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 $X_1, X_2, \ldots, X_n$,

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

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,

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

where $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ is the covariance. Only when $\mathrm{Cov}(X, Y) = 0$—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 = (number of heads) − (number of tails). Thus, $X = +2$ for HH, 0 for HT or TH, and -2 for TT.
- $Y = \begin{cases} +1 & \text{if both tosses are the same (i.e., } HH \text{ or } TT\text{)}, \\ -1 & \text{if the tosses differ (i.e., } HT \text{ or } TH\text{)}. \end{cases}$

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

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

because $\mathrm{Cov}(X,Y) = 0$. 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 $\mathrm{Cov}(X,Y) = 0$, 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 $H(X \mid Y = -1) = 0$.

Thus, $H(X \mid Y) < H(X)$, 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:
- $X =$ (number of heads) $-$ (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:

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

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

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 $H(X \mid Y)$

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

- **When** $Y = +1$ (probability 1/2): Outcomes: $HH \to X = +2$, $TT \to X = -2$ So $P(X = +2 \mid Y = +1) = \frac{1/4}{1/2} = 1/2$, $P(X = -2 \mid Y = +1) = 1/2$, $P(X = 0 \mid Y = +1) = 0$

  Entropy:

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

- **When** $Y = -1$ (probability 1/2): Outcomes: $HT, TH \to$ both give $X = 0$ So $P(X = 0 \mid Y = -1) = 1$

  Entropy:

  $$H(X \mid Y = -1) = -[1 \cdot \log_2 1] = 0 \text{ bits}$$

Now average:

$$H(X \mid Y) = P(Y = +1)H(X \mid Y = +1) + P(Y = -1)H(X \mid Y = -1) = \frac{1}{2}(1) + \frac{1}{2}(0) = 0.$$

$$\boxed{H(X \mid Y) = \frac{1}{2} \text{ bit}}$$

Step 4: Mutual Information $I(X;Y)$

$$I(X;Y) = H(X) - H(X \mid Y) = \frac{3}{2} - \frac{1}{2} = 1 \text{ bit}$$

We can verify via $H(Y) - H(Y \mid X)$:

- **When** $X = 0$ (prob 1/2): then outcome is HT or $TH \to Y = -1$ with certainty $\to H(Y \mid X = 0) = 0$
- **When** $X = +2$ (prob 1/4): outcome is $HH \to Y = +1 \to H(Y \mid X = +2) = 0$
- **When** $X = -2$ (prob 1/4): outcome is $TT \to Y = +1 \to H(Y \mid X = -2) = 0$

So $H(Y \mid X) = 0$, and $I(X;Y) = H(Y) - 0 = 1$ bit. ✓ Consistent.

$$\boxed{I(X;Y) = 1 \text{ bit}}$$

Final Results

- $H(X) = 1.5$ bits
- $H(Y) = 1$ bit
- $H(X \mid Y) = 0.5$ bits
- $H(Y \mid X) = 0$ bits
- **Mutual Information**: $I(X;Y) = 1$ bit

Interpretation

- Knowing Y reduces uncertainty about X by **1 bit** (from 1.5 to 0.5 bits).
- Knowing X **completely determines** Y (since if $X = 0$, $Y = -1$; if $X = \pm 2$, $Y = +1$), so $H(Y \mid X) = 0$.
- Despite **zero correlation**, there is **1 full bit of shared information**—a substantial dependence.

This quantifies the philosophical point: **uncorrelated does not mean unrelated**.

Mutual Information Example: Dice Roll and Sum

Let X and Y be discrete random variables with joint distribution $p(x, y)$, and marginals $p(x)$, $p(y)$.

1. **Entropy**
   Measures the average uncertainty in a random variable:
$$H(X) = -\sum_{x} p(x) \log_2 p(x)$$

2. **Conditional Entropy**
   Average uncertainty in X given knowledge of Y:
$$H(X \mid Y) = -\sum_{y} p(y) \sum_{x} p(x \mid y) \log_2 p(x \mid y) = \sum_{x,y} p(x, y) \log_2 \frac{1}{p(x \mid y)}$$

3. **Mutual Information**
   Reduction in uncertainty about X due to knowing Y (symmetric):
$$I(X; Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$
   Equivalently, in terms of joint and marginals:
$$I(X; Y) = \sum_{x,y} p(x, y) \log_2 \left(\frac{p(x, y)}{p(x)p(y)} \right)$$

All quantities are measured in **bits** (logarithm base 2).

Natural Dice Example: One Die and the Sum of Two Dice

Setup

Roll two fair six-sided dice, denoted D_1 and D_2. Define:

- $X = D_1$ (the outcome of the first die)
- $Y = D_1 + D_2$ (the sum of both dice) We compute $I(X; Y)$: how much does knowing the sum tell us about the first die?

