International Journal of Mathematics, Statistics and Operations Research
Vol. 5; Number 2; 2025, pp. 287-302
https://doi.org/10.47509/IJMSOR.2025.v05i02.07

ARF INDIA
Academic Open Access Publishing
www.arfjournals.com

Predicting Outcomes of Random Phenomena

Jyotirmoy Sarkar[a] and Quyen Tran[b]

[a] *Indiana University Indianapolis, IN, USA*
[b] *DePauw University, Greencastle, IN, USA*

ABSTRACT
In a course in *Inferential Statistics*, students usually learn point estimation, confidence interval estimation, and hypothesis testing. Although they learn quite a bit about the parameters in a model based on a random sample, they still know very little about how to predict a future random outcome. This paper explains the difficulty involved and proposes a way to predict efficiently to minimize the overall cost associated with the error of prediction.

KEYWORDS
Binomial, Multinomial, Discrete, Continuous, Penalty function

1. Introduction

Predicting the next outcome in a sequence—be it a coin flip, a categorical label, or a real-valued observation—is a core problem in statistics and machine learning [1, 4, 8]. In practical applications, the true data-generating process is often unknown, and decisions must be made in real time based on limited information [6]. As a result, many prediction strategies emerge, ranging from random guessing to methods based on running estimates or memory-based heuristics [1].

Our goal in this paper is to classify these strategies under a common analytical framework [2] and compare their performance systematically. Whether a strategy is randomized, fixed from the start, or continuously updated using observed data, we evaluate how well it performs under uncertainty and how it responds to different loss functions that are powers of the absolute difference between the predicted and the observed values.

We begin with the binary (Bernoulli) prediction problem [7] and introduce a family of strategies, including fixed predictors, delayed estimators, and memory-based rules like myopic guessing. We then generalize to the three-outcome case, analyzing how empirical frequency estimation extends to more complex categorical data or a discrete variable. We also study the prediction of real-valued variables that are mixtures of normal distributions or skewed distributions. In each scenario, we utilize both theoretical analysis and simulation results to highlight the accuracy of prediction measured by the expected long-run penalty.

CONTACT Author[a]. Email: jsarkar@iu.edu

Article History
Received : 24 August 2025; Revised : 10 October 2025; Accepted : 23 October 2025; Published : 10 November 2025

To cite this paper
Jyotirmoy Sarkar and Quyen Tran (2025). Predicting Outcomes of Random Phenomena. *International Journal of Mathematics, Statistics and Operations Research.* 5(2), 287-302.

This comparative approach prompts us to ask a central question: Should we fix the prediction, or change it adaptively as we gather data, or additionally randomize to attain the optimal predictive behavior? This paper advocates a structured approach to prediction strategies with practical and theoretical implications for sequential learning systems.

2. The Bernoulli Problem

2.1. *Problem Description*

Consider a biased coin with an unknown probability of heads $p \in (0, 1)$. We are about to observe a sequence of flips $X_1, X_2, \dots$, each $X_i \in \{0, 1\}$, which are independent and identically distributed (IID) as Bernoulli(p) [7]; that is, each X_i takes values 0 and 1, with $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = 0) = q = 1 - p$. Our task is to predict each flip, just before we see the outcome, so that we maximize the number of correct predictions. To predict each future outcome, we can use the results of the previous flips. How shall we proceed?

2.2. *Prediction Strategies*

We analyze the following prediction strategies:

(1) **Always Stick**: Predict the same outcome (Head or Tail) for all trials, ignoring the results of the previous flips.
(2) **Delayed Estimate-Based Fixed**: Choose a $k \geq 1$. Estimate p by $\hat{p}_k$, the ratio of the number of heads from the first k flips. If $\hat{p}_k \geq 1/2$, predict all future outcomes as heads; otherwise, as tails.
(3) **Estimate-Based Adaptive**: Continually update the estimate $\hat{p}_n$ after $n \geq 1$ flips. Predict the next outcome as heads if $\hat{p}_n \geq 1/2$; otherwise, as tails.
(4) **Repeat Adherence**: Predict the next outcome to be the same as the current (the latest) outcome.
(5) **Repeat Avoidance**: Predict the next outcome to be the opposite of the current (the latest) outcome.
(6) **Randomize Always**: Flip another fair coin on the side and predict the next outcome of the biased coin as whatever the fair coin shows.

Strategies 4 and 5 are myopic rules because they are based on the current outcome.

2.3. *Theoretical Analysis*

2.3.1. *Always Stick*

For a fixed p, the probability of a correct prediction when always guessing Heads is p, and when always guessing Tails is $1 - p$. If p were known to be greater than half, always guessing heads would be preferable. Similarly, if p were known to be less than half, always guessing tails would be preferable. The maximum theoretical accuracy is $\max\{p, 1-p\}$. However, if p is unknown and one always guesses heads blindly, then the expected accuracy is p. Similarly, one always guesses tails blindly, then the expected accuracy is $1 - p$. "Always heads" strategy is suboptimal for $p \in (0, .5)$, and "always tails" strategy is suboptimal for $p \in (.5, 1)$.

2.3.2. Delayed Estimate-Based Fixed

Let $k \geq 1$ denote the burn-in period. During this phase, the predictions are arbitrary. When the outcomes $X_1, \ldots, X_k$ are recorded, define the sample proportion of heads $\hat{p}_k = (\sum_{i=1}^{k} X_i)/k$. By the Strong Law of Large Numbers [5], $\hat{p}_k \to p$ almost surely, as $k \to \infty$. After k trials, always predict heads if $\hat{p}_k \geq .5$, and tails otherwise. For large k, the prediction accuracy is near $\max\{p, 1-p\}$, except when $|p - .5|$ is small.

