April 01, 2024

Ephrata Getachew

16 Barrett Hill Dr

Amherst, MA 01002

Dear Chief Intelligence Officer Professor Wagman,

We are writing to present the results of our analysis regarding the estimation of the number of tanks in the Gold army's tank fleet, as assigned to us by the Purple army's intelligence division. Our investigation involved both theoretical derivations and a simulation study to assess the performance of various estimators under different scenarios.

Theoretical Derivations

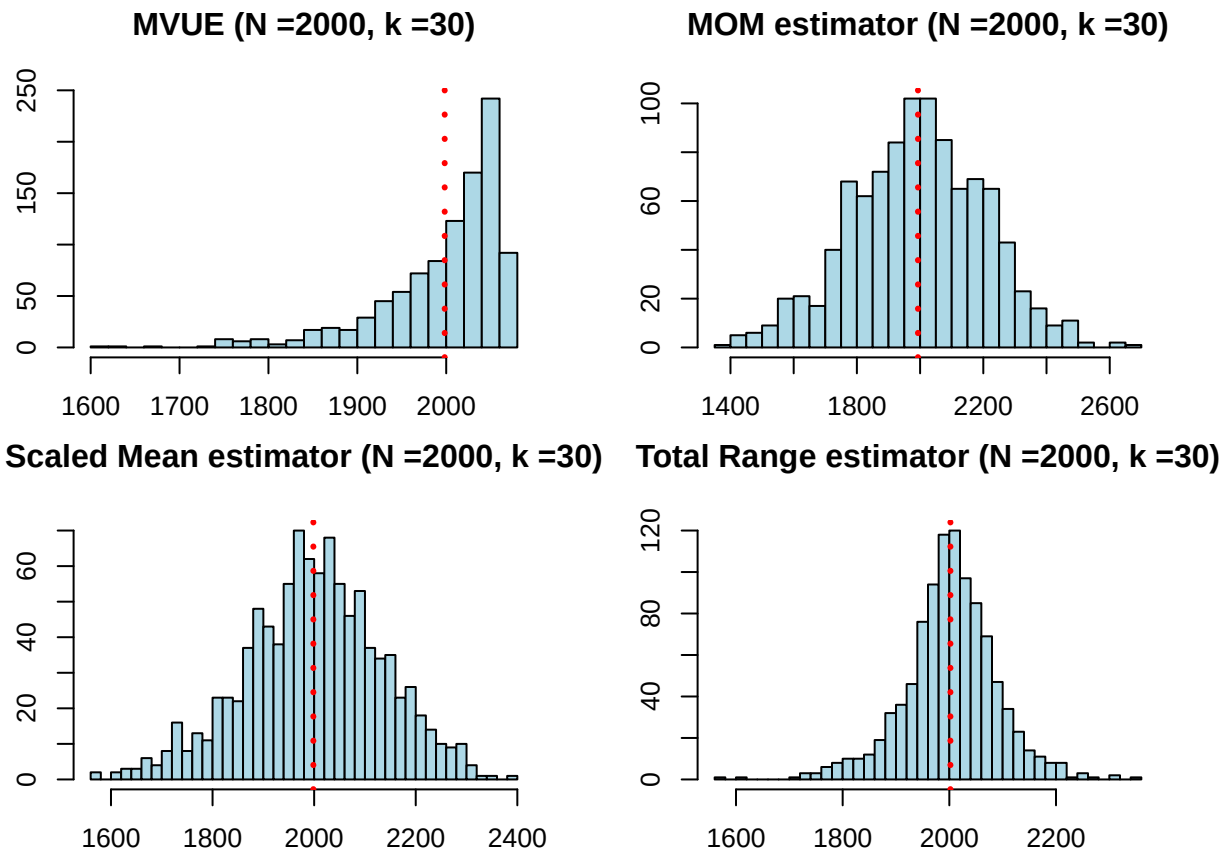
In our theoretical derivations, we explored the method of moments estimator (MoM) and the minimum variance unbiased estimator (MVUE) for estimating the number of tanks in the Gold army's fleet. By analyzing the probability density function (pdf) and cumulative distribution function (cdf) of a single observation from a discrete uniform distribution, we derived expressions for the expected value and variance of the observation. We then derived the MoM estimator and the MVUE for number of tanks. Furthermore, we identified a sufficient statistic for the number of tanks and investigated the unbiasedness and efficiency of the estimators, guiding us in selecting our preferred estimator. The detailed derivations and findings can be found on page 7 of the appendix.

