Lesson 1 — Statistical Modeling

Random Variables, Distributions, LLN, CLT

Applied Statistics Course

September 16, 2025

Outline

Why Probability? From Questions to Models

Natural language $\rightarrow$ **mathematical model** via events and random variables.

- Web A/B test: "Is variant B better than A?" $\Rightarrow$ Encode clicks as Bernoulli trials; compare p_A vs p_B.
- Manufacturing: "Are defects random and rare?" $\Rightarrow$ Model counts with Poisson; check fit.

Language of sets and probability

Events are sets (subsets of Ω). Questions like "did a click occur?" or "defects ≤ 3" translate to $A \in \mathcal{F}$ and $\mathbb{P}(A)$.

Probability theory provides mathematical tools designed to **formalize uncertainty** and enable rigorous **reasoning about random phenomena**.

From Natural Language to Events and RVs

Natural question	**Event / RV**	**Typical model**
"Roll of a fair die?"	$X \in \{1, \ldots, 6\}$, event $E = \{X \text{ even}\}$	X uniform on $\{1, \ldots, 6\}$; $\mathbb{P}(X = k) = 1/6$
"User clicks?"	$C = \{\text{click}\}$, indicator $X = \mathbf{1}_C$	$X \sim \text{Bernoulli}(p)$; compare p_A vs p_B
"Defects in a batch?"	Count $D \in \{0, 1, 2, \ldots\}$	$D \sim \text{Poisson}(\lambda)$ (rare, independent)
"Time until failure?"	Continuous $T \geq 0$	$T \sim \text{Exponential}(\lambda)$ (memoryless)

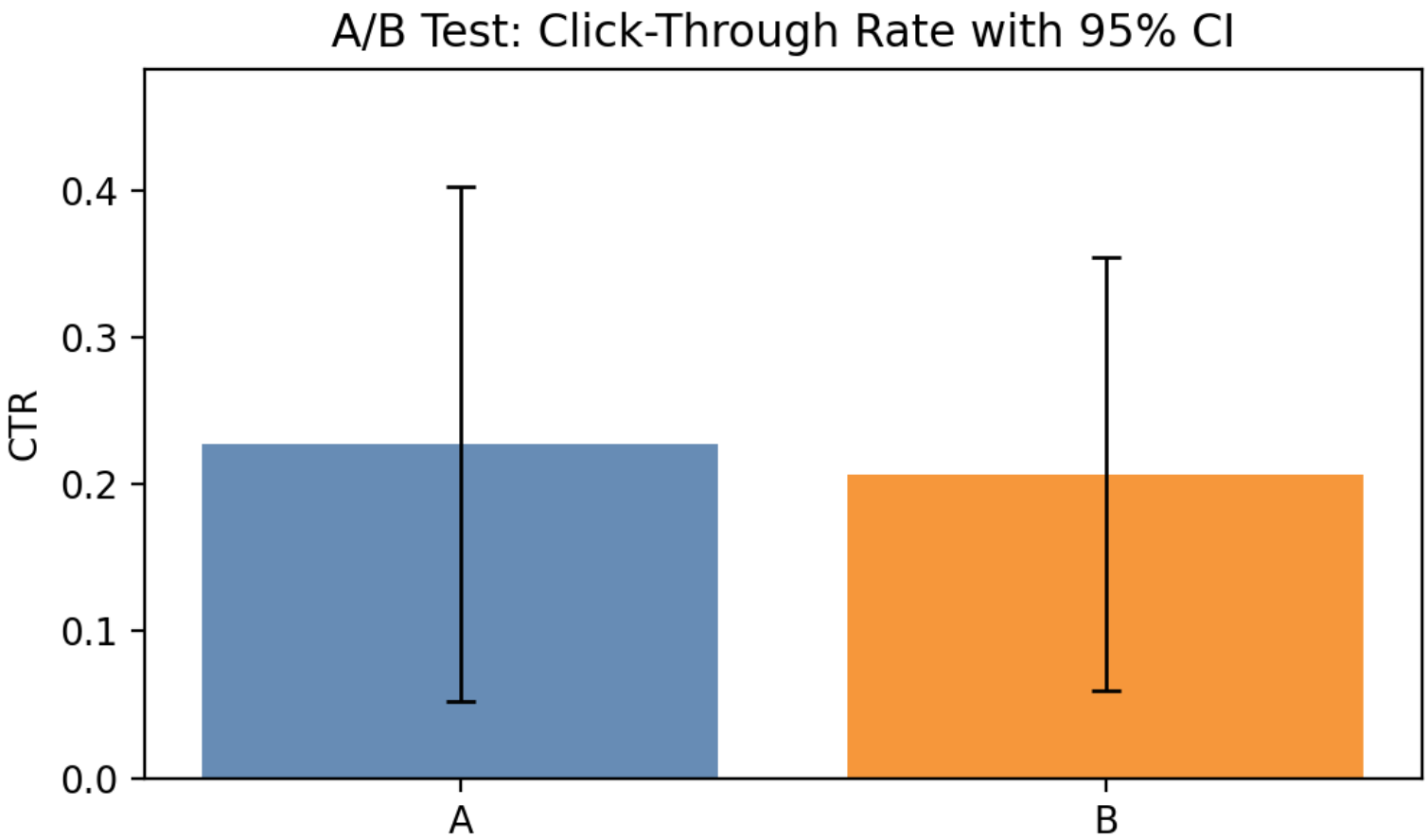
Set operations encode logic:

$A \cup B$ = "A or B", $A \cap B$ = "A and B"

A^c = "not A"

Example: A/B Test as Bernoulli/Binomial

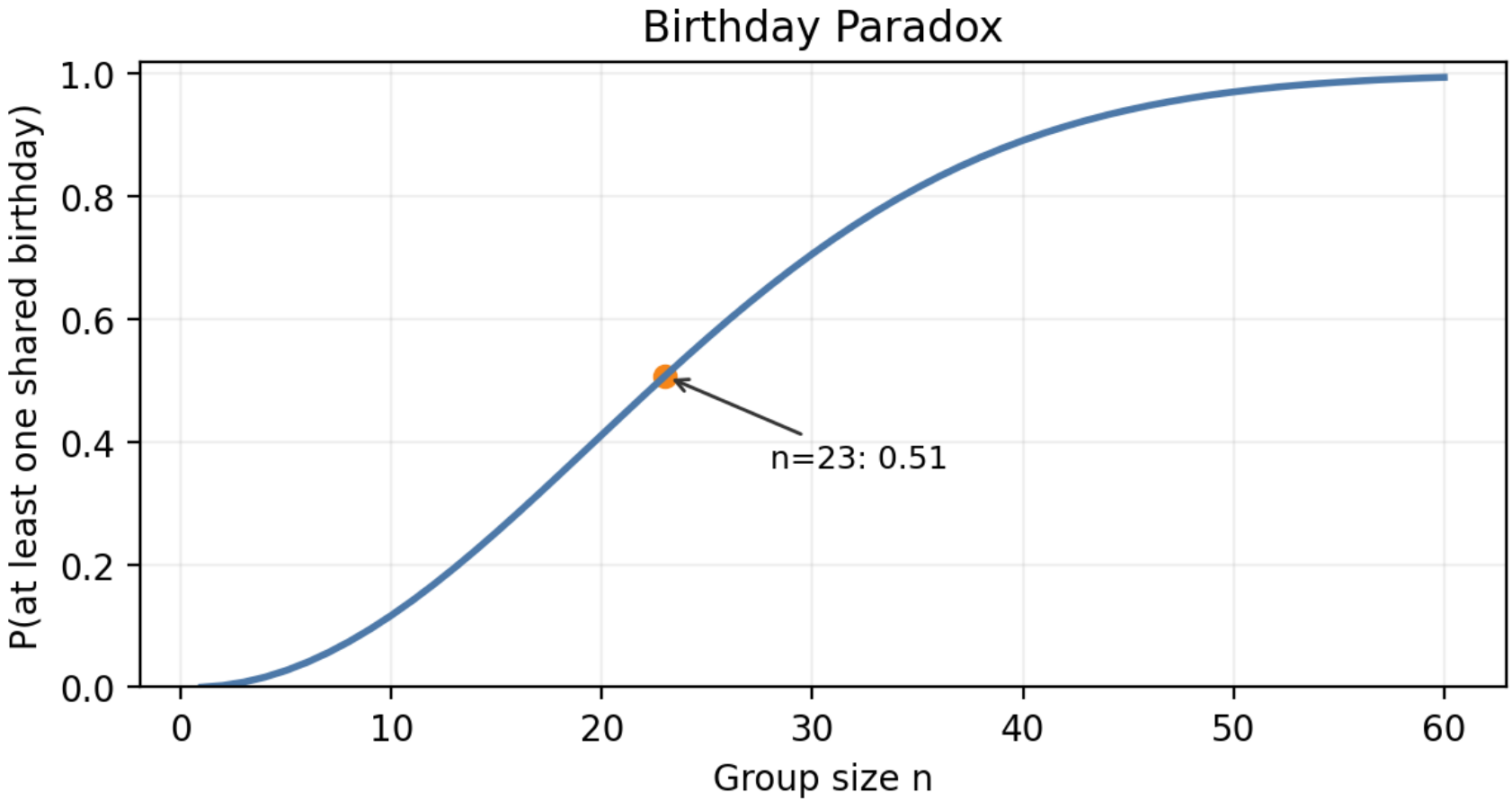
Each impression is a trial: click $= 1$, no click $= 0$. Variant-level CTR estimates $\hat{p} = \frac{\text{clicks}}{\text{impressions}}$.



Confidence intervals visualize uncertainty from finite samples.

Catchy: The Birthday Paradox

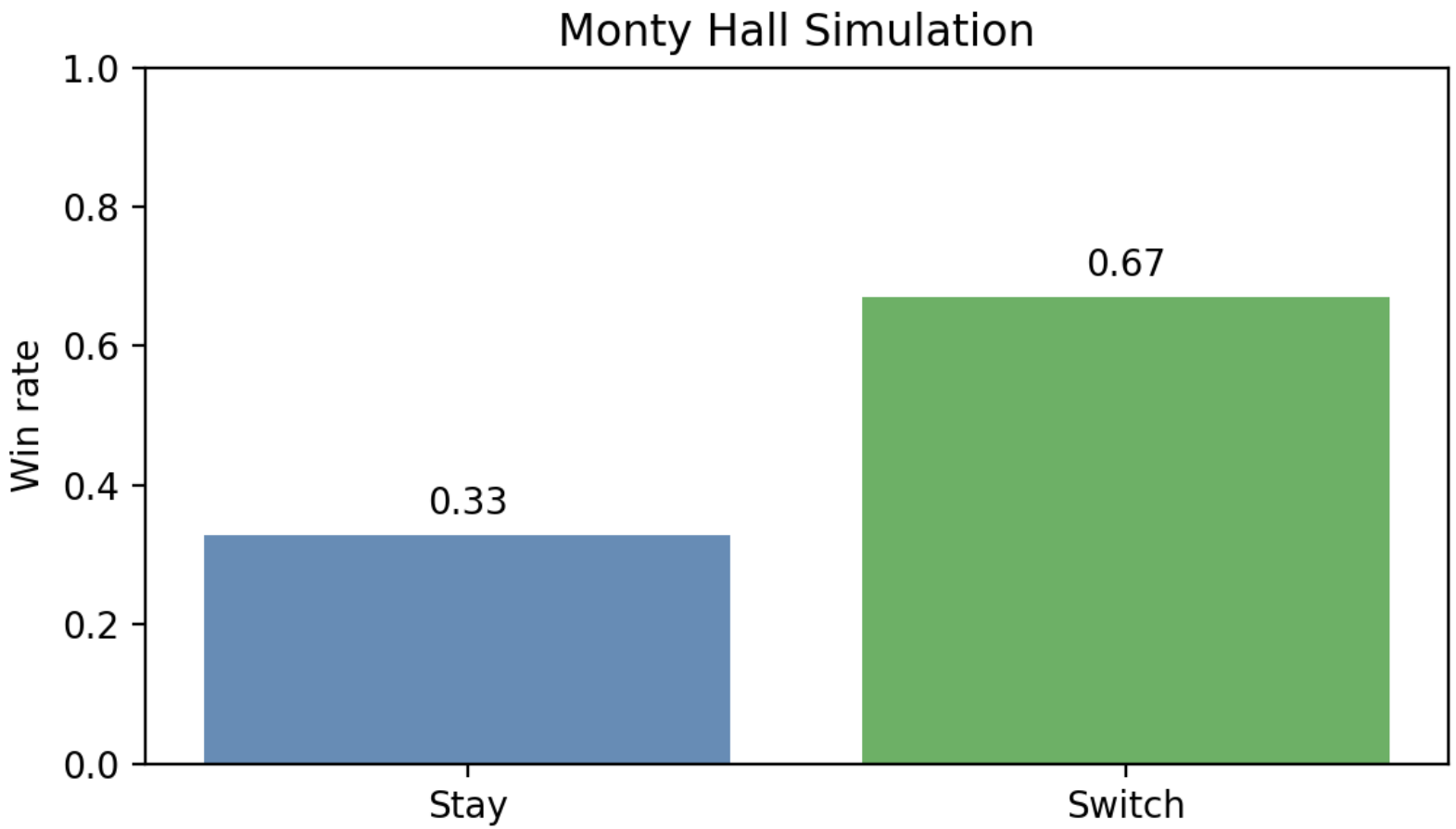
In a class of n students, what's the chance at least two share a birthday? Surprisingly, at $n = 23$ it's about 0.5.



Shows how multiplicative complements and event counting yield unintuitive results.

Catchy: Monty Hall — Switch or Stay?

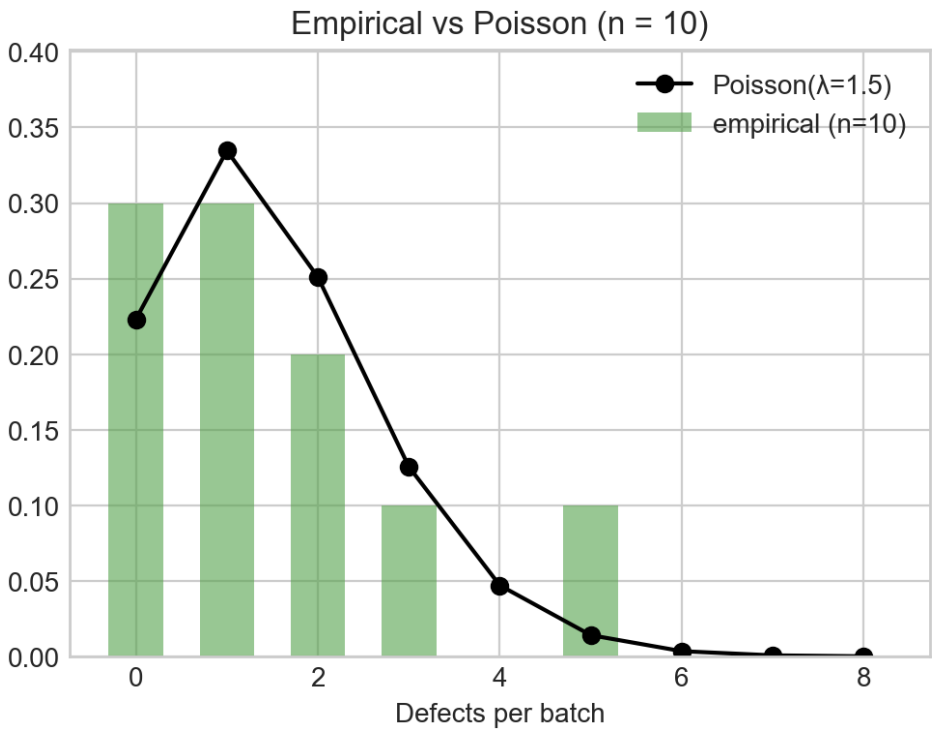
Game show setup: switching doors wins with probability $\approx 2/3$; simulation confirms the model.

