stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE STATISTIEK SW 86/82 AUGUSTUS
(DEPARTMENT OF MATHEMATICAL STATISTICS)

A.W. AMBERGEN & W. SCHAAFSMA

THE ASYMPTOTIC VARIANCE OF ESTIMATORS FOR POSTERIOR PROBABILITIES

Preprint

kruislaan 413 1098 SJ amsterdam

Printed at the Mathematical Centre, 413 Kruislaan, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

1980 Mathematics subject classification: 62H30

The asymptotic variance of estimators for posterior probabilities [*)]

by

A.W. Ambergen & W. Schaafsma[**)]

ABSTRACT

In this paper asymptotic variances of estimators for the posterior probability that an individual belongs to one of $k \geq 2$ populations are presented. It is assumed that a set of k prior probabilities and a $p \geq 1$ dimensional vector of scores of the individual are given. In the model the populations are represented by multivariate normal distributions. The case with the assumption of homogeneity of dispersion matrices as well as the case without this assumption are treated. Further we give some crude preliminary proposals to deal with the general case '$k \geq 2$, $p \geq 1$' where no normality assumptions are made. The last part of this paper deals with a case from physical anthropology.

KEY WORDS & PHRASES: *estimating posterior probabilities, confidence intervals for posterior probabilities*

[*)] This report will be submitted for publication elsewhere.

[**)] Rijksuniversiteit te Groningen, Postbus 800, 9700 AV Groningen

0. SUMMARY

During long discussions with research workers it became clear to the authors that there are many diagnostic situations where the involvement of statistician and computer should be restricted to the computation of point and interval estimates for posterior probabilities. Specification of prior probabilities should be left to the genuine decision maker who will also have to bear full responsibility for the interpretation, in terms of future actions and decisions, of the generated confidence intervals for the posterior probabilities.

Let $x \in \mathbb{R}^p$ denote the vector of scores of the individual under investigation and let $\rho_1,\ldots,\rho_k$ denote the corresponding prior probabilities. The posterior probabilities $\rho_{1|x},\ldots,\rho_{k|x}$ have to be estimated on the basis of training samples from the k populations (or subpopulations) involved. In the first part (Section 2,3,4) we assume that x has a multivariate normal distribution for each of the k populations. In addition to this we shall sometimes postulate that homogeneity of dispersions holds. In any case it is easy to derive maximum likelihood estimators $R_{1|x},\ldots,R_{k|x}$ for the posterior probabilities (or other asymptotically equivalent and asymptotically efficient estimators). The limit distribution (as $n \to \infty$) of $n^{\frac{1}{2}}(R_{t|x}-\rho_{t|x})$ is used to characterize the accuracy of the estimator for the t-th posterior probability (n denotes the total sample size).

Section 2 is devoted to the case "$k \geq 2$, $p \geq 1$, normality and homogeneity of dispersions", Section 3 to "$k \geq 2$, $p \geq 1$, normality" and Section 4 to a comparison of the results in the sections 2 and 3.

In the second part (Sections 5 and 6) attention is paid to some crude preliminary proposals to deal with the general case "$k \geq 2$, $p \geq 1$" where no normality assumptions are made. The results of sections 3 and 5 are compared in Section 6 in order to suggest how many observations it costs to drop the assumption of normality.

The third part (Section 7) is devoted to a case from physical antropology.

1. INTRODUCTION

The intuitive background of our subject becomes clear if its historical perspective is described. Discriminant analysis originated from practical needs of research workers. This is obvious for Hotelling's generalized T^2 and very obvious for Fisher's contributions. Its history, like that of mathematical statistics at large, has been influenced by controversies concerning the probability concept. Note that in broad outline three competing definitions exist: (1) the classical definition which is very convenient if games of chance are studied, (2) the frequentistic definition where probabilities of events are mathematical idealizations of relative frequencies, (3) the subjectivistic definition where probabilities of events and also of statements are largely intended to describe personal degrees of belief. Fisher, Neyman-Pearson, Wald and many others created the classical objectivistic approach to statistics by working within the framework of definition 2. The attention was focussed on the construction of procedures and their comparison by means of performance characteristics which usually are functions of the underlying unknown parameter θ. Examples of such functions are the power function in the Neyman-Pearson theory of testing statistical hypotheses and the risk function in Wald's theory of statistical decision functions. Within this objectivistic approach one may consider "weight functions" $\rho(\theta)$ and construct "Bayes procedures" which minimize the corresponding weighted average of e.g. the risk function. However, one usually refuses to regard θ as the outcome of a random variable with known prior distribution, simply because this makes no sense if θ has the character of a universal constant whose true value is unknown.

Jeffreys, Savage and many others preferred the subjectivistic approach because there are many problem areas where the other definitions are too restrictive. One might think of situations where (almost) no data is available and where "opinions" have to be characterized, confronted and possibly reconciled. Within the subjectivistic framework there has been a tendency to rediscover the results of classical statistics by working with convenient "diffuse", "non-informative", "improper" priors. This provided new and deeper insight in the objectivistic approach (see Stein's work on the (in) admissibility of various classical procedures).

The consequences of the objectivistic approach for discriminant analysis were that the attention was focussed on the classification aspect. Fisher's discriminant functions were regarded as useful tools whose relevancy follows from their occurence in optimal classification procedures. WALD (1944) is an early attempt to deal with the obvious fact that the true densities $f_1,\dots,f_k$ for the vector of scores X in the k involved populations will be unknown in practice. On the basis of the observation vector X_0 for the individual under classification and the independent random samples $X_{h1},\dots,X_{hn_h}$ from $f_h(h=1,\dots,k)$, the individual has to be assigned to one of the k populations. Restricting the attention to $[k=2,\ f_h$ the p.d.f. of $N_p(\mu_h,\Sigma)]$, Wald, Anderson, Okamoto, Sitgreaves a.o. constructed classification procedures and studied the corresponding misclassification probabilities as function of the underlying unknown parameters (μ_1,μ_2,Σ), usually by giving asymptotic expansions. The second author has thought for a very long time that this work should be regarded as the core of discriminant analysis. He tried to contribute by deriving exact results for the univariate case (see SCHAAFSMA-VAN VARK(1977) for further references). Note that any specification of prior probabilities and prior distributions is avoided in this objectivistic approach. One works with "plug-in", "maximum likelihood", "minimax risk" and "invariance" considerations.

The consequences of the subjectivistic approach for discriminant analysis were that an elegant "fully Bayesion" approach came into being, based on diffuse improper prior distributions for the underlying unknown parameter $(\mu_1,\dots,\mu_k,\Sigma)$ (see GEISSER (1964)). PRESS (1972) seems to prefer this approach over the objectivistic one on the basis of the simplicity of the results.

Long discussions with research workers, in particular the physical-anthropologist Van Vark, have convinced the second author that he should abandon his attitude to avoid the specification of the prior probabilities $\rho_t = P(T=t)$ $(t=1,\dots,k)$ where T is the random variable labeling the population to which the individual under investigation belongs. Note that "objecvistic" results, based on "plugging in", maximum likelihood", "minimax risk" or "invariance", lose much of their appeal if it is clear that $(\rho_1,\dots,\rho_k)$ is not in the neighbourhood of $(k^{-1},\dots,k^{-1})$. In such situations one will either have to study the consequences of such specifications of $(\rho_1\dots,\rho_k)$

or one will have to adopt a Neyman-Pearson type formulation where certain error probabilities are controlled. Note that the prior probabilities $\rho_1,\ldots,\rho_k$ have a clear frequentistic interpretation if the individual under classification can be regarded as one from many potential individuals, ρ_t denoting the relative frequency of membership of population t. The basic difficulty lies in the specification of these prior probabilities. In Van Vark's problems of diagnosing the sex of prehistoric human skeletal remains nobody will object if $\rho_1 = \rho_2 = \frac{1}{2}$ is taken. In other anthropological investigations one would like to take into account where and when the investigated individual lived. This will usually not lead to a unique choice of $\rho_1,\ldots,\rho_k$. One will have to study the consequences of various different specifications in the hope that different opinions can be reconciled in the light of the data. It is interesting to remark that reasonable specifications of prior probabilities can often be obtained on the basis of past experiences if medical diagnosis problems are considered. GANESALINGAM and MCLACHLAN (1979) refer to a case study on haemophilia by Van der Broek, Habbema and Hermans where prior probabilities can be computed on the basis of family trees. This leads to different prior probabilities for different individuals. If the specification of $\rho_1,\ldots,\rho_k$ causes difficulties then *all conclusions should be formulated with respect to the introduced prior probabilities.*

Practical research workers understand very well that possible conclusions may depend on the specification of $\rho_1,\ldots,\rho_k$ and that they, and not the statistician or computer, should make this specification. They usually are very much interested in the corresponding *posterior probabilities*.

$$(1.1) \qquad \rho_{t|x} = P(T=t\,|X_0 = x) = \rho_t\, f_t(x) \,/\, \sum_{h=1}^{k} \rho_h f_h(x)$$

$(t=1,\ldots,k)$. Note that these probabilities are unknown parameters because they depend on $f_1,\ldots,f_k$. It has become common practice in certain circles to supplement the application of a crude classification procedure by computing an estimate for the posterior probability of the population to which the individual is assigned. This estimate is regarded as the probability that the individual has been classified correctly. There are some difficulties in this interpretation, especially because an estimate is taken for the true value. Other complications are that $\rho_1,\ldots,\rho_k$ need not be correct

assumptions are made. This will lead to larger asymptotic variances (see Section 4). Section 5 contains some specific proposals to deal with the case that no normality assumptions are made. This will lead to larger asymptotic m.s.e.'s than in Section 3 (see Section 6).

Physical anthropologists interested in basic concepts and applications are invited to continue with Section 7 before venturing into the technicalities of Sections 2,...,6.

