SANKHYĀ

THE INDIAN JOURNAL OF STATISTICS

Edited by : P. C. MAHALANOBIS

VOL. 17, PART 1 JUNE 1956

SOME CONTRIBUTIONS TO THE DESIGN OF SAMPLE SURVEYS

By TOSIO KITAGAWA

Kyusyu University, Fukuoka
and
Indian Statistical Institute, Calcutta

PART IV. EXACT SAMPLING THEORIES AND ANALYSIS OF VARIANCE SCHEMES ASSOCIATED WITH DESIGNS OF SAMPLE SURVEYS*

1. INTRODUCTION

The current theories are concerned at least as their first approximate approaches with the errors of estimates arising solely from variations of random sampling. In actual large-scale surveys, where various sources of errors should be expected, our theoretical formulations should be so broad as to cover some of the important features such as failures to cover some of the units in the chosen sample, errors of measurements and biases due to the observers. It would not be sound to consider all these features at once; we should rather gradually proceed from the simplest to the most complicated one. Thus in this part, we propose to discuss a certain mathematical formulation in which some sort of statistical inference theories concerning sample surveys could be established. It may be readily observed that our assumptions take into consideration errors of measurements on a unit and also some slight time-changes in the populations which may always be expected in actual situations. To make our analysis simpler we shall begin with the normality assumption, which, however, can be replaced by other more general assumptions. Our main point is to overcome certain characteristic difficulties associated with the finiteness of the population and sampling without replacement by formulating our finite population in a more realistic sense. Among several authors Cochran (1953) suggested the validities of confidence intervals

* Parts I to III published in *Sankhyā*, 14 ,317-362.

based on t-distributions and made some comments in favour of validity of the normal approximation. It should, however, be pointed out that the results due to several authors about finite population cannot be applied to demonstrate any validity of the t-distribution so far as they are concerned with the case where the size of the sample becomes infinity. In a previous paper (Kitagawa, 1950b) we have discussed an application of the two-sample theory to statistical inferences for finite populations. The standpoint of the two-sample theory may sometimes be useful in this formulation but there still remain some sort of artificialities. Another formulation which seems in some respects more natural is to appeal to a subsampling scheme in which we shall assume an infinite grand population Π from which our finite population N should be drawn and consequently our sample of size n should be recognised as a subsample from Π. In fact D. Basu in a seminar held at the Indian Statistical Institute in June 1953 pointed out our assumptions (1^a), (2^a) and (3^a) were just equivalent to this subsampling procedure from the normal grand population Π. One might not yet be perfectly satisfied with the assumption of the existence of one grand population. In fact so far as sample surveys are concerned, there are real difficulties in imagining possible infinite villages, districts and areas in crops. The new assumptions which will be introduced in §2 and from which our inference theories may be developed may form the basis of an approach in which both the characteristic features are taken into consideration, that is, the errors of measurements and the finiteness of the population.

The second aim of this part is to provide the analysis of variance schemes applicable in sample surveys which can be duly discussed only after establishing some sort of exact sampling theories, so far as their applications are treated from the stochastic standpoint.

2. Subsample formulations and validities of t and z distributions for a finite population

Let us introduce the following assumptions:

Assumption I: Let us consider a set of N grand populations

$$\{\Pi_k\} \ (k = 1, 2, \ldots, N)$$

Assumption II: From each of the N grand populations a sample of size one shall be drawn independently, which we shall denote by $y_k (k = 1, 2, \ldots, N)$.

Assumption III: Let us draw a sample of size n without replacement from the set of size N: $\{y_k\}$ $(k = 1, 2, \ldots, N)$, and let us denote this sample by $\{\dot{y}_i\}$ $(i = 1, 2, \ldots, n)$.

Assumption IV: The grand population Π_j has normal distribution $N(\xi_j, \sigma^2)$ $(j = 1, 2, \ldots, N)$ where σ^2 is common to all the N grand populations.

Lemma 4.1: *The joint probability distribution of* $(\dot{y}_1, \dot{y}_2, \ldots, \dot{y}_n)$ *is for any set of real numbers* $(x_1, x_2, \ldots, x_n)$ given by

$$Pr.\{\dot{y}_1 < x_1, \dot{y}_2 < x_2, \ldots, \dot{y}_n < x_n\}$$

$$= \frac{1}{{}_N P_n} \sum_{(i_1, i_2, \ldots i_n)} Pr.\{y_{i_1} < x_1\} Pr.\{y_{i_2} < x_2\} \ldots Pr.\{y_{i_n} < x_n\} \qquad \ldots (2.01)$$

where $(i_1, i_2, \ldots, i_n)$ *means a permutation taking n elements from* $(1, 2, \ldots, N)$ *and the summation runs through all the permutations* ${}_N P_n$.

(a) *The joint distribution of the sample mean and sample variance.* We shall give here the joint distribution of the sample mean $\bar{y} = (\dot{y}_1 + \dot{y}_2 + \ldots + \dot{y}_n)/n$ and sample variance $s^2 = \Sigma(\dot{y}_i - \bar{y})^2/(n-1)$.

In view of Weibull (1950) and Lemma 4.1, we shall observe

Lemma 4.2: *The characteristic function of the joint distribution of* $\bar{y}$ *and* s^2 *is given by*

$$\frac{1}{{}_N P_n} \sum_{\alpha} f_\alpha(t_1, t_2) \qquad \ldots (2.02)$$

where $\alpha = (i_1, i_2, \ldots, i_n)$ *runs through all permutations of n elements from* $(1, 2, \ldots, N)$ *and the characteristic function* $f_\alpha(t_1, t_2)$ *is defined for each fixed permutation* α *such that*

$$f_\alpha(t_1, t_2) = \left(1 - \frac{2\sigma^2 i t_2}{n-1}\right)^{-\frac{n-1}{2}} \exp\left\{\frac{\lambda_\alpha^2 i t_2}{1 - 2\sigma^2 i t_2 (n-1)^{-1}}\right\} \times$$

$$\times \exp\{\bar{\xi}_\alpha i t_1 - \sigma^2 t_1^2 (2n)^{-1}\}, \qquad \ldots (2.03)$$

where

$$\lambda_\alpha^2 = \sum_{j=1}^{n} (\xi_{i_j} - \bar{\xi}_\alpha)^2 \qquad \ldots (2.04)$$

and

$$\bar{\xi}_\alpha = n^{-1} \sum_{j=1}^{n} \xi_{i_j}. \qquad \ldots (2.05)$$

Our result seems at first glance to be very complicated, but it may be readily observed that the average shown in (2.02) will sometimes make our formulae simplified.

(b) *The t-distribution.* Now let us define the statistic t by

$$t = \sqrt{n}(\bar{y} - \mu)/s \qquad \ldots (2.06)$$

and let us find out the distribution function of t. For each fixed permutation $\alpha = (i_1, i_2, \ldots, i_n)$ we may write

$$t_\alpha = \sqrt{n}(\bar{y}_\alpha - \mu)/s_\alpha. \qquad \ldots (2.07)$$

Lemma 4.3: *For each fixed x, the probability density function of the joint distribution of $\bar{y}_n$ and s_n is given by*

$$\left(\frac{n}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left\{-n\frac{(\bar{y}_n-\bar{\xi}_n)^2}{2\sigma^2}\right\} \exp\left\{-\frac{(n-1)\lambda_n^2}{2\sigma^2}\right\} \times$$

$$\times \sum_{r=0}^{\infty} \frac{1}{r!}\left\{\frac{(n-1)\lambda_n^2}{2\sigma^2}\right\}^r \frac{(s_n^2)^{\frac{n-1}{2}+r-1} \exp\left\{-\frac{(n-1)s_n^2}{2\sigma^2}\right\}}{\Gamma\left(\frac{n-1}{2}+r\right)\left(\frac{2\sigma^2}{n-1}\right)^{\frac{n-1}{2}+r}}. \quad \ldots (2.08)$$

This is due to Weibull (1950).

Now we shall turn to the distribution of t_n which wears the character of non-centrality, as we shall show in Lemma 4.4.

Lemma 4.4: *For each fixed x the elementary probability function of the statistic $t_n^2/(n-1)$ is given by*

$$\phi_n(t_n^2/(n-1))d(t_n^2/(n-1)) = \exp\left\{-\frac{(n-1)\lambda_n^2}{2\sigma^2}\right\} \exp\left\{-\frac{n\delta_n^2}{2\sigma^2}\right\} \times$$

$$\times \sum_{r=0}^{\infty} \left\{\frac{(n-1)\lambda_n^2}{2\sigma^2}\right\}^r \frac{1}{\Gamma(r+1)} \left\{\sum_{k=0}^{\infty} (-1)^k \left(\frac{n\delta_n^2}{2\sigma^2}\right)^{\frac{k}{2}} \frac{1}{\Gamma\left(\frac{k}{2}+1\right)} \times\right.$$

$$\left. \times \frac{\Gamma\left(\frac{n+k}{2}+r\right)}{\Gamma\left(\frac{n-1}{2}+r\right)\Gamma\left(\frac{k+1}{2}\right)} \frac{(t_n^2(n-1)^{-1})^{\frac{k}{2}}}{\left(1+t_n^2(n-1)^{-1}\right)^{\frac{n+k}{2}+r}}\right\} d\left(t_n^2(n-1)^{-1}\right) \quad \ldots (2.09)$$

where

$$\delta_n = \bar{\xi}_n - \bar{\xi} = n^{-1}\sum_{j=1}^{n} \xi_{ij} - N^{-1}\sum_{h=1}^{N} \xi_h. \quad \ldots (2.10)$$

The proof may be obtained as follows. First let us write $\bar{y}_n = \mu + \delta_n + ts_n/\sqrt{n}$ and let us make a change of variables $(\bar{y}_n, s_n)$ into (t_n, s_n) in (2.08). The integration of (2.08) with respect to s_n in $0 < s_n < \infty$ can be done term by term in the expansion of $\exp\{-n^{\frac{1}{2}}\delta_n t_n s_n/\sigma^2\}$ into power series.

Lemma 4.5: *The characteristic function of the statistic $t = \sqrt{n}(\bar{y}-\bar{\xi})/s$ is given by*

$$f(\tau) = \frac{1}{N^n}\sum_{a=(i_1,\ldots,i_N)} E\{\exp(i\tau t_a)\} \quad \ldots (2.11)$$

where

$$E\{\exp(i\tau t_a)\} = \int_{-\infty}^{\infty} e^{i\tau t_a}\phi_a\left(\frac{t_a^2}{n-1}\right)\frac{2t_a}{n-1}\,dt_a \quad \ldots (2.12)$$

with $\phi_a(t_a^2/(n-1))$ enunciated in (2.09).

3. Examples of exact sampling distributions

Our methods are those which were adopted in Part VI of our previous paper (Kitagawa, 1951a) and which appeal to the essential assumptions (1°), (2°) and (3°). Here we shall enunciate the following fundamental exact sampling distributions.

In our finite population, inferences should be concerned with the one in which some inference about the finite population with the elements $(y_j)(j = 1,2,\ldots,N)$ should be given in view of $\{y_{i_i}\}(i = 1, 2, \ldots, n)$. In the present formulation depending upon our assumptions (1°) to (4°), we can divide $\{y_j\}$ into two classes of which one consists of $\{y_{i_i}\}$ $(j = 1, 2, \ldots, n)$, the other being those remaining for each permutation $\alpha = (i_1, i_2, \ldots, i_n)$.

Thus for each fixed permutation α, the difference between the mean of the finite population $\bar{y} = N^{-1}(y_1+\ldots+y_N)$ and the sample mean $\bar{y} = n^{-1}(y_{i_1}+y_{i_2}+\ldots+y_{i_n})$ is independently distributed according to the normal distribution $N(\bar{\xi}-\bar{\xi}_\alpha, \sigma^2(N-n)(Nn)^{-1})$.

Consequently we shall have immediately, in view of Lemmas 4.3 and 4.4, the following.

