BUSS975 Causal Inference in Financial Research

Ji-Woong Chung
chung_jiwoong@korea.ac.kr
Korea University Business School

Chapter 2

Statistical Foundations Review

Probability theory provides the formal framework for analyzing randomness and uncertainty. In this chapter, we will cover the basic definitions of probability spaces, random variables, distribution functions, common distributions, random vectors, conditional distributions, and independence. Throughout, we emphasize intuition and examples, building on solid theoretical foundations.

2.1 Probability Spaces and Events

Probability theory starts with the idea of an *experiment* with uncertain outcomes. The set of all possible outcomes is called the **sample space**, denoted by Ω. An individual outcome of the experiment is $\omega \in \Omega$, also called a realization. An **event** is any collection (subset) of outcomes, i.e. $E \subseteq \Omega$. We will often be interested in the probability of various events.

Example 2.1. Consider tossing a coin twice. The sample space is $\Omega = \{HH, HT, TH, TT\}$,where each element is a two-letter sequence indicating the result of the first and second toss (H for heads, T for tails). For instance, $\omega = HH$ means the outcome was heads on both tosses. An example of an event is "the first toss is tails," which corresponds to the subset $E = \{TH, TT\} \subseteq \Omega$. Another event could be "exactly one heads in two tosses," which is $E' = \{HT, TH\}$.

Often we are interested in whether or not an outcome ω belongs to a certain event E. For this purpose, it is convenient to use **indicator functions**.

Definition 2.2 (Indicator Function)**.** Let Ω be a sample space and $E \subseteq \Omega$ an event. The *indicator function* of event E, denoted $\mathbf{1}_E(\omega)$, is defined as

$$\mathbf{1}\{\omega \in E\} = \mathbf{1}_E(\omega) = \begin{cases} 1, & \text{if } \omega \in E, \\ 0, & \text{if } \omega \notin E\ , \end{cases}$$

for any outcome $\omega \in \Omega$. In words, $\mathbf{1}_E(\omega)$ equals 1 if the event E occurs (i.e. if the outcome is in E) and 0 otherwise.

Example 2.3. Continuing the coin-toss experiment, let $E_1 = \{TT\}$ be the event that both tosses are tails, and $E_2 = \{TH, TT\}$ be the event that the first toss is tails (as above). Then for outcome $\omega = TT$, we have $\mathbf{1}_{E_1}(TT) = 1$ and $\mathbf{1}_{E_2}(TT) = 1$ (since $TT \in E_1$ and also $TT \in E_2$). For $\omega = TH$, we find $\mathbf{1}_{E_1}(TH) = 0$ (since $TH \notin E_1$) but $\mathbf{1}_{E_2}(TH) = 1$ (since $TH \in E_2$). These indicators correctly reflect which events occur for each outcome.

Indicator functions allow us to answer "yes-or-no" questions about outcomes algebraically. They are especially handy as event descriptions grow more complicated, thanks to a few key properties:

Lemma 2.4 (Properties of Indicator Functions)**.** *For any events $E_1, E_2 \subseteq \Omega$ and any outcome $\omega \in \Omega$, the following properties hold:*

1. $\mathbf{1}_{E_1}(\omega)^k = \mathbf{1}_{E_1}(\omega)$ *for any exponent $k \neq 0$.*
2. $\mathbf{1}_{E_1^c}(\omega) = 1 - \mathbf{1}_{E_1}(\omega)$*, where E_1^c is the complement event (not E_1).*
3. $\mathbf{1}_{E_1 \cap E_2}(\omega) = \mathbf{1}_{E_1}(\omega)\, \mathbf{1}_{E_2}(\omega)$.
4. $\mathbf{1}_{E_1 \cup E_2}(\omega) = \mathbf{1}_{E_1}(\omega) + \mathbf{1}_{E_2}(\omega) - \mathbf{1}_{E_1 \cap E_2}(\omega)$.

2.1.1 Probability Measure

Now we introduce the central object of probability theory: the probability measure. The probability measure assigns a likelihood (a number between 0 and 1) to each event, respecting certain axioms.

Definition 2.5 (Probability Measure)**.** A *probability measure* P on a sample space Ω assigns a probability $P(E)$ to each event $E \subseteq \Omega$. The function $P : 2^\Omega \to [0, 1]$ must satisfy the following axioms:

1. $P(\Omega) = 1$. (The probability that *something* in the sample space occurs is 1.)

2. $P(E) \geq 0$ for all events $E \subseteq \Omega$. (Probabilities are nonnegative.)

3. If E_1 and E_2 are disjoint events ($E_1 \cap E_2 = \emptyset$), then $P(E_1 \cup E_2) = P(E_1) + P(E_2)$. More generally, for any countable collection of pairwise disjoint events $E_1, E_2, E_3, \ldots$,

$$P\Big(\bigcup_i E_i\Big) = \sum_i P(E_i).$$

   In particular, for two disjoint events E_1 and E_2, $P(E_1 \cup E_2) = P(E_1) + P(E_2)$. This property is called *additivity* (or *sigma-additivity* in the infinite case).

A triple $(\Omega, \mathcal{F}, P)$, where $\mathcal{F}$ is a collection of events (typically a σ-algebra) on Ω and P satisfies these properties on $\mathcal{F}$, is called a **probability space**.

In many simple settings, especially with a finite or countable Ω, one can specify P by first assigning probabilities to each elementary outcome $\omega \in \Omega$ such that $\sum_{\omega \in \Omega} P(\omega) = 1$, and then for any event E, $P(E)$ is the sum of $P(\omega)$ for $\omega \in E$. However, in continuous settings, one cannot usually define $P(\omega)$ for each individual outcome in a meaningful way — instead, probabilities are attached directly to events (often via integrals of density functions, as we'll see).

Example 2.6. If we toss a fair coin twice, a natural probability measure is the one that assigns equal probability to each of the four outcomes in $\Omega = \{HH, HT, TH, TT\}$. For example, we set $P(\{\omega\}) = 1/4$ for each $\omega \in \Omega$. Then by additivity, $P(\{HH\}) = 1/4$, $P(\{TT\}) = 1/4$, and $P(\{HT, TH\}) = P(\{HT\}) + P(\{TH\}) = 1/4 + 1/4 = 1/2$. This aligns with our intuition for independent fair coin flips.

