On Some Criteria for Stratification

Chikio HAYASHI
Fumiyuki MARUYAMA
Masatsugu D. ISHIDA

(Received August 10, 1950)

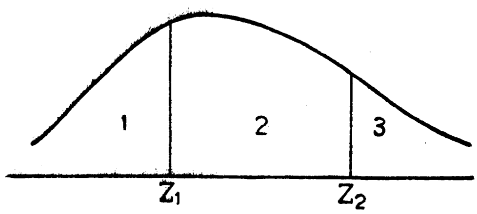
I. Generally speaking, Stratification is made to reduce the sampling variance of an estimate fòr a poulation parameter. In stratifying a population which is constructed by giving a certain sampling probability to every element of a universe in which we will obtain some propositions, the critèria for stratification have been decided in some cases from the qualitative standpoint, for example by the qualitative qualities of a universe and in other cases from the quantitative standpoint, for example by using the quantitative data of elements in a universe. But the criteria for stratification seem not to have been considered from the theoretical point of view, that is so say, the criteria dividing a population into strata seem to be chosen arbitrarily and conventionally. In this paper, the criteria dividing a population into strata will be theoretically considered, in some problems.

II. Let Y_i be the label of an element of population $i=1,\ldots, N$ where N is population size and supposed to be very large.

When we are estimating the population mean $\bar{Y}$ where $\bar{Y}=\frac{1}{N}\sum_{i=1}^{N} Y_i$, the population will be stratified by using the labels which are highly correlated with Y_i. Let X_i be the label correlated with Y_i. For example X_i is a quantity in the k-th year concerning a quality in question, Y_i is a quantity in the $(k+1)$th year concerning it. Suppose that we have some knowledge about X_i $(i=1,\ldots, N)$, for example the distribution of X in the population.

Using some knowledge about X_i, usually the population is divided into strata, because the result may be effective.
Now suppose that the distribution of X in the population is known. Then we consider the most effective and reasonable stratification from this situation.

(a) First problem



Let F(x) be, the distribution function, which is approximately expressed by a differentiable function, is which has density function f(x).

Let z_1, z_2 be the dividing points of population.

Elements the labels of which are smaller than z_1 belong to stratum 1. Elements the labels of which are larger than z_1, and smaller than z_2 belong to the stratum 2. Elements the labels of which are larger than z_2 belong to the stratum 3.

In some cases, sampling methods is adopted in strata, 1, 2 but complete survey method adopted in stratum 3. The underlying idea is perhaps as below.

Suppose that population mean is estimated. Samples of size n are drawn with equal sampling probability and n_i is the sample size allocated to i stratum. We assume that n_i is decided by the socalled Neymans' optimum allocation in a linear unbiased estimate of population mean, that is to say, $n_i \doteqdot n \dfrac{p_i\sigma_i}{\sum_{i=1}^{R} p_i\sigma_i}$ where σ_i is the variance of i stratum, N_i is the size of i stratum, p_i is $\dfrac{N_i}{\sum_{i=1}^{R} N_i}$, R is the number of strata. In our case, the following relations approximately hold,

$$p_1=\int_{-\infty}^{z_1} dF(x),\quad p_2=\int_{z_1}^{z_2} dF(x),\quad p_3=\int_{z_2}^{\infty} dF(x),$$

$$N_1=Np_1,\quad N_2=Np_2,\quad N_3=Np_3,\quad N=N_1+N_2+N_3$$

$$\sigma_1^2=\frac{\int_{-\infty}^{z_1} x^2 dF(x)}{\int_{-\infty}^{z_1} dF(x)}-\left(\frac{\int_{-\infty}^{z_1} x dF(x)}{\int_{-\infty}^{z_1} dF(x)}\right)^2$$

$$\sigma_2^2=\frac{\int_{z_1}^{z_2} x^2 dF(x)}{\int_{z_1}^{z_2} dF(x)}-\left(\frac{\int_{z_1}^{z_2} x dF(x)}{\int_{z_1}^{z_2} dF(x)}\right)^2$$

$$\sigma_3^2=\frac{\int_{z_2}^{\infty} x^2 dF(x)}{\int_{z_2}^{\infty} dF(x)}-\left(\frac{\int_{z_2}^{\infty} xdF(x)}{\int_{z_2}^{\infty} dF(x)}\right)^2$$

If $N_i \leqq n_i$, it may be reasonably recognized that i strata will be completely surveyed. In many cases σ^2 is usually large when x is large, for example, the variance of stratum 3 is larger. So $N_3 \leqq n_3$ may hold in some cases. Then sampling method in stratum 1, 2, and complete survey method in stratum 3. But this is meaningless if the above condition is not fulfilled.

