Universitat Politècnica de Catalunya

Facultat d'Informàtica de Barcelona (FIB)

GALTON BOARD SIMULATION AND DISTRIBUTION COMPARISON

Randomized Algorithms

IVÁN SÁNCHEZ GALA, SERGI GUIMERÀ ROIG

Conrado Martínez Parra

Barcelona, Spain
Oct, 2025

1 Introduction

1.1 Statement

In this first assignment of *Randomized Algorithms (RA–MIRI)*, we are required to simulate a Galton board (also known as a Galton box, quincunx, or bean machine). This device is a classical physical model used to illustrate the Central Limit Theorem (CLT) — the principle that, as the number of random independent events increases, the resulting distribution tends to approximate a normal (Gaussian) distribution, even if the original variables themselves are not normally distributed.

1.2 Galton board

In the Galton board, a large number of balls are dropped from the top of a board filled with interleaved rows of pegs. As each ball descends, it randomly bounces left or right with equal probability ($\frac{1}{2}$). After passing through n levels of pegs, the final position of each ball corresponds to the number of right bounces it took, which follows a binomial distribution $\mathrm{Bin}(n, \frac{1}{2})$. When the number of levels n and the number of balls N are sufficiently large, this discrete binomial distribution becomes well approximated by a normal distribution $\mathscr{N}(\mu = n/2, \sigma^2 = n/4)$.

1.3 Goal

The goal of this assignment is to simulate this experiment computationally and to study the match between the experimental binomial data and the theoretical normal values. In particular, we analyze how increasing the number of levels (n) and the number of balls (N) affects the accuracy of the normal approximation of the binomial distribution evaluated using the **Mean Squared Error (MSE)** and the **Chi-squared** (χ^2) **test**.

1.4 Implementation

A Python program was developed to both simulate the Galton box and perform the statistical experiments automatically, as there the experiments are not deterministic, we make repetitions for each experiment configuration (n and N values) and to allow reproducibility the seeds were previously chosen. The implementation stores the results (MSE, χ^2 p-values) in a CSV file for further analysis, and generates comparative plots showing the empirical (binomial) and theoretical (normal) distributions. As there are repetitions the mean for MSE and min for χ^2 are used, moreover, due to numerical reasons we plotted the $100 \cdot RMSE$ instead of the MSE.

The complete source code, together with setup instructions and generated plots, is available in the following GitHub repository:

`https://github.com/S3RXxX/RA_A1`

1.5 Report structure

The remainder of this report includes:

- **Mathematical background** section describing the relationship between the binomial and normal distributions and the formulas for used metrics.
- **Experimental analysis** of the simulation results.
- **Conclusions** on the convergence of the empirical data to the theoretical model.

2 Mathematical background

2.1 Binomial distribution

2.1.1 Probability of i

Consider a Galton board with n levels of pegs. Each ball starts from the top at position $(0,0)$ and, at every level, it encounters a peg that can deflect it either to the left or to the right.

At each step of its descent:

- The ball moves left with probability $\frac{1}{2}$,
- The ball moves right with probability $\frac{1}{2}$.

After n levels, the ball has performed exactly n independent binary trials. Let us define a random variable:

$$Y = \text{number of right moves after } n \text{ levels.}$$

Each move to the right can be modeled as a Bernoulli random variable (equivalently for the left as in the Python implementation):

$$Z_j = \begin{cases} 1, & \text{if the ball goes right at level } j, \\ 0, & \text{if the ball goes left at level } j. \end{cases}$$

with

$$\mathbb{P}(Z_j = 1) = \mathbb{P}(Z_j = 0) = \tfrac{1}{2}.$$

Since all n movements are independent,

$$Y = Z_1 + Z_2 + \cdots + Z_n$$

is the sum of n independent Bernoulli trials with success probability $p = \frac{1}{2}$.

