this model is taken as accurate, it leads to the conclusions that occupational and educational rank are negatively related to symptom level, that racial-ethnic rank is positively related to symptom level, and that all forms of sharp status inconsistency produce more or less equivalent increments in symptom level.

The second alternative model, Model II, was based on the hypothesis that racial-ethnic status *per se* has no effect on symptom level. It included terms for the additive effects of occupation and education and terms for effects of sharp inconsistency, but not for additive racial-ethnic effects. Like Model I, this analysis leads to conclusions that educational and occupational rank are negatively related to symptom level and that sharp occupation-education inconsistencies increase symptom levels. Unlike the previous model, Model II indicates that sharp (and perhaps moderate) inconsistency between high ascribed status and low achieved status has a much greater impact on symptom level than the converse form of discrepancy.

Of course, these two alternate models fall far short of exhausting the explanatory possibilities.[22] But they do constitute two plausible hypotheses for further test. What are the grounds for choosing between them at the present time? Model I, involving all three status dimensions and four types of sharp inconsistency, accounts for somewhat more variation than Model II, at the cost of adding two terms to the prediction equation. We prefer Model II, however, because we find the assumption that inconsistency between high ascribed and low achieved status is especially likely to lead to a symptomization response, theoretically more palatable, for reasons explained in the earlier report,[23] than the assumption that high racial-ethnic status is itself positively related to symptom level.

A last conclusion is the the method of analysis used herein should be admirably suited to much sociological research, in which low levels of measurement and an interest in interaction are frequently combined.[24]

[22] For example, mentally distressed people may tend both to become status inconsistent and to develop symptoms. This hypothesis does not, however, account for the relations between status inconsistency and other dependent variables, notably political liberalism.

[23] Jackson, *op. cit.*, see pp. 476–477.

[24] For another very useful method, see Leo A. Goodman, "Interactions in Multidimensional Contingency Tables," *The Annals of Mathematical Statistics*, 35 (1964), pp. 632–646 and "Simple Methods for Analyzing Three-Factor Interaction in Contingency Tables," *Journal of the American Statistical Association*, 59 (1964), pp. 319–352.

SAMPLING ORGANIZATIONS AND GROUPS OF UNEQUAL SIZES *

LESLIE KISH
University of Michigan

A methodological issue arises whenever groups of elements of greatly differing sizes serve as observational units. The value of each unit is observed and assigned a single group value $\bar{Y}_a$; but the values also have meaning for the parent population of elements. In studies of groups, organizations and ecological units, the unweighted group mean $\bar{Y}_g$ of the units is often computed automatically. But a mean weighted for significance should always be considered and usually preferred, particularly the element mean $\bar{Y}_e$, weighted by the numbers of elements in the units. For estimating $\bar{Y}_e$, it is efficient to select units with probabilities proportional to those numbers. Implications for several research designs are advanced. The difference $\bar{Y}_e - \bar{Y}_g = \bar{Y}_g R C_y C_n$ is discussed.

ONE of the frightening statements made about American education, around 1957, was that half of the high schools offered no physics, a quarter no chemistry, and a quarter no geometry. It was later noted that, although these back-

* Presented in September 1964 to the annual meeting of the American Sociological Association in Montreal. Examples are drawn from research at the Survey Research Center, Institute for Social Research. Support was received from the National Science Foundation Grant G-7571.

ward schools were numerous indeed, they accounted for only 2 per cent of all high school students. There were many more small schools than large ones, but the small proportion of large schools accounted for a large proportion of students. Moreover, the curricula and facilities of large and small schools can and do differ drastically. Hence, presenting average school characteristics gives a misleading picture of conditions facing the average student.

In this illustration, the population elements are students, since we are interested in their opportunities. The school serves as both sampling unit and observational unit, since a single observation made on a school reveals the survey variables: the opportunities offered by the school's curriculum to all its students.

This issue arises whenever the group characteristic of each unit is observed and assigned a single value, which is associated with all elements comprising the group. The researcher is usually interested in the effect of the unit characteristic on individual elements. Nevertheless, many researchers have assigned equal weights to the observational units automatically and mistakenly, as if these were simple population elements. In my experience, once a researcher suffering from this confusion has been shown the difference, he will prefer to say, for instance, "High schools without physics courses account for 2 per cent of students," rather than "Half of the high schools offer no physics."

