First and First

Random itself has the meaning like *unknown*, and *unpredictable*. It'll be a contradiction if we can define what is *randomness*. Or, from different aspect, we can only use *unknown* to define *random*.

Random Variable

A *random variable* is a function that maps the outcome of a random experiment to a real number. We use $\mathcal{C}$ to denote the set of all outcomes; c to denote a single outcome.

For example, let $C = \{\text{GPT}, \text{GAN}, \text{BERT}, \text{YOLO}\}$, then a random varialbe X can be

$$X(\text{GPT}) = 0, X(\text{GAN}) = 1, X(\text{BERT}) = 2, X(\text{YOLO}) = 3.$$

You can change to any number you want.

We cannot observe a random variable itself, i.e., the mapping X is unobservable. We can only define the mapping, and then observe the result of applying this mapping to an experiment outcome.

The *realization* of a random variable is the result of applying the random variable (i.e., mapping) to an observed outcome of a random experiment. This is what we actually observe.

Typically, we use lowercase to denote the realized number; uppercase to denote the random variable. e.g., x is a realization of X.

The *space* or *range* of X is a set of real numbers $\mathcal{D} = \{X(c) : c \in C\}$.

Probability Mass Function

A random variable X is said to be *descrete* if its space $\mathcal{D}$ is either finite or countable.

Let X be a discrete random variable with space $\mathcal{D}$. The *probability mass function* of X, $p_X(d_i)$, is defined by

$$p_X(d_i) = P[\{c : X(c) = d_i\}] = P[X = d_i],$$

for all $d_i \in \mathcal{D}$.

The induced probability distribution, $P_X(\cdot)$, of X is

$$P_X(D) = \sum_{d_i \in D} p_X(d_i) = \sum_{d_i \in D} P[\{c : X(c) = d_i\}] = \sum_{d_i \in D} P[X = d_i],\ D \subset \mathcal{D}$$

Note that the notation $P[X = d_i]$ is an abbreviation, since the outcome c is not actually important here.

Cumulative Distribution Function

The *cumulative distribution function*, $F_X(x)$, of X is defined by

$$F_X(x) = P_X((-\infty, x]) = P[\{c : X(c) \leq x\}] = P(X \leq x).$$

Cdf is also simply called the *distribution function*.

Probability Density Function

A random variable X is said to be *continuous* if its cdf $F_X(x)$ is continuous for all $x \in \mathbb{R}$.

Let X be a continuous random variable with interval $\mathcal{D} \subset \mathbb{R}$ as space. The *probability density function* of X, $f_X(x)$, is a function that satisfies

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt.$$

When there exits such a function $f_X(x)$, X is also called an *absolutely continuous* random variable.

If $f_X(x)$ is also continuous, we have

$$\frac{d}{dx}F_X(x) = f_X(x)$$

by the Fundamental Theorem of Calculus. Note that for any continuous random variable X, there are no points of discrete mass, hence

$$P(X = x) = 0,$$

for all $x \in \mathbb{R}$.

From this, we can also infer that

$$P(a < X \leq b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X < b)$$

Different random variable can have the same cdf

Let X has be a random variable that stands for a real random number randomly choosed from the interval $(0, 1)$, and we simply use the sample as the assigned number. In this case, the domain is $\mathcal{D} = (0, 1)$. Assign a probability on X,

$$P_X[(a, b)] = b - a, \text{ for } 0 < a < b < 1$$

Then the pdf of X is

$$f_X(x) = \begin{cases} 1 \; 0 < x < 1 \\ 0 \text{ elsewhere} \end{cases}$$

It's easy to show that the cdf is

$$F_X(x) = P(X \leq x) = \begin{cases} 0 \text{ if } x < 0 \\ x \text{ if } 0 \leq x < 1 \\ 1 \text{ if } x \geq 1 \end{cases}$$

Now consider $Y = 1 - X$,

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(1 - X \leq y) = P(X \geq 1 - y) = 1 - P(X < 1 - y) \\ &= \begin{cases} 0 \text{ if } y < 0 \\ y \text{ if } 0 \leq y < 1 \\ 1 \text{ if } 1 \leq y \end{cases} \end{aligned}$$

In this case, we said X and Y are equal in distribution and denote by $X \overset{D}{=} Y$.

