# Chapter 4: Expectation and Their Distributions

Introduction to Probability, 2nd Edition
Blitzstein & Hwang

DRME

December 22, 2025

# Chapter Overview

# What is Expectation?

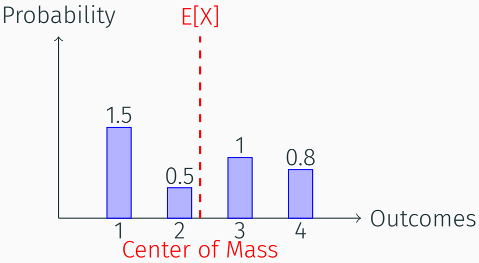
**Central Idea**

The **expectation** or **expected value** of a random variable is its long-run average value if the experiment is repeated many times.

**Analogy**: The expectation is the ”center of mass” of the probability distribution.

Wikipedia: Expected Value

# Visualizing Expectation: Center of Mass



The expectation balances the probability distribution like a seesaw.

# Definition of Expectation

# Expectation for Discrete Random Variables

**Definition (Expectation for Discrete RVs)**

For a discrete random variable $X$ with probability mass function $p_X$, the **expected value** is:

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

summed over all possible values of $X$.

**Interpretation**:

- Weighted average of all possible values
- Weights = probabilities of those values
- Only defined if the sum converges absolutely

# Discrete Expectation Example

**Example**: Fair die roll

$$E[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

**Why not 3.5?** Even though 3.5 isn’t a possible outcome, it represents the average over many rolls.

# Expectation for Continuous Random Variables

### Definition (Expectation for Continuous RVs)

For a continuous random variable $X$ with probability density function $f_X$, the **expected value** is:

$$E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx$$

provided the integral converges absolutely.

**Interpretation**:

- Continuous analog of weighted average
- Weight at each point = density $f_X(x)$

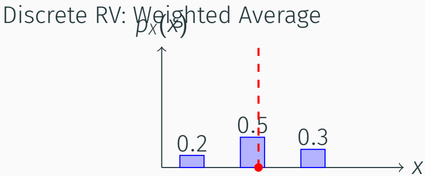
# Continuous Expectation Example

**Example**: $X \sim \text{Uniform}(a, b)$

$$E[X] = \int_a^b x \cdot \frac{1}{b-a}\, dx = \frac{1}{b-a} \cdot \frac{x^2}{2}\Big|_a^b = \frac{a+b}{2}$$

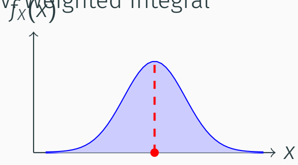
The expected value is exactly the midpoint, as we would intuitively expect.

# Visualizing Discrete Expectation



$$E[X] = 0.5 \times 0.2 + 1.5 \times 0.5 + 2.5 \times 0.3$$
$$= 0.1 + 0.75 + 0.75 = 1.6$$

# Visualizing Continuous Expectation



Continuous RV: Weighted Integral

$$E[X] = \int x f_X(x) dx$$

**Key Insight**: The expected value balances the probability mass, just like a seesaw balances weights.

# Properties of Expectation

# Linearity of Expectation

**Theorem (Linearity of Expectation)**

*For any random variables $X$ and $Y$, and constants $a$ and $b$:*

$$E[aX + bY] = aE[X] + bE[Y]$$

**Crucial Insight**: This property holds **always**, regardless of:

- Independence or dependence of $X$ and $Y$
- Discrete or continuous nature
- Any special relationship between them

# Proof Sketch of Linearity

**Proof.**

For discrete case:

$$\begin{aligned} E[aX + bY] &= \sum_x \sum_y (ax + by) p_{X,Y}(x, y) \\ &= a \sum_x x \sum_y p_{X,Y}(x, y) + b \sum_y y \sum_x p_{X,Y}(x, y) \\ &= aE[X] + bE[Y] \end{aligned}$$

□

Continuous case follows similarly using integrals instead of sums.

# Hat Check Problem: Setup

Example 1: Hat Check Problem (from Book)

- $n$ people put hats in a box
- Hats are randomly redistributed
- Let $X$ = number of people who get their own hat back
- Find $E[X]$

**Without Linearity**: Need full distribution of $X$ (complicated!)

**With Linearity**: Much simpler approach using indicator variables

# Hat Check Problem: Solution

Define indicator variables:

$$I_i = \begin{cases} 1 & \text{if person } i \text{ gets own hat} \\ 0 & \text{otherwise} \end{cases}$$