Lemma 4.6: (1) *For each fixed permutation α, the probability density function of the joint distribution of $\bar{y}_\alpha-\bar{y} = z_\alpha$ and s_α is given by*

$$\left(\frac{Nn}{2\pi(N-n)\sigma^2}\right)^{\frac{1}{2}} \exp\left\{-\frac{Nn(z_\alpha-\delta_\alpha)^2}{2(N-n)\sigma^2}\right\} \exp\left\{-\frac{(n-1)\lambda_\alpha^2}{2\sigma^2}\right\}\times$$

$$\times\sum_{r=0}^{\infty}\frac{1}{\Gamma(r+1)}\left(\frac{(n-1)\lambda_\alpha^2}{2\sigma^2}\right)^r \frac{(s_\alpha^2)^{\frac{n-1}{2}+r-1}\exp\left\{-\frac{(n-1)s_\alpha^2}{2\sigma^2}\right\}}{\Gamma\left(\frac{n-1}{2}+r\right)\left(\frac{2\sigma^2}{n-1}\right)^{\frac{n-1}{2}+r}} \qquad \ldots (3.01)$$

(2) *For each fixed permutation α, let us define the statistic t_α^* by*

$$t_\alpha^* = (\bar{y}_\alpha-\bar{y}) \Big/ s_\alpha\sqrt{\frac{N-n}{Nn}}. \qquad \ldots (3.02)$$

Then the probability density function of $t_\alpha^{*2}/(n-1)$ is given by

$$\varphi_\alpha^*\left(\frac{t_\alpha^{*2}}{n-1}\right) = \exp\left\{-\frac{(n-1)\lambda_\alpha^2}{2\sigma^2}\right\}\exp\left\{-\frac{nN\delta_\alpha^2}{2(N-n)\sigma^2}\right\}\times$$

$$\times\sum_{r=0}^{\infty}\left(\frac{(n-1)\lambda_\alpha^2}{2\sigma^2}\right)^r \frac{1}{\Gamma(r+1)}\sum_{k=0}^{\infty}(-1)^k\left(\frac{nN\delta_\alpha^2}{2(N-n)\sigma^2}\right)^{\frac{k}{2}}\frac{1}{\prod\left(\frac{k}{2}+1\right)}\times$$

$$\times\frac{\prod\left(\frac{n+k}{2}+r\right)}{\prod\left(\frac{n-1}{2}+r\right)\prod\left(\frac{k+1}{2}\right)}\frac{\left(t_\alpha^*(n-1)^{-1}\right)^{\frac{k}{2}}}{(1+t_\alpha^{*2}(n-1)^{-1})^{\frac{n+k}{2}+r}}. \qquad \ldots (3.03)$$

Consequently, we shall reach the following theorem which will play a fundamental role concerning the confidence interval of the population mean $\bar{y}$ and which will show how and under what conditions the current uses of the t-distribution remain valid.

Theorem 4.1: *The characteristic function of the statistic*

$$t^{\alpha} = (\bar{y}-\bar{y}) \Big/ s\sqrt{\frac{N-n}{Nn}} \qquad \ldots \quad (3.04)$$

is given by

$$f(\tau) = \frac{1}{N^{\alpha}P_{\alpha}} \sum_{\alpha=(i_1,\ldots,i_n)} E\{\exp(i\tau t_{\alpha}^{\alpha})\} \qquad \ldots \quad (3.05)$$

where $$E\{\exp(i\tau t_{\alpha}^{\alpha})\} = \int_{-\infty}^{\infty} e^{i\tau t_{\alpha}^{\alpha}} \phi_{\alpha}^{\alpha}\left(\frac{t_{\alpha}^{2}}{n-1}\right)^{\frac{n-3}{2}} dt_{\alpha} \qquad \ldots \quad (3.06)$$

with ϕ_{α}^{α} *defined in* (3.03).

Next let us proceed to make inferences about the variance of a finite population by means of the sample variance. For each assigned permutation $\alpha = (i_1, i_2, \ldots, i_n)$, let us denote for the sake of simplicity $(y_{ij})(j = 1, 2, \ldots, n)$ by $y_{11}^{(\alpha)}, y_{12}^{(\alpha)}, \ldots, y_{1n}^{(\alpha)}$ and the remaining ones by $y_{21}^{(\alpha)}, y_{22}^{(\alpha)}, \ldots, y_{2,N-n}^{(\alpha)}$. Let us put

$$ns_1^{(\alpha)2} = \sum_{k=1}^{n} (y_{1k}^{(\alpha)} - \bar{y}_{1.}^{(\alpha)})^2 = \chi_1^{(\alpha)2}, \qquad \ldots \quad (3.07)$$

$$(N-n)s_2^{(\alpha)2} = \sum_{k=1}^{N-n} (y_{2k}^{(\alpha)} - \bar{y}_{2.}^{(\alpha)})^2 = \chi_2^{(\alpha)2}, \qquad \ldots \quad (3.08)$$

$$s_3^{(\alpha)2} = (N-n)nN^{-1}(\bar{y}_{1.}^{(\alpha)} - \bar{y}_{2.}^{(\alpha)})^2 = \chi_3^{(\alpha)2}, \qquad \ldots \quad (3.09)$$

$$Ns^{(\alpha)2} = ns_1^{(\alpha)2} + (N-n)s_2^{(\alpha)2} + s_3^{(\alpha)2}. \qquad \ldots \quad (3.10)$$

Then what we have to infer from $s_1^{(\alpha)2}$ is concerned with $s^{(\alpha)2}$. These $ns_1^{(\alpha)2}$, $(N-n)s_2^{(\alpha)2}$ and $s_3^{(\alpha)2}$ are known to be independently distributed according to the non-central chi-square distributions whose characteristic functions are

$$(1-2\sigma^2 it)^{-\frac{n-1}{2}} \exp\{S_1^{(\alpha)2}it/(1-2\sigma^2 it)\}, \qquad \ldots \quad (3.11)$$

$$(1-2\sigma^2 it)^{-\frac{N-n-1}{2}} \exp\{S_2^{(\alpha)2}it/(1-2\sigma^2 it)\}, \qquad \ldots \quad (3.12)$$

$$(1-2\sigma^2 it)^{-\frac{1}{2}} \exp\{S_3^{(\alpha)2}it/(1-2\sigma^2 it)\} \qquad \ldots \quad (3.13)$$

where
$$S_1^{(s)} = \sum_{j=1}^{n} (m_{1j}^{(s)} - \bar{m}_{1.}^{(s)})^2, \qquad \ldots (3.14)$$

$$S_2^{(s)} = \sum_{j=1}^{N-n} (m_{2j}^{(s)} - \bar{m}_{2.}^{(s)})^2, \qquad \ldots (3.15)$$

$$S_3^{(s)} = n(\bar{m}_{1.}^{(s)} - m)^2 + (N-n)(\bar{m}_{2.}^{(s)} - m)^2. \qquad \ldots (3.16)$$

For our present purpose let us first notice the results (5.24) due to Weibull (1950) whose transformation will yield

Theorem 4.2: *The probability density function of the statistic*

$$w = \frac{(n-1)\sum_{i=1}^{n} (y_i - \bar{y})^2}{(N-1)\sum_{i=1}^{n} (\dot{y}_i - \bar{y})^2} \qquad \ldots (3.17)$$

is given by
$$\frac{1}{{}_NP_n} \sum_{s} g_s(w) \qquad \ldots (3.18)$$

with
$$g_s(w)dw$$

$$= \exp\left\{-\frac{S_1^{(s)} + S_2^{(s)} + S_3^{(s)}}{2\sigma^2}\right\} \left[\sum_{r=0}^{\infty} \sum_{s=0}^{\infty} \left(\frac{S_1^{(s)}}{2\sigma^2}\right)^r \left(\frac{S_2^{(s)} + S_3^{(s)}}{2\sigma^2}\right)^s \times \right.$$

$$\left. \times \frac{1}{r!s!} \cdot \frac{\Gamma\left(\frac{N-1}{2} + r + s\right)}{\Gamma\left(\frac{n-1}{2} + r\right)\Gamma\left(\frac{N-n}{2} + s\right)} \cdot \frac{N-n}{n-1} \cdot \frac{(z(N-n)(n-1)^{-1})^{\frac{N-n}{2}+r-1}}{(1+z(N-n)(n-1)^{-1})^{\frac{N-1}{2}+r+s}} \right] dz. \qquad \ldots (3.19)$$

where
$$\frac{N-n}{n-1} z = \frac{N-1}{n-1} w - 1 \qquad \ldots (3.20)$$

and z *runs through* $0 \leqslant z < \infty$ *while* $(n-1)(N-1)^{-1} \leqslant w < \infty$.

4. Analysis of variance applied to a finite population

The uses of analysis of variance in designs and analysis of sample surveys are well recognised. To establish some theory of inference we shall introduce the following assumptions:

Assumption 1. Let us consider a set of MN grand populations

$$\{\Pi_{rs}\} \ (r = 1, 2, \ldots, M; \quad s = 1, 2, \ldots, N).$$

Assumption II: From each grand population Π_r a sample of size one shall be independently drawn, which we shall denote by $y_{rs} (r = 1, 2, ..., M;\ s = 1, 2, ..., N)$.

Assumption III: Let us draw a sample of size m $(i_1, i_2, ..., i_m)$ from the set $(1, 2, ..., M)$ without replacement, and also a sample of size n $(j_1, j_2, ..., j_n)$ from the set $(1, 2, ..., N)$ without replacement. Let these two samplings be independent. Let us consider a sample of size mn $\{y_{i_k j_l}\} (k = 1, 2, ..., m;\ l = 1, 2, ..., n)$.

Assumption IV: The grand populations Π_r have the normal distributions $N(\xi_{rs}, \sigma^2)$, where σ^2 is common to all these MN distributions for $r = 1, 2, ..., M$; $s = 1, 2, ..., N$.

Thus our subsampling procedure will yield as a set of mn values which depend upon a combination γ of the two permutations $\alpha = (i_1, i_2, ..., i_m)$, and $\beta = (j_1, j_2, ..., j_n)$.

In order to simplify our notations, let it be assumed that $i_r = r$ $(r = 1, 2, ..., m)$ and $j_s = s$ $(s = 1, 2, ..., n)$. Let i, j, u and v be natural numbers such that $1 \leqslant i \leqslant M$, $1 \leqslant j \leqslant N$, $m+1 \leqslant u \leqslant M$ and $n+1 \leqslant v \leqslant N$ respectively. Thereafter a sample $y_{i_k j_l}$ will be denoted by $x_{rs}^{(\gamma)}$ while the other y's by $x_{ms}^{(\gamma)}$, $x_{us}^{(\gamma)}$ and $x_{uv}^{(\gamma)}$ respectively and their respective means by

$$\bar{x}_{rn}^{(\gamma)} = n^{-1} \sum_{s=1}^{n} x_{rs}^{(\gamma)}, \qquad \ldots (4.01)$$

$$\bar{x}_{un}^{(\gamma)} = n^{-1} \sum_{s=1}^{n} x_{us}^{(\gamma)}, \qquad \ldots (4.02)$$

$$\bar{x}_{r,n-N}^{(\gamma)} = (N-n)^{-1} \sum_{s=n+1}^{N} x_{rs}^{(\gamma)}, \qquad \ldots (4.03)$$

$$\bar{x}_{rN}^{(\gamma)} = N^{-1} \sum_{j=1}^{N} x_{rj}^{(\gamma)}, \qquad \ldots (4.04)$$

$$\bar{x}_{mN}^{(\gamma)} = N^{-1} \sum_{j=1}^{N} \bar{x}_{mj}^{(\gamma)} = (mN)^{-1} \sum_{r=1}^{m} \sum_{j=1}^{N} x_{rj}^{(\gamma)}, \qquad \ldots (4.05)$$

$$\bar{x}_{MN}^{(\gamma)} = (MN)^{-1} \sum_{i=1}^{M} \sum_{j=1}^{N} x_{ij}^{(\gamma)} = \bar{y}_{..} = \bar{y}_{MN}, \qquad \ldots (4.06)$$

and similarly for other means such as $\bar{x}_{mi}^{(\gamma)}$, $\bar{x}_{M-m,s}^{(\gamma)}$, $\bar{x}_{M-m,n}^{(\gamma)}$, $\bar{x}_{Mn}^{(\gamma)}$ and so on.

