The Annals of Statistics
1984, Vol. 12, No. 4, 1536–1542

THE NATURE OF SIMPLE RANDOM SAMPLING[1]

BY S.K. MITRA AND P. K. PATHAK

Indian Statistical Institute Delhi Centre[2,3], Queen's University[3] and University of New Mexico[3]

Dedicated to Colin R. Blyth on his sixtieth birthday.

As an estimator of the population mean, the sample mean based only on the distinct units possesses a remarkable invariance property. Under three forms of simple random sampling, viz. simple random sampling without replacement (SRSWOR), simple random sampling with replacement (SRSWR), and fixed cost simple random sampling (SRSFC), it is admissible and unbiased; and asymptotically normally distributed if and only if the Erdös–Rényi–Hájek condition is satisfied. An important implication of this invariance is that for estimating the population mean, these forms of simple random sampling are asymptotically equally cost-efficient. However, from a practical point of view SRSFC does seem to provide greater flexibility in large surveys.

1. Introduction. It is well-known that as an estimator of the population mean, the sample mean based only on the distinct units is both admissible and unbiased under three forms of simple random sampling, namely simple random sampling without replacement (SRSWOR), simple random sampling with replacement (SRSWR), and fixed cost simple random sampling (SRSFC) (cf. Basu [1], Hájek [3], Joshi [5], and Pathak [6]). The object of this paper is to show that under these three sampling schemes, it is also asymptotically normally distributed if and only if the Erdös–Rényi–Hájek condition is satisfied. In broad terms, the result on asymptotic normality can be stated as follows.

THEOREM 1.1. *Consider a sequential sampling scheme under which units are selected with equal probabilities either with or without replacement at each draw from a given population of size N. Let ν denote the observed number of distinct units drawn in the sample. Suppose that the effective sample size, namely the number ν, and the population size N approach infinity so that*

$$\lim E\nu = \infty \quad \text{and} \quad \lim(N - E\nu) = \infty. \tag{1.1}$$

(*The notion of limit in* (1.1) *can be made precise through a triangular array of populations by indexing both ν and N by a common suffix, say k. For reasons of brevity we have chosen not to use this extra suffix in this paper.*)

Received August 1981; revised June 1984.

[1] This research was supported in part by the National Science Foundation Grant INT-8020450 and by the Natural Sciences and Engineering Research Council of Canada Grant A8470.

[2] S. K. Mitra.

[3] P. K. Pathak.

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

Key words and phrases. Simple random sampling, sample mean, asymptotic normality, the Erdös–Rényi–Hájek condition, fixed cost sampling, cost-adjusted efficiency.

Then the sample mean based on the distinct units, $\bar{y}_\nu$ say, is asymptotically normally distributed with parameters

$$(\bar{Y},\ (1/E\nu - 1/N)(N-1)^{-1} \textstyle\sum (Y_j - \bar{Y})^2)$$

if and only if

$$\lim[\textstyle\sum_{\tau(\varepsilon)} (Y_j - \bar{Y})^2 / \sum (Y_j - \bar{Y})^2] = 0 \tag{1.2}$$

where $\tau(\varepsilon)$ denotes the summation over those population units which satisfy the inequality

$$(Y_j - \bar{Y})^2 > \varepsilon(E\nu/N)(1 - E\nu/N) \textstyle\sum (Y_j - \bar{Y})^2 \tag{1.3}$$

in which Y_j denotes the Y-variate value of the jth population unit, $\bar{Y} = N^{-1} \sum Y_j$ and the sum $\sum$ extends over all the N population units. (In the sequel we shall refer to the condition given by (1.2) *and* (1.3) *as the Erdös–Rényi–Hájek condition.)*

For SRSWOR, this theorem is due to Erdös and Rényi [2] and Hájek [4], and for SRSWR, it is due to Pathak [7]. We now proceed to show that Theorem 1.1 is also valid for SRSFC.

