CS145: Lecture 7 Outline

- Conditional Expectations
- Variance of Random Variables
Consider two random variables $X, Y$.
Suppose range of $X$ is size $N$, range of $Y$ is size $M$.

The joint probability mass function or joint distribution of two variables:

$$p_{XY}(x, y) = P(X = x \text{ and } Y = y)$$

$$p_{XY}(x, y) \geq 0, \quad \sum_x \sum_y p_{XY}(x, y) = 1.$$ 

The joint distribution is uniquely specified by $NM-1$ numbers.
Expectation of Multiple Variables

- The *expectation* or *expected value* of a function of two discrete variables:

$$E[g(X, Y)] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} g(x, y) p_{XY}(x, y)$$

- A similar formula applies to functions of 3 or more variables

- Expectations of sums of functions are sums of expectations:

$$E[g(X) + h(Y)] = E[g(X)] + E[h(Y)] = \left[ \sum_{x \in \mathcal{X}} g(x) p_X(x) \right] + \left[ \sum_{y \in \mathcal{Y}} h(y) p_Y(y) \right]$$

- This is always true, *whether or not X and Y are independent*

- Specializing to *linear functions*, this implies that:

$$E[aX + bY + c] = aE[X] + bE[Y] + c$$
Marginal Probability Distributions

The joint probability mass function or joint distribution of two variables:

\[ p_{XY}(x, y) = P(X = x \text{ and } Y = y) \]

The range of each variable defines a partition of the sample space, so the marginal distributions can be computed from the joint distribution:

\[ p_X(x) = P(X = x) = \sum_y p_{XY}(x, y) \]
\[ p_Y(y) = P(Y = y) = \sum_x p_{XY}(x, y) \]

The marginal distributions are defined by \((N-1)+(M-1)\) numbers. Many joint distributions may have the same marginals.
### Conditional Probability Distributions

<table>
<thead>
<tr>
<th>$X = 1$</th>
<th>$p_{XY}(x, y)$</th>
<th>$Y = 8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{X</td>
<td>Y}(x \mid 1)$</td>
<td></td>
</tr>
<tr>
<td>$p_{X</td>
<td>Y}(x \mid 8)$</td>
<td></td>
</tr>
</tbody>
</table>

- By the definition of conditional probability:

\[
P(X = x \mid Y = y) = \frac{P(X = x \text{ and } Y = y)}{P(Y = y)}
\]

- The **conditional probability mass function** is then:

\[
p_{X|Y}(x \mid y) = P(X = x \mid Y = y) = \frac{p_{XY}(x, y)}{p_Y(y)} = \frac{p_{XY}(x, y)}{\sum_{x'} p_{XY}(x', y)}
\]
Given that I observe $Y=y$, the conditional expectation of $X$ equals

$$E[X \mid Y = y] = \sum_{x \in \mathcal{X}} x \cdot p_{X \mid Y}(x \mid y)$$

If $X$ and $Y$ are not independent, observing $Y=y$ may change the mean of $X$. 

$$p_{X \mid Y}(x \mid y) = P(X = x \mid Y = y) = \frac{p_{XY}(x, y)}{p_Y(y)} = \frac{p_{XY}(x, y)}{\sum_{x'} p_{XY}(x', y)}$$
Given that I observe $Y=y$, the conditional expectation of $X$ equals

$$E[X \mid Y = y] = \sum_{x \in X} x p_{X \mid Y}(x \mid y)$$

- If $X$ and $Y$ are not independent, observing $Y=y$ may change the mean of $X$. 

Given $Y = \{X \geq 2\}$ is observed,

$$p_{X \mid Y}(x \mid y) =$$

$$E[X \mid Y] = 3$$

$$E[X] = 2.5$$
Total Expectation Theorem

Applying the definitions of joint, marginal, and conditional distributions:

$\mathbb{E}[X] = \sum_{y \in Y} p_Y(y) \mathbb{E}[X \mid Y = y]$  

Mean is a weighted average of (possibly simpler) conditional means.
Example 2.14. Messages transmitted by a computer in Boston through a data network are destined for New York with probability 0.5, for Chicago with probability 0.3, and for San Francisco with probability 0.2. The transit time $X$ of a message is random. Its mean is 0.05 secs if it is destined for New York, 0.1 secs if it is destined for Chicago, and 0.3 secs if it is destined for San Francisco. Then, $E[X]$ is easily calculated using the total expectation theorem as

$$E[X] = 0.5 \cdot 0.05 + 0.3 \cdot 0.1 + 0.2 \cdot 0.3 = 0.115 \text{ secs.}$$

*Do not need to know full distribution of transmission times to each city to compute overall mean.*

$$E[X] = \sum_{y \in Y} p_Y(y) E[X \mid Y = y]$$

*Mean is a weighted average of (possibly simpler) conditional means.*
A geometric random variable $X$ has parameter $p$, countably infinite range: $p_X(k) = (1 - p)^{k-1}p$, $\mathcal{X} = \{1, 2, 3, \ldots\}$

