Research article

ISSN: 2980-1982
2025, Vol. 3, No. 1, 11–34
DOI: 10.66147/lapm.20253129

Letters on
Applied and Pure Mathematics

Bernoulli distribution: Important properties and comparison of estimation methods

Sal Ly [1,*]

1. Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam

*Corresponding author email: lysal@tdtu.edu.vn

Abstract

The Bernoulli distribution (BD) is a fundamental model for binary data and a key building block in constructing zero-inflated distributions. This study explores the core properties of BD and compares three widely used estimation methods: Method of Moments (MM), Maximum Likelihood Estimation (MLE), and Bayesian Method (BM). Through systematic simulation studies across different sample sizes and parameter values, we evaluate the bias and mean squared error (MSE) of each estimator. Special attention is given to Bayesian estimation under various Beta priors, including uniform, Jeffreys, symmetric, and informative moment-matched priors. Results show that MLE is unbiased and consistent, while Bayes estimators achieve lower small-sample MSE through shrinkage toward the prior mean, where informative priors aligned with the true parameter yield the best finite-sample performance. These findings provide practical guidance for selecting estimation methods in applications where sample size and prior information vary.

Keywords: Bernoulli distribution, maximum likelihood estimation, Bayes method, point estimation, regression models
MSC (2020): 60A01, 60A05, 62E15, 62E20
Article history: Received 01 Sep 2025; Accepted 15 Nov 2025; Online 11 Dec 2025

1 Introduction

Binary data is among the most prevalent and essential types of data in scientific research, characterized by its simplicity and versatility. It frequently arises in practical applications, where events or outcomes can be encoded as 0 or 1 to represent two distinct states. The Bernoulli distribution (BD) is the most natural and widely accepted model for binary data, offering a straightforward yet powerful framework for analysis. Despite its foundational role, the BD's simplicity in formulation has led to relatively limited attention in the literature compared to more complex distributions.

However, its significance extends beyond basic modelling as it serves as a key component in constructing more advanced models, such as zero-inflated distributions, which are increasingly used in modern statistical applications. The purpose of this article is to provide a concise and unified presentation of the Bernoulli distribution, its classical estimators, and their empirical behavior. Although these methods are well known, our aim is to present them in an accessible way for educational and applied statistical use. Further, by comparing the performance of commonly used techniques, we seek to underscore the practical relevance of BD and encourage further exploration of its applications in statistical modelling.

In our view, one of the most prominent roles of the BD in scientific research is its use as a foundational component in constructing new distributions, particularly those designed to model zero-inflated (ZI) count data. ZI distributions are especially well-suited for datasets characterized by an excess of zero counts, which are common in fields such as healthcare, ecology, and insurance. The construction of a ZI distribution typically involves a two-part process: the first component models the occurrence of structural zeros using the BD, while the second component employs a discrete distribution to model the count data. This dual-process framework allows for greater flexibility and accuracy in capturing the underlying data structure.

Several well-known ZI distributions have been developed using this approach. These include the Zero-Inflated Poisson (ZIP) distribution [24], as well as its numerous extensions and variants, such as the Zero-Inflated Conway–Maxwell–Poisson [4], Zero-Inflated Generalized Poisson [11], Zero-Inflated Hyper-Poisson [22], Zero-Inflated Poisson–Ailamujia [1], and Zero-Inflated Poisson–Akash [27] distributions, among others. These models highlight the central role of the BD in advancing statistical methodologies for complex data types.

The aforementioned ZI distributions are typically constructed using a Bernoulli process to model the excess zeros and a Poisson or Poisson-related distribution to model the count data. As a result, many of these models include "Poisson" in their names, reflecting their foundational structure. However, the Poisson distribution has limitations, particularly in handling overdispersion, where the variance exceeds the mean. To address this, the Negative Binomial (NB) distribution is often employed as a more flexible alternative, offering a natural extension of the Poisson model. Consequently, a growing body of research has focused on developing ZI distributions based on the NB distribution.

Notable examples include the Zero-Inflated Negative Binomial (ZINB) and its variants, such as the ZINB Crack distribution [20], ZINB Generalized Exponential distribution [2], and ZINB Sushila distribution [21]. Beyond the Poisson and NB families, other innovative ZI models have emerged, including the Zero-Inflated Bell [15], Zero-Inflated Waring [19], and Zero-Inflated New Discrete Lindley distributions [25], among others. Once again, these developments underscore the central role of the Bernoulli distribution in the construction of modern ZI models and highlight its continued relevance in statistical modelling.

In today's data-driven era, the development of new statistical distributions has become increasingly important. As datasets grow in both size and complexity, there is a growing need for flexible and robust models capable of capturing diverse data characteristics. Among these, zero-inflated distributions have gained significant attention due to their ability to model excess zeros in count data. As we can see that the construction of ZI models fundamentally relies on the BD to model structural zeros; BD plays a crucial and indispensable role in modern statistical modelling.

Despite its foundational importance, the Bernoulli distribution has received relatively limited attention in the literature, particularly regarding its theoretical properties and parameter estimation techniques. This gap motivates the present study, which aims to provide a comprehensive examination of the BD and its estimation methods, thereby contributing to a more complete

understanding of this essential distribution.

Specifically, this work focuses on two main objectives: (1) to explore and present the key properties of the Bernoulli distribution in detail, and (2) to compare the performance of three widely used point estimation methods: Method of Moments (MM), Maximum Likelihood Estimation (MLE), and Bayesian Method (BM) in estimating its parameter. These methods are among the most commonly applied in statistical inference, and understanding their behavior in the context of BD is valuable for both theoretical and applied research. To enhance accessibility for readers without a strong statistical background, we also provide step-by-step algorithms for each estimation method.

The structure of this paper is as follows. Section 2 introduces the fundamental definitions and formulas of the Bernoulli distribution. Section 3 reviews the general principles of common point estimation methods. Section 4 applies these estimation techniques specifically to the BD. Section 5 discusses the key properties of the BD, including those revealed through estimation. Section 6 presents simulation studies to evaluate the performance of the estimation methods. Section 7 applies the BD to real-world datasets to demonstrate its practical utility. Section 8 concludes the paper with a summary of findings and suggestions for future research.

2 The Bernoulli Distribution

A random variable X is said to follow a Bernoulli distribution if it takes only two possible values: 1 (typically representing success) and 0 (representing failure), with probabilities p and $1-p$, respectively. The probability mass function (PMF) of the Bernoulli distribution is defined as:

$$f(x;p) = \begin{cases} p, & \text{if } x = 1, \\ 1-p, & \text{if } x = 0. \end{cases}$$

This PMF can be compactly expressed using the following formula, which is commonly used in statistical literature:

$$f(x;p) = p^x(1-p)^{1-x}, \quad \text{with } x \in \{0,1\}. \tag{2.1}$$

Alternatively, it can be written in a linear form:

$$f(x;p) = px + (1-p)(1-x), \quad \text{with } x \in \{0,1\}.$$

Among these, the exponential form (2.1) is the most widely used due to its simplicity and elegance in derivations.

Next, the cumulative distribution function (CDF) of the Bernoulli distribution is given by:

$$F(x;p) = \begin{cases} 0, & x < 0, \\ 1-p, & 0 \le x < 1, \\ 1, & x \ge 1. \end{cases}$$

Unlike the PMF, which can be expressed in multiple forms, the CDF of the Bernoulli distribution typically has a single standard representation.

Figure 1 illustrates the PMF of the Bernoulli distribution for three different values of p. Figure 1a: When $p = 0.7$, the probability of success ($x = 1$) is higher than that of failure ($x = 0$). This is reflected in the taller bar at $x = 1$. Figure 1b: When $p = 0.5$, the probabilities of success and failure are equal, resulting in bars of equal height. Figure 1c: When $p = 0.3$, the probability of failure (x=0) exceeds that of success (x=1), as shown by the taller bar at x=0.

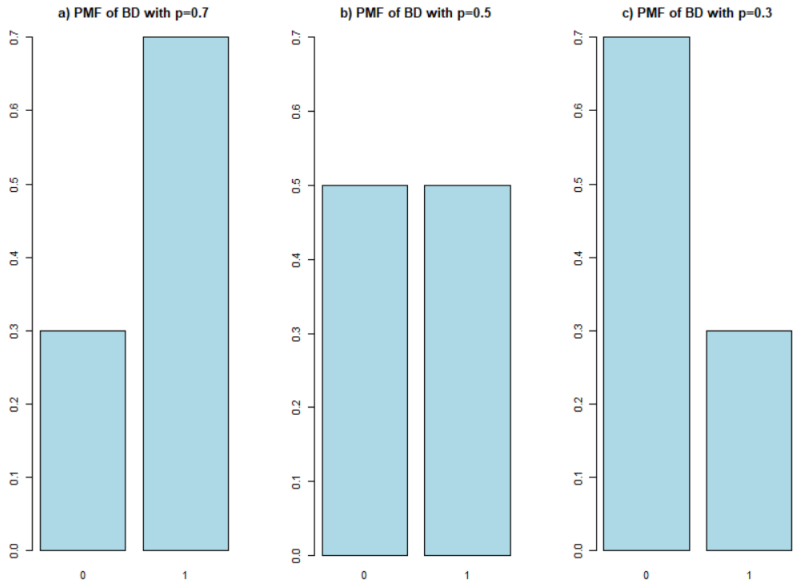


Figure 1: The PMF of the Bernoulli distribution.

