Random variables

Random variable: function from experiment state to real number. Denoted by capital letters, e.g. X. Not to confuse a function $X(s)$, which maps the experiment state to the real number, and the random variable X itself.

The probability function is defined on random variables.

Example (1) Flipping a coin.

The possible outcomes are head (H) and tail (T).

Sample space: $S = \{H, T\}$

Let us define a function $X(s)$, $s \in S$, such that:

$$X(s) = \begin{cases} 1 & s = H \\ 0 & s = T \end{cases}$$

$X(s)$ is a random variable that assigns values $\{0, 1\}$.

(2) Tossing a die: $S = \{1, 2, 3, 4, 5, 6\}$

A random variable defined on this sample space may be: $X = \begin{cases} 1 & \text{if } s \text{ is even} \\ 0 & \text{if } s \text{ is odd} \end{cases}$

Another random variable may be:

$$X(s) = \begin{cases} 1 & \text{if } s \text{ is even} \\ 0 & \text{if } s \text{ is odd} \\ -1 & \text{if } s \text{ is prime} \end{cases}$$

Discrete random variable: If a random variable $X(s)$ only takes discrete values, it is called a discrete random variable.

Example: $X(s)$ corresponding to flipping a coin or tossing a die.

Continuous random variable: Many of the experiment have continuous outcomes for example, measuring a person's weight, finding height of a new employee, and the year of graduation.

The outcome in this case takes any value in certain range.

For that, the sample space S is continuous due to the range of values $X(s)$.

The random variable X is not able to take discrete values in a certain range, rather, it can take any continuous random variable.

Cumulative distribution function (CDF): all discrete and continuous random variables have a cumulative distribution function (CDF) that is defined as:

$$F_X(x) = P(X \le x) \quad -\infty < x < \infty$$

The function $F_X(x)$ is called cumulative distribution function (CDF) of the random variable X.

Properties of a CDF:

(1) $0 \le F_X(x) \le 1$

(2) $F_X(-\infty) = P(X \le -\infty) = 0$

(3) $F_X(\infty) = P(X \le \infty) = 1$

Probability density function (pdf): $f_X(x)$ is called probability density function (PDF) of a continuous random variable X if it satisfies:

$$f_X(x) = \frac{d}{dx} F_X(x) \quad -\infty < x < \infty$$

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du$$

where u random variable X (discrete, the probability mass function) f_X is defined as:

$$f_X(x) = \sum_{u \le x} P(X = u)\, \delta(x - u)$$

Note $x, x_1, x_2, \ldots$ are the values taken by the random variable X. $\delta(x)$ is Dirac's delta function, $\delta(x) = \infty$ at $x = 0$ and zero elsewhere, and $\int_{-\infty}^{\infty} \delta(x)\, dx = 1$.

Example (1) plot the pdf and CDF for a random variable which defines the number of heads in the experiment of flipping a fair coin.

Sample space: $S = \{HH, HT, TH, TT\}$

The random variable $X(s)$ is number of heads

		$x \le y$ and $x \ge 0$
HH	2	x has been achieved
HT	1	x has been achieved
TH	1	x is distinct random variable
TT	0	

$$f_X(0) = P(X = 0) = \frac{1}{4}$$

$$f_X(1) = P(X = 1) = \frac{1}{2}$$

$$f_X(2) = P(X = 2) = \frac{1}{4}$$

So, the pdf plot is:

CDF: $F_X(0) = P(X \le 0)$

$$= P(X = 0) = \frac{1}{4}$$

$$F_X(1) = P(X \le 1)$$

$$= P(X = 0) + P(X = 1)$$

$$= \frac{1}{4} + \frac{1}{2} = \frac{3}{4}$$

$$F_X(2) = P(X \le 2)$$

$$= P(X = 0) + P(X = 1) + P(X = 2)$$

$$= \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1$$

(2) For a continuous random variable X, the CDF is given below:

$$F_X(x) = \begin{cases} 0 & x < 0 \\ \frac{x^2}{4} & 0 \le x < 2 \\ 1 & x \ge 2 \end{cases}$$