This stand point in sampling surveys is considered to be reasonable. We proceed to the next consideration. What is the reasonable method of diciding the point z_2, the stratum, the labels belonging to which are larger than z_2, being completely surveyed?

Following the idea mentioned above, z_2 is decided as the maximum value to satisfy the following relation, where population size is large,

$$\int_{z_2}^{\infty} dF(x) \geqq n \frac{\sigma_3 p_3}{\sigma_1 p_1+\sigma_2 p_2+\sigma_3 p_3}$$

this idea holds in general cases.

(b) Second Problem

Suppose that n is allocated to strata in proportion to the size of them, that is to say, $n_i=n \dfrac{N_i}{\sum_{i=1}^{R} N_i}$ where R is the number of strata.

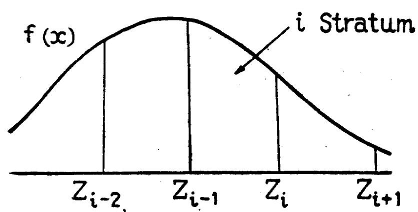
This allocation is very usuful in many cases, where the data obtained from sampling are analized from many stand points. In this case, the optimum criteria for stratification are as followings. Of course population mean will be estimated. We wish to estimate $\bar{Y}$, but consider about X_i and think of the optimum method to estimate the population mean $\bar{X}=\frac{1}{N}\sum_{i=1}^{N} X_i$ using the past knowledge, where $N=\sum_{i=1}^{R} N_i$. As Y_i is highly correlated with X_i, this may be well recognized in practise. Samples of size n_i are drawn from i stratum, and sample mean $\bar{x}_i$ is made.

To estimate unbiasedly population mean, $\bar{x}=\sum_{i=1}^{R} p_i \bar{x}_i$ is made. The variance of $\bar{x}$, $\sigma_{\bar{x}}^2$, is followings

$$\sigma_{\bar{x}}^2=\sum_{i=1}^{R} \frac{N_i-n_i}{N_i-1} \left(\frac{N_i}{N}\right)^2 \frac{\sigma_i^2}{n_i}$$

$$\doteqdot \text{Const.} \sum^{R} N_i \sigma_i^2$$

where N_i, σ^2_i is respectively size and variance of i stratum, and N_i is large.



$$N_i\doteqdot N\int_{z_{i-1}}^{z_i} dF(x)$$

$$\sigma_i^2=\frac{\int_{z_{i-1}}^{z_i} x^2 dF(x)}{\int_{z_{i-1}}^{z_i} dF(\alpha)}-\left(\frac{\int_{z_{i-1}}^{z_i} xdF(x)}{\int_{z_{i-1}}^{z_i} dF(\alpha)}\right)^2$$

Dividing points z_1, z_2,$\cdots$, z_{R-1} must be decided to minimize the variance $\sigma_{\bar{x}}^2$. This idea is the most reasonable. The dividing points z_1, z_2,..., z_{R-1} which are obtained from this stand point, are the optimum criteria for stratification.

Moreover using the relation

$$\sigma^2=\sum_{i=1}^{R} \frac{N_i}{N} \sigma_i^2+\sum_{i=1}^{R} \frac{N_i}{N} (\bar{X}-\bar{X}_i)^2$$

where $\bar{X}$, σ^2 are mean and variance of the population, $\bar{X}_i$ is mean of i stratum, i.e.,

$$\bar{X}=\int_{-\infty}^{\infty} xdF(x)$$

$$\sigma^2=\int_{-\infty}^{\infty} x^2dF(x)-\left(\int_{-\infty}^{\infty} xdF(x)\right)^2$$

