Week 12-13: Discrete Probability

April 22, 2021

1 Probability Space

There are many problems about chances or possibilities, called **probability** in mathematics. When we roll two dice there are possible outcomes (i, j), where $1 \le i, j \le 6$. The collection $\{(i, j) : 1 \le i, j \le 6\}$ is know as a sample space. A **sample space** is just a collection Ω of all possible outcomes. A subset $S \subseteq \Omega$ is a called an **event** of Ω. A sample space is called **discrete** if it is finite or countably infinite.

A **finite probability space** is a finite sample space Ω together with a **probability function** $P : \mathscr{P}(\Omega) \to [0, 1]$ satisfying

(P1) $P(\Omega) = 1$.

(P2) If A and B are disjoint events, then $P(A \cup B) = P(A) + P(B)$.

We often call finite sample space and finite probability space just as sample space and probability space without mentioning their finiteness.

Each probability function $P : \mathscr{P}(\Omega) \to [0, 1]$ induces a function $P : \Omega \to [0, 1]$ defined by

$$P(\omega) = P(\{\omega\}), \quad \omega \in \Omega.$$

Clearly, $\sum_{\omega \in \Omega} P(\omega) = 1$. Conversely, each function $P : \Omega \to [0, 1]$ satisfying $\sum_{\omega \in \Omega} P(\omega) = 1$ induces a probability function $P : \mathscr{P}(\Omega) \to [0, 1]$ defined by

$$P(A) = \sum_{\omega \in A} P(\omega), \quad A \subseteq \Omega.$$

We can redefine a **finite probability space** as a finite sample space Ω together with a probability function $P : \Omega \to [0, 1]$ such that $\sum_{\omega \in \Omega} P(\omega) = 1$.

Example 1.1. For a probability space (Ω, P), if $P(\omega) = 1/|\Omega|$ for each $\omega \in \Omega$, we say that P is **equally likely distributed**. Then

$$P(A) = |A|/|\Omega|, \quad A \subseteq \Omega.$$

Example 1.2. Consider rolling of two fair dice, one blue and one red. The collection of possible ordered pairs of numbers in the top faces of the dice is the space $\Omega = \{(i, j) : 1 \le i, j \le 6\}$, and the probability function P is given by $P(i, j) = 1/36$. The event E that $i + j$ is even is the subset

$$\begin{aligned} E &= \big\{(1,1),(1,3),(1,5),(2,2),2,4),(2,6),(3,1),(3,3),(3,5), \\ &\quad (4,2),(4,4),(4,6),(5,1),(5,3),(5,5),(6,2),(6,4),(6,6)\big\}. \end{aligned}$$

It turns out that $P(E) = 18/36 = 1/2$.

2 Independence in Probability

Let (Ω, P) be a probability space. Given an event S such that $P(S) > 0$. The **conditional probability** of an event E given S is defined as

$$P(E|S) = \frac{P(E \cap S)}{P(S)}, \quad E \subseteq \Omega.$$

It is easy to see that the function $P(\,\cdot\,|S)$ on $\mathscr{P}(S)$ is a probability function.

If $P(S) = 0$, the above definition of conditional probability $P(E|S)$ does not make sense; instead, we define $P(E|S) = 0$.

Two events A and B are said to be **independent** if $P(A \cap B) = P(A)P(B)$. If $P(B) > 0$, then independence of A and B is equivalent to

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

If events A and B are independent, so are the events A and the complement B^c of B. In fact,

$$\begin{aligned} P(A \cap B^c) &= P(A - A \cap B) = P(A) - P(A \cap B) \\ &= P(A) - P(A)P(B) = P(A)(1 - P(B)) \\ &= P(A)P(B^c). \end{aligned}$$

Events $A_1, \ldots, A_n$ are said to be **independent** if

$$P(A_1 \cap \cdots \cap A_n) = P(A_1) \cdots P(A_n).$$

Example 2.1. A TV show has three rooms, one room contains a car, each of the other two rooms contains a sheep, unknowing to the audience. The game is to choose a person from the audience, and the person is asked to select a room by luck, if a room with a car is selected, the person wins the car; if a room with a sheep is selected, the person wins nothing. At the time the person has selected a room, one of the other two rooms is opened with a sheep, and the person is asked if he/she would like to change mind to select the room unopened. **Question:** Is it worth for the person to change his/her mind to select the other unopened room? Let c and s denote Car and Sheep. The sample space is $\Omega = \{(c, s, s), (s, c, s), (s, s, c)\}$.

Clearly, the probability is 1/3 if the person doesn't change. However, if he changes mind, the only case he lost the car is that he had selected the room with a car. Then if he changes his mind, the probability to win the car is 2/3.

Example 2.2. Given a fair HK dollar coin whose number-side is denoted by 1 and whose flower-side is denoted by 0. Tossed the coin n times, the possible outcomes form the sample space $\Omega = \{0, 1\}^n$. What is the probability that the number-side appeared exactly r times.

$$P(\text{number-side appreas } r \text{ times in } n \text{ tosses}) = \binom{n}{r} \cdot \frac{1}{2^n}.$$

Let E_k denote the event that the kth toss is the number-side. Then $\bar{E}_k$ is the event that the kth toss is the flower-side. Since $E_1, \ldots, E_n$ are independent and $P(E_k) = P(\bar{E}_k) = 1/2$, we have

$$P\left(\bigcap_{k=1}^{n} E_k\right) = \prod_{k=1}^{n} P(E_k) = \frac{1}{2^n}.$$

Example 2.3. A company purchases cables from three firms and keep a record of how many are defective. The facts are summarized as the table:

Firm	A	B	C
Fraction of cables purchased	0.50	0.20	0.30
Fraction of defective cables	0.01	0.04	0.02

From the table 30% of the cables are purchased from firm C and 2% percent of them are defective, i.e.,

$$P(A) = 0.50, \quad P(B) = 0.20, \quad P(C) = 0.30.$$

Let D denote the event of defect cables.