Therefore, Y follows a **binomial distribution**:

$$Y \sim \text{Bin}(n, p = \frac{1}{2})$$

The probability that the ball makes exactly i right moves (and thus $n-i$ left moves) is given by the binomial probability mass function:

$$p_{i,n} = \mathbb{P}(Y = i) = \binom{n}{i} p^i (1-p)^{n-i} = \binom{n}{i} \left(\frac{1}{2}\right)^i \left(1 - \frac{1}{2}\right)^{n-i} = \binom{n}{i} \left(\frac{1}{2}\right)^n$$

Each unique sequence of i right and $n-i$ left moves corresponds to one possible path that ends in the cell $(i, n-i)$. The number of such distinct paths is given by the binomial coefficient $\binom{n}{i}$, since we can choose in $\binom{n}{i}$ ways which i of the n steps are to the right.

Thus, the probability that a ball starting at $(0,0)$ ends at cell $(i, n-i)$ after n levels is:

$$\boxed{p_{i,n} = \binom{n}{i} \left(\frac{1}{2}\right)^n.}$$

Now, we can see that if i is fixed while $n \to \infty$, the binomial coefficient grows polynomially in n (indeed $\binom{n}{i} \sim n^i/i!$), whereas $(1/2)^n$ decays exponentially. Hence for any fixed i,

$$p_{i,n} = \binom{n}{i} \left(\tfrac{1}{2}\right)^n \leq \frac{n^i}{i!} \left(\tfrac{1}{2}\right)^n \xrightarrow[n\to\infty]{} 0$$

Thus an individual point probability tends to zero. Intuitively, as n grows the distribution spreads on a larger support $\{0,\dots,n\}$, so mass at any particular fixed index vanishes.

2.1.2 Mean

For a single Bernoulli trial Z_i, the expected value is:

$$\mathbb{E}[Z_i] = 1 \cdot p + 0 \cdot (1-p) = p$$

By linearity of expectation

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

and since each trial has the same expectation p, we can formulate it as:

$$\mathrm{E}[Y] = \sum_{i=1}^{n} p = np \tag{1}$$

2.1.3 Variance

The variance of a sum of independent random variables equals the sum of their variances:

$$\mathrm{Var}[Y] = \sum_{i=1}^{n} \mathrm{Var}[Z_i].$$

So we only need to compute $\mathrm{Var}[Z_i]$ for one Bernoulli variable.

By definition:

$$\mathrm{Var}[X_i] = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2.$$

Since $Z_i \in \{0,1\}$, we have $Z_i^2 = Z_i$, and therefore:

$$\mathbb{E}[Z_i^2] = \mathbb{E}[Z_i] = p.$$

Then:

$$\mathrm{Var}[Z_i] = p - p^2 = p(1-p).$$

Finally, since all trials are independent:

$$\boxed{\mathrm{Var}[Y] = n\,\mathrm{Var}[Z_i] = np(1-p) = \frac{n}{4}} \tag{2}$$

2.1.4 Normal Approximation

The approximation is proven via the CLT and we will not discuss it here. As we have seen the distribution parameters in Eq. (1) and Eq. (2) the approximation would be $\mathcal{N}(np, np(1-p)) = \mathcal{N}(n/2, n/4)$.

2.2 Metrics

2.2.1 Mean Squared Error (MSE)

$$\text{MSE} = \frac{1}{n+1}\sum_{i=0}^{n}(\hat{p}_i - p_i)^2 \tag{3}$$

2.2.2 Root Mean Squared Error (RMSE)

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n+1}\sum_{i=0}^{n}(\hat{p}_i - p_i)^2} \tag{4}$$

2.2.3 Chi-squared Statistic (χ^2)

$$\chi^2 = \sum_{i=0}^{n}\frac{(O_i - E_i)^2}{E_i}, \tag{5}$$

where:

- O_i are the observed counts (from the experiment),
- E_i are the expected counts (from the theoretical normal distribution)

3 Experimentation

3.1 Studying the effects of large boards

As we can see in Fig. 1, Fig. 2,Fig. 3 and Fig. 4, the error between the experimental (binomial) and theoretical (normal) distributions decreases as either the number of balls N or the number of levels n increases (notice that error Fig. 1 > error Fig. 3, similarly for the other two figures).

