Nonparametric classification and nonparametric density function estimation
by Don Owen Loftsgaarden

A thesis submitted to the Graduate Faculty in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY in Mathematics
Montana State University

Abstract:
Two main problems are considered In this thesis. There is a definite need for nonparametric classification procedures, i.e. classification procedures which require no parametric assumptions.

This is the first problem considered here. A nonparametric classification procedure is presented and various properties of this procedure are discussed.

During the course of work on the above mentioned problem, a second problem was encountered which is important in its own right. One approach to the nonparametric discrimination problem depends on having a consistent nonparametric density function estimator available. Consequently, a considerable portion of the work in this thesis was concerned with finding a good nonparametric density function estimator. Such a density estimator is introduced and its asymptotic consistency is shown. In addition, a small empirical study was run on this estimator to obtain some feeling for its behavior for various finite values of n.

NONPARAMETRIC CLASSIFICATION AND NONPARAMETRIC DENSITY FUNCTION ESTIMATION

by

DON OWEN LOFTSGAARDEN

A thesis submitted to the Graduate Faculty in partial fulfillment of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Mathematics

Approved:

Head, Major Department

C. P. Zwennberg
Chairman, Examining Committee

K. Goering
Asst. Dean, Graduate Division

MONTANA STATE COLLEGE
Bozeman, Montana

June, 1965

ACKNOWLEDGMENT

The research in this thesis was supported in part by a National Defense Education Act Fellowship and in part by National Aeronautics and Space Administration Research Grant NsG-562.

The author is particularly indebted to Dr. C.P. Quesenberry, his thesis advisor, for the many hours he spent directing this thesis. Dr. F.S. McFeely made several suggestions on a preliminary draft of this thesis which improved the final copy immensely.

The author is also grateful to his wife for the long hours she spent in typing preliminary drafts of this thesis and to Mrs. Sarah Myrick for her excellent typing of the final copy.

TABLE OF CONTENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Two main problems are considered in this thesis. There is a definite need for nonparametric classification procedures, i.e. classification procedures which require no parametric assumptions. This is the first problem considered here. A nonparametric classification procedure is presented and various properties of this procedure are discussed.

During the course of work on the above mentioned problem, a second problem was encountered which is important in its own right. One approach to the nonparametric discrimination problem depends on having a consistent nonparametric density function estimator available. Consequently, a considerable portion of the work in this thesis was concerned with finding a good nonparametric density function estimator. Such a density estimator is introduced and its asymptotic consistency is shown. In addition, a small empirical study was run on this estimator to obtain some feeling for its behavior for various finite values of n.

CHAPTER I

INTRODUCTION

The classification or discrimination problem arises when an experimenter makes a number of measurements on an individual and wishes to classify the individual into one of several populations on the basis of these measurements. As an example, consider a lot of electronic parts. Each part is to be classified as defective or nondefective, the criterion being length of time to failure. If the length of time to failure is less than some preassigned period of time the part is classified as defective and otherwise as nondefective. Obviously, each of the parts cannot be tested to see whether it is defective or not as the necessary test would be destructive (i.e. the part would be tested until it failed). The discrimination approach to this problem is to make several measurements on a part and then classify the part as defective or nondefective on the basis of these measurements.

Some widely used classification procedures require parametric assumptions. An example of a procedure of this kind is the one based on the classical normal discriminant function, which may be found in Anderson (1958). Many people using this procedure have been concerned with the parametric assumptions that necessarily must be made. For this reason, nonparametric classification procedures not requiring parametric assumptions are desirable. However, very little work has been done in developing such nonparametric classification procedures. In this paper a nonparametric discrimination procedure is presented. Certain optimum properties of this procedure, mainly asymptotic, are shown.

In the process of developing this procedure, the problem of finding a nonparametric estimator for a probability density function arose. This is a problem of considerable importance in its own right. Such an estimator is introduced in this thesis and properties of this estimator are studied. The development of this density estimator is a principal part of this thesis. A nonparametric classification procedure is almost an immediate result of obtaining this density estimator.

Since one of the important properties of this nonparametric density estimator studied here is an asymptotic property, a small empirical study was made on the IBM 1620 computer. The purpose of this study was to see how the density estimator behaved in practical situations for finite sample sizes. The results of this study are included in this thesis.

CHAPTER II

REVIEW OF LITERATURE

2.1 Literature on Nonparametric Classification

The classification problem was introduced in Chapter I. In the example given in Chapter I there are only two populations involved, the defective or the nondefective. Two types of errors can be made in classifying a part. It can be classified as nondefective when it is actually defective or vice versa. The problem is to find discrimination procedures that minimize the probabilities of these errors.

The notation to be used in this simple two population case is now set down. Let F and G be the distribution functions of two random variables, X and Y respectively. F and G are assumed to be absolutely continuous. The corresponding probability density functions are f and g. X and Y are assumed to be p-dimensional. A p-dimensional random variable Z is assumed to have either the F or the G distribution. The problem is to decide on the basis of z, an observed value of Z, which of the two distributions Z has. In other words, the problem is to decide which population z belongs to.

Fix and Hodges (1951) have broken the classification problem into three subproblems according to the assumed knowledge about F and G. These are:

(1) F and G are completely known,

(2) F and G are known except for the values of one or more parameters, i.e. the functional forms of f and g are known except

for parameters,

(3) F and G are completely unknown except possibly for assumptions about existence of densities, etc.

Subproblem (1) is completely solved since the discrimination depends only on $f(z)/g(z)$ by the Neyman-Pearson Fundamental Lemma (cf. Welch (1939)). An appropriate c is chosen and then F or G is chosen as follows:

choose F if $f(z)/g(z) > c$

choose G if $f(z)/g(z) < c$

choose F or G arbitrarily if $f(z)/g(z) = c$.

Fix and Hodges (1951) call this the "likelihood ratio procedure" and use the notation L(c) for it. It is convenient to assume that $P\{f(z) = cg(z)\} = 0$ and this is done hereafter.

For subproblems (2) and (3) samples $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_m$ from F and G, respectively, are assumed given. In subproblem (2) it is further assumed that f and g are known except for the values of a vector parameter θ, where θ belongs to the set Ω. Thus f and g may be written as $f(\theta)$ and $g(\theta)$. The most common approach to subproblem (2) is to use the samples of x's and y's to obtain an estimate $\hat{\theta}$ for θ. Using $\hat{\theta}$ the values $f(\hat{\theta})$ and $g(\hat{\theta})$ are obtained and then one proceeds as if f and g are known as in case (1). An alternative method is to use the likelihood ratio criterion to set up a discrimination procedure.

Examples of both of these methods are given in Anderson (1958), pp. 137-142, for the case of underlying normal assumptions. The first method leads to the procedure based upon the classical discriminant function. These approaches seem reasonable if the underlying assumptions are valid. In the discriminant function case, for example, if there is much departure from the normality assumptions very little is known about the validity of the results.

From the above discussion the need for ways of solving subproblem (3) is seen. This is known as the nonparametric discrimination problem. Here there is a minimum of underlying assumptions. Fix and Hodges (1951) and (1952) were the first to work on subproblem (3). Their work is discussed in some detail here because it is pertinent to the succeeding work. Stoller (1954), Johns (1960) and Cooper (1963) have also considered this problem and brief mention of their work is also made.

No nonparametric classification procedure exists that is better than the optimum procedure available if the density functions f and g are assumed completely known (cf. Fix and Hodges (1951)). The L(c) procedure can therefore be used as a sort of limiting procedure to try to approach with a nonparametric procedure. Intuitively, it would seem that a good nonparametric procedure should have a limiting form which is in some sense consistent with L(c). Fix and Hodges (1951) have made these consistency notions precise. They formulate definitions in terms of sequences of decision functions.

Consider a finite decision space with elements $d_1, \ldots, d_s$. For two populations there are only two decisions, d_1 and d_2. Let $\{D'_n\}$ and $\{D''_n\}$ be two sequences of decision functions. For example, D'_n is a function which for each n chooses one of the decisions $d_1, \ldots, d_s$ (or in the setting of this thesis, D'_n chooses the population into which z is classified). These decision functions will usually depend on n observed values of some random variable. Two ways in which these two sequences can be thought of as being consistent with each other are now considered. First, for each n there is a probability that D'_n will make the decision d_1, that D'_n will make the decision d_2, etc. The same thing is true for D''_n. If the probabilities that D'_n and D''_n make the decision d_i are nearly the same for all i as n increases, then in this sense the sequences are consistent with each other. Secondly, it might happen that there is a high probability that D'_n and D''_n will make the same decision as n increases. These ideas motivate the following definitions of Fix and Hodges (1951).

<u>Definition 1</u> We shall say that the sequences $\{D'_n\}$ and $\{D''_n\}$ are consistent in the sense of performance characteristics if, whatever be the true distributions, and whatever be $\varepsilon > 0$, there exists a number N such that whenever $m > N$ and $n > N$,

$$|P\{D'_n = d_i\} - P\{D''_m = d_i\}| < \varepsilon$$

for every decision d_i.

<u>Definition 2</u> We shall say that the sequences $\{D_n'\}$ and $\{D_n''\}$ are consistent in the sense of decision functions if, whatever be the true distributions, and whatever be $\varepsilon > 0$, there exists a number N such that whenever $m > N$ and $n > N$,

$$P\{D_n' = D_m''\} > 1 - \varepsilon.$$

Consistency in the second sense implies consistency in the first sense, but not vice versa (cf. Fix and Hodges (1951)). It is generally convenient to use definition 2.

Earlier discussion has shown that in nonparametric discrimination it would be desirable to have a way of comparing a sequence of decision functions and the L(c) procedure. The following definition is a specialization of definition 2 which does this.

<u>Definition 3</u> A sequence $\{D_{n,m}\}$ of discrimination procedures, based on Z and on samples $x_1, x_2, \ldots, x_n$ from F and $y_1, y_2, \ldots, y_m$ from G, is said to be consistent with L(c) if, whatever be the distributions F and G, regardless of whether Z is distributed according to F or according to G, and whatever be $\varepsilon > 0$, we can assure

$$P\{D_{n,m} \text{ and } L(c) \text{ yield the same classification of } Z\} > 1 - \varepsilon$$

provided only that m and n are sufficiently large.

$D_{n,m}$ has the double subscript as each D is a function of $x_1, x_2, \ldots,$ x_n, $y_1, y_2, \ldots, y_m$ and Z. Both m and n get large in the sequence $\{D_{n,m}\}$.

Theorem 1 of Fix and Hodges (1951) concerns consistency in subproblem (2). It shows the relationship between consistency in estimation and the consistency of sequences of decision functions as defined above. The following additional notation is needed in order to state this theorem. Let P_1 denote probabilities computed assuming that Z is distributed according to F and P_2 probabilities computed assuming that Z is distributed according to G. Let $\mathcal{F} = \{f_\theta | \theta \epsilon \Omega\}$ and $\mathcal{G} = \{g_\theta | \theta \epsilon \Omega\}$ be classes of density functions. It is assumed that some notion of convergence is defined in Ω and θ may be a vector. Suppose that for each n and m there is an estimate $\hat{\theta}_{n,m}$ for θ which is a function of $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_m$.

<u>Theorem 2.1</u> If,

(a) the estimates $\{\hat{\theta}_{n,m}\}$ are consistent;

(b) for every θ, $f_\theta(z)$ and $g_\theta(z)$ are continuous functions of θ for every z except perhaps $z \epsilon Z_\theta$ where $P_i(Z_\theta) = 0$, $i = 1,2$, then the sequence of discrimination procedures $\{\hat{L}_{n,m}(c)\}$ obtained by applying the likelihood ratio principle with critical value $c > 0$ to $f_{\hat{\theta}_{n,m}}(z)$ and $g_{\hat{\theta}_{n,m}}(z)$ is consistent with L(c).

Fix and Hodges (1951) prove this theorem and the proof is similar to the proof of Theorem 2.2 given below.

With this background, the way is clear for discussion of subproblem (3). It should be recalled that the assumptions being made for this case are only that F and G are absolutely continuous. Consideration of the L(c) procedure and Theorem 2.1 suggests an approach to subproblem (3). In the L(c) procedure, once z is given, the only information needed to make the discrimination is f(z) and g(z). In Theorem 2.1, f(z) and g(z) are not directly available so they are estimated. This is done by first estimating θ and then using this estimate of θ to estimate f(z) and g(z). Likewise, in case (3), f(z) and g(z) are not directly available, but these densities are not characterized by a parameter θ. Therefore, it seems logical to try to estimate f(z) and g(z) themselves. Estimating the densities f and g at z is a problem of considerable merit in its own right. Consequently, a later section will be devoted to the literature of this subject. It will include work done by Fix and Hodges on this problem while they were investigating the nonparametric classification problem.

Suppose estimates $\hat{f}$ and $\hat{g}$ for f and g are available. If the L(c) procedure is applied using these estimates for f and g the resulting procedure is designated by $L^*(c,\hat{f},\hat{g})$. Theorem 2.2 below, due to Fix and Hodges (1951), states what type of estimates $\hat{f}$ and $\hat{g}$ should be in order for the procedures $L^*(c,\hat{f},\hat{g})$ to be consistent with L(c). This theorem is basic to the work of this paper. For this reason and since the proof does not explicitly appear in Fix and Hodges (1951) it is included here.

<u>Theorem 2.2</u> If $\hat{f}_{n,m}(z)$ and $\hat{g}_{n,m}(z)$ are consistent estimates for f(z) and g(z) for all z except possibly $z \varepsilon Z_{f,g}$ where $P_i(Z_{f,g}) = 0$, $i = 1,2$,

then $L^*_{n,m}(c,\hat{f},\hat{g})$ is consistent with $L(c)$.

Proof:

The $L(c)$ procedure depends on $f(z) - cg(z)$. If $f(z) - cg(z) > 0$ choose F and if $f(z) - cg(z) < 0$ choose G. The event $f(z) - cg(z) = 0$ is assumed to have probability zero under either distribution. Likewise, the $L^*_{n,m}(c,\hat{f},\hat{g})$ procedure will depend on the variable $\hat{f}_{n,m}(z) - c\hat{g}_{n,m}(z)$. To show that $L^*_{n,m}(c,\hat{f},\hat{g})$ is consistent with $L(c)$ it will be sufficient to show two things. The first is that $\hat{f}_{n,m}(z) - c\hat{g}_{n,m}(z) \xrightarrow{P} f(z) - cg(z)$ and the second is that $|f(z) - cg(z)|$ can be bounded away from zero with sufficiently high probability. The first condition assures that $f(z) - cg(z)$ and $\hat{f}_{n,m}(z) - c\hat{g}_{n,m}(z)$ are arbitrarily close to each other with high probability. The second condition assures that they are both positive or both negative with high probability. The two together show that both procedures will make the same decision with high probability.