$$\bar{X}_i=\frac{\int_{z_{i-1}}^{z_i} xdF(x)}{\int_{z_{i-1}}^{z_i} dF(x)}$$

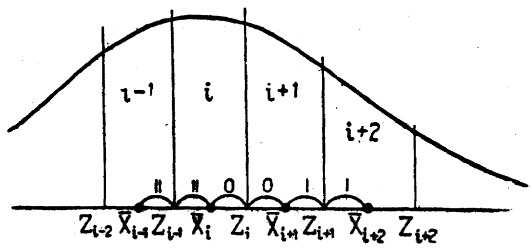
So

$$\sigma_{\bar{x}}^2\doteqdot\text{Const}\left\{\sigma^2-\sum_{i=1}^{R} \frac{N_i}{N} (\bar{X}-\bar{X}_i)^2\right\}$$

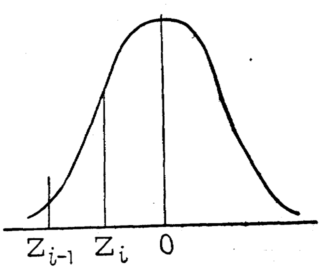
To minimize $\sigma_{\bar{x}}^2$, i.e. to maximize $\sum_{i=1}^{R} N_i (\bar{X}-\bar{X}_i)^2$, $z_1, \ldots, z_{R-1}$ must be decided.

This implies that $z_1, \ldots, z_{R-1}$ must be decided to maximize $Q=\sum_{i=1}^{R} \bar{X}_i^2 N_i$

So, from $\frac{\partial Q}{\partial z_i}=0$, $i=1, \ldots, R-1$, the relations $z_i=\frac{1}{2}(\bar{X}_i+\bar{X}_{i+1})$ $i=1, \ldots, R-1$, are obtained. If this solution is unique, $(z_1, \ldots, z_{R-1})$ is the optimum deviding points' set.

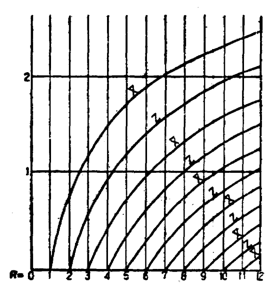


Illustration



(1) Let $dF(x)=\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$.

In this case $z_1, \ldots, z_{R-1}$ are shown in Table I and Graph I.



Graph I

(2) Let the distribution be rectangle. In this case, $z_0=0$, $z_i=\frac{i}{R}$, $\bar{X}_i=\frac{i-1}{R}+\frac{1}{2R}$, N is the population size, and n is the sample size.

