Continuous Distributions

Uniform, Normal, Exponential, and More

Scott Cunningham

Gov 2001 · Harvard University

Spring 2026

Where We Are

From discrete counts to continuous measurements

Monday: Discrete distributions

- Bernoulli, Binomial, Poisson
- Counting successes and rare events

Today: Continuous distributions — the **A-list** stars of statistics

- **Uniform** — equal probability, foundation for simulation
- **Normal** — the star of the show (CLT, regression, everything)
- **Exponential** — waiting times, survival analysis
- **Chi-square,** t — the B-list supporting cast for inference

Reading: Aronow & Miller §1.4–1.5, Blackwell 2.4–2.5

Recall My Pedagogy

How I approach each topic

My approach, usually in this order:

1. **History & narrative** — Who discovered it? What problem were they solving?
2. **Political science application** — Where does this show up in our world?
3. **Visuals** — What does it look like? Build intuition graphically
4. **Technical rigor** — The math that makes it precise

The "Netflix–Matt Damon" principle:

Netflix worked with Matt Damon on writing movies for their platform. Their advice: "Remind audiences of the plot regularly." Their data showed viewers constantly pausing, multitasking, distracted — they needed more help than a theater audience.

I'll do the same. Even for concepts from weeks ago, I'll remind you of the plot. **This isn't remedial — it's respect for how we actually learn** when juggling too much.

The Pattern Continues: Parameters Define Everything

Same roadmap as discrete, now with PDFs instead of PMFs

Monday's pattern for discrete distributions:

$$\text{Parameters} \rightarrow \text{PMF} \rightarrow \mathbb{E}[X], \text{Var}[X]$$

Today's pattern for continuous distributions:

$$\text{Parameters} \rightarrow \text{PDF} \rightarrow \text{CDF} \rightarrow \mathbb{E}[X], \text{Var}[X]$$

For each distribution today, we'll identify:

1. **Parameters**: What do you need to specify? $(\mu, \sigma^2, \lambda, a, b)$
2. **PDF**: The density function $f(x)$
3. **CDF**: The cumulative distribution $F(x) = \mathbb{P}(X \leq x)$
4. **Moments**: Expected value and variance

Once you know the parameters, everything else follows.

Wait — Why Is It Called $f(x)$ for Both PMF and PDF?

Same notation, different meanings

Historical convention: Both are "the function that characterizes the distribution," so mathematicians use the same letter. But they work differently:

	PMF (discrete)	PDF (continuous)
$f(x)$ means	$\mathbb{P}(X = x)$	density at x
What it is	An actual probability	**Not** a probability
Values	Always in $[0, 1]$	Can be *any* ≥ 0 (even > 1!)
To get $\mathbb{P}$	Read $f(x)$ directly	Must integrate: $\int_a^b f(x)\,dx$

The key difference: For continuous X, $\mathbb{P}(X = x) = 0$ for any specific x. Probability only exists over *intervals*.

Density Means "Probability Per Unit Length"

Why $f(x)$ can exceed 1

Think of $f(x)$ as **probability concentration** at point x.

Example: If $X \sim \text{Uniform}(0, 0.1)$, then:

$$f(x) = \frac{1}{0.1} = 10 \quad \text{for } x \in [0, 0.1]$$

That's a density of 10 — but the total probability is still:

$$\int_0^{0.1} 10\, dx = 10 \times 0.1 = 1 \quad \checkmark$$

Intuition: High density means probability is *concentrated*. Low density means probability is *spread out*.

Exponential with large λ has $f(0)$ very large — probability is concentrated near zero, not "more than 100% likely."

Most of Your Statistical Life Will Be Normal

The A-list and B-list of continuous distributions

A-list actors — you'll model data with these:

- **Normal** — regression errors, polling, heights, test scores
- **Exponential** — waiting times, survival, duration models
- **Uniform** — simulation, randomization, probability foundations

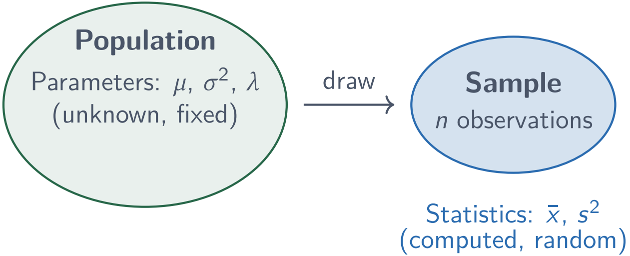
B-list actors — supporting roles for inference:

- **Chi-square** — variance estimation, goodness-of-fit
- t**-distribution** — hypothesis testing with estimated variance

The B-list are *derived* from Normal. You don't model data with them—you use them for inference.

Statistics Is Learning About Populations from Samples

The central distinction



The game: Use sample statistics to *estimate* population parameters.

Today's distributions describe *populations*. Estimation theory tells us how sample statistics behave.