2.3.3. Estimate-Based Adaptive

After the n-th trial, update the sample proportion of heads to $\hat{p}_n = (\sum_{j=1}^{n} X_j)/n$. Then predict trial $n+1$ as heads if $\hat{p}_n \geq 0.5$, and tails otherwise. Again, $\hat{p}_n \to p$ almost surely, as $n \to \infty$ [5]. Hence, the long-run accuracy converges to $\max\{p, 1-p\}$. Short-term oscillations may occur in finite samples.

2.3.4. Repeat Adherence

For each flip $n \geq 2$, predict that the outcome X_n will match the previous observed outcome X_{n-1}; that is, $\hat{X}_n = X_{n-1}$. The prediction is correct when $X_n = X_{n-1}$. Since the outcomes $X_1, X_2, \ldots$ are IID, we have

$$\mathbb{P}(X_n = X_{n-1}) = \mathbb{P}(X_n = 1 = X_{n-1}) + \mathbb{P}(X_n = 0 = X_{n-1}) = p^2 + (1-p)^2.$$

Thus, the long-run accuracy of this Myopic Strategy of repeat adherence is $.5 \leq p^2 + (1-p)^2 = 1 - 2p(1-p) \leq \max\{p, 1-p\}$.

2.3.5. Repeat Avoidance

For each flip $n \geq 2$, predict that the outcome X_n will differ from the previous observed outcome X_{n-1}; that is, $\hat{X}_n = 1 - X_{n-1}$. The prediction is correct when $X_n = 1 - X_{n-1}$. Since the outcomes $X_1, X_2, \ldots$ are IID, we have

$$\mathbb{P}(X_n = 1 - X_{n-1}) = \mathbb{P}(X_n = 1, X_{n-1} = 0) + \mathbb{P}(X_n = 0, X_{n-1} = 1) = 2p(1-p).$$

Thus, the long-run accuracy of this Myopic Strategy of repeat avoidance is $1 - \max\{p, 1-p\} \leq 2p(1-p) \leq 0.5$. This is a rather poor strategy. "What has happened currently must be compensated for by the opposite thing happening next" is not a logically tenable argument since the random process does not have any memory. The outcomes are, in fact, *independent*. Similarly, "What has happened currently, must happen again" is not tenable either. However, "What has happened currently is likely to be the more likely outcome next time." is a judicious choice.

2.3.6. Randomize Always

Flip another fair coin on the side whose outcome is Y_n. Predict the next outcome of the biased coin as whatever the outcome on this fair coin is; that is, $\hat{X}_n = Y_n$. The prediction is correct when $X_n = Y_n$. Since the outcomes X_n, Y_n are independent, we have

$$\mathbb{P}(X_n = Y_n) = \mathbb{P}(X_n = 1 = Y_n) + \mathbb{P}(X_n = 0 = Y_n) = (1/2)p + (1/2)(1-p) = 1/2.$$

Thus, the long-run accuracy of the Randomize Always strategy is 0.5, which is typically below $\max\{p, 1-p\}$, with equality if and only if $p = .5$.

2.4. *Simulation Study*

We conducted a simulation study assuming $p = .70$ and generating $n = 2000$ Bernoulli trials. Here are the sample paths of the running proportions of correct prediction θ_n under each strategy. Let $\theta = \theta_{2000}$.

- Always Stick strategy yields a θ near $\max\{p, 1-p\}$, if p were known. But if p is unknown and we always predict heads, then θ is near p. Finally, if p is unknown and we always predict tails, then θ is near $1-p$.
- Delayed Estimate-Based Fixed Strategy with $k = 100$ yields a θ near the optimal accuracy $\max\{p, 1-p\} = .70$.
- Estimate-Based Adaptive strategy also yields a θ near the optimal accuracy $\max\{p, 1-p\} = .70$.
- Randomize Always strategy yields a θ near .50.

Note that the θ_n curves are symmetrically away from .50 for (1) "Always stick to heads" and "Always stick to tails" strategies, and (2) "Repeat adherence" and "Repeat avoidance" strategies.