Several aspect of this problem are illustrated by the differences between the *group mean* $\bar{Y}_g$ and the *element mean* $\bar{Y}_e$ in the following situations.

a) We want to estimate the prevalence of swimming pools in the high schools of a state. $\bar{Y}_g$ per cent of the schools have swimming pools, but $\bar{Y}_e$ per cent of the students go to schools with pools. $\bar{Y}_e$ is considerably larger than $\bar{Y}_g$, because large schools have pools more often than small schools.

b) In a national voluntary organization $\bar{Y}_g$ per cent of the branches are in large metropolitan areas, but $\bar{Y}_e$ per cent of the members come from those areas. $\bar{Y}_e$ is much larger than $\bar{Y}_g$ because the branches have many more members in such areas. Most of us would prefer $\bar{Y}_e$ to $\bar{Y}_g$ as a measure of the extent to which this organization is metropolitan.

c) To forecast industrial employment and mobility in a state, the heads of a sample of manufacturing plants are interviewed regarding their plans to expand, to stay in the state, or move out of it. The results can be presented in terms of $\bar{Y}_g$ per cent of plants; but normally data for plants that account for $\bar{Y}_e$ per cent of employees will be more useful. The two means can diverge widely, because large and small plants differ.

d) In a certain industry $\bar{Y}_g$ per cent of firms operate with a specified type of organization (or leadership, or safety measures), and $\bar{Y}_e$ per cent of the employees are subject to it. I would generally prefer $\bar{Y}_e$ for measuring the prevalance of a given type of organization.

e) To estimate the prevalence of museums in the cities of a country, one can choose between $\bar{Y}_g$, the proportion of cities with museums, and $\bar{Y}_e$, the proportion of people living in cities with museums.

These are actual issues I have encountered repeatedly in situations concerned with significant social research. In each case the researcher chose the element mean after the difference was explained to him. On the other hand, analyses based on the group mean often occur in the social science literature. I suspect that in most such instances the issue was not faced squarely, so that the group mean was chosen automatically although the element mean would have been more relevant. Of course, in some situations the researcher may reasonably prefer and deliberately choose $\bar{Y}_g$ over $\bar{Y}_e$.[1]

WEIGHTED VERUS UNWEIGHTED MEANS

Suppose that a population comprises A units, and that $\bar{Y}_a$ is some value of the a-th unit. From the values $\bar{Y}_a$ of all A units in the population one can compute either the unweighted mean of the units, the *group mean,*

$$\bar{Y}_g = \frac{1}{A}\sum_a^A \bar{Y}_a, \qquad (1)$$

or the *element mean,*

$$\bar{Y}_e = \frac{1}{A}\Sigma\frac{N_a}{\bar{N}}\bar{Y}_a = \Sigma\frac{N_a}{\Sigma N_a}\bar{Y}_a = \frac{\Sigma N_a\bar{Y}_a}{\Sigma N_a}, \qquad (2)$$

[1] See discussion in Appendix 1 of Peter Townsend, *The Last Refuge,* London: Routledge and Kegan Paul, 1962.

using the numbers of elements in the units as weights. $\bar{N}=\Sigma N_\alpha/A$ is the mean unit size. Both means may be regarded as special cases of weighted means $\bar{Y}_w=\Sigma W_\alpha \bar{Y}_\alpha/A$, where $\Sigma W_\alpha=A$. For the group mean, $W_\alpha=1$ is assigned arbitrarily to all units, while the element mean utilizes the weights $W_\alpha=N_\alpha/\bar{N}$.

I have centered the discussion around the numbers of elements used as weights, but the gist of the argument is equally relevant when the N_α represent any measure of importance or size for the units comprising a population. *Choice of a measure of importance or size is a substantive problem for the researcher.*

Note that $N_\alpha \bar{Y}_\alpha=\sum_{\beta}^{N_\alpha} Y_{\alpha\beta}=Y_\alpha$ can be considered the aggregate of a variable $Y_{\alpha\beta}$ which has the same value $Y_{\alpha\beta}=\bar{Y}_\alpha$ for all N_α values in the $_\alpha$-th unit, Conversely, the unit aggregate Y_α may be considered divided into N_α equal portions $Y_\alpha/N_\alpha=\bar{Y}_\alpha$ among the N_α elements of the α-th unit. Viewed as an element variable, $\bar{Y}_\alpha$ represents complete uniformity and homogeneity of the variable $Y_{\alpha\beta}$ within units. From this view the element mean is simply the mean of the $\Sigma N_\alpha=N$ elements in the population:

$$\bar{Y}_e=\frac{\Sigma N_\alpha \bar{Y}_\alpha}{\Sigma N_\alpha}=\frac{Y}{N}. \qquad (2')$$

Although (2) and (2′) are equivalent, I shall stress the form (2) of a weighted mean of units in discussing the choice between an unweighted group mean $\bar{Y}_g$ and a weighted element mean $\bar{Y}_e$ to represent a population of A units.

