

A MEASURE OF DIVERGENCE AMONG SEVERAL POPULATIONS

By

William S. Cleveland and Peter A. Lachenbruch

State University of New York, Buffalo
and
University of North Carolina, Chapel Hill, N. C.

Institute of Statistics Mimeo Series No. 838

AUGUST 1972

A Measure of Divergence Among Several Populations

by

William S. Cleveland
(State University of New York, Buffalo)*

and

Peter A. Lachenbruch
(University of North Carolina, Chapel Hill, N.C. 27514)

1. Introduction

Measures of distance between two populations, or more generally of divergence of α populations, are widely used in statistics, for example as inputs to clustering and multidimensional scaling procedures or in the analysis of contingency tables. In this paper a measure of divergence of α populations is defined to be the probability of correctly assigning an item to one of the populations on the basis of a measurement X on the item, using a particular classification rule. An important property of this measure is that it possesses operational meaning by virtue of this classification interpretation. The argument of Goodman and Kruskal (1954, p. 733-735) that operational meaning is important for measures of association applies equally well to measures of divergence of populations.

The following is a brief summary of this paper. It will be supposed except in the last section that X is a categorical variable taking values labeled $1, 2, \dots, \beta$. Let π_a denote the a -th population and let $p_a(b)$ be the probability $X = b$ given the item is from population π_a . In the case where

*Present address: Bell Telephone Laboratories, Inc., Murray Hill, N. J. 07974 U.S.A.

the $p_a(b)$ are known, the measure of divergence is

$$\alpha^{-1} \sum_{b=1}^{\beta} \max_a p_a(b).$$

In the case where the $p_a(b)$ are unknown and a sample is available the measure has the same form but with $p_a(b)$ replaced by an estimate.

A convenient way of displaying all pairwise divergences of the populations is described and the definition of what it means for the populations to be in order is given in terms of this display. Other measures of divergence on the basis of a categorical variable for the case $\alpha=2$ which have appeared in the literature are described. It is argued that the measure of divergence can be used to measure association in a contingency table and comparisons are made with the Guttman-Goodman-Kruskal measures of association. The divergence measure is used in the analysis of three data sets: 1) Voting results in each state for Nixon, Humphrey, and Wallace in the 1968 U.S. Presidential election. 2) The distributions of birth weights of nonwhites and whites in the U.S. in 1950, 1955, 1960, and 1965. 3) Results of intelligence tests given to 5-year old children and their mothers. The paper is concluded by a few remarks about the measure of divergence in the general case where X may have any distribution.

2. The Measure of Divergence When the Distributions of X are Known

Suppose an item is equally likely to come from any one of the α populations π_1, \dots, π_α and the $p_a(b)$ are known. Suppose on the basis of its X measurement the item is to be classified in one of the populations.

A sensible procedure (the optimal one if the goal is to maximize the probability of a correct classification is, if $X = b$, to choose a π_a for which $p_a(b)$ is maximized over a . The measure of classification divergence of the α populations is defined to be the probability the item will be correctly classified.

Let p be the $\alpha \times \beta$ matrix whose (a,b) -th element is $p_a(b)$. Thus the a -th row of p is the distribution of X given the item is from π_a . The probability of a correct classification will be denoted either by $\Gamma(p)$, showing the dependence on the probabilities p , or by $\Gamma(\pi_1, \dots, \pi_\alpha)$, showing which populations are included.

An expression for $\Gamma(p)$ will now be derived. Let a_b , for $b=1, \dots, \beta$, be an index such that

$$p_{a_b} = \max_a p_a(b).$$

It should be noted that, if there are two or more possible values for a_b , $\Gamma(p)$ does not depend upon which value is chosen in the above classification procedure. Now

$$\Gamma(p) = \sum_{a=1}^{\alpha} \text{Prob}[\text{Correct Classification} | \pi_a] \text{Prob}[\pi_a].$$

Since the assumption is that the item is equally likely to come from any one of the populations, $\text{Prob}[\pi_a] = \alpha^{-1}$. Furthermore,

$$\text{Prob}[\text{Correct Classification} | \pi_a] = \sum_{b=1}^{\beta} p_a(b) \langle a_b = a \rangle ,$$

where $\langle a_b = a \rangle$ is equal to 1 if $a_b = a$ and 0 otherwise. Thus

$$\begin{aligned}\Gamma(p) &= \alpha^{-1} \sum_{b=1}^{\beta} \sum_{a=1}^{\alpha} p_a(b) \langle a_b = a \rangle \\ &= \alpha^{-1} \sum_{b=1}^{\beta} p_{a_b}(b) \\ &= \alpha^{-1} \sum_{b=1}^{\beta} \max_a p_a(b) .\end{aligned}$$

The following list of facts sheds some light on the behavior of $\Gamma(p)$.

