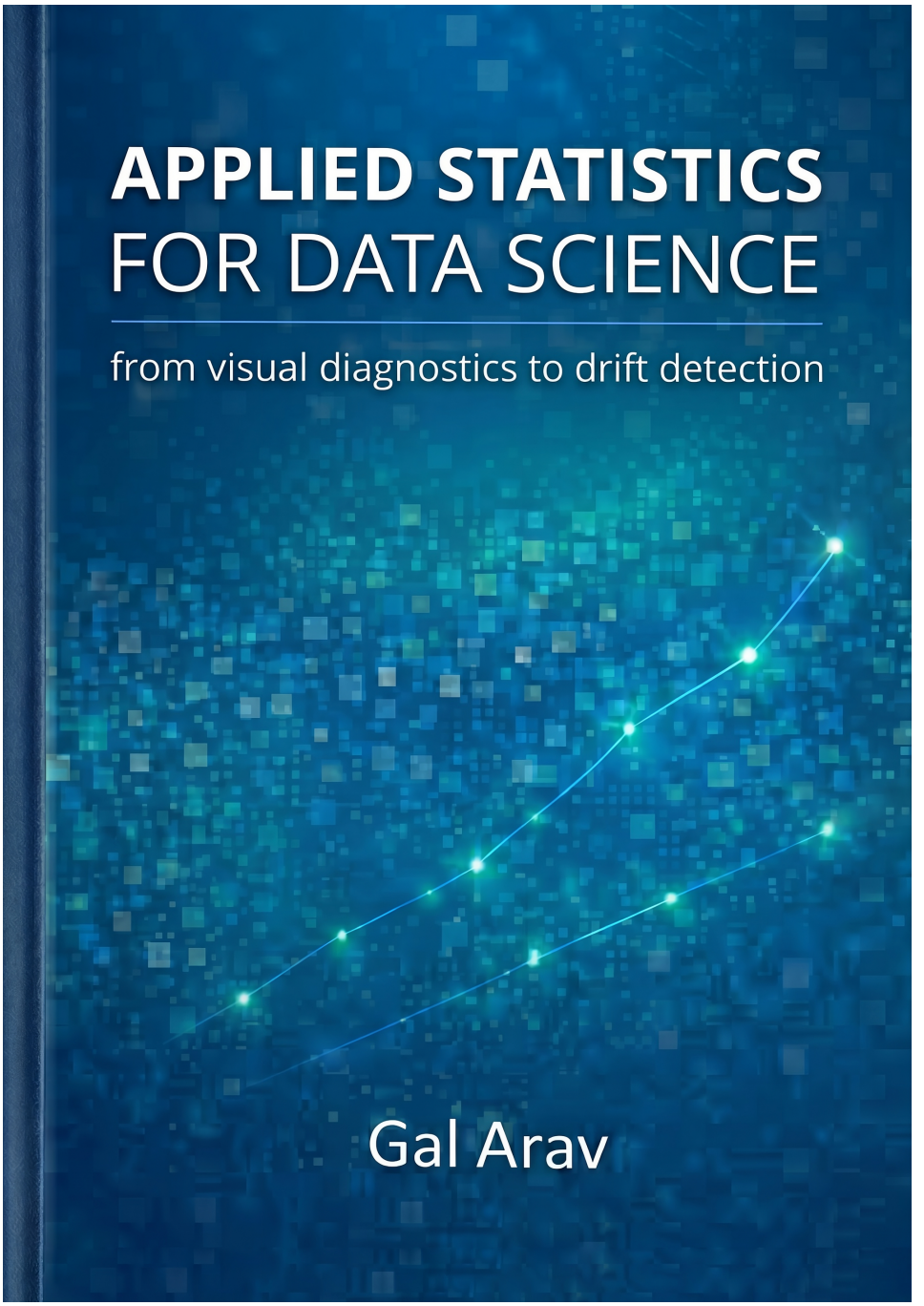


Copyright and Permissions

Permissions, Contact & Feedback

For permissions, requests, or inquiries, please contact the author via LinkedIn: Gal Arav

Readers are warmly encouraged to share feedback, suggest improvements or report any errors they discover. Your feedback will help make future editions clearer, more accurate and more useful for all readers.

Applied Statistics for Data Science

from visual diagnostics to drift detection

by Gal Arav M.Sc.

Author's website: qikly.com

Build Version, Date, Type: 1.0.2-free, 2026-06-18, Free Leanpub Edition

ISBN: [ISBN Number]

First Edition, 2026 — Self-Published

For errata and updates to this edition, please visit https://qikly.com/updates

Introduction

Modern machine learning relies on a solid grasp of probability, uncertainty, sampling behavior, and the ways data changes over time. This book develops that understanding through visual intuition and simulation-based reasoning, and it is supported by clear Python examples available on GitHub for each chapter.

Statistical thinking is presented as a practical tool that supports every stage of modeling, from selecting appropriate probability models to diagnosing prediction issues and monitoring systems in production. The aim is to make the core ideas of statistics feel intuitive and useful rather than abstract or overwhelming.

How this book is structured

This book is organized into three major parts as follows:

- **Part I: Foundations**
  - Chapter 1: Random Events, Variables & Probability Modeling
  - Chapter 2: Distribution Families & Shapes
  - Chapter 3: Sampling & Estimators
- **Part II: Core Statistical Tools**
  - Chapter 4: Hypothesis Testing & Statistical Comparison Methods
  - Chapter 5: Regression & Prediction Diagnostics
  - Chapter 6: Sampling Designs & Experiments
  - Chapter 7: Resampling & Permutation Based Inference
- **Part III: Drift, Reliability & Temporal Behavior**
  - Chapter 8: Nonparametric Drift Detection & Monitoring
  - Chapter 9: Parametric Drift Detection & Monitoring
  - Chapter 10: Survival Curves & Reliability Modeling

You can read the book from start to finish, but each chapter is designed to stand on its own. Code examples and visualizations make it easy to experiment as you go. Readers working in production environments may find themselves jumping directly to Parts II and III.

Who should read this book

This book is written for anyone who works with data and wants to understand why statistical methods behave the way they do.