Then:

$$X = I_1 + I_2 + \cdots + I_n$$

By linearity:

$$E[X] = E[I_1] + E[I_2] + \cdots + E[I_n]$$

Since each person has probability $\frac{1}{n}$ of getting their own hat:

$$E[I_i] = 1 \cdot \frac{1}{n} + 0 \cdot \frac{n-1}{n} = \frac{1}{n}$$

Thus:

$$E[X] = n \cdot \frac{1}{n} = 1$$

**Result**: Expected number of fixed points is 1, regardless of $n$!

# Basic Properties of Expectation

1. **Expectation of constant**: $E[c] = c$ for any constant $c$
2. **Monotonicity**: If $X \leq Y$ almost surely, then $E[X] \leq E[Y]$
3. **Non-negativity**: If $X \geq 0$, then $E[X] \geq 0$
4. **Triangle inequality**: $|E[X]| \leq E[|X|]$

# Expectation of Functions and Caution

**Expectation of function**: For any function $g$:

$$E[g(X)] = \begin{cases} \sum_x g(x)p_X(x) & \text{(discrete)} \\ \int_{-\infty}^{\infty} g(x)f_X(x)dx & \text{(continuous)} \end{cases}$$

(This is called the Law of the Unconscious Statistician, LOTUS)

**Important Caution**:

$$E[XY] \neq E[X]E[Y] \quad \text{in general}$$

Equality holds only if $X$ and $Y$ are uncorrelated (in particular, if independent).

# Law of the Unconscious Statistician (LOTUS)

**Theorem (LOTUS)**

*For any function g and random variable X:*

$$E[g(X)] = \begin{cases} \sum_x g(x)p_X(x) & \textit{(discrete)} \\ \int_{-\infty}^{\infty} g(x)f_X(x)dx & \textit{(continuous)} \end{cases}$$

**Why "Unconscious Statistician"?** Because you don't need to find the distribution of $g(X)$ first!

# LOTUS Example

**Example**: If $X \sim \text{Uniform}(0, 1)$, find $E[X^2]$

**Without LOTUS**: Need to find PDF of $Y = X^2$, then compute $E[Y]$

**With LOTUS**:

$$E[X^2] = \int_0^1 x^2 \cdot 1\, dx = \frac{x^3}{3}\Big|_0^1 = \frac{1}{3}$$

**Proof Idea**:

- For discrete case, group outcomes where $g(X)$ takes same value
- For continuous, use change of variables

# Variance and Standard Deviation

# Definition of Variance

### Definition (Variance)

The **variance** of a random variable $X$ measures its spread or dispersion:

$$\mathrm{Var}(X) = E[(X-\mu)^2] \quad \text{where } \mu = E[X]$$

### Definition (Standard Deviation)

The **standard deviation** is the square root of variance:

$$SD(X) = \sqrt{\mathrm{Var}(X)}$$

Wikipedia: Variance

# Interpretation of Variance

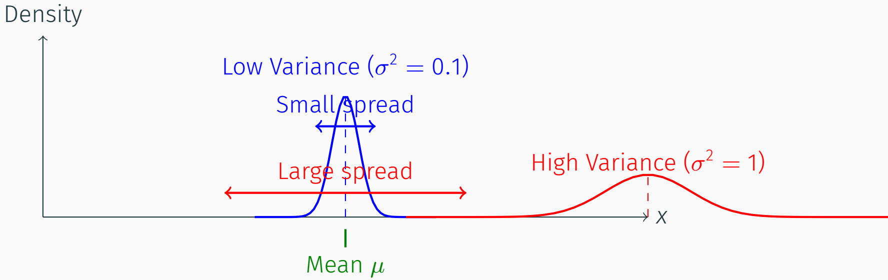
Interpretation:

- Variance = average squared distance from mean
- Standard deviation = typical distance from mean (in original units)
- Both are non-negative: $\text{Var}(X) \geq 0$, $SD(X) \geq 0$

Units:

- Variance has units squared (e.g., $\text{cm}^2$ if $X$ is in cm)
- Standard deviation has same units as $X$ (e.g., cm)

# Visualizing Variance


Density
Low Variance ($\sigma^2 = 0.1$)
Small spread
Large spread
High Variance ($\sigma^2 = 1$)
$x$
Mean $\mu$

**Key Insight**: Variance measures how ”spread out” the distribution is around the mean.

# Computational Formula for Variance

## Theorem (Alternative Variance Formula)

$$Var(X) = E[X^2] - (E[X])^2$$

Proof:

$$\begin{aligned}
\mathrm{Var}(X) &= E[(X-\mu)^2] \\
&= E[X^2 - 2\mu X + \mu^2] \\
&= E[X^2] - 2\mu E[X] + \mu^2 \quad \text{(by linearity)} \\
&= E[X^2] - 2\mu \cdot \mu + \mu^2 \\
&= E[X^2] - \mu^2
\end{aligned}$$

# Why the Computational Formula is Useful

**Why Useful?** Often easier to compute $E[X^2]$ and $(E[X])^2$ separately.

**Example**: For $X \sim \text{Bernoulli}(p)$:

$$\begin{aligned} E[X] &= 0 \cdot (1-p) + 1 \cdot p = p \\ E[X^2] &= 0^2 \cdot (1-p) + 1^2 \cdot p = p \\ \text{Var}(X) &= p - p^2 = p(1-p) \end{aligned}$$