The above conditions are shown here in reverse order. Fix $c > 0$ and $\varepsilon > 0$. Consider the random variable $|f(z) - cg(z)|$. For $i = 1,2$, $P_i\{|f(z) - cg(z)| \leq a\}$ is the distribution function of this random variable and is continuous on the right. By assumption $P_i\{|f(z) - cg(z)| = 0\} = 0$. Therefore, there exists a $\delta > 0$ such that

$$P_i\{|f(z) - cg(z)| < \delta\} < \varepsilon/2 \qquad i = 1,2.$$

This is condition two above.

By assumption $\hat{f}_{n,m}(z) \xrightarrow{P} f(z)$ and $c\hat{g}_{n,m}(z) \xrightarrow{P} cg(z)$. Therefore,

$\hat{f}_{n,m}(z) - c\hat{g}_{n,m}(z) \xrightarrow{P} f(z) - cg(z)$. That is, there exists an N and M such that if $n > N$ and $m > M$

$$P\{|(\hat{f}_{n,m}(z) - c\hat{g}_{n,m}(z)) - (f(z) - cg(z))| > \delta/2\} < \varepsilon/2.$$

This is condition one above.

Combining these two yields

$$P\{L^{*}_{n,m}(c,\hat{f},\hat{g}) \text{ and } L(c) \text{ yield the same classification of } Z\} > 1 - \varepsilon$$

if $n > N$ and $m > M$.

This ends the discussion of the work of Fix and Hodges for the moment. Further discussion of their work will be contained in the density estimation section of this chapter.

Stoller (1954) gives a distribution-free procedure for the uninvariate ($p=1$) two-population case and an estimate of the probability of correct classification. His approach to the problem is based on a paper by Hoel and Peterson (1949). He restricts the form of his procedure which makes it difficult to generalize it to higher dimensions than $p=1$. The procedure is to choose a point λ. If $z \leq \lambda$ classify z as from F and if $z > \lambda$ classify z as from G. Stoller shows that (1) the estimate of the probability of correct classification is a consistent estimate of the optimum probability of classification and (2) the probability of correct classification induced by this procedure converges in probability to the optimum probability of correct classification.

Johns (1960) has also considered the nonparametric discrimination problem. He uses a slightly more general decision theory setting than

Fix and Hodges do. In their setting the loss depends only on the probability of misclassification. This is a particular case of Johns' work which has a loss structure that takes into account the relative severity of misclassifications.

Cooper (1963) demonstrates that the multivariate classification procedure optimal for the multivariate normal distribution, the discriminant function, is optimum for broader classes of distributions. These classes are the multivariate extensions of the Pearson Type II and III distributions. Thus, this paper is not directly concerned with nonparametric classification.

2.2 Literature on Nonparametric Density Estimation

As indicated in the previous section, the problem of estimating density functions may arise in connection with the nonparametric classification problem. Some other authors who have worked on this problem were motivated by the problem of estimating the hazard, or conditional rate of failure, function $f(x)/\{1 - F(x)\}$. Here it is assumed that $F(x)$ is an absolutely continuous distribution function. The first part of this section concerns work done by Fix and Hodges (1951) and (1952). They applied this work to nonparametric discrimination. Consequently, this part of this section will consider both density estimation and the application to the discrimination problem. It may be thought of as a continuation of the previous section.

As before, a sample $x_1, x_2, \ldots, x_n$ is to be used to obtain an estimate of $f(x)$ at z. The idea is to take a small neighborhood of z for each n. If N is the number of x's that are in this neighborhood then N/n is an estimate of the probability in the neighborhood, and N/n divided by the measure of this neighborhood should provide an estimate of f(z). Lemma 2.3 below due to Fix and Hodges (1951) makes these ideas explicit.

Let μ donate Lebesgue measure in p-dimensional space and let $|X - Y|$ denote the Euclidean distance between X and Y in this same space.

Lemma 2.3 If f(x) is continuous at x = z, and if $\{\Delta_n\}$ is a sequence of sets such that

$$\lim_{n\to\infty} \sup_{d\in\Delta_n} |z - d| = 0$$

and $\lim_{n\to\infty} n\mu(\Delta_n) = \infty$, and if N is the number of $x_1, x_2, \ldots, x_n$ which lie in Δ_n, then $\{N/n\mu(\Delta_n)\}$ is a consistent estimate for f(z).

There are two main ideas involved in the proof and they are roughly stated here. First $P(\Delta_n) \approx f(z)\mu(\Delta_n)$ and secondly $P(\Delta_n) \approx N/n$. One key point to notice in the above lemma is that the measure of the sets Δ_n approaches zero but not as fast as $n\to\infty$, i.e. $\mu(\Delta_n)\to 0$ but $n\mu(\Delta_n)\to\infty$. This type of assumption will be very important in the construction of further density estimators which are consistent.

The nonparametric classification procedures based on these estimates are designated by $L^*(c, N/n\mu(\Delta_n), M/m\mu(\Lambda_m))$. By Theorem 2.2 these procedures are consistent with L(c). These procedures have one drawback. How should the regions Δ_n for f and the corresponding regions Λ_m for g be chosen? Choosing these regions too large or too small can have serious consequences on the density estimators. Consequently, Fix and Hodges have given an alternative approach to this density estimation problem in order to avoid some of these problems. This approach was designed specifically for the nonparametric discrimination problem. It is aimed at estimating two densities at a point z simultaneously.

Preliminary to discussing the alternative approach it is shown that the nonparametric classification problem can be reduced from p-dimensions to one. Let $\rho(x,z)$ be a metric. Suppose that $\rho(x,z)$ and $\rho(y,z)$ are random variables possessing densities, say $f_z(x)$ and $g_z(y)$, which are continuous and not both zero at zero. Setting $\rho(x,y)$ equal to p-dimensional Euclidean distance will work. The two samples $x_1,x_2,\ldots,x_n$ and $y_1,y_2,\ldots,y_m$ are replaced by $\rho(x_1,z)$, $\rho(x_2,z),\ldots,\rho(x_n,z)$ and $\rho(y_1,z)$, $\rho(y_2,z),\ldots,\rho(y_m,z)$ respectively. The problem now is to decide whether $f_z(0)$ or $cg_z(0)$ is larger. Hence, it may be assumed without loss of generality that f and g are densities of non-negative univariate random variables and that $z = 0$. This is done for the remainder of this chapter.

The main idea of Theorem 2.4 due to Fix and Hodges (1951) is to take an integer k for each n and m and take the k closest points, either x's or y's, to z. This then is the region Δ_n and the region Λ_m. Questions which arise are how should the k's behave and since $N+M = k$ what effect does this have on the estimates? Theorem 2.4 answers these questions.

<u>Theorem 2.4</u> Let X and Y be non-negative. Let f and g be positive and continuous at 0. Let $k(n,m)$ be a positive-integer-valued function such that $k(n,m) \to \infty$, $(1/n)k(n,m) \to 0$ and $(1/m)k(n,m) \to 0$, as n and m approach ∞. (These limits are restricted so that m/n is bounded away from 0 and ∞). Define

$U = k^{th}$ smallest value of combined samples of X's and Y's

N = Number of X's $\leq U$

M = Number of Y's $\leq U$.

Then N/nU is a consistent estimate for $f(0)$ and M/mU is a consistent estimate for $g(0)$.

The $L^*(c, N/nU, M/mU)$ procedure is then to

choose F if $N/n > cM/m$

and choose G if $N/n < cM/m$.

This procedure is consistent with $L(c)$.

The above procedure is seen to have optimum properties as n and m get large. One question which arises immediately is how do these procedures behave if n and m are small?

The procedure based upon the linear discriminant function is a reasonable procedure if (1) F and G are p-variate normal and (2) F and G have the same covariance matrix. This procedure involves much computation whereas the procedures proposed above are easy to apply. This brings up a second question. How much discriminating power is lost if the non-parametric procedure above is applied when assumptions (1) and (2) are valid? Fix and Hodges (1952) attempt to answer this question and the one in the preceding paragraph simultaneously. They compute the two types of error for p=1,2 and various values of n (mostly small) and compare these errors for the two procedures. The nonparametric procedure seems to compare very favorably.

This ends the review of the work of Fix and Hodges (1951) and (1952). A discussion of several papers concerned solely with density estimation follows.

Parzen (1962) has constructed a family of estimators for a density function f(x). His estimators are based on a sample of n independent and identically distributed random variables $x_1, x_2, \ldots, x_n$ with the same continuity assumptions as stated previously. His estimators are consistent in quadratic mean, (this type of consistency implies ordinary consistency). They are also asymptotically normal. A review of his work follows.

Let $F_n(x)$ be the sample distribution function.

$$
\begin{aligned}
F_n(x) &= 0 && \text{for } x < x_{(1)} \\
&= r/n && \text{for } x_{(r)} \leq x < x_{(r+1)} \qquad r=1,\ldots,n-1 \\
&= 1 && \text{for } x \geq x_{(n)}
\end{aligned}
\tag{2.1}
$$

where the lower subscript in parentheses indicates an order statistic. The random variable $F_n(x)$ is a binomial random variable with $EF_n(x) = F(x)$ and var $\{F_n(x)\} = \{(1/n)F(x)\}\{1-F(x)\}$. One possible estimate for f(z) which occurs is

$$\hat{f}_n(z) = \{F_n(z+h) - F_n(z-h)\}/2h. \tag{2.2}$$

Here h is a positive number which must get small as $n \to \infty$. How fast should h get small? How should h be chosen? These are question which must be answered in studying $\hat{f}_n(z)$.

These questions may appear new but they were encountered earlier. In fact, the estimator (2.2) is of the type proposed by Fix and Hodges in Lemma 2.3 for the univariate case. Once h is chosen, for a particular n, this determines an interval or region Δ_n about z of diameter 2h. $F_n(z+h) - F_n(z-h)$ is the number of x's which fall in Δ_n, say N, divided by the total number of x's, n. Thus, $\hat{f}_n(z)$ is equal to N/2hn which in turn is equal to $N/n\mu(\Delta_n)$. This is the Fix and Hodges estimator as claimed. The problem of choosing h is thus seen to be analogous to the earlier problem of choosing Δ_n.

Rewriting $\hat{f}_n(z)$ in an alternative form will suggest a whole class of estimators based on the empirical distribution function $F_n(x)$. It turns out that to study the estimate (2.2) and to try to answer the above questions concerning h it is just as easy to consider this whole class of estimators. Let

$$(2.3) \qquad \begin{aligned} K(y) &= 1/2 \qquad |y| \leq 1 \\ &= 0 \qquad |y| > 1. \end{aligned}$$

Then

$$\hat{f}_n(z) = \int_{-\infty}^{\infty} (1/h)K(\{z-y\}/h)dF_n(y) = (1/nh)\sum_{j=1}^{n} K(\{z-x_j\}/h).$$

Varying $K(y)$ will lead to other estimators.

A few of the results obtained by Parzen are now stated. This is not an exhaustive listing of his results, but a summary of results relevant to the work of this paper.

$K(y)$ will be restricted to be a weighing function, i.e. $K(y)$ will be an even function satisfying the following conditions:

$$(2.4)\qquad \begin{aligned} &\text{(i)} \quad \sup_{-\infty<y<\infty} |K(y)| < \infty \\ &\text{(ii)} \quad \int_{-\infty}^{\infty} |K(y)|\,dy < \infty \\ &\text{(iii)} \quad \lim_{y\to\infty}. |yK(y)| = 0 \\ &\text{(iv)} \quad \int_{-\infty}^{\infty} K(y)\,dy = 1. \end{aligned}$$

The first three of these conditions are necessary for applying a theorem in Bochner (1955) which is a key theorem needed in the proof of some of Parzen's results.

Since h depends on n let $h = h(n)$. If

$$(2.5)\qquad \lim_{n\to\infty}. h(n) = 0,$$

then (2.2) is asymptotically unbiased (i.e. $\lim_{n\to\infty}. E\,\hat{f}_n(z) = f(z)$). If in addition to (2.5) the h's satisfy

$$(2.6)\qquad \lim_{n\to\infty}. nh(n) = \infty,$$

then $\hat{f}_n(z)$ is consistent in quadratic mean (i.e. $\lim_{n \to \infty} E|\hat{f}_n(z) - f(z)|^2 = 0$).

We note that h(n) approaching zero is equivalent to the interval about z, Δ_n, getting small. In effect, (2.6) says that Δ_n must not get small as fast as $n \to \infty$. This same idea was noted previously in comments immediately following Lemma 2.3.

Here, then, is a class of estimators all of which are consistent. These are the type of estimators needed in Theorem 2.2. However, certain problems in choosing the h's still remain, as conditions (2.5) and (2.6) allow for wide variation. Parzen (1962) has obtained a theoretical optimum value of h which is of the form $cn^{-\delta}$ where c is a constant and $0 < \delta < 1$. Its practical value in applications is somewhat limited however, since knowledge of the unknown density is necessary in computing the optimum h. This problem of choosing h's seems to be basic to the density estimation problem. Intuitively, it seems that perhaps the choice of h should depend on the sample of n, $x_1, x_2, \ldots, x_n$, available. This makes a fundamental change in the estimators as h would then be a random variable rather than a constant. A close look at the alternative density estimators proposed by Fix and Hodges in Theorem 2.4 shows that this is exactly what has been done there. This will also be the case for a density estimator proposed in the work of this paper.

Rosenblatt (1956) was the original contributor to the nonparametric density estimation problem. He considered the estimator (2.2) in some detail. He also indicated how a class of estimators could be generated using different weighing functions and discussed them briefly. Parzen's paper (1962) is a continuation and generalization of Rosenblatt's work. One interesting fact which Rosenblatt proves is that all nonparametric density function estimators are biased.

Cacoullos (1964) is concerned with estimating a multivariate density. His paper is a generalization of Parzen's (1962) work with almost all results being multivariate analogies of Parzen's results.

Manija (1961) gives the two-dimensional generalization of Rosenblatt's paper (1956). In view of Cacoullos' paper (1964), mentioned above, this paper need not be considered here.

Some work on estimating density functions has been done using a somewhat different approach than those already discussed. In particular, Whittle (1958) and Watson and Leadbetter (1963) have used results from spectral analysis in constructing density estimators. Spectral analysis is concerned with analyzing a stationary time series. It turns out that it is much easier to work with the spectral density function, which is the Fourier transform of the autocovariance function of the time series, than to work with the time series directly. A problem which immediately arises is the estimation of this spectral density function. Several papers have been written on this subject, most of them prior to the papers on the density estimation problem. This estimation problem turns out to have several similarities with the one being discussed here.

