DOCUMENT RESUME

ED 026 737 24 EA 001 982

By-Bashaw, W. L., Ed.; Findley, Warren G., Ed.

Symposium on General Linear Model Approach to the Analysis of Experimental Data in Educational Research (Athens, Georgia, June 29-July 1, 1967). Final Report.

Georgia Univ., Athens.

Spons Agency-Office of Education (DHEW), Washington, D.C. Bureau of Research.

Bureau No-BR-7-8096

Pub Date 23 Aug 68

Contract-OEC-2-7-008096-0496

Note-243p.

EDRS Price MF-$1.00 HC-$12.25

Descriptors-Analysis of Variance, Computer Oriented Programs, *Data Analysis, *Educational Research, *Experiments, *Models, Research Methodology, *Statistics, Symposia

This volume contains the five major addresses and subsequent discussion from the Symposium on the General Linear Models Approach to the Analysis of Experimental Data in Educational Research, which was held in 1967 in Athens, Georgia. The symposium was designed to produce systematic information, including new methodology, for dissemination to the educational research community to (1) promote wider use of sound methodology and (2) provide caveats regarding limitations of this powerful approach. The authors and their papers are: (1) Graybill, Franklin A., "Introduction to the Use of General Linear Models in the Analysis of Experimental Data," (2) Ward, Joe H., Jr., "Synthesizing Regression Models--An Aid to Learning Effective Problem Analysis," (3) Winer, B. J., "Problems in the Use of General Linear Model Methods," (4) Bargmann, Rolf E., "A Survey of Appropriate Methods of Analysis of Factorial Designs," and (5) Bock, R. Darrell, "Remarks on Analysis of Variance and Analysis of Regression." The papers and ensuing discussion elucidate strengths and limitations of the general linear models approach, discuss procedures for handling computations, and present the independent views of major authorities on theory and of established practical authorities on the use and usability of methods. (HW)

SYMPOSIUM
ON GENERAL LINEAR MODEL APPROACH TO THE ANALYSIS OF EXPERIMENTAL DATA IN EDUCATIONAL RESEARCH

W. L. Bashaw and Warren G. Findley

Final Report
Project 7-8096
Contract No. OEC2-7-008096-0496

August 23, 1968

University of Georgia   Athens, Georgia

Final Report
Project No. 7-8096
Contract No. OEC2-7-008096-0496

SYPOSIUM ON GENERAL LINEAR MODEL APPROACH TO THE ANALYSIS OF EXPERIMENTAL DATA IN EDUCATIONAL RESEARCH

August 1968

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

SYMPOSIUM ON GENERAL LINEAR MODEL APPROACH TO THE ANALYSIS OF EXPERIMENTAL DATA IN EDUCATIONAL RESEARCH

Project 7-8096
Contract No. OEC2-7-008096-0496

W. L. Bashaw and Warren G. Findley

August 23, 1968

The research reported herein was performed pursuant to a contract with the Office of Education, U. S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

The University of Georgia

Athens, Georgia

TABLE OF CONTENTS

References

Anderson, T. W. The choice of the degree of a polynomial regression as a multiple decision problem. Annals of Mathematical Statistics, 1962, 33, 255-265

Bock, R. D. A computer program for univariate and multivariate analysis of variance. Proceedings of scientific symposium on statistics. (Thomas J. Watson Research Center) Yorktown Heights, New York: (1964).

Bock, R. D. Multivariate analysis of variance of repeated measures. In Harris, C. W. (Ed.). Problems in measuring change. Madison, Wisconsin: University of Wisconsin Press, 1963

Bock, R. D. Programming univariate and multivariate analysis of variance. Technometrics, 1963, 5, 95-117.

Bodewig, E.. Matrix calculus. Amsterdam: North-Holland Publishing Company, 1959.

Bose, R. C. Classnotes in least-squares analysis. Chapel Hill, North Carolina: University of North Carolina, 1960.

Box, G. E. P. and Hunter, G. S. Multifactor experimental designs for exploring response surfaces. Annals of Mathematical Statistics, 1957, 28, 195-241.

Bottenberg, Robert A. and Ward, Joe H., Jr. Applied multiple linear regression, Technical Documentary Report PRL-TDR-63-6. Lackland A.F.B., Texas: 6570th Personnel Research Laboratory, Aerospace Medical Division, Air Force Systems Command, March 1963.

Clyde, D. J., Cramer, E. M. and Sherin, R. J. Multivariate statistical programs. Coral Gables, Florida: Biometric Laboratory, University of Miami, 1966.

Cochran, W. G. Analysis of covariance: its nature and use. Biometrics, 1957, 13, 261-281.

Cochran, W. G. Some consequences when assumptions for the analysis of variance are not satisfied. Biometrika, 1947, 3, 22-38.

Cohen, Jacob. Multiple regression as a general data-analytic system. Mimeographed paper, March 1967. (Address: Graduate Psychology, N.Y.U.,21 Washington Place, New York, New York 10003)

Dixon, W. J. (Ed.) BMD biomedical computer programs. Los Angeles, Calif: Health Sciences Computing Faculty, Department of Preventive Medicine and Public Health, School of Medicine, University of California, September 1, 1965.

Draper, Norman R. and Smith, Harry, Jr. Applied regression analysis, New York: Wiley, 1966.

Durand, D. A note on matrix inversion by the square-root method. Journal of the American Statistical Association, 1956, 51, 288-292

Elston, R. C. and Bush, N. The hypotheses that can be tested when there are interactions in an analysis of variance model. Biometrics, 20, 1964, 681-698.

Finn, J. D. Multivariance: Fortran program for univariate and multivariate analysis of variance and covariance. Buffalo: School of Education, State University of New York at Buffalo, 1967.

Freese, Frank. Linear regression methods for forest research. Madison, Wisconsin: Forest Service, 1964.

Gaito, John and Wiley, David E. Univariate analysis of variance procedures in the measurement of change. In Harris, C. W. (Ed.) Problems in measuring change. Madison, Wisconsin: Univ. of Wisconsin Press, 1963.

Graybill, Franklin A. An introduction to linear statistical models, Volume I. New York: McGraw-Hill, 1961.

Harvey, Walter R. Least-squares analysis of data with unequal subclass numbers. Beltsville, Maryland: Biometrical Services, Agricultural Research Service, U. S. Department of Agriculture, Plant Industry Station, July 1960.

Hoffman, P. J. The paramorphic representation of clinical judgment. Psychological Bulletin, 1960, 57, 116-131.

Householder, A. S. Principles of numerical analysis. New York: McGraw-Hill, 1953.

Householder, A. S. The theory of matrices in numerical analysis. New York: Blaisdell, 1964.

Mood, Alexander M. and Graybill, Franklin A. Introduction to the theory of statistics. New York: McGraw-Hill, 1963.

Roy, S. N. and Bargmann, R. E. Tests of multiple independence and the associated confidence bounds. Annals of Mathematical Statistics, 1958, 29, 491-503.

Scheffe', H. A. The analysis of variance. New York: Wiley, 1960.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.

PREFACE

The Symposium on the General Linear Models Approach to the Analysis of Experimental Data in Educational Research was held in Athens, Georgia during June 29 - July 1, 1967. This report presents the major addresses and the discussion of particular methodological problems.

The Symposium was held to allow experts to discuss with each other the merits and limitations of the use of general linear models and least squares analyses in the analysis of experiments and quasi-experiments. The discussion is based on the consideration of related issues raised by the co-editors and the several participants. Thus, the discussion is indirectly related to the five major papers which are instructional in nature.

The five major papers were presented by five leading statisticians. Franklin A. Graybill, author of the definitive text An Introduction to Linear Statistical Models (Graybill, 1961) presented the introductory address. His comments throughout the meeting concerning practical considerations of analysis and interpretation should become well-quoted in the educational research literature.

The second paper, by Joe H. Ward, Jr., co-author of the widely used Applied Linear Regression (Bottenberg and Ward, 1963), was also instructional and was intended to show similarities between alternative analyses.

B. J. Winer, noted among educators and psychologists for his Statistical Principles in Experimental Design, was asked to discuss possible problems related to the linear models - least squares approach.

The fourth paper was by Rolf E. Bargmann, who was asked to outline particularly appropriate occasions for using linear models and least squares analyses. Bargmann presents some original research in this area which is not yet generally available elsewhere,

The last paper was by R. Darrell Bock. This paper draws together and critiques the previous four presentations. The reader will be especially interested in Bock's discussion of computer routines and his discussion of the proper analysis of repeated - measures designs.

The second major section of this book presents the discussion of relevant problems. This session was chaired by Warren G. Findley. The participants included the five major speakers and the following persons:

Harry E. Anderson, Jr., University of Georgia,

Elliot Cramer, University of North Carolina,

Robert Bottenberg, Personnel Laboratory, Lackland AFB,

Jacob Cohen, New York University,

Earl Jennings, University of Texas,

F. J. King, Florida State University,

Leslie McLean, Ontario Institute for Studies in Education, and

David E. Wiley, University of Chicago.

The discussion was tape-recorded and the transcription was edited by the co-editors. The speakers were not always identifiable, and in some cases, errors could have been made in the identification of persons making remarks. Moreover, the original intent of the speakers, in some cases, might have been distorted in the transcription and editing process.

No attempt was made to reorganize the discussion remarks. The reader will find many helpful suggestions and recommendations throughout this section. Hopefully, the editors have preserved in this document a little of the flavor and excitement of the discussion.

The editors would like to thank the many persons who shared in the work of this project. Rolf Bargmann and Larry Anderson were particularly helpful in the planning of the Symposium and the identification of other participants. In the long period between the original conception of the meeting and the preparation of this report, many typists and secretaries assisted the editors. We wish to express our special gratitude to

Mrs. Sherry Wilson,
Mrs. Carol Donaldson, and
Mrs. Patsy Jennings.

We wish to acknowledge the financial support of the U.S. Office of Education, whose grant to us made the Symposium possible. We are also grateful for the support of time, funds, and facilities of the University of Georgia and the Research and Development Center in Educational Stimulation.

August,. 1968

W.L. Bashaw and
Warren G. Findley

Introduction of Dr. Franklin A. Graybill

by

Clifford Cohen
University of Georgia

As a representative of the Statistics Department, I'd like to add my welcome to that which has already been extended to you who are participating here today. We are pleased that we are able to cooperate in at least a small way; and later in the program you will hear from a member of our department, Dr. Rolf Bargmann. I might say that Dr. Carl Kossack, who is chairman of our department, is unavoidably absent since conflicting schedules made it imperative that he be out of town at this time. It is regretful that he is unable to be with us.

Now, I would like to proceed to the task which I was requested to perform and it is indeed a pleasure to be called upon to introduce the speaker. He is, perhaps, best known for his very excellent book, An Introduction to Linear Statistical Models (McGraw-Hill, 1961) which has been quite widely distributed and very well received. Incidentally, this book is labeled volume one. Dr. Graybill tells me that volume two is coming along very nicely and will soon be released. Frankly, I think if he were inclined to do so, he could just rest on his laurels with his first volume but he's not that type of person. He is also quite well known for his work in a revised edition of one of the leading advanced texts, An Introduction to the Theory of Statistics (Mood, and Graybill, 1963). The first edition was written by A. M. Mood and the revision is largely the work of our speaker and Mood.

Now, besides these two works I have mentioned, which perhaps are the reasons for so many people knowing about him, he has published a number of research papers in the leading statistical journals, particularly the _Annals of Mathematical Statistics_ and the _Journal of the American Statistical Association_. He is a member of the Institute of Mathematical Statistics, the Biometric Society, and is a fellow of the American Statistical Association. His undergraduate work was done at William Penn College where he received his Bachelor of Science Degree and then he received his Master of Science from Oklahoma State University. His Ph.D. was from Iowa State University. He taught at Oklahoma State for several years before accepting a position as Chairman of the department at Colorado State University, a position which he still occupies. At Oklahoma State, he was, I guess you might say, Carl Marshall's right-hand man. I saw Carl shortly after it had been announced that Dr. Graybill was going to Colorado State and you would have thought Carl had lost his right arm. I just tell you that to let you know how much his former department head thought about him. Well, since going to Colorado he has been quite active; he's quite energetic in various and sundry programs and with that I will yield to our speaker, Dr. Frank Graybill.

INTRODUCTION TO THE USE OF GENERAL LINEAR MODELS
IN THE ANALYSIS OF EXPERIMENTAL DATA

Franklin A. Graybill
Colorado State University

You know one of the nice things about a meeting such as this is that you get to know each other on a first-name basis. I've been involved in a number of these with groups of geologists and biologists--none with education people, however. It seems that you run into each other from time to time at meetings and various places like this and I think it is a good opportunity to get together and find out what's going on. I'm a little dismayed to be the first speaker in a sense that I suppose being the first speaker is sort of like calisthenics in the morning--get them out of the way and get going.

I wanted to know what credentials I could bring to a group like this. You see, you have me at a disadvantage; you all know some statistics and I know nothing about education even though I'm heavily involved in the educational process or, at least, I think I am. I thought back as to what I could bring to increase my status with you and one thing I can say is that my undergraduate degree led to my receiving a high school teaching certificate, so I guess that's something.

Well, since this Symposium does involve linear models, I will say something about this. But I think that before linear models should enter, I must say a little bit about some of the techniques in statistics that I presume will be discussed and debated here today and tomorrow.

I'd like to preface my remarks with a few words about statistics in general since I believe that from time to time during the Symposium certainly our deliberations will lead us to some of the foundations upon which the theory of linear models must rest.

I presume I will not say anything today that is new but I may say something that is controversial. I hope everyone does; this is the way we can get ideas across. And even though I may say something that will be controversial to some people, I won't take time in every instance to point out every side of controversial statements. It is popular these days, and almost mandatory, for a statistician to declare his party affliation. By that I mean, in particular, we are seeing an influx of what we call Bayesian statisticians these days. I don't know whether you've been involved in Bayesian statistics or not, but in a political venacular, I'm an independent. You might say that I'm a fence strattler. I'm Bayesian when I think it's appropriate and non-Bayesian when I think it's demanded. So I guess in sophisticated language, I'm what you'd call a neo-Bayesian. But I think that we're all Bayesians in the sense that we must bring to bear upon each of our problems--not only in a scientific atmosphere, but in everyday affairs of men and women--to all the knowledge we have. One way to do this in statistical formulations is through what we might call Bayesian methods. Bayesian formulations are not very well defined yet and perhaps never will be. If we're interested in this topic in the symposium, we might say a little bit about it later.

Now statisticians, I think, and statistical communities today are divided into two groups--not by natural division or by any intended thing--but I think we are divided. There are the mathematical statisticians who really don't live very close to data nor who care much about data. Then there is a group, and I think a growing group, that feel that the real job of a statistician is to be a data analyzer. With the advent of the computer our problem is made not easier but perhaps more difficult; because we can make wrong decisions and use wrong methods and techniques much more quickly and make many more errors than we could a decade ago. But, be that as it may, I presume that we as data analyzers are to take a set of data and make some sense out of it.

We want to talk about linear models and so I will lay a ground work so that we might have a starting point. We might start off by saying that we live in two worlds. We have what we might call the real world and the abstract world. Now, the abstract world is a world of symbols, of conceptualization, and so forth. Those of you who have known mathematicians and have thought from time to time that mathematicians are out of this world, what you probably mean is that they are out of the real world; they're in this abstract world.

But, in contrast to this abstract world, I think it helps, at least it helps my thinking, to focus on the real world. This is the world of the senses, perhaps, the world of measurement--this is the world we really live in. In the abstract world we would include the world of thought--what our thoughts, our reasonings, and so forth are. I think the problem of modeling, not only linear modeling, is to dip freely back and forth from the real and abstract worlds.

For example, Galileo dropped rocks from the leaning tower, and from these acts he developed a formula relating time and the distance that a body falls under free flight. Now if he obtained this result by working in the real world, he looked at the data. I don't know if the data indicated 1/2 gt^2 or maybe it indicated 1/2 $gt^{1.99999}$. But anyway, he arrived at 1/2 gt^2. Now, by dipping into the abstract world of symbols, we can obtain the velocity at any given time and the acceleration at any given time. We can obtain a number of things like that. Galileo could have also done exactly the same thing by working completely in the real world.

But what I'm saying is that the reason for modeling is so that we can work in the abstract world and save a great deal of time. And, not only save time, but perhaps by working in this abstract world, which is much easier, we can fit in ideas, thoughts, perhaps new techniques, and so forth, that we may never be able to synthesize in the real world. It is a condensation of ideas using symbols. I think it is extremely important. But I think it is also important that if we go to the abstract world to do our manipulations from time to time we get back into the real world to check these calculations, to check these equations and symbols. This is, in my opinion, why we try to model either simple or complex situations.

Now, I would like to work toward the goal that I will call a fundamental proposition as far as modeling is concerned. You'll notice that I will almost never use the words, "cause" and "effect." I think that technically speaking cause and effect are very difficult to defend. But, nevertheless, I'd like to take the following as a rough proposition.

First, y is some measureable quantity in the real world and we want to predict it. However, it is something we'd like to predict without measuring it. We can think of a lot of examples for y. We take as a proposition that there exists a finite number of quantities that are not directly related to y, and a function, f, such that, if these quantities were known and if the function f were known, then I could predict y exactly.

There are some who will find fault with this fundamental proposition, but in spite of the fault that they find--and there is some--I think that this is the way a scientist acts. I think what scientists do is decide to predict or describe some quantity. Then they pick out some other factors that they think have a bearing on this quantity of interest--factors that they believe will be useful in understanding the system, that somehow determine or drive the system, and they try to find a mathematical model relating the factors and the quantity of interest. The use of this derived model puts us in the abstract world. We work in the real world with all of our knowledge to decide on the factors that contribute to driving the system that determines y. Then we put these factors with some kind of a formula or function and we use these to try to predict y.

Sometimes we use "functions" in our prediction formula in a very loose sense. For example, we know that when it rains, there must be clouds. When there are clouds, it is more apt to rain than when there are not clouds. If my father and mother were very tall, I would expect my children to be tall. If my father and mother were short and I'm short, then perhaps my children will be short.

If we find out that these statements or derived models do lead to some measure of predictability--but not perfect as, of course, they never would be--then what we try to do is to find other factors that also contribute to driving this system and we bring in these additional factors to try to have a better prediction under more and different varying circumstances.

I think this is an idealized way to look at modeling. We believe there are factors that we can find, that we can observe in the real world; we can use some kind of a symbolism or some kind of equation or formula, and use it to predict quantities in which we're interested. We believe these factors somehow help determine and drive this system of interest and yet they're not directly associated with it. As an example of this "indirect relationship," consider the prediction of a variable y. I'm going to measure the square root of y and square it--I've got a perfect predictor for y. One predictor, $\sqrt{y}$, is directly related to y. This is not what we're looking for; this is not of very much help if any help at all to us.

One of the objectives in science is to describe, predict, and relate quantities in the real world, and mathematics is used to describe connections between events, but mathematics doesn't prove the statements. Mathematics is not a science of truth; it's a science of logical reasoning. This is why we use mathematics to do the modeling. We certainly think logical reasoning is called for and mandatory. It'll tell us something about the relationships in the abstract world, but it will tell us absolutely nothing about the real world. And so the input to these formulas that we use in the abstract world are what is really important. Now, of course, it is important that logical reasoning be instituted and used in the best sense, so it is important that we know how to manipulate these quantities in the abstract world, but it isn't the whole answer.

Now, I'd like to continue this a little further to show you how it is related to linear models. Linear models are very special cases of more general models. Linear models are the only ones that have been developed very far and in some sense, perhaps, the only ones that ever will be, but we may be able to do a little bit more with non-linear models than we have in the past. I will use as an example something very simple--the prediction of the height of an individual. Let us assume that I'm trying to find some factors that will predict what the height of an individual will be when he reaches a certain age. Suppose there are n factors $X_1, X_2, \ldots, X_n$ that "determine" height. Suppose we find two factors, X_1 and X_2, that will be important. Maybe these are the heights of the parents of this particular person whose height we are trying to predict at a later age. We find some kind of a function of these factors, say $f(X_1, X_2)$. In other words, we're looking only at two factors in our prediction. The model can be written $y = f(X_1, X_2) + g(X_3, X_4, \ldots, X_n)$. We know that these two factors, X_1 and X_2, are not the only quantities that determine an individuals height at a certain age, because if we observe many people whose parents have the same height, these people would probably all differ in height. We know that there might be other relevant variables--X_3, X_4, ..., X_n, for example, diet, grandmother's height, etc. In other words, I make an observation of an X_1 and an X_2 for an individual. I observe another individual and he has the same X_1 and X_2 values, but the heights of the two individuals differ. The reason is that other factors are really affecting height but we haven't brought them into our model. We consider the non-used factors in the function $g(X_3, X_4, \ldots, X_n)$ and examine the way $g(X_3, X_4, \ldots, X_n)$ varies when X_1 and X_2 are held fixed. We treat $g(X_3, X_4, \ldots, X_n)$ as a random error and write $y^* = f(X_1, X_2) + e$.

This is the first approximation to understanding--we begin to lay it out, stretch it out, tear it apart, find out what factors drive this particular event of interest. We may collect some data and estimate the variance of an error term. If the variance is zero, this means that we have an exact predictor. This never really happens or I've never known it to happen, but the variance might be quite small. If so, we have a population of heights that can be predicted with quite good accuracy and perhaps enough accuracy to solve our problem.

I think we're never interested in predicting the height of an individual to the nearest one ten billionth of an inch, or closer, as we might if we were looking for what we call deterministic model or point deterministic model. We might settle for what we call an interval deterministic model. With an interval deterministic model we would predict, for example, height to within a millimeter, because for all practical purposes and even many impractical purposes, I'd have my problem solved. The deterministic models are formally stated and summarized in Table 1.

Table 1

Mathematical Deterministic Models Summary

Fundamental Proposition

For any Y there is a function f and variables $X_1, X_2, \ldots, X_n$ such that

$$y = f(X_1, X_2, \ldots, X_n).$$

The three "levels" of deterministic models

1. Point deterministic: $y = f(X_1, X_2, \ldots, X_n)$.
2. Interval deterministic: $y = f(X_1, X_2, \ldots, X_n)$,

   where each X_i is observed to be in an interval $X_i \pm \varepsilon_i$.
3. Level 3 is the same as 2, except that y is observed to be in

   an interval $y \pm \varepsilon$ with a probability less than one.

So the goodness of our model depends on the error variance. If it takes various individuals and measure height of their mother and father, and their predicted heights vary as much as 50 pounds for the same mother's height and the same father's height, then my variance is too large, my prediction is not precise enough. Therefore, I may decide there are other factors entering into this system or I may have the wrong function relating the factors.

Suppose I decide that there are other factors entering in and such as X_3 that has to do with diet. We find a different function and now try to measure y as a function of the three variables. Suppose I have several observations, each of which has the same mother's height, the same father's height, and the same nutritional measurement. If these individuals have different heights, then I do not have either the correct function or all the variables that go into exactly determining the height. However, again, if I can put a probability distribution on $g_2(X_4, X_5, \ldots, X_n)$ at least to a first approximation, g_2 varies and acts as a random variable, then I could write my model as $f_2\ (X_1, X_2, X_3)$ plus another random error which is different from the earlier one. Now, let's again examine the variance of that random error. In other words, what I examine really is how $g_2\ (X_3, X_4, \ldots, X_5)$ changes when X_1, X_2, and X_3 are held fixed to tell me more or less whether these three factors are contributing enough to the prediction of height. Suppose I had examined the error variance of the two-factor model and found it too large. It gave me, perhaps, a distribution that has a spread of let's say six inches. So when I measured the mother's height and the father's height, I still could not predict within less than six inches so it wasn't precise enough.

So I brought the third variable into my model and got a three-factor model. Now I examine the error variance of the new model. If the variance is a great deal smaller than the variance of the two-factor model, then the X_3 variable has done some good. It means instead of six inches of spread on the prediction, I have reduced it to something like, say, one inch, and if the variation is now one inch, perhaps I will decide that I've got a model that is good enough for my predictive purposes. I realize and recognize and know that I will not be able to predict the height of an individual exactly by only knowing these three quantities of input and this particular function of the model because there are other important things that will determine the height that causes individuals with the same X_1, X_2 and X_3 to have different heights. But if the heights don't vary too much--if the error variance is tolerable so to speak--then I've got a pretty good model and so I say "Here is the model I will use."

Now, to continue this a little bit further, we ordinarily don't use, at least in the initial stages of experimentation, just any function. We ordinarily use what we call a linear function. And when we say linear function, we mean linear in the unknown parameters. We don't care about linearity in the X's--it may be logarithms, exponentials, squares, cross-products, almost anything like that--but the model must be in the unknown parameters.

Now, let's digress a moment. I need to point out that in modeling of this kind, there are two types of errors. We may make measurement error in trying to observe our variables. For example, when any kind of continuous variable is involved we know we make measurement errors. If somebody says "What's my height?" I can't tell you exactly what it is.

There's a measurement error involved. The only time there isn't a measurement error involved is when we count. For example, how many people are in this room? I presume we could count the number of people in this room and everybody in here would agree. Now, if I had indentified the exact function and the exact factors that drive this function to predict our variable y, if there were errors involved in measuring the X's and an error involved in measuring the y, I would not predict the exact value of y's, just due to measurement error.

On the other hand, there is a second kind of error. Suppose I could measure all of the quantities X_1, X_2, and X_3 exactly--and yet when I try to predict y by using these three quantities and the function $f_2(X_1, X_2, X_3)$ I don't predict exactly because there are other things contributing to y. This is what I would call an equation error. It's an error in predicting because I don't have the correct equation. There are two general ways I can have an equation error. My equation may have important predictors omitted or the function chosen may be the wrong function.

I think the fact of the matter is that in every real world situation, the reason we don't predict some things exactly is because we make both types of error. We make measurement error and we also make an equation error. Now the question is which of these in a particular situation should we examine more closely; which should we try to take into account? What we've discussed primarily today is trying to take into account the equation error. Our first model might not be the right equation. The second model might not be the correct equation either, but it's more correct, let's say, than the first one if we have been good experimenters with good insight. There is still some equation error involved but there is also undoubtedly a measurement error involved.

Now, there are some sciences, and education is one, where measurement error is a very real and important error and may be equal in magnitude in many situations to the equation error. We may have a good equation but our measurement error may be so high that it somehow invalidates our model as a good prediction equation.

Now, let's try to use our model to relate the real world and the abstract world. If we can enter the abstract world and find this model then we can go through our mathematical manipulations and do many things and find out perhaps even some new things without doing all the experimentation that would be necessary if we worked only in the real world. Now, there are some reasons why we can't bridge the two worlds. There are many reasons, in fact. One, as I said, mathematics perhaps is too exact and maybe it is. But it is exacting and that's just the way mathematics finds itself. Another thing is real world quantities are not well defined. For example, height is not really well defined and we need some kind of operational definition. I'll say I'll take a certain kind of measurement device and I'll measure a thousand times and take the average. It's an operational definition, but height itself is not well defined at all. We could, I suppose, hold the whole symposium in discussing what we mean by the length of this table. If we mean length perhaps to the nearest foot, the problem is solved. But suppose I said I needed to know the length of this table to the nearest micron. Well, the problem is not solved and so we would have to solve it and there are very deep difficulties with something even as simple as that.

Another reason why perhaps it is difficult to relate the real world and the abstract world through mathematical modeling here is that there's measurement in the real world. We've got to live with it. Physical sciences sometimes solve this in many of their problems, so measurement error is not too important to them. Measurement errors have a way of accumulating when you take a number, when you square it, when you take logarithms, exponentials, and etc, and so even in the physical sciences where the measurement error is very small, it offers some problem.

Another difficulty in relating the two worlds is the difficulty in holding a certain quantity fixed in the real world while others vary. Mathematically you can do this very easily. Say, let's hold X_1 fixed, let X_2 vary, and see what happens. Well, you, perhaps, can't really do that in the real world. You can't hold height fixed and let weight vary. You couldn't let weight go up to 300 pounds and the height be 18 inches or something like that. It just doesn't make sense, so it's difficult to do that in the real world and so we have to be careful.

Another difficulty in relating the two worlds is that real world quantities may not be independent in mathematical sense. That is why we can't hold some constant while others vary. We can consider them mathematically independent, but in the real world that may not be the case. Now, this is one reason why we've got to be very careful when we use the computer. We run away with ourselves and come up quite often with nonsense.

Finally, there are no reasons why we should be able to relate these two worlds actually.

Now, the feeling is that we as mathematicians and as statisticians many times spend a great deal of time getting precise solutions to the wrong problem and I think we could, perhaps, spend time better getting approximate solutions to the correct problem.

Now, I'd like to classify these particular models. These models are called quantification models. They're models to be used when the variables are quantitative variables, variables that I can measure, variables that have interval or ratio scales. These are things like height and weight, things that have a pretty good definition of measurement even though maybe I can't necessarily measure them. Now, there are a number of classifications that could be made here and I would just make some of them before I start. We've got to bring in linear models here. We'll use X's to denote independent variables and y to denote the dependent variable.

The first model is where the independent variables are pre-selected real variables—not random, they're pre-selected. We use this all the time. We decide well, I'm going out and get somebody that weighs 50 pounds and his mother weighs 140 pounds and things like this. They're actually pre-selected. Now, of course, y is always a random component, a random variable.

The second case is where y, X_1, X_2, and so forth are jointly random variables. Now, this may be a situation where I go out and select people at random. Let's say if I'm talking about height and weight, I select people and they have a height and a weight. I select another person, he has a height and a weight, so the height and the weight, the y and X, are random variables. In the former case we purposely select somebody that weighs 50 pounds and measure his height, select somebody that weighs 60 pounds and measure his height, and select somebody that weighs 70 pounds and measure his height. This would be the first type. In the second case, I might take the person at random someway, but his height and weight would not be pre-determined by me.

Now, the third situation is either case one or two when some or all of the independent variables can not be observed. Now, that is really almost always the situation. For example, in the height and weight example, I can't really observe height and weight. You have a true height, and a true weight, but I can't observe it. What I observe is a measurement of your height and weight and if this is not truly your height and weight, it turns out that a different kind of model is appropriate than is needed in cases one and two. I don't observe these y's. I observe the X's plus some kind of random errors. You might say, "Well, these models almost look the same." But the difficulty is the appropriate solution to case three has never been found.

The particular problem here has never been adequately solved. There is some indication that it never will be, that the situation where I have errors in measuring my independent variables is not amenable to actual solution. Now, I say this so we might keep it in mind. Now, if the measurement error or the X variance of the measurement error is very, very small relative to the magnitude of the quantity measured, then the method we go through in the first two cases are very good and almost the same as would be if I measured them without error. But the problem exists and sometimes it is sort of glossed over and we don't recognize it. This is sometimes called "regression" or "linear model with error of observation in the dependent variable." It doesn't matter if we have errors of observation in y. It just increases our variance and we don't do quite as good a job, but we can still do the job we set out to do. If there are errors of observation on the X's, then that problem has not been solved and can offer some real difficulty. I think we need to be aware of that.

There are two things we are generally interested in doing in these models. One is predicting, the other is to estimate, and these are two different problems.

If we are interested in only the prediction problem, predict the height let's say of an individual, we don't care perhaps what the parameters in the model are as long as we have a good model for prediction. If all we want to do is predict the distance a body falls in a free flight in a vacuum under the influence of only gravity, we may just like to be able to predict it as a function of time. This is what I call a prediction problem.

Now, the estimation problem means that there are some reasons why the model parameters are important. They may be important in their own right, not only important in just being able to predict. For example, in Galileo's model the constant of the model is a measure of the gravitational constant. So while I would need to know the constant fairly accurately to have a good prediction, I need to know it for some reason in its own right and so I'd like to estimate it. Usually, regression constants may not have particularly important properties to me, but I'm just trying to obtain these factors to get a good prediction of y, so I'm interested in good estimation of weights only in so far as good estimation would lead me normally to good prediction.

The problems I think are quite clear. I think that there are times when we want to do one and times when we want to do the other. Well, that's all I have to say about this particular model--what I call the quantification model, or quantitative variables model.

Now, I'd like to turn just for a minute to what I call qualitative models. These have ordinarily been called analysis of variance models. Now, to a mathematician and a mathematical statistician, they're exactly the same, but you know in some ways, the interval from zero to one is exactly the same as the interval from 10,000 to 10,001, but if you're going to have your salary in dollars per year in one of those intervals, I think perhaps you wouldn't want to consider them the same. So, while to a mathematician and a mathematical statistician, what I call the qualitative or analysis of variance models can be obtained as a special case, mathematically speaking, from the models on the quantitative variables that I just discussed, I don't think there is much value in that except perhaps as a teaching aid. You just have to go through theory once, but to someone who is going to use it, even though as I say, mathematically they're equivalent, I think it's very important we go through these models independently.

Now, these models generally can be written

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

The μ's are what we call the means; the ε's are what we call the random component, and both the μ's and the ε's are unobservable. The y's are what we observe. The subscripts i and j take the place of the X's. Subscript i takes the place, for example, of one X; j of another, etc.

I shall classify these into three categories and I'm classifying on the nature of the μ's: one, the fixed effect models; two, the random effect models; and three, the mixed, or random and fixed effect models.

Now, let's just take the fixed effect model and sub-classify it. Let's consider a two-way classification of data. Let's use n_{ij} to mean that the number of observations in the ith row and jth column. If n_{ij}

equals zero it means there are no observations in the ith, jth group; that is, y_{ij} doesn't even exist. We say that a model is "complete" if all n_{ij} are positive. If at least one n_{ij} is zero, then we have an "incomplete" model. In other words, if all cells have something in them, we have a complete model; if there is at least one cell with no data we call the model "incomplete."

In our breakdown we could have a tree of classifications. The first breakdown is models with and without interaction. This is very important. Models specifying interaction, we call "non-additive." Suppose in order for discussion, we have one observation per cell. If we want to check for presence of interaction we have a test due to John Tukey and variations of that test. A second breakdown is by cell sample sizes. First, we could have one observation per cell as above. Second, suppose we have more observations in each cell, but the number of observations per cell is the same for each cell, say $n_{ij} = m > 1$. This can be analyzed by convential ANOV. Third, we may have an unequal number of observations per cell. There are two things we may want to do here. We may want to subclassify this. If we have equal numbers, things are pretty straight forward. You can check interaction, you can check what we might call main effects, and so forth. If we have a complete design with unequal numbers, let's say more than one in some cells at least, we can check for interaction and we can check main effects and I call both of those conventional methods. The main effects are estimated by unweighted means analysis.

Now, let's consider additive models, or models that specify no interaction. First, consider the case with one observation per cell. This has a conventional analysis. Second, consider additive models with equal numbers in each cell, but the n_{ij} are greater than one ($n_{ij} = m > 1$).

This model also has a conventional analysis. The third one is, of course, its counterpart, with unequal numbers in the cells. These various models are summarized in Table 2.

