

# Random Variables and Probability Distributions

From Data to Distributions: What Will You Earn?

Jake Anderson

# Outline

- 1 Motivation
- 2 Random Variables
- 3 Discrete Random Variables
- 4 Continuous Random Variables
- 5 Normal Distributions
- 6 Population vs. Sample
- 7 Looking Ahead

# Outline

- 1 Motivation
- 2 Random Variables
- 3 Discrete Random Variables
- 4 Continuous Random Variables
- 5 Normal Distributions
- 6 Population vs. Sample
- 7 Looking Ahead

# A Question for Every Senior

You're about to graduate from UCLA with an economics degree.

**What will your starting salary be?**

# A Question for Every Senior

You're about to graduate from UCLA with an economics degree.

**What will your starting salary be?**

You can't know the exact number. But you *can* say something useful:

- “Probably between \$45,000 and \$75,000”
- “Most likely around \$55,000”
- “Very unlikely to be below \$30,000”

# A Question for Every Senior

You're about to graduate from UCLA with an economics degree.

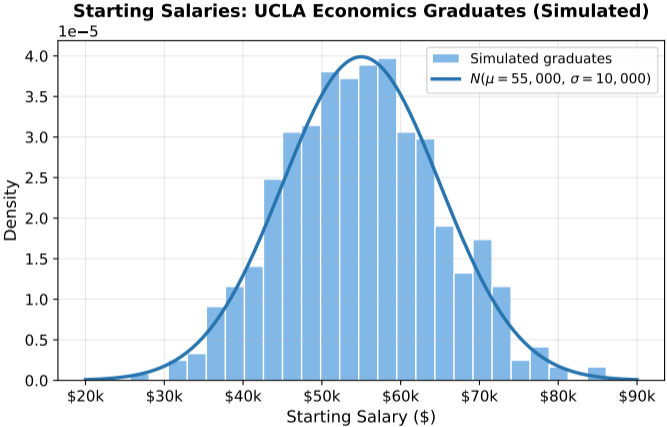
**What will your starting salary be?**

You can't know the exact number. But you *can* say something useful:

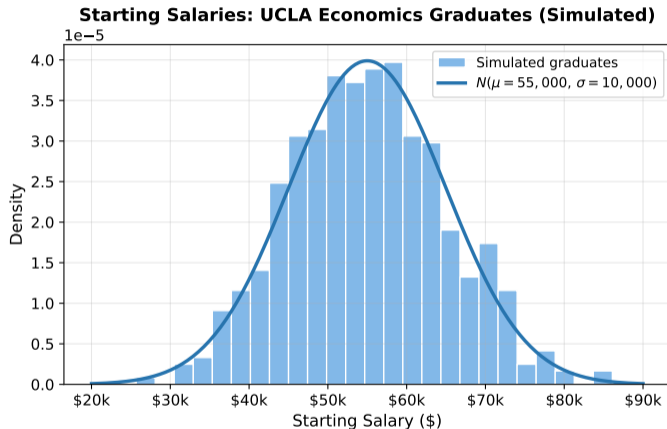
- “Probably between \$45,000 and \$75,000”
- “Most likely around \$55,000”
- “Very unlikely to be below \$30,000”

⇒ You're already thinking in terms of a **probability distribution**.

# Starting Salaries: UCLA Economics Graduates



# Starting Salaries: UCLA Economics Graduates



Each bar shows the fraction of graduates earning in that range. The shape of this histogram approximates the **probability distribution** of starting wages.

# Why Econometrics Cares About Distributions

Econometrics is about understanding **relationships between economic variables**:

- How does an extra year of education affect wages?
- Does using my phone less improve my mood?
- What is the effect of working out on mental and physical health?

# Why Econometrics Cares About Distributions

Econometrics is about understanding **relationships between economic variables**:

- How does an extra year of education affect wages?
- Does using my phone less improve my mood?
- What is the effect of working out on mental and physical health?

All of these variables are **uncertain** before we observe data.

# Why Econometrics Cares About Distributions

Econometrics is about understanding **relationships between economic variables**:

- How does an extra year of education affect wages?
- Does using my phone less improve my mood?
- What is the effect of working out on mental and physical health?

All of these variables are **uncertain** before we observe data.

⇒ To build a statistical model, we need a language for describing uncertainty. That language is **random variables** and **probability distributions**.

# Why Econometrics Cares About Distributions

Econometrics is about understanding **relationships between economic variables**:

- How does an extra year of education affect wages?
- Does using my phone less improve my mood?
- What is the effect of working out on mental and physical health?

All of these variables are **uncertain** before we observe data.

⇒ To build a statistical model, we need a language for describing uncertainty. That language is **random variables** and **probability distributions**.

Today: review the foundations you'll use in every chapter of this course.

# Outline

- 1 Motivation
- 2 Random Variables**
- 3 Discrete Random Variables
- 4 Continuous Random Variables
- 5 Normal Distributions
- 6 Population vs. Sample
- 7 Looking Ahead

# What Is a Random Variable?

A **random variable** is a variable whose value is unknown until it is observed.

# What Is a Random Variable?

A **random variable** is a variable whose value is unknown until it is observed.

Notation convention:

- **Uppercase**  $X, Y$ : the random variable (uncertain)
- **Lowercase**  $x, y$ : a specific realized value

# What Is a Random Variable?

A **random variable** is a variable whose value is unknown until it is observed.

Notation convention:

- **Uppercase**  $X, Y$ : the random variable (uncertain)
- **Lowercase**  $x, y$ : a specific realized value

Before you check your first paycheck: Wage is a random variable.