- **Data scientists and machine learning practitioners** who want stronger intuition behind the tools they use every day.
- **Engineers and analysts** who evaluate models, design experiments, or monitor systems in production.
- **Researchers, students and curious enthusiasts** who want a practical bridge between mathematical statistics and applied machine learning.
- **Professionals in reliability, operations, or risk modeling** who need to understand drift, survival curves, and temporal behavior through simple explanations and practical examples.

You do not need an advanced math background. Curiosity and a willingness to explore ideas through examples and simulation are enough.

How this book is different from other books on machine learning

Most machine learning texts focus on models, architectures, and performance benchmarks. This book emphasizes the statistical foundations that make those models reliable. Key differences:

- **Simulation first explanations** Concepts are demonstrated through repeated sampling, Python code, and visualizations rather than abstract proofs.
- **Model agnostic reasoning** The goal is not to teach a specific algorithm but to strengthen the statistical intuition that applies to all algorithms.
- **Practical diagnostics** You will learn how to evaluate models, detect failures, and design experiments that work in real systems.
- **Focus on uncertainty and temporal behavior** Drift detection, reliability modeling, and survival analysis are treated as key topics once the foundational ideas are in place.

Supporting website and code repository

All figures, simulations, and extended examples are available on the supporting website at qikly.com, along with a complete code repository. These resources allow you to reproduce every plot, explore variations, and experiment with the ideas introduced in each chapter. The Python code was developed using Google Colab, which provides a freely available, "zero-setup" environment for running Jupyter notebooks with popular statistics and data-science packages preinstalled. For this reason, Colab is the recommended platform for working through the examples in this book.

Chapter 1: Random Events, Variables & Probability Modeling

Introduction

This first chapter introduces the core ideas needed to understand how data behaves. We begin with random events and random variables and the basic quantities that describe them, such as expectation, variance and long run behavior. These ideas provide the foundation for thinking about uncertainty and variation.

We then move to an introduction to probability distributions. Using examples like the Uniform and Normal distributions, along with visual tools such as histograms, density estimates and cumulative distribution functions, we build intuition for concepts like location, spread, shape and tail behavior. Mixture models show how complex patterns can arise from simple components.

Next we connect these ideas to real data through the distinction between populations and samples. We look at how parameters describe underlying processes, how statistics summarize observed data and how sampling variability leads to results such as the Law of Large Numbers and the Central Limit Theorem. Simulation helps make these ideas concrete.

Finally we explore core probability distributions that model common random mechanisms, including Bernoulli trials, Binomial counts, Poisson arrivals and waiting time models such as the Geometric and Exponential distributions. These models show how simple rules can generate the wide range of patterns seen in practice.

A review of the Probability Laws and Counting Rules is provided in the Appendix for readers who wish to revisit these basic topics.

Random Events & Variables

When we cannot be sure how an outcome will turn out, probability gives us a way to describe the uncertainty. This section introduces the main ideas behind that description.

What Are Random Events?

Random events are outcomes that cannot be predicted with certainty. Each individual outcome is uncertain, but when the same situation is repeated many times, the overall pattern becomes stable and follows consistent probability laws.

Example (rolling a die):

Possible outcomes: {1, 2, 3, 4, 5, 6}

Random event: "The die shows an even number" → {2, 4, 6}

Random event: "The die shows a 6" → {6}

Random Variables: Discrete vs Continuous

To analyze random events mathematically, we represent them using **random variables**. These are numerical quantities whose values are determined by the **outcome** of a **random process**. A random variable provides a consistent numerical label for each possible outcome, allowing randomness to be summarized and modeled through probability distributions.

Probability models describe random processes by assigning probabilities or densities to the possible values of a random variable. These probabilities are numbers between 0 and 1 that reflect the chance of each outcome relative to all possible outcomes. For a quick review of the basic probability laws, readers may consult Appendix A.

Random variables may be **discrete**, such as counts or cycles, or **continuous**, such as time or measurement. They are typically denoted by symbols like X or T and are described by their distribution and the set of all possible outcomes, known as their **support**.

Discrete Examples of Random Variables

(a) Fair coin toss: $X = 1$ for Heads, 0 for Tails; Support: $\{0, 1\}$
→ Probabilities: $P(X = 1) = 1/2$, $P(X = 0) = 1/2$

(b) Fair die roll: X is the face value shown; Support: $\{1, 2, 3, 4, 5, 6\}$
→ Probabilities: $P(X = k) = 1/6$ for $k = 1, \ldots, 6$

In any discrete probability model, the probabilities assigned to all possible outcomes must sum to 1:

$$\sum_i P[X = x_i] = 1$$

X is a random variable
x_i are the possible values (outcomes) of X
$P[X = x_i]$ is the probability that X takes the value x_i

For example, for a fair die:

$$X \in \{1, 2, 3, 4, 5, 6\}$$
$$P[X = x_i] = \frac{1}{6} \text{ for each } x_i$$
$$\sum_{i=1}^{6} P[X = x_i] = 6 \times \frac{1}{6} = 1$$

Continuous Examples of Random Variables

(c) Waiting time (unbounded): t is the time until the next event; Support: $[0, \infty) \rightarrow$ Density: $f(t) = \lambda e^{-\lambda t}$ for $t \geq 0$ with λ as the event-rate parameter (Exponential model)

(d) Measurement within bounds: t is a temperature between 0 and 100; Support: $[0, 100] \rightarrow$ Density: $f(t) = 1/100$ for $0 \leq t \leq 100$ (Uniform model)

As in the discrete case, a continuous probability model must account for all possible outcomes, but now the total probability is obtained by integrating the density, and this integral must equal 1:

$$\int_{-\infty}^{\infty} f_T(t)\, dt = 1$$

T is a continuous random variable
t is a possible value of T
$f_T(t)$ is the probability density function (PDF) of T

For example, for a uniform temperature between 0 and 100:

$$T \in [0, 100]$$

$$f(t) = \frac{1}{100} \quad \text{for } 0 \leq t \leq 100$$

$$\int_0^{100} f(t)\, dt = \int_0^{100} \frac{1}{100}\, dt = 1$$

Note: Unlike discrete probabilities, density values $f_T(t)$ do not represent probabilities by themselves; only integrals over intervals do.

Expectation E(X)

Expectation represents the long-run average value of a random variable obtained by repeating the random process many times. Note that μ is simply another name for $E(X)$.

