STRATIFICATION IN SURVEYS ON FRUIT CROPS

BY

R. SETHUMADHAVI[1] AND B. V. SUKHATME[2]

I. A. S. R. I., New Delhi

(Received ; March, 1973)

An investigation of stratification is reported for a universe with high positive skewness. The method of constructing strata, the sample allocation, the number of strata and the optimum sample size are considered. Comparisons are made among four types of allocation in combination with the corresponding optimum stratifications. Gains from stratification are examined for the two estimation variables.

INTRODUCTION

On the initiative of the Ministry of Food and Agriculture the Institute of Agricultural Research Statistics initiated a series of sampling investigations on a co-ordinated basis on important fruit crops, such as Mango, Guava, Banana, Orange, Lime etc.

One such investigation on temperature fruit crops was carried out in Mahasu District of Himachal Pradesh during the year 1965-66. The object of this paper is to examine critically the data collected in this investigation with a view to study the various aspects involved in stratification as :

1. Construction of strata ;
2. type of sample allocation ;
3. number of strata ;
4. expected gains from stratification ; and
5. Determination of sample size.

The problems mentioned above are not new and have been considered by several authors. If information concerning the character under study or some correlated character is available from the past surveys, rules have been given by Dalenius for determining the optimum strata boundaries. The performance of these rules depends

[1] Statistical Officer, Dena Bank, Fort. Bombay.

[2] At present Professor of Statistics in Iowa State University, Ames, Iowa, U. S. A.

heavily on the distribution of the character under study. Several studies have been carried out in this direction, such as Cochran [1] Sethi [8]. These authors have considered the optimum construction of strata when the population under study is skew. However, the results obtained are not conclusive. As will be shown in the next section, the distributions considered in this paper are highly skewed and it is hoped that this investigation will throw further light on the usefulness and validity of the rules as also other problems involved in the construction of strata.

The Population Under Study

Mahasu is one of the most important centres in Himachal Pradesh growing fruits in well-cared orchards. Ten Tehsils *viz.*, Kumarsain, Rampur, Chopal, Rohru, Tubbal, Kotkhai, Theog, Kasumpti, Arki and Suni in the district were included in the survey. The information concerning area under fruits was available for all the villages in these ten tehsils, the number of villages having area under fruits being 703.

For the survey mentioned earlier, a random sample of 130 villages was selected and informations regarding area under fruits and number of fruits trees was collected for the sampled villages. We shall utilise this sample for further studies on stratification. Table 1 gives the frequency distribution of the 130 villages in the sample according to area under fruits. Figures in brackets denote the corresponding population values. The distribution has a high positive skewness and a long positive tail.

TABLE 1

Frequency Distribution of the sample villages according to area under fruits.

Area under Fruits (in Acres)	*No. of villages*
0— 4	65 (394)
4— 8	24 (132)
8—12	16 (73)
12—16	9 (35)
16—20	3 (14)
20—24	3 (16)
24—28	1 (13)
28—32	6 (11)
32—36	0 (3)
36—40	0 (3)
40—44	0 (0)
44—48	1 (1)
48—52	0 (0)
More than or equal to 52	2 (8)

CONSTRUCTION OF STRATA

The problem of constitution of strata when information is available concerning the frequency function of the variable under study has been consi dered by Dalenius, Gurney and others. Several approximate methods have been proposed in the literature because solving sets of equations to establish boundary points is impractical for general use. Five such rules to be compared here are as follows :

1. Dalenius and *Hodges' rule* to construct equal intervals on the cumulattive of

$$\sqrt{f(y)}$$

(where $f(y)$ is the frequency function and y the variable under study).

2. Ekman's rule which equalizes the product of the frequency within the stratum and the width of the stratum—that is

$$W_h(Y_h - Y_{h-1})$$

is equal to constant.

3. Fqualization of the aggragate outputs, or the product of the stratum weight W_h and the stratum mean μ made constant, suggested by Mahalanobis (1952) and Hansen, Hurvitz and Madow (1953).

