ETA, DISCO, ODISCO, and F

Guttman, Louis

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Historical Social Research, Vol.14 — 1989 — No. 1, 68-88

ETA, DISCO, ODISCO, AND F

Louis Guttman †*

Abstract: Two coefficients are proposed for measuring the extent of overlap in distributions as a direct function of the variance between the arithmetic means (»disco« and »odisco«). They are designed to answer such questions as: »Given the value of a numerical variable x, to which population should an individual be assigned so that minimum error would be incurred?« This is just the reverse of the question addressed by ANOVA. These coefficients are shown to be analytic in x and they are related to Pearson's eta and Fisher's F. Extensions of these coefficients (designed for univariate, one way discrimination) to k-way and multivariate discriminant analysis and measurement of »interaction« are suggested.

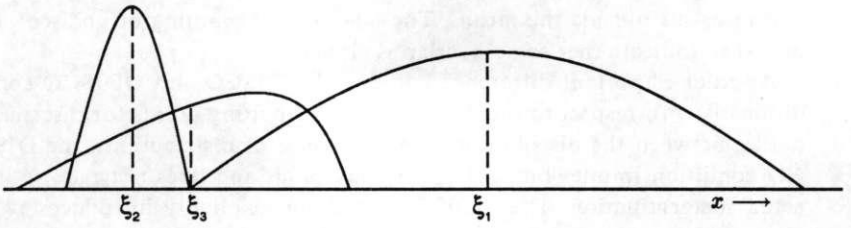
The Problem of Overlap

When two or more populations have distributions on the same numerical variable x, it is of interest to know to what extent these distributions overlap. One motivation for this interest is the problem of discriminant analysis. Suppose an individual has a known value of x, but his/her population is unknown. To which population should he/she be assessed to belong, with minimal expected error? The problem can be illustrated by Figure **1,** for the case of three finite populations. For each value of x there, to what extent can one correctly say to which population the individuals with that value belong? The population means are indicated on the x-axis, labelled £i, £2, and ξ_3, respectively. Overlap is indicated by the crossing of frequency density curves.

There has been no standard loss function for error of misclassification (as for populations **1** and 3, or 2 and 3 in Figure **1).** One popular way for handling this problem is merely to count the number (or proportion) of errors. The present paper is devoted to expressing the loss due to overlap as

* Reprinted by permission of *The Psychometric Society*. Printed in *Psychometrika,* 1988, 53, Nr. 3.

Figure 1: An Example of Three Population Distributions.



a direct function of the variance between the arithmetic means of the distributions.

Two coefficients will be presented for this purpose. One was developed elsewhere (Guttman, 1981); it is not yet very well-known, so it may be helpful to review it here.

The other is new, given here for the first time. The first has been called »disco« for »discrimination coefficient - by the Computer Center of the Hebrew University. The second is called »odisco«; it is more relaxed than disco in a certain sense of overlap. Both coefficients vary between 0 and 1. They equal 0 if there is no difference among the means (which is not necessarily true for coefficients based on counting or other coefficients). Each equals 1 if there is no loss (perfect discrimination holds) in its sense. Each is distribution-free, avoiding traditional assumptions of normality of population distributions and equality of variances within the populations. Such conventional assumptions are unrealistic and misleading in many cases.

The distinction between disco and odisco can be phrased as follows. For each pair of populations a and b, if £b<4«, l[illegible] largest value of x for b be denoted by max(x |b), and let the smallest value of x for a be denoted by min(x | a). Then disco asks whether or not

$$\text{(DISCO condition)} \quad \max(x|b) \leq \min(x|a) \quad (\xi_b < \xi_a).$$

In contrast, odisco asks whether or not two inequalities hold:

$$\text{(ODISCO condition)} \quad \begin{cases} \max(x|b) \leq \xi_a \\ \min(x|a) \geq \xi_b \end{cases},$$

for $\xi_b < \xi_a$. When the DISCO condition holds, then there is no overlap between the distributions (except possibly at a single point). When the ODISCO condition holds, no member of population b has an x-value above ξ_a, and no member of population a has an x-value below ξ_b: there may be overlap in the interval between the two means, but no overlap in the two intervals outside the means. The »**O**« at the beginning of »odisco« is meant to indicate that some overlap is allowed.

Another important difference between the DISCO and ODISCO conditions is with respect to the determination of cutting points for discriminating between the distributions. For each pair of distributions, the DISCO condition implies but a single cutting point and does not require its actual determination. The ODISCO condition explicitly introduces two cutting points, namely the arithmetic means themselves.

