RESEARCH IN MULTIVARIATE STATISTICAL ANALYSIS AT CHAPEL HILL

A progress report of research under Office of Naval Research Contract N7onr 284, Task Order II, as to the portion of work carried on at Chapel Hill under the direction of Harold Hotelling.

Submitted by the

INSTITUTE OF STATISTICS, UNIVERSITY OF NORTH CAROLINA

To: Chief of Naval Research
(Submitted for distribution)

October 1, 1948
CONTENTS

1. STAFF OF CHAPEL HILL PROJECT IN MULTIVARIATE ANALYSIS ................................. 1

2. ORGANIZATIONAL SETTING ......................................................................................... 1

3. NATURE AND BACKGROUND OF MULTIVARIATE ANALYSIS ................................. 1

4. CURRENT PROBLEMS AT CHAPEL HILL ................................................................. 3
   a. The $T_n$ Distribution ......................................................................................... 3
   b. Rank Correlation ............................................................................................... 3
   c. Distributions of Quadratic Forms and Functions of Them ......................... 4
   d. Serial and Lag Correlations ............................................................................ 5
   e. Shrinkage of Multiple Correlations, and Problems with Incomplete Data .... 6
   f. Problems of Prediction and Nuisance Parameters ........................................ 7
   g. Truncated Normal Distributions ..................................................................... 7

5. PROBLEMS ON WHICH EXPLORATORY WORK HAS BEGUN ............................. 8

6. MAJOR PROBLEMS FOR WHICH RESEARCH PLANS HAVE BEEN MADE ......... 9
   a. Assignment of Personnel to Groups ............................................................... 9
   b. Sampling of Variates ....................................................................................... 10
   c. Time Series and Stochastic Difference Equations ....................................... 11
   d. Sequentialization of Multivariate Tests ....................................................... 12
1. STAFF OF CHAPEL HILL PROJECT IN MULTIVARIATE ANALYSIS

Gertrude M. Cox, director of the Institute of Statistics.
Harold Hotelling, project director.
Isadore Blumen, research associate.
Glenn L. Burrows, research associate.
Wassily Hoeffding, research associate.
Dorothy J. Morrow, research associate.
Webb Stacy, research assistant.
Robert F. Tate, research assistant.
James W. Walker, research assistant.
Martha Jordan, secretary
Evelyn S. O’Neal, computer.

FORMER MEMBERS OF THE STAFF

Mary E. Armbruster
Max Halperin
P. L. Hsu
George E. Nicholson

2. ORGANIZATIONAL SETTING

The Institute of Statistics of the University of North Carolina is devoted to statistical research, teaching and service. The service function is discharged mainly through the Department of Experimental Statistics at Raleigh as regards direct practical applications. Research and graduate instruction in statistics are carried on both in it and in the Department of Mathematical Statistics at Chapel Hill, but chiefly in the latter as regards the more mathematical areas. Cooperation is close between the two departments. Task Order II of the Office of Naval Research Contract N7onr 284 is divided into two parts, one carried on in each department. This report pertains to the Chapel Hill project. A separate report is being submitted regarding the work in the Department of Experimental Statistics, "The Statistical Analysis of Groups of Experiments."

3. NATURE AND BACKGROUND OF MULTIVARIATE ANALYSIS

Gauss, Bravais and Karl Pearson laid the groundwork of multivariate statistical analysis. The method of least squares, correlation, and the multivariate normal distribution had by the end of the nineteenth century reached a state of high usefulness. French writers had explored the "probability ellipse" as a means of studying the accuracy of artillery fire with reference to errors in two dimensions, and Karl Pearson with biometric collaborators had developed an approximate probable error for the correlation coefficient and other contributions to multivariate statistical theory, having in view chiefly the mathematical study of evolution. R. A. Fisher turned the direction of development to the consideration of exact rather than asymptotically approximate tests, beginning with his paper of 1915 on the distribution of the correlation coefficient. Fisher proceeded to show by means of his theory of efficient statistics how the very beginnings of statistical procedure needed to be reorganized to take account of the twin needs of economy of observations and exact probabilities. By using geometrical methods and taking advantage of the invariance of the normal distribution under rotations, Fisher was able to reduce distributions of functions of n observations to integrals of low multiplicities independent of n, and thus in many cases to familiar functions. Further developments of these geometrical procedures as bases for the rigorous derivation of exact distributions of importance in multivariate analysis took place in the 1930's in this country and in India. On the other hand, many investigators have felt uneasy about geometrical methods in n dimensions, and have preferred purely analytical and algebraical approaches.
Further progress in multivariate analysis was rapid. Wishart in 1928 published the joint distribution of sample covariances from a multivariate normal distribution, Fisher in 1929 the distribution of the multiple correlation coefficient. In 1931 Hotelling published the distribution of the statistic $T$ appropriate for testing jointly the differences between two populations of the means of several variates and for other uses analogous to those of the Student $t$. Mahalanobis, Bose and Roy considered essentially the same statistic $T$ in its role as a measure of "distance" between populations, and later published its more general sampling distribution on the hypothesis that the population distance is not zero. Wilks in 1932 gave the distributions of a number of determinants and ratios of determinants generalizing in certain ways the analysis of variance.