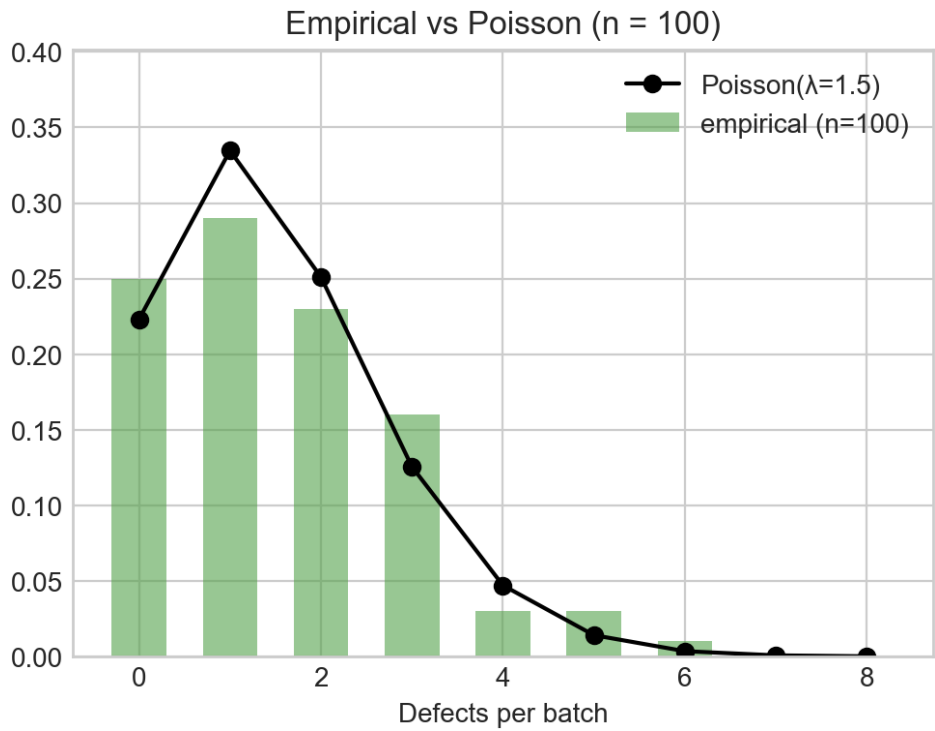


Encodes information and conditional probability; a great motivator for Bayes' rule.

Example: Defects as Poisson Counts

When events are rare and independent, counts in a fixed window often follow Poisson(λ). Compare empirical distribution vs Poisson model.

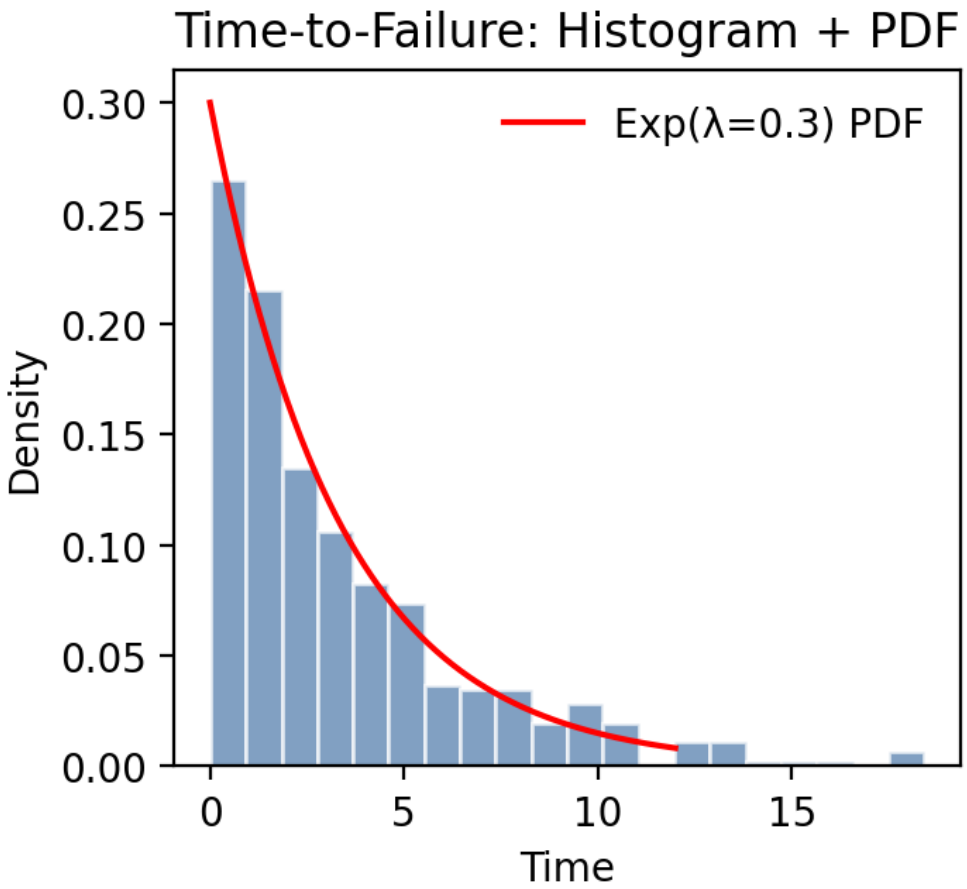


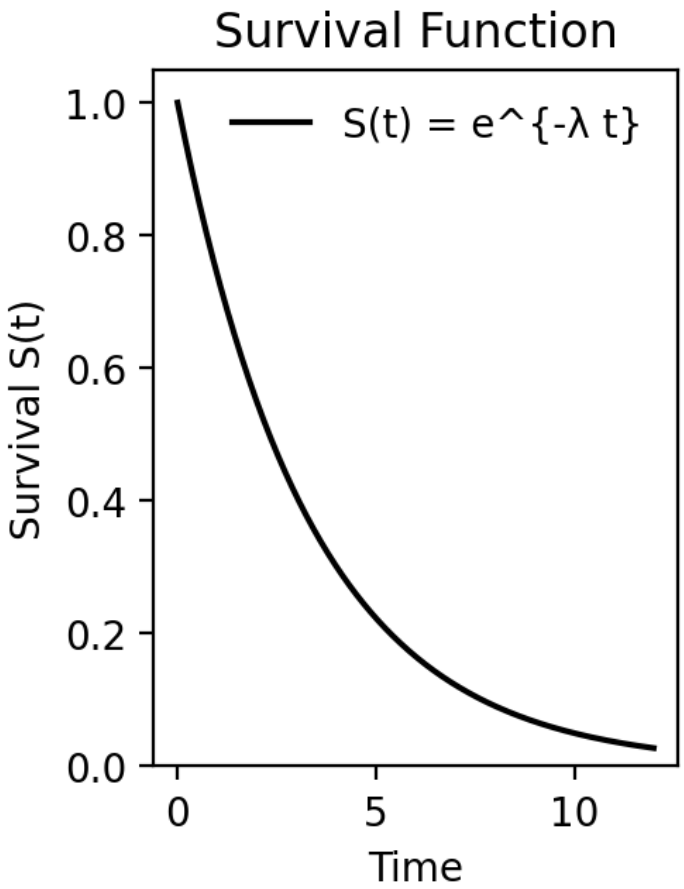


If the fit is reasonable, λ summarizes the average defects per batch, guiding quality control.

Example: Time-to-Failure as Exponential

Waiting times between independent events are often modeled as Exponential(λ). The survival is $S(t) = \mathbb{P}(T > t) = e^{-\lambda t}$.





Useful for reliability, queueing, and risk modeling.

Learning Objectives

- **Define random variables (RVs) and distributions**.
  Purpose: Map real-world uncertainty to mathematical objects and clarify what a distribution encodes.
- **Use PMF/PDF/CDF to compute probabilities/quantiles**.
  Purpose: Turn event and threshold questions into computations to answer "how likely?" and "what cutoff?".
- **Compute expectation, variance, and interpret moments**.
  Purpose: Summarize center and variability to compare models and quantify uncertainty/risk.
- **Understand LLN and CLT (intuition + statements)**.
  Purpose: Explain why averages stabilize and when normal approximations apply, enabling CIs and tests.
- **Connect descriptive statistics to probabilistic modeling**.
  Purpose: Bridge EDA to formal models so assumptions are explicit and limits understood.

Outline

Probability Space – Recap

Definition. A probability space is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where:

- Ω is the sample space; $\mathcal{F} \subseteq 2^{\Omega}$ is a σ-algebra;
- $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a probability measure, $\mathbb{P}(\Omega) = 1$, and $\mathbb{P}$ is countably additive.

Basic properties for events $A, B \in \mathcal{F}$

- Bounds: $0 \leq \mathbb{P}(A) \leq 1$, with $\mathbb{P}(\varnothing) = 0$ and $\mathbb{P}(\Omega) = 1$.
- Complement: $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
- Monotonicity: $A \subseteq B \;\Rightarrow\; \mathbb{P}(A) \leq \mathbb{P}(B)$.
- Union/intersection: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
- Disjoint additivity: if $A \cap B = \varnothing$, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$; extends to countable disjoint unions.

Events are the measurable statements about outcomes: elements of $\mathcal{F}$.

Random Variables and Laws

Random variable

A map X is **measurable** if for every Borel set $B \in \mathcal{B}(\mathbb{R})$, the preimage $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$.

A **real-valued random variable** is a measurable map $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Law (distribution)

For $B \in \mathcal{B}(\mathbb{R})$, the law of X is the pushforward measure

$$\mu_X(B) = \mathbb{P}(X \in B) = \mathbb{P}\big(X^{-1}(B)\big).$$

Types of laws

Discrete (countable support; PMF p_X), absolutely continuous (PDF f_X; $F_X' = f_X$ a.e.), and mixed.

Why Random Variables are Useful?

Same information, but easier to use:

- A random variable $X : \Omega \to \mathbb{R}$ induces a distribution $\mathbb{P}_X$.
- Working with X or $\mathbb{P}_X$ encodes the same probabilistic information.
- $\Rightarrow$ Random variables let us write *equations* instead of handling sets and events.

Advantages of random variables:

- *Algebraic manipulation:* can form $Y = 2X + 3$, $Z = X + Y$, etc.
- *Compounding:* joint r.v.s describe multiple processes easily.
- *Analytical tools:* enable expectations, variances, mgfs, characteristic functions.
- *Modeling:* regression, Bayesian models, stochastic processes all rely on r.v.s.

Key message: Random variables = distributions in disguise, but far more convenient for *calculation, composition, and modeling*.

Discrete random variables: PMF

What is a discrete RV?

X takes values in a countable set $A = \{x_1, x_2, \dots\}$. Probability mass sits on points; probabilities *add up* across values.

PMF (probability mass function)

The PMF $p_X(x) = \mathbb{P}(X = x)$ assigns a weight to each $x \in A$ with $p_X(x) \geq 0$ and $\sum_{x \in A} p_X(x) = 1$. For any measurable B,

$$\mathbb{P}(X \in B) = \sum_{x \in A \cap B} p_X(x).$$

Interpretation: p_X is a table (or bar chart) of *weights* at allowed values.

Examples: Bernoulli(p): $p(1) = p$, $p(0) = 1 - p$. Poisson(λ): $p(k) = e^{-\lambda}\, \lambda^k / k!$, $k = 0, 1, 2, \dots$.

Discrete random variables: CDF

CDF (cumulative distribution function)

$F_X(x) = \mathbb{P}(X \le x)$ accumulates mass up to x. For discrete X, F_X is a non-decreasing *step* function; at each support point x the jump size equals $p_X(x)$. Between support points F_X is flat.

Using the CDF

For $a < b$,

$$\mathbb{P}(a < X \le b) = F_X(b) - F_X(a), \qquad \mathbb{P}(X = x) = F_X(x) - F_X(x^-) = p_X(x).$$

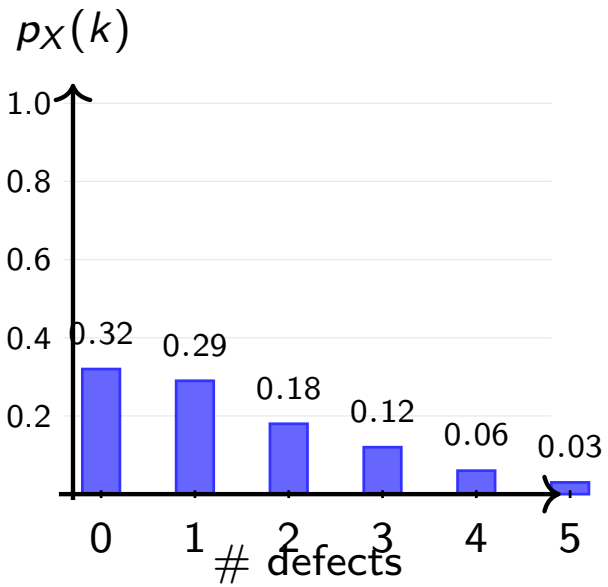
Properties: non-decreasing, right-continuous, $\lim_{x\to-\infty} F_X = 0$, $\lim_{x\to\infty} F_X = 1$.

When is the CDF meaningful?

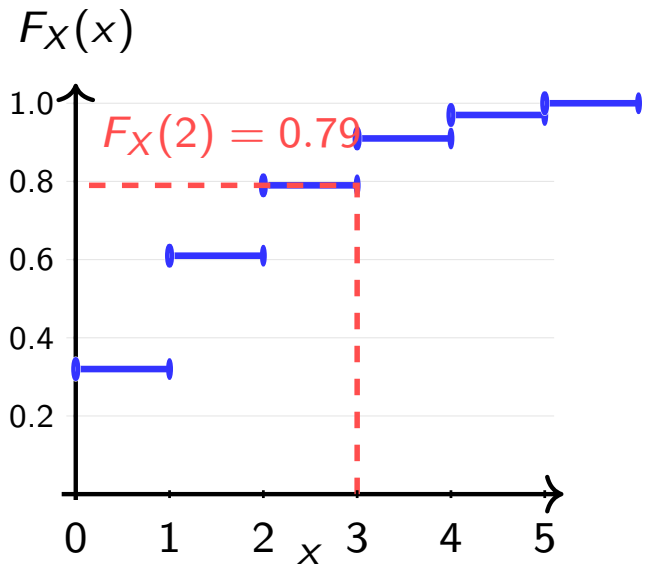
A CDF requires a meaningful order on the support. For nominal categories (unordered), prefer category probabilities (PMF); only define cumulative quantities once an inherent or agreed ordinal scale exists.

When the CDF *does* make sense (ordered counts)

PMF (Defect count)



CDF (cumulative steps)



Natural ordering

Values $0 < 1 < 2 < \ldots$ have clear meaning.
Events like $\{X \leq 2\}$ are interpretable.

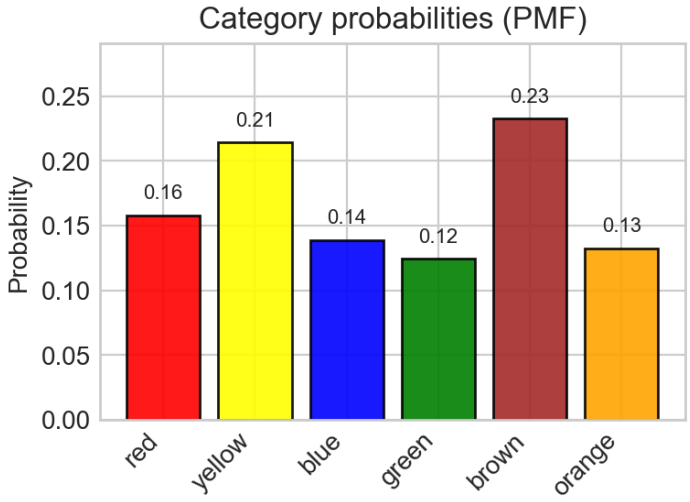
Meaningful cumulation

$F_X(k) = \mathbb{P}(X \leq k)$ answers concrete questions about defects.

