Arbeidsnotater

STATISTISK SENTRALBYRÅ

Dronningensgt. 16, Oslo-Dep., Oslo 1. Tlf. 41 38 20

WORKING PAPERS FROM THE CENTRAL BUREAU OF STATISTICS OF NORWAY

IO 76/25 3. august 1976

# ON THE EFFECT OF STRATIFICATION WHEN TWO STRATIFYING VARIABLES ARE USED

By

Ib Thomsen*)

## CONTENTS

Page

*) This paper will appear in Journal of the American Statistical Association.

Not for further publication. This is a working paper and its content must not be quoted without specific permission in each case. The views expressed in this paper are not necessarily those of the Central Bureau of Statistics.

*Ikke for offentliggjøring. Dette notat er et arbeidsdokument og kan siteres eller refereres bare etter spesiell tillatelse i hvert enkelt tilfelle. Synspunkter og konklusjoner kan ikke uten videre tas som uttrykk for Statistisk Sentralbyrås oppfatning.*

## ABSTRACT

Most of the literature on survey sampling deals with a single stratifying variable. In this paper an attempt is made to study the effect of using two stratifying variables. We present an approximation to the variance of the study variable under the assumption of a linear regression on the two stratifying variables. This approximation depends only on the number of strata, the simultaneous density of the stratifying variables, and the correlations between the study variable and each of the stratifying variables. In Sections 3 and 4 we study the case in which the stratifying variables are independent. In Section 5 the stratifying variables are assumed to have a bivariate normal distribution. The results seem to indicate that in many practical situations the gains from using two stratifying variables over one seem to be nontrivial.

## 1. INTRODUCTION

Our aim is to estimate the population mean of some quantitative characteristic Y in a finite population. The population is partitioned into a number of strata, and from each stratum a simple random sample is selected. The variance of the usual stratified mean $\bar{y}_{st}$ evidently depends on how the strata are constructed and how the sample is allocated.

The effect of stratification using one stratifying variable has received considerable attention in the sampling literature. See Cochran (1963), Dalenius (1957), (1959), Ekman (1959), Kish (1964), Serfling (1968), Sethi (1963), and Thomsen (1976). One of the conclusions from these studies is that when stratification is done on an auxiliary variable, the optimal number of strata is somewhere between 1 and 10, except in cases with extremely large correlation between the stratifying variable and the study variable.

In this paper we shall demonstrate that under some conditions one can expect a considerable reduction of the variance by using two stratifying variables, and in this case the optimal number of strata is larger than when stratifying along one variable.

When stratification is done along one variable with density f, an efficient stratification and allocation method consists in choosing the boundary points such that they create equal intervals on the cum $\sqrt{f}$ scale, and allocate the sample with an equal number of observations to each stratum (Dalenius et. al, 1959; Ekman, 1959). Although the mathematics behind these rules is crude, they seem to work well for both theoretical and actual distributions (Cochran, 1961; Hess et. al., 1966).

In this paper we shall take the position that the population values of the two stratifying variables, X and Z, are generated from a background bivariate distribution with a joint density f and marginal densities $f_1$ and $f_2$. The population values of the study variable y are also assumed to be realizations of a stochastic background variable, and the regression of this variable on the stratification variables is assumed to be linear. The population is stratified into rs strata in the following way: Using the "cum $\sqrt{f_1}$" method, r strata are constructed alon X, and s strata are constructed by an equal partitioning of "cum $\sqrt{f_2}$". An element in the population belongs to stratum (i,j) if its X-value belongs to the i-th stratum along X and its Z-value belongs to the j-th stratum along Z. The sample is allocated with n/rs observations in each stratum. It is assumed that there are $N_{ij}$ units with Y-values (h=1,2,...,$N_{ij}$) in stratum (i,j). Their mean is

$$\overline{Y}_{ij} = N_{ij}^{-1} \sum_h Y_{ijh},$$

and their empirical variance is

$$s_{ij}^2(Y) = (N_{ij}-1)^{-1} \sum_h (Y_{ijh}-Y_{ij})^2.$$

The population mean is

$$\overline{Y} = N^{-1} \sum_{i,j,h} Y_{ijh},$$

where $N=\Sigma N_{ij}$. We denote the sample size in stratum (i,j) by $n_{ij}$, and the k-th observed Y-value in stratum (i,j) by $y_{ijk}$. The stratum mean in stratum (i,j) is

$$\overline{y}_{ij} = n_{ij}^{-1} \Sigma y_{ijk},$$

and

$$\overline{y}_{st} = \Sigma \overline{y}_{ij} N_{ij}/N$$

is an unbiased estimator of $\overline{Y}$.

