Econ 103: Introduction to Econometrics

Lecture 2 — Random Variables & Distributions

Ryan Longmuir

UCLA

Summer Session C, 2026

Reading: Hill, Griffiths & Lim (5th ed.), *Probability Primer*, §P.1–P.2.

Where we are

Last time (Lecture 1). What econometrics is; correlation vs. causation; a first look at data.

The next three lectures build the probability toolkit we need to do inference:

- Lecture 2 (today): random variables, distributions, pmf / pdf / cdf.
- Lecture 3: expectation, variance, covariance.
- Lecture 4: the Normal distribution, sampling, and the CLT.

Why bother?

A dataset is a *sample* drawn from a larger *population*. To learn about the population from the sample, we first need a language for *uncertainty*. That language is the random variable.

Today's plan

1. Random variables
2. Discrete vs. continuous
3. Discrete distributions: the pmf
4. Continuous distributions: the pdf
5. The cdf

Part 1

Random variables

What is a random variable?

Definition

A **random variable** is a variable whose value is unknown until it is observed — i.e. a numerical outcome that is not perfectly predictable.

Everyday examples:

- the score you will get on the next exam,
- tomorrow's value of a stock-market index,
- the number of games the football team wins next season,
- the wage of a randomly selected worker.

Convention. We write random variables with uppercase letters (X, Y, W) and the particular values they take with lowercase letters (x, y, w). So "$X = x$" reads: *the random variable X takes the value x.*

Why economists care

Think of the **population** of California adults. Pick one person at random and record their *education level*.

- The outcome is *not* deterministic — different people have different education.
- So education is a random variable: its **distribution** tells us the probability a randomly drawn person falls in each category, e.g. $\mathbb{P}(\text{bachelor's degree}) \approx 0.225$.

What is a "probability"?

The **probability** of an outcome is its long-run relative frequency. "$\mathbb{P}(\text{bachelor's}) \approx 0.225$" means that across many random draws, about 22.5% of those drawn hold a bachelor's degree.

The econometric problem in one sentence

We rarely know the true distribution. Econometrics uses a random *sample* to make inferences about the underlying distribution.

Part 2

Discrete vs. continuous

Outcome spaces: discrete vs. continuous

Every random variable comes with an **outcome space** $\mathcal{O}_X$: the set of all values it can take.

Discrete

$\mathcal{O}_X$ is countable (think: a list, possibly infinite).

- Coin flip: $\{H, T\}$
- Die roll: $\{1, 2, 3, 4, 5, 6\}$
- # of doctor visits: $\{0, 1, 2, \dots\}$

Continuous

$\mathcal{O}_X$ is uncountable: a whole interval of values.

- Sprint time (s): $[9.5, 10.5]$
- Income: $[0, \infty)$
- Interest rate, GDP, ...

Example (Indicator variables)

A yes/no answer ("college graduate?") is a special discrete variable taking only $\{0, 1\}$*. We will use these constantly for qualitative traits.*

Part 3

Discrete distributions: the pmf

Describing a discrete distribution

For a discrete random variable, the distribution is captured by the **probability mass function** (p.m.f.):

$$f_X(x) \;=\; \mathbb{P}(X = x).$$

The pmf assigns, to each possible value x, the probability that X equals exactly that value.

Property (Two rules every pmf obeys)

$$\textit{(1)} \quad 0 \le f_X(x) \le 1 \qquad\qquad \textit{(2)} \quad \sum_{x \in \mathcal{O}_X} f_X(x) = 1.$$

To get the probability of a *set* of outcomes A, just add up the masses:

$$\mathbb{P}(X \in A) = \sum_{x \in A} f_X(x).$$

Example: a fair die

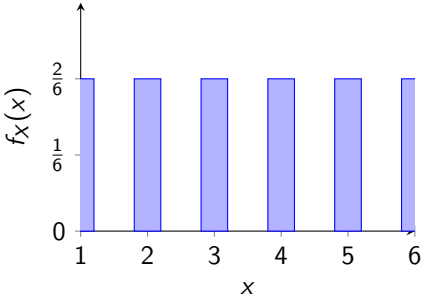
Let X be the result of a fair die roll. Its pmf is

$$f_X(x) = \begin{cases} \frac{1}{6} & x \in \{1,2,3,4,5,6\} \\ 0 & \text{otherwise.} \end{cases}$$

Probability of an even roll, $A = \{2,4,6\}$:

$$\begin{aligned} \mathbb{P}(X \in \{2,4,6\}) &= f_X(2) + f_X(4) + f_X(6) \\ &= \tfrac{1}{6} + \tfrac{1}{6} + \tfrac{1}{6} = \tfrac{1}{2}. \end{aligned}$$

Obvious here — but the *procedure* is what matters. With a loaded die we would follow exactly the same steps.



