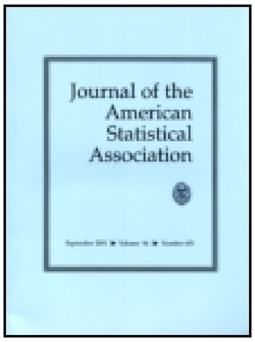

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Contribution to the Theory of Sampling Human Populations

J. Neyman

To cite this article: J. Neyman (1938) Contribution to the Theory of Sampling Human Populations, Journal of the American Statistical Association, 33:201, 101-116

To link to this article: https://doi.org/10.1080/01621459.1938.10503378

Published online: 08 May 2012.

Submit your article to this journal

Article views: 30

View related articles

Citing articles: 77 View citing articles

CONTRIBUTION TO THE THEORY OF SAMPLING HUMAN POPULATIONS

By J. Neyman
Department of Statistics, University College, London

1. INTRODUCTION

At a Conference on Sampling Human Populations held last April at the Department of Agriculture Graduate School in Washington, a problem was presented by Mr. Milton Friedman and Dr. Sidney Wilcox for which I could not offer a solution at the time. Since it seemed to be important and of general interest, I have considered it in some detail. The purpose of this paper is to present the results I have obtained.

2. STATEMENT OF THE PROBLEM

I shall start by describing the problem in much the same form as it was stated to me, without using any mathematical symbols. Then I shall formulate it in mathematical terms. The reader who does not wish to follow the mathematical processes may skip from equation (8) to the results and examples beginning with equation (52) on page 111.

A field survey is to be undertaken to determine the average value of some character of a population, for example, the amount of money which families spend for food in a population of families residing in a certain district. The collection of these data requires long interviews by specially trained enumerators and, hence, the cost per family is quite high. Since the total cost of the survey must be held within the amount appropriated for it, the data must be secured from a small sample of the population. In view of the great variability of the character, the sample appears to be too small to yield an estimate of the desired degree of accuracy.

Now the character is correlated with a second character which can be determined much more readily and at a low cost per family. Since a very accurate estimate of the second character can be secured at relatively small expense, and since for any given value of it, the variation of the original character will be smaller than it is in the whole population, a more accurate estimate of the original character may be obtained for the same total expenditure by arranging the sampling of the population in two steps. The first step is to secure data, for the second character only, from a relatively large random sample of the population in order to obtain an accurate estimate of the distribution of this character. The second step is to divide this sample, as in stratified sampling, into classes or strata according to the value of the second character and to draw at random from each of the strata a small sample for the costly intensive interviewing necessary to secure data regarding the first character.

An estimate of the first character based on these samples may be more accurate than one based on an equally expensive sample drawn at random without stratification. The question is to determine for a given expenditure, the sizes of the initial sample and the subsequent samples which yield the most accurate estimate of the first character.

Let us now enter into the details and introduce the necessary notation. Denote by π the population studied and by X the character of its individuals the average of which, say $\overline{X}$, is to be estimated. This is the character the collection of data on which is costly. Next let Y denote the second character, on which the collection of data is cheap, and which is assumed to be correlated with X. The range of variation of Y in π being more or less known, we shall divide it into s intervals, say

$$\text{from } Y_0 \text{ to } Y_1, \text{ from } Y_1 \text{ to } Y_2, \cdots, \text{ and from } Y_{s-1} \text{ to } Y_s. \tag{1}$$

Denote by π_i the part of the population π composed of the individuals for which

$$Y_{i-1} \leqq Y < Y_i \qquad (i = 1, 2, \cdots, s) ; \tag{2}$$

π_i will be called the ith stratum of the population π. Denote further by

$$p_1, p_2, \cdots, p_s \tag{3}$$

the proportions of the individuals of π belonging to the strata π_1, π_2, $\cdots$, π_s respectively.

In the following we shall have to consider three different processes of sampling which it is important to distinguish. The first two form the method described by Mr. Friedman and Dr. Wilcox, which I shall further describe as the method of double sampling. The third will serve as a standard of comparison of the accuracy of the method of double sampling. In order to avoid any misunderstanding let us describe all three in detail.

The method of double sampling consists of the following steps:

(i) Out of the population π we select at random N individuals and ascertain for them the values of the character Y. This sample will be denoted by S_1. The sample S_1 is meant to estimate the proportions p_i.

