SAMPLING FROM FINITE POPULATIONS: ACTUAL COVERAGE PROBABILITIES FOR CONFIDENCE INTERVALS

Jörgen Dalén, Statistics Sweden

1 Introduction

In survey sampling the prevailing estimation strategy for estimating a finite population parameter θ is based upon an (approximately) unbiased point estimator $\hat{\theta}$ and an (approximately) unbiased variance estimator $\hat{V}(\hat{\theta})$. Then a central limit theorem is referred to for the assumption that $\hat{\theta}$ is approximately normally distributed, and it is stated that the interval $\hat{\theta} \pm 1.96\{\hat{V}(\hat{\theta})\}^{1/2}$ covers the true value θ with a probability of approximately 95 %. Sometimes 1.96 is exchanged for the corresponding value taken from the Student's t table with an appropriate number of degrees of freedom.

In an individual survey it is not, however, very easy to establish how accurate this approximation is. It depends on a number of factors such as the type of estimator and design used, the underlying population sampled from, and the sample size. Consequently, there is a great need of increasing our knowledge of the coverage properties of the standard procedures for calculating confidence intervals in different set-ups and of working out simple rules of thumb useful for the survey practitioner.

For in sampling practice the standard approximation fails frequently. One example is sampling from populations of enterprises with very skewed variables such as production, employment, investment, export or import. Another example is small area estimation where most observations are set to zero.

In this paper we study the case of the sample mean under simple random sampling. For this case Erdös and Rényi (1959) and Hajek (1960) developed conditions for the sampling distribution to converge to normality.

Stenlund and Westlund (1975 and 1976), Barrett and Goldsmith (1976) and Hägglund (1978) studied this problem by means of Monte-Carlo experiments.

For populations in which the principal deviation from normality consists of a marked positive skewness, Cochran (1977) suggested the simple rule

$$n > 25G_1^2 ,$$

where n is the sample size and G_1 the usual measure of population skewness defined below. According to Cochran "this rule is designed so that a 95 % confidence probability statement will be wrong not more than 6 % of the time".

Robinson (1978) gave an asymptotic Edgeworth-type expansion for the sum of a simple random sample without replacement from a finite population. The crucial quantities in this expansion are skewness and kurtosis. He showed that, subject to a condition ensuring that the population distribution is "almost continuous", the absolute difference between the distribution function of the sample sum and the asymptotic expansion is bounded by a term containing the absolute fifth moment of the population distribution.

It would, perhaps, be possible to base a rule of thumb on this expansion, although it would have to be quite complicated, as it has to take into account both the skewness and the kurtosis as well as the absolute fifth moment of the population distribution. It would also have to exclude the lattice cases which sometimes occur in practice.

In this paper a simpler approach is used, inspired by Cochran's rule and Höglund (1978), who has derived the following remainder term estimate (the formula is slightly manipulated algebraically to serve our purpose):

$$\left| F(t) - \Phi\left[\frac{t-n\mu}{\sigma\sqrt{n(1-f)}}\right] \right| \leq \frac{CG_2}{\sqrt{n(1-f)}} , \text{ where (1.1)}$$

F is the distribution function of the sum of a sample of n units among the N population units $(x_1, x_2, \ldots, x_N)$, Φ is the standard normal distribution function, μ is the population mean, σ is the population standard deviation, $f = n/N$, C is an absolute constant (Quine (1985) shows that $C \leq 145$) and

$$G_2 = \frac{\sum_{j=1}^{N} |x_j-\mu|^3}{N\sigma^3} . \text{ From above, we have}$$

$$G_1 = \frac{\sum_{j=1}^{N} (x_j-\mu)^3}{N\sigma^3} .$$

We notice three things about (1.1):

i) The deviation from normality is bounded by a term containing the factor G_2, the standardized absolute third moment.

ii) The formula is symmetric in n and (N-n), indicating that the accuracy of the normal approximation could be expected to be equally good for these two sample sizes. In fact, as pointed out by Plane and Gordon (1982), the sampling distributions of the sample mean of n and (N-n) units are mirror images of each other except for a scale change. For this reason N/2 is the sample size where the sampling distribution is closest to the normal.

iii) If we wish to have an upper limit to the absolute difference (called ε) of formula (1.1) we obtain the condition $CG_2/\sqrt{n(1-f)} < \varepsilon \leftrightarrow n(1-f) > C^2G_2^2/\varepsilon^2$ or, if we consider large populations and set $K=C^2/\varepsilon^2$, $n > KG_2^2$. This provides a theoretical argument for a rule similar to Cochran's, although the population skewness is replaced by G_2. The constant K depends only on the maximum error allowed in the approximation. Of course, in constructing two-sided confidence intervals we are interested in the difference between the deviations in symmetric pairs of percentiles of the distribution, usually 2.5 and 97.5, and therefore ε is not necessarily equal to the difference between the nominal and actual coverage probability of the confidence interval.

