Biometrika (1964), 51, 1 and 2, p. 185
Printed in Great Britain

On inverse sampling with unequal probabilities†

BY P. K. PATHAK‡
University of Illinois

SUMMARY

Sampford (1962) has considered the following sampling scheme to select a sample with n distinct population units. Sampling with unequal probabilities with replacement is carried out until $(n+1)$ different population units are selected, the last unit is not recorded in the sample to insure some simplicity in the estimation procedure. The method is called inverse sampling with unequal probabilities. In this paper it is shown that this method of sampling is equivalent to sampling with unequal probabilities without replacement in some sense. The estimator of the population total given by Sampford is compared with other existing estimators. It is shown that there exist estimators which are uniformly better than the estimator given by Sampford. Finally, it is demonstrated that Des Raj's estimator (1956) has certain definite advantages over Sampford's estimator suggesting that sampling with unequal probabilities (without replacement) with Des Raj's estimator probably gives a better estimation procedure than inverse sampling with unequal probabilities.

1. INTRODUCTION AND PRELIMINARIES

Although sampling with unequal probabilities with replacement has the merit of simplicity, it has a major drawback that the number of distinct population units selected in a sample is a random variable and varies from sample to sample. Sampling with unequal probabilities without replacement does not have this drawback, but most estimators of the population total in this sampling scheme are either unwieldy or biased. Recently some workers have developed sampling schemes wherein the number of distinct population units selected in a sample is constant from sample to sample, and also the estimator of the population total is easy to compute and has some desirable properties like possessing a non-negative estimator of its variance, etc. Of special interest in this connexion are the works of Des Raj (1956), Stevens (1958), Rao, Hartley & Cochran (1962) and others. In cluster sampling, Sampford (1962) has considered a method of sampling with probabilities proportional to cluster size (with replacement) by which a fixed number of distinct clusters are included in the sample and has given an unbiased estimator of the population total and an estimator of its variance. A slightly different version of this method of sampling is considered here in detail and is referred to as inverse sampling with unequal probabilities.

The following notation and definitions are used in the subsequent sections.

Number of population units:	N.
jth population unit:	$U_j \quad (j=1,\ldots,N)$.
Value of a real-valued Y-characteristic of U_j:	$Y_j \quad (j=1,\ldots,N)$.
Probability of selection of U_j:	$P_j \quad \left(\sum_{j=1}^{N} P_j = 1\right)$.

† Work supported by the National Science Foundation.
‡ On leave from Indian Statistical Institute.

Z-value of U_j:	$Z_j = Y_j/P_j \quad (j=1,\ldots,N)$.
Population total of Y-characteristic:	$Y = \sum_{j=1}^{N} Y_j$.
Population variance:	$\sigma_z^2 = \sum_{j=1}^{N} P_j(Z_j - Y)^2$.
Number of distinct population units in a sample:	n.
Size of a sample:	$r \quad (r = n, n+1, \ldots)$.
An observed sample:	$s = (u_1, \ldots, u_r)$.

Small letters will refer to the sample and capital to the population, e.g. $u_1, u_2, \ldots, u_r$ are the r sample units selected in order of draw.

DEFINITION 1·1. Let $u_{(1)}, u_{(2)}, \ldots, u_{(n)}$ be the n distinct population units observed in a sample and arranged in order of their draw. Then $T_2 = [u_{(1)}, u_{(2)}, \ldots, u_{(n)}]$ is called the 'serial-statistic'.

DEFINITION 1·2. Let $u_{(1)}, u_{(2)}, \ldots, u_{(n)}$ be the n distinct population units observed in a sample and arranged in increasing order of their unit-indices. Then $T_1 = [u_{(1)}, u_{(2)}, \ldots, u_{(n)}]$ is called the 'order-statistic'.

The unit index of an individual whose population representation is U_j is j. The above definition of 'order-statistic' differs from its customary definition where the sample observations are arranged in order of magnitude. This definition has been adopted here to avoid possible ambiguity in defining the order-statistic' when a sample may contain two different population units of the same order of magnitude.