(ii) Now we proceed to sample the strata π_i and this is the second of the sampling processes mentioned. Out of each stratum π_i we select at random m_i individuals which form a sample to be denoted by $S_{2,i}$ and ascertain for each of these individuals, the value of the character X. The samples $S_{2,i}$ serve to estimate the mean value of X in each of the strata π_i. These estimates and the estimates of the proportions (3) obtained previously from the sample S_1, permits us to estimate the grand mean $\overline{X}$.

The combination of (i) and (ii) forms the method of double sampling. Denote by m_0 the sum of the sizes m_i of all the samples $S_{2,i}$, so that

$$m_0 = \sum_{i=1}^{s} m_i \tag{4}$$

and by A and B the costs of ascertaining for one individual the value of X and that of Y respectively. Finally, let C denote the total amount of money available for the collection of data. Then the numbers m_0 and N must be subject to the restriction

$$Am_0 + BN = C \, . \tag{5}$$

We shall consider what values of m_i, m_0 and N, satisfy the conditions (4) and (5), yield the greatest accuracy in estimating the mean value of X by the method of double sampling. This accuracy will then be compared with that attainable in the ordinary way, that is without the application of the method of double sampling. For this purpose we shall consider a third sampling process by which all the funds C available are spent on selecting at random a number, say M, of the elements of π and in ascertaining for each of them the value of X. Denote this third sample by S_0. Its size will have to be $M = C/A$. In order to get an idea of the utility of the method of double sampling we shall compare its accuracy with that of the ordinary mean value of X calculated from the sample S_0.

3. FIRST METHOD OF APPROACH

In the present paper[1] we shall make no assumption as to the character of the regression of X on Y in the population π. Denote by X_1, $X_2, \cdots, X_s$, the mean values of X in each of the strata. It follows that the grand mean of X which is to be estimated is

$$\overline{X} = \sum_{i=1}^{s} p_i X_i \, . \tag{6}$$

[1] The same problem, under the assumption that the regression of X on Y has a certain known form, *forming the second method of approach, will be considered in a later paper.*

Further denote by σ_i the standard deviation of X within the ith stratum.

Denote by n_i the number of individuals drawn in the first sample S_1 which fall within the ith stratum and introduce

$$r_i = n_i/N \ . \tag{7}$$

Let x_{ij} denote the value of X of the jth individual drawn from the ith stratum to form the sample $S_{2,i}$. Put

$$x_i = \frac{1}{m_i} \sum_{j=1}^{m_i} x_{ij} \ . \tag{8}$$

We shall start by considering what function F_1 of the observations, namely of the numbers (7) and of

$$x_{i1},\, x_{i2},\, \cdots,\, x_{im_i} \qquad \text{for} \qquad i = 1, 2, \cdots, s \tag{9}$$

would be suitable as an estimate of (6). We shall limit our considerations to homogeneous functions of second order, of the form

$$F_1 = \sum_{i=1}^{s} \sum_{j=1}^{s} \sum_{k=1}^{m_j} \lambda_{ijk} r_i x_{jk} \tag{10}$$

where λ_{ijk} is a constant coefficient. Out of all such functions we shall select and term the best unbiased estimate of $\overline{X}$, the one which has the following properties:

(i) The mathematical expectation of F_1 is identically equal to $\overline{X}$.

(ii) The variance of F_1 is smaller than that of any other function of the form (10) having the property (i).

Denoting by $\mathcal{E}(u)$ the mathematical expectation of any variable u, we may rewrite (i) in the following form.

$$\mathcal{E}(F_1) \equiv \sum_{i=1}^{s} \sum_{j=1}^{s} \sum_{k=1}^{m_j} \lambda_{ijk} \mathcal{E}(r_i x_{jk}) \equiv \sum_{i=1}^{s} p_i X_i \ . \tag{11}$$

When calculating expectations, we shall use the assumption that the population π and all of its strata are so large compared to the sample drawn that the particular drawings can be considered as mutually independent. We shall notice further that, in spite of the fact that the samples $S_{2,i}$ probably will be drawn out of the sample S_1 and not directly from the strata π_i, the variable x_{jk} is independent of r_i. This follows from the circumstance that when we draw the first sample S_1, we do so without any consideration of the values of X. It follows that

$$\mathcal{E}(r_i x_{jk}) = \mathcal{E}(r_i)\mathcal{E}(x_{jk}) = p_i X_j \,. \tag{12}$$

Substituting (12) in (11) and rearranging, we have

$$\sum_{i=1}^{s} p_i \left(\sum_{j=1}^{s} X_j \sum_{k=1}^{m_j} \lambda_{ijk} - X_i \right) \equiv 0 \,. \tag{13}$$

