Tutorial 5: Introduction to Data Science - Precision of SRS Estimator

Exercise 1: Mark and Recapture Sampling (Equal Probabilities)

Task (a): Expected Value of the Estimator $\hat{N}^{-1}$

In this task, we are asked to find the **expected value** of the inverse of the estimator $\hat{N}^{-1}$ as a function of N.

Let's calculate the expected value of $\hat{N}^{-1}$.

Given that:

$$\hat{N}^{-1} = \frac{X}{n \cdot M}$$

The expected value of $\hat{N}^{-1}$ is:

$$E(\hat{N}^{-1}) = E\left(\frac{X}{n \cdot M}\right)$$

Since $X \sim \text{Binomial}(n, p = \frac{M}{N})$, we know that the expected value of X is:

$$E(X) = n \cdot p = n \cdot \frac{M}{N}$$

Thus, the expected value of $\hat{N}^{-1}$ is:

$$E(\hat{N}^{-1}) = \frac{n \cdot \frac{M}{N}}{n \cdot M} = \frac{1}{N}$$

Therefore, the expected value of $\hat{N}^{-1}$ is:

$$E(\hat{N}^{-1}) = \frac{1}{N}$$

Task (b): Variance and Standard Deviation of the Estimator $\hat{N}^{-1}$

For this task, we are asked to calculate the **variance** and **standard deviation** of the estimator $\hat{N}^{-1}$ as a function of N.

Recall that:

$$\hat{N}^{-1} = \frac{X}{n \cdot M}$$

We want to calculate $\text{Var}(\hat{N}^{-1})$.

Since $X \sim \text{Binomial}(n, p = \frac{M}{N})$, the variance of X is:

$$\text{Var}(X) = n \cdot p \cdot (1 - p) = n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right)$$

Now, the variance of $\hat{N}^{-1}$ is:

$$\text{Var}(\hat{N}^{-1}) = \frac{\text{Var}(X)}{(n \cdot M)^2} = \frac{n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right)}{(n \cdot M)^2}$$

Simplifying this expression:

$$\text{Var}(\hat{N}^{-1}) = \frac{\frac{M}{N} \cdot \left(1 - \frac{M}{N}\right)}{n \cdot M^2}$$

Thus, the **variance** of $\hat{N}^{-1}$ is:

$$\text{Var}(\hat{N}^{-1}) = \frac{\frac{M}{N} \cdot \left(1 - \frac{M}{N}\right)}{n \cdot M^2}$$

To calculate the **standard deviation**, we take the square root of the variance:

$$\text{SD}(\hat{N}^{-1}) = \sqrt{\frac{\frac{M}{N} \cdot \left(1 - \frac{M}{N}\right)}{n \cdot M^2}}$$

Therefore, the standard deviation is:

$$\mathrm{SD}(\hat{N}^{-1}) = \sqrt{\frac{\frac{M}{N} \cdot \left(1 - \frac{M}{N}\right)}{n \cdot M^2}}$$

Task (c): Approximate Distribution of $\hat{N}^{-1}$

In this task, we are asked to describe the **approximate distribution** of $\hat{N}^{-1}$ using the **Central Limit Theorem**.

According to the **Central Limit Theorem (CLT)**, if $X \sim \mathrm{Binomial}(n, p)$, for large n, the distribution of X can be approximated by a **normal distribution**:

$$X \sim \mathcal{N}(n \cdot p, n \cdot p \cdot (1 - p))$$

Since $\hat{N}^{-1}$ is a linear transformation of X (i.e., $\hat{N}^{-1} = \frac{X}{n \cdot M}$), the **distribution of** $\hat{N}^{-1}$ is approximately **normal** for large n, with:

- Mean $E(\hat{N}^{-1}) = \frac{1}{N}$
- Variance $\mathrm{Var}(\hat{N}^{-1}) = \frac{\frac{M}{N} \cdot \left(1 - \frac{M}{N}\right)}{n \cdot M^2}$

Thus, the approximate distribution of $\hat{N}^{-1}$ is:

$$\hat{N}^{-1} \sim \mathcal{N}\left(\frac{1}{N}, \frac{\frac{M}{N} \cdot \left(1 - \frac{M}{N}\right)}{n \cdot M^2}\right)$$

Task (d): Standard Deviation of $\hat{N}$ Given $X = 12$

In this task, we are given that **12 marked animals** were recaptured, i.e., $X = 12$. We are asked to calculate the **standard deviation** of the estimator $\hat{N}$, based on this information.

Step 1: Estimate the Population Size $\hat{N}$

We use the formula:

$\hat{N} = \frac{n \cdot M}{X}$

Where:

- $n = 50$ is the sample size.
- $M = 100$ is the number of marked animals released.
- $X = 12$ is the number of marked animals recaptured.

