A quick review of probability theory

longhuan@sjtu.edu.cn

Outline

- Events and probability
- Bayes' rule
- Discrete random variables and expectation
- Moments and derivations

Definition of Probability

- **Experiment:** toss a coin twice
- **Sample space:** possible outcomes of an experiment
  - Ω={HH, HT,TH,TT}
- **Event:** a subset of possible outcomes.
  - A={HH}, B={HT, TH}
- **Probability of an event:** an number assigned to an event $\Pr(A)$
  - Axiom 1: $\Pr(A) \geq 0$
  - Axiom 2: $\Pr(\Omega) = 1$
  - Axiom 3: For every sequence of disjoint events $\Pr(\bigcup_i A_i) = \sum_i \Pr(A_i)$

Set notations

- $E_1 \cap E_2$ is the event that both E_1 and E_2 happen.
- $E_1 \cup E_2$ for the event that at least one of E_1 and E_2 happen.
- $E_1 - E_2$ for the occurrence of an event that is in E_1 but not in E_2.
- $\bar{E}$ stands for $\Omega - E$.

Lemma: for any two events E_1 and E_2:

$$\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2)$$

Proof. (Inclusion-exclusion principle)

Union Bound

Lemma: For any finite or countably infinite sequence of events $E_1, E_2, \ldots$

$$Pr\left(\bigcup_{i\geq 1} E_i\right) \leq \sum_{i\geq 1} \Pr(E_i).$$

Proof.

Independence

- Two events ***A and B are*** ***independent*** in case

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$$

- A set of events $\{A_1, A_2, \dots, A_k\}$ are **mutually independent** iff for any subset $I \subseteq [1, k]$

很强

$$\Pr\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \Pr(A_i)$$

Independence

Consider the experiment of tossing a coin twice

- **Example I.**
  - $A = \{HT, HH\}, B = \{HT\}$
  - Will event A independent from event B?
- **Example II.**
  - $A = \{HT\}, B = \{TH\}$
  - 非常相关 – Will event A independent from event B?
- Disjoint $\neq$ Independence → 不相关 两个极端 我发生你一定不发生
- If A is independent from B, B is independent from C, will A be independent from C?

Application1: Identify polynomials

$$(x+1)(x-2)(x+3)(x-4)(x+5)(x-6)$$
$$?= \ x^6 - 7x^3 + 25$$

- Generally $F(x) \,?= G(x)$

能否不展？ 带根

Probabilistic algorithm

- Assume $Max(Deg(G(x)), Deg\,(F(x))) = d$
- Algorithm
  - Choose an integer r uniformly at random in the range $\{1, \ldots., 100d\}$
  - Compute $F(r)$ and $G(r)$
  - If $F(r) = \; G(r)$ output Yes;
    otherwise, output No.

Analysis

- E: The event that the algorithm fails.
- The algorithm may fail iff
  - $F(x) \neq G(x)$ and $F(r) = G(r)$
  - r is the solution $of\ \ H(x) = F(x) - G(x) = 0.$
  - $H(x)$ has at most d solutions.
- $\Pr(E) \leq \frac{d}{100d} = \frac{1}{100}$
- **Idea** : If it keeps returning (Yes), we repeat the algorithm for *k* times.
  - The updated algorithm will fail iff every E_i fails for $1 \leq i \leq k$.

For $i = 1$ to k do

- Choose an integer r uniformly at random in the range $\{1, \dots, 100d\}$
- Compute $F(r)$ and $G(r)$
- $If\ F(r) = G(r)$ return Yes;
  otherwise stop and output No.