DEFINITION 1·3. Let $\alpha_{(1)}, \alpha_{(2)}, \ldots, \alpha_{(n)}$ be respectively the number of times $u_{(1)}, u_{(2)}, \ldots, u_{(n)}$ occur in the sample. Then $T_3 = [(u_{(1)}, \alpha_{(1)}), (u_{(2)}, \alpha_{(2)}), \ldots, (u_{(n)}, \alpha_{(n)})]$ is called the 'symmetric-statistic'.

Any statistic $T(s)$, a function of the sample observations, divides the collection of samples into sets such that the given statistic $T(s)$ is constant on each of these sets. This collection of sets is the partition induced by $T(s)$. Two samples are said to be equivalent if they contain the same population units. A sufficient statistic is thus defined as follows (Pathak, 1964).

DEFINITION 1·4. A statistic $T(s)$ is sufficient if the partition induced by $T(s)$ consists of sets of equivalent samples.

It is thus obvious that the above defined statistics, T_1, T_2 and T_3, are all sufficient.

For sampling schemes from finite populations, the above definition is a suitable version of the existing notion of sufficiency as defined by Fisher (1921), Halmos & Savage (1949), Bahadur (1954) and others. The following intuitive argument shows that this definition is equivalent to the definition of sufficiency in Fisher sense. Under a given sampling process, a selected sample gives us information about those population units selected in the sample. If two samples consist of the same population units they give the same information about the population units and are equally informative. Given one of these samples one can predict what the other sample is and vice versa. In the same way a sample from a given set of equivalent samples is as informative as any other sample from that set; and it is theoretically possible to work out all samples of this set given any one sample from it. A partition of samples into sets of equivalent samples divides the sample space into sets of

equally informative samples and is, therefore, sufficient. It may be noted that the sufficiency is with regard to the unknown population characteristics $Y_1, Y_2, \ldots, Y_N$.

It is possible to give a rigorous justification of this definition in abstract terms but this has been omitted here for brevity. The author (1964) has dealt with this problem at some length in a paper to be published in *Annals of Mathematical Statistics* soon.

DEFINITION 1·5. Given two estimators $t_1(s)$ and $t_2(s)$ of a population parameter, $t_1(s)$ is said to be uniformly better than $t_2(s)$ if for any convex loss function $t_1(s)$ has smaller expected loss than $t_2(s)$ for all $Y_1, \ldots, Y_N$ with strictly less expected loss for at least one $Y_1, \ldots, Y_N$.

It follows as a consequence of the Rao–Blackwell theorem (Pathak, 1964) that if $T(s)$ is a sufficient statistic and if $t_2(s)$ is an estimator of a parameter which does not depend on $T(s)$ then $t_1(s) = E[t_2(s)|T(s)]$ is a uniformly better estimator than $t_2(s)$ unless $t_1(s) = t_2(s)$ with probability one. If the loss function is the squared error then $t_1(s)$ has smaller mean-square error than $t_2(s)$; the decrease in the mean-square error is equal to $E[t_1(s) - t_2(s)]^2$.

DEFINITION 1·6. Random variables $u_1, u_2, \ldots, u_r$ are called interchangeable if their joint distribution is invariant under any permutation of $u_1, u_2, \ldots, u_r$.

2. INVERSE SAMPLING WITH UNEQUAL PROBABILITIES

In inverse sampling with unequal probabilities, population units are selected with unequal probabilities (with replacement), P_j being the probability of selection of U_j $(j = 1, \ldots, N)$, and the selection of units is stopped at the $(r+1)$th draw when the sample first contains $(n+1)$ different population units. The last unit is rejected and the recorded sample consists of sample units corresponding to n different population units. The rejection of the last sample unit introduces simplicity in the distribution of the sample units as then for a given sample size the recorded sample units behave as interchangeable random variables. An observed sample is recorded as

$$s = (u_1, u_2, \ldots, u_r) \quad (r = n, n+1, \ldots), \tag{1}$$

where $u_1, u_2, \ldots, u_r$ are respectively the first, second and the rth sample units.