Find the CDF?

$$f_X(x) = \frac{d}{dx} F_X(x) \text{ for } -\infty < x < \infty$$

$$f_X(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{2} & 0 \le x < 2 \\ 0 & x \ge 2 \end{cases}$$

CDF: $F_X(x) = \int_{-\infty}^{x} f_X(u)\, du$ for continuous random variable

$F_X(x) = \sum_{u \le x} f_X(u)$ for discrete random variable

Expectation and moments: Mean or expectation of a random variable X is defined as:

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx \quad \text{for continuous random variable}$$

$$E(X) = \sum_{x} x f_X(x) \quad \text{for discrete random variable}$$

We provide the long-term average of the random variable.

Assume we have a large number N of the random variable X we expect value as the number of occurrences of experiment times x.

We also measure as the first moment of X.

In general, we can measure the nth moment of X as:

$$E(X^n) = \int_{-\infty}^{\infty} x^n f_X(x)\, dx \quad \text{for CDF}$$

$$E(X^n) = \sum_{x} x^n f_X(x) \quad \text{for PMF}$$

Let $E(X) = m$, then

$$E[(X - m)^n] = \int_{-\infty}^{\infty} (x - m)^n f_X(x)\, dx$$

It called the nth central moment of the rv X, because the moment of X taken relative to the mean.

When $n = 2$, the central moment is called the variance of the rv X:

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x - m)^2 f_X(x)\, dx \quad \text{for CDF}$$

$$\sigma_X^2 = \sum_{x} (x - m)^2 f_X(x) \quad \text{for PMF}$$

Variance σ_X^2 provides a measure of the dispersion of the random variable X about mean, if variance is small, the random variable X cluster from its mean.

Higher the value of σ_X^2 larger is the difference between values taken by the random variable.

Standard deviation σ_X is defined as:

With the help of σ_X, we can find dispersion of random variable X relative to its mean.

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x - m)^2 f_X(x)\, dx = \int_{-\infty}^{\infty} (x^2 - 2mx + m^2) f_X(x)\, dx$$

$$= \int_{-\infty}^{\infty} x^2 f_X(x)\, dx - 2m \int_{-\infty}^{\infty} x f_X(x)\, dx + m^2 \int_{-\infty}^{\infty} f_X(x)\, dx$$

$$\sigma_X^2 = E(X^2) - (E(X))^2$$

Some useful probability distributions

1) Binomial distribution: It is a discrete probability distribution for the number of successes in a sequence of n independent experiments, each success of which that is successes with prob p or failure with prob $1-p$.

A single successful failure experiment is called a Bernoulli trial. The binomial distribution is a sum of independent trials of Bernoulli by the binomial theorem:

$$(a + b)^n = \binom{n}{0} a^0 b^{n-0} + \binom{n}{1} a^1 b^{n-1} + \binom{n}{2} a^2 b^{n-2} + \ldots + \binom{n}{n} a^n b^{n-n}$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad \text{is the binomial coefficient}$$

$$\binom{n}{0} = 1 \qquad \binom{n}{n} = 1$$

The probability distribution can be calculated as:

Let X be a random variable, X is a binomial random variable with parameters n and p, X can take every integer between zero to n inclusive. Hence there are $n+1$ possible values of distributing n successes in a sequence of n trials.

Example (1) Distribution of a distribution with equally likely outcomes. If the sample space has n elements, the distribution of random variable X is:

PMF of random distribution is:

$$f_X(x) = \begin{cases} \frac{1}{n} & \text{for } x \in \{1, 2, \ldots, n\} \\ 0 & \text{otherwise} \end{cases}$$

The CDF:

$$F_X(x) = \sum_{k=1}^{\lfloor x \rfloor} f_X(k) = \frac{\lfloor x \rfloor}{n} \quad x \in [1, n]$$

$$f_X(x) = \frac{1}{n} \langle \delta(x) + \delta(x-1) + \ldots + \delta(x-n) \rangle$$

Normal distribution: It is the most important continuous probability distribution. It is used in many fields. Many random variables are normally distributed, or can be approximated to a normal distribution, they are approximately normal due to the central limit theorem. Normal distribution is parameterized, it is very useful due to central limit theorem (this is discussed later).