(a) The probabilities that a cable was purchased from firm and was defective are given as follows:

$$P(A \cap D) = P(A)P(D|A) = 0.50 \times 0.01 = 0.005,$$

$$P(B \cap D) = P(B)P(D|B) = 0.20 \times 0.04 = 0.008,$$

$$P(C \cap D) = P(C)P(D|C) = 0.30 \times 0.02 = 0.006.$$

(b) The probability that a random cable is defective is

$$\begin{aligned} P(D) &= P(A)P(D|A) + P(B)P(D|B) + P(C)P(D|C) \\ &= 0.005 + 0.008 + 0.006 = 0.019. \end{aligned}$$

Theorem 2.1 (Total Probability Formula)**.** *Let $A_1, \ldots, A_n$ be a partition of the sample space Ω and $P(A_i) > 0$ for all i. Then for each event B we have*

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i).$$

Proof. Since $P(B|A_i) = P(B \cap A_i)/P(A_i)$, we have $P(B \cap A_i) = P(B|A_i)P(A_i)$. Note that $B = \bigsqcup_{i=1}^{n}(B \cap A_i)$ (disjoint union). Thus

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(B|A_i)P(A_i).$$

□

Theorem 2.2 (Bayes' Formula)**.** *Let Ω be a sample space partitioned into events $A_1, \ldots, A_n$ such that $P(A_i) > 0$ for all i. If S is an event with $P(S) > 0$, then*

$$P(A_i|S) = \frac{P(S|A_i)P(A_i)}{P(S)}, \quad i = 1, \ldots, n$$

where $P(S) = \sum_{i=1}^{n} P(S|A_i)P(A_i)$.

Proof. Since $P(S|A_i) = P(S \cap A_i)/P(A_i)$, i.e., $P(A_i \cap S) = P(S|A_i)P(A_i)$, we have

$$P(A_i|S) = \frac{P(A_i \cap S)}{P(S)} = \frac{P(S|A_i)P(A_i)}{P(S)}.$$

□

Example 2.4. In the previous example, assume that defective cables are 19 per thousand in record. Now when a defective cable happens in someday. What are the chances that the particular defective cable comes from the three firms A, B, C respectively?

We are to computer the conditional probabilities given $P(D) = 0.019$:

$$P(A|D) = \frac{P(D|A)P(A)}{P(D)} = \frac{0.01 \times 0.5}{0.019} \approx 0.26,$$

$$P(B|D) = \frac{P(D|B)P(B)}{P(D)} = \frac{0.04 \times 0.2}{0.019} \approx 0.42,$$

$$P(C|D) = \frac{P(D|A)P(A)}{P(D)} = \frac{\times 0.02 \times 0.3}{0.019} \approx 0.32.$$

3 Random Variable

A **random variable** is a function from the sample space Ω to the set $\mathbb{R}$ of real numbers, usually denoted by capital letters X, Y, Z, etc. A random variable X is said to be **discrete** if the set of values

$$X(\Omega) = \{X(\omega) : \omega \in \Omega\}$$

can be listed as a (finite or infinite) sequence.

A coin is said to be **unfair** (or **biased**) if the probability p of the number-side is different from $\frac{1}{2}$. We imagine an experiment with one possible outcome of interest, traditionally called **success**; the complementary event is called **failure**. We assume that $P(\text{success}) = p$ for some p, $0 < p < 1$. We set $q = P(\text{failure})$, so that $p + q = 1$.

The sample space of a HK dollar coin tossed n times is $\Omega = \{0, 1\}^n$. Assume that the number-side appears at probability p, $0 < p < 1$. Then the probability function on Ω is given by

$$P(a_1,\ldots,a_n)=\prod_{i=1}^{n}p^{a_i}q^{1-a_i},\quad (a_1,\ldots,a_n)\in\Omega.$$

The probability of the event A that the number-side appears exactly k times is

$$P(A)=\binom{n}{k}p^kq^{n-k},\quad k=0,1,\ldots,n.$$

Let X be a random variable on a sample space Ω. Let C be a condition on the values $X(\omega)$. We adopt the following standard convention (notation) of probability theory:

$$P(X\in C)=P(\{\omega\in\Omega:X(\omega)\text{ satisfies }C\}).$$

Example 3.1. (a) A natural random variable X on the sample space Ω of outcomes when two dice are tossed is the one that gives the sum of the values shown on the top faces of two dice, i.e.,

$$X(i,j)=i+j,\quad (i,j)\in\Omega.$$

The probability that the sum is 8 is

$$\begin{aligned}P(X=8)&=P(\{(i,j)\in\Omega:i+j=8\})\\&=P(\{(2,6),(3,5),(4,4),(5,3),(6,2)\})=\frac{5}{36}.\end{aligned}$$

(b) Consider the sample space Ω of tossing a fair coin n times. One natural random variable X is the count of the number-sides come up. Thus

$$X(a_1,\ldots,a_n)=a_1+\cdots+a_n.$$

Let X_i denote the indicator function on Ω such that $X_i(\omega)=1$ if the ith toss is the number-side and $X_i(\omega)=0$ otherwise. Then $X=X_1+\cdots+X_n$ and

$$P(X=k)=\binom{n}{k}\cdot\frac{1}{2^n},\quad k\in\{0,1,\ldots,n\}.$$

(c) Consider the sample space Ω of words of 0 and 1 of length n and X counts the number of times that consecutive 1's appeared. Then X has values

