Attention is called to the importance of intraclass correlation in various kinds of survey studies where subsampling or natural clusters may be present. The problem is analyzed and some examples are taken from the morbidity data of a study of housing environment and family life.

INTRAFAMILY CORRELATION AND ITS SIGNIFICANCE IN THE INTERPRETATION OF SAMPLE SURVEYS

Carl E. Hopkins, Ph.D., M.P.H., F.A.P.H.A.; Rosabelle P. Walkley; Daniel M. Wilner, Ph.D., F.A.P.H.A.; and Theodore T. Gold

IN SURVEYS involving natural clusters, such as families, school classes, sororities, villages, health districts, and the like, the problem of intraclass correlation is frequently ignored, with the likely consequence of overestimating the precision and statistical significance of results obtained.

The purpose of this paper is to review briefly some circumstances in which correlation within clusters has been encountered and the various statistical methods that have been used to deal with the problem. Special reference will be made to a study of the housing environment and family life,*[1-4] a longitudinal assessment of the effects of housing on morbidity and mental health. In this study, with which the authors are associated, the problem of intraclass correlation arose in the statistical analysis of morbidity data and made a substantial difference in the interpretation of the results. The outcome of this analysis is reported in order to call attention to the importance of considering this useful, although somewhat burdensome, statistical refinement.

Intraclass correlation or correlation within clusters is commonly taken into account in designed sample surveys. Thus in a housing quality survey of a city, a two-stage cluster sampling plan might be adopted. First a random sample of blocks might be selected, followed by selection of a random sample of houses from each sample block. In this context the block is called the "primary sampling unit," the individual house is called a "listing unit" or "subsample unit." The houses sampled in a block comprise a "cluster."

Some correlation (similarity) is likely among the members of a cluster: that is, the houses in the same block. If houses on a block were quite homogeneous, this would be reflected in a high positive within-cluster correlation coefficient. Or, in analysis of variance terms, the variance within the clusters would be small compared with that between the clusters. In such a circumstance, the correct sampling strategy would be to take a large sample of blocks and a very small sub-

* The study was conducted during the period 1954-1960 at the Johns Hopkins University School of Hygiene and Public Health where Dr. Wilner and Mrs. Walkley held appointments. Additional analyses of the data are currently being made by the authors at the University of California School of Public Health, Los Angeles. The study has been supported by a research grant from the National Institutes of Health, Public Health Service, U. S. Department of Health, Education, and Welfare.

sample of houses within blocks, since in the homogeneous block one house would give nearly all the information there is in that block.

The within-cluster correlation might, on the other hand, be low positive or near zero. This would be the case if there were great heterogeneity within blocks, so that observation of a single house would give very little information about other houses in the block. In this circumstance, one would wish to obtain a fairly large sample of houses within a block, with fewer blocks sampled.

Negative correlation within a block (or cluster) might arise in some unusual circumstances. Such would be the case in a survey of families in which the sex of the adult members was the variable of interest. If the sex of the respondent is male, the sex of the other adult family member is almost certain to be not male. Hence there would be a high negative within-cluster correlation.

Within-cluster correlation when positive, as is usually the case, has the effect of deflating the variance of the mean per individual listing unit as estimated from the sample variance. To visualize this, we might think of a sample of 16 houses taken by simple random sampling from a city. The estimate of the mean of variable Y, say number of windows per house, will be $\Sigma Y/16$ and the variance of the mean in the population is estimated by the variance observed in the sample divided by the sample size, 16, since the 16 observations are assumed statistically independent. If, however, these 16 houses are all in one block selected as a primary sampling unit, they could, because of within-block correlation, be much more similar to one another than 16 houses taken at random from the entire city. The result would be an underestimation of the variance of the mean. If the correlation were high, for instance, these 16 observations would be nearly equivalent to only one independent observation, since all of the 16 would give nearly the same information about houses in the city generally. Since the effective sample size in the block could then be nearer to one than 16, the true variance of the mean in the population might be nearly 16 times as great as that estimated by the sample variance. The exact formula for obtaining an unbiased estimate of the variance of the mean is:

$$\text{Var}\,(\bar{y}) = \frac{S^2}{NM}\,[1+(M-1)R]\ , \qquad (1)$$

where $\bar{y}$ is the grand mean of all the individual observations,

- N is the number of clusters, all of size M,
- M is the number of individual elements in each cluster,
- S^2 is the variance of the individual observations treated as independent observations in a simple random sample, and
- R is the coefficient of intraclass correlation. (Formula (1) omits the finite population correction factor, which is not of interest in this connection.)

