Thesis CH
A method for 00070
3 1735 015 512 530

A METHOD FOR EXAMINING
PARTIAL ASSOCIATION IN A POPULATION

Due

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH LIBRARY

Manuscript theses

Unpublished essays and theses submitted for the Master's and Doctor's degrees and deposited in the University of Pittsburgh, Graduate School of Public Health Library, are to be used only with due regard to the rights of the authors. Bibliographical references may be noted, but passages must not be copied without permission of the authors, and without proper credit being given in subsequent written or published work.

This thesis by Sheehe has been used by the following persons, whose signatures attest their acceptance of the above restrictions.

NAME AND ADDRESS	DATE

A METHOD FOR EXAMINING
PARTIAL ASSOCIATION IN A POPULATION

By

Paul R. Sheehe

B.S.B.A., University of Buffalo, 1948

M.B.A., University of Buffalo, 1954

Submitted to the Graduate School of Public Health
of the University of Pittsburgh in partial fulfillment
of the requirements for the degree of
Doctor of Science
in Hygiene

Thesis

University of Pittsburgh

1959

SE 21 '59

PREFACE

I am grateful to Dr. Antonio Ciocco for his purposeful guidance and thoughtful criticism along the way, and for his helpful suggestions in the presentation of this subject. I wish, also, to express my thanks to Dr. Donovan Thompson for helping me over some rough spots and for his careful review of the manuscript in its various stages.

My thanks also go to Mr. Arthur LeGasse and to the personnel of the Computing Center, University of Pittsburgh, for their help in the use of their facilities, without which this work could not have been done.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

1.0. INTRODUCTION

If an experimenter can identify independent elements to which various treatments have been assigned at random, then he may argue that differences observed among treatment categories are due either to chance alone or to the combined effect of chance and the different treatments. If the observed differences would seldom occur by chance alone, then he may infer that the differences are, at least in part, the effects of treatments. A cause and effect relation between treatments and observations is thereby inferred and the logic of this inference is not disturbed by the fact that the composition of one treatment group differs from another in many aspects.

The purely observational study, on the other hand, is not subject to such simple interpretation. A population may be categorized according to some characteristic and differences may be observed among the categories. But any claim of a cause and effect relation may be challenged. For, in general, the randomization of categories to population elements, or of elements to categories, is absent. Observed differences among categories, that is, observed associations, may be considered to be due not to the characteristics on which the categories are based but to the varied composition of the categories with respect to other characteristics.

Is not experimentation preferred to observational studies? Possibly yes, when there is a choice between the two. But usually there is no choice. This is particularly true in the study of man. While genetic traits may be in part the result of random forces, how do we randomize race or age to individuals, or socio-economic status, religion, place of birth, social customs, and so forth? Granted that certain volunteer groups can sometimes be obtained for the experimental trial of therapeutic agents, or of some other kind of treatment, these groups are not the population of ultimate interest. The conditions under which the demonstration of a causal relation, through controlled experimentation, is feasible are usually not the conditions under which we seek to establish, ultimately, a causal relation. Consequently, even those experiments among volunteer or selected groups must be bolstered by observational studies of the population.

A cause and effect mechanism cannot be divorced from the conditions under which it is operative. In the experimental framework, the effect of a treatment is established only with reference to the materials and conditions of the experiment. If one were to sub-classify the experimental elements according to some characteristic and determine the effect of treatments within such sub-classes, he would often find that the effects of a given treatment varied with the sub-class of elements tested. Similarly, in the observational study, the association between a characteristic and an observation may often vary among sub-classes of the population under study, whether or not this association be causal. Some mechanisms, such as those based on Newton's laws of motion or those based on fundamental genetic principles, may operate under very general circumstances, but for the most part the conditions under which a cause operates in a predictable fashion are fairly specific, and it may be said that a causal mechanism is understood only to the extent that the conditions under which it is operative are known. For example, few, if any, persons would argue that infection by a specific virus is not the cause of clinically identifiable poliomyelitis. Yet in the United States, only about one in a hundred infections results in a clinically recognized case.[1] The conditions under which the virus causes clinical illness must be quite limited. As another example, if high speed on the roadway is accepted as a cause of accidents, it must also be granted that in most cases of speeding an accident does not occur; only in particular circumstances does an accident occur. Similarly, when we say heavy exertion causes heart failure, or that the bite of a mosquito causes yellow fever, and so forth, we realize that this is only true sometimes, and that the cause is understood only to the extent that we can specify the conditions under which it actually produces the stated effect.

If a causal relation has been demonstrated experimentally, then the function which an observational study of a population may serve is to determine under what conditions, if any, the effects are seen in the population. If experimental evidence is not available, because experimentation is not feasible or merely because experimentation has not been done, then the function of an observational study may be to test a causal hypothesis. In either case, determination of those conditions in the population under which the cause or hypothesized cause is operative is of prime importance. Consequently, the observer can never be satisfied to view the simple association of one variable with another. He must view the association under a wide variety of circumstances in the population under study. In other words, he must be concerned with partial association.

In theory, one may sub-classify a universe with respect to as many conditions as he chooses and examine the association between two variables within each of the sub-classes. But in practice, the degree of sub-classi-

[1] Maxcy, K. F., " Preventive Medicine and Public Health ", Appleton-Century-Crofts, Inc., 8th ed., N. Y., 1956.

fication is limited by the number of observations which are available: one soon reaches a stage of sub-classification beyond which a large number of sub-classes either contain no observations or contain so few observations that measurements of association within the sub-classes are practically meaningless when they are assessed on a probabilistic basis. Nevertheless, the need for taking account of a large number of conditioning variables remains. For example, it is not enough to measure the association between smoking and lung cancer in a sample from the population. It is of prime importance that the association be measured after adjusting for the conditioning influence of age, race, sex, socio-economic status, rural-urban status, occupational history, and other variables which may be thought to have a bearing on the incidence of lung cancer or on the smoking habits of the population. If there be an association of smoking with lung cancer, one wishes to know how many of such other variables, considered as conditions (rather than consequences of smoking), may be taken into account before the association is destroyed. On the other hand, if there be no strong simple association, one wishes to know whether a strong association would emerge if enough conditions were taken into account.

Similar problems are met when the development of illness in households is studied. It is a fundamental proposition that the health of an individual is determined by the interaction of his biological constitution with his environment, social as well as physical. The household ranks high in the social environment of the individual, and it is therefore reasonable to suppose that household stresses influence the health of its individual members. In particular, illness in one member can itself be viewed as a household stress which influences the health of other members. One of the most obvious of such influences is that of communicable disease. Less clear is the chain of events proceding from a chronic disease stress in the household. Indeed, the household health consequences of a non-communicable health stress may be manifold. The practical purpose of the present study is to examine those consequences in relation to that stress. Simply stated, the question is asked: does a household member with a non-communicable health problem increase the chances of subsequent health problems among the initially healthy members. This problem will be taken up in the first section to follow. In the course of analysis, it will become evident that a simple measure of association does not adequately answer the question. Conditioning factors will have to be taken into account: the household size, the sex distribution of its members, and age distribution. In addition, some account will have to be taken of the conditions under which the survey data were obtained: interviewer characteristics and sample strata. Among the many problems of analysis which will be met, one of the most difficult will involve the small frequencies which fall into the class categories. Thus our study will be faced with the problem of multiple conditioning variables.

This problem of multiple variables is often shunted aside by the investigator: he makes a decision that some variables are more important than others; he picks and chooses among the many factors which rise to mind to arrive at three or four. The data are classified, then analyzed on the basis of this restricted number of variables. Concurrently, the investigator assumes, or rather hopes, that the disturbing influence of other variables is negligible. Now it is true that we can never hope to account for all conceivable conditioning factors, for they are unlimited in number. However, it should be clear that the technique which can account for a larger number of variables than customary classification techniques is to be desired.

Multiple correlation, and, more generally, multivariate analysis and path analysis are well established techniques for handling several variables when examining partial relations. But they are not completely applicable to examining partial association in a manifold. Before applying them, one is forced to designate categories of all classes by quantities. Now the quantity is not, in general, a perfect representation of a category. Consequently, the absence of relation between quantities representative of categories of two factors is not full evidence of the absence of relation between the factors. Further, quantities applied to categories are necessarily discrete variables. But measures of significance associated with these techniques depend upon the assumption that continuous, normally distributed variables are involved. Therefore, significance levels applied to relations between discrete variables must be viewed as approximate, and the closeness of the approximation is often in doubt.

The analysis of variance is more generally applicable to the study of partial association in a manifold, provided that a dependent factor be numerically represented. Yet this technique is used relatively infrequently or applied only to a limited extent in observational studies. This is in contrast with the almost universal application of analysis of variance techniques in the statistical treatment of experiments. The reason for this is not that there is any difference between the ultimate goals of experimentation and observational studies, nor that randomization of treatments can be performed in experiments. Rather, analysis of variance is often discarded as an analytical tool for the study of survey data because the solutions of the least squares equations are much more time-consuming when orthogonality is absent, i.e. when conditioning factors are correlated with each other due to an imbalance in cell frequencies. The great majority of experimental designs are orthogonal; surveys almost always are not orthogonal with respect to all the factors among which it may be desired to study associations. However, with the advent of the electronic computer, the objection that solution of the least squares equations of the analysis of variance is too time-consuming is less valid.

Again, the classical tests of significance for the analysis of variance depend on the assumption of a normally distributed error or residual. If the dependent variable is a set of numerical values which are associated with two or more categories of the dependent factor, this assumption can never be precisely true. Nevertheless, Pitman, Welch, and others, have shown, in the experimental setting, that significance tests based on the normality assumption are rather close approximations to exact randomization tests when several observations in each category and several categories of the dependent variable are involved.[2,3,4] Consequently, it should not be surprising if the analysis of variance becomes a more popular analytical tool for the statistical estimation of partial association from survey data.

But the raw material of the analysis of variance is a numerical variable. When dealing with association in a two-factor contingency table or more generally with partial association in a manifold, it would seem to be more appropriate to approach the problem with the tabled frequencies as

[2] Pitman, E. J. G., "Significance tests which may be applied to samples from any populations III. the analysis of variance test", Biometrika, V29, pp. 322-335, 1937.

[3] Welch, B. L., "On the z-test in randomized blocks and latin squares", Biometrika, V29, pp. 21-52, 1937.

[4] Eden, T. and Yates, F., "On the validity of Fisher's z test when applied to an actual sample of non-normal data", J. of Agric. Science, V23, pp. 8-17, 1933.

the starting point. This has been done for the two-factor contingency table. In 1936, H. Hotelling defined canonical variates. In relation to contingency tables, they are sets of scores, representing the categories of each factor, such that the correlation between the two sets is a maximum, a stationary value, or a minimum.[5] In 1940, R. A. Fisher described an iterative method for arriving at stationary values of scores for the categories of a two-factor contingency table. The factors were eye color and hair color. Choosing arbitrary scores to represent the categories of eye color, scores for hair color were determined which maximized the correlation between the two sets of scores; using the hair color scores, new scores for eye color were determined which maximized the correlation, and this alternating procedure was continued until the two sets of scores stabilized.[6] In 1941, K. Maung was able to show that, in any g by h contingency table, $g \geq h$, there exist $(h - 1)$ canonical correlations and that they correspond to all the maximum and minimum values of the product-moment correlations between all possible scores assigned to the categories of the two factors. He showed in detail a direct method for determining the correlations and the corresponding scores. Further, he showed that the sum of squares of the correlations was equal to chi-square divided by the total number of observations, i. e. Pearson's mean square contingency, ϕ^2.[7] In 1952, E. J. Williams presented a paper dealing with tests of significance

[5] Hotelling, H., "Relations between two sets of variates", Biometrika, V28, pp. 321-377, 1936.

[6] Fisher, R. A., "The precision of discriminant functions", Annals of Eugenics, V10, pp. 422-429, 1940.

[7] Maung, K., "Measurement of association in a contingency table with special reference to the pigmentation of hair and eye colour of Scottish school children", Annals of Eugenics, V11, pp. 189-223, 1941.

of canonical relations applied to contingency tables. As introductory material, he presented a clear summary of the concepts and techniques discussed above; further, he gave a generalization of Lancaster's method of partitioning the chi-square of a contingency table into component parts.[8,9]

These studies by Hotelling, Fisher, Maung, and Williams take classified frequencies as their starting point. In contrast to this, the analysis of variance, as customarily conceived, has its roots in a measured variable. The distinction is basic. We learn to classify things first. We cannot number until we first know how to classify. We cannot measure without using numbers. Thus the assignment of scores to the categories of contingency tables in order to describe a relation is basic. Being basic, it is the more general approach. In fact, Williams, et al., have shown that for two-factor tables the least squares formulae used in the analysis of variance is a particular case of the formulae for maximum correlation scores. When at least one of the two factors is a dichotomy, or when arbitrary scores are assigned to one of the two factors, the maximum correlation scores for the remaining factor are equivalent to least squares effects.

Now the maximum correlation score technique has been applied only to two-factor contingency tables. But in practical survey work it has been emphasized that multiple factors must be considered. Therefore, it is our intention to show that the maximum correlation scores technique can be applied to multiple factor classifications. Least squares formulae will be shown to bear the same equivalence to maximum correlation score formulae

[8] Williams, E. J., "Use of scores for the analysis of association in contingency tables", Biometrika, V39, pp. 274-289, 1952.

[9] Lancaster, H. O., "The derivation and partition of χ^2 in certain discrete distributions", Biometrika, V36, pp. 117-129, 1949.

for the multiple factor classification as exists for the two-factor classification. Instead of a two-factor contingency table, a multi-factor table will be considered in which there are several independent factors and in which there is one dependent factor. The general case of determining scores for all factors will not, however, be considered; instead, attention will be restricted to the case for which the dependent factor is either a dichotomy or a classification whose categories can be represented by quantitative characters. No such restriction will be placed on the independent factors. It will be shown that the scores developed in this approach may be interpreted as the partial effects of the factor categories. A chi-square test of significance of effects will be proposed. This proposal will be tested on the basis of empirical results of sampling from known populations, using the IBM 650 computer for this work.

But we must not lose sight of the practical problem which has created the need for such theoretical development. That is the problem which was underscored in the opening paragraphs: extending the number of variables which can be handled in the analysis of survey data. In the following section which deals with the stress of household health problems, we shall attempt to illustrate the technique for extending the number of variables and to show how this technique is integrated into the general analytical method. In subsequent sections the theoretical development of the technique is taken up.

2.0. HEALTH PROBLEMS AS AN INDICATOR OF SUBSEQUENT HEALTH PROBLEMS IN THE HOUSEHOLD

Do health problems 'run' in households? Few persons will contest that some diseases do. Communicable disease is a notable example: it is reasonable to believe that transmission of a biological agent of disease from one household member to another, in general, is more easily accomplished than direct transmission to persons outside the household. It is also granted that some hereditary diseases, such as diabetes, may tend to cluster in households. In addition, poor housing conditions and other environmental stresses common to all members may be responsible for the clustering of health problems in households. Again, the psychiatrist may relate the development of mental illness to household tensions. Further, 'accident-prone' persons, those who have 'too many' accidents, may very well cause injury to others in the household. And, finally, it may be that chronic disease in one member sometimes induces tensions among other members, leading to diseases of varied kinds. All these examples of familial aggregation illustrate the generally held idea that illness in one member of a household implies an increased tendency for illness to be present in other members. It is not the intent of the present study either to substantiate or to disprove this. Rather, the relation of household health problems as a whole to antecedent non-communicable health problems in the household is to be investigated.

The distinction may be made clear by comparing our study to a study by Downes, "Illness in the Chronic Disease Family".[10] The data were

[10] Downes, J., "Illness in the chronic disease family", American Journal of Public Health, V32, pp. 589-600, 1942.

compiled from monthly interviews of families in the Eastern Health District of Baltimore, from mid-1939 to mid-1940. Individuals who had no chronic disease were divided into two groups: Group I, those having, and Group II, those not having, a chronically ill person in the immediate family. The illness rate for Group I was found to be greater than that for Group II. This finding indicates that a sort of familial aggregation of health problems was present in the families which were indexed by the presence of a chronically ill person. Determination of the presence of chronic disease and other illness was, however, concurrent. It is not known, in general, whether chronic disease existed first, followed by an increase of illness in other household members, or whether chronic disease merely occurred more often in households characterized by a high general illness rate.

The study we are about to take up is oriented differently. Rather than just chronic disease problems, all kinds of non-communicable household health problems which occur during a one-year interval are the index of a household health stress. Households with health problems in the following year, occurring to members who had no health problem in the first year, are studied in relation to that stress. The object of interest is antedated by the stress. Under this approach, concurrent familial aggregation does not contribute to the relation. Rather, present health problems are viewed, sometimes as an indicator, sometimes as a cause of household health problems in the future.

It seems reasonable to expect that, because of familial aggregation, a positive relation between antecedent and subsequent health problems should exist in households taken from the general population. Nevertheless, it should be pointed out that the truth of this hypothesis does not follow necessarily. This is because a selective, negative force is actuated by every occurrence of a health problem stress. In our ignorance of all the manifold conditions under which health problems develop, we can say that some persons in a household are more susceptible to illness than others. If the more susceptible persons come down with illness in a given interval of time, then the less susceptible persons remain. These less susceptible persons constitute a 'preferred risk'. It is therefore entirely possible that the selective, negative force of disease within the household, due to differences in susceptibility of its members, may balance or outweigh the supposed familial aggregation of illness, due either to characteristic differences of households within the community or to the stress which a health problem sets up in the household. The resultant of these opposing tendencies may vary under different conditions. In this event, it becomes important not only to examine the overall relation of household health problems in a population, but also to study how this relation changes under various circumstances.

In the study we take up here, the simple relation between antecedent and subsequent household health problems is investigated first. Then certain characteristics, i. e. size, average age and sex distribution of the household, which from _a priori_ consideration influence selection within the household, are taken into account. Also, certain circumstances in respect to the sample design, i. e. stratification and interviewer characteristics, are taken into account. The customary classification techniques are found to be inadequate in accounting for all these variables. The new technique developed in subsequent sections is employed. After adjustment for household size, average age, and sex distribution as well as stratum and interviewer characteristics by this technique, the adjusted relation of household health problems can be interpreted. Following this, the question of household health problems as a cause of subsequent health problems is investigated. This is done by analysis of the variation of the adjusted household health problem relation among size-age-sex specific groups of households. Statistical properties of measures of partial association, developed in the later sections of this work, are applicable in that analysis.

2.1. Source and Limitations of the Data

The material for this study comes from a survey of a district of Pittsburgh, Pennsylvania. In July 1951, the Graduate School of Public Health, University of Pittsburgh, conducted a survey of some 3000 households in the central portion of the Arsenal Health District of the Pittsburgh Health Department. The area sampled by the survey comprised 22 of the 194 census tracts in the city and had a total population of about 80,000. On the basis of average monthly rental data, households in the study area were judged to be well below the rest of the city in income level. One of the principal objectives of the study was to measure the health characteristics of households in the area by means of personal interview of responsible members of the selected households.

A detailed discussion of the sampling design is given by Horvitz.[11] Some of the main features are described below. The 468 blocks in the district were classified into three strata, according to the number of dwelling units occupied (1940): Stratum I, 100 or more; Stratum II, 50 and

[11] Horvitz, D. G., "Sampling and field procedures of the Pittsburgh Morbidity Survey", Public Health Reports, V67, pp. 1003-1012, 1952.

less than 100; Stratum III, less than 50. The primary sampling unit within each stratum was the block. For each selected block a proportion of the dwelling units in that block was chosen for interview. The block sampling ratios and dwelling unit sampling ratios within a given block were so chosen that each dwelling unit in the entire study area had an equal chance (approx. 2/15) of entering the sample. Both blocks and dwelling units within blocks were selected by systematic sampling with a random start. (For illustrative purposes in the preliminary analysis, the sample data are treated as a simple random sample; but in the refined analysis account is taken of the classification into three strata.) In June 1952, approximately one year later, the same households as were selected in 1951 were interviewed again, unless the household had moved or refused to cooperate in the second survey.

The response problem was given particular attention in the design of the first interview. Horvitz relates that there were 18 enumerators. Ten were male medical students, two were female medical students, one was a female worker experienced in health surveys, and 5 were non-medical female college students. Through the assignment of interviewers to blocks in a random fashion, differences in response, according to interviewers, could be assessed. Horvitz concludes, from a study of illness rates in relation to interviewer groups, that there was conclusive evidence of differences in illness rates elicited by the various interviewers, differences not ascribable to chance. (Consequently, our analysis adjusts for certain interviewer characteristics as well as for strata.)

One important limitation of the data, as for most survey data, is non-response. Table 1 shows that, of 2957 households initially selected for interview, 166 failed to cooperate in the first interview; another 220

Table 1

An Account of Non-respondent and Excluded Households

Total Households Initially Selected for Interview:			2957
Less: Failed to Cooperate, 1st Interview -		166	
Non-respondent, 2nd Interview -			
Failed to cooperate	220		
Moved away	251	471	
Total Non-response, Either Interview:			637
Households for which Some Data Available from Both Surveys:			2320
Less: Information Incomplete for One or More Members -			49
Households for which Complete Data Available from Both Surveys:			2271
Less Exclusions:			
No Members Free of Health Problem in 1st Year -		58	
Communicable Disease Present at 1st Interview -		71	
Total Exclusions:			129
Total Households Available for Analysis:			2142

which cooperated in the first failed to do so in the second. An additional 251 households moved after first interview and were lost from observation. For 49 of the remaining households, information on one or more members is incomplete. This leaves 2271 households for which complete information is available from both surveys. Of these, there are 58 in which no members were free of health problems as of the first interview. An additional 71 households are excluded from the analysis because communicable disease was present at first interview. Thus, we are left with 2142 of an original 2957 selected households. In a strict statistical sense, then, the results of this study are not generalized to include those households in the community,

(a) which would not have cooperated,

(b) which would have moved,

(c) which would have failed to give certain pertinent information,

(d) which would have no members initially free of health problems,

or (e) which would have reported a communicable disease at the first interview. It is noted, however, that communicable disease in the second year is not excluded as a sequel to non-communicable health stress in the first.

A further limitation of the data is that their accuracy is limited to the accuracy of respondents' recall.

2.2 Definitions

In order to analyze the problem in specific terms, certain measures must be defined. And, as Ciocco puts it, "The important point to keep in mind is that the criteria (for classification) adopted and the resulting classifications have an important bearing on the interpretations to be drawn from the findings."[12] Consequently, it is doubly important that the definitions which follow be made clear at the outset.

2.21. Household

A household is consisted of a number of persons living at a common dwelling unit. In the great majority of cases, the household is equivalent to a family, or a single person living alone, but there are a few cases for which more than one family, or several unrelated persons constitute a household.

[12] Ciocco, A., Densen, P., and Horvitz, D., "On the association between health and social problems in the population", The Milbank Memorial Fund Quarterly, V31, pp. 265-290, 1953.

2.22. A Person with a Health Problem

A person with a health problem is:

one with a reported physical impairment or chronic disease existing within one year prior to interview;
one reported hospitalized at any time in the year prior to interview, except hospitalization for delivery without complications, and except for routine check-up;
or, one reported to be semi-ambulatory or confined to bed for any interval during the month prior to interview.

This definition is made in view of the limitations of the survey. It was felt that respondents could not give accurate histories of illness beyond one month prior to interview, except for hospitalization and chronic illness. This definition of a person with a health problem conforms closely to that adopted in earlier studies based on the Arsenal Survey.[13]

2.23. A person with a communicable disease

A person with a communicable disease is any reported person with a health problem who, in the month prior to interview, had an illness coded (International Statistical Classification of Diseases, Injuries and Causes of Death, 6th Revision, 1948):

001-138	infective and parasitic,
470-475	acute upper respiratory infection,
480-483	influenza,
490-493	pneumonia.

2.24. Age of person

The age of a person is taken as the reported age in years to last birthday prior to first interview.

2.25. A Propositus

A propositus is a reported person, over one year of age at first interview, who had no health problem during the first year, and who did

[13] *Ibid.*, p. 272.

not leave the household in the second year, except by reason of death or illness.

2.3. Preliminary Analysis

In this preliminary analysis, the second year health problem rate for propositi is studied. The rates are compared for propositi from households with, versus those from households without, a first year health stress. A negative relation is observed; that is, the second year health problem rate is greater for propositi from households without initial health stress. This is true, despite the fact that an overall familial aggregation of health problems is observed in the first year. It is shown that these two findings are not contradictory, that the difference may be explained as due to greater negative selective forces acting on the serial (one-year time lagged) relation than on the concurrent (aggregation within first year) relation. Attention is then turned to the household as the unit of analysis. That part of the household consisting of propositi (a propositus household) is studied. Two groups of propositus household, those indexed by the existence or non-existence of a household health problem in the first year, are compared. Again a negative relation is found: relatively more propositus households which were free of a household health stress in the first year developed health problems in the second year. The measure of the strength of this relation is taken as R, the product-moment correlation between health status in the first year and health status in the second year. It is shown that R is an adjusted difference in second year health problem rates between the two groups of propositus households. This observed negative relation is significant on the .1% level, which indicates that an overall preponderance of negative forces existed in the population.

2.31. Comparison of Propositi in Households, I, Without and, II, With First Year Health Problems

Although the individual propositus is not, in a strict statistical sense, the proper unit for analysis, it may be informative to compare individual propositi in households having a first year health problem with propositi in households free of first year problems. Table 2 shows this comparison.

Table 2

Number of Health Problems and Health Problem Rate in the Second Year for Propositi in Households, I, Without and, II, With a Health Problem in the First Year

Household Status 1st Year	Number of Propositi	No. of 2nd Year Health Problems for Propositi	2nd Year Health Problem Rate (Hlth. Pbs. per 100 Propositi)
I Without Hlth.Pbs.	4809	1632	33.9
II With Hlth. Pbs.	1796	552	30.7

On a relative basis, the propositi of Group I experienced the higher rate of health problems, 33.9, as compared to the 30.7 rate for propositi of Group II. This is an observed negative relation, quite different from the positive relation of chronic disease to concurrent health problems seen by Downes.

It is emphasized, however, that the negative relation seen here is not contradictory to Downes' finding, nor is it contradictory to the general idea that there is familial aggregation of illness. This is because any familial aggregation which may have been present in the first year does not, in itself, contribute to the serial, that is the one-year time lagged, relation seen in Table 2. Consider, as a simplified hypothetical example of aggregation, the following. Suppose we are dealing only with 100 households of size three. Also, taking a rather obvious illustration of familial aggregation, suppose that first year household health problems occurred only in pairs in 50 of the households. Table 3 indicates this situation.

Table 3

Hypothetical Distribution of Number of Members (X)
with First Year Health Problems,
for Households of Size S = 3

	No. of Persons with Hlth. Pbs. in the Household, 1st Year (X)				Total No. of Households (n)
	0	1	2	3	
Frequency of Households (f)	50	0	50	0	100

As the measure of familial aggregation, we take the observed variance of X (the number of persons with health problems) in ratio to the variance in X which would be expected if health problems were distributed randomly to individuals; this we shall call the coefficient of familial aggregation. If the coefficient is greater than 1, then there is an observed familial aggregation; if less than 1, then there is an observed dis-aggregation. In the present hypothetical case, there are 50 x 2 = 100 persons with health problems out of a total of 100 x 3 = 300 persons. Then the overall proportion of persons with health problems is P = 100/300 = 1/3. If these persons had been distributed by chance, then the expected distribution would be the terms of the binomial, $(P + Q)^S$, which, in this case, is $(1/3 + 2/3)^3$. The variance of this expected distribution would be $V_e = SPQ = 3(1/3)(2/3) = 2/3$. Actually, the observed variance, V_o, as calculated from Table 3 is: $V_o = (\Sigma fX^2/n) - (\Sigma fX/n)^2 = (50 \times 2^2/100) - (50 \times 2/100)^2 = 2 - 1 = 1$. The ratio of observed to expected variance is therefore $1 \div 2/3 = 1.5$. This ratio greatly exceeds one; hence there is a strong observed familial aggregation.