Now our objects of statistical inferences concerning finite populations will be concerned with all or some of the sums of squares

$$S_T(M, N) \equiv \sum_{i=1}^{M} \sum_{j=1}^{N} (y_{ij} - \bar{y}_{MN})^2 \qquad \ldots (4.07)$$

$$S_R(M, N) \equiv N \sum_{i=1}^{M} (\bar{y}_{iN} - \bar{y}_{MN})^2 \qquad \ldots (4.08)$$

$$S_C(M, N) \equiv M \sum_{j=1}^{N} (\bar{y}_{Mj} - \bar{y}_{MN})^2 \qquad \ldots (4.09)$$

$$S_W(M, N) \equiv \sum_{i=1}^{M} \sum_{j=1}^{N} (\bar{y}_{ij} - \bar{y}_{iN} - \bar{y}_{Mj} + \bar{y}_{MN})^2 \qquad \ldots (4.10)$$

while our estimators will be all or some of the following sums of squares which can be calculated from our sample

$$S_T^{(\gamma)}(m, n) \equiv \sum_{r=1}^{m} \sum_{s=1}^{n} (x_{rs}^{(\gamma)} - \bar{x}_{mn}^{(\gamma)})^2 \qquad \ldots (4.11)$$

$$S_R^{(\gamma)}(m, n) \equiv n \sum_{r=1}^{m} (\bar{x}_{rn}^{(\gamma)} - \bar{x}_{mn}^{(\gamma)})^2 \qquad \ldots (4.12)$$

$$S_C^{(\gamma)}(m, n) \equiv m \sum_{s=1}^{n} (\bar{x}_{ms}^{(\gamma)} - \bar{x}_{mn}^{(\gamma)})^2 \qquad \ldots (4.13)$$

$$S_W^{(\gamma)}(m, n) \equiv \sum_{r=1}^{m} \sum_{s=1}^{n} (x_{rs}^{(\gamma)} - \bar{x}_{rn}^{(\gamma)} - \bar{x}_{ms}^{(\gamma)} + \bar{x}_{mn}^{(\gamma)})^2 \qquad \ldots (4.14)$$

respectively.

The relation between $S_T(M, N)$ and $S_T^{(\gamma)}(m, n)$ is quite simple. Indeed the division of the totality of $M \times N$ elements into the four parts which consist of $m \times n$, $m \times (N-n)$, $(M-m) \times n$ and $(M-m) \times (N-n)$ elements respectively, will lead to the following analysis of variance:

$$\begin{aligned} S_T(M, N) = {} & S_T^{(\gamma)}(m, n) + S_T^{(\gamma)}(m, N-n) + \\ & + S_T^{(\gamma)}(M-m, n) + S_T^{(\gamma)}(M-m, N-n) + \\ & + S_B^{(\gamma)}(m, n; M, N), \end{aligned} \qquad \ldots (4.15)$$

where

$$S_T^{(\gamma)}(m, N-n) = \sum_{r=1}^{m} \sum_{v=n+1}^{N} (x_{rv}^{(\gamma)} - \bar{x}_{m,N-n}^{(\gamma)})^2 \qquad \ldots (4.16)$$

$$S_T^{(\gamma)}(M-m, n) = \sum_{u=m+1}^{M} \sum_{s=1}^{n} (x_{us}^{(\gamma)} - \bar{x}_{M-m,n}^{(\gamma)})^2 \qquad \ldots (4.17)$$

$$S_B^{(\gamma)}(M-m, N-n) = \sum_{\alpha=m+1}^{M} \sum_{\gamma=n+1}^{N} (x_{\alpha\gamma}^{(\gamma)} - \bar{x}_{M-m,N-n}^{(\gamma)})^2 \quad \ldots (4.18)$$

$$S_R^{(\gamma)}(m, n; M, N) = S_B^{(\gamma)}(m, n) + S_B^{(\gamma)}(M-m, n) + S_H^{(\gamma)}(m, N-n) + S_B^{(\gamma)}(M-m, N-n) \quad \ldots (4.19)$$

in which $$S_B^{(\gamma)}(m, n) \equiv mn(\bar{x}_{mn}^{(\gamma)} - \bar{x}_{MN}^{(\gamma)})^2 \quad \ldots (4.20)$$

$$S_B^{(\gamma)}(m, N-n) \equiv m(N-n)(\bar{x}_{m,N-n}^{(\gamma)} - \bar{x}_{MN}^{(\gamma)})^2 \quad \ldots (4.21)$$

$$S_B^{(\gamma)}(M-m, n) \equiv (M-m)n(\bar{x}_{M-m,n}^{(\gamma)} - \bar{x}_{MN}^{(\gamma)})^2 \quad \ldots (4.22)$$

$$S_B^{(\gamma)}(M-m, N-n) \equiv (M-m)(N-n)(\bar{x}_{M-m,N-n}^{(\gamma)} - \bar{x}_{MN}^{(\gamma)})^2. \quad \ldots (4.23)$$

As to the relation between $S_R(M, N)$ and $S_W(m, n)$, a similar decomposition may be applied to the quantities $Z_{ij}^{(\gamma)} = x_{ij}^{(\gamma)} - \bar{x}_{iN}^{(\gamma)} - \bar{x}_{Mj}^{(\gamma)} + \bar{x}_{MN}^{(\gamma)}$ which yield us

$$\begin{aligned} S_W(M, N) = {} & S_W^{(\gamma)}(m, n) + S_W^{(\gamma)}(m, N-n) + \\ & + S_W^{(\gamma)}(M-m, m) + S_W^{(\gamma)}(M-m, N-n) + \\ & + S_{WR}^{(\gamma)}(m; n, N-n) + S_{WR}^{(\gamma)}(M-m; n, N-n) + \\ & + S_{WC}^{(\gamma)}(n; m, M-m) + S_{WC}^{(\gamma)}(N-n; m, M-m), \end{aligned} \quad \ldots (4.24)$$

where

$$S_{WR}^{(\gamma)}(m; n, N-n) \equiv \frac{n(N-n)}{N} \sum_{\alpha=1}^{M} \left(\bar{x}_{\alpha n}^{(\gamma)} - \bar{x}_{mn}^{(\gamma)} - \bar{x}_{\alpha,N-n}^{(\gamma)} + \bar{x}_{m,N-n}^{(\gamma)} \right)^2 \quad \ldots (4.25)$$

$$S_{WR}^{(\gamma)}(M-m; n, N-n) \equiv \frac{n(N-n)}{N} \sum_{\alpha=m+1}^{M} \left(\bar{x}_{\alpha n}^{(\gamma)} - \bar{x}_{M-m,n}^{(\gamma)} - \bar{x}_{\alpha,N-n}^{(\gamma)} + \bar{x}_{M-m,N-n}^{(\gamma)} \right)^2 \quad \ldots (4.26)$$

$$S_{WC}^{(\gamma)}(n; m, M-m) \equiv \frac{m(M-m)}{M} \sum_{\beta=1}^{n} \left(\bar{x}_{m\beta}^{(\gamma)} - \bar{x}_{m\beta}^{(\gamma)} - \bar{x}_{M-m,\beta}^{(\gamma)} + \bar{x}_{M-m,\beta}^{(\gamma)} \right)^2 \quad \ldots (4.27)$$

$$S_{WC}^{(\gamma)}(N-n; m, M-n) \equiv \frac{m(M-m)}{M} \sum_{\beta=n+1}^{N} \left(\bar{x}_{m\beta}^{(\gamma)} - \bar{x}_{m,N-n}^{(\gamma)} - \bar{x}_{M-m,\beta}^{(\gamma)} + \bar{x}_{M-m,N-n}^{(\gamma)} \right)^2 \quad \ldots (4.28)$$

Regarding $S_R(M, N)$ and $S_R^{(\gamma)}(m, n)$ their relation becomes more complicated so that we cannot separate out $S_R^{(\gamma)}(m, n)$ as an independent component from $S_R(M, N)$. Indeed we can write merely

$$\begin{aligned} S_R(M, N) = {} & S_R^{(\gamma)}(m; n, N-n) + S_R^{(\gamma)}(M-m; n, N-n) + \\ & + S_R^{(\gamma)}(m; n, N-n) + S_R^{(\gamma)}(M-m; n, N-n) \end{aligned} \quad \ldots (4.29)$$

where

$$S_R^{(\gamma)}(m;\, n,\, N-n) \equiv N\sum_{r=1}^{m}\left(\frac{n(x_{rn}^{(\gamma)}-\bar{x}_{mn}^{(\gamma)})}{N}+\frac{(N-n)(x_{r,N-n}^{(\gamma)}-\bar{x}_{m,N-n}^{(\gamma)})}{N}\right)^2 \quad \ldots \quad (4.30)$$

$$S_R^{(\gamma)}(M-m;\, n,\, N-n) \equiv N\sum_{u=m+1}^{M}\left(\frac{n(x_{un}^{(\gamma)}-\bar{x}_{M-m,n}^{(\gamma)})}{N}+\frac{(N-n)(x_{u,N-n}^{(\gamma)}-\bar{x}_{M-m,\,N-n}^{(\gamma)})}{N}\right)^2$$

$$\ldots \quad (4.31)$$

$$S_R^{(\gamma)}(m;n,\, N-n) \equiv Nm\left(\frac{n(\bar{x}_{mn}^{(\gamma)}-\bar{x}_{MN}^{(\gamma)})}{N}+\frac{(N-n)(\bar{x}_{m,N-n}^{(\gamma)}-\bar{x}_{MN}^{(\gamma)})}{N}\right)^2 \quad \ldots \quad (4.32)$$

$$S_R^{(\gamma)}(M-m\,;\, n,\, N-n) \equiv N(M-m)\left(\frac{n(\bar{x}_{M-m,n}^{(\gamma)}-\bar{x}_{MN}^{(\gamma)})}{N}+\frac{(N-n)(\bar{x}_{M-m,N-n}^{(\gamma)}-\bar{x}_{MN}^{(\gamma)})}{N}\right)^2 \quad (4.33)$$

and similarly for $S_Q(M, N)$.

Here let it be observed that (1) in each of the right-hand sides of (4.15), (4.24) and (4.29) all the summands are mutually independent among themselves and that (2) each summand is distributed according to the non-central chi-square distribution with its respective non-centrality parameter, where the degrees of freedom are equivalent to those valid in the case of null-hypothesis, while the non-centrality parameters are those corresponding to the finite population formulation and depending upon our permutation. For example the characteristic function of $S_T^{(\gamma)}(m,\, n)$ is given by

$$(1-2\sigma^2 it)^{-\frac{mn-1}{2}}\exp\left\{-\frac{\mu_T^{(\gamma)}(m,\,n)^2\,2\sigma^2 it}{1-2\sigma^2 it}\right\} \quad \ldots \quad (4.34)$$

where

$$\mu_T^{(\gamma)}(m,\,n)^2 \equiv \frac{1}{2\sigma^2}\sum_{r=1}^{m}\sum_{s=1}^{n}(\xi_{rs}^{(\gamma)}-\bar{\xi}^{(\gamma)})^2. \quad \ldots \quad (4.35)$$

The definitions of $\xi_{rs}^{(\gamma)}$, $\bar{\xi}_{mn}^{(\gamma)}$ are similar to those of $x_{rs}^{(\gamma)}$, $\bar{x}_{mn}^{(\gamma)}$ derived from $\{y_{ij}\}$. Indeed these are defined by operating our permutation on the set $\{\xi_{ij}\}$. As to the inferences to S_T (M,N) by means of $S_T(m,n)$ and those to $S_W(M,N)$ by means of $S_W(m,n)$ there are many common features among them and an ordinary analysis of variance. Our inferences concerning these two cases will depend upon the ratios of two independent non-central chi-squares. Let S_1 and S_2 be two independent non-central chi-squares with the degrees of freedom f_1 and f_2 and with the non-centrality parameters λ_1 and λ_2 respectively.