# Basic Properties of Variance

1. **Non-negativity**: $\mathrm{Var}(X) \geq 0$, with equality iff $X$ is constant
2. **Scaling**: $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$
   - Adding constant doesn't change spread
   - Multiplying by constant scales variance by square of constant

# Variance of Sums

**Variance of sum**: For any $X$ and $Y$:

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$$

**Independent case**: If $X$ and $Y$ are independent:

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

**Variance of sample mean**: If $X_1, \ldots, X_n$ are i.i.d. with variance $\sigma^2$:

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{\sigma^2}{n}$$

# Example: Variance of Binomial Distribution

**Problem**: Find variance of $X \sim \text{Bin}(n, p)$

**Step 1: Represent as sum of indicators**

$$X = I_1 + I_2 + \cdots + I_n, \quad I_i \sim \text{Bernoulli}(p) \text{ i.i.d.}$$

# Example: Variance of Binomial (continued)

Step 2: Use variance properties

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{Var}(I_1 + I_2 + \cdots + I_n) \\ &= \mathrm{Var}(I_1) + \mathrm{Var}(I_2) + \cdots + \mathrm{Var}(I_n) \quad \text{(independence)} \\ &= n \cdot \mathrm{Var}(I_1) \end{aligned}$$

Step 3: Compute variance of Bernoulli

$$\mathrm{Var}(I_1) = p(1-p)$$

Step 4: Final answer

$$\mathrm{Var}(X) = np(1-p)$$

# Covariance and Correlation

# Definition of Covariance

### Definition (Covariance)

The **covariance** between two random variables $X$ and $Y$ measures their linear relationship:

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - E[X]E[Y]$$

where $\mu_X = E[X]$, $\mu_Y = E[Y]$.

Wikipedia: Covariance

# Interpretation of Covariance

**Interpretation**:

- $\text{Cov}(X, Y) > 0$: $X$ and $Y$ tend to be above/below their means together
- $\text{Cov}(X, Y) < 0$: When $X$ is above mean, $Y$ tends to be below mean
- $\text{Cov}(X, Y) = 0$: Uncorrelated (but not necessarily independent!)

**Key Insight**: Covariance measures **linear** relationship.

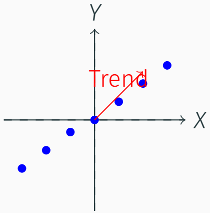
# Properties of Covariance

Properties:

- $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
- $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
- $\mathrm{Cov}(aX + b, cY + d) = ac\,\mathrm{Cov}(X, Y)$
- Bilinearity: $\mathrm{Cov}(X + Y, Z) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)$

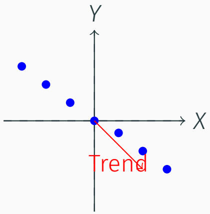
# Visualizing Positive Covariance

Positive Covariance



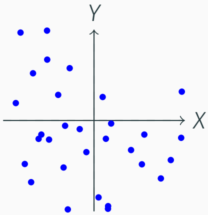
# Visualizing Negative Covariance

Negative Covariance



# Visualizing Zero Covariance

Zero Covariance (Uncorrelated)



# Example: Computing Covariance - Setup

**Problem**: Roll two fair dice. Let $X$ = number on first die, $Y$ = sum of both dice. Find Cov($X$, $Y$).

**Step 1: Compute $E[X]$ and $E[Y]$**

$$E[X] = 3.5$$
$$E[Y] = E[X_1 + X_2] = E[X_1] + E[X_2] = 3.5 + 3.5 = 7$$

# Example: Computing Covariance - Calculations

Step 2: Compute $E[XY]$

$$E[XY] = E[X(X_1 + X_2)] = E[X^2 + XX_2] = E[X^2] + E[X]E[X_2]$$

since $X$ and $X_2$ are independent.

Step 3: Compute $E[X^2]$

$$E[X^2] = \frac{1}{6}(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) = \frac{91}{6}$$

# Example: Computing Covariance - Final Answer

Step 4: Put it all together

$$E[XY] = \frac{91}{6} + 3.5 \times 3.5 = \frac{91}{6} + \frac{49}{4} = \frac{329}{12}$$

$$\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y] = \frac{329}{12} - 3.5 \times 7 = \frac{35}{12} \approx 2.92$$

Positive covariance makes sense: higher first die tends to give higher sum.

# Correlation Coefficient

### Definition (Correlation)

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

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Wikipedia: Correlation

# Properties of Correlation

Properties:

1. $-1 \leq \rho(X, Y) \leq 1$
2. $\rho(X, Y) = 1$ iff $Y = aX + b$ with $a > 0$ (perfect positive linear relationship)
3. $\rho(X, Y) = -1$ iff $Y = aX + b$ with $a < 0$ (perfect negative linear relationship)
4. $\rho(X, Y) = 0$: Uncorrelated (no linear relationship)
5. Invariant to scaling: $\rho(aX + b, cY + d) = \text{sign}(ac) \cdot \rho(X, Y)$

# Why Use Correlation Instead of Covariance?

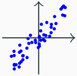
Why use correlation instead of covariance?

