Classifying Random Walks: Continuous and Discrete Probabilistic Models for Human vs. Computer Patterns

Zeynep Eylül YAĞCIOĞLU

CS 109 Challenge
Fall 2024

1 Motivating Problem

The goal of this project is to classify a given random walk as either human-generated (H) or computer-generated (C). Let W represent the observed random walk consisting of 100 steps, and let r_i denote the i-th run-length within the walk.

1.1 Bayesian Framework for Classification

Using Bayes' theorem, the posterior odds of H versus C given the observed walk W can be expressed as:

$$\frac{P(H \mid W)}{P(C \mid W)} = \frac{\frac{P(W|H)P(H)}{P(W)}}{\frac{P(W|C)P(C)}{P(W)}} = \frac{P(W \mid H)P(H)}{P(W \mid C)P(C)}.$$

Assuming equal priors ($P(H) = P(C)$), the expression simplifies to:

$$\frac{P(H \mid W)}{P(C \mid W)} = \frac{P(W \mid H)}{P(W \mid C)}.$$

This means the classification decision depends on comparing the likelihood of the observed walk under the human model versus the computer model. To calculate these likelihoods, probabilistic models for both $P(W \mid H)$ and $P(W \mid C)$ are required.

1.2 Modeling Computer-Generated Walks

Computer-generated random walks are governed by well-defined probabilistic patterns due to their underlying mechanics. Key characteristics include:

Runs as Bernoulli Trials: A random walk can be represented as a sequence of Bernoulli trials, where each step is either forward ($+1$) or backward (-1) with equal probability ($p = 0.5$). The total displacement after 100 steps can thus be modeled as a Binomial random

variable:

$$X \sim \text{Binomial}(n = 100, p = 0.5),$$

where X represents the net displacement (number of forward steps minus backward steps).

Run-Lengths Modeled as a Geometric Distribution: The length of consecutive steps in the same direction (run-lengths) can be modeled using a geometric distribution:

$$R \sim \text{Geo}(p = 0.5),$$

where R represents the number of consecutive forward or backward steps in a run. The probability mass function (PMF) is given by:

$$P(R = r) = (1 - p)^{r-1}p, \quad r \geq 1.$$

For $p = 0.5$, this simplifies to:

$$P(R = r) = 0.5^r.$$

Using these theoretical distributions, the likelihood $P(W \mid C)$ for a given walk W can be calculated by decomposing W into its constituent runs and steps.

1.3 Challenges with Human-Generated Walks

In contrast to computer-generated walks, human-generated random walks lack a clear underlying mathematical structure. Human behavior introduces variability, biases, and patterns that cannot be easily modeled using standard probabilistic distributions. This poses two primary challenges:

1. Unknown Distribution: Unlike the geometric and binomial models for computer-generated walks, there is no predefined probabilistic model for human-generated runs. The characteristics of $P(W \mid H)$ must be empirically derived from collected data.

2. Classifying Walks: How can we build a human model that accurately distinguishes human-generated walks from computer-generated walks? This requires identifying unique

statistical features of human walks and incorporating them into the model.

1.4 Proposed Solution

To address these challenges, I developed two probabilistic models for human-generated random walks:

Continuous Probability Model: This model assumes that run-lengths within a walk and between different walks are independent and identically distributed (IID). Using this assumption, run-lengths are modeled as a continuous random variable, and their distribution is estimated using bootstrapping and CLT.

Discretized Probability Model: This model builds an empirical probability mass function (PMF) for human run-lengths by discretizing them into predefined, non-uniform buckets with varying sizes. The bucket ranges are designed to emphasize differences between human and computer-generated walks by tailoring the structure to highlight patterns unique to human behavior.

2 Methodology

2.1 Data Preprocessing and Terminology

Before delving into the Continuous Probability Analysis: Human-Generated Walks and Discretized PMF Analysis, this section outlines the preprocessing steps and introduces the terminology used throughout the models.

2.1.1 Raw Walks

Each trajectory is initially represented as a sequence of positions (integers). For example, a trajectory may look like:

$$w_1 = \{0, 1, 2, 1, 0, -1, -2\}$$

where w_1 represents one walk of 100 steps.