Fact: The minimum value of $\Gamma(p)$ is α^{-1} . This occurs when X has the same distribution for each population; that is, the rows of p are identical.

Fact: If $\beta \geq \alpha$ the maximum value of $\Gamma(p)$ is 1. This occurs when the α distributions of X given π_a are concentrated on mutually disjoint sets, that is, each column of p has at most one nonzero element. If $\beta < \alpha$ the maximum value of $\Gamma(p)$ is β/α , which occurs when there is a subset of β populations, each having its own single X value which no other population in the subset can take; that is, there is a $\beta \times \beta$ submatrix of p with exactly one 1 in each column.

Fact: If

$$p = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ s & 1-s \end{bmatrix} ,$$

where $0 \leq s \leq 1$, then $\Gamma(p) = 2/3$ does not depend on s . This is a little upsetting. However, this phenomenon of total lack of dependence of $\Gamma(p)$

on the values of some row of p can only occur if $\beta < \alpha$.

Fact: Let p and q be $\alpha \times \beta$ matrices whose rows are distributions of X . Suppose the rows of q lie in the convex region generated by the rows of p . Intuitively one feels that the rows of q diverge less than those of p . In fact, it is true that

$$\Gamma(q) \leq \Gamma(p).$$

To see this note that

$$q_a(b) = \sum_{m=1}^{\alpha} c_{am} p_m(b)$$

where $\sum_{m=1}^{\alpha} c_{am} = 1$ for each $a=1, \dots, \alpha$. Thus

$$q_a(b) \leq \max_a p_a(b)$$

so that

$$\max_a q_a(b) \leq \max_a p_a(b),$$

which verifies the fact.

Fact: If some of the categories of X are merged, the divergence decreases or stays the same since $\max_a p_a(b_1) + \max_a p_a(b_2) \geq \max_a p_a(b_1) + p_a(b_2)$. For example suppose

$$p = \begin{bmatrix} .2 & 0 & .8 & 0 \\ 0 & .8 & .1 & .1 \end{bmatrix}.$$

Suppose the 2nd, 3rd, and 4th categories of X are merged so that the resulting distributions are

$$q = \begin{bmatrix} .2 & .8 \\ 0 & 1 \end{bmatrix} .$$

Then $.5 = \Gamma(p) > \Gamma(q) = .6$. In some cases it will be desirable to use $\Gamma(p)$ directly to measure divergence. In others it may be desirable to relocate and rescale $\Gamma(p)$ so that it ranges from 0 to 1 as p ranges over all possible values for fixed α and β . This new measure is

$$\Delta(p) = \frac{\Gamma(p) - \alpha^{-1}}{\alpha^{-1} \min(\alpha, \beta) - \alpha^{-1}} .$$

$\Delta(p)$ obscures the operational meaning somewhat but may enable more effective comparison of divergences of two groups of populations with different values of α . For $I=2$, $\Delta(p)$ is a metric when the probability distributions of X are viewed as points in Euclidean β -space, since

$$\Delta(p) = \frac{1}{2} \sum_{b=1}^{\beta} |p_1(b) - p_2(b)| .$$

3. The Measure of Divergence When the Population Distributions of X are Unknown

Suppose the true population distributions of X are unknown but the X values and population origins of a sample of items is available. Let $n_a(b)$ be the number of items from the a -th population whose X values are b . Let $n_a(\cdot) = \sum_{b=1}^{\beta} n_a(b)$. $\Gamma(p)$ cannot be used to measure the distance of the populations since the $p_a(b)$ are unknown. What is needed is a measure which is in the same spirit as $\Gamma(p)$ but is based on the sample information. This will be done, first from the sampling theory point of view and then from the Bayesian point of view.