Reminder: Why We Care About Distributions

The link to inference

The roadmap:

1. **Population**: Described by a distribution with unknown parameters
2. **Sample**: Data we observe (drawn from the population)
3. **Estimation**: Use data to learn about parameters (μ, σ^2, λ)
4. **Uncertainty**: Quantified via sampling distributions (which we derive from population distributions)

Today's distributions matter because:

- The **Normal** is the sampling distribution of the mean (CLT)
- The **Chi-square** appears when estimating variance
- The t-**distribution** is what we use for hypothesis tests with estimated variance

Everything connects. Today we're building the vocabulary you'll use for inference.

The Uniform Distribution

Simple but Fundamental

Notation Warning: Parameters Mean Different Things

Another inconsistency I didn't invent

Compare these two "standard" distributions:

Distribution	Notation	What the numbers mean
Standard Normal	$Z \sim N(0, 1)$	$N(\text{mean}, \text{variance})$
"Standard" Uniform	$U \sim \text{Uniform}(0, 1)$	$\text{Uniform}(\text{lower}, \text{upper})$

For Normal: the parameters *are* the mean and variance.
For Uniform: the parameters are the *interval endpoints* — you derive the moments:

$$\mathbb{E}[U] = \frac{0+1}{2} = 0.5, \quad \text{Var}[U] = \frac{(1-0)^2}{12} = \frac{1}{12}$$

Like PMF vs PDF using the same $f(x)$: inconsistent notation that evolved historically. Just be aware.

Why the Difference? Distributions Parameterize Different Concepts

Moments vs. bounds vs. rates

The deeper point: Every distribution needs parameters, but what those parameters *represent* varies:

Distribution	Parameters	What they represent
$X \sim N(\mu, \sigma^2)$	μ, σ^2	Moments (mean, variance)
$X \sim \text{Uniform}(a, b)$	a, b	Support bounds (endpoints)
$X \sim \text{Exponential}(\lambda)$	λ	Rate (events per unit time)

The general Normal: $X \sim N(\mu, \sigma^2)$

- $\mu =$ population mean (can be any real number)
- $\sigma^2 =$ population variance (must be positive)
- The "standard" Normal $N(0, 1)$ is just the special case $\mu = 0$, $\sigma^2 = 1$

We'll use μ and σ^2 throughout. When you see specific numbers, you're seeing a specific distribution from the family.

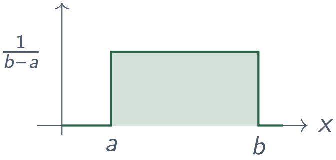
The Simplest Distribution: Equal Probability Everywhere

Political example: When does the voter arrive?

Example: A voter arrives at a polling station sometime between 8am and 8pm. If arrivals are "uniformly distributed," any moment is equally likely.

Definition: $X \sim \text{Uniform}(a, b)$ has PDF:

$$f(x) = \frac{1}{b-a} \quad \text{for } x \in [a, b]$$



Key formulas: $\mathbb{E}[X] = \frac{a+b}{2}$ (midpoint), $\quad \text{Var}[X] = \frac{(b-a)^2}{12}$

The Standard Uniform: $U \sim \textbf{Uniform}(0, 1)$

The building block for everything

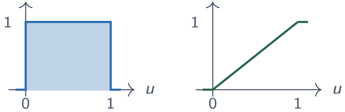
The **standard uniform** $U \sim \text{Uniform}(0, 1)$ is special:

- PDF: $f(u) = 1$ for $u \in [0, 1]$ *(a flat horizontal line at height 1)*
- CDF: $F(u) = u$ for $u \in [0, 1]$ *(output = input, so $y = x$, a diagonal line)*
- $\mathbb{E}[U] = 0.5, \quad \text{Var}[U] = 1/12$

Why is it fundamental?

- Every random number generator starts with Uniform(0,1)
- Randomization in experiments: “treat if $U < 0.5$”

PDF: $f(u) = 1$ CDF: $F(u) = u$



What Does the CDF Actually Tell You?

Reading the plot

$F(u) = \mathbb{P}(U \leq u)$ = "the probability of getting a value at or below u"

For Uniform(0,1), $F(u) = u$ means:

- $F(0.3) = 0.3 \quad \longrightarrow \quad$ "30% chance of being below 0.3"
- $F(0.7) = 0.7 \quad \longrightarrow \quad$ "70% chance of being below 0.7"

The slope of the CDF tells you where outcomes concentrate:

- **Steep slope** = probability accumulating fast → outcomes cluster there
- **Flat slope** = probability not accumulating → outcomes rare there
- **Constant slope** (diagonal line) = probability accumulates evenly everywhere

A diagonal CDF is the visual definition of "uniform" — no region is more likely than any other.

Support Tells You Where Outcomes Are Possible

You worked with this on Problem Set 2

Definition: The **support** of a random variable is where its PDF is positive:

$$\text{Supp}[X] = \{x : f(x) > 0\}$$

Three types of support we'll see today:

Distribution	Support	Type
Uniform(a, b)	$[a, b]$	Bounded (finite interval)
Normal(μ, σ^2)	$(-\infty, +\infty)$	Unbounded (whole line)
Exponential(λ)	$[0, +\infty)$	Half-line (non-negative)

PS2 Q2 asked you to find support. This concept matters for specifying models correctly.