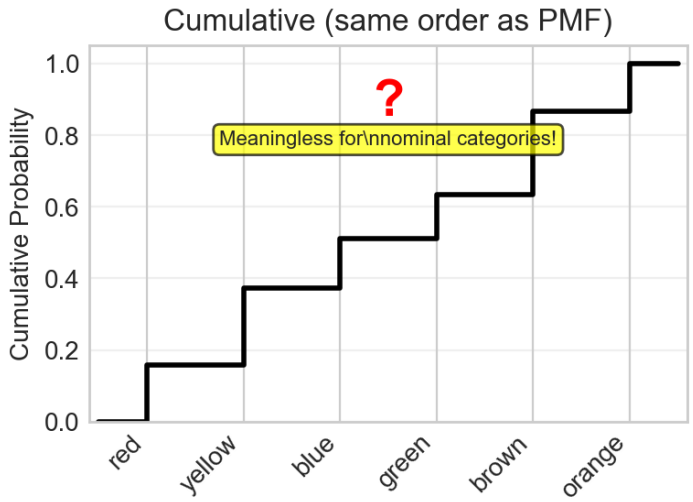
Key insight: Natural ordering $\Rightarrow$ meaningful cumulative probabilities

When CDF is *meaningless* (nominal categories)

PMF for M&M colors



Arbitrary "CDF" (alphabetical)



No natural order

Colors are *nominal categories*.
No meaningful $<$ relationship exists.

Why this fails

Different orderings give different "CDFs".
$\mathbb{P}(\text{color} \leq \text{Green})$ is nonsense!

Key insight: No natural order $\Rightarrow$ CDF is arbitrary and meaningless

Continuous random variables: PDF

What is a continuous RV?

No atoms: $\mathbb{P}(X = a) = 0$ for every $a \in \mathbb{R}$. Probabilities live on intervals/sets, not single points.

PDF (probability density function)

A nonnegative function f_X with $\int_{\mathbb{R}} f_X(x)\,dx = 1$ such that for any measurable B, $\mathbb{P}(X \in B) = \int_B f_X(x)\,dx$. For $a < b$,

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

Intuition: f_X is a *height*; probability is *area under the curve*. The PDF itself is not a probability.

Examples: Uniform(a, b): $f = 1/(b-a)$ on $[a, b]$; Exponential(λ): $\lambda e^{-\lambda x}$ for $x \geq 0$; Normal(μ, σ^2): bell-shaped.

Continuous random variables: CDF

CDF (cumulative distribution function)

$F_X(x) = \mathbb{P}(X \le x) = \int_{-\infty}^{x} f_X(t)\,dt$. For $a < b$,

$$\mathbb{P}(a < X \le b) = F_X(b) - F_X(a) = \int_a^b f_X(x)\,dx.$$

Properties: non-decreasing, right-continuous, $\lim_{x\to-\infty} F_X = 0$, $\lim_{x\to\infty} F_X = 1$.

Remember

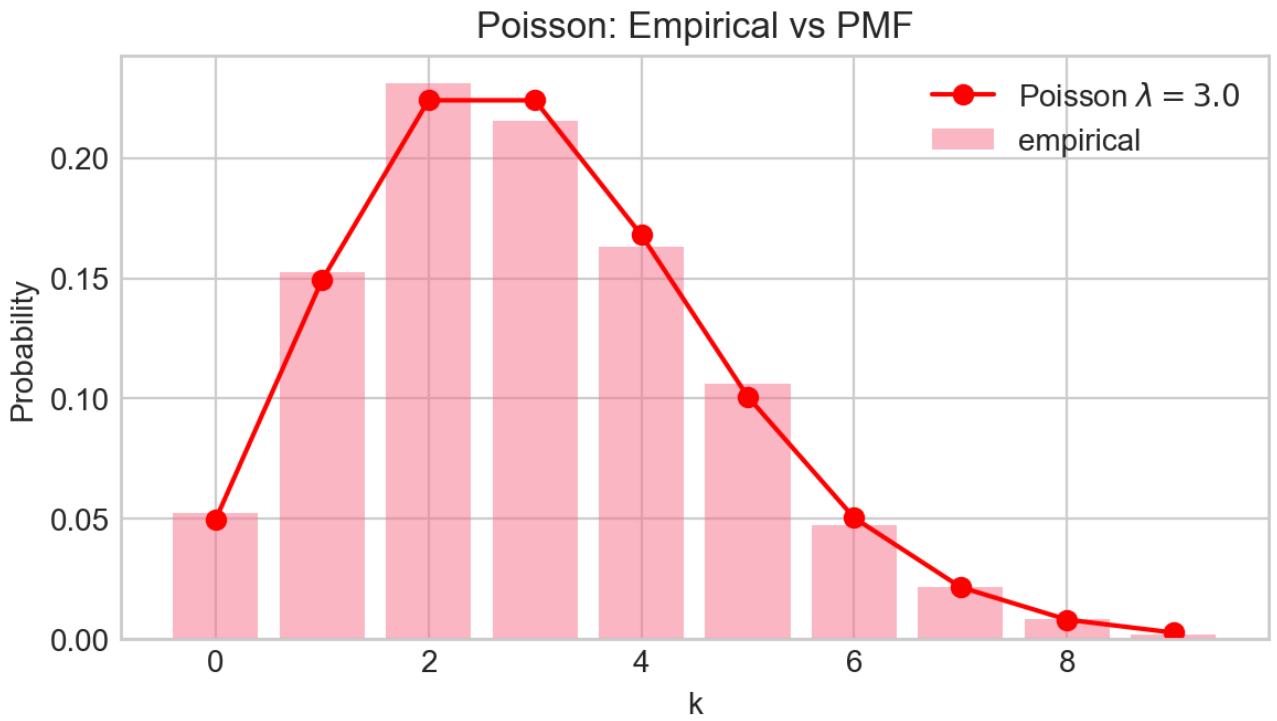
The PDF is the *derivative* of the CDF: $f_X(x) = F_X'(x)$ almost everywhere. The CDF is bounded in $[0, 1]$; the PDF is a *slope* and need not be ≤ 1 — probabilities come from *areas* under f_X, not heights.

Quantiles

For $q \in (0, 1)$, a q-quantile is any x with $F_X(x) \ge q$. If F_X is strictly increasing, the q-quantile is $F_X^{-1}(q)$.

Key Distributions (Discrete)

- Bernoulli(p): $\Pr(X = 1) = p$, $\Pr(X = 0) = 1 - p$, $\mathbb{E}[X] = p$, $\mathrm{Var}(X) = p(1 - p)$
- Binomial(n, p): sum of n iid Bernoulli; $\mathbb{E}[X] = np$, $\mathrm{Var}(X) = np(1 - p)$
- Poisson(λ): $\Pr(X = k) = e^{-\lambda}\lambda^k/k!$, $\mathbb{E}[X] = \lambda$, $\mathrm{Var}(X) = \lambda$

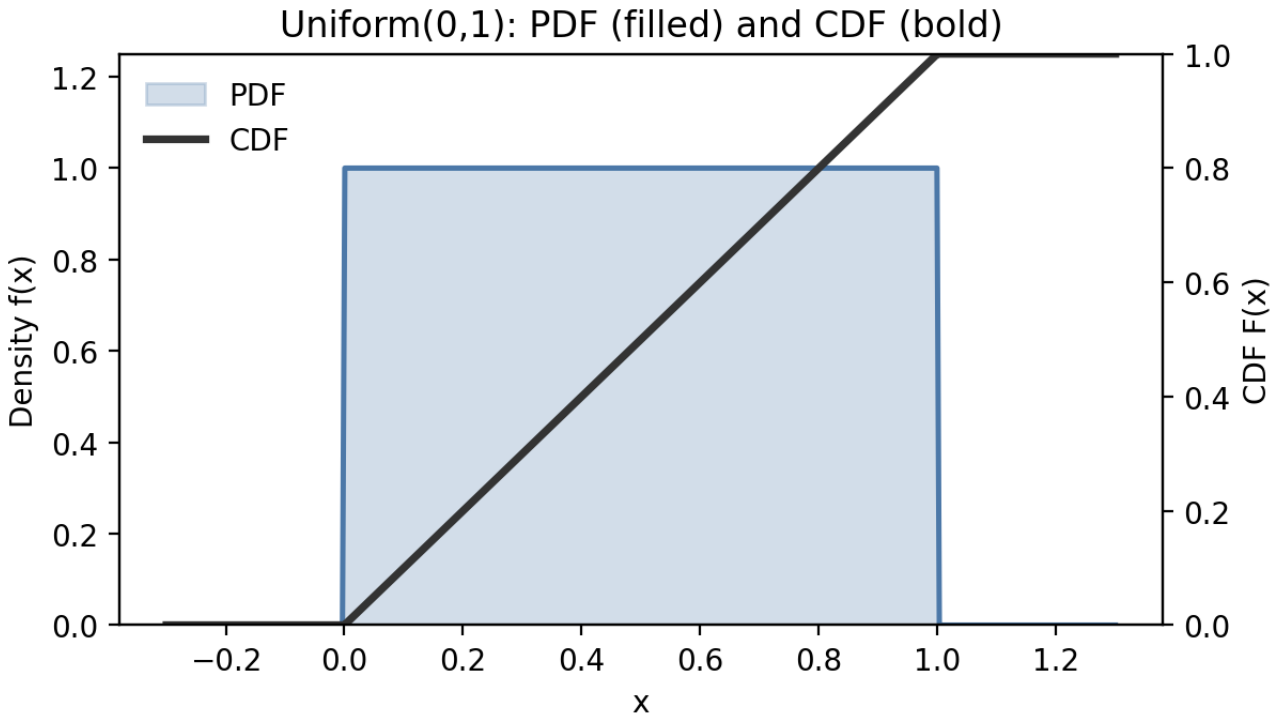


Uniform (a, b)

Formulas

$$\text{PDF: } f(x) = \begin{cases} \dfrac{1}{b-a}, & a \le x \le b, \\ 0, & \text{otherwise} \end{cases} \qquad \text{CDF: } F(x) = \begin{cases} 0, & x < a, \\ \dfrac{x-a}{b-a}, & a \le x \le b, \\ 1, & x > b \end{cases}$$

$$\text{Mean: } \mathbb{E}[X] = \frac{a+b}{2} \qquad \text{Variance: } \operatorname{Var}(X) = \frac{(b-a)^2}{12}$$



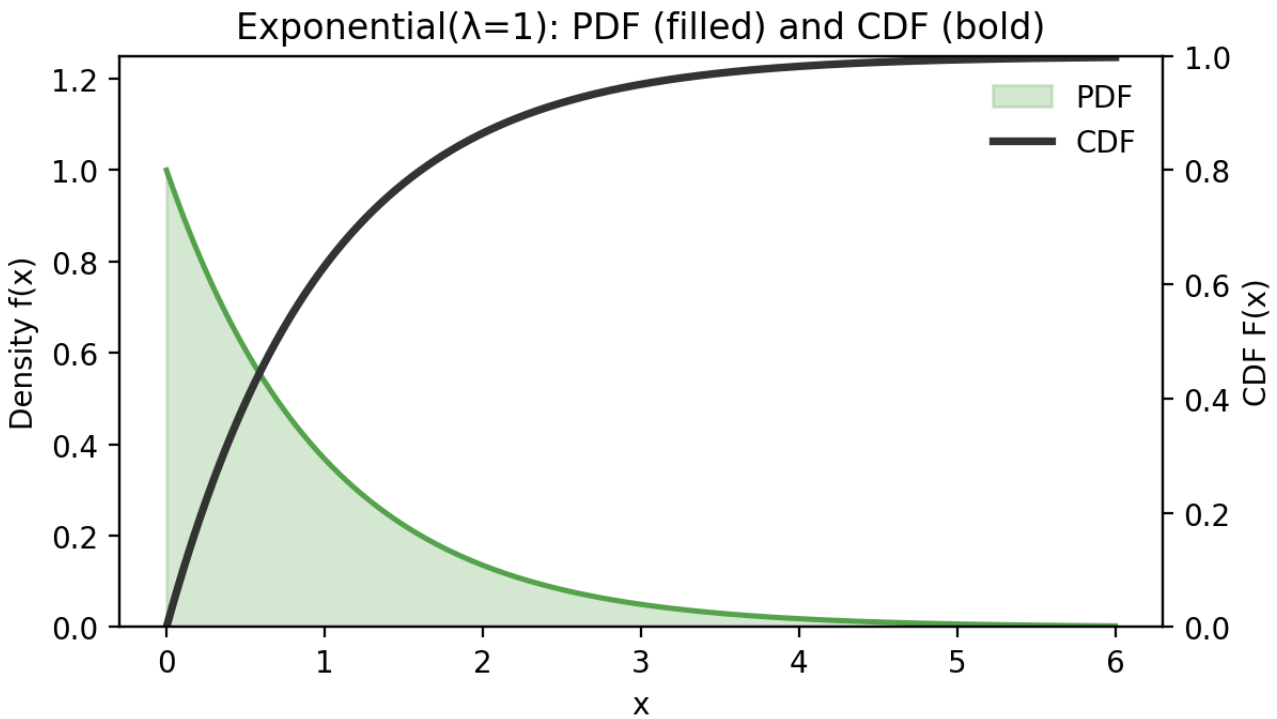
Exponential (λ)

Formulas

Support: $x \geq 0$ PDF: $f(x) = \lambda e^{-\lambda x}$ CDF: $F(x) = 1 - e^{-\lambda x}$

Memoryless: $\mathbb{P}(X > s + t \mid X > s) = e^{-\lambda t}$

Mean: $\mathbb{E}[X] = \dfrac{1}{\lambda}$ Variance: $\mathrm{Var}(X) = \dfrac{1}{\lambda^2}$

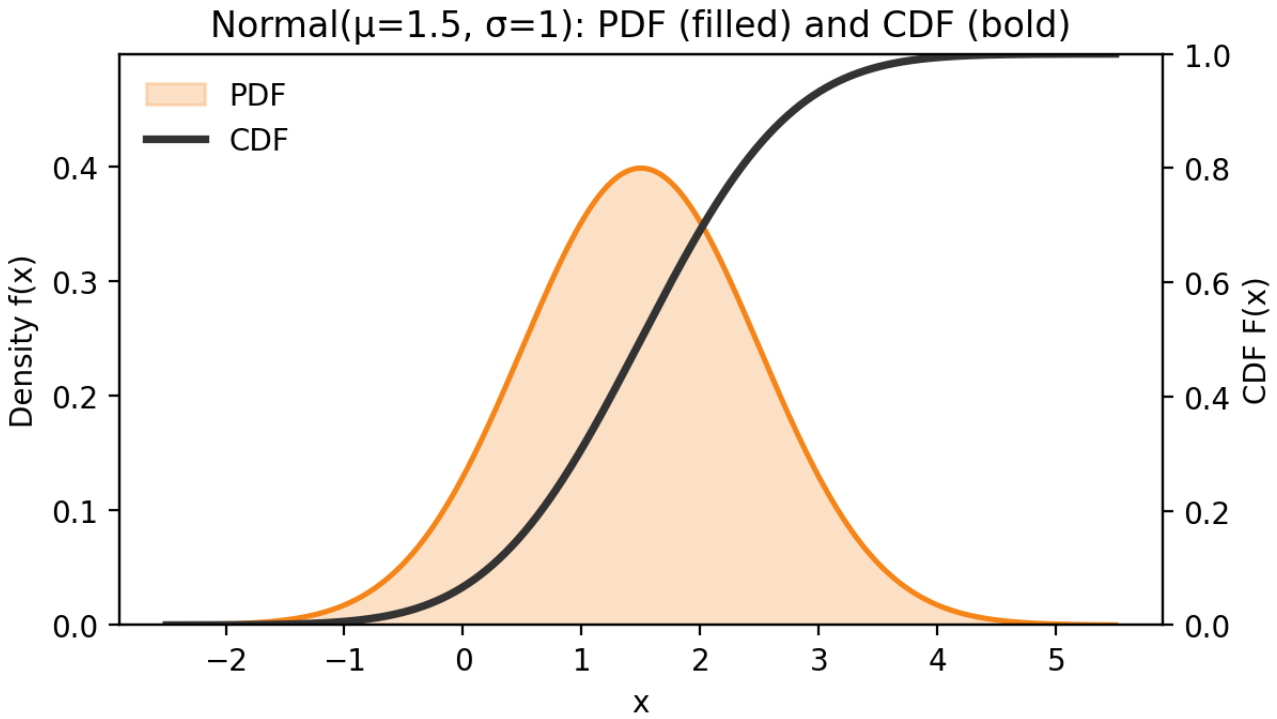


Normal (μ, σ^2)

Formulas

PDF: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$ CDF: $F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$ (no closed form)

Mean: $\mathbb{E}[X] = \mu$ Variance: $\mathrm{Var}(X) = \sigma^2$ Standardization: $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$



Outline

What is a Statistic?