For a **discrete** random variable X taking values x_i with probabilities p_i:

$$E(X) = \sum_i x_i\, p_i$$

A concise way to see where the discrete formula comes from is to imagine repeating the experiment many times. In the long run, the proportion of times X takes value x_i approaches p_i, which matches the structure of the expectation formula above.

For a **continuous** random variable with probability density $f(x)$:

$$E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

For continuous variables, the integral computes the average by weighting each value according to its density.

Example Calculation: E(X) for a Fair Die

Probability for each face: $p_i = \frac{1}{6}$ for $i = 1, \ldots, 6$

$$E(X) = 1p_1 + 2p_2 + 3p_3 + 4p_4 + 5p_5 + 6p_6$$

$$E(X) = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right)$$

$$E(X) = \frac{1+2+3+4+5+6}{6}$$

$$E(X) = \frac{21}{6} = 3.5$$

Variance Var(X)

Variance measures how far the values of a random variable tend to deviate from their mean. Note that σ^2 is simply another name for $\mathrm{Var}(X)$, and σ is its square root.

It is defined as:

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

Expanding the square gives:

$$(X-\mu)^2 = X^2 - 2\mu X + \mu^2, \qquad \mu = E(X)$$

Taking expectations term by term:

$$\mathrm{Var}(X) = E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - \mu^2$$

Thus the **computational formula** is:

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

Example Calculation: Var(X) for a Fair Die

Probability for each face: $p_i = 1/6$ for $i = 1, \ldots, 6$

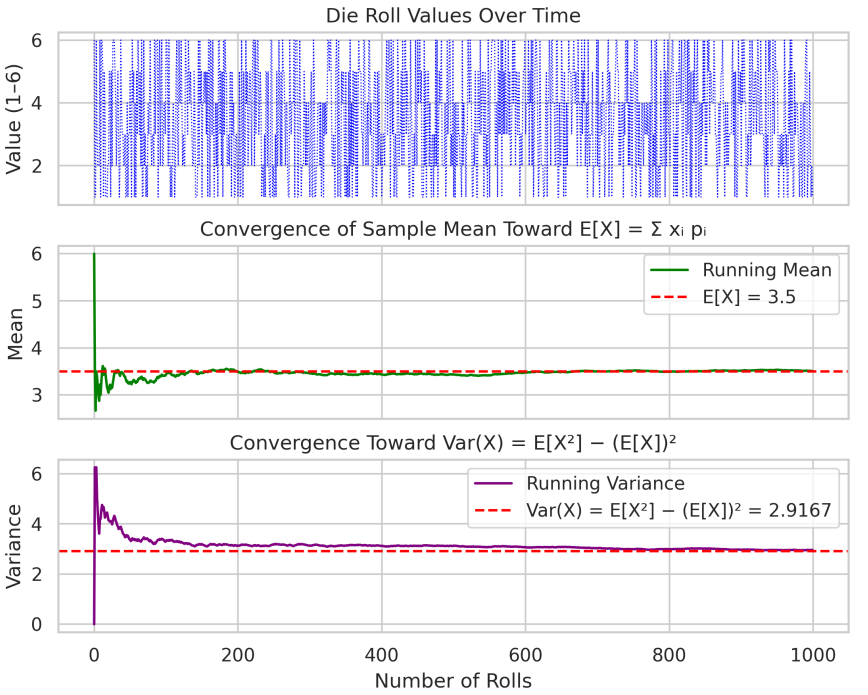
- $\mathrm{E}[X] = (1 + 2 + 3 + 4 + 5 + 6)/6 = 21/6 = 3.5$
- $\mathrm{E}[X^2] = (1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2)/6 = 91/6 \approx 15.1667$
- $(\mathrm{E}[X])^2 = 3.5^2 = 12.25$
- $\mathrm{Var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = 15.1667 - 12.25 = 2.9167$

Long Run Behavior of E(X) and Var(X)

Understanding a random variable begins with two key ideas: its long-run average and the variability around that average.

For something simple like rolling a fair die, these ideas emerge from its **independent** and **equal probability structure**, where each value in the support {1, 2, 3, 4, 5, 6} is equally likely, and from the natural spread created by those equally likely outcomes. By simulating many rolls, we can watch these quantities settle into their true **long run behavior**, providing a concrete illustration of the Law of Large Numbers (formally defined later in the chapter).

Example Long Run Behavior of a Fair Die



Here we see three key patterns emerge as the number of rolls increases:

1. Raw values of a discrete random variable: an independent, equally likely outcome whose support is the set $\{1, 2, 3, 4, 5, 6\}$, with probability $1/6$ assigned to each value.
2. Convergence of the sample mean toward $\mathrm{E}[X] = \sum x_i p_i = 3.5$
3. Convergence of the sample variance toward $\mathrm{Var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = 2.9167$

i.i.d. Samples

Many probability models rely on the assumption that observations are **independent and identically distributed (i.i.d.)**.

- **Independence** means that one observation does not influence another.
- **Identical distribution** means that all observations arise from the same underlying probability law.

Together, these conditions make it possible to describe data using simple probability models and to reason clearly about uncertainty.

In practice, it is important to check for potential dependence, such as time-ordered data or batch effects, before applying methods that rely on the i.i.d. assumption.

Introduction to Probability Distributions

Probability distributions describe uncertain outcomes by distinguishing continuous processes from those that occur in discrete, countable increments.

Continuous & Discrete Distributions

Continuous distributions have a **PDF (Probability Density Function)**, which describes how probability is spread over a continuum and always integrates to 1 over its entire support (the full range of continuous values). A PDF represents **density** rather than literal probability at a point.