Have You Ever Wondered How R Generates Random Numbers?

Where do Bernoulli, Binomial, Poisson, Normal come from?

When you type `rbinom(1, 10, 0.5)` in R, what's actually happening?

The answer: **Everything comes from the Uniform.**

- The **Uniform** is the raw randomness — the stochastic part
- The **parameters** (p, n, λ) shape that randomness into the distribution you want

You choose the parameters, R transforms the Uniform:

`rbinom(1, n, p)`	Uniform + cutpoint p, repeated n times $\longrightarrow$ Binomial
`rpois(1, lambda)`	Uniform + staircase shaped by λ $\longrightarrow$ Poisson
`rexp(1, lambda)`	Uniform + inverse CDF shaped by λ $\longrightarrow$ Exponential

The parameters are the recipe. The Uniform is the ingredient.

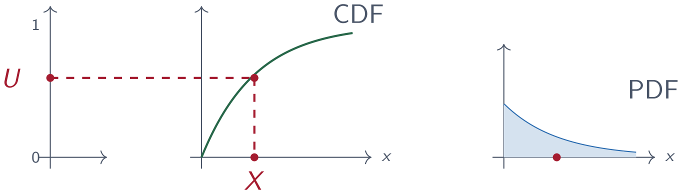
Next: Let's see how this works under the hood.

Generating Any Distribution: Start With the Picture

The inverse CDF method, visualized

Goal: You want to simulate X from some distribution F. All you have is $U \sim \text{Uniform}(0, 1)$.

1. Draw U2. Find where U hits CDF 3. That's your draw!



The recipe: Draw $U \sim \text{Uniform}(0, 1)$ on the y-axis $\rightarrow$ trace horizontally to CDF $\rightarrow$ drop down to x-axis $\rightarrow$ that's $X = F^{-1}(U)$

This works because the CDF maps outcomes to $[0, 1]$, so the inverse maps $[0, 1]$ back to outcomes.

Why Does This Produce the Right Distribution?

The CDF slope determines where outcomes land

Key insight: The shape of the CDF controls the transformation.

- **Where CDF is steep**: Many different U values map to a *narrow* range of X
  $\longrightarrow$ Outcomes **cluster** there (high density)
- **Where CDF is flat**: Few U values map to that region of X
  $\longrightarrow$ Outcomes are **rare** there (low density)
- **Uniform's 45° line**: No stretching or compression — equal in, equal out
  $\longrightarrow$ This is why Uniform is the "undistorted" starting point

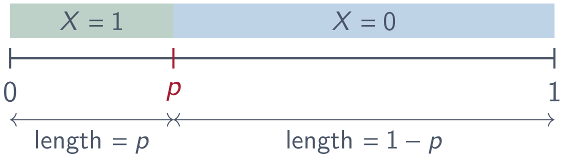
The inverse CDF "stretches" the Uniform to match whatever density you want.

The Uniform is like a blank canvas. The inverse CDF is the paint.

Universality for Discrete: Bernoulli

The simplest case — one cutpoint

Goal: Generate $X \sim \text{Bernoulli}(p)$ from $U \sim \text{Uniform}(0, 1)$



Algorithm:

1. Draw $U \sim \text{Uniform}(0, 1)$ — a random point on $[0, 1]$
2. If $U < p$: return $X = 1$ (success)
3. If $U \geq p$: return $X = 0$ (failure)

Why it works: $\mathbb{P}(U < p) = p$ and $\mathbb{P}(U \geq p) = 1 - p$ — exactly Bernoulli!

Universality for Discrete: Binomial

Just repeat the Bernoulli trick n times

Goal: Generate $X \sim \text{Binomial}(n, p)$ from Uniform draws

Recall: $X \sim \text{Binomial}(n, p)$ counts successes in n independent Bernoulli(p) trials.

Algorithm:

1. Draw $U_1, U_2, \ldots, U_n$ independently from Uniform$(0, 1)$
2. For each U_i: count as "success" if $U_i < p$
3. Return X = total number of successes

Example ($n = 5$, $p = 0.5$):

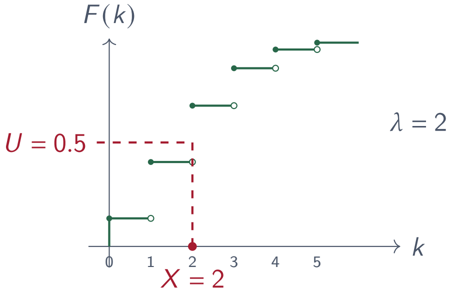
$$\begin{array}{ccccc} U_1 = 0.23 & U_2 = 0.81 & U_3 = 0.15 & U_4 = 0.67 & U_5 = 0.42 \\ < 0.5 \checkmark & \geq 0.5 & < 0.5 \checkmark & \geq 0.5 & \geq 0.5 \end{array} \Rightarrow X = 2$$

This is exactly what `rbinom(1, n, p)` does in R.