Simulation Study Setup

In our simulation study, we aimed to assess the performance of four different estimators (MVUE, MoM, Scaled Mean, and Total Range estimators) in estimating the number of tanks in the Gold army's fleet under varying conditions of sample size (k) and the number of tanks (N). We conducted simulations across four different combinations of sample sizes (k = 15, 20, 25, 30) and the number of tanks (N = 1000, 1500, 2000, 2500), generating 1000 replications for each combination to calculate the values of the estimators. To maintain reproducibility, we initialized a seed at the beginning of the study. Within each iteration of the loop, we generated estimator values for the MVUE, MoM, Scaled Mean(mean(sample) + 1.74 * standard deviation), and Total Range (maximum(sample) + minimum(sample)) methods. We then calculated the means and standard deviations of these estimators and stored the results in a dataframe. Subsequently, histograms were plotted for each estimator, with the mean estimator value highlighted to provide a reference point. The code is provided in (Appendix Code for Simulation)

Evaluation of Estimators

The histograms below display the distribution of estimates for different values of k and $N = 2000$ using the four estimators. The red dotted line represents the average value, serving as our point estimate. Although simulations were conducted for four different values of N, we only show results for $N = 2000$ for illustrative purposes.



Looking at the histograms, the MoM estimates demonstrate symmetrical distributions centered around the mean estimate. Referring to (Appendix MoM estimator plot) as the sample size (k) increases in our estimation of the number of tanks in the Gold army's fleet, the distribution of the MoM estimate tends to approach normality and bell-shaped. This means that as we gather more sequential serial numbers from the field, our estimation of the fleet size becomes more reliable. This phenomenon is in line with the Central Limit Theorem. Similarly, the Scaled Mean and Total Range estimator is exhibits a distribution somehow close to that of the MoM estimate.

The MVUE estimates display a left skewed distribution. Since it is based on the maximum value observed in the sample, it tends to be influenced by extreme values, pulling the distribution towards the lower end and resulting in left skewness. However, as k values increase, the spread of estimates diminishes, implying better precision in estimation.

Overall, we found that our four estimator N are nearly unbiased as the true values is near the mean = 2000. (Appendix Mean Estimates). As we would expect with a higher k the mean becomes close to the true population size 2000 and the standard deviation (and hence the variance) decreases significantly as we increase the sample size.

However, looking at the Table 1, we observed that MVUE consistently exhibited lower standard deviations compared to MoM, Total Range, and Scaled Mean estimator. The lower standard deviations indicate greater precision in estimating the true value of N, making it more consistent and an accurate estimator for the true value of N. By minimizing the spread of estimates, it offers a more focused and reliable assessment of the number of tanks, which is crucial in military intelligence operations where precision and accuracy are paramount. Thus, in scenarios where minimizing estimation error is critical, the MVUE emerges as the preferred choice due to its superior precision and reliability.

Table 1: Standard Deviation of Estimators for N = 2000

k	MoM_sd	MVUE_sd	Scaled_Mean_sd	Total_Range_sd
15	300.631	121.936	191.989	169.886
20	269.200	92.964	166.992	132.212
25	235.946	80.155	144.623	114.993
30	213.371	62.040	138.551	93.926