Disco is but one member of a relatively new class of coefficients for the discrimination problem (Guttman, 1981). These all equal 1 if there is perfect discrimination, but differ in how they weight error. The present disco is actually a special case of the monotonicity coefficient μ_2 (Raveh, 1978; Guttman, 1986a).

Pearson's Eta and the »Analysis of Variance«

The study of differences among means is typically thought to be a problem of »analysis of variance« (ANOVA). This should not be confused with analysis of discrimination. The ANOVA problem may be regarded as addressing the reverse of the question stated above for discriminant analysis, namely: »Suppose the population of an individual is known, but not his x-value. What is the best predicted value of x?« (cf. Guttman, 1941). The answer - when expected square deviation is taken as the loss-function - is the arithmetic mean of that population. Over all individuals, zero loss occurs only if there is no variation within any of the populations: each distribution is degenerate, being concentrated at but a single point. More generally, the size of the loss is expressed by the variances within the populations - which are traditionally compared with the variance between the means. A standardized coefficient for this comparison purpose is Pearsons's classical correlation ratio »eta« (Pearson, 1905) - or, equivalently, Pearson's pointbiserial r for the special case of only two populations. Eta varies between 0 and 1, equaling 0 when there is no difference among the means and equaling 1 for the degenerate case of no variation within each of the population.

When eta = 1, there is, of course, perfect discrimination. However, the converse is not true; perfect discrimination can exist even when eta is

small. For example, in Figure 1 there is no overlap between the distributions of populations 1 and 2. Perfect discrimination holds even though eta here is less than 1.

For the x-prediction problem of ANOVA, the means and variances enter »naturally«. This is not true for the discrimination problem. Both problems are alike in that the defininition of their respective loss-functions are distribution-free.

While the problems of eta, disco, and odisco refer to the population distributions, only sample data are available in practice. ANOVA conventionally focuses largely on the null hypothesis of no difference between the population means, and usually calculates R. A. Fisher's F statistic instead of Karl Pearson's earlier eta. In a way, this is quite surprising, since F is a simple transformation of eta:

$$F = \frac{(n-m)\eta^2}{(m-1)(1-\eta^2)},$$

where n is the total sample size and m is the number of population distributions. (More generally, (n - m) is the degree of freedom of the denominator, while (m - 1) is the degrees of freedom of the numerator). The tables of probabilities could just as well have been made in terms of eta, making F superfluous. Fisher himself essentially implies this in his discussion of testing eta for »significance« (Fisher, 1950, p. 256). F is a peculiar statistic that estimates no population parameter (which may be why non-mathematical students have trouble with it). Historically, when Fisher introduced ANOVA, he prepared probability tables by use of his z statistic for technical calculation reasons. To be more user-friendly, he later adopted Snedecor's transformation of z into F by the formula: $F = e^{2z}$. But the transformation could have been made just as easily from z directly into eta, which would be even more user-friendly; students and practitioners would not have to learn F any more than the now forgotten $z = \ln(s_1/s_2)$.

In any event, we shall not be concerned here with the testing of null hypotheses and all its problematics (cf. Cowger, 1984; Guttman, 1977; and many others), but rather with consistent estimation. As discussed elsewhere (cf. Muller, 1982, p. 342; Guttman, 1985; Ross, 1985) consistent estimation is necessary and sufficient for cumulative science. As Muller states in a context parallel to ANOVA, »This article will ignore questions associated with significance testing. For purposes of estimation, the focus of this article, only the usual least squares assumptions will be required.« Estimates are always only tentative in cumulative science, to be continually improved upon by further data gathering at other times and places.

Experience has shown that, in practice, eta generally is much less than 1 (which may be a reason for many authors not to publish it with their

ANOVA results, giving only F instead). Similarly, generally neither of the DISCO and ODISCO conditions holds exactly, and one must consider loss functions. Before going on to the algebra of the respective loss functions, it may be useful to look at two numerical examples for showing the different roles played by the various coefficients. The first example is of artificial data, while the second is of actually observed data. Following these presentations, we shall develop the algebra for the coefficients within a systematic framework. It will be shown that eta, disco, and odisco all have the same numerator; they differ only in their denominators, and in such a way that always

(The INEQUALITIES) $odisco \geq disco \geq eta,$

the equalitiy between any two of the coefficients holding if and only if the one on the right equals 0 or 1. The discrimination coefficients can be quite large even when eta is small, as will be illustrated next.

