AD NUMBER

ADA800391

NEW LIMITATION CHANGE

TO

Approved for public release, distribution unlimited

FROM

Distribution authorized to U.S. Gov't. agencies and their contractors; Administrative/Operational Use; AUG 1952. Other requests shall be referred to US Air Force Air University, Attn: School of Aviation Medicine, Randolph AFB, TX.

AUTHORITY

AFSAM ltr, 12 May 1970

ADA800391

L-C 647

15933

ADA800391

UNCLASSIFIED

ATI 159 333 (Copies obtainable from ASTIA-DSC)

University of California, Berkeley

Discriminatory Analysis - Nonparametric Discrimination: Small Sample Performance - Project No. 21-49-004 - Project Report

Fix, Evelyn; Hodges, J.L., Jr. Aug'52 43pp. tables, graphs

USAF School of Aviation Medicine, Randolph Air Force Base, Tex., USAF Contr. No. 41(128)-31 (Report No. 11)

Statistical analysis

Mathematics (55)
Statistics (10)

UNCLASSIFIED

UNITED STATES AIR FORCE

AIR UNIVERSITY

School of AVIATION MEDICINE

DISCRIMINATORY ANALYSIS

Nonparametric Discrimination: Small Sample Performance

PROJECT NUMBER 21-49-004
REPORT NUMBER 11

PROJECT REPORT

THIS REPORT CONCERNS . . . .
a classification procedure and its accuracy.

IT IS FOR THE USE OF . . . .
statisticians and personnel responsible for developing classification procedures.

THE APPLICATION FOR THE AIR FORCE IS . . . .
a possible optimum classification procedure wherein the probabilities of possible misclassifications are known under certain conditions.

USAF School of Aviation Medicine, Project No. 21-49-004, Report No. 11.

Discriminatory Analysis—Nonparametric Discrimination: Small Sample Performance.

Evelyn Fix, and J.L. Hodges, Jr., University of California, Berkeley.

40 pp & ill. 11 illus. 27 cm. UNCLASSIFIED

A classification procedure is worked out for the following situation: Two large samples, one from each of two populations, have been observed. An individual of unknown origin is to be classified as belonging to the first population if the majority of a specified odd number of individuals closest to the individual in question belong to the first population. This method has optimum properties when the number of closest individuals is permitted to be very large. For certain cases involving multivariate normal distributions with the same covariance matrix, the probabilities of possible misclassification have been computed and compared with those of the discriminant function method.

1. Mathematical statistics. 2. Selection and classification.

I. Fix, Evelyn. II. Hodges, J.L., Jr. August 1952

USAF School of Aviation Medicine, Project No. 21-49-004, Report No. 11.

Discriminatory Analysis—Nonparametric Discrimination: Small Sample Performance.

Evelyn Fix, and J.L. Hodges, Jr., University of California, Berkeley.

40 pp & ill. 11 illus. 27 cm. UNCLASSIFIED

A classification procedure is worked out for the following situation: Two large samples, one from each of two populations, have been observed. An individual of unknown origin is to be classified as belonging to the first population if the majority of a specified odd number of individuals closest to the individual in question belong to the first population. This method has optimum properties when the number of closest individuals is permitted to be very large. For certain cases involving multivariate normal distributions with the same covariance matrix, the probabilities of possible misclassification have been computed and compared with those of the discriminant function method.

1. Mathematical statistics. 2. Selection and classification.

I. Fix, Evelyn. II. Hodges, J.L., Jr. August 1952

Best Available Copy

USAF School of Aviation Medicine, Project No. 21-49-004, Report No. 11.

Discriminatory Analysis—Nonparametric Discrimination: Small Sample Performance.

Evelyn Fix, and J.L. Hodges, Jr., University of California, Berkeley.

40 pp & ill. 11 illus. 27 cm. UNCLASSIFIED

A classification procedure is worked out for the following situation: Two large samples, one from each of two populations, have been observed. An individual of unknown origin is to be classified as belonging to the first population if the majority of a specified odd number of individuals closest to the individual in question belong to the first population. This method has optimum properties when the number of closest individuals is permitted to be very large. For certain cases involving multivariate normal distributions with the same covariance matrix, the probabilities of possible misclassification have been computed and compared with those of the discriminant function method.

1. Mathematical statistics. 2. Selection and classification.

I. Fix, Evelyn. II. Hodges, J.L., Jr. August 1952

USAF School of Aviation Medicine, Project No. 21-49-004, Report No. 11.

Discriminatory Analysis—Nonparametric Discrimination: Small Sample Performance.

Evelyn Fix, and J.L. Hodges, Jr., University of California, Berkeley.

40 pp & ill. 11 illus. 27 cm. UNCLASSIFIED

A classification procedure is worked out for the following situation: Two large samples, one from each of two populations, have been observed. An individual of unknown origin is to be classified as belonging to the first population if the majority of a specified odd number of individuals closest to the individual in question belong to the first population. This method has optimum properties when the number of closest individuals is permitted to be very large. For certain cases involving multivariate normal distributions with the same covariance matrix, the probabilities of possible misclassification have been computed and compared with those of the discriminant function method.

1. Mathematical statistics. 2. Selection and classification.

I. Fix, Evelyn. II. Hodges, J.L., Jr. August 1952

DISCRIMINATORY ANALYSIS
Nonparametric Discrimination: Small Sample Performance

EVELYN FIX, Ph.D.
J.L. HODGES, Jr., Ph.D.
University of California, Berkeley

Contract No. AF 41(128)—31

PROJECT NUMBER 21-49-004
REPORT NUMBER 11

Air University
USAF SCHOOL OF AVIATION MEDICINE
RANDOLPH FIELD, TEXAS
August 1952

DISCRIMINATORY ANALYSIS

Nonparametric Discrimination: Small Sample Performance

1. Introduction

In an earlier paper [1] concerned with the problem of nonparametric discrimination, the present authors proposed several classes of nonparametric discrimination procedures and proved that these procedures have asymptotic optimum properties for large samples. The ideas and results of [1] are briefly summarized in section 2 for the convenience of the reader.

The present paper is concerned with the performance of some of these procedures where the samples are small. While the large sample optimum properties given in [1] are general, the investigation of small sample properties is necessarily special since small sample performance depends greatly upon a number of variables connected with the underlying distributions assumed. We have examined in detail certain special cases which seemed of interest and have tried to give some indication of the performance in others. The scope of the present study is given in section 3. The results obtained are presented in the remaining sections.

A related paper, "Nonparametric Discrimination: Consistency Properties," was published as Report No. 4 of this project, February 1951.

2. A class of nonparametric discriminators and their large sample properties

In the present section we summarise some of the ideas and results of [1]. Let $X_1, X_2, \cdots, X_m$ be a sample from the p-variate distribution F and let $Y_1, Y_2, \cdots, Y_n$ be a sample from the p-variate distribution G. We do not suppose that F and G are known, nor even that their parametric form is known. Let Z be an observation known to be either from F or from G; our problem is to decide which. To this end, define in the p-dimensional space a notion of "distance," in terms of which the $m + n$ observations in the combined samples can be ranked according to their "nearness" to Z. The general idea of the discrimination procedures of [1] is that Z should be assigned to F if most of the nearby observations are X's; otherwise Z should be assigned to G. To simplify matters, suppose the sample sizes are equal ($m = n$), and select an odd integer k. A specific procedure of the general class is obtained by assigning Z to that distribution from which came the majority of the k nearest observation.

In [1], it was shown that several classes of these nonparametric discriminators have asymptotically optimum performance as m and n tend to infinity, in the sense that the probabilities of misclassification,

$$P_1 = P\{Z \text{ is assigned to } G | Z \text{ came from } F\},$$

$$P_2 = P\{Z \text{ is assigned to } F | Z \text{ came from } G\},$$

tend, as m and n tend to infinity, to the theoretical minimum values which they could have even if F and G were completely known. The results do not require any restrictive assumptions on the form of F and G, or on the definition of nearness which is used.

3. Scope of the present study

The optimum large sample property mentioned above, together with the applicational simplicity of the procedures, suggests that nonparametric discriminators may be useful alternatives to the commonly employed linear discriminant function. The latter is a reasonable procedure if (i) F and G are p-variate normal distributions and (ii) F and G have the same covariance matrix. Many users and also potential users of the linear discriminant function have been disturbed by the apparent and often considerable failure to satisfy conditions (i) and/or (ii) in cases where the procedure has been applied. In the absence of knowledge of the performance of the linear discriminant function under other conditions than (i) and/or (ii), such uneasiness leads to an interest in methods whose theoretical justification is free of these restrictions.

It would not be reasonable, however, to propose an alternative to the linear discriminant function solely on the basis of asymptotic properties. In particular, it is necessary to ask how much discriminating power is lost through the use of a nonparametric procedure when samples are small and when assumptions (i) and (ii) are valid so that the linear discriminant function is appropriate. The answer to this question requires a comparison of the probabilities of error, P_1 and P_2, which result when the linear discriminant function is used with the corresponding probabilities P_1 and P_2 obtained when some alternative discriminating procedure is used.

The number of parameters on which these probabilities of error seem to depend is considerable: (i) the dimensionality p of the observation space (that is, the number of measurements made on each individual), (ii) the $\frac{p(p+1)}{2}$ parameters of the common covariance matrix, (iii) the $2p$ coordinates of the two vector expectations and, finally, (iv) the specification of the distance function used in the nonparametric procedures to order the sample observations according to their nearness to Z.

We may note that the distance function does not need to be a metric although any metric will serve. All that is required is that, of two points u and v, the distance function specify which is closer to a point z. Geometrically, this amounts to establishing for each point z a system of loci, each locus consisting of those points at the same distance from z. For example, if we use Euclidean distance, the loci are just the surfaces of p-dimensional hyperspheres centered at z. As a second example, consider the distance defined by

$$\Delta(x, z) = \max_{i=1}^{p} |x_i - z_i|.$$

Here the locus of points at a given distance from z consists of the surface of a hypercube, centered at z, with faces parallel to the coordinate hyperplanes. The distance $\Delta(x, z)$ has the advantage of being easily computed. It is, incidentally, a metric.

We now observe that the problem can be substantially reduced by considering linear transformations on the observation space. First, it is always possible by such a transformation to insure that F and G will have the identity covariance matrix; that is, that the p transformed measurements are independent in each population, and that each measurement has unit variance. Second, we can put the expectation vector of F at the origin and the expectation vector of G on the positive first axis. Thus, only two parameters are required to specify the transformed populations, namely, p and λ where

λ = E(first coordinate of Y)

= distance between the means of the transformed populations.