The probability of selecting a sample s is given by

$$P(s) = kf(s)\, p_1 p_2 \ldots p_r (1 - p_{(1)} - p_{(2)} - \ldots - p_{(n)}), \tag{2}$$

where p_i denotes the probability of selection of the ith sample unit, u_i $(i = 1, \ldots, r)$,

$$f(s) = \begin{cases} 1 & \text{if } s \text{ contains } n \text{ distinct population units,} \\ 0 & \text{otherwise,} \end{cases}$$

$p_{(1)}, \ldots \ldots, p_{(n)}$ are the probabilities of selection associated with n different population units selected in the sample, and k is to be determined such that the summation of the right side over all possible samples equals unity.

From (2) it is evident that for a given r, $u_1, u_2, \ldots, u_r$ are interchangeable random variables.

Under this sampling scheme, the order-statistic, the serial statistic and the symmetric-statistic are all sufficient. The probability distribution of the symmetric-statistic

$$T_3 = [(u_{(1)}, \alpha_{(1)}), \ldots, (u_{(n)}, \alpha_{(n)})]$$

is easily seen to be given by

$$P[T_3] = kf(s) \frac{r!}{\alpha_{(1)}! \ldots \alpha_{(n)}!} p_{(1)}^{\alpha_{(1)}} \ldots p_{(n)}^{\alpha_{(n)}} (1 - p_{(1)} - \ldots - p_{(n)}), \tag{3}$$

where k and $f(s)$ have been defined in (2) and $p_{(i)}$ is the probability of selection of $u_{(i)}$ $(i = 1, \ldots, n)$.

A little consideration will show that the probability distribution of the serial-statistic $T_2 = [u_{(1)}, u_{(2)}, \ldots, u_{(n)}]$ is given by

$$P[T_2] = p_{(1)} \frac{p_{(2)}}{(1-p_{(1)})} \cdots \cdots \frac{p_{(n)}}{(1-p_{(1)}-\ldots-p_{(n-1)})}, \tag{4}$$

where $p_{(i)}$ is the probability of selection of $u_{(i)} (i = 1, \ldots, n)$.

Lastly the probability distribution of the order-statistic $T_3 = [u_{(1)}, \ldots, u_{(n)}]$ is given by

$$P[T_3] = \Sigma' p_{(1)} \frac{p_{(2)}}{(1-p_{(1)})} \cdots \cdots \frac{p_{(n)}}{(1-p_{(1)}-\ldots-p_{(n-1)})}, \tag{5}$$

where the summation Σ' extends over all posaible orderings of $p_{(1)}, \ldots, p_{(n)}$.

It can be seen from (4) that the distribution of the serial-statistic is the same as that of the serial-statistic in sampling with unequal probabilities (without replacement). Thus in this sense the two sampling schemes are equivalent. The author (1961) has earlier proved a similar result. The equivalence can be made precise as follows. Suppose $g_1(s)$ and $g_2(s)$ are estimators of the population total based on inverse sampling with unequal probabilities and sampling with unequal probabilities (without replacement) respectively. Now if to the statistician the outcome of only one of these sampling schemes is given, he could, if he wished, compute with the help of a random device both $g_1(s)$ and $g_2(s)$ such that their probability distributions will be identical with their original probability distributions under the two given sampling schemes. Thus theoretically, the knowledge of the outcome of one of these sampling schemes enables the statistician to produce an outcome of the second sampling scheme and therefore the two sampling schemes can be said as equivalent. This also shows that in inverse sampling with unequal probabilities if one is using estimators which are based on the serial-statistic then sampling can be stopped as soon as the nth population unit has been selected.

3. Estimation of the population total

Sampford (1962) has considered the following estimator of the population total

$$\bar{z} = \frac{1}{r} \sum_{i=1}^{r} z_i, \tag{6}$$

where z_i is the Z-value of u_i, and gave

$$v(\bar{z}) = \frac{1}{r(r-1)} \sum_{i=1}^{r} (z_i - \bar{z})^2 \tag{7}$$

as an unbiased estimator of $V(\bar{z})$.

Since $E(z_1) = Y$, and for a given $r, z_1, \ldots, z_r$ are interchangeable random variables,

$$E[\bar{z}] = E_r E[\bar{z}|r] = E_r E[z_1|r] = E[z_1] = Y. \tag{8}$$

This proves that $\bar{z}$ is an unbiased estimator of the population total Y.

