4: Expectation

Naijia Liu

Spring 2025

Road map

- We've defined random variables and their distributions.
- Distributions give full information about the probabilities of an r.v.
- Today: begin to summarize distributions with expectation.

Motivation: Causal Effects

- Consider a hypothetical intervention such as "door-to-door get out the vote."

Motivation: Causal Effects

- Consider a hypothetical intervention such as "door-to-door get out the vote."
- We'll define two **potential outcomes**:

Motivation: Causal Effects

- Consider a hypothetical intervention such as "door-to-door get out the vote."
- We'll define two **potential outcomes**:
  - $Y_i(1)$: whether person i would vote (1) or not (0) if they **received** canvassing.

Motivation: Causal Effects

- Consider a hypothetical intervention such as "door-to-door get out the vote."
- We'll define two **potential outcomes**:
  - $Y_i(1)$: whether person i would vote (1) or not (0) if they **received** canvassing.
  - $Y_i(0)$: whether person i would vote (1) or not (0) if they **didn't receive** the canvassing.

Motivation: Causal Effects

- Consider a hypothetical intervention such as "door-to-door get out the vote."
- We'll define two **potential outcomes**:
  - $Y_i(1)$: whether person i would vote (1) or not (0) if they **received** canvassing.
  - $Y_i(0)$: whether person i would vote (1) or not (0) if they **didn't receive** the canvassing.
- The individual causal effect of canvassing then is

$$\tau_i = Y_i(1) - Y_i(0)$$

Motivation: Causal Effects

- Consider a hypothetical intervention such as "door-to-door get out the vote."
- We'll define two **potential outcomes**:
  - $Y_i(1)$: whether person i would vote (1) or not (0) if they **received** canvassing.
  - $Y_i(0)$: whether person i would vote (1) or not (0) if they **didn't receive** the canvassing.
- The individual causal effect of canvassing then is

$$\tau_i = Y_i(1) - Y_i(0)$$

- We can think of $Y_i(1)$ and $Y_i(0)$ as rvs and so τ_i is a rv as well.

Motivation: Causal Effects

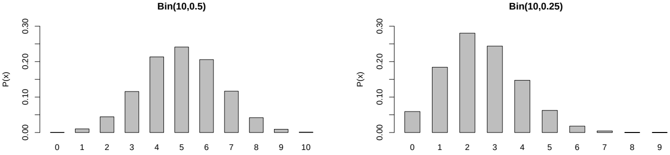
- Consider a hypothetical intervention such as "door-to-door get out the vote."
- We'll define two **potential outcomes**:
  - $Y_i(1)$: whether person i would vote (1) or not (0) if they **received** canvassing.
  - $Y_i(0)$: whether person i would vote (1) or not (0) if they **didn't receive** the canvassing.
- The individual causal effect of canvassing then is

$$\tau_i = Y_i(1) - Y_i(0)$$

- We can think of $Y_i(1)$ and $Y_i(0)$ as rvs and so τ_i is a rv as well.
- How should we summarize the distribution of causal effects?

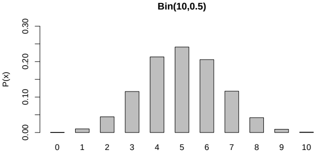
How can we summarize distributions?

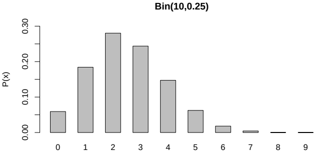
- Probability distributions describe the uncertainty about r.v.s.



How can we summarize distributions?

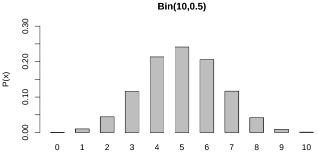
- Probability distributions describe the uncertainty about r.v.s.
- Can we summarize probability distributions?

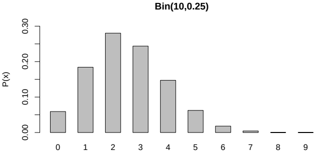




How can we summarize distributions?

- Probability distributions describe the uncertainty about r.v.s.
- Can we summarize probability distributions?
- **Question:** What is the difference between these two p.m.f.s? How might we summarize this difference?





Goals for summarizing

1. **Central tendency:** where the center of the distribution is.

Goals for summarizing

1. **Central tendency:** where the center of the distribution is.
    - We'll focus on the mean/expectation.

Goals for summarizing

1. **Central tendency:** where the center of the distribution is.
   - We'll focus on the mean/expectation.
2. **Spread:** how spread out the distribution is around the center.

Goals for summarizing

1. **Central tendency:** where the center of the distribution is.
   - We'll focus on the mean/expectation.
2. **Spread:** how spread out the distribution is around the center.
   - We'll focus on the variance/standard deviation.

Goals for summarizing

1. **Central tendency:** where the center of the distribution is.
   - We'll focus on the mean/expectation.
2. **Spread:** how spread out the distribution is around the center.
   - We'll focus on the variance/standard deviation.

- These are **population parameters** so we don't get to observe them.

Goals for summarizing

1. **Central tendency:** where the center of the distribution is.
   - We'll focus on the mean/expectation.
2. **Spread:** how spread out the distribution is around the center.
   - We'll focus on the variance/standard deviation.

- These are **population parameters** so we don't get to observe them.
  - We won't get to observe them...

Goals for summarizing

1. **Central tendency:** where the center of the distribution is.
   - We'll focus on the mean/expectation.
2. **Spread:** how spread out the distribution is around the center.
   - We'll focus on the variance/standard deviation.

- These are **population parameters** so we don't get to observe them.
  - We won't get to observe them...
  - but we'll use our sample to learn about them.

Two ways to calculate averages

- Calculate the average of: $\{1, 1, 1, 3, 4, 4, 5, 5\}$

$$\frac{1+1+1+3+4+4+5+5}{8} = 3$$

Two ways to calculate averages

- Calculate the average of: $\{1, 1, 1, 3, 4, 4, 5, 5\}$

$$\frac{1+1+1+3+4+4+5+5}{8} = 3$$

- Alternative way to calculate average based on **frequency weights**:

$$1 \times \frac{3}{8} + 3 \times \frac{1}{8} + 4 \times \frac{2}{8} + 5 \times \frac{2}{8} = 3$$

Two ways to calculate averages

- Calculate the average of: $\{1, 1, 1, 3, 4, 4, 5, 5\}$

$$\frac{1+1+1+3+4+4+5+5}{8} = 3$$

- Alternative way to calculate average based on **frequency weights**:

$$1 \times \frac{3}{8} + 3 \times \frac{1}{8} + 4 \times \frac{2}{8} + 5 \times \frac{2}{8} = 3$$

  - Each value times how often that value occurs in the data.