Whittle (1958) follows an approach that he used in an earlier paper on spectral density estimation. Watson and Leadbetter (1963) use the techniques of Parzen from an earlier paper on spectral density estimation. Whittle (1958) considers estimators of the form

$$(2.7) \qquad \hat{f}(x) = (1/n) \sum_{j=1}^{n} w_x(x_j),$$

where $w_x(y)$ is chosen to minimize " expected mean square error ". A key assumption in his work is " that the curve f(x) being estimated is one of a whole population of curves, and that the population correlation coefficient of f(x) and f(x+ε) tends to unity as ε tends to zero ".

Watson and Leadbetter (1963) consider estimators of the form

$$(2.8) \qquad \hat{f}_n(x) = (1/n) \sum_{i=1}^{n} \delta_n(x-x_i),$$

with δ_n assumed square integrable. They use various criteria from the Parzen paper on spectral density estimation including minimizing the " mean integrated square error ". Two broad classes are defined in terms of the behavior of the characteristic function of the distribution. It is shown that the class of estimators proposed by Parzen (1962) falls into one of these categories. Some consistency criteria are defined, again following Parzen, and the estimates (2.8) are discussed in terms of these definitions.

Discussion of these last two papers in greater detail here does not seem warranted as they are not closely related to the work of this paper. The papers of Fix and Hodges (1951) and Parzen (1962) are more closely related. In addition, the discussion of these latter two, in some detail, serves as good background material for the density estimation problem, as it will be treated here. It should be noted that only papers dealing with nonparametric density estimation have been reviewed. Nothing is said about parametric density estimation.

CHAPTER III

THEORY OF COVERAGES

The theory of coverages plays a key role in the development of the density estimator of this thesis. The results needed will be developed in this chapter. The notation set down in Chapter III will continue to be used throughout the later chapters.

Let $x_1, x_2, \ldots x_n$ be a sample from a population with an absolutely continuous distribution function $F(x)$. Let the ordered values be $x_{(1)}$, $x_{(2)}, \ldots, x_{(n)}$, i.e. $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$. The intervals $(-\infty, x_{(1)}]$, $(x_{(1)}, x_{(2)}], \ldots, (x_{(n)}, +\infty)$ are called sample blocks and designated by $B_1^{(1)}, \ldots, B_1^{(n+1)}$. The coverages $c_1, \ldots, c_{n+1}$ of these blocks are defined as $F(x_{(1)})$, $F(x_{(2)}) - F(x_{(1)}), \ldots, 1 - F(x_{(n)})$, respectively. Since $\sum_{i=1}^{n+1} c_i = 1$ only the first n coverages are usually considered. The subscript on the B's indicates that these blocks are for one-dimensional variables. Note: $P(B_1^{(i)}) = c_i$ and the c_i are random variables. The following theorem on coverages is easily proved using the theory of the dirichlet distribution and the theory of order statistics (see Wilks (1962)).

<u>Theorem 3.1</u> The sum of any k of the coverages $c_1, c_2, \ldots, c_{n+1}$ has the beta distribution with parameters k and n-k+1.

The generalization of this theorem in terms of multi-dimensional coverages is used in some of the proofs of this paper. Multi-dimensional coverages are now introduced and the generalization of Theorem 3.1 is given.

Let $x_1, x_2, \ldots, x_n$ be a sample of n from a p-dimensional distribution. $x_i = (x_{1i}, x_{2i}, \ldots, x_{pi})$ for $i = 1,2,\ldots,n$. An ordering function φ is introduced where $w = \varphi(x)$ is a univariate random variable which has a continuous distribution function $H(w)$. Consider the new random variables $w_1, w_2, \ldots, w_n$ where $w_i = \varphi(x_i)$ for $i = 1,2,\ldots,n$. Order these w_i obtaining $w_{(1)}, w_{(2)}, \ldots, w_{(n)}$. The coverages are now defined as $c_1 = H(w_{(1)})$, $c_2 = H(w_{(2)}) - H(w_{(1)}), \ldots, c_n = H(w_{(n)}) - H(w_{(n-1)})$. They correspond to the p-dimensional sample blocks $B_p^{(1)}, B_p^{(2)}, \ldots, B_p^{(n+1)}$. The blocks are the disjoint parts that the p-dimensional space is divided into by the ordering curves $\varphi(x) = w_{(i)}$ for $i = 1,2,\ldots,n$. As before, $c_i = P(B_p^{(i)})$, $i = 1,2,\ldots,n$.

<u>Theorem 3.2</u> The sum of any k of the p-dimensional coverages $c_1, c_2, \ldots, c_{n+1}$ has the beta distribution with parameters k and n-k+1.

Further discussion and proofs may be found in Wilks (1962), Wilks (1941), Wald (1943) and Tukey (1947).

An example is now given. It illustrates how this theorem will be used in this paper. Let $p = 2$, and consider a sample of n two-dimensional random variables $x_1, \ldots, x_n$. Let z be a point in two-dimensional Euclidean space. Define $\varphi(x)$ as the two-dimensional Euclidean distance between x and z, i.e. $w = \varphi(x) = |x-z|$. Using φ we obtain a new set of ordered

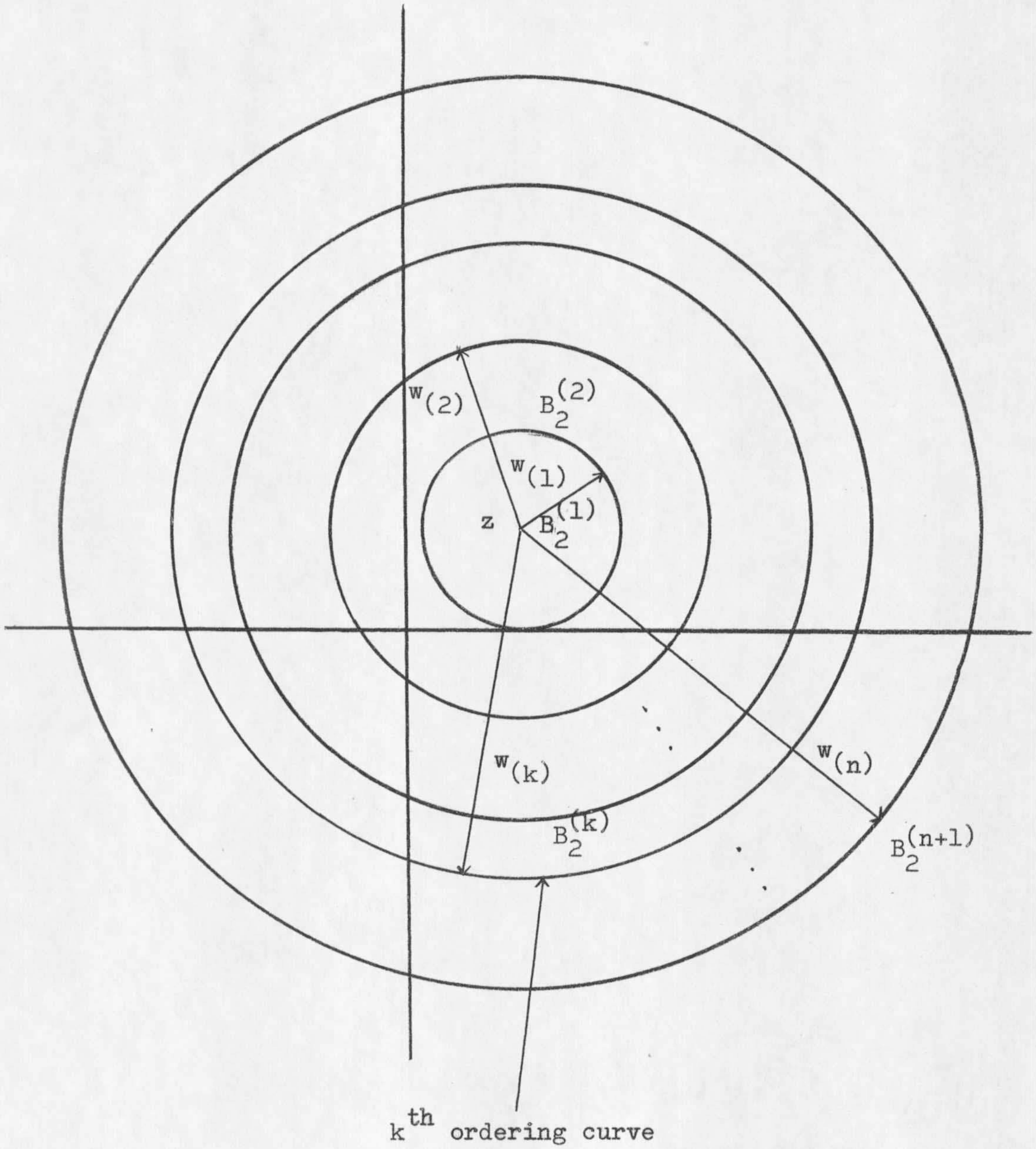

FIGURE 1

SAMPLE BLOCKS AND COVERAGES FOR p = 2

variables $w_{(1)},\ldots,w_{(n)}$. These induce the sample blocks $B_2^{(1)},\ldots,B_2^{(n+1)}$. The distance from z to the x_i closest to it is $w_{(1)}$. Therefore, $B_2^{(1)}$ consists of those points inside a circle about z of radius $w_{(1)}$. This circle is the ordering curve $\varphi(x) = w_{(1)}$. The distance from z to the x_i that is 2[nd] closest to z is $w_{(2)}$. Therefore, $B_2^{(2)}$ consists of those points inside a circle about z of radius $w_{(2)}$ but which are not in $B_2^{(1)}$. $B_2^{(k)}$ consists of those points which are inside a circle of radius $w_{(k)}$ about z but which are not in $B_2^{(1)}, B_2^{(2)},\ldots, B_2^{(k-1)}$ for $k = 1,2,\ldots,n$. $B_2^{(n+1)}$ consists of those points outside the circle of radius $w_{(n)}$ about z. See Figure 1. As previously, c_i is equal to $P(B_2^{(i)})$, $i = 1,2,\ldots,n$. The sum of k blocks $B_2^{(1)} + \ldots + B_2^{(k)}$ consists of those points inside a circle of radius $w_{(k)}$ about z. The sum of corresponding coverages is $c_1 + \ldots + c_k$. By Theorem 3.2 this sum of coverages has the beta distribution with parameters k and n-k+1. Analogous results are available for arbitrary dimension p except that the circles are replaced by p-dimensional hyperspheres.

The following results concerning the beta distribution should be recalled as they are used in the sequel. If m has the beta distribution with parameters k and n-k+1 then

$$\begin{aligned} E(m) &= k/(n+1) \\ \text{var}\ (m) &= k(n-k+1)/(n+1)^2(n+2). \end{aligned} \tag{3.1}$$

CHAPTER IV

A NONPARAMETRIC DENSITY FUNCTION ESTIMATOR

An estimator for a multivariate density function is now proposed. In one sense, the estimator proposed in Theorem 2.4 by Fix and Hodges may be considered a special case of this estimator. The two estimators will be compared. The motivation and development of this estimator is by different methods than those used by Fix and Hodges. The theory of coverages contained in Chapter III is used in the proof of the consistency of this estimator.

Let $x_1, \ldots, x_n$ be a sample of n p-dimensional observations on a random variable $X = (X_1, X_2, \ldots, X_p)$. An observation x_i is $x_i = (x_{1i}, \ldots, x_{pi})$. Assume X has an absolutely continuous distribution function, $F(x) = F(x_1, x_2, \ldots, x_p)$, with corresponding density function $f(x) = f(x_1, x_2, \ldots, x_p)$, which necessarily exists almost everywhere.

$$f(x_1, \ldots, x_p) = \frac{\partial^p F(x_1, \ldots, x_p)}{\partial x_1 \partial x_2 \ldots \partial x_p} . \tag{4.1}$$

An estimate is desired for the density f at a point of continuity, $z = (z_1, z_2, \ldots, z_p)$, where f is also positive.

4.1 Preliminary Work and Notation

Let h_i, $i=1,2,\ldots,p$ be positive constants.

$$\begin{aligned} &P\{z_1-h_1 \leq X_1 < z_1+h_1, \ldots, z_p-h_p \leq X_p < z_p+h_p\} \\ &= F(z_1+h_1, \ldots, z_p+h_p) - F(z_1+h_1, \ldots, z_{p-1}+h_{p-1}, z_p-h_p) - \ldots \end{aligned} \tag{4.2}$$

$$-F(z_1-h_1, z_2+h_2, \ldots, z_p+h_p) + \ldots + (-1)^p F(z_1-h_1, \ldots, z_p-h_p)$$

$$\equiv \Delta_p F(z_1, \ldots, z_p).$$

In particular, for the case p=2

$$\Delta_2 F(z_1, z_2) = F(z_1+h_1, z_2+h_2) - F(z_1+h_1, z_2-h_2) - F(z_1-h_1, z_2+h_2) + F(z_1-h_1, z_2-h_2).$$

Now by definition and the original assumptions

$$f(z_1, \ldots, z_p) = \lim_{\substack{h_i \to 0 \\ i=1,\ldots,p}} \frac{\Delta_p F(z_1, \ldots, z_p)}{2^p h_1 h_2 \ldots h_p} \tag{4.3}$$

$$= \lim_{\substack{h_i \to 0 \\ i=1,\ldots,p}} \frac{P\{z_1-h_1 \leq X_1 < z_1+h_1, \ldots, z_p-h_p \leq X_p < z_p+h_p\}}{2^p h_1 h_2 \ldots h_p}$$

It is convenient to write (4.3) in an alternative form. For this the following notation is necessary. Let $d(X,Z)$ represent the p-dimensional Euclidean distance function $|X-Z|$. A p-dimensional hypersphere about z of radius r will be designated by $S_{r,z}$, i.e. $S_{r,z} = \{x \mid d(x,z) \leq r\}$. The volume or measure of the hypersphere $S_{r,z}$ will be called $A_{r,z}$. $A_{r,z}$ is equal to $2r^p\pi^{p/2}/p\Gamma(p/2)$.

Briefly

(4.4) $\quad d(X,Z) = |X-Z|$

(4.5) $\quad S_{r,z} = \{x \mid d(x,z) \leq r\}$

(4.6) $\quad A_{r,z} = \text{Measure of } S_{r,z} = 2r^p\pi^{p/2}/p\Gamma(p/2).$

Both $S_{r,z}$ and $A_{r,z}$ appear later with identifying sub-subscripts.