The difference between the two means is

$$\bar{Y}_e-\bar{Y}_g=\frac{1}{A}\sum_{\alpha}^{A}\frac{N_\alpha}{\bar{N}}\bar{Y}_\alpha-\bar{Y}_g$$
$$=\frac{1}{\bar{N}}\Big[\frac{1}{A}\sum_{\alpha}^{A}N_\alpha\bar{Y}_\alpha-\bar{N}\bar{Y}_g\Big]=\frac{1}{\bar{N}}\mathrm{Cov}(N_\alpha,\bar{Y}_\alpha)$$
$$=\frac{1}{\bar{N}}R\sigma_n\sigma_y=\bar{Y}_g R C_n C_y. \qquad (3)$$

Here $\mathrm{Cov}(N_\alpha, \bar{Y}_\alpha)$ and $R=R_{ny}$ represent the covariance and the correlation coefficient between the variables N_α and $\bar{Y}_\alpha$. The symbols σ_n and σ_y denote standard deviations, and C_n and C_y denote coefficients of variation for the two variables. Thus $\sigma^2_y=\sum_{\alpha}(\bar{Y}_\alpha-\bar{Y}_g)^2/A$ and $C_y=\sigma_y/\bar{Y}_g$.

The two means, $\bar{Y}_e$ and $\bar{Y}_g$, can differ emphatically when R and C_n are large. This occurs typically when the distribution of the sizes N_α is very skewed, so that a small proportion of units accounts for a large proportion of elements, and the variable $\bar{Y}_\alpha$ is strongly correlated with the sizes N_α. The element mean $\bar{Y}_e$ is greater or less than the group mean $\bar{Y}_g$, according to a positive or negative correlation R_{ny}.

All the examples of the first section dealt with group values $\bar{Y}_\alpha$ that are dichotomous: a school either has physics courses or not; it either has swimming pools or not; a city either has museums or not. Possession of the characteristic is denoted with $\bar{Y}_\alpha=1$, and lack of it with $\bar{Y}_\alpha=0$. Dichotomous variables are common in social research (probably too common) and they readily illustrate the large correlations R_{ny} that may exist between sizes N_α and the group values $\bar{Y}_\alpha$ of the units. For dichotomous variables the group mean is a proportion, $\bar{Y}_g=P_g$, and its variance is $\sigma^2_y=P_g(1-P_g)$.

The entire argument is equally valid, however, for the frequent case when $\bar{Y}_\alpha$ is not dichotomous. For example, $\bar{Y}_\alpha$ may be the number of different physics courses offered by a school; or it may be the average number of physics courses taken by the school's students. For the branches of a voluntary organization, $\bar{Y}_\alpha$ may denote the size of the city, rather than an arbitrary "metropolitan" division; or it may denote the mean income of its members. For cities $\bar{Y}_\alpha$ may denote the number of museums, or the per capita attendance.

Proper representation of the group values $\bar{Y}_\alpha$ is a primary issue in many areas of social and economic research. The sources of group values vary. They may represent simply the means of individual values: for example, the mean income or the proportion of home owners, as characteristics of cities. They may be group values arising less directly from individuals: for example, the population sizes and densities of cities. Or, the values may belong specifically to the group, without direct effect from its elements: for example, the climate, altitude, or age of a city, its form of government, or the presence of museums. Whatever the origin of the group values, the choice between the element mean $\bar{Y}_e$ and the group mean $\bar{Y}_g$ should depend primarily on which conveys a more meaningful summary value.

Issues concerning the origin of group values are important, and more empirical studies should be made of measurements performed directly on social groups and organizations, rather than relying on the better known methods of measuring and averaging individual values. But methods of ascertaining and assigning group values are not my subject here. Group values may originate in measurements either on individuals (elements) or directly on groups; and the same measurements may also serve as individual values. Whatever its origin, a measurement can always serve in the double capacity of a group value and a value for each of the individuals comprising the group. A city may either be old or have citizens with a high average age. Its inhabitants can also be characterized either by the venerable age of their city or by the high mean age of its citizens. Four types of variables result from the two origins and the two uses of the data.