We realise that this stratification and allocation need not be optimal. Other stratification and allocation methods can be studied by an approach similar to the one used in this paper.

In Section 2 we give an approximation to the variance of the stratified mean. This approximation of the variance only depends on n, r, s, the correlation coefficients between the study variable and the stratifying variables, and the simultaneous density of X and Z. In

Sections 3 and 4 we study the case where X and Z are independet, and find the optimal choice of n, r and s for fixed cost. In Section 5, X and Z are assumed to have a bivariate normal distribution.

The results in this study should be combined with results from real-life data and/or artificial data before any final conclusions about the efficiency of using more than one stratifying variable can be made, but some results seem to be indicated already: When one chooses to construct many strata, the gain from using two stratifying variables instead of one seems to be non-trivial when the correlations between the study variable and each of the stratifying variables are of some size, and the correlation between the two stratifying variables is small. Under the same conditions the results indicate that for a given number of strata, it is more efficient to use two stratifying variables and make a few strata long each variable, as compared with using only the "best" stratifying variable and make optimal stratification along this variable.

## 2. AN APPROXIMATION TO THE VARIANCE

For any two stochastic variables U and V with joint density g(u,v), and marginal densities $g_1(u)$ and $g_2(v)$ respectively, we define

$$K(U) = \int_{-\infty}^{\infty} [g_1(u)]^{\frac{1}{2}} \, du,$$

$$\sigma^2(U) = \int_{-\infty}^{\infty} u^2 g_1(u)\,du - \{\int_{-\infty}^{\infty} u g_1(u)\,du\}^2,$$

$$M(U,V) = \frac{K^3(U)K(V)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g^2(u,v) g_1^{-\frac{3}{2}}(u) g_2^{-\frac{1}{2}}(v) \, du \, dv}{12\sigma^2(U)},$$

and

$$N(U,V) = K(U)K(V)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g^2(u,v) g_1^{-\frac{1}{2}}(u) g_2^{-\frac{1}{2}}(v) \, du \, dv.$$

When finding an approximation to the variance of the stratified mean, we shall confine our attention to a finite square $[a,b] \times [c,d]$ outside of which f(x,z) may be assumed to be zero with negligible error. Let $p_1(X)<\ldots<p_{r-1}(X)$, and $q_1(Z)<\ldots<q_{s-1}(Z)$ be the boundary points defining r strata along X and s strata along Z. $p_o(X)=a$, $p_r(X)=b$, $q_o(Z)=c$, and $q_s(Z)=d$. Denote

$$A_i(X) = \int_{p_{i-1}(X)}^{p_i(X)} [f_1(x)]^{\frac{1}{2}} dx, \qquad A_j(Z) = \int_{q_{j-1}(Z)}^{q_j(Z)} [f_2(z)]^{\frac{1}{2}} dz,$$

$$W_{ij} = W_{ij}(X,Z) = \int_{p_{i-1}(X)}^{p_i(X)} \int_{q_{j-1}(Z)}^{q_j(Z)} f(x,z) ds,$$

$$W_i(X) = \int_{p_{i-1}(X)}^{p_i(X)} f_1(x) dx, \qquad W_j(Z) = \int_{q_{j-1}(Z)}^{q_j(Z)} f_2(z) dz,$$

$$\sigma_i^2(X) = \int_{p_{i-1}(X)}^{p_i(X)} x^2 \frac{f_1(x)}{W_i(X)} dx - \{\int_{p_{i-1}(X)}^{p_i(X)} x \frac{f_1(x)}{W_i(X)} dx\}^2$$

$$\sigma_j^2(Z) = \int_{p_{j-1}(Z)}^{q_j(Z)} z^2 \frac{f_2(z)}{W_j(Z)} dz - \{\int_{q_{j-1}(Z)}^{q_j(Z)} z \frac{f_2(z)}{W_j(Z)}\}^2 .$$

