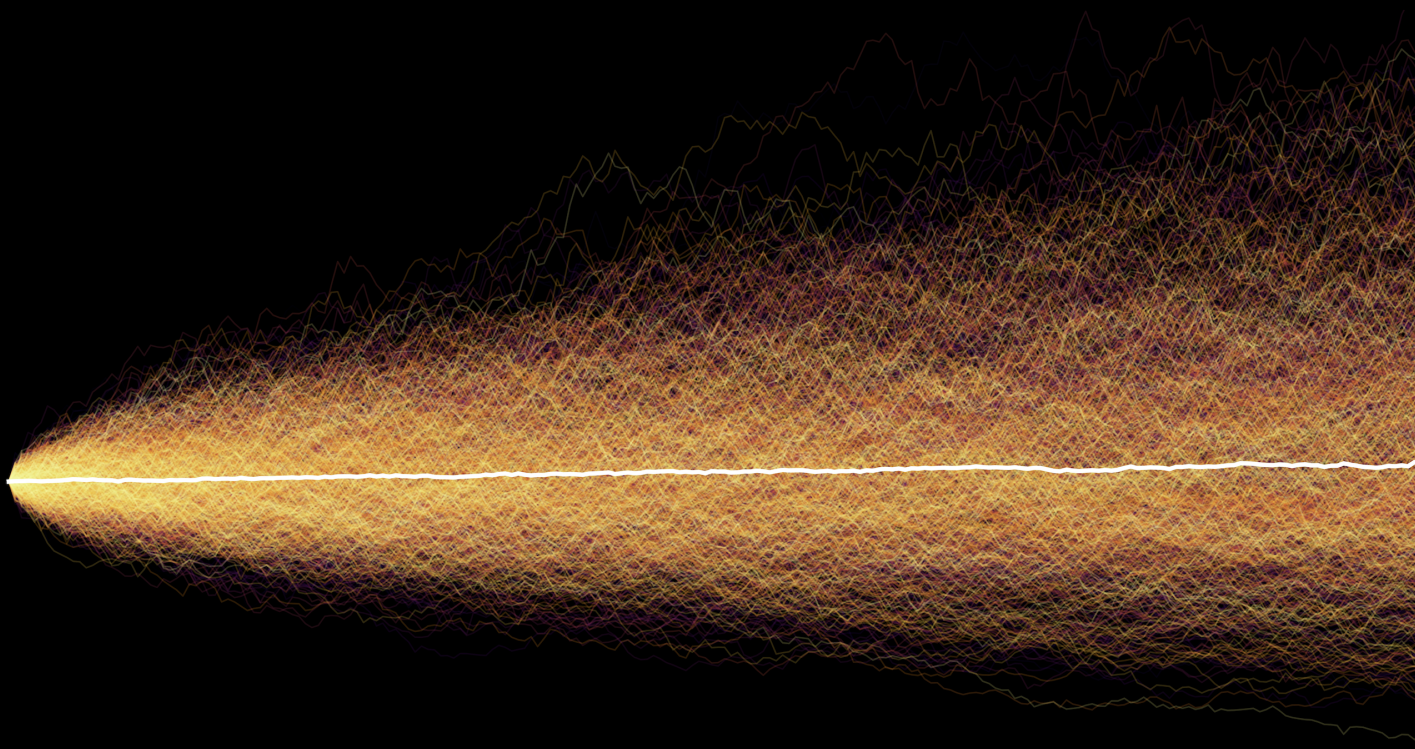
Probability and Stochastic Processes

Hannah Attar



Simulated Geometric Brownian Motions

Preface

This book is intended as a concise reference for the central concepts and results of probability and stochastic processes. It is not designed for instruction or self-contained learning, but as a structured handbook for readers who already possess familiarity with the material and require a clear, reliable way to refresh definitions, recall key formulas, and connect ideas in applied settings. The emphasis throughout is on essential structure and final expressions rather than derivations or examples, with brief conceptual notes included only where they support interpretation or use.

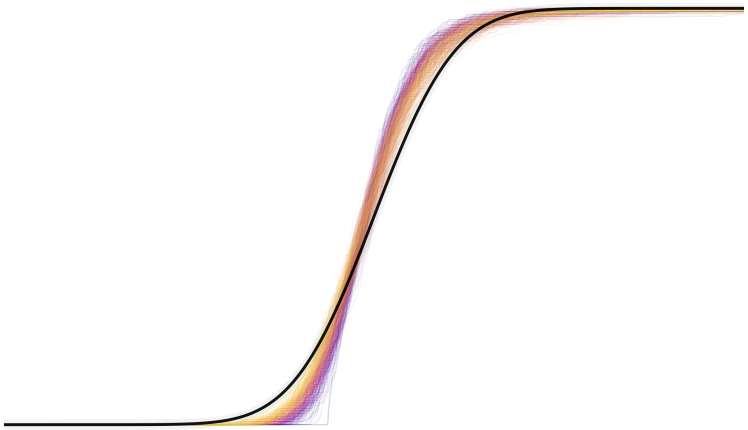
The material is organized linearly, beginning with probability theory and progressing into stochastic processes, reflecting the natural conceptual dependency between the two. Topics are presented with an emphasis on canonical formulas, limiting results, and modeling primitives that frequently arise in quantitative finance, financial engineering, machine learning, and related applied fields. Visualizations are used in place of worked examples to reinforce intuition and highlight behavior that is most relevant in practice.

The content is compiled and synthesized from coursework and publicly available sources. While the underlying theory is standard, this work represents a deliberate effort to curate, structure, and distill the material into a reference format optimized for rapid lookup and practical application. All figures and visualizations were generated programmatically. This document is an evolving work and reflects an ongoing process of refinement; as such, errors or omissions may remain.

Contents

Part I Probability

Probability is the language used whenever outcomes are uncertain but decisions still must be made. It appears wherever data are noisy, systems are complex, and perfect information is unavailable, which is to say, almost everywhere. From finance and engineering to machine learning, physics, and economics, probability provides the structure that allows uncertainty to be modeled, quantified, and acted upon.



What makes probability compelling is not that it predicts individual outcomes, but that it reveals order beneath randomness. Seemingly erratic behavior gives rise to stable patterns when viewed at the right scale. Aggregates behave differently from individuals, averages become reliable, and variability itself follows precise laws. These regularities make it possible to reason about risk, infer hidden structure from data, and design systems that perform reliably in uncertain environments.

This section develops the mathematical machinery that supports those ideas. The concepts introduced here form the foundation for inference, estimation, stochastic modeling, and learning from data. Probability does not remove uncertainty, but it makes uncertainty manageable, and, in many settings, exploitable.

Chapter 1: Events, Information, & Conditioning

When we talk about probability, we are really talking about events, how they relate to one another, and how information alters their likelihood. At its core, this subject concerns the structure of uncertainty: what it means for occurrences to be compatible, independent, or conditioned on further knowledge. The formulas collected here express these relationships in their most compact form. They serve as the basic mechanisms by which probability is assigned, combined, and updated, independent of any specific application.

1 Events, Independence, & Conditional Probability

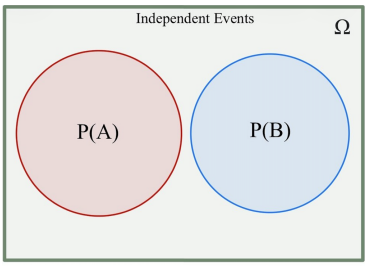
Independence. Two events A and B are *independent* if

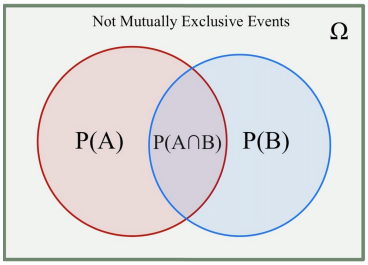
$$P(A \cap B) = P(A)\,P(B).$$

This equality expresses that the occurrence of one event carries no information about the other: the probability of the intersection factors into the product of the individual probabilities. Equivalently,

$$P(B \mid A) = P(B) \quad \text{and} \quad P(A \mid B) = P(A),$$

whenever the conditional probabilities are defined.





Conditional Probability. For events A and B with $P(A) > 0$, the conditional probability of B given A is

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}.$$

Conditioning restricts attention to the portion of the sample space where A occurs, and renormalizes probabilities accordingly.

2 Total Probability & Bayes' Rule

Law of Total Probability. If $\{H_k\}$ is a partition of the sample space Ω, then

$$P(E) = \sum_k P(H_k)\,P(E \mid H_k).$$

This expresses the probability of E as a weighted combination over disjoint scenarios. Each term corresponds to the contribution of E within H_k, scaled by $P(H_k)$.

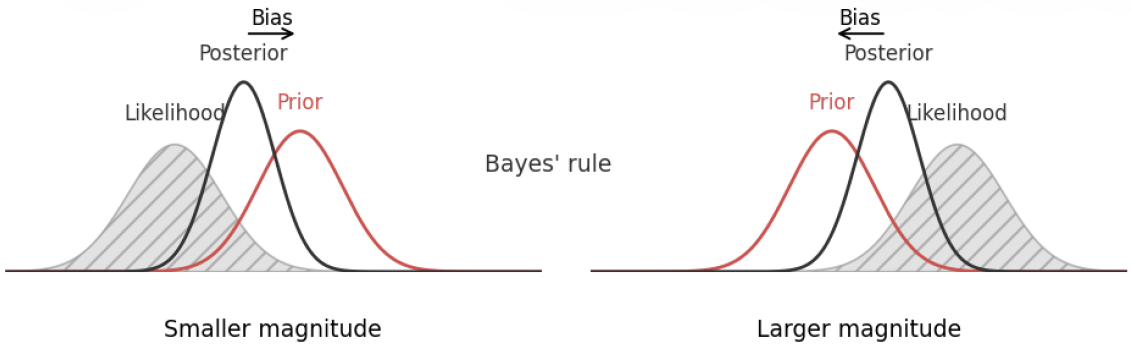
Special case: when $\Omega = A \cup A^c$,

$$P(B) = P(A)\,P(B \mid A) + P(A^c)\,P(B \mid A^c).$$

Bayes' Theorem. For a partition $\{H_k\}$ of Ω and an event E with $P(E) > 0$,

$$P(H_j \mid E) = \frac{P(E \mid H_j)\,P(H_j)}{\sum_k P(E \mid H_k)\,P(H_k)}.$$

Conceptually, Bayes' theorem reweights the prior probabilities $\{P(H_j)\}$ by how compatible each hypothesis is with the observed evidence E, via the likelihoods $\{P(E \mid H_j)\}$, and then renormalizes so that the updated probabilities $\{P(H_j \mid E)\}$ again sum to one.



Bayes shifts beliefs toward data & disagreement between prior and likelihood shows up as bias

Example. If $\{H_1, H_2\}$ partitions Ω and

$$P(H_1 \mid E) = \frac{a}{a+b}, \qquad P(H_2 \mid E) = \frac{b}{a+b},$$

then a and b represent the relative contributions of H_1 and H_2 to the likelihood of E. The posterior probabilities are obtained by normalizing these contributions.

Issue Spotting Sequence

P: Is there a partition?

$$\{H_k\} \text{ or } A \cup A^c$$

U: Is an unconditional probability needed?

$$P(B)$$

T: Use total probability:

$$P(E) = \sum_k P(H_k)\, P(E \mid H_k)$$

C: Is a conditional probability required?

