ANALYTIC STUDIES OF SURVEY DATA

By: H. O. Hartley, Iowa State University

1. *The problem of group comparisons in sample surveys*

One of the main objectives of a sample survey is the computation of estimates of (say) means and totals of a number of characteristics attached to the units of a population. More often than not, however, the data are also used for what is known as an 'analytic study' or a 'critical analysis' of a survey. Such an analysis usually involves the comparison of means and totals of certain subgroups of the population. To fix the idea, a farm-economic survey in the State of Iowa may have been primarily planned to obtain estimates of totals of numerous farm-economic items such as annual hog sales or annual bushels of corn sealed for the total population of farms in Iowa. In a subsequent analytic study one may then be concerned with the comparison of some of these items for certain subgroups of farms such as 'owner operators' and 'tenant operators'. Such subgroups of the total population have been termed 'domains of study' by the U. N. Subcommission on Sampling and this term is also used by Yates (1949, 1953) who provides certain formulas for the estimation of their means and variances in the more elementary survey designs.

Although such domains are usually fairly well defined, it will often not be known until after sampling which of the domains any particular unit belongs to. Thus the domains with which an analytic study is concerned are normally not represented in the sample in prescribed fixed sample proportions and the number of sampled units in each domain will itself be a random variable. This is perhaps the most characteristic difference between 'domains' in analytic studies and 'treatment groups' of experiments: but there are others, and the main departures from standard analysis of variance conditions may be summarized under three headings:

1. The number of units in the domains (subgroups) are random variables.
2. The population from which the samples are drawn is finite.
3. Sampling is often not simple random but stratified and/or multistage resulting in correlations of the characteristics of units in the same domain as well as of units in different domains.

To illustrate these points and, at the same time, introduce notation required later we give below three examples of simple survey designs.

Example 1. Simple random sample.

Yates (1949, p. 152) gives data* for a simple random sample of 1/20 Hertfordshire farms. The $n = 125$ sampled farms were, after selection, classified into 7 districts and the number of farms and their total acreages (of crops and grass) are shown in Table 1 below.

TABLE 1

Numbers and total acreages for 125 Hertfordshire farms classified in 7 districts after selection

	District Number							Total (mean)
	1	2	3	4	5	6	7	
No. of farms in sample .	15	8	40	24	4	24	10	125
Total acreage.	1,935	1.385	4.851	4,034	335	2.027	547	15.114
Mean acreage.	129.0	173.1	121.3	168.1	83.8	84.5	54.7	120.9

* Yates (1949) stresses that these data are for illustration only, see his explanations pg. 30-31.

The notation for the entries in Table 1 are shown in Table 1a below.

TABLE 1a

General notation for random sample of n units classified in k domains after sampling

	Domain No.				Total (mean)
	1	2.	j	k	
No. of units in sample .	${}_1n$	${}_2n$	${}_jn$. . . .	${}_kn$	n
Total of character . . .	${}_1y$	${}_2y$	${}_jy$. . . .	${}_ky$	y
Mean of character . . .	${}_1\bar{y}$	${}_2\bar{y}$	${}_j\bar{y}={}_jy/n_j$.	${}_k\bar{y}$	$\bar{y} = y/n$

If we visualize the whole population as subdivided into the domains we reach the notation set out in Table 1b.

TABLE 1b

General notation for population values of domains from which a random sample was drawn

	Domain No.				Total (mean)
	1	2.	j	k	
No. of units in pop. domain	${}_1N$	${}_2N$. . . .	${}_jN$	${}_kN$	N
Total of characteristic . .	${}_1Y$	${}_2Y$. . . .	${}_jY$	${}_kY$	Y
Mean of characteristic . .	${}_1\bar{Y}$	${}_2\bar{Y}$. . . .	${}_j\bar{Y}$	${}_k\bar{Y}$	$\bar{Y} = Y/N$

The main purpose of an analytic study would now consist in estimating the domain means and totals ${}_j\bar{Y}$ and ${}_jY$ from the sample and to provide errors for such estimates. In the present simplest case of a random sample it is easy to guess (as will, in fact, be established later) that the domain population means ${}_j\bar{Y}$ may be estimated by the corresponding sample means ${}_j\bar{y}$. However, the computation of the errors of means of the 'single classification' given in Tables 1 and 1a by standard analysis of variance technique would take no account of either the sampling procedure by which the data were collected or of the actual population for which estimates are required and would, in fact, introduce the assumption of an artificial model not necessarily relevant to the data. That faulty inferences* can be drawn from the application of standard analysis of variance procedures is obvious if we consider the special case when sampling is 100 %, i.e. $n = N$ and when the ${}_j\bar{Y} \equiv {}_j\bar{y}$ are in fact estimated without error.

Example 2. Simple stratified sampling (Yates, 1949, p. 154).

Only the first two entries in each cell, the number of units ${}_jn_h$ and the totals ${}_jy_h$ are shown in the example Table 2; the means ${}_j\bar{y}_h = {}_jy_h/{}_jn_h$ are not entered.

If we visualize the whole population of units (farms) as likewise classified by strata and domains the population numbers and totals corresponding to those in the sample of Table 2a would be denoted by capital letters, i.e.

$$ {}_jN_h\,,\ {}_jY_h\,,\ {}_j\bar{Y}_h = {}_jY_h/{}_jN_h\,;\quad N_h\,,\ Y_h\,,\ \bar{Y}_h = Y_h/N_h\,; $$

$$ {}_jN,\ {}_jY,\ {}_j\bar{Y} = {}_jY/{}_jN,\quad N,\ Y,\ \bar{Y} = Y/N\,; $$

This simple example shows the occurrence of correlation (say) between units in the same domain: Of the ${}_jn$ units giv-

* Note also discussion in Section 8.

TABLE 2

Numbers of farms and total wheat acreages for a sample of $n = 135$ Hertfordshire farms stratified by 'size-group' and classified by District after selection

Size group stratum (acres)	Population size N_h	District Number 1	2	3	4	5	6	7	Total
6- 20	519	0 0	1 0	0 0	1 0	0 0	1 0	0 0	3 0
21- 50	357	1 0	1 10	2 0	1 17	0 0	1 0	0 0	6 27
51-150	519	3 36	3 40	5 40	5 65	2 0	5 19	3 14	26 214
151-300	400	4 63	5 213	10 270	8 305	5 112	6 140	2 60	40 1163
301-500	215	4 320	10 1074	11 659	12 989	3 234	2 0	1 16	43 3292
501-	51	1 114	4 487	2 315	9 1937	0 0	1 72	0 0	17 2925
	Number	13	24	30	36	10	16	6	135
Total	Wheat acreage	533	1824	1284	3313	346	231	90	7621

TABLE 2a

General notation for sample of n units stratified in L strata and classified in k domains after selection

Stratum number	Domain Number	1 . . .	j . . .	k	Total (mean)
	No. of units	${}_1n_1$	${}_jn_1$	${}_kn_1$	n_1
$h = 1$	Total.	${}_1y_1$	${}_jy_1$	${}_ky_1$	y_1
	Mean.	${}_1\bar{y}_1$	${}_j\bar{y}_1$	${}_k\bar{y}_1$	$\bar{y}_1$
		. . . . .	. . . . .	. . . . .	. . .
	No. of units	${}_1n_h$	${}_jn_h$	${}_kn_h$	n_h
h	Total.	${}_1y_h$	${}_jy_h$	${}_ky_h$	y_h
	Mean.	${}_1\bar{y}_h$	${}_j\bar{y}_h$	${}_k\bar{y}_h$	$\bar{y}_h$
		. . . . .	. . . . .	. . . . .	. . .
	No. of units	${}_1n_L$	${}_jn_L$	${}_kn_L$	n_L
L	Total.	${}_1y_L$	${}_jy_L$	${}_ky_L$	y_L
	Mean.	${}_1\bar{y}_L$	${}_j\bar{y}_L$	${}_k\bar{y}_L$	$\bar{y}_L$
	No. of units	${}_1n$	${}_jn$	${}_kn$	n
Total	Total.	${}_1y$ . .	${}_jy$ . .	${}_ky$	y
(Mean)	Mean.	${}_1\bar{y}$	${}_j\bar{y}$	${}_k\bar{y}$	$\bar{y}$

ing rise to the domain mean ${}_j\bar{y}$ 'a cluster' of ${}_jn_a$ will be found in stratum h and will usually be positively correlated. The correlation is similar to that found in a two-way table of an analysis of variance with unequal cell frequencies but, of course, cannot here be assumed to have been caused by an additive model.