Then the probability density function of the distribution of S_1/S_2 will be given by

$$k(z; f_1, \lambda_1; f_2, \lambda_2)$$

$$= e^{-\frac{\lambda_1+\lambda_2}{2}} \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} \frac{\lambda_1^r \lambda_2^s}{r!\, s!} \frac{\Gamma\left(\frac{f_1+f_2}{2}+r+s\right)}{\Gamma\left(\frac{f_1}{2}+r\right)\Gamma\left(\frac{f_2}{2}+s\right)} \frac{z^{\frac{f_1}{2}+r-1}}{(1+z)^{\frac{f_1+f_2}{2}+r+s}}, \quad \ldots \quad (4.36)$$

We shall here observe immediately

Theorem 4.3: *The characteristic function of the statistic $S_T(M, N)/S_T(m, n)-1$ and that of $S_W(M, N)/S_W(m, n)-1$ are given by*

$$E\left[\exp\left\{it\left(\frac{S_T(M, N)}{S_T(m, n)}-1\right)\right\}\right]$$

$$= \frac{1}{{}_MP_m \times {}_NP_n} \sum_{\gamma} E\left[\exp\left\{it\left(\frac{S_T(M, N)}{S_T^{(\gamma)}(m, n)}-1\right)\right\}\right], \quad \ldots \quad (4.37)$$

$$E\left[\exp\left\{it\left(\frac{S_W(M, N)}{S_W(m, n)}-1\right)\right\}\right]$$

$$= \frac{1}{{}_MP_m \times {}_NP_n} \sum_{\gamma} E\left[\exp\left\{it\left(\frac{S_W(M, N)}{S_W^{(\gamma)}(m, n)}-1\right)\right\}\right], \quad \ldots \quad (4.38)$$

where

$$E\left[\exp\left\{it\left(\frac{S_T(M, N)}{S_T^{(\gamma)}(m, n)}-1\right)\right\}\right]$$

$$= \int_{-\infty}^{\infty} e^{itz}\, k\left(z; MN-1, \mu_T'^{(\gamma)}(M, N)^2-\mu_T'^{(\gamma)}(m, n)^2; mn-1, \mu_T'^{(\gamma)}(m, n)^2\right) dz \quad \ldots \quad (4.39)$$

$$E\left[\exp\left\{it\left(\frac{S_W(M, N)}{S_W^{(\gamma)}(m, n)}-1\right)\right\}\right]$$

$$= \int_{-\infty}^{\infty} e^{itz}\, k\Big(z; (M-1)(N-1); \mu_W'^{(\gamma)}(M, N)^2-\mu_W'^{(\gamma)}(m, n)^2;$$

$$(m-1)(n-1), \mu_W'^{(\gamma)}(m, n)^2\Big)\, dz \quad \ldots \quad (4.40)$$

with

$$\mu_x^{(\gamma)}(M, N)^2 = (2\sigma^2)^{-1}\sum_{i=1}^{M}\sum_{j=1}^{N}(\xi_{ij}-\bar{\xi}_{MN})^2 \qquad \ldots (4.41)$$

$$\mu_x^{(\gamma)}(m, n)^2 = (2\sigma^2)^{-1}\sum_{r=1}^{m}\sum_{s=1}^{n}(\xi_{rs}^{(\gamma)}-\bar{\xi}_{mn}^{(\gamma)})^2, \qquad \ldots (4.42)$$

$$\mu_W^{(\gamma)}(M, N)^2 = (2\sigma^2)^{-1}\sum_{i=1}^{M}\sum_{j=1}^{N}(\xi_{ij}-\bar{\xi}_{iN}-\bar{\xi}_{Mj}+\bar{\xi}_{MN})^2, \qquad \ldots (4.43)$$

$$\mu_W^{(\gamma)}(m, n)^2 = (2\sigma^2)^{-1}\sum_{r=1}^{m}\sum_{s=1}^{n}(\xi_{rs}^{(\gamma)}-\bar{\xi}_{rn}^{(\gamma)}-\bar{\xi}_{ms}^{(\gamma)}+\bar{\xi}_{mn}^{(\gamma)})^2. \qquad \ldots (4.44)$$

To proceed to $S_M(M; N)$ we have to prepare the following

Lemma 4.7: *Let* $\{x_i\}, \{y_i\}$ $(i = 1, 2, \ldots, k)$ *be a set of* $2k$ *mutually independent stochastic variables where each* x_i *and* y_i *are distributed in* $N(a_i, \sigma^2)$ *and in* $N(b_i, \sigma^2)$ *respectively. Let* p *and* q *be non-negative real numbers such that* $p^2+q^2 = 1$.

Let us now define

$$S_1 \equiv \sum_{i=1}^{k}(x_i-\bar{x}_\cdot)^2, \qquad \ldots (4.45)$$

$$S_2 \equiv \sum_{i=1}^{k}(y_i-\bar{y}_\cdot)^2, \qquad \ldots (4.46)$$

$$S_3 \equiv \sum_{i=1}^{k}\left\{p(x_i-\bar{x}_\cdot)+q(y_i-\bar{y}_\cdot)\right\}^2. \qquad \ldots (4.47)$$

Then the characteristic function of the joint distribution of S_1, S_2, *and* S_3 *is given by*

$$E\left[\exp\{i(t_1S_1+t_2S_2+t_3S_3)\}\right] = \frac{1}{\Delta_1^{(k-1)/2}}\exp\left\{-\frac{\Delta}{\Delta_1}\right\}, \qquad \ldots (4.48)$$

where

$$\Delta_1 \equiv \{1-2\sigma^2i(t_1+p^2t_3)\}\{1-2\sigma^2i(t_2+q^2t_3)\}+4p^2q^2\sigma^2t_3^2, \qquad \ldots (4.49)$$

$$\Delta = \frac{1}{2\sigma^2}\sum_{j=1}^{k}{}' \begin{vmatrix} 1-2\sigma^2i(t_1+p^2t_3) & -2\sigma^2ipqt_3 & -(a_j-\bar{a}) \\ -2\sigma^2ipqt_3 & 1-2\sigma^2i(t_2+q^2t_3) & -(b_j-\bar{b}) \\ -(a_j-\bar{a}) & -(b_j-\bar{b}) & (a_j-\bar{a})^2+(b_j-\bar{b})^2 \end{vmatrix} \qquad \ldots (4.50)$$

where $\bar{a}$ and $\bar{b}$ are the arithmetic means of $\{a_j\}$ and $\{b_j\}$ respectively.

The proof can be obtained by a direct calculation of the characteristic function. This lemma and the following corollary seem to us to be very important in successive designs of experiments.

Corollary 4.1: *The characteristic function of the joint distribution function of S_1 and S_2 is given by*

$$E[\exp\{i(S_1t_1+S_2t_2)\}]$$

$$= (1-2\sigma^2it_1-2\sigma^2it_2-4\sigma^4q^2t_1t_2)^{-\frac{k-1}{2}} \exp\left\{\frac{2\sigma^2i\sigma_a^2t_1 + 2\sigma^2it_2\sum_{j=1}^{k}(p\delta a_j+q\delta b_j)^2}{1-2\sigma^2it_1-2\sigma^2it_2-4\sigma^4q^2t_1t_2}\right\} \quad \ldots (4.51)$$

where
$$\sigma_a^2 = (2\sigma^2)^{-1}\sum_{j=1}^{k}(a_j-\bar{a})^2, \quad \ldots (4.52)$$

$$\delta a_j = (2^{\frac{1}{2}}\sigma)^{-1}(a_j-\bar{a}), \quad \ldots (4.53)$$

$$\delta b_j = (2^{\frac{1}{2}}\sigma)^{-1}(b_j-\bar{b}). \quad \ldots (4.54)$$

The essential point to note here is that there are some characteristic features of the exact sampling theory for finite populations which make it necessary to use somewhat coherent statistics as in (4.51).

5. SUMMARY

We have established a certain set of exact sampling distributions under certain assumptions which in our opinion will give a more correct picture of real situations. Here we have been content with giving theoretical considerations which in combination with some abbreviated numerical calculations will yield us some useful results applicable to practical cases. Specially, the average taken over all possible permutations will make it clear how and under what conditions ordinary uses of t- and F-distributions may be justified. It is also to be noted that our fundamental idea is to appeal to subsample and two-sample formulations, and that, on the contrary, assumptions of normality are rather artificial.

PART V. OPERATIONAL FORMULATION OF PROBLEMS OF STATISTICAL INFERENCES IN SAMPLE SURVEYS

1. Introductory

Different types of errors occur in the case of large-scale sample surveys. Deming (1944, 1950) gave a detailed listing and description of the different types of errors which should be taken into consideration both in designing and analysing sample surveys. Recently Hansen, Hurwitz, Marks, and Mauldin (1951) discussed response errors which are important factors influencing accuracies of surveys. In this Part V, we shall consider the main sources of error in large-scale sample surveys. For this purpose the classification of different types of error into three types made by Mahalanobis (1944, 1946) is specially suited at least for general considerations. Mahalanobis (1944) mentions the following three types of error: (1) sampling fluctuation, (2) observational error and (3) gross inaccuracies, where (1) and (2) may be presumed to follow probabilistic schemes exactly or at least approximately and hence to be amenable to statistical treatment which, however, does not apply to errors of type (3).

"...In actual practice, however, it is difficult to separate these latter two groups, and it is necessary to pool together the second and the third types under one common head which may be called recording mistakes arising from the human factor" (Mahalanobis, 1944). He also "revealed the great importance of controlling and eliminating as far as possible the mistakes which occurred at the stage of the field survey" (Mahalanobis, 1944, p.409). "One way of doing this would be to organise the sample survey in the form of two or more interpenetrating subsamples" (Mahalanobis, 1944, p. 381). In fact, in spite of many controversies over the usefulness of interpenetrating samples, their diagnostic power can only be duly recognised after one has taken into consideration all the three types of errors and not merely (1) and (2). If we consider—and we must consider—situations in which errors of all these three types should be duly treated, our usual formulation will be found to be too narrow and our usual theories of statistical inference will not be sufficient to cover all the problems. How to consider this type of error will be discussed in § 2, whereas § 3 will be devoted to the problem of statistical inference and controls for guarding against this type of error.

2.* State, operator and scheme

We now consider a case where the third type of errors viz., gross inaccuracies should be taken into consideration in order that we should be able to give a more realistic formulation of the problem of sampling design. If we confine ourselves to the first type of error, viz. sampling fluctuations, then each element of our population has a definite value which may be a scalar but may sometimes be vectorial. If we consider both the first and the second types of error, then to each element of our

population there should correspond a stochastic variable, so far as the second type of error may be amenable to probabilistic approach.

The third type of error, on the other hand, is actually a very broad type which includes all errors belonging to neither of these two, and naturally involves various kinds of errors; inaccuracies and falsehood in statements and recording, tricks and so on. Thus the sound approach to make a step forward is not to give a too broad (and obscure) formulation aiming to cover all types of errors that could be imagined, but rather to choose, corresponding to each stage of theoretical and practical development, some restricted domain we should take and we could take into consideration to make an adequate and effective improvement of our designs and analysis.

For this purpose, we propose to introduce here the notions of state, operator and scheme.

In almost all sampling surveys, we shall be able to introduce the notions of respondents, investigators and scheme of surveys. To each respondent there corresponds an objective existence which we call a state and which can be recognised to be an existence independent of our surveys. Responses obtained (if possible) from a respondent through some procedure by investigators may result in vector values, because they will give information on each question item of the schedule, giving answers of the nature of either attributes or variables. There will, however, occur problems of non-response and also of possible interference between investigators and respondents. An abstract idea of investigators which may include any other type of questionnaires such as mail or telephone or interview survey should be more relevantly represented by the notion of operators. Thus there is a state ξ of objective existence and to each state an operator α will be applied so that it may give us variables under a certain scheme S. The domain of α in which, under the scheme S, we may be able to observe some variable corresponding to a state ξ does not necessarily cover the whole of possible states. If it be defined for a certain set of state ξ, operation α and scheme S, then we shall denote the variable by $S(\alpha, \xi)$.

An abstract idea of $S(\alpha, \xi)$ will be so broad as to be associated with or to be subject to falsehood, deliberations and even strategies, for which there could not be any objective approach, unless we restrict ourselves to certain realms of S, α and ξ. Some kinds of falsehood, deliberations and strategies may have naturally various sources. It is impossible to suggest *a priori* a method by which we should be able to measure or to control all sorts of these errors. Nevertheless we think some or all of the following approaches would be particularly useful in dealing with this third type of error.

(a) *Restriction within a certain domain of types of error:* In taking into consideration all sorts of errors at the same time, our attitude should be gradually progressive. At the first stage we may consider the types of error for which the following procedures (b)—(d) are comparatively easier than others.

(b) *Application of randomisation:* The principal object of randomisation is to introduce a probabilistic scheme so that valid statistical inferences can be made. One cannot help appealing to this principle in order to establish any useful results applicable to some class of errors belonging to the third type. Indeed the scheme which Hansen, Hurwitz, Marks and Mauldon (1951) used in discussing response errors in surveys and the scheme which Sukhatme and Seth (1952) used in studying non-sampling errors in surveys have this as a common feature that each of them appeals to the principle of randomisation (although the models are different).