Conclusion

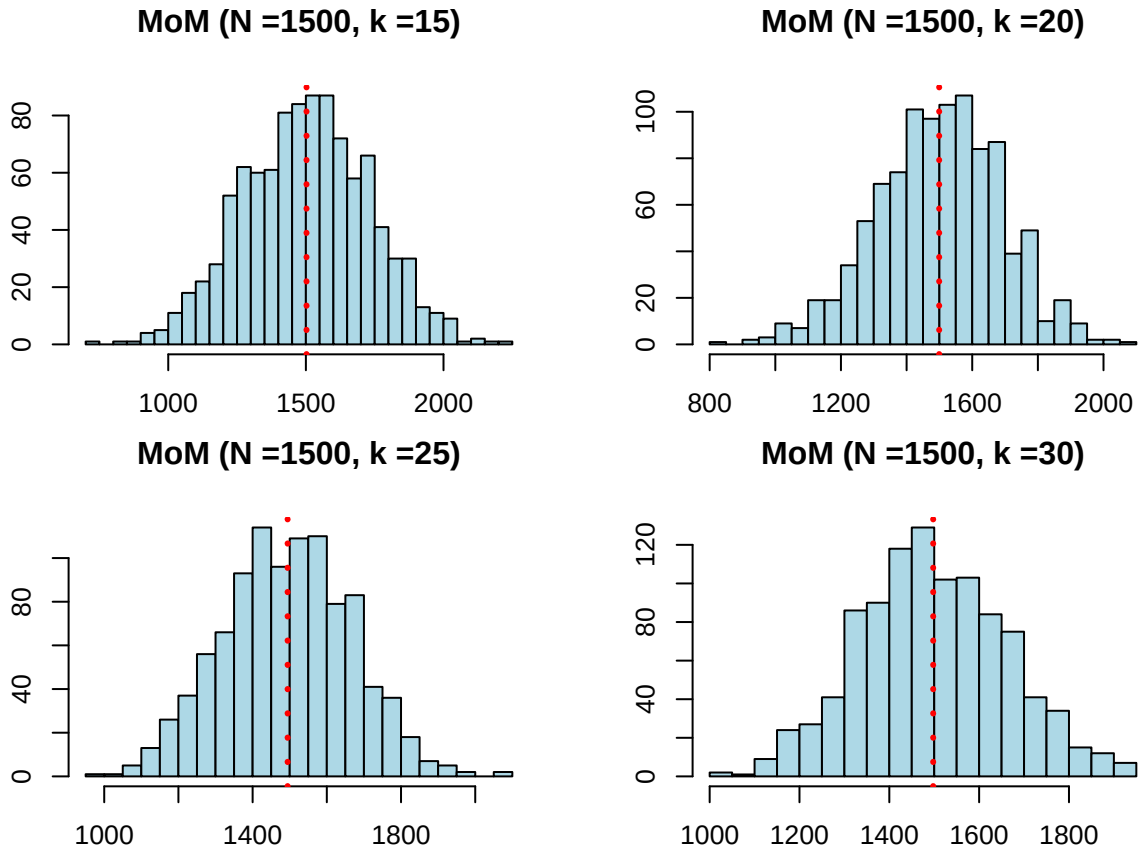
Based on our theoretical derivations and simulation study findings, we propose the MVUE as the preferred estimator for estimating N, the total number of tanks in the Gold army's fleet due to its consistency and accurate estimation(low varaince). Using our preferred MVUE estimator and the provided sample of Gold tank serial numbers, our best estimate for the total number of tanks in the Gold army's fleet is 2190.

Sincerely,

Ephrata Getachew

Apendix

MoM estimator plot (Return To Report)