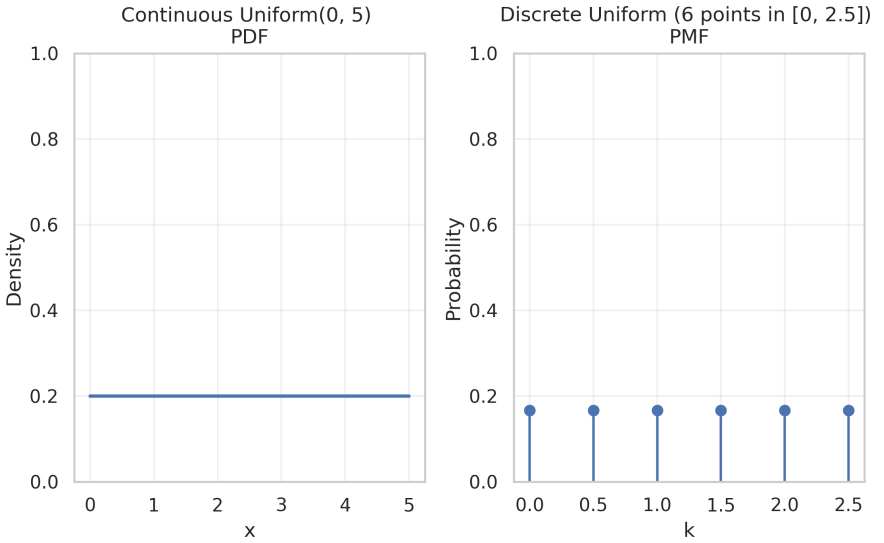
Discrete distributions have a **PMF (Probability Mass Function)**, which assigns probability to individual countable values and always sums to 1 across all possible outcomes. A PMF gives the **individual probability** of each discrete value.

Support

The **support** of a probability distribution is the set of all values where the distribution assigns non-zero probability for discrete distributions or non-zero density for continuous distributions.

Example PDF vs PMF

Many common families include both continuous and discrete forms, reflecting similar structural ideas expressed in different domains. The chart below shows an example for the **Uniform distribution**.



In this figure, the y-axis shows how **probability** is allocated:

- **For the continuous Uniform PDF**, the probability density ("Density") is $p(x) = 0.2$, a constant probability value of 1/5 over the interval $[0, 5]$. The total probability (the area under the curve) equals 1.
- **For the discrete Uniform PMF**, the probability mass ("Probability") for the 6 sample points is $p(k) = 1/6 \approx 0.1666$ for each of the discrete sample outcomes in $\{0, 0.5, 1, 1.5, 2, 2.5\}$. The total probability is the sum of all probability masses, which equals 1.

Uniform Distribution

Now that we have some intuition for the Uniform Distribution's PDF and PMF, we can formalize the definition.

Continuous Uniform Distribution

$$X \sim \text{Unif}(a, b)$$

$$\text{Support: } x \in [a, b]$$

$$f_X(x) = \frac{1}{b-a}, \quad a \le x \le b$$

$$a < b, \quad X \text{ is equally likely anywhere in } [a, b]$$

Discrete Uniform Distribution

$$X \sim \text{Unif}\{1, \dots, n\}$$

$$\text{Support: } k \in \{1, 2, \dots, n\}$$

$$\Pr(X = k) = \frac{1}{n}, \quad k = 1, 2, \dots, n$$

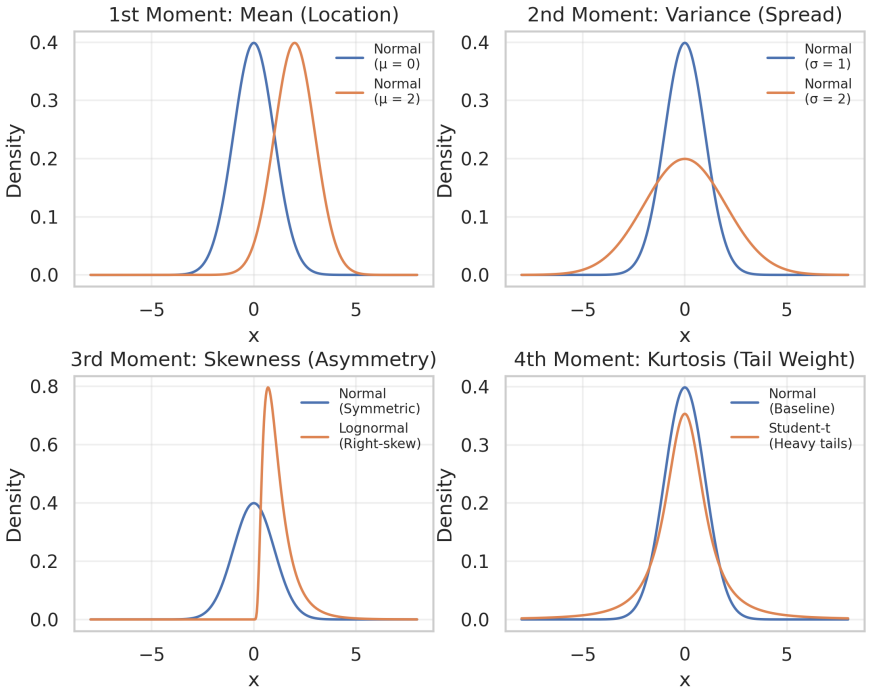
$$k \text{ is one of the } n \text{ equally likely discrete outcomes}$$

The **Uniform distribution** describes situations where all possible outcomes have the same likelihood, whether the values form a continuous interval or a finite set of discrete points. In the continuous case every value in the interval $[a, b]$ has the same density $\frac{1}{b-a}$, while in the discrete case each of the n outcomes has probability $\frac{1}{n}$.

The Four Moments of Distribution Shape

When we refer to the **shape** of a distribution, we are really describing four fundamental numerical descriptors known as the **statistical moments**.

PDFs for the Four Moments



The Four Moments

They are called moments because each one is built from a power of the variable, just like physical moments in mechanics that describe balance around a pivot.

Moment	Uses (Power of X)	What It Weights	What It Reveals
1st moment (Mean)	X^1	Values proportionally	Center **location**
2nd moment (Variance)	X^2	Larger deviations	**Spread** around the mean
3rd moment (Skewness)	X^3	Positive vs. negative deviations	**Asymmetry**
4th moment (Kurtosis)	X^4	Extreme values	Influence of the **tails** and outliers

Mean (μ) and Standard Deviation (σ)

Among the four moments, the **mean (μ)** and **standard deviation (σ)** are especially important, since they define a distribution's central location and overall spread; σ is simply the square root of the **variance**, making it the more intuitive measure of how widely values are dispersed.

Suppose we have a range of individual values x_i, then the mean and standard deviation of these values are defined as:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

N is the number of values,
x_i are the individual values,
μ is their mean,
σ is their standard deviation.

Note: Here we divide by N. In some contexts $N-1$ appears instead; this distinction depends on whether the values represent an entire population or a sample drawn from it. We will return to this idea later when introducing samples and populations.

