Tank Estimators

Jo Hardin

Fall 2022

How can a random sample of integers between 1 and N (with N unknown to the researcher) be used to estimate N? This problem is known as the German tank problem and is derived directly from a situation where the Allies used maximum likelihood to determine how many tanks the Axes had produced. See https://en.wikipedia.org/wiki/German_tank_problem.

1. The tanks are numbered from 1 to N. Working with your group, randomly select five tanks, without replacement, from the bowl. The tanks are numbered:
2. Think about how you would use your data to estimate N. (Come up with at least 3 estimators.) Come to a consensus within the group as to how this should be done. One person from your group will report out after the warm-up is over. Ideally, the person to report out will be someone who has not yet spoken in class this semester. Step-up if you haven't yet spoken. Step back if you speak regularly.

The estimates of N are:

The rules or formulas for the estimators of N based on a sample of n (in your case 5) integers are:

Assuming the random variables are distributed according to a discrete uniform. (Tbh, this model is with replacement, but the answers you get aren't much different than without replacement if $n << N$.)

$$X_i \sim P(X = x|N) = \frac{1}{N} \qquad x = 1, 2, \dots, N \qquad i = 1, 2, \dots, n$$

3. What is the method of moments estimator of N?
4. What is the maximum likelihood estimator of N? (Hint: draw a picture!)

Theoretical Mean Squared Error

Most of our estimators are made up of four basic functions of the data: mean, median, min, and max. Fortunately, we know something about their moments:

g($\underline{X}$)	E(g($\underline{X}$))	Var(g($\underline{X}$))
$\overline{X}$	$\frac{N+1}{2}$	$\frac{(N+1)(N-1)}{12n}$
median($\underline{X}$) = M	$\frac{N+1}{2}$	$\frac{(N-1)^2}{4n}$
min($\underline{X}$)	$\frac{(N-1)}{n}+1$	$\left(\frac{N-1}{n}\right)^2$
max($\underline{X}$)	$N-\frac{(N-1)}{n}$	$\left(\frac{N-1}{n}\right)^2$

Using the information on expected value and variance, we can calculate the MSE for 4 of the estimators that we have derived. (Remember that MSE = Variance + Bias^2.)

$$\begin{aligned} \text{MSE } (2\cdot\overline{X}-1) &= \frac{4(N+1)(N-1)}{12n} + \left(2\left(\frac{N+1}{2}\right)-1-N\right)^2 \\ &= \frac{4(N+1)(N-1)}{12n} \qquad (1) \end{aligned}$$

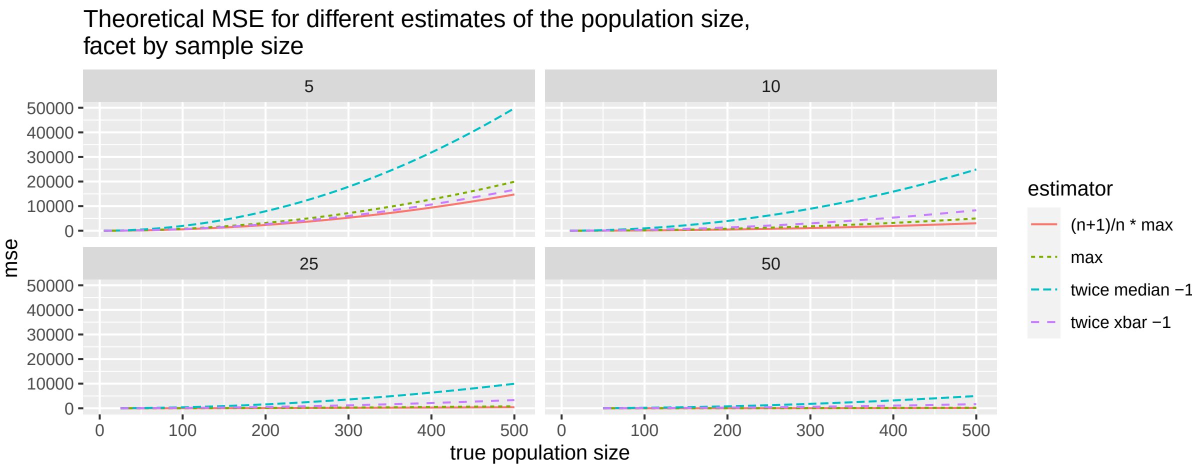
$$\begin{aligned} \text{MSE } (2\cdot M-1) &= \frac{4(N-1)^2}{4n} + \left(2\left(\frac{N+1}{2}\right)-1-N\right)^2 \\ &= \frac{4(N-1)^2}{4n} \qquad (2) \end{aligned}$$