From the sampling theory point of view a sensible procedure would be to choose an estimate of $p_a(b)$ and use $\Gamma(p)$ with the $p_a(b)$ replaced by the estimates, as a sample measure of divergence. An estimate of $p_a(b)$, and the one which will be used in this paper, is the sample frequency

$$f_a(b) = \frac{n_a(b)}{n_a(\cdot)} .$$

Thus, in this case the sample measure of divergence will be

$$\alpha^{-1} \sum_{b=1}^{\beta} \max_a f_a(b) .$$

In some cases, particularly when the counts are low, one should consider adding a flattening constant to the counts. (cf. (Fienberg and Halland, 1972), (Good, 1965, p. 24), (Mosteller, 1968, p. 23)).

It is mildly annoying that we must regard this sample measure of divergence as an estimate of a probability rather than a probability itself. The situation can, however, be remedied by viewing from the Bayesian angle. Suppose your prior distribution (it is an improper one) is proportional to

$$\prod_{b=1}^{\beta} [p_a(b)]^{-1} .$$

Then the posterior is proportional to

$$\prod_{b=1}^{\beta} [p_a(b)]^{n_a(b) - 1} .$$

The probability a future item from population π_a will have $X=b$ is, given the sample, equal to $f_a(b)$. The optimal rule for classifying a future item with $X=b$, if you gain 1 from a correct classification and 0 otherwise and if your prior probabilities are α^{-1} for each π_a , is to choose a population which maximizes $f_a(b)$ over a . Your probability the item will be correctly classified under this rule is the above sample measure of divergence.

4. The Pairwise Divergence Triangle and Populations Which are In Order

For most practical examples it will be informative to calculate, the $\binom{\alpha}{2}$ divergences $\Delta(\pi_i, \pi_j)$ between each distinct pair of populations from the set π_1, \dots, π_α . One way of displaying these numbers puts them in a lower triangular matrix with the population labels along the main diagonal. However a rotation of 45° produces a more congenial (at least to the eye) diagram as in Display 6. We shall refer to this diagram as the pairwise divergence triangle of the populations. The two populations corresponding to a particular entry may be found by moving upward along the two diagonals which meet at that entry. For example, in Display 6 the divergence between π_2 and π_4 is $\Delta(\pi_2, \pi_4) = .40$.

In a number of practical examples we will want to check a particular ordering of the populations, say π_1, \dots, π_α , to see if the pairwise divergences are "in order". By this it is meant that for each $i=1, \dots, \alpha$, $\Delta(\pi_i, \pi_j)$ increases as j runs from $i+1$ to α and increases as j runs from $i-1$ to 1. In terms of the pairwise divergence triangle this means the divergences increase as you move down any one of the 2α diagonals. For example, the pairwise divergence triangles in Displays 4 and 6 are both

in order, whereas the triangle in Display 7 is not in order.

The typical case where one expects the populations to be in order is where the categorical variable represents the grouping of measurements of a numerical variable. This is illustrated in the examples of Section 7.

5. Other Measures of Divergence for the Case $\alpha=2$

Rao (1952, p. 352) suggests a measure of distance which is also based on the probability of a correct classification. However his classification rule is the minimax strategy rather than the rule used here which maximizes the probability of a correct classification. For example, if $\beta=2$ and

$$p = \begin{bmatrix} s & 1-s \\ t & 1-t \end{bmatrix}$$

where $0 \leq t \leq s \leq 1$, then Rao's measure is $\frac{s}{s+t}$, whereas $\Gamma(p) = \frac{1}{2}(s+1-t)$.

Gini (1914-1915) has suggested using

$$\Delta(p) = \frac{1}{2} \sum_{b=1}^{\beta} |p_1(b) - p_2(b)|$$

in contingency tables to measure the distance between two columns (or two rows), where $p_1(b)$ is the conditional distribution in one column and $p_2(b)$, the conditional distribution in the other. It does not appear, however, that Gini discussed the classification interpretation of the measure.

Over the past few years there has been discussion regarding distance

between populations based on gene frequencies by several authors. Balakrishnan and Sangvhi (1968) proposed a measure based on the χ^2 statistic

$$G_s^2 = \sum_{j=1}^r \sum_{k=1}^{s_j+1} (p_{1jk} - p_{2jk})^2 / p_{jk}$$

where j indexes the character and the j -th character has s_j+1 states. Kurzynski (1970) introduced a measure $D_k^2 = (p_1 - p_2)' S^{-1} (p_1 - p_2)$ analogous to the Mahalanobis D^2 . Edwards and Cavalli-Sforza (cf. Edwards (1971)) proposed a measure based on the angular transformation

$$E^2 = \frac{8(1 - \sum \sqrt{p_{1i} p_{2i}})}{(1 + \sum \sqrt{p_{1i}/s})(1 + \sum \sqrt{p_{2i}/s})}$$

If there are several independent loci (categorical variables) they propose adding the E^2 values.

