1 Probability Review

1.1 Preliminaries

To start with, I’ll assume that you know the following things:

- The set of all possible outcomes of an experiment is the sample space $\Omega$.
- Any subset $E$ of a sample space $\Omega$ is an event.
- The set of all possible events is denoted by $\mathcal{F}$, which is called a sigma field of $\Omega$.

For example, if $\Omega = \{H, T\}$, then $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$.

A sigma field must satisfy the following:

1. $A \in \mathcal{F}$ implies the complement $\bar{A} \in \mathcal{F}$.
2. $A_1, A_2, \ldots \in \mathcal{F}$ implies $\bigcup_{j=1}^{\infty} A_j \in \mathcal{F}$.

- The probability function $P(\cdot)$ must satisfy 3 axioms:

1. For any event $E \in \mathcal{F}$, we must have $0 \leq P(E) \leq 1$
2. $P(\Omega) = 1$
3. For any disjoint sequence of events $E_1, E_2, \ldots$ (i.e., $E_i \cap E_j = \phi$ if $i \neq j$), we have $P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$.

- A probability space is the triple $(\Omega, \mathcal{F}, P)$.

1.2 Conditional Probability and Independence

Definition: If $P(B) > 0$, then $P(A|B) \equiv P(A \cap B)/P(B)$ is the conditional probability of $A$ given $B$. 
Example: Toss a fair die. Let $A = \{1, 2, 3\}$ and $B = \{3, 4, 5, 6\}$. Then
\[
P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{4/6} = \frac{1}{4}. \quad \Diamond
\]

Definition: If $P(A \cap B) = P(A)P(B)$, then $A$ and $B$ are independent events.

Theorem: If $A$ and $B$ are independent, then $P(A \mid B) = P(A)$.

Proof: Easy. \quad \Diamond

Example: Toss two dice. Let $A = \text{“Sum is 7”}$ and $B = \text{“First die is 4”}$. Then $P(A) = 1/6$, $P(B) = 1/6$, and $P(A \cap B) = P((4, 3)) = 1/36 = P(A)P(B)$; so $A$ and $B$ are independent. \quad \Diamond

### 1.3 Random Variables

Definition: A random variable (RV) $X$ is a function from $S$ to the real line $\mathbb{R}$, i.e., $X: S \to \mathbb{R}$.

Example: Let $X$ be the sum of two dice rolls. Then $X((4, 6)) = 10$. In addition,
\[
P(X = x) = \begin{cases} 
1/36 & \text{if } x = 2 \\
2/36 & \text{if } x = 3 \\
\vdots \\
1/36 & \text{if } x = 12 \\
0 & \text{otherwise}
\end{cases} \quad \Diamond
\]

Definition: If the number of possible values of a RV $X$ is finite or countably infinite, then $X$ is a discrete RV. Its probability mass function (pmf) is $f(x) \equiv P(X = x)$.

Example: Flip 2 coins. Let $X$ be the number of heads.
\[
f(x) = \begin{cases} 
1/4 & \text{if } x = 0 \text{ or } 2 \\
1/2 & \text{if } x = 1 \\
0 & \text{otherwise}
\end{cases} \quad \Diamond
\]

Examples: Here are some well-known discrete RV’s that you should review: Bernoulli($p$), Binomial($n, p$), Geometric($p$), Negative Binomial, Poisson($\lambda$), etc.

Definition: A continuous RV is one with probability zero at every individual point. A RV is continuous if there exists a probability density function (pdf) $f(x)$ such that
\[ P(X \in A) = \int_A f(x) \, dx \] for every set \( A \).

**Example:** Pick a random number between 3 and 7. Then

\[ f(x) = \begin{cases} 
1/4 & \text{if } 3 \leq x \leq 7 \\
0 & \text{otherwise}
\end{cases} \] \(\blacklozenge\)

**Examples:** Here are some well-known continuous RV’s that you should review: Uniform\((a, b)\), Exponential\((\lambda)\), Normal\((\mu, \sigma^2)\), etc.

**Definition:** For any RV \( X \) (discrete or continuous), the **cumulative distribution function** (cdf) is

\[ F(x) \equiv P(X \leq x) = \begin{cases} 
\sum_{y \leq x} f(y) & \text{if } X \text{ is discrete} \\
\int_{-\infty}^{x} f(y) \, dy & \text{if } X \text{ is continuous}
\end{cases} \]

For convenience, we’ll henceforth write \( F(x) = \int_{-\infty}^{x} dF(y) \) to denote both the discrete and continuous cases.