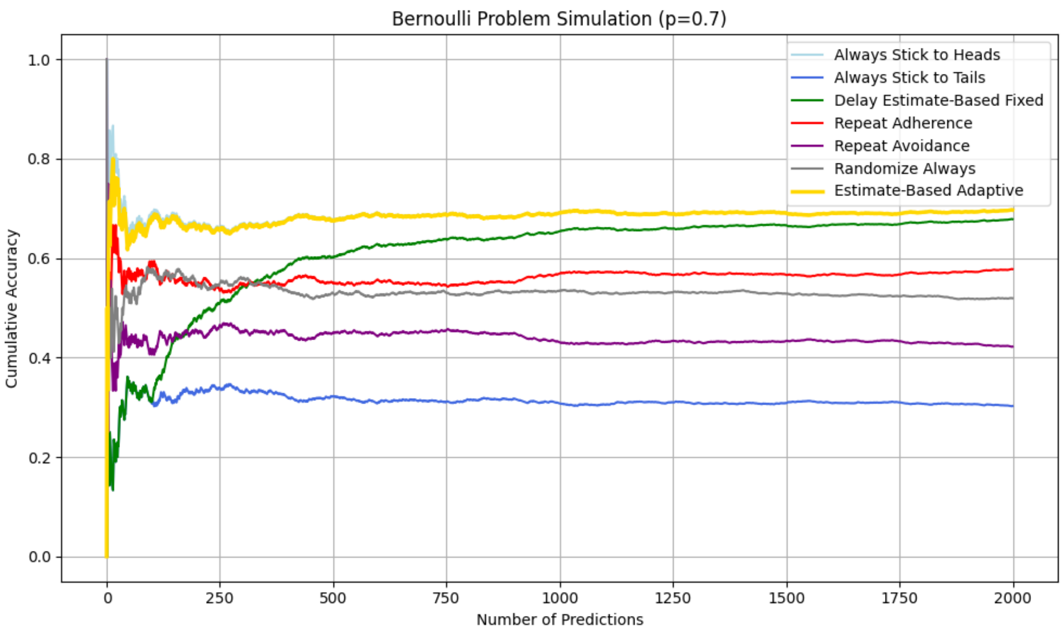

Figure 1. Comparing accuracy of various strategies based on running proportions

Next, we replicated the process of generating the 2000 Bernoulli(p) variables for $M = 1000$ times and recorded the values of $\theta = \theta_{2000}$ in each replicate. Figure 2 displays these values as histograms, and Table 1 documents their means and SD's. The Estimate-Based Adaptive strategy performs the best both in terms of high accuracy given by high mean(θ) and high precision given by low SD(θ).

Strategy	Mean	Standard Deviation
Always Stick to Heads	0.6998	0.0098
Always Stick to Tails	0.3002	0.0098
Delay Estimate-Based Fixed	0.6799	0.0100
Repeat Adherence	0.5791	0.0127
Repeat Avoidance	0.4209	0.0127
Randomize Always	0.4999	0.0112
Estimate-Based Adaptive	0.6992	0.0099

Table 1. Mean and standard deviation of final cumulative accuracies after 2000 predictions, based on 1000 replicates, when $p = .7$

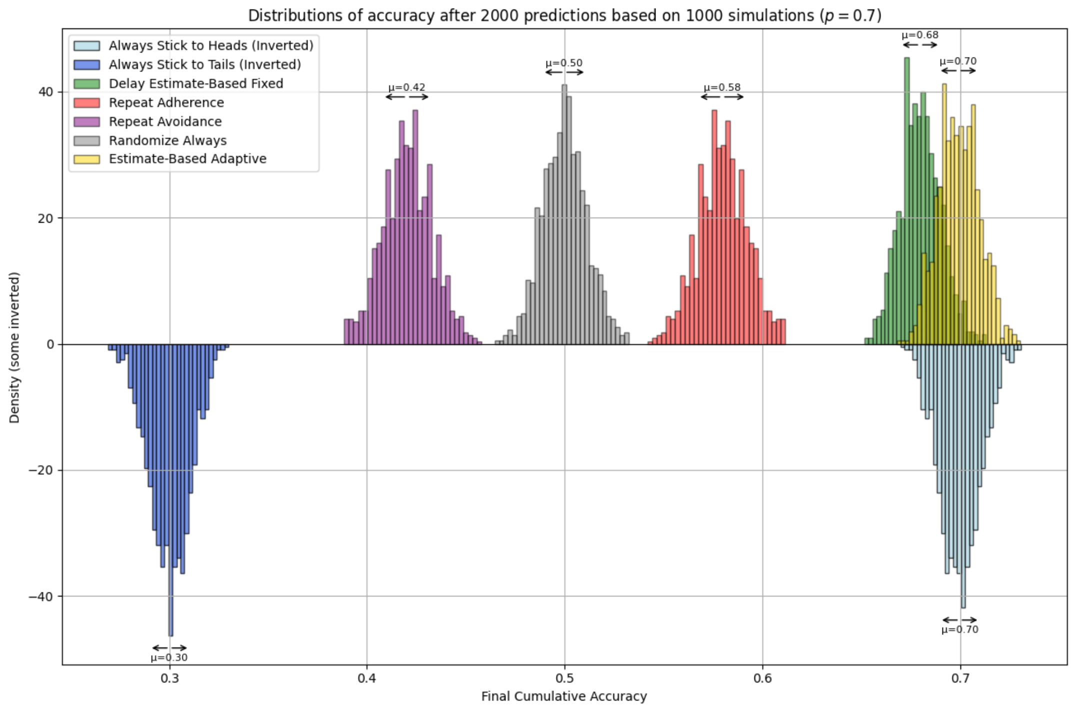

Figure 2. Histograms of accuracy of various prediction strategies based on 1000 replicates, when $p = .7$

3. Predicting a Ternary Outcome

3.1. *Problem Description*

Consider a random variable X with **three** possible outcomes $\{A, B, C\}$, where $p_A = \mathbb{P}(X = A), p_B = \mathbb{P}(X = B), p_C = \mathbb{P}(X = C)$ satisfy $p_A + p_B + p_C = 1$. Suppose that $X_1, X_2, X_3, \ldots$ form an IID sample of these three outcomes. Just before each outcome is revealed, we must predict it based on the data we have seen so far. How should we predict to maximize the proportion of correct predictions?

Note that a ternary outcome is a special case of a multinomial distribution [3].

3.2. *Prediction Strategies*

We analyze the same six strategies introduced for the Bernoulli problem, adapted to three outcomes.