2.2 Random Variables and Distribution Functions

In many applications, we are not directly interested in the outcome ω itself, but rather in some numerical quantity determined by the outcome. **Random variables** formalize this notion.

Definition 2.7 (Random Variable)**.** A *random variable* X is a function $X : \Omega \to \mathbf{R}$ that assigns a real number $X(\omega)$ to each outcome $\omega \in \Omega$. In other words, a random variable is a numerical summary of the outcome of an experiment.

Example 2.8. In the coin-toss experiment, define $X(\omega)$ to be the number of heads that occur in the outcome ω. Then X is a random variable taking values in $\{0, 1, 2\}$. For example, if $\omega = TH$ (first toss tails, second toss heads), then $X(\omega) = 1$. If $\omega = TT$, then $X(\omega) = 0$.

Note: We have defined a random variable in a somewhat informal way as a function from outcomes to $\mathbb{R}$. In a more rigorous mathematical treatment, one requires X to be a measurable function with respect to the event σ-algebra, but that level of detail is beyond our scope here. It suffices for our purposes to treat $X(\omega)$ as a real number associated with outcome ω.

Every random variable X induces a probability distribution on the real line, describing how the total probability mass of 1 is distributed over the values that X can take. A key descriptor of this distribution is the cumulative distribution function:

Definition 2.9 (Cumulative Distribution Function)**.** The **cumulative distribution function (CDF)** of a random variable X is the function $F_X : \mathbb{R} \to [0, 1]$ defined by $F_X(x) = P(X \leq x), \forall x \in \mathbb{R}$. For each real number x, $F_X(x)$ gives the probability that the random variable X will take a value less than or equal to x.

Notation: We typically use uppercase letters (like X, Y) for random variables, and lowercase letters (like x, y) for specific values or realizations of those variables. If F_X is the CDF of X, we sometimes denote the distribution by writing $X \sim F_X$. In the special case that X has a known named

distribution (e.g. normal with mean μ and variance σ^2), we use notation like $X \sim N(\mu, \sigma^2)$ to denote its distribution.

The CDF $F_X(x)$ is a non-decreasing function in x that satisfies $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$. It encapsulates all the information about the probability distribution of X. In fact, the CDF uniquely determines the probability law of X, as formalized by the following theorem:

Theorem 2.10 (CDF Characterizes Distribution)**.** *Let X and Y be two random variables with CDFs F_X and F_Y respectively. If $F_X(x) = F_Y(x)$ for all $x \in \mathbb{R}$, then X and Y have the same distribution. In particular, $P(X \in E) = P(Y \in E)$ for every event $E \subseteq \mathbb{R}$ (for every subset of real numbers).*

When X and Y have the same CDF (hence the same distribution), we say they are **identically distributed**. This is sometimes denoted by $X \overset{d}{=} Y$ (read as "X equals Y in distribution"). Identically distributed random variables need not be equal to each other as numbers; they simply behave the same probabilistically.

Example 2.11. Let X be the number of heads in two fair coin tosses, and let Y be the number of tails in two fair coin tosses. Intuitively, X and Y have the same distribution (since in two tosses, the distribution of "number of heads" is the same as the distribution of "number of tails"). We can confirm this by writing out the CDF of X explicitly:

$$F_X(x) = P(X \le x) = \begin{cases} 0, & x < 0, \\ 1/4, & 0 \le x < 1, \\ 3/4, & 1 \le x < 2, \\ 1, & 2 \le x. \end{cases}$$

because X can only take the values 0, 1, or 2, with probabilities $P(X = 0) = 1/4$, $P(X = 1) = 1/2$, $P(X = 2) = 1/4$. Now Y = number of tails $= 2 - X$ in this experiment, so Y can also be $0, 1, 2$ with the same probabilities $1/4, 1/2, 1/4$. Indeed one can check $F_Y(x) = P(Y \le x)$ is exactly the same as $F_X(x)$ above. Thus $F_X(x) = F_Y(x)$ for all x, and by the theorem X and Y are identically distributed ($X \overset{d}{=} Y$). However, clearly $X \neq Y$ for any given outcome (if there are x heads there are $2 - x$ tails, so $Y = 2 - X$). Two random variables can have the same distribution without being equal as random quantities.

2.2.1 Discrete Random Variables

Random variables come in different types. If a random variable can take at most a countable number of distinct values (like 0,1,2 or a finite list), we call it **discrete**. The distribution of a discrete random variable can be described by its probability mass function:

Definition 2.12 (Discrete Random Variable and PMF)**.** A random variable X is *discrete* if it takes values in a countable set $\{x_1, x_2, x_3, \dots\}$ (finite or countably infinite). The **probability mass function (pmf)** of X is the function $f_X : \mathbb{R} \to [0, 1]$ defined by $f_X(x) = P(X = x)$ for each $x \in \mathbb{R}$. (For values of x not in the support of X, $f_X(x) = 0$.) The **support** of X is the set of values that X can actually take with positive probability: $\text{supp}(X) = x \in \mathbb{R} : P(X = x) > 0$.

By the laws of probability, for a discrete X we must have $f_X(x) \geq 0$ for all x and $\sum_{x \in \text{supp}(X)} f_X(x) = 1$ (all the probability mass sums to 1).

Once we have the pmf, the cumulative distribution function can be obtained by summing probabilities. In fact, for any x,

$$F_X(x) = P(X \leq x) = \sum_{x' \in \text{supp}(X)} f_X(x')\, \mathbf{1}\{x' \leq x\},$$

i.e. we sum the probabilities of all support points x' that are $\leq x$.

Example 2.13. Again let X be the number of heads in two fair coin tosses. The support of X is $0, 1, 2$. Its pmf is

$$f_X(x) = \begin{cases} 1/4, & x = 0, \\ 1/2, & x = 1, \\ 1/4, & x = 2, \\ 0, & \text{otherwise.} \end{cases}$$

We can use this pmf to compute the CDF at, say, $x = 1$:

$$F_X(1) = P(X \leq 1) = P(X = 0) + P(X = 1) = f_X(0) + f_X(1) = 1/4 + 1/2 = 3/4,$$

which matches the earlier CDF piecewise definition for $F_X(x)$.

