2.1: Random variables and probability mass functions

For a probability space $(\Omega, \mathcal{F}, P)$, a random variable $X$ is a real-valued function on $\Omega$. In other words, $X$ maps $\omega$, the outcome of a probability experiment, to a real value. For example, $X$ can be the sum of the numbers showing on a pair of fair dice when they are rolled. Here, $\Omega = \{\omega = (i, j) : 1 \leq i \leq 6, 1 \leq j \leq 6\}$ and $X(\omega) = i + j$.

Discrete random variables. A random variable $X$ is said to be of discrete-type if it can take finitely many values $x_1, \cdots, x_n$ or countably infinitely many values $x_1, x_2, \cdots$.

Probability mass function. The probability mass function (PMF) for a discrete-type random variable $X$, $p_X$ is defined as $p_X(x) = P(X = x)$. Note that the PMF always sums to 1, $\sum_i p_X(x_i) = 1$, and for any event $A \in \mathcal{F}$, $P(X \in A) = \sum_{x_i \in A} p_X(x_i)$.

2.2: The mean and variance of a random variable

The mean (also called expectation) of a random variable $X$ with PMF $p_X$ is denoted by $E[X]$ and is defined by $E[X] = \sum_{x} x p_X(x)$, where $x_1, x_2, \cdots$ is the list of values that $X$ can take. For example, if $X$ is the sum of the two numbers showing on a pair fair dice when they are rolled, then $E[X] = 7$.

The law of the unconscious statistician (LOTUS). If $Y = g(X)$, then $E[Y] = \sum_i g(x_i)p_X(x_i)$, where the summation is taken over the list of values that $X$ can take. For example, if $X_1$ and $X_2$ are the two numbers showing when two fair dice are rolled, and $Y = X_1X_2$, then $E[Y] = (1/36)\sum_{i=1}^6 \sum_{j=1}^6 ij = 12.25$.

Linearity of the expectation operator. If $g(X)$ and $h(X)$ are functions of $X$, and $a$, $b$, and $c$ are constants, then $ag(X) + bh(X) + c$ is also a function of $X$, and

$$E[ag(X) + bh(X) + c] = aE[g(X)] + bE[h(X)] + c.$$  

The expectation operator $E[.]$ is a linear one: $E[aX + b] = aE[X] + b$.

Variance and standard deviation. The variance of a random variable $X$ is a measure of how spread out the PMF of $X$ is. Let $\mu_X = E[X]$, the variance is defined by $\text{Var}(X) = E[(X - \mu_X)^2]$. An alternative expression for the variance is given by $\text{Var}(X) = E[X^2] - \mu_X^2$. Sometimes, $\text{Var}(X)$ is referred to as the mean square deviation of $X$ around its mean. The variance is often denoted by $\sigma_X^2$, where $\sigma_X = \sqrt{\text{Var}(X)}$ is called the standard deviation of $X$. The variance operator $\text{Var}(\cdot)$ is not linear. In fact,

$$\text{Var}(X + b) = \text{Var}(X)$$
$$\text{Var}(aX) = a^2\text{Var}(X).$$

The random variable $Y = \frac{X - \mu_X}{\sigma_X}$ is called the standardized version of $X$. This is because no matter what $\mu_X$ and $\sigma_X$ are, $\mu_Y = 0$ and $\sigma_Y = 1$.

The moment of a random variable. For an integer $i \geq 1$, the $i^{th}$ moment of $X$ is defined to be $E[X^i]$. Note that the variance of $X$ is equal to its second moment minus the square of its first moment (its mean).

The mode of a random variable. The mode of a random variable $X$ is the value $x$ with the highest probability. In other words, if $x$ is the mode of $X$, then $p_X(x) \geq p_X(u)$ for all $u$. 


\[2.3: \text{Conditional probabilities}\]

Let \( A \) and \( B \) be two events in \( \mathcal{F} \) for some probability experiment \((\Omega, \mathcal{F}, P)\). The conditional probability of \( B \) given \( A \) is defined by

\[
P(B|A) = \begin{cases} 
P(A \cap B) / P(A) & \text{if } P(A) > 0 \\ 
\text{undefined} & \text{if } P(A) = 0. 
\end{cases}
\]

In general, \( P(B|A) \) can be smaller than, larger than, or equal to \( P(B) \). For example, if we roll two fair dice, and let \( A = \text{“the sum is six”} \) and \( B = \text{“the numbers are not equal”} \). Then \( P(B) = 5/6 \) while \( P(B|A) = 4/5 \). Therefore, \( P(B|A) < P(B) \), which should be interpreted as: if we know that \( \omega \in A \), then the probability of \( \omega \) being in \( B \) decreases.

**Properties of conditional distributions.** The following properties follow from the definition of conditional probabilities.