Using this notation and noting that $A_{r,z} \longrightarrow 0$ if and only if $r \longrightarrow 0$, $f(z_1,z_2,\ldots,z_p)$ may be written

$$(4.7) \qquad f(z_1,\ldots,z_p) = \lim_{r\to 0} P\{X \varepsilon S_{r,z}\}/A_{r,z},$$

i.e. there exists an R such that if $r < R$ then

$$(4.8) \qquad |P\{X\varepsilon S_{r,z}\}/A_{r,z} - f(z_1,\ldots,z_p)| < \varepsilon,$$

for arbitrary $\varepsilon > 0$.

4.2 A Nonparametric Density Function Estimator

This section will include a general discussion which will serve as the motivation for writing down a density estimator $\hat{f}_n(z)$. In the following section the consistency of this estimator will be shown.

According to (4.8), $P\{X\varepsilon S_{r,z}\}/A_{r,z}$ can be made as near $f(z_1,\ldots,z_p)$ as one chooses by letting r approach zero. $P\{X\varepsilon S_{r,z}\}$ is unknown since it depends on the density f which is being estimated. Therefore, if a good estimate of $P\{X\varepsilon S_{r,z}\}$ can be found, it can be substituted in the expression $P\{X\varepsilon S_{r,z}\}/A_{r,z}$ and this should be a good estimate of the density f at z. This is the approach which will be used here.

In the example given in the chapter on coverages it was seen that the sum of k blocks $B_p^{(1)} + B_p^{(2)} + \ldots + B_p^{(k)}$ was a p-dimensional hypersphere $S_{r_k,z}$ where $r_k = w_{(k)}$, and $w_{(k)}$ is as defined in the example of Chapter III. If the hypersphere $S_{r,z}$ in the preceding paragraph is replaced by

$S_{r_k,z}$, a sum of blocks, then $P\{X \varepsilon S_{r,z}\}$ is the sum of the corresponding coverages. Theorem 3.2 on coverages gives the distribution of this sum of coverages. If this sum of coverages converges in probability to its expectation, then approximating it by this expectation seems reasonable. This is basically the idea used here to obtain a density estimator.

More specifically, let $\{k(n)\}$ be a sequence of integers (this can be generalized to more general $k(n)$ with minor difficulty) such that

$$(4.9) \qquad \begin{aligned} &\lim_{n\to\infty} k(n) = \infty \\ &\lim_{n\to\infty} k(n)/n = 0. \end{aligned}$$

From now on, $k(n)$ is set equal to k unless something to the contrary is stated, i.e. $k(n) = k$.

Let the ordering function be $\varphi(x) = w = |x-z|$. The ordered variables are $w_{(1)}, w_{(2)},\ldots,w_{(n)}$, corresponding to the set of p-dimensional x's, $x_1,\ldots,x_n$. According to the example of Chapter III, the ordering curves $\varphi(x) = w_{(i)}$, $i=1,\ldots,n$, are surfaces of p-dimensional hyperspheres of radius $w_{(i)}$, $i=1,\ldots,n$, about z. The corresponding blocks are $B_p^{(1)},\ldots, B_p^{(n+1)}$ and the coverages are $c_1,c_2,\ldots,c_n$. Consider the sum of k of these blocks, $B_p^{(1)} + \ldots + B_p^{(k)}$. This sum is a p-dimensional hypersphere $S_{r_k,z}$ where $r_k = w_{(k)}$. Let U_k equal the corresponding sum of coverages $c_1 + \ldots + c_k$.

$$U_k = c_1 + \ldots + c_k = P\{X \varepsilon S_{r_k, z}\}. \tag{4.10}$$

By Theorem 3.2, U_k has the beta distribution with parameters k and n-k+1. Therefore, using (3.1) it is easily shown that U_k minus $(k-1)/n$ converges in probability to zero, i.e. $U_k - (k-1)/n \xrightarrow{P} 0$. Thus, $(k-1)/n$ is an approximation for $U_k = P\{X \varepsilon S_{r_k, z}\}$. This leads to the proposed estimator

$$\begin{aligned} \hat{f}_n(z) &= \{(k-1)/n\}\{1/A_{r_k, z}\} \\ &= \{(k-1)/n\}\{p\Gamma(p/2)/2r_k^p\pi^{p/2}\}. \end{aligned} \tag{4.11}$$

The reason for using $(k-1)/n$ rather than $k/(n+1)$ is discussed in Chapter VI. Theorem 4.1, below, supplies the details of the preceding discussion.

4.3 Consistency of the Nonparametric Density Estimator $\hat{f}_n(z)$

<u>Lemma 4.1</u> If $c_n = a_n/b_n \xrightarrow{P} c \neq 0$ and $a_n \xrightarrow{P} 1$, then $b_n \xrightarrow{P} 1/c$.

Proof:

$a_n \xrightarrow{P} 1$ and $c_n \xrightarrow{P} c \neq 0$. Therefore, $a_n/c_n = b_n \xrightarrow{P} 1/c$.

This elementary lemma is used in a key step of the proof of Theorem 4.1 below; thus it is convenient to have it explicitly set down here.

<u>Theorem 4.1</u> The density estimator $\hat{f}_n(z)$ as given in (4.11) is consistent.

Proof:

The first step in this proof is to show that $f(z_1, \ldots, z_p)$ can be approximated by $P\{X \varepsilon S_{r_k, z}\}/A_{r_k, z}$. This is done by showing that

$P\{X \varepsilon S_{r_k,z}\}/A_{r_k,z} \xrightarrow{P} f(z_1,\ldots,z_p).$

$P\{X \varepsilon S_{r_k,z}\} = U_k$ has the beta distribution with parameters k and n-k+1. This can be used to show that $U_k \xrightarrow{P} 0$. By (3.1)

$$(4.12) \qquad \begin{aligned} EU_k &= k/(n+1) \\ \text{var } \{U_k\} &= k(n-k+1)/(n+1)^2(n+2). \end{aligned}$$

Using the Tchebysheff inequality and for arbitrary $\varepsilon > 0$

$$P\{|U_k - k/(n+1)| \geq \varepsilon\} \leq \text{var } (U_k)/\varepsilon^2$$

$$= k(n-k+1)/(n+1)^2(n+2)\varepsilon^2.$$

Using the conditions (4.9), the right hand side is seen to converge to zero. Thus, for large n, the right hand side can be made arbitrarily small. That is, $U_k - k/(n+1) \xrightarrow{P} 0$. Using (4.9) again gives $k/(n+1) \longrightarrow 0$. Combining these two results gives

$$(4.13) \qquad U_k = P\{X \varepsilon S_{r_k,z}\} \xrightarrow{P} 0.$$

However, this can happen only if the measure of $S_{r_k,z}$, viz. $A_{r_k,z}$, converges in probability to zero, by the continuity assumptions. This, in turn, can occur if and only if $r_k \xrightarrow{P} 0$.

Let R be as defined in (4.8). Since $r_k \xrightarrow{P} 0$, there exists an N such that if n > N, and for arbitrary $\eta > 0$

$$(4.14) \qquad P\{r_k < R\} > 1 - \eta.$$

Using (4.8) and (4.14) the following statement can be made. If $n > N$

$$P\{|P\{X\varepsilon S_{r_k,z}\}/A_{r_k,z} - f(z_1,\ldots,z_p)| < \varepsilon\} > 1 - \eta,$$

i.e.

$$P\{X\varepsilon S_{r_k,z}\}/A_{r_k,z} \xrightarrow{P} f(z_1,\ldots,z_p). \tag{4.15}$$

This concludes the first part of the proof.

The concluding portion of the proof goes as follows. By (4.15) $U_k/A_{r_k,z} \xrightarrow{P} f(z_1,\ldots,z_p)$ or, rewriting this,

$$\{n/(k-1)\}U_k/\{n/(k-1)\}A_{r_k,z} \xrightarrow{P} f\,(z_1,\ldots,z_p). \tag{4.16}$$

If it can be shown that the numerator of (4.16), viz. $\{n/(k-1)\}U_k$, converges in probability to 1, then it will follow that the denominator, viz. $\{n/(k-1)\}A_{r_k,z}$, will converge in probability to $1/f(z_1,\ldots,z_p)$. This is so because of Lemma 4.1, when the following associations are made: $a_n = \{n/(k-1)\}U_k, b_n = \{n/(k-1)\}A_{r_k,z}$, $c_n = a_n/b_n$ and $c = f(z_1,\ldots,z_p)$. But the above statement is equivalent to

$$\{(k-1)/n\}\{1/A_{r_k,z}\} \xrightarrow{P} f(z_1,\ldots,z_p). \tag{4.17}$$

This is the desired conclusion of the theorem. It remains to show that $\{n/(k-1)\}U_k \xrightarrow{P} 1$.

Consider $\{n/(k-1)\}U_k$.

$$(4.18)\quad \begin{aligned} E\left[\{n/(k-1)\}U_k\right] &= \{n/(k-1)\}\{k/(n+1)\} \\ \operatorname{var}\left[\{n/(k-1)\}U_k\right] &= \{n^2/(k-1)^2\}\{k(n-k+1)/(n+1)^2(n+2)\}. \end{aligned}$$

The variance in (4.18) approaches zero by (4.9). Using the Tchebysheff inequality and for arbitrary $\varepsilon > 0$

$$(4.19)\quad \begin{aligned} P\left[\,|\{n/(k-1)\}U_k - \{n/(k-1)\}\{k/(n+1)\}| \geq \varepsilon\right] &\leq \operatorname{var}\left[\{n/(k-1)\}U_k\right]/\,\varepsilon^2 \\ &= \{n^2/(k-1)^2\}\{k(n-k+1)/(n+1)^2(n+2)\}\{1/\varepsilon^2\}. \end{aligned}$$

Thus, $\{n/(k-1)\}U_k - \{nk/(k-1)(n+1)\} \xrightarrow{P} 0$. Also, $nk/(k-1)(n+1) \xrightarrow{P} 1$. Combining these two results gives

$$(4.20)\qquad \{n/(k-1)\}U_k \xrightarrow{P} 1.$$

Thus, (4.17) follows from the argument above and the theorem is proved. Written out in detail, (4.17) looks as follows:

$$(4.21)\qquad \{(k-1)/n\}\{p\Gamma(p/2)/2r_k^p\pi^{p/2}\} \xrightarrow{P} f(z_1,\ldots,z_p).$$

4.4 How Some Basic Problems Were Solved for $\hat{f}_n(z)$

Earlier, the problem of choosing neighborhoods (or equivalently h's) was mentioned. In particular, it was mentioned in the discussion following Lemma 2.3 and in the paragraph following (2.5) and (2.6) in the review of Parzen's (1962) paper. There, it was pointed out that in order to have a consistent density estimator it was necessary for the h's to get small as n increases but not as fast as n increases. These conditions are made explicit by (2.5) and (2.6) in the review of

Parzen's (1962) paper. It seems necessary to point out that these requirements are satisfied in the estimator of Theorem 4.1 and why they are required for the consistency of this estimator.

The conditions which are equivalent to (2.5) and (2.6) are (4.9). The first condition of (4.9) requires that the sequence of constants get large as $n \longrightarrow \infty$ and the second requires that this sequence be of lower order of infinity than n. These conditions are used in the discussion immediately following (4.13). There it is seen that U_k approaching zero implies the same thing for A_{r_k} which, in turn, implies the same thing for r_k. Thus, r_k can be thought of as h(n) and if U_k does not get small as fast as $n \longrightarrow \infty$ then neither does r_k. By (4.12) U_k is of order $k/(n+1)$ and hence converges to zero, but not as fast as $n \longrightarrow \infty$, since k also approaches ∞. Thus, r_k also gets small, but not as fast as $n \longrightarrow \infty$.

So far, it has been pointed out that the conditions (4.9) are equivalent in some sense to (2.5) and (2.6). It remains to show that they were necessary conditions in the proof of the theorem. This is done simply by pointing out where they were used. The second of the conditions (4.9) was used immediately preceding (4.13) in order to arrive at (4.13). The first of the conditions (4.9) was used in the application of the Tchebysheff inequality immediately following (4.18). It has the effect of making the variance in (4.18) go to zero as $n \longrightarrow \infty$.

It may appear that the estimator $\hat{f}_n(z)$ could have been simplified by fixing k at some constant value for all n. However, this is now ruled out by the above discussion since $\hat{f}_n(z)$ would then not be a consistent estimator.

4.5 The Univariate Case and a Comparison of It with Theorem 2.4

In this section, the discussion will be restricted to the case where X is a univariate random variable with $X \geq 0$, $f(0) > 0$ and f is continuous on $[0,\infty)$. These are the conditions used in Theorem 2.4 and are interesting in view of the discussion in the second paragraph preceding Theorem 2.4. An estimate of f(0) is desired. The specialization of (4.11) to this case is given. The estimators of Theorem 2.4 are discussed and briefly compared to the specialization of (4.11) given here.

By definition, f(0) is

$$(4.22) \qquad f(0) = \lim_{h \to 0} (F(h) - F(0))/h = \lim_{h \to 0} F(h)/h.$$

This statement is equivalent to (4.7) in the general case. In this case, h is the length of the interval $[0,h)$. There is not a central interval about the point zero as $X \geq 0$. The general density estimator is based on a central interval about 0 of length 2h. Thus, some slight modifications must be made in the general density estimator to handle this case. The reason for these modifications being that if f is defined as 0 for $X < 0$, then $X = 0$ is a point of discontinuity of f. However, f is assumed continuous on the interval $[0,\infty)$. The development is sketched here, the verification being almost identical to the proof of Theorem 4.1.