The necessary and sufficient condition for this equality to hold good identically, that is to say, whatever the unknown proportions p_1, $p_2, \cdots, p_3$ may be, is that the coefficients of the p_i vanish, i.e.

$$\sum_{j=1}^{s} X_j \sum_{k=1}^{m_j} \lambda_{ijk} - X_i \equiv 0 \qquad \text{for} \qquad i = 1, 2, \cdots, s \,. \tag{14}$$

As we do not know the values of the X_j, these equalities should again hold good identically, that is to say, whatever the values of the X_j. The equation (14) can be rewritten in the form

$$\sum_{j=1}^{i-1} X_j \sum_{k=1}^{m_j} \lambda_{ijk} + X_i \left(\sum_{k=1}^{m_i} \lambda_{iik} - 1 \right) + \sum_{j=i+1}^{s} X_j \sum_{k=1}^{m_j} \lambda_{ijk} \equiv 0 \tag{15}$$

and its identical fulfilment is easily seen to require that

$$\begin{aligned} &\sum_{k=1}^{m_j} \lambda_{ijk} = 0 \qquad \text{for any} \qquad j \neq i;\, i, j = 1, 2, \cdots, s \qquad \text{and} \\ &\sum_{k=1}^{m_i} \lambda_{iik} = 1 \qquad \text{for} \qquad i = 1, 2, \cdots, s \,. \end{aligned} \tag{16}$$

Equations (16) express the necessary and sufficient conditions for the function F_1 to be unbiased, considered as an estimate of $\overline{X}$. Obviously, there is an infinite number of systems of coefficients λ_{ijk} satisfying (16) and therefore an infinity of unbiased estimates of $\overline{X}$ of the form (10). We shall now determine the one that we agreed to call the "best," i.e., that which has the smallest possible variance. Let us assume that the values of the λ_{ijk} are fixed somehow satisfying the conditions (16) and calculate the variance of F_1. Denoting it by V_1 we shall have, owing to (6)

$$\begin{aligned} V_1 &= \mathcal{E}(F_1 - \overline{X})^2 \\ &= \mathcal{E}\left(\sum_{i=1}^{s} (r_i \xi_i - p_i X_i) \right)^2 \end{aligned} \tag{17}$$

where

$$\xi_i = \sum_{j=1}^{s} \sum_{k=1}^{m_j} \lambda_{ijk} x_{jk} \tag{18}$$

is again independent of r_i. We have further

$$\begin{aligned} V_1 = {} & \sum_{i=1}^{s} \mathcal{E}((r_i\xi_i - p_iX_i)^2) \\ & + 2\sum_{i=1}^{s-1} \sum_{h=i+1}^{s} \mathcal{E}\{(r_i\xi_i - p_iX_i)(r_h\xi_h - p_hX_h)\}. \end{aligned} \tag{19}$$

But

$$\begin{aligned} \mathcal{E}(r_i\xi_i - p_iX_i)^2 &= \mathcal{E}(\{(r_i - p_i)\xi_i + p_i(\xi_i - X_i)\}^2) \\ &= \mathcal{E}((r_i - p_i)^2\xi_i^2) + 2p_i\mathcal{E}\{(r_i - p_i)(\xi_i^2 - X_i\xi_i)\} \\ &\quad + p_i^2\mathcal{E}\{(\xi_i - X_i)^2\} \\ &= \mathcal{E}\{(r_i - p_i)^2\}\mathcal{E}(\xi_i^2) + p_i^2\mathcal{E}\{(\xi_i - X_i)^2\} \end{aligned} \tag{20}$$

owing to the independence of ξ_i and r_i and to the fact that $\mathcal{E}(r_i) = p_i$. Now it is known that

$$\mathcal{E}\{(r_i - p_i)^2\} = \mathcal{E}(r_i^2) - p_i^2 = p_iq_i/N \tag{21}$$

with $q_i = 1 - p_i$. Since[2]

$$\mathcal{E}\{(\xi_i - X_i)^2\} = \mathcal{E}(\xi_i^2) - X_i^2\,, \tag{22}$$

to calculate (20) it will be sufficient to calculate $\mathcal{E}\{(\xi_i - X_i)^2\}$ or the variance of ξ_i. Applying the usual formula for the variance of a linear function of independent variables and remembering that the variance of x_{jk} is denoted by σ_j^2, we have

$$\mathcal{E}\{(\xi_i - X_i)^2\} = \sum_{j=1}^{s} \sigma_j^2 \sum_{k=1}^{m_j} \lambda^2_{ijk}\,. \tag{23}$$

It follows that

$$\begin{aligned} \mathcal{E}\{(r_i\xi_i - p_iX_i)^2\} = {} & \frac{p_iq_i}{N}\left(\sum_{j=1}^{s} \sigma_j^2 \sum_{k=1}^{m_j} \lambda^2_{ijk} + X_i^2 \right) \\ & + p_i^2\sum_{j=1}^{s} \sigma_j^2 \sum_{k=1}^{m_j} \lambda^2_{ijk}\,. \end{aligned} \tag{24}$$

[2] Owing to (16) the expectation of ξ_i is obviously equal to X_i.

