PROBABILITY AND STATISTICS IN PSYCHOMETRIC RESEARCH: ITEM ANALYSIS AND CLASSIFICATION TECHNIQUES

HERBERT SOLOMON
TEACHERS COLLEGE, COLUMBIA UNIVERSITY

1. Introduction

As in other disciplines, statistics and probability have a strong impact on the growth and development of the psychological sciences; and the substantive problems of the psychological sciences, in turn, are instrumental in the growth and development of concepts and methodology in statistics and probability. Moreover, this relationship now also exists, to a much smaller degree, between other mathematical disciplines and the psychological sciences. This marriage of the two subjects has certainly been a legitimate affair since Galton's genius operated in the second half of the nineteenth century.

Probability continually looms as an important tool for measurement problems in psychology, and statistics does its share once a probabilistic model has been formulated. This paper was sandwiched between two others[1] whose titles illustrate this point very well. In the previous paper by Mosteller you were invited to look at recent thinking in probabilistic models in learning theory and some of the statistical problems connected with the estimation of the parameters of the models. If I may assume one of the prerogatives of the statistician, I should like to "predict" that you will be treated to a similar repast in the paper by Anderson in connection with "common factor" factor analytic models, one of the earliest attempts at measurement of mental ability. However, the important point to be brought out by the comparison is that in the old sometimes maligned model discussed by Anderson only questions of inference for parameters of the model are analyzed in a statistically sophisticated manner with not much attention to the validity of the model; while in Mosteller's discussion of a rather new measurement device in learning theory the emphasis is on the validity of the model.

At this point one might add a trite statement, namely, that the statistician or probabilist should work closely with the psychologist so that the inference machinery of the statistician, the tools of the probabilist, and the imagination and experience of the psychologist can be jointly brought to bear on the simultaneous construction of both a valid model and the inference devices necessary to demonstrate that the model is tenable on the basis of sample observations. However, who is to determine whether the stat-

Supported in part by the School of Aviation Medicine, Randolph Field, Texas, under Contract AF 18(600)-941.

[1] At the Symposium, this paper followed the paper by Mosteller and preceded the one by Anderson.

istician's emphasis is to be placed on the validity of models or the inference aspects connected with measurement models? The fundamentals of a prospering research system demand that only intellectual motivation be a consideration before statisticians and probabilists are arrayed for battle with psychological models. Obviously both types of activity are necessary and desirable. I should like to add at this point that my position on the program was not attained by design so that I could make this opening sally (friendly as Mosteller and Anderson are to me, and as I hope they will remain after this paper). It just happened and I decided to exploit it.

It would certainly be presumptuous now to make the grandiose statement that I shall discuss the mutual impact of statistics and probability and the psychological sciences. The American Psychological Association has some 15 active chapters and each of these considers itself an aggregate of many areas of inquiry. Some time ago, it was suggested to me that for this Symposium I consider either some topics in psychology which have been mistreated by psychologists through the use of incorrect statistical methods especially when appropriate techniques are available, or some topics for which appropriate techniques are not available, and yet decisions must be made even though one cannot calculate the risks of the decision process. This latter point reminds me of my government and industrial experiences in the field of acceptance sampling where decision rules for acceptance or rejection of manufactured products were intuitively constructed and used, since obviously some procedure was necessary. Quantitative analyses of the risks inherent in the use of some of these rules were later made by statisticians but there is usually quite a lag between the actual use of a decision rule and the measurement of its risks. While the literature of psychology (and other disciplines) probably abounds with situations where Student's t, or chi-square, or the analysis of variance F are used indiscriminately, I will not concern myself with this issue. Accordingly, I prefer to consider the second point raised and choose a specific, though not necessarily small, area of psychology and discuss a few current issues. The topics I choose to discuss fall under two general headings: (1) item analysis and (2) classification techniques. These two topics are, of course, related since the first represents a fundamental study in test theory and is therefore a natural precursor of the second topic which is based on the results of testing. I should like to reiterate that the emphasis for these two topics will be on *unsolved* or *unclear* issues.

2. Item analysis

The subject of item analysis has received considerable attention in psychological and other journals. We can embark on this subject in much the same manner as most authors do by assuming the testing situation to be a rather simple one and then notice that we are rather quickly confronted with intricate if not deep problems. Suppose that a test is composed of K items and the response to a test item is either right or wrong and can thus receive the value one or zero. Then the test score is simply the sum of the number of correct responses. Before going any further let us begin to mathematize the situation. Let X_g be the ability to answer item g, $g = 1, 2, \cdots, K$, and assume item g is answered correctly when $X_g \geqq a_g$. Let $S_1, S_2, \cdots, S_K$ denote the chance variables which are

$$S_g = 1 \text{ if item } g \text{ is answered correctly}$$

$$S_g = 0 \text{ if item } g \text{ is answered incorrectly}$$

and let X_g be normally distributed with zero mean and unit variance (this will cause no loss of generality). Then P_g where

$$P_g = P\{S_g = 1\} = 1 - \Phi(a_g), \tag{1}$$

with

$$\Phi(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{X} e^{-t^2/2} dt, \tag{2}$$

is termed the difficulty of item g. For further theoretical development of item analysis to which we shall soon return the aforementioned description of item difficulty can suffice; however, the estimation of item difficulty can quickly lead to interesting problems.