Universality for Discrete: Poisson

The CDF is a staircase — find which step U lands on

Goal: Generate $X \sim \text{Poisson}(\lambda)$ from $U \sim \text{Uniform}(0, 1)$



Algorithm: Find smallest k such that $F(k) \geq U$
Example ($\lambda = 2,\ U = 0.5$): $F(1) = 0.406 < 0.5$ but $F(2) = 0.677 \geq 0.5 \Rightarrow X = 2$
Key insight: Height of each jump = PMF at that point. **Bigger jumps = more likely outcomes.**

The Formal Statement: Universality of the Uniform

Now that you've seen it work, here's the theorem

Theorem (Probability Integral Transform):
Let X be a continuous random variable with CDF F. Then:

$$F(X) \sim \text{Uniform}(0, 1)$$

The converse (the useful direction for simulation):
If $U \sim \text{Uniform}(0, 1)$ and F is any CDF, then:

$$X = F^{-1}(U) \text{ has CDF } F$$

This is why Uniform is the "mother of all distributions": From one Uniform, you can generate *any* distribution — continuous or discrete.

PS3 asks you to prove $F(X) \sim \text{Unif}(0, 1)$ directly for Exponential. Hint: Find $\mathbb{P}(F(X) \leq u)$.

The Big Picture: What We Just Learned

Universality of the Uniform, informally

The Uniform is special because its CDF is a 45° line — no distortion.

Every other CDF "stretches" the Uniform:

- Steep regions $\rightarrow$ outcomes cluster there
- Flat regions $\rightarrow$ outcomes are rare there

The inverse CDF method:

- Draw $U \sim \text{Uniform}(0, 1)$ — your "raw randomness"
- Apply F^{-1} — the CDF of whatever distribution you want
- Out comes $X \sim F$ — a draw from your target distribution

This works for everything: Bernoulli, Binomial, Poisson, Normal, Exponential — any distribution you'll ever meet.

When you call `rnorm()`, `rpois()`, or `rexp()` in R, this is what's happening under the hood.

Part II

The Normal Distribution

The Star of the Show

De Moivre Discovered It; Gauss Got the Credit

Stigler's Law of Eponymy

Stigler's Law: "No scientific discovery is named after its original discoverer."

The Normal distribution is called "Gaussian"—but Gauss didn't discover it.

- **Abraham de Moivre (1733)**: French Huguenot exile in London, surviving by tutoring aristocrats in gambling mathematics. First derived the normal curve in *The Doctrine of Chances*.
- **Pierre-Simon Laplace (1774–1812)**: Developed the theory systematically. Proved early versions of the Central Limit Theorem.
- **Carl Friedrich Gauss (1809)**: Applied it to astronomical errors. Got the credit. But Gauss himself called it the "Laplacian curve."

De Moivre died impoverished. Gauss is called the "Prince of Mathematicians." Life isn't fair.

Gauss, Legendre, and Least Squares

The drama continues

Stigler's Law strikes again: Least squares wasn't discovered by Gauss either.

- **Legendre (1805)**: Published the method of least squares
- **Gauss (1809)**: Published it four years later, but claimed he'd been using it since 1795 — when he was 18
- Legendre was *furious*. Gauss offered no proof of his earlier use.

But here's Gauss's real contribution:
He showed that *if* errors are normally distributed, *then* least squares gives the best estimates (maximum likelihood).

The Normal distribution justifies least squares. That's why they're forever linked.

When you run a regression, you're relying on Gauss's 1809 insight — even if Legendre got there first.

The Normal Distribution: The Star of the Show

Application first: Where do you see it?

Examples:

- Heights of adults, test scores, measurement errors
- **Polling errors** — why we talk about "margin of error"
- **Regression residuals** — the foundation of inference

Definition: $X \sim \text{Normal}(\mu, \sigma^2)$ has PDF:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Why everywhere? Three reasons:

1. **Central Limit Theorem**: Sample means are approximately normal
2. **Closure**: Sums of normals are normal
3. **Tractability**: Easy to compute probabilities

Standardization Converts Any Normal to Z

Definition: $Z \sim N(0,1)$ is the **standard normal**.

Standardization: If $X \sim N(\mu, \sigma^2)$, then:

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$

What does Z mean? It's the deviation from the mean, scaled by the standard deviation.

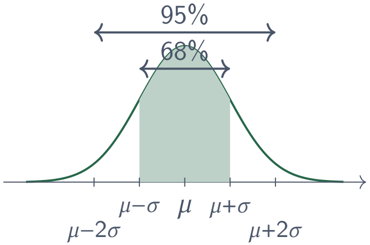
- $Z = 0$: you're at the mean
- $Z = 1$: you're one SD above the mean
- $Z = -2$: you're two SDs below the mean

Key notation: $\phi(z)$ = standard normal PDF; $\Phi(z) = \mathbb{P}(Z \leq z)$ = standard normal CDF

Tables, software, and formulas are all in terms of Φ. Standardize first.