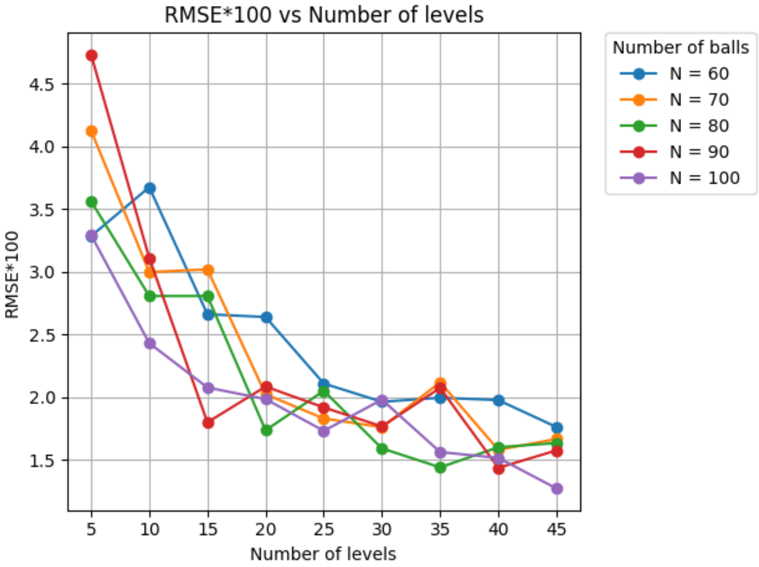

Figure 1: Evolution of the error when the number of levels is increased between $[5, 45]$ for number of balls $[60, 100]$

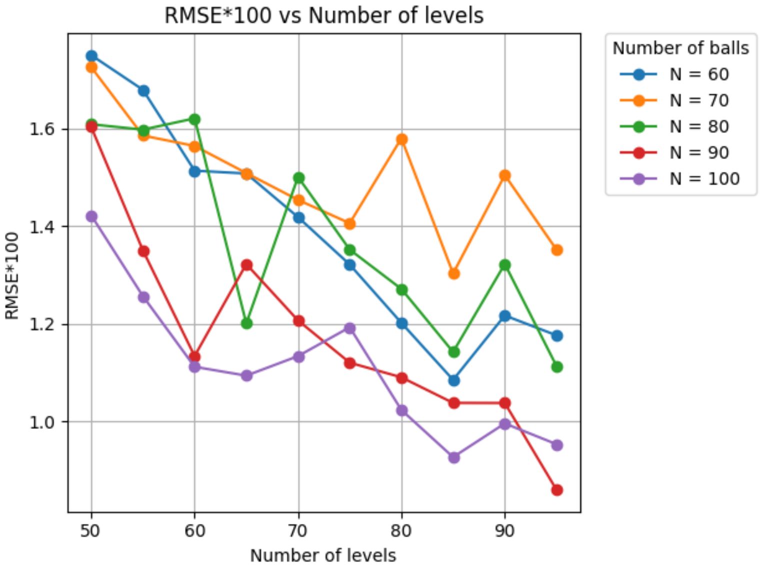

Figure 2: Evolution of the error when the number of levels is increased between $[50, 95]$ for number of balls $[60, 100]$

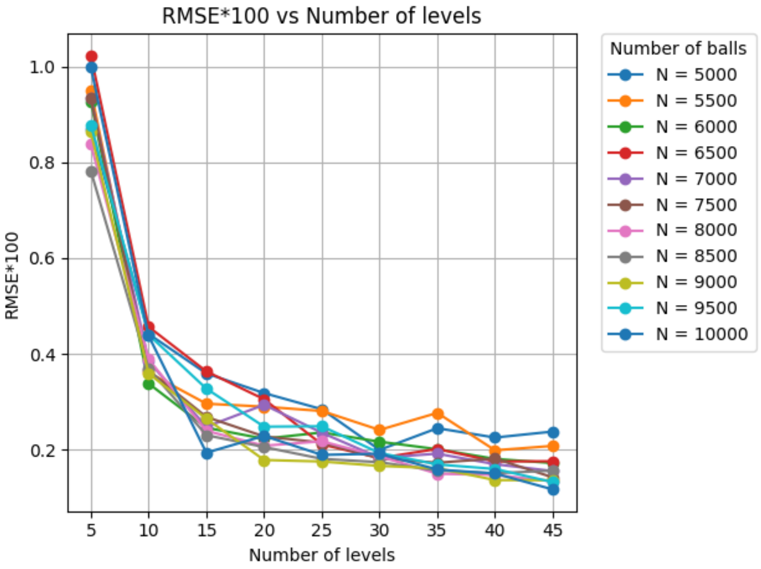

Figure 3: Evolution of the error when the number of levels is increased between $[5, 45]$ for number of balls $[5000, 10000]$

