REPLICATION OF SAMPLES FROM FINITE POPULATIONS

Barry L. Ford, USDA

1. Introduction

In the context of a simple random sample replication is randomly dividing a sample into groups so that each group is capable of estimating a population parameter. Replication has become an important strategy in sampling theory. Not only does replication simplify the calculations involved in a complex sampling scheme, but it also yields unbiased estimates of the variance of complex, nonlinear estimators.

When one has an infinite population or is sampling with replacement, a rationale for the number of replicates is given by Des Raj in his sampling text (2). The purpose of this paper is to extend the formulas to finite populations. Furthermore, it is demonstrated that one only needs a moderate population size in order to ignore the process of sampling without replacement and use the simpler formulas of sampling with replacement as an approximation.

When the sample design is a simple random sample of size n, the population total:

$$\tau = \sum_{i=1}^{N} x_i$$

is usually estimated by the sample statistic:

$$T_{simple} = \frac{N}{n} \sum_{i=1}^{n} x_i = N\bar{x}.$$

Another estimator of τ results from the technique of replication. If r replicates of size m are selected, replication yields the estimator;

$$T_{replicate} = \frac{\sum_{i=1}^{r} t_i}{r}$$

where

$$t_i = \frac{N}{m} \sum_{j=1}^{m} x_{ij}$$

Thus, the subscript of x refers to the j^{th} element in replicate i.

Obviously, the expected value of $T_{replicate}$ is:

$$E(T_{replicate}) = E(t)$$

$$= \frac{N}{m} \sum_{i=1}^{m} E(x)$$

$$= N \cdot E(x) \qquad (1.1)$$

$$= E(T_{simple}). \qquad (1.2)$$

When one selects the sample units *with replacement*, the replicates are independent and it is obvious that the variance of $T_{replicate}$ is:

$$Var(T_{replicate}) = \frac{M_2(t_i)}{r}$$

where $M_i(\cdot)$ represents the i^{th} central moment. Thus, when n = mr:

$$Var(T_{replicate}) = \frac{1}{r} M_2 \left(\frac{N}{m} \sum_{j=1}^{m} x_{ij}\right)$$

$$= \frac{N^2}{mr} M_2(x)$$

$$= \frac{N^2}{n} M_2(x) \qquad (1.3)$$

$$= Var(T_{simple}). \qquad (1.4)$$

If $u = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$, then one can estimate Var (T_{simple}) unbiasedly by:

$$Var(T_{simple}) = \frac{N^2}{n} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1} \qquad (1.5)$$

$$= \frac{N^2}{n} u \qquad (1.6)$$

where $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$. Also, by allowing $u_t = \frac{\sum_{i=1}^{r} (t_i - \bar{t})^2}{r - 1}$, one can estimate Var $(T_{replicate})$ unbiasedly with:

$$var(T_{replicate}) = \frac{\sum_{i=1}^{r} (t_i - \bar{t})^2}{r(r - 1)} \qquad (1.7)$$

$$= \frac{u_t}{r} \qquad (1.8)$$

where $\bar{t} = \frac{\sum_{i=1}^{r} t_i}{r}$.

Up to this point there is no loss in efficiency by adopting a replicated design. However, *there is a loss of efficiency in replicated designs caused by a decrease in the precision of the variance estimate* (Raj, pg. 194). Remembering (1.6) one sees that the squared coefficient of variation of $var(T_{simple})$ is:

$$CV^2 \left[var(T_{simple})\right] = CV^2 (u) \qquad (1.9)$$

and

$$CV^2 \left[var(T_{replicate})\right] = CV^2 (u_t). \qquad (1.10)$$

Because the right side of expressions (1.9) and (1.10) are more easily written and comprehended, they are used in the following comparisons. It is well known (Raj, pg. 190) that:

$$CV^2 (u) = \frac{1}{n} \left\{\beta_x - \frac{n - 3}{n - 1}\right\} \qquad (1.11)$$

where $\beta = \frac{M_4(\cdot)}{M_2^2(\cdot)}$, the kurtosis of a distribution.

