of a random variable X, denoted by $Var(X)$ is defined as:

$$Var(X) = \mathbb{E}(X - \mathbb{E}X)^2$$

Variance

- Variance of a random variable X, denoted by $Var(X)$ is defined as:
$$Var(X) = \mathbb{E}(X - \mathbb{E}X)^2$$

- Is a measure of dispersion

Variance

- Variance of a random variable X, denoted by $Var(X)$ is defined as:

$$Var(X) = \mathbb{E}(X - \mathbb{E}X)^2$$

- Is a measure of dispersion
- The following two properties follow easily from the definitions of expectation and variance:

Variance

- Variance of a random variable X, denoted by $Var(X)$ is defined as:

$$Var(X) = \mathbb{E}(X - \mathbb{E}X)^2$$

- Is a measure of dispersion
- The following two properties follow easily from the definitions of expectation and variance:
  1. $Var(X) = \mathbb{E}X^2 - (\mathbb{E}\mathbb{X})^2$

Variance

- Variance of a random variable X, denoted by $Var(X)$ is defined as:

$$Var(X) = \mathbb{E}(X - \mathbb{E}X)^2$$

- Is a measure of dispersion
- The following two properties follow easily from the definitions of expectation and variance:
  1. $Var(X) = \mathbb{E}X^2 - (\mathbb{E}\mathbb{X})^2$
     (Proof: Write $\mathbb{E}X = \mu$. Expanding $Var(X) = \mathbb{E}(x - \mu)^2 = \mathbb{E}(X^2 - 2\mu X + \mu^2)$. Using linearity of expectation yields $\mathbb{E}(X^2) - \mu^2$)

Variance

- Variance of a random variable X, denoted by $Var(X)$ is defined as:
$$Var(X) = \mathbb{E}(X - \mathbb{E}X)^2$$

- Is a measure of dispersion
- The following two properties follow easily from the definitions of expectation and variance:
  1. $Var(X) = \mathbb{E}X^2 - (\mathbb{E}\mathbb{X})^2$
  (Proof: Write $\mathbb{E}X = \mu$. Expanding $Var(X) = \mathbb{E}(x - \mu)^2 = \mathbb{E}(X^2 - 2\mu X + \mu^2)$. Using linearity of expectation yields $\mathbb{E}(X^2) - \mu^2$)
  2. $Var(aX + b) = a^2 Var(X)$

Variance

- Variance of a random variable X, denoted by $Var(X)$ is defined as:

$$Var(X) = \mathbb{E}(X - \mathbb{E}X)^2$$

- Is a measure of dispersion
- The following two properties follow easily from the definitions of expectation and variance:
  1. $Var(X) = \mathbb{E}X^2 - (\mathbb{E}\mathbb{X})^2$
     (Proof: Write $\mathbb{E}X = \mu$. Expanding $Var(X) = \mathbb{E}(x - \mu)^2 = \mathbb{E}(X^2 - 2\mu X + \mu^2)$. Using linearity of expectation yields $\mathbb{E}(X^2) - \mu^2$)
  2. $Var(aX + b) = a^2 Var(X)$
     (Proof: $Var(aX + b) = \mathbb{E}(aX + b - (a\mu + b))^2 = \mathbb{E}(a^2(X - \mu)^2) = a^2 Var(X)$)

Joint Distributions

- Let $X_1, \ldots, X_n$ be discrete random variables. The function f defined by $f(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$ is called the joint probability mass function of $X_1, \ldots, X_n$

Joint Distributions

- Let $X_1, \ldots, X_n$ be discrete random variables. The function f defined by $f(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$ is called the joint probability mass function of $X_1, \ldots, X_n$
- $X_1, \ldots, X_n$ are independent if and only if $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_) = \mathbb{P}(X_1 = x_1) \ldots \mathbb{P}(X_n = x_n)$ for all $x_1, x_2, \ldots, x_n$