There is considerable difficulty in interpreting these last three measures. We are not able to develop a feeling for a distance between two sets of probabilities calculated in the ways that have been proposed, primarily due to the lack of any operational meaning. D_k^2 is quite subject to major problems if the proportions are far apart. In that case, no sensible estimate of S is obtainable since the individual cells will have different variances. The maximum of E^2 depends on the number of states which makes comparisons difficult.

We would suggest that the measures G_s^2 and E^2 might be improved by modifying their treatment of several categorical variables assumed to be independent. Instead of calculating the distances from the marginals and

summing it would seem more reasonable to estimate cell probabilities by the products of the marginals and treat the computations as one large multinomial.

It has been called to our attention that a measure similar to the one proposed here has recently been developed by Powell, Levene, and Dobzhansky. It is to appear in Evolution in 1973.

6. A Measure of Predictive Association

Suppose A and B are two categorical variables whose joint distribution is denoted

$$r_{ab} = \Pr[A=a, B=b]$$

for $a=1, \dots, \alpha$ and $b=1, \dots, \beta$. Let $r_{.a}$ denote the marginal distribution of A and $r_{.b}$, the marginal of B. Let r be the $\alpha \times \beta$ matrix whose elements are r_{ab} . Let $r[B|A]$ be the $\alpha \times \beta$ matrix whose (a,b) -th element is $r_{ab} \div r_{.a}$. Thus the rows of $r[B|A]$ are the α conditional distributions of B given A.

$\Delta(r[B|A])$ measures the divergence among the α conditional distributions of B given A and therefore may be regarded as a measure of dependence of B on A. It is defined provided no row of r contains all zeros. It ranges between 0 and 1 and is 0 if and only if A and B are independent, that is, $r_{ab} = r_{.a} \cdot r_{.b}$. $\Delta(r[B|A])$ is equal to the Guttman-Goodman-Kruskal (cf. (Goodman and Kruskal, 1954, Section 5.1)) measure of dependence of B on A, if and only if the marginal of B has equal probabilities.

Now

$$\Omega(r) = \frac{1}{2}\Delta(r[A|B]) + \frac{1}{2}\Delta(r[B|A])$$

is a measure of classification association between A and B. The following facts about $\Omega(r)$ give some insight into its behavior.

Fact: $\Omega(r)$ is defined provided no row or no column of r contains all zeros.

Fact: For fixed α and β , $\Omega(r)$ ranges between 0 and 1, as r ranges over all possible distributions.

Fact: $\Omega(r) = 0$ if and only if A and B are independent.

Fact: If $\alpha > \beta$, $\Omega(r) = 1$ if and only if there is exactly one nonzero probability in each row of r (and no column has all zero probabilities).

For example, if

$$r = \begin{bmatrix} 1/3 & 0 \\ 0 & 1/3 \\ 1/3 & 0 \end{bmatrix},$$

$\Omega(r) = 1$. A similar statement holds when $\beta \leq \alpha$.

Let $\Lambda(r)$ denote the Guttman-Goodman-Kruskal (cf. Goodman and Kruskal (1954, Section 5.2)) measure of association between A and B. In our notation, $\Lambda(r) = \frac{1}{2}(\sum_b \max_a r_{ab}^{-\alpha}) / (1 - \alpha^{-1}) + \frac{1}{2}(\sum_a \max_b r_{ab}^{-\beta}) / (1 - \beta^{-1})$. $\Omega(r)$ has the advantage that it is 0 if and only if A and B are independent, which is not true of $\Lambda(r)$. $\Lambda(r)$ sometimes gives results that are at first sight quite surprising. Consider, for example, the table

$$r = \begin{bmatrix} 1-3 \times 10^{-10} & 10^{-10} \\ 10^{-10} & 10^{-10} \end{bmatrix}.$$

Intuitively one might feel there is a high degree of association. Yet $\Lambda(r) = 0$, whereas $\Omega(r) \approx .5$.