2.1.2 Step Conversion

I converted the positional data into directional steps, denoted as $\hat{w}_i$, where each step is assigned:

$$\text{Step}[i] = \begin{cases} +1 & \text{if Position}[i] > \text{Position}[i-1] \\ -1 & \text{if Position}[i] < \text{Position}[i-1]. \end{cases}$$

For the trajectory above, the corresponding steps become:

$$\hat{w}_1 = \{+1, +1, -1, -1, -1, -1\}.$$

2.1.3 Walk Division

The dataset is divided into smaller walks, each containing 100 steps, resulting in a dataset $\hat{W}$:

$$\hat{W} = \{\hat{w}_1, \hat{w}_2, \hat{w}_3, \ldots, \hat{w}_n\}, \quad \text{where } |\hat{w}_i| = 100.$$

2.1.4 Run-Length Extraction

For each walk $\hat{w}_i$, I identified **run-lengths**—defined as sequences of consecutive steps in the same direction. The run-lengths for a given walk are denoted as:

$$R_i = \{r_1, r_2, \ldots, r_k\},$$

where R_i represents the run-lengths of walk $\hat{w}_i$, and r_j is the length of the j^{th} run. Note that the number of runs k may vary for each walk.

For example, given the steps:

$$\hat{w}_1 = \{+1, +1, -1, -1, -1, +1\},$$

the run-lengths become:

$$R_1 = \{2, 3, 1\}.$$

2.2 Identifying the Distinguishing Features

Identifying distinguishing features was a critical challenge in this project. In the literature, much of the work on random walks focuses on displacement as a primary feature. I initially considered several quantities as potential distinguishing features, including:

- **Displacement:** The net difference in position after 100 steps.
- **Maximum Displacement:** The largest deviation from the origin observed during the walk.
- **Distribution of Run Directions:** The balance of forward and backward runs in each walk.
- **Time or Speed of Input:** The rate at which steps are entered or recorded.

However, when I performed p-tests on these variables, their respective p-values exceeded 0.05. This supported the null hypothesis that these variables do not exhibit statistically significant differences between human- and computer-generated walks.

2.2.1 Run-Lengths as Distinguishing Features

As a result, I focused on **mean run-lengths** and **run-lengths** as the primary distinguishing features. These variables not only yielded p-values less than 0.05, indicating statistically significant differences between human- and computer-generated walks, but also offered compatibility with both continuous and discrete analyses, as further discussed in Evaluation

of Using the Average Version of CLT and Evaluation of Using Varying Bucket Sizes in the Discrete Model. Thus, I identified run-lengths as the distinguishing feature for classifying human- and computer-generated walks.

