Confidence Intervals for Clustered Samples

Author(s): Leslie Kish

Source: *American Sociological Review*, Apr., 1957, Vol. 22, No. 2 (Apr., 1957), pp. 154-165

Published by: American Sociological Association

Stable URL: http://www.jstor.com/stable/2088852

REFERENCES
Linked references are available on JSTOR for this article:
http://www.jstor.com/stable/2088852?seq=1&cid=pdf-reference#references_tab_contents
You may need to log in to JSTOR to access the linked references.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms

American Sociological Association and *Sage Publications, Inc.* are collaborating with JSTOR to digitize, preserve and extend access to *American Sociological Review*

This content downloaded from
132.174.254.12 on Wed, 10 Jun 2020 21:48:30 UTC
All use subject to https://about.jstor.org/terms

(above) becomes phase 1 in the second research cycle.

The advantages of this procedure may be summarized as follows:

1. It minimizes the likelihood that the investigator will present to himself and the world a prematurely coherent set of propositions in which contradictory propositions, however plausible, are ignored. The procedure requires that a range of contradictory propositions be made explicit before the empirical investigation, and it should conduce to theoretical humility after the fact. Projects are less likely to culminate in particularistic interpretations in terms of a single theory.

2. This procedure increases the likelihood of the researcher's building into his research-design provisions for the test of a variety of theoretical interpretations of a number of possible empirical findings.

3. The proposed procedure makes the researcher more aware of the total significance of his empirical findings. Where, as in the usual procedure, the investigator is concerned with upholding or refuting a particular theory, he may be completely unaware of the fact that his empirical findings actually add confirmation, or doubt, as the case may be, to numerous other theoretical propositions extant in the area or in related areas.

4. The procedure employed here makes for continuity of research. At the present state of theoretical development in sociology it is rare that clear-cut, unambiguous interpretations are possible after a single research project. Where there is little in the way of theory to guide interpretation, subsequent empirical investigations of alternative interpretations are often necessary. Where alternate interpretations are made explicit from the beginning of the project, they are more likely to survive as alternatives after the fact of investigation: the present procedure encourages research programs rather than isolated projects.[8]

[8] E.g., in the case of present research program, projects are now underway which will enable us to select with greater confidence a narrower range of interpretations. The presupposed empirical relationships of these subsequent projects existed as alternative interpretations in previous phases.

CONFIDENCE INTERVALS FOR CLUSTERED SAMPLES *

Leslie Kish

University of Michigan

Standard statistical literature has been developed almost entirely in terms of independent observations obtained by simple random sampling (s.r.s.). The statistical tests and confidence intervals one finds in textbooks and in journal articles are based on the assumption that the sample observations were selected independently, at random. Although this assumption either appears (unnoticed) in "fine print" or not at all, it constitutes the basis of all confidence intervals and tests of hypotheses; their validity is based on it. The ubiquity of the s.r.s. assumption in statistical theory can be explained by the basic nature of s.r.s. and by its facilitation of interesting theoretical results.

On the other hand, most social research, especially survey work, is in fact carried out by means of complex sample designs. Simple random selection of human populations is a rare phenomenon, limited to small and confined populations. Socially important human populations are usually large and widely scattered, so that a simple random selection of them would prove in most cases to be uneconomical and impractical. The complexities of good sample design occur in the attempt to tailor economically the available research resources to the wide and irregular dispersal of important populations.

Thus the complexities of sampling problems are often reflected openly and clearly in sample design; but important statistical consequences of complex designs are often ignored. By complexities I mean departures from s.r.s. By designs or sample designs I

* Presented in part at a joint session of the American Psychological Association and the Psychometric Society, Chicago, September, 1956.

This content downloaded from
132.174.254.12 on Wed, 10 Jun 2020 21:48:30 UTC
All use subject to https://about.jstor.org/terms

mean the combined effects of selection and estimation either in a survey or an experiment. The problems caused by complex design, by the lack of independence in selection, may also arise in experimental design. Although the arguments are developed chiefly in terms of the problems of sample survey design they apply as well to experimental design.[1]

The principal methods of departure from simple random selection are clustering, stratification, unequal probabilities of selection and systematic sampling. Among these methods clustering most frequently causes the largest effects. "Clustering" is the selection of observational units (elements, cases, individuals) in clusters or groups, rather than individually. The researcher uses these clusters as sampling units to make the procedures of selection and/or observation more economical and convenient. For example: (1) In studies of student attitudes and behavior, classes of students are selected into the sample. (2) In national interview surveys, counties, blocks and households frequently serve as clusters of persons. (3) In researches dealing with employees of a firm, work groups are used as sampling units. In each of these situations the sampling units comprise clusters of the individuals who form the basic elements of the investigation for drawing inferences from the sample to the population.

The researcher generally finds it economical and convenient to use for his sampling units existing clusters (blocks, cities, counties, work groups, platoons, school classes, etc.). These clusters usually exist as ecological, perhaps geographical units, and sometimes as "psychological groups" as well. The individuals in these units tend to resemble each other—there is usually some *homogeneity* of characteristics, of attitudes, of behavior—but homogeneity is generally not complete. It may be due to common selective factors, or to joint exposure to the same effects, or to mutual influence (interaction), or to some combination of these three causes.[2] Because of this homogeneity, the use of these clusters for sampling units has definite consequences: it destroys the independence of the characteristics of the sample elements. The correspondence with the "well-mixed urn," inherent in the assumption of independence, is negated; and formulas that depend on that assumption fail to apply.

