Random Variables

From Outcomes to Numbers

Scott Cunningham

Harvard University
Department of Government

February 3, 2026

Where Are We?

Last week: Probability foundations

- Sample spaces, events, axioms
- Conditional probability, Bayes' Rule
- Independence of events

Today: Random variables

- Moving from events to *numbers*
- PMFs, PDFs, CDFs
- Joint distributions and independence of random variables

Wednesday: Famous distributions (Bernoulli, Binomial, Normal, Poisson)

Random variables are how we actually *do* statistics.

Random Variables

Turning outcomes into numbers

The Problem with Events

Events like "roll a six" or "candidate wins" are useful, but limited.

We want to work with *numbers*:

- What's the *average* income in a population?
- How much does vote share *vary* across districts?
- What's the *expected* number of protests per year?

To answer these questions, we need to convert outcomes into numbers.

That's what **random variables** do.

Random Variables

The intuition

Think of a **random variable** as a **container** or **placeholder** for a quantity that has yet to be determined by a random process.

Example: "The number showing when I roll this die."

- We don't know what it will be yet
- We know what values it *could* take (1, 2, 3, 4, 5, 6)
- We know how likely each value is

The random variable gives us a way to talk about uncertain quantities **before** we observe them.

Random Variables

The formal definition

Mathematically, a **random variable** is a function from outcomes to numbers:

$$X : \Omega \longrightarrow \mathbb{R}$$

Example: Roll a die. Define X = "the number showing."

- Sample space: $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$
- Random variable: $X(\omega_i) = i$

Key insight: Despite the name, a random variable is **neither random nor a variable**—it's a function. The randomness comes from which ω nature selects.

Random Variables

More examples

Flip two coins: $\Omega = \{HH, HT, TH, TT\}$

- X = number of heads: $X(HH) = 2$, $X(HT) = X(TH) = 1$, $X(TT) = 0$

Survey a voter: $\Omega = \{\text{all possible voters}\}$

- X = age of selected voter
- Y = 1 if Democrat, 0 otherwise
- Z = feeling thermometer toward Biden (0–100)

Key insight: Many random variables can be defined on the same sample space.

The sample space is about *what happens*. Random variables are about *what we measure*.

Notation Convention

Capital letters for random variables: X, Y, Z

Lowercase letters for specific values: x, y, z

Example:

- "X" = the random variable (a function)
- "$X = 3$" = the event $\{\omega \in \Omega : X(\omega) = 3\}$
- "$x = 3$" = a specific number

We write:

$$\mathbb{P}(X = x) \quad \text{or} \quad \mathbb{P}(X \leq x)$$

This is shorthand for "the probability of the event where X takes value x."

Two Types of Random Variables

Discrete: Takes on a finite or countably infinite set of values.

- Number of votes, count of protests, party ID (coded 1, 2, 3)
- We use **probability mass functions** (PMFs)

Continuous: Can take any value in an interval.

- Income, vote share, feeling thermometer
- We use **probability density functions** (PDFs)

The distinction matters for how we compute probabilities.

Functions vs. Operators

A preview of what's coming

Functions of random variables produce new random variables:

- If X is income, then $Y = \log(X)$ is also a random variable
- $g(X) = X^2$ transforms X into another uncertain quantity

Operators on random variables produce numbers:

- $\mathbb{E}[X]$ (expected value) $\longrightarrow$ a single number
- $\text{Var}(X)$ (variance) $\longrightarrow$ a single number

We'll spend next week on operators like $\mathbb{E}[\cdot]$ and $\text{Var}(\cdot)$.

For now, just note the distinction: $g(X)$ is still random; $\mathbb{E}[X]$ is not.

Part II

Discrete Random Variables

Probability mass functions

Probability Mass Function (PMF)

Definition

For a discrete random variable X, the **PMF** is:

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

The PMF tells us the probability that X takes each possible value.

Properties:

1. $f_X(x) \geq 0$ for all x
2. $\sum_x f_X(x) = 1$ (probabilities sum to 1)

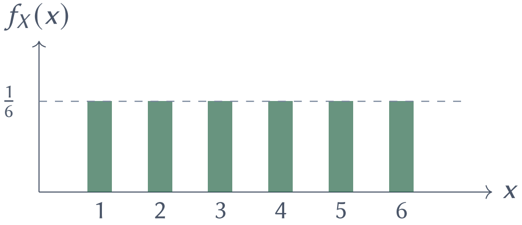
The PMF completely describes the distribution of a discrete random variable.

Example: Fair Die

Let X = result of rolling a fair die.

PMF:

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



Example: Number of Heads in Two Flips

Flip a fair coin twice. Let X = number of heads.

Sample space: $\{HH, HT, TH, TT\}$, each with probability $\frac{1}{4}$.