The standard error of the mean would, of course, be the square root of (1). It is evident from this relation that use of the cluster instead of the element as the primary sampling unit multiplies the variance of the mean by the factor $[1+(M-1)R]$. If R is positive, the variance is increased; and if R is negative, it is decreased. With zero R, there is no change. The amount of change depends also on the cluster size, M. Large clusters have a greater effect than small. This formula is derived and thoroughly discussed in Cochran,[6] section 9.7.

When sample surveys are planned for a single purpose, such as measurement of housing quality, interest focuses on planning an optimum combination of M and N, cluster size and number of clusters—optimum in the sense of minimizing the variance of the mean to be estimated. Most of the discussion of intraclass correlation in references on sample survey methods[5-8] centers on such considerations.

There are, nevertheless, numerous

studies and surveys, especially in the public health field, in which the sampling, while primarily simple random sampling, has multiple purposes and includes, either inadvertently or unavoidably, data from subclasses or elements of the primary sampling units. Thus in the typical household diet or consumption survey, natural clusters of household members are present, and estimates of characteristics per household member will be subject to the influence of within-household correlation. Still another component of within-household correlation is added if a single respondent gives information for all household members. The single informant could be adding some adventitious correlation in reporting data for other household members.

The Kinsey studies of sexual behavior of human beings, for example, were beset with very considerable natural clustering, as a result of the method of recruiting volunteer subjects in pre-existing groups, such as college fraternities, women's clubs, and the like. Failure to take into account probable within-cluster correlation in most items makes it most difficult to assess the precision of the averages obtained and the significance of differences observed.[9,10]

In the study of the housing environment and family life[1-4] the problem arose because of the multiple levels of data collection. A sample of 300 families in substandard slum housing moved into new public housing of good quality. These families were matched one for one to 300 slum families who were eligible but, because of limitations of the public housing accommodations, did not get into the new housing. The "test" and "control" families were measured on a large number of variables before and for several years after the move. Some of the variables such as housing quality, social adjustment, style of life, and so forth, were true family observations and for them the family was the ultimate sampling unit as well as the primary. But for some other variables such as morbidity, the ultimate sampling element was the individual family member. Accordingly, estimates of mean number of episodes of sickness in test and control families had to take into account the possible within-family correlation. Moreover, since the clusters arose "naturally" rather than by sampling design, they were of different sizes. Thus M was also a random variable, making computation of within-cluster correlations quite complex.

A FORTRAN program was written for the IBM 7090 computer which:

1. sorts the individual persons into family clusters and forms a frequency distribution of the number of families of each size;
2. computes the analysis of variance for each family size, arriving at the coefficient of intrafamily correlation via the Fisher-Haggard formula[11];
3. computes the grand mean of the variable across all families in the sample;
4. computes the ordinary variance of this mean (treating each individual family member as an independent random observation); and
5. computes the variance of the grand mean corrected for the intrafamily correlations observed in the sample.

The corrected variance is computed from a formula derived by J. J. Gart, for computing convenience, from Sukhatme[5,p.267]:

$$\text{Var}(\bar{y}) = \frac{(\Sigma M_i)^2 \Sigma Y_i^2 - 2\Sigma M_i \Sigma Y_i \Sigma M_i Y_i + (\Sigma Y_i)^2 \Sigma M_i^2}{(\Sigma M_i)^3 (\Sigma M_i - 1)}, \qquad (2)$$

which weights the means of the unequal clusters by the square of their size.