The observations $x_{(1)},\ldots,x_{(n)}$ are already ordered as to distance from $z = 0$ since $X \geq 0$ and X is univariate. Therefore, the ordering function $d(x,z)$ is not necessary here. Consider the blocks $B_1^{(1)} = [0,x_{(1)}]$, $B_1^{(2)} = (x_{(1)},x_{(2)}],\ldots, B_1^{(n+1)} = (x_{(n)},\infty)$. Corresponding to these blocks are coverages $c_1 = F(x_{(1)})$, $c_2 = F(x_{(2)}) - F(x_{(1)}),\ldots,$ $c_{n+1} = 1 - F(x_{(n)})$. Now

$$\begin{aligned} U_k &= c_1 + \ldots + c_k \\ &= F(x_{(k)}) \end{aligned}$$

and
$$B_1^{(1)} + \ldots + B_1^{(k)} = [0,x_{(k)}].$$

The measure of this last sum of blocks is $x_{(k)}$. U_k has the same distribution as it had in previous work. Therefore, the procedure now is the same as that in developing $\hat{f}_n(z)$ in the general case. This leads to

$$\hat{f}_n(0) = \{(k-1)/n\}\{1/x_{(k)}\}. \tag{4.23}$$

In Theorem 2.4, Fix and Hodges have two nonnegative univariate random variables X and Y, respectively. Assuming $f(0) > 0$ and $g(0) > 0$ the problem considered there is estimating $f(0)$ and $g(0)$ simultaneously from samples $x_1,\ldots,x_n$ from f and $y_1,\ldots,y_m$ from g. Their procedure is to choose a sequence of integers $k(n,m)$, such that $k(n,m) \longrightarrow \infty$, $(1/n)k(n,m) \longrightarrow 0$ and $(1/m)k(n,m) \longrightarrow 0$ as m and $n \longrightarrow \infty$, and to define

$U = k^{\text{th}}$ smallest value of combined samples of X's and Y's

N = Number of X's $\leq U$

M = Number of Y's $\leq U$.

Note: $M + N = k(n,m) = k$.

Then

$$\hat{f}(0) = N/nU \quad \text{and} \quad \hat{g}(0) = M/mU. \tag{4.24}$$

No use of the theory of coverages was made in this development.

There is some similarity between (4.23) and (4.24). They both make use of the observations in determining h. In the first case, h is $x_{(k)}$ and in the second case, it is U. This is in contrast to work such as that in Parzen (1962) and Watson and Leadbetter (1963) where h is a constant that depends on n but not on $x_1,\ldots,x_n$. The estimators (4.24) are available only for the specialized case considered in this section. This is very restrictive as far as density estimation in cases other than the classification problem is concerned. On the other hand, the estimator (4.11) is available for any multivariate density satisfying the continuity assumptions previously set down.

4.6 $E\hat{f}_n(z)$ and var $\hat{f}_n(z)$ for the Uniform Distribution in One Case

Let X be a random variable uniformly distributed on the interval $[0,1]$. Then

$$f(x) = 1 \qquad x \in [0,1] \tag{4.25}$$
$$= 0 \qquad \text{otherwise.}$$

Let $z = 0$. Consider the estimator of $f(0)$, $\hat{f}_n(0)$. This is a particular case of (4.23), i.e.

$$\hat{f}_n(0) = \{(k-1)/n\}\{1/x_{(k)}\}.$$

In this section, $E\hat{f}_n(0)$ and var $\hat{f}_n(0)$ are examined and the behavior of $\hat{f}_n(0)$ is studied for this special case.

The distribution function for x is

$$F(x) = 0 \qquad x < 0$$
$$= x \qquad 0 \le x \le 1 \tag{4.26}$$
$$= 1 \qquad x > 1.$$

Using (4.26), the distribution of the k^{th} order statistic can be written down. See Sarhan and Greenberg (1962) or use the results of Chapter III.

$$\Psi(x_{(k)})dx_{(k)} = \frac{n!}{(k-1)!\,(n-k)!}(x_{(k)})^{k-1}(1-x_{(k)})^{n-k}dx_{(k)}. \tag{4.27}$$

Therefore,

$$E\left[1/x_{(k)}\right] = \int_0^1 \frac{n!}{(k-1)!\,(n-k)!}(x_{(k)})^{k-2}(1-x_{(k)})^{n-k}dx_{(k)}$$

$$= \frac{\Gamma(k-1)\,\Gamma(n+1-k)}{\Gamma(n)} \cdot \frac{n!}{(k-1)!\,(n-k)!} \tag{4.28}$$

$$= n/(k-1).$$

Likewise,

$$E\left[1/x_{(k)}^2\right] = \int_0^1 \frac{n!}{(k-1)!\,(n-k)!}(x_{(k)})^{k-3}(1-x_{(k)})^{n-k}dx_{(k)}$$

$$= [n(n-1)]/[(k-1)(k-2)]. \tag{4.29}$$

Using (4.28) and (4.29)

$$\begin{aligned} E\hat{f}_n(0) &= \{(k-1)/n\}E\{1/x_{(k)}\} \\ &= 1 \\ &= f(0), \end{aligned} \tag{4.30}$$

$$\begin{aligned} E\hat{f}_n^2(0) &= \{(k-1)^2/n^2\}E\{1/x_{(k)}^2\} \\ &= \{(k-1)/n\}\{(n-1)/(k-2)\} \end{aligned} \tag{4.31}$$

and

$$\begin{aligned} \operatorname{var}\hat{f}_n(0) &= E\hat{f}^2(0) - \{E\hat{f}_n(0)\}^2 \\ &= \{(k-1)/n\}\{(n-1)/(k-2)\} - 1 \\ &= \frac{n-k+1}{n(k-2)} \end{aligned} \tag{4.32}$$

Using the conditions (4.9), the following results are immediate

$$\begin{aligned} &E\hat{f}_n(0) = 1 \\ &\operatorname{var}\hat{f}_n(0) \longrightarrow 0 \\ &\hat{f}_n(0) \xrightarrow{P} 1 = f(0). \end{aligned} \tag{4.33}$$

This is exactly as guaranteed by Theorem 4.1. For this special case, if the factor $k/(n+1)$ had been used rather than $(k-1)/n$ in defining $\hat{f}_n(0)$, then $\hat{f}_n(0)$ would have been biased upward. This is the first indication that using the factor $(k-1)/n$ in defining $\hat{f}_n(z)$ will lead to a better estimator than using $k/(n+1)$ will. The asymptotic results are the same in either case.

CHAPTER V

NONPARAMETRIC CLASSIFICATION

A nonparametric classification procedure is introduced in this chapter. It will be based on the density estimators developed in the preceding chapter and on Theorem 2.2. Invariance properties of the procedure will be discussed. Special cases will be considered and some comparison with the procedure given by Fix and Hodges based on Theorem 2.4 will be given.

5.1 A Nonparametric Classification Procedure

Let f and g be two density functions corresponding to absolutely continuous distribution functions F and G, respectively. F and G are unknown. Let z be a p-variate observation which is to be classified as belonging to F or to G. Samples $x_1,\ldots,x_n$ and $y_1,\ldots,y_m$ are given from F and G respectively. Consistent estimates for the densities f and g at z, $\hat{f}_n(z)$ and $\hat{g}_m(z)$, are given by Theorem 4.1. Let $\{k_1(n) = k_1\}$ and $\{k_2(m) = k_2\}$ be sequences of integers such that

$$(5.1)\qquad \begin{array}{ll} \lim_{n\to\infty} k_1 = \infty & \lim_{m\to\infty} k_2 = \infty \\ \lim_{n\to\infty} k_1/n = 0 & \lim_{m\to\infty} k_2/m = 0. \end{array}$$

Essentially what (5.1) says is choose k_1 and k_2 large but small compared to n and m respectively.

Let r_{k_1} and r_{k_2} be defined analogously to the way in which r_k was defined in section 4.2 of Chapter IV, (r_{k_1} is defined using the x's and

r_{k_2} is defined using the y's). Likewise, define $A_{r_{k_1},z}$, $A_{r_{k_2},z}$, $S_{r_{k_1},z}$ and $S_{r_{k_2},z}$ as $A_{r_k,z}$ and $S_{r_k,z}$ were defined in section 4.2 of Chapter IV.

Then

$$\hat{f}_n(z) = \{(k_1-1)/n\}\{1/A_{r_{k_1},z}\}$$

$$= \{(k_1-1)/n\}\{p\Gamma(p/2)/2r_{k_1}^p\,\pi^{p/2}\}.$$

(5.2) and

$$\hat{g}_m(z) = \{(k_2-1)/m\}\{1/A_{r_{k_2},z}\}$$

$$= \{(k_2-1)/m\}\{p\Gamma(p/2/2r_{k_2}^p\,\pi^{p/2}\}.$$

It should be noted that even if k_1 and k_2 are equal the r in $\hat{f}_n$ and the r in $\hat{g}_m$ are not in general equal. The notation will not be further complicated, at this time, in order to distinguish the r's. The classification procedure is then as follows:

(5.3)

$$\text{choose } F \text{ if} \quad \hat{f}_n(z)/\hat{g}_m(z) > c$$

$$\text{choose } G \text{ if} \quad \hat{f}_n(z)/\hat{g}_m(z) < c.$$

This procedure will henceforth be designated by $L(c,\hat{f}_n,\hat{g}_m)$. The choice of c depends on the relative importance of the two types of errors: classifying z as being from G when it is from F and vice-versa. For example, if c is chosen so that the probabilities of these two types of errors are equal the procedure is called a minimax procedure. The choice of c is not considered here but it is assumed to be given.

According to Theorem 2.2 and Theorem 4.1, $L(c,\hat{f}_n,\hat{g}_m)$ is consistent with $L(c)$. Since $L(c)$ is optimum with respect to minimizing the probabilities of the errors of misclassification, $L(c,\hat{f}_n,\hat{g}_m)$ will share these optimum properties as m and n get large. Thus, (5.3) provides a nonparametric classification procedure that has asymptotically optimum properties. It is not restricted to the one-dimensional situation or to any special cases.

If $k_1 = 1$ or $k_2 = 1$, the procedure (5.3) does not depend on the observations. Therefore, if $k_1 = 1$ or $k_2 = 1$ it is convenient to alter (5.3) slightly by changing the density estimators (5.2). The following change, which will be made, does not affect the asymptotic properties of the estimators (5.2). Thus, it does not affect the asymptotic properties of (5.3) either. If $k_1 = 1$ or $k_2 = 1$ the factors $(k_1-1)/n$ and $(k_2-1)/m$ will be changed to $k_1/(n+1) = 1/(n+1)$ and $k_2/(m+1) = 1/(m+1)$, respectively. These changes are assumed in effect for the following work.

5.2 Special Cases and a Comparison with the Fix and Hodges Procedure

(5.3) can be written in the following alternative form.

$$\text{Choose } F \text{ if} \quad \frac{(k_1-1)m}{(k_2-1)n} \cdot \frac{r^p_{k_2}}{r^p_{k_1}} > c$$

(5.4) and

$$\text{choose } G \text{ if} \quad \frac{(k_1-1)m}{(k_2-1)n} \cdot \frac{r^p_{k_2}}{r^p_{k_1}} < c\ .$$

Suppose that m = n. It seems reasonable then to choose $k_1 = k_2 = k$. Let ${}_1r$ and ${}_2r$ designate an r coming from samples from F and G respectively. Then $L(c,\hat{f}_n,\hat{g}_m)$ is as follows:

$$\text{(5.5)} \qquad \begin{array}{ll} \text{choose F if} & {}_2r_k^p/{}_1r_k^p > c \\ \text{choose G if} & {}_2r_k^p/{}_1r_k^p < c\,. \end{array}$$

In order to compare this procedure with that based on Theorem 2.4, X and Y are assumed to be univariate nonnegative random variables. Also, z is taken to be zero and f(0) and g(0) are assumed to be greater than zero. In view of the discussion that leads to (4.23), (5.4) can be simplified to the following:

$$\text{(5.6)} \qquad \begin{array}{ll} \text{choose F if} & \dfrac{(k_1-1)m}{(k_2-1)n} \cdot \dfrac{y_{(k_2)}}{x_{(k_1)}} > c \\ \text{choose G if} & \dfrac{(k_1-1)m}{(k_2-1)n} \cdot \dfrac{y_{(k_2)}}{x_{(k_1)}} < c\,. \end{array}$$

Under the further assumptions that m = n and that $k_1 = k_2 = k$, (5.6) can be simplified to the following:

$$\text{(5.7)} \qquad \begin{array}{ll} \text{choose F if} & y_{(k)}/x_{(k)} > c \\ \text{choose G if} & y_{(k)}/x_{(k)} < c \end{array}$$

The procedure based on Theorem 2.4 is $L^*(c,N/nU,M/mU)$. It is as follows:

$$(5.8)\quad \begin{array}{ll} \text{choose } F \text{ if} & N/n > c(M/m) \\ \text{choose } G \text{ if} & N/n < c(M/m). \end{array}$$

N and M are as defined in Theorem 2.4. If $n = m$, (5.8) reduces to the following:

$$(5.9)\quad \begin{array}{ll} \text{choose } F \text{ if} & N > cM \\ \text{choose } G \text{ if} & N < cM. \end{array}$$

For the moment, let $c = 1$. Then (5.7) says to choose F if the k^{th} ordered x, $x_{(k)}$, is less than the k^{th} ordered y, $y_{(k)}$, and choose G otherwise. (5.9) says choose F if out of the $k(n,m) = N + M$ x's and y's nearest zero there are more x's than y's and choose G otherwise. If the further restrictions $k = 1$ and $k(n,m) = 1$ are made, then both procedures are equivalent. Each procedure consists of classifying F if the closest observation to zero is an x and vice versa.

Fix and Hodges (1952) have considered this latter procedure for small values of n and have compared it to the procedure based upon the discriminant function under assumptions necessary for this latter procedure. It compares very favorably. Hence, the classification procedure of this chapter will also compare favorably to the discriminant function procedure in this last case.

5.3 Invariance Properties of $L(c,\hat{f}_n,\hat{g}_m)$

The classification procedure $L(c,\hat{f}_n,\hat{g}_m)$ has certain invariance properties that are examined in this section. Since $L(c,\hat{f}_n,\hat{g}_m)$ depends on the samples $x_1,\ldots,x_n$ and $y_1,\ldots,y_m$ only through $\hat{f}_n$ and $\hat{g}_m$ it follows that whenever $\hat{f}_n$ and $\hat{g}_m$ are invariant then $L(c,\hat{f}_n,\hat{g}_m)$ is also. Consequently, the invariance properties of $\hat{f}_n(z)$ are considered first.

It is reasonable to require that certain transformations of the n p-dimensional observations $x_1,x_2,\ldots,x_n$ and z will leave $\hat{f}_n(z)$ unchanged. Assume $x_1,\ldots,x_n$ and z are 1 x p vectors. Let b be a 1 x p vector of constants, i.e. $b = (b_1,\ldots,b_p)$. Consider the linear translation $z + b$, $x_1 + b,\ldots,x_n + b$. This translation has the effect of moving the whole density to a new location. Hence, the estimate of the density at the translated z point should be the same as that at the original point z. That is, $\hat{f}_n(z)$ should be the same as $\hat{f}^*_n(z + b)$, where the * on f indicates that the estimate is based on the translated variables $x_1 + b,\ldots,x_n + b$. This then is the first invariance property that $\hat{f}_n(z)$ should have.