Following Cochran (1963; p. 130), we shall approximate f(x,z) with a constant within each stratum. Hence

$$W_{ij}(X,Z) = \xi_{ij}[p_i(X)-p_{i-1}(X)] [q_j(Z)-q_{j-1}(Z)],$$

where $\xi_{ij}$ is the "constant" value of f(x,z) within stratum (i,j). Similarly,

$$\sigma_i^2(X) \doteq [p_i(X) - p_{i-1}(X)]^2/12,$$

$$\sigma_j^2(Z) \doteq [q_j(Z) - q_{j-1}(Z)]^2/12$$

$$A_i(X) \doteq \xi_i^{\frac{1}{2}}(X)[p_i(X) - p_{i-1}(X)]$$

and

$$A_j(Z) \doteq \xi_j^{\frac{1}{2}}(Z)[q_j(Z) - q_{j-1}(Z)],$$

where $\xi_i(X)$ and $\xi_j(Z)$ are "constant" values of the marginal densities within the i-th stratum in X and the j-th stratum in Z respectively.

The model adopted in this paper is similar to the model adopted in Thomsen (1976), and differs slightly from others in the literature, in that we have assumed explixitly that the population values of Y, X, Z are generated by a background distribution. As in Thomsen (1976) it can be shown that this makes no difference to the mathematics.

Suppose now that the regression of Y and X and Z is linear, i.e., that

$$Y_{i,j,k} = c + \alpha X_{i,j,k} + \beta X_{i,j,k} + e_{i,j,k},$$

where $E(e_{i,j,k})=0$, $\mathrm{var}(e_{i,j,k})=\sigma^2$, and the $i_{ijk}$ are uncorrelated with each other and with $X_{ijk}$ and $Z_{ijk}$. If we for convenience let $\mathrm{var}(\bar{y}_{st})$ denote the expected value of the conditional variance of $\bar{y}_{st}$ given all population Y-values, and use a similar approach as in lemma 1 in Thomsen (1976), we find under the assumptions given that

$$(2.1) \qquad \mathrm{var}\,(\bar{y}_{st}) = \frac{rs}{n} \sum_{i,j} W_{ij}^2 \sigma_{ij}^2(Y)$$

$$= \frac{rs}{n} \sum_{i,j} W_{ij}^2 (\alpha^2 \sigma_{ij}^2(X) + \beta^2 \sigma_{ij}^2(Y) + \sigma_{ij}^2(e) + 2\alpha\beta\ \mathrm{cov}_{ij}(X,Z)),$$

where $\mathrm{cov}_{ij}(X,Z)$ is the covariance between X and Z within stratum (i,j). We shall assume that $\sigma_{ij}^2(e) = \sigma^2(e)$ and $\mathrm{cov}_{ij}(X,Z) = 0$ for all i,j. Then

$$(2.2) \qquad \mathrm{var}(\bar{y}_{st}) = \frac{rs}{n}\{\alpha^2 \sum_{i,j} W_{ij}^2 \sigma_{ij}^2(X) + \beta^2 \sum_{i,j} W_{ij}^2 \sigma_{ij}^2(Z) + \sum_{i,j} W_{ij}^2 \sigma^2(e)\}.$$

Using the approximations suggested above, we find that

$$\sum_{i,j} W_{ij}^2 \sigma_{ij}^2(X) = \sum_{i,j} \xi_{ij}^2 [p_i(X) - p_{i-1}(X)]^4 [q_j(Z) - q_{j-1}(Z)]^2/12$$

$$(2.3) \qquad \doteq \sum_{i,j} \xi_{ij}^2 \xi_i^{-\frac{3}{2}}(X) \xi_j^{-\frac{1}{2}}(Z) [p_i(X) - p_{i-1}(X)]$$

$$[q_j(X) - q_{j-1}(X)]\ A_i^3(X) A_j(Z)/12$$

The strata along X and Z are constructed such that $A_i(X)$ is equal to the constant $K(X)/r$, and such that $A_j(Z)$ is equal to the constant $K(Z)/s$. Inserting these values into (2.3), we find that