The above arguments provide the logical foundation for the empirical investigations presented in this paper. Here we calculate the exact coverage probabilities of confidence intervals based on the normal and the t-distribution for dichotomous populations, where these probabilities have a simple hypergeometric distribution. No other distributions are known, where these probabilities are easily calculated for arbitrary sample sizes and degrees of skewness. Moreover, there are strong reasons to believe that this distribution, because of its extreme lattice character, represents more or less the worst case. This was actually proved by Esseen (1956) in the i.i.d. case. The t-distribution is studied together with the normal because it is recommended by many textbook authors, although a solid theoretical argument based on a limit theorem is lacking.

The structure of the type of rule of thumb that we investigate is therefore

$$n > K_\alpha G_2^2, \qquad (1.2)$$

the interpretation being that if we know G_2 exactly and are prepared to allow an actual coverage probability of α, we must choose a sample size greater than $K_\alpha G_2^2$. It is studied to what extent the K_α:s are stable for different degrees of skewness of the dichotomous population and for finite realizations of some continuous parametric distributions. Throughout we study confidence intervals with 95% nominal level.

For very skewed populations $G_2 \approx G_1$ and then this rule coincides with Cochran's but at the other extreme, for symmetric populations, $G_1=0$ and Cochran's rule reduces to $n>0$ and is therefore obviously unsuitable. For this reason G_1 is not used in the empirical investigations below. On the contrary $G_2 \geq 1$ for all populations with equality, if and only if, the population is dichotomous and symmetric as shown in Dalén (1985). For reference $G_2=4/\sqrt{2\pi}\approx 1.6$ for the normal distribution and $\sqrt{27/4}\approx 1.3$ for the uniform distribution.

2 The dichotomous population

For the dichotomous population studied the following notations are used:

Value	Number of units in the population	Number of units in the sample
0	N - M	n - m
1	M	m
Total	N	n

The population has the following characteristics: Population mean = μ = P, population variance = σ^2 = P - P^2, $G_1 = (1-2P)/(P-P^2)^{0.5}$ and $G_2 = (1-2P+2P^2)/(P-P^2)^{0.5}$, where P = M/N.

Notice that $G_2 = G_1 + 2P^{1.5}/(1-P)^{0.5}$ so that $\lim_{P\to 0}(G_2-G_1)=0$ and that $G_1=0$ and $G_2=1$ when P=0.5.

The sample has the following characteristics: Sample mean = X = m/n and sample variance = $s^2 = (m-m^2/n)/(n-1)$.

A nominal 95% confidence interval for μ based on the sample outcome would now be

$$\bar{X} - t_{0.975}\, s\sqrt{(1-f)/n} < \mu < \bar{X} + t_{0.975}\, s\sqrt{(1-f)/n},$$

where $t_{0.975}$ is either 1.96 or the corresponding quantity from the t-distribution with (n-1) degrees of freedom.

(Since we are interested in how bad the approximation could be at worst, the continuity correction is not used. If it was, the constants K_α needed would become much lower but would be more difficult to generalize to other types of populations.)

Now, let I_m be the indicator of this confidence interval statement as a function of the sample outcome. That is:

$$I_m = \begin{cases} 1 \text{ for those m where the confidence interval contains } \mu \\ 0 \text{ otherwise} \end{cases}$$

The actual coverage probability (ACP) is now defined as the probability for a sample of a certain size n from our population to produce a confidence interval statement containing μ, that is $ACP(N,M,n) = \sum_{m=0}^{n} I_m p(m)$, where $p(m) = \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}$ according to the hypergeometric distribution.

Computer programs were written which computed these probabilities for various combations of N, n and P. The programming language was SIMULA, and the IMSL procedures MDBIN, MDHYP and MDSTI were used.

In diagram 1, a typical example is given of how the ACP varies with n up to N/2. We see that the ACP does not increase monotonously with n. Typically there are intervals of increase (shorter and shorter as n increases), followed by downward jumps. This is of course due to the discontinuous character of the population studied. Up to 85-90% the increase is rapid, but after that there are oscillations around a mean, which gets closer and closer to 95%.