3 Ubiquitous Estimation Methods

In this section, we revisit three of the most widely used methods for point estimation in statistics: the Method of Moments (MM), Maximum Likelihood Estimation (MLE), and the Bayesian Method (BM). These methods have been extensively studied and applied across various statistical contexts [6–10, 12]. Respectively, we use the following notations of estimators throughout the paper: $\widehat{p}_{\mathrm{MM}}, \widehat{p}_{\mathrm{MLE}}, \widehat{p}_{\mathrm{B}}$.

3.1 Method of Moments (MM)

The Method of Moments (MM) is one of the earliest techniques for parameter estimation, first introduced by Karl Pearson in 1902 [17]. It remains a foundational approach in both theoretical and applied statistics due to its simplicity and minimal assumptions. The core idea of MM is to equate the theoretical (population) moments of a distribution to the corresponding empirical (sample) moments, and then solve the resulting system of equations to estimate the unknown parameters [6].

Let $X_1, X_2, ..., X_n$ be a random variables from population X, with the probability density function (PDF) or probability mass function (PMF) $f(x; \alpha_1, \alpha_2, ..., \alpha_n)$, where $\alpha_1, \alpha_2, ..., \alpha_n$ are the parameters to be estimated.

Then, for a continuous random variable, the k-th population moment is defined as:

$$M_k = E(X^k) = \int_{-\infty}^{\infty} x^k f(x; \alpha_1, \alpha_2, ..., \alpha_n) dx. \tag{3.1}$$

For a discrete random variable, the k-th population moment is:

$$M_k = E(X^k) = \sum_{i=1}^{n} x^k P(X = x). \tag{3.2}$$

The corresponding sample moment of order k is:

$$\widehat{M_k} = \frac{1}{n} \sum_{i=1}^{n} x_i^k. \tag{3.3}$$

To apply the MM, we follow these steps:

- **Step 1**: Compute the population moments $M_1, M_2, \ldots, M_k$ as functions of the unknown parameters.
- **Step 2**: Compute the sample moments $\widehat{M}_1, \widehat{M}_2, \ldots, \widehat{M}_k$ from observed data.
- **Step 3**: Set up and solve the system of equations $M_k = \widehat{M}_k$, for $k = 1, 2, \ldots, n$ to obtain the parameter estimates.

Although MM estimators are generally consistent and easy to compute, they may not always be efficient or unbiased. Moreover, they are not guaranteed to be sufficient, meaning they might not capture all the information available in the sample.

3.2 Maximum Likelihood Estimation Method (MLE)

Maximum Likelihood Estimation (MLE) is one of the most widely used methods in statistical inference for estimating unknown parameters based on observed data. The core idea of MLE is to determine the parameter values that maximize the likelihood function, which measures the probability of observing the given sample under different parameter settings [7, 12].

Let $X_1, X_2, \ldots, X_n$ be a random sample drawn from a population with PDF $f(x; \alpha)$, where α is an unknown parameter. The associated likelihood function $L(\alpha)$ is defined as the joint PDF of the sample:

$$L(\alpha) = \prod_{i=1}^{n} f(x_i; \alpha). \tag{3.4}$$

The value of α that maximizes the likelihood function is called the maximum likelihood estimator, denoted by $\widehat{\alpha}$. Formally, it is defined as:

$$\widehat{\alpha} = \arg\sup_{\alpha \in \Omega} L(\alpha), \tag{3.5}$$

where Ω is the parameter space.

MLE Procedure:. The steps to obtain the MLE are as follows:

- **Step 1:** Assume that $x_1, x_2, \ldots, x_n$ is a random sample from a population with PDF $f(x; \alpha)$.
- **Step 2:** Construct the likelihood function $L(\alpha)$ as shown in Equation (3.4).
- **Step 3:** Maximize $L(\alpha)$ (or equivalently, the log-likelihood function) with respect to α to obtain the MLE $\widehat{\alpha}$, as defined in Equation (3.5).

MLE is known for producing estimators that are consistent, asymptotically normal, and efficient under regularity conditions. However, it may require numerical optimization techniques when the likelihood function is complex or lacks a closed-form solution.

3.3 Bayesian Method (BM)

The Bayesian Method (BM) is a powerful approach to parameter estimation that incorporates prior knowledge or beliefs about the parameters through the use of a prior distribution. In contrast to frequentist methods, BM seeks to minimize the posterior expected loss or, equivalently, to maximize the posterior expected utility. This method is widely used not only in statistics but also in fields such as economics, social sciences, finance, and technology. Despite its broad applicability, the Bayesian approach is often considered more complex and less intuitive than MM and MLE.

Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with PDF $f(x \mid \alpha)$, where α is an unknown parameter to be estimated [9, 10]. The prior belief about α is represented by a

Table 1: Comparison of MM, MLE, and BM Estimation Methods

Criteria	MM	MLE	BM
Conceptual Basis	Matches sample moments to population moments	Maximizes the likelihood function based on observed data	Updates prior beliefs using observed data via Bayes' theorem
Computational Complexity	Generally simple and straightforward	Moderate; may require numerical optimization	Often complex; may require integration or simulation (e.g., MCMC)
Assumptions	Minimal assumptions; relies on existence of moments	Requires assumptions about the likelihood function	Requires specification of prior distribution and likelihood
Estimator Properties	Often biased; not necessarily efficient or sufficient	Consistent, asymptotically normal and efficient under regularity conditions	Depends on prior and loss function; can be optimal under specific criteria
Interpretability	Easy to understand and apply	Widely accepted and interpretable in frequentist framework	Provides full probabilistic interpretation of uncertainty
Use Cases	Quick approximations; simple models	Most common in statistical modeling and inference	Complex models; when prior information is available or needed

prior distribution $h(\alpha)$. Given the observed data, the posterior distribution of α is obtained using Bayes' theorem:

$$g(\alpha \mid x_1, \ldots, x_n) = \frac{h(\alpha) \prod_{i=1}^{n} f(x_i \mid \alpha)}{\int_{-\infty}^{\infty} h(\alpha) \prod_{i=1}^{n} f(x_i \mid \alpha)\, d\alpha}. \tag{3.6}$$

The posterior distribution $g(\alpha \mid x_1, \ldots, x_n)$ combines the prior information with the likelihood of the observed data, providing a complete probabilistic description of the uncertainty about α after observing the sample.

BM Procedure:. The steps to apply the Bayesian Method are as follows:

- **Step 1:** Specify the prior distribution $h(\alpha)$ based on prior knowledge or assumptions.
- **Step 2:** Compute the likelihood function $\prod_{i=1}^{n} f(x_i \mid \alpha)$ and use it to derive the posterior distribution as shown in Equation (3.6).
- **Step 3:** Minimize the posterior expected loss (or maximize the posterior expected utility) to obtain the Bayesian estimator of α.

Bayesian estimators are often derived under specific loss functions, such as the squared error loss, which leads to the posterior mean as the optimal estimator. While Bayesian methods offer a flexible and coherent framework for inference, they may require computational techniques such as Markov Chain Monte Carlo (MCMC) when analytical solutions are not feasible.

3.4 Comparison of the Three Estimation Methods

Below we provide a comparative summary table of MM, MLE, and BM as shown in Tab. 1. It summarizes the strengths and limitations of each method.

4 Applying Estimation Methods to the BD

4.1 Method of Moments (MM) for BD

To apply the Method of Moments (MM) to the Bernoulli distribution, we first note that the BD is characterized by a single parameter p, which represents the probability of success. Therefore, only the first-order moment is required for estimation.

The theoretical (population) first moment of a Bernoulli random variable $X \in \{0, 1\}$ is given by:

$$M_1 = \mathbb{E}[X] = p \cdot 1 + (1 - p) \cdot 0 = p. \tag{4.1}$$

The corresponding empirical (sample) first moment is computed as:

$$\widehat{M_1} = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{4.2}$$

where $x_1, x_2, \ldots, x_n$ are observed values from the sample.

According to the MM procedure, we solve equation $M_1 = \widehat{M_1}$ for the parameter p. This yields the MM estimator:

$$\widehat{p}_{\text{MM}} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{4.3}$$

Thus, the MM estimator for the Bernoulli distribution is simply the sample mean, which provides a straightforward and intuitive estimate of the success probability p.

4.2 Maximum Likelihood Estimation (MLE) for BD

Assume that $X_1, X_2, \ldots, X_n$ are independent and identically distributed Bernoulli random variables, i.e., $X_i \sim \text{Ber}(p)$. The probability mass function (PMF) of each X_i is given by:

$$f(x_i; p) = p^{x_i}(1 - p)^{1-x_i}, \quad x_i \in \{0, 1\}.$$

The likelihood function $L(p)$ for the sample is the product of the individual PMFs:

$$\begin{aligned} L(p) &= \prod_{i=1}^{n} f(x_i; p) = \prod_{i=1}^{n} p^{x_i}(1 - p)^{1-x_i} \\ &= p^{\sum_{i=1}^{n} x_i}(1 - p)^{n - \sum_{i=1}^{n} x_i}. \end{aligned}$$

Taking the natural logarithm of the likelihood function yields the log-likelihood function:

$$\begin{aligned} \ln L(p) &= \ln \left[p^{\sum x_i}(1 - p)^{n - \sum x_i} \right] \\ &= \left(\sum_{i=1}^{n} x_i \right) \ln p + \left(n - \sum_{i=1}^{n} x_i \right) \ln(1 - p). \end{aligned} \tag{4.4}$$

To find the MLE, we differentiate the log-likelihood with respect to p and set the derivative to zero:

$$\frac{d}{dp} \ln L(p) = \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1 - p} = 0.$$

Solving this equation gives:

$$\sum_{i=1}^{n} x_i = np \quad \Rightarrow \quad \widehat{p}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

To confirm that this solution corresponds to a maximum, we examine the second derivative:

$$\begin{aligned} \frac{d^2}{dp^2} \ln L(p) &= -\frac{\sum x_i}{p^2} - \frac{n - \sum x_i}{(1 - p)^2} \\ &= \frac{-n}{p(1 - p)} < 0, \end{aligned}$$

which is always negative for $0 < p < 1$, confirming that $\hat{p}_{\text{MLE}}$ is indeed the maximum.