After you see it:  $wage = \$58,000$  is a realization.

# What Is a Random Variable?

A **random variable** is a variable whose value is unknown until it is observed.

Notation convention:

- **Uppercase**  $X, Y$ : the random variable (uncertain)
- **Lowercase**  $x, y$ : a specific realized value

Before you check your first paycheck: Wage is a random variable.

After you see it:  $wage = \$58,000$  is a realization.

Two types:

- **Discrete**: countable number of values
- **Continuous**: any value in an interval

# Discrete Random Variables

A **discrete** random variable takes a countable number of values.

## Economic examples:

- Number of job offers a graduate receives:  $X \in \{0, 1, 2, 3, \dots\}$
- Number of cars owned by a household:  $X \in \{0, 1, 2, 3, \dots\}$
- Number of children in a family:  $X \in \{0, 1, 2, \dots\}$

A **discrete** random variable takes a countable number of values.

**Economic examples:**

- Number of job offers a graduate receives:  $X \in \{0, 1, 2, 3, \dots\}$
- Number of cars owned by a household:  $X \in \{0, 1, 2, 3, \dots\}$
- Number of children in a family:  $X \in \{0, 1, 2, \dots\}$

A special case: **indicator (dummy) variables** take only values 0 or 1.

- $D = 1$  if the person is a college graduate,  $D = 0$  otherwise
- $D = 1$  if the firm exports,  $D = 0$  otherwise

A **discrete** random variable takes a countable number of values.

**Economic examples:**

- Number of job offers a graduate receives:  $X \in \{0, 1, 2, 3, \dots\}$
- Number of cars owned by a household:  $X \in \{0, 1, 2, 3, \dots\}$
- Number of children in a family:  $X \in \{0, 1, 2, \dots\}$

A special case: **indicator (dummy) variables** take only values 0 or 1.

- $D = 1$  if the person is a college graduate,  $D = 0$  otherwise
- $D = 1$  if the firm exports,  $D = 0$  otherwise

⇒ You can always list the possible values (even if the list is long).

# Continuous Random Variables

A **continuous** random variable can take any value in an interval.

## Economic examples:

- Starting salary:  $Wage \in (0, \infty)$
- Stock return:  $R \in (-1, \infty)$
- GDP growth rate:  $g \in (-\infty, \infty)$
- Household expenditure:  $Exp \in (0, \infty)$

# Continuous Random Variables

A **continuous** random variable can take any value in an interval.

## Economic examples:

- Starting salary:  $Wage \in (0, \infty)$
- Stock return:  $R \in (-1, \infty)$
- GDP growth rate:  $g \in (-\infty, \infty)$
- Household expenditure:  $Exp \in (0, \infty)$

The distinction from discrete:

- Discrete: “How many?” (countable values)
- Continuous: “How much?” (any value in a range)

# Continuous Random Variables

A **continuous** random variable can take any value in an interval.

## Economic examples:

- Starting salary:  $Wage \in (0, \infty)$
- Stock return:  $R \in (-1, \infty)$
- GDP growth rate:  $g \in (-\infty, \infty)$
- Household expenditure:  $Exp \in (0, \infty)$

The distinction from discrete:

- Discrete: “How many?” (countable values)
- Continuous: “How much?” (any value in a range)

⇒ Most economic variables we study in econometrics (wages, prices, GDP, returns) are treated as continuous.

# Outline

- 1 Motivation
- 2 Random Variables
- 3 Discrete Random Variables**
- 4 Continuous Random Variables
- 5 Normal Distributions
- 6 Population vs. Sample
- 7 Looking Ahead

# Probability Mass Function (pmf): Discrete Case

The **probability mass function (pmf)** gives the probability of each possible value:

$$f(x) = P(X = x)$$

# Probability Mass Function (pmf): Discrete Case

The **probability mass function (pmf)** gives the probability of each possible value:

$$f(x) = P(X = x)$$

Two properties every pmf must satisfy:

- 1  $0 \leq f(x) \leq 1$  for all  $x$
- 2 Probabilities sum to one:  $\sum_{\text{all } x} f(x) = 1$

# Probability Mass Function (pmf): Discrete Case

The **probability mass function (pmf)** gives the probability of each possible value:

$$f(x) = P(X = x)$$

Two properties every pmf must satisfy:

- 1  $0 \leq f(x) \leq 1$  for all  $x$
- 2 Probabilities sum to one:  $\sum_{\text{all } x} f(x) = 1$

**Example:** Suppose the number of job offers  $X$  for an econ graduate has pmf:

$x$	0	1	2	3
$f(x) = P(X = x)$	0.10	0.30	0.40	0.20

# Probability Mass Function (pmf): Discrete Case

The **probability mass function (pmf)** gives the probability of each possible value:

$$f(x) = P(X = x)$$

Two properties every pmf must satisfy:

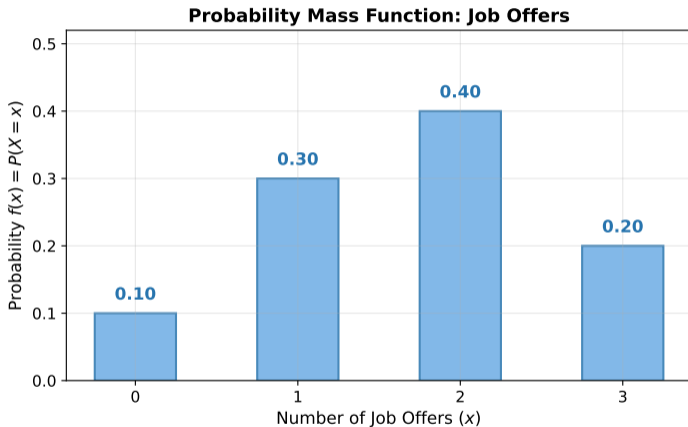
- 1  $0 \leq f(x) \leq 1$  for all  $x$
- 2 Probabilities sum to one:  $\sum_{\text{all } x} f(x) = 1$