To prove the unbiasedness of $v(\bar{z})$ it suffices to observe that

$$\begin{aligned} E[v(\bar{z})] &= E\left[\bar{z}^2 - \frac{1}{r(r-1)} \sum_{i \neq i'=1}^{r} z_i z_{i'}\right] \\ &= E[\bar{z}^2] - E_r E\left[\frac{1}{r(r-1)} \sum_{i \neq i'=1}^{r} z_i z_{i'} | r\right] \\ &= E[\bar{z}^2] - E_r E[z_1 z_2 | r] \\ &= E[\bar{z}^2] - E[z_1 z_2] = E[\bar{z}^2] - Y^2 \\ &= V(\bar{z}). \end{aligned} \tag{9}$$

Further if $c_1, c_2, \ldots, c_r$ are r variables depending on r only and such that $\sum_{i=1}^{r} c_i = 1$, then it is easily verified in a similar manner that $\sum_{i=1}^{r} c_i z_i$ is also an unbiased estimator of Y. It is proved below that $\bar{z}$ is the best estimator among this class of estimators.

THEOREM 1. $\bar{z}$ is uniformly better than any estimator of the above type.

Proof. Since the symmetric statistic

$$T_3 = [(u_{(1)}, \alpha_{(1)}), \ldots, (u_{(n)}, \alpha_{(n)})]$$

is sufficient, it follows that an estimator uniformly better than $\sum_{i=1}^{r} c_i z_i$ is given by

$$E\left[\sum_{i=1}^{r} c_i z_i | T_3\right].$$

Further, when T_3 is given (r is fixed), $c_1, c_2, \ldots, c_r$ are fixed and also $z_1, z_2, \ldots, z_r$ are interchangeable. Therefore

$$E\left[\sum_{i=1}^{r} c_i z_i | T_3\right] = E\left[\left\{\sum_{i=1}^{r} c_i\right\} z_1 | T_3\right] = E[z_1 | T_3]$$

$$= \sum_{i=1}^{n} z_{(i)} P[u_1 = u_{(i)} | T_3]. \tag{10}$$

It can be easily verified that

$$P[u_1 = u_{(i)} | T_3] = \frac{P[u_1 = u_{(i)} \cap T_3]}{P[T_3]} = \frac{\alpha_{(i)}}{r}, \tag{11}$$

so that

$$E[\Sigma c_i z_i | T_3] = \frac{1}{r} \sum_{i=1}^{r} \alpha_{(i)} z_{(i)} = \bar{z}. \tag{12}$$

This completes the proof of the theorem.

COROLLARY 1·1. On taking $c_1 = c_2 = \ldots = c_n = 1/n$ and $c_{n+1} = \ldots = c_r = 0$, it is seen that $\bar{z}$ is uniformly better than the corresponding estimator $\bar{z}_n = \frac{1}{n} \sum_{i=1}^{n} z_i$ of sampling with unequal probabilities (with replacement).

4. VARIANCE OF $\bar{z}$

From (9), it is clear that

$$V(\bar{z}) = E\left[\frac{1}{r(r-1)} \sum_{i=1}^{r} (z_i - \bar{z})^2\right]$$

$$= E\left[\frac{1}{2r^2(r-1)} \sum_{i \neq i'=1}^{r} (z_i - z_{i'})^2\right]$$

$$= E_r E\left[\frac{1}{2r^2(r-1)} \sum_{i \neq i'=1}^{r} (z_i - z_{i'})^2 | r\right]$$

$$= E_r E\left[\frac{1}{2r} (z_1 - z_2)^2 | r\right], \tag{13}$$

since for a given r, $z_1, \ldots, z_r$ are interchangeable. Thus

$$V(\bar{z}) = E\left[\frac{(z_1 - z_2)^2}{2r}\right]$$

$$= \sum_{j \neq j'=1}^{N} \tfrac{1}{2} P_j P_{j'} (Z_j - Z_{j'})^2 C_{jj'}, \tag{14}$$

where

$$C_{jj'} = E[1/r | u_1 = U_j, u_2 = U_{j'}] \quad (j \neq j' = 1, \ldots, N).$$

In the particular case when $n = 2$

$$C_{jj'} = -\frac{(1-P_j-P_{j'})}{(P_j+P_{j'})^2}[\log(1-P_j-P_{j'})+P_j+P_{j'}] \quad (j \neq j' = 1, \ldots, N). \tag{15}$$

