The Annals of Statistics
1983, Vol. 11, No. 2, 702–706

SUCCESSIVE SAMPLING IN LARGE FINITE POPULATIONS

BY LOUIS GORDON[1]

Energy Information Administration, U.S. Department of Energy

The permutation distribution induced upon a finite population by the order of selection under successive sampling is closely related to the order statistics of independent exponentially distributed waiting times. This characterization is applied to obtain necessary and sufficient conditions for asymptotic normality of the sum of characteristics observed in a successive sample from a finite population. The necessary and sufficient conditions generalize previous results for simple random sampling without replacement, and apply to sampling fractions close to 0 or 1.

1. Introduction and summary. Consider the finite population of labels $U_N = \{1, 2, \cdots, N\}$ with associated characteristics $\{y_{1N}, \cdots, y_{NN}\}$ and positive measures of size $\{x_{1N}, \cdots, x_{NN}\}$. We sample the fraction f_N of the population without replacement according to the following scheme. The first unit to be sampled is taken with probability proportional to the size measures x_j. Subsequent units are taken sequentially from among those not yet sampled with probability proportional to size relative to all units not yet sampled. This sampling scheme is called successive sampling and has been studied by Rosen (1972), Holst (1973) and Raj (1968, page 57). Sen (1979) studies another related successive sampling scheme which results in a sample having a randomly determined number of selected units.

Holst generalizes the successive sampling scheme with fixed sample size as follows. To each unit is assigned a cost of observation $\{c_{1N}, \cdots, c_{NN}\}$. The cost of a census of the full population is $c_{+N} = \sum c_{jN}$. For a given sampling fraction f_N between 0 and 1, we sample successively with probability proportional to the x_j until the total cost of the sample first exceeds $c_{+N} \cdot f_N$.

We refer to the former scheme (which results in a sample of fixed size) as uniform cost successive sampling (UCSS) and the second scheme (in which a cost ceiling is fixed) as variable cost successive sampling (VCSS). Note the UCSS is a generalization of simple random sampling without replacement, and that VCSS is a generalization of UCSS.

Hajek (1960) gives necessary and sufficient conditions that a simple random sample of size Nf_N has a limiting Gaussian distribution centered at its sampling expectation and scaled by its sampling standard deviation. The condition specializes in the case of sampling fractions bounded away from 0 and 1 to the Noether condition, which requires that the range of the centered y_j be of order smaller than $N^{1/2}$. The result is proved necessary and sufficient whenever both Nf_N and $N(1-f_N)$ grow arbitrarily large.

Holst (1973) shows that the Noether condition is sufficient for asymptotic normality under VCSS, and hence under UCSS, when the sampling fraction f_N is uniformly bounded away from 0 and 1, and when both costs and size measures are uniformly bounded away from 0 and infinity. In this paper, we show that when the costs and size measures are bounded away from 0 and infinity, a variant of the Noether condition is necessary and sufficient for asymptotic normality under both UCSS and VCSS. Our theorem generalizes the Hajek result to VCSS, extending the Holst result to sampling fractions close to 0 or to 1, and showing the necessity of the conditions we propose.

Our proof is close in spirit to that of Hajek, in that we realize VCSS in such a manner that the total of characteristics from a VCSS sample is very close to a sum of independent

Received June 1981; revised October 1982.

[1] The opinions and conclusions expressed herein are solely those of the author, and should not be construed as representing the opinions or policy of any agency of the United States Government.

AMS 1980 *subject classification.* Primary 60F05; secondary 62D05.

Key words and phrases. Central limit theorem, Noether condition.

random variables. We then follow Hajek's ideas in using the Lindeberg-Feller theorem to show sufficiency and the Cramer-Levy theorem to show necessity.

2. Preliminaries. In this section we establish the notation and terminology that we use in the remainder of the paper. Wherever convenient and unambiguous, we suppress the explicit dependence upon N. We assume throughout the following regularity conditions:

$$1 \leq c_j \leq a, \tag{2.1}$$

$$1 \leq x_j \leq b, \tag{2.2}$$

$$N^{-1} \sum (y_j - \bar{y})^2 = 1 \quad \text{where} \quad \bar{y} = N^{-1} \sum y_j. \tag{2.3}$$

