

Lecture 4

*Instructor: Sruthi Sekar**Scribe: Priyanshu Singh*

Topics for Today

- Sum of Random Variables (RVs)
- Central Limit Theorem

References

- Introduction to Probability Theory by Feller
- Ryan O'Donnell's Lecture - 4

Chalk and Talk Topics

1. **Topic 1:** An elementary proof of Stirling's formula
 - Group Size: 2
 - Presentation Date: February 18th (during class)

The Erdős-Kac Theorem

The Erdős-Kac theorem is a remarkable result in probabilistic number theory that demonstrates the normal distribution of the number of prime factors of integers.

Theorem 1 (Erdős-Kac, 1940). *Let $\omega(n)$ denote the number of distinct prime factors of n . Then for any real number x ,*

$$\lim_{N \rightarrow \infty} \frac{1}{N} \# \left\{ n \leq N : \frac{\omega(n) - \log \log n}{\sqrt{\log \log n}} \leq x \right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$$

Remark This theorem essentially states that the distribution of $\omega(n)$ follows a normal distribution with:

- Mean $\approx \log \log n$
- Variance $\approx \log \log n$

Example 2. *For a large integer n , the probability that $\omega(n)$ deviates from $\log \log n$ by more than $\lambda \sqrt{\log \log n}$ is approximately the same as the probability that a standard normal random variable exceeds λ in absolute value.*

The Galton Board Experiment

Rules and Description

The Galton board consists of a vertical board with interleaved rows of pegs. A ball is dropped from the top and bounces off the pegs as it falls, eventually landing in one of the collection bins at the bottom.

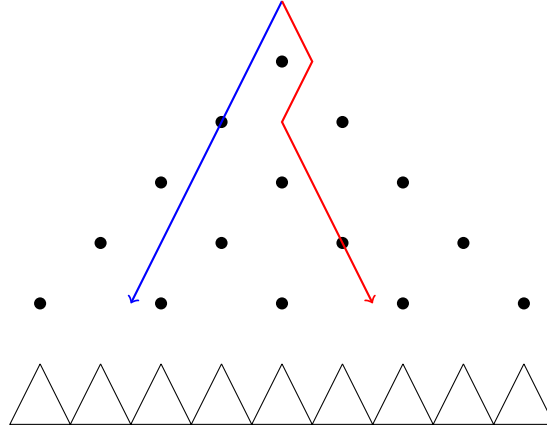


Figure 1: Galton Board (Quincunx)

Mathematical Model

Let X_i be the random variable representing the direction of the i th bounce:

$$X_i = \begin{cases} +1 & \text{if the ball goes right} \\ -1 & \text{if the ball goes left} \end{cases}$$

Each X_i is a Bernoulli random variable with $P(X_i = +1) = P(X_i = -1) = \frac{1}{2}$ (assuming the board is perfectly balanced).

Key Assumptions

1. **Independence:** Each bounce is independent of all previous bounces
2. **Identical Distribution:** Each bounce has the same probability distribution ($p = 1/2$)

Final Position

The final position of the ball after n bounces is given by:

$$S_n = \sum_{i=1}^n X_i$$

Note that S_n now represents the net displacement from the center, where:

- $E[X_i] = 0$
- $Var(X_i) = 1$
- $S_n \sim \text{Binomial}(n, -1, +1)$

By the Central Limit Theorem, as n becomes large:

$$\frac{S_n}{\sqrt{n}} \xrightarrow{d} N(0, 1)$$

Properties of Variance

Let X and Y be random variables, c be a constant, and let σ denote the standard deviation.

1. **Variance of Sum of Independent RVs:**

For independent random variables X and Y :

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$$

2. **Variance with Constant Addition:**

For any constant k :

$$\text{Var}[X + k] = \text{Var}[X]$$

3. **Variance with Scalar Multiplication:**

For any constant c :

$$\text{Var}[cX] = c^2 \text{Var}[X]$$

4. **Definition of Variance:**

Let $\mu = E[X]$, then:

$$\text{Var}[X] = E[(X - \mu)^2]$$

5. **Standard Deviation:**

The standard deviation is the square root of variance:

$$\sigma = \sqrt{\text{Var}(X)}$$

Remark These properties are fundamental in probability theory and are particularly useful when working with sums of random variables, as in the Central Limit Theorem.

Standardization of Random Variables

General Process of Standardization

1. **Centering:** Subtract expected value to center around 0

$$S_n - E[S_n]$$

2. **Normalizing Variance:** Divide by standard deviation to get variance 1

$$Z_n = \frac{S_n - E[S_n]}{\sigma}$$

3. **Recovery Formula:** Original variable can be recovered using

$$S_n = Z_n \sigma + E[S_n]$$

Application to Galton Board

Consider the Galton board where $p = q = \frac{1}{2}$ and $S_n = X_1 + X_2 + \dots + X_n$ represents the final position.