**Example:** Flip 2 coins. Let \( X \) be the number of heads.

\[ F(x) = \begin{cases} 
0 & \text{if } x < 0 \\
1/4 & \text{if } 0 \leq x < 1 \\
3/4 & \text{if } 1 \leq x < 2 \\
1 & \text{if } x \geq 2
\end{cases} \] \(\blacklozenge\)

**Example:** Suppose \( X \sim \text{Exp}(\lambda) \) (i.e., \( X \) has the exponential distribution with parameter \( \lambda \)). Then \( f(x) = \lambda e^{-\lambda x}, \ x \geq 0 \), and the cdf is \( F(x) = 1 - e^{-\lambda x}, \ x \geq 0 \). \(\blacklozenge\)

### 1.4 Expectation

**Definition:** The **expected value** (or mean) of a RV \( X \) is

\[ \mathbb{E}[X] \equiv \int_{R} x \, dF(x) = \begin{cases} 
\sum_{x} xP(X = x) & \text{if } X \text{ is discrete} \\
\int_{R} xf(x) \, dx & \text{if } X \text{ is continuous}
\end{cases} \]

**Example:** Suppose that \( X \sim \text{Bernoulli}(p) \). Then

\[ X = \begin{cases} 
1 & \text{with prob. } p \\
0 & \text{with prob. } 1 - p
\end{cases} \]

and we have \( \mathbb{E}[X] = \sum_{x} xf(x) = p \). \(\blacklozenge\)
Example: Suppose that $X \sim \text{Uniform}(a, b)$. Then

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

and we have $E[X] = \int_R x f(x) dx = (a + b)/2$. ◊

"Definition" (the “law of the unconscious statistician”): Suppose that $g(X)$ is some function of the RV $X$. Then $E[g(X)] \equiv \int_R g(x) dF(x)$.

Example: Suppose $X$ is the following discrete RV:

<table>
<thead>
<tr>
<th>$x$</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f(x)$</td>
<td>0.3</td>
<td>0.6</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Then $E[X^3] = \sum_x x^3 f(x) = 8(0.3) + 27(0.6) + 64(0.1) = 25$. ◊

Example: Suppose $X \sim \text{U}(0, 2)$. Then $E[X^n] = \int_R x^n f(x) dx = 2^n/(n + 1)$. ◊

Definitions: $E[X^n]$ is the $n$th moment of $X$. $E[(X - E[X])^n]$ is the $n$th central moment of $X$. $\text{Var}(X) \equiv E[(X - E[X])^2]$ is the variance of $X$.

Theorem: $\text{Var}(X) = E[X^2] - (E[X])^2$.

Proof: Easy. ◊

Example: Suppose $X \sim \text{Bern}(p)$. Recall that $E[X] = p$. Further,

$$E[X^2] = \sum_x x^2 f(x) = 0^2(1 - p) + 1^2 p = p$$

and

$$\text{Var}(X) = E[X^2] - (E[X])^2 = p - p^2 = p(1 - p)$$. ◊

Example: Suppose $X \sim \text{U}(0, 2)$. By previous examples, $E[X] = 1$ and $E[X^2] = 4/3$. So $\text{Var}(X) = E[X^2] - (E[X])^2 = 1/3$. ◊

Theorem: $E[aX + b] = aE[X] + b$ and $\text{Var}(aX + b) = a^2\text{Var}(X)$.

Proof: Easy. ◊
Definition: The moment generating function (mgf) of $X$ is $M_X(t) \equiv E[e^{tX}]$.

Example: Suppose $X \sim \text{Bern}(p)$. $M_X(t) = \sum_x e^{tx} f(x) = pe^t + q$. ◊

Example: Suppose $X \sim \text{Exp}(\lambda)$. $M_X(t) = \int_0^\infty e^{tx} \lambda e^{-\lambda x} \, dx = \lambda/(\lambda - t), \ t < \lambda$. ◊

Theorem (why we call them moment generating functions): Under certain technical conditions,
$$E[X^k] = \left. \frac{d^k}{dt^k} M_X(t) \right|_{t=0}, \ k = 1, 2, \ldots.$$  

“Proof” We’ll just do the first moment. (The others are similar.)
$$\frac{d}{dt} M_X(t) = \frac{d}{dt} E[e^{tX}] \overset{\text{“=”}}{=\text{ E}} \left[ \frac{d}{dt} e^{tX} \right] \overset{\text{“=”}}{=\text{ E}} [X e^{tX}].$$

If you believe the above steps, then
$$\left. \frac{d}{dt} M_X(t) \right|_{t=0} = E[X].$$ ◊

Theorem: If $X$ and $Y$ have the same mgf, then they have the same distribution (at least in this course).