We define the following scalar quantities: $c_+ = \sum_J c_j$, t_f is the solution to $\sum_j c_j \exp(-x_j t_f) = (1 - f)c_+$,

$$r_f = \sum_J y_j \exp(-x_j t_f), \tag{2.4}$$

$$\bar{y}_f = \{\sum_J y_j x_j \exp(-x_j t_f)\} / \{\sum_i c_i x_i \exp(-x_i t_f)\}, \tag{2.5}$$

$$\sigma_f^2 = \sum_J (y_j - c_j \bar{y}_f)^2 \exp(-x_j t_f)\{1 - \exp(-x_j t_f)\}, \tag{2.6}$$

$$n_f = \sum_J \exp(-x_j t_f). \tag{2.7}$$

We also need the following random quantities: $W_1, \cdots W_N$ are i.i.d. uniform exponential random variables with mean 1,

$$J_j(t) = I_{\{W_j > x_j t\}},$$

$$T_f = \sup\{t \mid \sum_J c_j J_j(t) > (1 - f)c_+\}, \tag{2.8}$$

$$C_f^* = \sum_j c_j J_j(T_f) - (1 - f)c_+,$$

$$R_f = \sum_J y_j J_j(T_f), \tag{2.9}$$

$$M(t; s, m) = \sum (y_j - c_j m)\exp(-s x_j)\{J_j(t)\exp(t x_j) - 1\}. \tag{2.10}$$

Note that T_f is the first time that the total cost of observation equals or exceeds fc_+, so that $C_f^* \leq 0$.

3. The approximation. In this section, we show that the sum of characteristics obtained in a successive sample can be approximated by a sum of independent random variables. The main results are presented in Theorems 1 and 2. Theorem 1, which can be obtained easily by induction on Nf_N, characterizes successive sampling in terms of exponential waiting times. It is stated without proof.

THEOREM 1. *Denote by y_+ the sum of characteristics y_j over the entire population. The random vector*

$$(y_+ - R_{jN})_{j=1}^N$$

has the joint distribution of the N partial sums obtained by successively sampling the entire finite population U_N.

Theorem 2 provides the basic approximation. It is proved by reference to Lemmas 3.1 and 3.2. In order to simplify the notation, we state and prove the approximation results in terms of the successive remainders R_f, rather than in terms of the successive sums of characteristics actually observed.

LEMMA 3.1(a) *Given $\varepsilon > 0$, there exists a constant $K(a, b, \varepsilon)$ for which*

$$P\{|T_f - t_f| > K(a, b, \varepsilon)\, n_f^{-1/2}\} < \varepsilon \text{ whenever } n_f > K^2(a, b, \varepsilon).$$

(b) *Under* VCSS, $(1 - f)^b \leq \exp(-x_j t_f) \leq (1 - f)^{1/b}$.

PROOF. (a) Let d be a scalar with $|d| < 1$. Note that

$$|E(\sum_j c_j J_j(t_f + d)) - (1 - f)c_+| > |d|(1 - f)c_+/2$$

and that

$$\mathrm{Var}\{\sum_j c_j J_j(t_f + d)\} < ae^b(1 - f)c_+.$$

From Chebychev's inequality, $P\{|T_f - t_f| > d\} < 8ae^b/(d^2 n_f)$. Hence, we may take $K^2(a, b, \varepsilon) = 8ae^b/\varepsilon$.

(b) Let $g = \exp(-t_f)$. Because of the regularity assumptions,

$$(1 - f)c_+ > \sum c_j g^b = c_+ g^b, \quad (1 - f)c_+ < \sum c_j g = c_+ g.$$

Hence $(1 - f) < g < (1 - f)^b$. The result follows from regularity condition (2.2).

LEMMA 3.2. (a) *Let* $\mathscr{F}_{tN} = \sigma\{J_{jN}(s) \mid s \le t \text{ and } j = 1, \cdots, N\}$, *then* $\{M_N(t; s, m), \mathscr{F}_{tN}, 0 \le t\}$ *is a martingale.*

(b) $$(R_f - r_f) - M(T_f; t_f, \bar{y}_f) = C_f^* \bar{y}_f + \sum_j (y_j - c_j \bar{y}_f) J_j(T_f)[1 - \exp\{x_j(T_f - t_f)\}].$$

PROOF. Assertion (a) follows immediately from the lack of memory of the exponential distribution. Assertion (b) is an immediate consequence of the definitions (2.8), (2.9) and (2.10).

THEOREM 2. *If* $n_f \to \infty$, *then* $(R_f - r_f) - M(t_f; t_f, \bar{y}_f) = C_f^* \bar{y}_f + O_p(\sigma_f(fn_f)^{-1/2})$.