2.2.2 Continuous Random Variables

Another important class of random variables are **continuous** random variables, which have an uncountable range and are described by a density function rather than point probabilities.

Definition 2.14 (Continuous Random Variable and PDF)**.** A random variable X is *continuous* if its distribution can be described by a nonnegative **probability density function (pdf)** $f_X : \mathbb{R} \to [0, \infty)$ such that for any interval $[a, b]$,

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

The pdf must satisfy $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$. Given a pdf, the CDF is obtained by integration:

$$F_X(x) = P(X \leq x) = \int_{\infty}^{x} f_X(t)\,dt,$$

and conversely, if F_X is differentiable, then $f_X(x) = \frac{d}{dx} F_X(x)$.

In simpler terms, for a continuous random variable X, probabilities are given by areas under the density curve. An important consequence is that for a truly continuous distribution, the probability of X taking any exact value is zero. In fact, if X is continuous, for any specific number c we have

$$P(X = c) = \int_c^c f_X(x)\,dx, = 0$$

All probability is in intervals or ranges of values, not at points.

Example 2.15. Consider choosing a number uniformly at random between 0 and 1. This can be modeled by a continuous random variable X with pdf

$$f_X(x) = \begin{cases} 1, & 0 \leq x \leq 1, \\ 0, & \text{otherwise}, \end{cases}$$

Clearly $f_X(x) \geq 0$ everywhere and $\int_{-\infty}^{\infty} f_X(x)\,dx = \int_0^1 1\,dx = 1$. For any subinterval $[a, b] \subseteq [0, 1]$,

$$P(a < X \leq b) = \int_a^b = b - a,$$

which matches our intuition of a uniform pick. The CDF in this case is

$$F_X(x) = \begin{cases} 0, & x < 0, \\ x, & 0 \le x \le 1, \\ 1, & x > 1. \end{cases}$$

since $P(X \le x) = 0$ for $x < 0$, equals x for $0 \le x \le 1$ (because the proportion of the interval $[0, 1]$ up to x is x itself), and is 1 for $x \ge 1$. We denote $X \sim U(0, 1)$, read "X is uniformly distributed on $[0, 1]$." This $U(0, 1)$ distribution is often called the *standard uniform distribution*.

Caveat (Discrete vs. Continuous): It is important not to confuse the roles of pmfs and pdfs. For a discrete random variable, $P(X = x)$ is given directly by the pmf $f_X(x)$. In contrast, for a continuous random variable, $P(X = x) = 0$ for every x, *even though* we might have $f_X(x) > 0$. The number $f_X(x)$ for a continuous distribution does *not* equal $P(X = x)$; rather, $f_X(x)$ is a density height such that probabilities of intervals are integrals of f_X. Also, note that a pdf can sometimes take values larger than 1 (or even be unbounded) without issue, as long as the area under the curve is 1. For example, if $X \sim U(0, 0.5)$ (uniform on $[0, 0.5]$), then $f_X(x) = 2$ for $0 \le x \le 0.5$, which is greater than 1. Similarly, a pdf like $f_X(x) = \frac{1}{2\sqrt{x}}$ for $0 < x < 1$ (and 0 elsewhere) is unbounded as $x \to 0$, yet it is a valid density (you can check $\int_0^1 \frac{1}{2\sqrt{x}} dx = 1$). By contrast, for discrete distributions the probabilities $f_X(x) = P(X = x)$ can never exceed 1.

The CDF is useful for calculating probabilities of various events. Here are some common formulas that follow directly from properties of CDFs:

Lemma 2.16 (Using the CDF)**.** *Let X be a random variable with CDF $F(x) = P(X \le x)$. Then for any real numbers $a < b$, the following hold:*

1. $P(a < X \le b) = F(b) - F(a)$.
2. $P(X > x) = 1 - P(X \le x) = 1 - F(x)$.
3. *If X is continuous, then $P(a < X < b) = P(a \le X \le b) = F(b) - F(a)$, because for continuous distributions the probability of endpoints $P(X = a)$ or $P(X = b)$ is zero.*

Another way to characterize a distribution is via its **quantile function**, which is essentially the inverse of the CDF.

Definition 2.17 (Quantile Function)**.** Given a random variable X with CDF F_X, the **quantile function** (also called the inverse CDF) F_X^{-1} is defined for $q \in [0, 1]$ by

$$F_X^{-1}(q) = \inf\{x \in \mathbb{R} : F_X(x) \geq q\}.$$

In words, $F_X^{-1}(q)$ is the smallest number x such that $P(X \leq x) \geq q$. Equivalently, $F_X^{-1}(q)$ is a value such that $P(X \leq F_X^{-1}(q)) = q$ (for continuous and strictly increasing F_X, this is the unique value with that property).

If F_X is continuous and strictly increasing, then F_X^{-1} is the usual inverse function and indeed $P(X \leq F^{-1}(q)) = q$. Common terminology: $F^{-1}(0.5)$ is the median of X, $F^{-1}(0.25)$ is the first quartile, $F^{-1}(0.75)$ the third quartile, etc. If you haven't seen the inf (infimum) operator before, you can think of it loosely as a minimum value; for continuous distributions the infimum is attained as a minimum when F is continuous.

Example 2.18. For a standard uniform $U(0, 1)$ random variable U, the CDF is $F_U(x) = x$ for $x \in [0, 1]$. The quantile function is thus $F_U^{-1}(q) = q$ for $q \in [0, 1]$. In this trivial case, it so happens that X and $F^{-1}(U)$ have the same distribution F for any distribution F. In fact, an important general result is that if $U \sim U(0, 1)$ and we define $X = F_X^{-1}(U)$ using any continuous CDF F_X, then X follows the distribution F_X. This is known as the *inverse CDF method* and is fundamental in random variable generation: one can simulate any distribution by first simulating a uniform random variable and then transforming it by the quantile function.

2.3 Important Univariate Distributions

Now that we have the general language of pmfs, pdfs, and CDFs, we discuss a few specific probability distributions that are especially common or useful. Some of these we have already encountered informally (e.g. uniform, binomial). As a financial research student, you will likely see these distributions appear in modeling or data analysis contexts.