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

B: Apply Bayes' theorem:

$$P(H_j \mid E) = \frac{P(E \mid H_j)\, P(H_j)}{\sum_k P(E \mid H_k)\, P(H_k)}$$

$$P(H_j) = \text{prior}, \quad P(H_j \mid E) = \text{posterior}, \quad P(E \mid H_j) = \text{likelihood}.$$

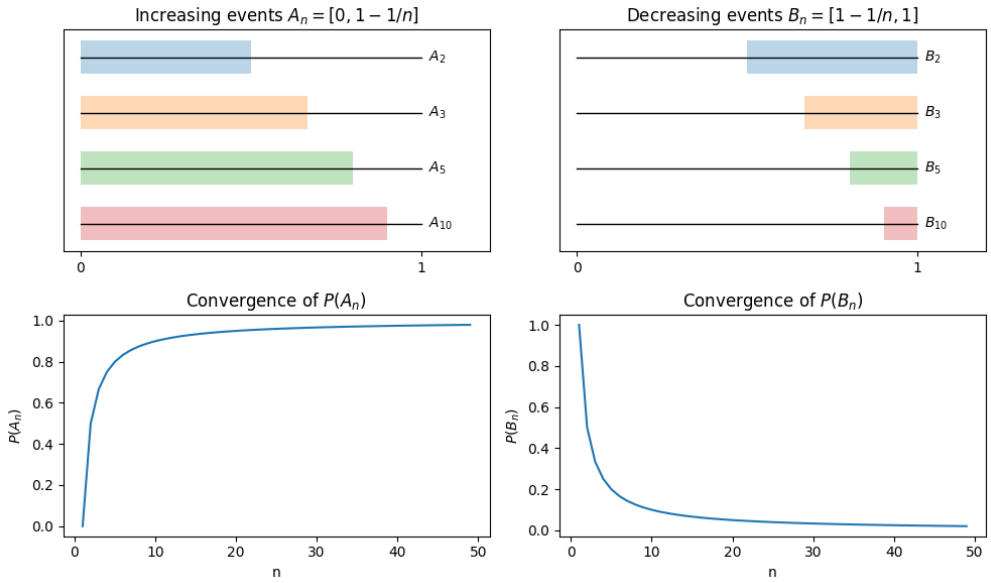
3 Long-Run Events & Borel–Cantelli

Probability Limits. Sequences of events play the role of limits for sets, describing what eventually happens as n increases. For a sequence $\{A_n\}$,

$$\lim_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} A_n \quad \text{if } A_k \subseteq A_{k+1} \quad \text{(increasing sequence)},$$

$$\lim_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} A_n \quad \text{if } A_k \supseteq A_{k+1} \quad \text{(decreasing sequence)}.$$

In the increasing case, the limit collects everything that eventually occurs; in the decreasing case, it retains only what persists forever.



Increasing and decreasing events converge in both sets and probabilities.

Continuity of Probability. Continuity of probability states that limits of events and limits of their probabilities are compatible for monotone sequences. If $\{A_n\}$ is increasing, then

$$P\Big(\lim_{n\to\infty} A_n\Big) = \lim_{n\to\infty} P(A_n) = P\left(\bigcup_{n=1}^{\infty} A_n\right).$$

If $\{A_n\}$ is decreasing, then

$$P\Big(\lim_{n\to\infty} A_n\Big) = \lim_{n\to\infty} P(A_n) = P\left(\bigcap_{n=1}^{\infty} A_n\right).$$

Thus, for monotone sequences, taking limits and taking probabilities commute.

Boole's Inequality. Boole's inequality provides a general upper bound on the probability of a union of events:

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} P(A_n).$$

The probability of at least one A_n occurring cannot exceed the sum of their individual probabilities; overlapping events only make the union smaller, never larger.

Infinitely Often Events. A sequence of events $\{A_k\}$ is said to occur *infinitely often (i.o.)* if

$$\forall n \geq 1,\ \exists k \geq n \text{ such that } A_k \text{ occurs},$$

which is represented by the event

$$\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k.$$

This event captures the long-run behavior that A_k keeps happening for arbitrarily large indices k, rather than eventually stopping.

Borel–Cantelli Lemma. Let $\{A_n\}$ be a sequence of events.

1. If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then
$$P\left(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k\right) = 0.$$
In this case, the total probability mass assigned to the A_n is finite, and the events occur only finitely many times almost surely.

2. If $\sum_{n=1}^{\infty} P(A_n) = \infty$ and the events $\{A_n\}$ are independent, then
$$P\left(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k\right) = 1.$$
Here, the accumulated probability is infinite and, under independence, the events occur infinitely often almost surely.

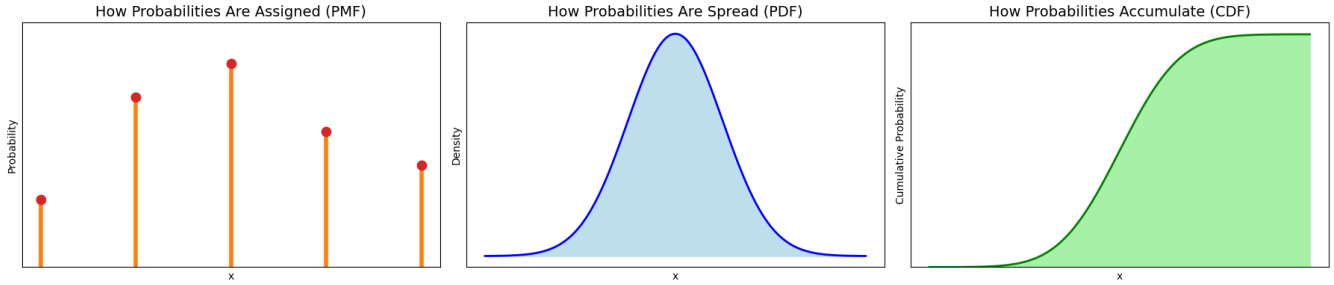
The lemma thus separates sequences that eventually die out from those that continue to occur indefinitely.

Chapter 2: Distributions - How Randomness is Shaped

Randomness is not only about whether events occur, but also how often and in what patterns. A distribution assigns weight to possible outcomes and determines the characteristic "shape" of uncertainty. Some distributions concentrate around a typical value, others decay slowly, and still others describe rare, sporadic events.

1 Distribution Functions

A random variable can be characterized in several ways. The correct object depends on whether the support is discrete, continuous, or mixed. The three fundamental descriptions are the *probability mass function*, the *probability density function*, and the *cumulative distribution function*.



The PMF assigns probability at distinct points, the PDF spreads probability smoothly across values so areas under the curve give interval probabilities, and the CDF records the cumulative total, rising from 0 to 1 as x increases.

Probability Mass Function (PMF). For a **discrete** random variable X with countable support $\mathcal{X}$, the PMF assigns probability to individual outcomes:

$$f_X(x) = P(X = x), \qquad x \in \mathcal{X}.$$

It satisfies

$$f_X(x) \geq 0, \qquad \sum_{x \in \mathcal{X}} f_X(x) = 1.$$

Probabilities of events are obtained by summing over the relevant points of the support.

Cumulative Distribution Function (CDF). The CDF gives the probability that the random variable takes a value less than or equal to x:

$$F_X(x) = P(X \leq x), \qquad x \in \mathbb{R}.$$

Every random variable has a CDF. It is non-decreasing, right-continuous, with limits

$$\lim_{x \to -\infty} F_X(x) = 0, \qquad \lim_{x \to +\infty} F_X(x) = 1.$$

For discrete distributions, $F_X(x)$ is a step function with jumps at the support points. For continuous distributions, $F_X(x)$ is smooth and increases continuously.

Probability Density Function (PDF). A distribution is **absolutely continuous** if there exists a function f such that

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

In this case,

$$f(x) = F_X'(x), \qquad f(x) \geq 0, \qquad \int_{-\infty}^{\infty} f(x)\,dx = 1.$$

The PDF describes how probability is distributed along the real line. Interval probabilities are obtained by integration:

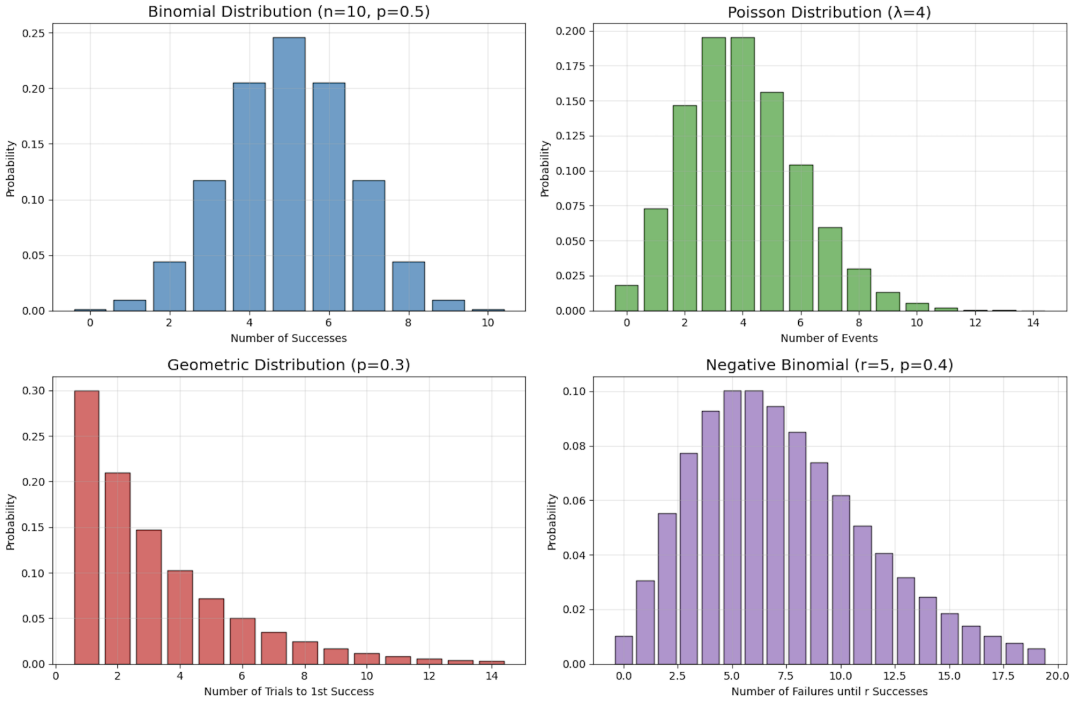
$$P(a \leq X \leq b) = \int_a^b f(x)\,dx.$$

A PDF exists only when the distribution is absolutely continuous. Discrete or mixed distributions may not admit a single global density, but every distribution always has a valid CDF.

Object	Support Type	Formula	Use
PMF	Discrete	$f_X(x) = P(X = x)$	Probabilities by summation
PDF	Continuous	$f_X(x) = F_X'(x)$	Probabilities by integration
CDF	All	$F_X(x) = P(X \leq x)$	Universal description of distribution

2 Core Discrete Distributions

Discrete distributions assign probability to individual points rather than ranges, so uncertainty is organized over a countable set where every outcome can, in principle, be listed. They arise naturally whenever we observe counts, trials, or selections, and many familiar models are connected.



Bernoulli (p)

- Single trial with two outcomes: "success" with probability p and "failure" with probability $1-p$.
- Support $\{0,1\}$, where 1 usually denotes success.
- Basic building block for Binomial, Geometric, and Negative Binomial models.

Binomial (n,p)

- Number of successes in n independent Bernoulli(p) trials.
- Support $\{0,1,\ldots,n\}$.
- Symmetric about $n/2$ when $p=1/2$; right-skewed if $p<1/2$, left-skewed if $p>1/2$.
- Approximations:

$$\text{Binomial}(n,p) \approx \text{Poisson}(\lambda = np) \quad \text{when } n \text{ large and } p \text{ small,}$$

$$\text{Binomial}(n,p) \approx \text{Normal}(np,\, np(1-p)) \quad \text{when } np,\, n(1-p) \gtrsim 10.$$

Geometric (p)

- Number of trials required until the first success in i.i.d. Bernoulli(p) trials.
- Only discrete distribution with the *memoryless* property:

$$P(X > m+n \mid X > m) = P(X > n).$$

- Typically highly right-skewed with a long tail.

Hypergeometric (N_1, N_2, n)

- Number of "successes" in a sample of size n drawn *without replacement* from a finite population with N_1 successes and N_2 failures $(N = N_1 + N_2)$.
- Variance includes a finite-population correction factor $\frac{N-n}{N-1}$.
- For large N with $n \ll N$, Hypergeometric is well approximated by Binomial with $p = N_1/N$.