1.5 Functions of a RV

Problem: Suppose we have a RV $X$ with p.d.f./p.m.f. $f(x)$. Let $Y = h(X)$. Find $g(y)$, the p.d.f./p.m.f. of $Y$.

Discrete case: If $X$ is discrete, then $Y$ will be discrete, in which case
$$g(y) = \Pr(Y = y) = \Pr[h(X) = y] = \Pr\{x : h(x) = y\} = \sum_{x : h(x) = y} f(x).$$

Example: Let $X$ denote the number of $H$’s from two coin tosses. We want the p.m.f. for $Y = X^2 - X$.

<table>
<thead>
<tr>
<th>$x$</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f(x)$</td>
<td>1/4</td>
<td>1/2</td>
<td>1/4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$y = x^2 - x$</th>
<th>0</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f(x)$</td>
<td>1/4</td>
<td>3/4</td>
</tr>
</tbody>
</table>

This implies that $g(0) = \Pr(Y = 0) = \Pr(X = 0 \text{ or } 1) = 3/4$ and $g(2) = \Pr(Y = 2) = 1/4$. In other words,
$$g(y) = \begin{cases} 
3/4 & \text{if } y = 0 \\
1/4 & \text{if } y = 2 \\
0 & \text{otherwise}
\end{cases}.$$ ◊
Continuous Case: We’ll assume that if $X$ is continuous, then so is $Y$. The usual method is to first compute the c.d.f. of $Y$,

$$G(y) = \Pr(Y \leq y) = \Pr[h(X) \leq y] = \int_{\{x: h(x) \leq y\}} f(x) \, dx,$$

and then take the derivative, $g(y) = G'(y)$.

Example: Suppose $X$ has p.d.f. $f(x) = |x|$, $-1 \leq x \leq 1$. Find the p.d.f. of $Y = X^2$.

First of all, the c.d.f. of $Y$ is

$$G(y) = \Pr(Y \leq y) = \Pr(X^2 \leq y) = \Pr(-\sqrt{y} \leq X \leq \sqrt{y})$$

$$= \int_{-\sqrt{y}}^{\sqrt{y}} |x| \, dx = y, \quad 0 < y < 1.$$

Thus, the p.d.f. of $Y$ is $g(y) = G'(y) = 1$, $0 < y < 1$, indicating that $Y \sim \text{Unif}(0, 1)$.

Here is a more-direct method for dealing with functions of RV’s…

Theorem: Suppose that $X$ has p.d.f. $f(x)$, $a \leq x \leq b$. Let $Y = h(X)$ be a monotone function (either increasing or decreasing) of $X$. Then the p.d.f. of $Y$ is

$$g(y) = f(h^{-1}(y)) \left| \frac{d}{dy} h^{-1}(y) \right|, \quad h(a) \leq y \leq h(b) \quad \text{(or} \ h(b) \leq y \leq h(a))\text{).}$$

Remarks: (i) Warning: You can only use this method of $h(x)$ if monotone! (The p.d.f. $f(x)$ doesn’t have to be monotone.) (ii) Think of the inverse function $h^{-1}(y) = x$, and the quantity in the $| \cdot |$ as the Jacobian of the transformation.

Example: Suppose that $f(x) = 3x^2$, $0 \leq x \leq 1$. Find the p.d.f. of $Y = h(X) = X^2$.

Note that $f(x)$ is only defined on the domain $0 \leq x \leq 1$; and on this range, $h(x)$ is monotone increasing — so it’s OK to use the wonderful theorem.

First, we have $x = h^{-1}(y) = \pm \sqrt{y} = \sqrt{y}$ (since we’re only concerned with positive $x$’s). The theorem then implies that

$$g(y) = f(\sqrt{y}) \left| \frac{d}{dy} \sqrt{y} \right|, \quad h(0) \leq y \leq h(1)$$

$$= 3y \times \frac{1}{2\sqrt{y}} = \frac{3y}{2\sqrt{y}}, \quad 0 \leq y \leq 1. \quad \diamond$$

Remark: We can also look at functions of $\geq 2$ RV’s, but this takes more work. See any probability text for more info on this important topic.
1.6 Jointly Distributed RV’s

Definition: The joint cdf of X and Y is \( F(x, y) \equiv P(X \leq x, Y \leq y) \), for all \( x, y \).

Remark: The marginal cdf of \( X \) is \( F_X(x) = F(x, \infty) \). (We use the \( X \) subscript to remind us that it’s just the cdf of \( X \) all by itself.) Similarly, the marginal cdf of \( Y \) is \( F_Y(y) = F(\infty, y) \).