There are problems we need to solve after we have the model. We need to examine the data for interaction--where and what kind. This is basic, not only find where the interaction is, but what kind of interaction is present. Is it a strengthening kind or is it a reversal kind? These are very important problems, it appears to me. We perhaps want to examine for row and column effects. If there are effects, where are they? What's operating here? What's pushing the system? More important than any of these is to examine individual cell effects.

The first thing I think should be done in a situation is to find the sufficient statistics. Now, if you're not well trained in mathematical statistics, you may not know what that means. But what it means in a nutshell is this--reduce the data as far as possible without losing any information. For example, if you have data here that involves, let's say 500 observations, you may be able to reduce that to 50 observations without losing any information under the model which you are assuming. Again, don't take the model as gospel too much. You should go back and examine the model. You should use your data not only to check what you started to determine, but you should also use it to examine the model in which you use it. I think one should reduce the data as far as possible without losing any information--to what we might call the smallest set of sufficient statistics. It's just a matter of simplicity, it seems to me, it's easier to look at 50 numbers and read something out than it is to look at 500.

Table 2

A Summary of Some Fixed Effect, Complete Models[1]

A. Interaction (Non-Additive)

1. One observation per cell; $n_{ij} = 1$; Tukey test for interaction
2. Equal cell sample sizes: $n_{ij} = m > 1$; conventional analysis
3. Unequal cell sample sizes; n_{ij} not equal for all i and j

    Interaction--conventional analysis

    Main effects--unweighted means analysis

B. No Interaction (Additive)

1. One observation per cell; $n_{ij} = 1$; conventional analysis
2. Equal cell sample sizes; $n_{ij} = m > 1$; conventional analysis
3. Unequal cell sample sizes; n_{ij}: conventional analysis

[1] This breakdown can also be applied to random components models, and incomplete models.

I think the important thing in a two-way classification is to estimate the cell means. You may not want to do this; you may have other things you want to do. But this is what we should estimate to begin with, it seems to me. Preliminary analyses are very useful.

Another thing is to use some kind of a technique to milk the data after you do the things for which you set up the experiment, because this is where you get the new ideas.

I think something else should be done as you use the data to examine the model. Use the data in another way to decide how you could go on to another experiment in sort of a sequential fashion to improve on the result--to either confirm or deny your conclusions.

There is another problem that is important and deals with the Bayesian point-of-view. If I use a linear model, it's a very formalized thing. You all have a great deal of knowledge to bring to bear on a problem you cannot model. I think that this is where the idea of what I shall call the target and sample population was developed. There are two populations. You sample populations, and from that sample, using statistical techniques, you can obtain probability inferences to that population that was sampled. That perhaps is not the real population you want to discuss. The real population you want to discuss is called the target population. You sample one population. You can draw valid probabilistic statistical inferences to the sample population. Then the population you're really interested in, the target population, must be given consideration. After you have the information on the sample population, the inferences you draw to the target population are perhaps non-probabilistic, more personalistic.

Now, I recommend some "don'ts" and I wish we could discuss these. First, "Don't use a statistical test." There is a one-to-one correspondence between statistical tests and statistical confidence intervals. For example, mathematically, in a sense, they're equivalent; but the way we think about these things is very, very far from equivalent. I would like to discourage the use of any kind of a statistical test and even talking of the phrases "statistical significance" or "tests of hypothesis." Use confidence intervals where possible.

The other "don'ts" are "Don't be restricted by absolute pre-conceived linear models," and "Don't reduce the data too far." Show the cell means. Show the sufficient statistics.

Introduction of Dr. Joe H. Ward, Jr.

by

W. L. Bashaw
University of Georgia

I would like to introduce Dr. Joe H. Ward, Jr., who is the next speaker. Dr. Ward is a Texan who earned his Ph.D. at the University of Texas. Our major reason for choosing him to speak is his co-authorship of *Applied Linear Regression*. This is a book that has been very useful to many of us since it was one of the few documents that has been available on the subject over the last several years. A secondary reason for asking Dr. Ward to participate is that he has, over the last few years, been traveling around the country giving lectures and workshops on general linear methods. We certainly would have been amiss if we had not gotten him on the program.

Some of you will be surprised at his institutional affiliation. All of you will identify him with Lackland Air Force Base's Personnel Research Laboratory. I understand he has been at the Personnel Lab now some seventeen years. This year he is on a one year leave of absence so he has not broken his connection with the Lab. At the present time he is Program Director for the Southwest Educational Development Laboratory.

A few of his research areas might interest some of you, in addition to myself. One, of course, is application of linear models, a second is homogeneous multivariate grouping. This is a set of grouping techniques that was developed for grouping many things. My interest primarily is grouping people although the techniques are used for such things as grouping regression systems and things of this sort. Third, many of you will know him primarily for his computer programming, particularly for the Persub Programming System. This is a very complete set of matrix subroutines that can be tied together to do any set of matrix operations.

Finally, I would like to say that Dr. Ward has consented to give us the one day workshop on Saturday. This was his idea and I would like to repeat, for those of you who have not been informed, everyone is welcome to attend.

Dr. Ward's paper today is "Synthesizing regression models, an aid to learning effective problem analysis."

SYNTHESIZING REGRESSION MODELS--
AN AID TO LEARNING EFFECTIVE PROBLEM ANALYSIS

Joe H. Ward, Jr.
Southwest Educational Development Laboratory

Regression models can be used to assist in the analysis of a wide variety of problems. However, the power of regression models is not widely utilized. There are two major reasons for the lack of use of general regression models. First, there have been too few attempts by teachers to develop the behaviors in students that are necessary to effectively create models appropriate to the particular problem of interest. Second, many of the models that *should* be utilized for a particular problem require the use of a computer, but many research workers do not have effective software systems to facilitate communication with the computer.

These two problems can be helped by 1) providing an instructional system that will develop in students the capability of defining regression models appropriate to their problems of interest; and 2) providing computational software that facilitates the analysis by a high speed computer.

Even though both of the above areas are important, the first--defining appropriate models--is the most important and difficult behavior to bring about in research workers. The following presentation will be devoted to the discussion of several aspects of this problem. First, a few general comments will be made about the general problem of teaching (and learning) techniques of model generation. This will be followed by a specific example of an instructional approach--a description of a synthesis of several different regression models.

The Generation of Models

Some of the intermediate behavioral objectives that lead to effective model generation are:

1. A research worker should be able to define the vector of interest (i.e., dependent vector) appropriate to his problem. This is developed by extensive practice on problems with increasing difficulty.
2. A research worker must develop the capability of expressing his vector of interest (call it Y), as a linear combination of appropriately defined vectors (call them X(1), X(2), ..., X(k)) plus an error vector (call it E). Extensive practice in defining vectors is required to develop the desired capability. The research worker should think "I need to find 'another name for Y' so that the statements that I make about this 'other name' will be relevent to my problem." A student should have extensive practice in defining those vectors which are to be used in the "remaining" of Y.
3. After the vector of interest (Y) has been expressed as a linear combination of the new vectors (X(1), X(2), ..., X(k)) plus an error, the research worker can then make statements (or hypotheses) about "expected" or "predicted" values of Y. This involves the translation of the research question from natural language into the language of the mathematical

representation. This translation process is sometimes quite difficult, and the student should have extensive practice, using simple problems in the beginning.

4. The translation of statements about the model leads to the imposition of restrictions on the model. The student should impose his restrictions on the model to determine the effects of the restrictions on the error. It is sometimes helpful to view these restrictions as the "giving up" of information.
5. After the student has imposed the restrictions it is important that the student verify that the restricted model actually does possess the properties imposed by the restrictions. This serves as a check for the student's substitution. It also provides frequent insights into previously unrecognized properties of various models.

A Synthesis of Regression Models

The following illustration is designed to show the idea that is common to four regression models that are often treated quite separately in our instructional programs. The basic problem of interest is the same in all four models; however, the models appear quite different due to the differences in the original assumptions that were made for the four models.

For our example, we consider four different research workers who are studying the effects that different amounts of practice have on typing proficiency. Furthermore, there is some concern by these research workers for the possible "contaminating" effect of the age of the students on the research results. Each research worker feels that something should be done to "hold age constant" or "take out the effects of age." However, a couple of the researchers aren't quite sure what they must do to "take out the effects."

Now the first research worker was located at a university (ANOVA U.) where there was strong emphasis on analysis of variance--with very little instruction in covariance analysis, or multiple correlation and regression. And this research worker was particularly fond of the "two-factor design." The second researcher was at a university (COVARIA U.) which to no one's surprise was really strong on covariance analysis. This school had a complete course in covariance analysis to stress its importance. The third worker received his education many years ago at a university (MULCOR U.) that had only taught multiple correlation and regression analysis. The analysis of variance and covariance course was started the year after he completed his statistics course.

The fourth researcher had attended a university (VARICO U.) that had stressed a slightly different approach which they described as "a sort of reverse covariance analysis" which they have named VARICO ANALYSIS.

All four of these research workers have conceptualized a common problem. First, they are all interested in studying the effect of practice on typing proficiency while "controlling" or "taking out the effects" of age. Furthermore since they are dealing with the age information, they all wish to test for interaction since it may be that the effects of amount of practice are different for students of different ages. All four are interested in, first, testing for interaction, and then if there is no interaction they will test for the effect of amount of practice.

Even though these four research workers have a common problem, each one would probably perform quite different analysis because of the varied educational emphasis. Also, they might each argue that they are doing quite different analyses. These analyses appear even more different because the computational procedures appear quite different.

The four different approaches will be presented below in a form that will emphasize that there are basic ideas common to all.

Approach 1 - (ANOVA U.)

The research worker at ANOVA U. wishes to think of his problem as a "two-factor design."

We assume for all problems that there are observed 15 levels of practice (16, 17, ..., 30 hours) and that there are 5 ages (14, 15, ..., 18 years).

We define the following vectors:

Y = a vector containing the typing proficiency scores of the n students in the study

X(i,j) = a set of vectors with elements defined as 1 if the corresponding element of Y is observed from a person with practice hours i, and age j; and 0 otherwise, (i = 16,17,...,30), (j = 14,15,..., 18)

Notice that if some X(i,j) vectors are null they are not included in the model.

E = a residual vector

Then the full model is

$$Y = \sum_{i}\sum_{j} a_{ij} X(i,j) + E$$

or in the extended form

$$\begin{aligned}(1)\quad Y = {} & a_{16,14}X(16,14) + a_{16,15}X(16,15) + \ldots + a_{16,18}X(16,18) + \\ & a_{17,14}X(17,14) + a_{17,15}X(17,15) + \ldots + a_{17,18}X(17,18) \\ & \quad\vdots \qquad\qquad\qquad\quad \vdots \qquad\qquad\qquad\quad \vdots \\ & + a_{30,14}X(30,14) + a_{30,15}X(30,15) + \ldots + a_{30,18}X(30,18) + E\end{aligned}$$

Figure 1

Extended Form of Vectors of Model (1)

Y	X(16,14)	X(16,15)...	...X(30,14)	X(30,15)...	...X(30,18)
$y_{16,14,1}$	1	0	0	0	0
$y_{16,14,2}$	1	0	0	0	0
$y_{16,14,3}$	1	0	0	0	0
$y_{16,15,1}$	0	1	0	0	0
$y_{16,15,2}$	0	1	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮
$y_{30,14,1}$	0	0	1	0	0
$y_{30,15,1}$	0	0	0	1	0
$y_{30,15,2}$	0	0	0	1	0
$y_{30,15,3}$	0	0	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮
$y_{30,18,1}$	0	0	0	0	1
$y_{30,18,2}$	0	0	0	0	1
$y_{30,18,3}$	0	0	0	0	1

Predicted (or expected) value for an individual who practiced 16 hours and who is 15 years old.

$$E(16,15) = (a_{16,14}* 0) + (a_{16,15}* 1) + (a_{16,16}* 0) + \ldots + (a_{30,18}* 0)$$

$$E(16,15) = a_{16,15}$$

In this discussion different symbols will not be used to distinguish between the unknown parameters and their least squares estimators. In the model above the symbols a_{ij} will be used to represent both the unknown parameters and the estimators.

Consider four different students having the following characteristics:

Student	Hours of Practice	Age in years
1	r	p
2	s	p
3	r	q
4	s	q

The hypothesis of no interaction can be stated as follows:

The difference between expected (or predicted) typing performance of the two students at age p but with different practice levels r and s is equal to the difference between the expected (or predicted) typing performance of the two students at age q but with different practice levels r and s. This is hypothesized for all values of p, q, r, and s where $p \neq q$ and $r \neq s$.

Calling the four expected values E(rp), E(sp), E(rq), and E(sq) *the hypothesis of no interaction is*

$$E(rp) - E(sp) = E(rq) - E(sq).$$

Now in the model employed by the ANOVA U. research worker

$$E(rp) = a_{rp}$$

$$E(sp) = a_{sp}$$

$$E(rq) = a_{rq}$$

$$E(sq) = a_{sq}.$$

Then we see that the *hypothesis of no interaction is*

$$a_{rp} - a_{sp} = a_{rq} - a_{sq} \qquad \begin{array}{l} p \neq q \\ q = 14, \ldots, 18 \\ r \neq s \\ s = 16, \ldots, 30 \end{array}$$

When we impose these 56 restrictions on the model the restricted model can be written as

$$\text{(2)} \quad Y = a_{16}\,X(16) + a_{17}\,X(17) + \ldots + a_{30}\,X(30)$$
$$+ b_{14}\,Z(14) + b_{15}\,Z(15) + \ldots + b_{17}\,Z(17) + R$$

where

$X(i) = 1$ if the corresponding element of Y is from a student who practiced i hours;

0 otherwise $(i = 16,17,\ldots,30)$

$Z(j) = 1$ if the corresponding element of Y is from a student having age j.

R = the residual vector

a_i and b_i = unknown coefficients

Notice that $Z(18)$ is not included in this model since it is a linear combination of the other vectors.

Let

$q_1 = \sum_i E_i^2$, the sum of squares of errors in the full model

$q_2 = \sum_i R_i^2$, the sum of squares of the error in the restricted model

Then if the F statistic is desired to test the hypothesis we have

$$F = \frac{(q_2 - q_1)/(75-19)}{q_1/(n - 75)}$$

Now we will consider the situation in which the no-interaction hypothesis has been accepted as true.

Then we use the model

$$Y = a_{16}\ X(16) + a_{17}\ X(17) + \ldots + a_{30}\ X(30)$$
$$+ b_{14}\ Z(14) + b_{15}\ Z(15) + \ldots + b_{17}\ Z(17) + R$$

The next hypothesis (the effects of practice) is that the difference between the expected typing performance for two students at the same age p but who have practiced different amounts r and s is squal to zero. This must be true for all ages.

Then we consider the two expected values $F(rp)$ and $F(sp)$. The hypotheses is

$$E(rp) - F(sp) = 0$$

Now in the above model we find that

$$E(rp) = a_r + b_p$$

$$F(sp) = a_s + b_p$$

Then we see that the hypothesis is

$$(a_r + b_p) - (a_s + b_p) = 0$$

$$a_r - a_s = 0 \qquad r \neq s$$

$$r,s = 16,17,\ldots,30$$

Then we impose these 14 restrictions on the model; the new restricted model can be written as

(3) $Y = b_{14}\ Z(14) + b_{15}\ Z(15) + \ldots + b_{18}\ Z(18) + G$

Notice that this restricted model has no information to distinguish amounts of practice; i.e., we have given up the information about differences in amounts of practice.

Let $q_3 = \Sigma G_i^{\,2}$, the sum of squares of the error in the new restricted model.

Then if desired we have

$$F = \frac{(q_3 - q_2) / (19-5)}{q_2 / (n-19)}$$

Approach 2 - (COVARIA U.)

Since the research worker from COVARIA U. likes to do covariance analysis it is necessary to have his "contaminating" or covariable in "continuous" form. Therefore, it is necessary to accept a certain hypothesis about the model used in the ANOVA approach (model 1) above. Before beginning his analysis this research worker must make the following assumptions:

$$
\begin{aligned}
a_{16,14} &= c_{16} + d_{16} * 14 \\
a_{16,15} &= c_{16} + d_{16} * 15 \\
&\vdots \\
a_{16,18} &= c_{16} + d_{16} * 18 \\
a_{17,14} &= c_{17} + d_{17} * 14 \\
a_{17,15} &= c_{17} + d_{17} * 15 \\
&\vdots \\
a_{17,18} &= c_{17} + d_{17} * 18 \\
&\vdots \\
a_{30,18} &= c_{30} + d_{30} * 18
\end{aligned}
$$

or

$$a_{ij} = c_i + d_i * j \qquad \begin{aligned} i &= 17,18,\ldots,30 \\ j &= 14,15,\ldots,18 \end{aligned}$$

where

c_i and d_i are unknown parameters to be estimated by the least squares method. These assumptions then lead to the acceptance of the following model:

$$(4) \quad Y = c_{16}\ X(16) + d_{16}\ A(16) + c_{17}\ X(17) + d_{17}\ A(17) + \ldots + c_{30}\ X(30) + d_{30}\ A(30) + E$$

where

$X(i) = 1$ if the corresponding element of Y is from a student who practiced i hours; 0 otherwise

$A(i) =$ the age of the student if the corresponding element of Y is from a student who practiced i hours; 0 otherwise

Now, we emphasize the basic common element between the two approaches. The COVARIA approach is now stated exactly the same as the ANOVA approach, i.e., the hypothesis of no interaction is the difference between the expected (or predicted) typing performance of the two students at age p but with different practice levels r and s is equal to the difference between the expected (or predicted) typing performance of the two students at age q but with different practice levels r and s.

This is hypothesized for all values of p, q, r, and s where $p \neq q$ and $r \neq s$.

Exactly as in approach 1 the hypothesis of no interaction is

$$E(rp) = E(sp) = E(rq) - E(sq).$$

Notice that the two research workers are thinking about the problem in the same way.

Now we proceed to find that the expected values in the COVARIA approach are

$$E(rp) = c_r + d_r * p$$

$$E(sp) = c_s + d_s * p$$

$$E(rq) = c_r + d_r * q$$

$$E(sq) = c_s + d_s * q$$

Then the hypothesis is

$$(c_r + d_r * p) - (c_s + d_s * p) = (c_r + d_r * q) - (c_s + d_s * q)$$

$$(d_r - d_s) * (p - q) = 0$$

But since $p \neq q$

Then it is necessary that

$$d_r - d_s = 0$$

or $$d_r = d_s \qquad \begin{array}{c} r \neq s \\ r,s = (16,17,\ldots,30) \end{array}$$

Then imposing the restrictions on the full model (4) we have:

(5) $Y = c_{16}\ X(16) + c_{17}\ X(17) + \ldots + c_{30}\ X(30) + d_0\ A + R$

where

d_0 = a new unknown parameter which represents the coefficient common to all practice categories.

A = a vector containing the ages associated with the elements in Y.

Let

$$q_1 = \Sigma_i\ E_i^{\ 2}$$
$$q_2 = \Sigma_i\ R_i^{\ 2}$$

Then the F statistic

$$F = \frac{(q_2 - q_1)\ /\ (30 - 16)}{q_1\ /\ (n - 30)}$$

can be computed as a test for the interaction.

As before we consider the case where the research worker accepts the above hypothesis.

The hypotheses of the effects of practices is thought of in the same way as in the previous approach. The difference between the expected typing performance for two students at the same age p but who have practiced different amounts r and s is equal to zero.

Then the hypothesis is written exactly as in Approach 1.

$$E(rp) - E(sp) = 0$$

But in this COVARIA model we have

$$E(rp) = c_r + d_0 P$$

$$E(sp) = c_s + d_0 P$$

Then by substitution the hypothesis is

$$(c_r + d_0 p) - (c_s + d_0 p) = 0$$

or $$c_r = c_s \qquad \begin{matrix} r \neq s \\ r,s = (16,17,\ldots,30) \end{matrix}$$

Then imposing the restrictions on model (5) we have

(6) $$Y = c_0 U + d_0 A + G$$

where U = the unit vector of all 1's

c_0 = an unknown coefficient associated with the unit vector

Notice that in this model all information about practice has been eliminated.

Then

$$q_3 = \Sigma G_i^{\;2}$$

and $$F = \frac{(q_3 - q_2^{'}) \;/\; (16 - 2)}{q_2 \;/\; (n - 16)}$$

Approach 3 - (MULCOR U.)

Now the research worker who was trained at MULCOR U. needs to have all his information in a continuous form since that's what is required in his approach.

Then before the MULCOR man can start he assumes not only the restrictions represented by the COVARIA worker (see model 4) but in addition he must assume in model (4) that the following are true:

$$\begin{aligned}
c_{16} &= t_0 + t_1 * 16\\
d_{16} &= w_0 + w_1 * 16\\
c_{17} &= t_0 + t_1 * 17\\
d_{17} &= w_0 + w_1 * 17\\
&\vdots\\
c_{30} &= t_0 + t_1 * 30\\
d_{30} &= w_0 + w_1 * 30
\end{aligned}$$

or

$$c_i = t_0 + t_1 * i \qquad (i = 16,17,\ldots,30)$$

Imposing these assumptions on model (4) we develop a starting model as follows:

$$\begin{aligned}
Y = {} & (t_0 + t_1 * 16)\ X(16) + (w_0 + w_1 * 16)A(16)\\
& + \ldots + (t_0 + t_1 * 30)\ X(30) + (w_0 + w_1 * 30)A(30) + E\\
Y = {} & t_0[X(16) + \ldots + X(30)]\\
& + t_1\ [16 * X(16) + \ldots + 30 * X(30)]\\
& + w_0\ [A(16) + \ldots + A(30)]\\
& + w_1\ [16 * A(16) + \ldots + 30 * A(30)] + E
\end{aligned}$$

Then

(7) $Y = t_0 U + t_1 P + w_0 A + w_1 (P*A) + E$

Where U = the unit vector of all 1's

P = a vector containing hours of practice

A = a vector containing ages

$P*A$ = a vector whose elements are the product of the corresponding elements of P and A. This is called the direct product of P and A.

Again we emphasize that the MULCOR research worker thinks about the problem in the same manner as the previous two. His hypothesis of no interaction is as before

$$E(rp) - E(sp) = E(rq) - E(sq)$$

Now we determine the expected values in the model assumed by the MULCOR researcher. Looking at model (7) we find

$$E(rp) = t_0 + t_1 * r + w_0 * p + w_1 * (r*p)$$
$$E(sp) = t_0 + t_1 * s + w_0 * p + w_1 * (s*p)$$
$$E(rq) = t_0 + t_1 * r + w_0 * q + w_1 * (r*q)$$
$$E(sq) = t_0 + t_1 * s + w_0 * q + w_1 * (s*q)$$

Then the hypothesis becomes

$$[t_0 + t_1 * r + w_0 * p + w_1 * (r*p)] -$$
$$[t_0 + t_1 * s + w_0 * p + w_1 * (s*p)] =$$
$$[t_0 + t_1 * r + w_0 * q + w_1 * (r*q)] -$$
$$[t_0 + t_1 * s + w_0 * q + w_1 * (s*q)]$$

or $w_1[p - q][r - s] = 0$

Then for this to be always true the hypothesis is

$$w_1 = 0$$

Imposing this restriction on the assumed model (7) we obtain the restricted model (8).

(8) $$Y = t_0 U + t_1 P + w_0 A + R$$

Then we can compute

$$q_1 = \sum_i E_i^2$$

$$q_2 = \sum_i R_i^2$$

and the F statistic is

$$F = \frac{(q_2 - q_1) / (4 - 3)}{q_1 / (n - 4)}$$

Now if model (8) is accepted as true then we proceed to test the effects of practice as in the previous two approaches. The hypothesis is as before

$$E(rp) - E(sp) = 0$$

Now in model (8) we have

$$E(rp) = t_0 + t_1 * r + w_0 * P$$

$$E(sp) = t_0 + t_1 * s + w_0 * p$$

And our hypothesis is

$$(t_0 + t_1 * r + w_0 * p) - (t_0 + t_1 * s + w_0 * p) = 0$$

$$t_1 (r-s) = 0$$

and since $r \neq s$ then the hypothesis is

$$t_1 = 0$$

Imposing this restriction on model (8) we have

(9) $$Y = t_0 U + w_0 A + G$$

Then if

$$q_3 = \sum_i G_i^2$$

we have

$$F = \frac{(q_3 - q_2) / (3-2)}{q_2 / (n-3)}$$

Approach 4 - (VARICO U.)

The research worker for VARICO U. has always preferred to have his data in a different form from the others. He wishes to assume that model (1) of the ANOVA approach has the following restrictions:

$$\begin{array}{l}
a_{16,14} = k_{14} + m_{14} * 16 \\
a_{17,14} = k_{14} + m_{14} * 17 \\
\quad\vdots \qquad\quad \vdots \qquad \vdots \\
a_{30,14} = k_{14} + m_{14} * 30 \\
\quad\vdots \qquad\quad \vdots \qquad \vdots \\
a_{30,18} = k_{18} + m_{18} * 30
\end{array}$$

or

$$a_{ij} = k_j + m_j * i \qquad \begin{array}{l} i = 17,18,\ldots,30 \\ j = 14,15,\ldots,18 \end{array}$$

where k_j and m_j are unknown parameters to be estimated by the least squares method. These assumptions lead to the acceptance of the following model:

$$\begin{array}{ll}
(10) & Y = k_{14}\,Z(14) + m_{14}\,P(14) + k_{15}\,Z(15) + m_{15}\,P(15) \\
 & \quad + \ldots + k_{18}\,Z(18) + m_{18}\,P(18) + E
\end{array}$$

where

$Z(i)$ = 1 if the corresponding element of Y is from a student who is i years of age; 0 otherwise and

$P(i)$ = the hours of practice of the student if the corresponding element of Y is from a student who is i years of age; 0 otherwise.

Again, we emphasize the idea that is common to all four approaches.

The hypothesis of no interaction is still stated as

$$E(rp) - E(sp) = E(rq) - E(sq)$$

Then we obtain these expected values from model (10).

$$E(rp) = k_p + m_p * r$$

$$F(sp) = k_p + m_p * s$$

$$E(rq) = k_q + m_q * r$$

$$E(sq) = k_q + m_q * s$$

Then the hypothesis is

$$(k_p + m_p * r) - (k_p + m_p * s) = (k_q + m_q * r) - (k_q + m_q * s)$$

$$(m_p - m_q) * (r - s) = 0$$

But since $r \neq s$ then the hypothesis must be

$$m_p - m_q = 0 \qquad p \neq q$$

or

$$m_p = m_q \qquad p,q = (14, 15, \ldots, 18)$$

Imposing these restrictions on model (10) we obtain the restricted model

(11) $$Y = k_{14}\, Z(14) + k_{15}\, Z(15) + \ldots + k_{18}\, Z(18) + m_0\, P + R$$

where

m_0 = a new unknown parameter which is common to <u>all</u> ages.

P = a vector containing the practice hours associated with the elements in Y.

Then we can compute

$$q_1 = \Sigma E_i^2$$

$$q_2 = \Sigma R_i^2$$

and

$$F = \frac{(q_2 - q_1) / (10 - 6)}{q_1 / (n - 10)}$$

can be computed to test the hypothesis.

As in the three previous approaches, we next explore the case of no interactions and hypothesize that the difference in typing performance for two students at the same age p but who have practiced different amounts r and s is equal to zero. Again, the hypothesis is the same in all three previous approaches.

$$E(rp) - E(sp) = 0$$

Then we obtain these expected values in our VARICO model.

$$E(rp) = k_p + m_0 r$$

$$E(sp) = k_p + m_0 s$$

Then the hypothesis is

$$(k_p + m_0 r) - (k_p + m_0 s) = 0$$

or

$$m_0 = 0$$

Then imposing this restriction on model (11) we have

$$Y = k_{14} Z(14) + k_{15} Z(15) + \ldots + k_{18} Z(18) + G$$

computing

$$q_3 = \Sigma G_i^2$$

we can determine

$$F = \frac{(q_3 - q_2) / (6 - 5)}{q_2 / (n - 6)}$$

Summary

The ideas that were emphasized above are:

1. In all four approaches the statement of the hypothesis of no interaction was the same in the original thinking about the problem. Not until the specific assumed model was introduced did the approaches appear different.

2. In all four approaches the hypothesis testing the effects

of practice (which followed the acceptance of no interaction) was the same. Not until the specific model was introduced did the approaches appear different.

3. The assumed models in all four approaches were obtained by accepting assumptions about the first approach (ANOVA).

   If desired the research worker from COVARIA, MULCOR, and VARICO could test their assumed models to determine if these starting models are appropriate.

4. Even though the computational aspects were not emphasized, it can be observed that computing procedures required in all four approaches are quite similar.

It is interesting to notice that the original model of approach number one was basic to all others, and that the last three research workers chose to accept assumptions about the first model. Now the predictor vectors in this basic model that was the originator (or parent) of all others are binary coded, mutually exclusive vectors. Sometimes these basic vectors are called dummy vectors. This seems to imply that there is something "not quite right" or "bad" about these vectors. These binary vectors are really the parents of the other vectors and are in effect the most "brilliant" of them all. I would think that they should be called the "bright" vectors, and the other vectors might be called "dummy".

My guess is that since the binary (parent) vectors were recognized much later than their offspring, there was some attempt to apologize for the introduction of the parent.

Also, since many early studies were thought of in a multivariate normal setting, there existed more need for users of these binary vectors to apologize for their use since they were not multivariate normal.

The first three approaches, ANOVA, COVARIA, and MULCOR are frequently treated quite separately in the education of research workers. The fourth approach VARICO is not likely to appear at all.

I urge those teachers who are interested in developing in their students the capability of effective research analysis to consider carefully the objectives presented in the earlier part of this paper on page one. Then I suggest that the specific synthesis of models that has been presented will contribute to the development of the research capabilities that are desired of research workers.

Introduction of Dr. B. J. Winer

by

Joseph Hammock
University of Georgia

It is a pleasure to welcome the next speaker to the University of Georgia and Athens. He really needs no introduction because of his text, Statistical Principles in Experimental Design. It has had phenomenal success. He certainly needs no introduction for students who have been around Purdue and have taken his graduate courses in statistics.

It's rare, and I'm sure you will go along with this, to hear a really excellent statistics teacher compete with the other excellent teachers on the campus. At Purdue, I am told by all the students who have been in and around psychology and statistics, Dr. Winer is the master teacher. I think that's very important.

You may not know that he is a native of Oregon. He earned his Bachelor's and Master's degrees at the University of Oregon. He got all involved with World War II and worked with the Adjutant General's Office of the Army and the Civil Service Commission. He returned to Ohio State and earned his Ph.D. there with Wherry. For a couple of years he was at North Carolina where he worked with Cox and others. He has now been at Purdue University as Professor of Psychology for several years. It is a real pleasure to introduce to you Dr. Ben J. Winer.

PROBLEMS IN THE USE OF GENERAL LINEAR MODEL METHODS

B. J. Winer
Purdue University

The topic that I was originally assigned was the misuse of general linear models, but on the program it got translated "Problems in the Use of the General Linear Model." I think that, historically, the use of the general linear model as we know it today stems from R. A. Fisher. Certainly the general linear model, as it was taught to me by Professor Bose and others at Chapel Hill, very clearly indicated the connection between the experimental use of the general linear model and the ideas of R. A. Fisher in his analysis of variance methodology.

Before I talk about problems in the use of the general linear model, let me quote from R. A. Fisher and essentially indicate what Fisher thought about statistics in general, which is going to be relevant to what I have to say. I am quoting from a paper by Fisher entitled "On the Mathematical Foundaticns of Theoretical Statistics" which appeared in the Philosophical Translations of the Royal Society of London in 1922. Fisher says the following:

> The problems which arise in the reduction of data may be conveniently divided into three types. (1) Problems of specifications. These arise with the choice of the mathematical form of the population. (2) Problems of estimation. These involve the choice of methods of calculating from a sample statistical derivatives, or as we shall call them 'statistics' which are designed to estimate the values of parameters of the hypothetical population. (3) Problems of distribution.

> These include discussions of the distribution of the statistics derived from samples or, in general, any functions of quantities whose distribution is known.

It will be clear that when we know what parameters are required to specify the population from which a sample is drawn, how best to calculate the sample estimates of these parameters, and the exact form of the distribution of our derived statistics in different samples, then the theoretical aspect of the treatment of any particular body of data has been completely elucidated.

I want you to pay particular attention to the next quotation. Fisher was very much concerned with the application of statistics; he was a shrewd man, experimentally. Here is what he had to say about problems of specification. This is essentially the problem of which Dr. Graybill spoke.

> As regards problems of specification, these are entirely a matter for the practical statistician. Those cases where the quantity and nature of the hypothetical population is known do not involve any problem of this type. In other cases, we may know by experience what forms are likely to be suitable, and the adequacy of our choice may be tested a priori or a posterori. We must confine ourselves to those forms which we know how to handle or for which any tables which may be necessary have been constructed.

Let me repeat that sentence for emphasis. I may contradict Fisher a little later in this regard. I think Fisher is just being a little rigid here. "We must confine ourselves to those forms which we know how to handle or for which any tables which may be necessary have been constructed. More or less elaborate forms will be suitable according to the volume of the data." Here Fisher shows some flexibility. "The volume of data in turn," he says here, "determines which forms may or may not be suitable."

Evidently these are considerations the nature of which may change greatly during the work of a single generation. This was written in 1922. This is approximately 1967. I think we are certainly through a single generation and I think some things have changed and we have solved some of the distributional problems. But I am not sure, in my own mind, that the problems of specification are any nearer solution than they were in Fisher's time.

Let me quote again briefly from Fisher. This is a quote from "The Logic of Inductive Inference," which was published in 1935.

> I have called my paper 'The Logic of Inductive Inference.' It might just as well have been called 'A Making Sense of Figures.' For everyone who does habitually attempt the difficult task of making sense of figures is in fact essaying a logical process of the kind we call 'inductive' and, that is, attempting to draw emphasis from the particular to the general, or as we more usually say in statistics, from the sample to the population. Such inferences we recognize are not mathematically rigorous inferences; they are, however, statistically rigorous because they contain within themselves an adequate specification of the nature and extent of the uncertainly involved.

Now Dr. Graybill and Dr. Ward both very deliberately avoided tests of significance. I think they avoided this because it is a difficult problem. Since I have been assigned the topic "problems," I don't think I can avoid the topic although I'm going to try to do so, believe it or not. Let me at the start here again give you Fisher's view of tests of significance. According to Neyman, Fisher never did make a test of significance; of course, Fisher is first to admit it.

Let's see what Fisher has to say. This is a quote from the Design of Experiments.

> The improvement of natural knowledge, that is, learning by experience or by planned chains of experiments, conculsions are always provisional and in the nature of progress reports, interpreting and embodying the evidence so far accrued. Convenient as it is to note that hypothesis is contradicted at some familiar level of significance such as five per cent, or one per cent, we do not in inductive inference ever need to lose sight of the exact strength which the evidence has in fact reached or to ignore the fact that with further trial it might come to be stronger or weaker.