$0, 1, 2, \ldots, \lceil n/2 \rceil$. For instance, for $n = 5$, we have $X(00000) = 0$, $X(10101) = 3$, $X(01100) = 1$, $X(01101) = 2$, etc. The event $\{X = 1\}$ has 15 members as

$$10000, 01000, 00100, 00010, 00001; 11000, 01100, 00110, 00011;$$

$$11100, 01110, 00111; 11110, 01111; 11111.$$

The event $\{X = 2\}$ has 15 members as

$$10100, 10010, 10001, 01010, 01001, 00101;$$

$$11010, 11001, 10110, 10011, 01101, 01011; 11101, 11011, 10111.$$

(d) **Joke: Random variable is neither random nor a variable.**

Two random variables X and Y on a sample space Ω are said to be **independent** if any two events, described by X and Y respectively, are independent, more specifically,

$$\{X \in I\} = \{\omega \in \Omega : X(\omega) \in I\}, \quad \{\omega \in \Omega : Y(\omega) \in J\} = \{Y \in J\}$$

are independent for all choices of intervals I and J of $\mathbb{R}$. This definition is equivalent to saying that the events

$$\{X \leq a\} = \{\omega \in \Omega : X(\omega) \leq a\}, \quad \{Y \leq b\} = \{\omega \in \Omega : Y(\omega) \leq b\}$$

are independent for all real numbers a and b. In case that $X(\Omega)$ and $Y(\Omega)$ are finite, then X and Y are independent if and only if the events

$$\{X = a\} = \{\omega \in \Omega : X(\omega) = a\}, \quad \{Y = b\} = \{\omega \in \Omega : Y(\omega) = b\}$$

are independent for all real numbers a and b.

4 Expectation and Standard Deviation

Experience suggests that, if we toss a fair die many times, then the various possible outcomes 1, 2, 3, 4, 5, and 6 will each happen about the same number of times, and the average value of these outcomes will be about the average of six numbers $1, 2, \ldots, 6$, i.e., $(1 + \cdots + 6)/6 = 3.5$. More generally, if X is a random variable on a finite sample space Ω with all outcomes equally likely, then the average value

$$A = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} X(\omega)$$

of X on Ω has a probabilistic interpretation: If members ω of Ω are selected at random many times and the values $X(\omega)$ are recorded, then the average of these values will probably close to a number A. This statement is actually a theorem that needs proof, but we accept it reasonably intuitive at the moment.

The **expectation** (or **expected value** or **mean**) of a random variable X on a finite sample space Ω is defined as

$$E(X) = \mu = \sum_{\omega \in \Omega} X(\omega) P(\omega). \tag{1}$$

If all outcomes are equally likely, then $P(\omega) = 1/|\Omega|$ for all $\omega \in \Omega$, so $E(X)$ is exactly the average value A discussed above.

In Example 3.1(c) with $n = 5$, the expectation of the random variable X is

$$E(X) = 0 \cdot \frac{1}{32} + 1 \cdot \frac{15}{32} + 2 \cdot \frac{15}{32} + 3 \cdot \frac{1}{32} = \frac{3}{2}.$$

For random variables X and Y on a sample space Ω, there are random variables aX, $X + Y$ and XY on Ω defined by

$$(aX)(w) = aX(\omega), \quad (X+Y)(\omega) = X(\omega) + Y(\omega),$$

$$(XY)(\omega) = X(\omega)Y(\omega), \quad \omega \in \Omega.$$

Theorem 4.1. (a) $E(X+Y) = E(X) + E(Y)$.

(b) $E(aX) = aE(X)$ *for real numbers* a.

(c) $E(c) = c$ *for any constant random variable* c *on* Ω.

(d) $E(X - \mu) = 0$, *where* $\mu = E(X)$.

Proof.

$$\begin{aligned}
E(X+Y) &= \sum_{\omega\in\Omega}(X+Y)(\omega)P(\omega)\\
&= \sum_{\omega\in\Omega}\big(X(\omega)+Y(\omega)\big)P(\omega)\\
&= \sum_{\omega\in\Omega}X(\omega)P(\omega)+\sum_{\omega\in\Omega}Y(\omega)P(\omega)\\
&= E(X)+E(Y).
\end{aligned}$$

$$\begin{aligned}
E(aX) &= \sum_{\omega\in\Omega}(aX)(\omega)P(\omega)=\sum_{\omega\in\Omega}aX(\omega)P(\omega)\\
&= a\sum_{\omega\in\Omega}X(\omega)P(\omega)=aE(X).
\end{aligned}$$

$$E(c)=\sum_{\omega\in\Omega}cP(\omega)=c\sum_{\omega\in\Omega}P(\omega)=cP(\Omega)=c.$$

$$E(X-\mu)=E(X)-E(\mu)=E(X)-\mu=0.$$

□

Theorem 4.2. *Let X be a random variable on a finite sample space Ω. If $f:\mathbb{R}\to\mathbb{R}$ be a function, then $f(X)=f\circ X$ is a random variable on Ω, and*

$$E(f(X))=\sum_{k\in X(\Omega)} f(k)\cdot P(X=k). \tag{2}$$

Proof. Notice that $\{X=k\}=\{\omega\in\Omega : X(\omega)=k\}$ is an event and

$$\Omega=\bigcup_{k\in X(\Omega)}\{X=k\}\ \text{(disjoint)}.$$

We have

$$\begin{aligned}
E(f(X)) &= \sum_{\omega\in\Omega} f(X(\omega)) \cdot P(\omega) \\
&= \sum_{k\in X(\Omega)} \sum_{\omega\in\{X=k\}} f\big(X(\omega)\big) \cdot P(\omega) \\
&= \sum_{k\in X(\Omega)} f(k) \sum_{\omega\in\{X=k\}} P(\omega) \\
&= \sum_{k\in X(\Omega)} f(k) \cdot P(X=k).
\end{aligned}$$