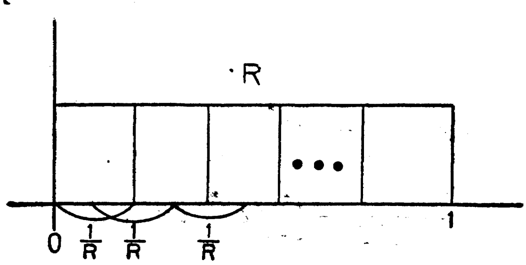


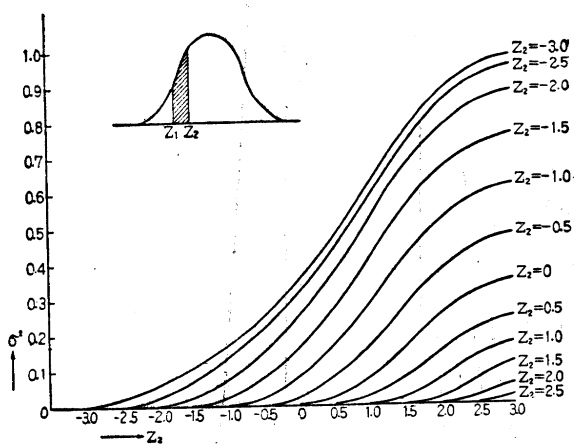
Table I

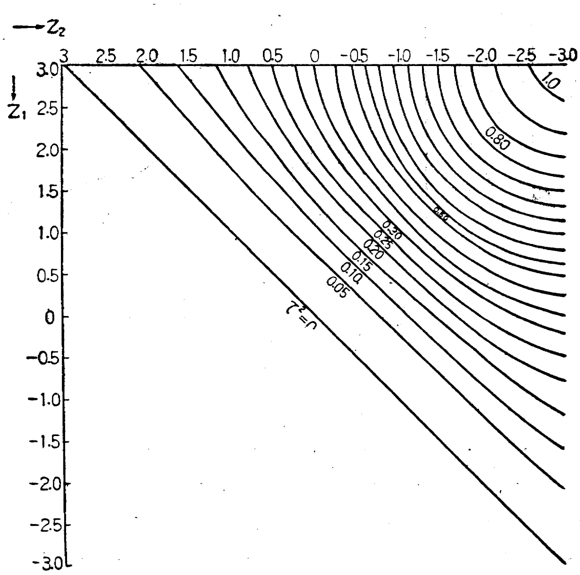
R	2	3	4	5	6	7	8	9	10	11	12
$\bar{X}$	0.798	1.224	1 510	1.724	1 894	2.033	2.152	2.256	2.346	2.426	2.499
Z	0	0.612	0.982	1.244	1.447	1.611	1.748	I.867	1.970	2.059	2.141
φ	0.399	0.331	0.246	0.184	0.140	0.109	0.087	0.070	0.057	0.048	0.040
Φ	0.5	0.730	0.837	0.893	0.926	0.946	0.960	0.969	0.976	0.980	0.984
$\bar{X}$		0	0.453	0.765	1.000	1.188	1.344	1.478	1.593	1.692	1.783
Z			0	0.382	0.659	0.874	1.050	1.799	1.326	1.436	1.535
φ			0.399	0.371	0.321	0.272	0.230	0.194	0.166	0.142	0.123
Φ			0.5	0.649	0.745	0.809	0.853	0.885	0.908	0.924	0.938
$\bar{X}$				0	0.318	0.561	0.756	0.920	1.059	1.179	1.286
Z					0	0.780	0.501	0.682	0.835	0.966	1.081
φ					0.399	0.384	0.352	0.316	0.281	0.250	0.222
Φ					0.5	0.610	0.692	0.753	0.798	0.833	0.860
$\bar{X}$						0	0.245	0.444	0.611	0.752	0.877
Z							0	0.222	0.406	0.560	0.695
φ							0.399	0.389	0.367	0.341	0.313
Φ							0.5	0.588	0.657	0.712	0.756
$\bar{X}$								0	0.200	0.368	0.512
Z									0	0.184	0.340
φ									0.399	0.392	0.376
Φ									0.5	0.573	0.633
$\bar{X}$										0	0.169
Z											0
φ											0.399
Φ											0.5

In Table I

$$\varphi=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$$

$$\Phi=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{x^2}{2}}dx$$

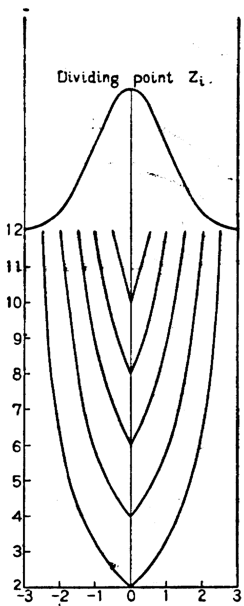


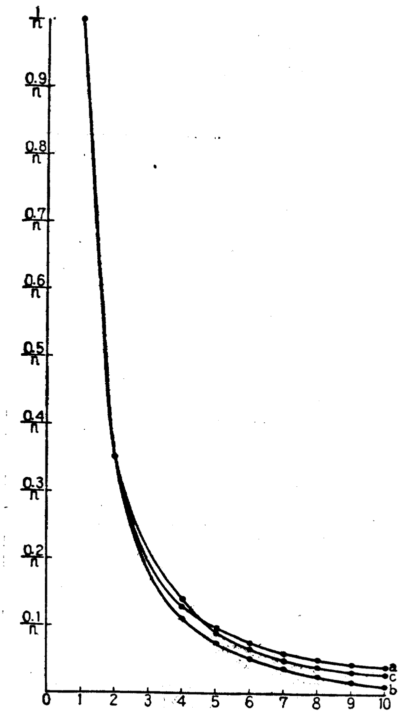


The variance of sample mean $\bar{x}$ is $\sigma_{\bar{x}}^2$

$$\sigma_{\bar{x}}^2 \doteqdot \left(\frac{1}{n}-\frac{1}{N}\right)\frac{1}{12}\,\frac{1}{R^2} \doteqdot \frac{1}{n}\,\frac{1}{12}\,\frac{1}{R^2}$$

