Lecture 4: Discrete Random Variables
Definition 17. The probability mass function (PMF) of a discrete RV $X$ taking on values in the countable set $\mathcal{X} = \{x_1, x_2, \ldots\} \subset \mathbb{R}$ is the collection of “jump” values of its cdf $F_X$. In particular,

$$p_X[i] = F_X(x_i) - \lim_{h \downarrow 0} F_X(x_i - h)$$

for all $i = 1, 2, \ldots$. Equivalently,

$$p_X[i] = \mathbb{P}(X = x_i)$$
Notation

- Shortcut equivalent notation: \( \{ p_X(x) : x \in \mathcal{X} \} \).

- Also, we shall occasionally use \( p = (p_1, p_2, \ldots, p_n) \) to indicate a probability vector such that

\[
p_i = p_X[i] = \mathbb{P}(X = x_i)
\]

- Also, we shall occasionally use \( \{ p_i \} = \{ p_1, p_2, \ldots \} \) to indicate a sequence of probabilities such that, again \( p_i = p_X[i] = \mathbb{P}(X = x_i) \).

- When \( |\mathcal{X}| = n < \infty \), the pmf probability vector \( p \) lives in the \( n - 1 \) dimensional simplex embedded in \( \mathbb{R}^n \), defined by

\[
p_i \geq 0, \quad i = 1, \ldots, n, \quad \sum_{i=1}^{n} p_i = 1
\]
Example: binomial distribution

Let \( \{X_i : i = 1, \ldots, n\} \) denote a collection of i.i.d. Bernoulli RVs with pmf \((p, 1 - p)\). Then,

\[
X = \sum_{i=1}^{n} X_i
\]

Each configuration with \( k \) ones and \( n - k \) zeros has identical probability \( p^k(1 - p)^{n-k} \).

The event \( \{X = k\} \) is the union of all configurations with \( k \) ones. There are exactly \( \binom{n}{k} \) such configurations, therefore

\[
\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}
\]

The pmf \( p_X[k] = \binom{n}{k} p^k (1 - p)^{n-k} \) for \( k = 0, 1, \ldots, n \) is called Binomial distribution.
Example: Poisson distribution

Consider the same setting as before, but now let \( p = \frac{\lambda}{n} \), for fixed \( \lambda \) and \( n \to \infty \). The Binomial probability becomes

\[
\binom{n}{k} \left( \frac{\lambda}{n} \right)^k \left( 1 - \frac{\lambda}{n} \right)^{n-k} \to \frac{e^{-\lambda} \lambda^k}{k!}
\]

The pmf \( p_X[k] = \frac{e^{-\lambda} \lambda^k}{k!} \) for \( k = 0, 1, 2, \ldots \) is called Poisson distribution.
Example: multinomial distribution

Suppose we conduct \( n \) independent trials, each of which can have \( r \) possible outcomes, with probability \( (p_1, p_2, \ldots, p_r) \). Let \( X_j \) denote the number of outcomes equal to \( j \). Then, for each \( j_1, \ldots, j_r \) such that \( j_i \geq 0 \) and \( \sum_{i=1}^{r} j_i = n \), the probability of having \( j_1 \) outcomes equal to 1, \( j_2 \) outcomes equal to 2 etc ..., is given by

\[
p_{X_1, \ldots, X_r}[j_1, \ldots, j_r] = \mathbb{P}(X_1 = j_1, X_2 = j_2, \ldots, X_r = j_r) = \frac{n!}{j_1! j_2! \cdots j_r!} \prod_{i=1}^{r} p_i^{j_i}
\]

This joint pmf is called multinomial distribution.
Example: Geometric distribution

Consider to flip a biased coin and stop when we find $T$. Hence, the possible sequences or experiments, or “runs”, are given by

$$T, HT, HHT, HHHT, HHHHT, \ldots$$

Letting $p$ denote the probability of $T$, we have

$$\mathbb{P}(T) = p, \quad \mathbb{P}(HT) = (1-p)p, \quad \mathbb{P}(HHT) = (1-p)^2p, \ldots$$

The associated RV has pmf $p_X[k] = (1-p)^{k-1}p$, called Geometric distribution. Notice that $\mathbb{P}(X > k) = (1-p)^k$.