Most Probability Concentrates Near the Mean

For $X \sim N(\mu, \sigma^2)$:



- **68%** of values within 1 SD of mean
- **95%** within 2 SDs (more precisely: 1.96)
- **99.7%** within 3 SDs

Normal Has Unbounded Support — But Tails Vanish Fast

Theoretically infinite, practically finite

Support: $\text{Supp}[X] = (-\infty, +\infty)$ — any value is *theoretically* possible.

But probabilities decay exponentially in the tails:

- Outside 3 SDs: only 0.3% of probability
- Outside 4 SDs: only 0.006% of probability
- Outside 5 SDs: essentially zero (1 in 3.5 million)

Practical implication: For heights (mean 170cm, SD 10cm):

- Normal says negative heights are "possible" — but $P(X < 0) \approx 0$
- The model is an approximation; we accept tiny errors in exchange for tractability

Contrast with Exponential: support $[0, \infty)$ *enforces* non-negativity.

A Warning: The Normal Can Be Misused

"The Bell Curve" controversy

1994: Herrnstein & Murray publish *The Bell Curve*, claiming IQ differences between racial groups are genetic and immutable.

The statistical sin: They treated the Normal distribution as *destiny* rather than *description*.

James Heckman's critique (Nobel laureate, 2000):

- IQ is not fixed — it responds to environment and intervention
- The authors confused *description* with *explanation*
- Selection bias: who takes the tests, when, under what conditions?

Lesson: The Normal describes many phenomena. It doesn't explain them. Distributions are tools, not theories of causation.

Statistics without causal reasoning is dangerous.

Normal Closure Properties

Sums and linear combinations stay normal

Property 1 (Scaling and shifting):
If $X \sim N(\mu, \sigma^2)$, then for constants a, b:

$$aX + b \sim N(a\mu + b,\ a^2\sigma^2)$$

Property 2 (Sum of independent normals):
If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ are **independent**, then:

$$X + Y \sim N(\mu_X + \mu_Y,\ \sigma_X^2 + \sigma_Y^2)$$

The key result for inference: If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\theta, \sigma^2)$, then:

$$\bar{Y} \sim N\left(\theta, \frac{\sigma^2}{n}\right)$$

PS3 asks: How does $\mathbb{P}(|\bar{Y} - \theta| < \varepsilon)$ change with n? This formula is your starting point.

Galton and "Regression to the Mean"

Why tall parents have shorter children (on average)

Francis Galton (1886): Studied heights of fathers and sons.

Finding: Sons of very tall fathers were tall — but not *as* tall as their fathers. Sons of very short fathers were short — but not *as* short.

Galton called this "regression toward mediocrity":

- Extreme observations tend to be followed by less extreme ones
- This is a *statistical phenomenon*, not a biological force
- It happens whenever two variables are imperfectly correlated

Why "regression"? This is literally where the term comes from. Galton was "regressing" son's height on father's height.

The Normal distribution quantifies this: extreme Z-scores are rare by definition.

Part III

The Exponential Distribution

Waiting for an Event

Where Does the Exponential Come From?

A story of outsiders, death, and waiting

Benjamin Gompertz (1825): A Jewish mathematician in London, barred from university because of his religion. Self-taught from Newton's writings. His brother-in-law founded an insurance company and made Gompertz the actuary.

His question: *How do we price life insurance?* He needed to model how long people live — and discovered that mortality risk increases exponentially with age. The exponential function became central to survival analysis.

"Event history analysis" and "survival analysis" — now workhorses of modern social science — trace back to 19th-century actuaries modeling death.

Gompertz was elected to the Royal Society despite being denied a university degree.

How Long Until the Next Supreme Court Vacancy?

Application first: Waiting times in politics

Political science questions that involve waiting:

- How long until the next Supreme Court vacancy?
- How long will this ceasefire last?
- How long until a cabinet collapse?
- Time between terrorist attacks in a region?

Historical data: Supreme Court vacancies occur at rate $\lambda \approx 0.5$ per year.

$\Rightarrow$ Average wait: about 2 years between vacancies.

The **Exponential distribution** models these waiting times.

