NATIONAL CENTER For HEALTH STATISTICS | Series 2 Number 31

VITAL and HEALTH STATISTICS

DATA EVALUATION AND METHODS RESEARCH

Pseudoreplication

Further Evaluation and Application of the Balanced Half-Sample Technique

Uses of the method in analysis of data from complex surveys.

DHEW Publication No. (HSM) 73-1270

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Public Health Service
Health Services and Mental Health Administration

Washington, D.C. January 1969

DHEW Publication No. (HSM) 73-1270

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C. 20402 - Price 35 cents

NATIONAL CENTER FOR HEALTH STATISTICS

Public Health Service Publication No. 1000-Series 2-No. 31

Library of Congress Catalog Card Number 68-62232

PREFACE

In an earlier publication of the National Center for Health Statistics (Series 2, No. 14) the method of balanced half-sample pseudoreplication was introduced as a device for estimating precision of estimates which come from surveys of modern complex design. The technique was developed by Philip J. McCarthy of Cornell University. The report and McCarthy's work are facets of a series of efforts by the Center to find appropriate methods for analyzing data from complex social and health surveys, in which the assumptions of standard classical techniques rarely are satisfied.

The present study describes a number of internal evaluation techniques which aid in understanding the behavior of balanced half-sample replication as it relates not only to the variance of linear estimators but also to precision of ratios. And it extends the method to applications of a more analytic character, including adaptations to the sign test and to contingency tables.

Replication and pseudoreplication are concepts that have deep roots in statistical theory and that have influenced the thinking of many people. Acknowledgment of all who have contributed is not feasible, but the list of those who have had special influence on the research presented here includes W. Edwards Deming, John Tukey, M.H. Quenouille, Leslie Kish, and several members of the staff of the U.S. Bureau of the Census. Professor McCarthy carried out the research and wrote the report. Walt R. Simmons has coordinated Center activities in the area of pseudoreplication and, as Project Officer has maintained close contact with the present undertaking.

SYMBOLS

Data not available------------------------	---
Category not applicable-------------------	...
Quantity zero----------------------------	-
Quantity more than 0 but less than 0.05----	0.0
Figure does not meet standards of reliability or precision-----------------	*

CONTENTS

A REPORT of the National Center for Health Statistics (Series 2, No. 14) introduced a pseudoreplication procedure, called balanced half-sample replication, as a basis for statistical analysis when dealing with data derived from complex sample designs and estimation processes. It appeared particularly suited for variance estimation in connection with the commonly used multistage design of two primary units per stratum, with the individual observations being subjected to various kinds of ratio and poststratification adjustments.

The exact behavior of balanced half-sample replication as a variance-estimating procedure is known only for a simple linear situation where it has no real utility. The first portions of this report review this theory and describe a number of evaluation techniques which can provide evidence concerning the behavior of balanced half-sample replication for a particular set of data. They are illustrated on data derived from the Center's Health Examination Survey (Series 11, No. 1) relating to a variety of body measurements for a sample of approximately 3,000 U.S. adult males and involving estimates of population means, multiple regression coefficients, and multiple correlation coefficients. These techniques are principally concerned with providing presumptive evidence relating to (1) the extent to which the average of a set of balanced half-sample estimates "exhausts" the information in the entire set of 2^L half-samples, (2) the existence of bias in the entire-sample estimate, and (3) the adequacy of the variance estimates produced by balanced half-sample replication. Applications to the Health Examination Survey data seem to indicate that balanced half-sample replication is performing well as a variance estimating procedure. Also, as might be expected on the basis of the sample size, there is no evidence of bias in the estimates. The variables and subpopulations are less extensive than one might desire for an exploratory investigation, but this limitation was imposed by the fact that these tabulations were produced as a byproduct of another study.

The last two sections of the report suggest several applications of balanced half-sample replication to problems of a more analytic character. Thus an investigation of its use to test the hypothesis of independence in a contingency table is described, and the relationship of balanced half-sample replication to the sign test is explored.

PSEUDOREPLICATION

FURTHER EVALUATION AND APPLICATIONS OF THE BALANCED HALF-SAMPLE TECHNIQUE

Philip J. McCarthy, *New York State School of Industrial and Labor Relations, Cornell University*

INTRODUCTION

The extensive literature on theory of sample surveys well documents the fact that commonly used complex sample designs and estimation procedures lead to approximate and extremely complicated expressions for variance estimation. Furthermore, the survey data often do not satisfy the conditions required for the application of even elementary statistical techniques of analysis. In the absence of simple methods for handling these problems, the ordinary researcher will ignore the complexities of design and merely treat the data as though they had been obtained by simple random sampling; the more sophisticated researcher will be frustrated in not being able to account for all the effects of design in his analysis. These points have been discussed in some detail in a publication of the National Center for Health Statistics (Series 2, No. 14).

Durbin (1967) has developed methods for designing multistage surveys in such a manner that error computations will be relatively simple, where the emphasis is on the estimation of population means, percentages, or totals. However, it is unlikely that these methods can ever be extended to handle more complicated problems of statistical analysis, such as the use of multiple regression techniques. Another possible avenue of approach to these problems is through the use of independent replications of the sample design, variously referred to as interpenetrating samples, duplicated samples, or random groups. Deming, for example, has been a consistent advocate of replicated sampling. He first wrote of it as the Tukey plan (1950), and his recent book (1960) presents references and descriptions of the applications of replicated sampling to many different situations and contains a wide variety of ingenious devices that he has developed for solving particular problems.

A major disadvantage of replicated sampling arises from the difficulty of obtaining a sufficient number of independent replicates to provide adequate sampling stability for estimating variances and for other purposes. Thus the commonly used design of two primary units per stratum (frequently obtained by collapsing strata from each of which only a single unit has been drawn) gives only two independent replicates. To overcome this problem of having only one degree of freedom available for variance estimation, the U.S. Bureau of the Census originated a pseudoreplication scheme called half-sample replication. The scheme was adapted and modified by the staff of the National Center for Health Statistics (NCHS) and has been used in their Health Examination Survey reliability measurements. A brief description of this approach is given in a report of the U.S. Bureau of the Census (Technical Paper No. 7, 1963, p. 57), and a reference to the Census method of half-sample replication was made by

Kish (1957, p. 164). The basic characteristics of this procedure have been described by Gurney (1962) and in the NCHS publication (Series 2, No. 14).

A refinement of this pseudoreplication technique, called *balanced half-sample replication*, that increases the precision of variance estimates was also introduced in NCHS (Series 2, No. 14) and was applied to the analysis of data derived from the Health Examination Survey (NCHS, Series 11, No. 1). The present report carries further the evaluation of the balanced half-sample technique, with illustrative data drawn from the Health Examination Survey. In addition, several applications to problems of a more analytic character are suggested. Thus an investigation of the use of balanced half-sample replication to test the hypothesis of independence in a contingency table is described, and the relationship of balanced half-sample replication to the sign test is investigated.

BACKGROUND FOR EVALUATION STUDIES

In simplest terms, the idea of half-sample replication is as follows. Consider a stratified sampling procedure where two independent selections are made from within each of L strata. Denote the strata weights by W_h, $h = 1, 2, \ldots, L$, and the observations by y_{hi}, $h = 1, 2, \ldots, L$ and $i = 1, 2$. Under these circumstances, a half-sample replicate is obtained by choosing one of y_{11} and y_{12}, one of y_{21} and $y_{22}, \ldots$, and one of y_{L1} and y_{L2}. A half-sample estimate of the population mean is $\bar{y}_{hs} = \sum_h W_h y_{hi}$, where i is *either* 1 or 2 for each h. There are 2^L possible half samples. If the estimate made from the entire sample is denoted by $\bar{y}_{st} = \sum_h W_h (y_{h1} + y_{h2})/2$, then it can be demonstrated that if k half samples are independently selected from the entire set of 2^L possible half samples, with means denoted by $\bar{y}_{hs,1}, \bar{y}_{hs,2}, \ldots, \bar{y}_{hs,k}$, then $E[\sum_{i=1}^{k} (\bar{y}_{hs,i} - \bar{y}_{st})^2/k]$ is equal to the ordinary variance estimate of $\bar{y}_{st}$, where the expectation is taken over the entire set of 2^L half samples. As is explained in NCHS (Series 2, No. 14), a balanced set of half samples is obtained by appropriately choosing the half samples so as to eliminate a between-strata contribution which influences the stability of this estimate of variance.

The basic characteristics of balanced half-sample replication, as outlined in the preceding paragraph, have been investigated in relation to a linear situation where the method obviously has no real utility, at least as far as ordinary variance estimation is concerned. Under these circumstances the method merely reproduces results that can be obtained by direct and simple methods of analysis. If, however, more complicated methods of sampling and estimation are employed, then direct methods of analysis may not be available or may require a prohibitive amount of computation in comparison with half-sample replication. For example, suppose that a multiple regression coefficient is estimated from a complex sampling and estimation operation. No known theory leads to an estimate of its sampling variability. Nevertheless, one may compute a value of the regression coefficient from each balanced half sample and use the comparisons among the computed regression coefficients to estimate the variability of the overall coefficient. Troublesome "domain of study" problems, as described by Hartley (1959) or Kish (1965), may also be handled in the same way. The simplicity of the approach is most appealing in that one has only to apply the appropriate estimation procedure to each half sample, followed by a simple variance computation on the separate half-sample estimates.

Since balanced half-sample replication does permit the "easy" computation of variance estimates which intuitively seem to mirror most of the complexities in survey sampling, estimation, and analysis, one would like to argue that it can "correctly" account for all of these features. This is clearly not the case. For example:

1. If the y_{hi} represent values of a variable associated with individual population elements, if the two elements per stratum are selected with equal probabilities and without replacement, and if the sampling fractions are the same in all strata, then a finite population correction can easily be inserted at the end of the variance

estimation process. The effect of differing sampling fractions in the separate strata could be taken into account by working with W'_h's where W'_h is equal to W_h multiplied by the square root of the appropriate finite population correction, but even this simple change effectively divorces the estimation procedure from the variance estimation process. The situation becomes more involved if the two units are selected with unequal probabilities and without replacement.

2. If the y_{hi} represent estimates of a cluster characteristic, as in two-stage sampling, then one must also worry about correctly accounting for the within-cluster contribution to the total variance. As is evident in Cochran (1963, ch. 10), the problem is not substantial if the first-stage units are selected with equal probabilities and with replacement and the first-stage sampling fractions are small, or if the first-stage units are selected with unequal probabilities and with replacement. The situation in which two units are selected with unequal probabilities and without replacement is the variance estimating problem to which Durbin (1967) addressed himself.

3. Most estimation problems are not linear but involve the use of ratios in one form or another. Under these circumstances, it will not even be necessarily true that the estimate made from the entire sample will be equal to the average of the balanced half-sample estimates and thus new dimensions of complexity will be involved.

As implied by the foregoing examples, the exact characteristics of estimates of variance obtained from balanced half-sample replication are, for the most part, unknown. These characteristics may be investigated in a variety of ways, including the following: (a) *exact analytic*, in which one assumes the functional form of a distribution (or a joint distribution) and obtains straightforward, although possibly not simple, answers; (b) *approximate analytic*, in which one uses Taylor series approximations, obtains bounds by analytic methods, and the like; (c) *empirical studies*, in which one employs data obtained from surveys to investigate the behavior of statistical procedures, frequently using such an approach to check on results obtained in (b); and (d) *Monte Carlo sampling* from synthetic populations, again usually to check on results obtained in (b).

