1 Probability and Random Variable

Suppose we run an experiment like throwing a die a number of times. Sometimes six dots show up on the top face, but more often fewer than six dots show up. We refer to six dots appearing at the top face as event $A$. Common sense tells us that the probability of event $A$ occurring is $1/6$, because every face is equally likely to appear at the top after a throw. The probability of event $A$ is defined as the ratio of the number of times $A$ occurs to the total number of outcomes.

We call an experiment a procedure that yields one of a given set of possible outcomes. The sample space, denoted $S$, of the experiment is the set of possible outcomes. An event $A$ is a subset of the sample space. Laplace’s definition of probability of an event $A$ is

$$\Pr(A) = \frac{|A|}{|S|}.$$  

Let $A$ and $B$ be events with $\Pr(B) > 0$. The conditional probability of event $A$ given event $B$ with non-zero probability is defined as

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

**Example 1.** Consider throwing a die a number of times. Suppose $A$ refers to the appearance of 4 on a die, and $B$ refers to the appearance of an even number. We have $\Pr(A) = 1/6$. But if we know that the die has an even number showing up at the top, then the probability becomes $1/3$. Formally, we know that $\Pr(B) = 1/2$, thus

$$\Pr(A|B) = \frac{1/6}{1/2} = \frac{1}{3}.$$  

The *a priori* probability of $A$ is 1/6. But the *a posteriori* probability of $A$ given $B$ is 1/3.

A *random variable* is a function from the sample space $S$ of an experiment to the set of real numbers. Namely, it assigns a real number to each possible outcome. For example, the roll of a die can be viewed as a random variable that maps the appearance of one dot to 1, the appearance of the two dots to 2, and so on. Of course, after a throw, the value of the die is no longer a random variable; it is certain. So the outcome of a particular experiment is not a random variable.

A random variable can be either continuous or discrete. The throw of a die is a discrete random variable, whereas the high temperature tomorrow is a continuous random variable whose outcome
takes on a continuous set of values. Given a random variable $X$, its cumulative distribution function (CDF) is defined as

$$D(x) = \Pr(X \leq x).$$  \hfill (1)

In the case of a discrete probability $\Pr(x)$, that is, the probability that a discrete random variable $X$ assumes value $x$, we have

$$D(x) = \sum_{X \leq x} \Pr(X).$$

Some trivial properties of the CDF are listed below:

$$D(x) \in [0, 1],$$
$$D(-\infty) = 0,$$
$$D(\infty) = 1,$$
$$D(a) \leq D(b), \text{ if } a < b,$$
$$\Pr(a < X \leq b) = D(b) - D(a).$$

The probability density function (PDF) $P(x)$ of a continuous random variable is defined as the derivative of the cumulative distribution function $D(x)$:

$$P(x) = \frac{d}{dx}D(x)$$  \hfill (2)

So

$$D(x) \equiv \int_{-\infty}^{x} P(\xi) \, d\xi.$$

Some properties of the PDF can be obtained from the definition:

$$P(x) \geq 0,$$
$$\int_{-\infty}^{\infty} P(x) \, dx = 1,$$
$$D(a < x \leq b) = \int_{a}^{b} P(x) \, dx.$$

A uniform distribution has constant PDF. The probability density function and cumulative distribution function for a continuous uniform distribution on the interval $[a, b]$ are

$$P(x) = \begin{cases} 0, & \text{for } x < a, \\ \frac{1}{b-a}, & \text{for } a \leq x \leq b, \\ 0, & \text{for } x > b; \end{cases}$$

$$D(x) = \begin{cases} 0, & \text{for } x < a, \\ \frac{x-a}{b-a}, & \text{for } a \leq x \leq b, \\ 1, & \text{for } x > b. \end{cases}$$
2 Generating a Continuous Distribution

A programming language often has some built-in function for generating uniformly distributed pseudo-random numbers. For instance, in C++, we can use the function `rand()` to generate a pseudo-random number between 0 and 65535 after seeding the built-in random number generator `srand(n)`, where `n` is an integer.

