Foundations

A mathematical framework for probability is defined by the **probability space** $(\Omega, \mathcal{F}, P)$:

- Ω (**Sample Space**): The set of all possible outcomes ω of a random experiment.
- $\mathcal{F}$ (σ-**Algebra**): A collection of subsets of Ω, called measurable events, that is closed under complements and countable unions.
- P (**Probability Measure**): A mapping $P : \mathcal{F} \to [0, 1]$ that assigns probabilities to events, with $P(\Omega) = 1$ and countable additivity over pairwise disjoint events.

Random Variable

Despite the name, a random variable is not an ordinary variable, but a function mapping outcomes from the sample space to the real numbers:

$$X : \Omega \to \mathbb{R}.$$

For any outcome $\omega \in \Omega$, $X(\omega) \in \mathbb{R}$ assigns a numerical value to that outcome. For example, if X counts the number of heads in a sequence of coin flips, then $X(\text{THTHT}) = 2$.
For a more thorough coverage of these topics, I encourage you to take a look at the following video playlists:

- Probably Optimal – Probability Course
- Carlos Fernandez-Granda – Probability and Statistics for Data Science

Conditional Probability

For two events A and B in a sample space Ω, the conditional probability of A given B is defined as:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad \text{when } P(B) > 0.$$

In other words, we restrict our attention to outcomes in B and determine what fraction of their probability corresponds to outcomes that also belong to A.

Cumulative Distribution Function (CDF)

The CDF of a random variable X gives the probability that X takes a value less than or equal to x:

$$F_X(x) = P(X \leq x), \quad \text{for all } x \in \mathbb{R}$$

Key Properties:

- **Interval Probabilities:** For any $a \leq b$, the probability that X falls in $(a, b]$ is given by:

$$P(a < X \leq b) = F_X(b) - F_X(a)$$

- **Monotonicity:** $F_X(x)$ is non-decreasing and takes values in $[0, 1]$.
- **Limits:** $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$.
- **Right-Continuity:** $F_X(x)$ is right-continuous.

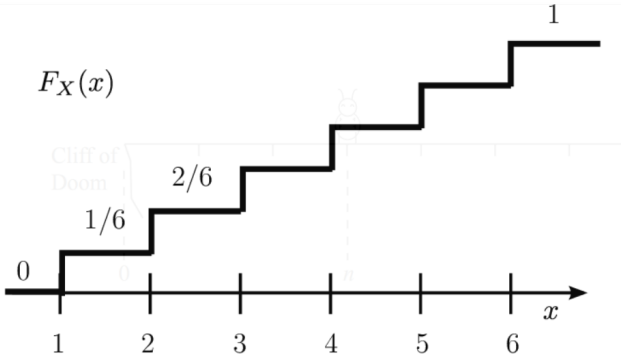


Figure 1: CDF of a fair die.

Probability Mass Function (PMF)

For a **discrete** random variable X with range $R_X = \{x_1, x_2, x_3, \dots\}$ (finite or countably infinite), the PMF gives the probability that X takes a specific value:

$$p_X(x_k) = P(X = x_k), \quad k = 1, 2, 3, \dots$$

The probabilities satisfy:

$$\sum_{x \in R_X} p_X(x) = 1$$

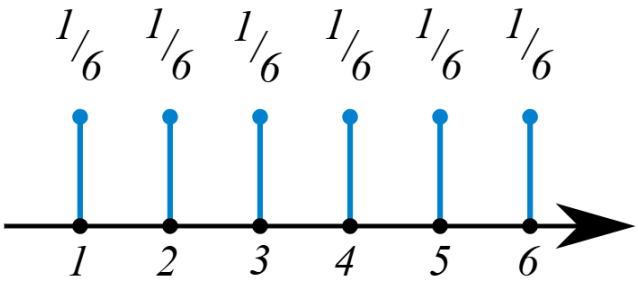


Figure 2: PMF of a fair die.

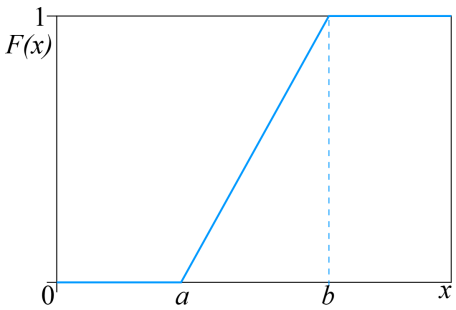
Probability Density Function (PDF)

For a random variable X with an absolutely continuous CDF $F_X(x)$, the probability density function $f_X(x)$ satisfies