3. Item difficulty

Suppose there are f possible responses to an item (multiple choice test) and the item is marked zero for an incorrect response and one for a correct response. This means zero can occur in $(f - 1)$ ways. If T is the total number of persons responding to the item and R is the number of persons correctly responding to the item then, in general, R/T may not be a reasonable estimate of P_g as defined above for there may be individuals with correct responses who either from a lack of knowledge guessed correctly or from partial knowledge narrowed down the possible choices and then guessed correctly. On the other hand there may be those who responded incorrectly because of misinformation, but who would have obtained the correct answer otherwise. Obviously, knowledge of the underlying information model is necessary to obtain reasonable estimates of P_g. If the individuals are basing their responses on the tosses of an f-faced die then one should reduce the number of observed correct responses by the average number of incorrect responses over the $(f - 1)$ incorrect categories of response before dividing by the total number of responses or

$$P_g = \frac{1}{T_g}\left(R_g - \frac{W_g}{f-1}\right) \tag{3}$$

where R_g is number of correct answers to the gth item

W_g is number of wrong answers to the gth item

T_g is total number of answers to the gth item.

However, it is unrealistic to assume the choices are equally attractive to the examinees.

An interesting, more realistic, model was proposed by Horst [6] some twenty years ago. Horst assumed that those who do not know the correct answer fall into various subgroups. The first subgroup is composed of those who know nothing about the alternatives in question and distribute their responses equally to all of the f possible answers. A second subgroup is composed of those who know that one of the alternatives is wrong and distributes its answers uniformly over the remaining $(f - 1)$ choices. The sth subgroup is composed of those who know that s of the alternatives are wrong $(0 < s \leqq f - 1)$. Then Horst showed that

$$P_g^* = \frac{1}{T_g}(R_g - W_g^*) \tag{4}$$

where W_g^* is the number of persons selecting the most frequent incorrect answer is a reasonable estimate of P_g.

On the basis of empirical observations one could say that where the correct responses to an item are less than one would expect by chance, then apparently misinformation is operating while conversely better than chance responses indicate a partial information framework. Thus, a correction based on blind guessing can seriously underestimate or overestimate item difficulty when partial information or misinformation respectively are operating. Beyond this statement nothing more exact can be done unless the underlying situation can be quantified. Horst succeeded twenty years ago in formulating a simple probabilistic model of the use of partial information. It seems to me that some effort in formulating other models incorporating partial information or misinformation would be quite useful in the determination of item difficulty. In line with my opening remarks this effort, rather than the tenability of existing models, obviously deserves primary attention in this area.

4. Item parameters and item relationships

Now to return to our general development of the subject. One of the prime purposes of item analysis is to gain as much information as possible about the distribution of test scores resulting from our K item test. Usually this information is desired in terms of the item difficulty P_g and the intercorrelation between any two items g and h, namely ρ^*_{gh}. In the test literature, the K items are usually measuring an underlying trait y (assumed to be normally distributed with zero mean and unit variance) with the understanding that the item intercorrelation matrix has rank one since the items are saturated with the underlying ability only and thus the Spearman single factor model is operating between the items and the underlying trait. A corollary of this model is that for a fixed ability

$$P\{S_g = 1, S_h = 1|y\} = P\{S_g = 1|y\}P\{S_h = 1|y\}\,, \tag{5}$$

or the partial correlation $\rho^*_{gh.y} = 0$. This also leads to $\rho^*_{gh} = \rho^*_{gy}\,\rho^*_{hy}$ and thus relates item intercorrelation to item-ability correlation. We may also write

$$\rho^{*2}_{gy} = \frac{\rho^*_{ge}\rho^*_{gf}}{\rho^*_{ef}} \tag{6}$$

where ρ^*_{ge}, ρ^*_{gf}, and ρ^*_{ef} are any three elements in the item correlation matrix subject only to the restrictions imposed by the subscripts. Since X_g and y have bivariate normal distributions with known covariance matrices, the marginal distribution of X_g given y is normal with mean $(\rho^*_{gy}y)$ and variance $(1 - \rho^{*2}_{gy})$. The joint distribution of $X_1, X_2, \cdots, X_g, \cdots, X_K, y$ can now be written. Since

$$p(X_1, X_2, \cdots, X_K, y) = p(X_1, X_2, \cdots, X_K|y)p(y)\,, \tag{7}$$

thus

$$p\,(X_1, X_2, \cdots, X_K, y) = C \exp\left\{-\frac{1}{2}\left[\sum_{g=1}^{K}\frac{(X_g - \rho^*_{gy}y)^2}{1-\rho^{*2}_{gy}} + y^2\right]\right\} dX_1 dX_2, \cdots, dX_K dy\,, \tag{8}$$

where

$$C = \frac{1}{(2\pi)^{(K+1)/2} \prod_{g=1}^{K} (1 - \rho_{gv}^{*2})^{1/2}}. \tag{9}$$

By means of this terminology we can now discuss what has been labeled the *item characteristic curve* (ICC). This curve expresses the relationship between the probability of a correct item response and a fixed point on the ability continuum. Thus we get

$$\text{ICC} = P\{S_g = 1 \mid y\} = \Phi\left(\frac{\rho_{gv}^{*} y - a_g}{\sqrt{1 - \rho_{gv}^{*2}}}\right). \tag{10}$$

This should not be confused with either the item difficulty P_g or the regression of X_g on y which by the above discussion must be linear. It is interesting to note that in latent structure analysis, Lazarsfeld [8] attempts somewhat the reverse procedure of deriving the distribution of the underlying trait from a known distribution of test scores plus some assumptions about the ICC function; he calls the ICC the *trace line* of the item.

To return to our basic discussion it seems likely that information on $P_{g_1 g_2 \cdots g_n}$ $(n = 1, 2, \cdots, K)$ the proportion of persons answering all of the items $g_1, g_2, \cdots, g_n$ correctly will be necessary to determine the moments of the distribution of test scores. As we shall soon see it will be necessary to determine $P_{g_1 g_2}$ in order to get the interitem covariance from test scores.