It is well known that P_1 and P_2 for the linear discriminant function are unchanged by this transformation. Thus, in so far as the linear discriminant function is concerned, there is no loss of generality.

What about the nonparametric procedures? Associated with each z and each distance from z, there was a locus of points in the original space. We may consider the transformed loci, in the new space, as providing a transformed distance function. Since the totality of possible distance functions in the original space is mapped one-one into the totality in the new space, our transformation loses no generality for the nonparametric procedure either. Therefore, it is sufficient to consider the transformed populations with the two parameters, p and λ.

It is clear that the totality of possible distance functions forms a very large class; in fact, it is not even a parametric class. It is also easy to see that the values of P_1 and P_2 will depend very heavily upon the distance function used. For example, if we use

$$\delta(x, z) = |x_2 - z_2|$$

as distance (remembering that in the transformed populations the expectation vector of F is at the origin and the expectation vector of G is on the positive first axis), we would have no discriminating power at all and $P_1 = P_2 = 1/2$. At the other extreme,

$$\delta'(x, z) = |x_1 - z_1|$$

would give quite good discrimination, even with small samples (see section 4). In using the nonparametric discriminators proposed here, the judgment of the statistician as to the relative importance of the various measurements is of great consequence. In a sense, the linear discriminant function makes great demands on the populations being discriminated but asks of the statistician only a routine (though lengthy) computation--while the nonparametric discriminators which ask little or nothing of the populations demand considerable judgment on the part of the statistician. Of course, this is not a clear cut distinction since, for instance, with the linear discrimination function, judgment is needed to decide whether or not assumptions (i) and (ii) are sufficiently true in the case under consideration to permit its use.

We are now able to define the scope of the present study. Throughout the entire paper we assume that the sizes of the samples taken from each population are equal, $m = n$. Most of the computations have been made using Δ (defined in section 3) as distance function. Also a great part of the work has dealt with the case where Z is assigned to that population from which came the individual of the pooled samples who most closely resembles Z, that is, $k = 1$. The values of $P_1 = P_2$, when Δ is used as distance function, are given in sections 4 and 5 for $p = 1$ and 2; $\lambda = 1, 2, 3$; $n = 1, 2, 3, 4, 5, 10, 20, 50$ and ∞; $k = 1$. In section 6, values of $k > 1$ have been considered. Section 7 has a discussion of the effect of distance function alternative to Δ. A brief investigation for $p > 3$ is reported in section 8.

Unfortunately, we are unable to say how the values of $P_1 = P_2$ obtained here compare with those of the linear discriminant function, since the latter is not yet tabled. A preliminary survey indicated that an adequate treatment of the performance characteristic of the linear discriminant function would require a large computational program. The result would be of great value and interest but was beyond our means at this time. We have given the results in the univariate case (section 4) where it is easily obtained.

4. Univariate case

When $p = 1$, the obvious and natural distance function is ordinary Euclidean distance which in this case coincides with Δ. The linear discriminant function is also greatly simplified, since no matrix computation enters. One simply computes the arithmetic mean of the sample means,

$$\frac{\bar{X} + \bar{Y}}{2},$$

and assigns Z to that population whose sample mean lies on the side of $(\bar{X} + \bar{Y})/2$ as does Z itself. In this case the probabilities of error of the linear discriminant function are easily computed and this we now proceed to do.

From the symmetry of the problem it is clear that $P_1 = P_2$, so it suffices to compute P_1, that is, we assume that Z is distributed according to P. As shown in section 3, we lose no generality by putting $E(X) = 0$, $E(Y) = \lambda > 0$, $\sigma_X^2 = \sigma_Y^2 = 1$. Introduce the new variables

$$U = \bar{Y} - \bar{X}, \qquad V = \bar{X} + \bar{Y} - 2Z$$

where $n\bar{X} = \sum_{i=1}^{n} X_i$, $n\bar{Y} = \sum_{i=1}^{n} Y_i$. Since, as is well known, $\bar{Y} - \bar{X}$ and $\bar{X} + \bar{Y}$ are independent, we see that U and V are independent normal random variables, with

$$E(U) = \lambda, \quad \sigma_U^2 = 2/n, \quad E(V) = \lambda, \quad \sigma_V^2 = 4 + 2/n.$$

Furthermore, an error is committed by the linear discriminant function if and only if

$$\text{(i)} \quad Z > \frac{\bar{X} + \bar{Y}}{2} \quad \text{and} \quad \bar{Y} > \bar{X}$$

or

$$\text{(ii)} \quad Z < \frac{\bar{X} + \bar{Y}}{2} \quad \text{and} \quad Y < \bar{X}.$$

Thus, an error occurs if and only if $UV < 0$. Therefore, it follows that for the linear discriminant function, when $p = 1$,

$$\text{(4.1)} \quad P_1 = P_2 = \left[1 - \Phi\left(-\frac{\sqrt{n}\lambda}{\sqrt{2}}\right)\right] \Phi\left(-\frac{\sqrt{n}\lambda}{\sqrt{2+4n}}\right) + \Phi\left(-\frac{\sqrt{n}\lambda}{\sqrt{2}}\right)\left[1 - \Phi\left(-\frac{\sqrt{n}\lambda}{\sqrt{2+4n}}\right)\right]$$

where

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du.$$

The limiting value for $n = \infty$ is $\Phi(-\lambda/2)$ since with infinite samples the population means become surely known and P_1 is just the probability that Z exceeds $\lambda/2$. Table I gives the values of $P_1 = P_2$ for various values of n and λ. The results are pictured graphically in figures 1 and 2.

Let us now consider nonparametric discriminators. The simplest of these procedures is the one corresponding to $k = 1$ which consists in assigning Z to that population from which came the sample individual nearest to Z. This "rule of nearest neighbor" has considerable elementary intuitive appeal and

Table I

Probability of error, linear discriminant function, univariate normal distributions

n	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
1	.4175	.25[illegible]	.1235
2	.3821	.1999	.0910
3	.3611	.1819	.0826
4	.3472	.1744	.0787
5	.3376	.1[illegible]	.0763
10	.3175	.1646	.0716
20	.3110	.1616	.0692
50	.3094	.1599	.0678
∞	.3085	.1587	.0668

n = size of sample taken from each population

λ = distance between the means of the two populations

Probability of error = $P\{Z \text{ is assigned to } G \mid Z \text{ came from } F\}$

$= P\{Z \text{ is assigned to } F \mid Z \text{ came from } G\}$

(see formula 4.1)

probably corresponds to practice in many situations. For example, it is possible that much medical diagnosis is influenced by the doctor's recollection of the subsequent history of an earlier patient whose symptoms resemble in some way

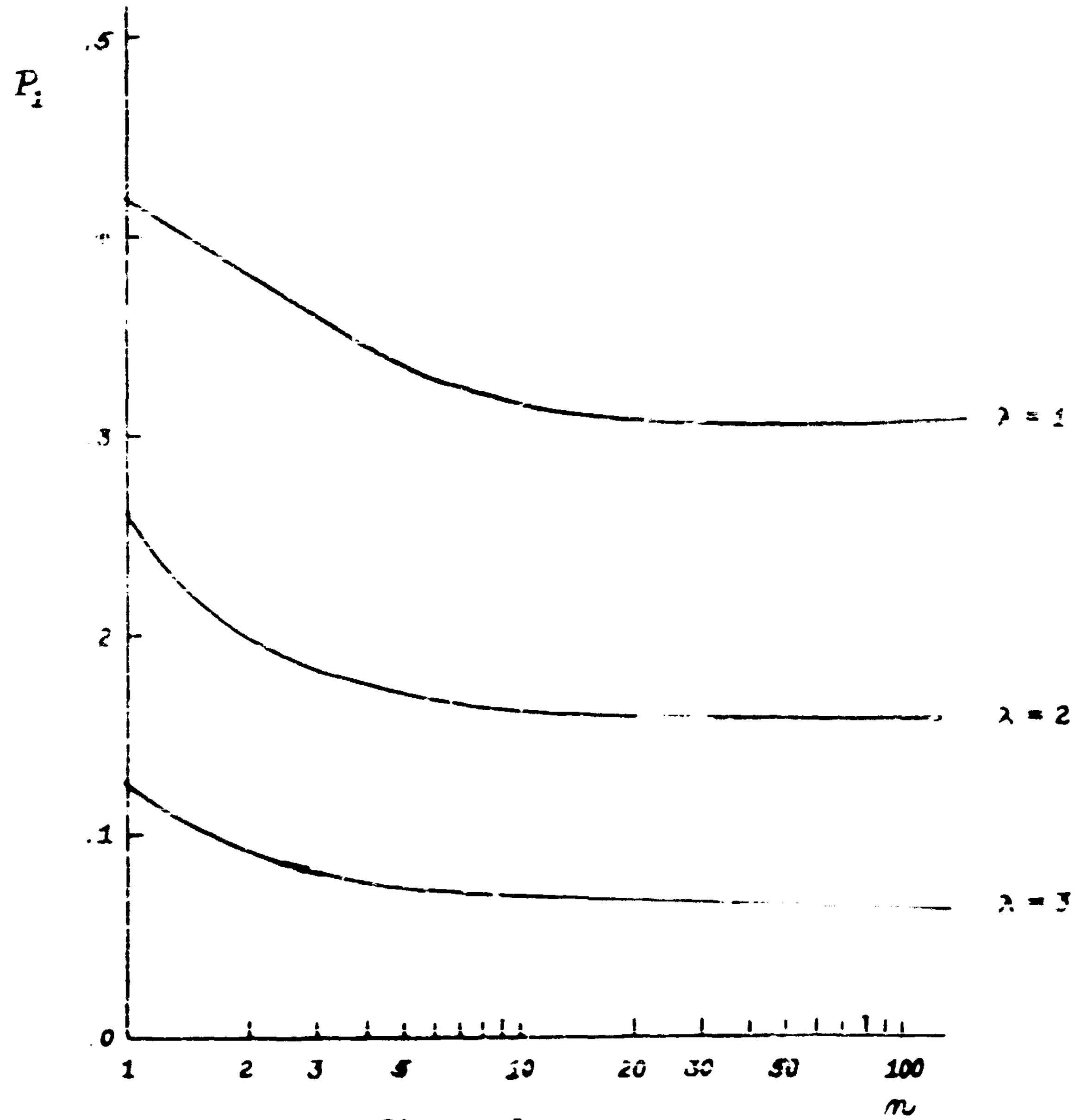

Figure 1

Probability of error P_1 of the linear discriminant function for two univariate normal distributions with distance between means = λ. n = size of sample from each population.

those of the current patient. At any rate it seemed advisable to begin computations with the simplest procedure, that is, to begin with the computation of the probability P_1 that the nearest neighbor to Z is one of the Y's, given that Z has the distribution of an X.