(1) **Always Stick**: Always predict the same outcome (A, B, or C) for every trial.
(2) **Delayed Estimate-Based Fixed**: Use the first k trials to estimate the probabilities, then always predict the most frequent outcome observed so far.
(3) **Estimate-Based Adaptive**: Update the frequency estimates after each trial and predict the most frequent outcome at that time.
(4) **Repeat Adherence**: Predict that the next outcome will be the same as the most recent one.
(5) **Repeat Avoidance**: Predict that the next outcome will be *different* from the most recent one, and is equally likely to be one of the remaining two.
(6) **Randomize Always**: Predict an outcome at random, uniformly from $\{A, B, C\}$.

3.3. *Theoretical Analysis*

3.3.1. Always Stick

If one always predicts outcome A, the long-run accuracy is p_A; likewise, for B and C. Hence, the best one can achieve with full knowledge of (p_A, p_B, p_C) is $\max\{p_A, p_B, p_C\}$. Without this knowledge, blindly sticking to one choice is suboptimal.

3.3.2. Delayed Estimate-Based Fixed

After observing the first k trials, we estimate

$$\hat{p}_A = \frac{\#A}{k}, \quad \hat{p}_B = \frac{\#B}{k}, \quad \hat{p}_C = \frac{\#C}{k}.$$

We then lock all our predictions onto the most frequent outcome. By the Strong Law of Large Numbers [5], $\hat{p}_i \to p_i$ almost surely, so for large k, the strategy selects the correct outcome with the highest p_i, achieving asymptotic accuracy $\max\{p_A, p_B, p_C\}$.

3.3.3. Estimate-Based Adaptive (Continual Updating)

At each trial n, update the estimates:

$$\hat{p}_A^{(n)} = \frac{\#A}{n}, \quad \hat{p}_B^{(n)} = \frac{\#B}{n}, \quad \hat{p}_C^{(n)} = \frac{\#C}{n},$$

and predict the next outcome to be the most frequent outcome so far. As $n \to \infty$, the proportion of correct prediction approaches the optimal accuracy $\max\{p_A, p_B, p_C\}$. Short-term fluctuations may enforce some suboptimal predictions early on.

3.3.4. Repeat-Adherence Strategy

Predict the first outcome arbitrarily. Then, always predict the next event to be the same as the current outcome:

$$\widehat{X}_i = X_{i-1} \quad (\text{for } i \geq 2).$$

Since the trials are IID,

$$\mathbb{P}(X_i = X_{i-1}) = \mathbb{P}(X_i = A = X_{i-1}) + \mathbb{P}(X_i = B = X_{i-1}) + \mathbb{P}(X_i = C = X_{i-1}) = p_A^2 + p_B^2 + p_C^2.$$

In general,

$$p_A^2 + p_B^2 + p_C^2 \leq \max\{p_A, p_B, p_C\}$$

with equality when the three probabilities are 1/3 each, or close to equality when one outcome has a probability very close to 1. Hence, the Myopic Strategy is inferior to the strategies that converge to $\max\{p_A, p_B, p_C\}$.

3.3.5. Repeat-Avoidance Strategy

Instead of repeating the previous outcome, this strategy deliberately avoids it by randomly choosing one of the other two outcomes, say, by flipping a fair coin. If the latest event is A, the next prediction is selected randomly between B and C. So, the probability of correctly predicting the next outcome is:

$$\frac{1}{2}p_B + \frac{1}{2}p_C = \frac{1 - p_A}{2}.$$

Similarly, if the current outcome is B, the probability of a correct prediction is $(1 - p_B)/2$. If the current outcome is C, the probability of a correct prediction is $(1 - p_C)/2$. Since the probability of the previous outcome being A, B, or C is simply p_A, p_B, or p_C, the overall probability of a correct prediction is:

$$P(\text{correct}) = p_A \frac{1 - p_A}{2} + p_B \frac{1 - p_B}{2} + p_C \frac{1 - p_C}{2} = \frac{1}{2}\left[1 - (p_A^2 + p_B^2 + p_C^2)\right].$$

3.3.6. Randomize All The Time

At each trial, rather than committing to a fixed prediction, this strategy selects an outcome completely at random with equal probability for all three outcomes. If predictions are made uniformly at random, the expected accuracy is:

$$\frac{1}{3}(p_A + p_B + p_C) = \frac{1}{3}.$$

3.4. ***Simulation Study***

We conducted a simulation study assuming $(p_A, p_B, p_C) = (0.5, 0.3, 0.2)$ and generating $n = 2000$ ternary trials. Figure 3 shows the sample paths of the running proportions of correct prediction θ_n under each of the following strategies.

- **Always Stick**:
  - Always guess A: Accuracy = 0.50.
  - Always guess B: Accuracy = 0.30.
  - Always guess C: Accuracy = 0.20.
  - If you *knew* p_A was the largest, you should always guess A and achieve a 50% accuracy.
- **Delayed Estimate-Based Fixed**:
  - After $k = 100$ burn-in trials, we will have $\hat{p}_A > \hat{p}_B, \hat{p}_C$ with a very high probability .9866[1]
  - Lock in to guessing A for all future outcomes. Then the long-run accuracy converges to 0.50.
- **Estimate-Based Adaptive**:
  - As n grows, almost surely, A will occur most often. Thereafter, you will guess A. So, the accuracy converges to $p_A = 0.50$.
- **Repeat Adherence**:

$$\mathbb{P}(X_i = X_{i-1}) = 0.5^2 + 0.3^2 + 0.2^2 = 0.25 + 0.09 + 0.04 = 0.38.$$

  This is noticeably lower than $p_A = 0.50$.
- **Repeat Avoidance**: The expected probability of a correct prediction is

$$P(\text{correct}) = \frac{1}{2}(1 - 0.38) = \frac{1}{2} \cdot 0.62 = 0.31,$$

  which is further lower than $p_A = 0.50$.
- **Randomize All The Time:** Each outcome is predicted with equal probability. So, the expected accuracy is:

$$\frac{1}{3}(0.5 + 0.3 + 0.2) = \frac{1}{3} \approx 0.333.$$

  This is worse than the strategies that adapt to the most probable outcome.