Roots of certain determinantal equations gained statistical application in an early paper of Pearson and in one by Thurstone in 1930. These were the latent roots of matrices of correlations or covariances. A detailed treatment of these with reference to psychological applications, with an iterative computational procedure and a beginning of the probability testing of hypotheses by means of them, was given by Hotelling in 1933 and amplified by later work of Girshick and Hotelling. A different type of determinantal equation was encountered by Hotelling in studying relations between two sets of variates invariant under internal linear transformations of each set, such as the linear functions of a number of test scores of applicants for jobs best for predicting success, and simultaneously the linear function of various measures of success that can be predicted most accurately. Several psychological, meteorological and economic types of application were suggested in this paper, published in 1936. The critical sampling distributions were derived in part in it, and later in greater generality by Girshick. In 1939 a still more general distribution, applicable to these and also many other determinantal equations, was obtained independently by P. L. Hsu and S. N. Roy. The distribution of Hsu and Roy was, however, a joint distribution of all the roots, and difficult integrations are required to obtain in usable form the distributions of various functions such as the greatest root and the sum of the roots, which are needed for many practical purposes.

In 1934 a Naval Proving Ground problem in the evaluation of accuracy of test bombing led to the search for suitable functions taking account simultaneously of errors in range and in deflection, errors having very different variances and possibly correlated with each other. Such measures should lend themselves to comparison of accuracy with different types of equipment and personnel. Moreover, the extremely high cost of test bombing called for the utmost statistical efficiency, in the sense of using functions of observations conserving to the full the costly information contained in them, even if this should require a little extra computing. Exact probabilities, not functions of unknown parameters, were also desired. Three statistics based on the earlier $T$ were developed that possessed all the required properties, and could be used to test the accuracy of the deviation of the centroid or mean point of impact (MPI) from the target, or of the individual bombs from the MPI, or of the individual bombs from the target, -- three measures sensitive to different types of defects in bombsights and techniques. These were called respectively $T_M$, $T_D$, and $T_Q$. The distribution of $T_M$ is the same as that of $T$. Those of $T_D$ and $T_Q$ are equivalent to each other with a change of parameters, and their distribution was found for the case then immediately important of deviations in just two dimensions. The question was at once raised of generalizing the distribution to more than two dimensions; this has proved to involve mathematical complexities of much higher order, and is now one of the concerns of the Chapel Hill project.
4. CURRENT PROBLEMS AT CHAPEL HILL

a. The $T_0$ Distribution

An extensive study is now nearly completed of the $T_0$ distribution in more than two dimensions. Miss Morrow has developed as a mathematical tool for obtaining the requisite probabilities a series which converges rapidly for some values of the parameters of practical interest, but much more slowly in other cases which therefore require special treatment. She has obtained a different series for direct calculation of percentage points, of which a short table is being prepared. She has also studied the asymptotic behavior of the distribution and other features, and has explored various applications, particularly in connection with the control of quality of manufactured products where several variables are measured. The final report of this enterprise is now being written.

b. Rank Correlation

The great work of Karl Pearson and his followers, and of R. A. Fisher, on correlation was for the most part based on an assumed normal distribution. Some of the results are closely accurate even when the underlying population is of other than normal form, but others are wide of the mark. To deal with this situation in practical problems, where some distributions are known not to be normal and in many other cases normality is in doubt, methods are needed for studying correlation without assumptions of normality.