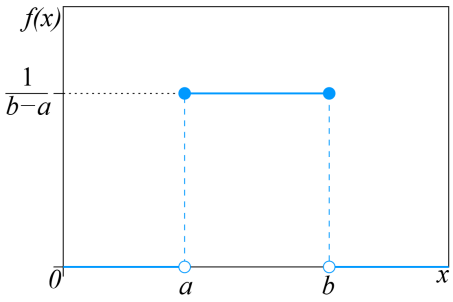
$$f_X(x) = F'_X(x)$$

wherever the derivative exists (and hence almost everywhere). The CDF can be recovered by integrating the PDF:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt.$$



(a) CDF of a uniform distribution



(b) PDF of a uniform distribution

Independence

Two events A and B are independent if the occurrence of one does not affect the probability of the other. Formally,

$$P(A \cap B) = P(A)P(B).$$

When $P(B) > 0$, this is equivalent to

$$P(A \mid B) = P(A).$$

Two random variables X and Y are independent if their joint distribution factorizes into the product of their marginal distributions. In particular:

- **Discrete:**
$$P(X = x, Y = y) = P(X = x)P(Y = y), \quad \forall x, y$$
- **Continuous:**
$$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$
almost everywhere.

Common Mistake: Disjoint vs. Independent Events

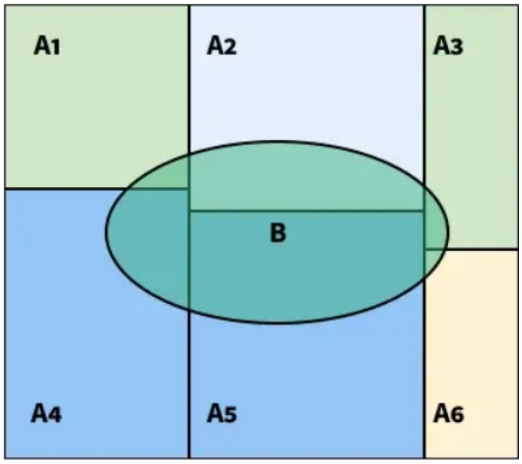
Concept	Meaning	Key Formulas
Disjoint	A and B cannot occur at the same time.	$A \cap B = \emptyset \implies P(A \cap B) = 0$ $P(A \cup B) = P(A) + P(B)$
Independent	Knowing whether one event occurred provides no information about the other.	$P(A \cap B) = P(A)P(B)$ $P(A \mid B) = P(A) \quad \text{for } P(B) > 0$

Key Distinction: If two events with positive probabilities ($P(A) > 0$, $P(B) > 0$) are **disjoint**, knowing that B occurred tells you that A cannot occur ($P(A \mid B) = 0$). Thus, disjoint events with non-zero probability are **never** independent!

Law of Total Probability

If $A_1, A_2, A_3, \ldots$ form a **partition** of the sample space Ω (i.e., they are pairwise disjoint and their union is Ω), then for any event B:

$$P(B) = \sum_i P(B \cap A_i) = \sum_i P(B \mid A_i) P(A_i)$$



Probability Distributions

Bernoulli Distribution

A random variable $X \sim \text{Bernoulli}(p)$ models a single trial with binary outcomes (success/failure), where p is the probability of success ($0 \leq p \leq 1$):

$$p_X(x) = \begin{cases} p & \text{for } x = 1 \\ 1 - p & \text{for } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

Geometric Distribution

A random variable $X \sim \text{Geometric}(p)$ models the number of independent trials required to obtain the **first success**, where each trial has success probability p ($0 < p \leq 1$):

$$p_X(k) = \begin{cases} p(1-p)^{k-1} & \text{for } k = 1, 2, 3, \ldots \\ 0 & \text{otherwise} \end{cases}$$

Binomial Distribution

A random variable $X \sim \text{Binomial}(n, p)$ models the total number of successes in n independent Bernoulli trials, each with success probability p ($n \geq 1$, $0 \leq p \leq 1$):

$$p_X(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{for } k = 0, 1, 2, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

■ **Example:** Let $X \sim \text{Binomial}(n, p)$ and $Y \sim \text{Binomial}(m, p)$ be independent random variables. Define $Z = X + Y$. Find the PMF of Z.

The Intuitive Approach

Recall that a Binomial(n, p) random variable counts the number of successes in n independent Bernoulli(p) trials.

- X represents the number of successes in n independent flips of a coin with success probability p.
- Y represents the number of successes in m independent flips of the same coin.

Since X and Y are independent, adding them ($Z = X + Y$) counts the total number of successes across $n + m$ independent flips of the exact same coin.