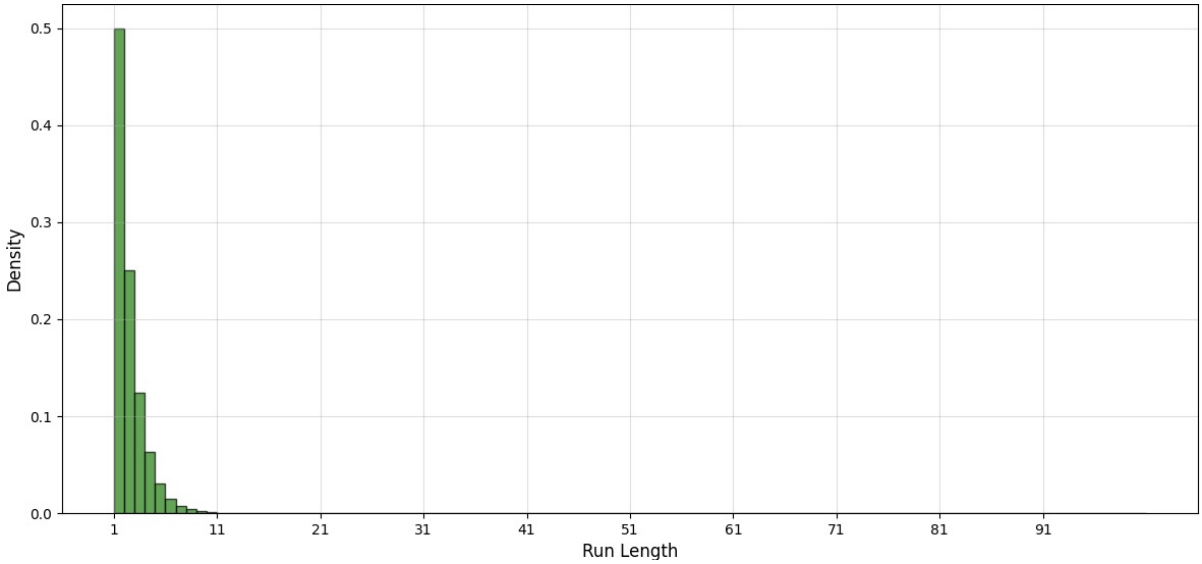

Figure 1: Histogram of Run Lengths in Computer-Generated 100-Step Walks

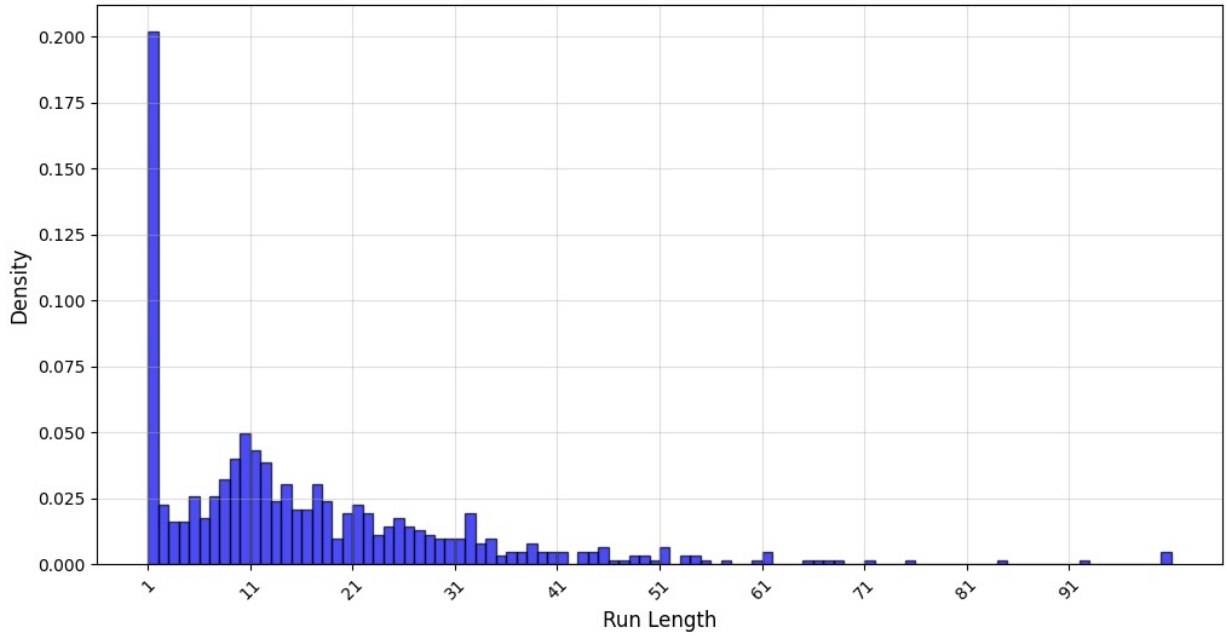

Figure 2: Histogram of Run Lengths in Human-Generated 100-Step Walks

2.3 Continuous Probability Analysis: Human-Generated Walks

2.3.1 Bootstrap Estimation of Model Parameters

Key Random Variables: Each run-length is treated as a random variable. The **mean run-length per walk** is defined as:

$$\overline{R}_i = \frac{\text{Sum of run lengths in walk } \hat{w}_i}{\text{Number of runs in walk } \hat{w}_i}.$$

Here, $\overline{R}_i$ represents the mean run-length of walk $\hat{w}_i$.