Definition of a Statistic

- Let $X_1, \ldots, X_n$ be a sample of random variables taking values in $\mathcal{X}$.
- A **statistic** is a measurable function

$$T : \mathcal{X}^n \longrightarrow \mathcal{Y},$$

  where typically $\mathcal{Y} = \mathbb{R}^k$.
- For a realization $(x_1, \ldots, x_n)$, the statistic produces $T(x_1, \ldots, x_n)$.

Intuition:

- A statistic is a *summary* of the sample.
- It compresses n observations into a simpler object (number or vector).

Examples:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad \max(X_1, \ldots, X_n).$$

Descriptive vs Inferential Statistics

Descriptive statistics

- *Purpose:* summarize and present the features of the observed dataset; no claims beyond the data.
- *Mathematical view:* summary functionals $S(\mathbf{x})$ of a sample $\mathbf{x} = (x_1, \ldots, x_n)$ capturing location, spread, shape, and dependence.

Examples

- Mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, median, mode.
- Variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, standard deviation, range, IQR.
- Counts/proportions (categorical), histograms, boxplots, scatter plots.

Inferential statistics

- *Purpose:* generalize from the sample to the population using probability models; quantify uncertainty.
- *Outputs:* estimates with standard errors, confidence intervals, hypothesis tests, predictive statements.

Examples

- Confidence intervals; hypothesis tests (e.g., t-test); p-values.
- Parametric/GLM/Regression inference; model comparison/selection.
- Bootstrap-based uncertainty quantification.

Key distinction: *descriptive* = what this dataset looks like, *inferential* = what we can say about the population (with uncertainty).

Expectation via measurable functions

Discrete case (PMF p_X)

For measurable g with $\mathbb{E}[|g(X)|] < \infty$,

$$\mathbb{E}[g(X)] = \sum_{x \in \text{supp}(X)} g(x)\, p_X(x) \quad \text{(absolute convergence)}.$$

In particular, take $g(x) = x$ to get $\mathbb{E}[X] = \sum_x x\, p_X(x)$.

Continuous case (PDF f_X)

If X admits a density f_X w.r.t. Lebesgue measure and $\mathbb{E}[|g(X)|] < \infty$,

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x)\, f_X(x)\, dx.$$

In general, with law μ_X, $\mathbb{E}[g(X)] = \int g\, d\mu_X$. Setting $g(x) = x$ yields $\mathbb{E}[X]$.

Note: $\mathbb{E}[g(X)]$ is defined whenever $\int |g|\, d\mu_X < \infty$.

Expectation and variance: definitions and properties

Definitions

For integrable X, the mean of the random variable is defined as

$$\mathbb{E}[X] = \int_{\Omega} X \, d\mathbb{P} = \int_{\mathbb{R}} x \, d\mu_X(x)$$

If $\mathbb{E}[X^2] < \infty$, then the variance of the random variable is defined as

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

Algebraic properties

- Linearity: $\mathbb{E}[aX + b] = a\,\mathbb{E}[X] + b$ and $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.
- Affine scaling: $\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X)$; non-negativity: $\mathrm{Var}(X) \geq 0$ with equality iff X is a.s. constant.
- Second-moment identity: $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ when $\mathbb{E}[X^2] < \infty$.

Higher-Order Moments: Skewness and Kurtosis

Definitions (when finite)

Let $\mu = \mathbb{E}[X]$, $\sigma^2 = \mathrm{Var}(X)$, and central moments $\mu_k = \mathbb{E}\big[(X-\mu)^k\big]$. Skewness $\gamma_1 := \mu_3/\sigma^3$; Kurtosis $\beta_2 := \mu_4/\sigma^4$; Excess kurtosis $\gamma_2 := \beta_2 - 3$.

Relevance and interpretation

- Skewness (γ_1): *asymmetry*. $\gamma_1 > 0$ right-skewed; $\gamma_1 < 0$ left-skewed; $\gamma_1 = 0$ for symmetric laws.
- Kurtosis (β_2) and excess (γ_2): tail weight/peakedness. Normal: $\beta_2 = 3$ hence $\gamma_2 = 0$.
- Practice: diagnose non-normality, tail risk, outlier propensity; both are outlier-sensitive (kurtosis especially).
- Sampling: for sample means, standardized skew decays $\propto n^{-1/2}$ and excess kurtosis $\propto n^{-1}$ (i.i.d., finite moments).

Example: Bernoulli (p)

Random variable and summary

Let $X \sim \text{Bernoulli}(p)$ with support $\{0, 1\}$, $\mathbb{P}(X = 1) = p$, $\mathbb{P}(X = 0) = 1 - p$. Then $\mathbb{E}[X] = p$ and $\text{Var}(X) = p(1 - p)$.

Computation.

$$\mathbb{E}[X] = \sum_{x \in \{0,1\}} x \, p_X(x) = 0 \cdot (1 - p) + 1 \cdot p = p.$$

$$\mathbb{E}[X^2] = \sum_{x \in \{0,1\}} x^2 \, p_X(x) = 0^2 \cdot (1 - p) + 1^2 \cdot p = p \quad (\text{since } X^2 = X).$$

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = p - p^2 = p(1 - p).$$

Example: Exponential (λ)

Random variable and summary

Let $X \sim \text{Exponential}(\lambda)$ with density $f_X(x) = \lambda e^{-\lambda x}\,\mathbf{1}_{\{x \geq 0\}}$. Then $\mathbb{E}[X] = \dfrac{1}{\lambda}$ and $\text{Var}(X) = \dfrac{1}{\lambda^2}$.

Computation.

$$\mathbb{E}[X] = \int_0^\infty x\, f_X(x)\, dx = \int_0^\infty x\, \lambda e^{-\lambda x}\, dx = \frac{1}{\lambda} \quad \text{(IBP or } \Gamma \text{ function)}.$$

$$\mathbb{E}[X^2] = \int_0^\infty x^2\, f_X(x)\, dx = \int_0^\infty x^2\, \lambda e^{-\lambda x}\, dx = \frac{2}{\lambda^2}.$$

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Example: Normal (μ, σ^2)

Random variable and summary

Let $X \sim \mathcal{N}(\mu, \sigma^2)$ with density $f_X(x) = \dfrac{1}{\sigma\sqrt{2\pi}} \exp\Big(- \dfrac{(x-\mu)^2}{2\sigma^2} \Big)$. Then $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

Computation. Let $Z = \dfrac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$. Then

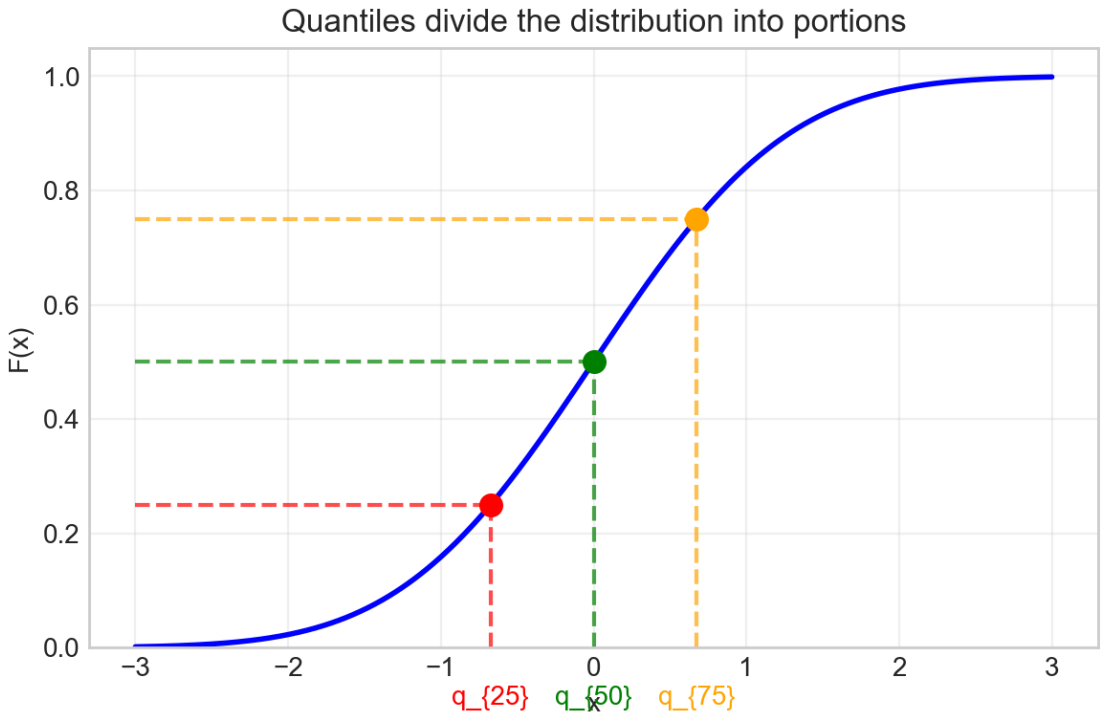
$$\mathbb{E}[X] = \mathbb{E}[\mu + \sigma Z] = \mu + \sigma\, \mathbb{E}[Z] = \mu,$$
$$\mathrm{Var}(X) = \mathrm{Var}(\mu + \sigma Z) = \sigma^2\, \mathrm{Var}(Z) = \sigma^2 \quad (\mathrm{Var}(Z) = 1).$$

Definition of Quantiles

Quantile

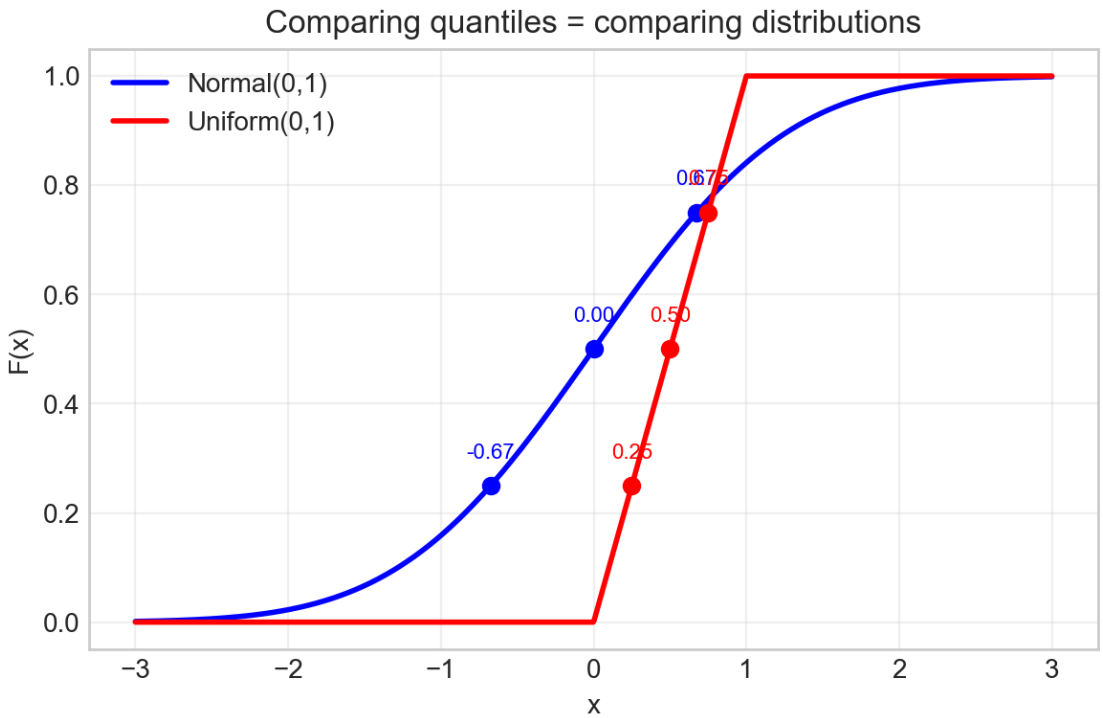
The p-quantile is $q_p = \inf\{x : F(x) \geq p\}$, where F is the CDF.

Intuition: "the value below which a fraction p of the data lies."



Why are quantiles useful?

- Quantiles summarize distributional shape (center, spread, tails).
- Robust to outliers (median, quartiles).
- Natural way to compare distributions.

