Estimating the Size of a Population

KEYWORDS:
Teaching;
Simulation;
Minitab.

Roger W. Johnson
Carleton College, Northfield, Minnesota, U.S.A.

Summary
Several estimates of an unknown population size are compared.

◆ INTRODUCTION ◆

SUPPOSE we have a population of objects labelled 1, 2, 3, ..., N with N unknown. From a random sample $X_1, X_2, X_3 ..., X_n$ of size n without replacement from this population we consider how to estimate N. One may estimate the number of runners in a race, taxicabs in a city, or concession booths at a fair, for instance, based upon a seeing just a sample of these labelled items. Ruggles and Brodie (1947) summarise Allied estimates of German weapons production (e.g. tanks) during World War II from factory serial number markings using the above model and extensions of it. Ruggles (1991) also indicates that German intelligence "conducted extensive factory markings analysis on Soviet military equipment" during WWII and "factory markings on Soviet equipment were also analysed" during the U.S.-Korean War. In this article we present several estimates of N using elementary arguments and show how Minitab might be used to select which one is "best." At the end of the article we state some results which could be verified by students with a more advanced probability and statistics background.

◆ THE ESTIMATES ◆

We now derive a number of estimates of N using only a little common sense (c.f. Noether (1990), pp.33-43). First of all, suppose we knew the middle value m in the list 1, 2, ..., N. Then there would be m–1 values below m and m–1 values above m in the list. So, including the middle value m, we would have $N = (m-1) + 1 + (m-1) = 2m-1$. Now, since we don't know m, it is certainly reasonable to replace it by an estimate of the middle such as the median or mean. If $\tilde{X}$ and $\overline{X}$ denote the median and mean, respectively, of our observed sample of labels $X_1, X_2, X_3 ..., X_n$, this gives the estimates

$$\hat{N}_1 = 2\tilde{X} - 1 \quad \text{and} \quad \hat{N}_2 = 2\overline{X} - 1.$$

(These estimates and subsequent estimates, if not integers, should be rounded to the nearest integer.) Unfortunately, both $\hat{N}_1$ and $\hat{N}_2$ can be less than the largest label, $X_{(n)}$, that we see in our sample! (In what follows we let $X_{(1)} < X_{(2)} < \ldots < X_{(n-1)} < X_{(n)}$ denote the ordered values of our sample $X_1, X_2, \ldots, X_n$.) Note, for instance, if $X_1 = 2$, $X_2 = 10$, and $X_3 = 3$ in a sample of size $n = 3$, then $\hat{N}_1 = 5$, and $\hat{N}_2 = 9$, but $N \geq X_{(3)} = 10$. We now derive some estimates which are always at least as big as the largest sample label we see. By symmetry, one would suppose that the number of unobserved labels above $X_{(n)}$ should be about equal to the number of unobserved labels below $X_{(1)}$. So, setting

$$N - X_{(n)} = X_{(1)} - 1 \quad \text{yields} \quad \hat{N}_3 = X_{(n)} + X_{(1)} - 1.$$

Extending the above reasoning it would seem reasonable to set the number of unobserved labels above $X_{(n)}$ to be the average of: the number of unobserved labels below $X_{(1)}$, the number of unobserved labels between $X_{(1)}$ and $X_{(2)}$, the number of unobserved labels between $X_{(2)}$ and $X_{(3)}, \ldots$, and the number of unobserved labels between $X_{(n-1)}$ and $X_{(n)}$. That is, set $N - X_{(n)}$ equal to

$$\left\{\left[X_{(1)} - 1\right] + \left[X_{(2)} - X_{(1)} - 1\right] + \left[X_{(3)} - X_{(2)} - 1\right] + \ldots + \left[X_{(n)} - X_{(n-1)} - 1\right]\right\} / n$$

which reduces to $\left[X_{(n)} / n\right] - 1$ giving

$$\hat{N}_4 = \frac{(n+1)}{n} X_{(n)} - 1.$$

Neither of the estimates $\hat{N}_3$ or $\hat{N}_4$ have the undesirable property that they can be less than the largest sample label seen.