This problem does not exist if the units or groups (independently and randomly selected) are themselves the elements of the analysis. It arises only when the sampling units used in the selection process are clusters of the elements that constitute the units of analysis. As a consequence there are often large mistakes in the construction of confidence intervals and of tests of hypotheses.

NATURE AND MAGNITUDE OF THE PROBLEM

Let us suppose that we have a body of data obtained from a complex sample design,

[1] In survey research the complexities of sample design have recently received explicit recognition and theoretical development in accord with the explicit definitions of the population from which the sample is drawn. There are now five fine textbooks dealing with the methods of survey sampling: W. G. Cochran, *Sampling Techniques,* New York: Wiley and Sons, 1953; W. E. Deming, *Some Theory of Sampling,* New York: Wiley and Sons, 1947; M. H. Hansen, W. N. Hurwitz and W. G. Madow, *Sample Survey Methods and Theory,* New York: Wiley and Sons, 1953; P. V. Sukhatme, *Sampling Theory of Surveys with Applications,* Ames: Iowa State College Press, 1954; F. Yates, *Sampling Methods for Censuses and Surveys,* New York: Hafner, 2nd. ed., 1953.

A brief, simplified treatment was attempted in Leslie Kish, "Selection of A Sample," Ch. 5 in Leon Festinger and Daniel Katz, editors, *Research Methods in the Behavioral Sciences,* New York: Dryden Press, 1953.

Experimental design has flourished in the numbers and complexities of the effects it undertakes to investigate simultaneously and efficiently, and in the intricate control of many variables. However, it has largely neglected until recently the joint problems of sampling and of specifying the population to which its inferences apply. See M. B. Wilk and O. Kempthorne, "Fixed, Mixed, and Random Models," *Journal of the American Statistical Association,* 50 (December, 1955), pp. 1144–1167.

[2] The homogeneity of individuals within units may be measured by *rho,* the coefficient of intraclass correlation; for most units and for most characteristics it is found to be positive. It measures that proportion of the total variance among the elements of a population which is "explainable" by belonging to the clusters. This homogeneity increases the variance of the mean, and of similar statistics, when compared with s.r.s. selections of the same number of elements (see below). Brief descriptions of the intraclass correlation may be found in Deming, *op. cit.;* in Hansen, Hurwitz and Madow, *op. cit.;* in Kish, *op. cit.;* and L. Kish, "Differentiation in Metropolitan Areas," *American Sociological Review,* 19 (August, 1954), pp. 391–392.

This content downloaded from
132.174.254.12 on Wed, 10 Jun 2020 21:48:30 UTC
All use subject to https://about.jstor.org/terms

in which the assumptions of independence have not been fulfilled. The usual textbook formulas that depend on these assumptions cannot then be applied validly to the data at hand. Is it dangerous to violate the assumption of independence? Some laxity in statistical rigor may arise from the knowledge of two facts. First, this assumption is very commonly disregarded. Second, most statistics involve other assumptions; these, too, are often disregarded. For example, moderate departures from normality generally have little effect on the confidence coefficients; the "finite population correction" $\sqrt{1-\frac{n}{N}}$ results in the negligible factor .99, even when the sampling rate $\frac{n}{N}$ is as large as $\frac{1}{50}$; using n insted of (n—1) in obtaining the sample variance will cause little trouble for samples as small as 50; and so on.

But we may not disregard the assumption of independence without serious consequences. In the social sciences the use of s.r.s. formulas on data from complex samples is now the most frequent source of gross mistakes in the construction of confidence statements and tests of hypotheses.[3] The mistakes are frequent because complex clustered samples provide much of the best research data in the social sciences. The mistakes are often large because the homogeneity among the elements of the cluster, ignored in the use of s.r.s. formulas for variances, often results in gross mistakes in the construction of confidence intervals.

Table 1 shows the nature of hypothetical mistakes made when s.r.s. formulas are used to compute confidence intervals for data from complex samples. Columns 1–3 give respectively, the amount of distortion of the confidence intervals due to using s.r.s. formulas when they do not apply: column 1, the ratio of the actual variance to s.r.s. variance; column 2, the inverse of this ratio; column 3, the square roots of the numbers of column 2, gives the ratios by which the standard errors (and the confidence intervals) are shortened by mistake if s.r.s. variances are used in place of the correct variance formulas. The results of the mistakes are given in columns 4–7. They show the probabilities of excluding the true population value for different levels of the confidence coefficient. For example, column 5 deals with the P=.95 level of confidence. The use of ±1.96 standard errors, computed correctly, allows for errors .05 of the time. The last row shows that if by mistake one is actually using 1.96×.32=.63 standard errors, then he makes errors .533 of the time.

TABLE 1. EFFECTS OF DISTORTED CONFIDENCE INTERVALS ON PROBABILITY STATEMENTS

Amount of Distortion of the Confidence Interval			Probabilities of Incorrect Statements When Two-Sided Confidence Coefficients Are Aimed At:			
$\frac{\text{True Var.}}{\text{s.r.s. Var.}}$	$\frac{\text{s.r.s. Var.}}{\text{True Var.}}$	$\frac{\text{s.r.s. S.E.}}{\text{True S.E.}}$	P=.90	P=.95	P=.99	P=.999
(1)	(2)	(3)	(4)	(5)	(6)	(7)
0.8	1.25	1.12	.064	.028	.004	.0001
1.0	1.00	1.00	.100	.050	.010	.001
1.2	.83	.92	.133	.074	.019	.003
1.5	.67	.82	.179	.110	.035	.007
2.0	.50	.71	.242	.164	.067	.019
3.0	.33	.58	.342	.258	.137	.057
5.0	.20	.45	.462	.381	.249	.141
10.0	.10	.32	.603	.533	.415	.298

Values of $\frac{\text{True Var.}}{\text{s.r.s. Var.}}$ of 1.2 to 2 are found frequently in well designed samples; mistaken use of s.r.s. formulas would lead to incorrect statements .074 to .164 of the time,

[3] A rival for this dubious honor is that other, somewhat more widely known, *bête noire:* hunting through masses of comparisons until one finds a rare "significant" difference, and then treating the "test of significance" as if the experiment had been designed explicitly for that test. This indictment may be extended to the biological and other sciences by persons more familiar with research in those fields.