□

The expectation of a random variable X gives us its probabilistic average. However, it doesn't tell us how close the average we are likely to be. We need another measurement describe this. A natural choice is the probabilistic average distance of X from its mean μ. This is the "mean deviation" $E(|X-\mu|)$, i.e., the mean of all deviations $|X(\omega)-\mu|$, $\omega \in \Omega$. While the measurement is sometimes used, it turns out that a similar measure, called the standard deviation, is much more manageable and useful technically.

The **standard deviation** of a random variable X on a sample space Ω is

$$\sigma_X = \sqrt{E((X-\mu)^2)} \tag{3}$$

and the **variance** of X is

$$V(X) = \sigma_X^2 = E((X-\mu)^2). \tag{4}$$

Theorem 4.3. *For a discrete random variable X with mean μ, we have*

$$V(X) = \sum_{k\in X(\Omega)} (k-\mu)^2 \cdot P(X=k) = E(X^2) - \mu^2.$$

Proof. Since $(X-\mu)^2 = X^2 - 2\mu X - \mu^2$, we have

$$\begin{aligned}
V(X) = E((X-\mu)^2) &= E(X^2) - 2\mu E(X) + \mu^2 \\
&= E(X^2) - 2\mu^2 + \mu^2 \\
&= E(X^2) - \mu^2.
\end{aligned}$$

□

Theorem 4.4. *If X and Y are independent random variables, then*

$$E(XY) = E(X) \cdot E(Y). \tag{5}$$

Proof. We restrict to discrete random variables. We have

$$\begin{aligned}
E(XY) &= \sum_{m \in XY(\Omega)} m \cdot P(XY = m) \\
&= \sum_{m \in XY(\Omega)} m \sum_{k \in X(\Omega),\, l \in Y(\Omega),\, kl = m} P(X = k, Y = l) \\
&= \sum_{k \in X(\Omega),\, l \in Y(\Omega)} kl \cdot P(X = k, Y = l) \\
&= \sum_{k \in X(\Omega),\, l \in Y(\Omega)} kl \cdot P\big((X = k) \cap (Y = l)\big) \\
&= \sum_{k \in X(\Omega),\, l \in Y(\Omega)} kl \cdot P(X = k) \cdot P(Y = l) \\
&= \sum_{k \in X(\Omega)} k \cdot P(X = k) \sum_{l \in Y(\Omega)} l \cdot P(Y = l) \\
&= E(X) \cdot E(Y).
\end{aligned}$$

□

Theorem 4.5. *If $X_1, \ldots, X_n$ are independent random variables, then*

$$V(a_1 X_1 + \cdots + a_n X_n) = a_1^2 V(X_1) + \cdots + a_n^2 V(X_n).$$

Proof. Note that $E(aX) = a\mu$ with $\mu = E(X)$. We see that

$$V(aX) = E((aX - a\mu)^2) = E(a^2(X - \mu)^2) = a^2 E((X - \mu)^2) = a^2 V(X).$$

We only give proof for two independent random variables X and Y. Let μ_X denote the mean of X and μ_Y the mean of Y. Then

$$\begin{aligned}
V(aX + bY) &= E((aX + bY)^2 - (a\mu_X + b\mu_Y)^2) \\
&= E(a^2 X^2 + 2abXY + b^2 Y^2) \\
&\quad -(a^2 \mu_X^2 + 2ab\mu_X \mu_Y + b^2 \mu_Y^2) \\
&= a^2 E(X^2) + 2abE(X)E(Y) + b^2 E(Y^2) \\
&\quad -a^2 \mu_X^2 - 2ab\mu_X \mu_Y - b^2 \mu_Y^2.
\end{aligned}$$

Since $V(X) = E(X^2) - \mu_X^2$, $V(Y) = E(Y^2) - \mu_Y^2$, and $E(XY) = E(X)E(Y)$, we have

$$V(aX + bY) = a^2V(X) + b^2V(Y).$$

$\square$

Example 4.1. Let S_n denote denote the random variable on the sample space of a biased HK dollar coin tossed n times with probability p of the number-side, counting the number of times that the number-side appeared in the n tosses. Then

$$E(S_n) = np, \quad V(S_n) = npq.$$

Proof. Let X_i denote the indicator function that the ith toss is success, $i = 1, \ldots, n$. Note that

$$E(X_i) = 1 \cdot P(X_i = 1) + 0 \cdot P(X_i = 0) = p,$$

$$V(X_i) = E(X_i^2) - (E(X_i))^2 = p - p^2 = p(1-p) = pq,$$

and $S_n = X_1 + \cdots + X_n$. We have

$$E(S_n) = \sum_{i=1}^{n} E(X_i) = np,$$

$$V(S_n) = \sum_{i=1}^{n} V(X_i) = npq.$$

$\square$

5 Probability Distributions

For a random variable X on a probability space Ω, the **cumulative distribution function** (**cdf**) of X is a function $F : \mathbb{R} \to [0, 1]$ defined by

$$F(y) = P(X \leq y), \quad y \in \mathbb{R}.$$

If X is a discrete random variable, then F sums the values, i.e.,

$$F(y) = \sum_{k \leq y} P(X = k).$$

It is clear that $F(+\infty) = F(X < +\infty) = 1$. The function F is a non-decreasing function, i.e., $F(x) \leq F(y)$ for $x \leq y$. In fact,

$$\begin{aligned} F(x) &= P(X \leq x) \\ &\leq P(X \leq x) + P(x < X \leq y) \\ &= P(X \leq y) = F(y). \end{aligned}$$