PROOF. Given $\varepsilon > 0$, choose $K = K(a, b, \varepsilon)$ as in Lemma 3.1, and write $d_N = Kn_f^{-1/2}$. Take N large so that $bd_N < 1$. Then let $I^* = I_{\{|T_f - t_f| < d\}}$, and use Lemma 3.1 and Taylor series to write

$$\begin{aligned}
I^*|R_f - (r_f + M(t_f; t_f, \bar{y}_f) + C_f^* \bar{y}_f)| &< I^*|T_f - t_f| \, |\sum (y_j - c_j \bar{y}_f) x_j J_j(t_f)| \\
&\quad + I^*|T_f - t_f| \sum |y_j - c_j \bar{y}_f| x_j |J_j(t_f + d) - J_j(t_f - d)| \\
&\quad + I^*|T_f - t_f|^2 \sum |y_j - c_j \bar{y}_f| \, x_j^2 J_j(t_f - d) e^{bd} \\
&\quad + I^* \max_{|t - t_f| < d} |M(t; t_f, \bar{y}_f) - M(t_f - d; t_f, \bar{y}_f)| \\
&= I^*|T_f - t_f| \, |Q_1| + I^*|T_f - t_f| Q_2 \\
&\quad + I^*|T_f - t_f|^2 Q_3 + I^* \max_t |Q_4(t)| \\
&< d|Q_1| + dQ_2 + d^2 Q_3 + \max|Q_4(t)|.
\end{aligned}$$

Note that $\bar{y}_f$ was chosen exactly to make $EQ_1 = 0$, and that $EQ_1^2 \le b^2 \sigma_f^2$.

From Lemma 3.1, $1 - \exp(-x_j t_f) > f/b$, so that

$$\sigma_f^2 > \sum (y_j - c_j \bar{y}_f)^2 \exp(-x_j t_f) f/b.$$

Hence $EQ_2 < 2Kbe^b f^{-1/2} \sigma_f$, and $dEQ_3 < Kb^2 e^b f^{-1/2} \sigma_f$.

Because M is a martingale, the martingale maximum inequality applies to Q_4, once we evaluate $EQ_4^2(t_f + d)$. For scalar t, $Q_4(t)$ is a sum of independent random variables having zero mean. Because $EQ_4^2(t_f + d) < 3Kbf^{-1/2} \sigma_f$, the lemma is proved.

4. Necessary and sufficient conditions. In this section, we state and prove analogs to the Noether conditions which apply to both UCSS and VCSS. We first introduce some additional notation which we will use in the statement and proof of the main theorem.

Let $y_j' = y_j - c_j \bar{y}_f$. Denote by R_f', r_f', $\bar{y}_f'$, and σ_f' quantities as defined in (2.9), (2.4), (2.5), and (2.6), with y_j' substituted for y_j. Note that $r_f' = r_f - \bar{y}_f'(1 - f)c_+$, that $\bar{y}_f' = 0$, and that $\sigma_f' = \sigma_f$.

Theorem 3 generalizes the theorem of Hajek (1960). Certain of the technical difficulties

in the statement of the lemmas are necessitated by our need to prove the result for sampling fractions close to unity.

THEOREM 3. *Given regularity conditions* (2.1) *to* (2.3), *and that* $Nf(1-f) \to \infty$, *we may conclude that the pair of statements*

(4.1) $\sigma_f^{-1}(R_f - r_f)$ *is asymptotically standard normal*

(4.2) $\sigma_f^{-1}(R'_f - r'_f)$ *is asymptotically standard normal*

is equivalent to the pair of statements

(4.3) $$\sigma_f^{-1} C_f^* \bar{y}_f \to_P 0$$

(4.4) $$\sigma_f^{-2} \sum_J (y_J - c_J \bar{y}_f)^2 p_J (1-p_J) I_{\{|y_J - c_J \bar{y}_f| > h\sigma_f\}} \to 0$$

for all $h > 0$, *where* $p_J = \exp(-x_J t_f)$.

PROOF. From (2.7), n_f lies between $(1-f)N/a$ and $(1-f)Na$, so that $n_f \to \infty$. Note also that, from (2.8),

(4.5) $$(R_f - r_f) - (R'_f - r'_f) = C_f^* \bar{y}_f.$$

Assume (4.3) and (4.4). From (4.3) and (4.5), we know that the limiting distributions defined in (4.1) and (4.2) are identical, if they exist. From Theorem 2 and (4.3), the limiting distribution of (4.1) is the same as that of $\sum y'_{jN} B_j$, where $B_j = J_j(t_f) - p_j$, and $p_j = \exp(-x_j t_f)$. The latter sum is, however, a sum of independent Bernoulli random variables, centered at expectations, scaled by the sum's standard deviation.

