Chapters 1–2: Probability and Random Variables

Donghyun Ko

May 27, 2026

These notes provide a detailed and graduate-level explanations of Chapters 1 and 2 of *Mathematical Statistics and Data Analysis* (John A. Rice, 3rd ed.). The focus is on building intuition, formal definitions, and computational tools for probability models and random variables.

Contents

1 Probability **2**
- 1.1 Why probability? . . . 2
- 1.2 Sample space and events . . . 2
- 1.3 Axioms of probability . . . 3
- 1.4 Basic consequences . . . 3
- 1.5 Finite sample spaces and counting . . . 4
- 1.6 Conditional probability . . . 4
- 1.7 Multiplication rule . . . 4
- 1.8 Law of total probability . . . 4
- 1.9 Bayes' rule . . . 5
- 1.10 Independence . . . 5

2 Random Variables **5**
- 2.1 The cumulative distribution function . . . 5
- 2.2 Discrete random variables . . . 6
- 2.3 Continuous random variables . . . 7
- 2.4 Relationship between CDF and PDF . . . 9
- 2.5 Expectation . . . 9
- 2.6 Variance . . . 10
- 2.7 Linear transformations . . . 10
- 2.8 Transformations of continuous random variables . . . 10

1 Probability

1.1 Why probability?

Probability theory is used to model situations in which outcomes are uncertain. Even when an experiment is repeated under identical conditions, the result may vary. Probability does not remove randomness; instead, it provides a mathematical framework for quantifying uncertainty and making consistent predictions. A probability model consists of:

- a description of all possible outcomes, and
- a rule that assigns probabilities to events.

1.2 Sample space and events

A *probability model* starts by clearly describing *what outcomes are possible.*

- **Sample space** Ω: the set of all possible outcomes of the experiment.
  - Example 1. two coin tosses: $\Omega = \{HH, HT, TH, TT\}$.
  - Example 2. die roll: $\Omega = \{1, 2, 3, 4, 5, 6\}$.
  - Example 3. waiting time: $\Omega = [0, \infty)$, as a waiting time can be nonnegative real number.
- **Outcome** ω: a single element of Ω (one realized result). For instance, in two coin tosses, $\omega = HT$ is one outcome.
- **Event** A: a subset of Ω. An event is a *statement* about the outcome that can be true or false. The event occurs if the realized outcome ω lies in the set A (i.e., $\omega \in A$).
  - Example. die roll: $A = \{2, 4, 6\}$ means "the die shows an even number." If the outcome is 4, then A occurs; if the outcome is 3, then A does not occur.
- **Empty event** $\emptyset$: the event with no outcomes in it. It represents an impossible event.
- **Certain event** Ω: the event that always occurs (since the outcome must be in Ω).

Because events are sets, we use set operations to build new events from old ones. These operations match familiar logical words like "not," "and," and "or."

- **Complement** A^c:
$$A^c = \{\omega \in \Omega : \omega \notin A\}.$$
This is the event "A does not occur."
Example: if $A = \{2, 4, 6\}$ for a die, then $A^c = \{1, 3, 5\}$ ("odd outcome").
- **Intersection** $A \cap B$:
$$A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}.$$
This is the event "both A and B occur."
Example: if $A = \{\text{even}\}$ and $B = \{4, 5, 6\}$ ("at least 4"), then $A \cap B = \{4, 6\}$.

- **Union** $A \cup B$:
$$A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}.$$
This is the event "at least one of A or B occurs" (possibly both).
Example: if $A = \{\text{even}\}$ and $B = \{\text{at least } 4\}$, then $A \cup B = \{2, 4, 5, 6\}$.

- **Disjoint (mutually exclusive) events**: Events A and B are disjoint if $A \cap B = \emptyset$. This means they cannot happen at the same time.
Example: on a die, $A = \{1, 2, 3\}$ and $B = \{4, 5, 6\}$ are disjoint.

