Chapter 4: Expected Values

Donghyun Ko

May 27, 2026

In Chapter 3, we described randomness using distributions (PMF/PDF/CDF), including joint distributions and conditioning. Chapter 4 asks a different question: *how can we summarize a distribution by a few numbers that are easy to interpret and compute?* The two most important summaries are the mean (expected value) and the variance (spread). We then learn inequalities (Markov and Chebyshev) that control tail probabilities using only these summaries. Finally, we extend these ideas to multiple random variables (covariance and correlation), conditional expectation (as a tool for simplification and prediction), and moment generating functions.

Contents

1 What you should be able to do

After reading this note, you should be able to:

- Compute $\mathbb{E}[X]$ and $\mathbb{E}[g(X)]$ in discrete and continuous settings;
- Use linearity of expectation to simplify computations (no independence needed);
- Compute $\mathrm{Var}(X)$ and $\mathrm{SD}(X)$ and use $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mu^2$;
- Apply Markov and Chebyshev inequalities to obtain distribution-free tail bounds;
- Compute covariance/correlation and $\mathrm{Var}(aX + bY)$;
- Use conditional expectation and the identities

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y \mid X]], \qquad \mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y \mid X)] + \mathrm{Var}(\mathbb{E}[Y \mid X]);$$

- Use MGFs to compute moments and recognize distributions (when the MGF exists).

Notation reminder: write $F_X(\cdot)$ for a CDF as a function ("placeholder" $\cdot$).

2 Expected value (mean)

2.1 Motivation: why expectation matters

Suppose you repeat a random experiment many times and record values $X_1, X_2, \ldots, X_m$. A central empirical summary is the sample average

$$\overline{X}_m = \frac{1}{m} \sum_{i=1}^{m} X_i.$$

When the experiment is stable and m is large, $\overline{X}_m$ tends to settle near a constant. That constant is what probability theory calls the *expected value* $\mathbb{E}[X]$. So you can think of expectation as the theoretical long-run average.

2.2 Definition (discrete vs. continuous)

There are two common situations:

- **Discrete:** X takes isolated values (like counts).
- **Continuous:** X takes values on intervals (like time, length).

In both cases, $\mathbb{E}[X]$ is a weighted average, using probability as the weight.

Theorem

Expected value (definition). Let X be a random variable.

- If X is discrete with PMF $p_X(x) = \mathbb{P}(X = x)$, then

$$\mathbb{E}[X] = \sum_{x \in \text{Supp}(X)} x\, p_X(x), \quad \text{provided } \sum_x |x| p_X(x) < \infty.$$

- If X is continuous with PDF $f_X(x)$, then

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x)\, dx, \quad \text{provided } \int_{-\infty}^{\infty} |x| f_X(x)\, dx < \infty.$$

If the absolute-sum / absolute-integral condition fails, the mean is *undefined*.

2.3 Expectation of a function: $\mathbb{E}[g(X)]$

A major practical idea in Rice is the *LOTUS* principle (Law Of The Unconscious Statistician): to compute $\mathbb{E}[g(X)]$, you integrate (or sum) $g(x)$ against the distribution of X. You do *not* need the distribution of $g(X)$.

Theorem

LOTUS (expectation of a transformed variable). Let $Y = g(X)$.

- Discrete:

$$\mathbb{E}[g(X)] = \sum_{x \in \text{Supp}(X)} g(x)\, p_X(x), \quad \text{provided } \sum_x |g(x)| p_X(x) < \infty.$$

- Continuous:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx, \quad \text{provided } \int |g(x)|\, f_X(x)\, dx < \infty.$$

2.4 Example 1 (discrete): compute $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$ carefully

Assume $X \in \{0, 1, 2\}$ with

$$\mathbb{P}(X = 0) = 0.2, \quad \mathbb{P}(X = 1) = 0.5, \quad \mathbb{P}(X = 2) = 0.3.$$

Start by writing the definition:

$$\mathbb{E}[X] = \sum_x x\, p_X(x) = 0 \cdot 0.2 + 1 \cdot 0.5 + 2 \cdot 0.3.$$

Compute term-by-term:

$$0 \cdot 0.2 = 0, \quad 1 \cdot 0.5 = 0.5, \quad 2 \cdot 0.3 = 0.6.$$

So

$$\mathbb{E}[X] = 0 + 0.5 + 0.6 = 1.1.$$

Now take $g(x) = x^2$. LOTUS says:

$$\mathbb{E}[X^2] = \sum_x x^2 p_X(x) = 0^2(0.2) + 1^2(0.5) + 2^2(0.3).$$

Compute:

$$0^2(0.2) = 0, \quad 1^2(0.5) = 0.5, \quad 2^2(0.3) = 4(0.3) = 1.2,$$

hence

$$\mathbb{E}[X^2] = 0 + 0.5 + 1.2 = 1.7.$$

These two numbers, $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$, will soon give us the variance.