Suppose we wish to get the interitem correlation from test scores. Call this ρ_{gh} and notice that it is different from ρ_{gh}^{*} which can be determined from our underlying model. Of course, ρ_{gh} and ρ_{gh}^{*} will be related and this relationship will help explain,[2] as we shall see shortly, an alleged paradox in test theory, apparently first discussed explicitly by Tucker [19], and called the "attenuation paradox" by Loevinger [9]. We know that

$$\rho_{gh} = \frac{P_{gh} - P_g P_h}{\sqrt{P_g (1 - P_g) P_h (1 - P_h)}}, \tag{11}$$

and since P_g and P_h are item difficulties we must still determine P_{gh}. Now

$$P\{S_g = 1, S_h = 1 \mid y\} = \int_{a_g}^{\infty} \int_{a_h}^{\infty} \frac{\exp\left\{-\frac{1}{2}\left[\frac{(X_g - \rho_{gv}^{*} y)^2}{1 - \rho_{gv}^{*2}} + \frac{(X_h - \rho_{hv}^{*} y)^2}{1 - \rho_{hv}^{*2}}\right]\right\}}{2\pi (1 - \rho_{gv}^{*2})^{1/2} (1 - \rho_{hv}^{*2})^{1/2}} \, dx_g dx_h, \tag{12}$$

and thus

$$P_{gh} = P\{S_g = 1, S_h = 1\} = \int_{a_g}^{\infty} \int_{a_h}^{\infty} \frac{\exp\left\{-\frac{1}{2(1 - \rho_{gh}^{*2})}[X_g^2 - 2\rho_{gh}^{*} X_g X_h + X_h^2]\right\}}{2\pi (1 - \rho_{gh}^{*2})^{1/2}} \, dx_g dx_h, \tag{13}$$

where $\rho_{gh}^{*} = \rho_{gv}^{*} \rho_{hv}^{*}$.

Thus ρ_{gh} is a function of ρ_{gh}^{*} and the item difficulties and their relationship can be depicted graphically through the use of the bivariate normal tables in *Tables for Stat-*

[2] Much of the work presented here and related to the attenuation paradox in test theory will be contained in a report on "Statistical Formulation of the Attenuation Paradox in Test Theory" prepared under Air Force Contract AF 18(600)-941 (School of Aviation Medicine) by Dr. Rosedith Sitgreaves.

isticians and Biometricians. When the item difficulties are equal, that is, $P_g = P_h = 1 - \Phi(a)$ and $\sigma^2 = \Phi(a)[1 - \Phi(a)]$ then

$$\rho_{gh} = \frac{P_{gh} - [1 - \Phi(a)]^2}{\Phi(a)[1 - \Phi(a)]}. \tag{14}$$

5. The attenuation paradox

An important study in item analysis centers on the choice of items which optimize two operating characteristics of a test of K items. These two characteristics are the "reliability of a test" and the "validity of a test." The meanings of reliability and validity in test theory are essentially the same as the meanings of precision and accuracy, respectively, in any measuring device. By reliability we usually mean the stability of test scores after many repetitions of the test and measure it by the correlation of two parallel tests. Validity, also measured by a correlation coefficient, is the relationship between the test and ability (the single common factor). The discussion thus far has been a prelude to the consideration of the "attenuation paradox" in test theory. Briefly, several authors have demonstrated that if item intercorrelations and item difficulties are the same for all items, then as item intercorrelations increase, the reliability of a test continually increases but that a point is reached in the process where validity, which is also increasing, reaches a maximum and then decreases. It is difficult to imagine at first why one should construe this to be a paradox since precision and accuracy (or reliability and validity) need have no relationship to each other. However, the underlying single factor model between item and ability can give some justification to the feeling in this instance. The higher the item-ability correlations the higher the validity of the test. But higher item-ability correlation means higher interitem correlation since $\rho^*_{gh} = \rho^*_{gy}\rho^*_{hy}$; and this in turn leads to higher reliability. Since reliability and validity of a test as we shall specifically define them can be written in terms of item intercorrelations and item difficulties, it should be possible to "explain" the paradox just from the mathematical manifestations of the formulations made by researchers in test theory.

The reliability of test scores ρ_{SS} is defined in the following way. Let $S = \sum_{i=1}^{K} S_i$ be the test score, then

$$\rho_{SS} = \frac{K}{K-1}\left[1 - \frac{\sum_{g=1}^{K}\sigma_{gg}}{\sum_{g=1}^{K}\sum_{h=1}^{K}\sigma_{gh}}\right] \tag{15}$$

where $\sigma_{gg} = V(S_g) = P_g(1 - P_g)$ and $\sigma_{gh} = \operatorname{cov}(S_g S_h) = P_{gh} - P_g P_h$. Then we may write

$$\rho_{SS} = \frac{K}{(K-1)}\left\{\frac{\sum_{g=1}^{K}\sum_{\substack{h=1 \\ g \neq h}}^{K}\rho_{gh}\sigma_{gg}^{1/2}\sigma_{hh}^{1/2}}{\sum_{g=1}^{K}\sigma_{gg} + \sum_{g=1}^{K}\sum_{\substack{h=1 \\ g \neq h}}^{K}\rho_{gh}\sigma_{gg}^{1/2}\sigma_{hh}^{1/2}}\right\}. \tag{16}$$

At this point the reliability is a function of number of test items, intercorrelations of items based on test scores, and item difficulties. Now to examine the alleged paradox, let $\rho_{gh} = \rho$; $g, h = 1, 2, \cdots, K$; $\sigma_{gg} = \sigma^2$; $g = 1, 2, \cdots, K$; then

$$\rho_{SS} = \frac{K\rho}{1 + (K-1)\rho}. \tag{17}$$

Thus as the equal item intercorrelations increase, the reliability increases.