The two rank correlation coefficients, of which one was first studied by Spearman, Gosset, Esher, and Pearson and the other by Lindenberg and Kendall, may be used for this purpose. They are, however, only two of an infinite class of statistics that may be used as tests of independence vs. correlation, or as estimates of parameters measuring correlation or degree of relationship when it exists. These all suffer in varying measure from the defect that, if applied to data from a population that is really normal, they are statistically less efficient than the ordinary product-moment correlation coefficient $r$. They are in case of true normality less accurate measures than $r$ of the underlying correlation parameter, and when used as tests of independence with a fixed probability level such as .05 for false rejection of independence, they are less likely than $r$ to detect any correlation that may exist. Moreover, being based on rank orders, they have distributions involving number-theoretic functions of integers, which are much less easily amenable to simple practical treatment than the continuous functions expressing the distributions of the product-moment correlation in samples from a normal population and of related statistics.

Even in samples from non-normal populations, some of these potential rank order statistics are much more powerful than others in detecting correlation, and presumably are more efficient also in measuring the degree of relationship when correlation exists, at least for small correlations. A suitable choice among these infinitely many possible coefficients is, therefore, of great importance. A further complication lies in the fact that the results obtained through choices based on the asymptotically normal distribution of the estimates realized in most cases may differ materially from those appropriate for very small samples. A still further difficulty lies in the possibility that some of the estimates of correlation based on ranks may not have even asymptotically normal distributions.

Dr. Hoeffding has been developing a theory of functionals that may be taken to represent correlation in bivariate and multivariate distributions, and that possess unbiased estimates having minimum variance in a large class of parent distributions. Being asymptotically normally distributed, they can be considered as most efficient unbiased estimates. A manuscript embodying a considerable number
of results pertaining to this problem was completed several months ago and sub-
mitted with an earlier report; it is due to appear in the September issue of the
Annals of Mathematical Statistics. In the meantime, Dr. Hoeffding has proceeded
with the slightly different problem of asymptotically shortest confidence inter-
vals for rank order statistics and has obtained a number of relevant theorems.
Several shorter papers including these are expected to be published soon.

The same rank order problems arising in connection with the correlation
between two variates occur also in the more complex situations involving many
variates. It is hoped that, after getting the theory of simple correlation into
a satisfactory state from this standpoint, a more thoroughgoing attack can be made
on the corresponding problems of general multivariate analysis.

c. Distributions of Quadratic Forms and Functions of Them

A considerable part of multivariate analysis depends on the distributions of
functions of various quadratic forms in normally distributed variates. Most of
the forms considered in the past have independent distributions of the chi-square
type, and in many practical problems the only distribution needed is of the ratio
of two such forms.

Many needs have arisen recently for distributions both of quadratic forms
not of the simple chi-square type, and of functions of quadratic forms more com-
plicated than the simple ratio. In time-series analysis, whether in economics,
meteorology, astronomy, or anti-aircraft tracking, successive observations are
correlated, and serial correlation and regression coefficients can supply valuable
help in utilizing such observations. These statistics are in the simplest cases
ratios of quadratic forms, of which one has the chi-square distribution and the
other does not. In more complicated cases, corresponding to multiple regression
analysis and systems of stochastic difference equations, the statistics which
must be studied are functions of numerous quadratic forms which for the most part
are not of the chi-square type, and the functions are much less simple than mere
ratios. Other uses for quadratic forms not having the chi-square distribution
arise in connection with uses of correlation, analysis of variance, and other
familiar techniques where the observations are of unequal accuracy, and still
other uses are in prospect.

The time seems ripe, therefore, for an attack on a wider front than the
searches for distributions of particular simple functions of quadratic forms such
as have been carried on independently in connection with many scattered applied
problems. As a beginning, the distribution of a single arbitrary positive definite
quadratic form in normally distributed variates with zero means has been investi-
gated. A series of Laguerre polynomials seems to provide a representation of the
distribution highly satisfactory for many purposes, but like other orthogonal ex-
pansions subject to the drawback that the error in stopping with any finite number
of terms is for any particular value of the variable difficult to estimate. This
series lends itself also to convolution so as to give the distributions of inde-
finite quadratic forms, and also of ratios of forms as convergent series. Some
useful inequalities have been derived. For certain special cases a finite sum of
exponentials gives the distribution.