This is a natural dependence: for example, if $Y = 2$, then X must be 1; if $Y = 7$, then X could be any value from 1 to 6.

Joint and Marginal Distributions

Since the dice are independent and fair,

$$P(D_1 = x, D_2 = d) = \frac{1}{36}, \quad x, d \in \{1, \dots, 6\}.$$

Because $Y = X + D_2$, we have

$$p(x, y) = P(X = x, Y = y) = \begin{cases} \frac{1}{36} & \text{if } 1 \le x \le 6 \text{ and } 1 \le y - x \le 6, \\ 0 & \text{otherwise.} \end{cases}$$

The marginal of X is uniform:

$$p(x) = \frac{1}{6}, \quad x = 1, \dots, 6.$$

The marginal of Y (sum of two dice) is well known:

y	2	3	4	5	6	7	8	9	10	11	12
$p(y)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Computing Mutual Information

We use the identity:

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

Entropy of X

Since X is uniform over 6 outcomes,

$$H(X) = \log_2 6 \approx 2.58496 \text{ bits}.$$

Conditional Entropy $H(X \mid Y)$

Given $Y = y$, the possible values of X are those for which $1 \le y - x \le 6$, i.e., $x \in [\max(1, y - 6), \min(6, y - 1)]$. The number of such values is:

$$n_y = \begin{cases} 1 & y = 2 \text{ or } 12, \\ 2 & y = 3 \text{ or } 11, \\ 3 & y = 4 \text{ or } 10, \\ 4 & y = 5 \text{ or } 9, \\ 5 & y = 6 \text{ or } 8, \\ 6 & y = 7. \end{cases}$$

Because all valid (x, y) pairs have equal probability $(1/36)$, the conditional distribution $p(x \mid y)$ is uniform over the n_y possibilities. Hence,

$$H(X \mid Y = y) = \log_2 n_y.$$

Therefore,

$$H(X \mid Y) = \sum_{y=2}^{12} p(y) \log_2 n_y.$$

Grouping symmetric terms:

$$\begin{aligned} H(X \mid Y) = &\frac{2}{36} \cdot \log_2 1 + \frac{4}{36} \cdot \log_2 2 + \frac{6}{36} \cdot \log_2 3 + \frac{8}{36} \cdot \log_2 4 \\ &+ \frac{10}{36} \cdot \log_2 5 + \frac{6}{36} \cdot \log_2 6. \end{aligned}$$

Using $\log_2 1 = 0$, $\log_2 2 = 1$, $\log_2 4 = 2$, and numerical values:

$$\begin{aligned} \log_2 3 &\approx 1.58496, \\ \log_2 5 &\approx 2.32193, \\ \log_2 6 &\approx 2.58496, \end{aligned}$$

we compute:

$$\begin{aligned} H(X \mid Y) &\approx \frac{4}{36}(1) + \frac{6}{36}(1.58496) + \frac{8}{36}(2) + \frac{10}{36}(2.32193) + \frac{6}{36}(2.58496) \\ &= \frac{4}{36} + \frac{9.5098}{36} + \frac{16}{36} + \frac{23.2193}{36} + \frac{15.5098}{36} \\ &= \frac{68.2389}{36} \approx 1.8955 \text{ bits.} \end{aligned}$$

Mutual Information

Finally,

$$I(X;Y) = H(X) - H(X \mid Y) \approx 2.58496 - 1.8955 = \boxed{0.6895 \text{ bits}}.$$

Interpretation

- Initially, the first die has $H(X) \approx 2.585$ bits of uncertainty. - After observing the sum Y, uncertainty reduces to $H(X \mid Y) \approx 1.896$ bits. - Thus, the sum reveals approximately 0.69 bits of information about the first die. -

This reflects intuition: extreme sums (e.g., 2 or 12) fully determine X, while moderate sums (e.g., 7) leave significant uncertainty.

This example uses only standard, fair dice—no hidden variables or artificial constructions—and illustrates how mutual information quantifies dependence in a natural probabilistic setting.

Mutual Information and Bayes' Theorem: Information Gain from Bayesian Updating

Mutual information and Bayes' theorem are deeply connected:

- **Bayes' theorem** tells you how to update a *single prior* $P(X)$ to a *posterior* $P(X \mid Y = y)$ after observing a specific outcome $Y = y$.
- **Mutual information** $I(X;Y)$ tells you the *expected reduction in uncertainty* about X *on average* over all possible outcomes y, weighted by their likelihood.

In short:

Mutual information = Expected information gain from Bayesian updating.