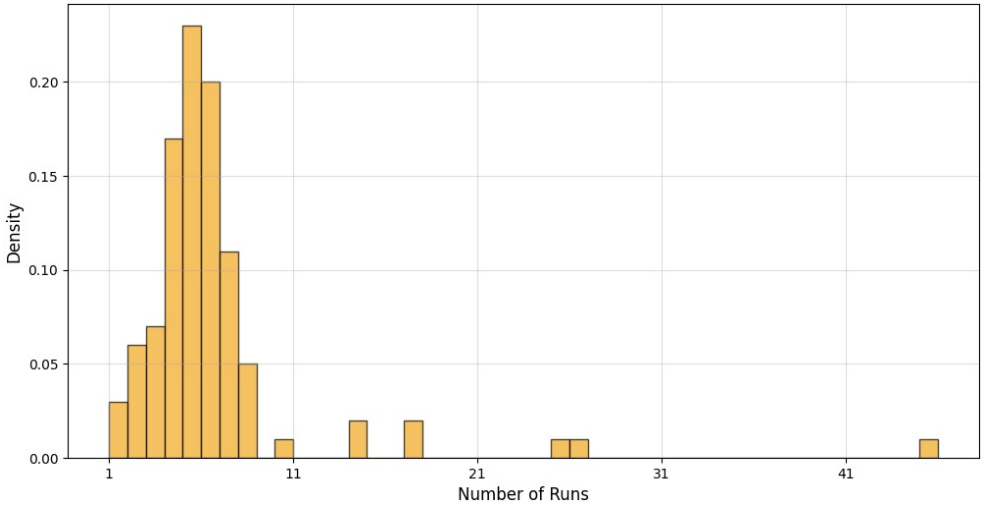

Figure 3: Histogram of Number of Runs In 100-Step Walks

A Challenge: Non-constant Number of Runs: The number of runs per walk (n_i) varies across walks. This creates variability in estimating the variance of mean run-lengths ($\overline{R}_i$). To address this, I performed the following steps during bootstraping:

- Sample a random n_i (number of runs in a walk) from the dataset.
- Sample n_i run-lengths (with replacement) to form a synthetic walk.
- Compute the mean ($\overline{R}_i$) for this synthetic walk.

This process was repeated for 100,000 iterations to estimate:

$$\mu_H = \text{Average of bootstrapped means}, \quad \sigma_H^2 = \text{Average of bootstrapped variances}.$$

Innovative Adjustment Using CLT The Central Limit Theorem (CLT) states that the distribution of the sample mean ($\overline{R}_i$) approaches normality as n_i increases. To calculate the standard deviation of the mean run-lengths, I introduced n_i (a random variable) into the variance formula:

$$\sigma_H^2 = \frac{\text{Bootstrap Variance}}{\mathbb{E}[n_i]}.$$

Incorporating n_i as a random variable in the variance calculation significantly improves the robustness of the model. This adjustment ensures that the variance accounts for variability in the number of runs per walk.

2.3.2 Human Model Construction

Using results from the bootstrap analysis, I found **Mean Run-Length** as $\mu_H = 16.05$ and **Adjusted Variance** as $\sigma_H^2 = 5.55$ (derived using the CLT with varying n_i, accounting for the variability in the number of runs per walk). Then, the normal distribution representing human-generated walks can be expressed as:

$$\overline{R}_i \sim \mathcal{N}(\mu_H = 16.05, \sigma_H^2 = 5.55).$$

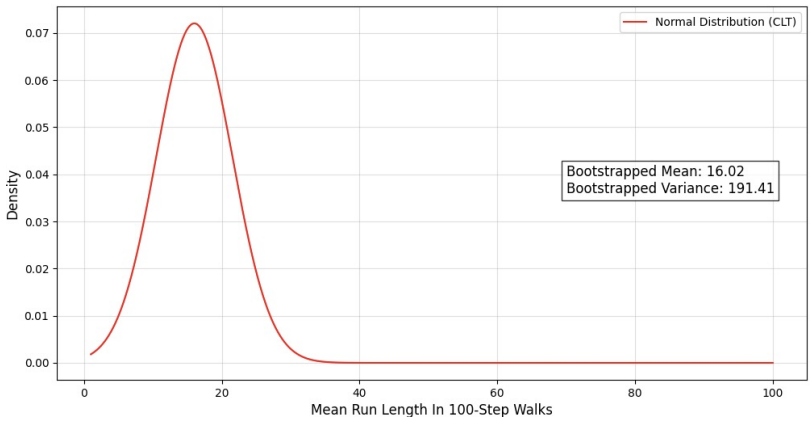

Figure 4: CLT Approximation for Human-Generated Walks

2.3.3 Likelihood Computation

Random Variables and Assumptions I assumed that the run-lengths within a walk and across different walks are independent and identically distributed (IID). This assumption simplifies the likelihood calculation.

Likelihood for Human Model For a given walk, with observed run-lengths $\{r_1, r_2, \ldots, r_k\}$, the likelihood under the human model is:

$$P(W \mid H) = \prod_{i=1}^{k} f_H(r_i), \quad \text{where } f_H(r_i) = \text{PDF}_{\mathcal{N}}(r_i; \mu_H, \sigma_H^2).$$

Log-Likelihood For computational stability, I calculated the log-likelihood:

$$\log P(W \mid H) = \sum_{i=1}^{k} \log f_H(r_i).$$

2.4 Continuous Probability Analysis: Computer-Generated Walks

By applying CLT to both human and computer models, I maintained consistency in how the likelihoods were calculated.

2.4.1 Theoretical Properties of Computer Walks

Run-Length Distribution: Run-lengths in computer-generated walks are modeled using the geometric distribution Geo($p = 0.5$), which represents the number of consecutive steps in the same direction. The probability mass function (PMF) for a run of length r is given by:

$$P(R = r) = (1 - p)^{r-1} \cdot p, \quad r \geq 1.$$

For $p = 0.5$, this simplifies to:

$$P(R = r) = 0.5^r.$$

Aggregating Run-Lengths in 100-Step Walks: In a 100-step walk, the overall structure can be analyzed by aggregating multiple runs. To model the *mean run-length per walk*, we use the CLT, which approximates the distribution of the sample mean as normal:

$$\overline{R}_C = \frac{\sum_{i=1}^{n} r_i}{n},$$

where r_i are the individual run-lengths, and n is the number of runs in the walk. From the properties of the geometric distribution (Geo($p = 0.5$)), we know:

- The **mean** is $\mathbb{E}[\overline{R}_C] = \mu_C = 2$.
- The **variance** of the sample mean is $\text{Var}(\overline{R}_C) = \frac{\sigma_C^2}{n}$, where $\sigma_C^2 = 2$.
- The **standard deviation** of the sample mean is $\sigma_{\overline{R}_C} = \sqrt{\frac{\sigma_C^2}{n}}$.

Adjusting for Empirical n: The parameter n, representing the number of runs in a walk, varies between theoretical assumptions and empirical observations:

Theoretical Assumption: Under ideal conditions for computer-generated walks, where the expected run-length is 2, the total number of runs in a 100-step walk is:

$$n = \frac{\text{Total Steps}}{\text{Expected Run Length}} = \frac{100}{2} = 50.$$

Using this value of n, the standard deviation of the mean run-length is:

$$\sigma_{\overline{R}_C} = \sqrt{\frac{2}{50}} = 0.2.$$

Empirical Observation: While analyzing human-generated walks, I observed that the mean number of runs per walk was significantly smaller, with $n = 6.24$. To ensure consistency in comparisons between human and computer models, I applied this empirical n to the computer-generated walk analysis. Although the theoretical properties of computer-generated walks remain unchanged, using the empirical n aligns the models and accounts for

variability in real-world observations. With $n = 6.24$, the standard deviation becomes:

$$\sigma_{\overline{R}_C} = \sqrt{\frac{2}{6.24}} \approx 0.57.$$

This approach ensures that the comparison between $P(W \mid H)$ and $P(W \mid C)$ is conducted under consistent conditions, highlighting the distinguishing characteristics of human and computer-generated random walks.