The Gart-Sukhatme formula does not explicitly generate the intraclass correlation coefficients. These are computed from relations given by Fisher and formulated by Haggard[11] for the case of clusters of equal size:

$$R = \frac{MSB - MSW}{MSB + (M-1)MSW}, \qquad (3)$$

where MSB is the mean square between the classes or clusters, MSW is the mean square within clusters, and M is the cluster size (equal for all clusters), in the usual Analysis of Variance format.

For the case of unequal cluster sizes, the over-all R can be approximated from the same formula by replacing M by $\overline{M}$, a mean cluster size calculated from

$$\overline{M} = \frac{1}{N-1}\left(\Sigma M_i - \frac{\Sigma M_i^2}{\Sigma M_i}\right). \qquad (4)$$

How good this plausible approximation [taken from Snedecor[12]] is in operation has not been investigated. In our computations it sometimes resulted in good agreement with the Gart-Sukhatme result for Var $(\bar{y})$, sometimes not. Hence it seems advisable to use the Gart-Sukhatme formula for computing the variance of $\bar{y}$ and the Fisher-Haggard analysis of variance for computing the individual intraclass correlations.

Table 1 shows a specimen page of computer output. This computation was for males under age 20 in the Test sample. A total of 412 individuals were in this sample, which included 108 families with only one male member under 20, 67 families with two such members, 39 clusters of three members, and so forth. The intraclass correlations for each size of within-family cluster are shown. At bottom are given the grand mean and its variance (0.0117), taking into account the increase due to the within-cluster correlation. Also shown, for comparison, is the variance of the grand mean (0.0072) as it would appear if computed without accounting for the within-cluster correlation.

In Table 2 the correlation coefficients are shown for test and control families before and after the housing move, for two main morbidity variables: total episodes of illness (during a 12-month period) and total days of disability. These are for the subclass of family members under 20 years old, separated by sex. In the "before" period, the test families with a single male under 20 numbered 108, and of course the within-family correlation was zero for these solitaires. There were 67 families containing clusters of two males, and they generated a correlation coefficient of 0.504. The 39 clusters of size 3 had R=0.526, and the eight size 4 clusters had R=0.612. Throughout the table similarly substantial positive coefficients are seen, indicating that a considerable shrinkage of effective sample size and consequently of precision can be anticipated. These coefficients, generally in the vicinity of 0.50, indicate that about $(0.50)^2$ or 25 per cent of the variance of a random family member's morbidity is predictable from the morbidity of his age-sex similars in his family.

These rather high correlations are not particularly astonishing in view of the fact that a large proportion of sickness in the younger ages is due to contagious infections, which add some "direct cause" correlation to the indirect correlation that would be normally expected within closely associated groups. In older age classes, e.g., 20 to 60, males, the within-family correlation had much less effect, since there were so few families containing more than one male 20 to 60 years old.

Confidence intervals can be formed for these estimates of R, using tables supplied by Haggard.[11] For example, the 90 per cent confidence interval for the R=0.504 of Table 2, I., cluster size M=2, would be 0.36 to 0.66. For the less numerous family clusters of size 4, the 90 per cent confidence limits around R=0.612 would be 0.33 and 0.81. It is seen that even with the smaller frequencies of the larger clusters, the lower limit of the estimate of R is definitely not negligible.

Sukhatme[5,p.247] mentions that the size of R usually diminishes as cluster size increases. This may be true for randomly formed (as in census surveys) clusters, but for the family, as an existing natural cluster, this attenuation does