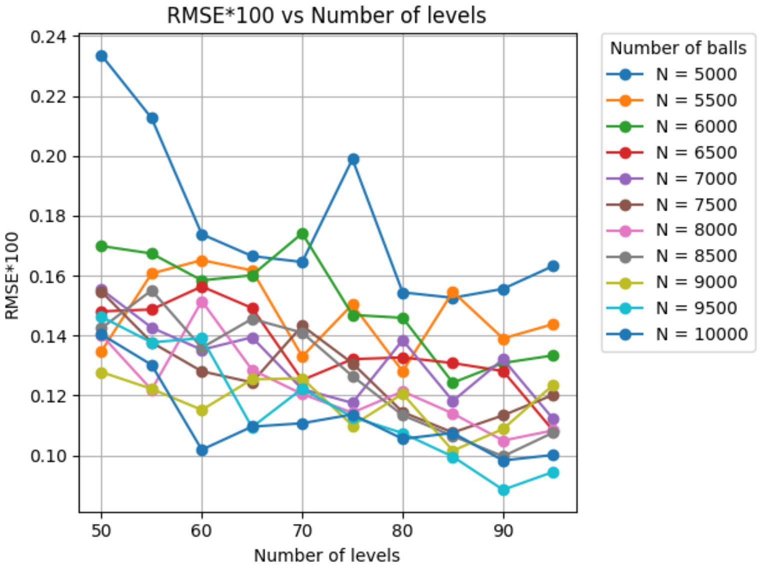

Figure 4: Evolution of the error when the number of levels is increased between $[50,95]$ for number of balls $[5000,10000]$

The effect of the number of levels n can be understood from the properties of the binomial distribution. As n increases, the variance of the distribution grows linearly according to Eq. (2) and therefore the probability associated with each individual outcome $p_{i,n}$ becomes smaller as shown in Fig. 5 and Fig. 6. This makes the discrete binomial distribution smoother and its shape more similar to the continuous normal curve. Consequently, the approximation error between the two distributions decreases. Furthermore, since the error metric (e.g. the RMSE) is normalized by the number of possible outcomes $(n+1)$, as shown in Eq. (4), increasing n effectively distributes any residual discrepancy over a larger number of bins, thus reducing the overall average error.

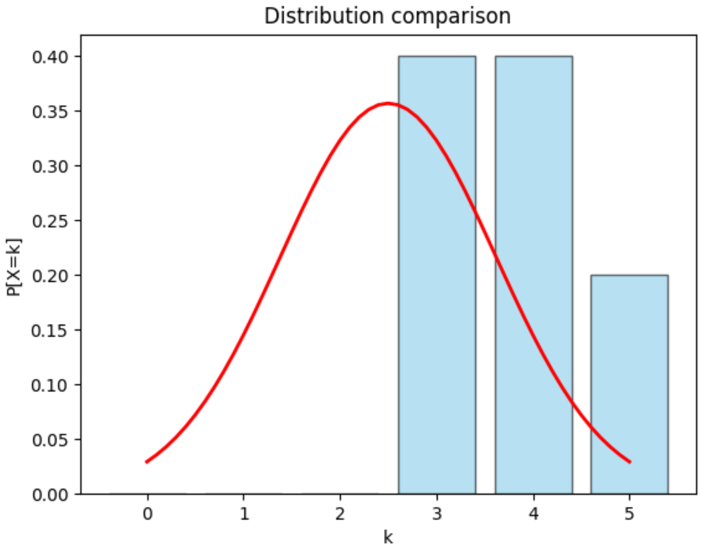

Figure 5: Sample distribution vs theoretical distribution with 5 levels and throwing 5 balls.