Two ways to calculate averages

- Calculate the average of: $\{1, 1, 1, 3, 4, 4, 5, 5\}$

$$\frac{1 + 1 + 1 + 3 + 4 + 4 + 5 + 5}{8} = 3$$

- Alternative way to calculate average based on **frequency weights**:

$$1 \times \frac{3}{8} + 3 \times \frac{1}{8} + 4 \times \frac{2}{8} + 5 \times \frac{2}{8} = 3$$

  - Each value times how often that value occurs in the data.
  - We'll use this intuition to create an average/mean for r.v.s.

Expectation

Definition

The **expected value** (or **expectation** or **mean**) of a discrete r.v. X with possible values, $x_1, x_2, \ldots$ is

$$\mathbb{E}[X] = \sum_{j=1}^{\infty} x_j \mathbb{P}(X = x_j)$$

- Weighted average of the values of the r.v. weighted by the probability of each value occurring.

Expectation

Definition

The **expected value** (or **expectation** or **mean**) of a discrete r.v. X with possible values, $x_1, x_2, \ldots$ is

$$\mathbb{E}[X] = \sum_{j=1}^{\infty} x_j \mathbb{P}(X = x_j)$$

- Weighted average of the values of the r.v. weighted by the probability of each value occurring.
  - $\mathbb{E}[X]$ is a constant!

Expectation

Definition

The **expected value** (or **expectation** or **mean**) of a discrete r.v. X with possible values, $x_1, x_2, \ldots$ is

$$\mathbb{E}[X] = \sum_{j=1}^{\infty} x_j \mathbb{P}(X = x_j)$$

- Weighted average of the values of the r.v. weighted by the probability of each value occurring.
  - $\mathbb{E}[X]$ is a constant!
- Example: $X \sim \text{Bern}(p)$, then $\mathbb{E}[X] = 1p + 0(1 - p) = p$.

Expectation

Definition

The **expected value** (or **expectation** or **mean**) of a discrete r.v. X with possible values, $x_1, x_2, \dots$ is

$$\mathbb{E}[X] = \sum_{j=1}^{\infty} x_j \mathbb{P}(X = x_j)$$

- Weighted average of the values of the r.v. weighted by the probability of each value occurring.
  - $\mathbb{E}[X]$ is a constant!
- Example: $X \sim \text{Bern}(p)$, then $\mathbb{E}[X] = 1p + 0(1-p) = p$.
- If X and Y have the same distribution, then $\mathbb{E}[X] = \mathbb{E}[Y]$.

Expectation

Definition

The **expected value** (or **expectation** or **mean**) of a discrete r.v. X with possible values, $x_1, x_2, \ldots$ is

$$\mathbb{E}[X] = \sum_{j=1}^{\infty} x_j \mathbb{P}(X = x_j)$$

- Weighted average of the values of the r.v. weighted by the probability of each value occurring.
  - $\mathbb{E}[X]$ is a constant!
- Example: $X \sim \text{Bern}(p)$, then $\mathbb{E}[X] = 1p + 0(1-p) = p$.
- If X and Y have the same distribution, then $\mathbb{E}[X] = \mathbb{E}[Y]$.
  - Converse isn't true!

Example - number of treated units

- Randomized experiment with 3 units. X is the number of treated units.

x	$p_X(x)$	$x \cdot p_X(x)$
0	$1/8$	0
1	$3/8$	$3/8$
2	$3/8$	$6/8$
3	$1/8$	$3/8$

Example - number of treated units

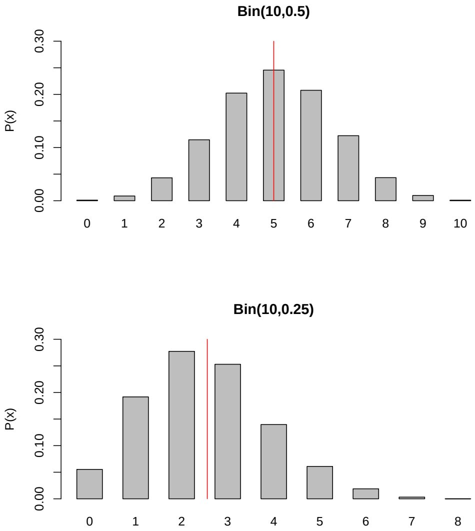
- Randomized experiment with 3 units. X is the number of treated units.

x	$p_X(x)$	$x \cdot p_X(x)$
0	$1/8$	0
1	$3/8$	$3/8$
2	$3/8$	$6/8$
3	$1/8$	$3/8$

- Calculate the expectation of X:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{j=1}^{k} x_j \mathbb{P}(X = x_j) \\
&= 0 \cdot \mathbb{P}(X = 0) + 1 \cdot \mathbb{P}(X = 1) + 2 \cdot \mathbb{P}(X = 2) + 3 \cdot \mathbb{P}(X = 3) \\
&= 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{12}{8} = 1.5
\end{aligned}$$

Expectation as balancing point



Properties of the expected value

- We can derive expectation of **transformations** of other r.v.s

Properties of the expected value

- We can derive expectation of **transformations** of other r.v.s
- Possible for **linear** functions because expectation is **linear**:

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

$$\mathbb{E}[aX] = a\mathbb{E}[X] \quad \text{if } a \text{ is a constant}$$

Properties of the expected value

- We can derive expectation of **transformations** of other r.v.s
- Possible for **linear** functions because expectation is **linear**:

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

$$\mathbb{E}[aX] = a\mathbb{E}[X] \quad \text{if } a \text{ is a constant}$$

- True even if X and Y are dependent!

Properties of the expected value

- We can derive expectation of **transformations** of other r.v.s
- Possible for **linear** functions because expectation is **linear**:

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

$$\mathbb{E}[aX] = a\mathbb{E}[X] \quad \text{if } a \text{ is a constant}$$

- True even if X and Y are dependent!
- But this isn't always true for nonlinear operations:

Properties of the expected value

- We can derive expectation of **transformations** of other r.v.s
- Possible for **linear** functions because expectation is **linear**:

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

$$\mathbb{E}[aX] = a\mathbb{E}[X] \quad \text{if } a \text{ is a constant}$$

- True even if X and Y are dependent!
- But this isn't always true for nonlinear operations:
  - $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$ unless $g(\cdot)$ is a linear function.

Properties of the expected value

- We can derive expectation of **transformations** of other r.v.s
- Possible for **linear** functions because expectation is **linear**:

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

$$\mathbb{E}[aX] = a\mathbb{E}[X] \quad \text{if } a \text{ is a constant}$$

- True even if X and Y are dependent!
- But this isn't always true for nonlinear operations:
  - $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$ unless $g(\cdot)$ is a linear function.
  - $\mathbb{E}[XY] \neq \mathbb{E}[X]\mathbb{E}[Y]$ unless X and Y are independent.

Examples

- $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$ unless $g(\cdot)$ is a linear function.