$$\sum_{i,j} W_{ij}^2 \sigma_{ij}^2(X) \doteq \frac{K^3(X) K(Z)}{r^3 s\ 12} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f^2(x,z) f_1^{-\frac{3}{2}}(x) f_2^{-\frac{1}{2}}(z)\ dxdz$$

$$(2.4) \qquad = \frac{M(X,Z)\sigma^2(X)}{r^3 s}.$$

By symmetry we find that

$$(2.5) \qquad \sum_{i,j} W_{ij}^2 \sigma_{ij}^2(Z) \doteq \frac{M(Z,X)\sigma^2(Z)}{s^3 r}$$

Again using the approximations suggested above, we find that

(2.6) $$\sum_{i,j} W_{ij}^2 \doteq \sum_{i,j} \xi_{ij}^2 (p_i(X) - p_{i-1}(X))^2 (q_j(Z) - q_{j-1}(Z))^2$$

$$\doteq \sum_{i,j} \xi_{ij}^2 \xi_i^{-\frac{1}{2}}(X) \xi_j^{-\frac{1}{2}}(Z)(p_i(X) - p_{i-1}(X))(q_j(Z) - q_{j-1}(Z)) A_i(X) A_j(Z)$$

$$\doteq \frac{K(X)K(Z)}{rs} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f^2(x,z) f_1^{-\frac{1}{2}}(x) f_2^{-\frac{1}{2}}(z)\, dxdz = N(X,Z)/rs.$$

Inserting (2.4), (2.5) and (2.6) into (2.2) we find that

(2.7) $$\text{var}(\bar{y}_{st}) \doteq \frac{rs}{n}\{\alpha^2 \frac{M(X,Z)\sigma^2(X)}{r^3 s} + \beta^2 \frac{M(Z,X)\sigma^2(Z)}{s^3 r} + \sigma^2(e) \frac{N(X,Z)}{rs}\}$$

For any pair (u,v) of stochastic variables, let $r_{uv}$ denote the correlation coefficient between U and V. Then

(2.8) $$\alpha = (r_{xy} - r_{yz} r_{xz})\sigma(Y)/\{(1 - r_{xz}^2)\sigma(X)\},$$

(2.9) $$\beta = (r_{zy} - r_{xz} r_{xy})\sigma(Y)/\{(1 - r_{xz}^2)\sigma(Z)\},$$

and

(2.10) $\sigma^2(e) = (1 - R_{y\cdot xz}^2)\sigma^2(Y)$, where $R_{y\cdot xz}^2$ denotes the multiple correlation coefficient.

Inserting (2.8), (2.9) and (2.10) into (2.7) gives

(2.11) $$\text{var}(\bar{y}_{st}) = \frac{\sigma^2(Y)}{n} \Big[M(X,Z)(r_{xy} - r_{yz} r_{xz})^2/\{(1-r_{xz}^2)r^2\} +$$

$$+ M(Z,X)(r_{zy} - r_{xz} r_{xy})^2/\{(1-r_{xz}^2)s^2\} + N(X,Z)(1-R_{y\cdot xz}^2)\Big]$$

<u>Remark 1</u>: For r and s large, it follows from (2.11) that

$$\frac{\text{var}(\bar{y}_{st})}{\text{var}(\bar{y})} \doteq N(X,Z)(1-R_{y\cdot xz}^2),$$

where $\text{var}(\bar{y})$ is the variance of the sample mean when the sample is simple random.

In the next section we shall study (2.11) under the assumption that $f(x,z) = f_1(x)f_2(z)$.