- $\Pr(E) = \Pr(E_1 \cap E_2 \cap \cdots \cap E_k)$

$$= \Pr(E_1) \cdot \Pr(E_2) \cdot \cdots \cdot \Pr(E_k)$$

$$\leq \left(\frac{1}{100}\right)^k$$

Conditioning

- If E and F are events with $\Pr(F) > 0$, the **conditional probability of E given F** is

$$\Pr(E|F\} = \frac{\Pr(E \cap F)}{\Pr(F)}$$

- If E and F are independent

$$\Pr(E|F) = \frac{\Pr(E \cap F)}{\Pr(F)} = \frac{\Pr(E)\Pr(F)}{\Pr(F)} = \Pr(E)$$

Application

- Example: Drug test

	Women	Men
Success	200	1800
Failure	1800	200

A = {Patient is a Women}

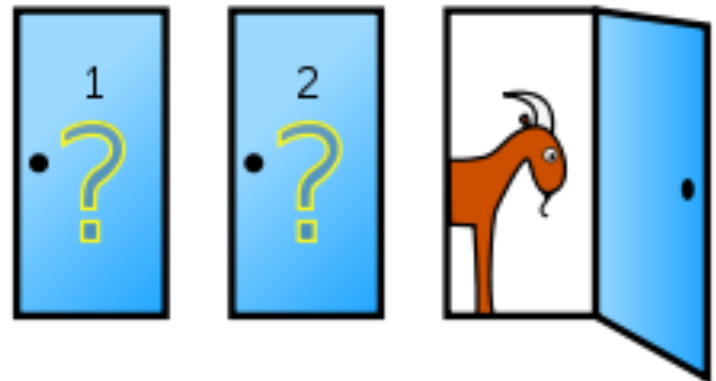
B = {Drug fails}

$\Pr(B|A) = ?$

$\Pr(A|B) = ?$

Application 2: Monty Hall problem

- Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?



Behind door 1	Behind door 2	Behind door 3	Result if staying at door #1	Result if switching to the door offered
Car	Goat	Goat	**Wins car**	Wins goat
Goat	**Car**	Goat	Wins goat	**Wins car**
Goat	Goat	**Car**	Wins goat	**Wins car**

Tuesday boy problem

- "I have two children. One is a boy born on a Tuesday. What is the probability I have two boys?"

<BTU, girl>	7
<girl, BTU>	7
<BTU, boy>	7
<boy, BTU>	7-1= 6

(7+6)/(7+7+7+6)=13/27

Drug Evaluation

	Women		Men	
	Drug I	Drug II	Drug I	Drug II
Success	200	10	19	1000
Failure	1800	190	1	1000

Simpson's Paradox: View I

	Women		Men	
	Drug I	Drug II	Drug I	Drug II
Success	200	10	19	1000
Failure	1800	190	1	1000

Drug II is better than Drug I

	Drug I	Drug II
Success	219	1010
Failure	1801	1190

A = {Using Drug I}

B = {Using Drug II}

C = {Drug succeeds}

Pr(C|A)= 219/2020 ~ 10%

Pr(C|B)=1010/2200 ~ **50%**

Simpson’s Paradox: View II

	Women		Men	
	Drug I	Drug II	Drug I	Drug II
Success	200	10	19	1000
Failure	1800	190	1	1000

Drug I is better than Drug II

Female Patient

A = {Using Drug I}

B = {Using Drug II}

C = {Drug succeeds}

Pr(C|A) ~ **10%**

Pr(C|B) ~ 5%

Male Patient

A = {Using Drug I}

B = {Using Drug II}

C = {Drug succeeds}

Pr(C|A) ~ **100%**

Pr(C|B) ~ 50%

Another version: Berkeley gender bias case (1973)