Expectation

The *expectation* of X is defined by

$$E[X] = \begin{cases} \sum x_i p(x_i) \text{ if } X \text{ is discrete with pmf } p(x) \text{ , and} \sum |x| p(x) < \infty \\ \int x f(x) dx \text{ if } X \text{ is continuous with pdf } f(x) \text{ , and} \int |x| f(x) dx < \infty \end{cases}$$

Expectation is also called *mean*, or *expected value*, and mostly denoted by μ.

The expection can reflect the transformation of random variable. Let $Y = g(X)$, then

$$E(Y) = E(g(X)) = \sum g(x)p(x)$$
$$E(Y) = E(g(X)) = \int g(x)f(x)dx$$

The expection is linear with respect to random variable,

$$E[k_1 g_1(X) + k_2 g_2(X)] = k_1 E[g_1(X)] + k_2 E[g_2(X)]$$

Variance and Standard Deviation

Let X be a random variable with finite mean μ and $E\left[(X-\mu)^2\right]$ is also finite. The variance of X is defined by

$$\text{Var }[X] = E\left[(X-\mu)^2\right]$$

Variance is mostly denoted by σ^2. The single σ is called the *standard deviation*. The number σ is sometimes interpreted as a measure of the dispersion of the points of the space relative to the mean value μ.

Note that

$$\begin{aligned}\sigma^2 = E\left[(X-\mu)^2\right] &= E(X^2 - 2X\mu + \mu^2) \\ &= E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - \mu^2\end{aligned}$$

Random Vector

Consider two random variables X_1 and X_2 on the same sample space $\mathcal{C}$, that they assign each element c of $\mathcal{C}$ one and only one ordered pair of numbers $X_1(c) = x_1$, $X_2(c) = x_2$. Then we say that (X_1, X_2) is a random vector. The *space* of (X_1, X_2) is the set of ordered pairs $\mathcal{D} = \{(x_1, x_2) : x_1 = X_1(c), x_2 = X_2(c), c \in \mathcal{C}\}$.

Probability Mass Function

A discrete random vector (X_1, X_2) with finite or countable space $\mathcal{D}$. The *joint probability mass function* of (X_1, X_2), $p_{X_1,X_2}(x_1, x_2)$, is defined by

$$p_{X_1,X_2}(x_1, x_2) = P[X_1 = x_1, X_2 = x_2]$$

for all $(x_1, x_2) \in \mathcal{D}$.

Cumulative Distribution Function

The cumulative distribution function of (X_1, X_2), $F_{X_1,X_2}(x_1, x_2)$, is defined by

$$F_{X_1,X_2}(x_1, x_2) = P[\{X_1 \leq x_1\} \cap \{X_2 \leq x_2\}],$$

for all $(x_1, x_2) \in \mathbb{R}$. This is also called *joint cumulative distribution function*.

We'll also abbreviate $P[\{X_1 \leq x_1\} \cap \{X_2 \leq x_2\}]$ to $P[X_1 \leq x_1, X_2 \leq x_2]$.

Probability Density Function

A random vector (X_1, X_2) with space $\mathcal{D}$ is said to be continuous if

$$F_{X_1,X_2}(x_1, x_2) = P[\{X_1 \leq x_1\} \cap \{X_2 \leq x_2\}]$$

is continuous.