Another form of the distribution of an arbitrary positive definite quadratic
form has been devised by Professor Herbert Robbins of the Institute of Statistics
at Chapel Hill, who was consulted on this problem. His solution, a power series
with readily calculable coefficients and a satisfactory bound for the error in
stopping with any particular term, has now been published in the Annals of Mathema-
tical Statistics for June, 1948, pp. 266–270.
The present hope is to extend these results to the joint distributions of several quadratic forms, and thence by integration to various functions of these forms.

d. Serial and Lag Correlations

Other approaches are being tried in connection with the particular quadratic forms involved in serial and lag correlations of time series. Certain of the joint distributions have been studied, but are of a complicated nature, with different analytical forms in different regions into which the space of the variable forms is subdivided. How to reduce these involved distributions to usable shape is now a subject of consideration. Mr. Blumen, and formerly Mr. Halperin, have been working on this part of the project.

The methods used in this attack on distributions important for time series analysis (and also incidentally for certain problems involving spatially distributed variables, such as fertility elements in a field) include geometrical, algebraic and analytical procedures. The well known representation of a correlation coefficient as the cosine of an angle in n-space supplies the basis for geometrical methods of great power, particularly when a normal distribution is assumed with independence among the observations, since then the probability distribution in the n-space has spherical symmetry and is, therefore, invariant under all rotations. This and other invariantive properties provide means of attacking important classes of distribution functions, including those here in question.

The notion of circular serial and lag correlation coefficients was introduced by the project director many years ago as a device for taking advantage of an invariantive property that seems to open the way to determination of exact distributions. Properties of determinants of the type known as circulants play a part in this kind of research. It was in this way that R. L. Anderson obtained the exact distribution of the circular serial correlation in 1942 on the hypothesis of independence, which W. G. Madow generalized in 1945. Both Anderson and Madow are now members of the Institute of Statistics of the University of North Carolina, and have been helpful with suggestions. Mr. Blumen has been exploiting further the geometrical approach and the use of circulants, of which determinants he has pushed the theory well beyond the published treatments.

The latent roots of circulants are very simply expressible in terms of cosines of multiples of the same angle. This fact makes possible the simplification of the contour integrals giving the probability densities in terms of the joint characteristic functions for several serial correlations with different lags. However, the expressions obtained, though individually of elementary character, are different in the many different regions into which the plane of variation of two serial correlations is subdivided by twelve straight lines. For a larger number of serial correlations the regions within which the probability density takes different analytic forms are greatly multiplied. Mr. Halperin, before leaving the group in August, carried a study of the joint distribution of two serial correlations far enough to make these features clear.

Use of the circular serial correlation represents a slight loss of efficiency in detecting true serial correlations of the types usually considered. However, in discussing the joint distributions of several correlations, it is found mathematically advantageous to retain the circular definition despite the loss of efficiency. The advantages arise partly from the applicability of a theorem ascribed to Frobenius, which states that a necessary and sufficient condition for the simultaneous reducibility of a set of quadratic forms to weighted sums of squares by the same orthogonal transformation is that their matrices be
commutative. This requirement is satisfied when the circular definition is used, but not with other conventional formulae for serial correlations and mean square successive differences.

e. Shrinkage of Multiple Correlations, and Problems with Incomplete Data

Users of regression equations in personnel placement and other fields have often been pained to discover that the correlations between prediction and fulfillment have been far less in the material to which they applied a regression equation than the multiple correlation in the sample from which the regression function was calculated. Some shrinkage of this kind is to be expected on theoretical grounds, since the regression coefficients are chosen so as to minimize the sum of squares of deviations in the old sample but not necessarily in the future one. A more exact knowledge of the probability distribution of this shrinkage has been desired by personnel workers, and has been an objective of a piece of research under way for the past year. This has now been brought to a tentative conclusion, and a detailed report is in preparation.

Only approximate results have been obtained, and these on rather special assumptions. Different formulations are possible. The particular data in hand may be compared with a hypothetical population of samples in each of which the predictors take the same set of values as those observed, a procedure brilliantly successful in the ordinary theory of least squares. On the other hand, predictors and predictand may be thought of as drawn from a multivariate normal distribution, with different values in different samples, as in Fisher's "\text{A}" distribution of the multiple correlation coefficient. The success of the first formulation in obtaining tests relating to regression coefficients and free of nuisance parameters is well known; the second formulation leads to less happy results so far as regression coefficients are concerned because their distribution in the broader population then assumed contains unknown nuisance parameters. For this reason, attention has in the present study been concentrated on the first formulation, which assumes fixed sets of predictors in different samples. Since the problem concerns an old and a new sample, some variation of this scheme is appropriate. The problem as finally formulated leads to a distribution for the shrinkage in a form free of nuisance parameters. Unfortunately, the distribution is that of a certain rational function of four independent variates, each having the chi-square distribution, and no tables are available from which the relevant probabilities can be read. Hence, pending the fuller study of functions of quadratic forms mentioned elsewhere in this report (p. 4), and after the failure of attempts to apply geometrical and other methods to obtain simple expressions, it seems necessary to leave this distribution for the present with some approximations which have been derived, but which will presumably be superseded when more is known about the general subject of distributions of functions of quadratic forms.