Examples

- $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$ unless $g(\cdot)$ is a linear function.
  - Let $g(X) = X^2$

Examples

- $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$ unless $g(\cdot)$ is a linear function.
  - Let $g(X) = X^2$
  - $\mathbb{E}[g(X)] = \mathbb{E}[X^2]$

Examples

- $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$ unless $g(\cdot)$ is a linear function.
  - Let $g(X) = X^2$
  - $\mathbb{E}[g(X)] = \mathbb{E}[X^2]$
  - $g(\mathbb{E}[X]) = (\mathbb{E}[X])^2$

Examples

- $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$ unless $g(\cdot)$ is a linear function.
  - Let $g(X) = X^2$
  - $\mathbb{E}[g(X)] = \mathbb{E}[X^2]$
  - $g(\mathbb{E}[X]) = (\mathbb{E}[X])^2$
- $\mathbb{E}[XY] \neq \mathbb{E}[X]\mathbb{E}[Y]$ unless X and Y are independent.

Examples

- $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$ unless $g(\cdot)$ is a linear function.
  - Let $g(X) = X^2$
  - $\mathbb{E}[g(X)] = \mathbb{E}[X^2]$
  - $g(\mathbb{E}[X]) = (\mathbb{E}[X])^2$
- $\mathbb{E}[XY] \neq \mathbb{E}[X]\mathbb{E}[Y]$ unless X and Y are independent.

Examples

- $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$ unless $g(\cdot)$ is a linear function.
  - Let $g(X) = X^2$
  - $\mathbb{E}[g(X)] = \mathbb{E}[X^2]$
  - $g(\mathbb{E}[X]) = (\mathbb{E}[X])^2$
- $\mathbb{E}[XY] \neq \mathbb{E}[X]\mathbb{E}[Y]$ unless X and Y are independent.
  - Let $X = 2Y$

Examples

- $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$ unless $g(\cdot)$ is a linear function.
  - Let $g(X) = X^2$
  - $\mathbb{E}[g(X)] = \mathbb{E}[X^2]$
  - $g(\mathbb{E}[X]) = (\mathbb{E}[X])^2$
- $\mathbb{E}[XY] \neq \mathbb{E}[X]\mathbb{E}[Y]$ unless X and Y are independent.
  - Let $X = 2Y$
  - $\mathbb{E}[XY] = \mathbb{E}[2Y^2] = 2\mathbb{E}[Y^2]$

Examples

- $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$ unless $g(\cdot)$ is a linear function.
  - Let $g(X) = X^2$
  - $\mathbb{E}[g(X)] = \mathbb{E}[X^2]$
  - $g(\mathbb{E}[X]) = (\mathbb{E}[X])^2$
- $\mathbb{E}[XY] \neq \mathbb{E}[X]\mathbb{E}[Y]$ unless X and Y are independent.
  - Let $X = 2Y$
  - $\mathbb{E}[XY] = \mathbb{E}[2Y^2] = 2\mathbb{E}[Y^2]$
  - $\mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[2Y]\mathbb{E}[Y] = 2(\mathbb{E}[Y])^2$

Expectation of a binomial

- Let $X \sim \text{Bin}(n, p)$, what's $\mathbb{E}[X]$? Could just plug in formula:

$$\mathbb{E}[X] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = ??$$

Expectation of a binomial

- Let $X \sim \text{Bin}(n, p)$, what's $\mathbb{E}[X]$? Could just plug in formula:

$$\mathbb{E}[X] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = ??$$

- Use the story of the binomial as a sum of n Bernoulli $X_i \sim \text{Bern}(p)$

$$X = X_1 + \cdots + X_n$$

Expectation of a binomial

- Let $X \sim \text{Bin}(n, p)$, what's $\mathbb{E}[X]$? Could just plug in formula:

$$\mathbb{E}[X] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = ??$$

- Use the story of the binomial as a sum of n Bernoulli $X_i \sim \text{Bern}(p)$

$$X = X_1 + \cdots + X_n$$

- Use linearity:

$$\mathbb{E}[X] = \mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = np$$

Expectation of the sample mean

- Let $X_1, \ldots, X_n$ be identically distributed with $\mathbb{E}[X_i] = \mu$.

Expectation of the sample mean

- Let $X_1, \ldots, X_n$ be identically distributed with $\mathbb{E}[X_i] = \mu$.
- Define the **sample mean** to be

$$\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i.$$

Expectation of the sample mean

- Let $X_1, \ldots, X_n$ be identically distributed with $\mathbb{E}[X_i] = \mu$.
- Define the **sample mean** to be

$$\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i.$$

  - $\bar{X}$ is a r.v.!

Expectation of the sample mean

- Let $X_1, \ldots, X_n$ be identically distributed with $\mathbb{E}[X_i] = \mu$.
- Define the **sample mean** to be

$$\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i.$$

  - $\bar{X}$ is a r.v.!
- We can find the expectation of the sample mean using linearity:

$$\mathbb{E}[\bar{X}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n} n\mu = \mu$$

Expectation of the sample mean

- Let $X_1, \ldots, X_n$ be identically distributed with $\mathbb{E}[X_i] = \mu$.
- Define the **sample mean** to be

$$\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i.$$

  - $\bar{X}$ is a r.v.!

- We can find the expectation of the sample mean using linearity:

$$\mathbb{E}[\bar{X}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}n\mu = \mu$$

- **Intuition**: on average, the sample mean is equal to the population mean.

Monotonicity of expectations

- Expectations don't have to be in the support of the data.

Monotonicity of expectations

- Expectations don't have to be in the support of the data.
  - $X \sim \text{Bern}(p)$ has $\mathbb{E}[X] = p$ which isn't 0 or 1.

Monotonicity of expectations

- Expectations don't have to be in the support of the data.
  - $X \sim \text{Bern}(p)$ has $\mathbb{E}[X] = p$ which isn't 0 or 1.
- But it must be between the highest and lowest possible value of an r.v.

Monotonicity of expectations

- Expectations don't have to be in the support of the data.
  - $X \sim \text{Bern}(p)$ has $\mathbb{E}[X] = p$ which isn't 0 or 1.
- But it must be between the highest and lowest possible value of an r.v.
  - If $\mathbb{P}(X \geq c) = 1$, then $\mathbb{E}[X] \geq c$.

Monotonicity of expectations

- Expectations don't have to be in the support of the data.
  - $X \sim \text{Bern}(p)$ has $\mathbb{E}[X] = p$ which isn't 0 or 1.
- But it must be between the highest and lowest possible value of an r.v.
  - If $\mathbb{P}(X \geq c) = 1$, then $\mathbb{E}[X] \geq c$.
  - If $\mathbb{P}(X \leq c) = 1$, then $\mathbb{E}[X] \leq c$.

Monotonicity of expectations

- Expectations don't have to be in the support of the data.
  - $X \sim \text{Bern}(p)$ has $\mathbb{E}[X] = p$ which isn't 0 or 1.
- But it must be between the highest and lowest possible value of an r.v.
  - If $\mathbb{P}(X \geq c) = 1$, then $\mathbb{E}[X] \geq c$.
  - If $\mathbb{P}(X \leq c) = 1$, then $\mathbb{E}[X] \leq c$.
- Useful application of linearity: expectation is **monotone**.