- Covariance depends on units of measurement
- Correlation is dimensionless and bounded between -1 and 1
- Easier to interpret strength of relationship

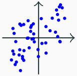
Interpretation Guide:

- $|\rho| \approx 0$: Weak linear relationship
- $|\rho| \approx 0.5$: Moderate linear relationship
- $|\rho| \approx 0.8$: Strong linear relationship
- $|\rho| \approx 1$: Very strong linear relationship

# Visualizing Different Correlation Values

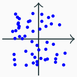


$0.9 = 0.9$ (Strong positive)

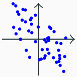


$0.5 = 0.5$ (Moderate positive)

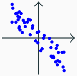
# Visualizing Different Correlation Values (Continued)



$0 = 0$ (No linear relationship)



$-0.5 = -0.5$ (Moderate negative)



$-0.9 = -0.9$ (Strong negative)

# Important Notes about Correlation

Important Notes:

- $\rho = 0$ doesn't imply independence (only no linear relationship)
- $\rho$ measures only linear relationships (nonlinear relationships can have $\rho = 0$)
- Independence $\Rightarrow \rho = 0$, but $\rho = 0 \not\Rightarrow$ independence

Example of nonlinear relationship with $\rho = 0$: Let $X \sim \text{Uniform}(-1, 1)$ and $Y = X^2$. Then $\text{Cov}(X, Y) = E[X^3] - E[X]E[X^2] = 0 - 0 \cdot E[X^2] = 0$, but $X$ and $Y$ are clearly dependent.

# Conditional Expectation

# Definition of Conditional Expectation

### Definition (Conditional Expectation for Discrete RVs)

For discrete random variables $X$ and $Y$, the **conditional expectation** of $Y$ given $X = x$ is:

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

**Interpretation**: The average value of $Y$ when we know $X = x$.

# Conditional Expectation for Continuous RVs

### Definition (Conditional Expectation for Continuous RVs)

For continuous random variables $X$ and $Y$, the **conditional expectation** of $Y$ given $X = x$ is:

$$E[Y \mid X = x] = \int_{-\infty}^{\infty} y \cdot f_{Y|X}(y \mid x)\, dy$$

where $f_{Y|X}(y \mid x)$ is the conditional density.

**Note**: $E[Y \mid X]$ is itself a random variable (function of $X$).

# Basic Properties of Conditional Expectation

1. **Linearity**: $E[aY + bZ \mid X = x] = aE[Y \mid X = x] + bE[Z \mid X = x]$
2. **Taking out what's known**: If $g(X)$ is a function of $X$ only:

$$E[g(X)Y \mid X = x] = g(x)E[Y \mid X = x]$$

3. **Independence**: If $X$ and $Y$ are independent:

$$E[Y \mid X = x] = E[Y] \quad \text{for all } x$$

# Important Properties of Conditional Expectation

4. **Law of Total Expectation (Tower Property)**:

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

5. **Best predictor**: $E[Y \mid X]$ is the function of $X$ that minimizes $E[(Y - g(X))^2]$ over all functions $g$

# Example: Conditional Expectation in Two Dice Problem

**Problem**: Roll two fair dice. Let $X$ = number on first die, $Y$ = sum of both dice. Find $E[Y \mid X = 3]$.

**Step 1: Identify conditional distribution** Given $X = 3$, $Y = 3 + X_2$ where $X_2$ is second die.

**Step 2: Compute conditional expectation**

$$E[Y \mid X = 3] = E[3 + X_2] = 3 + E[X_2] = 3 + 3.5 = 6.5$$

# Example: General Conditional Expectation

**Step 3: Generalize** For any $x$:

$$E[Y \mid X = x] = E[x + X_2] = x + 3.5$$

**Step 4: Verify Law of Total Expectation**

$$\begin{aligned} E[Y] &= E[E[Y \mid X]] = E[X + 3.5] = E[X] + 3.5 = 3.5 + 3.5 = 7 \\ &= \text{Direct computation: } E[Y] = 7 \quad \checkmark \end{aligned}$$

# Law of Total Expectation (Tower Property)

**Theorem (Law of Total Expectation)**

*For any random variables X and Y:*

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

# Proof of Law of Total Expectation

**Proof (discrete case).**

$$\begin{aligned} E[E[Y \mid X]] &= \sum_x E[Y \mid X = x] P(X = x) \\ &= \sum_x \left( \sum_y y P(Y = y \mid X = x) \right) P(X = x) \\ &= \sum_x \sum_y y P(Y = y, X = x) \\ &= \sum_y y \sum_x P(Y = y, X = x) \\ &= \sum_y y P(Y = y) = E[Y] \end{aligned}$$

□

# Interpretation of Law of Total Expectation

Interpretation: To compute overall average of $Y$, average conditional averages weighted by probability of conditioning events.