Validity is defined as the correlation between S and y; thus we can write

$$\rho_{Sy} = \frac{E\left(y\sum_{g=1}^{K} S_g\right) - E(y)E\left(\sum_{g=1}^{K} S_g\right)}{\sigma_y \sigma_S}. \tag{18}$$

This reduces to

$$\rho_{Sy} = \frac{\sum_{g=1}^{K} E(S_g y)}{\sigma_S}, \tag{19}$$

since $E(y) = 0$ and $\sigma_y = 1$. Now $E(S_g y) = E_y[yE(S_g|y)]$ so that

$$E(S_g y) = \int_{-\infty}^{\infty}\int_{a_g}^{\infty} \frac{y}{2\pi(1-\rho_{gy}^{*2})^{1/2}} \cdot \exp\left[\frac{1}{2(1-\rho_{gh}^{*2})}(X_g^2 - 2\rho_{gy}^{*}X_g y + y^2)\right] dx_g dy \tag{20}$$

or

$$E(S_g y) = \int_{a_g}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot \exp\left[\frac{-X_g^2}{2}\right] dx_g \int_{-\infty}^{\infty} \frac{y}{\sqrt{2\pi}(1-\rho_{gy}^{*2})^{1/2}} \exp\left[\frac{-1}{2(1-\rho_{gy}^{*2})}(y - \rho_{gy}^{*}X_g)^2\right] dy. \tag{21}$$

This reduces to

$$E(S_g y) = \frac{\rho_{gy}^{*} e^{-a_g^2/2}}{\sqrt{2\pi}}, \tag{22}$$

which then leads to

$$\rho_{Sy} = \frac{\sum_{g=1}^{K} \rho_{gy}^{*} e^{-a_g^2/2}}{\sqrt{2\pi}\left(\sum_{g=1}^{K}\sum_{h=1}^{K} \sigma_{gh}\right)^{1/2}}, \tag{23}$$

or

$$\rho_{Sy}^2 = \frac{\sum_{g=1}^{K}\sum_{h=1}^{K} e^{-(a_g^2+a_h^2)/2} \rho_{gy}^{*}\rho_{hy}^{*}}{2\pi\sum_{g=1}^{K}\sum_{h=1}^{K}\sigma_{gh}}. \tag{24}$$

Since from the underlying model $\rho^*_{gy}\,\rho^*_{hy} = \rho^*_{gh}$, we can write

$$\rho^2_{Sy} = \frac{\sum_{g=1}^{K}\sum_{h=1}^{K} e^{-(a_g^2+a_h^2)/2}\rho^*_{gh}}{2\pi\left[\sum_{g=1}^{K}\sigma_{gg} + \sum_{\substack{g=1\\ g\neq h}}^{K}\sum_{h=1}^{K}\rho_{gh}\sigma_{gg}^{1/2}\sigma_{hh}^{1/2}\right]}. \tag{25}$$

Now since the item difficulties and item intercorrelations are equal we write $a_g = a_h = a$, $\rho^*_{gh} = \rho^*$, $\rho_{gh} = \rho$ and thus in this situation the validity ρ_{Sy} is defined by

$$\rho^2_{Sy} = \frac{e^{-a^2}K\rho^*}{2\pi\sigma^2[1+(K-1)\rho]}. \tag{26}$$

Now recall, of course, that ρ is a function of ρ^* and we can thus examine the validity ρ_{Sy} when item difficulties are equal and the interitem correlations increase. Suppose we consider the parameters usually quoted by test theorists in examining the attenuation paradox, namely the item difficulties are all at the fifty per cent level; this means $a = 0$, $\sigma = \frac{1}{2}$ and

$$\rho^2_{Sy} = \frac{2}{\pi}\left[\frac{\rho^*}{1+(K-1)\rho}\right]. \tag{27}$$

Moreover, when $a = 0$, $\sigma = \frac{1}{2}$, we get $\rho = (2/\pi)\arcsin\rho^*$. This relationship is plotted in figure 1 for three values of the item difficulty. Notice that the second factor on the

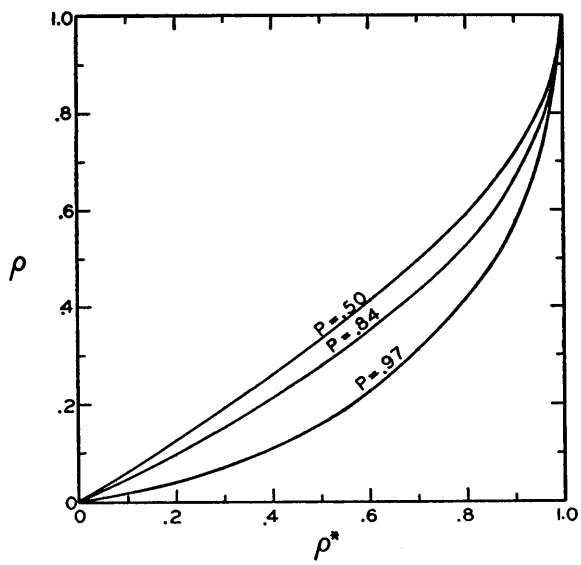

FIGURE 1

Relation between ρ and ρ^*, where ρ is the correlation between two item scores and ρ^* is the correlation between the underlying abilities to answer these items, for three levels of item difficulty.

right-hand side of the equation is very similar to ρ_{SS}; the only difference resulting from the fact ρ^* appears in the numerator of ρ_{Sy}. This is, of course, also true for any a or σ^2. However, it is now possible to demonstrate that certainly ρ_{SS} increases with increasing ρ while ρ_{Sy} will reach a maximum and then decrease except as is obvious for the special

case $K = 1$ when ρ_{Sy} will also be an always increasing function. The point at which the maximum validity will occur can be obtained by differentiating ρ_{Sy} with respect to ρ or ρ^* and setting the resulting equation equal to zero. This leads to

$$1 + (K-1)\,\rho - \frac{2\,(K-1)\,\rho^*}{(1-\rho^{*2})^{1/2}} = 0\,. \tag{28}$$

That is, the maximum value of the validity occurs for that value of ρ^* such that

$$\rho = \frac{2\rho^*}{\pi\,(1-\rho^{*2})^{1/2}} - \frac{1}{K-1}\,. \tag{29}$$

Thus as reliability and validity are defined we do specifically come to a paradox. In figures 2 and 3, the relationship between reliability and validity is portrayed for $P_g = .50$ and $P_g = .84$. Similar graphs appear in Tucker's article [19] since he apparently treats the same probabilistic model but his presentation is rather tedious for the statistician to follow.