The geometric distribution has the following memoryless property: for $k \geq 0$ and $m \geq 0$,

$$\mathbb{P}(X = m + k | X > m) = \frac{\mathbb{P}(X=m+k, X>m)}{P(X>m)} = \frac{\mathbb{P}(X=m+k)}{\mathbb{P}(X>m)} = \frac{(1-p)^{m+k-1}p}{(1-p)^m} = (1-p)^{k-1}p = \mathbb{P}(X = k)$$
Example: hypergeometric distribution

Consider the experiment of extracting $r$ balls from a bin that contains $n_1$ red balls and $n - n_1$ white balls. Let $X$ denote the number of red balls in the sample. Then, for $k = 0, \ldots, \min\{n_1, r\}$ we have

$$p_X[k] = \frac{\binom{n_1}{k} \binom{n - n_1}{r - k}}{\binom{n}{r}}$$

This is called hypergeometric distribution.

Notice that for $r \leq n_1$ we have

$$\sum_{k=0}^{r} \binom{n_1}{k} \binom{n - n_1}{r - k} = \binom{n}{r}$$
Example: negative binomial distribution

This is a generalization of the Geometric distribution. The Geometric distribution may be interpreted as the probability distribution of the “waiting time” \( X \), indicating the time till the first success event occurs in a sequence of independent Bernoulli trials.

Let \( X_r \) denote the waiting time for the \( r \)-th success. For example, consider the sequence of biased coin flips: we stop when the \( r \)-th “T” occurs. Then,

\[
p_{X_r}[k] = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \quad k = r, r+1, r+2, \ldots
\]

In order to see this, notice that for \( X_r = k \) it must be that the \( k \)-th event is the \( r \)-th success. Then, in the remaining \( k-1 \) events we have \( r-1 \) successes. The probability of this is precisely given by the binomial term

\[
\binom{k-1}{r-1} p^{r-1} (1-p)^{k-1-(r-1)}
\]
Finally, we multiply by $p$ (the probability of the $r$-th success occurring at the $k$ trial) and we have the above pmf.

We notice also that letting $W_1$ denote the waiting time for the first success, $W_2$ the inter-success time between first and second and so on, we have

$$X_r = W_1 + W_2 + \cdots + W_r$$

It is easy to see that each $W_i$ is a geometric RV with parameter $p$. Furthermore, because of the memoryless property said before, the RVs $W_1, \ldots, W_r$ are independent.
Consider $r$ bins and $n$ balls. The possible partitions of $n$ balls into $r$ bins are $r^n$. In fact, each partition (assignment of a ball to a bin) corresponds to a vector of length $n$ of labels in $\{1, \ldots, r\}$. These partitions, or arrangements, are assumed equiprobable. The number of placements with occupancy numbers $j_1, \ldots, j_r$ is given by the multinomial coefficient $\frac{n!}{j_1! \cdots j_r!}$, as seen before. Define the random vector $(X_1, \ldots, X_r)$ such that $X_i$ is the content of bin $i$. This random vector may correspond to the state of some physical system. Then, its pmf is given by

$$p_{X_1, \ldots, X_r}[j_1, \ldots, j_r] = \frac{n!}{j_1! \cdots j_r!} r^{-n}$$

Notice that this is a special case of the multinomial distribution for the case $p_i = 1/r$.

This pmf is called the Maxwell-Boltzmann distribution, and has been proposed to model the distribution of particles in a region, partitioned into $r$ small regions (bins).
Example: Bose-Einstein distribution

Consider a physical system with $n$ particles, and partition the space into $r$ non-overlapping subregions (bins). The state of the system is described by the distribution of particles in the bins. Differently from the Maxwell-Boltzmann case, where the particles are *distinguishable* (you can think of each particle as a ball with a unique identified number from 1 to $n$ attached to it), the Bose-Einstein distribution assumes that the particles are *indistinguishable* and all *distinguishable* configurations are equiprobable. The balls arrangements can differ only by their occupancy numbers $j_1, \ldots, j_r$ such that $\sum_{i=1}^{r} j_i = n$.