**Example:** Suppose the number of job offers  $X$  for an econ graduate has pmf:

$x$	0	1	2	3
$f(x) = P(X = x)$	0.10	0.30	0.40	0.20

Check:  $0.10 + 0.30 + 0.40 + 0.20 = 1 \checkmark$

# Visualizing a Discrete Distribution



The height of each bar is the probability  $P(X = x)$ . You can read probabilities directly from the graph.

# Cumulative Distribution Function (CDF)

The **cumulative distribution function (cdf)** gives the probability that  $X$  is *at most* a given value:

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

# Cumulative Distribution Function (CDF)

The **cumulative distribution function (cdf)** gives the probability that  $X$  is *at most* a given value:

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

For our job-offers example:

$x$	0	1	2	3
$f(x) = P(X = x)$	0.10	0.30	0.40	0.20
$F(x) = P(X \leq x)$	0.10	0.40	0.80	1.00

# Cumulative Distribution Function (CDF)

The **cumulative distribution function (cdf)** gives the probability that  $X$  is *at most* a given value:

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

For our job-offers example:

$x$	0	1	2	3
$f(x) = P(X = x)$	0.10	0.30	0.40	0.20
$F(x) = P(X \leq x)$	0.10	0.40	0.80	1.00

Reading the cdf:

- $P(X \leq 1) = F(1) = 0.40$ : a 40% chance of getting at most 1 offer
- $P(X > 1) = 1 - F(1) = 0.60$ : a 60% chance of getting more than 1 offer

# Cumulative Distribution Function (CDF)

The **cumulative distribution function (cdf)** gives the probability that  $X$  is *at most* a given value:

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

For our job-offers example:

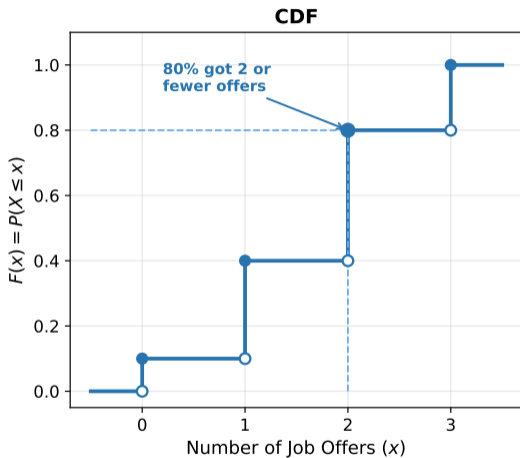
$x$	0	1	2	3
$f(x) = P(X = x)$	0.10	0.30	0.40	0.20
$F(x) = P(X \leq x)$	0.10	0.40	0.80	1.00

Reading the cdf:

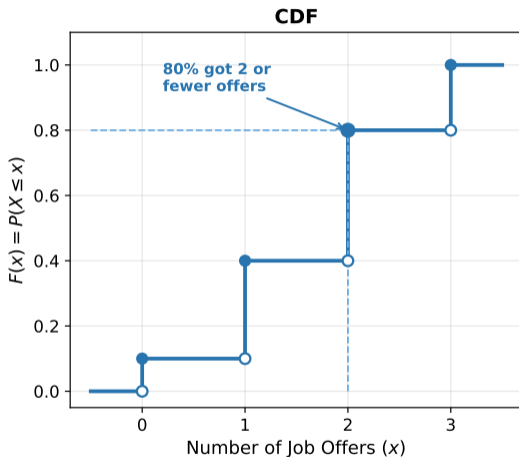
- $P(X \leq 1) = F(1) = 0.40$ : a 40% chance of getting at most 1 offer
- $P(X > 1) = 1 - F(1) = 0.60$ : a 60% chance of getting more than 1 offer

$\implies$  The cdf is non-decreasing and reaches 1 at the largest possible value.

# PMF vs. CDF: Two Views of the Same Distribution



# PMF vs. CDF: Two Views of the Same Distribution



The PMF tells you the probability of *each* value. The CDF tells you the probability of getting *that value or less*. Same distribution, different questions.

# Outline

- 1 Motivation
- 2 Random Variables
- 3 Discrete Random Variables
- 4 Continuous Random Variables**
- 5 Normal Distributions
- 6 Population vs. Sample
- 7 Looking Ahead

## From Discrete to Continuous: What Breaks?

For discrete variables, we compute  $P(X = x)$  directly from the pmf. Can we do the same for continuous variables?

# From Discrete to Continuous: What Breaks?

For discrete variables, we compute  $P(X = x)$  directly from the pmf. Can we do the same for continuous variables?

**Try it:** What is  $P(\text{Wage} = \text{exactly } \$58,000.00)$ ?

# From Discrete to Continuous: What Breaks?

For discrete variables, we compute  $P(X = x)$  directly from the pmf. Can we do the same for continuous variables?

**Try it:** What is  $P(\text{Wage} = \text{exactly } \$58,000.00)$ ?

Think about what “exactly” means: not \$58,000.01, not \$57,999.99, but \$58,000.000...