The First (Artificial) Numerical Example

Figure 1 illustrates the theoretical discrimination problem in terms of population distributions. These distributions, of course, are generally not available in practice. To illustrate the discrimination problem in terms of sample data, suppose that a sample P_1 of four individuals is drawn from population 1, a sample P_2 of five is drawn from population 2, and a sample P_3 of five is drawn from population 3. Suppose the respective sample values of the numerical variable x turn out to be as in Table 1, where the sample means, standard deviations, and sizes are as indicated.

Let us begin by comparing the first two distributions. The means of P_1 and P_2 are 54 and 29, respectively. But how much overlap is there between the two distributions? Inspection of the first two columns of Table 1 shows that there is no overlap at all: the smallest value in P_1 is 37, while the largest value in P_2 is 36. Not only is the mean of P_1 larger than that of P_2, but every member of P_1 is larger than every member of P_2: the DISCO condition is fulfilled. Mere inspection of the difference between means cannot reveal the perfect discrimination. Comparing the difference in means with the standard deviations within the distributions is more informative about the overlap, but still leaves something to be desired.

Disco for these first two distributions equals 1. (Since odisco is never less than disco, odisco also equals 1 in this case). Eta (pointbiserial r) is far from 1, having the value .80. Even less informative is Fisher's F, which equals 12.46.

Table 1

An Example of Three Sample Distributions.

P_1	P_2	P_3
59	29	54
72	22	21
48	36	28
37	27	47
	31	35
$\bar{x}_1 = 54$	$\bar{x}_2 = 29$	$\bar{x}_3 = 37$
$s_1 = 13.0$	$s_2 = 4.6$	$s_3 = 12.1$
$n_1 = 4$	$n_2 = 5$	$n_3 = 5$

Now let us compare the first and third distributions. The respective means are 54 and 37. While such a difference may be regarded as substantial, there is nevertheless overlap between distributions P_1 and P_3. Disco equals .84 here. However, the overlap is of a certain limited kind. Every value of P_1 is greater than or equal to the mean of P_3 and every value of P_3 id less than or equal to the mean of P_1: the ODISCO condition is perfectly satisfied. For the data of P_1, and P_3, odisco = 1.00. Thus, odisco supplements disco by assessing the extent to which overlap is bounded by the means. Odisco can equal 1 when disco does not.

Eta and F are even less informative about the situation addressed by odisco. For comparing P_1 with P_3, eta = .56 and F = 3.20.

Proceeding with the remaining comparison of P_2 with P_3, none of the coefficients reaches 1. The respective values are: odisco = .82, disco = .65, eta = .40 (F = 1.53).

By looking more closely at the last two distributions in Table 1, we can see a certain asymmetry in discrimination not revealead even by odisco. While only three of the five values of P_3 are above the mean of P_2, *all* of the values of P_2 are below the mean of P_3. There is perfect one-sided discrimination here in the odisco sense. Odisco itself essentially averages the errors of discrimination of both sides.

Such an averaging process occurs when more than two populations are compared simultaneously. Just as eta (and F) can be computed over all three distributions in Table 1 at the same time, so can disco and odisco. The respective values are: odisco = .98 and disco = .91. Neither of these overall indices equals 1, even though perfect discrimination exists between

some of the pairs. For this reason, the DISCO computer program (part of HUDAP - the Hebrew University Data Analysis Package) gives all the pairwise comparisons as well as the overall coefficients.

Note that the simultaneous estimation of many (and interrelated) coefficients remains consistent. In contrast, simultaneous or stepwise testing of null hypotheses vitiates the traditional calculations of probabilities for F (cf. Muller, 1982, p. 342; Guttman, 1977). Therefore, the DISCO program calculates F along with the other coefficients only for historical reasons, and refrains from printing »probabilities« (or stars).

It may be helpful to review all the numerical coefficient values just cited by putting them into a single table, as in Table 2. Each row of Table 2 shows how eta < disco< odisco (unless one of these equals 1, as in the first row; also unless one equals 0, implying that all equal 0). F cannot be compared directly with these since it has no upper bound. Each column of Table 2 indicates that the overall coefficient in the last row is some weighted average of the corresponding ones in the first three rows.

Table 2

The Values of the Coefficients for the Data of Table 1.

Samples Compared	Eta	Disco	Odisco	F
Pi, Pi	.80	1.00	1.00	12.46
Pi,P₃	.56	.84	1.00	3.20
P2,P₃	.40	.65	.82	1.53
Pl,P2,P3	.70	.91	.98	5.17

The reader may have noticed by now that the sample data in Table 1 could well have been drawn from the population distributions of Figure 1. Inspection of Figure 1 suffices to show that disco = 1 for populations 1 and 2, while odisco = 1 for populations 1 and 3 as well as for 1 and 2. Clearly, if disco = 1 for populations, it must equal 1 for any samples drawn from them. However, even if odisco = 1 for populations, it need not equal 1 in the samples. Conversely, if disco and/or odisco = 1 for samples, this does not necessarily mean that the same must be true in the populations - sampling error has to be considered.