Histograms, KDEs and CDFs

Before diving into statistical concepts for specific distributions, we introduce three essential visualization tools:

- **Histograms** show empirical frequency bins.

- **Kernel Density Estimate (KDE)** smooths the histogram into a continuous curve, giving an empirical approximation of the PDFs, hence they are also referred to as **Empirical PDFs**.
- **Cumulative Distribution Function (CDF)** for both PDFs and PMFs increases from 0 at the lower end of the support to 1 at the upper end, **accumulating probability** as the variable increases. It is obtained by integrating the PDF for continuous distributions or by cumulatively summing the PMF for discrete distributions. The CDF is especially useful for comparing tail behavior and understanding how probability accumulates across the range of all possible values.

These tools form the backbone of distribution diagnostics.

Normal Distribution

The **Normal distribution** describes continuous data that cluster symmetrically around a central mean and form a smooth bell-shaped curve. It is defined by two parameters, the mean **μ** and the variance **σ²**, which determine its **location** and **spread**. It is often referred to as the **Gaussian distribution**, named after Carl Friedrich Gauss, the "Prince of Mathematicians," who formalized its properties and helped establish its central role in probability and statistics. Many natural and human-driven phenomena — such as measurement noise, biological traits, and aggregated random processes — tend to follow a Normal distribution, which is why it serves as such a useful reference point and remains one of the most widely used models in statistics.

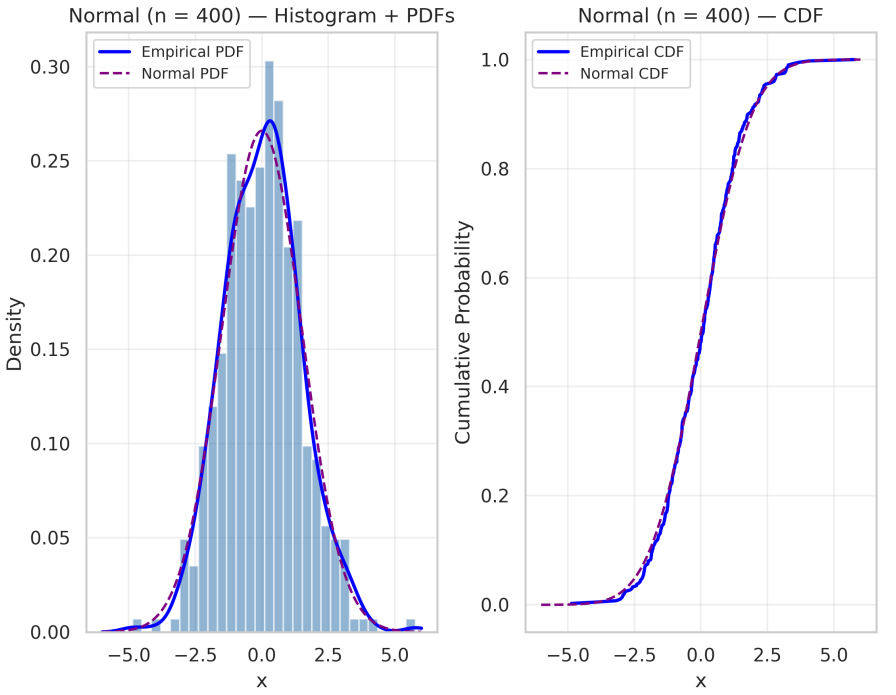
The Normal distribution is **symmetric**, **unimodal** (single peak), and **light-tailed** (few extreme values). Together, these properties make it ideal for illustrating how shape changes when parameters vary.

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

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

$\mu \in \mathbb{R}$ is the mean, $\quad \sigma^2 > 0$ is the variance

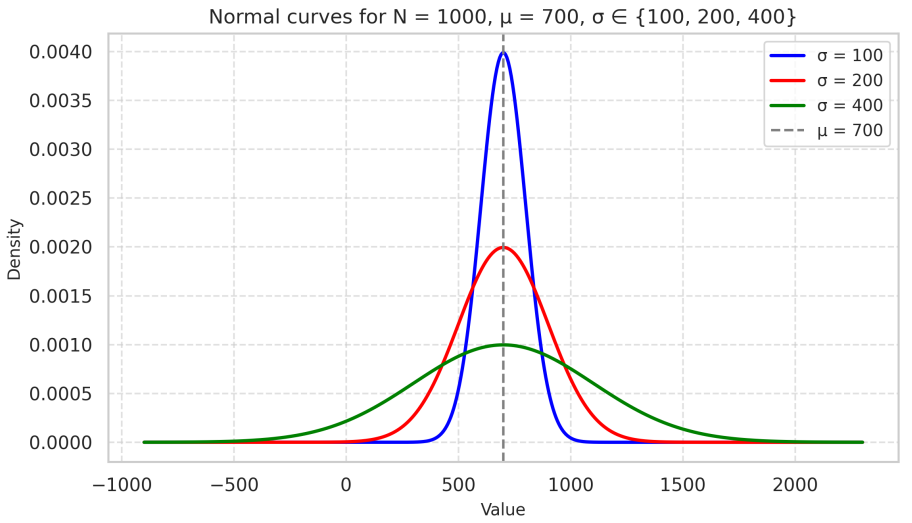
Normal Distribution for Two Simulated Sample Sizes



- The normal distribution's bell shape is evident in the histogram and KDE, while the CDF shows how probability accumulates symmetrically around the mean.
- As sample size increases, the empirical curves become more faithful to the true distribution. Larger samples make the empirical PDF smoother and the empirical CDF closer to the theoretical curve because more data reduces noise, jaggedness and sampling variability.