Code for Simulation Code for simulation (Return To Report)

```
set.seed(231)

# Parameters
N <- c(1000, 1500, 2000, 2500)
k <- c(15, 20, 25, 30)
reps <- 1000

par(mfrow = c(2, 2), mar = c(4, 4, 2, 2))

# Create a dataframe to store results
results <- data.frame(N = numeric(), k = numeric(), mean = numeric(),
                      sd = numeric())

# Initialize vectors with 0s
mvue_estimates <- rep(0, reps)
mom_estimates <- rep(0, reps)
scaled_mean_estimates <- rep(0, reps)
range_estimates <- rep(0, reps)
```

```r
# Loop through each combination of N and k
for (n in N) {
  for (k_val in k) {
    # Generate estimator values for each
    for (i in 1:reps) {
      sample <- runif(k_val, 0, n)
      mvue_estimates[i] <- max(sample) * ((k_val + 1) / k_val)
      mom_estimates[i] <- 2 * mean(sample) - 1
      scaled_mean_estimates[i] <- mean(sample) + 1.74*sd(sample)
      range_estimates[i] <- max(sample) + min(sample)
    }

     # Calculate mean
    mvue_estimates <- mean(mle_estimates)
    mom_mean <- mean(mom_estimates)
    scaled_mean <- mean(scaled_mean_estimates)
    range_mean <- mean(range_estimates)

    # Calculate standard deviations
    mvue_estimates <- sd(mle_estimates)
    mom_sd <- sd(mom_estimates)
    scaled_mean_sd <- sd(scaled_mean_estimates)
    range_sd <- sd(range_estimates)

    # Add results to the dataframe
    results <- rbind(results, data.frame(N = n, k = k_val,
                                         method = c("MVUE", "MoM",
                                                    "Scaled Mean", "Total Range"),
                                         sd = c(mvue_estimates, mom_sd, scaled_mean_sd,
                                                range_sd)))

    # Plot histogram for each estimator
    hist(mvue_estimates, breaks = 50, main = paste("MVUE (N =", n, ", k = ",
                                                   k_val, ")", sep = ""),
         xlab = "Estimator Value", col = "lightblue")
    abline(v = mean(mvue_estimates), col = "red", lwd = 2, lty = "dotted")

    hist(mom_estimates, breaks = 50, main = paste("MoM (N =", n, ", k = ",
                                                  k_val, ")", sep = ""),
         xlab = "Estimator Value", col = "lightblue")
    abline(v = mean(mom_estimates), col = "red", lwd = 2, lty = "dotted")

    hist(scaled_mean_estimates, breaks = 50, main = paste("Scaled Mean (
                                      N =", n, ", k = ", k_val, ")", sep = ""),
         xlab = "Estimator Value", col = "lightblue")
    abline(v = mean(scaled_mean_estimates), col = "red", lwd = 2,
                                                    lty = "dotted")
    hist(range_estimates, breaks = 50, main = paste("Total Range (N =", n, ", k = ",
                                                    k_val, ")", sep = ""),
         xlab = "Estimator Value", col = "lightblue")
    abline(v = mean(range_estimates), col = "red", lwd = 2, lty = "dotted")
  }
}
```

Code for table

```
results_n2000 <- results %>% filter(N == 2000)
table_n2000 <- results_n2000 %>%
  select(k, method, sd) %>%
  spread(key = method, value = sd) %>%
  rename(MVUE_sd = "MVUE", MoM_sd = "MoM", Scaled_Mean_sd = "Scaled Mean",
         Range_sd = "Total Range") %>%
  mutate_at(vars(MVUE_sd, MoM_sd, Scaled_Mean_sd, Range_sd), function(x)
    format(round(x, 3), nsmall = 3))

kable(table_n2000, caption = "Standard Deviation of Estimators for N = 2000")
```

Mean Estimates Mean Estimates (Return To Report)

```
results_mean_n2000 <- results_mean %>% filter(N == 2000)

table_mean_n2000 <- results_mean_n2000 %>%
  select(k, method, mean) %>%
  spread(key = method, value = mean) %>%
  rename(MVUE_mean = "MVUE", MoM_mean = "MoM",
         Scaled_Mean_mean = "Scaled Mean", Total_Range_mean = "Total Range") %>%
  mutate_at(vars(MVUE_mean, MoM_mean, Scaled_Mean_mean, Total_Range_mean),
            function(x) format(round(x, 3), nsmall = 3))

kable(table_mean_n2000, caption = "Mean Estimates of Estimators for N = 2000")
```

Table 2: Mean Estimates of Estimators for N = 2000

k	MoM_mean	MVUE_mean	Scaled_Mean_mean	Total_Range_mean
15	1991.211	1999.402	2001.348	2003.303
20	1995.987	1997.497	2001.081	2003.991
25	2003.341	1996.135	2002.647	1998.625
30	1998.784	1999.018	1994.056	1997.806