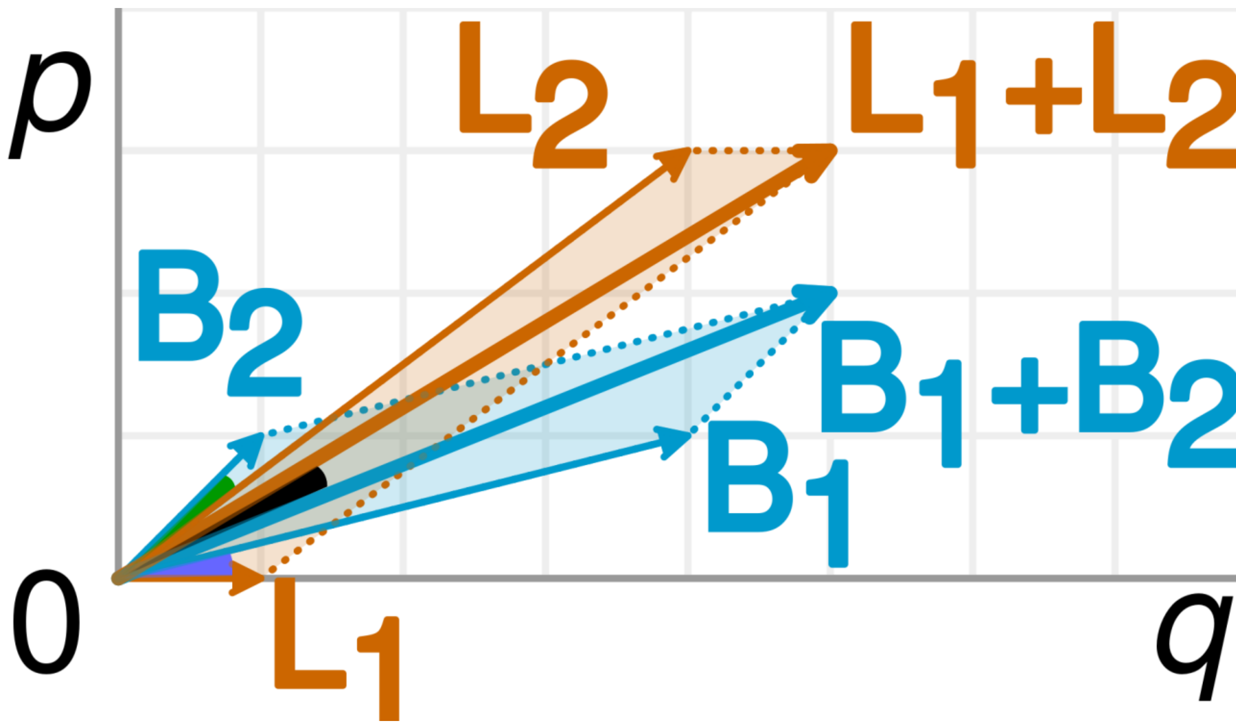
	Applicants	Admitted
Men	8442	**44%**
Women	4321	35%

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	272	6%	341	**7%**

Simpson's paradox - Wikipedia

Vector interpretation of Simpson's paradox

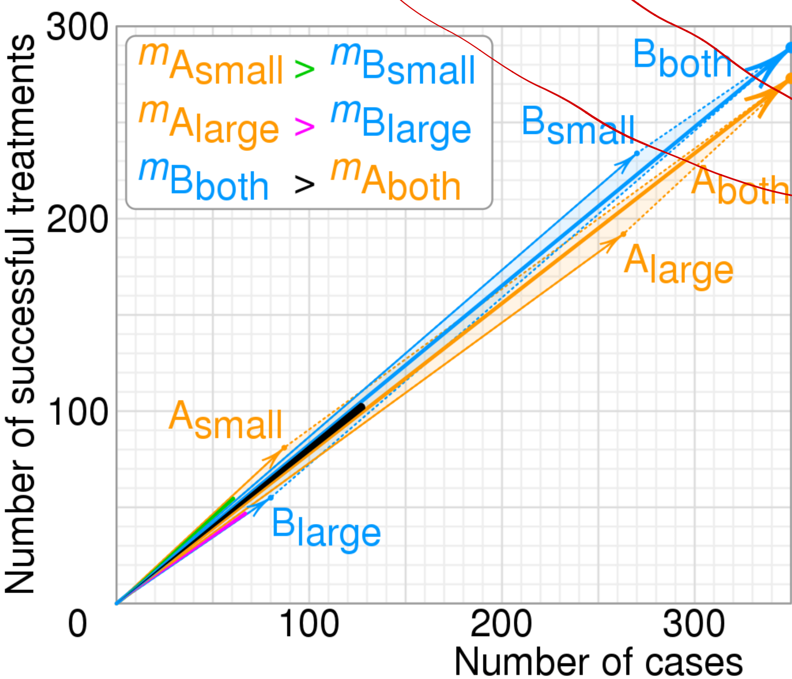
L2数据量大，斜率拉上去了



Simpson's paradox - Wikipedia

A real-life example from a medical study comparing the success rates of two treatments for kidney stones.