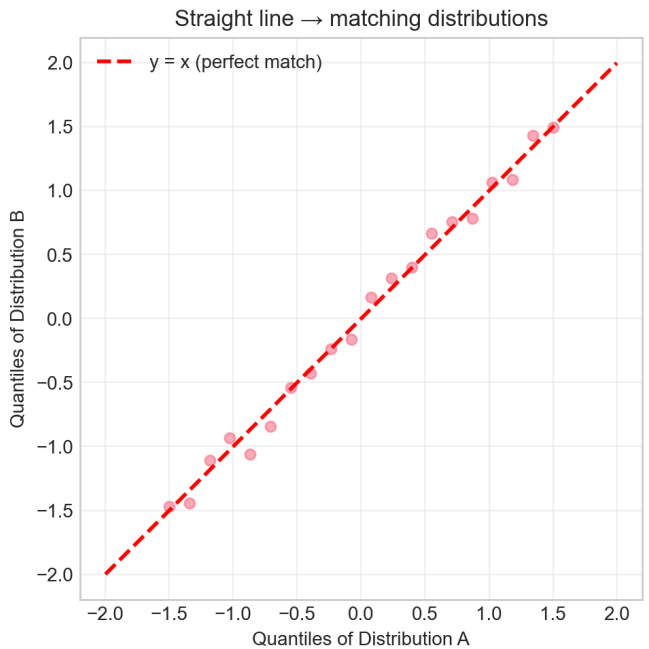


Definition of QQ-Plot

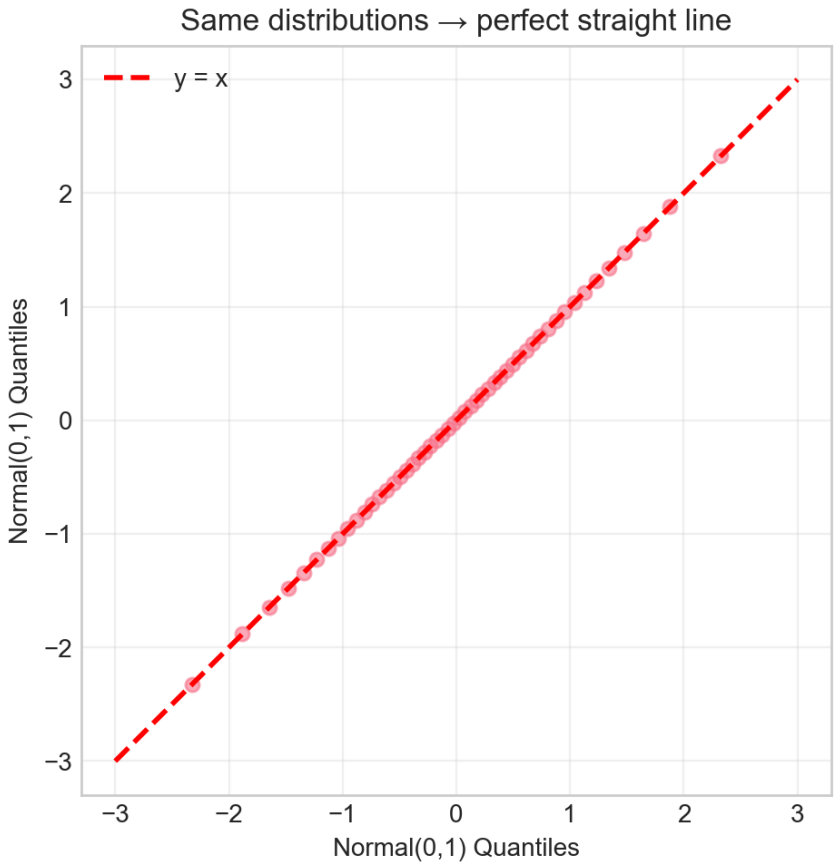
QQ-Plot

QQ-plot = plot quantiles of one distribution against quantiles of another.

- If distributions are the same $\rightarrow$ points lie on line $y = x$.
- Deviations show differences in location, scale, or tails.

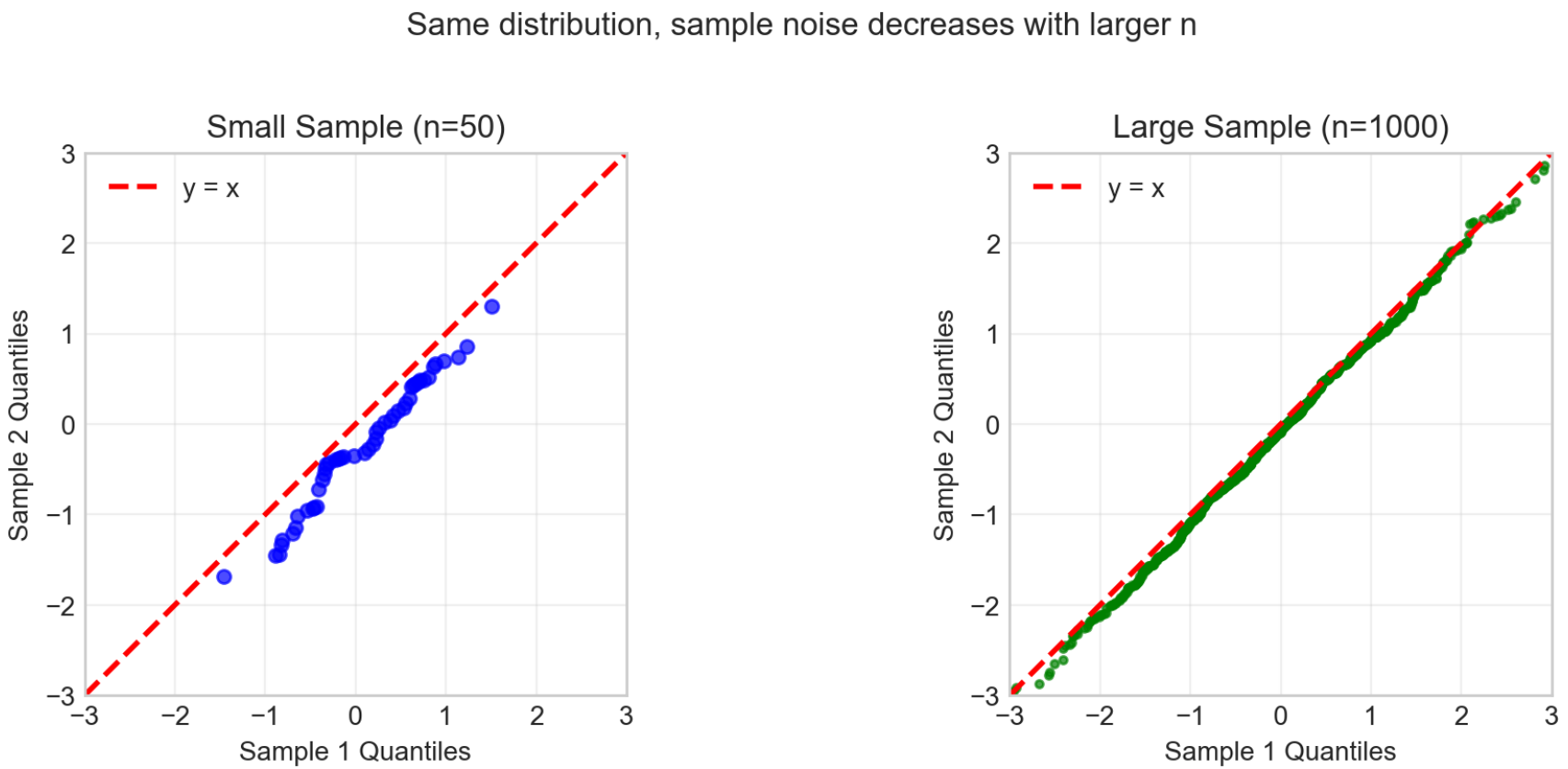


QQ-Plot: Normal vs Normal (Theoretical)



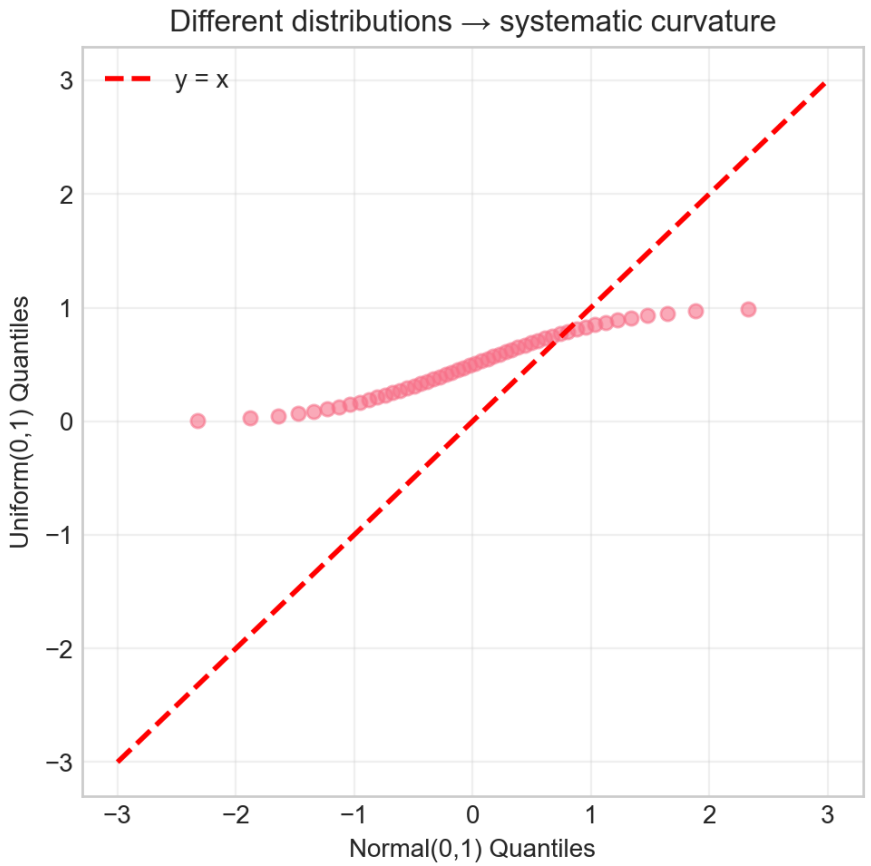
Same distributions → perfect straight line.

QQ-Plot: Normal vs Normal (Empirical)



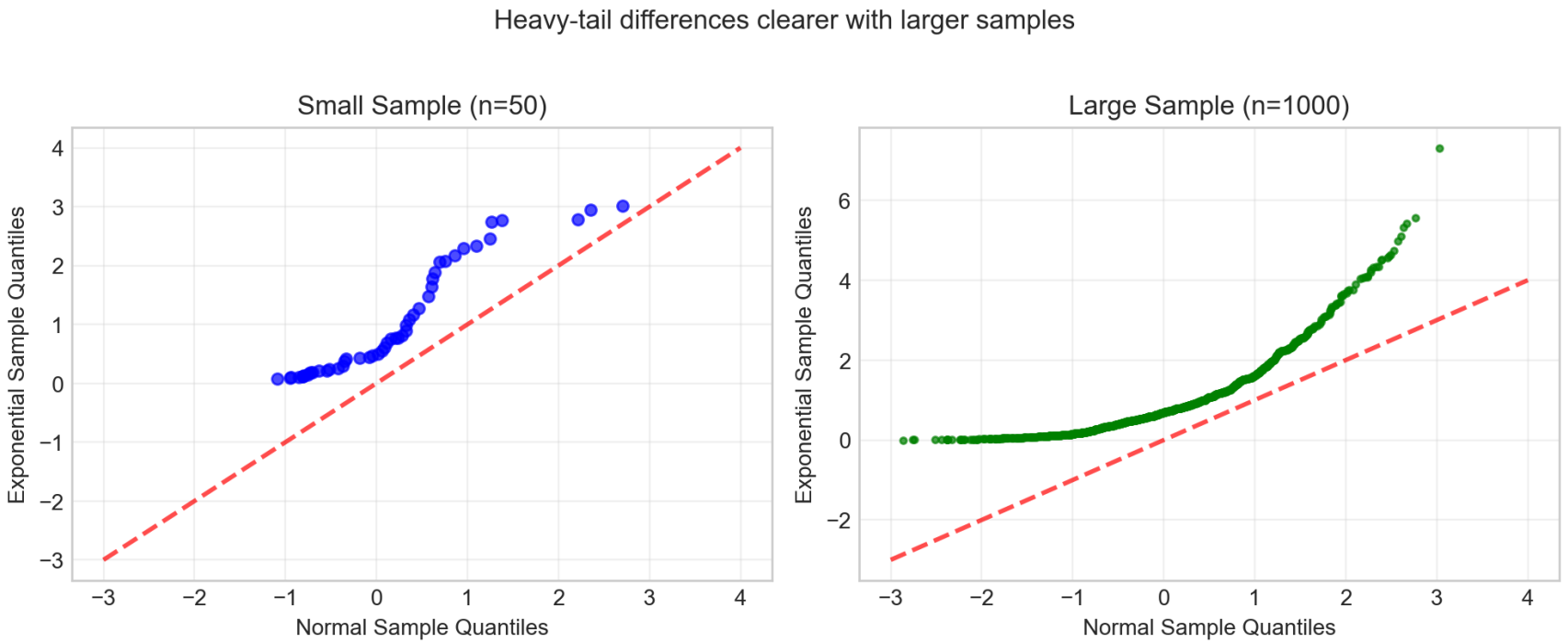
Same distribution: larger samples reduce noise, approach theoretical line.

QQ-Plot: Normal vs Uniform (Theoretical)



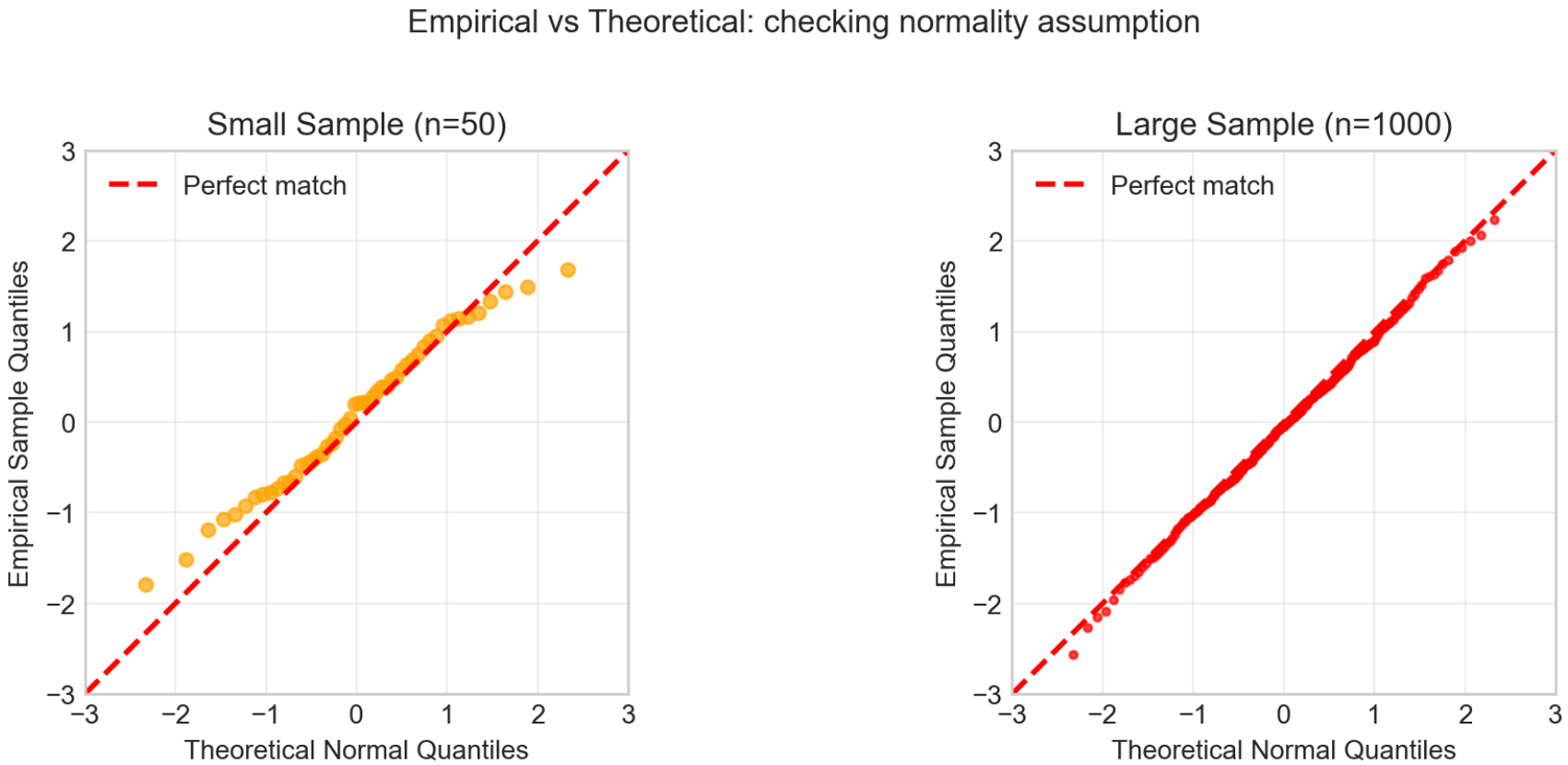
Different distributions → systematic curvature.