Negative Binomial (r,p)

- Number of trials (or failures) required to observe r successes in i.i.d. Bernoulli(p) trials.
- Geometric is the special case $r=1$.
- Useful for overdispersed count data where variance exceeds the mean (in contrast to Poisson).
- Right-skewed; heavier tail than Poisson with the same mean.

Poisson (λ)

- Counts the number of events occurring in a fixed interval of time or space when events arrive independently at rate λ.
- Additive under independent summation:

$$X_1 \sim \text{Poisson}(\lambda_1),\ X_2 \sim \text{Poisson}(\lambda_2) \Rightarrow X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2).$$

- Approximates Normal(λ, λ) when λ is large.

Discrete Uniform on $\{1, \ldots, m\}$

- All outcomes in the finite set $\{1, \ldots, m\}$ are equally likely.
- Natural model for an ideal die, random index, or symmetric label.
- Mean $(m+1)/2$; variance $(m^2-1)/12$ reflects symmetric spread around the center.

Multinomial $(n, \mathbf{p})$

- Joint distribution of counts $(X_1, \ldots, X_k)$ from n independent trials with k outcome categories and category probabilities $\mathbf{p} = (p_1, \ldots, p_k)$.
- Generalizes the Binomial to more than two categories; each marginal X_i is Binomial(n, p_i).
- Covariances between counts are negative:

$$\mathrm{Cov}(X_i, X_j) = -np_ip_j, \quad i \neq j,$$

  reflecting competition between categories.
- Appears in contingency tables, categorical frequency counts, and empirical histograms.

3 Choosing a Discrete Model: Sampling vs Outcomes

	With Replacement	**Without Replacement**
2 Outcomes	Geometric Negative Binomial Bernoulli Binomial	Hypergeometric
> 2 **Outcomes**	Multinomial	Multivariate Hypergeometric

Counting? $\Rightarrow$ Poisson

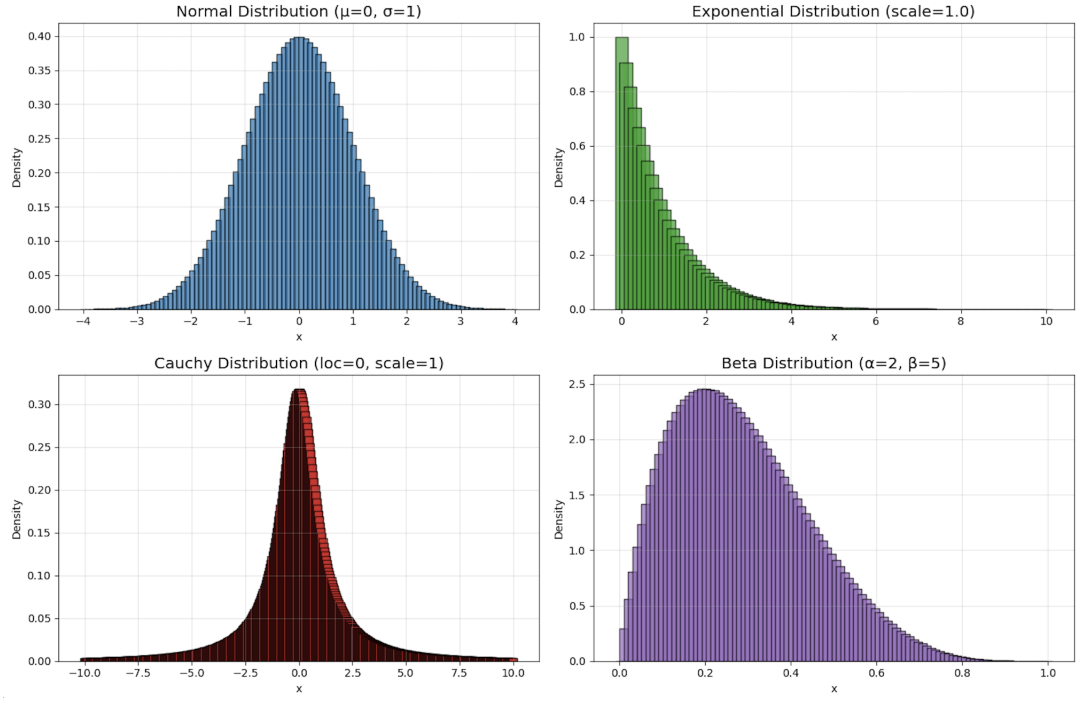
1. Decide how sampling is done:
   - **With replacement** $\Rightarrow$ independent trials
   - **Without replacement** $\Rightarrow$ dependent trials
2. Count the number of outcome categories:
   - **Two outcomes** (success/failure)
     - With replacement: Bernoulli, Binomial, Geometric, Negative Binomial
     - Without replacement: Hypergeometric
   - **More than two outcomes**
     - With replacement: Multinomial
     - Without replacement: Multivariate Hypergeometric
3. If the task is **counting events in time** rather than sampling, use Poisson.

Table of Discrete Distributions & Their Moments

Distribution	**PMF** $f(x)$	**Mean** μ	**Variance** σ^2
Bernoulli	$p^x(1-p)^{1-x},\ x=0,1$	p	$p(1-p)$
Binomial	$\binom{n}{x}p^x(1-p)^{n-x},\ x=0,1,\ldots,n$	np	$np(1-p)$
Geometric	$(1-p)^{x-1}p,\ x=1,2,3,\ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Hypergeometric	$\frac{\binom{N_1}{x}\binom{N_2}{n-x}}{\binom{N}{n}},\ x\le n,\ x\le N_1,\ n-x\le N_2$	$n\frac{N_1}{N}$	$n\frac{N_1}{N}\frac{N_2}{N}\frac{N-N_1}{N-1}$
Negative Binomial	$\binom{x-1}{r-1}p^r(1-p)^{x-r},\ x=r,r+1,\ldots$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$
Poisson	$\frac{\lambda^x e^{-\lambda}}{x!},\ x=0,1,2,\ldots$	λ	λ
Uniform (discrete)	$\frac{1}{m},\ x=1,2,\ldots,m$	$\frac{m+1}{2}$	$\frac{m^2-1}{12}$
Multinomial	$\frac{n!}{x_1!\cdots x_k!}p_1^{x_1}\cdots p_k^{x_k},\ \sum x_i=n,\ \sum p_i=1$	np_i	$np_i(1-p_i)$

4 Core Continuous Distributions

Continuous distributions assign probability through a density over an interval or the entire real line, so individual points have probability zero and only ranges carry mass. The key features are the shape of the density, the support, and how the parameters control concentration, tail behavior, and skewness.



Uniform (a, b)

- All points in $[a, b]$ are equally likely; the density is constant.
- Represents complete indifference over a finite range.

Exponential (λ)

- Models waiting time until the first event in a Poisson process.
- Support $[0, \infty)$; sharply right–skewed.
- Only continuous memoryless distribution:

$$P(X > s + t \mid X > s) = P(X > t).$$

Gamma (α, β)

- Sum of α independent Exponential(β) variables (integer α).
- Support $[0, \infty)$; shape controlled by α, scale by β.
- Special cases: Exponential, Chi–square.
- Gamma function identities:

$$\Gamma(n) = (n-1)! \quad \text{for integers,} \qquad \Gamma\left(\tfrac{1}{2}\right) = \sqrt{\pi},$$

$$\Gamma(n+1) = n\,\Gamma(n).$$

Chi-square (k)

- Distribution of $\sum_i Z_i^2$ for standard Normals Z_i.
- Support $[0, \infty)$; right–skewed, approaches Normal as k increases.
- Special case of Gamma: $\chi^2_k \sim \Gamma\left(\frac{k}{2}, \frac{1}{2}\right)$.

Beta (α, β)

- Defined on $[0, 1]$; models proportions and probabilities.
- Highly flexible shapes (U–shaped, uniform, symmetric, skewed).
- Beta function:
$$B(\alpha, \beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\, du = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$
- Conjugate prior for Bernoulli/Binomial data; α, β act like prior counts.

Normal (μ, σ^2)

- Symmetric bell–shaped distribution centered at μ with spread σ^2.
- Arises as the limit of sums/averages of many small independent effects (CLT).
- Completely determined by mean and variance; closed under affine transformations.
- Standardization: $Z = (X - \mu)/\sigma$.
- Symmetry fact: the standard Normal density is symmetric about 0.

Cauchy (x_0, γ)

- Heavy–tailed distribution; ratio of independent standard Normals.
- No finite mean or variance; moments are undefined.
- Sample averages do not converge; illustrates failure of typical limit laws.

Table of Continuous Distributions & Their Moments

Distribution	**PDF** $f(x)$	**CDF** $F(x)$	**Range**	**Mean** μ	**Variance** σ^2
Beta (α, β)	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	$I_x(\alpha, \beta)$	$0 < x < 1$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Chi-square (r)	$\frac{1}{2^{r/2}\Gamma(r/2)} x^{r/2-1} e^{-x/2}$	$P\left(\frac{r}{2}, \frac{x}{2}\right)$	$x \geq 0$	r	$2r$
Exponential (θ)	$\frac{1}{\theta} e^{-x/\theta}$	$1 - e^{-x/\theta}$	$x \geq 0$	θ	θ^2
Gamma (α, θ)	$\frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} e^{-x/\theta}$	$P\left(\alpha, \frac{x}{\theta}\right)$	$x \geq 0$	$\alpha\theta$	$\alpha\theta^2$
Normal (μ, σ^2)	$\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$	$\Phi\left(\frac{x-\mu}{\sigma}\right)$	$x \in \mathbb{R}$	μ	σ^2
Uniform (a, b)	$\frac{1}{b-a}$	$\frac{x-a}{b-a}$	$a \leq x \leq b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Cauchy (m, d)	$\frac{1}{\pi d\left[1 + \left(\frac{x-m}{d}\right)^2\right]}$	$\frac{1}{\pi} \tan^{-1}\left(\frac{x-m}{d}\right) + \frac{1}{2}$	$x \in \mathbb{R}$	undefined	undefined

5 Normal & Standard Normal

Normal Distribution. A random variable X is Normal with mean μ and variance σ^2 if

$$X \sim \mathcal{N}(\mu, \sigma^2), \qquad f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

It is symmetric about μ, completely determined by its mean and variance, and closed under linear transformations.

CDF.

$$F_X(x) = P(X \le x) = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{x} \exp\left(-\frac{1}{2}\left(\frac{w-\mu}{\sigma}\right)^2\right) dw.$$

Standardization. Any Normal variable can be mapped to a Standard Normal by

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

Standard Normal.

$$\Phi(z) = P(Z \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-w^2/2}\, dw, \qquad \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

$$\Phi(-z) = 1 - \Phi(z), \qquad P(Z > z) = 1 - \Phi(z).$$

Tables or numerical routines for $\Phi(z)$ allow probability calculations for any Normal variable via standardization.

Jointly Gaussian. If (X, Y) are jointly Normal, independence is equivalent to zero correlation:

$$(X,Y) \text{ jointly Gaussian}, \quad \rho = 0 \iff X \text{ independent of } Y.$$

The joint density is

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}.$$

Normal Approximations. Many discrete or skewed distributions become approximately Normal when parameters are large:

$$\begin{aligned}
\text{Binomial:} &\quad B(n,p) \approx \mathcal{N}(np,\, npq) \\
\text{Negative Binomial:} &\quad NB(n,p) \approx \mathcal{N}\left(\frac{n}{p},\, \frac{nq}{p^2}\right) \\
\text{Poisson:} &\quad P(\lambda) \approx \mathcal{N}(\lambda,\, \lambda) \\
\text{Gamma:} &\quad \Gamma(n,\theta) \approx \mathcal{N}(n\theta,\, n\theta^2) \\
\text{Chi-square:} &\quad \chi^2(n) \approx \mathcal{N}(n,\, 2n).
\end{aligned}$$

These approximations arise from the Central Limit Theorem.

6 Signal Detection & Hypothesis Testing.

Signal detection treats observation as a noisy measurement that may or may not contain a structured effect, and frames the problem as choosing between competing hypotheses based on that observation. The goal is to design a decision rule (here, a threshold on a Normal measurement) that balances false alarms and missed detections, and to quantify this tradeoff through error probabilities and the power of the test.