So this is what Fisher has to say about tests in an overall way. Incidently, notice that Fisher said nothing about the word "power" here at all. Neyman is right; Fisher didn't consider power as such, but I'm not sure it is really necessary. You see, I come from a Fisherian background. It is hard to contradict the old master, in a certain sense.

Now, I've come to talk about possible misuse of the linear model. Misuse and problems, I suppose, are synonymous in some sense. What I'd like to talk about first in the way of problems or misuses is the difference between the analysis of planned data and the analysis of unplanned data. I suppose this is the difference between a designed experiment and one which has not been designed. There are many dangers in trying to interpret regression analyses, particularly, to try and get any kind of causal relationship. Dr. Graybill spoke of that. An example that Dr. Box of the University of Wisconsin has used illustrates, I think, the major issues involved. Box distinguishes between what he calls observable variables and latent variables, and very carefully distinguishes between regression analysis associated with a planned experiment versus regression analysis obtained through historical data. Let me give you his example. The example is from the chemical industry, as any example from Box would be. I think the constructs will carry over to any field in which one is asked to interpret data or to make sense of figures, to use Fisher's term. Box cites an example in which the criterion is the production of a final product in a production system. He has a regression of the form $Y = \beta_1X_1 + \beta_2X_2$. The estimated productivity, Y, is a linear function of two variables. He gives this very interesting example in which historically whenever frothing is observed in the chemical process the operator is told to apply pressure to the system. So X_1 is pressure applied to the vat in which the process is going on. Pressure is applied historically whenever there are impurities. This is the operational rule--if frothing occurs, increase pressure.

Other variables in the system are latent variables, latent in the sense that they cannot be measured easily. Through historical data one finds a relationship. Let me change the notation here and replace X_2 by X_k. It can't be measured. There are other variables here such as temperature control and so forth. We cannot observe X_k directly and get a fairly decent prediction of productivity, historically. Notice that there is a set of procedures that define what is done at various stages.

Another question arises here, can one increase productivity in this sort of setup by merely increasing pressure? There is a high correlation between the increase of pressure and the increase of productivity. Historically the correlation is very high; the validity is very high on a historical basis. But this correlation is not of any use at all in building a system or in revising the production system to increase the production. If one increased pressure when the impurities were not present, it would not increase productivity.

I want to make the point from this example that historical correlation may indeed be an accident of procedure. If one had run a controlled experiment, such that one independently manipulated the variables in the regression equation, one could certainly be dealing with uncorrelated variables. Again, the experiment is part of the history.

The important message that I want to give here with respect to the use of a linear model is this: to a large extent correlations are man made. When we conduct an orthogonal experiment as opposed to a non-orthogonal one, almost at will we can increase correlation or reduce it to zero.

The question that arises in any particular application of the linear model is this: to what population are we really trying to generalize? As Box puts it, if you are trying to predict what will happen if you alter variables in your prediction system, then you have to build your regression equation from an experiment in which these variables were altered. You cannot alter at will variables in a regression equation computed from historical data and hope that just by increasing the numerical value of X_1 (if it has a positive regression weight) that production will be increased. It might up to a point, but certainly it would be very limited. That is, the utility of X_1 in the prediction system really depends on the latent variable X_k.

Let me repeat, and this has a great deal to do with disproportionalities in data which is analyzed, correlations to a considerable extent, particularly in design work, depend upon how one designs the experiment and blind use of a regression model or any kind of analysis of variance model can be quite misleading.

I have a certain distaste for applying the principles developed within experimental design to data which were not obtained experimentally. I said I had a distaste for it;

that doesn't mean I don't do it or that I don't advise others to do it. But there are many dangers there, and I think perhaps it's a misuse of design models.

We must not lose sight of the population to which one is trying to generalize; we must not lose sight of what one is trying to predict. Methods suitable for one kind of data may be quite inappropriate for methods of another kind; that is, methods of analysis of data gathered one way may be quite inappropriate for data gathered another way.

Now let me return to Fisher here. In terms of what we call misuse of the linear model, I don't think that applied statisticians are essentially erring in distribution problems. I don't think they are erring in estimation procedures. They know these. Perhaps the error is applying these estimation and distribution principles to a model which is inappropriate. Misuse of the linear model lies primarily in the specification aspect.

Let me go on to another point. Joe Ward spoke at considerable length this morning about the general regression model and the difference in treatment that this model would have depending upon whether you were from ANOV University, Purdue University, University of Georgia, or what.

Let's look at an example in which perhaps the principles about which Ward talked are utilized in perhaps a slightly more effective way.

Table 1

Two-factor Numerical Example of Cell Totals (Cell N's = 5)

		Levels of factor B			
		1	3	5	7
Levels of factor A	0	10	26	58	106
	2	24	52	96	156
	4	38	78	134	206
	6	52	104	172	256

I have a set of data that has been classified into rows and columns, a two-factor problem. We have an A factor at levels 0, 2, 4, and 6. Notice the levels are equally spaced. The B factor is at levels 1, 3, 5, and 7, also equally spaced. Now let's suppose these data are obtained from an experiment. One is trying to evaluate the relative effectiveness of factors A and B, if they were applied to, say, a collection of elements which are untreated. So at the beginning of this experiment there is no population, except conceptually, to which we are trying to generalize. But what we are doing in the experiment is creating from a specified state of elements a new population, a population of which elements have been treated. One can draw inferences with reference to this new population in our experiment. This seems strange to some to talk this way. It seems strange to me even to say this--that in an experiment, we actually create a population. We create a sample from a non-existent population. Well, the elements that we work on are drawn at random early in this case. Now, notice that I have randomization, in this case, on a population of elements and we want to draw inferences about the hypothetical population of treated elements.

Berkson in two or three different papers has distinguished between the experimental creation of a population as opposed to drawing samples from an already existing population. I think the latter is essentially historical research, as opposed to experimental research in the former.

Fisher had a great deal to say about the inadequacies of historical research. I remember back a few years that Fisher came through Indiana and talked about the relationship between smoking and cancer. He quoted all these figures that had been gathered by his esteemed colleagues in England. The data if you merely looked at them, in terms of cell frequencies, were enough to convince almost anyone about the relationship between smoking and cancer. While he was delivering his talk, he was chain smoking. I think the hint there is a good one; he wasn't convinced. The brochure that he put out was rather vehement against historical data as relating to cause and effect in any way. I think that some of the objections that he had in this survey data have since been supplemented by direct experimentation, more direct causal relations.

Well, let's get back to the example. There are a series of numbers in the cells. I have said that there are five observations in each cell. If I had not specified cell sample sizes, it would be almost irrelevant. I would probably analyze these data in exactly the same way even if they were disproportionate.

The population to which I want to generalize is hypothetical, a hypothetical population of treated individuals. I want to know which of the treatments should be used for the basic population of untreated elements with which we started. In this case it makes no sense to assign any kind of differential weighting to any of the cells, even though by accident the cell frequencies might be highly disproportional. Any good experimenter would want equal precision with respect to each of the cells and therfore, would assign equal cell frequency. If he didn't want equal cell precision with respect to each of the cells, he might deliberately assign different numbers of observations to each.

That doesn't mean that in certain aspects of the analysis, we will necessarily have to use certain procedures. Let me emphasize this one point. There are generally several ways in which one can analyze exactly the same data. The different ways are used to look at different aspects of the data.

Let me take the easy way out. There are equal frequencies in each of the 16 cells in the example. I am just going to be concerned for the moment with the between-cell variation. I'm a Fisherian, and being in a Fisherian mood, I'm not going to take on anything that Pearson did first. Let's look at the analysis of variance approach to these 16 cells and remember here that we are interested in making sense of these figures.

Well, let's examine the between-cell variation. The summary is in Table 2. We want a picture to determine whether interaction is present or not.

Table 2

Between Cell Analysis of Variance of the Numerical Example

Source		SS		df
Between Cell		15,020.80		15
A		4,096.00		3
A_{linear}	4,096.00		1	
All other	0.00		2	
B		10,204.80		3
B_{linear}	10,000.00		1	
$B_{quadratic}$	204.80		1	
B_{cubic}	0.00		1	
A x B		720.00		9
$A_{lin.}$ x $B_{lin.}$	720.00		1	
All other	0.00		8	

Now, what happens if we sum over the levels of B and just look at the marginal effects of A? First of all, the between-cell variation is 15,021 with 15 degrees of freedom. Focusing our attention on just the marginal totals, the variation of the main effect due to A is 4,096 units, the main effect of the variations due to B is 10,205 units and the variation due to the interaction is 720 units. I'm treating the data as if they were orthogonal. They are orthogonal as far as this is a designed experiment with equal cell frequencies and there is no problem about orthogonality there. Incidentally, the A-by-B interaction is the culprit there.

Now, let's focus our attention on the linear comparison among the marginal totals of the A--that accounts for all the variations of A. Now focus attention on the marginal totals for B and look at the linear comparison associated with the B's. The linear trend accounts for 10,000 units of the 10,204.80, and the quadratic component accounts for the remaining 204.80 units. All of the interaction is concentrated in a single comparison, namely the linear-by-linear with 720 units of variation. From this table, we have a description of the between-cell variation.

Table 3 shows the residuals. It is the residual table after we use the original table for the main effects. This is essentially that part of the original table which is not predictable from the marginal totals alone, the residual between-cell variation. What we do essentially in the analysis of variance is to successively residualize what we are working with. If things are orthogonal, it is very readily done.

Table 3

Cell Totals Adjusted for Main Effects and Over-all Mean

		Levels of factor B			
		1	3	5	7
	0	27	9	-9	-27
Levels of	2	9	3	-3	- 9
factor A	4	- 9	-3	3	9
	6	-27	-9	9	27

Now, what I'd like to do in the next table (Table 4) is to relate the partition of the between-cell variation in the analysis of variance sense to a correlational approach to exactly the same data.

Table 4

Correlations Obtainable from Analysis of Variance

Source		r^2	r
A_{lin}	$SS_{A_{lin}}/SS_{b.cell}$	.27268	.52218
B_{lin}	$SS_{B_{lin}}/SS_{b.cell}$	.66574	.81593
B_{quad}	$SS_{B_{quad}}/SS_{b.cell}$	.01363	.11672
$A_{lin} \times B_{lin}$	$SS_{A_{lin} \times B_{lin}}/SS_{b.cell}$	.04793	.21896

We have 15,021 units of variation, plus the amount of uncertainty in the cell totals (or the cell means, which is the same thing, essentially, if we have equal cell frequencies). Let's partition all the between cell variation into single contrasts, single contrasts carrying single degrees of freedom. If we take the ratio of the linear component in the main effect of A to the total between-cell variation we get a squared correlation of .273, or a correlation of .522.

What this means in terms of prediction is this: If we paid attention only to the linear trend in A, the correlation between what we could predict and what we observe is .522. If we add to our regression system a second term that corresponds to the linear component of B divided by the sum of squares for between cell, we get, in this case, .666, or a correlation of .82. The main effects are orthogonal in this case, hence the overall predictability is $r^2_{y.12} = .273 + .666 = .938$.

The marginal totals for B are a much better predictor of the cell frequencies than are the marginal totals of A. Now, we add to this the ratio of sum of squares for B quadratic to sum of squares between cell. We have the effect of adding another orthogonal variable to our regression system. The additional predictability is .014. Of course, that is not a product-moment correlation. It is something called a semi-partial correlation, the additional contribution of the quadratic component to a system already containing the linear component of B. Then of course, we can add in the contribution of the linear-by-linear. That has a squared correlation of .048 or a correlation of .22.

In figure 1 I have actually sketched the relatively simple surface represented by the initial data. It is a surface that is linear in one direction and quadratic in the second direction.

If we were to take the usual regression approach, and this is done in Table 5, we would have to compute the correlations between the X's. I want to point out when I am through the fact that, in a certain sense, I can make these correlations anything I want them to be. Here are our variables. X_1 is the level of A and X_2 is the level of B. Now, if you'll notice in the analysis of variance, only the linear component of A has any contribution whatsoever to the 15,021 units of variation between cells. The B linear as well as the B quadratic make contributions. Let me define a dummy variate to be $X_3 = X_2^2$. Since these are quantitative variates (they are usually handled this way), X_3 assumes the values 1, 9, 25, and 49. Now, if you will notice in the analysis of the interaction, there is only one component there--the linear-by-linear--so we can define a variate X_4 equal to the simple product $X_1 \times X_2$. In actually setting up the intercorrelation matrix associated with this sort of analysis, each cell in the matrix would be described by some combination of these. For example, let's consider the 16 observation points, one corresponding to each cell. Associated with

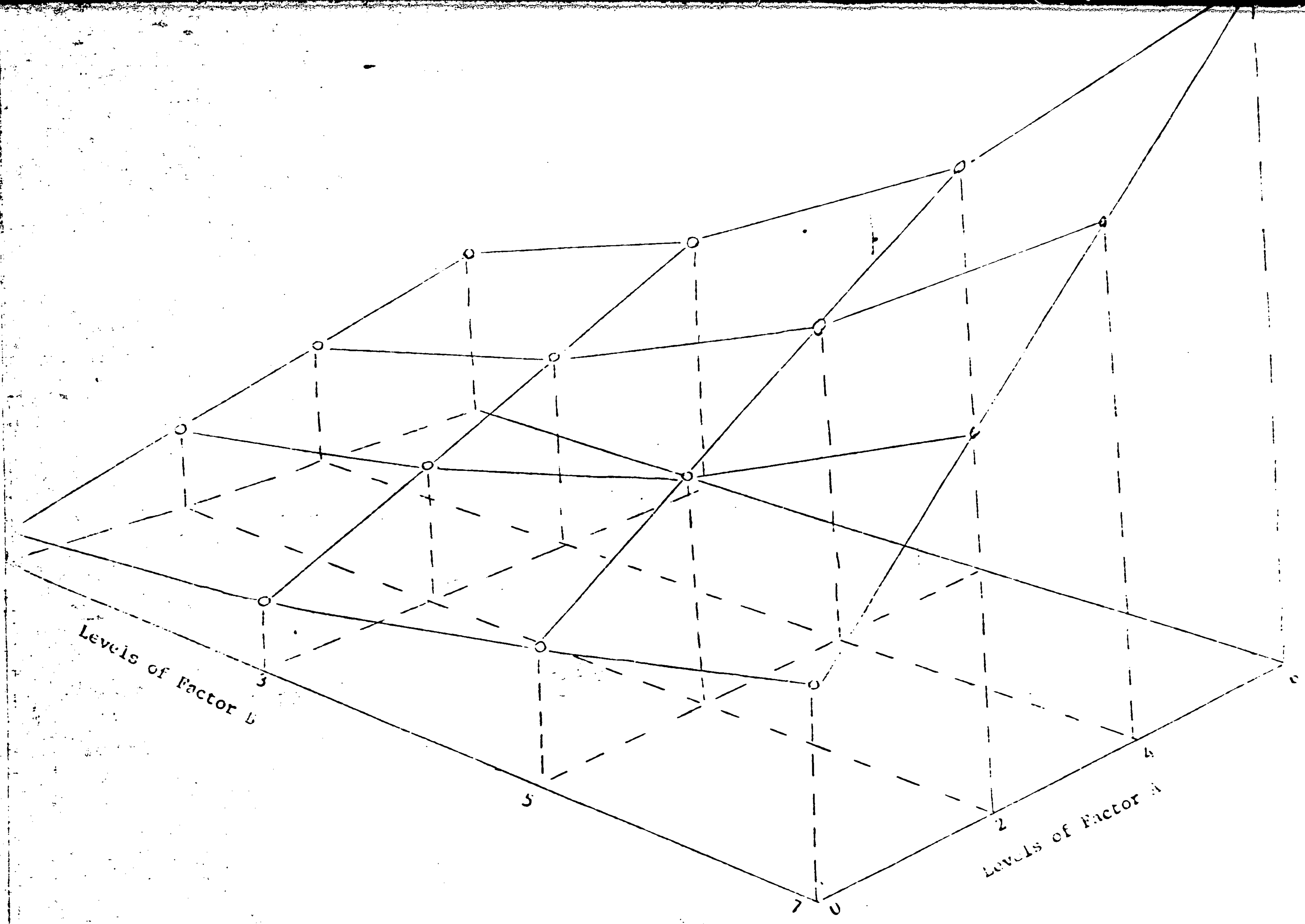


FIGURE 1. Response Surface Corresponding to Cell Totals

Table 5

Intercorrelations of Regression Variables

X_1 = levels of factor A

X_2 = levels of factor B

$X_3 = X_2^2$

$X_4 = X_1X_2$

	X_1	X_2	X_3	X_4	Y
X_1	1.00000	.00000	.00000	.73209	.52218
X_2		1.00000	.97590	.54772	.81593
X_3			1.00000	.53451	.82174
X_4				1.00000	.91764
Y					1.00000

the cell (0, 1), one has $X_1 = 0$, $X_2 = 1$, $X_3 = 1^2 = 1$, $X_4 = 0 \cdot 1 = 0$. The actual observation of Y is 10; so here is an observation vector, $(X_1, X_2, X_3, X_4; Y) = (0, 1, 1, 0; 10)$, and there would be 16 of these. One can compute the intercorrelation matrix and in the traditional sense get the validity coefficients. The latter are indicated in the Y column of Table 5, the validities of X_1, X_2, X_3, and X_4 as predictors of Y.

These kinds of data can be very easily converted to the usual correlational form. This can be taken as a starting point. I don't think they should be. I think that they should be taken as an end point instead of the starting point. Ward and I may have some arguments about this.

There are a whole sequence of regression equations that can be derived whether you start with the entire matrix of intercorrelations and then define hypotheses by restrictions or start with the analysis of variance table and then try to fit an appropriate surface to the data. This may be a matter of taste. In terms of operations, it is simpler to first find the analysis of variance. One should first find what the possible contributors to the prediction system are. The analysis of variance will do this in a well designed experiment, or does this for you orthogonally to begin with. A little later in our prediction system we may introduce correlated variates. Some of these artificial variates that we set up may be indeed correlated.

Now, in Tables 6 and 7 I have a series of regression equations that can be used for prediction. Although the analysis of variance shows, if you look at each of these separately, that you have some variation predicted from each, we are going to find, indeed, that these four predictors are redundant in spite of the fact that in the analysis of variance each of them essentially contributes in terms of non-error variation. In terms of final prediction, to include all four of them will prove redundant. We are going to find that with these particular data there is a linear dependency. The entire augmented matrix including the Y column is a singular matrix, so that from the point of view of efficiency of prediction, all of these are not needed--one drops out.

Table 6

Regression Equations and Correlations

(1)		$Y^* = .52218\ X_1^* + .81593\ X_2^*$
	$r^2_{Y.12} = .93841$	$r_{Y1} = .52218$
		$r_{Y2} = .81593$
(2)		$Y^* = .52218\ X_1^* + .29387\ X_2^* + .53495\ X_3^*$
	$r^2_{Y.123} = .95204$	$r_{Y1} = .52218$
		$r_{Y2} = .81593$
		$r_{Y(3.12)} = .11672$
(3)		$Y^* = .13055\ X_1^* + .52219\ X_2^* + .53629\ X_4^*$
	$r^2_{Y.124} = .98635$	$r_{Y1} = .52218$
		$r_{Y2} = .81593$
		$r_{Y(4.12)} = .21895$
(4)		$Y^* = .13043\ X_1^* + .53501\ X_3^* + .53643\ X_4^*$
	$r^2_{Y.134} = 1.00000$	$r_{Y1} = .52218$
		$r_{Y3} = .82174$
		$r_{Y(4.13)} = .22819$

Table 7

Analysis of Regression

	Source	SS	df	ANOV Source
	X_1, X_2	$r^2_{Y.12}SS_Y = 14,096$	2	$A_{lin} + B_{lin}$
	X_1	$r^2_{Y1}\ SS_Y = 4,096$	1	A_{lin}
(1)	X_2	$r^2_{Y2}\ SS_Y = 10,000$	1	B_{lin}
	X_1, X_2, X_3	$r^2_{Y.123}SS_Y = 14,300$	3	$A_{lin} + B_{lin} + B_{quad}$
	X_1	$r^2_{Y1}\ SS_Y = 4,096$	1	A_{lin}
(2)	X_2	$r^2_{Y2}\ SS_Y = 10,000$	1	B_{lin}
	$X_{3.12}$	$r^2_{Y(3.12)}SS_Y = 205$	1	B_{quad}
	X_1, X_2, X_4	$r^2_{Y.124}SS_Y = 14,816$	3	$A_{lin} + B_{lin} + A_{lin} + B_{lin}$
	X_1	$r^2_{Y1}\ SS_Y = 4,096$	1	A_{lin}
(3)	X_2	$r^2_{Y2}\ SS_Y = 10,000$	1	B_{lin}
	$X_{4.12}$	$r^2_{Y(4.12)}SS_Y = 720$	1	$A_{lin} \times B_{lin}$
	X_1, X_3, X_4	$r^2_{Y.134}SS_Y = 15,021$	3	$A_{lin} + B_{lin} + A_{lin} \times B_{lin}$
	X_1	$r^2_{Y1}\ SS_Y = 4,096$	1	A_{lin}
(4)	X_3	$r^2_{Y3}\ SS_Y = 10,143$	1	$B_{lin} + B_{quad}$
	$X_{4.13}$	$r^2_{Y(4.13)}SS_Y = 782$	1	$+ A_{lin} \times B_{quad}$
	X_1, X_2, X_3, X_4	$r^2_{Y.1234}SS_Y = 15,021$	4	$A_{lin} + B_{lin} + B_{quad} + A_{lin} \times B_{lin}$
	X_1	$r^2_{Y1}\ SS_Y = 4,096$	1	A_{lin}
	X_2	$r^2_{Y2}\ SS_Y = 10,000$	1	B_{lin}
(5)	$X_{3.12}$	$r^2_{Y(3.12)}SS_Y = 205$	1	B_{quad}
	$X_{4.123}$	$r^2_{Y(4.123)}SS_Y = 720$	1	$A_{lin} \times B_{lin}$

Need we strive for complete efficiency? I think there are instances in which we may want to carry redundant variables along. Think about this for a while. Need we eliminate all the redundancy? If you will look in detail at the regression equations that have been computed here and the various correlations that are recorded in Tables 6 and 7, you'll find that one can do a more efficient prediction job in terms of fewer variates. The data will give you a multiple correlation of unity. These are artificial data.

I said nothing about hypothesis testing here, but I may be forced to shortly. This is just a small numerical example that I think indicated an alternative way of combining estimation problems with prediction problems. In this case, notice what I have done, the details are unimportant. I have obtained estimates of various sources of variation and then built a prediction pin-pointed at just these sources. But the prediction system that I would build just by looking at these sources is redundant and we can eliminate one of them if we want to. But from the point of view of predicting, taking statistical pictures on the original table, all four of these variants were relevant. For purposes of predicting the 16 cell totals, if this is all we want, these four variables do not constitute a minimal set. But notice here that in the analysis of variance in the original description of these 16 cell entries I added eight more entries, the row marginals and the column marginals. For the complete description of the 16 cell entries plus the eight additional sums, one needed the entire set of four variants. Well, the reason for that is that certain of these cross-product terms become constant when you sum down the columns and that is what makes one of these variables drop out.

Dr. Graybill mentioned problems associated with errors in observation. This is a very serious problem in the area of behavioral sciences. The observations that we make are considerably less than perfectly reliable. Repetition of the same instrument to the same individual does not give one, in most cases, a distribution of error which is insignificant compared to the magnitude of what we are trying to estimate.

There is a description of how one can handle in part this problem in Graybill's book and there is a somewhat more extensive treatment of this problem in the literature. This is the problem, I think, education and psychology people have called the combination of reliability, that is, the error of measurement, and experimental error. They are quite distinct conceptionally and the problems of handling them are also distinct.

Madansky had a fundamental article a few years ago in the Journal of the American Statistical Association on various ways of combining the usual regression approach with problems of reliability. More recently, in connection with design of experiments, Box and some of Kempthorne's students have tackled the following problems. (This is very closely related to the problem in which we in psychology and education have been concerned for a number of years). It comes up under a new name. Box as well as Ziskind and Kempthorne call this the problem of errors in the levels of factors. That is, let's suppose that in the experiment we have levels a_1 and a_2 of some factor. Upon replication of this experiment, levels a_1 and a_2 aren't quite the same sort of thing. That is, a_1 and a_2 involve some instrument setting, a specified dosage which can only be measured within X units of accuracy. In a certain sense, upon replicating a_1 and a_2, we really have as a_1 the first attempt at replicating a_{11}, a_{12}, and a_{13}. They're all supposed to be a_1, but they are slightly different.

Now most of you in psychometrics will find after some study that this is almost completely equivalent to the testing problem. You want to measure a trait A and a trait B; a_{11}, a_{12}, and a_{13} are simply items in trait A. They are all different; they are all supposed to be pin-pointed at a_1, but they are not exactly a_1. They differ from it by the fact that a_1 can have three different aspects, or it can be measured in three different ways, or there are three different items, each pertaining to a_1. So all the factors associated with trait A essentially enter in the measurement of the level of the factor and we see that the problem is directly relevant to the work in psychometrics. It doesn't appear to be on the surface, but it is. I think it is due time that the experimental-design oriented people realize that this type of error is worthy of considerable study. We have it particularly in trying to replicate control conditions in which, in behavioral sciences, a_1 and a_2 are groups, or correspond to high-versus-low in terms of a factor such as "anxiety". Let me conclude this point by saying this is relevent. It is one of the problems that we face in trying to experiment, in using the ordinary linear model, when the levels of the factors themselves cannot be reproduced exactly from one replication to the next.

Incidentally, this book that I wrote has been a problem for me, but it has been the source of my getting all sorts of letters. Some of these letters are very strange sorts of things. It's really amazing the things that can be put in print and can be misinterpreted the way they are. One of the central topics, one of the repeating themes of the types of letters I get, concerns how one estimates variance components. One measures these variance components to evaluate what Fisher calls the strength of relationship. Essentially, a correlation is nothing more than a ratio of two components.

If something is statistically significant, what is the strength of association? What is the measure of association? What the variance components are depends very, very much on what the proper specification of the model is. In practice, I don't hesitate to use several different estimation procedures. I don't hesitate to get estimations based upon the assumption that certain factors are fixed. I don't hesitate to get another estimate under the assumption that certain factors are random. I take solace again, being a Fisherian, in the statement that you'll find a Fisher somewhere. I found it somewhere, but I couldn't find it when I was looking for it the other day. The fact that Fisher says, and this is sloppy, sloppy from some points of view, that any data no matter how it's gathered is a sample from some population.

Sampling theory people tell you that the population comes first, then the sample. There is no such thing in a certain sense. Some data is better than none. This does, in part, justify one's intuitive attempt to make the most sense possible out of any set of data. I don't think there is such a thing as an appropriate model; "appropriate for what" would be a better designation. One model may be more appropriate than another for interpreting given data one way.

As I say, the problem of estimating variance components does not have a single answer. It does have a unique answer only if you have complete specification of the initial model.

One of the purposes of experimentation is to specify sequentially a model. I think we lack, in our general experimental approach, the facility to tackle data, especially in education and psychology, sequentially. One of the advantages of this type of approach is that one can sequentially specify a model. Models are guides, not bins.

Now, one of the problems, and my book is just full of this kind of problem, is that in which the same experimental unit is used under a variety of treatment conditions. In five cents jargon, it is a repeated measures problem if we are dealing with individual subjects. To what extent does the fact that we've used the same experimental unit, the same subject, under a sequence of treatment conditions affect various models that we might present?

Let me give you an example. First of all, let me indicate why under certain assumptions, it is a multivariate problem. Here you have an individual--individual one, individual two, individual three, etc. You make a series of observations on this individual under several treatments. These may be over different time periods under the same treatment, it makes no difference. And what you have here essentially is an observation vector having several components. You have a series of observation vectors. There is correlation between, say, between period one and period two, or treatment one and treatment two.

In the traditional agricultural setting, this may be viewed as a split-plot design of some form. This is the jargon in the agricultural field. The whole plot is the individual; split-plots are observations within the individual.

Call π the individual difference component associated with the individual subject. As long as the π's remain constant, then it can be shown that the variance-covariance matrix associated with these data must have a very special pattern. So if π is a constant, if π does not change under these treatments, you get a very highly patterned covariance matrix which has just two distinct roots. The usual approaches, the analysis of variance approach and the multivariate analysis approach, will give you exactly the same answer. It doesn't look like it, but it is true. You are really doing a very special case of a multivariate analysis, but it looks like a univariate analysis, because of the fact that in the model we can assume that π does not change. In other words, that individual number one is exactly the same (after the treatment effect wears off) when you look at him under each of the other treatments.

But, of course, if π is not a constant, then this very highly structured variance-covariance matrix takes on one of these complex forms that Bargmann talks about.

Note the use of the same subject under repeated conditions. This is the central problem about which I get many letters. Suppose I have a repeated measures design and I don't get anything near the variance-covariance matrix I should under the assumption that π remains constant. What can I do about it? Well, let me reverse this problem. I think that it is more usual that an individual who uses this method to begin with probably knows what he can do about it, but may not want to do it. Many ask this question: How much am I wrong by assuming π to be a constant when it really is not? How much violation is created by approximating one covariance pattern by another? I don't think anyone knows the answer to that yet. On the other hand, again this is a computer age and we have Bock in our audience. Bock has a whole series of programs for which you just press a button and you'll get all sorts of multivariate analyses of variance for your data. I am not sure that this kind of button-pressing is the solution.

I think we will increase in our facility in the interpretation of multivariate analysis of variance once we gain experience with multivariate data and essentially its generalization to canonical analysis. That is, so far, we are working here with only one dependent variable.

Interpretations of the outcomes of multivariate analysis of variance is difficult. I think they are difficult mainly because we have had no experience with them. I think this is a coming problem in the field.

The computer revolution solves some problems, but I think creates as many as it solves. This is one instance I think where our cortexes have not quite kept up with the computer facility for doing operations. Before computers were available to us, we thought our problems would be solved once we could do the operations. Now that we can, the interpretation is quite difficult, and this leads me to the concluding section of what I want to talk about.

That is problems of interpretation, particularly in the analysis of variance setting, and particularly with respect to hypothesis testing. As I say, since I was assigned the topic of problems, this is a problem area and it's been complicated in recent years. I know that when I was at Chapel Hill in my student days we made an F-test, we made another F-test, we made another F-test and at that time the systematic use of various kinds of error rates was not quite in vogue. Within the last ten years, there has been a series of developments, among others, those of Scheffe.

We are, in our hypothesis testing work, plagued with this problem. This is another one about which I receive letters all the time. Which of several possible error rates in hypothesis testing should be used? Should it be an individual comparison error rate, a per-experiment error rate, an experiment-wise error rate--and I dare say, that there are all kinds of additional error rates that are in the process of being published right now.

This is, I think, an important theoretical question. Should the unit with which we are concerned for error rate be just a subcollection of the entire experiment or the entire experiment? Should one be concerned with one row and test hypotheses with an error rate which has as its base unit all possible comparisons within any one row, or within any one column, or within any two rows? Suppose I have four rows in the table. One row might be an early phase, or another row might be a late phase, or there just might be only two rows. Well, you might just say that Winer produced a half experimental unit. Does it make logical sense to describe the data in terms of a unit of error rate which considers all possible differences or all possible generalized contrasts? Are you going to work with a Scheffe level of significance or a kind of Tukey level?

This is a problem that I think has no answer. No answer in purely mathematical terms. I dare say, though, in practice it has an answer. Remember that the purpose of our statistics is to make sense of data. If it makes sense to set an error rate cn just a portion of the data, to focus on this portion, and to set a separate error rate on a different portion of data, by all means, do so.

If you read Tukey carefully, you will find that he is, ultimately, very much in the same spirit as Fisher. As a matter of fact, mathematically Tukey is much more sloppy than Fisher ever dreamed of being. Yet, I think that Tukey is probably the outstanding mathematical statistician in this country today. If you'll read Tukey very carefully, you will find that he does things that I would never advise my students to do. Except, that he indicates very carefully just in what kind of sense this particular procedure applies.

There are many problems in applying the linear model. Having computer programs available to us, I think, simplifies the computational task. In a certain sense it makes much more difficult the specification task, the task of the individual experimenter. This, perhaps, can be computerized, but it hasn't yet. But, perhaps, if one could write a program in which all the relevant inputs could be coded and all the possible utility functions associated with each input specified, then I think the computer can give you alternative specifications. We are not yet quite at this stage, so that I think that the problem of specification and the problem of interpretation are linked.

As Fisher noted we are limited in our choice of models. But don't hesitate to use the existing models even though none is ideal for your purpose. Instead of throwing up your hands in utter despair, the chances are any tractable model would be an over-simplification of real data in education and psychology. But remember, it is only a first approximation, and a series of approximations gets us much closer to the information we want than no activity at all.

Introduction of Dr. Rolf E. Bargmann

by

Harry E. Anderson, Jr.
University of Georgia

I have the pleasure of introducing Dr. Rolf Bargmann of the University of Georgia's Department of Mathematical Statistics. He was formerly manager of the information sciences in IBM research; Professor of Statistics at Virginia Polytech Institute; Head of the Statistics Department, Institute for International Education Research at Frankfurt, Germany. He did his undergraduate work in Chemistry at the University of Berlin and did graduate study in physical chemistry at the University of Hamburg. He did psychometric work with Thurstone in Chicago and took his Ph.D. in mathematical statistics at the University of North Carolina. He is a fellow in the American Association for the Advancement of Science, member of the American Statistical Association, Econometric Society, Psychometric Society, and others. It is a great pleasure to introduce Dr. Rolf Bargmann.

A SURVEY OF APPROPRIATE METHODS OF ANALYSIS OF FACTORIAL DESIGNS

Rolf E. Bargmann
University of Georgia

The model for a two way or two factor design

$$E(y_{ijk}) = \mu + \alpha_i + \beta_j + \delta_{ij}$$

$$i = 1,2,\ldots r;\ j = 1,2,\ldots s$$

plays a key role in a majority of experimental situations, both as a self-contained model, and as an easily interpretable submodel of a more complicated design. Also, in most industrial and educational applications, the number of observations in each cell can be quite irregular, often leaving whole cells empty.

It must be understood, first, that the above model is insufficiently specified. It is easily shown that, with this generality, comparisons or contrasts in α-effects or β-effects are non-estimable. This is probably best illustrated by a perfectly legitimate assumption that all α_i and all β_j are zero. Under this assumption (a condition on the model, not a constraint on the estimates) we would have a one-way classification model with r x s groups. This model specification is obviously not what we wanted, else we would not have postulated a two factor model in the first place.