4. Durbin's rule to construct equal intervals on the cumulative of

$$\tfrac{1}{2}[r(y) + f(y)]$$

[where $r(y) = \dfrac{F(y_L)}{y_L - y_o}$; $y_0, y_1, y_2 \ldots\ldots y_L$

are the strata boundaries and $F(y)$ is the cumulative of $f(y)$].

5. Dalenius' rule for proportional allocation (for which Sethi gave an iterative method).

With reference to the problem considered in this paper and described in Section 1 no information is available concerning the production of fruits. The only other variables which are likely to be correlated with the total production and in respect of which

some amount of information was available are the number of fruit trees and the area under fresh fruits. Of these, the information in respect of area under fresh fruits was available for all the 703 villages constituting the population under survey. As such we shall make use of this information in the construction of strata.

In the course of this study we shall restrict ourselves to estimate two characters only *viz.*, the total number of trees and total area under fruits. Using the information on area under fruits, the optimum points of stratification were determined by using the five approximate methods described earlier, for number of strata varying from 2 to 4. These are presented in table-2. The highly skew population discussed in this paper resembles chi-square distribution with one degree of freedom. Sethi has prepared tables for optimum stratification points for some standard distributions. The columns 6 to 9 of table-2 are based on this. It will be seen from Table-2, that except Sethi's OPS (optimum points of stratification) for optimum allocation for χ_2^2 and Dalenius and Hodges' rule no two methods for the construction of strata lead to similar stratification.

TABLE 2

Value of distribution function at stratification points for various methods of stratification

No. of strata	*Daleñius Hodges*	*Ekman*	*Equalisation of strata total*	*Durbin*	*OPS for proportional allocation*		*Optimum allocation*	
					χ_1^2	χ_2^2	χ_1^2	χ_2^2
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
2	.7482	.9018	.8805	.8051	.8620	.7981	.8077	.7404
3	.5605	.7212	.7852	.5605	.7458	.6321	.6572	.5507
	.8806	.9616	.9445	.9218	.9516	.9257	.9112	.8775
4	.5605	.6273	.7212	.5605	.6572	.5276	.5614	.4230
	.7483	.8805	.8805	.8052	.8787	.8173	.8077	.7404
	.9218	.9829	.9358	.9617	.9760	.9612	.9488	.9257

Again it is seen that the methods based on equalization of strata and Ekman's rule lead to wider initial strata which is not the case with the Dalenius and Hodges' rule closely followed by Durbin's rule.

To examine the efficiency of the different methods of construction of strata from the points of view of precision, ratios of the variance of the estimate under stratified sampling to that under simple random sampling were obtained for the number of strata varying from two to four (the results are not presented he re). While calculating the variances under stratified sampling it is assumed that the distribution of the sample among the different strata is made according to optimum allocation.

The methods of stratification give almost identical results. However, Ekman's method excels but for two strata. Equalisation of trata totals is next in performance, followed closely by the cum root of (f) rule, Durbin's rule and Sethi's iterative method (for proportional allocation). The cum root of (f) rule and Durbin's rule yield almost equal precision. Dalenius' rule for proportional allocation is the least efficient, as is to be expected. When strata are formed by the method of Dalenius and Hodges, the contribution to the variance from the several strata is almost the same ; furthermore, for a given number of strata, L, each stratum approximately accounts for $1/L$ of the total variance. Stratification by other rules leads to departure from this pattern of equal contribution to variance from different strata. They lead to the construction of top strata that are too wide making large contributions to the total variance. Eventhough Ekman's method gives the least variance it is more laborious especially when the number of strata is large. On the other hand eventhough Dalenius and Hodges' rule gives somewhat higher variances, it can easily be adopted in practice. We shall, therefore, examine the gain in efficiency by Ekman's rule over Dalenius and Hodges' rule that we can find whether Ekman's rule is worth following in spite of the large labour involved in that.

Allocation of Sample to the Strata

For the purpose of our study, we consider the following allocations :

(i) Neyman allocation : $n=\frac{np_is_i}{\Sigma p_is_i}$ where s_i is the standard deviation of the i-th stratum for number of trees.

