RA - Assignment 1: Galton board

Jairo El Yazidi Ríos
https://github.com/JairoRY/Galton-board

October 2025

1 Introduction

The Galton board, also known as the *quincunx*, is a classical statistical experiment that illustrates the emergence of binomial and normal distributions through a stochastic process.

In this assignment, a C++ simulation was implemented to emulate the Galton board. The program allows the user to specify both the number of levels n, the number of balls N, and the number of repetitions x (for experimental purposes) and computes the empirical distribution of the final positions. The goal is to compare the observed frequencies with the theoretical predictions of probability theory and to measure their agreement quantitatively.

2 Methods

2.1 Simulation Model

The Galton board is represented in the program as a one-dimensional array of counters, each corresponding to one of the possible final positions of a ball after passing through the n levels. Specifically, the array has size $n + 1$, where the k-th element counts how many balls end up with exactly k rightward moves. Repeating this process for N balls produces a complete empirical histogram of the final positions.

2.2 Theoretical Distributions

Two theoretical distributions are used for comparison:

- **Binomial Distribution:** Each path through the Galton board corresponds to a sequence of n Bernoulli trials, where a "success" is a move to the right. The probability mass function (PMF) is the following:

$$P_B(k) = \frac{\binom{n}{k}}{2^n} \tag{1}$$

  The expected number of balls at position k is then $E_B(k) = N \cdot P_B(k)$.

- **Normal Distribution:** For large n, the binomial distribution approaches the normal distribution $N(\mu, \sigma^2)$ with

$$\mu = \frac{n}{2}, \quad \sigma^2 = \frac{n}{4}.$$

The probability density function (PDF) is as follows:

$$f_N(k) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(k-\mu)^2}{2\sigma^2}} \tag{2}$$

As in the binomial case, the expected counts are scaled by N: $E_N(k) = N \cdot f_N(k)$.

2.3 Mean Squared Error (MSE)

The degree of similarity between distributions is quantified using the mean squared error:

$$\mathrm{MSE}(A, B) = \frac{1}{n+1} \sum_{k=0}^{n} (A_k - B_k)^2. \tag{3}$$

Here, A and B are vectors of expected counts at each position k, namely the observed, binomial and normal ones.

2.4 Data Structure and Repetition Scheme

Each simulation of the Galton board returns an `ExperimentResult` structure containing:

- The observed frequencies (as probabilities).
- The expected binomial probabilities.
- The expected normal probabilities.
- The three MSE values comparing them.

To ensure statistical robustness, the main program repeats the simulation x times. The observed frequencies and MSE values are averaged across repetitions, providing mean estimates of empirical behavior. The results are then printed in tabular format, showing the comparison between the observed and theoretical distributions, as well as a summary table of MSE values.

3 Results

The results obtained during the execution of the experiments with the implementation of the Galton board are shown below with a series of graphs (always with $x = 20$). Figures 1 to 4 aim to show the similarity between the experimental data and the predictions of probability theory, while also verifying the agreement between the binomial and normal distributions.

Results for $N = 50$ balls and $n = 6$ levels

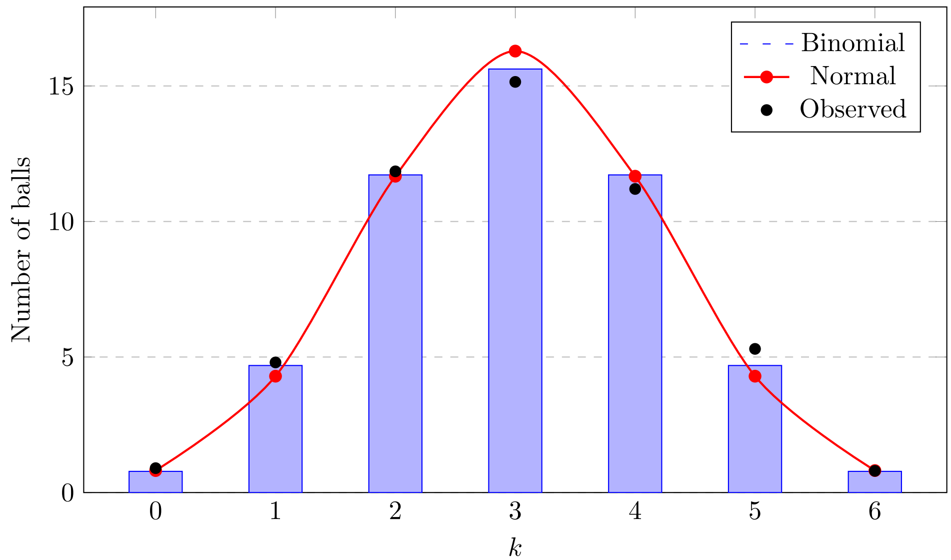

Figure 1: Comparison between distributions and observed data.