(c) *Application of the principle of 'transformation':* In actual cases it frequently occurs that we are really concerned not with the variables $A(x, \xi)$ themselves but some differences among these. In these cases we may be able to eliminate certain unmanagable factors, thanks to the operations of differences. For instance, suppose we are concerned with the states at two different times, say, $A(x, \xi_{t_1})$ and $A(x, \xi_{t_2})$. In spite of the fact that for some set of A, x and ξ, $A(x, \xi_{t_1})$ and $A(x, \xi_{t_2})$ cannot be obtained, it may be possible in such cases to obtain the variables $A(x, \xi_{t_1} - \xi_{t_2})$ defined for every set of A, x and ξ.

This principle belongs to the realm of logic and may be regarded as a prototype of analysis of variance, but it also has an intimate relation with the following principle α. Generally speaking there may sometimes be another function ϕ such that $A(x, \phi(\xi_{t_1}, \xi_{t_2}))$ can be defined throughout the whole domain or at least in a broader domain. Moreover there are possibilities of making use of a couple or a team of operators by which $A((x, \beta), \xi)$ may be defined throughout a broader domain. The combination (x, β) means the co-operation of two types of investigators x and β where x is a proper investigator who wants to obtain necessary data from respondents while β is an auxiliary person who is not well-trained as an investigator but who has intimate knowledge of the respondents and will serve to make respondents confident enough to answer correctly to α. Similarly the questionnaire may sometimes contain some set B of questions which has a similar effect on the respondents as this auxiliary person β. This can be expressed by the symbol (A, B) (x, ξ). Instead of starting direct questionning about domestic economics of households, it is often more effective to speak about general topics which lead them naturally to answer the desired questions.

Both before and in course of the sequence of surveys, there sometimes arises the need of some enlightenment and 'education' for respondents by which we can expect to enlarge the domain of α for which $A(x, \xi)$ are defined. This domain should be actually denoted by $A(x, L\xi)$ where L stands for enlightenment to the respondents. These principles or procedures seem to belong to some sort of expert techniques, but it will not only be possible but also necessary to give theoretical considerations and also to analyse real data. By suitable formulation the efficiencies of such transformations and the costs for executing them should be discussed in a manner similar to the discussion in parts I-III of costs of surveys and variances of estimates where the latter was concerned with the first type of error only.

(d) *Application of operational view-points:* After all our efforts of making use of (a), (b) and (c) there may still remain certain cases in which the domain of ξ for which $A(x, \xi)$ is defined for every x and A is not coincident with the whole space of ξ. For example, let us consider a sampling survey on living costs which requires of each sample household to write in their diary the daily expenditures on and the quantities consumed of each item of food in suitable units. All households cannot be expected to agree with writing this sort of diary for several months. Thus, broadly speaking, there arises the problem of non-response so far as we adhere to such uses of diaries. The social and economic circumstances which cause this type of non-response have some relationship with, say, the living standard and we cannot deny that the biases may not be negligible, if our survey should be confined to the diary records. In such a situation our attitude (to be justified from operational point of view) should be to divide the aims of our surveys into two types; the first type will be concerned with the diary reports, while the second type will aim to investigate the circumstances under which some households cannot or will not respond and also to study the relationship of this with their ways of living. So far as it is accompanied by the latter surveys, the first part of the survey is useful for a certain class of operational problems, for example, for wage agreements between the entrepreneur and the labour union. The estimates of living costs obtained from a sample survey of the first part with auxiliary data of the second part may be recognised to have operational value. This was the method adopted by Kitagawa and Fujita (1951) for a sampling survey of living costs of coal miners in 1948.

3. Operational formulation of statistical inference

In dealing with the third type of errors our general principles illustrated in § 2 should be introduced both in designing the sample survey and in the analysis of data. Regarding the latter it seems necessary (and adequate) to appeal to an operational formulation of the problem of statistical inference. An operational formulation means an elaboration of our statistical decisions so as to deal with some sort of previous knowledge and/or problems of prognosis in a more comprehensive way as we have done in two previous papers (Kitagawa, 1953a, 1953b).

In a previous paper (Kitagawa, 1953a) it was pointed out that certain operational formulations of previous knowledge would sometimes be useful, because of the fact that previous knowledge may be derived from various sources, not necessarily from sampling, or from designed experiments to which probabilistic approaches are possible, but also sometimes from current literature and from obscure sources. Such situations will surely occur when we shall take into consideration the third type of errors discussed in § 1 and § 2. Since there remains usually a lack of objective knowledge enough to appeal to averaging process, however, we may endeavour to apply certain principles enunciated in § 2, and also some statistical procedures which are both operational and objective. Some elaboration of statistical inferences such

as pooling of data (i.e. estimation after preliminary testing of hypothesis) as well as a more general procedure of successive process of statistical inferences discussed in our papers (Kitagawa, 1950a, 1953b) will be found to be specially useful in such situations. In the next section, we shall give an example of applications in connection with interpenetrating samples. Detailed designs of experiments developed in agricultural experimentation may be also helpful in analysing some types of errors in conjunction with psychometric, sociometric and econometric studies on sources of such errors to which sampling survey and design of experiments are now being applied by various research workers.

Successive procedures seem to be extremely useful in dealing with the broad class of errors belonging to the third type. Indeed a follow-up procedure dealing with the non-response problem adopted by Hansen and Hurwitz (1943) and an interesting method of finding sample size due to Birnbaum and Sirken (1950) when non-response is present, appeal to double sampling or to two sample procedure each of which is a special case of successive procedure.

4. Uses of interpenetrating samples

The method of interpenetrating net-works of samples in sample surveys uses a type of design in which the sample units are arranged in two or more independent sets of samples each supplying an independent estimate of the variate under study. This brilliant idea is originally due to Mahalanobis (1940,1946) and has been recommended by the United Nations Sub-commission on Statistical Sampling. Its purpose is "to provide statistical controls for detecting and guarding against such recording mistakes" (Mahalanobis, 1944) where "recording mistakes" in his terminology comprise the second and the third types of error in §2.

Indeed there have been several authors on both the practical and the theoretical sides who do not recognise this method as a useful statistical control. The present author is unable to discuss the effects of the application of the method, examining in detail the data available. It is, however, intended to show that some of the theoretical objections against the method are not quite appropriate. This is due to not taking the whole role of inter-penetrating samples in consideration fully. Thus although the calculations of the efficiency per unit cost given by Mokashi (1949) shows one of essential loss of efficiency due to the interpenetrating samples, his considerations do not seem to cover the different roles and the functions of these samples. There are various aspects of interpenetrating samples which sometimes lead us to some sort of confusion of notions. To make our views clear in distinction with some authors, let us quote here the description of Ghosh (1949) who discussed the functions of interpenetrating samples in some details and who was not convinced of the actual uses of these samples. According to Ghosh (1949) the basic roles of these samples are enunciated as follows. "Basically the method consists in having two

(or more) samples from the same population so designed that (i) each of these samples will furnish a valid estimate (with its confidence limits) of a common population characteristic; (ii) it will be possible to make statistically valid comparisons between the different samples and (iii) in case the samples (as observed) are not significantly different from one another (as is expected if the samples are enumerated correctly) it will be possible to construct a joint estimate (with confidence limits) of the population characteristic by pooling the information from all the samples". (Ghosh, 1949, pp.108-109). Furthermore he points out (quite appropriately) the following: "It may be noted that the stronger the (positive) correlation between the samples the greater will be the sensitiveness (or discriminating power) of the comparison between the samples, mentioned under (ii) above and the lower will be the efficiency (or precision) of the joint estimate from all the samples, mentioned under (iii)." (Ghosh 1949, p.109). It is thus clear that the uses of interpenetrating samples can be judged not only by the efficiency of the joint estimate but also by the discriminating power of the comparison between the samples, and also that once we enter into the latter, what we are really concerned with is not only the null hypothesis but also the alternative hypothesis, that is, the possibility of different populations must be taken into consideration. Consequently the description of Ghosh seems to us somewhat unsatisfactory.

The discriminating power of interpenetrating samples should be emphasised. The calculation of efficiencies given by Mokashi seems therefore to be inadequate (1949) who takes into consideration the role (iii) merely. "If comparisons between the different investigators by means of interpenetrating samples have been arranged, the comparative results must be available as quickly as possible, in order that effective action may be taken if discrepancies are discovered". (F. Yates, 1949, p.107). Yates also indicates the following important points:'Interpenetrating samples are of value if the survey or census has to be carried out by successive stages. This is frequently necessary when preliminary results are required quickly". (Yates, 1949, p. 44.).

In summing up, our conclusion is that the true merits of interpenetrating samples can only be suitably discussed from the point of view of successive process of statistical inferences and controls in which we must and we shall discuss some elaborated inferences and also effects of statistical controls. In what follows we shall state briefly what we want to mean by this assertion. For the sake of simplicity emphasising the essential points of our views, we shall assume in what follows an infinite normal population, although in real situations we are concerned with finite populations. This simplification is justified by the theory which we have already shown in Part IV as a means of focussing our essential points. The assumption of the normality of our parent population is also a matter of mathematical technique. On the other hand, certain characteristic features of interpenetrating samples should be carefully formulated and introduced in the Assumptions.

5. Examples of statistical inference connected with interpenetrating samples

Our Assumptions will be as follows:

Assumption 1. Let (x, y) be distributed in a bivariate normal distribution Π with the density function

$$(2\pi\sigma_1\sigma_2)^{-1}(1-\rho^2)^{-\frac{1}{2}}\exp\{-Q/2(1-\rho^2)\}, \qquad \ldots \quad (5.01)$$

where

$$Q \equiv \frac{(x-a)^2}{\sigma_1^2} - 2\rho\frac{(x-a)(y-b)}{\sigma_1\sigma_2} + \frac{(y-b)^2}{\sigma_2^2}. \qquad \ldots \quad (5.02)$$

Assumption 2. Let $O_n : \{(x_i, y_i)\ i = 1, 2, \ldots, n\}$ be a random sample of size n from the population Π.

Assumption 3. The total cost for observing the sample O_n is equal to $c_1 f_1(n_1) + c_2 f_2(n_2)$ where c_1 and c_2 are costs per single observation of x and y respectively.

The correlation ρ represents a certain degree of intraclass correlation between the members of a pair, x_i and y_i, for the same i. If we adopt the functional form of the cost of journeys due to Mokashi (1949), then $f_1(n) = f_2(n) = kn^{\frac{1}{2}}$, k being a certain constant.

In the discussion in §4 the statistical problem connected with interpenetrating samples have been shown to be one of estimation after testing of hypotheses, that is, of the so-called pooling of data. There may be various procedures which resemble real operations actually adopted in practical circumstances. Whatever they may be, the common features are to test statistical hypotheses as a first step. In our formulation we may consider, for example, three different null hypotheses

$$(1^0)\ H_1 : a = b,\ \sigma_1^2 = \sigma_2^2;\ (2^0)\ H_2 : a = b;\ (3^0)\ H_3 : \sigma_1^2 = \sigma_2^2,$$

each of which is composite. The testing of these null hypotheses is the first step. The second step will depend on the results of these. If the tests do not show significant differences the final estimates should be the pooled ones. If, however, differences are significant there may be various procedures possible. For instance, we may consider the following three procedures:

Procedure A: If there are certain a priori reasons for preferring the x-observations to the y-observations, our procedure will be to adopt the set of observations $(x_1, x_2, \ldots, x_n)$, the set $(y_1, y_2, \ldots, y_n)$ being completely ignored.

Procedure B: There is a possibility of appealing to a successive process of pooling such as was discussed in Part III of the present paper.

Procedure C: This process is to draw from the population another random sample of a suitable size which should be regarded with more confidence than either of the samples; $(x_1, x_2, \ldots, x_n)$ or $(y_1, y_2, \ldots, y_n)$.

For each of these procedures we may give a formulation according to which our statistical procedure will be defined. Parts I and III of the present paper may be of some use after certain generalisations. These generalisations will be in two respects, firstly concerned with the intercorrelation ρ, while the second one with cost considerations.

Let us now confine ourselves to one special case where $\sigma_1 = \sigma_2 = \sigma$ (unknown) will be assumed and where the statistical procedure will be of the type (A). This case will arise when there is no difference between the variances of two operators but the presence of some bias of the less trained one which gives us y-values is expected.

