BUSS975 Causal Inference in Financial Research

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

Chapter 3

Statistical Foundations Review

3.1 Expectation and Moments of Random Variables

In the previous chapter, we focused on probability distributions as the fundamental characterization of random variables. We saw that knowing a random variable's cumulative distribution function (CDF) or probability density/mass function (pdf/pmf) gives a complete description of its behavior. However, in practice we often do not need the *entire* distribution of a random variable. Instead, we are typically interested in certain key features or summary measures of the distribution, especially in applications like causal inference or financial analysis. For example, recall the returns-to-education study where we defined the average treatment effect on the treated (ATT) as:

$$\tau_{ATT} = E[Y_i(1) - Y_i(0) \mid D_i = 1],$$

the expected difference in potential outcomes for those who attended college ($D_i = 1$). Here $Y_i(1)$ and $Y_i(0)$ are potential wages with and without college, and D_i indicates college attendance. In this causal estimand, we care only about an *expected value* (a single number summarizing the average causal effect for college-goers) rather than the full distribution of $Y_i(1) - Y_i(0)$. This illustrates a common scenario: we often seek the "center" or other moments of a distribution as opposed to a full description of all probabilities.

In this chapter, we introduce the concept of **expectation** (also called the *mean* or *first moment*) and related measures like variance, covariance, and correlation, which together summarize important features of probability distributions. We will also discuss their conditional counterparts (conditional expectation and variance) which characterize distributions under given information. Finally, we introduce the idea of *mean independence*, a weaker notion than full statistical independence that is particularly relevant in econometric contexts (for instance, when discussing exogeneity of regressors).

3.2 Features of a Probability Distribution

3.2.1 Expectation

Definition 3.1 (Expected Value)**.** The *expectation* or *expected value* of a random variable X (with respect to its own distribution) is defined as a weighted average of its possible values, using probabilities or densities as weights. Formally:

$$E[X] = \begin{cases} \sum_{x \in \text{supp } X} x\, f_X(x), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} x\, f_X(x)\, dx, & \text{if } X \text{ is continuous,} \end{cases}$$

provided these sums or integrals converge absolutely. Here $f_X(x)$ denotes the pmf or pdf of X, and supp(X) is the support (set of values with nonzero probability or density). We will also use the notation $E_X[X]$ or $\mathbb{E}[X]$ interchangeably for $E[X]$. The expected value $E[X]$ is a single number (a constant, not a random variable) that intuitively represents the "center of gravity" or long-run average of X's distribution.

If $E[|X|] < \infty$, we say the expectation of X *exists* (is well-defined and finite). In this course, we will always assume that the random variables we deal with have well-defined, finite expectations.

The expectation $E[X]$ is a weighted average of the possible values of X, weighted by their probability or density. It is a single number that describes the center of the distribution of X. For example, if X is the number of heads in 2 fair coin tosses, we found $E[X] = 0 \cdot (1/4) + 1 \cdot (1/2) + 2 \cdot (1/4) = 1$. If $Y \sim U(0,1)$, then

$$E[Y] = \int_0^1 y\,(1)dy = \frac{1}{2}.$$

In general for $U \sim U(a,b)$, $E[U] = \frac{a+b}{2}$, the midpoint of the interval.

Expectations are very convenient to work with algebraically, because they have nice linear properties. One does not need the full distribution of X to evaluate many expectations; often it's easier to use algebraic rules or known formulas.

Theorem 3.2 (Linearity of Expectation)**.** *For any random variable X and constants $a, b \in \mathbb{R}$,*

$$E[a + bX] = a + b\,E[X].$$

More generally, if $X_1, \ldots, X_n$ are random variables (not necessarily independent), then for any constants $b_1, \ldots, b_n$,

$$E\Big[\sum_{i=1}^{n} b_i X_i\Big] = \sum_{i=1}^{n} b_i\,E[X_i].$$

Linearity of expectation is extremely useful. It means, for example, we can calculate the expected sum of 1000 random quantities without computing a convolution of 1000 distributions—we can just sum their individual expectations. No matter what dependence or correlation might exist between X_i and X_j, the linearity property $E[X_i + X_j] = E[X_i] + E[X_j]$ *always* holds. As a simple but important special case: if X and Y are independent, $E[X + Y] = E[X] + E[Y]$. (Again, this holds even without independence.)

The Law of the Unconscious Statistician (LOTUS)

Often we are interested not directly in X itself, but in some function of X. For example, if X represents a return on investment, we might be interested in $Y = g(X)$ which could be a nonlinear function like utility $g(X) = \ln(1+X)$ or an indicator of whether the return exceeds a threshold $g(X) = \mathbf{1}X > c$. The distribution of Y can be complicated to derive from that of X. Fortunately, there is a shortcut to finding $E[Y] = E[g(X)]$ without explicitly finding Y's distribution. This is provided by an important result humorously named the **Law of the Unconscious Statistician (LOTUS)**

Theorem 3.3 (Law of the Unconscious Statistician)**.** *Let X be a random variable with known distribution, and let $Y = g(X)$ be some function of X. Then, as long as $E[|g(X)|]$ exists and is finite, we can compute the expectation of Y by integrating (or summing) over the distribution of X:*

$$E[Y] = E[h(X)] = \begin{cases} \sum_{x \in supp\ X} g(x)\, f_X(x), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx, & \text{if } X \text{ is continuous,} \end{cases}$$

provided the sum/integral converges absolutely.