The use of variables for either group values or individual values must be distinguished from the origins of their measurements.[2] Origin refers to the *units of observation;* use refers to the *units of analysis,* which are often called the elements. Distinct from both is the issue of *sampling units,* which I shall discuss in the next section.

Analysis based on individual elements may utilize both element values and group values.[3] For example, one may correlate a personal health measure with city altitude, or a personal attitude measure with the city's latitude. Of course, both element and group variables may represent vectors of many variables.

When groups are the units of analysis, group values can be freely used, regardless of the origins of their measurement. For example, the mean of the element values within units becomes a single value for each unit. This may be denoted as $\bar{Y}_\alpha$ and used as a group value. But it may also be used as an individual value for all elements comprising the unit. The mean income of a city may be used to characterize either the city or each of its residents.

The terms group, organization, and unit are not distinguished here and are used interchangeably. I assume that each population element belongs to one and only one unit, that the population has been partitioned into a defined set of units without overlaps or gaps. The main argument can be extended to overlapping memberships of elements belonging to several units; but this extension would detract from the simplicity of the presentation.

In this paper I deal with means, and comparison of means follows readily. More complex measures of association, such as regression, would require separate treatment, and the structural model would have to be specified in each case.

SAMPLING FOR GROUP VALUES

Choice of the proper mean should precede the sample design, since the design depends on which mean is more appropriate. We should first ask ourselves: If we had all A values of N_α and $\bar{Y}_\alpha$ for the entire population, which mean would we choose? Since the element mean will most often be chosen, I shall consider samples designed for it.

The sampling problem resembles that of selecting entire clusters of unequal sizes—but with some special features. First, evaluation of $\bar{Y}_\alpha$ requires only a single observation on each cluster. Second, the homogeneity is extreme, because $\bar{Y}_\alpha$ is the same for all elements within the units; the intraclass correlation is perfect. Third, variations in the unit sizes (N_α) must be accepted as given, because the units are fixed entities; they cannot be divided by subsampling.

[2] One area of confusion has been well noted: William S. Robinson, "Ecological Correlations and the Behavior of Individuals," *American Sociological Review,* 15 (1959), pp. 351–357; Leo A. Goodman, "Some Alternatives to Ecological Correlation," *American Journal of Sociology,* 64 (1959), pp. 610–625. See also James A. Davis, Joe L. Spaeth and Carolyn Hudson, "Analyzing Effects of Group Composition," *American Sociological Review,* 26 (1961), pp. 215–225.

[3] "In terms of the actual analysis the matter can be restated in the following terms: just as we can classify people by demographic variables or by their attitudes, we can also classify them by the kind of environment in which they live." Some of these issues are discussed by Patricia L. Kendall and Paul F. Lazarsfeld, "Problems of Survey Analysis," in Robert K. Merton and Paul F. Lazarsfeld (eds.), *Continuities in Social Research,* Glencoe, Ill.: The Free Press, 1950.

Two further special assumptions are implicit in the following brief statistical treatment. First, that the variance of $\bar{Y}_\alpha$ is similar among units, regardless of unit size. Second, that the unit cost of obtaining $\bar{Y}_\alpha$ is similar for large and small units. Compared with the variations in the sizes N_α, differences in variance and cost factors often can be reasonably assumed small. When they are not, disproportionate allocation may be introduced.

If a units are selected with equal probability a/A, the simple mean of the sample $\bar{y}_{gs}=\sum_{\alpha}^{a}\bar{y}_\alpha/a$ is an unbiased estimate of $\bar{Y}_g$.[4] To estimate the element mean $\bar{Y}_e$, one must introduce the weights N_α and use the ratio mean

$$\bar{y}_{es}=\sum_{\alpha}^{a}N_\alpha\bar{y}_\alpha/\sum_{\alpha}^{a}N_\alpha. \qquad (4)$$

Its variance merely applies a formula for a ratio mean, appropriate to the method used for selecting the a units in the sample. If the N_α are grossly unequal, the weights will render this estimate inefficient: a few large selections will tend to dominate the estimate and its variance. In this extreme situation, selection with *probabilities proportional to size* (PPS) seems particularly appropriate.