Therefore, by physical intuition:

$$\mathbf{Z} \sim \text{Binomial}(\mathbf{n} + \mathbf{m}, \mathbf{p})$$

$$\implies P_Z(k) = \binom{n+m}{k} p^k (1-p)^{n+m-k}, \quad \text{for } k = 0, 1, \ldots, n+m$$

The Formal Derivation

We can also prove this algebraically using conditioning and the Law of Total Probability:

$$\begin{aligned}
P_Z(k) &= P(X + Y = k) \\
&= \sum_{i=0}^{n} P(X + Y = k \mid X = i) P(X = i) && \text{(Law of Total Probability)} \\
&= \sum_{i=0}^{n} P(Y = k - i \mid X = i) P(X = i) \\
&= \sum_{i=0}^{n} P(Y = k - i) P(X = i) && \text{(Independence of } X \text{ and } Y\text{)} \\
&= \sum_{i=0}^{n} \binom{m}{k-i} p^{k-i} (1-p)^{m-(k-i)} \cdot \binom{n}{i} p^i (1-p)^{n-i} && \text{(Binomial PMFs)} \\
&= p^k (1-p)^{m+n-k} \sum_{i=0}^{n} \binom{m}{k-i} \binom{n}{i}
\end{aligned}$$

Since $\sum_{i=0}^{n} \binom{m}{k-i}\binom{n}{i} = \binom{m+n}{k}$, we get the same result:

$$P_Z(k) = \binom{m+n}{k} p^k (1-p)^{m+n-k} \implies \mathbf{Z} \sim \text{Binomial}(\mathbf{m} + \mathbf{n}, \mathbf{p})$$

Expected Value

The expected value $\mathbb{E}[X]$ represents the probability-weighted average value of a random variable.

- **Discrete Random Variable:** For X with range $R_X = \{x_1, x_2, x_3, \dots\}$:

$$\mathbb{E}[X] = \sum_{x_k \in R_X} x_k P(X = x_k) = \sum_{x_k \in R_X} x_k p_X(x_k)$$

- **Continuous Random Variable:** For X with PDF $f_X(x)$:

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

Expectation is a linear operator, regardless of whether the random variables involved are independent or dependent. When the expectations exist:

- **Scaling and Shift:** $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$, for all $a, b \in \mathbb{R}$.
- **Sum of Random Variables:**

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

To emphasize that linearity of expectation does not require independence, consider the following setup.

■ **Example:** Let X be the outcome of rolling a fair six-sided die, so $X \in \{1, 2, 3, 4, 5, 6\}$, with equal probability $1/6$.

Define Y as an indicator variable based on X:

$$Y = \begin{cases} 0 & \text{if } X \text{ is odd } (X \in \{1, 3, 5\}), \\ 1 & \text{if } X \text{ is even } (X \in \{2, 4, 6\}). \end{cases}$$

Notice that Y is entirely determined by X.

Step 1: Compute individual expectations

- $\mathbb{E}[X] = \sum_{k=1}^{6} k \cdot \frac{1}{6} = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3.5$
- $\mathbb{E}[Y] = 0 \cdot P(Y = 0) + 1 \cdot P(Y = 1) = 0 \cdot \frac{3}{6} + 1 \cdot \frac{3}{6} = 0.5$

Sum of individual expectations:

$$\mathbb{E}[X] + \mathbb{E}[Y] = 3.5 + 0.5 = 4.$$

Step 2: Compute $\mathbb{E}[X+Y]$ directly from the six possible outcomes
Let $Z = X + Y$. The possible values of X, Y, and Z are:

X	Y	$Z = X + Y$	Probability
1	0	1	1/6
2	1	3	1/6
3	0	3	1/6
4	1	5	1/6
5	0	5	1/6
6	1	7	1/6

Calculating the expectation directly:

$$\mathbb{E}[X+Y] = \frac{1+3+3+5+5+7}{6} = \frac{24}{6} = 4.$$

Conclusion:

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y] = 4.$$

Even though Y is entirely determined by X, linearity of expectation still holds!

Functions of Random Variables & LOTUS

If X is a random variable and $Y = g(X)$, then Y is also a random variable, with range $R_Y = \{g(x) \mid x \in R_X\}$. To find the PMF of Y, sum the probabilities over all inputs x that map to y:

$$p_Y(y) = P(Y = y) = P(g(X) = y) = \sum_{x:g(x)=y} p_X(x)$$

Law of the Unconscious Statistician (LOTUS):
LOTUS allows us to calculate the expected value of a function of random variables directly from their PMF/PDF, without first computing the distribution of the transformed variable.

- **Single Discrete Variable:**

$$E[g(X)] = \sum_{x \in R_X} g(x) p_X(x)$$

- **Two Discrete Variables:**

$$E[g(X,Y)] = \sum_{(x,y) \in R_{XY}} g(x,y) p_{XY}(x,y)$$

- **Continuous Case:**

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

$$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y)\, dx\, dy$$

Here, p_{XY} and f_{XY} denote the joint PMF and joint PDF, respectively. These formulas do not require X and Y to be independent.

■ **Example:** Let $X \sim \text{Uniform}(0,1)$ and let $Y = e^X$. Find:

1. The CDF of Y, $F_Y(y)$.
2. The PDF of Y, $f_Y(y)$.
3. The expected value EY.

Directly finding the PDF $f_Y(y)$ of a non-linear transformation $Y = g(X)$ can be tricky. So we start by the CDF.