This content downloaded from
132.174.254.12 on Wed, 10 Jun 2020 21:48:30 UTC
All use subject to https://about.jstor.org/terms

instead of the .05 level the researcher intended to use. But we have also found values from 3 to 10 times the s.r.s. variance; ignorance of the true formula would lead the researcher to substitute unknowingly for the .05 level of confidence some level from .258 to .533!

One may look at the effect of clustering in terms of the increase of the variance over that of a simple random sample of the same number of elements. This is approximately

$$\frac{\text{True Var.}}{\text{s.r.s. Var.}}=[1+\text{rho}(\bar{a}-1)]$$

where rho is the coefficient of intraclass correlation and $\bar{a}$ is the number of elements in the sample cluster. This formula also holds fairly well when $\bar{a}$ is an *average* cluster size. Neither a small rho nor a small $\bar{a}$ alone gives assurance that the factor of clustering may be neglected safely. With a rho as small as .04 a sample of infantry companies of size $\bar{a}=101$ would lead to a ratio of $\frac{\text{True Var.}}{\text{s.r.s. Var.}}=5$, hence to the distortion of the confidence level from .05 to .381. With work groups as small as $\bar{a}=6$, a rho of 0.2 results in a ratio of 2, hence in distorting the confidence level from .05 to .164.

The distortions caused by the improper use of s.r.s. formulas are presented in terms of confidence intervals. For tests of hypotheses there are similar effects in rejecting the null hypothesis too often.

How is it that a procedure that leads to such large mistakes is still widely practiced? One answer may lie in the fact that this procedure has been relatively successful in the the physical sciences and has consequently gained sanction. There may be three reasons why this matter has received less notice in physical sciences. (1) In some of the physical sciences (especially in those that serve as "models") the error due to uncontrolled causes is sometimes relatively unimportant. (2) Frequently s.r.s. selection or a good approximation may be used with ease. (3) When clustered samples are taken, the intraclass correlation of the units in the clusters is negligible. In other words, the characteristic to be measured is "fairly randomly" distributed in the clusters. For similar reasons the crudest judgment sampling methods have often yielded reliable results in the physical sciences—methods not acceptable in the social sciences, because of their potentially large biases. On the other hand, in social science research all of these three factors are often absent; hence, the effect of clustering must be considered for confidence intervals and tests of hypotheses to be valid.

EMPIRICAL EVIDENCE

The reader may wish to see the magnitude of the effects as they arise in some specific situations. We have computed the variances for a variety of survey characteristics over the past ten years in the Sampling Section of the Survey Research Center. Table 2 gives a few examples to illustrate some contrasting situations. Other organizations with good probability samples have had similar experiences. In particular the relases of the *Current Population Surveys* of the U. S. Census Bureau present interesting tables of sampling errors.

For Item 1 actual variances range in magnitude from the same as s.r.s. variance to 1.8 times s.r.s. variance. Hence the actual standard error is up to $\sqrt{1.8}=1.34$ times as large as that of a simple random sample of the same number (2000) of elements. If the researcher, ignoring the correct formula, used s.r.s. formulas, he would be using instead of the correct interval of 1.96, actual intervals as small as $1.96/1.34=1.46$. This would result in his statements being incorrect 14 per cent of the time instead of the 5 per cent level he wanted to use.

When pairs of subclasses are compared, and the variances are computed for the differences of the two means, the ratio of $\frac{\text{True Var.}}{\text{s.r.s. Var.}}$ is similar to those found for single means. Here too, the actual variance may be as much as 1.8 times the value that s.r.s. formulas would yield. In these comparisons, however, we find ratios as low as 0.8; at times the actual variance is a little less than s.r.s. formulas would indicate. At present we lack good theoretical justification for this result.

In Item 2 the subject matter is quite different from that in 1, and there are also changes in the numbers of different kinds of sampling units. (About 8000 dwellings had to be included to find 2900 respondents; the other 5100 contained no women of childbear-

This content downloaded from
132.174.254.12 on Wed, 10 Jun 2020 21:48:30 UTC
All use subject to https://about.jstor.org/terms

