STATISTICS, MEASUREMENT, AND EDUCATIONAL RESEARCH

Henry F. Kaiser, University of Wisconsin

This afternoon I should like to make a short survey of some random ruminations and polemical personal pontifications regarding educational measurement and statistics, and point out possible implications of my remarks with respect to the scientific study of education.

Preliminary Thoughts

Because of a failure to distinguish measurement problems from statistical problems, traditional training has often confused and thus confounded these two distinct aspects of quantitative methodology; witness the unfortunate synonymous use of "reliability" and "significance." Current thought attempts to react against the earlier confusion to separate, probably to too large an extent, measurement from statistics. For in one sense such a distinction is a little artificial: almost invariably in educational research the problem of devising measurement instruments on one hand and the statistical problems of sampling and design cannot each be considered in isolation.

I think it could be said that the problem of measurement is closer to the concerns of the educational researcher than is statistics. Statistics, as an independent discipline, is only a formal set of procedures for analyzing data, while the scientist must take upon himself the principal responsibility of devising his own measurement tools--for only he knows, however vaguely, the concepts with which he is concerned.

In educational research, perhaps the largest methodological difficulties stem from a failure to plan ahead with sufficient care. Once we perceive a problem, we are tempted to blast forward in an ill-conceived fashion to attempt to solve it. Oftentimes, technical problems of measurement and statistics are only vaguely conceived of *a priori*, being dismissed with the thought, "We can cross that bridge when we come to it." It is found only later that data so enthusiastically gathered cannot be analyzed in a systematic fashion. Some people who have been denied this allegation, suggesting that if enough data is gathered with enough enthusiasm, solutions to problems will surely come forth like a bolt from the blue. Perhaps so; but quite likely these solutions will be to the wrong problems. Rather than "cross the bridge when you come to it," a better maxim would be, "Look before you leap." Careful thought on what to measure and how to measure it, considered simultaneously with appropriate methods of statistical analysis is the sound way to do business. Yet also I would never suggest that such planning should inhibit subsequent effort; one should not be stunned into silence because of difficulties in planning. A struggling start is certainly better than no start at all.

Let us think about statistics in a little more detail. As I suggested before, statistics is an independent discipline, having nothing necessarily to do with any science. From this viewpoint, statistical methods are capable only of providing us with decisions about the probability distributions of random variables. It is the responsibility of the scientist, as a scientist and not as a statistician, to consider the relationship of statistical decisions to scientific decisions in education. In playing the role of statistical consultant, there is nothing more distressing to me than having a purported scientist ask me to state his problems; who am I to speak with authority about the problems of the administration of secondary school guidance for curriculum workers in a laboratory school setting?

Scales of Measurement

The relationship of kind or level of scale of measurement of the educator's data to statistical procedures appropriate for working with these data persists as a topic of great controversy. In one school statistics is considered as above: formal discipline which bears no necessary relationship to the real world. This school of thought--to which I must admit for the most part I adhere--asserts that considerations of scales of measurement are irrelevant to statistical procedures. Actually, this is more than an assertion: it is a fact. Statistically, we can do anything we please perfectly "legally" --so long as the formal statistical assumptions are more or less met. But, whether the statistical results have any scientific meaning is an entirely different question, and should be thought of as such. *For example*, we have often heard the statement that a variable must be measured on an interval scale in order to compute a mean; really, this is hogwash. However, for our scientific *interpretation* of a mean, such considerations *may* be of importance. Pertinent here is the distinction between a scientific and a statistical hypothesis. While statistical hypothesis is nothing more than a statement about the probability distribution of a random variable, a scientific hypothesis is a statement about something in the real world. For example, the question, "Are boys smarter than girls?" is not a statistical hypothesis. However, corresponding to this, as scientific hypothesis, there may be a reasonable isomorphism to a statistical hypothesis; in this case, it could be the assertion that the population mean IQ of boys is greater than that of girls, given that IQ both for boys and girls is normally distributed and that each of these distributions has the same variance. This statistical statement obviously leads to a traditional *t*-test. What it appears that we do in practice, then, as scientists using statistics, is first to state a scientific hypothesis, then translate this to a seemingly reasonable statistical hypothesis, formally test this statistical hypothesis, make a statistical decision, and finally make a corresponding equally reasonable or useful scien-

tific decision. In making these reasonable translations, it would appear that in some not-well-defined way measurement considerations are of importance.