Step 1: Find the CDF of Y

First, we determine the support/range of Y. Since $X \in [0,1]$, $Y = e^X$ spans $[e^0, e^1] = [1, e]$.

- For $y < 1$: $F_Y(y) = 0$
- For $y \geq e$: $F_Y(y) = 1$
- For $y \in [1, e]$:

$$\begin{aligned} F_Y(y) = P(Y \leq y) &= P(e^X \leq y) \\ &= P(X \leq \ln y) && \text{(Since } e^x \text{ is strictly monotonic)} \\ &= F_X(\ln y) = \ln y && \text{(Since } F_X(x) = x \text{ for } x \in [0,1]\text{)} \end{aligned}$$

Thus, the CDF is:

$$F_Y(y) = \begin{cases} 0 & \text{for } y < 1 \\ \ln y & \text{for } 1 \leq y < e \\ 1 & \text{for } y \geq e \end{cases}$$

Step 2: Find the PDF of Y via Differentiation

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \begin{cases} \frac{1}{y} & \text{for } 1 \leq y \leq e \\ 0 & \text{otherwise} \end{cases}$$

Step 3: Calculate EY

Method A: Direct LOTUS

$$EY = E[e^X] = \int_{-\infty}^{\infty} e^x f_X(x)\, dx = \int_0^1 e^x\, dx = [e^x]_0^1 = e - 1$$

Method B: Using the derived PDF $f_Y(y)$

$$EY = \int_{-\infty}^{\infty} y f_Y(y)\, dy = \int_1^e y \cdot \frac{1}{y}\, dy = \int_1^e 1\, dy = e - 1$$

LOTUS allows computing EY immediately from X, but the CDF method is indispensable when the density $f_Y(y)$ itself is required.

Variance & Standard Deviation

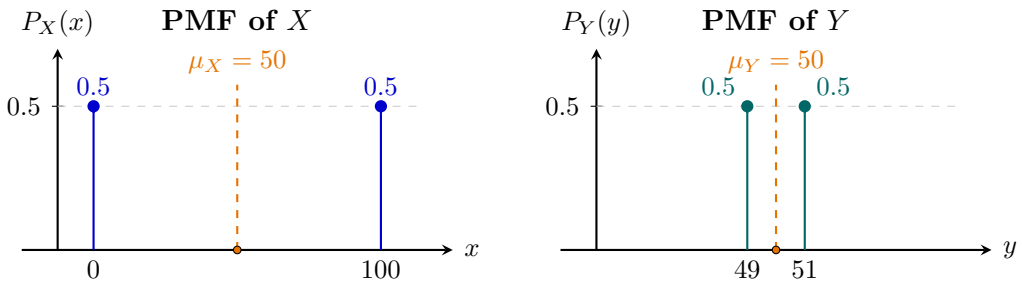
Motivating Variance:
Suppose we have two discrete random variables, X and Y, representing two financial investments with outcomes in dollars:

- X takes values 0 or 100 with equal probability: $P_X(0) = 0.5$, $P_X(100) = 0.5$.
- Y takes values 49 or 51 with equal probability: $P_Y(49) = 0.5$, $P_Y(51) = 0.5$.

Calculating their expected values:

$$EX = 0(0.5) + 100(0.5) = 50$$

$$EY = 49(0.5) + 51(0.5) = 50$$



Both distributions have the **exact same expected value**, yet they behave very differently in terms of risk. X fluctuates widely around its mean, whereas Y remains tightly clustered near 50. Expectation alone fails to capture the spread of a random variable.

Measuring Deviations from the Mean
To quantify spread, let's look at the deviation of a random variable from its mean $\mu_X = EX$:

$$\text{Deviation} = X - \mu_X$$

Can we simply take the expected value of these deviations, $E[X - \mu_X]$? Using linearity of expectation:

$$E[X - \mu_X] = EX - E[\mu_X] = \mu_X - \mu_X = 0$$

Because positive and negative deviations cancel each other out, $E[X - \mu_X]$ is always zero whenever the expectation exists.
To prevent positive and negative deviations from canceling, we can eliminate the sign. Two natural options are:

- **Absolute Deviations:** $E[|X - \mu_X|]$. This is intuitive and measures the average absolute distance from the mean.
- **Squared Deviations:** $E[(X - \mu_X)^2]$. Squaring guarantees non-negativity, penalizes larger deviations more heavily, and makes the resulting expression easier to manipulate algebraically.

Formal Definitions

The **variance** of a random variable X with mean $\mu_X = EX$ is defined as the expected squared deviation from its mean:

$$\mathrm{Var}(X) = E\left[(X - \mu_X)^2\right]$$

Units Warning (Variance vs. Standard Deviation):

Because variance squares the difference $(X - \mu_X)^2$, its units are the square of the original units. For example, if X is measured in dollars, $\mathrm{Var}(X)$ is measured in squared dollars ($\$^2$).

