4: Expectation

Fall 2023

Matthew Blackwell

Gov 2002 (Harvard)

Where are we? Where are we going?

- 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 a few numbers.

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

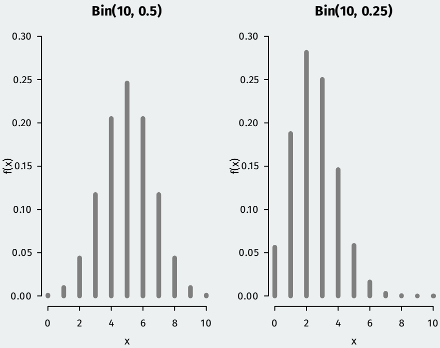
$$\tau_i = Y_i(1) - Y_i(0)$$

- 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?

1/ Definition of Expectation

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.
    - 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: $\{1, 1, 1, 3, 4, 4, 5, 5\}$

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

- 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.
  - $E[X]$ 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

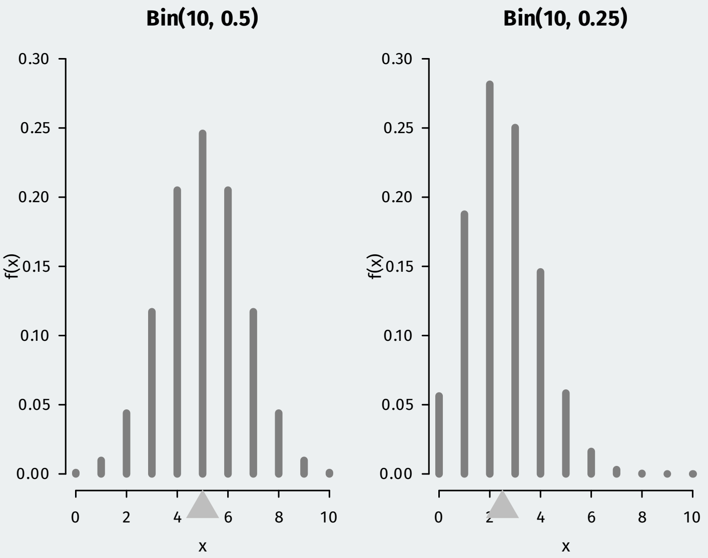
- Randomized experiment with 3 units. X is number of treated units.

x	$p_X(x)$	$xp_X(x)$
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



2/ Linearity of Expectations

Properties of the expected value

- Often want to derive expectation of **transformations** of other r.v.s
- Possible for **linear** functions because expectation is **linear**:

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

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

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 $X_1, \dots, X_n$ be identically distributed with $\mathbb{E}[X_i] = \mu$.
- Define the **sample mean** to be $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$.
    - $\overline{X}$ is a r.v.!
- We can find the expectation of the sample mean using linearity:

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

- 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.
    - $X \sim \text{Bern}(p)$ has $E[X] = p$ 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.
  - Probability of reaching $X = k$ is:

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

- 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) $\rightsquigarrow \mathbb{E}[Y] = 41$
  - Risk avoidance/concave utility $U = Y^{1/2} \rightsquigarrow \mathbb{E}[U(Y)] \approx 2.41$

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.

3/ Indicator Variables

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.

Using indicators to find expectations

- Suppose we are assigning n units to k treatments and all possibilities 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.

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

- Thus, we have $\mathbb{E}\left[\sum_j I_j\right] = k(1 - 1/k)^n$.

4/ 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:

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

- 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?

Defintion

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

- 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.
2. If a is a constant, $\mathbb{V}[aX] = a^2\mathbb{V}[X]$.
3. If X and Y are **independent**, then $V[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y]$.
    - 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?
  - 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:

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

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

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

Variance of the sample mean

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

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

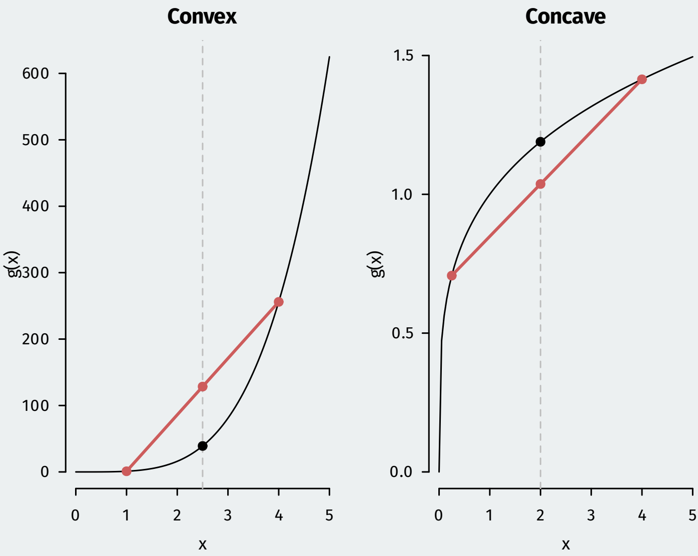
- Note: we needed independence and identically distributed for this.
- $SD(\overline{X}_n) = \sigma/\sqrt{n}$
- Under i.i.d. sampling we know the expectation and variance of $\overline{X}_n$ without any other assumptions about the distribution of the X_i!
    - We don't know what distribution it takes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

5/ Inequalities

Inequalities

- Bounds are very important 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.

Concave and convex



Jensen's inequality

Jensen's inequality

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

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

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:
  - $\mathbb{E}[|X|] \geq |\mathbb{E}[X]|$
  - $\mathbb{E}[1/X] \geq 1/\mathbb{E}[X]$
  - $\mathbb{E}[\log(X)] \leq \log(\mathbb{E}[X])$

6/ Poisson Distribution

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:

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

- 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.: $\sum_{k=0}^{\infty} \lambda^k / k! = e^{\lambda}$.

Poisson properties

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

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

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

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

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