The Exponential Distribution

The math behind waiting times

Definition: $T \sim \text{Exponential}(\lambda)$ has PDF:

$$f(t) = \lambda e^{-\lambda t} \quad \text{for } t \geq 0$$

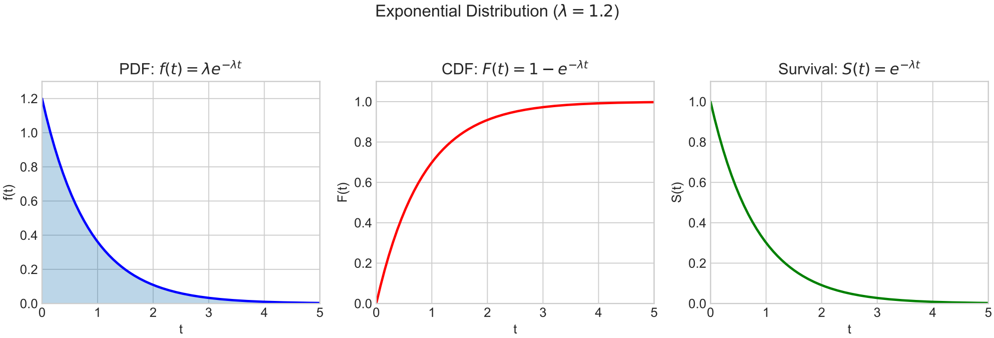
where $\lambda > 0$ is the **rate parameter**.

Support: $[0, +\infty)$ — waiting times are always non-negative.

PS2 Connection: Problem 7(c) asked you to find c for $f(y) = ce^{-2y}$. That's an Exponential($\lambda = 2$)! The answer was $c = 2$.

The Survival Function Decays Exponentially

PDF, CDF, and Survival — three views of the same distribution



Survival function: $S(t) = \mathbb{P}(T > t) = 1 - F(t) = e^{-\lambda t}$

The probability of “surviving” (not yet experiencing the event) past time t decays exponentially.

Average Wait Is the Inverse of the Rate

Key properties of the Exponential

For $T \sim \text{Exponential}(\lambda)$:

Expected value: $\mathbb{E}[T] = \frac{1}{\lambda}$
Variance: $\text{Var}[T] = \frac{1}{\lambda^2}$

Interpretation: If events occur at rate λ per unit time, the average wait is $1/\lambda$.

Example: Supreme Court vacancies at rate $\lambda = 0.5$ per year $\rightarrow$ average wait $=$ 2 years.

The Exponential Distribution Has No Memory

How long you've waited doesn't affect how much longer you'll wait

Property: For $T \sim \text{Exponential}(\lambda)$:

$$\mathbb{P}(T > s + t \mid T > s) = \mathbb{P}(T > t)$$

In words: Given that you've already waited s units, the probability of waiting *another* t units is the same as if you'd just started waiting.

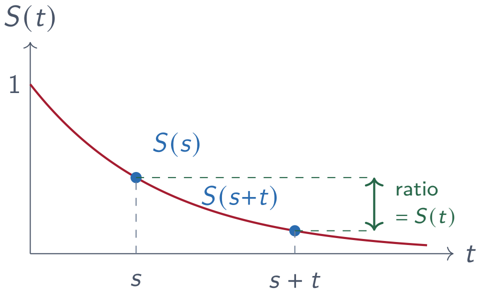
Proof:

$$\mathbb{P}(T > s + t \mid T > s) = \frac{\mathbb{P}(T > s + t)}{\mathbb{P}(T > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = \mathbb{P}(T > t)$$

Only the exponential (continuous) and geometric (discrete) have this property.

Memorylessness Visualized

The survival curve “restarts” at any point



The *ratio* of survival probabilities depends only on the additional wait t, not on s.

Why Does "Memory" Even Make Sense Here?

It's the question being asked

Bernoulli: "What happened?" — heads or tails, 0 or 1.

- Resolves instantly. No duration. No elapsed time.
- You can't ask "given that I've been partially through this coin flip..."
- The concept of memory has no surface to attach to.

Exponential: "How long until something happens?"

- Time is explicitly in the picture. You're sitting there waiting.
- You *can* ask: "I've waited 3 years — does that change my forecast?"
- For the Exponential, the answer is *no*. That's memorylessness.

The question determines whether memory is even a coherent concept.

The Geometric Bridge

Connecting Bernoulli to Exponential

The **Geometric distribution** asks a waiting-time question about Bernoulli trials:

"How many trials until the first success?"

- Each Bernoulli trial has no time dimension — it just happens
- But counting *how many* trials introduces a duration concept
- Now the past (failed trials) *could* inform the future
- And the answer is: it doesn't. **Geometric is memoryless too.**

Bernoulli $\longrightarrow$ **Geometric**

"What happened?" "How many until success?"

$\downarrow$

Exponential

"How long until event?"

Geometric is discrete memorylessness; Exponential is continuous memorylessness.

Part IV

The Poisson–Exponential Connection

Two Sides of One Process

A Brief History: How Did We Get Here?

Three people, three problems, one insight

Poisson (1837): French mathematician studying rare events — wrongful convictions in court trials. Asked: "If something rarely happens, how do we model how *many* times it occurs?" Developed the Poisson distribution, but didn't connect it to waiting times.