Now in the foregoing situation there are 50 propositi from the households with first year health problems and 150 propositi from households without first year health problems. If, say, 1/5th of each of the two propositus groups come down with a health problem in the second year, then the health problem rate for each propositus group is 20 per 100 propositi, and no serial relation exists. Thus, the mere existence of familial aggregation in the first year tells us nothing, necessarily, of the serial relation.

These same considerations apply to the more complex case with which we are dealing, in which households of various size are involved. Table 4 shows the distribution of number of persons with first year health problems in households of size 3,4, ..., 11, as observed in the Arsenal data. From the data for each household size, an estimate of the proportion of persons with health problems, P_i, is made in the same manner as illustrated for Table 3. (However, a correction is made for the fact that households in which all members have a health problem are excluded;[14] see exclusions, Table 1.) Then the variance, $(V_o)_i$, of the observed distribution of X is computed and divided by the variance, $(V_e)_i$, of the expected distribution of X to form the coefficient of familial aggregation, A_i, in the same fashion as illustrated for Table 3. (Again, a correction is made because of the truncated distributions.) Households of size 1 and 2 are omitted because the observed coefficient of aggregation must always be unity for those cases.

[14] For a method of estimating P when the frequencies in the extreme class of binomial distributions are unknown, see:
Li, C. C., "Segregation of recessive offspring", Methods in Medical Research, V6, pp. 3-16, The Year Book Publishers, Inc., Chicago, 1954.

Table 4

Distribution of Number of Members with First Year Health Problems, Expected Chance Distribution, and Coefficient of Familial Aggregation, for Size-Specific Groups of Households*

Household Size	Total No. of Hh.	Observed and (Expected) No. of Members with 1st Year Health Problems (X)						Obs./Exp. Var. Var.	Coeff. of Agg.
(s_i)	(n_i)	0	1	2	3	4	5	$(V_o)_i/(V_e)_i$	(A_i)
3	694	415 (404)	149 (242)	130 (48)	X X	X X	X X	.617/.450	1.37
4	432	281 (280)	126 (128)	24 (22)	1 (1.7)	X X	X X	.367/.367	1.00
5	236	156 (156)	68 (67)	10 (12)	2 (.9)	0 (.0)	X X	.376/.367	1.02
6	128	72 (75)	50 (42)	6 (10)	0 (1.3)	1 (.1)	0 (.0)	.437/.471	.93
7	41	23 (13)	12 (16)	6 (8.5)	0 (2.5)	0 (.4)	0 (.0)	.536/.888	.60
8	20	9 (9.6)	9 (7.4)	1 (2.5)	1 (.4)	0 (.1)	0 (.0)	.610/.640	.95
9	9	4 (4.0)	3 (3.4)	2 (1.3)	0 (.3)	0 (.0)	0 (.0)	.617/.710	.87
10	4	2 (.8)	1 (1.4)	0 (1.1)	0 (.5)	0 (.2)	1 (.0)	4.32/1.28	3.32
11	6	4 (4.3)	2 (1.5)	0 (.2)	0 (.0)	0 (.0)	0 (.0)	.222/.323	.69

* Households in which all members have health problems are omitted, per Table 1.

In general, when observed frequencies exceed the expected at the extreme values of X, aggregation is present; when observed frequencies tend to cluster about the center, dis-aggregation is present. For the 694 households of size three, strong aggregation is present, A = 1.37, while only very slight or no aggregation is present for households of size four and five. Households of larger size tend to show dis-aggregation, with the exception of the four households of size 10. But the frequencies for the larger sized households, seven and up, are quite small and fluctuations in the observed coefficient of aggregation can be expected to be large. The aggregation picture is really dominated by households of size 3 through 6, these constituting the great majority of households. As an overall, summary measure of aggregation in the whole sample, we take the average coefficient of aggregation, $A_{ave} = \sum n_i (V_o)_i / \sum n_i (V_e)_i$. This is the ratio of observed variance to expected variance, averaged over all sizes of household, with number of households as weights. In our case, $A_{ave} = 1.16$; thus, overall, there is a moderate familial aggregation of health problems in the first year.

Summarizing thus far, an overall negative serial relation is observed (Table 2) in spite of the fact that an overall familial aggregation is observed in the first year (Table 4). This is because the serial relation is subject not only to familial forces but also to others. We may hypothesize two kinds of such forces, positive and negative:

(1) positive

a) familial - the fact that some households had health problems in the first year may be the result of poor housing, poor social environment, genetic traits, etc., acting in common upon all members of such households; the households free of problems the first year may be in a relatively safe environment; then, from such familial forces, we may expect the propositi in the households with initial health stress to experience a higher rate of second year health problems than the other group of propositi;

b) direct causal - the occurrence of health problems may act as a stress tending, on average, to cause future health problems in propositi in households containing such stress; again, this would tend to show up as a positive relation;

(2) negative

a) selective - propositi in households with first year health problems did not come down with health problems for as much as a full year, when confronted with the same positive familial and directly causal forces as less healthy members; this may mean that these propositi have individual characteristics which make them less susceptible to health problem stresses than the propositi in households without first year health problems; in this event, a negative serial relation tends to be present;

b) direct causal - it is conceivable that the occurrence of health problems in some members of the household acts to change the mode of life of the remaining propositi such that propositi are exposed to less risk of health problems in the future; in this case, health problems would constitute a negative force on the propositi.

Probably a mixture of all these tendencies is at play in the actual population, but apparently negative forces outweigh the positive forces, producing the observed negative relation seen in Table 2.

2.32. Comparison of Households, I, Without and, II, With First Year Health Problems

Now we turn from the individual propositi as the analytical units to the household. The household was the sample element in the survey. Hence statistical measures of significance are more valid when the household, rather than the individual propositus, is considered to be the analytical unit. Furthermore, the analysis is concerned with the household health problem relation for various classes of households rather than for classes of individuals.

In order that the customary definition of a household, given in 2.21, be distinguished from that part of the household consisting of propositi only, the latter is defined as a propositus household. Table 5 is a four-fold classification of propositus households according to health problem status in the first year and in the second year.

Table 5

Four-fold Classification of Households According to Household Health Problem Status at First Interview and Propositus Household Health Problem Status at Second Interview

First Interview	Second Interview		Total Households
	No Hlth. Pb. in Prop. Hh.	Hlth. Pb. in Prop. Hh.	
I. No Household Hlth. Pb.	600	882	1482
II. Household Hlth. Pb.	321	339	660
Total Households	921	1221	2142

It is first noted that, while only 660 households were found with health problems in the first interview, 1221, nearly twice as many, propositus households had health problems in the second interview. At first glance, this information would appear rather surprising, since persons with health problems in the first interview do not contribute to the count of health problems in propositus households in the second year. Other things being equal, it would be expected that the number of propositus households with health problems in the second year would be something less than 660. However, this apparently large increase can be explained as due, for the most part, to differential response in the two interviews. The questionnaire used in the first interview contained only general questions about the occurrence of health problems, whereas in the second survey a long list of specified illnesses was presented on the questionnaire. This revision in procedure was apparently successful in improving respondent recall of illness. Of course, other factors are at play to create a difference in health problems for the two periods; actual illnesses, as distinguished from reported illness, may have been at different levels in the two years, for propositi are a full year older at time of second interview, epidemics may have had greater influence in the population in the second year, and so forth. However, it seems that these other factors could explain only a minor part of the increase, and that this large increase illustrates the magnitude of the problem of differential response.

Now attention is directed to a comparison of propositus households free of first year health stress (Group I) with those subjected to first year health stress (Group II). Rather than choosing the simple difference between proportion of households with second year health problems as a measure of the comparison, we choose the product-moment correlation, $R = \sqrt{\chi^2/n}$, as the measure of relation. In 3.6 it will be shown that, for interpretative purposes,

$$R = (p_2 - p_1)\sqrt{p'q'/pq} \quad ,$$

where $(p_2 - p_1)$ is the difference in proportion of Group II and Group I households, resp., with second year health problems,

and p is the proportion of all households with second year health problems, $(q = 1 - p)$,

and p' is the proportion of all households with a first year health problem stress, $(q' = 1 - p')$.

Thus the product-moment correlation is the simple difference between proportions, $(p_2 - p_1)$, adjusted by the factor $\sqrt{p'q'/pq}$.

The denominator, $\sqrt{pq}$, of the adjusting factor may be termed the 'inherent' variability of rates due to the level of p_1 and p_2. This is because p is a weighted average of p_1 and p_2. Now p describes a binomial universe with standard deviation of $\sqrt{pq}$. Then, when $(p_2 - p_1)$ is divided by $\sqrt{pq}$, the result is a standardized difference in proportions. For a given difference in proportions, $(p_2 - p_1)$, a stronger relation is indicated when the average of these two proportions is far from .5 than when the average is near to .5. For example, if the difference between p_2 and p_1 is .1, a stronger relation is indicated when p = .2 than when p = .5, as

$(p_2 - p_1)/\sqrt{pq} = .1/.4 = .25$ in the first instance,

and $(p_2 - p_1)/\sqrt{pq} = .1/.5 = .20$ in the second instance.

This standardized difference in proportions, standardized by the inherent variability attached to the average level of the proportions, is multiplied by the numerator, $\sqrt{p'q'}$, of the adjusting factor. The numerator term is a maximum when half of the households fall in each of the two categories being compared, and decreases uniformly as the number of households in one category increases beyond .5 of the total number of households. Thus, when a given standardized difference in proportions is observed between two equal sized groups sampled from the population, a stronger relation is considered to be indicated than when the majority of elements falls in one category and a few in the other. This weighting can be rationalized on the basis of a selection principle. Classification into two categories constitutes a 'selection' of cases for each category. Now, a criterion for selection which produces a relatively few elements in one category, as compared to a criterion which produces approximately equal numbers of elements in each category, would be expected to produce larger differences more often. Thus, the factor, $\sqrt{p'q'}$, standardizes for inherent variability of differences between p_2 and p_1

due to unequal size of categories.

In summary, the adopted measure of relation, R, is a difference in rates, adjusted for inherent variability in the rates due to differences in category size, and due to the average level of the rates.

Now with reference to Table 5, $(p_2 - p_1) = (339/660) - (882/1482) = .513 - .595 = -.082$, or -8.2%. When this difference is adjusted by the factor, $\sqrt{p'q'/pq} = \sqrt{\frac{(660/2142)(1482/2142)}{(1221/2142)(921/2142)}} = .933$, the adjusted difference in proportions, that is the product-moment correlation, is

$$R = -.082(.933) = -.077, \text{ or } -7.7\%.$$

Thus, with households as the unit of analysis, the serial relation is negative, as previously found for the serial relation based on individual propositi as the analytical unit. Again, the sample reflects that the net influence of familial, selective and direct causal forces, described in detail in 2.31, is negative. But how confident are we that this negative relation existed in the population from which the sample was taken? To answer this question, we must know the distributional properties of R. In section 3.0, the properties of the maximum correlation in the general two-factor contingency table are reviewed. In the particular case of the four-fold table, such as Table 5, it is shown that the maximum correlation reduces to the product-moment correlation, R, and that R is approximately normally distributed, with variance, $1/n$, where n is the sample size. Then, testing the observed R against the null hypothesis,

$$(R - 0)/\sqrt{1/n} = \sqrt{n}R = \sqrt{2142}(-.077) = -3.57.$$

Referring this value, -3.57, to a normal table, we find that a value as great or greater would occur by chance less than one time in a thousand under repetitive sampling, if there were no relation in the population. Thus we can infer with a high degree of confidence that a negative serial relation truly existed in the Arsenal population.

As stated before, we can attribute this negative relation to an excess of negative forces. Discarding, for the moment, the possibility that there is a preponderance of negative direct causal forces, this suggests that the negative relation can be explained by a preponderance of selective forces. In the next sub-section, certain selective and familial forces are specified. Some of those forces can be taken into account by sub-classification of the Arsenal sample data. But there are others which cannot be taken into account by further sub-classification because too small frequencies in the ultimate categories would be encountered. In order to take all specified forces into account, a new measure of partial serial relation is used. This measure is quite analogous to the simple measure, R, above. Its distributional properties are also found to be approximately normal. Consequently, tests of significance, similar to that applied in the above preliminary analysis, can be applied in that more refined analysis.

2.4. The Serial Relation, Adjusted for Various Conditioning Factors

Adjustment for the influence of conditioning factors is taken up here. The specific factors are household size, average age, sex distribution, atratum and interviewer characteristics. It is desired to obtain the serial relation of household health problems for sub-groups of households which are comparable in all five of the above respects. The traditional way to accomplish this is to classify the sample data according to all five factors and to compute the serial relation within each ultimate size-age-sex-stratum-interviewer-specific category. But it is found, because of the size of the available sample, that the ultimate category frequencies would be much too small. In a majority of ultimate categories, there would be no frequencies present; thus, not even a measure of serial relation could be computed for these categories. Furthermore, in the remaining categories for which a measure could be computed, the frequencies would be so small that the computed serial relations would be practically meaningless when assessed on a probabilistic basis.

Therefore two essentially different techniques for adjusting the serial correlation are employed. The first is the traditional classification: the data are classified according to household size, average age and sex distribution. Most of the ultimate size-age-sex categories contain frequencies which are fairly large, 30 or more. But classification stops here. In order to further adjust for stratum and interviewer characteristics, the second, new, technique is applied. For this new technique, each ultimate size-age-sex category containing 30 or more frequencies is treated as a sample from the corresponding size-age-sex specific sub-population. Within each of these categories a partial serial relation is computed. The computation of this partial serial relation is accomplished by applying the results of the theoretical development in subsequent sections of this work, sections 4 and 5. The partial serial relation is the serial relation which exists after the influence of stratum and interview characteristics is adjusted for. The procedure for adjustment is quite analogous to the balancing procedure carried out in the ordinary factorial experimental design; the only difference is that the factorial experimental design usually adjusts for conditioning influences by balancing the number of observations taken at each level of the factors, whereas in our case, we must take the observations from the population as they come. The disturbing influence of conditioning factors is adjusted, not through controlling the number of observations taken, but through correcting the observed relation for the influence of the unbalanced frequencies. This adjustment technique is the formal equivalent of the familiar least squares analysis in experimental work. However, the approach differs from least squares in two respects, one conceptual, the other practical. Conceptually, the technique is founded on the principle of maximizing the squared correlation between scores assigned to the categories of contingency tables, rather than minimizing the squared error for a continuous variable. Practically, the method differs from the usual straight-forward application of the least-squares formulae in that an adjustment for continuity is applied to all the frequencies used in the computation formulae. This adjustment for continuity comes from the development in section 5; there it is shown, by actual experimental trial on the IBM 650 computer, that an adjustment is necessary to increase the validity of measures of partial relation based on least-squares formulae.

Thus, the adjustment procedure combines two different techniques: first, classification as far as possible; then, mathematical adjustment of remaining conditioning factors. The partial, i. e. adjusted, serial relations in each ultimate category of the classification systems are then combined (averaged) to produce a single, overall, serial relation adjusted for all the conditioning variables. Conceptually, this adjusted serial relation is based on comparisons of households only in comparable size, age, sex, stratum and interviewer groups. From the known statistical properties of the size-age-sex specific relations, as determined in section 5 of this work, the significance of the overall adjusted serial relation can be assessed, and an inference can be made.

2.41. Choice of Conditioning Variables to be Taken into Account

The very process of a health problem coming into existence creates a selective negative force on the serial relation of household health problems. The following illustration which considers selection involving the factors, household size, age and sex distribution, is over-simplified, but, being such, emphasizes this selective force.

First, let us suppose that, ceteris paribus, two persons constitute a greater exposure to health problems than does one person, that health problems occur more often to older persons than younger persons, and that the adult female has health problems more often than the adult male.

Now consider two households composed of three individuals, father, mother and daughter, aged 40, 40, and 16, resp. In household number I, suppose no one has a health problem in the first year. In household number II, suppose mother has a health problem in the first year. Then the propositi of household I are:

1 father, age 40
2 mother, age 40
3 daughter, age 16.

Thus there are three propositi; their average age is 32; and they are predominantly (2/3) female. In household II, due to mother having a health problem, the propositi are:

1 father, age 40
2 daughter, age 16.

Here there are only two propositi; their average age is 28; and the female does not predominate.

Consider now the likelihood, ceteris paribus, of a second year health problem occurring to one or more propositi in each household:

(1) by reason of there being more propositi in household I than in II, the likelihood of a health problem occurring to at least one of the three propositi in I is greater than that for at least one of the two

propositi in II;

(2) the average age of propositus household I is greater than for propositus household II, there being two 40 year olds in the former to one 40 year old in the latter; then, since it is assumed that health problems occur more often to older persons, propositus household I is more likely to have a second year health problem than propositus household II, because of the age difference brought about by occurrence of the first year health problem;

(3) propositus household I has a predominance of females while II does not; I has an adult female propositus, II does not; then, assuming it is true that the adult female more often has a health problem than the adult male, household I is again more likely to have a second year health problem than household II.

The influence of each of these factors, size, age, sex, is shown here to be the result of the very process of a health problem having occurred. Each of these factors illustrates the selective process which occurs due to differences among the individuals within the household.

Now, we can turn about and say that size, age and sex distribution can also be positive familial forces rather than negative selective forces. Thus, consider household III, comprised of five females, average age 40, and household IV, comprised of a married couple, average age 30. Then household III is more likely to have a first year health problem because,

(1) there are more individuals at risk,

(2) the individuals at risk are older,

(3) a greater proportion of females is at risk.

Further, if the more likely event occurs, household III does have a first year health problem, while IV does not. Then propositus household III will still have the greater likelihood of a second year health problem because the household III propositi are still

(1) greater in number,

(2) older,

(3) more predominantly female.

This, then, illustrates that differences in size, age and sex distribution may tend to create a positive serial relation. But isn't this line of reasoning contradictory to the former? Not at all. For in the latter illustration, differences among households, not differences among individuals, account for the positive force. In the former illustration, differences among individuals within households, not differences among households, account for the negative force.

If, then, we classify propositus households so that comparisons are made only for households of comparable size, age and sex distribution, we should be removing sources of both negative and positive nature. If the serial relation becomes more negative, after such adjustment for size, age and sex, then a greater positive (familial) force than a negative (selective) force should have been removed, and vice versa.

There are additional factors of classification, the influence of which it would appear desirable to adjust in this study. These relate to the conditions under which the survey was carried out, and they are, firstly, strata, and secondly, interviewer characteristics.

Adjustment is made for strata, by the 3-category classification,

I 100 or more dwelling units per block,

II 50 and less than 100 units per block,

III less than 50 units per block.

There are two reasons for adjusting the serial relation for possible influences of strata: firstly, to bring the analysis more into line with the sampling design, as described in 2.1; secondly, to account for a factor with socio-economic overtones, since blocks with larger numbers of dwelling units in the Arsenal Health District may, in general, be associated with lower socio-economic status. Thus, a presumed positive familial force due to socio-economic status is partly removed by classification on this factor.

Finally, it has been mentioned that Horvitz found differences in response attributable in part to interviewer characteristics in the first interview. Of the 18 enumerators in the first interview,

ten were male medical students - group A,
two were female medical students - group B,
one was a female non-student - group C,
five were female non-medical students - group D.

In the second survey, all sixteen interviewers were male medical students, group A. In the analysis which follows, groups A and B are combined and groups C and D are combined. The resultant two groups, AB and CD, are nearly confounded for sex and type of school. That is, in group AB, the interviewers are medical students, predominantly male, while in group CD the interviewers are predominantly non-medical students and are female. Therefore, by the use of only two categories of interviewer, AB and CD, the small categories, B and C, are eliminated, but sex and type of school are both fairly well preserved. Then adjustment on the basis of this classification should account for most of the possible spurious influence of both interviewer sex and type of school on the serial relation of household health problems.

2.42. The Serial Relation, Adjusted for Propositus Household Size, Average Age, and Sex Distribution, and Adjusted for Stratum and Interviewer Characteristics

Classification of propositus households by size, age and sex yields the frequencies shown in Table 6.

Table 6

Classification of Propositus Households According to Size, Average Age and Sex Distribution

Age in Years (Ave.)	Sex Dist.*	Size				
		1	2	3	4	5&up
15-29	M	21	135	104	143	158
	F	17	14	117	54	92
30-44	M	30	124	99	96	57
	F	25	19	97	27	45
45-54	M	14	97	33	15	4
	F	36	12	42	5	1
55&up	M	59	179	24	10	1
	F	72	23	35	5	1
Total		284	603	551	355	359

* F denotes propositus household in which the females predominate, M denotes that in which half or more of the members are males.

Frequencies of 30 and above are underlined. Some of the interrelationships of size, age and sex are evident in this table: propositus households of larger sizes are very rare in the older age groups; predominantly female propositus households are rare in the younger age groups. The underlined frequencies total 1904; the remainder total 238. Thus almost 90% of the sample is represented in the 22 underlined categories, while the remaining 10% is distributed among the 18 other categories. The analysis which follows ignores the latter rare categories because the statistical properties of the measures of partial association within them are probably not well approximated by the method developed in subsequent sections, and because, numerically, these categories are relatively unimportant in the population. Also it is believed better to discard these categories rather than to combine them with other neighboring categories in a more gross classification system. For example, it is believed better to discard the 21 female households of size one in the 15-29 year age group rather than to produce extreme heterogeneity of categories, with consequent difficulties in interpretation, by putting these 21 households in the same category as the 135 predominantly male households of size two in the 15-29 year age group.

The frequency in each of the size-age-sex specific categories of Table 6 can be considered as a sample from each corresponding size-age-sex specific sub-population of the Arsenal District. Then, a four-fold table of household health status in first and second year can be constructed for each such category, in the same manner as Table 5 was constructed for the sample as a whole. For example, take the category of propositus households of size one, average age 30-44, not predominantly female. (In this particular case we are dealing with individual male propositi between the ages of 30 and 44, incl.) From Table 6, there are 30 such households in the sample. When classified on health problem status, first and second year, these 30 households are distributed as shown in Table 7. Just as for Table 5, the product-moment correllation, R, can be calculated for Table 7. This is,

$$R = (p_2 - p_1)\sqrt{p'q'/pq} = (1/3 - 9/27)\sqrt{(27)(3)/(20)(10)} = 0.$$

However, this measure of serial relation, while being specific for size, age and sex categories, has not been adjusted for the influence of stratum and interviewer characteristics. Furthermore, the sample size, 30, is no longer very large; therefore, the serial relation should also be adjusted for continuity.

Table 7

Four-fold Classification of Propositus Households of Size One, Average Age 30-44, Male, According to Household Health Status at First Interview and Propositus Household Health Status at Second Interview

First Interview	Second Interview		Total No. of Hh.
	No Hlth. Pb. in Prop. Hh.	Hlth. Pb. in Prop. Hh.	
No Hh. Hlth. Pb.	2	1	3
Hh. Hlth. Pb.	18	9	27
Total No. of Hh.	20	10	30

The detailed procedure to adjust Table 7 for the possible spurious influence of stratum and interviewer characteristics, as well as for continuity, is illustrated in 5.9. For present purposes, only the results of such adjustments are presented, Table 8.

Table 8

Four-fold Classification of Propositus Households of Size One, Average Age 30-44, Male, According to Health Status at at First and Second Interview; 'Frequencies' Adjusted for Stratum and Interviewer Characteristics, as well as for Continuity

First Interview	Second Interview		Total No. of Hh.
	No Hlth. Pb. in Prop. Hh.	Hlth. Pb. in Prop. Hh.	
No Hh. Hlth. Pb.	2.13	1.26	3.39
Hh. Hlth. Pb.	17.71	8.90	26.61
Total No. of Hh.	19.84	10.16	30.00

Now the partial serial association, having been adjusted for stratum and interviewer characteristics, and adjusted for continuity, is

$$r = (1.26/3.39 - 8.90/26.61)\sqrt{(3.39)(26.61)/(19.84)(10.16)}$$

$$= -.025, \text{ or } -2.5\%.$$

This compares with an unadjusted R of 0.0%.

Similarly, all the adjusted serial relations are computed for the 22 size-age-sex specific categories in which the frequency of households equals or exceeds 30. These adjusted relations are shown in rank order in Table 9.

Table 9

Rank Ordered Partial Serial Relation (r_α) of Household Health Problems, Adjusted for Stratum and Interviewer Group, and Adjusted for Continuity, for Size-Age-Sex (α) Categories of Propositus Households

α	r_α	Size	Age	Sex	n_α
1	-.322	3	45-54	M	33
2	-.208	4	30-44	M	96
3	-.175	5&up	30-44	M	57
4	-.114	2	55&up	M	179
5	-.109	3	15-29	F	117
6	-.085	5&up	15-29	F	92
7	-.074	3	15-29	M	104
8	-.046	2	45-54	M	97
9	-.043	2	30-44	M	124
10	-.043	5&up	15-29	M	158
11	-.040	2	15-29	M	135
12	-.038	3	30-44	M	99
13	-.025	1	30-44	M	30
14	+.000	1	45-54	F	36
15	+.013	3	45-54	F	42
16	+.029	4	15-29	F	54
17	+.038	3	55&up	F	35
18	+.045	1	55&up	M	59
19	+.052	1	55&up	F	72
20	+.078	3	30-44	F	97
21	+.101	5&up	30-44	F	45
22	+.196	4	15-29	M	143

In section 5 it will be shown that each of such adjusted, or partial, serial relations has an approximately normal sampling distribution, with variance equal to $1/n_\alpha$, the reciprocal of the total frequency in the size-age-sex specific category. Note that these statistical properties are essentially the same as for R, the unadjusted serial relation for large samples.

These properties are applied to the present analysis in the various tests of significance and confidence intervals which ensue.

Now an overall measure of the adjusted serial relation of household health problems in the Arsenal Health District can be obtained. This is done by averaging the 22 size-age-sex specific relations according to the formula:

$$r_{ave} = \sum n_\alpha r_\alpha / \sum n_\alpha \quad ;$$

r_{ave} is a weighted average of the 22 values of r_α, with the n_α as weights. This turns out to be

$$r_{ave} = -.035 = -3.5\%.$$

This overall measure is the adjusted difference between the rate of second year health problems for propositus households, I, with, versus those, II, without, a first year health stress. Conceptually, this means that only households of the same size, the same average age class, the same class of sex distribution, the same stratum and seen by the same class of interviewer, enter into the comparison of health problem rates for the propositus households of groups I and II. This adjusted comparison of rates, -3.5%, is not found to be significant on the 5% level. (For detailed illustration of the computation of r_{ave} and of the test of significance, see 5.10.) It is recalled that the unadjusted relation, R, is -7.7% and is significant on the 0.1% level (2.32). This significant unadjusted relation has been explained as due to an excess of negative selective forces. Since the adjusted relation is less negative, this means that a net negative force has been removed by adjustment for household size, age, sex, stratum and interviewer characteristics. The 95% confidence estimate of the strength of the partial serial relation which existed in the Arsenal population is: -8.0% to +1.1%. (See illustration, 5.10.) Therefore, we can infer that, overall, the remaining unspecified positive and negative forces on the serial relation nearly balanced each other in the Arsenal population.

2.5. First Year Health Problems as a Cause of Second Year Health Problems in the Household

In the introduction, section 1.0, it was pointed out that any causal interpretation of observational material, such as we are dealing with here, is subject to question. But it was also stated that, often, one of the prime purposes of the survey is to test a causal hypothesis and to see how the effects of the supposed cause vary under different circumstances.