## 3. INDEPENDENCE BETWEEN THE STRATIFYING VARIABLES

Following Serfling (1968), we define for any stochastic variable U with density g we define

$$K^*(U) = \int_{-\infty}^{\infty} [g(u)]^{3/2} du,$$

$$k_u = K^4(U)/12\sigma^2(U), \quad k_u^* = K(U)K^*(U).$$

Under the assumption that X and Z are independent (2.11) simplifies to

$$(3.1) \qquad \mathrm{var}(\overline{y}_{st}) = (\sigma^2(Y)/n)[k_z^* k_x r_{xy}^2/r^2 + k_x^* k_z r_{zy}^2/s^2 + k_x^* k_z^* (1-R_{y\cdot xz}^2)]$$

Formula (3.1) parallels formula (5.A.24) in Cochran (1963; p. 134) and formula(2.15) in Serfling (1968), both derived for the case with one stratifying variable. The following values of $k_u$ and $k_u^*$ for different distributions of U are given by Serfling (1968).

| Distribution | $k_u$ | $k_u^*$ |
|---|---|---|
| Rectangular | 1 | 1 |
| Normal | $2\Pi/3 \doteq 2.09$ | $2/\sqrt{3} \doteq 1.16$ |
| Exponential | 4/3 | 4/3 |
| Gamma | 1.64 | 1.21 |

In addition he shows that K and $K^*$ are invariant under a change of location with fixed scale and that k and $k^*$ are invariant under either scale or location.

In Table 1 below we give the ratio of the variances of a stratified sample, and a simple random sample, when both stratifying variables are rectangular. In Table 2 the same ratios are given, but here the distribution of X is normal and the distribution of Z is exponential.

Remark 2. For r and s large we have

$$\mathrm{var}(\bar{y}_{st})/\mathrm{var}(\bar{y}) \doteq k_x^{*} k_z^{*}(1 - R^2_{y\cdot xz}).$$

The results in 'tables 1 and 2 seem to suggest that, for å given number of strata, it is more efficient to use more than one stratifying variable, and make a few strata along each variable, as comared with using only the "best" stratifying variable and make many strata along this variable. In cases with a large budget, meaning many strata, and good information on a number of variables, the gain from using more than one variable seems to be non-trivial.

Table 1. Ratio of variances, when $r_{xy}$=0.85, $r_{zy}$=0.50, and the distributions of X and Z are both rectangular

| s<br>r | 1 | 2 | 3 | 4 | 5 | 6 | ∞ |
|---|---|---|---|---|---|---|---|
| 1 | 1.000 | 0.812 | 0.778 | 0.766 | 0.760 | 0.757 | 0.750 |
| 2 | 0.458 | 0.271 | 0.236 | 0.224 | 0.218 | 0.215 | 0.208 |
| 3 | 0.358 | 0.170 | 0.136 | 0.123 | 0.118 | 0.115 | 0.108 |
| 4 | 0.323 | 0.135 | 0.101 | 0.088 | 0.083 | 0.080 | 0.073 |
| 5 | 0.306 | 0.119 | 0.094 | 0.072 | 0.066 | 0.063 | 0.056 |
| 6 | 0.298 | 0.110 | 0.075 | 0.063 | 0.058 | 0.055 | 0.048 |
| ∞ | 0.277 | 0.090 | 0.055 | 0.043 | 0.038 | 0.034 | 0.028 |

Table 2*). Ratio of variances, when $r_{xy}$=0.85, $r_{zy}$=0.50, the distribution of X is normal, and the distribution of Z is exponential

| s<br>r | 2 | 3 | 4 | 5 | 6 | ∞ |
|---|---|---|---|---|---|---|
| 2 | 0.6283 | 0,5821 | 0,5658 | 0.5583 | 0.5542 | 0.5450 |
| 3 | 0.3480 | 0.3018 | 0.2855 | 0.2780 | 0.2739 | 0.2647 |
| 4 | 0.2509 | 0.2047 | 0.1884 | 0.1809 | 0.1768 | 0.1676 |
| 5 | 0.2058 | 0.1596 | 0.1433 | 0.1358 | 0.1317 | 0.1225 |
| 6 | 0.1814 | 0.1352 | 0.1189 | 0.1114 | 0.1073 | 0.0981 |
| ∞ | 0.1258 | 0.0796 | 0.0633 | 0.0558 | 0.0517 | 0.0425 |

*) Cases where r or s is equal to one are not included in this and the following tables as the approximations $rs\Sigma W_{ij}=k_x^{*}K_z^{*}$, $s\Sigma W_j^2=k_z^{*}$, and $r\overset{r}{\underset{j}{\Sigma}}W_i^2=k_x^{*}$ obviously are wrong in these cases, except when both are rectangular.