2. THE CLASSICAL CASE OF NORMALITY AND HOMOGENEITY OF DISPERSION MATRICES

Regarding $x \in \mathbb{R}^p$ and $\rho_1,\dots,\rho_k$ as given prescribed constants, we are interested in estimating the posterior probabilities $\rho_{1|x},\dots,\rho_{k|x}$ on the basis of training samples, if f_h is the p.d.f. of the p-variate normal distribution $N_p(\mu_h,\Sigma)$. Hence $\theta = (\mu_1,\dots,\mu_k,\Sigma)$ plays the part of the unknown parameter in this section. Note that

$$(2.1) \qquad \rho_{t|x} = \rho_t \exp\left(-\tfrac{1}{2}\Delta^2_{x:t,t}\right) / \left\{\sum_{k=1}^{k} \rho_h \exp\left(-\tfrac{1}{2}\Delta^2_{x:h,h}\right)\right\}$$

where

$$(2.2) \qquad \Delta^2_{x:h,t} = (x-\mu_h)^T \Sigma^{-1} (x-\mu_t)$$

Let $X_{h1},\dots,X_{hn_h}$ denote the h-th training sample which means that these random vectors are independent, and with $N_p(\mu_h,\Sigma)$ distributions. Let $n = n_1 + \dots + n_k$ denote the total sample size. Note that $\rho_{t|x}$ is a function of $\theta = (\mu_1,\dots,\mu_k,\Sigma)$ which can be estimated, e.g. by means of the maximum likelihood method. (Various other methods were considered in AMBERGEN-SCHAAFSMA (1981); these methods will not be repeated here because they did not lead to a significant improvement over the maximum likelihood method.) Using the notations

$$X_{h.} = n_h^{-1} \sum_{i=1}^{n_h} X_{hi} \quad \text{and} \quad S = \sum_{h=1}^{k} \sum_{i=1}^{n_k} (X_{hi} - X_{h.})(X_{hi} - X_{h.})^T$$

for the h-th sample mean and the pooled matrix of cross-products, we see that

$$\hat{\theta} = (X_{1.},\ldots,X_{k.},n^{-1}S)$$

is the maximum likelihood estimator for θ, while

$$R_{t|x} = \hat{\rho}_{t|x} = \rho_t \exp(-\tfrac{1}{2}\hat{\Delta}^2_{x;t,t})/\{\sum_{h=1}^{k} \rho_h \exp(-\tfrac{1}{2}\hat{\Delta}^2_{x;h,h})\}$$

is the maximum likelihood estimator for $\rho_{t|x}$, where

$$\hat{\Delta}^2_{x;h,t} = n(x-X_{h.})^T S^{-1}(x-X_{t.})$$

The exact distribution of the estimator $R_{t|x}$ around the true value $\rho_{t|x}$ can be studied by means of simulation experiments. Approximations based on the following limit theorem are reliable if the sample sizes are not smaller than, say 30 and if they are based on the idea that $R_{t|x}$ is approximately normally distributed with expectation $\rho_{t|x}$ and variance $n^{-1}(\psi\,\Gamma\,\psi)_{tt}$. The reliability of the approximations follows from Ambergen's simulation experiments, see AMBERGEN-SCHAAFSMA (1982).

THEOREM 2.1. *If* $n \to \infty$ *and* $n_h n^{-1} \to b_h > 0$ $(h=1,\ldots,k)$, *then*

$$L\, n^{\frac{1}{2}}(R_{.|x} - \rho_{.|x}) \to N_k(0,\psi\,\Gamma\,\psi) \tag{2.3}$$

where $R_{.|x} = (R_{1|x},\ldots,R_{k|x})^T$, $\rho_{.|x} = (\rho_{1|x},\ldots,\rho_{k|x})^T$ *and* Γ *is determined by*

$$\begin{aligned} \Gamma_{h,h} &= 4\, b_h^{-1}\, \Delta^2_{x;h,h} + 2\,\Delta^4_{x;h,h} \\ \Gamma_{h,t} &= 2\,\Delta^4_{x;h,t} \qquad (h\neq t) \end{aligned} \tag{2.4}$$

while ψ *is determined by*

$$\begin{aligned} \psi_{h,h} &= \tfrac{1}{2}\,\rho_{h|x}(1-\rho_{h|x}) \\ \psi_{h,t} &= -\tfrac{1}{2}\,\rho_{h|x}\,\rho_{t|x} \qquad (h\neq t) \end{aligned} \tag{2.5}$$

PROOF. See AMBERGEN-SCHAAFSMA (1982) which generalized previous results in SCHAAFSMA-VAN VARK (1977) for "$k = 2$, $p = 1$, $\sigma_1^2 = \sigma_2^2$" and SCHAAFSMA-VAN VARK (1979) for "$k = 2$, $p \geq 1$, $\Sigma_1 = \Sigma_2$" to the case $k > 2$.

3. NORMALITY BUT NOT NECESSARILY HOMOGENEITY OF DISPERSION MATRICES

Regarding $x \in \mathbb{R}^p$ and $\rho_1,\ldots,\rho_k$ as given prescribed constants, we are now interested in estimating $\rho_{1|x},\ldots,\rho_{k|x}$ if f_h is the p.d.f. of $N_p(\mu_h,\Sigma_h)$. Hence $\theta = (\mu_1,\ldots,\mu_k,\Sigma_1,\ldots,\Sigma_k)$ will now play the part of the unknown parameter and $\rho_{t|x}$ is determined by (1.1) where

$$(3.1) \qquad f_h(x) = |2\pi\,\Sigma_h|^{-\frac{1}{2}} \exp(-\tfrac{1}{2}\,\Delta^2_{x;h})$$

with

$$(3.2) \qquad \Delta^2_{x;h} = (x-\mu_h)^T\,\Sigma_h^{-1}\,(x-\mu_h)$$

Again using the maximum likelihood method, we obtain

$$\hat{\theta} = (X_{1.},\ldots,X_{k.},n_1^{-1}S_1,\ldots,n_k^{-1}S_k)$$

where

$$X_{h.} = n_h^{-1}\sum_{i=1}^{n_h} X_{hi} \text{ and } S_h = \sum_{i=1}^{n_h}(X_{hi}-X_{h.})(X_{hi}-X_{h.})^T$$

Hence the maximum likelihood estimators

$$\hat{\Delta}^2_{x;h} = n_h(x-X_{h.})\,S_h^{-1}(x-X_{h.})$$

$$\hat{f}_h(x) = |2\,\pi\,n_h^{-1}\,S_h|^{-\frac{1}{2}} \exp(-\tfrac{1}{2}\,\hat{\Delta}^2_{x;h})$$

and

$$(3.3) \qquad R_{t|x} = \hat{\rho}_{t|x} = \rho_t\,\hat{f}_t(x)\,/\,\{\sum_{h=1}^{k}\rho_h\hat{f}_h(x)\}$$

are immediately obtained.

LEMMA 3.1. *If* $n_h \to \infty$, *then*

$$(3.4) \qquad L \; n_h^{\frac{1}{2}} (\hat{f}_h(x) - f_h(x)) \to N(0, \tfrac{1}{2}(\Delta^4_{x;h} + p) f_h^2(x))$$

PROOF. See AMBERGEN-SCHAAFSMA (1982).

THEOREM 3.2. *If* $n \to \infty$ *and* $n_h n^{-1} \to b_h > 0 (h=1,\dots,k)$, *then*

$$(3.5) \qquad L \; n^{\frac{1}{2}} (R_{.|x} - \rho_{.|x}) \to N_k(0, \Psi\,\Theta\,\Psi)$$

where $R_{.|x}, \rho_{.|x}$ 2

$$(3.6) \qquad \Theta_{h,h} = 2\, b_h^{-1} (\Delta^4_{x;h} + p)$$

$$\Theta_{h,t} = 0 \qquad\qquad (h \neq t)$$

PROOF. This is an immediate consequence of the lemma and (3.3). It is of some interest to remark that we had expected that the case of this section would be more difficult than that of Section 2. In fact it is easier because of the independence of the density estimators $\hat{f}_1, \dots, \hat{f}_k$.

4. WHAT DOES IT COST TO DROP THE HOMOGENEITY ASSUMPTION?

It follows from general arguments that the estimators of Section 2 are better than those of Section 3 if the homogeneity assumptions of Section 2 are satisfied. Hence $\Sigma_1 = \dots = \Sigma_k = \Sigma$ implies that $\Psi(\Theta - \Gamma)\Psi$ is nonnegative definite. Mathematical verification of this result would provide a check on the validity of (2.3) and (3.5). This verification is not very easy because $\Theta - \Gamma$ is not always nonnegative definite. Up to now we succeeded only in verifying $\Psi(\Theta - \Gamma)\,\Psi \geq 0$ for $k = 2$ (Case 1 provides an example where equality holds, this is an indication of the complications which appear if one tries to give a general proof).

Practical research workers will wonder how large the approximate standard deviations $n^{-\frac{1}{2}}\{(\Psi\,\Gamma\,\Psi)_{t,t}\}^{\frac{1}{2}}$ and $n^{-\frac{1}{2}}\{(\Psi\,\Theta\,\Psi)_{t,t}\}^{\frac{1}{2}}$ of $R_{t|x}$ will be. Is the second approximate standard deviation much larger than the first one?

How many observations does "not knowing $\Sigma_1 = \dots = \Sigma_k$" cost? What happens if the sample sizes are modified? What if the dimensionality p is increased, x is changed, or $\rho_1, \dots \rho_k$ are modified? In practice one will not know what the true values of the underlying unknown parameters are. One will have to estimate the approximate standard deviations. However, if one needs an intuitive feeling for the magnitude of the standard deviations, then it suffices to elaborate on a number of theoretical cases. The following cases were selected in order to suggest possible answers to the above-mentioned questions.

Case 1. Suppose $p = 2$, $k = 2$, $n_1 = n_2 = \frac{1}{2} n$, $\mu_1 = (1,0)^T$, $\mu_2 = (-1,0)^T$, $\Sigma = I_2$, $x = (0,1)^T$ (the reader should draw a picture and notice that $x - \mu_1$ is perpendicular to $x - \mu_2$), $\rho_1 = \rho_2 = \frac{1}{2}$. Note that the consequences of drawing training samples are studied without actually drawing them. The above-mentioned specifications imply $\rho_{1|x} = \rho_{2|x} = \frac{1}{2}$, $b_1 = b_2 = \frac{1}{2}$, $\Delta^2_{x;h} = \Delta^2_{x;h,h} = 2$, $\Delta^2_{x;h,t} = 0$ $(h \neq t)$, $\Gamma_{hh} = 4 \times 2 \times 2 + 2 \times 2^2 = 24 = 2 \times 2 \times (2^2 + 2) = \Theta_{h,h}$, $\Gamma_{h,t} = \Theta_{h,t} = 0 (h \neq t)$. Hence $\Gamma = \Theta$ and for this very special situation both estimators are asymptotically equivalent. It follows from

$$\Psi \, \Gamma \, \Psi = \Psi \, \Theta \, \Psi = 64^{-1} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 24 & 0 \\ 0 & 24 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

that the approximate standard deviations of all $R_{t|x}$'s are $n^{-\frac{1}{2}}(64^{-1}48)^{\frac{1}{2}} = .87 \, N^{-\frac{1}{2}}$. If n_1 is $n_2 = 25$, then we obtain $.87(50)^{-\frac{1}{2}} = .12$ for the approximate standard deviation. This considerable in comparison with the estimated true values $\rho_{1|x} = \rho_{2|x} = .50$.