In the area of present concern, there appears to be little hope of obtaining exact analytic solutions to problems. In particular, exact solutions require quite restrictive models as evidenced by a portion of the work of Durbin (1959) and the work of Rao and Webster (1965) on the Quenouille version of a ratio estimate. Their developments start with the assumption that $y_i = \alpha + \beta x_i + u_i$, $E(u_i | x_i) = 0$, $V(u_i | x_i) = n\delta$ where δ is a constant for each n such that $n\delta$ is bounded, and the variates x_i/n have the gamma distribution with parameter h. Such investigations are extremely valuable in providing clues concerning the behavior of a particular method, but it is not possible to devise specific models for the many different types of data studied by large survey organizations.

Taylor series approximations have also been used extensively to investigate the properties of a wide variety of ratio estimators, both with and without models. One of the more recent accounts and summaries of this type of approach is that of Tin (1965). In general, results are obtained on bias and variance estimation which one wishes to check against data obtained either from actual studies or from Monte Carlo experiments, particularly when the samples are small or moderate in size. Thus Kish, Namboodiri, and Pillai (1962) use actual survey data while Tin (1965) and Lauh and Williams (1963) use Monte Carlo techniques.

In this report, the general view is adopted that the behavior of estimates and of variance estimates should be investigated, if possible, on the basis of actual survey results. This enables one to consider a wide variety of types of data that are of current interest and avoids the problem of having to construct "representative" synthetic populations and having the results refer only to these populations. Even the construction of appropriate synthetic populations, to which one could apply complex sample designs and estimation procedures, would be a difficult undertaking.

DESCRIPTION AND APPLICATION OF EVALUATION TECHNIQUES

This section of the report will describe a number of internal evaluation techniques which can provide evidence concerning the behavior of balanced half-sample replication. These techniques will be illustrated on data derived from the Health Examination Survey (NCHS, Series 11, No. 1, and the appendix to NCHS, Series 2, No. 14) relating to a variety of body measurements for a sample of approximately 3,000 U.S. adult males. (The total sample also contained approximately 4,000 U.S. adult females.) Two different types of estimates will be considered: (1) Estimates of population means and (2) Estimates of multiple regression coefficients and multiple correlation coefficients. The original tabulations upon which these results are based were provided by the Survey Research Center of the University of Michigan as a portion of the work performed under a contract between the Survey Research Center and the National Center for Health Statistics. The variables and subpopulations are less extensive than one might desire for an exploratory investigation, but this limitation was imposed by the fact that these tabulations were produced as a byproduct of another study.

The Health Examination Survey is an example of the combination of a highly complex sample design and estimation procedure. In brief summary, the roughly 1,900 primary sampling units (PSU's) of the Current Population Survey were grouped into 42 strata, some of which contained only a single PSU that was then included in the sample with certainty. From each of the strata containing more than one PSU, a single PSU was selected and then subsampled in accordance with customary practice to obtain a sample of about 160 persons (75 males). The estimation procedure included four principal operations: (1) Inflation by the reciprocal of the probability of selection, (2) a first-stage ratio adjustment to 1960 population for eight geographic and population concentration classes, (3) an adjustment for nonresponse carried out in 294 age-sex-PSU cells, and (4) a poststratification by 12 age-sex cells. These observations were forced into a "two observations from each of 27 strata" format by pairing "adjacent" noncertainty strata and by defining "similar" subsamples from within the certainty PSU's.

Average of Half-Sample Estimates

As mentioned earlier, the formal consideration of balanced half samples has been restricted to the ordinary linear situation. One can move away from the linear case in a variety of ways, but only the ordinary combined ratio estimate will be discussed here. The points raised, however, apply to any nonlinear situation and will be illustrated in Health Examination Survey data. In the case of stratified simple random sampling, with two units selected from within each of L strata, there are a totality of 2^L half samples, from each of which one may compute a combined ratio estimate, say r_i. For the linear case, it has been shown that analyses carried out on a set of balanced half samples give the same results as would be obtained by carrying out the same half-sample analyses on the entire set of 2^L elements. We shall now investigate the manner in which this carries over to the case of half-sample ratio estimates.

In the linear case, denote the means of a balanced set of half samples by $\bar{y}_{hs,1}, \bar{y}_{hs,2}, \ldots, \bar{y}_{hs,k}$, and the mean of the entire sample by $\bar{y}_{st}$. Denote the means of the complementary set of half samples by $\bar{y}'_{hs,1}, \bar{y}'_{hs,2}, \ldots, \bar{y}'_{hs,k}$, where $\bar{y}'_{hs,i}$ is the mean of the half sample obtained using the L elements left out of the ith original half sample. If k is greater than L—so that each observation appears in half of the samples—certain relationships hold among these means. In particular, we have that $(\bar{y}_{hs,i} + \bar{y}'_{hs,i})/2 = \sum_{i=1}^{k} \bar{y}_{hs,i}/k = \sum_{i=1}^{k} \bar{y}'_{hs,i}/k = \bar{y}_{st}$. In a sense, we can say that the set of k balanced half samples "exhausts" the information in the totality of 2^L half samples, as far as estimation is concerned. The question can then be raised of whether the half-sample combined ratio estimators $r_1, r_2, \ldots, r_k$, separately from or in combination with their complementary estimators $r'_1, r'_2, \ldots, r'_k$, also "exhaust" the totality of all 2^L half-sample ratio estimators. Notice that we are here asking about the relationship between $\sum_{i=1}^{k} r_i/k$, $\sum_{i=1}^{k} r_i'/k$, and the average of the r_i's for all 2^L half samples and *not* about the relationship between these quantities and $\hat{R}$, the combined ratio estimator for the entire sample. The latter is a separate question which will be discussed shortly.

It appears that this problem can be attacked empirically by the following argument. Consider by way of illustration, the three-strata example discussed in the NCHS report (Series 2, No. 14, p. 17), where the observations are denoted by (x_{hi}, y_{hi}), $h = 1, 2, 3$, $i = 1, 2$, and $W_1 = W_2 = W_3 = 1/3$. Translated into combined ratio estimate terms, for a set of four balanced half samples, this is

$$r_1 = \frac{y_{11} + y_{21} + y_{31}}{x_{11} + x_{21} + x_{31}} \qquad r_1' = \frac{y_{12} + y_{22} + y_{32}}{x_{12} + x_{22} + x_{32}}$$

$$r_2 = \frac{y_{11} + y_{22} + y_{32}}{x_{11} + x_{22} + x_{32}} \qquad r_2' = \frac{y_{12} + y_{21} + y_{31}}{x_{12} + x_{21} + x_{31}}$$

$$r_3 = \frac{y_{12} + y_{22} + y_{31}}{x_{12} + x_{22} + x_{31}} \qquad r_3' = \frac{y_{11} + y_{21} + y_{32}}{x_{11} + x_{21} + x_{32}}$$

$$r_4 = \frac{y_{12} + y_{21} + y_{32}}{x_{12} + x_{21} + x_{32}} \qquad r_4' = \frac{y_{11} + y_{22} + y_{31}}{x_{11} + x_{22} + x_{31}}$$

In the r_i set, note that each observation (x_{hi}, y_{hi}) occurs in two of the four samples. Furthermore, if we look at any two of the three strata, say strata h and k, then each of the four possible pairs of observations

(x_{h1}, y_{h1}) and (x_{k1}, y_{k1})

(x_{h1}, y_{h1}) and (x_{k2}, y_{k2})

(x_{h2}, y_{h2}) and (x_{k1}, y_{k1})

(x_{h2}, y_{h2}) and (x_{k2}, y_{k2})

appears in one of the four samples. It is this "balancing of pairs of observations" which characterizes any set of balanced half samples and, in the linear case, makes them so suitable as the basis for variance-estimation. The set of r_i's is not, however, balanced on triplets of observations. Thus the triplet (x_{12}, y_{12}), (x_{22}, y_{22}), (x_{32}, y_{32}), obtained by taking one observation from each of the three strata, does not appear in the r_i set. Similar observations hold for the r_i' set. The combined set of r_i's and r_i''s in this case does account for all possible half-sample ratio estimates and is therefore balanced on triplets. If, in a particular example involving three strata, we found that $\bar{r}$ and $\bar{r}'$, the averages of the r_i's and r_i''s, respectively, were in close agreement, this would indicate that the effect of the balancing of triplets was negligible. Of course, the average of $\bar{r}$ and $\bar{r}'$ is in this instance the actual value that one wishes to estimate. Even for the artificial and extreme example given in NCHS report (Series 2, No. 14, p. 27), $E(\bar{r}) = 1.1985$ and $E(\bar{r}') = 1.1968$, thus indicating that the effect of triplet balancing is extremely small.

In general, a set of k half samples, where k is a multiple of four, will be balanced on the number of strata corresponding to the highest power of two that is an exact divisor of k. Thus if a set of eight balanced half samples is used for five strata, then the r_i and r_i' sets will each be balanced on singles, pairs, and triplets of observations while the combined set will also be balanced on quadruplets. Under these circumstances, if $\bar{r}$ and $\bar{r}'$ agree, then one can argue that quadruplet balancing is unnecessary, although this can be achieved by combining the two sets.

When k is not exactly equal to two raised to an integer power, and the Plackett-Burman (1943-46) method of constructing an orthogonal matrix is used, the balancing situation is not as clear cut. For example, the case of $k = 12$ has been completely investigated. Thus the r_i and r_i' sets are both balanced for pairs of observations since 12 is divisible by four. Furthermore, the combined r_i and r_i' sets, which contain 24 half samples, are balanced on triplets of observations. On the other hand, for the case of $k = 24$ which is given in NCHS report (Series 2, No. 14, p. 19), the r_i and r_i' sets are each balanced on triplets of observations, since 24 is divisible by eight, but the combined set of 48 half samples is not balanced on quadruplets of observations even though 48 is divisible by 16. Nevertheless, a small investigation[1] has indicated that the combined sets will be better balanced on quadruplets than will either set separately.

In ordinary practice one would perform all computations on a single set of k balanced half samples and would not be concerned with the complementary set. The two sets can, however, be compared if there is any doubt concerning the adequacy of a single set. This comparison has been made, for methodological purposes, with data derived from the Health Examination Survey which, as noted earlier, was based on approximately 3,000 adult males. The analyses were performed on a set of 28 balanced half samples, the sample being treated as if two independent selections had been made within each of 27 strata, together with the complementary set of 28 half samples. Two different types of estimates were considered:

1. Ratio estimates of population means were made for 15 physical body characteristics (e.g., chest girth, waist girth, and knee height).

2. Each of eight body characteristics were separately regressed on the independent variables of age, height, and weight. Estimates were made of the partial regression coefficients and of the multiple correlation coefficient.

In each case—15 means, 24 partial regression coefficients, and eight multiple correlation coefficients—the average of the estimates made from the original set of 28 balanced half samples can be compared with the average of the estimates made from the complementary set. These comparisons are given in tables 1 and 2. In only seven of the 47 comparisons do the two means differ by more than .03 percent of their average and these instances could well have arisen from rounding error. Hence we conclude that for this situation half-sample estimates, based on a set of 28 half samples, effectively give the same results as would be obtained by working with the entire set of 2^{27} half samples.

[1]Ten groups of four columns were selected randomly from the population of ${}_{21}C_4$ possible groups of four columns. Perfect balance was achieved by combining the r_i and r_i' sets in five of the 10. Improved balance was achieved in four of the 10; and in only one of the 10 did no improvement occur.

Bias of a Ratio Estimator

The estimate that is customarily used for the type of sample design to which the present discussion refers is the combined ratio estimator which, when based on the total sample, will be designated by $\hat{R}$. This estimator may be biased, especially for small samples, and this topic has been extensively researched in a wide variety of ways, e.g., Kish, Namboodiri, and Pillai (1962) and Tin (1965). When balanced half samples are used for variance estimation, it is possible to obtain, as a byproduct, an empirical check on the existence and magnitude of bias.