To obtain a measure of spread in the same units as the original random variable, we take the square root of the variance to define the standard deviation:

$$\mathrm{SD}(X) = \sigma_X = \sqrt{\mathrm{Var}(X)}$$

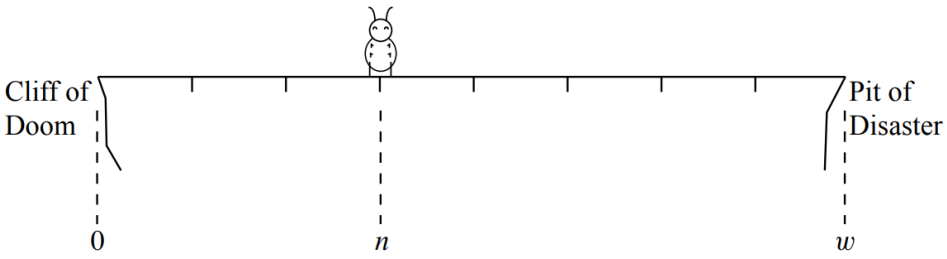
Computational Formula for Variance

In practice, evaluating $E[(X - \mu_X)^2]$ directly can be tedious. We expand the expression and use linearity of expectation:

$$\begin{aligned}
\mathrm{Var}(X) &= E\left[(X - \mu_X)^2\right] \\
&= E\left[X^2 - 2\mu_X X + \mu_X^2\right] \\
&= E[X^2] - 2\mu_X E[X] + E[\mu_X^2] \quad \text{(Linearity of Expectation)} \\
&= E[X^2] - 2\mu_X^2 + \mu_X^2 \quad \text{(Since } \mu_X \text{ is a constant)} \\
&= E[X^2] - \mu_X^2.
\end{aligned}$$

$$\boxed{\mathrm{Var}(X) = E[X^2] - (EX)^2}$$

"A Big Island" Random Walk Problem[1]



In this problem, **Stencil** (a bug) performs a 1D random walk on an island of width w inches. He starts at position n (where $0 \leq n \leq w$):

- **Position** 0**:** The *Cliff of Doom* (falling off the left edge).
- **Position** w**:** The *Pit of Disaster* (falling off the right edge).
- At each step, Stencil hops 1 inch **left** with probability 1/2 or 1 inch **right** with probability 1/2.

We analyze two key metrics for this random walk:

1. **Absorption Probability (R_n):** The probability that Stencil reaches the Pit of Disaster (w) before falling off the Cliff of Doom (0), given he starts at position n.
2. **Life Expectancy (X_n):** The expected number of total hops Stencil takes before hitting either boundary, starting at position n.

Part 1: Absorption Probability (R_n)

Let R_n be the probability of reaching position w before position 0 starting from position n.

1. Boundary Conditions:

- $R_0 = 0$ (Starts at 0, falls off Cliff of Doom immediately).
- $R_w = 1$ (Starts at w, falls into Pit of Disaster immediately).

2. Setting Up the Linear Recurrence: For $0 < n < w$, conditioning on the first hop (Law of Total Probability):

$$R_n = \frac{1}{2}R_{n-1} + \frac{1}{2}R_{n+1}$$

Rearranging into standard second-order linear recurrence form:

$$R_{n+1} - 2R_n + R_{n-1} = 0$$

3. Solving the Recurrence: The characteristic equation is $x^2 - 2x + 1 = 0$, which yields a **repeated root at** $x = 1$. Thus, the general solution takes the form:

$$R_n = (a + bn)(1)^n = a + bn$$

[1]Source: MIT 6.042J

Applying the boundary conditions:

- $R_0 = 0 \implies a = 0$
- $R_w = 1 \implies a + bw = 1 \implies b = \frac{1}{w}$

$$\boxed{\mathbf{R_n = \frac{n}{w}}}$$

Intuition: The probability of reaching the right edge is purely proportional to Stencil's starting distance from the left edge!

Part 2: Life Expectancy (X_n)

Let X_n be the expected number of total hops before hitting either edge when starting from position n.

1. Boundary Conditions:

$$X_0 = 0, \quad X_w = 0 \quad \text{(0 hops remaining at either boundary)}$$

2. Setting Up the Recurrence: By the Law of Total Expectation and adding 1 for the current hop:

$$X_n = 1 + \frac{1}{2}X_{n-1} + \frac{1}{2}X_{n+1} \quad \text{for } 0 < n < w$$

Rearranging into standard form:

$$X_{n+1} - 2X_n + X_{n-1} = -2$$

3. Solving the Recurrence:

- **Homogeneous Solution:** As before with repeated root $x = 1$, $X_n^{(h)} = a + bn$.
- **Particular Solution:** Because the constant term -2 conflicts with the homogeneous roots, we try $X_n^{(p)} = en^2$. Substituting en^2 into $e(n+1)^2 - 2en^2 + e(n-1)^2 = -2$ gives:

  $$e(n^2 + 2n + 1 - 2n^2 + n^2 - 2n + 1) = 2e = -2 \implies e = -1$$

  So $X_n^{(p)} = -n^2$.