Case 2. Suppose $p = 2$, $k = 4$, $n_1 = n_2 = n_3 = n_4 = 4^{-1} n$, $\mu_1 = (1,0)^T$, $\mu_2 = (0,1)^T$, $\mu_3 = (-1,0)^T$, $\mu_4 = (0,-1)^T$, $\Sigma = I_2$ and $x = (0,0)^T$ (the reader should draw a picture and notice that everything has been arranged nicely around the origin to facilitate computations). With $\rho_1 = \dots = \rho_k = .25$ we obtain $\rho_{t|x} = .25$ for the posterior probabilities $(t=1,\dots,4)$ because $\Delta^2_{x;h,h} = \Delta^2_{x;2,3} = 1$. Note that $\Delta^2_{x;1,3} = \Delta^2_{x;2,4} = -1$ while $\Delta^2_{x;1,2} = \Delta^2_{x;1,4} = \Delta^2_{x;2,3} = \Delta^2_{x;3,4} = 0$. Hence

$$\Psi = 32^{-1}\begin{bmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & -3 \end{bmatrix}, \Gamma = \begin{bmatrix} 18 & 0 & 2 & 0 \\ 0 & 18 & 0 & 2 \\ 2 & 0 & 18 & 0 \\ 0 & 2 & 0 & 18 \end{bmatrix}, \Theta = \begin{bmatrix} 24 & 0 & 0 & 0 \\ 0 & 24 & 0 & 0 \\ 0 & 0 & 24 & 0 \\ 0 & 0 & 0 & 24 \end{bmatrix}$$

Notice that $\Theta - \Gamma \geq 0$ and hence $\Psi\ (\Theta - \Gamma)\ \Psi \geq 0$ in this case. If $\Psi\,\Gamma\,\Psi$ and $\Psi\,\Theta\,\Psi$ are computed, then the approximate standard deviations

$$n^{-\frac{1}{2}}\{(\Psi\Gamma\Psi)_{t,t}\}^{\frac{1}{2}} = .45\,n^{-\frac{1}{2}}\ , n^{-\frac{1}{2}}\{\ \Psi\ \Theta\ \Psi\ )_{t,t}\}^{\frac{1}{2}} = .53\,n^{-\frac{1}{2}}$$

are obtained for the estimators $R_{t|x}$. Note that the second standard deviation is a bit less than 1.2 times as large as the first one: it costs about 40% of the observations if the homogeneity assumption $\Sigma_1 = \ldots = \Sigma_4$ is removed (the sample sizes should be about 1.4 times as large if the same standard deviation is required). If $n_1 = \ldots = n_4 = 25$, then $n = 100$ and the respective approximate standard deviations .045 and .053 are considerable in comparison with the true values .25 of the posterior probabilities (a comparison with Case 1 requires that the same observation vector $x = (0,0)^T$ is considered).

Case 3. We modify Case 2 by taking $n_1 : n_2 : n_3 : n_4 = 1 : 2 : 3 : 4$ ($b_1 = .10$, $b_2 = .20$, $b_3 = .30$, $b_4 = .40$) and we leave the other specifications unchanged. Hence $p = 2$, $k = 4$, $\mu_1 = (1,0)^T$, $\mu_2 = (0,1)^T$, $\mu_3 = (-1,0)^T$, $\mu_4 = (0,-1)^T$, $\Sigma = I_2$, $x = (0,0)^T$, $\rho_1 = \ldots = \rho_4 = .25$, Ψ as in Case 2. For Γ and Θ we obtain

$$\Gamma = \begin{bmatrix} 24 & 0 & 2 & 0 \\ 0 & 22 & 0 & 2 \\ 2 & 0 & 15.33 & 0 \\ 0 & 2 & 0 & 12 \end{bmatrix}, \quad \Theta = \begin{bmatrix} 60 & 0 & 0 & 0 \\ 0 & 30 & 0 & 0 \\ 0 & 0 & 20 & 0 \\ 0 & 0 & 0 & 15 \end{bmatrix}$$

Notice that $\Theta - \Gamma \geq 0$ and hence $\Psi(\Theta - \Gamma)\ \Psi \geq 0$ in this case. Elaborating on $\Psi\,\Gamma\,\Psi$ we obtained the approximate standard deviations

$$n^{-\frac{1}{2}}\{\Psi\Gamma\Psi)_{tt}\}^{\frac{1}{2}} = .64n^{-\frac{1}{2}},\quad .51\,n^{-\frac{1}{2}},\ .45\,n^{-\frac{1}{2}}\ ,\ .42\,n^{-\frac{1}{2}}$$

for $t = 1, 2, 3, 4$ respectively. The average of these values is larger than the value $.45\, n^{-\frac{1}{2}}$ in Case 2. The smallest value $.42\, n^{-\frac{1}{2}}$ is smaller than in Case 2 and this is caused by the large value of b_4. Elaborating on $\Psi\ \Theta\ \Psi$ we obtained approximate standard deviations for the $R_{t|x}$'s which are again about 1.2 times as large as for the estimators based on the homogeneity assumptions.

Case 4. We modify Case 2 by increasing the dimensionality p. It is interesting from a theoretical point of view to do this by introducing variables which contain no discrimination information. In that case one will find larger approximate standard deviations and this may lead to new insight in the problem of selecting variables in discriminant analysis.

We take $p = 4$, $k = 4$, $n_1 = n_2 = n_3 = n_4 = 4^{-1}n, \mu_1 = (1,0,0,0)^T$, $\mu_2 = (0,1,0,0)^T$, $\mu_3 = (-1,0,0,0)^T$, $\mu_4 = (0,-1,0,0)^T$, $\Sigma = I_4, x = (0,0,x_3,x_4)^T$ and $\rho_1 = \ldots = \rho_4 = .25$. Note that any specification of x_3, x_4 leads to $\Delta^2_{x;1} = \ldots = \Delta^2_{x;4}$ and $\rho_{1|x} = \ldots = \rho_{4|x} = .25$. If we take $x_3 = x_4 = 0$ then Ψ, Γ and Θ are as in Case 2. If we take $x_3 = x_4 = 1$, then the matrix of inner-products $\Delta^2_{x;h,t} = (x-\mu_h)^T\ (x-\mu_t)$ $(h,t = 1,\ldots,4)$ becomes

$$\begin{bmatrix} -1 & 0 & 1 & 1 \\ 0 & -1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 2 & 1 & 2 \\ 2 & 3 & 2 & 1 \\ 1 & 2 & 3 & 2 \\ 2 & 1 & 2 & 3 \end{bmatrix}$$

Hence Ψ is still as in Case 2 but

$$\Gamma = \begin{bmatrix} 66 & 8 & 2 & 8 \\ 8 & 66 & 8 & 2 \\ 2 & 8 & 66 & 8 \\ 8 & 2 & 8 & 66 \end{bmatrix} ; \ \Theta = \begin{bmatrix} 104 & 0 & 0 & 0 \\ 0 & 104 & 0 & 0 \\ 0 & 0 & 104 & 0 \\ 0 & 0 & 0 & 104 \end{bmatrix}$$

Notice that again there is no question about $\Theta - \Gamma \geq 0$. Elaborating on $\Psi\ \Gamma\ \Psi$ we obtained the approximate standard deviations $.84\, n^{-\frac{1}{2}}$ for the estimators $R_{t|x}$ of Section 2. Computing the diagonal elements of $\Psi\ \Theta\ \Psi$ we obtain the approximate standard deviations $1.10\, n^{-\frac{1}{2}}$ for the estimators $R_{t|x}$ of Section 3. For $n_1 = n_2 = n_3 = n_4 = 25$ we now obtain the respective approximate standard deviations .084 and .110 for the estimators $R_{t|x}$. The values are very large with respect to the estimated true values $\rho_{t|x} = .25$. Not knowing that

$\Sigma_1 = \ldots = \Sigma_4$ now costs about $(110/84)^2 - 1 = 72\%$ of the observations. This is not unexpected because increasing the dimensionality without supplying relevant extra information causes extra confusion, especially if $\Sigma_1 = \ldots \Sigma_4$ cannot be postulated.

Case 5. We modify Case 2 by breaking the symmetry such that the posterior probabilities differ from the prior ones. The specifications $p = 2$, $k = 4$, $n_1 = n_2 = n_3 = n_4 = 4^{-1}n, \mu_1 = (1,0)^T$, $\mu_2 = (0,1)^T$, $\mu_3 = (-1,0)^T$ $\mu_4 = (0,-1)^T$ and $\Sigma = I_2$ are left unchanged but $x = (1,0)^T$ is taken such that the first posterior probability $\rho_{1|x_2}$ (again $\rho_1 = \ldots = \rho_4 = .25$) is close to .50. The matrix of inner-products $\Delta^2_{x;h,t} = (x-\mu_h)^T (x-\mu_t)$ $(h,t=1,\ldots,4)$ becomes

$$\begin{bmatrix} 0 & 0 \\ 1 & -1 \\ 2 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 2 & 1 \\ 0 & -1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 2 & 2 & 0 \\ 0 & 2 & 4 & 2 \\ 0 & 0 & 2 & 2 \end{bmatrix}$$

with the consequence that

$$\rho_{1|x} : \rho_{2|x} : \rho_{3|x} : \rho_{4|x} = e^0 : e^{-1} : e^{-2} : e^{-1}$$

and hence

$$\rho_{1|x} = .53 \; ; \; \rho_{2|x} = .20 \; ; \; \rho_{3|x} = .07 \; ; \; \rho_{4|x} = .20$$

The matrices Ψ, Γ and Θ become

$$\Psi = \tfrac{1}{2} \begin{bmatrix} .25 & -.11 & -.04 & -.11 \\ -.11 & .16 & -.01 & -.04 \\ -.04 & -.01 & .07 & -.01 \\ -.11 & -.04 & -.01 & .16 \end{bmatrix}, \Gamma = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 40 & 8 & 0 \\ 0 & 8 & 96 & 8 \\ 0 & 0 & 8 & 40 \end{bmatrix}, \Theta = \begin{bmatrix} 16 & 0 & 0 & 0 \\ 0 & 48 & 0 & 0 \\ 0 & 0 & 144 & 0 \\ 0 & 0 & 0 & 48 \end{bmatrix}$$

Notice that again there is no question about $\Theta - \Gamma \geq 0$. If we compute $\Psi \Gamma \Psi$ then we obtain the respective approximate standard deviations $.56\ n^{-\frac{1}{2}}$, $.51\ n^{-\frac{1}{2}}$, $.34\ n^{-\frac{1}{2}}$ and $.51\ n^{-\frac{1}{2}}$ for the estimators $R_{t|x}$ $(t = 1,\ldots,4)$ of Section 2 and based on the assumption of homogeneity of dispersion matrices. Notice that the largest posterior probability $\rho_{1|x} = .53$ has an estimator $R_{1|x}$ with approximate standard deviation .06 if $n_1 = n_2 = n_3 = n_4 = 25$, this accuracy is quite satisfactory in our opinion. If we compute $\Psi \ominus \Psi$, then we obtain the approximate standard deviations $.77\ n^{-\frac{1}{2}}$, $.62\ n^{-\frac{1}{2}}$, $.43\ n^{-\frac{1}{2}}$, $.62\ n^{-\frac{1}{2}}$.