The other school of thought on the question on the relationship of scales of measurement to statistical procedures is due to S. S. Stevens, who is responsible for the insightful taxonomy of scales of measurement into nominal, ordinal, interval, and ratio scales. Although after the fact, this taxonomy seems particularly obvious, it was not fully articulated until the late '40s. It may surely be considered one of the major landmarks in the theory of measurement. However, it would seem that Stevens has gone a wee bit too far. This is represented by his pontifications on what you can't do. For example, you can't do a t-test unless the variables are measured on an interval scale. As suggested above, of course you can; what I think Stevens means to say is if you want to make scientific sense out of the results of your t-test it is really sufficient that the variables be measured on an interval scale--and it may be necessary.

Perhaps my disenchantment with these prescriptions stems from the experience of dealing with educationists who have taken Stevens too literally, i.e., I have often been confronted with insecure and terrorized consultees with hollow and haunted eyes, so frightened that they will do something "wrong" that they are inclined catatonically to do nothing. After all, many of us are scientists first and statisticians second. To allow statistics to repress our ideas is an anathema of the worst kind. The tail should never wag the dog.

To the extent that research conclusions hold up, one has pragmatic evidence of the efficacy of his scaling assumptions. For example, the massive weight of evidence would indicate that most intelligence tests are essentially measured on an interval scale. As an educational psychologist I would say with a high degree of confidence, that a difference in IQ of 75 and 85 may be considered the "same" as the difference between 105 and 110. To have established this statement a priori is like proving the existence of God. On the other hand, if one were newly to devise a rating scale, say, and handle the data as if it were on an interval scale, the generalities of the scientific translations could well be suspect--although the statistics was perfectly dandy.

Thus it would seem that the critical issue in scaling is the scientific generality of the resulting conclusions. The more general the scaling, the more general the scientific (as opposed to statistical) results will be. A careful consideration of levels of measurement would then seem essential to scientific conclusions, although such cogitations are irrelevant to statistical procedures.

Significance Testing

Let me turn now to some comments on the applications of statistics. In reading the research literature in education, I am impressed --I might say appalled--by the relative frequency with which tests of significance are performed as a matter of thoughtless ritual. Although I have strained for years to understand the meaningfulness of the seemingly ever-popular significance test, I remain convinced that there is little relationship in this ritualistic procedure to the scientific thinking of educational investigators. First, typically one tests a null hypothesis against all possible alternatives. Appealing to subjective probability, such hypotheses are simply absurd. What scientist, on this green earth, would ever state that girls and boys are exactly equally bright? Or that the Dandy-Dan method of teaching arithmetic is exactly as effective as the Johnson-Kleinsohn procedure? Testing such closely specified null hypotheses against omnibus alternatives simply doesn't make sense, for such null hypotheses will be rejected, or not, simply as a function of the sample size and the power of the test used. Even were boys much, much brighter than girls, a sample of size 2 would rarely show significance; or, if boys and girls were essentially equally bright--but not quite--a sample of size 4,000,000 would almost certainly pick up the negligible difference. Significance testing is a myopic way of doing business.

It seems to me that the only time when testing procedures in statistics are valid is for the purpose of final adjudication between two or more equally specific theories, where each can be translated into statistical hypotheses of the same dimensionality in the parameter space. Thus, rather than having "everything-else" alternatives, the scientist should state a particular difference which he, as a scientist, considers to be educationally significant. Once this is done, he should assert precisely what sort of risks he is willing to take for all possible errors. Then, the standard application of statistics (à la Neyman-Pearson) will do the deed. But in areas such as education and psychology--the behavioral sciences generally--studies which are concerned with such final adjudication would seem rare indeed.