At the end of each simulation, we compute $\theta = \theta_{2000}$. Then we repeat the simulation 1000 times and save all θ values. These are shown in the form of histograms in Figure 4, and their mean and SD's are tabulated in Table 2.

4. Predicting Outcomes from a Finite Set

4.1. *Problem Description*

Consider a basket containing n balls with distinct numbers, $x_1 < x_2 < \cdots < x_n$, written on them. The balls are shuffled, and one is drawn randomly *with replacement.* The number on the chosen ball is not revealed until after you make a prediction. Your task is to predict the number on the ball as closely as possible.

[1]Letting N_A, N_B, N_C denote the frequencies of A, B, C among the first 100 outcomes, either $N_A > 50$, or $34 \leq N_A \leq 50$ and $N_B, N_C < N_A$. This happens with probability $P\{N_A > 50\} + \sum_{m=34}^{50} P\{N_A = m\} \sum_{i=101-2m}^{m-1} P\{N_B = i | N_A = m\}$, and can be computed using a software since N_A follows a binomial$(100, p_A)$ distribution and $N_B | N_A = m$ follows a binomial$(m, p_B/(1 - p_A))$ distribution.

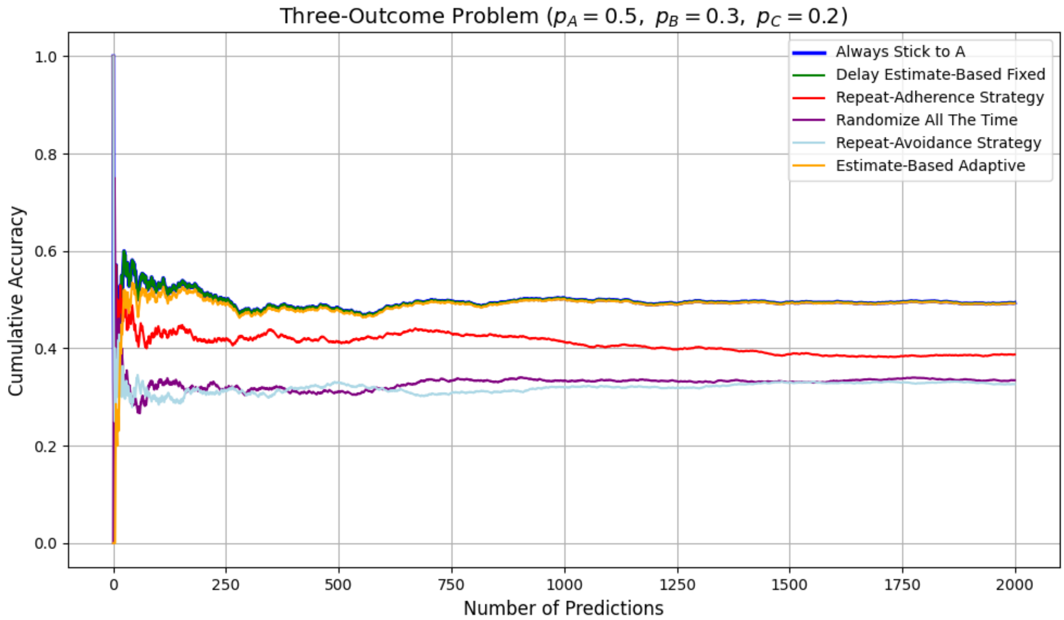

Figure 3. Comparing the accuracy of prediction strategies in the three-outcome prediction problem, based on 2000 trials when ($p_A = .5, p_B = .3, p_C = .2$)

Strategy	**Mean**	**Standard Deviation**
Always Stick to A	0.499825	0.011240
Delay Estimate-Based Fixed	0.498886	0.017787
Estimate-Based Adaptive	0.498807	0.011351
Repeat-Adherence Strategy	0.380178	0.011577
Randomize All The Time	0.333435	0.010359
Repeat-Avoidance Strategy	0.309652	0.010649

Table 2. Mean and standard deviation of the final cumulative accuracy of each strategy, based on 1000 replicates when ($p_A = .5, p_B = .3, p_C = .2$)

4.2. *Penalty Rules*

In Sections 2 and 3, the penalty was 0 for correct prediction and 1 for incorrect prediction. That is, all incorrect predictions were equally costly. Here, we allow different penalties for different incorrect predictions. The penalties for incorrect predictions are calculated according to the following rules:

(1) **Penalty Rule 1:** The penalty is the absolute difference between the prediction and the outcome of the draw, or

$$L(y, \hat{y}) = |y - \hat{y}|.$$

(2) **Penalty Rule 2:** The penalty is the squared difference between the prediction

