Statistical Inference in a Cocaine Drug Bust

DATA 602, Winter 2025

David Fakolujo, Akinyemi Apampa, Ravin Jayasuriya, Prince Oloma and Joshua Ogunbo

Background

A cocaine bust in Calgary yielded $\mathbf{N = 496}$ suspected cocaine plastic packets. To convict the suspected drug traffickers, the **Alberta Crown Prosecution Service (ACPS)** and the **Calgary Police Service (CPS)** had to prove that there was **genuine cocaine** in (at least one of) the packets.

Apparently, drug traffickers have been **mixing "clean" packets** (i.e., packets that are negative for cocaine, often containing corn starch) with **"dirty" packets** (i.e., packets that are positive for cocaine) to confound the police.

Due to **budget limitations** or a **lack of resources**, law enforcement is often restricted to testing **a smaller sample** of the total shipment. This raises the question:

> *How can statistical inference be used to estimate the number of contaminated packets with minimal testing?*

By analyzing this problem using **statistical methods**, we aim to demonstrate how **statistical inference can be utilized in law enforcement** to assist in decision-making.

The goal of this project is to develop a **statistical inference method** to determine the **total number of contaminated packets** within a cocaine shipment. This estimation is crucial for **legal proceedings**, as it determines whether there is **sufficient evidence** for conviction.

Methods

To estimate θ, the total number of contaminated packets, we will use the following statistical methods:

(1) **Hypergeometric Distribution:** Models the probability of selecting **dirty packets** in a **limited sample** without replacement.

$$P_\theta(X_1 = x_1) = \frac{\binom{\theta}{x_1}\binom{N_1-\theta}{n_1-x_1}}{\binom{N_1}{n_1}}$$

where:

- N_1 - The total number of packets,
- n_1 - The sample size,
- θ is the unknown number of dirty packets,

- x_1 represents the observed number of dirty packets,
- $N_1 - \theta$ is the number of clean packets in the population,
- $n_1 - x_1$ is the number of clean packets from the sample.

(2) **Maximum Likelihood Estimation (MLE):** This was used to determine the total number of dirty packets,$hat\theta$, in the parameter space containing all the possible values of θ, which has the highest probability in the hypergeometric distribution and negative hypergeometric distribution

$$\hat{\theta}_1 = \underset{\theta \in \Theta(x_1)}{\arg\max} P_\theta(X_1 = x_1).$$

(3) **Monte Carlo Simulation**: Monte Carlo simulation was used to evaluate and compare the performance of the two estimation approaches by repeatedly generating synthetic data and applying the estimation methods. The process involved the following steps:

- **Hypergeometric Distribution:**
  - Generated N random samples of dirty packets based on the observed sample size.
  - Applied the **hypergeometric function** to each sample to estimate the **maximum likelihood estimate (MLE)** of the total number of contaminated packets (θ) across various sample sizes.
- **Negative Hypergeometric Distribution:**
  - Similarly, generated N random samples of clean packets picked before picking a dirty packet under a **negative hypergeometric** framework.
  - Applied the **negative hypergeometric function** to estimate θ for each generated sample using MLE.

These simulations provided insights into the **bias, variance, and accuracy** of the estimation methods, ensuring that the chosen approach performs well under repeated sampling conditions.

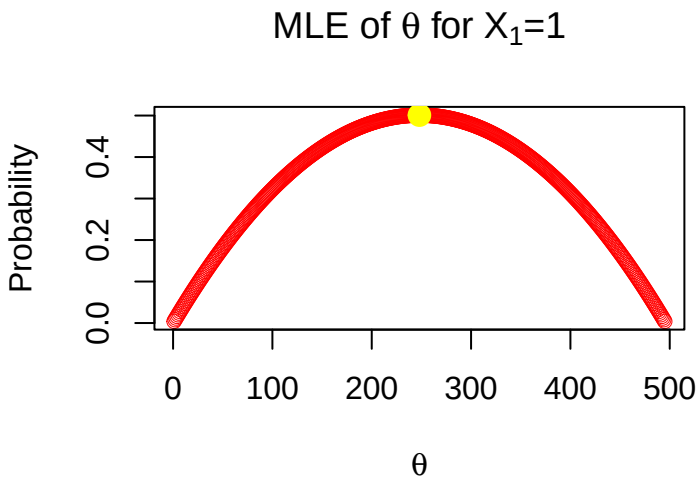
Analysis and Results

Scenario 1

Our analysis involves taking a fixed sample of size $n = 2$ from a population of $N = 496$. We then evaluate different cases where the number of dirty packets in the sample, x_1, can take values from $\{0, 1, 2\}$.