The number of possible distinguishable arrangements is given by the coefficient (already introduced)

$$A_{n,r} = \binom{r + n - 1}{n}$$
Hence, letting again the state of the system be denoted by the random vector 
\((X_1, \ldots, X_r)\), the Bose-Einstein distribution corresponds to the pmf

\[
p_{X_1, \ldots, X_r}[j_1, \ldots, j_r] = \begin{cases} 
\frac{1}{A_{n,r}} & \text{if } \sum_{i=1}^{r} j_i = n, \ j_i \geq 0 \\
0 & \text{elsewhere}
\end{cases}
\]
Consider again a physical system with \( n \) indistinguishable particles, and partition the space into \( r \) non-overlapping subregions (bins). Under the Fermi-Dirac assumption the particles are repulsive, such that each bin cannot contain more than one particle. This implies \( n \leq r \). Then, all valid configurations are equiprobable. The resulting Fermi-Dirac distribution is given by

\[
p_{X_1,\ldots,X_r}[j_1,\ldots,j_r] = \begin{cases} 
\frac{1}{\binom{r}{n}} & \text{if } \sum_{i=1}^{r} j_i = n, \quad j_i \in \{0, 1\} \\
0 & \text{elsewhere}
\end{cases}
\]
Example: Negative hypergeometric distribution

Consider a population of \( n \) fishes. We capture \( r \) fishes, mark them with a red spot and release them in the same lake again. Then, we catch fishes again, sequentially and without releasing them again, and we stop when we have captured \( m \) marked fishes. The number \( X \) of captured fishes is a discrete RV that takes on the values in \( \{m, m + 1, \ldots, n - r + m\} \).

The number of distinct arrangements of fishes (they are distinguishable only by their red dot), is given by \( \binom{n}{r} \). In order to stop at the \( k \)-th capture, we need that in the first \( k - 1 \) captured fishes we have found \( m - 1 \) red dots, and the \( k \)-th fish has a red dot. The remaining fishes can be arranged in any arbitrary manner. We have precisely

\[
\binom{k-1}{m-1} \binom{n-k}{r-m}
\]

ways of making such arrangements. It follows that

\[
p_X[k] = \frac{\binom{k-1}{m-1} \binom{n-k}{r-m}}{\binom{n}{r}} = \frac{r \binom{n-r}{k-m} \binom{r-1}{m-1}}{n \binom{n-1}{k-1}}
\]
Independence

Definition 18. Two RVs $X$ and $Y$ are independent if and only if the events
$\{X \leq x\}$ and $\{Y \leq y\}$ are independent for all $x, y \in \mathbb{R}$.

- In the case of discrete RVs, we can always write

$$X = \sum_i x_i I_{A_i}$$

where $I_A$ denotes the indicator function of the event $A$ and where $A_i = \{X = x_i\}$ by construction.

- Let $X = \sum_i x_i I_{A_i}$ and $Y = \sum_i y_i I_{B_i}$. Then, $X$ and $Y$ are independent if and only if the collection of events $\{A_1, A_2, \ldots, B_1, B_2, \ldots\}$ is an independent set.

- As already said, if $X$ and $Y$ are independent, then

$$p_{X,Y}[i,j] = p_X[i]p_Y[j]$$
• More in general, for any countable collection $X_1, X_2, \ldots$ of independent RVs, we have that

$$
P(X_i = x_i : i \in S) = \prod_{i \in S} P(X_i = x_i)$$

and for any finite index set $S \subset \{1, 2, \ldots\}$ and any $|S|$-tuple of values $\{x_i : i \in S\}$.

**Lemma 17.** If $X$ and $Y$ are discrete and independent, and $f, g$ are functions $\mathbb{R} \to \mathbb{R}$, then $X' = f(X)$ and $Y' = g(Y)$ are also independent.

**Example:** Poisson coin flips. A coin is tossed $N$ times, where $N$ is a Poisson RV with parameter $\lambda$. Let $X$ denote the number of “heads” and $Y$ denote the number of “tails”. Then, $X$ and $Y$ are independent. (Notice that for fixed deterministic $N$ this is clearly not true since $X = N - Y$).
Expectation

• Let $a_1, a_2, \ldots, a_N$ denote a sequence of real values. The arithmetic mean of these values is defined as

$$m = \frac{1}{N} \sum_{j=1}^{N} a_j$$

• Now, suppose to perform a random experiment whose outcome can be represented by a discrete RV $X$, taking on values in the set $\{x_1, \ldots, x_n\}$. We record $N$ realizations of $X$, and call them $a_1, a_2, \ldots, a_N$.