2. Fixed cost simple random sampling (SRSFC). Consider a population $P = (U_1, \cdots, U_j, \cdots, U_N)$ of N units in which the jth unit $U_j = (j, Y_j, C_j)$, where j denotes its label, Y_j its unknown Y-characteristic value under study and C_j the unknown cost of ascertaining the value of Y_j, $1 \le j \le N$. We assume that the cost characteristic has been suitably scaled so that $\bar{C} = N^{-1} \sum C_j = 1$. Unless stated otherwise, we also assume that there exists a universal constant Δ such that $0 \le C_j \le \Delta$ for all j. We assume that the total cost, n, to be spent on sample selection is fixed in advance and is an integer. The noninteger case can be easily reduced to the integer case by replacing n by the greatest integer contained in n. We assume that $\min(n, N - n)$ is at least as large as 2Δ. This last assumption is a technicality and is needed to ensure that the sample size in SRSFC is at least two and no more than $(N - 2)$.

Briefly, under SRSFC sample units are drawn sequentially with equal probabilities and without replacement (WOR) until the total accumulated cost of sampling reaches the preassigned level n. This version of simple random sampling completely eliminates the randomness of the total cost of sample selection, as well as ensures collection of maximum possible information at a given cost. Under SRSFC, we record the observed sample as $\mathscr{S}_\nu = (u_1, u_2, \cdots, u_\nu)$, where $u_i = (y_i, c_i)$ is the ith sample unit with y_i being its Y-variate value and c_i its cost of selection. We define the stopping variable ν in a somewhat nontraditional way as follows: $\nu = r$ if and only if $\sum_1^r c_i \le n$ and $\sum_1^{r+1} c_i > n$. Since under this stopping rule, given ν, the conditional distribution of $u_1, \cdots, u_\nu$ is a symmetric function of ν, we shall henceforth refer to ν as a *symmetric stopping rule*. We hope that the prefix "symmetric" would keep the readers from inadvertently mistaking it for a stopping rule in the customary sense.

In SRSFC, the sample mean $\bar{y}_\nu = (1/\nu) \sum_i y_i$ is an unbiased estimator of the corresponding population mean $\bar{Y}$. Similarly the sample variance $s_\nu^2 = (\nu - 1)^{-1} \cdot \sum_i (y_i - \bar{y}_\nu)^2$ is an unbiased estimator of the corresponding population variance $S^2 = (N - 1)^{-1} \sum (Y_j - \bar{Y})^2$. Thus despite the fact that the sample size ν is random, the customary estimators of the population mean and variance continue to remain unbiased. Admissibility of the sample mean is also preserved. Besides the customary estimator of the variance of the sample mean, viz. $(1/n - 1/N)s^2$, with n replaced by ν continues to be an unbiased estimator of $V(\bar{y}_\nu)$ even though the actual form of $V(\bar{y}_\nu)$ is quite complicated. We refer the reader to [6] for these details. Perhaps the most intriguing aspect of SRSFC is that under very mild restrictions the necessary and sufficient condition for asymptotic normality of the sample mean under SRSFC is identical to those for the sample mean under SRSWOR. In order to present this result, we first establish a few moment inequalities.

LEMMA 2.1. *Let* $\max C_j \leq \Delta$. *Let* ν_n *denote the sample size for SRSFC of cost* n. *Then*

$$|E\nu_n - n| \leq 1 + \Delta. \tag{2.1}$$

PROOF. Let $\mathbf{u} = (u_1, u_2, \cdots, u_N)$ denote a random permutation of the N population units. Let $\nu_n(\mathbf{u})$ be the observed value of ν_n based on $\mathbf{u}$ and c_i the cost of observing the ith sample unit u_i in $\mathbf{u}$. Let $S(\nu_n)$ denote the actual cost of observing an SRSFC sample of cost n based on $\mathbf{u}$. Then $S(\nu_n) = c_1 + c_2 + \cdots + c_{\nu_n}$. Clearly

$$ES(\nu_n) = E[E(c_1 + \cdots + c_{\nu_n} \mid \nu_n)]. \tag{2.2}$$

Since ν_n depends on the c_i's up to the time ν_n only through the sum $S(\nu_n)$, a symmetric function of $c_1, \cdots, c_{\nu_n}$, it follows that $E(c_1 \mid \nu_n) = \cdots = E(c_{\nu_n} \mid \nu_n)$. Consequently