The utility of pseudo-random numbers can be wielded for simulating a given cumulative distribution function $D(x)$. Here we present a method for generating continuous distributions as offered in [2, p. 121]. By its definition (1), the function $D(x)$ increases monotonically from zero to one. Suppose $D(x)$ is continuous and strictly increasing, there exists an inverse function $D^{-1}(y)$ such that, for $0 < y < 1$,

$$ y = D(x) \quad \text{if and only if} \quad x = D^{-1}(y). $$

We can compute a random variable $X$ with distribution $D(x)$ by setting

$$ X = D^{-1}(Y), $$

where $Y$ is a random variable with uniform distribution over $[0, 1]$. The reasoning is as follows:

$$ \Pr(X \leq x) = \Pr(D^{-1}(Y) \leq x) \\
= \Pr(Y \leq D(x)) \\
= D(x) $$

**Example 2 (uniformly distributed random points).** Given a rectangle with length $l$ and width $w$, we can easily generate random points uniformly distributed inside the rectangle. Simply plot a point at the position $(X, Y)$, where $X$ and $Y$ are random variables uniformly distributed within the intervals $[0, l]$ and $[0, w]$, respectively.

Suppose we want to generate uniform random points inside a circle of radius $\rho$ and centered at the origin. We can represent the location of such a point as $R(\cos \Theta, \sin \Theta)$, where $R$ and $\Theta$ are two random variables with ranges $[0, \rho]$ and $[0, 2\pi]$, respectively. The value of $\Theta$ follows the uniform distribution since the polar angle of a random point is equally likely to take on any value in $[0, 2\pi]$. But the radius variable $R$ does not. If we were to generate its values according to a uniform distribution over $[0, \rho]$, then the resulting points would be more concentrated near the origin than far from it.

Hence, we need to find a distribution for the radius variable $R$. First, we compute the cumulative distribution function:

$$ D(r) = \Pr(R \leq r) = \frac{\pi r^2}{\pi \rho^2} = \frac{r^2}{\rho^2}. $$

Clearly, $0 \leq D(r) \leq 1$. We let $s = D(r)$ and obtain $r = \rho \sqrt{s}$. Introduce a random variable $S$ with uniform distribution over $[0, 1]$. So $R$ can be computed as $R = \rho \sqrt{S}$. 

---

`P(x)` and `D(x)` are shown in the diagram with `a` and `b` as the start and end points for generating a continuous distribution within the interval `[a, b]`. The graph for `D(x)` is depicted increasing from zero to one, highlighting the monotonic nature of the cumulative distribution function.
In summary, random points should be generated at positions $\rho \sqrt{S} (\cos \Theta, \sin \Theta)$, where $S$ and $\Theta$ are random variables with uniform distributions over $[0, 1]$ and $[0, 2\pi]$.

3 Expected Value and Standard Deviation

The *expected value*, or *mean*, of a random variable $X$ is its average value over a large number of experiments. Suppose we run the experiment $N$ times and observe a total of $m$ different outcomes. Among them, the outcome $x_1$ occurs $n_1$ times, $x_2$ occurs $n_2$ times, $\ldots$, and $x_m$ occurs $n_m$ times. Then the expected value is computed as

$$E(X) = \frac{1}{N} \sum_{i=1}^{m} x_i n_i.$$

**Example 3.** Suppose we roll a die an infinite number of times. We expect that each number appears $1/6$ of the time. So the expected value is

$$E(X) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{6} i \cdot \frac{n}{6} = \frac{7}{2}.$$

For a continuous random variable $X$ with probability density function $P(x)$, its expected value is given as

$$E(X) = \int_{-\infty}^{\infty} x P(x) \, dx. \quad (3)$$

A discrete random variable with $N$ possible values $x_1, \ldots, x_N$ and mean $\mu$ has *variance*

$$\text{var}(X) = \sum_{i=1}^{N} \Pr(x_i) (x_i - \mu)^2,$$

where $\Pr(x_k)$ is probability of the value $x_k$, for $1 \leq k \leq N$. The variance measures the dispersion of these values taken by $X$ around its mean value. We often write $\text{var}(X) = \sigma^2$ and call $\sigma$ the *standard derivation*. For a continuous distribution, the variance is given by

$$\text{var}(X) = \int_{-\infty}^{\infty} P(x)(x - \mu)^2 \, dx. \quad (4)$$

Note that the variance can be written as

$$\sigma^2 = E((X - \mu)(X - \mu))$$
$$= E(X^2 - 2X\mu + \mu^2)$$
$$= E(X^2) - 2\mu^2 + \mu^2$$
$$= E(X^2) - \mu^2.$$
A random variable $X$ with mean $\mu$ and variance $\sigma^2$ has Gaussian distribution or normal distribution if its probability density function is given by

$$P(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}.$$  \hspace{1cm} (5)

This next figure\(^1\) plots the function with $\mu = 0$ and $\sigma = \sqrt{\frac{2}{\pi}}$, after being scaled by $\sqrt{\pi}$.