As a key idea to remember, an event is a *set of outcomes*, and the probability assigns a number to the events (sets), not to individual outcomes in isolation.

1.3 Axioms of probability

A probability measure P assigns a numerical value to each event $A \subset \Omega$. These assignments are governed by a small set of axioms that ensure internal consistency and agreement with intuitive notions of chance.

- **Normalization:** $P(\Omega) = 1$.
The probability of the entire sample space is one, since some outcome in Ω must occur whenever the experiment is performed.

- **Nonnegativity:** if $A \subset \Omega$, then $P(A) \geq 0$.
Every event is assigned a nonnegative probability. In particular, this axiom rules out negative probabilities and ensures that probabilities can be interpreted as measures of likelihood. As a consequence, impossible events will later be shown to have probability zero.

- **Additivity:** if $A_1, A_2, \ldots$ are disjoint events (i.e., $A_i \cap A_j = \emptyset$ for $i \neq j$), then
$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i).$$
In particular, if two events A and B are disjoint ($A \cap B = \emptyset$), then
$$P(A \cup B) = P(A) + P(B).$$
This expresses the idea that when events cannot occur together, the probability that one of them occurs is obtained by simple addition.

1.4 Basic consequences

Although the axioms are simple, they imply many useful rules that are applied repeatedly in probability calculations. The following properties all follow directly from the axioms above.

- **Complement rule:** $P(A^c) = 1 - P(A)$.
If an event does not occur, its complement must occur, and together they exhaust all outcomes.

- **Empty event:** $P(\emptyset) = 0$.
An impossible event has zero probability.

- **Monotonicity:** if $A \subset B$, then $P(A) \leq P(B)$.
  An event contained inside another cannot be more likely than the larger event.

- **Inclusion–exclusion:**

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

  When two events overlap, the probability of their union is obtained by adding their probabilities and subtracting the overlap to avoid double counting.

1.5 Finite sample spaces and counting

When the sample space Ω is finite and all outcomes are equally likely, probabilities can be computed by simple counting arguments. In this case,

$$P(A) = \frac{|A|}{|\Omega|},$$

where $|A|$ denotes the number of outcomes in the event A. As a result, many probability problems reduce to counting the number of favorable outcomes and the total number of possible outcomes, often using the multiplication rule, permutations, and combinations.

1.6 Conditional probability

In many situations, additional information is available before an event of interest is evaluated. Conditional probability formalizes how probabilities should be updated once we know that another event has occurred. For events A and B with $P(B) > 0$, the conditional probability of A given B is defined by

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

This definition restricts the sample space to the subset B and rescales probabilities so that $P(B) = 1$ within the conditioned experiment.

1.7 Multiplication rule

Rearranging the definition of conditional probability yields the multiplication rule:

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

This identity provides a practical way to compute joint probabilities by considering events in sequence: first B occurs, and then A occurs given B. The rule extends naturally to more than two events by repeated conditioning.

1.8 Law of total probability

Often, an event A can occur through several mutually exclusive scenarios. Let $B_1, \ldots, B_n$ be a partition of the sample space Ω with $P(B_i) > 0$. Then the probability of A can be decomposed as

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i)\, P(B_i).$$

This formula expresses $P(A)$ as a weighted average of conditional probabilities, with weights given by the probabilities of the scenarios B_i.

1.9 Bayes' rule

Bayes' rule follows by combining the multiplication rule with the law of total probability. For a partition $\{B_1, \ldots, B_n\}$ of Ω,

$$P(B_j \mid A) = \frac{P(A \mid B_j)\, P(B_j)}{\sum_{i=1}^{n} P(A \mid B_i)\, P(B_i)}.$$

This formula reverses the direction of conditioning: it allows probabilities of the underlying causes B_j to be updated after observing evidence A. Bayes' rule plays a central role in statistical inference and decision making.