Example 3. Two stage sample, primaries drawn with replacement and probabilities proportional to size (p. p. s.) and secondaries drawn without replacement and with equal probability.

Table 3 below gives data from a 'consumer preference survey' carried out by the Statistical Laboratory of Iowa State College in the City of Des Moines. The survey was arranged as a stratified two-stage design with 50 city blocks as strata from each of which were sampled 2 'segments' (primaries) each containing an expected number of 5 households (secondaries). After completion of the survey the data in each stratum were classified by segment number $t = 1, t = 2$ and in 6 different Income Groups $j = 1, 2, \ldots, 6$. Actually Table 3 gives the totals for 5 strata-groups combined by pooling the answers for all segments 1 and all segments 2 for the 10 strata in each group. The table shows for each stratum group:

(*a*) the number of households in each classification cell (${}_jm_t$, top line)

(*b*) the total number of persons ${}_jy_t$ in households in the j^{th} income group and in segments $t = 1$ and segments $t = 2$ respectively

(*c*) the average number of persons per household (${}_j\bar{y}_t$) for each cell of the two-way classification.

Table 3a shows the notation in the general case of a stratum from which n primaries ($t = 1, 2, \ldots, n$) were drawn with the sample results subdivided into k domains ($j = 1, 2, \ldots, k$). The stratum subscript h has been omitted from all symbols.

TABLE 3

Number of households (a), number of persons (b) and number of persons per household (c) for 467 households sampled from the City of Des Moines and arranged in 5 'strata groups', 6 'income groups' and 2 segments

Stratum group	Segment		Income group 1 Less than $25 per week	2 $25-$50 per week	3 $50-$75 per week	4 $75-$100 per week	5 $100-$125 per week	6 More than $125 per week	Total
1	1	a	—	7	5	13	14	8	47
		b	—	21	14	42	55	25	157
		c	—	3.00	2.80	3.23	3.93	3.12	3.34
	2	a	2	8	6	12	5	13	46
		b	2	23	16	45	21	50	157
		c	1.00	2.88	2.67	3.75	4.20	3.85	3.41
2	1	a	2	9	13	9	10	13	56
		b	2	20	42	26	37	43	170
		c	1.00	2.22	3.23	2.89	3.70	3.31	3.04
	2	a	2	9	9	13	7	10	50
		b	2	14	20	48	30	34	148
		c	1.00	1.56	2.22	3.69	4.28	3.40	2.96
3	1	a	11	8	15	9	—	4	47
		b	21	24	50	31	—	10	136
		c	1.91	3.00	3.33	3.44	—	2.50	2.89
	2	a	1	8	11	7	5	3	35
		b	1	17	36	26	19	17	116
		c	1.00	2.12	3.27	3.71	3.80	5.67	3.31
4	1	a	4	5	15	5	1	12	42
		b	8	12	47	13	3	36	119
		c	2.00	2.40	3.13	2.60	3.00	3.00	2.83
	2	a	3	2	20	9	2	9	44
		b	6	1	50	26	3	28	115
		c	2.00	2.00	2.50	2.89	1.50	3.11	2.61
5	1	a	8	10	10	9	9	12	58
		b	15	26	37	32	27	45	182
		c	1.88	2.60	3.70	3.56	3.00	3.75	3.14
	2	a	—	7	14	8	8	5	42
		b	—	15	50	34	27	9	135
		c	—	2.14	3.57	4.25	3.38	1.80	3.21

If we visualize the whole population of units as likewise classified by primaries and domains the population numbers and totals corresponding to those in the sample of Table 3a would be denoted by capital letters, i.e. by

$$N; \quad {}_jM_t, {}_jY_t, {}_j\bar{Y}_t = {}_jY_t/{}_jM_t; \quad M_t, Y_t, \bar{Y}_t = Y_t/M_t;$$

$${}_jM, {}_jY, {}_j\bar{Y} = {}_jY/{}_jM; \quad M, Y, \bar{Y} = Y/M;$$

TABLE 3a

General notation for two stage sample of n primaries $t = 1, 2, \ldots, n$ containing respectively m_t sampled secondaries which are classified into k domains after sampling

Primary Number		Domain No. 1 . . .	Domain No. j . . .	Domain No. k	Total
	No. of units	${}_1m_1$	${}_jm_1$	${}_km_1$	m_1
$t=1$	Total	${}_1y_1$	${}_jy_1$	${}_ky_1$	y_1
	Mean	${}_1\bar{y}_1$	${}_j\bar{y}_1$	${}_k\bar{y}_1$	$\bar{y}_1$
	No. of units	${}_1m_t$	${}_jm_t$	${}_km_t$	m_t
t	Total	${}_1y_t$	${}_jy_t$	${}_ky_t$	y_t
	Mean	${}_1\bar{y}_t$	${}_j\bar{y}_t$	${}_k\bar{y}_t$	$\bar{y}_t$
	No. of units	${}_1m_n$	${}_jm_n$	${}_km_n$	m_n
n	Total	${}_1y_n$	${}_jy_n$	${}_ky_n$	y_n
	Mean	${}_1\bar{y}_n$	${}_j\bar{y}_n$	${}_k\bar{y}_n$	$\bar{y}_n$
Total	No. of units	${}_1m$	${}_jm$	${}_km$	m
	Total	${}_1y$	${}_jy$	${}_ky$	y
	Mean	${}_1\bar{y}$	${}_j\bar{y}$	${}_k\bar{y}$	$\bar{y}$

The primary index $t = 1, 2, \ldots, N$ now runs through the complete population of primary units. If the population consists of L strata the stratum index $h = 1, 2, \ldots, L$ will precede the primary index t in all symbols.

2. *Estimation of domain totals, variance formulas and variance estimation*

We begin with the estimation of domain totals leaving that for domain means for the next section. It is clear that any theory of estimating domain totals or means will have to cover, as a special case, the situation when the 'domain' consists of the total population, and this is, of course, the well known theme dealt with in the literature on sample survey methodology. As is well known, this generally accepted theory of estimation is essentially distribution free, does not accept any particular models, is in fact based on the first two moments only and can be roughly described as 'unbiased estimation with optimum variance properties'.

In the following we shall accept this theory of estimation also for the more general problem of estimating domain totals and means.

We recall here certain basic formulas for the estimation of population totals which may be found in most text books on the subject: Appropriate to any particular design there are three basic formulas pertinent to the estimation of the population total. They are in general notation:

(*a*) The estimate of the population total Y: —

$$\hat{Y} = \hat{Y}(y_i) \qquad [1]^*$$

(*b*) The population variance formula for $\hat{Y}$: —

$$\mathrm{Var}(\hat{Y}) = V(y_i) \qquad [2]^*$$

(*c*) The estimated variance of $\hat{Y}$: —

$$\mathrm{var}(\hat{Y}) = \mathrm{v}(y_i) \qquad [3]^*$$

The multi-variable functions $\hat{Y}(y_i)$ $V(y_i)$ and $\mathrm{v}(y_i)$ will depend on the particular design which has given rise to the sample of y_i *.