In other words, to find $E[g(X)]$ you do *not* need to determine the distribution of $Y = g(X)$ first. You can simply compute the weighted average of $g(x)$ using the original distribution of X. This greatly simplifies working with transformations of random variables.

Why the whimsical name "unconscious statistician"? Because many students (and practitioners) use this law instinctively — plugging $g(x)$ into the expectation integral of X — without realizing they are invoking a theorem. In fact, many textbooks simply treat the LOTUS formula as the *definition* of $E[g(X)]$. But formally, the definition of expectation is in terms of Y's distribution, and LOTUS is a derived result that lets us avoid the intermediate step of finding f_Y. The nickname is a reminder that while it may feel like common sense, one should be conscious that this step is justified by a proven result.

Example 3.4. (Indicator function) A very useful special case of LOTUS is when $g(x)$ is an indicator of an event. Let $\mathcal{A}$ be some event (a subset of the real line, in the context of X). Consider $Y = \mathbf{1}\{X \in \mathcal{A}\}$, a random variable that equals 1 if X falls in $\mathcal{A}$ and 0 otherwise. Then by LOTUS,

$$E[\mathbf{1}\{X \in \mathcal{A}\}] \;=\; \int_{-\infty}^{\infty} \mathbf{1}\{x \in \mathcal{A}\}\, f_X(x)\, dx \;=\; \int_{\mathcal{A}} f_X(x)\, dx \;=\; P(X \in \mathcal{A})\,.$$

But the left-hand side is also $P(X \in \mathcal{A})$ by the definition of expectation of an indicator (since an indicator's expectation *is* the probability of the event). This simple equality $E[\mathbf{1}\{X \in A\}] = P(X \in A)$ holds for any event A. Although trivial seeming, it is extremely handy: it means that probabilities can often be manipulated or derived by treating them as expectations of indicator variables and then using linearity or other expectation rules. We will use this trick frequently in derivations.

Example 3.5. Suppose X is a continuous random variable and we define a new random variable $Y = X^2$. Finding the distribution of Y might involve some effort (especially if X's distribution is complicated). However, if we only want $E[Y] = E[X^2]$, LOTUS tells us we can bypass finding f_Y. We can compute

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x)dx,$$

straight from X's density. For instance, if $X \sim$ Uniform$(-1, 1)$, then by symmetry we know $E[X] = 0$ but $E[X^2]$ is not zero. We can find it by integrating:

$$E[X^2] = \int_{-1}^{1} x^2 \frac{1}{2} dx = \frac{1}{3}.$$

We never explicitly derived $Y = X^2$'s distribution (which in fact is Beta$(1/2, 1/2)$ shape on $[0, 1]$), and we didn't need to. As promised, LOTUS saved us work.

3.2.2 Variance and Standard Deviation

While the expectation $E[X]$ captures the central tendency of a random variable, it does not tell the whole story about the distribution. Two very different distributions can have the same mean. A natural next question: how spread out is the distribution around that mean? This brings us to the **variance**.

Definition 3.6 (Variance and Standard Deviation)**.** The *variance* of a random variable X (with finite mean $E[X] = \mu_X$) is defined as

$$Var(X) = E\Big[(X - E[X])^2\Big].$$

In words, $Var(X)$ is the expected squared deviation of X from its own mean. The *standard deviation* of X is defined as $sd(X) = \sqrt{Var(X)}$ the (positive) square root of the variance. Standard deviation has the advantage of being in the same units as X, making it easier to interpret in context.

The variance $Var(X)$ is a nonnegative number (in fact, variance is zero if and only if X is almost surely constant). It quantifies the variability or dispersion of the distribution of X around its mean. A large variance means X tends to take values far from the mean (high spread), whereas a small variance means X is tightly clustered around its mean. By expanding the definition $(X - \mu_X)^2 = X^2 - 2\mu_X X + \mu_X^2$ and taking expectation, one can show an equivalent formula:

$$Var(X) = E[X^2] - (E[X])^2.$$

This is often a more convenient formula for calculations: the variance is the difference between the second moment $E[X^2]$ and the square of the first moment. Be careful to never confuse $E[X^2]$ (which is $E[X^2]$) with $(E[X])^2$ (the square of $E[X]$); the latter is typically smaller unless the distribution has no spread.