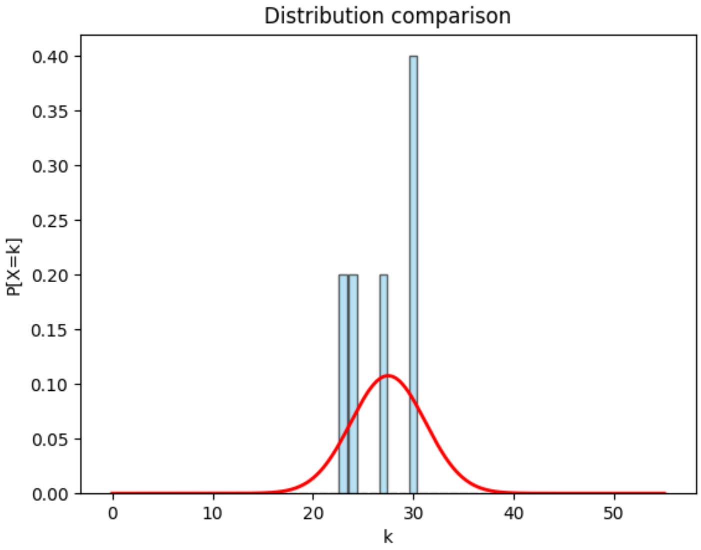

Figure 6: Sample distribution vs theoretical distribution with 55 levels and throwing 5 balls.

Since we had problems calculating the χ^2 (due to insufficient levels with less than 5 samples) for number of levels equal to 5 or 10, we imputed the missing p-value to 0 and later verified by checking the plots (that can be found in the git repository) that these levels follow the same pattern as the others (when the number of balls increases, it resembles more to a normal distribution). Once we had all the values, we obtained Fig. 7, Fig. 8 and Fig. 9 detecting that when the number of balls is lower than the number of levels, the experimental data is less likely to follow a normal distribution, hence the approximation is worse.

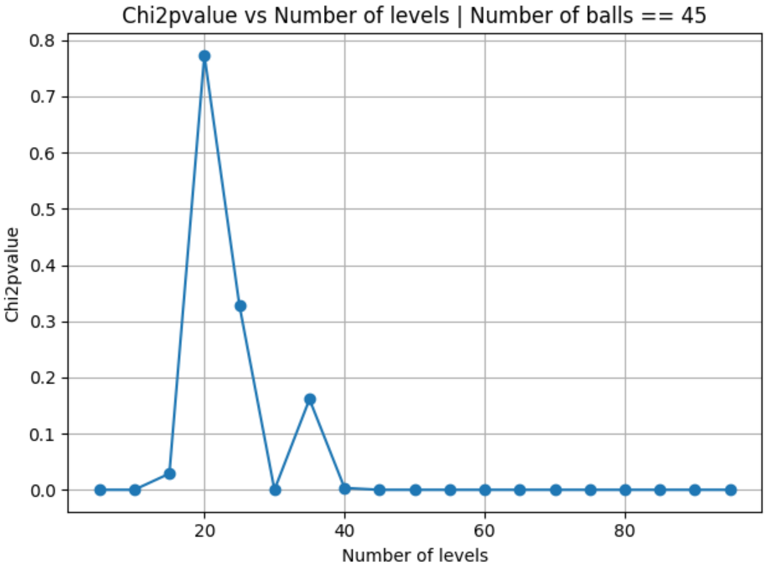

Figure 7: χ^2 p-value change as number of levels increases when number of balls is 45.