3 Average ACP

ACP is a measure of the goodness of the normal or t-approximation. If the nominal confidence level is 95% we consider the approximation to be good if we can count on a coverage probability α sufficiently close to 95%.

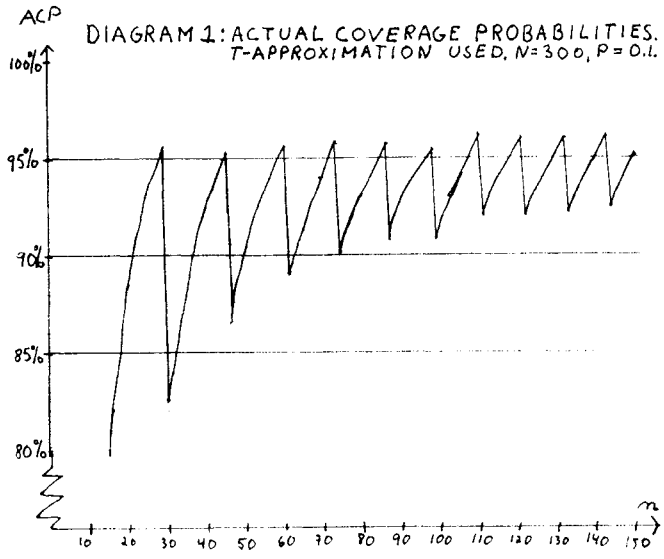

DIAGRAM 1: ACTUAL COVERAGE PROBABILITIES. T-APPROXIMATION USED. N=300, P=0.1.

However, it is not possible in an individual case to promise a "guaranteed" ACP. This is partly because we do not know the population characteristics exactly but also because of the oscillations in the ACP-level as n varies as shown in diagram 1. A device intended to deal with the latter problem is the concept of an average ACP.

Definition 1: α is an average ACP for a certain sample size n in a certain population if

$$\sum_{j=0}^{s} ACP(n+j)/(s+1) \geq \alpha$$

for all integers s such that $0 \leq s \leq N/2-n$.

Definition 2: For a certain population the sample size n_α required for an average ACP of α is the smallest n for which α is an average ACP.

These two definitions also give a unique value of the constant K in (1.2) for a certain population, namely

$$K_\alpha = n_\alpha / G_2^2 .$$

In table 1 values of these constants are presented for some combinations of N and P including the binomial case ($N=\infty$) and for α=94%. Two comments to the calculation of this tables should be made:

i) In the binomial case definition 1 could not be applied exactly, since N is infinite. Instead we had to choose a maximum sample size up to which we calculated the ACP and which was equated to N/2 in definition 1. This sample size was in all cases greater than $100G_2^2$.

Table 1: Constants for an average ACP of 94%.

P	Normal approximation N =					t approximation N =				
	500	1000	2000	5000	∞	500	1000	2000	5000	∞
0.01	-	(5.1)	-	18.6	22.5	-	(5.1)	-	18.6	22.5
0.02	(5.2)	-	16.4	20.0	23.3	(5.2)	-	16.4	19.9	23.2
0.03	(8.2)	(16.2)	18.6	22.1	24.0	(8.2)	13.9	18.6	22.1	21.3
0.04	(11.3)	17.3	20.7	22.7	23.1	(11.3)	16.0	19.6	20.0	23.1
0.05	(14.4)	16.9	20.1	23.4	23.8	(14.4)	16.9	19.0	19.4	23.8
0.1	19.4	24.8	25.3	23.8	25.8	19.4	20.1	20.5	22.4	22.5
0.15	24.1	23.2	26.9	25.3	27.6	18.6	17.2	19.3	21.4	19.8
0.2	31.8	28.4	30.8	30.8	26.3	17.0	19.4	15.2	22.1	17.6
0.25	25.9	24.0	31.7	36.5	36.5	11.0	13.9	19.2	14.4	16.8
0.3	30.6	28.7	36.8	33.7	31.8	8.7	11.9	11.9	15.0	15.0
0.35	29.1	35.2	30.6	26.8	43.7	13.0	10.0	10.0	10.0	10.0
0.4	43.5	30.2	28.4	28.4	30.2	6.2	6.2	6.2	6.2	6.2
0.45	26.2	26.2	36.9	36.9	34.9	5.8	5.8	5.8	5.8	5.8
0.5	36	37	29	34	34	14	9	9	9	9

ii) In those cases where $N/2-n_\alpha < 50$ we have put brackets around the value of the constant. This is because those values may be considered to be accidental from a global point of view. (The number 50 is, of course, to a certain extent arbitrary.)