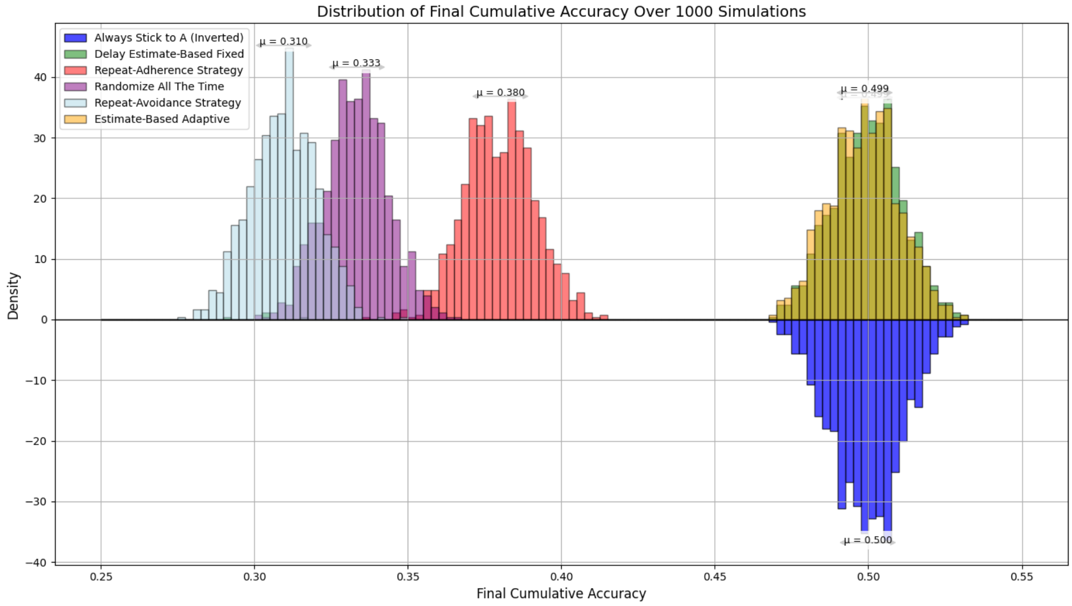

Figure 4. Histograms of accuracy of various ternary prediction strategies based on 1000 replicates when ($p_A = .5, p_B = .3, p_C = .2$)

and the outcome of the draw, or

$$L(y, \hat{y}) = (y - \hat{y})^2.$$

(3) **Penalty Rule 3:** The penalty is the positive square root of the absolute difference between the prediction and the outcome of the draw, or

$$L(y, \hat{y}) = \sqrt{|y - \hat{y}|}.$$

(4) **Penalty Rule 4:** The penalty is the absolute difference between the prediction and the outcome of the draw, raised to the power 3/2, or

$$L(y, \hat{y}) = |y - \hat{y}|^{3/2}.$$

Thus, the penalty rules defined by various powers of the absolute difference between the prediction and the outcome, such as powers $1, 2, 1/2$, and $3/2$.

4.3. *Predictive Strategies*

We analyze the following predictive strategies to minimize the expected penalty under each penalty rule:

(1) **Always Predict the Mean:** Predict the mean value of the numbers in the basket:

$$\hat{y} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

This strategy minimizes the expected penalty for the squared difference.

(2) **Always Predict the Median:** Predict the median value of the numbers in the basket. For an odd n, the median is the middlemost value; for an even n, it is the simple average of the two middlemost values:

$$\hat{y} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd,} \\ \frac{x_{(n/2)}+x_{(n/2)+1}}{2} & \text{if } n \text{ is even.} \end{cases}$$

This strategy minimizes the expected penalty for the absolute difference.

(3) **Flip between Mean and Median:** Predict either the mean or the median by flipping a fair coin.

(4) **Random Choice:** Randomly select one of the n numbers as the prediction. This strategy is not particularly good, but it provides a baseline for comparison.

(5) **Minimizer Power 1/2:** Predict outcomes as $\nu_{1/2} = \arg\min_t \sum |x_i - t|^{1/2}$.

(6) **Minimizer Power 3/2:** Predict outcomes as $\nu_{3/2} = \arg\min_t \sum |x_i - t|^{3/2}$.

The reader should check that $\nu_1 = \arg\min_t \sum |x_i - t|$ is the median and $\nu_2 = \arg\min_t \sum |x_i - t|^2$ is the mean.

4.4. *Theoretical Analysis*

The performance of each strategy depends on the penalty rule:

- For Penalty Rule 1 (absolute difference), the median minimizes the expected penalty.
- For Penalty Rule 2 (squared difference), the mean minimizes the expected penalty.
- For Penalty Rule 3 (square root of the absolute difference), both the mean and median are reasonable choices, though their performance may differ slightly. The best predictor is $\nu_{1/2} = \arg\min_t \sum |x_i - t|^{1/2}$.
- For Penalty Rule 4 (absolute difference raised to 3/2), strategies close to the mean tend to perform slightly better than those close to the median. The best predictor is $\nu_{3/2} = \arg\min_t \sum |x_i - t|^{3/2}$.

4.5. *Simulation Study*

Suppose that a basket contains five balls with values 10, 18, 21, 30, 36. Then the mean is 23, the median is 21, the minimizer power 1/2 is $\nu_{1/2} = 21.00$, and the minimizer power 3/2 is $\nu_{3/2} = 22.09$. Note that the random prediction is always the worst under any penalty rule. Moreover, additional randomness (by flipping a coin) to make a prediction increases the overall penalty.