Analogy: To find average grade in a class:

- Compute average grade for each section
- Weight each section average by number of students in that section
- Sum the weighted averages

# Example: Law of Total Expectation Application - Setup

**Problem (from Book)**: Suppose we have a stick of length 1. Break it at a random point $X \sim \text{Uniform}(0, 1)$. Then break the longer piece at a random point. What's the expected length of the final longest piece?

**Step 1: Define variables** Let $X$ = first break point, $Y$ = length of final longest piece.

**Step 2: Use law of total expectation**

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

# Example: Law of Total Expectation Application - Conditional

Step 3: Compute $E[Y \mid X = x]$

If $x \geq 1/2$, left piece is longer. Break it at point $U \sim \text{Uniform}(0, x)$. Longest piece length = $\max(U, x - U)$. By symmetry, $E[\max(U, x - U) \mid X = x] = \frac{3x}{4}$.

If $x < 1/2$, right piece is longer. Similar argument gives $E[Y \mid X = x] = \frac{3(1-x)}{4}$.

# Example: Law of Total Expectation Application - Final

Step 4: Compute overall expectation

$$\begin{aligned}
E[Y] &= \int_0^{1/2} \frac{3(1-x)}{4} dx + \int_{1/2}^{1} \frac{3x}{4} dx \\
&= \frac{3}{4}\left[\int_0^{1/2} (1-x) dx + \int_{1/2}^{1} x dx\right] \\
&= \frac{3}{4}\left[\left(\frac{1}{2} - \frac{1}{8}\right) + \left(\frac{1}{2} - \frac{1}{8}\right)\right] \\
&= \frac{3}{4} \times \frac{3}{4} = \frac{9}{16}
\end{aligned}$$

# Moment Generating Functions

# Definition of Moment Generating Function

### Definition (Moment Generating Function)

The **moment generating function (MGF)** of a random variable $X$ is:

$$M_X(t) = E[e^{tX}]$$

defined for all $t$ where the expectation exists.

Wikipedia: Moment Generating Function

# Why ”Moment Generating”?

The MGF generates moments because:

$$M_X^{(n)}(0) = \left.\frac{d^n}{dt^n} M_X(t)\right|_{t=0} = E[X^n]$$

The $n$-th derivative at 0 gives the $n$-th moment.

**Example**: For $X \sim \text{Exponential}(\lambda)$:

$$M_X(t) = E[e^{tX}] = \int_0^\infty e^{tx} \lambda e^{-\lambda x} dx = \frac{\lambda}{\lambda - t}, \quad t < \lambda$$

Then $E[X] = M_X'(0) = \frac{1}{\lambda}$, $E[X^2] = M_X''(0) = \frac{2}{\lambda^2}$.

# Basic Properties of Moment Generating Functions

1. **Uniqueness**: If $M_X(t) = M_Y(t)$ for all $t$ in neighborhood of 0, then $X$ and $Y$ have the same distribution.
2. **Linear transformation**: For $Y = aX + b$:

$$M_Y(t) = e^{bt}M_X(at)$$

3. **Sum of independent RVs**: If $X$ and $Y$ are independent:

$$M_{X+Y}(t) = M_X(t)M_Y(t)$$

# More Properties of Moment Generating Functions

4. **Moments from MGF**:

$$E[X^n] = M_X^{(n)}(0)$$

5. **Relationship to other functions**:
   - Characteristic function: $\phi_X(t) = E[e^{itX}]$
   - Probability generating function (for discrete): $G_X(z) = E[z^X]$

# Example: MGF of Normal Distribution - Setup

**Problem**: Find MGF of $X \sim N(\mu, \sigma^2)$.

**Step 1: Standardize** Let $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$. Then $X = \mu + \sigma Z$.

# Example: MGF of Standard Normal

Step 2: Find MGF of standard normal

$$
\begin{aligned}
M_Z(t) = E[e^{tZ}] &= \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(z^2 - 2tz)\right) dz \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}[(z-t)^2 - t^2]\right) dz \\
&= e^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(z-t)^2/2} dz \\
&= e^{t^2/2} \quad \text{(integral of normal PDF = 1)}
\end{aligned}
$$

# Example: MGF of Normal Distribution - Final

Step 3: Transform back

$$M_X(t) = E[e^{t(\mu+\sigma Z)}] = e^{\mu t} M_Z(\sigma t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$$