In one attempt to explain the paradox Lord [10] essentially states that ρ_{Sy} is not a good index of validity in the region of the paradox. Lord, however, only discusses the curvilinear regression between S_g and y. While $E\,(S \mid y) = \sum_{g=1}^{K} E\,(S_g \mid y)$, it is the regression of test score S on ability y which is important for this discussion.

Now consider the situation for equal item difficulties and equal item intercorrelations and write the conditional distribution of test scores $S = \sum_{g=1}^{K} S_g$ for given ability y. This is

$$P\{S = j \mid y\} = \binom{K}{j}\left[\Phi\left(\frac{\rho^* y - a}{(1-\rho^*)^{1/2}}\right)\right]^j\left[\Phi\left(\frac{a - \rho^* y}{(1-\rho^*)^{1/2}}\right)\right]^{K-j}. \tag{30}$$

Thus the regression of S on y is

$$E\,(S \mid y) = K\Phi\left[\frac{\rho^{*1/2} y - a}{(1-\rho^*)^{1/2}}\right]. \tag{31}$$

This is a curvilinear relationship between S and y whose curvature increases with increasing ρ^*. Thus in the region of the paradox, ρ_{Sy} which stems from a linear regression model and is our usual measure of validity seems to be most ill-fitted for its task and a new measure of validity could be more meaningful and possibly eliminate the paradox. To further explore this issue and possibly eliminate the paradox one could look for a relationship index between S and y which would satisfy the curvilinear regression. One possibility which immediately suggests itself is η^2, the correlation ratio, which is the usually indicated measure of relationship for this type of regression situation.

6. Classification techniques

The problem of classification is an old one in scientific method. Especially in this topic do we find procedures continually proposed with very little analytical evaluation. This does not mean that entirely reasonable procedures cannot be derived on a purely descriptive basis even though it may be mathematically impossible to perform analytical evaluations of them. However, quite a few classification procedures have been proposed and are used that are capable of analytical assessment. At this point, for the purpose of this discussion, I should like to define two subareas which together probably comprise

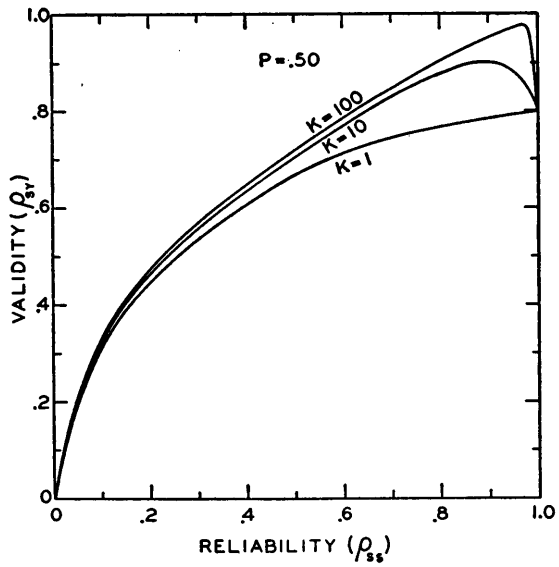

FIGURE 2

Relation between reliability and validity coefficients when all item difficulties are .50 and the interitem correlations are equal. [Note: The reliability coefficient is a monotonically increasing function of the interitem correlation. See equation (16).]

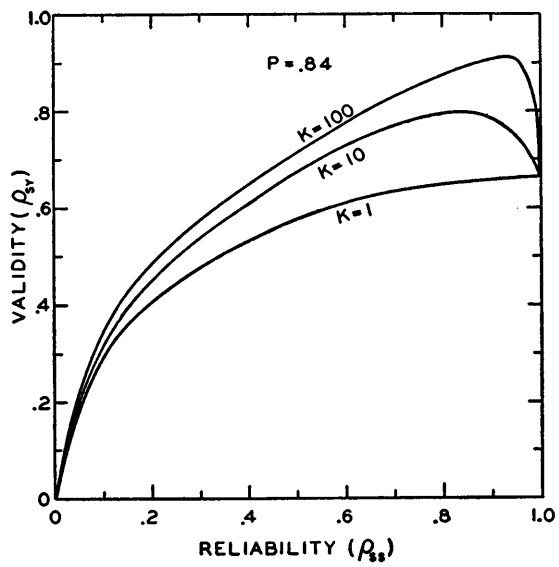

FIGURE 3

Relation between reliability and validity coefficients when all item difficulties are .84 and the interitem correlations are equal. [Note: The reliability coefficient is a monotonically increasing function of the interitem correlation. See equation (16).]

the whole area of classification research. One is the situation where an element is to be assigned to one of k groups where k is known but information on the probability distribution of observables for the groups runs the gamut from complete ignorance of the functional form of the distribution to the point where the functional form and all parameters are known. One part of this problem is the selection of observables which maximize appropriately defined distances among the k groups, a study which has received emphasis in the writings of R. A. Fisher and C. R. Rao. The other situation can be characterized in the same manner except that the value of k is unknown. For the purposes of this exposition, we can refer to the first situation briefly as the "assignment" problem, and the latter situation as the "clustering" problem.