REMARK. It would be interesting to combine various modifications of Case 2 and to see what happens if the prior probabilities are changed. Modifications of the latter kind have a considerable effect on the parameters to be estimated: the "bias" which appears if wrong prior probabilities are used may be much larger than the approximate standard deviations of the estimators $R_{t|x}$ for $\rho_{t|x}$. Conclusions should always be formulated with respect to the introduced prior probabilities.

5. DROPPING THE ASSUMPTION OF NORMALITY

Regarding $x \in \mathbb{R}^p$ and $\rho_1, \ldots, \rho_k$ as given prescribed constants we shall now try to estimate the posterior probabilities (1.1) if $f_1, \ldots, f_k$ are arbitrary continuous densities or more generally if $f_1, \ldots, f_k$ are nice Radon-Nikodym derivatives with respect to some appropriate σ-finite measure λ on $\mathbb{R}^p$. Such generalisations are needed if the measurement vector has discrete components.

Two natural approaches present themselves: (1) adapt the theory of non-parametric density estimation, (2) see what happens if the theory of Section 3 is applied.

With respect to Approach (1).

Since ROSENBLATT (1956) started the subject, many contributions were made. A recent review is BEAN-TSOKOS (1980). It is emphasized that the posterior probabilities (1.1) depend only upon the values of the densities in the particular point x. This implies that results based on the integrated mean square error as performance characteristic are not useful. Another complication is that the mean square error loses much of its appeal if one wants to construct confidence intervals instead of estimators. This makes clear why we have to adapt the existing theory to our particular needs. We restrict the attention to Rosenblatt's original idea to use a window estimator (Parzen's kernel estimators are ignored because we want to make one simple definite proposal to construct confidence intervals for posterior probabilities). Let U denote some neighbourhood of x. The value $f_h(x)$ of the density in x is approximately equal to p_h/λ (U) where $\lambda(U)$ is the measure of U, and $p_h = \int_U f_h \, d\lambda = P(X_{hi} \in U)$. Hence the posterior probability $\rho_{t|x}$ is approximately equal to

$$(5.2) \qquad \bar{\rho}_{t|x} = \rho_t \, p_t \, / \sum_{h=1}^{k} \rho_h \, p_h$$

The natural estimators for p_h and $\bar{\rho}_{t|x}$ are

$$(5.3) \qquad \hat{p}_h = \# \{i \, ; X_{hi} \in U\} \, / \, n_h \; ; \; \bar{R}_{t|x} = \rho_t \, \hat{p}_t \, / \sum_{h=1}^{k} \rho_h \, \hat{p}_h$$

and the central limit theorem for relative frequencies implies that the following results hold if U is a *fixed* neighbourhood of x.

LEMMA 5.1. *If* $n_h \to \infty$, *then*

$$(5.4) \qquad L \, n_h^{\frac{1}{2}} \, (\hat{p}_h - p_h) \to N(0, p_h(1-p_h))$$

THEOREM 5.2. *If* $n \to \infty$ *and* $n_h \, n^{-1} \to b_h > 0$ $(h=1,\ldots,k)$, then

$$(5.5) \qquad L \, n^{\frac{1}{2}} \, (\bar{R}_{./x} - \bar{\rho}_{./x}) \to N_k \, (0, \Psi\Delta\Psi)$$

where Ψ is determined by (2.5) and Δ by

$$(5.6) \qquad \Delta_{h,h} = 4 b_h^{-1} \, p_h^{-1} (1-p_h) \quad ; \quad \Delta_{h,t} = 0 \quad (h \neq t)$$

CONCLUSION. We propose to use

$$(5.7) \qquad \bar{R}_{t|x} \pm 1.96 \quad n^{-\frac{1}{2}} \, \{(\hat{\Psi}\hat{\Delta}\hat{\Psi})_{t,t}\}^{\frac{1}{2}}$$

as a 95% confidence interval, both for $\bar{\rho}_{t|x}$ and for $\rho_{t|x}$. $\hat{\Psi}$ and $\hat{\Delta}$ are obtained by plugging in estimates for the unknown parameters.

The basic problem is how to choose the neighbourhood U. Two conflicting aspects can be distinguished: (1) U should be "small"because this guarantees that $p_n/\lambda(U)$ is close to $f_h(x)$ such that one may hope that $\bar{\rho}_{t|x}$ is close to the true posterior probability $\rho_{t|x}$ (notice that difficulties appear if all p_h's are small), (2) U should not be "small" because the standard deviations of the estimators get large if the probabilities (5.1) are small. Notice that $\Delta_{h,h}$ in (5.6) explodes if $p_h \to 0$. Usually an appropriate choice of U cannot be made without looking at the data and the sample sizes. This data dependence invalidates not only the proofs of Lemma (5.1) and Theorem (5.2) but even their formulation, because the normal distributions in the right hand sides of (5.4) and (5.5) will depend on the sample sizes if U is shrinked as $n \to \infty$, which is quite natural. Hence some mathematical plasterwork is needed to prove that

$$(5.8) \qquad \lim_{n\to\infty} P(\rho_{t|x} \in \text{interval (5.7)}) = .95$$

holds for the sequence of data-dependent neighbourhoods U which will be proposed. Mathematical theory for choosing U is relegated to the appendix because this theory is not completely relevant since it focusses the attention on constructing confidence intervals for a density instead of for the posterior probabilities. However, the following procedure is recommended.

Step 1. We mentioned that two conflicting aspects are involved in the choice of U. It is intuitively clear that reconciliation in the light of the data appears if the sample sizes are sufficiently large. However, if p is not small and $n = n_1 + \ldots + n_k$ is not large then it will not be possible to keep bias (aspect 1) and variance (aspect 2) within reasonable bounds. The remark following (8.15) suggests that n is certainly sufficiently large if the strong requirement (8.16) is satisfied and that lousy results have to be expected if even the weak requirement (8.17) is not satisfied. In the latter case one should try to reduce the dimension p or increase the sample sizes.

Step 2. Compute the outcomes of

$$(5.9) \qquad X_{..} = n^{-1} \sum \sum X_{hi} \quad ; \quad M = (n-1)^{-1} \sum \sum (X_{hi}-X_{..})(X_{hi}-X_{..})^T$$

in order to characterize the general situation of the data. Using the total sample covariance matrix M, the space $\mathbb{R}^p$ is endowed with the inner-product

$(y,z)_m = y^t M^{-1} z$, norm $\| y \|_M = \{(y,y)_M\}^{\frac{1}{2}}$ and metric $d_M(y,z) = \| y - z \|_M$.

Step 3. Compute the Mahalanobis distances

(5.10) $$d = d_M\,(x,X_{..}) = [(x-X_{..})^T M^{-1} (x-X_{..})]^{1/2}$$

and

(5.11) $$d_m(x,X_{hi}) = [(x-X_{hi})^T M^{-1} (x-X_{hi})]^{1/2}$$

$(i=1,\dots,n_h\ ; h=1,\dots,k)$.

Step 4. Compute

(5.12) $$r(p,0,n) = [2^{\frac{1}{2}p+1}(p+2)^2\ \Gamma(\tfrac{1}{2}p)/\{p\, n\,(\ell n\, n)^2\}]^{1/(p+4)}$$

and $r = r(p,d,n)$ by requiring

(5.13) $$P(\chi^2_{p;d^2} \leqq r^2) = P(\chi^2_p \leqq r^2(p,0,n))$$

where χ^2_p $(\chi^2_{p;d^2})$ has (non-central) chi-square distribution with p degrees of freedom (and non-centrality parameter d^2). Next the neighbourhood of x is defined by

(5.14) $$U = \{y \in \mathbb{R}^p\ ;\ d_M\,(y,x) \leqq r(p,d,n)\}$$

The radius $r(p,0,n)$ of ball U in case $x = X_{..}$ has been determined such that the bias of a density estimator satisfies certain requirements. The extension (5.13) is such that $r(p,d,n)$ is an increasing function of $d = d_M(x,X_{..})$. The rationale behind (5.13) is that we want to avoid the situation that all p_h's are small because then (5.2) is very unstable.

Step 5. Continue along the lines of (5.3),...,(5.7) and note that (5.11) is exploited in

$$\#\,\{i\ ;\ X_{hi} \in U\} = \#\,\{i\ ; d_M(x,X_{hi}) \leqq r(p,d,n)\}$$

With respect to Approach (2).

If one applies Section 3 to situations with non-normal densities, then a non-vanishing bias will appear: the procedures of Section 3 are far from "robust". Note that a similar behaviour would be displayed by the confidence interval (5.7) if U is kept fixed. By shrinking U as $n \to \infty$ we created the theoretic possibility to satisfy (5.8). This possibility does not exist if Section 3 is to be applied.