**Special case**: For standard normal ($\mu = 0, \sigma^2 = 1$):

$$M_Z(t) = e^{t^2/2}$$

# Application 1: Sum of Independent Normals

**Application**: Sum of Independent Normals

If $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, independent, then:

$$\begin{aligned} M_{X+Y}(t) &= M_X(t)M_Y(t) \\ &= \exp\left(\mu_1 t + \frac{\sigma_1^2 t^2}{2}\right) \exp\left(\mu_2 t + \frac{\sigma_2^2 t^2}{2}\right) \\ &= \exp\left((\mu_1 + \mu_2)t + \frac{(\sigma_1^2 + \sigma_2^2)t^2}{2}\right) \end{aligned}$$

Thus $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

# Application 2: Method of Moments Estimation

**Application**: Method of Moments Estimation

Given i.i.d. sample $X_1, \ldots, X_n$, method of moments:

1. Compute sample moments: $m_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$
2. Express population moments in terms of parameters: $\mu_k(\theta) = E[X^k]$
3. Solve $\mu_k(\theta) = m_k$ for parameters $\theta$

**Example**: For $X \sim$ Exponential$(\lambda)$, $E[X] = 1/\lambda$. Method of moments estimate: $\hat{\lambda} = 1/\bar{X}$.

# Important Inequalities

# Markov's Inequality

**Theorem (Markov's Inequality)**

*For any nonnegative random variable $X$ and any $a > 0$:*

$$P(X \geq a) \leq \frac{E[X]}{a}$$

Wikipedia: Markov's Inequality

# Proof of Markov's Inequality

**Proof.**

$$
\begin{aligned}
E[X] &= \int_0^\infty x f_X(x) dx \\
&\geq \int_a^\infty x f_X(x) dx \quad \text{(integral over smaller domain)} \\
&\geq \int_a^\infty a f_X(x) dx \quad \text{(since } x \geq a \text{ in this region)} \\
&= aP(X \geq a)
\end{aligned}
$$

□

# Interpretation of Markov's Inequality

Interpretation: Probability of large values is controlled by mean.

Example: If average income is \$50,000, at most 10% can have income ≥ \$500,000.

Limitation: Very conservative bound, often not tight.

# Chebyshev’s Inequality

**Theorem (Chebyshev’s Inequality)**

*For any random variable $X$ with finite mean $\mu$ and variance $\sigma^2$, and any $k > 0$:*

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Wikipedia: Chebyshev’s Inequality

# Proof of Chebyshev's Inequality

**Proof.**

Apply Markov's inequality to $(X-\mu)^2$:

$$P(|X-\mu| \geq k\sigma) = P((X-\mu)^2 \geq k^2\sigma^2) \leq \frac{E[(X-\mu)^2]}{k^2\sigma^2} = \frac{1}{k^2}$$

□

# Interpretation of Chebyshev’s Inequality

Interpretation:

- For $k = 2$: At most 25% of probability is more than 2 SDs from mean
- For $k = 3$: At most 11% of probability is more than 3 SDs from mean
- For $k = 10$: At most 1% of probability is more than 10 SDs from mean

Example: If test scores have mean 70, SD 10, at most 25% scored below 50 or above 90.

# Cauchy-Schwarz Inequality for Expectations

**Theorem (Cauchy-Schwarz Inequality)**

*For any random variables $X$ and $Y$ with finite second moments:*

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

*Equality holds iff $Y = aX$ almost surely for some constant $a$.*

# Proof of Cauchy-Schwarz Inequality

**Proof.**

Consider $E[(tX + Y)^2] \geq 0$ for all $t$:

$$E[X^2]t^2 + 2E[XY]t + E[Y^2] \geq 0$$

This quadratic in $t$ has at most one real root, so discriminant ≤ 0:

$$(2E[XY])^2 - 4E[X^2]E[Y^2] \leq 0$$

which gives the inequality. □

# Application of Cauchy-Schwarz Inequality

**Application**: Shows $|\rho(X, Y)| \leq 1$:

$$|\mathrm{Cov}(X, Y)| = |E[(X - \mu_X)(Y - \mu_Y)]| \leq \sqrt{E[(X - \mu_X)^2]E[(Y - \mu_Y)^2]} = \sigma_X \sigma_Y$$

Thus:

$$|\rho(X, Y)| = \frac{|\mathrm{Cov}(X, Y)|}{\sigma_X \sigma_Y} \leq 1$$

# Jensen’s Inequality

### Theorem (Jensen’s Inequality)

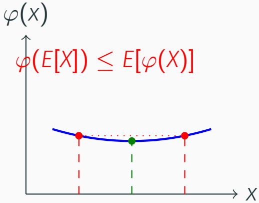
*If $\varphi$ is a convex function and $X$ is a random variable, then:*

$$\varphi(E[X]) \leq E[\varphi(X)]$$

*If $\varphi$ is concave, the inequality reverses.*

Wikipedia: Jensen’s Inequality

# Visualizing Jensen's Inequality



# Examples of Jensen's Inequality

Examples:

- $\varphi(x) = x^2$ convex: $E[X]^2 \leq E[X^2]$ (already knew from variance ≥ 0)
- $\varphi(x) = e^x$ convex: $e^{E[X]} \leq E[e^X]$
- $\varphi(x) = \log(x)$ concave: $E[\log(X)] \leq \log(E[X])$

Application: In information theory, concavity of log gives:

$$E[\log(X)] \leq \log(E[X])$$

which is used in proving properties of entropy.