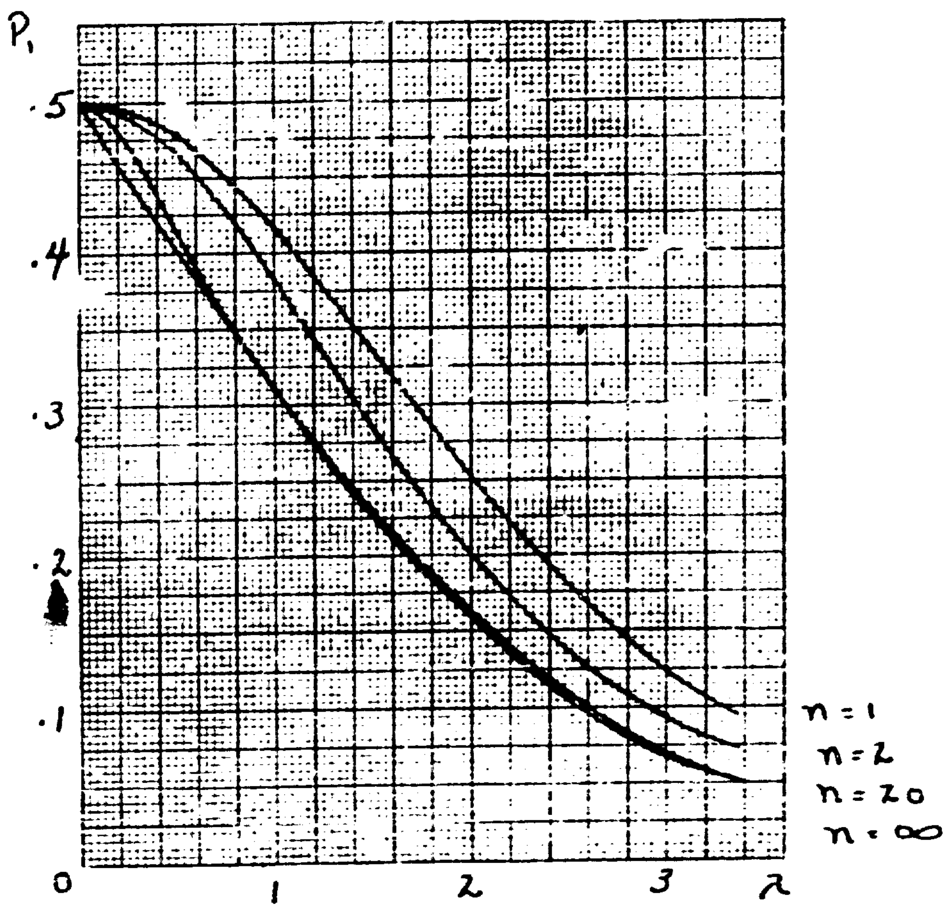

Figure 2

Probability of error P_1 of the linear discriminant function for two univariate normal distributions with distance between the means = λ, plotted as a function of λ. n = size of sample from each population.

Our technique for performing this computation is as follows. Suppose it is given that $Z = z$, and let $P_1(z)$ denote the conditional probability that the nearest of the $2n$ sample observations to Z is a Y, given that $Z = z$. Then

$$(4.2) \qquad P_1 = E[P_1(Z)] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} z^2} P_1(z)\, dz.$$

The calculation of $P_1(z)$ is quite straight forward. Let

$$(4.3) \qquad H_z(\delta) = P\{|X - z| < \delta\}$$

$$= P\{z - \delta < X < z + \delta\}$$

$$= \Phi(z + \delta) - \Phi(z - \delta),$$

while

$$(4.4) \qquad K_z(\delta) = P\{|Y - z| < \delta\}$$

$$= P\{z - \lambda - \delta < Y - \lambda < z - \lambda + \delta\}$$

$$= \Phi(z - \lambda + \delta) - \Phi(z - \lambda - \delta).$$

The event, "the nearest sample value to z is a Y" may be classified into the n exclusive events, "the nearest sample value to z is Y_i", $i = 1, 2, \cdots, n$. By symmetry these n events are equiprobable. The event, "the nearest sample value to z is Y_1" may be broken down according to the distance from z to Y_1. Thus,

$$(4.5) \quad P_1(z) = n \int_0^{\infty} [1 - H_z(\delta)]^n [1 - K_z(\delta)]^{n-1} \, dK_z(\delta).$$

Formulae (4.2) and (4.5) are the basis of all our computations for the "nearest neighbor rule," no matter what the value of p. If $p > 1$, $H_z(\delta)$ and $K_z(\delta)$ are not, of course, given by the explicit formulae (4.3) and (4.4). Their definition is analogous if one replaces $P\{|X - z| < \delta\}$ by $P\{\text{the distance of } X \text{ from } z < \delta\}$ in (4.3) and similarly $P\{|Y - z| < \delta\}$ by $P\{\text{the distance of } Y \text{ from } z < \delta\}$ in (4.4). The specific evaluation depends then upon the distance function used.

Aside from the case $p = 1$, $n = 1$, which is given explicitly by formula (4.1) with $n = 1$, the bulk of the computation was carried out by straightforward numerical integration. For $p = 1$,

$$dK_z(\delta) = \frac{1}{\sqrt{2\pi}} \left[e^{-\frac{(z-\lambda+\delta)^2}{2}} + e^{-\frac{(z-\lambda-\delta)^2}{2}} \right] d\delta .$$

The values of $H_z(\delta)$, $K_z(\delta)$ and $dK_z(\delta)$ were taken from tables [2] and [3]. In the calculation of $P_1(z)$ the fineness of the mesh and the quadrature rule used depended to some extent on the location of z. After the values of $P_1(z)$ had been obtained, a final quadrature (4.2) was effected to obtain the value of P_1. The results given in table II were computed in this way.

The computations that led to the values recorded in table II are quite heavy. This is especially true in the bivariate case, $p = 2$, with which we began computations. Therefore a search for a simple and sufficiently accurate approximate method was instituted. Of the numerous approximate formulae tried, the following was the most successful. Let δ denote the distance from z at which the nearest sample value lies. The conditional value of $P_1(z)$, given δ, may be seen to be

$$\frac{\dfrac{dK_z(\delta)}{1 - K_z(\delta)}}{\dfrac{dK_z(\delta)}{1-K_z(\delta)} + \dfrac{dH_z(\delta)}{1-H_z(\delta)}} = q(z, \delta). \qquad (4.6)$$

It is notable that $q(z, \delta)$ is independent of n. The idea of the approximation is that $P_1(z)$ may be replaced by its conditional value, $q(z, \delta^*)$ where δ^* is in some reasonable sense an average value of δ. In order that $q(z, \delta)$ be an adequate replacement for $P_1(z)$, it is clear that δ^* will be a monotonic decreasing function of n. The function of δ^* which served best was arrived at by treating the n observations from each population as a pooled sample of size $2n$. An average value of δ was thought to be one which would make the probability that at least one of the combined sample values would fall within the prescribed δ distance of z equal to the probability that a sample value would fall outside this prescribed distance. The value of

δ^* for a given n was then chosen to satisfy the following equation:

$$(4.7)\quad \left[\frac{\{1-H_z(\delta^*)\} + \{1-K_z(\delta^*)\}}{2}\right]^{2n} = e^{-\frac{\{1-H_z(\delta^*)\} + \{1-K_z(\delta^*)\}}{2}}$$

It was found easier to solve the above equation for the value of n, say n^*, corresponding to a given value of δ. Then, if $q(z, \delta)$ is regarded as a function of n^*, the value of $q(z, \delta)$ corresponding to a given n can be found by interpolation, using Aitken's method. Table II was extended to larger values of n in this way and the results are shown in table II-A. Figure 3 is based on the combined data of tables II and II-A.

The approximation by means of (4.6) and (4.7) was developed specifically for the bivariate case and it appears to be a better approximation for small n under these conditions than in the univariate case. Time permitted us to make only a limited search for an approximation which would be more satisfactory for the univariate normal distributions. It may be of some interest to give the first terms of the expansion of (4.5). We are indebted to Mr. T. A. Jeeves of this Laboratory for bringing this expansion to our attention. In this connection, see [4] and [5].

Table II
Probability of error, nonparametric discriminator
with k = 1, univariate normal distribution

n	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
1	.4175	.2532	.1235
2	.4086	.2364	.1084
3	.4052	.2307	.1036
4	.4032	.2280	.1014

Table II-A
Approximate probability of error, nonparametric discriminator
with k = 1, univariate normal distribution

n	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
4	.403	.226	.102
5	.401	.225	.100
10	.399	.223	.098
20	.398	.224	.098
50	.398	.225	.098
∞	.398	.225	.098

n = size of sample from each population

λ = distance between the means of the two populations

k = odd integer such that Z is assigned to that population from which came the majority of the k nearest observations-- k = 1 is the "rule of nearest neighbor."

$$\text{Probability of error} = P\{Z \text{ is assigned to } G \mid Z \text{ came from } F\}$$
$$= P\{Z \text{ is assigned to } G \mid Z \text{ came from } G\}$$

(see formulae 4.2 - 4.5)

Distance function = $\Delta(x,z) = |x - z|$

$$P_1(z) = \frac{dK_z(0)}{dH_z(0)+dK_z(0)}$$

$$- \frac{1}{n} \frac{dH_z(0)dK_z(0)\left[dH_z(0)-dK_z(0)\right]}{\left[dH_z(0)+dK_z(0)\right]^2}$$

$$+ \frac{1}{n^2\left[dH_z(0)+dK_z(0)\right]^3} \left\{ \frac{d^3K_z(0)dH_z(0)-dK_z(0)d^3H_z(0)}{dH_z(0)+dK_z(0)} \right.$$

$$\left. + \frac{dH_z(0)dK_z(0)\left[dH_z(0)-dK_z(0)\right]\left[(dH_z(0))^2-4dH_z(0)dK_z(0)+(dK_z(0))^2\right]}{\left[dH_z(0)+dK_z(0)\right]^2} \right\}$$

$$+ O(n^{-3}).$$

The limiting value for $n \to \infty$ may be approached in another way. When n is large, δ will be small, so that in the limit, $P_1(z)$ will simply be $q(z,0) = \dfrac{dK_z(0)}{dK_z(0)+dH_z(0)} = \dfrac{g(z)}{f(z)+g(z)}$, where f and g are the density functions corresponding to F and G, respectively. This argument is quite general: for large n,

$$P_1 \cong E\left[\frac{g(Z)}{g(Z)+f(Z)}\right] = \int_{-\infty}^{\infty} \frac{f(z)g(z)}{f(z)+g(z)}\, dz.$$

A simple application of Schwartz's inequality shows the latter integral to be at most 1/2. We can thus assert that, whatever be the populations being discriminated, the "rule of nearest neighbor" will have in the limit, as $m = n \rightarrow \infty$, equal probabilities of error at most 1/2. While this remark is of no practical interest, it is theoretically interesting because the "optimum" maximum likelihood rule, "assign Z to that population with the larger density at z," possesses no such nontrivial general bound on the individual probabilities of error.