Substitute these values:

$\hat{N} = \frac{50 \cdot 100}{12} = \frac{5000}{12} \approx 416.67$

Step 2: Calculate the Standard Deviation of $\hat{N}^{-1}$

Using the standard deviation formula from **Task (c)**, but replacing N with $\hat{N}$:

$\mathrm{SD}(\hat{N}^{-1}) = \sqrt{\frac{\frac{M}{\hat{N}} \cdot \left(1 - \frac{M}{\hat{N}}\right)}{n \cdot M^2}}$

Substitute $M = 100$, $\hat{N} = 416.67$, and $n = 50$:

$\mathrm{SD}(\hat{N}^{-1}) = \sqrt{\frac{0.24 \cdot 0.76}{50 \cdot 100^2}} = \sqrt{\frac{0.1824}{500000}} = \sqrt{3.648 \times 10^{-7}} \approx 0.000603$

Task (e): 95% Confidence Interval for N^{-1}

In this task, we are asked to calculate a **95% confidence interval** for N^{-1}, using the distribution derived in **Task (d)**.

Step 1: Calculate $\hat{N}^{-1}$

From **Task (d)**, we estimated the population size $\hat{N} = 416.67$. Therefore, the estimate for $\hat{N}^{-1}$ is:

$$\hat{N}^{-1} = \frac{1}{\hat{N}} = \frac{1}{416.67} \approx 0.0024$$

Step 2: Use the Standard Deviation from Task (d)

From **Task (d)**, we know that the standard deviation of $\hat{N}^{-1}$ is approximately:

$$\mathrm{SD}(\hat{N}^{-1}) \approx 0.000603$$

Step 3: Calculate the 95% Confidence Interval

To calculate the 95% confidence interval, we use the formula (you can also see it in the figure at the bottom of the question sheet):

$$\hat{N}^{-1} \pm 1.96 \times \mathrm{SD}(\hat{N}^{-1})$$

Substituting the values:

$$0.0024 \pm 1.96 \times 0.000603$$

This gives:

$$0.0024 \pm 0.00118188$$

Thus, the 95% confidence interval for N^{-1} is:

$$[0.00121812, 0.00358188]$$

Conclusion:

We are 95% confident that the true value of N^{-1} lies within the interval $[0.00121812, 0.00358188]$.

Side Exercise: Do the same but for 90% confidence interval.

Task (f): 95% Confidence Interval for N

In this task, we are asked to invert the confidence interval for N^{-1}, calculated in **Task (e)**, to find a 95% confidence interval for N, and determine whether this interval includes the true population size $N = 500$.

Step 1: Recall the 95% Confidence Interval for N^{-1}

From **Task (e)**, the 95% confidence interval for N^{-1} is:

$$[0.00121812, 0.00358188]$$

Step 2: Invert the Interval to Find the Confidence Interval for N

To find the confidence interval for N, we take the reciprocal of both endpoints of the confidence interval for N^{-1}:

- Lower bound for N:

$$\frac{1}{0.00358188} \approx 279.18$$

- Upper bound for N:

$$\frac{1}{0.00121812} \approx 820.86$$

Thus, the 95% confidence interval for N is:

$$[279.18, 820.86]$$

Step 3: Does the Interval Include 500?

Yes, the interval includes the true population size $N = 500$, because 500 lies between the lower bound 279.18 and the upper bound 820.86.

In [7]:
```
# Task (g)

import numpy as np

np.random.seed(0)

# Number of simulations
n = 1000
```

```python
# Simulating binomial data: x is the number of marked animals recaptured in each sample
x = np.random.binomial(50, 100 / 500, n)

# Estimating N^{-1} (inverse population size estimate)
nm1 = x / 5000      # Question for you: Why have I divided by 5000 here?

# Calculating the standard deviation for N^{-1}
sds = np.sqrt(x / 50 * (1 - x / 50) * 50 / 5000**2)

# Calculating 95% confidence intervals for N
cil = 1 / (nm1 + 1.96 * sds)  # Lower bound of the confidence interval
ciu = 1 / (nm1 - 1.96 * sds)  # Upper bound of the confidence interval

# Proportion of intervals that contain N = 500
proportion_in_interval = np.sum((cil < 500) & (ciu > 500)) / n

# Output the proportion
print(f"Proportion of samples where N = 500 is within the 95% confidence interval: {proportion_in_interval:.3f}")
```

```
Proportion of samples where N = 500 is within the 95% confidence interval: 0.945
```

Task (h): Relative Frequency of $N = 500$ in the Confidence Interval

In this task, we are asked to determine the **relative frequency** with which the confidence intervals calculated in **Task (g)** include the true population size $N = 500$.

Relative Frequency Calculation

From the Python simulation in **Task (g)**, the proportion of confidence intervals that included $N = 500$ was:

$$\text{Proportion} = 0.945$$

This means that **94.5%** of the 95% confidence intervals contained the true population size $N = 500$.

Did We Expect This Relative Frequency?

Yes, since we are calculating **95% confidence intervals**, we expect that approximately **95%** of the intervals should contain the true population size $N = 500$.

The observed relative frequency of **94.5%** is very close to the expected value of **95%**, which is consistent with the performance of a 95% confidence interval. The small deviation from exactly 95% is due to random variation in the sampling process.