◆ COMPARING THE ESTIMATES ◆

To compare these estimates one can simulate drawing from the population 1,2,...,N with a known N using almost any statistical package. (If a statistical package is not available but one knows some programming language with a random number generator, this simulation can be performed with the aid of Bebbington (1976)). After computing each estimate for a number of simple random samples of size n from the population of size N, one would examine the distribution of estimates obtained. This may be accomplished using the Minitab commands below:

```
MTB > store 'compare.mtb'
STOR> noecho
STOR> sample k1  c1 c2
STOR> let k2 = maximum(c2)
STOR> let k3 = minimum(c2)
STOR> let k4 = median(c2)
STOR> let k5 = mean(c2)
STOR> let k6 = 2*k4–1
STOR> let k7 = 2*k5–1
STOR> let k8 = k2 + k3–1
STOR> let k9 = (k1+1)*k2/k1–1
STOR> stack k6 c3 c3
STOR> stack k7 c4 c4
STOR> stack k8 c5 c5
STOR> stack k9 c6 c6
STOR> end
MTB > name c3 'N1' c4 'N2' c5 'N3' c6 'N4'
MTB > erase c1–c6
MTB > set c1
DATA>1:5000
DATA> end
MTB > note: Here N=5000
MTB > let k1 = 100
MTB > note: Here n=100
MTB > note: Now take, say, 50 simple random
MTB > note: samples of size k1 from our
MTB > note: population and compare the results:
MTB > execute 'compare.mtb' 50
MTB > print c3-c6
MTB > describe c3-c6
MTB > histogram c3-c6
```

The output of the describe command will tend to provide evidence that each of the estimates is unbiased, and that the standard deviations of the estimates $\hat{N}_i$ decrease with i (with those of $\hat{N}_1$ and $\hat{N}_2$ considerably more than those of $\hat{N}_3$ and $\hat{N}_4$). Consequently, $\hat{N}_4$ would appear to be the best estimate.

◆ THEORETICAL RESULTS ◆

A student with a knowledge of elementary combinatorics who is adept at manipulating sums may verify the entries in Table 1 below. Note that the estimates are unbiased, and that the standard deviations of the estimates $\hat{N}_i$ do decrease with i.

Table 1. Expectation and Variance of Estimates.

i	$\hat{N}_i$	$E(\hat{N}_i)$	$Var(\hat{N}_i)$
1	$2\bar{X}-1$	N	$\dfrac{(N-n)(N+1)}{(n+2)}$ (n odd) $\dfrac{n}{(n+1)}\cdot\dfrac{(N-n)(N+1)}{(n+2)}$ (n even)
2	$2\overline{X}-1$	N	$\dfrac{(n+2)}{3n}\cdot\dfrac{(N-n)(N+1)}{(n+2)}$
3	$X_{(n)}+X_{(1)}-1$	N	$\dfrac{2}{(n+1)}\dfrac{(N-n)(N+1)}{(n+2)}$
4	$\dfrac{(n+1)}{n}X_{(n)}-1$	N	$\dfrac{1}{n}\dfrac{(N-n)(N+1)}{(n+2)}$

In this article we will be content to verify only that $E(\hat{N}_4)=N$. (As a simple check on the variance calculations, however, note that they agree in the case n=1, for which $\hat{N}_1=\hat{N}_2=\hat{N}_3=\hat{N}_4$, and correctly vanish in the case n=N.)