One advantage of $\Lambda(r)$, however, is that it is not as sensitive as $\Omega(r)$ to a single observation in a row (or column) with a small number of counts. For example, consider the two contingency tables

$$\begin{bmatrix} 10^{10} & 1 \\ 10^{10} & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 10^{10} & 0 \\ 10^{10} & 1 \end{bmatrix} .$$

For both these tables Λ is zero whereas Ω is zero for the first and .5 for the second.

The lesson in all this (it has been taught many times) is that no single measure of association will do everything one wants it to.

7. Examples

Example 1: 1968 Presidential Election.

The World Almanac (1972, p. 718) gives the number of votes cast in each of the 51 states of the U.S. (including the District of Columbia) in the 1968 Presidential election for each of the three candidates, Humphrey, Nixon, and Wallace. The items will be the voters, the three populations will be the three sets of each candidates voters, and the state in which an item voted will be the categorical variable X. Thus in studying the divergence measures we shall be studying how the candidates differed according to how their vote was distributed among 51 states. It should be noted that these measures are independent of the total number of votes cast for each candidate.

The various divergences are shown in Display 1. The three candidates appear to diverge a moderate amount. Nixon and Humphrey are substantially closer together than either is to Wallace, not a surprising result.

Example 2: Birth Weight Distributions.

Display 2 exhibits birth weight distributions of eight populations in the U.S. which are cross classified according to race and year. The data was obtained from Chase and Byrnes, (1972, p. 40-41).

Display 3 shows six interesting divergences. The divergence $\Delta(\pi_5, \pi_6, \pi_7, \pi_8)$ of the four nonwhite distributions is about six times as large as the divergence $\Delta(\pi_1, \pi_2, \pi_3, \pi_4)$ of the four white distributions. A close look at Display 2 reveals a definite pattern in the non-white distributions; the birth weights are shifting toward the left (getting lower) through time. The divergences between the two populations white and non-white for each of the four years show that, in addition, the nonwhite distributions are shifting away from the white distributions.

Display 4 gives the pairwise divergence triangle for the four non-white populations. It is interesting to note that the populations are in order, that is, the divergences increase as you move down any diagonal. Furthermore, as you move from left to right along any row of the triangle, the values decrease. This is an indication that the change through time in the nonwhite distributions is slowing down.

Example 3: Intelligence Testing of Mothers and Children.

Barber (1972) has reported the results of an I.Q. test given to 92 5-year-old children and an intelligence test given to their mothers. The data is shown in Display 5.

The χ^2 statistic is 28.03, which as a statistical test for the existence of association between the two intelligence tests is near the .05 level of significance. But this seems to tell us little. That we should be suspicious of a significance level based on asymptotic results

when so many low counts occur is almost beside the point. It seems pointless to ask if association exists when we feel so certain that it must. It would only be a question of gathering enough data to prove such existence. The interesting questions would seem to be what is the strength of the association, what is the pattern of the association, and how well does the data measure these two items.

The divergence of the four populations of mothers when the categorical variable X is taken to be the child's test score category is

$$\Delta(\pi_1, \dots, \pi_4) = .24.$$

The divergence of the seven populations of children when the categorical variable X is taken to be the mother's test score category is

$$\Delta(\theta_1, \dots, \theta_7) = .40.$$

Thus the measure of association Ω is .32, but again we should be somewhat skeptical of the precision of this value in view of the low counts in several cells.

Display 6 gives the pairwise divergence triangle for the four mother populations π_1, \dots, π_4 and Display 7 for the seven child populations $\theta_1, \dots, \theta_7$. We would fully expect the true population pairwise divergence triangles to be in order. This in fact is the case in Display 6 but not in Display 7. The pairwise divergences involving either θ_1 or θ_7 are somewhat erratic, presumably due to the small number of counts in each of these populations. Merging θ_1 with θ_2 and θ_6 with θ_7 seemed therefore sensible. The new child populations will be denoted by $\theta'_1, \dots, \theta'_5$.