Thus,

$$CV^2(u_t) = \frac{1}{r}\left\{\beta_t - \frac{r-3}{r-1}\right\} \quad (1.12)$$

An easy calculation (Raj, 1964) yields:

$$\beta_t = \frac{1}{m}\left\{\beta_x + 3(m-1)\right\} \quad (1.13)$$

and therefore:

$$CV^2(u_t) = \frac{1}{n}\left\{\beta_x - 3 + \frac{2}{r-1}\right\}. \quad (1.14)$$

The result (Raj, pg. 195) is that:

$$CV^2(u_t) - CV^2(u) = \frac{2(n-r)}{(n-1)(r-1)} > 0. \quad (1.15)$$

For example, in a simple random sample of total size n = 100 with r = 10 the variance estimate using the replicates has a squared coefficient of variation which is approximately 0.20 greater than the squared coefficient of variation of the variance estimate, u. One can observe from (1.15) that r should be as large as possible.

2. The Stability of Variance Estimates When Sampling a Finite Population

Suppose a sample of size n is drawn *without replacement* from a population of size N. Then the most common estimator of τ:

$$T = N\bar{x}$$

remains of the same form as when sampling with replacement but the variance of T becomes:

$$Var(T) = \left(1 - \frac{n}{N}\right)\left(\frac{N}{N-1}\right)\frac{M_2(x)}{n}. \quad (2.1)$$

An unbiased estimator of $\frac{N}{N-1}M_2(x)$ (usually referred to by sampling theory texts as S^2) is:

$$u = \sum_{i=1}^{n}\frac{(x_i - \bar{x})^2}{n-1}. \quad (2.2)$$

One must derive the variance of u under the condition of sampling without replacement. The details are not given in this paper, but if they are requested will be furnished by the author. One can derive:

$$E(u^2) = \frac{1}{n}M_4(x) + \frac{1}{N-1}\left(\frac{n-1}{n}\right)\left(\frac{2}{(n-1)^2} + 1\right)\left\{NM_2^2(x) - M_4(x)\right\} + \frac{1}{N-1}\left(\frac{4}{n}\right)\left\{M_4(x)\right\} + \left(\frac{2(n-2)(n-3)}{(N-1)(N-2)(n)(n-1)}\right)\left\{NM_2^2(x) - 2M_4(x)\right\} + \left(\frac{3(n-2)(n-3)}{(N-1)(N-2)(N-3)(n)(n-1)}\right)\left\{NM_2^2(x) - 2M_4(x)\right\}. \quad (2.3)$$

The result, (2.3) can be found in a different form in a sampling text (Hansen, Hurwitz, Madow; page 101; Volume II).

By substracting the term $E(u)^2$ from both sides of (2.13) and remembering that:

$$E(u) = \frac{N}{N-1}M_2(x) \quad (2.4)$$

One finds:

$$Var(u) = M_4(x)\left\{\frac{N-n+2}{n(N-1)} \cdot \frac{2(n-2)(n-3)(2N-3)}{n(n-1)(N-1)(N-2)(N-3)} - \frac{2}{n(n-1)(N-1)}\right\} + NM_2^2(x)\left\{\frac{2}{n(n-1)(N-1)} + \frac{n-1}{n(N-1)} + \frac{(n-2)(n-3)(2N-3)}{n(n-1)(N-1)(N-2)(N-3)} - \frac{N}{(N-1)^2}\right\} \quad (2.5)$$

After a great deal of algebra (2.5) can be simplified into the form:

$$Var(u) = D_1 M_4(x) + D_2 M_2^2(x) \quad (2.6)$$

where:

$$D_1 = \frac{N(N-n)\left\{(N-1)(n-1) - 2\right\}}{n(n-1)(N-1)(N-2)(N-3)} \quad (2.7)$$

$$D_2 = \frac{N(N-n)(3N^2 - nN^2 - 6N + 3n + 3)}{n(n-1)(N-1)^2(N-2)(N-3)} \quad (2.8)$$

There are a few properties of D_1 and D_2 that should be pointed out.