2.5 Example 2 (continuous): uniform distribution

Let $X \sim \text{Unif}(0, 4)$, so

$$f_X(x) = \frac{1}{4}, \quad 0 < x < 4, \qquad f_X(x) = 0 \text{ otherwise.}$$

By definition,

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_0^4 x \cdot \frac{1}{4}\, dx.$$

Factor out $\frac{1}{4}$:

$$\mathbb{E}[X] = \frac{1}{4}\int_0^4 x\, dx = \frac{1}{4}\left[\frac{x^2}{2}\right]_0^4 = \frac{1}{4}\cdot\frac{16}{2} = \frac{1}{4}\cdot 8 = 2.$$

Now define $Y = 3X + 1$. You *could* compute $\mathbb{E}[Y]$ via an integral, but linearity (next section) makes it immediate:

$$\mathbb{E}[3X + 1] = 3\mathbb{E}[X] + 1 = 3 \cdot 2 + 1 = 7.$$

3 Linearity of expectation

Linearity is one of the most important tools in probability. It means that for linear expressions, expectation behaves like ordinary algebra. The key point is that **no independence assumption is needed.**

Theorem

Linearity of expectation (with full proof). Let X and Y be random variables such that $\mathbb{E}[|X|]$ and $\mathbb{E}[|Y|]$ are finite. Let a, b, c be constants. Then

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

Proof (discrete case). Assume (X, Y) is discrete with joint PMF $p(x, y)$. Then, by definition,

$$\mathbb{E}[aX + bY + c] = \sum_x \sum_y (ax + by + c)\, p(x, y).$$

Distribute the sum:

$$\sum_{x,y}(ax + by + c)p(x, y) = a\sum_{x,y} xp(x, y) + b\sum_{x,y} yp(x, y) + c\sum_{x,y} p(x, y).$$

Now observe:

$$\sum_{x,y} xp(x,y) = \sum_x x \left(\sum_y p(x,y) \right) = \sum_x x\, p_X(x) = \mathbb{E}[X],$$

and similarly $\sum_{x,y} yp(x,y) = \mathbb{E}[Y]$. Also $\sum_{x,y} p(x,y) = 1$. Therefore,

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

Proof (continuous case). Assume (X, Y) is a continuous random vector with joint PDF $f(x, y)$ such that $\mathbb{E}[|X|] < \infty$ and $\mathbb{E}[|Y|] < \infty$ (so all integrals below are well-defined and finite). By definition of expectation for a function of (X, Y),

$$\mathbb{E}[aX + bY + c] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (ax + by + c)\, f(x, y)\, dx\, dy.$$

Using linearity of the double integral and factoring out constants,

$$\mathbb{E}[aX + bY + c] = a \iint x f(x, y)\, dx\, dy + b \iint y f(x, y)\, dx\, dy + c \iint f(x, y)\, dx\, dy.$$

The last integral equals 1 because f is a joint density:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1.$$

For the first term, integrate out y to obtain the marginal density of X:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy.$$

Then (by Tonelli/Fubini, justified by the integrability assumption),

$$\iint x f(x, y)\, dx\, dy = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x f(x, y)\, dy \right] dx = \int_{-\infty}^{\infty} x \left[\int_{-\infty}^{\infty} f(x, y)\, dy \right] dx = \int_{-\infty}^{\infty} x f_X(x)\, dx = \mathbb{E}[X].$$

Similarly, define the marginal density of Y:

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx,$$

and compute

$$\iint y f(x, y)\, dx\, dy = \int_{-\infty}^{\infty} y f_Y(y)\, dy = \mathbb{E}[Y].$$

Substituting these results back,

$$\mathbb{E}[aX + bY + c] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y] + c \cdot 1 = a\mathbb{E}[X] + b\mathbb{E}[Y] + c.$$

□

4 Variance and standard deviation

4.1 Motivation: center vs. spread

Two distributions can have the same mean but very different variability. Variance measures how far X tends to be from its mean. Because positive and negative deviations cancel, we square them:

$$(X-\mu)^2.$$

Variance is the expected squared deviation; standard deviation is its square root, returning to the original units.