Combining homogeneous and particular solutions:

$$X_n = a + bn - n^2$$

Applying boundary conditions $X_0 = 0$ and $X_w = 0$:

- $X_0 = 0 \implies a = 0$
- $X_w = 0 \implies bw - w^2 = 0 \implies b = w$

$$\boxed{\mathbf{X_n = wn - n^2 = n(w - n)}}$$

Intuition: Stencil's expected lifespan is simply the **product of his distances to the two edges** (n and $w - n$)! For example, if $n = 4$ and $w = 12$, $X_4 = 4 \times 8 = 32$ hops.

Part 3: The Infinite Plateau Paradox

What if there is a Cliff of Doom at position 0, but the island extends infinitely to the right?

1. Probability of Eventually Falling Off:
Let T be the number of hops until Stencil reaches position 0, starting from position $n = 1$. On a finite island with boundaries at 0 and w, the probability of reaching w before 0 is $R_1 = 1/w$. For any positive integer w, a walk that never reaches 0 must either reach w before 0 or remain forever in the finite set $\{1, \ldots, w-1\}$. The latter event has probability zero. Therefore,

$$P(T = \infty) \leq P(\text{reach } w \text{ before } 0) = \frac{1}{w}.$$

Since this holds for every positive integer w, letting $w \to \infty$ gives

$$P(T = \infty) = 0, \qquad P(T < \infty) = 1.$$

2. Expected Lifespan on an Infinite Plateau:
For a finite island, the expected number of hops starting at $n = 1$ is

$$X_1 = w - 1.$$

As $w \to \infty$, this expectation grows without bound. In fact, the expected hitting time on the infinite plateau is infinite.

The Resolution:
Stencil is almost surely guaranteed to eventually fall off the cliff, yet his expected lifespan is infinite. There is no contradiction: although the probability of surviving for a very long time becomes small, it decays slowly enough that the expected number of hops diverges.

Covariance & Correlation Coefficient

When dealing with two random variables X and Y, we often want to measure how they co-vary (move together).

1. Covariance
The **covariance** measures how X and Y vary together:

$$\mathrm{Cov}(X,Y) = E[(X-EX)(Y-EY)] = E[XY] - (EX)(EY)$$

- **Positive Covariance:** Values of X above its mean tend to be associated with values of Y above its mean.
- **Negative Covariance:** Values of X above its mean tend to be associated with values of Y below its mean.
- **Independence Property:** If X and Y are independent, then $E[XY] = (EX)(EY)$, so $\mathrm{Cov}(X,Y) = 0$. However, zero covariance does *not* necessarily imply independence.

We should note that covariance only captures **linear dependence** between X and Y.

2. Correlation Coefficient
Covariance depends on the scales and units of X and Y. To obtain a dimensionless measure bounded between -1 and 1, we normalize the covariance by the standard deviations:

$$\rho_{XY} = \rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$$

provided that $0 < \sigma_X, \sigma_Y < \infty$. The correlation coefficient satisfies

$$-1 \le \rho_{XY} \le 1.$$

Extension to Higher Dimensions: The Covariance Matrix
For a random vector $\mathbf{X} = [X_1, X_2, \ldots, X_d]^T \in \mathbb{R}^d$, the pairwise covariances are organized into a $d \times d$ matrix called the **covariance matrix** $\boldsymbol{\Sigma}$. Defining the mean vector as $\boldsymbol{\mu} = E[\mathbf{X}]$:

$$\boldsymbol{\Sigma}_{ij} = \mathrm{Cov}(X_i, X_j), \qquad \boldsymbol{\Sigma} = E\left[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T\right]$$

- The diagonal entries are the individual variances: $\boldsymbol{\Sigma}_{ii} = \mathrm{Var}(X_i)$.
- The off-diagonal entries are the pairwise covariances: $\boldsymbol{\Sigma}_{ij} = \mathrm{Cov}(X_i, X_j)$.
- **Key Property:** $\boldsymbol{\Sigma}$ is symmetric ($\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^T$) and positive semidefinite (PSD): $\mathbf{z}^T\boldsymbol{\Sigma}\mathbf{z} \ge 0$ for all $\mathbf{z} \in \mathbb{R}^d$. This is fundamental in machine learning, multivariate Gaussian distributions, and PCA.