Predictive Strategy	Rule 1	Rule 2	Rule 3	Rule 4
Mean (23.00)	8.202	86.278	2.741	26.073
Median (21.00)	7.834	89.918	2.447	26.165
Flip Mean/Median	8.076	89.610	2.607	26.425
Random Choice	10.337	171.541	2.749	41.173
$\nu_{1/2}$ (21.00)	7.834	89.918	2.447	26.165
$\nu_{3/2}$ (22.09)	8.034	86.944	2.663	25.934

Table 3. Average penalty of various prediction strategies for 1000 trials under each penalty rule when the basket values are {10, 18, 21, 30, 36}.

5. Predicting Data from a Continuous Bimodal Distribution

5.1. *Problem Description*

Suppose that we must predict a random variable X representing the height of the next person, drawn from an unknown distribution. Empirical observations suggest the variable height has a continuous bimodal distribution, indicating two dominant subgroups (perhaps men and women) within the data. Imagine that at a graduation ceremony the graduates are being called in alphabetical order of their names. Before you hear the name (so that you do not know the person's gender), you must predict the height of the graduate. Your goal is to predict the height of the next person as accurately as possible.

5.2. *Simulation Setup*

To investigate the behavior of height prediction strategies, we conduct the following simulation study:

(1) Sample the height data of size 1000 from a bimodal distribution shown by the red density curve in Figure 5. The bimodal distribution consists of 50% values from $N(175, SD = 10)$ and 50% values from $N(155, SD = 6)$ distributions. The blue histogram in Figure 5 shows a typical sample of size 1000.
(2) Make a prediction just before each data point is observed.
(3) Sequentially reveal the heights (but not the subgroup), one at a time. Revealed data may be used in predicting the next value.
(4) Compute the average penalty under each penalty function.

5.3. *Penalty Functions*

The penalties are computed using one of the four loss functions given in Section 4: Rule 1=absolute error (AE), Rule 2=square error (SE), Rule 3=square-root error (SRE) and Rule 4=cube of square-root error (CSRE).

5.4. *Predictive Strategies*

We explore various strategies to predict the next height based on statistical measures of the observed values.

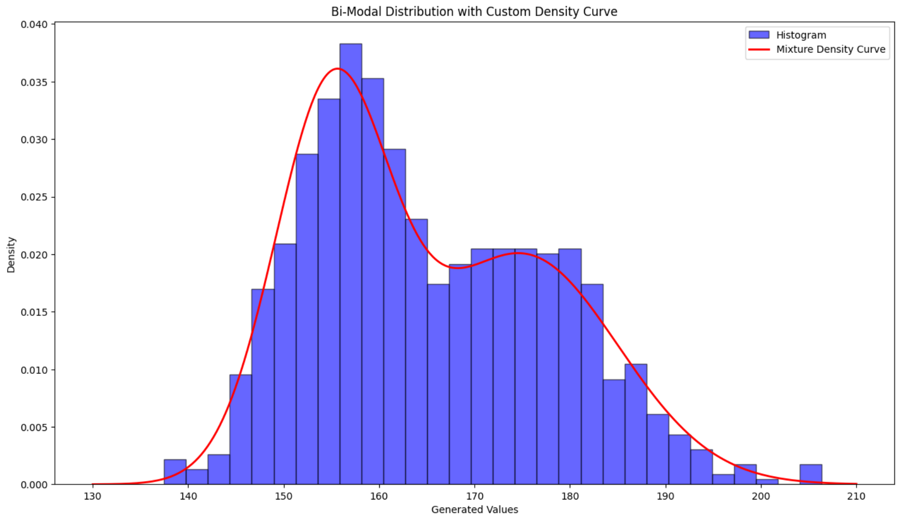

Figure 5. A blue histogram of 1000 values generated from a bimodal distribution of heights (shown in red)

(1) **Mean Prediction**: Predict the next observation using the sample mean of the observed heights.
(2) **Median Prediction**: Predict the next observation using the median, which is more robust to outliers.
(3) **Random Normal Prediction**: Predict the next observation using a randomly generated value from a normal distribution with mean and standard deviation estimated by the sample mean and the sample standard deviation.
(4) **Random Middle 50% Prediction**: Predict the next observation using a randomly chosen value within the middle 50% of the observed data.
(5) **Trimmed Mean Prediction**: Predict the next observation using the mean of the middle 50% of the sample.

5.5. *Result*

The mean penalty (out of 1000 trials) under each prediction strategy and penalty rule is tabulated in Table 4. The minimum penalty under each penalty rule is underlined. Again, note that strategies involving additional randomization are never the best.

6. Predicting Data from a Continuous Right-Skewed Distribution

6.1. *Problem Description*

Charitable donation data often exhibits non-normal characteristics, such as positive skewness and heavy tails. Understanding how different strategies perform in predicting data from a skewed distribution is crucial for accurately modeling the donation amounts and for optimizing fundraising efforts.

Predictive Strategy	Rule 1 MAE	Rule 2 MSE	Rule 3 MSRE	Rule 4 MCSRE
Mean Prediction	10.88	166.83	3.10	41.35
Median Prediction	10.75	174.81	3.04	41.84
Random Normal Prediction	14.58	328.15	3.52	66.74
Random Middle 50% Prediction	11.58	200.71	3.16	46.65
Trimmed Mean Prediction	10.76	169.66	3.06	41.34

Table 4. Average penalty of height prediction strategies under different penalty functions, based on 1000 trials

6.2. *Simulation Setup*

To investigate the prediction strategies for charitable donation amounts, we conduct the following simulation study:

(1) We generate 1000 donation amounts from a right-skewed and heavy-tailed distribution exhibiting two donor subgroups: Subgroup 1 (regular donors) constitutes 80% of all donors. They donate an amount modeled by a gamma distribution with mean 100 and standard deviation **31.19** (shape=10, scale=10). Subgroup 2 (generous donors) constitutes the remaining 20% of all donors. Their donation amounts follow a gamma distribution with mean 400 and standard deviation **146.27** (shape=8, scale=50). In Figure 6, the density is shown in red, and the blue histogram shows a typical sample of size 1000.
(2) Sequentially reveal the donation amounts given by donors 1 through 1000, without revealing whether the donor belongs to the regular or the generous subgroup.
(3) Predict the next donation amount based on all previous donation amounts.
(4) Compute the average penalty under each loss function.