The Second (Empirical) Numerical Example

The second example is of actual data, relating a psychological variable to a medical variable (Rogentine, et al., 1979). The population sampled was of patients diagnosed to have malignant melanoma. The psychological variable was self-assessment by the patients of the amount of adjustment needed to cope with their illness, the scores ranging from 1 to 100. The medical variable was dichotomous: relapse or non-relapse of the illness within one year after the psychological self-assessment. The problem here posed is to estimate the extent to which the numerical psychological variable can discriminate between the two subpopulations defined by the medical variable: relapsers and nonrelapsers.

The authors of this data have shown them in graphic form, as reproduced in Figure 2. The numerical variable is shown here as the vertical axis. The authors performed a number of analyses on aspects of these and related data. Among other things, they noted the discrimination role of the means of the psychological variable in Figure 2: the vast majority of the nonrelapsers are above the mean of the relapsers, while the vast majority of the relapsers are below the mean of the nonrelapsers. This is the type of discrimination addressed by odisco. It is not studied by any of the standard techniques for discriminant analysis: the latter are focussed largely on the problem addressed by disco.

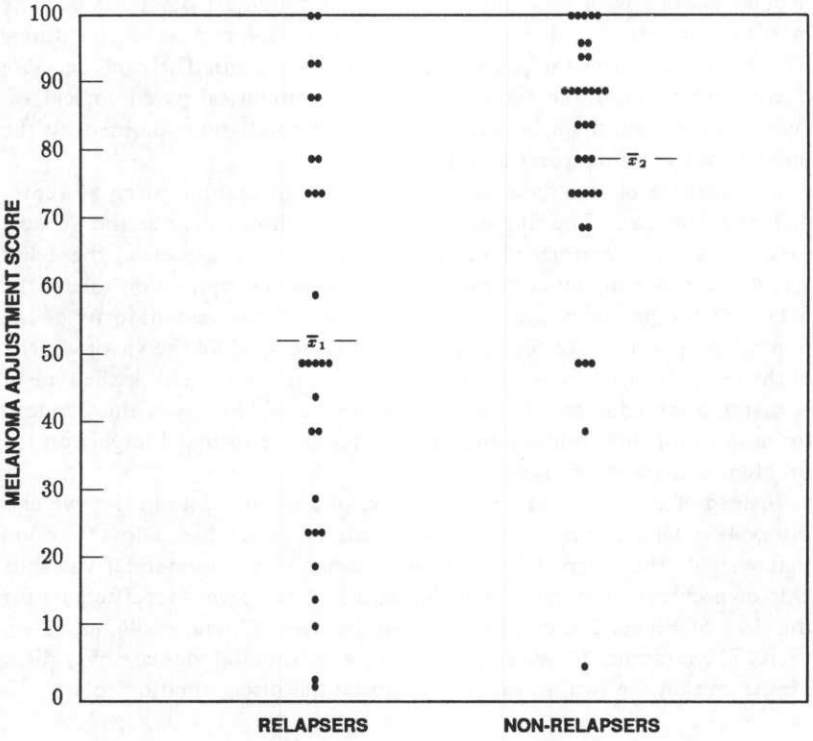
Instead of simply counting the number of aberrant cases in the two end intervals outside the means, as do the authors, odisco uses a loss function that weights the extent of deviation in terms of the numerical variable. Our own calculations show that the values of the several coefficients for the data of Figure 2 are: odisco = .84, disco = .73, eta = .46, and F = 17.10. The large size of odisco here - and the substantial increase over disco indicate that the two means play a substantial discrimination role.

The Algebraic Notation

Each of the sample coefficients eta, disco, and odisco is a consistent estimate of its corresponding population parameter. (In contrast, as remarked above, F estimates no population parameter; it was not devised for descriptive purposes). It will be convenient to develop the algebraic structure of all these coefficients in sample terms. Restatement in population terms is easily done by using expected values in place of mean values throughout.

We consider the general case of m populations, with a sample from each. Let P_a denote the set of sample values from population a (a = 1, 2, ..., m) and let n_a denote the number of individuals in P_a. Individuals will be

Figure 2: Distributions of Melanoma Adjustment Scores for Relapsers and Non-relapsers (from Rogentine, et al., 1979, p.650).



denoted by p, q, ..., etc. If individual p belongs to P_a we write p £ P_a.