(*ii*) Neyman allocation : $n_i = \frac{np_i s_i}{\Sigma p_i s_i}$ where s_i is the standard deviation of the i-th stratum for area under fruits.

(*iii*) Proportional allocation : $n_i = \frac{nN_i}{N}$ where N_i is the total number of villages in the i-th stratum.

(*iv*) Proportional allocation : $n_i = \frac{nA_i}{A}$ where A_i is the area under fruits in the i-th stratum.

As we do not have the knowledge of the population S_i^s we estimate them from a sample of 130 villages and these estimated values, s_i^s , are utilised to allocate the sample to the different strata.

We shall first consider the case when the character to be estimated is the total area under fruits. Table-3 gives variance ratios (denoted by Var_L/Var_1) for different types of allocations when the strata are constructed by Ekman's rule and Dalenius and Hodges' rule.

Table 3 shows us that the two allocations *viz.*, optimum and proportional based on area under fruits yield equally good results and give maximum precision when the strata are constructed through Dalenius and Hodges' rule. When the strata are constructed by Ekman's rule the table clearly shows that optimum allocation based on area under fruits is certainly more efficient than proportional allocation based on area under fruits. However, the gain in efficiency is only marginal. It is also to be noted that with increasing number of strata the differences between the variance ratios for various allocations disappear to a great extent and considerable reduction in variance is achieved through stratified sampling.

We shall now consider the problem of estimating the total number of trees and investigate the efficiency of the four allocations considered above. Table-4 gives the ratios of variances under stratified sampling to that under simple random sampling for these four allocations when the strata are constructed through Ekman's rule and Dalenius and Hodges' rule. We see that the two allocations *viz.* optimum and proportional based on area under fruits are inefficient as the variances are higher than those of simple random sampling except for optimum allocation based on area under fruits when the strata are formed by Ekman's rule. In this case the variances are no doubt reduced but the reduction is hardly significant. Of the remaining two allocations *viz.* optimum allocation

TABLE 3

VAR_L/VAR_I for estimated total area under fruits, by number of strata, method of constructing strata, sample size and type of sample allocation

No. of strata	Sample size	STRATIFICATION RULE: $W_h(Y_h - Y_{h-1}) = Constant$ (Ekman) — Allocation				Equal intervals on cum $\sqrt{f}$ (Dalenius and Hodges) — Allocation			
		Optimum according to No. of trees	Optimum according to area under fruits	Proportional to No. of villages	Proportional to area under fruits	Optimum according to No. of trees	Optimum according to area under fruits	Proportional to No. of villages	Proportional to area under fruits
2	50	.208	.208	.264	.250	.232	.162	.376	.162
	70	.207	.207	.265	.250	.228	.156	.377	.156
	100	.205	.204	.266	.250	.221	.145	.378	.145
	120	.202	.202	.265	.249	.215	.137	.378	.137
	150	.198	.198	.264	.247	.207	.124	.377	.123
3	50	.086	.061	.109	.069	.105	.072	.203	.072
	70	.086	.059	.109	.068	.102	.068	.204	.068
	100	.085	.057	.109	.066	.098	.062	.205	.062
	120	.084	.055	.109	.062	.094	.057	.204	.057
	150	.083	.052	.109	.057	.087	.049	.204	.048
4	50	.070	.043	.081	.050	.058	.036	.122	.038
	70	.070	.042	.082	.049	.056	.033	.122	.036
	100	.070	.041	.082	.046	.053	.029	.122	.031
	120	.069	.039	.082	.044	.051	.025	.122	.028
	150	.068	.035	.082	.041	.047	.014	.122	.024
5	50	.038		.048	.034	.040		.087	.026
	70	.038		.048	.033	.039		.088	.024
	100	.037		.048	.031	.037		.088	.021
	120	.037		.048	.030	.035		.088	.019
	150	.036		.048	.027	.032		.087	.016

TABLE 4

VAR_L/VAR_I for estimated total number of treees, by number of starta, method of constructing strata, sample size and type of sample allocation

