Probabilistic Theory: I

李昂生

Discrete Mathematics
U CAS
3, May, 2018

Outline

1. Finite probability
2. Basics
3. Applications
4. Bayes'
5. Expectation and variance
6. Moments and deviations
7. Exercises

Big Picture - New understanding

- 一切事物都由必然性和偶然性构成
- 科学研究必然性
- 概率研究偶然性
- 信息论研究必然性和偶然性的区分.
  Shannon's theory not achieved this yet, leaving a grand challenge to the 21st century.
- 概论和经典信息论都只研究无结构对象，客观世界对象可能无结构，也可能有结构，大多数是有结构的. 是否有一个研究结构中的偶然性的理论?
- **随机性**是上帝赐予人类的宝贵资源
  This is still a grand challenge in the 21st century. 为什么?
  随机性在人类进化、人工智能、博弈等的作用

Notions

- **Experiment**: a *procedure* that yields one of a finite set of *possible outcomes*
- **Sample space**: the set of all the possible outcomes of an experiment
- **Event**: a subset of the sample space.

Laplace's definition

Definition 1

If S is a finite nonempty sample space of equally likely possible outcomes, and E is an event, i.e., a subset of S, then the *probability* of event E is:

$$\Pr[E] = \frac{|E|}{|S|}, \tag{1}$$

that is, if x is uniformly picked at random, written $x \in_{\mathrm{R}} S$, then the probability that $x \in E$ is $\frac{|E|}{|S|}$.

We may restate the definition as

$$\Pr_{x \in_{\mathrm{R}} S}[x \in E] = \frac{|E|}{|S|}. \tag{2}$$

Remarks

- In the Laplace's definition, it is assumed that all the possible outcomes in the sample space occur with equal probability
- The probability is defined by the sizes of various sets, so sets are the basic notions of probability, so probability can be defined by using the notions of sets
- According to the definition, probability is naturally accompanying with counting problems
- There is no structure in the sample space

Complement

Theorem 2

Let E be an event in a sample space S, and let $\bar{E} = S \setminus E$. Then

$$\Pr[\bar{E}] = 1 - \Pr[E]. \tag{3}$$

Union

Theorem 3

Let E_1 and E_2 be two events in the sample space S. Then,

$$\Pr[E_1 \cup E_2] = \Pr[E_1] + \Pr[E_2] - \Pr[E_1 \cap E_2]. \tag{4}$$

Probability distribution

Let S be a sample space of possible outcomes.

Definition 4

A *probability distribution* over S is a function

$$p: \; S \to [0,1]$$

such that the following properties are satisfied:

(1) For each $s \in S$,

$$0 \le p(s) \le 1,$$

(2)

$$\sum_{s \in S} p(s) = 1.$$

(How to deal with the case, if the sum is $\neq 1$?)

Uniform distribution

Definition 5

Suppose that the sample space S has size n. Then the *uniform distribution* over S is to define

$$p(i) = \frac{1}{n}$$

for each $i \in S$.

Probability of an event

Suppose that S is a sample space, E is an event of S, and p is a probability distribution over S. Then, the *probability of the event E* is:

$$\Pr[E] = \sum_{s \in E} p(s). \tag{5}$$

Disjoint events

Theorem 6

Suppose that $E_1, E_2, \cdots$ are pairwise disjoint events in a sample space S. Then, for $E = \cup_{i\geq 1} E_i$,

$$\Pr[E] = \sum_{i\geq 1} \Pr[E_i]. \tag{6}$$

Conditional probability

Definition 7

Let E, F be events with $\Pr[F] > 0$. We define the **conditional probability** of E under the condition of F, written, $\Pr[E|F]$, as follows:

$$\Pr[E|F] = \frac{\Pr[E \cap F]}{\Pr[F]}. \tag{7}$$

By definition, if $\Pr[E] \cdot \Pr[F] > 0$, then

$$\Pr[E|F] \cdot \Pr[F] = \Pr[E \cap F] = \Pr[F|E] \cdot \Pr[E]. \tag{8}$$

Independency

Definition 8

Given events E, F, we say that E and F are **independent**, if:

$$\Pr[E \cap F] = \Pr[E] \cdot \Pr[F]. \tag{9}$$

Pairwise independency

We say that the events $E_1, E_2, \cdots, E_n$ are *pairwise independent*, if for all $i \neq j$,

$$\Pr[E_i \cap E_j] = \Pr[E_i] \cdot \Pr[E_j].$$

Mutual independency

We say that the events $E_1, E_2, \cdots, E_n$ are **mutually independent**, if for any set $X \subset [n]$,

$$\Pr[\cap_{x \in X} E_x] = \prod_{x \in X} \Pr[E_x].$$

Bernoulli Trails

A Bernoulli trail is an execution of an experiment that has only two possible outcomes, written 0 and 1, respectively.

- 1: success, true, head
- 0: failure, false, tail

Generally, the possible outcomes of a Bernoulli trail are 1 and 0 for which the probability that 1 occurs is p, and the probability that 0 occurs is $q = 1 - p$.

Binomial distribution theorem

Theorem 9

The probability of exactly k successes in n independent Bernoulli trails with probability p of 1, and $q = 1 - p$ of 0, is

$$\binom{n}{k} p^k (1-p)^{n-k}. \tag{10}$$

Proof

For the Bernoulli trail, let $p(1) = p$, and $p(0) = q = 1 - p$. Then, every string $a \in \{0, 1\}^n$ is a possible outcome of the n independent Bernoulli trails.
For every possible outcome $a = a_1 a_2 \cdots a_n$, by the independency,

$$p(a) = \prod_{i=1}^{n} p(a_i).$$

Let E be the event that there are exactly k 1's in n independent Bernoulli trails. Then, for every $a \in E$,

$$p(a) = p^k (1 - p)^{n-k}.$$

Proof - continued

By definition, $|E| = \binom{n}{k}$. Therefore,

$$\Pr[E] = \sum_{a \in E} p(a) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{11}$$

We write

$$b(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k},$$

called the *binomial distribution*, since

$$\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (p+q)^n = 1,$$

for $q = 1 - p$.

Applications

- Primality test
- Fingerprinting
- Error correcting code
- Hash functions
- And more to come

Bayes' Theorem

Theorem 10

Given events E, F in a sample space S, if $\Pr[E] \cdot \Pr[F] > 0$*, then*

$$\Pr[F|E] = \frac{\Pr[E|F] \cdot \Pr[F]}{\Pr[E|F] \cdot \Pr[F] + \Pr[E|\bar{F}] \cdot \Pr[\bar{F}]}. \tag{12}$$

Intuition

The probability of F under the condition of event E can be expressed by the probability of E under the conditions of both the event F and the complement of F.

Proof

(1)
$$\Pr[E|F] \cdot \Pr[F] = \Pr[E \cap F].$$

(2)
$$\Pr[F|E] \cdot \Pr[E] = \Pr[E \cap F].$$

(3)
$$\Pr[E|F] \cdot \Pr[F] + \Pr[E|\bar{F}] \cdot \Pr[\bar{F}] = \Pr[E].$$

(3) follows from

$$E = E \cap S = E \cap (F \cup \bar{F}) = (E \cap F) \cup (E \cap \bar{F}), \tag{13}$$

for disjoint sets $E \cap F$ and $E \cap \bar{F}$.
By the definition of conditional probability,

$$\Pr[F|E] = \frac{\Pr[F \cap E]}{\Pr[E]}. \tag{14}$$

The theorem follows from (1) - (3).

Generalised Bayes' Theorem

Theorem 11

Suppose

(i) *E is an event in sample space S,*

(ii) *$F_1, F_2, \cdots, F_n$ are events that form a partition of S, and*

(iii) $\Pr[E] \cdot \prod_{i=1}^{n} \Pr[F_i] > 0$.