Using Maximum Likelihood Estimation (MLE), we compute the likelihood function for all values of θ, and identify the value of θ that maximizes the likelihood.

Case 1: Estimating θ When $X_1 = 1$




```
## [1] "The maximum likelihood estimate for theta is: 248"
```

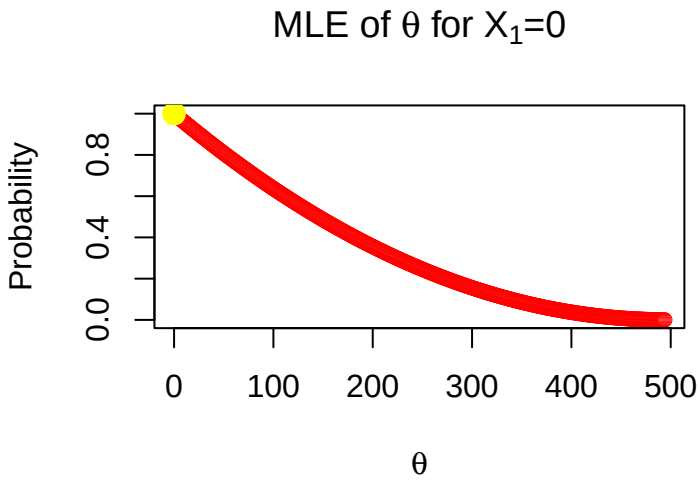
```
## [1] "The maximum probability is: 0.501010101010101"
```

From the output, the **Maximum Likelihood Estimate (MLE)** indicates that the most likely number of dirty packets in the population is $\theta = 248$.

This means that, given a sample of **2 packets**, where **1 was found to be dirty**, the best estimate for the total number of dirty packets in the entire population is **248**.

Additionally, the **likelihood** of observing exactly **1 dirty packet** in the sample, assuming that $\theta = 248$, is approximately **50.1%**.

Case 2: Estimating θ When $X_1 = 0$



```
## [1] "The maximum likelihood estimate for theta is: 0"
```

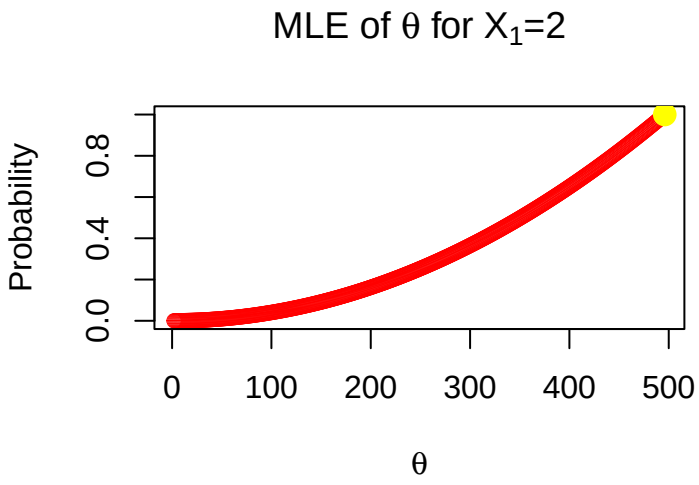
```
## [1] "The maximum probability is: 1"
```

When no dirty packets ($x_1 = 0$) are observed in the sample, the **Maximum Likelihood Estimate (MLE)** suggests that the most likely value of θ (the total number of dirty packets in the population) is **0**.

This conclusion makes intuitive sense if none of the selected packets tested positive for cocaine, the best estimate is that there are **no dirty packets** in the entire population. In other words, based on the sample data, it is most probable that all packets in the population are clean.

