Methods Camp

Day 2

Instructor: Casey F. Breen TA: Leng Seong Che

Department of Sociology
University of Texas at Austin

August 18, 2026

Welcome

- **Congratulations** on making it to Day 2 of orientation week!
- Today, we will introduce some of the **foundations of statistics**.
- Again, our goal is to help you become comfortable with the basic language of quantitative research that you will see frequently in methods courses and papers.

Warmup Exercise (small group)

- A city has 50,000 residents. Following a hurricane, 40,000 residents remain at the end of the year. Assume there are no births or deaths.
  1. How many residents moved away?
  2. What is the out-migration rate per 1,000 residents?
  3. By what percentage did the city's population decline?

Warmup Exercise: Solutions

- **Number who moved away:** $50{,}000 - 40{,}000 = 10{,}000$
- **Out-migration rate:**

$$\frac{10{,}000}{50{,}000} \times 1{,}000 = 200$$

$$\boxed{\text{200 per 1,000 residents}}$$

- **Percent decline in population:**

$$\frac{40{,}000 - 50{,}000}{50{,}000} \times 100 = -20\%$$

Warmup Exercise (small group)

- A population has 10,000 people at time $t = 0$ and grows at an annual rate of 2%.
  - Assuming **discrete** population growth, what will the population be after 10 years?
  - Assuming **continuous** population growth, what will the population be after 10 years? Hint, use formula: (baseline pop) $\times e^{rt}$ where r $=$ 0.02 and t $=$ 10
  - Why are the two answers different?

Warmup Exercise: Solutions

- **Discrete growth:** $N_{10} = 10{,}000(1.02)^{10} \approx 12{,}190$
- **Continuous growth:** $N(10) = 10{,}000e^{0.02(10)} \approx 12{,}214$
- The continuous-growth population is slightly larger because growth occurs continuously rather than at discrete intervals.

What is Probability?

- **Probability** is a number between 0 and 1 that measures the likelihood of an event occuring.
- $P(A) = 0$: the event A **never** happens.
- $P(A) = 1$: it **always** happens.

Example: Two Coin Tosses

What is the probability of getting exactly one head in 2 tosses of a fair coin?

Example: Two Coin Tosses

- Flip a fair coin twice. Each flip is Heads (H) or Tails (T), equally likely.

1st Coin	2nd Coin	Probability	# of Heads
H	H	$1/4$	2
H	T	$1/4$	1
T	H	$1/4$	1
T	T	$1/4$	0

- All four outcomes are equally likely: $P(\text{outcome}) = 1/4$ each.

Example: Two Coin Tosses

- Flip a fair coin twice. Each flip is Heads (H) or Tails (T), equally likely.

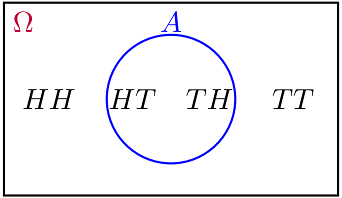
1st Coin	2nd Coin	Probability	# of Heads
H	H	$1/4$	2
H	T	$1/4$	1
T	H	$1/4$	1
T	T	$1/4$	0

- All four outcomes are equally likely: $P(\text{outcome}) = 1/4$ each.
- The event **"exactly one head"** contains *two* outcomes (HT, TH) so we **add** their probabilities:

$$P(\text{exactly one head}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

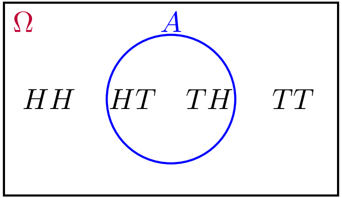
Sample Space and Events

- The **sample space**, Ω, is the set of *all* possible outcomes of a random process.



Sample Space and Events

- The **sample space**, Ω, is the set of *all* possible outcomes of a random process.
- An **event** is a subset of the sample space — one or more outcomes we care about.



Random Variables

What is a Random Variable?

- A **random variable** is a variable whose value is a numerical outcome of some random process.
- It assigns a number to each possible outcome of a **sample space**.