The easiest and most vivid method of comparing the figures of tables I, II and II-A is graphically. Therefore, in figure 4, the probabilities of misclassification for paired values of λ are plotted against n while figure 5 shows the same values plotted this time against λ for selected values of n. It seems needless to discuss the graphs at length since in any practical case the experimenter must make up his mind whether or not the simplicity of operation given by the nonparametric discriminator makes up for the loss of efficiency. In the univariate case the question seems somewhat pointless since the linear discriminant function is easy to compute and also it is little work to derive its performance characteristic. The univariate investigation was undertaken for the sake of completeness of presentation and because it provides a simple case on which to illustrate the use on nonparametric discriminators.

Next to the "rule of nearest neighbor," the simplest nonparametric discriminator is obtained by setting $k = 3$ and using the "rule of two out of three," that is, assign

Z to that population from which came the majority of the nearest three observations in the pooled samples. For finite n, the problem of misclassification reduces to the following.

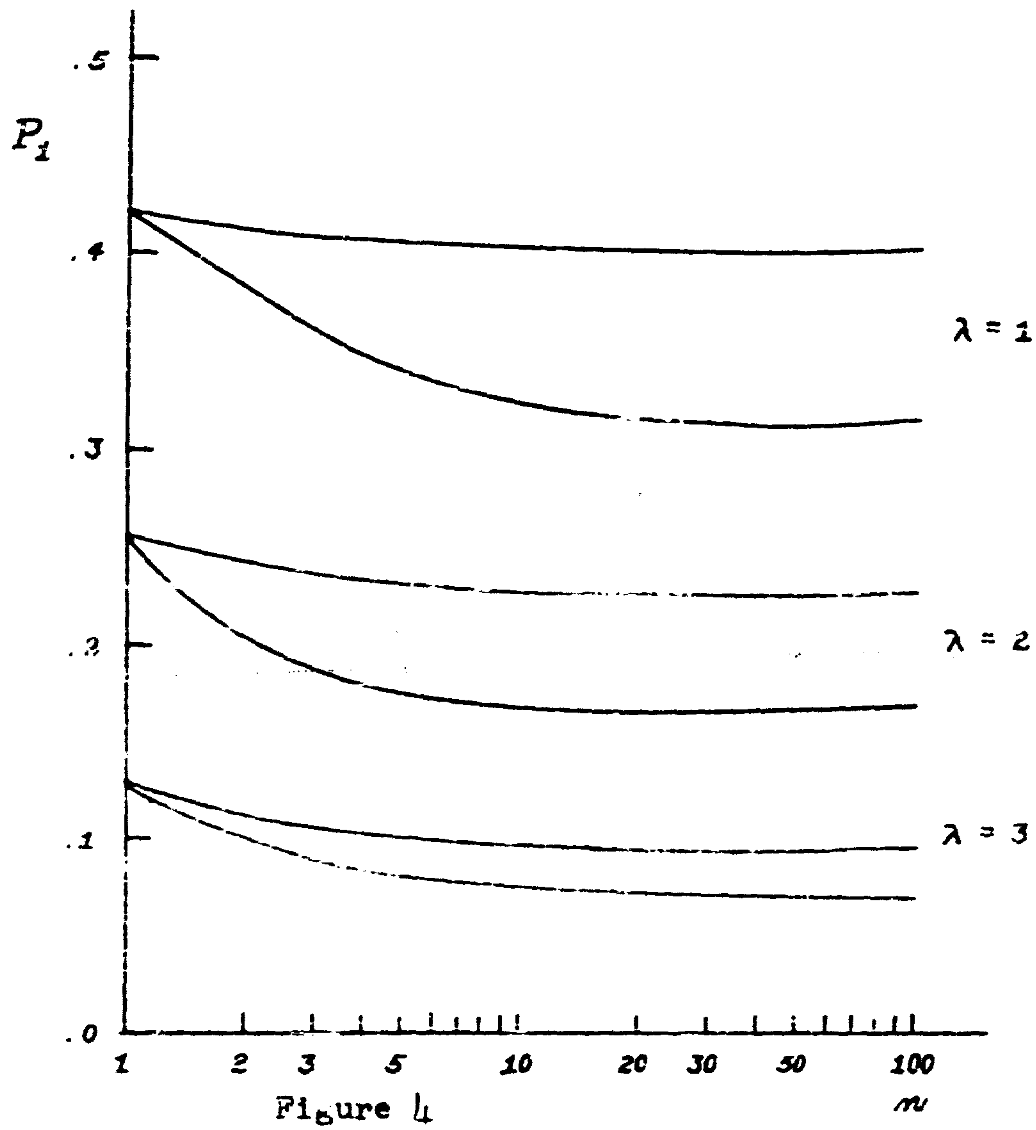

Figure 4

Comparison of the probability of error P_1 as a function of n for the linear discriminant function and the non-parametric discriminator, distance function $= \Delta$, $k = 1$, for two normal univariate populations with distance between means $= \lambda$. n = size of sample from each population.

Let X_1, X_2, X_3 denote the values obtained from F and Y_1, Y_2, Y_3 the values from G. Then the conditional probability that two of the three values nearest to Z will be Y's given that Z belongs to F and $Z = z$ is

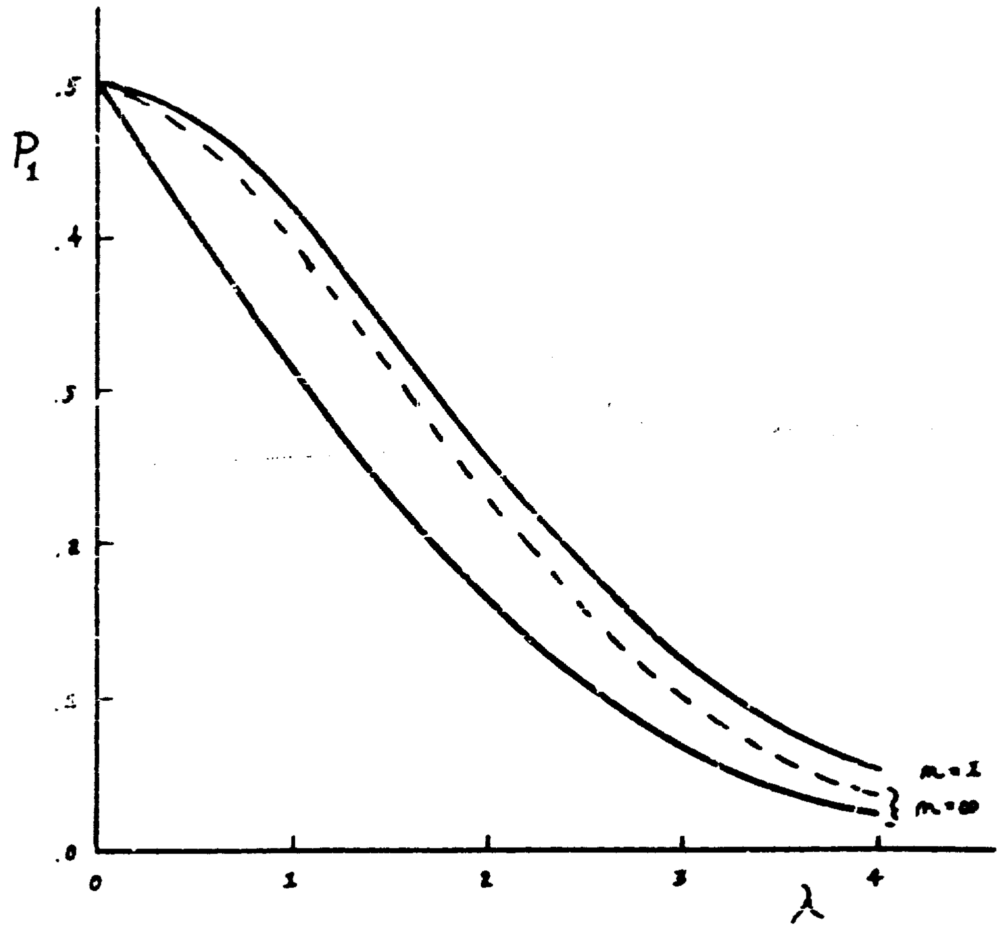

Figure 5

Comparison of the probability of error P_1 as a function of λ, the distance between the means, for the linear discriminant function and the nonparametric discrimnator, distance function = Δ, $k = 1$, for two normal univariate populations n = size of sample from each population. $n = 1$ is identical for both. --- indicates the nonparametric procedure.

$P_1^{(3)}(z) = 6P\{$all Y's and X_1 are nearer to z than X_2 while X_3 is farther from z than $X_2\}$

$+ 18P\{Y_1, Y_2, X_1$ are nearer z than X_2 while X_3 and Y_3 are farther from z than $X_2\}$

$$= 6 \int_0^\infty K_z^3(\delta)\, H_z(\delta)\, [1 - H_z(\delta)]\, dH_z(\delta)$$

$$+ 18 \int_0^\infty K_z^2(\delta)\, H_z(\delta)\, [1 - H_z(\delta)][1 - K_z(\delta)]\, dH_z(\delta).$$

Then, as before,

$$P_1^{(3)} = E\left[P_1^{(3)}(z)\right].$$

As $n \longrightarrow \infty$, $P_1^{(3)}$ may be shown (the argument is similar to the one used when $n \longrightarrow \infty$, $k = 1$) to approach

$$P_1^{(3)} = \int_{-\infty}^{\infty} \frac{[g(z)]^3 + 3[g(z)]^2\, f(z)}{[f(z) + g(z)]^3}\, f(z)\, dz.$$

It is noteworthy that as $n \longrightarrow \infty$, the value of $P_1^{(3)}$ for fixed values of k, however small, are independent of the dimensionality p of the sample space.