For example in stratified sampling (*Example* 2) we would have

$$\hat{Y} \equiv \hat{Y}(y_{hi}) = \sum_{h=1}^{L} N_h \bar{y}_h \qquad [4]$$

$$\mathrm{Var}(\hat{Y}) \equiv V(y_{hi}) = \sum_{h=1}^{L} N_h^2 (1 - n_h/N_h)\, n_h^{-1} S_h^2 \qquad [5]$$

$$\mathrm{var}(\hat{Y}) \equiv \mathrm{v}(y_{hi}) = \sum_{h=1}^{L} N_h^2 (1 - n_h/N_h)\, n_h^{-1} s_h^2 \qquad [6]$$

$$\text{where } S_h^2 = (N_h - 1)^{-1} \sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2 \qquad [7]$$

$$s_h^2 = (n_h - 1)^{-1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2 \qquad [8]$$

and these formulas for $L = 1$ (one stratum) would yield as a special case the corresponding formulas for simple sampling (*Example* 1). The appropriate formulas for the two stage design of *Example* 3 (primaries p.p.s. equal take $\bar{m}$ of secondaries)

$$\hat{Y} \equiv \hat{Y}(y_{ti}) = M\bar{y} \qquad [9]$$

$$\mathrm{Var}(\hat{Y}) \equiv V(y_{ti}) = \frac{M}{n} \sum_{t=1}^{N} \frac{M_t}{\bar{m}} (1 - \bar{m}/M_t) S_t^2 + \frac{M^2}{n} \sum_{t=1}^{N} \frac{M_t}{M} (\bar{Y}_t - \bar{Y})^2 \qquad [10]$$

$$\mathrm{var}(\hat{Y}) \equiv \mathrm{v}(y_{ti}) = \frac{M^2}{n(n-1)} \sum_{t=1}^{n} (\bar{y}_t - \bar{y})^2 \qquad [11]$$

$$\text{where } S_t^2 = (N_t - 1)^{-1} \sum_{i=1}^{M_t} (y_{ti} - \bar{Y}_t)^2 \qquad [12]$$

$$s_t^2 = (n_t - 1)^{-1} \sum_{i=1}^{\bar{m}} (y_{ti} - \bar{y}_t)^2 \qquad [13]$$

In general $\hat{Y}$ and $\mathrm{v}(\hat{Y})$ will depend on the sampled y_i only, and these are the formulas used in practical evaluation of the estimate and its variance, whilst $V(y_i)$ depends on the entire population of y_i values and is used for the theoretical evaluation of the merits of the estimator. The three formulas are applicable to *any* set of characteristics y_i attached to the population units. This enables us to use these same formulas for the estimation of domain totals and their variances: — To this end we introduce the following characteristics ${}_jy_i$ which we shall attach to *all* units in the population.

$${}_jy_i = \begin{cases} y_i \text{ if the ith unit belongs to jth group (domain)} \\ 0 \text{ otherwise} \end{cases} \qquad [14]$$

Now the group total ${}_jY$ of our j^{th} group is seen to be the population total of the ${}_jy_i$ and standard sample survey theory, therefore, provides the following estimators:

(*a*) The estimate of the domain total ${}_jY$: —

$${}_j\hat{Y} = \hat{Y}({}_jy_i) \qquad [15]$$

(*b*) The population variance formula for ${}_j\hat{Y}$: —

$${}_jV = V({}_jy_i) \qquad [16]$$

* The arguments y_i denote the characteristics attached to the *individual* units; in stratified sampling they would of course show a double subscript y_{hi}, in multistage sampling a triple subscript y_{hti}, etc.

(*c*) The estimated variance of ${}_j\hat{Y}$: —

$$ {}_j\mathrm{v} = \mathrm{v}\,({}_jy_i) \qquad [17] $$

together with the assurance of unbiasedness

$$ E\,({}_j\hat{Y}) = {}_jY \qquad [18] $$

$$ E\,({}_jV) = {}_jV_i \qquad [19] $$

The meaning of formulas [15] to [17] is simply that the standard formulas for estimation [4] to [13] be applied to the characteristics [14]. The ' spelling out ' of these formulas often results in simplifications: In Example 2 (stratified sampling) we find that formulas [15] to [17] can be written as

(*a*) The estimate of the domain total:

$$ {}_j\hat{Y} = \sum_h N_h\, {}_jy_h/n_h \qquad [20] $$

(*b*) The population variance formula for ${}_j\hat{Y}$

$$ {}_jV = \sum_h \frac{N_h\,(N_h - n_h)}{n_h\,(N_h - 1)} \left\{ \sum_{i=1}^{{}_jN_h} {}_jy_{hi}^2 - {}_jY_h^2/N_h \right\} \qquad [21] $$

(*c*) The estimated variance of ${}_j\hat{Y}$

$$ {}_j\mathrm{v} = \sum_h \frac{N_h\,(N_h - n_h)}{n_h\,(n_h - 1)} \left\{ \sum_{i=1}^{{}_jn_h} {}_jy_{hi}^2 - {}_jy_h^2/n_h \right\} \qquad [22] $$

The last formula may be compared with one given in the 2nd edition of Yates' book (Yates 1953, 301 formula [9.3.e]) which corrects an earlier (faulty) formula given in his 1st edition (Yates 1949, 202). To make this comparison we use the sample variance of the units in the h^{th} stratum and in the j^{th} domain, we write

$$ {}_js_h^2 = \sum_{i=1}^{{}_jn_h} ({}_jy_{hi} - {}_j\bar{y}_h)^2/({}_jn_h - 1) $$

and hence obtain [22] in the form

$$ {}_jV = \sum_h \frac{N_h\,(N_h - n_h)}{n_h\,(n_h - 1)} \left\{ ({}_jn_h - 1)\,{}_js_h^2 + \left(\frac{1}{{}_jn_h} - \frac{1}{n_h} \right) {}_jy_h^2 \right\} \qquad [23] $$

which agrees with Yates' formula [9.3.e]. Our formulas [22] and [23] have been proved to be unbiased estimates of the exact variance of ${}_j\hat{Y}$. The second term of [23] which is proportional to the square of the domain total ${}_jy_h^2$, is characteristic of estimators of *totals* of a population domain whose size is unknown and this feature will, of course, disappear in the estimation of domain *means*.

In Example 3, (two stage sampling with primaries drawn p.p.s.) if we assume that an equal take of $\bar{m}$ secondaries are sampled from each primary, formulas [15] and [17] spell out as follows:

(*a*) The estimate of the domain total:

$$ {}_j\hat{Y} = M\,{}_jy/m = M \left(\frac{{}_jm}{m} \right) {}_j\bar{y} \qquad [24] $$

(*c*) The estimated variance of ${}_j\hat{Y}$

$$ {}_j\mathrm{v} = \frac{M^2}{n\,(n-1)} \sum_{t=1}^{n} \left(\frac{{}_jy_t}{m_t} - \frac{{}_jy}{m} \right)^2 \qquad [25] $$

The interpretation of [24] is simple: The sample domain mean ${}_j\bar{y}$ estimates the corresponding population mean ${}_j\bar{Y}$ and the fraction ${}_jm/m$ the corresponding population fraction ${}_jM/M$, the unbiasedness of the product being assured by [18].