Results for $N = 1000$ balls and $n = 6$ levels

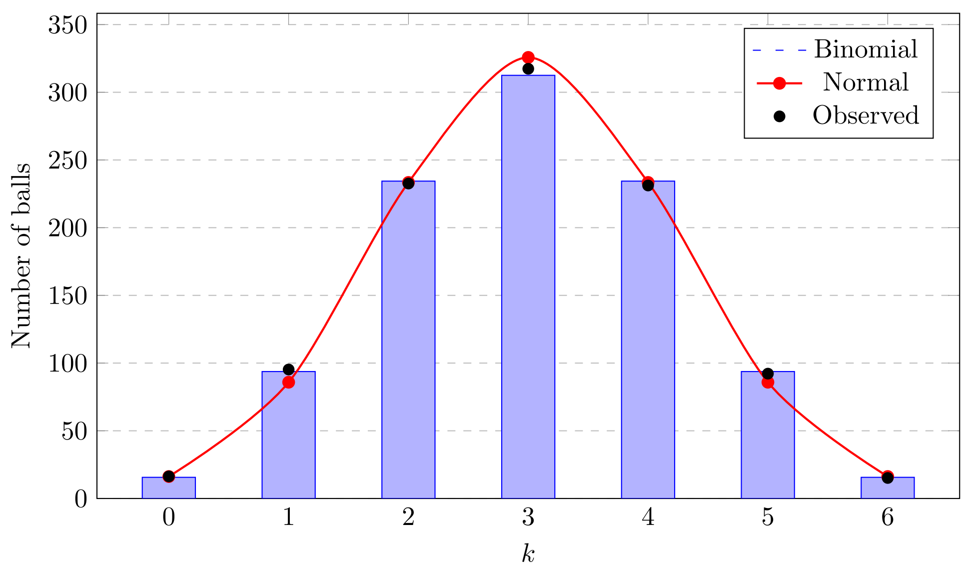

Figure 2: Comparison between distributions and observed data.

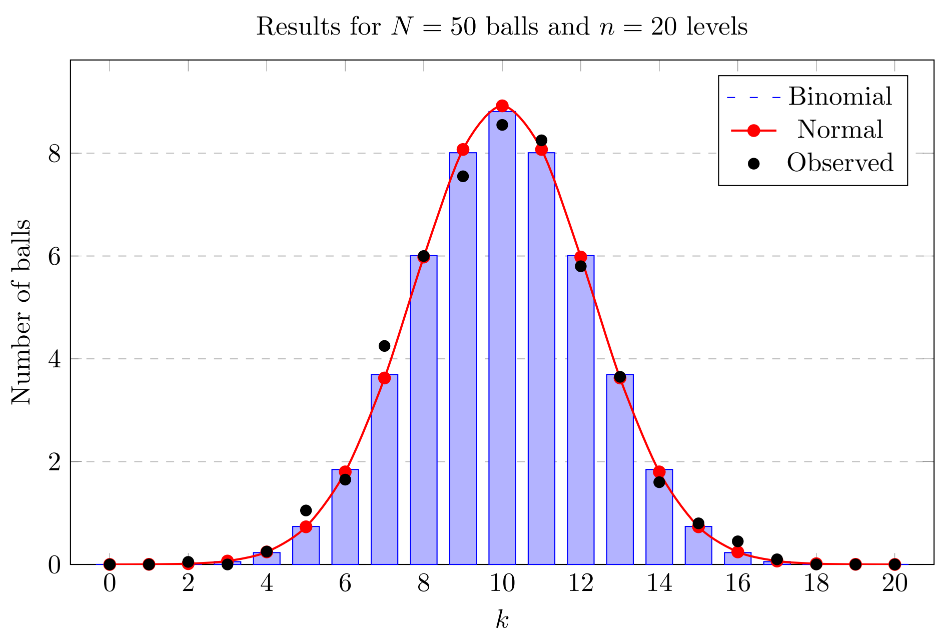

Figure 3: Comparison between distributions and observed data.