TABLE 2. EFFECTS OF COMPLEXITY OF DESIGN ON CONFIDENCE INTERVALS *

Items	Nature of the Characteristics and of the Population	Numbers of Sampling Units	Type of Estimates	No. of Tests Conducted	Ratio True var. / s.r.s. Var.	Confidence Levels
1	Attitudes and expectations on consumer items	2000 respondents in 2000 families, in 1000 segments, in 150 places, in 34 psu's †	Proportions based on 2000 and on subclasses of 150–1000	144	1 to 1.8	.05 to .14
	U. S. family heads and wives		Comparisons for pairs of subclasses	48	0.8 to 1.8	.03 to .14
2	Numbers of actual and expected children; family planning behavior	2900 women from 2900 dwellings, in 2000 segments, in 250 places, in 66 psu's †	Means and proportions based on 2700 and on subclasses of 400–1800	48	1 to 2.5	.05 to .21
	U. S. women of child-bearing age		Comparisons of the above for pairs of subclasses	24	0.7 to 2.5	.02 to .21
3	Travel behavior; frequencies of trips and choice of transportation	8500 adults in 4100 dwellings, in 2000 segments, in 250 places, in 66 psu's †	Proportions based on 8500	10	2 to 6	.16 to .42
	U. S. adults		Proportions based on subclasses of 300–3000	60	1 to 6	.05 to .42
			Comparisons of the above for pairs of subclasses	30	0.8 to 5	.03 to .38
4	Perceptions of supervisory practices	685 workers in 59 work sections	Proportions based on groups of 10 to 25 sections	15	1 to 5	.05 to .38
	Workers in public utility firm					
5	Perceptions of production norms and attitudes of satisfaction		Proportions based on entire population and on groups of sections	20	1 to 4	.05 to .33
	Production workers in large manufacturing company	1458 workers in 57 work sections				

* From Sample Surveys conducted by the Survey Research Center.

† "Psu's" are "primary sampling units," the first and largest units selected in the multi-stage sample. Here they were counties or groups of counties. The "places" were cities, towns or rural areas.

ing age, 18–40, as defined for the survey.) The effect of the complexities of the sample is similar to that found for Item 1, but somewhat larger. Again, the variances of differences show effects similar to the variances of single estimates.

Item 3 is an an example of the very large effects that high homogeneity can produce. For some items the travel behavior of the adults of the same household is very similar. Thus the presence of an average of 2.1 adults per household will have strong clustering effects—as compared with a similar number of adults selected independently. When 4100 respondents (one per household) were analyzed separately, the effects of the sample design appeared to be not much greater than for Item 2. (The information about the "other" adults in the household was not "worth" statistically and relatively very much per individual adult; however, the cost of this information per adult was very low

This content downloaded from
132.174.254.12 on Wed, 10 Jun 2020 21:48:30 UTC
All use subject to https://about.jstor.org/terms

too. Thus the information per dollar was still worthwhile.) Note that the comparisons of two means can be subject to very large effects too.

In Items 4 and 5 we examine the effects of clustering in two situations involving work groups. There are strong effects in Item 4 in spite of the fact that the clusters averaged only 11 or 12; the large homogeneity of attitudes within work sections (rho going up to 0.4) caused large effects. If the researcher had used s.r.s. formulas he might have been wrong 38 per cent of the time instead of 5 per cent.

For most of the national surveys of the SRC the actual variances are from 1 to 2 times as great as the s.r.s. variances. This means that the "effective size" of a sample of 2,000 elements is equivalent to 1,000 s.r.s. elements. Of course, the sample is designed in the expectation that the savings in cost per element will be greater than the loss in information. Effects as large as those of Item 3 are rare in good surveys, because if one can anticipate the large effects they can usually be avoided in good design. However, Marks found a ratio of 11 in his pioneer investigations of this problem.[4] The study he criticized involved intelligence tests using school classes as sampling units. Using s.r.s. formulas in that case distorted confidence levels from .01 to .50! Furthermore, we found effects of similar magnitudes for the confidence intervals of the *differences* of pairs of means. From these we may expect that *analytical statistics* in general can be subject to effects of this kind.

The distribution of the $\frac{\text{True Var.}}{\text{s.r.s. Var.}}$ ratios are not uniform between their limits of 1 and 2. The greatest concentration is near the lower end, perhaps near 1.2. Then there are fewer and fewer as one gets near 2. Furthermore, the computed ratios are subject to considerable sampling errors; our "true variances" are based usually on 66 primary sampling units. Because of this, in setting up the above limits, we excluded an occasional stray value that fell well beyond the bounds that included most computations.

[4] Eli S. Marks, "Sampling in the Revision of the Stanford-Binet Scale," *Psychological Bulletin*, 44 (September, 1947), pp. 413–434.

THE PRESENT SITUATION

In attempting to state in broad and general terms the present situation on the existence of valid statistics for data from complex samples, I shall assume that the sample has come from a good probability design, and also that it is either self-weighting or that the proper weights have been introduced to counteract unequal probabilities of selection.

1. For estimating a population mean, proportion, median, or total, the corresponding sample value may serve as a good estimator. Here the s.r.s. assumption does not lead to large errors, and the ordinary sample values are usually unbiased estimators or have only negligible bias. (It is understood that the sample total is multiplied by some "raising factor" to estimate the population total; that a proportion is but one form of the mean; and that frequently the sample mean is really a ratio estimator and calls for some caution.)

2. The variances of the above statistics (mean, proportion, total, and median) can be estimated to a good approximation for properly designed samples. The formulas for the variances differ from one sample design to another; generally they are quite different from s.r.s. formulas. The substitution of s.r.s. formulas (such as σ^2/n or pq/n) in place of the appropriate variance estimators can lead to large mistakes; for clustered samples it typically leads to underestimation that may be serious (as shown above). The literature of the social sciences shows many of these mistakes.

For the difference of two means or totals the variance estimators are but extensions of the foregoing; thus we may construct confidence intervals and tests of hypotheses for those differences. Furthermore, the above estimators of means and totals, and of their variances, are available not only for the entire population covered by the sample, but also for its subclasses; and the difference of two subclasses can also be tested. With that, however, we exhaust the list of statistics available for data from complex sample designs.[5]

[5] The several texts on sampling (see footnote 1) contain valuable information on the estimation of means, proportions, and totals. They also discuss at length the variances of those statistics; these are rather complicated for some complex designs.