QQ-Plot: Normal vs Exponential (Empirical)



Different distributions: larger samples reveal clearer patterns.

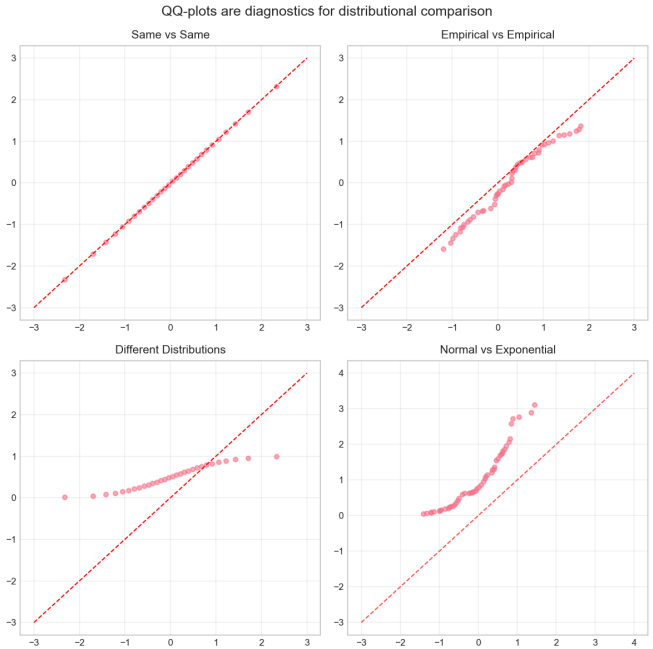
QQ-Plot: Empirical vs Theoretical Normal



Testing normality assumption: straight line = good fit to normal distribution.

Why QQ-Plots Matter

- Tool to check if a dataset matches a theoretical distribution (empirical vs theoretical).
- Tool to compare two datasets (empirical vs empirical).
- Straight line $\rightarrow$ matching distributions.
- Curvature or deviation $\rightarrow$ differences in location, spread, or tails.



Moment Generating Function (MGF)

Definition

For a real-valued random variable X, the moment generating function (when finite) is

$$M_X(t) = \mathbb{E}\left[e^{tX}\right], \qquad t \in D := \{t \in \mathbb{R} : \mathbb{E}[e^{tX}] < \infty\}.$$

Key properties and relevance

- Normalization: $M_X(0) = 1$. If M_X exists on a neighborhood of 0, it uniquely determines the law of X.
- Moments: whenever $\mathbb{E}[|X|^k] < \infty$, $M_X^{(k)}(0) = \mathbb{E}[X^k]$.
- Affine and sums: $M_{aX+b}(t) = e^{bt}\, M_X(at)$; if $X \perp\!\!\!\perp Y$, then $M_{X+Y}(t) = M_X(t)\, M_Y(t)$.
- Uses: identify distributions (e.g., Normal $M_X(t) = \exp(\mu t + \frac{1}{2}\sigma^2 t^2)$), compute moments, and obtain distributions of sums of independent variables.
- Caveat: may fail to exist near 0 for heavy-tailed laws (e.g., lognormal has $M_X(t) = \infty$ for $t > 0$).

Characteristic Function

Definition

The characteristic function of X is

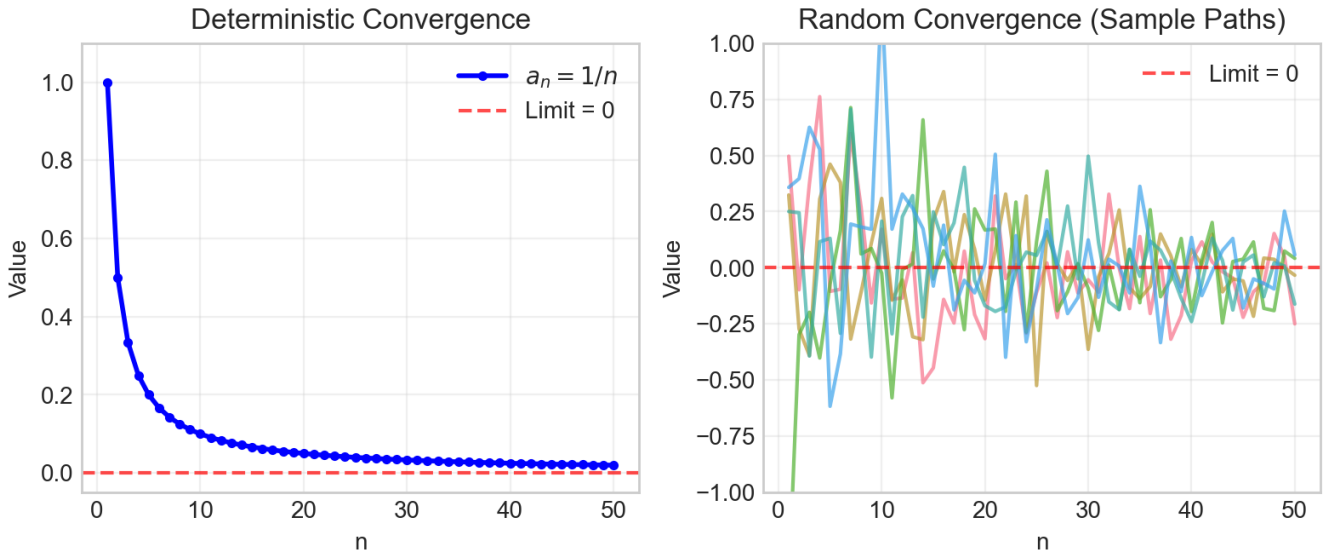
$$\varphi_X(t) = \mathbb{E}\left[e^{itX}\right], \qquad t \in \mathbb{R}.$$

Key properties and relevance

- Always exists; $\varphi_X(0) = 1$, $|\varphi_X(t)| \leq 1$; uniformly continuous and positive-definite.
- Uniqueness and inversion: φ_X uniquely determines the law; inversion formulas recover the CDF/PDF under mild regularity.
- Moments: if $\mathbb{E}[|X|^k] < \infty$, then $\varphi_X^{(k)}(0) = i^k\,\mathbb{E}[X^k]$.
- Affine and sums: $\varphi_{aX+b}(t) = e^{ibt}\,\varphi_X(at)$; if $X \perp\!\!\!\perp Y$, then $\varphi_{X+Y}(t) = \varphi_X(t)\,\varphi_Y(t)$.
- Limit theory: Lévy continuity theorem links pointwise convergence of φ_{X_n} to $X_n \xrightarrow{\mathcal{D}} X$; core tool for CLT and stable laws.

Why Convergence Matters in Probability

- Deterministic sequences: $a_n \to a$ means terms get arbitrarily close to a fixed limit.
- For random variables (X_n), multiple notions of *getting closer* exist.
- Hierarchy: a.s. $\Rightarrow$ in probability $\Rightarrow$ in distribution (no converses in general).

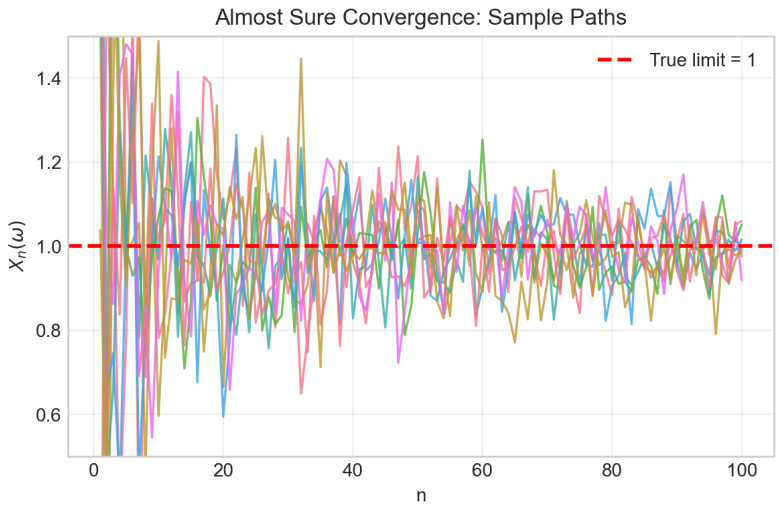


Convergence Almost Surely

Definition

$$X_n \xrightarrow{\text{a.s.}} X \quad \Longleftrightarrow \quad \mathbb{P}\Big(\lim_{n\to\infty} X_n = X\Big) = 1.$$

- Intuition: *pathwise* convergence for almost every outcome ω.
- Strongest of the three: implies in probability (hence in distribution).



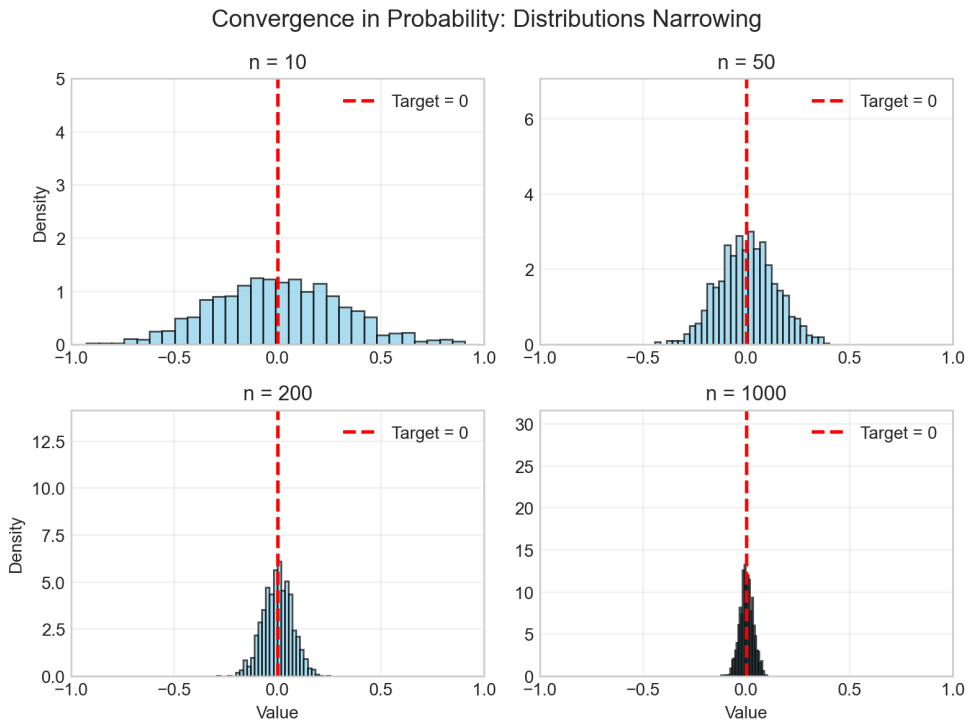
Example: $X_n = X + Y\,\mathbf{1}_A$ with $\mathbb{P}(A) = p$ small and fixed, and $Y \to 0$ deterministically. Then $X_n \xrightarrow{\text{a.s.}} X$ when the pathwise perturbations vanish almost surely.

Convergence in Probability

Definition

$$X_n \xrightarrow{\mathbb{P}} X \quad \Longleftrightarrow \quad \forall\, \varepsilon > 0,\ \mathbb{P}(|X_n - X| > \varepsilon) \to 0.$$

- Intuition: large deviations become rare; individual paths may still oscillate.
- Weaker than a.s.; stronger than in distribution (in general settings).



Example: $X_n = \mathbf{1}_{\{U \le 1/n\}}$ with $U \sim \mathrm{Uniform}(0,1)$. Then $X_n \xrightarrow{\mathrm{P}} 0$ but $X_n \not\xrightarrow{\text{a.s.}} 0$.

Convergence in Distribution

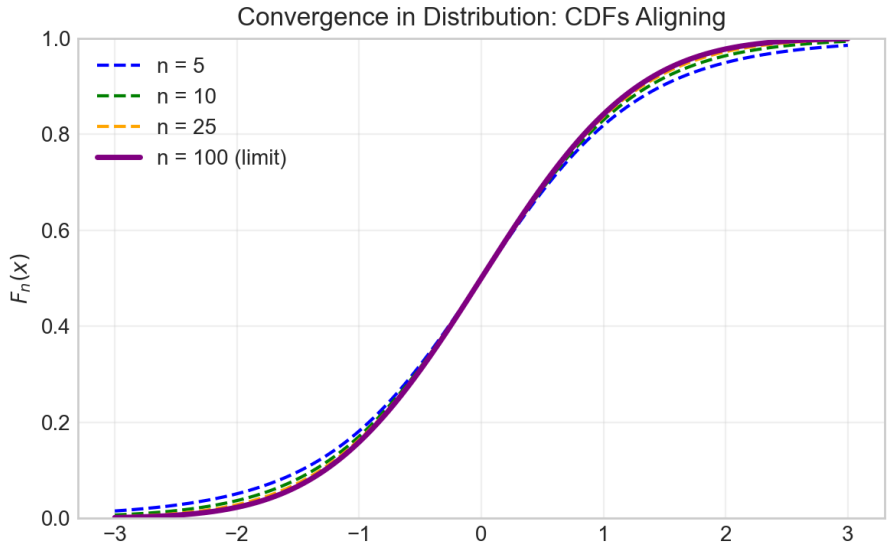
CDF-based definition

$$X_n \xrightarrow{d} X \quad \Longleftrightarrow \quad F_{X_n}(t) \to F_X(t) \text{ at all continuity points of } F_X.$$

Portmanteau (equivalent)

$$X_n \xrightarrow{d} X \quad \Longleftrightarrow \quad \mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)] \quad \text{for all bounded, } \textit{continuous } f.$$

- Converging *shape* of the distribution; samples need not get close pointwise.
- Do not replace "continuous" by "measurable"; that would demand a stronger mode.



Summary & Intuition Map

Mode	Definition	Meaning	Visual	Implies
a.s.	$\mathbb{P}(\lim X_n = X) = 1$	Pathwise for a.e. outcome	paths settle	in P
in P	$\mathbb{P}(\lvert X_n - X\rvert > \varepsilon) \to 0$	Large deviations rare	histograms narrow	in $\mathcal{D}$
in $\mathcal{D}$	$F_{X_n} \to F_X$ at cont. pts	Laws converge	CDFs align	—

Key Hierarchy

Almost surely $\Rightarrow$ In probability $\Rightarrow$ In distribution

Takeaway

Stronger modes give more pathwise closeness; weaker modes only guarantee distribution-level agreement.