Then, for every $j \in [n]$,

$$\Pr[F_j|E] = \frac{\Pr[E|F_j] \cdot \Pr[F_j]}{\sum_{i=1}^{n} \Pr[E|F_i] \cdot \Pr[F_i]}. \tag{15}$$

Understanding

Suppose that

1. E is a cancer, say, lung cancer
2. $F_1, F_2, \cdots, F_n$ are all the causes of lung cancer
3. The known data include: for each i
   – the probability of every cause F_i,
   – the probability of lung cancer occurs when cause i occurs

The generalised Bayes' Theorem allows to compute the probability that a lung cancer is caused exactly by cause F_j, for each j.
This understanding allows the theorem to be applied in a wide range of applications in engineering, and data mining etc.

Expectation

Definition 12

Let S be a sample space. A **random variable** on S is a function of the form:

$$X : \ S \to \mathbb{R}^{\geq 0}. \tag{16}$$

Definition 13

Let S be a sample space, and X be a random variable on S. Then the *expectation* of X, written $E[X]$, is defined by

$$E[X] = \sum_{s \in S} p[s] \cdot X(s). \tag{17}$$

The *deviation of X at $s \in S$* is:

$$X(s) - E[X].$$

Basic property

Theorem 14

If X is a random variable, $p(s)$ is a probability distribution of sample space S, then

(1) *For every $r \in \mathbb{R}^{\geq 0}$,*

$$\Pr[X = r] = \sum_{X(s)=r, s \in S} p(s).$$

(2)

$$E[X] = \sum_{r} \Pr[X = r] \cdot r.$$

Proof

Proof.

(1) is by definition. For (2).

$$E[X] = \sum_{s \in S} p(s)X(s) = \sum_{r} \sum_{s \in S, X(s)=r} p(s) \cdot r$$
$$= \sum_{r} r \cdot \Pr[X = r].$$

$\square$

Linearity of Expectation

Theorem 15

If X_i, $i = 1, 2, \cdots, n$ are random variables on S, not necessarily independent, for $X = \sum_{i=1}^{n} X_i$, and for $\alpha, \beta \geq 0$,

(1)

$$E[X] = \sum_{i=1}^{n} E[X_i].$$

(2)

$$E[\alpha X + \beta] = \alpha E[X] + \beta.$$

Proof

Proof.
For (1).

$$
\begin{aligned}
E[X] &= \sum_{s \in S} p(s)X(s) \\
&= \sum_{s \in S} p(s)(X_1(s) + \cdots + X_n(s)) \\
&= \sum_{i=1}^{n} \sum_{s \in S} p(s)X_i(s) \\
&= \sum_{i=1}^{n} E[X_i].
\end{aligned}
$$

For (2). Similarly by definition. □

Expectation of Bernoulli trails

Theorem 16

The expected number of successes when n independent Bernoulli trails are performed, where p is the probability of success on each trail, is

$$np.$$

Proof.

By the linearity of expectation. $\square$

The Geometric Distribution

Suppose that the probability that a coin comes up tails is p. The coin is flipped repeatedly until it comes up tails. What is the expected number of flips?
Let X be the random number of times of flips that come up tail for the first time. Then:

$$\Pr[X = k] = (1 - p)^{k-1} p. \tag{18}$$

This leads to

Definition 17

A random variable X has a *geometric distribution with parameter p*, if:

$$\Pr[X = k] = (1 - p)^{k-1} p, \; k \geq 1. \tag{19}$$

Expectation of Geometric distribution

Theorem 18

If the random variable X has the geometric distribution with parameter p, then

$$E[X] = \frac{1}{p}.$$

Proof.

$$\begin{aligned} E[X] &= \sum_{k=1}^{\infty} k \cdot \Pr[X = k] \\ &= \sum_{k \geq 1} k \cdot (1-p)^{k-1} p \\ &= \frac{1}{p}. \end{aligned}$$

Expectation of Geometric distribution - understanding

If the probability that an event occurs, is p, then on the average, $\frac{1}{p}$ many times experiments will make sure that, the event must occur.