Table 1—Specimen Computer Output

	1 No.	1 INDS.	MSQ	Y-Total	MY	YSQ	TOTSS	BCSS	WSS	BCMS	WMS	COR
1	108.0	108.0	108.0	254.0	254.0	980.0	0.0	0.0	0.0	0.0	0.0	0.0
2	67.0	134.0	268.0	271.0	542.0	1607.0	340.933	255.433	85.500	3.870	1.276	0.5041
3	39.0	117.0	351.0	212.0	636.0	1844.0	339.863	230.530	109.333	6.067	1.402	0.5259
4	8.0	32.0	128.0	67.0	268.0	781.0	80.719	54.969	25.750	7.853	1.073	0.6124
5	3.0	15.0	75.0	28.0	140.0	290.0	31.733	5.733	26.000	2.867	2.167	0.0607
6	1.0	6.0	36.0	23.0	138.0	529.0	10.833	0.0	10.833	0.0	2.167	−0.2000
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	−0.0	0.0	−0.0
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	−0.0	0.0	−0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	−0.0	0.0	−0.0
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	−0.0	0.0	−0.0
11	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	−0.0	0.0	−0.0
12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	−0.0	0.0	−0.0
13	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	−0.0	0.0	−0.0
14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	−0.0	0.0	−0.0
15	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	−0.0	0.0	−0.0
	226.0	412.0	966.0	855.0	1978.0	6031.0						

Grand average 2.0752.
Variance of grand average—Uncorrected 0.0072.
Variance of grand average—Corrected 0.0117.

Table 2—Some Within-Family Correlation Coefficients Found in the Housing Study: Morbidity Variables

I. Total Episodes of Illness (Family members under 20 years old)

Within-Family Correlation Coefficient (n=frequency)

Period	Sex	Sample	Cluster Size 1	2	3	4	5
Before	M	Test	0 (108)	0.504 (67)	0.526 (39)	0.612 (8)	0.061 (3)
Before	M	Control	0 (85)	0.496 (70)	0.496 (49)	0.727 (14)	0.085 (3)
After	M	T	0 (108)	0.474 (67)	0.440 (39)	0.710 (8)	0.237 (3)
After	M	C	0 (85)	0.603 (70)	0.566 (49)	0.826 (14)	0.255 (3)
Before	F	T	0 (101)	0.452 (76)	0.298 (38)	0.338 (13)	0.129 (5)
Before	F	C	0 (95)	0.558 (60)	0.624 (32)	0.327 (18)	0.661 (7)
After	F	T	0 (101)	0.500 (76)	0.414 (38)	0.273 (13)	0.759 (5)
After	F	C	0 (95)	0.591 (60)	0.764 (32)	0.526 (18)	0.331 (7)

90% Confidence Limit Examples: for R=0.504 (line 1): 0.360 to 0.661; for R=0.612 (line 1): 0.335 to 0.808.

II. Total Days of Disability (Family members under 20 years old)

Before	M	Test	0 (108)	0.367 (67)	0.169 (39)	0.474 (8)	0.013 (3)
Before	M	Control	0 (85)	0.040 (70)	0.284 (49)	0.004 (14)	0.000 (3)
After	M	T	0 (108)	0.675 (67)	0.437 (39)	0.297 (8)	0.614 (3)
After	M	C	0 (85)	0.574 (70)	0.274 (49)	0.443 (14)	0.845 (3)
Before	F	T	0 (101)	0.249 (76)	0.042 (38)	0.760 (13)	0.187 (5)
Before	F	C	0 (95)	0.075 (60)	0.372 (32)	0.191 (18)	0.477 (7)
After	F	T	0 (101)	0.372 (76)	0.412 (38)	0.096 (13)	0.164 (5)
After	F	C	0 (95)	0.303 (60)	0.145 (32)	0.213 (18)	0.035 (7)

not appear to occur. The correlation coefficients observed are quite variable for the larger cluster sizes, because of small frequencies, but generally hold up in magnitude. An interesting exception is that in a few instances in which rather large clusters of six or seven individuals in the same family were observed, the correlation was negative, suggesting some confusion in our definition of "family." These turned out to be households shared by more than one family.

Table 3 shows the effects of these same within-family correlations on the variance of the estimated mean and on the t-test comparison of the test vs. control means. The means are, of course, fairly precise because of the large sample sizes, but their variances are markedly inflated (generally about doubled) by the within-family correlation. This has the consequence of shrinking the t-ratio for the test-control difference by about one-fourth, which is sufficient in two out of three instances to change the interpretation from "significant difference" to "no difference observed."