From this formula, the middle column of table III was computed. Corresponding results from tables I and II-A are repeated for comparison. As shown in [1], as $n \longrightarrow \infty$ and

$k \to \infty$ (more slowly, however, than n), the linear discriminant function and the nonparametric discriminators have a common limiting behavior, shown in line three of table III. Thus, for $\lambda = 2$, $p = 1$, and large n, the "rule of two out of three" has a 19.2 per cent chance of misclassification as against 15.9 per cent for the optimum. Figure 6 illustrates these results graphically.

Table III

Limiting probabilities of error as $n \to \infty$, for the p-variate normal distribution

λ	k = 1	k = 3	k = ∞
0	.500	.500	.500
1	.398	.368	.309
2	.225	.192	.159
3	.098	.080	.067
4	.034	.027	.023
5	.009	.007	.006

n = size of sample from each population

= distance between the means of the transformed populations

k = odd integer such that Z is assigned to that population from which λ came the majority of the k nearest observations.

Probability of error = P{Z is assigned to G | Z came from F}

= P{Z is assigned to F | Z came from G}.

The probability of error for n large is independent of p.

5. Bivariate normal distribution

For $p = 2$, we have employed methods analogous to those described in section 4, to obtain the probabilities of error for the nonparametric discriminators with $k = 1$; $\lambda = 1, 2, 3$;

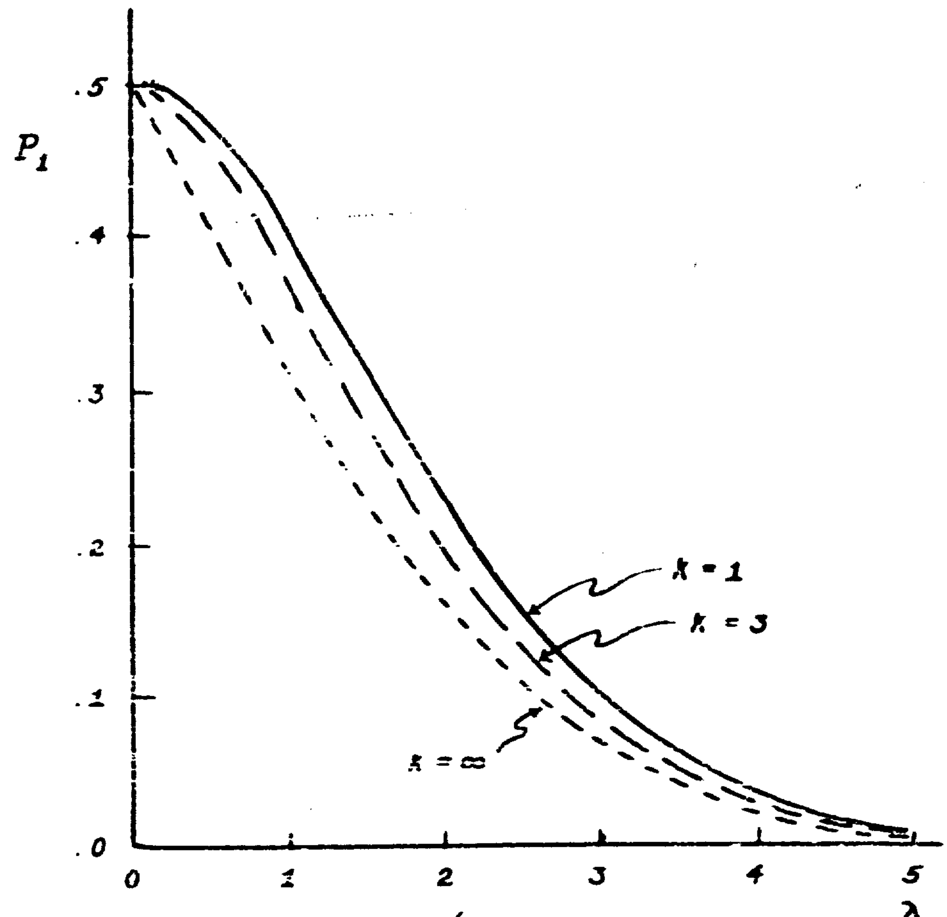

Figure 6

Limiting probabilities of error P_1 as n, the size of sample from each population, $\to\infty$, for two p-variate normal distributions. Distance function = Δ. k = number of nearest individuals on which the nonparametric procedure is based.

and $n = 1, 2, 3, 4, 5, 10, 20, 50, \infty$. The results are summarized in table IV. All finite values of $n > 4$ were obtained by the approximate method discussed in the last section. A comparison of the values obtained by numerical integration with those given by the approximation are shown in table IV-A.

To enable the reader to get a clearer picture of the change in probabilities of misclassification with a change in λ, figure 7 shows the values of table IV plotted against λ.

Unfortunately we do not have available the comparable figures for the linear discriminant function. However, as a measure of the efficiency of the nonparametric discriminators we have included the optimum limiting behavior to which both the nonparametric discriminator and the linear discriminant function tend.

6. $k \geqq 3$ for the univariate and bivariate normal distributions

As k is increased the computations become much more laborious, so much so that the actual numerical integrations were carried out in only a very few instances for the "two out of three rule." The following method may, however, be used to estimate the effect of $k \geqq 3$. Let us consider an alternative discriminator which we shall denote as (r, n', k'). Suppose $k = rk'$ and $n = rn'$. Partition the $2n$ sample values at random into r sets of $2n'$ each and for each set observe the population-or-origin of the majority of the k' observations nearest to Z. Assign Z to that population whose elements are in the majority for a majority of the r sets. It is easy to show that this discriminator will determine the

Table IV

Probabilities of error, nonparametric discriminators, k = 1, bivariate normal distribution

n	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
1	.435	.292	.157
2	.429	.269	.135
3	.423	.259	.125
4	.420	.252	.120
5	.417	.250	.117
10	.411	.240	.109
20	.406	.234	.104
50	.402	.230	.100
∞	.398	.225	.098

Table IV-A

Comparison of the values obtained by numerical integration with those obtained by the approximate method

n	λ	numerical integration	approximation
1	1	.4354	.4370
1	2	.2920	.2951
2	2	.2693	.2721
3	2	.2566	.2612
4	2	.2525	.2548
1	3	.1572	.1560

n = size of sample from each population

λ = distance between the means of the transformed populations

Probability of error = P{Z is assigned to F | Z came from G}

= P{Z is assigned to G | Z came from F}

k = odd integer such that Z is assigned to that population from which came the majority of the k nearest observations. k = 1 is "nearest neighbor rule."

Distance function = $\Delta = \max\{|x_1 - z_1|, |x_2 - z_2|\}$

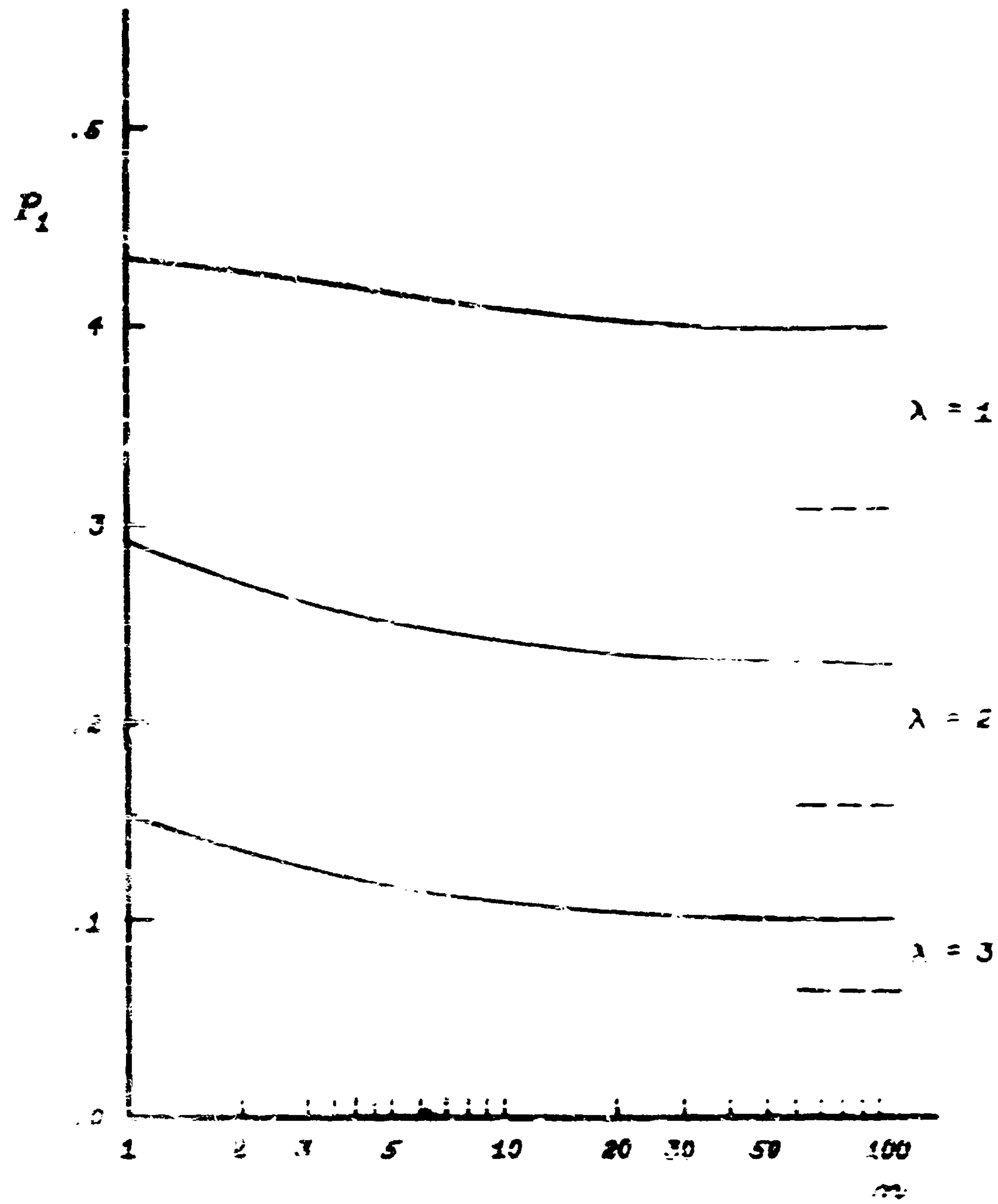

Figure 7

Probability of error P_1 of the nonparametric discriminator with Δ as distance function, for two bivariate normal distributions with distance between means = λ. n = size of sample from each population. k = 1, the rule of nearest neighbor. --- indicates the optimum likelihood ratio procedure.

assignment of Z on the basis of observations less close to Z than would be the case if we employed the ordinary discriminator using the k closest of the entire sample of 2n. Hence it is intuitive that the probabilities of error of (r, n', k') will exceed those of the usual rule (n,k). We do not know a proof of this, however.