3. *Estimation of domain means, variance formulas and variance estimation*

In the preceding section we gave estimators ${}_j\hat{Y}$ of the population total ${}_jY$ of the j^{th} domain. We now turn to the estimation of the domain means ${}_j\bar{Y}$ (given by ${}_jY/{}_jN$ in the notation of Example (1 and 2 and Table 1 b, and by ${}_jY_j/M$ in the notation of Example 3). Even if the number of units in the j^{th} domain ${}_jN$ (or ${}_jM$) were known it would usually be unwise to use ${}_j\hat{Y}/{}_jN$ (or in two-stage sampling ${}_j\hat{Y}/{}_jM$) as an estimator of ${}_j\bar{Y}$, since it is well known from experience with such estimators of population means that their variances are large. In any case ${}_jN$ or ${}_jM$ will often be unknown (as are, for example, the number of tenant-operators in Iowa) and in these circumstances we are forced to estimate *both* the unknown numerator ${}_jY$ and unknown denominator ${}_jN$ (or ${}_jM$). This direct approach automatically leads to what is known as the " combined ratio estimator " of ${}_j\bar{Y}$. Other ratio estimators, available in special situations, are briefly discussed in section 6. In order to estimate, then, the denominator ${}_jN$ (or ${}_jM$) by precisely the same method used for the numerator ${}_jY$ we introduce the " count variates "

$$ {}_ju_i = \begin{cases} 1 \text{ if } i^{th} \text{ unit is in } j^{th} \text{ domain} \\ 0 \text{ otherwise} \end{cases} \qquad [26] $$

and these obtain the estimate of the domain size ${}_jN$ (or ${}_jM$) as

$$ {}_j\hat{U} = \hat{Y}\,({}_ju_i). \qquad [27] $$

Provided that ${}_j\hat{U} > 0$, that is provided there is at least one sampled unit in the j^{th} domain, we can then use

(*a*) the combined ratio estimator of the domain mean ${}_j\bar{Y}$

$$ {}_j\hat{y} = {}_j\hat{Y}/{}_j\hat{U} \qquad [28] $$

In general this estimator will be biased and conditions as to when this bias is negligible are known from the literature (see e.g. Cochran, 1953, pp. 114-8). Unbiased ratio estimators are briefly discussed in section 6. Standard approximate formulas (see e.g. Cochran, 1953, pp. 114-120) for the variance and estimated variance of the ratio estimator likewise yield

(*b*) The approximate variance formula for ${}_j\hat{y}$

$$ \mathrm{Var}\,({}_j\hat{y}) \doteq {}_jN^{-2}\,V\,({}_jy_i - {}_j\bar{Y}\,{}_ju_i) \qquad [29] $$

(*c*) The approximate estimate of the variance of ${}_j\hat{y}$

$$ \mathrm{var}\,({}_j\hat{y}) = {}_j\hat{U}^{-2}\,\mathrm{v}\,({}_jy_i - {}_j\hat{y}\,{}_ju_i) \qquad [30] $$

The spelling out of these formulas [28], [29] and [30] will show that the domain mean ${}_j\bar{Y}$ is usually estimated by the simple sample domain mean ${}_j\bar{y}$. This will certainly be so in Examples 1 and 3 as indeed with all self-weighting survey designs. Likewise the estimated variance of ${}_j\hat{y}$ will often simplify:

In Example 2 we have the following formulas

(*a*)
$$ {}_j\hat{y} = \sum_h \frac{N_h}{n_h}\, {}_jy_h \Big/ \sum_h \frac{N_h}{n_h}\, {}_jn_h \qquad [31] $$

which, in the case of proportional allocation $N_h/n_h = N/n$ will reduce to

(*a*)
$$ {}_j\hat{y} = {}_j\bar{y}. \qquad [32] $$

The estimated variance in this example from [30] and [23] is given by

$$ \mathrm{var}\,({}_j\hat{y}) = {}_j\hat{U}^{-2} \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{n_h\,(n_h - 1)} \left\{ {}_j\Sigma_h + {}_jq_h\,{}_jn_h\,({}_j\bar{y}_h - {}_j\hat{y})^2 \right\} \qquad [33] $$

where ${}_j\Sigma_h = ({}_jn_h - 1)\,{}_js_h^2$ is the sum of squares of deviations of

the y in the j^{th} domain of the h_{th} stratum about their mean $_j\bar{y}_h$ and $_jq_h = (n_h - {}_jn_h)/n_h$ the proportion of units sampled outside the j^{th} domain. Formula [33] agrees with Yates' (1953, p. 301 formula [9.3.c]) except that the latter appears to be restricted to a proportional allocation of the sample to strata. The formula for Example 1 (simple sampling) is obtained as a special case of [33] when $L = 1$ and $_j\bar{y}_h = {}_j\bar{y}$ yielding

$$\text{var } {}_j\hat{y} = \left(1 - \frac{n}{N}\right)\left(\frac{n}{n-1}\right)\left(\frac{{}_jn - 1}{{}_jn}\right){}_js^2/{}_jn \qquad [34]$$

which is, approximately, equal to the familiar variance of a mean of a sample of size $_jn$.

It should be noted that the case of stratified sampling is not derivable from the simple sampling case by summation over strata since this would only yield the first terms inside the $\{\}$ of [33]. The second terms allow for the fact that the strata proportions $_jN_h/{}_jN$ for the j^{th} domain are not known.

A numerical example:

We illustrate the above formulas by evaluating $_j\hat{y}$ and var $(_j\hat{y})$ for Example 2 using the data of Table 2 to evaluate an estimate of the mean wheat acreage for District $j = 1$.

The N_h and n_h are given in the second and last (top entry) columns of Table 2 and the $_1y_h$ and $_1m_h$ in the third column. For the individual wheat acreage values $_1y_{hi}$ of the farms in the first district reference is made to Yates (1949, p. 154). We compute

$$_1\hat{Y} = 3291\,, \quad {}_1\hat{U} = 183\,, \quad {}_1\hat{y} = 17.98 \quad , \quad \text{var } (_1\hat{y}) = 61.52\,.$$

Finally the spelling out of formulas [30] and [28] for Example 3 yields

$$\text{var } (_j\hat{y}) \doteq \frac{n}{n-1}\sum_{i=1}^{n}\left(\frac{{}_jm_i}{{}_jm}\right)^2 ({}_j\bar{y}_i - {}_j\hat{y})^2 \qquad [35]$$

where $_j\hat{y} = {}_j\bar{y}$.

4. *Correlation between the estimates of domain totals and means and variances of differences between means*

In the preceding sections we derived formulas for the estimation of domain totals and means and their variances. We now turn to the task of the "comparison" of two estimated domain means, and to this end require estimates of their covariances in order to obtain variance estimates for the *difference of two* estimated domain means.

Again, we first deal with the estimates of domain-totals and afterwards with the domain means. It is convenient to write all formulas for the comparison of two particular domains which, without loss of generality, may be taken as domains $j = 1$ and $j = 2$.

The following formulas are obvious by arguments similar to those of section 2.

$$\text{Var } (_1\hat{Y} - {}_2\hat{Y}) = V\,(_1y_i - {}_2y_i) \qquad [36]$$

$$2\ \text{Cov } (_1\hat{Y}\,,\ {}_2\hat{Y}) = V\,(_1y_i) + V\,(_2y_i) - V\,(_1y_i - {}_2y_i) \qquad [37]$$

$$\text{var } (_1\hat{Y} - {}_2\hat{Y}) = \text{v}\,(_1y_i - {}_2y_i) \qquad [38]$$

$$2\ \text{cov } (_1\hat{Y}\,,\ {}_2\hat{Y}) = \text{v}\,(_1y_i) + \text{v}\,(_2y_i) - \text{v}\,(_1y_i - {}_2y_i) \qquad [39]$$

Formula [38] gives directly the expression required for estimating the variance of the difference between two estimates $_1\hat{Y}$ and $_2\hat{Y}$ of the two domain totals. Formula [39], however, is required for proving that cov $(_1\hat{Y}, {}_2\hat{Y})$ is given by the bilinear form in the variates $_1y_i$ and $_2y_i$ which corresponds to the symmetric quadratic form v (y_i). The same remarks apply to the variance formulas [37] and [36].