The Lindeberg-Feller theorem assures us that the limiting normal distribution of the sum is standard normal if, for each $h > 0$, the sums in (4.6) below converge to zero. See, for example, Chung (1968, page 187).

(4.6)
$$\sigma_f^{-2} \sum E(y'_j B_j)^2 I_{\{|y'_j B_j| > h\sigma_f\}}$$
$$= \sigma_f^{-2} \sum (y'_j)^2 p_j (1-p_j)[p_j I_{\{|y'_j| p_j > h\sigma_f\}} + (1-p_j) I_{\{|y'_j|(1-p_j) > h\sigma_f\}}].$$

The latter sum converges to 0 for all h if and only if condition (4.4) is satisfied. Hence we have shown that (4.3) and (4.4) imply (4.1) and (4.2).

We now prove the converse. Assume (4.1) and (4.2). Note that $C_f^* \leq 0$. Because of (4.5) and the convergence in (4.1) and (4.2) to identical limiting laws, (4.3) follows. Choose and fix $h > 0$, and let

$$R_N^{**} = \sigma_f^{-1} \sum y'_j B_j I_{\{(y'_j)^2 p_j (1-p_j) > h\sigma^2\}}$$

and

$$R_N^* = \sigma_f^{-1} \sum y'_j B_j I_{\{(y'_j)^2 p_j (1-p_j) \leq h\sigma^2\}}.$$

Because (4.3) is true, $R_N^{**} + R_N^*$ has a limiting standard normal distribution. Further, the summands are stochastically independent, with variances bounded by 1. Hence, by the Cramer-Levy theorem, we may assume without loss of generality that both summands converge in law to normal distributions. See, for example, Feller (1971, page 525).

Because R_N^{**} has at most $2^{1/h}$ atoms, we may conclude that its limiting law has at least one atom of mass greater than $2^{-1/h}$. Therefore, its limiting Gaussian law must place all its mass at 0.

Hence, R_N^* converges in law to a standard normal distribution. If R_N^{**} were not eventually 0, R_N^* would have variance bounded above by $1-h$, and would converge in law to a standard normal distribution on some subsequence. This contradicts the Fatou lemma applied to the variances of the R_N^*. Hence R_N^{**} must eventually equal 0. The summands

are therefore individually negligible, and so the Lindeberg-Feller condition must be satisfied. Therefore, the sums (4.6) converge to 0 for each $h > 0$. It follows that condition (4.4) holds, and we have shown that (4.1) and (4.2) imply (4.3) and (4.4).

Note that condition (4.3) is essentially relevant only to VCSS with very large sampling fractions. Under UCSS, we always can make C_f^* identically 0 by choosing the sampling fraction f_N so that Nf_N is an integer. Under VCSS, $\bar{y}_f/\sigma_f$ converges to 0 whenever $Nf(1-f)^b \to \infty$ and the coefficient of variation of the y_j/c_j's is bounded away from 0. The former condition is trivially established when the sampling fraction is bounded away from 1.

Note also, that as in Hajek (1960), if the sampling fractions f_N lie between ε and $1-\varepsilon$, for some $\varepsilon > 0$, then condition (4.4) is equivalent to $\max |y_j - c_j\bar{y}_f|/\sigma_f \to 0$. This condition is the obvious analog to the Noether condition in the case of VCSS.

REFERENCES

CHUNG, K. L. (1968). *A Course in Probability Theory.* Harcourt, Brace and World, New York.

FELLER, W. (1971) *An Introduction to Probability Theory and Its Applications II.* Wiley, New York.

HAJEK, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publ. Math. Inst. Hungar. Acad. Sci.* **5** 361–374.

HOLST, L. (1973). Some limit theorems with applications in sampling theory. *Ann. Statist.* **1** 644–658.

RAJ, D. (1968). *Sampling Theory.* McGraw-Hill, New York.

ROSEN, B. (1972). Asymptotic theory for successive sampling with varying probabilities without replacement, I and II. *Ann. Math. Statist.* **43** 373–397; 748–776.

SEN, P. K. (1979). Invariance principles for the coupon collector's problem: a martingale approach. *Ann. Statist.* **7** 372–380.

ENERGY INFORMATION ADMINISTRATION
U.S. DEPARTMENT OF ENERGY
WASHINGTON, D.C. 20858