The computation of P_1 for (r, n', k') once P_1 has been obtained for (n', k') is relatively easy. For fixed z, the r sets can be regarded as r independent trials each with constant probability $P_1(z)$ for (n', k') of success (success is here defined as the event that Z will be misclassified). The values of $P_1(z)$ for (r, n', k') can then be found from the tables of the binomial distribution [6].

Tables V and VI give the results for the univariate and bivariate normal distributions, respectively. The first line in table VI has the values calculated for the two out of three rule. The second line gives the probability of error when a sample of three observations from each population is considered as a set of three independent trials and the individual Z is assigned to that population in which the majority of the trials placed him. One notices that while the corresponding probabilities in the two lines are extremely close the figures bear out one's intuition mentioned above. The tables have been arranged so that comparison between different uses of the same total number of individuals in the sample will be convenient and an idea of the most effective discriminator (r, n', k') can be obtained. The same results are illustrated graphically in figures 8 and 9.

Table V

Probabilities of error, nonparametric discriminator, univariate normal distribution

n	r	n'	k'	$\lambda = 1$	$\lambda = 2$
3	1	3	1	.405	.231
3	3	1	1	.385	.203
9	9	1	1	.345	.173
10	1	10	1	.399	.223
29	29	1	1	.324	.164
50	1	50	1	.398	.225

n = total size of sample from each population

r = number of sets in the partition of the total sample

n' = size of each of the r sets; n = n'r

k' = 1 = rule of nearest neighbor

λ = distance between the means of the transformed populations

Probability of error = $P\{Z$ is assigned to $G | Z$ came from $F\}$

= $P\{Z$ is assigned to $F | Z$ came from $G\}$

Distance function = Δ .

Table VI

Probabilities of error, nonparametric discriminator, bivariate normal distribution

n	r	n'	k'	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
3	1	3	3	.408	.238	.110
3	3	1	1	.408	.239	.112
15	1	15	1	.408*	.2[illegible]6*	
15	3	5	1	.386	.207	
15	5	3	1	.375	.198	
15	15	1	1	.353*	.188*	
5	5	1	1	.391	.218	.096
12	3	4	1	.389	[illegible]	.090
29	29	1	1	.332	.184	
30	3	10	1	.379	.201	.083
150	3	50	1	.371	.195	.077

n = total size of sample from each population

r = number of sets in the partition of the total sample

n' = size of each of the r sets; n = n'r

k' = 1 = rule of nearest neighbor

k' = 3 = rule of two out of three

λ = distance between the means of the transformed populations

Probability of error = $P\{Z$ is assigned to $G \mid Z$ came from $F\}$

$= P\{Z$ is assigned to $F \mid Z$ came from $G\}$

Distance function = Δ .

*The starred values were read from graphs

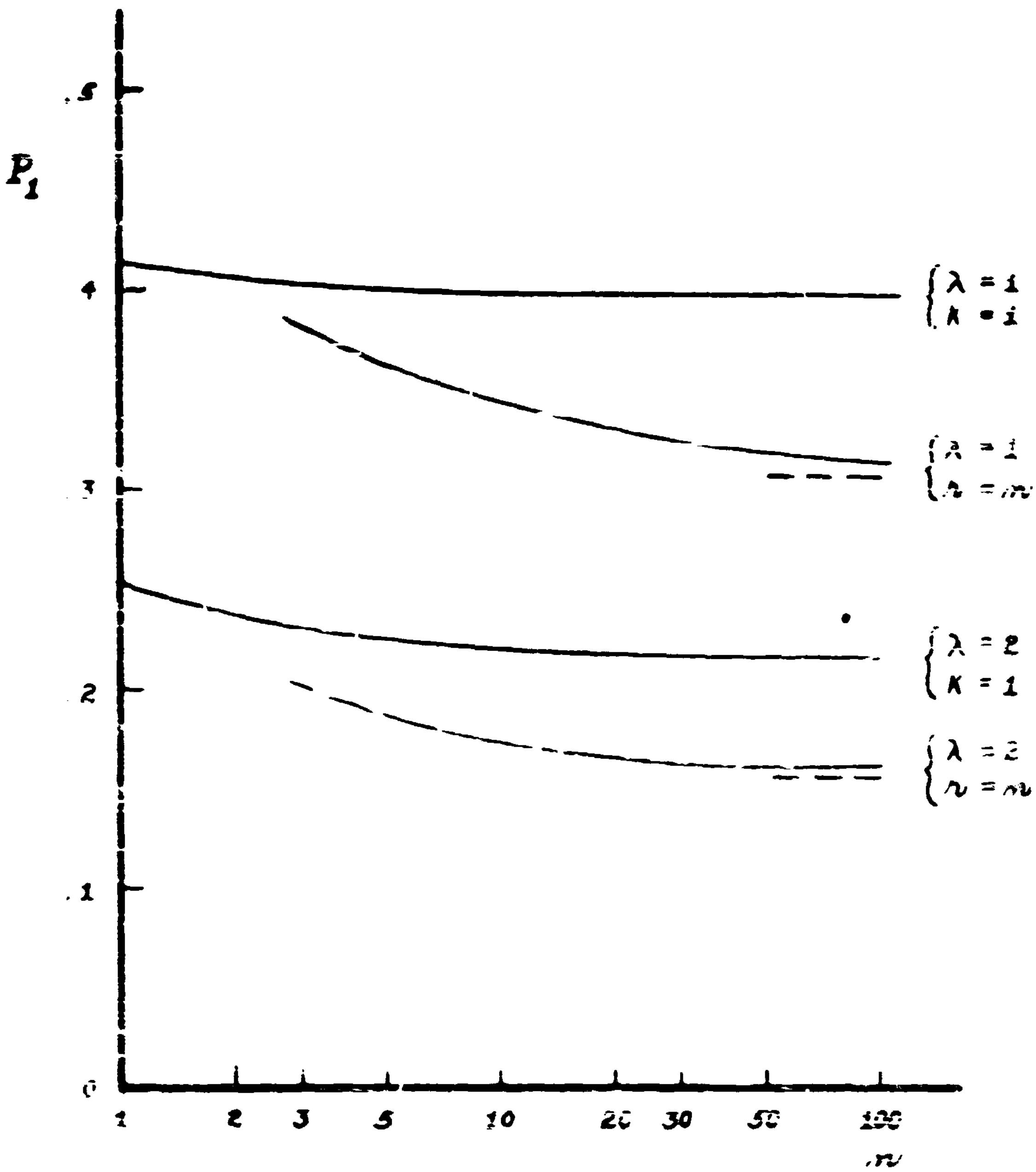

Figure 8

Probability of error P_1 of the nonparametric discriminator, distance function = Δ, for two bivariate normal populations with distance between means = λ. n = size of sample from each population. k = k' = 1 and r = 1 for k = 1; r = n for k' = 1. --- indicates the optimum procedure.