The total sample is the union of all the P_a, and will be denoted by P. The number of individuals in the total sample will be denoted by n:

$$n = \sum_{a=1}^{m} n_a.$$

It will also be convenient to use a characteristic function e_{ap} to indicate to which sample p belongs:

$$e_{ap} = \begin{cases} 1, & \text{if } p \in P_a \\ 0, & \text{otherwise.} \end{cases}$$

Each individual belongs to one and only one sample, so for each p ε P:

$$\sum_{a=1}^{m} e_{ap} = 1.$$

Furthermore, for each a,

$$n_a = \sum_{p \in P} e_{ap}.$$

The summation in the right could just as well have been written as over P_a instead of P; indeed this would be more convenient for actual calculations. However, using P here and below is more convenient for the algebraic exposition - emphasizing the overall sample as the point of departure.

The value of the numerical variable x for individual p will be denoted by x_p. The arithmetic mean of x for P_a will be denoted by $\bar{x}_a$:

$$\bar{x}_a = \frac{1}{n_a} \sum_{p \in P} e_{ap} x_p.$$

The total mean over P will be denoted by $\bar{x}$:

$$\bar{x} = \frac{1}{n} \sum_{p \in P} x_p = \frac{1}{n} \sum_{a=1}^{m} n_a \bar{x}_a.$$

The »between« variance for the means will be denoted by s^2_B:

$$s_B^2 = \frac{1}{n} \sum_{a=1}^{m} n_a (\bar{x}_a - \bar{x})^2. \tag{1}$$

An equivalent formula for s^2_B that will be more convenient for the developments below is:

$$s_B^2 = \frac{1}{2n^2}\sum_{a=1}^{m}\sum_{b=1}^{m} n_a n_b (\bar{x}_a - \bar{x}_b)^2. \tag{2}$$

The equivalence is easily established by expanding the right member of (2). Such an equivalence is well-known for variances in general (cf. Kendall, 1943; p. 42). For example, if the total variance of x over all samples, denoted by s^2_x, is defined by

$$s_x^2 = \frac{1}{n}\sum_{p\in P} (x_p - \bar{x})^2, \tag{3}$$

then a formula equivalent to (3) is

$$s_x^2 = \frac{1}{2n^2}\sum_{p\in P}\sum_{q\in P} (x_p - x_q)^2. \tag{4}$$

The Algebraic Structure of Disco

In sample terms, the DISCO condition above can be regarded as asking, for each P_a and P_b: if $\bar{x}_a > \bar{x}_b$, to what extent does this imply that $x_p > x_q$ for each $p \in P_a$ and each $q \in P_b$? Let u_{abpq} be definded by

$$u_{abpq} = e_{ap} e_{bq} (x_p - x_q)(\bar{x}_a - \bar{x}_b). \tag{5}$$

An algebraic condition for perfect discrimination is that the following inequality be satisfied for all p,q,a and b:

$$u_{abpq} \geq 0. \tag{6}$$

According to this inequality, $x_p - x_q$ must have the same sign as $\bar{x}_a - \bar{x}_b$ whenever $p \in P_a$ and $q \in P_b$, unless one of these differences is 0. The inequality does allow the two distributions to meet at a single point.

Let u_{ab} and u be defined respectively by:

(7) $$u_{ab} = \sum_{p \in P} \sum_{q \in P} u_{abpq}$$

(8) $$u = \sum_{a=1}^{m} \sum_{b=1}^{m} u_{ab}.$$

When m = 2, u = $2u_{12}$.

Now, a condition equivalent to inequality (6) is the equality:

(9) $$u_{abpq} = |u_{abpq}|.$$

Let v_{ab} and v be defined respectively by

(10) $$v_{ab} = \sum_{p \in P} \sum_{q \in P} |u_{abpq}|$$

(11) $$v = \sum_{a=1}^{m} \sum_{b=1}^{m} v_{ab}.$$

When m = 2, v = $2v_{12}$.