We may now go on and calculate the expectation of the other type of term in (19). We have

$$\begin{aligned}\mathcal{E}\{(r_i\xi_i - p_iX_i)(r_h\xi_h - p_hX_h)\} &= \mathcal{E}(r_ir_h\xi_i\xi_h) - p_ip_hX_iX_h \\ &= \mathcal{E}(r_ir_h)\mathcal{E}(\xi_i\xi_h) - p_ip_hX_iX_h\end{aligned} \tag{25}$$

again owing to the independence of ξ_i and r_i. It is known that

$$\mathcal{E}(r_ir_h) = p_ip_h\left(1 - \frac{1}{N}\right). \tag{26}$$

Further

$$\mathcal{E}(\xi_i\xi_h) = \mathcal{E}\left\{\sum_{j=1}^{s}\sum_{k=1}^{m_j}\lambda_{ijk}x_{jk}\sum_{g=1}^{s}\sum_{u=1}^{m_g}\lambda_{hgu}x_{gu}\right\}. \tag{27}$$

Remembering that

$$\mathcal{E}(x_{jk}) = X_j \qquad \text{and} \qquad \mathcal{E}(x^2_{jk}) = \sigma_j^2 + X_j^2 \tag{28}$$

and that the x_i are assumed to be mutually independent, we have

$$\mathcal{E}(\xi_i\xi_h) = \sum_{j=1}^{s}\sigma_j^2\sum_{k=1}^{m_j}\lambda_{ijk}\lambda_{hjk} + \left(\sum_{j=1}^{s}X_j\sum_{k=1}^{m_j}\lambda_{ijk}\right)\left(\sum_{g=1}^{s}X_g\sum_{u=1}^{m_g}\lambda_{hgu}\right). \tag{29}$$

Until the present moment we have not used the conditions (16) for the unbiased character of the estimate F_1. Therefore the formula for the variance V_1 which we could obtain by substituting (24), (25), (26) and (29) into (19) would be perfectly general. We shall use it in our second method of approach. Now however we shall simplify (29) by substituting (16). We have

$$\mathcal{E}(\xi_i\xi_h) = \sum_{j=1}^{s}\sigma_j^2\sum_{k=1}^{m_j}\lambda_{ijk}\lambda_{hjk} + X_iX_h\,. \tag{30}$$

Now

$$\begin{aligned}V_1 = \sum_{i=1}^{s}\left(p_i^2 + \frac{p_iq_i}{N}\right)\sum_{j=1}^{s}\sigma_j^2\sum_{k=1}^{m_j}\lambda^2_{ijk} + \sum_{i=1}^{s}\frac{p_iq_i}{N}X_i^2 \\ + \frac{2}{N}\sum_{i=1}^{s-1}\sum_{h=i+1}^{s}p_ip_h\left\{(N-1)\sum_{j=1}^{s}\sigma_j^2\sum_{k=1}^{m_j}\lambda_{ijk}\lambda_{hjk} - X_iX_h\right\}.\end{aligned} \tag{31}$$

Without attempting to simplify this expression at the present stage, let us select the λ_{ijk} so as to minimize (31) while keeping the relations

(16) satisfied. For this purpose we will differentiate with respect to λ_{ijk} the expression

$$f = V_1 - 2\sum_{i=1}^{s}\sum_{j=1}^{s} \alpha_{ij} \sum_{k=1}^{m_j} \lambda_{ijk} \tag{32}$$

where the α_{ij} are Lagrange arbitrary multipliers, and equate the derivatives to zero. After some rearrangement, we get the following equation:

$$\frac{1}{N} p_i\sigma_j^2\left(\lambda_{ijk} + (N-1)\sum_{h=1}^{s} p_h\lambda_{hjk}\right) = \alpha_{ij} . \tag{33}$$

Summing both sides with respect to k from zero to m_j and taking into account (16), we get

$$\begin{aligned} \frac{N-1}{N} p_jp_i\sigma_j^2 &= m_j\alpha_{ij} \qquad \text{for} \qquad i \neq j \\ \frac{1}{N} p_i\sigma_j(1 + (N-1)p_j) &= m_j\alpha_{jj} . \end{aligned} \tag{34}$$

