Galton Board Simulation and Probability Analysis

Miquel Sala Franci
Aftab Ahmed Choudhry

Randomized Algorithms

October 2025

Abstract

This report presents a numerical simulation of a Galton board experiment using Python in Google Colab. The goal is to study the relationship between empirical results obtained from random experiments and theoretical predictions based on the Binomial and Normal distributions. We examine how the agreement improves as the number of board levels n and the number of balls N increase. The source code and notebooks used for these experiments are available on GitHub: https://github.com/Aftab5550/galton-board-simulation.

Contents

Chapter 1

Galton Board Simulation and Probability Analysis

1.1 Purpose and Initial Function Definitions

The purpose of this project is to simulate the behavior of a **Galton board** (also known as a quincunx) and to compare the *experimental results* obtained from simulation with the *theoretical predictions* derived from probability theory. In particular, the study aims to examine how the outcomes of many independent trials of falling balls correspond to the Binomial distribution, and how the Binomial distribution approaches the Normal distribution as the number of levels increases, while the accuracy of the experimental results improves as the number of simulated balls increases.

Before performing the experiments, two fundamental functions were defined: `single_ball_path` and `experiment`.

Function `single_ball_path(levels)`

This function simulates the trajectory of a *single ball* as it travels through a Galton board with a given number of levels. Each level represents a point where the ball can move either to the **left** or to the **right** with equal probability (50%).

The function starts at the initial position $(0, 0)$ and iteratively decides the next step based on a random number generated at each level:

- If the random probability is less than 0.5, the ball moves to the left.
- Otherwise, it moves to the right.

At the end of the process, the function returns a list of coordinate pairs that represent the complete path of the ball as it moves down the board. Each coordinate (row, col) indicates the vertical and horizontal position of the ball at a given level.

Function `experiment(n, N)`

The second function is responsible for simulating the behavior of *many balls*. It performs N independent trials, each representing a different ball falling through the same Galton board with n levels. For each ball, it calls the `single_ball_path` function to determine its final position and records the outcomes in order to compute experimental frequencies and compare them with theoretical probabilities.

1.2 Verification of the `single_ball_path` Function

Before proceeding to large-scale experiments, a preliminary test was carried out to verify that the `single_ball_path` function behaves correctly and produces random trajectories.

For this test, the number of levels was set to $n = 10$, and the simulation was repeated for 20 different balls. Each ball's path was printed to observe the sequence of moves it followed through the board. This allowed us to confirm that:

- Each ball begins at the same starting point $(0, 0)$.
- The total number of movements corresponds to the number of levels ($n = 10$).
- At the final step, the sum of the two coordinates equals the number of levels ($row + col = n$).
- Each successive coordinate differs from the previous one by exactly 1 unit in either the horizontal or vertical direction.
- The sequence of left and right movements varies randomly between balls.

An excerpt of the output obtained from this verification is shown below:

```
The path for the 0 ball is:  [(0, 0), (0, 1), (0, 2), (1, 2), (1, 3), (1, 4), (1,
5), (2, 5), (2, 6), (3, 6), (3, 7)]
The path for the 1 ball is:  [(0, 0), (1, 0), (1, 1), (2, 1), (3, 1), (3, 2), (3,
3), (3, 4), (4, 4), (5, 4), (6, 4)]
...  (remaining 18 balls omitted for brevity)
```

The results clearly demonstrate that the random component of the algorithm works as intended: Each ball follows a unique path determined by independent random decisions. This step served as an essential verification of the correctness and randomness of the simulation before moving on to the statistical analysis of results.

1.3 Comparison Between Empirical and Theoretical Distributions

After verifying that the simulation behaves as expected for individual balls, we extended the analysis to a large number of trials in order to compare the empirical results with the corresponding theoretical models. The main objective of this section is to quantify the difference between the experimental probabilities obtained from simulation and the theoretical probabilities predicted by the Binomial and Normal distributions.

Experimental Setup

Two parameters were varied systematically:

- The number of levels in the Galton board, denoted by n, taking values in {`5, 10, 30, 50`}.
- The number of balls (or experiments), denoted by N, taking values in {`50, 100, 500, 1000, 5000, 10000, 20000, 50000`}.

