THE PROBLEM OF NORMALITY IN STATISTICAL INFERENCE FOR SAMPLE SURVEYS

J. C. Koop, Research Triangle Institute

SUMMARY

The sampling distribution of an estimator that subsumes all the estimators considered in the literature of sample surveys is defined. A new expression for kurtosis which is much simpler than the usual one based on the central moments or the conditional central moments is presented. In general the attainment of approximate normality of distributional form or otherwise, depends on all features of the sampling design of which overall sample size is one. More caution, and less reliance on asymptotic normality, is suggested by this result in making probability statements about the characteristics of a finite universe.

1. INTRODUCTION

In large-scale sample surveys many hundreds of ultimate sampling units are observed and there is a view that because the total sample size is large, the distribution of an estimate, particularly when it is linear in the variates involved, will be somewhat like the normal distribution thereby justifying its use, for example, in setting confidence limits. This traditional view persists despite Cochran's (1963, pp. 38-43) note of caution in his famous text book. His illuminating comments on the work of Hájek (1960) in regard to the conditions that are necessary and sufficient to ensure that the sample mean, in simple random sampling from a finite universe, tends to normality with increasing sample size are interesting. The substance of his remarks is that subasymptotic normality is more relevant. These remarks would also apply to more recent work, e.g., Fuller (1975) and Krewski and Rao (1978).

The work of Hastings (1974) for one-stage stratified sampling, and Koop (1963) for two-stage sampling, both with equal probabilities and without replacement, shows that normality can be seriously disturbed, as judged by the values of the skewness and kurtosis coefficients of the relevant linear estimators.

The question which arises is whether or not such important reservations hold for all sampling designs and all estimators, linear or nonlinear, and whatever the total sample size, short of the total number of ultimate sampling units in the finite universe. The main purpose of this note is to answer this question.

2. THEORY AND RESULTS

For the purpose in mind, we define the sampling distribution of an estimator of a universe value or estimand, both of which subsume all the estimands and their corresponding estimators considered in the literature of sampling theory. Note that this would include linear and nonlinear estimators, e.g., studentized or t-like statistics in multi-stage sampling. Before we do so we present some basic definitions and notation.

2.1. Definitions and Notation

There is a finite universe U consisting of N different identifiable units where u_i is the ith unit, i.e.,

$$U = \{ u_i : i = 1, 2, . . ., N \} \qquad (2.1.1)$$

with a set of N corresponding vectors of ℓ real-valued components, (x, y, z, . . .) for each unit, i.e.:

$$\{ (x_i, y_i, z_i, . . .) : i = 1, 2, . . ., N \} . \qquad (2.1.2)$$

The units of U may be clustered in a hierarchy of units, but for the sake of generality we do not specify the structure of U. The frame F identifies the units of U and is assumed to be perfect.

We have a physical randomization procedure R (e.g., tested random numbers and playing cards) for selecting samples from U through the frame F and according to a set of rules H (for drawing the units) and a probability system P that defines the respective selection probabilities of the units and which in turn is implemented by R. See Koop (1979) for explanations about these definitions.

The sampling procedure for selecting samples from U is defined by the combination

$$(P, R) . \qquad (2.1.3)$$

Through the application of (P, R) a sample of distinct units, i.e.,

$$s = \{ u_i, u_j, . . . , u_m \} \qquad (2.1.4)$$

is eventually realized with probability p(s). The number of units in s, i.e., the effective sample size, is n(s). The entire possible collection of s is designated

$$\mathcal{S} = \{ s: s \subset U \} . \qquad (2.1.5)$$

The probability p(s), regarded as a function defined on $\mathcal{S}$, is also called the sampling design.

We are interested in estimating A(U) which is a real-valued function of the variate values of U given by (2.1.2).

To estimate A(U) we define a real function a(s), defined for all $s \varepsilon \mathcal{S}$, that is a function of all the variate values of the distinct sample s and with a certain number of undetermined constants and also such that

$$a(s)_{min} < A(U) < a(s)_{max} \qquad (2.1.6)$$

The constants in a(s) are determined by some method. This function naturally represents and subsumes all the estimators considered in the literature of sample surveys.