There are uncountably many possible values in any interval. If each one got positive probability, the total would be infinite.

## From Discrete to Continuous: What Breaks?

For discrete variables, we compute  $P(X = x)$  directly from the pmf. Can we do the same for continuous variables?

**Try it:** What is  $P(\text{Wage} = \text{exactly } \$58,000.00)$ ?

Think about what “exactly” means: not \$58,000.01, not \$57,999.99, but \$58,000.000...

There are uncountably many possible values in any interval. If each one got positive probability, the total would be infinite.

$\implies P(X = x) = 0$  for every single value  $x$ .

## From Discrete to Continuous: What Breaks?

For discrete variables, we compute  $P(X = x)$  directly from the pmf. Can we do the same for continuous variables?

**Try it:** What is  $P(\text{Wage} = \text{exactly } \$58,000.00)$ ?

Think about what “exactly” means: not \$58,000.01, not \$57,999.99, but \$58,000.000...

There are uncountably many possible values in any interval. If each one got positive probability, the total would be infinite.

$\implies P(X = x) = 0$  for every single value  $x$ .

Just as a single point has no length on a number line, a single value gets no probability from a continuous distribution.

## From Discrete to Continuous: What Breaks?

For discrete variables, we compute  $P(X = x)$  directly from the pmf. Can we do the same for continuous variables?

**Try it:** What is  $P(\text{Wage} = \text{exactly } \$58,000.00)$ ?

Think about what “exactly” means: not \$58,000.01, not \$57,999.99, but \$58,000.000...

There are uncountably many possible values in any interval. If each one got positive probability, the total would be infinite.

$\implies P(X = x) = 0$  for every single value  $x$ .

Just as a single point has no length on a number line, a single value gets no probability from a continuous distribution.

$\implies$  We need a different tool: probabilities as **areas under a curve**.

# The Probability Density Function (pdf): Continuous Case

For continuous random variables, probabilities are **areas under the pdf curve**:

$$P(a < X < b) = \text{area under } f(x) \text{ between } a \text{ and } b$$

# The Probability Density Function (pdf): Continuous Case

For continuous random variables, probabilities are **areas under the pdf curve**:

$$P(a < X < b) = \text{area under } f(x) \text{ between } a \text{ and } b$$

The pdf satisfies:

- 1  $f(x) \geq 0$  for all  $x$
- 2 Total area under the curve = 1

# The Probability Density Function (pdf): Continuous Case

For continuous random variables, probabilities are **areas under the pdf curve**:

$$P(a < X < b) = \text{area under } f(x) \text{ between } a \text{ and } b$$

The pdf satisfies:

- 1  $f(x) \geq 0$  for all  $x$
- 2 Total area under the curve = 1

Notice:  $f(x)$  is *not* a probability. It is a **density**. Only the area under  $f(x)$  over an interval gives a probability.

# The Probability Density Function (pdf): Continuous Case

For continuous random variables, probabilities are **areas under the pdf curve**:

$$P(a < X < b) = \text{area under } f(x) \text{ between } a \text{ and } b$$

The pdf satisfies:

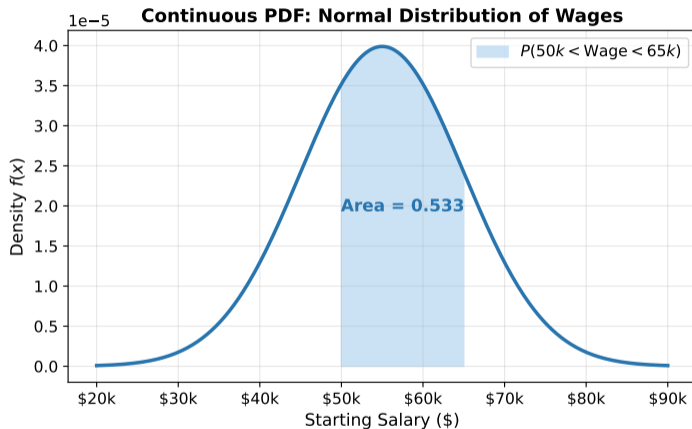
- 1  $f(x) \geq 0$  for all  $x$
- 2 Total area under the curve = 1

Notice:  $f(x)$  is *not* a probability. It is a **density**. Only the area under  $f(x)$  over an interval gives a probability.

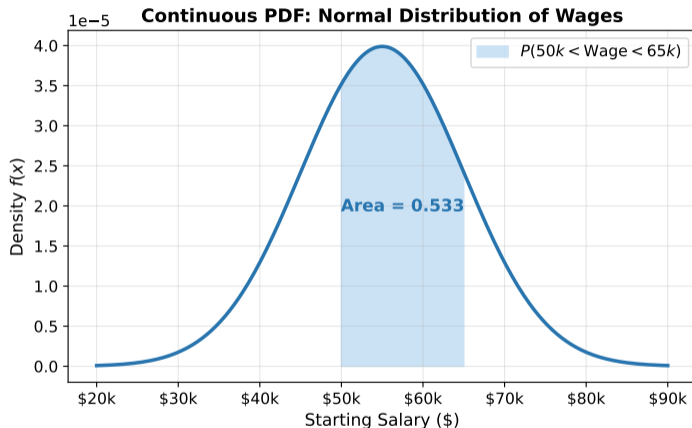
For continuous  $X$ : since  $P(X = c) = 0$ , strict vs. weak inequalities give the same answer:

$$P(a < X < b) = P(a \leq X \leq b)$$

# A Continuous PDF: Starting Wages



# A Continuous PDF: Starting Wages



The shaded area represents  $P(50,000 < Wage < 65,000)$ . This is the probability that a randomly chosen graduate earns between \$50,000 and \$65,000.