2.4.2 CLT Approximation & Likelihood for Computer-Generated Walks

As a result, the mean run-length per walk follows a normal distribution under CLT:

$$\overline{R}_C \sim \mathcal{N}(\mu_C = 2, \sigma_C^2 = 0.57^2).$$

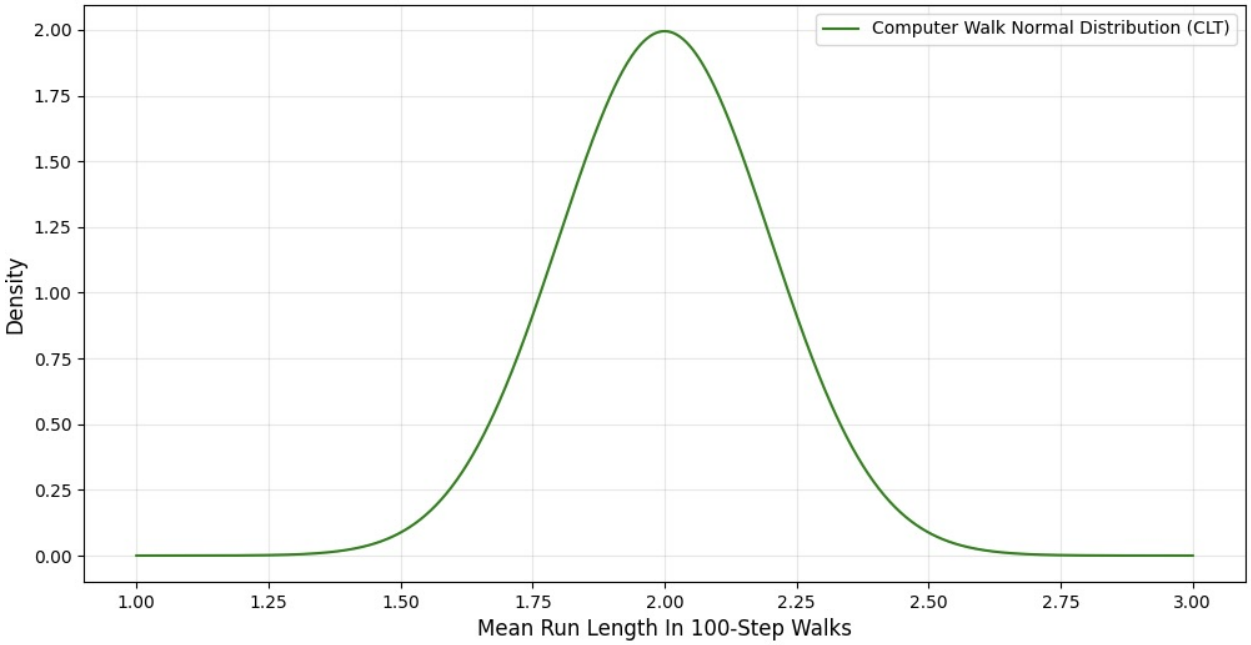

Figure 5: CLT Approximation for Computer-Generated Walks (Geo(0.5))

Likelihood for Computer Model: For a given walk, with observed run-lengths $\{r_1, r_2, \ldots, r_k\}$, the likelihood under the computer model is:

$$P(W \mid C) = \prod_{i=1}^{k} f_C(r_i), \quad \text{where } f_C(r_i) = \text{PDF}_{\mathcal{N}}(r_i; \mu_C = 2, \sigma_C^2 = 0.57^2).$$

Log-Likelihood: To avoid numerical underflow, I calculated the log-likelihood:

$$\log P(W \mid C) = \sum_{i=1}^{k} \log f_C(r_i).$$

2.5 Evaluation of Continuous Model

With the final model, when an input walk is provided on the website, the model identifies the observed run-lengths and calculates the joint probability of observing these run-lengths under the human- and computer-generated models. The likelihood ratio is expressed as:

$$\frac{P(W \mid H)}{P(W \mid C)} = \prod_{i=1}^{n} \frac{f_H(r_i)}{f_C(r_i)},$$

where $f_H(r_i)$ and $f_C(r_i)$ are the respective probability density functions for the human and computer models.