We compare two hypotheses:

$$H_0 : \text{No signal (just noise)}, \quad X \sim \mathcal{N}(0, \sigma^2)$$

$$H_1 : \text{Signal is present}, \quad X \sim \mathcal{N}(1, \sigma^2)$$

A **decision threshold** T determines which hypothesis we accept:

$$\begin{cases} X \leq T \Rightarrow \text{Accept } H_0 & \text{(No signal)} \\ X > T \Rightarrow \text{Accept } H_1 & \text{(Signal detected)} \end{cases}$$

Error Probabilities.

$$\alpha \triangleq P[\text{Type I Error}] = P[\text{reject } H_0 | H_0 \text{ true}] = P[X > T | H_0]$$

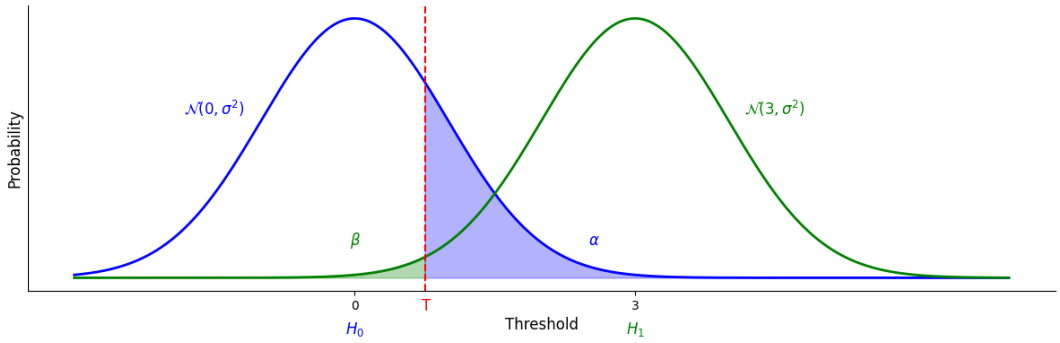
$$\beta \triangleq P[\text{Type II Error}] = P[\text{accept } H_0 | H_1 \text{ true}] = P[X \leq T | H_1]$$

Here, α is the *false alarm probability*, and β is the *miss probability*.

Power.

$$\text{Power} = 1 - \beta = P[\text{correct detection} \mid H_1]$$

The power increases as the distributions separate, or as T moves left and measures how likely we are to correctly say "there is a signal" when one truly exists.



Normal Formulas for α & β via Standardization. Under $H_0 : X \sim \mathcal{N}(0, \sigma^2)$ and $H_1 : X \sim \mathcal{N}(1, \sigma^2)$, standardization gives

$$\alpha = P[X > T \mid H_0] = 1 - \Phi\left(\frac{T}{\sigma}\right), \qquad \beta = P[X \leq T \mid H_1] = \Phi\left(\frac{T-1}{\sigma}\right).$$

As T increases, α decreases and β increases, reflecting the tradeoff between false alarms and missed detections. The separation

$$d' \triangleq \frac{1-0}{\sigma} = \frac{1}{\sigma}$$

measures how distinguishable the two hypotheses are: larger d' yields lower error probabilities for a well-chosen threshold.

Chapter 3: Moments & Dependence

Moments are the language of structure in probability. They describe not only where a distribution is centered, but how tightly it spreads, how asymmetrically it leans, and how heavy its tails may be. Dependence extends this language from single variables to pairs, capturing whether two quantities tend to rise and fall together, or whether they vary independently despite sharing the same environment. Together, moments and dependence provide a concise summary of the behavior of random variables, isolating the features that matter across models, data, and applications.

1 Expectation & Variance

Expectation summarizes the typical level of a random variable, while variance quantifies how much values fluctuate around that level. Together they are the primary numerical descriptors of a distribution's location and spread, and they extend naturally to conditional settings and to estimators built from data.

Expectation

Definition. The expectation $E[X]$ of a random variable X is its average value under the distribution of X:

$$E[X] = \mu$$
$$= \begin{cases} \displaystyle\sum_x x\, f_X(x), & \text{discrete,} \\ \displaystyle\int_{-\infty}^{\infty} x\, f_X(x)\, dx, & \text{continuous.} \end{cases}$$

Conceptually, $E[X]$ is the "center of mass" of the distribution: more probable outcomes pull the average more strongly.

Linearity and sums. Expectation is linear, so constants and sums pass through:

$$E[aX + b] = a\, E[X] + b, \qquad E\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i\, E[X_i].$$

No independence is required; this is purely an algebraic property of the integral/sum.

Products under independence. If X and Y are independent,

$$E[XY] = E[X]\, E[Y].$$

Independence removes any interaction term: on average, the product behaves like the product of averages.

Conditional expectation. Conditional expectation refines the average once additional information $Y = y$ is known:

$$E[X \mid Y = y] = \begin{cases} \displaystyle\sum_x x\, p_{X|Y}(x \mid y), & \text{discrete,} \\ \displaystyle\int_{-\infty}^{\infty} x\, f_{X|Y}(x \mid y)\, dx, & \text{continuous.} \end{cases}$$

For jointly distributed (X, Y),

$$E[X] = \iint x\, f_{X,Y}(x, y)\, dx\, dy = \int E[X \mid Y = y]\, f_Y(y)\, dy,$$

which links the joint density, the conditional expectation, and the marginal density of Y.

Law of total expectation. Averaging conditional expectations over Y recovers the unconditional mean:

$$E[X] = E_Y[E[X \mid Y]].$$

Operationally: Compute the mean of X in each scenario $Y = y$, then average these scenario-wise means according to how likely each y is.

Expectation Toolbox

Topic	**Formula**	**Interpretation**
Definition	$E[X] = \sum_x x f_X(x)$ or $E[X] = \int x f_X(x)\,dx$	Center / average value
Linearity	$E[aX + b] = aE[X] + b$	Shift and scale pass through
Sums	$E\left[\sum_i a_i X_i\right] = \sum_i a_i E[X_i]$	Expectation distributes over sums
Independence	$X \perp Y \Rightarrow E[XY] = E[X]E[Y]$	Product of means when independent
Conditional mean	$E[X \mid Y = y]$	Mean with $Y = y$ fixed
Total expectation	$E[X] = E_Y[E[X \mid Y]]$	Average of scenario-wise means

Variance

Definition. The variance of X measures the typical squared deviation from its mean:

$$\begin{aligned} V[X] &= \sigma_X^2, \\ &= E[(X - E[X])^2], \\ &= E[X^2] - (E[X])^2. \end{aligned}$$

A large variance means realizations of X are widely scattered around $E[X]$; a small variance means they are tightly clustered.

Scaling and shifts. Adding a constant does not change variability, while scaling stretches or contracts it:

$$V[aX + b] = a^2\, V[X].$$

Sums and linear combinations. For any collection $\{X_i\}_{i=1}^n$,

$$V\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n V[X_i] \;+\; 2 \sum_{1 \le i < j \le n} \operatorname{cov}(X_i, X_j).$$

Variance of a sum consists of the individual variances plus all pairwise covariance terms. If the X_i are independent, all covariances vanish and

$$V\Big[\sum_{i=1}^n X_i\Big] = \sum_{i=1}^n V[X_i].$$

For two variables and constants a, b,

$$V[aX + bY] = a^2V[X] + b^2V[Y] + 2ab\ \mathrm{cov}(X, Y),$$

which is the basic variance formula for linear combinations.

Product of independent variables. When X and Y are independent,

$$V[XY] = E[X^2]E[Y^2] - \big(E[X]E[Y]\big)^2,$$

expressing the variability of a product in terms of second moments and means of the factors.

Conditional variance and total variance. Conditional variance quantifies spread once Y is fixed:

$$V[X \mid Y] = E\big[(X - E[X \mid Y])^2 \mid Y\big] = E[X^2 \mid Y] - \big(E[X \mid Y]\big)^2.$$

The law of total variance decomposes overall variability into within-scenario and between-scenario parts:

$$V[X] = E_Y[V[X \mid Y]] + V_Y(E[X \mid Y]).$$

Bias–variance decomposition for estimators. For an estimator $\widehat{\theta}$ of a parameter θ,

$$\mathrm{MSE}(\widehat{\theta}) = E[(\widehat{\theta} - \theta)^2] = V[\widehat{\theta}] + \big(E[\widehat{\theta}] - \theta\big)^2,$$

with

$$\text{Squared Bias} = \big|E[\widehat{\theta}] - \theta\big|^2.$$

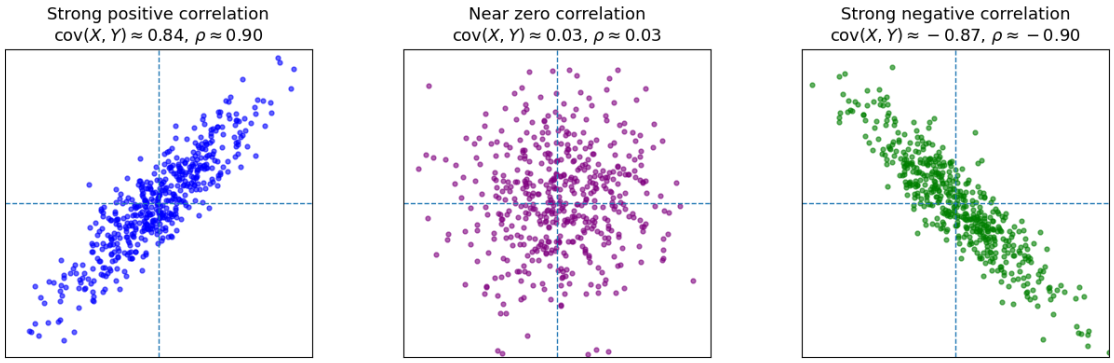
Here $V[\widehat{\theta}]$ captures random fluctuation of the estimator, while the squared bias measures systematic offset from the true parameter.

Variance Toolbox

Topic	**Formula**	**Interpretation**
Definition	$V[X] = E[(X - E[X])^2]$	Spread around the mean
Moment form	$V[X] = E[X^2] - (E[X])^2$	Uses first and second moments
Scaling	$V[aX + b] = a^2V[X]$	Shifts irrelevant, scaling matters
Sums	$V\left[\sum_i X_i\right] = \sum_i V[X_i] + 2\sum_{i<j} \mathrm{cov}(X_i, X_j)$	Variances plus covariances
Independence	X_i indep $\Rightarrow V\left[\sum_i X_i\right] = \sum_i V[X_i]$	Covariances vanish
Total variance	$V[X] = E[V[X \mid Y]] + V(E[X \mid Y])$	Within + between variability
Bias–variance	$\mathrm{MSE}(\hat{\theta}) = V[\hat{\theta}] + (E[\hat{\theta}] - \theta)^2$	Random + systematic error

2 Covariance and Correlation

Covariance and correlation summarize how two random variables change together. Covariance measures the *direction and magnitude* of joint variation; correlation rescales it to a *dimensionless* index between -1 and 1. Both concepts extend naturally to conditional settings.



The shape of each scatter shows the sign and strength of linear dependence.

Covariance

Definition. The covariance of (X, Y) is

$$\begin{aligned}\operatorname{cov}(X,Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - E[X]E[Y].\end{aligned}$$

It is positive when X and Y move together, negative when they move oppositely, and zero when they show no linear co-variation.

Basic properties. Covariance behaves linearly in both arguments:

$$\operatorname{cov}(aX + b,\ cY + d) = ac\ \operatorname{cov}(X,Y).$$

It is symmetric and identifies variance as a special case:

$$\operatorname{cov}(X,Y) = \operatorname{cov}(Y,X), \qquad \operatorname{cov}(X,X) = V[X].$$

Linearity over sums. For linear combinations,

$$\operatorname{cov}\left(\sum_i a_i X_i,\ \sum_j b_j Y_j\right) = \sum_i \sum_j a_i b_j\ \operatorname{cov}(X_i, Y_j).$$

Variance of a sum is a sum of variances and covariances:

$$V\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} V[X_i]\ +\ 2 \sum_{1 \le i < j \le n} \operatorname{cov}(X_i, X_j).$$

Relation to product expectation.

$$E[XY] = \operatorname{cov}(X,Y) + E[X]\,E[Y].$$

Covariance captures the deviation from $E[X]E[Y]$.