Display 8 shows the pairwise divergences between the mother populations after the merging. The populations are still in order. As predicted by a Fact in Section 2 each value in this new triangle is less than or equal to the corresponding value in the old triangle in Display 6. Display 9 shows the pairwise divergences between the child populations $\theta'_1, \dots, \theta'_5$. This new triangle is now in order except for one near miss. $.23 = \Delta(\theta'_3, \theta'_4)$ is slightly larger than $.22 = \Delta(\theta'_2, \theta'_4)$. Thus the merging seems justified.

We now have that

$$\Delta(\pi_1, \dots, \pi_4) = \Delta(\theta'_1, \dots, \theta'_5) = .24$$

so that Ω , the measure of association between the mothers' test scores and the children's test scores is .24.

8. The Measure of Divergence for Any Distribution of X

The measure of divergence can easily be extended, in a manner entirely analogous to the categorical case, to allow for any distribution of X. If $d_a(x)$ is the density of X with respect to a measure m , given the item is from the a-th population, then the measure of divergence, Γ , is

$$\alpha^{-1} \int \max_a d_a(x) dm.$$

Suppose $\alpha=2$, and $d_1(x)$ and $d_2(x)$ are multivariate normal densities with the same covariance matrix, then

$$\Delta = 2 \Phi(D/2) - 1$$

where D^2 is the Mahalanobis distance function (cf. Rao (1952, p. 354)) and Φ is the standard normal distribution function. Thus Γ is a monotone function of D^2 in this case.

Suppose $I=2$ and $d_a(x) = \lambda_a \exp - \lambda_a x$ for $a=1,2$. If $\lambda_1 > \lambda_2$ then

$$\Delta = e^{-\lambda_2 c} - e^{-\lambda_1 c}$$

where $c = \log \frac{\lambda_1 \lambda_2}{(\lambda_1 - \lambda_2)}$.

Acknowledgement: One of us was supported by research career development award HD-46344. W. S. Cleveland was supported by the Office of Naval Research.

Display 1. Divergences for Three Candidates in 1968 Presidential Election.

| <u>Populations</u> | <u>Γ</u> | <u>Δ</u> |
|------------------------|----------------------------|----------------------------|
| Humphrey-Nixon-Wallace | .48 | .22 |
| Humphrey-Nixon | .54 | .08 |
| Humphrey-Wallace | .68 | .36 |
| Nixon-Wallace | .69 | .38 |

Display 2. Percentage Distribution of Birth Weights for Eight Populations.

| <u>Populations</u> | <u>Birth Weight (Grams)</u> | | | | | |
|--------------------|-----------------------------|-----------|-----------|-----------|-----------|------------|
| | Below 2001 | 2001-2500 | 2501-3000 | 3001-3500 | 3501-4000 | Above 4000 |
| White | | | | | | |
| (π_1) 1950 | 2.3 | 4.8 | 17.8 | 38.3 | 27.5 | 9.3 |
| (π_2) 1955 | 2.4 | 4.6 | 17.5 | 38.4 | 28.0 | 9.3 |
| (π_3) 1960 | 2.3 | 4.5 | 17.3 | 38.1 | 28.2 | 9.6 |
| (π_4) 1965 | 2.4 | 4.8 | 18.3 | 38.5 | 27.1 | 8.9 |
| Nonwhite | | | | | | |
| (π_5) 1950 | 3.5 | 6.8 | 21.4 | 35.4 | 22.8 | 10.2 |
| (π_6) 1955 | 4.0 | 7.6 | 23.7 | 36.7 | 20.6 | 7.4 |
| (π_7) 1960 | 4.7 | 8.3 | 25.3 | 37.1 | 18.9 | 5.9 |
| (π_8) 1965 | 5.0 | 8.8 | 26.2 | 37.3 | 17.8 | 4.9 |

Display 3. Various Divergences for the Birth Weight Distributions.

| <u>Populations</u> | <u>Γ</u> | <u>Δ</u> |
|--|----------------------------|----------------------------|
| Whites ($\pi_1, \pi_2, \pi_3, \pi_4$) | .255 | .006 |
| Nonwhites ($\pi_5, \pi_6, \pi_7, \pi_8$) | .276 | .034 |
| White and Nonwhite in | | |
| 1950 (π_1, π_5) | .538 | .077 |
| 1955 (π_2, π_6) | .554 | .108 |
| 1960 (π_3, π_7) | .571 | .141 |
| 1965 (π_4, π_8) | .573 | .145 |

Display 4. Pairwise Divergence Triangle for Nonwhite Birth Weight Distributions π_5, \dots, π_8 .