A necessary and sufficient condition for inequality (6) to hold for all p and q (for fixed a and b) is

(12) $$u_{ab} = v_{ab}.$$

This is so because, if (6) is violated even once by some p and q, then it must be that $u_{ab} < v_{ab}$. Similarly, for equality (12) to hold over all a and b, a necessary and sufficient condition is that

(13) $$u = v.$$

We shall define a loss function for misclassification between P_a and P_b to

be $v_{ab} - u_{ab}$. Similarly, we shall define a loss function for misclassification over all the samples to be v - u. Each of these loss functions is nonnegative, and equals zero if there is perfect discrimination (except possibly at endpoints). The disco coefficient simply restates these loss functions in a standardized form. Thus, for discriminating between P_a and P_b,

$$\text{disco}(P_a, P_b) = \frac{u_{ab}}{v_{ab}}. \tag{14}$$

For overall discrimination among the m samples,

$$\text{disco} = \frac{u}{v}. \tag{15}$$

When m = 2, (14) and (15) are equivalent. In each case, for any m,

$$\text{disco} \leq 1, \tag{16}$$

the equality on the right holding if and only if the corresponding loss is zero.

We now also want to show that disco is always nonnegative:

$$\text{disco} \geq 0. \tag{17}$$

To this end, perform the summations in the right of definition (7) to see that

$$u_{ab} = n_a n_b (\bar{x}_a - \bar{x}_b)^2, \tag{18}$$

whence the numerator in the right of (14) in nonnegative. Since always $v_{ab} \geq 0$, the ratio defining disco in the right of (14) is always nonnegative.

To see the same for u, sum both members of (18) over a and b, and recall (2) and (8) to establish that

(19) $$u = 2n^2 s^2_{\bar{x}}$$

According to (19), the numerator of disco is nonnegative, which establishes (17). An important further consequence of (19) is that disco vanishes if and only if there is no difference among the m means.

To see how disco is a monotonicity coefficient, let us assign to each p a score on a new numerical variable y by the following formula:

(20) $$y_p = \sum_{a=1}^{m} e_{ap}\bar{x}_a.$$

According to (20). if p £ P_a then $y_p = \bar{x}_a$. Instead of p being characterized by P_a only in a qualitative way, it is also characterized by the numerical score $\bar{x}_a$. Each p **8** P now has two numerical values, one on x and one on y. Definition (5) becomes equivalent to

(21) $$u_{abpq} = e_{ap}e_{bq}(x_p - x_q)(y_p - y_q).$$

Summing over subscripts a and b in the right of (21) cancels the two e's. Summing further over p and q yields

(22) $$u = \sum_{p\in P}\sum_{q\in P}(x_p - x_q)(y_p - y_q).$$

Similarly, from the definition of v and (21), it follows that

(23) $$v = \sum_{p\in P}\sum_{q\in P}|x_p - x_q||y_p - y_q|.$$

Hence, u/v or disco, has precisely the structure of monotonicity coefficient ^2 (cf. Raveh, 1978; Guttman, 1986a).

A further interesting feature is that

(24) $$\frac{u}{2n^2} = \operatorname{cov}(x, y),$$

which follows from expanding the right member of (22). Furthermore, it follows from (19), (24), and (2) that

(25) $$s_B^2 = \operatorname{cov}(x, y) = s_y^2.$$

The last equality is established by rewriting the right member of (2) as

$$\frac{1}{2n^2} \sum_{a=1}^{m} \sum_{b=1}^{m} \sum_{p \in P} \sum_{q \in P} e_{ap} e_{bq} (\overline{x}_a - \overline{x}_b)^2$$

and using definition (20).

The Structure of Odisco

Odisco will be developed along lines parallel to disco, but starting with u_{abp} defined by

(26) $$u_{abp} = \sum_{q \in P} u_{abpq}.$$

Summing the right member of (5) over q shows that

(27) $$u_{abp} = e_{ap} n_b (x_p - \overline{x}_b)(\overline{x}_a - \overline{x}_b).$$

This differs from (5) in the first parenthesis on the right. In (27), x_p is compared with the *mean* of the sample of q, and not with x_q itself as in (5). To have odisco = 1 requires both parentheses in the right of (27) to have the same sign - unless one of these vanishes - for all a, b, and p. Analogously to (6) and (9), a necessary and sufficient condition for such perfect discrimination is that, for all a, b, and p,

(28) $$u_{abp} \geq 0,$$

or equivalently,

(29) $$u_{abp} = |u_{abp}|.$$

Summing both members of (27) over p, and recalling (18), shows that

(30) $$u_{ab} = \sum_{p \in P} u_{abp}.$$

Thus, despite the differences in their definition, u_{abpq} and u_{abp} yield the same u_{ab} when summed over the individuals concerned. This will provide the numerator for odisco as well as for disco. The denominator of odisco will, however, be different. If we denote the new denominator by v^*_{ab}, then it is defined by

(31) $$v^*_{ab} = \sum_{p \in P} |u_{abp}|.$$

Note that v^*_{ab} is not necessarily equal to v^*_{ba} (the reader can verify that, for the data of Table 1, $v^*_{23} \neq v^*_{32}$).