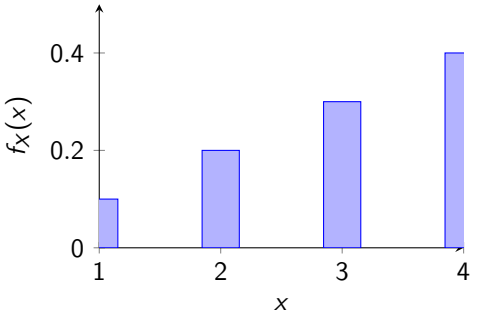
Each bar's *height* is a probability.

The pmf as a table

A discrete distribution is often easiest to read as a table. Consider X with

$$f_X(1) = 0.1, \quad f_X(2) = 0.2, \quad f_X(3) = 0.3, \quad f_X(4) = 0.4.$$

x	$f_X(x)$
1	0.1
2	0.2
3	0.3
4	0.4
sum	**1.0**



The probabilities are non-negative and sum to one — a valid pmf. We return to this X when we build its cdf.

A special case: the indicator (Bernoulli) variable

The most important discrete variable in this course takes only two values, 0 and 1. It is called an **indicator** (or **dummy**, or **Bernoulli**) variable, and it encodes a yes/no trait.

Let $D = 1$ if a randomly drawn person is a college graduate and $D = 0$ if not. With $p = \mathbb{P}(D = 1)$, its pmf is

$$f_D(d) = \begin{cases} p & d = 1 \\ 1 - p & d = 0 \\ 0 & \text{otherwise,} \end{cases}$$

the **Bernoulli**(p) distribution. A single number, p, says everything.

Why we care

Indicators encode qualitative traits — sex, race, treatment status, whether a policy is in place.

Example (A preview)

The mean of a 0/1 variable is just the proportion of ones: $\mathbb{E}[D] = p$. We show this next lecture — it is why regressions on indicators recover group shares and treatment effects (L19–L20).

Part 4

Continuous distributions: the pdf

Describing a continuous distribution

For a continuous random variable we *cannot* use a pmf. Why?

The key fact

A continuous variable can take *uncountably* many values, so the probability of any *single* exact value is zero:

$$\mathbb{P}(X = x) = 0 \quad \text{for every } x.$$

Instead we describe the distribution with a **probability density function** (p.d.f.), $f_X(x)$. Probabilities are areas under the density:

$$\mathbb{P}(a \leq X \leq b) \;=\; \int_a^b f_X(x)\, dx.$$

Notation note: following HGL we write f_X for *both* the discrete pmf and the continuous pdf. Same symbol, different meaning — for a discrete variable $f_X(x)$ *is* a probability, while for a continuous variable it is a *density* (only its *area* is a probability).

Density is not probability

A density $f_X(x)$ can exceed 1 — it is *not* a probability. Only the area under it is.

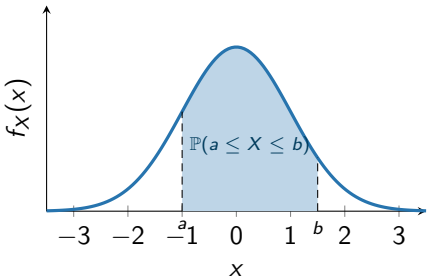
Property (What makes f_X a valid pdf)

$$f_X(x) \geq 0 \qquad \textit{and} \qquad \int_{-\infty}^{\infty} f_X(x)\,dx = 1.$$

The total area under any density is one — the continuous analog of "the masses sum to one."

Because single points carry zero probability, endpoints don't matter:

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a < X < b).$$



Probability = shaded *area*.

Example: the Uniform$[0, 1]$ distribution

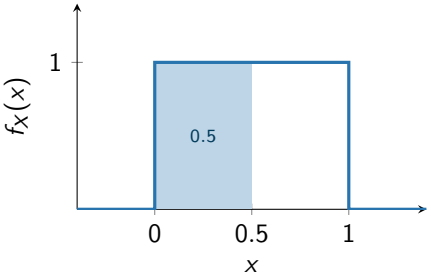
Let X be **uniform** on $[0, 1]$, with density

$$f_X(x) = \begin{cases} 1 & 0 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

What is $\mathbb{P}(0 \le X \le 0.5)$?

$$\begin{aligned} \mathbb{P}(0 \le X \le 0.5) &= \int_0^{0.5} f_X(x)\, dx \\ &= \int_0^{0.5} 1\, dx \\ &= 0.5 - 0 = 0.5. \end{aligned}$$

The area is a rectangle: width $0.5\times$ height $1 = 0.5$ — half the probability sits in the left half, exactly what "uniform" means.



