The sample space $\Omega$ is the set of possible outcomes of an experiment. Points $\omega$ in $\Omega$ are called sample outcomes, realizations, or elements. Subsets of $\Omega$ are called events.

An event is denoted by a capital letter near the beginning of the alphabet ($A, B, \ldots$). The probability that $A$ occurs is denoted by $P(A)$.

**Example.** If we toss a coin twice then $\Omega = \{HH, HT, TH, TT\}$. The event that the first toss is heads is $A = \{HH, HT\}$. 
Probability satisfies the following elementary properties, called axioms; all other properties can be derived from these.

1. \(0 \leq P(A) \leq 1\) for any event \(A\);
2. \(P(\text{not } A) = 1 - P(A)\);
3. \(P(A \text{ or } B) = P(A) + P(B)\) if \(A\) and \(B\) are mutually exclusive events (i.e. \(A\) and \(B\) cannot both happen simultaneously).

More generally, if \(A\) and \(B\) are any events then

\[
P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B). \quad (1)
\]

If \(A\) and \(B\) are mutually exclusive, then
\(P(A \text{ and } B) = 0\) and (1) reduces to axiom 3.
Conditional probability. If $B$ is known to have occurred, then this knowledge may affect the probability of another event $A$. The probability of $A$ once $B$ is known to have occurred is written $P(A|B)$ and called “the conditional probability of $A$ given $B$,” or, more simply, “the probability of $A$ given $B.”$ It is defined as

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$ \hspace{1cm} (2)

provided that $P(B) \neq 0$.

Independence. The events $A$ and $B$ are said to be independent if

$$P(A \text{ and } B) = P(A) P(B).$$ \hspace{1cm} (3)

By (2), this implies $P(A|B) = P(A)$ and $P(B|A) = P(B)$. Intuitively, independence means that knowing $A$ has occurred provides no information about whether or not $B$ has occurred and vice-versa.
Random variables

A random variable is the outcome of an experiment (i.e. a random process) expressed as a number. We use capital letters near the end of the alphabet (X, Y, Z, etc.) to denote random variables. Random variables are of two types: discrete and continuous.

Continuous random variables are described by probability density functions (PDF). For example, a normally distributed random variable has a bell-shaped density function like this:

![Bell-shaped density curve](image)

The probability that X falls between any two particular numbers, say a and b, is given by the area under the density curve $f(x)$ between a and b,

$$P(a \leq X \leq b) = \int_{a}^{b} f(x) \, dx.$$
The two continuous random variables that we will use most are the normal and the $\chi^2$ (chisquare) distributions. Areas under the normal and $\chi^2$ density functions are tabulated and widely available in textbooks. They can also be computed with statistical computer packages (e.g. Minitab).

*Discrete random variables* are described by probability mass functions (PMF), which we will also call “distributions.” For a random variable $X$, we will write the distribution as $f(x)$ and define it to be

$$f(x) = P(X = x).$$

In other words, $f(x)$ is the probability that the random variable $X$ takes the specific value $x$. For example, suppose that $X$ takes the values 1, 2, and 5 with probabilities $1/4$, $1/4$, and $1/2$ respectively. Then we would say that $f(1) = 1/4$, $f(2) = 1/4$, $f(5) = 1/2$, and $f(x) = 0$ for any $x$ other than 1, 2, or 5:

$$f(x) = \begin{cases} .25 & x = 1, 2 \\ .50 & x = 5 \\ 0 & \text{otherwise} \end{cases}$$
A graph of $f(x)$ has spikes at the possible values of $X$, with the height of a spike indicating the probability associated with that particular value:

Note that $\sum_x f(x) = 1$ if the sum is taken over all values of $x$ having nonzero probability. In other words, the sum of the heights of all the spikes must equal one.

**Joint distribution.** Suppose that $X_1, X_2, \ldots, X_n$ are $n$ random variables, and let $X$ be the entire vector

$$X = (X_1, X_2, \ldots, X_n).$$

Let $x = (x_1, x_2, \ldots, x_n)$ denote a particular value that $X$ can take. The joint distribution of $X$ is

$$f(x) = P(X = x) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n).$$
In particular, suppose that the random variables $X_1, X_2, \ldots, X_n$ are _independent and identically distributed_ (iid). Then $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ are independent events, and the joint distribution is

$$f(x) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X_1 = x_1)P(X_2 = x_2)\cdots P(X_n = x_n) = f(x_1)f(x_2)\cdots f(x_n) = \prod_{i=1}^{n} f(x_i)$$

where $f(x_i)$ refers to the distribution of $X_i$.

**Moments**

The _expectation_ of a discrete random variable $X$ is defined to be

$$E(X) = \sum_x x f(x)$$

where the sum is taken over all possible values of $X$. $E(X)$ is also called the _mean_ of $X$ or the _average_ of $X$, because it represents the long-run average value if the experiment were repeated infinitely many times.
In the trivial example where $X$ takes the values 1, 2, and 5 with probabilities $1/4$, $1/4$, and $1/2$ respectively, the mean of $X$ is

$$E(X) = 1(.25) + 2(.25) + 5(.5) = 3.25.$$ 

In calculating expectations, it helps to visualize a table with two columns. The first column lists the possible values $x$ of the random variable $X$, and the second column lists the probabilities $f(x)$ associated with these values:

<table>
<thead>
<tr>
<th>$x$</th>
<th>$f(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.25</td>
</tr>
<tr>
<td>2</td>
<td>.25</td>
</tr>
<tr>
<td>5</td>
<td>.50</td>
</tr>
</tbody>
</table>

To calculate $E(X)$ we merely multiply the two columns together, row by row, and add up the products: $1(.25) + 2(.25) + 5(.5) = 3.25$.