Theoretical Derivation

We started by deriving the PDF & CDF which represent a single observation from a discrete uniform distribution on 1 to N, denoted as $f_X(x)$ & $F_X(x)$ respectively, we can utilize these functions to understand the distribution of individual tank serial numbers within the given range

$$f_X(x) = P(X=x) = \begin{cases} \frac{1}{N} & 1 \le X \le N \\ 0 & \text{otherwise} \end{cases}$$

$$F_X(x) = \sum_{i=1}^{N} P(X=i) = \begin{cases} 0 & X < 1 \\ \frac{X}{N} & 1 \le X \le N \\ 1 & X > N \end{cases}$$

expected value

Having found $\hat{E}(X) = \frac{N+1}{2}$ & $\text{Var}(X) = \frac{N^2-1}{12}$ we gain insights into the central tendency and variability of individual tank serial numbers.

Building upon this, we explored different estimation approaches to infer the total number of tanks N based on a sample of k tank serial numbers. One approach is the Method of Moment (MoM) estimator

$$\mu_1' = E(X)$$

$$\mu_1 = \frac{1+N}{2}$$

$$m = \frac{1}{k} \sum_{i=1}^{k} x_i$$

$$\frac{1}{k} \sum_{i=1}^{n} x_i = \frac{1+N}{2}$$

$2\bar{x} - 1 = N \Rightarrow \boxed{\hat{N}_{mom} = 2\bar{x} - 1}$

$$E(\hat{N}) = E(2\bar{x} - 1) = 2E(\bar{x}) - 1 = 2\left(\frac{1+N}{2}\right) - 1 = \frac{N}{2}$$

Thus the estimator is unbiased

$$Var(\hat{N}) = Var(2\bar{x} - 1) = 4\,Var(\bar{x}) = 4\left(\frac{N^2 - 1}{12k}\right) = \frac{N^2 - 1}{3k}$$

Another estimator I considered is Maximum Likelihood estimator (MLE),

$$L(N) = L(x_1, x_2, \ldots x_k | N) = \prod_{i=1}^{k} \frac{1}{N} = \frac{1}{N^k}$$

$$= \begin{cases} \frac{1}{N^k} & \text{for } N \geq \max(x_1, x_2, \ldots, x_k) \\ 0 & \text{otherwise} \end{cases}$$

N must be at least as large as the maximum observed serial number ($\max(x_1, x_2, \ldots x_k)$), i.e. the MLE of N is $\hat{N}_{MLE} = \max(x_1, x_2, \ldots, x_k)$

Now $f_n(x|N) = \frac{1}{N^n}$, $x \in (0, N)^n$

let $q[\max(x), N] = 1$, if $\max \leq N$, 0 otherwise

$$f_n(x|N) = \frac{1}{N^n}\, q[\max(x), N] \times 1$$

$$g(t, N) = \frac{1}{N^n}\, q[\max(x), N] \text{ and } h(x) = 1$$

Now that we have identified the maximum of the sample, $X_{(k)}$, as a sufficient statistic for N,

we can use this information to streamline the estimation process.

Now that we explored the distribution of a single observation X from discrete uniform distribution, let's consider the maximum of of our sample as an estimate for N.

$$F_{X(k)}(x) = P(X_{(k)} \leq x) = \left(\frac{x}{N}\right)^k$$

$$P(X_{(k)} \leq x) = \begin{cases} 0 & \text{if } x < 1 \\ \left(\frac{x}{N}\right)^k & \text{if } 1 \leq x \leq N \\ 1 & \text{if } x > N \end{cases}$$