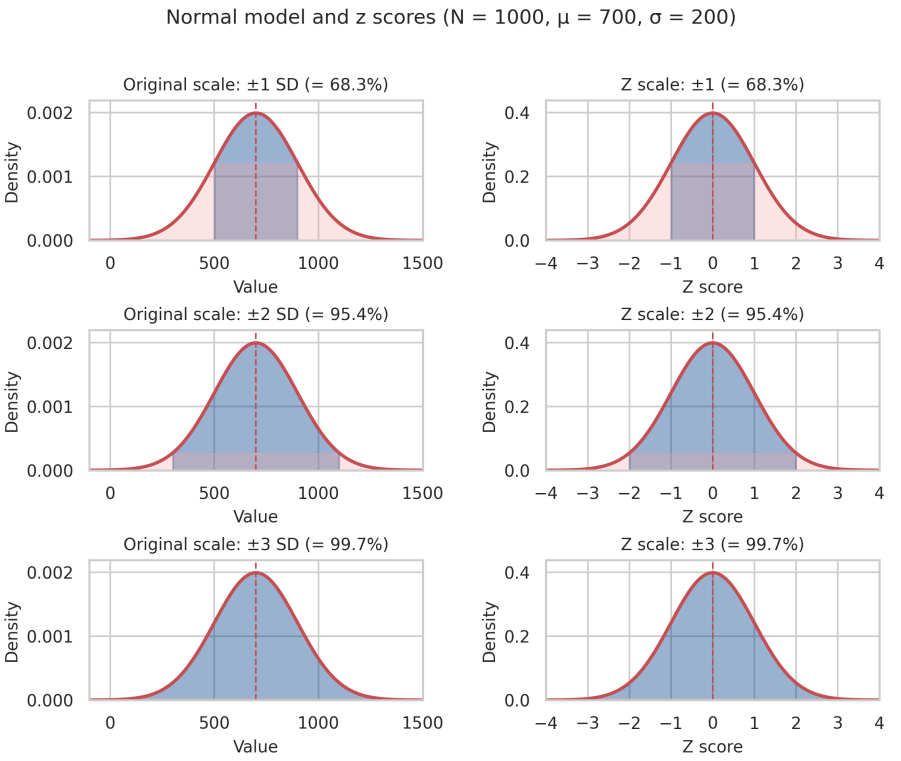
Standard Deviation and z Scores

The **standard deviation** describes how spread out the values are around the mean. On a Normal curve, it determines the width of the bell: a small standard deviation produces a tall, narrow curve with values tightly clustered near the mean, while a large standard deviation produces a wider, flatter curve with values more dispersed.



It also marks predictable probability regions in the Normal distribution: about 68% of values lie within one standard deviation of the mean, about 95% within two, and about 99.7% within three. These percentages apply only when the data are approximately Normal.

The standard deviation sets the scale for standardized scores, also called **z scores**. A z score indicates how many standard deviations a value is from the mean. Any dataset can be converted into standardized scores, regardless of the original units or the shape of the distribution.



The probability meaning of z scores, such as locating a value on a Normal curve or using the above 68–95–99.7 rule, applies only when the data follow a distribution that is close to Normal. When the data are not Normal, standardized scores still describe relative position within the dataset, but they no longer match the probabilities implied by the Normal curve, because those probabilities depend on the data having a shape that is close to Normal.

Gaussian Mixture Model (GMM)

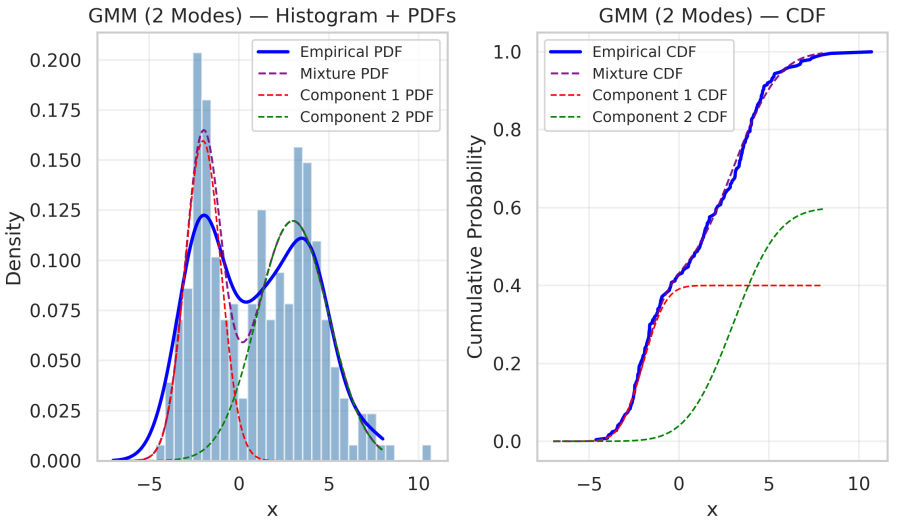
Real data often contains **multiple subpopulations**. A Gaussian Mixture Model (**GMM**) is the simplest way to illustrate this.

A GMM gives us two complementary perspectives on a distribution: a smooth **theoretical density** obtained by blending its Gaussian components, and an **empirical density** that arises from the actual samples it produces.

Two Mode GMM Example

The mixture PDF/CDF describe the true underlying model, while the empirical histogram/CDF show how a finite, noisy dataset approximates that model in practice. This contrast builds intuition for mode drift, cluster overlap and the presence of hidden subpopulations.

GMM PDF/CDF comparison



The chart shows how a two-component GMM compares to the data it generates: the left panel contrasts the empirical histogram/KDE with the model's mixture and component PDFs, while the right panel shows how their CDFs accumulate probability differently.

Continue Reading

References

This book is based primarily on the author's original work, experience, and experimentation. Additional inspiration came from standard texts in statistics, probability, and machine learning. Portions of the writing and code development were supported by AI tools which were used to validate ideas, refine structure, and check content flow. All final explanations, examples, and implementations were authored and reviewed by the writer. For readers interested in deeper study, a curated list of recommended resources is provided below.

Selection of Classic Texts

- **Keynes — *A Treatise on Probability* (1921)**
  A seminal text introducing the logical interpretation of probability as a rational relationship between propositions.
- **Fisher — *Statistical Methods for Research Workers* (1925)**
  A transformative work that popularized the p-value and the frequentist framework for scientific research.
- **von Mises — *Probability, Statistics and Truth* (1928)**
  A foundational work establishing the frequentist interpretation of probability through the concept of collectives.
- **Kolmogorov — *Foundations of the Theory of Probability* (1933)**
  The definitive work that established the modern axiomatic foundation of probability theory, unifying its mathematical structure.