Discrete vs. Continuous Random Variables

Discrete

Takes a countable set of values.

$$0, 1, 2, 3, \ldots$$

- number of children
- number of arrests
- voted: 0 or 1

Continuous

Can take any value in an interval.

$$2.1,\ 2.14,\ 2.141, \ldots$$

- income
- age
- response time

Random Variable Example: Voting

For a binary random variable X:

$$X = \begin{cases} 1 & \text{if a respondent votes} \\ 0 & \text{if a respondent does not vote} \end{cases}$$

Random Variable Example: Voting

For a binary random variable X:

$$X = \begin{cases} 1 & \text{if a respondent votes} \\ 0 & \text{if a respondent does not vote} \end{cases}$$

Suppose:

$$\Pr(X = 1) = 0.70, \qquad \Pr(X = 0) = 0.30$$

Random Variable Example: Voting

For a binary random variable X:

$$X = \begin{cases} 1 & \text{if a respondent votes} \\ 0 & \text{if a respondent does not vote} \end{cases}$$

Suppose:

$$\Pr(X = 1) = 0.70, \qquad \Pr(X = 0) = 0.30$$

- Probabilities range from **0 to 1**.
- Probabilities across all possible outcomes sum to **1**.

Expectations of Random Variables

- The **expectation** (or **expected value**) of a random variable is its long-run average value, weighted by probability.

Expectations of Random Variables

- The **expectation** (or **expected value**) of a random variable is its long-run average value, weighted by probability.
- For a discrete random variable X with possible values $x_1, x_2, \ldots$:

$$E[X] = \sum_i x_i \, P(X = x_i)$$

Expectations of Random Variables

Now going back to the voting example:

$$X = \begin{cases} 1 & \text{if a respondent votes} \\ 0 & \text{if a respondent does not vote} \end{cases}$$

$$\Pr(X = 1) = 0.70, \qquad \Pr(X = 0) = 0.30$$

Expectations of Random Variables

Now going back to the voting example:

$$X = \begin{cases} 1 & \text{if a respondent votes} \\ 0 & \text{if a respondent does not vote} \end{cases}$$

$$\Pr(X = 1) = 0.70, \qquad \Pr(X = 0) = 0.30$$

What is the expected value of X?

Expectations of Random Variables

$$\mathsf{E}[X] = (1)(0.70) + (0)(0.30) = 0.70$$

Properties of Expectation

- Expectation is **linear**: for constants a, b,

$$E[aX + b] = aE[X] + b$$

- Expectation of a sum is the sum of expectations:

$$E[X + Y] = E[X] + E[Y]$$

Exercise 1: Random Variables & Expectation

A survey asks respondents how many times they attended religious services last month. Suppose the distribution is:

Times attended (x)	0	1	2	4
$P(X = x)$	0.4	0.3	0.2	0.1

(a) Is X discrete or continuous? Why?

(b) Compute $E[X]$ by hand.

(c) In your own words, what does $E[X]$ tell us?

(d) If attendance is recoded as $Y = 2X + 1$, what is $E[Y]$? (Use the linearity property.)

Exercise 1: Solutions

(a) X is **discrete**: it only takes on a **countable** set of values $(0, 1, 2, 4)$, not any value in a continuous interval.

(b)

$$E[X] = 0(0.4) + 1(0.3) + 2(0.2) + 4(0.1) = 0 + 0.3 + 0.4 + 0.4 = 1.1$$

(c) $E[X] = 1.1$ is the **long-run average** across *many* respondents — if we surveyed thousands of people and averaged their attendance, we'd land close to 1.1.

(d) Using linearity, $E[aX + b] = aE[X] + b$:

$$E[Y] = E[2X + 1] = 2E[X] + 1 = 2(1.1) + 1 = 3.2$$

Distributions

What does a Distributions Tell?

1. What values can the variable take?

What does a Distributions Tell?

1. What values can the variable take?

2. How likely is each value or range of values?

What does a Distributions Tell?

1. What values can the variable take?
2. How likely is each value or range of values?
3. Where are values centered, and how spread out are they?

What is a Distribution?

- A **probability distribution** describes the possible values of a random variable and their probabilities.
- Two key tools describe a distribution:
  - The **PDF** (probability density/mass function)
  - The **CDF** (cumulative distribution function)

PMF and PDF

Discrete: PMF

The **probability mass function** gives

$$\Pr(X = x).$$

Bars can have positive probability.

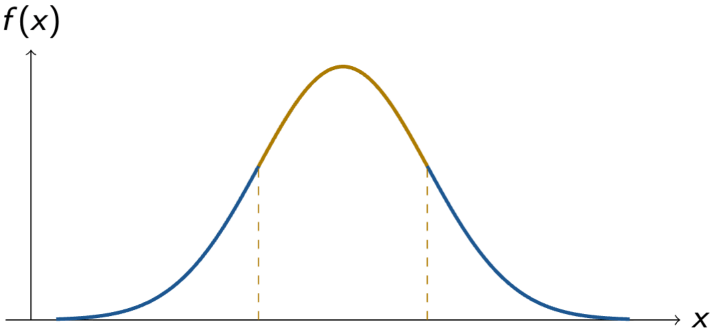
Continuous: PDF

The **probability density function** $f(x)$ describes density.

$$\Pr(a < X < b) = \int_a^b f(x)\, dx$$

Areas represent probabilities.