$$ES(\nu_n) = E[E(\nu_n c_1 \mid \nu_n)] = E[c_1 \nu_n]. \tag{2.3}$$

Now given the random permutation $\mathbf{u} = (u_1, \cdots, u_N)$ of the N populations units, let $\mathbf{v} = (v_1, \cdots, v_N)$ be another permutation of the N units obtained from $\mathbf{u}$ by placing u_1 at random in one of the empty spaces marked by the asterisks in the arrangement $* \, u_2 * u_3 * \cdots * u_N *$. Then like $\mathbf{u}$, $\mathbf{v}$ is also a random permutation of the N population units. Moreover, it is easily seen that for each j, $1 \leq j \leq N$, $P(u_1 = U_j \mid \mathbf{v}) = N^{-1}$ which is free of $\mathbf{v}$. Consequently $u_1(\mathbf{u})$ and $\mathbf{v}$ are independently distributed. Let $\nu_n(\mathbf{v})$ denote the observed value of ν_n based on $\mathbf{v}$. It is clear that $\nu_n(\mathbf{v}) \geq \nu_n(u) - 1$. Therefore

$$c_1(\mathbf{u})\nu_n(\mathbf{u}) \leq c_1(\mathbf{u})(1 + \nu_n(\mathbf{v})). \tag{2.4}$$

Taking expectations on both sides of (2.4) and invoking the independence between $c_1(u_1)$ and $\nu_n(\mathbf{v})$, we obtain

$$Ec_1\nu_n \leq (Ec_1)(1 + E\nu_n) = 1 + E\nu_n \tag{2.5}$$

since by assumption $Ec_1 = N^{-1} \sum C_j = 1$. The definition of ν_n implies that

$S(\nu_n) > n - \Delta$. So from (2.5) and (2.3), we get

$$E\nu_n \geq n - (1 + \Delta). \tag{2.6}$$

To establish the reverse inequality, consider an SRSFC of cost $(n + \Delta)$. Then a similar analysis, invoking the inequality $\nu_{n+\Delta}(\mathbf{u}) \geq \nu_n(\mathbf{v})$, yields

$$(n + \Delta) \geq ES(\nu_{n+\Delta}) = Ec_1\nu_{n+\Delta} \geq Ec_1 E\nu_n = E\nu_n. \tag{2.7}$$

The lemma follows from (2.6) and (2.7).

LEMMA 2.2. *$E\nu_n^2 \leq (n + 2\Delta)^2/Ec_1c_2$.*

PROOF. Consider an SRSFC of cost $(n + 2\Delta)$. Then the total cost of sample selection admits the representation: $S(\nu_{n+2\Delta}) = \sum_1 c_i$, in which the sum $\sum_1$ runs over $1 \leq i \leq \nu_{n+2\Delta}$. A technique similar to that of Lemma 2.1 yields

$$\begin{aligned} ES^2(\nu_{n+2\Delta}) &= Ec_1^2\nu_{n+2\Delta} + Ec_1c_2\nu_{n+2\Delta}(\nu_{n+2\Delta} - 1) \\ &= Ec_1c_2\nu_{n+2\Delta}^2 + E(c_1 - c_2)^2\nu_{n+2\Delta}/2 \geq Ec_1c_2\nu_{n+2\Delta}^2. \end{aligned} \tag{2.8}$$

Again let $\mathbf{u} = (u_1, u_2, \cdots, u_N)$ denote a random permutation of the N population units. Let $\nu_{n+2\Delta}(\mathbf{u})$, $c_1(\mathbf{u})$ and $c_2(\mathbf{u})$ be the respective observed values of these variables based on $\mathbf{u}$. Next given $\mathbf{u}$, let $\mathbf{v} = (v_1, v_2, \cdots, v_N)$ be a second permutation obtained from $\mathbf{u}$ by replacing u_1 and u_2 at random in the $(N - 1)$ empty spaces marked by the asterisks in the arrangement $* \ u_3 \ * \ \cdots \ * \ u_N \ *$. It is easily seen that the pair $(c_1(\mathbf{u}), c_2(\mathbf{u}))$ and $\mathbf{v}$ are independently distributed. Let $\nu_n(\mathbf{v})$ denote the observed value of the symmetric stopping time based on $\mathbf{v}$ of cost n. Then $\nu_{n+2\Delta}^2(\mathbf{u})c_1(\mathbf{u})c_2(\mathbf{u}) \geq \nu_n^2(\mathbf{v})c_1(\mathbf{u})c_2(\mathbf{u})$. Independence of (c_1, c_2) and $\nu_n(\mathbf{v})$ now implies that