The expected value (proved earlier by summation identity) equals:

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} k(1 - p)^{k-1}p = \frac{1}{p}$$

Interpretation: How many trials are needed before the first occurrence of an event with probability $p$?
A geometric random variable $X$ has parameter $p$, countably infinite range:

$$p_X(k) = (1 - p)^{k-1} p$$

$$\mathcal{X} = \{1, 2, 3, \ldots\}$$

**Memoryless:** For any integer $c > 0$, if I observe that $X > c$, then $Y = X - c$ has the same geometric PMF:

$$p_{X-c}(k \mid X > c) = (1 - p)^{k-1} p, \quad k = 1, 2, \ldots$$

**Interpretation:** Because coin tosses are independent, if I tell you that the first $c$ tosses are tails, it does not reduce the expected number of future tosses until first head.
A geometric random variable $X$ has parameter $p$, countably infinite range:

$$p_X(k) = (1 - p)^{k-1}p$$

$\mathcal{X} = \{1, 2, 3, \ldots\}$

Memoryless: For any integer $c > 0$, if I observe that $X > c$, then $Y = X - c$ has same geometric PMF:

$$p_{X-c}(k \mid X > c) = (1 - p)^{k-1}p, \quad k = 1, 2, \ldots$$

Split computation of mean into two cases:

$$E[X] = P(X = 1)E[X \mid X = 1] + P(X > 1)E[X \mid X > 1]$$

$$E[X] = p \times 1 + (1 - p) \times (1 + E[X])$$

$$E[X] = \frac{1}{p}$$
CS145: Lecture 7 Outline

- Conditional Expectations
- Variance of Random Variables
The expectation or expected value of a discrete random variable is:

$$E[X] = \sum_{x \in \mathcal{X}} x p_X(x)$$

The expectation is a single number, not a random variable. It encodes the “center of mass” of the probability distribution:

$$x_{\text{min}} \leq E[x] \leq x_{\text{max}}$$

$$x_{\text{min}} = \min\{x \mid x \in \mathcal{X}\}$$

$$x_{\text{max}} = \max\{x \mid x \in \mathcal{X}\}$$
Variance

- The **variance** is the *expected squared deviation* of a random variable from its mean (we will show later that these definitions are equivalent):

  \[
  \text{Var}[X] = E[(X - E[X])^2] = \sum_{x \in \mathcal{X}} (x - E[X])^2 p_X(x)
  \]

  \[
  \text{Var}[X] = E[X^2] - E[X]^2 = \left[\sum_{x \in \mathcal{X}} x^2 p_X(x)\right] - \left[\sum_{x \in \mathcal{X}} x p_X(x)\right]^2
  \]

- Intuition: If the variance is large, then it is more frequent (probable) for random variable $X$ to take values distant from its mean.

- The **variance** is always *non-negative*:

  \[
  \text{Var}[X] \geq 0 \text{ because } (x - E[X])^2 \geq 0 \text{ for all } x.
  \]
The variance is the expected squared deviation of a random variable from its mean (we will show later that these definitions are equivalent):
\[
\text{Var}[X] = E[(X - E[X])^2] = \sum_{x \in \mathcal{X}} (x - E[X])^2 p_X(x)
\]

\[
\text{Var}[X] = E[X^2] - E[X]^2 = \left[ \sum_{x \in \mathcal{X}} x^2 p_X(x) \right] - \left[ \sum_{x \in \mathcal{X}} x p_X(x) \right]^2
\]

By definition, the standard deviation is the square root of the variance:
\[
\sigma_X = \text{Std}[X] = \sqrt{\text{Var}[X]}
\]