2.3.1 Common Discrete Distributions

Definition 2.19 (Discrete Uniform Distribution)**.** Let $k \geq 1$ be an integer. A random variable X has a **discrete uniform distribution** on $1, 2, \ldots, k$, denoted $X \sim U1, \ldots, k$, if

$$P(X = x) = \frac{1}{k} \quad \text{for } x = 1, 2, ...k,$$

and $P(X = x) = 0$ for any other x. In other words, X is equally likely to be any of the integers from 1 to k.

This describes the scenario of picking an element at random from a set of k elements with equal probability. A simple example is a fair k-sided die (for $k = 6$, the die outcomes 1 through 6 are each 1/6).

Definition 2.20 (Bernoulli Distribution)**.** Let $0 < p < 1$ be a fixed probability. A random variable X has a **Bernoulli**(p) distribution if

$$P(X = 1) = p, \quad P(X = -) = 1 - p,$$

and those are the only two values X can take. We write $X \sim$ Bernoulli(p).

A Bernoulli(p) random variable is the simplest non-trivial discrete random variable: it represents a single trial with probability p of "success" (outcome 1) and $1-p$ of "failure" (outcome 0). For example, the result of one coin flip can be modeled as Bernoulli(p) with $p = 0.5$ if we let 1 represent heads and 0 tails. It is sometimes convenient to express the Bernoulli pmf in a formula: for $x \in 0, 1$, $P(X = x) = p^x(1 - p)^{1-x.}$ This formula will generalize to the binomial distribution.

Definition 2.21 (Binomial Distribution)**.** If we perform n independent trials, each with probability p of success, and let X be the total number of successes, then X is said to have a **Binomial**(n, p) distribution. The pmf is

$$f_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, 2, \ldots, n, \\ 0, & \text{otherwise.} \end{cases}$$

We write $X \sim$ Binomial(n, p). The binomial distribution generalizes the coin toss example: it is the distribution of the number of heads in n independent coin flips (with probability p of heads on each flip).

Here $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ is the binomial coefficient, which counts the number of ways to choose x items out of n. The binomial distribution can be understood as the sum of n independent Bernoulli(p) trials. It represents the total number of successes in n trials, each with success probability p. For example, if you flip a biased coin n times (each flip with probability p of heads), then the number of heads follows Binomial(n, p). The presence of $\binom{n}{x}$ in the pmf accounts for the fact that those x successes can occur in any $\binom{n}{x}$ distinct positions among the n trials.

The previous discrete distributions are all related: Bernoulli(p) is a special case of Binomial(n, p) with $n = 1$, and the binomial(n, p) is the distribution of a sum of n independent Bernoulli(p) variables.

Definition 2.22 (Poisson Distribution)**.** Let $\lambda > 0$ be a given rate parameter. A random variable X has a **Poisson**(λ) distribution if

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

We denote this $X \sim \text{Poisson}(\lambda)$.

The Poisson distribution is often used to model the number of occurrences of some random event in a fixed interval of time or space, when those events happen at a constant average rate λ and independently of each other. For instance, X could be the number of trades made by a high-frequency trader in one millisecond, or the number of insurance claims in a day. The mean of a Poisson(λ) is λ (as is the variance), meaning λ is the expected count of events. The pmf formula $e^{-\lambda}\lambda^x/x!$ indeed sums to 1 over $x = 0$ to ∞. One reason the Poisson is important is that it arises as an approximation or limit of the binomial distribution in the regime of rare events. Specifically, if n is large and p is small such that $np = \lambda$ (fixed), then Binomial(n, p) is approximately Poisson(λ).

2.3.2 Common Continuous Distributions

We have already encountered the continuous uniform distribution in the $U(0, 1)$ example. Here it is defined in general:

Definition 2.23 (Continuous Uniform Distribution)**.** If X is equally likely to lie anywhere in the interval $[a, b]$, we say X has a **Uniform**(a, b) distribution and write $X \sim U(a, b)$. Its pdf is

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

The CDF is $F_X(x) = \frac{x-a}{b-a}$ for $a \leq x \leq b$. The special case $U(0, 1)$ we have already encountered.

The uniform distribution is fundamental in simulation (as noted, $U(0, 1)$ is the basis of generating other distributions). However, the most celebrated and ubiquitous distribution in probability and statistics is the **normal distribution** (also called the Gaussian distribution).

Definition 2.24 (Normal (Gaussian) Distribution)**.** A random variable X is said to have a **Normal** distribution with mean μ and variance σ^2 if its pdf is

$$f_X(x) = \frac{1}{\sqrt{2\pi\,\sigma^2}} exp\Big(- \frac{(x-\mu)^2}{2\sigma^2}\Big), \qquad \forall x \in \mathbb{R}.$$

We write $X \sim N(\mu, \sigma^2)$.

The normal distribution (also called Gaussian) is arguably the most important distribution in probability and statistics. It is bell-shaped and symmetric about μ. If $X \sim N(\mu, \sigma^2)$, then the probability of X falling within one standard deviation of the mean is about 68%, within two standard deviations about 95%, and within three about 99.7% (this is the 68-95-99.7 rule, stemming from the properties of the normal CDF). In particular, $P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.68$, and $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95$. Because of the Central Limit Theorem (discussed later), the normal often provides a good approximation to the distributions of sums or averages of many independent random factors. Many statistical procedures are built on the normal distribution as a convenient approximation.

The normal distribution is incredibly important in probability, statistics, and financial modeling for several reasons:

- Many natural processes and measurement errors tend to be approximately normal (by virtue of the Central Limit Theorem, which says

roughly that the sum of many small independent effects is approximately normal).

- The normal is analytically convenient: linear combinations of normal variables are normal; many statistical methods assume normality for tractability.

- In finance, asset returns are often modeled as normal (or at least used to be in classical models, though in practice returns have heavier tails than normal).

One quirk of the normal distribution is that the CDF does not have a closed-form antiderivative. In other words,

$$\Phi(x) = P(X \leq x) = \frac{1}{\sqrt{2\pi\,\sigma^2}} \int_{-\infty}^{x} exp\Big(- \frac{(t-\mu)^2}{2\sigma^2} \Big) dt$$

cannot be expressed in terms of elementary functions. Instead, values of $\Phi(x)$ (often standardized to the $N(0,1)$ case) are obtained via numerical tables or software.