Theorem

Variance identity (with full proof). Assume $\mu = \mathbb{E}[X]$ exists and $\mathbb{E}[X^2] < \infty$. Then

$$\mathrm{Var}(X) = \mathbb{E}\big[(X-\mu)^2\big] = \mathbb{E}[X^2] - \mu^2.$$

Proof. Expand:

$$(X-\mu)^2 = X^2 - 2\mu X + \mu^2.$$

Take expectation and use linearity:

$$\mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mu^2.$$

But $\mathbb{E}[X] = \mu$, so

$$\mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2] - 2\mu^2 + \mu^2 = \mathbb{E}[X^2] - \mu^2.$$

□

4.2 Example: compute mean, variance, and SD (step by step)

Use the discrete example from earlier:

$$\mu = \mathbb{E}[X] = 1.1, \qquad \mathbb{E}[X^2] = 1.7.$$

Then

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mu^2 = 1.7 - (1.1)^2.$$

Compute $(1.1)^2 = 1.21$, so

$$\mathrm{Var}(X) = 1.7 - 1.21 = 0.49, \qquad \mathrm{SD}(X) = \sqrt{0.49} = 0.7.$$

Interpretation: values of X typically differ from the mean by about 0.7 (in the units of X).

Theorem

When variance is zero (with full proof). If $\mathrm{Var}(X) = 0$, then $X = \mathbb{E}[X]$ with probability 1.

Proof. Let $\mu = \mathbb{E}[X]$. Then $(X-\mu)^2 \geq 0$ always and

$$\mathbb{E}[(X-\mu)^2] = \mathrm{Var}(X) = 0.$$

A nonnegative random variable can have expectation 0 only if it equals 0 almost surely. Therefore $(X-\mu)^2 = 0$ a.s., hence $X = \mu$ a.s. □

5 Tail bounds: Markov and Chebyshev

5.1 Why inequalities are useful

Sometimes you do not know the full distribution of X, but you know one or two moments such as $\mathbb{E}[X]$ and $\mathrm{Var}(X)$. Markov and Chebyshev give bounds on tail probabilities that hold for *every* distribution satisfying those moment conditions.

Theorem

Markov's inequality. If $X \geq 0$ almost surely and $\mathbb{E}[X] < \infty$, then for all $t > 0$,

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Proof.
Fix $t > 0$ and consider the event $\{X \geq t\}$. Let $\mathbf{1}_{\{X \geq t\}}$ denote the indicator random variable of this event, defined by

$$\mathbf{1}_{\{X \geq t\}} = \begin{cases} 1, & X \geq t, \\ 0, & X < t. \end{cases}$$

On the event $\{X \geq t\}$, we have $X \geq t$, and thus

$$X\,\mathbf{1}_{\{X \geq t\}} \geq t\,\mathbf{1}_{\{X \geq t\}}.$$

On the complement $\{X < t\}$, both sides are equal to 0. Hence, the inequality holds pointwise:

$$X\,\mathbf{1}_{\{X \geq t\}} \geq t\,\mathbf{1}_{\{X \geq t\}} \quad \text{almost surely.}$$

Taking expectations of both sides yields

$$\mathbb{E}\big[X\,\mathbf{1}_{\{X \geq t\}}\big] \geq t\,\mathbb{E}\big[\mathbf{1}_{\{X \geq t\}}\big].$$

Since $\mathbb{E}[\mathbf{1}_{\{X \geq t\}}] = \mathbb{P}(X \geq t)$, this becomes

$$\mathbb{E}\big[X\,\mathbf{1}_{\{X \geq t\}}\big] \geq t\,\mathbb{P}(X \geq t).$$