Odisco for discriminating between the distributions of P_a and P_b is defined as

(32) $$\text{odisco}(P_a, P_b) = \frac{2u_{ab}}{(v^*_{ab} + v^*_{ba})}.$$

This coefficient varies between 0 and 1. It equals 0 if both means are equal [by virtue of (18)], and equals 1 if - for each sample - all members are on the same side of the mean of the other sample, or the ODISCO condition above is fulfilled exactly for the samples.

To define odisco over all m samples simultaneously, let v* be - analogous to v in (11):

(33) $$v^* = \sum_{a=1}^{m}\sum_{b=1}^{m} v^*_{ab}.$$

Then

(34) $$\text{odisco} = \frac{u}{v^*}.$$

When m = 2, u = $2u_{12}$ (as for disco), whereas $v^* = v^*_{12} + v^*_{21}$.

To show that odisco is never less than disco, we must show that always

(35) $$v^* \leq v.$$

This can be done by comparing the right member of (31) with that of (10). Clearly the former is never greater than the latter, the equality holding if and only if condition (9) holds throughout - or disco = 1. Summing both over a and b establishes (35). Hence, always odisco $\geq$ disco, the equality holding if and only if disco equals 0 or 1.

, The Algebraic Structure of Eta

To help round out the picture, it may be useful to review the structure of eta. In the present notation, Karl Pearson's eta can be defined as

(36) $$\eta = \frac{s_B}{s_x},$$

where s_B and s_x are respective square roots of the variances defined in (1) and (3). As is well-known, an equivalent definition is as a Pearson correlation coefficient r:

(37) $$\eta = r_{xy} = \frac{\text{cov}(x,y)}{s_x s_y},$$

where y is as defined in (20). The equivalence of (36) and (37) follows from (24) and (25). For present purposes, an important apparent difference between (36) and (37) is in their numerators. By virtue of (25), the numerator in (37) is equivalent to s^2_B; however, the numerator in (36) is s_B itself.

Since the numerators of both disco and odisco are equal to $2n^2s^2_B$, these coefficients are more conveniently compared with eta through (37) than through (36). Only the denominators have to be compared. It has shown elsewhere that always disco $\geq$ eta. The proof consists in using the Cauchy-Schwarz inequality in (23) to see that the denominator $v/2n^2$ is never greater than s_1s_2. Since it has already been shown above that odisco $\geq$ disco, this completes the proof of the continued INEQUALITIES asserted in the second section above. Note that the INEQUALITIES involve eta itself and not the square of eta.

Extensions

It has already been remarked that disco is but a special case of a whole family of discrimination coefficients, the difference among the coefficients lying in the way they weight error of misclassification. Similarly, odisco uses but a special case of weighting error as well as in choosing cutting points. Coefficients parallel to odisco can be constructed for pairs of cutting points other than the two means. The present choices were made in order to bring the arithmetic means in explicitly and to have the INEQUALITIES above hold. Other choices do not necessarily revolve about the arithmetic means, nor do they lead to such neat inequalities. More importantly, disco and odisco are analytic in x, facilitating multivariate extensions.

The treatment given above has been for the case of oneway discrimination by a single numerical variable x. Extensions can be made in at least two directions: oneway discrimination for more than one numerical variable, and k-way ANOVA for a single variable x.

For the case of discrimination from more than one numerical variable, a standard approach is to seek an optimal linear function of the variables. Thus, if $x_1, x_2, \ldots, x_k$ are k given numerical variables, the discriminant function x is of the form

$$x = c_1x_1 + c_2x_2 + \cdots + c_kx_k, \tag{38}$$

where the c's are constants to be determined optimally according to a given loss function. For disco, a convenient way is to maximize u/v, where u is

given by (19) and v by (23) (cf. Guttman, 1986a, p. 86). The maximization problem concerns u/v*, which again is nonstandard. It would be desirable to have efficient computer programs for these maximization problems.

Extension to k-way ANOVA leads to concepts other than discrimination, and in particular to »interaction«. It is curious that no coefficient strictly comparable to eta has been proposed for expressing the size of interaction (Hay's **CO**= comes closest to paralleling η^2). At a recent meeting of the Israel Statistical Association (Guttman, 1986b), I showed how to determine upper bounds for interaction, using the absolute value approach (in the spirit of the present paper for establishing upper bounds for s^2_B). Dividing an iteraction variance by its upper bound gives a meaningful coefficient which varies between 0 and 1, being 0 if there is no interaction, and equaling 1 if the condition for the maximum holds.