Example 5.1. Consider the sample space of rolling a pair of dice, one is colored black and the other white. Let X_b denote the number on the top face of the black die, X_w the number on the top face of the white die, and X_s the sum of two numbers on the top faces of the dice, i.e., $X_s = X_b + X_w$. Then X_b and X_w has the cdf

$$F(y) = \begin{cases} 0 & \text{for } \; y < 1 \\ k/6 & \text{for } \; k \leq y < k+1 \; (k = 1, \ldots, 5) \\ 1 & \text{for } \; y \geq 6 \end{cases}$$

The random variable $X_s = X_b + X_w$ has the cdf

$$F(y) = \begin{cases} 0 & \text{for } \; y < 2 \\ 1/36 & \text{for } \; 2 \leq y < 3 \\ 3/36 & \text{for } \; 3 \leq y < 4 \\ 6/36 & \text{for } \; 4 \leq y < 5 \\ 10/36 & \text{for } \; 5 \leq y < 6 \\ 15/36 & \text{for } \; 6 \leq y < 7 \end{cases} \qquad F(y) = \begin{cases} 21/36 & \text{for } \; 7 \leq y < 8 \\ 26/36 & \text{for } \; 8 \leq y < 9 \\ 30/36 & \text{for } \; 9 \leq y < 10 \\ 33/36 & \text{for } \; 10 \leq y < 11 \\ 35/36 & \text{for } \; 11 \leq y < 12 \\ 1 & \text{for } \; y \geq 12 \end{cases}$$

Example 5.2 (Cumulative Binomial Distribution). Let S_n denote the random variable on the sample space of tossing a coin n times with success probability p, counting the number of successes in the n experiments. The probability function P is given by

$$P(S_n = k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, \ldots, n, \quad \text{where } q = 1 - p.$$

The cdf for S_n is

$$F(y) = \sum_{k \leq y} \binom{n}{k} p^k q^{n-k}, \quad -\infty < y < \infty.$$

Example 5.3 (**Uniform Distribution).** What it means when people talk about choosing a random number on the interval $[0, 1)$? One may state it, of course, as that all numbers in $[0, 1]$ are equally likely to be chosen? But this doesn't make sense, since the the probability of choosing a given number in the interval is 0. What we mean instead is that the probability of choosing a number in any given sub-interval $[a, b)$ is proportional to the length of the sub-interval. The probability of choosing the number in $[0, 1)$ is 1, so the probability of choosing it in $[a, b)$ is $b-a$. Let U denote the random variable on $[0, 1)$ that gives the value of the number chosen. Then $P(U \in [0, x)) = x$ for $0 \leq x < 1$. Since $P(U = x) = 0$, we see that $P(U \in [0, x]) = P(U \in [0, x))$. Thus the cdf F_U is given by

$$F_U(y) = P(U \leq y) = \begin{cases} 0 & \text{for } \ y < 0 \\ y & \text{for } \ 0 \leq y < 1 \\ 1 & \text{for } \ y \geq 1 \end{cases}$$

Example 5.4. (a) Consider the random variable X that records the value obtained when a single fair die is tossed. Thus $P(X = k) = 1/6$, $k = 1, \ldots, 6$. Let us define $f(k) = P(X = k)$, $k = 1, \ldots, 6$. Consider the function

$$f(x) = \begin{cases} 0 & \text{for } \ x < 0 \\ 1/6 & \text{for } \ 0 \leq x \leq 6 \\ 0 & \text{for } \ x > 6 \end{cases}$$

We see that the cdf F is given by

$$\begin{aligned} F(k) = P(X \leq k) &= \text{area under } f \text{ over } (-\infty, k] \\ &= \int_{-\infty}^{k} f(x)\mathrm{d}x, \quad k = 1, \ldots, 6. \end{aligned}$$

However, $F(y) = P(X \leq y)$ does not hold for non-integer $y \in [0, 6]$.

(b) Consider the random variable S_n on the sample space of tossing a coin n times with success probability p. Setting $f(k) = P(S_n = k) = \binom{n}{k} p^k q^{n-k}$, $k = 0, 1, \ldots, n$, where $q = 1 - p$. Define the function f by

$$f(x) = \begin{cases} 0 & \text{for } \ x \leq -1 \\ \binom{n}{k} p^k q^{n-k} & \text{for } \ k-1 < x \leq k \in [0, n] \cap \mathbb{Z} \\ 0 & \text{for } \ x > n \end{cases}$$

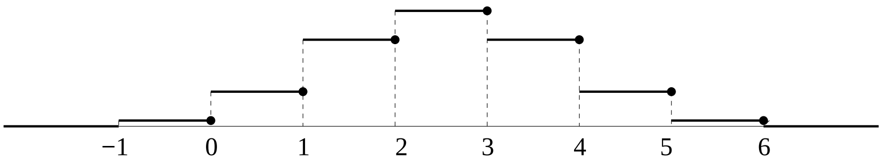
The cdf F of X can be given by integration

$$\begin{aligned} F(k) = P(S_n \leq k) &= \text{area under } f \text{ over } (-\infty, x] \\ &= \int_{-\infty}^{k} f(x)\mathrm{d}x, \quad k = 0, 1, \ldots, n. \end{aligned}$$