For computational efficiency, this is simplified by taking the logarithm:

$$\log \left(\frac{P(H \mid W)}{P(C \mid W)} \right) = \sum_{i=1}^{n} \log \left(\frac{f_H(r_i)}{f_C(r_i)} \right).$$

Evaluation of Using the Average Version of CLT In this model, the reference is not the individual run-length itself but, for a given run-length in a walk, how likely it is to match the mean of a walk. By leveraging the Central Limit Theorem (CLT), the model transitions from analyzing individual run-lengths to focusing on their averages, which decreases the variance in both the human- and computer-generated models. This reduction in variance is advantageous for this application because it enhances the model's ability to differentiate between the two classes.

When comparing the histograms of human- and computer-generated walks:

- Both human and computer walks include smaller run-lengths.
- However, overall, human-generated walks exhibit larger run-lengths compared to the computer model.

Thus:

- While calculating $P(W \mid H)$, the longer runs in a walk are rewarded more heavily, while shorter runs contribute significantly less.
- Conversely, while calculating $P(W \mid C)$, the shorter runs are rewarded more heavily, while longer runs contribute much less.

This behavior aligns with the empirical patterns observed in the histograms of human- and computer-generated data. Therefore, using the mean and the CLT effectively highlights the key distinguishing features of the two classes, significantly enhancing the accuracy of the probability analysis.

Reflections on Design Choices: Initially, I worried that focusing on the mean might obscure valuable details about individual run-lengths, potentially weakening the model. However, when comparing the histograms and the expected patterns, it became clear that focusing on the mean was a strength. The mean run-length provides a robust and interpretable feature that effectively differentiates human- and computer-generated walks while ensuring computational efficiency in the likelihood calculations.

2.6 Discretized PMF Analysis

For human-generated walks, I constructed an empirical PMF using observed run-lengths. For computer-generated walks, the PMF is derived directly from the theoretical properties of the geometric distribution, Geo(0.5). A unique aspect of this analysis is the use of non-uniform bucket sizes, tailored to capture distinct decay patterns observed in the geometric distribution compared to human-generated data.

2.6.1 Data Processing for Human PMF

Run-Length Extraction: Each trajectory was divided into 100-step walks. Run-lengths, defined as sequences of consecutive steps in the same direction, were extracted for all walks, resulting in a dataset of run-lengths.

Non-Bucketed PMF: I calculated the frequency of each run-length directly from the data:

$$P_{\text{human}}(r) = \frac{\text{Count of runs with length } r}{\text{Total runs in } R_{\text{human}}},$$

where R_{human} is the set of all run-lengths extracted from the human-generated dataset.

Non-Uniform Bucket Ranges: By employing varying bucket sizes tailored to the exponential decay of the geometric model, I rewarded the distinguishing characteristics of human-generated patterns more against the computer-generated patterns while constructing the PMF for human-generated walks.

bucket_ranges $= \{(1,1),(2,2),(3,3),(4,5),(6,7),(8,10),(11,15),(16,20),(21,30),(31,50),(51,100)\}$