Mathematically, this means that the **likelihood function** reaches its highest value when $\theta = 0$, reinforcing the idea that the absence of dirty packets in the sample strongly suggests their absence in the entire population.

Case 3: Estimating θ When $X_1 = 2$




```
## [1] "The maximum likelihood estimate for theta is: 496"
```

```
## [1] "The maximum probability is: 1"
```

When two dirty packets ($x_1 = 2$) are observed in the sample, the **Maximum Likelihood Estimate (MLE)** suggests that the most likely value of θ (the total number of dirty packets in the population) is **496**.

This conclusion makes intuitive sense if two of the selected packets tested positive for cocaine, the best estimate is that there are **496 dirty packets** in the entire population. In other words, based on the sample data, it is most probable that all packets in the population are dirty

Mathematically, this means that the **likelihood function** reaches its highest value when $\theta = 496$, reinforcing the idea that the presence of dirty packets in the sample strongly suggests their prsence in the entire population.

Scenario 2

We analyzed the joint sampling distribution of two dependent discrete random variables, X_1 and X_2, representing the number of dirty packets found in two successive rounds of random sampling without replacement. Initially, $n_1 = 2$ packets were tested from a total of $N_1 = 496$, and both were found clean. Subsequently, the investigators selected $n_2 = 4$ additional packets from the remaining $N_2 = 494$ packets, hoping to detect at least one dirty packet to support their case.

After the initial test of $n_1 = 2$ packets resulted in $X_1 = 0$ dirty packets, additional $n_2 = 4$ packets were tested to gather sufficient evidence, and 2 packets were found to be dirty. Let X_2 represent the number of dirty packets in this second sample.

Joint Probability of X_1 and X_2

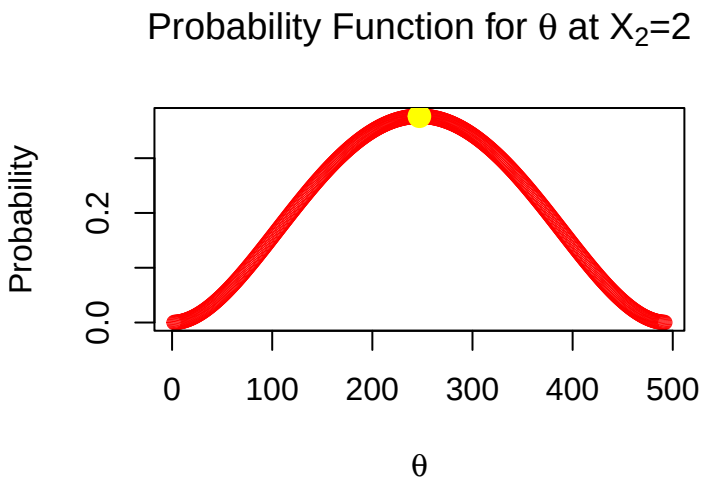
The joint probability function is given by:

$$P_\theta(X_1 = x_1, X_2 = x_2) = P_\theta(X_2 = x_2 | X_1 = x_1) P_\theta(X_1 = x_1)$$

Given that:

- $X_1 \sim \text{Hypergeometric}(N_1 = 496, \theta, n_1 = 2)$,
- $X_2 | X_1 = x_1 \sim \text{Hypergeometric}(N_2 = 496 - 2, \theta - x_1, n_2 = 4)$,

we define the possible values of θ.




```
## [1] "The Maximum Likelihood Estimate for theta is: 247"
```

```
## [1] "The maximum Probability is: 0.376525945724873"
```

```
## [1] "The smallest possible value of theta for the joint sampling distribution is: 2"
```

```
## [1] "The largest possible value of theta for the joint sampling distribution is: 492"
```

Scenario 3

Given that $x_2 = 2$ out of the additional $n_2 = 4$ sampled packets were found to be dirty, the conviction of the accused is now certain. However, the total number of dirty packets, θ, remains unknown and needs to be estimated.