We use the notation $N(\mu, \sigma^2)$ for a Gaussian distribution.

4 Multiple Random Variables

Let $X$ and $Y$ be random variables with cumulative distribution functions

$$G(x) = \Pr(X \leq x) \quad \text{and} \quad H(y) = \Pr(Y \leq y),$$

respectively. The probability for both $X \leq x$ and $Y \leq y$ is defined as the joint cumulative distribution function

$$D(x, y) = \Pr(X \leq x \land Y \leq y).$$

Below are some properties of this function:

\begin{align*}
D(x, y) & \in [0, 1], \\
D(x, -\infty) = D(-\infty, y) & = 0, \\
D(\infty, \infty) & = 1, \\
D(a, c) & \leq D(b, d), \quad \text{if } a \leq b \text{ and } c \leq d, \\
\Pr(a < x \leq b \land c < y \leq d) & = D(b, d) + D(a, c) - D(a, d) - D(b, c), \\
D(x, \infty) & = G(x), \\
D(\infty, y) & = H(y).
\end{align*}

\(^1\)from the Wolfram MathWorld at http://mathworld.wolfram.com/GaussianFunction.html.
The joint probability density function is defined as the following second order partial derivative:

\[ P(x, y) = \frac{\partial^2}{\partial x \partial y} D(x, y). \]

It possesses some properties which can be obtained from the definition:

\[
D(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} P(w, z) \, dw \, dz,
\]

\[
\Pr(a < x \leq b \land c < y \leq d) = \int_{c}^{d} \int_{a}^{b} f(x, y) \, dx \, dy,
\]

\[
P_X(x) = \int_{-\infty}^{\infty} P(x, y) \, dy,
\]

\[
P_Y(y) = \int_{-\infty}^{\infty} P(x, y) \, dx.
\]

where \( P_X(x) \) and \( P_Y(y) \) are the probability density functions of \( X \) and \( Y \), respectively.

Two random variables \( X \) and \( Y \) are statistically independent if they satisfy the following relation:

\[
\Pr(X \leq x \land Y \leq y) = \Pr(X \leq x) \cdot \Pr(Y \leq y), \tag{6}
\]

for all \( x, y \in \mathbb{R} \). Hence, the following hold for the joint cumulative distribution and probability density functions:

\[
D(x, y) = G(x)H(y),
\]

\[
P(x, y) = P_X(x)P_Y(y).
\]

In fact, the sum of independent random variables tends toward a Gaussian random variable, regardless of their individual probability density functions. A random variable in nature often appears to follow a Gaussian distribution because it is the sum of many individual and independent random variables. For instance, the high temperature on any given day in any given location is affected by clouds, precipitation, wind, air pressure, humidity, etc. It has a Gaussian probability density function.

The covariance of random variables \( X \) and \( Y \) with means \( \mu_X \) and \( \mu_Y \), respectively, is defined as

\[
\text{cov}(X, Y) = \mathbb{E} \left( (X - \mu_X)(Y - \mu_Y) \right) = \mathbb{E}(XY) - \mu_X\mu_Y.
\]

When \( X \) and \( Y \) are the same variable, we see that \( \text{cov}(X, X) = \sigma_X^2 \) is the variance of \( X \) with \( \sigma_X \) the standard deviation. The correlation coefficient of the two variables is

\[
\text{cor}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y},
\]

where \( \sigma_Y \) is the standard deviation of \( Y \). The correlation coefficient gives the strength of the relationship between the two random variables.
If $X$ and $Y$ are independent, then we have

$$E(XY) = \int \int xyP(x,y) \, dx \, dy$$

$$= \int \int xyP_X(x)P_Y(y) \, dx \, dy$$

$$= \int xP_X(x) \int yP_Y(y) \, dy$$

$$= E(X)E(Y).$$

Therefore,

$$\text{cov}(X, Y) = \text{cor}(X, Y) = 0.$$

Two random variables $X$ and $Y$ are \textit{uncorrelated} if

$$\text{cov}(X, Y) = 0,$$

or equivalently, if

$$E(XY) = E(X)E(Y). \quad (7)$$

Independence implies uncorrelatedness, but not necessarily vice versa.