Monotonicity of expectations

- Expectations don't have to be in the support of the data.
  - $X \sim \text{Bern}(p)$ has $\mathbb{E}[X] = p$ which isn't 0 or 1.
- But it must be between the highest and lowest possible value of an r.v.
  - If $\mathbb{P}(X \geq c) = 1$, then $\mathbb{E}[X] \geq c$.
  - If $\mathbb{P}(X \leq c) = 1$, then $\mathbb{E}[X] \leq c$.
- Useful application of linearity: expectation is **monotone**.
  - If $X \geq Y$ with probability 1, then $\mathbb{E}(X) \geq \mathbb{E}(Y)$.

St. Petersburg Paradox

- Game of chance: stranger pays you \$ 2^X where X is the number of flips with a fair coin until the first heads.

St. Petersburg Paradox

- Game of chance: stranger pays you \$ 2^X where X is the number of flips with a fair coin until the first heads.

St. Petersburg Paradox

- Game of chance: stranger pays you \$ 2^X where X is the number of flips with a fair coin until the first heads.
  - Probability of reaching $X = k$ is:

$$\begin{aligned}\mathbb{P}(X = k) &= \mathbb{P}(T_1 \cap T_2 \cap \cdots \cap T_{k-1} \cap H_k) \\ &= \mathbb{P}(T_1)\mathbb{P}(T_2)\ldots\mathbb{P}(T_{k-1})\mathbb{P}(H_k) \\ &= \frac{1}{2^k}\end{aligned}$$

St. Petersburg Paradox

- How much would you be willing to pay to play the game?

St. Petersburg Paradox

- How much would you be willing to pay to play the game?
- Let payout be $Y = 2^X$, we want $\mathbb{E}[Y]$:

St. Petersburg Paradox

- How much would you be willing to pay to play the game?
- Let payout be $Y = 2^X$, we want $\mathbb{E}[Y]$:

St. Petersburg Paradox

- How much would you be willing to pay to play the game?
- Let payout be $Y = 2^X$, we want $\mathbb{E}[Y]$:

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

- Two ways to resolve the "paradox":

St. Petersburg Paradox

- How much would you be willing to pay to play the game?
- Let payout be $Y = 2^X$, we want $\mathbb{E}[Y]$:

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

- Two ways to resolve the "paradox":
  - No infinite money: max payout of 2^{40} (around a trillion) $\Rightarrow \mathbb{E}[Y] = 41$

St. Petersburg Paradox

- How much would you be willing to pay to play the game?
- Let payout be $Y = 2^X$, we want $\mathbb{E}[Y]$:

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

- Two ways to resolve the "paradox":
  - No infinite money: max payout of 2^{40} (around a trillion) $\Rightarrow \mathbb{E}[Y] = 41$
  - Risk avoidance/concave utility $U = Y^{1/2} \Rightarrow \mathbb{E}[U(Y)] \approx 2.41$

Undefined expectations

- We saw $\mathbb{E}[X]$ can be infinite, but it can also be undefined.

Undefined expectations

- We saw $\mathbb{E}[X]$ can be infinite, but it can also be undefined.
- Example: X takes 2^k and -2^k each with prob 2^{-k-1}.

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} 2^k 2^{-k-1} - \sum_{k=1}^{\infty} 2^k 2^{-k-1} = \sum_{k=1}^{\infty} \frac{1}{2} - \sum_{k=1}^{\infty} \frac{1}{2} = \infty - \infty$$

Undefined expectations

- We saw $\mathbb{E}[X]$ can be infinite, but it can also be undefined.
- Example: X takes 2^k and -2^k each with prob 2^{-k-1}.

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} 2^k 2^{-k-1} - \sum_{k=1}^{\infty} 2^k 2^{-k-1} = \sum_{k=1}^{\infty} \frac{1}{2} - \sum_{k=1}^{\infty} \frac{1}{2} = \infty - \infty$$

- Often, both of these are assumed away by assuming $\mathbb{E}[|X|] < \infty$ which implies $\mathbb{E}[X]$ exists and is finite.

Indicator variables/fundamental bridge

- The probability of an event is equal to the expectation of its indicator:

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

Indicator variables/fundamental bridge

- The probability of an event is equal to the expectation of its indicator:

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

- **Fundamental bridge** between probability and expectation

Indicator variables/fundamental bridge

- The probability of an event is equal to the expectation of its indicator:

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

- **Fundamental bridge** between probability and expectation
- Makes it easy to prove probability results like **Bonferroni's inequality**

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) \leq \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n)$$

Indicator variables/fundamental bridge

- The probability of an event is equal to the expectation of its indicator:

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

- **Fundamental bridge** between probability and expectation
- Makes it easy to prove probability results like **Bonferroni's inequality**

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) \leq \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n)$$

- Use the fact that $\mathbb{I}(A_1 \cup \cdots \cup A_n) \leq \mathbb{I}(A_1) + \cdots + \mathbb{I}(A_n)$ and then take expectations.

Proof of Bonferroni's Inequality

Bonferroni's Inequality: For any finite set of events $A_1, A_2, \ldots, A_n$, we have:

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) \leq P(A_1) + P(A_2) + \cdots + P(A_n)$$

Proof: Using indicator variables.

- Define the indicator random variable for each event:

$$I(A_i) = \begin{cases} 1, & \text{if } A_i \text{ occurs} \\ 0, & \text{otherwise} \end{cases}$$

Proof of Bonferroni's Inequality

$$I(A_i) = \begin{cases} 1, & \text{if } A_i \text{ occurs} \\ 0, & \text{otherwise} \end{cases}$$

- Consider the indicator for the union:

$$I(A_1 \cup A_2 \cup \cdots \cup A_n) \leq I(A_1) + I(A_2) + \cdots + I(A_n)$$

Proof of Bonferroni's Inequality

$$I(A_i) = \begin{cases} 1, & \text{if } A_i \text{ occurs} \\ 0, & \text{otherwise} \end{cases}$$

- Consider the indicator for the union:

$$I(A_1 \cup A_2 \cup \cdots \cup A_n) \leq I(A_1) + I(A_2) + \cdots + I(A_n)$$

- Taking expectations on both sides:

$$E[I(A_1 \cup A_2 \cup \cdots \cup A_n)] \leq E[I(A_1)] + E[I(A_2)] + \cdots + E[I(A_n)]$$

Proof of Bonferroni's Inequality

$$I(A_i) = \begin{cases} 1, & \text{if } A_i \text{ occurs} \\ 0, & \text{otherwise} \end{cases}$$

- Consider the indicator for the union:

$$I(A_1 \cup A_2 \cup \cdots \cup A_n) \leq I(A_1) + I(A_2) + \cdots + I(A_n)$$

- Taking expectations on both sides:

$$E[I(A_1 \cup A_2 \cup \cdots \cup A_n)] \leq E[I(A_1)] + E[I(A_2)] + \cdots + E[I(A_n)]$$

- Since $E[I(A_i)] = P(A_i)$, we get:

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) \leq P(A_1) + P(A_2) + \cdots + P(A_n)$$

Using indicators to find expectations

- Suppose we are assigning n units to k treatments and all possibilities are equally likely. What is the expected number of treatment conditions without any units?