Moreover, because $X \geq 0$ and $0 \leq \mathbf{1}_{\{X \geq t\}} \leq 1$, we have

$$0 \leq X\,\mathbf{1}_{\{X \geq t\}} \leq X,$$

and therefore

$$\mathbb{E}\big[X\,\mathbf{1}_{\{X \geq t\}}\big] \leq \mathbb{E}[X].$$

Combining the inequalities above, we obtain

$$t\,\mathbb{P}(X \geq t) \leq \mathbb{E}[X].$$

Dividing both sides by $t > 0$ gives

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t},$$

which completes the proof. □

Theorem

Chebyshev's inequality. If $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2 < \infty$, then for all $t > 0$,

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

Proof.
We begin by recalling that the variance of X is defined as

$$\mathrm{Var}(X) = \mathbb{E}\big[(X - \mu)^2\big] = \sigma^2,$$

which is finite by assumption.

Step 1: Define a nonnegative random variable. Consider the random variable

$$Y = (X - \mu)^2.$$

Since a square is always nonnegative, we have $Y \geq 0$ almost surely. Moreover,

$$\mathbb{E}[Y] = \mathbb{E}\big[(X - \mu)^2\big] = \sigma^2 < \infty.$$

Thus, the assumptions of Markov's inequality are satisfied.

Step 2: Apply Markov's inequality. By Markov's inequality, for any $a > 0$,

$$\mathbb{P}(Y \geq a) \leq \frac{\mathbb{E}[Y]}{a}.$$

We now choose $a = t^2$ with $t > 0$. Then

$$\mathbb{P}\big((X - \mu)^2 \geq t^2\big) \leq \frac{\mathbb{E}[(X - \mu)^2]}{t^2} = \frac{\sigma^2}{t^2}.$$

Step 3: Rewrite the event. Observe that

$$(X - \mu)^2 \geq t^2 \quad \Longleftrightarrow \quad |X - \mu| \geq t.$$

Therefore,

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

This proves Chebyshev's inequality. □

Remark (intuition). Chebyshev's inequality states that the probability of a random variable deviating far from its mean is controlled by its variance. A smaller variance forces the distribution of X to concentrate more tightly around μ, regardless of the exact shape of the distribution.

6 Two random variables: covariance, correlation, and linear combinations

6.1 Covariance and correlation (interpretation first)

Variance measures spread of one variable. Covariance measures how two variables move together.

- If $\mathrm{Cov}(X,Y) > 0$, then large X tends to occur with large Y (positive association).
- If $\mathrm{Cov}(X,Y) < 0$, then large X tends to occur with small Y (negative association).
- If $\mathrm{Cov}(X,Y) = 0$, they are *uncorrelated*. This does not necessarily imply independence.

Correlation rescales covariance to lie in $[-1,1]$ and is unit-free.

Theorem

Covariance and correlation. Assume $\mathbb{E}[X^2]$ and $\mathbb{E}[Y^2]$ are finite. Let $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$. Define

$$\mathrm{Cov}(X,Y) = \mathbb{E}\big[(X-\mu_X)(Y-\mu_Y)\big].$$

Then

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mu_X\mu_Y.$$

If $\mathrm{Var}(X) > 0$ and $\mathrm{Var}(Y) > 0$, define

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} \in [-1,1].$$

6.2 Example: compute covariance and correlation step by step

Suppose (X,Y) takes values:

$$(0,0) \text{ with prob } 0.5, \qquad (2,1) \text{ with prob } 0.5.$$

First compute the means:

$$\mu_X = \mathbb{E}[X] = 0(0.5) + 2(0.5) = 1, \qquad \mu_Y = \mathbb{E}[Y] = 0(0.5) + 1(0.5) = 0.5.$$

Next compute $\mathbb{E}[XY]$:

$$\mathbb{E}[XY] = 0 \cdot 0 \cdot 0.5 + 2 \cdot 1 \cdot 0.5 = 1.$$

Thus

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mu_X\mu_Y = 1 - (1)(0.5) = 0.5.$$

To compute correlation we need variances. Compute $\mathbb{E}[X^2]$ and $\mathbb{E}[Y^2]$:

$$\mathbb{E}[X^2] = 0^2(0.5) + 2^2(0.5) = 0 + 4(0.5) = 2, \quad \mathrm{Var}(X) = 2 - 1^2 = 1,$$

$$\mathbb{E}[Y^2] = 0^2(0.5) + 1^2(0.5) = 0 + 0.5 = 0.5, \quad \mathrm{Var}(Y) = 0.5 - (0.5)^2 = 0.5 - 0.25 = 0.25.$$

Therefore

$$\mathrm{Corr}(X,Y) = \frac{0.5}{\sqrt{1}\sqrt{0.25}} = \frac{0.5}{0.5} = 1.$$

Interpretation: in this example Y is a perfect linear function of X (indeed $Y = \frac{1}{2}X$).