In many industrial applications, the lowest level of each factor designates its absence, hence an absent factor could not "interact" with another one. The specification to (1) would thus be

$$\delta_{1j} = 0 \text{ for all } j \text{ and } \delta_{i1} = 0 \text{ for all } i.$$

This leads to an extremely simple estimation procedure for, if Y_{ij} denotes the cell total in cell (i,j) and n_{ij} the number of observations in that cell,

$$(\hat{\alpha}_i - \hat{\alpha}_k) = Y_{i1}/n_{i1} - Y_{k1}/n_{k1}$$

$$(\hat{\beta}_j - \hat{\beta}_\ell) = Y_{1j}/n_{1j} - Y_{1\ell}/n_{1\ell}$$

$$\hat{\delta}_{ij} = Y_{ij}/n_{ij} - \hat{\alpha}_i - \hat{\beta}_j - Y_{11}/n_{11}$$

with obvious generalizations to more than two factors. Thus, row effects are estimated only from data in the first column, and column effects only from data in the first row, while the remaining information is used to estimate interaction effects.

There has been searching, and entirely unjustified, criticism of this simplest of all analyses. Why, its opponents agree, should I estimate the main effects--in which I am vitally interested--from so few observations only? The fact of the matter is, of course, that this model brings us back to the classical assumptions of experimental planning: If factors can, potentially, interact with other factors, one should attempt to study each factor in the absence of others in order to ascertain its own, isolated, effect. In educational research, alas, this principle is seldom applicable. We cannot plan an experiment with "absence" of intelligence or "absence" of color, to name a few absurd examples.

One set of specifications for the description of the model has attained rather widespread acceptance. It is usually stated as follows.

$$\sum_j n_{ij} \quad \delta_{ij} = 0 \text{ for all } i$$

and

$$\sum_i n_{ij} \quad \delta_{ij} = 0 \text{ for all } j$$

The name "natural constraints" given to these r + s - 1 conditions is poorly chosen. These specifications are neither natural, nor are they constraints (the latter must be arbitrary specifications of estimates). It is, in fact, a set of conditions which maximizes the non-centrality parameter in the joint chi-square statistic used for testing the hypothesis of equality of all main effects, i.e.,(Sum of Squares for Columns, unadjusted + Sum of Squares for Rows, adjusted)/ σ^2, for fixed σ^2. The classical formulas, involving adjusted normal equations, for the irregular design without interaction, provide the estimation procedure in this case.

When the interaction effect becomes significantly large in this instance, the additive model is a poor representation of the data. Outliers in individual cells may, of course, produce such significant interaction effects, and can easily be detected. If outliers do not explain the departure from additivity, two modifications of the model are available.

(1) The interactions affect the cell means in an irregular fashion; one assumes that they will show different patterns in repeated experiments. The proper model would then be a "Random Interaction Model".

(2) The interaction effects show some definite trend from level to level of the main effects. Here it is necessary to make an assumption regarding the mechanism (specifying the trend except for proportionality constants). The residual portion of the interaction can again be regarded as fixed effects or random components. These models will be called Covariance Models.

The Random Model

$$y_{ijk} = \mu + \alpha_i + \beta_j + d_{ij} + e_{ijk}$$

$$E(d_{ij}) = 0; \quad E(e_{ijk}) = 0$$

$\text{var}\ (d_{ij}) = \sigma_d^2$, $\text{var}\ (e_{ijk}) = \sigma_e^2$, all covariances zero.

The analysis is quite similar to the irregular fixed effect analysis, with the incidence matrix replaced by a matrix of weights

$$w_{ij} = 1/(\rho^2 + 1/n_{ij})$$

where $\rho^2 = \sigma_d^2 / \sigma_e^2$. Unbiased estimates and approximate confidence bounds for ρ^2 are available from the fixed model analysis. They utilize the F-statistic for interaction of the fixed model. The coefficients are somewhat involved, in the irregular design, but are easily programmed. Confidence bounds are based upon the improved variance-stabilizing ($\cosh^{-1}$) transformation of non-central F.

It may be noted that the weights become quite similar to each other if ρ^2 is large and/or if the smallest non-empty cell is appreciably large. In this case then, the "unweighted means" analysis (regular if all cells are occupied, the usual irregular "treatment-block" analysis if some cells are empty) is the limiting form of the Random Model analysis.

The Covariance Models

(a) Fixed Residuals:

$$y_{ijk} = \mu + \alpha_i + \beta_j + f(i,j|\gamma_1,\gamma_2\ldots\gamma_m) + e_{ijk}$$

$$E(e_{ijk}) = 0 \quad \text{var}\ (e_{ijk}) = \sigma_e^2 ,$$

where $f(i, j|\gamma_1,\gamma_2 \ldots \gamma_m)$ is a non-linear function of concentrations of α- levels i and concentrations of β-levels j. A simple form (neutralization) would be $f = -\gamma c_i c_j$, where c are concentrations. The analysis proceeds like an irregular design analysis with one or more covariates. The latter represent the function or functions of the levels which are assumed to explain the mechanism of the interaction effects. A linear combination of i and j (e.g., $\gamma_1 c_i^2 + \gamma_2 c_j$) is ruled out, since these covariates would affect (and be assumed to affect) main effects.

(b) Random Residuals:

$$y_{1jk} = \mu + \alpha_i + \beta_j + d_{ij} + e_{ijk}, \text{ where}$$

$$E(e_{ijk}) = 0, \text{ var } (e_{ijk}) = \sigma_e^2, \; E(d_{ij}) = f(i,j|\gamma_1 \ldots \gamma_m)$$

$$\text{var } (d_{ij}) = \sigma_d^2, \text{ all covariances zero,and } \rho^2 = \sigma_d^2/\sigma_e^2.$$

This is a weighted irregular covariance analysis, with weights equal to those in the random model without covariates. The estimates and confidence interval estimates for ρ^2 are based upon the F for interaction in the fixed residual covariance model. The coefficients are somewhat more involved than those in the random model without interaction.

Demonstration Studies:*

Two computer programs have been written which serve as tools to determine whether some of the two-way classification models considered above provide the best estimation of main effects when the interaction effect possesses certain characteristics. For instance, when the interaction is significant and possesses a random character, or when the interaction is significant and is directed or biased in a definite manner, it would be expected that one of the methods of analysis should be superior. One program

*From F.C.Clark, "The Role of Interaction in Two-Way Classification Models", unpublished Ph.D. dissertation, University of Georgia, 1967.

performs the complete classical maximum main effect analysis, the Random model analysis using the unbiased estimate of the variance component ratio ρ^2, as well as the lower and upper $1-\alpha$ confidence points for ρ^2, and finally an unweighted mean analysis. The other program performs the so-called Covariance Model analysis. Special variants of these programs make use of a data-generating subroutine which utilizes a random normal number generating program. Values for the main effects, interaction term, error term, and entries of the incidence matrix are fed in as input information for the data-generating subroutine.

The studies which make use of the first program were designed in such a way that the interaction is significant under the maximum main effect model and such that the interaction effect possesses a random character.

In the ensuing four studies we will make use of the following symbols for purposes of convenience: "Av" will denote the average value of the estimate over 10 data sets. "T" will denote the true value of the effect. "S.D." will denote the standard deviation of the estimate. "R.M.S." will denote the root mean square of the estimate. "Fixed" will denote the estimate assuming the fixed, maximum main effect model. "Unbiased" will denote the estimate under the Random Model using the unbiased estimate of ρ^2. "Lower" will denote the estimate assuming the Random Model but using the lower $1-\alpha$ confidence point of ρ^2. "Upper" will denote the estimate assuming the

Random Model but using the upper 1-α confidence point of ρ^2. "Unweighted' will denote the estimate assuming the unweighted mean analysis.

For the first study we have the following details:

Factors: A, B

Levels : A @ 5 levels, B @ 10 levels

Incidence Matrix

3	28	3	4	39	43	42	25	36	23	246
20	41	38	36	37	22	28	35	26	36	319
0	25	3	37	6	7	42	14	39	32	205
43	39	38	38	25	9	0	25	36	2	255
37	39	42	0	38	39	1	37	25	26	284
103	172	124	115	145	120	113	136	162	119	1309

An interaction term of one multiplied by a random normal (0,1) number is added to each cell. This gives d_{ij} a random effect. An error term of one times a random normal (0,1) number is added to each cell. Thus, $\rho^2 = 1$. The mean was chosen to be 20. The main effects which were fed into the program are: $\alpha_1 = -2.0$, $\alpha_2 = -1.0$, $\alpha_3 = 0.0$, $\alpha_4 = 1.0$, $\alpha_5 = 2.0$, $\beta_1 = -2.5$, $\beta_2 = -2.0$, $\beta_3 = -1.5$, $\beta_4 = -1.0$, $\beta_5 = -0.5$, $\beta_6 = 0.0$, $\beta_7 = 0.5$, $\beta_8 = 1.0$, $\beta_9 = 1.5$, $\beta_{10} = 2.0$.

Only scme of the parameters are shown in the tables. The others exhibit similar properties. "B" denotes the best value and "W" denotes the worst value.

Main effect estimates

α_1	T	Av	S.D.	R.M.S
Fixed	-2.000	-1.889	.3427	.3600B
Unbiased	-2.000	-1.885	.3549	.3728
Lower	-2.000	-1.886	.3603	.3779W
Upper	-2.000	-1.884	.3531	.3716
Unweighted	-2.000	-1.878	.3415	.3629

α_3	T	Av	S.D.	R.M.S.
Fixed	0.0	-.0647	.2959	.3029W
Unbiased	0.0	-.0278	.2362	.2378
Lower	0.0	-.0272	.2374	.2389
Upper	0.0	-.0284	.2249	.2267B
Unweighted	0.0	-.0321	.2274	.2297

α_5	T	Av	S.D.	R.M.S
Fixed	2.000	1.968	.3894	.3907
Unbiased	2.000	1.979	.3303	.3309
Lower	2.000	1.979	.3348	.3354W
Upper	2.000	1.977	.3264	.3272
Unweighted	2.000	1.974	.2898	.2909B

β_1	T	Av	S.D.	R.M.S.
Fixed	-2.324	-2.281	.3518	.3544B
Unbiased	-2.324	-2.222	.3819	.3953
Lower	-2.324	-2.229	.3758	.3879
Upper	-2.324	-2.219	.3949	.4084
Unweighted	-2.324	-2.195	.4263	.4453W

β_4	T	Av	S.D.	R.M.S
Fixed	-0.824	-.8579	.4039	.4053
Unbiased	-0.824	-.7218	.4547	.4660W
Lower	-0.824	-.8049	.3975	.3979B
Upper	-0.824	-.8047	.3991	.3995
Unweighted	-0.824	-.8112	.4097	.4099

β_7	T	Av	S.D.	R.M.S.
Fixed	0.676	.7011	.6230	.6235W
Unbiased	0.676	.6701	.5385	.5385
Lower	0.676	.6696	.4989	.4989
Upper	0.676	.6703	.5302	.5302
Unweighted	0.676	.6608	.4564	.4567B

β_{10}	T	Av	S.D.	R.M.S.
Fixed	2.676	2.884	.5513	.5891B
Unbiased	2.676	2.983	.5493	.6292
Lower	2.676	2.979	.5502	.6283
Upper	2.676	2.997	.5491	.6359W
Unweighted	2.676	2.994	.5440	.6299

In the second study we have the following information:

Factors: A, B.

Levels : A @ 5 levels, B @ 10 levels.

Incidence Matrix

3	28	3	4	39	43	42	25	36	23	246
20	41	38	36	37	22	28	35	26	36	319
0	25	3	37	6	7	42	14	39	32	205
43	39	38	38	25	9	0	27	36	2	255
37	39	42	0	38	39	1	37	25	26	284
103	172	124	115	145	120	113	136	162	119	1309

An interaction term of 4 multiplied by a random normal (0,1) number is added to each cell. An error term of 8 multiplied by a random normal (0,1) number is added to each cell. Thus $\rho^2 = \frac{1}{4}$. The mean was chosen to be 20.

The main effects which were fed into the program are: $\alpha_1 = -2.0$, $\alpha_2 = -1.0$, $\alpha_3 = 0.0$, $\alpha_4 = 1.0$, $\alpha_5 = 2.0$, $\beta_1 = -2.0$, $\beta_2 = -1.5$, $\beta_3 = -1.0$, $\beta_4 = -0.5$, $\beta_5 = 0.0$, $\beta_6 = 0.5$, $\beta_7 = 1.0$, $\beta_8 = 1.5$, $\beta_9 = 2.0$, $\beta_{10} = 2.5$. This study was taken over 10 data sets. Under the maximum main effect analysis, F for interaction versus error is 6.5 (Average). Best values are denoted by "B" and worst values are denoted by "W".

Main effect estimates

α_1	T	Av	S.D.	R.M.S.
Fixed	-2.000	-1.540	1.526	1.593
Unbiased	-2.000	-1.609	1.449	1.501
Lower	-2.000	-1.534	1.492	1.563
Upper	-2.000	-1.509	1.523	1.599W
Unweighted	-2.000	-1.442	1.363	1.473B

α_3	T	Av	S.D.	R.M.S.
Fixed	0.0	-.1454	1.250	1.259W
Unbiased	0.0	-.2513	1.113	1.141B
Lower	0.0	-.2361	1.133	1.157
Upper	0.0	-.2502	1.119	1.146
Unweighted	0.0	-.3495	1.175	1.226

α_5	T	Av	S.D.	R.M.S.
Fixed	2.000	2.065	1.534	1.536
Unbiased	2.000	1.963	1.508	1.508
Lower	2.000	1.960	1.539	1.540W
Upper	2.000	1.965	1.389	1.389
Unweighted	2.000	1.993	1.183	1.183B

β_1	T	Av	S.D.	R.M.S.
Fixed	-2.324	-2.159	1.493	1.502
Unbiased	-2.324	-2.177	1.482	1.489
Lower	-2.324	-2.213	1.455	1.459B
Upper	-2.324	-2.135	1.529	1.541
Unweighted	-2.324	-1.888	1.898	1.948W

β_4	T	Av	S.D.	R.M.S.
Fixed	-0.824	-.9259	1.714	1.717
Unbiased	-0.824	-.7988	1.673	1.673
Lower	-0.824	-.8101	1.666	1.666B
Upper	-0.824	-.7899	1.683	1.683
Unweighted	-0.824	-.7748	1.795	1.796W

β_7	T	Av	S.D.	R.M.S.
Fixed	0.676	.7233	2.463	2.464
Unbiased	0.676	.6175	2.444	2.445
Lower	0.676	.6089	2.477	2.477W
Upper	0.676	.6271	2.409	2.409
Unweighted	0.676	.7776	1.961	1.963B

β_{10}	T	Av	S.D.	R.M.S.
Fixed	2.676	3.494	2.344	2.482B
Unbiased	2.676	3.765	2.441	2.673
Lower	2.676	3.733	2.385	2.609
Upper	2.676	3.813	2.548	2.789
Unweighted	2.676	3.973	2.495	2.812W

In the third study we have the following information:

Factors: A, B.

Levels : A @ 3 levels, B @ 5 levels.

Incidence Matrix

0	9	3	1	4	17
3	1	0	5	6	15
9	0	6	3	0	18
12	10	9	9	10	50

An interaction term of $\sqrt{0.5}$ multiplied by a random normal (0,1) number is added to each cell. An error term of $\sqrt{1.5}$ multiplied by a random normal (0,1) number is added to each cell. Thus $\rho^2 = 1/3$. The mean was chosen to be 20. The main effects which were fed into the program are : $\alpha_1 = -1.0$, $\alpha_2 = 0.0$, $\alpha_3 = 1.0$, $\beta_1 = -2.0$, $\beta_2 = -1.0$, $\beta_3 = 0.0$, $\beta_4 = 1.0$, $\beta_5 = 2.0$. The study was taken over 10 data sets. Under the maximum main effect analysis the F for interaction versus error was equal 2.12. "B" means best estimate and "W" means worst estimate.

Main effect estimates

α_1	T	Av	S.D.	R.M.S
Fixed	-1.000	-.9630	.3946	.3963
Unbiased	-1.000	-.9440	.4921	.4952W
Lower	-1.000	-.9521	.3899	.3928
Upper	-1.000	-.9399	.3445	.3497B
Unweighted	-1.000	-.9328	.3540	.3603

α_3	T	Av	S.D.	R.M.S.
Fixed	1.0	.9192	.4256	.4332B
Unbiased	1.0	.8991	.4359	.4474
Lower	1.0	.9128	.4282	.4369
Upper	1.0	.9391	.5024	.5061W
Unweighted	1.0	.8674	.4562	.4751

β_1	T	Av	S.D.	R.M.S.
Fixed	-1.920	-1.687	.5788	.6238W
Unbiased	-1.920	-1.843	.6084	.6131
Lower	-1.920	-1.732	.5928	.6219
Upper	-1.920	-1.925	.5989	.5989
Unweighted	-1.920	-1.922	.5925	.5925B

β_3	T	Av	S.D.	R.M.S
Fixed	0.08	.2393	.5489	.5715W
Unbiased	0.08	.2164	.5081	.5260
Lower	0.08	.2507	.5139	.5415
Upper	0.08	.1710	.4995	.5077B
Unweighted	0.08	.1910	.5373	.5488

β_5	T	Av	S.D.	R.M.S
Fixed	2.080	2.232	.4962	.5188B
Unbiased	2.080	2.037	.5636	.5652
Lower	2.080	2.188	.7668	.7743W
Upper	2.080	1.969	.5799	.5902
Unweighted	2.080	1.976	.6138	.6226

The fourth study contains the following results:

Factors: A, B.

Levels : A @ 5 levels, B @ 10 levels.

Incidence Matrix

3	28	3	4	39	43	42	25	36	23	246
20	41	38	36	37	22	28	35	26	36	319
0	25	3	37	6	7	42	14	39	32	205
43	39	38	38	25	9	0	27	36	2	255
37	39	42	0	38	39	1	37	25	26	284
103	172	124	115	145	120	113	136	162	119	1309

An interaction term of $\sqrt{0.2}$ multiplied by a random normal (0,1) number is added to each cell. An error term of 1 multiplied by a random normal (0,1) number is added to each cell. Thus $\rho^2 = 0.2$. The mean is taken to be 20.

The main effects are arbitrarily chosen as: $\alpha_1 = -2.0$, $\alpha_2 = -1.0$, $\alpha_3 = 0.0$, $\alpha_4 = 1.0$, $\alpha_5 = 2.0$, $\beta_1 = -2.0$, $\beta_2 = -1.5$, $\beta_3 = -1.0$, $\beta_4 = -0.5$, $\beta_5 = 0.0$, $\beta_6 = 0.5$, $\beta_7 = 1.0$, $\beta_8 = 1.5$, $\beta_9 = 2.0$, $\beta_{10} = 2.5$. The study was taken over ten such data sets. The F, interaction versus error under the maximum main effect analysis is 8.86. The best value is denoted by "B" and the worst value is denoted by "W".

Main effect estimates

α_1	T	Av	S.D.	R.M.S.
Fixed	-2.000	-1.927	.1919	.2053W
Unbiased	-2.000	-1.926	.1755	.1905
Lower	-2.000	-1.928	.1790	.1929
Upper	-2.000	-1.925	.1781	.1932
Unweighted	-2.000	-1.916	.1604	.1811B

α_3	T	Av	S.D.	R.M.S.
Fixed	0.0	-.0848	.1079	.1372
Unbiased	0.0	-.0943	.1149	.1486
Lower	0.0	-.0939	.1145	.1481
Upper	0.0	-.0857	.1293	.1551W
Unweighted	0.0	-.0861	.1042	.1352B

α_5	T	Av	S.D.	R.M.S.
Fixed	2.000	2.029	.3745	.3756W
Unbiased	2.000	2.029	.2326	.2344
Lower	2.000	2.031	.2346	.2366
Upper	2.000	2.027	.2320	.2336B
Unweighted	2.000	2.011	.2339	.2342

β_1	T	Av	S.D.	R.M.S.
Fixed	-2.324	-2.276	.2091	.2075B
Unbiased	-2.324	-2.272	.2110	.2173
Lower	-2.324	-2.281	.2065	.2109
Upper	-2.324	-2.259	.2162	.2258
Unweighted	-2.324	-2.215	.2612	.2830W

β_4	T	Av	S.D.	R.M.S.
Fixed	-0.824	-.7002	.3075	.3315W
Unbiased	-0.824	-.7049	.2649	.2904
Lower	-0.824	-.7068	.2714	.2956
Upper	-0.824	-.7045	.2600	.2861
Unweighted	-0.824	-.7181	.2361	.2588B

β_7	T	Av	S.D.	R.M.S.
Fixed	0.676	.8919	.2881	.3600
Unbiased	0.676	.8564	.3114	.3598B
Lower	0.676	.8659	.3123	.3655
Upper	0.676	.8748	.3220	.3784
Unweighted	0.676	.8809	.3530	.4082W

β_{10}	T	Av	S.D.	R.M.S.
Fixed	2.676	2.672	.1731	.2277W
Unbiased	2.676	2.722	.1688	.2137
Lower	2.676	2.718	.1703	.2149
Upper	2.676	2.725	.1678	.2017
Unweighted	2.676	2.694	.1715	.1905B

Additional studies were made using this program with smaller and larger samples. They yielded the same pattern of results as the foregoing ones. Also, due to the simplicity of the procedure involved in the setting up of a study, it would be quite easy to process as many studies as is desired.

Fach study was made using ten sets of data, each of which being the same for a given study, except fot the random numbers used in the interaction and error terms. For each main effect estimate, and for each type of analysis, the average value, the standard deviation, and the root mean square of the estimate was calculated over the ten data sets. The root mean square of the estimate was taken as the basis for comparison among the different estimating procedures. The root mean square of the estimate may be said to serve as a type of "goodness of fit" statistic. Tables of the row and column contrasts were printed out for each procedure. Also, tables of standardized row and column contrasts were printed out. These could also serve as a basis for comparison of the estimates of the different models.

In the first study the F value for interaction versus error under the maximum main effect analysis is quite significant. The F value for interaction versus error in the second study is only slightly significant. In the third study the F value for interaction versus error is slightly significant at the .01 level and non-significant at lower levels. In the fourth study the F value is significant.

Considering all of the studies made it is clearly evident that there is no best procedure to use for estimating the main effects. The computer results plainly state that one method is as good as the next. Logically, if for a given experiment the interaction effects appear significant after a fixed, maximum main effect analysis, and if the interaction effect possesses a random character, then one of the random type analyses should be best suited for estimating the main effects. However, as these and many other studies have shown, differences in estimates are quite small for the different techniques. Logical consistency would clearly earmark the random interaction analysis as best. In practice, the entire range of procedures, from fixed model least squares to unweighted means, differed so little that none of the techniques can be marked as consistently superior.

The covariance studies are designed in such a way that the interaction is significant under the maximum main effect model, and such that the interaction effect possesses a definite direction.

In one of the covariance studies the covariate was chosen in the following way: $x_{ij} = i.j/10$, where i = 1, ... 3, j = 1, 2, ..., 5, or $E(d_{ij}) = i.j/10$.
The incidence matrix has the form,

0	90	30	10	40	170
30	10	0	50	60	150
90	0	60	30	0	180
120	100	90	90	100	500

An interaction term of $\sqrt{0.2}$ multiplied by a random normal (0,1) number simulates the residual effect, in addition to the covariance term. An error term of one times a random normal (0,1) number is added to each observation. The main effects that were fed into the program are, $\alpha_1 = -1.0$, $\alpha_2 = 0.0$, $\alpha_3 = 1.0$, $\beta_1 = -2.0$, $\beta_2 = -1.0$, $\beta_3 = 0.0$, $\beta_4 = 1.0$, $\beta_5 = 2.0$. This particular study is taken over 10 sets of data.

The results of this study are as follows: Under the fixed, maximum main effect analysis, F for interaction versus error is very significant, with an average value over the ten data sets of 49.05.

Main effect contrasts

Rows:	$\hat{\alpha}_2 - \hat{\alpha}_1$	$\hat{\alpha}_3 - \hat{\alpha}_2$	$\hat{\alpha}_3 - \hat{\alpha}_1$
True Value	1.000	1.000	2.000
Covariance Analysis	.5350	1.199	1.734
	.9402	1.201	2.141
	.7457	.4058	1.052
	.5195	.5714	1.091
	1,509	1.639	3.149
	1.378	1.298	2.836
	1.748	2.318	4.066
	- .2322	.5251	.2929
	1.326	1.115	2.863
	1.209	.4878	1.697
Average	.9678	1.076	2.092
Fixed, Maximum Main Effect Analysis	-3.358	- .9172	-4.275
	-2.935	- .9054	-3.840
	-2.811	-1.528	-4.339
	-3.244	-1.474	-4.718
	-3.036	- .8311	-3.867
	-3.358	-1.364	-4.722
	-3.541	- .5566	-4.098
	-3.252	-1.116	-4.368
	-2.899	-1.411	-4.311
	-2.940	-1.768	-4.709
Average	-3.137	-1.18	-3.854

Columns:	$\hat{\beta}_2 - \hat{\beta}_1$	$\hat{\beta}_3 - \hat{\beta}_2$	$\hat{\beta}_4 - \hat{\beta}_3$	$\hat{\beta}_5 - \hat{\beta}_4$	$\hat{\beta}_5 - \hat{\beta}_1$
True Value	1.000	1.000	1.000	1.000	4.000
Covariance Analysis	1.502	- .3306	1.329	.6248	3.126
	.5880	.9871	.9481	.7920	3.833
	.8944	.0520	.9807	.8039	2.628
	.7717	.8658	.3777	.8218	2.837
	1.911	.2812	1.579	.9540	4.725
	1.466	.6617	1.625	1.599	5.351
	3.445	.2759	1.503	1.442	7.306
	- .7513	.3850	.5349	1.249	1.417
	1.482	1.921	.4671	1.313	5.183
	- .0860	2.287	.8134	1.569	4.583
Average	1.122	.8473	1.015	1.116	4.098
Fixed, Maximum Main Effect Analysis	-3.067	-1.254	- .2974	-.3645	-4.984
	-3.960	.5850	- .6717	-.1922	-4.239
	-3.282	- .8962	- .5060	-.0993	-4.783
	-3.646	- .0273	-1.195	-.1342	-5.003
	-3.426	- .7976	- .3201	-.2016	-4.745
	-4.283	- .5003	- .4219	.3557	-4.850
	-2.765	- .3385	- .7077	.0981	-3.713
	-4.297	- .3316	- .7272	.4811	-4.874
	-3.974	.8184	-1.476	.1329	-4.499
	-4.958	1.301	- .9216	.5149	-4.064
Average	-3.766	-1.441	- .7245	.0591	-4.575

The covariance studies all indicate that if for a given experiment the interaction is clearly directed then the fixed, maximum main effect analysis leads to erroneous results. This can be seen by considering the table of main effect contrasts of the previous study and observing the true values, the values obtained by the covariance analysis, and the values obtained by using the maximum main effect analysis.

We have already witnessed from the results of the studies made using the random analysis without covariance that it would not improve the situation relative to main effect estimates to perform the random covariance analysis.

In order to study the effect of making a wrong assumption on the interaction bias, that is, using incorrect values for the X's, we simulated a model where the interaction is positive if either α or β is at a low level and where the interaction turns negative at higher levels of α and β. This can be expressed algebracially by the following equation:

$$E(d_{ij}) = \begin{cases} i,j, \text{ if either } i \text{ or } j \leq 2. \\ -i,j, \text{ if either } i,\ j \geq 3. \end{cases}$$

In the analysis we make the wrong assumption that the expected value of interaction components is proportional to -i.j, for all i and j. The results of this study are given below.

Incidence Matrix

0	90	30	10	40	170
30	10	0	50	60	150
90	0	60	30	0	180
35	52	6	34	21	148
155	152	96	124	121	648

A residual (interaction) variance of .5 and error variance of 1.0 is used in this study. The F value (average) for interaction versus error is very significant under the fixed, maximum main effect analysis.

Main effect contrasts

Rows	$\hat{\alpha}_2 - \hat{\alpha}_1$	$\hat{\alpha}_3 - \hat{\alpha}_2$	$\hat{\alpha}_4 - \hat{\alpha}_3$	$\hat{\alpha}_4 - \hat{\alpha}_1$
True Value	1.000	1.000	1.000	3.000
Covariance Analysis	18.819	- .0520	9.403	28.170
	19.527	1.212	9.961	30.700
	18.890	.0490	9.631	28.570
	18.393	- .1550	9.532	27.770
	18.338	.0740	9.278	27.690
	19.105	1.131	8.904	29.140
	19.709	- .5740	11.305	30.440
	17.539	- .0350	8.696	26.200
	18.908	.6884	8.788	28.380
	18.582	.1040	9.474	28.180
Average	18.781	.2358	9.497	28.524

Columns	$\hat{\beta}_2 - \hat{\beta}_1$	$\hat{\beta}_3 - \hat{\beta}_2$	$\hat{\beta}_4 - \hat{\beta}_3$	$\hat{\beta}_5 - \hat{\beta}_4$	$\hat{\beta}_5 - \hat{\beta}_1$
True Value	1.000	1.000	1.000	1.000	4.000
Covariance Analysis	16.262	.5179	9.402	9.938	36.120
	17.113	.7766	10.121	10.604	38.620
	14.782	1.472	9.377	8.869	34.500
	15.493	.4795	8.581	9.857	34.410
	15.591	1.588	9.056	9.775	36.010
	16.939	- .5245	9.229	10.437	36.080
	15.319	1.634	8.978	9.748	35.680
	14.132	.4110	9.128	10.019	33.790
	15.394	- .5750	9.227	11.034	35.080
	15.983	.4672	8.233	8.477	34.940
Average	15.701	.6247	9.133	9.876	35.523

It is apparent from these results that it is very crucial that we make a reasonably good assumption of the X values. The results also point out that an erroneous assumption on the trend of expected values of interaction may yield as inadequate values as the assumption that they are all zero, as a comparison between this and the previous study indicates.

Thus, in summary, it may be concluded that, if interaction effects are significant but show no systematic trend from level to level, the method of analysis is, in practice, irrelevant. The entire range, from fixed effects to random effect to unweighted means analysis, yields similar results, quite adequate in all cases under study.

If the interaction effects are directed, the presence-absence or covariance analysis, with good assumptions regarding the interaction mechanism, are the only adequate methods of analysis. Disregard of, or a wrong assumption relative to the trend of such effects, leads to entirely erroneous estimates of main effect comparisons.

Introduction of Dr. R. Darrell Bock

by

Rolf E. Bargmann
University of Georgia

It is now my privilege to introduce our next speaker, Professor Darrell Bock. Dr. Bock is Professor of Education and Human Development at the University of Chicago and was formerly a Professor at the Psychometric Laboratory at the University of North Carolina at Chapel Hill. Dr. Bock received his B.S. in Chemistry at Carnegie Tech and his M.A. and Ph.D. in Educational Psychology at the University of Chicago. He is on the Board of Trustees of the Psychometric Society and on the Board of Regional Advisors of the Biometric Society. His research preference is psychometrics and psychological statistics, and with his permission I am dropping the word psychological, so shall we say, psychometrics and statistics, and computation. Dr. Bock will give a summary, but he will introduce it by some rather important research concerned with the analysis of variance in non-experimental settings in which Dr. Bock will present some of his ideas.

REMARKS ON ANALYSIS OF VARIANCE AND ANALYSIS OF REGRESSION

R. Darrell Bock
University of Chicago

Our hosts, Dr. Findley and Dr. Bashaw, have given me the assignment of commenting upon the excellent papers which we heard yesterday from Professor Graybill, Dr. Ward, and Professors Winer and Bargmann. The task will be an easy one because our speakers touched on so many of the problems which arise in the use of linear models in data analysis that I have a wide field to play on. They were also considerate enough to leave at least a few questions unanswered, thus, giving me the opportunity to interject my own opinions here and there. You can be sure I will not let this splendid opportunity pass me by. I would like, however, to have the privilege of speaking on a selection of topics suggested by yesterday's papers, rather than the more difficult task of discussing each paper as a whole. If you will permit me this, I will begin by directing some comments to Professor Winer.

Analysis of repeated measurements data

Let me correct slightly Professor Winer's reference to the computer programs which we use at Chicago for analyzing repeated measurement data by means of multivariate analysis of variance. Actually, we have only one program--a highly general univariate and multivariate analysis of variance program. It gives exact least-square analyses in the case of missing and unequal numbers of observations within subclasses, and includes provisions for analysis of covariance and analysis of regression. This program, which bears the cryptic title MFSA 95, was written by Jeremy Finn, now with the Department of Educational Psychology, State University of New York at Buffalo. The program is based on flow diagrams which I prepared for the IBM Computer Symposium on Statistics, 1963 (Bock, 1965). The original version of the program utilized special features of the Chicago operating system and could not readily be used at other installations. In the meantime, however, Finn has prepared a new version, called "MULTIVARIANCE", which is written entirely in FORTRAN IV and should operate on any machine which has FORTRAN IV capability (Finn, 1967). Finn now has this program and its documentation ready for distribution. As many of you know, a similar program, called "MANOVA," has been prepared by Dr. Elliot Cramer, and is available from the Biometric Laboratory, University of Miami (Clyde, Cramer, and Sherin, 1966).

Professor Winer remarked correctly that the multivariate analysis of variance program can be used to analyze repeated measurements data under more general models than those assumed in the mixed model analysis. He added the qualification, however, that multivariate analysis of variance is difficult to interpret. Actually, this has not been our experience at Chicago. We find that, if the person using the program has some familiarity with univariate analysis of covariance and with component and factor analysis, he has little trouble in understanding such multivariate content of the program as the "step-down" F-statistics, discriminant functions, canonical variates, or multivariate tests of the joint significance of multiple dependent variables. His difficulties in understanding the analysis occur more often in regard to the analysis of variance as such, rather than with its multivariate aspect. Typically, his problems concern the interpretation of significant interactions, or how to judge the importance of a significant main effect, how to interpret the adjustments made in the analysis of covariance, how to decide what is testable or estimable when there are significant interactions, how to choose the appropriate error term in mixed-model analysis, etc. In addition, there are, in the case of unequal subclass numbers, special problems in the interpretation of non-orthogonal analysis of variance which come as an unpleasant surprise to persons who feel they are expert in orthogonal analysis of variance.

In short, our experience has been that, if we can assume on the part of the user a good knowledge of analysis of variance, then we can offer him in the multivariate analysis of variance program a convenient and hightly flexible vehicle for the analysis of repeated measures designs. This is particularly true when the design is some complex form such as the Lindquist type VI design, with a crossed or nested classification of both measures and subjects. In the multivariate treatment, the investigator need only concern himself with setting up the appropriate analysis for the design of the _subject_ classification, and of specifying a certain linear transformation of the repeated measurements. Once these two types of information are supplied, the appropriate analysis, including the choice of error terms, falls out of the analysis in a natural way.