This content downloaded from
132.174.254.12 on Wed, 10 Jun 2020 21:48:30 UTC
All use subject to https://about.jstor.org/terms

3. In general, formulas are lacking for the analytical treatment of data arising from complex sample designs.[6] There is a great variety of statistical tools now available for s.r.s. selection: correlation and regression analysis, discriminant analysis, analysis of variance for differences among several (3 or more) means, tests of contingency tables, nonparametric tests, and so forth. All the formulas for these tests involve the assumption of n independent selections, that is, s.r.s. selection. However, they are not valid for data from complex sample designs.

For some statistics the erroneous use of s.r.s. formulas leads to only small errors. For example, the sample correlation coefficient is usually a good estimate of the population correlation coefficient. But the s.r.s. formula (something like $1/\sqrt{n-k-1}$) for the standard error of that coefficient can be a very poor estimate; for clustered samples it may be a serious underestimate. Thus the use of s.r.s. formulas on data from complex sample can lead to gross mistakes in probability statements. I fear that such mistakes are frequently made.[7]

For an attempt to provide a simplified and unified approach for computing variances of means and of their differences, applicable to most practical designs, see Leslie Kish and Irene Hess, "On Variances of Ratios and Their Differences in Multi-Stage Samples," *The Journal of the American Statistical Association*, in press.

For the variances of medians see Hansen, Hurwitz and Madow, *op. cit.*, pp. 448–449. For the treatment of differences of subclasses from a stratified sample of elements see Yates, *op. cit.*, Ch. 9.

[6] "The required sampling assumptions are, in general, extremely restrictive; the majority of tests of significance still relate only to extremely simple Bernoullian samples and the sampling procedures now in general use ('cluster' sampling, two-stage sampling, etc.) remain, in the present state of theory, outside the scope of these tests." [P. Thionet, "Mathematical Methods in Public Opinion Polls," *International Social Science Bulletin*, 6 (December, 1954), p. 652.]

[7] The situation of the correlation coefficient is somewhat similar to that of the mean. The sample value is often a good estimator of the population value. But the sampling distribution of the estimator can be subject to much larger variance than s.r.s. selection, because of the correlation among the units selected from the same cluster. Then the use of s.r.s. variance formulas results in gross underestimation of the true variance. For these opinions I assume responsibility, but in discussions with fellow statisticians I found support and no contradiction.

Let us recapitulate and assume agreement on three things: (1) That using simple random sampling formulas on data obtained from complex clustered sample designs may lead to large mistakes and should be avoided. (2) That for properly designed samples good approximations are available for estimators of population means, proportions, totals, and medians, and also for estimating the variances of those estimators. Furthermore, for the difference of two of these estimators the variances, too, can be well approximated. These things can be done for the entire sample and for its sub-classes. The computations of these estimates are only moderately cumbersome, and are now generally available. (3) Present statistical theory fails to provide confidence intervals and tests of hypotheses for analytical statistics if the data arise from complex samples.

What are the alternatives facing the research scientist who wants to construct confidence intervals and tests of hypotheses for analytical statistics? Below are seven procedures that are appropriate in different circumstances. The first three amount to ways of avoiding the main problem through simplified designs. The last four are attempts at obtaining useful approximations for complex samples. The presentation is organized around the problem of providing the estimators of variances for the construction of valid confidence intervals. This problem can be translated into methods for testing hypotheses.

WAYS OF AVOIDING THE PROBLEM

Here follow three ways, one of which may help the researcher in a specific situation to avoid the difficult problems of constructing valid confidence intervals on data obtained from a complex sample design. I refrain from including the alternative of postponing the computations until the proper formulas are derived by mathematical statisticians. Also I object strongly to ignoring the problem by making the hypothesis and the "universe" fit the sample. In this inverted procedure the statistical inference is to some unspecified universe, or to a vague and artificially contrived one. The resulting confusion often hides unsuspected nonsense. It is common to find the researcher ignoring the nature and the very existence of the "universe" that

This content downloaded from
132.174.254.12 on Wed, 10 Jun 2020 21:48:30 UTC
All use subject to https://about.jstor.org/terms

must lie behind every statement about "significant differences."

1. *The researcher may take a simple random sample of elements.* Under this heading I would include sample designs which do not depart importantly from s.r.s., e.g. proportionate stratified random samples of elements, or systematic samples of elements. This procedure can be followed with little or no sacrifice of efficiency when the listing and the travel costs are not prohibitive. The researcher can afford moderate increases in the costs of the selection and of the measurement (the "field work") if the s.r.s. design brings compensations in easy computations of variances and of analytical statistics.

The complexities of cluster samples are sometimes introduced into sample design without sufficient justification. Here are three examples of situations where a return to simpler methods may be advisable: (a) A simple sample of elements may be best for mailed questionnaires when good mailing lists are available. (b) For dwellings of a city, instead of an area sample clustered by blocks, one might well use a systematic selection of individual lines from the city directory supplemented by an area sample of the small proportion missed by the directory.[8] (c) In sampling the students of a university, one can often take students rather than classes as the sampling unit.