6.3 Variance of a linear combination

When we add random variables, spread depends not only on individual variances but also on dependence.

Theorem

Variance of a linear combination. Assume that X and Y have finite second moments. For constants $a, b \in \mathbb{R}$,

$$\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X, Y).$$

In particular, if X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$ and

$$\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y).$$

Proof.
Recall the definition of variance:

$$\mathrm{Var}(Z) = \mathbb{E}\big[(Z - \mathbb{E}[Z])^2\big].$$

Let $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$. By linearity of expectation,

$$\mathbb{E}[aX + bY] = a\mu_X + b\mu_Y.$$

Step 1: Center the random variable. We write

$$aX + bY - \mathbb{E}[aX + bY] = a(X - \mu_X) + b(Y - \mu_Y).$$

Hence,

$$\mathrm{Var}(aX + bY) = \mathbb{E}\Big[\big(a(X - \mu_X) + b(Y - \mu_Y)\big)^2\Big].$$

Step 2: Expand the square. Using the algebraic identity $(u + v)^2 = u^2 + v^2 + 2uv$, we obtain

$$\big(a(X - \mu_X) + b(Y - \mu_Y)\big)^2 = a^2(X - \mu_X)^2 + b^2(Y - \mu_Y)^2 + 2ab(X - \mu_X)(Y - \mu_Y).$$

Step 3: Take expectations term by term. Since expectation is linear,

$$\begin{aligned}\mathrm{Var}(aX + bY) = a^2\mathbb{E}\big[(X - \mu_X)^2\big] + b^2\mathbb{E}\big[(Y - \mu_Y)^2\big] \\ + 2ab\,\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]\,.\end{aligned}$$

Step 4: Identify variance and covariance. By definition,

$$\mathrm{Var}(X) = \mathbb{E}\big[(X - \mu_X)^2\big]\,, \qquad \mathrm{Var}(Y) = \mathbb{E}\big[(Y - \mu_Y)^2\big]\,,$$

and

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

Substituting these into the expression above yields

$$\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X, Y).$$

Step 5: Independent case. If X and Y are independent, then

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y],$$

which implies

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

Therefore, when X and Y are independent,

$$\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y).$$

□

6.4 Example: same marginal spreads, different dependence

Suppose $\mathrm{Var}(X) = 4$, $\mathrm{Var}(Y) = 9$, and $\mathrm{Cov}(X, Y) = 3$. Then

$$\mathrm{Var}(X + Y) = 4 + 9 + 2(3) = 19.$$

If instead X and Y*were* independent (so covariance 0), then $\mathrm{Var}(X + Y) = 13$. This shows: the uncertainty of a sum depends strongly on dependence.

7 Conditional expectation: simplifying calculations and prediction

7.1 What is $\mathbb{E}[Y \mid X]$?

For a fixed value $X = x$, the conditional expectation $\mathbb{E}[Y \mid X = x]$ is a number: the mean of the conditional distribution of Y given $X = x$. If we let X vary randomly, then $\mathbb{E}[Y \mid X]$ becomes a *random variable* (a function of X). This is why identities like $\mathbb{E}[\mathbb{E}[Y \mid X]]$ make sense: we are taking the expectation of a random variable.

Theorem

Law of total expectation and total variance. Assume all expectations below exist.