Using indicators to find expectations

- Suppose we are assigning n units to k treatments and all possibilities are equally likely. What is the expected number of treatment conditions without any units?
- Use indicators! $I_j = 1$ if jth condition is empty. So $I_1 + \cdots + I_k$ is the number of empty conditions.

$$\begin{aligned}
\mathbb{E}[I_j] &= \mathbb{P}(\text{cond } j \text{ empty}) \\
&= \mathbb{P}(\{\text{unit 1 not in cond } j\} \cap \cdots \cap \{\text{unit } n \text{ not in cond } j\}) \\
&= \mathbb{P}(\{\text{unit 1 not in cond } j\}) \dots \mathbb{P}(\{\text{unit } n \text{ not in cond } j\}) \\
&= \left(1 - \frac{1}{k}\right)^n
\end{aligned}$$

Using indicators to find expectations

- Suppose we are assigning n units to k treatments and all possibilities are equally likely. What is the expected number of treatment conditions without any units?
- Use indicators! $I_j = 1$ if jth condition is empty. So $I_1 + \cdots + I_k$ is the number of empty conditions.

$$\begin{aligned}\mathbb{E}[I_j] &= \mathbb{P}(\text{cond } j \text{ empty}) \\ &= \mathbb{P}(\{\text{unit 1 not in cond } j\} \cap \cdots \cap \{\text{unit } n \text{ not in cond } j\}) \\ &= \mathbb{P}(\{\text{unit 1 not in cond } j\}) \ldots \mathbb{P}(\{\text{unit } n \text{ not in cond } j\}) \\ &= \left(1 - \frac{1}{k}\right)^n\end{aligned}$$

- Thus, we have :

$$\mathbb{E}\left[\sum_j I_j\right] = k\left(1 - \frac{1}{k}\right)^n.$$

Variance

- The **variance** measures the spread of the distribution:

$$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

Variance

- The **variance** measures the spread of the distribution:

$$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

- Could also use $\mathbb{E}[|X - \mathbb{E}[X]|]$ but more clunky as a function.

Variance

- The **variance** measures the spread of the distribution:

$$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

- Could also use $\mathbb{E}[|X - \mathbb{E}[X]|]$ but more clunky as a function.
- Weighted average of the squared distances from the mean.

Variance

- The **variance** measures the spread of the distribution:

$$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

- Could also use $\mathbb{E}[|X - \mathbb{E}[X]|]$ but more clunky as a function.
- Weighted average of the squared distances from the mean.
  - Larger deviations ($+$ or $-$) $\rightsquigarrow$ higher variance.

Variance

- The **variance** measures the spread of the distribution:

$$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

- Could also use $\mathbb{E}[|X - \mathbb{E}[X]|]$ but more clunky as a function.
- Weighted average of the squared distances from the mean.
  - Larger deviations ($+$ or $-$) $\rightsquigarrow$ higher variance.
- The **standard deviation** is the (positive) square root of the variance:

$$SD(X) = \sqrt{\mathbb{V}[X]}$$

Variance

- The **variance** measures the spread of the distribution:

$$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

- Could also use $\mathbb{E}[|X - \mathbb{E}[X]|]$ but more clunky as a function.
- Weighted average of the squared distances from the mean.
  - Larger deviations ($+$ or $-$) $\rightsquigarrow$ higher variance.
- The **standard deviation** is the (positive) square root of the variance:

$$SD(X) = \sqrt{\mathbb{V}[X]}$$

- Useful equivalent representation of the variance:

$$\mathbb{V}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

LOTUS

- How do we calculate $\mathbb{E}[X^2]$ since it's nonlinear?

Definition

The **Law of the Unconscious Statistician**, or LOTUS, states that if $g(X)$ is a function of a discrete random variable, then

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

LOTUS

- How do we calculate $\mathbb{E}[X^2]$ since it's nonlinear?

Definition

The **Law of the Unconscious Statistician**, or LOTUS, states that if $g(X)$ is a function of a discrete random variable, then

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

- Example: $\mathbb{E}[X^2]$ where $X \sim \text{Bin}(n, p)$.

$$\mathbb{E}[X] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}$$

LOTUS

- How do we calculate $\mathbb{E}[X^2]$ since it's nonlinear?

Definition

The **Law of the Unconscious Statistician**, or LOTUS, states that if $g(X)$ is a function of a discrete random variable, then

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

- Example: $\mathbb{E}[X^2]$ where $X \sim \text{Bin}(n, p)$.

$$\mathbb{E}[X] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}$$

LOTUS

- How do we calculate $\mathbb{E}[X^2]$ since it's nonlinear?

Definition

The **Law of the Unconscious Statistician**, or LOTUS, states that if $g(X)$ is a function of a discrete random variable, then

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

- Example: $\mathbb{E}[X^2]$ where $X \sim \text{Bin}(n, p)$.

$$\mathbb{E}[X] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}$$

$$\mathbb{E}[X^2] = \sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k}$$

Example - number of treated units

- Use LOTUS to calculate the variance for a discrete r.v.: $\mathbb{V}[X] = \sum_{j=1}^{k} (x_j - \mathbb{E}[X])^2 \mathbb{P}(X = x_j)$

x	$p_X(x)$	$x - \mathbb{E}[X]$	$(x - \mathbb{E}[X])^2$
0	1/8	-1.5	2.25
1	3/8	-0.5	0.25
2	3/8	0.5	0.25
3	1/8	1.5	2.25

Example - number of treated units

- Use LOTUS to calculate the variance for a discrete r.v.: $\mathbb{V}[X] = \sum_{j=1}^{k}(x_j - \mathbb{E}[X])^2 \mathbb{P}(X = x_j)$

x	$p_X(x)$	$x - \mathbb{E}[X]$	$(x - \mathbb{E}[X])^2$
0	1/8	-1.5	2.25
1	3/8	-0.5	0.25
2	3/8	0.5	0.25
3	1/8	1.5	2.25

- Let's go back to the number of treated units to figure out the variance of the number of treated units:

$$\begin{aligned}
\mathbb{V}[X] &= \sum_{j=1}^{k}(x_j - \mathbb{E}[X])^2 p_X(x_j) \\
&= (-1.5)^2 \times \frac{1}{8} + (-0.5)^2 \times \frac{3}{8} + (0.5)^2 \times \frac{3}{8} + (1.5)^2 \times \frac{1}{8} \\
&= 2.25 \times \frac{1}{8} + 0.25 \times \frac{3}{8} + 0.25 \times \frac{3}{8} + 2.25 \times \frac{1}{8} = 0.75
\end{aligned}$$

Properties of variances

1. $\mathbb{V}[X + c] = \mathbb{V}[X]$ for any constant c.