$$\begin{aligned} \text{MSE } (\max(\underline{X})) &= \left(\frac{N-1}{n}\right)^2 + \left(N-\frac{(N-1)}{n}-N\right)^2 \\ &= \left(\frac{N-1}{n}\right)^2 + \left(\frac{N-1}{n}\right)^2 = 2*\left(\frac{N-1}{n}\right)^2 \qquad (3) \end{aligned}$$

$$\text{MSE } \left(\left(\frac{n+1}{n}\right)\max(\underline{X})\right) = \left(\frac{n+1}{n}\right)^2\left(\frac{N-1}{n}\right)^2 + \left(\left(\frac{n+1}{n}\right)\left(N-\frac{N-1}{n}\right)-N\right)^2 \qquad (4)$$

Empirical MSE

We don't need to know the theoretical expected value or variance of the functions to approximate the MSE. We can visualize the sampling distributions and also calculate the actual empirical MSE for any estimator we come up with.



By changing the population size and the sample size, we can assess how the estimators compare and whether one is particularly better under a given setting.

Some possible estimators of N are:[1]

$$\begin{aligned}
\hat{N}_1 &= 2 \cdot \overline{X} - 1 \quad \text{the MOM} \\
\hat{N}_2 &= 2 \cdot \text{median}(\underline{X}) - 1 \\
\hat{N}_3 &= \max(\underline{X}) \quad \text{the MLE} \\
\hat{N}_4 &= \frac{n+1}{n} \max(\underline{X}) \quad \text{less biased version of the MLE} \\
\hat{N}_5 &= \max(\underline{X}) + \min(\underline{X}) \\
\hat{N}_6 &= \frac{n+1}{n-1} [\max(\underline{X}) - \min(\underline{X})]
\end{aligned}$$

When $n = 5$, which means the sample size is 5 (whereas the population number is 447), the measurements are shown below. We want empirical MSE to be lowest and bias close to 0. The modified maximum has the lowest MSE and thus this estimator is the best when $n = 5$. The histograms estimating the sampling distributions are illustrated below.

```
calculate_N <- function(nsamp,npop){
  mysample =  sample(1:npop,nsamp,replace=F)  # what does this line do?
  xbar2 <- 2 * mean(mysample) - 1
  median2 <- 2 * median(mysample) - 1
```

[1] Note that the MOM and MLE estimators were derived under the assumption that the data are *iid* from a population of discrete uniform values. Because our data is sampled without replacement, we don't have an *iid* model. However, if $n <<< N$, the *iid* discrete uniform is a reasonable model for the situation at hand.

```r
  samp.max <- max(mysample)
  mod.max <- ((nsamp + 1)/nsamp) * max(mysample)
  sum.min.max <- min(mysample) + max(mysample)
  diff.min.max <- ((nsamp + 1)/(nsamp - 1)* (max(mysample) - min(mysample)))
  data.frame(xbar2, median2, samp.max, mod.max, sum.min.max, diff.min.max,nsamp,npop)
}

reps <- 2
nsamp_try <- 5
npop_try <- 447
map_df(1:reps, ~map2(nsamp_try, npop_try, calculate_N))
```

```
  xbar2 median2 samp.max mod.max sum.min.max diff.min.max nsamp npop
1 375.0     351      434   520.8         486          573     5  447
2 333.4     265      373   447.6         398          522     5  447
```