## 4. OPTIMAL CHOICE OF r, s, AND n FOR FIXED COSTS

In the case with two stratifying variables the form of the cost function does not seem to have been studied. We shall assume a cost function of the form .

$$C = c_o + c_1 n + c_2(r+s). \tag{4.1}$$

We want to minimize

$$\text{var } (\bar{y}_{st}) = \sigma^2(Y)n^{-1}\left[k_z^* k_x r_{xy}^2/r^2 + k_x^* k_z r_{zy}^2/s^2 + k_x^* k_z^*(1 - R_{y\cdot xz}^2)\right]$$

for fixed C.

Using Lagrange's method we find the following three equations:

$$(s/r)^3 = Q^2,$$

$$c_2 r^3 + 3c_2\gamma(Q^{2/3}+1)r - 2(C-c_0)\gamma = 0,$$

and

$$n = (C - c_0 - (r+s)c_2)/c_1 \text{ ,}$$

where

$$Q^2 = \frac{k_x^* k_z}{k_x k_z^*}\,\frac{r_{zy}^2}{r_{xy}^2}\text{ ,} \quad \text{and} \quad \gamma = \frac{k_x r_{xy}^2}{k_x^*(1-R_{y\cdot xz}^2)}\text{ .}$$

Below is given the optimal choice of n, r, and s when $k_x = k_x^* = k_z = k_z^* = 1$, $r_{xy}^2 = 0.7225$, and $r_{zy}^2 = 0.25$ for three different values of $C-c_0$, $c_1 = 10$, and $c_2 = 200$ (Serfling, 1968).

| | Optimal choice | | |
|---|---|---|---|
| $C-c_0$ | r | s | n |
| 2 000 | 4 | 3 | 60 |
| 4 000 | 6 | 4 | 200 |
| 6 000 | 8 | 6 | 320 |

Using the variable X alone, the following choice is optimal by formula (4.21) in Serfling (1968).

| $C-c_0$ | Optimal choice | |
|---|---|---|
| | r | n |
| 2 000 | 3 | 140 |
| 4 000 | 4 | 320 |
| 6 000 | 5 | 500 |

It is seen that the optimal number of strata is substancially larger when using two stratifying variables than when using only one.

In practice one is seldom able to find two stratifying variables with values that can be considered to be realisations of two independent stochastic variables. Before using the two variables one can therefore transform them, so that the empirical correlation coefficient between the transformed variables is zero, and the results in section 3 are relevant. The sampler, however, would often try to avoid any transformation of the stratifying variables especially if some of the strata using the original stratifying variables are domains of study, which is often the case. In the next section therefore we shall study (2.11) when the stratifying variables have a bivariate normal distribution.

## 5. THE STRATIFYING VARIABLES DISTRIBUTED AS BIVARIATE NORMAL

When $f(x,z)=(2\Pi)^{-1}(\sigma\tau)^{-1}(1-\rho^2)^{-\frac{1}{2}}\exp\{\frac{-1}{2(1-\rho^2)}((\frac{x-\xi}{\sigma})^2-$

$$\frac{2\rho(x-\xi)(z-\eta)}{\sigma\tau}+(\frac{z-\eta}{\tau})^2)\}$$

we find that

$$(5.1) \qquad M(X,Z) = M(Z,X) = 4\Pi/\{3\sqrt{3}\ (1-\rho^2)\}$$

and

$$(5.2) \qquad N(X,Z) = 4((\rho^2+3)^2 - 16\rho^2)^{-\frac{1}{2}}.$$

Thus, both n(X,Z) and N(X,Z) are independent of $\xi$, $\eta$, $\sigma$ and $\tau$. Inserting (5.1) and (5.2) into (2.11) we find that

$$\text{var}(\bar{y}_{st}) = \frac{\sigma^2(Y)}{n}\left[\frac{4\Pi(r_{xy}-r_{yz}\rho)^2}{3\sqrt{3}(1-\rho^2)^2r^2}+\frac{4\Pi(r_{zy}-r_{xz}\rho)^2}{3\sqrt{3}(1-\rho^2)^2s^2}+\frac{4(1-R^2_{y\cdot xz})}{(\rho^4-10\rho^2+9)^{\frac{1}{2}}}\right]$$

In Table 3 is given the ratio between the variance of a stratified sample and that of a simple random sample when the stratifying variables are indepandent and normally distributed. In Table 4 the same ratio is given when the stratifying variables are bivariate normal with $\rho=0.20$.