Linear transformations. If $Y = aX + b$ then

$$\operatorname{cov}(X, Y) = a\, V[X].$$

More generally,

$$V[aX + bY] = a^2 V[X] + b^2 V[Y] + 2ab\, \operatorname{cov}(X, Y).$$

Conditional covariance and total covariance. Conditioning on Z,

$$\operatorname{cov}(X, Y \mid Z) = E[XY \mid Z] - E[X \mid Z]E[Y \mid Z],$$

and averaging yields the total covariance decomposition:

$$\operatorname{cov}(X, Y) = E_Z[\operatorname{cov}(X, Y \mid Z)] + \operatorname{cov}_Z(E[X \mid Z],\, E[Y \mid Z]).$$

The first term is "within-Z" co-variation; the second is co-variation of conditional means.

Correlation

Definition. Correlation rescales covariance by the marginal standard deviations:

$$\rho_{XY} = \frac{\operatorname{cov}(X, Y)}{\sigma_X\, \sigma_Y}, \qquad -1 \le \rho_{XY} \le 1.$$

It measures direction and strength of linear dependence on a unitless scale.

Interpretation.

$\rho_{XY} = 1$ (perfect positive), $\rho_{XY} = -1$ (perfect negative), $\rho_{XY} = 0$ (no linear association).

Relation back to covariance.

$$\operatorname{cov}(X, Y) = \rho_{XY}\, \sigma_X \sigma_Y.$$

Covariance captures raw co-variation; correlation captures standardized strength.

Key Differences of Covariance vs Correlation

	Covariance	**Correlation**
What it measures	Direction of the relationship (positive, negative, or none)	Direction and strength of the relationship (scaled between -1 and 1)
Scale	Depends on the units and magnitude of X and Y	Standardized, does not depend on units
Range	Unbounded, can take any real value	Always between -1 and 1
Interpretation	Shows direction and rough magnitude of co-movement	Shows direction and strength of linear relationship
Use case	Understanding the direction of a relationship within one dataset	Comparing relationships across different datasets

3 Higher Moments & Central Moments

Moments extend the ideas of mean and variance by describing the global shape of a distribution. Raw moments capture the average value of X^k relative to the origin, while central moments capture the average value of $(X-\mu)^k$ relative to the mean. Increasing k emphasizes the contribution of extreme values, so higher moments are naturally connected to tail behavior, asymmetry, and peakedness.

Raw Moments

Definition. The kth moment (raw moment) of X is

$$\mathbb{E}[X^k] = \begin{cases} \displaystyle\sum_x x^k\, p(x), & \text{discrete,} \\ \displaystyle\int_{-\infty}^{\infty} x^k\, f(x)\, dx, & \text{continuous.} \end{cases}$$

Raw moments describe how the distribution is positioned relative to the origin.

Central Moments

Definition. The kth central moment is the moment of deviations from the mean:

$$\mathbb{E}[(X-\mu)^k] = \begin{cases} \displaystyle\sum_x (x-\mu)^k\, p(x), & \text{discrete,} \\ \displaystyle\int_{-\infty}^{\infty} (x-\mu)^k\, f(x)\, dx, & \text{continuous.} \end{cases}$$

Central moments measure shape *relative to the mean*, rather than absolute values.

A moment exists only if the corresponding expectation is finite; heavy-tailed distributions may have lower-order moments but fail to have higher-order ones.

Key Cases

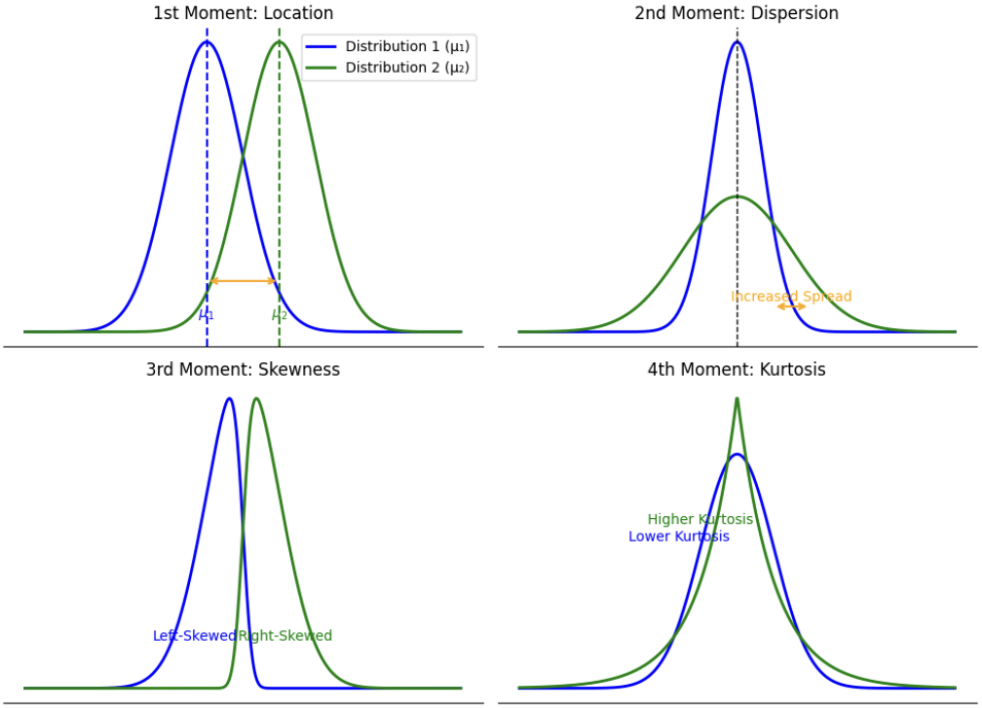
1. **First moment:** $\mathbb{E}[X] = \mu$
   Location of the distribution.
2. **Second central moment:** $\mathbb{E}[(X-\mu)^2] = \sigma^2$
   Overall spread around the mean.
3. **Third central moment:** describes *skewness*
   Positive skew: longer right tail. Negative skew: longer left tail.
4. **Fourth central moment:** describes *kurtosis*
   Measures tail thickness and peak sharpness relative to a normal distribution.

Moments and central moments are complementary: raw moments locate the distribution relative to zero, central moments describe how it stretches and bends around its mean.

Standardized moments divide the kth central moment by σ^k, removing scale and producing unitless measures such as skewness ($k=3$) and kurtosis ($k=4$).

The figure below summarizes kth moments and central moments for standard distributions.

Figure: First Four Central Moments



The first four moments successively describe location, dispersion, asymmetry, and tail behavior. Higher moments emphasize extremes, and for heavy-tailed distributions they may fail to exist, which is itself informative.

On the next page is a table displaying the explicit formulas for $\mathbb{E}[X^k]$ and $\mathbb{E}[(X-\mu)^k]$ for the discrete and continuous families covered.

kth Moments & Central Moments Table

Distribution	**kth moment** $\mathbb{E}[X^k]$	**kth central moment** $\mathbb{E}[(X-\mu)^k]$
Bernoulli (p)	p	$(1-p)(-p)^k + p(1-p)^k$
Binomial (n,p)	$\sum_{j=0}^{n} j^k \binom{n}{j} p^j (1-p)^{n-j}$	$\sum_{j=0}^{n} (j-np)^k \binom{n}{j} p^j (1-p)^{n-j}$
Geometric (p)	$p\sum_{x=1}^{\infty} x^k (1-p)^{x-1}$	$p\sum_{x=1}^{\infty} (x-\frac{1}{p})^k (1-p)^{x-1}$
Poisson (λ)	$e^{-\lambda}\sum_{x=0}^{\infty} \frac{x^k \lambda^x}{x!}$	$e^{-\lambda}\sum_{x=0}^{\infty} \frac{(x-\lambda)^k \lambda^x}{x!}$
Negative Binomial (r,p)	$\sum_{x=0}^{\infty} x^k \binom{r+x-1}{x} (1-p)^x p^r$	$\sum_{x=0}^{\infty} (x-\frac{r(1-p)}{p})^k \binom{r+x-1}{x} (1-p)^x p^r$
Discrete Uniform $\{1,\ldots,n\}$	$\frac{1}{n}\sum_{x=1}^{n} x^k$	$\frac{1}{n}\sum_{x=1}^{n} \left(x-\frac{n+1}{2}\right)^k$
Beta (α,β)	$\frac{\Gamma(\alpha+k)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha+\beta+k)}$	$\sum_{j=0}^{k} \binom{k}{j} (-\frac{\alpha}{\alpha+\beta})^{k-j} \frac{\Gamma(\alpha+j)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha+\beta+j)}$
Chi-square (r)	$2^k \frac{\Gamma(\frac{r}{2}+k)}{\Gamma(\frac{r}{2})}$	$\sum_{j=0}^{k} \binom{k}{j} (-r)^{k-j} 2^j \frac{\Gamma(\frac{r}{2}+j)}{\Gamma(\frac{r}{2})}$
Exponential (θ)	$k!\,\theta^k$	$\sum_{j=0}^{k} \binom{k}{j} (-\theta)^{k-j} j!\,\theta^j$
Gamma (α,θ)	$\theta^k \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)}$	$\sum_{j=0}^{k} \binom{k}{j} (-\alpha\theta)^{k-j} \theta^j \frac{\Gamma(\alpha+j)}{\Gamma(\alpha)}$
Normal (μ,σ^2)	$\sum_{j=0}^{\lfloor k/2 \rfloor} \frac{k!}{(k-2j)!\,j!\,2^j} \mu^{k-2j}\sigma^{2j}$	$\begin{cases} 0, & k \text{ odd}, \\ (k-1)!!\,\sigma^k, & k \text{ even}. \end{cases}$
Uniform (a,b)	$\frac{b^{k+1}-a^{k+1}}{(k+1)(b-a)}$	$\frac{1}{b-a}\int_a^b \left(x-\frac{a+b}{2}\right)^k dx$
Cauchy (m,d)	undefined (no finite moments)	undefined (no mean)

4 Sample Statistics & Standardization

Sample statistics are empirical counterparts of population quantities. They are computed from data and used to estimate moments, dependence, and standardized values. As sample size grows, these estimates converge to their population limits.

Population vs Sample Quantities

Population parameters describe the underlying distribution:

$$\mu_X, \qquad \sigma_X^2, \qquad \sigma_X, \qquad \sigma_{XY}.$$

Sample statistics estimate them from observations:

$$\bar{X}_n \to \mu_X, \qquad s_X^2 \to \sigma_X^2, \qquad s_X \to \sigma_X, \qquad s_{XY} \to \sigma_{XY}.$$

Sample covariance is bounded by Cauchy–Schwarz:

$$s_{XY}^2 \le s_X^2 s_Y^2.$$

Sample Statistics (from $\mathrm{X}_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$)

Sample mean:
$$\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k,$$

Sample variance:
$$s_X^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2,$$

$$s_X = \sqrt{s_X^2},$$

Sample covariance:
$$s_{XY} = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X}_n)(Y_k - \bar{Y}_n).$$

These are empirical estimates of mean, variance, and covariance.

The factor $1/(n-1)$ in s_X^2 (instead of $1/n$) makes the sample variance an unbiased estimator of σ_X^2 when the observations are i.i.d.

Sample Mean and Sample Sum (i.i.d. case)

Assume $X_1, \ldots, X_n$ are i.i.d. with

$$E[X_i] = \mu_X, \qquad V[X_i] = \sigma_X^2 < \infty.$$

Sample mean.

$$\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k, \qquad E[\bar{X}_n] = \mu_X, \qquad V[\bar{X}_n] = \frac{\sigma_X^2}{n}.$$

The sample mean is unbiased, and its variance decreases with n.