A random variable X is said to be normal distributed or Gaussian distributed if it takes the form:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

The normal distribution has the following features:

1) The mean and σ is the standard deviation.

$$E(X) = m, \quad E[(x-m)^2] = \sigma^2$$

2) We denoted this that X is a random variable, X is the normal (m, σ^2) form $\sim$ X is equal to $\sim$

$$\int_{-\infty}^{\infty} f_X(x)\, dx = 1$$

3) The PDF is symmetric with respect to $x = m$, it has a bell shaped curve.

4) The pdf has a zero area that $m \pm \sigma$ or $m \pm 2\sigma$ area.

5) The CDF is:

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(u-m)^2}{2\sigma^2}}\, du$$

Let $u = \frac{x-m}{\sigma}$, $du = \frac{dx}{\sigma}$ $\quad$ σ is the standard deviation

$$F_X(x) = \int_{-\infty}^{\frac{x-m}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}\, du$$

$$F_X(x) = Q\left(\frac{x-m}{\sigma}\right)$$

The function $Q\left(\frac{x-m}{\sigma}\right)$ is frequently used for finding the area under the tail of the normal distribution.

For $m = 0$ and $\sigma = 1$, $f_X(x)$ is known as the standard normal distribution.

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad -\infty < x < \infty$$

The CDF of the standard normal random variable

$$f_X(x) = Q(x)$$

6) The standard normal is used for normal distribution, it is used to standardize the normal distribution:

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}\, du$$

7) Finding about the normal distribution:

$$P = P_X(x_1 \le X \le x_2) = P_X(x_1 \le X \le x_2)$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x_2} e^{-\frac{(x-m)^2}{2\sigma^2}} dx\right) - \left(\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x_1} e^{-\frac{(x-m)^2}{2\sigma^2}} dx\right)$$

$$= \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x_2-m}{\sigma}} e^{-\frac{u^2}{2}} du\right) - \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x_1-m}{\sigma}} e^{-\frac{u^2}{2}} du\right)$$

$$P = \frac{1}{\sqrt{2\pi}} \left(\int_{-\infty}^{\frac{x_2-m}{\sigma}} e^{-\frac{u^2}{2}} du - \int_{-\infty}^{\frac{x_1-m}{\sigma}} e^{-\frac{u^2}{2}} du\right)$$

$$P = \frac{1}{\sqrt{2\pi}} \left(\int_{\frac{x_1-m}{\sigma}}^{\frac{x_2-m}{\sigma}} e^{-\frac{u^2}{2}} du\right) = \frac{1}{\sqrt{2\pi}} \int_{\frac{x_1-m}{\sigma}}^{\frac{x_2-m}{\sigma}} e^{-\frac{u^2}{2}} du$$

Central limit theorem

Let $X_1, X_2, X_3, \ldots$ be independent and identically distributed (iid) random variables with mean m and finite variance σ^2.

Let Y_n be defined as the normalized sum of the first n variables:

$$Y_n = \frac{X_1 + \ldots + X_n}{\sqrt{n}}$$

The random variable Y_n is linearly dependent, so obtaining the mean of a random variable X from a number of experiments as the limit of n tends to infinity:

$$E(Y_n) = \frac{m}{\sqrt{n}} \quad \text{and} \quad \sigma^2(Y_n) = \sigma^2$$

$$\lim_{n \to \infty} E(Y_n) = \lim_{n \to \infty} \frac{m}{\sqrt{n}} = 0$$

when n is around, we can calculate for the mean m, we can find its equivalent value to 0 and the variance remains as it is, minimum. It means the uncertainty in the outcomes of experiment is increasing.

The central limit theorem states that the sum of statistically independent and identically distributed (iid) random variables approaches to a normal (Gaussian) distribution as n tends to infinity.

The sum tends to a normal distribution regardless of the original distribution of the individual random variables. The distribution of the population is not necessarily a normal distribution.