Definition: If \( X \) and \( Y \) are discrete, then the joint pmf of \( X \) and \( Y \) is \( f(x, y) \equiv P(X = x, Y = y) \).

Remark: The marginal pmf of \( X \) is
\[
f_X(x) = P(X = x) = \sum_y f(x, y).
\]
The marginal pmf of \( Y \) is
\[
f_Y(y) = P(Y = y) = \sum_x f(x, y).
\]

Example: Suppose the following table gives the joint pmf of \( X \) and \( Y \), along with the accompanying marginals.

<table>
<thead>
<tr>
<th>( X )</th>
<th>( X = 2 )</th>
<th>( X = 3 )</th>
<th>( X = 4 )</th>
<th>( f_Y(y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y = 4 )</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
<td>0.6</td>
</tr>
<tr>
<td>( Y = 6 )</td>
<td>0.1</td>
<td>0.2</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td>( f_X(x) )</td>
<td>0.4</td>
<td>0.4</td>
<td>0.2</td>
<td>1</td>
</tr>
</tbody>
</table>

Definition: If \( X \) and \( Y \) are continuous, then the joint pdf of \( X \) and \( Y \) is \( f(x, y) \equiv \frac{\partial^2}{\partial x \partial y} F(x, y) \).

Remark: The marginal pdf of \( X \) is
\[
f_X(x) = \int_R f(x, y) \, dy.
\]
The marginal pdf of \( Y \) is
\[
f_Y(y) = \int_R f(x, y) \, dx.
\]

Example: This example shows that you have to be careful about “funny” limits when computing marginals. Suppose the joint pdf is
\[
f(x, y) = \frac{21}{4} x^2 y, \quad x^2 \leq y \leq 1.
\]
Then the marginal pdf's are:

\[ f_X(x) = \int_R f(x, y) \, dy = \int_{x^2}^{1} \frac{21}{4} x^2 y \, dy = \frac{21}{8} x^2 (1 - x^4), \quad -1 \leq x \leq 1 \]

and

\[ f_Y(y) = \int_R f(x, y) \, dx = \int_{\sqrt[4]{y}}^{\sqrt[4]{y}} \frac{21}{4} x^2 y \, dx = \frac{7}{2} y^{5/2}, \quad 0 \leq y \leq 1. \]

### 1.7 Independent RV’s

**Definition:** X and Y are independent RV’s if \( f(x, y) = f_X(x) f_Y(y) \) for all \( x, y \).

**Examples:** If \( f(x, y) = cxy \) for \( 0 \leq x \leq 2, 0 \leq y \leq 3 \), then X and Y are independent. If \( f(x, y) = \frac{21}{4} x^2 y \) for \( x^2 \leq y \leq 1 \), then X and Y are not independent. If \( f(x, y) = \frac{c}{x + y} \) for \( 1 \leq x \leq 2, 1 \leq y \leq 3 \), then X and Y are not independent.

**Definition:** If \( f_X(x) > 0 \), then \( f(y|x) = f(x, y) / f_X(x) \) is the conditional pdf (or pmf) of Y given \( X = x \).

**Example:** Suppose \( f(x, y) = \frac{21}{4} x^2 y \) for \( x^2 \leq y \leq 1 \). By a previous example, we find that

\[ f(y|x) = \frac{f(x, y)}{f_X(x)} = \frac{\frac{21}{4} x^2 y}{\frac{21}{8} x^2 (1 - x^4)} = \frac{2y}{1 - x^4}, \quad x^2 \leq y \leq 1. \]

“**Definition**”: Suppose that \( h(X, Y) \) is some function of the RV’s X and Y. Then

\[ \mathbb{E}[h(X, Y)] = \left\{ \begin{array}{ll}
\sum_x \sum_y h(x, y)f(x, y) & \text{if } (X, Y) \text{ is discrete} \\
\int_R \int_R h(x, y)f(x, y) \, dx \, dy & \text{if } (X, Y) \text{ is continuous}
\end{array} \right. \]

**Example/Theorem:** Whether or not X and Y are independent, we have \( \mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] \). In fact, if \( X_1, X_2, \ldots \) are RV’s, then \( \mathbb{E}[\sum_i X_i] = \sum_i \mathbb{E}[X_i] \).

**Theorem:** If X and Y are independent, then \( \mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y] \) and \( \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) \).