When $\mu = 0$ and $\sigma^2 = 1$, we call that the **standard normal** distribution, denoted $Z \sim N(0,1)$. We typically use $\phi(x) = \frac{1}{\sqrt{2\pi}} exp(-x^2/2)$ for its pdf and $\Phi(x) = P(Z \leq x)$ for its CDF. Again, there is no simple closed form for $\Phi(x)$, but it has been tabulated and built into statistical software. Some quantiles of the standard normal that are frequently used in practice: $\Phi^{-1}(0.975) \approx 1.96$, $\Phi^{-1}(0.95) \approx 1.64$, $\Phi^{-1}(0.05) \approx -1.64$, and $\Phi^{-1}(0.025) \approx -1.96$. (These values are often memorized because of their role in 5% significance tests—more on that later.)

The normal family has some very useful properties. For example, it is closed under linear transformations: If $X \sim N(\mu, \sigma^2)$ and a, b are constants with $b \neq 0$, then $Y = a + bX$ also has a normal distribution: specifically $Y \sim N(a + b\mu,\ b^2\sigma^2)$. In particular, any normal X can be standardized to a $Z \sim N(0,1)$ by

$$Z = \frac{X - \mu}{\sigma},$$

and conversely X can be represented as $X = \mu + \sigma Z$ where $Z \sim N(0,1)$.

These relationships imply, for instance,

$$P(a < X \leq b) = P\Big(a < \mu + \sigma Z \leq b\Big) = P\Big(\frac{a-\mu}{\sigma} < Z \leq \frac{b-\mu}{\sigma}\Big)$$
$$= \Phi\Big(\frac{b-\mu}{\sigma}\Big) - \Phi\Big(\frac{a-\mu}{\sigma}\Big).$$

2.4 Random Vectors and Joint Distributions

So far we have focused on a single random variable at a time (univariate distributions). However, in causal inference and financial applications, we usually deal with multiple random variables simultaneously and their relationships. For example, we might have a pair of random variables (X, Y) representing two different measurements on the same experimental unit (e.g. X = treatment status, Y = outcome). We need ways to describe the joint behavior of (X, Y), including whether and how they are related (correlated or independent, etc.).

2.4.1 Joint, Marginal, and Conditional Distributions

A **random vector** is a vector whose components are random variables. Formally, a d-dimensional random vector $\mathbf{X}$ is a function $\mathbf{X} : \Omega \to \mathbb{R}^d$ that maps each outcome to a d-tuple of real numbers. For example, a bivariate random vector (X, Y) is just a pair of random variables. Each component X and Y has its own distribution (called the *marginal distributions*), but together they have a *joint distribution* that can capture any dependence between them. For simplicity, we will discuss the bivariate case $d = 2$. The generalization to higher dimensions ($d > 2$) is conceptually straightforward (just more notation).

Definition 2.25 (Joint CDF)**.** The **joint cumulative distribution function** of a pair (X, Y) is

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$$

for $x, y \in \mathbb{R}$. This gives the probability that X is at most x *and* simultaneously Y is at most y.

Just as the univariate CDF characterizes the distribution of one variable, the joint CDF characterizes the joint distribution of (X, Y). From the joint CDF one can in principle recover all probabilities of events concerning X and Y. Similar to the univariate case, if (X, Y) takes values in a discrete set, we describe it with a joint pmf, and if it's continuous in $\mathbb{R}^2$, we describe it with a joint pdf.

Definition 2.26 (Joint PMF (discrete case))**.** If X and Y are discrete random variables, the **joint probability mass function** is

$$f_{X,Y}(x, y) = P(X = x,\ Y = y),$$

for all values (x, y) in the (countable) support of (X, Y). For any two values x, y, $f_{X,Y}(x, y)$ gives the probability that X equals x *and* Y equals y at the same time.

Example 2.27. Suppose (X, Y) can take the following four combinations of values with the probabilities given in the table:

	$Y = 0$	$Y = 1$
$X = 0$	$1/5$	$1/10$
$X = 1$	$3/10$	$2/5$

This table defines a joint pmf: for instance, $P(X = 0, Y = 1) = 1/10$, $P(X = 1, Y = 0) = 3/10$, etc. We can verify the probabilities sum to 1 (as they must): $\frac{1}{5}+\frac{1}{10}+\frac{3}{10}+\frac{2}{5} = 1$. Using this pmf, one can compute probabilities of more involved events; e.g. $P(X < Y) = P(X = 0, Y = 1) = 1/10$ in this case.

From the joint pmf, we can get the individual distribution of X or Y by summing over the other variable. These are called *marginal distributions*.

Definition 2.28 (Marginal PMF)**.** If (X, Y) has joint pmf $f_{X,Y}(x, y)$, the **marginal pmf** of X is obtained by summing out Y:

$$f_X(x) = P(X = x) = \sum_y P(X = x, Y = y) = \sum_y f_{X,Y}(x, y),$$

where the sum is over all y in the support of Y. Likewise, $f_Y(y) = \sum_x f_{X,Y}(x, y)$ gives the marginal pmf of Y.

In other words, to find the probability $X = x$, consider all joint outcomes where $X = x$ (with any Y) and add up those probabilities. This is an instance of the **Law of Total Probability**, which in general states that

$$P(X = x) = \sum_y P(X = x \text{and} Y = y)$$

summing over any partition (here partitioned by values of Y). In continuous form it will be an integral.

Example 2.29. Using the previous table, the marginal distribution of X is:

$$P(X = 0) = 1/5 + 1/10 = 3/10, \quad P(X = 1) = 3/10 = 2/5 = 7/10.$$

We can present this calculation in an augmented table by adding the row/column totals:

	$Y = 0$	$Y = 1$	$P(X = x)$
$X = 0$	$1/5$	$1/10$	$3/10$
$X = 1$	$3/10$	$2/5$	$7/10$
$P(Y = y)$	$1/2$	$1/2$	1

From the table we also see the marginal distribution of Y: $P(Y = 0) = 1/2$ and $P(Y = 1) = 1/2$.