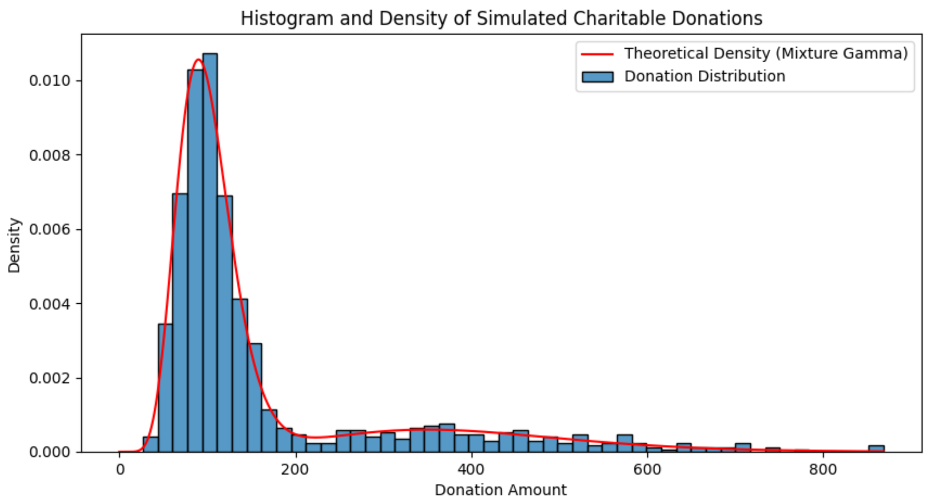

Figure 6. A right-skewed, heavy-tailed distribution for the donation amount

6.3. *Penalty Functions*

Consistent with the previous sections, the penalties are computed using one of the four loss functions given in Section 4: Rule 1= absolute error (AE), Rule 2=square error (SE), Rule 3=square-root error (SRE) and Rule 4=cube of square-root error (CSRE).

6.4. *Predictive Strategies*

We study the same five predictive strategies as in Section 5.

6.5. *Result*

The mean penalty (out of 1000 predictions) under each prediction strategy and penalty rule is tabulated below. Again, additional randomization increases the mean penalty.

Prediction Strategies	**Rule 1 MAE**	**Rule 2 MSE**	**Rule 3 MSRE**	**Rule 4 MCSRE**
Median	83.11	24616.51	7.19	1318.98
Random Middle 50%	89.09	24022.13	7.79	1344.13
Mean	104.55	21603.44	9.37	1386.68
Random Guess	126.08	41160.71	9.24	2146.50
Trimmed Mean	83.32	24229.27	7.26	1307.95

Table 5. Comparison of donation prediction strategies using different penalty functions, based on 1000 trials

7. Conclusion

Having compared different prediction strategies under various penalty rules, our central finding is clear: The prediction must be a function of the penalty rule and the observed data. It must not involve any additional randomness, either inherent in the data or extraneous. Such prediction strategies consistently outperform strategies that involve full or partial randomization across all settings. Of course, the non-random strategy depends on the observed data and critically on the penalty rule.

For categorical outcomes, such as binary or ternary variables, strategies that predict the next outcome based on the most frequent observed category achieve near-optimal accuracy in the long run, closely matching the penalty when the relative frequencies of the categories are known. Thus, accumulated data overcomes the shortcoming of a lack of knowledge! For real-valued outcomes, especially under skewed or heavy-tailed distributions, predictions based on robust statistics—such as the median or the trimmed mean—perform better than those based on the mean, particularly when the loss function is sensitive to large deviations.

We have skipped the more advanced topic of prediction based on regression models. Interested readers may see [4].

In summary, data-dependent predictions are generally more reliable than random guesses. However, the choice of the appropriate function of the observed data depends on the data distribution and the penalty rule.

Acknowledgments

The authors wish to thank Dr. Mamunur Rashid for facilitating this research.

References

[1] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* Wiley-Interscience, 2nd edition, 2006.
[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *Model Assessment and Selection*, pages 219–259. Springer, New York, 2009.
[3] Robert V. Hogg, Joseph W. McKean, and Allen T. Craig. *Introduction to Mathematical Statistics.* Pearson, 8th edition, 2019.
[4] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R.* Springer, New York, 2013.
[5] Michel Ledoux and Michel Talagrand. *The Strong Law of Large Numbers*, pages 178–195. Springer Berlin Heidelberg, Berlin, Heidelberg, 1991.
[6] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective.* MIT Press, 2012.
[7] Vijay K. Rohatgi and A. K. Md. Ehsanes Saleh. *An Introduction to Probability and Statistics.* Wiley Series in Probability and Statistics. John Wiley & Sons, 2015.
[8] Hristos Tyralis and Georgia Papacharalampous. A review of predictive uncertainty estimation with machine learning. *Artificial Intelligence Review*, 57(4):94, 2024.