Example 3.7. Coin Tosses (Variance). Revisit the experiment of tossing a fair coin twice with X = number of heads, where we found $E[X] = 1$. We can compute the variance:

$$E[X^2] - (E(X))^2 = (0^2(1/4) + 1^2(1/2) + 2^2(1/4)) - 1 = 0.5$$

The standard deviation is $\sqrt{0.5} \approx 0.707$. This measures the typical deviation from the mean (which was 1) in the number of heads: roughly speaking, about \$0.707 heads off. (Of course, in reality you can only be 0 or 1 heads away from the mean of 1, but the standard deviation gives a kind of "root mean square" deviation.)

It's worth noting that X in this example follows a Binomial$(n = 2, p = 0.5)$ distribution, for which general formulas give $E[X] = np = 1$ and $Var(X) = np(1 - p) = 2 \cdot 0.5 \cdot 0.5 = 0.5$, consistent with our calculation.

Example 3.8. If X is Bernoulli(p) (taking value 1 with probability p and 0 with probability $1 - p$), then $E[X] = p$ and

$$E[X^2] - (E(X))^2 = p - p^2 = p(1 - p)$$

For instance, a fair coin toss ($p = 0.5$) has mean 0.5 and variance 0.25.

Linear change of variables: Variance has a simple behavior under linear transformations, albeit not as simple as expectation (which was exactly linear). If $Y = a + bX$, then $E[Y] = a + bE[X]$ as we know. What about $Var(Y)$? We find

$$Var(Y) = E[(a + bX - (a + bE[X]))^2] = b^2 Var(X).$$

So adding a constant a does not change the variance at all (shifting does nothing to spread), and scaling by b multiplies the variance by b^2. In particular, if you change the units of measurement of X (say from dollars to cents, where $b = 100$), the variance is scaled by the square of that conversion factor (so dollars → cents multiplies variance by 10000!). This reminds us that variance's units are the square of the original units, which is why we often prefer to talk about standard deviation (back in original units). Standard deviation of $a+bX$ would be $|b|\ \mathrm{sd}(X)$ since we take the positive square root of b^2.

3.2.3 Covariance

So far, we have introduced two fundamental descriptors of a single random variable's distribution: the mean and variance (or standard deviation). However, in research we often deal with multiple random variables at once and are interested in how they relate to each other. A key measure of joint variability between two random variables is the **covariance**. Covariance extends the idea of variance (which was for one variable) to two variables:

Definition 3.9 (Covariance)**.** For random variables X and Y with means $E[X] = \mu_X$ and $E[Y] = \mu_Y$, the *covariance* between X and Y is defined as

$$Cov(X, Y) = E[](X - \mu_X)(Y - \mu_Y)].$$

Equivalently (by expanding the product),

$$Cov(X, Y) = E[XY] - E[X]E[Y].$$

Covariance measures the extent to which X and Y "co-vary" or move together. If X and Y tend to be above their means at the same time (and below their means at the same time), the covariance will be positive. If one tends to be above its mean when the other is below its mean (and vice versa), the covariance will be negative. If knowing that X is above/below its mean gives no clue about whether Y is above or below its mean, then the covariance will be around zero.

Important properties of covariance: $-\,Cov(X, Y) = Cov(Y, X)$ (symmetric in its arguments). $-\,Cov(X, X) = Var(X)$. $-$ If a, b are constants, then $Cov(a+$

$bX,\ Y) = b\,Cov(X,Y)$ (linearity in each argument). Similarly $Cov(X,\ c + dY) = d\,Cov(X,Y)$. In particular, shifting one variable by a constant does not change the covariance. – $Cov(X,Y)$ can be positive, negative, or zero. In principle it could be large or small in magnitude depending on the units and scales of X and Y.

Example 3.10. (Discrete joint distribution) Suppose X and Y are two discrete random variables with the joint distribution given by the following table:

	$Y = 0$	$Y = 1$	Total $P(X = x)$
$X = 0$	0.20	0.10	0.30
$X = 1$	0.30	0.40	0.70
Total $P(Y = y)$	0.50	0.50	1

Here $P(X = 0, Y = 0) = 0.20$, $P(X = 0, Y = 1) = 0.10$, $P(X = 1, Y = 0) = 0.30$, $P(X = 1, Y = 1) = 0.40$. From the table, the marginals are $P(X = 1) = 0.7$ (so $P(X = 0) = 0.3$) and $P(Y = 1) = 0.5$ (so $P(Y = 0) = 0.5$). Let's compute means first: $E[X] = 0(0.3) + 1(0.7) = 0.7$ and $E[Y] = 0(0.5) + 1(0.5) = 0.5$. Next, $E[XY] = \sum_{(x,y)} xyP(X = x, Y = y) = 0.40$. Now we plug into the covariance formula: $E[XY] - E[X]E[Y] = 0.05$. So the covariance is $+0.05$ in these units. This is a small positive number, indicating a slight tendency for X and Y to be high or low together. Indeed, if we examine the joint probabilities: X and Y are both 1 with probability 0.4, which is a bit higher than one might expect under independence (which would have been $0.7 \times 0.5 = 0.35$). That is the reason covariance came out positive.