Sample sum.

$$S_n = \sum_{k=1}^{n} X_k, \qquad E[S_n] = n\,\mu_X, \qquad V[S_n] = n\,\sigma_X^2.$$

The total scales linearly with n, both in expectation and in variance.

Both $\bar{X}_n$ and s_X^2 are unbiased estimators of their corresponding population quantities under the i.i.d. assumption.

Sample Correlation

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}.$$

This is a sample-based estimate of ρ_{XY} using empirical variances and covariance.

Standardization

Standardization rescales variables to zero mean and unit variance.

Population standardization.

$$Z = \frac{X - \mu_X}{\sigma_X}, \qquad E[Z] = 0, \quad V[Z] = 1.$$

Sample z-scores.

$$z_k = \frac{X_k - \bar{X}_n}{s_X}, \qquad w_k = \frac{Y_k - \bar{Y}_n}{s_Y}.$$

These satisfy

$$\frac{1}{n}\sum_{k=1}^{n} z_k = 0, \qquad \frac{1}{n-1}\sum_{k=1}^{n} (z_k - \bar{z})^2 = 1.$$

After standardization, sample correlation becomes the sample covariance of the z-scores:

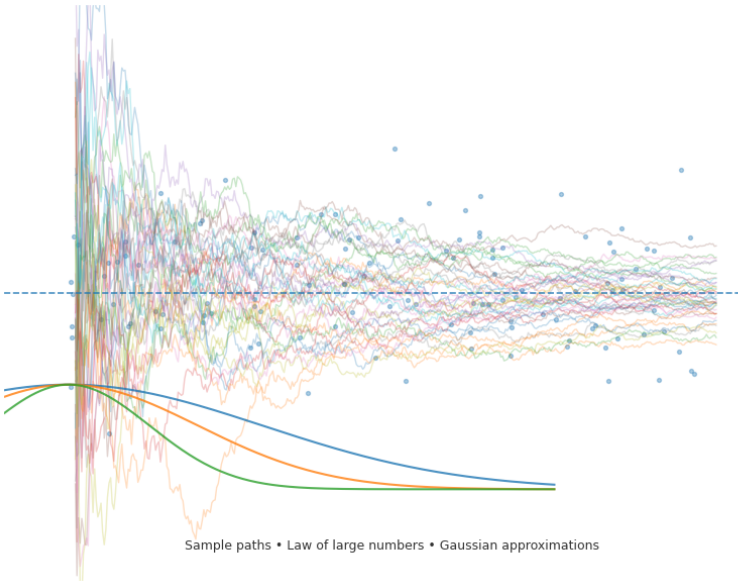
$$r_{XY} = \frac{1}{n-1}\sum_{k=1}^{n} z_k w_k.$$

Standardization enables comparison across variables, since all are brought to a common scale (of zero mean and unit variance) before further analysis.

Chapter 4: Limits, Averages & Approximations

Randomness is often described as unpredictable, yet long sequences of random outcomes exhibit a striking regularity. As observations accumulate, averages stabilize, fluctuations acquire a characteristic scale and shape, and crude uncertainty gives way to quantitative control.

This chapter formalizes that transition from noise to structure: we describe how sequences of random variables can converge, how the law of large numbers turns repeated sampling into a reliable notion of "typical behavior," and how the central limit theorem explains the ubiquitous appearance of the normal distribution.



1 Modes of Convergence

Sequences of real numbers admit a single, familiar notion of convergence. For random variables, however, there are several natural ways to express that a sequence "gets close" to a limit. Each mode of convergence captures a different balance between pointwise behavior, probabilistic control, and moment information. These distinctions matter for limit theorems, because laws like LLN and CLT are stated in specific modes.

Recall first the usual definition of convergence for a numerical sequence

$$\{a_n,\ n \geq 0\}, \quad a_n \in \mathbb{R},$$

which converges to $a^* \in \mathbb{R}$ if, for every $\varepsilon > 0$, there exists N such that $|a_n - a^*| < \varepsilon$ for all $n \geq N$.

In probability, we have a sequence of random variables

$$\{X_n,\ n \geq 0\},$$

and we want to define what it means for this sequence to converge to a random variable X. Each X_n is a measurable function from the sample space Ω to $\mathbb{R}$, so we are now tracking convergence of entire random functions rather than single numbers. This leads to several distinct, but related, notions of stochastic convergence.

Stochastic Convergence

Mode of Convergence	**Definition / Mathematical Condition**
Almost Surely (a.s.) / Probability One	$X_n \xrightarrow{o} X \quad \Leftrightarrow \quad P(\lim_{n\to\infty} X_n = X) = 1$
In Probability (p)	$X_n \xrightarrow{p} X \quad \Leftrightarrow \quad \lim_{n\to\infty} P(\lvert X_n - X\rvert > \varepsilon) = 0,\ \forall \varepsilon > 0$
In Mean Square (m.s.)	$X_n \xrightarrow{m} X \quad \Leftrightarrow \quad \lim_{n\to\infty} E\left[(X_n - X)^2\right] = 0$
In Mean of Order p	$X_n \xrightarrow{L^p} X \quad \Leftrightarrow \quad \lim_{n\to\infty} E[\lvert X_n - X\rvert^p] = 0,\ \text{for } p \geq 1$
In Distribution (d)	$X_n \xrightarrow{d} X \quad \Leftrightarrow \quad \lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ at continuity points of F_X

Conceptual Descriptions.

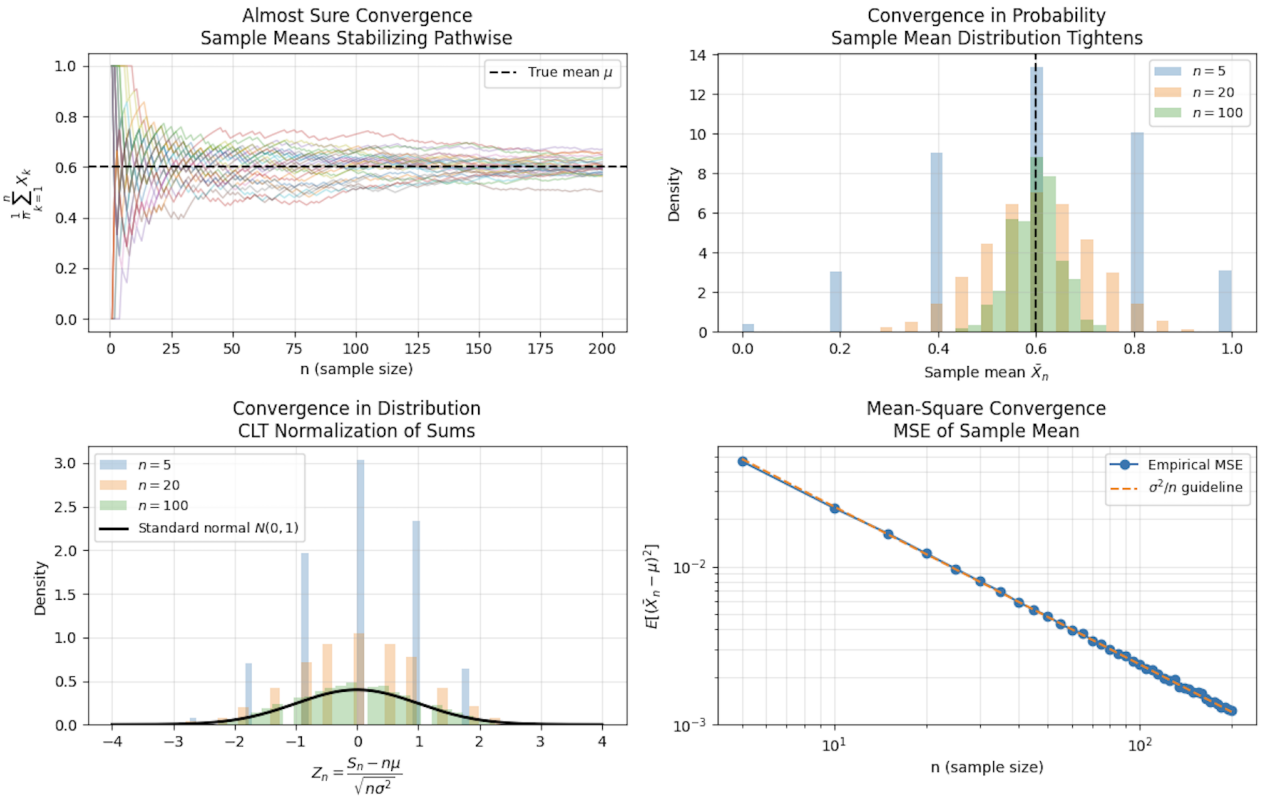
- **Almost sure convergence** requires that for almost every outcome ω, the sample-path values $X_n(\omega)$ eventually stay arbitrarily close to $X(\omega)$. It is the strongest pointwise notion and speaks about what happens on individual realizations, except on a set of probability zero.
- **Convergence in probability** demands that the probability of a noticeable deviation $|X_n - X| > \varepsilon$ becomes negligible as n grows. It does not insist on pathwise stabilization, but it ensures that large discrepancies become rare.
- **Mean-square (and L^p) convergence** controls the *size* of the error in terms of moments. Mean-square convergence requires the second moment of the error to vanish; more generally, L^p convergence requires the p-th moment of the error to go to zero. These are strong forms of convergence that quantify both proximity and tail behavior.
- **Convergence in distribution** only asks that the distributions of X_n approach the distribution of X at the level of CDFs. It does not require X_n and X to be defined on the same space or to be comparable on a sample-path basis. This is the weakest mode, but it is central in limit theorems such as the CLT.

Implication Structure. These modes of convergence are not equivalent, but there are standard implications:

$$\begin{aligned} \text{convergence in m.s.} &\Rightarrow \text{convergence in probability} \Rightarrow \text{convergence in distribution,} \\ \text{convergence a.s.} &\Rightarrow \text{convergence in probability} \Rightarrow \text{convergence in distribution.} \end{aligned}$$

Mean-square and almost sure convergence therefore provide stronger guarantees, while convergence in distribution is often the minimal notion needed to describe asymptotic shapes of distributions.

Visualizing Modes of Stochastic Convergence



Deviation Inequalities and Convergence. To turn qualitative convergence statements into quantitative control, we often need bounds on the probability of large deviations. Simple inequalities relate tail probabilities to expectations and variances, and they are the basic tools behind many proofs of convergence in probability and almost sure convergence (for example, in the law of large numbers).

Markov's Inequality. For a nonnegative random variable $X \geq 0$ and any threshold $a > 0$,

$$P(X \geq a) \leq \frac{E[X]}{a}.$$

Conceptually, if the mean of X is small relative to a, then the probability that X exceeds a must be small. Markov's inequality is very general (it uses only $E[X]$), and it is often the first step in bounding tail probabilities and establishing convergence in probability.

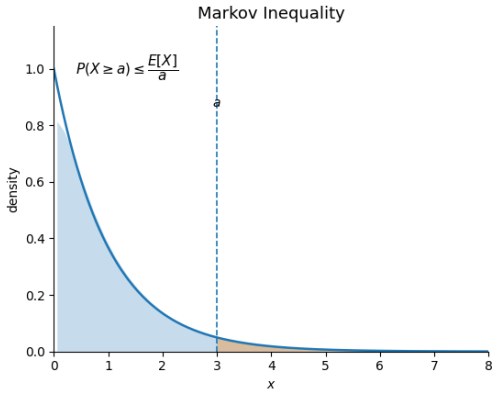
Chebyshev's Inequality. For a random variable X with finite mean $\mu = E[X]$ and variance $\sigma^2 = \mathrm{Var}(X)$,

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}, \quad k > 0.$$

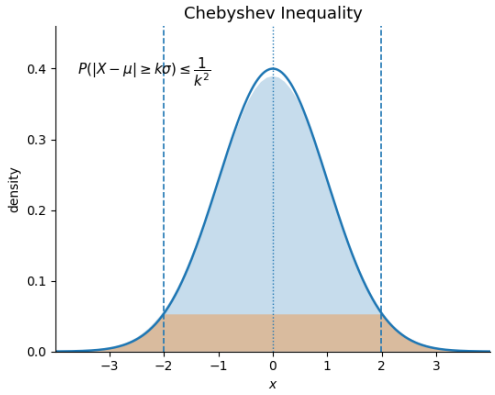
Equivalently, for any $\varepsilon > 0$,

$$P(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

Chebyshev's inequality refines Markov's inequality by using second-moment information. It quantifies how rarely X can stray far from its mean, given its variance. Applied to sample averages, Chebyshev's inequality yields a direct proof of the weak law of large numbers: as the sample size increases, the probability of a large deviation of the average from the true mean must go to zero.