If the units are selected with probabilities proportional to the N_α, an unbiased estimate of the sample is the simple mean

$$\bar{y}_{ep}=\frac{1}{a}\sum_{\alpha}^{a}\bar{y}_\alpha. \qquad (5)$$

This demonstrates the chief advantage of selecting units with PPS to estimate the element mean: the ordinary self-weighting mean of sample observations is simple to compute, and is efficient under the special assumptions described above. If a selections have been drawn with replacement from the entire population, the variance of the sample mean may be estimated simply as

$$\text{var}(\bar{y}_{ep})=\frac{1}{a(a-1)}\sum_{\alpha}^{a}(\bar{y}_\alpha-\bar{y}_{ep})^2. \qquad (6)$$

This can readily be perceived by considering a selections drawn with replacement from a population of N elements, each with its own value $\bar{Y}_\alpha$. In this population N_1 elements have the value $\bar{Y}_1$, N_2 have $\bar{Y}_2$, N_α have $\bar{Y}_\alpha$, etc. To retain a simple presentation, I assume that the units are selected with replacement; if a unit is selected two or more more times, it remains in the sample and in the estimate as often as it was selected.[5]

We typically use stratification in these situations, resorting to formulas appropriate for stratified samples. Suppose that the units have been sorted into H strata, and that a_h units have been selected with PPS from the h-th stratum. The weight of the stratum is W_h, which ordinarily represents N_h/N, the proportion of elements it contains, and $\sum_h W_h=1$. The element mean is estimated by

$$\bar{y}_{ep}=\sum_{h}^{H}W_h\bar{y}_h,\text{ where }\bar{y}_h=\frac{1}{a_h}\sum_{\alpha}^{a_h}\bar{y}_{h\alpha}. \qquad (7)$$

Assuming that the a_h selections in the strata were made with replacement, the estimate of the variance is

$$\text{var}(\bar{y}_{ep})=\sum_{h}^{H}\frac{W^2{}_h}{a_h(a_h-1)}\sum_{\alpha}^{a_h}(\bar{y}_{h\alpha}-\bar{y}_h)^2. \qquad (8)$$

The sample can be made self-weighting if the number of selections is made proportional to the statum sizes ($a_h=kW_h$). Then for a total sample size of $a=\sum a_h$ we have

$$\bar{y}_{ep}=\frac{1}{a}\sum_{h}^{H}\sum_{\alpha}^{a_h}\bar{y}_{h\alpha}, \qquad (9)$$

and

$$\text{var}(\bar{y}_{ep})=\frac{1}{a^2}\sum_{h}^{H}\frac{a_h}{a_h-1}\sum_{\alpha}^{a_h}(\bar{y}_{h\alpha}-\bar{y}_h)^2. \qquad (10)$$

If $H=a/2$ strata of equal sizes are formed, and paired selections (1 and 2) are drawn from each stratum, the last formulas become

$$\bar{y}_{ep}=\frac{1}{a}\sum_{h}^{H}(\bar{y}_{h1}+\bar{y}_{h2})$$

$$\text{and var}(\bar{y}_{ep})=\frac{1}{a^2}\sum_{h}^{H}(\bar{y}_{h1}-\bar{y}_{h2})^2. \qquad (11)$$

[4] Leslie Kish, *Survey Sampling,* New York: John Wiley, 1965, Section 2.8.

[5] If unconvinced by this argument, see another in William G. Cochran, *Sampling Techniques* (2nd ed.), New York: John Wiley, 1963, Section 9.10.

Some large units may be larger than N/a, the designed population size per selection, and some almost as large. These should be taken into the sample with certainty; they will not contribute to the variance. At the other extreme, the strata may be filled with many units, each much smaller than the strata; for these it will matter little whether selection is with or without replacement, because the finite population correction (1-f) may be disregarded in computing the variance. In between, the strata may contain units somewhat, but not much, smaller than the strata, where the factor (1-f) should be considered.