An exact expression for $C_{jj'}$ is derived in the Appendix given at the end of the paper. When $P_j = 1/N\,(j = 1, \ldots, N)$, $V(\bar{z})$ can be expressed as (Sampford)

$$V(\bar{z}) = \sigma_z^2 \binom{N-2}{n-1} \sum_{k=1}^{n} (-)^{n-k} \binom{n-1}{k-1} \left(\frac{1}{k} + \frac{N(k-1)}{k^2} \log\left(1-\frac{k}{N}\right)\right). \tag{16}$$

An asymptotic expression for $V(\bar{z})$ in this case in ascending powers of $1/N$ is given by

$$V(\bar{z}) = \frac{1}{n}\sigma_z^2\left\{1 - \frac{1}{2N}\left[n - \frac{2}{n+1}\right] - \frac{1}{12N^2}\left[n^2 + \frac{n^3+15n^2+2n-24}{(n+1)(n+2)}\right] \ldots\right\}. \tag{17}$$

5. Comparison of Sampford's estimator with Des Raj's estimator

If the serial-statistic is recorded then it is seen that Des Raj's estimator (1956) is obtained from

$$\left.\begin{aligned} t_1 &= \frac{y_{i_1}}{p_{i_1}}, \\ t_2 &= y_{i_1} + \frac{y_{i_2}}{p_{i_2}}(1-p_{i_1}), \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ t_t &= y_{i_1} + \ldots + y_{i_{t-1}} + \frac{y_{i_t}}{p_{i_t}}(1-p_{i_1}-\ldots-p_{i_{t-1}}), \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ t_n &= y_{i_1} + \ldots + y_{i_{n-1}} + \frac{y_{i_n}}{p_{i_n}}(1-p_{i_1}-\ldots-p_{i_{n-1}}). \end{aligned}\right\} \tag{18}$$

Using the fact that $t_1, \ldots, t_n$ are uncorrelated, Des Raj considered

$$\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i \tag{19}$$

as an unbiased estimator of the population total and gave

$$v(\bar{t}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(t_i - \bar{t})^2 \tag{20}$$

as an unbiased estimator of $V(\bar{t})$.

The above estimator of the population total has an advantage over $\bar{z}$ that it does not take into account the number of times a particular unit u_{i_i} is included in the sample whereas the Sampford's estimator does. Des Raj's estimator, however, takes into account the order in which the distinct population units are included in the sample and Sampford's estimator does not. This is an advantage of Sampford's estimator over Des Raj's. It is felt that perhaps the advantage of Des Raj's estimator is more desirable than that of Sampford's. A still more desirable estimator than Des Raj's and Sampford's would be one which disregards both the order and the number of times different population units are included in the sample (Theorem 2).

It seems rather difficult to give a direct comparison of the relative efficiencies of $\bar{t}$ and

$\bar{z}$ in the general case. However, when $n = 2$, and if the loss function is the squared error, $V(\bar{t})$ is given by

$$\begin{aligned} V(\bar{t}) &= E[v(\bar{t})] = E[\tfrac{1}{4}(t_1 - t_2)^2] \\ &= \tfrac{1}{4}E\left[\left(\frac{y_{(1)}}{p_{(1)}} - \frac{y_{(2)}}{p_{(2)}}\right)^2 (1 - p_{(1)})^2\right] \\ &= \frac{1}{8}\sum_{j \neq j'=1}^{N} (Z_j - Z_{j'})^2 P_j P_{j'}(2 - P_j - P_{j'}). \end{aligned} \tag{21}$$

From this, it is obvious that $V(\bar{t}) < V(\bar{z})$ provided

$$2 - P_j - P_{j'} < -4\frac{(1 - P_j - P_{j'})}{(P_j + P_{j'})^2}[\log(1 - P_j - P_{j'}) + P_j + P_{j'}] \tag{22}$$

for all $j \neq j' = 1, \ldots, N$.

It can be seen that this will be so if $(P_j + P_{j'}) < \frac{1}{2}$ for all $j \neq j' = 1, \ldots, N$. Thus while sampling from a reasonably big population where this condition is automatically satisfied Des Raj's estimator will have smaller variance.