The comparative study in behavioral research

My next comment is directed to both Professor Graybill and Professor Winer. They have expressed a distrust of statistical analysis applied to what they called "historical research," that is, to research carried out in a natural, as opposed to an experimental, setting. In my view, the term "historical" is not entirely accurate here. The events being studied are not fixed in the historical past. The studies can be replicated and the systematic nature of the events can be established. Since the objective of these studies is to compare the responses of subjects under various identifiable conditions, it would be more accurate to call them "comparative studies." I do not believe that we can rule out the comparative study in biological and behavioral research without sacrificing sources of information which are potentially of great importance. If we were to exclude comparative studies from biological research, we would for example, sacrifice much of the field of epidemiology, where comparative studies have been spectacularly successful. Can you imagine Edward Jenner discounting, because it was based on comparative study, the observation that the incidence of smallpox was lower among milkmaids than in the population generally? Fortunately for us, he did not ignore this datum but went on to make the connection that milkmaids were likely to have had cowpox, and that cowpox produces immunity to smallpox, and that possibly people could be protected from smallpox by vaccination with cowpox virus.

Or to take an example from behavioral science, I think that most of us would admit that the remarkable constancy of the rate of incidence of schizophrenia in different countries, in different socio-economic classes, and in different historical periods, as revealed by comparative studies, has an important bearing on where we should look for the causes of this disorder. Certainly, it discourages a theory exclusively based on response environmental stress, which must differ from one population to another, and suggests we explore instead a biogenic mechanism which is at a more-or-less steady state in these populations.

I believe that we should accept the comparative study in behavioral research, but we must be conscious of its limitations. The most that such a study can achieve is a description of systematic differences in the responses of different classes of subjects. It may demonstrate that certain responses and certain characteristics are associated, but it does not tell us we can change subjects' responses by changing their characteristics. It takes an experimental study following up the comparative study to establish this kind of practical knowledge.

Unhappily, the causal interpretations which are gratuitously imputed to mere associations give comparative studies a bad name. I recently came across an example of this kind of abuse which is so flagrant that one is left stunned. In summarizing responses to a questionnaire item contained in the Coleman Report (Equality of Educational Opportunity), one reviewer stated: "The most telling factor in achievement is the attitude of students toward themselves. When students feel they have control over their environment and destiny they achieve more." Apparently, the converse interpretation--that students who achieve more feel that they have more control over their environment and destiny and will express this feeling on a questionnaire item--did not occur to the reviewer.

Obviously, the data are incapable of distinguishing between the two interpretations.

If we discount this kind of over-zealous embracing of causal explanations, however, we can accept the comparative study as a useful research strategy in early stages of investigation where we are seeking ideas which may be followed up in more experimentally-oriented studies. In behavioral science especially, where most research is at the stage of preliminary investigation, we can expect the comparative study to be widely used, as indeed it is at the present time.

When discussing the comparative study, in which the investigator deliberately sets out to contrast certain populations or sub-populations, I consider it important to distinguish this type of study from a survey, in which the investigator attempts to describe a single population on the basis of a sample of subjects. Before selecting the sample in the comparative study, the investigator identifies the characteristics which identify the various subclasses which he wishes to compare. He then goes into the population and, from among those subjects who fall into a particular subclass, he draws a random sample and measures some response of the subjects he has selected. In effect then, each subclass constitutes a separate population, and it is only within these populations that the investigator has to maintain random sampling. This is often advantageous since it is usually easier to randomly sample within a narrow class, where the subjects may be geographically more accessible, or where lists of all subjects in the population are available. Furthermore, the numbers of subjects to be selected for the subclasses are largely at the disposal of the investigator, and he may choose them so as to obtain the best possible precision in estimating the comparisons of interest.

The procedure in a survey is much different. The investigator samples randomly from the general population, and then classifies subjects according to the characteristics which he wishes to associate with the response variables. In this case a number of subjects in the subclasses is a random variable which reflects the population portions for the subclass. In a survey study, it makes sense to collapse data over various ways of classification and to describe differences between certain groups ignoring other ways of classification. Since each way of classification samples the entire original population, statistics based on the collapsed data can be identified with definite population parameters. In a comparative study, where the number of subjects in the various classes are arbitrary, it is in general meaningless to collapse some of the ways of classification because the resulting groups of subjects do not represent any real population. This means that the statistics such as correlation coefficients or correlation ratios, which neasure sone variance components with respect to the total variance, are meaningless in a comparative study, because the total variance is arbitrary and does not refer to any real population. Thus, the device of quoting the per cent of variance accounted for (i.e., the multiple r^2), which is popular among people who use regression methods to analyze this form of data, has no general meaning when applied to comparative study. What is really of interest in these studies is the size of the effects associated with the various classes of subjects or the interactions of classes. If the metric of the response measures is arbitrary, so that these differences have no clear absolute meaning, then the best we can do is to compare effects estimated for some classes or some variables with other classes and other variables. Thus, it may be possible to say that while both A- and B-way-of-classification in a comparative study clearly have statistically significant effects, the effects of the B-way are, say, an order of magnitude smaller than those of the A-way. It must be understood that we cannot get this information by comparing, let us say, F-statistics for the respective ways of classification, because the F-statistics reflect the precision with which the effects are estimated and not the magnitudes of the effects themselves. The actual least-squares estimates of the effects need to be estimated and examined.

Considerations of interpretation have an important connection with the prior question of how the investigator should set up the analysis of data from a comparative study assuming, as we are here, that the analysis is based on a linear statistical model. From the point of statistical method, it would be convenient if we could assume that the investigator always selects equal numbers of subjects for the subclasses of design. Then an orthogonal analysis would be possible and the computation and interpretation would be simplified. But a comparative study seldom works out this neatly. Sometimes the design includes subclasses for which few or no representatives can be found. Sometimes subjects withdraw from the sample before the data are collected, or sometimes data are found to be erroneous or mixed up and cannot be replaced. Inevitably, lack of time or money prevents the investigator from filling out the design, and he may elect to analyze the data he has on hand. If so, he can take two approaches to the analysis, and they are well represented by the papers given yesterday by Professor Bargman and Dr. Ward.

Analysis of variance vs. regression analysis

Professor Bargmann, supposing in his paper that the investigator has set up the analysis in the form of an analysis of variance, raises the question of how accurate are alternative methods for estimating main class effects assuming unequal subclass numbers and the presence of interaction. His calculations show that the exact method, the method of fitting constants, is not appreciably different from the approximate method of unweighted means in terms of the accuracy with which they recover main class effects in Monte Carlo data. He expresses some surprise that the adding of an interaction in the form of a random cell effect shows very little influence on the estimation of main effects. I find less reason for surprise, however, since adding a random cell effect is essentially the same as adding a random sampling error to observations within cells so far as main effects are concerned. The difference is only that the variance contributed by the cell effect is larger in proportion to the number of observations within the subclasses. The random cell effects tend to average out, especially when large numbers of levels are involved as in Bargmann's examples, and are not readily observable in the estimates of main effects. In real data, however, they are systematic rather than random and, as some of Bargmann's later examples show, the main effects are not estimable in the presence of such systematic interactions.

As to the relative merits of the exact analysis versus unweighted means, Bargmann's calculations show that it is not in estimation, as such, that the difference between the two methods is brought out. After all, both methods give unbiased estimates of main class effects. The important difference is that, in unweighted means, one has only the crude estimate of the error variance computed using the harmonic mean of the subclass numbers as a compromise figure for sample size. In the least-square analysis one has a best quadratic estimate of the error variance and, hence, can calculate exact F-statistics and confidence intervals. I, for one, would not wish to give up these advantages of exact analysis of variance merely to avoid more complex computation and somewhat more difficult interpretation. My preference is to make use of programs capable of performing an exact analysis in the non-orthogonal case and to learn how the non-orthogonal analysis differs in its interpretation from the orthogonal analysis with which we all are familiar.

The other approach to the analysis of data from a comparative study is closely identified with the work of Dr. Ward (even though in his paper he gives "equal time" to analysis of variance and analysis of covariance). In this approach the investigator is advised to set up his analysis as a regression problem, using dummy variables to represent the various classes in the design and their interactions. As Dr. Ward points out, the analysis which results is numerically equivalent to an exact least-square analysis whether or not the subclass numbers are equal or unequal. I think it must be admitted that a good part of the motivation for using this approach is the fact that good regression programs have been available for computers longer than have been flexible analysis-of-variance programs. But let's suppose the investigator has available both a regression program and an analysis of variance program such as MULTIVARIANCE or MANOVA. Is there any advantage to his taking the pure regression approach to data obtained from a designed experiment or comparative study?

Reparameterization of design models

A problem with the pure regression approach is that it forces the investigator to rely almost entirely on F-ratios and tests of significance, while making it difficult for him to make use of the estimates of the effects represented in the fitted regression coefficients. This is true because when applied to design models, the regression analysis has the effect of transforming the parameters of the original linear model without giving the user any indication of what transformation is involved. Let me illustrate this by a simple example. Consider the problem of analyzing data from a two-by-three cross-classification. Suppose we attempt to fit a model

$$E(y_{jk}) = \mu + \alpha_j + \beta_k \qquad (j = 1,2;\ k = 1,2,3)$$

where y_{jk} is the measurement of the response of a randomly selected subject from the j-th A-class and the k-th B-class;

μ is a constant term which incorporates the arbitrary origin of measurement on the response scale;

α_j is the effect on the response associated with membership in the j-th class of the A-way-of-classification and

β_k is the effect on the response associated with membership in the k-th level of the B-way-of-classification.

To anyone acquainted with multiple regression analysis, it will be clear that, since there are no limitations on the independent variables of a regression problem, except that they be real numbered variables, the problem of fitting a model may be cast as a regression problem. This may be done by deploying independent variables, the quantities x_0, x_1, x_2, x_3, x_4, x_5, which take on values 1 or 0 according to whether the associated effect is present or absent, that is,

$$E(y_{jk}) = x_0 + \alpha_1 x_1 + \alpha_2 x_2 + \beta_1 x_3 + \beta_2 x_4 + \beta_3 x_5 \text{ .}$$

This means that we can perform the regression analysis on data which takes the form shown in Table 1.

A more efficient analysis from the point-of-view of computation, however, may be formulated as a weighted least-squared fitting of the subclass means, where the weights are the subclass numbers, n_{11}, n_{12}, n_{13}, n_{21}, n_{22}, and n_{23}. All information in the data necessary for this solution may be summarized in the form of Table 2. This form of analysis calls attention to the estimated cell means and variances, which as Professor Graybill stresses, should be available for the investigator's inspection.

Table 1

Data for Least-Squares Analysis

Dependent Variable	Independent Variables					
$Y_{(i)jk}$	x_0	x_1	x_2	x_3	x_4	x_5
$Y_{(1)11}$	1	1	0	1	0	0
$Y_{(2)11}$	1	1	0	1	0	0
⋮			⋮			
$Y_{(n_{11})11}$	1	1	0	1	0	0
$Y_{(1)12}$	1	1	0	0	1	0
$Y_{(2)12}$	1	1	0	0	1	0
⋮			⋮			
$Y_{(n_{12})12}$	1	1	0	0	1	0
⋮			⋮			
$Y_{(1)23}$	1	0	1	0	0	1
$Y_{(2)23}$	1	0	1	0	0	1
⋮			⋮			
$Y_{(n_{23})23}$	1	0	1	0	0	1

Table 2

Summary Data for Analysis of Variance

Subclasses	Subclass Numbers (Weights)	Variances	Dependent Variable $y_{\cdot jk}$	Subclass means Independent Variables x_0	x_1	x_2	x_3	x_4	x_5
1. 1,1	n_{11}	s^2_{11}	$y_{\cdot 11}$	1	1	0	1	0	0
2. 1,2	n_{12}	s^2_{12}	$y_{\cdot 12}$	1	1	0	0	1	0
3. 1,3	n_{13}	s^2_{13}	$y_{\cdot 13}$	1	1	0	0	0	1
4. 2,1	n_{21}	s^2_{21}	$y_{\cdot 21}$	1	0	1	1	0	0
5. 2,2	n_{22}	s^2_{22}	$y_{\cdot 22}$	1	0	1	0	1	0
6. 2,3	n_{23}	s^2_{23}	$y_{\cdot 23}$	1	0	1	0	0	1

In matrix notation, the weighted regression solution for these data can be expressed in compact form for a general model involving 6 subclass means and 6 effects to be estimated. Required in the solution is the 6 x 6 matrix of independent variables, X, shown at the right of Table 2, the 6 x 1 vector of subclass means $\underline{y}$ shown in the center of Table 2, and a 6 x 6 diagonal matrix, D, whose elements are the subclass numbers in the order shown at the left. In this example, let the effects to be estimated be the 6 x 1 vector $\underline{\xi}$, with elements μ, α_1, α_2, β_1, β_2 and β_3. Then the n = 6 equations of the model can be expressed in the matrix equation

$$E(\underline{y}.) = X\underline{\xi} \quad .$$

It can be shown that the least-squares estimate of $\underline{\xi}$ is a solution of the so-called "normal" equations,

$$X'DX\underline{\xi} = X'D\underline{y} \quad .$$

Solving the normal equations is complicated, however, by the fact that independent variables for experimental design models are subject to linear dependencies because certain columns of the model matrix are the sums of other columns. In the example, $x_1 + x_2 = x_3 + x_4 + x_5 = x_0$ in all rows of the model matrix, X. We say then that the model is of "deficient rank," or "not of full rank," where rank refers to the number of linearly independent columns.

Many workers are aware that sophisticated regression algorithms are capable of giving a solution in spite of the linear dependencies, but few understand how these procedures work or how the model is altered in process. To explain this, it is first necessary to establish that the normal equations have a solution when the rank of X is less than m, the number of parameters. A solution is assured by a theorem of elementary matrix algebra which gives the necessary condition for a consistent solution of a system of linear equations. The condition is that the constant terms on the right be subject to the same linear dependencies as the matrix of coefficients. If it satisfies this condition, a system of n equations in m unknowns, with matrix of coefficients of rank $\ell \leq n$, has a solution for ℓ of the m unknowns in terms of the remain $n - \ell$ unknowns. If values for the latter unknowns are arbitrarily assigned, the system has an actual numerical solution. Clearly the solution is not unique because it depends on the $n-\ell$ arbitrarily assigned quantities. It is easy to show in the context of linear statistical models, that the normal equations fulfill this condition, and that the arbitrary assignment of unknowns merely amounts to choosing the origin of the scale of measurement of the effects. Because we are all accustomed to scales with arbitrary origins, e.g., the Fahrenheit and Celsius scales of temperature, and because most statistical procedures are invariant under translation of scale from one origin to another, this form of non-uniqueness can be tolerated.

The simplest method of solving the normal equations for models not of full rank is to omit redundant variables as they are encountered in the forward part of the solution of systems of linear equations by elimination, as in the Gauss-Doolittle, square-root, Gauss-Jordan, or simple bordering methods (See Bodewig, 1959; Householder, 1953). As the eliminations are performed, null rows will be encountered when a redundant variable is reached. If these rows, and the corresponding columns of X'DX are dropped, the remaining ℓ equations are of full rank and may be solved in the conventional back solution. This is the solution most frequently used.

It is essential for the worker to understand, however, that an alteration of the original model is implied in this procedure. It amounts to arbitrarily setting to zero the <u>last</u> effect encountered in each way of classification. This is easy to demonstrate in the present example. It is clear that the parametric form of the model can be altered in the following way without distrubing the equalities,

$$
\begin{array}{lllllll}
E(y_{11}) = (\mu + \alpha_2 + \beta_3) & + (\alpha_1 - \alpha_2) & & + (\beta_1 - \beta_3) & & \\
E(y_{12}) = (\mu + \alpha_2 + \beta_3) & + (\alpha_1 - \alpha_2) & & & + (\beta_2 - \beta_3) & \\
E(y_{13}) = (\mu + \alpha_2 + \beta_3) & + (\alpha_1 - \alpha_2) & & & & + (\beta_3 - \beta_3) \\
E(y_{21}) = (\mu + \alpha_2 + \beta_3) & & + (\alpha_2 - \alpha_2) & + (\beta_1 - \beta_3) & & \\
E(y_{22}) = (\mu + \alpha_2 + \beta_3) & & + (\alpha_2 - \alpha_2) & & + (\beta_2 - \beta_3) & \\
E(y_{23}) = (\mu + \alpha_2 + \beta_3) & & + (\alpha_2 - \alpha_2) & & & + (\beta_3 - \beta_3)
\end{array}
$$

Now let: $\mu + \alpha_2 + \beta_3 = \mu_{(c)}$; $\alpha_1 - \alpha_2 = \alpha_{1(c)}$; $\beta_1 - \beta_3 = \beta_{1(c)}$ and $\beta_2 - \beta_3 = \beta_{2(c)}$. Since $(\alpha_2 - \alpha_2)\, x_2 = 0$ and $(\beta_3 - \beta_3)\, x_5 = 0$, the model may be expressed in terms of the 3 parameters rather than 5, and the linear dependencies are eliminated:

$$E(y_{11}) = \mu_{(c)} + \alpha_{1(c)} + \beta_{1(c)}$$

$$E(y_{12}) = \mu_{(c)} + \alpha_{1(c)} \qquad + \beta_{2(c)}$$

$$E(y_{13}) = \mu_{(c)} + \alpha_{1(c)}$$

$$E(y_{21}) = \mu_{(c)} \qquad + \beta_{1(c)}$$

$$E(y_{22}) = \mu_{(c)} \qquad + \beta_{2(c)}$$

$$E(y_{23}) = \mu_{(c)}$$

In matrix notation

$$E(\underline{y}) = K_{(c)}L_{(c)}\,\underline{\xi}$$

$$= K_{(c)}\underline{\theta}_{(c)}, \text{ say,}$$

where

$$K_{(c)} = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \text{ and } L_{(c)} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

In the corresponding normal equations

$$K_{(c)}'DK_{(c)}L_{(c)}\,\underline{\xi} = K_{(c)}'Dy. \;,$$

the $\ell \times \ell$ matrix of coefficients $K_{(c)}\,'DK_{(c)}$ is of rank ℓ and has an inverse. The least-squares solution may therefore by expressed as

$$L_{(c)}\underline{\xi} = (K_{(c)}'DK_{(c)})^{-1}K'_{(c)}D\underline{y}.$$

$$= M\underline{y}.$$

$$= \underline{\hat{\theta}}_{(c)} \;;$$

i.e., $\underline{\hat{\theta}}_{(c)}$ is the least-squares estimate of $L_{(c)}\underline{\xi}$.

The ℓ x n matrix M has been called the "estimation" matrix for the design (Bock, 1963).

We may illustrate this solution numerically with the data of Table 3.

Table 3

Artificial Data

Subclass	Classification A	B	Subclass Number	Mean	Standard Deviation
1	1	1	10	21.2	5.83
2	1	2	10	23.4	5.56
3	1	3	9	28.7	5.65
4	2	1	7	20.1	5.87
5	2	2	9	21.3	5.36
6	2	3	10	23.5	5.92

The least-square estimate of $\underline{\theta}_{(c)}$ is:

$$\hat{\underline{\theta}}_{(c)} = \begin{bmatrix} 24.600 \\ 2.877 \\ -5.546 \\ -3.709 \end{bmatrix} \begin{array}{l} \text{Constant term } (\mu + \alpha_2 + \beta_3) \\ \alpha_1 - \alpha_2 \\ \beta_1 - \beta_3 \\ \beta_2 - \beta_3 \end{array}$$

The foregoing is a simple example of the reparameterization of a linear model. It is a linear reparameterization and the transformation is represents is specified by the ℓ x m matrix L. The new parameters $\underline{\theta}_{(c)}$ are called linear parametric functions of the original parameters (Bose, 1960).

Reparameterization in the partition of the total sum of squares

If it is true, as I have tried to show, that the reparameterization of the design model implied in certain regression procedures should be made explicit so that the user will know what is estimated, then it is doubly true that the further reparameterization implied in the associated analysis of variance must be explicated if the user is to know what is being tested. If the user does not understand the logic behind the calculation of the F-statistics in the regression analysis, he is in danger of misinterpreting the tests of significance based on them, especially in the analysis of design models. I have in mind, in particular, those computing algorithms which produce an F-statistic for each parameter in the model, eliminating all other parameters, or, equivalently, those which produce partial correlations between the dependent variable and each independent variable, while holding fixed all remaining independent variables.

Fundamentally, these procedures involve the additive partition of the total sum of squares which was introduced into statistical practice by R. A. Fisher. There are many ways, geometric and algebraic, to understand the meaning of this partition, but perhaps none is clearer than an explanation in terms of the corresponding reparameterization of the design model.

The objective of this "second" reparameterization is the construction of certain parametric functions whose estimators are uncorrelated and have common variance. Applied to data, these estimators yield what may be called _orthogonal estimates_ (to use terminology suggested by Bargmann), although Durand has called them _semi-partial_ regression coefficients

(Durand, 1956). If the sampling errors are similarly, independently, and normally distributed, the squares of the orthogonal estimates are (possibly non-central) independent chi-square variates, and the F-distribution applies.

The reparameterization implied in the partition of the total sum of squares repays study because it makes utterly clear the meaning of the analysis of variance. The reparameterization is accomplished by factoring the basis matrix for the design into the product of a matrix orthonomal by columns (possibly with respect to a matrix of weights), and an upper triangular matrix:

$$\underset{n \times \ell}{K} = \underset{n \times \ell}{P}\,\underset{\ell \times \ell}{T}$$

(schema)

where P'DP = I. Numerically, the factorization may be carried out by a generalized Gram-Schmidt process (Householder, 1953, p. 72) or by Householder's orthogonal triangularization (Householder, 1964, pp. 133-134). (Finn's MULTIVARIANCE program uses the former and Cramer's MANOVA program the latter.) Whatever the numerical method, the reparameterized model becomes,

$$\begin{aligned} E(\underline{y}) &= PT\underline{\theta} \\ &= P\underline{\upsilon}\text{ , say.} \end{aligned}$$

The least-square estimate of $\underline{\upsilon}$ is

$$\begin{aligned} \hat{\underline{\upsilon}} &= (P'DP)P'D\underline{y} \\ &= P'D\underline{y} \\ &= \underline{u}\text{ , say.} \end{aligned}$$

Its vector expectation and variance covariance matrix are, respectively,

$$E(\underline{u}) = \underset{\ell x1}{\underline{\upsilon}} = T\underline{\Theta} \text{ and}$$

$$V(\underline{u}) = \underset{\ell x\ell}{I\sigma^2} ,$$

where σ^2 is the common error variance.

The triangularity of T is crucial here:

$$T = \begin{bmatrix} t_{11} & t_{12} & & t_{1,\ell-1} & t_{1\ell} \\ 0 & t_{22} & \cdots & t_{2,\ell-1} & t_{2\ell} \\ . & . & & . & . \\ 0 & & & t_{\ell-1,\ell-1} & t_{\ell-1,\ell} \\ 0 & 0 & \cdots & 0 & t_{\ell\ell} \end{bmatrix}$$

It means that the last element of $\underline{\upsilon}$ involves only the last element of $\underline{\Theta}$, (i.e., $\upsilon_\ell = t_{\ell\ell}\Theta_\ell$); that the next-to-the-last element of $\underline{\upsilon}$ involves the last two elements of $\underline{\Theta}$ (i.e., $\upsilon_{\ell-1} = t_{\ell-1, \ell-1}\,\Theta_{\ell-1} + t_{\ell-1,\ell}\Theta_\ell$) , and so on, until the first element of $\underline{\upsilon}$ involves all elements $\underline{\Theta}$.

It is clear, then, that a test of the hypothesis that $\nu_\ell = 0$ is equivalent to testing the hypothesis $\Theta_\ell = 0$, since if T is of rank ℓ, $t_{\ell\ell}$ cannot be zero.

A test of the hypothesis $\upsilon_{\ell-1} = 0$, on the other hand, is equivalent to a test of $\Theta_{\ell-1} = 0$ if, and only if, one or both of two conditions are met--either $t_{\ell-1,\ell} = 0$ or $\Theta_\ell = 0$. The salient difference between an orthogonal and non-orthogonal analysis is that, in an orthogonal analysis, if Θ_ℓ and $\Theta_{\ell-1}$ are effects of different ways of classification or different interactions, the first of these conditions is met, i.e., $t_{\ell-1,\ell} = 0$. It is because of the vanishing of certain above-diagonal elements of T that the independent interpretation of each line of the analysis of variance table, with which we are all familiar, is possible.

If the analysis is non-orthogonal, on the other hand, none of the above-diagonal elements of T is zero independent of the arbitrary subclass numbers. This means that an hypothesis such as $\Theta_{\ell-1} = 0$ is testable in this reparameterization if, and only if, Θ_ℓ is assumed null. In this case, only "stepwise" testing of hypotheses about effects in the reparameterized model, $E(y) = K\underline{\Theta}$, is possible. If an independent estimate of the error variance is available, e.g., from the replications within cells, then a variance ratio corresponding to each orthogonal estimate, or for two or more of the estimates pooled together, may be inspected for statistical significance. The inspection begins with the last orthogonal estimate and proceeds in order to the first. When one of the variance ratios is found to exceed a predetermined critical value, the process is terminated and the number of parameters to be included in the model is established.

This procedure provides a decision rule for determining the most parsimonious model which is consistent with the data. Its statistical justification has been given by Roy and Bargmann (1958) and T. W. Anderson (1962), who show that, under the null hypothesis the tests at each stage are stochastically independent. The overall error rate of the procedure is therefore easy to calculate. Specifically, if a test with type I error equal to α_i is made at the i-th stage, then the probability α of accepting the null hypothesis at each stage when it is in fact false for at least one of m stages is,

$$\alpha = 1 - \prod_{i=1}^{m} (1 - \alpha_i) .$$

To see how the foregoing results apply to the type of regression analyses mentioned above, in which an F-statistic is computed for each parameter eliminating all others, we observe that this analysis is equivalent to orthogonalizing K m times (if there are m parameters in question) with each parameter in turn in the last position. Viewed in this way, we see that there are a number of pitfalls associated with this procedure which may trap the unwary user:

First, the Roy-Bargmann and Anderson result applies to only one such ordering. Since the results of multiple orderings are not independent, the calculation of error rates is hopelessly complicated. This presents a problem in the analysis of non-orthogonal factorial designs. A single partition of the sum of squares for the non-orthogonal design does not have the same effect as the partition for the corresponding orthogonal designs. To obtain the effect of an orthogonal analysis, we must perform as many partitions as there are factors. In these partitions, we would order the bases vectors in K so that vectors corresponding to each main-effect appear last in one of the partitions. In practical work, we may need to test all factors and thus may be obliged to proceed in this manner. If we do so, we will be able to make a probability statement which is correct for any given partition, but not a statement which applies to the partitions jointly. Admittedly, most practical workers will not be greatly disturbed by this limitation (although perhaps they should be), because in factorial analyses they usually make their probability statements for each effect separately rather than jointly. Strictly speaking, however, it would be preferable to identify before the analysis the effect which is to be tested critically and to employ a single partition in which that effect enters last. An exact probability statement will then be possible.

Second, when design models with interactions are involved, there is no logical justification for placing each possible parametric function last in the ordering and testing it there. Consider, for example, the 2 x 3 design discussed above. There are only two orderings possible for this design, and they are shown, together with the numerical results for the artificial data, in Table 4. The only room for choice is in the order in which the two main effects are introduced.

Table 4

Partition of the total sum of squares
for the 2 x 3 design

Source	d.f.	Sum of Squares	Source	d.f.	Sum of Squares
Constant Term	1	29404.0151	Constant Term	1	29404.0151
A-classes, ignoring B	1	83.1732	B-classes ignoring A	2	259.0878
B-classes, eliminating A	2	288.3770	A-classes eliminating B	1	112.5323
Interaction	2	41.4238	Interaction	2	41.4238
Within subclasses	49	1594.9993	Within subclasses	49	1594.9993
Total	55		Total	55	

It would be illogical to attempt to eliminate interaction from the main effects for the following reason. There are actually six interactive parameters in the original model. The two degrees of freedom for interaction in Table 4 represent merely the two possible linear functions of these six parameters which are linearly independent of the main effects. If one wanted to eliminate the interactive effects, he would assign one degree of freedom to the constant term and five to the interactions. But this would exhaust the degrees of freedom for the design without the main effects having been included. It would, in fact, reduce the analysis to a one-way design, which is, of course, precisely what the interaction terms amount to. Obviously, any computing procedure which would include interactive functions ahead of main effect functions when orthogonalizing the models (explicitly or implicitly) in the partition of the total sum of squares is erroneous.

Third, there is no logical justification for testing individual degrees of freedom within main effects or interactions unless the classes in the way of classification are ordered or structured in some way. It should not be thought that, say, the simple contrasts of each class with the last class (as in the artificial example above) could be tested separately by means of the corresponding orthogonal estimate if the classes are nominal. One must remember that a further reparameterization of the model occurs in the orthogonalization. Indeed, when the subclass numbers are equal, the simple contrasts are turned into "Helmert" contrasts in the orthogonalization. (A Helmert contrast is the difference between the effect of, say, the i-th group and the mean of the effects of groups i + 1 to n when the n groups are ordered in some way.) If the structure of the groups is meaningful, then "one-degree of freedom" tests in that order may be useful; if not, the degrees of freedom and the corresponding squared orthogonal estimates for the way of classification should be pooled. Of course, when the groups are ordered with known spacing, the one-degree of freedom analysis using orthogonal polynomial contrasts is meaningful and often valuable. Other types of structuring are possible. I am indebted to Dr. Elliot Cramer for the following example. Suppose two related drugs and a placebo are being tested in three independent groups of subjects. I would consider it correct and meaningful to perform single-degree-of-freedom tests on two ordered contrasts among the three groups. The first would contrast the placebo group with the mean of the two drug groups. The second would contrast the two drug groups. In a non-orthogonal analysis, the effect for second contrast would be tested eliminating that for the first. If this test should show that the drugs differed in their effects, the analysis would terminate. A drug effect of some kind would have been demonstrated and the estimated effects would be examined to determine the nature of the effect. On the other hand, if the test of the second contrast did *not* show the drugs to differ, then it would be necessary to test the first contrast, ignoring the second, in order to determine if the assumed equal effects of the two drugs are different from the placebo effect. This formulation of the analysis assumes a linear model in which the effect of a drug is the sum of a placebo effect and a true drug effect, whereas the placebo has only the placebo effect.

Lastly, I would like to point out that when the independent variables are *random* variables, the sum of squares of the *orthogonal estimates*, divided by the total sum of squares, is a squared multiple correlation coefficient (r^2). The separate terms in this sum are valuable in that they show how much r^2 will increase when the corresponding variable is added to the regression equation, given that all variables preceding it in the ordering are already in the equation. Notice that this interpretation depends upon the arbitrary order in which the orthogonalization is carried out. It is deplorable, but true of the psychological literature, that the product of the standardized regression coefficients and the corresponding first order correlation has been advocated as an index of the proportion of variance accounted for by each independent variable which does not depend on the order of independent variables (Hoffman, 1960). Actually, there is no sense in which one of these products can be regarded as proportional to the percentage of variance determined by an independent variable. In the first place, the product can be negative, which is not admissible for a proportion; and in the second place, it does not correspond to the proportionate reduction of r^2 when the variable is removed from the regression, which is the only possible way to make a statement about the contribution of a variable independent of order. Let us hope that this erroneous index has not found its way into computer programs that statistically naive educationists or psychologists might use.

Interpreting main-effects and interactions

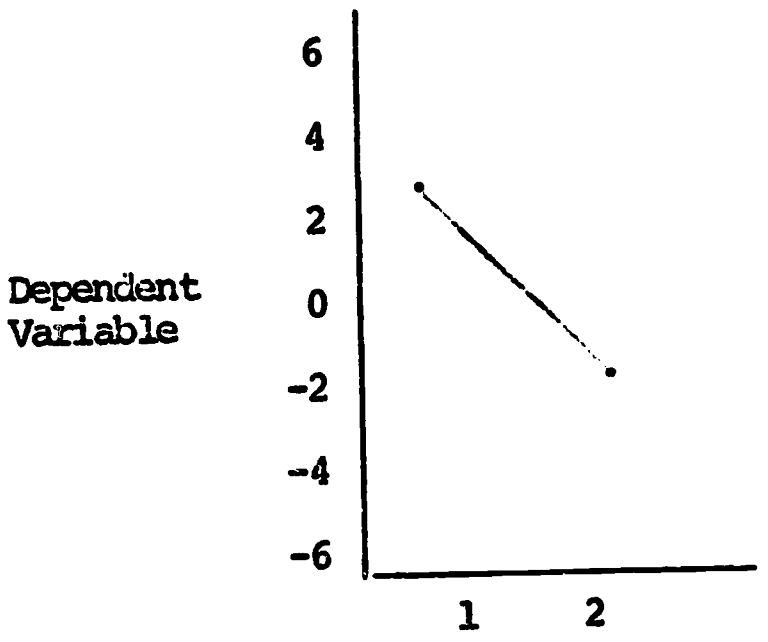
A final problem which I see in the routine use of general regression programs for designed studies concerns the interpretation of effects. Personally, I find no difficulty in interpreting the quantities which are actually estimable in the design models, namely, the contrasts of effects between classes and subclasses. But after five years of unrelieved failure to get any applied worker to think in terms of effect-contrasts, I am beginning to get the message: the natural way to interpret effects in a designed study is in terms of the estimated means of the relevant main-classes and subclasses.

No doubt this fixation on marginal means is the result of constant exposure to orthogonal analysis of variance where the marginal means are, in fact, best unbiased estimates of effects in the model under conventional restrictions. Unfortunately, this practice can be carried over to nonorthogonal analysis only in a study where the data as a whole constitute a probability sample of the defined population. In this case, the marginal means are best unbiased estimates of the population means for some classes of the design when other ways of classification are ignored.

If, on the other hand, the subclass numbers are arbitrary, either by design or by attrition in the sample, the marginal means do not estimate the corresponding mean effects. I have tried (unsuccessfully) to convince users that they should fix attention on the effects, rather than on means, and display best estimates of effects when reporting on a study.

