# 2

# Summarizing Distributions

> *Probability is expectation founded upon partial knowledge. A perfect acquaintance with all the circumstances affecting the occurrence of an event would change expectation into certainty, and leave neither room nor demand for a theory of probabilities.*
>
> —George Boole

In this chapter, we discuss some important summary features of random variables. Summarizing distributions is a key aspect of an agnostic approach to statistical inference—it allows us to precisely describe features of potentially infinitely complex objects without making further assumptions about the data generating process. Thus, this chapter will be foundational to defining inferential targets that will guide our approach to estimation as the book proceeds.

We begin by defining measures of the "center" and "spread" of the distribution of a single random variable, which will be central to our discussion of estimation in Chapter 3. In the remainder of the chapter, we will discuss summary features of joint distributions, which allow us to describe the relationships between random variables. We will focus on two primary ways of characterizing the relationship between two variables: the conditional expectation function and best linear predictor.[1] These concepts will be central to our treatment of regression in Chapter 4. As with Chapter 1, we conclude by considering multivariate generalizations.

[1] We are fond of an anonymous reviewer's characterization that "we are building the tools needed to relate two variables to one another, which is something we might want to do if we wish to describe general patterns in data. With these tools and these tools only, we have a powerful way to describe what might in principle be very complicated things."

## 2.1 SUMMARY FEATURES OF RANDOM VARIABLES

In Chapter 1, we presented some examples of simple probability distributions (such as Bernoulli, uniform, and normal). In the real world, the full probability distribution of a random variable may be highly complex and therefore may not have a simple mathematical representation. Thus, we will want to be able to describe at least certain key features of such distributions *nonparametrically*, that is, without assuming that these distributions can be fully characterized by a distribution function with a finite number of parameters. (Chapter 5 contains details on the promise, and perils, of parametric models.) Even for very complex distributions, these characteristics have substantive meaning and practical applications. We begin with perhaps the simplest of these summary features: the *expected value*.

### 2.1.1 Expected Values

The expected value (also known as the *expectation* or *mean*) of a random variable can be thought of as the value we would obtain if we took the average over many, many realizations of that random variable. It is the most commonly used measure of the "center" of a probability distribution. The expected value will be very important as we proceed, since our inferential targets can very often be written in terms of expectations. (See, e.g., Sections 3.1, 3.2.2, and 3.2.3.)

**Definition 2.1.1.** *Expected Value*
For a discrete random variable $X$ with probability mass function (PMF) $f$, if $\sum_x |x| f(x) < \infty$,[2] then the *expected value* of X is

$$\mathrm{E}[X] = \sum_x x f(x).$$

For a continuous random variable X with probability density function (PDF) $f$, if $\int_{-\infty}^{\infty} |x| f(x) dx < \infty$, then the *expected value* of X is

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x) dx.$$

The expected value is an operator (hence the $[\cdot]$ notation—see Definition 1.2.3, *Operator on a Random Variable*); it takes as an input a random variable and returns a number. So when we compute the expected value of a random variable, we are applying the *expectation operator*. The following examples illustrate how the expectation operator works in practice.

[2] This regularity condition (and the corresponding one in the continuous case) is known as *absolute convergence*. It is virtually always satisfied in practice, so we omit this technical condition from all subsequent discussion of expectations.

**Example 2.1.2.** *A Fair Die Roll*
Consider, again, a roll of one fair (six-sided) die. Let $X$ be the value of the outcome of the die roll. Then the expected value of $X$ is

$$\mathrm{E}[X] = \sum_{x=1}^{6} xf(x) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{7}{2}.$$

Note that a random variable does not necessarily take on its expected value with positive probability. In this example, $\Pr[X = \mathrm{E}[X]] = \Pr[X = \frac{7}{2}] = f(\frac{7}{2}) = f(3.5) = 0$. △

The example of a Bernoulli distribution is particularly helpful here, and will be used frequently when discussing binary data (see, e.g., Sections 5.1.2, 6.1.4, and 7.1.6).

**Example 2.1.3.** *Bernoulli Distribution*
Let $X$ be a Bernoulli random variable with probability $p$. (Recall that, as shown in Example 1.2.7, we can think of such a random variable as a potentially biased coin flip.) Then

$$\mathrm{E}[X] = \sum_{x=0}^{1} xf(x) = 0 \cdot (1 - p) + 1 \cdot p = p.$$

Notice that this implies a convenient feature of Bernoulli random variables: $\mathrm{E}[X] = \Pr[X = 1]$. △

Finally, we can show that another important distribution, the standard normal distribution, has the property of being centered on zero.

**Example 2.1.4.** *Standard Normal Distribution*
Let $X \sim N(0,1)$. Then the expected value of $X$ is

$$\begin{aligned}
\mathrm{E}[X] &= \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-\frac{x^2}{2}} dx \\
&= \frac{1}{\sqrt{2\pi}} \left( -e^{-\frac{x^2}{2}} \right) \Big|_{-\infty}^{\infty} = 0. \ \triangle
\end{aligned}$$

Since functions of random variables are themselves random variables, they too have expected values. The following theorem establishes how we can compute the expectation of a function of a random variable $g(X)$ without actually deriving the PMF or PDF of $g(X)$.

**Theorem 2.1.5.** *Expectation of a Function of a Random Variable*

- If $X$ is a discrete random variable with PMF $f$ and $g$ is a function of $X$, then

$$\mathrm{E}\big[g(X)\big] = \sum_x g(x)f(x).$$

- If $X$ is a continuous random variable with PDF $f$ and $g$ is a function of $X$ then

$$\mathrm{E}\big[g(X)\big] = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

We omit the proof of this theorem, but it highlights how an operator can be applied to a function of a random variable.

When dealing with functions of a random variable, we will often simply write, for instance, $X^2 + 3X$ to denote the random variable $g(X)$, where $g(X) = X^2 + 3X$. In general, any mathematical expression containing a random variable $X$ denotes the function of $X$ (itself a random variable) defined by that expression.

The following theorem states two basic properties of the expectation operator, the proofs of which will be helpful for understanding what it means to apply operators more generally.

**Theorem 2.1.6.** *Properties of Expected Values*
For a random variable $X$,

- $\forall c \in \mathbb{R}, \mathrm{E}[c] = c.$
- $\forall a \in \mathbb{R}, \mathrm{E}[aX] = a\mathrm{E}[X].$

**Proof:** A constant $c$ can be considered as a discrete random variable $X$ with the PMF

$$f(x) = \begin{cases} 1 & : \quad x = c \\ 0 & : \quad \text{otherwise.} \end{cases}$$

(This is known as a *degenerate distribution* or *degenerate random variable*.) Thus, $\forall c \in \mathbb{R}$,

$$\mathrm{E}[c] = \sum_x xf(x) = cf(c) = c \cdot 1 = c.$$

Now, let $a \in \mathbb{R}$, and let $g(X) = aX$. If $X$ is discrete with PMF $f$, then by Theorem 2.1.5,

$$\mathrm{E}[aX] = \mathrm{E}\big[g(X)\big] = \sum_x g(x)f(x) = \sum_x axf(x) = a\sum_x xf(x) = a\mathrm{E}[X].$$

Likewise, if $X$ is continuous with PDF $f$, then by Theorem 2.1.5,

$$\begin{aligned}\mathrm{E}[aX] &= \mathrm{E}\big[g(X)\big] \\ &= \int_{-\infty}^{\infty} g(x)f(x)dx \\ &= \int_{-\infty}^{\infty} axf(x)dx \\ &= a\int_{-\infty}^{\infty} xf(x)dx \\ &= a\mathrm{E}[X]. \ \square\end{aligned}$$

These results will be fundamental when dealing with expectations going forward.

We need not stop with the univariate case. We can generalize the concept of expected value to the bivariate case in a couple of ways. (Again, further generalization to the case of three or more random variables can be done analogously.) Since each of the elements of a random vector is just a random variable, the expected value of a random vector $(X, Y)$ is defined as the vector of expected values of its components.

**Definition 2.1.7.** *Expectation of a Bivariate Random Vector*
For a random vector $(X, Y)$, the *expected value* of $(X, Y)$ is

$$\mathrm{E}\big[(X, Y)\big] = \big(\mathrm{E}[X], \mathrm{E}[Y]\big).$$

This definition is rarely used, but it illustrates how an operator can be applied to a random vector. More importantly, we can compute the expected value of a function of two random variables, since a function of random variables is itself a random variable.

**Theorem 2.1.8.** *Expectation of a Function of Two Random Variables*

- For discrete random variables $X$ and $Y$ with joint PMF $f$, if $h$ is a function of $X$ and $Y$, then

$$\mathrm{E}\big[h(X,Y)\big] = \sum_x \sum_y h(x,y) f(x,y).$$

- For jointly continuous random variables $X$ and $Y$ with joint PDF $f$, if $h$ is a function of $X$ and $Y$, then

$$\mathrm{E}\big[h(X,Y)\big] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y) f(x,y) dy dx.$$

We omit the proof of this theorem. As in the univariate case, we will often simply write, for instance, $X^2 + 3Y + XY$ to denote the random variable $h(X,Y)$, where $h(X,Y) = X^2 + 3Y + XY$. So we might therefore write $\mathrm{E}[X^2 + 3Y + XY]$ to denote $\mathrm{E}[h(X,Y)]$. In general, any mathematical expression containing one or more random variables denotes the function of random variables (itself a random variable) defined by that expression.

A consequence of Theorem 2.1.8 is the following generalization of Theorem 2.1.6 (*Properties of Expected Values*).

**Theorem 2.1.9.** *Linearity of Expectations*
Let $X$ and $Y$ be random variables. Then, $\forall a, b, c \in \mathbb{R}$,

$$\mathrm{E}[aX + bY + c] = a\mathrm{E}[X] + b\mathrm{E}[Y] + c.$$

**Proof:** Let $X$ and $Y$ be either discrete random variables with joint PMF $f$ or jointly continuous random variables with joint PDF $f$,[3] and let $a, b, c \in \mathbb{R}$. Let $h(X,Y) = aX + bY + c$. If X and Y are discrete, then by Theorem 2.1.8,

$$\begin{aligned}
\mathrm{E}[aX + bY + c] &= \mathrm{E}\big[h(X,Y)\big] \\
&= \sum_x \sum_y h(x,y) f(x,y) \\
&= \sum_x \sum_y (ax + by + c) f(x,y) \\
&= a \sum_x \sum_y x f(x,y) + b \sum_x \sum_y y f(x,y) + c \sum_x \sum_y f(x,y)
\end{aligned}$$

[3] This theorem, and every subsequent theorem that we only prove for these two cases, also holds when one random variable is discrete and the other continuous. We omit the formal proofs for the mixed case because they require measure theory.

$$
\begin{aligned}
&= a\sum_x x\sum_y f(x,y) + b\sum_y y\sum_x f(x,y) + c\sum_x\sum_y f(x,y) \\
&= a\sum_x x f_X(x) + b\sum_y y f_Y(y) + c\sum_x f_X(x) \\
&= a\mathrm{E}[X] + b\mathrm{E}[Y] + c\cdot 1 \\
&= a\mathrm{E}[X] + b\mathrm{E}[Y] + c.
\end{aligned}
$$

Likewise, if $X$ and $Y$ are jointly continuous, then by Theorem 2.1.8,

$$
\begin{aligned}
\mathrm{E}[aX+bY+c] &= \mathrm{E}\big[h(X,Y)\big] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y)f(x,y)dydx \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (ax+by+c)f(x,y)dydx \\
&= a\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf(x,y)dydx + b\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf(x,y)dydx \\
&\quad + c\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dydx \\
&= a\int_{-\infty}^{\infty} x\left(\int_{-\infty}^{\infty} f(x,y)dy\right)dx + b\int_{-\infty}^{\infty} y\left(\int_{-\infty}^{\infty} f(x,y)dx\right)dy \\
&\quad + c\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dydx \\
&= a\int_{-\infty}^{\infty} xf_X(x)dx + b\int_{-\infty}^{\infty} yf_Y(y)dy + c\int_{-\infty}^{\infty} f_X(x)dx \\
&= a\mathrm{E}[X] + b\mathrm{E}[Y] + c\cdot 1 \\
&= a\mathrm{E}[X] + b\mathrm{E}[Y] + c. \quad \square
\end{aligned}
$$

Linearity of expectations is a useful implementation of the properties of expected values (Theorem 2.1.6) and will be key in how we define our primary inferential target in Section 7.1 (*Identification with Potential Outcomes*).

### 2.1.2 Moments, Variances, and Standard Deviations

The expected value is one of the most commonly used summary features of a random variable. We can generalize this concept to further characterize the features of a distribution. We begin with a simple case: *raw moments*, which include the expected value as a special case.

**Definition 2.1.10.** *$j^{th}$ Raw Moment*
For a random variable $X$ and $j \in \mathbb{N}$, the *$j^{th}$ raw moment* of $X$ is

$$\mu'_j = \mathrm{E}\left[X^j\right].$$

The $j^{\text{th}}$ raw moment of a random variable $X$ is the expected value of $X^j$. The expected value is therefore the first raw moment. Raw moments provide summary information about a distribution, describing elements of its shape and location. Sometimes, however, we might seek to have a summary measure that purely reflects the shape and spread of a distribution, and does not depend on its expected value.[4] For $j > 1$, the *$j^{th}$ central moment* generally provides more useful information about the spread and shape of a distribution than the regular $j^{\text{th}}$ moment.

**Definition 2.1.11.** *$j^{th}$ Central Moment*
For a random variable $X$ and $j \in \mathbb{N}$, the *$j^{th}$ central moment* of $X$ is

$$\mu_j = \mathrm{E}\left[(X - \mathrm{E}[X])^j\right].$$

This is referred to as a central moment because it is centered on $\mathrm{E}[X]$.