PMF:

- $f_X(0) = \mathbb{P}(X = 0) = \mathbb{P}(\{TT\}) = \frac{1}{4}$
- $f_X(1) = \mathbb{P}(X = 1) = \mathbb{P}(\{HT, TH\}) = \frac{2}{4} = \frac{1}{2}$
- $f_X(2) = \mathbb{P}(X = 2) = \mathbb{P}(\{HH\}) = \frac{1}{4}$

Check: $\frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1$ ✓

This is a **Binomial(2, 0.5)** distribution—more on Wednesday.

Computing Probabilities from PMFs

Once we have the PMF, we can compute any probability:

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

Example: For a fair die, what's $\mathbb{P}(X \leq 3)$?

$$\mathbb{P}(X \leq 3) = f_X(1) + f_X(2) + f_X(3) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

Example: What's $\mathbb{P}(X \text{ is even})$?

$$\mathbb{P}(X \in \{2, 4, 6\}) = f_X(2) + f_X(4) + f_X(6) = \frac{3}{6} = \frac{1}{2}$$

Continuous Random Variables

Probability density functions

The Problem with Continuous Variables

For continuous random variables, $\mathbb{P}(X = x) = 0$ for any specific x.

Why? Uncountably many possible values. If each had positive probability, they'd sum to more than 1.

Example: What's the probability someone's height is *exactly* 5.7832941... feet?

Zero. But we can ask: What's the probability their height is *between* 5.5 and 6 feet?

For continuous variables, we only assign probabilities to **intervals**.

Probability Density Function (PDF)

Definition

For a continuous random variable X, the **PDF** $f_X(x)$ satisfies:

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

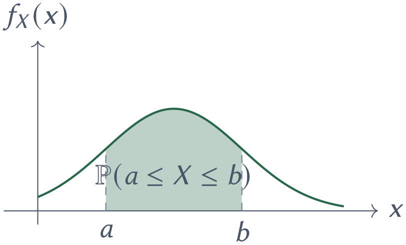
Properties:

1. $f_X(x) \geq 0$ for all x
2. $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$

Key insight: The PDF is *not* a probability. It's a *density*.

Probabilities are **areas under the curve**.

PDF Intuition



The shaded area equals $\mathbb{P}(a \leq X \leq b)$.

The total area under the curve equals 1.

Example: Uniform Distribution

$X \sim \text{Uniform}(0, 1)$ means X is equally likely to be anywhere in $[0, 1]$.

PDF:

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

Example: What's $\mathbb{P}(0.3 \leq X \leq 0.7)$?

$$\mathbb{P}(0.3 \leq X \leq 0.7) = \int_{0.3}^{0.7} 1\, dx = 0.7 - 0.3 = 0.4$$

For the uniform distribution, probability = length of interval.

Cumulative Distribution Function

A unifying concept

Cumulative Distribution Function (CDF)

Definition

The **CDF** of a random variable X is:

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

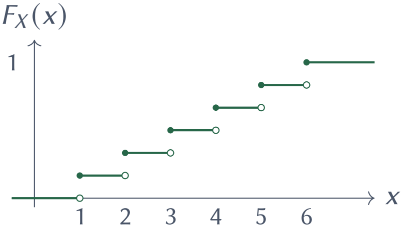
Properties:

1. $F_X(x)$ is non-decreasing
2. $\lim_{x \to -\infty} F_X(x) = 0$
3. $\lim_{x \to \infty} F_X(x) = 1$
4. $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$

Every random variable has a CDF. It's the universal language.

CDF for Discrete Variables

For a fair die: $F_X(x) = \sum_{k \leq x} f_X(k) = \sum_{k \leq x} \frac{1}{6}$



The CDF is a step function for discrete random variables.

CDF for Continuous Variables

For continuous X with PDF $f_X(x)$:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$$

And conversely: $f_X(x) = \frac{d}{dx} F_X(x)$

Example: For $X \sim \text{Uniform}(0, 1)$:

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}$$

Why the CDF Matters

The CDF is the **universal** way to describe any distribution.

- Every random variable has a CDF
- Not every random variable has a PMF (continuous ones don't)
- Not every random variable has a PDF (discrete ones don't)

Quantiles come from the CDF:

- Median: $F_X^{-1}(0.5)$ — the value where half the probability is below
- 95th percentile: $F_X^{-1}(0.95)$

Statistical software uses CDFs constantly: `pnorm()`, `qnorm()`, etc.

Part V

Joint Distributions

Multiple random variables together

Joint Distributions

Why we need them

In practice, we care about *relationships* between variables:

- How does education relate to income?
- How does campaign spending relate to vote share?
- Are two variables independent?

To answer these, we need to describe *two or more* random variables *together*.

This is the **joint distribution**.