In the present analysis, we have adjusted in one way or another for five conditioning factors: household size, average age, sex distribution, stratum and interviewer characteristics. This is probably a mininum of conditioning variables which should be specified and accounted for before entertaining the causal hypothesis that health problems in the first year constitute a stress leading on average to increased second year health problems in the household. There is no argument with those readers who would not wish to entertain a direct causal hypothesis of this nature without specifying and adjusting for other conditioning variables. For example, there might be good reason to adjust for factors such as occupation, race, and health history of propositi. Or again, some readers would wish to classify the health problems by diagnosis before bringing a causal interpretation to any part of the analysis. The first point we are making here is that the framework in which a person is willing to entertain a causal hypothesis is subjective; it varies from person to person. Therefore, using observational data, we would not attempt either to 'prove' or 'disprove' the existence of a cause to the satisfaction of all, or even necessarily to a few, readers.

A second point which we are making is that, most often, for any causal hypothesis to be credible, a large number of conditioning variables must be taken into account in one way or another. The methods developed in subsequent sections of this work provide a beginning for extending the number of conditioning variables taken into account by the analytical procedure. The present analysis illustrates these methods.

Then, as a working hypothesis, we assume that first year health problems can be a cause of second year health problems. We have already seen in 2.42 that there is no significant serial relation in the sample as a whole, when adjusted for the five specified conditioning factors; this carries the interpretation that, if first year health problems are a cause of subsequent health problems in the household, such a relation does not show up in the population as a whole. But we may also test the variation in effects of first year health problems under varying conditions. If significant patterns of variation in these effects, i. e. in the partial serial relations of Table 9, are found, then these can be given a causal interpretation, in the framework which has been chosen. No attempt is made to secure general agreement that first year health problems are directly responsible for such patterns of variation, if they exist. General agreement could only be achieved by a whole series of studies by various investigators, at various times, in various human populations. Rather, the objective of the present analysis is less ambitious: to determine whether patterns of variation in the serial relation are consistent with a causal hypothesis.

The investigation of variations in the partial serial relations under varying conditions, i. e. for varying size, age, and sex categories shown in Table 9, is divided into three parts:

(1) a test of the overall variation of the 22 size-age-sex

specific adjusted serial relations - if this variation is greater than could reasonably be expected by chance alone, then it is likely that something other than chance has caused it; such a finding would be consistent with the hypothesis that health problems, under certain circumstances related to household size, age, and sex, exert a stronger force than under other such circumstances;

(2) a test of a specific *a priori* hypothesis which anticipates a greater force in certain size-age-sex specific categories than in others; this test, if significant against the null alternative, would directly substantiate the hypothesis;

(3) *a posteriori* scrutiny of the variations in serial relation in an attempt to abstract a meaningful, unifying pattern of variation - such a pattern would furnish a new causal hypothesis which could be tested in future surveys; the degree of belief, or credibility, in such an *a posteriori* finding, if one exists, would depend on the 'significance' of the pattern and on the level of residual variation left unexplained by the *a posteriori* hypothesis.

2.51. The Overall Variation of the 22 Size-Age-Sex Specific Adjusted Serial Relations

The 22 serial relations in Table 9 show variation; they range from -.322 to +.196, with a concentration of values about -.040. The question is: is this variation consistent with the level of variation which would be seen if the 22 relations had been randomly selected from a common universe; or is this variation greater than could reasonably be expected on just a chance basis.

A test of significance which answers the above question is readily available. Since each $\sqrt{n_\alpha}\, r_\alpha$ is approximately normally distributed with variance equal to one, the sum of squared deviations of r_α from r_{ave}, each weighted by n_α is

$$\sum n_\alpha r_\alpha{}^2 - nr_{ave}{}^2 \text{ , where } n = \sum n_\alpha \text{ .}$$

On the hypothesis that the population serial relations are the same for all 22 size-age-sex groups, the above expression is distributed approximately as a chi-square with 21 degrees of freedom. In the present case, the computed chi-square value is

$\chi^2 = 22.34 - 2.27 = 20.07$. (See illustration, 5.10, for detailed computations.)

On referring this to a chi-square distribution with 21 degrees of freedom, it is found that an observed value of 20.07 or more would occur slightly more than half the time in repeated random sampling. Thus the observed level of variation, 20.07, is very close to the value, 21.00, which would be expected in random sampling.

On the basis of this test, then, the variations of the serial relations in Table 9 are much the same as would be expected by chance alone. This is one indication that the net resultant of familial, selective and direct causal forces is practically the same, no matter which category of households we may choose. However, in the following sub-section, there remains to be made an *a priori* test of a specific causal hypothesis. If that hypothesized relation should prove significant, then a certain amount of credence can be assigned to a direct causal interpretation, as distinguished from an interpretation based on familial or selective influences. This is so for two reasons:

(1) a significant pattern of variation will have been found through *a priori* direct causal considerations, rather than familial or selective considerations;

(2) because the level of variation is already low, any residual variation left over as unexplained by the causal hypothesis will be non-significant; being such, this would lend substance to the interpretation that the net resultant of familial and selective forces is practically the same in every size-age-sex category.

For it would indeed seem to be a rare coincidence if the pattern of variation due to familial and selective forces were to match so closely the pattern predicated by direct causal considerations.

2.52. Test of the Specific Causal Hypothesis

If the mechanism through which a cause produces an effect is clearly understood, then there is little difficulty in specifying the hypothesized variations of effects under varying circumstances. However, we must admit in the present case that the causal mechanism is not clearly understood; consequently, the choice of a specific hypothesis with respect to variations in effects is difficult. Without doubt the specific hypothesis we make here is not the 'best' which could possibly be made, and is probably somewhat at variance with what another investigator might choose on _a priori_ grounds. Perhaps, in situations like this, it would be helpful to poll several experts on the subject in an attempt to arrive at some mutually agreeable viewpoint. Therefore, it is with some misgivings that the following hypothesis is presented. Nevertheless, it will serve to illustrate the rigorous kind of test which this author believes is necessary to make, if a causal conclusion is to be taken seriously. To meet the test, four criteria should be satisfied:

(1) a fairly large number of varying conditions should be available for analysis, giving effects an opportunity to manifest themselves in various ways;

(2) an _a priori_ hypothesis about the kinds and level of varia-

tion under the varying conditions should be specified;

(3) the a priori hypothesis should account for a significant amount of variation;

(4) the residual variation, after removal of variation 'explained' by the hypothesis, should be low enough to be assigned to chance fluctuation; that is, residual variation should not be significantly large. Failure to meet any one of these four criteria would, this author believes, favor a non-causal interpretation of any association which might be seen.

The stress of non-communicable health problems in mind causes an alteration of some kind in the lives of the propositi. This alteration in activities and responses may or may not result in a health problem for a given propositus. By reason of the stress, some propositi may escape a health problem, while others may acquire one, while still others may escape one problem and acquire another, and the health problem status of still others may be completely unaffected by the stress. Evidently, the mechanism of response is manifold. Consequently it is quite impossible to specify any single mechanism. We can, however, specify those conditions under which we hypothesize a greater average response than under other conditions.

One of the conditions available for analysis is household size. It seems reasonable to expect that, if health problems constitute a stress, this stress should be greater if borne by a single propositus than if by more than one person. Therefore, it is hypothesized that the serial relation should increase in a positive direction as household size decreases.

Secondly, age is a conditioning variable. The stress in mind, it seems reasonable, would not be so great for younger, more flexible propositi than for the older propositi. Then the relation should increase with increasing average age of propositi.

Finally, sex distribution is a conditioning variable. If most of the propositi are females, then the existence of a health problem stress may well cause a greater ghange in the lives of the propositi than if more males were present in the household. For example, in a husband and wife household, if the husband comes down with serious illness, it may be necessary for the wife to change from housewife to breadwinner, and the transition may result in health problems for the wife. Also, the female is often more dependent on other members of the household for her sense of security; if illness in the household disturbs that sense of security, health problems may result.

In order to account for these three hypothesized influences, we assign a value to the various categories of household size, age and sex, in accord with the hypothesis; a high value reflects an hypothesized high serial relation, while a low value reflects a supposed lower serial relation. These are as follows:

Size	Value	Age	Value	Sex	Value
1	4	15-29	0	M	0
2	2	30-44	1	F	4
3	1	45-54	2		
4	0	55&up	4		
5&up	0				

These values represent the relative importance put on each category of the conditioning factors, as well as the relative importance between factors. For a household of given size, age and sex distribution, the appropriate values are added together to yield an index score for the hypothesized level of the serial relation. For example, in propositus households of size one, age 60, female, a total score of 4 + 4 + 4 = 12 is given; and so on. Table 10 shows the index score, together with observed serial relation, for each of the 22 categories being tested.

Table 10

Comparison of Causal Hypothesis (Index Score) with Observed Partial Serial Relation for Size-Age-Sex Specific Groups of Propositus Households

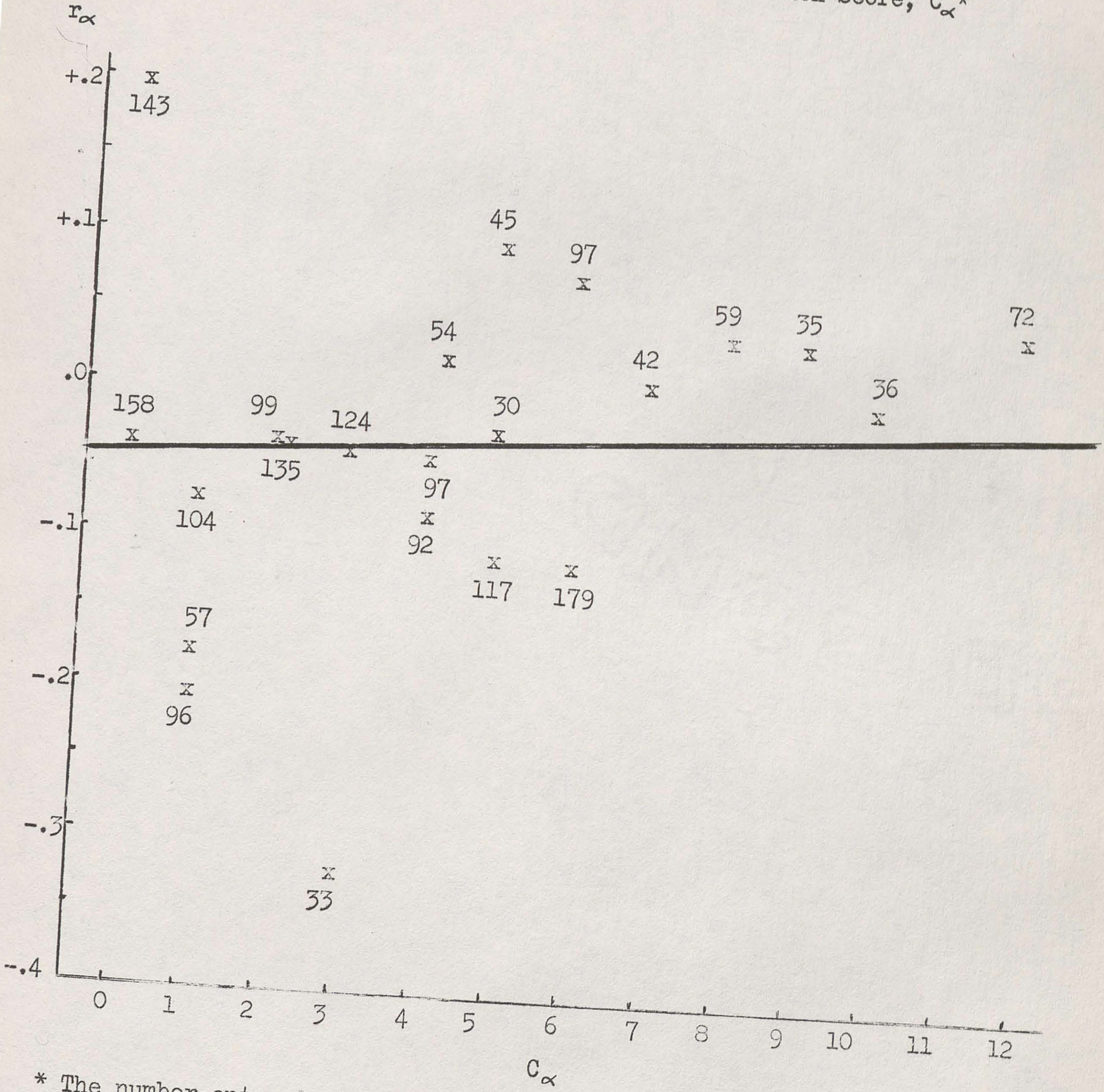
α	Size	Age	Sex	n_α	Index (C_α)	r_α
1	5&up	15-29	M	158	0	-.043
2	4	15-29	M	143	0	+.196
3	4	30-44	M	96	1	-.208
4	5&up	30-44	M	57	1	-.175
5	3	15-29	M	104	1	-.074
6	2	15-29	M	135	2	-.040
7	3	30-44	M	99	2	-.038
8	3	45-54	M	33	3	-.322
9	2	30-44	M	124	3	-.043
10	5&up	15-29	F	92	4	-.085
11	2	45-54	M	97	4	-.046
12	4	15-29	F	54	4	+.029
13	3	15-29	F	117	5	-.109
14	1	30-44	M	30	5	-.025
15	5&up	30-44	F	45	5	+.101
16	2	55&up	M	179	6	-.114
17	3	30-44	F	97	6	+.078
18	3	45-54	F	42	7	+.013
19	1	55&up	M	59	8	+.045
20	3	55&up	F	35	9	+.038
21	1	45-54	F	36	10	+.000
22	1	55&up	F	72	12	+.052

By inspection of Table 10, there appears to be some degree of agreement between hypothesis and observation. For the lower index scores, 5 and below, negative serial relations predominate, while for the index scores higher than 5, positive relations predominate. The most serious exceptions seem to be the +.196 relation observed for an index score of 0, and the very low -.322 observed for a score of 3.

The degree of agreement is more easily grasped from Figure 10, which is a plot of the serial relations, r_α, as a function of the corresponding index scores, C_α. Perfect agreement would be indicated if each of the plotted points fell along a straight line with positive slope. The least-squares regression line plotted on the graph indicates that, on average, the slope is indeed positive, indicating some measure of agreement.

Figure 10

Serial Relation, r_α, as a Function of Index Score, C_α*



* The number entered with each plotted point is n_α; the least-squares straight line fitted to these points takes n_α as weights.

But the scatter of points about the line is relatively great in comparison to the slope of the line. Hence this graphic representation indicates that the agreement of hypothesis with observation is not very strong. By the procedures illustrated in detail in 5.10, we may apply a test of significance to the amount of agreement, as determined by the slope of the regression line in relation to the variability in the observed serial relations. It is found that the slope, +.00333, is far from being significantly great; a slope this great or greater would occur 33 times in a 100 by chance alone. Thus, the hypothesis does not account for a significant amount of variation in the serial relations.

Also by the test illustrated in 5.10, the residual variation (the variation of the serial relation about the line of regression) is well within the realm of chance variation.

How, then, does the causal hypothesis measure up to the four test criteria specified above? Taken one by one:

(1) effects have an opportunity to manifest themselves differently in 22 sub-groups of the population, varying with respect to household size, average age, and sex distribution; then the first criterion is met;

(2) the second criterion is met, since an a priori causal hypothesis has been constructed;

(3) the third criterion is not met, for the hypothesis has not accounted for a significant amount of variation in effects;

(4) the residual variation can be assigned to chance fluctuations; thus the fourth criterion is met.

Failure to meet the third criterion, then, enables us to conclude that a causal relation between household health problems in the first year and the second year has not been substantiated. Failure to meet any one of the four criteria would have been interpreted in a similar way.

2.53. A Posteriori Analysis of the Partial Serial Relation

The truth of an a posteriori relation discovered by the analyst may be seriously questioned; those relations which lack all form or pattern are most vulnerable in this respect. Thus, if the highest serial correlation in Table 9, +.196, is compared with the lowest, -.322, the difference between them is 'significant'; but this is no basis whatever for concluding that there was a difference between the serial relations for the two population categories to which these measures apply. For in any large set of measures from a single population, such 'significant' differences would almost always be found; and they would be meaningless. Further, the existence of such differences in measurements is hardly an adequate basis for the formulation of hypotheses for future testing. Rather, a degree of credibility can be assigned to a posteriori findings only when observed variations are consistent with some unifying rule. There seems to be no single method which can be applied to find such patterns in sets of data; it seems that this is more a matter of trial and error. In the present case, several attempts to find a pattern of some kind among the size-age-sex categories were made. Rather than burden the reader with an account of all these attempts to 'explain' the variations of the data on an a posteriori basis, it should suffice to present the one 'significant' pattern which has been found.

The one pattern found to be of significance is based on household sex distribution. In Table 9, it is seen that only two of the 13 negative serial relations are for predominantly female households, while seven of the nine positive relations are for predominantly female households. An inspection of the table shows that the two female groups which show negative correlations are in the very young category, age 15-29. It can be assumed that several of the households in these two groups contain very young children and, therefore, that female children happen to predominate. Also, among the positive correlations, there are only two male categories. One of these, which happens to be the highest positive correlation, is for a household in the 15-29 age category. Now if it is argued that the stress, if any, of health problems is not differentially active in the young males and females, but only in older persons, then a comparison on sex distribution would be more sensitive if the very young households were omitted. When such comparison is made, the average relation for male households, of age groups older than 15-29, is -.094. For predominantly female households, older than 15-29, the average relation is +.054. The difference, .148, is found to be significant on the 3% two-tailed level. (See illustration, 5.10.) Furthermore, the residual variation is neither significantly high nor significantly low.

How shall we interpret this a posteriori finding? We were unable to formulate, a priori, the correct causal hypothesis; that is, the relation has not been put to the test as a predictor. Then, certainly, to apply a causal interpretation would be premature. But even an interpretation of the finding as a measure of true association in the population, with no attempt to discriminate between familial, selective, or direct causal forces, is not valid. For, as has been previously pointed out, in almost any series of observations a 'significant' relation can be found if we only look long enough for one. However, the fact remains that of the 9 groups of adult male households 8 showed negative serial relations; and of the 6 groups of adult female households, all 6 showed positive serial relations. This is a fact which ought not to be dismissed lightly. For to do so would be to throw away a possibly rewarding hypothesis.

While we cannot dignify the a posteriori finding as a conclusion, we can offer this finding as an a priori hypothesis in future studies. Strictly, this would not constitute a causal hypothesis as it stands. For the finding may well have been due to familial and selective influences. However, the finding is in general agreement with causal considerations. Then it would be an adequate causal hypothesis if it were used in future studies as a part of an hypothesis which would predict variations under new varying conditions (such as varying diagnosis, varying health history of propositi, etc.).

But, as for now, we cannot discriminate between a causal or merely associative finding. Moreover, as for now, we cannot discriminate between whether or not this finding is a chance event. All we can say is that, in view of the a posteriori significance of the finding, and in view of the fact that residual variation left unexplained by the finding is neither too high nor too low to be ascribed to chance, it is a plausible hypothesis for future study.

2.6. Conclusion

It has been shown that the serial, or one-year time-lagged, relation is determined not only by familial forces but also selective and direct causal forces. In the Arsenal Health District Survey of 1951-1952, the net resultant of these forces was negative. That is, a significantly greater proportion of households without first year health problems, as compared to those with first year health problems, acquired new health problems in the second year.

But when households comparable on the basis of size, average age, sex distribution, stratum, and type of interviewer were compared, the serial relation was found not to be significantly negative. Thus, the above specified conditioning factors account for the significantly negative overall serial relation.

Variations in the serial relation among size, age and sex groups was no more than would be expected of a set of randomly chosen relations. Furthermore, a specific causal hypothesis with respect to variations among these groups was not substantiated. Therefore, a cause and effect relation, as distinguished from associative familial and selective relations, was not demonstrated. Nor was any association between the serial relation and household size, age or sex distribution demonstrated to exist in the Arsenal District.

However, on an _a posteriori_ basis, it was found that all the predominantly female adult households (average age over 29) showed a positive serial relation, while all but one of the remaining adult households showed a negative serial relation. This latter finding is not a conclusion, but rather points to the future investigation of a new causal hypothesis: that adult females, or adult households containing a majority of females, are particularly susceptible to ill effects from the non-communicable health problems of other household members.

This brings to an end the present study of the serial relation of health problems in households. But there remains to be shown that the various measures in this illustrative analysis do have the statistical properties which have been applied to the analysis. This more abstract development is taken up in subsequent sections.

3.0. TESTS OF ASSOCIATION IN A TWO-FACTOR CONTINGENCY TABLE

The two-factor (g x h) contingency table is considered in this section. From generalized measures of association, we proceed to discuss the partitioning of chi-square when one of the factors is quantified and the further partitioning when both factors are quantified. Also, proceeding from the general case of the (g x h) table, the particular cases of the (g x 2) table and the (2 x 2) table are discussed in detail in order to illustrate the meaning of the general approach in terms of these frequently occurring situations.

3.1. Definition of Symbols

The general two-factor contingency table is presented in Table 11.

Table 11

The (g x h) Contingency Table

		Y				
	j \ i	1	2	. .	h	$n_{i.}$
X	1	n_{11}	n_{12}	. .	n_{1h}	$n_{1.}$
	2	n_{21}	n_{22}	. .	n_{2h}	$n_{2.}$
	.	.	.	. .	.	.
	.	.	.	. .	.	.
	g	n_{g1}	n_{g2}	. .	n_{gh}	$n_{g.}$
	$n_{.j}$	$n_{.1}$	$n_{.2}$	. .	$n_{.h}$	n

Let the first factor, X, be considered arbitrarily as the independent factor, and let the second factor, Y, be dependent on X. X is divided into g categories and Y into h categories, denoted by i and j, respectively. For a sample of size n from a universe, the frequency of observations in any cell of the contingency table is denoted by n_{ij}. Marginal frequencies are denoted by $n_{i.}$ and $n_{.j}$ for rows and columns, respectively.

If the n observations have been obtained randomly from a universe in which no association exists between X and Y, the expected value of any n_{ij} is

$E(n_{ij}) = (n_{i.} n_{.j})/n$, where $n_{i.}$ and $n_{.j}$ are regarded as being fixed. Due to sampling fluctuations, the n_{ij} will deviate from their expected values by

$$d_{ij} = n_{ij} - (n_{i.} n_{.j})/n$$

The quantity,

$\sum\sum n d_{ij}^2/(n_{i.} n_{.j})$, is distributed approximately as χ^2 with $(g - 1) \times (h - 1)$ degrees of freedom (henceforth sometimes denoted as $\chi^2_{(g - 1) \times (h - 1)}$), provided the cell expectations are not too small.[15]

3.2. Association when Neither Factor is Quantified

Karl Pearson defined $\phi^2 = \chi^2/n$ (using χ^2 to mean the computed quantity) as the mean square contingency.[16] As a measure of divergence from independence, Pearson proposed that $C = \sqrt{\phi^2/(1 + \phi^2)}$ be used.[16] Several other measures of association based on deviations from independence have been proposed, but none seems to be a completely satisfactory estimate of association in the population.[16] Even if a satisfactory estimate were

[15] See Cochran, W. G., "Some methods for strengthening the common χ^2 tests", <u>Biometrics</u>, V10, pp. 417-451, 1954, for a discussion of frequency requirements in a contingency table for good approximations to chi-square.

[16] Kendall, M. G., "The Advanced Theory of Statistics", V1, ch. 13, Charles Griffin and Co., Ltd., 4th ed., London, 1948.

available, the meaning of such a measure would not be very specific. For departures from independence of the two factors can be of many kinds, and such measures which lump together all the departures from expectations fail to discriminate among the various kinds of departure.

The catch-all nature of $\chi^2_{(g - 1) \times (h - 1)}$ was cause for the development of methods by which it could be partitioned into more meaningful component parts. Lancaster, in 1949, showed that $\chi^2_{(g - 1) \times (h - 1)}$ can be partitioned into (g - 1) x (h - 1) different component chi-square values, each with 1 degree of freedom. Each such value corresponds to a (2 x 2) table, and each is asymptotically independent of the others as the sample size becomes large.[17] These component values could be combined to test for specific relations within the contingency table.

Williams[18] generalized Lancaster's methods to show a partitioning of $\chi^2_{(g - 1) \times (h - 1)}$ into (h - 1) component parts, each of (g - 1) degrees of freedom. The (h - 1) canonical correlations which can be determined from the contingency table correspond to such a partitioning. More specifically, if the (h - 1) canonical correlations are designated by R_k, k = 1, 2, .., (h - 1), then each nR_k^2 is a χ^2 with (g - 1) degrees of freedom, and

$$\sum nR_k^2 = \chi^2_{(g - 1) \times (h - 1)};$$

that is, the sum of the squared correlations equals ϕ^2, as was shown by Maung.[19] To each canonical correlation, there corresponds a unique set of scores for

[17] Lancaster, H. O., "The derivation and partition of χ^2 in certain discrete distributions", Biometrika, V36, pp. 117-129, 1949.

[18] Williams, E. J., "Use of scores for the analysis of association in contingency tables", Biometrika, V39, pp. 274-289, 1952.

[19] Maung, K., "Measurement of association in a contingency table with special reference to the pigmentation of hair and eye colour of Scottish school children", Annals of Eugenics, V11, pp. 189-223, 1941.

the categories of X and of Y. The set of scores which corresponds to the largest of the correlations, R_{max}, would be chosen in practice as the best numerical values to place on the categories of the two factors. Unfortunately, Williams provides the reader with no interpretation of such scores. We can imagine, however, certain practical uses to which such scores might be put:

(1) if _a priori_ numerical characters could be assigned to the categories of the two factors, then these _a priori_ scores might be compared with the m.c. (maximum-correlation) scores; in this way, the _a priori_ scores could be evaluated as to their adequacy in describing a linear association;

(2) if data pertaining to the relation between two factors of qualitative character were available for a wide variety of circumstances, the m.c. scores developed from such data might be used to establish quantification of the factors with respect to their mutual relationship; thus, in the future, the essentially qualitative factors might take on quantitative aspects. But these considerations are beyond the scope of this presentation. They are mentioned as possible subjects for investigation, because of their fascinating implications as to the meaning of numbers applied to classes of things, and because they are generalizations of the methods to be discussed.

3.3. Association when the Dependent Factor is Quantified

When each of the categories of the dependent factor, Y, is characterized by a numerical quantity, y'_j, then a set of scores, x_i, can be determined for the independent categories such that the squared product moment correlation between x_i and y'_j is a maximum. As will be shown, each of the m.c. scores for X can be interpreted as the observed 'effect' of the independent category on the dependent variable. Further, the squared correlation times n is distributed asymptotically as χ^2 with (g - 1) degrees of freedom, assuming no association between X and Y in the universe from which the contingency table is derived.