Secondly, let Γ be a p x p orthogonal matrix. Consider the transformation $z\Gamma$, $x_1\Gamma,\ldots,x_n\Gamma$. This is a rotation and since it does not change the relative positions of the points $x_1,\ldots,x_n$ and z it should satisfy the same criterion that the translation did, namely that $\hat{f}_n(z)$ and $\hat{f}^*_n(z\Gamma)$ are equal.

Looking at (4.11) it is seen that $\hat{f}_n(z)$ depends on the sample $x_1,\ldots,x_n$ and z only through r_k. But, r_k is $d(x,z)$ for some x_i, $i=1,\ldots,n$. Therefore, it is only necessary to look at $d(x,z)$ in order to study the behavior of $\hat{f}_n(z)$ under the transformations discussed above. If $d(x,z)$ is invariant, so is $\hat{f}_n(z)$. But, $d(x,z)$ may be written as follows:

$$d(x,z) = \sqrt{(x-z)(x-z)'} \; . \tag{5.10}$$

Consider the translation by the vector b and let $d^*(x,z)$ designate the distance between the two translated variables. Then

$$\begin{aligned} d(x,z) &= \sqrt{(x-z)(x-z)'} \\ &= \sqrt{(x+b-(z+b))(x+b-(z+b))'} \\ &= d^*(x,z). \end{aligned} \tag{5.11}$$

Thus, $\hat{f}_n(z)$ is invariant under translations, since $d(x,z)$ is.

Let Γ be an orthogonal matrix of dimension p x p. Then

$$\begin{aligned} d(x,z) &= \sqrt{(x-z)(x-z)'} \\ &= \sqrt{(x-z)\Gamma\,\Gamma'(x-z)'} \\ &= \sqrt{(x\Gamma - z\Gamma)(x\Gamma - z\Gamma)'} \\ &= d(x\Gamma, z\Gamma) \\ &= d^*(x,z) \end{aligned} \tag{5.12}$$

Thus $\hat{f}_n(z)$ is seen to be invariant under an orthogonal transformation.

Since $\hat{f}_n$ and $\hat{g}_m$ are invariant under a translation or orthogonal transformation then $L(c,\hat{f}_n,\hat{g}_m)$ is also invariant under these transformations.

Now let c be a scalar. Transform $x_1,\ldots,x_n$, $y_1,\ldots,y_m$ and z to $cx_1,\ldots,cx_n$, $cy_1,\ldots,cy_m$ and cz. First, examine the result of this transformation on d(x,z) as above.

$$\begin{aligned} d(x,z) &= \sqrt{(x-z)(x-z)'} \\ &= (1/c)\sqrt{(cx-cz)(cx-cz)'} \qquad (5.13) \\ &= (1/c)d^*(x,z). \end{aligned}$$

Thus, $\hat{f}_n(z)$ is transformed to $(1/c)\hat{f}_n(z)$ and $\hat{g}_m(z)$ is transformed to $(1/c)\hat{g}_m(z)$ since $\hat{f}_n$ and $\hat{g}_m$ depend on $1/d(x,z)$. Since $L(c,\hat{f}_n,\hat{g}_m)$ depends only on the ratio $\hat{f}_n/\hat{g}_m$ it is invariant under this scale transformation. The 1/c in the numerator and denominator of the ratio cancel out. This concludes the discussion of the invariance properties of the classification procedure $L(c,\hat{f}_n,\hat{g}_m)$.

CHAPTER VI

AN EMPIRICAL STUDY OF $\hat{f}_n(z)$

When actually using the estimator $\hat{f}_n(z)$ proposed in this paper, several questions arise.

i) Since $\hat{f}_n(z) \xrightarrow{P} f(z)$, how large should n be for $\hat{f}_n(z)$ to be very near $f(z)$?

ii) How should one choose the constant k used in the estimator $\hat{f}_n(z)$?

iii) How does this estimator compare with other estimators? In particular, how does it compare with an estimator obtained by assuming a particular parametric form for the density function f and then using maximum likelihood estimation to obtain estimates for the unknown parameters in the density function?

In order to illuminate some of these questions, a small empirical study was carried out on an IBM 1620 computer. Three specific distributions (the uniform, the exponential and the normal) were treated in this study. The purpose was not to find concrete answers to the above questions, but to try to obtain some feeling for the behavior of the estimator $\hat{f}_n(z)$ and at the same time perhaps get some indication of what the answers to the above questions might be. The exact study carried out, the results, etc. will be presented in the following sections.

6.1 Generation of Random Variables from Uniform, Exponential and Normal Distributions

Much of the work in this empirical study involved generating random samples from the above three distributions. The techniques used are briefly discussed here.

The three density functions involved are:

i) Uniform

$$f(x) = 1 \qquad 0 \leq x \leq 1$$
$$= 0 \qquad \text{otherwise;}$$

ii) Exponential

$$f(x;\beta) = (1/\beta)e^{-(x/\beta)} \qquad (\beta > 0;\ 0 \leq x \leq \infty)$$
$$= 0 \qquad \text{otherwise;}$$

iii) Normal

$$f(x;\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[\frac{-1}{2}\frac{(x-\mu)^2}{\sigma^2}\right] \qquad \begin{array}{l} -\infty < x < \infty \\ -\infty < \mu < \infty \\ 0 < \sigma^2. \end{array}$$

There is a random number generator subroutine built into the IBM 1620 Fortran programming system. This subroutine generates numbers uniformly on the interval (0,1) and this subroutine was used to generate random numbers from the uniform distribution.

A sample $x_1,\ldots,x_n$ from the exponential distribution was generated by using the probability integral transformation. First, a sample $u_1,\ldots,u_n$ was generated from the uniform distribution using the subroutine described above. Set $U = F(x;\beta)$ where F is the distribution function for the exponential distribution, i.e. $U = 1 - e^{-x/\beta}$. Then

$x = -\beta \ln(1 - U)$ is distributed as an exponential variate, hence $x_1 = -\beta \ln(1 - u_1)$, $x_2 = -\beta \ln(1 - u_2), \ldots, x_n = -\beta \ln(1 - u_n)$ is a sample of n from the exponential distribution.

In the normal case, it is not possible to use the technique described in the preceding paragraph as the distribution function F for the normal distribution does not have a closed form as the exponential distribution function does. That is, $F(x;\mu,\sigma^2)$ for the normal distribution involves an integral which can only be evaluated numerically. A search for some alternative method of generating random normal variables led to a method proposed by Box and Muller (1958) which will be described below. Muller (1959) indicates that this is a satisfactory procedure.

The Box and Muller technique of generating normal random variables again makes use of the uniform random number generator subroutine. Let U_1, U_2 be independent randon variables from the same uniform density function on the interval $[0,1]$. Let

$$x_1 = (-2\ln U_1)^{1/2} \cos 2\pi U_2$$

$$x_2 = (-2\ln U_1)^{1/2} \sin 2\pi U_2.$$

Then (x_1,x_2) will be a pair of independent random variables from a normal distribution with mean zero, and variance one. These two normal variables can now be transformed to two other normal variables having any desired mean and variance. Proceeding in this manner, we can generate a random sample of any desired size.

6.2 Maximum Likelihood Estimates for Normal and Exponential Distributions

The maximum likelihood estimate of the parameter β in the exponential distribution, based on a sample $x_1,\ldots,x_n$ from this distribution is $\hat{\beta} = \bar{x}$. The maximum likelihood estimates of the parameters μ and σ^2 in the normal distribution, based on a sample of $x_1,\ldots,x_n$ from this distribution are $\hat{\mu}_n = \bar{x}$ and $\hat{\sigma}^2_n = s^2$ respectively, where

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

By well known properties of maximum likelihood estimators, $\hat{\beta}_n \xrightarrow{P} \beta$, $\hat{\mu}_n \xrightarrow{P} \mu$, $\hat{\sigma}^2_n \xrightarrow{P} \sigma^2$, $f(x;\hat{\beta}_n) \xrightarrow{P} f(x;\beta)$ and $f(x;\hat{\mu}_n,\hat{\sigma}^2_n) \xrightarrow{P} f(x;\mu,\sigma^2)$.

6.3 Description of the Empirical Study and the Results

The results of this empirical study had a significant effect on earlier chapters of this thesis and that effect is now pointed out prior to discussing the empirical study. In this empirical study, the estimator $\hat{f}_n(z)$ as given in (4.11) was not used. Instead, an estimator using the factor $k/(n+1)$ rather than $(k-1)/n$ was used. Using this estimator for f led to estimates which were biased upward a large percentage of the time. This will readily be seen when the results are presented later in this section. As a result of this, a search was started for an estimator which would not have this strong tendency to be biased upward. In the last section of Chapter IV a special case was considered and there it was observed that using the factor $(k-1)/n$ instead of $k/(n+1)$ led to an unbiased estimator at a certain point. Thus, the estimates obtained in this study were converted to new estimates based on the estimator (4.11) rather than that originally used. It was observed that approximately one-half of all estimates obtained in this study were now lower than their theoretical value and approximately one-half were now higher. This seemed to be sufficient evidence to recommend using (4.11) rather than some other asymptotically equivalent estimator for f, such as the one described above.

The empirical study made use of the results in the previous sections of this chapter and of the work on the proposed estimator $\hat{f}_n(z)$ in Chapter IV. A description of the empirical study undertaken will be presented first and it will be divided into three parts as will the presentation of the results. The three parts will consist of 1) the uniform distribution case, 2) the exponential distribution case and 3) the normal distribution case.

No attempt was made in the uniform distribution case to compare the estimator of this paper with the maximum likelihood estimator. Table I of the Appendix contains the various problems considered. This table will be explained in some detail, as Tables III, V, VII, and IX are very similar. Consider problem U-3 in Table I. A sample of 25 was generated from the uniform distribution and estimates of f(x) were calculated at the points $z = .05, .25, .50, .75$ and $.95$ using $\hat{f}_n(z)$ with $k = 3$, i.e. $\hat{f}_{25}(.05)$, $\hat{f}_{25}(.25)$, $\hat{f}_{25}(.50)$, $\hat{f}_{25}(.75)$ and $\hat{f}_{25}(.95)$ were calculated. The estimator $\hat{f}_n(z)$ is as described in the first paragraph of this section, rather than (4.11), and this is also true for the normal and exponential cases considered in the following two paragraphs. This procedure was repeated ten times giving ten different estimates of $f(x)$ at each of the five z points. The means at each of these points, $\bar{\hat{f}}_{25}(.05)$, $\bar{\hat{f}}_{25}(.25)$, $\bar{\hat{f}}_{25}(.50)$, $\bar{\hat{f}}_{25}(.75)$ and $\bar{\hat{f}}_{25}(.95)$, were calculated as well as the sample variances of each of these means. These sample variances will be designated by $\widehat{\text{var}}\ \bar{\hat{f}}_n(z)$. In all cases k is chosen near $n^{1/2}$. For the cases where the number of observations per sample is 25 and 100, k is varied slightly above and slightly below $n^{1/2}$ as well (e.g. when the number of observations per sample is 25, 3 values of k, 3, 5 and 7, are used). The results of the problems in Table I are given in Table II. Problem numbers in Table II correspond to those in Table I. Any time the same problem was run with 10 samples and also with 50 samples, these results appear side by side. This occurs in almost all cases as can be seen by looking at Table I. The same procedure will be followed in Tables IV, VI, VIII and X. Table II is self-explanatory when used with Table I. For convenience, the theoretical values of the density at each point where it was estimated are listed here. They are $f(.05) = f(.25) = f(.50) = f(.75) = f(.95) = 1$.

The exponential case was considered next. Problems were run with the parameter β in the exponential distributions equal to .5 and 2.0. Using the nonparametric estimator $\hat{f}_n(z)$, essentially the same type of problems were run as in the uniform case for both $\beta = .5$ and $\beta = 2.0$. The problems are given in Table III, which reads the same as Table I. Next, maximum likelihood estimation was considered. All of the problems run corresponded to similar situations in the nonparametric case, and they are listed in Table V. Consider problem number E2-3 in Table V. A sample of 25 was generated from an exponential distribution with $\beta = .5$. $\bar{x}$, the maximum likelihood estimate of β, was calculated. $\bar{x}$ was substituted in the exponential density function for β and estimates of $f(x)$ at $z = .25$, $.50$, 1.0 and 2.0 were calculated. This was repeated ten times, giving ten estimates of $f(x)$ at these four z points. The mean of these estimates at each of the points, designated by $\bar{\hat{f}}(z)$, was calculated as well as the sample variance of each of these means, $\widehat{\text{var}}\ \bar{\hat{f}}(z)$. The results of problem number E2-3 are comparable to those of E1-3, E1-5 and E1-7 in the nonparametric case. The results of the problems listed in Tables III and V are given in Tables IV and VI, respectively. The theoretical values of $f(x)$ at the points where it was estimated are: $\beta = .5$, $f(.25) = 1.21306$, $f(.50) = .73576$, $f(1.00) = .27067$, $f(2.00) = .03663$; $\beta = 2.0$, $f(.25) = .44125$, $f(.50) = .38940$, $f(1.00) = .30327$, $f(2.00) = .18394$. These values as well as those for the normal distribution given below are also given on the first page of the appendix for convenience.

The normal distribution was the last case considered. Only problems for the normal distribution with mean zero and variance one were considered. The problems run for the nonparametric estimator for the normal distribution are very similar to those run for the uniform and exponential distributions and are given in Table VII. The problems run for the maximum likelihood case correspond to the nonparametric problems the same as in the exponential case. The maximum likelihood procedure here is exactly as in the exponential case except that $\bar{x}$ and s^2, the mean and sample variance, are substituted in the normal density function each time for μ and σ^2 rather than $\bar{x}$ for β. The estimator s^2 is not the precise maximum likelihood estimator but is unbiased. The problems run are listed in Table IX and the results of problems in Tables VII and IX are given in Tables VIII and X. The values of f(x) at the points where it was estimated are: $f(-2.5) = f(2.5) = .01753$, $f(-1.0) = f(1.0) = .24197$ and $f(0) = .39894$.

6.4 General Remarks on the Empirical Study

As was pointed out earlier, it is readily seen that a fairly large proportion of the estimates as originally obtained in the empirical study were high. It was not surprising that the estimator originally used was biased in view of the statement proved by Rosenblatt (1956). He proved that any nonparametric density function estimator is biased. However, the nonparametric density estimator (4.11) seems to have less bias than the one used in this empirical study.