Independence and Covariance: If X and Y are independent random variables, then $E[XY] = E[X]\,E[Y]$ (because the joint pdf/pmf factorizes), and thus $Cov(X,Y) = 0$. So independence $\Longrightarrow$ zero covariance. However, the converse is *not* true in general: $Cov(X,Y) = 0$ does not guarantee independence. Zero covariance means X and Y are *uncorrelated*, but they could still have a nonlinear relationship. A classic counterexample is: take X uniform on $[-1,1]$ and let $Y = X^2$. Then $E[X] = 0$ (symmetric distribution) and $E[Y] = E[X^2] > 0$. We find $E[XY] = E[X \cdot X^2] = E[X^3] = 0$ (again by symmetry, since X^3 is an odd function). Therefore $Cov(X,Y) = 0$ –

$(0)(E[Y]) = 0$. Yet X and Y are clearly *not* independent: knowing Y (which is X^2) tells us X is either the positive or negative square root of Y rather than any value in $[-1, 1]$; in fact Y is completely determined by X (functional dependence). This example shows how two variables can have no linear association (zero covariance) but still be strongly related in a nonlinear way.

Variance of a sum: Covariance provides a convenient formula for the variance of the sum of two random variables:

$$Var(X+Y) = E[(X+Y-E[X]-E[Y])^2] = E[(X-E[X]+Y-E[Y])^2].$$

Expanding the square yields

$$Var(X+Y) = E[(X-\mu_X)^2] + 2E[(X-\mu_X)(Y-\mu_Y)] + E[(Y-\mu_Y)^2],$$

which simplifies to

$$Var(X+Y) = Var(X) + 2Cov(X,Y) + Var(Y).$$

In short, the variance of a sum is the sum of variances plus twice the covariance. This formula generalizes: for any two random variables,

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y).$$

If X and Y are *independent*, then $Cov(X,Y) = 0$, and the formula reduces to

$$Var(X+Y) = Var(X) + Var(Y).$$

This result can then be extended by induction: the variance of a sum of *mutually independent* random variables is the sum of their variances. (If the variables are not independent, you must also account for all the covariance terms between each pair.)

Example 3.11. If $X_1, X_2, \ldots, X_n$ are independent returns of n different investments, and you invest equally in all of them, then your total return is $T = \sum_{i=1}^{n} X_i$. The variance of the total return is $Var(T) = \sum_{i=1}^{n} Var(X_i)$, since independence kills all covariance terms. This illustrates one benefit of diversification: if individual variances don't all move in sync, the total variance is spread out across terms (in fact if they were identical and independent, total variance would grow linearly with n, whereas investing all money in one would scale variance by n^2 for the same expected return $nE[X_i]$).

The Cauchy–Schwarz Inequality and Correlation

The covariance has units equal to the product of the units of X and Y. For example, if X = height (in cm) and Y = weight (in kg), then $Cov(X, Y)$ would have units cm·kg, which is not easy to interpret directly. Moreover, the numerical value of covariance can be hard to compare across different pairs of variables or datasets because it depends on the arbitrary scaling of variables. We often prefer a normalized measure of linear association: the **correlation coefficient**.

Definition 3.12 (Correlation Coefficient)**.** The (Pearson) correlation between two random variables X and Y is defined as

$$\rho_{X,Y} = \frac{Cov(X, Y)}{sd(X)sd(Y)},$$

provided sd(X) and sd(Y) are both finite and nonzero. In other words, $\rho_{X,Y}$ is covariance scaled by the standard deviations of each variable.

Correlation is a unitless number, since the units cancel out in the ratio. It always lies between -1 and 1:$-1 \leq \rho_{X,Y} \leq 1$. This bound is a consequence of the famous **Cauchy–Schwarz inequality**. In particular, $\rho_{X,Y} = \pm 1$ if and only if X and Y have an exact linear relationship between them (with probability 1), i.e. $Y = a + bX$ for some constants a, b (with $b > 0$ giving $+1$ correlation and $b < 0$ giving -1). If $\rho_{X,Y} = 0$, we say X and Y are *uncorrelated*, which as noted is weaker than being independent (though for jointly Normal random variables, zero correlation does imply independence, a special case often encountered in statistics).

Proof. (Sketch) To see why $|\rho_{X,Y}| \leq 1$, consider the Cauchy–Schwarz inequality in the form: $(E[UV])^2 \leq E[U^2]E[V^2]$, which holds for any random variables U and V (this is analogous to the inequality $\langle u, v\rangle^2 \leq |u|^2|v|^2$ for vectors). Take $U = X - E[X]$ and $V = Y - E[Y]$. Then $E[U] = E[V] = 0$. The inequality becomes: $(E[(X - \mu_X)(Y - \mu_Y)])^2 \leq E[(X - \mu_X)^2]E[(Y - \mu_Y)^2]$. The left side is $Cov(X, Y)^2$ and the right side is $Var(X)\,Var(Y)$. So $Cov(X, Y)^2 \leq Var(X)Var(Y)$. Taking square roots (and noting variances are positive) yields $|Cov(X, Y)| \leq sd(X)sd(Y)$. Now dividing both sides by the product sd(X) sd(Y) (assuming these are nonzero) gives exactly $|\rho_{X,Y}| \leq 1$. The equality case occurs if and only if U and V are linearly dependent (one is a scalar multiple of the other), which translates to X and Y having a perfect linear relationship. □