Substituting these results in (33), we obtain

$$\lambda_{ijk} = \frac{N-1}{m_j} p_j - (N-1)\sum_{h=1}^{s} p_h\lambda_{hjk} = \lambda_{.jk} \quad \text{(say)} \tag{35}$$

$$\lambda_{jjk} = \lambda_{.jk} + 1/m_j . \tag{36}$$

Substituting in (35) the values of λ_{hjk} thus obtained, we easily get

$$\begin{aligned} \lambda_{ijk} &= \lambda_{.jk} = 0 \qquad \text{for} \qquad i \neq j \\ \lambda_{jjk} &= 1/m_j . \end{aligned} \tag{37}$$

Substituting these values into (10) we obtain the following expression for the best unbiased estimate of $\overline{X}$:

$$F_1 = \sum_{i=1}^{s} r_ix_i . \tag{38}$$

The formula for the variance, V_1, of F_1 is obtained by substituting (37) in (31)

$$V_1 = \sum_{i=1}^{s}\left\{\left(p_i^2 + \frac{p_iq_i}{N}\right)\frac{\sigma_i^2}{m_i} + \frac{p_iq_i}{N} X_i^2\right\} - \frac{2}{N}\sum_{i=1}^{s-1}\sum_{j=i+1}^{s} p_ip_jX_iX_j \tag{39}$$

which immediately reduces to the following form most convenient for finding the system of values of N and the m_i that assure the greatest accuracy is estimating $\overline{X}$:

$$V_1 = \frac{1}{m_0}\left(\sum_{i=1}^{s}\sigma_i\sqrt{p_i^2 + p_iq_iN^{-1}}\right)^2 \tag{40}$$
$$+ \sum_{i=1}^{s} m_i\left(\frac{\sigma_i\sqrt{p_i^2 + p_iq_iN^{-1}}}{m_i} - \frac{\sum_{i=1}^{s}\sigma_i\sqrt{p_i^2 + p_iq_iN^{-1}}}{m_0}\right)^2$$
$$+ \frac{1}{N}\sum_{i=1}^{s} p_i(X_i - \overline{X})^2 .$$

It is seen that none of the three terms in the right hand side can be negative. There is only one term which depends directly on m_1, m_2, $\cdots$, m_s, namely the second, the others being dependent on $m_0 = \sum_{i=1}^{s} m_i$ and on N. It follows that once N and m_0 are fixed in one way or another the value of V_1 depends on the m_i and the value they ascribe to the second term. It is easily seen that its minimum value is zero and that this is attained, whenever for each value of $i = 1, 2, \cdots, s$

$$m_i = m_0\sigma_i\sqrt{p_i^2 + p_iq_iN^{-1}}\Big/\sum \sigma_i\sqrt{p_i^2 + p_iq_iN^{-1}} . \tag{41}$$

Owing to the fact that the m_i are integers, this ideal seldom can be attained exactly, but it may be approached as far as possible. We shall further assume that the m_i are selected in closest agreement with (41) and that the second term in (40) is negligible compared with the remaining two.

We must now consider what values of m_0 and N satisfying (5) are likely to give the smallest value to the sum of only two terms in (40), say

$$V'_1 = \frac{1}{m_0}\left(\sum_{i=1}^{s}\sigma_i\sqrt{p_i^2 + p_iq_iN^{-1}}\right)^2 + \frac{1}{N}\sum_{i=1}^{s} p_i(X_i - \overline{X})^2 . \tag{42}$$

Owing to the complex structure of the first of these terms, an accurate solution of the problem is difficult to attain. However, it is easy to get an approximate solution which will probably in most cases be sufficient.

In most cases, whenever we do not make any special assumption concerning the character of the regression of X on Y, we shall probably classify the population π into only a few strata whence it may be assumed that the proportions p_i will not be very small and consequently