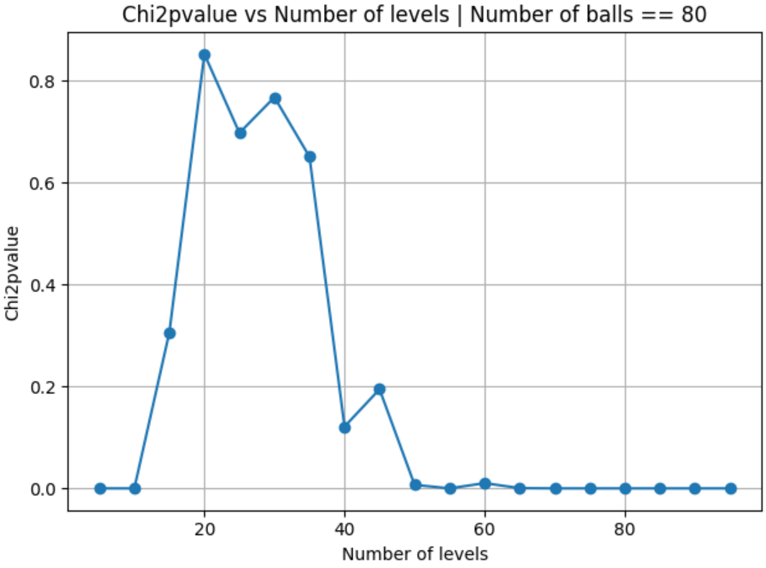

Figure 8: χ^2 p-value change as number of levels increases when number of balls is 400.

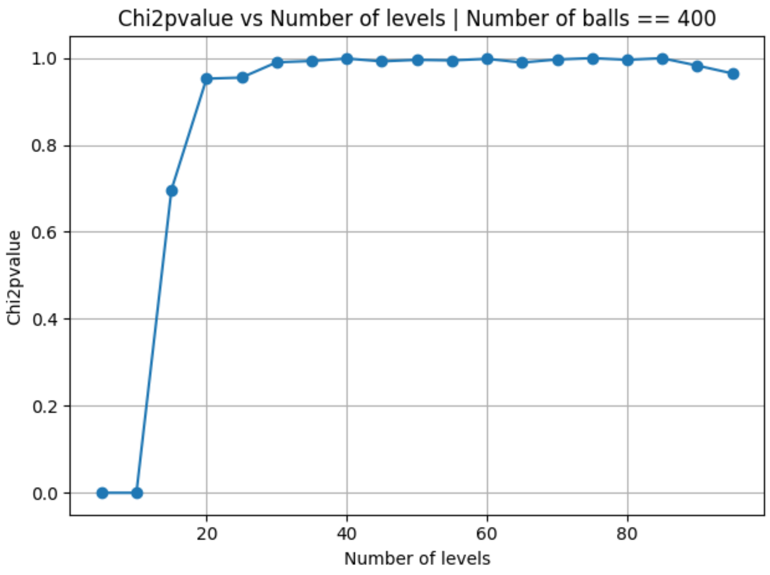

Figure 9: χ^2 p-value change as number of levels increases when number of balls is 400.

3.2 Studying the effects of a large number of experiments

As shown in Fig. 10 and Fig. 11, we can observe that when the number of balls increase, the error decrease, hence doing better approximations.

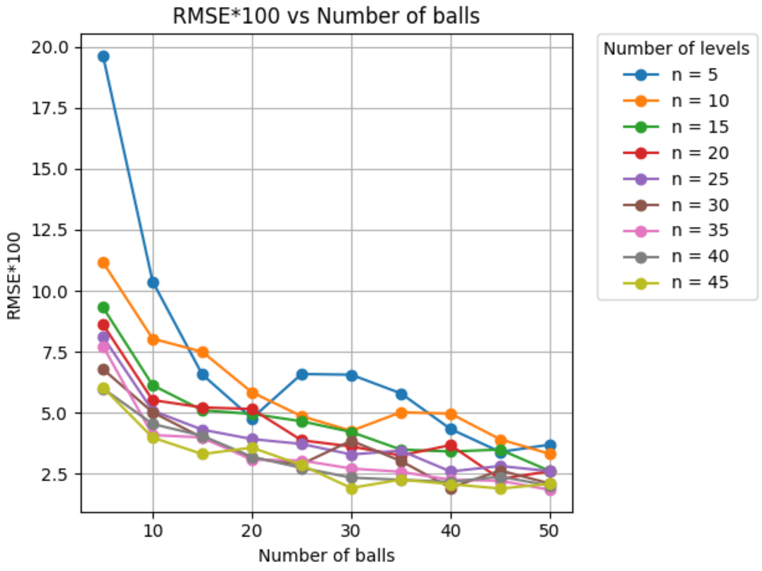

Figure 10: Evolution of the error when the number of balls is increased between $[5,50]$ for number of levels $[5,45]$