References

Cowger, C. D. (1984): Statistical significance tests: Scientific ritualism or scientific method? *Social Service Review*, 58, 358 - 372.

Fisher, R. A. (1959): *Statistical Methods for Research Workers.* London: Oliver & Boyd, (11th ed.).

Guttman, L. (1941): *An outline of the statistical theory of prediction.* In: P. Horst, Wallin, P., Guttman, L., Wallin, F. B., Clausen, J. A., Reed, R., & Rosenthal, E. (Contributors): *The Prediction of Personal Adjustment.* New York: Social Science Research Council, cf. pp. 264-268.

Guttman, L. (1977): What is not what in statistics. *The Statistician*, 26, 81-107. Also in: I. Borg (ed., 1981). *Multidimensional Data Representations: When and Why.* Ann Arbor: Mathesis Press, pp. 20-46.

Guttman, L. (1981): Efficacy coefficients for differences among averages. In: I. Borg (ed.). *Multidimensional Data Representations: When and Why.* Ann Arbor: Mathesis Press, pp. 1-10.

Guttman, L. (1985): The illogic of statistical inference for cumulative science. *Applied Stochastic Models and Data Analysis,* 1, 3-10.

Guttman, L. (1986a): Coefficients of polytonicity and monotonicity. In: S. Kotz & Johnson, N. L. (eds.). *Encyclopedia of Statistical Science,* Vol. 7, New York: John Wiley & Sons. pp. 80-86.

Guttman, L. (1986b): Effective analysis of frequency distributions and designed experiments (for cumulative science). Israel Institute of Applied Social Research (stencil).

Kendall, M. G. (1941): *The Advanced Theory of Statistics*, Vol. I, London: Charles Griffin.

Muller, K. E. (1982): Understanding canonical correlation through the general linear model and principal components, *The American Statistician,* 36, 342-354.

Pearson, K. (1905): On the general theory of skew correlation and non-linear regression. *Drapers Company Research Memoirs, Biometric Series II.*

Raveh, A.(1978): Guttman's regression-free coefficients of nomotonicity and polytonicity. In: S. Shye (ed.). *Theory Construction and Data Analysis in the Behavioral Sciences.* San Francisco: Jossey Bass, pp. 387-389.

Rogentine, G. N., van Kampen, D. P., Fox, B. H., Docherty, J. P., Rosenblatt, J. E., Boyd, S. C, & Bunney, W. E. (1979): Psychological factors in the prognosis of malignant melanoma: A prospective study. *Psychosomatic Medicine,* 41, 647-655.

Ross, J. (1985): Misuse of statistics on social sciences. *Nature,* 318, 12.

OBITUARY

Louis Guttman, 1916- 1987

The name of Louis Guttman is most closely associated with the development of scaling theory. But his pioneering and seminal work stretches far beyond the invention of the famous »Guttmann Scaling Method« and his innovations in smallest space analysis or facet theory. He always stressed the point that measurement is not just the assignment and manipulation of numbers but, above all, an exercise in the application and construction of social theory.

The broad scope of his scientific concerns was reflected by the title »Professor of Social and Psychological Assessment at the Hebrew University of Jerusalem (to which he was appointed in 1955) and by the many awards and honors given to him - among them: the Rothschild prize for Social Sciences in Israel; his election to membership in the American Academy of Arts and Sciences; the Outstanding Achievement Award of the University of Minnesota.

Guttman emphasized, in theory and practice, the service function of social theory and research to the public. In his case this meant, above all, the application of social science to the buildung of the new state of Isreal. From 1941 - 1954 he was an expert consultant at the Research Branch of the Information and Education Division of the War Department (see Vol. 4

of the »American Soldier«); later he volunteered his services to the Israel Defense Force and, still later, founded and directed the Israel Institute of Applied Social Research.

Although historians may wonder, why they should be concerned about choosing Eta or rather Disco or Odisco, we feel honored that Louis Guttman submitted this paper (one of his last, as we have to note now) to our journal. His presentation is a beautiful example of how one can go about translating, step by step, a substantive analytical problem into algebraic language. It addresses a problem which is of interest not only to psychologists or sociologists, but to social historians as well:

What is the best way of discriminating populations (e.g., groups of people) which may have overlapping distributions on some variable X; and how, with what amount of error, does one assign individuals to preconceived groups on the basis of their individual measures xi? For example, to what extent are parliamentary groupings (factions) separated on certain dimensions of political ideology revealed by roll call behavior? Although Guttman takes his examples from outside social history, their analytical structure is obviously generalizable to problems in other fields of research.