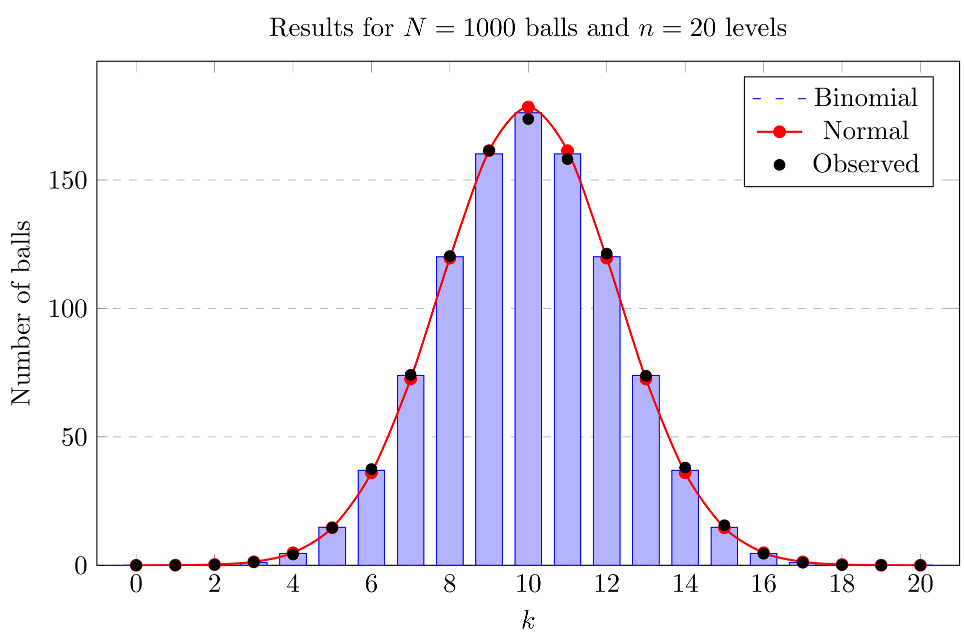

Figure 4: Comparison between distributions and observed data.

Finally, the last graphs below (Figures 5 to 7) visually display the Mean Squared Error between the empirical and theoretical results, as well as between the two reference distributions.

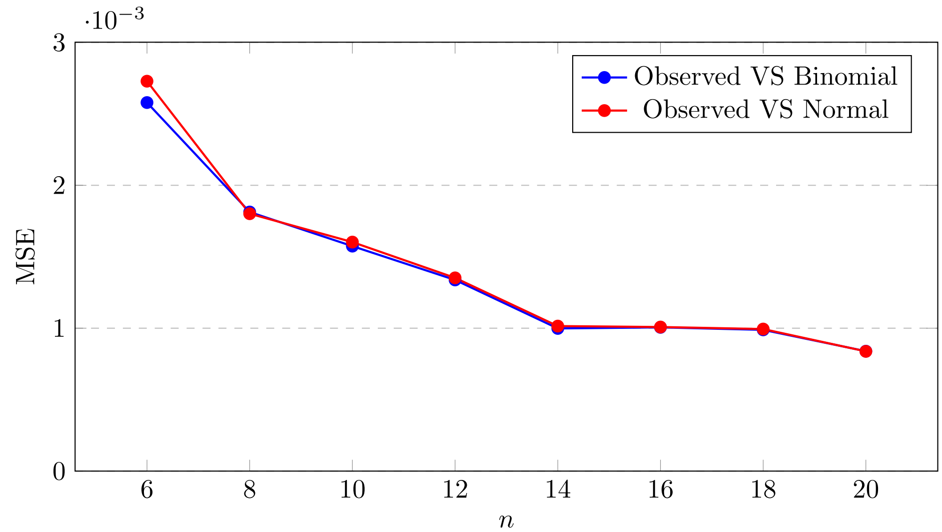

Figure 5: Evolution of MSE between empirical and theoretical distributions with respect to n.