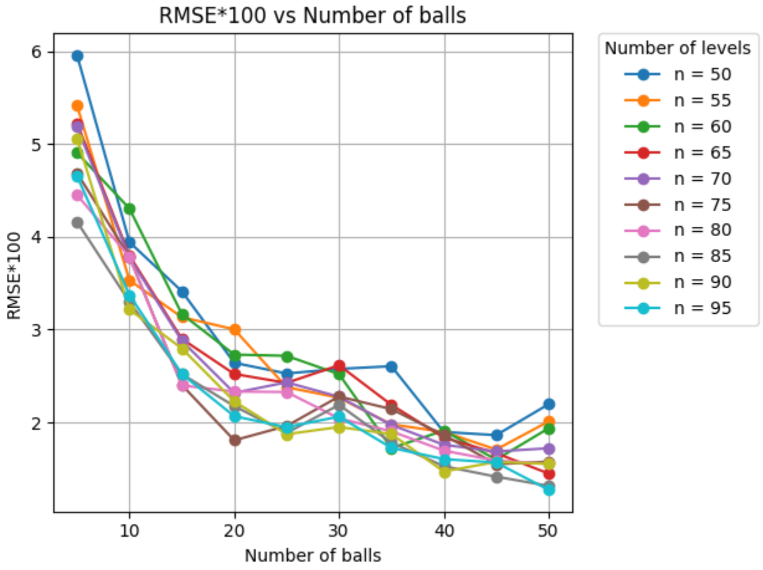

Figure 11: Evolution of the error when the number of balls is increased between $[5,50]$ for number of levels $[50,95]$

Checking the p-values of the χ^2 test in Fig. 12, we see that when the number of balls increases, then it is more similar to a normal distribution. For every level, the corresponding plot is similar to the one in Fig. 12, with the exception of number of levels equal to 5 or 10 that has the problem mentioned above.

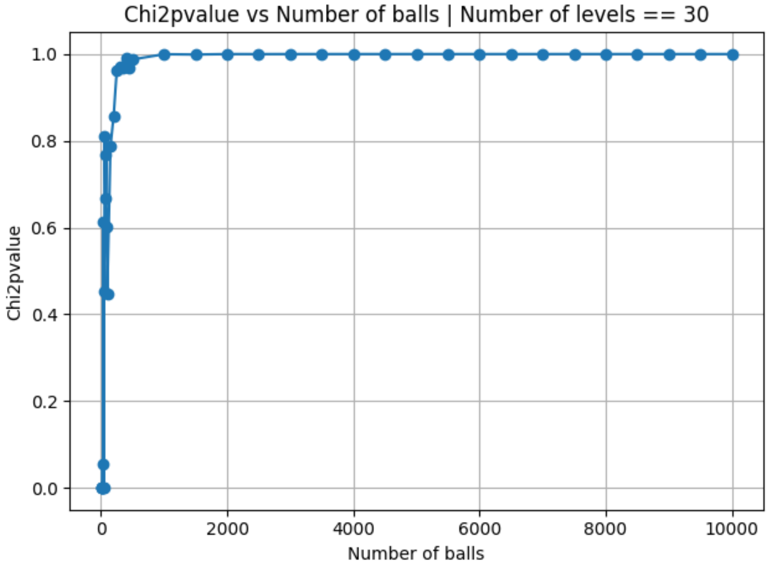

Figure 12: χ^2 p-value change as number of balls increases when number of levels is 30

4 Conclusions

With this study we have experimentally illustrated the Central Limit Theorem (CLT) through the simulation of a Galton Board. By performing repeated experiments and comparing the resulting empirical (binomial) distributions with the corresponding theoretical normal distributions, we observed that the approximation improves significantly as both the number of balls N and the number of levels n increase.

In particular, when the number of balls is sufficiently large relative to the number of levels, and both are large, the empirical distribution of the final ball positions closely follows a normal distribution with parameters $\mu = \frac{n}{2}$ and $\sigma^2 = \frac{n}{4}$. This convergence was quantitatively verified using the χ^2 test and the RMSE.

Overall, the experimental results confirm the theoretical prediction that the binomial distribution $\mathrm{Bin}(n, \frac{1}{2})$ tends toward the normal distribution $\mathscr{N}(\frac{n}{2}, \frac{n}{4})$ as n grows, thereby providing a concrete computational demonstration of the Central Limit Theorem in practice.