Remark:. For the Bernoulli distribution, both the Method of Moments and Maximum Likelihood Estimation yield the same estimator:

$$\widehat{p}_{\mathrm{MM}} = \widehat{p}_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{4.5}$$

Given this equivalence and the desirable properties of MLE, subsequent analysis will focus primarily on the MLE approach.

4.3 Bayesian Method (BM) for BD

To apply the Bayesian approach to the Bernoulli distribution, we begin by specifying a prior distribution for the unknown parameter p, denoted by $h(p)$. For analytical convenience and conjugacy, we assume the prior follows a Beta distribution:

$$h(p) \propto p^{r-1}(1-p)^{s-1}, \quad \text{where } r, s > 0.$$

Given a sample $X_1, X_2, \ldots, X_n \sim \mathrm{Ber}(p)$, the likelihood function is:

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i}.$$

Combining the prior and likelihood, the unnormalized posterior distribution becomes:

$$h(p) \cdot L(p) \propto p^{r+\sum x_i - 1}(1-p)^{s+n-\sum x_i - 1}.$$

The normalizing constant is the Beta function:

$$\begin{aligned}\int_0^1 p^{r+\sum x_i-1}(1-p)^{s+n-\sum x_i-1}\,dp = \\ = B(r+\sum x_i, s+n-\sum x_i).\end{aligned}$$

Thus, the posterior distribution is:

$$\begin{aligned}g(p \mid x_1, \ldots, x_n) &= \frac{p^{r+\sum x_i-1}(1-p)^{s+n-\sum x_i-1}}{B(r+\sum x_i, s+n-\sum x_i)} \\ &\sim \mathrm{Beta}(r+\sum x_i, s+n-\sum x_i).\end{aligned}$$

The Bayesian estimator of p, under squared error loss, is the mean of the posterior distribution:

$$\widehat{p}_B = \mathbb{E}[p|x] = \frac{r+\sum x_i}{r+s+n}, \tag{4.6}$$

with the uncertainty captured by the posterior variance:

$$\mathrm{Var}(p \mid x) = \frac{(r+\sum_{i=1}^n x_i)(s+n-\sum_{i=1}^n x_i)}{(r+s+n)^2(r+s+n+1)}. \tag{4.7}$$

This completes the Bayesian estimation for the Bernoulli distribution. Next, we explore its fundamental statistical properties.

5 Important Properties of the BD

5.1 Basic Features: Mean, Variance, Skewness, Kurtosis

The Bernoulli distribution exhibits several key statistical properties:

Mean:.

$$\begin{aligned}\mu = E(X) &= \sum_{x=0}^{1} x f\left(x;p\right) = \sum_{x=0}^{1} x p^x (1-p)^{1-x} \\ &= 0 + p^1(1-p)^{1-1} = p.\end{aligned}$$

Higher-Order Moments:. For any integer $k \geq 1$, the k-th moment is:

$$\begin{aligned}E\left(X^k\right) &= \sum_{x=0}^{1} x^k f\left(x;p\right) = \sum_{x=0}^{1} x^k p^x \left(1-p\right)^{1-x} \\ &= 0 + p^1(1-p)^{1-1} = p, \quad k = 1, 2, \ldots\end{aligned} \tag{5.1}$$

Variance:.

$$\begin{aligned}\sigma^2 = \mathrm{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\ &= p - p^2 = p(1-p).\end{aligned} \tag{5.2}$$

Note that for Bernoulli random variables $X \in \{0,1\}$, the identity $X^k = X$ holds for all integers $k \geq 1$, which immediately implies $\mathbb{E}[X^k] = p$. Using this property, the skewness and kurtosis are determined as:

Skewness:.

$$\begin{aligned}\gamma_1 &= E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{E\left(X-\mu\right)^3}{\sigma^3} \\ &= \frac{E\left(X^3 - 3X^2\mu + 3X\mu^2 - \mu^3\right)}{\sigma^3} \\ &= \frac{E(X^3) - 3E(X^2)\mu + 3E(X)\mu^2 - \mu^3}{\sigma^3} \\ &= \frac{p - 3pp + 3pp^2 - p^3}{p(1-p).\sqrt{p(1-p)}} \\ &= \frac{1 - 3p + 3p^2 - p^2}{(1-p).\sqrt{p(1-p)}} = \frac{1 - 3p + 2p^2}{(1-p).\sqrt{p(1-p)}} \\ &= \frac{(1-p).(1-2p)}{(1-p).\sqrt{p(1-p)}} = \frac{1-2p}{\sqrt{p(1-p)}}.\end{aligned}$$

Kurtosis:.

$$\begin{aligned}
\gamma_2 &= E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{E\left(X-\mu\right)^4}{\sigma^4} \\
&= \frac{E\left(X^4 - 4X^3\mu + 6X^2\mu^2 - 4X\mu^3 + \mu^4\right)}{\sigma^4} \\
&= \frac{E(X^4) - 4E(X^3)\mu + 6E(X^2)\mu^2 - 4E(X)\mu^3 + \mu^4}{\sigma^4} \\
&= \frac{p - 4pp + 6pp^2 - 4pp^3 + p^4}{p(1-p).p(1-p)} \\
&= \frac{1 - 4p + 6p^2 - 4p^3 + p^3}{(1-p).p(1-p)} \\
&= \frac{1 - 4p + 6p^2 - 3p^3}{(1-p).p(1-p)} \\
&= \frac{(1-p)(1 - 3p + 3p^2)}{(1-p).p(1-p)} \\
&= \frac{1 - 3p + 3p^2}{p(1-p)}.
\end{aligned}$$

These expressions highlight the simplicity and symmetry of the Bernoulli distribution, making it a foundational model in binary data analysis.

5.2 Estimation Properties: Unbiasedness and Consistency

In statistical inference, an estimator is considered reliable if it satisfies two fundamental properties: **unbiasedness** and **consistency**. These criteria are widely accepted as essential standards for evaluating the quality of an estimator. In this section, we briefly review these concepts and apply them to the estimators of the Bernoulli distribution.

5.2.1 Unbiasedness

An estimator $\widehat{p}$ is said to be *unbiased* if its expected value equals the true parameter it estimates. Mathematically, this is expressed as:

$$\mathbb{E}[\widehat{p}] = p.$$

The bias of an estimator is defined as:

$$\text{Bias}(\widehat{p}, p) = \mathbb{E}[\widehat{p}] - p. \tag{5.3}$$

A bias close to zero indicates a good estimator, and a bias of exactly zero implies perfect unbiasedness.

Unbiasedness of MLE:. Recall that the MLE for the Bernoulli distribution is:

$$\widehat{p}_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Its expected value is:

$$\begin{aligned}
\mathbb{E}[\widehat{p}_{\text{ML}}] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[x_i] \\
&= \frac{1}{n}\cdot n \cdot p = p.
\end{aligned}$$

Hence, $\widehat{p}_{\text{ML}}$ is an unbiased estimator of p.

Unbiasedness of Bayesian Estimator:. The Bayesian estimator is given by:

$$\widehat{p}_B = \frac{r + \sum_{i=1}^{n} x_i}{r + s + n}.$$

Its expected value is:

$$\mathbb{E}[\widehat{p}_B] = \frac{r + \mathbb{E}\left[\sum_{i=1}^{n} x_i\right]}{r + s + n} = \frac{r + np}{r + s + n} \neq p. \tag{5.4}$$

Thus, $\widehat{p}_B$ is a biased estimator. However, it is *asymptotically unbiased*, as shown below:

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}[\widehat{p}_B] &= \lim_{n\to\infty} \left(\frac{r + np}{r + s + n}\right) \\ &= \lim_{n\to\infty} \left(\frac{\frac{r}{n} + p}{\frac{r}{n} + \frac{s}{n} + 1}\right) = p.\end{aligned}$$

This implies that the bias of $\widehat{p}_B$ diminishes as the sample size increases, making it a reliable estimator in large samples.

5.2.2 Consistency

In statistical estimation, an estimator is said to be *consistent* if it converges in probability to the true parameter value as the sample size increases. One common way to assess consistency is through the **Mean Squared Error (MSE)**, which combines both the variance and the squared bias of the estimator.

The MSE of an estimator $\widehat{p}$ for a parameter p is defined as:

$$\begin{aligned}\mathrm{MSE}(\widehat{p}) &= \mathbb{E}[(\widehat{p} - p)^2] \\ &= \mathrm{Var}(\widehat{p}) + \left[\mathrm{Bias}(\widehat{p}, p)\right]^2.\end{aligned} \tag{5.5}$$

Since MSE is derived from the squared Euclidean distance, it is always non-negative and ideally approaches zero as the sample size increases.