Correlation provides a standardized measure of linear association. A correlation near $+1$ means X and Y are almost perfectly linearly increasing together; near -1 means as one increases, the other decreases in a near-perfect linear way; near 0 means knowing one gives essentially no linear prediction of the other. However, be cautious: correlation specifically measures *linear* relationships. It is possible for two variables to have a strong nonlinear relationship but zero correlation (like the X and $Y = X^2$ example earlier: the data would form a parabola shape if plotted, which is symmetric and yields $\rho = 0$). Always visualize or consider the possibility of nonlinear associations in data rather than relying solely on correlation.

3.3 Features of Conditional Distributions

So far we have discussed expectation, variance, etc., with respect to the entire distribution of a random variable (sometimes called "unconditional" or "marginal" expectation/variance). In many situations, we have additional information or conditioning events. For example, we might be interested in the average outcome Y given some condition on X. This leads us to **conditional expectation** and **conditional variance**, which are key concepts especially in regression analysis and causal inference.

3.3.1 Conditional Expectation

Definition 3.13 (Conditional Expectation)**.** For random variables X and Y, the *conditional expectation of* Y *given* $X = x$ (assuming it exists) is

$$E[Y \mid X = x] = \begin{cases} \sum_y y\, f_{Y|X}(y|x), & \text{if } Y \text{ is discrete,} \\ \int_{-\infty}^{\infty} y\, f_{Y|X}(y|x)\, dy, & \text{if } Y \text{ is continuous.} \end{cases}$$

where $f_{Y|X}(y|x)$ is the conditional pmf/pdf of Y given $X = x$.

The conditional expectation $E[Y|X = x]$ is simply the expected value of Y when X is known to equal x. As a function of x, it is often called the **conditional expectation function (CEF)**: $m(x) := E[Y|X = x]$. If we treat X as random, then $E[Y|X]$ denotes the random variable $m(X)$ which equals $m(x)$ whenever $X = x$. In other words, $E[Y|X]$ is a function of X. We sometimes write $E[Y|X] = g(X)$ to emphasize that it is a random variable measurable in terms of X.

Conditional expectation is extremely important in econometrics: it provides the best prediction of Y given X in a mean-squared error sense. In fact, $E[Y|X]$ minimizes $E[(Y - h(X))^2]$ over all functions $h(X)$, which is why linear regression aims to estimate $E[Y|X]$ (the true conditional mean of Y given X). If X contains all relevant observed information, $E[Y|X]$ is the optimal predictor of Y using that information.

Example 3.14. Let X be uniformly distributed on $[0, 1]$. Now suppose that conditional on $X = x$, the random variable Y is uniformly distributed on $[x, 1]$. In other words, we *first* draw $X \sim \text{Uniform}(0, 1)$, and then we draw Y from a uniform distribution that depends on X (the support of Y shifts with X). We can compute $E[Y|X = x]$ from the definition: $E[Y \mid X = x] = \int_{y=x}^{1} y \cdot \frac{1}{1-x} dy$, for $0 \leq x \leq 1$. (The conditional density of Y given $X = x$ is $1/(1 - x)$ on the interval $[x, 1]$.) Carrying out the integration:

$$E[Y \mid X = x] = \frac{1}{1-x} \int_{x}^{1} y dy = \frac{1}{1-x} \left[\frac{y^2}{2} \right]_{y=x}^{y=1} = \frac{1+x}{2}$$

So the conditional expectation function is $m(x) = \frac{1+x}{2}$. If $X = 0.2$, then $E[Y|X = 0.2] = 0.6$; if $X = 0.8$, then $E[Y|X = 0.8] = 0.9$, etc. Notice $E[Y|X]$ is itself a random variable equal to $(1+X)/2$. The distribution of this random variable $E[Y|X]$ is just the distribution of $(1 + X)/2$. Since X was Uniform(0,1), $(1+X)/2$ is Uniform(0.5, 1). But keep in mind the distinction: $E[Y|X]$ (with capital X) is a random variable that can take different values depending on the realized X, while $E[Y|X = x]$ (with lowercase x) is a specific number, the conditional mean for that fixed x.

Conditional expectations share linearity properties similar to regular expectations, with one important caveat: when taking expectation conditional on X, any function of X can be treated like a “constant” (since given X, that function’s value is known). Formally: – $E[a + bY \mid X] = a + bE[Y \mid X]$ for constants a, b. – If $h(X)$ is any function of X alone, then $E[h(X)Y \mid X] = h(X)E[Y \mid X]$. More generally, for any two functions $h_1(X)$ and $h_2(X)$ of X, $E[h_1(X) + h_2(X)g(Y) \mid X] = h_1(X) + h_2(X)E[g(Y) \mid X]$.