The value of k was chosen near $n^{1/2}$. It was varied slightly in a few cases. In these cases it appears that $k = n^{1/2}$ or $k > n^{1/2}$ gave better results than $k < n^{1/2}$. The evidence is not too strong, especially since few cases were considered. The choice of an optimum k is a point which would be worth investigating.

As the number of observations per sample increases, the trend is for the various nonparametric estimates of $f(x)$ to improve. This is expected in view of the asymptotic properties of the estimate used. No definite conclusion can be drawn concerning how large n should be for the estimate to reach a certain accuracy. This point appears worthy of further investigation.

A few comparisons of the estimates obtained by using the nonparametric estimator and those obtained by using maximum likelihood estimation shows that the maximum likelihood procedure appears to be substantially better. This is true even if the nonparametric estimates obtained in this study are converted to the estimates of (4.11). This is not at all surprising because of the large amount of additional information assumed in the maximum likelihood case. However, in those situations where it appears that there is simply not enough evidence to warrant any parametric assumptions, $\hat{f}_n(x)$ is a reasonable estimator to use.

CHAPTER VII

SUMMARY

The first part of Chapter II was a review of work done in the area of nonparametric classification. It was pointed out that the nonparametric classification problem can be closely related to the problem of nonparametric density estimation. As a result of this, a major portion of the work in this thesis was concerned with nonparametric density estimation.

In Chapter IV, a nonparametric density function estimator was introduced. It was shown that it was basically quite different than most other nonparametric density estimators appearing in the literature. The main result of this chapter is contained in Theorem 4.1, which shows the consistency of the nonparametric density estimator $\hat{f}_n(z)$. The similarities between this estimator and one introduced by Fix and Hodges (1951) are discussed for a special case. The expected value and variance of $\hat{f}_n(z)$, for a special case, are also found.

A nonparametric classification procedure based on $\hat{f}_n(z)$, as introduced in Chapter IV, is given in Chapter V. This procedure is shown to have certain optimum properties asymptotically. A brief comparison of the similarities between this procedure and one introduced by Fix and Hodges (1951) is made for a special case. Invariance properties of this classification procedure are discussed briefly.

An empirical study was made on the nonparametric estimator $\hat{f}_n(z)$. It provided some feeling for the behavior of $\hat{f}_n(z)$ for finite sample size. This study also led to a density estimator which appears to be less biased than one that had been considered earlier.

Several questions concerning $\hat{f}_n(z)$ which are as yet unanswered were mentioned in the last section of Chapter VI. Similar questions can be asked concerning the nonparametric classification procedure given in Chapter V. For example, what is the behavior of this classification procedure for finite n? Several of these questions are also worthy of further investigation.

APPENDIX

For convenience in reading the tables on the following pages of this appendix the theoretical values are given here for all points where a density was estimated.

Uniform Distribution

$$f(.05) = f(.25) = f(.50) = f(.75) = f(.95) = 1$$

Exponential Distribution

i) $\beta = .5$ $\quad f(.25) = 1.21306 \quad f(.50) = .73576$

$\quad f(1.00) = .27067 \quad f(2.00) = .03663$

ii) $\beta = 2.0$ $\quad f(.25) = .44125 \quad f(.50) = .38940$

$\quad f(1.00) = .30327 \quad f(2.00) = .18394$

Normal Distribution

$$\mu = 1 \qquad \sigma^2 = 1$$

$$f(-2.5) = f(2.5) = .01753$$

$$f(-1.0) = f(1.0) = .24197$$

$$f(0) = .39894$$

TABLE I

SUMMARY OF UNIFORM DISTRIBUTION PROBLEMS RUN IN THE EMPIRICAL STUDY FOR THE NONPARAMETRIC CASE

Prob. no.	No. of obs. per sample=n	No. of samples	k	The point z where the density was estimated				
U-1	10	10	3	.05	.25	.50	.75	.95
U-2	10	50	3	.05	.25	.50	.75	.95
U-3	25	10	3	.05	.25	.50	.75	.95
U-4	25	50	3	.05	.25	.50	.75	.95
U-5	25	10	5	.05	.25	.50	.75	.95
U-6	25	50	5	.05	.25	.50	.75	.95
U-7	25	10	7	.05	.25	.50	.75	.95
U-8	25	50	7	.05	.25	.50	.75	.95
U-9	50	10	7	.05	.25	.50	.75	.95
U-10	50	50	7	.05	.25	.50	.75	.95
U-11	100	10	7		.25			.95
U-12	100	50	7		.25			.95
U-13	100	10	10		.25			.95
U-14	100	50	10		.25			.95
U-15	100	10	12		.25			.95
U-16	100	50	12		.25			.95
U-17	500	10	22		.25			.95
U-18	1000	10	30		.25			.95

TABLE II

RESULTS OF UNIFORM DISTRIBUTION PROBLEMS LISTED IN TABLE I

Prob. no.	z	$\bar{\hat{f}}_n(z)$	$\widehat{var}\ \bar{\hat{f}}_n(z)$	Prob. no.	z	$\bar{\hat{f}}_n(z)$	$\widehat{var}\ \bar{\hat{f}}_n(z)$
	.05	.96231	.04377		.05	.94235	.01936
	.25	1.26147	.02382		.25	1.48632	.03612
U-1	.50	1.11745	.04544	U-2	.50	1.29170	.01691
	.75	1.14999	.06103		.75	1.29347	.01167
	.95	.78459	.02389		.95	.94966	.01382
	.05	1.46334	.13073		.05	1.20370	.01517
	.25	1.65453	.02308		.25	1.54005	.02064
U-3	.50	1.61989	.24283	U-4	.50	1.41832	.01839
	.75	1.94361	.20868		.75	2.54434	.73322
	.95	.86447	.02306		.95	1.71656	.05921
	.05	1.20246	.09253		.05	.97372	.01784
	.25	1.30236	.01402		.25	1.08661	.00331
U-5	.50	1.03680	.01831	U-6	.50	1.27768	.01231
	.75	1.15362	.01925		.75	1.12560	.00741
	.95	.89503	.05301		.95	1.09554	.01557
	.05	.73812	.01143		.05	.71917	.00384
	.25	1.04738	.01621		.25	1.10768	.00374
U-7	.50	.96931	.00353	U-8	.50	1.10590	.00253
	.75	1.23992	.01328		.75	1.11329	.00356
	.95	.77807	.01024		.95	.79193	.00862
	.05	1.12740	.03412		.05	1.02342	.00988
	.25	1.18992	.03548		.25	1.09320	.00308
U-9	.50	.96097	.00504	U-10	.50	1.20062	.00599
	.75	1.10176	.02456		.75	1.14133	.00679
	.95	1.04352	.01272		.95	1.08909	.00481

TABLE II (cont.)

Prob. no.	z	$\bar{\hat{f}}_n(z)$	$\widehat{\text{var}}\ \bar{\hat{f}}_n(z)$	Prob. no.	z	$\bar{\hat{f}}_n(z)$	$\widehat{\text{var}}\ \bar{\hat{f}}_n(z)$
U-11	.25	1.05055	.01117	U-12	.25	1.09559	-.00257
	.95	1.13575	.01031		.95	1.06274	.00362
U-13	.25	1.08446	.00562	U-14	.25	1.07590	.00246
	.95	1.18658	.02186		.95	1.00803	.00196
U-15	.25	.88705	.00591	U-16	.25	1.07085	.00145
	.95	.73289	.00811		.95	1.01370	.00238
U-17	.25	1.26483	.00904	U-18	.25	1.00901	.00702
	.95	.91601	.00161		.95	1.06097	.00327

TABLE III

SUMMARY OF EXPONENTIAL DISTRIBUTION PROBLEMS RUN IN THE EMPIRICAL STUDY FOR THE NONPARAMETRIC CASE

Prob. no.	No. of obs. per sample=n	No. of samples	k	β	The points z where the density was estimated			
E1-1	10	10	3	.5	.25	.50	1.0	2.0
E1-2	10	50	3	.5	.25	.50	1.0	2.0
E1-3	25	10	3	.5	.25	.50	1.0	2.0
E1-4	25	50	3	.5	.25	.50	1.0	2.0
E1-5	25	10	5	.5	.25	.50	1.0	2.0
E1-6	25	50	5	.5	.25	.50	1.0	2.0
E1-7	25	10	7	.5	.25	.50	1.0	2.0
E1-8	25	50	7	.5	.25	.50	1.0	2.0
E1-9	50	10	7	.5	.25	.50	1.0	2.0
E1-10	50	50	7	.5	.25	.50	1.0	2.0
E1-11	100	10	7	.5	.25		1.0	
E1-12	100	50	7	.5	.25		1.0	
E1-13	100	10	10	.5	.25		1.0	
E1-14	100	50	10	.5	.25		1.0	
E1-15	100	10	12	.5	.25		1.0	
E1-16	100	50	12	.5	.25		1.0	
E1-17	500	10	22	.5	.25		1.0	
E1-18	1000	10	30	.5	.25		1.0	
E1-19	10	10	3	2.0	.25	.50	1.0	2.0
E1-20	10	50	3	2.0	.25	.50	1.0	2.0
E1-21	25	10	3	2.0	.25	.50	1.0	2.0
E1-22	25	50	3	2.0	.25	.50	1.0	2.0
E1-23	25	10	5	2.0	.25	.50	1.0	2.0
E1-24	25	50	5	2.0	.25	.50	1.0	2.0
E1-25	25	10	7	2.0	.25	.50	1.0	2.0
E1-26	25	50	7	2.0	.25	.50	1.0	2.0
E1-27	50	10	7	2.0	.25	.50	1.0	2.0
E1-28	50	50	7	2.0	.25	.50	1.0	2.0
E1-29	100	10	7	2.0	.25		1.0	
E1-30	100	50	7	2.0	.25		1.0	
E1-31	100	10	10	2.0	.25		1.0	
E1-32	100	50	10	2.0	.25		1.0	
E1-33	100	10	12	2.0	.25		1.0	
E1-34	100	50	12	2.0	.25		1.0	
E1-35	500	10	22	2.0	.25		1.0	
E1-36	1000	10	30	2.0	.25		1.0	

TABLE IV

RESULTS OF EXPONENTIAL DISTRIBUTION PROBLEMS LISTED IN TABLE III

Prob. no.	z	$\bar{\hat{f}}_n(z)$	var $\bar{\hat{f}}_n(z)$	Prob. No.	z	$\bar{\hat{f}}_n(z)$	var $\bar{\hat{f}}_n(z)$
	.25	1.98562	.07842		.25	2.37902	.38881
El-1	.50	.75498	.00670	El-2	.50	1.21984	.02899
	1.00	.34732	.00244		1.00	.31804	.00030
	2.00	.10313	.00003		2.00	.10922	.00001
	.25	1.38593	.04492		.25	1.46637	.02495
El-3	.50	1.02351	.06263	El-4	.50	.99896	.01185
	1.00	.40374	.00211		1.00	.43792	.00500
	2.00	.06157	.00001		2.00	.07828	.00004
	.25	1.51422	.01986		.25	1.48670	.01173
El-5	.50	.91269	.01425	El-6	.50	.90736	.00686
	1.00	.28414	.00054		1.00	.37968	.00146
	2.00	.09105	.00009		2.00	.08709	.00001
	.25	1.54477	.05097		.25	1.35898	.00498
El-7	.50	.73138	.00210	El-8	.50	.87525	.00198
	1.00	.35359	.00087		1.00	.30805	.00009
	2.00	.10219	.00002		2.00	.10095	.00000
	.25	1.29015	.03320		.25	1.65895	.01368
El-9	.50	.64764	.00184	El-10	.50	.74732	.00113
	1.00	.35592	.00085		1.00	.35554	.00204
	2.00	.07535	.00002		2.00	.07228	.00000
El-11	.25	1.23271	.03611	El-12	.25	1.44500	.00953
	1.00	.27065	.00116		1.00	.28623	.00018
El-13	.25	1.16631	.01363	El-14	.25	1.34312	.00487
	1.00	.30122	.00032		1.00	.29822	.00021

TABLE IV (cont.)

Prob. no.	z	$\bar{\hat{f}}_n(z)$	$\widehat{\text{var}}\ \bar{\hat{f}}_n(z)$	Prob. no.	z	$\bar{\hat{f}}_n(z)$	$\widehat{\text{var}}\ \bar{\hat{f}}_n(z)$
E1-15	.25	1.27754	.00988	E1-16	.25	1.27768	.00390
	1.00	.27436	.00029		1.00	.31312	.00025
E1-17	.25	1.20023	.01380	E1-18	.25	1.25589	.00521
	1.00	.27561	.00012		1.00	.26545	.00026
	.25	.60371	.01230		.25	.56046	.00680
E1-19	.50	.54145	.01248	E1-20	.50	.51860	.00274
	1.00	.41814	.00380		1.00	.39516	.00121
	2.00	.23648	.00101		2.00	.26877	.00071
	.25	.56115	.01761		.25	.66830	.00573
E1-21	.50	.95150	.12177	E1-22	.50	.66625	.02194
	1.00	.64018	.08192		1.00	.41535	.00152
	2.00	.19830	.00043		2.00	.25973	.00047
	.25	.43824	.00321		.25	.51852	.00293
E1-23	.50	.55152	.01024	E1-24	.50	.48579	.00132
	1.00	.34382	.00247		1.00	.30814	.00032
	2.00	.20135	.00074		2.00	.21617	.00019
	.25	.42013	.00473		.25	.43471	.00089
E1-25	.50	.43635	.00166	E1-26	.50	.45066	.00039
	1.00	.31229	.00052		1.00	.36595	.00029
	2.00	.19696	.00007		2.00	.20278	.00007
	.25	.47175	.00285		.25	.51363	.00089
E1-27	.50	.38321	.00129	E1-28	.50	.45625	.00062
	1.00	.31698	.00138		1.00	.32658	.00021
	2.00	.22060	.00094		2.00	.19267	.00007

TABLE IV (cont.)