Consistency of MLE:. Recall that the MLE for the Bernoulli distribution is:

$$\widehat{p}_{\mathrm{ML}} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

And it is an unbiased estimator, so:

$$\mathrm{Bias}(\widehat{p}_{\mathrm{ML}}, p) = 0.$$

The variance of $\widehat{p}_{\mathrm{ML}}$ is:

$$\begin{aligned}\mathrm{Var}(\widehat{p}_{\mathrm{ML}}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) &= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(x_i) \\ = \frac{np(1-p)}{n^2} &= \frac{p(1-p)}{n}.\end{aligned}$$

Thus, the MSE is:

$$\mathrm{MSE}(\widehat{p}_{\mathrm{ML}}) = \frac{p(1-p)}{n} \to 0 \quad \text{as } n \to \infty.$$

This confirms that $\widehat{p}_{\mathrm{ML}}$ is a consistent estimator of p.

Consistency of Bayesian Estimator:. The Bayesian estimator is:

$$\widehat{p}_B = \frac{r + \sum_{i=1}^{n} x_i}{r + s + n}.$$

Its bias is:

$$\text{Bias}(\widehat{p}_B, p) = \frac{r + np}{r + s + n} - p,$$

And its variance is:

$$\begin{aligned}\text{Var}(\widehat{p}_B) &= \text{Var}\left(\frac{r + \sum x_i}{r + s + n}\right) \\ &= \frac{1}{(r + s + n)^2} \sum_{i=1}^{n} \text{Var}(x_i) = \frac{np(1-p)}{(r + s + n)^2}.\end{aligned}$$

Therefore, the MSE is:

$$\begin{aligned}\text{MSE}(\widehat{p}_B) &= \frac{np(1-p)}{(r + s + n)^2} + \left(\frac{r + np}{r + s + n} - p\right)^2 \\ &= \frac{np(1-p)}{(r + s + n)^2} + \left(\frac{r - rp - sp}{r + s + n}\right)^2 \quad \to 0, \quad \text{as} \quad n \to \infty.\end{aligned}$$

This confirms that $\widehat{p}_B$ is also a consistent estimator of p, despite being biased in finite samples.

6 Systematic Study of Beta(r,s) Priors

6.1 How to Choose Hyperparameters of the Prior Beta(r, s)?

Unlike frequentist estimation, which relies solely on the observed data $(x_1, \ldots, x_n)$ to estimate the unknown parameter p of a Bernoulli distribution via methods such as maximum likelihood (MLE) or method of moments (MM), the Bayesian approach incorporates prior beliefs about p. Specifically, we assume a prior distribution

$$p_{\text{prior}} \sim \text{Beta}(r, s), \tag{6.1}$$

which encodes our belief about the unknown parameter p. After observing data $(x_1, \ldots, x_n)$, we update the estimation by constructing the posterior distribution given by:

$$p_{\text{posterior}} \sim \text{Beta}\left(r + \sum_{i=1}^{n} x_i,\ s + n - \sum_{i=1}^{n} x_i\right). \tag{6.2}$$

Under squared-error loss, the optimal Bayes estimator is the posterior mean:

$$\widehat{p}_B = \frac{r + \sum_{i=1}^{n} x_i}{r + s + n} = w \cdot \frac{r}{r + s} + (1 - w) \cdot \frac{\sum_{i=1}^{n} x_i}{n}, \tag{6.3}$$

which is a weighted sum of the prior mean and sample mean with the weight defined by $w = \dfrac{r + s}{r + s + n}$, showing shrinkage toward the prior mean $\dfrac{r}{r + s}$. The choice of hyperparameters (r, s) determines the prior's informativeness and directly influences the shrinkage properties of the Bayes estimator. Additionally, to capture the uncertainty of the Bayes estimator, one can use the posterior variance:

$$\text{Var}(p \mid x) = \frac{(r + \sum_{i=1}^{n} x_i)(s + n - \sum_{i=1}^{n} x_i)}{(r + s + n)^2 (r + s + n + 1)}. \tag{6.4}$$

Below is the guidelines for selecting hyperparameters of prior Beta(r, s),

- Use Beta$(1,1)$ when no prior information is available, i.e. Beta$(1,1)$ is used as a noninformative prior, corresponding to a uniform distribution over $p \in [0,1]$.
- Use Beta$(1/2,1/2)$ (known as Jeffreys prior) for objective Bayesian inference, especially when invariance and minimax risk properties are important.
- Use Beta(r,r) when prior knowledge or symmetry considerations suggest shrinkage toward 0.5, with r controlling the strength of prior belief.
- Use **Beta**$(ap, a(1-p))$ when we have informative prior encoding prior knowledge about p, with concentration controlled by a.

For a general prior distribution Beta(r,s), suppose we have prior beliefs that the true parameter p has mean π and standard deviation σ. The hyperparameters (r,s) can then be determined by solving the moment-matching equations:

$$\mathbb{E}[p_{\text{prior}}] = \pi, \qquad \text{Var}(p_{\text{prior}}) = \sigma^2. \tag{6.5}$$

Since

$$\mathbb{E}[p_{\text{prior}}] = \frac{r}{r+s}, \qquad \text{Var}(p_{\text{prior}}) = \frac{rs}{(r+s)^2(r+s+1)}, \tag{6.6}$$

we obtain the system

$$\begin{cases} \dfrac{r}{r+s} = \pi, \\ \dfrac{rs}{(r+s)^2(r+s+1)} = \sigma^2. \end{cases} \iff \begin{cases} r = \dfrac{\pi^2}{\sigma^2}(1-\pi) - \pi, \\ s = \dfrac{1-\pi}{\pi}\, r. \end{cases} \tag{6.7}$$

Special Case: Informative Prior Beta$(ap, a(1-p))$

A useful parametrization arises when we set $r = ap$ and $s = a(1-p)$, yielding the prior

$$p_{\text{prior}} \sim \text{Beta}(ap,\ a(1-p)). \tag{6.8}$$

In this case, the prior mean is exactly $p = \pi$, while the concentration parameter a controls the strength of belief around π. Matching the variance to σ^2 gives us a formula to find

$$a = \frac{\pi(1-\pi)}{\sigma^2} - 1. \tag{6.9}$$

Thus, larger values of a correspond to stronger prior concentration (smaller variance), while smaller values of a yield more diffuse priors. In Section 7, we will provide numerical illustrations to compare Bayes estimator under several priors above with varying sample sizes.

6.2 Asymptotic Risk Analysis and Minimaxity

We consider the risk function under squared-error loss:

$$\text{MSE}(p, \widehat{p}) = \mathbb{E}\left[(\widehat{p} - p)^2\right]. \tag{6.10}$$

As $n \to \infty$, we have the weight $w \to 0$, and thus $\widehat{p}_B \to \hat{p}_{MLE}$, equalizing asymptotic risk. For small finite sample sizes, n, Bayes estimators typically have lower risk due to shrinkage, while the MLE remains unbiased but with higher variance. Table 2 highlights risk behavior comparison of Bayes estimators under some selected Beta(r,s) priors [5, 14]. Further, Bayesian decision theory provides conditions under which Bayes estimators are minimax or admissible. For example, Jeffreys prior Beta$(1/2,1/2)$) yield minimax estimators under squared-error loss, while informative priors such as Beta$(ap, a(1-p))$ are Bayes rules when prior knowledge about p exists [5, 14]. In Section 7, we shall provide simulation studies to illustrate risk curves for different priors compared to the MLE.

Table 2: Comparison of Bayes estimators under selected Beta(r, s) priors in the Bernoulli model.

Prior	Shrinkage target	Concentration	Finite-sample bias	Risk behavior
Beta(1/2, 1/2)	0.5	Very low	Very low	Often strong frequentist properties; near-minimax under many losses
Beta(2, 2)	0.5	Moderate	Moderate toward 0.5	Reduced variance vs. MLE; improved small-n MSE near 0.5
Beta$(ap, a(1-p))$	p	Tunable via a	Toward p (magnitude $\propto a$)	When p is well specified, dominates MLE in small n via lower MSE

7 Simulation Studies

All computations in this study, including simulation experiments and real data analyses, were conducted using the R statistical software. To perform simulation studies for the Bernoulli distribution, we generate synthetic data. In R, this is conveniently done using the function `rbinom(n, size = 1, prob)`, where:

- `n` is the sample size (number of observations),
- `size` is the number of trials per observation (set to 1 for BD),
- `prob` is the probability of success, with values in the interval $(0, 1)$.

To ensure a comprehensive evaluation, the probability parameter p is varied across nine values: $0.1, 0.2, \ldots, 0.9$, while the sample sizes n considered are: $25, 50, 100, 200, 500, 1000, 2000, 5000$. Additionally, each simulation scenario is repeated 2000 times to ensure statistical stability.