An unexpected by-product of this study is a neat contribution to the estimation of correlations and other parameters from samples consisting of matrices of observations with vacant spaces corresponding to missing observations. This carries further the study by S. S. Wilks in the \textit{Annals of Mathematical Statistics}, vol. 3 (1932), pp. 163-195, and will be reported upon separately.

This part of the project has been in the hands of Mr. Nicholson, who, however, has resigned from the project to accept an assistant professorship in the Department of Mathematical Statistics here. He is now making final revisions of the two reports mentioned above.
f. Problems of Prediction and Nuisance Parameters.

This part of the project, which is being conducted by Mr. Burrows, is a continuation of the study reported in "The selection of variates for use in prediction, with some comments on the general problem of nuisance parameters," by Harold Hotelling, Annals of Mathematical Statistics, vol. 11 (1940), pp. 271-283. In that paper exact tests, independent of unknowns which were there for the first time christened "nuisance parameters," were obtained for the significance of the differences in correlation with a predictand of different proposed predictors, of which only one could be used.

In the practical application of these tests a difficulty arises in that the goodness of prediction is essentially measured by the absolute value or by the square of the correlation, so that the appropriate test statistic would seem to be proportional to the difference of absolute values of the correlations with two predictors, instead of proportional to the difference of the correlations themselves as in the earlier paper. For example, if in a certain region the yield of wheat has a correlation .6 with June soil moisture and a correlation -.6 with average June temperature, a significant difference might be asserted by the earlier test for some samples, though the indicated predicting power is exactly the same for the two variables. If no information outside the sample correlation coefficients were available as to the sign of the true coefficients, there might be a possibility of this test becoming misleading. Such cases are doubtless numerous, and it seems desirable to devise a test invariant under reversal of sign of either correlation.

Several plausible test functions have been examined for this purpose. Each has proved to have an objectionable feature in that its distribution involves unknown nuisance parameters. This has happened so consistently as to raise the suspicion that this difficulty is inherent in the nature of the problem, and a proof is now being sought that this is the case. The question is one that reaches deep into the foundations of statistical theory. It deserves attention for this reason as well as because of its direct practical importance. Mr. Burrows is, therefore, seeking new approaches, including both sequential methods and possible extensions of Neyman's work on regions similar to the sample space.

g. Truncated Normal Distributions

Several military workers in personnel selection problems have stated that they need an amplification of the limited existing theory of truncated normal distributions in a multiplicity of variables. Similar problems arise in measuring the accuracy of fire, for example, on a towed rectangular target. The numbers both of applications and of distinct mathematical problems in this field appear to be considerable.

Karl Pearson applied the method of moments to fit univariate normal distributions truncated at one end, but the efficiency of this procedure was demonstrated only in 1931 by R. A. Fisher in his introduction to the British Association for the Advancement of Science table of Hermite functions, which Fisher showed how to use effectively for this particular problem. A generalization of Fisher's theorem has been obtained and will be published after completion of some auxiliary investigations regarding the technique of fitting truncated multivariate distributions. (A similar theorem has now been reported by Z. W. Birnbaum also.) It is found that the method of moments is efficient for estimating the parameters of a multivariate normal distribution even though only the cases falling into a particular measurable set of combinations of values can be observed, and moreover that it gives a unique result. This, however, does not dispose of the problem, since difficult integrations must be dealt with even for relatively simple cases. Much remains to be done in the development of practical techniques of fitting.
5. PROBLEMS ON WHICH EXPLORATORY WORK HAS BEGUN

The problems listed in this section have been examined and work has begun on them in a tentative way, with a view to developing new methods of approach. Several of these problems have arisen out of needs connected with mental testing, personnel classification, and similar activities. Any results obtained should have applications in these fields, and also in considerably more numerous areas of quite different kinds.