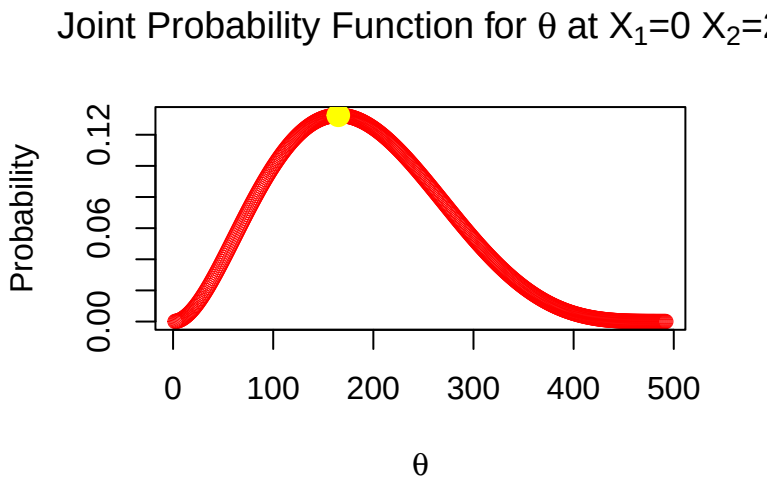
Maximum Likelihood Estimation (MLE)

We determine θ, the **maximum likelihood estimate (MLE)** by maximizing the joint probability $P_\theta(X_1 = x_1, X_2 = x_2)$ as a function of θ over its feasible parameter space, $\Theta(x_1, x_2)$, which depends on the observed values x_1 and x_2. Since both $X_1(\theta)$ and $X_2(\theta)$ are functions of θ, the possible values of θ must be considered. ## Finding the MLE We obtain The MLE, denoted as $\hat{\theta}_2$, by evaluating $P_\theta(X_1 = 0, X_2 = 2)$ for all possible values of θ in $\Theta(0, 2)$ and selecting the value that maximizes this probability. This approach ensures that the estimated θ is the most likely given the observed data.

With $x_2 = 2$ dirty packets found in the second test, the conviction is certain, but the actual number of dirty packets remains unknown. To estimate θ, we maximize the joint probability function:

$$\hat{\theta}_2 = \underset{\theta \in \Theta(0,2)}{\arg\max} P_\theta(X_1 = 0, X_2 = 2)$$

The new parameter space $\Theta(0, 2)$ depends on the observed values of X_1 and X_2.




```
## [1] "The Maximum Likelihood Estimate for theta (joint) is: 165"
```