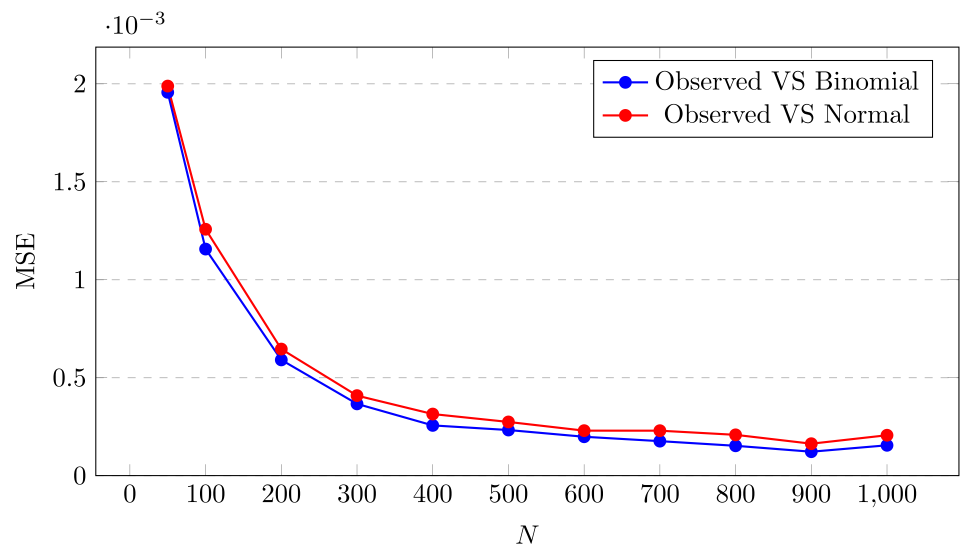

Figure 6: Evolution of MSE between empirical and theoretical distributions with respect to N.

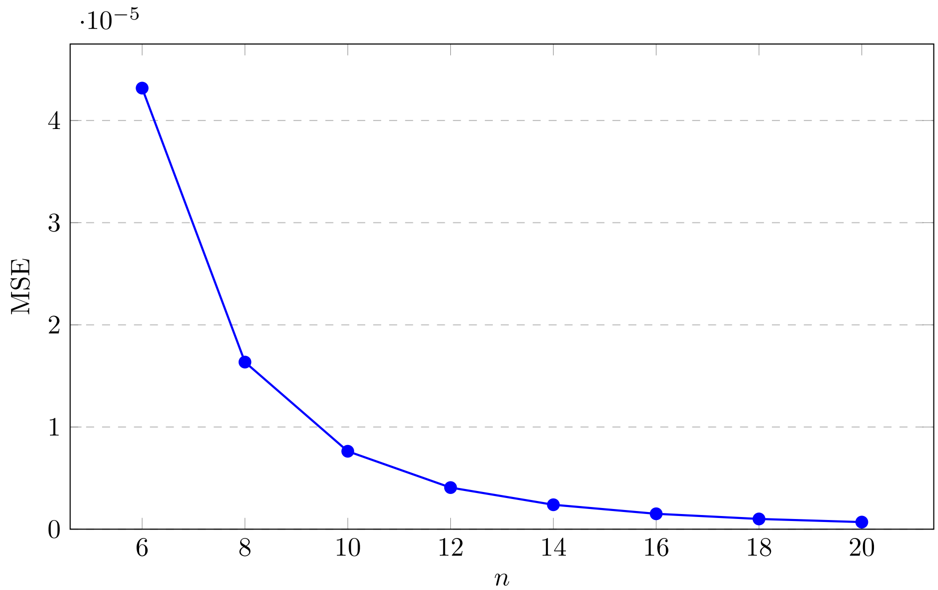

Figure 7: Evolution of MSE between Binomial and Normal distributions.

4 Discussion

The results obtained from the simulations clearly demonstrate the progressive convergence between the empirical data and the theoretical predictions as the number of balls N and the number of levels n increase. In Figures 1 and 2, where the board has $n = 6$ levels, the observed frequencies already exhibit a roughly binomial shape. However, some deviations are visible when $N = 50$, which can be attributed to sampling noise. When N increases to 1000, these discrepancies decrease significantly and the empirical distribution closely matches the binomial and normal curves. This confirms that a larger number of trials improves the statistical reliability of the experiment.

Figures 3 and 4, corresponding to $n = 20$ levels, provide additional information. For small N, the empirical histogram still approximates the theoretical predictions, but random fluctuations remain noticeable due to limited sampling. When $N = 1000$, the resulting distribution becomes smooth and aligns almost perfectly with both the binomial and normal models. This outcome illustrates the Central Limit Theorem in practice: as n grows, the binomial distribution tends toward normality, and thus the difference between both theoretical models becomes negligible.

The Mean Squared Error (MSE) plots in Figures 5 to 7 quantitatively support these observations. Figure 5 shows that the MSE between the observed and theoretical distributions decreases as n increases, confirming that deeper boards produce results that are more consistent with theoretical expectations. Similarly, Figure 6 shows that the MSE decreases with higher N, as larger sample sizes reduce statistical noise. Finally, Figure 7 reveals that the MSE between the binomial and normal distributions is very small (on the order of 10^{-5}) and decreases with n, highlighting their increasing similarity for large numbers of levels. The reason why the graph of this MSE as a function of N has not been included is that this parameter does not influence it at all, so for fixed n the graph would have been a horizontal line.

In general, these findings validate both the accuracy of the simulation implemented and the theoretical principles underlying the Galton board. The empirical results reproduce the expected transition from discrete binomial behavior to a continuous normal approximation, confirming that the simulation accurately models the stochastic dynamics of the system.