The quantities $\bar{r}$, $\bar{r}'$, and their average, denoted by $\bar{r}^*$, are of exactly the same form as $\hat{R}$, but are based on half samples. Since the bias of a ratio estimator decreases with increasing sample size, we can expect $\bar{r}^*$ to be subject to greater bias than $\hat{R}$. If, for a given set of data, $\bar{r}$, $\bar{r}'$, $\bar{r}^*$ and $\hat{R}$ are essentially the same, this is presumptive evidence that the differential bias is close to zero and it therefore follows that the absolute bias of either estimator is essentially zero. For the artificial example used in NCHS report (Series 2, No. 14, p. 27), this is clearly not the case since $E(\bar{r}) = 1.1985$, $E(\bar{r}') = 1.1968$, $E(\hat{R}) = 1.1666$, and $R = 1.1548$. It is true that the bias in $\hat{R}$ may be small even though there is differential bias between $\bar{r}$ and $\hat{R}$.

On the basis of previously cited empirical investigations and on the basis of the size of the sample upon which the illustrative data of this paper are based, one would expect little, if any, evidence of bias. This expectation is borne out by the summary results presented in table A, which is based on the individual values given in column 7 of table 1 and column 8 of table 2. As can be observed from the table, the differential bias of the 15 ratio estimates of population means is negligible, although one does observe that the sign of the difference is negative in 12 of the 15 cases, possibly indicating some small residual bias in the half-sample estimators. The results for the estimates of partial regression coefficients and multiple correlation coefficients lead to the same general conclusion, although the

Table A. Differential bias of $\bar{r}^*$ and $\hat{R}$, as a fraction of the estimated standard error of $\hat{R}$ [1]

$(\bar{r}^* - \hat{R})/s(\hat{R})$	Ratio estimates of population means	Estimates of partial regression coefficients and multiple correlation coefficients
-.30 to -.25-----	-	1
-.25 to -.20-----	-	1
-.20 to -.15-----	-	-
-.15 to -.10-----	-	2
-.10 to -.05-----	-	4
-.05 to .00------	13	11
.00 to +.05------	2	5
+.05 to +.10-----	-	7
+.10 to +.15-----	-	1
Total-----------	15	32
Mean------------	-.016	-.018
Median----------	-.020	-.011

[1] See the text for a description of the data upon which these estimates of differential bias are used. Individual values from which these distributions are derived are given in column 7 of table 1 and column 8 of table 2.

values of the differential bias (relative to the standard error) tend to be somewhat larger than in the former case. The three largest (in absolute value) of these arise from the independent variable, height, but no specific reason for this could be found. In general, one must conclude that, for this type of data and for this size of sample, there is no need to be concerned about bias in either the half-sample estimators or, more particularly, in the entire-sample estimator.

If the possibility of "serious" bias appears to exist for any or all of $\bar{r}$, $\bar{r}'$, $\bar{r}^*$ and $\hat{R}$, as evidenced by large differences among these quantities in a specific situation, one can consider the alternative of moving to a Quenouille-type (or Jackknife-type) transformation. There is some evidence, primarily based on Taylor series investigations of the ordinary ratio estimate, that such transformations may reduce bias without having deleterious effects on the variance of the estimate. (See, for example, Brillinger (1964), Durbin (1959), NCHS (Series 2, No. 14), Quenouille (1956), Rao (1965), and Tin (1965). There have been no theoretical or empirical investigations of complex design and estimation situations similar to the one being discussed in this paper.

Estimating the Variance of $\bar{r}$

Under ordinary circumstances, one has a set of k balanced half samples and the accompanying half-sample ratio estimates, $r_1, r_2, \ldots, r_k$. Carrying over the argument outlined earlier for the linear case, the variance of $\hat{R}$ (ratio estimate computed from the entire sample) would be estimated as $\sum_{i=1}^{k}(r_i - \hat{R})^2/k$. To keep the argument in simplest terms, we shall first restrict attention to the problem of estimating the variance of $\bar{r}$ instead of $\hat{R}$ and hence use the variance estimate $\sum_{i=1}^{k}(r_i - \bar{r})^2/k$, here denoted by $\hat{V}_{BHS}(\bar{r})$. If the data for the complementary set of half samples are also processed, as is the case for the illustrative example being used in this paper, then a second estimate of the same form can be computed from $r_1', r_2', \ldots, r_k'$, namely $\hat{V}_{BHS}(\bar{r}')$. In the linear case, these two estimates are identical.

Another estimate of variance with which to compare $V_{BHS}(\bar{r})$ and $V_{BHS}(\bar{r}')$ may be obtained by the following argument. Viewing a particular half sample and its complement as two independent samples, the variance of a half-sample ratio can be estimated with a single degree of freedom as

$$(r_i - \frac{r_i + r_i'}{2})^2 + (r_i' - \frac{r_i + r_i'}{2})^2$$

$$= (\tfrac{1}{2})(r_i - r_i')^2$$

The average of r_i and r_i', say $\bar{r}_i^*$, makes use of all the sample information and has estimated variance

$$(\tfrac{1}{4})(r_i - r_i')^2$$

If this estimate of variance is obtained from a single half sample and its complement, then it is the simplest form of a replicated estimate of variance and has often been suggested as a crude first approximation to such variances. In the linear case, the value of $\bar{r}_i^*$ would be equal to the whole-sample estimator for any i and hence we shall regard $(1/4)\,(r_i - r_i')^2$ as an approximation to the variance of the whole-sample estimator, whether it be computed as $\bar{r}$, $\bar{r}'$, or as $\bar{r}^*$.

If the foregoing estimate of variance could be computed for each of the 2^{L-1} complementary pairs of half samples in the entire set of 2^L half samples, then the average of these estimates, each of which has the correct expected value, would appear to provide an "excellent" estimate of the variance $\bar{r}^*$. Since this is impossible, we shall substitute its computation over a set of k balanced half samples, together of course with the complementary set. Thus

$$\hat{V}_{CBHS}(\bar{r}^*) = (1/4)\sum_{i=1}^{k}(r_i - r_i')^2/k\ .$$

It seems likely, on the basis of the argument previously set forth on the balancing of singles, pairs, and triplets, and of computations carried out on simple examples, that these estimates are extremely close to those that would be obtained by working with the entire set of 2^L half samples.

Although no way has been found to put the foregoing considerations on a formal basis, it would appear that $\hat{V}_{CBHS}(\bar{r}^*)$ is a "better" estimate of variance than either $\hat{V}_{BHS}(\bar{r})$ or $\hat{V}_{BHS}(\bar{r}')$. In practice, of course, one would have available only one of the latter two estimates. In the

Table B. Difference between $\hat{V}_{BHS}(\bar{r})$, or $\hat{V}_{BHS}(\bar{r}')$, and $\hat{V}_{CBHS}(\bar{r}^*)$, as a fraction of $\hat{V}_{CBHS}(\bar{r}^*)$

Fractional difference	Ratio estimates of population means		Estimates of partial regression coefficients and multiple correlation coefficients	
	$\hat{V}_{BHS}(\bar{r})$	$\hat{V}_{BHS}(\bar{r}')$	$\hat{V}_{BHS}(\bar{b})$	$\hat{V}_{BHS}(\bar{b}')$
-.08 to -.07	-	-	-	1
-.07 to -.06	-	-	-	-
-.06 to -.05	-	-	2	1
-.05 to -.04	-	-	1	1
-.04 to -.03	2	-	1	2
-.03 to -.02	5	-	2	2
-.02 to -.01	2	-	2	6
-.01 to .00	2	2	4	5
.00 to .01	2	3	2	2
.01 to .02	1	3	6	4
.02 to .03	1	5	5	1
.03 to .04	-	1	-	2
.04 to .05	-	1	4	2
.05 to .06	-	-	-	1
.06 to .07	-	-	2	2
.07 to .08	-	-	-	-
.08 to .09	-	-	1	-
Total	15	15	32	32
Mean	-.011	+.019	+.010	+.001
Median	-.018	+.017	+.012	-.001

present investigation, all three estimates have been computed for the 15 ratio estimates of population means and for the previously described regression coefficents and multiple correlation coefficients. Agreement of these three variance estimators will be taken as presumptive evidence that the balanced half-sample variance estimation procedure is appropriate for a given set of data. A summary of the results of this comparison is given in table B, while the actual values on which these distributions are based are presented in tables 3 and 4. It will be observed that the distributions of fractional differences show the three estimates to be in relatively close agreement. Several further observations concerning these distributions are:

1. For any given estimate $\hat{V}_{CBHS}(\bar{r}^*)$ tends to be between $\hat{V}_{BHS}(\bar{r})$ and $\hat{V}_{BHS}(\bar{r}')$.

2. For the ratio estimates of population means (physical body characteristics of U.S. adult males), the estimates $\hat{V}_{BHS}(\bar{r})$ tend to be on the same side of $\hat{V}_{CBHS}(\bar{r}^*)$. In other words, the values of $\hat{V}_{BHS}(\bar{r})$ are correlated with one another. This effect is not as pronounced for the estimates of partial regression coefficients and multiple correlation coefficients.

3. The effects of (1) and (2) show up in the two distributions on the left of table B, one distribution being located somewhat below zero and the other somewhat above zero.

4. If the comparison were made on the basis of estimated standard errors, the distributions would show less dispersion. This follows immediately from the fact that $(\sqrt{a}-\sqrt{b})/\sqrt{b}=(\sqrt{a}/\sqrt{b})-1$ is always smaller in absolute value than is $(a-b)/b=(a/b)-1$.

The foregoing results can be presented in a slightly different manner by introducing the concept of correlated variables. Walsh (1947) summarizes the following results. If $x_1, x_2, \ldots, x_n$ are observations where $E(x_i)=\mu$, $V(x_i)=\sigma^2$, and $E(x_i-\mu)(x_j-\mu)=\rho\sigma^2$, then $V(\bar{x})=\frac{\sigma^2}{n}[1+(n-1)\rho]$ and $E[\Sigma(x_i-\bar{x})^2/(n-1)]=\sigma^2(1-\rho)$. Furthermore, if the x_i's are normal, the quantity

$$\frac{(\bar{x}-\mu)[(1-\rho)/(1+(n-1)\rho)]^{1/2}}{[\Sigma(x_i-\bar{x})^2/n(n-1)]^{1/2}}$$

has a Student's distribution with $(n-1)$ degrees of freedom. For present purposes, we replace the x_i's by the balanced half-sample ratio estimators $r_1, r_2, \ldots, r_k$. The r_i's do have a common variance, and they are correlated since the half samples within a balanced set do have elements in common. For the moment, the complementary set is ignored.

The ordinary practice for balanced half-sample replication has been to take $\Sigma(r_i-\hat{R})^2/k$ as an estimate of the variance of $\hat{R}$. As has been done earlier in this discussion, let us consider $\bar{r}$ rather than $\hat{R}$ and see under what circumstances we can treat

$$\frac{\bar{r}-\mu_r}{\sqrt{\frac{\Sigma(r_i-\bar{r})^2}{k}}}$$

as having a Student's t distribution with $(k-1)$ degrees of freedom, at least as far as variance considerations are concerned.

If the aforementioned equality is to hold, then it is necessary that

$$\frac{1+(k-1)\rho}{(1-\rho)}=(k-1)$$

which means that

$$\rho=\frac{k-2}{2(k-1)}.$$

This condition can be shown to hold, through the use of a common-element argument set forth in NCHS (Series 2, No. 14, pp. 20-21), if

1. We are dealing with a linear situation in which $W_1=W_2=\ldots=W_L$, and $S_1^2=S_2^2=\ldots=S_L^2$.

2. The number of strata, L, is one less than a multiple of four and the set of $k=L+1$ balanced half samples is obtained by deleting the column whose sign entries are all the same from a $k\times k$ orthogonal matrix.

In the present instance, and in more complicated ones, these conditions cannot be expected to hold exactly. The strata weights and variances will not be equal; the half samples may not have exactly the same number of elements in common, which was required for the "common element" argument leading to a theoretical value of ρ; and a ratio estimator may behave in a different fashion from a linear estimator. Nevertheless, one would expect these results to hold in some average sense, especially when there is a reasonable number of strata. It is possible to investigate this matter empirically.