The tetrachoric correlation coefficient introduced by Karl Pearson is an estimate of the true correlation in an assumed bivariate normal distribution on the basis of the four frequencies obtained by a double dichotomy. It is widely used in item analysis in test construction, though Pearson's examples pertained to applications of other kinds. Much attention has been given to simplifying the calculation of tetrachoric \( r \) and its probable error by means of tables and nomograms. Its theory, however, remains in a deplorable state. The obsolete notation and concepts and the non-rigorous methods of Pearson's early work on tetrachoric \( r \) seem never to have been corrected, though his similar work on the product-moment coefficient has now been thoroughly overhauled and perfected, mainly as a result of the brilliant contributions of R. A. Fisher to its theory. The Chapel Hill group is now making a modest exploration of this situation. As a first objective, it is hoped to find a transformation of the tetrachoric correlation coefficient effecting results similar to those of Fisher's \( z \)-transformation of the product-moment coefficient. This inquiry may later be broadened to cover other phases of item analysis and selection for tests.

A similarly tentative inquiry has been begun regarding reliability coefficients and correlations corrected for attenuation. As to the latter the outstanding paradox consists in the fact that the formula commonly used in an attempt to improve the estimate of the correlation sometimes gives values greater than unity, which cannot possibly be the true values. This suggests an appeal to fundamental principles of estimation and leads to the idea of applying the method of maximum likelihood. In this case, however, the maximum likelihood is not found directly by differentiation, but takes place on the boundary of the region in the parameter space corresponding to positive definite matrices, which are the only ones capable of being correlation matrices. This is an atypical situation, quite unlike others in which maximum likelihood has been used.

Another tentative inquiry under way concerns prediction on the basis of two or more samples with unequal accuracies. This is related to the studies c, e, and f of Section 4 above.

Attention has been given in the group to a problem propounded by Admiral G. L. Schuyler, U. S. N., Retired, as to the best methods of voting, and of combining votes, when a board or committee must choose a limited number of projects, or a number of projects with limited total cost, from among a larger set of proposals. This is related to the election problems that have been studied by certain political scientists, and also to certain problems encountered by industrial psychologists. Admiral Schuyler has shown that some plausible systems have objectionable features, and has proposed a particular solution which appears reasonable. Questions of probability pertaining to the accuracy of choice according to such a system fall in the same category as those involved in the study of rank correlation. It is hoped that these questions can be investigated in detail.
6. MAJOR PROBLEMS FOR WHICH RESEARCH PLANS HAVE BEEN MADE

The problems described in this section have received longer study in the group than those of the preceding section, though still in the planning stage, and methods of attack on them have been outlined. The questions listed below are of substantial importance for the further development of statistical theory, and also from the standpoint of numerous applications. They are beyond the capacity of beginners in statistical research excepting under the supervision of mathematical statisticians who have reached a certain degree of maturity in research, and the carrying out of these projects is conditional upon getting and retaining the services of relatively mature or at least intermediate, scholars in the field. Each of these projected studies is a logical extension of work now under way in Chapel Hill and calls for specialized knowledge now possessed by members of the group. The need for additional personnel to supervise their execution arises partly from the preoccupation of the present senior staff with the work already described and the recent resignations of two research associates to accept research positions elsewhere at much higher salaries, partly from the desirability of bringing to these studies certain additional research techniques specially relevant to them.

Multivariate analysis has been the central theme of the work of this group. Rank correlation is a part of multivariate analysis, but differs from the other topics studied in its concern with functions of integers and permutations, and forms a bridge with other branches of statistical theory concerned with these same discontinuous functions. The first of the projects described below is in this category. The others are closer to the multivariate studies based on continuous functions.

a. Assignment of Personnel to Groups

When a large number of persons is to be divided into groups of, say, five, who are to live or work together, results are apt to be much more satisfactory if account is taken of mutual likes and dislikes in making the assignments. Questionnaires to ascertain these preferences, and methods of assignment based on the responses, have been devised by the Sociometric Institute, which provides service of this kind to schools, businesses and other organizations. The importance of some such procedure for smooth-working human relations points to considerable extension of such activities. Certain mathematical questions are, therefore, of interest which affect the appropriate form of the questionnaire, the method of using the replies, and the estimated improvement likely to result.