| 1 | 2 | 3 | 4 |
|---|------|------|------|
| | .050 | .032 | .021 |
| | | .082 | .053 |
| | | | .103 |

Display 5. Cross Classification of Child-Mother Scores on Intelligence Tests.

| <u>Children's Scores</u> | <u>Mothers' Scores</u> | | | |
|--------------------------|------------------------|-----------------|-----------------|--------------------|
| | (π_1) Below 60 | (π_2) 60-69 | (π_3) 70-79 | (π_4) Above 79 |
| (θ_1) Below 85 | 3 | 1 | 2 | 0 |
| (θ_2) 85-94 | 1 | 2 | 0 | 1 |
| (θ_3) 95-104 | 5 | 5 | 9 | 3 |
| (θ_4) 105-114 | 2 | 6 | 6 | 3 |
| (θ_5) 115-124 | 2 | 4 | 10 | 9 |
| (θ_6) 125-134 | 0 | 2 | 5 | 6 |
| (θ_7) Above 134 | 0 | 1 | 0 | 4 |

Display 6. Pairwise Divergence Triangle for the Mother Populations
 π_1, \dots, π_4 .

| 1 | 2 | 3 | 4 |
|---|-----|-----|-----|
| | .33 | .24 | .30 |
| | | .35 | .40 |
| | | | .58 |

Display 7. Pairwise Divergence Triangle for the Child Populations
 $\theta_1, \dots, \theta_7$.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|-----|-----|-----|-----|-----|-----|
| | .58 | .41 | .17 | .23 | .10 | .39 |
| | | .27 | .35 | .22 | .32 | .48 |
| | | | .38 | .51 | .33 | .62 |
| | | | | .43 | .60 | .66 |
| | | | | | .51 | .55 |
| | | | | | | .83 |

Display 8. Pairwise Divergence Triangle After Merging for the Mother
Populations π_1, \dots, π_4 .

| 1 | 2 | 3 | 4 |
|---|-----|-----|-----|
| | .31 | .18 | .26 |
| | | .35 | .40 |
| | | | .58 |

Display 9. Pairwise Divergence Triangle After Merging for the Child
Populations $\theta'_1, \dots, \theta'_5$.

| 1 | 2 | 3 | 4 | 5 |
|---|-----|-----|-----|-----|
| | .25 | .17 | .23 | .20 |
| | | .28 | .22 | .38 |
| | | | .46 | .42 |
| | | | | .53 |

References

- Balakrishnan, V. and Sanghvi, L. D., "Distance between populations on the basis of attribute data", Biometrics 24 (1968), 859-865.
- Barber, C. Renate, "A note on the results of intelligence testing during a longitudinal survey", Journal of Biosocial Science 4 (1972), 37-39.
- Chase, Helen C., and Byrnes, Mary E., "Trends in 'Prematurity.' United States: 1950-67", Vital and Health Statistics Analytic Studies (1972), Series 3, Number 15, U.S. Department of Health, Education, and Welfare.
- Edwards, A. W. F., "Distances between populations on the basis of gene frequencies", Biometrics 27 (1971), 873-881.
- Fienberg, Stephen E., and Holland, Paul W., "On the choice of flattening constants for estimating multinomial probabilities", Journal of Multivariate Analysis 2 (1972), 127-134.
- Gini, Corrado, "Di una misura della dissomiglianza tra due gruppi di quantita e delle sue applicazioni allo studio delle relazioni statistiche", Atti Del Reale Istituto Veneto di scienze, lettere ed arti, Tomo LXXIV, Parte seconda (1914-1915), 185-213.
- Good, Irving John, The Estimation of Probabilities (1965), The M.I.T. Press, Cambridge.
- Goodman, Leo A. and Kruskal, William H., "Measures of association for cross-classifications", Journal of the American Statistical Association 49 (1954), 732-764.
- Kurczynski, T. W., "Generalized distance and discrete variable", Biometrics 26 (1970), 525-534.
- Mosteller, Frederick, "Association and estimation in contingency tables", Journal of the American Statistical Association 63 (1968), 1-28.
- Rao, C. R., Advanced Statistical Methods in Biometric Research (1952), Wiley, New York.
- The World Almanac (1972), Luman H. Long, editor, Newspaper Enterprise Association, Inc., New York.