In practical situations the sample sizes are essentially fixed and one should not jump to the conclusion that (5.7) based on (5.14) is to be preferred to the interval based on Section 3, if aberrations from normality are evident. In fact it is an urgent and very difficult problem to draw reasonable dividing-lines between (1) the applications where Section 2 is preferable, (2) those where Section 3 has to be recommended and (3) those where the nonparametric approach is to be chosen. It is obvious that the situation of these dividing-lines should depend on the sample sizes. If these are small then one will not worry much about the bias because the estimated standard deviations will be large. If the sample sizes are very small then one will prefer the approach of Section 2. If sample sizes are increased then preference shifts towards Section 3 and the nonparametric approach. Hence asymptotic arguments cannot be conclusive. In fact the subject of this whole paper becomes irrelevant if sample sizes get very large: the genuine decision maker will not worry about estimated standard deviations and a possible bias if these are no more than a few per cent. His main concern will then be regarding the uncertainties in his prior probabilities $\rho_1,\ldots,\rho_k$ and, more fundamentally, the specification of his decision situation: is it allowed to remain undecided and what is an appropriate loss-function?

Focussing on the reliability of the procedures of Section 3, the following asymptotic considerations are useful. Assuming that $\mu_h = E\,X_{hi}$ and $\Sigma_h = E(X_{hi}-\mu_h)(X_{hi}-\mu_h)^T$ exist, it follows from the law of large numbers that $X_{h.} \to \mu_h$,

$$n^{-1}S_h \to \Sigma_h\ ,\ \hat{\Delta}^2_{x;h} \to \Delta^2_{x,h}\ ,\ \hat{f}_h(x) \to \tilde{f}_h(x) \quad \text{and}$$

$$(5.15) \qquad R_{t|x} \to \rho_t\,\tilde{f}_t(x) \Big/ \sum_{h=1}^{k} \rho_h\,\tilde{f}_h(x) \;=\; \tilde{\rho}_{t|x} \quad \text{if } n \to \infty.$$

Here $\hat{\Delta}^2_{x;h}$, $\hat{f}_h(x)$ and $R_{t|x}$ are the estimators defined in Section 3 while $\Delta^2_{x;h}$ is defined in (3.2) and $\tilde{f}_h(x)$ denotes the right-hand side of (3.1). Note that $\tilde{f}_h$ is some normal approximation for the true density f_h. Normality of $f_1,\ldots,f_k$ implies that $\tilde{\rho}_{t|x}$ coincides with the true posterior probability (1.1). Deviations from normality will usually lead to $\tilde{\rho}_{t|x} \neq \rho_{t|x}$. In such cases, (5.15) shows that the mid-point $R_{t|x}$ of the confidence interval of Section 3 converges towards the wrong value $\tilde{\rho}_{t|x}$. The length tends to 0 and the true value $\rho_{t|x}$ is contained in its confidence interval with probability tending to 0. This is the price to be paid if one works with a wrong model. It has been made clear before that one should be willing to pay this price if the asymptotic bias $\tilde{\rho}_{t|x} - \rho_{t|x}$ is only a few per cent, and also if it is larger provided that it is small in comparision with the standard deviation of the corresponding estimator.

6. WHAT DOES IT COST TO DROP THE NORMALITY ASSUMPTION?

Comparing the estimators of Section 3 with those of Section 5 under the assumption of normality, one will expect that those of Section 3 will be best "on the average". This is not always true because if x is such that prior and posterior probabilities coincide then the m.s.e. of the estimator of Section 5 will tend to 0 if the radius r tends to infinity (we ignore the restrictions which were recommended in Section 5 because we only want to suggest that there are no general arguments why the estimators of Section 3 will always be better than those of Section 5).

In Section 4 we focussed on the question how many observations it costs if the assumption $\Sigma_1 = \ldots = \Sigma_k$ is not used whereas it holds. Now we make no assumptions of this kind but we focus on the question how many observations it costs if the assumption of normality is not used whereas it holds. The answer depends on the proposal (5.11) which requires that for each value of n new computations are needed to determine the radius r(p,d,n). We restrict the attention to "Case 1 with $n_1 = n_2 = 50$" and "Case 2 with $n_1 = \ldots = n_4 = 25$", the cases being specified in Section 4.

Case 1 with n = 100.

The combined sample will look like a sample from the mixture

$\frac{1}{2}N_2(\mu_1,I_2) + \frac{1}{2}N_2(\mu_2,I_2)$ which has expectation 0 and covariance matrix $\begin{pmatrix}2 & 0\\ 0 & 1\end{pmatrix}$ with the consequence that in (5.9) $X_{..} \approx \begin{pmatrix}0\\0\end{pmatrix}$ and $M \approx \begin{pmatrix}2 & 0\\ 0 & 1\end{pmatrix}$ so that $d_M^2(x,X_{..}) \approx 1$. Specification of $p = 2$ and $n = 100$ gives $r(2,0,100) = .50$ so that $r(2,1,100) = .63$ and $p_h = P(d_m(X_{hi},x) \leq r(p,d,n)) = .10$. Substituting $p_h = .10$ into (5.6) gives $\Delta_{h,h} = 88.9$. This value should be compared with $\Theta_{h,h} = 24$ obtained in Section 4. The approximate standard deviation in the non-parametric case will be 1.92 times as large as when normality is assumed and we get that the standard deviation is $1.674\ n^{-\frac{1}{2}}$. With $n = 100$ this gives .1674 which value should be compared with the .087 of Section 4. Note that one needs 3.7 times as many observations in the non-parametric approach to get the same accuracy as in the case with the assumption of normality but without homogeneity of covariances. The above-mentioned standard deviation of .1674 means that one may find point estimates like .80 for a posterior probability of $\rho_{1|x} = .50$. It is obviously of the utmost importance that such inaccurate estimates are equipped with a standard deviation. The genuine decision maker's action may be different if he hears that $\rho_{1|x} = .80$ or if he hears that $\rho_{1|x} = .80 \pm .17$ or even $\rho_{1|x} \in [.47,1.0]$ by the use of an approximate 95% confidence interval.

Case 2 with n = 100.

In this case the combined sample looks like a sample from $\frac{1}{4}\sum_{h=1}^{4} N_2(\mu_h,I_2)$ which has expectation $\begin{pmatrix}0\\0\end{pmatrix}$ and covariance matrix $(3/2)I_2$. Now we have $x = 0$ so that $d = 0$ and $r(2,0,100) = .497$. Next we have to determine $p_h = P(d_M(X_{hi},x) \leq .497) = P(2/3\ \chi'^2_{2;1} \leqq .497^2) = .11$. Substituting into (5.7) gives $\Delta_{h,h} = 129$. This value should be compared with $\Theta_{h,h} = 24$ in Section 4 and it results in an approximate standard deviation of $1.23\ n^{-\frac{1}{2}}$ which is 2.32 times as large as the $.53\,n^{-\frac{1}{2}}$ in the case of normality without the homogeneity of covariances. For $n = 100$ these standard deviations are .123 resp. .053. So we see that we may not expect that the estimates of the posterior probabilities would be very reliable.

7. APPLICATIONS TO PHYSICAL ANTHROPOLOGY

We like to continue the non-technical discussions of Section 1 because such discussions can be very informative, they at least were very revealing to us.

The idea that posterior probabilities should be equipped with standard deviations appeared during discussions of the second author with Van Vark who is involved in the physical evolution of our own kind. This subject like physical anthropology at large, goes beyond the group membership discussions where posterior probabilities may be useful. The most important applications of our theory will be sought in the area of medical diagnoses where such discussions are central. Nevertheless group membership is important in physical anthropology and we shall suggest the implications of our theory, or rather those of the way of thinking of classical statisticians, for actual anthropological practice by elaborating on the following case which was suggested to us by Van Vark.

It was in 1940 that W.E. Horton, while digging for guano at Border Cave (near the boundary between Swaziland and Zululand, South-Africa) found fragments of a human cranium. More of the adult cranium was found in his dump during 1941-42. It was supposed on the basis of circumstantial evidence, e.g. artifacts, that the cranium has belonged to a Middle Stone Age hominid. The fragments of the cranium have been set in a plaster reconstruction by A.R. Hughes. Rightmire took $p = 11$ measurements on the original fossil and employed multiple discriminant analysis, i.e. canonical variates, to compare Border Cave with crania drawn from $k = 8$ recent African populations (Bushman males and females, Hottentot males, Zulu males and females, Sotho males and females and Venda males). When all discriminants are considered, Border Cave lies closest to the Hottentot centroid and is contained within the .05 limits of this distribution. This assignment should not be interpreted in a strict sense to exclude it from all Bushman populations. These and many more interesting sentences can be found in Rightmire (1979) which moreover contains comments by various interested scholars from different disciplines. CAMPBELL (1980) put the statistical comments in perspective and made useful suggestions which led to RIGHTMIRE (1981). Though we welcome the progress made by concentrating on typicality indices (from the F distribution) and the posterior probabilities (based on multivariate t densities as suggested by the semi-Bayesian approach), we are not completely satisfied because we prefer a classical statistical approach where these indices and probabilities are regarded as estimates for basic unknown parameters, estimates which should be equipped with standard deviations if not replaced by confidence intervals.

Our evaluation of Border Cave is based on a comparison with crania drawn from k = 8 recent African populations (see Table 1). We used samples from Van Vark's data bank. It is a pity that Hottentots were not available because Rightmire had concluded that Border Cave is closest to the Hottentot centroid. The figures in Table 1 were obtained by converting those of RIGHTMIRE (1979) Table 2 into Howell's measurement system.

The reader is invited to make univariate comparisons by looking at Table 1 and the column of standard deviations in Table 2 (Student's 2 sample test is the correct tool). Rightmire's Table 2 shows that Hottentot males and Bushman males are very similar so that we do not worry much about the missing Hottentots. It is clear from our Table 1 that most Border Cave scores are too large to fit one of the samples. Border Cave is not very "typical" for one of the 8 populations involved, in fact it looks rather "atypical". Hence prior and posterior probabilities should be regarded with suspicion because they are based on the assumption that Border Cave has randomly been drawn from one of the k populations involved. This complication deepened our insight. In fact we are grateful that we have chosen Border Cave for illustrative purposes and not some standard example from the area of medical diagnosis or anthropological sex-diagnosis. Physical anthropology has always been rich in motivating statisticians. We believe that the basic reason for this phenomenon is that sample sizes etc. are essentially limited in anthropology. In many other areas of application one can increase sample sizes easily and the interest in what can be proved on the basis of available data is much less tense. It is pertinent to anthropological discussions that uncertainties are expressed as clearly as possible. Anthropologists are keen on systematic errors, measurement errors, statistical errors, etc. They are always aware of the fact that population parameters are unknown and are insufficiently revealed by the available data which cannot be extended as easily as in experimental sciences. Missing data problems in anthropology are not caused by lazy or careless experimenters but by the true nature of the problem. It makes no sense to ignore Border Cave because Horton should have been more careful.