Bounds & Probability Inequalities

1. Markov's Inequality
Statement: If X is a **non-negative** random variable ($X \geq 0$), then for any threshold $a > 0$:

$$\mathbf{P(X \geq a) \leq \frac{EX}{a}}$$

Proof (Continuous Case): For a non-negative continuous random variable X with PDF $f_X(x)$:

$$\begin{aligned} EX = \int_{-\infty}^{\infty} x f_X(x)\,dx &= \int_{0}^{\infty} x f_X(x)\,dx \qquad \text{(Since } X \geq 0\text{)} \\ &\geq \int_{a}^{\infty} x f_X(x)\,dx \qquad \text{(Restricting integration domain for } a > 0\text{)} \\ &\geq \int_{a}^{\infty} a f_X(x)\,dx \qquad \text{(Since } x \geq a \text{ inside the integral)} \\ &= a \int_{a}^{\infty} f_X(x)\,dx \\ &= aP(X \geq a) \end{aligned}$$

Dividing both sides by a yields $P(X \geq a) \leq \frac{EX}{a}$.

2. Union Bound
Statement: For any sequence of events $A_1, A_2, \ldots, A_n$:

$$\mathbf{P}\left(\bigcup_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{A_i}\right) \leq \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{P(A_i)}$$

Proof: Let X be the random variable counting how many events A_i occur:

$$X = \sum_{i=1}^{n} \mathbb{I}_{A_i}$$

By linearity of expectation, $EX = \sum_{i=1}^{n} E[\mathbb{I}_{A_i}] = \sum_{i=1}^{n} P(A_i)$. Since X counts occurrences, X is a non-negative integer-valued random variable ($X \geq 0$). Applying Markov's inequality with $a = 1$:

$$P(X \geq 1) \leq \frac{EX}{1} = \sum_{i=1}^{n} P(A_i)$$

Note that $P(X \geq 1)$ is the probability that at least one event A_i occurs, $P(X \geq 1) = P\left(\bigcup_{i=1}^{n} A_i\right)$.

3. Chebyshev's Inequality
Statement: Let X be a random variable with finite mean $\mu = EX$ and finite variance $\text{Var}(X)$. For any $b > 0$:

$$\mathbf{P(|X - EX| \geq b) \leq \frac{Var(X)}{b^2}}$$

Proof (via Markov's Inequality): Define the transformed random variable $Y = (X - EX)^2$. Since Y is a square, it is non-negative ($Y \geq 0$) with expectation $EY = E[(X - EX)^2] = \text{Var}(X)$. Applying Markov's inequality to Y with threshold $a = b^2$:

$$P(Y \geq b^2) \leq \frac{EY}{b^2} = \frac{\text{Var}(X)}{b^2}$$

Noting that $P(Y \geq b^2) = P\left((X - EX)^2 \geq b^2\right) = P(|X - EX| \geq b)$ gives us Chebyshev's inequality.

4. Chernoff Bounds
Statement: Let X be a real-valued random variable and let $a \in \mathbb{R}$.
For any $s > 0$ such that $E[e^{sX}] < \infty$,

$$P(X \geq a) \leq e^{-sa} E[e^{sX}].$$

Similarly, for any $s < 0$ such that $E[e^{sX}] < \infty$,

$$P(X \leq a) \leq e^{-sa} E[e^{sX}].$$

Proof (via Markov's Inequality):
For $s > 0$, the exponential function e^{sx} is increasing, so $X \geq a$ if and only if $e^{sX} \geq e^{sa}$. Applying Markov's inequality to the non-negative random variable e^{sX} gives

$$P(X \geq a) = P(e^{sX} \geq e^{sa}) \leq \frac{E[e^{sX}]}{e^{sa}} = e^{-sa} E[e^{sX}].$$

For $s < 0$, the exponential function is decreasing, so $X \leq a$ if and only if $e^{sX} \geq e^{sa}$. Applying Markov's inequality in the same way gives

$$P(X \leq a) \leq e^{-sa} E[e^{sX}].$$

Since these bounds hold for any admissible s of the appropriate sign, we can optimize over s to obtain the strongest bound.

Sample Mean, WLLN, and the Central Limit Theorem

1. The Sample Mean

For n i.i.d. random variables $X_1, X_2, \ldots, X_n$ with mean $EX_i = \mu$ and variance $\text{Var}(X_i) = \sigma^2$, the **sample mean** $\overline{X}$ is defined as:

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Its expectation and variance are given by:

$$E[\overline{X}] = \frac{EX_1 + EX_2 + \cdots + EX_n}{n} = \frac{n\mu}{n} = \mu$$

$$\text{Var}(\overline{X}) = \frac{\text{Var}(X_1 + \cdots + X_n)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Key Takeaway: As $n \to \infty$, the variance of the sample mean vanishes, so the sample mean becomes increasingly concentrated around μ.