$p_iq_iN^{-1}$ will be considerably smaller than any of the p_i^2. If so, then the value of the square root

$$(43) \qquad \sqrt{p_i^2 + p_iq_iN^{-1}}$$

will be very much the same as that of p_i. For example, if $p_i=.1$, $q_i=.9$ and $N=100$, it is .1044 and if the value of Np_i were somewhat larger, the agreement would be still better. Therefore, instead of trying to minimize (42) we may usefully start by trying to minimize, say

$$(44) \qquad \begin{aligned} V_1'' &= \frac{1}{m_0}\left(\sum_{i=1}^{s} p_i\sigma_i\right)^2 + \frac{1}{N}\sum_{i=1}^{s} p_i(X_i - \overline{X})^2 , \qquad \text{or} \\ V_1'' &= \frac{a^2}{m_0} + \frac{b^2}{N} \end{aligned}$$

for short. Denote by v_1 and v_2 the smallest numbers of selections into the first and the second sample respectively, the total cost of which is the same, so that

$$(45) \qquad v_1B = v_2A \ .$$

If m'_0 and N' are the integer numbers minimizing (44) and satisfying (5), then any change of these values by taking instead of them either

$$(46) \qquad \begin{matrix} m_0' - v_2 & \text{and} & N' + v_1 & \text{or} \\ m_0' + v_2 & \text{and} & N' - v_1 & \end{matrix}$$

will increase the value of (44). This means that m_0' and N' satisfy the inequalities

$$(47) \qquad \frac{a^2}{m'_0 + v_2} + \frac{b^2}{N' - v_1} > \frac{a^2}{m'_0} + \frac{b^2}{N'} < \frac{a^2}{m'_0 - v_2} + \frac{b^2}{N' + v_1} \cdot$$

These inequalities reduce easily to the following ones

$$(48) \qquad \frac{1 - \dfrac{v_2}{m'_0}}{1 + \dfrac{v_1}{N'}} < \frac{a^2v_2}{m_0'^2}\,\frac{N'^2}{b^2v_1} < \frac{1 + \dfrac{v_2}{m'_0}}{1 - \dfrac{v_1}{N'}}$$

showing that in order to minimize (44) while keeping (5) fixed, we have to select m_0 and N as nearly as possible proportionately to $a\sqrt{v_2}$ and $b\sqrt{v_1}$ respectively. Putting for a moment

$$(49) \qquad m_0 = N(a/b)\sqrt{v_2/v_1}$$

and substituting it in (5), we get

$$N = Cb\sqrt{v_1}/(Aa\sqrt{v_2} + Bb\sqrt{v_1}) \tag{50}$$

which gives

$$m_0 = Ca\sqrt{v_2}/(Aa\sqrt{v_2} + Bb\sqrt{v_1}) \ . \tag{51}$$

Using (45) and eliminating v_1 and v_2 we may rewrite (50) and (51) in the final form

$$N = Cb/(a\sqrt{AB} + bB) \tag{52}$$

$$m_0 = Ca/(aA + b\sqrt{AB}) \tag{53}$$

where

$$a = \sum_{i=1}^{s} p_i\sigma_i \tag{54}$$

and

$$b^2 = \sum_{i=1}^{s} p_i(X_i - \overline{X})^2 \ . \tag{55}$$

Here we must remember the following circumstances:

(1) that both m_0 and N are integers and therefore formulae (52) and (53) should be calculated to the nearest integer;

(2) that a change in m_0 by one unit must be compensated by a change in N by several units;

(3) that the solutions which would be obtained by taking exact values of (52) and (53) would minimize the value of (44) with a as given in (54), whereas the value of the variance in (42) depends on

$$a_1 = \sum_{i=1}^{s} \sigma_i\sqrt{p_i^2 + p_iq_iN^{-1}} \tag{56}$$

instead of a.

It follows that the integers nearest to (52) and (53) may not necessarily minimize (42), but since the difference between a and a_1 is slight, they may be considered as the first approximations. Frequently these first approximations will also be the accurate values.

In order to find the second approximation, we may calculate a_1 as in (56) substituting N as calculated from (52) and then substitute the value obtained into (53) to get a new value of m_0. This sometimes will indicate the necessity of increasing the original m_0 by unity. However, owing to the fact that both m_0 and N must be integers, the real check of what values do give the minimum is obtained simply by substituting into (42) both the first approximations to m_0 and N and a few neighboring systems of values, e.g., m_0-1 and m_0+1 and the corresponding values of N.