Discussion

Similar problems of correlation within naturally occurring clusters occur in many kinds of experiments and surveys. In a survey of blood types of Basques in Idaho,[13] a large sample of 165 persons was collected, who on their genealogies turned out to be so consanguineous that they really represented something like only 25 independent "natural" clusters. The correlation within these clusters was somewhat awkwardly taken into account by computing a coefficient of relationship[14] and then randomly drawing a single person from each cluster showing significant relationship.

Longitudinal studies of growth characteristically face a similar problem. Successive measurements of a growing child, for example, are likely to be correlated. In a sense, the child is the primary sampling unit and his M successive observations form a cluster of M elements. The variance of estimates made from such data has been treated under the headings of "serial correlation"[15] and "sampling on successive occasions."[6,7,16] Tanner[17] has offered an analytical model for the typical mixed longitudinal growth study in which different children are observed for different lengths of time, with replacement and overlapping. Some of the restrictions and limitations of the Tanner approach have been relieved by a more general model developed by Henderson, et al.[18] Terzaghi[19] has developed a FORTRAN routine that makes the Henderson model usable by committing the heavy arithmetic to the electronic computer.

Haggard in a recent monograph[11] shows the relationship of intraclass correlation to the analysis of variance, taking up where Fisher[20] left off. Numerous examples are shown of applications in psychology, education, etc., where such problems are encountered as the scores of a given pupil (a cluster) on each of several tests (elements).

Numerous other areas of biometric investigation encounter problems of intraclass correlation, notably genetic and plant and animal breeding experiments, and appropriate statistical methods have been developed for their handling.[21-26]

Summary

Attention is called to the importance of intraclass correlation in various types of survey studies in which subsampling or natural clusters may be present. The presence of any substantial positive correlation within clusters may have a large effect on the variance of the estimated mean per individual. Ignoring of the intraclass correlation can lead to erroneous conclusions. A FORTRAN program is now available for the heavy computations involved in estimating and correct-

Table 3—The Variance of the Grand Mean Corrected for Within-Family Correlation

I. Total Episodes of Illness (Family members under 20 years old)

Period	Sex	Sample	Grand Mean	Families	Persons	Variance of Mean		t for T-C Difference	
						Uncorrected	Corrected	Uncorrected	Corrected
Before	M	Test	2.075	226	412	0.007	0.012	1.43	1.06
Before	M	Control	1.911	222	449	0.006	0.012		
After	M	T	2.655	226	412	0.013	0.020	2.63*	1.84
After	M	C	3.094	222	449	0.015	0.037		
Before	F	T	2.074	233	444	0.006	0.010	1.38	1.08
Before	F	C	1.922	213	424	0.006	0.010		
After	F	T	2.750	233	444	0.012	0.018	1.30	0.99
After	F	C	2.955	213	424	0.013	0.025		

II. Total Days of Disability (Family members under 20 years old)

Before	M	Test	3.056	226	412	0.110	0.140	2.08*	1.93
Before	M	Control	2.105	222	449	0.099	0.102		
After	M	T	3.806	226	412	0.117	0.176	3.64*	3.01*
After	M	C	5.706	222	449	0.156	0.224		
Before	F	T	2.732	233	444	0.127	0.153	0.71	0.62
Before	F	C	2.436	213	424	0.048	0.073		
After	F	T	4.813	233	444	0.127	0.172	0.80	0.71
After	F	C	5.281	213	424	0.214	0.265		

* To be interpreted as significant difference at $P=0.05$ level.

ing for intraclass correlations. Some examples of intraclass correlation coefficients and their effect on the variance of the mean are taken from the morbidity data of a study of housing environment and family life.

ACKNOWLEDGMENT is made of use of the IBM 7090 electronic computer of The Computing Facility, University of California, Los Angeles.