For each simulated dataset, the parameter p is estimated using two methods:

- Maximum Likelihood Estimation (MLE): $\widehat{p}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i$,
- Bayesian Method (BM): $\widehat{p}_B = \frac{r+\sum_{i=1}^{n} x_i}{r+s+n}$, under several priors $\beta(r, s)$ discussed in Section 6.1. Especially, we determine (r, s) using the moment-matching formula (6.7), which aligns the prior mean and variance with specified beliefs. For example:
  - When the true proportion is $p = 0.3$ or $p = 0.5$, we assume a prior belief centered at $\pi = 0.4$ with standard deviation $\sigma = 0.10$. Solving (6.7) yields $r = 9$ and $s = 14$, corresponding to a concentration parameter $a = 23$.
  - When the true proportion is $p = 0.7$, we set the prior belief at $\pi = 0.65$ with $\sigma = 0.10$. This results in $r = 14$ and $s = 18$, giving a concentration parameter $a = 22$.

To assess the performance of these estimators, three metrics are computed:

- **Bias**, calculated using Equation (5.3),
- **Mean Squared Error (MSE)**, calculated using Equation (5.5),
- **Standard Deviation (SD)**, computed using the `sd()` function in R.

7.1 Comparison of Estimators under Single Non-informative Prior Beta(1,1)

The results of the simulation studies with uniform prior $r = s = 1$, are presented in Table 3, given in the appendix. We now discuss the findings presented in Table 3. Overall, the results for all three evaluation metrics: Bias, SD, and MSE are consistently small across different values of p and sample sizes. This indicates that both the Maximum Likelihood Estimation and Bayesian Method provide highly reliable estimates for the Bernoulli distribution.

Bias:. The bias values for both estimators are generally low, with MLE showing slightly better performance than BM in most cases. However, the differences are minor and statistically insignificant. As expected, bias tends to decrease as the sample size increases, confirming the asymptotic unbiasedness of both estimators.

An interesting observation is that for values of $p \geq 0.6$, the bias tends to be negative. Moreover, the bias values for p and $1-p$ are nearly symmetric, which is intuitive given the binary nature of the Bernoulli distribution. Since p represents the probability of success (i.e., the proportion of ones), this symmetry reflects the inherent balance between the two outcomes (0 and 1). It is worth noting that the sign of the bias is not critical; what matters is its magnitude, which can be assessed using absolute values.

Standard Deviation (SD):. The SD values are uniformly small across all scenarios, further supporting the reliability of both estimation methods. For smaller sample sizes (i.e., $n < 200$), BM tends to yield slightly lower SD values than MLE, although the difference is negligible. For larger sample sizes, the SD values from both methods converge and become nearly identical, demonstrating the robustness of both approaches. The empirical standard deviations closely match the theoretical values $\sqrt{p(1-p)/n}$, confirming the validity of the simulation setup.

Mean Squared Error (MSE):. The MSE results reinforce the conclusions drawn from the bias and SD analyses. Both MLE and BM produce low MSE values across all settings. For extreme values of p (i.e., 0.1 and 0.9) with small sample sizes, MLE slightly outperforms BM. Conversely, for moderate values of p, BM shows marginally better MSE performance. However, these differences are minimal and do not affect the overall conclusion.

7.2 Comparison of Estimators under Several Prior Beta(r,s)

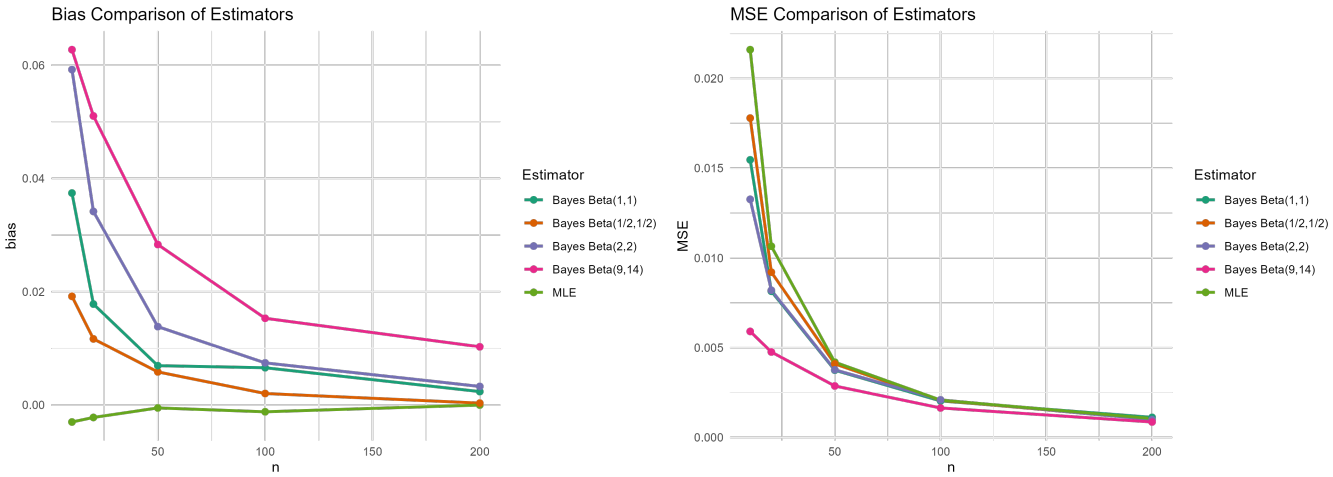
In this subsection, we analyze the simulation results presented in Figures 3–5, which compare the bias and MSE (risk) of the MLE and Bayes estimators under different Beta priors. The priors considered include the uniform prior Beta$(1,1)$, Jeffreys prior Beta$(1/2,1/2)$, symmetric prior Beta$(2,2)$, and informative priors of the form Beta$(ap, a(1-p))$ determined via moment matching.

7.2.1 Case $p = 0.3$: Figure 2

- **Bias:** The MLE remains nearly unbiased across all sample sizes. Bayes estimators with symmetric priors (Beta$(1,1)$, Beta$(1/2,1/2)$, Beta$(2,2)$) exhibit positive bias at small n due to shrinkage toward 0.5. The informative prior Beta$(9,14)$ reduces this bias since its prior mean (≈ 0.39) is closer to the true $p = 0.3$.
- **MSE:** Bayes estimators achieve lower MSE than MLE for small n by reducing variance. The informative prior Beta$(9,14)$ performs best, highlighting the advantage of aligning the prior mean with the true parameter.
- **Conclusion:** Informative priors near the true p improve small-sample performance, while symmetric priors induce more bias when p is far from 0.5.

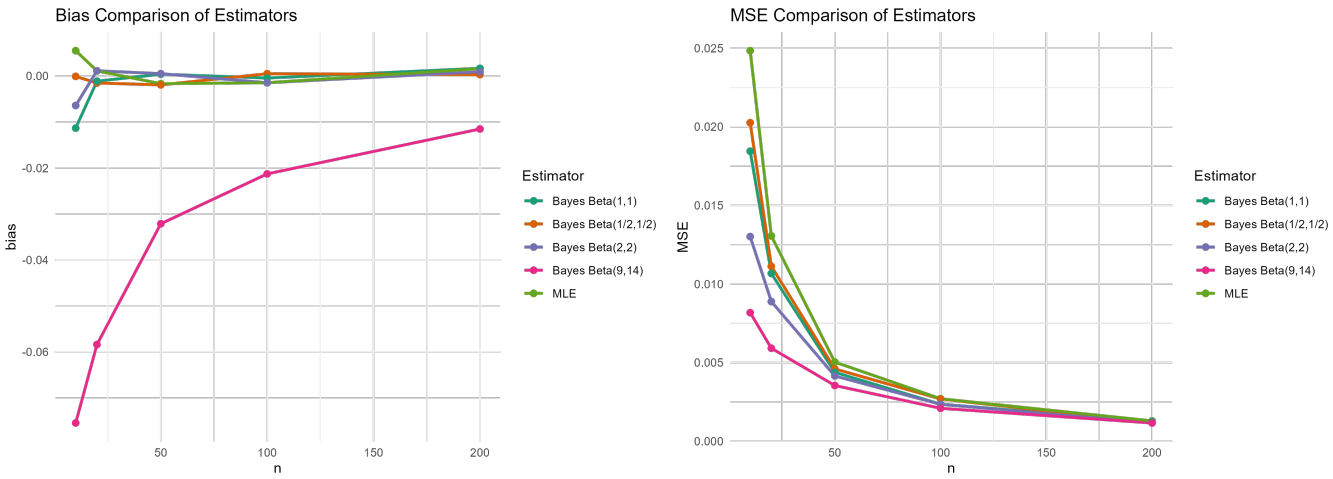
7.2.2 Case $p = 0.5$: Figure 3

- **Bias:** The MLE is unbiased. Symmetric priors centered at 0.5 yield minimal bias, which vanishes quickly as n increases. The informative prior Beta$(9,14)$ introduces negative bias at small n since its mean (≈ 0.39) is misspecified relative to $p = 0.5$.
- **MSE:** Symmetric priors (Beta$(1,1)$, Beta$(1/2,1/2)$, Beta$(2,2)$) yield the lowest MSE for small n because their shrinkage target matches the true parameter. The misspecified