$$Ec_1c_2\nu_{n+2\Delta}^2 \geq Ec_1c_2E\nu_n^2. \tag{2.9}$$

The lemma follows from (2.8), (2.9) and the fact that $S(\nu_{n+2\Delta}) \leq n + 2\Delta$.

LEMMA 2.3.

$$V(\nu_n) \leq 8\Delta n + 8n^2S_c^2/N \tag{2.10}$$

where $S_c^2 = (N - 1)^{-1}(\sum C_j^2 - 1)$.

This follows from the preceding lemma and the observations that $Ec_1c_2 - 1 = \mathrm{Cov}(c_1, c_2) = -S_c^2/N$ and that

$$Ec_1c_2 = Ec_1E(c_2 \mid c_1) \geq Ec_1(N - \Delta)/(N - 1) \geq (Ec_1)/2 = 1/2.$$

The complement of an SRSWOR sample of size n is an SRSWOR sample of size $(N - n)$. It is natural to ask if this duality is also shared by SRSFC. If it indeed were true, we should then be able to establish a dual of Lemma 2.3 with n replaced by $(N - n)$ on the right side of (2.10). This can be done only if we assume that the cost-characteristic assumes only positive values.

LEMMA 2.4. *Suppose that $C_j > 0$ for all j. Then*

$$V(\nu_n) \le 20[\Delta(\min(n, N-n)) + (S_c^2/N)(\min(n, N-n))^2]. \tag{2.11}$$

PROOF. Let $\mathbf{u}$ be a random permutation of the N population units. Let $\mathbf{w}$ be the permutation $\mathbf{u}$ in the reverse order. Let $\nu_n(\mathbf{u})$ denote the symmetric stopping time based on $\mathbf{u}$ for SRSFC of cost n and define $\nu_{N-n}(\mathbf{w})$ similarly for SRSFC of cost $(N-n)$. The lemma follows from Lemma 2.3 and the identity: $\nu_n(\mathbf{u}) = N - \nu_{N-n}(\mathbf{w}) - \theta$, where $\theta = 0$ or 1 with $\theta = 0$ only if $S(\nu_n) = n$.

We turn now to the investigation of asymptotic normality in SRSFC.

3. Asymptotic normality in SRSFC. Let R_c denote an SRSFC sampling experiment of selecting ν units of cost n from the given population P, and R_s refer to an SRSWOR sampling experiment of selecting n units from P. Let $\bar{y}_\nu$ and $\bar{y}_n$ denote the sample means under SRSFC and SRSWOR respectively. We first show that $\bar{y}_\nu$ and $\bar{y}_n$ are asymptotically equivalent. To do so, we perform the following random experiment.

RANDOM EXPERIMENT.

1. First draw an SRSWOR sample of size n from P as follows: Let $\mathbf{u} = (u_1, \cdots, u_N)$ be a random permutation of P. Let $\mathscr{S}_n$ denote the observed SRSWOR based on the first n coordinates of $\mathbf{u}$ and let $S(n)$ denote the total cost of selecting this sample.
2. If $S(n) > n$, select an SRSFC subsample $\mathscr{S}_\nu$ of cost n from $\mathscr{S}_n$, treating $\mathscr{S}_n$ as a population in its own right.
3. If $S(n) \le n$, select sequentially additional units from the remaining units in P until an SRSFC sample $\mathscr{S}_\nu$ of cost n has been selected.
4. Repermute the ν units in $\mathscr{S}_\nu$ at random. For notational simplicity, we denote this repermuted sample also by the same symbol $\mathscr{S}_\nu$.

Under this experiment $\mathscr{S}_n$ is an SRSWOR sample of size n and $\mathscr{S}_\nu$ an SRSFC sample of cost n.