STRATIFICATION RULE

No. of strata	Sample size	$W_h(Y_h - Y_{h-1}) = Constant$ (Ekman)				Equal interval on cum $\sqrt{f}$ (Dalcnius and Hodges)			
		Allocation				Allocation			
		Optimum according to No. of trees	Optimum according to area under fruits	Proportional to No. of villages	Proportional to area under fruits	Optimum according to No. of trees	Optimum according to area under fruits	Proportional to No. of villages	Proportional to area under fruits
2	50	.724	.726	.835	.914	.767	1.260	.805	1.251
	70	.720	.722	.835	.917	.765	1.275	.805	1.265
	100	.714	.716	.835	.921	.763	1.297	.805	1.287
	120	.710	.713	.835	.923	.762	1.314	.805	1.305
	150	.704	.706	.835	.929	.761	1.343	.806	1 331
3	50	.627	.862	.741	1.227	.677	1.199	.819	1.306
	70	.624	.866	.741	1.295	.673	1.211	.820	1.322
	100	.618	.873	.741	1.322	.665	1.231	.820	1.348
	120	.614	.877	.741	1.272	.660	1.244	.819	1.366
	150	.607	.883	.742	1.233	.663	1.403	.820	1.396
4	50	.599	.957	.779	1.493	.671	1.108	.814	1.369
	70	.593	.963	.780	1.517	.666	1.118	.814	1.387
	100	.584	.972	.780	1.475	.659	1.133	.814	1.407
	120	.575	.978	.781	1.470	.654	1.144	.814	1.437
	150	.566	.914	.780	1.482	.645	1.053	.814	1.381
5	50	.665		.867	1.337	.514		.730	1.185
	70	.658		.869	1.326	.507		.730	1.200
	100	.648		.869	1.302	.496		.731	1.223
	120	.640		.868	1.302	.488		.731	1.228
	150	.628		.869	1.302	.474		.730	1.148

according to number of trees and allocation proportional to villages, we find that optimum allocation always gives smaller variance than proportional allocation whatever may be the rule for stratification. Since our object is to estimate both the total number of trees as also the area under fruits it is clear from the above discussion that the optimum allocation based on number of trees will give us the maximum precision in the case of both the variables.

The Optimum Number of Strata, Gain Due to Stratification and Sample Size

We now consider the problem of determining the optimum number of strata. This can best be done by observing the reduction in variance effected by the addition of another stratum. That is, the variance from L strata is compared with the variance resulting from $L-1$ strata. As we have seen in Section 3, the optimum allocation based on number of trees is most efficient for the problem considered here. We shall therefore consider the problem of determining the optimum number of strata based on this allocation only. Table 5 gives the ratios $Var\ L/Var_{L-1}$ for both the estimation variables with optimum allocation based on number of trees. These results are given for number of strata varying from two to five when they are constructed by Ekman's rule and Dalenius and Hodges' rule. Here we see that regarding the stratification variable *viz.*, area under fruits, the formula $V_L/V_{L-1}=(L-1)^2/L^2$ is followed well when the strata are constructed by Dalenius and Hodges' rule though this is not so the case with Ekman's rule. With regard to the variable area under fruits we see that there is considerable reduction in variance with the addition of every stratum, whatever may be the stratification rule, Ekman's or Dalenius and Hodges'. Regarding the variable number of fruit trees, under Ekman's rule there is hardly any reduction in variance if we proceed after three strata whereas under Dalenius and Hodges' rule considerable reduction is seen with the addition of every stratum.

To know the amount of reduction in variance we refer to Table-6. Table-6 gives us the percentage gain due to stratification for the number of strata varying from two to five when the strata are constructed through Ekman's rule and Dalenius and Hodges' rule. The sample is distributed according to optimum allocation based on number of fruit trees. The results derived from Table-5 are clearly brought out in Table-6. We see that as far as area under fruits is concerned, the percentage gain due to stratification is very high, whatever be the sample size and mode of constructing the strata.