(1) **Law of total expectation:**

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y \mid X]]\,.$$

Proof. We show the discrete case (the continuous case is analogous with integrals). Assume X is discrete. Then,

$$\mathbb{E}[\mathbb{E}[Y \mid X]] = \sum_x \mathbb{E}[Y \mid X = x]\mathbb{P}(X = x).$$

But by the definition of conditional expectation in the discrete case,

$$\mathbb{E}[Y \mid X = x] = \sum_y y\,\mathbb{P}(Y = y \mid X = x).$$

Substitute this into the previous sum:

$$\sum_x \left(\sum_y y\,\mathbb{P}(Y = y \mid X = x) \right) \mathbb{P}(X = x) = \sum_x \sum_y y\,\mathbb{P}(Y = y \mid X = x)\mathbb{P}(X = x).$$

Use $\mathbb{P}(Y = y \mid X = x)\mathbb{P}(X = x) = \mathbb{P}(X = x, Y = y)$:

$$\sum_x \sum_y y\, \mathbb{P}(X = x, Y = y) = \sum_y y \left(\sum_x \mathbb{P}(X = x, Y = y) \right) = \sum_y y\, \mathbb{P}(Y = y) = \mathbb{E}[Y].$$

So $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y \mid X]]$. □

(2) **Law of total variance:**

$$\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y \mid X)] + \mathrm{Var}(\mathbb{E}[Y \mid X])\,.$$

Proof. Start from $\mathrm{Var}(Y) = \mathbb{E}[Y^2] - \{\mathbb{E}[Y]\}^2$. Apply the law of total expectation to Y^2:

$$\mathbb{E}[Y^2] = \mathbb{E}\big[\mathbb{E}[Y^2 \mid X]\big]\,.$$

Also $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y \mid X]]$. Hence

$$\mathrm{Var}(Y) = \mathbb{E}\big[\mathbb{E}[Y^2 \mid X]\big] - (\mathbb{E}[\mathbb{E}[Y \mid X]])^2\,.$$

Now use the identity (true for each fixed X):

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

Take expectations:

$$\mathbb{E}[\mathrm{Var}(Y \mid X)] = \mathbb{E}\big[\mathbb{E}[Y^2 \mid X]\big] - \mathbb{E}\big[\{\mathbb{E}[Y \mid X]\}^2\big]\,.$$

Also,

$$\mathrm{Var}(\mathbb{E}[Y \mid X]) = \mathbb{E}\big[\{\mathbb{E}[Y \mid X]\}^2\big] - (\mathbb{E}[\mathbb{E}[Y \mid X]])^2\,.$$

Add the last two equations: the $\mathbb{E}[\{\mathbb{E}[Y \mid X]\}^2]$ terms cancel, leaving

$$\mathbb{E}[\mathrm{Var}(Y \mid X)] + \mathrm{Var}(\mathbb{E}[Y \mid X]) = \mathbb{E}\big[\mathbb{E}[Y^2 \mid X]\big] - (\mathbb{E}[\mathbb{E}[Y \mid X]])^2 = \mathrm{Var}(Y).$$

□

7.2 Example: mixture model with full calculation

Let $Y \in \{0, 1\}$ with $\mathbb{P}(Y = 1) = 0.3$ and $\mathbb{P}(Y = 0) = 0.7$. Given Y, define

$$X \mid (Y = 0) \sim \mathrm{Unif}(0, 2), \qquad X \mid (Y = 1) \sim \mathrm{Unif}(2, 4).$$

Step 1: conditional means. For a uniform $\mathrm{Unif}(a, b)$, the mean is $\frac{a+b}{2}$. Thus

$$\mathbb{E}[X \mid Y = 0] = \frac{0 + 2}{2} = 1, \qquad \mathbb{E}[X \mid Y = 1] = \frac{2 + 4}{2} = 3.$$

Step 2: unconditional mean via total expectation.

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]] = 1 \cdot 0.7 + 3 \cdot 0.3 = 0.7 + 0.9 = 1.6.$$

Step 3: conditional variances. For $\mathrm{Unif}(a, b)$, $\mathrm{Var} = \frac{(b-a)^2}{12}$. Hence

$$\mathrm{Var}(X \mid Y = 0) = \frac{(2 - 0)^2}{12} = \frac{4}{12} = \frac{1}{3}, \quad \mathrm{Var}(X \mid Y = 1) = \frac{(4 - 2)^2}{12} = \frac{4}{12} = \frac{1}{3}.$$

So

$$\mathbb{E}[\mathrm{Var}(X \mid Y)] = \frac{1}{3} \cdot 0.7 + \frac{1}{3} \cdot 0.3 = \frac{1}{3}.$$

Step 4: between-group variance. The random variable $\mathbb{E}[X \mid Y]$ takes values 1 (w.p. 0.7) and 3 (w.p. 0.3). Compute

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = 1.6, \quad \mathbb{E}\left[(\mathbb{E}[X \mid Y])^2\right] = 1^2(0.7) + 3^2(0.3) = 0.7 + 2.7 = 3.4.$$

Thus

$$\mathrm{Var}(\mathbb{E}[X \mid Y]) = 3.4 - (1.6)^2 = 3.4 - 2.56 = 0.84.$$

Step 5: total variance.