The of (X_1, X_2), $f_{X_1,X_2}(x_1, x_2)$, is defined to satisfy

$$F_{X_1,X_2}(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f_{X_1,X_2}(w_1, w_2) dw_1 dw_2$$

for all $(x_1, x_2) \in \mathbb{R}$. Then

$$\frac{\partial^2 F_{X_1,X_2}(x_1, x_2)}{\partial x_1 \partial x_2} = f_{X_1,X_2}(x_1, x_2)$$

For an event $A \subset \mathcal{D}$, we have

$$P[(X_1, X_2) \in A] = \int \int_A f_{X_1,X_2}(x_1, x_2) dx_1 dx_2$$

Marginals

Let (X_1, X_2) be a random vector. Recall that

$$\begin{aligned} \{X_1 \leq x_1\} &= \{c : X_1(c) \leq x_1\} = \{c : X_1(c) \leq x_1\} \cap \{c : -\infty < X_2 < \infty\} \\ &= \{X_1 \leq x_1, -\infty < X_2 < \infty\}, \end{aligned}$$

hence,

$$F_{X_1}(x_1) = P[X_1 \leq x_1, -\infty < X_2 < \infty],$$

for all $x_1 \in \mathbb{R}$. By the property of cdf, we can get

$$F_{X_1}(x_1) = \lim_{x_2 \to \infty} F_{X_1,X_2}(x_1, x_2).$$

This is exactly where we connect the cdf, pdf, pmf between random variable and random vector.

Discrete

For discrete (X_1, X_2). Let $\mathcal{D}_{X_1}$ be the support of X_1, i.e., $\mathcal{D}_{X_1} = \left\{x \in \mathcal{D} : p_{X_1}(x) \neq 0\right\}$ where $\mathcal{D}$ is the space of X_1. For $x_1 \in \mathcal{D}_{X_1}$

$$\begin{aligned} F_{X_1}(x_1) &= P[X_1 \leq x_1, -\infty < X_2 < \infty] \\ &= \textstyle\sum\sum_{w_1 \leq x_1, -\infty < x_2 < \infty} p_{X_1,X_2}(w_1, x_2) \\ &= \sum_{w_1 \leq x_1} \left\{ \sum_{x_2 < \infty} p_{X_1,X_2}(w_1, x_2) \right\} \end{aligned}$$

By uniqueness of cdfs, we know the pmf of X_1 must be

$$p_{X_1}(x_1) = \sum_{x_2 < \infty} p_{X_1,X_2}(x_1, x_2),$$

for all $x_1 \in \mathcal{D}_{X_1}$. This is called the *marginal pmf* of X_1. We can get similar result for X_2.

Continuous

For continuous (X_1, X_2). We use the same notation as the discrete one. Then

$$F_{X_1}(x_1) = \int_{-\infty}^{x_1} \int_{-\infty}^{\infty} f_{X_1,X_2}(w_1, x_2) dx_2 dw_1 = \int_{-\infty}^{x_1} \left\{ \int_{-\infty}^{\infty} f_{X_1,X_2}(w_1, x_2) dx_2 \right\} dw_1,$$

for all $x_1 \in \mathcal{D}_{X_1}$. The pdf of X_1 must be

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2) dx_2$$

Expectation

From above, we have

$$\begin{aligned} E(X_1) &= \int x_1 f_{X_1}(x_1) dx_1 \\ &= \int x_1 \left\{ \int f_{X_1,X_2}(x_1, x_2) dx_2 \right\} dx_1 \\ &= \int \int x_1 f_{X_1,X_2}(x_1, x_2) dx_2 dx_1 \end{aligned}$$

Let $\mathbf{X} = (X_1, X_2)'$ be a random vector. The expectation $E(\mathbf{X})$ exists if the expectations X_1 and X_2 exist, and, is computed by

$$E[\mathbf{X}] = \begin{bmatrix} E(X_1) \\ E(X_2) \end{bmatrix}$$

It's easy to verify that $E[\mathbf{X}]$ is linear.

Conditional Distributions and Expectations

Let $f_{X_1,X_2}(x_1, x_2)$ be the joint pdf of two random variables X_1 and X_2. Let $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$ denote the marginal pdf of $X1$ and X_2, respectively. Observe that

$$\int_{-\infty}^{-\infty} \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_1}(x_1)} dx_2 = \frac{1}{f_{X_1}(x_1)} \int_{-\infty}^{-\infty} f_{X_1,X_2}(x_1, x_2) dx_2 = \frac{1}{f_{X_1}(x_1)} f_{X_1}(x_1) = 1$$

That is, $\frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)}$ satisfies the properties of a pdf of one continous random variable on the support of $X1$. We called this the *conditional pdf* of X_2, given $X_1 = x_1$.