Suppose that after selecting with probabilities proportional to the size measures N_α, one finds different desired "true" sizes n'_α for the selected units. The difference may be due to changes in size, discrepancies in the units of measurement, etc. Although the n'_α and N_α should be highly correlated, the differences may not be negligible; for example, one could underrepresent fast-growing units. Then weight each $\bar{y}_\alpha$ in the sample with $\bar{x}'_\alpha = n'_\alpha/N_\alpha$ and, using $\bar{y}_\alpha = \bar{x}'_\alpha \bar{y}_\alpha$, compute the ratio mean

$$\bar{y}_{ep} = \sum_\alpha \bar{y}'_\alpha / \sum_\alpha \bar{x}'_\alpha \qquad (12)$$

For stratified selection, and with the selections proportional to stratum sizes, the ratio mean becomes

$$\bar{y}_{ep} = \sum_h \sum_\alpha \bar{y}'_{h\alpha} / \sum_h \sum_\alpha \bar{x}'_{h\alpha}. \qquad (13)$$

Here $\bar{y}'_{h\alpha} = \bar{y}_{h\alpha} \bar{x}'_{h\alpha} = \bar{y}_{h\alpha} n'_{h\alpha}/N_{h\alpha}$, the variable corrected for the change in desired size from the measure of size. The variances of this ratio mean may be found in several text books.

With proper weighting, one can also estimate the group mean $\bar{Y}g$, when required. If the units were selected with probabilities proportional to $N\alpha$, each selected mean should be weighted with $1/N\alpha$. The ratio mean for a proportionate stratified selection would be

$$\bar{y}_{gp} = \sum_h \sum_\alpha (\bar{y}_{h\alpha}/N_{h\alpha}) / \sum_h \sum_\alpha 1/N_{h\alpha}, \qquad (14)$$

and its variance is that of a stratified ratio mean.[6]

Let us recapitulate the estimation of the group mean $\bar{Y}_g$ and the element mean $\bar{Y}_e$ from observations made on a sample of group values $\bar{y}_\alpha$. If the groups are selected with equal probabilities, the simple mean estimates $\bar{Y}_g$; but to estimate $\bar{Y}_e$, the sample values $\bar{y}_\alpha$ should be weighted by their sizes N_α. On the contrary, if groups are selected with PPS, the simple mean estimates $\bar{Y}_e$; but to estimate $\bar{Y}_g$, the values $\bar{y}_\alpha$ should be weighted with $1/N_\alpha$.

FURTHER SAMPLING CONSIDERATIONS

Selecting groups with PPS is particularly well suited to studies in which both individuals and the groups to which they belong are used as units in separate analyses. For example, in a study of a large organization some of the results concern individual members, others deal with the groups (units, branches) of the organization. By selecting groups with probability proportional to their size measures, and then subsampling elements within groups with probabilities inversely proportional to the same measure, one obtains an equal-probability selection of elements, also an equal number of elements per unit. The simple mean of individual values estimates their population mean $\bar{Y}$; and the simple mean of the sample group values

TABLE 1. ESTIMATES OF THE GROUP MEAN $\bar{Y}_g$ AND THE ELEMENT MEAN $\bar{Y}_e$ FROM SAMPLES SELECTED WITH EQUAL PROBABILITY AND WITH PROBABILITEIS PROPORTIONAL TO SIZE (PPS)

Selection Probabilities	Simple mean		Weighted Mean	
Equal: a/A	$\bar{y}_{gs} = \sum_\alpha^a \bar{y}_\alpha / a$	$E(\bar{y}_{gs}) = \bar{Y}_g$	$\bar{y}_{es} = \sum_\alpha^a N_\alpha \bar{y}_\alpha / \sum_\alpha^a N_\alpha$	$E(\bar{y}_{es}) = \bar{Y}_e$
PP: $aN_\alpha/\Sigma N_\alpha$	$\bar{y}_{ep} = \sum_\alpha^a \bar{y}_\alpha / a$	$E(\bar{y}_{ep}) = \bar{Y}_e$	$\bar{y}_{gp} = \sum_\alpha^a \frac{\bar{y}_\alpha}{N_\alpha} / \sum_\alpha^a \frac{1}{N_\alpha}$	$E(\bar{y}_{gp}) = \bar{Y}_g$

[6] The preceding arguments and formulas can be found in sampling texts. See, for example, Ch. 6 and 7 in Kish, *op. cit.*

$\bar{y}_a$ estimates the element mean $\bar{Y}_e$ of the group values.[7]

Note, however, that selecting all N_a elements (or any constant fraction) from groups selected with PPS would not yield an equal probability of selection for individuals, nor a self-weighting mean for estimating $\bar{Y}$.

Subsampling with PPS is also generally efficient when the group values are computed as means of the n_a individuals selected from the group, $\bar{y}_a = \sum_{\beta}^{n_a} y_{a\beta}/n_a$. Because the group values $\bar{y}_a$ are based on (approximately) equal numbers n_a of elements, they are subject to approximately equal variances. This typically enhances the efficiency of the statistical analysis of group values. These may be related to, and analyzed together with, other group values which can be measured directly on the groups.