7. The assignment problem

The assignment problem has received the most sophisticated statistical treatment in exactly those instances which are probably most unrealistic. For the situation where an observation is to be classified as coming from one of two multivariate normal populations with unequal means but equal, although unknown, convariance matrices, Wald [20], Anderson [1], and Sitgreaves [14] have made significant contributions to the development of the distribution theory of some classification statistics. This is important for the computation of probabilities of misclassification. The case of two equal but unknown covariance matrices is not too realistic in practice although it does simplify the mathematical treatment.

We can describe the problem briefly as follows. Consider $N_1 + N_2 + 1$ independent p-dimensional chance vectors. The first N_1 vectors are observations from a population Π_1, the following N_2 observations are from population Π_2 and the $(N_1 + N_2 + 1)$st observation is from a population Π where Π is either Π_1 or Π_2. The probability distribution in both Π_1 and Π_2 is multivariate normal with the same covariance matrix Σ; the vector of expected values is $\mu^{(1)}$ in Π_1 and $\mu^{(2)}$ in Π_2. The values of $\mu^{(1)}$, $\mu^{(2)}$, and Σ are not known. Let X denote the $p \cdot (N_1 + N_2 + 1)$ matrix of observations. On the basis of X we want to classify the last observation $X_{N_1+N_2+1}$ as coming from Π_1 or Π_2. When the parameter values are known, the class of Bayes solutions is easily found, resulting in pairs of classification regions of the form

$$T^* \leqq T_0 \quad \text{and} \quad T^* > T_0 , \tag{32}$$

where

$$T^* = X'_{N_1+N_2+1}\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) - \tfrac{1}{2}(\mu^{(1)} + \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) . \tag{33}$$

Both Wald and Anderson proposed the use of classification procedures derived from (33) by substituting estimates for the unknown parameter values. Wald proposed principally the statistic

$$U = X'_{N_1+N_2+1}S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}) , \tag{34}$$

where

$$\bar{X}^{(1)} = \frac{1}{N_1}\sum_{t=1}^{N_1} X_t, \; \bar{X}^{(2)} = \frac{1}{N_2}\sum_{t=N_1+1}^{N_1+N_2} X_t , \tag{35}$$

and

$$(36)\quad S=\frac{1}{N_1+N_2-2}\cdot\left[\sum_{t=1}^{N_1}(X_t-\bar{X}^{(1)})(X_t-\bar{X}^{(1)})'+\sum_{t=N_1+1}^{N_1+N_2}(X_t-\bar{X}^{(2)})(X_t-\bar{X}^{(2)})'\right].$$

Anderson proposed the statistic

$$(37)\quad W=X'_{N_1+N_2+1}S^{-1}(\bar{X}^{(1)}-\bar{X}^{(2)})-\tfrac{1}{2}(\bar{X}^{(1)}+\bar{X}^{(2)})S^{-1}(\bar{X}^{(1)}-\bar{X}^{(2)}).$$

Now it can be shown that $U = kV$, where $V = Y_1A^{-1}Y_2$, the p components of the vectors Y_1 and Y_2 have multivariate normal distributions with different mean vectors but identity covariance matrices, A is a $p \times p$ positive definite matrix with a Wishart distribution involving $N_1 + N_2 - 2$ degrees of freedom, and Y_1, Y_2 and A are independently distributed. We can also write $W = aY_1'A^{-1}Y_2 + bY_2'A^{-1}Y_2$ where a and b are known scalars and Y_1, Y_2 and A are defined as before. In this situation the mean vectors of Y_1 and Y_2 are proportional. When $N_1 = N_2$, $b = 0$; and thus the Wald and Anderson statistics are equivalent. In order to determine the risks, that is, the probability of misclassification, in using say the Anderson statistic W, we must determine the distribution of W. Life will be a little more complicated because the distribution of W contains a nuisance parameter $2\lambda^2$ which is the square of the distance between the centroids of the two multivariate normal populations, that is,

$$(38)\qquad 2\lambda^2=\sum_{i=1}^{p}\sum_{j=1}^{p}\sigma^{ij}(\mu_i^{(1)}-\mu_i^{(2)})(\mu_j^{(1)}-\mu_j^{(2)}).$$

The exact distribution of W is a formidable exercise. When we know all the parameters, it is normally distributed with mean $2\lambda^2$ and variance $4\lambda^2$ if the $(N_1 + N_2 + 1)$st observation is from one population, or normally distributed with mean $-2\lambda^2$ and variance $4\lambda^2$ if the $(N_1 + N_2 + 1)$st observation is from the other population. Thus for large samples we can say it is approximately normally distributed since the estimates used in W are consistent estimates. However, the approach to normality in large samples depends on p as well as $n = N_1 + N_2 - 2$.

An attempt was made to obtain the distribution empirically by machine methods at the Institute for Numerical Analysis, University of California, Los Angeles, for $p = 1(1)15$; $n = 50, 100, 200$; $\lambda = 0(.5)2$. The first four moments were also obtained in this manner. A better glimpse at the approach to normality can then be obtained especially insofar as the value of p influences it. Moreover, approximations of the Cornish-Fisher type can be employed for moderate size n and then compared with the exact results.

Since the distribution of W contains $2\lambda^2$, values of this parameter must be generated in some way so that in a practical situation the probabilities of misclassification can be computed. Since $2\lambda^2$ will usually be unknown, information contained in the sample observations should be used to generate confidence intervals for $2\lambda^2$. If this were done, then probabilities of misclassification could be computed twice, once for each end point of the confidence interval, and if the larger of the two resulting probabilities is relatively small, or if the two are not far apart, a meaningful result is obtained. Where they are too far apart, one can either obtain more observations or decrease the confidence coefficient or do both in order to yield a shorter confidence interval. However, the generation of

confidence intervals is not an easy matter for this situation. When $p = 1$, tables of the noncentral t distribution are necessary, and when $p > 1$ tables of the noncentral F distribution must be available. For the former case, the now classical Johnson-Welch tables [7] can be employed to produce an adequate graphical device for general use. I also understand that rather comprehensive tables of the noncentral t are being computed at the Applied Mathematics and Statistics Laboratory, Stanford University. For the latter situation some tables and graphs of the noncentral F exist [12], [15] but I understand that the Statistical Engineering Laboratory of the National Bureau of Standards is preparing more extensive tables which could aid in producing a graphical device for obtaining confidence intervals when $p > 1$.