4. EXAMPLE I

It may be useful to illustrate the above theory by some simple examples. Assume that there are only three strata, so that $s=3$. Assume further the following values of the constants involved:

$$\begin{aligned} p_1 &= \tfrac{1}{4}\,, & p_2 &= \tfrac{2}{4}\,, & p_3 &= \tfrac{1}{4}, \\ X_1 &= 1\,, & X_2 &= 3\,, & X_3 &= 6, \\ \sigma_1 &= 1\,, & \sigma_2 &= 2\,, & \sigma_3 &= 4, \\ A &= 4\,, & B &= 1\,, & C &= 500\,. \end{aligned} \tag{57}$$

In order to calculate the values of m_0 and N, we calculate

$$a = 2.25, \tag{58}$$

$$\bar{X} = 3.25, \tag{59}$$

$$b^2 = 3.1875 = (1.7854)^2\,. \tag{60}$$

It follows that

$$N = 142 \qquad \text{(to the nearest integer)} \tag{61}$$

and accordingly

$$m_0 = 89\,. \tag{62}$$

It will be seen that the necessity of taking m_0 to the nearest integer permits an increase in the value of N to 144, without exceeding the limit of expense, 500 units. Let us now see how $m_0=89$ should be distributed between the three strata. Easy calculations give

$$\begin{aligned} \sigma_1\sqrt{p_1{}^2 + p_1q_1N^{-1}} &= .2526 \\ \sigma_2\sqrt{p_2{}^2 + p_2q_2N^{-1}} &= 1.0035 \\ \sigma_3\sqrt{p_3{}^2 + p_3q_3N^{-1}} &= 1.0104 \\ \sum_{i=1}^{3}\sigma_i\sqrt{p_i{}^2 + p_iq_iN^{-1}} &= 2.2664\,. \end{aligned} \tag{63}$$

Hence, using (41) and taking the nearest integers, we get

$$m_1 = 10\,, \qquad m_2 = 39\,, \qquad m_3 = 40\,. \tag{64}$$

With this system of the m_i the middle term of formula (40) would have the value

$$(65)\qquad \sum_{i=1}^{s} m_i \left(\frac{\sigma_i\sqrt{p_i^2 + p_i q_i N^{-1}}}{m_i} - \frac{\sum \sigma_i\sqrt{p_i^2 + p_i q_i N^{-1}}}{m_0} \right)^2$$

$$= .0000048564\ .$$

The total value of V_1 in (40) is found to be

$$(66)\qquad V_1 = .079855$$

and it follows that by using (64) the value of the middle term is for all practical purposes negligible. It is interesting to compare this value with the one which could be obtained without adjusting the numbers m_i to the variability and the size of the strata, i.e., without using (41). Putting arbitrarily $m_1 = 29$, $m_2 = m_3 = 30$, we get

$$(67)\qquad V_1 = .091927\ .$$

Comparing this with (66) we see that neglecting to adjust the m_i according to formula (40) results, in this particular example, in an increase of the variance by over 15 per cent, which is a considerable and unnecessary loss in accuracy.

This is the situation if we use for m_0 and N the values found as first approximations. Substituting 144 for N in (56) and calculating a_1 and then using this value instead of a to calculate the second approximation of m_0, we get

$$(68)\qquad m_0 = 89.6783$$

which suggests that the best integer values of m_0 and N are $m_0 = 90$ and $N = 140$. However using them we obtain

$$(69)\qquad V_1 = .079866.$$

Again using $m_0 = 88$ and $N = 148$ we get

$$(70)\qquad V_1 = .079888$$

and it appears that the first approximation gives in fact the best possible result, but the actual difference is negligible.

We must now see whether this result, the best that could be obtained by the method of double sampling is actually better than what could be obtained by spending all the money available to collect as much data on X as possible, i.e. by drawing the unrestricted random sample S_0 (see p. 103).

The best linear estimate of $\overline{X}$ calculated from the sample S_0 would be the sample mean $\bar{x}$. Its variance, V_0, is known to be connected with the symbols of this paper by means of the formula

$$V_0 = \frac{1}{M}\left(\sum_{i=1}^{s} p_i\sigma_i^2 + \sum_{i=1}^{s} p_i(X_i - \overline{X})^2\right). \tag{71}$$

It is easy to find that in our example

$$M = C/A = 125 \tag{72}$$

and

$$V_0 = .0755\ . \tag{73}$$

It follows that in this particular case the method of double sampling even supplemented by the optimum adjustment of the numbers of sampling, is equivalent to a certain loss of accuracy of the final result. Taking the ratio of the variances (73) and (66)

$$V_1/V_0 = 1.058 \tag{74}$$

we see that this loss of accuracy amounts to nearly 6 per cent. This unfavourable result is, of course, due to the fact that the differentiation between the strata with respect to the values of X is small compared with the variability of the strata themselves and to the fact that the difference in the cost of obtaining data on X and Y is comparatively small. To illustrate this point let us consider the following examples.