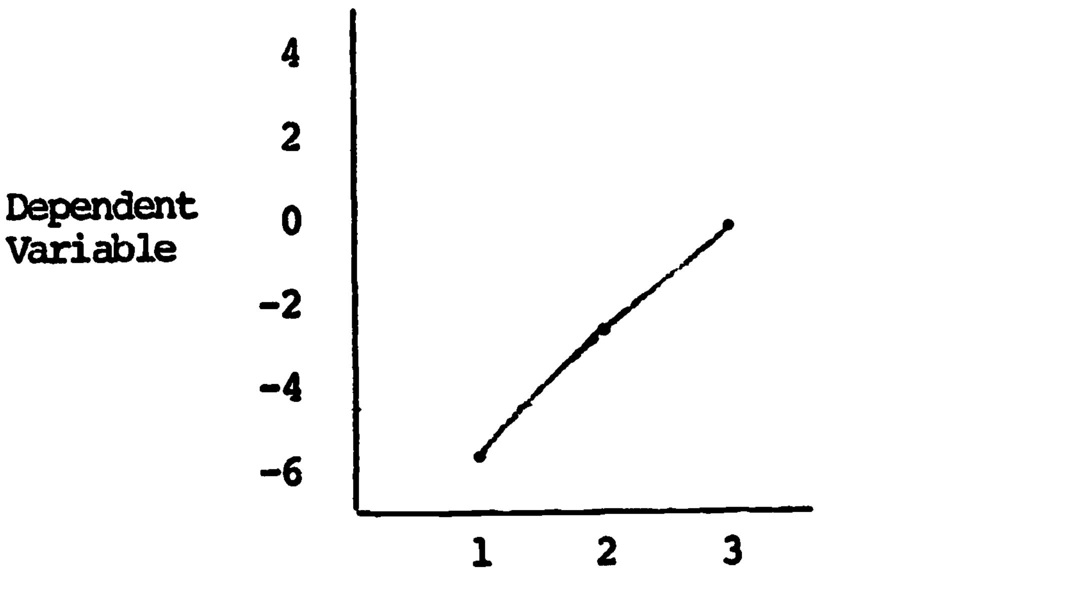
Thus, if we wished to display the estimated effects in our artificial example, we would recall that these contrasts amount to setting the last class in each way of classification to zero, and we would depict the effects for each class separately as shown in Figures 1 and 2.

Figure 1



A-class Effects (Arbitrary Origin)

Figure 2



B-class Effects (Arbitrary Origin)

The points plotted in figures 1 and 2 are entirely plausible (the connecting lines are intended to guide the eye and do not imply functional relationship), but the scale is peculier because it is not in the range of the original data. We need to change the arbitrary origin of the scale to some other point. A natural convention would be to choose the origin so that the effects appear as they would in an orthogonal analysis where the effects are estimated by the marginal means. A general method of estimating such means is, first, to estimate the subclass means from the fitted model, i.e.,

$$\hat{\underline{y}} = K\hat{\underline{\Theta}} \quad ,$$

and, second, to calculate the marginal means from the fitted cell means.

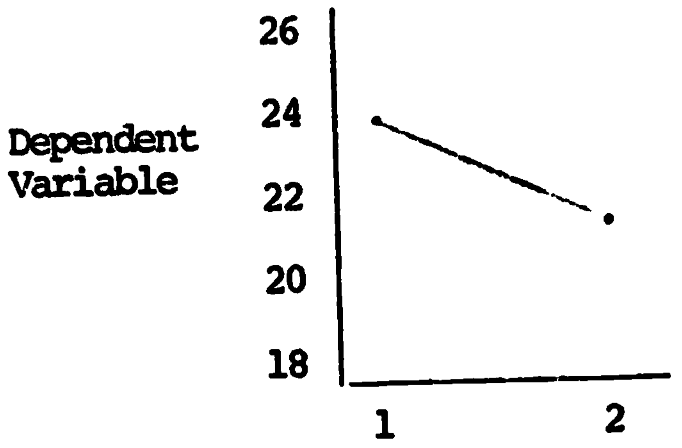
For the example, the matrix K is given on page 126 and the estimated l.p.f. on page 127. The reproduced cell means are

$$\hat{y}_{.11} = 21.931$$
$$\hat{y}_{.12} = 23.768$$
$$\hat{y}_{.13} = 27.477$$
$$\hat{y}_{.21} = 19.054$$
$$\hat{y}_{.22} = 20.891$$
$$\hat{y}_{.23} = 24.600$$

The marginal means for A-classes 1 and 2 are 24.588 and 21.515; for B-classes 1, 2, and 3 the means are 20.493, 22.329 and 26.039. The graphs on this scale appear in Figures 3 and 4.

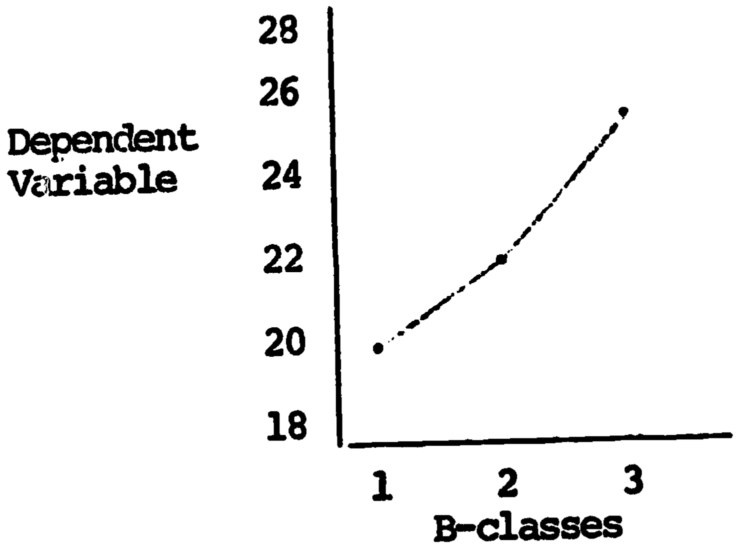
It must be understood, however, that only the differences between the points in Figures 3 and 4 actually have meaning. The numerical values of the points do not estimate means of any naturally existing population. They estimate means for a hypothetical population in which each subclass is equally represented.

Figure 3



A-class Effects
(Natural Origin)

Figure 4



B-class Effects
(Natural Origin)

Interactions

In orthogonal analysis of variance, a widely used aid to interpreting interactions is a graph of the marginal or cell means for the interacting ways of classification. This device can also be adapted to non-orthogonal analysis. The procedure is to fit a model including the significant interactions, and to reproduce the cell means or marginal means with the fitted model.

If the highest order interaction in the design is significant, the rank of the model fitted must equal the number of non-vacant cells in the design, and the cell means can then be fitted exactly. In other words, the cell means in this case are their own best estimates. Thus, when all interactions are significant, it is the cell means which are to be plotted in both orthogonal and non-orthogonal analyses.

In multiway designs, on the other hand, it will frequently happen that a low-order interaction is significant while higher-order interactions are not. In this case, a model of which the rank is equal to (1 + d.f. for main effects + d.f. for significant interactions) should be fitted and the best estimates of the cell means calculated. The marginal means can then be calculated for the interacting dimensions and plotted.

For an illustration of these calculations, let us extend our artificial example by adding a third way of classification. In terms of subclass means and standard deviations, the data might appear as in Table 5.

Table 5

Artificial data for a 2x2x3 design

Subclass	Classification A	B	C	Subclass Number	Observed Mean	Standard Deviation	Estimated Mean
1	1	1	1	10	22.1	3.125	22.43
2	1	1	2	9	23.5	2.986	23.59
3	1	1	3	8	29.8	3.271	28.90
4	1	2	1	7	20.0	3.111	20.27
5	1	2	2	10	21.5	2.895	21.58
6	1	2	3	10	22.9	3.175	22.94
7	2	1	1	6	24.7	3.250	24.15
8	2	1	2	9	25.4	3.011	25.31
9	2	1	3	10	29.9	2.943	30.62
10	2	2	1	9	22.2	2.751	21.90
11	2	2	2	8	23.4	3.167	23.30
12	2	2	3	10	24.7	3.112	24.66

An analysis of variance of these data appears in Table 6. We see an indication of a significant B x C interaction, but no evidence whatsoever of other interactions. Since there is also evidence of a significant A effect, it appears that the simplest model capable of describing the data is of rank 7 and may be expressed in terms of the constant term, the four main-effect contrasts and the two B x C interactive contrasts. If this model is fitted to the data and the subclass means estimated by $K\hat{\underline{\Theta}}$, the figures shown in the right hand column of Table 5 are obtained. From these estimates, the various marginal means shown in Table 7 are calculated. The marginal means are simple, unweighted averages of the estimated subclass means. In effect, they predict what the investigator would have obtained for marginal means had he been able to obtain equal numbers of observations in the subclasses.

Table 6

Analysis of variance for 2x2x3 design

Source	d.f.	Mean Square	F	p
Constant term	1	62053.6411		
C, ignoring A, B, BC AC, AB, and ABC	2	194.1985	29.75	.0000
A, eliminating C and ignoring B, BC, AC, AB and ABC	1	81.2782	8.67	.0040
B, eliminating C and A, and ignoring BC, AC, AB and ABC	1	318.7697	34.0679	.0000
BC, eliminating C, A and B and ignoring AC, AB and ABC	2	45.7028	4.8844	.0096
AC, eliminating C, A, B and BC, and ignoring AB and ABC	2	4.3448	.4643	.6300
AB, eliminating C, A, B, BC and AC, and ignoring ABC	1	1.6044	.1715	.6798
ABC, eliminating C, A, B, BC, AC and AB	2	2.7515	.2941	.7459
Within subclasses	94	9.3569		

Table 7

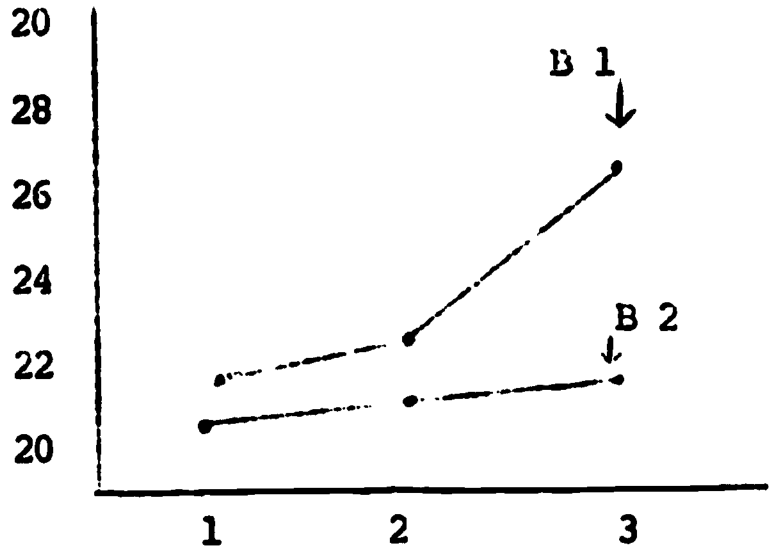
Subclass and marginal means predicted by the rank 7 model

		A		
		1	2	
B	1	24.97	26.69	25.83
	2	21.60	23.32	22.46
		23.28	25.01	24.15

		A		
		1	2	
C	1	21.33	23.07	22.20
	2	22.59	24.31	23.45
	3	25.92	27.64	26.78
		23.28	25.01	24.14

		B		
		1	2	
C	1	23.78	21.13	22.20
	2	24.45	22.44	23.45
	3	29.76	23.80	26.78
		25.83	22.46	24.14

Figure 5



C-classes
Interactive effects (natural origin)

We are primarily interested in the B x C table, which contains information about the significant interaction. Plotting in Figure 5 the entries in the body of the B x C table, we see that the interaction may be attributed to an excessive response in class B1 under C3. No evidence of specific B-class effects under C1 and C2 is evident.

I hope these examples serve to make clear my present views on the question of whether we should take a "pure" regression approach to the analysis of comparative studies by means of the linear model, or whether we should retain the approaches and terminology that have grown up in the application of analysis of variance to designed experiments. Since the two approaches are formally identical and lead to the same result if properly carried out, the question becomes a matter of precedent, convenience, and taste. Precedence certainly favors the analysis of variance formulation where designed studies are concerned. It is the only treatment which appears in the widely used statistical texts. The fact that the theory of experimental design is formulated in analysis of variance terms is also important here. In terms of convenience to the user, the analysis of variance approach, which uses the subclass means as the summary form of the data and deals with effects in terms of class effects and contrasts among class effects, seems easier to apply and interpret. In particular, construction of the bases matrix K is much easier than the construction of the deficient rank matrix for the original model. Furthermore, as I mentioned in my remarks to Professor Winer, if we extend analysis of variance to the multivariate case, we have a convenient method of handling the mixed-model analysis. This analysis is virtually impossible by the direct regression method if the random dimensions have many classes.

As for taste, I can only agree with Dr. Ward that it seems to be accounted for largely by what one has been taught and thus varies dramatically from one university to another. This is not the most satisfactory state of affairs. Clearly, it is our responsibility in teaching, not to treat multiple regression and analysis of variance as if they were unrelated topics. There is really no excuse for doing so when an integrated account of these subjects is available in Professor Graybill's excellent text, *Introduction to Linear Statistical Models, Volume I*. On this optimistic note let me thank you for your attention and end my remarks here.

DISCUSSION OF SPECIFIC PROBLEMS
RELATED TO THE USE OF THE GENERAL LINEAR MODEL

Discussion of Specific Problems
Related to the Use of the General Linear Model

Dr. Findley:

As a basis for significant discussion, a series of questions and illustrative problems is offered as a point of departure. The first question is

1. Can designs in which one or more cells contain no data be analyzed by the general linear model? Are there rules of thumb that one could apply regarding the number or pattern of missing cells that could be allowed in an analysis? Should one use formulas for estimating missing data from row and column means?

We are all familiar with the practical rules of thumb given in basic statistical texts on chi-square as to what you may do. I recall that in a text I have used about five very specific guide rules on the use of chi-square were listed. I think what we are asking is whether we can pass on to the educational research community similar guidelines on linear models, though perhaps not of such specificity.

Dr. Graybill:

The answer to the first question is "yes." The general linear model with missing cells is of no consequence except that it makes the computational problem more difficult. As to using formulas for estimating missing data, the only advantage of such a formula is computational. Let us say that you have a two-way design with one observation in all cells but one. In that case, it is easier to go through a missing data procedure than it is to invert the matrix or solve the system of equations and so I would use that. However, if the thing gets more complicated, then perhaps it would be easier to go ahead and do a regular least squares analysis. But, by and large, the missing data estimations are just for computational ease and the data can be analyzed with general linear models whether cells are missing or not.

While I am talking about this, let me add that I am not clear what is meant here in this conference by disproportionate subclass numbers. Does this mean that if they are proportionate in the analysis, it is easier, or something like that? I have heard this referred to a number of times. But if you perform an analysis and the subclasses are proportional, but not equal, and you perform an analysis by proportional subclasses, you have weighted your effects, so you need to be very careful not to do a proportional subclass analysis if you do not want to weight means for the subclass numbers. So I am not sure what is meant here. The fact that proportionality gives orthogonality is completely by the way. You should make the analysis that is meaningful to you relative to the cell means and if you do it by proportional analysis, if they are not equal, but proportional, you weight the cell means. So if you do not have a survey, you may not get what you want. I think you should be very careful about proportional analysis. I do not know what you mean, but I have heard talk of disproportionality. If there is something easier about proportionality, that is nice, but you may have the wrong analysis.

Dr. Bargmann:

I agree with Dr. Graybill and would underline even more strongly that the least squares analysis of the two-way classification, even irregular, even disproportionate, even with missing cells, is very easy and I am sure all my students will agree. If you use adjusted normal equations, there are only very small equations you have to solve. This is a lot easier than using Snedecor approximations for other things. As for the question, specifically, how many empty cells should be or could be tolerated before one would discard this whole model, there is, of course, a very simple indicator: if the degrees of freedom for interaction become zero, if you have missing cells, your subtotal degrees of freedom always get reduced by one for each missing cell and you come to a point where there is no more room for interaction in the degrees of freedom. At that stage, I would say consider the reformulation of your model and assumptions, because otherwise confounding or aliasing is going on in the main effects which makes interpretation extremely hazardous.

Dr. Wiley:

You might point out that Dr. Graybill's comment is appropriate here, because missing cells are a form of disproportionality. The pattern of missing cells may be as important as particular effects, so that it may be important which cells are missing. Some rules of thumb are found in Elston and Bush (1964). There are some problems related to that article that have not been brought up here which I think are very important and need to be discussed. One is relative to the models you are comparing in the nonorthogonal analysis of variance and relative to the fact that, when you are talking about testing a certain main effect, you have to be very specific because there are many tests of the very same main effect depending on which models you are comparing. I believe Elston and Bush were concerned with testing a main effect when there is an interaction in the model and I believe their test of a main effect was a test eliminating both of the other main effects and the interaction. The point that would have to be considered here is what tests of main effects are appropriate in the situation where you are allowing for interaction. Is it proper generally to test your main effects, ignoring the interactions after you have tested them and found them to be null, or should you do what some people say and get so-called clean tests of main effects where you are eliminating interaction? I think this is a question that has not been brought up and is a very important one that should be discussed.

Dr. Findley:

It might come up in discussing the specific problem that is mentioned next.

Dr. Cramer:

Could I make a comment on this question? I think Dr. Bargmann's point was a good one in that when the number of missing cells is quite large, you have a high degree of confounding in your estimation of main effects versus interaction. One kind of thing that can be done that is quite feasible given the modern computation equipment is to actually calculate alias matrices where you show exactly how your estimates are aliased by other effects.

Dr. Bargmann:

Not quite. Main effects are always aliased with interaction, but if the number of missing cells gets too large, then there is no degree of freedom left and main effects start to be aliased with other main effects.

Dr. McLean:

A latin square design is a complete design with a lot of empty cells. The main effects are completely confounded with certain interactions.

Dr. Wiley:

I wonder if one should use formulas for estimating missing data from row and column means? That also involves the question of the way you are going to test your effects. For if you use the exact formula for estimating missing data, you are automatically saying that what you are going to do is test main effects, eliminating your interaction. So this is really tied in with the more general problem.

Dr. Cohen:

It happens that the problem, particularly in the educational context, arises frequently in surveys. Probably the most significant part of such data is the correlation between the two factors in a two-way design implied by the pattern of missing observations. We sometimes lose sight of the fact that this means that the main effects are correlated. If one effect is educational level and the other is income, trying to get estimates of means gets to be relatively meaningless. The first point is that these two factors are themselves correlated. Indeed they relate to the independent variable, but not only do each of them do so in an overlapped way, so does their interaction. Rather than be so concerned about the question of how one makes estimates of marginal mean differences in circumstances like this, I think a more meaningful procedure is to understand that you are not dealing with an experiment -- that the nonorthogonality or the correlation among your factors is a real phenomenon that needs to be incorporated into the analysis of the data.

Dr. Graybill:

I would like to say "Amen" to that and, in fact, I think we get hidebound by this $\upsilon + \alpha_i + \beta_j$. This is what the experimenter in agricultural experiments has invented. This is wrong. We should look at the cell means. We can estimate the cell means and, if we have repeated observations, we can use something like the studentized maximum modulus and put multiple confidence intervals on these. From there do what you want to do--do not be hidebound by some pre-conveived model somebody twenty years ago invented. I think this is a very serious mistake. I would not let the statisticians shove something down my throat. I would do exactly what Dr. Cohen said.

Dr. Bargmann:

For guidance on this point, the researcher should be referred to the more standard textbooks. If a survey contains missing cells or highly disproportionate or irregular entries, and if this is a reflection of the proportion of such combinations in the population, why not simply subject the data to a contingency table test which can be found in any textbook and from such a contingency table test infer what kind of association exists between these two principles of classification?

Dr. Bock:

I want to disagree with Dr. Graybill's comment on cell means. I think that the whole point of analysis is to try to see if something that looks complicated can be explained in a simpler manner. If you have ten thousand cells in the design, eventually your model is going to contain these ten thousand cell means.

Dr. Graybill:

What if your problem is such that means for rows and columns really do not mean anything--but there is a diagonal that means something? If you do not look at the cell means, you really are not analyzing your data. Now, I am not saying you should not look at row and column means. These may be important, but they are not sacred.

New Speaker:

That would show up in your interaction and then you would begin to think, "Obviouslv this is not a simple additive situation with nothing else going on. Let us try to find ..." Somewhere you want a model, I think, and it ought to have fewer parameters than cells.

Dr. Graybill:

You see, that is part of the problem in data analysis. If you know a great deal about your subject, you do model it and estimate it, but in initial research stages it seems to me you have data in search of a model.

New Speaker:

That is why you need a preliminary work specification for which the analysis of variance can be quite helpful.

Dr. Cohen:

Let me explain my notion with an example. Suppose for some dependent variable you have an independent variable A that can be measured continuously or nominally or whatever, and you have another independent variable B. Now, if A and B are things like education and age, you know very well that the overlap between the two of them is not similar to the disproportionality of cell frequencies that happens in a laboratory when some fool technician drops a tray of test tubes. It means something. Education and age are related to each other; they both overlap the criterion independently; each accounts for portions of the variance, but there is in addition some overlapping area for which they both account. Now, it is purely a matter of your theory whether you are interested in how much of the criterion is accounted for by age added to education, or how much of the criterion they jointly account for, or whether the joint contribution is to be split between them in some fashion. All of this has to do with what we call models, but in any case, it should be an expression of what you as a researcher have in mind.

Dr. Findley:

Well, I think we have explored the first question rather well and I hope we can do that much with other topics. I put here second a rather straight-forward type of study with which I am familiar. It is representative of many others that have been made at many other institutions.

> In studies of "native" vs. transfer students for evidence of academic success later in college, it is common to use all the data of a given period. Native students are compared with transfers from other institutions: state system junior colleges, other state system four-year colleges, outside-of-system colleges, and so on. Sex is also ordinarily an independent variable. Scores on common entrance tests and similarly computed high school averages are continuous variables. How appropriate is the use of a linear model approach like Harvey's "Least Squares Analysis" (Harvey 1960)?

That is the one used in the study I recall. If so, should all interactions be automatically tried first? What overlap on the continuous variables of test scores and high school averages should be required, if any?

Dr. Bargmann:

I am not familiar with that particular computer program that USDA uses. My answer is that this is a straight-forward two-way analysis with covariates. In fact, one factor is sex, the other factor is the school system from which the subjects came, and the continuous covariates or concomitant variables are the entrance test scores. The computer program with which I am familiar is one from the University of Illinois which we have on the 360 at Georgia. This and other programs certainly handle this case, a special case of the linear analysis. Incidentally, it is not a matter of setting up the entire design or model matrix and then running through, who knows, 50 by 50 inversions. It resolves to perhaps a 2 by 2 inversion. In this particular case it is one equation and one unknown - because one of these factors has only one degree of freedom. So there is not even inversion involved. It is that simple.

The other question was whether interactions should be automatically tried first. I would say that most programs do it this way. Certainly, the total effect (all effects combined) should be tried first, because if an F-ratio shows no significance with all effects combined, then there is no need to break the data down into components. Beyond that, an interaction test is always quite useful because it may indicate to you many things.

Of course, the most important thing, let me again concur with Dr. Graybill -- and our computer programs certainly do it -- is to state the cell means. State the unadjusted cell means and state the cell means adjusted for the continuous covariaties. But after that, if you find significant interaction, you may first of all consider whether there may be some outliers in some cells. You can detect them right away. You eliminate them and then look at the cell means to see if there is some kind of a trend or not. Again, the sum of squares for interaction can be explained. It is for this reason that I think the sum of squares for interaction is an important indicator to have. It will tell you at once if there are outliers or not or if there is still a lot of trend in the data or not. Eliminate that if you can. If there is still something left (e.g., non-linearity) it will tell if we should make a transformation. Yes, I would say this little number, sum of squares for interaction, is very useful.

Dr. Anderson:

I find Dr. Bargmann's comments very interesting. Such a testing of the overall cell variation might well lead to a Type II error. It seems to run contrary to the present day tendency to go ahead and run specific contrasts or perhaps orthogonal contrasts in spite of the fact the the overall F on cell variation may be insignificant.

Dr. Cohen:

Certainly one could partition the cells in a design like this into three major families like row, column, and interaction, rather than use an overall test in which one subset might be washing out another subset of logically distinct pieces.

New Speaker:

It would wash it out of the estimate, but it certainly would not wash it out of the statistical testing. That is unaffected.

Dr. Bock:

If you have any prior information that leads you to think that if there is interaction it is going to be a simple one--linear x linear--you ought to look at it.

Dr. Findley:

Well, the question asked here is, since you have all of these variables in the picture, all the different colleges from which the subjects come, sex, and the continuous variables, whether you should as a matter of rule deal with all of the interactions first before proceeding further. Or should I take your last comment, Dr. Bock, to suggest that only if you suspect interaction you should check it? That I certainly would. I am asking whether I should cheek interactions even if I do not suspect any.

Dr. Winer:

I have heard of this approach proposed under the guise of "cleaning up the model." You build everything by inference and you have no real a priori guide. In going through a series of F-tests with no real systematic procedure, you arrive at something by essentially a trial and error process. This aspect of the complete regression model I abhor. I think in handling this kind of problem one should be guided by the natural classifications, those which are meaningful a priori. Either by tests of hypotheses, or by actually inspecting the sources of variation, identify the relative sources of variation, then build the model from those sources of variation which are identifiable, which can be expressed. Now, any source of variation with more than one degree of freedom can be broken up in several different ways. Parsimony, if this is the only thing you can fall back upon, is a good guide. But if there is some premise as to the nature of the underlying relationships, by all means use it rather than the polynomial. Polynomials should be used only as a last resort--particularly one involving anything above a second degree term. This is, again, an appeal to parsimony.

As I tried to emphasize yesterday, a good guide is to break down the total variation into orthogonal components or orthogonal sets of components. However, this may not be necessary at all. And then for prediction purposes, yes, formulate perhaps a regression model which included linear terms or non-linear terms as the need may be.

Dr. Wiley:

There can be real danger here. If one has a rather highly cross-classified design, say a two-to-the-tenth factorial arrangement, then the likelihood is quite high that some of the contrasts are automatically going to appear to have low probability in terms of their size. In fact it's quite likely that when you have considered a complete model and tested each of the contrasts for all the main effects and interactions in a two-to-the-tenth factorial design individually, you will find about 50 of them significant at the .05 level.

New Speaker:

It seems to me highly important, especially when the design has a large number of cells, to engage in some kind of overall test to make sure that you are not simply dealing with complete error.

Dr. Graybill:

I do not know enough about this particular problem to know whether you would want to do this initially but it seems to me that finally one ought to use multiple procedures. I do not like to say "multiple decision," but it is sometimes used in cases where you can snip your data, take all your comparisons, and look at them. You know exactly what your protection levels are, so you can look at the data in terms of confidence intervals and not as tests.

For example, if a confidence interval misses zero by just a very small amount but is in a very narrow interval, it is practically insignificant even though a statistical test may indicate significance at a very high level. That's the main reason why I do not trust tests of significance. I trust them more than I do tests of hypotheses, but I don't trust either of them.

A good experimenter, it appears to me, wants to look at his data, to know what overall protection he enjoys, to know the confidence interval of the estimates, to put these together, and to begin to read contours and stories out of the data, realizing that he has a finite set of data.

Statisticians have put over tests of hypotheses on the data-analyzing public for the reason that mathematically they are much easier to teach. If we take the approach I have just suggested, we do not have to teach separately the concepts of confidence intervals, tests of hypotheses, and tests of significance. We can teach one thing. Neyman wrote a book on this. Nobody in data analysis paid attention to Neyman's book. But one who studies statistics should because this is the economical way to study data. I hope some time before we end this symposium we can have a very heated debate on tests of significance. I think it is extremely important that we data analysts not let tests keep us from doing what we want to do. But I would also look at the ove^rall procedures to give you an overall protection in dealing with such problems.

Dr. King:

I would like to agree with one thing Dr. Graybill says. We have not looked very much here at the multiple correlations that may be derived from our data. That is a very useful kind of information to know--what proportion of the dependent variable is being accounted for by the others. As a matter of fact, Dr. Ward's early work was to test multiple correlations and I always liked that approach.

Dr. Bock:

If you are free to choose your subjects, you can make these correlations practically anything. You can leave out the middle group which is very large. A correlation only has meaning in reference to a population.

Dr. Findley:

We see that any question that starts from a specific example can lead in all kinds of directions. As I indicated, this is a typical study in which you take all the cases that are there, all the students who transferred during a two year period and the ones who had been enrolled all along during that same period so we are not talking about doing what we want to with the data. We are asking what we may properly do with the data that has more or less defined itself to us.

Dr. McLean:

To a data analyst, the idea of looking at an array of cell means and pondering what forces are operating is attractive. But we are dealing with examples in which optimal cell means are very differently determined. Some cells will have many observations and others will have few. Now, how does the data analyst protect himself in cases like this where he is trying to compare means with unequal precision?

Dr. Bargmann:

He could put confidence bounds on each cell as a first guide to protecting himself against over-interpreting. For example, if he has a cell mean based on a very small sub-class, he finds a very wide confidence region that spans many of the means of the larger, better defined sub-classes.

Dr. Anderson:

Would you say something like a Duncan's multiple range test might be appropriate?

Dr. Bargmann:

That is certainly not indicated for interpreting the cell means. A multiple range test would involve contradictions due to the disproportionality---some effects that are way out are not significant whereas smaller ones are, so a multiple range test at this point would not be indicated. I would rather take a least significant difference approach or look at every possible mean comparison and put a plus-minus on it.

Dr. Graybill:

Use the studentized maximum modulus.

New Speaker:

I think that you have presented an excellent example here because it brings missing cells that result in this particular situation from the fact that some students do not have high enough high school and other test scores to get into the university, so they go to the junior college first.

Dr. Findley:

That is exactly the case and that is the reason for the last question to which none of us have spoken directly.

New Speaker:

There are other problems. You are going to have different variability within your cells because if you accept students from a junior college you may accept only the A or perhaps B students. So their range of first two year scores is going to be very restricted while the group you are comparing them with, who were already in your college, will have practically a full range. There will be other groups of colleges where the students will have full ranges, so you have non-homogeneity of variance within cells.

Dr. Wiley:

Professor Bargmann's implicit model for that procedure does not have the complications that were indicated in that he was using covariates which were the entrance examination scores and the high school grade point average. Certainly there would be complete data in the cell design for the source-by-sex classification. The essential question would be something like, "Would you expect an interaction between high school grade point average or test scores and source of the student?" If you expected that interaction, you would probably set up a different model than the one Professor Bargmann proposed for this situation.

Dr. Findley:

Well, we do have the specific situation cited where the selective admissions at one institution are less severe than the other and one of the questions implied in this last question here is what we are to do if we get a regression line for one group and a parallel one for another group but they are over different ranges of the independent variable. Does that affect our ability to interpret?

Dr. Cramer:

Given a model you think fits, it does not affect your interpretation. That is, if you assume that there is no interaction, then you can certainly use that model. On the other hand, if the ranges are so disproportionate that there is minimal information about any common area in the regression line, there is no ability to test the parallelism of those lines.

Dr. Cohen:

Dr. Bargmann made reference to outliers a moment ago. Of course, they are quite troublesome, not only in situations of this sort, but in other cases. I wonder if he would say just a word about how he would proceed in that regard.

Dr. Bargmann:

There are just about as many tests for outliers as there are statisticians. Each has his own ideas and I think there is no standard objective definition as to what constitutes an outlier, so one's subjective definition determines what kind of a test he makes in order to determine if something is an outlier or not. In nice, two-way classification presentations of data, in each cell you can see the mean and the standard deviation, and look at the pattern of meaas. If you find a rather peculiar mean, first of all the computer says "I find a very high interaction effect," so I look at these peculiar means and at the same time I look at the corresponding standard deviations. Now if a deviate mean, one that does not fit into any trend, is also associated with a large standard deviation in the same cell, I think you have proof-positive that you should look at those raw data once more. This is as snoopy an indicator for outliers as I know and it is a fast one. Other techniques can be used.

Dr. Anderson:

You keep saying to look at the cell means and yet you tell us not to interpret chance.

Dr. Bargmann:

I look at the complete effects first; I mean I look at the F-test of all effects combined first. If that is nonsignificant, if that is very small, I have a perfectly good, plausible, parsimonious model. There is only a general effect. Why should I look for anything more complicated?

Dr. Anderson:

What you are saying, then, is if you do an overall test and it is insignificant, you cannot do something like Duncan's multiple range.

Dr. Findley:

The third question is one submitted by Dr. King:

"The following problem was encountered by Dr. Garrett Foster. In a general sense, it is the problem of measuring change (pre-post design) on the dependent variable as a function of the interaction of two independent variables. This is often done by putting the pretest data in as a predictor and testing that contribution of the interaction of the independent variables which is, in fact, independent of the pretest vector. The problem arises when one finds that the estimated regression weights for the pretest data vary (interact) with one of the independent variables (e.g., there are significantly different regression weights for the pretest by group product vectors). I have worked out several possible solutions, such as testing the triple order interaction among the predictors (e.g., pretest, SCAT, and school) and plotting the results when significant.

Dr. Bargmann:

I think the answer is extremely simple. The question is not about a univariate analysis, but it is a clear case of the multivariate analysis model. I think we should try to point out that in a mathematical model we say "y is equal to the function of x or several x's." We say "the right hand side is independent and the left hand side is dependent" and this defines it. In a statistical model where we say "expected value of y is a function of x's," we do not have the subdivision into independent and dependent sets of variables. In fact, we assume in the univariate model that the right hand side variables are known without error or at least knowable without error. They are prespecified. They are part of the design. They are concomitant variables. It is on the left side that we have the random variable. Now, in multivariate analysis we have precisely the same model except that on the left hand side we have, in this case, two random variables, the pretest and the posttest variables. One of them could be dependent; the other one could be independent. I can take the left hand side in a multivariate model and split it into dependent and independent variables. The analysis tools for this are quite well known and already available to the practicing statisticians and applied statisticians. Morrison (1967) has described these methods quite fully. The situation described is multivariate. There is a very great number of tests, confidence intervals, and statements you can make if you view it appropriately. If you confuse the idea of a concomitant variable and a random variable--Dr. Bock was hinting at this--you can prove anything you please. We must not confuse these two. This is a case requiring multivariate analysis and not univariate.

Dr. Wiley:

There is an interesting point here. You can see this situation as an experimental situation where you have random assignments, say in the independent variables of interest other than the pretest, or design factors by which the subjects are randomly assigned to the groups. If so, the circumstance can not possibly occur, so in one sense it has to be a pseudo situation in a natural situation treating the covariate or the pretest variable as an independent variable. But the real implication of this is that the pretest score is being affected in some way by the other independent variables in the design. That can not possibly happen in an experimental situation. So that is why no one would run across this problem.

New Speaker:

I am not sure that I understand this. I wonder if one might have a situation in which we could have two treatment groups and random assignment to them, ana a pre-measure. Now, does the question here relate to whether the regression coefficients are the same in the two groups?

New Speaker:

The groups will not differ in their composition because of random assignment. They will not differ in their composition on the pretest measure, at least in expectation.

New Speaker:

You should no more expect them to do that than to differ in mean on the pretest measure if randomly assigned.

New Speaker:

But could not the effect of the treatment be to change the regression coefficient?

Dr. Bock:

That would just be giving the test of means then because the groups would start out at the same initial point on the average.

Dr. King:

Well, obviously this is not an experiment. This arose from an attempt to evaluate school program and the subjects were randomly selected within the schools but obviously could not be assigned to schools randomly. So it would be possible in this case for the pretest to interact in the manner you speak of, but it is not an experiment.

Dr. Findley:

Let us turn to another question that may be the same problem but stated just a bit differently.