Note that $\mathrm{E}[X]$ is the first raw moment, *not* the first central moment. The first central moment of any distribution is $\mathrm{E}\left[X - \mathrm{E}[X]\right] = \mathrm{E}[X] - \mathrm{E}[X] = 0$. Note that, when $\mathrm{E}[X] = 0$, then all raw and central moments agree. The sole distinction between raw and central moments lies in whether or not the expected value of $X$ is subtracted before calculations. One of the most frequently employed central moments is the second central moment, also known as the *variance*. Whereas the expected value of a distribution characterizes its location or center, variance characterizes its variability or spread. Formally, variance measures the expected value of the squared difference between the observed value of X and its mean. Consequently, higher variance implies greater unpredictability.[5]

[4] In other words, suppose that $Y = X + c$. Then any given raw moment for $X$ and $Y$ may differ, even though the distribution of $Y$ is identical to that of $X$, just shifted to the right by $c$. So we may want summary measures that reflect the fact that the distributions of $X$ and $Y$ have the same shape.

[5] There are also an infinite number of higher moments, from which one can compute additional features of the shape of a distribution: skewness, kurtosis, etc. In practice, we are very unlikely to make use of moments higher than the second moment of a distribution (unless we need

**Definition 2.1.12.** *Variance*
The *variance* of a random variable $X$ is

$$\mathrm{V}[X] = \mathrm{E}\Big[\big(X - \mathrm{E}[X]\big)^2\Big].$$

In words, the variance is the average squared deviation from the expected value. The following theorem gives an alternative formula for the variance that is often easier to compute in practice.

**Theorem 2.1.13.** *Alternative Formula for Variance*
For a random variable $X$,

$$\mathrm{V}[X] = \mathrm{E}\big[X^2\big] - \mathrm{E}[X]^2.$$

**Proof:**

$$\begin{aligned}
\mathrm{V}[X] &= \mathrm{E}\Big[\big(X - \mathrm{E}[X]\big)^2\Big] \\
&= \mathrm{E}\big[X^2 - 2X\mathrm{E}[X] + \mathrm{E}[X]^2\big] \\
&= \mathrm{E}\big[X^2\big] - 2\mathrm{E}\big[X\mathrm{E}[X]\big] + \mathrm{E}\big[\mathrm{E}[X]^2\big] \\
&= \mathrm{E}\big[X^2\big] - 2\mathrm{E}[X]\mathrm{E}[X] + \mathrm{E}[X]^2 \\
&= \mathrm{E}\big[X^2\big] - 2\mathrm{E}[X]^2 + \mathrm{E}[X]^2 \\
&= \mathrm{E}\big[X^2\big] - \mathrm{E}[X]^2. \quad \square
\end{aligned}$$

Notice that $\mathrm{E}[X]$ is a *constant* and is therefore treated as such each time we apply linearity of expectations above. The same holds for the variance operator, $\mathrm{V}[\cdot]$. The following theorem states some basic properties of the variance operator. Note carefully how these differ from the properties of expected values (Theorem 2.1.6).

**Theorem 2.1.14.** *Properties of Variance*
For a random variable $X$,

- $\forall c \in \mathbb{R}, \mathrm{V}[X + c] = \mathrm{V}[X]$.
- $\forall a \in \mathbb{R}, \mathrm{V}[aX] = a^2\mathrm{V}[X]$.

conditions for asymptotics). Counterexamples presented to us by readers will be met with prompt, but perhaps insincere, apologies.

**Proof:** Let $a, c \in \mathbb{R}$. Then

$$\begin{aligned}
\mathrm{V}[X+c] &= \mathrm{E}\left[\left(X+c-\mathrm{E}[X+c]\right)^2\right] \\
&= \mathrm{E}\left[\left(X+c-\mathrm{E}[X]-c\right)^2\right] \\
&= \mathrm{E}\left[\left(X-\mathrm{E}[X]\right)^2\right] \\
&= \mathrm{V}[X],
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{V}[aX] &= \mathrm{E}\left[\left(aX-\mathrm{E}[aX]\right)^2\right] \\
&= \mathrm{E}\left[\left(aX-a\mathrm{E}[X]\right)^2\right] \\
&= \mathrm{E}\left[a^2\left(X-\mathrm{E}[X]\right)^2\right] \\
&= a^2\mathrm{E}\left[\left(X-\mathrm{E}[X]\right)^2\right] \\
&= a^2\mathrm{V}[X]. \quad \square
\end{aligned}$$

While the variance is one of the most common measures of the "spread" of a distribution, perhaps even more common is the *standard deviation*, which is the square root of the variance.

**Definition 2.1.15.** *Standard Deviation*
The *standard deviation* of a random variable $X$ is

$$\sigma[X] = \sqrt{\mathrm{V}[X]}.$$

We can begin to see why the standard deviation might be preferred when we consider the basic properties of this operator, which are stated in the following theorem.

**Theorem 2.1.16.** *Properties of Standard Deviation*
For a random variable $X$,

- $\forall c \in \mathbb{R}, \sigma[X+c] = \sigma[X]$.
- $\forall a \in \mathbb{R}, \sigma[aX] = |a|\sigma[X]$.

**Proof:** Take the square root of both sides of each equation from Theorem 2.1.14. □

The standard deviation is often preferable to the variance, since it is on the same scale as the random variable of interest. We illustrate this with an example.

**Example 2.1.17.** *A Fair Die Roll*
Consider, again, a roll of one fair (six-sided) die. Let $X$ be the value of the outcome of the die roll. Then

$$\begin{aligned}\mathrm{V}[X] &= \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 \\ &= \sum_{x=1}^{6}\left(x^2 \cdot \frac{1}{6}\right) - \left(\sum_{x=1}^{6} x \cdot \frac{1}{6}\right)^2 \\ &= \frac{91}{6} - \left(\frac{21}{6}\right)^2 \\ &= \frac{35}{12} \approx 2.92.\end{aligned}$$

So

$$\sigma[X] = \sqrt{\mathrm{V}[X]} = \frac{\sqrt{105}}{6} \approx 1.71.$$

Now let $Z = 100X$, which is equivalent to rolling a fair six-sided die with faces labeled 100, 200, 300, 400, 500, and 600. Then

$$\mathrm{V}[Z] = \mathrm{V}[100X] = 100^2\mathrm{V}[X] = 10{,}000\mathrm{V}[X] = \frac{87{,}500}{3} \approx 29{,}200,$$

and

$$\sigma[Z] = \sigma[100X] = |100|\sigma[X] = \frac{50\sqrt{105}}{3} \approx 171. \ \triangle$$

When we scale up the random variable, the standard deviation remains in the same order of magnitude as the spread of outcomes. By contrast, the variance of a random variable can "blow up" when we rescale the random variable. Variance is thus more difficult to interpret than standard deviation, since its magnitude does not clearly correspond to the magnitude of the spread of the distribution.

We can infer some other features of a distribution just by knowing its mean and standard deviation. For example, *Chebyshev's*[6] *Inequality* allows us to put

[6] Also sometimes spelled Chebychev, Chebysheff, Chebychov, Chebyshov, Tchebychev, Tchebycheff, Tschebyschev, Tschebyschef, or Tschebyscheff.

an upper bound on the probability that a draw from the distribution will be more than a given number of standard deviations from the mean.

**Theorem 2.1.18.** *Chebyshev's Inequality*
Let $X$ be a random variable with finite[7] $\sigma[X] > 0$. Then, $\forall \epsilon > 0$,

$$\Pr\Big[\big|X - \mathrm{E}[X]\big| \geq \epsilon\sigma[X]\Big] \leq \frac{1}{\epsilon^2}.$$

We omit the proof here, but see Goldberger (1968, p. 31) for a simple proof via Markov's Inequality. This theorem will be important later for showing that estimators converge to the "right" value (see Theorem 3.2.5, *Chebyshev's Inequality for the Sample Mean*, and Definition 3.2.17, *Consistency*). Notice that what we learn about the distribution from Chebyshev's Inequality is driven by $\sigma[X]$ rather than $\mathrm{V}[X]$.

While useful for theoretical reasons, the bounds provided by Chebyshev's Inequality are somewhat limited in their practical applicability. When we know more about the distribution, knowledge of its expected value and standard deviation can be even more informative. For example, in the case of the *normal distribution*, knowing just these two quantities tells us *everything*.

**Definition 2.1.19.** *Normal Distribution*
A continuous random variable $X$ follows a *normal distribution* if it has PDF

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \forall x \in \mathbb{R},$$

for some constants $\mu, \sigma \in \mathbb{R}$ with $\sigma > 0$. We write $X \sim N(\mu, \sigma^2)$ to denote that $X$ follows a normal distribution with parameters $\mu$ and $\sigma$.

The following theorem implies that knowing the mean and standard deviation of a normal distribution tells us everything about the distribution. Recall that the standard normal distribution, introduced in Example 1.2.19, has a PDF with the same form, but with mean, $\mu$, specified as zero and standard deviation, $\sigma$, specified as one.

[7] Moments, including the expected value, can be infinite or undefined; this is why we invoked the condition of absolute convergence in Definition 2.1.1 (Expected Value). To see this, try taking the expected value of $\frac{1}{X}$, where $X \sim U(0,1)$. Note also that, if the $j^{\text{th}}$ moment of a random variable $X$ is non-finite, then all higher moments (that is, all $k^{\text{th}}$ moments where $k > j$) are also non-finite.

**Theorem 2.1.20.** *Mean and Standard Deviation of the Normal Distribution*
If $X \sim N(\mu, \sigma^2)$, then

- $\text{E}[X] = \mu$.
- $\sigma[X] = \sigma$.

We omit the proof of this theorem. The parameters $\mu$ and $\sigma$ of a normal distribution are its mean and standard deviation, respectively. A normal distribution is thus uniquely specified by its mean and standard deviation.[8] Furthermore, this is why $N(0,1)$ is the *standard* normal distribution: it has the "nice" properties of being centered on zero ($\mu = 0$) and having a standard deviation (and variance) of 1 ($\sigma = \sigma^2 = 1$).

The normal distribution has many nice properties, two of which we state in the following theorem.

**Theorem 2.1.21.** *Properties of the Normal Distribution*
Suppose $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$. Then

- $\forall a, b \in \mathbb{R}$ with $a \neq 0$, if $W = aX + b$, then $W \sim N\big(a\mu_X + b, a^2\sigma_X^2\big)$.
- If $X \perp\!\!\!\perp Y$ and $Z = X + Y$, then $Z \sim N\big(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2\big)$.

We omit the proof. This theorem has an interesting and immediate implication: any linear combination of any number of mutually independent normal random variables must itself be normal.[9]

### 2.1.3 Mean Squared Error

We often want to characterize how well a random variable $X$ approximates a certain value $c$. In order to do so, we will need a metric for how far off $X$ is from $c$, on average. The most commonly used metric is the *mean squared error (MSE)*, which is the expected value of the squared difference between the observed value of $X$ and $c$.[10] As we proceed, we will see that MSE has a

[8] Caution: when we see, for example, $X \sim N(2,9)$, this means $X$ is normally distributed with mean $\mu = 2$ and *variance* $\sigma^2 = 9$, so $\sigma = \sqrt{9} = 3$. This convention is used because a multivariate normal distribution is characterized by its mean and covariance matrix (see Definition 2.3.1).

[9] This result, implying that the normal distribution is a "stable distribution," is closely related to the Central Limit Theorem (Theorem 3.2.24). We refer interested readers to DasGupta (2008, Ch. 5).

[10] MSE is most commonly used as a term to quantify the precision of an estimator; as we will discuss in Chapter 3, this is in fact the same definition in a different context (see Definition 3.2.14, *MSE of an Estimator*). There are alternative metrics, such as *mean absolute error* (MAE), which is the expected value of the absolute difference between the observed value of $X$ and $c$.

number of appealing properties for the applied researcher. Notice that the MSE is closely related to other concepts we have introduced in this chapter. The MSE about zero is the same as the second raw moment, and the MSE about E[$X$] is the same as the second central moment, which is also the variance.

**Definition 2.1.22.** *Mean Squared Error (MSE) about c*
For a random variable $X$ and $c \in \mathbb{R}$, the *mean squared error* of $X$ about $c$ is $\mathrm{E}[(X-c)^2]$.

A closely related quantity is the *root mean squared error* about $c$, $\sqrt{\mathrm{E}[(X-c)^2]}$. Much as the standard deviation rescales the variance in a way that ensures that it remains on the same scale as the range of outcomes for $X$, root MSE rescales the MSE so that it remains on the same scale as $X-c$. For these reasons, researchers often prefer to report the root MSE rather than the MSE.

We can apply Theorem 2.1.13 (*Alternative Formula for Variance*) to derive an alternative formula for MSE, which will help elucidate how MSE relates to the expected value and variance of $X$.

**Theorem 2.1.23.** *Alternative Formula for MSE*
For a random variable $X$ and $c \in \mathbb{R}$,

$$\mathrm{E}\left[(X-c)^2\right] = \mathrm{V}[X] + \left(\mathrm{E}[X]-c\right)^2.$$

**Proof:**

$$\begin{aligned}
\mathrm{E}\left[(X-c)^2\right] &= \mathrm{E}\left[X^2 - 2cX + c^2\right] \\
&= \mathrm{E}\left[X^2\right] - 2c\mathrm{E}[X] + c^2 \\
&= \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 + \mathrm{E}[X]^2 - 2c\mathrm{E}[X] + c^2 \\
&= \left(\mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2\right) + \left(\mathrm{E}[X]^2 - 2c\mathrm{E}[X] + c^2\right) \\
&= \mathrm{V}[X] + \left(\mathrm{E}[X]-c\right)^2. \quad \square
\end{aligned}$$

Our decomposition also gives some insight into the utility of the expected value for approximating a random variable $X$. Theorem 2.1.23 directly implies that the expected value, E[$X$], has an interpretation as the best predictor of $X$ in terms of MSE:

**Theorem 2.1.24.** *The Expected Value Minimizes MSE*
For a random variable $X$, the value of $c$ that minimizes the MSE of $X$ about $c$ is $c = \mathrm{E}[X]$.