5. EXAMPLE II

Assume that the values of the p_i, X_i and σ_i are exactly as in Example I and put

$$A = 40\ , \qquad B = 1\ , \qquad C = 5000 \tag{75}$$

so that the process of obtaining data on Y is now 40 times cheaper than that on X, while the ratio of C/A is the same as formerly. It follows that V_0 in this case will be exactly the same as formerly (73), but the minimum value of V_1 will change. We shall have

$$m_0 = 111\ , \qquad N = 560 \tag{76}$$

and, assuming that the m_i are fixed according to (41), we get finally

$$V_1 = .05147 \tag{77}$$

and it is seen that this value is exceeded by V_0 by more than 46 per cent!

6. EXAMPLE III

Here we shall keep the values of the p_i, the σ_i, and those of A, B and C as in Example I but change the values of the X_i so as to increase the value of b, namely put

(78) $$X_1 = 1 \,, \qquad X_2 = 6 \,, \qquad X_3 = 11 \,.$$

Then

(79 $$b^2 = 12.5 = (3.53553)^2$$

and

(80) $$V_0 = .1500 \,.$$

On the other hand, applying the method of double sampling and taking the optimum system of numbers of samplings, viz.,

(81) $$m_1 = 8 \,, \quad m_1 = 31 \,, \quad m_3 = 31 \,, \quad m_0 = 70 \,, \quad N = 220$$

we get

(82) $$V_1 = .1298 ,$$

a gain in accuracy in comparison with (80) of about 15 per cent

7. CONCLUSIONS

(i) The examples II and III show that under favorable conditions the method of double sampling is a very powerful tool of statistical research.

(ii) However, the advantages of methods are but rarely universal and in certain cases, as for instance in the above example I, the direct unrestricted sampling may be more efficient than the method of double sampling.

(iii) Without a certain previous knowledge of the properties of the population sampled it is impossible to say which of the two methods will be more efficient.

(iv) It is also impossible to tell in advance what the values of N, m_0, and of the m_i should be to assure the greatest accuracy of the double sampling method.

(v) On the other hand, if certain properties of the sampled population π are known, or can be estimated, then it is possible to estimate the values of m_0 and N and also those of the m_i by which the method of double sampling gives the greatest possible accuracy. The properties of population π needed for this purpose are the values of the p_i, σ_i and X_i. They could be estimated by means of a preliminary inquiry on the lines suggested by me during the conference at the U. S. Department of Agriculture Graduate School and also in my previous publications on sampling human populations.[3] Once approximate values of the p_i, σ_i and X_i are obtained, they should be substituted into formulae (52), (53) and (41) to obtain the approximations of the optimum values of m_0, N and the m_i.

(vi) Before deciding whether to apply the method of double sampling, we should see that the prospects are that it will give better results than the direct unrestricted sampling of values of X.

For this purpose the approximate values of the p_i, σ_i and X_i should be substituted into (40) and (71) to obtain the approximate values of variance V_1 and V_0. The decision to apply the method of double sampling should be taken only if the approximate value of V_1 proves to be considerably smaller than that of V_0.

(vii) The steps described in (iii) and (iv) are possible only if some previous knowledge of the population π is available. This may be obtained from various sources: from some previous experience concerning the population π, or from a specially arranged preliminary inquiry. Such a preliminary inquiry consists of drawing from π a relatively small unrestricted random sample of individuals and in ascertaining for all of them the values of both characters under consideration X and Y. The data thus obtained should be used to estimate the p_i, the σ_i and the X_i.

In order to exemplify the kind of previous experience which may be used to plan future inquiries on the lines as indicated in (v) and (vi), I may mention a recent extensive Study of Consumer Purchases, a Federal Works Project administered by the Bureau of Labor Statistics, U. S. Department of Labor and the Bureau of Home Economics, U. S. Department of Agriculture, in cooperation with the National Resources Committee and the Central Statistical Board.[4] This inquiry was carried out by method of double sampling and therefore, in the process of working out the data, both the proportions p_i and the means X_i corresponding to particular strata and to many a character X must have been estimated. Probably the values of σ_i are also available. These figures could be used as pointed out in (v) and (vi) when planning any new inquiry concerning the same characters and the same or some similar population.

[3] J. Neyman: "An Outline of the Theory and Practice of Representative Method Applied in Social Research." Institute for Social Problems, Warsaw 1933. Polish with an English Summary.

J. Neyman: "On the Two Different Aspects of the Representative Method." J.R.S.S. 1934, pp. 558–625.

See also P. V. Sukhatme: "Contribution to the Theory of the Representative Method." Supplement to the J.R.S.S., Vol. II, 1935, pp. 253–268.

[4] This JOURNAL, Vol. XXXI, 1936, p. 135 and Vol. XXXII, 1937, p. 311.