Once we leave the rather special case of two multivariate normal populations with equal covariance matrices, sophisticated statistical treatments diminish. Fix and Hodges [3], [4] discuss a rather interesting nonparametric approach to the assignment problem. Their general idea is to assign the observation to Π_1 if most of the nearby observations are from Π_1; otherwise it is assigned to Π_2. Let us simplify matters by supposing the sample sizes are equal. Then if we choose an odd integer r, a specific procedure of their general class of procedures is obtained by assigning the observation to that distribution from which the majority of the r nearest observations came. Their procedure depends on a notion of "distance" in a p-dimensional space in terms of which the $N_1 + N_2$ observations in the combined samples can be ranked according to their "nearness" to $X_{N_1+N_2+1}$. Fix and Hodges [4] evaluate the operating characteristics (probabilities of misclassification) of some special cases falling out of their general procedure but more important, they have opened up some very interesting problems in both the mathematical and quantitative aspects of classification theory.

Von Mises [11] in one of the earliest analytical attempts at classification considered the assignment of an individual into one of k completely specified univariate groups. He considered the risk of equal probabilities of misclassification and demonstrated how to choose classification regions which minimized these equal risks; essentially he obtained the minimax solution for the special case considered. Along the lines of classification into one of k groups, Tiedemann [18] treats an interesting model. His model is as follows. In a given population Π_l, p psychological measurements made on individuals in the population have a joint multivariate normal distribution with means $\mu_1^{(l)}, \mu_2^{(l)}, \cdots, \mu_p^{(l)}$ and covariance matrix $[\sigma_{ij}^{(l)}]$. Two individuals in the population with measurements $(x_1, x_2, \cdots, x_p)$ and $(y_1, y_2, \cdots, y_p)$, respectively, are said to have similar profiles if

$$(39)\quad Q = \sum_{i=1}^{p}\sum_{j=1}^{p} \sigma^{ij(l)}\,(x_i - \mu_i^{(l)})\,(x_j - \mu_j^{(l)}) = \sum_{i=1}^{p}\sum_{j=1}^{p} \sigma^{ij(l)}\,(y_i - \mu_i^{(l)})\,(y_j - \mu_j^{(l)}).$$

The distribution of these quadratic forms in Π_l is a chi-square distribution with p degrees of freedom. For any specified value of Q, Tiedemann defines the corresponding "centour" score as the probability in per cent of exceeding the specified value of Q. Thus when $Q = 0$, the centour score is 100. Consider now k populations $\Pi_1, \Pi_2, \cdots, \Pi_k$ in each of which the p measurements are jointly normally distributed but with different means and different covariance matrices. The actual values of the means, variances, and covariances are unknown but estimated from the samples drawn from the populations. It is desired to assign an observation vector to one of these populations. With the use of the estimated parameters, the centour score for the observed individual is determined for each of the k populations and the individual is assigned to that population

for which he has the highest centour score. This looks like a reasonable thing to do but it would be interesting to assess the costs of misclassification through the use of this classification procedure. Moreover, if one just considers the general k population assignment problem in the simplest case, that is, k completely known multivariate populations (the generalization of the von Mises study), the choice of classification regions, let alone assessment of probabilities of misclassification, is still unexplored.

8. The clustering problem

Especially in the psychological sciences has there been an emphasis on what may be termed the clustering problem. This can be looked at geometrically in the following sense. We are usually given N points in a multi-dimensional space and the problem is to delineate either an optimal number of clusters or a specific number of clusters. This also means identifying the foci and the boundaries of each cluster. The important problem in each situation where this is to be accomplished is the designation of the actual procedure to be used to assign an element to a group. A somewhat well known situation occurs in bi-factor analysis where tests are grouped before the analysis is started. Holzinger [5] proposed in this case the use of a coefficient of belonging, B, which is the ratio of the average correlation between all tests in one group to the average correlation between a test in the group and a test outside the group. Obviously the higher values of B indicate "more" clustering. If the addition of a test to the group causes B to diminish, then the test is not considered part of the group. A big question here is what constitutes a significant drop in B. Obtaining the exact distribution of B is too formidable, yet there is no reason why it should not be used with a decision as to what constitutes stability in B based either on subjective feeling or previous experience. Other procedures have been initiated in this context; for instance, Thomson [16] suggests that the five highest elements in each row of the correlation matrix resulting from the test battery be labeled and then all those tests which have the same five columns labeled be put in one group. This sounds reasonable, as did B, but assessing the risks of the procedure also provides quite a problem.

A typical problem, simplified somewhat for exposition purposes, that may yield to analytical evaluation is as follows. Suppose we consider a psychiatric institute in which each of N patients is subjected to p tests. Assume the test score for the ith test is either zero or one with probabilities Q_i or P_i ($i = 1, 2, \cdots, p$), respectively. Obviously, two individuals having exact scores over all p tests can be considered to have the same test score profile. However, all N individuals will not usually present appropriately delineated profiles. It appears that in this special situation some index of the test scores could be formulated whose risks could be assessed. What might operationally be hoped for in the clustering index would be its selection of k clusters, where within each cluster, for instance, the period of hospitalization required is roughly the same. This is also, of course, the item analysis problem where the responses to the items are dichotomous but a total test score obtained by algebraic summation is not meaningful. This occurs in personality and biographical inventories.