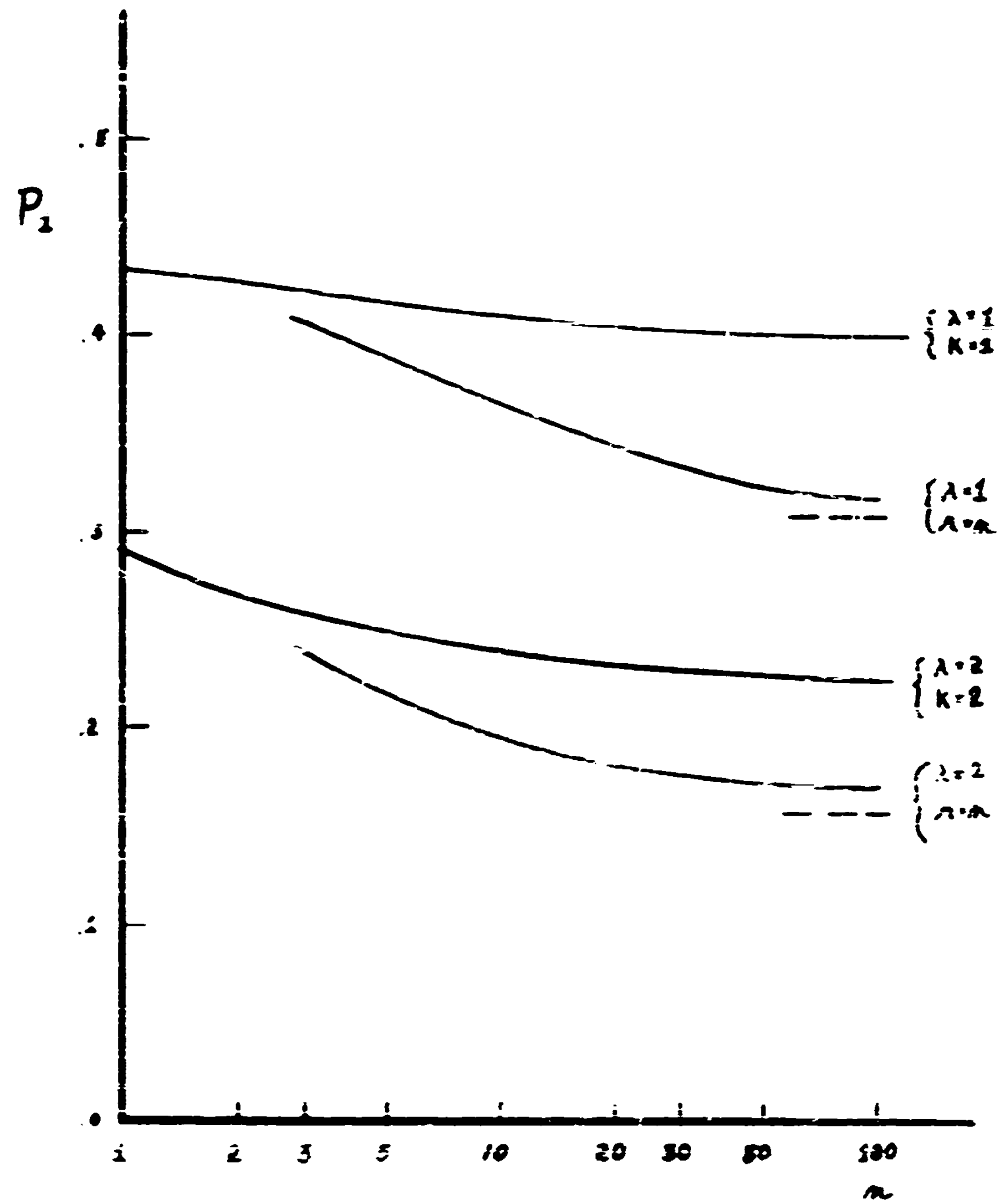

Figure 9

Probability of error P_1 of the nonparametric discriminator, distance function = Δ , for two bivariate normal distributions, $r = 1$ with $k' = 1$ and $r = n$ with $k' = 1$.

7. Alternative distance functions for the normal bivariate distribution

The dependence of P_1 on the distance function was emphasized in section 3. The numerical results which are given in this section are intended to show the magnitude of the effect on P_1 of certain moderate changes in the distance function.

During the computations which are reported in section 5, we noticed that the value of $P_1(0,0)$, the conditional probability of error given that Z is at the origin (the expected position of Z), was remarkably consistent with the value of P_1. Since we felt that it would be more worthwhile to survey a larger area of problems than to concentrate on the complete answer to one, we decided to make use of the fact noted above and to recalculate the values of $P_1(0,0)$ for various distance functions. In table VII, the values of $P_1(0,0)$ and P_1 are given, together with the difference $P_1 - P_1(0,0)$. The fourth column gives and approximation for P_1 obtained by adding a crude correction term to $P_1(0,0)$, namely,

$$\frac{1}{3}\,\frac{.5 - P_1(0,0)}{\lambda/2},$$

.5 being the value of $P_1(\lambda/2, x_2)$. It is our belief that the order of the magnitude of the change in P_1 with the change of distance function will be shown by the effect of the distance function on $P_1(0,0)$.

Table VII

Comparison of the probabilities of error P_1 with the conditional probability of error $P_1(0,0)$ given that Z is at the origin. Nonparametric discriminator, normal p-variate distribution, p = 1, 2, k = 1, 3.

n	P_1	$P_1(0,0)$	D	$\tilde{P}_1$	P_1	$P_1(0,0)$	D	$\tilde{P}_1$
	λ = 1, p = 2, k = 1				λ = 2, p = 2, k = 1			
1	.435	.391	.044	.464	.292	.184	.108	.289
2	.429	.385	.044	.462	.269	.159	.110	.273
3	.423	.383	.040	.461	.259	.149	.110	.266
4	.420	.382	.038	.461	.252	.142	.110	.261
5	.417	.381	.036	.460	.250	.139	.111	.259
10	.411	.379	.032	.460	.240	.130	.110	.253
20	.406	.379	.027	.460	.234	.125	.109	.250
50	.402	.378	.024	.459	.230	.121	.109	.247
∞	.398	.378	.020	.459	.225	.119	.106	.246
	λ = 1, p = 2, k = 3				λ = 2, p = 2, k = 3			
3	.408	.339	.069	.446	.238	.088	.150	.225
	λ = 3, p = 2, k = 1				λ = 1, p = 1, k = 1			
1	.157	.051	.106	.151	.418	.365	.053	.455
2	.135	.027	.108	.132	.408	.365	.043	.455
3	.125	.024	.101	.130	.405	.367	.038	.455
4	.120	.021	.099	.127	.403	.368	.035	.456
5	.117	.019	.098	.126	.401	.369	.032	.456
10	.109	.015	.094	.123	.399	.373	.026	.458
20	.104	.013	.091	.121	.398	.375	.023	.458
50	.100	.012	.088	.120	.398	.376	.021	.459
∞	.098	.011	.087	.120	.398	.378	.020	.459
	λ = 3, p = 2, k = 3							
3	.110	.007	.103	.117				
	λ = 2, p = 1, k = 1				λ = 3, p = 1, k = 1			
1	.253	.145	.108	.263	.124	.033	.090	.137
2	.236	.125	.112	.250	.108	.019	.090	.126
3	.231	.119	.112	.246	.104	.015	.089	.123
4	.228	.116	.111	.244	.101	.013	.088	.121
5	.225	.115	.109	.243	.099	.012	.087	.120
10	.223	.115	.108	.243	.098	.011	.087	.120
20	.224	.116	.108	.244	.098	.011	.087	.120
50	.225	.119	.105	.246	.098	.011	.088	.120
∞	.225	.119	.106	.246	.098	.011	.087	.120

$D = P_1 - P_1(0,0)$

$$\tilde{P}_1 = P_1(0,0) + \frac{2[.5 - P_1(0,0)]}{3}$$

In selecting the alternative distances, we had in mind that, in general, we were dealing with a transformed space and were interested in the effect on the probability of error if Δ itself were used as a distance function in the original space. The distance functions chosen are as follows. The definition of Δ is repeated for the sake of completeness.

$$\text{(i)} \quad \Delta[(x_1, x_2), (z_1, z_2)] = \max \{|x_1 - z_1|, |x_2 - z_2|\}.$$

The locus of points at a given distance from z is a square, centered at z, with sides parallel to the axes.

$$\text{(ii)} \quad \Delta_1[(x_1, x_2), (z_1, z_2)] = \sqrt{(x_1 - z_1)^2 + (x_2 - z_2)^2}.$$

Thus Δ_1 is ordinary Euclidean distance (perhaps a more natural distance function than Δ). The locus of points at a given distance from z is a circle centered at z.

$$\text{(iii)} \quad \Delta_2[(x_1, x_2), (z_1, z_2)] = \max \{|x_1 - z_1|, 3|x_2 - z_2|\}.$$

The locus of points at a given distance from z is a rectangle centered at z whose sides are parallel to the axes and in the ratio of one to three.

$$\text{(iv)} \quad \Delta_3[(x_1, x_2), (z_1, z_2)] = \max \{3|x_1 - z_1|, |x_2 - z_2|\}.$$

The locus of points at a given distance from z is a rectangle centered at z with sides parallel to the axes and in the ratio of three to one.

Distance functions Δ_2 and Δ_3 are the transforms if the original distributions is independent normal bivariate but the variances of the two measurements are unequal.

(v) Distance functions denoted by $\Delta(\rho = a)$. The locus of points is a square centered at z but whose sides are not parallel to the axes. The values of ρ are a = .25, .50, .75. This is the transform of Δ if the original distribution is joint normal bivariate with the two variates having unit variances and covariance = ρ.

The comparison of values of $P_1(0,0)$ for the various distance functions is given in table VIII and in figure 10. It will be seen that for all practical purposes it makes no difference whether Δ or Δ_1 (Euclidean distance) is used. However, the effect of the other distance functions is marked. This bears out the statement made previously that a burden is placed upon the statistician for selecting the appropriate distance function.

8. $p \geqq 2$ for the p-variate normal distribution

This section is an attempt to give an indication of the influence that an increase in p, the number of dimensions of the sample space, will produce on the probability of misclassification. We have again computed only the conditional probability $P_1(0)$ for z at the origin. Two alternative distance functions were used, namely,

Table VIII

Probabilities of error, nonparametric discriminator, normal bivariate distribution, $k = 1$.

		Δ		Δ_1		Δ_2		Δ_3	
λ	n	$P_1(0,0)$	$\tilde{P}_1$	$P_1(0,0)$	$\tilde{P}_1$	$P_1(0,0)$	$\tilde{P}_1$	$P_1(0,0)$	$\tilde{P}_1$
1	1	.391	.435	.389	.463				
2	1	.184	.292	.184	.289	.383	.422	.146	.225
2	2	.300	.269			.300	.367	.128	.211
2	3	.254	.259			.254	.336	.123	.207
2	4	.226	.252			.226	.317	.122	.206
3	1	.051	.157	.053	.152				
		$\Delta = \Delta(\rho = 0)$		$\Delta(\rho = .25)$		$\Delta(\rho = .50)$		$\Delta(\rho = .75)$	
2	1	.184	.292	.179	.286	.165	.277	.137	.258
	2	.159	.269	.152	.268	.129	.253	.084	.222
	3	.149	.259	.140	.260	.112	.241	.062	.208
	4	.142	.252	.133	.255	.102	.235	.049	.199

λ = distance between the means of the transformed populations

n = size of sample from each population

k = 1 = nearest neighbor rule

$P_1(0,0)$ = conditional probability that for z at the origin, Z will be misclassified

P_1 = probability of misclassification

$\tilde{P}_1$ = rough estimate of P_1. The figures to be compared are the $P_1(0,0)$

Distance functions Δ's are as defined in the preceding paragraphs

$$\Delta[(x_1,\cdots,x_p),(z_1,\cdots,z_p)] = \max_{i=1}^{p} |x_i - z_i|$$

and

$$\Delta_1[(x_1,\cdots,x_p),(z_1,\cdots,z_p)] = \sqrt{\sum_{i=1}^{p} (x_i - z_i)^2},$$

and the computations were carried out for $n = 1$, $k = 1$.