$\sigma_{\bar{x}}^2$ is proportionate to $\frac{1}{R^2}$. So $\sigma_{\bar{x}}$ is reduced to $\frac{1}{4}$ if the number of strata becomes twice, while $\sigma_{\bar{x}}^2$ is reduced to $\frac{1}{2}$ if the sample size becomes twice.





III. In some problems, it is required that every stratum has the same variance. Then $z_1, z_2, \ldots, z_{R-1}$ may be decided satisfying the relation

$$\sigma_1^2=\sigma_2^2=\ldots=\sigma_R^2=\tau^2$$

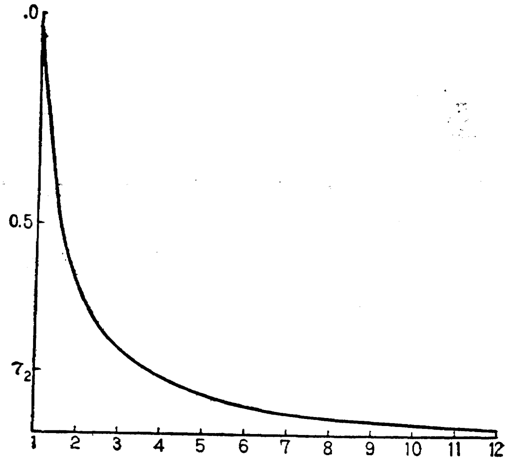
This stratification is useful not only to allocate samples (in this case, Neyman's Method is identifyed with the size proportionate allocation), but also to analyse the results. Moreover we can apply this method for typification of a universe and controlling the groups.

Then we will show the some properties concerning this method. We take up the normal distribution

$$F(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{x^2}{2}}\,dx$$

(for convenience we consider the interval $-3\leq x\leq +3$, approximately), and decide, $z_1, z_2, \ldots, z_{R-1}$. For this purpose, we calculated the variance τ^2 between z_1 and z_2

GRAPH VII



a Allocating by Size Proportionate Metned in Equall size Strata.

b Allocating by Neyman's Method in Equall size Strata.

c Allocating by Size Proportionate Method in Equall Variance Strata. (Sample size: n)

$$\tau^2=\frac{\int_{z_1}^{z_2} x^2 dF(x)}{\int_{z_1}^{z_2} dF(x)}-\left(\frac{\int_{z_1}^{z_2} x dF(x)}{\int_{z_1}^{z_2} dF(x)}\right)^2$$

and drew Graph III, IV. From this graphs, we can acquire the deviding point $z_1, z_2, \ldots, z_{R-1}$, by some numerical calculation of interporation. Graph V show this deviding points $z_1, z_2, \ldots, z_{R-1}$ and Graph VI show τ^2(variance of each stratum) about n strata. We can easily see that if the number of strata is more than eight, the dividing points are almost equal interval,

$$z_2-z_1=z_3-z_2=\ldots$$

When the distribution is not normal, so far as the number of strata is large enough, be sufficient for this method to divide the intervals equally. This method will be applicable.

Applying this dividing method in sampling design, the sampling variances $\sigma_{\bar{x}}^2$ of the sample mean $\bar{x}$ are shown in Graph VII.

Let R be the number of strata.

(a) $N'_1=N'_2=\ldots=N'_R$ (The size of every stratum is equall.

Allocating n samples by size proportionate method

$$n_i=n\frac{N'_i}{\sum_{i=1}^{R} N'_i}$$

(b) $$N_1' = N_2' \dots = N_R'$$

Allocating by Neyman's method.

$$n_i = n \frac{\sigma_i}{\sum_{i=1}^{R} \sigma_i}$$

(c) Equal variance in each stratum. Allocating by size proportionate method.

Institute of Statistical Mathematics.