In other words, inside a conditional expectation $E[\cdot|X]$, you can pull out factors that depend only on X (since conditioning on X makes them known). This is analogous to factoring constants out of an expectation, but here the "constant" is any X-measurable function.

Perhaps the most important property is the **Law of Iterated Expectations (LIE)**, also known as the **tower property**:

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

This says that if you first take the conditional expectation of Y given X, and then take the expectation of that (averaging out X), you recover the overall expectation of Y. Symbolically, $E(E(Y|X)) = E(Y)$. The law of iterated expectations is intuitive: $E[Y|X]$ is the best guess of Y when you know X. If you then average that guess over all possible values of X, weighted by how likely those X are, you should get the average of Y overall. A discrete proof is straightforward:

$$E_X[E[Y \mid X]] = \sum_x E[Y \mid X = x]P(X = x) = \sum_x \sum_y yP(Y = y \mid X = x)P(X = x).$$

But $\sum_x P(Y = y \mid X = x)P(X = x) = P(Y = y)$ by the law of total probability. Thus the inner sum becomes $\sum_y y\,P(Y = y) = E[Y]$. The continuous version is analogous with integrals.

Using LIE: Simpson's Paradox. The law of iterated expectations is very useful for breaking complicated expectations into parts. It can also illuminate phenomena like Simpson's paradox, where aggregated data can mislead. Simpson's paradox occurs when a comparison of two groups reverses sign upon conditioning on a third factor. For example, it's possible that $E[Y \mid \text{Gender} = M] > E[Y \mid \text{Gender} = F]$ (men have higher average Y than women overall), yet within each category of a third variable (say country of origin or department), women have higher averages: $E[Y \mid \text{Gender} = M, \text{Origin} = A] < E[Y \mid \text{Gender} = F, \text{Origin} = A]$ and likewise for Origin B. How can this happen?

The law of iterated expectations provides the clue:

$$E[Y \mid \text{Gender} = M] = E_{\text{Origin}}[E[Y \mid \text{Gender} = M, \text{Origin}]]$$

This expands to a weighted average of the conditional expectations within each origin group:

$$\begin{aligned} E[Y \mid G = M] &= E[Y \mid G = M, O = A] \cdot P(O = A \mid G = M) \\ &\quad + E[Y \mid G = M, O = B] \cdot P(O = B \mid G = M). \end{aligned}$$

Similarly

$$\begin{aligned} E[Y \mid G = F] &= E[Y \mid G = F, O = A] \cdot P(O = A \mid G = F) \\ &\quad + E[Y \mid G = F, O = B] \cdot P(O = B \mid G = F). \end{aligned}$$

Even if $E[Y|M, A] < E[Y|F, A]$ and $E[Y|M, B] < E[Y|F, B]$ (women outperform men in both origin groups), the overall averages can be reversed if the weightings (the $P(O = \cdot|G = \cdot)$ terms) differ. For instance, if most men are from origin A (where everyone tends to have high Y) and most women from origin B (where everyone has lower Y), the aggregate might show men ahead. This is exactly what happened in the famous UC Berkeley graduate admissions case, where women applied to more competitive departments (low admission rates) than men did, creating a paradox in the aggregated data. The lesson: always consider relevant covariates (like origin or department) before concluding that one group inherently has a higher expectation than another. The law of iterated expectations helps to formally relate the conditional and marginal expectations.

3.3.2 Conditional Variance

Just as we defined conditional expectation to characterize the mean of Y given X, we can define the **conditional variance** of Y given X to characterize the uncertainty or variability of Y around that conditional mean.

Definition 3.15 (Conditional Variance)**.** The *conditional variance* of Y given X is

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

Equivalently, $Var(Y \mid X) = E[Y^2 \mid X] - (E[Y \mid X])^2$.

Here $Var(Y \mid X)$ is a function of X (hence a random variable when X is random). For each realized value of $X = x$, $Var(Y \mid X = x)$ tells us how spread out Y is around the mean $E[Y \mid X = x]$. In regression terms, if Y is an outcome and X are covariates, $Var(Y \mid X)$ is the "noise" or idiosyncratic variance that remains in Y after accounting for X. When $Var(Y \mid X)$ is not constant but depends on X, we have heteroskedasticity (variance of the residuals depends on X).

Example 3.16. Education and Wages. Consider again Y = hourly wage, and D = college degree indicator (1 if college graduate, 0 if not). Then $E[Y \mid D = 1]$ is the average wage of college grads, and $E[Y \mid D = 0]$ that of non-grads. But what about $Var(Y \mid D = 1)$ versus $Var(Y \mid D = 0)$? Which do you think is higher? One might guess that the variance in wages among college graduates is higher. College-educated workers might have a broad range of outcomes: some end up in highly paid professional jobs, others in lower paid jobs, giving a wide spread. Non-college workers may be more concentrated in a narrower band of lower-skilled jobs (all earning similarly modest wages), so their wage distribution could be tighter. Indeed, data often show greater wage dispersion for higher-educated groups. Thus $Var(Y \mid D = 1)$ is likely larger than $Var(Y \mid D = 0)$. This is an example of how conditional variance can yield insights: the effect of education is not only to raise average wages but also potentially to increase the inequality or variability of wages among those who attain higher education.