Selection of Modern Texts

- **Efron & Tibshirani — *An Introduction to the Bootstrap* (1993)**
  The classic reference on resampling and bootstrap methods.
- **Hastie, Tibshirani & Friedman — *The Elements of Statistical Learning* (2001; 2nd ed. 2009)**
  A foundational reference for statistical learning theory.
- **Bishop — *Pattern Recognition and Machine Learning* (2006)**
  A standard graduate-level text on probabilistic machine learning.
- **Murphy — *Machine Learning: A Probabilistic Perspective* (2012)**
  A modern, comprehensive treatment of probabilistic ML.
- **Goodfellow, Bengio & Courville — *Deep Learning* (2016)**
  The canonical text on deep learning.

Appendix A: Probability Laws

This concise refresher provides the classical probability laws that serve as background knowledge.

1. Set Notation

- Ω — sample space
- A, B — events
- A^c — complement of A
- $A \cup B$ — union (A or B or both)
- $A \cap B$ — intersection (A and B)
- $\emptyset$ — empty set

Key identity:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

2. Probability Axioms

1. **Non-negativity:**

$$P(A) \geq 0$$

2. **Normalization:**

$$P(\Omega) = 1$$

3. **Additivity (disjoint events):**
   If $A \cap B = \emptyset$, then

$$P(A \cup B) = P(A) + P(B)$$

Useful consequence:

$$P(A^c) = 1 - P(A)$$

3. Conditional Probability

Definition:

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

- $P(A \mid B)$ — probability that A occurs given B
- $P(A \cap B)$ — probability that A and B occur together
- $P(B)$ — probability that B occurs

Rearranged:

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

4. Independence

Events A and B are independent if:

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

- $P(A \cap B)$ — probability A and B occur together
- $P(A)$ — probability A occurs
- $P(B)$ — probability B occurs

Equivalent forms:

$$P(A \mid B) = P(A)$$

$$P(B \mid A) = P(B)$$

5. Law of Total Probability

If $\{A_1, A_2, \ldots, A_k\}$ is a **partition** of the sample space (mutually exclusive and exhaustive), then for any event B:

$$P(B) = \sum_{i=1}^{k} P(B \mid A_i)\, P(A_i)$$

Interpretation of each element:

- A_i — a possible case or scenario
- $P(A_i)$ — probability that case A_i occurs
- $P(B \mid A_i)$ — probability of B assuming case A_i
- $P(B \mid A_i)P(A_i)$ — contribution of case A_i to B
- $P(B)$ — evidence or marginal likelihood: overall probability of observing B

6. Bayes' Rule

Definition:

$$P(A_i \mid B) = \frac{P(B \mid A_i)\, P(A_i)}{P(B)}$$

where:

$$P(B) = \sum_{j} P(B \mid A_j)\, P(A_j)$$

Interpretation of each element:

- A_i — a specific hypothesis or class
- $P(A_i)$ — **prior**: belief in A_i before observing B
- $P(B \mid A_i)$ — **likelihood**: probability of observing B if A_i is true
- $P(B \mid A_i)P(A_i)$ — **joint contribution** of hypothesis A_i and evidence B

- $P(B)$ — **evidence** or **marginal likelihood**: overall probability of observing B
- $P(A_i \mid B)$ — **posterior**: updated belief in A_i after observing B

7. Counting Rules

Multiplication Rule

$$\text{Total outcomes} = n_1 n_2 \cdots n_k$$

- $n_1, n_2, \ldots, n_k$ — number of choices at each step
- **Total outcomes** — number of possible combined outcomes

Permutations (order matters)

$$P(n, k) = \frac{n!}{(n-k)!}$$

- n — total number of items
- k — number of items selected and arranged
- $P(n, k)$ — number of ordered arrangements of k items from n

Combinations (order does not matter)

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- n — total number of items
- k — number of items selected
- $\binom{n}{k}$ — number of distinct groups of size k (ignoring order)

Appendix B: Probability Distributions

Group 1: Basic Distributions

Distribution	PDF / PMF	Mean	Variance
Uniform (Continuous) $X \sim \text{Unif}(a, b)$	$a < bx \in [a, b] f(x) = \frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Uniform (Discrete) $X \sim \text{Unif}\{1, \ldots, n\}$	$n \in \mathbb{N} k \in \{1, \ldots, n\} P(X = k) = \frac{1}{n}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$
Normal $X \sim \mathcal{N}(\mu, \sigma^2)$	$\sigma > 0 x \in \mathbb{R} f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$	μ	σ^2
Bernoulli $X \sim \text{Bern}(p)$	$0 \leq p \leq 1 x \in \{0, 1\} P(X = 1) = p,\ P(X = 0) = 1 - p$	p	$p(1-p)$
Binomial $X \sim \text{Bin}(n, p)$	$n \in \mathbb{N},\ 0 \leq p \leq 1 k \in \{0, \ldots, n\} P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$	np	$np(1-p)$
Geometric $X \sim \text{Geom}(p)$	$0 < p \leq 1 k \in \{1, 2, \ldots\} P(X = k) = p(1-p)^{k-1}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson $X \sim \text{Pois}(\lambda)$	$\lambda > 0 k \in \{0, 1, \ldots\} P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$	λ	λ

Group 2: Exponential Family & Related Discrete Models

Distribution	PDF / PMF	Mean	Variance
Exponential$X \sim \text{Exp}(\lambda)$	$\lambda > 0 x \geq 0 f(x) = \lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma$X \sim \Gamma(k, \theta)$	$k, \theta > 0 x \geq 0 f(x) = \frac{x^{k-1} e^{-x/\theta}}{\Gamma(k) \theta^k}$	$k\theta$	$k\theta^2$
Beta$X \sim \text{Beta}(\alpha, \beta)$	$\alpha, \beta > 0 x \in [0, 1] f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Negative Binomial$X \sim \text{NegBin}(r, p)$	$r > 0,\ 0 < p < 1 k \in \{0, 1, ...\} P(X = k) = \binom{k+r-1}{k} p^r (1-p)^k$	$\frac{r(1-p)}{p}$	$\frac{r(1-p)}{p^2}$
Hypergeometric$X \sim \text{Hypergeom}(N, K, n)$	$N, K, n \in \mathbb{N} k \in \{0, \dots, n\} P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$	$n\frac{K}{N}$	$n\frac{K}{N}\left(1 - \frac{K}{N}\right)\frac{N-n}{N-1}$