Turning now to the domain means we require an approximate expression for the covariance between two ratios y/u and x/v: This follows on identical lines as the familiar variance formulas and yields:

$$\text{Cov}\left(\frac{y}{u}\,,\ \frac{x}{v}\right) \doteq (E\,(u)\ E\,(v))^{-1}\,\text{Cov}\left(y - \frac{E\,(y)}{E\,(u)}u\,,\ x - \frac{E\,(x)}{E\,(v)}v\right) \qquad [40]$$

The combination of [36] and [39] can be shown to yield

$$\text{Var } (_1\hat{y} - {}_2\hat{y}) \equiv \text{Var}\left(\frac{{}_1\hat{Y}}{{}_1\hat{U}} - \frac{{}_2\hat{Y}}{{}_2\hat{U}}\right)$$
$$\doteq V\left(\frac{1}{{}_1N}\,(_1y_i - {}_1\bar{Y}\ {}_1u_i) - \frac{1}{{}_2N}\,(_2y_i - {}_2\bar{Y}\ {}_2u_i)\right) \qquad [41]$$

The corresponding approximation for the estimated variance yields

$$\text{var } (_1\hat{y} - {}_2\hat{y}) \doteq \text{v}\left(\frac{1}{{}_1\hat{U}}\,(_1y_i - {}_1\hat{y}\ {}_1u_i) - \frac{1}{{}_2\hat{U}}\,(_2y_i - {}_2\hat{y}\ {}_2u_i)\right) \qquad [42]$$

where the variables $_jy_i$ and $_ju_i$ are given by [14] and [26] respectively, the estimates $_j\hat{U}$ and $_j\hat{y}$ by [27] and [28] and the multivariable function v is defined in [3]. This function v will of course depend on the particular survey design but there is no difficulty in spelling it out in any particular case. We proceed to do so for our examples 1, 2 and 3 when v (y_i) is given by equations (6) and (11) respectively. For example 2, simple stratified sampling with varying sampling fractions, we obtain the formula

$$\text{v}\,(_1\hat{y} - {}_2\hat{y}) = \sum_h a_h\left\{\frac{{}_1\Sigma_h}{{}_1\hat{U}^2} + \frac{{}_2\Sigma_h}{{}_2\hat{U}^2} + 2\,n_h\,{}_1p_h\,{}_2p_h\,\frac{(_1\bar{y}_h - {}_1\hat{y})}{{}_1\hat{U}}\,\frac{(_2\bar{y}_h - {}_2\hat{y})}{{}_2\hat{U}} + \right.$$
$$\left. + n_h\,{}_1p_h\,{}_1q_h\,(_1\bar{y}_h - {}_1\hat{y})^2 + n_h\,{}_2p_h\,{}_2q_h\,\frac{(_2\bar{y}_h - {}_2\hat{y})^2}{{}_2\hat{U}^2}\right\} \qquad [43]$$

where

$$a_h = \frac{N_h\,(N_h - n_h)}{n_h\,(n_h - 1)}\,, \quad {}_jp_h = \frac{{}_jn_h}{n_h}\,, \quad {}_jq_h = 1 - {}_jp_h$$

$${}_j\Sigma_h = \sum_{i=1}^{{}_jn_h}({}_jy_{hi} - {}_j\bar{y}_h)^2$$

Formula [43] is in agreement with that for $\hat{\text{cov}}\,(_1\hat{y}\,,\ {}_2\hat{y})$ given by Yates (1953, p. 301, formula 9.3.d) except that the latter appears to be restricted to stratified sampling with proportional allocation. The first two terms in [43] are the within domain within strata components, whilst the last three terms have the form of multinomial variances and covariances and allow for the fact that the strata proportions $_1N_h/{}_1N$ and $_2N_h/{}_2N$ are unknown. These latter terms disappear in the special case of a single stratum $L = 1$ which yields the answer for example 1 in the form

$$V\,(_1\hat{y} - {}_2\hat{y}) = \left(1 - \frac{n}{N}\right)\left(\frac{n}{n-1}\right)\left\{\frac{{}_1n - 1}{{}_1n}\,\frac{{}_1s^2}{{}_1n} + \frac{{}_2n - 1}{{}_2n}\,\frac{{}_2s^2}{{}_2n}\right\} \qquad [44]$$

This expression is approximately equal to $(_1s^2/{}_1n + {}_2s^2/{}_1n)$ i.e. the familiar variance formula for the difference of two means based on fixed samples of size $_1n$ and $_2n$.

Finally, the spelling out of formula [42] for Example 3 (two stage sampling primaries drawn p.p.s. equal take of secondaries) yields the surprisingly simple formula

$$\text{v}\,(_1\hat{y} - {}_2\hat{y}) \doteq \frac{n}{n-1}\sum_{i=1}^{n}\left\{\frac{{}_1m_i}{{}_1m}\,(_1\bar{y}_i - {}_1\hat{y}) - \frac{{}_2m_i}{{}_2m}\,(_2\bar{y}_i - {}_2\hat{y})\right\}^2 \qquad [45]$$

In the important case of $n = 2$ primaries this formula simplifies further to

$$v\,({}_1y - {}_2y) = 4\,({}_1w\,({}_1\bar{y}_1 - {}_1\bar{y}_2) - {}_2w\,({}_2\bar{y}_1 - {}_2\bar{y}_2))^2 \qquad [46]$$

where

$${}_iw = {}_im_1\,{}_im_2/({}_im_1 + {}_im_2)^2$$

Certain special cases of formulas [44] and [45] were recently obtained by an independent derivation by L. Kish and Irene Hess (1955).

5. *The domain total expressed as a proportion of the population total.*

In many analytic studies of sample survey data we are interested in the proportion of a measured characteristic which falls into a specified domain. For example in a consumer study we may be interested in the proportion of say the total milk consumption which is attributable to families of a particular income group. Or, again, in a soil survey we may be interested in the proportion of the total farm land which is of a particular soil type. In the preceding sections we have estimated the total ${}_iY$ of the variate values y_i which fall into the j^{th} domain by the estimator $\hat{{}_iY}$. It is therefore suggested that the proportion ${}_iY/Y$ be estimated by the ratio

$${}_i\hat{p} = \hat{{}_iY}/\hat{Y} \qquad [47]$$

The properties of this ratio estimator can be evaluated by a method similar to that used in Section 3: We attach two variates to each unit in the population viz
the numerator variable given (by [14] i.e.) by

$${}_iy_i = \begin{cases} y_i \text{ if the } i^{th} \text{ unit belongs to the } j^{th} \text{ group} \\ 0 \text{ otherwise} \end{cases} \qquad [48]$$

and the denominator variable given by y_i. Since $\hat{{}_iY}$ and $\hat{Y}$ are, respectively, the estimates of the population totals for the variates ${}_iy_i$ and y_i the estimator ${}_i\hat{p}$ is seen to be the standard (combined) ratio estimator for ${}_iP = {}_iY/Y$ of the respective population totals. Its variance may therefore be obtained by the standard (approximate) formula viz.

$$\mathrm{Var}\,{}_i\hat{p} = Y^{-2}\,V\,({}_iy_i - {}_iP\,y_i) \qquad [49]$$

where the variance function $V(y)$ is defined by [2]. Let us spell out the general formula [49] for the particular case of a simple stratified design and in terms of the notation used in Table 2a and that described immediately following Table 2a. We obtain the formula

$$\mathrm{Var}\,{}_i\hat{p} = Y^{-2}\sum \frac{N_h^2}{n_h\,(N_h - 1)} \sum_{i=1}^{N_h} (d_{hi} - \bar{D}_h)^2 \qquad [50]$$

where $d_{hi} = {}_iy_{hi} - {}_iP\,y_{hi}$ and $\bar{D}_h$ is the stratum mean of the d_{hi}.

Yates (1953) p. 304 only discusses this case of a stratified sample and offers a formula only for the special case when the domains do not cut across the strata. In the general case of domains cutting across the strata Yates gives hints for variance computation and we have not attempted to identify the result of these instructions with our simple formula [50].

For the purpose of variance comparison, an alternative form of [50] may be more useful. After some algebra we reach the formula

$$\mathrm{Var}\,({}_i\hat{p}) = Y^{-2}\sum \frac{N_h^2}{n_h\,(N_h - 1)} \{{}_iQ^2\,{}_i\Sigma_h + {}_iP^2\,{}_i\Sigma_h + N_h^{-1}\,{}_iN_h\,{}_jN_h\,({}_jQ\,{}_i\bar{Y}_h + {}_iP\,{}_j\bar{Y}_h)^2\} \qquad [51]$$

where

${}_iN_h$ is the number of units in the h^{th} stratum falling within the j^{th} domain

${}_jN_h$ is the number of units in the h^{th} stratum not falling within the j^{th} domain

${}_iY_h$ is the y-total for the h^{th} stratum of units in the j^{th} domain

${}_jY_h$ is the y-total for the h^{th} stratum of units not in the j^{th} domain.