The results are shown in table IX and figure 11. As one would expect the results depend rather heavily on the dimen-

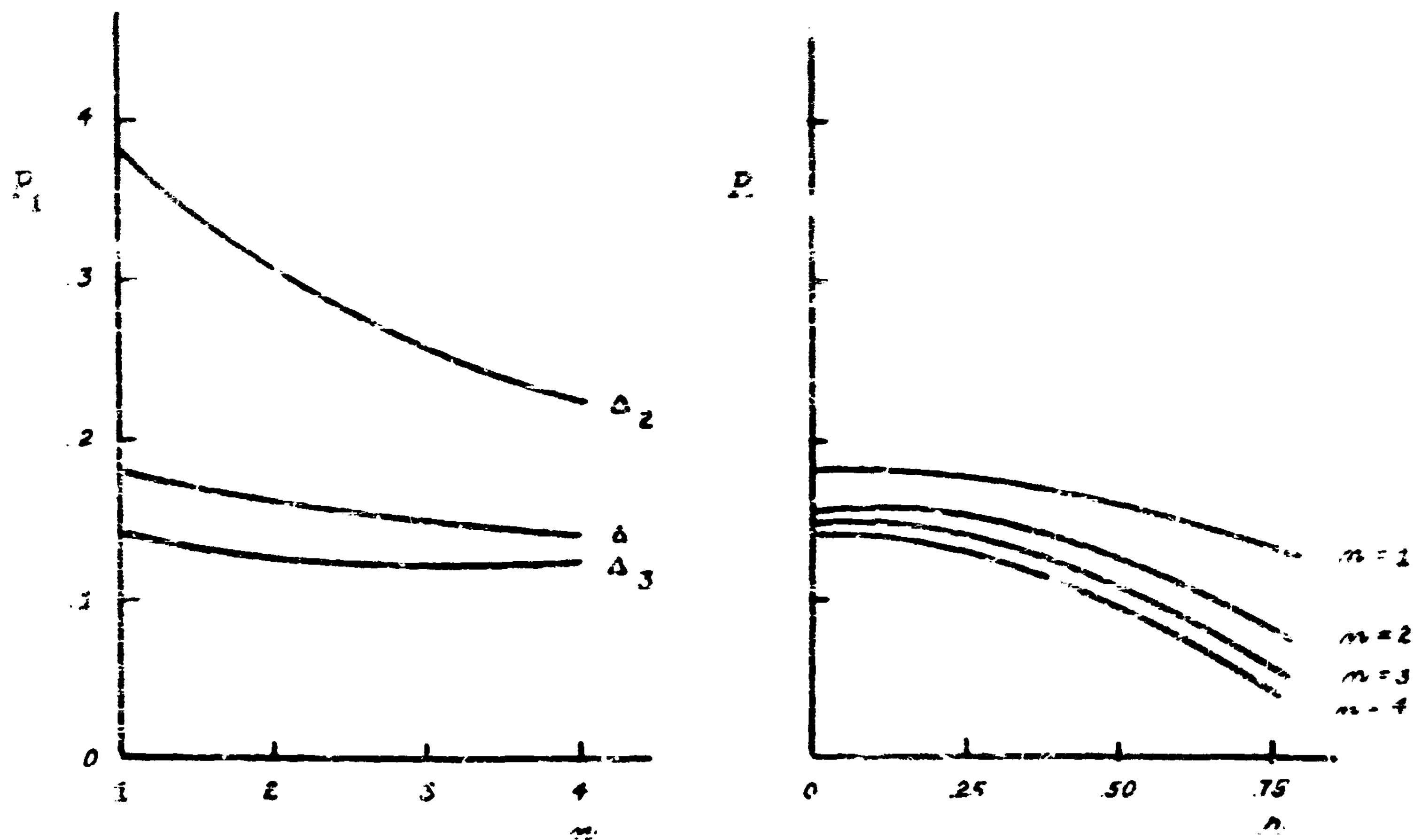

Figure 10

Probability of error P_1 of nonparametric discriminator for two bivariate normal distributions with distance between the means equal to 2 for various distance function. $n = k = 1$.

Table IX

Conditional probabilities $P_1(0,0)$ of error given that Z is at the origin, nonparametric discriminator, normal p-variate distribution, $n = 1$, $k = 1$

	$\lambda = 1$	$\lambda = 2$		$\lambda = 3$
p	Δ_1	Δ	Δ_1	Δ_1
2	.389	.184	.184	.053
4	.414	.228	.230	.082
6	.427	.255	.259	.105
10		.288	.295	

sionality of the space when n is fixed. Admittedly, this is a most cursory glance at the situation for p dimensions. The fact that the figures refer to $n = 1$, $k = 1$, means of course that the figures have no practical value. Nevertheless, we decided to include them since it seemed that the behavior in this simplest case might provide some indication of what might be expected as the dimensionality p is increased.

9. Conclusion

The choice between parametric and nonparametric rules will in any given situation depend upon (i) the strength of the statistician's belief in his parametric model, (ii) the loss he would suffer by using the nonparametric rule if in fact the parametric form is correct and (iii) the loss he would suffer by using the parametric rule if the actual densities depart from the parametric form assumed. In [1], it was ascertained that if the sample size increases and at the same time the number of nearest neighbors on which the non-parametric procedures base its decision is increased but more slowly than the sample size, then in the limit the probabilities of error will be those of the optimum likelihood ratio rule whatever the population densities. However,

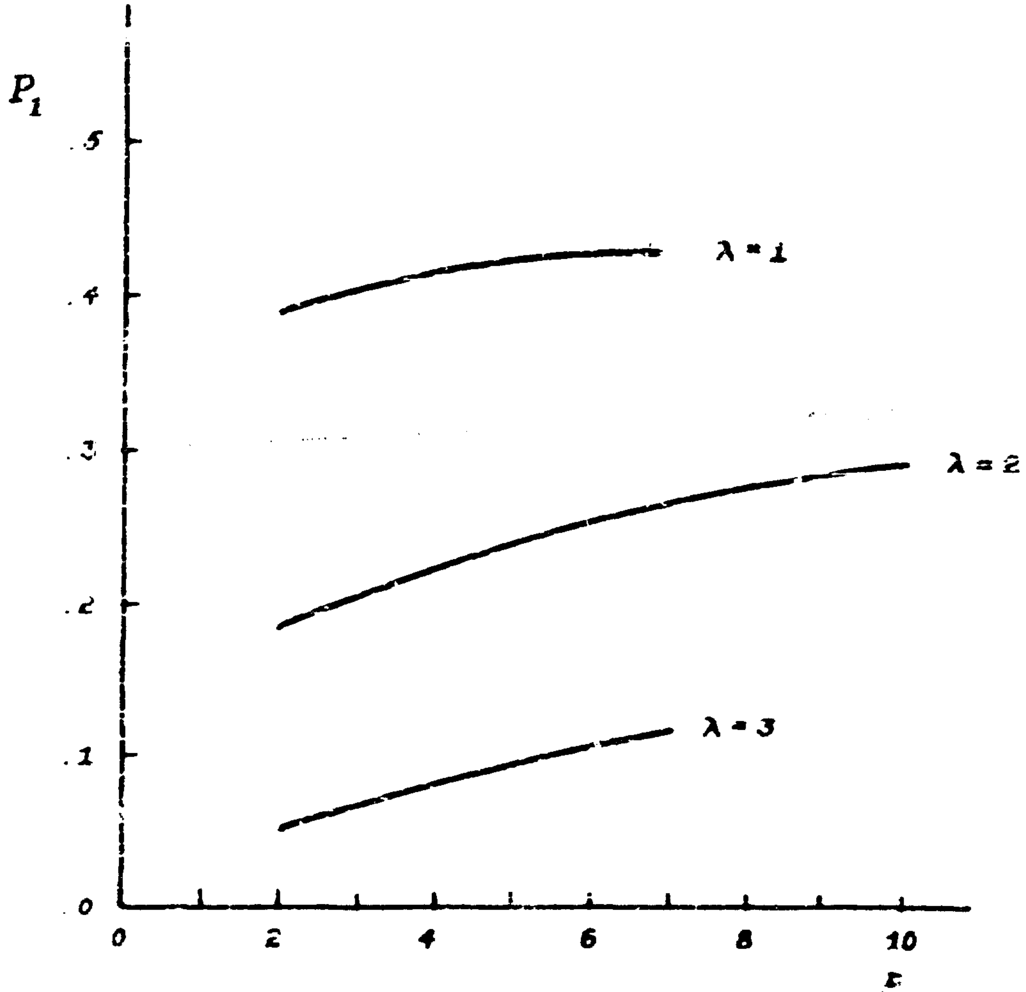

Figure 11

Probability of error P_1 of nonparametric discriminator, distance function = Δ, for two p-variate distributions with distance between the means = λ. n = 1, k = 1.

the matter of greatest practical interest is the performance of the rules when the samples are small.

In this paper, we have been concerned with (ii) for the special case of greatest interest, the linear discriminant function. We succeeded in finding the probabilities of misclassification for some nonparametric procedures. However, the computation of the performance characteristic of the linear discriminant function proved to be too lengthy. It would be of extreme value, especially when one thinks of the wide use to which the linear discriminant function is put if its probabilities of misclassification in representative situations would be tabulated.

In summary, let us indicate the nature of the situations in which a nonparametric discriminator may be preferable to the linear discriminant function, and conversely. If the populations to be discriminated are well known, and have been investigated to establish that the normal distribution gives a good fit and that the variances and correlations do not change much when the means are changed, and if the classification to be made warrants the labor of matrix inversion, then the linear discriminant function should certainly be used. If on the other hand, the populations are either not well known, or are known not to be approximately normal, or to have very different covariance matrices; or if the discrimination is one in which small decreases in probability of error are not worth extensive computations, then the simple nonparametric rule, perhaps with $k \geqq 3$, seems to have the edge.

In conclusion, we would like to express our appreciation to the many people to whom we are indebted for help in the preparation of this paper. Especially we would like to thank Mrs. Jeanne Lovasich and Mrs. Eloise Putz, who computed the tables for us.

REFERENCES

[1] EVELYN FIX and J. L. HODGES, Jr., "Nonparametric discrimination. I. Consistency properties," Technical report prepared under contract between University of California and School of Aviation Medicine, Randolph Field, Texas, 22 pp.

[2] Tables of probability functions. Vol. 2, Federal Works Agency, Mathematical Tables Project, National Bureau of Standards, 1942.

[3] KARL PEARSON, Tables for statisticians and biometricians parts 1 and 2. Cambridge University Press, 1930.

[4] T. A. JEEVES, "Functions of large numbers. Asymptotic expansions," from Probability methods in physics--seminar by Michel Loeve and appendix by participants, pp. 187-201. Mimeographed, University of California, 1948-1949.

[5] O. PERRON, K. Akad. der Wis., Munick, Sitz. Math. Phy. Kl., 1917, pp. 191-219.