Outline

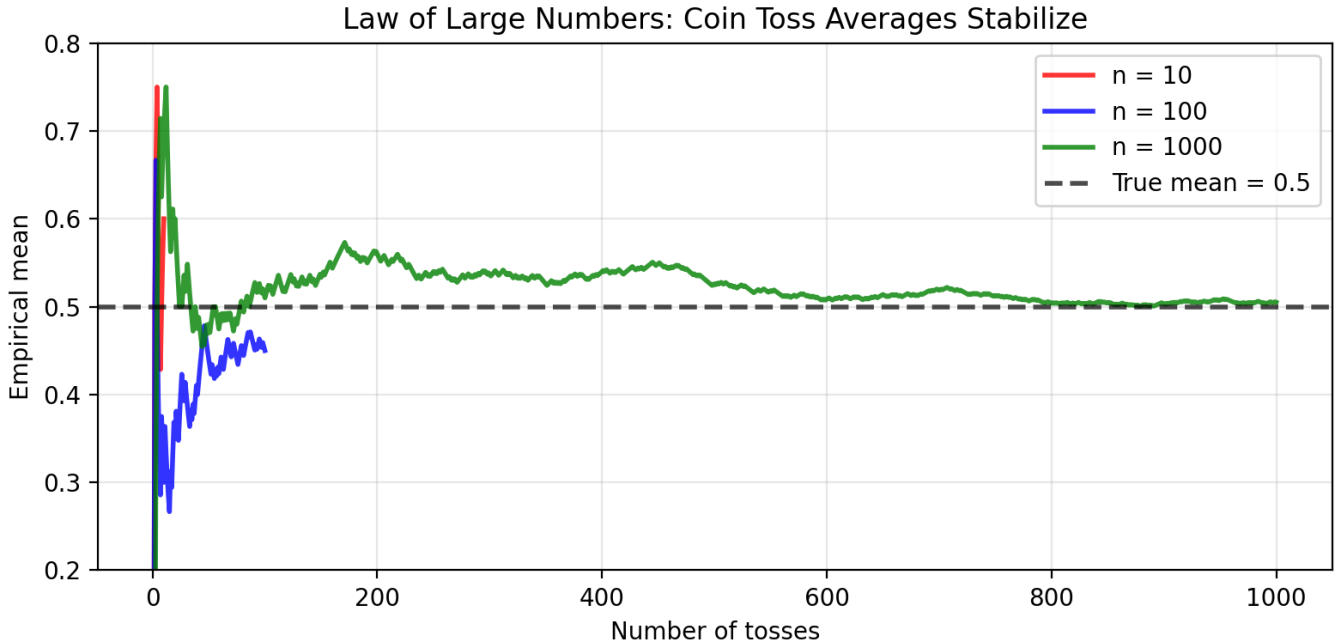
Why do averages stabilize?

Individual outcomes are random and fluctuate:

- A single coin toss: completely unpredictable (heads or tails).
- But the average of many tosses becomes remarkably predictable.

Key insight:

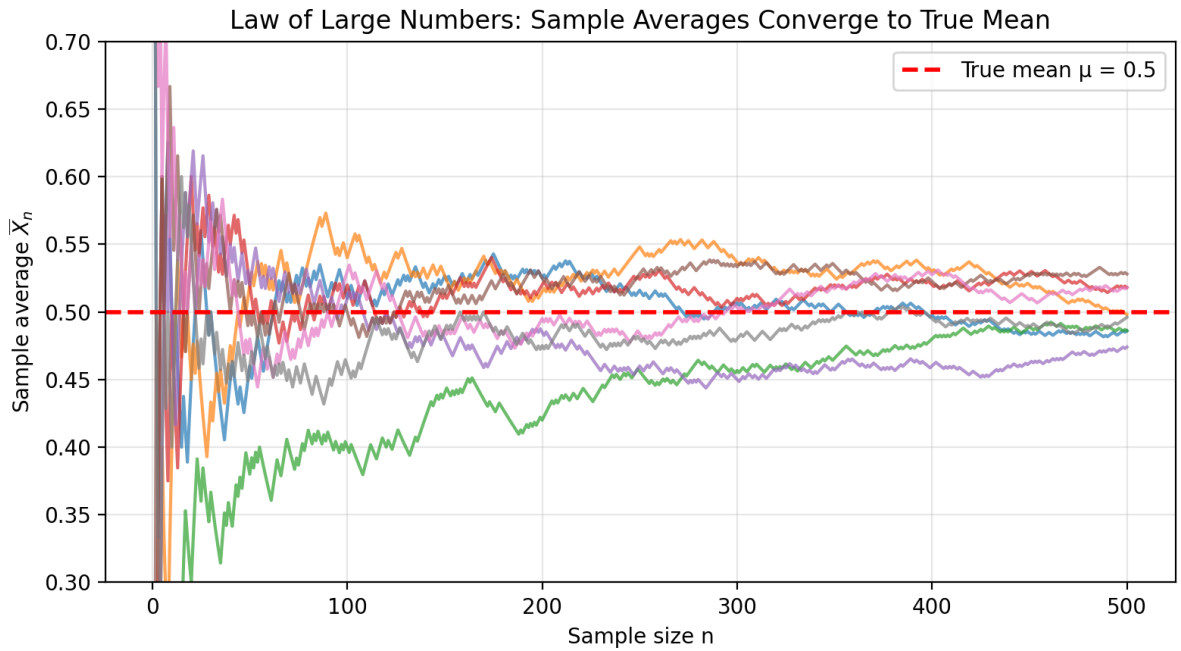
- Averaging across many samples cancels out randomness.
- The average becomes more predictable as sample size grows.
- This stabilization is guaranteed by mathematical laws.



Empirical averages get closer to the true mean

Multiple sample paths of Bernoulli(0.5):

- Early averages fluctuate wildly between different sample paths.
- As n increases, all paths converge toward the true mean $\mu = 0.5$.
- Individual variation decreases; central tendency emerges.



Key message

"The law of large numbers tells us this stabilization is guaranteed."

Law of Large Numbers (Informal)

Informal statement

"When we take more and more independent observations, the sample average converges to the true mean."

In everyday language:

- "Data averages out."
- Random fluctuations cancel when you have enough data.
- Larger samples $\Rightarrow$ more reliable estimates.

Why this matters for statistics:

- Justifies using sample means to estimate population parameters.
- Explains why polling works (with sufficient sample size).
- Foundation for Monte Carlo methods and simulation.

Law of Large Numbers (Formal)

Formal statement (Weak LLN)

Let $X_1, X_2, \ldots$ be i.i.d. random variables with expectation $\mu = \mathbb{E}[X_1]$.

Define the sample mean: $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Then: $\overline{X}_n \xrightarrow{\mathbb{P}} \mu$.

Translation to plain language:

- For any $\varepsilon > 0$: $\mathbb{P}(|\overline{X}_n - \mu| > \varepsilon) \to 0$ as $n \to \infty$.
- The probability of large deviations from the true mean vanishes.
- Sample averages concentrate around the population mean.

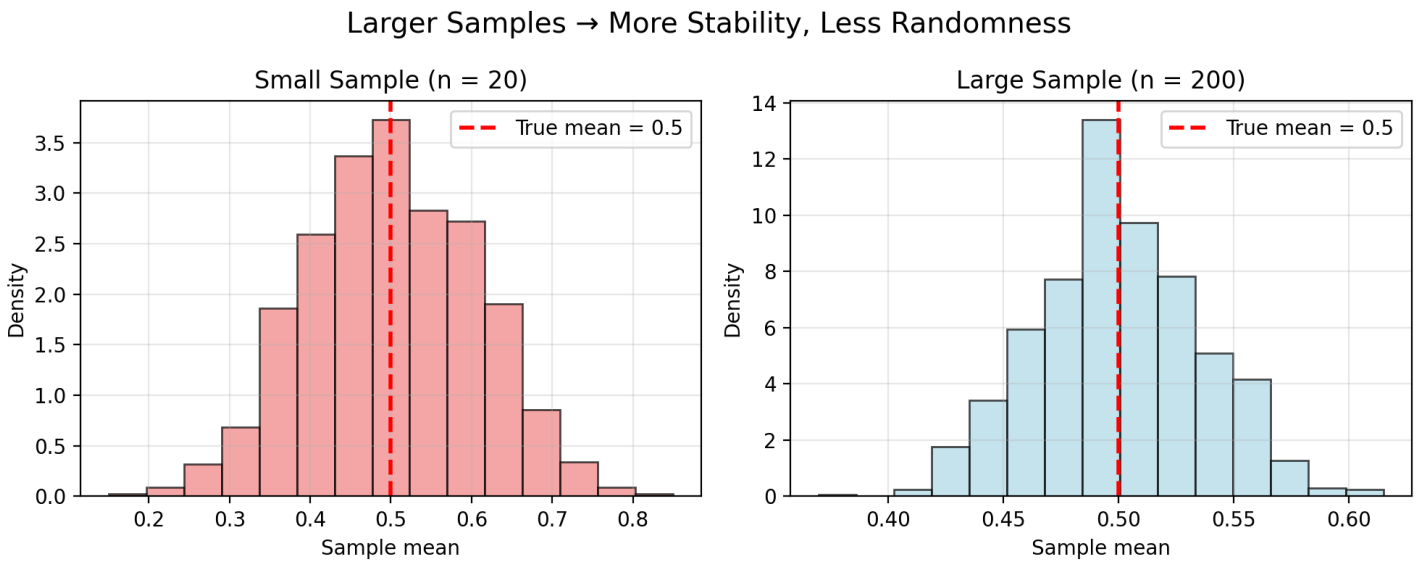
Why LLN Matters for Applied Statistics

Core justification for statistical inference:

- Sample averages (and estimators) approximate population averages.
- Foundation for estimation theory.
- More data $\Rightarrow$ more reliable results.

Key principles:

- Larger samples $\Rightarrow$ more stability, less randomness.
- Random quantities become "deterministic-looking" with enough data.
- Basis for confidence intervals, hypothesis tests, and prediction.

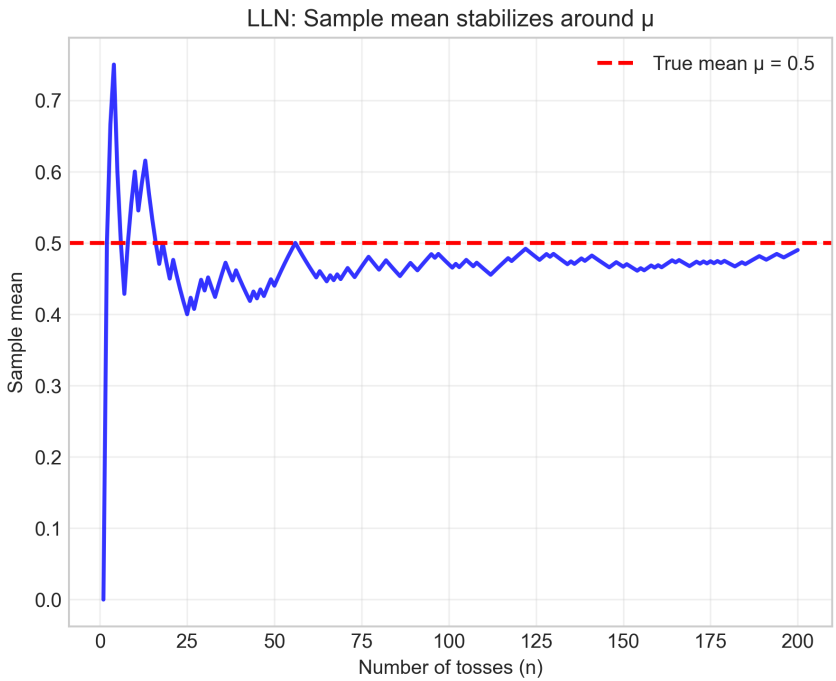


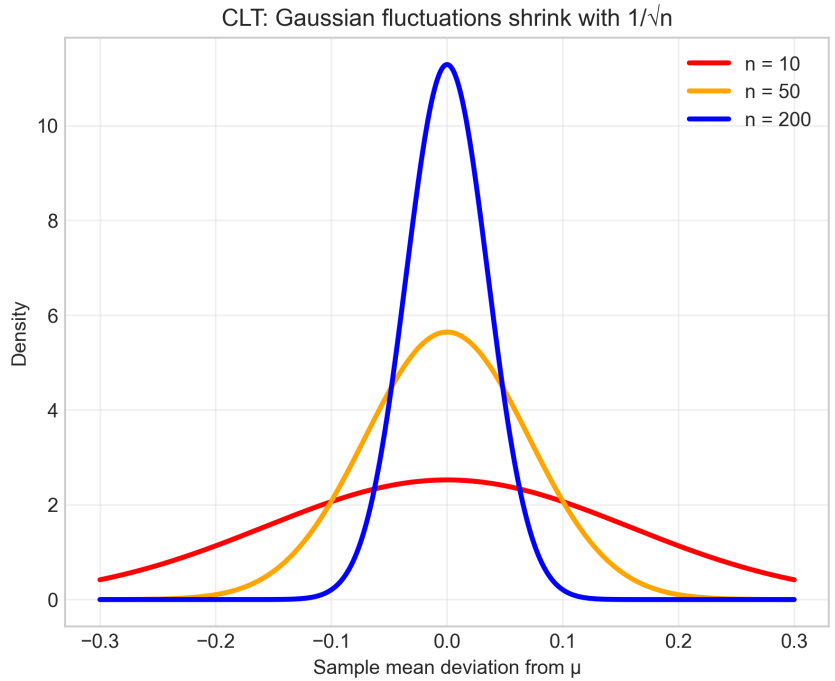
Small sample vs large sample: histogram comparison showing reduced variability.

From stabilization to distributional shape

Motivation & Intuition

- LLN: sample mean $\overline{X}_n$ converges to the true mean μ.
- CLT: answers *how fast* (scaling by $\sqrt{n}$) and *with what shape* (Gaussian).





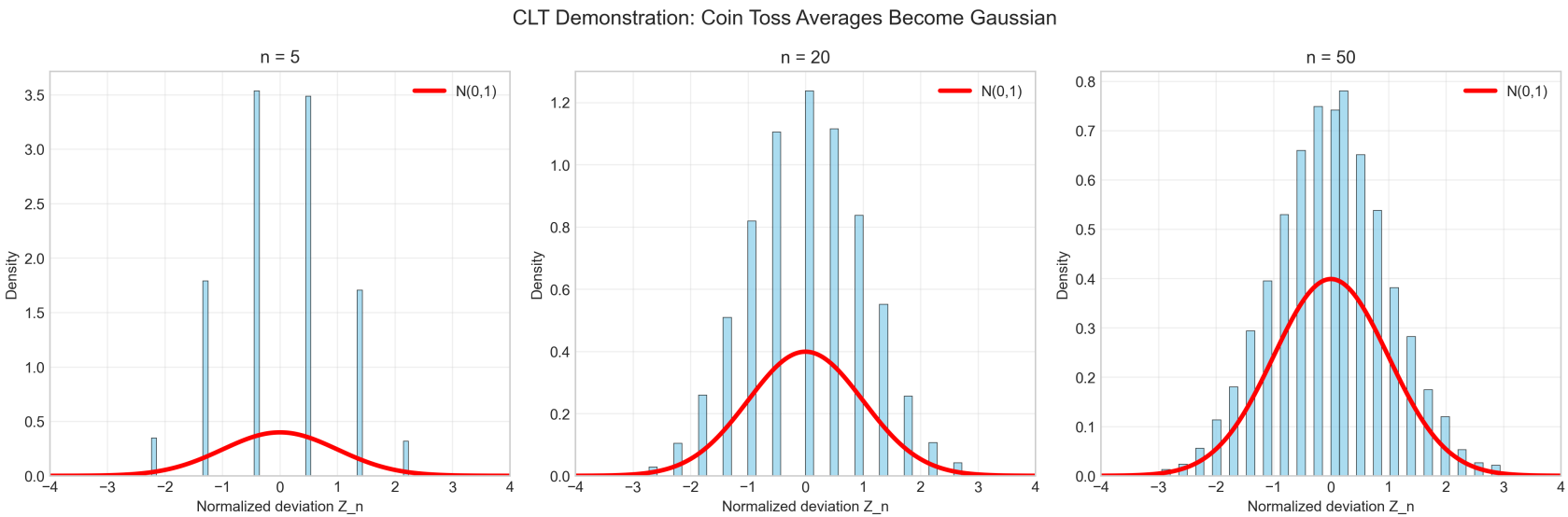
Key insight: LLN = convergence to mean. CLT = Gaussian fluctuations shrinking with $1/\sqrt{n}$.

Coin toss averages become Gaussian

Visual Demonstration

Experiment: Toss a fair coin (0/1), repeat many times.

- Simulate 10,000 repetitions of sample means for $n = 5, 20, 50$.
- Compute normalized deviations: $Z_n = \sqrt{n}(\overline{X}_n - 0.5)$.