Properties of variances

1. $\mathbb{V}[X + c] = \mathbb{V}[X]$ for any constant c.
2. If a is a constant, $\mathbb{V}[aX] = a^2\mathbb{V}[X]$.

Properties of variances

1. $\mathbb{V}[X + c] = \mathbb{V}[X]$ for any constant c.
2. If a is a constant, $\mathbb{V}[aX] = a^2\mathbb{V}[X]$.
3. If X and Y are **independent**, then $\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y]$.

Properties of variances

1. $\mathbb{V}[X + c] = \mathbb{V}[X]$ for any constant c.
2. If a is a constant, $\mathbb{V}[aX] = a^2\mathbb{V}[X]$.
3. If X and Y are **independent**, then $\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y]$.
    - But this doesn't hold for dependent r.v.s.

Properties of variances

1. $\mathbb{V}[X + c] = \mathbb{V}[X]$ for any constant c.
2. If a is a constant, $\mathbb{V}[aX] = a^2\mathbb{V}[X]$.
3. If X and Y are **independent**, then $\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y]$.
   - But this doesn't hold for dependent r.v.s.
4. $\mathbb{V}[X] \geq 0$ with equality holding only if X is a constant, $\mathbb{P}(X = b) = 1$.

Binomial variance

- Clunky to use LOTUS to calculate variances. Other ways?

Binomial variance

- Clunky to use LOTUS to calculate variances. Other ways?
  - Use stories and indicator variables!

Binomial variance

- Clunky to use LOTUS to calculate variances. Other ways?
  - Use stories and indicator variables!
- $X \sim \text{Bin}(n, p)$ is equivalent to $X_1 + \cdots + X_n$ where $X_i \sim \text{Bern}(p)$.

Binomial variance

- Clunky to use LOTUS to calculate variances. Other ways?
  - Use stories and indicator variables!
- $X \sim \text{Bin}(n, p)$ is equivalent to $X_1 + \cdots + X_n$ where $X_i \sim \text{Bern}(p)$.
- Variance of a Bernoulli:

$$V[X_i] = \mathbb{E}[X_i^2] - (\mathbb{E}[X_i])^2 = p - p^2 = p(1-p)$$

Binomial variance

- Clunky to use LOTUS to calculate variances. Other ways?
  - Use stories and indicator variables!
- $X \sim \text{Bin}(n, p)$ is equivalent to $X_1 + \cdots + X_n$ where $X_i \sim \text{Bern}(p)$.
- Variance of a Bernoulli:

$$V[X_i] = \mathbb{E}[X_i^2] - (\mathbb{E}[X_i])^2 = p - p^2 = p(1-p)$$

  - (Used $X_i^2 = X_i$ for indicator variables)

Binomial variance

- Clunky to use LOTUS to calculate variances. Other ways?
  - Use stories and indicator variables!
- $X \sim \text{Bin}(n, p)$ is equivalent to $X_1 + \cdots + X_n$ where $X_i \sim \text{Bern}(p)$.
- Variance of a Bernoulli:

$$V[X_i] = \mathbb{E}[X_i^2] - (\mathbb{E}[X_i])^2 = p - p^2 = p(1-p)$$

  - (Used $X_i^2 = X_i$ for indicator variables)
- Binomials are the sum of **independent** Bernoulli r.v.s so:

$$V[X] = V[X_1 + \cdots + X_n] = V[X_1] + \cdots + V[X_n] = np(1-p)$$

Variance of the sample mean

- Let $X_1, \ldots, X_n$ be i.i.d. with $\mathbb{E}[X_i] = \mu$ and $V[X_i] = \sigma^2$.

Variance of the sample mean

- Let $X_1, \ldots, X_n$ be i.i.d. with $\mathbb{E}[X_i] = \mu$ and $V[X_i] = \sigma^2$.
  - Earlier we saw that $\mathbb{E}[\bar{X}_n] = \mu$, what about $V[\bar{X}_n]$?

Variance of the sample mean

- Let $X_1, \ldots, X_n$ be i.i.d. with $\mathbb{E}[X_i] = \mu$ and $V[X_i] = \sigma^2$.
  - Earlier we saw that $\mathbb{E}[\bar{X}_n] = \mu$, what about $V[\bar{X}_n]$?
- We can apply the rules of variances:

$$V[\bar{X}_n] = V\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} V[X_i] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Variance of the sample mean

- Let $X_1, \dots, X_n$ be i.i.d. with $\mathbb{E}[X_i] = \mu$ and $V[X_i] = \sigma^2$.
  - Earlier we saw that $\mathbb{E}[\bar{X}_n] = \mu$, what about $V[\bar{X}_n]$?
- We can apply the rules of variances:

$$V[\bar{X}_n] = V\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} V[X_i] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

  - Note: we needed independence and identical distribution for this.

Variance of the sample mean

- Let $X_1, \ldots, X_n$ be i.i.d. with $\mathbb{E}[X_i] = \mu$ and $V[X_i] = \sigma^2$.
  - Earlier we saw that $\mathbb{E}[\bar{X}_n] = \mu$, what about $V[\bar{X}_n]$?
- We can apply the rules of variances:

$$V[\bar{X}_n] = V\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} V[X_i] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

  - Note: we needed independence and identical distribution for this.
  - $SD(\bar{X}_n) = \sigma/\sqrt{n}$.

Variance of the sample mean

- Let $X_1, \ldots, X_n$ be i.i.d. with $\mathbb{E}[X_i] = \mu$ and $V[X_i] = \sigma^2$.
  - Earlier we saw that $\mathbb{E}[\bar{X}_n] = \mu$, what about $V[\bar{X}_n]$?
- We can apply the rules of variances:

$$V[\bar{X}_n] = V\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} V[X_i] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

  - Note: we needed independence and identical distribution for this.
  - $SD(\bar{X}_n) = \sigma/\sqrt{n}$.
- Under i.i.d. sampling we know the expectation and variance of $\bar{X}_n$ without any other assumptions about the distribution of the X_i!

Variance of the sample mean

- Let $X_1, \ldots, X_n$ be i.i.d. with $\mathbb{E}[X_i] = \mu$ and $V[X_i] = \sigma^2$.
  - Earlier we saw that $\mathbb{E}[\bar{X}_n] = \mu$, what about $V[\bar{X}_n]$?
- We can apply the rules of variances:

$$V[\bar{X}_n] = V\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} V[X_i] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

  - Note: we needed independence and identical distribution for this.
  - $SD(\bar{X}_n) = \sigma/\sqrt{n}$.
- Under i.i.d. sampling we know the expectation and variance of $\bar{X}_n$ without any other assumptions about the distribution of the X_i!
  - We don't know what distribution it takes though!