PDF: Probability is Area

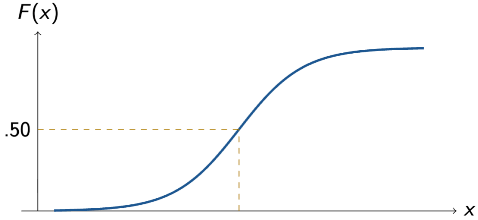


$\Pr(-1 < X < 1)$ is the **area under the gold segment**.

CDF

$$F(x) = \Pr(X \leq x)$$

The **cumulative distribution function** gives the probability of a value at or below x.



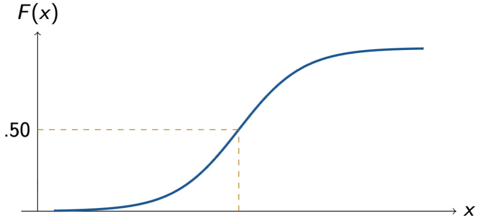
For a symmetric distribution centered at 0, $F(0) = .50$.

CDF

$$F(x) = \Pr(X \leq x)$$

The **cumulative distribution function** gives the probability of a value at or below x.

- $F(x)$ never decreases.
- It begins near 0 and approaches 1.
- It works for both discrete and continuous variables.



For a symmetric distribution centered at 0, $F(0) = .50$.

Mean of a Continuous Random Variable

For a continuous random variable X with probability density function $f(x)$, the mean (or expected value) is

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

- x is a possible value of the random variable.
- $f(x)$ tells us the relative likelihood of values near x.
- The integral is a **weighted average** over all possible values of X.
- Values that are more likely receive more weight.

Aside: Calculating the Mean from Empirical Data

Suppose we observe the number of hours five students studied last week:

$$x_1 = 2, \quad x_2 = 4, \quad x_3 = 5, \quad x_4 = 6, \quad x_5 = 8$$

The sample mean is the sum of all observations divided by the number of observations:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Here, $n = 5$, so:

$$\bar{x} = \frac{2+4+5+6+8}{5} = \frac{25}{5} = 5$$

- The mean amount of time studied is **5 hours**.

Variance and Standard Deviation

- **Variance** measures how spread out a distribution is around its mean:

$$\mathsf{Var}(X) = E\left[(X - E[X])^2\right]$$

Variance and Standard Deviation

- **Variance** measures how spread out a distribution is around its mean:

$$\text{Var}(X) = E\big[(X - E[X])^2\big]$$

- **Standard deviation** is just the square root of variance:

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

Mean and Variance Describe Different Features

- The **mean** describes the **center** of a distribution.
- The **variance** and **standard deviation** describe how widely values are **spread around the mean**.

Mean and Variance Describe Different Features

- The **mean** describes the **center** of a distribution.
- The **variance** and **standard deviation** describe how widely values are **spread around the mean**.

Two distributions can have the same mean but very different amounts of variation.

Common Distributions: Bernoulli

The **Bernoulli distribution** models a single yes/no (0/1) event.

Common Distributions: Bernoulli

The **Bernoulli distribution** models a single yes/no (0/1) event.

$$X \sim \text{Bernoulli}(p)$$

$$\Pr(X = 1) = p, \qquad \Pr(X = 0) = 1 - p$$

Common Distributions: Bernoulli

The **Bernoulli distribution** models a single yes/no (0/1) event.

$$X \sim \text{Bernoulli}(p)$$

$$\Pr(X = 1) = p, \qquad \Pr(X = 0) = 1 - p$$

Examples: employed/not employed; voted/did not vote

Mean: $\mathsf{E}[X] = p$

Variance: $\mathsf{V}(X) = p(1 - p)$

Common Distributions: Binomial

- The **Binomial distribution** counts the number of "successes" across n **independent** Bernoulli trials, each with probability p.

Common Distributions: Binomial

- The **Binomial distribution** counts the number of "successes" across n **independent** Bernoulli trials, each with probability p.
- Example: number of voters among 20 independently sampled eligible residents.

