1.1 Conditional probability and independence

Let B be an event with non-zero probability. The conditional probability of any event A given B is defined as,

$$P(A|B) \triangleq \frac{P(A \cap B)}{P(B)}$$

In other words, $P(A|B)$ is the probability measure of the event A after observing the occurrence of event B. Two events are called independent if and only if $P(A \cap B) = P(A)P(B)$ (or equivalently, $P(A|B) = P(A)$). Therefore, independence is equivalent to saying that observing B does not have any effect on the probability of A.

2.1 Cumulative distribution functions

In order to specify the probability measures used when dealing with random variables, it is often convenient to specify alternative functions (CDFs, PDFs, and PMFs) from which the probability measure governing an experiment immediately follows. In this section and the next two sections, we describe each of these types of functions in turn.

A **cumulative distribution function (CDF)** is a function $F_X : \mathbb{R} \to [0, 1]$ which specifies a probability measure as,

$$F_X(x) \triangleq P(X \leq x). \tag{1}$$

By using this function one can calculate the probability of any event in $\mathcal{F}$.[3] Figure **??** shows a sample CDF function.

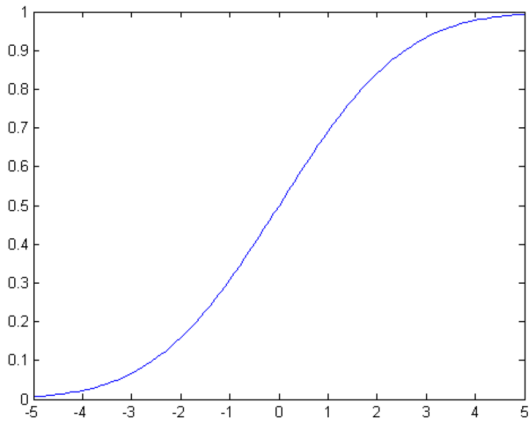

Figure 1: A cumulative distribution function (CDF).

- $0 \leq F_X(x) \leq 1$.
- $\lim_{x \to -\infty} F_X(x) = 0$.
- $\lim_{x \to \infty} F_X(x) = 1$.
- $x \leq y \Longrightarrow F_X(x) \leq F_X(y)$.

2.2 Probability mass functions

When a random variable X takes on a finite set of possible values (i.e., X is a discrete random variable), a simpler way to represent the probability measure associated with a random variable is to directly specify the probability of each value that the random variable can assume. In particular, a *probability mass function (PMF)* is a function $p_X : \Omega \to \mathbb{R}$ such that

$$p_X(x) \triangleq P(X = x).$$

In the case of discrete random variable, we use the notation $Val(X)$ for the set of possible values that the random variable X may assume. For example, if $X(\omega)$ is a random variable indicating the number of heads out of ten tosses of coin, then $Val(X) = \{0, 1, 2, \ldots, 10\}$.

Properties:

- $0 \leq p_X(x) \leq 1$.
- $\sum_{x \in Val(X)} p_X(x) = 1$.
- $\sum_{x \in A} p_X(x) = P(X \in A)$.

2.3 Probability density functions

For some continuous random variables, the cumulative distribution function $F_X(x)$ is differentiable everywhere. In these cases, we define the **Probability Density Function** or **PDF** as the derivative of the CDF, i.e.,

$$f_X(x) \triangleq \frac{dF_X(x)}{dx}. \tag{2}$$

Note here, that the PDF for a continuous random variable may not always exist (i.e., if $F_X(x)$ is not differentiable everywhere).

According to the properties of differentiation, for very small Δx,

$$P(x \leq X \leq x + \Delta x) \approx f_X(x)\Delta x. \tag{3}$$

Properties:

- $f_X(x) \geq 0$.
- $\int_{-\infty}^{\infty} f_X(x) = 1$.
- $\int_{x \in A} f_X(x)dx = P(X \in A)$.

2.4 Expectation

Suppose that X is a discrete random variable with PMF $p_X(x)$ and $g : \mathbb{R} \longrightarrow \mathbb{R}$ is an arbitrary function. In this case, $g(X)$ can be considered a random variable, and we define the **expectation** or **expected value** of $g(X)$ as

$$E[g(X)] \triangleq \sum_{x \in Val(X)} g(x)p_X(x).$$

If X is a continuous random variable with PDF $f_X(x)$, then the expected value of $g(X)$ is defined as,

$$E[g(X)] \triangleq \int_{-\infty}^{\infty} g(x)f_X(x)dx.$$

Intuitively, the expectation of $g(X)$ can be thought of as a "weighted average" of the values that $g(x)$ can taken on for different values of x, where the weights are given by $p_X(x)$ or $f_X(x)$. As a special case of the above, note that the expectation, $E[X]$ of a random variable itself is found by letting $g(x) = x$; this is also known as the **mean** of the random variable X.

Properties:

- $E[a] = a$ for any constant $a \in \mathbb{R}$.
- $E[af(X)] = aE[f(X)]$ for any constant $a \in \mathbb{R}$.
- (Linearity of Expectation) $E[f(X) + g(X)] = E[f(X)] + E[g(X)]$.
- For a discrete random variable X, $E[1\{X = k\}] = P(X = k)$.