Suppose we consider a set of balanced half samples and their complements, together with the half-sample ratio estimators $r_1, r_2, \ldots, r_k$ and $r'_1, r'_2, \ldots, r'_k$. Then take

$$2\hat{V}_{CBHS}(\bar{r}^*) = (1/2)\sum_{i=1}^{k}(r_i - r'_i)^2/k$$

as an estimate of σ^2, the variance of a single half-sample estimator. Furthermore, take

$$\frac{\sum_{i=1}^{k}(r_i - \bar{r})^2}{k-1} \quad \text{and} \quad \frac{\sum_{i=1}^{k}(r'_i - \bar{r}')^2}{k-1}$$

as estimators of $\sigma^2(1-\rho)$. The correlation ρ can then be estimated as

$$\hat{\rho} = 1 - \frac{\sum_{i=1}^{k}(r_i - \bar{r})^2/(k-1)}{(1/2)\sum_{i=1}^{k}(r_i - r'_i)^2/k}$$

or as

$$\hat{\rho}' = 1 - \frac{\sum_{i=1}^{k}(r'_i - \bar{r}')^2/(k-1)}{(1/2)\sum_{i=1}^{k}(r_i - r'_i)^2/k}$$

and these values, or their average, can be checked against the theoretical value of

$(k-2)/2(k-1)$.

The foregoing estimates of ρ have been computed for the illustrative data used in this paper. Their distributions are shown in table C and the individual values are given in tables 3 and 4. The left portion of table C presents the results for ratio estimates of mean body characteristics of U.S. adult males, while the right portion gives the same data for regression coefficients and multiple correlation coefficients. The theoretical value of ρ with which these can be compared is $(k-2)/2(k-1) = 26/54 = .4815$. The general agreement of the empirical results with the theoretical value again leads to the conclusion that, for the illustrative data being used in this paper, one can take $\sum_{i=1}^{k}(r_i - \bar{r})^2/k$ as an approximation to

$$\frac{\sum_{i=1}^{k}(r_i - \bar{r})^2}{k(k-1)} \cdot \frac{[1+(k-1)\rho]}{(1-\rho)}$$

in estimating the variance of $\bar{r}$. The general observations made concerning the relationship among distributions in table B also apply here since the $\hat{\rho}$'s are simple functions of $\hat{V}_{BHS}(\bar{r})$, $\hat{V}_{BHS}(\bar{r}')$, and $\hat{V}_{CBHS}(\bar{r}^*)$.

Estimating the Variance of $\hat{R}$

Thus far the discussion of variance estimation has focused on $\bar{r}$, $\bar{r}'$, and $\bar{r}^*$, rather than $\hat{R}$, for the simple reason that we are actually dealing with half-sample ratio estimators. The ordinary procedure in either half-sample or balanced half-sample replication has been to use the quantity $\sum_{i=1}^{k}(r_i - \hat{R})^2/k$ as an estimator of the variance of $\hat{R}$. This estimate of variance is larger than the estimate $\sum_{i=1}^{k}(r_i - \bar{r})^2/k$ by the amount $(\bar{r} - \hat{R})^2$. Since we have already observed that $\bar{r}$ and $\bar{r}'$ do not differ by an appreciable amount from $\hat{R}$ for the illustrative data of this paper, it makes little difference which estimate is used. In general, it would seem that $\sum_{i=1}^{k}(r_i - \hat{R})^2/k$ would overestimate even the mean square error of $\hat{R}$ since it contains a contribution for the differential bias of $\bar{r}$ and $\hat{R}$, and this differential bias may well be larger than the absolute bias in $\hat{R}$.

It is possible to obtain a formal relationship connecting $V(\bar{r})$ and $V(\hat{R})$. Consider the expression $E(\bar{r} - \hat{R})^2$. It follows that

$$\begin{aligned} E(\bar{r}-\hat{R})^2 &= V(\bar{r}) + V(\hat{R}) - 2\,cov(\bar{r},\hat{R}) \\ &\quad + [E(\bar{r}) - E(\hat{R})]^2 \\ &= V(\bar{r}) + V(\hat{R}) - 2\,\rho_{\bar{r},\hat{R}}\sqrt{V(\bar{r})}\sqrt{V(\hat{R})} \\ &\quad + [E(\bar{r}) - E(\hat{R})]^2 . \end{aligned}$$

Table C. Estimated value of the intraclass correlation coefficient for a set of 28 balanced half-sample estimators

Estimated value of ρ	Ratio estimates of population means		Estimates of partial regression coefficients and multiple correlation coefficients	
	$\hat{\rho}$	$\hat{\rho}'$	$\hat{\rho}$	$\hat{\rho}'$
.43 to .44	-	-	1	-
.44 to .45	-	-	2	2
.45 to .46	-	1	3	2
.46 to .47	-	6	5	4
.47 to .48	3	4	8	6
.48 to .49	5	4	6	10
.49 to .50	6	-	4	5
.50 to .51	1	-	3	1
.51 to .52	-	-	-	1
.52 to .53	-	-	-	1
Total	15	15	32	32
Mean	.4876	.4734	.4762	.4811
Median	.4890	.4724	.4754	.4838

Regrouping of terms leads to

$$V(\bar{r}) + V(\hat{R}) - 2\rho_{\bar{r},\hat{R}}\sqrt{V(\bar{r})}\sqrt{V(\hat{R})}$$
$$= E(\bar{r}-\hat{R})^2 - [E(\bar{r}) - E(\hat{R})]^2 .$$

Now we observe that $\rho_{\bar{r},\hat{R}} \leq 1$. Actually, it appears that this correlation will be very close to one under almost any circumstances, and this conjecture has been confirmed by direct computation in several "small," synthetic examples. Thus

$$V(\bar{r}) + V(\hat{R}) - 2\sqrt{V(\bar{r})}\sqrt{V(\hat{R})}$$
$$\leq E(\bar{r}-\hat{R})^2 - [E(\bar{r}) - E(\hat{R})]^2$$
$$\leq E(\bar{r}-\hat{R})^2$$

or

$$\left[\sqrt{V(\bar{r})} - \sqrt{V(\hat{R})}\right]^2 \leq E(\bar{r}-\hat{R})^2.$$

Since, as shown earlier, we can estimate $V(\bar{r})$ from a single sample, and since $(\bar{r}-\hat{R})^2$ is an approximation for $E(\bar{r}-\hat{R})^2$ this last expression provides a crude way of comparing $V(\bar{r})$ and $V(\hat{R})$ on the basis of a single sample. The bounds obtained will not be particularly good if the neglected term $[E(\bar{r}) - E(\hat{R})]^2$ is at all appreciable in relation to $E(\bar{r}-\hat{R})^2$, and this has been the case with various synthetic examples that have been investigated. On the other hand, this may provide quite reasonable bounds in practical situations where the neglected term may be expected to be small. It should be observed that the development of the final expression includes sampling without replacement from finite populations. Furthermore, equality will hold throughout if $(r_i + r_i')/2 = \hat{R}$ for each i, as it does in the linear case, and we then have the result that $V(\bar{r}) = V(\hat{R})$. For the illustrative data of this paper, as summarized in table A, all indications are that we are obtaining "good" estimates of $V(\hat{R})$ as well as of $V(\bar{r})$.

Design Effect

As a final point, we note that column 2 in table 3 and column 3 in table 4 contain estimates of variance, labeled $\hat{V}_{RAN}(.)$, computed as if the observations had arisen from simple random sampling. Since the estimate $\hat{V}_{CBHS}(\cdot)$ takes into account most of the survey features (stratification, clustering, poststratification estimation, etc.), the ratio of $\hat{V}_{CBHS}(\cdot)$ to $\hat{V}_{RAN}(\cdot)$ is an estimate of the design effect. These ratios are given in column 8 of table 3 and column 9 in table 4. In table 3, which refers to estimates of mean body characteristics based on approximately 3,000 adult males, these ratios range from 1.158 to 5.729 with an average value of 3.223. In table 4, which refers to estimates of regression coefficients and multiple correlation coefficients, the ratios range from .570 to 3.675 with an average value of 1.807. The fact that design effect is smaller for the regression coefficients than for the mean body characteristics is consistent with other studies.

APPROXIMATE TEST OF INDEPENDENCE BASED ON A SET OF BALANCED HALF-SAMPLE REPLICATES

The statistical analysis of data that have been produced by the use of a complex sample design and then processed through an involved estimation procedure poses many difficult problems. Although it appears that the application of some form of pseudoreplication will provide a reasonable solution to the problem of estimating the sampling variability of a statistic, it is not immediately apparent that one can use these same techniques to solve certain other commonly occurring problems of statistical analysis. This section gives a summary of the results of an investigation carried out by Chapman (1966) concerning the use of balanced half-sample replications to test the hypothesis of independence in a contingency table.

The contingency-table test of independence is ordinarily phrased in the following manner. Observations are classified jointly on each of two qualitative variables, the categories of variable one being denoted by i, $i = 1, 2, \ldots, r$, and those of variable two by j, $j = 1, 2, \ldots, c$. If p_{ij} denotes the probability of an observation falling in the ith category of variable one and in the jth category of variable two, then the hypothesis of independence states that $p_{ij} = p_{i.}\, p_{.j}$ where $p_{i.}$ refers to the marginal distribution of the first variable and $p_{.j}$ refers to the marginal distribution of the second variable. Assuming that a "large" sample of n independent observations is available, the statistic

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(n_{ij} - \frac{n_{i.}\, n_{.j}}{n}\right)^2}{\frac{n_{i.}\, n_{.j}}{n}}$$

is approximately distributed as Chi-square with $(r-1)(c-1)$ degrees of freedom when the hypothesis is true, and the hypothesis is rejected for "large" values of X^2. Data derived from a complex survey operation certainly do not conform to this model, particularly in view of the dependencies that are introduced by the use of stratification and clustering techniques. Furthermore, the effects of commonly used estimation procedures would add to these complications.

One might attempt to find a solution to this problem in the following manner. If large samples are used, the distribution of the sample estimates of $(rc-1)$ of the p_{ij}'s, which will be denoted by $\hat{p}_{ij}$, can be approximated by a nondegenerate, normal, multivariate distribution. The exponent (ignoring the - 1/2) in this distribution is a quadratic form in the $\hat{p}_{ij}$'s and the coefficients of this form are the elements in the inverse of the variance-covariance matrix of the $\hat{p}_{ij}$'s. The variance-covariance matrix of the $\hat{p}_{ij}$'s can be estimated by a pseudoreplication method, such as balanced half samples, and these estimates would mirror the effects of the design and estimation. It should be observed, however, that the volume of computation would be large since even for $r = 3$ and $c = 3$ there would be eight variances and 28 covariances. It would then be necessary to invert this matrix and evaluate the quadratic form to take the hypothesis $p_{ij} = p_{i.}\, p_{.j}$ into account. The resulting statistic would be treated as a Chi-square variable with some appropriate number of degrees of freedom. This approach has not been pursued since we shall now describe a "simpler" alternative.