• For independent experiments, it is likely that a fraction $N_i \approx N p_X[i]$ of times we are going to observe the value $x_i$. It follows that, for very large $N$,

$$m = \frac{1}{N} \sum_{j=1}^{N} a_j = \frac{1}{N} \sum_{i=1}^{n} x_i N_i \approx \sum_{i=1}^{n} x_i p_X[i]$$

Copyright G. Caire (Sample Lectures)
This intuition yields the following rigorous definition:

**Definition 19.** The *mean value or expectation or expected value* of a discrete RV $X$ is given by

$$
\mathbb{E}[X] = \sum_x x p_X(x)
$$

*whenever this sum is absolutely convergent.*

*Notice:* if the above sum is not absolutely convergent, then we say that the expectation of $X$ does not exist.
Lemma 18. If $X$ has pmf $p_X(x)$ and $g : \mathbb{R} \rightarrow \mathbb{R}$, then

$$E[g(X)] = \sum_x g(x)p_X(x)$$

whenever the above sum is absolutely convergent.

- The operator $E[\cdot]$ is called expectation operator, and has properties summarized by the following

Lemma 19. a) if $X \geq 0$ (i.e., $P(X \geq 0) = 1$), then $E[X] \geq 0$.

b) linearity: for given constants $a, b \in \mathbb{R}$, $E[aX + bY] = aE[X] + bE[Y]$.

c) a constant RV $X = a$ (i.e., $p_X(a) = 1$), has $E[X] = a$.

Lemma 20. If $X$ and $Y$ are independent, then $E[XY] = E[X]E[Y]$. 
**Definition 20.** If $k$ is a positive integer, the $k$-th moment of a RV $X$ is defined as

$$m_k = \mathbb{E}[X^k]$$

The $k$-th central moment of $X$ is defined as

$$\mu_k = \mathbb{E}[(X - \mathbb{E}[X])^k]$$

- The second central moment $\mu_2$ is very important and has a special name: it is called variance.
- There are a few usual equivalent expressions for the variance

$$\mu_2 = \text{Var}(X) = \sigma_X^2 = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = m_2 - m_1^2$$
**Definition 21.** Two RVs $X$ and $Y$ are called **uncorrelated** if $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. 

**Lemma 21.** For two RVs $X$ and $Y$,

a) $\text{Var}(aX) = a^2\text{Var}(X)$, for $a \in \mathbb{R}$ constant.

b) $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$. The term $\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$ is called **covariance** of $X$ and $Y$.

c) If $X$ and $Y$ are uncorrelated, then $\text{Cov}(X, Y) = 0$.

**Definition 22.** The **correlation coefficient** of two RVs $X$ and $Y$ is given by

$$
\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}
$$
Theorem 1. For two jointly distributed RVs $X$ and $Y$, we have

$$|\mathbb{E}[XY]|^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$$

with equality if and only if $\mathbb{P}(aX = bY) = 1$ for some $a, b \in \mathbb{R}$ of which at least one is non-zero.
Indicator functions

- A very useful class of discrete (binary) random variables are the indicator functions of events, defined as follows:

\[ I_A = \begin{cases} 
1 & \text{if } A \text{ occurs} \\
0 & \text{otherwise} 
\end{cases} \]

- A fundamental property of indicator functions is the following:

\[ \mathbb{P}(A) = \mathbb{E}[I_A] \]

- We shall spend a few examples in order to illustrate the use of indicator functions.
Conditional distributions

Let $X$ and $Y$ be two jointly distributed discrete RVs defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

**Definition 23.** The conditional pmf of $Y$ given $X = x_i$, written as $p_{Y|X}[j|i]$ or, more briefly, as $p_{Y|X}(y_j|x_i)$, is defined as

$$p_{Y|X}(y_j|x_i) = \mathbb{P}(Y = y_j|X = x_i) = \frac{\mathbb{P}(Y = y_j, X = x_i)}{\mathbb{P}(X = x_i)}$$

and it is defined for any $x_i$ such that $\mathbb{P}(X = x_i) > 0$. 

Notice: you have to think of $p_{Y|X}(y|x)$ as a two-dimensional array with elements indexed by all possible values $y$ and $x$, that is, we have a column for each value $x$ such that $\mathbb{P}(X = x) > 0$. Also, summing each column with respect to $y$, we have $\sum_y p_{Y|X}(y|x) = 1$, since for each conditioning value $x$, $p_{Y|X}(\cdot|x)$ is a well-defined pmf for $Y$.