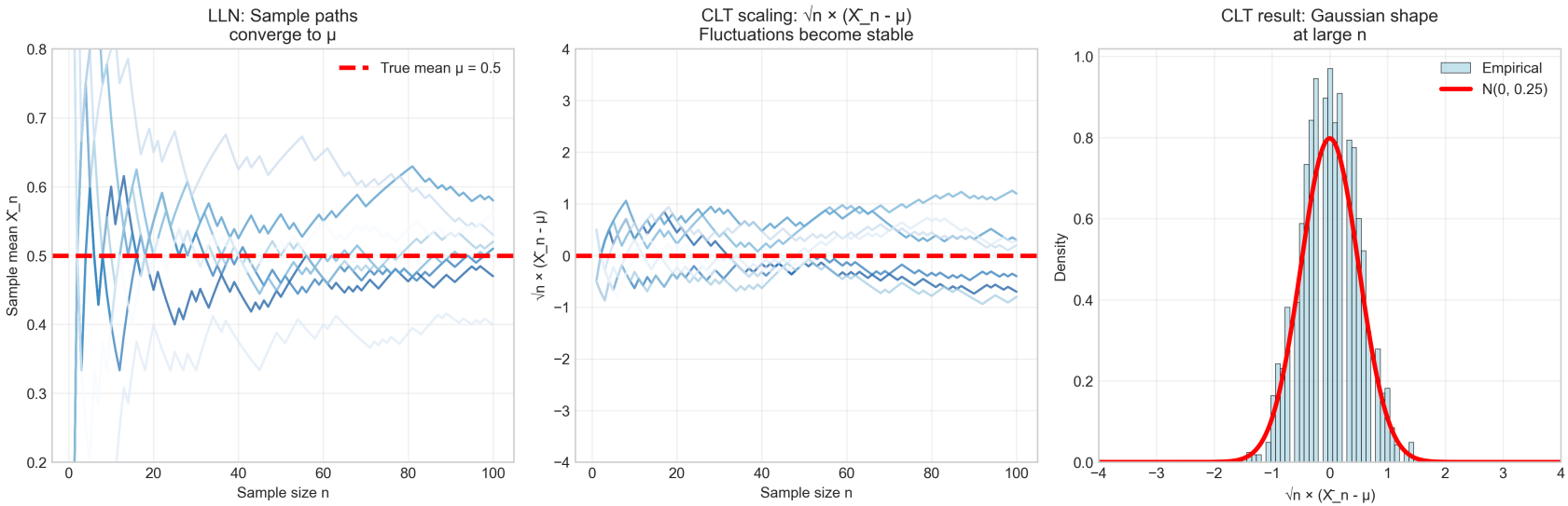
As n grows, Z_n looks increasingly Gaussian.

Central Limit Theorem (Informal)

Informal Statement

"LLN says averages stabilize. CLT says that if we zoom in at the scale $\sqrt{n}$, fluctuations are Gaussian."

Key insight: CLT provides both *speed* and *shape* of convergence.



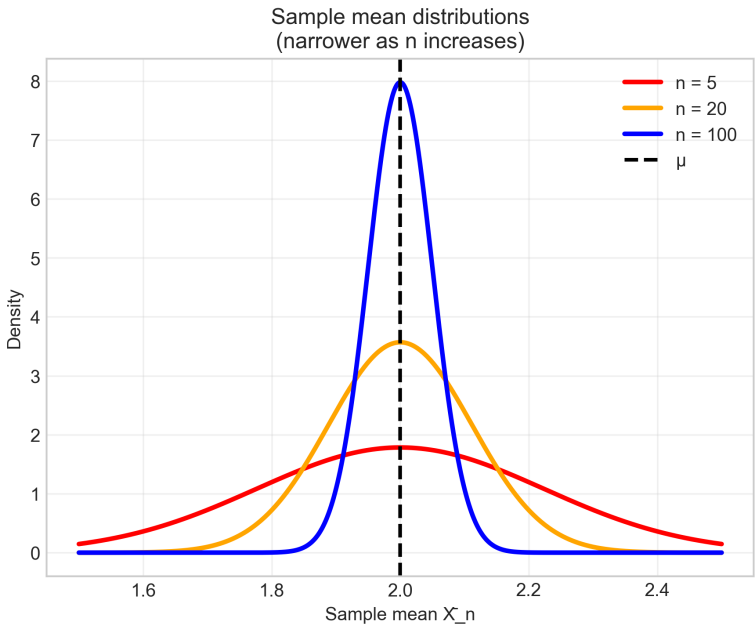
Left: Multiple sample paths converge to μ. **Middle:** After $\sqrt{n}$-scaling, fluctuations stabilize. **Right:** At large n, the scaled deviations are Gaussian.

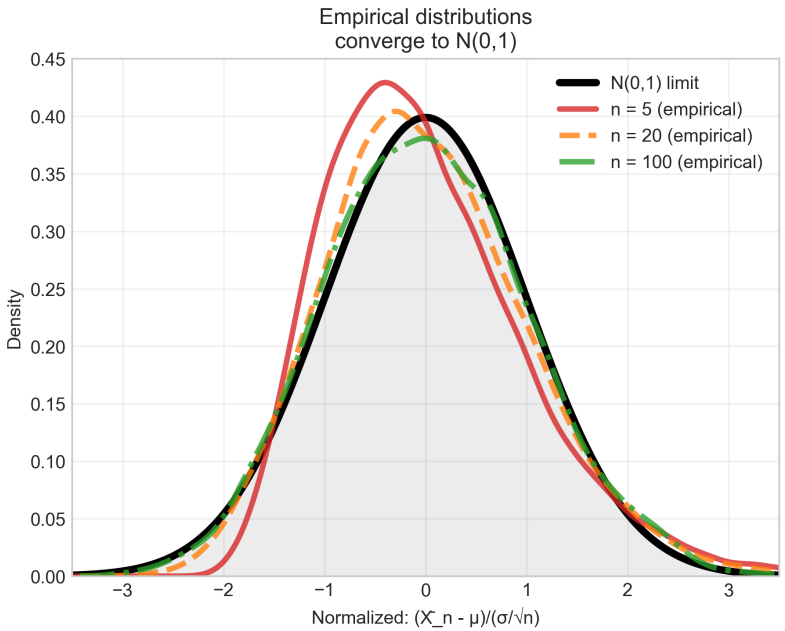
Central Limit Theorem (Formal)

Formal Statement

Let $X_1, X_2, \ldots$ i.i.d. with mean μ, variance $\sigma^2 < \infty$.
Sample mean: $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.
Then: $\sqrt{n}\,\frac{\overline{X}_n - \mu}{\sigma} \xrightarrow{d} \mathcal{N}(0,1)$.

Interpretation: Fluctuations shrink at rate $1/\sqrt{n}$ and are Gaussian-shaped.

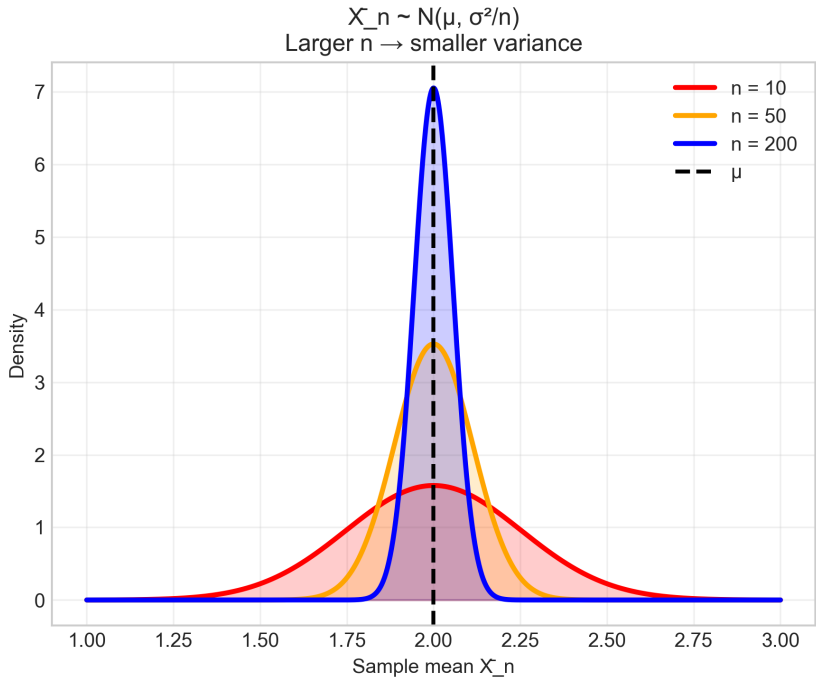


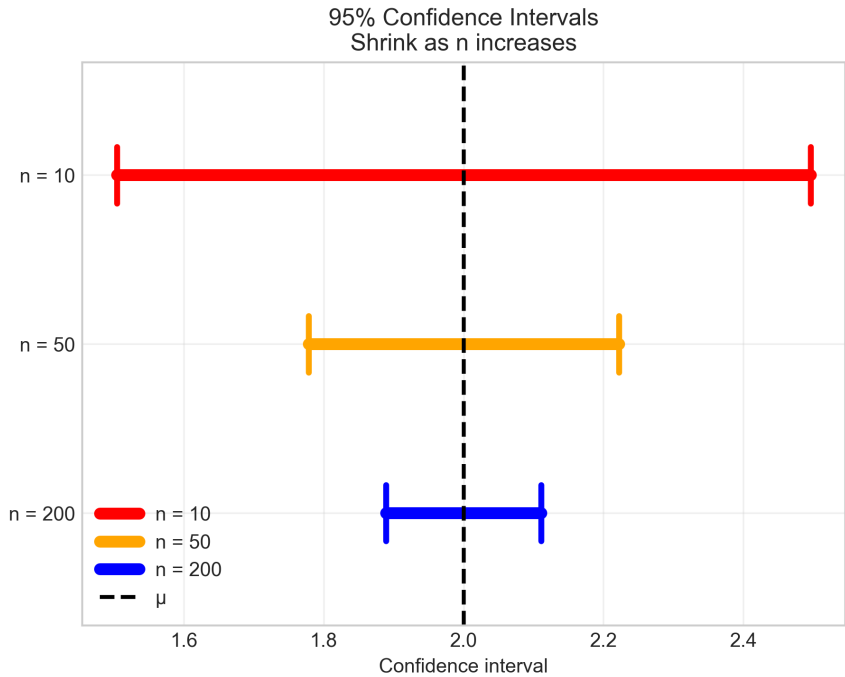


Why the CLT matters for Applied Statistics

Practical Consequences

- Approximation: $\overline{X}_n \approx \mathcal{N}(\mu, \sigma^2/n)$.
- Larger $n \Rightarrow$ smaller variance $\Rightarrow$ narrower confidence intervals.
- CLT underpins confidence intervals and hypothesis tests.

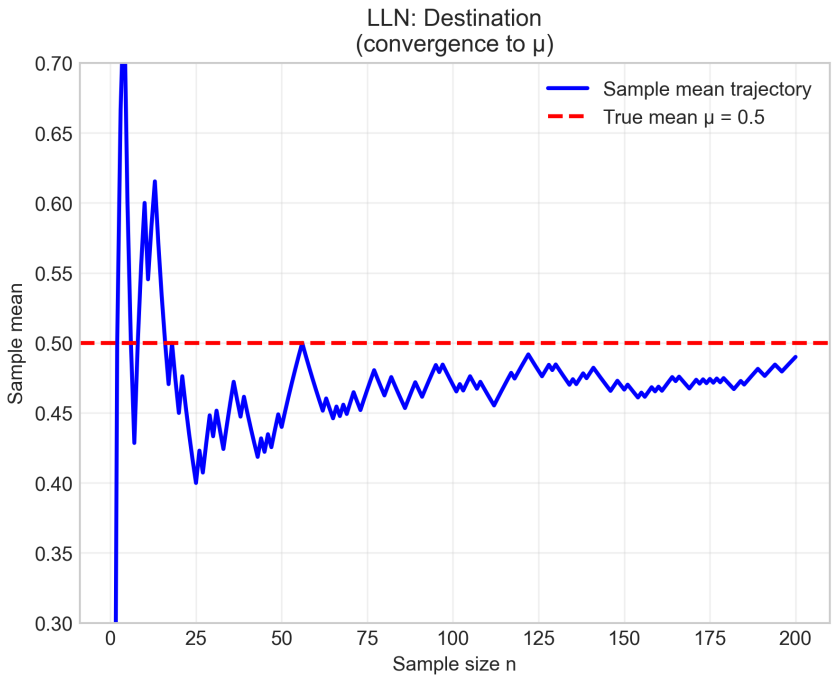


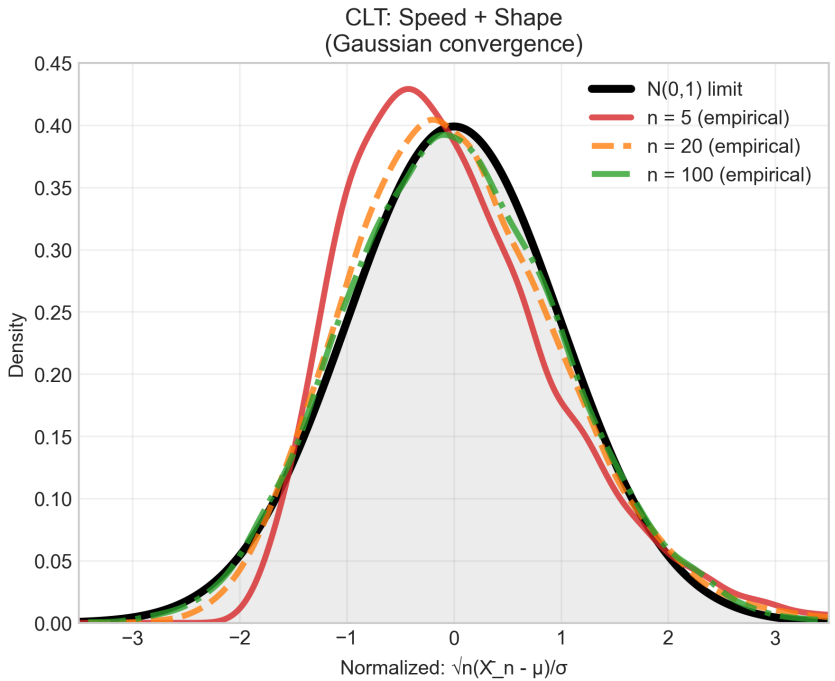


Key Insights from the CLT

Takeaways

- LLN: averages converge to μ.
- CLT: fluctuations shrink at rate $1/\sqrt{n}$ and are Gaussian-shaped.
- This dual perspective makes inference possible.





Remember: LLN = destination. CLT = speed + shape of convergence.

Outline

Exercises (Theory)

1. If $X \sim \mathrm{Uniform}(0,1)$, compute $\mathbb{E}[X]$, $\mathrm{Var}(X)$, median, and $q_{0.9}$.
2. If $X \sim \mathrm{Poisson}(3)$, compute $\mathbb{P}(X \leq 1)$ and $\mathbb{E}[X(X-1)]$.
3. Show that for $X \sim \mathrm{Bernoulli}(p)$, skewness $\gamma_1 = \dfrac{1-2p}{\sqrt{p(1-p)}}$.
4. Prove WLLN using Chebyshev for i.i.d. with finite variance.

Practical Preview

- EDA on heights dataset
- CLT simulation; Poisson fit to defects

Reference

See practical tasks in the lesson materials and starter code in the repository.

Outline

Summary

- Bridge descriptive to probabilistic modeling

References

- Casella and Berger, Statistical Inference
- Wasserman, All of Statistics
- Grimmett and Stirzaker, Probability and Random Processes