1.10 Independence

Two events A and B are said to be independent if the occurrence of one does not affect the probability of the other. Formally, A and B are independent if

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

When $P(B) > 0$, this condition is equivalent to

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

which shows that conditioning on B provides no additional information about A.

2 Random Variables

In Chapter 1, probability was assigned directly to events, which are sets of outcomes. In many applications, however, we are interested not in the outcomes themselves but in numerical quantities derived from them. A **random variable** X is a real-valued function defined on the sample space Ω, mapping each outcome ω to a number $X(\omega)$. Random variables allow us to describe randomness using numbers and to summarize uncertainty through probability distributions.

2.1 The cumulative distribution function

Every random variable X, whether discrete or continuous, is completely characterized by its **cumulative distribution function (CDF)**, defined as

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

The CDF gives the probability that the random variable takes a value less than or equal to x and provides a unified description of distributions. The CDF satisfies the following basic properties:

- F is non-decreasing, since probabilities of larger sets cannot decrease;
- $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$;
- for any $a < b$, $P(a < X \leq b) = F(b) - F(a)$.

2.2 Discrete random variables

A random variable is called **discrete** if it takes values in a finite or countably infinite set. In this case, probability is assigned directly to individual numerical values. Discrete random variables typically arise when we *count* outcomes.

Example. Consider tossing a fair coin twice and let X denote the number of heads observed. The sample space is $\Omega = \{HH, HT, TH, TT\}$, and the possible values of X are

$$\{0, 1, 2\}.$$

Here, X is a discrete random variable because it takes only finitely many values. We can compute probabilities directly, for example, $P(X = 0) = P(TT) = 1/4$ and $P(X = 2) = P(HH) = 1/4$.

More generally, discrete random variables are used to model quantities such as the number of successes in repeated trials, the number of arrivals in a time interval, or the number of failures before a success occurs.

Probability mass function (PMF) The **probability mass function** (PMF) of a discrete random variable X is defined by

$$p(x) = P(X = x),$$

for each value x in the support of X. The PMF must satisfy the normalization condition

$$\sum_{x \in \mathrm{Supp}(X)} p(x) = 1,$$

where the *support* of X is

$$\mathrm{Supp}(X) = \{x : P(X = x) > 0\}.$$

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

Key properties of a PMF:

- $p(x) \geq 0$ for all x,
- $\sum_x p(x) = 1$,
- $P(X \in A) = \sum_{x \in A} p(x)$ for any subset A of the support.

Common discrete distributions Many important discrete random variables have standard forms that arise naturally from repeated experiments or counting processes.

- **Bernoulli**(p): models a single trial with success probability p.

$$P(X = x) = p^x (1 - p)^{1-x}, \quad x = 0, 1.$$

- **Binomial**(n, p): counts the number of successes in n independent Bernoulli trials.

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, 1, \ldots, n.$$

  If $Z_1, \ldots, Z_n$ are i.i.d. Bernoulli(p), then

$$X = \sum_{i=1}^{n} Z_i \sim \mathrm{Binomial}(n, p).$$

- **Geometric**(p): counts the number of trials until the first success.

$$P(X = x) = p(1-p)^{x-1}, \quad x = 1, 2, \ldots$$

  This distribution has infinite support and satisfies $\sum_{x=1}^{\infty} P(X = x) = 1$.

- **Negative Binomial**(r, p): counts the number of trials needed to achieve r successes.

$$P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \quad x = r, r+1, \ldots$$

- **Poisson**(λ): models the number of events occurring in a fixed interval, when events occur independently at rate λ.

$$P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

  The Poisson distribution can be derived as a limit of the Binomial(n, p) distribution as $n \to \infty$ and $p \to 0$ with $np \to \lambda$.

These distributions form the basic toolkit for modeling discrete data. Which model is appropriate depends on how the random quantity is generated. The broad examples are single trials (Bernoulli), repeated trials (Binomial), waiting times for success (Geometric or Negative Binomial), or rare events over time or space (Poisson).