For each combination of (n, N), the experiment was repeated independently. The function `experiment(n, N)` was used to simulate the fall of N balls through a Galton board with n levels. Each ball follows a random sequence of left or right moves, and its final position determines the bin where it lands at the bottom.

Empirical Probabilities

Once all N balls have been dropped, we count how many balls landed in each possible final position. The empirical probability of a ball landing in position k is therefore given by:

$$P_{\text{emp}}(k) = \frac{\text{Number of balls in position } k}{N}, \quad k = 0, 1, \ldots, n.$$

This forms the **experimental probability distribution**, which can be directly compared to theoretical models.

Theoretical Models

Two theoretical distributions were considered for comparison:

- **Binomial Distribution:** Each ball has an equal chance of moving right (success) or left (failure) at each level, making the number of right moves follow a Binomial distribution with parameters $(n, p = 0.5)$. The corresponding probability mass function (PMF) is:

$$P_{\text{binom}}(k) = \binom{n}{k}(0.5)^k(0.5)^{n-k}.$$

- **Normal Approximation:** For large n, as we will see in Section 1.6, the Binomial distribution can be approximated by a Normal distribution with mean $\mu = \frac{n}{2}$ and standard deviation $\sigma = \sqrt{\frac{n}{4}}$. To make the comparison meaningful at discrete points, the probabilities were computed using the *continuity correction*:

$$P_{\text{norm}}(k) = \Phi(k + 0.5;\, \mu, \sigma) - \Phi(k - 0.5;\, \mu, \sigma),$$

where $\Phi(x;\, \mu, \sigma)$ denotes the cumulative distribution function of a Normal distribution with mean μ and standard deviation σ.

Error Quantification

To evaluate how well the theoretical models fit the empirical results, we computed the **Mean Squared Error (MSE)** between the empirical probabilities and the theoretical predictions:

$$\text{MSE}_{\text{binom}} = \frac{1}{n+1}\sum_{k=0}^{n}(P_{\text{emp}}(k) - P_{\text{binom}}(k))^2,$$

$$\text{MSE}_{\text{norm}} = \frac{1}{n+1}\sum_{k=0}^{n}(P_{\text{emp}}(k) - P_{\text{norm}}(k))^2.$$

These errors quantify how closely the simulated data follow the expected theoretical patterns.

Visualization of Results

For each (n, N) pair, the results were visualized using bar plots and line graphs:

- Blue bars represent the **empirical probabilities** obtained from simulation.
- Solid dots connected by lines represent the **Binomial PMF**.
- Red dashed lines represent the **Normal approximation**.

1.4 Convergence Analysis and Discussion of Results

Hypothesis

According to the **Law of Large Numbers**, as the number of balls N increases, the empirical distribution of outcomes should converge to the true theoretical distribution. Moreover, as the number of levels n increases, the **Binomial distribution** that describes the Galton board outcomes approaches a **Normal distribution** as we will discuss in the Section 1.6. Therefore, our initial hypotheses are:

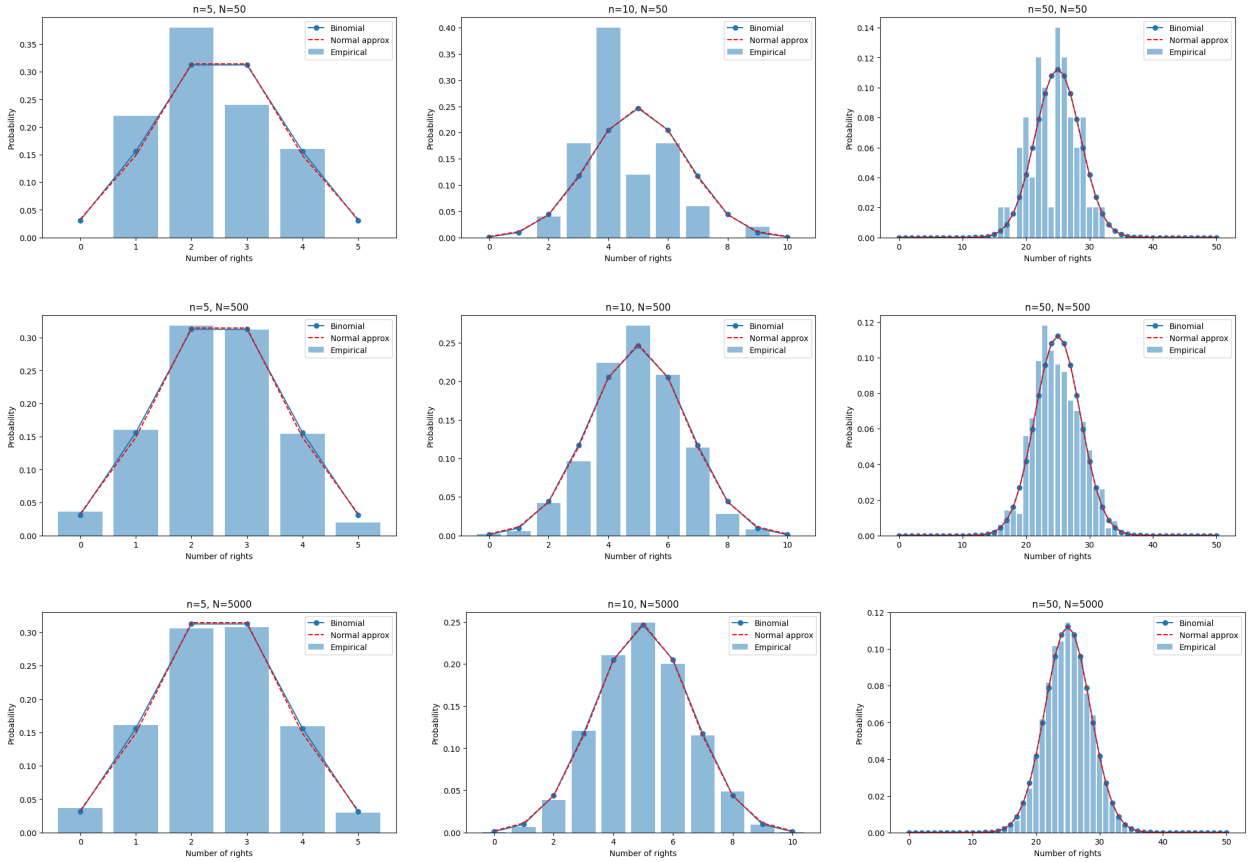
1. For fixed n, increasing N should make the empirical distribution more closely match the Binomial distribution.

2. For large n, the Binomial distribution itself should closely resemble the Normal distribution.

These hypotheses will be tested by comparing empirical, Binomial, and Normal probabilities across different values of n and N.

Visual Comparison of Distributions

The following figures present the empirical results (blue bars), the theoretical Binomial probabilities (solid dots), and the Normal approximations (red dashed lines) for each pair of parameters (n, N):



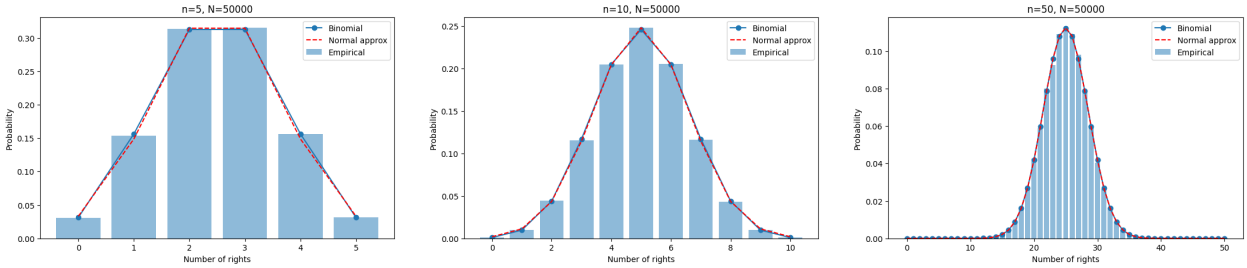


Figure 1.2: Comparison between empirical, Binomial, and Normal distributions for different numbers of levels (n) and balls (N).

For small values of N, the empirical histograms show noticeable random fluctuations due to limited sampling. As N increases, these irregularities smooth out, and the empirical distribution aligns almost perfectly with the Binomial prediction. When n is large (e.g., $n = 30$ or $n = 50$), the Normal curve becomes visually indistinguishable from the Binomial distribution, confirming the expected asymptotic behavior.