The method for determining the m.c. scores is developed as follows:

let $y_j = (y'_j - \bar{y}')/s_{y'}$, where $\bar{y}'$ and $s_{y'}$ are the mean and standard deviation, resp., of y'; thus, $\bar{y} = 0$ and $s_y = 1$, i. e., y is scaled to a zero mean and unit variance; a set of scores, x_i, also scaled to a zero mean and unit variance, is to be determined such that the squared correlation, R^2, between x_i and y_j is a maximum; the correlation between x_i and y_j is, by definition,

$$R = \sum\sum n_{ij} x_i y_j / n \quad ; \tag{1}$$

then we may solve for x_i by maximizing the expression,

$n^2R^2 = (\sum\sum n_{ij} x_i y_j)^2$, subject to the restrictions that

$\sum n_{i.} x_i = 0$, and

$\sum n_{i.} x_i^2 = n$, i. e., that x_i have zero mean and unit variance; using LaGrange multipliers, we maximize the expression,

$$(\sum\sum n_{ij} x_i y_j)^2 - 2L_1 \sum n_{i.} x_i - L_2 \sum n_{i.} x_i^2 \quad ;$$

taking derivatives with respect to the x_i and setting equal to zero,

$$2(\sum\sum n_{ij} x_i y_j)(\sum_j n_{ij} y_j) - 2L_1 n_{i.} - 2L_2 n_{i.} x_i = 0 \quad ; \tag{2}$$

summing with respect to i,

$$2(\sum\sum n_{ij} x_i y_j)(\sum\sum n_{ij} y_j) - 2L_1 n - 2L_2 \sum n_{i.} x_i = 0 \quad ;$$

but,

$\sum\sum n_{ij} y_j = \sum n_{.j} y_j = 0$, and

$\sum n_{i.} x_i = 0$, so

$2L_1 n = 0$, and therefore $L_1 = 0$;

then (2) simplifies to

$$(\sum\sum n_{ij} x_i y_j)(\sum_j n_{ij} y_j) - L_2 n_{i.} x_i = 0 \quad ; \tag{3}$$

multiplying (3) by x_i and summing with respect to i,

$$(\sum\sum n_{ij}x_iy_j)(\sum\sum n_{ij}x_iy_j) - L_2\sum n_{i.}x_i^2 = 0 \quad ;$$

that is,

$$n^2R^2 - L_2n = 0 \text{ , so that}$$

$$L_2 = nR^2 \quad ; \tag{4}$$

substituting (4) in (3), we have

$$nR(\sum_j n_{ij}y_j) - nR^2n_{i.}x_i = 0 \text{ , and solving for } x_i,$$

$$x_i = (1/R)(\sum_j n_{ij}y_j/n_{i.}) = \bar{y}_i/R \quad ; \tag{5}$$

thus, the x_i are proportional to the means of the dependent variable within the respective categories of X.

It remains to determine the value of R (arbitrarily taken as always being positive). Squaring both sides of (5) and multiplying by $n_{i.}$,

$$n_{i.}x_i^2 = (1/R^2)n_{i.}\bar{y}_i^2 \quad ;$$

summing with respect to i,

$$\sum n_{i.}x_i^2 = (1/R^2)\sum n_{i.}\bar{y}_i^2 \text{ , that is,}$$

$$R^2 = \sum n_{i.}\bar{y}_i^2/\,n = s_{\bar{y}_i}^2 \quad . \tag{6}$$

One cannot fail to note that equations (5) and (6) are equivalent to those used in the analysis of variance. Since $\bar{y} = 0$, $\bar{y}_i$ measures the deviation of the average for the ith group from the grand average. This, in the experimental setting, is termed the observed effect of the ith treatment. Thus, the x_i of equation (5) are proportional to the 'effects' of the categories of X; they differ from the usual measure of effects only in scale. Also, from equation (6), R^2 is merely the observed variance of the means.

Consequently, as Williams states, an F test could be performed to test for the significance of the association between the categories of X and the values of Y.

Alternatively for large n, since nR^2 is distributed as a chi-square with (g - 1) degrees of freedom, a chi-square test of association is appropriate. For future purposes, the χ^2 test will be adopted because it is simpler to perform.

3.4. Association when Quantities are Assigned to Both Factors

In accord with Williams' demonstration of the partitioning of χ^2, the $\chi^2_{(g - 1)}$ of the previous sub-section, determined when Y is quantified, can be partitioned into parts. In particular, if an arbitrary set of scores be assigned to the categories of X, then n times the squared correlation between the X scores and the Y scores corresponds to a χ^2 with one degree of freedom.

3.5. Particular Case: Association when the Dependent Factor is a Dichotomy

In the particular case of a (g x 2) contingency table, the overall chi-square has (g - 1) x (h - 1) = (g - 1) degrees of freedom. Further, there are (h - 1) = 1 canonical correlations with (g - 1) degrees of freedom. As in the general case, if any arbitrary pair of different numbers is assigned to the two categories of Y, then the maximum squared correlation between x_i and y_j corresponds to a χ^2 with (g - 1) degrees of freedom. Thus the overall χ^2, n times the squared canonical correlation, and n times the squared maximum correlation with the assigned y_j quantities are all identical. In other words, there is no loss in generality if the analysis is performed by assigning arbitrary numbers to the dependent dichotomy.

Equations (5) and (6), sub-section 3.3, give the values for x_i and for R^2, as follows:

$$x_i = \bar{y}_i/R$$

$$R^2 = \sum n_{i.}\bar{y}_i^{\,2}/n = s_{\bar{y}_i}^{\,2} \; .$$

However, when dealing with a dependent dichotomy, one customarily expresses relations in terms of rates, or proportions. Therefore, with reference to the (g x 2) contingency table below, the usual symbols for proportions will be defined, and equations for x_i and R^2 will be re-stated in terms of proportions.

Table 12

The (g x 2) Contingency Table

		Y		
	i \ j	1	2	$n_{i.}$
X	1	n_{11}	n_{12}	$n_{1.}$
	2	n_{21}	n_{22}	$n_{2.}$
	.	.	.	.
	.	.	.	.
	g	n_{g1}	n_{g2}	$n_{g.}$
	$n_{.j}$	$n_{.1}$	$n_{.2}$	n

Let the first class of Y be the condition of not having a stated quality and the second class as having the stated quality, such as death, having a disease, etc. Let $p = n_{.2}/n$ be the proportion of all sampled elements which have the quality. And let $p_i = n_{i2}/n_{i.}$ be the proportion observed to have the quality in category i of X. Similarly, q and q_i are the proportions not having the quality, so that

$$p + q = 1, \text{ and}$$

$$p_i + q_i = 1 \text{ for all } i.$$

Let values of y_1 and y_2 be chosen such that the mean of y_1 and y_2 is zero and their variance is unity. That is,

$$n_{.1}y_1 + n_{.2}y_2 = 0$$

$$n_{.1}y_{.}^2 + n_{.2}y_2^2 = n \quad .$$

Solving the above for y_1 and y_2,

$$y_1 = \pm\sqrt{n_{.2}/n_{.1}} = \pm\sqrt{p/q}, \text{ and}$$

$$y_2 = \mp\sqrt{n_{.1}/n_{.2}} = \mp\sqrt{q/p} \quad .$$

We arbitrarily associate a negative value with the condition of not having the quality and a positive value with having the quality, hence,

$$y_1 = -\sqrt{p/q}$$

$$y_2 = +\sqrt{q/p} \quad .$$

Now, $\bar{y}_i = (n_{i1}y_1 + n_{i2}y_2)/n_{i.} = p_iy_2 + q_iy_1$;

substituting $(1 - p_i)$ for q_i and simplifying, we get

$$\bar{y}_i = (p_i - p)/\sqrt{pq} \quad ; \text{ then,}$$

$$x_i = (p_i - p)/R\sqrt{pq} \quad . \qquad (7)$$

That is to say, the effect, x_i, is proportional to the deviation of the ith category proportion from the overall proportion. Again, except for scale, x_i is identical to the usual concept of an effect: the amount by which the proportion having a specified quality in a sub-class deviates from the corresponding pooled proportion for all the classes.

By direct substitution,

$$R^2 = \sum n_{i.}\bar{y}_i^{\,2}/n = (1/pq)\sum n_{i.}(p_{i.} - p)^2/n = s_{p_i}^{\,2}/pq \quad , \qquad (8)$$

is proportional to the variance in the observed proportions.

It is also noted that, since $\chi^2_{(g-1)} = nR^2$, $R^2 = \phi^2$, the mean square contingency for the (g x 2) table. Consequently, for a (g x 2) table, the mean square contingency is clearly interpreted as being proportional to the variance in the observed proportions. It is of further interest to note that the proportionality factor, $1/pq$, puts the variance of observed proportions in relation to the expected level of variation on the basis of chance alone. Thus, the χ^2 test of nR^2 is a test of whether the variance in observed proportions is significantly greater than would be expected on the null hypothesis.

3.6. Particular Case: Association in the (2 x 2) Table; Illustration

For the (2 x 2) table, $nR^2 = \chi^2_{(g-1)} = \chi^2_1$. But n times the squared product moment correlation is also χ^2_1 ; consequently, R is identical to the product moment correlation in the (2 x 2) table. Also, we have from equation (8),

$$nR^2 = (1/pq)\sum n_{i.}(p_i - p)^2 \text{ ; in a (2 x 2) table,}$$

$$p = (n_{1.}p_1 + n_{2.}p_2)/n \text{ ; thus,}$$

$$p_1 - p = p_1 - (n_{1.}p_1 + n_{2.}p_2)/n = n_{2.}(p_1 - p_2)/n \text{ ;}$$

Similarly,

$$p_2 - p = n_{1.}(p_2 - p_1)/n \text{ .}$$

Then,

$$nR^2 = (1/pq)\left[(n_{1.}n_{2.}{}^2/n^2)(p_1 - p_2)^2 + (n_{2.}n_{1.}{}^2/n^2)(p_2 - p_1)^2\right].$$

Letting $p' = n_{2.}/n$ and $q' = n_{1.}/n$, the above expression reduces to

$$nR^2 = n(p_2 - p_1)^2(p'q'/pq) \text{ , which in turn equals } \chi^2_1 \text{ or} \qquad (9)$$

z^2, the square of a standard normal deviate.

Thus, the z test of a difference between two proportions, the test for significance of χ^2, and the χ^2 test of the significance of the product moment correlation are all equivalent for the (2 x 2) table.

Dividing (9) by n and taking the square root, we get

$$R = (p_2 - p_1)\sqrt{p'q'/pq} \text{ .} \qquad (10)$$

This is the measure of relation which was adopted and interpreted in 2.32. Since nR^2 is the square of a normal deviate with unit variance, $\sqrt{n}R$ is a normal deviate with unit variance, and R, the adopted measure of simple relation in section 2, is a normal deviate with variance equal to 1/n. These statistical properties of R, when n is large, were used in 2.32 to test the significance of the observed relation computed from Table 5.

Some of the identities developed above can be illustrated by use of the (2 x 2) table presented in section 2.32, Table 5. This is the fourfold classification of households according to household health status at first and second interview (see Table 5 for full description):

i \ j	1	2	
1	600	882	1482
2	321	339	660
	921	1221	2142

To compute χ_1^2 according to definition (see 3.1), we take

$$\chi_1^2 = \sum\sum n d_{ij}^2/n_{i.}n_{.j} = 2142(37.2^2)/(921)(1482) + 2142(37.2^2)/(1221)(1482)$$

$$+ 2142(37.2^2)/(921)(660) + 2142(37.2^2)/(1221)(660)$$

$$= 12.36 \; .$$

Then to compute R^2,

$R^2 = \chi^2/n = 12.36/2142 = .00577$, and finally,

$R = -.076$ (negative sign taken arbitrarily in view of the subject matter of Table 5) .

There are numerous alternative ways to calculate the above measures, such as:

$$\chi_1^2 = 2142(600 \cdot 339 - 321 \cdot 882)^2/ \, 921 \cdot 1221 \cdot 1482 \cdot 660 = 12.38,$$

which agrees with the first method, except for rounding.

Another alternative is to assign arbitrary scores of 0 and 1 to the categories of i and of j, as follows,

	0	1	
0	600	882	1482
1	321	339	660
	921	1221	2142

,

and compute the correlation between the two sets of scores. This computation is the product moment correlation and reduces to,

$$R = (339 \cdot 2142 - 660 \cdot 1221)/\sqrt{921 \cdot 1221 \cdot 1482 \cdot 660} = -.076, \text{ as above.}$$

From this result, χ^2 can be determined, of course, by

$$\chi^2 = nR^2 = 12.38, \text{ also as found above.}$$

Or, again, R can be computed directly from the identity, (10), which is:

$$R = (p_2 - p_1)\sqrt{p'q'/pq} = (p_2 - p_1)\sqrt{n_{1.}n_{2.}/n_{.1}n_{.2}}$$
$$= (.514 - .596)\sqrt{(1482/921)(660/1221)} = (.514 - .596)\sqrt{1.61 \cdot .541}$$
$$= -.082(.933) = -.077, \text{ which again agrees with prior computed}$$

results, except for rounding.

This (2 x 2) table can also be used to illustrate identity (6), $R^2 = s_{\overline{y}_i}^{\ 2}$, which says that the variance of means is equal to the squared product moment correlation. First we find y_1 and y_2:

$$y_1 = -\sqrt{p/q} = -\sqrt{n_{.2}/n_{.1}} = -\sqrt{1221/921} = -1.15;$$
$$y_2 = \sqrt{q/p} = \sqrt{n_{.1}/n_{.2}} = \sqrt{921/1221} = .87 \quad .$$

Then,

$$\overline{y}_1 = (600(-1.15) + 882(.87))/1482 = .0520 \text{ , and}$$
$$\overline{y}_2 = (321(-1.15) + 339(.87))/660 = -.1121 \quad .$$

Note that the mean of $\overline{y}_1$ and $\overline{y}_2$ is

$$(1482(.0520) + 660(-.1121))/2142 = 0 \quad .$$

Therefore the variance of $\overline{y}_1$ and $\overline{y}_2$ is:

$$(1482(.0520)^2 + 660(-.1121)^2)/2142 = .00574 \quad .$$

Taking R = -.076, as computed before, and squaring, we get $R^2 = .00578$, which agrees with the variance of $\overline{y}_i$, above, except for rounding.

3.7. Summary

If the dependent classification can be characterized by a set of quantities, y_j, $j = 1, \ldots, h$, with zero mean and unit variance, then a set of scores for the categories of the independent factor may be determined by means of the formula,

$x_i = \overline{y}_i/R$, $\quad i = 1, \ldots, g$, where x_i are the scores for the independent categories, $\overline{y}_i$ are the mean values of the dependent variable for corresponding categories of the independent variable, and R is the product moment correlation between the x_i and the y_j.

These scores are such that R^2 is a maximum. The term, nR^2, is distributed as a χ^2 with $(h - 1)$ degrees of freedom when no association exists in the population from which the observations are assumed to be derived.

If a set of arbitrary scores is assigned to the independent categories and the product moment correlation, r, between these scores and the dependent variable is compared, the nr^2 is distributed as χ^2 with one degree of freedom.

In the particular case of a dependent dichotomy, quantities can always be applied to the two dependent classes with no loss of generality.

The scores, x_i, being proportional to the deviations of the respective class averages from the grand mean of the dependent variable, are analogous to 'effects' in experiments. Finally, the variance of these 'effects' is equal to R^2, so that R, as a measure of relation, increases with increasing observed differences of the dependent variable among the classes of the independent variable.

For dependent dichotomies, these results may be expressed in terms of rates, and in these terms, R^2 reduces to familiar expressions closely related to those often used for testing differences in rates and for measuring degree of association.

In the following sections, these results will be extended to a contingency table having multiple independent factors.

4.0. PARTIAL ASSOCIATION

Tests of association in a (g x h) contingency table, as discussed in section 3, apply to the single relation, termed the simple association, which exists between two factors, X and Y, in the sampled population. Partial association, on the other hand, refers to the simple association between two factors in various sub-classes of the population. Thus, the term, partial association, implies that there are more than two factors involved: two factors between which some simple relation exists in each sub-class, and one or more other factors used as the basis for sub-classification. Examples of partial association are innumerable: the relation of sex to incidence of death within specified age groups; the relation of type of housing to communicable disease incidence, within specified income groups; the incidence of yellow fever, by geographic location, within classes defined by calendar time; the occurrence of automobile accidents, by time of day, within classes defined by type of road, weather conditions, and calendar time; the occurrence of heart deaths by history of salt intake, within classes defined by age, race, sex, weight, and within classes of dietary factors other than salt; and so on. Invariably, the study of a simple association in a population is a prelude to partial association, and partial association with respect to one set of classes leads to partial association with respect to other sets of classes; for the mind wants to know why a relation between two variables exists, and there is no explanation except in terms of other factors.

4.1. Definition of Symbols for the Multiple Factor Contingency Table

In order to cope with the additional factors of classification inherent in the study of partial association, additional symbols are required.

Table 13

The (g x h) Contingency Table,

Where g = a x b _ x c

Factor										$n_{i.} =$ $n_{kl\_m.}$
X	U	V	_	V'	W	Y				
Index										
i	k	l	_	l'	m	j				
						1	2	...	h	
1					1	n_{11_11}	n_{11_12}		n_{11_1h}	$n_{11\_1.}$
2				1	:	:	:		:	:
:					c	n_{11_c1}	n_{11_c2}		n_{11_ch}	$n_{11\_c.}$
:		1	_	:	:	:	:		:	:
:					1	:	:		:	:
:				b'	:	:	:		:	:
:					c	:	:		:	:
:	1	:		:	:	:	:		:	:
:					1	:	:	$(n_{ij} =$	:	:
:				1	:	:	:		:	:
:					c	:	:	n_{kl_mj}	:	:
:		b	_	:	:	:	:	in each	:	:
:					1	:	:	corresponding	:	:
:				b'	:	:	:	cell)	:	:
:					c	:	:		:	:
:	:	:		:	:	:	:		:	:
:					1	:	:		:	:
:				1	:	:	:		:	:
:					c	:	:		:	:
:		1	_	:	:	:	:		:	:
:					1	:	:		:	:
:				b'	:	:	:		:	:
:					c	:	:		:	:
:	a	:		:	:	:	:		:	:
:					1	:	:		:	:
:				1	:	:	:		:	:
:					c	:	:		:	:
:		b	_	:	:	:	:		:	:
:					1	n_{ab_11}	n_{ab_12}		n_{ab_1h}	$n_{ab\_1.}$
:				b'	:	:	:		:	:
g					c	n_{ab_c1}	n_{ab_c2}		n_{ab_ch}	$n_{ab\_c.}$
$n_{.j} = n_{.._.j}$						$n_{.._.1}$	$n_{.._.2}$	......	$n_{.._.h}$	n

Therefore, the general multiple independent factor contingency table is presented in Table 13. Despite its clumsy appearance, if the factors, U, V, _, V', and W are ignored, Table 13 is identical to Table 11, section 3, being a (g x h) contingency table. However, factor X is now considered to be a complex of the multiple factors, U, V, _, V', and W. Factor U is composed of categories k = 1, ..., a; similarly for V, l = 1, ..., b; for V', l' = 1, ..., b', and for W, m = 1, ..., c. Within each category of U, there is a full set of the categories of V, and so on up to the second-last factor, V', within each category of which there is a full set of the categories of W. Thus, in all, there are g = a x b _ x c categories of the independent factors. For each of these categories, there are h categories of Y, for a total of a x b _ x c x h cells in which frequencies may fall. The frequency within each cell, previously identified as n_{ij}, is now identified as n_{kl_mj}. The dot notation, which was previously applied to indicate a marginal total, continues to apply. Thus, for example, $n_{.l\_mj}$ is the total frequency in the l_mj marginal cell, $n_{..\_.j}$ is the total frequency in the jth category of Y, and $n_{kl\_m.}$ is the total frequency in the kl_m category of the independent factors. Finally, n denotes the size of the sample.

The multiple factor table is more easily visualized if a particular example is taken. Table 14 is a particular case for which there are three independent factors, taken from the subject-matter of section 2: stratum, interviewer group, and household health problem status at 1st interview. These are denoted by U, V and W, respectively. The dependent factor, Y, is the health problem status of households at second interview. There are three strata; thus, a = 3. Similarly, b = 2, c = 2, and h = 2. The factor, X, is the composite of stratum-interviewer group-1st interview health status,

Table 14

Household Health Status at 2nd Interview
in a Sample of the Arsenal Health District of Pittsburgh, Pa.,
1951-1952, by Stratum, Interviewer Group
and Health Problem Status at 1st Interview

(X) (i)	(U) Stratum (k)	(V) Inter-viewer Group (l)	(W) Health Status 1st Int. (m)	(Y) Health Status, 2nd Interview (j)		$(n_{klm.})$ Total Households
				1	2	
(1)	I	(1) AB	(1) No H.P.	n_{1111}	n_{1112}	$n_{111.}$
(2)			(2) H.P.	n_{1121}	n_{1122}	$n_{112.}$
(3)		(2) CD	(1) No H.P.	n_{1211}	n_{1212}	$n_{121.}$
(4)			(2) H.P.	n_{1221}	n_{1222}	$n_{122.}$
(5)	II	(1) AB	(1) No H.P.	n_{2111}	n_{2112}	$n_{211.}$
(6)			(2) H.P.	n_{2121}	n_{2122}	$n_{212.}$
(7)		(2) CD	(1) No H.P.	n_{2211}	n_{2212}	$n_{221.}$
(8)			(2) H.P.	n_{2221}	n_{2222}	$n_{222.}$
(9)	III	(1) AB	(1) No H.P.	n_{3111}	n_{3112}	$n_{311.}$
(10)			(2) H.P.	n_{3121}	n_{3122}	$n_{312.}$
(11)		(2) CD	(1) No H.P.	n_{3211}	n_{3212}	$n_{321.}$
(12)			(2) H.P.	n_{3221}	n_{3222}	$n_{322.}$
$(n_{...j})$ Total Households				$n_{...1}$	$n_{...2}$	n

and has g = 3 x 2 x 2 = 12 categories. The n_{klmj} indicate the particular frequencies which fall in the corresponding cells, and marginal frequencies are denoted by the dot notation.

4.2. Failure of χ^2 to be Closely Approximated when g is Large

For a given sample size, n, if g becomes large relative to n, then the χ^2 approximations discussed in section 3 become very poor. Suppose, as an extreme example, that g = n, and that at least one observation falls in each category of X. Then only one observation falls in each of the g categories of X, since $g = n$. It will be found that the maximized squared correlation between x_i and y_j is $R^2 = 1$. If the formula,

$$\chi^2_{(g - 1)} = nR^2 \text{ , were applied, then}$$
$$\chi^2_{(n - 1)} = n \quad .$$

Thus, the computed χ^2 would be equal to n, with $(n - 1)$ 'degrees of freedom'. But the expected value of χ^2 with $(n - 1)$ degrees of freedom is $(n - 1)$. Consequently, the difference between the computed χ^2 and the expected value would be negligible, and no matter how large n might be, the computed χ^2 could never be judged 'significant'. As a less extreme example, suppose that $n = 40$ and $g = 30$. The largest possible value of the computed χ^2 (for quantified categories of Y, or for Y as a dichotomy) is achieved when $R^2 = 1$, that is,

$$\chi^2 = nR^2 = n = 40.$$

The 5% significance level of χ^2 with $(g - 1) = 29$ degrees of freedom is 42.6. Thus a true χ^2 exceeds this value 5% of the time. But the computed χ^2 can never exceed 40. Consequently, no matter how great an association exists in a population, a sample of size 40, with 30 categories of X could never show significance by this 'test'.

As a traditional rule of thumb, it is often stated that χ^2 is sufficiently well approximated when no expected frequency in the table is less than 5. Cochran[20] liberalizes this requirement, allowing a few of the expectations to be as low as 1. But by either set of rules, the sample size must be considerably greater than the number of categories.

Now in reference to Table 13, if there are either a large number of independent factors, or a large number of categories within the independent

[20] Op. cit., p. 57.

factors, or both, $g = a \times b \_ \times c$ is very large relative to the sample size, and the use of the χ^2 approximations developed in section 3 are invalid unless they are revised. In the remainder of this section, the principle of m.c. scores will be applied to obtain a revised set of solutions for scores applicable to the multiple factor table. The scores appropriate to each factor will be interpreted as observed partial effects. In section 5, to follow, a χ^2 test for the significance of these partial effects will be proposed, and this proposition will be evaluated on the basis of empirical sampling distributions of the measures of partial association.

4.3. A Restriction on x_i

In section 3.3, the scores, x_i, were free to vary, subject to the restriction that their mean value be zero. Now let us specify an additional restriction: that the x_i be the sum of scores for the categories of U, V, _, and W, where the mean value of these latter scores is zero for each of the factors, U, V, _, W; in symbols,

$x_i = u_k + v_l + \_ + w_m$, where

u_k, v_l, _, and w_m are scores placed on the categories of U, V, _, and W, resp., and where

$\sum n_{k.\_..} u_k = \sum n_{.l\_..} v_l = \_ = \sum n_{..\_m.} w_m = 0$, and where

each value of i corresponds, as in Table 13, to a particular set of values for kl_m.

Since the mean value of the scores for each factor is zero, the mean value of x_i remains zero. Also, let the variance of x_i be unity, as in section 3. Thus, if we know the (a) values of u_k, the (b) values of v_l and so on to the (c) values of w_m, then the $g = a \times b \_ \times c$ values of x_i are determined.

In a manner which is completely analogous to that is section 3.3, the derivation of scores, x_i, or rather u_k, v_l, _, and w_m, for which the squared correlation between x_i and y_j is a maximum is presented in the following sub-section, after which it will be shown that such scores may be interpreted as 'partial effects'.

4.4. Derivation of M.C. Scores for the Multiple Factors

As in section 3.3, each of the categories of the dependent factor, Y, is characterized by a numerical quantity, y'_j. The scale and level of y'_j are standardized by means of the transformation,

$y_j = (y'_j - \overline{y}')/s_{y'}$, where $\overline{y}'$ and $s_{y'}$ are the mean and standard deviation, resp., of y'_j, such that the mean of y_j is zero and the variance is one.