Theorem 2.1: If n = N, then $D_1 = D_2 = 0$.

The proof is obvious.

Theorem 2.2: With simple random sampling from a a finite population:

$$\lim Var(u) = \frac{1}{n}M_4(x) - \frac{(n-3)}{(n-1)}M_2^2(x). \quad (2.9)$$

Again the proof is obvious. As expected, (2.9) is the variance of u when sampling with replacement (Raj, pg. 190) and will be used as a large size approximation to Var(u).

Theorem 2.3: If n strictly increases, the variance of u strictly decreases.

The proof is accomplished by showing that both D_1 and D_2 decrease as n increases. By rewriting (2.8):

$$D_2 = \frac{N(N-n)\left\{3N^2 - 6N - n(N^2-3) + 3\right\}}{n(n-1)(N-1)^2(N-2)(N-3)}$$

it is evident that as n increases the denominator increases and the numerator decreases if N > 3. (The restriction on N is inconsequestial because N must be greater than 3 to prevent division by zero.)

It is also true that as n increases, D_i decreases, but the proof required is more tedious. Suppose n increases by one then from (2.7)

$$(D_1|n) = \frac{N(N-n)\left\{(N-1)(n-1) - 2\right\}}{n(n-1)(N-1)(N-2)(N-3)} \quad (2.10)$$

$$(D_1|n+1) = \frac{N(N-n-1)\left\{(N-1)(n) - 2\right\}}{n(n+1)(N-1)(N-2)(N-3)} \quad (2.11)$$

Ignoring common factors, to prove $(D_1|n) > (D_1|n+1)$ one needs to show that:

$$\frac{(N-n)(nN-N-n-1)}{n-1} > \frac{(N-n-1)(nN-n-2)}{n+1} \quad (2.12)$$

Algebraically, (2.12) is equivalent to:

$$\frac{N^2n^2 - Nn - Nn^3 + n^3 + 2n^2 - N^2 - Nn^2 + n}{(n-1)(n+1)}$$

$$> \frac{N^2n^2 - Nn^2}{(n-1)(n+1)}$$

$$- \frac{Nn^3 + n^3 - 2n^2 - nN^2 + 2N - n - 2}{(n-1)(n+1)} \qquad (2.13)$$

After subtracting all terms on the left side of (2.12), one need only show for $n > 1$:

$$Q = N^2(n-1) - N(n+2) + 4n^2 + 2n + 2 > 0. \qquad (2.14)$$

When $n = 2$, (2.14) becomes $Q = N^2 - 4N + 22 > 0$ which is true for all N.

If n increases by one, then the change in Q, ΔQ, is:

$$\Delta Q = N^2 - N + 8n + 8 > 0$$

for $N > 0$, $n > 1$. Thus, one proves (2.14) which proves (2.12) which in turn proves that C_1 strictly decreases as n strictly increases given $N > 3$, $n > 1$. Therefore, one has the property that as n strictly increases, the variance of u strictly decreases.

Knowing the variance of u, formula (2.6) and remembering that $E(u) = \frac{N}{N-1} M_x(x)$; one finds:

$$CV^2(u) = C_1 \beta_x + C_2$$

where:

$$C_1 = \frac{(N-1)(N-n)\{(N-1)(n-1)-2\}}{n(n-1)N(N-2)(N-3)}$$

$$C_2 = \frac{(N-n)\{3N^2 - nN^2 - 6N + 3n + 3\}}{n(n-1)N(N-2)(N-3)}$$

It is easy to see that the limit as $N \to \infty$ of formula (2.15) is formula (1.11), the formula for with replacement sampling. Table 1 displays the values of N where the difference in these two formulas in less than 0.01. Thus, for population sizes larger than those in the table one can forget the condition of with replacement sampling and use the simpler formulas of without replacement sampling.