# Detailed Examples

# Example 1: St. Petersburg Paradox - Setup

**Problem**: A casino offers a game. Flip a fair coin until it lands heads. If first heads occurs on $n$-th toss, you win $2^n$ dollars. How much would you pay to play?

**Step 1: Define random variable** Let $X$ = payout. $P(\text{heads on toss } n) = (1/2)^n$.

# Example 1: St. Petersburg Paradox - Calculation

Step 2: Compute expected value

$$E[X] = \sum_{n=1}^{\infty} 2^n \cdot \left(\frac{1}{2}\right)^n = \sum_{n=1}^{\infty} 1 = \infty$$

Step 3: Paradox Expected value is infinite! But would you pay $1,000 to play? Probably not.

# Example 1: St. Petersburg Paradox - Resolution

### Step 4: Resolution

- Utility theory: Money has diminishing marginal utility
- Use log utility: $E[\log(X)] = \sum_{n=1}^{\infty} \log(2^n) \cdot (1/2)^n = \log(4)$
- Casino has finite wealth
- People are risk-averse

# Example 2: Coupon Collector Problem - Setup

**Problem**: There are $n$ different coupons. Each box contains one coupon, uniformly random. How many boxes to collect all coupons?

**Step 1: Define variables** Let $T$ = total boxes needed. Let $T_i$ = boxes to get $i$-th new coupon after having $i - 1$.

# Example 2: Coupon Collector Problem - Analysis

**Step 2: Analyze** $T_i$ After $i-1$ coupons, probability new coupon in next box = $\frac{n-(i-1)}{n}$. So $T_i \sim \text{Geometric}(p_i)$ with $p_i = \frac{n-i+1}{n}$. Thus $E[T_i] = \frac{1}{p_i} = \frac{n}{n-i+1}$.

# Example 2: Coupon Collector Problem - Solution

Step 3: Use linearity

$$
\begin{aligned}
E[T] = E[T_1 + T_2 + \cdots + T_n] = \sum_{i=1}^{n} E[T_i] = \sum_{i=1}^{n} \frac{n}{n-i+1} \\
= n \sum_{j=1}^{n} \frac{1}{j} = nH_n \approx n(\log n + \gamma)
\end{aligned}
$$

where $H_n$ is $n$-th harmonic number, $\gamma \approx 0.577$ is Euler-Mascheroni constant.

**Result**: Need about $n \log n$ boxes on average.

# Example 3: Random Walk Expectation - Setup

**Problem**: Start at 0. Each step, move +1 with probability $p$, -1 with probability $q = 1 - p$. After $n$ steps, what's expected position?

**Step 1: Define variables** Let $X_i$ = step $i$: $X_i = \begin{cases} +1 & \text{prob } p \\ -1 & \text{prob } q \end{cases}$ Position after $n$ steps: $S_n = X_1 + X_2 + \cdots + X_n$.

# Example 3: Random Walk Expectation - Solution

Step 2: Compute expectation

$$E[X_i] = 1 \cdot p + (-1) \cdot q = p - q$$

By linearity:

$$E[S_n] = \sum_{i=1}^{n} E[X_i] = n(p - q)$$

**Note**: This uses linearity despite $X_i$ not being independent!

# Example 3: Random Walk Expectation - Special Cases

Step 3: Special cases

- Fair coin ($p = q = 1/2$): $E[S_n] = 0$
- Biased ($p = 0.6$): $E[S_n] = n(0.6 - 0.4) = 0.2n$
- As $n \to \infty$ with $p > 1/2$: $E[S_n] \to \infty$

# Example 4: Waiting for Patterns - Setup

**Problem**: Flip fair coin until pattern HTH appears. What's expected number of flips?

**Step 1: Define states** Let $E$ = expected flips from start. Let:

- $E_H$ = expected flips given we just saw H
- $E_{HT}$ = expected flips given we just saw HT

# Example 4: Waiting for Patterns - Equations

**Step 2: Set up equations** From start:

$$E = 1 + \frac{1}{2}E + \frac{1}{2}E_H$$

From state H:

$$E_H = 1 + \frac{1}{2}E_{HT} + \frac{1}{2}E_H \quad \text{(if T, go to HT; if H, stay at H)}$$

From state HT:

$$E_{HT} = 1 + \frac{1}{2} \cdot 0 + \frac{1}{2}E_H \quad \text{(if H, done; if T, go to H)}$$