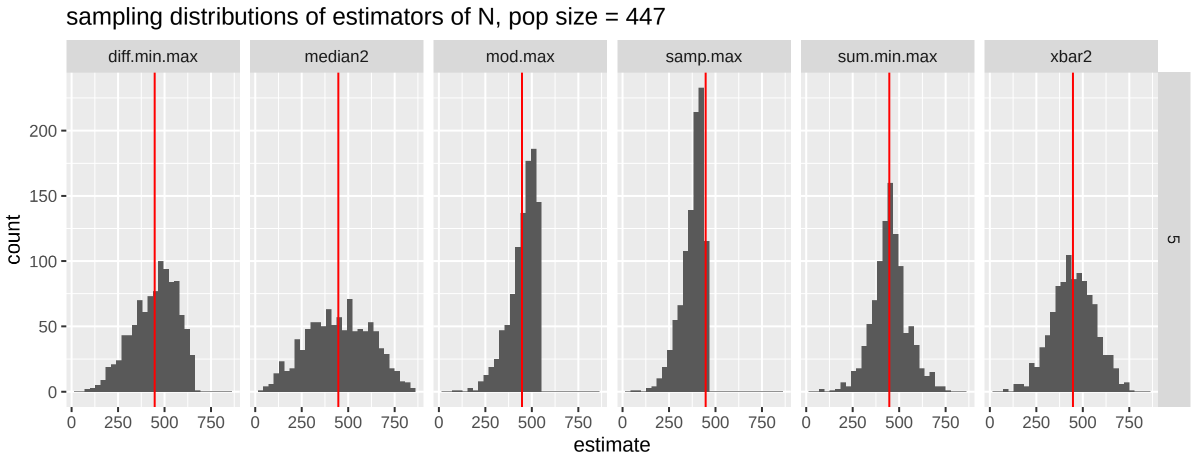
```r
reps <- 1000
results <- map_df(1:reps, ~map2(nsamp_try, npop_try, calculate_N))

# making the results long instead of wide:
results_long <- results %>%
  pivot_longer(cols = xbar2:diff.min.max,
               names_to = "estimator",
               values_to = "estimate")

# how is results different from results_long?  let's look at it:
results_long %>% head()
```

```
# A tibble: 6 x 4
  nsamp  npop estimator    estimate
  <dbl> <dbl> <chr>           <dbl>
1     5   447 xbar2            163.
2     5   447 median2          141
3     5   447 samp.max         186
4     5   447 mod.max          223.
5     5   447 sum.min.max      204
6     5   447 diff.min.max     252
```

```
results_long %>%
  group_by(nsamp, npop, estimator) %>%
  summarize(mean = mean(estimate),
            median = median(estimate),
            bias = mean(estimate - npop),
            var = var(estimate),
            mse = (mean(estimate - npop))^2 + var(estimate))
```

```
# A tibble: 6 x 8
# Groups:   nsamp, npop [1]
  nsamp  npop estimator     mean median   bias    var    mse
  <dbl> <dbl> <chr>        <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
1     5   447 diff.min.max  449.   466.   1.77 14316. 14319.
2     5   447 median2       451.   451    4.48 29129. 29150.
3     5   447 mod.max       450.   468    2.76  5423.  5431.
4     5   447 samp.max      375.   390  -72.2   3766.  8979.
5     5   447 sum.min.max   450.   449    3.42  9114.  9126.
6     5   447 xbar2         451.   450.   3.58 13344. 13356.
```

```
results_long %>%
  ggplot(aes(x = estimate)) +
  geom_histogram() +
  geom_vline(aes(xintercept = npop), color = "red") +
  facet_grid(nsamp ~ estimator) +
  ggtitle("sampling distributions of estimators of N, pop size = 447")
```


sampling distributions of estimators of N, pop size = 447
diff.min.max
median2
mod.max
samp.max
sum.min.max
xbar2
count
estimate