Joint Distributions

- Let $X_1, \ldots, X_n$ be discrete random variables. The function f defined by $f(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$ is called the joint probability mass function of $X_1, \ldots, X_n$
- $X_1, \ldots, X_n$ are independent if and only if $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_{)} = \mathbb{P}(X_1 = x_1) \ldots \mathbb{P}(X_n = x_n)$ for all $x_1, x_2, \ldots, x_n$
- If$X_1, \ldots, X_n$ are independent, then $\mathbb{E}X_1, X_2, \ldots, X_n = \mathbb{E}X_1\mathbb{E}X_2, \ldots, \mathbb{E}X_n$ (Also: If X and Y are independent, then $Var(X+Y) = Var(X) + Var(Y)$)

Joint Distributions

- Let $X_1, \ldots, X_n$ be discrete random variables. The function f defined by $f(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$ is called the joint probability mass function of $X_1, \ldots, X_n$
- $X_1, \ldots, X_n$ are independent if and only if $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_) = \mathbb{P}(X_1 = x_1) \ldots \mathbb{P}(X_n = x_n)$ for all $x_1, x_2, \ldots, x_n$
- If$X_1, \ldots, X_n$ are independent, then $\mathbb{E}X_1, X_2, \ldots, X_n = \mathbb{E}X_1 \mathbb{E}X_2, \ldots, \mathbb{E}X_n$ (Also: If X and Y are independent, then $Var(X+Y) = Var(X) + Var(Y)$)
- *Covariance:* The covariance of two random variables X and Y is defined as the number $Cov(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y)$

Joint Distributions

- Let $X_1, \ldots, X_n$ be discrete random variables. The function f defined by $f(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$ is called the joint probability mass function of $X_1, \ldots, X_n$
- $X_1, \ldots, X_n$ are independent if and only if $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_) = \mathbb{P}(X_1 = x_1) \ldots \mathbb{P}(X_n = x_n)$ for all $x_1, x_2, \ldots, x_n$
- If$X_1, \ldots, X_n$ are independent, then $\mathbb{E}X_1, X_2, \ldots, X_n = \mathbb{E}X_1 \mathbb{E}X_2, \ldots, \mathbb{E}X_n$ (Also: If X and Y are independent, then $Var(X + Y) = Var(X) + Var(Y)$)
- *Covariance:* The covariance of two random variables X and Y is defined as the number $Cov(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y)$
- It is a measure for the amount of linear dependency between the variables

Joint Distributions

- Let $X_1, \ldots, X_n$ be discrete random variables. The function f defined by $f(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$ is called the joint probability mass function of $X_1, \ldots, X_n$
- $X_1, \ldots, X_n$ are independent if and only if $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_)) = \mathbb{P}(X_1 = x_1) \ldots \mathbb{P}(X_n = x_n)$ for all $x_1, x_2, \ldots, x_n$
- If$X_1, \ldots, X_n$ are independent, then $\mathbb{E}X_1, X_2, \ldots, X_n = \mathbb{E}X_1\mathbb{E}X_2, \ldots, \mathbb{E}X_n$ (Also: If X and Y are independent, then $Var(X+Y) = Var(X) + Var(Y)$)
- *Covariance:* The covariance of two random variables X and Y is defined as the number $Cov(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y)$
- It is a measure for the amount of linear dependency between the variables
- If X and Y are independent, the covariance is zero

Some Important Discrete Distributions

Bernoulli Distribution: Coin Tossing

- We say X has a Bernoulli Distribution with success probability p if X can only take values 0 and 1 with probabilities

$$\mathbb{P}(X = 1) = p = 1 - \mathbb{P}(X = 0)$$

Bernoulli Distribution: Coin Tossing

- We say X has a Bernoulli Distribution with success probability p if X can only take values 0 and 1 with probabilities

$$\mathbb{P}(X = 1) = p = 1 - \mathbb{P}(X = 0)$$

- Expectation:

Bernoulli Distribution: Coin Tossing

- We say X has a Bernoulli Distribution with success probability p if X can only take values 0 and 1 with probabilities

$$\mathbb{P}(X = 1) = p = 1 - \mathbb{P}(X = 0)$$

- Expectation: $\mathbb{E}X = 0\mathbb{P}(X = 0) + 1\mathbb{P}(X = 1)p$

Bernoulli Distribution: Coin Tossing

- We say X has a Bernoulli Distribution with success probability p if X can only take values 0 and 1 with probabilities

$$\mathbb{P}(X = 1) = p = 1 - \mathbb{P}(X = 0)$$

- Expectation: $\mathbb{E}X = 0\mathbb{P}(X = 0) + 1\mathbb{P}(X = 1)p$
- Variance:

Bernoulli Distribution: Coin Tossing

- We say X has a Bernoulli Distribution with success probability p if X can only take values 0 and 1 with probabilities

$$\mathbb{P}(X = 1) = p = 1 - \mathbb{P}(X = 0)$$

- Expectation: $\mathbb{E}X = 0\mathbb{P}(X = 0) + 1\mathbb{P}(X = 1)p$
- Variance:
$Var(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \mathbb{E}X - (\mathbb{E}X)^2 = p(1 - p)$

Binomial Distribution

- Consider a sequence of n coin tosses. Suppose X counts the total number of heads. If the probability of "heads" is p, then we say X has a binomial distribution with parameters n and p and write $X \sim Bin(n, p)$

Binomial Distribution

- Consider a sequence of n coin tosses. Suppose X counts the total number of heads. If the probability of "heads" is p, then we say X has a binomial distribution with parameters n and p and write $X \sim Bin(n, p)$
- The pmf is

$$f(x) = \mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \text{ with } x = 0, 1, \ldots, n$$

Binomial Distribution

- Consider a sequence of n coin tosses. Suppose X counts the total number of heads. If the probability of "heads" is p, then we say X has a binomial distribution with parameters n and p and write $X \sim Bin(n, p)$
- The pmf is

$$f(x) = \mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \text{ with } x = 0, 1, \ldots, n$$

- Expectation:

Binomial Distribution

- Consider a sequence of n coin tosses. Suppose X counts the total number of heads. If the probability of "heads" is p, then we say X has a binomial distribution with parameters n and p and write $X \sim Bin(n, p)$
- The pmf is

$$f(x) = \mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \text{ with } x = 0, 1, \ldots, n$$

- Expectation: $\mathbb{E}X = np$. Could evaluate the sum, but that is messy. Use linearity of expectation instead (X can be viewed as a sum $X = X_1 + X_2, \ldots, X_n$ of n independent Bernoulli random variables).
- Variance:

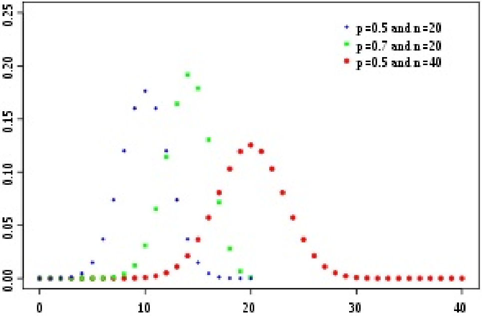
Binomial Distribution

- Consider a sequence of n coin tosses. Suppose X counts the total number of heads. If the probability of "heads" is p, then we say X has a binomial distribution with parameters n and p and write $X \sim Bin(n, p)$
- The pmf is

$$f(x) = \mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \text{ with } x = 0, 1, \ldots, n$$

- Expectation: $\mathbb{E}X = np$. Could evaluate the sum, but that is messy. Use linearity of expectation instead (X can be viewed as a sum $X = X_1 + X_2, \ldots, X_n$ of n independent Bernoulli random variables).
- Variance: $Var(X) = np(1-p)$ (showed in a similar way to the expectation)

Binomial Distribution



Geometric Distribution

- Again look at coin tosses, but count a different thing: Number of tosses before the first head