If (X, Y) is instead jointly continuous (loosely speaking, they have a two-dimensional density), we define a joint pdf analogously:

Definition 2.30 (Joint PDF (continuous case))**.** A pair (X, Y) of continuous random variables has a **joint probability density function** $f_{X,Y}(x, y)$ if $f_{X,Y}(x, y) \geq 0$ for all (x, y) and

$$P((X, Y) \in A) = \iint_{(X,Y)\in A} f_{X,Y}(x, y)dxdy,$$

for A in $\mathbb{R}^2$. In particular, over the whole plane $\iint_{\mathbb{R}^2} f_{X,Y}(x, y)\, dx\, dy = 1$.

This is just the higher-dimensional analog of a one-dimensional density. If we integrate the joint density over some area A, we get the probability that (X, Y) falls in that area.

Example 2.31. Suppose (X, Y) is chosen uniformly at random from the unit square $[0,1] \times [0,1]$. Then

$$f_{X,Y}(x,y) = \begin{cases} 1, & 0 \le x \le 1,\ 0 \le y \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

This means X and Y are each $U(0,1)$ by themselves. In fact $f_X(x) = \int_0^1 1\,dy = 1$ for $0 \le x \le 1$ (and 0 outside that), so $X \sim U(0,1)$; similarly $Y \sim U(0,1)$. For any region A inside the unit square, $P((X,Y) \in A)$ is just the area of A. For instance, $P(X \le 0.5,\ Y \le 0.5) = 0.5 \times 0.5 = 0.25$ since that corresponds to a 0.5×0.5 square quarter of the unit square.

For continuous random vectors, the **marginal density** of X is obtained by integrating out Y:

Definition 2.32 (Marginal PDF)**.** If (X, Y) has joint pdf $f_{X,Y}(x,y)$, then

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$$

is the marginal pdf of X, and similarly $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx$.

Example 2.33. For the uniform distribution on $[0,1]^2$ above, the marginal density of X is

$$f_X(x) = \int_0^1 dy = 1(1-0) = 1$$

for $0 \le x \le 1$ (and 0 outside that range). Thus $X \sim U(0,1)$ marginally. The same is true for Y. Intuitively, if you pick a random point in the unit square, the x-coordinate by itself is uniform on $[0,1]$, and so is the y-coordinate by itself.

So far we have described how to get single-variable distributions (marginals) from a joint distribution. Another important concept is the **conditional distribution**. This describes the distribution of X given some information about Y (or vice versa). It tells us how X and Y relate: if X and Y are dependent, the distribution of X will generally change when we know $Y = y$ has occurred.

Definition 2.34 (Conditional PMF)**.** If (X, Y) is a discrete random vector with joint pmf $f_{X,Y}(x, y)$, the **conditional pmf** of X given $Y = y$ is

$$P(X = x \mid Y = y) = \frac{P(X = x,\ Y = y)}{P(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)},$$

for any x such that $P(Y = y) > 0$. (If $P(Y = y) = 0$, the conditional probability is undefined for that y.) This is denoted $f_{X|Y}(x|y)$. Likewise one can define $f_{Y|X}(y|x) = P(Y = y \mid X = x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$.

This is just the discrete version of Bayes' rule: $P(X|Y) = \frac{P(X,Y)}{P(Y)}$. It formalizes how we update probabilities for X once we know $Y = y$.

Example 2.35. Using our earlier joint pmf table, let's compute a conditional probability. We found $P(Y = 0) = 1/2$. Then

$$P(X = 0 \mid Y = 0) = \frac{P(X = 0, Y = 0)}{P(Y = 0)} = (1/5)(1/2) = 0.4.$$

So if we know $Y = 0$ happened, there's a 40% chance that X was 0. On the other hand,

$$P(Y = 0 \mid X = 0) = \frac{P(X = 0, Y = 0)}{P(X = 0)} = (1/5)(3/10) = 0.667.$$

These two conditional probabilities are not the same (0.4 vs 0.667), which already hints that X and Y are not independent (knowing one changes the distribution of the other).

For continuous random vectors, conditional density is defined similarly via ratio:

Definition 2.36 (Conditional PDF)**.** If (X, Y) has joint pdf $f_{X,Y}(x, y)$, the **conditional density** of X given $Y = y$ (assuming $f_Y(y) > 0$) is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}, \qquad \text{for any } y \text{ with } f_Y(y) > 0.$$

for any $x \in \mathbb{R}$. This satisfies $P(X \in A \mid Y = y) = \int_{x \in A} f_{X|Y}(x|y)\, dx$ for any region A. Similarly define $f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$.

From these definitions, one can always relate the joint and conditional distributions by the formula:

$$f_{X,Y}(x,y) = f_{X|Y}(x|y)\, f_Y(y) = f_{Y|X}(y|x)\, f_X(x).$$

which is just a rearrangement of the definition. In words: *joint = conditional* $\times$ *marginal*. This is another way to express the law of total probability and Bayes' rule in a general form.

2.4.2 Independence

A particularly important relationship between random variables is **independence**. Intuitively, X and Y are independent if knowing the value of one gives no information about the other. Formally:

Definition 2.37 (Independence)**.** Two random variables X and Y are *independent* if for *every* pair of events $A, B \subseteq \mathbb{R}$,

$$P(X \in A,\ Y \in B) = P(X \in A)\, P(Y \in B).$$

We often denote independence by $X \perp Y$.

This means the joint probability of any event about X and any event about Y factorizes into the product of the separate probabilities. In particular, taking $A = x$ and $B = y$ (in discrete cases), we get $P(X = x, Y = y) = P(X = x)P(Y = y)$. Likewise for densities, we'd require $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all x, y. The definition as stated requires checking an infinite collection of sets A and B (all Borel sets, technically). This can be tedious, but fortunately a simpler characterization exists in terms of the joint pdf/pmf:

Theorem 2.38 (Factorization Criterion)**.** *If X and Y have a joint pmf (discrete case) or joint pdf (continuous case) $f_{X,Y}(x,y)$, then*

$$X \perp Y \quad \Longleftrightarrow \quad f_{X,Y}(x,y) = f_X(x)\, f_Y(y) \text{ for all } x, y\ ,$$

provided $f_X(x)$ and $f_Y(y)$ are the marginal distributions.