The basic idea upon which the proposed test procedure rests is the following. Suppose that an estimate $\hat{p}_{ij}$ is obtained from a sample and that a product estimate, $\hat{p}_{i.}\, \hat{p}_{.j}$, is obtained from an independent sample. Form the difference of these two estimates $(\hat{p}_{ij} - \hat{p}_{i.}\ \hat{p}'_{.j})$, and consider the sign of the difference. Define the variable S_{ij} to be one when the sign is plus and zero when it is minus. If the hypothesis of independence does hold, then this is the difference of two independent estimates of the parameter, p_{ij}, and one would intuitively expect the sign of this difference to be plus or minus with equal probability. Actually, even with two independent random samples, each of size n, it can be shown that this result does not generally hold. In particular,

1. If n is "small," then there may be an appreciable probability that the difference is exactly equal to zero.
2. The distributions of $\hat{p}_{ij}$ and $\hat{p}'_{i.}\ \hat{p}'_{.j}$ will not be identical and therefore the distribution of the difference may not have a median at zero. For example, it can be demonstrated that

$$V(\hat{p}_{ij}) = (1/n)\, p_{ij}(1 - p_{ij})$$

$$V(\hat{p}'_{i.}\ \hat{p}'_{.j}) = (1/n)\, p_{ij}(1 - p_{ij}) - [(n-1)/n^2]\, p_{ij}(1 - p_{i.})(1 - p_{.j}).$$

On the other hand, if we can assume that both $\hat{p}_{ij}$ and $\hat{p}'_{i.}\ \hat{p}'_{.j}$ are approximately normally distributed, then their difference will also be approximately normally distributed and the mean of this distribution will be zero when the hypothesis of independence is true. Hereafter, we shall make this assumption of approximate normality.

Now suppose that it were possible to replicate this situation k times and thus obtain k independent observations on the variable S_{ij}. The sum of the k S_{ij}'s, X_{ij}, is a binomial variable with

$$E(X_{ij}) = (k/2)$$

$$V(X_{ij}) = (1/4)k$$

when the null hypothesis is true and, if k is "large," the quantity

$$\frac{(X_{ij} - \frac{k}{2})^2}{(1/4)k}$$

is distributed approximately as Chi-square with one degree of freedom.

In general it would be impossible to consider using this form of analysis since it requires $2k$ independent replicates, each of which is large enough to ensure that $\hat{p}_{ij}$ and $\hat{p}'_{i.}\ \hat{p}'_{.j}$ are approximately normally distributed. Furthermore, it is also desirable that k be sufficiently large so that the test will have adequate power. However, if one is using a sample design for which a balanced half-sample replication method of estimating variances is appropriate, it appears that these circumstances can be realized in an approximate manner. Consider, as an example, a situation in which there are two independent selections from within each of seven strata and that one is therefore using a set of eight balanced half samples. Denoting the first element selected in each stratum by a + and the second by a −, the following table shows one set of eight balanced half samples and their complements:

Half samples	*Strata* 1	2	3	4	5	6	7	*Complements*	*Strata* 1	2	3	4	5	6	7
1	+	+	+	+	+	+	+	1	−	−	−	−	−	−	−
2	−	+	−	+	−	+	−	2	+	−	+	−	+	−	+
3	+	−	−	+	+	−	−	3	−	+	+	−	−	+	+
4	−	−	+	+	−	−	+	4	+	+	−	−	+	+	−
5	+	+	+	−	−	−	−	5	−	−	−	+	+	+	+
6	−	+	−	−	+	−	+	6	+	−	+	+	−	+	−
7	+	−	−	−	−	+	+	7	−	+	+	+	+	−	−
8	−	−	+	−	+	+	−	8	+	+	−	+	−	−	+

It will be observed that a half sample and its complement have no elements in common, that a half sample and any other half sample in the same set have three elements in common, and that a half sample and any half sample from the complementary set, other than the complement itself, have four elements in common. In general, if there are L strata and one can use a set of $(L + 1) = k$ balanced half samples, where k is a multiple of four, these numbers of common elements will be zero, $(k/2) - 1$, and $(k/2)$, respectively.

Now consider the following simplified version of the stated problem:

1. ${}_1\hat{p}_{ij}$ is the estimate of p_{ij} made from half-sample one and ${}_1\hat{p}'_{i.}\ {}_1\hat{p}'_{.j}$ is the corresponding product estimate obtained from the complementary half sample. These are independent estimators.

2. ${}_2\hat{p}_{ij}$ is the estimate of p_{ij} made from half sample two and ${}_2\hat{p}'_{i.}\ {}_2\hat{p}'_{.j}$ is the corresponding product estimate obtained from the complementary half sample.

3. Assume that ${}_1\hat{p}_{ij}$ and ${}_2\hat{p}_{ij}$ are ordinary linear estimates based on L strata of equal weight and that the within-strata estimates have common variance S_w^2. It then follows, using the argument set forth in NCHS (Series 2, No. 14, p. 21), that

$$Cov({}_1\hat{p}_{ij}, {}_2\hat{p}_{ij}) = \left\{[(L-1)/2]\right\}\left\{1/L^2\right\}S_w^2$$

$$= \frac{(L-1)}{2L^2}\ S_w^2 .$$

4. Assume that the product estimates ${}_1\hat{p}'_{i.}\ {}_1\hat{p}'_{.j}$ behave in exactly the same manner as the $\hat{p}_{ij}$'s. As a result, we have

$$Cov({}_1\hat{p}'_{i.}\ {}_1\hat{p}'_{.j}, {}_2\hat{p}'_{i.}\ \hat{p}'_{.j}) = \frac{(L-1)}{2L^2}\ S_w^2$$

and

$$Cov({}_1\hat{p}'_{ij}, {}_2\hat{p}'_{i.}\ {}_2\hat{p}'_{.j}) = Cov({}_2\hat{p}'_{ij}, {}_1\hat{p}'_{i.}\ {}_1\hat{p}'_{.j})$$

$$= \left\{(L+1)/2\right\}\left\{1/L^2\right\}S_w^2$$

$$= \frac{(L+1)}{2L^2}\ S_w^2 .$$

5. Using the foregoing results and assumptions, we find that

$$Var({}_1\hat{p}_{ij} - {}_1\hat{p}'_{i.}\ {}_1\hat{p}'_{.j}) = Var({}_2\hat{p}_{ij} - {}_2\hat{p}'_{i.}\ {}_2\hat{p}'_{.j}) = \frac{2}{L}\ S_w^2$$

$$Cov[({}_1\hat{p}_{ij} - {}_1\hat{p}'_{i.}\ {}_1\hat{p}'_{.j}), ({}_2\hat{p}_{ij} - {}_2\hat{p}'_{i.}\ {}_2\hat{p}'_{.j})] = -\frac{2}{L^2}\ S_w^2$$

and that the correlation between the two differences is equal to $-1/L$. For any reasonable number of strata this correlation will be very close to zero. Since the various estimates are assumed to be approximately normally distributed, it follows that the k differences $(\hat{p}_{ij} - \hat{p}'_{i.}\ \hat{p}'_{.j})$, as computed from a set of balanced half samples, can be treated as independent observations. Some evidence concerning the reasonableness of this type of argument is provided by the data on the intraclass correlation coefficient given in table C. It was shown that the estimated values agreed quite well with the theoretical value predicted on the basis of the "common element" approach.

The application of the foregoing procedure to a set of k balanced half samples leads to an observed value of X_{ij} for each of the rc cells in the contingency table, where

$$E(X_{ij}) = k/2 \qquad \left\{\begin{array}{l} i = 1, 2, \ldots, r \\ j = 1, 2, \ldots, c \end{array}\right\}$$

$$V(X_{ij}) = (1/4)k$$

when the hypothesis of independence is true. We would, of course, like to combine these rc values into one overall test. However, such a combination cannot be carried out in a straightforward manner since the X_{ij} are not independent of one another. Chapman (1966) presents an extensive analysis of this problem which culminates in the suggestion of an approximate test statistic that appears to have reasonable properties and which is simple to use in practice. The basic steps in the development of the statistic are:

1. The rc variables X_{ij} are assumed to have a nondegenerate, normal, multivariate distribution. Although there are constraints on the X_{ij}'s, namely that they cannot all be simultaneously equal to zero or one, these constraints are not of a form that reduces the rank of the quadratic form.

2. Approximations to the covariance and the correlation of any two X_{ij}'s are obtained in the following manner:

   a. If two normal variables, with zero means and unit variances, Y_i and Y_j, have a normal bivariate distribution with correlation $\rho(Y_i, Y_j)$, then the correlation $\rho(S_i, S_j)$ between their signs—i.e., $S_i = 1$ if $Y_i > 0$, and $S_i = 0$ if $Y_i < 0$, and similarly for Y_j —is given by

$$\rho(S_i, S_j) = \frac{2}{\pi} \text{ arc sin } \rho(Y_i, Y_j).$$

This result is given in Cramér (1946, p. 290).

b. Assuming that $\hat{p}_{ij}$ and $\hat{p}'_{i.}\hat{p}'_{.j}$ are based on random samples, each of size n, from a population in which $p_{ij} = (1/rc)$, $p_{i.} = (1/r)$, $p_{.j} = (1/c)$, it can be shown that the *average* correlation between $(\hat{p}_{ij} - \hat{p}'_{i.}\hat{p}'_{.j})$ and $(\hat{p}_{ht} - \hat{p}'_{h.}\hat{p}'_{.t})$, where either or both of $i \neq h$ and $j \neq t$ are true, is equal to $-1/(rc-1)$.

c. Thus we have the result that

$$\rho(X_{ij}, X_{ht}) \doteq \frac{2}{\pi} \text{ arc sin } [-1/(rc-1)]$$

where either or both of $i \neq h$ and $j \neq t$ hold. Chapman (1966) investigates some aspects of this approximation procedure. It appears to be a reasonable one except under extreme circumstances, e.g., where one or more of the p_{ij} are extremely small and other p_{ij} are extremely large.

3. Walsh (1947) proves that if one has n normal variables, $X_1, X_2, \ldots, X_n$, each with mean μ and variance σ^2 and with a common correlation coefficient ρ then

$$T = \frac{1}{\sigma^2(1-\rho)} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

has a χ^2 distribution with $(n-1)$ degrees of freedom and

$$T' = \frac{n(\bar{X} - \mu)^2}{\sigma^2[1 + (n-1)\rho]}$$

has an independent χ^2 distribution with one degree of freedom.

4. Combining the preceding results, we obtain the test statistic

$$T = \frac{\sum_{i=1}^{r} \sum_{j=1}^{r} (X_{ij} - \bar{X})^2}{(1/4)k\left[1 - \frac{2}{\pi} \text{ arc sin}\left(-\frac{1}{rc-1}\right)\right]}$$

which, when the hypothesis of independence is true, has approximately a χ^2 distribution with $(rc-1)$ degrees of freedom. The hypothesis of independence is rejected for "significantly" large values of T. Since arc sin $x \doteq x$ when x is small, this can usually be simplified to

$$T = \frac{\sum_{i=1}^{r} \sum_{j=1}^{c} (X_{ij} - \bar{X})^2}{(1/4)k\left[1 + (2/\pi)(1/(rc-1))\right]}.$$

5. The use of the independent Chi-square with one degree of freedom

$$T' = \frac{rc(\bar{X} - \frac{k}{2})^2}{(1/4)k\left[1 + (rc-1)\frac{2}{\pi} \text{ arc sin}\left(-\frac{1}{rc-1}\right)\right]}$$

is not recommended since $E(\bar{X})$ will tend to be close to $k/2$ whether the hypothesis is true or not. Under these circumstances, the inclusion of T' in the test statistic will weaken the power of the test.

ORDER STATISTICS OF A SET OF BALANCED HALF-SAMPLE ESTIMATES

A discussion of the relationship between balanced half-sample replication and the sign test is presented in NCHS (Series 2, No. 14, pp. 20-23). The problem can briefly be stated as follows. Suppose that each half-sample replication provides an estimate, say $d_{hs,i}$, of a population difference, and that one wishes to test the hypothesis that the $d_{hs,i}$'s were drawn from a population whose median is equal to zero. If the $d_{hs,i}$'s were independent of one another, then one solution to this problem would be obtained by a straightforward application of the ordinary sign test. The half-sample estimates are not, however, independent of one another, as is observed in the earlier sections of this paper, and the question can be raised of whether it is possible to devise some suitable modification of the sign test.