Table 3. Ratio of variances when $r_{xy}=0.85$, $r_{zy}=0.50$, $\rho=0$, and the distributions of X and Z are both normal

| s<br>r | 2 | 3 | 4 | 5 | 6 | ∞ |
|---|---|---|---|---|---|---|
| 2 | 0.628 | 0.543 | 0.514 | 0.500 | 0.493 | 0.476 |
| 3 | 0.385 | 0.300 | 0.271 | 0.257 | 0.250 | 0.233 |
| 4 | 0.299 | 0.214 | 0.185 | 0.171 | 0.164 | 0.147 |
| 5 | 0.260 | 0.175 | 0.146 | 0.132 | 0.125 | 0.108 |
| 6 | 0.238 | 0.153 | 0.124 | 0.110 | 0.103 | 0.086 |
| ∞ | 0.190 | 0.105 | 0.076 | 0.062 | 0.055 | 0.038 |

Table 4. Ratio of variances when $r_{xy}=0.85$, $r_{zy}=0.50$, $\rho=0.20$, and the simultaneous distribution of X and Z is bivariate normal

| s<br>r | 2 | 3 | 4 | 5 | 6 | ∞ |
|---|---|---|---|---|---|---|
| 2 | 0.609 | 0.569 | 0.555 | 0.549 | 0.545 | 0.537 |
| 3 | 0.404 | 0.364 | 0.350 | 0.344 | 0.338 | 0.332 |
| 4 | 0.332 | 0.292 | 0.278 | 0.272 | 0.268 | 0.260 |
| 5 | 0.299 | 0.259 | 0.245 | 0.239 | 0.235 | 0.227 |
| 6 | 0.281 | 0.241 | 0.127 | 0.221 | 0.217 | 0.209 |
| ∞ | 0.240 | 0.200 | 0.186 | 0.180 | 0.176 | 0.168 |

As one would expect, the reduction of the variance due to stratification is smaller when the stratifying variables are correlated than when they are uncorrelated. However, for the case r=s=2 the reduction of the variance is larger in Table 4 than in Table 3. This seems unreasonable, and it is probably due to the crudeness of the approximations when the number of strata is as small as in this case.

## 6. ACKNOWLEDGEMENTS

I am deeply grateful to Professor Leslie Kish, University of Michigan, for suggesting this problem, and for useful comments on earlier drafts of this paper. Thanks are due also to Professor Jan M. Hoem, University of Copenhagen, for helpful comments on previous drafts of this paper.

## 7. REFERENCES

[1] Cochran, W.G. (1961): Comparison of Methods for Determining Stratum Boundaries. Bull. Int. Stat. Int. 38, 2, 345-58.

[2] Cochran, W.G. (1963): Sampling Techniques. John Wiley & Sons, Inc.

[3] Dalenius, T. (1957): Sampling in Sweden Contributions to the methods and theories of sample survey practice. Almqvist and Wicksell, Stockholm.

[4] Dalenius, T and Hodges, J.L.Jr. (1959): Minimum variance stratification. JASA. 54, 88-101.

[5] Ekman (1959): An approximation useful in univariate stratification. Ann. Math. Statist. 30, 219-29.

[6] Hess, I., Sethi, V.K. and Balakrishnan, T.R. (1966): Stratification: A practical investigation J.A.S.A. 61, 74-90.

[7] Kish, L. (1965): Survey Sampling. Wiley, New York.

[8] Neyman. J. (1934): On the two different aspects of stratified sampling and the method of purposive selection J.R.S.S. 97, 558-606.

[9] Serfling, R.J. (1968): Approximately Optimal Stratification, JASA 63, 1298-1309.

[10] Sethi, V.K. (1963): A Note on Optimum Stratification of Populations for Estimating the Population Mean. Aust. J. Stat. 5, 20-33.

[11] Thomsen, I. (1976): A Comparison of Approximately Optimal Stratification given porportional Allocation with other Methods of stratification and allocation. Metrika 23, 15-25.