The next step in evaluating Border Cave is to take into account the multivariate character of the data, e.g. by computing Mahalanobis distances and canonical variates and by performing Hotelling tests instead of Student tests. Table 3 presents some of our results. Note that $\hat{\Delta}^2_{x;h,h}$ was defined in Section 2 and $\hat{\Delta}^2_{x;h}$ in Section 3. The null-hypothesis H_h can be tested that Border Cave is from the same population as the h'th sample by referring the Hotelling T^2 statistic

$$(7.1) \qquad n_h(n-k-p+1)n^{-1}(n_h+1)^{-1}\ p^{-1}\ \hat{\Delta}^2_{x;h,h}$$

to the F(p,n-k-p+1) distribution if normality is postulated together with homogeneity of covariances (see Section 2 and RAO (1965) 8.b.2.XII) and by testing

$$(7.2) \qquad (n_h-p)(n_h+1)^{-1}\ p^{-1}\ \hat{\Delta}^2_{x;h}$$

in the $F(p,n_h-p)$ distribution if normality is postulated but no assumptions are made concerning the covariances (if one tries to verify that (7.1) and (7.2) coincide if $k = 2$ and $n = n_h + 1$ then one should notice that $n_h\ \hat{\Delta}^2_{x;h,h} = n\ \hat{\Delta}^2_{x;h}$: the m.l.estimators of the sections 2 and 3 are different because the underlying models differ).

Of course it is informative to compute the probability that the observed outcome is exceeded. These F-probabilities can be found in Table 3 and are called "typicality probabilities" in the recent publications of Campbell and Rightmire. Note that all these F-probabilities are smaller than .05 which means that Border Cave is significantly atypical at the 5% level for all recent African populations considered. This is in perfect agreement with expectations based on the earlier univariate comparisons.

We would like to add the following idea. The above mentioned typicality probabilities measure the typicality of Border Cave with respect to the *sample* from population h. If anything should be called the typicality of Border Cave with respect to *population* h, then this should be the *unknown parameter*

$$(7.3) \qquad \alpha_h(x) = P(G > (x-\mu_h)^T \textstyle\sum_h^{-1}(x-\mu_h))$$

where G has the $\chi^2(p)$ distribution. This entails the problem to construct a *confidence interval* for the typicality probability $\alpha_h(x)$ of Border Cave with respect to population h. The required confidence interval is easily obtained by transforming the confidence interval for the unknown parameter (2.2), if $\Sigma_1 = \ldots = \Sigma_k$ is postulated, or (3.3) of homogeneity of covariances is not required. Now x is regarded as a prescribed constant and not as a random drawing as in the theory behind (7.1) and (7.2). An exact confidence interval for $\Delta^2_{x;h,h}$, under the assumptions of Section 2, follows from the distributional result that

$$(7.4) \qquad n_h(n-k-p+1)\, n^{-1}\, p^{-1}\, \hat{\Delta}^2_{x;h,h}$$

has the noncentral F distribution with p and n - k - p + 1 d.f.'s and noncentrality parameter $n_h\, \Delta^2_{x;h,h}$ (see e.g. RAO (1965) 8.b.2.XII). In practice one might content oneself with approximate results based on the unbiasedness of

$$(7.5) \qquad (n-k-p-1)^{-1}\, n^{-1}\, \hat{\Delta}^2_{x;h,h} - n_h^{-1}\, p$$

as an estimator for $\Delta^2_{x;h,h}$ and the corresponding variance

$$(7.6) \qquad (n-k-p-3)^{-1}\, \{2\, \Delta^4_{x;h,h} + 4(n-k-1)n_h^{-1}\, \Delta^2_{x;h,h} + 2p(n-k-1)n_h^{-2}\}$$

We applied this approach to Border Cave and Zulus. Starting from $\hat{\Delta}_{x;3,3} = 6.66$ we obtained the approximate outcomes 42 and 13 for (7.5) and (7.6) with the consequence that [35,49] is an approximate confidence interval for $\Delta^2_{x;3,3}$. Converting this by means of (7.3) delivers a confidence interval for $\alpha_3(x)$ "left of everything". This makes very clear that Border Cave is very atypical for the Zulu population, at least if $\Sigma_1 = \ldots = \Sigma_k$ is postulated.

If the assumptions of Section 2 are weakened to those of Section 3, then the uncertainty is increased considerably because Σ_h has to be estimated on the basis of sample h only. Instead of (7.4) we now obtain that

$$(7.7) \qquad (n_h-p)\, p^{-1}\, \hat{\Delta}^2_{x;h}$$

has the noncentral $F(p;n_h-p;n_h\, \Delta^2_{x;h})$ distribution and that

(7.8) $(n_h-p-2)n_h^{-1}\;\hat{\Delta}^2_{x;h} - n_h^{-1}\;p$

is an unbiased estimator for $\Delta^2_{x;h}$ with variance

(7.9) $(n_h-p-4)^{-1}\{2\;\Delta^4_{x;h} + 4\;(n_h-2)n_h^{-1}\;\Delta^2_{x;h} + 2p\;(n_h-2)n_h^{-2}\}$

Applying this approach to Border Cave and Zulus we obtain from $\hat{\Delta}_{x;3} = 6.14$ in Table 3 and $n_h = 55$, $p = 11$ that (7.8) and (7.9) are approximately equal to 28.5 and 5.6. Hence [23.7,33.3] is an approximate confidence interval for $\Delta^2_{x;3}$. Conversion by means of χ^2_{11} delivers [.0005,.02] as approximate confidence interval for the typicality probability (7.3) of Border Cave with respect to the *population* of Zulus. The impression made by this result differs from that made by the F-probability in Table 3 though a common feature is that everything is at the left of .05: whichever way we turn, Border Cave is certainly not random drawing from the population of Zulus.

The next step in evaluating Border Cave is to compute approximate confidence intervals for posterior probabilities because (1) RIGHTMIRE (1981) computed similar estimates, (2) we like to illustrate our theory. It is obvious that these computations are not very relevant because F-tests and confidence intervals for typicality probabilities show that Border Cave cannot be regarded as a random drawing from one of the populations involved. Results of the computations are presented in Table 3. The enormous standard deviations for Bushman males and Zulu males show that *if Border Cave were known to be Bushman, Zulu, Dogon or Teita,* then it will either be a Bushman male or a Zulu male. It is impossible to discriminate between these two possibilities.

Our final step in evaluating Border Cave is to conclude from Border Cave's atypicality that no interpretations should be based on the confidence intervals for the posterior probabilities. One should rather reexamine Rightmire's original conclusion that Border Cave lies closest to the Hottentot centroid. His conclusion refers to the centroid of the Hottentot *sample*. It seems very relevant to us to pose questions of the following kind. Does sufficient evidence exist for the statement that the Border Cave *specimen* is closer to the Hottentot *population* than to e.g. that of Teita males? The underlying concept of distance requires that $\Sigma_1 = \ldots = \Sigma_k$ is postulated.

We become interested in the testing problem where the null-hypothesis

$$H: \quad \Delta_{x;1,1} = \Delta_{x;7,7}$$

is to be tested that "the triangle with apex x (=Border Cave) and other vertices μ_1(=Bushman males ≈ Hottentot males) and μ_7(=Teita males) has equal legs". Similar problems (with Σ known) were met with in Van Vark's research. The reader should notice that testing H is an extremely complicated affair. We were not able to develop satisfactory exact theory. So we took refuge in asymptotic theory. Instead of reporting the corresponding results we content ourselves, and hopefully the reader by using the following crude approach. In (7.5) and (7.6) we learned how to construct an approximate confidence interval for $\Delta^2_{x;h,h}$. It is true that certain dependencies are hidden behind the intervals for h = 1 (Bushman males) and h = 7 (Teita males). The abovementioned asymptotic theory accounts for these dependencies. However, if approximate confidence intervals are available for $\Delta^2_{x;1,1}$ and $\Delta^2_{x;7,7}$ then the interpretation will be clear. Starting from $\hat{\Delta}_{x;1,1} = 6.48$ ($\hat{\Delta}_{x;7,7} = 7.38$) we obtained the approximate outcomes 39 and 13 (52 and 20) for (7.5) and (7.6) with the consequence that [32,46] is an approximate confidence interval for $\Delta^2_{x;1,1}$ (and [43,61] for $\Delta^2_{x;7,7}$). These crude computations suggest that the difference between $\hat{\Delta}_{x;1,1} = 6.48$ and $\hat{\Delta}_{x;7,7} = 7.38$ is on the verge of being significant. *The Border Cave specimen is closer to Bushman males than to Bushman females, Dogon females and Teita females. It is very likely that Border Cave is also closer to Bushman males than to Teita males.*

We are reluctant to make further definite statements especially because we are of opinion that it is less interesting to compare Border Cave itself with other populations than it is to compare the *population* from which Border Cave is drawn with other *populations*. In that case not much more can be concluded than that the population, from which Border Cave is drawn, differs from the recent African populations considered. We want to prove this assertion by showing that the difference between the Border Cave *population* and that of Dogon females is on the verge of being significantly larger than the difference between the Border Cave *population* and that of Bushman males. Let μ denote the vector of expectations for the Border Cave population. Earlier in this section we learned that $H_h : \mu = \mu_h$ can be tested by referring (7.1) to the F(p,n-k-p+1) distribution. Now we are interested in a confidence interval for $(\mu-\mu_h)^T \Sigma^{-1}(\mu-\mu_h) = \Delta_h^2$. We want to show that the confidence intervals for Δ_1^2 and for Δ_6^2 show so much overlap that $H : \Delta_1^2 = \Delta_6^2$ is on the verge of rejection. An exact confidence interval for Δ_h^2 follows immediately by noting that (7.1) has the noncentral F'(p,n-k-p+1, $n_h (n_h+1)^{-1} \Delta_h^2$) distribution. We are content with crude approximate results based on the consequence of the above-mentioned result that

$$(7.10) \qquad n^{-1}(n-k-p-1)\ \hat{\Delta}^2_{x;h,h} - n_h^{-1}(n_h+1)\ p$$

is an umbiased estimator for Δ_h^2 with variance

$$(7.11) \qquad \begin{aligned} &n^{-2}\ p^{-2}(n-k-p-3)^{-1}(n-k-p+1)^2\{2p(n-k-1) + \\ &+ 4\ n_h(n_h+1)^{-1}(n-k-1)\ \Delta_h^2 + 2\ n_h^2(n_h+1)^{-2}\ \Delta_h^4\ \} \end{aligned}$$

Starting from $\hat{\Delta}_{x;1,1} = 6.48$, $n_1 = 41$, p= 11 and n = 375 we obtain the outcomes 29 and 1.1 for (7.10) and (7.11). Hence [27,31] is an approximate confidence interval for Δ_1^2. Starting from $\hat{\Delta}_{x;6,6} = 8.30$ and $n_6 = 53$, we similarly obtain outcomes 54 and 2.2 for (7.10) and (7.11). Hence [49,59] is an approximate confidence interval for Δ_6^2. We see that our supposition was wrong. The Border Cave population is definitely different from that of recent Dogon females.