As soon as one introduces dependence among the members of a set of observations, it is no longer possible in general to devise nonparametric or distribution-free procedures. If, however, it is reasonable to assume that the $d_{hs,i}$'s have a normal multivariate distribution, then existing tables can, under certain circumstances, be applied to test the stated hypothesis. Thus, Gupta (1963, p. 817) provides a table which gives the "probability that N standard normal random variables with common correlation ρ are simultaneously less than or equal to H." In particular, if H is taken to be zero, the appropriate probabilities form the basis for obtaining the levels of significance for an "extreme" sign test, i.e., all signs plus or all signs minus.

For any reasonably large number of balanced half-sample estimates, the Gupta tables are not especially helpful. To obtain meaningful levels of significance, one must be able to move in from the extremes on the distribution of the number of plus signs. Since such tables of the normal multivariate distribution do not presently exist, we shall here present a procedure for approximating these probabilities. The degree of approximation appears to be quite acceptable for ordinary applications of the sign test.

Consider n independent and normally distributed variables $X_1, X_2, \ldots, X_n$ with $E(X_i) = 0$ and $E(X_1^2) = 1$. If the collection $\{X_i\}$ is ordered so that $X^{(1)} \geq X^{(2)} \geq X^{(n)}$, then certain moments of the $X^{(i)}$ have been extensively tabulated. In particular, Teichroew (1956) has tabulated $E(X^{(i)})$ and $E(X^{(i)}X^{(j)})$ for $n = 1\ (1)\ 20$ and Harter (1960) has tabulated $E(X^{(i)})$ for $n = 2\ (1)\ 100$, plus values for a number of additional $n \leq 400$. Furthermore, Owen and Steck (1962) have shown that the moments of the order statistics of a sample from the equicorrelated, multivariate, normal distribution can be readily obtained from the corresponding moments for n independent variables. Thus, if $Z_1, Z_2, \ldots, Z_n$ are jointly distributed random variables with

$$E(Z_i) = 0,\ E(Z_i^2) = 1, \text{ and } E(Z_i Z_j) = \rho \text{ for } i \neq j,$$

$$E(Z^{(i)}) = (1-\rho)^{1/2} E(X^{(i)})$$

and

$$E[(Z^{(i)}) - E(Z^{(i)})]^2 = \rho + (1-\rho)\, E[X^{(i)} - E(X^{(i)})]^2.$$

Our present concern is with the probability that U out of the n values $Z_1, Z_2, \ldots, Z_n$ are positive, where U takes on the values $0, 1, 2, \ldots$. For example, for the first three values, it is clear that

$$Pr(U = 0) = Pr(Z^{(1)} \leq 0)$$

$$Pr(U = 0 \text{ or } 1) = Pr(Z^{(2)} \leq 0)$$

$$Pr(U = 0 \text{ or } 1 \text{ or } 2) = Pr(Z^{(3)} \leq 0).$$

Under ordinary circumstances the calculation of the probabilities on the right side of these expressions would be extremely difficult since they must be obtained from the distributions of $Z^{(1)}, Z^{(2)}, \ldots$, and this is the extreme-value problem that has been considered in some detail by Gumbel (1958). In the present problem, however, moderate values of ρ, i.e., in the neighborhood of 1/2, appear to ensure that the distributions of the $Z^{(i)}$ can be approximated reasonably well by a normal distribution. Some third and fourth moments of the extreme order statistic, for various values of ρ, are exhibited by Owen and Steck (1962), and these values confirm this observation. Owen and Steck note that this behavior is to be expected as ρ increases since $Z^{(i)}$ has a unit normal distribution when $\rho = 1$.

As an illustration of this approach consider the case where $n = 20$ and $\rho = .50$. From Teichroew's table

$$E(X^{(1)}) = 1.86748$$
$$\sigma^2(X^{(1)}) = .27568,$$

therefore

$$E(Z^{(1)}) = 1.32051$$
$$\sigma^2(Z^{(1)}) = .63784.$$

Table D. Probability that U or fewer, out of n normal correlated observations, will be greater than their common mean (less than their common mean)

Number of obser-vations	16		20		24		48	
	.40	.50	.40	.50	.40	.50	.40	.50
U								
0----------	.0361 [1](.0335)	.0603 [1](.0588)	.0272 [1](.0251)	.0492 [1](.0476)	.0235 [1](.0197)	.0433 [1](.0440)	.0096 -	.0227 -
1----------	.0806	.1178	.0607	.0956	.0487	.0809	.0193	.0415
2----------	.1346	.1746	.1015	.1430	.0808	.1204	.0321	.0615
3----------	-	-	-	-	-	-	.0467	.0818
4----------	-	-	-	-	-	-	.0628	.1022
5----------	-	-	-	-	-	-	.0805	.1226

[1]The values in parentheses are exact and are from Gupta's tables.

Assuming a normal distribution for $Z^{(1)}$, we obtain

$$Pr(Z^{(1)} \leq 0) = .0492.$$

The correct value for this probability, from Gupta's tables, is .0476. This appears to be quite acceptable accuracy for ordinary purposes. Furthermore, the approximation should be better for $Z^{(2)}, Z^{(3)}, \ldots$ since the distributions of these variables cannot be as skewed as is the distribution of $Z^{(1)}$. For $n > 20$, it was also necessary to approximate $\sigma^2(Z^{(i)})$. This was done by converting $E(X^{(i)})$, from Harter's table, to a cumulative percentage and using the asymptotic variance formula for the percentage point of a distribution (see Wilks, 1962, p. 273).

The values of these probabilities for $n = 16$, 20, 24, and 48, and for $\rho = .40$ and .50 have been computed and are presented in table D. Where possible, the exact values from Gupta's table are given for purposes of comparison. Not only do these values lead to a sign test, but they also lead to confidence intervals for the population mean (or median). Thus with 16 half-sample estimates, and with a ρ of .50, the largest and smallest estimates provide a $1 - (2 \times .06) = 88$ percent confidence interval; similarly, the second smallest and second largest estimates provide a $1 - (2 \times .1178) = 76$ percent confidence interval.

In order to apply the foregoing theory it is, of course, necessary to have a value of ρ with which to enter the tables. The data given in table C on the intraclass correlation coefficient suggest that a reasonable approach is to assume that ρ has the value suggested by the "common element" argument. That is, if a set of k balanced half samples is used for $(k - 1)$ strata, ρ would be assigned the value $(k - 2)/2(k - 1)$.

Finally, we observe that if one is willing to use the set of complementary half samples, then still another approach to this problem is suggested by some of the earlier arguments of this report. Suppose one wishes to compare the means of two variables X and Y. Let $\hat{X}_i$ be the estimate of $\bar{X}$ made from the ith member of a set of k balanced half samples and $\hat{Y}_i'$ be the estimate of $\bar{Y}$ made from the complementary half sample. The development on p. 15 suggests that the correlation between any two of the k differences $(\hat{X}_i - \hat{Y}_i')$ will be very small and that it would not be unreasonable to treat them as k independent observations. Thus the ordinary sign-test tables could be applied to the signs of the k differences.

REFERENCES

Brillinger, D. R.: The asymptotic behavior of Tukey's general method of setting confidence limits (the Jacknife) when applied to maximum likelihood estimates. *Review of International Statistical Institute* 32: 202-206, 1964.

Chapman, D. W.: An Approximate Test of Independence Based on Replications of a Complex Sample Survey Design. M.S. thesis, New York State School of Industrial and Labor Relations, Cornell University, 1966.

Cochran, W. G.: *Sampling Techniques,* ed. 2. New York. John Wiley & Sons, Inc., 1963.

Cramér, H.: *Mathematical Methods of Statistics*. Princeton. Princeton University Press, 1946.

Deming, W. E.: *Some Theory of Sampling*. New York. John Wiley & Sons, Inc., 1950.

Deming, W. E.: *Sampling Design in Business Research.* New York. John Wiley & Sons, Inc., 1960.

Durbin, J.: A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika* 46: 447-480, 1959.

Durbin, J.: Design of multi-stage surveys for the estimation of sampling errors. *Applied Statistics* 16: 152-164, 1967.

Gumbel, E. J.: *Statistics of Extremes*. New York. Columbia University Press, 1958.

Gupta, S. S.: Probability integrals of multivariate normal and multivariate *t*. *Ann.math.Statist.* 34: 792-828, 1963.

Gurney, M.: The Variance of the Replication Method for Estimating Variances for the CPS Sample Design. Unpublished memorandum, U.S. Bureau of the Census, 1962.

Harter, H. L.: *Expected Values of Normal Order Statistics*. Aeronautical Research Lab. Technical Report 60-292. Wright-Patterson Air Force Base, Ohio, 1960.

Hartley, H. O.: *Analytic Studies of Survey Data*. Instituto di Statistica, Rome, Volume in onora di Corrado Gini. Ames, Iowa. Statistical Laboratory, Iowa State University of Science and Technology. Reprint Series 63. 1959.

Kish, L.: Confidence intervals for clustered samples. *American Sociological Review* 22(2): 154-165, Apr. 1957.

Kish, L.: *Survey Sampling*. New York. John Wiley & Sons, Inc., 1965.

Kish, L., Namboodiri, N. K., and Pillai, P. K.: The ratio bias in surveys. *J.Am.Statist.A.* 56(300): 863-876, Dec. 1962.

Lauh, E. and Williams, W. H.: Some small sample results for the variance of a ratio. *Proceedings of the Social Statistics Section of the American Statistical Association,* 1963. pp. 273-283.

National Center for Health Statistics: Cycle I of the Health Examination Survey, sample and response, United States, 1960-62. *Vital and Health Statistics*. PHS Pub. No. 1000-Series 11-No. 1. Public Health Service. Washington. U.S. Government Printing Office, Apr. 1964.

National Center for Health Statistics: Replication, an approach to the analysis of data from complex surveys. *Vital and Health Statistics*. PHS Pub. No. 1000-Series 2-No. 14. Public Health Service. Washington. U.S. Government Printing Office, Apr. 1966.

Owen, D. B., and Steck, G. P.: Moments of order statistics from the equicorrelated multivariate normal distribution. *Ann.math.Statist.* 33: 1286-1291, 1962.

Plackett, R. L.: and Burman, J. P.: The design of optimum multifactorial experiments. *Biometrika* 33: 305-325, 1943-46.

Quenouille, M. H.: Notes on bias in estimation. *Biometrika* 43: 353-360, 1956.

Rao, J. N. K.: A note on estimation of ratios by Quenouille's method. *Biometrika* 52: 647-649, 1965.

Rao, J. N. K.: and Webster, J. T.: On Two Methods of Bias Reduction in the Estimation of Ratios. Unpublished paper, 1965.

Teichroew, D.: Tables of expected values of order statistics for samples of size twenty and less from the normal distribution. *Ann.math.Statist.* 27: 410-426, 1956.

Tin, M.: Comparison of some ratio estimators. *J.Am.Statist.A.* 60(309): 294-307, 1965.

U.S. Bureau of the Census: *The Current Population Survey, A Report on Methodology*. Technical Paper No. 7. Washington. U.S. Government Printing Office, 1963.

Walsh, J. E.: Concerning the effect of intraclass correlation on certain significance tests. *Ann.math.Statist.* 18(1): 88-96, 1947.

Wilks, S. S.: *Mathematical Statistics*. New York. John Wiley & Sons, Inc., 1962.