As $E(\hat{N}_4)=E\left(\dfrac{(n+1)}{n}X_{(n)}-1\right)=\dfrac{(n+1)}{n}E(X_{(n)})-1$,

to show $E(\hat{N}_4)=N$ it suffices to show $E(X_{(n)})=\dfrac{n(N+1)}{(n+1)}$. Now, as we are taking a simple random sample of size n from a population of size N, each of the $C(N,n)$ samples are equally likely. (Here, $C(r,s)=\dfrac{r!}{s!(r-s)!}$.) The event $\{X_{(n)}=j\}$ implies $X_{(1)},X_{(2)},\ldots,X_{(n-1)}$ must be selected from 1, 2, ..., j–1, and because there are $C(j-1,n-1)$ such selections,

$$P(X_{(n)}=j)=\frac{C(j-1,n-1)}{C(N,n)} \quad j=n,n+1,\ldots,N. \qquad (1)$$

Consequently, $E(X_{(n)})=\sum_{k=n}^{N}kP(X_{(n)}=k)$

$$=\frac{1}{C(N,n)}\sum_{k=n}^{N}kC(k-1,n-1)$$

but

$$kC(k-1,n-1) = k\frac{(k-1)!}{(n-1)!(k-n)!} = n\frac{k!}{n!(k-n)!}$$
$$= nC(k,n)$$

so the above sum becomes

$$E(X_{(n)}) = \frac{n}{C(N,n)}\sum_{k=n}^{N} C(k,n). \qquad (2)$$

To simplify this, note that the probabilities in (1) must sum to 1. So

$$\sum_{j=n}^{N} C(j-1,n-1) = C(N,n)$$

which implies

$$\sum_{k=n}^{N} C(k,n) = C(N+1,n+1). \qquad (3)$$

Combining (2) and (3) and going through some simple algebra we see

$$E(X_{(n)}) = \frac{n}{C(N,n)}\,C(N+1,n+1) = \frac{n(N+1)}{(n+1)}$$

as required.

Looking back at the table, we see that each of the estimates is unbiased and, among the four, $\hat{N}_4$ is the one with minimum variance. It can be shown, in fact, that $\hat{N}_4$ is the uniformly minimum variance unbiased estimate (UMVUE) of N. We should also note that $\hat{N}_2$ is the method of moments estimate of N, and $\hat{N}_4$ is a scaled and shifted version of the maximum likelihood estimate, $X_{(n)}$, of N.

◆ CONCLUDING COMMENTS ◆

There are, of course, problems in which one would like to estimate the size of a population but for which the labels do not run from 1 to N. Suppose, for instance, that the population elements are labelled sequentially from S to N with both S and N unknown. Ruggles and Brodie (1947, p81), for example, indicate that this was the case with gearbox markings on the German Mark V ("Panther") tanks during WWII. Students may wish to consider how to estimate the number of items in this population (here, $X_{(1)}$ and $X_{(n)}$ are the sufficient statistics.) Again, estimates could be analysed theoretically and/or by simulation. The interested reader should see Ruggles and Brodie (1947) for other, more complicated, labellings of population elements.

We close with an anecdote of Colonel Trevor Dupuy (1991). "In the Middle East a few years ago I was given permission by Israeli military authorities to go through the entire Merkava Tank production line. At one time I asked how many Merkavas had been produced, and I was told that this information was classified. I found it amusing, because there was a serial number on each tank chassis."

References

Bebbington, A.C. (1976). A Simple Method of Drawing a Sample Without Replacement. *Applied Statistics*, **24**, 136.

Dupuy, T.N. (1991), President of Hero-TNDA Publishers & Researchers, Inc, 1324 Kurtz Road, McLean, VA 22101. Personal Correspondence.

Noether, G.E. (1990). *Introduction to Statistics: The Nonparametric Way*. Springer-Verlag, New York.

Ruggles, R. and Brodie, H. (1947). An Empirical Approach to Economic Intelligence in World War II. *Journal of the American Statistical Association*, **42**, 72-91.

Ruggles, R. (1991), Department of Economics, Yale University. Personal Correspondence.

Acknowledgment

Thanks to G. Fred Kramer, Naval Command, Control and Ocean Surveillance Center and to Jim Maar, National Security Agency, for help in tracking down sources.