- Expected Value: $E[S_n] = \frac{n}{2}$
- Variance: $\text{Var}[S_n] = \frac{n}{4}$
- Standard Deviation: $\sigma = \frac{\sqrt{n}}{2}$

The standardized variable is:

$$Z_n = \frac{S_n - \frac{n}{2}}{\frac{\sqrt{n}}{2}} = \frac{1}{\sqrt{n}}[(2X_1 - 1) + (2X_2 - 1) + \dots + (2X_n - 1)]$$

Probability Distribution Analysis

For even n , using Stirling's approximation:

$$\Pr[S_n = \frac{n}{2}] = \Pr[B_n = 0] \approx \sqrt{\frac{2}{\pi}} \cdot \frac{1}{\sqrt{n}} = \Theta(\frac{1}{\sqrt{n}})$$

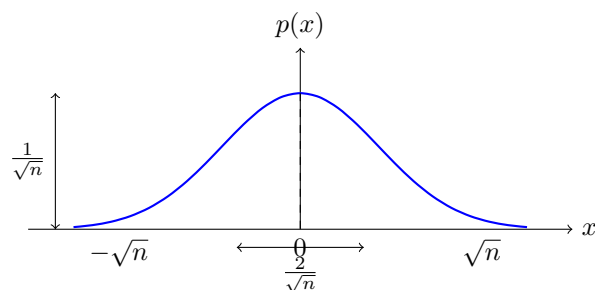


Figure 2: Probability Distribution of Standardized Galton Board Outcomes

Analysis of Key Points

- **Peak Height** ($\frac{1}{\sqrt{n}}$): Follows from Stirling's approximation
- **Bin Width** ($\frac{2}{\sqrt{n}}$): Represents standardized step size
- **Extremes** ($\pm\sqrt{n}$): Maximum deviation possible when all steps go in same direction:

$$\pm\sqrt{n} = \frac{n - \frac{n}{2}}{\frac{\sqrt{n}}{2}} = \frac{\pm\frac{n}{2}}{\frac{\sqrt{n}}{2}}$$

Step Size Analysis

In the standardized form Z_n , let's calculate the difference between consecutive positions:

Original step size = 1 (in S_n)

$$\text{In standardized form } Z_n = \frac{S_n - \frac{n}{2}}{\frac{\sqrt{n}}{2}}$$

$$\begin{aligned} \text{Therefore, one step in } Z_n &= \frac{1}{\frac{\sqrt{n}}{2}} \\ &= \frac{2}{\sqrt{n}} \end{aligned}$$

This means when we move one position in the original Galton board (S_n changes by 1), the standardized variable Z_n changes by $\frac{2}{\sqrt{n}}$.

Remark This standardized step size $\frac{2}{\sqrt{n}}$ becomes smaller as n increases, which is one reason why the distribution looks increasingly smooth for large n .

Gaussian Distribution

Basic Definition

If $X \sim N(\mu, \sigma^2)$, then X follows a Gaussian distribution with probability density function:

$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Building Intuition

Let's understand how the Gaussian shape emerges:

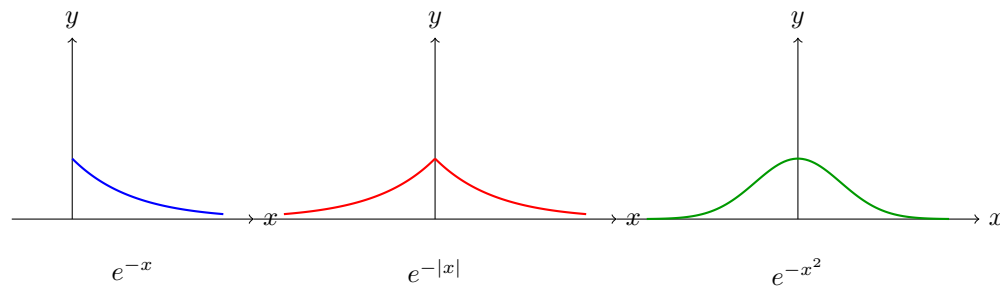


Figure 3: Evolution of the Gaussian shape

Proving $\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$

Consider the double integral:

$$I = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)} dx dy$$

Using polar coordinates:

$$I = \int_0^{\infty} 2\pi r e^{-r^2} dr = \pi$$

Due to independence:

$$I = \left(\int_{-\infty}^{\infty} e^{-x^2} dx \right)^2$$

Therefore:

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

Radial Symmetry

The Gaussian distribution exhibits radial symmetry, meaning:

$$\phi(x - \mu) = \phi(\mu - x)$$

This property was crucial in our proof using polar coordinates.