**Proof:**

$$\begin{aligned}\operatorname*{arg\,min}_{c\in\mathbb{R}} \mathrm{E}\big[(X-c)^2\big] &= \operatorname*{arg\,min}_{c\in\mathbb{R}}\Big(\mathrm{V}[X] + \big(\mathrm{E}[X]-c\big)^2\Big)\\ &= \operatorname*{arg\,min}_{c\in\mathbb{R}}\big(\mathrm{E}[X]-c\big)^2\\ &= \mathrm{E}[X]. \ \square\end{aligned}$$

In other words, if we had to pick one number as a prediction of the value of $X$, the "best" choice (in terms of minimizing MSE) would be $\mathrm{E}[X]$. We illustrate Theorem 2.1.24 by minimizing MSE for a fair coin flip in the example below.

**Example 2.1.25.** *A Fair Coin Flip*
Consider, again, a fair coin flip. Let $X = 0$ if the coin comes up heads and $X = 1$ if the coin comes up tails. What is the *minimum MSE* guess for the value of $X$? The PMF of $X$ is

$$f(x) = \begin{cases} \frac{1}{2} & : \ x \in \{0,1\} \\ 0 & : \ \text{otherwise,} \end{cases}$$

so the MSE about $c$ is

$$\begin{aligned}\mathrm{E}\big[(X-c)^2\big] &= \frac{1}{2}(0-c)^2 + \frac{1}{2}(1-c)^2\\ &= \frac{1}{2}(c^2+1+c^2-2c)\\ &= \frac{1}{2}(1+2c^2-2c)\\ &= \frac{1}{2}+c^2-c.\end{aligned}$$

The first-order condition is thus

$$0 = \frac{d}{dc}\mathrm{E}\big[(X-c)^2\big] = \frac{d}{dc}\left(\frac{1}{2}+c^2-c\right) = 2c-1,$$

which is solved by $c = \frac{1}{2}$. This is $\mathrm{E}[X]$ (see Example 2.1.3, *Bernoulli Distribution*). △

Figure 2.1.1 illustrates this solution: the value that minimizes MSE is, as expected, $c = \mathrm{E}[X] = \frac{1}{2}$.

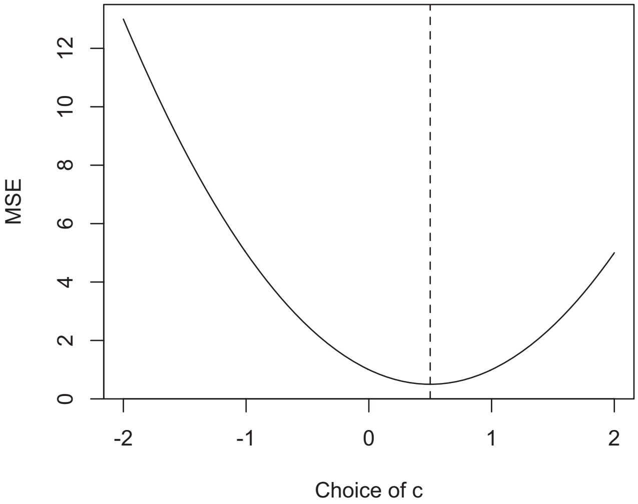

**FIGURE 2.1.1.** *Minimum MSE Solution for a Fair Coin Flip*

## 2.2 SUMMARY FEATURES OF JOINT DISTRIBUTIONS

One of the primary uses of statistics in the social and health sciences is to describe relationships between random variables. For example, we often want to know: if one quantity, $X$, is high, is another, $Y$, more or less likely to be high? There exist many useful summary features of the joint distribution of two or more random variables. The simplest of these are covariance and correlation.

### 2.2.1 Covariance and Correlation

We will show that a natural generalization of variance to the bivariate case is the *covariance*. Covariance measures the extent to which two random variables "move together." If $X$ and $Y$ have positive covariance, that means that the value of $X$ tends to be large when the value of $Y$ is large and small when the value of $Y$ is small. If $X$ and $Y$ have negative covariance, then the opposite is true: the value of $X$ tends to be small when the value of $Y$ is large and large when the value of $Y$ is small.

**Definition 2.2.1.** *Covariance*
The *covariance* of two random variables $X$ and $Y$ is

$$\text{Cov}[X, Y] = \text{E}\Big[\big(X - \text{E}[X]\big)\big(Y - \text{E}[Y]\big)\Big].$$

As with expected value and variance, $\text{Cov}[\cdot, \cdot]$ is an operator, not a function of the values its arguments take on, so $\text{Cov}[X, Y]$ is a constant, not a random variable. As with variance, there is an alternative formula for covariance that is generally easier to compute in practice.

**Theorem 2.2.2.** *Alternative Formula for Covariance*
For random variables $X$ and $Y$,

$$\text{Cov}[X,Y] = \text{E}[XY] - \text{E}[X]\text{E}[Y].$$

**Proof:**

$$\begin{aligned}\text{Cov}[X,Y] &= \text{E}\Big[\big(X-\text{E}[X]\big)\big(Y-\text{E}[Y]\big)\Big] \\ &= \text{E}\big[XY - X\text{E}[Y] - Y\text{E}[X] + \text{E}[X]\text{E}[Y]\big] \\ &= \text{E}[XY] - \text{E}[X]\text{E}[Y] - \text{E}[X]\text{E}[Y] + \text{E}[X]\text{E}[Y] \\ &= \text{E}[XY] - \text{E}[X]\text{E}[Y]. \quad \square\end{aligned}$$

We will put this representation of the covariance into immediate use in the proof of the following theorem, which generalizes Theorem 2.1.14 (*Properties of Variance*) to the bivariate case, allowing us to draw a fundamental link between variances and covariances.

**Theorem 2.2.3.** *Variance Rule*
Let $X$ and $Y$ be random variables. Then

$$\text{V}[X+Y] = \text{V}[X] + 2\text{Cov}[X,Y] + \text{V}[Y].$$

More generally, $\forall a,b,c \in \mathbb{R}$,

$$\text{V}[aX+bY+c] = a^2\text{V}[X] + 2ab\text{Cov}[X,Y] + b^2\text{V}[Y].$$

**Proof:**

$$\begin{aligned}\text{V}[X+Y] &= \text{E}\big[(X+Y)^2\big] - \text{E}[X+Y]^2 \\ &= \text{E}\big[X^2 + 2XY + Y^2\big] - \big(\text{E}[X]+\text{E}[Y]\big)^2 \\ &= \text{E}\big[X^2\big] + 2\text{E}[XY] + \text{E}\big[Y^2\big] - \text{E}[X]^2 - 2\text{E}[X]\text{E}[Y] - \text{E}[Y]^2 \\ &= \text{E}\big[X^2\big] - \text{E}[X]^2 + 2\big(\text{E}[XY] - \text{E}[X]\text{E}[Y]\big) + \text{E}\big[Y^2\big] - \text{E}[Y]^2 \\ &= \text{V}[X] + 2\text{Cov}[X,Y] + \text{V}[Y].\end{aligned}$$

The proof of the more general version is omitted. □

Note that, unlike expected values, we do *not* have linearity of variances: $\text{V}[aX+bY] \neq a\text{V}[X] + b\text{V}[Y]$. To remember the variance rule, it may be helpful to recall the algebraic formula $(x+y)^2 = x^2 + 2xy + y^2$. Indeed, the above proof shows that this similarity is not coincidental.

We now derive a few properties of covariance.

**Theorem 2.2.4.** *Properties of Covariance*
For random variables $X$, $Y$, $Z$, and $W$,

- $\forall c, d \in \mathbb{R}, \text{Cov}[c, X] = \text{Cov}[X, c] = \text{Cov}[c, d] = 0$.
- $\text{Cov}[X, Y] = \text{Cov}[Y, X]$.
- $\text{Cov}[X, X] = \text{V}[X]$.
- $\forall a, b, c, d \in \mathbb{R}, \text{Cov}[aX + c, bY + d] = ab\text{Cov}[X, Y]$.
- $\text{Cov}[X + W, Y + Z] = \text{Cov}[X, Y] + \text{Cov}[X, Z] + \text{Cov}[W, Y] + \text{Cov}[W, Z]$.

**Proof:** Let $a, b, c, d \in \mathbb{R}$. Then

$$\text{Cov}[c, X] = \text{E}[cX] - \text{E}[c]\text{E}[X] = c\text{E}[X] - c\text{E}[X] = 0.$$

$$\text{Cov}[X, c] = \text{E}[Xc] - \text{E}[X]\text{E}[c] = c\text{E}[X] - c\text{E}[X] = 0.$$

$$\text{Cov}[c, d] = \text{E}[cd] - \text{E}[c]\text{E}[d] = cd - cd = 0.$$

$$\begin{aligned}\text{Cov}[X, Y] &= \text{E}[XY] - \text{E}[X]\text{E}[Y] \\ &= \text{E}[YX] - \text{E}[Y]\text{E}[X] \\ &= \text{Cov}[Y, X].\end{aligned}$$

$$\text{Cov}[X, X] = \text{E}[XX] - \text{E}[X]\text{E}[X] = \text{E}\left[X^2\right] - \text{E}[X]^2 = \text{V}[X].$$

$$\begin{aligned}\text{Cov}[aX + c, bY + d] &= \text{E}\Big[\big(aX + c - \text{E}[aX + c]\big)\big(bY + d - \text{E}[bY + d]\big)\Big] \\ &= \text{E}\Big[\big(aX + c - a\text{E}[X] - c\big)\big(bY + d - b\text{E}[Y] - d\big)\Big] \\ &= \text{E}\Big[a\big(X - \text{E}[X]\big)b\big(Y - \text{E}[Y]\big)\Big] \\ &= ab\text{E}\Big[\big(X - \text{E}[X]\big)\big(Y - \text{E}[Y]\big)\Big] \\ &= ab\text{Cov}[X, Y].\end{aligned}$$

$$\begin{aligned}\text{Cov}[X + W, Y + Z] &= \text{E}\big[(X + W)(Y + Z)\big] - \text{E}[X + W]\text{E}[Y + Z] \\ &= \text{E}[XY + XZ + WY + WZ] \\ &\quad - \big(\text{E}[X] + \text{E}[W]\big)\big(\text{E}[Y] + \text{E}[Z]\big) \\ &= \text{E}[XY] + \text{E}[XZ] + \text{E}[WY] + \text{E}[WZ] \\ &\quad - \text{E}[X]\text{E}[Y] - \text{E}[X]\text{E}[Z] - \text{E}[W]\text{E}[Y] - E[W]\text{E}[Z]\end{aligned}$$

$$
\begin{aligned}
&= \big(\mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]\big) + \big(\mathrm{E}[XZ] - \mathrm{E}[X]\mathrm{E}[Z]\big) \\
&\quad + \big(\mathrm{E}[WY] - \mathrm{E}[W]\mathrm{E}[Y]\big) + \big(\mathrm{E}[WZ] - \mathrm{E}[W]\mathrm{E}[Z]\big) \\
&= \mathrm{Cov}[X,Y] + \mathrm{Cov}[X,Z] + \mathrm{Cov}[W,Y] + \mathrm{Cov}[W,Z]. \quad \square
\end{aligned}
$$

Although many of these properties will be helpful for our calculus, perhaps the most noteworthy is the third: $\mathrm{Cov}[X,X] = \mathrm{V}[X]$. Variance is effectively a special case of covariance: the covariance of a random variable with itself is its variance. Thus, covariance is indeed a natural generalization of variance.

Another measure of the relationship between two random variables is their *correlation*. Readers are likely to be familiar with the idea of correlation in conversational use; we provide a formal definition in the statistical setting.

**Definition 2.2.5.** *Correlation*
The *correlation* of two random variables $X$ and $Y$ with $\sigma[X] > 0$ and $\sigma[Y] > 0$ is

$$\rho[X,Y] = \frac{\mathrm{Cov}[X,Y]}{\sigma[X]\sigma[Y]}.$$

Much like standard deviation rescales variance, correlation rescales covariance to make its interpretation clearer. Importantly, since the denominator of the expression for correlation is positive, correlation is positive when covariance is positive and negative when covariance is negative. (Note that, unlike covariance, correlation is undefined when the variance of either variable is zero.)

*Linear dependence* describes the relationship between two random variables where one can be written as a linear function of the other; $X$ and $Y$ are linearly dependent if $\exists a, b \in \mathbb{R}$ such that $Y = aX + b$. Even if two random variables are not exactly linearly dependent, they can be more or less nearly so. Correlation measures the degree of linear dependence between two random variables and is bounded in $[-1, 1]$, where a correlation of one represents perfect positive linear dependence, a correlation of $-1$ represents perfect negative linear dependence, and a correlation of zero implies no linear relationship.[11] We formalize these facts (except for the last one) in the following theorem.

[11] For this reason, some people prefer to call $\rho[\cdot,\cdot]$, the *linear correlation*. Some nonlinear measures of correlation include the Spearman and Kendall rank correlations, discussed in Wackerly, Mendenhall, and Scheaffer (2008, Ch. 15.10).

**Theorem 2.2.6.** *Correlation and Linear Dependence*
For random variable $X$ and $Y$,

- $\rho[X,Y] \in [-1,1]$.
- $\rho[X,Y] = 1 \iff \exists a,b \in \mathbb{R}$ with $b > 0$ such that $Y = a + bX$.
- $\rho[X,Y] = -1 \iff \exists a,b \in \mathbb{R}$ with $b > 0$ such that $Y = a - bX$.