The statistical procedure will be as follows:

(i) Let the statistic t be defined by

$$t = \frac{\sqrt{n}\, d}{s_d}, \qquad \ldots \quad (5.03)$$

where we have put

$$d = n^{-1} \sum_{i=1}^{n} d_i = n^{-1} \sum_{i=1}^{n} (x_i - y_i), \qquad \ldots \quad (5.04)$$

$$s_d = \{(n-1)^{-1} \sum_{i=1}^{n} (d_i - d)^2\}^{\frac{1}{2}}. \qquad \ldots \quad (5.05)$$

(ii) The estimate $\bar{\bar{x}}$ of α will be defined in the following manner:

(a) If $| t | < t_{n-1}(\alpha)$, then

$$\bar{\bar{x}} = 2^{-1}(\bar{x} + \bar{y}) \qquad \ldots \quad (5.06)$$

and (b) $\bar{\bar{x}} = \bar{x}$, if otherwise, $\qquad \ldots \quad (5.07)$

where $t_{n-1}(\alpha)$ denotes the value of t with $n-1$ degrees of freedom for a significance level α, $0 < \alpha < 1$.

Theorem 5.1: *Let z be any assigned real number. The distribution of $\bar{\bar{x}}$ is given by*

$$Pr.\{\bar{\bar{x}} < z\} = Pr.\{\bar{\bar{x}} < z, | t | < t_{n-1}(\alpha)\} + Pr.\{\bar{\bar{x}} < z, | t | > t_{n-1}(\alpha)\} \ldots \quad (5.08)$$

where we have

$$Pr.\{\bar{\bar{x}} < z, | t | < t_{n-1}(\alpha)\} = Pr.\left\{\frac{x+y}{2} < z\right\} Pr.\{| t | < t_{n-1}(\alpha)\} \qquad \ldots \quad (5.09)$$

with

$$Pr.\left\{\frac{x+y}{2} < z\right\} = \frac{(2n)^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}\sigma(1+\rho)^{\frac{1}{2}}} \int_{-\infty}^{z} \exp\left\{-\frac{n(u-2^{-1}(a+b))^2}{\sigma^2(1+\rho)}\right\} du \qquad \ldots (5.091)$$

and

$$Pr.\{|t| < t_{n-1}(\alpha)\} = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{1}{2}\right)} \exp\left\{-\frac{n(a-b)^2}{2\sigma^2(1-\rho)}\right\} \times \qquad \ldots (5.092)$$

$$\times \int_{0}^{t_{n-1}^2(\alpha)/(n-1)} \frac{f^{-\frac{1}{2}}}{(1+f)^{\frac{n}{2}}} {}_1F_1\left(\frac{n}{2}, \frac{f}{2(1+f)}, \frac{n(a-b)^2}{2\sigma^2(1-\rho)}\right) df, \qquad \ldots (5.10)$$

$$Pr.\{\bar{x} < z, |t| \geqslant t_{n-1}(\alpha)\} = \int_{-\infty}^{z} \frac{n^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}\sigma} \exp\left\{-\frac{n(\bar{x}-a)^2}{2\sigma^2}\right\} d\bar{x} \times$$

$$\times \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{1}{2}\right)} \exp\left\{-\frac{n[a-b+(1-\rho)(x-a)]^2}{2\sigma^2(1-\rho^2)}\right\} \times$$

$$\times \int_{t_{n-1}^2(\alpha)/(n-1)}^{\infty} \frac{f^{-\frac{1}{2}}}{(1+f)^{\frac{n}{2}}} {}_1F_1\left(\frac{n}{2}, \frac{f}{2(1+f)}, \frac{n[(a-b)+(1-\rho)(x-a)]^2}{2\sigma^2(1-\rho^2)}\right) df.$$

Here ${}_1F_1$ *is the confluent hypergeometric function given by*

$${}_1F_1\left(\frac{n}{2}, \frac{1}{2}, x\right) = \sum_{r=0}^{\infty} \frac{\Gamma\left(\frac{n}{2}+r\right)\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(r+\frac{1}{2}\right)\Gamma(r+1)} x^r. \qquad \ldots (5.11)$$

Proof: Let us consider the transformation of (x_i, y_i) into (u_i, v_i):

$$u_i = 2^{-\frac{1}{2}}(x_i - y_i), \quad v_i = 2^{-\frac{1}{2}}(x_i + y_i) \qquad \ldots (5.12)$$

for $i = 1, 2, \ldots, n$.

Then we shall have

$$q_i = (2\sigma^2(1-\rho^2))^{-1}((x_i-a)^2 - 2\rho(x_i-a)(y_i-b) + (y_i-b)^2)$$
$$= (2\sigma^2(1-\rho))^{-1}(u_i - 2^{-\frac{1}{2}}(a-b))^2 + (2\sigma^2(1+\rho))^{-1}(v_i - 2^{-\frac{1}{2}}(a+b))^2. \qquad \ldots (5.13)$$

Consequently the elementary probability associated with the sample of size n i.e.

$$\prod_{i=1}^{n} \frac{1}{2\pi\sigma^2(1-\rho^2)^{\frac{1}{2}}} \exp\left\{-q_i\right\} dx_i dy_i \qquad \ldots (5.14)$$

will be transformed into

$$\frac{n^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}\sigma(1+\rho)^{\frac{1}{2}}} \exp\left\{-\frac{n(\bar{v}-2^{-\frac{1}{2}}(a+b))^2}{2(1+\rho)\sigma^2}\right\} d\bar{v} \cdot \frac{n^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}\sigma(1-\rho)^{\frac{1}{2}}} \exp\left\{-\frac{n(\bar{u}-2^{-\frac{1}{2}}(a-b))^2}{2(1-\rho)\sigma^2}\right\} d\bar{u} \times$$

$$\times\ G\left(\frac{1}{2(1-\rho)\sigma^2},\ \frac{n-1}{2},\ S^2\right) dS^2 \qquad \ldots (5.14.1)$$

where we have put

$$\bar{v} = n^{-1}\sum_{i=1}^{n} v_i, \quad \bar{u} = n^{-1}\sum_{i=1}^{n} u_i, \qquad \ldots (5.15)$$

$$S^2 = \sum_{i=1}^{n} (u_i - \bar{u})^2, \qquad \ldots (5.16)$$

and the function $G(x, p, x)$ denotes the gamma distribution

$$G(x, p, x) = \frac{\alpha^p}{\Gamma(p)} e^{-\alpha x} x^{p-1}. \qquad \ldots (5.17)$$

We now calculate the first and the second term of the right hand side (5.08) separately.

Regarding the first term it will suffice to note that

$$2^{-\frac{1}{2}}(\bar{x}+\bar{y}) = 2^{-\frac{1}{2}}\bar{v} \qquad \ldots (5.18)$$

$$t^2 = (n-1)n\bar{u}^2 S^{-2}, \qquad \ldots (5.19)$$

and to make use of the non-central t-distribution.

Regarding the second term, we shall first rewrite

$$Q \equiv \frac{n}{2\sigma^2(1-\rho)}\left(\bar{u} - \frac{a-b}{2^{\frac{1}{2}}}\right)^2 + \frac{n}{2\sigma^2(1+\rho)}\left(\bar{v} - \frac{a+b}{2^{\frac{1}{2}}}\right)^2 \qquad \ldots (5.20)$$

$$= \frac{n}{2\sigma^2}(\bar{x}-a)^2 + \frac{n}{\sigma^2(1-\rho^2)}\left(\bar{u} - \frac{(a-b)+(1-\rho)(\bar{x}-a)}{2^{\frac{1}{2}}}\right)^2$$

which gives us

$$Pr.\{\bar{x}<z, |t| \geqslant t_{n-1}(\alpha)\} \qquad \ldots \quad (5.21)$$

$$= \int\int\int_{D_1} \frac{n^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}\sigma} \exp\left\{-\frac{n(x-a)^2}{2\sigma^2}\right\} dx \,.\, G\left(\frac{1}{2(1-\rho^2)\sigma^2},\ \frac{n-1}{2},\ S^2\right) dS^2 \times$$

$$\times \frac{(2n)^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}\sigma(1-\rho^2)^{\frac{1}{2}}} \exp\left\{-\frac{2n}{2\sigma^2(1-\rho^2)}\left(\bar{u}-\frac{(a-b)+(1-\rho)(x-a)}{2^{\frac{1}{2}}}\right)^2\right\} d\bar{u}$$

where the domain of integration is defined by

$$D_1 : \begin{cases} x < z, \\ (n-1)n\bar{u}^2S^{-2} \geqslant t^2_{n-1}(\alpha). \end{cases} \qquad \ldots \quad (5.22)$$

But the application of the non-central t-distribution gives us the relation (5.10) where we have introduced $f = n\bar{U}^2S^{-2}$, $\bar{U}^2$ being equal to $2^{\frac{1}{2}}u$.

We can readily observe from Theorem 5.1 the following:

Theorem 5.2: *The k-th moment of the statistic $\bar{x}$ is given by*

$$E\{\bar{x}^k\} = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \left(\frac{a+b}{2} + \left(\frac{1+\rho}{2n}\right)^{\frac{1}{2}}\sigma g\right)^k \exp\left\{-\frac{g^2}{2}\right\} dg \,.\, \{1-H_{n-1}(z, \delta^2(1-\rho)^{-1})\} +$$

$$+(2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} (a+n^{-\frac{1}{2}}\sigma h)^k \exp\left\{-\frac{h^2}{2}\right\} . H_{n-1}(\alpha^*, (\delta+h)^2(1-\rho^2)^{-1}) dh, \quad \ldots \quad (5.23)$$

where we have put

$$H_{n-1}(\alpha, \xi^2) \equiv \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{1}{2}\right)} e^{-\frac{n\xi^2}{2}} \int_{(n-1)^{-\frac{1}{2}}t_{n-1}(\alpha)}^{\infty} \frac{f^{-\frac{1}{2}}}{(1+f)^{\frac{n}{2}}} \, {}_1F_1\left(\frac{n}{2}, \frac{f}{2(1+f)}, \frac{n\xi^2}{2}\right) df$$

$$\ldots \quad (5.24)$$

$$\delta \equiv n^{\frac{1}{2}}(a-b)\sigma^{-1} \qquad \ldots \quad (4.25)$$

while α^* is defined by the central t-distribution with $n-1$ degrees of freedom such that α^* is the point whose significance value is equal $2^{\frac{1}{2}}t_{n-1}(\alpha)$, that is to say

$$t_{n-1}(\alpha^*) = 2^{\frac{1}{2}}t_{n-1}(\alpha). \qquad \ldots \quad (5.26)$$

Corrolary 5.1: *The mean and the variance of the statistic* $\bar{\bar{x}}$ *are given by*

$$E(\bar{\bar{x}}) = \frac{a+b}{2}(1-H_{n-1}) + a\overline{H^*_{n-1}} + \frac{\sigma}{n}\overline{hH^*_{n-1}} \qquad \ldots (5.27)$$

$$\sigma^2(\bar{\bar{x}}) = \frac{1+\rho}{2n}\sigma^2(1-H_{n-1}) + \frac{\sigma^2}{n}(\overline{h^2H^*_{n-1}} - \overline{hH^*_{n-1}})^2 -$$

$$-2\frac{\sigma}{n}\left\{\frac{a+b}{2}(1-H_{n-1}) + \overline{aH^*_{n-1}}\right\}\overline{hH^*_{n-1}} + \left(\frac{a+b}{2}\right)^2 H_{n-1}(1-H_{n-1})$$

$$-2a\frac{a+b}{2}\overline{H^*_{n-1}}(1-H_{n-1}) + a^2\overline{H^*_{n-1}}(1-\overline{H^*_{n-1}}). \qquad \ldots (5.28)$$

where we have put for the sake of brevity

$$H_{n-1} = H_{n-1}(x, \delta^2(1-\rho)^{-1}) \qquad \ldots (5.29)$$

$$h^\nu H^*_{n-1} \equiv (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} h^\nu \exp\left\{-\frac{h^2}{2}\right\} H_{n-1}\left(x^*, \frac{(\delta+h)^2}{1-\rho^2}\right) dh, \qquad \ldots (5.30)$$

for $\nu = 0, 1, 2$.