2.5 Variance

The **variance** of a random variable X is a measure of how concentrated the distribution of a random variable X is around its mean. Formally, the variance of a random variable X is defined as

$$Var[X] \triangleq E[(X - E(X))^2]$$

Using the properties in the previous section, we can derive an alternate expression for the variance:

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

where the second equality follows from linearity of expectations and the fact that $E[X]$ is actually a constant with respect to the outer expectation.

Properties:

- $Var[a] = 0$ for any constant $a \in \mathbb{R}$.
- $Var[af(X)] = a^2 Var[f(X)]$ for any constant $a \in \mathbb{R}$.

Example Calculate the mean and the variance of the uniform random variable X with PDF $f_X(x) = 1, \ \forall x \in [0,1]$, 0 elsewhere.

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx = \int_0^1 x dx = \frac{1}{2}.$$

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) dx = \int_0^1 x^2 dx = \frac{1}{3}.$$

$$Var[X] = E[X^2] - E[X]^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

2.6 Some common random variables

Discrete random variables

- $X \sim Bernoulli(p)$ (where $0 \leq p \leq 1$): one if a coin with heads probability p comes up heads, zero otherwise.

$$p(x) = \begin{cases} p & \text{if } p = 1 \\ 1 - p & \text{if } p = 0 \end{cases}$$

- $X \sim Binomial(n, p)$ (where $0 \leq p \leq 1$): the number of heads in n independent flips of a coin with heads probability p.

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

- $X \sim Geometric(p)$ (where $p > 0$): the number of flips of a coin with heads probability p until the first heads.

$$p(x) = p(1-p)^{x-1}$$

- $X \sim Poisson(\lambda)$ (where $\lambda > 0$): a probability distribution over the nonnegative integers used for modeling the frequency of rare events.

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

Continuous random variables

- $X \sim Uniform(a, b)$ (where $a < b$): equal probability density to every value between a and b on the real line.

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

- $X \sim Exponential(\lambda)$ (where $\lambda > 0$): decaying probability density over the nonnegative reals.

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

- $X \sim Normal(\mu, \sigma^2)$: also known as the Gaussian distribution

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

Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
$Binomial(n, p)$	$\binom{n}{k} p^k (1-p)^{n-k}$ for $0 \leq k \leq n$	np	npq
$Geometric(p)$	$p(1-p)^{k-1}$ for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$e^{-\lambda}\lambda^x / x!$ for $k = 1, 2, \ldots$	λ	λ
$Uniform(a, b)$	$\frac{1}{b-a} \quad \forall x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x} \quad x \geq 0, \lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

3.1 Joint and marginal distributions

Suppose that we have two random variables X and Y. One way to work with these two random variables is to consider each of them separately. If we do that we will only need $F_X(x)$ and $F_Y(y)$. But if we want to know about the values that X and Y assume simultaneously during outcomes of a random experiment, we require a more complicated structure known as the **joint cumulative distribution function** of X and Y, defined by

$$F_{XY}(x, y) = P(X \leq x, Y \leq y)$$

It can be shown that by knowing the joint cumulative distribution function, the probability of any event involving X and Y can be calculated.

The joint CDF $F_{XY}(x, y)$ and the joint distribution functions $F_X(x)$ and $F_Y(y)$ of each variable separately are related by

$$\begin{aligned} F_X(x) &= \lim_{y \to \infty} F_{XY}(x, y)dy \\ F_Y(y) &= \lim_{x \to \infty} F_{XY}(x, y)dx. \end{aligned}$$

Here, we call $F_X(x)$ and $F_Y(y)$ the **marginal cumulative distribution functions** of $F_{XY}(x, y)$.

Properties:

- $0 \leq F_{XY}(x, y) \leq 1$.
- $\lim_{x,y \to \infty} F_{XY}(x, y) = 1$.
- $\lim_{x,y \to -\infty} F_{XY}(x, y) = 0$.
- $F_X(x) = \lim_{y \to \infty} F_{XY}(x, y)$.

3.2 Joint and marginal probability mass functions

If X and Y are discrete random variables, then the **joint probability mass function** $p_{XY} : \mathbb{R} \times \mathbb{R} \to [0, 1]$ is defined by

$$p_{XY}(x, y) = P(X = x, Y = y).$$

Here, $0 \leq P_{XY}(x, y) \leq 1$ for all x, y, and $\sum_{x \in Val(X)} \sum_{y \in Val(Y)} P_{XY}(x, y) = 1$.

How does the joint PMF over two variables relate to the probability mass function for each variable separately? It turns out that

$$p_X(x) = \sum_y p_{XY}(x, y).$$

and similarly for $p_Y(y)$. In this case, we refer to $p_X(x)$ as the **marginal probability mass function** of X. In statistics, the process of forming the marginal distribution with respect to one variable by summing out the other variable is often known as "marginalization."