A set of scores, $x_i = u_k + v_l \_ + w_m$, is to be determined such that the squared correlation between x_i and y_j is a maximum. The correlation between x_i and y_j is, by definition,

$$R' = \sum_{ij} n_{ij} x_i y_j/n = \sum_{kl\_mj} n_{kl\_mj}(u_k + w_l \_ + w_m)y_j/n \ ;$$

then we may solve for u_k, v_l, _, w_m (which determine the x_i) by maximizing the expression,

$$n^2R'^2 = (\sum_{kl\_mj} n_{kl\_mj}(u_k + v_l \_ + w_m)y_j)^2 \ , \text{ subject to the}$$

restrictions that

$$\sum n_{k.\_..}u_k = \sum n_{.l\_..}v_l = \_ = \sum n_{..\_m.}w_m = 0 \ , \text{ and}$$

$$\sum n_{i.}x_i^2 = \sum_{kl\_m} n_{kl\_m.}(u_k + v_l \_ + w_m)^2 = n \ , \text{ i. e., that}$$

all component factors have zero mean, and that the sum of component scores have unit variance; using La Grange multipliers, we maximize the expression,

$$(\sum_{kl\_mj} n_{kl\_mj}(u_k + v_l \_ + w_m)y_j)^2 - 2L_1\sum n_{k.\_..}u_k - 2L_2\sum n_{.l\_..}v_l$$
$$- \_ - 2L_f\sum n_{..\_m}w_m - L_{f+1}\sum_{kl\_m} n_{kl\_m.}(u_k + v_l \_ + w_m)^2 \ ; \qquad (11)$$

taking derivatives with respect to u_k, and setting equal to zero,

$$2(nR')\sum_{l\_mj} n_{kl\_mj}y_j - 2L_1 n_{k.\_..} - 2L_{f+1}\sum_{l\_m} n_{kl\_m.}(u_k + v_l \_ + w_m) = 0 \; ; \qquad (12)$$

summing with respect to k,

$$2(nR')\sum_{kl\_mj} n_{kl\_mj}y_j - 2nL_1 - 2L_{f+1}\sum_{kl\_m} n_{kl\_m.}(u_k + v_l \_ + w_m) = 0 \; ,$$

but,

$$\sum_{kl\_mj} n_{kl\_mj}y_j = \sum_{ij} n_{ij}y_j = \sum n_{.j}y_j = 0 \; , \text{ and}$$

$$\sum_{kl\_m} n_{kl\_m.}(u_k + v_l \_ + w_m) = \sum n_{k.\_..}u_k + \sum n_{.l\_..}v_l + \_ + \sum n_{..\_m.}w_m = 0 \; , \text{ so}$$

$2nL_1 = 0$ and therefore $L_1 = 0$; in precisely the same manner, differentiating (11) with respect to v_l, setting equal to zero, and summing with respect to l, we find $L_2 = 0$; similarly for all factors up to W, for which $L_f = 0$; then (12) simplifies to

$$(nR')\sum_j n_{k.\_.j}y_j - L_{f+1}\sum_{l\_m} n_{kl\_m.}(u_k + v_l \_ + w_m) = 0 \qquad (A)$$

and similarly, the derivatives with respect to $v_l \_ w_m$ simplify to

$$\left.\begin{array}{ll} (nR')\sum_j n_{.l\_.j}y_j - L_{f+1}\sum_{k.\_m} n_{kl\_m.}(u_k + v_l \_ + w_m) = 0 & (B) \\ \vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots & \vdots \\ (nR')\sum_j n_{..\_mj}y_j - L_{f+1}\sum_{kl\_} n_{kl\_m.}(u_k + v_l \_ + w_m) = 0 & (F) \end{array}\right\} (13) \; ;$$

now multiplying (13)(A) by $(u_k + v_{l*} \_ + w_{m*})$, where the starred subscript distinguishes a particular value of the subscript, and similarly, multiplying (B) by $(u_{k*} + v_l \_ + w_{m*})$, and so on to (13)(F), which is multiplied by $(u_{k*} + v_{l*} \_ + w_m)$, we get

$$(nR')\sum_j n_{k.\_.j}y_j(u_k + v_{l*} \_ + w_{m*}) - L_{f+1}\sum_{l\_m}(u_k + v_l \_ + w_m)(u_k + v_{l*} \_ + w_{m*}) = 0 \; ,(A)(14)$$

and similar terms for (14)(B) through (F);

summing (14)(A) over k, l*, _, m*,

$$(nR')\sum_{kl*\_m*j} n_{k.\_.j}y_j(u_k + v_{l*} \_ + w_{m*}) - L_{f+1}\sum_{kl*\_m*l\_m}(u_k + v_l \_ + w_m)(u_k + v_{l*} \_ + w_{m*}) = 0 \; ,$$

and simplifying, we get

$$(nR')(b)\_(c)\sum_{kj} n_{k.\_.j} u_k y_j - L_{f+1}(b)\_(c)\sum_{kl\_m} n_{kl\_m}(u_k^2 + u_k v_l \_ + u_k w_m) = 0 \; ; \quad (A)(15)$$

letting Q = (a)(b)_(c), Q/a then equals (b)_(c), so (15)(A) becomes

$$(nR'Q/a)\sum_{kj} n_{k.\_.j} u_k y_j - (L_{f+1}Q/a)\sum_{kl\_m} n_{kl\_m.}(u_k^2 + u_k v_l \_ + u_k w_m) = 0; \quad (A)(16)$$

similarly for (B), _, (F);

multiplying (16)(A) by a, (16)(B) by b, and so on to (16)(F), which is multiplied by c, then summing all these equations, we get

$$nR'Q(\sum_{kj} n_{k.\_.j} u_k y_j + \sum_{lj} n_{.l\_.j} v_l y_j + \_ + \sum_{mj} n_{..\_mj} w_m y_j)$$
$$- L_{f+1}Q\sum_{kl\_m} n_{kl\_m.}(u_k + v_l \_ + w_m)^2 = 0;$$

but the first term is

$nR'Q(nR')$, by definition, and the second term is

$L_{f+1}Qn$, by definition, so

$n^2R'^2Q - L_{f+1}Qn = 0;$

that is,

$$L_{f+1} = nR'^2 \; .$$

Substituting this value of L_{f+1} in (13), we have

$$(nR')\sum_{j} n_{k.\_.j} y_j - nR'^2\sum_{l\_m} n_{kl\_m.}(u_k + v_l + \_ + w_m) = 0 \; , \quad (A)$$

and similar expressions for (B) through (F); consequently, omitting dot and dash notations on the subscripts (these being understood),

$$\left.\begin{aligned}
n_k u_k + \sum_l n_{kl} v_l + \_ + \sum_m n_{km} w_m &= \sum_j n_{kj} y_j / R' \\
\sum_k n_{kl} u_k + n_l v_l + \_ + \sum_m n_{lm} w_m &= \sum_j n_{lj} y_j / R' \\
\vdots \qquad\quad \vdots \qquad\quad \vdots \qquad &\quad \vdots \\
\sum_k n_{km} u_k + \sum_l n_{lm} v_l + \_ + n_m w_m &= \sum_j n_{mj} y_j / R'
\end{aligned}\right\} \; ;$$

finally, letting

$\sum_l n_{kl} v_l / n_k = \bar{v}_k \; ; \sum_j n_{kj} y_j / n_k = \bar{y}_k$; and so forth, we have,

$$\left.\begin{array}{l} u_k + \bar{v}_k + \ldots + \bar{w}_k = \bar{y}_k/R' \\ \bar{u}_l + v_l + \ldots + \bar{w}_l = \bar{y}_l/R' \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ \bar{u}_m + \bar{v}_m + \ldots + w_m = \bar{y}_m/R' \end{array}\right\} \qquad (17)$$

Equations (17) are (a + b + ... + c) simultaneous linear equations in the unknowns, u_k, v_l, ..., w_m, and, together with the restriction that the mean values of u_k, v_l, ..., w_m each be zero, can be solved. As a practical method of solution, each equation may be multiplied through by R'; then the equations,

$$\left.\begin{array}{l} R'u_a + R'\bar{v}_a + \ldots + R'\bar{w}_a = \bar{y}_a \\ R'\bar{u}_b + R'v_b + \ldots + R'\bar{w}_b = \bar{y}_b \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ R'\bar{u}_c + R'\bar{v}_c + \ldots + R'w_c = \bar{y}_c \end{array}\right\} \quad ,$$

are replaced, respectively, by

$$\sum_k n_k u_k = 0 \; ; \sum_l n_l v_l = 0 \; ; \quad \ldots \quad ; \sum_m n_m w_m = 0 \quad ;$$

the system is then solved for $R'u_k$, $R'v_l$, and so on to $R'w_m$. Finally, R'^2 is found by computing the variance of $R'x_i = R'u_k + R'v_l + \ldots + R'w_m$, for,

$\text{Var}(R'x_i) = R'^2 \text{Var } x_i = R'^2$, since Var x_i is unity. Taking the positive square root of R'^2 and dividing this into $R'u_k$, and so on, the values of u_k, v_l, ..., and w_m are determined. These are the m.c. scores for the U, V, ..., W component factors.

4.5. Interpretation of the M.C. Scores

In the simplest case, there would be but one component of X, say W. Then X ≡ W, i = m, and g = c, and since w_m and x_i are both defined as the scores which maximize the squared correlation, $x_i \equiv w_m$. But the x_i are interpreted as the effects of factor X; so the w_m are also interpreted as the effects of W.

However, the interpretation of u_k, v_1, ..., and w_m as effects is not quite so obvious in the general case. In order to arrive at such an interpretation, consider the following (c x h) table, which is derived as a marginal table from Table 13:

Table 15

Marginal Contingency Table

(Factors W and Y)

	Y				
j m	1	2	. .	h	n_m
1	$n_{\cdot\cdot\_11}$	$n_{\cdot\cdot\_12}$	. .	$n_{\cdot\cdot\_1h}$	$n_{\cdot\cdot\_1\cdot}$
2	$n_{\cdot\cdot\_21}$	$n_{\cdot\cdot\_22}$	. .	$n_{\cdot\cdot\_2h}$	$n_{\cdot\cdot\_2\cdot}$
.	.	.	. .	.	.
.	.	.	. .	.	.
c	$n_{\cdot\cdot\_c1}$	$n_{\cdot\cdot\_c2}$	. .	$n_{\cdot\cdot\_ch}$	$n_{\cdot\cdot\_c\cdot}$
n_j	$n_{\cdot\cdot\_\cdot1}$	$n_{\cdot\cdot\_\cdot2}$	. .	$n_{\cdot\cdot\_\cdot h}$	n

The tabled frequencies are summations over k, l, _, l' . Now, since Table 15 is a (c x h) contingency table, the technique of section 3 could be applied to test the simple association between W and Y. In particular, the simple observed effects of W would be given by

$w'_m = \bar{y}_m/R_f$, by formula (5), where w'_m has mean zero and unit variance, and where R_f is the correlation between w'_m and y_j.

However, we know that the value of $\bar{y}_m$ may be affected not only by the mth category of W, but also by the factors, U, V, _, V'. Suppose, for example, that the effects of the categories of U alone are the known values, u^*_k. Then we may compute the net effect of factor U on $\bar{y}_m$ as follows: (when the meaning is clear without use of dot and dash notation on the subscripts, they are omitted)

for $k = 1$, the effect of U is u^*_1; this effect applies to $n_{1.\_..}$ elements in the sample; for $k = 2$, the effect of U is u^*_2 and applies to $n_{2.\_..}$ elements; then the effect of U on the combined cells, 1._.. and 2._.. is

$(n_{1.\_..}u^*_1 + n_{2.\_..}u^*_2)/(n_{1.\_..} + n_{2.\_..})$, that is, a weighted average of the two effects; similarly, the effect of U on all cells combined is

$$\sum n_k u^*_k/n \ ;$$

but, by scale requirements,

$\sum n_k u^*_k = 0$, so the net effect of all categories of U on the whole sample is

$$\sum n_k u^*_k/n = 0 \ ; \qquad (18)$$

equation (18) may be written,

$$\sum n_k u^*_k/n = \sum_{mk} n_{mk}u^*_k/n = \sum_m (n_m/n)\sum_k (n_{mk}u^*_k/n_m) = 0 \ ;$$

but $\sum_k n_{mk}u^*_k/n_m$ is merely the weighted average of the effects of U on the mth category of W; in the same way as argued above, this is the net effect of all categories of U on the mth category of W, and may be denoted $\overline{u}^*_m$; similarly, $\overline{v}^*_m = \sum_l n_{ml}v^*_l/n_m$ is the net effect of V on the mth category of W; and so on. Now in the mth category of W, assuming no interaction of effects and no random effects on the dependent variable, the total observed effect, $\overline{y}_m/R_f$, must be the sum of the effects of each of the factors in the mth category of W; in symbols, that is

$$\overline{v}^*_m + \overline{u}^*_m + \ldots + w^*_m = \overline{y}_m/R_f \ ;$$

by the same argument, we could write,

$$\left.\begin{array}{l} u^*_k + \bar{v}^*_k + \dots + \bar{w}^*_k = \bar{y}_k/R_1 \\ \bar{u}^*_l + v^*_l + \dots + \bar{w}^*_l = \bar{y}_l/R_2 \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ \bar{u}^*_m + \bar{v}^*_m + \dots + w^*_m = \bar{y}_m/R_f \end{array}\right\} \quad ;$$

multiplying, appropriately, by R_1, or R_2, ..., or R_f, and dividing each expression by R', we get

$$\left.\begin{array}{l} u_k + \bar{v}_k + \dots + \bar{w}_k = \bar{y}_k/R' \\ \bar{u}_l + v_l + \dots + \bar{w}_l = \bar{y}_l/R' \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ \bar{u}_m + \bar{v}_m + \dots + w_m = \bar{y}_m/R' \end{array}\right\} \quad ,$$

where $u_k = R_1 u^*_k/R'$, and so on; but this last system of equations is identical to (17), sub-section 4.4, where u_k, v_l, ..., and w_m are the m.c. scores. Thus the m.c. scores differ from the defined partial effects only in scale, the proportionality factor being R_1/R' for the u_k, R_2/R' for the v_l, and so on to R_f/R' for the w_m. We may therefore refer to the m.c. scores of equations (17) as the observed partial effects of the factors.

4.6. Equivalence of M.C. Scores with Least-Squares Estimates of Effects

An additive effects model may be set up and the 'normal' equations for estimating effects may be developed on the principle of least-squares, as is usually done for the analysis of variance in experiments.

Let,

$$y_j = u^*_k + v^*_l + \dots + w^*_m + e_{kl\_mj} \ ,$$

where y_j is a numerical value with mean zero,

u^*_k, v^*_l, ..., and w^*_m are the estimated effects,

and e_{kl_mj} is the difference between y_j and the sum of the estimated effects for given levels of k, l, ..., m and j. By the principle of least-squares, the estimated effects are found by minimizing the variance of e_{kl_mj}, that is, by minimizing

$$\sum_{kl\_mj} n_{kl-mj}(e_{kl\_mj})^2 = \sum_{kl\_mj} n_{kl\_mj}(y^*_j - u^*_k - v^*_l - \dots - w^*_m)^2 .$$

Taking derivatives with respect to u^*_k, and setting equal to zero,

$$-2\sum_{l\_mj} n_{kl\_mj}(y^*_j - u^*_k - v^*_l - \dots - w^*_m) = 0 ;$$

this simplifies to, (omitting dot and dash notation on subscripts)

$$n_k u^*_k + \sum_l n_{kl} v^*_l + \dots + \sum_m n_{km} w^*_m = \sum_j n_{kj} y^*_j \text{ , which in turn can}$$

be expressed as

$$\text{similarly,}\quad \left.\begin{array}{l} u^*_k + \overline{v}^*_k + \dots + \overline{w}^*_k = \overline{y}_k \quad ; \\ \overline{u}^*_l + v^*_l + \dots + \overline{w}^*_l = \overline{y}_l \\ \vdots \\ \overline{u}^*_m + \overline{v}^*_m + \dots + w^*_m = \overline{y}_m \end{array}\right\} \quad . \qquad (19)$$

Equations (19), together with the arbitrary scale requirement that the grand mean of each set of effects be zero, have unique solutions for u^*_k, v^*_l, ..., and w^*_m.

Dividing equations (19) by R', we again get

$$\left.\begin{array}{l} u_k + \overline{v}_k + \dots + \overline{w}_k = \overline{y}_k/R' \\ \overline{u}_l + v_l + \dots + \overline{w}_l = \overline{y}_l/R' \\ \vdots \\ \overline{u}_m + \overline{v}_m + \dots + w_m = \overline{y}_m/R' \end{array}\right\} \quad ,$$

where $u_k = u^*_k/R'$, etc., and this system of equations is identical to (17).

Thus, the least squares solutions differ from m.c. scores only by the factor, R'. Again, as in the simple case in section 3, it is seen that the equations for finding m.c. scores are equivalent to those used in the analysis of variance for finding observed 'effects'.

4.7. The Relation of R' to Variation Among the Dependent Means

In section 3, it was shown that R, defined as the correlation between m.c. scores, x_i, and the dependent values, y_j, was equal to the standard deviation of the observed means of y (see equation 6). An analogous relation is found to exist in the multi-factor case now under discussion, because: from equations (17), we may write

$R'u_k = \bar{y}_k - R'(\bar{v}_k + \ldots + \bar{w}_k)$, and similar expressions for $R'v_1$ through $R'w_m$; the right-hand side of the relation is equal to the observed mean of y, less a set of corrections for the net effects of v, ..., w; therefore, we define

$\bar{y}_{kadj} = \bar{y}_k - R'(\bar{v}_k + \ldots + \bar{w}_k)$ as the adjusted value of $\bar{y}_k$;

thus,

$$R'u_k = \bar{y}_{kadj} \;;\; R'v_1 = \bar{y}_{1adj} \;;\; \ldots \;;\; R'w_m = \bar{y}_{madj} \;;$$

now, the variance of the sum of partial scores has been specified to be unity (section 4.4); that is,

$$(1/n)\sum_{kl\_m} n_{kl\_m.}(u_k + v_1 + \ldots + w_m)^2 = 1 \quad ;$$

substituting $\bar{y}_{kadj}/R'$ for u_k, and similarly for the other scores,

$$(1/nR'^2)\sum_{kl\_m} n_{kl\_m.}(\bar{y}_{kadj} + \bar{y}_{1adj} + \ldots + \bar{y}_{madj})^2 = 1 ,$$

that is,

$$R'^2 = \mathrm{var}(\bar{y}_{kadj} + \bar{y}_{1adj} + \ldots + \bar{y}_{madj}) \quad . \qquad (20)$$

Thus, it is found that R' is equal to the standard deviation of the sum of adjusted means.

4.8. Particular Case: Partial Association when the Dependent Factor is a Dichotomy

In the particular case of a dependent dichotomy, that is, $j = 1,2$, equations (17) of course are equally applicable as for three or more dependent categories. As in section 3, we may express the equations in the more familiar terms of proportions.

As before,

$$y_1 = -\sqrt{p/q}\ , \text{ and } y_2 = \sqrt{q/p}\ , \text{ where } p = n_{\cdot\cdot\_\cdot 2}/n \text{ and } p + q = 1.$$

Since

$$\bar{y}_k = \sum_j n_{kj} y_j / n_k = (n_{k\cdot\_\cdot 1} y_1 + n_{k\cdot\_\cdot 2} y_2)/n_{k\cdot\_\cdot\cdot}\ , \text{ then}$$

$$\bar{y}_k = -(n_{k\cdot\_\cdot 1}/n_{k\cdot\_\cdot\cdot})\sqrt{p/q} + (n_{k\cdot\_\cdot 2}/n_{k\cdot\_\cdot\cdot})\sqrt{q/p}\ .$$

Denoting $p_k = n_{k\cdot\_\cdot 2}/n_{k\cdot\cdot\_\cdot\cdot}$ as the observed proportion of elements in the second category of Y, for the kth category of U, we have

$$\bar{y}_k = p_k\sqrt{q/p} - (1 - p_k)\sqrt{p/q}\ .$$

Simplifying, we get

$\bar{y}_k = (p_k - p)/\sqrt{pq}$; that is, $\bar{y}_k$ is proportional to the deviation of p_k from the overall p in the sample; and since p and q are constant for the sample, p_k corresponds to $\bar{y}_k$. Therefore, for the case of a dependent dichotomy, $\bar{y}_{kadj}$ may be expressed as

$$\bar{y}_{kadj} = (p_{kadj} - p)/\sqrt{pq}\ .$$

Then equations (17) become

$$\left.\begin{array}{l} u_k = (p_{kadj} - p)/\sqrt{pq} \\ v_1 = (p_{1adj} - p)/\sqrt{pq} \\ \cdot \qquad\qquad \cdot \\ \cdot \qquad\qquad \cdot \\ w_m = (p_{madj} - p)/\sqrt{pq} \end{array}\right\} \qquad (21)$$

Knowing p and q from the sample, and having solved for the partial effects of, say, factor W, it would then be a simple matter for one to express the observed results in terms of adjusted proportions by the following relations:

$$p_{madj} = p + w_m/\sqrt{pq} \quad . \tag{22}$$

If, for example, factor U were stratum, V were interviewer group, W were 1st interview health status, and Y denoted 2nd interview health status, as in Table 14, the p_{madj} would be the stratum-interviewer adjusted proportion of households with 2nd year health problems for the two groups of households under factor W. For purposes of presentation, it would also be possible to construct an adjusted 'contingency' table to show the observed partial relation between 1st and 2nd year health problems. (See sub-section 2.42, Table 8, for an illustration of this.)

4.9. Summary

If the dependent classification can be characterized by a set of quantities, y_j, $j = 1, \ldots, q$, with zero mean and unit variance, then a set of scores for the categories of each independent factor may be determined by means of the formulae,

$$\left.\begin{array}{ll} u_k = \overline{y}_{kadj}/R' \;, & k = 1, \ldots, a \\ v_l = \overline{y}_{ladj}/R' \;, & l = 1, \ldots, b \\ \cdot \quad\quad \cdot & \\ \cdot \quad\quad \cdot & \\ w_m = \overline{y}_{madj}/R' \;, & m = 1, \ldots, c \end{array}\right\} \quad ;$$

where $u_k, v_l, \ldots, w_m$ are scores for the independent factors, U, V, ..., W, resp.;

where $\overline{y}_{kadj} = \overline{y}_k - R'(\overline{v}_k + \ldots + \overline{w}_k)$, and similar expressions in subscripts l, ..., m;

and where R' is the product moment correlation between y_j and the sum of $u_k, v_l, \ldots, w_m$.

These scores are such that R'^2 is a maximum. They are analogous to least-squares effects which would be determined in non-orthogonal experiments. These scores are termed 'partial effects', because each set of scores for a given factor is derived after adjusting for the influence attributed to the remaining factors. Finally, the variance of the sum of 'partial effects' is equal to R'^2.

For dependent dichotomies, these results may be expressed in terms of rates. Adjusted 'contingency' tables also may be constructed to illustrate the observed partial association between any given factor and the dependent factor.

In this section, any reference to properties of the observed measures as estimators has been studiously avoided. This is because there are several aspects of the partial association problem which do not match that of simple association. These aspects will be discussed in the following section. A χ^2 test of partial association will then be proposed, and this proposal will be evaluated on the basis of empirical sampling distributions.

5.0. SAMPLING DISTRIBUTIONS OF MEASURES OF PARTIAL ASSOCIATION; ESTIMATOR PROPERTIES

The observed simple association in a (g x h) contingency table is given by R, the maximized correlation. When no association exists in the universe from which the sample has been obtained, nR^2 is distributed asymptotically as χ^2 with (g - 1) degrees of freedom, as n becomes large. Further, if an arbitrary set of values is assigned to the categories of X, the correlation between these values and the dependent variable has a normal distribution, asymptotically as n becomes large, when there is no association in the universe. (See section 3.) The approach to a χ^2 or normal distribution, as the case may be, is rapid enough so that, when most cell expectations are 5 or more, the theoretical distributions may be used for all practical purposes. But it should be emphasized that this rule applies to samples from a universe in which no association is present. When association is present to an appreciable degree, then the sample size may need to be considerably larger before the theoretical distributions become close approximations to correct sampling distributions.[21] In general, then, valid confidence statements based on the asymptotically approached distributions require larger samples than do valid tests of the null hypothesis.

In the case of partial association, it may be desired either to test a null hypothesis or to make a confidence statement for the partial association of, say, factor W with factor Y. When a test of the null hypothesis is made, it is pointless to assume no universe association for factors U, V, ..., V', as well as for W, because one of the fundamental reasons for

[21] _Op_. _cit_., p. 57.

studying partial association is to adjust for the supposed influence of the extraneous factors. Rather, the null hypothesis must be of the form: there is no association of W with Y in the universe, but U, V, ..., V' may be associated with Y. Similarly, when confidence statements are made for the influence of W, not only U, V, ..., V', but also W may be associated with Y in the population. Therefore, an association, usually of appreciable magnitude, almost always is present in the universe.

Clearly, one face of the general problem posed by the existence of universe association is to determine some rule for minimal sample sizes. But even if one were successful in determining such a rule, it might be of limited value if, by that rule, inordinately large samples were required. Consequently, there remains the more general task of determining the bias and error variance and distributional form of partial association measures developed from samples of given sizes, as related to universes of varied types. Completely general answers to these problems will not be coming forth here; but, by the selection of prototypic populations and by the generation of empirical sampling distributions from them, some insight is gained.

Before presenting the empirical results, it is necessary to adopt a means of characterizing different universes and to define explicitly what is meant by a given level of association in the universe. Also, an outline of the procedures used for selecting samples and for generating empirical distributions on the IBM 650 digital computer will be given. Following this, the empirical results will be presented and discussed.

3.1. Universe Description; Universe Parameters

The schema presented in sections 3 and 4 is readily adapted to the specification of the universe and to the definition of universe parameters. For the universe, the cell entries of Table 13 become $P_{ij} = P_{kl\_mj}$, in place of $n_{ij} = n_{kl\_mj}$, where P_{kl_mj} is the proportion of universe elements in any given cell, such that $\sum_{kl\_mj} P_{kl\_mj} = 1$. Marginal universe relative frequencies are denoted by dot notation corresponding in every respect to the notation employed in section 4 for sample frequencies.

Considering the universe first as a (g x h) contingency table, we then define the squared total universe relation between the composite independent factor and the dependent factor as

$$\underline{R}^2 = \text{var } \overline{\underline{y}}_i \ . \qquad (23)$$

The underscore notation above and in what follows denotes universe measures, in contrast to sample measures for which no underscores are made.

Now considering the universe as an (a x b x ... x c x h) contingency table, we define the additive partial 'effects' of the various factors by the equations:

$$\left.\begin{array}{l} \underline{u}_k + \overline{\underline{v}}_k + \dots + \overline{\underline{w}}_k = \overline{\underline{y}}_k/\underline{R}' \\ \overline{\underline{u}}_l + \underline{v}_l + \dots + \overline{\underline{w}}_l = \overline{\underline{y}}_l/\underline{R}' \\ \cdot \quad\ \cdot \qquad \cdot \qquad \cdot \qquad \cdot \\ \cdot \quad\ \cdot \qquad \cdot \qquad \cdot \qquad \cdot \\ \overline{\underline{u}}_m + \overline{\underline{v}}_m + \dots + \underline{w}_m = \overline{\underline{y}}_m/R' \end{array}\right\} ,$$

or more simply,

$$\left.\begin{array}{l} \underline{u}_k = \overline{\underline{y}}_{kadj}/\underline{R}' \\ \underline{v}_l = \overline{\underline{y}}_{ladj}/\underline{R}' \\ \cdot \qquad \cdot \\ \cdot \qquad \cdot \\ \underline{w}_m = \overline{\underline{y}}_{madj}/\underline{R}' \end{array}\right\} , \qquad (24)$$

completely analogous to equations (17), section 4.4.

The squared total relation due to additive effects is given by

$$\underline{R}'^2 = \text{var}(\overline{\underline{y}}_{\text{kadj}} + \overline{\underline{y}}_{\text{ladj}} + \ldots + \overline{\underline{y}}_{\text{madj}}) \text{ , which is the} \qquad (25)$$

expression analogous to (20), section 4.7.

Also, we may now define the squared amount of relation not due to additive effects, i. e. the squared total interaction in the universe, as the difference between $\underline{R}^2$ and $\underline{R}'^2$:

$$\underline{I}^2 = \underline{R}^2 - \underline{R}'^2 \text{ .} \qquad (26)$$

The simple, i. e. unadjusted, 'effects' of a given independent factor in the universe may also be defined by forming a universe table analogous to Table 15, section 4.5. Then, by analogy with (6), section 3.3, the squared amount of simple association between, say, factor W and Y is given by

$$\underline{R}^2_{\text{wy}} = \text{var } \overline{\underline{y}}_{\text{m}} \text{ .} \qquad (27)$$

By further analogy, we define the squared amount of partial association due to factor W as the variance of the adjusted means, $\overline{\underline{y}}_{\text{madj}}$:

$$\underline{R}'^{\ 2}_{\text{wy}} = \text{var } \overline{\underline{y}}_{\text{madj}} \text{ .} \qquad (28)$$

But since $\underline{w}_{\text{m}} = \overline{\underline{y}}_{\text{madj}}/\underline{R}'$, then var $\overline{\underline{y}}_{\text{madj}} = \underline{R}'^2$ var $\underline{w}_{\text{m}}$. (29)

Substituting this expression in (28), we re-define the squared total amount of partial association due to W as

$$\underline{R}'^{\ 2}_{\text{wy}} = \underline{R}'^2 \text{ var } \underline{w}_{\text{m}} \text{ .} \qquad (30)$$

Finally, the squared total amount of partial association due to factor W may be partitioned into (c - 1) components, where, it is remembered, c is the number of categories of factor W. In particular, we can assign arbitrary values, w'_{m}, to the categories of W and denote $r_{w'\underline{w}}$ as the correlation between the arbitrary scores and the partial effects of W. Then $r_{w'\underline{w}}^{\ 2}$ is the proportion of the squared total amount of partial association due to the arbitrary values of W. Consequently, we write

$$\underline{R}'_{w'y}{}^2 = \underline{R}'^2 \, r_{w'\underline{w}}{}^2 \text{ var } \underline{w}_m \tag{31}$$

as the squared component of partial association attributable to the arbitrary values of W. If the w'_m values happen to be numerical descriptions of the categories of W, in some context, then $\underline{R}'_{w'y}{}^2$ is the squared linear partial association, in that context, of W with Y. The sign of $r_{w'\underline{w}}$ determines whether such linear association is negative or positive.