Joint analysis of group variables $\bar{Y}_a$ and individual variables $X_{a\beta}$ is possible. Each population element possesses both kinds of variables. A sample of elements permits their joint analysis.

Results are often tabulated for domains defined by size classes of units. Researchers tend to define domains roughly equal in numbers of elements, because size is a measure of the domain's importance. The domains of large units typically contain fewer units than domains comprised of numerous small units. Under these conditions PPS selection has another advantage over selection with equal probabilities for all units. If the latter is used, the domains of large units, though important in numbers of elements, contain few units and receive very few selections. With PPS selection the several domains receive equal numbers of selections to the degree that they contain equal numbers of elements. These domains can also serve as strata. If the number of selections in domains based on size are proportional to their element sizes, and if variations in size are small within the strata, the efficiency of selecting with equal probabilities within strata will be roughly equal to that of selecting with PPS.

Suppose that one has selected with equal probability ($f = n/N$) a sample of n persons directly from a population of N individuals, without using the groups in the selection. Now, suppose that this sample of individuals is also used to obtain the group variables $\bar{y}_a$ concerning the groups to which they belong. For example, from a selection of n individuals one can obtain characteristics of the family, or county, or university to which they belong. For a group of size N_a, the expected representation in the sample is fN_a; the actual size n_a from the group is a random variable. The simple mean of the values of $\bar{y}_a$ taken for all the n sample elements will be an unbiased estimate of the element mean $\bar{Y}_e$. To estimate the group mean $\bar{Y}_g$ one should use the weights $1/N_a$ in the estimate $\bar{y}_g = \sum(\bar{y}_a/N_a)/\sum_a(1/N_a)$, summed over the n sample cases.

For example, a sample of n voters can be selected with equal probabilities from an area selection of counties and segments. Suppose one ascertains for each sample voter a value $\bar{y}_a$ about the Congressional District to which he belongs; for example, the vote of his Congressman on a bill. The simple mean of the n voters will estimate the element mean $\bar{Y}_e$ for Congressional Districts; in this mean each District is weighted by $N_a/\bar{N}$, with N_a representing the number of voters. If a somewhat different weight n'_a is preferred, formulas (12) and (13) should be used. To estimate the simple mean $\bar{Y}_g$ of Congressional Districts one can weight the values of sample voters with $1/N_a$. But it would be a mistake to accept the simple mean of all group values that happen to appear in the sample, whether once or several times. In a national area sample, Districts from large cities tend to come into the sample with high probabilities but low n_a; the reverse is the case for rural Districts with sparse populations.

A SPECIAL CASE

A curious special case of the unit variable $\bar{Y}_a$ occurs when it represents the unit size N_a. Suppose, for example, we were to ascertain for each element in the population the size of his household (or his city, or

[7] *Ibid.*, Section 7.3.

his organization). For the variable N_a the element mean is

$$\bar{Y}_e = \frac{\Sigma N_a N_a}{\Sigma N_a} = \frac{\Sigma N^2{}_a/A}{\bar{N}} \qquad (15)$$
$$= \frac{\sigma^2{}_n}{\bar{N}} + \bar{N} = \bar{N}(C^2{}_n + 1).$$

This also follows from (3) for the special case when $N_a = \bar{Y}_a$; hence $\bar{Y}_g = \bar{N}$, $R = 1$, and $C_n = C_y$. (The variance can be computed, as for any ratio mean, for the variables $N^2{}_a$ and N_a.) The variable N_a can be obtained from an equal probability selection from the $\Sigma N_a = N$ population elements; the sample mean will estimate $\bar{Y}_e$. Clearly when the relative variance $C^2{}_n$ of unit sizes is great, this mean, sometimes called the *"contraharmonic mean"* of the unit sizes N_a, is much larger than the simple arithmetic mean $\bar{N}$. In other words, the variable representing the size of the unit to which elements belong has a larger mean than the mean size of the units. Although the mean number of adults per household is only 2.02, the mean number of household members is 2.24 for the average adult. The greater size ranges of large organizations produce more striking effects. In 1960, 50 million people lived in 130 U. S. cities that had 100,000 or more population; in this population, the average city size was 0.39 million, but the size of the city in which the average person lived was 2.0 millions. Using medians does not help: the median city size was 0.19 million, but the median person lived in a city of 0.62 million.