# The CDF for Continuous Variables

The cdf is defined the same way as for discrete variables:

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

# The CDF for Continuous Variables

The cdf is defined the same way as for discrete variables:

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

Probabilities over intervals come from the cdf:

$$P(a < X < b) = F(b) - F(a)$$

# The CDF for Continuous Variables

The cdf is defined the same way as for discrete variables:

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

Probabilities over intervals come from the cdf:

$$P(a < X < b) = F(b) - F(a)$$

**Example:** If  $\text{Wage} \sim N(55,000, 10,000^2)$ , then  $F(65,000) = 0.84$  and  $F(50,000) = 0.31$ :

$$P(50,000 < \text{Wage} < 65,000) = 0.84 - 0.31 = 0.53$$

# The CDF for Continuous Variables

The cdf is defined the same way as for discrete variables:

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

Probabilities over intervals come from the cdf:

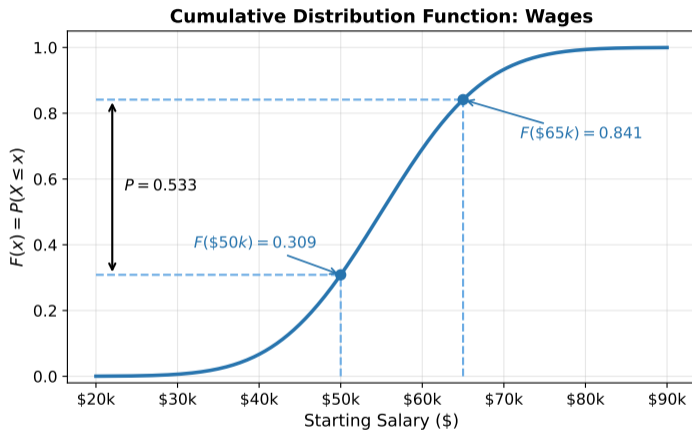
$$P(a < X < b) = F(b) - F(a)$$

**Example:** If  $\text{Wage} \sim N(55,000, 10,000^2)$ , then  $F(65,000) = 0.84$  and  $F(50,000) = 0.31$ :

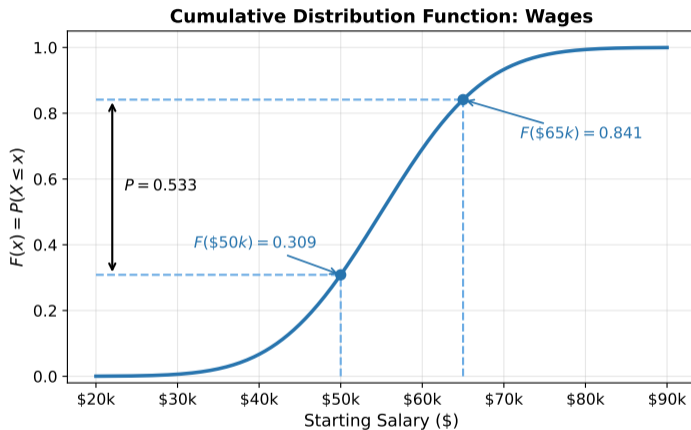
$$P(50,000 < \text{Wage} < 65,000) = 0.84 - 0.31 = 0.53$$

$\implies$  A 53% chance of earning between \$50,000 and \$65,000.

# Visualizing the CDF



# Visualizing the CDF



The CDF starts at 0 (for very low wages) and approaches 1 (for very high wages). It is always non-decreasing. The probability of an interval is the vertical distance between two points on this curve.

# Properties of Every CDF

Whether  $X$  is discrete or continuous, the cdf  $F(x) = P(X \leq x)$  always satisfies:

- 1  $0 \leq F(x) \leq 1$  for all  $x$

# Properties of Every CDF

Whether  $X$  is discrete or continuous, the cdf  $F(x) = P(X \leq x)$  always satisfies:

- 1  $0 \leq F(x) \leq 1$  for all  $x$
- 2  $F(x)$  is **non-decreasing**: if  $a < b$ , then  $F(a) \leq F(b)$

# Properties of Every CDF

Whether  $X$  is discrete or continuous, the cdf  $F(x) = P(X \leq x)$  always satisfies:

- 1  $0 \leq F(x) \leq 1$  for all  $x$
- 2  $F(x)$  is **non-decreasing**: if  $a < b$ , then  $F(a) \leq F(b)$
- 3  $F(x) \rightarrow 0$  as  $x \rightarrow -\infty$  and  $F(x) \rightarrow 1$  as  $x \rightarrow +\infty$

# Properties of Every CDF

Whether  $X$  is discrete or continuous, the cdf  $F(x) = P(X \leq x)$  always satisfies:

- 1  $0 \leq F(x) \leq 1$  for all  $x$
- 2  $F(x)$  is **non-decreasing**: if  $a < b$ , then  $F(a) \leq F(b)$
- 3  $F(x) \rightarrow 0$  as  $x \rightarrow -\infty$  and  $F(x) \rightarrow 1$  as  $x \rightarrow +\infty$
- 4  $P(X > a) = 1 - F(a)$  (complement rule)

# Properties of Every CDF

Whether  $X$  is discrete or continuous, the cdf  $F(x) = P(X \leq x)$  always satisfies:

- 1  $0 \leq F(x) \leq 1$  for all  $x$
- 2  $F(x)$  is **non-decreasing**: if  $a < b$ , then  $F(a) \leq F(b)$
- 3  $F(x) \rightarrow 0$  as  $x \rightarrow -\infty$  and  $F(x) \rightarrow 1$  as  $x \rightarrow +\infty$
- 4  $P(X > a) = 1 - F(a)$  (complement rule)
- 5  $P(a < X \leq b) = F(b) - F(a)$  (interval probability)

# Properties of Every CDF

Whether  $X$  is discrete or continuous, the cdf  $F(x) = P(X \leq x)$  always satisfies:

- 1  $0 \leq F(x) \leq 1$  for all  $x$
- 2  $F(x)$  is **non-decreasing**: if  $a < b$ , then  $F(a) \leq F(b)$
- 3  $F(x) \rightarrow 0$  as  $x \rightarrow -\infty$  and  $F(x) \rightarrow 1$  as  $x \rightarrow +\infty$
- 4  $P(X > a) = 1 - F(a)$  (complement rule)
- 5  $P(a < X \leq b) = F(b) - F(a)$  (interval probability)

$\implies$  The CDF is a complete description of the distribution. If you know  $F(x)$ , you can compute any probability involving  $X$ .

# Discrete vs. Continuous: Side by Side

	<b>Discrete</b>	<b>Continuous</b>
Question	“How many?”	“How much?”
Probabilities from $P(X = x)$	pmf: $f(x) = P(X = x)$	pdf: area under $f(x)$
Probabilities sum/integrate to	Can be $> 0$	Always $= 0$
CDF	$\sum f(x) = 1$	$\int f(x) dx = 1$
Economic examples	Step function	Smooth curve
	Job offers, children	Wages, GDP, returns

# Discrete vs. Continuous: Side by Side

	Discrete	Continuous
Question	“How many?”	“How much?”
Probabilities from $P(X = x)$	pmf: $f(x) = P(X = x)$ Can be $> 0$	pdf: area under $f(x)$ Always $= 0$
Probabilities sum/integrate to	$\sum f(x) = 1$	$\int f(x) dx = 1$
CDF	Step function	Smooth curve
Economic examples	Job offers, children	Wages, GDP, returns

⇒ The CDF works identically in both cases:  $F(x) = P(X \leq x)$ .

# Outline

- 1 Motivation
- 2 Random Variables
- 3 Discrete Random Variables
- 4 Continuous Random Variables
- 5 Normal Distributions**
- 6 Population vs. Sample
- 7 Looking Ahead

# Why the Normal Distribution?

Three reasons this distribution dominates econometrics:

# Why the Normal Distribution?

Three reasons this distribution dominates econometrics:

## 1. Central Limit Theorem

Sample averages are approximately normal for large samples, under mild conditions. Since many econometric estimators are averages (or functions of averages), normality shows up everywhere.

# Why the Normal Distribution?

Three reasons this distribution dominates econometrics:

## 1. Central Limit Theorem

Sample averages are approximately normal for large samples, under mild conditions. Since many econometric estimators are averages (or functions of averages), normality shows up everywhere.

## 2. Convenient mathematics

Sums of *independent* normal random variables are normal. This makes deriving estimator distributions convenient.

# Why the Normal Distribution?

Three reasons this distribution dominates econometrics:

## 1. Central Limit Theorem

Sample averages are approximately normal for large samples, under mild conditions. Since many econometric estimators are averages (or functions of averages), normality shows up everywhere.

## 2. Convenient mathematics

Sums of *independent* normal random variables are normal. This makes deriving estimator distributions convenient.

## 3. Empirical regularity

Many economic variables (log wages, measurement errors, test scores) have approximately bell-shaped distributions.

# The Normal Distribution: Definition

If  $X \sim N(\mu, \sigma^2)$ , then  $X$  is normally distributed with:

- **Mean**  $\mu$ : center of the distribution
- **Variance**  $\sigma^2$ : spread of the distribution

# The Normal Distribution: Definition

If  $X \sim N(\mu, \sigma^2)$ , then  $X$  is normally distributed with:

- **Mean**  $\mu$ : center of the distribution
- **Variance**  $\sigma^2$ : spread of the distribution

The pdf:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right], \quad -\infty < x < \infty$$

# The Normal Distribution: Definition

If  $X \sim N(\mu, \sigma^2)$ , then  $X$  is normally distributed with:

- **Mean**  $\mu$ : center of the distribution
- **Variance**  $\sigma^2$ : spread of the distribution

The pdf:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right], \quad -\infty < x < \infty$$

You will never need to evaluate this formula by hand. Software and tables handle it. What you need to remember: a normal distribution is fully determined by  $\mu$  and  $\sigma^2$ .

# The Normal Distribution: Definition

If  $X \sim N(\mu, \sigma^2)$ , then  $X$  is normally distributed with:

- **Mean**  $\mu$ : center of the distribution
- **Variance**  $\sigma^2$ : spread of the distribution

The pdf:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right], \quad -\infty < x < \infty$$

You will never need to evaluate this formula by hand. Software and tables handle it. What you need to remember: a normal distribution is fully determined by  $\mu$  and  $\sigma^2$ .