**Proof:** Easy algebra.

**Theorem:** Suppose that \( X_1, \ldots, X_n \) are independent RV’s. If \( Y = \sum_{i=1}^n X_i \), then

\[ M_Y(t) = \mathbb{E}[e^{tY}] = \mathbb{E}\left[e^{t\sum X_i}\right] = \prod_{i=1}^n \mathbb{E}[e^{tX_i}] = \prod_{i=1}^n M_{X_i}(t). \]
**Definition:** $X_1, \ldots, X_n$ form a random sample from $f(x)$ is

1. $X_1, \ldots, X_n$ are independent, and
2. Each $X_i$ has the same pdf (or pmf) $f(x)$.

**Notation:** $X_1, \ldots, X_n \overset{iid}{\sim} f(x)$. (The term “iid” reads independent and identically distributed)

**Corollary:** $X_1, \ldots, X_n$ iid implies that $M_Y(t) = [M_{X_i}(t)]^n$.

**Example:** Suppose $X_1, \ldots, X_n \overset{iid}{\sim}$ Bern($p$). Then $M_{\sum_{i=1}^n X_i}(t) = (pe^t + q)^n$. It turns out that this is the mgf for the Bin($n, p$) distribution. Thus, by a previous theorem, we have $\sum_{i=1}^n X_i \sim$ Bin($n, p$). ♦

**Example:** If $X_1, \ldots, X_n \overset{iid}{\sim} f(x)$ and $\bar{X} \equiv \sum_{i=1}^n X_i/n$, then $E[\bar{X}] = E[X_i]$ and $\text{Var}(\bar{X}) = \text{Var}(X_i)/n$. Thus, the variance decreases. ♦

### 1.8 Covariance and Correlation

**Definition:** The covariance between $X$ and $Y$ is $\text{Cov}(X, Y) \equiv E[(X - E[X])(Y - E[Y])]$. Note that $\text{Var}(X) = \text{Cov}(X, X)$.


**Proof:** Easy. ♦

**Theorem:** If $X$ and $Y$ are independent RV’s, then $\text{Cov}(X, Y) = 0$.

**Proof:** Since $X$ and $Y$ are independent, we have $E[XY] = E[X]E[Y]$. ♦

**Remark:** $\text{Cov}(X, Y) = 0$ does not imply that $X$ and $Y$ are independent!

**Theorem:** If $a$ and $b$ are constants, the $\text{Cov}(aX, bY) = ab\text{Cov}(X, Y)$.

**Definition:** The correlation between $X$ and $Y$ is

$$\rho \equiv \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.$$
Theorem: $-1 \leq \rho \leq 1$.

Proof: Follows from the Cauchy-Schwarz inequality. ♦

Remark: If $\rho \approx 1$, we say that $X$ and $Y$ have “high positive” correlation. If $\rho \approx 0$, $X$ and $Y$ have “low” correlation. If $\rho \approx -1$, there is “high negative” correlation.

Example: Suppose that $X$ is the average yards per carry gained by a University of Georgia fullback and $Y$ is his IQ. Further suppose that the joint pmf $f(x, y)$ is given in the following table.

<table>
<thead>
<tr>
<th></th>
<th>$X = 2$</th>
<th>$X = 3$</th>
<th>$X = 4$</th>
<th>$f_Y(y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y = 40$</td>
<td>0.00</td>
<td>0.20</td>
<td>0.10</td>
<td>0.3</td>
</tr>
<tr>
<td>$Y = 50$</td>
<td>0.15</td>
<td>0.10</td>
<td>0.05</td>
<td>0.3</td>
</tr>
<tr>
<td>$Y = 60$</td>
<td>0.30</td>
<td>0.00</td>
<td>0.10</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Then we have $E[X] = 2.8$, $\text{Var}(X) = 0.66$, $E[Y] = 51$, $\text{Var}(Y) = 69$, $E[XY] = \sum_x \sum_y xyf(x, y) = 140$, and

$$\rho = \frac{E[XY] - E[X]E[Y]}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = -0.415.$$  ♦

Theorem: $\text{Var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \text{Var}(X_i) + 2 \sum \sum_{i<j} a_i a_j \text{Cov}(X_i, X_j)$.

Corollary: $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$.

Corollary: $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)$.

1.9 Some Fun Distributions

First, some discrete distributions...

1.9.1 Bernoulli

$X \sim \text{Bernoulli}(p)$.

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

$E[X] = p$, $\text{Var}(X) = p(1 - p)$, $M_X(t) = pe^t + q$.