So that

$${}_i\bar{Y}_h = {}_iY_h/{}_iN_h \quad {}_j\bar{Y}_h = {}_jY_h/{}_jN_h$$

are the corresponding means.

Further

$${}_iP = \sum_h {}_iY_h/Y \quad {}_iQ = \sum_h {}_jY_h/Y = 1 - {}_iP$$

$${}_i\Sigma_h = \sum_{i=1}^{{}_iN_h} ({}_iy_{hi} - {}_i\bar{Y}_h)^2$$

is the sum of squares of deviations of the y values in the h^{th} stratum and j^{th} domain from their mean and

$${}_j\Sigma_h = \sum_{i=1}^{{}_jN_h} ({}_iy_{hi} - {}_j\bar{Y}_h)^2$$

is the corresponding sum of squares for the y-values in the h^{th} stratum and *not* in the j^{th} domain.

To interpret the three terms in [51] we may write the estimator in the form

$$1/{}_i\hat{p} = 1 + (\hat{{}_jY}/\hat{{}_iY}) \qquad [52]$$

where $\hat{{}_jY} = \hat{Y} - \hat{{}_iY}$ is an estimate for the y-total *not* in the j^{th} domain. It can be shown that the first two terms of [51] are contributed by variation in y-values only, holding the ratios ${}_in_h/n_h$ constant at their expected values $\frac{{}_iN_h}{N_h}\,n_h$. The third term allows for the variability in the proportion ${}_in_h/n_h$ and is seen to be a binomial type of variance being approximately equal to

$$\sum_h \left(\frac{{}_iN_h}{N_h}\right)\left(\frac{{}_jN_h}{N_h}\right)\frac{1}{n_h}\,Y^{-2}\,N_h^2\,({}_j\bar{Y}_h\,{}_iP + {}_i\bar{Y}_h\,{}_jQ)^2\,. \qquad [53]$$

In this equation [53] the term

$$\left(\frac{{}_iN_h}{N_h}\right)\left(\frac{{}_jN_h}{N_h}\right)\frac{1}{n_h}$$

is the binomial variance of ${}_in_h/n_h$ and the term inside the brackets $\{\}$, is the square of the expected value of the coefficient of ${}_in_h/n_h$ in the expansion of the estimator $\hat{{}_iY}/\hat{Y}$.

For the computation of an estimate of variance we, again, rely on standard formulas for ratio estimators. The standard estimate (not necessarily unbiased) of the variance given by [51] is computed from

$$\hat{\mathrm{var}}\,({}_i\hat{p}) = \hat{Y}^{-2}\,v\,({}_iy_i - {}_i\hat{p}\,y_i) \qquad [54]$$

where the sample variance function is given by [3].

If we spell out this general formula in the particular case of a stratified sample we obtain

$$\mathrm{var}\,{}_i\hat{p} = \hat{Y}^{-2}\sum_h \frac{N_h^2}{n_h\,(n_h - 1)} \sum_{i=1}^{n_h} (d_{hi} - \bar{d}_h)^2 \qquad [55]$$

where $d_{hi} = {}_iv_{hi} - {}_i\hat{p}\,y_{hi}$ and $\bar{d}_h$ is their stratum sample means and $\hat{Y}$ is the estimate of the population total i.e. $\hat{Y} = \sum_h N_h\,\bar{y}_h$ and ${}_i\hat{p} = \hat{{}_iY}/\hat{Y}$ the estimate of the proportion. An alternative

formula which does not require the computation of the d_{hi} can be obtained, after some algebra as follows.

$$\operatorname{var}{}_j\hat{p} = \hat{Y}^{-2}\sum_h \frac{N_h^2}{n_h(n_h-1)}\left\{{}_jq^2\,{}_jT_h + {}_j\hat{p}^2\,{}_{\bar{j}}T_h + n_h^{-1}\,{}_jn_h\,{}_{\bar{j}}n_h\,({}_jq_j\bar{y}_h + {}_j\hat{p}\;{}_{\bar{j}}\bar{y}_h)^2\right\}. \qquad [56]$$

Here

${}_jn_h$ is the number of sampled units in the h^{th} stratum falling within the j^{th} domain.

${}_{\bar{j}}n_h$ is the number of sampled units in the h^{th} stratum not falling within the j^{th} domain.

${}_jy_h$ is the y-total for the h^{th} stratum of units in the j^{th} domain.

${}_{\bar{j}}y_h$ is the y-total for the h^{th} stratum of units not in the j^{th} domain.

So that

$${}_j\bar{y}_h = {}_jy_h/{}_jn_h \qquad {}_{\bar{j}}\bar{y}_h = {}_{\bar{j}}y_h/{}_{\bar{j}}n_h$$

are the corresponding means. Further

$${}_j\hat{p} = {}_j\hat{Y}/\hat{Y} \quad , \quad {}_jq = 1 - {}_j\hat{p}$$

$${}_jT_h = \sum_{i=1}^{{}_jn_h}({}_jy_{hi} - {}_j\bar{y}_{\cdot})^2$$

is the sum of squares of deviations of the sampled y-values in the h^{th} stratum and j^{th} domain from their mean and

$${}_{\bar{j}}T_h = \sum_{i=1}^{{}_{\bar{j}}n_h}({}_{\bar{j}}y_{hi} - {}_{\bar{j}}\bar{y}_h)^2$$

is the corresponding sum of squares for the y-values in the h^{th} stratum and not in the j^{th} domain.

6. *Domain means adjusted for concomitant variables.*

We now turn to the analogue to « analysis of variance and covariance » for domain means. First let us briefly recall the essentials of "analysis of covariance" as it is practiced with the adjustment of (say) two group means from experimental designs. It is usually assumed that:

(*a*) The true (or population) means for the concomitant variable x are identical for the two groups, so that any differences in observed x-group means are due to sampling. In popular language the failure of this condition to be satisfied is sometimes described by the warning that "Analysis of Covariance should not be misused to correct away real treatment differences in the x-means."

(*b*) The true (or population) regression lines for the two groups are parallel.

Condition (*b*) is, of course, not essential, but in the situations in which it is satisfied (or approximately satisfied) it is almost universally invoked and simplifies the analysis. If condition (*a*) is not satisfied a meaningful generalization of analysis of covariance can still be employed if the two population group x-means ${}_j\bar{X}$ are known, and estimates of the y-group means can be obtained as the ordinates of the respective estimated group regression lines evaluated at the respective abscissa values ${}_j\bar{X}$. Now with situations as are usually encountered in "analytic studies" of sample survey data one will hardly ever be able to assume identity (or even approximate identity) of the x-domain means ${}_j\bar{X}$; on the other hand, there are situations when these means are known and we shall therefore here attempt to develop an analogue to analysis of covariance in this case.

We therefore deal with the following situation: A sample survey provides paired data $y_i\,x_i$ for a sample of n units sampled from a population of N units. After sampling the n units are classified into k domains $j = 1, 2, \ldots, k$ for which the population means ${}_j\bar{X}$ are known. It is now required to estimate the (unknown) domain y-means ${}_j\bar{Y}$ utilizing the ${}_j\bar{X}$.

The question then arises as to which estimator of ${}_j\bar{Y}$ should be used. The regression theory in classical analysis of covariance arises as the maximum likelihood estimation and results from the assumption of a linear model. Although the validity of such a model, even for very large finite populations, will often be in doubt, the use of regression estimation may still result in a gain of precision. Nevertheless we shall here *not* employ regression estimators. The reason for this is *not* that we consider regression theory inappropriate, but that this theory for finite populations requires considerable development before it can be applied in the present situation. On the other hand, ratio estimators are easily adapted to the estimation of domain means but in using these we should stress their well known limitation, namely, that they are likely to be effective only if the y and/or x scales can be so chosen that the population regression will intersect near the origin.

