CS145: Lecture 8 Outline

- Variance of Sums of Random Variables
- Cumulative Distribution Functions
- Limits and Continuous Variables
The **variance** is the *expected squared deviation* of a random variable from its mean (the following definitions are equivalent):

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

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

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

\[
\text{Var}[X] \geq 0 \text{ because } (x - E[X])^2 \geq 0 \text{ for all } x.
\]

By definition, the **standard deviation** is the square root of the variance:

\[
\sigma_X = \text{Std}[X] = \sqrt{\text{Var}[X]}
\]
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:

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

Compute second moment via two cases:

$$E[X^2] = pE[X^2 \mid X = 1] + (1 - p)E[X^2 \mid X > 1]$$

$$E[X^2] = p + (1 - p)E[(X + 1)^2]$$

$$E[X^2] = \frac{2 - p}{p^2}$$
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 mean and variance of the geometric distribution then equal:

$$E[X^2] = \frac{2 - p}{p^2}$$

$$E[X] = \frac{1}{p}$$

$$\text{Var}[X] = \frac{1 - p}{p^2}$$
Define

\[ P_X(x) = \begin{cases} \frac{1}{2^k} & x = 2^k \text{ for } k = 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases} \]

\( P_X(k) \) defines a proper probability distribution: \( \sum_{k \geq 1} \frac{1}{2^k} = 1 \)

But

\[ E[X] = \sum_{k \geq 1} \frac{2^k}{2^k} = \infty \]

\( X \) has no \textbf{first} moment (expectation)
Define

\[ P_Y(y) = \begin{cases} 
\frac{1}{2^k} & y = 2^{k/2} \text{ for } k = 1, 2, \ldots \\
0 & \text{otherwise}
\end{cases} \]

\( P_y(k) \) defines a proper probability distribution

\[
E[Y] = \sum_{k \geq 1} \frac{2^{k/2}}{2^k} = \sum_{k \geq 1} \frac{1}{\sqrt{2}^k} = \frac{1}{\sqrt{2} - 1}
\]

\[
E[Y^2] = \sum_{k \geq 1} \frac{2^k}{2^k} = \infty
\]

Y has first moment but no second moment (and variance)
Sums of Independent Variables

- If $Z = X + Y$ and random variables $X$ and $Y$ are independent, we have

\[E[Z] = E[X] + E[Y]\]
\[\text{Var}[Z] = \text{Var}[X] + \text{Var}[Y]\]

*For any variables $X$, $Y$. Only for independent $X$, $Y$."

- Interpretation: Adding independent variables increases variance

\[\text{Var}[Z] \geq \text{Var}[X]\]
\[\text{and}\]
\[\text{Var}[Z] \geq \text{Var}[Y]\]

- The *standard deviation* of a sum of independent variables is then

\[\sigma_Z = \sqrt{\sigma_X^2 + \sigma_Y^2}\]
\[\sigma_X = \sqrt{\text{Var}[X]}, \sigma_Y = \sqrt{\text{Var}[Y]}, \sigma_Z = \sqrt{\text{Var}[Z]}\]

- Identity used in proof: If $X$ and $Y$ are *independent* random variables,

\[E[XY] = E[X]E[Y]\]

*This equality does not hold for general, dependent random variables.*
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
Suppose you flip \(n\) coins with bias \(p\), count number of heads

A **binomial** random variable \(X\) has parameters \(n, p\):

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

If \(X_i\) is a Bernoulli variable indicating whether toss \(i\) comes up heads, then

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

Then because tosses are **independent**:

\[ E[X] = np \]
\[ \text{Var}[X] = np(1 - p) \]
Variance of Sums of Random Variables
Cumulative Distribution Functions
Limits and Continuous Variables
Recall *probability mass function* (PMF):

\[ p_X(x) = P(X = x) \]

The *cumulative distribution function* (CDF) is the cumulative sum of the PMF:

\[ F_X(x) = P(X \leq x) = \sum_{k \leq x} p_X(k) \]

The CDF equals 0 below the range of X, 1 above the range of X, and is *monotonically increasing*:

\[ F_X(x_2) \geq F_X(x_1) \text{ if } x_2 > x_1. \]

The CDF allows quick computation of the probability of *intervals*:

\[ P(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1) \]
Binomial Probability Distribution

- You flip $n$ coins with bias $p$, count number heads.
- A binomial probability mass function (PMF) has parameters $n$, $p$:
  \[ p_X(k) = \binom{n}{k} p^k (1 - p)^{n-k} \]
  \[ \mathcal{X} = \{0, 1, 2, \ldots, n\} \]
- By definition, the binomial cumulative distribution function (CDF) equals
  \[ F_X(x) = \sum_{k=0}^{\lfloor x \rfloor} \binom{n}{k} p^k (1 - p)^{n-k} \]

No simple closed form, evaluate numerically.
A geometric probability mass function (PMF) has parameter $p$, countably infinite range:

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

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

The geometric cumulative distribution function (CDF) then equals:

$$ F_X(x) = 1 - P(X > x) = 1 - (1 - p)^x $$

Note that the CDF is strictly less than 1.0 for all finite $x$, because the range is unbounded above.
Quantiles of Distributions

- For $0 < p < 1$, the *p-quantile* of distribution of random variable $X$ is the *smallest* $x$ for which

$$F_X(x) \geq p$$

- The *median* is the 0.5-quantile. This is the “center” of the distribution, which sometimes (but not always) equals the mean.

- The 0.25-quantile and 0.75-quantile are sometimes called *quartiles*.

- Often we are interested in “extreme” quantiles, which give the probabilities of rare events:
  - $p = 0.9, 0.99, 0.999, \ldots$
  - $p = 0.1, 0.01, 0.001, \ldots$
CS145: Lecture 8 Outline

- Variance of Sums of Random Variables
- Cumulative Distribution Functions
- Limits and Continuous Variables
The Discrete Uniform Distribution is described by the probability mass function (PMF) and cumulative distribution function (CDF) as follows:

\[ p_X(x) = \frac{1}{n} \quad \text{if} \quad a \leq x \leq b \]

\[ n = b - a + 1 \]

\[ F_X(x) = \frac{\lfloor x \rfloor - a + 1}{n} \quad \text{if} \quad a \leq x \leq b \]
Take a discrete random variable uniformly distributed between 0 and \( n-1 \), and multiply by \( 1/n \) to get a variable taking values between 0 and 1.

What does this random variable approach as \( n \) becomes large?