The procedures in current use call for each of N persons to choose K-1 others from among the N with whom he would like to be associated in a group of K. The results may be expressed by a matrix H in which the element in the i th row and j th column is unity if the i th person has chosen the j th person, but is otherwise zero. The sum of the elements in each row has the fixed value K-1. Matrices of this type have received mathematical attention, for instance by V. Romanovsky in his study of Markoff chains, who has classified them according to their degree of separability into submatrices of the same type, a classification eminently suited to the sociometric application. Such matrices occur also in the study of circuits in topology. Any rule for the assignment of persons on the basis of these questionnaires may be expressed in terms of some function of this matrix. The mathematical problem arises of finding functions of such matrices that possess certain desirable properties.

Any particular assignment of the N persons into c groups of K (N = cK) may be represented by a matrix J of N rows and c columns, in which the element in the i th row and j th column is unity if the i th individual is put into the j th group, but is otherwise zero. In the product matrix HJ each of the c
columns corresponds to one of the groups. In the column corresponding to the r th
group, the element in the i th row, if the i th individual has been put into this

group, is the number of other persons chosen by this individual who have also been
put into it. This element may, therefore, be taken as a measure of the good for-
tune of this individual. The good fortune of all those in the rth group, defined
as the sum of those measures for the individuals in it, is the sum of all those


elements in the r th column of HJ which are in rows pertaining to individuals who
have been put into this group. This sum equals the r th diagonal element of


J'HJ. Hence the total of the measures of good fortune for all the N persons is
the sum of the diagonal elements, or trace, of J'HJ, and equals also the sum of


the latent roots of J'HJ.


Thus an objective worth investigating appears to be the choice of an as-


signment matrix J, given H, which will minimize the trace of J'HJ. This is remi-
niscent of the problem discussed in several recent papers in the Annals of


Mathematical Statistics of selecting matrices with elements 0 or 1 to represent
efficient combinations of objects to be weighed in groups.


Relations between as well as within groups may be examined with the help of
the matrix H. This might be important if certain groups were to be assigned to


cooperate with other groups, as in a military operation. As between any two
mutually exclusive subsets of the population of N, a criterion of the popularity
of one with the other may be based on the number of choices by persons in the
second of persons in the first group. The two criteria thus relating the two
groups may be averaged to provide a symmetrical measure of mutual esteem, and this
can be expressed as an index by dividing by its greatest possible value. If each
group consists of only one person, the index takes the value 1, 1/2, or 0 accord-
ing as each chooses the other, one chooses the other but is not chosen by him, or
neither chooses the other.


Probability distributions of this index, and of functions of the matrices
that may be used in making assignments, need to be studied under various hypo-
theses. The simplest such hypothesis is mere randomness of choice. Dr. Hoeffding
has made a beginning by working out the mean and variance of the index of mutual
esteem on this hypothesis, and has shown that its limiting distribution is of the
Poisson form if N and the number in one group become large while the number in the
other group remains fixed. He has also calculated the mean and variance on the
basis of more realistic considerations regarding clustering among the N persons,
and has conjectured that the limiting distributions are of normal form if N and
the numbers in the groups compared increase in fixed proportions. If this con-
jecture is correct, it could doubtless be verified by a study of the moments.


Other aspects of this problem that seem worthy of investigation include the
question how to treat partial non-response, the distribution of the indexes under
hypotheses other than those mentioned, and the use of questionnaire modifications
to permit expression of degree of preference.


b. Sampling of Variates


When p tests have been applied to N persons, the correlation matrix has
been calculated, and an analysis of the complex into components of some kind
("factor analysis") has been carried out, the question is inevitably raised as to
the effect on the results if one or more of the tests were to be replaced by
others not originally included. There has been a strong feeling among psycholo-
gists that the significance of the components found must rest heavily on some sort
of invariance, at least approximate, of the results under changes in the battery
of tests. A part of this feeling may have arisen from a foredoomed quest for
genetic factors which was at one time undertaken on the basis of correlation
matrices among tests given to miscellaneous sample of the population without
regard to relationships among the individuals tested. There is, however, a very real and substantial basis for the contention that inferences from a correlation matrix should be supported by the results of applying the same mathematical procedures to other sets of tests set up in a like manner and administered to the same or a comparable group of persons.

The only reasonable way to go about answering the question of the amount of variation to be expected by reason of a change in the tests seems to be to regard each test as drawn at random from a larger aggregate of tests constructed in an essentially similar manner. The aggregates from which the drawings are made will usually in practical situations vary from test to test of the battery, but in a first approach it seems a desirable simplification to regard all the tests as drawn from the same large hypothetical aggregate over which there is a definite probability distribution.