2.3 Continuous random variables

A random variable is called **continuous** if it can take values on a continuum, typically an interval or union of intervals on the real line. In contrast to discrete random variables, probabilities are not assigned to individual points but to *intervals* of values.

Example. Let X denote the waiting time (in hours) until the next customer arrives at a service desk. Possible values of X form the interval $[0, \infty)$, and it is natural to model X as a continuous random variable rather than assigning probabilities to exact times such as $X = 2.317$ hours.

The distribution of a continuous random variable X is described by a **probability density function (PDF)** $f(x)$. The PDF is a nonnegative function that satisfies:

- $f(x) \geq 0$ for all x,
- $\displaystyle\int_{-\infty}^{\infty} f(x)\, dx = 1$,
- for any $a < b$, $P(a < X < b) = \int_a^b f(x)\, dx$.

The PDF itself is *not* a probability. Instead, probabilities are obtained by integrating $f(x)$ over intervals. Regions where $f(x)$ is large indicate values of X that are more likely to occur. A fundamental property of continuous random variables is that

$$P(X = x) = 0 \quad \text{for all } x.$$

This does not mean that X never equals a specific value; rather, the probability of any single point is negligible compared to the continuum of possible values. The **support** of X is defined as

$$\mathrm{Supp}(X) = \{x : f(x) \neq 0\},$$

and determines the region over which probabilities and expectations are computed. Correctly identifying the support is essential for setting up integrals.

Common continuous distributions. Many practical models are built from a small number of standard continuous distributions. We summarize the most important ones in the below.

- **Uniform**(a, b). Models a random choice over a finite interval $[a, b]$ with no preference for any subinterval of equal length.

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \leq x \leq b, \\ 0, & \text{otherwise.} \end{cases}$$

  The cumulative distribution function (CDF) is

$$F(x) = \begin{cases} 0, & x < a, \\ \dfrac{x-a}{b-a}, & a \leq x \leq b, \\ 1, & x > b. \end{cases}$$

  All intervals of equal length have equal probability. *Example*: choosing a point at random on a stick of length $b - a$.

- **Exponential**(λ). Models the waiting time until the first occurrence of an event, with rate $\lambda > 0$.

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

  Its CDF is

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

  The exponential distribution satisfies the *memoryless property*:

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

  meaning that the remaining waiting time does not depend on how long one has already waited. *Example*: time until the next phone call arrives at a call center.

- **Gamma**(α, λ). Models the waiting time until α events occur, where $\alpha > 0$ is the shape parameter and $\lambda > 0$ is the rate.

$$f(x) = \begin{cases} \dfrac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

  The CDF does not have a simple closed form for general α and is usually computed numerically. When $\alpha = 1$, the Gamma distribution reduces to the Exponential(λ) distribution. *Example*: time required to receive α customer arrivals.

- **Normal**(μ, σ^2). Models natural variability and measurement error in physical and social phenomena.
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$
The CDF is
$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt,$$
which has no closed form and is evaluated using tables or software. If $Z \sim N(0,1)$, then $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$. *Example*: measurement noise around a true value μ.

- **Beta**(α, β). Defined on the interval $[0,1]$, commonly used to model proportions and probabilities.
$$f(x) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & 0 \le x \le 1, \\ 0, & \text{otherwise.} \end{cases}$$
The shape of the distribution is controlled by α and β, allowing for a wide range of behaviors (uniform, U-shaped, skewed). The CDF has no simple closed form in general. *Example*: modeling an unknown probability of success in Bayesian inference.

Important connection. Exponential, Gamma, and Poisson distributions are closely related. If interarrival times are i.i.d. Exp(λ), then:

- the number of arrivals in one unit of time follows Poisson(λ),
- the number of arrivals by time t follows Poisson(λt),
- the time until k arrivals follows Gamma(k, λ).

This relationship provides a unified framework for modeling counts and waiting times.