LEMMA 3.1. *Under the given experiment*

$$E(\bar{y}_\nu - \bar{y}_n)^2 = E\,|\,1/\nu - 1/n\,|\, s_\nu^2 \tag{3.1}$$

where s_ν^2 denotes the sample variance based on $\mathscr{S}_\nu$.

PROOF. Observe that given $S(n) \le n$ and $\mathscr{S}_\nu$, $\mathscr{S}_n$ is an SRSWOR subsample of size n from $\mathscr{S}_\nu$. This implies that given $S(n) \le n$ and

$$\mathscr{S}_\nu,\ E[(\bar{y}_n - \bar{y}_\nu)^2 \mid S(n) \le n, \mathscr{S}_\nu] = (n^{-1} - \nu^{-1})s_\nu^2.$$

Next given $S(n) > n$ and $\mathscr{S}_n$, $\mathscr{S}_\nu$ is an SRSFC subsample of cost n from $\mathscr{S}_n$. This implies that $(\nu^{-1} - n^{-1})s_\nu^2$ is an unbiased estimator of $E[(\bar{y}_n - \bar{y}_\nu)^2 \mid S(n) > n, \mathscr{S}_n]$ (cf. Pathak, Theorem 2.1, page 1014, [6]). From these considerations, the lemma follows.

LEMMA 3.2. *Under the given experiment*

$$E(\bar{y}_\nu - \bar{y}_n)^2 \le (10\Delta/n^2)[\Delta + \sqrt{V(\nu)}]\sigma^2 \tag{3.2}$$

where $\sigma^2 = N^{-1} \sum (Y_j - \bar{Y})^2$.

PROOF. Techniques similar to that of Lemmas 2.1 and 2.2 show that

$$E(\bar{y}_\nu - \bar{y}_n)^2 = E\left[\frac{|\nu - n|(y_1 - y_2)^2}{2\nu n}\right] \le \frac{10\Delta}{n^2}[\Delta + \sqrt{V(\nu)}]\sigma^2. \tag{3.3}$$

LEMMA 3.3. *Under the given experiment*

$$H(\bar{y}_\nu, \bar{y}_n) \equiv \frac{E(\bar{y}_\nu - \bar{y}_n)^2}{V(\bar{y}_n)} \le 10\Delta[\Delta + \sqrt{V(\nu)}]\left[\frac{1}{n} + \frac{1}{N-n}\right]. \tag{3.4}$$

This follows from Lemma 3.2 *and the fact that* $V(\bar{y}_n) = [(N-n)/n(N-1)]\sigma^2$.

We refer to $H(\bar{y}_\nu, \bar{y}_n)$ as Hâjek's *measure of disparity* between $\bar{y}_\nu$ and $\bar{y}_n$. Now suppose that $H(\bar{y}_\nu, \bar{y}_n)$ approaches zero asymptotically, i.e., as $E\nu \approx n \to \infty$ *and* $(N - n) \to \infty$,

$$\lim H(\bar{y}_\nu, \bar{y}_n) = 0. \tag{3.5}$$

The implications of (3.5) are quite remarkable. First, it implies that $\lim V(\bar{y}_\nu)/V(\bar{y}_n) = 1$, showing that both $\bar{y}_\nu$ and $\bar{y}_n$ are asymptotically equally efficient. Second, it implies the asymptotic equivalence of the limiting distributions of $\bar{y}_\nu$ and $\bar{y}_n$ in the following sense. Suppose that under a certain condition $\bar{y}_n$ is asymptotically normally distributed with parameters $(E\bar{y}_n, V(\bar{y}_n))$. Then under the same condition $\bar{y}_\nu$ is asymptotically normally distributed with parameters $(E\bar{y}_\nu, V(\bar{y}_\nu))$ and vice versa (cf. Hâjek, [4]).