Just as there was a law of total expectation, there is a companion formula called the **Law of Total Variance**. It provides a useful decomposition of the overall variance of Y into explained and unexplained parts:

$$Var(Y) = E[Var(Y \mid X)] + Var(E[Y \mid X]).$$

This is a neat identity. The second term $Var(E[Y|X])$ is the variance of the conditional mean $E[Y|X]$ when X varies — effectively, how much of Y's variability is due to differences in the conditional expectation across different X. The first term $E[Var(Y|X)]$ is the average of the within-X variances — basically, the expected leftover variance of Y that remains after accounting for X. Sometimes this is described as: total variance = "explained variance" + "unexplained variance." If X accounts for a lot of the variation in Y, then $E[Var(Y|X)]$ will be small and $Var(E[Y|X])$ will be large.

Proof. Starting from the definition, we have:

$$Var(Y) = E[Y^2] - (E[Y])^2.$$

By the Law of Iterated Expectations, $E[Y] = E(E[Y|X])$. Now add and subtract $E[(E[Y|X])^2]$ inside the expression:

$$\begin{aligned}Var(Y) &= E[E[Y^2|X]] - (E[E[Y|X]])^2 \\ &= E[E[Y^2|X]] - E[(E[Y|X])^2] + E[(E[Y|X])^2] - (E[Y])^2.\end{aligned}$$

Observe that $E[E[Y^2|X]] = E[Y^2]$ and $E[(E[Y|X])^2] - (E[Y])^2 = Var(E[Y|X])$. Also $E[Y^2|X] - (E[Y|X])^2 = Var(Y|X)$ by definition. So the first two terms become $E[\,Var(Y|X)\,]$. Thus we end up with

$$Var(Y) = E[Var(Y \mid X)] + Var(E[Y \mid X]).$$

which is the desired result. □

Interpretation: If you think of predicting Y given X, $E[Y|X]$ is the prediction (the part "explained" by X). The variance of that prediction as X varies is basically how much of Y's variance is explained by X. The expected conditional variance is the remaining variance not explained by X. In extreme cases: – If Y is almost a deterministic function of X (very little noise), then $Var(Y|X)$ is nearly 0 always, so $E[Var(Y|X)] \approx 0$ and $Var(Y) \approx Var(E[Y|X])$. All variance in Y comes from differences in X. – If knowing X tells you almost nothing about Y (very weak relationship), then $E[Y|X]$ is nearly constant (equal to $E[Y]$), so $Var(E[Y|X]) \approx 0$ and $Var(Y) \approx E[Var(Y|X)]$. The total variance is just the average variance within each conditional distribution (nothing is explained by X). This decomposition is useful in analysis of variance (ANOVA) and in understanding R^2 in regression (which is the fraction of variance explained by X).

3.4 Mean Independence

We conclude this chapter with an important concept that lies between the extremes of full independence and mere uncorrelatedness: **mean independence**. In many econometric contexts, assumptions are made about zero conditional mean of errors, etc., which are essentially mean independence assumptions.

Definition 3.17 (Mean Independence)**.** We say a random variable Y is *mean independent* of X if

$$E[Y \mid X = x] = E[Y] \quad \text{"for all"}\ x,$$

i.e. the conditional expectation of Y given X is constant (equal to the unconditional expectation). Equivalently, $E[Y \mid X] = E[Y]$ as a random variable (almost surely).

Mean independence means that knowing X has no effect on the expected value of Y. In other words, X carries no information about the mean of Y (though it could affect higher moments or the distribution of Y in other ways). This is a much weaker condition than full independence, which would require Y's entire distribution to be the same regardless of X. Here we only demand the first moment is the same.

Trivially, if X and Y are independent, then $E[Y|X] = E[Y]$ (because the conditional distribution of Y given any $X = x$ is just the marginal distribution of Y). So independence $\implies$ mean independence. However, the converse is not true: there are dependent random variables which are mean independent. All that's required for mean independence is a cancellation in the first moment.

Example 3.18. (Mean independence without full independence): Take X uniformly distributed on $[-1, 1]$ and let $Y = X^2$. Then clearly Y depends on X (indeed Y is completely determined by X). They are not independent. But $E[X] = 0$, and for any given $Y = y$, the two possible X values are $+\sqrt{y}$ or $-\sqrt{y}$, symmetric about 0. Thus $E[X \mid Y = y] = 0$ as well. In other words $E[X|Y] = 0 = E[X]$. Here X is mean independent of Y. (Note: Y is *not* mean independent of X in this example, since $E[Y|X] = X^2$ which is not constant.) This construction shows mean independence does not imply independence: X and Y are quite dependent, yet the mean of X given Y is always the same as the overall mean of X. The key was the symmetry that made the mean wash out.