3.3 Joint and marginal probability density functions

Let X and Y be two continuous random variables with joint distribution function F_{XY}. In the case that $F_{XY}(x, y)$ is everywhere differentiable in both x and y, then we can define the **joint probability density function**,

$$f_{XY}(x, y) = \frac{\partial^2 F_{XY}(x, y)}{\partial x \partial y}.$$

Like in the single-dimensional case, $f_{XY}(x, y) \neq P(X = x, Y = y)$, but rather

$$\iint_{x \in A} f_{XY}(x, y) dx dy = P((X, Y) \in A).$$

Note that the values of the probability density function $f_{XY}(x, y)$ are always nonnegative, but they may be greater than 1. Nonetheless, it must be the case that $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) = 1$.

Analagous to the discrete case, we define

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy,$$

as the **marginal probability density function** (or **marginal density**) of X, and similarly for $f_Y(y)$.

3.4 Conditional distributions

Conditional distributions seek to answer the question, what is the probability distribution over Y, when we know that X must take on a certain value x? In the discrete case, the conditional probability mass function of X given Y is simply

$$p_{Y|X}(y|x) = \frac{p_{XY}(x, y)}{p_X(x)},$$

assuming that $p_X(x) \neq 0$.

In the continuous case, the situation is technically a little more complicated because the probability that a continuous random variable X takes on a specific value x is equal to zero[4]. Ignoring this technical point, we simply define, by analogy to the discrete case, the *conditional probability density* of Y given $X = x$ to be

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)},$$

provided $f_X(x) \neq 0$.

3.5 Bayes's rule

A useful formula that often arises when trying to derive expression for the conditional probability of one variable given another, is **Bayes's rule**.

In the case of discrete random variables X and Y,

$$P_{Y|X}(y|x) = \frac{P_{XY}(x, y)}{P_X(x)} = \frac{P_{X|Y}(x|y)P_Y(y)}{\sum_{y' \in Val(Y)} P_{X|Y}(x|y')P_Y(y')}.$$

If the random variables X and Y are continuous,

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)} = \frac{f_{X|Y}(x|y)f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y}(x|y')f_Y(y')dy'}.$$

3.7 Expectation and covariance

Suppose that we have two discrete random variables X, Y and $g : \mathbf{R}^2 \longrightarrow \mathbf{R}$ is a function of these two random variables. Then the expected value of g is defined in the following way,

$$E[g(X,Y)] \triangleq \sum_{x \in Val(X)} \sum_{y \in Val(Y)} g(x,y) p_{XY}(x,y).$$

For continuous random variables X, Y, the analogous expression is

$$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y) dx dy.$$

We can use the concept of expectation to study the relationship of two random variables with each other. In particular, the **covariance** of two random variables X and Y is defined as

$$Cov[X,Y] \triangleq E[(X - E[X])(Y - E[Y])]$$

Using an argument similar to that for variance, we can rewrite this as,

$$\begin{aligned} Cov[X,Y] &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY - XE[Y] - YE[X] + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y]. \end{aligned}$$

Properties:

- (Linearity of expectation) $E[f(X,Y) + g(X,Y)] = E[f(X,Y)] + E[g(X,Y)]$.
- $Var[X+Y] = Var[X] + Var[Y] + 2Cov[X,Y]$.
- If X and Y are independent, then $Cov[X,Y] = 0$.
- If X and Y are independent, then $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$.

4.1 Basic properties

We can define the **joint distribution function** of $X_1, X_2, \ldots, X_n$, the **joint probability density function** of $X_1, X_2, \ldots, X_n$, the **marginal probability density function** of X_1, and the **conditional probability density function** of X_1 given $X_2, \ldots, X_n$, as

$$\begin{aligned} F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots x_n) &= P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) \\ f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots x_n) &= \frac{\partial^n F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots x_n)}{\partial x_1 \ldots \partial x_n} \\ f_{X_1}(X_1) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots x_n) dx_2 \ldots dx_n \\ f_{X_1|X_2,\ldots,X_n}(x_1|x_2, \ldots x_n) &= \frac{f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots x_n)}{f_{X_2,\ldots,X_n}(x_1, x_2, \ldots x_n)} \end{aligned}$$

To calculate the probability of an event $A \subseteq \mathbb{R}^n$ we have,

$$P((x_1, x_2, \ldots x_n) \in A) = \int_{(x_1,x_2,\ldots,x_n) \in A} f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots x_n) dx_1 dx_2 \ldots dx_n \quad (4)$$

Chain rule: From the definition of conditional probabilities for multiple random variables, one can show that

$$\begin{aligned} f(x_1, x_2, \ldots, x_n) &= f(x_n|x_1, x_2 \ldots, x_{n-1}) f(x_1, x_2 \ldots, x_{n-1}) \\ &= f(x_n|x_1, x_2 \ldots, x_{n-1}) f(x_{n-1}|x_1, x_2 \ldots, x_{n-2}) f(x_1, x_2 \ldots, x_{n-2}) \\ &= \ldots = f(x_1) \prod_{i=2}^{n} f(x_i|x_1, \ldots, x_{i-1}). \end{aligned}$$