Starting from $\hat{\Delta}_{x;2,2} = 7.59$ we obtain also that the approximate confidence intervals show no overlap though they are very close to each other.

Conclusion. Border Cave specimen is closer to Bushman males than to Bushman males, Dogon females and Teita females and probably also to Teita males. If one considers the mean μ of the population from which Border Cave is regarded as a random drawing, then the extra uncertainty is less than we had expected. The same conclusions can be made for the population parameter μ as for the score vector x.

Table 1

	measurements	Border Cave	Bushmen		Zulu		Dogon		Teita	
			males N_1 = 41	females N_2 = 49	males N_3 = 55	females N_4 = 46	males N_5 = 48	females N_6 = 53	males N_7 = 34	females N_8 = 49
1	SOS, Supraorbital projection	10	6.73	5.69	6.18	5.24	5.40	4.08	6.44	4.94
2	FMB, Bifrontal breadth	112	97.27	93.90	101.98	97.74	99.54	94.34	100.06	95.43
3	NAS, Nasio-frontal subtense	15	15.41	16.20	17.84	16.48	16.46	15.45	18.79	17.12
4	NFA, Nasio-frontal angle	150	143.20	143.65	141.51	142.70	143.46	143.68	138.88	140.49
5	WMH, Check height	21	20.93	19.84	20.73	20.06	21.21	19.96	22.21	20.18
6	FRC, Nasion-bregma chord	116	109.17	105.10	111.69	109.39	110.00	105.66	108.71	105.76
7	FRS, Nasion-bregma subtense	32	28.46	28.22	27.71	27.70	26.69	25.64	26.62	27.02
8	FRF, Nasion-subtence fraction	51	47.59	45.08	47.16	46.04	47.88	44.62	48.82	47.37
9	FRA, Frontal angle	122	124.29	122.73	126.33	125.33	127.58	127.28	127.41	125.43
10	OBB, Orbit breadth, left	45	39.27	37.67	40.44	39.20	39.71	38.08	39.65	37.76
11	MDH, Mastoid height	26	25.24	21.61	28.42	25.61	29.06	25.21	29.09	24.18

Measurements of Border Cave compared with means for eight modern African populations.

Table 2

		standard deviation	Correlation-matrix										
			1	2	3	4	5	6	7	8	9	10	11
1	SOS	1.18	1.00										
2	FMB	3.43	0.29	1.00									
3	NAS	2.21	0.25	0.33	1.00								
4	NFA	4.43	-0.18	-0.07	-0.96	1.00							
5	WMH	2.17	0.06	0.26	0.03	0.04	1.00						
6	FRC	4.63	0.04	0.22	0.09	-0.04	0.22	1.00					
7	FRS	2.62	-0.01	0.06	-0.12	0.15	-0.02	0.60	1.00				
8	FRF	3.46	-0.03	0.10	0.02	0.01	0.22	0.53	0.18	1.00			
9	FRA	3.78	0.03	0.04	0.19	-0.19	0.14	-0.18	-0.89	0.18	1.00		
10	OBB	1.65	0.08	0.63	0.25	-0.09	0.04	0.11	-0.01	0.02	0.06	1.00	
11	MDH	3.14	0.11	0.18	0.04	0.01	0.16	0.12	0.03	0.00	0.02	0.12	1.00

Standard deviations and correlation matrix for the eleven measurements in the eight populations for the case with homogeneity of dispersion matrices.

Table 3

			N_i	with homogeneity of dispersion matrices				without homogeneity of dispersion matrices			
				$\hat{\Delta}_{x;h,h}$	F-prob	posterior probability	standard dev. post-prob.	$\hat{\Delta}_{x,h}$	F-prob	posterior probability	standard dev. post. prob.
1	Bushmen	males	41	6.48	0.025	0.749	0.283	6.50	0.044	0.056	0.311
2		females	49	7.59	0.008	0.000	0.000	7.69	0.014	0.000	0.000
3	Zulu	males	55	6.66	0.021	0.231	0.271	6.14	0.046	0.943	0.313
4		females	46	7.13	0.013	0.009	0.013	7.34	0.020	0.001	0.004
5	Dogon	males	48	7.13	0.013	0.009	0.013	7.45	0.018	0.000	0.001
6		females	53	8.30	0.004	0.000	0.000	9.76	0.002	0.000	0.000
7	Teita	males	34	7.38	0.009	0.001	0.002	9.87	0.004	0.000	0.000
8		females	49	7.93	0.006	0.000	0.000	10.96	0.001	0.000	0.000

Mahalanobis distances, F-probabilities, posterior probabilities and standard deviations of posterior probabilities for the two cases with and without homogeneity of dispersion matrices.

8. APPENDIX

This appendix is devoted to the rationale behind the choice of the neighbourhood (5.14) recommended in the first part of Section 5. We recall that the problem of choosing U has two conflicting aspects: (1) U should be small in order to guarantee that the bias $\bar{\rho}_{t|x} - \rho_{t|x}$ is small, (2) U should be large because (5.6) shows that the asymptotic variance explodes if p_h gets small.

This appendix will be based on asymptotic theory, partly suggested by the theory of nonparametric density estimation. There is one interesting difference. Most results in the latter theory deal with the minimization of the (integrated) *mean square error* which equals variance + squared bias of the involved *estimators*. The corresponding solutions are such that variance and squared bias are of the same order of magnitude. Our main concern is the construction of *reliable confidence intervals*. A basic requirement in this connection is (5.8). This requirement will not be satisfied if the bias $\bar{\rho}_{t|x} - \rho_{t|x}$ is of the same order of magnitude as the estimated standard deviation $n^{-1/2}\{\hat{\Psi}\,\hat{\Delta}\,\hat{\Psi}_{t,t}\}^{1/2}$ in (5.7). When dealing with confidence intervals, U has to shrink slightly faster than when dealing with pure estimation.

The inner-product $(y,z)_M = y^T M^{-1} z$ based on the total-sample covariance matrix M is such that orthogonalization processes can be carried out leading to transformed measurement vectors e.g. $\tilde{X}_{hi\cdot} = M^{-1/2}(X_{hi\cdot} - X_{\cdot\cdot})$, with mean $\tilde{X}_{\cdot\cdot} = 0$ and total-sample covariance matrix I_p. The $\tilde{X}_{hi\cdot}$'s satisfy $\tilde{X}_{\cdot\cdot} = 0$ and $(n-1)^{-1}\,\Sigma\Sigma\,\tilde{X}_{hi\cdot}\,\tilde{X}_{hi\cdot}^T = I$. Their outcomes in $\mathbb{R}^p$ are scattered around the origin somewhat like a sample of n elements from a distribution with expectation 0, covariance matrix I and density f. The transformed vector $\tilde{x}$ of scores for the investigated individual e.g. $\tilde{x} = N^{-1/2}(x - x_{\cdot\cdot})$, is at Euclidian distance $\|\tilde{x}\| = d_M(x, x_{\cdot\cdot}) = d$ from the origin (see (5.10)). In Section 5 a neighbourhood U of x was needed for the construction of interval estimates for the posterior probabilities. The basic question was which choice should be made for the radius $r = r(p,d,n)$ of the the ball U around x. This extremely difficult problem is tackled by replacing it for the following related, but certainly not equivalent, problem.

Let $X_1,\dots,X_n$ be an i.r.s. from the unknown density f on $\mathbb{R}^p$. Let x be a point in $\mathbb{R}^p$ at distance $\|x\| = d$ from the origin. The value f(x) of the density in x is estimated by means of

$$(8.1) \qquad f_n(x) = \# \{i \; ; \; x_i \in U_n\} / \{n \, \lambda(U_n)\}$$

where

$$(8.2) \qquad U_n = \{y \in \mathbb{R}^p , \; \|y-x\| \leqq r(p,d,n)\}$$

is the ball with centre x and radius r(p,d,n)

and

$$(8.3) \qquad \lambda(U_n) = 2\pi^{\frac{1}{2}p}\{r(p,d,n)\}^p/\{p \, \Gamma(\tfrac{1}{2}p)\}$$

is the volume of U_n.

The numerator in (8.1) has binomial distribution $B(n,p_n)$ with

$$(8.4) \qquad p_n = P(X_i \in U_n) = \int_{U_n} f(y)d\lambda(y) \sim f(x)\lambda(U_n)$$

Bias and variance of the estimator (8.1) for f(x) are of the utmost importance. Using

$$f(y) \sim f(x) + \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_p}\right)(y-x) + \tfrac{1}{2}(y-x)^T A(y-x)$$

where $[A_{ij}] = \left[\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right]$ is a symmetric matrix whose eigenvalues will be denoted by $\lambda_1,\ldots,\lambda_p$, we obtain from (8.4) that the bias

$$(.8.5) \qquad Ef_n(x) - f(x) \sim \tfrac{1}{2}\int_{U_n} (y-x)^T A(y-x)d\lambda(y)/\lambda(U_n) = c_1(x)\{r(p,d,n)\}^2$$

where

$$(8.6) \qquad \tfrac{1}{2}\min(\lambda_1,\ldots,\lambda_p) \leqq c_1(x)p^{-1}(p+2) \leqq \tfrac{1}{2}\max(\lambda_1,\ldots,\lambda_p)$$

For the variance we obtain

$$(8.7) \qquad \begin{aligned} \mathrm{Var}\, f_n(x) &= n^{-1}p_n(1-p_n)/\{\lambda(U_n)\}^2 \\ &\sim n^{-1}f(x)/\lambda(U_n) = c_2 \, n^{-1}\{r(p,d,n)\}^{-p} \end{aligned}$$

where

(8.8) $$c_2 = \tfrac{1}{2}f(x)p\ \Gamma(\tfrac{1}{2}p)\ \pi^{-\frac{1}{2}p}$$

It follows immediately from (8.5) and (8.7) that the rate at which the mean square error $E(f_n(x)-f(x))^2$ tends to zero as $n \to \infty$, is maximized if

(8.9) $$r(p,d,n) \sim c_3(p,d)\ n^{-1/(4+p)}$$

because then both the variance (8.7) and the squared bias (8.5) are of the same order of magnitude as $n^{-4/(4+p)}$ whereas otherwise one of the two contributions to m.s.e. show a slower rate of convergence.