Summarizing, we have the following universe parameters, in addition to the additive partial effects given by (24):

$\underline{R}$, the total association of the composite of independent factors with Y;

$\underline{R}'$, the total association due to the purely additive 'effects' of the independent factors;

$\underline{I}$, the total interactive association;

$\underline{R}_{wy}$, the partial association of W with Y; and similar expressions for factors U, V, ..., V';

$\underline{R}'_{w'y}$, the linear partial association of values, assigned to categories of W, with Y; and similar expressions for factors U, V, ..., V'.

By setting up universes with different values for these various parameters, it will be possible to examine the influence of these parameters on the properties of empirical sampling distributions of measures of partial association.

Each of the above universe parameters, it is noted, is determined by the set of P_{kl_mj}, for the universe. Thus, by the selection of the proper values for the P_{kl_mj}, a universe with any desired values of these parameters may be specified.

5.2. Observed Measures are Consistent Estimates of Universe Parameters

Universe measures and sample measures of partial effects are computed in identical manner. Also, as n becomes large, the relative magnitudes of sample cell frequencies converge stochastically to the universe relative frequencies. Then, considering that the measures of association are in the form of averages and variances, the limiting values of the measures of association are equal to the universe values, as n approaches infinity. That is to say, the sample values are consistent estimates of the universe parameters.[22]

5.3. Proposed χ^2 Test for the Significance of Partial Association

Referring again to Table 15, the significance of the simple association between W and Y could be tested by

$$\chi_{c-1}^2 = nR_{wy}^2 = ns_{\bar{y}_m}^2 \text{, from (6), Section 3.}$$

Now, from equations (17),

$$w_m = \bar{y}_{madj}/R' .$$

Then by analogy with the simple case, we hypothesize that

$\chi_{c-1}^2 = nR'^2_{wy} = ns_{\bar{y}_{madj}}^2$, where R'_{wy} is the correlation between w_m and y_j . (32)

Since $w_m = \bar{y}_{madj} / R'$, $s_{\bar{y}_{madj}}^2 = R'^2 s_{w_m}^2$; so from (32) we get the relation,

$$R'_{wy} = R's_{w_m} , \tag{33}$$

that is, the observed partial correlation between W and Y equals the total additive association due to all factors times the standard deviation of the observed partial effects of W.

[22] See Slutsky's theorem, p. 255, of Cramer, H., "Mathematical Methods of Statistics", Princeton University Press, 1946.

Finally, by substitution of (33) in equation (32),

$$\chi^2_{(c - 1)} = nR'^2 s_{w_m}^{\ 2} \ . \tag{34}$$

Equation (34) expresses that, when the squared universe partial association, $\underline{R}'^2 \text{var } \underline{w}_m$, equals zero, n times the observed squared sample association, $R'^2 s_{w_m}^{\ 2}$, is asymptotically distributed as a χ^2 with (c - 1) degrees of freedom. This statement is not proven here, but is merely hypothesized. As an hypothesis it will be tested against empirically generated distributions of $nR'^2 s_{w_m}^{\ 2}$ when the universe values of $\underline{R}'^2 \text{var } \underline{w}_m = 0$.

5.4. Hypothesized Asymptotic Distribution of the Observed Linear Partial Association

Taking $R'_{wy}{}^2 = \text{var } \bar{y}_{madj}$ as the observed squared partial association of W with Y, we may obtain its linear component when arbitrary values, w'_m, are assigned to the categories of W as follows:

let $r_{w'w}$ be the correlation between w'_m and w_m; then $r_{w'w}{}^2$ is the proportion of the squared partial association which is attributable to the arbitrary values of W; so $R'_{wy}{}^2 r_{w'w}{}^2$ is the observed squared linear partial association of W with Y, denoted $R'_{w'y}{}^2$; then, since

$$R'_{wy}{}^2 = R'^2 s_{w_m}^{\ 2} \ ,$$
$$R'_{w'y}{}^2 = R'^2 s_{w_m}^{\ 2} r_{w'w}{}^2 \ ;$$

finally, if it is true that

$$\chi^2_{(c - 1)} = nR'^2 s_{w_m}^{\ 2} \ , \text{ from the hypothesized equation (34),}$$

then

$$\chi^2_1 = nR'^2 s_{w_m}^{\ 2} r_{w'w}{}^2 \ . \tag{35}$$

Equation (35) expresses that, when the universe squared partial association, $\underline{R}'^2 r_{w'\underline{w}}{}^2 \text{var } \underline{w}_m$, equals zero, n times the observed squared sample association, $R'^2 s_{w_m}^{\ 2} r_{w'w}{}^2$, is asymptotically distributed as χ^2

with one degree of freedom. This statement is true only if (34) is true in the preceding sub-section 5.3. If it is true, then $nR'r_{w'w}s_{w_m}$ is asymptotically distributed as a standard normal deviate. This, too, will be tested as an hypothesis against empirically generated distributions of $nR'r_{w'w}s_{w_m}$ when the universe value of $\underline{R}'r_{w'\underline{w}}\sqrt{\text{var}\underline{w}_m} = 0$. However, we may go further than this; we may hypothesize that, when the universe value of $\underline{R}'r_{w'\underline{w}}\sqrt{\text{var }\underline{w}_m} \neq 0$, $nR'r_{w'w}s_{w_m}$ is asymptotically distributed as a normal deviate with unit variance and mean equal to $n\underline{R}'r_{w'\underline{w}}\sqrt{\text{var }\underline{w}_m}$.

5.5. Procedure for Generating Empirical Distributions on the IBM 650 Computer

The details of programming on the IBM 650 digital computer are too lengthy and technical to recount here. Nevertheless, the following is a general outline of the procedure used in generating empirical distributions.

The universe from which samples of a given size are to be taken is chosen as a 3^4 universe, that is, a universe with four factors of three categories each. This allows us to denote one of the factors as the dependent variable, Y, and three factors as independent variables, U, V, and W. Within these limitations, we may select universes containing no relations, containing various levels of additive partial association of one, two, or all three independent factors with Y, or containing not only additive associations but also interactive associations. The universe contains $3^4 = 81$ cells; to each cell corresponds a proportion, P_{klmj}, the full set of which adds to 1; this full set completely determines not only the universe partial 'effects', but also all the other universe parameters listed in section 5.1.

The 81 cells of the universe are identified with 81 corresponding cells in the computer memory. In the first computer cell, the first universe proportion, P_1, is placed; in the second computer cell, $P_1 + P_2$ is placed; in the third, $P_1 + P_2 + P_3$; and so on up to the 81st cell which contains $P_1 + P_2 + \ldots + P_{81} = 1$. These proportions are shown to the third decimal place. For example, we might have $P_1 = .005$, $P_2 = .052$, $P_3 = .025$, etc. Then in the computer cells, we would have .005, .057, .082, etc., resp., all the way to 1.000.

Now a correspondence is made between the computer cell contents and the random numbers between 000 and 999: the random numbers, 000 through 004, correspond to the first cell; 005 through 056 to the second; and so on. Thus, to each cell corresponds a number of random elements in proportion to the universe relative frequency.

Once the cumulated universe relative frequencies have been placed in the computer memory, a random number from 000 to 999 is selected by the computer. The correspondence between the random number and the appropriate universe cell is made by the computer, and a frequency of one is then placed in a sample region of the computer memory. This sample region also contains 81 cells, each of which corresponds to a universe cell. For example, if the first random digit is 056, a count of one is placed in the second cell of the sample region. Following this, a second random digit is selected, the correspondence is made, and a frequency of one is added to the proper sample cell. This process continues until a sample of size n has been generated.

Once the sample has been generated, the partial effects and measures of partial and total association are computed, according to the equations which have been presented. These sample values are punched on a card, the machine clears itself for a new sample, and the process of selecting a new random sample of the same size, n, is begun again.

The time required to select a sample of size 50 and solve for measures of association is about 2 1/6 minutes. Consequently, it takes about $3\frac{1}{2}$ hours to obtain 100 sample results for samples of size 50. This program utilizes only the basic 650 machine. If the program were optimally programmed with respect to time and if recently acquired auxiliary computing mechanisms were used, the time could be reduced by at least a factor of 3; thus the 2 minute cycle is 'slow'.

5.6. Empirical Results

Observation and experimentation are indispensible to the advancement of the natural sciences. In the past few centuries, it has been recognized that logic alone, unsupported by observable evidence, cannot capture the complexities of nature. Nor is experimentation and observation a stranger to the world of mathematics. Unlike nature, it may be theoretically possible to solve certain extremely complex problems in mathematics directly. However, the theoretical tools needed to solve a given problem may not yet have been invented, or are not practicable, or are not available. When this is the case, the mathematical problem may be approached on an experimental basis. This has been an approach familiar to the statistician in the last two centuries, and before him, the gambler.

The method consists of constructing a physical analogue of the mathematical problem in such a way that elements of the physical analogue can be sampled at random. From a sample of elements, the solution of the problem can be inferred. The following examples are not very complex, but they serve to illustrate the method:

Example 1.

Problem: Will the house eventually win a profit on the roulette table?

Solution: Bet x dollars n times (n large) and see who wins.

Example 2.

Problem: What is the area under a given curve?

Solution: Draw curve on square piece of paper with unit length and width. Color area under curve. Cut paper into a large number of small squares and number each square. Select small square at random and record $x = 1$ if colored, $x = 0$ if not colored, $x = \frac{1}{2}$ if partly colored. Replicate selection procedure n times. Then estimated area is $\sum x/n$. The larger n is, the more precise is the estimate.

Prior to the last 20 or 30 years, this technique was sometimes referred to as 'model sampling'. But more recently, the body of such technique has been unified, systematized and extended, and it is now termed the 'Monte Carlo method', in deference to its ancient origin.

The procedures outlined in the preceding sub-section, 3.5., are of this nature. The physical analogue, in that case, is set up in a computing machine, admirably suited to this type of work, and random sampling is achieved through use of a table of random digits fed into the machine. Thus the results which follow are experimental, i. e. empirical.

This approach has been taken because of the enormous complexity of the problem before us: to learn something about the statistical properties of measures of partial relation determined by finite samples from a variety of universes.

5.6.1. Universe 1

The first universe to be considered contains no association whatever, so that $\underline{R} = 0$, $\underline{R}' = 0$, $\underline{R}_{wy} = 0$, $\underline{R}'_{wy} = 0$, and $\underline{R}'_{w'y} = 0$. This is a trivial case, as discussed in section 5.0., in that we shall never be concerned with hypothesizing a universe of this type, even for a null hypothesis. However, this universe has been set up in order to test for the hypothetical χ^2 and normal distributions of the sample measures of partial association under the most favorable conditions. 65 samples, each of size 50, were taken, and the squared partial association of W with Y was computed for each sample. Each of the values was multiplied by n, for, as the reader will recall, it is hypothesized that nR'^{2}_{wy} is distributed asymptotically as a chi-square. In the particular case at hand, there being c = 3 categories of W, this chi-square has 2 degrees of freedom.

The grouped results, as compared to expectations, are shown in Table 16. With the exception of the third class ($.446 \leq nR'^{2}_{wy} \leq .712$) and the seventh ($3.219 \leq nR'^{2}_{wy} \leq 4.604$), the agreement between observed and expected frequencies is very good. In particular, the observed frequencies in the upper tail of the distribution, i. e. the last three classes, correspond very closely to expectation. Using the chi-square 'goodness of fit' test ($\sum(\text{obs.} - \text{exp.})^2/\text{exp.}$), the chi-square value of 8.65 is found to be not significant ($.7 > P > .5$); the observed deviations are well within the realm of chance variation. Evidently, a test for the significance of partial association in a random sample of size 50 from Universe 1, utilizing the chi-square distribution at, say, the 5% or 1% level, would be nearly correct.

In Table 17, the grouped results for $\sqrt{n}R'_{w'y}$ are presented. By hypothesis, these values are asymptotically distributed as a standard

Table 16

Empirical Sampling Distribution of n Times the Squared Partial Association of W with Y as Compared to χ^2 Expected Frequencies; Universe 1*; n = 50

nR'^{2}_{wy}	Frequency	
	Observed	Expected
0 - .210	7	6.5
.211 - .445	7	6.5
.446 - .712	11	6.5
.713 - 1.385	13	13.0
1.386 - 2.407	9	13.0
2.408 - 3.218	8	6.5
3.219 - 4.604	2	6.5
4.605 - 5.990	4	3.2
5.991 - 7.823	3	2.0
7.824 & up	1	1.3
All observations	65	65.

'Goodness of fit': $\chi^2_{10} = 8.65$; one-tailed P: .7 > P > .5

Table 17

Empirical Sampling Distribution of $\sqrt{n}$ Times the Linear Partial Association of W with Y as Compared to Standard Normal Expected Frequencies; Universe 1*; n = 50

$\sqrt{n}R'_{w'y}$	Frequency	
	Observed	Expected
-1.97 & less	1	1.6
-1.96 to -1.45	3	3.2
-1.44 to - .94	6	6.5
- .93 to - .60	11	6.5
- .59 to - .32	3	6.5
- .31 to - .00	7	8.1
+ .00 to .31	12	8.1
.32 to .59	7	6.5
.60 to .93	6	6.5
.94 to 1.44	3	6.5
1.45 to 1.96	4	3.2
1.97 & up	2	1.6
All observations	65	65.

'Goodness of fit': $\chi^2_{12} = 9.57$
One-tailed P:
.7 > P > .5

Test of mean: hyp. .000
obs. -.018
two-tailed P: P = .89

Test of var.: hyp. 1
obs. .89
two-tailed P: P = .62

*Universe 1: 3^4 cells; $\underline{R} = 0$; $\underline{R}' = 0$; $\underline{R}_{wy} = 0$; $\underline{R}'_{wy} = 0$; $\underline{R}'_{w'y} = 0$.

normal deviate. Inspection of Table 17 reveals that the empirical results agree well with expectations, particularly at the two tails. The 'goodness of fit' test yields a probability between .7 and .5, again well within the realm of chance variation. Also, as indicated below Table 17, the observed mean and variance of the 65 sample values are quite close to the hypothesized true values of zero and one, resp. These results tend to substantiate the hypothesis that, for samples of size 50 from Universe 1, the observed linear partial association is distributed approximately as a normal distribution with mean zero (no bias) and with variance equal to 1/n.

5.62. Universe 2

The second universe contains a total relation, $\underline{R}$, equal to .453. There is no interaction, as indicated by the fact that $\underline{R}'$ also equals .453. None of this total relation is due to the partial effects of W, since $\underline{R}'_{wy}$ as well as $\underline{R}'_{w'y}$ equals 0. But the universe frequencies are in a state of imbalance, as indicated by an appreciable simple association between W and Y, $\underline{R}_{wy} = .194$. It is desirable that, notwithstanding an appreciable simple association, the sample measures of the partial and linear partial association of W with Y be distributed according to the hypothesized chi-square and normal distributions which would be indicative of no partial association in the universe.

The squared partial associations observed for 91 samples, each of size 50, are presented in Table 18. For the smaller χ^2_2 values, the observed frequency is consistently lower than expectation; also, in the extreme upper tail (7.824 & up) the observed frequency is much higher than expectation. The 'goodness of fit' test yields a one-tailed probability between .10 and .05, of borderline significance. Consequently, the fit of the data to the hypothetical chi-square distribution is under suspicion.

Table 18

Empirical Sampling Distribution of n Times the Squared Partial Association of W with Y as Compared to χ^2 Expected Frequencies; Universe 2*; n = 50

nR'^{2}_{wy}	Frequency	
	Observed	Expected
0 - .210	6	9.1
.211 - .445	8	9.1
.446 - .712	5	9.1
.713 - 1.385	20	18.2
1.386 - 2.407	17	18.2
2.408 - 3.218	7	9.1
3.219 - 4.604	15	9.1
4.605 - 5.990	4	4.6
5.991 - 7.823	3	2.7
7.824 & up	6	1.8
All observations	91	91.

'Goodness of fit': $\chi^2_{10} = 17.5$; one-tailed P: $.10 > P > .05$

Table 19

Empirical Sampling Distribution of $\sqrt{n}$ Times the Linear Partial Association of W with Y as Compared to Standard Normal Expected Frequencies; Universe 2*; n = 50

$\sqrt{n}R'_{w'y}$	Frequency	
	Observed	Expected
-1.97 & less	6	2.3
-1.96 to -1.45	5	4.5
-1.44 to - .94	8	9.1
- .93 to - .60	3	9.1
- .59 to - .32	5	9.1
- .31 to - .00	6	11.4
+ .00 to .31	7	11.4
.32 to .59	8	9.1
.60 to .93	14	9.1
.94 to 1.44	10	9.1
1.45 to 1.96	12	4.5
1.97 & up	7	2.3
All observations	91	91.

'Goodness of fit': $\chi^2_{12} = 31.2$ Test of mean: hyp. .000   Test of var.: hyp. 1
One-tailed P: obs. .288   obs. 1.62
$P < .01$ two-tailed P: $P < .01$ two-tailed P: $P < .01$

*Universe 2: 3^4 cells; $\underline{R} = .453$; $\underline{R}' = .453$; $\underline{R}_{wy} = .194$; $\underline{R}'_{wy} = 0$; $\underline{R}'_{w'y} = 0$.

Table 19, showing the results for the linear component of the partial association, leaves little doubt that the actual sampling distribution does not conform to hypothesis. A probability well below .01 is obtained for the 'goodness of fit' test. Further, the grand mean of the observed values is .288, significantly greater ($P < .01$) than the hypothesized zero; finally, the sampling variance is observed to be 1.62 in contrast to the hypothetical unity, a significant departure ($P < .01$). Since the mean of $\sqrt{n}R'_{w'y} = .288$, the mean of $R'_{w'y}$ is $.288/\sqrt{50}$, or .04. Thus, the bias in this estimator is considered to be in the neighborhood of +.04. If this bias is compared to the simple association of .194, it is evident that a major portion of the extraneous influences of factors U and V has been eliminated. Nevertheless, when one considers that the two extreme tails of the empirical distribution contain 13 of the 91 observations, or more than 14%, in contrast to an expected 4.6 observations, or 5%, it is also evident that the use of the hypothetical normal curve in a test of significance at the 5% level would be unsatisfactory.

According to hypothesis, as the sample size increases, the sampling distributions should more nearly be approximated by chi-square and normal distributions, as the case may be. This is borne out when the sample size for samples from Universe 2 is increased from 50 to 150. Table 20 indicates that, for samples of size 150, the observed distribution of squared partial associations does not deviate significantly from hypothesis ($.7 > P > .5$); the fit is particularly good at the upper tail. This is in contrast to the very poor fit in the upper tail region for samples of size 50 (Table 18). Also, as shown in Table 21, the distribution of the linear partial association of W with Y does not deviate significantly from the hypothesized standard normal curve ($.2 > P > .1$). The grand mean of $\sqrt{n}R'_{w'y}$ for all forty

Table 20

Empirical Sampling Distribution of n Times the Squared Partial Association of W with Y as Compared to χ^2 Expected Frequencies; Universe 2*; n = 150

nR'^{2}_{wy}	Frequency Observed	Frequency Expected
0 - .210	6	4.0
.211 - .445	7	4.0
.446 - .712	3	4.0
.713 - 1.385	9	8.0
1.386 - 2.407	5	8.0
2.408 - 3.218	3	4.0
3.219 - 4.604	2	4.0
4.605 - 5.990	1	2.0
5.991 - 7.823	2	1.2
7.824 & up	2	.8
All observations	40	40.

'Goodness of fit': χ^2_{10} = 8.80 ; one-tailed P: .7 > P > .5

Table 21

Empirical Sampling Distribution of $\sqrt{n}$ Times the Linear Partial Association of W with Y as Compared to Standard Normal Expected Frequencies; Universe 2*; n = 150

$\sqrt{n}R'_{w'y}$	Frequency Observed	Frequency Expected
-1.97 & less	0	1
-1.96 to -1.45	2	2
-1.44 to - .94	2	4
- .93 to - .60	4	4
- .59 to - .32	3	4
- .31 to - .00	8	5
+ .00 to .31	7	5
.32 to .59	1	4
.60 to .93	4	4
.94 to 1.44	4	4
1.45 to 1.96	1	2
1.97 & up	4	1
All observations	40	40.

'Goodness of fit': χ^2_{12} = 16.6 Test of mean: hyp. .000 Test of var.: hyp. 1
One-tailed P: obs. .22 obs. .953
.20 > P > .10 two-tailed P: P = .16 two-tailed P: P = .97

*Universe 2: 3^4 cells; $\underline{R}$ = .453; $\underline{R}'$ = .453; $\underline{R}_{wy}$ = .194; $\underline{R}'_{wy}$ = 0; $\underline{R}'_{w'y}$ = 0.

samples is closer to the hypothesized zero than is the corresponding mean for samples of size 50, and furthermore, it does not differ significantly from zero. The observed variance, .953, is very close to the hypothesized unity. Finally, the extreme tails contain 4, or 10%, of the 40 sample values, closer to the expected 5% than was the case for samples of size 50. Evidently, by the increase in sample size, the sampling distributions are more nearly approximated by the hypothesized distributions, although some bias in the mean and variance may still remain.

5.63. Universe 3

In the third universe, the total relation is $\underline{R} = .534$, greater than in Universe 2. Again as in Universe 2, there is no interaction, as indicated by the fact that $\underline{R}' = \underline{R}$. However, a part of this total relation is due to factor W, as well as factors U and V, as indicated by the universe partial association value of $\underline{R}'_{wy} = .285$. This partial association of W with Y is entirely linear, i. e. $r_{w'\underline{w}} = 1$, as indicated by the fact that $\underline{R}'_{w'y}$ also equals .285. We do not hypothesize that nR'^{2}_{wy} is distributed as a chi-square with two degrees of freedom in this case, because the association existing in the universe will increase the observed values. Rather, we hypothesize that, if the linear component of nR'^{2}_{wy} is removed from nR'^{2}_{wy}, the resulting non-linear component will be distributed asymptotically as a chi-square with one degree of freedom, since there is no non-linear universe partial association. Further, as indicated in 5.4, we hypothesize that $\sqrt{n}$ times the deviation of $R'_{w'y}$ from $\underline{R}'_{w'y}$ is distributed asymptotically as a standard normal deviate.

Table 22 shows the results for the non-linear component of the squared partial association between W and Y for 65 samples, each of size n = 50. Aside from an apparent hiatus in the second class, the fit is very

Table 22

Empirical Sampling Distribution of n Times the Squared Non-Linear Partial Association of W with Y as Compared to χ^2 Expected Frequencies; Universe 3*; n = 50

$n(R'^2_{wy} - R'^2_{w'y})$	Frequency Observed	Frequency Expected
0 - .016	9	6.5
.017 - .064	1	6.5
.065 - .148	7	6.5
.149 - .455	11	13.0
.456 - 1.074	15	13.0
1.075 - 1.642	6	6.5
1.643 - 2.706	8	6.5
2.707 - 3.841	7	3.25
3.842 - 5.412	0	1.95
5.413 & up	1	1.30
All observations	65	65.

'Goodness of fit': $\chi^2_{10} = 13.01$; one-tailed P: $.3 > P > .2$

good ($.3 > P > .2$). Particular attention is called to the behavior of the distribution in the upper tail, in which the observed frequencies compare favorably with expectation.

Results for the observed linear partial association are shown in Table 23. The 'goodness of fit' test yields a non-significant probability between .2 and .1. Nevertheless, it is observed that there appears to be a consistently high concentration of observed frequencies at the center of the distribution, with frequencies at the two tails being consistently less than expectation. The hypothetical mean value of the deviation of $\sqrt{n}R'_{w'y}$ from $\sqrt{n}\underline{R}'_{w'y}$ is, of course, zero; the observed mean of -.052 is quite close to this hypothesized value, such that this discrepancy can be attributed to chance ($P = .58$). But the observed .566 variance, in contrast to the hypothesized unity, is significantly low ($P < .01$), confirming the visual impression of a high concentration of observed frequencies at the center of the distribution. It would appear, therefore, that the bias, if any, is

Table 23

Empirical Sampling Distribution of $\sqrt{n}$ Times the Deviation of the Observed from the Universe Linear Partial Association of W with Y, as Compared to Standard Normal Expected Frequencies; Universe 3*; n = 50

$\sqrt{n}(R'_{w'y} - \underline{R}'_{w'y})$	Frequency	
	Observed	Expected
-1.97 & less	1	1.625
-1.96 to -1.45	4	3.25
-1.44 to - .94	3	6.5
- .93 to - .60	6	6.5
- .59 to - .32	3	6.5
- .31 to - .00	16	8.125
+ .00 to .31	10	8.125
.32 to .59	9	6.5
.60 to .93	7	6.5
.94 to 1.44	4	6.5
1.45 to 1.96	1	3.25
1.97 & up	0	1.625
All observations	65	65.

'Goodness of fit': χ^2_{12} = 17.43	Test of mean: hyp. .000	Test of var.: hyp. 1
One-tailed P:	obs. -.052	obs. .566
.2 > P > .1	two-tailed P: P = .58	two-tailed P: P < .01

*Universe 3: 3^4 cells; $\underline{R}$ = .534; $\underline{R}'$ = .534; $\underline{R}'_{wy}$ = .285; $\underline{R}'_{w'y}$ = .285.

quite small, and that the sampling variance, if not unity, is actually less than hypothesis. Consequently, for this universe, one would feel quite confident in using the standard normal curve as the basis of an interval estimate for the partial linear association.

5.64. Universe 4

None of the preceding universes carries interaction. In order to test the possible influence of universe interaction on sample measures, the fourth universe is constructed to contain a rather high interaction, I = .442. However, all additive partial effects are zero, and all simple effects are zero, such that $\underline{R}'$ = 0, $\underline{R}_{wy}$ = 0, $\underline{R}'_{wy}$ = 0, and $\underline{R}'_{w'y}$ = 0. The two hypotheses with respect to the partial association of W with Y are now that nR'^{2}_{wy}

Table 24

Empirical Sampling Distribution of n Times the Squared Partial Association of W with Y as Compared to χ^2 Expected Frequencies; Universe 4*; n = 50

nR'^2_{wy}	Frequency	
	Observed	Expected
0 - .210	7	7.6
.211 - .445	5	7.6
.446 - .712	6	7.6
.713 - 1.385	9	15.2
1.386 - 2.407	11	15.2
2.408 - 3.218	16	7.6
3.219 - 4.604	7	7.6
4.605 - 5.990	10	3.8
5.991 - 7.823	2	2.3
7.824 & up	3	1.5
All observations	76	76.

'Goodness of fit': $\chi^2_{10} = 25.96$; one-tailed P: $P < .01$

*Universe 4: 3^4 cells; $\underline{R} = .442$; $\underline{R}' = 0$; $\underline{R}_{wy} = 0$; $\underline{R}'_{wy} = 0$; $\underline{R}'_{w'y} = 0$.

is distributed approximately as a chi-square with two degrees of freedom for large enough n, and that $\sqrt{n}R'_{w'y}$ is approximately a standard normal deviate for large enough n. The results for 76 samples, each of size 50, are presented in Tables 24 and 25.

In Table 24, the observed frequencies for smaller values of χ^2_2 are consistently below hypothesis. As χ^2_2 increases toward the upper tail, the observed frequencies become greater than expectation, as a rule. These deviations from the hypothetical distribution indicate a poor fit ($P < .01$).