> "In a study, two sets of intact classes are taught by three teachers. The classes are not matched, but students were assigned in essentially alternate fashion to classes meeting at the same hour. Shifts of classes after registration and some attrition account for different subclass numbers. Can inferential statistics be properly applied here? What differences in numbers would give pause to comparisons?"

There is another element to this question, but isn't this the kind of situation you have in mind where you actually have these people in intact classes?

New Speaker:

"In alternate fashion" doesn't imply a random sample.

Dr. Findley:

Well, I think so. If you send the first registering to class number one, the second registering to number six, and the third to number nine, etc.

New Speaker:

Well, they arrive in random order, and therefore, they are assigned in random positions.

New Speaker:

You are dealing with an experimental situation; however, one can point out that in this case each treatment is applied to each class as a whole and the class as a whole is the sampling unit.

New Speaker:

The classes could be regarded as blocks in this design and we have a very irregular treatment-block situation. We can now regard these block effects--class effects--as fixed ones or we can say they represent a random sample from many more classes and treat them as random effects.

Dr. Graybill:

You need to be very careful about using the within-class variance as error, as I think Dr. Bock pointed out, and this generally would underestimate error if you use it.

New Speaker:

Use treatment-block interaction.

New Speaker:

The means would be the adequate statistics and classes would all be pretty much the same size.

Dr. Findley:

I did not mean to draw us away from the other question, but it does seem to me that we are beginning to verge on some of what was in the other question. Is there any further discussion of this point?

Dr. King:

There is one thing I would like to ask. Possibly this could be directed to Dr. Bargmann. This he says is a multivariate case because the pretest is not known without error. I can certainly see that this is true, but is it not true very often that we use independent variables such as I.Q. and so forth? These we do not know without error, so if we use that criterion, are we illegitimate very often?

Dr. Bargmann:

There is a very simple distinction. If you use the pretest in order to select your subjects, for example, if you use the pretest in such a way as to take five subjects with this score, five subjects with this score, five subjects with this score, you make it a regressional, univariate problem because the very fact of selecting on the score makes it a concomitant variable. All regression that we are talking about is a conditional expectation. We ask what is the expected value of y in the posttest score given x in the pretest score. Now, this is the case as soon as you start assigning by the actual pretest score.

This is, in fact, a univariate study, the pretest score is a concomitant variable. On the other hand, if you did not use the pretest to select students, but merely obtained pretest measures on all subjects, then both tests are random variables. For every experimental unit, for every student, you observe randomly those two scores.

It may be of interest that the simplest of all well-known multivariate tests is the t-test. The t-test, when you come to look at it closely, is a multivariate test. You have two random variables, you take the differences. In this case, it reduces to a univariate problem. In this sense then, the inference of what happens between pretest and posttest is either made by a regression or prediction equation. This presupposes that you have actually selected students on their pretests or you make your inference in terms of a correlation between pre- and posttest, or shall we say confidence bounds with certain differences or weighted differences of these two. The crucial thing is whether or not you select on this variable. If you do, then it certainly is a concomitant variable. If you do not, then it is a random variable and multivariate procedures should be advised.

Dr. Ward:

Is it possible that multivariate procedures are too restrictive, so that you may be getting far away from the problem at hand?

Dr. Bargmann:

Why should they be more restrictive? Actually they are less restrictive. They contain more information than the univariate procedures. They take into account all of the relationships and overlaps. When you make confidence statements in multivariate cases, you do not simply make confidence statements on each variable separately, but you determine joint confidence regions and talk about the joint probability of the set of parameters having certain values. I would say multivariate analysis is not more restrictive but less restrictive. It takes a little practice, I assure you, to make sense of it and it usually takes two pages of relevant comments to explain one or two confidence intervals that a computer puts out.

Dr. Ward:

In a multivariate case where you think product or squared terms are relevant, how does that affect the assumptions that are involved in multivariate analyses? Suppose you actually did not collect a large number of variables but it was closer to your thinking about the problem to generate squares and products of a few variables that do not yield multivariate normal joint distributions?

Dr. Bargmann:

The interpretation will probably become so messy that I would have to program a computer to interpret it.

Dr. Bock:

Well, you might concentrate on what kind of transformation you might make on the variables.

Dr. Bargmann:

Transformations very frequently reduce matters to a single point. Then I don't know what this point means. A log transformation that all of a sudden matches your data to a single point--I'm lost.

Dr. Bock:

But there may be some useful transformation with the independent variables.

Dr. Winer:

I think one needs a balance between trying to stay close to the problem at hand and looking at other situations, then deciding upon which procedure is appropriate. As was mentioned yesterday, an exact solution to the wrong problem is worse than an approximate solution to the problem you are interested in.

Dr. Findley:

May we turn to question 4 submitted by Dr. Jennings?

I think most of the elementary texts typically used in an educational statistics course do a very good job of giving the student both the computational tools and an intuitive understanding of the meaning of the comparison between two means. As the designs become more complicated, however, it seems to me that the student is simply asked to accept the fact that a particular computational procedure produces a good number called "the main effect" or a "linear component" without much guidance as to the inferences one might draw from the presence of such an "effect." The student is thus encouraged to learn the computational procedure and the "names" of his questions without knowing in many cases what the questions actually are.

"Let me give you an example. It is not uncommon for texts that deal with so-called trend analysis to separate the total sum of squares into an error term, a term called "deviations from linear regression," and a term called the "linear component." Frequently recourse to a table of orthogonal polynomials is required. In no treatment I have seen are the basic model and the restrictions on the parameters identified. The inference I draw from the text is that a significant "deviation" implies that the means do not lie on a straight line and that the presence of a significant "linear component" implies that the means do lie on a straight line, although I have never seen that stated in so many words. What inference is to be drawn when they are both significant? It seems to me that if a researcher is encouraged to formulate his questions in terms of parameter restrictions, he defines operationally, by means of the restrictions he imposes, what he means by a "linear component," and the problem just never comes up unless he had a question to go along with it."

Dr. Cohen:

The significance of linear components I have always understood simply means that the best fitting line is not horizontal. A function can be anything you like but if you set a straight line to it, the rejection of the hypothesis of linearity means simply rejection of the hypothesis that the best fitting line is flat, so both can be simultaneously true.

Dr. Wiley:

If you insist upon fitting a polynomial model, and if you get significant departure from linearity, you cannot consider any component individually. You have to consider just the curve that is the best fit curve. Consequently, if you get departure from linearity and you insist upon going on, then you fit more terms. The usual procedure is to fit terms to account for the curvelinearity or whatever it might be. Then you would want to plot the predicted response from whatever level the analysis indicated.

Dr. Findley:

Further possible comments here?

Dr. Bock:

Well, regarding the general question of curve fitting, I did not quite understand Dr. Winer yesterday. I understood what he said, but the tables offered suggested two routes that yielded identical results--either the orthogonal polynomials fit or one must compute the regression values. His tone seemed to imply that the orthogonal polynomial procedures were in some sense better, although I could not see why. Then earlier, Joe Ward presented the five ages by 15 practice sessions data matrix in which the analysis that he proceeded to use implicitly involved an effort to fit all 75 of those cell means so that in effect 74 degrees of freedom were being used for this purpose when it seems to me no reasonable model about this would suggest the loss of more than at most six degrees of freedom. Let us say a linear, quadratic, and cubic term for each of the two variables, and indeed you could even cheaply use nine more degrees of freedom for the interactions of these and get a very complicated surface. By going the route of taking out all means, writing a model at the level of cell means in this situation, he was, in effect, from a practical point of view, giving up 74 degrees of freedom to account for this variation. It seems to me almost certain, given the nature of the data that were posited, that no more than nine or ten would be needed for snooping purposes to account for what was very likely going on in that entire complex surface. The problem can be handled entirely by a regression analysis that does not in any way depend upon either the equality of the intervals or the equality of the sample sizes. Set up terms like age, age-squared, age-cubed (which is probably more than you need) and then practice, practice-squared, practice-cubed, and then, if you wish, the vector product of these two for nine more terms. Ward's initial N obviously had to be enormous to have any cell replication. This approach uses N-15 degrees of freedom instead of N-74. Again, I find the regression to be not only simpler, but in many ways more powerful in at least the fact that it does not overfit, it keeps things as close to the data as you want them, and it is relevant to the nature of the variables that you are using. The design in the example that was offered would have been no different if the five-by-fifteen factors were purely nominal. But there was no attempt at all to take advantage of the fact that the data were in fact continuous, interval kinds of data.

Dr. Bargmann:

I think I can make this even stricter. You explained in very lucid terms what I think I hinted at yesterday. If you have an interval scale in the back of your mind, go ahead and do your multiple linear regression and curvilinear regression right away. May we perhaps dramatize the situation by pointing out that if your data or your levels are nominal, then by reordering and placing these levels on some X axis, you can always draw a perfect straight line through the means. This may give people some food for thought who would like to fit curves to levels, treating them as equally spaced or spaced in some known order even if they are really nominal for a certain length. Even in ordinal data you can do a lot of juggling and produce a practically straight line as all people do in bioassay, growth curves, and learning curves, a more meaningful approach than curvilinear regression.

Dr. Jennings:

My main question about this particular kind of formulation is that I wonder how many of us have actually tried to get a solution with these kinds of predictors? My experience has been that when you get beyond quadratics into cubes and interaction terms, the possibility of getting a very accurate solution is not good. Now, of course, this depends upon the computer program you are using and so forth.

Dr. Graybill:

Using the formulation that Dr. Bock has indicated, where one reparameterizes the model explicitly and develops orthogonal polynomials in a computer, one has absolutely no problems whatsoever.

Dr. Bock:

You always have problems, but the most accurate way to handle these problems is to orthogonalize the basis. There are some routines due to Householder which Cramer used in his program package. This is probably the best way for doing the orthogonalization and the subsequent least squares analysis even in single precision.

Dr. Cramer:

In Dr. Bargmann's situation, we would choose our metric. If we really knew what metric was necessary to fit a straight line, we could use it and reduce the number of parameters.

Dr. Ward:

I think it would be good for computing to realize that, if you had an orthogonal system that had to error in it, it would help to get from your original system to the orthogonal system. Then things would be better because you would just have to be sure to remember that the computing procedure involves going from the original set to the orthogonal set. As a matter of fact, some ways of getting to an orthogonal basis involve exactly the same procedures that are involved in solving simultaneous equations. You have to be careful since you may have exactly the same numerical inadequacy because you separate the two systems somewhat artificially.

Dr. Cramer:

I think that this is not so. One does not have the same inadequacies although it is true that you are doing the same thing as solving equations. There have been presentations of methods of solving a least squares problem using the orthogonalization procedures which are extremely well conditioned and one does not run into the kinds of problems you get from using the ordinary inversion procedures. So, if one works with orthogonalization methods, one is much better off in either situation.

Dr. Graybill:

If I understand this problem, the complaint is that Dr. Ward used a two-way classification rather than multiple regression. It seems to me that the multiple regression technique would be very good, but the two-way classification is also good. You know, the geologists do something that I think is very, very good. They take such a two-way classification and look at the cell means and then they contour the means. This is not very respectable statistically, but you get a lot more out of it. Some of you are suggesting fitting complex terms in a model. I would discourage this unless I had no other alternative. There is a great deal of merit in taking a row-column classification, looking at spikes of interaction in the cell means and contouring them. Both methods would be good. I would not throw out the row-column analysis. There is a great deal to be said for it.

New Speaker: No one is arguing that it should be thrown out. The only argument is that once you deal with the row-column presentation you should try to find a parsimonious model to describe the data.

Dr. Findley:

The next problem area is suggested by Dr. McLean. "Let's talk about more ways to check on the adequacy of the model, _e.g._, the examination of residuals." Maybe Dr. McLean would like to add to that brief statement before we enter into the discussion.

Dr. McLean:

Thank you. People have hinted or made comments from time to time in our discussions about checking on the adequacy of the model, _i e._, whether this method or that is a good way to test whether a model is good to explain the data. But these statements have been somewhat offhand. In particular, I do not think the residuals have been mentioned at all. I was just wondering why that should be so.

Dr. Graybill:

The residuals are an excellent method in my opinion to check the model. In fact, using cell means, you might look at interaction in each cell. You can even plot them on a half normal plot, for example. Many think the theory of residuals is open to criticism, but I think residuals are very valuable. In fact, I think it was Winer who said that what we really do is examine residuals all the time.

Dr. McLean:

I am concerned about something that is a little more uncommon. I am not speaking about the residual variance that is left after you take everything else out. The residual, I suggest, is the difference between the value the model predicts for the particular model with which you are working, and the actual values you obtain with various values of the independent variables. A separate calculation of these is not always done in the computation routine, and if it is, it is often summarized as the residual sum of squares used for the estimate of error. There is no option to print out the cell residuals.

Dr. Jennings:

A fairly recent book by Draper and Smith (1966) presents examples for discussion of how to treat or locate a sequence to evaluate the model.

Dr. Bock:

The Mesa A-5 program has as an option the computing of residuals about the cell means in any model that you may fit. These are presented in standardized form, that is, residuals are divided by the standard deviations, and presented as t statistics. This is extremely effective. If you have an interaction you want to try to figure out, you fit the model including everything else except that interaction. You might even include other interactions, but you omit the one that you want to interpret. You then look at the residuals and ususally you find a systematic trend and sign that shows you what is going on. As I said, residuals for interactions are in reality systematic.

Dr. Bargmann:

I am very glad indeed that the point has just been made that residuals and error terms differ. In our development of the general linear model, we regard this last term as error with expected value zero and common variance. Ideally this holds only if you can repeat the experiments under identical conditions. In many applications what goes into the error are merely high order interactions. I think even a study of the errors of replication would lend itself to this type of analysis for the same reason that when you say you are repeating the experiment under identical conditions you have to qualify your results by assuming identical conditions. Some condition may have changed and the residual may very well tell you what condition was affected.

Dr. Graybill:

Another thing that geologists do that is useful is to get the residuals of the several cells and then contour them. It is a very effective method to look for anomalies. You can use the half normal plot. Or you can use something like studentized maximum modulus even here and really get in and examine the residuals. I think this is very effective. It has not been played up enough. It is really just of recent origin that it has begun to appear in books, but some of you might want to look into it.

Another point, with regard to examining the model. If you take sufficient statistics to summarize the data, then whatever is left over is free of all parameters. Therefore, under the model given, they are very effective for examining the model. This is what I do when I use my sufficient statistics under the model that I postulate. I take what is left over from the sufficient statistics, which is free of all parameters if the model is true, and can do many different things to reexamine the model. Fisher has done this.

Dr. Winer:

I think this should also be said. In any specification of a model, no matter how complete or incomplete, any test of goodness --- there may be, for example, two tests of goodness which indicate that the model is fit equally well by both situations, but the pattern of residuals can be quite different -- provides supplementary information to look at to help decide which way to modify a model. Suppose we start with an incompletely specified model and eventually look at the residuals of the model. I do not think tests, in many cases, are sufficiently sensitive to tell us where to go next.

Dr. Wiley:

Let me just make one small point. That is very much a function of the design of the experiment. That is, if you have a model, you base the design on the model that you are postulating and you allow in the design of the experiment for specific components of lack of fit or a priori possibilities for the lack of fit in the model. Then you can have a fairly exclusive direction to go if the model in fact does not fit.

Dr. Bargmann:

I think the goodness-of-fit dilemma is due to the fact that we are attempting to reduce the quality of fit to a single number and to a single index. For example, in chi-square, when we want to have a little bit more information to use for pointers, we should have a few more indexes. We might, for example, consider doing the exact goodness-of-fit test, which is a multinomial. In the non-central case, of course, it would have as many parameters as there are classes. In that particular case, we would have several numbers, and a pattern of these. I do not quite know how to interpret them, but I can imagine that there would be information in the pattern of these numbers that would direct further modification.

Dr. Graybill:

I think this points up my objection to tests of significance. You summarize your data too far. You summarize it to one number. In the case of goodness-of-fit, the empirical distribution function is very easy to inspect--maybe a normal plot or something like it. It seems to me very important to be alert constantly and look at the data, all the data. Summarize as far as possible by sufficient statistics, but then when you amalgamate everything together in some way to look at one number, I think you lose too much. You are throwing away too much. Goodness-of-fit is a case in point. Even though I must say I use it, I do not like it.

Dr. Findley:

There is one small point on the last numbered question we distributed to which we did not speak specifically, although perhaps you gave me an answer to it:

"What differences in numbers would give pause to comparisons?" Is there anything over and above the very specific insistence that we use treatment blocks with regard to these variations in sizes of groups which come about, so far as we can tell, by operation of unrelated factors?

Dr. Bock:

If you are using group means of the statistics, their precision is going to differ in only a minor way due to differences in sample size. The samples you listed vary from 17 to 32.

Dr. Findley:

Suppose it were 17 to 149, how about it? Do we have any kinds of rules of thumb or helpful suggestions?

Dr. Bock:

If the difference in N is ten-fold, you might consider doing some type of weighted mean analysis.

Dr. Graybill:

I think it depends on what the variance is. For example, the coefficient of variation might play a role here. If the variance is very small, seventeen cases give you a very good estimate and your precision of the mean is very good. Of course, a sample of 100 is better, but 17 is sufficiently precise anyway. In fact, it makes little difference what the actual variance is then. If I am talking about a score that goes from 100 to 110 and my measurements are very precise, an N of five may be enough.

Dr. Findley:

I was not thinking so much in terms of additional sampling as I was of the situation in which you take natural groups. If the sample sizes differ from one to the other, is there some point at which you decide you had better in some way randomly sample any large group so as not to give it a disproportionate weight in what you are doing? How else do you deal with it?

Dr. Bock:

If you are using the class means, they are not weighted by the numbers of students in the classes. In fact, they are under-weighted a little insofar as efficiency is concerned, so there is no bias involved. There is only a question of efficiency.

Dr. McLean:

You might watch out for unequal variances, especially in a case like this. We do know that you are hurt worst if your assumption of equal variances is wrong in addition to having unequal cell sample sizes.

Dr. Graybill:

You should not use your within variance for your experimental error.

Take an extreme case. Suppose that every one of the 26 students in one class had the same grade. The variance would be zero, and you would not worry about unequal numbers. So if the variance is small, you do not need to worry about differences in sample sizes, since you are not going to use the within variance for error estimation.

Dr. McLean:

I probably should correct my earlier statement. The fear I would have is if the problem differed and you have a lot of classrooms in one treatment and only a few in another. That is the only situation to which my comment would apply.

New Speaker:

This is a basic problem in educational research. Is the classroom the sampling unit or are the students independent replications within classrooms? I believe the classroom is used as a unit because one thinks the students are correlated within replications and therefore their differences are underestimates.

New Speaker:

Would you ever test to see if there is an underestimate? In other words, test the unit against the within classroom?

Dr. Bargmann:

Unfortunately, there is only one way. You must be able to separate the variation within the class from the variation of a student under test and retest situations. I have always advised those doing research in school situations *never* to treat classes as units, but, if at all possible, to treat *students as blocks* and to make sure to obtain more than one observation on each student in the form of a test-retest or some kind of a verification measure. The same situation holds perhaps with even more force in the transition from model 1 to model 2 in paired comparisons using the Thurstone approach, where you treat one judge making N judgments as equivalent to N judges making one judgment. Well, this is not true. In any case, an experiment can get directly at the effects--especially if we are dealing with educational tests--by treating students as blocks and splitting the test into parallel forms to produce two scores. Thus, we can make sure that we can regard individual differences as block effects and not as an error component.

Dr. Ward:

I would like to throw in a word of support because I think this is particularly imoortant in experimental situations. You have pre- and post-measures and you ought to consider making equivalent forms out of the pre- and posttests so you can do exactly what Dr. Bargmann is advising.

Dr. Findley:

To return to an earlier point, Dr. Bock suggested that when one does not have all the classes filled, he could go out and fill the classes in advance from various sources. I wonder if there might not be some danger in this kind of selection that would be perpetrated upon the data by the fact that you would have to look in certain places in order to find data to fill out those classes. Is that a fair problem to raise? It seems to me this question is often raised in disucssions of the relative merits of matched samples and the analysis of covariance.

Dr. Bock:

If a universe is described by characteristics A and B, and there are not many such people, it may take a while to find them, but I do not see why that would lead to a biased sample.

Dr. Findley:

I don't know that it would have to.

Dr. Bock:

Unless, for some reason, to find cases you went to a different locality or something like that. It is conceivable. One is assuming here that the rare people have some special characteristics.

Dr. McLean:

Here is a perfect example of this: the problem of relating lifetime income with education. We call it post hoc reasoning because almost all the rich kids go to college. If you try to get data on a sample of rich kids who do not go to college, you select a very peculiar group. So while you might be able to fill the cells by seeking these people out, there would be so many contributing factors involved that it is just not good procedure.

Dr. Bock:

That is where the model comes in. If there are several other factors involved, you will not be able to predict the cell response by the general factors that are in your model. So you will find something special is operating and know that you have to look further and elaborate the model quite a bit.

Dr. Cohen:

If it is true that you have a lot of trouble finding rich kids who do not go to college, does this not almost certainly mean that there is some profound interaction operating?

Dr. McLean:

There may be interaction between the independent variables associated with the fact that some combinations are hard to find. But you are not primarily studying relationships between those variables. You are studying the effects they have on the response variables.

Dr. Cohen:

I am not trying to talk about real phenomena. Suppose that you are interested in some dependent variable which is a function of things like education and income. It seems to me the mathematics does not dictate this by any means, but I would almost certainly expect a relatively profound interaction if you had trouble finding rich kids who did not get to college. On almost any dependent variable that you are interested in where income and education were independent variables, substantial interaction would be likely to occur. It is the rarity of this phenomenon of the rich kid that does not go to college that makes this probable.

Dr. Bock:

Nevertheless, because these two factors do not occur very often does not mean that they do not affect the responses predictably.

Dr. Findley:

May we move on to the questions Dr. Bottenberg has put before us. Is it your notion that the rather algebraic fashion of expressing relationships of beginning statistics courses would be helped if we used a more geometric model? Is that the essence of your point?

Dr. Bottenberg:

Well, I don't know if I would call it a more geometric model. For some time, it has seemed to me that if the statement of the model is given by actually writing the entire array or, at least, representative sets of rows and columns of the independent vectors, it is a good deal easier for a learner or a beginner to understand what his model says. He can go into his model and see how, for a particular combination of characteristics, this is what his model says the expected value for that experimental unit is.

On the other hand, it has seemed to me with just the formulation of μ, α_i, and β_j, that these terms in the beginning of training are foreign to a potential educational research worker who is not primarily interested in acquiring a high level of competence in mathmatical statistics. Formulations in terms of parameters alone tend to be confusing and impractical to the potential research worker who wants to acquire some capability in statistics. So, when the model is displayed with the predictor values and he has had some practice in the development of statements of what the expected values are for different categories and combinations of categories, he is in a position to ask himself questions that are a lot more meaningful to him--such as whether he thinks specified categories have comparable differences, or whether they are equal, or any of a variety of kinds of relationships he can formulate. These questions would be difficult for him to ask in the context of a model like $y = \mu + \alpha_i + \beta_j$.

Dr. Cramer:

I want to disagree with that point most heartily. I do not think the choice is between these two approaches; rather, if one has a student who is familiar with ordinary, simple orthogonal analysis of variance, one can talk in terms of what you put down in an analysis of variance table--A, B, and AB--and one can express his model in terms of those effects that are in the model. In the non-orthogonal case, one can completely express the model in terms of the order in which one writes these effects. We need not get to the $\alpha_i + \beta_j$ idea and certainly one need never get to the point of writing down columns of artificial variables, because they really are artificial variables, and I do not think that they convey any great amount of information to students. Furthermore, in my experience students have a great deal of difficulty in routinely putting them down and putting in the restrictions. It seems so unnecessary.

In the manner that Dr. Bock has formulated, specifying that we parameterize or specify main effect contrasts of interest, you can deal with a completely symbolic notation with which students already are familiar if they know something about analysis of variance. The formulation with regression variables is the foreign one. It hides the basic differences that exist between an analysis of variance model and the regression model.

Dr. Bargmann:

Dr. Cramer is addressing himself to a very minor subset of the question that was raised. The question that was raised was "Should we teach our students parameterization or the algebraic expression as such?" I would say "non-parametrics" according to Savage means there are too many parameters. In non-parameterics, and especially in the tests of fit that we use in queuing theory, the representation is not so much interms of formal algebraic models. Let us use a queue as an example. Let's use an aborting queue so that people will not add to it any more if it gets too long. You now have a simple two decision rule: you watch your actual cases and, to compare them, you have a simple non-central chi-square test. The goodness of fit and the contingency tables are a step in the direction indicated by Dr. Bottenberg: the representation of your situation in expectancy tables of some sort, and vectors of certain variables.

On the other hand, I saw something today to which I think statisticians, and perhaps teachers, have paid too little attention. How much information is there in the Venn diagrams? Can we translate the Venn diagrams into some kind of parametric function? Everything is clear if we have very few effects. Then we can graphically represent the model as proportionate areas and this will give us all the information we want. But as soon as we go into very high dimensions, we must somehow translate this graphical approach into some parametric formulation, which may very well be the linear model in any of its ramifications. In any case, I would certainly invite people who are interested in the teaching of statistics to see how they can translate pseudo-graphical-visual displays such as the Venn diagram into a model which represents the situation.

Bear in mind that there is nothing holy or even unique about any formulation that we present. There are many ways to represent the same underlying mechanism. What we must have is something that takes our interpretation back to the physical or educational situation. And if the situation is educational, it is perfectly meaningless to say these variables have a correlation of, say, .78 against this criterion. Please tell us what is the influence. How far does each variable overlap the criterion? How much does this variable contribute? How much are these related to each other? I would say as soon as we treat correlation, or regression, or these parameters, these factors we invent in order to make our problem solvable in a computer, as soon as we treat them as entities having their own life, we are making a mistake. We are here dealing with symbolism only.

Dr. Findley:

May we ask, Dr. Bottenberg, if we are moving into your second question here, "What is the most appropriate way to report predictive efficiency?" I detect in your question the suggestion that one check on this point would be how the means of successive intervals increase when you use a kind of expectancy table.

Dr. Bottenberg:

In regard to prediction systems, I often think of the worker in educational research. Workers are not fully trained in mathematical statistics. One need is some method of evaluating effectiveness of a prediction system or a system they have for representing or predicting or accounting for a criterion variable of interest. One of the ways that has been used very widely in the past is a multiple R, or a multiple R-squared. It has seemed to me after dealing with the problem for some time that this is a relatively uninformative item of informatiun. Some information with a great deal more impact, as far as the experimental worker is concerned, would be to display the criterion mean value for successive valu3s of the predictor score scale. You can demonstrate the impact of how the criterion does change as the predictor, or the composite predictor, changes.

Dr. Findley:

How does this differ from the concept of the expectancy table, such as we use for predicting achievement in college? Are we saying to ourselves that once again we are in peril when we depart from tabulated data and compress things into an index?

Dr. Graybill:

Dr. Bottenberg, are you asking the question: When the multiple correlation is .25 tell us something about whether the predictors are of value or not?

Dr. Bottenberg:

Yes, I am trying to get at that.

Dr. Graybill:

Let me tell you how I get at it. You see I don't believe you can answer the question you ask with what you are doing. Let me take an example and talk about the height of people in Athens, Georgia. What I am looking for is a representative number to call "height of the people in Athens." Well, the most representative number would be the mean. Another question is how good that representation is. Well, if the variance is zero, if everybody is the same height, the mean is a very good measure to use. However,if the variance is five feet, the mean is not a very good predictor. I see somebody down the street and I want to be able to predict his height. Perhaps I can learn his weight. I carry a bathroom scale around with me all the time. So if I can get his weight, I can perhaps predict his height. So I look at the correlation coefficient. I am talking about that of the population, now, not samples. I have this whole population under study and the correlation coefficient is such that if I stratify on weight, so that now the variance within these subpopulations is one inch and I can live with one inch, then I have a very good predictor. So, it seems to me, what we are talking about is by how much I decrease the variance.The correlation coefficient enables me to tell by what percentage I decrease it. But I have to have more than that percentage. I have to know if I decrease it enough to live with the result. For example, if I know that all persons with a specific weight in this town have heights that differ only about an inch, then you tell me someone's weight and I can predict his height accurately enough. The tolerance is the important thing.

Now, when you work with samples you have estimation problems. I think we should first always think of the population. By and large, sample values reflect the population. But the thinking should be done first, it seems to me, in regard to the population. What would I do if I had all my population values available to me? What would my thinking be? Now, I don't have that population value, but I would like to get as near it as I can, so I sample. We have to be careful about making complete reflections of the sample to the population, but this is the way I think it should be done. So, I think many times of how I can reduce my variance. We will, of course, have a variance estimate if it is truly a multivariate situation.

Dr. Wiley:

You have to be very careful to differentiate between two different kinds of problems. One is where you actually want predictive efficiency, and the other is where you are trying to elucidate some basic mechanism.

As an example, I was consulting on a study where the investigator was using a testing variable criterion with item sampling. He was giving only three or four items to each individual unit of population. This makes the error of measurement very large. But he had an extremely large sample and he had very accurate determination of the regression weights. If he cross-validated this regression equation on a new sample that is based on a hundred-item test, the multiple correlation would change radically. You have to be very careful whether you are estimating a regression weight in a system where there is a lot of measurement error versus trying to maximize predictive efficiency in the system that you are currently working in with the current precision of measurement.

Dr. Graybill:

The answer is the same. Here you are working with a different population. You have to be careful at which population you are looking.

Dr. Cohen:

In references that we frequently see in statistics texts where the applied areas are physical sciences, we in the behavioral sciences are left in trouble. Unless we talk in terms of proportion of variance accounted for, we can not express ourselves meaningfully. Confidence intervals do not help us because our units do not mean anything. Units are usually somebody's pets which may be used for the first and last time in the particular study under consideration. It does not help us to know how large the estimation unit is. It helps us to indicate how much variability we can account for by our model in this ad hoc, maybe first and last time used, procedure.

Dr. Graybill:

In that case you would use tolerance intervals. You would say, "Here is the population. What percentage of my population is in these values?" You may not be interested in means, if means have no particular importance for you.

Dr. Cohen:

I find the proportion of variance accounted for a relatively pure measure that covers all kinds of circumstances, since I can not attach any meaning to the units with which I am dealing.

Dr. Graybill:

Well, if you can not attach meaning to the unit, then how can you attach meaning to the variance?

Dr. Cohen:

I can attach meaning to a proportion of the variance that this system accounts for or that one feature of the system accounts for.

Dr. Graybill:

If someone takes your units and multiplies them all by a constant, he will certainly change the variance.

Dr. Cohen:

It would not change the proportion of variance. That is why we would want to use that index.

Dr. Graybill:

A variance ratio is a kind of correlation.

Dr. Cohen:

Right, it is a squared correlation, the coefficient of determination.

Dr. Graybill:

That is what you use to express how much you have reduced the variance. But the problem is that you may reduce it a hundred percent and it might still be so big you could not live with it.

Dr. Cohen:

Well, it depends on how meaningful your units are. If the problem is to predict freshman grade point average, then the unit is one we understand. But if your problem is a theoretical one and the unit is meaningless, then you must finally fall back on how much or what proportion of the variance is accounted for. Our units are meaningless very often.

New Speaker:

It is not exactly that they are meaningless; it is that we have not had enough experience with them to know what they mean. On "Joe's New Test for Social Skills for the Mentally Retarded" we do not know what the units mean.

But we do know if we can account for sixty percent of the variance in this test on the basis of certain variables or characteristics.

Dr. Bargmann:

I will first of all wholeheartedly underline what Dr. Graybill said--look at the population. But I think there is a communication problem. Outside of census figures, a population in other applications rarely is a collection of things. A population is a conceptual unit. A population can be $S = \frac{1}{2} gt^2$. A population is a mathematical model. We may assume actual test scores to be distributed around some true score. The true score in this case, even if it is a single one, represents a population in that sense. The important thing is that as soon as you are dealing with indices--indices that are supposed to reflect how well your data agree with some concept, how well your data serve for predictive inference, how well they explain the mechanism--you should say, "What would this look like in the population?" I do not mean in the population of 100,000 students; rather, I mean, if the concept were exactly true. Now, since I am taking a sample, how far can this fluctuate? What do I have to do in order to condense or to expand the indices that I am looking at into terms that have information for me? Insofar as we deal with arbitrary raw scores in testing, we certainly cannot use confidence intervals in inches. In fact, if we deal with multivariate analysis, we cannot do anything except scale standardized measures because what is our unit then? Is our unit inch-pounds? You see, at one step we must standardize to some kind of statistical unit. The idea is simply not to look at the sample that you get, but say, "What would happen in the conceptual unit called population?" Then say, "How much inference can I draw from the sample?"

Dr. Cohen:

Dr. Bargmann, would you agree that variance proportion terms are acceptable?

Dr. Bargmann:

I like variance proportion very much, but I do not like your inference that it is the square of the multiple correlation. The variance ratio is a statistic that happens to have the same distribution as the multiple correlation under the null hypothesis. As soon as the null hypothesis is not true, then one has a non-central F; the other one is hypergeometric. So you see that they are intrinsically somewhat different. But they have enough similarity to convey the same meaning for those people who have been living with correlation as the last word. A psychologist who reads a .70 correlation has about the same feeling as I would have to hear 70°. I feel comfortable with a temperature of 70°, and I feel comfortable with .70 as a validity. This is about all it is--a convention that has been in practice so long that people think it has a lot of information; but in many cases we condense so much in this one coefficient that it is not going to help us any more either in predictive efficiency or in understanding the mechanism. It is in this case that I would like to go along with Dr. Bottenberg to say that we must have something more visual. Despite anyone's misgivings, I like Venn diagrams.

Dr. Ward:

All I want to say about the Venn diagram is that its purpose is to interpret additive parts of variance accounted for. I am not sure that this is what it does· it can be very misleading. In general, we do not have the additivity implied in the Venn diagram.

Dr. Cohen:

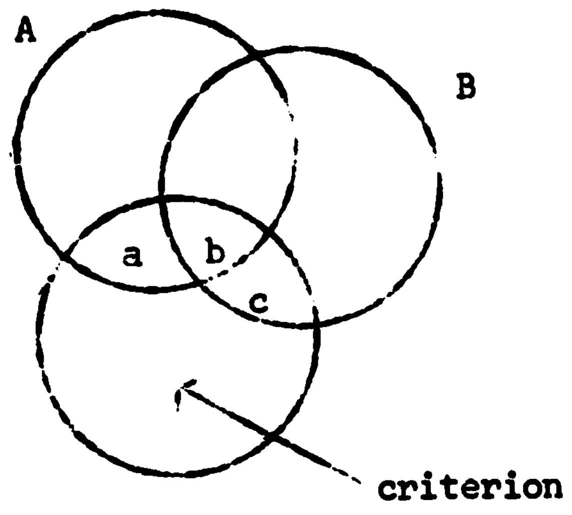
I think it is there. The problem is that a piece of the diagram can be negative.