If \( X \) is the location in meters of a self-driving car along a road, then \( E[X] \) is the average location (in meters), \( \text{Var}[X] \) is the average squared distance from the mean (in meters squared), and \( \text{Std}[X] \) is a positive distance (in meters).
The variance is the expected squared deviation of a random variable from its mean (we will show later that these definitions are equivalent):

\[
\text{Var}[X] = E[(X - E[X])^2] = \sum_{x \in X} (x - E[X])^2 p_X(x)
\]

\[
\text{Var}[X] = E[X^2] - E[X]^2 = \left[ \sum_{x \in X} x^2 p_X(x) \right] - \left[ \sum_{x \in X} x p_X(x) \right]^2
\]

Terminology: Moments of random variables

\[
E[X] = \text{first moment or mean of } X
\]

\[
E[X^2] = \text{second moment of } X
\]

\[
E[X^p] = p^{\text{th}} \text{ moment of } X
\]
Example: Uniform distribution on \( \{0, 1, \ldots, n\} \)

First moment:

\[
E[X] = 0 \times \frac{1}{n+1} + 1 \times \frac{1}{n+1} + \cdots + n \times \frac{1}{n+1} = \frac{n(n+1)}{2(n+1)} = \frac{n}{2}
\]

Second moment:

\[
E[X^2] = \frac{1}{n+1} \left[ 0^2 + 1^2 + 2^2 + \cdots + n^2 \right] = \frac{n(n+1)(2n+1)}{6(n+1)} = \frac{n(2n+1)}{6}
\]

Variance:

\[
Var[X] = E[X^2] - E[X]^2 = \frac{n(n+2)}{12}
\]

\[
Var[X] = E \left[ \left( X - \frac{n}{2} \right)^2 \right]
\]
Bernoulli Probability Distribution

- A Bernoulli or indicator random variable $X$ has one parameter $p$:
  $p_X(1) = p, \quad p_X(0) = 1 - p, \quad X = \{0, 1\}$

- For an indicator variable, expected values are probabilities:
  $E[X] = p$

Examples:
- Flip a possibly biased coin with probability of coming up heads $p$
- A user answers a true/false question in an online survey
- Does it snow or not on some day
Bernoulli Probability Distribution

- A **Bernoulli** or *indicator* random variable $X$ has one parameter $p$:
  \[ p_X(1) = p, \quad p_X(0) = 1 - p, \quad \mathcal{X} = \{0, 1\} \]
- For an indicator variable, *expected values are probabilities*:
  \[ E[X] = p \]
- Variance of Bernoulli distribution:
  \[ \text{Var}[X] = E \left[ (X - p)^2 \right] = p(1 - p) \]
  \[ E[X^2] = p \]
- Fair coin ($p=0.5$) has largest variance
- Coins that always come up heads ($p=1.0$), or always come up tails ($p=0.0$), have variance 0
Two Ways to Compute Variances

\[ m = E[X] = \sum_{x \in X} xp_X(x) \]

- The **variance** is the **expected squared deviation** from the mean:

\[ \text{Var}[X] = E[(X - m)^2] = E[X^2] - m^2 \]

- Prove equivalence via the **linearity of expectations**:

\[ E[(X - m)^2] = E[X^2] - 2E[Xm] + E[m^2] = E[X^2] - 2mE[X] + m^2 \]
Consider a linear function: \( Y = g(X) = aX + b \)

Example: Change of units (temperature, length, mass, currency, …)

In this special case, mean of \( Y \) is the linear function applied to \( E[X] \):

\[
E[Y] = g(E[X]) = aE[X] + b
\]

For linear functions, the variance of \( Y \) is proportional to the variance of \( X \):

\[
\text{Var}[Y] = a^2 \text{Var}[X]
\]

Prove via algebraic manipulation. First step:

\[
E[Y^2] = a^2 E[X^2] + 2abE[X] + b^2
\]

Example: Conversion from degrees Celsius to Fahrenheit:

\[
Y = 1.8X + 32 \quad \text{Var}[Y] = (1.8)^2 \text{Var}[X] = 3.24 \text{Var}[X]
\]
Variance of Shifted Discrete Uniform

- Example: Uniform distribution on \( \{0, 1, \ldots, n\} \)

Instead consider a uniform distribution on \( \{a, a+1, \ldots, a+n\} \):

\[
E[X] = \frac{n}{2}
\]

\[
Var[X] = E[X^2] - E[X]^2 = \frac{n(n+2)}{12}
\]

- Instead consider a uniform distribution on \( \{a, a+1, \ldots, a+n\} \):

\[
Y = X + a
\]

\[
E[Y] = \frac{n}{2} + a = \frac{a + (a + n)}{2}
\]

\[
Var[Y] = Var[X] = \frac{n(n+2)}{12}
\]