Independent Random Variables

Definition 19

We say that random variables X and Y on S are *independent*, if:

$$\Pr[X = x \;\&\; Y = y] = \Pr[X = x] \cdot \Pr[Y = y]. \qquad (20)$$

Theorem 20

If X and Y are independent random variables on a sample space S, then

$$E[X \cdot Y] = E[X] \cdot E[Y].$$

Variance

Definition 21

Let X be a random variable on a sample space S. The *variance* of X, denoted by $\mathrm{Var}[X]$, is defined by

$$\mathrm{Var}[X] = \sum_{s \in S} (X(s) - E[X])^2 p(s). \tag{21}$$

The **standard deviation** of X, written $\sigma(X)$, is defined by

$$\sigma(X) = \sqrt{\mathrm{Var}[X]}. \tag{22}$$

Theorem of Deviation

Theorem 22

If X is a random variable on a sample space S, then

$$\mathrm{Var}[X] = E[X^2] - (E[X])^2 = E[(X - E[X])^2]. \qquad (23)$$

Proof.

$$\begin{aligned}
\mathrm{Var}[X] &= \sum_{s \in S} (X(s) - E[X])^2 p(s) \\
&= \sum_{s \in S} X^2(s)p(s) - 2E[X] \sum_{s \in S} X(s)p(s) + (E[X])^2 \sum_{s \in S} p(s) \\
&= E[X^2] - (E[X])^2 \\
&= E[(X - E[X])^2].
\end{aligned}$$

$\square$

Bienaymé's formula

Theorem 23

1. *If X, Y are independent random variables on a sample space S, then*

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y].$$

2. *If $X_1, X_2, \cdots, X_n$ are pairwise independent random variables on S, then*

$$\mathrm{Var}[X_1 + \cdots + X_n] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n].$$

Proof

$$\begin{aligned}
& \mathrm{Var}[X_1 + \cdots X_n] \\
= \;& E[(X_1 + \cdots + X_n)^2] - (E[X_1 + \cdots + X_n])^2 \\
= \;& \sum_{i=1}^{n} E[X_i^2] + 2E_{i<j}[X_i X_j] - (\sum_{i=1}^{n} E[X_i])^2 \\
= \;& \sum_{i=1}^{n} (E[X_i^2] - (E[X_i])^2), \text{ using pairwise independency} \\
= \;& \sum_{i=1}^{n} \mathrm{Var}[X_i].
\end{aligned}$$

The variance of n independent Bernoulli trails

1. $$E[X_i] = p.$$
2. $$X_i^2 = X_i$$
3. $$\mathrm{Var}[X_i] = E[X_i^2] - (E[X_i])^2 = p - p^2 = p(1-p).$$
4. $$\mathrm{Var}[X_1 + X_2 + \cdots X_n] = npq, q = 1 - p.$$

Occupancy problem

Given m balls and n bins, each ball is randomly put in one of the n bins.

Questions

1) What is the maximum number of balls in any bin?
2) What is the expected number of bins with k balls in them?

Sum Principle

For arbitrary events $E_1, E_2, \cdots, E_n$, not necessarily independent,

$$\Pr[\cup_{i=1}^{n} E_i] \leq \sum_{i=1}^{n} \Pr[E_i]. \tag{24}$$

We will use this simple principle to answer the questions.

The events

Consider the case $m = n$.

- For each i, $1 \leq i \leq n$, define
  X_i to be the number of balls in the ith bin.
  Clearly, $E[X_i] = 1$.
- Define the event:
  $E_j(k)$: Bin j has k or more balls.

$E_1(k)$

First, the probability that bin 1 has exactly i balls is:

$$
\begin{aligned}
& \binom{n}{i}(\frac{1}{n})^i(1-\frac{1}{n})^{n-i} \\
\leq\ & \binom{n}{i}(\frac{1}{n})^i \\
\leq\ & (\frac{ne}{i})^i(\frac{1}{n})^i \\
\leq\ & (\frac{e}{i})^i.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\Pr[E_1(k)] &\leq \sum_{i=k}^{n}(\frac{e}{i})^i \\
&\leq (\frac{e}{k})^k(1+\frac{e}{k}+(\frac{e}{k})^2+\cdots).
\end{aligned}
$$

The probability estimation

Set $k^* = \frac{3 \ln n}{\ln \ln n}$, then

$$\Pr[E_1(k^*)] \le (\frac{e}{k^*})^{k^*} \frac{1}{1 - e/k^*} \le n^{-2}. \tag{25}$$