Table 1 and corresponding tables for other α-levels not published here are summarized by the following table, showing the range of the constants for each level.

α	Normal	t
85	1.6 - 5.3	1.6 - 4.4
90	1.9 - 8.4	1.9 - 5.1
93	7.9 -19.5	4.9 -12.1
94	16.4 -43.7	5.8 -23.8
94.5	25.2 -81.5	5.8 -46.5

We notice that, for the higher α-levels and less skew populations, the t-approximation gives much smaller constants. This provides an empirical argument for the use of this approximation instead of the normal one.

4 Almost continuous populations

The K_α-values obtained in the dichotomous case were also tried on populations of an "almost continuous" type. The K-values chosen were 3, 5, 11, 20 and 40, corresponding roughly to the five α-levels of the summary table above.

It is to be expected that the convergence rate be more rapid for such populations than for the dichotomous population with its pronounced lattice character. If this is correct the α-levels should generally be exceeded if we choose the above K-values.

The populations used were based on fixed percentiles of the beta, lognormal, power function and Weibull distributions. For each of these four distributions, six different finite populations were generated with different degrees of skewness by taking the percentiles from 0.001 to 0.999 with intervals of 0.002, making the population size 500. The reference used for these distributions was Patel et al (1976).

For each population five different sample sizes were chosen so that they corresponded as closely as possible to the K-values above. This means that n was chosen so that $n \geq KG_2^2 > n-1$ for values of K of 3, 5, 11, 20 and 40 respectively.

For every sample size 1000 simple random samples without replacement were made. For each sample the population mean was estimated and a confidence interval based on the sample standard deviation and the t-distribution was calculated. The number of cases when this interval covered the true population mean was counted. This figure divided by 1000 became our estimated actual coverage probability (EACP). EACP is of course stochastic in this case with a standard error of 0.7% to 1.1% when the ACP ranges from 95% to 85%.

In table 2 the outcome of these Monte-Carlo trials is presented in terms of the EACP for a certain combination of population and sample size for the lognormal case. This table and corresponding tables for the other populations show, as expected, that in almost all the cases the α-levels obtained from the studies of the dichotomous population are exceeded, for α = 85 and 90% by large margins. The convergence rate up to 90-92% seems in general to be rapid. Only in 5 cases out of 90 are the presupposed levels not obtained (those cases are indicated with an asterisk). The EACPs are in these cases 0.1-0.4% below the expected level. One case is for α=93% (0.1 below), three cases are for α =94% (0.1-0.3 below) and one case is for α = 94.5 (0.4 below). The deviations may very well be entirely due to the stochastic effect of the Monte-Carlo trials.

Table 2: EACPs for 1000 random samples from populations based on the lognormal distribution.

K	α-level(%)	G_2=1.597	n	G_2=1.982	n	G_2=2.943	n	G_2=3.883	n	G_2=4.695	n	G_2=6.331	n
3	85	95.9	8	92.7	12	92.1	26	89.6	46	90.6	67	89.5	121
5	90	94.7	13	93.6	20	91.4	44	92.8	76	91.8	111	91.7	201
11	93	95.9	28	95.6	44	94.4	96	95.2	166	94.5	243		
20	94	94.8	51	95.5	79	93.9*	174						
40	94.5	95.4	102	96.2	158								

5 Conclusions

Our empirical investigations into the problem of how large the sample size must be to allow a standard 95% confidence interval to be calculated for a simple random sample from a finite population support the following tentative conclusions:

i) When the difference is of any significance, the confidence interval should be based on the Student's t distribution with (n-1) degrees of freedom, making the convergence rate more rapid.

ii) A rule of thumb of the Cochran type (1.2) is useful to the practicing statistician, if he has a reasonably good knowledge of G_2. A choice of K = 20 should in most cases allow him to count on an ACP of 94% for a nominal 95% confidence interval. For "almost continuous" populations, a K greater than 3 should be enough for an ACP of around 90%. For some symmetric populations, i.e. those close to uniform, even more liberal limits will do.

The rule of thumb could be used a priori to assist a decision on sample size. If our knowledge of G_2 is insufficient before the survey, the rule could be used to evaluate the quality of a standard confidence interval based on the sample data after the sample is drawn.