However, we are not primarily interested in the construction of an optimal (sequence of) estimator(s). We need a confidence interval for f(x) such that the analogue of (5.8) is satisfied. It is obvious that this cannot be established unless $\text{bias}^2/\text{var} \to 0$. If we need a reliable confidence interval then, instead of (8.9), we shall have to require that

(8.10) $$\lim_{n\to\infty} n^{1/(4+p)} r(p,d,n) = 0$$

Further specification such that r(p,0,n) *is completely determined.* If f is the p.d.f of $N_p(0,I)$ and $x = 0$, then a considerable negative bias $E\ f_n(0) - f(0)$ may be expected because $f(0) > f(y)$ for all $y \neq 0$. Focussing on this crucial example we shall require that

(8.11) $$|\text{bias}|\ /\text{standard deviation} \sim \{\ell n\ n\}^{-1}$$

The only rationale behind this formula is that $\text{bias}^2/\text{var} \to 0$ is needed for the confidence interval to be asymptotically reliable. On the other hand, the slower this convergence the larger the rate at which the mean square error tends to 0. The right-hand side provides such slow convergence. The value .22 for n = 100 does not seem unrealistic.

The previous order of magnitude considerations leading to (see (8.5) and (8.6)

$$(8.12) \qquad \text{bias } f_n(0) \sim -\tfrac{1}{2}(2\pi)^{-\frac{1}{2}p} p(p+2)^{-1}\{r(p,0,n)\}^2$$

and

$$(8.13) \qquad \text{var } f_n(0) \sim 2^{-\frac{1}{2}p-1}\, \pi^{-p}\, p\, \Gamma(\tfrac{1}{2}p)\, n^{-1}\{r(p,o,n)\}^{-p}$$

can be replaced by rigorous limit theorems based on

$$p_n = P(X_i \in U_n) = P(\chi_p^2 \leqq r^2(p,0,n)) = \int_0^{\frac{1}{2}r^2} e^{-y}\, y^{\frac{1}{2}p-1} dy/\, \Gamma(\tfrac{1}{2}p)$$

and

$$(8.14) \qquad \lim_{r\to 0} \left\{ \int_0^{\frac{1}{2}r^2} e^{-y}\, y^{\frac{1}{2}p-1}\, dy - \int_0^{\frac{1}{2}r^2} (1-y)\, y^{\frac{1}{2}p-1}\, dy \right\} /r^{p+3} = 0$$

By combining (8.11), (8.12) and (8.13), with equality sign = instead of the wiggles, we obtain that

$$(8.15) \qquad r(p,0,n) = \left[2^{\frac{1}{2}p+1}(p+2)^2\, \Gamma(\tfrac{1}{2}p) \,/ \left\{ p\, n(\ell n\, n)^2 \right\} \right]^{1/(p+4)}$$

REMARK. Crude computations delivered $r(1,0,100) = .43$ with $p_n = P(-.43 \leq U \leq .43) = .33$;$r(2,0,100) = .50$ with $p_n = P(\chi_2^2 \leqq .25) = 1 - e^{-.125} = .12$;$r(4,0,100) = .66$ with $p_n = P(\chi_4^2 \leqq .44) = .02$. The interpretation of the rapid decrease of p_n as p gets larger than 2 or 3 is obvious. If we control the bias by requiring (8.11), then the variance of the estimator gets large. If both bias and variance are controlled by requiring (8.11) and that the standard deviation of $f_n(0)$ is no more than say 20% of the true value $f(0) = (2\pi)^{-\frac{1}{2}p}$, then one will see that these conflicting aims cannot be satisfied unless n is sufficiently large. Crude evaluation by means of (8.15) and (8.13) leads to the requirement that n should satisfy

$$(8.16) \qquad n^4(\ell n\, n)^{-2p} \geqq 25^{p+4}\, 2^{-4}\, p^{2p+4}(p+2)^{-2p}\{\Gamma(\tfrac{1}{2}p)\}^4$$

This formula should not be taken too seriously. If (8.11) is changed by multiplying the righthand side with a factor 2 and if one impairs the other restriction by only requiring that the standard deviation of $f_n(0)$ is no more than 30% of $f(0)$, then (8.16) is drastically modified because the radius gets $4^{1/(p+4)}$ times as large and the factor $25^{p+4} = (.04)^{-(p+4)}$ is replaced by

$11^{p+4} = (.09)^{-(p+4)}$. In fact one obtains

$$(8.17) \qquad n^4(\ell n\, n)^{-2p} \geqq 11^{p+4}\, 2^{-2p-4}\, p^{2p+4}(p+2)^{-2p}\{\Gamma(\tfrac{1}{2}p)\}^4$$

Defining r(p,d,n) *by extending* r(p,0,n). The idea behind (5.13) is clear. If $x \neq 0$ is at a distance $d = \|x\|$ from the origin then we define the ball U around x such that its probability is the same as that for the ball around 0 with radius r(p,0,n), if f is the density of $N_p(0,I)$.

ACKNOWLEDGEMENTS

We wish to thank Dr. G.N. Van Vark for suggesting the Border Cave application and for the use of his data of the South African Negro crania. We thank Dr. G.P. Rightmire for his aid in transforming the Border Cave data into Howell's measurement system.

REFERENCES

AMBERGEN, A.W. (1981), *Approximate confidence intervals for posterior probabilities*. Report TW-224. Depart. of Math., postbox 800, Groningen.

AMBERGEN, A.W. & W. SCHAAFSMA (1982). *Interval estimates for posterior probabilities*. To be published.

ANDERSON, T.W. (1973), *An asymptotic expansion of the distribution of the studentized classification statistic* W. Ann. of Stat., Vol 1, No. 5, 964-972.

BEAN, S.J. & C.P. TSOKOS (1980), *Developments in nonparametric density estimation*. Int. Stat. Review, 48,267-287.

BOWKER, A.H. & R. SITGREAVES (1961). *An asymptotic expansion for the distribution function of the W-classification statistic*. In Solomon, H.

CACOULLOS, T. (1966), *Estimation of a multivariate density*. Ann. of the Inst. of Stat. Math., 18, 178-189.

CACOULLOS, T. (1973), *Discriminant analysis and applications,* Ac. Press.

CAMPBELL, N.A. (1980), *On the study of the Border Cave Remains: Statistical Comments.* Current Anthropology, Vol. 21, No. 4, August 1980, 532-535.

FISHER, R.A. (1936), *The use of multiple measurements in taxonomic problems,* Ann. Eugenics 7, 179-188.

GANESALINGAM, S. & G.L. MCLACHLAN, (1979). *A case study of two clustering methods based on maximum likelihood,* Statistica Neerlandica, Vol. 33, NR 2.

GEISSER, S.(1964), *Posterior odds for multivariate normal classifications.* Journal of the Royal Statistical Society, Series B, 26, 69-76.

HABBEMA, J.D.F. & J. HERMANS, (1978). *Statistical methods for clinical decision making.* Thesis, Leiden University.

HABBEMA, J.D.F., J. HERMANS, & K VAN DER BROEK, (1974), *A stepwise discriminant analysis program using density estimation.* In G. Bruckmann (Ed.), Compstat 1974, Proceedings in Computational Statistics. (Physica Verlag, Wien).

HOWELLS, W.W. (1973), *Cranial variation in man. A study by multivariate analysis of patterns of difference among recent human populations.* Papers of the Peabody Museum 67.

MCLACHLAN, G.J. (1979). *A comparison of the estimate and predictive methods of estimating posterior probabilities.* Commun. Statist.Theor. Meth. A8(9), 919-929.

OKAMOTO, M. (1964). *An asymptotic expansion for the distribution of the linear discriminant function.* Ann. Math. Stat., 34, 1286-1301, correction 39, 1358-1359.

PARZEN E. (1962). *On estimation of a probability density function and mode,* AMS, 33, 1065-1076.

ROSENBLATT, M. (1956), *Remarks on some nonparametric estimates of a density function.* AMS, 27, 832-837.

RAO, C.R. (1965), *Linear statistical inference and its applications*, Wiley.

RIGHTMIRE, G.P. (1979), *Implications of Border Cave skeletal remains for Later Pleistocene Human Evolution*, Current Anthropology, Vol. 20, No. 1, March 1979, 23-35.

RIGHTMIRE, G.P. (1981), *More on the study of the Border Cave remains*, Current Anthropology, Vol. 22, No.2, 199-200.

SCHAAFSMA, W. (1976), *The asymptotic distribution of some statistics from discriminant analysis*, Report TW-176, Dept. Math, Postbox 800, Groningen.

SCHAAFSMA, W. (1982), *Selecting variables in discriminant analysis for improving upon classical procedures*. To appear in: Kanal, L. and Krishnaiah, P.R. (eds.), Handbook of Statistics 2. North Holland, Amsterdam.

SCHAAFSMA, W. & T. STEERNEMAN (1981), *Classification and discrimination procedures when the number of features is unbounded*, IEEE Transac. SMC 11/2, 144-151.

SCHAAFSMA, W. & G.N. VAN VARK (1977), *Classification and discrimination problems with applications*, *part* I, Statistica Neerlandica 31, 25-45.

SCHAAFSMA, W. & G.N. VAN VARK (1979), *Classification and discrimination problems with applications*, part II^{a}, Statistica Neerlandica 33, 91-126.

SITGREAVES, R. (1961), *Some results on the distribution of the* W-*classification statistic*. In Solomon, H.

SOLOMON, H. (1961), *Studies in item analysis and prediction*, Stanford Univ. Press.

STEIN, CH. (1966), *Multivariate analysis*, (mimeographed notes recorded by M.L. Eaton). Dept. Stat, Stanford.

VAN VARK, G.N. (1970), *Some statistical procedures for the investigation of prehistoric human skeletal material*, Thesis, Groningen University.

VAN VARK, G.N. & P.G.M. VAN DER SMAN (1982), *New discrimination and classification techniques in anthropological practice*, Zeitschrift für Morphologie und Anthropology, Vol. 3, No.1, 21-36.