In words, X and Y are independent if and only if their joint density (or mass) function factors into a product of a function of x alone and a function of y alone.

Corollary 2.39. *Equivalently, X and Y are independent if and only if the conditional distribution of X given Y is the same as the marginal distribution of X. That is,*

$$X \perp Y \iff f_{X|Y}(x|y) = f_X(x) \text{ for all } x, y \text{ (with } f_Y(y) > 0).$$

Proof. Starting from the factorization criterion: if $f_{X,Y}(x, y) = f_X(x)f_Y(y)$, then

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{f_X(x)f_Y(y)}{f_Y(y)} = f_X(x),$$

showing the conditional equals the marginal. Conversely, if $f_{X|Y}(x|y) = f_X(x)$ for all x, y, multiply both sides by $f_Y(y)$ to recover $f_{X,Y}(x, y) = f_X(x)f_Y(y)$. This holds in both discrete and continuous settings. □

In practice, to check independence one often checks if the joint pmf/pdf factorizes as above or if one conditional equals the marginal (as a quick test at some values). If any counterexample is found (like one conditional probability not matching the marginal), then the variables are not independent.

Example 2.40. Recall the joint pmf table earlier. We found $P(X = 0) = 0.3$ and $P(X = 0 \mid Y = 0) = 0.4$. Since $P(X = 0 \mid Y = 0) \neq P(X = 0)$, it follows X and Y are *not* independent (knowing $Y = 0$ changed the probability of $X = 0$). Indeed, if you compare the table to the product of marginals, they differ:

$$f_X(0)f_Y(0) = 0.3 \times 0.5 = 0.15,$$

but

$$f_{X,Y}(0, 0) = 0.2.$$

So the factorization fails as well (0.2 vs 0.15). On the other hand, suppose we had a joint pmf like:

	$Y = 0$	$Y = 1$	$P(X = x)$
$X = 0$	1/4	1/4	1/2
$X = 1$	1/4	1/4	1/2
$P(Y = y)$	1/2	1/2	1

In this hypothetical case, $P(X = x, Y = y) = 1/4$ for each combination with $x \in 0, 1, y \in 0, 1$. We can see that $f_{X,Y}(x, y) = (1/2) \times (1/2) = f_X(x)f_Y(y)$ in each cell. So here X and Y *are* independent. Each outcome pair is exactly the product of the marginals ($0.5 \times 0.5 = 0.25$).

Independence has an important implication: if $X \perp Y$, any function of X is independent of any function of Y. This means once we break the link between X and Y, even non-linear transformations won't induce dependence.

Corollary 2.41. *If X and Y are independent, then for any (deterministic) function h, the random variable $h(Y)$ is also independent of X. That is, $X \perp Y$ implies $X \perp h(Y)$. Likewise $g(X)$ is independent of Y for any function g.*

Proof. Let $\mathcal{A}$ be any event concerning X and $\mathcal{B}$ any event concerning $h(Y)$. The event $h(Y) \in \mathcal{B}$ can be expressed in terms of Y: it is $Y \in h^{-1}(\mathcal{B})$ where $h^{-1}(\mathcal{B}) = \{y : h(y) \in \mathcal{B}\}$. Using independence of X and Y, we have

$$P(X \in \mathcal{A}, h(Y) \in \mathcal{B}) = P(X \in \mathcal{A}, Y \in h^{-1}(\mathcal{B}) = P(X \in \mathcal{A})P(Y \in h^{-1}(\mathcal{B})),$$

since $Y \in h^{-1}(\mathcal{B})$ is an event about Y. But $P(Y \in h^{-1}(\mathcal{B})) = P(h(Y) \in \mathcal{B})$. Therefore

$$P(X \in \mathcal{A}, h(Y) \in \mathcal{B}) = P(X \in \mathcal{A})P(h(Y) \in \mathcal{B}),$$

showing X and $h(Y)$ satisfy the definition of independence. □

Example 2.42. In a causal inference context, suppose D is a treatment indicator (1 if treated, 0 if control) and U represents all other unknown factors affecting the outcome. If we assume D is randomly assigned (independent of U), then by the above corollary D is also independent of any function of U. In particular, if $Y(0) = g(D = 0, U)$ and $Y(1) = g(D = 1, U)$ represent the potential outcomes (outcomes under control and treatment, respectively, as functions of U), random assignment implies $D \perp Y(0)$ and $D \perp Y(1)$. This means the treatment assignment is independent of what the outcome would have been either way, which is a crucial condition for unbiased causal effect estimation.

2.4.3 The Bivariate Normal Distribution

As an important example of a joint distribution, we highlight the **bivariate normal distribution**. Many results in statistics assume joint normality of variables, and it has nice properties (e.g., any marginal or conditional distribution of a normal vector is normal).

Definition 2.43 (Bivariate Normal)**.** A pair (X, Y) is said to have a **bivariate normal distribution** if there exist parameters $\mu_X \in \mathbb{R}$, $\mu_Y \in \mathbb{R}$, and

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \rho\,\sigma_X\sigma_Y \\ \rho\,\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

with $\sigma_X > 0, \sigma_Y > 0$ and σ_{XY} satisfying $\sigma_{XY}^2 < \sigma_X^2\sigma_Y^2$, such that the joint pdf of (X, Y) is given by

$$f_{X,Y}(x, y) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\begin{pmatrix} x-\mu_X \\ y-\mu_Y \end{pmatrix}^\top \Sigma^{-1} \begin{pmatrix} x-\mu_X \\ y-\mu_Y \end{pmatrix}\right),$$

for all $x, y \in \mathbb{R}$. We then write $(X, Y) \sim N(\mu, \Sigma)$, where $\mu = (\mu_X, \mu_Y)^\top$ and Σ is the covariance matrix.

This definition is a bit heavy on linear algebra; an equivalent characterization is that any linear combination $aX + bY$ is normally distributed (for all constants a, b). The parameters μ_X, μ_Y are the means of X and Y, $\sigma_X^2 = Var(X)$, $\sigma_Y^2 = Var(Y)$, and $\sigma_{XY} = Cov(X, Y)$, the covariance between X and Y. The condition $\sigma_{XY}^2 < \sigma_X^2\sigma_Y^2$ ensures Σ is positive-definite (a valid covariance matrix, equivalently the correlation $\rho = \sigma_{XY}/(\sigma_X\sigma_Y)$ satisfies $-1 < \rho < 1$).