1.5 Error Trends and Quantitative Analysis

After verifying the correctness of the simulation, the next step was to study how the empirical distributions converge to their theoretical counterparts as both the number of levels n and the number of balls N increase. For this purpose, the Mean Squared Error (MSE) between the experimental data and the theoretical Binomial and Normal distributions was computed.

Experimental Configuration

The experiments were performed for four different board sizes, $n = [5, 10, 30, 50]$, and eight different sample sizes, $N = [50, 100, 500, 1000, 5000, 10000, 20000, 50000]$. For each (n, N) pair:

- N balls were simulated using the `experiment(n, N)` function.
- The empirical probability distribution was computed using normalized frequency counts.
- Theoretical distributions were obtained from the Binomial model $\text{Bin}(n, 0.5)$ and the Normal approximation $\mathcal{N}(\mu = n/2, \sigma^2 = n/4)$.
- MSE values were calculated as the mean squared deviation between experimental and theoretical probabilities.

Each simulation also generated plots showing the agreement between the empirical distribution (bars), the theoretical Binomial distribution (solid line), and the Normal approximation (dashed line). These plots confirmed that the fit improves as either n or N increases.

Error Evolution with Sample Size

To analyze convergence quantitatively, the MSE values were plotted as a function of the number of balls N, for each fixed n. The results are displayed on a logarithmic scale to better capture the order-of-magnitude variation in the errors.

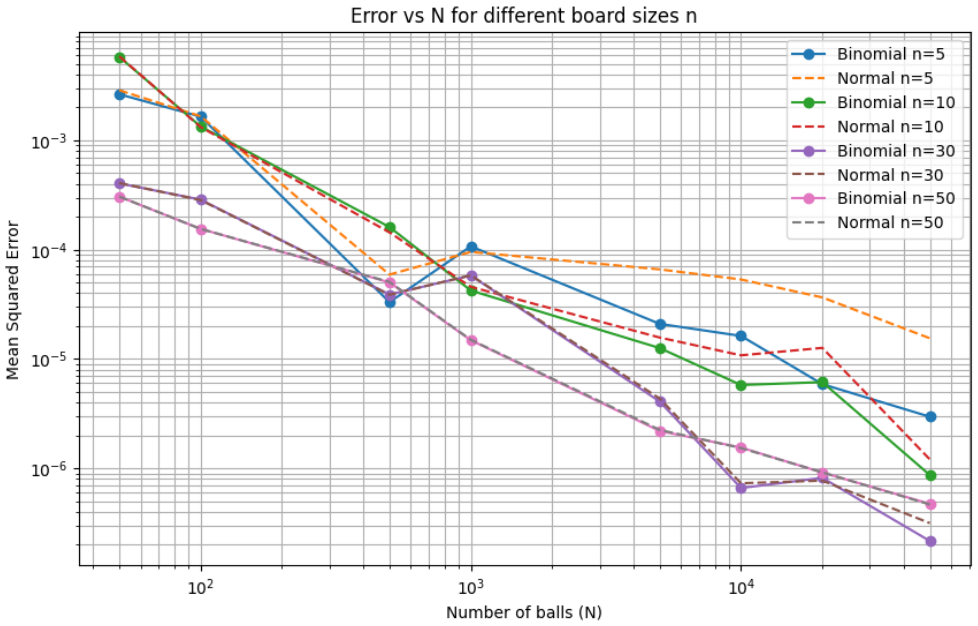


Figure 1.3: Mean Squared Error (MSE) as a function of the number of balls N for different board sizes n. Solid lines correspond to the Binomial model and dashed lines to the Normal approximation. Both axes are logarithmic.

Quantitative Results

Table 1.1 summarizes the MSE values obtained from all combinations of n and N. The numerical results confirm that both theoretical models converge to the empirical data as N increases, with the Normal approximation improving significantly for higher n.

n	N	$\text{MSE}_{\text{Binomial}}$	$\text{MSE}_{\text{Normal}}$
5	50	2.64×10^{-3}	2.87×10^{-3}
10	1000	4.23×10^{-5}	4.58×10^{-5}
30	5000	4.09×10^{-6}	4.35×10^{-6}
50	50000	4.73×10^{-7}	4.67×10^{-7}

Table 1.1: Representative MSE values showing convergence behavior for increasing N and n.