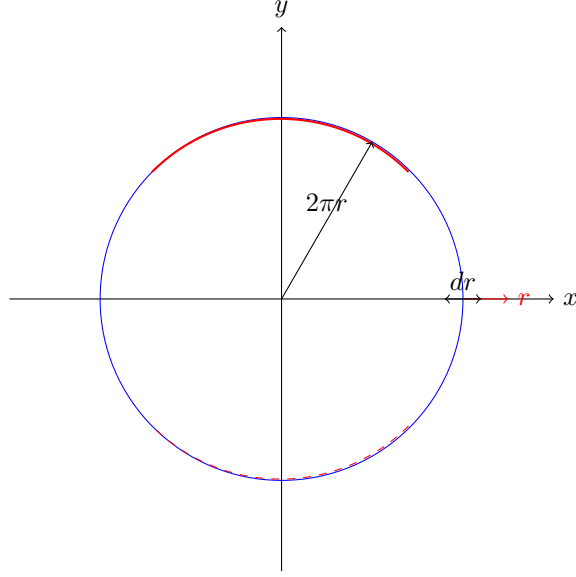


Figure 4: Cylindrical shell for integration

Sum of Independent Gaussians

If $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ are independent, then:

$$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Proof. Let $Z = X_1 + X_2$. By independence, the joint density is the product of individual densities:

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_{X_1}(x) f_{X_2}(z-x) dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \cdot \frac{1}{\sigma_2 \sqrt{2\pi}} e^{-\frac{(z-x-\mu_2)^2}{2\sigma_2^2}} dx \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2} - \frac{(z-x-\mu_2)^2}{2\sigma_2^2}\right) dx \end{aligned}$$

Complete the square in the exponent:

$$\begin{aligned} &\frac{(x-\mu_1)^2}{2\sigma_1^2} + \frac{(z-x-\mu_2)^2}{2\sigma_2^2} \\ &= \frac{\sigma_2^2(x-\mu_1)^2 + \sigma_1^2(z-x-\mu_2)^2}{2\sigma_1^2\sigma_2^2} \\ &= \frac{(\sigma_1^2 + \sigma_2^2)}{2\sigma_1^2\sigma_2^2} \left(x - \frac{\sigma_2^2\mu_1 + \sigma_1^2(z-\mu_2)}{\sigma_1^2 + \sigma_2^2}\right)^2 + \frac{(z-(\mu_1+\mu_2))^2}{2(\sigma_1^2 + \sigma_2^2)} \end{aligned}$$

After integration:

$$f_Z(z) = \frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}} e^{-\frac{(z-(\mu_1+\mu_2))^2}{2(\sigma_1^2 + \sigma_2^2)}}$$

This is the density function of $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. □

Central Limit Theorem

Let $X_1, X_2, \dots, X_n$ be independent, identically distributed random variables with mean μ and variance σ^2 . Define:

$$S_n = X_1 + X_2 + \dots + X_n$$
$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

Then, for any real number u :

$$\lim_{n \rightarrow \infty} \Pr[Z_n \leq u] = \Pr[X \leq u]$$

where $X \sim N(0, 1)$.

Berry-Esseen Theorem

The Berry-Esseen theorem provides a quantitative version of the Central Limit Theorem, giving bounds on the rate of convergence.

Theorem 3 (Berry-Esseen). *Let $X_1, \dots, X_n$ be independent random variables with:*

- $E[X_i] = 0$ for all i
 - $\sum_{i=1}^n \sigma_i^2 = 1$, where $\sigma_i^2 = \text{Var}[X_i]$
- Let $S_n = X_1 + \dots + X_n$, then $E[S_n] = 0$ and $\text{Var}[S_n] = 1$.
For $Z \sim N(0, 1)$ and any $u \in \mathbb{R}$:*

$$|\Pr[S_n \leq u] - \Pr[Z \leq u]| \leq c \cdot Y$$

where:

- $Y = \sum_{i=1}^n E[|X_i|^3]$
- $c \leq 0.56$ is a universal constant

Remark If $|X_i| \leq \varepsilon$ with probability 1, then:

$$Y = \sum_{i=1}^n E[|X_i|^3] \leq \varepsilon \sum_{i=1}^n E[|X_i|^2] \leq \varepsilon$$

Example: Fair Coin Tosses

Consider fair coin tosses where:

$$X_i = \begin{cases} +\frac{1}{\sqrt{n}} & \text{with probability } \frac{1}{2} \\ -\frac{1}{\sqrt{n}} & \text{with probability } \frac{1}{2} \end{cases}$$

Let's calculate:

$$\sigma_i^2 = \text{Var}[X_i] = \frac{1}{n}$$
$$\sum_{i=1}^n \sigma_i^2 = 1$$
$$E[|X_i|^3] = \frac{1}{n^{3/2}}$$
$$Y = \sum_{i=1}^n E[|X_i|^3] = \frac{1}{\sqrt{n}}$$

For counting heads in n fair coin tosses:

- Original count: $H_n = \sum_{i=1}^n \mathbf{1}_{X_i = +\frac{1}{\sqrt{n}}}$
- Standardized version: $\frac{H_n - \frac{n}{2}}{\sqrt{n}/2}$
- Berry-Esseen bound: $|F_n(x) - \Phi(x)| \leq \frac{0.56}{\sqrt{n}}$

where F_n is the CDF of the standardized count and Φ is the standard normal CDF.