Joint PMF

For discrete random variables

For discrete random variables X and Y, the **joint PMF** is:

$$f_{X,Y}(x, y) = \mathbb{P}(X = x \text{ and } Y = y)$$

Properties:

1. $f_{X,Y}(x, y) \geq 0$ for all x, y
2. $\sum_x \sum_y f_{X,Y}(x, y) = 1$

The joint PMF is often displayed as a table.

Example: Joint PMF

Roll two dice. Let X = first die, Y = second die.

Since the dice are independent, $f_{X,Y}(x, y) = \frac{1}{36}$ for all pairs.

More interesting: Let X = first die, S = sum of both dice.

	$X = 1$	$X = 2$	$X = 3$	$X = 4$	$X = 5$	$X = 6$
$S = 2$	$\frac{1}{36}$	0	0	0	0	0
$S = 3$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	0
$S = 4$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

X and S are **not** independent—knowing X tells you something about S.

Marginal Distributions

Recovering individual distributions

Given a joint PMF, we can recover the **marginal PMF** of each variable:

$$f_X(x) = \sum_y f_{X,Y}(x, y)$$

$$f_Y(y) = \sum_x f_{X,Y}(x, y)$$

Intuition: Sum over all possible values of the other variable.

"Marginalize out" the variable you don't care about.

Example: Computing Marginals

Joint distribution of X (party) and Y (vote):

	$Y = 0$ (No)	$Y = 1$ (Yes)	$f_X(x)$
$X = 0$ (Rep)	0.30	0.15	0.45
$X = 1$ (Dem)	0.10	0.45	0.55
$f_Y(y)$	0.40	0.60	1.00

Marginal of X: $f_X(0) = 0.30 + 0.15 = 0.45$

Marginal of Y: $f_Y(1) = 0.15 + 0.45 = 0.60$

The marginals are the row and column sums.

Conditional Distributions

Distributions given information

The **conditional PMF** of Y given $X = x$ is:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

From our example: What's the distribution of vote (Y) among Democrats ($X = 1$)?

$$f_{Y|X}(0|1) = \frac{0.10}{0.55} \approx 0.18 \qquad f_{Y|X}(1|1) = \frac{0.45}{0.55} \approx 0.82$$

Among Democrats, 82% vote Yes. Among Republicans: $0.15/0.45 \approx 33\%$.

Independence of Random Variables

When knowing one tells you nothing about the other

Independence of Random Variables

Definition

Random variables X and Y are **independent** if:

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y) \quad \text{for all } x, y$$

Equivalent statements:

- $f_{Y|X}(y|x) = f_Y(y)$ for all x, y
- Knowing X tells you nothing about the distribution of Y

Notation: $X \perp\!\!\!\perp Y$

This extends independence of events to random variables.

Testing Independence from a Joint PMF

Question: Are X (party) and Y (vote) independent?

	$Y=0$	$Y=1$	$f_X(x)$
$X=0$	0.30	0.15	0.45
$X=1$	0.10	0.45	0.55
$f_Y(y)$	0.40	0.60	1.00

Check: Does $f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$?

For $X=0, Y=0$: $f_X(0) \cdot f_Y(0) = 0.45 \times 0.40 = 0.18$

But $f_{X,Y}(0,0) = 0.30 \neq 0.18$

Not independent. Party and vote are related.

What Independence Would Look Like

If X and Y were independent with the same marginals:

	$Y = 0$	$Y = 1$	$f_X(x)$
$X = 0$	$0.45 \times 0.40 = 0.18$	$0.45 \times 0.60 = 0.27$	0.45
$X = 1$	$0.55 \times 0.40 = 0.22$	$0.55 \times 0.60 = 0.33$	0.55
$f_Y(y)$	0.40	0.60	1.00

Each cell would equal (row marginal) $\times$ (column marginal).

Under independence, party wouldn't predict vote at all.

Today's Key Ideas

1. **Random variables**: Functions mapping outcomes to numbers
2. **PMF** (discrete): $f_X(x) = \mathbb{P}(X = x)$
3. **PDF** (continuous): Probability = area under the curve
4. **CDF** (both): $F_X(x) = \mathbb{P}(X \leq x)$
5. **Joint distributions**: Describe multiple variables together
6. **Independence**: $f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$

Now we can describe *what* we want to learn about populations.

Looking Ahead

Wednesday: Famous distributions

- Bernoulli and Binomial (counting successes)
- Poisson (rare events)
- Uniform and Normal (continuous)

Next week: Expected value and variance

- Summarizing distributions with numbers
- The most important summary: the mean

Wednesday's distributions will show up constantly—they're the building blocks.

For Wednesday

Reading:

- Aronow & Miller, §1.2 (finish): Support, bivariate distributions
- Blackwell, Chapter 2.3–2.4: Plug-in estimators

Problem Set 1 will be posted this week.

- Covers probability, conditional probability, Bayes' Rule
- Includes working with joint PMFs and testing independence
- Due: February 14

Questions?