The deviation of empirical results for the linear partial association in Table 25 is of borderline significance ($.1 > P > .05$) when tested for 'goodness of fit'. The behavior of observed frequencies at the extreme tails is poor. The mean of $\sqrt{n}R'_{w'y}$ is not quite significantly low ($P = .06$); but the variance is well within the realm of chance ($P = .17$). Table 25 gives the impression that the distribution rises to a peak at a negative

Table 25

Empirical Sampling Distribution of $\sqrt{n}$ Times the Linear Partial Association of W with Y as Compared to Standard Normal Expected Frequencies; Universe 4*; n = 50

$\sqrt{n}R'_{w'y}$	Frequency	
	Observed	Expected
-1.97 & less	6	1.9
-1.96 to -1.45	4	3.8
-1.44 to - .94	13	7.6
- .93 to - .60	4	7.6
- .59 to - .32	10	7.6
- .31 to - .00	8	9.5
+ .00 to .31	7	9.5
.32 to .59	6	7.6
.60 to .93	8	7.6
.94 to 1.44	5	7.6
1.45 to 1.96	2	3.8
1.97 & up	3	1.9
All observations	76	76.

'Goodness of fit': χ^2_{12} = 18.83	Test of mean: hyp. .000	Test of var.: hyp. 1
One-tailed P:	obs. -.240	obs. 1.21
.1 > P > .05	two-tailed P: P = .06	two-tailed P: P = .17

*Universe 4: 3^4 cells; $\underline{R}$ = .442; $\underline{R}'$ = 0; $\underline{R}_{wy}$ = 0; $\underline{R}'_{wy}$ = 0; $\underline{R}'_{w'y}$ = 0.

value, rather than the hypothesized zero, and is skewed positively. Apparently, the introduction of interaction in an otherwise null association universe has caused the sample results to deviate slightly more from hypothesis than was the case for Universe 1, which contained no association whatever. The contrived high interaction in the universe appears to exert a relatively weak influence on the non-interactive measures of association.

5.7. An Adjustment for Continuity

Because there apparently are cases, notably Universe 2, for which the assumed asymptotic distributions are rather poor approximations to the actual distributions of sample measures, it is of importance to find some means of improving the approximation. To this end, we employ a concept which is similar in some respects to 'maximum likelihood'; but we are dealing with discrete sampling distributions and unknown population forms, so that the resemblance is purely superficial; we therefore do not claim that the adjustment to be developed makes sample measures unbiased (they are, in fact, biased), and the terms 'efficiency' and 'sufficiency' do not apply. It is merely our purpose to develop a basis, admittedly intuitive, for an adjustment which may give more accuracy to our sample measures. (Here, the term, accuracy, means the reciprocal of the root mean square error of sample estimates from the true universe value.)

Consider a universe consisting of N elements. Let this universe be sampled at a rate, r, such that the distribution of the number of elements, n, which fall into a sample is a binomial, of the form

$$\Pr(n) = [N!/n!(N - n)!]r^n(1 - r)^{N-n} .$$

Now, letting N become large and r become small in such a way that $m = Nr$ remains constant, the limiting distribution is a Poisson, of the form

$$\Pr(n) = (e^{-m}m^n)/n! \quad .$$

(In order to restrict n to values of zero or greater, the following development applies only for values of $m \geq 1$.)

Replacing n! in the expression on the right by $\Gamma(n + 1)$, we have a smooth function of n which passes through each of the points on the discrete Poisson:

$$f(n) = (e^{-m}m^n)/\Gamma(n + 1) \quad .$$

Now, we find the value of $n = n^*$, such that $f(n^*) = f(n^* + 1)$, as follows:

$$f(n^* + 1)/f(n^*) = [e^{-m}m^{n^*+1}/\Gamma(n^*+ 2)]\cdot[\Gamma(n^* + 1)/e^{-m}m^{n^*}] = 1; \quad (36)$$

since $\Gamma(n^* + 2) = (n^* + 1)\Gamma(n^* + 1)$, (36) reduces to

$$m/(n^* + 1) = 1 .$$

So,

$$n^* + 1 = m \text{ , and}$$

$$n^* = m - 1 \text{ .}$$

Now, since both Pr(n) and f(n) have one maximum, and since f(n) passes through the maximum point of Pr(n), the maximum, and only the maximum, of Pr(n) must lie on or between n* and n* + 1, that is

$m - 1 \leq n_{max} \leq m$, where n_{max} is the integer for which Pr(n) is a maximum. Now consider a given sample which contains n elements. If n is n_{max}, then

$$n \leq m \leq n + 1 \text{ .}$$

Thus, if n is n_{max}, the universe m can be anywhere on or between n and (n + 1). For the sake of consistency, we take m to be midway between n and (n + 1), that is $(n + \frac{1}{2})$, and we designate $m_{max} = n + \frac{1}{2}$ as the 'average maximum likelihood' value of m for the given sample.

If we have two universes, such that $N_1/N_2 = C$, then independently selected samples, selected at the sampling rate r, containing n_1 and n_2 elements, resp., form the basis of estimating m_1 and m_2:

$$m_{1\ max} = n_1 + \tfrac{1}{2}$$

$$m_{2\ max} = n_2 + \tfrac{1}{2} \text{ .}$$

Then,

$$(m_{1\ max}/m_{2\ max}) = (rN_1)_{max}/(rN_2)_{max} = C_{max} = (n_1 + \tfrac{1}{2})/(n_2 + \tfrac{1}{2}) \text{ ;}$$

and, in general, for h universes,

$$N_{1\ max}:N_{2\ max} \ldots :N_{h\ max} = (n_1 + \tfrac{1}{2}):(n_2 + \tfrac{1}{2}) \ldots :(n_h + \tfrac{1}{2}) \text{ ,}$$

$$\text{or } P_{1\ max}:P_{2\ max} \ldots :P_{h\ max} = (n_1 + \tfrac{1}{2}):(n_2 + \tfrac{1}{2}) \ldots :(n_h + \tfrac{1}{2}) \text{ ,}$$

where $P_{s\ max} = N_{s\ max}/ N$, s = 1, 2, ..., h.

Thus, for independent samples, one from each of h universes, or sub-classes of a universe, each taken at a constant sampling rate, r, the values of

$(n_s + \frac{1}{2})$, s = 1, 2, ..., h, stand in the same relative magnitude as the corresponding values of $P_{s\ max}$. This result is subject to the restriction that $m_s \geqq 1$ (see previous development, p. 110), so that, in practice, when several sample cell frequencies are zero, the addition of $\frac{1}{2}$ to cell frequencies is probably not advisable.

The correction for continuity, then, involves adding $\frac{1}{2}$ to each cell frequency. In order to evaluate this correction empirically, samples of fixed size are to be taken. This introduces an interrelationship, of the order $(P_s \times P_t)$, for s and t equal 1, 2, ..., h, $s \neq t$, between any pair of sampling distributions. However, it is assumed that, when h is fairly large, such that $(P_s \times P_t)$ is small relative to $(P_s(1 - P_s))$, this interdependence is negligible.

For a 3^4 cell universe, i. e. 81 cells, many if not most of the cells must have a zero frequency for samples as small as size 50. Consequently, the continuity correction could not be validly applied to each cell. However, this does not seem necessary, for, consider equations (17), 4.4. This set of equations determines the additive partial effects of each factor on the basis of all the two-factor marginal cell frequencies that can be constructed, i. e., n_{kl}, ..., n_{km}, ..., n_{lm}, n_{kj}, n_{lj}, ..., n_{mj}. It is these frequencies and the n_k, n_l, ..., n_m, n_j, not the n_{kl_mj}, which determine the partial effects; and in turn the measures of partial association and linear partial association are determined by the partial effects. In general, n_{kl}, etc., are much higher values than n_{kl_mj}, so that if the $+\frac{1}{2}$ correction is applied to these marginal values, the requirement for few zero frequencies of n_{kl}, etc., can be met in the great majority of practical situations.

The rationale for making this correction is as follows: 1) the addition of $\frac{1}{2}$ to each marginal frequency puts the observed frequencies in approximately the same relative magnitude as the 'average maximum likelihood' values of the corresponding universe class proportions; 2) using these adjusted frequencies in the calculation of sample measures of partial association is then equivalent, approximately, to using the 'average maximum likelihood' proportions; 3) if the 'average maximum likelihood' proportions are more accurate estimates than the relative values of the observed uncorrected frequencies, then it is hoped that functions of them, such as our measures of squared partial association, also are more accurate estimates of the universe values.

In order to test whether this correction improves the accuracy of sampling distributions of partial association measures, the continuity correction has been applied to the same samples from the same universes which have been discussed. (See illustration 5.9 for the details of making the continuity correction to the marginal frequencies.) The results are presented in Table 26.

Table 26 shows that the adjustment for continuity has improved the fit of the observed squared partial association. For Universe 1, the 'goodness of fit' chi-square value is reduced from 8.65 to 4.36; for Universe 2, from 17.5 to 5.23; for Universe 3, from 13.0 to 3.13; for Universe 4, from 26.0 to 4.81. The adjusted values show a very nice fit to the hypothesized distribution, as indicated by 'goodness of fit' probabilities in excess of .8 in every case.

Table 27, for the linear partial association, gives less conclusive results for the over-all fit. For Universe 1, the 'goodness of fit' chi-square value increases slightly from 9.57 to 10.4; for Universe 2, it

Table 26

Frequency Distributions of Observed Chi-Square Values, Formed from Measures of Partial Association, for Repeated Random Samples of Various Universes; Comparison of Theoretically Expected Frequencies with Observed Frequencies of Unadjusted and Adjusted Chi-Square Values

A

χ^2 Class Intervals Expressed in Percentiles	U_1			U_2			U_3			U_4		
	Un-adj.	Adj.	Exp.	Un-adj.	Adj.	Exp.	Un-adj.	Adj.	Exp.	Un-adj.	Adj.	Exp.
0 up to 10	7	6	6.5	6	8	9.1	9	6	6.5	7	9	7.6
10 up to 20	7	6	6.5	8	7	9.1	1	7	6.5	5	4	7.6
20 up to 30	11	10	6.5	5	9	9.1	7	5	6.5	6	5	7.6
30 up to 50	13	12	13.0	20	17	18.2	11	13	13.0	9	13	15.2
50 up to 70	9	15	13.0	17	18	18.2	15	16	13.0	11	18	15.2
70 up to 80	8	3	6.5	7	13	9.1	6	6	6.5	16	9	7.6
80 up to 90	2	6	6.5	15	9	9.1	8	7	6.5	7	9	7.6
90 up to 95	4	3	3.2	4	4	4.6	7	4	3.2	10	5	3.8
95 up to 98	3	2	2.0	3	2	2.7	0	1	2.0	2	2	2.3
98 to 100	1	1	1.3	6	4	1.8	1	0	1.3	3	2	1.5
All Values	65	65	65.	91	91	91.	65	65	65.	76	76	76.

B

Universe	'Goodness of fit'			
	Unadjusted Distribution		Adjusted Distribution	
	χ^2_{10}	Probability	χ^2_{10}	Probability
U_1	8.65	$.70 > P > .50$	4.36	$.95 > P > .90$
U_2	17.5	$.10 > P > .05$	5.23	$.90 > P > .80$
U_3	13.0	$.30 > P > .20$	3.13	$.98 > P > .95$
U_4	26.0	$P < .01$	4.81	$.95 > P > .90$

Table 27

Frequency Distributions of Observed 'Normal Deviates', Formed from Measures of Linear Partial Association, for Repeated Random Samples of Various Universes; Comparison of Theoretically Expected Frequencies with Observed Frequencies of Unadjusted and Adjusted 'Normal Deviates'

A

Normal Deviate Class Intervals Expressed in Percentiles	U_1			U_2			U_3			U_4		
	Un-adj.	Adj.	Exp.	Un-adj.	Adj.	Exp.	Un-adj.	Adj.	Exp.	Un-adj.	Adj.	Exp.
0 up to 2.5	1	1	1.6	6	1	2.3	1	1	1.6	6	2	1.9
2.5 up to 7.5	3	1	3.2	5	7	4.6	4	4	3.2	4	6	3.8
7.5 up to 17.5	6	8	6.5	8	7	9.1	3	5	6.5	13	12	7.6
17.5 up to 27.5	11	11	6.5	3	6	9.1	6	6	6.5	4	7	7.6
27.5 up to 37.5	3	4	6.5	5	1	9.1	3	5	6.5	10	10	7.6
37.5 up to 50.0	7	6	8.1	6	9	11.4	16	19	8.1	8	8	9.5
50.0 up to 62.5	12	12	8.1	7	9	11.4	10	9	8.1	7	6	9.5
62.5 up to 72.5	7	7	6.5	8	4	9.1	9	8	6.5	6	8	7.6
72.5 up to 82.5	6	7	6.5	14	17	9.1	7	6	6.5	8	7	7.6
82.5 up to 92.5	3	3	6.5	10	10	9.1	4	1	6.5	5	6	7.6
92.5 up to 97.5	4	4	3.2	12	16	4.6	1	1	3.2	2	1	3.8
97.5 to 100.	2	1	1.6	7	4	2.3	0	0	1.6	3	3	1.9
All Values	65	65	65.	91	91	91.	65	65	65.	76	76	76.

B

	'Goodness of fit'			
Universe	Unadjusted Distribution		Adjusted Distribution	
	χ^2_{12}	Probability	χ^2_{12}	Probability
U_1	9.57	.70 > P > .50	10.4	.70 > P > .50
U_2	31.2	P < .01	51.4	P < .01
U_3	17.4	.20 > P > .10	24.0	.05 > P > .02
U_4	18.8	.10 > P > .05	11.4	.50 > P > .30

increases from 31.2 to 51.4, both values being highly significant; for Universe 3, it increases from 17.4 to 24.0; and for Universe 4, it decreases from 18.8 to 11.4. Thus, if χ^2_{12} is taken as an over-all indicator of goodness of fit, the adjusted values are generally more at variance with hypothesis than the unadjusted values.

However, in making tests of significance and interval estimates, we are most frequently concerned with one tail or another, or both tails, of the sampling distribution. Consequently, it is not so important that the form of the hypothetical sampling distribution be as good an approximation at the central values as it be a good approximation at the tails. For the distributions in Table 26, we are concerned most with the upper tail, let us say the upper 10%. For the distributions in Table 27, we are concerned most with the upper and lower tails, let us say the lower 7.5% and the upper 7.5%. Examination of Tables 26 and 27 reveals that in general the unadjusted frequencies tend to be higher than expectation, and that the adjustment tends to reduce the frequencies in the tails. This general improvement is illustrated in Table 28, in which the upper 10% frequencies from Table 26 and the upper and lower 7.5% frequencies from Table 27 have been pooled from all four sampling distributions.

The unadjusted frequencies for the upper tail of the hypothetical chi-square distribution run significantly high ($P < .01$), while the adjusted frequencies fit very nicely ($.90 > P > .80$). A similar improvement for the fit of adjusted frequencies is seen in the tails of the hypothetical normal, from high unadjusted frequencies of borderline significance ($.05 > P > .02$) to a fairly good fit for the adjusted frequencies ($.30 > P > .20$), due principally to improvement of fit at the extreme tails (the 0 to 2.5 and the 97.5 to 100 percentile classes).

Table 28

Comparison of Theoretical Expected Tail Frequencies with Observed Tail Frequencies from Unadjusted and Adjusted Samples of Size n = 50, from U_1, U_2, U_3, and U_4 Combined

10% Upper Tail Chi-square Class Intervals, in Percentiles (from Table 26)	Unadj.	Adj.	Exp.	'Goodness of fit'	
				Unadj.	Adj.
90 up to 95	25	16	14.9	$\chi^2_3 = 11.4$	$\chi^2_3 = 0.62$
95 up to 98	8	7	8.9		
98 to 100	11	7	5.9	$P < .01$	$.90 > P > .80$
7.5% Upper and Lower Normal Deviate Class Intervals, in Percentiles (from Table 27)					
0 up to 2.5	14	5	7.4	$\chi^2_4 = 9.82$	$\chi^2_4 = 4.95$
2.5 up to 7.5	16	18	14.9		
92.5 up to 97.5	19	22	14.9	$.05 > P > .02$	$.30 > P > .20$
97.5 to 100	12	8	7.4		

The general conformance of the adjusted measures to the hypothesized distribution for most practical purposes can be illustrated as follows. A total of 297 samples, each of size 50, have been taken from four different universes. If the adjusted partial association for each of these samples had been tested by an appropriate one-tail chi-square test of the null (true) hypothesis, the null (true) hypothesis would have been rejected 14 times, that is, 4.7% of the time, on the nominal 5% significance level. This represents excellent agreement. If the adjusted linear partial association had been tested by a two-tail normal distribution, on the nominal 5% level, the null (true) hypothesis would have been rejected 13 times, that is, 4.4% of the time. Or, viewing the linear association as an estimating problem, if a nominal 95% confidence interval had been constructed from each sample, 95.6% of such intervals would have included the true value. Again, agreement is very good.

5.8. Summary

If the dependent classification in a multiply classified universe can be characterized by a set of quantities, $\underline{y}_j$, $j = 1, \ldots, q$, with zero mean and unit variance, then the total universe association, $\underline{R}$, is given by the standard deviation of the universe means (appropriately weighted by class frequencies):

$$\underline{R} = \sqrt{\operatorname{var} \overline{\underline{y}}_i} \;.$$

The additive, i. e. non-interactive, universe association, $\underline{R}'$, is given by the standard deviation of the sum of the adjusted universe means:

$$\underline{R}' = \sqrt{\operatorname{var}(\overline{\underline{y}}_{kadj} + \overline{\underline{y}}_{ladj} + \cdots + \overline{\underline{y}}_{madj})} \;.$$

The partial association due to a given factor, say W, is given by

$$\underline{R}'_{wy} = \sqrt{\operatorname{var} \overline{\underline{y}}_{madj}} \;,$$

that is, the standard deviation of the adjusted means for the categories of W. The linear partial association due to factor W is given when a set of quantities, w'_m, is assigned to the categories of W. Denoting $r_{w'\underline{w}}$ as the correlation between w'_m and the universe partial effects, $\underline{w}_m$, the universe linear partial association, $\underline{R}'_{w'y}$, is given by

$$\underline{R}'_{w'y} = \underline{R}' r_{w'\underline{w}} \sqrt{\operatorname{var} \underline{w}_m} \;.$$

When random samples of size $n = 50$ were taken, and when a skew correction was made by adding $\frac{1}{2}$ to each two-factor contingency cell frequency in each sample, the observed values of nR'^{2}_{wy} were found, by empirical trials from a variety of universes, to be distributed approximately as chi-square with $(c - 1)$ degrees of freedom, when the universe value of $\underline{R}'_{wy} = 0$. The observed values of $\sqrt{n}R'_{w'y}$ were found to be approximately normally distributed, with unit variance, about the true universe value, $\sqrt{n}\underline{R}'_{w'y}$.

Thus, for practical purposes, the observed value of the partial association may be tested for significance by a chi-square test, and the observed linear partial association may either be given a normal test of significance or may be the basis of a confidence interval estimate of the corresponding universe parameter.

These conclusions must be qualified in several respects. Firstly, while a very good fit was observed in the tails of the distributions, the approximation by hypothetical chi-square and normal distributions was not always good in the mid-portions of the distributions. This would limit the usefulness of the approximations if high significance levels, such as 15% or more, were to be used; but for the usual significance levels of 5% or less, the chi-square or normal assumption appears to be well approximated.

Secondly, the continuity correction can be validly applied only when expected two-factor cell frequencies are $\geqq 1$. As a practical rule, then, if several observed two-factor cell frequencies were zero, the continuity correction could not be applied. This limitation implies that when very high associations among the dependent and independent variables are encountered, the continuity correction can not be applied to samples of such small size as 50. This is because very high associations can be achieved only through the existence of several very low two-factor cell frequencies. This does not seem to be a very strong limitation, however, since very high associations are seldom encountered in sample surveys.

Thirdly, the empirical results presented here necessarily cover a limited number of possible universes. An attempt has been made to employ a variety of different universes, but it is quite possible that some universes could be found for which the empirical sampling distribution would be poorly approximated by the hypothetical chi-square and normal curves. There is no question that, in the future, more experimentation should be done on universes with a greater number of classification factors, and with a greater variety of factor interrelationships, than has been done here.

5.9. Illustration: Computation of the Linear Partial Association of First and Second Year Health Problems in Sampled Households of a Specific Size, Average Age, and Sex Distribution Group

The manner in which computations are made for sample measures of partial association can be illustrated with sample data from the Arsenal Study, section 2. In 2.42, Tables 7 and 8 show the unadjusted and 'adjusted' frequencies, resp., of households with and without health problems in the first and second years, for households of size one, age 30-44, male. Table 8 has been adjusted for continuity and for the effects of strata and interviewer groups. The details of making these adjustments have been deferred until now because they would not be fully understood without first knowing of the procedures developed in sections 4 and, this section, 5.

The first step is to form a classification of the 30 households in the size-age-sex specific group according to stratum, interviewer group, first year health problems and second year health problems. This classification is shown in Table 29. Note that Table 29 is a specific instance of the general table shown in section 4.1, Table 14. Thus, with reference to Table 14, $n_{1111} = 0$, $n_{1122} = 2$, $n_{...2} = 10$, etc. Note also that Table 7 can be formed from Table 29 by obtaining $n_{..11}$, $n_{..21}$, $n_{..12}$, and $n_{..22}$. Thus, Table 7 is a specific instance of the general marginal table, Table 15, section 4.5. Computations for the unadjusted association in Table 7 are given in 2.42, and similar computations are illustrated more fully in 3.6.

Table 29

Classification of Propositus Households of Size One, Age 30-44, Male, According to Health Problem Status at First and Second Interview, Within Categories of Stratum and Interviewer Characteristics

U	V	W	Y		
Stratum (k)	Interviewer Group (l)	Hlth. Pbs. First Interview (m)	Hlth. Pbs. 2nd Interview (j)		Total
			Zero	Non-zero	
I	AB	Zero	0	0	0
		Non-zero	7	2	9
	CD	Zero	0	0	0
		Non-zero	1	2	3
II	AB	Zero	0	1	1
		Non-zero	6	2	8
	CD	Zero	2	0	2
		Non-zero	1	1	2
III	AB	Zero	0	0	0
		Non-zero	2	2	4
	CD	Zero	0	0	0
		Non-zero	1	0	1
Total			20	10	30

Now, returning to Table 29, the problem is to obtain the measure of partial relation between first and second year health problems, and to obtain the corresponding Table 8. The procedures which follow would be found quite laborious, if done by hand, but, being done by the IBM 650 computer, they take only about one minute.

As the first step in finding the adjusted, that is partial, serial relation, all the two-factor tables of the data of Table 29 are formed to give n_{kl}, n_{km}, n_{lm}, n_{kj}, n_{lj}, and n_{mj}, as shown in Table 30.

Table 30

(A)
n_{kl} for
factors U and V

k \ l	1	2	n_k
1	9	3	12
2	9	4	13
3	4	1	5
n_l	22	8	30

(B)
n_{km} for
factors U and W

k \ m	1	2	n_k
1	0	12	12
2	3	10	13
3	0	5	5
n_m	3	27	30

(C)
n_{lm} for
factors V and W

l \ m	1	2	n_l
1	1	21	22
2	2	6	8
n_m	3	27	30

(D)
n_{kj} for
factors U and Y

k \ j	1	2	n_k
1	8	4	12
2	9	4	13
3	3	2	5
n_j	20	10	30

(E)
n_{lj} for
factors V and Y

l \ j	1	2	n_l
1	15	7	22
2	5	3	8
n_j	20	10	30

(F)
n_{mj} for
factors W and Y

m \ j	1	2	n_m
1	2	1	3
2	18	9	27
n_j	20	10	30

Next, the continuity adjustment is made by adding a frequency of .5 to each two-factor cell frequency, which yields the following adjusted tables:

Table 31

(A)
Adjusted n_{kl}

k \ l	1	2	Tot.
1	9.5	3.5	13
2	9.5	4.5	14
3	4.5	1.5	6
Tot.	23.5	9.5	33

(B)
Adjusted n_{km}

k \ m	1	2	Tot.
1	.5	12.5	13
2	3.5	10.5	14
3	.5	5.5	6
Tot.	4.5	28.5	33

(C)
Adjusted n_{lm}

l \ m	1	2	Tot.
1	1.5	21.5	23
2	2.5	6.5	9
Tot.	4	28	32

(D)
Adjusted n_{kj}

k \ j	1	2	Tot.
1	8.5	4.5	13
2	9.5	4.5	14
3	3.5	2.5	6
Tot.	21.5	11.5	33

(E)
Adjusted n_{lj}

l \ j	1	2	Tot.
1	15.5	7.5	23
2	5.5	3.5	9
Tot.	21	11	32

(F)
Adjusted n_{mj}

m \ j	1	2	Tot.
1	2.5	1.5	4
2	18.5	9.5	28
Tot.	21	11	32

Note that frequencies in those tables involving factor U, index k, total 33 whereas the other tables total 32. This is because factor U has three levels rather than two levels as for each of V, W and Y. In order to put all the cell entries on a comparable relative basis, entries in tables which total 33 are multiplied by 30/33, while the other table entries are multiplied by 30/32. This would not have been necessary if all factors had an equal number of categories.

The following tables have been adjusted by the appropriate factor:

Table 32

(A)

Adjusted n_{kl}

k l	1	2	Total
1	8.64	3.18	11.82
2	8.64	4.09	12.73
3	4.09	1.36	5.45
Tot.	21.37	8.63	30

(B)

Adjusted n_{km}

k m	1	2	Total
1	.45	11.37	11.82
2	3.18	9.55	12.73
3	.45	5.00	5.45
Tot.	4.08	25.92	30

(C)

Adjusted n_{lm}

l m	1	2	Total
1	1.41	20.16	21.57
2	2.34	6.09	8.43
Tot.	3.75	26.25	30

(D)

Adjusted n_{kj}

k j	1	2	Total
1	7.73	4.09	11.82
2	8.64	4.09	12.73
3	3.18	2.27	5.45
Tot.	19.55	10.45	30

(E)

Adjusted n_{lj}

l j	1	2	Total
1	14.53	7.03	21.56
2	5.15	3.29	8.44
Tot.	19.68	10.32	30

(F)

Adjusted n_{mj}

m j	1	2	Total
1	2.34	1.41	3.75
2	17.34	8.91	26.25
Tot.	19.68	10.32	30

In solving for the partial effects of categories of U, V and W, it is necessary to use n_k, n_l, n_m, and n_j, in addition to the two-factor cell frequencies, in order to satisfy scale requirements that the mean partial effects be zero. Therefore, the relative marginal frequencies are also adjusted for continuity, as for example:

Adjustment of n_k

k	Unadjusted	Adjusted	Adjusted x 30/31.5
1	12	12.5	11.90
2	13	13.5	12.86
3	5	5.5	5.24
Tot.	30	31.5	30.

Similar adjustment for all marginal frequencies yields:

Table 33

Specific Index	n_k	n_l	n_m	n_j
1	11.90	21.77	3.39	19.84
2	12.86	8.23	26.61	10.16
3	5.24	X	X	X
Tot.	30.	30.	30.	30.