Two variables $X$ and $Y$ are \textit{orthogonal} if

$$E(XY) = 0. \quad (8)$$

If they are orthogonal, they may or may not be uncorrelated. If they are uncorrelated, then they are orthogonal only if $E(X) = 0$ or $E(Y) = 0$.

\textbf{Example 4.} Let $X$ and $Y$ represent two throws of the dice. The two variables are independent because the outcome of one throw does not affect that of the other. We have that

$$E(X) = E(Y) = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5.$$

All possible outcomes of the two throws are $(i, j), 1 \leq i, j \leq 6$. Each combination has probability $\frac{1}{36}$. The mean of $XY$ is

$$E(XY) = \frac{1}{36} \sum_{i=1}^{6} \sum_{j=1}^{6} ij$$

$$= 12.25$$

$$= E(X)E(Y).$$

Thus $X$ and $Y$ are uncorrelated. However, they are not orthogonal since $E(XY) \neq 0$.

\textbf{Example 5.} Consider the circuit in the left figure. The input voltage $V$ is uniformly distributed on $[-1, 1]$ in volts. The two currents, in amps, are

$$I_1 = \begin{cases} 0 & \text{if } V > 0, \\ V & \text{if } V \leq 0; \end{cases}$$

$$I_2 = \begin{cases} V & \text{if } V \geq 0, \\ 0 & \text{if } V < 0. \end{cases}$$

\textsuperscript{2}from Example 2.11 in [3, p. 65].
So $I_1$ is uniformly distributed on $[-1, 0]$ and $I_2$ is uniformly distributed on $[0, 1]$. The expected values of the three random variables are as follows:

$$E(V) = 0,$$
$$E(I_1) = -\frac{1}{2},$$
$$E(I_2) = \frac{1}{2}.$$ 

The random variables $I_1$ and $I_2$ are related in that exactly one of them is zero at any time instant. Since $I_1I_2 = 0$ always holds, $E(I_1I_2) = 0$. So the two variables are orthogonal. Because $E(I_1)E(I_2) = -\frac{1}{4} \neq E(I_1I_2) = 0$, $I_1$ and $I_2$ are correlated.

5 Multivariate Statistics

Let $X = (X_1, X_2, \ldots, X_n)^T$ be a vector of $n$ random variables such that, for $1 \leq i \leq n$,

$$\mu_i = E(X_i).$$

The covariance matrix is $\Sigma$ whose $(i, j)$-th entry is as

$$(\Sigma)_{ij} = \text{cov}(X_i, X_j)$$
$$= E((X_i - \mu_i)(X_j - \mu_j))$$
$$= E(X_iX_j) - \mu_i\mu_j.$$ 

Denoting $\mu = (\mu_1, \mu_2, \ldots, \mu_n)^T$, we can simply write the covariance matrix as

$$\Sigma = E((X - \mu)(X - \mu)^T) = EXX^T - \mu\mu^T.$$ 

The correlation matrix has in the $(i, j)$-th entry the correlation coefficient

$$\text{cor}(X_i, X_j) = \frac{\text{cov}(X_i, X_j)}{\sigma_i\sigma_j}.$$ 

When the random variables are normalized, i.e., with unit standard variations, the correlation matrix is the same as the covariance matrix.

Whereas theoreticians are primarily interested in the covariance matrix, practitioners prefer the correlation matrix, because a correlation coefficient is more intuitive than a covariance. Other than that, both matrices have essentially the same properties.

An $n$-element $X$ is Gaussian (or normal) if its probability density function is given by

$$\frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(X - \mu)^T\Sigma^{-1}(X - \mu)\right).$$
6 Stochastic Process

A *stochastic process* \( X(t) \) is a random variable that changes with time. Time may be continuous or discrete. The value of the random variable may be continuous at every time instant or discrete at every time instant. So a stochastic process can be one of four types.