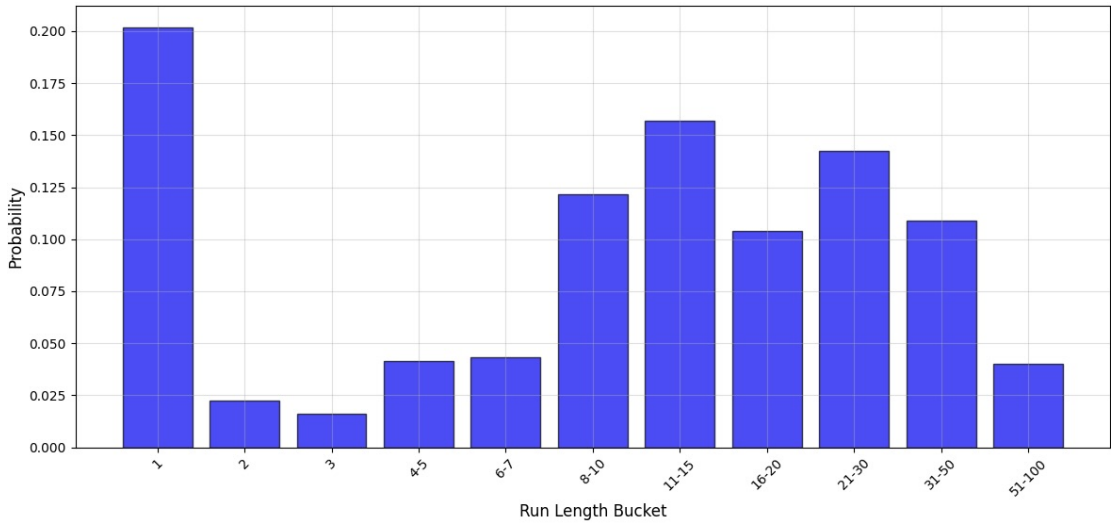

Figure 6: PMF with Varying Buckets for Run Lengths of Human-Generated 100-Step Walks

Bucketed PMF: Using the bucket ranges, I aggregated the probabilities for each bucket:

$$P_{\text{human}}(B_i) = \sum_{r \in B_i} P_{\text{human}}(r),$$

where B_i represents the range of run-lengths in the i-th bucket.

2.6.2 Theoretical PMF for Computer Walks

Geometric Distribution: The run-lengths for computer-generated walks follow the geometric distribution, Geo(0.5), with the probability mass function:

$$P_{\text{computer}}(r) = 0.5^r, \quad r \geq 1.$$

2.6.3 Classification Using Bucketed PMFs

Log-Likelihood Calculation: For a given walk, composed of observed run-lengths $\{r_1, r_2, \ldots, r_k\}$, the log-likelihood under each model is computed as:

$$\log P(W \mid H) = \sum_{i=1}^{k} \log P_{\text{human}}(r_i),$$

$$\log P(W \mid C) = \sum_{i=1}^{k} \log P_{\text{computer}}(r_i).$$

Likelihood Ratio: The log-likelihood ratio quantifies the relative likelihood of a walk being human or computer-generated:

$$\log \frac{P(W \mid H)}{P(W \mid C)} = \log P(W \mid H) - \log P(W \mid C).$$

Decision Rule:

If $\log P(W \mid H) > \log P(W \mid C)$, classify the walk as human. Otherwise, classify it as computer.

2.7 Evaluation of Using Varying Bucket Sizes in the Discrete Model

In the discrete model, I experimented with different bucketing strategies. Although bootstrapping with $n = 100,000$ would have allowed me to construct a PMF with a one-to-one mapping for each run-length up to 100, I opted against this approach for two key reasons:

1. **Potential Bias in the Data:** Using a 1-to-1 PMF could inadvertently incorporate biases present in the dataset.

2. **Emphasizing Distinctive Features:** My primary goal was to highlight and reward features that effectively differentiate between human and computer-generated models.

Key Insight on Longer Runs: One significant distinction between the geometric distribution and human-generated patterns lies in the behavior of longer runs. In the geometric distribution, the probability of observing $R > 9$ converges to zero. In contrast, human-generated patterns show occurrences of runs exceeding 20 or more. To capture this difference, I used bucketed ranges to emphasize and reward higher run-lengths as distinctive features, instead of discretizing each individual value and diluting their impact. This approach significantly enhanced the accuracy of the discrete model.

Handling Shorter Runs: For shorter run-length values, I deliberately avoided bucketing. In the geometric distribution, shorter runs are highly probable, whereas in human patterns, they are nearly as likely as longer runs but occur less frequently overall. Aggregating shorter runs (e.g., combining values 1 to 4 into a single bucket) would have disproportionately increased their weight, bringing their probability closer to that of computer-generated walks. This would have diminished the model's ability to distinguish between the two patterns.

Optimizing Bucket Sizes: By employing varying bucket sizes tailored to the exponential decay of the geometric model, I preserved the distinguishing characteristics of human-generated patterns. This strategy proved to be the most effective approach for this task, as it balanced the representation of both shorter and longer runs while maintaining the distinctive nature of the human-generated data.

3 Impact

References