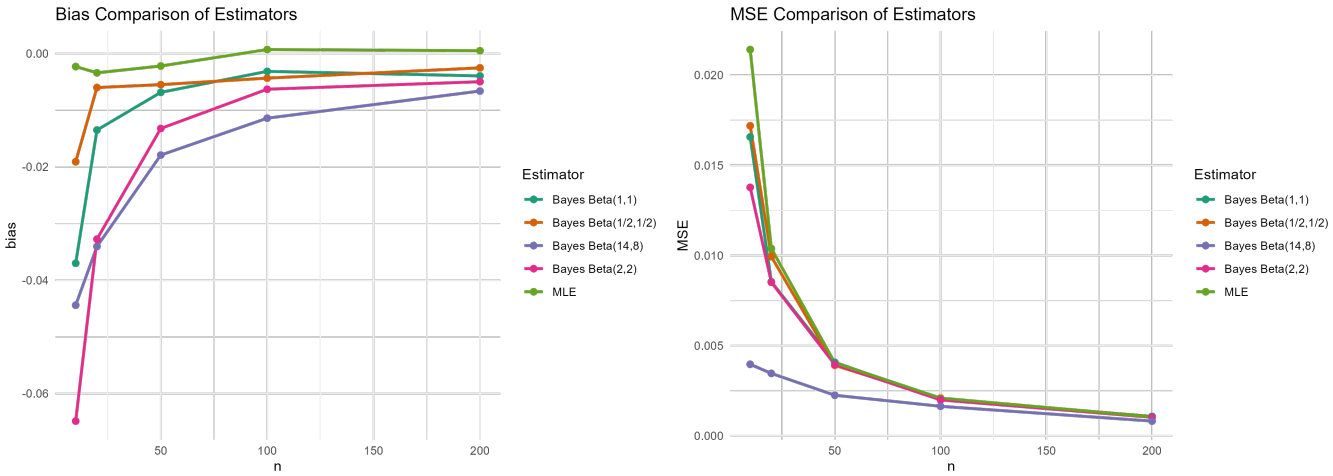
(a) Bias **(b)** MSE

Figure 2: Comparison of estimators under several priors, where the bias and MSE are as functions of the sample size n for p=0.3.



(a) Bias **(b)** MSE

Figure 3: Comparison of estimators under several priors, where the bias and MSE are as functions of the sample size n for p=0.5.



(a) Bias **(b)** MSE

Figure 4: Comparison of estimators under several priors, where the bias and MSE are as functions of the sample size n for p=0.7.

informative prior has higher MSE initially but converges as n grows.

- **Conclusion:** When $p \approx 0.5$, symmetric priors are optimal. Informative priors with incorrect mean can degrade small-sample performance.

7.2.3 Case $p = 0.7$: Figure 4

- **Bias:** The MLE remains unbiased. Symmetric priors shrink toward 0.5, producing negative bias at small n. The informative prior Beta$(14, 18)$ (mean ≈ 0.66) reduces bias by shrinking toward a value closer to $p = 0.7$.
- **MSE:** The informative prior achieves the lowest MSE for small n, outperforming symmetric priors whose shrinkage target is suboptimal. As n increases, all estimators converge.
- **Conclusion:** For p far from 0.5, informative priors centered near the true value yield superior small-sample risk properties.

7.2.4 Cross-Case Insights

- Bayes estimators improve small-sample MSE primarily by reducing variance, at the expense of bias toward the prior mean.
- The magnitude and direction of bias depend on the gap between the prior mean and the true p; bias diminishes as n grows.
- Symmetric priors are most effective when $p \approx 0.5$, while informative priors are preferable when p is far from 0.5.
- Asymptotically, all estimators converge, so prior choice matters most in small to moderate samples.

The simulation results demonstrate that both MLE and BM are effective and reliable methods for estimating the parameter p in the Bernoulli distribution. While MLE is slightly more accurate in terms of bias and MSE for certain scenarios, BM performs comparably well and offers additional flexibility through prior incorporation. These findings validate the use of both methods in practical applications involving binary data.

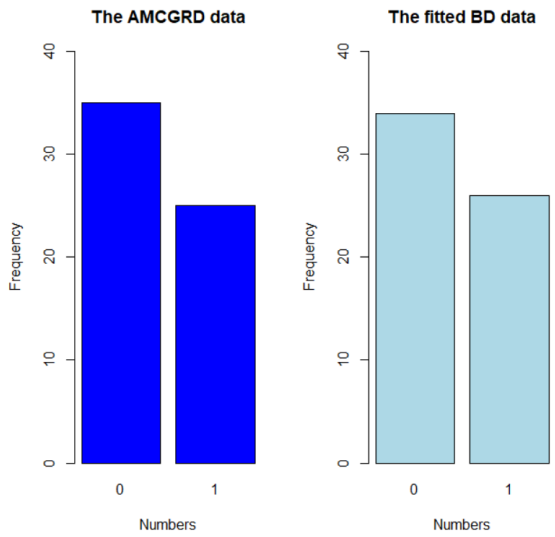


Figure 5: PMF of the AMCGRD data and the fitted Bernoulli distribution.

8 Practical Examples

Based on the findings from the simulation studies, the Maximum Likelihood Estimation method demonstrates slightly greater stability and computational simplicity compared to the Bayesian Method. For this reason, MLE is selected for analyzing real-world datasets in this section. To provide a comprehensive evaluation, four actual datasets are considered.

Note that all original datasets were binarized to illustrate the Bernoulli modeling framework. As this transformation simplifies the original data structure, the conclusions apply specifically to the binary representation.

8.1 Data 1: AMCGRD Dataset

The first dataset pertains to the number of mistakes made in copying groups of random digits, referred to as AMCGRD. This dataset was analyzed in the study by Atikankul (2023) [3]. A detailed description of the dataset is available in the cited work.

Although the AMCGRD dataset originally consists of count data, it can be transformed into binary data by encoding errors as 1 and non-errors as 0. This transformation allows the Bernoulli distribution to be applied. The dataset contains a total of 60 observations, of which 35 are zeros.

Using R statistical software, the MLE for the Bernoulli parameter is computed as:

$$\widehat{p}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i = 0.4167.$$

Figure 5 displays the probability mass function (PMF) of the AMCGRD data alongside the fitted Bernoulli distribution using the estimated parameter $\widehat{p}_{\mathrm{ML}} = 0.4167$.

As shown in Figure 5, the fitted Bernoulli distribution closely matches the empirical distribution of the AMCGRD data. Specifically, the proportion of zeros in the AMCGRD dataset is $\frac{35}{60} \approx 0.5833$, while the fitted Bernoulli model predicts approximately $\frac{34}{60} \approx 0.5667$ zeros. The similarity in these proportions and the visual alignment of the PMFs indicate that the Bernoulli distribution provides a good fit for the AMCGRD data. This supports its applicability in modelling binary outcomes in scientific research.

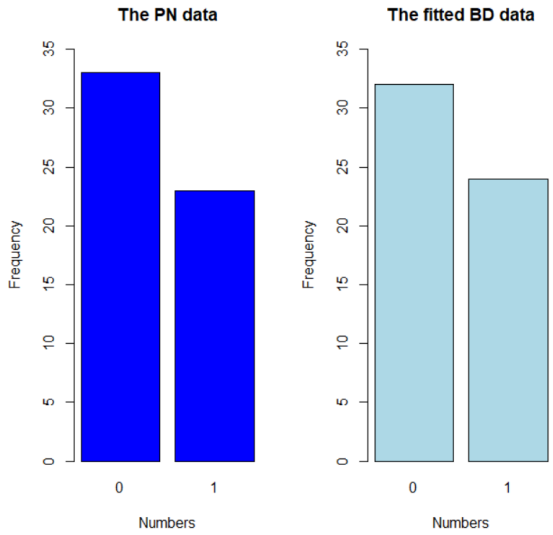


Figure 6: PMF of the PN data and the fitted Bernoulli distribution.

8.2 Data 2: Pyrausta Nublilalis (PN) Dataset

The second dataset analyzed is the Pyrausta Nublilalis (PN), also referenced in the study by Atikankul (2023) [3]. Similar to the AMCGRD dataset, PN originally consists of count data and is transformed into a binary format for analysis. The encoding is straightforward: if PN is absent (i.e., no occurrence), it is recorded as 0; otherwise, it is recorded as 1. The dataset comprises 56 survey subjects, with 23 instances of PN occurrence (i.e., ones). Given its binary nature, the Bernoulli distribution is appropriate for modelling this data.

Using R statistical software, the MLE for the Bernoulli parameter is calculated as:

$$\widehat{p}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i = 0.4107.$$

Figure 6 presents the PMF of the PN data alongside the fitted Bernoulli distribution using the estimated parameter $\widehat{p}_{\mathrm{ML}} = 0.4107$.

As illustrated in Figure 6, the fitted Bernoulli distribution closely approximates the empirical distribution of the PN data. Specifically, the number of zeros in the PN dataset is 33 out of 56 observations, while the fitted Bernoulli model predicts approximately 32 zeros. The visual similarity between the two PMFs confirms that the Bernoulli distribution provides a good fit for the PN data, reinforcing its utility in modeling binary outcomes in applied research.

8.3 Data 3: Fishing Dataset

The third dataset analyzed is the Fishing Data Set (FDS), a commonly used binary dataset that was examined in the study by Pho (2023) [18]. The dataset consists of responses from 250 individuals regarding the number of fish caught during fishing activities. The data is encoded into binary format: a value of 0 indicates no fish caught, while a value of 1 indicates at least one fish caught. Out of the 250 observations, 142 are zeros, indicating no catch. Given its binary nature, the Bernoulli distribution is suitable for modelling this dataset.

Using R statistical software, the MLE for the Bernoulli parameter is computed as:

$$\widehat{p}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i = 0.432.$$

Figure 7 displays the PMF of the FDS data alongside the fitted Bernoulli distribution using the estimated parameter $\widehat{p}_{\mathrm{ML}} = 0.432$.