If $X_1, X_2, \ldots, X_n$ are i.i.d. Bern($p$), we say that they form a series of Bernoulli($p$) trials.
1.9.2 Binomial

$X \sim \text{Binomial}(n, p)$.

$$f(x) = \binom{n}{k} p^x (1 - p)^{n-x}, \quad x = 0, 1, \ldots, n.$$  

$E[X] = np$, $\text{Var}(X) = np(1 - p)$, $M_X(t) = (p e^t + q)^n$. If $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bern}(p)$, then $\sum_{i=1}^{n} X_i \sim \text{Bin}(n, p)$.

1.9.3 Geometric

$X \sim \text{Geom}(p)$ is the number of Bern$(p)$ trials until a success occurs. For example, “FFFS” implies that $X = 4$.

$$f(x) = (1 - p)^{x-1} p, \quad x = 1, 2, \ldots.$$  

$E[X] = 1/p$, $\text{Var}(X) = q/p^2$.

1.9.4 Negative Binomial

$X \sim \text{NegBin}(r, p)$ is the sum of $r$ i.i.d. Geom$(p)$ RV’s, i.e., the time until the $r$th success occurs. For example, “FFFSSFS” implies that NegBin$(3, p) = 7$.

$$f(x) = \binom{x-1}{r-1} (1 - p)^{x-r} p^r, \quad x = r, r + 1, \ldots.$$  

$E[X] = r/p$, $\text{Var}(X) = qr/p^2$.

1.9.5 Poisson

A counting process $N(t)$ tallies the number of “arrivals” observed in $[0, t]$. A Poisson process is a counting process satisfying the following.

i. Arrivals occur one-at-a-time.

ii. Independent increments, i.e., the numbers of arrivals in disjoint time intervals are independent.

iii. Stationary increments, i.e., the distribution of the number of arrivals only depends on the length of the time interval under observation.
$X \sim \text{Pois}(\lambda)$ is the number of arrivals that a Poisson processes experiences in one time unit, i.e., $N(1)$.

$$f(x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, \ldots.$$  

$\mathbb{E}[X] = \lambda = \text{Var}(X)$.

Now, some continuous distributions...

1.9.6 Uniform

$X \sim \text{Unif}(a, b)$.

$$f(x) = \frac{1}{b-a}, \quad a \leq x \leq b.$$  

$\mathbb{E}[X] = \frac{a+b}{2}, \quad \text{Var}(X) = \frac{(b-a)^2}{12}, \quad M_X(t) = \frac{e^{tb}-e^{ta}}{t}$.

1.9.7 Exponential

$X \sim \text{Exp}(\lambda)$.

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$  

$\mathbb{E}[X] = 1/\lambda, \quad \text{Var}(X) = 1/\lambda^2, \quad M_X(t) = \frac{\lambda}{\lambda-t}, \quad t < \lambda.$

**Theorem.** The exponential distribution has the *memoryless property*, i.e., for $s, t > 0$,  

$$\Pr(X > s + t | X > s) = \Pr(X > t).$$

By the way, the Exp($\lambda$) is the only continuous distribution with this property.

1.9.8 Gamma

$X \sim \text{Gamma}(\alpha, \lambda)$.

$$f(x) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}, \quad x \geq 0,$$

where the gamma function is  

$$\Gamma(z) \equiv \int_0^\infty t^{z-1} e^{-t} dt.$$  

$\mathbb{E}[X] = \alpha/\lambda, \quad \text{Var}(X) = \alpha/\lambda^2, \quad M_X(t) = \left(\frac{\lambda}{\lambda-t}\right)^\alpha$. If $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$, then $Y \equiv \sum_{i=1}^n X_i \sim \text{Gamma}(n, \lambda)$. The Gamma$(n, \lambda)$ is also called the Erlang$_n(\lambda)$. It has c.d.f.  

$$F_Y(y) = 1 - e^{-\lambda y} \sum_{j=0}^{n-1} \frac{(\lambda y)^j}{j!}, \quad y \geq 0.$$
1.9.9 Normal

\( X \sim \text{Nor}(\mu, \sigma^2) \).

\[
f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[ -\frac{(x - \mu)^2}{2\sigma^2} \right], \quad x \in \mathbb{R}.
\]

\( \mathbb{E}[X] = \mu, \ \text{Var}(X) = \sigma^2, \ M_X(t) = \exp\{\mu t + \frac{1}{2}\sigma^2 t^2\} \).