Common Distributions: Binomial

$$X \sim \text{Binomial}(n, p)$$

- n: number of independent trials
- p: probability of success on each trial
- X: number of successes

Common Distributions: Binomial

$$X \sim \text{Binomial}(n, p)$$

- n: number of independent trials
- p: probability of success on each trial
- X: number of successes

$$\mathsf{E}[X] = np, \qquad \mathsf{V}(X) = np(1-p)$$

Common Distributions: Poisson

- The **Poisson distribution** models **counts of rare events** over a fixed interval (time, space, etc.).
- Example: number of protests a city experiences in a year.

Common Distributions: Poisson

- The **Poisson distribution** models **counts of rare events** over a fixed interval (time, space, etc.).
- Example: number of protests a city experiences in a year.
- $X \sim \text{Poisson}(\lambda)$, where λ is the average rate.

$$\mathsf{E}[X] = \lambda, \qquad \mathsf{V}(X) = \lambda$$

Common Distributions: Normal

- The **Normal (Gaussian) distribution** is the familiar bell curve, used constantly for **continuous** variables.
- $X \sim N(\mu, \sigma^2)$, fully described by two parameters:
  - $\mu =$ mean
  - $\sigma^2 =$ variance

Common Distributions: Normal

- The **Normal (Gaussian) distribution** is the familiar bell curve, used constantly for **continuous** variables.
- $X \sim N(\mu, \sigma^2)$, fully described by two parameters:
  - $\mu =$ mean
  - $\sigma^2 =$ variance
- Many statistical tests **assume** normality.

Exercise 2: Distributions

(a) For each scenario below, name the **most appropriate distribution** (Bernoulli, Binomial, Poisson, or Normal) and briefly justify your choice:

i. Whether a single respondent owns a home (yes/no)

ii. The number of home-owners among 200 randomly sampled respondents

iii. The number of hate-crime incidents reported in a city per month

iv. Adult height in a large population

(b) For scenario (ii), if $p = 0.65$, what is $E[X]$?

Exercise 2: Solutions

(a)

i. **Bernoulli** — a single yes/no trial for one respondent.

ii. **Binomial** — the count of "successes" (home-owners) across $n = 200$ independent Bernoulli trials, each with the same probability p.

iii. **Poisson** — a count of relatively rare events occurring over a fixed interval (a month), without a natural fixed number of "trials."

iv. **Normal** — a continuous measurement that tends to cluster symmetrically around an average, the classic bell-curve shape.

(b) Using $E[X] = np$ for a Binomial:

$$E[X] = 200 \times 0.65 = 130$$

Regression

Unit of Observation

- The **unit of observation** is the entity that each observation (or row) in a dataset represents.

Unit of Observation

- The **unit of observation** is the entity that each observation (or row) in a dataset represents.
- Common units of observation include:
  - **Individuals**: one row per person
  - **Households**: one row per household
  - **Countries**: one row per country
  - **Country-years**: one row per country in each year

Unit of Observation

- The **unit of observation** is the entity that each observation (or row) in a dataset represents.
- Common units of observation include:
  - **Individuals**: one row per person
  - **Households**: one row per household
  - **Countries**: one row per country
  - **Country-years**: one row per country in each year
- We often use the subscript i to index observations:

$$X_i = \text{value of } X \text{ for observation } i.$$

Why Regression is the Core Tool

- Regression is how sociologists ask: **"How does Y change as X changes, holding other things constant?"**
- Many empirical papers you read will use some form of regression.

The Basic Linear Regression Model

A simple regression model relates an outcome Y to a predictor X:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Y_i: outcome for observation i (person, country, year, etc)

The Basic Linear Regression Model

A simple regression model relates an outcome Y to a predictor X:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Y_i: outcome for observation i (person, country, year, etc)

X_i: predictor for observation i]pause

β_0: intercept

The Basic Linear Regression Model

A simple regression model relates an outcome Y to a predictor X:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Y_i: outcome for observation i (person, country, year, etc)

X_i: predictor for observation i]pause

β_0: intercept

β_1: slope

The Basic Linear Regression Model

A simple regression model relates an outcome Y to a predictor X:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Y_i: outcome for observation i (person, country, year, etc)

X_i: predictor for observation i]pause

β_0: intercept

β_1: slope

ε_i: error term for observation i

The Basic Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

β_0 and β_1 are **fixed for the whole population**.

Y_i, X_i, ε_i vary by observation.

The Basic Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

The regression line describes the **expected** value of Y at each value of X.

The Basic Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

The regression line describes the **expected** value of Y at each value of X.

Individual observations can lie **above or below** that line.