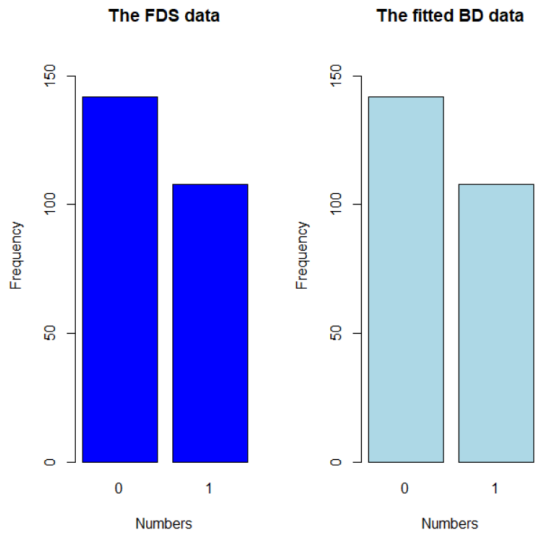


Figure 7: PMF of the FDS data and the fitted Bernoulli distribution.

As shown in Figure 7, the fitted Bernoulli distribution closely matches the empirical distribution of the FDS data. Specifically, the number of zeros in the actual dataset is 142, while the fitted Bernoulli model predicts approximately 141 zeros out of 250 observations. The minimal difference between the two distributions confirms that the Bernoulli model provides a strong fit for the FDS data, further validating its use in modelling binary outcomes in applied research.

8.4 Data 4: Infected Blood Cells (IBC) Dataset

The fourth dataset analyzed is the count of infected blood cells (IBC), as discussed in the study by Lemonte, Moreno-Arenas, and Castellares (2020) [15]. Like the previous datasets, the IBC data originally consists of count values and is transformed into a binary format for analysis. The encoding is defined as follows: if an individual has no infected blood cells, the value is recorded as 0; otherwise, it is recorded as 1. The dataset includes a total of 511 observations, with 314 zeros indicating no infection. Given its binary nature, the Bernoulli distribution is applied to model the data.

Using R statistical software, the MLE for the Bernoulli parameter is computed as:

$$\widehat{p}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i = 0.3856.$$

Figure 8 presents the PMF of the IBC data alongside the fitted Bernoulli distribution using the estimated parameter $\widehat{p}_{\mathrm{ML}} = 0.3856$.

As shown in Figure 8, the fitted Bernoulli distribution closely approximates the empirical distribution of the IBC data. Specifically, the actual data contains 314 zeros, while the fitted model predicts approximately 316 zeros out of 511 observations. The difference is minimal and visually indistinguishable in the PMF plot. This further supports the effectiveness of the Bernoulli distribution in modelling binary outcomes, especially in datasets with clear dichotomous characteristics.

9 Discussions and Conclusions

Although this study is conceptually straightforward, it addresses a highly significant topic in statistical modelling - the Bernoulli distribution (BD), which is foundational in the family of binary distributions. The BD plays a crucial role in constructing zero-inflated (ZI) count distributions, which are widely used in modelling data with excess zeros. These ZI models

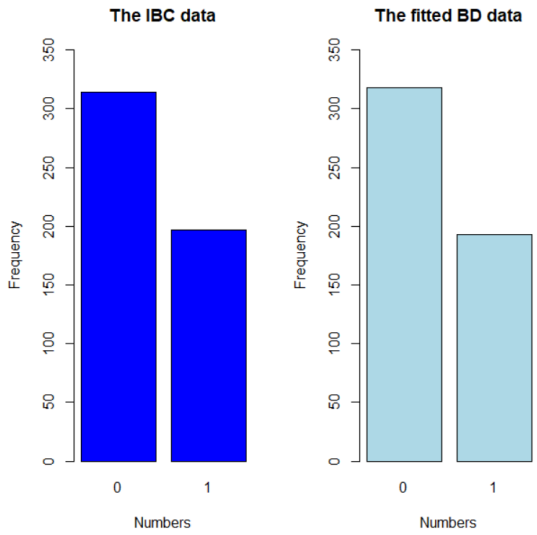


Figure 8: PMF of the IBC data and the fitted Bernoulli distribution.

typically involve two components, one of which is governed by the BD. Therefore, understanding the BD in depth is essential for developing new distributions and enhancing analytical tools for real-world data.

In this work, we have presented the general formulation of the BD and examined three widely used point estimation methods: the Method of Moments (MM), Maximum Likelihood Estimation (MLE), and the Bayesian Method (BM). For each method, we provided clear algorithms to facilitate practical implementation. These methods were then specifically applied to the BD, and their performance was evaluated through both theoretical properties and empirical analysis.

We explored the fundamental statistical features of the BD, including its mean, variance, skewness, and kurtosis. Additionally, we investigated the properties of the estimators: unbiasedness and consistency, through analytical derivations. The simulation studies demonstrated that both MLE and BM yield highly reliable estimates, with MLE offering slightly better performance in terms of bias and computational simplicity.

A key insight is that the choice of prior in Bayesian estimation critically influences performance. Jeffreys prior Beta$(1/2, 1/2)$ exhibits near-minimax properties and serves as a robust default. Symmetric priors Beta(r, r) are most effective when the true parameter is near 0.5, while informative priors Beta$(ap, a(1-p))$ aligned with prior knowledge yield superior small-sample performance when p is far from 0.5. These findings provide practical guidance for selecting estimation methods depending on sample size and the availability of prior information.

To further validate the applicability of the BD, we analyzed four real-world datasets. In each case, the BD provided a strong fit to the binary-transformed data, reinforcing its utility in practical research. The MLE was used for estimation due to its simplicity and stability, and the results consistently showed minimal differences between empirical and fitted distributions.

It is important to note that both MLE and BM require complete datasets and cannot be directly applied to data with missing values. The challenges and solutions related to missing data are discussed in works such as Schafer and Graham (2002) [23], Little (1992) [16], and Lee, Pho, and Li (2021) [13].

From a computational perspective, MLE is particularly advantageous due to its compatibility with pre-programmed optimization functions in R, such as `optim` [26], `nleqslv`, and `maxLik`. Among these, the `optim` function is noted for its speed and efficiency. In contrast, implementing BM often requires custom coding and more complex calculations, making MLE a more practical

choice for many scientific applications.

Final Remarks. This study highlights the importance of the Bernoulli distribution in both theoretical and applied statistics. By providing a comprehensive overview of its properties, estimation methods, and practical applications, we hope to encourage further exploration and utilization of the BD in statistical modelling, especially in the development of new ZI distributions and binary data analysis.

10 Declarations

Funding

This research received no external funding.

Competing Interests

The author declare no conflict of interest.

Ethical Approval

Not applicable

Authors's Contributions

Not applicable

Data Availability Statement

Not applicable.

References

[1] Yasin Altinisik and Emel Cankaya, *New zero-inflated regression models with a variant of censoring*, Brazilian Journal of Probability and Statistics **36** (2022), no. 4, 641–674.

[2] Sirinapa Aryuyuen, Winai Bodhisuwan, and Thidaporn Supapakorn, *Zero inflated negative binomial-generalized exponential distribution and its applications*, Songklanakarin Journal of Science and Technology **36** (2014), no. 4, 483–491.

[3] Yupapin Atikankul, *A New Poisson Generalized Lindley Regression Model*, Austrian Journal of Statistics **52** (2023), no. 1, 39–50.

[4] Gladys DC Barriga and Francisco Louzada, *The zero-inflated Conway–Maxwell–Poisson distribution: Bayesian inference, regression modeling and influence diagnostic*, Statistical Methodology **21** (2014), 23–34.

[5] James O Berger, *Statistical decision theory and Bayesian analysis*, Springer Science & Business Media, 2013.

[6] Whitten Betty Jones and A Clifford Cohen, *Further considerations of modified moment estimators for Weibull and lognormal parameters*, Communications in Statistics-Simulation and Computation **25** (1996), no. 3, 749–784.

[7] JV Carnahan, *Maximum likelihood estimation for the 4-parameter beta distribution*, Communications in Statistics-Simulation and Computation **18** (1989), no. 2, 513–536.

[8] Zdeněk Fabián, *Score function of distribution and revival of the moment method*, Communications in Statistics-Theory and Methods **45** (2016), no. 4, 1118–1136.

[9] Malay Ghosh, *Nonparametric empirical bayes estimation of certain funtionals*, Communications in Statistics-Theory and Methods **14** (1985), no. 9, 2081–2094.

[10] Md Mobarak Hossain, Tomasz J Kozubowski, and Krzysztof Podgórski, *A novel weighted likelihood estimation with empirical Bayes flavor*, Communications in Statistics-Simulation and Computation **47** (2018), no. 2, 392–412.

[11] Kirtee K Kamalja and Yogita S Wagh, *Estimation in zero-inflated Generalized Poisson distribution*, Journal of Data Science **16** (2018), no. 1, 183–206.

[12] S Lalitha and Anand Mishra, *Modified maximum likelihood estimation for Rayleigh distribution*, Communications in Statistics-Theory and methods **25** (1996), no. 2, 389–401.

[13] Shen-Ming Lee, Kim-Hung Pho, and Chin-Shang Li, *Validation likelihood estimation method for a zero-inflated Bernoulli regression model with missing covariates*, Journal of Statistical Planning and Inference **214** (2021), 105–127.

