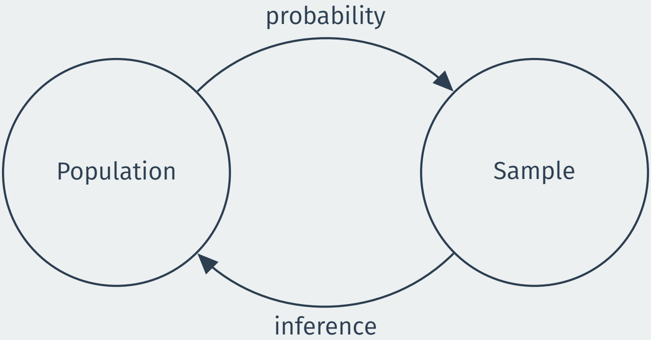
Gov 51: Expectation, Variance, and Sample Means

Matthew Blackwell

Harvard University

Remember our goal



- We want to learn about the chance process that generated our data.
- Last time: entire probability distributions. Is there something simpler?

How can we summarize distributions?

- Two numerical summaries of the distribution are useful.
    1. **Mean/expectaion**: where the center of the distribution is.
    2. **Variance/standard deviation**: how spread out the distribution is around the center.
- 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

- We write $\mathbb{E}(X)$ for the **mean** of an r.v. X.
- For discrete $X \in \{x_1, x_2, \dots, x_k\}$ with k levels:

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

  - Weighted average of the values of the r.v. weighted by the probability of each value occurring.
- If X is age of randomly selected registered voter, then $\mathbb{E}(X)$ is the average age in the population of registered voters.
- Notation notes:
  - Lots of other ways to refer to this: **expectation** or **expected value**
  - Often called the **population mean** to distinguish from the sample mean.

Properties of the expected value

- We use properties of $\mathbb{E}(X)$ to avoid using the formula every time.
- Let X and Y be r.v.s and a and b be constants.

1. $\mathbb{E}(a) = a$
    - Constants don't vary.
2. $\mathbb{E}(aX) = a\mathbb{E}(X)$
    - Suppose X is income in dollars, income in $10k is just: $X/10000$
    - Mean of this new variable is mean of income in dollars divided by 10,000.
3. $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$
    - Expectations can be distributed across sums.
    - X is partner 1's income, Y is partner 2's income.
    - Mean household income is the sum of the each partner's income.

Variance

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

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

- Weighted average of the squared distances from the mean.
  - Larger deviations ($+$ or $-$) $\rightsquigarrow$ higher variance
- If X is the age of a randomly selected registered voter, $\mathbb{V}[X]$ is the usual sample variance of age in the population.
  - Sometimes called **population variance** to contrast with sample variance.
- **Standard deviation**: square root of the variance: $SD(X) = \sqrt{\mathbb{V}[X]}$.
  - Useful because it's on the scale of the original variable.

Properties of variances

- Some properties of variance useful for calculation.

1. If b is a constant, then $\mathbb{V}[b] = 0$.
2. If a and b are constants, $\mathbb{V}[aX + b] = a^2\mathbb{V}[X]$.
3. In general, $\mathbb{V}[X + Y] \neq \mathbb{V}[X] + \mathbb{V}[Y]$.
    - If X and Y are independent, then $\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y]$

Sums and means are random variables

- If X_1 and X_2 are r.v.s, then $X_1 + X_2$ is a r.v.
  - Has a mean $\mathbb{E}[X_1 + X_2]$ and a variance $\mathbb{V}[X_1 + X_2]$
- The **sample mean** is a function of sums and so it is a r.v. too:

$$\overline{X} = \frac{X_1 + X_2}{2}$$

- Example: the average age of two randomly selected respondents.

Distribution of sums/means

	X_1	X_2	$X_1 + X_2$	$\overline{X}$
draw 1	44	32	76	38
draw 2	27	50	77	38.5
draw 3	34	48	82	41
draw 4	68	28	96	48
⋮	⋮	⋮	⋮	⋮
			distribution of the sum	distribution of the mean

Independent and identical r.v.s

- **Independent and identically distributed** r.v.s, $X_1, \dots, X_n$
    - Random sample of n respondents on a survey question.
    - Written "i.i.d."
- **Independent**: value that X_i takes doesn't affect distribution of X_j
- **Identically distributed**: distribution of X_i is the same for all i
    - $\mathbb{E}(X_1) = \mathbb{E}(X_2) = \cdots = \mathbb{E}(X_n) = \mu$
    - $\mathbb{V}(X_1) = \mathbb{V}(X_2) = \cdots = \mathbb{V}(X_n) = \sigma^2$

Distribution of the sample mean

- **Sample mean** of i.i.d. random variables:

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

- $\overline{X}_n$ is a random variable, what is its distribution?
    - What is the expectation of this distribution, $\mathbb{E}[\overline{X}_n]$?
    - What is the variance of this distribution, $\mathbb{V}[\overline{X}_n]$?

Properties of the sample mean

Mean and variance of the sample mean

Suppose that $X_1, \dots, X_n$ are i.i.d. r.v.s with $\mathbb{E}[X_i] = \mu$ and $\mathbb{V}[X_i] = \sigma^2$. Then:

$$\mathbb{E}[\overline{X}_n] = \mu \qquad \mathbb{V}[\overline{X}_n] = \frac{\sigma^2}{n}$$

- Key insights:
  - Sample mean is on average equal to the population mean
  - Variance of $\overline{X}_n$ depends on the population variance of X_i and the sample size
- Standard deviation of the sample mean is called its **standard error**:

$$SE = \sqrt{\mathbb{V}[\overline{X}_n]} = \frac{\sigma}{\sqrt{n}}$$