Inequalities

- Bounds are very important in establishing unknown probabilities.

Inequalities

- Bounds are very important in establishing unknown probabilities.
  - Also very helpful in establishing limit results later on.

Inequalities

- Bounds are very important in establishing unknown probabilities.
  - Also very helpful in establishing limit results later on.
- Remember that $\mathbb{E}[a + bX] = a + b\mathbb{E}[X]$ is linear, but $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$ for nonlinear functions.

Inequalities

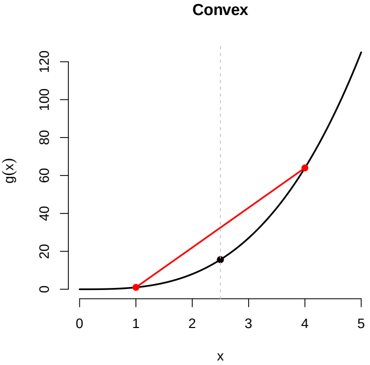
- Bounds are very important in establishing unknown probabilities.
  - Also very helpful in establishing limit results later on.
- Remember that $\mathbb{E}[a + bX] = a + b\mathbb{E}[X]$ is linear, but $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$ for nonlinear functions.
- Can we relate those? Yes, for **convex** and **concave** functions.

Convex and Concave

Definition (Convex Function). A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *convex* if for all $x, y \in \mathbb{R}^n$ and for all $\lambda \in [0, 1]$, we have

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y).$$

This means that the line segment connecting any two points on the graph of f lies above or on the graph.

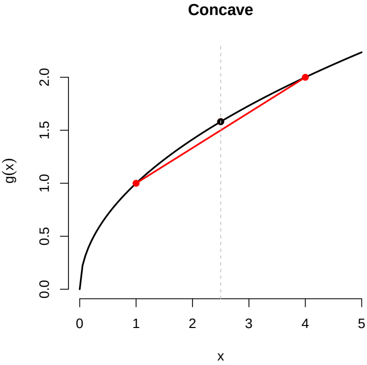


Convex and Concave

Definition (Concave Function). A function $g : \mathbb{R}^n \to \mathbb{R}$ is said to be *concave* if for all $x, y \in \mathbb{R}^n$ and for all $\lambda \in [0, 1]$, we have

$$g(\lambda x + (1 - \lambda)y) \geq \lambda g(x) + (1 - \lambda)g(y).$$

This means that the line segment connecting any two points on the graph of g lies below or on the graph.



Jensen's inequality

Jensen's inequality

Let X be a r.v. Then, we have

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]) \quad \text{if } g \text{ is convex}$$

$$\mathbb{E}[g(X)] \leq g(\mathbb{E}[X]) \quad \text{if } g \text{ is concave}$$

with equality only holding if g is linear.

- Makes proving variance positive simple.

Jensen's inequality

Jensen's inequality

Let X be a r.v. Then, we have

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]) \quad \text{if } g \text{ is convex}$$

$$\mathbb{E}[g(X)] \leq g(\mathbb{E}[X]) \quad \text{if } g \text{ is concave}$$

with equality only holding if g is linear.

- Makes proving variance positive simple.
  - $g(x) = x^2$ is convex, so $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$.

Jensen's inequality

Jensen's inequality

Let X be a r.v. Then, we have

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]) \quad \text{if } g \text{ is convex}$$

$$\mathbb{E}[g(X)] \leq g(\mathbb{E}[X]) \quad \text{if } g \text{ is concave}$$

with equality only holding if g is linear.

- Makes proving variance positive simple.
  - $g(x) = x^2$ is convex, so $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$.
- Allows us to easily reason about complicated functions:

Jensen's inequality

Jensen's inequality

Let X be a r.v. Then, we have

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]) \quad \text{if } g \text{ is convex}$$

$$\mathbb{E}[g(X)] \leq g(\mathbb{E}[X]) \quad \text{if } g \text{ is concave}$$

with equality only holding if g is linear.

- Makes proving variance positive simple.
  - $g(x) = x^2$ is convex, so $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$.
- Allows us to easily reason about complicated functions:
  - $\mathbb{E}[|X|] \geq |\mathbb{E}[X]|$

Jensen's inequality

Jensen's inequality

Let X be a r.v. Then, we have

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]) \quad \text{if } g \text{ is convex}$$

$$\mathbb{E}[g(X)] \leq g(\mathbb{E}[X]) \quad \text{if } g \text{ is concave}$$

with equality only holding if g is linear.

- Makes proving variance positive simple.
  - $g(x) = x^2$ is convex, so $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$.
- Allows us to easily reason about complicated functions:
  - $\mathbb{E}[|X|] \geq |\mathbb{E}[X]|$
  - $\mathbb{E}[1/X] \geq 1/\mathbb{E}[X]$

Jensen's inequality

Jensen's inequality

Let X be a r.v. Then, we have

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]) \quad \text{if } g \text{ is convex}$$

$$\mathbb{E}[g(X)] \leq g(\mathbb{E}[X]) \quad \text{if } g \text{ is concave}$$

with equality only holding if g is linear.

- Makes proving variance positive simple.
  - $g(x) = x^2$ is convex, so $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$.
- Allows us to easily reason about complicated functions:
  - $\mathbb{E}[|X|] \geq |\mathbb{E}[X]|$
  - $\mathbb{E}[1/X] \geq 1/\mathbb{E}[X]$
  - $\mathbb{E}[\log(X)] \leq \log(\mathbb{E}[X])$

Poisson

Definition

An r.v. X has the **Poisson distribution** with parameter $\lambda > 0$, written $X \sim \text{Pois}(\lambda)$ if the p.m.f. of X is:

$$\mathbb{P}(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

- One more discrete distribution is very popular, especially for counts.

Poisson

Definition

An r.v. X has the **Poisson distribution** with parameter $\lambda > 0$, written $X \sim \text{Pois}(\lambda)$ if the p.m.f. of X is:

$$\mathbb{P}(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

- One more discrete distribution is very popular, especially for counts.
  - Number of contributions a candidate for office receives in a day.

Poisson

Definition

An r.v. X has the **Poisson distribution** with parameter $\lambda > 0$, written $X \sim \text{Pois}(\lambda)$ if the p.m.f. of X is:

$$\mathbb{P}(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

- One more discrete distribution is very popular, especially for counts.
  - Number of contributions a candidate for office receives in a day.
- Key calculus fact that makes this a valid p.m.f.: $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$.