Dr. Ward:

Yes, but what does that mean?

Dr. Cohen:

Look at this diagram. The areas a, b, and c add up to the squared multiple R and a is positive because it is essentially a squared beta kind of measure, and c is positive, but I wish I could tell you that b was positive. Indeed, I do not know whether it is positive or not. In some instances it is; in some instances it is not. That is where the problem lies. Two years ago I thought one could add them up and interpret the sum as a variance proportion until I started running into negative b's, which will happen when you have a substantial amount of correlation in the system. What I think this says conceptually is that there is a certain logical or substantive priority that goes into this scheme. It matters whether you ask if set B is adding to what set A gives or if set A is adding to what set B gives, because you are talking about different proportions



of variance. The covariance problem for me is essentially deciding whether A or B is the covariate. The covariate takes its variance out first. That is what we mean by a covariance problem. A given substantive problem, depending on who is doing what kind of research, may either want set A to be the covariate and set B to be over and above the covariate, or the converse, and they can make equally good sense either way. It does not matter how we define set A and set B. Set A can be class membership, it can be purely nominal, or it can be x, x^2, and x^3 for that matter. I do not mean to interpret these as variance proportions because we get uncomfortable about negative additions. The relationship

$$\Sigma\beta r = R^2$$

looks like a great way to partition R squared into proportions that are additive, and it seems like a final and ideal solution to the question, "How much does each of the independent variables, however defined, contribute to the variance of the criterion?" Unforfunately, it does not work. The algebra is true, but some of the pieces can be negative.

New Speaker: Some of the pieces can be greater than one.

Dr. Cohen:

Which shows why some must be negative.

Dr. Anderson:

There are also other ways to partition the system; you can compute all of the semi-partial correlation coefficients, square them, add them up, and you get R^2.

Dr. Cohen:

No you don't.

Dr. Bock:

With semi-partials you do.

Dr. Cohen:

No, no! Only with successive ones.

Dr. Anderson:

Use higher order semi-partials.

Dr. Cohen:

That depends on what you mean.

Dr. Bock:

The term refers to a partial on an orthogonalized basis. Now, if you want to call the semi-partial orthogonalizing in all different ways, then of course they won't add up.

Dr. Findley:

We agree, then, that in order to interpret correlation one needs to know more than the correlation coefficient. Let's proceed to another topic. Let us turn to another of Dr. Bottenberg's questions:

"What meaning does a test of main effects have in the presence of interaction? How should it be tested?"

Dr. Bock:

The theory is sound only if the order is prescribed beforehand. Since we do have this strong theory for stepwise tests for prescribed order, I think we should strive mightily to prescribe the order. In my experience, this is not so difficult. Certainly in a model that contains terms of low degree and terms of high degree we would like to throw out whatever terms we can. In many substantive problems there are clearly some terms which, as I said, are problematical, things that you really want to test, and other things that you are pretty sure do have some effect so you want to put them in first. The problematical things, should go in last. I do not think you can do much more than that.

Dr. Cramer:

What I had in mind was a situation in which the investigator would have liked to have designed an orthogonal experiment, and meant to design an orthogonal experiment, but he just did not happen to get equal N's. What he would like to do is draw the conclusions that he would have drawn had he gotten equal N's. So he has two factors, A and B. He is certainly interested in making statements about interaction of A and B, and also statements about A and B themselves.

Dr. Bock:

You have to do them in all orders then.

Dr. Cohen:

You have to use least squares, orthogonalize to get A' and B'. A' and B' are not quite A and B. Orthogonalize them and you see how much each takes.

That is all right if you have missing data because some idiot dropped the test tubes. It is not good when A and B are education and income because, if you orthogonalize in that case, you are studying not quite education and not quite income, but forced vectors that may be least squares approximations for education and income, but are indeed neither.

Dr. Cramer:

Let us stick to the case where some idiot dropped them. We still have several tests that we can perform. If we are interested in testing A, we can test A, eliminating B and ignoring AB. We can test A, ignoring both B and AB. We can test A, eliminating both B and AB, but you have three tests in that situation. If you have more factors, you have many more tests of the same thing. I wonder if there would be any agreement here as to which of these tests are appropriate for what.

Dr. Findley:

Now, there is one type I thought Dr. Graybill was citing yesterday where you use two variables that permit a certain degree of prediction. Then he put a third one in, cut down the error, and improved the prediction. In that case he had a chronological sequence, sort of a natural sequence like we talked about before in predicting grades. You had high school averages a long time before you had test scores, for example. It seems to me the natural question to ask is "How much do the test scores improve the prediction after you have used high school grades?"

Dr. Bargmann:

I will say I wholeheartedly concur with Dr. Bock's presentation that some ordering of the variables should be studied and is indicated. This gives us a lot of information provided that such ordering is based on the physical content of the variables and is meaningful to the educator. Do not do formalistic manipulation of all possible orders to try to find which one gets the greatest increment in F or which one has the least contribution to make. We must not overlook the fact that in our stepwise procedures the best two predictore do not necessarily include the best single one. They may be different.

Now, I would say ordering, in the interest of interpretation, is clearly indicated if the order itself has been established by some interpreter's value criterion. I can very well imagine that in an educational setting you can say the teachers of mathmatics or physics or chemistry represent one particular block. Teachers of English and social sciences, or of English and foreign languages, we treat as another block. In each block I may have a more important, more prevailing criterion of ordering which means something educationally. As soon as we go to partitioning and to the contribution of the last x, unless we have a polynomial fit, this becomes rather hard and, I might say, esoteric or even metaphysical.

Dr. Wiley:

Let me rephrase Dr. Cramer's question, because I think it got lost. The real question is basically "Is it ever legitimate to test the extra due to a main effect above and beyond the interaction?"

Dr. Bock:

It is basically illogical, I think. It amounts to "Are you willing to entertain the function X\ as your model instead of X + Y?"

Dr. Cramer:

The question arises in a practical context. Suppose you have an analysis of variance that was designed to be orthogonal and you believe there is going to be interaction, but you want to get the same treatment precision with respect to all levels of B, which would be something like a blocking variable except with interaction. You want to make a practical applied decision of giving a single treatment level to everybody on the basis of the data even under the circumstances when a large interaction is present.

Dr. Bock:

Well, then you just ignore the other way of classification. There would not be any interaction. If you are in a practical situation of just wanting to know what the population sums are for two classes, you do not need to worry about the other classification.

Dr. Cramer:

What is the appropriate error measure?

Dr. Bock:

I assume you have replications within class. You have only two cells now.

Dr. Cramer:

In your replication, then, do you ignore the nonorthogonality of A and AB?

Dr. Bock:

Yes.

Dr. Cramer:

So the implication of this is that you never hold the interaction constant. You would not correct for it.

Dr. Bock:

Yes.

Dr. Cohen:

You specified that. You said that you are not interested in interaction.

Dr. Cramer:

But I am interested in it.

Dr. Bock:

Let us look at another case. When would you want to use the expression Z = aXY as a model? Well, there are models like that. The gas law is like that. But it is an application where you have well defined variables and it turns out that that kind of functional relationship is a very good one.

Dr. Cramer:

Let us say that I am a naive psychologist and do not know anything about models. I am used to doing my orthogonal analysis of variance, where I do not have to worry about any of these complications, but somebody lost some of my observations. Now, what is to be done?

Dr. Graybill:

I would like to go back to the cell means. Look at the cell means and suppose you, for a moment, pretended that you actually knew the cell means. Now, the question is; "What would you do if you knew the answer to the question?" That is what I would look at. The fact is that you have to estimate these cell means. Whether you estimate them by an orthogonal or a non-orthogonal technique is by the way. What do you want to do with the real population cell means that you had? Do you want to average them over the A classification and make general recommendations? This is one possibility. Do you want to average over B? Do it! You may say "I am going to pick out just one mean and make a recommendation based only on it." Do that! Do whatever you want to do. Don't let the analysis influence what you want to do as a result. After you decide what you want to do, then, in my opinion, use the best analysis possible to get the estimation for the result of what you want to do.

Dr. Anderson:

This is not what Dr. Bargmann would suggest.

Dr. Graybill:

That is why I am suggesting it.

Dr. Anderson:

But Dr. Bargmann says, "Thou shall not interpret chance."

Dr. Bock:

This is a pseudo-question because if you do it in this order when you re-orthogonalize, you will still get the same interaction term. You have just exchanged interaction vectors as the basis for main effect vectors. The same basis is there when you orthogonalize.

Dr. Anderson:

But you can always decompose a system into one term for each degree of freedom.

Dr. McLean:

Yes, but the question is that if you wanted to make the decision between the levels of A, regardless of B, why did you design the experiment with B in it? You may have wanted to make it more precise, but what is the precision going to do for you, if you are going to ignore B anyway?

Dr. Cramer:

Different investigators may want to use the same data for different purposes.

Dr. Bargmann:

An interaction is a term that the model builder puts into the model. I presented strong evidence yesterday that the model itself must be specified.

We must now make additional assumptions about these interactions. You must say, for example, that they are all zero or try to keep them as small as possible in relation to some main effect, or assume some trend neutralization or reinforcement. You must make the statement: "Would someone please tell me, I don't know?" Why is there always a mix-up here? What does orthogonality or non-orthogonality have to do with the presence or absence of interaction? The two are completely different concepts. It may be harder to interpret interaction if you have a very irregular design, but I do not know what the two have to do with each other. Computationally, both are very simple -- an irregular design analysis takes less time than estimating missing data.

Dr. McLean:

We do not understand the meaning of a main effect in the non-orthogonal case, when we have had to do some sort of stepwise procedure to arrive at the sum of squares. Whereas, we feel that in the orthogonal case, we can partition the sums of squares into independent chunks and look at the interaction separately.

Dr. Bargmann:

Do you visualize anything when you hear "sum of squares?" What is a "sum of squares?" Why must they be additive? I often ask myself this because I remember Clyde Cramer doing goodness-of-fit statistics and then after he did marginal goodness-of-fit, he had a residual, of which he said, "Well, I call that a residual." It certainly was not the test for interaction, it was not the interaction sum of squares. If you want it negative, you can have it negative; you can have it anything. It is simply a matter of symbolism. What is a "sum of squares?" In two words?

New Speaker:

Very well, It is a useful simple tool if it is uniquely ascribable by the design to a particular source.

Dr. Ward:

If one does not know what question to ask in the proportional or equal cell case, then he has this problem. But if you know what the question is in the first place, it does not matter to you whether it is an orthogonal or non-orthogonal case.

Dr. Bargmann:

Yes, it is just harder to get the variance components. Variance components mean something to me. They involve conditional inversions and all kinds of things. That means something to me, but "sum of squares" is just as meaningless as, let's say, applying least squares in the case in which you have correlated variables. It is a formalistic expression to which I cannot give any meaning.

Dr. Bock:

Well, you have to evaluate its expectation and see what parameter is involved.

Dr. Bargmann:

Put it into quadratic form and get your estimate, unbiased perhaps, in some confidence region of the variance components. That means something.

Dr. Cramer:

In the example I am talking about, in the orthogonal case, I think the questions I ask are clear. Suppose my experiment actually consisted of two random samples of subjects, and I have two different drugs and two different dose levels. This is a two-by-two design and the question that I might ask in the orthogonal case would be: Does the dose matter? Does the drug matter? Now what I want to know is how I can test them.

Dr. Graybill:

We are not going to test them. We will look at the means.

Dr. Cramer:

Are you saying that it is not valid for me to say, "I want to know if these two drugs are different?"

New Speaker: It is going to be invalid unless you specify what it is supposed to be.

New Speaker: I would have you test interaction rather than make an assumption about interaction.

Dr. Graybill:

Suppose you have a two-by-two design in which we will call the classifications A and B on which you have observations. You can estimate the cell means with confidence intervals. Suppose you knew the cell means; let us talk about what it would mean if I knew the population cell means. What questions do you want to ask? If you want, you may ask: "How close is $(\mu_{11} + \mu_{12}) / 2$ to $(\mu_{21} + \mu_{22}) / 2$?" That is a very nice question to ask. It can be asked, whether there is orthogonality or not, so long as you do not have all data gone. Well, suppose I want to ask another question: "How similar are $(\mu_{11} + \mu_{12}) / 2$ and $(\mu_{12} + \mu_{22}) / 2$?" This is a perfectly legitimate question. Yes, I would look at this using data from the same experiment. There is another question that I might want to ask, "Is $\mu_{11} - \mu_{12}$ the same as $\mu_{21} - \mu_{22}$?" That is, is there interaction? This is a legitimate question to ask right along with the other questions in the same experiment. You may want to ask the question, "What is the value of a particular cell mean?" That is a perfectly legitimate question that can be asked right along with everything else. As I said this morning, I would use the studentized maximum modulus, ask all these questions and answer them all in the terms of one sample with a known protection level of error rate. So that is why I use the modulus in any analysis and then say, "What questions do I want to ask?" Don't be limited in what questions you ask by whether the data are orthogonal or non-orthogonal. This is immaterial, by the way, if you have a computer. If you have a desk calculator and you have to have your answers by 2:30 this afternoon, you have to ask different questions. Even if some means are missing, you can still ask about the effect of A in the presence of this level of B or maybe one mean versus the average of two others. So what I do first is say, "Do I know the population? Now, what question do I want to ask?" I write down the questions. Now, I say "What is the best answer to those questions with the protection level that I am going to deal with?" You can do it. Orthogonality and non-orthogonality are beside the point.

Dr. Bargmann:

You are going to do it for a two-by-two design, why not for a ten-by-twenty.

Dr. Graybill:

Yes, the same thing.

Dr. Bargmann:

You want to write down trillions of different questions that you might have?

Dr. Graybill:

If I have those questions, yes. If I have the questions, I do.

Dr. Anderson:

Before you ask any specific question like this, would you test the overall cell variation? Are you interpreting chance?

Dr. Graybill:

No, absolutely not. The reason I would not test is this. Suppose I have a mean I am interested in and suppose I come up with a confidence interval like I talked about before. Let's use as an example the average height of people in Athens. Suppose that I come up with 5.11 to 5.12 feet for the confidence interval of the average. But suppose I am testing whether it is 5.10 feet. I would reject the hypothesis. Yet, when I use a practical point of view, I say for all practical purposes, if 5.10 is something sacred, yes, I'd just as soon use 5.10. Now, my result is significantly different from 5.10 at the .001 percent level. I would reach the same conclusion if I had obtained an interval of six to seven feet. So, you see, a significance test throws away too much data. It summarizes your data too far -- way past the sufficient statistic. The confidence interval if you are not a Bayesian, or the probability interval if you are a Bayesian, does throw away a little information, but not very much. I think the onus is on the experimenter to not say if the average is different from 5.10 or not, but to say, "Look at this, I have to make meaningful decisions on the basis that I believe quite strongly that μ is some place in this interval, that μ_{12} is some place in this interval and μ_{21} is in this interval. Now on the basis of that, I am going to ask various questions and then I am going to do the best job I can today to answer those questions. Let us not do just "yes-no" tests of significance or hypotheses. You take a lot of time to collect the data and summarize it in one little number. You are really sacrificing your data. I really think tests of hypotheses are the worst thing of all, tests of significance are the next to the worst thing of all. A test of an hypothesis is like this. Suppose I test $\mu=0$ and decide to use the five per cent level. If it is above five per cent I reject, if it is below five per cent I accept. That is a two-decision problem.

Now, the test of significance says "Test $\mu=0$" and I ask the question "At what level of significance do I reject $\mu=0$?" Maybe rejection at the .001 level gives me a lot more information than rejection at the five per cent level. You want to know at what level I rejected. This is not a decision problem. And the power of a test has meaning in tests of hypotheses, but does not have meaning in tests of significance.

Dr. Bargmann:

Assume that you state your model in terms of μ_{ij}'s and you ask questions in terms of these μ_{ij}'s. Then by all means the confidence bounds found have all the information that you want--the confidence bounds on the questions that you ask in the various comparisons.

Dr. Graybill:

The F-test involves something in the numerator and something in the denominator. Now, when you put them together in an F-ratio, you throw the two individual parts away. You throw the individual parts, the numerator and the denominator, into one factor. Suppose I have a model in which I am testing for a quadratic effect. Now, suppose I look at the data, and it is such that there is no doubt in my mind--in fact, it is such that any experimenter would say there is a quadratic effect operating. That is one possibility. But suppose an atypical observation is included. This increases my variance so that I get non-significance in spite of the obvious quadratic pattern.

Dr. Bargmann:

If you have extreme outliers, you should edit your data first.

Dr. Graybill:

I am trying to make it dramatic here. When you use the F-test only as a test, you overcondense your data. You condense your data too far.

New Speaker:

The maximum modulus has a denominator, too, though.

Dr. Graybill:

But the maximum modulus would be used to set limits on each point.

New Speaker:

Are you saying that just because somebody misuses the F-test here, we should throw out tests of significance?

Dr. Graybill:

I am saying that it is always misused.

Dr. Bargmann:

I agree that it gives too little information. I said that you should not interpret what you get in a test of significance. I merely say that a test of an hypothesis that all effects are equal is a test of a perfectly legitimate, simple model. If you are left with accepting that particular model, then I think you are an astrologer if you say, "Now, I'm going to take these things and break them into components." You can break them into components in infinitely many ways. But it is a matter of parsimony, the simplest model. All of them are equal. The rest is just random. This is, by far, enough to explain. Why look for a mechanism or even say later, "My confidence bounds are proof of a mechanism?" This, I think, is dangerous, that's all. It is a verboten sign. If you do not reach a certain level, it's verboten to go on. That would be my emphatic statement.

Dr. Graybill:

Now, I would like to ask you experimenters a question. Suppose I spend $100,000 getting data and I make an F-test and it is not significant. Are you going to shove that data in the drawer and say nothing else can be done. I don't think so. You are going to milk that data for information.

Dr. Cramer:

I quite agree with everything that Dr. Graybill was saying. Things should be phrased in terms of confidence intervals rather than tests of significance. Let us take our two-by-two table of means and say that the statements I am interested in making about the true means are about the differences of the sum of the diagonal element minus the other two diagonal elements, the sum of the first two row means minus the sum of the last two row means, the sum of the first two column means minus the sum of the last two column means. Now, perhaps these have a profound relationship with the parameters of the original linear model which I might have written down to start with. I am interested in confidence limits for these and I can get confidence limits or tests of significance under various circumstances. What they amount to doing is comparing different models. I am not clear which I should be doing.

Dr. Bock:

O. K., so these are not independent. You don't care. Use some method of judging all contrasts that protects you even though they are not independent.

Dr. Cohen:

I'd like to take issue with Dr. Bargmann. Here is an R-by-C matrix and the data mean something. I won't specify what, but they mean something. They account for criterion variance. However, I know that some of it is accounted for by the R variable and some of it by the C variable, and some of it by the RC interaction. Now, if there were no problems of power in this system, I would have no difficulty at all going along with what you say. I'll test the whole RC set and if I do not get significance, I am prepared to accept the simplest model. I can't help but see the F-test on cells as sort of "communism" in the extreme--taking all these effects and dividing them equally, each according to its single degree of freedom as it were. Then it makes a decision about the average of them. If C happens to be a weak effect in the population, R is a null affect in the population, and the interaction is a null effect, then this operation will almost certainly leave me with an F-test that will meet hardly any criterion. It won't meet the five percent level. It may not meet even the ten percent lèvel. I would prefer to think of these three sets, the R set, the C set, and the interaction set, as families, at least. Within these families I would be quite prepared, if the R factor set was not significant, not to pursue that further, then look at the C effect, and so on, rather than throw them all into a single RC conglomerate.

Dr. Bargmann:

This is really a very simple question. You observe only what happens under a combination of R and C. You postulate in your mind that this is a combined effect of R plus C, plus some interaction between the two. But that is not enough. You now suddenly also say, "I want to make the contribution of R and C as large as I can and deal with whatever remains as being non-additivity." This is what you have in your mind. Your data can give the sufficient statistics, namely the cell means and standard deviations, all this is available. But what of the compound F-test? The compound F-test tells you whether the full R by C design could have been brought about by chance, if all effects have the same true μ. Suppose I get an F that comes to the .10 level.

You want to say that you have a little of something because you assume that you had a weak effect in C, practically none in R, and none in whatever remained. Consequently, haven't you thrown something away?

I merely say, "Yes." This may be true. You have evidence. You may go on and say. "I have a hunch that there may be something in C which is somewhat obliterated." I think the only way that you can demonstrate the hunch to outsiders and make it convincing is to say, "In the presence of these weak effects, I have got to take more data. I have to collect additional data until I reach a somewhat reasonable significance level, .10 let's say. The effect presumably always works in the same direction, however weak. If you assume that such a weak effect exists, then if you take enough data you will eventually also prove it, prove it beyond a shadow of a doubt--and a shadow of a doubt is .10, .05, .01, you name it.

New Speaker:

You didn't get to Dr. Graybill's problem though. He spent $100,000 to get the data and wants to get something from it.

Dr. Bargmann:

I may treat it as a completely different problem. I may consider a completely different scoring system, for example. Psychologists don't seem to realize that as soon as they reject, as we say, or accept, a certain null hypothesis, this means complete randomness. In many personality studies this may be true. But they haven't thrown away the data. They haven't done everything in vain. They just have a poor scoring technique. Go back to your old data and find some other scoring principles, graphology, for instance! Look for more meaningful ways to handle the basic data till you find something. The scoring has been insufficient. Eventually you can always take the data again, re-evaluate them and re-quantify them in different ways and then you may come up with something.

But to try interpreting or breaking the total set down into row, column, and interaction, if you know that the three could very well add up to zero, does not seem scientifically tenable and does not hold up in court.

Dr. King:

I recently encountered the opposite situation. I did a three variable regression problem where the overall F was significant, but when I tested the individual regression coefficients none of them was.

Dr. Eber:

(from floor) You are testing parts and each one of the parts excludes the contribution of the others, including the joint contributions. I have a specific example in mind. In some rehabilitation studies there are ten factors we are studying. These are composite factor scores. Five of them represent characteristics that the client brings to the counselor to begin with. The sixth one represents college training that is given to the client. Now, if college training turns out to be insignificant in its long-range effect on the client, in terms of this model where the first five have been partialed out beforehand, what we are saying is that college training as such is not a significant influence. College training together with what there was about this client that made us decide to give him this kind of training may be quite significant. This is precisely the answer to the question that we want because what the counselor is asking is, "If I go through this set of plans randomly and sent everyone to college, will it really help?" No! That's what the model is saying. No, it won't.

Dr. Bock:

I think that is a little too causal an interpretation. What you are saying is that it can be attributed to that, that it is associated with it. I am not too sure about vour causal implication.

Dr. Eber:

Perhaps it is too causal. But the point is vou can at least make some hypothesis about what is going on psychologically in terms of what variables exist in the system. So the statistical procedure is giving you the right answers.

Dr. Findley:

There is a question that I think is very much in order. The third question of Dr. Bottenberg is on the use of a binary criterion. This is rather distinct from the other points that we have discussed. The use of a binary criterion in the context of the general linear regression model is seldom discussed.

Dr. Bottenberg:

Schemes based on discriminant analysis and likelihood ratios are ordinarilv suggested as the proper approaches to classification problems. But those techniques are usually difficult to understand. Those engaged in educational research are interested in classification problems. They may well consider the use of linear regression models with a binarv criterion. Some empirical results obtained at the Personnel Research Laboratory indicate that a regression model approach can be used effectively in dealing with classification problems. One can evaluate the effectiveness of the approach in terms of a hit table, displaying the correct classification count weighted by payoffs and costs.

Dr. Bargmann:

Certainly the regression approach, if you have several variables in the binary classification, is the correct approach. What is happening here is that again we are dealing with the left side and the right side of the equation. The right side in this case consists of 0's and 1's, or perhaps if we have more categories 0's, 1's, 2's. These are the design constants. The left side is the information that you have on all the random variables. Now we can set up discriminant functions, we can start classifying. It so happens that if we just want to classify the binary way--0 or 1--or if we want to explain the best way to total a score so that the score leads to the best classification, the multiple regression approach gives the correct answer. This is a mere coincidence. It happens to be the discriminant function. Take another instance. In weighting items for college selection, we need to find the weights in such a way as to discriminate best between those who will succeed in college and those who will fail in college. The discriminant function identifies the best set of weights, which happen to be the regression weights.

Now, there can be more than two groups. For example, when you have people who are talented for office training, people for technical training, people for general training, and people for KP. Now your set of tests have to discriminate among four groups. In this particular case, the solution is no longer multiple regression. The solution happens to be the vector associated with the largest root of a matrix product. It is a multivariate analysis that yields the discriminant function.

I would say that the claim that statisticians have not paid attention to this problem is not correct. They have paid very close attention to it and it is perhaps important to note that there exist multivariate techniques which are well known, well developed, easy to obtain, and, I would say, available to practicing people in education.

There is, of course, need for more study, especially in the case where allocation to one group or another is a random variable. Box is doing quite a bit of the research on this. There are many ways to handle the problem and practical ways are available.

Dr. Anderson:

You are all right as long as the means of the groups are in a straight line in the hyperdimensional space. If they are not, then you do have a problem.

Dr. Bargmann:

A straight line is not a requirement.

Dr. Anderson:

But it is. The means of the groups certainly enter into the calculation and if you do something like Rao does, you get an entirely different set of weights for each group. You do not get the root and vector.

Dr. Bargmann:

You can make a pairwise split--A against B, A against C, A against D, B against C, etc., and get useful information. If that is meaningful, then by all means do it, but if, as in education, someone wants a total score, what weights should be used? Usually these weights should be given in terms of a criterion set by the educator. But sometimes the weights are to be determined in such a way as to give the sharpest discrimination between certain types of groups, for example, interest differences between certain types of professions. In this case, the means are not really on a straight line. You should look at what I will call a calibration sample. Perhaps you have groups of extremely successful physicians, extremely successful accountants, undertakers, and so on, and you get a weighted total score of your various measures. With this weighted total score, you find out which one to take in order to discriminate best between the groups. Then it is evaluated for each individual and probabilities are minimized.

Dr. Anderson:

If you do it by the canonical approach, do you get a root and a weight for every variable and the same weights for every group?

Dr. Bargmann:

They may be the same but they may be only proportional: they are arbitrary.

Dr. Anderson:

Now, if you do it by the Rao approach, do you get different sets of weights for the variables for each group?

preparing for college, or some other group? Without stating what you mean, I can give no answer."

Dr. Findley:

Let us take time to go around the table for questions we may have missed.

Dr. Wiley:

We have glossed over a couple of things in regard to models and design. One essential point with respect to the usual general linear hypothesis class of models is that there are distinctions between them. When Dr. Ward is fitting the four models that he gives in his paper, those are four very different kinds of models and are useful for very different kinds of purposes. There ought to be some attention given to the question of under what circumstances one formulates what models for what purposes. In a very special area Dr. Bargmann has done that in his paper. In general, I still think there is wide residual confusion among the audience and among the readership about what in fact are the important distinctions. I would like one of the speakers to comment on this.

Dr. Bargmann:

The only quick answer that we could give is that the investigation of the plausibility of models--formulation A versus formulation B--of course, falls into the domain of exploratory analysis, to which statistical tools are sometimes applied. The entire area of effect analysis, and I think I ought to restrict it to effect analysis, is essentially an intent to try to indicate what kind of models would be plausible and what would not be plausible. I am sure there are many others in the entire area of spectral analysis, he entire area of fitting curves of unknown degrees. Many stochastic processes are made for the explicit purpose of providing pointers as to how to formulate a proper model which you can later test versus various alternatives in a confirmatory sense. So, attention has been given to it. I must admit, though, the attention given to it has been highly esoteric. In fact, some of the concepts that have been introduced, concerning the nature of latent variables and their relationship to observable ones, are quite difficult. When a psychologist talks about a factor loading, chances that the biologists, who may have a very similiar problem, would not even know what he is talking about. If one person talks about a correlation meaning something to him, the next person might say, "I don't know what it means." So, I would accept the challenge as one member of the statistical profession, that we can do a lot more by exploratory tools in order to help people get pointers to formulate models.

Eventually, however, all these tools, whatever numbers may come up, whatever vectors, whatever vector leadings or interesting thoughts, will have to be translated back into the actual originally measured variable and not to a principal component or to a latent factor score or something that no one can really interpret, but to what you really measured in the beginning. Once you have done this, I think you have done a service to the problem of formulating a model which under the circumstances seems plausible. It is never correct or incorrect, true or false, but can be either plausible or implausible.

Dr. Winer:

One thing that we have lost sight of here is that all of these models are man made. We have to admit this initially. The use of statistical procedures to add or subtract from a model, I think, is somewhat inappropriate in that essentially these models really establish a probability metric that we use at one stage or another. I just don't know the real answer. Fisher gave the job of model specification to the applied man, not because he wanted to give up any of his real mathematical work, but because he was, in a sense, incompetent to deal with this aspect of the problem.

Dr. Ward:

I agree with many of the comments that have been made, particularly Dr. Graybill's in the sense that we have to look at the problem that we have generated. The thing we have to do is develop the capability of our research workers so they can formulate appropriately their own problems instead of asking someone else to answer the question, "What is the right thing to do here?" What you want to do is give them the capability of conceptualizing things in their own situations. Now, related to this, I want to get back to Dr. Bottenberg's first point. If we continue the practice of getting people to understand well-developed models that may not be relevant to their problems, we are being fundamentally inconsistent. I hope we keep on telling people to go out and formulate their own problems. That is the only fair thing we can do. We may want to continue the way we have been doing. But let's not keep on trying to convince people they ought to think about problems in a very restrictive way so that they get in trouble one time after another. Let us try to determine the individual's objectives, but ask this person to think about his own problems and not worry about proportionality unless it is relevant.

Dr. Bock:

I had one other comment to make. It is the question of units of measurement. A lot of this would be solved if we would stop denying the fact that behavioral variables are almost entirely qualitative. We get around it by using a test score which is the sum of qualitative responses. On such items either you pass, fail, or omit. We hope that we can treat this test as a continuous variable and model it with a linear model that defines values in terms of real life. In a sense, it is quite a fiction; it is quite removed from the things with which we are actually concerned, that is, the qualitative response the examinee makes. If you deal with qualitative data, then you very naturally want to state the probability that he will respond this way or that way, or that this individual will be classified this way rather than that way. So all of the quantitative aspects of the discussion is in terms of probabilities. They are nice because they vary from zero to one on a well defined metric. They are not so nice because they are confined in this way: you cannot easily construct ordinary linear models for probabilities because the ordinary linear models will quite often give you a negative value or a value of over one. It is, however, possible to retain the usefulness of linear models if you introduce the concept of a response law--a non-linear function relating a linear model to probability of response, mapping this real line that the linear model defines in the (0,1) interval. In the last few years, I think we have made some very good progress on working with these types of models. I have been concerned with two types.

One of these is for contingency type data. I have been using a generalization of the logit, the multivariate form logit transformation and using maximum likelihood to estimate parameters of an underlying linear model connected through the response law to the observable responses. More recently, a number of us have been working on models for dichotomously scored test responses--one and zero responses--which treat this vector of ones and zeros as a qualitative entity. We do not sum them to get a score or weight that is treated as continuous. Again, the underlying model is linear, but it is connected to the response probabilities by non-linear models and, again, maximum likelihood serves very well.

I really think this is the direction in which we should move whenever we can and this is the thing that really fits our kind of data. This idea was really Thurstone's. We owe to him the idea of working with psychological data in this form. I hope we can push on and extend it.

Dr. Bargmann:

I do not want to detract from the importance or value of this approach of postulating a linear response, logits, probits, etc. They are very useful. They have one horrible handicap--they always fit much too well even if the data is random. But I will say we are still dealing with two entirely different problems. Why do you concentrate on behavioral sciences phenomena? The concept of heat is extremely qualitative. Some esoteric mercurial scale was developed, based on the behavior of a mercury column. Then someone introduced, somewhat loosely and poorly, a concept of temperature. Temperature is now measured on some scale and applied to this varying qualitative concept heat. It has nothing to do with subjective heat because what I call cold in summer may have a different temperature point from what I would call cold in winter. But it is this concept of an almost artificial scale that we are using for communication that has enabled us to establish relations between chemical reaction and this scale for heat, between energies and heat, even between physiological phenomena and the temperature scale. The approach of getting probabalistic or linearized models--standardized, linearized, normalized, logit, ranked--is useful, but the other approach eventually will be more effective. We want to find a scale, perhaps a combination of the test scores, perhaps a scaling of the test scores, that enables us to predict.

Dr. Bock:

I do not say the independent variables or the conceptual variables should not be quantitative. The response may be quantitative if it is how heavy a weight you can lift or something like that, but it is not very often quantitative.

Dr. Findley:

Don't we often now in the field of measurement think of a person doing a number of tasks successfully rather than simply a single task? Now each single task is a qualitative response, but when we think of a person being competent in any particular area, we seldom think of it as being made up of the separate actions that the person does. There is a continuity of scaling. If we define competence as all of the things he can do, then we have an infinite set of actions that make a scale. The scale becomes continuous, in effect.

Dr. Bock:

That is the distinction that we live with, but we do not really ever have this infinity of things and we do not know in many cases that we should regard it as a metric measure. The only behavior variables we have that are really quantitative are things like response time. They are really rare once you think about it. Most of the variables are really qualitative.

Dr. Findley:

You seem to reject the notion I was taught when I was first introduced to the use of bi-serial r--in item analysis. The assumption was that being right or wrong was a normalized trait and that either you had just enough of it to do the task or more, or you did not have quite enough to do it.

Dr. Bock:

No, I am not objecting to that. That is exactly what I am saying, but the response is to pass or fail. The underlying trait is latent; you cannot observe it, but it is in the model. The trait is on the scale for which the linear model is set up. It is not the kind of linear model that we are talking about here today because it has to be connected through the non-linear response law to the pass-fail response.

Dr. Bargmann:

But it is a form of scaling. You take regression models.

New Speaker:

In other words the crux of a lot of this is really the scaling.

Dr. Wiley:

I have one more comment about design. I think that has probably been the most neglected subject in this discussion. No one seems to be talking about design and the really fantastic economy one can achieve by having a planned observational scheme or a designed experiment. I was thinking about the example that Dr. Winer gave of the two quantitative independent variables which were classified into a cross-classification and the one that Dr. Ward gave which was 15 by 25, where the variables on which the classification was based were inherently quantitative and were scaled on a reasonable metric. In such cases, factorial designs are not the most efficient method of detecting or fitting the model, testing the model, or looking for lack of fit in the model. So that if one has a general quadratic model which one hypothesizes for a phenomenon, something like a composite design would be an excellent way to fit the model with a great deal more economy of observation, and a great deal more efficiency for a similar sample size, than going to the problem of a whole factorial array of data whether or not it is a designed experiment or a planned observational scheme.

Dr. Findley:

This concludes our discussion. Our thanks to both the panelists and the questioners.