Lemma 3.3. implies that Hâjek's disparity between $\bar{y}_\nu$ of SRSFC and $\bar{y}_n$ of SRSWOR can be made to approach zero asymptotically provided $n \to \infty$ and $(N - n) \to \infty$ such that $\lim V(\nu)/n^2 = 0$ and $\lim V(\nu)/(N - n)^2 = 0$. Lemma 2.3 implies that all SRSFC satisfy the first of these two conditions. The second condition is, however, more restrictive. For example, under the added assumption that the cost-characteristic is strictly positive, Lemma 2.4 holds. Then both of these conditions are satisfied. It is worth noting that inverse simple random sampling (SRSI) becomes a special case of SRSFC if we allow the cost-characteristic to be simply nonnegative so that it can also assume the value zero. Under SRSI units are drawn sequentially (WOR) until a preassigned number n of units with a specified trait are included in the sample. For such SRSFC schemes $V(\nu)/(N - n)^2$ does not approach zero without added restrictions on the rate of growth of n and $(N - n)$. A simple additional condition that suffices is to require that $\lim n/(N - n)^2 = 0$ which is satisfied if for example the sampling fraction remains strictly less than one.

Thus through the preceding analysis we have, among other things, established the following main results.

THEOREM 3.1. *Suppose that the cost-characteristic is strictly positive and bounded. Let* $\lim n = \lim(N - n) = \infty$. *Then under SRSFC of cost n, the sample mean* $\bar{y}_\nu$ *is asymptotically normally distributed if and only if the Erdös–Rényi–Hájek condition of Theorem* 1.1 *holds.*

THEOREM 3.2. *Suppose that the cost-characteristic is nonnegative and bounded, then the conclusion of the above theorem goes through if we require that* $\lim n = \lim(N - n) = \infty$ *and* $\lim \sup(n/N) < 1$.

The proof of these theorems are immediate consequences of Theorem 1.1 and the discussions following Theorem 1.1 and Lemma 3.3.

It should now be clear from the above theorems that as the cost of sampling gets larger and larger, the cost-adjusted relative efficiency of SRSFC versus SRSWOR approaches one. Our theorems provide for the first time a truly rigorous justification for results of this nature first anticipated by Basu [1] and others. In broad terms a second and perhaps equally important implication of our results is that asymptotically conditional inference based on simple random sampling is likely to be just as efficient as their unconditional counterparts provided the reference set for conditional inference is chosen so as to ensure that $\lim V(\nu)[1/(E\nu)^2 + 1/(N - E\nu)^2] = 0$. We would like to thank Professor Marvin Zelen for bringing this point to our attention.

4. Acknowledgments. We are grateful to the Associate Editor for his keen interest and thoughtful comments which have led to considerable improvements in the exposition of our results.

REFERENCES

[1] BASU, D. (1958). On sampling with and without replacement. *Sankhya* **20** 287–294.

[2] ERDÖS, P. and RÉNYI, A. (1959). On the central limit theorem for samples from a finite population. *Publ. Math. Inst. Hung. Acad. Sci.* **A4** 49–61.

[3] HÁJEK, J. (1959). Optimum strategy and other problems in probability sampling. *Casopis Pest. Mat.* **84** 387–425.

[4] HÁJEK, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publ. Math. Inst. Hung. Acad. Sci.* **A5** 361–374.

[5] JOSHI, V. M. (1966). Admissibility and Bayes estimation in sampling finite populations—IV. *Ann. Math. Statist.* **37** 1658–1670.

[6] PATHAK, P. K. (1976). Unbiased estimation in fixed cost sequential sampling schemes. *Ann. Statist.* **4** 1012–1017.

[7] PATHAK, P. K. (1982). Asymptotic normality of the average of distinct units in simple random sampling with replacement. *Essays in Honour of C. R. Rao* 567–573. G. Kallianpur et al., Eds. North Holland, Amsterdam.

[8] RÉNYI, A. (1957). On the asymptotic distribution of the sum of a random number of independent random variables. *Acta Math. Acad. Sci. Hungar.* **8** 193–199.

[9] SETH, C. R. and RAO, J. N. K. (1964). On the comparison of simple random sampling with and without replacement. *Sankhya A* **26** 85–86.

INDIAN STATISTICAL INSTITUTE DELHI CENTRE
7 SJSS MARG
NEW DELHI 110016
INDIA

DEPT. OF MATHEMATICS AND STATISTICS
UNIVERSITY OF NEW MEXICO
ALBUQUERQUE, NEW MEXICO 87131