3. Determining the Number of Replicates

Now one should consider two situations that often arise in relicated sampling.

Case 1: *R relicates of size m are constructed (perhaps in a nonrandom manner) from a population. Assuming a without replacement structure within each replicate, how many replicates are needed to achieve a desired level of the coefficient of variation?*

A good example of this situation is where the population is ordered according to some arbitrary criteria and replicates are formed systematically. When replicates are not formed randomly, the obvious method of estimating any coefficient of variations is to consider each replicate as a sampling unit and to estimate the distribution of the replicates. Thus, to estimate the coefficient of variation of u_t (1.7 and 1.8), one uses (2.26) and substitutes the corresponding parameters from the population of replicates.

Case 2: *Suppose one must randomly select r replicates. Within each replicate units are chosen without replacement. However, each replicate may contain any unit in the population.*

One still uses the with replacement formula of Raj:

$$CV^2(u_t) = \frac{1}{r}\left\{\beta_t - \frac{r-3}{r-1}\right\} \qquad (3.1)$$

Now β_t is also subject to the laws of without replacement sampling. It is possible to derive β_t in terms of β_x. The derivation is again quite tedious, but details will be furnished by the author upon request.

One can derive:

$$M_4(t) = \frac{N^4}{m^4}\Bigg[m\,M_4(x) + \frac{3m(m-1)}{N-1}\left\{N M_2^2(x) - M_4(x)\right\} - 4\frac{m(m-1)}{(N-1)} M_4(x) + \frac{6m(m-1)(m-2)}{(N-1)(N-2)}\left\{2M_4(x) - NM_2^2(x)\right\} + \frac{3m(m-1)(m-2)(m-3)}{(N-1)(N-2)(N-3)}\left\{NM_2^2(x) - 2M_4(x)\right\}$$

Algebra yields the result:

$$M_4(t) = \frac{N^4(N-m)}{m^3(N-1)(N-2)(N-3)}\Big[3N(m-1)(N-m-1)M_2^2(x) + (N^2 - 6mN + 6m^2 + N)M_4(x)\Big] \qquad (3.2)$$

Thus, by dividing expression (3.2) by the square of $M_2(t)$ one finds that:

$$\beta_t = \frac{(N-1)}{m(N-m)(N-2)(N-3)}\Big[(N^2 - 6mN + 6m^2 + N)\beta_x + 3N(m-1)(N-m-1)\Big] \qquad (3.3)$$

One should note that when $m = 1$, $\beta_t - \beta_x$ and as N approaches ∞, expression (3.3) becomes (1.13). Table 2 shows those values of N such that the without replacement formula for β_t, (3.3), can be approximated by the with replacement formula, (1.13). These values of N are extremely low.

One should also note that in both Tables 1 and 2 the formulas are not monotonic functions of N but curves. When computer programs were written to compute the tables, this fact showed up as irregularities in the tables, However, corrections were made, and calculations were performed to insure that N was large enough to compensate for curves in the functions.

Table 1: Population sizes for which the coefficient of variation of the estimate of the population variance $CV(u_t)$, can be approximated by the formula $CV(u) = [\frac{1}{n}(\beta_x - \frac{n-3}{n-1})]^{\frac{1}{2}}$

β_x = Kurtosis

n = Sample Size	1.0	2.0	3.0	4.0	5.0	10.0	20.0	50.0	100.0
2	100	102	104	111	116	139	178	263	363
3	116	102	100	112	117	140	179	264	364
4	123	110	113	119	125	154	203	306	425
5	127	120	129	139	150	195	264	405	567
6	130	129	145	160	174	233	321	499	701
10	135	164	199	230	257	364	514	812	1146
15	138	200	255	300	339	490	699	1110	1570
25	140	258	342	408	465	681	978	1559	2210
50	145	368	500	603	691	1020	1470	2347	3328
100	147	528	725	877	1006	1488	2146	3426	4706
500	502	1260	1714	2065	2097	2657	3601	6237	9170
1000	1002	1468	1715	2066	2543	3543	5433	9398	13850

Table 2: Values of N, population size, for which the calculation of β_t (Kurtosis of replicates) differs by less than 0.1 between the with replacement and the without replacement formulas.