For $p = 1/2$, $n = 6$, we have

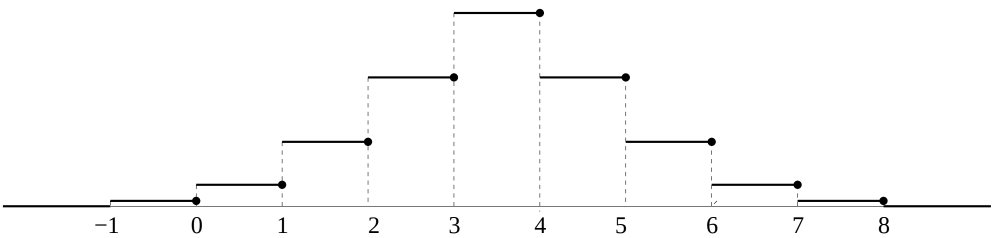
$$f(0) = 1/64, \quad f(1) = 6/64, \quad f(2) = 15/64, \quad f(3) = 20/64,$$
$$f(4) = 15/64, \quad f(5) = 6/64, \quad f(6) = 1/64.$$

The the graph of the function f is



For $p = 1/2$ and $n = 12$, we have

x	0	1	2	3	4	5	6	7	8	9	10
$f(x)$	$\frac{1}{512}$	$\frac{10}{512}$	$\frac{45}{512}$	$\frac{120}{512}$	$\frac{210}{512}$	$\frac{252}{512}$	$\frac{210}{512}$	$\frac{120}{512}$	$\frac{45}{512}$	$\frac{10}{512}$	$\frac{1}{512}$



(c) The uniform distribution F_U on $[0, 1)$ can be given as an integral of a function f as $F_U(y) = \int_{-\infty}^{y} f(x)\mathrm{d}x$, where is defined by

$$f(x) = \begin{cases} 0 & \text{for } \; x < 0 \\ 1 & \text{for } \; 0 \leq x < 1 \\ 0 & \text{for } \; x \geq 1 \end{cases}$$

which is known as the **density function** of the random variable U.

Definition 5.1. The **normalization** of a random variable X on a sample space Ω, having mean μ and standard deviation $\sigma > 0$, is the random variable

$$\tilde{X} = \frac{X - \mu}{\sigma}.$$

Theorem 5.2. *Let X be a random variable with mean μ, standard deviation $\sigma > 0$, and cumulative distribution function F. Let $\tilde{X}$ denote the normalization of X, and let $\tilde{F}$ denote the cdf for $\tilde{X}$. Then*

(a) $E(\tilde{X}) = 0$, $V(\tilde{X}) = 1$, *and* $\sigma_{\tilde{X}} = 1$.

(b) $F(y) = \tilde{F}(\frac{y-\mu}{\sigma})$ *for all* $y \in \mathbb{R}$.

(c) $\tilde{F}(y) = F(\sigma y + \mu)$ *for all* $y \in \mathbb{R}$.

Proof. (a) $E(\tilde{X}) = E\big(\frac{X-\mu}{\sigma}\big) = \frac{1}{\sigma}E(X-\mu) = \frac{1}{\sigma}\big(E(X)-\mu\big) = \frac{1}{\sigma}(\mu-\mu) = 0$. Note that

$$V(X+c) = E\Big(\big((X+c)-(\mu+c)\big)^2\Big) = E\big((X-\mu)^2\big) = V(X)$$

and

$$V(aX) = E\big((aX-a\mu)^2\big) = E\big(a^2(X-\mu)\big) = a^2E\big((X-\mu)^2\big) = a^2V(X).$$

We have

$$V(\tilde{X}) = V\Big(\frac{X-\mu}{\sigma}\Big) = V\Big(\frac{X}{\sigma}\Big) = \frac{1}{\sigma^2}V(X) = 1.$$

(b) Since $X \le y$ iff $X-\mu \le y-\mu$ iff $\frac{X-\mu}{\sigma} \le \frac{y-\mu}{\sigma}$ iff $\tilde{X} \le \frac{y-\mu}{\sigma}$, we have

$$F(y) = P(X \le y) = P\Big(\tilde{X} \le \frac{y-\mu}{\sigma}\Big) = \tilde{F}\Big(\frac{y-\mu}{\sigma}\Big).$$

(c) Since $\tilde{X} \le y$ iff $\frac{X-\mu}{\sigma} \le y$ iff $X \le \sigma y + \mu$, we have

$$\tilde{F}(y) = P(\tilde{X} \le y) = P(X \le \sigma y + \mu) = F(\sigma y + \mu).$$

□

Example 5.5. Let S_n be the random variable on the sample space of tossing a biased coin n times with success probability p, and failure probability $q = 1-p$. The corresponding normalized random variable is

$$\tilde{S}_n = \frac{S_n - \mu}{\sigma} = \frac{S_n - np}{\sqrt{npq}}.$$

The value set of S_n is $\{0, 1, \ldots, n\}$. While the value set of $\tilde{S}_n$ is more complicated:

$$\Big\{\frac{-np}{\sqrt{npq}}, \frac{-np+1}{\sqrt{npq}}, \frac{-np+2}{\sqrt{npq}}, \ldots, \frac{-np+n}{\sqrt{npq}}\Big\}.$$