If a combined ratio estimator is used to estimate ${}_j\bar{Y}$ with the help of the known x-domain mean ${}_j\bar{X}$ the theory is almost identical with that developed for the ratio estimation of Section 3, the only essential difference is that the concomitant variable x will now take the place of the "count variable" ${}_ju_i$ of [26]. Accordingly we introduce the variate

$${}_jx_i = \begin{cases} x_i \text{ if } i^{th} \text{ unit is in } j^{th} \text{ domain} \\ 0 \quad \text{otherwise} \end{cases} \qquad [57]$$

and obtain the ratio estimator of ${}_j\bar{Y}$ in the form:

(*a*) The combined ratio estimator of the domain mean ${}_j\bar{Y}$:

$$\tilde{{}_jy} = {}_j\bar{X}\,({}_j\hat{Y}/{}_j\hat{X}) \qquad [58]$$

where the ${}_j\bar{X}$ are the fixed and known domain means of x, ${}_j\hat{Y}$ is defined by [15] and ${}_j\hat{X}$ by analogy.

Employing standard results from the ratio estimator theory on similar lines as in section 3 we further obtain

(*b*) The approximate variance formula of $\tilde{{}_jy}$

$$\operatorname{Var}(\tilde{{}_jy}) = {}_jN^{-2}\,V({}_jy_i - {}_j\bar{Y}\,{}_jx_i/{}_j\bar{X}) \qquad [59]$$

(*c*) The approximate estimate of the variance of $\tilde{{}_jy}$

$$\operatorname{var}(\tilde{{}_jy}) = {}_j\hat{U}^{-2}\,\mathrm{v}({}_jy_i - {}_j\hat{Y}\,{}_jx_i/{}_j\hat{X}) \qquad (60)$$

where the multivariable functions V and v are defined by [2] and [3] the variates ${}_jy_i$ and ${}_jx_i$ by [14] and [57], ${}_j\hat{Y}$ and ${}_j\hat{U}$ by [15] and [27] and ${}_j\hat{X}$ by analogy.

Formulas [59] and [60] show the familiar result that the variance of $\tilde{{}_jy}$ depends on the residuals ${}_jy_i - ({}_j\bar{Y}/{}_j\bar{X})\,{}_jx_i$ of a "regression" with slope ${}_j\bar{Y}/{}_j\bar{X}$ passing through the origin. For the majority of units in the population ${}_jy_i = {}_jx_i = 0$ and no contribution is made to the residuals; for the units in the j^{th} domain the above residuals will be small only if the y, x data of this domain alone satisfy the usual conditions required for the effectiveness of ratio estimators, namely that the y, x correlation should be high and that the y, x regression should intersect near the origin. If the latter condition is not satisfied little is gained by using regression estimators. For if the y, x data in the j^{th} domain have a regression which does not intersect near

the origin then the addition of the large number of units (outside the j^{th} domain) for which ${}_jy_i = {}_jx_i = 0$ would generate large residuals from the best regression fitted to all N pairs of ${}_jy_i$, ${}_jx_i$.

The spelling out of formula (60) follows on the same lines as in section 3. For example 1 (random sampling) we find in analogy to (34)

$$\text{var}\,(\hat{{}_j\bar{y}}) \doteq \left(1 - \frac{n}{N}\right) \frac{n}{{}_jn^2\,(n-1)} \sum_{i=1}^{{}_jn} \left({}_jy_i - \frac{{}_j\bar{y}}{{}_j\bar{x}}\,{}_jx_i\right)^2 \qquad [61]$$

In a similar manner the formulas for the variance of a difference of two ratio estimators can be obtained following on the lines of the arguments in section 4. We reach the general formula

$$\text{var}\,(\hat{{}_1\bar{y}} - \hat{{}_2\bar{y}}) \doteq \text{v}\left(\frac{1}{\hat{{}_1U}}\left({}_1y_i - \frac{\hat{{}_1\bar{Y}}}{\hat{{}_1\bar{X}}}\,{}_1x_i\right) - \frac{1}{\hat{{}_2U}}\left({}_2y_i - \frac{\hat{{}_2\bar{Y}}}{\hat{{}_2\bar{X}}}\,{}_2x_i\right)\right) \qquad [62]$$

which, in the case of example 1 (simple sampling) spells out as

$$\text{var}\,(\hat{{}_1\bar{y}} - \hat{{}_2\bar{y}}) \doteq \left(1 - \frac{n}{N}\right)\left(\frac{n}{n-1}\right)\left\{\frac{{}_1\Sigma}{{}_1n^2} + \frac{{}_2\Sigma}{{}_2n^2}\right\} \qquad [63]$$

where

$${}_j\Sigma = \sum_{i=1}^{{}_jn}\left({}_jy_i - \frac{{}_j\bar{y}}{{}_j\bar{x}}\,{}_jx_i\right)^2$$

7. Unbiased ratio estimators for domain means and exact variance formulas.

In sections 3-6 we have freely used analogies to the well known combined ratio estimators and their approximate variance formulas. For a discussion of the magnitude of the bias and the precision of the approximate variance formulas we must refer to the literature. Since there may well be cases in which these approximations are too inaccurate we wish to put here on record a method which avoids these disadvantages under certain circumstances.

The situations in which our exact formulas will apply are characterized by the following conditions:

(*a*) The number of units in the j^{th} domain, ${}_jN$, must be known.

(*b*) Primary sampling units (or, if the design is single stage, the sampling units) must have been drawn *with* replacements.

Two remarks concerning these conditions may be pertinent: Concerning (*b*), although there are many surveys in which primaries are actually drawn with replacement, this condition becomes of little importance when the sampling fraction of primaries is small as then there is practically no difference between drawing with — and drawing without — replacement. Small primary sampling fractions are very common in survey designs.

Concerning the very important condition (*a*), if the ${}_jN$ are known it is of course possible to obtain unbiased estimates of ${}_j\bar{Y}$ by dividing the unbiased estimates of the ${}_jY$ (developed in the preceding sections) by the known ${}_jN$. It is, however, well known from experience with such estimates that for most survey designs and populations they would have large variances. Confirmation of this may be sought by comparing for the examples quoted in the preceding sections ${}_jN^{-2}\,\text{Var}\,\hat{{}_jY}$ with $\text{Var}\,(\hat{{}_j\bar{y}})$. Roughly speaking the former depends on the variation of ${}_jy$ — totals whilst the latter depends on the variation of ${}_j\bar{y}$ — means.

Turning, therefore, again to the device of ratio estimation we will describe first the unbiased ratio estimators for population means translating them into estimators of domain means later.

Let us consider a survey design consisting of L strata ($h = 1, 2, \ldots, L$) with population proportions P_h and let us assume that n_h primaries $t = 1, 2, \ldots, n_h$ have been drawn with replacement from the h^{th} stratum. Denote by y_{hti} a character [+] attached to the i^{th} unit in the t^{th} primary of the h^{th} stratum and by $\hat{y}_{ht}$ the unbiased estimate of $\bar{Y}_h$ from the t^{th} primary only. The functional form of $\hat{y}_{ht}$ will depend on the design details. Denote by

$$\bar{y}_h = n_h^{-1} \sum_{t=1}^{n_h} \hat{y}_{ht} \qquad [54]$$

and by

$$\bar{y}_{st} = \sum_{h=1}^{L} P_h\,\bar{y}_h \qquad [55]$$

the stratum means of the $\hat{y}_{ht}$ and the unbiased estimator of $\bar{Y}$.

Denote further by $\hat{x}_{ht}$, $\bar{x}_h$ and $\bar{x}_{st}$ the corresponding means for a second characteristic x for which the population mean $\bar{X}$ is known and for which we assume that either $x_{hti} \geq 0$ or $x_{hti} = y_{hti} = 0$.