Interpreting the intercept

$$\beta_0 = \mathsf{E}[Y \mid X = 0]$$

Interpreting the intercept

$$\beta_0 = \mathsf{E}[Y \mid X = 0]$$

Example:

$$\widehat{\mathsf{wage}} = 8 + 1.5(\mathsf{education})$$

$$\hat{\beta}_0 = 8$$

is the predicted wage when education equals 0.

Interpreting the intercept

$$\beta_0 = \mathsf{E}[Y \mid X = 0]$$

Example:

$$\widehat{\text{wage}} = 8 + 1.5(\text{education})$$

$$\hat{\beta}_0 = 8$$

is the predicted wage when education equals 0.

The intercept may not be substantively meaningful if $X = 0$ is impossible or far outside the data.

Interpreting the slope

$\beta_1 =$ change in expected Y for a one-unit increase in X

Interpreting the slope

$$\beta_1 = \text{change in expected } Y \text{ for a one-unit increase in } X$$

For:

$$\widehat{\text{wage}} = 8 + 1.5(\text{education}),$$

one additional year of education is associated with a **$1.50 increase in predicted hourly wage**, on average.

Error Terms

The **error term** ε_i captures everything that affects Y_i that is **not** explained by X_i.

It may reflect:

- Omitted factors (things we didn't measure)
- Measurement error

Error Term vs Residual

Error ε_i	**Residual** e_i
Part of the population model	Calculated after fitting a sample model
Unobserved	Observed from the fitted model
Distance from the true line	Distance from the estimated line

Model Example: OLS

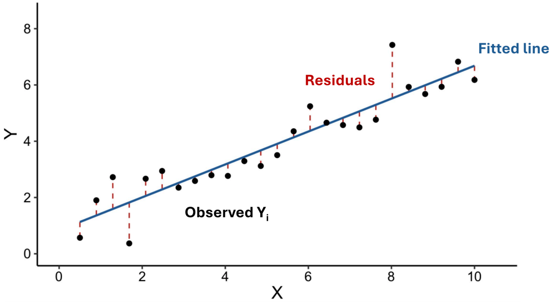
- Many possible lines could be drawn through the data.

Model Example: OLS

- Many possible lines could be drawn through the data.
- Each possible line produces different **predicted values** and **residuals**.

Model Example: OLS

- Many possible lines could be drawn through the data.
- Each possible line produces different **predicted values** and **residuals**.
- **Ordinary least squares (OLS)** selects the line with the smallest sum of squared residuals.



Estimates, Estimators, and Estimands: Why the Distinction Matters

- **Estimand**: the **quantity of interest**—the true, unknown population parameter, β_1.

Estimates, Estimators, and Estimands: Why the Distinction Matters

- **Estimand**: the **quantity of interest**—the true, unknown population parameter, β_1.
- **Estimator**: the **procedure or formula** used to estimate the estimand from data, such as OLS.

Estimates, Estimators, and Estimands: Why the Distinction Matters

- **Estimand**: the **quantity of interest**—the true, unknown population parameter, β_1.
- **Estimator**: the **procedure or formula** used to estimate the estimand from data, such as OLS.
- **Estimate**: the **numerical result** obtained by applying the estimator to a specific dataset, $\hat{\beta}_1 = 2{,}300$.

Exercise 3: Regression

Suppose a researcher wants to examine occupational prestige score (Y) as a linear function of their parents' years of schooling (X) across all U.S. families.

(a) Write the regression equation using index notation. What does the subscript i represent?

(b) Interpret the slope coefficient.

(c) Identify the **estimand**, a plausible **estimator**, and an example of an **estimate**.

(d) Name one plausible factor that could be included in the **error term**, ε_i.

Exercise 3: Solutions

(a)

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

The subscript i identifies a particular adult child or family.

(b) β_1 represents the expected difference in the child's occupational prestige score associated with one additional year of parental schooling.

(c)

- **Estimand**: the population slope β_1
- **Estimator**: ordinary least squares (OLS)
- **Estimate**: the value obtained from the sample, such as $\hat{\beta}_1 = 1.5$

(d) Parental income, neighborhood conditions, or the local labor market.

Hands-on Practice

For each paper,

- What is the unit of observation?
- What is the (main) dependent variable?
- What is the (main) independent variable?
- What is the estimand? What model do they use?

Papers:

1. *The Dynamics of Mass Migration. Massey and Zenteno, 1999.*
2. *The Long Arm of Childhood: The Influence of Early-Life Social Conditions on Men's Mortality. Hayward and Gorman, 2004.*
3. *The Political Economy of Incarceration in the Cotton South, 1910–1925. Muller and Schrage, 2021.*