	Treatment A	Treatment B
Small Stones	Group 1 **93%** (81/87)	Group 2 87% (234/270)
Large Stones	Group 3 **73%** (192/263)	Group 4 69% (55/80)
Both	78% (273/350)	**83%** (289/350)



到底用啥？

因素的选择不一定合理

Vector representation in which each vector‘s slope denotes its success rate.

[Simpson's paradox - Wikipedia]

Law of total probability

- Let $E_1, E_2, \ldots, E_n$ be mutually disjoint events in the sample space Ω, and let $\bigcup_{i=1}^{n} E_i = \Omega$, then

$$\Pr(B) = \sum_{i=1}^{n} \Pr(B \cap E_i)$$

$$= \sum_{i=1}^{n} \Pr(B|E_i) \Pr(E_i)$$

Conditional Independence

- Event A and B are ***conditionally independent given C*** in case

$$\Pr(A \cap B|C) = \Pr(A|C) \cdot \Pr(B|C)$$

Or equivalently,

$$\Pr(A| B \cap C) = \Pr(A|C)$$ 等价式

条件独立性

$$\frac{P(ABC)}{P(BC)} = \frac{P(AC)}{P(C)}$$

- Example: There are three events: A, B, C
  - $\Pr(A) = \Pr(B) = \Pr(C) = \frac{1}{5}$
  - $\Pr(A \cap C) = \Pr(B \cap C) = \frac{1}{25}, \Pr(A \cap B) = \frac{1}{10}$
  - $\Pr(A \cap B \cap C) = \frac{1}{125}$
  - Whether A, B are conditionally independent given C?
  - Whether A, B are independent?

- Example: There are three events: A, B, C
  - $\Pr(A) = \Pr(B) = \Pr(C) = \frac{1}{5}$
  - $\Pr(A \cap C) = \Pr(B \cap C) = \frac{1}{25}, \Pr(A \cap B) = \frac{1}{10}$
  - $\Pr(A \cap B \cap C) = \frac{1}{125}$
  - Whether A, B are conditionally independent given C? Yes
  - Whether A, B are independent? No
- A and B are independent

  $\neq$ A and B are conditionally independent

Outline

- Events and probability
- Bayes' rule
- Discrete random variables and expectation
- Moments and derivations

Bayes' Rule

- Given two events A and B and suppose that $\Pr(A) > 0$. Then

$$\Pr(B \mid A) = \frac{\Pr(AB)}{\Pr(A)} = \frac{\Pr(A \mid B)\Pr(B)}{\Pr(A)}$$

- Example:

Pr(W|R)	R	¬R
W	0.7	0.4
¬W	0.3	0.6

R: It is a rainy day

W: The grass is wet

Pr(R|W) = ?