2. *The selection procedures may be designed specifically to yield simple estimates of the variance or other needed statistics.* This idea can be illustrated with three examples. (a) One approach consists of weaving into the entire design a small number of independent replications of it. Each of these replications then gives an independent estimate of the research objectives, and their comparisons yield estimates of the variance of the entire sample.[9] (b) Relatively simple methods are available if the sample consists of *independently* selected clusters that are equal in size, or roughly equal (see below). The sample may sometimes be designed accordingly. If the natural clusters vary greatly in size, one may create subsamples of equal size after selecting the clusters with probabilities proportional to size.[10] (c) Another interesting approach is being investigated by Nathan Keyfitz of the Dominion Bureau of Statistics of Canada, and William Hurwitz and Joseph Steinberg at the U. S. Census Bureau. This procedure consists in exploiting the considerable simplification and standardization that can be obtained by taking exactly two selections from each stratum. The variance is built up simply from units like $(y_{n1}-y_{n2})^2$ where y_{n1} and y_{n2} are sample values of the two selections from the h–th stratum. For situations of experimental design Keyfitz shows somewhat analogous advantages for the 2^n factorial design.[11]

These are three examples of the general method of designing the sample with the object of obtaining the desired complicated statistics reliably and simply. To obtain these results, one needs to use foresight and care and, perhaps even to sacrifice some of the theoretically available statistical efficiency.

3. *The researcher may refrain from constructing confidence intervals and from making probability statements.* He thus allows the inference from sample to population value to remain indefinite. In these cases reliance on the research result comes not through statistical inference but through the successful repetition of similar results in separate investigations. In a sense, one may consider the separate investigations as if they were so many different sampling units. This approach may be quite acceptable in some situations, especially if the conditions are forced on the researcher, as is often the case in the social sciences, in geology, in astron-

[8] Theodore D. Woolsey, "Sampling Methods for a Small Household Survey," *Public Health Monograph* No. 40, Issued concurrently with *Public Health Reports,* 71, (August, 1956).

[9] For example, 10 subsamples may be drawn from the same population, independently and using the same probability design. Then if the mean of one of the ten subsamples is denoted as $\bar{y}_i$, and the mean of the entire sample is $\bar{y}=\frac{1}{10}\sum^{10}\bar{y}_i$, then simply variance of $\bar{y}=\frac{1}{10}\frac{1}{9}\sum^{10}(\bar{y}_i-\bar{y})^2$

See Deming, *op. cit.*, pp. 352–355; also W. Edwards Deming, "On Simplifications of Sampling Design Through Replication with Equal Probabilities and Without Stages," *Journal of the American Statistical Association,* 51 (March, 1956), pp. 22–53.

[10] Kish, *op. cit.*, pp. 226–230.

[11] Nathan Keyfitz, "A Factorial Arrangement of Comparisons of Family Size," *American Journal of Sociology,* 58 (March, 1953), pp. 470–480. See also Model I in Kish and Hess, *op. cit.;* and Deming, *op. cit.*, pp. 34–39.

This content downloaded from
132.174.254.12 on Wed, 10 Jun 2020 21:48:30 UTC
All use subject to https://about.jstor.org/terms

omy, in biology, and so forth. Perhaps most scientific advance has been made in this way.

In some of these situations the researcher may compute from s.r.s. formulas a lower limit on the size of the confidence interval. This is to serve as a check on the reader's imagination and as a precautionary measure against overinterpreting the data. But then the reader should be warned that these are not true confidence intervals but only likely lower limits. In terms of testing a null hypothesis: If the s.r.s. computations fail to reject the null hypothesis, it remains unrejected; but the reverse does not hold.

APPROXIMATIONS FOR ANALYTICAL STATISTICS

We trust that gradually mathematical statisticians will provide an increasing number of specific formulas for many useful analytical statistics and for different kinds of complex designs, but the researcher who has interesting data from a complex sample does not want to wait for that millenium. He wants to use some of the powerful tools of analytical statistics now. Furthermore, if he is conscious of the mistakes that may result from using s.r.s. formulas, he wants to avoid them. Some alternative approaches are suggested here, in the hope that they may be helpful. These approaches have various kinds of weaknesses that are more or less serious in different situations; and the mathematical foundations of some are not as secure as we wish. Nevertheless, they are frequently preferable to the s.r.s. formulas now in wide use.

1. *When an exact formula is not available, the researcher should look (or ask a statistician) for a good approximation.* One may construct from the available statistics some model that will fit the situation fairly well; perhaps much better than the s.r.s. model. In choosing a model, judgment must be based on statistical theory, on relevant statistical experience, and on the substantive knowledge of the nature of the data under treatment.

Consider a sample consisting of clusters that have been drawn with independent random choices: classes have been drawn at random, and these become clusters of the students who form the elements of the population under study. Sometimes satisfactory statistical treatment results from treating the cluster means as the units in the statistical analysis. This treatment *may* permit the use of the standard s.r.s. formulas on these units.

However, there are obstacles to this approach, and these may illustrate the needed sources of good judgment. First, we cannot consider the different cluster means of the analysis as independent if they are made up of elements that originate in common sampling units.[12] Second, gross inequalities in the sizes of the clusters may call for the use of weights for the cluster means.[13] Third, the use of the cluster means may be improper on substantive grounds. For example, correlations taken among cluster means has a different meaning than correlation among the elements.[14]

2. *It is sometimes useful to translate the problem into another problem for which the statistical solution is available.* Sometimes merely reformulating the statistics will provide a solution. For example, the general formulas for chi-square tests of contingency tables are lacking for data from complex samples. But a 2×2 chi-square test can usually be translated into a test of the hypothesis that there is zero difference between the proportions possessing a certain characteristic in two populations. This difference of two proportions can be tested because the formula of its variance is known.

To clarify this illustration examine a 2×2 table on which a chi-square test of independence is to be conducted:

---

[12] This point may be clarified through the example of a simple analysis of variance test among the means of k classes. If each of the k classes consists of *entire* clusters then the clustering does not introduce correlation among the means. But if the class lines cut across the cluster lines, the correlation due to clustering interferes with the contrasts among the cluster means. The difficult statistical problems of dealing with the correlations among the units of analysis, as these problems arise in complex sample designs, still await solution.