6 Some comments for the practical application

The practical application of a rule of thumb like (1.2) raises a number of questions, two of which are commented on below.

I G_2 is not known. In practice no population parameters are known exactly and G_2 is no exception. Estimating G_2 from the sample is not easy. No unbiased estimator is known and the corresponding sample quantity

$$g_2 = \sqrt{n}\ \sum_{i=0}^{n}|x_i-\bar{x}|^3 \ / \{\sum_{i=0}^{n}(x_i-\bar{x})^2\}^{3/2} \leq (n^2-2n+2)/n\sqrt{n-1} < \sqrt{n}$$

and therefore underestimates G_2 with probability one as soon as $n < G_2^2$, as shown in Dalén (1985).

Moreover, if the population consists of two subsets A and B where B contains a few large-value units with a small probability of showing up in a sample of size n, and G_2' calculated over AuB is much greater than G_2 calculated over A, then in most sample outcomes we would in a sense estimate G_2' rather than G_2 and our rule of thumb based on g_2 instead of G_2 would become seriously misleading.

It is therefore necessary to know more about G_2 than what can be inferred from a sample. If, for example, we know that the range of population values is not much greater than the range of sample values we would be on safer ground using g_2 or a similar estimator.

II Stratified samples. In the presence of a skewed population, estimation by the sample mean under simple random sampling is certainly not the best strategy available. In such situations the prevailing strategy at central statistical offices is stratified random sampling using the weighted mean with the stratum sizes as weights. However, due to the lack of a sufficiently good auxiliary variable we sometimes end up with very skewed subpopulations in many strata. It then becomes an issue when the normal approximation is reasonable in stratified samples.

Some empirical studies of this problem have been done, but due to the many dimensions involved (number of strata, stratum sizes, sample sizes, variances and degrees of skewness in each stratum), results are difficult to present in a systematic way. There are indications that a rule like $n > K_\alpha \Sigma w_i G_{2i}^2$, where summation is over strata, G_{2i} is G_2 in stratum i and w_i are

properly chosen weights such that $\Sigma w_i = 1$, would work satisfactorily. If a Neyman allocation is used, $w_i = N_i\sigma_i / \Sigma N_i\sigma_i$ seems to work in many cases. (N_i is the size and σ_i^2 the variance of stratum i.)

Acknowledgements

This paper has been part of a Statistics Sweden project on sampling and estimation led by Bengt Swensson. He, Carl-Erik Särndal, Jan Hagberg and Thomas Höglund have also contributed to this paper with helpful comments and suggestions.

References

Barrett, J.P. & Goldsmith, L. (1976), When is n sufficiently large?, The American Statistician, Vol 20, No 2.

Cochran, W.G., (1977), Sampling Techniques, third edition, Wiley.

Dalén, J., (1985), Bounds on standardized moments in samples and finite populations, Stockholms universitet, Statistiska institutionen, 1985:4.

Erdös, P. & Rényi, A., (1959), On the central limit theorem for samples from a finite population. Publ. Math. Inst. Acad. Sci. 4,

Esseen, C.G., (1956), A moment inequality with an application to the central limit theorem, Skandinavisk aktuarietidskrift, 1-2.

Hajek, J. (1960), Limiting distributions in simple random sampling from a finite population, Publ. Math. Inst. Hung. Acad. Sci. 5, 361-374.

Hägglund, G., (1978), Monte Carlo studies of the normal approximation in stratified samling (In Swedish), University of Uppsala.

Höglund, T., (1978), Sampling from a finite population. A remainder term estimate. Scand J Statist 5:69-71.

Patel, J.K., Kapadia, C.H. & Owen, D.B., (1976), Handbook of statistical distributions, Marcel Dekker, Inc.

Plane, D.R. & Gordon, K.R., (1982), A simple proof of the non-applicability of the central limit theorem to finite populations. The American Statistician, Vol. 36, No. 3, Part 1.

Quine, M.P., (1985) Remainder term estimates in a conditional central limit theorem for integer-valued random variables, Journal of the Australian Mathematical Society

Robinson, J., (1978), An asymptotic expansion for samples from a finite population, The Annals of Statistics, Vol 6, No. 5.

Stenlund, H. and Westlund, A., (1975), A Monte Carlo study of simple random sampling from a finite population, Scand J Statist 2:106-108.

Stenlund, H. and Westlund, A., (1976), On the asymptotic normality of the mean estimator, Scand J Statist 3:127-131.