β_x = Kurtosis

m = Replicate Size	1.0	2.0	3.0	4.0	5.0	10.0	20.0	50.0	100.0
2	22	12	34	60	84	210	460	1210	2460
3	30	15	45	78	111	279	612	1611	3276
4	32	16	52	88	124	312	688	1812	3688
5	35	20	55	95	135	335	735	1935	3935
6	36	18	54	96	138	348	762	2016	4098
10	30	30	60	110	150	380	830	2180	4430
15	30	30	60	105	165	390	855	2265	4590
25	50	50	75	125	175	400	875	2325	4725
50	100	100	100	150	200	400	900	2400	4850
100	200	200	200	200	200	400	900	2400	4900
500	1000	1000	1000	1000	1000	1000	1000	2500	5000
1000	2000	2000	2000	2000	2000	2000	2000	2000	5000

Case 3: First one selects n units without replacement from the population. These n units are then randomly selected without replacement to form r replicates of size m.

Now one must use formula (2.16) to form:

$$CV^2(u_t) = \beta_t C_1 + C_2 \qquad (3.4)$$

where:

$$\beta_t = \frac{(N-1)}{m(N - m(N-2)(N-3)} \left[(N^2 - 6mN + 6m^2 + N)\beta_x + 3N(m-1)(N-m-1)\right] \qquad (3.5)$$

$$C_1 = \frac{(R-1)(R-r)(rR-R-r-1)}{r(r-1)R(R-2)(R-3)} \qquad (3.6)$$

$$C_2 = \frac{(R-r)(3R^2 - rR - 6R + 3r + 3)}{r(r-1)R(R-2)(R-3)} \qquad (3.7)$$

Question 1: If m is fixed, what should r be to insure a specific level, a, of $CV^2(u)$?

Because of theorem 2.3 it is possible to find the lowest value of r which satisfies the CV requirement by using a simple computer program. The computer program would use the method of bisection. It would:

1: solve equation (3.12) for r* = 2 (if $CV^2(u_t) \leq \alpha$ at r* = 2, then r* is the solution and the problem is solved)

2: solve for r** = A, where A is a large even number

3: solve (3.12) for $r' = \frac{r^{**} + r^*}{2}$ if r' is even and for $r' = \frac{r^{**} + r^*}{2} - 1$ otherwise

4: if $CV^2(u_t) \leq \alpha$ at r', then r** is set equal r' and return to step 3

5: if $CV^2(u_t) \geq \alpha$ at r, then r* is set equal to r' and return to step 3

6: continue until r** - r* = 2 and then set r' - r* + 1

7: if $CV^2(u_t) > \alpha$ at r', then r** is the solution; if $CV^2(u_t) \leq \alpha$ at r', then r' is the solution.

If N is large enough (see Table 1 and Table 2), one may use the simpler, with replacement formulas for β_t, C_1, and C_2.

If one can make the large population assumption, one has:

$$CV^2(u_t) = C_1 \beta_t + C_2 \qquad (3.8)$$

where:

$$C_1 = \frac{1}{r} \qquad (3.9)$$

$$C_2 = \frac{1}{r}\left(\frac{3 - r}{r - 1}\right) \qquad (3.10)$$

in place of equations (3.6) and (3.7) and:

$$\beta_t = \lambda_1 \beta_x + \lambda_2$$

where:

$$\lambda_1 = \frac{1}{m} \qquad (1.11)$$

$$\lambda_2 = \frac{3(m - 1)}{m} \qquad (1.12)$$

in place of equation (3.3). Thus, one can solve:

$$CV^2(u_t) = \frac{1}{r}\{(\lambda_1 \beta_x + \lambda_2) + \frac{3 - r}{r - 1}\}$$

for r by use of the quadratic formula:

$$r = \frac{1}{2a}[\alpha + \lambda_1 \beta_x + \lambda_2 - 1 \pm \{(\alpha + \lambda_1 \beta_x + \lambda_2 - 1)^2 - 4\alpha^2 (\lambda_1 \beta_x + \lambda_2 - 3)\}^{\frac{1}{2}}]$$