[13] Paul Meier, "Variance of a Weighted Mean," *Biometrics*, 9 (March, 1953), pp. 59–73; William G. Cochran and Sarah P. Carroll, "A Sampling Investigation of the Efficiency of Weighting Inversely as the Estimated Variance," *Biometrics*, 9 (December, 1953), pp. 447–459.

[14] W. S. Robinson, "Ecological Correlations and the Behavior of Individuals," *American Sociological Review*, 15 (June, 1950), pp. 351–357.

This content downloaded from
132.174.254.12 on Wed, 10 Jun 2020 21:48:30 UTC
All use subject to https://about.jstor.org/terms

	1	2	Totals
a	n_{a1}	n_{a2}	n_a
b	n_{b1}	n_{b2}	n_b
Totals	n_1	n_2	n

In the usual situation one wants to test the null hypothesis of no difference in the proportions of a's occurring among the individuals of samples 1 and 2, respectively. Instead of the common chi-square test we may conduct a *t* test on the difference between $p_1=n_{a1}/n_1$ and $p_2=n_{a2}/n_2$. If the samples 1 and 2 were selected independently by s.r.s., the variance of this difference would be roughly $p(1-p)(\frac{1}{n_1}+\frac{1}{n_2})$, where $p=n_a/n$. The difference between the chi-square and the t-test is of practical importance only for small samples: the "Yates correction" for the former and the factor of $(n_1+n_2)/(n_1+n_2-2)$ for the latter. When the sample is not s.r.s. but complex, neither of the above two tests is valid; however, for the difference of the two proportions, valid estimates of the variance may be computed from appropriate formulas (see footnote 5).

In other situations a different test is called for because the rows and columns represent merely two measurements on the same sample of *n*. For example, in a "test-retest" situation the *a*'s and *b*'s may stand for the "yes" and "no" answers, respectively, *before* the experiment; and the 1's and 2's for the "yes" and "no" answers *after* the experiment. The null hypothesis—that the proportion of "yes" answers after the experiment is the same as before it—may be tested with a modification of the chi-square test based on the assumption of s.r.s.[15] This test, however, may be made also as a t-test of no difference between the two proportions $p_{a2}=n_{a2}/n$ and $p_{b1}=n_{b1}/n$. If the n elements came from an s.r.s. selection the variance of the difference $(p_{a2}-p_{b1})$ would be approximately $(p_{a2}+p_{b1})/n$. For a complex sample, again, we can compute an appropriate variance for the difference of the two proportions.

Another example arises from the problem of confidence intervals for medians and other quantiles. This has been solved by translating the problem into one of confidence intervals for proportions.[16] At the Survey Research Center we are working on an analogous confidence interval for the difference of two medians based on the difference of two proportions.

When the researcher cannot find solutions to problems by merely reformulating the statistics, he may attempt another approach by recasting the substance of his question in an altogether different mold. This approach is less desirable since statistical analysis should follow the researcher's interest rather than lead it. At times, however, this method may succeed where others fail. In particular, the difference of two means may serve in the most diverse places as an analytical tool. We may place high hopes on further extensions of the "multiple comparison tests," designed by Tukey and Scheffé, especially after its adaptation to complex samples.[17]

3. *From computations on available statistics inferences may be made by analogy to other statistics for which statistics are not available.* Take as an example the case where one wants to test the null hypothesis that k (greater than 2) means are samples from the same population of means. First, the researcher can compute the variance of the difference of two of these means; and he does this for several pairs, perhaps for all pairs. Second, he finds the ratio of these variances to what he would expect with s.r.s. Third, he infers that the comparison of the several means would show the same kind of increase over s.r.s. Sometimes tests will disclose that the s.r.s. assumption is fairly satisfactory, and the researcher can base his tests on the standard formulas with some confidence.

In this kind of situation, one often com-

[15] Quinn McNemar, *Psychological Statistics*, New York: Wiley and Sons, 1949, pp. 200–207.

[16] Hansen, Hurwitz and Madow, *op. cit.*, pp. 448–449.

[17] Henry Scheffé, "An Analysis of Variance for Paired Comparisons," *Journal of the American Statistical Association*, 47 (September, 1952), pp. 381–400.

This content downloaded from
132.174.254.12 on Wed, 10 Jun 2020 21:48:30 UTC
All use subject to https://about.jstor.org/terms

putes a whole series of variances for different items and then reduces them to some standard form so that they may be compared under assumptions that should be reasonable on the bases both of their logical content (their mathematical derivations) and of empirical evidence obtained from similar data. Probably the most popular standard form for comparison among sampling statisticians is the one mentioned above: noting the ratio of the increase of the variance over the s.r.s. formula. Sometimes this can be expressed conveniently in terms of the measure of homogeneity, the coefficient of intraclass correlation.[18] Sometimes some other standards of comparison may be useful; one is the coefficient of variation or its square, the "relvariance."

There are some important problems to be undertaken here. What is an analogous test? What criteria should one use for averaging a set of these comparisons? One may use some central value, but some prefer to use a so-called "safe" confidence interval; that is, something close to the topmost value of a series of computed ratios to s.r.s. Recently we have used two limits in our sampling error tables. These two limits are near the lower and upper limits of a set of ratios to s.r.s. The lower limit is generally that of s.r.s.; for proportions this is $\sqrt{pq/n}$. The upper limit quite frequently approaches $\sqrt{2pq/n}$ (but it tends to be lower for smaller *n*, the size of the subsample on which it is based). Within the two limits, tests of hypotheses are conducted as follows. If the lower limit is not exceeded, do not reject the null hypothesis. If the upper limit is exceeded, reject the null hypothesis. If the statistic falls between the lower and upper limits, suspend judgment and try to conduct a more exact test.