If $g(X)$ is a function of $X$ (e.g. $g(X) = \log X$, $g(X) = X^2$, etc.) then $g(X)$ is also a random variable. It’s expectation is

$$E(g(X)) = \sum_{x} g(x)f(x).$$  \hspace{1cm} (4)
Visually, in the table containing $x$ and $f(x)$, we can simply insert a third column for $g(x)$ and add up the products $g(x)f(x)$. In our example, if $Y = g(X) = X^3$, the table becomes

<table>
<thead>
<tr>
<th>$x$</th>
<th>$f(x)$</th>
<th>$g(x) = x^3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.25</td>
<td>$1^3 = 1$</td>
</tr>
<tr>
<td>2</td>
<td>.25</td>
<td>$2^3 = 8$</td>
</tr>
<tr>
<td>5</td>
<td>.50</td>
<td>$5^3 = 125$</td>
</tr>
</tbody>
</table>

and

$$E(Y) = E(X^3) = 1(.25) + 8(.25) + 125(.5) = 64.75.$$ 

If $Y = g(X) = a + bX$ where $a$ and $b$ are constants, then $Y$ is said to be a **linear function** of $X$, and $E(Y) = a + bE(X)$. An algebraic proof is

$$E(Y) = \sum_y yf(y)$$

$$= \sum_x (a + bx)f(x)$$

$$= \sum_x af(x) + \sum_x bx f(x)$$

$$= a \sum_x f(x) + b \sum_x xf(x)$$

$$= a \cdot 1 + bE(X).$$
That is, if $g(X)$ is linear, then $E(g(X)) = g(E(X))$. Note, however, that this does not work if the function $g$ is nonlinear. For example, $E(X^2)$ is not equal to $E(X)^2$, and $E(\log X)$ is not equal to $\log E(X)$. To calculate $E(X^2)$ or $E(\log X)$, we need to use (4).

**Variance.** The variance of a discrete random variable, denoted by $V(X)$, is defined to be

$$V(X) = E((X - E(X))^2) = \sum_x (x - E(X))^2 f(x).$$

That is, $V(X)$ is the average squared distance between $X$ and its mean. Variance is a measure of dispersion, telling us how “spread out” a distribution is. For our simple random variable, the variance is


A slightly easier way to calculate the variance is to use the well-known identity

$$V(X) = E(X^2) - (E(X))^2.$$
Visually, this method requires a table with three columns: $x$, $f(x)$, and $x^2$.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$f(x)$</th>
<th>$x^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.25</td>
<td>$1^2 = 1$</td>
</tr>
<tr>
<td>2</td>
<td>.25</td>
<td>$2^2 = 4$</td>
</tr>
<tr>
<td>5</td>
<td>.50</td>
<td>$5^2 = 25$</td>
</tr>
</tbody>
</table>

First we calculate

$E(X) = 1(.25) + 2(.25) + 5(.50) = 3.25$ and

$E(X^2) = 1(.25) + 4(.25) + 25(.50) = 13.75$. Then

$V(X) = 13.75 - (3.25)^2 = 3.1875$.

It can be shown that if $a$ and $b$ are constants, then

$V(a + bX) = b^2 V(X)$.  

In other words, adding a constant $a$ to a random variable does not change its variance, and multiplying a random variable by a constant $b$ causes the variance to be multiplied by $b^2$.

Another common measure of dispersion is the standard deviation, which is merely the positive square root of the variance,

$$SD(X) = \sqrt{V(X)}.$$
Mean and variance of a sum of random variables.

Expectation is always additive; that is, if $X$ and $Y$ are any random variables, then

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

If $X$ and $Y$ are independent random variables, then their variances will also add:

$$V(X + Y) = V(X) + V(Y) \text{ if } X, Y \text{ independent.}$$

More generally, if $X$ and $Y$ are any random variables, then

$$V(X + Y) = V(X) + V(Y) + 2\text{Cov}(X, Y)$$

where $\text{Cov}(X, Y)$ is the covariance between $X$ and $Y$,

$$\text{Cov}(X, Y) = E((X - E(X))(Y - E(Y))).$$

If $X$ and $Y$ are independent (or merely uncorrelated) then $\text{Cov}(X, Y) = 0$. This additive rule for variances extends to three or more random variables; e.g.,

$$V(X + Y + Z) = V(X) + V(Y) + V(Z)$$

$$+ 2\text{Cov}(X, Y) + 2\text{Cov}(X, Z) + 2\text{Cov}(Y, Z),$$

with all covariances equal to zero if $X$, $Y$, and $Z$ are mutually uncorrelated.