It remains to calculate distribution function of $\bar{\bar{x}}$ and its moments numerically. Nevertheless the Theorem and its corollary enable us to make the following observations:

(1) Several authors e.g. Ghosh (1949) points out that the stronger the (positive) correlation ρ between x and y the greater the sensitiveness (or discriminating power) of the comparison between the two samples. Our results bear out these in quantitative terms, for it may be seen from the terms of the non-central t-distribution that the parameters of non-centrality are $\delta^2(1-\rho)^{-1}$ and $(\delta+h)^2/(1-\rho^2)$, where $\delta = n^{\frac{1}{2}}(b-a)\sigma^{-1}$.

(2) Further, several authors e.g. Mokashi (1949), Ghosh (1949) point out also at the same time that the stronger the (positive) correlation ρ between x and y the lower is the efficiency of the joint estimate. Indeed the former showed that the variance of the joint estimate will be $(1+\rho)$ times that of an estimate based on an independent sample of size $2n$. This is not accurate unless we have definitely assigned our statistical procedure. On the contrary, we assert that at least under our formulation the circumstances are not so simple as supposed. One must observe the real situations from (5.28). Indeed there is one term which takes the form $(1+\rho)\sigma^2 k$, that is, the first term on the right-hand side of (5.28), but it is merely one constituent term of the contributions due to ρ.

(3) Unless the population means a and b are coincident with each other the more comparison between the variances is not adequate, because the joint estimate

is not an unbiassed estimate of a. It is evident that the mean bias of our estimate $\bar{x}$ is always less than that of the joint estimate without any preliminary testing of hypotheses.

It is worth while to notice the meaning of the correlation ρ. Indeed in view of general considerations on effects of biases developed in Cochran (1953, chapter 13) and specially his formulations on interpenetrating subsamples, the formulation due to Hansen, Hurwitz, Marks and Mauldin (1951) and also of our formulation developed in Chapter IV of the present paper, the existence and the magnitude of the correlation ρ reflect the characteristic features of both finite population and state, operator and scheme formulation, of which more detailed description should be required. We have only pointed out the main features rather in an abstract formulation. The generalised model of analysis of variance discussed in our paper (Kitagawa, 1953) has found its adequacy here.

The correlation ρ derives sometimes from these generalised models. Thus normal regression theory in the presence of interclass correlation discussed by Halperin (1951) and in a different way from the point of view of successive process of statistical inferences in our paper (Kitagawa, 1953a) will be useful in developing the detailed discussion of interpenetrating samples.

The numerical calculations involved and the detailed description of conditions under which our present formulation will give an approximative picture of real situations will be postponed to another occassion.

PART VI. THE EFFECTS OF STRATIFICATION

1. Introductory to the problem of stratification

The problem of constructing a system of strata is different from that of optimum allocation when a certain stratification is given. In this Part we shall consider various problems arising in practice restricting ourselves, however, as a first approach, to non-sequential theories. Our problems will be discussed by introducing a certain mathematical model relating to the objectives and by comparing the effects of various stratifications with reference to these models. It has been a tradition of experts to rely largely on intuition and experience. We shall discuss some of these with the purpose of investigating their merits under the theoretical formulations.

2. Effective methods of preliminary stratifications — Mahalanobis method

In spite of the elegant theory of optimum allocation in stratified random sampling, there is one serious difficulty, especially at the beginning of a sequence of sample surveys, which may prevent any attempt to adopt the optimum allocation

given in current literature. This is due to the lack of previous knowledge concerning the population which would be necessary for adopting any system of stratification, in particular the knowledge of the within stratum variances. In such a situation there is a method of stratification advocated by Mahalanobis. Broadly speaking, he suggests that when the number of strata is assigned a practical method of possible stratification is to stratify the whole population Π into a set of k strata $\{\Pi_i\}$ $(i = 1, 2, \ldots, k)$ such that the stratum sums are expected to be equal at least approximately, that is, in the notation of Part I

$$N_1\bar{x}_1 = N_2\bar{x}_2 = \ldots = N_k\bar{x}_k. \qquad \ldots \quad (2.01)$$

Since we have no accurate knowledge of the stratum sums, it is clear that we must make use of rough estimates or some values highly correlated with these stratum sums.

One justification of the method may be derived from the elementary fact that under an assigned sum of qth powers of $\{x_i\}$, say $\sum_{i=1}^{k} x_i^q = A$, (1) the minimum value of $\sum_{i=1}^{k} x_i^p$ for a certain $p > q$ and (2) the maximum value of $\sum_{i=1}^{k} x_i^p$ for a certain $p < q$ will be attained when $x_1 = x_2 = \ldots = x_k$ (for each fixed k). Thus we easily get

Lemma 6.1. *Under the general assumptions in §3 in Part II, let us assume that there exists a real number $r > 1$ such that*

$$\sum_{i=1}^{k} (Af c_i)^{\frac{r}{r+1}} = \text{const, say, } M, \qquad \ldots \quad (2.02)$$

for all possible stratifications now under consideration.

Then among all these systems of stratification the minimum value of the variance with the respective optimum allocations is given by the one which will satisfy

$$Af c_1 = Af c_2 = \ldots = Af c_k, \qquad \ldots \quad (2.03)$$

provided that a system of stratification satisfying (2.03) *belongs to the class of all systems of stratification considered.*

The stratifications for which the optimum allocation satisfy (2.03) are those for which equal cost will be divided into each stratum in their optimum allocation. The case when $pr/(1+p) = 1/2$ is specially important because they are concerned with linear combinations of stratum means, and it may be worth noting that there will appear another restriction to Mahalanobis's advocacy as to their cost function.

Indeed we have

Corollary 6.1: *In Lemma 6.1, let us assume in particular that* $r = (p+1)/2p$, *that is to say,*

$$\sum_{j=1}^{k} (A_j^p C_j)^{\frac{1}{2p}} = Const. \qquad \ldots \quad (2.04)$$

Then our conclusion to Lemma 6.1 holds true, so far as $0 < p < 1$.

We are now in a position where either of two courses can be taken into consideration. The first course is to justify the advocacy of Mahalanobis, that is, to give some sufficient conditions in verifying the advocacy which we shall enunciate in Theorem 6.1. The second course is to investigate in more detail a more generalised principle which our mathematical analysis will suggest us and of which Mahalanobis has been surely conscious as seen in Lemma 6.1, that is, the *principle of equipartition* of total cost in each stratum. As to the first course, we may enunciate

Theorem 6.1: *Let k be an assigned positive integer. Let* $S = (S_r)$ *be a set of stratifications with the k strata for which the cost function* (3.02) *in Part II and the condition* (2.02) *in this Part VI will hold true with a certain power p such that* $0 < p < 1$. *Let the cost per unit be assumed to be a common value c which is independent of strata for any* S_r *belonging to S. Let* S_0 *be a stratification for which* $A_1 = A_2 = \ldots = A_k$ *holds true.*

Then S_0 *is a stratification for which the variance of our estimates in its optimum allocation is not greater than any one of* S_r *belonging to S.*

Furthermore specially when $A_i = (N_i\sigma_i)^2$ *and when at the same time, it holds true that*

$$\frac{\sigma_1}{\bar{x}_1} = \frac{\sigma_2}{\bar{x}_2} = \ldots = \frac{\sigma_k}{\bar{x}_k} \qquad \ldots \quad (2.05)$$

for any stratification S_r *belonging to S, then* S_0 *can be charactrised by* (2.01).

This theorem is of course a formal one for which some detailed discussions should be required, because we have no precise knowledge concerning every population parameter and cost functions in the beginning of our surveys, but it may be still useful to make clear the underlying conditions in justifying Mahalanobis' ideas.

Let us now turn to the second course. In our real situations we have to appeal to a set of approximate values $\{A_j'\}$ and $\{c_j'\}$ in order to make stratification so as to satisfy the conditions at least approximately

$$A_1'^p c_1' = A_2'^p c_2' = \ldots = A_k'^p c_k' = \left(\frac{M}{k}\right)^{\frac{p+1}{r}} \qquad \ldots \quad (2.06)$$

Consequently our actual variance to be expected under the adoption of equi-distribution principle should be

$$V_1 = \frac{1}{C^2}\left(\frac{M}{k}\right)^{\frac{p+1}{p}}\left(\sum_{j=1}^{k}\frac{c_j}{c_j'}\right)^{\frac{1}{p}}\left(\sum_{j=1}^{k}\frac{A_j^{\frac{p+1}{p}}}{A_j^{\frac{1}{r}(p+1)}}\right). \qquad \ldots \quad (2.07)$$

On the other hand when we have complete knowledge about parameters, we have

$$V_2 = \frac{1}{C^2}\left(\frac{M}{k}\right)^{\frac{p+1}{rp}} k^{\frac{p+1}{p}} \qquad \ldots \quad (2.08)$$

3. PRINCIPLE OF EQUIPARTITION OF TOTAL COST INTO EACH STRATUM

In view of (2.06) and (2.07) we can observe the following assertions which may have some practical interest in themselves.

3.1. *The optimum determination of the number of strata*: Let us consider the case when all c_j are coincident with a common value $c_0(k)$ and hence all A_j with a common value $A_0(k)$. Under our assumption of having applied the principle of equipartition of total cost into each stratum we have

$$kA_0(k)^{\frac{pr}{p+1}}c_0(k)^{\frac{r}{p+1}} = M. \qquad \ldots \quad (3.01)$$

Let us also put

$$kA_0(k)^{\frac{pr}{p+1}} = T, \qquad \ldots \quad (3.02)$$

and assuming that T is independent upon of k, we shall have

$$V_2 = \frac{T^{\frac{p+1}{pr}}}{C^{1/p}}\left\{c_0(k)k^{\frac{r-1}{r}(1+p)}\right\}^{\frac{1}{p}} \qquad \ldots \quad (3.03)$$

The assumption that T is independent of k holds true under the conditions of Theorem 6.1.

Now the problem how the number k of strata may be determined is seen to be reduced to minimizing

$$\psi(k) = c_0(k)k^{\frac{r-1}{r}(1+p)} \qquad \ldots \quad (3.05)$$

where k runs through the domain $k \geq 1$.

The investigation of cost functions will be done in various sampling surveys, and here k may be considered roughly inversely proportional to the area. Thus as a first approximation we may put here

$$c_0(k) = a + \frac{b}{k^q} \qquad \ldots \quad (3.06)$$

where a, b and q are positive constants independent of k. The value of k which minimizes (3.06) and hence V_p is now given by

$$k_0 = \left\{ \frac{b}{a} \left(\frac{qr}{(r-1)(1+p)} - 1 \right) \right\}^{\frac{1}{q}} \qquad \ldots \quad (3.07)$$

provided that $q > (r-1)(1+p)r^{-1}$ and k_0 gives us an answer so far as $k_0 > 1$.

3.2. *The loss of efficiency by making use of approximate values* $\{A'_j\}$ and $\{c'_j\}$ *in designing sample surveys on the principle of equipartition*: This can be defined by

$$e(A', c'; A, c) \equiv \frac{V_p}{V_t} = \frac{k^{\frac{1+p}{p}}}{\left(\sum_{j=1}^{k} \frac{c_j}{c'_j} \right)^{\frac{1}{p}} \left(\sum_{j=1}^{k} \frac{A'^{\frac{1}{p+1}}_j}{A^{\frac{1}{p+1}}_j} \right)}$$

$$= \frac{k^{\frac{1+p}{p}}}{\left(\sum_{j=1}^{k} \frac{A'^{p}_j}{A^{p}_j} \right)^{\frac{1}{p}} \left(\sum_{j=1}^{k} \frac{A'^{\frac{1}{p+1}}_j}{A^{\frac{1}{p+1}}_j} \right)} \qquad \ldots \quad (3.08)$$

In the case of Mahalanobis principle, we shall have

$$e(A', c'; A, c)$$

$$= \frac{k^{\frac{1+p}{p}}}{\left(\sum_{j=1}^{k} \frac{\sigma_j^{2p}}{\sigma'^{2p}_j} \right)^{\frac{1}{p}} \left(\sum_{j=1}^{k} \frac{\sigma_j^{\frac{2}{p+1}}}{\sigma'^{\frac{2}{p+1}}_j} \right)} \qquad \ldots \quad (3.09)$$

and under the assumption that (2.04) should be assumed not only for the population

parameter $\{\sigma_j\}$ and $\{\bar{x}_j\}$ but also for their respective approximate values $\{\sigma'_j\}$ and $\{\bar{x}'_j\}$, we may write more briefly

$$e(A', c'; A, c) = \frac{k^{\frac{1+p}{p}}}{\left(\sum_{j=1}^{k}(\bar{x}_j\bar{x}_j'^{-1})^{2p}\right)^{\frac{1}{p}}\left(\sum_{j=1}^{k}(\bar{x}'_j\bar{x}_j^{-1})^{p+1}\right)^{\frac{2}{p+1}}} \quad \ldots \quad (3.10)$$

4. Construction of populations with nearly constant coefficients of variation in different strata

Our purpose in this section is to give examples of populations possessing the properties stated in the title.