The expression $E(Y) = \sum_{i=1}^{\infty} P(Y \geq i)$ sums up the probabilities of selecting tanks w serial number greater than or equal to i is $P(Y \geq i)$. To prove:

$$\begin{aligned} E(Y) &= \sum_{i=1}^{\infty} i \cdot P(Y=i) \\ &= 1 \cdot P(Y=1) + 2P(Y=2) + 3P(Y=3) + \cdots \\ &= \sum_{y=1}^{\infty} \left(\sum_{i=1}^{y} P(Y=y) \right) \\ &= \sum_{i=1}^{\infty} \left(\sum_{y=i}^{\infty} P(Y=y) \right) \quad \text{b/c } P(Y \geq i) = \sum_{y=i}^{\infty} P(Y=y) \\ &= \sum_{i=1}^{\infty} P(Y \geq i) \end{aligned}$$

Using the probability result we can establish a series as expression for the expected value of $X_{(k)}$

$$\begin{aligned} E(X_k) &= \sum_{i=1}^{N} P(X_{(k)} \geq i) \\ &= \sum_{i=1}^{N} (1 - P(X_{(k)} < i)) \end{aligned}$$

$$= \sum_{i=1}^{N} (1 - F_{X(k)}(i))$$
$$= \sum_{i=1}^{N} 1 - \left(\frac{i}{N}\right)^k$$

To approximate $E(X_{(k)})$, we can utilize left Riemann sum. To prove:

$$\Delta z = \frac{b-a}{n} = \frac{1}{n}$$

$$z_i = a + i\Delta z = \frac{i}{n}$$

$$\int_0^1 (1 - z^k)\, dz = \lim_{n\to\infty} \sum_{i=1}^{n} f(z_i)\Delta x$$
$$= \lim_{n\to\infty} \sum_{i=1}^{n} \left(1 - \left(\frac{i}{n}\right)^k\right) \frac{1}{n}$$

From the above we found $E(X_{(k)}) = \sum_{i=1}^{N} 1 - \left(\frac{i}{N}\right)^k$ and using Riemann sum we showed

$$\frac{E(X_{(k)})}{N} = \lim_{n\to\infty} \sum_{i=1}^{n} \left(1 - \left(\frac{i}{N}\right)^k\right) \frac{1}{n}$$

Now we can use this to approximate $E(X_{(k)})$

$$E(X_{(k)}) \approx N \int_0^1 (1 - z^k)\, dz$$
$$\approx N \left[z - \frac{z^{k+1}}{k+1} \right]_0^1$$
$$\approx N\left(1 - \frac{1}{k+1}\right) = N\left(\frac{k}{k+1}\right) //$$

This indicates the maximum is biased a estimator for N.

$$\hat{N}_{(k)} = \frac{k+1}{k} \cdot (X_{(k)})$$

This is the unbiased estimator for N.

$$\text{Var}(\hat{N}_{(k)}) = \left(\frac{k+1}{k}\right)^2 \text{Var}(X_{(k)})$$

$$= \frac{(k+1)^2}{k^2} \frac{(N+1)(N-k)k}{(k+1)^2(k+2)}$$

$$= \frac{(N+1)(N-k)}{k(k+2)}$$

MVUE for N is the function of MLE, which is the maximum observed value

① $\hat{N}_{(k)}$ is unbiased estimator for N

② The maximum observed value in the sample contains all the information about the parameter N (sufficiency)

As $\lim_{k\to\infty} V(\hat{N}_{(k)}) = \lim_{k\to\infty} \frac{(N+1)(N-k)}{k(k+2)} = 0$

Thus $\hat{N}_{(k)}$ is a consistent estimator for N.

Now using relative efficiency, we can compare the precision of MOM estimator to MVUE for N

$$RE(MoM, MVUE) = \frac{\text{Var MVUE}}{\text{Var MoME}} = \frac{\left(\frac{(N+1)(N-k)}{k(k+2)}\right)}{\left(\frac{N^2-1}{3k}\right)}$$

$$= \frac{3(N+1)(N-k)}{(k+2)(N^2-1)}$$

Using the preferred estimator which is MVUE, we can calculate the estimate number of Golden tanks.

$$\hat{N}(k) = \left(\frac{k+1}{k}\right) X(k) = \left(\frac{15+1}{15}\right) 2053 \approx 2190 \text{ tanks}$$