Prob. no.	z	$\hat{\bar{f}}_n(z)$	$\widehat{var}\ \hat{\bar{f}}_n(z)$	Prob. no.	z	$\hat{\bar{f}}_n(z)$	$\widehat{var}\ \hat{\bar{f}}_n(z)$
E1-29	.25	.52935	.00423	E1-30	.25	.50426	.00150
	1.00	.33075	.00076		1.00	.36398	.00042
E1-31	.25	.39791	.00092	E1-32	.25	.48466	.00033
	1.00	.32693	.00045		1.00	.33608	.00028
E1-33	.25	.45036	.00053	E1-34	.25	.47853	.00037
	1.00	.33195	.00056		1.00	.32523	.00025
E1-35	.25	.51841	.00097	E1-36	.25	.48259	.00070
	1.00	.33125	.00057		1.00	.28969	.00021

TABLE V

SUMMARY OF EXPONENTIAL DISTRIBUTION PROBLEMS RUN IN THE EMPIRICAL STUDY FOR THE MAXIMUM LIKELIHOOD CASE

Prob. no.	No. of obs. per sample=n	No. of samples	β	The points z where the density was estimated			
E2-1	10	10	.5	.25	.50	1.0	2.0
E2-2	10	50	.5	.25	.50	1.0	2.0
E2-3	25	10	.5	.25	.50	1.0	2.0
E2-4	25	50	.5	.25	.50	1.0	2.0
E2-5	50	10	.5	.25	.50	1.0	2.0
E2-6	50	50	.5	.25	.50	1.0	2.0
E2-7	100	10	.5	.25		1.0	
E2-8	100	50	.5	.25		1.0	
E2-9	500	10	.5	.25		1.0	
E2-10	1000	10	.5	.25		1.0	
E2-11	10	10	2.0	.25	.50	1.0	2.0
E2-12	10	50	2.0	.25	.50	1.0	2.0
E2-13	25	10	2.0	.25	.50	1.0	2.0
E2-14	25	50	2.0	.25	.50	1.0	2.0
E2-15	50	10	2.0	.25	.50	1.0	2.0
E2-16	50	50	2.0	.25	.50	1.0	2.0
E2-17	100	10	2.0	.25		1.0	
E2-18	100	50	2.0	.25		1.0	
E2-19	500	10	2.0	.25		1.0	
E2-20	1000	10	2.0	.25		1.0	

TABLE VI

RESULTS OF EXPONENTIAL DISTRIBUTION PROBLEMS LISTED IN TABLE V

Prob. no.	z	$\bar{\hat{f}}_n(z)$	$\widehat{\text{var}}\ \bar{\hat{f}}_n(z)$	Prob. no.	z	$\bar{\hat{f}}_n(z)$	$\widehat{\text{var}}\ \bar{\hat{f}}_n(z)$
	.25	1.28463	.00251		.25	1.19045	.00060
E2-1	.50	.65745	.00128	E2-2	.50	.70168	.00003
	1.00	.20387	.00114		1.00	.26226	.00012
	2.00	.02783	.00007		2.00	.04535	.00002
	.25	1.26504	.00114		.25	1.22792	.00035
E2-3	.50	.71880	.00005	E2-4	.50	.71313	.00002
	1.00	.23912	.00029		1.00	.25257	.00008
	2.00	.02953	.00003		2.00	.03690	.00001
	.25	1.21988	.00082		.25	1.17902	.00012
E2-5	.50	.72762	.00001	E2-6	.50	.72896	.00000
	1.00	.26330	.00018		1.00	.28195	.00002
	2.00	.03667	.00003		2.00	.04393	.00000
E2-7	.25	1.16868	.00053	E2-8	.25	1.21115	.00006
	1.00	.28685	.00009		1.00	.26996	.00001
E2-9	.25	1.19919	.00005	E2-10	.25	1.21698	.00004
	1.00	.27652	.00001		1.00	.26870	.00001
	.25	.42249	.00218		.25	.45828	.00035
E2-11	.50	.36899	.00120	E2-12	.50	.39664	.00018
	1.00	.28342	.00031		1.00	.29922	.00004
	2.00	.17174	.00001		2.00	.17473	.00000
	.25	.42936	.00066		.25	.44530	.00042
E2-13	.50	.37852	.00037	E2-14	.50	.39118	.00023
	1.00	.29478	.00010		1.00	.30230	.00006
	2.00	.18016	.00000		2.00	.18150	.00000

TABLE VI (cont.)

Prob. no.	z	$\bar{\hat{f}}_n(z)$	$\widehat{\text{var}}\ \bar{\hat{f}}_n(z)$	Prob. no.	z	$\bar{\hat{f}}_n(z)$	$\widehat{\text{var}}\ \bar{\hat{f}}_n(z)$
	.25	.44273	.00024		.25	.45475	.00006
E2-15	.50	.38985	.00014	E2-16	.50	.39865	.00003
	1.00	.30250	.00004		1.00	.30665	.00001
	2.00	.18262	.00000		2.00	.18216	.00000
E2-17	.25	.44462	.00008	E2-18	.25	.44412	.00003
	1.00	.30413	.00001		1.00	.30348	.00000
E2-19	.25	.44873	.00003	E2-20	.25	.45276	.00000
	1.00	.30602	.00000		1.00	.30769	.00000

TABLE VII

SUMMARY OF NORMAL DISTRIBUTION PROBLEMS RUN IN THE EMPIRICAL STUDY FOR THE NONPARAMETRIC CASE

Prob. no.	No. of obs. per sample=n	No. of samples	k	The points z where the density was estimated				
N1-1	10	10	3	-2.5	-1.0	0.0	1.0	2.5
N1-2	10	50	3	-2.5	-1.0	0.0	1.0	2.5
N1-3	26	10	3	-2.5	-1.0	0.0	1.0	2.5
N1-4	26	50	3	-2.5	-1.0	0.0	1.0	2.5
N1-5	26	10	5	-2.5	-1.0	0.0	1.0	2.5
N1-6	26	50	5	-2.5	-1.0	0.0	1.0	2.5
N1-7	26	10	7	-2.5	-1.0	0.0	1.0	2.5
N1-8	26	50	7	-2.5	-1.0	0.0	1.0	2.5
N1-9	50	10	7	-2.5	-1.0	0.0	1.0	2.5
N1-10	50	50	7	-2.5	-1.0	0.0	1.0	2.5
N1-11	100	10	7	-2.5			1.0	
N1-12	100	50	7	-2.5			1.0	
N1-13	100	10	10	-2.5			1.0	
N1-14	100	50	10	-2.5			1.0	
N1-15	100	10	12	-2.5			1.0	
N1-16	100	50	12	-2.5			1.0	
N1-17	500	10	22	-2.5			1.0	
N1-18	1000	10	30	-2.5			1.0	

TABLE VIII

RESULTS OF NORMAL DISTRIBUTION PROBLEMS LISTED IN TABLE VII

Prob. no.	z	$\hat{\bar{f}}_n(z)$	$\widehat{var}\ \hat{\bar{f}}_n(z)$	Prob. no.	z	$\hat{\bar{f}}_n(z)$	$\widehat{var}\ \hat{\bar{f}}_n(z)$
	-2.5	.07076	.00001		-2.5	.07834	.00001
	-1.0	.31022	.00130		-1.0	.31198	.00089
N1-1	0.0	.47612	.01226	N1-2	0.0	.57118	.00270
	1.0	.24924	.00061		1.0	.33845	.00129
	2.5	.07609	.00002		2.5	.07775	.00003
	-2.5	.04278	.00001		-2.5	.04722	.00000
	-1.0	.34028	.00557		-1.0	.38081	.00177
N1-3	0.0	.63971	.01897	N1-4	0.0	.70590	.01704
	1.0	.26811	.00535		1.0	.33897	.00078
	2.5	.05436	.00004		2.5	.05266	.00001
	-2.5	.05935	.00002		-2.5	.05842	.00000
	-1.0	.24620	.00179		-1.0	.29180	.00041
N1-5	0.0	.46142	.00482	N1-6	0.0	.48260	.00153
	1.0	.27469	.00064		1.0	.31206	.00046
	2.5	.06506	.00003		2.5	.06167	.00000
	-2.5	.06987	.00001		-2.5	.07317	.00000
	-1.0	.23109	.00044		-1.0	.26401	.00020
N1-7	0.0	.45013	.00150	N1-8	0.0	.44572	.00060
	1.0	.28510	.00070		1.0	.24580	.00011
	2.5	.07328	.00001		2.5	.07021	.00000
	-2.5	.05114	.00001		-2.5	.04900	.00000
	-1.0	.35095	.00213		-1.0	.29391	.00036
N1-9	0.0	.49571	.00464	N1-10	0.0	.49898	.00156
	1.0	.24445	.00032		1.0	.26113	.00019
	2.5	.04725	.00000		2.5	.05182	.00000

TABLE VIII (cont.)

Prob. no.	z	$\hat{\bar{f}}_n(z)$	$\widehat{\text{var}}\ \hat{\bar{f}}_n(z)$	Prob. no.	z	$\hat{\bar{f}}_n(z)$	$\widehat{\text{var}}\ \hat{\bar{f}}_n(z)$
N1-11	-2.5	.03365	.00000	N1-12	-2.5	.03680	.00000
	1.0	.25548	.00133		1.0	.30344	.00074
N1-13	-2.5	.04337	.00000	N1-14	-2.5	.04169	.00000
	1.0	.27202	.00041		1.0	.23958	.00008
N1-15	-2.5	.04644	.00000	N1-16	-2.5	.04538	.00000
	1.0	.23296	.00018		1.0	.25792	.00012
N1-17	-2.5	.02983	.00000	N1-18	-2.5	.02287	.00000
	1.0	.24754	.00031		1.0	.27420	.00078

TABLE IX

SUMMARY OF NORMAL DISTRIBUTION PROBLEMS RUN IN THE EMPIRICAL STUDY FOR THE MAXIMUM LIKELIHOOD CASE

Prob. no.	No. of obs. per sample	No. of samples	The points z where the density was estimated				
N2-1	10	10	-2.5	-1.0	0.0	1.0	2.5
N2-2	10	50	-2.5	-1.0	0.0	1.0	2.5
N2-3	26	10	-2.5	-1.0	0.0	1.0	2.5
N2-4	26	50	-2.5	-1.0	0.0	1.0	2.5
N2-5	50	10	-2.5	-1.0	0.0	1.0	2.5
N2-6	50	50	-2.5	-1.0	0.0	1.0	2.5
N2-7	100	10	-2.5			1.0	
N2-8	100	50	-2.5			1.0	
N2-9	500	10	-2.5			1.0	
N2-10	1000	10	-2.5			1.0	

TABLE X

RESULTS OF NORMAL DISTRIBUTION PROBLEMS LISTED IN TABLE IX

Prob. no.	z	$\bar{\hat{f}}(z)$	$\widehat{\text{var}}\ \bar{\hat{f}}(z)$	Prob. no.	z	$\bar{\hat{f}}(z)$	$\widehat{\text{var}}\ \bar{\hat{f}}(z)$
	-2.5	.02175	.00004		-2.5	.02329	.00001
	-1.0	.23056	.00163		-1.0	.22012	.00015
E2-1	0.0	.38799	.00076	E2-2	0.0	.40475	.00022
	1.0	.24068	.00114		1.0	.23774	.00015
	2.5	.02746	.00010		2.5	.02570	.00001
	-2.5	.02772	.00004		-2.5	.01901	.00000
	-1.0	.24610	.00015		-1.0	.23125	.00005
E2-3	0.0	.38229	.00025	E2-4	0.0	.40400	.00010
	1.0	.22999	.00017		1.0	.23938	.00005
	2.5	.02066	.00001		2.5	.02276	.00001
	-2.5	.01821	.00001		-2.5	.01704	.00000
	-1.0	.24935	.00010		-1.0	.23407	.00003
E2-5	0.0	.40841	.00011	E2-6	0.0	.40041	.00003
	1.0	.23139	.00009		1.0	.24534	.00003
	2.5	.01477	.00001		2.5	.02053	.00000
E2-7	-2.5	.01569	.00000	E2-8	-2.5	.01822	.00000
	1.0	.24031	.00003		1.0	.23822	.00001
E2-9	-2.5	.01705	.00000	E2-10	-2.5	.01759	.00000
	1.0	.24418	.00001		1.0	.24027	.00001

LITERATURE CITED

Anderson, T.W. (1958). An Introduction to Multivariate Statistical Analysis. Wiley, New York.

Bochner, S. (1955). Harmonic Analysis and the Theory of Probability. Univ. of California.

Box, G.E.P. and Muller, M.E. (1958). A note on the generation of random normal deviates. Ann. Math. Statist. 29 610-611.

Cacoullos, T. (1964). Estimation of a Multivariate Density. Univ. of Minnesota, Dept. of Statistics, Technical Report No. 40.

Cooper, P.W. (1963). Statistical classification with quadratic forms. Biometrika 50 439-448.

Cramer, H. (1946). Mathematical Methods of Statistics, Princeton University Press, Princeton.

Fix, E. and Hodges, J.L.Jr. (1951). Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties. USAF School of Avaition Medicine, Project No. 21-49-004, Report No. 4.

Fix, E. and Hodges, J.L.Jr. (1952). Discriminatory Analysis, Nonparametric Discrimination: Small Sample Performance. USAF School of Avaition Medicine, Project No. 21-49-004, Report No. 11.

Hoel, P.G. and Peterson, R.P. (1949). A solution to the problem of optimum classification. Ann. Math. Statist. 20 433-438.

Johns, M.V. (1960). An empirical bayes approach to nonparametric two-way classification. Studies in Item Analysis and Prediction, Edited by Solomon, chapter 13, Stanford.

Manija, G.M. (1961). Remarks on non-parametric estimates of a two-dimensional density function, (Russian). Soobsc. Akad. Nauk Gruzin SSR27 385-390.

Muller M.E. (1959). A comparison of methods for generating normal deviates on digital computers. J. Assoc. Comp. Mach. 6 376-383.

Parzen, E. (1962). On estimation of a probability density function and mode. Ann. Math. Statist. 33 1065-1076.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. Ann. Math. Statist. 27 832-837.

Sarhan, A.E. and Greenberg, B.G. (1962). Contributions to Order Statistics. Wiley, New York.

Stoller, D.S. (1954). Univariate two-population distribution-free discrimination. J. Amer. Statist. Assoc. 49 770-777.

Tukey, J.W. (1947). Non-parametric estimation II. Statistically equivalent blocks and tolerance regions - the continuous case. Ann. Math. Statist. 18 529-539.

Wald, A. (1943). An extension of Wilks' method for setting tolerance limits. Ann. Math. Statist. 14 45-55.

Watson, G.S. and Leadbetter, M.R. (1963). On the estimation of the probability density, I. Ann. Math. Statist. 34 480-491.

Welch, B.L. (1939). Note on discriminant functions. Biometrika 31 218-220.

Whittle, P. (1958). On the smoothing of probability density functions. J. Roy. Statist. Soc., Ser. B 20 334-343.

Wilks, S.S. (1941). Determination of sample sizes for setting tolerance limits. Ann. Math. Statist. 12 91-96.

Wilks, S.S. (1962). Mathematical Statistics. Wiley, New York.