Discussion of Results

Several consistent patterns emerge from the data:

- **Error decreases with sample size:** For all n, both $\text{MSE}_{\text{Binomial}}$ and $\text{MSE}_{\text{Normal}}$ decrease steadily as N increases. This reflects the Law of Large Numbers, as the empirical frequencies approach their true probabilities with more samples.
- **Normal approximation improves with n:** For small boards (e.g., $n = 5$), the Normal approximation is less accurate than the Binomial model. However, for larger n (such as $n = 50$), the MSE values of both models become nearly identical, confirming the validity of the Normal approximation under the Central Limit Theorem.

- **Log–log linearity:** The MSE curves in Figure 1.3 show an approximately linear decay on a log–log scale, suggesting a power-law relationship MSE $\propto N^{-\alpha}$ with α close to 1. This indicates that the estimation error decreases inversely with sample size.

Conclusion

Overall, these results demonstrate strong agreement between simulation and theory:

1. The simulation accurately reproduces the probabilistic behavior of the Galton board.
2. Increasing N leads to consistent convergence toward the theoretical Binomial and Normal distributions.
3. The Normal approximation becomes nearly exact for sufficiently large n.

This quantitative validation completes the study, confirming both the correctness of the implementation and the theoretical predictions of probability theory regarding the relationship between Binomial and Normal distributions.

1.6 Formal Justification of the Normal Approximation of the Binomial Distribution

We now formally justify why the binomial distribution converges to the normal distribution as the number of trials increases.

Lemma 1.6.1 (Linear Transformation of a Normal Random Variable)**.** *Let $Z \sim \mathcal{N}(\mu, \sigma^2)$ be a normal random variable, and let $a, b \in \mathbb{R}$ be constants. Then the linear transformation*

$$Y = aZ + b$$

is also normally distributed:

$$Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2).$$

Proof. Let $f_Z(z)$ denote the probability density function (PDF) of Z. Then $Y = aZ + b$ implies $Z = (Y - b)/a$. Using the change of variables formula for densities:

$$\begin{aligned} f_Y(y) &= \frac{1}{|a|} f_Z\Big(\frac{y-b}{a}\Big) \\ &= \frac{1}{|a|\sqrt{2\pi\sigma^2}} \exp\left(- \frac{\left(\frac{y-b}{a} - \mu\right)^2}{2\sigma^2} \right) \\ &= \frac{1}{\sqrt{2\pi(a^2\sigma^2)}} \exp\left(- \frac{(y - (a\mu + b))^2}{2a^2\sigma^2} \right) \end{aligned}$$

which is exactly the PDF of $\mathcal{N}(a\mu + b, a^2\sigma^2)$. □

Theorem 1.6.2 (Central Limit Theorem)**.** *Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with mean μ and variance $\sigma^2 < \infty$. Define the standardized sum*

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

Then, as $n \to \infty$, the distribution of Z_n converges in distribution to the standard normal:

$$Z_n \xrightarrow{d} \mathcal{N}(0,1).$$

Proposition 1.6.3 (De Moivre–Laplace Theorem)**.** *Let $S_n \sim \text{Binomial}(n,p)$ be a binomial random variable. Then, for large n, S_n can be approximated by a normal distribution:*

$$S_n \approx \mathcal{N}\big(np,\, np(1-p)\big).$$

Proof. Let $S_n = X_1 + \cdots + X_n$, where $X_i \sim \text{Bernoulli}(p)$ are independent. Then

$$E[S_n] = np, \qquad \text{Var}[S_n] = np(1-p).$$

Applying the Central Limit Theorem 1.6.2 to the standardized sum

$$Z_n = \frac{S_n - np}{\sqrt{np(1-p)}},$$

we have

$$Z_n \xrightarrow{d} \mathcal{N}(0,1) \quad \text{as } n \to \infty.$$

Solving for S_n, we obtain

$$S_n = np + \sqrt{np(1-p)}\, Z_n.$$

Since $Z_n \sim \mathcal{N}(0,1)$ in the limit, by Lemma 1.6.1 we have that

$$S_n \sim \mathcal{N}\big(np,\, np(1-p)\big),$$

□