If we ask people, "How many siblings do you have?", the answer will be $N_a - 1$; the mean of these is one less than the mean of N_a, or $\bar{N}C^2{}_n + \bar{N} - 1$.

BIAS VERSUS VARIANCE IN GROUP ANALYSIS

Frequently our main research interest is in units rather than elements. The units of analysis are not persons, but cities, schools, or some other group, organization, or ecological unit. It is a common mistake in these situations to use $\bar{Y}_g$ automatically. Use of $\bar{Y}_g$ can only be justified as the result of an appropriate and deliberate choice of the specific set of uniform weights, $1/A$. In most cases, however, I prefer $\bar{Y}_e$ to $\bar{Y}_g$, and the weights $N_a/\bar{N}$ appear as most appropriate, or as reasonable approximations.

These ideas lead to selection with probabilities proportional to size (PPS) when a sample of units is needed. However, this method is subject to a severe test when the research is based on a relatively small number A of units, say a few dozen or hundred, that comprise an entire population. If data are readily available for all A units, sampling then would waste information.

Whenever the A unit values $\bar{Y}_a$ represent all the necessary information without sampling error, the issue is simple. If $\bar{Y}_e$ is appropriate and $\bar{Y}_g$ is used in its place, the bias is known from (3):

$$\bar{Y}_g - \bar{Y}_e = \bar{Y}_g R\, C_n C_y. \qquad (16)$$

This can be large, and it should be avoided if no reasonable countervailing arguments arise. If the A values represent a sample from a larger universe, however, the bias should be balanced against sampling errors. It may well be good research strategy to regard the A units as a simple random sample from a hypothetical infinite universe of such units. Then the variance of the element mean may be written[8] as approximately

$$\mathrm{Var}\Big(\frac{1}{A}\sum_a^A \frac{N_a}{\bar{N}} \bar{Y}_a\Big) = \frac{\bar{Y}^2{}_g}{A}[C^2{}_y + C^2{}_n + 2RC_nC_y]; \qquad (17)$$

or

$$\mathrm{Var}(\bar{Y}_e) = \mathrm{Var}(\bar{Y}_g) + \frac{\bar{Y}^2{}_g}{A}[C^2{}_n + 2RC_nC_y]. \qquad (18)$$

The second term expresses the excess of variance of $\bar{Y}_e$ over that for $\bar{Y}_g$; it will be positive when $C_n > 2RC_y$. We should balance this against the bias of $\bar{Y}_g$. This can be done most readily by comparing the two

[8] The variance for a single random unit can be obtained from the formula for the product of two random variables:

$$\sigma^2(N_a\overline{Y}_a) = \overline{N}^2\sigma^2{}_y + \overline{Y}_g\sigma^2{}_n + 2\overline{NY}_g R\sigma_n\sigma_y.$$

For the mean of A random selections then, and after dividing by the mean values $\bar{Y}_g$ and $\bar{N}$ of the variables $\bar{Y}_a$ and N_a, we get (17). See *ibid.,* Section 6.6.D.

mean-squared-errors; the mean-squared-error equals Variance+Bias². Hence

$$\begin{aligned}
&\mathrm{MSE}(\bar{Y}_g) - \mathrm{MSE}(\bar{Y}_e) \\
&\quad = [\mathrm{Var}(\bar{Y}_g) + \mathrm{Bias}^2(Y_g)] - \\
&\qquad [\mathrm{Var}(\bar{Y}_e) + 0] \\
&\quad = [\mathrm{Var}(\bar{Y}_g) + \bar{Y}^2{}_g(RC_nC_y)^2] - \\
&\qquad [\mathrm{Var}(\bar{Y}_g) + \frac{\bar{Y}^2{}_g}{A}[C_n{}^2 + 2RC_nC_y] \\
&\quad = \bar{Y}^2{}_g[(RC_nC_y)^2 - \frac{1}{A}(C^2{}_n + 2RC_nC_y)].
\end{aligned} \tag{19}$$

The increase in the variance of $\bar{Y}_e$ decreases with A, the number of units, hence the bias of $\bar{Y}_g$ tends to predominate unless the correlation R is very small. If the correlation is small, the bias may become negligible, and the term $C^2{}_n/A$ can predominate in (19); but in this case the two means and their variances will be similar. If A is very small and the $C^2{}_n$ large, $\bar{Y}_g$ may appear more reliable than $\bar{Y}_e$; but for a very small "sample," other issues of inference also arise.