By definition, the sampling distribution of a(s), i.e., the distribution generated by the randomization procedure R, is given by the entire set of such values with their corresponding p(s)-values namely,

$$\{(a(s),\ p(s)):\ s\varepsilon \mathcal{S} \text{ and } a(s)_{\min} < A(U) < a(s)_{\max}\}. \qquad (2.1.7)$$

Note that (i) $p(s)$, for all $s\varepsilon\mathcal{S}$, is calculable as a numerical value, and in simple random sampling it is equal to $1/\binom{N}{n}$, and (ii) the distribution (2.1.7), unlike its classical analogues (density functions), is not characterized by parameters, unless by convention we choose to regard $A(U)$ as a parameter.

The properties of this distribution may be described by its first moment and its second and higher central moments, but particularly by its skewness and kurtosis coefficients. Thus we have the first moment

$$E\{a(s)\} = \sum_{s\varepsilon\mathcal{S}} p(s)\ a(s), \qquad (2.1.8)$$

and the higher central moments

$$\mu_r\{a(s)\} = \sum_{s\varepsilon\mathcal{S}} p(s)\ [a(s) - E\{a(s)\}]^r, \qquad (2.1.9)$$

for r = 2, 3, 4, ... By definition

$$\beta_1\{a(s)\} = \mu_3^2\{a(s)\} \ /\ \mu_2^3\{a(s)\}, \qquad (2.1.10)$$

and

$$\beta_2\{a(s)\} = \mu_4\{a(s)\} \ /\ \mu_2^2\{a(s)\}\ . \qquad (2.1.11)$$

If this generalized distribution has an approximate normal form, and this may be the case when $n(s)$, its effective sample size, is sufficiently large, then $\beta_1\{a(s)\}$ should be close to zero and $\beta_2\{a(s)\}$ should be approximately 3. Let us see whether this is so.

2.2. Results

We shall deal with the kurtosis coefficient first as it is relatively more tractable. By Lagrange's identity we find

$$\sum_{s\varepsilon\mathcal{S}} p(s) \sum_{s\varepsilon\mathcal{S}} p(s)\ [a(s) - E\{a(s)\}]^4$$

$$= \Big[\sum_{s\varepsilon\mathcal{S}} p(s)[a(s) - E\{a(s)\}]^2\Big]^2$$

$$+ \sum_{i>j} \Big([a(s_i) - E\{a(s)\}]^2\ \{p(s_i)p(s_j)\}^{\frac{1}{2}}$$

$$- [a(s_j) - E\{a(s)\}]^2\ \{p(s_j)p(s_i)\}^{\frac{1}{2}}\Big)^2. \qquad (2.2.1)$$

In (2.2.1) we have attached subscripts i and j to s; these subscripts should run from 1 to $\mathcal{C}$, where $\mathcal{C}$ is the number of possible samples in the set $\mathcal{S}$. Thus the summation indicated by $\sum_{i>j}$ is over $\mathcal{C}(\mathcal{C}-1)/2$ terms.

On the right-hand side of (2.2.1) the second term is zero if and only if

$$[a(s) - E\{a(s)\}]^2 = \text{a constant for all } s\varepsilon\mathcal{S}\ . \qquad (2.2.2)$$

Dividing both sides of (2.2.1) by $\mu_2^2\{a(s)\}$ we find

$$\beta_2\{a(s)\} = 1 + \sum_{i>j} p(s_i)p(s_j)[a(s_i) - a(s_j)]^2 .$$

$$[a(s_i) + a(s_j) - 2E\{a(s)\}]^2 / \mu_2^2\{a(s)\}. \qquad (2.2.3)$$

Note that the term following unity in (2.2.3) is always positive, except when (2.2.2) holds, in which case it is zero. Then $\beta_2\{a(s)\} = 1$; this result is certainly not trivial.

For example when

(i) $\mathcal{C} = 4M$, where M is a large postive integer,

(ii) $p(s) = 1/4M$ for all $s\varepsilon\mathcal{S}$,

(iii) $a(s_i) = h$, $i = 1, 2, \ldots, M$ and h is a real number,

(iv) $a(s_i) = k > h$, $i = M+1, M+2, \ldots, 2M$,

(v) $a(s_i) = h+(k-h)/g$, $g>1$ and $i = 2M+1, \ldots, 3M$, and

(vi) $a(s_i) = k+(k-h)/g$, $i = 3M+1, \ldots, 4M$,

then after computations according to (2.1.11) we find

$$\beta_2\{a(s)\} = 1 + \frac{4}{1+g^2} - \frac{4}{(1+g^2)^2}\ . \qquad (2.2.4)$$

Note that $\beta_2 \to 1$ when $g \to \infty$. For values of g less than 10 we find:

g	$\beta_2\{a(s)\}$
1.5	1.85
2	1.64
3	1.36
4	1.22
4.899	1.15
9.95	1.04 .