Exponential density with shaded tail region $P(X \geq a)$; Markov's inequality bounds this tail probability by $E[X]/a$, a very general but often loose bound that uses only the mean.



Bell-shaped density with shaded tails $\{|X - \mu| \geq k\sigma\}$; Chebyshev's inequality bounds this probability by $1/k^2$, ensuring most of the mass lies within a few standard deviations of μ, regardless of the exact distribution.

2 Law of Large Numbers

The law of large numbers formalizes the intuitive "law of averages": when we repeat the same random experiment many times, the average outcome becomes stable and representative of the underlying distribution. Random fluctuations do not disappear, but they become increasingly small relative to the number of observations, so the sample mean behaves more and more like a deterministic quantity.

Setup. Let $X_1, X_2, \ldots$ be i.i.d. random variables with

$$\mu_X = E[X_k], \qquad \sigma_X^2 = \mathrm{Var}(X_k) < \infty.$$

The sample mean based on the first n observations is

$$\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

By linearity of expectation and basic variance rules,

$$E[\bar{X}_n] = \mu_X, \qquad \mathrm{Var}(\bar{X}_n) = \frac{\sigma_X^2}{n}.$$

Thus the sample mean is an unbiased estimator of μ_X, and its variance shrinks at rate $1/n$; the typical size of its fluctuations is of order $1/\sqrt{n}$.

Law of Large Numbers (LLN). Under the assumptions above, the sample mean converges to the population mean as $n \to \infty$:

$$\bar{X}_n \longrightarrow \mu_X.$$

This expresses that for large samples the empirical average is close to the true mean with high reliability.

Modes of convergence. Different versions of the LLN specify the sense in which $\bar{X}_n$ converges:

Law of Large Numbers Variant	**Limit Statement**
Weak LLN (in probability)	$\bar{X}_n \xrightarrow{P} \mu_X \iff \forall\, \varepsilon > 0,\ P\Big(\lvert\bar{X}_n - \mu_X\rvert > \varepsilon\Big) \xrightarrow[n\to\infty]{} 0.$
Strong LLN (almost surely)	$\bar{X}_n \xrightarrow{a.s.} \mu_X \iff P\Big(\lim_{n\to\infty} \bar{X}_n = \mu_X\Big) = 1.$
Mean-square LLN	$\bar{X}_n \xrightarrow{m} \mu_X \iff E\Big[(\bar{X}_n - \mu_X)^2\Big] \xrightarrow[n\to\infty]{} 0.$ In our i.i.d. setting, $E\Big[(\bar{X}_n - \mu_X)^2\Big] = \mathrm{Var}(\bar{X}_n) = \frac{\sigma_X^2}{n} \longrightarrow 0,$ so mean-square convergence holds automatically when $\sigma_X^2 < \infty$.

Connection to Deviation Inequalities. Using Chebyshev's inequality,

$$P\Big(\lvert\bar{X}_n - \mu_X\rvert \geq \varepsilon\Big) \ \leq\ \frac{\mathrm{Var}(\bar{X}_n)}{\varepsilon^2} = \frac{\sigma_X^2}{n\varepsilon^2},$$

which tends to zero as $n \to \infty$. This provides a simple proof of the weak LLN: as we collect more data, the probability of a noticeable deviation of $\bar{X}_n$ from μ_X becomes arbitrarily small.

All versions of the LLN capture the same qualitative message: empirical averages are reliable summaries of long-run behavior. For large n, the sample mean $\bar{X}_n$ is tightly concentrated around μ_X, and limit theorems such as the central limit theorem (next section) refine this picture by describing the shape and scale of its remaining fluctuations.

3 Central Limit Theorem & Standardization

The law of large numbers explains that sample averages stabilize around their mean; it does not, however, describe *how* they fluctuate. The central limit theorem (CLT) refines this picture by showing that after proper rescaling, these fluctuations acquire a universal, approximately normal shape. This is what justifies the use of Gaussian approximations, z-scores, and standard normal tables in probability and statistics.

Setup. Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mu_X = E[X_k]$ and $\sigma_X^2 = \mathrm{Var}(X_k) < \infty$, and define the sample mean and sum

$$\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k, \qquad S_n = \sum_{k=1}^{n} X_k.$$

From the previous section (LLN) we know that $\bar{X}_n \to \mu_X$ as $n \to \infty$, and that $\mathrm{Var}(\bar{X}_n) = \sigma_X^2/n$. The CLT describes the limiting *distribution* of the centered and scaled quantities

$$\sqrt{n}\,(\bar{X}_n - \mu_X) \quad \text{or} \quad \frac{S_n - n\mu_X}{\sqrt{n}}.$$

Central Limit Theorem (CLT). Under the i.i.d. finite-variance assumptions above,

$$Z_n = \frac{\bar{X}_n - \mu_X}{\sigma_X/\sqrt{n}} = \frac{\sqrt{n}\,(\bar{X}_n - \mu_X)}{\sigma_X} \xrightarrow{d} Z \sim \mathcal{N}(0,1),$$

or equivalently,

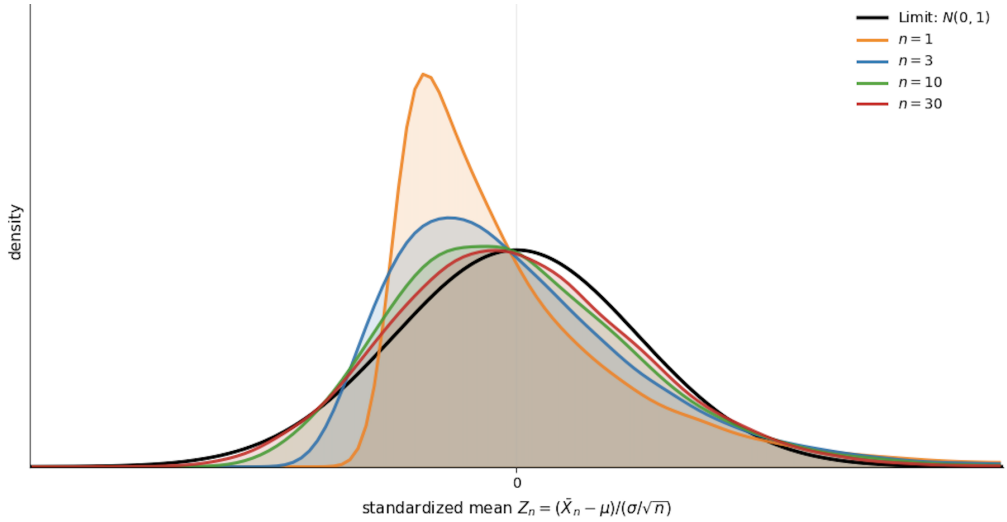
$$Z_n = \frac{S_n - n\mu_X}{\sqrt{n}\,\sigma_X} \xrightarrow{d} \mathcal{N}(0,1).$$

That is, as $n \to \infty$, the standardized sample mean (or sum) converges *in distribution* to a standard normal random variable. For moderately large n, we can already use the approximation

$$\frac{\bar{X}_n - \mu_X}{\sigma_X/\sqrt{n}} \approx Z \sim \mathcal{N}(0,1),$$

even when the individual X_k are not Gaussian.

CLT: Standardized Sample Means from a Skewed Distribution



As n increases, the curves collapse onto $N(0,1)$, illustrating the central limit theorem in action.

Notes.

- The LLN tells us $\bar{X}_n \to \mu_X$, so the average settles near its mean.
- The CLT adds that the *shape* of the fluctuations around μ_X is approximately normal, with typical size of order $\sigma_X/\sqrt{n}$.
- **Convergence is in distribution:** Probabilities for events involving $\bar{X}_n$ can be approximated using the standard normal CDF for large n, but there is no pathwise convergence of Z_n to a fixed random variable.

Standardization. Standardization is the simple affine transformation that converts a random variable to zero mean and unit variance. It is the basic operation that connects raw variables to the standard normal distribution in the CLT.

Single observation.

$$Z = \frac{X - \mu_X}{\sigma_X}, \qquad E[Z] = 0, \quad \mathrm{Var}(Z) = 1.$$

If $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$, then $Z \sim \mathcal{N}(0,1)$ *exactly*. For general X, this is just a rescaling; no normality is implied.

Sample mean.

$$Z = \frac{\bar{X}_n - \mu_X}{\sigma_X/\sqrt{n}}, \qquad E[Z] = 0, \quad \mathrm{Var}(Z) = 1.$$

By the CLT, for large n,

$$Z \approx \mathcal{N}(0,1),$$

so probabilities involving $\bar{X}_n$ can be approximated using standard normal quantiles. This is the foundation for confidence intervals and many classical hypothesis tests.

Sample sum.

$$Z = \frac{S_n - n\mu_X}{\sigma_X\sqrt{n}}, \qquad E[Z] = 0, \quad \mathrm{Var}(Z) = 1,$$

with the same asymptotic normality:

$$Z \approx \mathcal{N}(0,1) \quad \text{for large } n.$$

This form is convenient when working directly with aggregate quantities (e.g., total claims, total demand, total return).

Standardization extracts a dimensionless, unit-variance view of random fluctuations. The central limit theorem then asserts that, under mild conditions, the standardized fluctuations of sums and averages become approximately Gaussian.

4 Confidence Intervals and Sample Size

The central limit theorem turns the sample mean into an approximately normal quantity after standardization. Confidence intervals exploit this fact in the reverse direction: instead of asking for the distribution of $\bar{X}_n$ given μ_X, we construct a random interval around $\bar{X}_n$ that is designed to contain the unknown mean μ_X with a prescribed long-run frequency.

Setup. When σ_X is known and n is sufficiently large, the CLT gives the approximation

$$Z_n = \frac{\bar{X}_n - \mu_X}{\sigma_X/\sqrt{n}} \approx Z \sim \mathcal{N}(0,1).$$

Confidence intervals for the mean (known variance). For a desired confidence level $1-\alpha$, we choose $z_{\alpha/2}$ so that

$$P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1-\alpha, \quad Z \sim \mathcal{N}(0,1).$$

Using the CLT approximation for Z_n,

$$P\left(-z_{\alpha/2} \le \frac{\bar{X}_n - \mu_X}{\sigma_X/\sqrt{n}} \le z_{\alpha/2}\right) \approx 1-\alpha.$$

Rearranging the inequalities yields

$$P\left(\bar{X}_n - z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} \le \mu_X \le \bar{X}_n + z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}\right) \approx 1-\alpha.$$

Thus a $(1-\alpha)$ confidence interval for μ_X is

$$\mu_X \in \left[\bar{X}_n - z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}},\ \bar{X}_n + z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}\right]$$

The confidence level $1-\alpha$ is the long-run proportion of intervals that would contain μ_X if we repeatedly drew new samples of size n and rebuilt the interval each time. For a given dataset, the interval is random while μ_X is fixed; the statement concerns the procedure's coverage, not a probability distribution on μ_X.

Determining sample size. Often we specify both a target confidence level $(1-\alpha)$ and a desired precision $\varepsilon > 0$, interpreted as the maximum acceptable half-width:

$$z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} \le \varepsilon.$$

Solving for n gives the sample size requirement

$$\varepsilon = z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} \quad \Longrightarrow \quad n \ge \left(\frac{z_{\alpha/2}\sigma_X}{\varepsilon}\right)^2.$$

Larger confidence levels (smaller α) or smaller tolerances ε require larger n, reflecting the basic trade-off between certainty and precision.

Common confidence levels and critical values.