Properties:

- **Symmetric** and **bell-shaped** around  $\mu$
- Larger  $\sigma^2 \implies$  wider, flatter bell
- Smaller  $\sigma^2 \implies$  taller, narrower bell

# Standardization: The Z-Score

Any normal variable can be converted to a **standard normal**  $Z \sim N(0, 1)$ :

$$Z = \frac{X - \mu}{\sigma}$$

# Standardization: The Z-Score

Any normal variable can be converted to a **standard normal**  $Z \sim N(0, 1)$ :

$$Z = \frac{X - \mu}{\sigma}$$

This does two things:

- 1 **Centers** the variable at zero (subtracts the mean)
- 2 **Scales** to unit variance (divides by the standard deviation)

# Standardization: The Z-Score

Any normal variable can be converted to a **standard normal**  $Z \sim N(0, 1)$ :

$$Z = \frac{X - \mu}{\sigma}$$

This does two things:

- 1 **Centers** the variable at zero (subtracts the mean)
- 2 **Scales** to unit variance (divides by the standard deviation)

**Interpretation:**  $Z$  tells you how many standard deviations  $X$  is from  $\mu$ .

- $Z = 1.5 \implies$  the value is 1.5 standard deviations above the mean
- $Z = -2 \implies$  the value is 2 standard deviations below the mean

# Standardization: The Z-Score

Any normal variable can be converted to a **standard normal**  $Z \sim N(0, 1)$ :

$$Z = \frac{X - \mu}{\sigma}$$

This does two things:

- 1 **Centers** the variable at zero (subtracts the mean)
- 2 **Scales** to unit variance (divides by the standard deviation)

**Interpretation:**  $Z$  tells you how many standard deviations  $X$  is from  $\mu$ .

- $Z = 1.5 \implies$  the value is 1.5 standard deviations above the mean
- $Z = -2 \implies$  the value is 2 standard deviations below the mean

$\implies$  You only need one table (or one function in your software) to handle *every* normal distribution.

# Computing Normal Probabilities with $\Phi$

Define  $\Phi(z) = P(Z \leq z)$ : the cdf of the standard normal distribution.

# Computing Normal Probabilities with $\Phi$

Define  $\Phi(z) = P(Z \leq z)$ : the cdf of the standard normal distribution.

For any  $X \sim N(\mu, \sigma^2)$ , standardize and use  $\Phi$ :

$$P(X \leq a) = \Phi\left(\frac{a - \mu}{\sigma}\right)$$

$$P(a \leq X \leq b) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)$$

## Computing Normal Probabilities with $\Phi$

Define  $\Phi(z) = P(Z \leq z)$ : the cdf of the standard normal distribution.

For any  $X \sim N(\mu, \sigma^2)$ , standardize and use  $\Phi$ :

$$P(X \leq a) = \Phi\left(\frac{a - \mu}{\sigma}\right)$$

$$P(a \leq X \leq b) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)$$

Values of  $\Phi$  come from:

- A standard normal table (Appendix in your textbook)
- Software: `pnorm(z)` in R, `stats.norm.cdf(z)` in Python

## Computing Normal Probabilities with $\Phi$

Define  $\Phi(z) = P(Z \leq z)$ : the cdf of the standard normal distribution.

For any  $X \sim N(\mu, \sigma^2)$ , standardize and use  $\Phi$ :

$$P(X \leq a) = \Phi\left(\frac{a - \mu}{\sigma}\right)$$

$$P(a \leq X \leq b) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)$$

Values of  $\Phi$  come from:

- A standard normal table (Appendix in your textbook)
- Software: `pnorm(z)` in R, `stats.norm.cdf(z)` in Python

$\implies$  Every normal probability reduces to: standardize, then look up  $\Phi$ .

## Example: Computing a Normal Probability

Suppose starting wages follow  $\text{Wage} \sim N(55,000, 10,000^2)$ , so  $\mu = 55,000$  and  $\sigma = 10,000$ .

## Example: Computing a Normal Probability

Suppose starting wages follow  $\text{Wage} \sim N(55,000, 10,000^2)$ , so  $\mu = 55,000$  and  $\sigma = 10,000$ .

**What fraction of graduates earn between \$45,000 and \$70,000?**

## Example: Computing a Normal Probability

Suppose starting wages follow  $\text{Wage} \sim N(55,000, 10,000^2)$ , so  $\mu = 55,000$  and  $\sigma = 10,000$ .

**What fraction of graduates earn between \$45,000 and \$70,000?**

Standardize both endpoints:

$$z_1 = \frac{45,000 - 55,000}{10,000} = -1.0$$

$$z_2 = \frac{70,000 - 55,000}{10,000} = 1.5$$

## Example: Computing a Normal Probability

Suppose starting wages follow  $\text{Wage} \sim N(55,000, 10,000^2)$ , so  $\mu = 55,000$  and  $\sigma = 10,000$ .

**What fraction of graduates earn between \$45,000 and \$70,000?**

Standardize both endpoints:

$$z_1 = \frac{45,000 - 55,000}{10,000} = -1.0$$

$$z_2 = \frac{70,000 - 55,000}{10,000} = 1.5$$

Look up (or compute):

$$P(45,000 \leq \text{Wage} \leq 70,000) = \Phi(1.5) - \Phi(-1.0) = 0.9332 - 0.1587 = 0.7745$$

## Example: Computing a Normal Probability

Suppose starting wages follow  $\text{Wage} \sim N(55,000, 10,000^2)$ , so  $\mu = 55,000$  and  $\sigma = 10,000$ .

**What fraction of graduates earn between \$45,000 and \$70,000?**

Standardize both endpoints:

$$z_1 = \frac{45,000 - 55,000}{10,000} = -1.0$$

$$z_2 = \frac{70,000 - 55,000}{10,000} = 1.5$$

Look up (or compute):

$$P(45,000 \leq \text{Wage} \leq 70,000) = \Phi(1.5) - \Phi(-1.0) = 0.9332 - 0.1587 = 0.7745$$

$\implies$  About 77% of graduates earn between \$45,000 and \$70,000.