REFERENCES

1. Wilner, D. M.; Walkley, R. P.; Glasser, M.; and Tayback, M. The Effects of Housing Quality on Morbidity. A.J.P.H. 48:1607–1615, 1958.
2. Wilner, D. M.; Walkley, R. P.; Schram, T.; Pinkerton, T.; and Tayback, M. Housing as an Environmental Factor in Mental Health: The Johns Hopkins Longitudinal Study. Ibid. 50:55–63, 1960.
3. Wilner, D. M.; Walkley, R. P.; Pinkerton, T.; and Tayback, M. The Housing Environment and Family Life. Working paper prepared for Expert Committee on the Public Health Aspects of Housing, Geneva, World Health Organization, 1961.
4. ————. The Housing Environment and Family Life: A Longitudinal Study of the Effects of Housing on Morbidity and Mental Health. Baltimore, Md.: Johns Hopkins University Press, 1962.
5. Sukhatme, Pandurang V. Sampling Theory of Surveys with Applications. Ames, Iowa: Iowa State University Press, 1954.
6. Cochran, William G. Sampling Techniques. New York, N. Y.: Wiley, 1953.
7. Yates, Frank. Sampling Methods for Censuses and Surveys. New York, N. Y.: Hafner, 1949, pp. 175, 260, 263.
8. Hansen, Morris; Hurwitz, William N.; Madow, William G. Sampling Survey Methods and Theory. (Vol. I). New York, N. Y.: Wiley, 1953, pp. 13, 199.
9. Hopkins, C. E. A Critique of Kinsey's "Sexual Behavior in the Human Female." West. J. Surg. 62:177–188 (Mar.), 1954.
10. Cochran, W. G.; Mosteller, F.; and Tukey, J. W. Statistical Problems of the Kinsey Report. J. Am. Statist. A. 48:673–716, 1953.
11. Haggard, Ernest A. Intraclass Correlation and the Analysis of Variance. New York, N. Y.: Dryden Press, 1958.
12. Snedecor, George W. Statistical Methods (5th ed.). Ames, Iowa: Iowa State College Press, 1956, p. 379ff.
13. Laughlin, W. S.; Gray, M. P.; and Hopkins, C. E. Blood Group Genetics of the Basques of Idaho. Acta Genetica et Statistica Medica 6:536–548, 1956/57.
14. Li, C. C. Population Genetics. Chicago, Ill.: University of Chicago Press, 1955.
15. Kendall, M. G. The Advanced Theory of Statistics. Vol. II. London, England: Charles Griffin and Co., Ltd., 1948.
16. Patterson, H. D. Sampling on Successive Occasions with Partial Replacement of Units. J. Royal Statist. Soc. 2:241, 1950.
17. Tanner, J. M. Some Notes on Reporting of Growth Data. Human Biol. 23:93–159 (May), 1951.
18. Henderson, C. R.; Kempthorne, O.; Searle, S. R.; and von Krosigk, C. M. The Estimation of Environmental and Genetic Trends from Records Subject to Culling. Biometrics 15:192–218 (June), 1959.
19. Terzaghi, Jean, and Hopkins, C. E. A FORTRAN Computing Program for Analysis of Mixed Longitudinal Data. (In manuscript.)
20. Fisher, R. A. Statistical Methods for Research Workers (Tenth ed.). New York, N. Y.: Hafner, 1948, p. 224.
21. Moser, C. A. Survey Methods in Social Investigation. London, England: William Heinemann, Ltd., 1958, p. 91.
22. Leech, F. B., and Healy, M. J. R. The Analysis of Experiments on Growth Rate. Biometrics 15:98–106 (Mar.), 1959.
23. Nordskog, A. W. Note on Optimum Group Size for Progeny Tests. Ibid. 15:513–517 (Dec.), 1959.
24. Robertson, A. The Sampling Variance of the Genetic Correlation Coefficient. Ibid. 15:469–485 (Sept.), 1959.
25. Wearden, S. The Use of the Power Function to Determine an Adequate Number of Progeny per Sire in a Genetic Experiment Involving Half-Sibs. Ibid. 15:417–423 (Sept.), 1959.
26. Robertson, A. Experimental Design in the Evaluation of Genetic Parameters. Ibid. 15:219–226 (June), 1959.

The authors are associated with the University of California School of Public Health, Los Angeles, Calif.