

Why Expectation Is Blind to Dependence—but Variance Is Not

One of the most powerful and frequently used properties in probability theory is the **linearity of expectation**: for any random variables X_1, X_2, \dots, X_n ,

$$\mathbb{E}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbb{E}[X_i].$$

Remarkably, this identity holds *regardless of whether the variables are independent, dependent, or even deterministically linked*. No assumptions about joint distributions are needed. This robustness makes expectation an indispensable tool in probabilistic analysis—especially when dealing with complex or unknown dependencies.

In stark contrast, the **variance of a sum** is highly sensitive to the relationships between variables. For two random variables X and Y ,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y),$$

where $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ is the covariance. Only when $\text{Cov}(X, Y) = 0$ —i.e., when X and Y are **uncorrelated**—does variance become additive. Since independence implies zero covariance (but not vice versa), independence is a sufficient—but not necessary—condition for variances to add.

This asymmetry arises because expectation is a *linear* operator, while variance is *quadratic*. The expectation of a sum depends only on marginal averages, but the variance of a sum involves the joint behavior through $\mathbb{E}[XY]$. Thus, while expectation “ignores” dependence, variance “detects” it—specifically, its linear component.

To illustrate this distinction concretely, consider the following natural example based on a simple random experiment.

A Coin-Toss Example: Dependent but Uncorrelated Variables

Toss a fair coin twice. The sample space is

$$\Omega = \{HH, HT, TH, TT\},$$

with each outcome having probability $1/4$. Define two random variables:

- $X = (\text{number of heads}) - (\text{number of tails})$. Thus, $X = +2$ for HH , 0 for HT or TH , and -2 for TT .
- $Y = \begin{cases} +1 & \text{if both tosses are the same (i.e., } HH \text{ or } TT), \\ -1 & \text{if the tosses differ (i.e., } HT \text{ or } TH). \end{cases}$

The joint distribution is summarized below:

Outcome	X	Y	Probability
HH	$+2$	$+1$	$1/4$
HT	0	-1	$1/4$
TH	0	-1	$1/4$
TT	-2	$+1$	$1/4$

Dependence

The variables are clearly dependent: if $Y = +1$ (tosses match), then X must be ± 2 ; if $Y = -1$ (tosses differ), then $X = 0$ with certainty. Formally,

$$P(X = 0, Y = +1) = 0 \quad \text{but} \quad P(X = 0)P(Y = +1) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} \neq 0,$$

so X and Y are **not independent**.

Zero Correlation

Now compute the covariance:

$$\mathbb{E}[X] = (+2) \cdot \frac{1}{4} + 0 \cdot \frac{1}{2} + (-2) \cdot \frac{1}{4} = 0,$$

$$\mathbb{E}[Y] = (+1) \cdot \frac{1}{2} + (-1) \cdot \frac{1}{2} = 0,$$

$$\mathbb{E}[XY] = (+2)(+1) \cdot \frac{1}{4} + (0)(-1) \cdot \frac{1}{4} + (0)(-1) \cdot \frac{1}{4} + (-2)(+1) \cdot \frac{1}{4} = \frac{2}{4} - \frac{2}{4} = 0.$$

Hence,

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 - 0 = 0.$$

So X and Y are **uncorrelated**, despite being dependent.

Implications for Sums

Now consider the sum $S = X + Y$. By linearity of expectation—which *requires no independence*—we have

$$\mathbb{E}[S] = \mathbb{E}[X] + \mathbb{E}[Y] = 0 + 0 = 0.$$

However, the variance of S is

$$\text{Var}(S) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y) = \text{Var}(X) + \text{Var}(Y),$$

because $\text{Cov}(X, Y) = 0$. In this case, variances *do* add—but *not because X and Y are independent* (they are not!), but because they happen to be uncorrelated.

This underscores a crucial point: ****variance additivity depends on uncorrelatedness, not independence****. If we had chosen different functions of the coin tosses that were correlated, the covariance term would be nonzero, and the variance of the sum would reflect their interaction.

Conclusion

Expectation is remarkably agnostic to dependence: the expected value of a sum is always the sum of expected values. Variance, however, encodes how variables fluctuate *together*. While independence guarantees zero covariance, the converse is false—as our coin-toss example shows. Thus, when analyzing sums of random variables:

- Use linearity of expectation freely—even under complex dependence.
- Exercise caution with variance: always consider possible covariance.

This distinction is not merely technical; it shapes how we model uncertainty, design experiments, and interpret data in statistics, machine learning, and the physical sciences.

Dependence as Information Sharing (Shannon’s View)

In classical probability, two random variables X and Y are **statistically independent** if their joint distribution factors:

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y) \quad \text{for all } x, y.$$

If this fails *anywhere*, they are **dependent**.

But this is a *mathematical* definition. The **philosophical and informational meaning**—thanks to Claude Shannon’s information theory—is more intuitive:

X and Y are dependent if observing one changes your knowledge (i.e., your probability assessment) about the other.

This change in knowledge is quantified as a **reduction in uncertainty**, and uncertainty is measured by **entropy**.

Entropy: The Measure of Uncertainty

- The **entropy** of X , denoted $H(X)$, is the average uncertainty (in bits) before observing X :

$$H(X) = - \sum_x P(x) \log_2 P(x).$$

High entropy = high unpredictability.

- The **conditional entropy** $H(X \mid Y)$ is the average uncertainty about X *after* observing Y .