[14] Erich Leo Lehmann and George Casella, *Theory of point estimation*, Springer, 1998.

[15] Artur J Lemonte, Germán Moreno-Arenas, and Fredy Castellares, *Zero-inflated Bell regression models for count data*, Journal of Applied Statistics (2020).

[16] Roderick JA Little, *Regression with missing X's: a review*, Journal of the American statistical association **87** (1992), no. 420, 1227–1237.

[17] Karl Pearson, *On the systematic fitting of curves to observations and measurements*, Biometrika **1** (1902), no. 3, 265–303.

[18] Kim-Hung Pho, *Zero-inflated probit Bernoulli model: A new model for binary data*, Communications in Statistics-Simulation and Computation (2023), 1–21.

[19] Luisa Rivas and Francisca Campos, *Zero inflated Waring distribution*, Communications in Statistics-Simulation and Computation **52** (2023), no. 8, 3676–3691.

[20] Pornpop Saengthong, Winai Bodhisuwan, and Ampai Thongteeraparp, *The zero inflated negative binomial–Crack distribution: some properties and parameter estimation*, Songklanakarin J. Sci. Technol **37** (2015), no. 6, 701–711.

[21] K. M. Sakthivel, C. S. Rajitha, and K. B. Alshad, *Zero-inflated negative binomial-Sushila distribution and its application*, International Journal of Pure and Applied Mathematics **117** (2017), no. 13, 117–126.

[22] C Satheesh Kumar and Rakhi Ramachandran, *On zero-inflated hyper-Poisson distribution and its applications*, Communications in Statistics-Simulation and Computation **51** (2020), no. 3, 868–881.

[23] Joseph L Schafer and John W Graham, *Missing data: our view of the state of the art.*, Psychological methods **7** (2002), no. 2, 147.

[24] S. Singh, *A note on inflated Poisson distribution*, Journal of the Indian Statistical Association **33** (1963), no. 1, 140–144.

[25] Caner Tanış, Haydar Koç, and Ahmet Pekgör, *A new zero–inflated discrete Lindley regression model*, Communications in Statistics-Theory and Methods **53** (2024), no. 12, 4252–4271.

[26] Buu-Chau Truong, Van-Buol Nguyen, Hoang-Vinh Truong, and Thi Diem-Chinh Ho, *Comparison of optim, nleqslv and MaxLik to estimate parameters in some of regression models*, Journal of Advanced Engineering and Computation **3** (2019), no. 4, 532–550.

[27] Mohammad Kafeel Wani and Peer Bilal Ahmad, *Zero-inflated Poisson-Akash distribution for count data with excessive zeros*, Journal of the Korean Statistical Society **52** (2023), no. 3, 647–675.

Appendix

Table 3: Simulation results.

		MLE			BM		
p	n	Bias	SD	MSE	Bias	SD	MSE
0.1							
	25	0.0004	0.0613	0.0037	0.0300	0.0567	0.0041
	50	0.0006	0.0423	0.0017	0.0159	0.0406	0.0019
	100	< 0.0001	0.0296	0.0008	0.0077	0.0290	0.0009
	200	0.0005	0.0213	0.0004	0.0044	0.0211	0.0004
	500	< 0.0001	0.0132	0.0001	0.0016	0.0132	0.0001
	1000	< 0.0001	0.0095	< 0.0001	0.0007	0.0095	< 0.0001
	2000	< 0.0001	0.0067	< 0.0001	0.0004	0.0067	< 0.0001
	5000	< 0.0001	0.0041	< 0.0001	0.0001	0.0041	< 0.0001
0.2							
	25	0.0010	0.0796	0.0063	0.0231	0.0737	0.0059
	50	0.0012	0.0570	0.0032	0.0126	0.0548	0.0031
	100	< 0.0001	0.0395	0.0015	0.0058	0.0388	0.0015
	200	0.0012	0.0283	0.0008	0.0041	0.0280	0.0008
	500	-0.0001	0.0177	0.0003	0.0010	0.0177	0.0003
	1000	-0.0002	0.0123	0.0001	0.0003	0.0123	0.0001
	2000	0.0001	0.0090	< 0.0001	0.0004	0.0090	< 0.0001
	5000	0.0001	0.0055	< 0.0001	0.0003	0.0055	< 0.0001
0.3							
	25	0.0041	0.0907	0.0082	0.0186	0.0840	0.0074
	50	0.0023	0.0652	0.0042	0.0099	0.0627	0.0040
	100	0.0006	0.0459	0.0021	0.0045	0.0450	0.0020
	200	0.0013	0.0328	0.0010	0.0033	0.0325	0.0010
	500	0.0001	0.0206	0.0004	0.0009	0.0206	0.0004
	1000	-0.0001	0.0144	0.0002	0.0002	0.0143	0.0002
	2000	< 0.0001	0.0104	0.0001	0.0002	0.0104	0.0001
	5000	< 0.0001	0.0064	< 0.0001	0.0001	0.0064	< 0.0001
0.4							
	25	0.0048	0.0965	0.0093	0.0118	0.0894	0.0081
	50	0.0027	0.0691	0.0047	0.0064	0.0664	0.0044
	100	0.0016	0.0493	0.0024	0.0036	0.0483	0.0023
	200	0.0005	0.0355	0.0012	0.0015	0.0351	0.0012
	500	< 0.0001	0.0222	0.0004	0.0003	0.0221	0.0004
	1000	-0.0003	0.0155	0.0002	-0.0001	0.0155	0.0002
	2000	< 0.0001	0.0111	0.0001	< 0.0001	0.0111	0.0001
	5000	< 0.0001	0.0070	< 0.0001	< 0.0001	0.0070	< 0.0001
0.5							
	25	0.0031	0.0993	0.0098	0.0029	0.0919	0.0084
	50	0.0019	0.0704	0.0049	0.0018	0.0676	0.0045
	100	0.0018	0.0501	0.0025	0.0018	0.0491	0.0024
	200	0.0004	0.0353	0.0012	0.0004	0.0349	0.0012
	500	-0.0001	0.0223	0.0005	-0.0001	0.0222	0.0004
	1000	-0.0002	0.0154	0.0002	-0.0002	0.0154	0.0002
	2000	0.0001	0.0115	0.0001	0.0001	0.0115	0.0001
	5000	0.0001	0.0071	< 0.0001	0.0001	0.0071	< 0.0001
0.6							
	25	-0.0048	0.0965	0.0093	-0.0118	0.0894	0.0081
	50	-0.0027	0.0691	0.0047	-0.0064	0.0664	0.0044
	100	-0.0016	0.0493	0.0024	-0.0036	0.0483	0.0023
	200	-0.0005	0.0355	0.0012	-0.0015	0.0351	0.0012
	500	< 0.0001	0.0222	0.0004	-0.0003	0.0221	0.0004
	1000	0.0003	0.0155	0.0002	0.0001	0.0155	0.0002
	2000	< 0.0001	0.0111	0.0001	< 0.0001	0.0111	0.0001
	5000	< 0.0001	0.0070	< 0.0001	< 0.0001	0.0070	< 0.0001
0.7							
	25	-0.0041	0.0907	0.0082	-0.0186	0.0840	0.0074
	50	-0.0023	0.0652	0.0042	-0.0099	0.0627	0.0040
	100	-0.0006	0.0459	0.0021	-0.0045	0.0450	0.0020
	200	-0.0013	0.0328	0.0010	-0.0033	0.0325	0.0010
	500	-0.0001	0.0206	0.0004	-0.0009	0.0206	0.0004
	1000	0.0001	0.0144	0.0002	-0.0002	0.0143	0.0002
	2000	< 0.0001	0.0104	0.0001	-0.0002	0.0104	0.0001
	5000	< 0.0001	0.0064	< 0.0001	-0.0001	0.0064	< 0.0001
0.8							
	25	-0.0010	0.0796	0.0063	-0.0231	0.0737	0.0059
	50	-0.0012	0.0570	0.0032	-0.0126	0.0548	0.0031
	100	< 0.0001	0.0395	0.0015	-0.0058	0.0388	0.0015
	200	-0.0012	0.0283	0.0008	-0.0041	0.0280	0.0008
	500	0.0001	0.0177	0.0003	-0.0010	0.0177	0.0003
	1000	0.0002	0.0123	0.0001	-0.0003	0.0123	0.0001
	2000	-0.0001	0.0090	< 0.0001	-0.0004	0.0090	< 0.0001
	5000	-0.0001	0.0055	< 0.0001	-0.0003	0.0055	< 0.0001
0.9							
	25	-0.0004	0.0613	0.0037	-0.0300	0.0567	0.0041
	50	-0.0006	0.0423	0.0017	-0.0159	0.0406	0.0019
	100	< 0.0001	0.0296	0.0008	-0.0077	0.0290	0.0009
	200	-0.0005	0.0213	0.0004	-0.0044	0.0211	0.0004
	500	< 0.0001	0.0132	0.0001	-0.0016	0.0132	0.0001
	1000	< 0.0001	0.0095	< 0.0001	-0.0007	0.0095	< 0.0001
	2000	< 0.0001	0.0067	< 0.0001	-0.0004	0.0067	< 0.0001
	5000	< 0.0001	0.0041	< 0.0001	-0.0001	0.0041	< 0.0001