It's just that in the vast majority of educational research, theory has not reached a level of sophistication which allows scientists to make precise quantitative predictions for alternative hypotheses. For these studies, a more appropriate statistical procedure would be to estimate the differences of interest or the degree of relationship rather than dichotomously succeeding, or failing to succeed, to see "truth." Thus, the first concern would be a point estimate of the parameter for the problem in question. (I must admit that this major interest would seem implicit in the somewhat irrational defenses of so-called "descriptive" statistics.) After this primarily important point estimate is made, it would seem nice to jazz it up by putting an interval about it and indicate the degree of confidence which we have in the interval. Finally, but least important, we might sneak a peek to see if our interval covers zero. It is most unfortunate that many popular texts emphasize significance testing first --not last...

Nonparametric Methods

At this point, it seems appropriate to comment on the recent rise of nonparametric methods in educational statistics. I consider the stampede to these procedures unfortuante. First, most nonparametric methods emphasize the significance testing viewpoint. Usually, in nonparametric procedures, distributions are computable only under a traditional null hypothesis--have you ever heard of the sampling distribution of the rank-order correlation coefficient for a population rank-order coefficient different from zero?--and thus, to a certain extent, my previous diatribe about the thoughtless use of significance testing applies. A second consideration is related to the question of scales of measurement. For there are many poor souls who are driven into a dark corner by the imprecations of the overly serious scales of measurement boys and are unwilling ever to accept the notion of an interval scale and thus, at best, apply the much less informative nonparametric methods to their safe and sure ordinal data. Third, we often hear the cry that it is so important to meet the formal assumptions in statistical procedures. The question, "Have the assumptions for the t-test been met?" is an example of the watchword of these folk. Well, of course, the statistical assumptions have not been met. Nor have the assumptions of the corresponding nonparametric approach been met. For the assumptions in a formal statistical model are abstract assumptions and never can exactly be met in the real world. There is no such thing as normal distribution in Nature. Or, for most nonparametric methods, who ever heard of a continuous distribution existing in Nature? The correct question of "just" meeting assumptions is somewhat more difficult. It is simply a matter of how closely one comes. And to assess how close one must be seems a subjective, almost arbitrary decision. Fortunately, the problem of meeting statistical assumptions--considering that they never can be met exactly--is not really so bad for many traditional procedures using metric data. Empirically it is well known that standard, very useful things are robust, i.e., relatively insensitive to the underlying statistical assumptions, and thus one can blatently go forward with only slight distortion in his probabilistic conclusions.

Multivariate Analysis

In the often encountered situation in the behavioral sciences where there are a number of criterion or dependent variables, there has been built up in recent years a large number of techniques subsumed under the general title of multivariate analysis. Unquestionably the most common multivariate procedure in use today is factor analysis, a technique which, at the exploratory level, has probably done more than anything to bring some sort of preliminary order out of chaos.

I can't resist the opportunity to get in a plug for some recent developments in traditional factor analysis. First, regarding the communality problem, Chester Harris of Wisconsin has recently published some remarkable results linking the important statistical work of Rao with the important psychometric work of Guttman. His paper has clearly demonstrated the crucial notion of scale-free solutions--solutions which are metric invariant, so that we are no longer tied to the traditional normalization of observations. With regard to the transformation problem, Harris and I have invented methods which can yield all possible solutions--involving correlated or uncorrelated factors--using orthogonal transformations only. This seems important, for we can now attack the general problem with tractable machinery, for the first time. Finally, a mathematical statistician, Karl-Gustav Jöreskog of Uppsala, has begun to look at the "right" problems in factor analysis (from a psychometrician's viewpoint) and those things that we have been doing with such great gusto for so many years are now being annointed with the propriety of sampling distributions, etc.

But, of course, factor analysis is not the only multivariate technique. Generalizations of the t-test and of the analysis of variance have been made to the multivariate case. The ultimate fruitfulness of these approaches is probably yet more or less an unknown quantity. In educational research, undoubtedly the most vigorous activity in the application of the multivariate analysis of variance has been led by Professor Darrel Bock of the University of North Carolina. It will surely be interesting to continue to watch the progress of this provocative area of statistical methodology in education.