Geometric Distribution

- Again look at coin tosses, but count a different thing: Number of tosses before the first head
- $\mathbb{P}(X = x) = (1-p)^{x-1}p,$ for $x = 1, 2, 3....$ X is said to have a geometric distribution with parameter p, $X \sim G(p)$

Geometric Distribution

- Again look at coin tosses, but count a different thing: Number of tosses before the first head
- $\mathbb{P}(X = x) = (1-p)^{x-1}p,$ for $x = 1, 2, 3....$ X is said to have a geometric distribution with parameter p, $X \sim G(p)$
- Expectation:

Geometric Distribution

- Again look at coin tosses, but count a different thing: Number of tosses before the first head
- $\mathbb{P}(X = x) = (1-p)^{x-1}p,$ for $x = 1, 2, 3....$ X is said to have a geometric distribution with parameter p, $X \sim G(p)$
- Expectation: $\mathbb{E}X = \frac{1}{p}$

Geometric Distribution

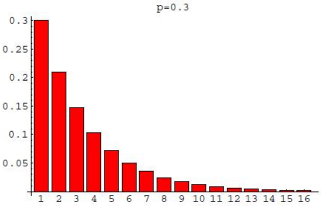
- Again look at coin tosses, but count a different thing: Number of tosses before the first head
- $\mathbb{P}(X = x) = (1-p)^{x-1}p,$ for $x = 1, 2, 3....$ X is said to have a geometric distribution with parameter p, $X \sim G(p)$
- Expectation: $\mathbb{E}X = \frac{1}{p}$
- Variance:

Geometric Distribution

- Again look at coin tosses, but count a different thing: Number of tosses before the first head
- $\mathbb{P}(X = x) = (1-p)^{x-1}p,$ for $x = 1, 2, 3....$ X is said to have a geometric distribution with parameter p, $X \sim G(p)$
- Expectation: $\mathbb{E}X = \frac{1}{p}$
- Variance: $Var(X) = \frac{1-p}{p^2}$

Geometric Distribution

- Again look at coin tosses, but count a different thing: Number of tosses before the first head
- $\mathbb{P}(X = x) = (1-p)^{x-1}p$, for $x = 1, 2, 3....$ X is said to have a geometric distribution with parameter p, $X \sim G(p)$
- Expectation: $\mathbb{E}X = \frac{1}{p}$
- Variance: $Var(X) = \frac{1-p}{p^2}$



Poisson Distribution

- A random variable X for which:

$$\mathbb{P}(X = x) = \frac{\lambda^x}{x!} \exp^{-\lambda}, \; x = 0, 1, 2, ...$$

  for fixed $\lambda > 0$

Poisson Distribution

- A random variable X for which:

$$\mathbb{P}(X = x) = \frac{\lambda^x}{x!} \exp^{-\lambda}, \; x = 0, 1, 2, ...$$

  for fixed $\lambda > 0$
- We write $X \sim Poi(\lambda)$
- Can be seen as a limiting distribution of $Bin(n, \frac{\lambda}{n})$

Law of Large Numbers

- To discuss the law or large numbers, we will first prove Chebyshev Inequality

Law of Large Numbers

- To discuss the law or large numbers, we will first prove Chebyshev Inequality

Theorem (Chebyshev Inequality)

Let X *be a discrete random variable with* $\mathbb{E}X = \mu$*, and let* $\epsilon > 0$ *be any positive real number. Then*

$$\mathbb{P}(|X - \mu| \geq \epsilon) \leq \frac{Var(X)}{\epsilon^2}$$

Law of Large Numbers

- To discuss the law or large numbers, we will first prove Chebyshev Inequality

Theorem (Chebyshev Inequality)

Let X be a discrete random variable with $\mathbb{E}X = \mu$, and let $\epsilon > 0$ be any positive real number. Then

$$\mathbb{P}(|X - \mu| \geq \epsilon) \leq \frac{Var(X)}{\epsilon^2}$$

- Basically states that the probability of deviation from the mean of more than k standard deviations is $\leq \frac{1}{k^2}$

Law of Large Numbers

Proof.