The *conditional probability* is then defined by

$$P(a < X_2 < b \mid X_1 = x_1) = \int_a^b \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_1}(x_1)} dx_2,$$

furthermore, the *conditional expectation*,

$$E[X_2 \mid x_1] = \int_{-\infty}^{-\infty} x_2 \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_1}(x_1)} dx_2,$$

the *conditional variance*,

$$\begin{aligned} \text{Var } [X_2 \mid x_1] &= E[(X_2 - E[X_2 | x_1])^2 | x_1] \\ &= \int_{-\infty}^{-\infty} (x_2 - E[X_2 \mid x_1])^2 \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_1}(x_1)} dx_2 \\ &= E(X_2^2 \mid x_1) - [E(X_2 \mid x_1)]^2 \end{aligned}$$

and, for $u(X_2)$ be a function of X_2,

$$E[u(X_2) \mid x_1] = \int_{-\infty}^{-\infty} u(x_2) \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)} dx_2,$$

When the realization x_1 is not that important, we'll denote the above concepts by $E[X_2 \mid X_1]$, Var $[X_2 \mid X_1]$, and $E[u(X_2) \mid X_1]$.

Important Theorem

Let (X_1, X_2) be a random vector such that the variance of X_2 is finite. Then

$$\begin{aligned} &E[E(X_2 \mid X_1)] = E(X_2), \\ &\text{Var }[E(X_2 \mid X_1)] \leq \text{Var }(X_2) \end{aligned}$$

Proof: Consider

$$\begin{aligned} E(X_2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_2 f(x_1,x_2) dx_2 dx_1 \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x_2 \frac{f(x_1,x_2)}{f_1(x_1)} dx_2 \right] f_1(x_1) dx_1 \\ &= \int_{-\infty}^{\infty} E(X_2 \mid x_1) f_1(x_1) dx_1 \\ &= E[E(X_2 \mid X_1)], \end{aligned}$$

where $f(x_1,x_2) = f_{X_1,X_2}(x_1,x_2)$ and $f_1(x_1) = f_{X_1}(x_1) = \int f(x_1,x_2) dx_2$.

For the second result, let $\mu_2 = E(X_2)$, consider

$$\begin{aligned} \text{Var }(X_2) &= E\left[(X_2 - \mu_2)^2\right] \\ &= E\left[(X_2 - E(X_2 \mid X_1) + E(X_2 \mid X_1) - \mu_2)^2\right] \\ &= E\left[(X_2 - E(X_2 \mid X_1)^2\right] + E\left[(E(X_2 \mid X_1) - \mu_2)^2\right] \\ &\quad +2E[(X_2 - E(X_2 \mid X_1))(E(X_2 \mid X_1) - \mu_2)], \end{aligned}$$

the last term is equal to

$$\begin{aligned} &2\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_2 - E(X_2 \mid X_1))(E(X_2 \mid X_1) - \mu_2) f(x_1,x_2) dx_2 dx_1 \\ &= 2\int_{-\infty}^{\infty} (E(X_2 \mid X_1) - \mu_2) \left\{ \int_{-\infty}^{\infty} (x_2 - E(X_2 \mid X_1)) \frac{f(x_1,x_2)}{f_1(x_1)} dx_2 \right\} f_1(x_1) dx_1, \end{aligned}$$

where the integral inside the curly braces is zero. Hence the variance of X_2 is

$$\text{Var }(X_2) = E\left[(X_2 - E(X_2 \mid X_1)^2\right] + E\left[(E(X_2 \mid X_1) - \mu_2)^2\right]$$

The first term is non negative, the second term is

$$E\left[(E(X_2 \mid X_1) - E(X_2))^2\right] = E\left[(E(X_2 \mid X_1) - E[E(X_2 \mid X_1)])^2\right] = \text{Var }[E(X_2 \mid X_1)],$$

we get the result Var $[E(X_2 \mid X_1)] \leq$ Var (X_2).

This theorem tells us that,

Random Sample