We omit the proof of this theorem.[12] The following theorem states some other important properties of correlation.

**Theorem 2.2.7.** *Properties of Correlation*
For random variables $X$, $Y$, and $Z$,

- $\rho[X,Y] = \rho[Y,X]$.
- $\rho[X,X] = 1$.
- $\rho[aX+c, bY+d] = \rho[X,Y]$, $\forall a,b,c,d \in \mathbb{R}$ such that either $a,b > 0$ or $a,b < 0$.
- $\rho[aX+c, bY+d] = -\rho[X,Y]$, $\forall a,b,c,d \in \mathbb{R}$ such that either $a < 0 < b$ or $b < 0 < a$.

**Proof:**

$$\rho[X,Y] = \frac{\text{Cov}[X,Y]}{\sigma[X]\sigma[Y]} = \frac{\text{Cov}[Y,X]}{\sigma[Y]\sigma[X]} = \rho[Y,X].$$

$$\rho[X,X] = \frac{\text{Cov}[X,X]}{\sigma[X]\sigma[X]} = \frac{\text{V}[X]}{\sigma[X]^2} = \frac{\text{V}[X]}{\left(\sqrt{\text{V}[X]}\right)^2} = \frac{\text{V}[X]}{\text{V}[X]} = 1.$$

Let $a,b,c,d \in \mathbb{R}$ with $a,b \neq 0$. Then

$$\rho[aX+c, bY+d] = \frac{\text{Cov}[aX+c, bY+d]}{\sigma[aX+c]\sigma[bY+d]} = \frac{ab\text{Cov}[X,Y]}{|a||b|\sigma[X]\sigma[Y]} = \frac{ab}{|ab|}\rho[X,Y].$$

[12] The fact that correlation is bounded in $[-1,1]$ is equivalent to the well-known *Cauchy–Schwarz Inequality*, which, in one form, states that, for random variables $X$ and $Y$, $\text{Cov}[X,Y]^2 \leq \text{V}[X]\text{V}[Y]$. This equivalence is shown as follows:

$$\text{Cov}[X,Y]^2 \leq \text{V}[X]\text{V}[Y] \iff \frac{\text{Cov}[X,Y]^2}{\text{V}[X]\text{V}[Y]} \leq 1 \iff \sqrt{\frac{\text{Cov}[X,Y]^2}{\text{V}[X]\text{V}[Y]}} \leq 1 \iff \frac{\sqrt{\text{Cov}[X,Y]^2}}{\sqrt{\text{V}[X]}\sqrt{\text{V}[Y]}} \leq 1$$

$$\iff \frac{|\text{Cov}[X,Y]|}{\sigma[X]\sigma[Y]} \leq 1 \iff -1 \leq \frac{\text{Cov}[X,Y]}{\sigma[X]\sigma[Y]} \leq 1 \iff -1 \leq \rho[X,Y] \leq 1.$$

If $a, b > 0$ or $a, b < 0$ (that is, $a$ and $b$ have the same sign), then $\frac{ab}{|ab|} = 1$, so

$$\rho[aX + c, bY + d] = \rho[X, Y],$$

and if $a < 0 < b$ or $b < 0 < a$ (that is, $a$ and $b$ have opposite signs), then $\frac{ab}{|ab|} = -1$, so

$$\rho[aX + c, bY + d] = -\rho[X, Y]. \ \square$$

These results agree with our intuitions. In particular, $X$ is perfectly (positively) correlated with itself, and linear transformations of $X$ or $Y$ do not change their absolute degree of linear dependence.

### 2.2.2 Covariance, Correlation, and Independence

Independence, covariance, and correlation are tightly connected topics. In this brief section, we show how independence relates to correlation and covariance. We begin by deriving some properties of independent random variables.

**Theorem 2.2.8.** *Implications of Independence (Part II)*
If $X$ and $Y$ are independent random variables, then

- $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$.
- Covariance is zero: $\mathrm{Cov}[X, Y] = 0$.
- Correlation is zero: $\rho[X, Y] = 0$.
- Variances are additive: $\mathrm{V}[X + Y] = \mathrm{V}[X] + \mathrm{V}[Y]$.

**Proof:** Let $X$ and $Y$ be either two discrete independent random variables with joint PMF $f(x,y)$ or two jointly continuous independent random variables with joint PDF $f(x,y)$. Then, $\forall x, y \in \mathbb{R}$, $f(x,y) = f_X(x)f_Y(y)$. So, if $X$ and $Y$ are discrete, then

$$\begin{aligned}
\mathrm{E}[XY] &= \sum_x \sum_y xyf(x,y) \\
&= \sum_x \sum_y xyf_X(x)f_Y(y) \\
&= \sum_x xf_X(x) \sum_y yf_Y(y) \\
&= \mathrm{E}[X]\mathrm{E}[Y].
\end{aligned}$$

Likewise, if $X$ and $Y$ are jointly continuous, then

$$\begin{aligned}
\mathrm{E}[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dydx \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)f_Y(y)dydx \\
&= \int_{-\infty}^{\infty} xf_X(x)dx \int_{-\infty}^{\infty} yf_Y(y)dy \\
&= \mathrm{E}[X]\mathrm{E}[Y].
\end{aligned}$$

Thus,

$$\mathrm{Cov}[X,Y] = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] = \mathrm{E}[X]\mathrm{E}[Y] - \mathrm{E}[X]\mathrm{E}[Y] = 0,$$

and

$$\rho[X,Y] = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{V}[X]}\sqrt{\mathrm{V}[Y]}} = \frac{0}{\sqrt{\mathrm{V}[X]}\sqrt{\mathrm{V}[Y]}} = 0.$$

Finally, by Theorem 2.2.3 (*Variance Rule*),

$$\begin{aligned}
\mathrm{V}[X+Y] &= \mathrm{V}[X] + 2\mathrm{Cov}[X,Y] + \mathrm{V}[Y] \\
&= \mathrm{V}[X] + 2\cdot 0 + \mathrm{V}[Y] \\
&= \mathrm{V}[X] + \mathrm{V}[Y]. \ \square
\end{aligned}$$

Recall that random variables $X$ and $Y$ are independent if and only if knowing the outcome for one provides no information about the probability of any outcome for the other. In other words, independence means there is *no relationship* between the outcome for $X$ and the outcome for $Y$. Thus, it is not surprising that no relationship implies no linear relationship. However, unlike in Theorem 1.3.17 (*Implications of Independence, Part I*), the converses of the statements in Theorem 2.2.8 do *not* necessarily hold—lack of correlation does *not* imply independence. The following example illustrates this fact.

**Example 2.2.9.** *Zero Correlation Does Not Imply Independence*
Let $X$ be a discrete random variable with marginal PMF

$$f_X(x) = \begin{cases} \frac{1}{3} & : \quad x \in \{-1,0,1\} \\ 0 & : \quad \text{otherwise,} \end{cases}$$

and let $Y = X^2$, so $Y$ has marginal PMF

$$f_Y(y) = \begin{cases} \frac{1}{3} & : \quad y = 0 \\ \frac{2}{3} & : \quad y = 1 \\ 0 & : \quad \text{otherwise,} \end{cases}$$

and the joint PMF of $X$ and $Y$ is

$$f(x,y) = \begin{cases} \frac{1}{3} & : \quad (x,y) \in \big\{(0,0),(1,1),(-1,1)\big\} \\ 0 & : \quad \text{otherwise.} \end{cases}$$

Clearly, $X$ and $Y$ are not independent; for instance, $f(0,1) = 0 \neq \frac{2}{9} = \frac{1}{3} \cdot \frac{2}{3} = f_X(0)f_Y(1)$. Yet

$$\begin{aligned} \text{Cov}[X,Y] &= \text{E}[XY] - \text{E}[X]E[Y] \\ &= \sum_x \sum_y xyf(x,y) - \sum_x xf_X(x) \sum_y yf_Y(y) \\ &= \left(-1 \cdot 1 \cdot \frac{1}{3} + 0 \cdot 0 \cdot \frac{1}{3} + 1 \cdot 1 \cdot \frac{1}{3}\right) \\ &\quad - \left(-1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3}\right)\left(0 \cdot \frac{1}{3} + 1 \cdot \frac{2}{3}\right) \\ &= 0 - 0 \cdot \frac{2}{3} \\ &= 0. \end{aligned}$$

Thus, $\rho[X,Y] = \dfrac{\text{Cov}[X,Y]}{\sigma[X]\sigma[Y]} = 0$. So we have $\rho[X,Y] = 0$, but $X \not\perp\!\!\!\perp Y$.[13] △

[13] While $\rho(X,Y) = 0$ does not generally imply $X \perp\!\!\!\perp Y$ even when $X$ and $Y$ are both normal, it does in the special case where the joint distribution of $X$ and $Y$ is *bivariate normal*, that is, $X$ and $Y$ have the joint PDF

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]}$$

for some $\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho \in \mathbb{R}$ with $\sigma_X > 0$, $\sigma_Y > 0$, and $-1 \leq \rho \leq 1$. As with the univariate normal distribution, it can be shown that the parameters correspond to the appropriate summary features of the distribution: $\mu_X = \text{E}[X]$, $\mu_Y = \text{E}[Y]$, $\sigma_X = \sigma[X]$, $\sigma_Y = \sigma[Y]$, and $\rho = \rho[X,Y]$.

### 2.2.3 Conditional Expectations and Conditional Expectation Functions

Just as we were able to compute the conditional probability of an event (Definition 1.1.8) and the conditional PMF (Definition 1.3.7) or PDF (Definition 1.3.14) of a random variable, we can compute the *conditional expectation* of a random variable, that is, the expected value of a random variable given that some other random variable takes on a certain value.

The idea of a conditional expectation is key to how we will characterize the relationship of one distribution to another in this book. Conditional expectations allow us to describe how the "center" of one random variable's distribution changes once we condition on the observed value of another random variable. To calculate a conditional expectation of a random variable, we replace the (marginal) PMF/PDF with the appropriate conditional PMF/PDF in the formula for expected value.

**Definition 2.2.10.** *Conditional Expectation*
For discrete random variables $X$ and $Y$ with joint PMF $f$, the *conditional expectation* of $Y$ given $X = x$ is

$$\mathrm{E}[Y|X=x] = \sum_y y f_{Y|X}(y|x), \forall x \in \mathrm{Supp}[X].$$

For jointly continuous random variables $X$ and $Y$ with joint PDF $f$, the *conditional expectation* of $Y$ given $X = x$ is

$$\mathrm{E}[Y|X=x] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy, \forall x \in \mathrm{Supp}[X].$$

Recall that an (unconditional) expected value is an operator, meaning it takes a random variable as an input and returns a number that describes a feature of the distribution of that random variable. In the case of conditional expectations, $\mathrm{E}[Y|X=x]$ is a family of operators on the random vector $(X, Y)$ indexed by $x$. That is, for every $x \in \mathrm{Supp}[X]$, $\mathrm{E}[Y|X=x]$ is an operator that summarizes a particular cross section of the joint distribution of $X$ and $Y$, namely, the conditional distribution of $Y$ given $X = x$. So, for example, if $X$ is a Bernoulli random variable, so that $\mathrm{Supp}[X] = \{0,1\}$, then we have two operators: $\mathrm{E}[Y|X = 0]$ and $\mathrm{E}[Y|X = 1]$, where $\mathrm{E}[Y|X = 0]$ gives the long-run average value of $Y$ among all realizations where $X = 0$, and $\mathrm{E}[Y|X = 1]$ gives the long-run average value of $Y$ among all realizations where $X = 1$. Of course, if $X$ is continuous, then we have an infinite number of operators, each of which describes the center of the conditional distribution of $Y$ given $X = x$ for a particular value of $x$.

As with regular expectations (see Theorem 2.1.5, *Expectation of a Function of a Random Variable* and Theorem 2.1.8, *Expectation of a Function of Two*

*Random Variables*), we can apply conditional expectations to functions of one or more random variables.

**Theorem 2.2.11.** *Conditional Expectation of a Function of Random Variables*
For discrete random variables $X$ and $Y$ with joint PMF $f$, if $h$ is a function of $X$ and $Y$, then the conditional expectation of $h(X,Y)$ given $X=x$ is

$$\mathrm{E}\big[h(X,Y)\,\big|\,X=x\big]=\sum_{y}h(x,y)f_{Y|X}(y|x), \forall x\in\mathrm{Supp}[X].$$

For jointly continuous random variables $X$ and $Y$ with joint PDF $f$, if $h$ is a function of $X$ and $Y$, then the conditional expectation of $h(X,Y)$ given $X=x$ is

$$\mathrm{E}\big[h(X,Y)\,\big|\,X=x\big]=\int_{-\infty}^{\infty}h(x,y)f_{Y|X}(y|x)dy, \forall x\in\mathrm{Supp}[X].$$

We omit the proof of this theorem. Everything else carries through just as though we were operating on a univariate PMF/PDF. For example, we can define the *conditional variance* of $Y$ given $X=x$.

**Definition 2.2.12.** *Conditional Variance*
For random variables $X$ and $Y$, the *conditional variance* of $Y$ given $X=x$ is

$$\mathrm{V}[Y|X=x]=\mathrm{E}\Big[\big(Y-\mathrm{E}[Y|X=x]\big)^2\,\Big|\,X=x\Big], \forall x\in\mathrm{Supp}[X].$$

As with regular variance, we can derive an alternative formula for conditional variance that is generally easier to work with in practice.

**Theorem 2.2.13.** *Alternative Formula for Conditional Variance*
For random variables $X$ and $Y$, $\forall x\in\mathrm{Supp}[X]$,

$$\mathrm{V}[Y|X=x]=\mathrm{E}\big[Y^2\,\big|\,X=x\big]-\mathrm{E}[Y|X=x]^2.$$

The proof is essentially the same as the proof of Theorem 2.1.13 (*Alternative Formula for Variance*). Additionally, just like unconditional expectations, conditional expectations are linear.