For $p = 1/2$ and $n = 6$, we have

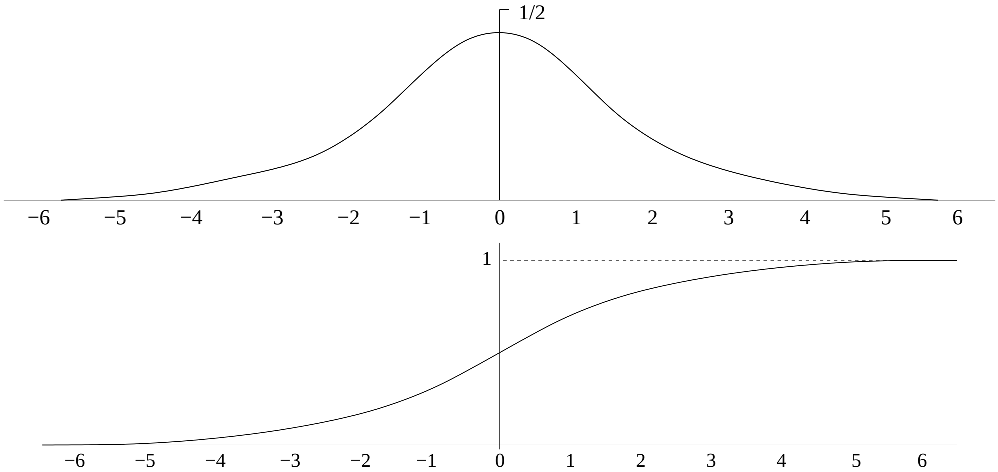
$$\left\{\frac{-3}{\sqrt{3/2}}, \frac{-2}{\sqrt{3/2}}, \frac{-1}{\sqrt{3/2}}, 0, \frac{1}{\sqrt{3/2}}, \frac{2}{\sqrt{3/2}}, \frac{3}{\sqrt{3/2}}\right\}$$
$$\approx \{-2.45, -2.31, -0.816, 0, 0.816, 2.31, 2.45\}$$

Let F_n and $\tilde{F}_n$ denote the cdf's of S_n and $\tilde{S}_n$ respectively. There exists a function $f_n(x)$ such that

$$\tilde{F}_n(y) = P(\tilde{S}_n \leq y) = \int_{-\infty}^{y} f_n(x)\mathrm{d}x.$$

Definition 5.3. The **Gaussian distribution** (or **standard normal distribution**) is the function Φ defined by

$$\Phi(y) = \int_{-\infty}^{y} \phi(x)\mathrm{d}x, \quad y \in \mathbb{R}, \quad \text{where} \quad \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R}.$$



A central result of probability theory is that $\tilde{F}_n(y) \approx \Phi(y)$ for large n and for all $y \in \mathbb{R}$. The distribution Φ does not depend on the success probability p. The following theorem states a much more general phenomena similar to the limit of $\tilde{F}_n$.

Example 5.6. In the Bernoulli space $\Omega = \{0, 1\}^n$ with $n = 10,000$ and success probability $p = 1/10$, the expected number of success is $\mu = np = 1000$. Estimate the chance that number of success is between 950 and 1050.

This is to compute $F_n(1050) - F_n(950)$. Note that

$$\sigma = \sqrt{npq} = \sqrt{10,000 \cdot \frac{1}{10} \cdot \frac{9}{10}} = 30.$$

We have

$$F_n(1050) = \tilde{F}_n\Big(\frac{1050 - 1000}{30}\Big) = \tilde{F}_n(1.7) \approx \Phi(1.7) \approx 0.955,$$

$$F_n(949) = \tilde{F}_n\Big(\frac{949 - 1000}{30}\Big) = \tilde{F}_n(-1.7) \approx \Phi(-1.7) \approx 0.045.$$

Thus

$$P(950 \le \text{number of success} \le 1050) \approx 0.955 - 0.045 = 0.91.$$

6 Covariance

Definition 6.1. The **covariance** of two random variables X, Y on a sample space Ω is the expected product of their deviations from their individual expected values, i.e.,

$$\operatorname{cov}(X, Y) = E\big((X - E(X))(Y - E(Y))\big),$$

a measure of the linear correlation between two random variables.

$$\begin{aligned}\operatorname{cov}(X, Y) &= E(XY - XE(Y) - E(X)Y + E(X)E(Y)) \\ &= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y).\end{aligned}$$

So X, Y are independent if and only if $\operatorname{cov}(X, Y) = 0$.

7 Limit Theorem

Proposition 7.1 (Markov's Inequality)**.** *If X is a nonnegative random variable, then for any value $\varepsilon > 0$,*

$$P\{X \ge \varepsilon\} \le \frac{E(X)}{\varepsilon}.$$

Proof.

$$\begin{aligned} E(X) &= \sum_{v \geq 0} vP\{X = v\} \geq \sum_{v \geq \varepsilon} vP\{X = v\} \\ &\geq \varepsilon \sum_{v \geq \varepsilon} P\{X = v\} = \varepsilon P\{X \geq \varepsilon\}. \end{aligned}$$

□

Proposition 7.2 (Chebyshev's Inequality)**.** *If X is a random variable with finite mean μ and variance σ^2, then for any value $\varepsilon > 0$,*

$$P\{|X - \mu| \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2}.$$

Proof. Since $(X - \mu)^2$ is a nonnegative random variable, Markov's inequality implies

$$P\{(X - \mu)^2 \geq \varepsilon^2\} \geq \frac{E((X - \mu)^2)}{\varepsilon^2}.$$

Note that $(X - \mu)^2 \geq \varepsilon^2$ if and only if $|X - \mu| \geq \varepsilon$. The desired inequality follows. □

Theorem 7.3 (Central Limit Theorem)**.** *Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables on a sample space Ω, each having mean μ and variance σ^2. Then the random variable*

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal distribution as $n \to \infty$. That is

$$\lim_{n \to \infty} P\left(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq y\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-x^2/2} \mathrm{d}x.$$