Confidence level	$\boldsymbol{\alpha}$	$\boldsymbol{z_{\alpha/2}}$
90%	0.10	1.645
95%	0.05	1.960
99%	0.01	2.576

These z-values are used throughout to translate desired coverage probabilities into concrete margins of error and sample size requirements via the normal approximation.

Chapter 5: Transformations & Generating Functions

Transformations and generating functions describe how probability laws respond to algebraic operations. Transformations track how distributions are reshaped under mappings of one or several random variables, preserving probability while redistributing mass and density. Generating functions encode an entire distribution into a single analytic object, turning operations on random variables into differentiation, multiplication, and limits. Together, these tools replace direct manipulation of densities with structural rules, providing a concise way to analyze sums, moments, and distributional behavior.

1 Transformations of Random Variables

Transformations describe how probability distributions change under deterministic mappings of random variables. Probability itself is conserved, but its representation, mass in the discrete case and density in the continuous case, is reshaped by the geometry of the transformation. The guiding principle is invariance of probability: events and their transformed images must carry the same probability.

Discrete Transformations. Let X be a discrete random variable with pmf $p_X(x)$, and suppose

$$Y = g(X)$$

is a one-to-one transformation. Each value y of Y corresponds uniquely to a value $x = g^{-1}(y)$ of X, so probability mass is transferred directly:

$$p_Y(y) = p_X(g^{-1}(y)).$$

When the transformation is not invertible or when the pmf of X is inconvenient to compute directly, the distribution of Y can always be obtained via the cumulative distribution function:

$$F_Y(y) = P(Y \le y) = P(g(X) \le y),$$

which remains valid regardless of the form of g.

Continuous Transformations. Let X be a continuous random variable with pdf $f_X(x)$, and let

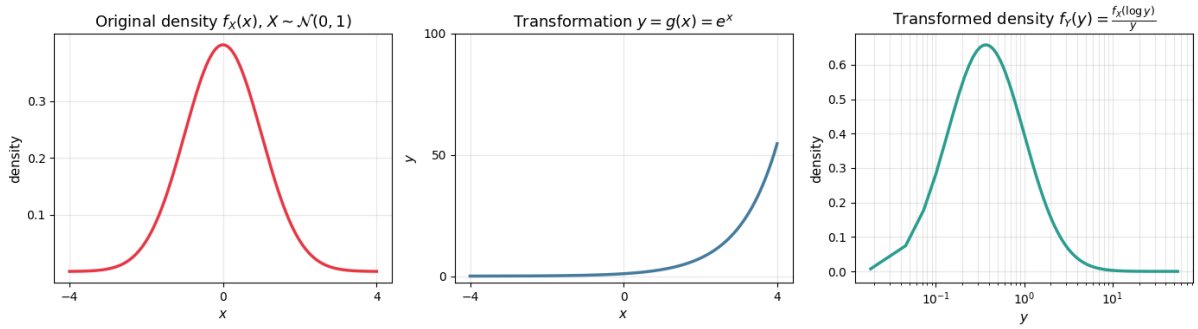
$$Y = g(X)$$

where g is monotonic and differentiable. Probability density transforms according to the change of variables:

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dx}{dy} \right|, \qquad x = g^{-1}(y).$$

The absolute derivative accounts for how intervals in the x-space expand or contract when mapped into the y-space. Regions where g stretches space correspond to lower density, while compression increases density, ensuring total probability is preserved.

Monotone Continuous Transformation



Monotone change of variables illustrating density reshaping under $y = g(x) = e^x$.

If g is *not* monotonic, multiple values of X may map to the same value of Y. In this case, the density of Y is obtained by summing the contributions from all preimages $\{x_k\}$ satisfying $y = g(x_k)$:

$$f_Y(y) = \sum_k f_X(x_k) \left| \frac{dx}{dy} \right|_{x=x_k}.$$

Each branch contributes additively, reflecting the fact that probability mass from different regions of the original space accumulates at the same point in the transformed space.

Joint Transformations and Change of Variables. Transformations naturally extend to multiple random variables. Let (X, Y) be a pair of continuous random variables with joint pdf $f_{X,Y}(x, y)$, and define new variables

$$U = g(X, Y), \qquad V = h(X, Y),$$

where the transformation is invertible with inverse

$$x = x(u, v), \qquad y = y(u, v).$$

The joint pdf of (U, V) is then

$$f_{U,V}(u, v) = f_{X,Y}(x(u, v),\, y(u, v)) \left| \frac{d(x, y)}{d(u, v)} \right|.$$

This formula generalizes the one-dimensional change of variables and ensures that probability assigned to regions in the (x, y)-plane matches the probability assigned to their images in the (u, v)-plane.

Jacobian Determinant. The Jacobian determinant is defined as

$$\frac{d(x, y)}{d(u, v)} = \det \begin{bmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\[2ex] \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{bmatrix}.$$

It measures the local area distortion induced by the transformation. Values greater than one indicate local expansion, while values less than one indicate compression. In higher dimensions, the Jacobian plays the same role, scaling volume elements so that total probability remains invariant under coordinate changes.

Formula Summary Table

Transformation Type	**Distribution Formula**
Discrete, one-to-one	$p_Y(y) = p_X(g^{-1}(y))$
Discrete (via CDF)	$F_Y(y) = P(Y \le y) = P(g(X) \le y)$
Continuous, monotone	$f_Y(y) = f_X(g^{-1}(y)) \left\lvert \dfrac{dx}{dy} \right\rvert$
Continuous, non-monotone	$f_Y(y) = \displaystyle\sum_k f_X(x_k) \left\lvert \frac{dx}{dy} \right\rvert_{x=x_k}, \quad y = g(x_k)$
Two-variable transformation	$f_{U,V}(u, v) = f_{X,Y}(x(u, v), y(u, v)) \left\lvert \dfrac{d(x, y)}{d(u, v)} \right\rvert$
Jacobian determinant	$\dfrac{d(x, y)}{d(u, v)} = \det \begin{bmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\[2ex] \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{bmatrix}$

2 Moment Generating Functions & Characteristic Functions

Transformations describe how distributions change under mappings of random variables. Generating functions provide a complementary perspective; instead of reshaping probability densities directly, they encode an entire distribution into a single analytic object. Operations on random variables then correspond to simple algebraic operations on these functions. This shift from densities to functions is especially powerful for studying moments, sums of independent variables, and distributional limits.

Moment Generating Function (MGF). The moment generating function of a random variable X is defined as

$$M_X(s) = E\Big[e^{sX}\Big],$$

whenever the expectation exists in an open neighborhood of $s = 0$. If it exists, the MGF uniquely determines the distribution of X.

The defining feature of the MGF is that its derivatives at the origin recover the raw moments of the distribution:

$$E[X^k] = M_X^{(k)}(0) = \left.\frac{d^k M_X(s)}{ds^k}\right|_{s=0}.$$

Thus, the entire sequence of moments is encoded in the local behavior of $M_X(s)$ near zero.

Remark: Raw vs. Central Moments. The moments generated by $M_X(s)$ are raw moments $E[X^k]$, not central moments. Central moments measure deviations from the mean,

$$E\Big[(X - E[X])^k\Big],$$

and must be obtained by algebraic manipulation of raw moments. Only when $E[X] = 0$ do raw and central moments coincide.

Characteristic Function (CF). The characteristic function of X is defined as

$$\phi_X(\omega) = E\Big[e^{i\omega X}\Big], \qquad \omega \in \mathbb{R}.$$

Unlike the MGF, the characteristic function always exists for every random variable, regardless of tail behavior. It plays an analogous role to the MGF but operates in the complex domain and is closely related to the Fourier transform of the distribution.

Relationship Between MGFs and CFs. When both exist, the MGF and CF are analytically connected. Moments can be recovered from either representation:

$$E[X^k] = \left.\frac{d^k}{ds^k} M_X(s)\right|_{s=0} = \frac{1}{i^k}\left.\frac{d^k}{d\omega^k}\phi_X(\omega)\right|_{\omega=0}.$$

In practice, MGFs are convenient when they exist, while CFs provide a universally valid alternative.

Key Structural Properties. Generating functions turn operations on random variables into algebraic rules:

1. *Additivity for independent sums.* If X and Y are independent,

$$M_{X+Y}(s) = M_X(s)\, M_Y(s), \qquad \phi_{X+Y}(\omega) = \phi_X(\omega)\, \phi_Y(\omega).$$

2. *Sums of i.i.d. variables.* For $S_n = \sum_{k=1}^n X_k$ with X_k i.i.d.,

$$M_{S_n}(s) = [M_X(s)]^n.$$

3. *Sample mean.* For $\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$,

$$M_{\bar{X}_n}(s) = [M_X(s/n)]^n.$$

4. *Linear transformations.* If $Y = aX + b$,

$$M_Y(s) = e^{sb}\, M_X(as), \qquad \phi_Y(\omega) = e^{i\omega b}\, \phi_X(a\omega).$$

Uses of Generating Functions.

- Compact derivation of moments
- Efficient analysis of sums of independent variables
- Distributional identification via uniqueness
- Proofs of limit theorems through functional convergence

Lévy's Continuity Theorem. If the characteristic functions of a sequence $\{X_n\}$ converge pointwise to the characteristic function of X,

$$\phi_{X_n}(\omega) \to \phi_X(\omega),$$

and ϕ_X is continuous at the origin, then

$$X_n \xrightarrow{d} X.$$

An analogous statement holds for MGFs when they exist in a neighborhood of zero.

Convergence of generating functions implies convergence in distribution. This principle underlies many classical limit results, including the Central Limit Theorem, and explains why generating functions serve as a bridge between finite-sample distributions and asymptotic behavior.

Formula Summary Table

Object / Operation	**Formula**
Moment generating function	$M_X(s) = E\left[e^{sX}\right]$
Raw moments from MGF	$E[X^k] = \left.\dfrac{d^k M_X(s)}{ds^k}\right\|_{s=0}$
Characteristic function	$\phi_X(\omega) = E\left[e^{i\omega X}\right]$
Moments from CF	$E[X^k] = \left.\dfrac{1}{i^k}\,\dfrac{d^k \phi_X(\omega)}{d\omega^k}\right\|_{\omega=0}$
Lévy continuity (MGF / CF)	$M_{X_n}(s) \to M_X(s)$ or $\phi_{X_n}(\omega) \to \phi_X(\omega) \;\Rightarrow\; X_n \xrightarrow{d} X$
Independent sum	$M_{X+Y}(s) = M_X(s)\, M_Y(s)$
Sum of i.i.d. variables	$M_{S_n}(s) = [M_X(s)]^n$
Sample mean	$M_{\bar{X}_n}(s) = [M_X(s/n)]^n$
Linear transformation	$M_{aX+b}(s) = e^{sb}\, M_X(as)$

Moment Generating Functions (MGFs)

Distribution	$M_X(s)$
Bernoulli (p)	$1 - p + pe^s$
Binomial (n, p)	$(1 - p + pe^s)^n$
Geometric (p)	$\frac{pe^s}{1-(1-p)e^s}, \quad s < -\ln(1-p)$
Poisson (λ)	$\exp[\lambda(e^s - 1)]$
Negative Binomial (r, p)	$\left(\frac{p}{1-(1-p)e^s}\right)^r, \quad s < -\ln(1-p)$
Discrete Uniform $\{1, \ldots, n\}$	$\frac{e^s(1-e^{ns})}{n(1-e^s)}$
Exponential (θ)	$\frac{1}{1-\theta s}, \quad s < \frac{1}{\theta}$
Gamma (α, θ)	$(1-\theta s)^{-\alpha}, \quad s < \frac{1}{\theta}$
Chi-square (r)	$(1-2s)^{-r/2}, \quad s < \frac{1}{2}$
Normal (μ, σ^2)	$\exp\left(\mu s + \frac{1}{2}\sigma^2 s^2\right)$
Uniform (a, b)	$\frac{e^{sb} - e^{sa}}{s(b-a)}$
Beta (α, β)	${}_1F_1(\alpha; \alpha+\beta; s)$ (via confluent hypergeometric function)
Cauchy (m, d)	does not exist (diverges for all $s \neq 0$)