Introduce now the ratio variates

$$\hat{r}_{ht} = \begin{cases} \hat{y}_{ht}/\hat{x}_{ht} & \text{if} \quad \hat{x}_{ht} > 0 \\ r^* & \text{if} \quad \hat{x}_{ht} = \hat{y}_{ht} = 0 \end{cases} \qquad [56]$$

where r^* is a constant conveniently chosen as discussed below. Finally introduce for these ratios the means $\bar{r}_h$ and $\bar{r}_{st}$ as above. The survey design can now be thought of as generating in each stratum an infinite joint population of $\hat{y}_{ht}$, $\hat{x}_{ht}$, $\hat{r}_{ht}$ by the infinitely repeatable process of

(*a*) drawing a single primary in accordance with the design,

(*b*) drawing from this primary the prescribed number of secondaries, tertiaries . . . in accordance with the design,

(*c*) computing $\hat{y}_{ht}$ $\hat{x}_{ht}$ $\hat{r}_{ht}$ for the drawn sample,

(*d*) replacing all units.

Following the lines of Hartley-Ross (1954) and Goodman-Hartley (1956) we can now construct the following estimators:

The unbiased ratio estimate of $\bar{Y}/\bar{X}$

$$r' = \bar{r}_{st} + \bar{X}^{-1}\,\hat{c} \qquad [67]$$

The unbiased estimate of $\bar{Y}$:

$$y' = \bar{X}\,\bar{r}_{st} + \hat{c} \qquad [68]$$

where

$$\hat{c} = (\bar{y}_{st} - \bar{x}_{st}\,\bar{r}_{st}) + \sum_h P_h^2\,(n_h - 1)^{-1}\left\{\bar{y}_h - \bar{x}_h\,\bar{r}_h\right\} \qquad [69]$$

is an unbiased estimate, computed from the stratified sample, of the population covariance of $\hat{x}_{ht}$ and $\hat{r}_{ht}$ defined by

$$C = \sum_h P_h\,E\,(\hat{x}_{ht} - \bar{X})\,(\hat{r}_{ht} - \bar{R}) \qquad [70]$$

in which E refers to the infinite populations described above.

The estimates r' and y' as given by [67] and [68] are unbiased no matter what value was chosen for the constant r^*. The choice of r^* will, however, affect the variance of these estimates. In the practical situations in which these estimators will be applied the chance of $\hat{x}_{ht} = \hat{y}_{ht} = 0$ to arise will be very small and hence any uncertainty in the choice of r^* will have a small effect. The best choice is to take r^* equal to the best estimate

[+] The notation applies directly to two stage designs, the index i denoting the secondary unit. In three or higher stage designs i must be replaced by a multiple subscript.

of $\bar{R}$ available in advance of sampling and must not be altered subsequently.

In order to apply these results to the estimation of the j^{th} domain mean we must use for the numerator variable

$$ {}_jy_{hti} = \begin{cases} y_{hti} \text{ if the } hti \text{ unit is in the } j^{th} \text{ domain} \\ 0 \text{ otherwise} \end{cases} $$

and for the denominator the count variable

$$ {}_ju_{hti} = \begin{cases} 1 \text{ if the } hti \text{ unit is in the } j^{th} \text{ domain} \\ 0 \text{ otherwise} \end{cases} $$

and substitute these in [66], [67] and [69]. It will be seen that for all selfweighting designs the leading term of [57] $\bar{r}_{st}$ will be of the form

$$ \bar{r}_{st} = \sum_h P_h n_h^{-1} \sum_t {}_j\bar{y}_{ht} $$

where ${}_j\bar{y}_{ht}$ is defined, as before, as the y- mean of the units (if any) in the j^{th} domain of the t^{th} primary of the h^{th} stratum and by r^* if there are no such units. The condition mentioned above, that the chance of ${}_j\hat{y}_{ht} = {}_j\hat{u}_{ht} = 0$ should be small means that the present method of estimation should only be used if there is a reasonable chance that all domains are represented in the sample from each primary.

The unbiased estimation of ${}_j\bar{Y}$ utilizing the knowledge of the domain mean ${}_j\bar{X}$ of a concomitant variable will be based on [68] when it should be noted that y' will estimate ${}_jN_j\bar{Y}/N$, that $\bar{X}$ will be ${}_jN\,{}_j\bar{X}/N$ so that the estimator of ${}_j\bar{Y}$ will be

$$ \frac{N}{{}_jN} y' = {}_j\bar{X}\,\bar{r}_{st} + \frac{N}{{}_jN}\hat{c} $$

where $\hat{c}$ is computed from [69] using the variates ${}_jy_{hti}$, ${}_jx_{hti}$ and the ratios ${}_j\hat{r}_{ht}$. In this case, therefore, both ${}_j\bar{X}$ and ${}_jN$ must be known.

The exact variance formulas for both r' and y' have recently been derived by Goodman-Hartley in the special case of simple sampling from one stratum ($L = 1$). Their results indicate that the variance of y' is of a similar order of magnitude (sometimes larger, sometimes smaller) than that of the combined ratio estimator. For the stratified case results are not as yet available.

8. *Inferences based on the variance formulas.*

Two questions concerning inferences arise:

The variance formulas of the preceding sections are based on finite population sampling and, in particular, have the property to become zero when sampling is 100 %. The obvious logic of this property is that when all units of the population are sampled all means of all domains in such a population are known without error so that all non-zero contrasts between domain means are "significant". It must not be forgotten, however, that such inferences can only apply to the particular finite population under investigation and are not of any wider significance. For instance, in the example of Table 3, had the population of Des Moines been sampled 100 % definite statements about differences in the number of persons / household of different Income Groups could have been made "without error" but such inferences would then only apply to the City of Des Moines at the time of the survey. If data from such a survey are to be used for (say) comparing "Income Groups" in "Cities like Des Moines" in "times like the present" in a more general sense, the finite population sampling theory must not be used. Whether the survey sample can be (artificially) regarded as one drawn from a wider populaton (in space or time) is a matter requiring special investigation and the onus of such an investigation falls upon those wishing to draw such wide inferences.

A second question concerns inferences to be drawn for the domains of the particular finite population for which the survey was planned:

In the preceding sections we have given estimates of domain means, their variances and estimates of these variances. We have also given variance formulas and estimates for the differences between estimated domain means. The question arises as to how these variance estimates are to be used for inferences to be drawn from the data. Even if it is not intended to carry out "tests of significance" there would remain the question of the computation of confidence intervals for the domain means and their contrasts. This question is, of course, not confined to the present issues but is an ever present difficulty in the theory of sample surveys. The customary procedure here is to employ normal theory approximations and to appeal to the respective central limit theorems for finite populations and to the relatively large sample sizes which are available in these situations. The results which are available on these problems are very restricted and require development and it is clearly not possible in this context to dwell upon this issue of much wider impact.

References

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

Goodman, L., Hartley, H. O., (1956). (Unpublished manuscript).

Hartley, H. O., Ross, A. — *Unbiased Ratio Estimators*, 1954. Nature, 174, 270.

Kish, L. and Hess, Irene, (1955). — *On Variances of Ratios and Their Differences in Multistage Sampling*. (Paper presented at the New York Meeting of A.S.A. Dec. 1955).

Stephan, F. F., (1941). — *Stratification in Representative Sampling*. J. of Marketing (1941), 38.

Yates, F., 1949 (2nd edition 1953). — *Sampling Methods for Censuses*. Surveys, London, Charles Griffin & Co.

ACKNOWLEDGMENT

This paper has appeared as a contribution to a volume published by the Institute of Statistics in honour of Professor Corrado Gini, upon the occasion of his retirement from the Chair of Statistics at the University of Rome. It is reproduced here through the kind permission of the Institute of Statistics and of its Director, Professor Vittorio Castellano. Professor Hartley presented a summary of this paper at the Annual Meeting.