2.4 Relationship between CDF and PDF

If a random variable X has a probability density function f, its CDF can be written as
$$F(x) = \int_{-\infty}^{x} f(t)\, dt.$$
When F is differentiable, the density is obtained by differentiation:
$$f(x) = F'(x).$$
Thus, the CDF represents accumulated probability, while the PDF describes how probability is locally distributed.

2.5 Expectation

The **expectation** of a random variable describes its average or central value. It is defined as

- Discrete case:
$$\mathbb{E}[X] = \sum_{x} x\, p(x),$$
- Continuous case:
$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f(x)\, dx.$$

Expectation is a linear operator and plays a fundamental role in statistical analysis.

2.6 Variance

The **variance** of a random variable measures the spread of its distribution around the mean. It is defined by

$$\mathrm{Var}(X) = \mathbb{E}[(X-\mu)^2], \qquad \mu = \mathbb{E}[X].$$

An equivalent and often convenient formula is

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

2.7 Linear transformations

If a random variable X is transformed linearly as $Y = aX + b$, then its mean and variance change in a simple way:

$$\mathbb{E}[aX+b] = a\,\mathbb{E}[X] + b, \qquad \mathrm{Var}(aX+b) = a^2\,\mathrm{Var}(X).$$

These properties explain how rescaling and shifting affect distributions.

2.8 Transformations of continuous random variables

Often we are interested not in the original random variable X itself, but in a new random variable defined as a function of X. For example, we may rescale, shift, or otherwise transform measurements. Suppose X is a continuous random variable with density f_X, and let

$$Y = g(X),$$

where g is a monotone and differentiable function. In this case, the distribution of Y can be obtained using the **change-of-variables formula**. If g is strictly monotone and has an inverse function g^{-1}, then the density of Y is

$$f_Y(y) = f_X\big(g^{-1}(y)\big)\left|\frac{d}{dy}g^{-1}(y)\right|.$$

The absolute value of the derivative accounts for the local stretching or compression of the scale induced by the transformation.

Example

Example 1 (Linear transformation). Let $X \sim \mathrm{Unif}(0,1)$ and define $Y = 2X + 1$.

Step 1: Identify the inverse transformation.

$$y = 2x + 1 \quad \Rightarrow \quad x = g^{-1}(y) = \frac{y-1}{2}.$$

Step 2: Compute the derivative of the inverse.

$$\frac{d}{dy}g^{-1}(y) = \frac{1}{2}.$$

Step 3: Apply the change-of-variables formula. Since $f_X(x) = 1$ for $0 \le x \le 1$,

$$f_Y(y) = 1 \cdot \frac{1}{2} = \frac{1}{2}, \quad \text{for } 1 \le y \le 3.$$

Thus $Y \sim \mathrm{Unif}(1,3)$. This example shows how linear transformations shift and rescale distributions.

Example

Example 2 (Nonlinear transformation). Let $X \sim \text{Exp}(1)$ and define $Y = \sqrt{X}$.

Step 1: Find the inverse transformation.

$$y = \sqrt{x} \quad \Rightarrow \quad x = g^{-1}(y) = y^2.$$

Step 2: Compute the derivative of the inverse.

$$\frac{d}{dy} g^{-1}(y) = 2y.$$

Step 3: Apply the formula. Since $f_X(x) = e^{-x}$ for $x \geq 0$,

$$f_Y(y) = e^{-y^2} \cdot 2y, \quad y \geq 0.$$

This density is no longer exponential and illustrates how nonlinear transformations can substantially change the shape of a distribution.

A particularly important special case arises when the transformation is the CDF itself. Let F be the CDF of a continuous random variable X, and define

$$Y = F(X).$$

Then

$$Y \sim \text{Unif}(0, 1).$$

This result is known as the **probability integral transform**. It provides the theoretical foundation for simulation: by generating $U \sim \text{Unif}(0, 1)$ and setting $X = F^{-1}(U)$, one can generate random samples from the distribution with CDF F.