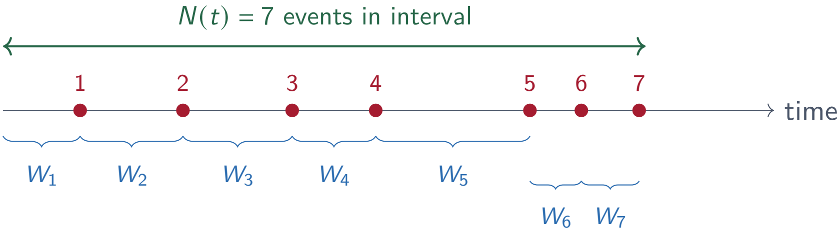
Bortkiewicz (1898): Russian economist, famously applied Poisson to Prussian cavalry soldiers killed by horse kicks. Showed rare events follow predictable patterns. Still focused on *counts*, not *durations*.

Erlang (1909): Danish engineer at the Copenhagen Telephone Exchange. His problem: how many phone lines do we need? He realized: calls arrive randomly (Poisson), but *time between calls* matters for capacity planning. **First to formally connect Poisson counts to exponential waiting times.**

The insight wasn't obvious. It took 70+ years from Poisson's distribution to Erlang's connection. Today, it's the foundation of queueing theory — from call centers to emergency rooms to internet servers.

Poisson Counts Events; Exponential Measures Waiting Times

Same process, different questions



- **Poisson question**: How many events in time t? $N(t) \sim \text{Poisson}(\lambda t)$
- **Exponential question**: How long until next event? $W_i \sim \text{Exp}(\lambda)$

Same rate λ. Same process. Different questions.

The Key Identity

Connecting Poisson and Exponential

Let T_1 be the time until the first event. Then:

$$\mathbb{P}(T_1 > t) = \mathbb{P}(\text{no events by time } t) = \mathbb{P}(N(t) = 0)$$

Using Poisson:

$$\mathbb{P}(N(t) = 0) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-\lambda t}$$

This is exactly the survival function of Exponential(λ)!

The Poisson count and exponential waiting time are two views of the same process.

Political Science Example

Supreme Court vacancies

Suppose vacancies occur at rate $\lambda = 0.5$ per year.

Poisson question: What's $\mathbb{P}$(at least 2 vacancies in a 4-year term)?

- $N(4) \sim \text{Poisson}(0.5 \times 4) = \text{Poisson}(2)$
- $\mathbb{P}(N \geq 2) = 1 - \mathbb{P}(N = 0) - \mathbb{P}(N = 1) = 1 - e^{-2} - 2e^{-2} \approx 0.59$

Exponential question: What's the average wait for the next vacancy?

- $T \sim \text{Exp}(0.5)$
- $\mathbb{E}[T] = 1/0.5 = 2$ years

Same λ, different questions, complementary answers.

Your Turn: Continuous Practice

Work through these with a partner

1. **Normal**: Adult heights are $N(170, 100)$ cm (mean 170, variance 100).
   - What's the standard deviation?
   - What range contains about 95% of heights?
2. **Exponential**: Congressional hearings occur at rate $\lambda = 3$ per month.
   - What's the expected wait for the next hearing?
   - What's $\mathbb{P}(\text{wait} > 1 \text{ month})$?

Answers: (1) SD = 10 cm; 150–190 cm. (2) $\mathbb{E}[T] = 1/3$ month; $\mathbb{P}(T > 1) = e^{-3} \approx 0.05$.

Part V

Chi-Square and t Distributions

The B-List: Supporting Actors for Inference

Chi-Square and t Are Inference Tools, Not Data Models

You don't model data with these—you use them for hypothesis testing

A-list vs B-list:

- **A-list** (Normal, Exponential, Uniform): You model *data* with these
- **B-list** (Chi-square, t): You use these for *inference about parameters*

Why do they exist?

- **Chi-square**: When you estimate variance from data, your estimate follows a χ^2
- t**-distribution**: When you test hypotheses using an estimated (not known) variance

The punchline: In a few weeks, when you run a regression and ask "is this coefficient statistically significant?"—the t-distribution will give you the answer.

We're planting seeds. You'll see these again in the regression unit.

The Chi-Square: Karl Pearson's 1900 Revolution

The birth of the goodness-of-fit test

Karl Pearson (1900): Statistician at University College London. Asked a simple question: *How do I know if my data actually fit a theoretical distribution?*

Before Pearson, researchers just assumed data were Normal. Pearson noticed real biological data were often skewed. He needed a formal test.

His solution: Sum up squared deviations between observed and expected counts. That sum follows a χ^2 distribution — and gives you a p-value.

The drama: Pearson got the degrees of freedom wrong. A young outsider named R.A. Fisher corrected him in 1922. Pearson refused to accept the correction and published a hostile "cooperative study" attacking Fisher. Fisher was furious, vowed never to publish in Pearson's journal again, and declared war on Pearson's entire approach to statistics.

The feud shaped 20th-century statistics. Fisher was right about the degrees of freedom.

Chi-Square Is a Sum of Squared Normals

Derived from Normal—support is $[0, \infty)$

Definition: If $Z_1, \ldots, Z_k \overset{\text{iid}}{\sim} N(0, 1)$, then:

$$X = Z_1^2 + Z_2^2 + \cdots + Z_k^2 \sim \chi_k^2$$

where k is the **degrees of freedom**.