Poisson properties

- A Poisson r.v. $X \sim \text{Pois}(\lambda)$ has an unusual property:

$$\mathbb{E}[X] = \mathbb{V}[X] = \lambda$$

Poisson properties

- A Poisson r.v. $X \sim \text{Pois}(\lambda)$ has an unusual property:

$$\mathbb{E}[X] = \mathbb{V}[X] = \lambda$$

- The sum of independent Poisson r.v.s is Poisson:

$$X \sim \text{Pois}(\lambda_1) \quad Y \sim \text{Pois}(\lambda_2) \quad \Rightarrow \quad X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$$

Poisson properties

- A Poisson r.v. $X \sim \text{Pois}(\lambda)$ has an unusual property:

$$\mathbb{E}[X] = \mathbb{V}[X] = \lambda$$

- The sum of independent Poisson r.v.s is Poisson:

$$X \sim \text{Pois}(\lambda_1) \quad Y \sim \text{Pois}(\lambda_2) \quad \Rightarrow \quad X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$$

- If $X \sim \text{Bin}(n, p)$ with n large and p small, then X is approximately $\text{Pois}(np)$.

Poisson Distribution as a Limiting Case of Binomial

Claim: If $X_n \sim \text{Bin}(n, p_n)$ with $n \to \infty$ and $p_n \to 0$ such that $np_n = \lambda$, then:

$$\lim_{n\to\infty} P(X_n = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

which is the **Poisson** distribution $X \sim \text{Pois}(\lambda)$.

Proof:

- The binomial probability mass function (p.m.f.) is:

$$P(X_n = k) = \binom{n}{k} p_n^k (1 - p_n)^{n-k}$$

Poisson Distribution as a Limiting Case of Binomial

Claim: If $X_n \sim \text{Bin}(n, p_n)$ with $n \to \infty$ and $p_n \to 0$ such that $np_n = \lambda$, then:

$$\lim_{n\to\infty} P(X_n = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

which is the **Poisson** distribution $X \sim \text{Pois}(\lambda)$.

Proof:

- The binomial probability mass function (p.m.f.) is:

$$P(X_n = k) = \binom{n}{k} p_n^k (1 - p_n)^{n-k}$$

- Using the factorial form:

$$P(X_n = k) = \frac{n!}{k!(n-k)!} p_n^k (1 - p_n)^{n-k}$$

Proof

- Approximate the terms as $n \to \infty$:

$$\frac{n!}{(n-k)!} \approx n^k$$

Proof

- Approximate the terms as $n \to \infty$:

$$\frac{n!}{(n-k)!} \approx n^k$$

- Using $p_n = \lambda/n$, rewrite the p.m.f.:

$$P(X_n = k) \approx \frac{n^k}{k!} \left(\frac{\lambda}{n}\right)^k (1 - \lambda/n)^{n-k}$$

Proof

- Approximate the terms as $n \to \infty$:

$$\frac{n!}{(n-k)!} \approx n^k$$

- Using $p_n = \lambda/n$, rewrite the p.m.f.:

$$P(X_n = k) \approx \frac{n^k}{k!} \left(\frac{\lambda}{n}\right)^k (1 - \lambda/n)^{n-k}$$

- Taking the limit as $n \to \infty$:

$$\lim_{n \to \infty} (1 - \lambda/n)^{n-k} = e^{-\lambda}$$

Proof

- Approximate the terms as $n \to \infty$:

$$\frac{n!}{(n-k)!} \approx n^k$$

- Using $p_n = \lambda/n$, rewrite the p.m.f.:

$$P(X_n = k) \approx \frac{n^k}{k!} \left(\frac{\lambda}{n}\right)^k (1 - \lambda/n)^{n-k}$$

- Taking the limit as $n \to \infty$:

$$\lim_{n\to\infty} (1 - \lambda/n)^{n-k} = e^{-\lambda}$$

- Therefore, the binomial probability simplifies to:

$$P(X_n = k) \to \frac{\lambda^k e^{-\lambda}}{k!}$$

Proof

- Approximate the terms as $n \to \infty$:

$$\frac{n!}{(n-k)!} \approx n^k$$

- Using $p_n = \lambda/n$, rewrite the p.m.f.:

$$P(X_n = k) \approx \frac{n^k}{k!} \left(\frac{\lambda}{n}\right)^k (1 - \lambda/n)^{n-k}$$

- Taking the limit as $n \to \infty$:

$$\lim_{n \to \infty} (1 - \lambda/n)^{n-k} = e^{-\lambda}$$

- Therefore, the binomial probability simplifies to:

$$P(X_n = k) \to \frac{\lambda^k e^{-\lambda}}{k!}$$

- Thus, X_n converges to a Poisson(λ) random variable.

Expectation of a Poisson Distribution

Claim: If $X \sim \text{Pois}(\lambda)$, then its expectation is:

$$E[X] = \lambda.$$

Proof:

- The expectation of X is given by:

$$E[X] = \sum_{k=0}^{\infty} kP(X = k).$$

Expectation of a Poisson Distribution

Claim: If $X \sim \text{Pois}(\lambda)$, then its expectation is:

$$E[X] = \lambda.$$

Proof:

- The expectation of X is given by:

$$E[X] = \sum_{k=0}^{\infty} kP(X = k).$$

- Using the Poisson p.m.f.:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

Expectation of a Poisson Distribution

Claim: If $X \sim \text{Pois}(\lambda)$, then its expectation is:

$$E[X] = \lambda.$$

Proof:

- The expectation of X is given by:

$$E[X] = \sum_{k=0}^{\infty} kP(X = k).$$

- Using the Poisson p.m.f.:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \dots$$

- Substituting:

$$E[X] = \sum_{k=1}^{\infty} k\frac{\lambda^k e^{-\lambda}}{k!}.$$

Expectation of a Poisson Distribution

- Since $k = k!/(k-1)!$, rewrite:
  $E[X] = \sum_{k=1}^{\infty} \frac{\lambda^k e^{-\lambda}}{(k-1)!}$.

Expectation of a Poisson Distribution

- Since $k = k!/(k-1)!$, rewrite: $E[X] = \sum_{k=1}^{\infty} \frac{\lambda^k e^{-\lambda}}{(k-1)!}$.
- Changing the index of summation by setting $j = k - 1$:

$$E[X] = \sum_{j=0}^{\infty} \frac{\lambda^{j+1} e^{-\lambda}}{j!}.$$

Expectation of a Poisson Distribution

- Since $k = k!/(k-1)!$, rewrite: $E[X] = \sum_{k=1}^{\infty} \frac{\lambda^k e^{-\lambda}}{(k-1)!}$.
- Changing the index of summation by setting $j = k - 1$:

$$E[X] = \sum_{j=0}^{\infty} \frac{\lambda^{j+1} e^{-\lambda}}{j!}.$$

- Factor out λ:

$$E[X] = \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!}.$$

Expectation of a Poisson Distribution

- Recognizing the power series for e^λ:

$$\sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = e^\lambda.$$

- Thus, we obtain:

$$E[X] = \lambda e^{-\lambda} \cdot e^\lambda = \lambda.$$

- **Conclusion:** The mean of a Poisson-distributed random variable is λ.