This is one to be expected since area under fruits is both the estimation variable as also the stratification variable. As regards the number of trees it is seen from Table-6 that there is considerable gain due to stratification and that in general it increases with increasing number of strata except when the strata are formed by Ekman's rule. The percentage gain due to stratification is maximum

TABLE 5

Var_L/Var_{L-1} for estimated total number of trees and estimated total area under fruits, by number of strata, stratification rule and sample size

No. of strata	*Sample size*	*STRATIFICATION RULE*			
		$W_h(Y_h - Y_{h-1})$ = Constant (Ekman)		*Equal intervals on cum $\sqrt{f}$ (Dalenius and Hodges)*	
		Estimation variable		*Estimation variable*	
		Area under fruits	*No. of fruit trees*	*Area under fruits*	*No. of fruits trees*
2	50	.208	.724	.232	.767
	70	.207	.720	.228	.765
	100	.205	.714	.221	.763
	120	.202	.710	.215	.762
	150	.198	.704	.207	.761
3	50	.415	.867	.453	.883
	70	.415	.866	.449	.879
	100	.416	.865	.441	.872
	120	.416	.864	.435	.867
	150	.417	.863	.423	.872
4	50	.812	.955	.551	.991
	70	.815	.961	.548	.991
	100	.819	.954	.543	.991
	120	.816	.937	.539	.990
	150	.826	.932	.531	.973
5	50	.539	1.110	.694	.766
	70	.537	1.109	.694	.761
	100	.535	1.111	.691	.753
	120	.536	1.114	.688	.746
	150	.530	1.111	.683	.735

and is of the order of 100 per cent when there are five strata and these are constructed by Dalenius and Hodges' rule. Taking this into account as also the percentage standard errors both in the case of area under fruits and the total number of trees it seems preferable to have about five strata and to use Dalenius and Hodges' rule for their construction.

TABLE 6

Percentage gain due to stratification for estimated total number of trees and estimated total area under fruits, by number of strata, stratification rule and sample size

No. of strata	*Sample size*	*STRATIFICATION RULE*			
		$W_h(y_h - y_{h-1}) = Constant$ (*Ekman*)		*Equal intervals on cum* $\sqrt{f}$ (*Dalenius and Hodges*)	
		Estimation Variable		*Estimation Variable*	
		Area under fruits	*No. of fruit trees*	*Area under fruits*	*No. of fruit trees*
2	50	380.8	38.1	331.0	30.4
	70	383.1	39.9	388.6	30.7
	100	390.2	40.1	352.5	31.1
	120	395.0	40.8	365.1	31.2
	150	405.1	42.0	383.1	31.4
3	50	1062.8	59.5	852.4	47.7
	70	1062.8	60.3	880.6	48.6
	100	1076.5	61.8	920.4	50.4
	120	1090.5	62.9	963.8	51.5
	150	1104.8	64.7	1049.4	50.8
4	50	1328.6	66.9	1624.1	49.0
	70	1328.6	68.8	1685.7	50.2
	100	1328.6	71.2	1786.8	51.7
	120	1343.9	73.9	1860.8	52.9
	150	1370.6	76.8	2027.7	55.0
5	50	2531.6	50.4	2400.0	94.6
	70	2531.6	52.0	2464.1	97.2
	100	2602.7	54.3	2602.7	101.6
	120	2602.7	56.3	2757.1	104.9
	150	2677.8	59.2	3025.3	111.0

Having investigated the best way of determining strata boundaries, the allocation of sample size to the different strata, the optimum number of strata and the percentage gain due to stratification, we shall now consider the determination of sample size. Obviously the sample size should be such that we can obtain maximum precision both in respect of area under fruits and total number of trees.

Toble 7 gives the standard errors for varying sample sizes for optimum allocation based on number of fruit trees when there are given strata constructed by Dalenius and Hodges' rule. It is seen that in the case of both the variables *viz.*, area under fruits and number of fruit trees the percentage standard error (ratio of the standard error of the estimate to the sample mean expressed in percentage) goes on decreasing as the sample size goes on increasing, the percentages S.E. being of the order of 6 per cent in the case of number of fruit trees and 2 per cent in the case of area under fruits for sample size 180. We should take the smallest sample size which satisfies the prescribed upper bounds of the percentage S.E.'s

TABLE 7

Percentage S.E. for estimated total number of trees and for estimated total area under fruits for optimum allocation based on number of fruit trees for various sample sizes.