**Theorem 2.2.14.** *Linearity of Conditional Expectations*
For random variables $X$ and $Y$, if $g$ and $h$ are functions of $X$, then $\forall x \in \text{Supp}[X]$,

$$\text{E}\big[g(X)Y + h(X) \,\big|\, X = x\big] = g(x)\text{E}[Y|X = x] + h(x).$$

**Proof:** Let $X$ and $Y$ be either discrete random variables with joint PMF $f$ or jointly continuous random variables with joint PDF $f$, and let $g$ and $h$ be functions of $X$. If $X$ and $Y$ are discrete, then by Theorem 2.2.11, $\forall x \in \text{Supp}[X]$,

$$\begin{aligned}
\text{E}\big[g(X)Y + h(X) \,\big|\, X = x\big] &= \sum_y \big(g(x)y + h(x)\big) f_{Y|X}(y|x) \\
&= g(x)\sum_y y f_{Y|X}(y|x) + h(x)\sum_y f_{Y|X}(y|x) \\
&= g(x)\text{E}[Y|X = x] + h(x)\cdot 1 \\
&= g(x)\text{E}[Y|X = x] + h(x).
\end{aligned}$$

Likewise, if $X$ and $Y$ are jointly continuous, then by Theorem 2.2.11, $\forall x \in \text{Supp}[X]$,

$$\begin{aligned}
\text{E}\big[g(X)Y + h(X) \,\big|\, X = x\big] &= \int_{-\infty}^{\infty} \big(g(x)y + h(x)\big) f_{Y|X}(y|x)dy \\
&= g(x)\int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy + h(x)\int_{-\infty}^{\infty} f_{Y|X}(y|x)dy \\
&= g(x)\text{E}[Y|X = x] + h(x)\cdot 1 \\
&= g(x)\text{E}[Y|X = x] + h(x). \quad \square
\end{aligned}$$

We now turn to a central definition of this book: the *conditional expectation function (CEF)*. The CEF is a function that takes as an input $x$ and returns the conditional expectation of $Y$ given $X = x$. The CEF is extremely useful, since it is a single function that characterizes all possible values of $\text{E}[Y|X = x]$. If we are interested in characterizing the way in which the conditional distribution of $Y$ depends on the value of $X$, the CEF is a natural summary feature of the joint distribution to target. Furthermore, we will see that the CEF can be closely linked to many topics that this book considers, including regression (Chapter 4), missing data (Chapter 6), and causal inference (Chapter 7).

**Definition 2.2.15.** *Conditional Expectation Function (CEF)*
For random variables $X$ and $Y$ with joint PMF/PDF $f$, the *conditional expectation function* of $Y$ given $X = x$ is

$$G_Y(x) = \text{E}[Y|X = x], \forall x \in \text{Supp}[X].$$

A few remarks on notation are in order here. We will generally write $\mathrm{E}[Y|X=x]$ to denote the CEF rather than $G_Y(x)$. The above definition is merely meant to emphasize that, when we use the term CEF, we are referring to the *function* that maps $x$ to $\mathrm{E}[Y|X=x]$, rather than the value of $\mathrm{E}[Y|X=x]$ at some specific $x$. It is also intended to clarify that the CEF, $\mathrm{E}[Y|X=x]$, is a univariate function of $x$. It is *not* a function of the random variable $Y$. So, for example, if $X$ is a Bernoulli random variable, then the CEF of $Y$ given $X$ is

$$G_Y(x)=\mathrm{E}[Y|X=x]=\begin{cases}\mathrm{E}[Y|X=0] & : \quad x=0\\ \mathrm{E}[Y|X=1] & : \quad x=1.\end{cases}$$

$G_Y(X)$ is a function of the random variable $X$ and is therefore itself a random variable[14] whose value depends on the value of $X$. That is, when $X$ takes on the value $x$, the random variable $G_Y(X)$ takes on the value $G_Y(x)=\mathrm{E}[Y|X=x]$.

We write $\mathrm{E}[Y|X]$ to denote $G_Y(X)$ (since $\mathrm{E}[Y|X=X]$ would be confusing). So, for example, if $X$ is a Bernoulli random variable with $p=\frac{1}{2}$, then $\mathrm{E}[Y|X]$ is a random variable that takes on the value of $\mathrm{E}[Y|X=0]$ with probability $\frac{1}{2}$ and the value of $\mathrm{E}[Y|X=1]$ with probability $\frac{1}{2}$. Also note that we can analogously define the *conditional variance function*, $H_Y(x)=\mathrm{V}[Y|X=x]$, which we will generally write as $\mathrm{V}[Y|X=x]$, and write $\mathrm{V}[Y|X]$ to denote the random variable $H_Y(X)$.

Theorem 1.3.2 (*Equality of Functions of a Random Variable*) implies that we can write statements about conditional expectation functions in the more compact $\mathrm{E}[Y|X]$ form. For example, Theorem 2.2.14 can equivalently be stated as follows: for random variables $X$ and $Y$, if $g$ and $h$ are functions of $X$, then

$$\mathrm{E}\big[g(X)Y+h(X)\,\big|\,X\big]=g(X)\mathrm{E}[Y|X]+h(X).$$

From this point on, we will generally state theorems involving conditional expectation functions using this more concise notation.

We now illustrate these concepts with our familiar example of flipping a coin and rolling a four- or six-sided die.

**Example 2.2.16.** *Flipping a Coin and Rolling a Die*
Consider, again, the generative process from Example 1.1.11. Recall from Example 1.3.5 that the conditional PMF of $Y$ given $X=x$ is

$$f_{Y|X}(y|x)=\begin{cases}\frac{1}{4} & : \quad x=0, y\in\{1,2,3,4\}\\ \frac{1}{6} & : \quad x=1, y\in\{1,2,3,4,5,6\}\\ 0 & : \quad \text{otherwise.}\end{cases}$$

[14] Ignoring some unimportant measure-theoretic regularity conditions, namely $X(\Omega)=\mathrm{Supp}[X]$.

Thus, the CEF of $Y$ given $X = x$ is

$$\begin{aligned} \mathrm{E}[Y|X=x] &= \sum_y y f_{Y|X}(y|x) \\ &= \begin{cases} \sum_{y=1}^{4} y \cdot \frac{1}{4} & : \quad x = 0 \\ \sum_{y=1}^{6} y \cdot \frac{1}{6} & : \quad x = 1 \end{cases} \\ &= \begin{cases} \frac{5}{2} & : \quad x = 0 \\ \frac{7}{2} & : \quad x = 1. \end{cases} \end{aligned}$$

Likewise, the conditional PMF of $X$ given $Y = y$ is

$$f_{X|Y}(x|y) = \begin{cases} \frac{3}{5} & : \quad x = 0, y \in \{1,2,3,4\} \\ \frac{2}{5} & : \quad x = 1, y \in \{1,2,3,4\} \\ 1 & : \quad x = 1, y \in \{5,6\} \\ 0 & : \quad \text{otherwise,} \end{cases}$$

so the CEF of $X$ given $Y = y$ is

$$\begin{aligned} \mathrm{E}[X|Y=y] &= \sum_x x f_{X|Y}(x|y) \\ &= \begin{cases} 0 \cdot \frac{3}{5} + 1 \cdot \frac{2}{5} & : \quad y \in \{1,2,3,4\} \\ 0 \cdot 0 + 1 \cdot 1 & : \quad y \in \{5,6\} \end{cases} \\ &= \begin{cases} \frac{2}{5} & : \quad y \in \{1,2,3,4\} \\ 1 & : \quad y \in \{5,6\}. \end{cases} \qquad \triangle \end{aligned}$$

We can now state one of the most important theorems in this book: the *Law of Iterated Expectations*. The Law of Iterated Expectations is important because it allows us to move between conditional expectations and unconditional expectations. Very often, conditional expectations are easier to work with—in such cases, we can treat some random variables as fixed, which allows for more tractable calculations. As a result, the Law of Iterated Expectations will be invoked frequently in proofs throughout the remainder of this book, particularly in Part III.

**Theorem 2.2.17.** *Law of Iterated Expectations*
For random variables $X$ and $Y$,

$$\mathrm{E}[Y] = \mathrm{E}\big[\mathrm{E}[Y|X]\big].$$[15]

**Proof:** Let $X$ and $Y$ be either two discrete random variables with joint PMF $f$ or two jointly continuous random variables with joint PDF $f$. If $X$ and $Y$ are discrete, then

$$\begin{aligned}
\mathrm{E}[Y] &= \sum_y y f_Y(y) \\
&= \sum_y y \sum_x f(x,y) \\
&= \sum_x \sum_y y f(x,y) \\
&= \sum_x \sum_y y f_{Y|X}(y|x) f_X(x) \\
&= \sum_x \left( \sum_y y f_{Y|X}(y|x) \right) f_X(x) \\
&= \sum_x \mathrm{E}[Y|X=x] f_X(x) \\
&= \mathrm{E}\big[\mathrm{E}[Y|X]\big].
\end{aligned}$$

Likewise, if $X$ and $Y$ are jointly continuous, then

$$\begin{aligned}
\mathrm{E}[Y] &= \int_{-\infty}^{\infty} y f_Y(y) dy \\
&= \int_{-\infty}^{\infty} y \left( \int_{-\infty}^{\infty} f(x,y) dx \right) dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x,y) dy dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{Y|X}(y|x) f_X(x) dy dx
\end{aligned}$$

[15] Note that $\mathrm{E}[Y|X]$ is a (univariate) function of the random variable $X$, so the outer expectation here takes the expected value of this random variable.

$$= \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy \right) f_X(x)dx$$

$$= \int_{-\infty}^{\infty} \mathrm{E}[Y|X=x] f_X(x)dx$$

$$= \mathrm{E}\big[\mathrm{E}[Y|X]\big]. \ \square$$

Put simply, the Law of Iterated Expectations implies that the unconditional expectation can be represented as a weighted average of conditional expectations, where the weights are proportional to the probability distribution of the variable being conditioned on. The Law of Iterated Expectations also directly yields the *Law of Total Variance*.

**Theorem 2.2.18.** *Law of Total Variance*
For random variables $X$ and $Y$,

$$\mathrm{V}[Y] = \mathrm{E}\big[\mathrm{V}[Y|X]\big] + \mathrm{V}\big[\mathrm{E}[Y|X]\big].$$

**Proof:**

$$\mathrm{V}[Y] = \mathrm{E}\big[Y^2\big] - \mathrm{E}[Y]^2$$

$$= \mathrm{E}\Big[\mathrm{E}\big[Y^2 \,|\, X\big]\Big] - \mathrm{E}\big[\mathrm{E}[Y|X]\big]^2$$

$$= \mathrm{E}\Big[\mathrm{E}\big[Y^2 \,|\, X\big] - \mathrm{E}[Y|X]^2 + \mathrm{E}[Y|X]^2\Big] - \mathrm{E}\big[\mathrm{E}[Y|X]\big]^2$$

$$= \mathrm{E}\Big[\mathrm{V}[Y|X] + \mathrm{E}[Y|X]^2\Big] - \mathrm{E}\big[\mathrm{E}[Y|X]\big]^2$$

$$= \mathrm{E}\big[\mathrm{V}[Y|X]\big] + \Big(\mathrm{E}\big[\mathrm{E}[Y|X]^2\big] - \mathrm{E}\big[\mathrm{E}[Y|X]\big]^2\Big)$$

$$= \mathrm{E}\big[\mathrm{V}[Y|X]\big] + \mathrm{V}\big[\mathrm{E}[Y|X]\big]. \ \square$$

Practically speaking, this theorem allows us to decompose the variability of a random variable $Y$ into the average variability "within" values of $X$ and the variability "across" values of $X$.[16]

[16] The Law of Total Variance is also referred to as the Analysis of Variance (ANOVA) identity or theorem. ANOVA can be extended to a broader range of models, considering different sources of variance, which can be used to test hypotheses about differences across groups. For more details, see, e.g., Wackerly, Mendenhall, and Scheaffer (2008, Ch. 13).

The Law of Iterated Expectations and the Law of Total Variance reveal the following properties of deviations from the CEF, which will be necessary to demonstrate its significance.

**Theorem 2.2.19.** *Properties of Deviations from the CEF*
Let $X$ and $Y$ be random variables and let $\varepsilon = Y - \mathrm{E}[Y|X]$. Then

- $\mathrm{E}[\varepsilon|X] = 0$.
- $\mathrm{E}[\varepsilon] = 0$.
- If $g$ is a function of $X$, then $\mathrm{Cov}\big[g(X), \varepsilon\big] = 0$.
- $\mathrm{V}[\varepsilon|X] = \mathrm{V}[Y|X]$.
- $\mathrm{V}[\varepsilon] = \mathrm{E}\big[\mathrm{V}[Y|X]\big]$.