In another special case when $P_j = 1/N\ (j = 1, \ldots, N)$, $V(\bar{t})$ is given (Murthy, 1957) by

$$V(\bar{t}) = \frac{\sigma_z^2}{n}\left\{1 - \frac{(n-1)}{N} + \frac{(n-1)(n-2)}{3N(N-1)}\right\}. \tag{23}$$

A comparison of (23) with the asymptotic expression for the variance of $\bar{z}$ shows that $\bar{t}$ has smaller variance than $\bar{z}$.

The above considerations lead the author to believe that Des Raj's estimator will probably be better than Sampford's estimator in practice.

6. An estimator uniformly better than Sampford's estimator

The theorem below gives an estimator uniformly better than Sampford's estimator.

Theorem 2. An estimator uniformly better than $\bar{z}$ is given by

$$\bar{z}^* = \sum_{i=1}^{n} c_{(i)} y_{(i)}, \tag{24}$$

where $c_{(i)} = \{P[T_1|u_1 = u_{(i)}]\}/P[T_1]$. $P[T_1]$ is the probability of getting the order-statistic $T_1 = [u_{(1)}, \ldots, u_{(n)}]$ and $P[T_1|u_1 = u_{(i)}]$ is the conditional probability of getting the order-statistic, T_1, given that $u_{(i)}$ has been selected at the first draw.

Proof. Since T_1 is sufficient, an estimator uniformly better than $\bar{z}$ is given by

$$\bar{z}^* = E\left[\frac{1}{r}\sum_{i=1}^{r} z_i | T_1\right] = E_r E\left[\frac{1}{r}\sum_{i=1}^{r} z_i | T_1, r\right]. \tag{25}$$

Evidently $z_1, \ldots, z_r$ are interchangeable for a given r and T_1, and therefore

$$\begin{aligned} \bar{z}^* &= E_r E[z_1|T_1, r] = E[z_1|T_1] \\ &= \sum_{i=1}^{n} \frac{y_{(i)}}{p_{(i)}} P[u_1 = u_{(i)}|T_1] \\ &= \sum_{i=1}^{n} c_{(i)} y_{(i)}. \end{aligned} \tag{26}$$

This proves the theorem.

The author (1961) had earlier proved that $\bar{z}^*$ is uniformly better than Des Raj's estimator. Murthy (1957) proved that $\bar{z}^*$ has smaller variance than Des Raj's estimator. This shows that $\bar{z}^*$ is uniformly better than Des Raj's as well as Sampford's estimator. Unfortunately $\bar{z}^*$ cannot be of much use in practice because of the cumbersome computation of the coefficients $c_{(i)}$. It can be seen that $c_{(i)} = N/n$ when $P_j = 1/N\ (j = 1, \ldots, N)$ with the help of (5) showing that in simple random sampling (without replacement) the estimator based on the sample mean is uniformly better than both Des Raj's and Sampford's estimators.

Other estimators uniformly better than Sampford's estimator can be suggested, as for example, $E[\bar{z}|T_1, r]$, $E[\bar{z}|T_2']$ and $E[\bar{z}|T_3, r]$, etc. The reader may find it an illustrative exercise to work out these estimators. The author's (1962) paper may be referred to for reference. It will be interesting to see how the above considered methods of estimation compare in two-state sampling.

The author wishes to thank a referee for some valuable suggestions.

APPENDIX

The probability distribution of r. The probability of getting the order-statistic $T_1 = [u_{(1)}, \ldots, u_{(n)}]$, for a given r, is given by (Pathak, 1962)

$$P[T_1, r] = [(p_{(1)} + \ldots + p_{(n)})^r - \Sigma_1(p_{(1)} + \ldots + p_{(n-1)})^r + \ldots(-)^{n-1}\Sigma_1 p_{(1)}^r][1 - p_{(1)} - \ldots - p_{(n)}], \quad (27)$$

where the summation Σ_1 is taken over all combinations of p's chosen out of $p_{(1)}, \ldots, p_{(n)}$. Therefore, the probability distribution of r is given by

$$P[r] = \Sigma_2[(p_{(1)} + \ldots + p_{(n)})^r - \Sigma_1(p_{(1)} + \ldots + p_{(n-1)})^r + \ldots(-1)^{n-1}\Sigma_1 p_{(1)}^r][1 - p_{(1)} - \ldots - p_{(n)}]], \quad (28)$$

where Σ_2 is taken over all $\binom{N}{n}$ combinations of p's chosen out of $P_1, P_2, \ldots, P_N$.