# Outline

- 1 Motivation
- 2 Random Variables
- 3 Discrete Random Variables
- 4 Continuous Random Variables
- 5 Normal Distributions
- 6 Population vs. Sample**
- 7 Looking Ahead

# Population vs. Sample

A probability distribution describes a **population**: the complete set of all possible values.

# Population vs. Sample

A probability distribution describes a **population**: the complete set of all possible values.

- The distribution of *all* possible starting salaries for econ graduates is the **population distribution**
- The salaries of the 200 graduates we surveyed are a **sample**

# Population vs. Sample

A probability distribution describes a **population**: the complete set of all possible values.

- The distribution of *all* possible starting salaries for econ graduates is the **population distribution**
- The salaries of the 200 graduates we surveyed are a **sample**

**Notation distinction:**

	<b>Population</b>	<b>Sample</b>
Mean	$\mu = E(X)$	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
Variance	$\sigma^2 = \text{Var}(X)$	$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

# Population vs. Sample

A probability distribution describes a **population**: the complete set of all possible values.

- The distribution of *all* possible starting salaries for econ graduates is the **population distribution**
- The salaries of the 200 graduates we surveyed are a **sample**

**Notation distinction:**

	Population	Sample
Mean	$\mu = E(X)$	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
Variance	$\sigma^2 = \text{Var}(X)$	$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

Greek letters ( $\mu, \sigma^2$ ) are **parameters**: fixed but unknown.

Roman letters ( $\bar{x}, s^2$ ) are **statistics**: computed from data, vary from sample to sample.

# The Sample Mean Is a Random Variable

We surveyed 200 graduates and computed  $\bar{x} = \$56,200$ .

# The Sample Mean Is a Random Variable

We surveyed 200 graduates and computed  $\bar{x} = \$56,200$ .

If we surveyed a *different* 200 graduates, we would get a different  $\bar{x}$ .

# The Sample Mean Is a Random Variable

We surveyed 200 graduates and computed  $\bar{x} = \$56,200$ .

If we surveyed a *different* 200 graduates, we would get a different  $\bar{x}$ .

⇒ Before we collect the data, the sample mean  $\bar{X}$  is itself a **random variable** with its own distribution.

# The Sample Mean Is a Random Variable

We surveyed 200 graduates and computed  $\bar{x} = \$56,200$ .

If we surveyed a *different* 200 graduates, we would get a different  $\bar{x}$ .

⇒ Before we collect the data, the sample mean  $\bar{X}$  is itself a **random variable** with its own distribution.

This idea is the foundation of statistical inference:

- How far is  $\bar{X}$  likely to be from  $\mu$ ?
- What is the distribution of  $\bar{X}$ ?
- How does increasing  $n$  affect the spread of  $\bar{X}$ ?

# The Sample Mean Is a Random Variable

We surveyed 200 graduates and computed  $\bar{x} = \$56,200$ .

If we surveyed a *different* 200 graduates, we would get a different  $\bar{x}$ .

⇒ Before we collect the data, the sample mean  $\bar{X}$  is itself a **random variable** with its own distribution.

This idea is the foundation of statistical inference:

- How far is  $\bar{X}$  likely to be from  $\mu$ ?
- What is the distribution of  $\bar{X}$ ?
- How does increasing  $n$  affect the spread of  $\bar{X}$ ?

⇒ These are the questions of Topic 3: expected values, variance, and the sampling distribution.

# Outline

- 1 Motivation
- 2 Random Variables
- 3 Discrete Random Variables
- 4 Continuous Random Variables
- 5 Normal Distributions
- 6 Population vs. Sample
- 7 Looking Ahead**

Today we reviewed the building blocks:

- Random variables: discrete vs. continuous
- Discrete pmf vs. continuous pdf: probabilities from bars vs. areas under curves
- CDFs: cumulative probabilities,  $F(x) = P(X \leq x)$
- The normal distribution and standardization
- Population parameters vs. sample statistics

Today we reviewed the building blocks:

- Random variables: discrete vs. continuous
- Discrete pmf vs. continuous pdf: probabilities from bars vs. areas under curves
- CDFs: cumulative probabilities,  $F(x) = P(X \leq x)$
- The normal distribution and standardization
- Population parameters vs. sample statistics

Next: **expected values, variance, and conditional expectation.**

Today we reviewed the building blocks:

- Random variables: discrete vs. continuous
- Discrete pmf vs. continuous pdf: probabilities from bars vs. areas under curves
- CDFs: cumulative probabilities,  $F(x) = P(X \leq x)$
- The normal distribution and standardization
- Population parameters vs. sample statistics

Next: **expected values, variance, and conditional expectation.**

The conditional expectation

$$E(\text{Wage} \mid \text{Educ} = 16)$$

asks: “What is the average wage among people with 16 years of education?”

Today we reviewed the building blocks:

- Random variables: discrete vs. continuous
- Discrete pmf vs. continuous pdf: probabilities from bars vs. areas under curves
- CDFs: cumulative probabilities,  $F(x) = P(X \leq x)$
- The normal distribution and standardization
- Population parameters vs. sample statistics

Next: **expected values, variance, and conditional expectation.**

The conditional expectation

$$E(\text{Wage} \mid \text{Educ} = 16)$$

asks: “What is the average wage among people with 16 years of education?”

⇒ That conditional mean is the regression function. Every model in this course is built on this idea.

Thank you!  
jakeanderson@g.ucla.edu