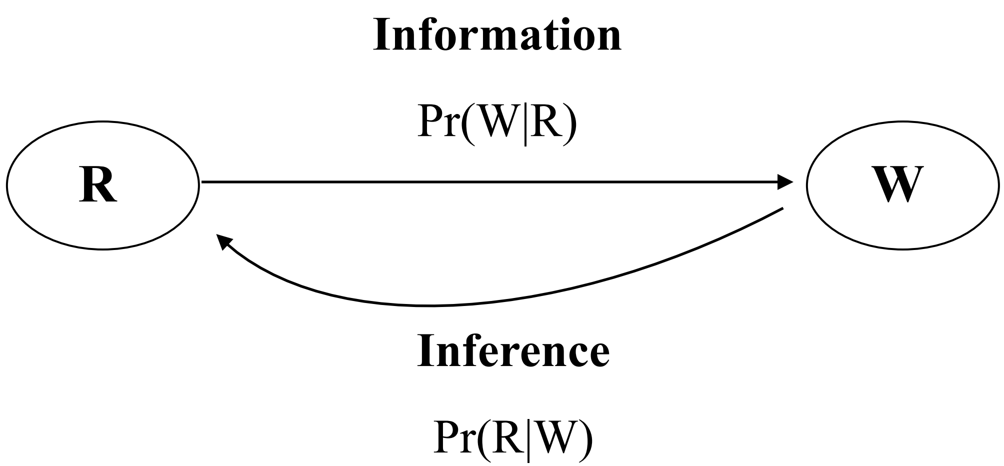
Pr(R) = 0.8

Bayes' Rule

	R	¬R
W	0.7	0.4
¬W	0.3	0.6

R: It rains

W: The grass is wet

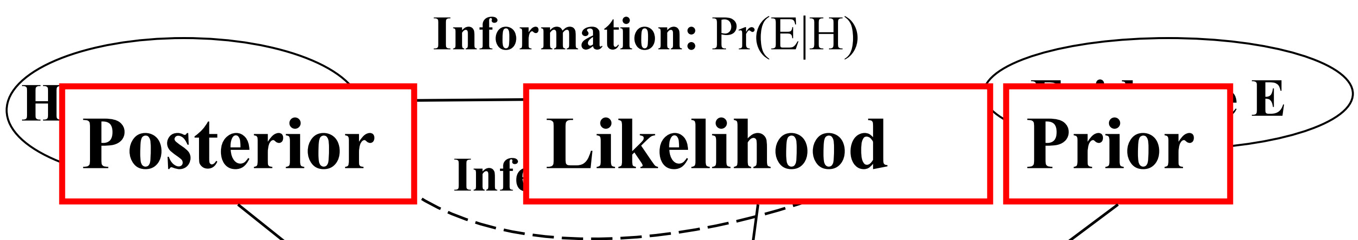


Bayes' Rule

	R	¬R
W	0.7	0.4
¬W	0.3	0.6

R: It rains

W: The grass is wet



$$\boxed{\Pr(H \mid E)} = \frac{\boxed{\Pr(E \mid H)}\,\boxed{\Pr(H)}}{\Pr(E)}$$

Bayes' Rule: More Complicated

Suppose that $B_1, B_2, \ldots B_k$ form a partition of S:

$$B_i \bigcap B_j = \varnothing; \quad \bigcup_i B_i = S$$

Suppose that $\Pr(Bi) > 0$ and $\Pr(A) > 0$. Then

$$\Pr(B_i \mid A) = \frac{\Pr(A \mid B_i)\Pr(B_i)}{\Pr(A)}$$

Bayes' Rule: More Complicated

Suppose that $B_1, B_2, \ldots B_k$ form a partition of S:

$$B_i \bigcap B_j = \varnothing; \quad \bigcup_i B_i = S$$

Suppose that $\Pr(Bi) > 0$ and $\Pr(A) > 0$. Then

$$\Pr(B_i \mid A) = \frac{\Pr(A \mid B_i)\Pr(B_i)}{\Pr(A)}$$

$$= \frac{\Pr(A \mid B_i)\Pr(B_i)}{\sum_{j=1}^{k} \Pr(AB_j)}$$

Bayes' Rule: More Complicated

Suppose that $B_1, B_2, \ldots B_k$ form a partition of S:

$$B_i \bigcap B_j = \varnothing; \quad \bigcup_i B_i = S$$

Suppose that $\Pr(Bi) > 0$ and $\Pr(A) > 0$. Then

$$\Pr(B_i \mid A) = \frac{\Pr(A \mid B_i)\Pr(B_i)}{\Pr(A)}$$

$$= \frac{\Pr(A \mid B_i)\Pr(B_i)}{\sum_{j=1}^{k} \Pr(AB_j)}$$

$$= \frac{\Pr(A \mid B_i)\Pr(B_i)}{\sum_{j=1}^{k} \Pr(B_j)\Pr(A \mid B_j)}$$

In all

E_0

Assume that $E_1, E_2, \ldots, E_n$ are mutually disjoint sets such that $\bigcup_{i=1}^{n} E_i = E$, then

$$\Pr\left(E_j \middle| B\right) = \frac{\Pr(E_j \cap B)}{\Pr(B)}$$

$$= \frac{\Pr\left(B \middle| E_j\right)\Pr(E_j)}{\sum_{i=0}^{n} \Pr\left(B \middle| E_i\right)\Pr(E_i)}$$

Example

E_i: the i^{th} coin is the biased one.