Group 3: Shape-Flexible Continuous Distributions

Distribution	PDF / PMF	Mean	Variance
Rayleigh$X \sim \text{Rayleigh}(\sigma)$	$\sigma > 0 x \geq 0 f(x) = \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)}$	$\sigma\sqrt{\frac{\pi}{2}}$	$\frac{4-\pi}{2}\sigma^2$
Weibull$X \sim \text{Weibull}(k, \lambda)$	$k, \lambda > 0 x \geq 0 f(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \exp\left[-\left(\frac{x}{\lambda}\right)^k\right]$	$\lambda\Gamma(1 + 1/k)$	$\lambda^2\Gamma(1 + 2/k) - \lambda^2\Gamma(1 + 1/k)^2$
Lognormal$X \sim \text{Lognormal}(\mu, \sigma^2)$	$\sigma > 0 x > 0 f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-(\ln x - \mu)^2/(2\sigma^2)}$	$e^{\mu+\sigma^2/2}$	$(e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$

Distribution	PDF / PMF	Mean	Variance
Shifted Lognormal $X = s + Y,\ Y \sim \text{Lognormal}(\mu, \sigma^2)$	$x > s f(x) = \frac{1}{(x-s)\sigma\sqrt{2\pi}} e^{-(\ln(x-s)-\mu)^2/(2\sigma^2)}$	$s + e^{\mu+\sigma^2/2}$	$(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$

Group 4: Heavy-Tailed & Power-Law Distributions

Distribution	PDF / PMF	Mean	Variance
Student's t $X \sim t_\nu$	$\nu > 0 x \in \mathbb{R} f(x) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}$	0 (if $\nu > 1$)	$\frac{\nu}{\nu-2}$ (if $\nu > 2$)
Shifted Student's t $X = \mu + \sigma T,\ T \sim t_\nu$	$x \in \mathbb{R} f(x) = \frac{1}{\sigma} f_T\left(\frac{x-\mu}{\sigma}\right)$	μ	$\sigma^2 \frac{\nu}{\nu-2}$ (if $\nu > 2$)
Cauchy $X \sim \text{Cauchy}(x_0, \gamma)$	$\gamma > 0 x \in \mathbb{R} f(x) = \frac{1}{\pi\gamma\left[1+\left(\frac{x-x_0}{\gamma}\right)^2\right]}$	undefined	undefined
Pareto $X \sim \text{Pareto}(x_m, \alpha)$	$x_m > 0,\ \alpha > 0 x \geq x_m f(x) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}}$	$\frac{\alpha x_m}{\alpha-1}$ (if $\alpha > 1$)	$\frac{\alpha x_m^2}{(\alpha-1)^2(\alpha-2)}$ (if $\alpha > 2$)

Group 5: Inference & Regression Distributions

Distribution	PDF / PMF	Mean	Variance
Chi-Square $X \sim \chi^2_k$	$k > 0 x \geq 0 f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$	k	$2k$
F-distribution $X \sim F_{d_1,d_2}$	$d_1, d_2 > 0 x > 0 f(x) = \frac{\sqrt{\frac{d_1 x^{d_1-1} d_2^{d_2}}{(d_1 x + d_2)^{d_1+d_2}}}}{x B(d_1/2, d_2/2)}$	$\frac{d_2}{d_2-2}$ (if $d_2 > 2$)	$\frac{2d_2^2(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$ (if $d_2 > 4$)

Distribution	PDF / PMF	Mean	Variance
Laplace (Double Exponential) $X \sim \text{Laplace}(\mu, b)$	$b > 0 x \in \mathbb{R} f(x) = \frac{1}{2b} e^{-\lvert x-\mu \rvert / b}$	μ	$2b^2$
Logistic $X \sim \text{Logistic}(\mu, s)$	$s > 0 x \in \mathbb{R} f(x) = \frac{e^{-(x-\mu)/s}}{s(1+e^{-(x-\mu)/s})^2}$	μ	$\frac{\pi^2 s^2}{3}$

Further Resources at qikly.com

Further resources are available on the author's website and GitHub pages (https://github.com/gal-a), including examples, visualizations, and all the code used to generate the figures in this chapter.

About the Author

Gal Arav is a data scientist with a wide-ranging career spanning industry, research and entrepreneurship. He has worked on data-intensive projects at NASA, Google, Verizon, AT&T, and General Motors. Most recently, he managed one of GM's EV battery laboratories to prevent thermal runaway and fire hazards using machine learning algorithms and previously led data science work for autonomous vehicle triage and simulation. His contributions at GM earned him the company's Critical Technical Talent Award.

Earlier in his career he founded an internet-based market research company that was featured in *Barron's* and *Bloomberg Businessweek*, and later went on to explore fintech, working as a quantitative researcher in global currency markets. His work with NASA included developing and installing eye-tracking systems at Langley Research Center. He also collaborated on fMRI medical research at leading hospitals and with Cornell and Duke Universities as part of his work at Applied Science Laboratories, a pioneering eye-tracking company founded by two MIT professors. Another highlight was his onsite work at AT&T's Kansas headend facility, where he integrated award winning video processing algorithms for cable broadcasting services he developed.

Gal holds a master's degree in applied mathematics from Tel Aviv University, specializing in Operations Research and Decision Theory. He is deeply curious about the rapid evolution of machine learning and artificial intelligence and has a keen interest in how ideas and actions across history have shaped the course of human progress. Outside of work he enjoys tennis, swimming, climbing and hiking with his kids and dog.

Applied Statistics for Data Science is a practical guide to the statistical foundations every analyst and researcher needs to work confidently with real-world data. Instead of overwhelming readers with theory, this book builds understanding through visual intuition, simulation and accompanying hands-on Python notebooks.

What You Will Learn

- Build a strong foundation in probability models, distribution families, and statistical intuition.
- Understand sampling, estimators, and the core ideas behind uncertainty and variability.
- Perform hypothesis tests, group comparisons, and regression diagnostics to support sound statistical reasoning.
- Design experiments, sampling strategies, and resampling-based inference workflows.
- Detect and monitor data drift using both parametric and non-parametric methods.
- Analyze survival curves, reliability patterns and time-to-event behavior in dynamic systems.

About the author

Gal Arav is a data scientist whose career spans industry, research, and entrepreneurship. He has led machine-learning and analytics projects at NASA, Google, Verizon, AT&T, and General Motors. Gal previously founded an internet-based market research company featured in Barron's and Bloomberg Businessweek and worked as a quantitative researcher in global currency markets.

Website: www.qikly.com