A beginning of a theory of sampling of tests or variates, which must be distinguished from the more familiar sampling theory for persons, was made by the project director in the Journal of Educational Psychology for 1933, pp. 504-520. The aggregate, finite or infinite, of possible tests was represented in terms of its principal components, with the coefficient of each order in this orthogonal expansion varying randomly according to a definite probability law. Thus each possible test was characterized by a sequence of these random coefficients of the same sequence of fundamental variates. These fundamental variates were considered uncorrelated in the population of persons tested, but any specific pair of tests would be correlated to an extent depending on the sum of products of corresponding coefficients. If a random finite sample of tests were drawn, the matrix of their correlations would determine various quantities, for example, the latent roots and their associated vectors, which would thus have a definite sampling distribution. These distributions were studied, partly through their moments, and applied to the problem of discriminating between the "sand" and "cobblestone" theories of the mind.

New applications have been emerging for these ideas, and the incomplete earlier study should be supplemented by fuller investigations. One class of needs for this kind of sampling theory is related to computational problems involving large matrices. If, for example, a quadratic form is to be computed whose matrix is the inverse of a given matrix of correlations among the variables appearing in the form, the labor is prohibitive with ordinary office machines when the number of variables is more than about twenty, and will be prohibitive with electronic machines when it exceeds some considerably higher limit. There is, however, ground for believing that a random sample of a moderate number of those variates can be used to compute a quadratic form similarly defined, with an enormous reduction of labor, and that this can serve as a basis for an unbiased estimate of the value originally sought, with an accuracy determinable in terms of probability and an error converging stochastically to zero as the number of variates in sample and aggregate increase in a fixed proportion. Further investigation along this line by a competent mathematical statistician should yield a beautiful new theory having widely diversified applications.

c. Time Series and Stochastic Difference Equations

Reference has been made in Section 3, Subsections c and d above, to the research now under way on lag and serial correlations. These are a part of the study of time series, whose proper treatment to extract a maximum of relevant information has long puzzled economists, meteorologists and others. Currently, the most hopeful method of attacking such problems seems to be through the use of stochastic difference equations, which undertake to explain the set of interacting variables considered as functions of their own previous values together with
chance elements. The testing of significance of fit of such equations to data requires the distributions mentioned earlier as objectives, and others. The accurate estimates of the parameters of those equations requires still other distributions of this general character.

The problem known as that of "identification", and relations with the theory of stochastic processes, need to be explored further. This cannot all be done at Chapel Hill; what is thought of as appropriate for this group is the part most closely connected with multivariate statistical analysis of the kind already cultivated here, particularly the probability distributions of functions of quadratic forms.

d. Sequentialization of Multivariate Tests

The advantages of sampling in such a way that the number of observations to be made is not fixed in advance but is determined by the outcome of the earlier trials have come to be well understood since the war through work of Wald and others. Nothing systematic appears, however, to have been done regarding the extension of sequential methods to observations on multiple variates through repeated experiments whose scale and design as well as number may also vary according to the outcome of the earlier observations, though outstanding cases exist of such sequential operations without previous complete design, such as the succession of pre-war samples of jute acreage in Bengal. Each of those surveys, conducted by able mathematical statisticians, covered a larger area than the preceding one, and was valuable not so much for the acreage estimates obtained for the year of the census as for the information it supplied on the cost and variance functions which made possible the more efficient design of the next year's sample. Informal sequentialization has indeed always been the rule in practice, but great gains in economy and accuracy are possible through the systematic application of theory with calculable probabilities.

Agricultural operations differ greatly from the industrial quality control schemes which heretofore have provided the chief application of systematic sequential methods, in that each experiment in agriculture usually requires at least a year, while for some crops and animals the period is longer. At the same time, the number of correlated variables to be considered is large.

In designing an experiment there is always an appeal to what is already known about the costs and variances, but data are often scanty and of doubtful relevance to the conditions of the projected experiment. A self-contained experiment is desirable in which the main operation is prefaced by one of the same kind on a smaller scale, designed to supply the information most needed for efficient design of the large experiment. The preliminary operation may in turn be preceded by a still smaller one, and so on for several steps, all designed in advance in the sense that rules have been set up governing the steps to be taken at each stage in every possible contingency.

These considerations apply to each of the many fields in which multivariate analysis is used, though with varying force. There is a great need of combining the extensive body of knowledge and techniques pertaining to multivariate analysis with sequential methods.