B: HHT

$\Pr(B|E_1) = \Pr(B|E_2)$

$$= \left(\frac{2}{3}\right) \cdot \left(\frac{1}{2}\right) \cdot \left(\frac{1}{2}\right) = \frac{1}{6}$$

$$\Pr(B|E_3) = \left(\frac{1}{2}\right) \cdot \left(\frac{1}{2}\right) \cdot \left(\frac{1}{3}\right) = \frac{1}{12}$$

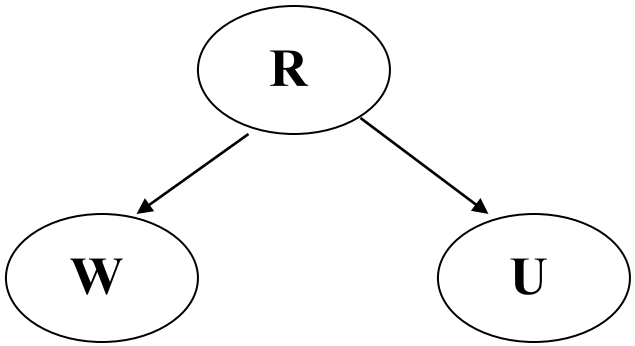
$$\Pr(E_i) = \frac{1}{3}$$

$\Pr(E_1\ |B) = 2/5 =$

$$\frac{(1/6)(1/3)}{2(1/6)(1/3)+(1/12)(1/3)}$$

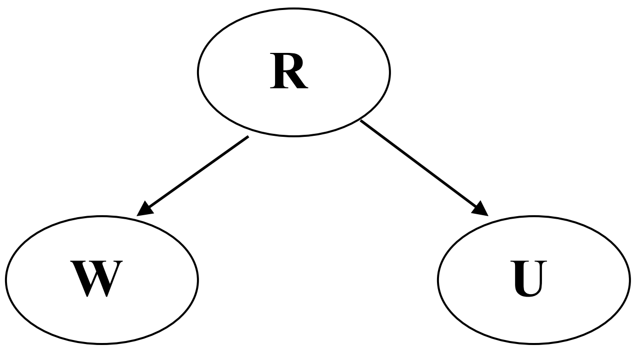
- We have three coins
  - Two of them: fair
  - The other one: $\Pr(H) = 2/3$
- Flip them we get: HHT
- Problem: What is the probability that the first coin is the biased one?

A More Complicated Example



R	It rains
W	The grass is wet
U	People bring umbrella

A More Complicated Example



R It rains

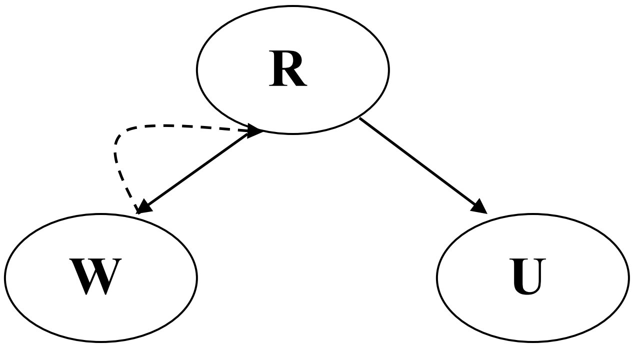
W The grass is wet

U People bring umbrella

$\Pr(UW|R)=\Pr(U|R)\Pr(W|R)$

$\Pr(UW|\neg R)=\Pr(U|\neg R)\Pr(W|\neg R)$

A More Complicated Example



$\Pr(R) = 0.8$

R It rains

W The grass is wet

U People bring umbrella

$\Pr(UW|R)=\Pr(U|R)\Pr(W|R)$

$\Pr(UW|\neg R)=\Pr(U|\neg R)\Pr(W|\neg R)$

Pr(W|R)	R	¬R
W	0.7	0.4
¬W	0.3	0.6

Pr(U|R)	R	¬R
U	0.9	0.2
¬U	0.1	0.8

$\Pr(U|W) = ?$

Outline

- Events and probability
- Bayes' rule
- Discrete random variables and expectation
- Moments and derivations
- The probabilistic method