# Example 4: Waiting for Patterns - Solution

**Step 3: Solve system** Solving: $E = 10$, $E_H = 8$, $E_{HT} = 6$.

**Result**: Expect 10 flips to see HTH.

**Note**: Different patterns have different expected waiting times!

# Summary and Key Formulas

# Key Formulas 1: Expectation and Variance

1. **Expectation**:

$$E[X] = \begin{cases} \sum_x x p_X(x) & \text{(discrete)} \\ \int_{-\infty}^{\infty} x f_X(x) dx & \text{(continuous)} \end{cases}$$

2. **Linearity**: $E[aX + bY] = aE[X] + bE[Y]$
3. **LOTUS**: $E[g(X)] = \sum g(x) p_X(x)$ or $\int g(x) f_X(x) dx$

# Key Formulas 2: Variance and Covariance

4. **Variance**: $\mathrm{Var}(X) = E[(X-\mu)^2] = E[X^2] - (E[X])^2$
5. **Variance properties**: $\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$
6. **Standard deviation**: $SD(X) = \sqrt{\mathrm{Var}(X)}$
7. **Covariance**: $\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y]$
8. **Correlation**: $\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{SD(X)SD(Y)}$

# Key Formulas 3: More Variance and Sums

9. **Variance of sum**: $\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)+2\mathrm{Cov}(X,Y)$
10. **Independent sum**: If $X, Y$ independent: $\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)$

# Key Formulas 4: Conditional Expectation

1. **Conditional expectation**:

$$E[Y \mid X = x] = \begin{cases} \sum_y y P(Y = y \mid X = x) & \text{(discrete)} \\ \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x) dy & \text{(continuous)} \end{cases}$$

2. **Law of Total Expectation**: $E[Y] = E[E[Y \mid X]]$

# Key Formulas 5: Moment Generating Functions

3. **MGF**: $M_X(t) = E[e^{tX}]$
4. **Moments from MGF**: $E[X^n] = M_X^{(n)}(0)$
5. **MGF of sum**: If $X, Y$ independent, $M_{X+Y}(t) = M_X(t)M_Y(t)$

# Key Formulas 6: Important Inequalities

1. **Markov's inequality**: $P(X \geq a) \leq \frac{E[X]}{a}$ for $X \geq 0$
2. **Chebyshev's inequality**: $P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$
3. **Cauchy-Schwarz**: $|E[XY]| \leq \sqrt{E[X^2]E[Y^2]}$
4. **Jensen's inequality**: $\varphi(E[X]) \leq E[\varphi(X)]$ for convex $\varphi$

# Common Expectations and Variances

| Distribution | PMF/PDF | $E[X]$ | $\mathrm{Var}(X)$ |
|---|---|---|---|
| Bernoulli($p$) | $p^x(1-p)^{1-x}$ | $p$ | $p(1-p)$ |
| Binomial($n, p$) | $\binom{n}{k}p^k(1-p)^{n-k}$ | $np$ | $np(1-p)$ |
| Poisson($\lambda$) | $\frac{e^{-\lambda}\lambda^k}{k!}$ | $\lambda$ | $\lambda$ |
| Geometric($p$) | $(1-p)^{k-1}p$ | $\frac{1}{p}$ | $\frac{1-p}{p^2}$ |
| Uniform($a, b$) | $\frac{1}{b-a}$ | $\frac{a+b}{2}$ | $\frac{(b-a)^2}{12}$ |
| Exponential($\lambda$) | $\lambda e^{-\lambda x}$ | $\frac{1}{\lambda}$ | $\frac{1}{\lambda^2}$ |
| Normal($\mu, \sigma^2$) | $\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ | $\mu$ | $\sigma^2$ |

# Problem-Solving Strategy 1: Finding $E[X]$

When asked to find $E[X]$:

1. Check if $X$ is sum of simpler RVs → use linearity
2. Check if can use LOTUS → compute $E[g(Y)]$ without finding distribution of $g(Y)$
3. Check if can use law of total expectation → condition on appropriate variable
4. Check if known distribution → use known formula

# Problem-Solving Strategy 2: Finding Var($X$)

When asked to find Var($X$):

1. Use formula $\mathrm{Var}(X) = E[X^2] - (E[X])^2$
2. If $X$ is sum, check if independent → variance adds
3. If $X$ is transformation of known RV, use properties

# Problem-Solving Strategy 3: Dependence and Inequalities

## When dealing with dependence:

1. Use covariance/correlation to quantify dependence
2. Remember: independence ⍰ uncorrelated, but converse false
3. For variance of sum, always include covariance term

## For inequalities:

1. Markov: for nonnegative RVs, bounds tail probability
2. Chebyshev: for any RV with finite variance, bounds deviation from mean
3. Cauchy-Schwarz: bounds covariance/correlation
4. Jensen: relates $E[g(X)]$ and $g(E[X])$ for convex/concave $g$

# End of Chapter 4

Thank You!

Complete problem sets and solutions available at http://stat110.net

Next: Chapter 5 - Limit Theorems