The distribution and density functions of a stochastic process are functions of time:

\[
D(x, t) = \Pr(X(t) \leq x),
\]

\[
P(x, t) = \frac{d}{dx}D(x, t).
\]

For a random vector \( \mathbf{X} = (X_1, \ldots, X_n) \), these functions are defined as

\[
D(x_1, \ldots, x_n, t_1, t_2) = \Pr(X_1(t_1) \leq x_1 \land \cdots \land X_n(t_2) \leq x_n),
\]

\[
P(x_1, \ldots, x_n, t_1, t_2) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n}D(x_1, \ldots, x_n, t_1, t_2).
\]

The mean and covariance of a stochastic process \( X(t) \), originally defined in (3) and (4) for a time-independent random variable, respectively, are also functions of time. The integrations are carried out over the values of the random variable.

A stochastic process \( X(t) \) at two different times \( t_1 \) and \( t_2 \) comprises two different random variables \( X(t_1) \) and \( X(t_2) \). So, we can talk about their joint distribution and joint density functions, defined as follows:

\[
D(x_1, x_2, t_1, t_2) = \Pr(X(t_1) \leq x_1 \land X(t_2) \leq x_2),
\]

\[
P(x_1, x_2, t_1, t_2) = \frac{\partial^2}{\partial x_1 \partial x_2}D(x_1, x_2, t_1, t_2).
\]

A stochastic process is called *stationary* if its probability density does not change with time.

**Example 6.** Tomorrow’s closing price of the Dow Jones Industrial Average might be a random variable with a certain mean and variance. However, a century ago the mean was much lower. The closing price is a random variable with generally increasing mean with time. It is not stationary.

7 Noise Simulation

If the random variables \( X(t_1) \) and \( X(t_2) \) are independent for all \( t_1 \neq t_2 \), then the stochastic process \( X(t) \) is called *white noise*. Otherwise, it is called *colored noise*. In optimal filtering research and experiments, we often have to simulate correlated white noise. Phrased in the discrete sense, we need to create random vectors whose elements are correlated with each other according to some predefined covariance matrix.

Suppose we want to generate an \( n \)-element random vector \( \mathbf{X} \) which has zero mean and covariance matrix:

\[
\Sigma = E((\mathbf{X} - \mathbf{0})(\mathbf{X} - \mathbf{0})^T) = \begin{pmatrix}
\sigma_1^2 & \cdots & \sigma_1^2 \\
\vdots & \ddots & \vdots \\
\sigma_n^2 & \cdots & \sigma_n^2
\end{pmatrix},
\]

\footnote{Example 2.12 in [3, p. 70].}
As a covariance matrix, $\Sigma$ must be positive semi-definite. Its eigenvalues are real and nonnegative, and denoted as $\mu_1^2, \ldots, \mu_n^2$. Let $d_1, \ldots, d_n$ be the corresponding eigenvectors which can be chosen orthogonal since $\Sigma$ is symmetric. By the Spectral Theorem [4, p. 273] in linear algebra, the matrix $\Sigma$ has a decomposition

$$
\Sigma = Q \Lambda Q^T,
$$

where $Q = (d_1, \ldots, d_n)$ is orthogonal, and

$$
\Lambda = \text{diag}(\mu_1^2, \ldots, \mu_n^2).
$$

We introduce a new random vector $Y = Q^{-1}X = Q^TX$ so that $X = QY$. Therefore,

$$
E(YY^T) = E(Q^TXX^TQ) = Q^T E(XX^T)Q = Q^T \Sigma Q = \Lambda.
$$

The covariance matrix of $X$ is as given:

$$
E(XX^T) = E((QY)(QY)^T) = E(QYY^TQ^T) = QE(YY^T)Q^T = Q\Lambda Q^T = \Sigma.
$$

The following steps summarize the algorithm of generating $X$ with zero mean and covariance matrix $\Sigma$.

1. Compute the eigenvalues $\mu_1^2, \ldots, \mu_n^2$ of $\Sigma$.
2. Find the corresponding eigenvectors $d_1, \ldots, d_2$.
3. For $i = 1, \ldots, n$, compute the random variable $Y_i = \mu_i Z_i$, where $Z_i$ is an independent random number with zero mean and unit variance.
4. Let $X = Q(Y_1, \ldots, Y_n)^T$.

We refer to [1, pp. 391–448] on eigenvalue computation for symmetric matrices.

References