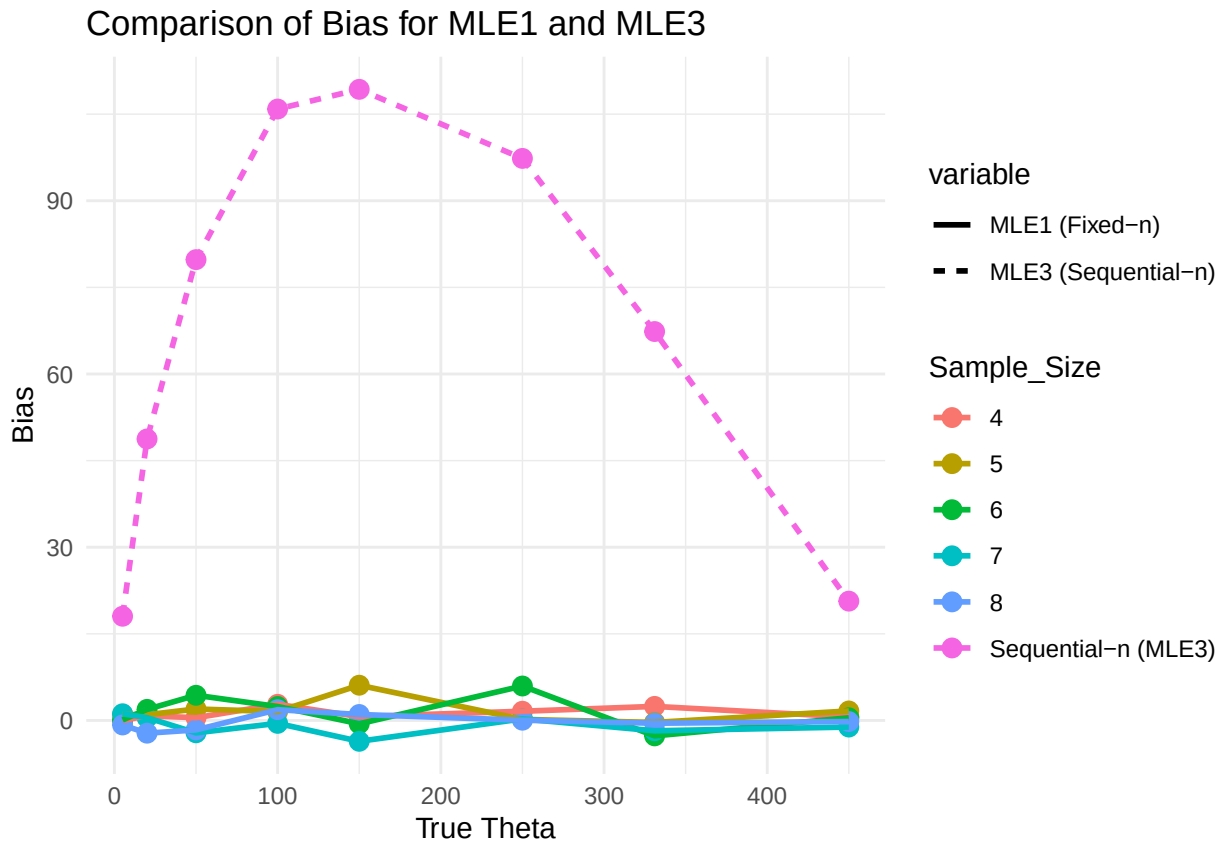
```
## [1] "The maximum Probability for joint distribution is: 0.1324900852217"
```