By the same reason, for each i, $1 \le i \le n$,

$$\Pr[E_i(k^*)] \le n^{-2}.$$

By the sum principle,

$$\begin{aligned} \Pr[\cup_{i \ge 1} E_i(k^*)] &\le \sum_{i=1}^{n} \Pr[E_i(k^*)] \\ &\le n \cdot n^{-2} = \frac{1}{n}. \end{aligned}$$

Theorem

Theorem 24

With probability at least $1 - \frac{1}{n}$*, no bin has more than* $k^* = \frac{e \ln n}{\ln \ln n}$ *balls in it.*

Corollary. For $m = n \log n$, with probability $1 - o(1)$, every bin contains $O(\log n)$ balls.

The Markov Inequality

Let X be a discrete random variable and $f(x)$ be any real-valued function. The *expectation* of $f(X)$ is defined by

$$E[f(X)] = \sum_{x} f(x) \cdot \Pr[X = x]. \tag{26}$$

Theorem 25

(Markov Inequality) Let Y be a random variable assuming only non-negative values. Then for all $t \in \mathbb{R}^+$,

$$\Pr[Y \geq t] \leq \frac{E[Y]}{t}. \tag{27}$$

Equivalently,

$$\Pr[Y \geq k \cdot E[Y]] \leq \frac{1}{k}. \tag{28}$$

Proof

Proof.

Define a function $f(y)$ by

$$f(y) = \begin{cases} 1, & \text{if } y \geq t \\ 0, & \text{o.w.} \end{cases} \tag{29}$$

Then,

$$\Pr[Y \geq t] = E[f(Y)].$$

Since $f(y) \leq \frac{y}{t}$ for all y,

$$E[f(Y)] \leq E[\frac{Y}{t}] = \frac{E[Y]}{t},$$

and the theorem follows. □

Chebyshev's Inequality

Theorem 26

(Chebyshev's Inequality) Let X be a random variable with expectation μ and standard deviation σ. Then for any $t \in \mathbb{R}^+$,

$$\Pr[|X - \mu| \geq t \cdot \sigma] \leq \frac{1}{t^2}. \tag{30}$$

Proof

Proof.
First,

$$\Pr[|X-\mu| \geq t\sigma] = \Pr[(X-\mu)^2 \geq t^2\sigma^2].$$

The random variable $Y = (X-\mu)^2$ has expectation σ^2, and applying the Markov inequality to Y bounds this probability from above by $\frac{1}{t^2}$. $\square$

Principle of Deferred Decision

Question: Order independency.
Random Subsum Principle: Let a be a nonzero element in $\mathrm{GF}(2)^n$. Then;

$$\Pr_{x \in_{\mathrm{R}} \mathrm{GF}(2)^n}[a \cdot x = 0] = \frac{1}{2}. \tag{31}$$

Proof.
Let $a = (a_1, a_2, \cdots, a_n)$ with $a_1 \neq 0$ say. For a random $x \in \mathrm{GF}(2)^n$,

$$\begin{aligned} a \cdot x = 0 &\iff a_1 x_1 + a_2 x_2 + \cdots a_n x_n = 0 \pmod 2 \\ &\iff x_1 = -(a_2 x_2 + \cdots a_n x_n) \mod 2. \end{aligned}$$

Case 1 $x_2, \cdots, x_n$ are chosen before x_1. Done
Case 2. Otherwise. The same result holds by the principle of deferred decision.

The Coupon Collection Problem

- There are n types of coupons
- At each trial, a coupon is picked randomly
- Let m be the number of trials.

Question: What is the relationship between m and the probability that each type of the coupons has been collected.
Let X be the random number of trails required to collect at least one copy of each of the coupons.
Let $C_1, C_2, \cdots, C_X$ denote the sequence of trials, where C_i denotes the type of the coupon that is picked by the ith trial.
We say that the ith trial is *successful*, if C_i is different from C_j for all $j < i$.
Clearly, C_1 and C_X are both successful.