Formal Connection

1. Bayes' Theorem and Information Gain for a Single Observation

Given a prior distribution $P(X)$, observing $Y = y$ yields the posterior via Bayes' rule:

$$P(X \mid Y = y) = \frac{P(Y = y \mid X)\, P(X)}{P(Y = y)}.$$

The **information gained** from this observation is measured by the Kullback–Leibler (KL) divergence from the prior to the posterior:

$$D_{\mathrm{KL}}\big(P(X \mid Y = y) \,\|\, P(X)\big) = \sum_x P(x \mid y) \log_2 \frac{P(x \mid y)}{P(x)}.$$

This quantifies how much the belief about X changed due to seeing $Y = y$.

2. Mutual Information as Expected KL Divergence

Mutual information is the expectation of this KL divergence over all possible outcomes y:

$$\boxed{I(X;Y) = \mathbb{E}_Y\big[D_{\mathrm{KL}}\big(P(X \mid Y) \,\|\, P(X)\big)\big] = \sum_y P(y) \sum_x P(x \mid y) \log_2 \frac{P(x \mid y)}{P(x)}.}$$

Using $P(x, y) = P(x \mid y)P(y)$, this is algebraically equivalent to the standard definition:

$$I(X;Y) = \sum_{x,y} P(x, y) \log_2 \left(\frac{P(x, y)}{P(x)P(y)} \right).$$

Thus, mutual information is precisely the **average information gain** from applying Bayes' rule.

Interpretation

[leftmargin=*]

- **KL divergence** $D_{\mathrm{KL}}(P(X|y)\|P(X))$:
  "How many extra bits would I waste if I encoded X using the prior instead of the correct posterior?"
  -¿ This is the *information gained* from observing $Y = y$.
- **Mutual information** $I(X;Y)$:
  "On average, how many bits do I gain about X per observation of Y?"
  -¿ This is the *expected information gain* from Bayesian updating.

Concrete Dice Example: Connecting Bayes + Mutual Information

Recall the setup:

- Roll two fair dice: D_1, D_2.
- Let $X = D_1$ (first die), $Y = D_1 + D_2$ (sum).
- Prior: $P(X = x) = \frac{1}{6}$ for $x = 1, \ldots, 6$.

Bayesian Updating for Specific Observations

1. **Observe** $Y = 2$:
   Only possible if $D_1 = 1, D_2 = 1$.
   Posterior: $P(X = 1 \mid Y = 2) = 1$, others 0.
   Information gain:
   $$D_{\mathrm{KL}}(P(X|2) \| P(X)) = \log_2 \frac{1}{1/6} = \log_2 6 \approx 2.585 \text{ bits}.$$
2. **Observe** $Y = 7$:
   All pairs $(1,6), (2,5), \ldots, (6,1)$ equally likely.
   Posterior: $P(X = x \mid Y = 7) = \frac{1}{6}$ for all x.
   Information gain:
   $$D_{\mathrm{KL}}(P(X|7) \| P(X)) = \sum_{x=1}^{6} \frac{1}{6} \log_2 \frac{1/6}{1/6} = 0 \text{ bits}.$$

Mutual Information as the Average Gain

Mutual information averages these gains over all possible sums:

$$I(X;Y) = \sum_{y=2}^{12} P(Y = y) \cdot D_{\mathrm{KL}}\big(P(X \mid Y = y) \parallel P(X)\big).$$

Using the known distribution of Y:

y	2	3	4	5	6	7	8	9	10	11	12
$P(Y = y)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
n_y (support size)	1	2	3	4	5	6	5	4	3	2	1
D_{KL}	$\log_2 6$	$\log_2 3$	$\log_2 2$	$\log_2 \frac{4}{3}$?	$\ldots$	0	$\ldots$				

More precisely, since $P(X \mid Y = y)$ is uniform over n_y values,

$$D_{\mathrm{KL}}\big(P(X \mid Y = y) \parallel P(X)\big) = \log_2 n_y - \log_2 6 = \log_2 \left(\frac{n_y}{6}\right) \quad \text{(but note: actually } = \log_2 n$$

Carrying out the full calculation (as in the earlier example) yields:

$$I(X;Y) \approx 0.6895 \text{ bits}.$$

This is exactly the **expected information gain** from observing the sum and applying Bayes' rule.

Why This Matters

[leftmargin=*]

- **Bayesian inference**: Every Bayes update provides information; mutual information quantifies its average value.
- **Experimental design**: Choose observations Y that maximize $I(X;Y)$ to learn most about X.
- **Machine learning**: In variational inference, mutual information appears in bounds on model evidence.
- **Cognitive science**: Models perception as Bayesian updating, with mutual information measuring sensory informativeness.

Summary

Concept	Role
Bayes' theorem	Updates prior $P(X) \to$ posterior $P(X \mid Y = y)$ fo
KL divergence $D_{\mathrm{KL}}(P(X\|y)\\|P(X))$	Information gained from that *specific* update.
Mutual information $I(X;Y)$	*Expected* information gain over all y.

Thus, mutual information provides an **information-theoretic foundation for Bayesian learning**: it measures how much an observation is expected to teach us about the world.