One can fabricate many such examples. A general recipe: first pick Y freely, then define a conditional distribution for X given each $Y = y$ that has mean equal to $E[X]$. For instance, let Y be any non-degenerate random variable (so X and Y will be dependent through this construction). Given $Y = y$, let X be equally likely to take two values symmetric around $E[X]$. This guarantees $E[X|Y = y] = E[X]$. Unless those two symmetric values coincide (which would make X degenerate), X and Y are not independent.

Mean Independence vs Uncorrelatedness: Mean independence of Y with respect to X implies $E[Y \mid X] = E[Y]$ for all X. If we take an expectation (over X) on both sides, we get $E[Y] = E[E[Y \mid X]] = E[Y]$, which is tautologically true. But if we multiply both sides by any function $h(X)$ and then take expectation, we also get:

$$E[h(X)E(Y \mid X)] = E[h(X)E(Y)] = E(Y)E[h(X)],$$

and since $E[h(X)E(Y \mid X)] = E[E(h(X)Y \mid X)] = E[h(X)Y]$, we have

$$E[h(X)Y] = E(Y)E[h(X)].$$

This holds for all $h(X)$. In particular, taking $h(X) = X$ itself gives

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

so $Cov(X, Y) = 0$. Thus mean independence $\implies$ zero covariance (uncorrelatedness). Again, the converse is not true: uncorrelatedness is weaker. For example, X and $Y = X^2$ earlier are uncorrelated (covariance 0) but X is not mean independent of Y (since $E[X|Y] = 0$, true, but Y not mean independent of X). So the hierarchy is: *Independence* $\implies$ *Mean independence* $\implies$ *Uncorrelatedness*, with each converse failing in general.

Why do we care about mean independence? In many econometric models, especially linear ones, we require the error term to be mean independent of the regressors. For instance, a key assumption for OLS regression to identify causal effects is $E[u \mid X] = 0$, meaning the error has mean zero given the regressors X. This is strictly weaker than assuming u is independent of X (which is often too strong; we allow u to be heteroskedastic or even depend on X's distribution in higher moments, as long as the mean given X is zero). Mean independence is enough to ensure X has no predictive power for Y's mean, which is exactly what's needed for unbiased estimation of a linear effect. It's a minimal "no omitted variable bias" condition in that sense.

To summarize: – If Y is mean independent of X, then knowing X does not change your best guess of Y's average value. X might still affect the variance or other aspects of Y's distribution, but not the mean. – Independence of X and Y would mean X tells you nothing about Y at all (not just the mean, but the entire distribution). – Uncorrelated (covariance zero) means no linear relationship in a global sense, but doesn't necessarily hold conditionally or in nonlinear ways.

3.5 Conclusion

In this chapter, we reviewed various summary characteristics of distributions:

- The **expectation** (mean) of a random variable, and how to compute it directly or via the Law of the Unconscious Statistician for transformed variables.

- The **variance** (and standard deviation) as a measure of spread, and related concepts like scaling properties and examples for common distributions.

- The **covariance** between two variables as a measure of their joint variability, and the derived concept of **correlation** which standardizes covariance to a $[-1, 1]$ scale. We saw that correlation is bounded by ± 1 (Cauchy–Schwarz) and captures linear association.

- We emphasized that $Cov(X, Y) = 0$ or $\rho_{XY} = 0$ does *not* imply independence except in special cases, even though independence always implies zero covariance.

- We explored **conditional expectation** $E[Y|X]$ as a random variable (function of X) giving the mean of Y for each value of X. This is central to regression analysis. We practiced using the **law of iterated expectations** $E(Y) = E(E(Y|X))$ and saw how failing to condition can lead to Simpson's paradox.

- We defined **conditional variance** $Var(Y|X)$ and presented the **law of total variance** $Var(Y) = E[Var(Y|X)] + Var(E[Y|X])$, which splits variance into explained and unexplained parts.

- Finally, we introduced **mean independence** as a weaker notion than full independence. We discussed its implications and its role in econometric assumptions (e.g. exogeneity of regressors requiring $E[u|X] = 0$).

With the tools from Part A (distributions) and Part B (expectations and moments) of our probability review, we are now well-equipped to handle the statistical concepts needed for causal inference. We can precisely define causal estimands as expectations (e.g. average treatment effects), and we can invoke assumptions like "selection on observables" or "instrument exogeneity" in terms of conditional independence or mean independence to identify those estimands. In the coming lectures, we will start examining **estimation**: how to use sample data to estimate these theoretical quantities. Our understanding of expectation will be crucial since an estimator is essentially a function of random sample data (hence itself a random variable) whose expectation we often set to a target parameter. Properties like variance and covariance will help us quantify estimation uncertainty and test hypotheses.