A SEQUENTIAL PROCEDURE FOR ESTIMATING THE SAMPLE SIZE NEEDED FOR NORMAL APPROXIMATION IN FINITE POPULATION SAMPLING

N. M. Lalu and P. Krishnan
University of Alberta, Canada

1. INTRODUCTION

It is often assumed that "the samples in surveys are often large enough so that an estimate made from them is approximately normally distributed" (Cochran, 1963:11). Cochran, while recognizing Hajek's (1960) results on the necessary and sufficient conditions under which the distribution of the sample mean tends to normality, for sampling with replacement from finite populations, bewails the "imposing body of knowledge on finite populations leaves something to be desired." In social sciences, the populations from which samples are drawn are generally marked by high degree of non-normality. This is true of some biological populations as well. To date, no safe general rule has been developed stipulating the sample size for normal approximation. Cochran (1963:41) notes a crude rule for populations in which the principal deviation from normality consists of marked positive skewness, viz.

$n>25\ G_1^2$ where G_1 is Fisher's measure of skewness, i.e., $G_1 = E(y_i - \bar{Y})^3/\sigma^3$.

An attempt is made here to develop an easier rule through a sequential procedure. We have not been able to prove our general result mathematically; simulation methodology has been utilized to verify this result.

2. GENERAL RESULT

We discuss here the sample size needed for ensuring the normality of the sample mean. Let us consider the universe to be large. The sampling is done sequentially; one unit is picked at random at a time. Let $\bar{X}_n$ be the sample mean at the nth stage (n=1,2,...). If $|\bar{X}_n - \bar{X}_{n+1}| \leq \varepsilon$ for a given ε at the $(n+1)^{st}$ stage (n=1,2,3,...), we stop the experiment. We claim that this sample size would permit the sample mean to be approximately normally distributed.

3. RESULTS FROM SIMULATION EXPERIMENTS

Since a mathematical proof is not immediately feasible, we present results from simulation experiments. The frequency distributions of the universe, assumed to be large, have been taken as (a) rectangular and (b) chi square with small degrees of freedom. The value of ε has been varied for all these instances and 100 simulations performed for each population and for a given ε. The findings are discussed below.

i. Uniform (rectangular) distribution

The density of this distribution is given by $f(x,a) = \frac{dx}{a}\ 0 \leq x \leq a$. We take a=10. Table 1 presents the values of ε, mean of sample means, and the average sample size in 100 simulations. In order to determine whether normality holds for the distribution of the means, we computed β_1 and β_2 of the empirical sampling distributions. These are shown in Table 2.

From these two tables we observe that the sampling distribution is centered to the theoretical value of five and is, roughly speaking, approaching the normal distributuon. We recognize that some positive skewness is present in the distribution generated by a small number (here 100) of replications. The kurtosis coefficient is close to three in one instance (ε=.05) and in all others greater than three depicting leptokurticity. We may remember the rule of thumb that the average of 12 uniformly distributed random variables is normal. This is also verified here.

ii. Chi square distribution

A chi square distribution with five degrees of freedom is considered in this section. We know the mean value for this distribution is five and the variance ten. As discussed earlier, for different values of ε, simulation experiments were carried out, the results of which are shown in Table 3. The central value of the empirical distribution of the sample mean is very close to the theoretical value. Now let us see how close it is to the normal distribution. Measures of skewness and kurtosis are shown in Table 4 for different values of ε. The degree of the skewness is indeed small in almost all the replications. The kurtosis measure is close to three in the majority of the cases. The closeness to normality is clear.

iii. Normal population

The means of samples from a normal population, we know, have a normal distribution. We need not have to present similar results on sampling from a normal universe. But we show in Table 5 β_1 and β_2 measures of the empirical sampling distribution to underscore the fact that due to sampling fluctuations β_2 may not always be close to three and β_1 close to zero. In view of the fact that we had only 100 replications for all populations for a given ε, the small size might have led to greater fluctuations.

4. SOME REMARKS

Obviously the sequential experiment has to terminate. Using Chebyshev's inequality, we can comment on the relationship between E(n) on the one hand and σ^2 and ε on the other. ε indicates how close we are covering the population mean. Since

$$\bar{X}_n - \bar{X}_{n+1} = \frac{1}{n+1}[\bar{X}_n - X_{n+1}]$$

we see $E[\bar{X}_n - \bar{X}_{n+1}]|n = 0$ and therefore,

$$E[\bar{X}_n - \bar{X}_{n+1}] = 0$$

$$V[\bar{X}_n - \bar{X}_{n+1}]|n = \frac{\sigma^2}{n(n+1)}$$

$$V[\bar{X}_n - \bar{X}_n+1] = \sigma^2 E[\frac{1}{n(n+1)}]$$

By Chebyshev's inequality,

$$P\{|\bar{X}_n - \bar{X}_{n+1}|<\varepsilon\}|_n \geq 1 - \frac{\sigma^2}{\varepsilon^2}[\frac{1}{n(n+1)}]$$

If σ^2 is large, for fixed ε, n and therefore E(n) have to be large in order that the probability may be close to unity. If ε is small, similarly E(n) has to be large.

REFERENCES

Cochran, W. G.
1963 Sampling Techniques (2nd Edition), New York: Wiley

Hajek, J.
1960 "Limiting distributions in simple random sampling from a finite populations:, Pub. Math. Inst. Hungarian Acad. Sci., 5:361-374. (Quoted in Cochran)

TABLE 1

Average Value of Sample Means and Sample Size in Sampling from a Uniform Distribution $f(x)dx = \frac{dx}{10}$

ε	Mean of Sample Size	Mean Sample Size
.01	5.016	27.8
.02	4.783	19.7
.03	4.843	16.5
.04	4.943	12.3
.05	4.923	12.6
.10	4.773	9.2

Source: Simulation experiments

TABLE 2

β_1 and β_2 of the Empirical Sampling Distribution of the Mean from a Uniform Distribution with Parameter a=10

ε	β_1	β_2
.01	0.0875	4.8597
.02	0.4923	5.8090
.03	1.7816	7.8218
.04	0.0075	4.9135
.05	0.0913	3.0183
.10	0.7861	4.9919

Source: See Table 1

TABLE 3

Mean of Sample Means and Mean Sample Size in Sampling from a Chi Square Distribution with Five Degrees of Freedom

ε	Mean of Sample Means	Mean Sample Size
.01	4.920	23.9
.02	4.976	17.5
.03	4.850	15.6
.04	4.975	13.1
.05	5.020	12.0
.10	4.667	7.9

Source: See Table 1

TABLE 4

β_1 and β_2 of the Empirical Distribution of Sample Means in Sampling From a Chi Square Distribution with Five Degrees of Freedom

ε	β_1	β_2
.01	.0711	6.7718
.02	.0053	3.9950
.03	.1863	5.6741
.04	.0364	2.9090
.05	.0295	2.9307
.10	.0510	3.0902

Source: See Table 1

TABLE 5

β_1 and β_2 of the Empirical Sampling Distribution of Means from a Normal Population with $\mu=5$ and $\sigma=1.5$

ε	β_1	β_2
.01	.1023	8.9240
.02	.1531	3.2461
.03	.1061	3.1788
.04	.1202	3.7429
.05	.1280	3.2122
.10	.00002	2.6556

Source: See Table 1