From (27) and (28), it can be seen that

$$P[T_1, r|u_1 = u_{(i)}, u_2 = u_{(i')}] = [(p_{(1)} + \ldots + p_{(n)})^{r-2} - \Sigma_1^{(i')}(p_{(1)} + \ldots + p_{(n-1)})^{r-2} + \ldots(-)^{n-2}(p_{(i)} + p_{(i')})^{r-2}][1 - p_{(1)} - \ldots - p_{(n)}]], \quad (29)$$

where $\Sigma_1^{(i')}$ stands for all combinations of p's containing $p_{(i)}$ and $p_{(i')}$ chosen out of $p_{(1)}, \ldots, p_{(n)}$.

From (29), it follows that

$$E[1/r|u_1 = u_{(i)}, u_2 = u_{(i')}, T_1] = \left\{\left[\frac{1}{(p_{(1)} + \ldots + p_{(n)})^2}\left(\log\frac{1}{(1 - p_{(1)} - \ldots - p_{(n)})} - p_{(1)} \ldots - p_{(n)}\right)\right.\right.$$
$$- \Sigma_1^{(i')}\frac{1}{(p_{(1)} + \ldots + p_{(n-1)})^2}\left(\log\frac{1}{(1 - p_{(1)} - \ldots - p_{(n-1)})} - p_{(1)} - \ldots - p_{(n-1)}\right)$$
$$\left.\left. + \ldots(-)^{n-2}\frac{1}{(p_{(i)} + p_{(i')})^2}\left(\log\frac{1}{(1 - p_{(i)} - p_{(i')})} - p_{(i)} - p_{(i')}\right)\right]\right.$$
$$\left.\times[1 - p_{(1)} - \ldots - p_{(n)}]\right\} = C_{(i),(i')}(T_1). \quad (30)$$

Finally, $C_{jj'}$, as required in (14), is given by

$$C_{jj'} = \Sigma_3^{(j')} C_{jj'}(T_1), \quad (31)$$

where $\Sigma_3^{(j')}$ is taken over all $\binom{N-2}{n-2}$ combinations of n P's chosen out of $P_1, P_2, \ldots, P_N$ and containing P_j and $P_{j'}$.

REFERENCES

BAHADUR, R. R. (1954). Sufficiency and statistical decision function. *Ann. Math. Statist.* **25**, 423–62.

DES RAJ (1956). Some estimators in sampling with varying probabilities without replacement. *J. Amer. Statist. Ass.* **51**, 269–84.

FISHER, R. A. (1921). On the mathematical foundations of theoretical statistics. *Phil. Trans.* A, **222**, 309–68.

HALMOS, P. R. & SAVAGE, L. J. (1949). Applications of the Radon–Nikodym theorem to the theory of sufficient statistics. *Ann. Math. Statist.* **20**, 225–41.

MURTHY, M. N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhyā*, **18**, 379–90.

PATHAK, P. K. (1961). Use of 'order-statistic' in sampling without replacement. *Sankhyā*, **A, 23**, 409–14.

PATHAK, P. K. (1962). On sampling with unequal probabilities. *Sankhyā*, **A, 24**, 315–26.

PATHAK, P. K. (1964). Sufficiency in sampling theory (to be published in *Ann. Math. Statist.*).

RAO, J. N. K., HARTLEY, H. O. & COCHRAN, W. G. (1962). On a simple procedure of unequal probability sampling without replacement. *J. R. Statist. Soc.* B, **24**, 482–91.

SAMPFORD, M. R. (1962). Methods of cluster sampling with and without replacement for clusters of unequal sizes. *Biometrika*, **49**, 27–40.

STEVENS, W. L. (1958). Sampling without replacement with probability proportional to size. *J. R. Statist. Soc.* B, **20**, 393–7.