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbb{E}[X \mid Y]) = \frac{1}{3} + 0.84 \approx 1.1733.$$

Interpretation: total variance = within-group variance + between-group variance.

8 Moment generating functions (MGFs)

8.1 What an MGF is (and what it is good for)

The MGF is a function that (when it exists near 0) encodes all moments of X. It is useful because:

- derivatives at 0 produce moments (mean, second moment, etc.),
- MGFs multiply for sums of independent variables, making them powerful for studying sums,
- in many common families, the MGF uniquely identifies the distribution.

Theorem

MGF and two key properties (with proofs). Let X be a random variable. The MGF is

$$M_X(t) = \mathbb{E}[e^{tX}],$$

defined for those t where the expectation is finite. If $M_X(t)$ exists on an open interval containing 0, then

$$M_X'(0) = \mathbb{E}[X], \qquad M_X''(0) = \mathbb{E}[X^2], \qquad M_X^{(r)}(0) = \mathbb{E}[X^r].$$

(1) **Linear transform.** If $Y = a + bX$, then $M_Y(t) = e^{at} M_X(bt)$.

Proof.

$$M_Y(t) = \mathbb{E}[e^{t(a+bX)}] = \mathbb{E}[e^{ta} e^{tbX}] = e^{at}\mathbb{E}[e^{(bt)X}] = e^{at} M_X(bt).$$

□

(2) **Sum of independent variables.** If X and Z are independent, then $M_{X+Z}(t) = M_X(t)M_Z(t)$.

Proof.

$$M_{X+Z}(t) = \mathbb{E}[e^{t(X+Z)}] = \mathbb{E}[e^{tX} e^{tZ}].$$

By independence, $\mathbb{E}[UV] = \mathbb{E}[U]\mathbb{E}[V]$ for $U = e^{tX}$ and $V = e^{tZ}$, hence

$$\mathbb{E}[e^{tX} e^{tZ}] = \mathbb{E}[e^{tX}]\,\mathbb{E}[e^{tZ}] = M_X(t)M_Z(t).$$

□

8.2 Example: Poisson MGF $\Rightarrow$ mean and variance

If $X \sim \text{Poisson}(\lambda)$, Rice gives (or you can derive) the MGF:

$$M_X(t) = \exp\{\lambda(e^t - 1)\}.$$

Differentiate once:

$$M_X'(t) = \lambda e^t \exp\{\lambda(e^t - 1)\}.$$

Evaluate at $t = 0$:

$$\mathbb{E}[X] = M_X'(0) = \lambda e^0 \exp\{\lambda(e^0 - 1)\} = \lambda \cdot 1 \cdot e^0 = \lambda.$$

Differentiate again (product rule). It is convenient to write $M_X'(t) = \lambda e^t M_X(t)$, so

$$M_X''(t) = \lambda e^t M_X(t) + \lambda e^t M_X'(t) = \lambda e^t M_X(t) + \lambda e^t(\lambda e^t M_X(t)) = \lambda e^t M_X(t) + \lambda^2 e^{2t} M_X(t).$$

Evaluate at 0:

$$\mathbb{E}[X^2] = M_X''(0) = \lambda \cdot 1 \cdot 1 + \lambda^2 \cdot 1 \cdot 1 = \lambda + \lambda^2.$$

Therefore

$$\text{Var}(X) = \mathbb{E}[X^2] - \{\mathbb{E}[X]\}^2 = (\lambda + \lambda^2) - \lambda^2 = \lambda.$$

8.3 Characteristic function (brief remark)

Even when the MGF fails to exist (heavy tails), the characteristic function always exists:

$$\varphi_X(t) = \mathbb{E}[e^{itX}], \qquad t \in \mathbb{R}.$$

Rice introduces this mainly as a robust alternative to MGFs.

This is a personal study purposed notes, based on a lecture slide given by Prof. Ana-Maria Staicu in ST 502, NC state university and Rice (3rd ed.) Chapter 4.