Sample sizes	*Area under fruits*	*No. of fruit trees*
50	4.22	13.60
70	3.45	11.25
100	2.73	9.08
120	2.39	8.09
150	1.99	6.95
180	1.68	6.08

Summary and Conclusions

This paper presents a critical study of the various aspects involved in stratification with reference to data collected in the sample survey conducted on temperate fruit crops in Mahasu District of Himachal Pradesh during the year 1965-66. The problems considered (*i*) construction ofstrata, (*ii*) type of sample allocation, (*iii*) the

number of strata, (*iv*) the expected gains from stratification and (*v*) determination of optimum sample size.

Area under fruits is chosen as the stratification variable as information regarding area is available for the entire population and it has a correlation of .58 with the other estimation variable *viz.* number of trees. Five methods of construction are examined for constructing the strata boundaries with optimum allocation based on number of fruit trees. The rules are (1) equal intervals on cum $\sqrt{f}$ (2) Wh $(y_h - Y_{h-1})$=constant, (3) Equalisation of strata totals (4) Durbin's rule and (5) Dalenius' rule for proportional allocation. The first four rules gave nearly identical results. However, beyond two strata, Ekman's rule excelled and equalisation of strata totals, cum $\sqrt{f}$ rule, Durbin's rule and Sethi's interative method followed in that order of performance. As Ekman's rule is more laborious in the construction of strata especially when the number of strata is large, we have considered cum $\sqrt{f}$ rule also along with Ekman's rule for further studies, cum $\sqrt{f}$ rule being quite simple in the matter of construction.

Four types of sample allocation are considered; optimum based on number of fruit trees, optimum based on area under fruits, proportional to number of villages and proportional to area under fruits. Optimum allocation based on number of trees gave the highest precision. Allocation proportional to number of villages showed only a little gain in efficiency. Optimum and proportional allocations based on area under fruits were found to be inefficient.

It was observed that gains from stratification by cum $\sqrt{f}$ rule with optimum stratification based on number of trees followed approximately the relationship Var L/Var $(L-1)=\left(\frac{L-1}{L}\right)^2$ when an unbiased estimate of the stratification variable was considered. The relationship weakened and the stratification gains diminished for the other estimation variable even though the correlation between the stratification and estimation variables was .58.

It was concluded that the population might be divided into five strata with a sample of size 200 allocated to the different strata according to optimum allocation based on number of trees so as to obtain fair degree of precision for both the estimation variables *viz*, area under fruits and number of fruit trees.

REFERENCES

[1] Cochran, W.G. (1961) : "Comparison of methods for determining stratum boundaries". Bulletin of International Statistical Institute, 38, 345-358.

[2] Dalanius, T. and Gurney, M. (1951) : "The problem of optimum stratification" II Skandinavik Aktuarietidskrift, 34, 133-148.

[3] Dalenius, T. and Hodges J.L. (1957) : "The choice of stratification points" Skandinavik Aktuarietidskrift, 40, 198-203.

[4] Dalenius, T. and Hodges, J.L. (1959) : "Minimum variance stratification". Jour. of American Statistical Association, 54, 88-101.

[5] Durbin, J. (1959) : "Review of sampling in Sweden". Journal of the Royal Statistical Society (A). 246-248.

[6] Ekman, G. (1959) : "An approximation useful in univariate stratification". The anals of Mathematical Statistics, 30, 219-229.

[7] Hansen, M.H., Hurvitz, W.N. and Madow W.G. (1953) : "Sample survey methods and theory". Vol. I and Vol. II. John Wiley and Sons Ltd., New York.

[8] Sethi, V.K. (1963) : "A note on optimum stratification of populations for estimating the population means". Australian Journal of Statistics, 5, 20-33.

[9] Hess, Irene, Sethi, V.K. and Balkrishnan, T.R. (1966) : "Stratiification a practical investigation" Journal of the American Statistical Association, 61, 70-90.

[10] Mahalanobis, P.C. (1952): "Some aspects of the design of sample surveys". Sankhya, , 1-7.