Random Variable and Distribution

- A **random variable *X*** is a numerical outcomes of a random experiment

$$X : \Omega \to R$$

- The **distribution** of a random variable is the collection of possible outcomes along with their probabilities:
  - Discrete case:

$$\Pr(X = a) = \sum_{s \in \Omega, X(s) = a} \Pr(s)$$

Random Variable: Example

- Let S be the set of all sequences of two rolls of a die. Let X be the sum of the number of dots on the two rolls.
- The event $X = 4$ corresponds to the set of basic $events$ $\{(1,3), (2,2), (3,1)\}$. Hence

$$\Pr(X = 4) = \frac{3}{36} = \frac{1}{12}$$

Independent random variable

- Two random variables X and Y are independent if and only if

$$\Pr((X = x) \ \cap \ (Y = y)) = \Pr(X = x) \cdot \Pr(Y = y)$$

Expectation

- A basic characteristic of a random variable is expectation.
- The expectation of a random variable is a weighted average of the values it assumes, where each value is weighted by the probability that the variable assumes that value.

Expectation

加权平均

- A random variable $X \sim \Pr(X = x)$. Then, its expectation is

$$E[X] = \sum_x x \Pr(X = x)$$

- In an empirical sample, $x_1, x_2, \ldots, x_N$,

$$E[X] = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Examples

- The expectation of the random variable X representing the sum of two dice is

$$E(X) = \frac{1}{36} \cdot 2 + \frac{2}{36} \cdot 3 + \frac{3}{36} \cdot 4 + \cdots + \frac{1}{36} \cdot 12 = 7$$

Examples

- The expectation of the random variable X representing the sum of two dice is

$$E(X) = \frac{1}{36}\cdot 2 + \frac{2}{36}\cdot 3 + \frac{3}{36}\cdot 4 + \cdots + \frac{1}{36}\cdot 12 = 7$$

- A random variable X that takes on the value 2^i with probability $1/2^i$ for i=1,2,…

$$E(X) = \sum_{i=1}^{\infty} \frac{1}{2^i} 2^i = \sum_{i=1}^{\infty} 1 = \infty$$

期望
不收斂

Linearity of expectations

- Expectation of sum of random variables

$$E(X) + E(Y) = E(X + Y)$$

Proof.

- Generally: For any finite collection of discrete random variables $X_1, X_2, \ldots, X_n$ with finite expectations.

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]$$

Example

- Recall: The expected sum of two dice.

Solution:

Let $X = X_1 + X_2$

where X_i represents the outcome of dice i for $i = 1,2$. Then

$$E(X_i) = \frac{1}{6}\sum_{j=1}^{6} j = \frac{7}{2}$$

$$E(X) = E(X_1) + E(X_2) = 7$$

Lemma

For any constant c and discrete random variable X

$$E[cX] = c \cdot E[X]$$

Proof.

$$\begin{aligned} E[cX] &= \textstyle\sum_j j \cdot \Pr(cX = j) \\ &= c \textstyle\sum_j (j/c) \cdot \Pr(X = j/c) \\ &= c \textstyle\sum_k k \cdot \Pr(X = k) \\ &= c \cdot E[X] \end{aligned}$$

Variance

- The **variance** of a random variable X is the expectation of $(X - E[X])^2$:

$$
\begin{aligned}
Var(X) &= E((X - E[X])^2) \\
&= E(X^2 + E[X]^2 - 2XE[X]) \\
&= E(X^2 - E[X]^2) \\
&= E[X^2] - E[X]^2
\end{aligned}
$$

Bernoulli Distribution

- The outcome of an experiment can either be success (i.e., 1) and failure (i.e., 0).
- $\Pr(X = 1) = p,\ \Pr(X = 0) = 1 - p$
- $E[X] = p,\ Var(X) = p(1 - p)$

Binomial Distribution

- Consider a sequence of n independent coin flips. What is the distribution of the number of heads in the entire sequence?
- n draws of a Bernoulli distribution. X stands for the number of successes in these experiments.
- Random variable X stands for the number of times that experiments are successful.

$$\Pr(X = x) = p_\theta(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 1, 2, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