——— o o o ———

DETAILED TABLES

Table 1. Estimates of mean body characteristics for population of U.S. adult males based on Health Examination Survey data

Variable (1)	$\hat{R}$ (2)	$\bar{r}$ (3)	$\bar{r}'$ (4)	$\bar{r}^*$ (5)	$(\bar{r}^* - \hat{R})$ (6)	$(\bar{r}^* - \hat{R})/\sqrt{\hat{V}(\hat{R})}$ (7)
1	39.76991	39.76854	39.76836	39.76845	-.00146	-.020
2	30.82899	30.82636	30.82650	30.82643	-.00256	-.025
3	99.57059	99.56900	99.56929	99.56914	-.00145	-.007
4	89.00548	89.00882	89.00932	89.00907	.00359	.016
5	1.26982	1.26907	1.26896	1.26902	-.00080	-.026
6	1.50193	1.50136	1.50143	1.50140	-.00053	-.028
7	86.63670	86.63346	86.63346	86.63346	-.00324	-.026
8	90.60693	90.60443	90.60411	90.60427	-.00266	-.022
9	54.32630	54.32346	54.32339	54.32342	-.00288	-.028
10	44.02929	44.02836	44.02843	44.02840	-.00089	-.013
11	14.51358	14.51357	14.51361	14.51359	.00001	.000
12	59.22556	59.22364	59.22371	59.22368	-.00188	-.022
13	49.45498	49.45439	49.45421	49.45430	-.00068	-.007
14	35.62331	---	---	---	---	---
15	42.22215	42.22332	42.22350	42.22341	.00126	.012
16	24.26486	24.26357	24.26364	24.26360	-.00126	-.012

Column 1 Variable

1. Biacromial diameter
2. Right arm girth
3. Chest girth
4. Waist girth
5. Right arm skinfold
6. Infrascapular skinfold
7. Sitting height (normal)
8. Sitting height (erect)
9. Knee height
10. Popliteal height
11. Thigh clearance height
12. Buttock-knee length
13. Buttock-popliteal length
14. Seat breadth
15. Elbow-elbow breadth
16. Elbow rest height

Column 2 $\hat{R}$ is the combined ratio estimate based on the entire sample, including adjustments for nonresponse and poststratification.

Column 3 The estimate $\hat{R}$ is obtained for each member of a set of 28 balanced half samples. These estimates are denoted by r_i, $i = 1, 2, \ldots, 28$, and $\bar{r}$ is their average.

Column 4 If r_i is the estimate from a half sample, then r_i' is the corresponding estimate made from the complementary half sample. $\bar{r}'$ is the average of these 28 complementary estimates.

Column 5 $\bar{r}^*$ is obtained by averaging columns 3 and 4.

Column 6 Column 5-column 2.

Column 7 The difference $(\bar{r}^* - \hat{R})$ expressed as a fraction of the estimated standard error of $\hat{R}$. $\hat{V}(\hat{R})$ is computed as $(1/4)\sum_{i=1}^{28}(r_i - r_i')^2/28$.

The actual values of the estimate of variance used in the computations are given in column 7 of table 3. In the text discussion, these estimates are denoted by $V_{CBHS}(\bar{r}^*)$.

Table 2. Estimates of regression coefficients and the multiple correlation coefficient for population of U.S. adult males based on Health Examination Survey data

Dependent variable (1)	Independent variable (2)	$\hat{\beta}$ (3)	$\bar{b}$ (4)	$\bar{b}'$ (5)	$\bar{b}^*$ (6)	$(\bar{b}^* - \hat{\beta})$ (7)	$(7)/\sqrt{\hat{V}(\hat{\beta})}$ (8)
1	Age-----------	-.02551	-.02555	-.02555	-.02555	-.00004	-.016
	Height--------	.05260	.05228	.05227	.05228	-.00032	-.056
	Weight--------	.02999	.02999	.02998	.02998	-.00001	-.011
	Multiple R----	.55033	.55024	.55028	.55026	-.00007	-.006
2	Age-----------	-.02724	-.02733	-.02734	-.02734	-.00010	-.038
	Height--------	-.12620	-.12689	-.12688	-.12688	-.00068	-.127
	Weight--------	.11140	.11142	.11142	.11142	.00002	.016
	Multiple R----	.88439	.88481	.88479	.88480	.00041	.103
3	Age-----------	.03889	.03885	.03884	.03884	-.00005	-.007
	Height--------	-.20126	-.20248	-.20247	-.20248	-.00122	-.075
	Weight--------	.28900	.28908	.28908	.28908	.00008	.026
	Multiple R----	.90431	.90470	.90468	.90469	.00038	.092
4	Age-----------	.20804	.20781	.20779	.20780	-.00024	-.028
	Height--------	-.38187	-.38291	-.38293	-.38292	-.00105	-.068
	Weight--------	.38092	.38117	.38119	.38118	.00026	.075
	Multiple R----	.91499	.91529	.91529	.91529	.00030	.069
5	Age-----------	-.00279	-.00275	-.00276	-.00276	.00003	.037
	Height--------	-.02281	-.02283	-.02284	-.00284	-.00003	-.011
	Weight--------	.01766	.01764	.01763	.01764	-.00002	-.037
	Multiple R----	.59132	.59274	.59282	.59278	.00146	.065
6	Age-----------	.00023	.00025	.00024	.00024	.00001	.011
	Height--------	-.03641	-.03650	-.03650	-.03650	-.00009	-.042
	Weight--------	.02313	.02311	.02311	.02311	-.00002	-.035
	Multiple R----	.77307	.77313	.77312	.77312	.00005	.015
7	Age-----------	.00037	.00010	.00009	.00010	-.00027	-.084
	Height--------	.34698	.34455	.34455	.34455	-.00243	-.232
	Weight--------	.02407	.02423	.02422	.02422	.00015	.058
	Multiple R----	.73569	.73549	.73544	.73546	-.00023	-.019
8	Age-----------	-.01787	-.01824	-.01824	-.01824	-.00037	-.101
	Height--------	.37403	.37154	.37152	.37153	-.00250	-.289
	Weight--------	.01639	.01654	.01653	.01654	.00015	.067
	Multiple R----	.78113	.78183	.78180	.78182	.00069	.079

Column 1 Dependent variable
1. Biacromial diameter
2. Right arm girth
3. Chest girth
4. Waist girth
5. Right arm skinfold
6. Infrascapular skinfold
7. Sitting height (normal)
8. Sitting height (erect)

Column 2 Independent variables as identified. Note that the last entry in each set is the multiple correlation coefficient.

Column 3 These are the regression coefficients and the multiple correlation coefficient as estimated from the entire sample, including adjustments for nonresponse and poststratification.

Column 4 The estimate $\hat{\beta}$ is obtained for each member of a set of 28 balanced half samples. These estimates are denoted by b_i, $i = 1, 2, \ldots, 28$, and $\bar{b}$ is their average.

Column 5 If b_i is the estimate from a half sample, then b_i' is the corresponding estimate made from the complementary half sample. $\bar{b}'$ is the average of these 28 complementary estimates.

Column 6 Average of columns 4 and 5.

Column 7 Column 6-column 3.

Column 8 The difference $(\bar{b}^* - \hat{\beta})$ expressed as a fraction of the estimated standard error of $\hat{\beta}$. $\hat{V}(\hat{\beta})$ is computed as $(1/4)\sum_{i=1}^{28}(b_i - b_i')^2/28$.

The actual values of the estimate of variance used in the computations are given in column 8 of table 4. In the text discussion, these estimates are denoted by $\hat{V}_{CBHS}(\bar{b}^*)$.

Table 3. Estimated sampling variability of mean body characteristics for population of U.S. adult males based on Health Examination Survey data, and estimated intraclass correlation coefficients among pseudo-replicate samples

Variable	$\hat{V}_{RAN}(\hat{R})$	$\hat{V}_{BHS}(\hat{R})$	$\hat{V}_{BHS}(\bar{r})$	$\hat{V}'_{BHS}(\hat{R})$	$\hat{V}_{BHS}(\bar{r}')$	$\hat{V}_{CBHS}(\bar{r}^*)$	(7)/(2)	Intraclass correlations		
								$\hat{\rho}$	$\hat{\rho}'$	$\hat{\rho}^*$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1-----------	.00143	.00551	.00551	.00564	.00564	.00556	3.888	.48651	.47392	.48021
2-----------	.00342	.01034	.01033	.01086	.01085	.01057	3.090	.49290	.46736	.48013
3-----------	.02265	.03740	.03740	.03943	.03943	.03833	1.692	.49413	.46661	.48037
4-----------	.04148	.04623	.04622	.05007	.05006	.04802	1.158	.50094	.45950	.48022
5-----------	.00019	.00096	.00096	.00100	.00100	.00098	5.158	.49490	.46939	.48214
6-----------	.00019	.00039	.00039	.00038	.00038	.00038	2.000	.47368	.48684	.48026
7-----------	.00449	.01591	.01590	.01599	.01598	.01592	3.546	.48210	.47959	.48084
8-----------	.00444	.01480	.01479	.01528	.01527	.01501	3.380	.48901	.47235	.48068
9-----------	.00282	.01012	.01011	.01083	.01082	.01045	3.706	.49809	.46316	.48062
10-----------	.00241	.00509	.00509	.00503	.00503	.00505	2.096	.47723	.48317	.48020
11-----------	.00096	.00556	.00556	.00546	.00546	.00550	5.729	.47545	.48545	.48045
12-----------	.00281	.00713	.00713	.00727	.00727	.00718	2.555	.48538	.47493	.48015
13-----------	.00310	.01061	.01061	.01111	.01111	.01084	3.496	.49262	.46863	.48062
14-----------	.00243	.01126	---	.01030	---	.01075	4.424	---	---	---
15-----------	.00707	.01109	.01109	.01156	.01156	.01129	1.596	.49070	.46900	.47985
16-----------	.00292	.01187	.01187	.01184	.01184	.01184	4.054	.48015	.48142	.48078

Column 1 Variable

1. Biacromial diameter
2. Right arm girth
3. Chest girth
4. Waist girth
5. Right arm skinfold
6. Infrascapular skinfold
7. Sitting height (normal)
8. Sitting height (erect)
9. Knee height
10. Popliteal height
11. Thigh clearance height
12. Buttock-knee length
13. Buttock-popliteal length
14. Seat breadth
15. Elbow-elbow breadth
16. Elbow rest height

Column 2 $\hat{V}_{RAN}(\hat{R})$ is computed as if the sample had been drawn as a simple random sample of approximately 3,000 adult males from the total U.S. population. The effects of clustering, stratification, and estimation have been ignored.

Column 3 $\hat{V}_{BHS}(\hat{R})$ is the variance estimate that is ordinarily computed from a set of 28 balanced half samples. If the individual half-sample estimates are denoted by r_i, $i = 1, 2, \ldots, 28$ then

$$\hat{V}_{BHS}(\hat{R}) = \sum_{i=1}^{28} (r_i - \hat{R})^2/28$$

Column 4 $\hat{V}_{BHS}(\bar{r})$ is similar to $\hat{V}_{BHS}(\hat{R})$, except that deviations are taken about the mean of the r_i's, namely $\bar{r}$.

$$\hat{V}_{BHS}(r) = \sum_{i=1}^{28} (r_i - \bar{r})^2/28$$

Column 5 $\hat{V}'_{BHS}(\hat{R})$ corresponds to $\hat{V}_{BHS}(\hat{R})$, except that it is computed from the complementary half-sample estimates. Denoting these by r'_i, $i = 1, 2, \ldots, 28$

$$\hat{V}'_{BHS}(\hat{R}) = \sum_{i=1}^{28} (r'_i - \hat{R})^2/28$$

Column 6

$$\hat{V}_{BHS}(r') = \sum_{i=1}^{28} (r'_i - \bar{r}')^2/28$$

Column 7 $V_{CBHS}(\bar{r}^*)$ is an estimate of variance based on the comparison of a half-sample estimate r_i with its complementary estimate r'_i. This estimate is not available in the ordinary application of balanced half-sample replication for variance estimation.

$$\hat{V}_{CBHS}(\bar{r}^*) = (1/4) \sum_{i=1}^{28} (r_i - r'_i)^2/28$$

Column 8 The ratio of column 7 to column 2 estimates the effect on sampling variability of sample design and estimation in comparison with simple random sampling.