Thus when g is a little more than 10, β_2 is virtually 1. The value of this example lies in the indications that it gives for sampling designs that produce discrete distributions which are somewhat rectangular in form. Such distributions would have β_2-values between 1 and 3.

We now consider the skewness coefficient. An expression that is suggestive of the values that it can assume cannot be derived as in the case of β_2. However, we can obtain its upper and lower bounds.

From Pearson (1916) we find that

$$\beta_2\{a(s)\} \geq \beta_1\{a(s)\} + 1. \qquad (2.2.5)$$

For ready reference see David and Barton (1962). If we denote by γ the expression following unity on the right-hand side of (2.2.3), then the limiting inequality for the skewness is given

$$-\gamma \leq \sqrt{\beta_1}\{a(s)\} \leq \gamma\ , \qquad (2.2.6)$$

in view of (2.2.5). Note that β_1=0 in the foregoing example.

From (2.2.3) and (2.2.6) we see that the values of the skewness and kurtosis coefficients for any finite universe, any sampling design and any estimator can be very different from 0 and 3, the values that would indicate subasymptotic normality[*] for distributions resulting from large-scale sample surveys with many hundreds or thousands of observations.

We also see from (2.2.3) that kurtosis, and therefore also skewness by virtue of (2.2.6), depends principally on the sampling design p and the form of the estimator or statistic a. The effective sample size n(s) only plays an indirect role, implicitly, through the sampling design.

3. COMMENTS

The practical upshot of the foregoing results is that we cannot always guarantee that t-like statistics based on the data of large-scale surveys would always yield reasonably accurate confidence intervals based on normality assumptions. On this question all the anomalies in the empirical work of Frankel (1971), Bean (1975), Neter and Loebbecke (1977), Campbell and Meyer (1978) are explained by the results at (2.2.3) and (2.2.6). With survey data, more caution and more empirical studies are needed before making meaningful probability statements based on normality assumptions.

Again it should be pointed out that with the method of independent replication exact probability statements can be made about the median without any assumptions, and for other statistics such statements can be made with assumptions or qualifications that are not restrictive (Koop, 1960). See also Murthy (1967). More consideration should be given to this method where in my view the advantages outweigh the disadvantages, taking into account nonsampling errors.

[*] In Koop (1979) it is argued that restrictions on randomization can be injurious to normailty.

REFERENCES

Bean, J. A. (1975). Distribution and properties of variance estimators for complex multistage probability samples. Vital and Health Statistics, Series 2, No. 65, U.S. Government Printing Office, Washington, D.C.

Campbell, C. and Meyer, M. (1978). Some properties of T confidence intervals for survey data. Proc. Sec. Survey Res. Methods, Amer. Statist. Assoc., 437-442.

Cochran, W. G. (1963). Sampling Techniques. John Wiley; New York.

David, F. N. and Barton, D. E. (1962). Combinatorial Chance. Charles Griffin; London.

Frankel, M. R. (1971). Inference From Survey Samples: An Empirical Investigation. Ann Arbor; Institute of Social Relations.

Fuller, W. A. (1975). Regression analysis for sample survey. Sankhya C, 37, 117-132.

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

Hastings, W. K. (1974). Variance reduction and non-normality. Biometrika, 61, 143-149.

Koop, J. C. (1960). On theoretical questions underlying the technique of replicated or interpenetrating samples. Proc. Social Statistics Sec., Amer. Statist. Assoc., 196-205.

Koop, J. C. (1963). On the axioms of sample formation and their bearing on the construction of linear estimators in sampling theory for finite universes, III. Metrika, 7, 165-204.

Koop, J. C. (1979). On statistical inference in sample surveys and the underlying role of randomization. Ann. Inst. Statist. Math., 31(A), 253-269.

Krewski, D. and Rao, J. N. K. (1978). Large sample properties of the linearization, jackknife and balanced repeated replication methods. Proc. Sec. Survey Res. Methods, Amer. Statist. Assoc., 457-461.

Neter, J. and Leobbecke, J. A. (1977). On the behavior of statistical estimators when sampling accounting populations. J. Amer. Statist. Assoc., 72, 501-507.

Murthy, M. N. (1967). Sampling Theory and Methods. Statistical Publishing Society; Calcutta.

Pearson, K. (1916). Mathematical contributions to the theory of evolution XIX: second supplement to a memoir on skew variation. Phil. Trans. Roy. Soc., London, A 216, 429-457.