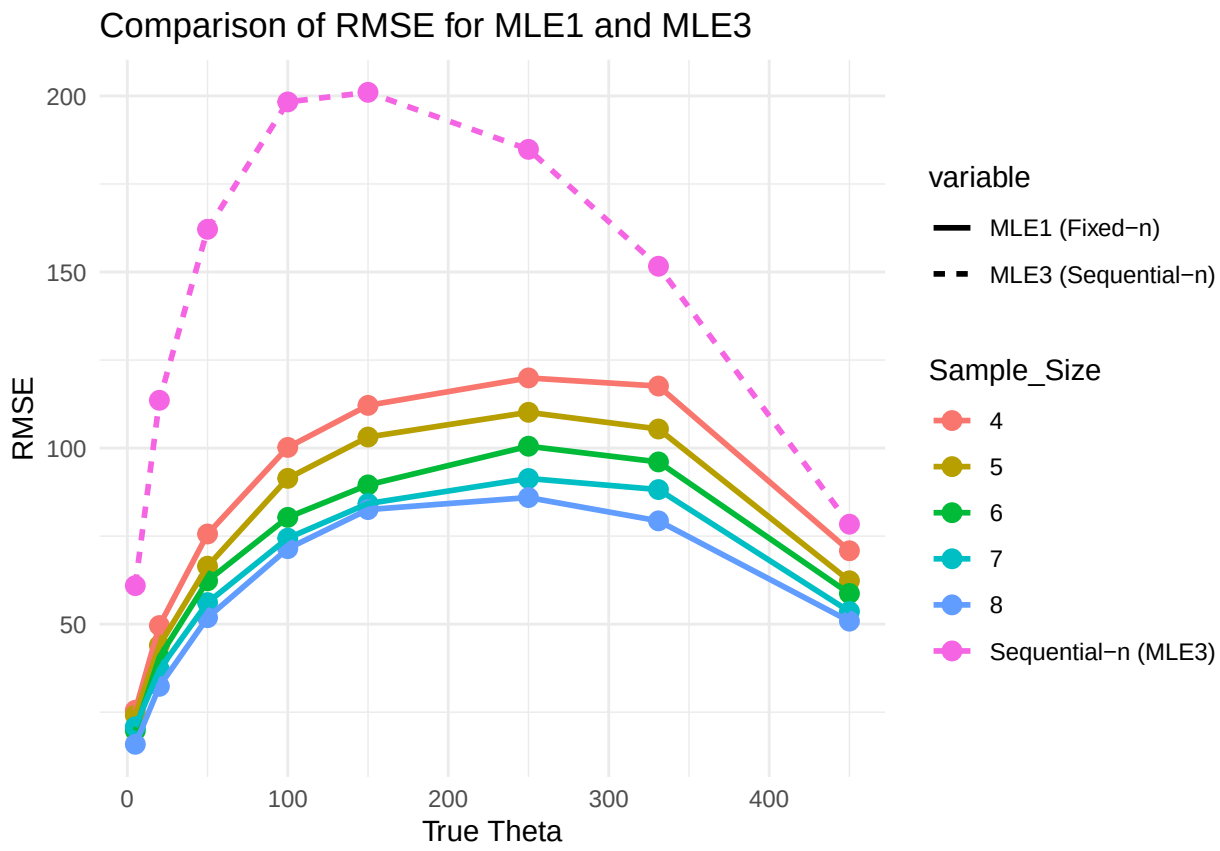
Scenario 4

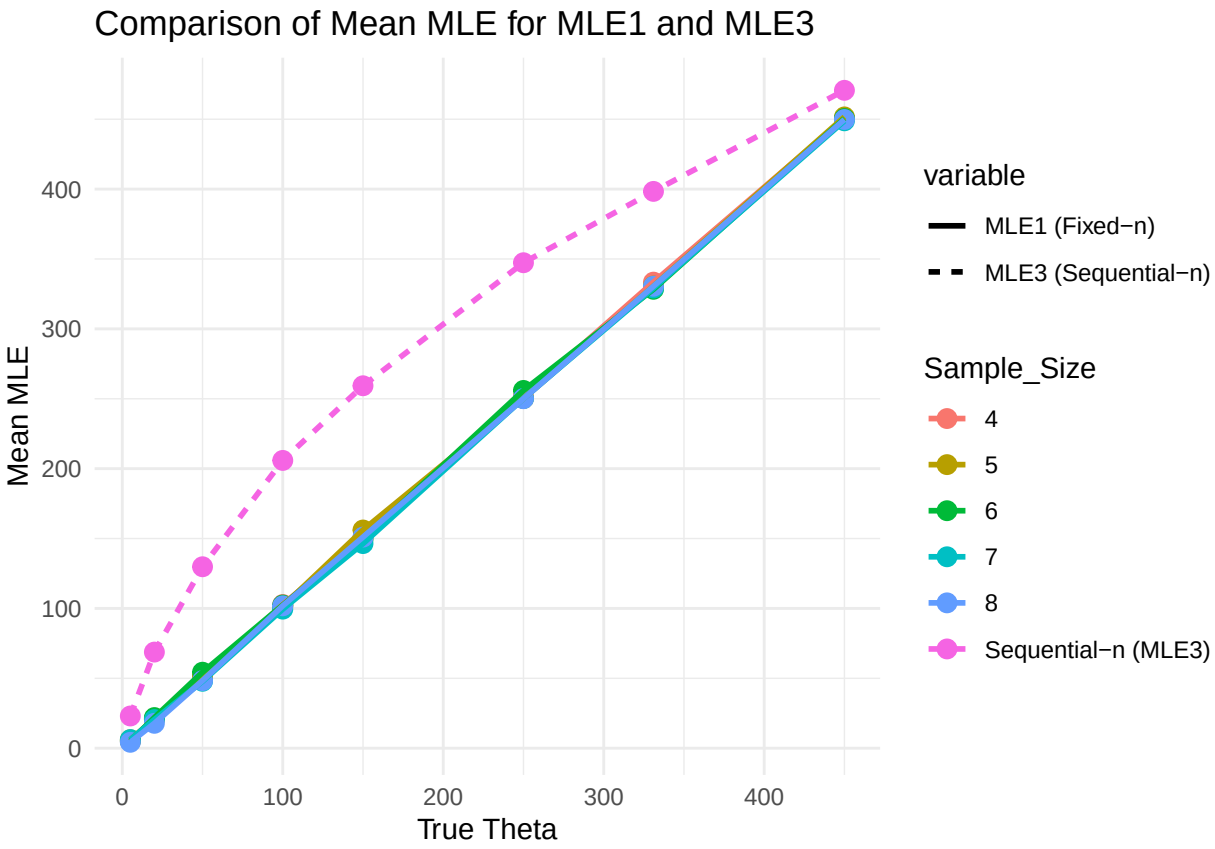
After analyzing the packets, where some packets were identified as "dirty," we estimated the total number of dirty packets (θ) in a set of 496 packets. Two different approaches can be used for this estimation:

1. **Fixed Sample Method**: Drawing a fixed number of packets at random and analyzing them.
2. **Sequential Sampling Method**: Drawing packets one by one until the first dirty packet is found.

Using statistical inference techniques, we compared these two estimation methods by evaluating their **bias** and **root mean squared error (RMSE)** through a **Monte Carlo simulation**.




Comparison of RMSE for MLE1 and MLE3
RMSE
200
150
100
50
0
100
200
300
400
True Theta
variable
MLE1 (Fixed−n)
MLE3 (Sequential−n)
Sample_Size
4
5
6
7
8
Sequential−n (MLE3)



Comparison of MLE1 and MLE3

In this study, we compared two different Maximum Likelihood Estimators (MLEs) for θ at different values of 5,20,50,100,150,250,331, and 450:
1. **MLE1 (Fixed-n Estimate)**, where the sample size n is pre-determined.
2. **MLE3 (Sequential-n Estimate)**, where sampling continues until at least one dirty packet is found. The average samples of clean packets picked before picking a dirty packet using different θ values were 81,23,9,5,3,2,2, and 1.

Comparison Metrics

- **Bias:** Measures the accuracy of the estimator. A lower bias means the estimator is closer to the true θ.
- **RMSE (Root Mean Square Error):** Accounts for both accuracy and variability. A lower RMSE means the estimator is more stable and reliable.

Conclusion from Our Analysis

Bias

- **MLE1 remains unbiased** across all values of θ, meaning it provides an accurate estimate of the true number of dirty packets.

- **MLE3 has a low bias when θ is small**, but **starts to overestimate significantly** as θ increases, before gradually reducing again.

RMSE

- **MLE1 has a lower and more stable RMSE**, meaning it consistently produces precise estimates.
- **MLE3 has a low RMSE when θ is small**, but **its RMSE increases as θ increases**, indicating higher variability and inconsistency, before gradually reducing again.

Key Findings

- **MLE3 performs well when θ is low**, as it **efficiently detects contamination without excessive sampling**. This makes it useful in **resource-constrained settings** where testing capacity is limited.
- **MLE3 becomes unreliable when θ is high**, as it **overestimates the number of contaminated packets**, leading to potential **false conclusions in forensic settings**.
- **MLE1 remains stable across all values of θ**, providing **consistent and dependable estimates**, making it a **preferred estimator when precision is critical**.

In a forensic or legal setting, where **accurate estimation of contamination levels is crucial**, a judge would likely prefer **MLE1** due to its **stability and lack of overestimation bias**. While **MLE3 can be useful for early detection** when contamination is rare, its tendency to **overestimate high contamination cases makes it unreliable for critical decision-making**.

Thus, **MLE1 is the recommended estimator for legal investigations** to ensure accurate and fair conclusions.

Group Members' Contributions

- **Akin** - Answered part E (S4). Prepared results section of the report.
- **David** - Answered question C and D. Prepared methods section of the report.
- **Joshua** - Answered part E (S3). Prepared results section of the report.
- **Prince** - Answered part E (S2). Prepared conclusion section of the report.
- **Ravin** - Answered A and B of the problem. Prepared background section of the report.

Reference

Shuster, J. J. (1991). The statistician in a reverse cocaine sting. The American Statistician, 45(2), 123–124