Proof. See any book on (advanced) probability theory. □

8 Page Rank

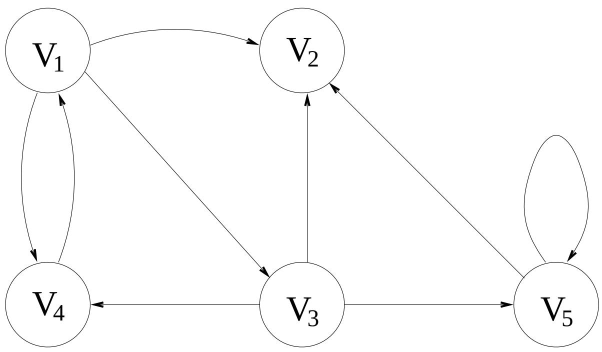
Internet can be viewed as a huge directed graph $G = (V, E)$ whose vertices are web pages and whose directed edges are links from one web page to the other.

A page rank is a kind of measure of importance of web pages. In practice, this measure is only given to a part of web pages of special interest. Anyway, we assume G is a subgraph of the huge internet graph. The importance of a page depends proportional to the number of pages linked to the page and the importance of those page linked the page.

Let A be the $V \times V$ adjacency matrix of digraph G, where (u, v)-entry of A, written a_{uv}, is the number of directed edges from u to v. A **page rank** is a nonnegative function $p : V \to \mathbb{R}$ such that

$$\begin{cases} p(v) = \sum_{u \in V} \frac{a_{uv}}{\operatorname{odeg}(u)} \cdot p(u), \quad v \in V, \\ \sum_{v \in V} p(v) = 1, \end{cases}$$

where $\operatorname{odeg}(u)$ is the out-degree of vertex u, the number of directed edges pointing away from u to all possible vertices, including u itself. If $\operatorname{odeg}(u) = 0$, i.e., u is not linked to any page, of course $a_{uv} = 0$, we assume $\frac{a_{uv}}{\operatorname{odeg}(u)} = 0$.



We introduce stochastic matrix $V \times V$ matrix $P = [p_{uv}]$ whose all row sums are 1, where

$$p_{uv} = \begin{cases} \frac{a_{uv}}{\operatorname{odeg}(u)} & \text{if } \operatorname{odeg}(u) \neq 0, \\ 1 & \text{if } \operatorname{odeg}(u) = 0, u = v, \\ 0 & \text{if } \operatorname{odeg}(u) = 0, u \neq v. \end{cases}$$

Let $\boldsymbol{p} = (p(v) : v \in V)$ be a row vector. Then the page rank problem is to find a vector in the simplex spanned by the coordinate vectors $\boldsymbol{e}_v$, $v \in V$, satisfying

$$\boldsymbol{p} = \boldsymbol{p}P, \quad \text{i.e.,} \quad \boldsymbol{p}(P - I) = \boldsymbol{0}.$$

So $\boldsymbol{p}$ is a left eigenvector of P for the eigenvalue 1.

Theorem 8.1 (Fundamental Theorem of Markov Chains)**.** *Let P be stochastic matrix whose row sums are* 1*. If every column of P is nonzero, then there exists a unique distribution $\boldsymbol{\pi}$ such that $\boldsymbol{\pi} = \boldsymbol{\pi}P$.*

Proof. Given an initial distribution $\boldsymbol{p}_0$. Define $\boldsymbol{p}_k = \boldsymbol{p}_{k-1}P$, $k \geq 1$. Then $\boldsymbol{p}_k = \boldsymbol{p}_0 P^k$ are distibutions. Let $\mathbf{1}$ denote the vector whose all entries are 1. Clearly, $P\mathbf{1} = \mathbf{1}$. Then

$$\langle \boldsymbol{p}_k, \mathbf{1} \rangle = \boldsymbol{p}_0 P^k \mathbf{1} = \boldsymbol{p}_0 \mathbf{1} = 1,$$

which shows that $\boldsymbol{p}_k$ is a distribution. Consider the average distribution

$$\boldsymbol{a}_k = (\boldsymbol{p}_0 + \boldsymbol{p}_1 + \cdots + \boldsymbol{p}_{k-1})/k = \boldsymbol{p}_0(I + P + \cdots + P^{k-1})/k.$$

Note that $\boldsymbol{a}_k(P - I) = (\boldsymbol{p}_1 + \cdots + \boldsymbol{p}_k)/k - (\boldsymbol{p}_0 + \boldsymbol{p}_1 + \cdots + \boldsymbol{p}_{k-1})/k$. Then $\boldsymbol{a}_k(P - I) = (\boldsymbol{p}_k - \boldsymbol{p}_0)/k \to \mathbf{0}$ $(k \to \infty)$. Thus

$$\boldsymbol{a}_k[P - I, \mathbf{1}] = [\boldsymbol{p}_k - \boldsymbol{p}_0, 1].$$

Recall that $\operatorname{rank}(P - I) = n - 1$. Let B denote the $n \times n$ submatrix of $[P - I, \mathbf{1}]$ by deleting its first column, and let $\boldsymbol{c}_k$ be the $(n-1)$-vector obtained from $(\boldsymbol{p}_k - \boldsymbol{p}_0)/k$ by deleting its first entry. Then $\operatorname{rank} B = n$ and $\boldsymbol{a}_k B = [\boldsymbol{c}_k, 1]$. Thus

$$\boldsymbol{a}_k = [\boldsymbol{c}_k, 1]B^{-1} \to [\mathbf{0}, 1]B^{-1} \ (k \to \infty).$$

$\square$