Analysis

For each i, define X_i to be the random number of trails that picks the $(i+1)$-th new type of coupons. Then

$$X_0 = 1$$

$$X = \sum_{i=0}^{n-1} X_i.$$

Let p_i be the probability of success on any trail of the ith epoch. Then

$$p_i = \frac{n-i}{n}$$

X_i is geometrically distributed with parameter p_i, therefore,

$$E[X_i] = \frac{1}{p_i},\ \mathrm{Var}[X_i] = \frac{1-p_i}{p_i^2}.$$

Analysis - continued

$$
\begin{aligned}
E[X] &= E[\sum_{i=0}^{n-1} X_i] = \sum_{i=0}^{n-1} E[X_i] \\
&= \sum_{i=0}^{n-1} \frac{n}{n-i} = n \sum_{i=0}^{n-1} \frac{1}{n-i} \\
&= n \cdot (1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}) = nH_n.
\end{aligned}
$$

H_n is the nth *Harmonic number*, which is asymptotically equal to $\ln n + \Theta(1)$, implying that

$$E[X] = n \ln n + O(n).$$

Analysis - continued

Since the X_i's are independent,

$$\begin{aligned}\sigma_X^2 &= \sum_{i=0}^{n-1} \sigma_{X_i}^2 = \sum_{i=0}^{n-1} \frac{ni}{(n-i)^2} \\ &= \sum_{i=1}^{n} \frac{n(n-i)}{i^2} = n^2 \sum_{i=1}^{n} \frac{1}{i^2} - nH_n.\end{aligned}$$

The sum $\sum_{i=1}^{n} \frac{1}{i^2}$ converges to $\frac{\pi^2}{6}$ as n goes to infinity, hence

$$\lim_{n\to\infty} \frac{\sigma_X^2}{n^2} = \frac{\pi^2}{6}.$$

Analysis - continued

By the Chebyshev's Inequality,

$$\Pr[|X - n \ln n| \geq n] \leq O(\frac{1}{n^2}).$$

The *k*th central moment

Definition 27

For $k \in \mathbb{N}$, the kth *moment* and the kth *central moment* of a random variable X are defined by

$$\mu_X^k = E[X^k]$$

$$\sigma_X^k = E[(X - E[X])^k].$$

The expected value is the first moment, the variance is the 2nd central moment.

Probability generating function

Definition 28

Let X be a non-negative integer-valued random variable with the density function p. The *probability generating function* of X is

$$G_X(z) = E[z^X] = \sum_{i=0}^{\infty} p(i)z^i. \tag{32}$$

Proposition: Let X be a non-negative integer-valued random variable with the probability generating function $G(z)$. Then:

1. $G(1) = 1$.
2. $E[X] = G'(1)$.
3. $E[X^2] = G''(1) + G'(1)$.
4. $\mathrm{Var}[X] = G''(1) + G'(1) - G'(1)^2$.

Distributions

1. Bernoulli distribution
   $E[X] = p$, $\mathrm{Var}[X] = pq$ and $G(z) = q + pz$, for $q = 1 - p$.
2. Binomial distribution
   $E[X] = np$, $\mathrm{Var}[X] = npq$, and $G(z) = (q + pz)^n$, for $q = 1 - p$.
3. Geometric distribution
   $E[X] = \frac{1}{p}$, $\mathrm{Var}[X] = q/p^2$, and $G(z) = pz/(1 - qz)$ for $q = 1 - p$.

Exercises

- Suppose that p and q are primes and $n = pq$. What is the probability that a randomly picked natural number less than n is not divisible by either p or q?
- Suppose that m and n are natural numbers. What is the probability that a randomly picked natural number less than mn is not divisible by either m or n?

A Card Game

- For fun and for better understanding

- Standard deck of 52 cards, each is randomly shuffled
- The pack is divided into 13 piles, each contains 4 cards
- Each pile is arbitrarily labeled by an index in $\{A, 1, 2, \cdots, 10, J, Q, K\}$
- The first move is to draw a card from the pile labeled K
- At each subsequent move, the card whose label is the face value of the last card is drawn
- The game is over when an attempt is made to draw a card from an empty pile

We win if at the end of the game, all cards were drawn, and lose otherwise.
What is the probability of winning the game?

谢谢！