Key facts:

- $\mathbb{E}[X] = k$
- $\text{Var}[X] = 2k$
- Support: $[0, \infty)$ — always non-negative (it's a sum of squares)

You'll see this when we estimate variance, test hypotheses about multiple coefficients, and compute R^2.

The t-Distribution: A Brewer's Secret

William Sealy Gosset and the Guinness brewery

William Sealy Gosset (1908): Chemist at the Guinness brewery in Dublin. His job: assess barley and hops quality. His problem: he only had small samples.

With small samples, the Normal distribution gives wrong answers — it's too confident. Gosset worked out the math for what happens when you estimate variance from limited data.

The catch: Guinness didn't allow employees to publish under their real names (they feared leaking trade secrets). So Gosset published as "Student" in 1908.

That's why it's called **Student's t-distribution** — not Gosset's.

The connection: Gosset spent 1906–07 studying with Karl Pearson in London. Pearson helped him with the mathematics. The chi-square and t are siblings.

Gosset's identity was only revealed publicly after his death in 1937.

The t Distribution Is Normal with Heavier Tails

What happens when you don't know the true variance

The problem: In real life, you don't know σ. You estimate it from data.

The consequence: Your estimate $\hat{\sigma}$ is uncertain. This makes extreme values more likely than the Normal predicts.

The solution: Use the t-distribution, which has heavier tails to account for this.

Definition: If $Z \sim N(0, 1)$ and $V \sim \chi^2_k$ are independent, then:

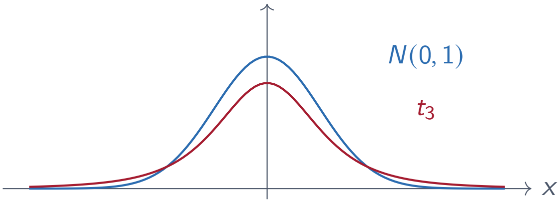
$$T = \frac{Z}{\sqrt{V/k}} \sim t_k$$

Key insight: Small k = more uncertainty about σ = heavier tails.

As $k \to \infty$, the t becomes Normal (you've estimated σ precisely).

This is why we use "t-tests" — they account for estimating variance from data.

Normal vs. t: Heavier Tails



The t distribution has more probability in the tails.

With small samples, extreme values are more likely — the t accounts for this.

Summary: Continuous Distributions

Distribution	$\mathbb{E}[X]$	$\text{Var}[X]$	**Use case**
Uniform(a, b)	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	Equal probability, simulation
Normal(μ, σ^2)	μ	σ^2	CLT, regression errors
Exponential(λ)	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	Waiting times, memoryless
χ^2_k	k	$2k$	Variance estimation, tests
t_k	0	$\frac{k}{k-2}$	Small-sample inference

Key connections:

- Uniform(0,1) $\rightarrow$ any distribution via inverse CDF
- Poisson $\leftrightarrow$ Exponential: counts vs. waiting times
- Normal $\rightarrow$ Chi-square (sum of squares) $\rightarrow$ t (ratio)

The A-List and B-List: A Summary

Which distributions model data? Which are for inference?

A-list actors — you model *data* with these:

- **Uniform**: Simulation, randomization, probability foundations
- **Normal**: CLT, regression errors, test scores, polling
- **Exponential**: Waiting times, survival analysis, duration

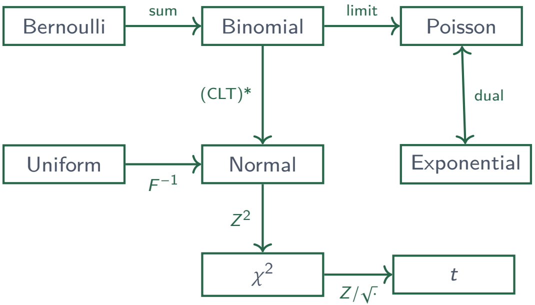
B-list actors — you use these for *inference*:

- **Chi-square**: Variance estimation, goodness-of-fit tests
- t: Hypothesis testing with estimated variance

The relationship: Normal $\xrightarrow{Z^2}$ Chi-square $\xrightarrow{Z/\sqrt{\cdot}}$ t

Most of your statistical life will be Normal. But when you estimate variance from data, the B-list appears.

How Distributions Connect: The Big Picture



Understanding these connections helps you see why certain distributions appear in certain contexts.

*CLT = Central Limit Theorem (Week 5). Sample means of *any* distribution approach Normal.

Looking Ahead

Next week: Joint distributions and the CEF

- Joint, marginal, and conditional distributions
- Covariance and correlation
- The Conditional Expectation Function (CEF)

Reading:

- Aronow & Miller, §1.3 and §2.2
- Blackwell, Chapter 2.4–2.5

Problem Set 3: Due next week, February 17th.