There is another interesting way of looking at this type of problem which has only recently been subjected to analysis. Suppose that instead of the profiles of N individuals based on p tests and obtained at a specific time we consider the profiles of N individuals based on one test (for simplicity) given over p time periods. Here each individual is again represented by p coordinates but the assessment of profile similarity now depends

on time series analysis, a methodology which has received great emphasis in the physical sciences, engineering, and economics. While recent work in spectral analysis provides some means for comparing two time series, a test for the homogeneity of N spectral functions obtained from N time series has not been considered. Moreover, if we consider more than one test for each individual over the p time periods, the problem becomes even more difficult.

Cronbach and Gleser [2] discuss several approaches to assessing similarity between profiles. One interesting phenomenon in their paper seems to be the preference for the Cartesian form to measure the distance between two points in multi-dimensional space rather than the Mahalanobis distance even though the coordinates are referred to a set of oblique axes. In an interesting article, R. L. Thorndike [17] presents an index of clustering based on Cartesian distances. His procedure is as follows. Assume a specific number of clusters 2, 3, 4, $\cdots$ then compute the average within cluster distance where distance between any two points 1 and 2, assuming a p-dimensional framework, is

$$D_{12}^2 = \sum_{j=1}^{p} (X_{j1} - X_{j2})^2 . \tag{40}$$

Then plot as abscissae the number of clusters and as ordinates the average within cluster distance. When this curve sharply approaches a point whose tangent has a slope equal to zero, the optimal number of clusters is reached. This is certainly a reasonable intuitive method but after trying it on some artificially constructed clusters to determine reproducibility of the artificial universe, Thorndike feels it is not too effective. It would be interesting to assess analytically Thorndike's procedure since it is an attempt which minimizes the amount of labor through the use of D^2. For instance, how much correlation can exist between coordinates and still not affect the clustering principle when D^2 is used as a measure of distance? Rao in his book [13] discusses the same procedure for clustering except that distance is measured by

$$D_{12}^{*2} = \sum_{i=1}^{p} \sum_{j=1}^{p} \sigma^{ij} (X_{i1} - X_{i2}) (X_{j1} - X_{j2}) . \tag{41}$$

It would also be interesting to assess analytically this procedure. In general, the statistician might find the clustering problem a fertile area ready for successful exploitation but probably only after much concerted effort.

REFERENCES

[1] T. W. Anderson, "Classification by multivariate analysis," *Psychometrika*, Vol. 16 (1951), pp. 31–50.

[2] L. J. Cronbach and C. Goldine Gleser, "Assessing similarity between profiles," *Psychological Bull.*, Vol. 50 (1953), pp. 456–473.

[3] Evelyn Fix and J. L. Hodges, Jr., "Discriminatory analysis; nonparametric discrimination: consistency properties," USAF School of Aviation Medicine Report No. 4, Randolph Field, Texas, 1951.

[4] ———, "Discriminatory analysis; nonparametric discrimination: small sample performance," USAF School of Aviation Medicine Report No. 11, Randolph Field, Texas, 1952.

[5] K. J. Holzinger and H. H. Harmon, *Factor Analysis*, Chicago, University of Chicago Press, 1941, pp. 23–34.

[6] PAUL HORST, "The difficulty of a multiple choice test item," *Jour. of Educational Psych.*, Vol. 24 (1933), pp. 229–232.

[7] N. L. JOHNSON and B. L. WELCH, "Applications of the non-central t-distribution," *Biometrika*, Vol. 31 (1939), pp. 362–389.

[8] P. F. LAZARSFELD, "The logical and mathematical foundation of latent structure analysis," *Measurement and Prediction*, Princeton, Princeton University Press, 1950, pp. 362–412.

[9] JANE LOEVINGER, "The attenuation paradox in test theory," *Psychological Bull.*, Vol. 51 (1954), pp. 493–504.

[10] FREDERICK LORD, "A theory of test scores and their relation to the trait measured," Princeton, Educational Testing Service, Bulletin 51–13, 1951.

[11] R. VON MISES, "On the classification of observation data into distinct groups," *Annals of Math. Stat.*, Vol. 16 (1945), pp. 68–73.

[12] E. S. PEARSON and N. O. HARTLEY, "Charts of the power function for analysis of variance tests, derived from the non-central F-distribution," *Biometrika*, Vol. 38 (1951), pp. 112–130.

[13] C. R. RAO, *Advanced Statistical Methods in Biometric Research*, New York, John Wiley and Sons, 1952, pp. 362–363.

[14] ROSEDITH SITGREAVES, "On the distribution of two random matrices used in classification procedures," *Annals of Math. Stat.*, Vol. 23 (1952), pp. 263–270.

[15] P. C. TANG, "The power function of the analysis of variance tests with tables and illustrations of their use," *Stat. Res. Memoirs*, Vol. 2 (1938), pp. 126–149.

[16] G. H. THOMSON, *The Factorial Analysis of Human Ability*, 5th ed., New York, Houghton Mifflin Co., 1951, pp. 20–21.

[17] R. L. THORNDIKE, "Who belongs in the family?" *Psychometrika*, Vol. 18 (1953), pp. 267–276.

[18] DAVID TIEDEMANN, "A model for the profile problem," *Proceedings of the Invitational Conference on Testing Problems*, Princeton, Educational Testing Service, 1954, pp. 54–75.

[19] LEDYARD R. TUCKER, "Maximum validity of a test with equivalent items," *Psychometrika*, Vol. 11 (1946), pp. 1–13.

[20] ABRAHAM WALD, "On a statistical problem arising in the classification of an individual into one of two groups," *Annals of Math. Stat.*, Vol. 15 (1944), pp. 145–162.