**Proof:** Noting that $\mathrm{E}[Y|X]$ is a function solely of X and applying Theorem 2.2.14,

$$\mathrm{E}[\varepsilon|X] = \mathrm{E}\big[Y - \mathrm{E}[Y|X]\,\big|\,X\big] = \mathrm{E}[Y|X] - \mathrm{E}[Y|X] = 0.$$

Applying the Law of Iterated Expectations,

$$\mathrm{E}[\varepsilon] = \mathrm{E}\big[\mathrm{E}[\varepsilon|X]\big] = \mathrm{E}[0] = 0.$$

Let $g$ be a function of $X$. Then

$$\begin{aligned}
\mathrm{Cov}\big[g(X), \varepsilon\big] &= \mathrm{E}\big[g(X)\varepsilon\big] - \mathrm{E}\big[g(X)\big]\mathrm{E}[\varepsilon] \\
&= \mathrm{E}\Big[g(X)\big(Y - \mathrm{E}[Y|X]\big)\Big] - \mathrm{E}\big[g(X)\big](0) \\
&= \mathrm{E}\big[g(X)Y - g(X)\mathrm{E}[Y|X]\big] \\
&= \mathrm{E}\big[g(X)Y\big] - \mathrm{E}\big[g(X)\mathrm{E}[Y|X]\big] \\
&= \mathrm{E}\big[g(X)Y\big] - \mathrm{E}\Big[\mathrm{E}\big[g(X)Y\,\big|\,X\big]\Big] \\
&= \mathrm{E}\big[g(X)Y\big] - \mathrm{E}\big[g(X)Y\big] \\
&= 0.
\end{aligned}$$

Recalling the definition of conditional variance (Definition 2.2.12),

$$\begin{aligned}
\mathrm{V}[\varepsilon|X] &= \mathrm{E}\Big[\big(\varepsilon - \mathrm{E}[\varepsilon|X]\big)^2\,\Big|\,X\Big] \\
&= \mathrm{E}\big[(\varepsilon - 0)^2\,\big|\,X\big]
\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{E}\big[\varepsilon^2 \,\big|\, X\big] \\
&= \mathrm{E}\Big[\big(Y - \mathrm{E}[Y|X]\big)^2 \,\Big|\, X\Big] \\
&= \mathrm{V}[Y|X].
\end{aligned}
$$

Finally, by the Law of Total Variance,

$$
\begin{aligned}
\mathrm{V}[\varepsilon] &= \mathrm{E}\big[\mathrm{V}[\varepsilon \,|\, X]\big] + \mathrm{V}\big[\mathrm{E}[\varepsilon \,|\, X]\big] \\
&= \mathrm{E}\big[\mathrm{V}[\varepsilon \,|\, X]\big] + \mathrm{V}[0] \\
&= \mathrm{E}\big[\mathrm{V}[\varepsilon \,|\, X]\big] \\
&= \mathrm{E}\big[\mathrm{V}[Y|X]\big]. \quad \square
\end{aligned}
$$

This result allows us to derive a key property of the CEF in the following section.

### 2.2.4 Best Predictors and Best Linear Predictors

The CEF is also of particular importance as a way to summarize the relationship between two random variables because it affords us some theoretical guarantees without having to impose further assumptions.

Suppose we knew the full joint cumulative distribution function (CDF) of $X$ and $Y$, and then someone gave us a randomly drawn value of $X$. What would be the guess of $Y$ that would have the lowest MSE? Formally, what function $g$ of $X$ minimizes $\mathrm{E}[(Y - g(X))^2]$? The answer is given by the CEF. The function $g(X)$ that best approximates $Y$ is the CEF.

**Theorem 2.2.20.** *The CEF is the Best Predictor*
For random variables $X$ and $Y$, the CEF, $\mathrm{E}[Y|X]$, is the best (minimum MSE) predictor of $Y$ given $X$.

**Proof:** Choose any $g(X)$ to approximate $Y$. Let $U = Y - g(X)$. By the definition of minimum MSE, our goal is to choose $g(X)$ to minimize $\mathrm{E}[U^2]$. Let $\varepsilon = Y - \mathrm{E}[Y|X]$ and $W = \mathrm{E}[Y|X] - g(X)$, so that $U = \varepsilon + W$. ($W$ is a function of $X$, so we can treat it like a constant in expectations conditioned on $X$.) Our goal is then to show that, by choosing $g(X) = \mathrm{E}[Y|X]$, we will minimize $\mathrm{E}[U^2]$. Now,

$$
\begin{aligned}
\mathrm{E}\big[U^2 \,\big|\, X\big] &= \mathrm{E}\big[(\varepsilon + W)^2 \,\big|\, X\big] \\
&= \mathrm{E}\big[\varepsilon^2 + 2\varepsilon W + W^2 \,\big|\, X\big]
\end{aligned}
$$

$$
\begin{aligned}
&= \mathrm{E}\big[\varepsilon^2 \,\big|\, X\big] + 2W\mathrm{E}[\varepsilon \,|\, X] + W^2 \\
&= \mathrm{E}\big[\varepsilon^2 \,\big|\, X\big] + 0 + W^2 \\
&= \mathrm{E}\big[\varepsilon^2 \,\big|\, X\big] - \mathrm{E}[\varepsilon \,|\, X]^2 + W^2 \\
&= \mathrm{V}[Y|X] + W^2,
\end{aligned}
$$

where the third and fifth lines follow from Theorem 2.2.19. Applying the Law of Iterated Expectations,

$$\mathrm{E}\big[U^2\big] = \mathrm{E}\Big[\mathrm{E}\big[U^2 \,\big|\, X\big]\Big] = \mathrm{E}\big[\mathrm{V}[Y|X] + W^2\big] = \mathrm{E}\big[\mathrm{V}[Y|X]\big] + \mathrm{E}\big[W^2\big].$$

$\mathrm{E}[\mathrm{V}[Y|X]]$ does not depend on the choice of $g(X)$. Additionally, $\mathrm{E}[W^2] \geq 0$, with equality if $g(X) = \mathrm{E}[Y|X]$. Therefore, choosing $g(X) = \mathrm{E}[Y|X]$ minimizes MSE. □

Thus, the CEF yields the best (minimum MSE) approximation of $Y$, conditional on the observed value of $X$. There is no better way (in terms of MSE) to approximate $Y$ given $X$ than the CEF. This makes the CEF a natural target of inquiry: if the CEF is known, much is known about how $X$ relates to $Y$. However, although it is the best predictor, the CEF can be extremely complicated—without further assumptions, the function can take any shape.

What if we were to restrict ourselves to just linear functions? Among functions of the form $g(X) = a + bX$, what function yields the best prediction of $Y$ given $X$? By choosing the $a$ and $b$ that minimize MSE, we obtain the *best linear predictor (BLP)* of $Y$ given $X$.[17] This naturally yields a much simpler target that retains much of the same interpretation. The BLP will have great importance for our discussion in Chapter 4 (*Regression*).

**Theorem 2.2.21.** *Best Linear Predictor (BLP)*
For random variables $X$ and $Y$, if $\mathrm{V}[X] > 0$,[18] then the best (minimum MSE) linear predictor of $Y$ given $X$ is $g(X) = \alpha + \beta X$, where

$$
\begin{aligned}
\alpha &= \mathrm{E}[Y] - \frac{\mathrm{Cov}[X,Y]}{\mathrm{V}[X]}\mathrm{E}[X], \\
\beta &= \frac{\mathrm{Cov}[X,Y]}{\mathrm{V}[X]}.
\end{aligned}
$$

[17] As with the CEF, the BLP is a univariate function of $x$, not a function of the random variable $Y$. Consequently, the function that provides us with the best linear prediction of $Y$ given $X$, when inverted, generally will not provide the best linear prediction of $X$ given $Y$.

[18] If $\mathrm{V}[X] = 0$, then any function such that $g(\mathrm{E}[X]) = \mathrm{E}[Y]$ would minimize MSE. Thus, unlike the CEF, the BLP is not necessarily uniquely defined.

**Proof:** Let $\varepsilon = Y - (a + bX)$. If $g(X) = \alpha + \beta X$ is the best linear predictor of $Y$ given $X$, then

$$(\alpha, \beta) = \underset{(a,b)\in\mathbb{R}^2}{\arg\min}\, \mathrm{E}\big[\varepsilon^2\big].$$

The first-order conditions yield the system of equations

$$\begin{aligned} 0 &= \frac{\partial \mathrm{E}[\varepsilon^2]}{\partial \alpha}, \\ 0 &= \frac{\partial \mathrm{E}[\varepsilon^2]}{\partial \beta}. \end{aligned}$$

By linearity of expectations and the chain rule, as well as the fact that derivatives pass through expectations (because they pass through sums and integrals), this becomes:

$$\begin{aligned} 0 &= \frac{\partial \mathrm{E}[\varepsilon^2]}{\partial \alpha} = \mathrm{E}\left[\frac{\partial \varepsilon^2}{\partial \alpha}\right] = \mathrm{E}\left[2\varepsilon \frac{\partial \varepsilon}{\partial \alpha}\right] = -2\mathrm{E}[\varepsilon], \\ 0 &= \frac{\partial \mathrm{E}[\varepsilon^2]}{\partial \beta} = \mathrm{E}\left[\frac{\partial \varepsilon^2}{\partial \beta}\right] = \mathrm{E}\left[2\varepsilon \frac{\partial \varepsilon}{\partial \beta}\right] = -2\mathrm{E}[\varepsilon X]. \end{aligned}$$

Now we solve $0 = \mathrm{E}[Y - (\alpha + \beta X)]$ and $0 = \mathrm{E}[(Y - (\alpha + \beta X))X]$.[19] From the first equation,

$$0 = \mathrm{E}\big[Y - (\alpha + \beta X)\big] = \mathrm{E}[Y] - \alpha - \beta \mathrm{E}[X],$$

[19] The second-order partial derivatives are

$$\begin{aligned} \frac{\partial^2 \mathrm{E}[\varepsilon^2]}{\partial \alpha^2} &= \frac{\partial}{\partial \alpha}\big(-2\mathrm{E}[\varepsilon]\big) = -2\mathrm{E}\left[\frac{\partial \varepsilon}{\partial \alpha}\right] = -2\mathrm{E}[-1] = 2, \\ \frac{\partial^2 \mathrm{E}[\varepsilon^2]}{\partial \beta^2} &= \frac{\partial}{\partial \beta}\big(-2\mathrm{E}[\varepsilon \mathbf{X}]\big) = -2\mathrm{E}\left[\frac{\partial \varepsilon}{\partial \beta} X\right] = -2\mathrm{E}\big[-X^2\big] = 2\mathrm{E}\big[X^2\big], \\ \frac{\partial^2 \mathrm{E}[\varepsilon^2]}{\partial \alpha \partial \beta} &= \frac{\partial}{\partial \beta}\big(-2\mathrm{E}[\varepsilon]\big) = -2\mathrm{E}\left[\frac{\partial \varepsilon}{\partial \beta}\right] = -2\mathrm{E}[-X] = 2\mathrm{E}[X]. \end{aligned}$$

So, since

$$\frac{\partial^2 \mathrm{E}[\varepsilon^2]}{\partial \alpha^2} = 2 > 0$$

and

$$\begin{aligned} \frac{\partial^2 \mathrm{E}[\varepsilon^2]}{\partial \alpha^2} \frac{\partial^2 \mathrm{E}[\varepsilon^2]}{\partial \beta^2} - \left(\frac{\partial^2 \mathrm{E}[\varepsilon^2]}{\partial \alpha \partial \beta}\right)^2 &= 4\mathrm{E}\big[X^2\big] - \big(2\mathrm{E}[X]\big)^2 = 4\mathrm{E}\big[X^2\big] - 4\mathrm{E}[X]^2 \\ &= 4\Big(\mathrm{E}\big[X^2\big] - \mathrm{E}[X]^2\Big) = 4\mathrm{V}[X] > 0, \end{aligned}$$

a unique solution will be an absolute minimum.

so $\alpha = \mathrm{E}[Y] - \beta\mathrm{E}[X]$. Then from the second equation,

$$\begin{aligned}
0 &= \mathrm{E}\Big[\big(Y - (\alpha + \beta X)\big)X\Big] \\
&= \mathrm{E}\big[YX - \alpha X - \beta X^2\big] \\
&= \mathrm{E}[XY] - \alpha\mathrm{E}[X] - \beta\mathrm{E}\big[X^2\big] \\
&= \mathrm{E}[XY] - \big(\mathrm{E}[Y] - \beta\mathrm{E}[X]\big)\mathrm{E}[X] - \beta\mathrm{E}\big[X^2\big] \\
&= \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] + \beta\mathrm{E}[X]^2 - \beta\mathrm{E}\big[X^2\big] \\
&= \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] - \beta\Big(\mathrm{E}\big[X^2\big] - \mathrm{E}[X]^2\Big) \\
&= \mathrm{Cov}[X, Y] - \beta\mathrm{V}[X].
\end{aligned}$$

Solving for $\beta$, we obtain

$$\beta = \frac{\mathrm{Cov}[X, Y]}{\mathrm{V}[X]}.$$

Finally, substituting this result back into $\alpha = \mathrm{E}[Y] - \beta\mathrm{E}[X]$ yields

$$\alpha = \mathrm{E}[Y] - \frac{\mathrm{Cov}[X, Y]}{\mathrm{V}[X]}\mathrm{E}[X]. \ \square$$

Note that $\alpha$ is the $y$-intercept of the BLP, and $\beta$ is its slope.[20] We note two important corollaries:

- The BLP is also the best linear approximation of the CEF; setting $a = \alpha$ and $b = \beta$ minimizes

$$\mathrm{E}\Big[\big(\mathrm{E}[Y|X] - (a + bX)\big)^2\Big].$$

- If the CEF is linear, then the CEF is the BLP.

Whereas the CEF might be infinitely complex, the BLP is characterized just by two numbers, $\alpha$ and $\beta$. The BLP is a simple approximation of the CEF, and one that operates on the same principle—find the function that minimizes MSE—but with the further restriction that the function must be linear.[21]

[20] Readers with prior training in statistics or econometrics may recognize the expression for the BLP as resembling the ordinary least squares (OLS) regression solution. This is not an accident. We will explain this similarity in Chapter 4.

[21] The BLP also demonstrates that covariance and correlation indeed measure the *linear* relationship between $X$ and $Y$ (as in Section 2.2.2): the BLP is the only linear function that satisfies $\mathrm{Cov}[X, g(X)] = \mathrm{Cov}[X, Y]$. Additionally, the solution to the BLP demonstrates why $\rho[X, Y] = 1$ if and only if $Y$ is a linear function of $X$ with a positive slope.