Let us assume that our whole population consists of an aggregate of some elementary clusters and that any stratum which may be our real concern is also an aggregate of these elementary clusters. Our fundamental idea is to set up elementary clusters whose distributions are all of the Gibrat type, that is, log-normal distributions such as

$$g_{ij}(x) = (2\pi)^{-\frac{1}{2}}(\sigma x)^{-1}\exp\{-(2\sigma^2)^{-1}(\log x - \xi_{ij})^2\} \quad \ldots \quad (4.01)$$

where σ^2 is a value common to all elementary clusters.

Let the size of this cluster Π_{ij} be denoted by N_{ij} and let us define a system of stratification where each stratum Π_i is an aggregate of $\{\Pi_{ij}\}$ $(j = 1, 2, \ldots, M_i)$ and by which our total population is now stratified into a sum of $\{\Pi_i\}$ $(i = 1, 2, \ldots, k)$. Since the mean and the variance of distribution of (4.01) are equal to

$$E\{X_{ij}\} = e^{\xi_{ij}+\frac{\sigma^2}{2}} \quad \ldots \quad (4.02)$$

$$V\{X_{ij}\} = e^{2\xi_{ij}+\sigma^2}(e^{\sigma^2}-1), \quad \ldots \quad (4.03)$$

the following may be readily observed:

(1) The mean of the i-th stratum is

$$E\{X_i\} = \sum_{j=1}^{M_i} p_{ij}\, E\{X_{ij}\} = e^{\frac{\sigma^2}{2}}\sum_{j=1}^{M_i} p_{ij}e^{\xi_{ij}}. \quad \ldots \quad (4.04)$$

(2) The variance within the i-th stratum is

$$V\{X_{i*}\} = \sum_{j=1}^{M_i} p_{ij}(E\{X_{ij}\} - E\{X_{i*}\})^2 + \sum_{j=1}^{M_i} p_{ij} V\{X_{ij}\}$$

$$= e^{\sigma^2} \{e^{\sigma^2} \sum_{j=1}^{M_i} p_{ij} e^{2\xi_{ij}} - (\sum_{j=1}^{M_i} p_{ij} e^{\xi_{ij}})^2\} \qquad \ldots \quad (4.05)$$

(3). The square of the coefficient of variation within the i-th stratum

$$\frac{V\{X_{i*}\}}{E^2\{X_{i*}\}} = \frac{e^{\sigma^2} \sum_{j=1}^{M_i} p_{ij}\ e^{2\xi_{ij}}}{(\sum_{j=1}^{M_i} p_{ij}\ e^{\xi_{ij}})^2} - 1 \qquad \ldots \quad (4.06)$$

where we have put for a moment

$$p_{ij} = N_{ij}(N_{i1} + N_{i2} + \ldots + N_{iM_i})^{-1} \qquad \ldots \quad (4.07)$$

for $\quad j = 1, 2, \ldots, M_i; \quad i = 1, 2, \ldots, k.$

Now concerning the ratios (4.06) we have the following assertions

(a) We have always

$$e^{\sigma^2} - 1 \leqslant \frac{V\{X_{i*}\}}{E^2\{X_{i*}\}} \leqslant e^{\sigma^2} \frac{\max_{1 \leqslant j \leqslant M_i} \{e^{2\xi_{ij}}\}}{(\sum_{j=1}^{M_i} p_{ij}\ e^{\xi_{ij}})^2} - 1 \qquad \ldots \quad (4.08)$$

(b) In particular when ξ_{ij} can be expressed as $\xi_{ij} = \mu_i + \nu_j$, the coefficients of variations are independent of μ_i and merely dependent upon ν_j and p_{ij} such as

$$\frac{V\{X_{i*}\}}{E^2\{X_{i*}\}} = \frac{e^{\sigma^2} \sum_{j=1}^{M_i} p_{ij} e^{2\nu_j}}{\left(\sum_{j=1}^{M_i} p_{ij} e^{\nu_j}\right)^2} - 1. \qquad \ldots \quad (4.09)$$

These assertions (a) and (b) are sufficient to observe under what conditions our stratifications will statisfy at least approximately the condition imposed in 1 of this Part VI. It may be interesting to consider a fine structure of our population by which a set of h systems of stratification each of which will approximately satisfy the

conditions of constant coefficients of variations will be suggested. Our fine structure may be defined by a decomposition of our whole population Π into the strata

$$\Pi = \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} \cdots \sum_{i_h=1}^{k_h} \Pi_{i_1 i_2 \ldots i_h} \qquad \ldots \quad (4.10)$$

where the sizes of $\Pi_{i_1 i_2 \ldots i_h}$ will be denoted by $N_{i_1 i_2 \ldots i_h}$ and they are assumed to be distributed in Gibrat distributions with the parameters $\xi_{i_1 i_2 \ldots i_h}$ and common σ^2. Our fine structure is now assumed to be as follows:

$$\xi_{i_1 i_2 \ldots i_h} = \mu_{1 i_1} + \mu_{2 i_2} + \cdots + \mu_{i_h h} \qquad \ldots \quad (4.11)$$

for all combinations of $i_1, i_2, \ldots, i_{h-1}$ and i_h.

Indeed this is nothing but a h-way classification which will suggest h systems of stratifications each of which will satisfy at least nearly our conditions.

Now we shall proceed to a multi-dimensional Gibrat distribution because we in actual practice consider other quantities which are highly correlated with our variable and which may be also recognised to be distributed in Gibrat distributions. Indeed in applying the principle of equipartitions and that of Mahalanobis, our reference will be concerned with some other variables highly correlated, for example, with the stratum totals of each possible stratification which can be derived from a previous survey. For the sake of simplicity, let us consider here two-dimensional formulations.

Let us now consider a two-dimensional Gibrat distribution according to which two variables (X_{ij}, Y_{ij}) of each elementary cluster Π_{ij} is distributed, namely, the distribution such that $\log X_{ij}$ and $\log Y_{ij}$ are distributed in a bivariate normal distribution with their means ξ_{ij} and η_{ij}, their variances σ_1^2 and σ_2^2 and their correlation coefficient ρ, where we shall assume σ_1^2, σ_2^2, and ρ are common to all elementary clusters. Then the covariance between X_{ij} and Y_{ij} is given by

$$\sigma(X_{ij}, Y_{ij}) = \sigma_{ij} = e^{\xi_{ij} + \frac{\sigma_1^2}{2} + \eta_{ij} + \frac{\sigma_2^2}{2}} (e^{\rho \sigma_1 \sigma_2} - 1). \qquad \ldots \quad (4.12)$$

Now what we are to discuss is the effect of mixturing upon the correlation coefficient $\rho(X_i, Y_i)$ between X_i and Y_i defined for the stratum Π_i. It may be observed that after an amalgation of these elementary clusters $\{\Pi_{ij}\}$ into $\{\Pi_i\}$ correlation coefficient ρ_i resembles closely that of the elementary cluster, because the latter is equal to

$$\rho(X_{ij}, Y_{ij}) = \frac{e^{\rho \sigma_1 \sigma_2} - 1}{(e^{\sigma_1^2} - 1)^{\frac{1}{2}} (e^{\sigma_2^2} - 1)^{\frac{1}{2}}} \qquad \ldots \quad (4.13)$$

while as to the latter we have

$$\rho(X_{ij}, Y_{ij}) = \frac{e^{\rho\sigma_\xi\sigma_\eta} q_{ij}(\xi, \eta)}{e^{\frac{\sigma_\xi^2 + \sigma_\eta^2}{2}} \{q_{ij}(\xi, \xi)\}^{\frac{1}{2}} \{q_{ij}(\eta, \eta)\}^{\frac{1}{2}}} \times$$

$$\times \frac{1 - q_{ij}(\xi, 1) q_{ij}(\eta, 1) q_{ij}(\xi, \eta)^{-1}}{\left(1 - e^{-\sigma_\xi^2} \frac{q_{ij}(\xi, 1)^2}{q_{ij}(\xi, \xi)}\right)^{\frac{1}{2}} \left(1 - e^{-\sigma_\eta^2} \frac{q_{ij}(\eta, 1)^2}{q_{ij}(\eta, \eta)}\right)^{\frac{1}{2}}} \quad \ldots . (4.14)$$

where we have put

$$q_{ij}(\xi, \eta) = \sum_{j=1}^{M_i} p_{ij} e^{\xi_{ij}} e^{\eta_{ij}}, \quad \ldots (4.15)$$

$$q_{ij}(\xi, 1) = \sum_{j=1}^{M_i} p_{ij} e^{\xi_{ij}} \quad \ldots (4.16)$$

and similarly for other notations. Similar assertions to those given in (a) and (b) hold true for $\rho(X_{ij}, Y_{ij})$ and analysis of variance scheme associated with a fine structnre will reveal how this value may depend upon ξ's and η's.

Refernences

Birnbaun, Z. W., and Sirken, M. G. (1950): Bias due to non-availability in sampling surveys. *J. Amer. Stat. Ass.*, 45, 98–111.

Cochran, W. G. (1953): *Sampling Techniques.* John Wiley and Sons, New York.

Deming, W. E. (1944): *Some Theory of Sampling.* John Wiley and Sons, New York.

——— (1950): On errors in surveys. *Amer. Sociological Review*, 9, 359–369.

Ghosh, B. (1949): Interpenotrating networks of samples. *Cal. Stat. Ass. Bull.*, 2, 108–110.

Halperin, M. (1951): Normal regression theory in the presence of interclass correlation. *Ann. Math. Stat.*, 22, 573–580.

Hansen, H. M. and Hurwitz, W. N. (1943): On the theory of sampling from finite populations. *Ann. Math. Stat.*, 14, 333–362.

Hansen, H. M., Hurwitz, W. N., Marks, E. S. and Mauldin, W. P. (1951): Responso errors in surveys. *J. Amer. Stat. Ass.*, 46, 147–190.

Kitagawa, T. (1950a): Successive process of statistical inforonces. (1) *Mem. Fract. Sci.*, Kyushu Univ., Ser. A., 5, 139–180.

——— (1950b): Estimation-formulne used in the sampling survey of the Fishing Catches in Fukuoka Preference: Appendix to Sampling survey of fishing catch in Fukuoka Prefecture. Statistics and Reseorch Division, Agricultural Improvement Bureau Ministry of Agriculture and Forestry, 1–30.

——— (1951a): Successive process of inferences. (2) *Mem. Frac. Sci.*, kyusyn Univ., Ser A. 6. 56–95.

——— (1953a) : Successive process of statistical inferences. (5) *Mem. Front. Sci.*, Kyusyu Univ., Ser. A., 7, 95–120.

——— (1953b) : Successive process of statistical inferences. (6) *Mem. Front. Sci.*, Kyusyu Univ., Ser. A., 8.

KITAGAWA, T. and FUJITA, T. (1951) : *Sampling Surveys of living costs of labourers in Miike Coal mine* (in Japanese). Toyokeizai Shimpo, Tokyo.

MAHALANOBIS, P. C. (1940) : A sampling survey of the acreage under jute in Bengal. *Sankhyā*, 4, 511–530.

——— (1944) : On large-scale sample surveys. *Phil. Trans. Roy. Soc.*, B231, 329–451.

——— (1946) : Recent experiments in statistical sampling in the Indian Statistical Institute. *J. Roy. Stat. Soc.*, 109, 325–370.

MOKASHI, V. K. (1949) : A note on interpenetrating samples. *J. Indian Soc. Agr. Stat.*, 2, 189–195.

SUKHATME, P. V. and SETH, G. R. (1952) : Non-sampling errors in surveys. *J. Indian Soc. Agr. Stat.*, 4, 5–41.

WISHART, M. (1950) : The distribution of the t and z variables in the case of stratified sample with individuals taken from normal parent populations with varying means. *Skandinavisk Aktuarietidskrift*, Heft 3–4, 137–167.

YATES, F. (1949) : *Sampling Methods for Censuses and Surveys*, Charles Griffin and Co., London.

Paper received : August, 1953.