1. \( P(B|A) \geq 0 \)
2. \( P(B|A) + P(B^c|A) = 1 \). More generally, if \( B_1, B_2, \ldots \) is a list of mutually exclusive sets, then \( P(B_1 \cup B_2 \cup \cdots |A) = P(B_1|A) + P(B_2|A) + \cdots \).
3. \( P(\Omega|A) = 1 \)
4. \( P(A \cap B) = P(A)P(B|A)\)
5. \( P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B)\)

The first three properties imply that \( P(\cdot|A) \) is a probability measure that satisfies the three probability axioms. Finally, observe that in general, \( P(B|A) \) need not be equal to \( P(A|B) \).

\[2.4: \text{Independence and the binomial distribution}\]

Two events \( A \) and \( B \) are said to be independent if \( P(B|A) = P(B) \). This condition is equivalent to \( P(A \cap B) = P(A)P(B) \). For example, if we roll a die and let \( A \) be the event that the outcome is even, and \( B \) be the event that the outcome is a multiple of 3. Then \( A \) and \( B \) are independent because \( P(A \cap B) = 1/6 \), \( P(A) = 1/2 \), and \( P(B) = 1/3 \). Note that the definition of independence is symmetric in \( A \) and \( B \). Therefore, if \( A \) is independent of \( B \), \( B \) is also independent of \( A \) (and vice versa). To emphasize this, some people like to say \( A \) and \( B \) are mutually independent.

**Physically independent events.** Two events \( A \) and \( B \) are physically independent if they depend on physically separated random experiments. For example, if a fair coin is flipped and a die is rolled, and \( A = \{\text{Heads}\} \) and \( B = \{\text{number showing} = 6\} \), then \( A \) and \( B \) are physically independent.

**Equivalent independence conditions.** The following 4 statements are equivalent.

1. \( A \) and \( B \) are independent
2. \( A^c \) and \( B \) are independent
3. \( A \) and \( B^c \) are independent
4. \( A^c \) and \( B^c \) are independent

Therefore, to show that \( A \) and \( B \) are independent, it suffices to prove that any one of the four above statements is true.

**Pairwise vs. mutual independence.** Events \( A, B, \) and \( C \) are pairwise independent if \( P(A \cap B) = P(A)P(B), P(A \cap C) = P(A)P(C), \) and \( P(C \cap B) = P(C)P(B) \). Pairwise independence does not imply
that any one of the events is independent of the intersection of the other two events! In order to have such
notion of independence, called \textit{mutual independence}, a stronger condition is needed. Events \(A, B,\) and \(C\) are
\textit{mutually independent} if

1. they are pairwise independent and
2. \(P(A \cap B \cap C) = P(A)P(B)P(C).\)

Suppose that \(A, B, C\) are mutually independent. Then \(A\) is independent of any event that can be made from 
\(B\) and \(C.\) For example, \(A\) is independent of \(B \cup C\) or 
\(B \cap C.\)

**Independence of random variables.** Let \(X\) and \(Y\) be two random variables defined over the same ran-
dom experiment. The \textit{joint} PMF of \(X\) and \(Y\) is given by 
\(P(X = i, Y = j).\) \(X\) and \(Y\) are independent if 
\(P(X = i, Y = j) = p_X(i)p_Y(j)\) for all \(i, j.\) Independence of random variables will be discussed in more detail in later lectures.

**Bernoulli distribution.** A random variable \(X\) is said to have a \textit{Bernoulli distribution} with parameter \(p,\)
where \(0 \leq p \leq 1,\) if \(X \in \{0, 1\},\) and \(P(X = 1) = p\) and \(P(X = 0) = 1 - p.\) Note that 
\(E[X] = p\) and \(Var(X) = p(1 - p).\)

**Binomial distribution.** Suppose that \(n\) independent Bernoulli trials are conducted, each resulting in a 1
with probability \(p\) and a 0 with probability \(1 - p.\) Let \(X\) be the total number of ones occurring in the \(n\)
trials. Then \(X \in \{0, \cdots, n\}\) follows a \textit{binomial distribution}. Moreover, the PMF of \(X\) is

\[ p_X(k) = \binom{n}{k} p^k (1 - p)^{n-k}. \]

Thus, the binomial distribution is parameterized by \(n\) (the number of Bernoulli trials) and \(p\) (the probability
of getting a 1 in each trial). The mean, variance of a binomial random variable \(X\) are given by

\[ E[X] = np \]
\[ Var(X) = np(1 - p) \]
\[ Mode(X) = \text{largest integer} \leq (n + 1)p. \]

**2.5: Geometric distribution**

**2.6: Random Process**

A discrete time random process \(X_k\) is a collection of random variables \(\{X_1, X_2, \cdots\},\) representing a sequence
of random experiments over time. The \(k^{th}\) random variable \(X_k\) indicates the outcome of the \(k^{th}\) random
experiment. The Bernoulli process is a special case where each \(X_k\) is the outcome of an independent
Bernoulli trial. Therefore, \(X_k = 1\) with probability \(p\) and \(X_k = 0\) with probability \(1 - p.\) More generally,
\(P(X_1 = 1, X_3 = 1, X_5 = 0, X_{12} = 1, X_{15} = 0) = p^3(1 - p)^2.\)