A remarkable fact is that any subset of a multivariate normal vector is also multivariate normal, and conditional distributions are normal as well. In particular, for a bivariate normal $(X, Y) \sim N(\mu, \Sigma)$:

- The marginal distribution of X alone is $N(\mu_X, \sigma_X^2)$, and Y alone is $N(\mu_Y, \sigma_Y^2)$. (So if two variables are jointly normal, each is univariate normal.)
- The conditional distribution of Y given $X = x$ is normal with

$$Y \mid X = x \sim N\left(\mu_Y + \frac{\sigma_{XY}}{\sigma_X^2}(x - \mu_X), \sigma_Y^2 - \frac{\sigma_{XY}^2}{\sigma_X^2}\right).$$

This formula says the mean of Y conditional on $X = x$ is linear in x, and the variance of the conditional is smaller than the marginal variance of Y (information about X reduces uncertainty about Y unless $\sigma_{XY} = 0$).

Another useful property: within a multivariate normal distribution, **zero correlation implies independence**. For general distributions, X and Y having zero covariance (or correlation) does *not* guarantee independence, but for normal it does.

Theorem 2.44. *If $(X, Y) \sim N(\mu, \Sigma)$ is bivariate normal, then*

$$X \perp Y \iff \sigma_{XY} = 0.$$

In other words, bivariate normal variables are independent precisely when their covariance is zero.

The linear structure of the normal also means any linear combination of jointly normal variables is normal:

Lemma 2.45 (Linear Combinations of Normals)**.** *If $(X, Y) \sim N(\mu, \Sigma)$ and a, b are constants, then the random variable $Z = aX + bY$ is distributed as*

$$Z \sim N(a\mu_X + b\mu_Y, a^2\sigma_X^2 = 2ab\sigma_{XY} = b^2\sigma_Y^2).$$

In particular, Z is normal with mean equal to $a\mu_X + b\mu_Y$ and variance $\begin{bmatrix} a & b \end{bmatrix} \Sigma \begin{bmatrix} a & b \end{bmatrix}^\top$. If moreover X and Y are independent (so $\sigma_{XY} = 0$), this variance simplifies to $a^2\sigma_X^2 + b^2\sigma_Y^2$.

Given these properties, one can derive other related distributions. For example, summing squares of independent standard normals yields the χ^2 distribution (chi-square), which appears often as a test statistic distribution in statistics:

Theorem 2.46 (χ^2 Distribution)**.** *If $Z_1, Z_2, \ldots, Z_m$ are independent $N(0, 1)$ variables, then*

$$Q = Z_1^2 + Z_2^2 + \cdots + Z_m^2 \sim \chi^2(m),$$

the chi-square distribution with m degrees of freedom. For instance, if $(Z_1, Z_2) \sim N(0, I_2)$ is bivariate standard normal (mean zero, independent components), then $Z_1^2 + Z_2^2 \sim \chi^2(2)$.

The parameter m (degrees of freedom) is usually an integer counting the number of squared normals in the sum. The chi-square distribution is a special case of the gamma distribution family and has mean m and variance $2m$. For m large, $\chi^2(m)$ becomes approximately normal (this can be seen by Central Limit Theorem, or by noting $\chi^2(m)$ is the sum of m independent $\chi^2(1)$ variables and $\chi^2(1)$ has mean 1 and variance 2). A useful corollary connects the chi-square with the distribution of a quadratic form of a normal vector:

Corollary 2.47. *If $X \sim N(\mu, \Sigma)$ is an m-dimensional normal vector (so $\mu \in \mathbb{R}^m$, Σ is $m \times m$ covariance matrix), then*

$$(\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2(m).$$

This result often appears in statistical theory as the distribution of the squared Mahalanobis distance for a sample from $N(\mu, \Sigma)$. It just says that if you standardize a multivariate normal vector by its covariance and mean, the sum of squares of coordinates has a chi-square distribution with degrees equal to the dimension.

Remark 2.48. The χ^2 distribution has a right-skewed shape for small m, but as m grows it starts to look more like a normal distribution. In fact, for large m, $\chi^2(m)$ is approximately $N(m, 2m)$ by the Central Limit Theorem (since it's a sum of m i.i.d. $\chi^2(1)$ variables). For moderate degrees of freedom, the distribution's mode is around $m - 2$ and it has a long right tail. This distribution comes up often in hypothesis testing (e.g., the chi-square test or as part of t and F distributions derivations).

2.5 Summary and What's Next

In this chapter, we reviewed how probability theory describes uncertainty. We introduced the formal setup of sample spaces and events, then defined random variables as functions on outcomes with corresponding distribution functions (CDFs). We saw that the CDF or the pmf/pdf completely characterizes a random variable's distribution. We covered several commonly used distributions — discrete ones like Bernoulli, Binomial, Poisson, and continuous ones like Uniform and normal — which will serve as building blocks or approximations in more complex models.

We then extended the discussion to multiple random variables, introducing joint distributions for random vectors and the concepts of marginal and conditional distributions. We defined independence and highlighted its significance: independent random variables have factorizing joint distributions and greatly simplify analysis (since knowing one tells you nothing about the other). We illustrated these ideas with the bivariate normal distribution, a case where computations are tractable and independence corresponds to zero covariance.

Up to this point, we have focused on describing the *full distribution* of random quantities. However, in practice, we often do not need the entire distribution of a random variable or vector — we might be interested in a few key summary measures (like the mean or variance), or relationships like correlation. For instance, in causal inference, we typically care about differences in expectations (average treatment effects) rather than the entire distribution of outcomes under each treatment.

In the next part of the review, we will explore those summary concepts: – Expectation (mean), variance, covariance, and correlation. – Functions of random variables and how to derive distributions or expectations for them. – Important limit theorems like the Law of Large Numbers and Central Limit Theorem, which justify why the normal distribution appears so often. These tools will allow us to concisely characterize random variables and make inferences, without needing to specify full distributions every time. They form the bridge from probability theory to statistical inference and causal effect estimation in financial research.