If X and Y are **independent**, then knowing Y tells you *nothing* about X , so:

$$H(X \mid Y) = H(X).$$

But if they are **dependent**, then:

$$H(X \mid Y) < H(X).$$

Your uncertainty about X **decreases** once you know Y .

Mutual Information: The Shared Knowledge

The amount of uncertainty reduced is called the **mutual information** between X and Y :

$$I(X; Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X).$$

- $I(X; Y) \geq 0$,

- $I(X; Y) = 0$ **if and only if** X and Y are independent,
- Larger $I(X; Y)$ means stronger dependence (more shared information).

Crucially, **mutual information captures *any* kind of dependence**—linear, nonlinear, deterministic, or stochastic.

Back to Our Coin-Toss Example

Recall:

- X = net heads minus tails,
- $Y = +1$ if tosses match, -1 if they differ.

We saw that $\text{Cov}(X, Y) = 0$, so they are **uncorrelated**.

But are they **informationally independent**?

No. Consider:

- Before seeing Y , X could be $-2, 0$, or $+2$.
- If you learn $Y = -1$ (tosses differ), you know **with certainty** that $X = 0$.
- Your uncertainty about X drops from $H(X) > 0$ to $H(X | Y = -1) = 0$.

Thus, $H(X | Y) < H(X)$, so $I(X; Y) > 0$.

They share information \rightarrow they are dependent, even though correlation is zero.

Philosophical Implications

1. Knowledge Is Probabilistic

Learning isn't about certainty—it's about **updating beliefs**. Dependence means your belief about one variable *should* change when you observe the other. Correlation misses this if the update isn't linear.

2. Correlation Is a Narrow Channel

Mutual information measures *total* shared information; correlation measures only the **linear component**. It's like judging a symphony by its average pitch—you miss harmony, rhythm, and timbre.

3. Causality Often Hides in Nonlinear Dependence

Many causal mechanisms (e.g., thresholds, feedback loops, logical rules) create **nonlinear dependencies**. If we only test for correlation, we may conclude “no link” where a deep causal structure exists.

4. Science Requires Richer Tools

Relying solely on correlation encourages a **linear worldview**. But biology, economics, and social systems thrive on nonlinear interdependence. Information theory gives us a language to detect and quantify those links.

A Deeper Unity

Shannon’s insight reveals a profound unity:

Statistical dependence = information flow.

This reframes probability not just as a calculus of chance, but as a **calculus of knowledge**. Every time two variables are dependent, there is a channel—however subtle—through which information passes.

Correlation is just one (limited) way to detect that channel. Mutual information, conditional entropy, and other information-theoretic tools let us **listen more carefully**.

In Summary

- **Dependence** means: *Knowing Y changes what you expect about X .*
- This change is a **reduction in uncertainty**, measured by entropy.
- The amount of shared information is **mutual information** $I(X; Y)$.
- **Correlation can be zero even when $I(X; Y) > 0$** —because correlation only sees linear patterns.
- Thus, **uncorrelated \neq independent**—not just mathematically, but **epistemologically**: the variables still “speak” to each other; we just need the right ears to hear it.

This is why modern data science increasingly turns to **information-theoretic methods** (like mutual information feature selection, entropy-based clustering, etc.)—to uncover the hidden conversations between variables that correlation silences.

Let's compute the **entropy** and **mutual information** for the coin-toss example:

- Toss a fair coin twice.
- Define:
- $X = (\text{number of heads}) - (\text{number of tails}) \rightarrow X \in \{-2, 0, +2\}$
- $Y = +1$ if tosses match (HH or TT), -1 if they differ (HT or TH)

From earlier, the joint distribution is:

Outcome	X	Y	Probability
HH	+2	+1	1/4
HT	0	-1	1/4
TH	0	-1	1/4
TT	-2	+1	1/4

Step 1: Marginal Distribution of X

- $P(X = -2) = P(TT) = 1/4$
- $P(X = 0) = P(HT \text{ or } TH) = 1/4 + 1/4 = 1/2$
- $P(X = +2) = P(HH) = 1/4$

So:

$$P_X(-2) = \frac{1}{4}, \quad P_X(0) = \frac{1}{2}, \quad P_X(+2) = \frac{1}{4}$$

Entropy of X :

$$H(X) = - \sum_x P(x) \log_2 P(x) = - \left[\frac{1}{4} \log_2 \frac{1}{4} + \frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{4} \log_2 \frac{1}{4} \right]$$

Compute:

- $\log_2(1/4) = -2$
- $\log_2(1/2) = -1$

So:

$$H(X) = - \left[\frac{1}{4}(-2) + \frac{1}{2}(-1) + \frac{1}{4}(-2) \right] = - \left[-\frac{2}{4} - \frac{1}{2} - \frac{2}{4} \right] = -[-0.5 - 0.5 - 0.5] = -(-1.5) = 1.5$$

$$H(X) = \frac{3}{2} \text{ bits}$$

Step 2: Marginal Distribution of Y

- $P(Y = +1) = P(HH \text{ or } TT) = 1/4 + 1/4 = 1/2$
- $P(Y = -1) = P(HT \text{ or } TH) = 1/2$

So Y is Bernoulli(1/2) over $\{-1, +1\}$.

Entropy of Y :

$$H(Y) = - \left[\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2} \right] = - \left[-\frac{1}{2} - \frac{1}{2} \right] = 1 \text{ bit}$$

$$\boxed{H(Y) = 1 \text{ bit}}$$

Step 3: Conditional Entropy $H(X | Y)$

We compute $H(X | Y = y)$ for each y , then average.

- **When $Y = +1$** (probability