Example: Suppose $\beta_x = 17$ and one desires an α of 0.30 (i.e. $CV(u_t) \approx 0.55$).

Then:

$$\beta_t = \frac{17}{2} + \frac{3}{2} = 10$$

and

$$r = \frac{1}{0.60}[0.30 + 10 - 1 \pm \{(0.30 + 10 - 1)^2 - 4(0.30)^2 (7)\}^{\frac{1}{2}}]$$

$$r = \frac{1}{0.60}[9.3 \pm (86.49 - 2.52)^{\frac{1}{2}}]$$

$$r = 30.8 \quad \text{or} \quad r = 0.20.$$

Thus, one would select 31 replicates.

Question II: Suppose n is fixed, but m is not fixed. What combination of r and m is best?

When R is small, one can find a minimum r (thus, maximum m) by using the computer program outlined above and substituting n/r for m in equation (3.5). To get a maximum r (thus, minimum m) one should substitute n/m = r and proceed iteratively (beginning with m = 2) through the calculations.

Suppose from Tables 1 and 2 that large population approximations are appropriate. $CV^2(u_t)$ is maximized for fixed n when m = 2. One can see that:

$$CV^2(u_t) = \frac{1}{r}\left[\frac{\beta_x}{m} + \frac{3(m - 1)}{m}\right] + \frac{1}{r}\left(\frac{3 - r}{r - 1}\right)$$

$$= \frac{\beta_x}{n} + \frac{3(m - 1)}{n} + \frac{m}{n}\left(\frac{3 - \frac{n}{m}}{\frac{n}{m} - 1}\right)$$

$$= \frac{\beta_x}{n} + \frac{3(m - 1)}{n} + \frac{m(3m - n)}{n(n - m)} \quad .$$

It is obvious that $CV^2(u_t)$ will increase with an increase in m. Therefore, if n is fixed, m = 2 will yield the lowest $CV(u_t)$. If one has other restrictions on the size of m, one can proceed inductively with larger values of m until these restrictions are met or until $CV(u_t)$ exceed an acceptable level.

Question III: If m, r, and n are unkown what values should they have to attain a specific level, a, of $CV^2(u_t)$?

Certainly a minimum n is determined by a desired accuracy on the mean or total estimate. From this minimum n one can compute the calculations of $CV^2(u_t)$ for m = 2 (when using the with replacement formula). If $CV^2(u_t)$ is greater than the desired α, one can continue to n + 1 and so forth because m = 2 yields the minimum $CV^2(u_t)$ for a specific n. When a certain n satisfies the requirements, then one can proceed inductively on m.

When using the with replacement formulas such principles can not be applied because it can not be shown that m = 2 yields a minimum $CV^2(u_t)$ for a fixed m.

5. Conclusions

From Table 1 one recognized the fact that most large sample surveys which sample *without replacement* may use the *with replacement* formulas of Raj as a good approximation. When the population sizes are small enough to require the exact formulas presented here, one can estimate the size and number of replicates needed to stabilize the variance estimator. These two factors--size and number--are determined by a specific precision requirement on the estimated variance of a total. This paper only presents work on simple random samples.

6. Bibliography

1. Hansen, Morris H.; Hurwitz, William N.; and Madow, William G. Sample Survey Methods and Theory, New York, Wiley and Sons. 1953.

2. Raj, Des. Sampling Theory. New York, McGraw-Hill Book Company. 1968