Let $f(x)$ denote the pmf for X. Then the probability that X differs from μ by ateast ϵ is given by $\mathbb{P}(|X - \mu| \geq \epsilon) = \sum_{|X-\mu|\geq\epsilon} f(x)$

Law of Large Numbers

Proof.

Let $f(x)$ denote the pmf for X. Then the probability that X differs from μ by ateast ϵ is given by $\mathbb{P}(|X-\mu| \geq \epsilon) = \sum_{|X-\mu|\geq\epsilon} f(x)$
We know that $Var(X) = \sum_x (x-\mu)^2 f(x)$, and this is at least as large as $\sum_{|x-\mu|\geq\epsilon}(x-\mu)^2 f(x)$ since all the summands are positive and we have restricted the range of summation.

Law of Large Numbers

Proof.

Let $f(x)$ denote the pmf for X. Then the probability that X differs from μ by ateast ϵ is given by $\mathbb{P}(|X-\mu| \geq \epsilon) = \sum_{|X-\mu|\geq\epsilon} f(x)$
We know that $Var(X) = \sum_x (x-\mu)^2 f(x)$, and this is at least as large as $\sum_{|x-\mu|\geq\epsilon}(x-\mu)^2 f(x)$ since all the summands are positive and we have restricted the range of summation. But this last sum is at least

$$\sum_{|x-\mu|\geq\epsilon} \epsilon^2 f(x) = \epsilon^2 \sum_{|x-\mu|\geq\epsilon} f(x) = \epsilon^2 \mathbb{P}(|x-\mu| \geq \epsilon)$$

So,

$$\mathbb{P}(|X-\mu| \geq \epsilon) \leq \frac{Var(X)}{\epsilon^2}$$

□

Law of Large Numbers(Weak Form)

Theorem (Law of Large Numbers)

Let $X_1, X_2, \ldots, X_n$ *be an independent trials process, with finite expected value* $\mu = \mathbb{E}X_j$ *and finite variance* $\sigma^2 = Var(X_j)$*. Let* $S_n = X_1 + X_2 + \cdots + X_n$*, then for any* $\epsilon > 0$

Law of Large Numbers(Weak Form)

Theorem (Law of Large Numbers)

Let $X_1, X_2, \ldots, X_n$ be an independent trials process, with finite expected value $\mu = \mathbb{E}X_j$ and finite variance $\sigma^2 = Var(X_j)$. Let $S_n = X_1 + X_2 + \cdots + X_n$, then for any $\epsilon > 0$

$$\mathbb{P}\Big(|\frac{S_n}{n} - \mu| \geq \epsilon\Big) \to 0$$

as $n \to \infty$ and equivalently

$$\mathbb{P}\Big(|\frac{S_n}{n} - \mu| < \epsilon\Big) \to 1$$

as $n \to \infty$

Sample average converges in probability towards expected value.

Proof.

Since $X_1, X_2, \ldots, X_n$ are independent and have the same distribution, we have $Var(S_n) = n\sigma^2$ and $Var(\frac{S_n}{n}) = \frac{\sigma^2}{den}$.

Proof.

Since $X_1, X_2, \ldots, X_n$ are independent and have the same distribution, we have $Var(S_n) = n\sigma^2$ and $Var(\frac{S_n}{n}) = \frac{\sigma^2}{den}$. We also know that $\mathbb{E}\frac{S_n}{n} = \mu$. By Chebyshev's inequality, for any $\epsilon > 0$

$$\mathbb{P}\Big(|\frac{S_n}{n} - \mu| \geq \epsilon\Big) \leq \frac{\sigma^2}{n\epsilon^2}$$

Thus for fixed ϵ, $n \to \infty$ implies the statement. □

Roadmap

- Today: Discrete Probability
- Next time: Continuous Probability