Notice: if $X$ and $Y$ are statistically independent, then $p_{Y|X}(y|x) = p_Y(y)$. 

Copyright G. Caire (Sample Lectures)
Definition 24. The conditional expectation of $Y$ given $X = x$ is defined as the expectation of the pmf $p_{Y|X}(\cdot|x)$, i.e.,

$$\mathbb{E}[Y|X = x] = \sum_{y} yp_{Y|X}(y|x)$$

where the conditional pmf is assumed to exist.

The value $\mathbb{E}[Y|X = x]$ in general depends on $x$. Hence, we can define the function of real argument $\psi(x) = \mathbb{E}[Y|X = x]$. We use the notation $\mathbb{E}[Y|X]$ to denote the function of random variable $\psi(X)$. Since $X$ is a RV, then also $\psi(X) = \mathbb{E}[Y|X]$ is a RV.
Theorem 2. *Iterated expectation.* The expectation of the conditional expectation is the expectation itself, i.e., in formulas,

\[ \mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]] \]

- More explicitly, we may rewrite the statement of the theorem as

\[ \mathbb{E}[Y] = \sum_x \mathbb{E}[Y|X = x] p_X(x) \]

- This theorem is very useful, since it allows to break down the computation of an expectation, which may be very complicated, into several possibly simpler steps.
Theorem 3. The conditional expectation $\psi(X) = \mathbb{E}[Y|X]$ satisfies

$$\mathbb{E}[\psi(X)g(X)] = \mathbb{E}[Yg(X)]$$

for any function $g : \mathbb{R} \rightarrow \mathbb{R}$ such that both expectations exist.
A hen lays $N$ eggs, where $N$ is a Poisson distributed RV with parameter $\lambda$. Each egg hatches with probability $p$ independently of the other eggs. Let $K$ be the number of resulting chicks. Find $\mathbb{E}[K|\mathcal{N}]$, $\mathbb{E}[K]$ and $\mathbb{E}[\mathcal{N}|K]$. 
Maximum and Minimum of a set of RVs

- Given a set $X^n = \{X_1, \ldots, X_n\}$ of RVs with joint pmf $p_{X^n}(x_1, \ldots, x_n)$, we wish to find the cdf of

$$Y = \max\{X_1, \ldots, X_n\}, \quad \text{and of} \quad Z = \min\{X_1, \ldots, X_n\}$$

- For the maximum, we have

$$\{Y \leq y\} = \{\max\{X_1, \ldots, X_n\} \leq y\} = \bigcap_{i=1}^{n} \{X_i \leq y\}$$

Hence

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}\left(\bigcap_{i=1}^{n} \{X_i \leq y\}\right)$$
• If the RVs are mutually independent, then

\[
F_Y(y) = \prod_{i=1}^{n} \mathbb{P}(X_i \leq y) = \prod_{i=1}^{n} F_{X_i}(y)
\]

• For the minimum, we have

\[
\{Z \leq z\} = \{\min\{X_1, \ldots, X_n\} \leq z\} = \bigcup_{i=1}^{n} \{X_i \leq z\}
\]
• Hence

\[ F_Z(z) = P(Z \leq z) = P\left( \bigcup_{i=1}^{n} \{X_i \leq z\} \right) = 1 - P\left( \bigcap_{i=1}^{n} \{X_i > z\} \right) \]

If the RVs are mutually independent, then

\[ F_Z(z) = 1 - \prod_{i=1}^{n} P(X_i > z) = 1 - \prod_{i=1}^{n} (1 - F_{X_i}(z)) \]
Sums of integer-valued RVs

**Theorem 4.** Let $X$ and $Y$ be integer-valued RVs, and define $Z = X + Y$. Then, we have

$$p_Z(z) = \mathbb{P}(X + Y = z) = \sum_x p_{X,Y}(x, z - x)$$

**Corollary 1.** If $X$ and $Y$ are independent, then

$$p_Z(z) = \sum_x p_X(x)p_Y(z - x)$$

which can be interpreted as the “discrete” convolution of two “discrete-time signals”, $p_X(k)$ and $p_Y(k)$, for $k \in \mathbb{Z}$, evaluated in $z$. 
End of Lecture 4