Column 9 $\hat{\rho}$ is an estimate of the intraclass correlation among the r_i's. It was computed as

$$\hat{\rho} = \frac{2\,\hat{V}_{CBHS}(\bar{r}^*) - (28/27)\,\hat{V}_{BHS}(\bar{r})}{2\,\hat{V}_{CBHS}(\bar{r}^*)}$$

Column 10 $\hat{\rho}'$ is an estimate of the intraclass correlation among the r'_i's. It was computed as

$$\hat{\rho}' = \frac{2\,\hat{V}_{CBHS}(\bar{r}^*) - (28/27)\,\hat{V}_{BHS}(\bar{r}')}{2\,V_{CBHS}(\bar{r}^*)}$$

Column 11 $\hat{\rho}^*$ is the average of $\hat{\rho}$ and $\hat{\rho}'$.

Table 4. Estimated sampling variability of regression coefficients and the multiple correlation coefficient for population of U.S. adult males based on Health Examination Survey data,[1] and estimated intraclass correlation coefficients among pseudoreplicate samples

Dependent variable	Independent variable	$\hat{V}_{RAN}(\hat{\beta})$	$\hat{V}_{BHS}(\hat{\beta})$	$\hat{V}_{BHS}(\bar{b})$	$\hat{V}'_{BHS}(\hat{\beta})$	$\hat{V}_{BHS}(\bar{b}')$	$\hat{V}_{CBHS}(\bar{b}^{*})$	(8)/(3)	Intraclass correlations		
									$\hat{\beta}$	$\hat{\beta}'$	$\hat{\beta}^{*}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
1	Age--------	.43432	.59823	.59807	.66892	.66876	.63165	1.454	.5091	.4510	.4800
	Height-----	.26396†	.32710†	.32608†	.33045†	.32936†	.32603†	1.235	.4814	.4762	.4788
	Weight-----	.15524	.08954	.08954	.08896	.08895	.08854	.570	.4756	.4791	.4774
	Multiple R-	.15748††	.16116††	.16115††	.15374††	.15374††	.15669††	.995	.4667	.4913	.4790
2	Age--------	.32395	.65564	.65483	.63545	.63445	.64223	1.982	.4713	.4878	.4796
	Height-----	.19689†	.31062†	.30586†	.28883†	.28421†	.29365†	1.491	.4599	.4982	.4790
	Weight-----	.11579	.17278	.17274	.15962	.15958	.16537	1.428	.4584	.4997	.4790
	Multiple R-	.01538††	.01514††	.01496††	.01695††	.01679††	.01575††	1.024	.5075	.4473	.4774
3	Age--------	.17959†	.39902†	.39900†	.35124†	.35122†	.37327†	2.078	.4458	.5121	.4790
	Height-----	.10915††	.28912††	.28763††	.24561††	.24415††	.26411††	2.420	.4353	.5207	.4780
	Weight-----	.64194	.93129	.93065	.94957	.94893	.93292	1.453	.4828	.4726	.4777
	Multiple R-	.01076††	.01762††	.01747††	.01702††	.01688††	.01710††	1.589	.4703	.4882	.4792
4	Age--------	.29386†	.71384†	.71331†	.73237†	.73174†	.72068†	2.452	.4868	.4735	.4802
	Height-----	.17860††	.25191††	.25083††	.23508††	.23396††	.24096††	1.349	.4603	.4966	.4784
	Weight-----	.10504†	.11596†	.11534†	.12938†	.12865†	.12097†	1.152	.5056	.4486	.4771
	Multiple R-	.00859††	.01870††	.01861††	.01994††	.01985††	.01914††	2.228	.4959	.4623	.4791
5	Age--------	.05323	.09286	.09270	.08917	.08908	.09038	1.698	.4682	.4890	.4786
	Height-----	.03235†	.05407†	.05407†	.05597†	.05596†	.05461†	1.688	.4866	.4687	.4776
	Weight-----	.01903	.04491	.04487	.04726	.04717	.04576	2.405	.4916	.4655	.4786
	Multiple R-	.13705††	.52915††	.52713††	.49128††	.48903††	.50365††	3.675	.4573	.4966	.4770
6	Age--------	.03276	.04977	.04973	.05018	.05017	.04982	1.521	.4824	.4779	.4802
	Height-----	.01991†	.04614†	.04606†	.04438†	.04430†	.04503†	2.262	.4696	.4899	.4798
	Weight-----	.01171	.03260	.03256	.03196	.03192	.03216	2.746	.4751	.4854	.4802
	Multiple R-	.05246††	.01384††	.01384††	.01416††	.01416††	.01390††	2.650	.4837	.4718	.4778
7	Age--------	.08967†	.11505†	.11432†	.10333†	.01255†	.10774†	1.202	.4498	.5065	.4782
	Height-----	.05450††	.11689††	.11099††	.11577††	.10987††	.10991††	2.017	.4764	.4817	.4790
	Weight-----	.32052	.73399	.73143	.71576	.71351	.71920	2.244	.4727	.4856	.4792
	Multiple R-	.06320††	.12980††	.12976††	.13933††	.13927††	.13375††	2.116	.4970	.4601	.4786
8	Age--------	.07538†	.13850†	.13713†	.13500†	.13486†	.13500†	1.791	.4733	.4820	.4776
	Height-----	.45812†	.82611†	.76411†	.80067†	.73767†	.74612†	1.629	.4690	.4874	.4782
	Weight-----	.26943	.48075	.47850	.47475	.47279	.47351	1.757	.4760	.4823	.4792
	Multiple R-	.04924††	.07338††	.07289††	.07896††	.07851††	.07524††	1.528	.4977	.4590	.4784

See footnotes on next page.

[1]In columns 3-8, the absence of a dagger means that the entry is to be multiplied by 10^{-5}, the presence of a single dagger signifies multiplication by 10^{-4}, and the presence of a double dagger signifies multiplication by 10^{-3}.

Column 1 Dependent variable
1. Biacromial diameter
2. Right arm girth
3. Chest girth
4. Waist girth
5. Right arm skinfold
6. Infrascapular skinfold
7. Sitting height (normal)
8. Sitting height (erect)

Column 2 Independent variables as identified. Note that the last entry in each set is the multiple correlation coefficient.

Column 3 $\hat{V}_{RAN}(\hat{\beta})$ is computed as if the sample had been drawn as a simple random sample of approximately 3,000 adult males from the total U.S. population. The effects of clustering, stratification, and estimation have been ignored.

Column 4 $\hat{V}_{BHS}(\hat{\beta})$ is the variance estimate that is ordinarily computed from a set of 28 balanced half samples. If the individual half-sample estimates are denoted by b_i, $i = 1, 2, \ldots, 28$ then

$$\hat{V}_{BHS}(\hat{\beta}) = \sum_{i=1}^{28} (b_i - \hat{\beta})^2/28$$

Column 5 $\hat{V}_{BHS}(\bar{b})$ is similar to $\hat{V}_{BHS}(\hat{\beta})$, except that deviations are taken about the mean of the b_i's, namely $\bar{b}$.

$$\hat{V}_{BHS}(b) = \sum_{i=1}^{28} (b_i - \bar{b})^2/28$$

Column 6 $\hat{V}'_{BHS}(\hat{\beta})$ corresponds to $\hat{V}_{BHS}(\hat{\beta})$, except that it is computed from the complementary half-sample estimates. Denoting these by b'_i, $i = 1, 2, \ldots, 28$

$$\hat{V}'_{BHS}(\hat{\beta}) = \sum_{i=1}^{28} (b'_i - \hat{\beta})^2/28$$

Column 7

$$\hat{V}_{BHS}(b') = \sum_{i=1}^{28} (b'_i - \bar{b}')^2/28$$

Column 8 $V_{CBHS}(\bar{b}^*)$ is an estimate of variance based on the comparison of a half-sample estimate b_i with its complementary estimate b'_i. This estimate is not available in the ordinary application of balanced half-sample replication for variance estimation.

$$\hat{V}_{CBHS}(\bar{b}^*) = (1/4) \sum_{i=1}^{28} (b_i - \bar{b}'_i)^2/28$$

Column 9 The ratio of column 8 to column 3 estimates the effect on sampling variability of sample design and estimation in comparison with simple random sampling.

Column 10 $\hat{\rho}$ is an estimate of the intraclass correlation among the b_i's. It was computed as

$$\hat{\rho} = \frac{2\,\hat{V}_{CBHS}(\bar{b}^*) - (28/27)\,\hat{V}_{BHS}(\bar{b})}{2\,\hat{V}_{CBHS}(\bar{b}^*)}$$

Column 11 $\hat{\rho}'$ is an estimate of the intraclass correlation among the b'_i's. It was computed as

$$\hat{\rho}' = \frac{2\,\hat{V}_{CBHS}(\bar{b}^*) - (28/27)\,\hat{V}_{BHS}(\bar{b}')}{2\,V_{CBHS}(\bar{b}^*)}$$

Column 12 $\hat{\rho}^*$ is the average of $\hat{\rho}$ and $\hat{\rho}'$.

———o o o———

☆ U. S. GOVERNMENT PRINTING OFFICE : 1973 515-214/68

VITAL AND HEALTH STATISTICS PUBLICATION SERIES

Originally Public Health Service Publication No. 1000

Series 1. Programs and collection procedures.—Reports which describe the general programs of the National Center for Health Statistics and its offices and divisions, data collection methods used, definitions, and other material necessary for understanding the data.

Series 2. Data evaluation and methods research.—Studies of new statistical methodology including: experimental tests of new survey methods, studies of vital statistics collection methods, new analytical techniques, objective evaluations of reliability of collected data, contributions to statistical theory.

Series 3. Analytical studies.—Reports presenting analytical or interpretive studies based on vital and health statistics, carrying the analysis further than the expository types of reports in the other series.

Series 4. Documents and committee reports.—Final reports of major committees concerned with vital and health statistics, and documents such as recommended model vital registration laws and revised birth and death certificates.

Series 10. Data from the Health Interview Survey.—Statistics on illness, accidental injuries, disability, use of hospital, medical, dental, and other services, and other health-related topics, based on data collected in a continuing national household interview survey.

Series 11. Data from the Health Examination Survey.—Data from direct examination, testing, and measurement of national samples of the civilian, noninstitutional population provide the basis for two types of reports: (1) estimates of the medically defined prevalence of specific diseases in the United States and the distributions of the population with respect to physical, physiological, and psychological characteristics; and (2) analysis of relationships among the various measurements without reference to an explicit finite universe of persons.

Series 12. Data from the Institutional Population Surveys.—Statistics relating to the health characteristics of persons in institutions, and their medical, nursing, and personal care received, based on national samples of establishments providing these services and samples of the residents or patients.

Series 13. Data from the Hospital Discharge Survey.—Statistics relating to discharged patients in short-stay hospitals, based on a sample of patient records in a national sample of hospitals.

Series 14. Data on health resources: manpower and facilities.—Statistics on the numbers, geographic distribution, and characteristics of health resources including physicians, dentists, nurses, other health occupations, hospitals, nursing homes, and outpatient facilities.

Series 20. Data on mortality.—Various statistics on mortality other than as included in regular annual or monthly reports—special analyses by cause of death, age, and other demographic variables, also geographic and time series analyses.

Series 21. Data on natality, marriage, and divorce.—Various statistics on natality, marriage, and divorce other than as included in regular annual or monthly reports—special analyses by demographic variables, also geographic and time series analyses, studies of fertility.

Series 22. Data from the National Natality and Mortality Surveys.—Statistics on characteristics of births and deaths not available from the vital records, based on sample surveys stemming from these records, including such topics as mortality by socioeconomic class, hospital experience in the last year of life, medical care during pregnancy, health insurance coverage, etc.

For a list of titles of reports published in these series, write to:

Office of Information
National Center for Health Statistics
Public Health Service, HSMHA
Rockville, Md. 20852