The values from Tables 32 and 33 may now be substituted into the normal equations for the partial effects, developed in section 4.4:

$$\left.\begin{aligned} &R'u_k + R'\overline{v}_k + R'\overline{w}_k = \overline{y}_k\ , \quad k = 1,\ 2. \\ &\sum n_k u_k = 0 \\ &R'\overline{u}_l + R'v_l + R'\overline{w}_l = \overline{y}_l\ , \quad l = 1. \\ &\sum n_l v_l = 0 \\ &R'\overline{u}_m + R'\overline{v}_m + R'w_m = \overline{y}_m\ , \quad m = 1. \\ &\sum n_m w_m = 0 \end{aligned}\right\} ,$$

where u_k, v_l, and w_m are scores for the factors U, V, and W. To determine $\overline{y}_k$, $\overline{y}_l$, and $\overline{y}_m$, it is first necessary to calculate y_j, j = 1, 2. Since we take n_j = 19.84 and 10.16 for j equal one and two, resp., (from Table 33)

$$\left.\begin{aligned} y_1 &= -\sqrt{10.16/19.84} = -\ .716 \\ y_2 &= +\sqrt{19.84/10.16} = +\ 1.40 \end{aligned}\right\} \quad \text{as illustrated in 3.6.}$$

Now, $\overline{y}_k = \sum_j n_{kj} y_j / \sum_j n_{kj}$, so

$$\overline{y}_{k\ (k=1)} = [7.73(-.716) + 4.09(1.40)]/(7.73 + 4.09) = +.0162.$$

In similar fashion, the remaining values of $\overline{y}_k$, $\overline{y}_l$, and $\overline{y}_m$ are determined.

Substituting values from Tables 32 and 33 and the values of $\overline{y}_k$, $\overline{y}_l$, and $\overline{y}_m$ into the normal equations and simplifying, we get a set of seven simultaneous linear relations. The solution of these equations yields values of $R'u_k$, $R'v_l$, and $R'w_m$. At this point, R' could be found by computing the standard deviation of the sums, $R'u_k + R'v_l + R'w_m$ (see section 4.7); however, since by (33), section 5.3, $R'_{wy} = R's_{w_m}$, where R'_{wy} is the observed partial correlation between W and Y, and s_{w_m} is the standard deviation of w_m, we obtain the observed partial correlation between W and Y directly by computing the standard deviation of $R'w_m$. Also, since there are only two categories of W in the present illustration, $R'_{wy} = \left| R'_{w'y} \right|$, that is R'_{wy} is equal to the absolute value of the linear partial association of an arbitrary set of values, w'_m, with y_j. The sign of the linear relation is given by $r_{w'w}$, which, in the case of two categories of w, is either plus one or minus one. Thus, the partial correlation between W and Y, together with sign, is determined.

The details of solving the normal equations and subsequent computations are not given here because they are lengthy, and because they involve standard mathematical operations. However, it should be appreciated that all the computations, beginning with the formation of all two-factor table, followed by the various adjustments for continuity, computation of $\overline{y}_k$, $\overline{y}_1$, and $\overline{y}_m$, solution of a set of linear equations with seven unknowns, and computation of the partial association measure would involve a great deal of labor if done by hand or by the aid of a desk calculator. Furthermore, these procedures yield the measure of partial serial association for only one size-age-sex class of household. In the Arsenal Study, there are 22 such calculated measures. Actually, all these computations are performed on the IBM 650 computer, and some realization of the efficiency of using this labor-saving device can be had by knowing that all 22 measures were obtained in less than 30 minutes of computing time.

In the particular size-age-sex class we have chosen for illustration, $R'_{w'y}$ turns out to be -.025, or -2.5%. The final step, if desired for the sake of presentation, is to develop Table 8. This is done simply by setting up the four-fold table, entering the total frequency, n = 30, and the marginal frequencies, n_m and n_j (adjusted for continuity per Table 33). Then the four cell frequencies in Table 8 are uniquely determined by requiring them to be such that the product moment correlation is equal to -.025. These cell frequencies are readily calculated by the use of the identity (10), section 3.6, plus the requirement that the cell frequencies must add to the marginal totals. Alternatively, any of the several identities illustrated in section 3.6 may be used.

The same procedures as outlined above are carried out for each specific sub-group (α) of the sample. Thus, for example, we have entered -.025 as the partial serial relation for the $\alpha = 13$ group, in Table 9, section 2.42, along with all the other partial relations for the remaining 21 sub-groups of the Arsenal data. (Note that in Table 9 and elsewhere in section 2, the linear partial association is denoted r_α, whereas in section 5 we have used the notation, $R'_{w'y}$, to denote the linear partial association of the specific factor, W, with Y. r_α is identical to $R'_{w'y}$ for any given sub-group, α, where it is understood that factor W is first year health status.)

5.10. Illustration: Analysis of Variation in the Partial Serial Relation of Household Health Problems

It has been shown that the observed linear partial association, when adjusted for continuity, is distributed approximately as a normal distribution with variance equal to $1/n_\alpha$, where n_α is the size of a simple random sample from the αth population. (However, see 5.8 for qualifications.) This statistical property proves very useful in the analysis of variations in the partial relation among a number of sub-populations. For an illustration of such an analysis, the reader is referred to section 2, particularly sub-sections 2.4 and 2.5. In those two sub-sections, emphasis was placed on the subject matter and on the analytical approach rather than on the detailed computations involved. These details have been reserved until now in order to show that each part of the analysis, i. e., the analysis of total variation, the test of significance (or confidence interval) for the mean partial relation, the test of significance of variations about the mean, the test of significance for the agreement of an a priori causal hypothesis, and the test of significance for residual variations, is a part of a unified whole. The reader who is familiar with the routine analysis of variance in experiments will find the following analysis closely analogous to that unified technique. In particular, the decomposition of a total chi-square into component parts is accomplished by exactly the same computations as in the analysis of experimental data. The only technical difference is that, in the usual experiment, the 'error' variance is estimated, and ratios of two chi-squares are formed in 'F' tests of significance, while in the following analysis, the 'error' variance (of measures of partial association) is 'known', so that chi-square tests of significance are performed.

The basic computations are set down in Table 34. α denotes a particular size-age-sex specific sub-population of the Arsenal Health District; n_α is the observed number of sample elements from the αth sub-population; and r_α is the observed linear partial association [previously denoted $R'_{w'y}$ in section 5; hence $r_\alpha = (R'_{w'y})_\alpha$]. These are shown in the first three columns of Table 34. (See also Table 9, section 2.42.) In the fourth column of Table 34, C_α is listed. This is the index of a priori hypothesized level of partial relation, as developed in section 2.52. (See also Table 10.)

Column (5) is the product of entries in columns (2) and (3), $n_\alpha r_\alpha$. Column (6) is the product of (2) and (4), $n_\alpha C_\alpha$. Column (7) is (3) x (5), $n_\alpha r_\alpha^2$.

Table 34

Basic Computations for Analysis of Variations

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
				(2)x(3)	(2)x(4)	(3)x(5)	(4)x(6)	(4)x(5)
α	n_α	r_α	C_α	$n_\alpha r_\alpha$	$n_\alpha C_\alpha$	$n_\alpha r_\alpha^2$	$n_\alpha C_\alpha^2$	$n_\alpha C_\alpha r_\alpha$
1	33	-.322	3	-10.626	99	3.422	297	- 31.878
2	96	-.208	1	-19.968	96	4.153	96	- 19.968
3	57	-.175	1	- 9.975	57	1.746	57	- 9.975
4	179	-.114	6	-20.406	1074	2.326	6444	-122.436
5	117	-.109	5	-12.753	585	1.390	2925	- 63.765
6	92	-.085	4	- 7.820	368	.665	1472	- 31.280
7	104	-.074	1	- 7.696	104	.570	104	- 7.696
8	97	-.046	4	- 4.462	388	.205	1552	- 17.848
9	124	-.043	3	- 5.332	372	.229	1116	- 15.996
10	158	-.043	0	- 6.794	0	.292	0	0
11	135	-.040	2	- 5.400	270	.216	540	- 10.800
12	99	-.038	2	- 3.762	198	.143	396	- 7.524
13	30	-.025	5	- .750	150	.019	750	- 3.750
14	36	.000	10	.000	360	.000	3600	0
15	42	.013	7	.546	294	.007	2058	3.822
16	54	.029	4	1.566	216	.045	864	6.264
17	35	.038	9	1.330	315	.051	2835	11.970
18	59	.045	8	2.655	472	.119	3776	21.240
19	72	.052	12	3.744	864	.195	10368	44.928
20	97	.078	6	7.566	582	.590	3492	45.396
21	45	.101	5	4.545	225	.459	1125	22.725
22	143	.196	0	28.028	0	5.493	0	0
Sum X	1904	X	X	-65.764	7089	22.335	43867	-186.371
Average				r_{ave}	C_{ave}			
				-.03454	3.7232	.01173	23.039	-.09788
Product Average				r_{ave}^2	C_{ave}^2			$(r_{ave})(C_{ave})$
				.001193	13.861			- .1286

Column (8) is (4) x (6), $n_\alpha C_\alpha^2$. Column (9) is the covariance column, the product of (4) and (5), $n_\alpha C_\alpha r_\alpha$.

There are 22 rows, one for each of the 22 sub-populations under analysis. The 22 entries in each column, with the exception of columns (1), (3) and (4), are summed to give the 'Sum' row. Thus, $\sum n_\alpha = n = 1904$; $\sum n_\alpha r_\alpha = -65.764$; $\sum n_\alpha C_\alpha = 7089$; $\sum n_\alpha r_\alpha^2 = 22.335$; $\sum n_\alpha C_\alpha^2 = 43867$; and $\sum n_\alpha C_\alpha r_\alpha = -186.371$. Now columns (5) through (9) are divided by n = 1904 to give the 'Average' row. Thus, $\sum n_\alpha r_\alpha / n = r_{ave} = -.03454$; $\sum n_\alpha C_\alpha / n = C_{ave} =$

3.7232; $\sum n_\alpha r_\alpha^2/n$ = .01173; $\sum n_\alpha C_\alpha^2/n$ = 23.039; and $\sum n_\alpha C_\alpha r_\alpha/n$ = -.09788. Below the 'Average' row, the 'Product Averages', r_{ave}^2, C_{ave}^2, and $(r_{ave})(C_{ave})$ appear.

Note that r_{ave} = -.03454 = -3.5%, and this appears in section 2.42, p. 41, as the adjusted relation of household health problems, adjusted for household size, age, sex and for stratum and interviewer characteristics.

Also, note that $\sum n_\alpha r_\alpha^2$ equals 22.335. This is distributed, under the null hypothesis, as a chi square with 22 degrees of freedom and is found to be non-significant. Thus, the 22 partial relations do not vary significantly from zero. $\sum n_\alpha r_\alpha^2$ = 22.34 appearsin section 2.5, p. 45.

Now a test of the significance of the mean, r_{ave}, can be performed. The total chi square with 22 degrees of freedom is made up partly of variation due to the mean and partly of variation about the mean. Variation due to the mean is

nr_{ave}^2 = 1904(.001193) = 2.271. This is a chi square with one degree of freedom under the null hypothesis, and it is not found to be significant on the 5% level. Thus, the observed mean, r_{ave} = -3.5%, is not significantly different from zero. This fact is reported in section 2.42, p. 41.

Now the variation about the mean is found by subtracting variation due to the mean from the total variation. Hence, variation about the mean is

$\sum n_\alpha r_\alpha^2 - nr_{ave}^2$ = 22.34 - 2.27 = 20.07, and is distributed as a chi square with 22 minus 1 = 21 degrees of freedom under the null hypothesis that there is no universe variation about the mean. This component of chi square is shown in 2.5, p. 45, where it is noted that again, on referring 20.07 to a chi square table with 21 degrees of freedom, the observed variation about the mean is far from significant.

Now we wish to decompose the variation about the mean, 20.07, into two components: the first component, with one degree of freedom, is that portion of the variation which is 'explained' by the a priori hypothesis; the second component, with 20 degrees of freedom, is the residual variation 'unexplained' by the a priori hypothesis. First, we obtain the slope of the regression of r_α on C_α by:

$$\text{slope} = \text{cov}(r_\alpha C_\alpha)/s_{C_\alpha}^{\;2} = (\Sigma n_\alpha C_\alpha r_\alpha/n - r_{ave}C_{ave})/(\Sigma n_\alpha C_\alpha^{\;2}/n - C_{ave}^{\;2})$$

$$= [-.0979 - (-.1286)]/(23.039 - 13.861) = .0307/9.178$$

$= .00334$. This slope, .00334, is reported in section 2.52, p. 51, and is the slope of the line of agreement plotted in Figure 10, p. 50. In order to determine the amount of variation in r_α due to the slope of the line of agreement, it is merely necessary to multiply the squared covariance of $r_\alpha C_\alpha$ by n, and divide by the variance of C_α, thus:

$n(\text{cov}\; r_\alpha, C_\alpha)^2/s_{C_\alpha}^{\;2} = 1904(.0307)^2/9.178 = .195$. This component of variance, .195, is distributed as a chi square with one degree of freedom on the null hypothesis. The test of significance, however, should not be performed by reference to a chi square distribution, because such a test is two tailed. In the case of testing the a priori hypothesis for agreement with observations, we are concerned with discriminating between a positive slope (agreement) and a negative slope (disagreement); hence we wish to test whether the slope is significantly positive. Therefore, taking $\sqrt{.195} = +.44$, we get a standard normal deviate (positive because the slope is positive). On referring +.44 to a normal table, we find that a value of +.44 would be exceeded 33 times in 100. Thus, the slope is far from being significantly positive, and it is concluded that the amount of agreement of a priori hypothesis with observed variations is not significant. This result is reported in 2.52, p. 51.

Finally, the residual variation, not accounted for by hypothesis is 20.07 - .195 = 19.87. On the null hypothesis, this is distributed as a chi square with 21 minus one, or 20 degrees of freedom, and it is found to be far from significant. Then we conclude that there is not a significantly great variation left unexplained by the *a priori* hypothesis. This result, too, is reported in 2.52, p. 51. Had the variation 'explained' by the *a priori* hypothesis been significant, this latter test of the 'unexplained' residual would have been the criticle test for discriminating between a causal and a non-causal interpretation. (See the fourth criterion, 2.52, p. 47 and p. 51.) A non-significant residual would have favored a causal interpretation of the significant agreement of the *a priori* hypothesis; whereas a significant residual would have favored the interpretation that the *a priori* hypothesis had merely 'picked up' a portion of the non-causal variation.

The analogy between the foregoing analysis and the analysis of variance can be clearly brought out by summarizing in a 'table of variations':

Table 35

Analysis of Variations in r_α

Source of Variation	d. f.	Observed Chi Square	Probability
Total $(\sum n_\alpha r_\alpha^2)$	22	22.34	.5 > P > .3 not sig.
Due to mean(nr_{ave}^2)	1	2.27	.2 > P > .1 not sig.
About the mean	21	20.07	.7 > P > .5 not sig.
Due to hypothesis	1	.20	(1 tailed normal) P = .33 not sig.
Residual	20	19.87	.5 > P > .3 not sig.

The computations for the *a posteriori* analysis are quite similar to those for the above analysis. On reference now to section 2.53, p. 52, it is seen that household groups in the 15-29 year average age groups were eliminated from the *a posteriori* comparison of predominantly female households with the households which were not predominantly female. Table 36 shows the basic computation for the 15 groups involved in the comparison.

Table 36

Basic Computations for <u>A Posteriori</u> Analysis of Variations in r_α

	α	n_α	r_α	C_α	$n_\alpha r_\alpha$	$n_\alpha C_\alpha$	$n_\alpha r_\alpha^2$	$n_\alpha C_\alpha^2$	$n_\alpha C_\alpha r_\alpha$
'Female' Households									
	1	36	.000	+1	.000	36	.000	36	.000
	2	42	.013	+1	.546	42	.007	42	.546
	3	35	.038	+1	1.330	35	.051	35	1.330
	4	72	.052	+1	3.744	72	.195	72	3.744
	5	97	.078	+1	7.566	97	.590	97	7.566
	6	45	.101	+1	4.545	45	.459	45	4.545
			Female Sub-total			327			17.731
				Female average					+ .054
'Male' Households									
	7	33	-.322	-1	-10.626	- 33	3.422	33	10.626
	8	96	-.208	-1	-19.968	- 96	4.153	96	19.968
	9	57	-.175	-1	- 9.975	- 57	1.746	57	9.975
	10	179	-.114	-1	-20.406	-179	2.326	179	20.406
	11	97	-.046	-1	- 4.462	- 97	.205	97	4.462
	12	124	-.043	-1	- 5.332	-124	.229	124	5.332
	13	99	-.038	-1	- 3.762	- 99	.143	99	3.762
	14	30	-.025	-1	- .750	- 30	.019	30	.750
	15	59	+.045	-1	2.655	- 59	.119	59	-2.655
			Male Sub-total			-774			72.626
				Male average					- .094
Sum	X	1101	X	X	-54.895	-447	13.664	1101	90.357
Average					- .0499	-.406	.0124	1	.0821
Product average					.0025	.1648			.0203

The rows and columns of Table 36 are set up just as for Table 34. However, in order to show an average for the 'female' and for the 'male' households, the table is divided into two parts. Note also that C_α is now +1 for the 'female' households and -1 for the 'male' households. This, of course, conforms to the <u>a posteriori</u> hypothesis that the serial relation for 'female' households is greater than for 'males'. Any two numbers other than +1 and -1 could have been selected and computations would have yielded valid tests, but by selecting unit values of C_α, the computation of separate means for 'male' and 'female' households is facilitated. The average relation for 'female' households is given by $\sum_{\alpha=1}^{6} n_\alpha C_\alpha r_\alpha / \sum_{\alpha=1}^{6} n_\alpha C_\alpha = \sum_{\alpha=1}^{6} n_\alpha r_\alpha / \sum_{\alpha=1}^{6} n_\alpha = .054$; on the other hand, for 'males', the average is $\sum_{\alpha=7}^{15} n_\alpha C_\alpha r_\alpha / \sum_{\alpha=7}^{15} n_\alpha C_\alpha = \sum_{\alpha=7}^{15} n_\alpha r_\alpha / \sum_{\alpha=7}^{15} n_\alpha = -.094$. These two values are reported and discussed in 2.53, p. 53. The analysis of variations for the _a posteriori_ analysis is constructed in precisely the same manner as before, and this is presented in Table 37.

Table 37

A Posteriori Analysis of Variations in r_α

Source of Variation	d. f.	Observed Chi Square	Probability
Total	15	13.66	.7 > P > .5 not sig.
Due to mean	1	2.75	.1 > P > .05 not sig.
About the mean	14	10.91	.7 > P > .5 not sig.
Due to hypothesis	1	5.04	(2 tailed) P = .024 sig. (1 tailed) P = .012 sig.
Residual	13	5.87	P = .95 not sig.

The total variation about zero is not significant, nor does the mean of the 15 groups differ significantly from zero. However, the variation due to _a posteriori_ hypothesis, 5.04, is significantly great (two-tailed normal P = .024, or one-tailed normal P = .012). The residual variation, unexplained by hypothesis, tends to be quite low (P = .95). Using a two-tailed chi square test of significance on the 5% level of significance, this is not, however, judged to be significantly low. If this residual variation had been significantly low, this would have added to the already tenuous basis of the _a posteriori_ hypothesis. This is because the conclusion would have been that the _a posteriori_ hypothesis 'explained' too much variation. If this had been the case, even less credibility could be assigned to the _a posteriori_ hypothesis than was given to it in the discussion in section 2.53.

6.0. CONCLUSION

In the application of statistical theory to the natural sciences, actual conditions do not meet the theoretical ideal required for strictly valid statistical inferences. For example, in experimental work, while close control and nearly ideal randomization usually can be effected, the conditions prevailing in the ultimate population of interest are not duplicated. On the other hand, observational studies of naturally occurring events usually can be better directed at the object of ultimate interest; but opportunities for control and randomization are absent. Thus, the experimentalist is faced with the problem of inferring well beyond the experimental situation. And the observationalist is faced with the problem of distinguishing cause and effect from casual association in a situation over which he has no control. Neither of these problems falls under the domain of statistical theory in its present state of development. Perhaps they are, in general, insoluble.

Purists dodge the issue. If the purist experimenter makes no direct inference to the population of ultimate interest, he nevertheless knows, and probably hopes, that, for his results to be of any practical value, others will have to make the unstated inference for him. So, too, the purist observer may avoid all direct mention of cause and effect, but by the very scheme of classification that he uses he lays the unstated causal inference at the feet of his readers.

This writer feels that the experimenter and observer, and mixtures of the two, have a responsibility either to make explicit the unstated inference and defend it as well as possible, or to discount the unstated inference and explicitly recognize the limited value of the work. The experimenter, on the one hand, needs to make known the population of ultimate interest and to show how well or how poorly his experimental situation measures up. On the other hand, the observationalist, who really classifies and adjusts data according to some causal framework, needs to make that framework known; and further, he needs to adopt some set of criteria for testing the observed relations against the causal hypothesis. We have dealt at length with this latter need.

In the attempt to test a causal hypothesis with observational data, it has been shown that a usually large number of conditioning or disturbing variables enter into an acceptable causal frame. Usually the number of conditioning variables is too great to be dealt with adequately by mere classification schemes. One can, however, classify on certain variables which are thought to be closely related to the hypothetical cause and effect mechanism. Further adjustment for other disturbing variables can be accomplished through the use of maximum correlation technique.

We have considered the principle of maximum correlation in the abstract. In the particular case of dependent dichotomies or of dependent classifications to the categories of which a set of numbers can be applied, the maximum correlation technique has been shown to be equivalent to the familiar least squares technique. In the analysis, however, it appears that an adjustment for continuity, which has been developed in this paper, and certain minimum frequency requirements, which have been discussed, are needed in order to apply normal theory.

The selection of variables for classification is done with a purpose: to have the resulting categories or sub-groups available for a test of _a priori_ variations in effects of the hypothesized cause. Also, the selection of further variables to be adjusted by maximum correlation (least squares) technique is done with a purpose: to remove variation due to additional disturbing influences. Then the test of the a priori causal hypothesis involves not just whether there is or is not a 'significant', overall, adjusted or unadjusted, effect observed in the sample. More than that, the test requires that the hypothesis correctly predict variations in effects among the categories of the sample. Further, the test requires that residual variations among the classes, not accounted for by hypothesis, be ascribable to chance. It is with respect to this latter part of the test that the selection of conditioning variables to be adjusted for by the maximum correlation (least squares) technique becomes important. If important disturbing factors have not been adjusted for, substantial residual variation may be unaccounted for by the causal hypothesis. If this be the case, then it is not clear whether the hypothesized cause is responsible for the 'explained' variations, or whether the hypothesized cause has merely 'picked up' a portion of the variation due to unadjusted factors. But if all parts of the foregoing exacting test are met, then one has a quite reasonable, though certainly not an irrefutable, basis for making a causal interpretation.

In the absence (or presence, for that matter) of a causal interpretation, a posteriori relationships can be sought. But these, needless to say, can only be offered as hypotheses for future testing. With respect to the a posteriori relations which may be found, this writer believes, it is most important that the analyst indicate clearly that his findings are indeed a posteriori and nothing more.

BIBLIOGRAPHY

Ciocco, A., Densen, P., and Horvitz, D., "On the association between health and social problems in the population", The Milbank Memorial Fund Quarterly, V31, pp. 265-290, 1953.

Cochran, W. G., "Some methods for strengthening the common χ^2 tests", Biometrics, V10, pp. 417-451, 1954.

Cramer, H., "Mathematical Methods of Statistics", Princeton University Press, 1946.

Downes, J., "Illness in the chronic disease family", American Journal of Public Health, V 32, pp. 589-600, 1942.

Eden, T. and Yates, F., "On the validity of Fisher's z test when applied to an actual sample of non-normal data", J. of Agric. Science, V23, pp. 8-17, 1933.

Fisher, R. A., "The precision of discriminant functions", Annals of Eugenics, V10, pp. 422-429, 1940.

Horvitz, D. G., "Sampling and field procedures of the Pittsburgh Morbidity Survey", Public Health Reports, V67, pp. 1003-1012, 1952.

Hotelling, H., "Relations between two sets of variates", Biometrika, V28, pp. 321-377, 1936.

Lancaster, H. O., "The derivation and partition of χ^2 in certain discrete distributions", Biometrika, V36, pp. 117-129, 1949.

Li, C. C., "Segregation of recessive offspring", Methods in Medical Research, V6, pp. 3-16, The Year Book Publishers, Inc., Chicago, 1954.

Maxcy, K. F., "Preventive Medicine and Public Health", Appleton-Century-Crofts, Inc., 8th ed., N. Y., 1956.

Maung, K., "Measurement of association in a contingency table with special reference to the pigmentation of hair and eye colour of Scottish school children", Annals of Eugenics, V11, pp. 189-223, 1941.

Pitman, E. J. G., "Significance tests which may be applied to samples from any populations III. the analysis of variance test", Biometrika, V29, pp. 322-335, 1937.

Welch, B. L., "On the z-test in randomized blocks and latin squares", Biometrika, V29, pp. 21-52, 1937.

Williams, E. J., "Use of scores for the analysis of association in contingency tables", Biometrika, V39, pp. 274-289, 1952.

A METHOD FOR EXAMINING PARTIAL ASSOCIATION IN A POPULATION

By

Paul R. Sheehe

In scientific studies of man it is simpler, often necessary, to observe the population of ultimate interest than to experiment with it. But observational studies are difficult to interpret because conditions in the population are not under the observer's control and because the number of conditions is unlimited. The observer can include only a limited number of conditions, or variables, in traditional classification schemes because of the small frequencies encountered in the sample categories. Therefore, a practicable analytical method which takes more variables into account can add to the value of the observational approach. Such a method is developed and applied in this study.

The method combines the traditional classification technique with a technique for further adjustment of multiple variables based on the principle of maximum-correlation-scores (choosing those scores for categories of independent variables such that the squared correlation between independent and dependent variables is a maximum). Also, the method employs a unified technique for analysis of variation among sample classes in order to test the significance of associations, of hypothesized *a priori* causal relations, and of *a posteriori* relations.

For illustrative purposes, the method is applied to a study of the development of health problems in households. The data constitute a sample of the Arsenal Health District of Pittsburgh, Pennsylvania. These data are the health histories of members of selected households, in the year preceding interviews in mid-1951 and in mid-1952. Persons with chronic disease, or physical impairment, and persons hospitalized during the year or recently

confined to bed are considered to have a health problem.

Specifically, the analysis is concerned with the question: does a household member with a non-communicable health problem in the first year increase the chances of subsequent health problems among the initially healthy members. A preliminary analysis shows an opposite tendency, that is a negative relation: households with, compared to those without, initial health problems develop subsequent health problems less often.

In a more refined analysis, the following household characteristics are chosen as conditioning variables: size; average age; sex distribution; stratum, defined according to density of neighboring households; and class of interviewer. Classification on these variables would produce class frequencies much too small to be analyzed by conventional techniques. But the proposed technique, utilizing the IBM 650 computer, can be employed. By this method the household health problem relation is adjusted for all the above conditions. Conceptually, this means that comparisons of households with and without initial health problems are made only between households of the same size, the same average age class, the same kind of sex distribution and the same stratum, as well as between households questioned by the same class of interviewer.

The negative relation practically disappears when these conditions are taken into account. Furthermore, the variation of the relation among the categories of households is no more than would be expected by chance. Consequently, the data do not support the hypothesized positive relation under any of the analyzed conditions.

Nevertheless, inspection of the data does reveal that, in households consisting predominantly of adults, the relation is positive when most of the initially healthy members are female. This suggests, among other possible hypotheses, that adult females, rather than males, may be responsive to health problem stresses.

A detailed development of the proposed analytical technique is taken up in the sections following the practical application. Properties of maximum-correlation-scores for two-factor contingency tables, as described by Williams in *Biometrika* (1952), are reviewed. This is followed by extension of the maximum-correlation-scores technique to multiple-factor tables. Statistical properties of measures of partial association, developed on the maximum-correlation-scores principle, are investigated by Monte Carlo methods on the IBM 650. It is found that an adjustment for continuity is required to obtain a good approximation of chi-square and normal distributions to the empirically generated sampling distributions. The use which is made of these statistical properties is illustrated in the foregoing practical application.