Area of the shaded rectangle $= 0.5$.

Part 5

The cdf

The cumulative distribution function

Both discrete and continuous variables share one common summary: the **cumulative distribution function** (c.d.f.),

$$F_X(x) = \mathbb{P}(X \leq x).$$

It accumulates probability from $-\infty$ up to x.

- Discrete: $F_X(x) = \sum_{t \leq x} f_X(t)$
- Continuous: $F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$

Property (Properties of any cdf)

- F_X *is non-decreasing, with* $\lim_{x \to -\infty} F_X(x) = 0$ *and* $\lim_{x \to +\infty} F_X(x) = 1$.
- $0 \leq F_X(x) \leq 1$.

Why the cdf is so useful

The cdf turns "probability of an interval" into simple subtraction.

The interval rule

$$\mathbb{P}(a < X \leq b) \;=\; F_X(b) - F_X(a).$$

And the complement rule:

$$\mathbb{P}(X > a) \;=\; 1 - F_X(a).$$

This is exactly how we will read probabilities off statistical tables and software later in the course (e.g. Normal and t probabilities):

$$\mathbb{P}(a < X < b) = F_X(b) - F_X(a).$$

We almost never integrate by hand — we look up or compute cdf values.

cdf of a discrete variable: a step function

Take our table from before: $f_X(1{:}4) = (0.1, 0.2, 0.3, 0.4)$. Accumulating,

$$F_X(1) = 0.1, \quad F_X(2) = 0.3, \quad F_X(3) = 0.6, \quad F_X(4) = 1.0.$$

Example (Reading the cdf)

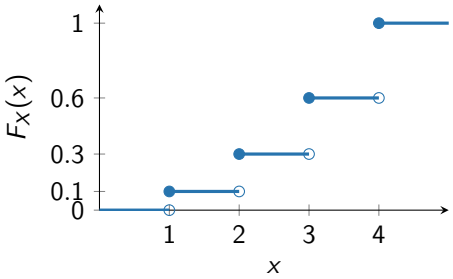
$$\mathbb{P}(X \leq 2) = F_X(2) = 0.1 + 0.2 = 0.3.$$

Even a value X can't take has a cdf:

$$F_X(2.5) = \mathbb{P}(X \leq 2.5) = 0.3.$$

Complement: $\mathbb{P}(X > 2) = 1 - F_X(2) = 0.7.$

The discrete cdf jumps at each possible value.



Jump size at each x equals $f_X(x)$.

cdf of a continuous variable: a smooth curve

For the Uniform$[0, 1]$, accumulate the area from the left:

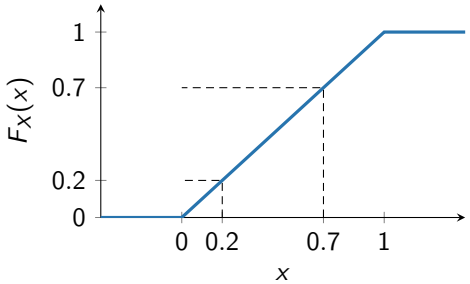
$$F_X(x) = \begin{cases} 0 & x < 0 \\ x & 0 \le x \le 1 \\ 1 & x > 1. \end{cases}$$

Check the interval rule:

$$\begin{aligned} \mathbb{P}(0.2 < X \le 0.7) &= F_X(0.7) - F_X(0.2) \\ &= 0.7 - 0.2 = 0.5. \end{aligned}$$

A continuous cdf is continuous (no jumps):

single points carry no probability, so there is nothing to jump by. Its slope is the density, $F'_X(x) = f_X(x)$.



Rises from 0 to 1 over $[0, 1]$.

Recap

Random variable

A numerical outcome that is unknown until observed; described by its *distribution*.

Discrete (countable $\mathcal{O}_X$)

- pmf: $f_X(x) = \mathbb{P}(X = x)$
- $\sum_x f_X(x) = 1$
- $\mathbb{P}(X \in A) = \sum_{x \in A} f_X(x)$

Continuous (interval $\mathcal{O}_X$)

- pdf: area gives probability
- $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$
- $\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\, dx$

cdf — the common language

$F_X(x) = \mathbb{P}(X \leq x)$, and $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$.

Next time: summarizing a distribution with a single number — expectation, then variance and covariance.

Questions?