2. The Weak Law of Large Numbers (WLLN)

For i.i.d. random variables $X_1, \ldots, X_n$ with $\mathbb{E}[X_i] = \mu$ and finite variance $\text{Var}(X_i) = \sigma^2$, the sample mean converges to μ in probability. That is, for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|\overline{X} - \mu| \geq \epsilon) = 0$$

Proof via Chebyshev:

$$P(|\overline{X} - \mu| \geq \epsilon) \leq \frac{\text{Var}(\overline{X})}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \xrightarrow{n\to\infty} 0.$$

Remark: A stronger result, known as the Strong Law of Large Numbers (SLLN), also exists, but we will not discuss its details here.

3. The Central Limit Theorem (CLT)

While the WLLN describes how the sample mean concentrates around μ as n grows, the CLT describes the shape of its distribution around μ for large n.

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $0 < \text{Var}(X_i) = \sigma^2 < \infty$. The standardized sample mean

$$Z_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma}$$

converges in distribution to a standard normal distribution $\mathcal{N}(0,1)$ as $n \to \infty$:

$$\lim_{n\to\infty} P(Z_n \leq x) = \Phi(x), \quad \text{for all } x \in \mathbb{R}.$$

Continuity Correction for Discrete Random Variables

When approximating an integer-valued discrete sum $Y = \sum_{i=1}^n X_i$ (such as a Binomial(n,p) variable) using a continuous Normal distribution via the CLT, we encounter a structural mismatch: discrete distributions place non-zero probability on individual points, whereas continuous distributions have zero probability at exact points.

The Key Idea: To bridge this gap, each integer value k is represented by the continuous interval $[k - 0.5, k + 0.5]$. For integer bounds l and u, the continuity correction approximates the probability of the discrete event by the corresponding interval under a continuous normal approximation:

$$P(l \leq Y \leq u) \approx P\left(l - \frac{1}{2} < W < u + \frac{1}{2}\right),$$

where W is a normal random variable approximating Y.

Applying the CLT with Continuity Correction: For $Y \sim \text{Binomial}(n, p)$ with mean $\mu = np$ and variance $\sigma^2 = np(1-p)$:

$$P(l \le Y \le u) \approx \Phi\left(\frac{u + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{l - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$$

■ **Example:**

Let $Y \sim \text{Binomial}\left(n = 20, p = \frac{1}{2}\right)$. We want to compute $P(8 \le Y \le 10)$.

1. Exact Probability Computation
Since $n = 20$ is small, we can calculate the exact binomial probability directly:

$$\begin{aligned} P(8 \le Y \le 10) &= \sum_{k=8}^{10} \binom{20}{k} \left(\frac{1}{2}\right)^k \left(1 - \frac{1}{2}\right)^{20-k} \\ &= \left[\binom{20}{8} + \binom{20}{9} + \binom{20}{10}\right] \left(\frac{1}{2}\right)^{20} = \mathbf{0.4565} \end{aligned}$$

2. Naive CLT Approximation
Represent $Y = \sum_{i=1}^{20} X_i$ where $X_i \sim \text{Bernoulli}\left(\frac{1}{2}\right)$.

- Mean: $n\mu = 20 \cdot \frac{1}{2} = 10$
- Variance: $n\sigma^2 = 20 \cdot \frac{1}{4} = 5 \implies \sqrt{n}\sigma = \sqrt{5} \approx 2.236$

Standardizing Y directly:

$$\begin{aligned} P(8 \le Y \le 10) &= P\left(\frac{8-10}{\sqrt{5}} \le \frac{Y-10}{\sqrt{5}} \le \frac{10-10}{\sqrt{5}}\right) \\ &\approx \Phi(0) - \Phi\left(\frac{-2}{\sqrt{5}}\right) \\ &\approx 0.5000 - 0.1855 = \mathbf{0.3145} \end{aligned}$$

Error: $0.4565 - 0.3145 = 0.1420$ (A severe 31% relative error!)

3. Improved Approximation (With Continuity Correction)
Expanding the integer bounds to $[8 - 0.5, 10 + 0.5] = [7.5, 10.5]$:

$$P(8 \le Y \le 10) \approx P(7.5 < W < 10.5),$$

where $W \sim \mathcal{N}(10, 5)$ is the normal approximation to Y.

Standardizing the adjusted bounds:

$$\begin{aligned} P(8 \le Y \le 10) &\approx P\left(\frac{7.5-10}{\sqrt{5}} < \frac{W-10}{\sqrt{5}} < \frac{10.5-10}{\sqrt{5}}\right) \\ &= \Phi\left(\frac{0.5}{\sqrt{5}}\right) - \Phi\left(\frac{-2.5}{\sqrt{5}}\right) \\ &\approx 0.5885 - 0.1318 = \mathbf{0.4567}. \end{aligned}$$

Takeaway: Continuity correction brought the approximation from 0.3145 to 0.4567, almost perfectly matching the true value of 0.4565!