4. *A general method now being investigated, promises to yield good approximations for the variances of many statistics.* Say that we need the variance for an estimate *u*. We select half of the entire sample in a manner that reproduces the full complexity of the sample. For many designs this means selecting at random half of the "primary sampling units" from each of the strata comprising the entire sample. Then we compute the same estimate as above, but based on the half sample; call this estimate $u_{1/2}$. It turns out the $(u_{1/2}-u)^2$ is an unbiased estimate of the variance of *u*. Since it is subject to large sampling error, one may make the split repeatedly to obtain several estimates, and then use a central value among them as the estimate of the variance of *u*.

In this way one may obtain estimates of the confidence intervals for which specific formulas are not now available. At the U. S. Census Bureau this method has been tried out on some complex estimates of means and totals for the Current Population Surveys. At the Survey Research Center we are now introducing it for computing standard errors for regression coefficients and other statistics.

SUMMARY

Standard statistical literaure has been developed in terms of the assumptions of *independent* observations obtained by simple random sampling. However, most social research is actually carried out by means of complex sample designs because s.r.s. is uneconomical and impractical. In the social sciences the use of s.r.s. formulas on data obtained from complex samples frequently results in serious mistakes. Among the various kinds of departures from the assumptions of s.r.s. *clustering* probably causes the largest and most frequent errors.

For both surveys and experiments the researcher is often compelled to select his individual observations in clusters in which the individuals tend to be more or less homogeneous. This homogeneity violates the assumption of independence inherent in most statistical formulas and has particularly serious consequences on the probability levels of confidence intervals and of tests of hypotheses. Many researchers who wish to use a 5 per cent level of confidence actually use levels ranging from 10–50 per cent! Empirical evidence is presented of the magnitude of these errors in different situations.

Several fine recent books deal with the problems of complex sampling, and contain proper formulas for estimating means, proportions, totals, and medians, and also the variances of these statistics. For the differ-

[18] John E. Walsh, "Concerning the Effect of Intraclass Correlation on Certain Significance Tests," *Annals of Mathematical Statistics*, 18 (March, 1947), pp. 88–96.

This content downloaded from
132.174.254.12 on Wed, 10 Jun 2020 21:48:30 UTC
All use subject to https://about.jstor.org/terms

ence of a pair of these estimates valid confidence intervals (and tests of hypotheses) can also be constructed. Beyond these, however, mathematical formulas are lacking for the analytical treatment of data arising from complex samples.

For the research scientist who wants to apply analytical treatment this article presents seven procedures for arriving at an approximate answer. These procedures are offered as preferable alternatives to the present common practice of relying on standard formulas based on the s.r.s. (independence) assumption. Of these alternatives three amount to evasions of the worst problems by designing simpler samples. The other four attempt approximations of analytical statistics for complex samples.

Appendix

To demonstrate the magnitude of the problem I examined briefly the 74 articles appearing in 1956 in one of the best journals in social science—the *American Sociological Review.*

Thirty-five were not based on probability samples and contained no probability statements. Of these 24 were not quantitative in content; 11, though quantitative in content, were not based on samples, but on small "chunks," that is, group(s) chosen by judgment, or on entire populations.

Eleven articles were not based on probability samples, but contained probability statements. Of these 2 were based on entire populations; and 9 were based on "chunks" selected by judgment.

The remaining 28 were based on probability samples.

- 3 were s.r.s. selections or close approximations
- 2 were complex selections and the probability statements were computed properly
- 3 were complex selections and probability statements were not made
- 15 were complex selections and the probability statements were computed improperly by s.r.s. formulas
- 5 contained s.r.s. probability statements, but the description of the selection method was not clear.

In summary, of the 28 articles based on probability samples, this discussion was pertinent to somewhere between 18 and 23.

In any journal devoted to quantitative social science it is the rule rather than the exception to find probability statements based on s.r.s. formulas for data obtained from samples that were complex, not s.r.s. We may hope that over the years the situation will improve. Meanwhile the reader may find rewarding exercise in searching *his* favorite journal for examples.

THE GROWTH OF METROPOLITAN SUBURBS *

Leo F. Schnore

Michigan State University

Decentralization is clearly one of the most significant movements in the long history of urban communities. In the United States the shift to the suburbs began around a few large cities toward the close of the nineteenth century,[1] but it is no longer confined to a mere handful of super-cities. As time has passed, the outward thrust of urban population has become characteristic of smaller and smaller places, and there is no indication that the movement is abating.

Fortunately, social scientists have charted the major trends involved in this suburban revolution during recent decades. The careful historical studies by Thompson, Bogue, and Hawley have described the outstanding population shifts from the turn of the century to the most recent census.[2] They show the

* Revised version of paper read at the annual meeting of the American Sociological Society, September, 1956. The data presented in this report were initially assembled while the writer held a Research Training Fellowship from the Social Science Research Council.

[1] Adna F. Weber, *The Growth of Cities in the Nineteenth Century,* New York: Columbia University Press, 1899.

[2] Warren S. Thompson, *The Growth of Metropolitan Districts in the United States, 1900–1940,* Washington: Government Printing Office, 1947; Donald J. Bogue, *Population Growth in Standard Metropolitan Areas, 1900–1950,* Washington: Gov-

This content downloaded from
132.174.254.12 on Wed, 10 Jun 2020 21:48:30 UTC
All use subject to https://about.jstor.org/terms