The CEF and BLP are very important, as each permits a principled and simple summary of the way in which the best prediction of one variable is related to the value of another variable—this is why we will sometimes refer to $X$ as an *explanatory variable* for $Y$. Additionally, both the CEF and BLP are generalizations of the simple expected value. It is also worth noting that some, but not all, properties of the CEF (Theorem 2.2.19) have analogous properties when speaking of the BLP.

**Theorem 2.2.22.** *Properties of Deviations from the BLP*
Let $X$ and $Y$ be random variables and let $\varepsilon = Y - g(X)$, where $g(X)$ is the BLP. Then

- $\mathrm{E}[\varepsilon] = 0$.
- $\mathrm{E}[X\varepsilon] = 0$.
- $\mathrm{Cov}[X, \varepsilon] = 0$.

We omit the proof of this theorem, but it is useful for understanding how the BLP's properties as a linear approximation yield linear analogs to the conditions met by the CEF. But whereas we were able to establish that $\mathrm{E}[\varepsilon|X] = 0$ for the CEF, this is not generally true for the BLP; instead, we have only the weaker statement that $\mathrm{E}[\varepsilon] = 0$. Perhaps the most commonly referenced of these properties is that $\mathrm{Cov}[X, \varepsilon] = 0$—there is no covariance between $X$ and the deviations.

We now consider an example to show how the CEF and BLP allow us to summarize bivariate relationships.

**Example 2.2.23.** *Plotting the CEF and BLP*
Let $X$ and $Y$ be random variables with $X \sim U(0,1)$ and $Y = 10X^2 + W$, where $W \sim N(0,1)$ and $X \perp\!\!\!\perp W$. We derive the CEF of $Y$ given $X$ as follows:

$$\mathrm{E}[Y|X] = \mathrm{E}\big[10X^2 + W\big|X\big].$$

By linearity of expectations,

$$\mathrm{E}\big[10X^2 + W\big|X\big] = 10\mathrm{E}\big[X^2\big|X\big] + \mathrm{E}[W|X] = 10\mathrm{E}\big[X^2\big|X\big] + 0 = 10X^2.$$

Thus, the CEF of $Y$ given $X$ is $\mathrm{E}[Y|X] = 10X^2$.

We now derive the BLP of $Y$ given $X$. The slope of the BLP is

$$\begin{aligned}\beta &= \frac{\mathrm{Cov}[X,Y]}{\mathrm{V}[X]} \\ &= \frac{\mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]}{\mathrm{E}\big[X^2\big] - \mathrm{E}[X]^2}\end{aligned}$$

$$= \frac{\mathrm{E}\left[X\left(10X^2+W\right)\right]-\mathrm{E}[X]\mathrm{E}\left[10X^2+W\right]}{\mathrm{E}\left[X^2\right]-\mathrm{E}[X]^2}$$

$$= \frac{\mathrm{E}\left[10X^3+XW\right]-\mathrm{E}[X]\mathrm{E}\left[10X^2+W\right]}{\mathrm{E}\left[X^2\right]-\mathrm{E}[X]^2}.$$

By linearity of expectations,

$$\beta = \frac{10\mathrm{E}\left[X^3\right]+\mathrm{E}[XW]-\mathrm{E}[X]\left(10\mathrm{E}\left[X^2\right]+\mathrm{E}[W]\right)}{\mathrm{E}\left[X^2\right]-\mathrm{E}[X]^2}$$

$$= \frac{10\mathrm{E}\left[X^3\right]+\mathrm{E}[XW]-10\mathrm{E}[X]\mathrm{E}\left[X^2\right]-\mathrm{E}[X]\mathrm{E}[W]}{\mathrm{E}\left[X^2\right]-\mathrm{E}[X]^2}$$

$$= \frac{10\left(\mathrm{E}\left[X^3\right]-\mathrm{E}[X]\mathrm{E}\left[X^2\right]\right)+\left(\mathrm{E}[XW]-\mathrm{E}[X]\mathrm{E}[W]\right)}{\mathrm{E}\left[X^2\right]-\mathrm{E}[X]^2}$$

$$= \frac{10\left(\mathrm{E}\left[X^3\right]-\mathrm{E}[X]\mathrm{E}\left[X^2\right]\right)+\mathrm{Cov}[X,W]}{\mathrm{E}\left[X^2\right]-\mathrm{E}[X]^2}.$$

By independence,

$$\beta = \frac{10\left(\mathrm{E}\left[X^3\right]-\mathrm{E}[X]\mathrm{E}\left[X^2\right]\right)+0}{\mathrm{E}\left[X^2\right]-\mathrm{E}[X]^2}$$

$$= \frac{10\left(\mathrm{E}\left[X^3\right]-\mathrm{E}[X]\mathrm{E}\left[X^2\right]\right)}{\mathrm{E}\left[X^2\right]-\mathrm{E}[X]^2}.$$

$X \sim U(0,1)$, so its PDF is $f_X(x)=1$, $\forall x \in [0,1]$, and $f(x)=0$ otherwise, so

$$\beta = \frac{10\left(\int_0^1 \left(x^3\cdot 1\right)dx-\left(\int_0^1 (x\cdot 1)dx\right)\left(\int_0^1 (x^2\cdot 1)dx\right)\right)}{\int_0^1 (x^2\cdot 1)dx-\left(\int_0^1 (x\cdot 1)dx\right)^2}$$

$$= \frac{10\left(\frac{1}{4}x^4\Big|_0^1-\left(\frac{1}{2}x^2\Big|_0^1\right)\left(\frac{1}{3}x^3\Big|_0^1\right)\right)}{\frac{1}{3}x^3\Big|_0^1-\left(\frac{1}{2}x^2\Big|_0^1\right)^2}$$

$$= \frac{10\left(\frac{1}{4}(1-0)-\frac{1}{2}(1-0)\frac{1}{3}(1-0)\right)}{\frac{1}{3}(1-0)-\left(\frac{1}{2}(1-0)\right)^2}$$

$$= \frac{10\left(\frac{1}{4} - \frac{1}{6}\right)}{\frac{1}{3} - \left(\frac{1}{2}\right)^2}$$

$$= \frac{10\left(\frac{1}{12}\right)}{\frac{1}{12}}$$

$$= 10.$$

Finally, the intercept is

$$\alpha = \mathrm{E}[Y] - \beta\mathrm{E}[X]$$

$$= \mathrm{E}\left[10X^2 + W\right] - \beta\mathrm{E}[X].$$

By linearity of expectations,

$$\alpha = 10\mathrm{E}\left[X^2\right] + \mathrm{E}[W] - \beta\mathrm{E}[X]$$

$$= 10\mathrm{E}\left[X^2\right] + 0 - \beta\mathrm{E}[X]$$

$$= 10\int_0^1 (x^2 \cdot 1)dx - 10\int_0^1 (x \cdot 1)dx$$

$$= 10 \cdot \frac{x^3}{3}\bigg|_0^1 - 10 \cdot \frac{x^2}{2}\bigg|_0^1$$

$$= \frac{10}{3} - \frac{10}{2}$$

$$= -\frac{5}{3}.$$

Thus, the BLP of $Y$ given $X$ is $g(X) = -\frac{5}{3} + 10X$. △

Figure 2.2.1 plots the CEF (in black) and BLP (in red). Here, the BLP approximates the CEF reasonably well over the domain of the data. While this is not always the case, it is very often the case in the social and health sciences. The BLP is thus a good "first approximation" in a very literal sense, in that it is an approximation with a first-order polynomial.[22] However, when the CEF is not linear, one must take care in interpreting the BLP as equivalent to the CEF. In particular, this may pose a problem when attempting to make inferences where $X$ has low probability mass over some parts of the domain of the CEF. This is because under nonlinearity, the BLP depends on the distribution of $X$. This is illustrated in the following example.

[22] We will discuss the approximation of the CEF with higher-order polynomials in Section 4.3.1.

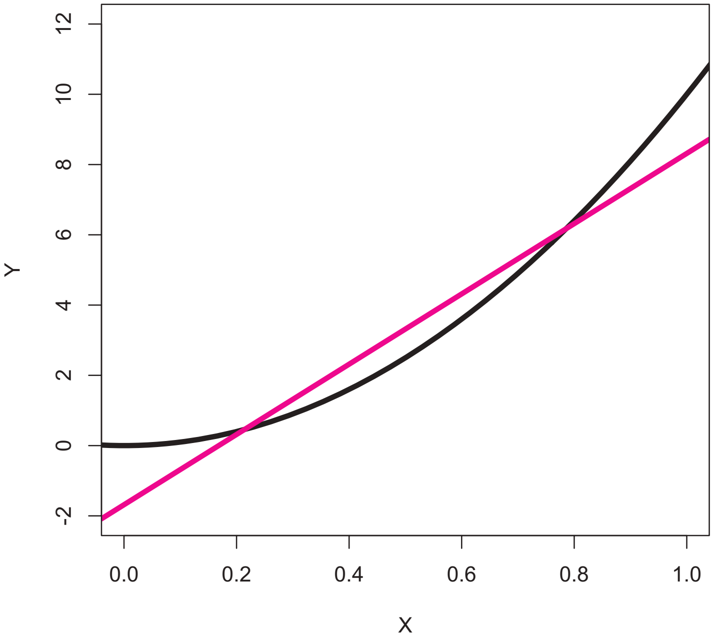

**FIGURE 2.2.1.** *Plotting the CEF and BLP*

**Example 2.2.24.** *BLP under Nonlinearity*
Suppose that, as in Example 2.2.23, $Y = 10X^2 + W$, $W \sim N(0,1)$ and $X \perp\!\!\!\perp W$. If $X$ is uniformly distributed between zero and one, the CEF of $Y$ given $X$ is the same as above: $\mathrm{E}[Y|X] = X^2$. However, we see in panel (a) of Figure 2.2.2 that the BLP approximates the CEF reasonably well only over certain areas. Additionally, if the distribution of $X$ changes, the BLP changes. Figure 2.2.2 demonstrates how, with the same (nonlinear) CEF (in black), the BLP (in red) depends on the distribution of $X$. $\triangle$

### 2.2.5 CEFs and BLPs under Independence

Although this section is very short, it contains results essential for our calculations in Part III. Here, we derive some additional properties of independent random variables as they relate to conditional expectations, the CEF, and the BLP.

**Theorem 2.2.25.** *Implications of Independence (Part III)*
If $X$ and $Y$ are independent random variables, then

- $\mathrm{E}[Y|X] = \mathrm{E}[Y]$.
- $\mathrm{V}[Y|X] = \mathrm{V}[Y]$.

- The BLP of $Y$ given $X$ is $\mathrm{E}[Y]$.
- If $g$ is a function of $X$ and $h$ is a function of $Y$, then
  - $\mathrm{E}\big[g(Y)\,\big|\,h(X)\big] = \mathrm{E}\big[g(Y)\big]$.
  - The BLP of $h(Y)$ given $g(X)$ is $\mathrm{E}\big[h(Y)\big]$.

We omit the proof of Theorem 2.2.25, noting that it directly follows from Theorems 1.3.17 and 2.2.8 (*Implications of Independence, Parts I and II*). These results imply that, when $Y$ and $X$ are independent, the CEF and BLP work as expected: they are both simply equal to $\mathrm{E}[Y]$, regardless of what value $X$ takes on. This is because $X$ provides no information as to

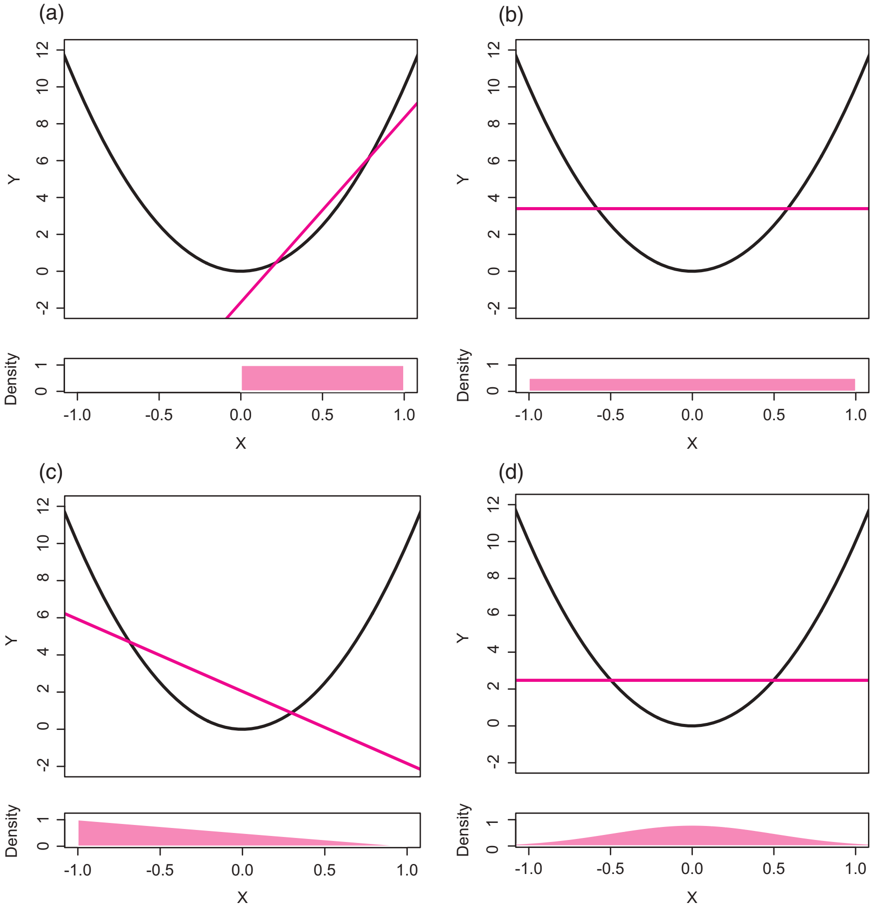

**FIGURE 2.2.2.** *Plotting the CEF and the BLP over Different Distributions of X*

the distribution of $Y$, leaving us simply with $\mathrm{E}[Y]$ as the best predictor of $Y$ (Theorem 2.1.24).