- $E[X] = np$ (by linearity), $Var(X) = np(1-p)$

几何分布

Geometric Distribution

- Suppose that we flip a coin $\boxed{\textbf{\textit{until}}}$ it lands on heads. What is the distribution of the number of flips?
- A geometric random variable X with parameter p is given by the following probability distribution on n=1,2,….:

$$\Pr(X = n) = (1-p)^{n-1}p$$

第一次成功

无记忆性　下一次成功的概率
与之前是否成功无关

Memoryless

- Geometric random variables are said to be *memoryless*: the probability that you will reach your first success *n* trials from now is independent of the number of failures you have experienced.
- Formally,

$$\Pr(X = n + k \mid X > k) \;=\; \Pr(X = n)$$

Proof.

$$
\begin{aligned}
\Pr(X = n + k \mid X > k) &= \frac{\Pr((X = n + k) \cap (X > k))}{\Pr(X > k)} \\
&= \frac{\Pr(X = n + k)}{\Pr(X > k)} \\
&= \frac{(1 - p)^{n+k-1} p}{\sum_{i=k}^{\infty} (1 - p)^{i} p} \\
&= \frac{(1 - p)^{n+k-1} p}{(1 - p)^{k}} \\
&= (1 - p)^{n-1} p \\
&= \Pr(X = n).
\end{aligned}
$$

Expectation

- Method 1: make use of the definitions.
- Method 2:

$$E[X] = p \cdot 1 + (1 - p) \cdot (E[X] + 1)$$

$$p \cdot E[X] = 1$$

$$E[X] = 1/p$$

Application: Coupon Collector's Problem

❖ Each box of cereal contain one of n different coupons.

❖ Once you obtain one of every type of coupon, you can send in for a prize.

❖ Coupons are distributed independently and uniformly at random from the n possibilities.

❖ **Question:** How many boxes of cereal must you buy before you obtain at least one of every type of coupon?

Solution

- Let X be the number of boxes bought until at least one of every type of coupon is obtained.
- X_i is the number of boxes bought while you had exactly $i-1$ different coupons.
- Clearly, $X=\sum_{1\leq i\leq n} X_i$
- X_i is a geometric random variable:
  - When exactly $i-1$ coupons have been found, the probability of obtaining a new coupon is $p_i = 1 - \frac{i-1}{n}$
  - $E[X_i] = \frac{1}{p_i} = \frac{n}{n-i+1}$
- By the linearity of expectations, we have

$$E[X] = E[\sum_{1\leq i\leq n} X_i] = \sum_{1\leq i\leq n} E[X_i] = \sum_{1\leq i\leq n} \frac{n}{n-i+1} = n\cdot\sum_{1\leq i\leq n} \left(\frac{1}{i}\right)$$

$$= n \cdot \ln n + \Theta(n)$$

(Where $\sum_{1\leq i\leq n} (\frac{1}{i})$= H(n) *harmonic number*)

Outline

- Events and probability
- Bayes' rule
- Discrete random variables and expectation
- Moments and derivations

Markov's Inequality

- Let X be a random variable that assumes only nonnegative values. Then for all $a > 0$

$$\Pr(X \geq a) \leq \frac{E[X]}{a}$$

- Proof.

Example

- Bound the probability of obtaining more than $\frac{3n}{4}$ heads in a sequence of n fair coin flips. Let $X_i = 1$ if the i^{th} coin flip is head, otherwise, $X_i = 0$.
  - Let $X = \sum_{1 \le i \le n} X_i$. It follows that $E[X] = \frac{n}{2}$
  - $\Pr\left(X \ge \frac{3n}{4}\right) \le \frac{E[X]}{\frac{3n}{4}} = 2/3$

Chebyshev's Inequality

- For any $a > 0$,

$$\Pr(|X - E(X)| \geq a) \leq \frac{Var[X]}{a^2}$$

- Proof.

Example: Coupon Collector's Problem

Recall: $E[X] = n \cdot Hn$

By Markov's inequality:

$$\Pr(X \geq 2n \cdot Hn) \leq 1/2$$

By Chebyshev's inequality, this can be improved to

$$\Pr(X \geq 2n \cdot Hn) \leq O\left(\frac{1}{(\ln n)^2}\right)$$