**Theorem (Additive Property of Normals):** Suppose that \( X_1, X_2, \ldots, X_n \) are independent with \( X_i \sim \text{Nor}(\mu_i, \sigma_i^2), \ i = 1, 2, \ldots, n \). Then

\[
Y = \sum_{i=1}^{n} a_i X_i + b \sim \text{Nor}\left( \sum_{i=1}^{n} a_i \mu_i + b, \sum_{i=1}^{n} a_i^2 \sigma_i^2 \right).
\]

**Proof:** Use m.g.f.'s.

\[
M_Y(t) = \mathbb{E}(e^{tY}) = \mathbb{E}\left( \exp\left\{ t \left( \sum_{i=1}^{n} a_i X_i + b \right) \right\} \right)
\]

\[
= e^{tb} \mathbb{E}\left( \exp\left\{ \sum_{i=1}^{n} (a_i t) X_i \right\} \right)
\]

\[
= e^{tb} \prod_{i=1}^{n} \mathbb{E}\left( e^{(a_i t) X_i} \right) \quad \text{(by independence)}
\]

\[
= e^{tb} \prod_{i=1}^{n} M_{X_i}(a_i t)
\]

\[
= e^{tb} \prod_{i=1}^{n} \exp\left\{ \mu_i (a_i t) + \frac{1}{2}\sigma_i^2 (a_i t)^2 \right\}
\]

\[
= \exp\left\{ \left( \sum_{i=1}^{n} \mu_i a_i + b \right) t + \frac{1}{2} \left( \sum_{i=1}^{n} a_i^2 \sigma_i^2 \right) t^2 \right\}. \quad \diamondsuit
\]

**Example:** Suppose \( X \sim \text{Nor}(3, 4), \ Y \sim \text{Nor}(4, 6) \), and \( X \) and \( Y \) are independent. Then

\[
2X - 3Y + 1 \sim \text{Nor}(2\mathbb{E}[X] - 3\mathbb{E}[Y] + 1, 4\text{Var}(X) + 9\text{Var}(Y)) \sim \text{Nor}(-5, 70). \quad \diamondsuit
\]

**Corollary:** If \( X \sim \text{Nor}(\mu, \sigma^2) \), then \( aX + b \sim \text{Nor}(a\mu + b, a^2\sigma^2) \).

**Corollary:** If \( X \sim \text{Nor}(\mu, \sigma^2) \), then \( Z \equiv \frac{X - \mu}{\sigma} \sim \text{Nor}(0, 1) \), the standard normal distribution.

**Notation:** The standard normal's p.d.f. is \( \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \), and the c.d.f. is \( \Phi(x) \), which is usually tabled. For example, \( \Phi(1.96) \approx 0.975 \).
1.10 A First Look at Some Limit Theorems

**Corollary** (of theorem on linear combinations of normals from previous subsection): If $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu, \sigma^2)$, then the sample mean

$$\bar{X} \equiv \frac{1}{n} \sum_{i=1}^{n} X_i \sim \text{Nor}(\mu, \sigma^2/n).$$

This is a special case of the *Law of Large Numbers*, which says that $\bar{X}$ approximates $\mu$ well as $n$ becomes large.

**Markov’s Inequality:** If $X$ is a non-negative RV, then for all $\epsilon > 0$, we have

$$\Pr(X \geq \epsilon) \leq \frac{\mathbb{E}[X]}{\epsilon}.$$

**Proof:** Since $X$ is non-negative,

$$\mathbb{E}[X] = \int_{0}^{\infty} xf(x) \, dx \geq \int_{\epsilon}^{\infty} xf(x) \, dx \geq \epsilon \int_{\epsilon}^{\infty} f(x) \, dx = \epsilon \Pr(X \geq \epsilon). \quad \diamond$$

**Chebychev’s Inequality:** For any RV $X$ and for all $\epsilon > 0$, we have

$$\Pr(|X - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\text{Var}(X)}{\epsilon^2}.$$

**Proof:** Uses Markov’s Inequality; see any probability test. \quad \diamond

**Bonus Generalization:** $\Pr(|X| \geq \epsilon) \leq \mathbb{E}[|X|^r]/\epsilon^r$.

**Remark:** These inequalities are usually pretty crude!

**Example:** Suppose that $X \sim \text{Unif}(0, 1)$. Then the probability that $X$ deviates from its mean by at least $1/4$ is exactly

$$\Pr\left(|X - \frac{1}{2}| \geq \frac{1}{4}\right) = 1 - \Pr\left(|X - \frac{1}{2}| < \frac{1}{4}\right) = 1 - \Pr\left(-\frac{1}{4} < X - \frac{1}{2} < \frac{1}{4}\right) = 1 - \Pr\left(\frac{1}{4} < X < \frac{3}{4}\right) = \frac{1}{2}.$$