## 2.3 MULTIVARIATE GENERALIZATIONS

As in Chapter 1, we conclude by sketching the generalizations of the concepts presented in this chapter to the multivariate case, with the goal of defining the CEF and BLP when we have more than one explanatory variable. We provide no examples and only one small proof in this section, as our goal is to illustrate that the concepts we have presented in detail in the bivariate case are easily extended. Essentially all omitted definitions, properties, and theorems can be generalized analogously.

The multivariate generalization of variance is the *covariance matrix* of a random vector.

**Definition 2.3.1.** *Covariance Matrix*
For a random vector $\mathbf{X}$ of length $K$, the *covariance matrix* $\mathrm{V}[\mathbf{X}]$ is a matrix whose $(k,k')^{\text{th}}$ entry is $\mathrm{Cov}[X_{[k]},X_{[k']}]$, $\forall i,j \in \{1,2,...,K\}$. That is,

$$\mathrm{V}[\mathbf{X}] = \begin{pmatrix} \mathrm{V}\big[X_{[1]}\big] & \mathrm{Cov}\big[X_{[1]},X_{[2]}\big] & \cdots & \mathrm{Cov}\big[X_{[1]},X_{[K]}\big] \\ \mathrm{Cov}\big[X_{[2]},X_{[1]}\big] & \mathrm{V}\big[X_{[2]}\big] & \cdots & \mathrm{Cov}\big[X_{[2]},X_{[K]}\big] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}\big[X_{[K]},X_{[1]}\big] & \mathrm{Cov}\big[X_{[K]},X_{[2]}\big] & \cdots & \mathrm{V}\big[X_{[K]}\big] \end{pmatrix}.$$

Note that this is a symmetric matrix (because $\mathrm{Cov}[X_{[k]},X_{[k']}] = \mathrm{Cov}[X_{[k']},X_{[k]}]$, by Theorem 2.2.4, *Properties of Covariance*) and the diagonal entries are variances (because $\mathrm{Cov}[X_{[k]},X_{[k]}] = \mathrm{V}[X_{[k]}]$, again by Theorem 2.2.4). This will be how we describe the variability of the regression estimator in Section 4.2.

We can also state a multivariate version of the variance rule (Theorem 2.2.3).

**Theorem 2.3.2.** *Multivariate Variance Rule*
For random variables $X_{[1]}$, $X_{[2]}$, ..., $X_{[K]}$,

$$V\big[X_{[1]} + X_{[2]} + \cdots + X_{[K]}\big] = V\left[\sum_{k=1}^{K} X_{[k]}\right] = \sum_{k=1}^{K}\sum_{k'=1}^{K} \mathrm{Cov}\big[X_{[k]},X_{[k']}\big].$$

As in the bivariate case, note how this (not coincidentally) resembles the algebraic identity

$$(x_1 + x_2 + \dots + x_n)^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j.$$

Theorem 2.3.2 will be key to deriving the variance of the cluster sample mean (see Theorem 3.5.5).

Conditional expectations are defined almost identically in the multivariate case as in the bivariate case (Definition 2.2.10).

**Definition 2.3.3.** *Conditional Expectation (Multivariate Case)*
For discrete random variables $X_{[1]}$, $X_{[2]}$, ..., $X_{[K]}$, and $Y$ with joint PMF $f$, the *conditional expectation* of $Y$ given $\mathbf{X} = \mathbf{x}$ is

$$\mathrm{E}[Y|\mathbf{X} = \mathbf{x}] = \sum_{y} y f_{Y|\mathbf{X}}(y|\mathbf{x}), \forall \mathbf{x} \in \mathrm{Supp}[\mathbf{X}].$$

For jointly continuous random variables $X_{[1]}$, $X_{[2]}$, ..., $X_{[K]}$, and $Y$ with joint PDF $f$, the *conditional expectation* of $Y$ given $\mathbf{X} = \mathbf{x}$ is

$$\mathrm{E}[Y|\mathbf{X} = \mathbf{x}] = \int_{-\infty}^{\infty} y f_{Y|\mathbf{X}}(y|\mathbf{x}) dy, \forall \mathbf{x} \in \mathrm{Supp}[\mathbf{X}].$$

Similarly, the conditional expectation function is analogously defined.

**Definition 2.3.4.** *CEF (Multivariate Case)*
For random variables $X_{[1]}$, $X_{[2]}$, ..., $X_{[K]}$, and $Y$ with joint PMF/PDF $f$, the *CEF* of $Y$ given $\mathbf{X} = \mathbf{x}$ is

$$G_Y(\mathbf{x}) = \mathrm{E}[Y|\mathbf{X} = \mathbf{x}], \forall \mathbf{x} \in \mathrm{Supp}[\mathbf{X}].$$

As before, this definition is merely meant to emphasize that the CEF refers to the *function* that maps $\mathbf{x}$ to $\mathrm{E}[Y|\mathbf{X} = \mathbf{x}]$; in general, we will write $\mathrm{E}[Y|\mathbf{X} = \mathbf{x}]$ to denote the CEF. The same notational conventions noted in the bivariate case apply: $\mathrm{E}[Y|\mathbf{X}]$ denotes $G_Y(\mathbf{X})$, which is a function of the random vector $\mathbf{X}$; the conditional variance function is $H_Y(\mathbf{x}) = \mathrm{V}[Y|\mathbf{X} = \mathbf{x}]$, which is generally written as $\mathrm{V}[Y|\mathbf{X} = \mathbf{x}]$; and $\mathrm{V}[Y|\mathbf{X}]$ denotes $H_Y(\mathbf{X})$.

In the multivariate case, the CEF is still the best (minimum MSE) predictor of $Y$ given $X_{[1]}$, $X_{[2]}$, ..., $X_{[K]}$. That is, suppose we knew the full joint CDF of $X_{[1]}$, $X_{[2]}$, ..., $X_{[K]}$, and $Y$, and then someone gave us a randomly drawn value of $(X_{[1]}, X_{[2]}, \dots, X_{[K]})$. What would be the guess of $Y$ that would have the lowest MSE? Formally, what function $g : \mathbb{R}^K \to \mathbb{R}$ minimizes $\mathrm{E}[(Y - g(X_{[1]}, X_{[2]}, \dots, X_{[K]}))^2]$? Again, the answer is given by the CEF.

**Theorem 2.3.5.** *The CEF Is the Minimum MSE Predictor*
For random variables $X_{[1]}$, $X_{[2]}$, ..., $X_{[K]}$, and $Y$, the CEF, $\mathrm{E}[Y|\mathbf{X}]$, is the best (minimum MSE) predictor of $Y$ given $\mathbf{X}$.

The proof is the same as in the bivariate case.

Finally, we can describe the BLP of $Y$ given $X_{[1]}$, $X_{[2]}$, ..., $X_{[K]}$. As before, the BLP is the *linear* function that minimizes MSE.

**Definition 2.3.6.** *BLP (Multivariate Case)*
For random variables $X_{[1]}$, $X_{[2]}$, ..., $X_{[K]}$, and $Y$, the *best linear predictor* of $Y$ given $\mathbf{X}$ (that is, the minimum MSE predictor of $Y$ given $\mathbf{X}$ among functions of the form $g(\mathbf{X}) = b_0 + b_1X_{[1]} + b_2X_{[2]} ... + b_KX_{[K]}$) is $g(\mathbf{X}) = \beta_0 + \beta_1X_{[1]} + \beta_2X_{[2]} ... + \beta_KX_{[K]}$, where

$$(\beta_0, \beta_1, \beta_2, ..., \beta_K) = \underset{(b_0,b_1,b_2,...,b_K)\in\mathbb{R}^{K+1}}{\arg\min} \mathrm{E}\left[\left(Y - \left(b_0 + b_1X_{[1]} + b_2X_{[2]} ... + b_KX_{[K]}\right)\right)^2\right].$$

That is, among functions of the form $g(\mathbf{X}) = b_0 + b_1X_{[1]} + ... + b_KX_{[K]}$, what function yields the best (minimum MSE) prediction of $Y$ given $X_{[1]}$, $X_{[2]}$, ..., $X_{[K]}$? Formally, what values of $b_0$, $b_1$, $b_2$, ..., $b_K$ minimize $\mathrm{E}[(Y - (b_0 + b_1X_{[1]} + b_2X_{[2]} + ... + b_KX_{[K]}))^2]$? In the multivariate case, we no longer have the nice, simple formulas $\alpha = \mathrm{E}[Y] - \frac{\mathrm{Cov}[X,Y]}{\mathrm{V}[X]}\mathrm{E}[X]$ and $\beta = \frac{\mathrm{Cov}[X,Y]}{\mathrm{V}[X]}$ for the solution (Theorem 2.2.21). Instead, the full solution is most simply represented using matrix algebra. We shall return to this in Chapter 4.

The BLP has many of the same properties in the multivariate case that it had in the bivariate case. As before:

- The BLP is also the best linear approximation of the CEF; setting $(b_0, b_1, b_2, ..., b_K) = (\beta_0, \beta_1, \beta_2, ..., \beta_K)$ minimizes

$$\mathrm{E}\left[\left(\mathrm{E}[Y|\mathbf{X}] - \left(b_0 + b_1X_{[1]} + b_2X_{[2]} + ... + b_KX_{[K]}\right)\right)^2\right].$$

  The proof is obtained by the same means as above.
- If the CEF is linear, then the CEF is the BLP.

One remarkable feature of the BLP is that it facilitates easy interpretation. The first coefficient, $\beta_0$, is the intercept (sometimes known as the constant), and represents the value the BLP would take on if all the variables were held at zero; that is, $g(0, 0, ..., 0) = \beta_0$. To see this, we need only note that $g(0, 0, ..., 0) = \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 0 + ... + \beta_K \cdot 0 = \beta_0$.

The remaining coefficients of the BLP represent partial derivatives. For example, $\beta_1$ represents how much the BLP of $Y$ would change if we moved one unit on $X_{[1]}$, holding all else equal. This can be demonstrated by considering a set of arbitrary values $(x_{[1]}, x_{[2]}, ..., x_{[K]})$, so that $g(x_{[1]}, x_{[2]}, ..., x_{[K]}) = \beta_0 + \beta_1 x_{[1]} + \beta_2 x_{[2]} ... + \beta_K x_{[k]}$, and $g(x_{[1]} + 1, x_{[2]}, ..., x_{[K]}) = \beta_0 + \beta_1(x_{[1]} + 1) + \beta_2 x_{[2]} ... + \beta_K x_{[k]}$. Therefore, $g(x_{[1]} + 1, x_{[2]}, ..., x_{[K]}) - g(x_{[1]}, x_{[2]}, ..., x_{[K]}) = \beta_1$.

In other words, $\beta_1$ is the slope of the BLP with respect to $X_{[1]}$, *conditional* on the values of all the other variables. If we held $X_{[2]}$, $X_{[3]}$, ..., $X_{[K]}$ fixed at some numbers, then how would the BLP change by moving $X_{[1]}$ by one? The same intuition holds for $\beta_2$, $\beta_3$, and so on. We formalize this in the following theorem.

**Theorem 2.3.7.** *Coefficients of the BLP Are Partial Derivatives*
For random variables $X_{[1]}$, $X_{[2]}$, ..., $X_{[K]}$, and $Y$, if $g(\mathbf{X})$ is the best linear predictor of $Y$ given $\mathbf{X}$, then $\forall k \in \{1, 2, ..., K\}$,

$$\beta_k = \frac{\partial g(\mathbf{X})}{\partial X_{[k]}}.$$

**Proof:** This follows immediately from linearity:

$$\frac{\partial g\big(X_{[1]}, X_{[2]}, ..., X_{[K]}\big)}{\partial X_{[k]}} = \frac{\partial\big(\beta_0 + \beta_1 X_{[1]} + \beta_2 X_{[2]} + ...\beta_K X_{[K]}\big)}{\partial X_{[k]}} = \beta_k. \ \square$$

When the CEF is well approximated by the BLP, these properties of the BLP afford us a simple way of describing features of the CEF.

Finally, we can state the multivariate generalization of Theorem 2.2.22 (*Properties of Deviations from the BLP*).

**Theorem 2.3.8.** *Properties of Deviations from the BLP (Multivariate Case)*
For random variables $X_{[1]}$, $X_{[2]}$, ..., $X_{[K]}$, and $Y$, if $g(\mathbf{X})$ is the best linear predictor of $Y$ given $\mathbf{X}$ and $\varepsilon = Y - g(\mathbf{X})$, then

- $\mathrm{E}[\varepsilon] = 0$.
- $\forall k \in \{1, 2, ..., K\}$, $\mathrm{E}\big[X_{[k]}\varepsilon\big] = 0$.
- $\forall k \in \{1, 2, ..., K\}$, $\mathrm{Cov}\big[X_{[k]}, \varepsilon\big] = 0$.

In Chapter 4, we will see that regression has analogous properties (Theorem 4.1.5, *Properties of Residuals from the OLS Regression Estimator*).

## 2.4 FURTHER READINGS

The further readings listed in Chapter 1 remain relevant here. Beyond these, we highly recommend Goldberger (1991) and Hansen (2013), which served as key inspiration for our discussion of the topics in Chapter 2. In particular, readers well versed in linear algebra may appreciate Hansen's representation of the BLP using projections. Angrist and Pischke (2009) also contains a concise but highly readable treatment of the topics here.