Meanwhile, by Chebychev (with $\mathbb{E}[X] = 1/2$, $\text{Var}(X) = 1/12$, and $\epsilon = 1/4$), we have

$$\Pr\left(|X - \frac{1}{2}| \geq \frac{1}{4}\right) \leq \frac{\text{Var}(X)}{\epsilon^2} = \frac{16}{12} = \frac{4}{3},$$

which is a very crude upper bound indeed! \quad \diamond
**Definition:** The sequence of random variables $Y_1, Y_2, \ldots$ with respective c.d.f.’s $F_{Y_1}(y), F_{Y_2}(y), \ldots$ converges in distribution to the random variable $Y$ having c.d.f. $F_Y(y)$ if $\lim_{n \to \infty} F_{Y_n}(y) = F_Y(y)$ for all $y$ belonging to the continuity set of $Y$ (i.e., the set of all points $y$ at which $F_Y(y)$ is continuous). Notation: $Y_n \xrightarrow{D} Y$. (Also sometimes called convergence in law or weak convergence.)

**Idea:** If $Y_n \xrightarrow{D} Y$, then you would expect to be able to approximate the distribution of $Y_n$ by the limiting distribution of $Y$, at least for large enough $n$.

**Central Limit Theorem:** If $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} f(x)$ with mean $\mu$ and variance $\sigma^2$, then

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{D} \text{Nor}(0,1),$$

where $\bar{X}_n$ is the sample mean. Thus, the c.d.f. of $Z_n$ approaches that of the standard normal as $n$ increases. The CLT usually works pretty well if the pdf/pmf is fairly symmetric and $n \geq 15$.

**Example:** Suppose that $X_1, X_2, \ldots, X_{100} \overset{\text{iid}}{\sim} \text{Exp}(1)$. Then

$$\Pr \left(90 \leq \sum_{i=1}^{100} X_i \leq 110\right) = \Pr \left(\frac{90 - 100}{\sqrt{100}} \leq Z_{100} \leq \frac{110 - 100}{\sqrt{100}}\right)$$

$$= \Pr(-1 \leq Z_{100} \leq 1) \approx \Pr(-1 \leq \text{Nor}(0,1) \leq 1) = 2\Phi(1) - 1 \approx 0.683. \Diamond$$

**Definition:** The sequence of random variables $Y_1, Y_2, \ldots$ is said to converge in probability to $Y$ (often a constant) if for all $\epsilon > 0$, $\Pr(|Y_n - Y| > \epsilon) \to 0$ as $n \to \infty$. Notation: $Y_n \xrightarrow{P} Y$.

**Theorem:** $Y_n \xrightarrow{P} Y$ implies $Y_n \xrightarrow{D} Y$. In other words, convergence in probability is a bit stronger than convergence in distribution.

**Weak Law of Large Numbers:** If $X_1, X_2, \ldots$ are i.i.d. with mean $\mu$, then $\bar{X}_n \xrightarrow{P} \mu$. Why is this called the weak LLN? Simply because there’s a stronger one coming up later.

**Continuous Mapping Theorem:** If $Y_n \xrightarrow{D} Y$ and $g(\cdot)$ is a nice, continuous function, then $g(Y_n) \xrightarrow{D} g(Y)$. The CMT is often useful for characterizing the convergence of nasty functions of the $Y_i$’s.

**Definition:** The sequence of random variables $Y_1, Y_2, \ldots$ is said to converge in $r$th mean to $Y$ (often a constant) if $\mathbb{E}[|Y_n - Y|^r] \to 0$ as $n \to \infty$. Notation: $Y_n \xrightarrow{r} Y$. 
Theorem: \( Y_n \xrightarrow{r} Y \) implies \( Y_n \xrightarrow{P} Y \). In other words, convergence in \( r \)th mean is a bit stronger than convergence in probability.

Proof: Follows immediately from bonus version of Chebychev. \( \diamond \)

Definition: The sequence of random variables \( Y_1, Y_2, \ldots \) converges almost surely (or with probability one) to \( Y \) if \( \Pr(Y_n \text{ converges to } Y) = 1 \) as \( n \to \infty \). Notation: \( Y_n \xrightarrow{a.s.} Y \).

Theorem: \( Y_n \xrightarrow{a.s.} Y \) implies \( Y_n \xrightarrow{P} Y \). In other words, convergence almost surely is a bit stronger than convergence in probability. Note that this says nothing about how almost sure and \( r \)th mean convergence relate to each other.

Strong Law of Large Numbers: If \( X_1, X_2, \ldots \) are i.i.d. with mean \( \mu \), then \( \bar{X}_n \xrightarrow{a.s.} \mu \).