The Annals of Statistics
1974, Vol. 2, No. 4, 763-774

BAYESIAN CLASSIFICATION: ASYMPTOTIC RESULTS[1]

BY C. P. SHAPIRO

Michigan State University

We have a population composed of two subpopulations whose probability properties are described by known univariate distribution functions, $G(x)$ and $H(x)$, respectively. The probability of observing an individual from the first population is θ, from the second is $1 - \theta$. We assume θ is a random variable with a prior distribution on (0, 1) and find the Bayes rule for classifying n observations as from G or from H when the loss function is equal to the number of misclassifications. The main results in the paper give the asymptotic properties of the Bayes rule and several proposed approximations.

1. Introduction. We have a population composed of two subpopulations whose probability properties are described by known univariate distribution functions $G(x)$ and $H(x)$, respectively. The probability of observing an individual from the first population is θ, from the second is $1 - \theta$. We assume θ is a random variable with prior distribution Λ on (0, 1) and find the Bayes rule for classifying n observations as from G or from H when the loss function is equal to the number of misclassifications. The main results in this paper concern the asymptotic properties of the Bayes rule and several proposed approximations. These limiting results are proven on the probability space conditional on $\theta = \theta_0$. Thus, while the rule being studied is a Bayes procedure, the examination of the rule is carried out from a frequentist point of view. That is, once the form of the rule is found, we consider the rule as a function of independent and identically distributed (i.i.d.) random variables and prove the asymptotic results conditional on $\theta = \theta_0$.

Specifically, we observe random variables $X_1, \cdots, X_n$ which, conditional on θ, are i.i.d. with distribution $F(x \mid \theta) = \theta G(x) + (1 - \theta)H(x)$ where G and H are known. Without loss of generality we assume F has density $f(x \mid \theta) = \theta g(x) + (1 - \theta)h(x)$ with respect to a sigma finite measure on the real line. The action space in the set of all sequences $\mathbf{a}^* = (a_1^*, \cdots, a_n^*)$ of 0's and 1's of length n and a *classification rule* $\mathbf{d}^* = (d_1^*, \cdots, d_n^*)$ is a measurable function from the sample space into the action space. We interpret $a_i^* = 1$ as classifying x_i from density g, $a_i^* = 0$ as classifying x_i from density h.

Define the *classification vector* $\mathbf{Z}^* = (Z_1^*, \cdots, Z_n^*)$ by

$$\begin{aligned} Z_i^* &= 1 \quad \text{if} \quad X_i \quad \text{is from} \quad g\,, \\ &= 0 \quad \text{if} \quad X_i \quad \text{is from} \quad h\,. \end{aligned}$$

Received October 1972; revised Juned 1973.

[1] Adapted from a portion of the author's dissertation at the University of Michigan.

AMS 1970 *subject classifications*. Primary 62C10; Secondary 62E20.

Key words and phrases. Bayesian inference, classification, asymptotic distribution, weak limit.

Thus, $\mathbf{Z}^*$ is the true classification of $\mathbf{X} = (X_1, \cdots, X_n)$. In this context, the joint density of $X_1, \cdots, X_n$ given $\mathbf{Z}^* = \mathbf{z}^*$ is

$$f(x_1, \cdots, x_n \mid \mathbf{Z}^* = \mathbf{z}^*) = \prod_{i=1}^n g(x_i)^{z_i^*} h(x_i)^{1-z_i^*},$$

and the density of $\mathbf{Z}^*$ given θ is

$$P(\mathbf{Z}^* = \mathbf{z}^* \mid \theta) = \theta^{\Sigma z_i^*}(1 - \theta)^{n-\Sigma z_i^*}.$$

The loss function is the number of misclassifications,

$$L(\mathbf{a}^*, \mathbf{Z}^*) = \sum_{i=1}^n (a_i^* - Z_i^*)^2.$$

In Section 2 we prove (Theorem 1) that the Bayes rule possesses a cut-type property which implies that it is completely determined by t_n, the proportion of observations classified to g. In Section 3 (Theorem 2), we find the limit of t_n which we denote by t_0, and hence the limiting form of the rule. In Section 4 we prove (Theorem 3) that under regularity conditions, $n^{\frac{1}{2}}(t_n - t_0)$ is asymptotically normal. In Section 5 we propose an approximation $\hat{t}_n$ of t_n and examine the rate at which $\hat{t}_n - t_n$ tends to zero (Theorem 4). In Section 6 we use iteration methods to obtain another approximation of t_n which is easy to compute and preserves the desired asymptotic properties of the first approximator. In Section 7, we remark on the application of our techniques to prove limiting results for a classical rule which is considered by Hannan and Robbins (1955). We also remark on extensions of our results to unknown g and h.

Throughout this paper we assume that $g \not\equiv h$ and that $g(x) = 0$ if and only if $h(x) = 0$.

2. The general form of the rule. Minimizing the posterior expected loss we find the Bayes rule is given by

$$a_i^* = 1 \quad \text{if} \quad P(Z_i^* = 1 \mid \mathbf{x}) > \tfrac{1}{2}, \qquad i = 1, \cdots, n,$$

where we take $a_i^* = 0$ if the probability is equal to $\frac{1}{2}$. Note that

$$P(Z_i^* = 1 \mid \mathbf{x}) = \sum_{\{\mathbf{z}^*:\, z_i^* = 1\}} P(\mathbf{Z}^* = \mathbf{z}^* \mid \mathbf{x})$$

where

$$P(\mathbf{Z}^* = \mathbf{z}^* \mid \mathbf{x}) = \frac{\prod_{i=1}^n g(x_i)^{z_i^*} h(x_i)^{1-z_i^*} \int \theta^{\Sigma z_i^*}(1 - \theta)^{n-\Sigma z_i^*}\, d\Lambda(\theta)}{\sum_{\mathbf{z}^*} \prod_{i=1}^n g(x_i)^{z_i^*} h(x_i)^{1-z_i^*} \int \theta^{\Sigma z_i^*}(1 - \theta)^{n-\Sigma z_i^*}\, d\Lambda(\theta)}.$$

To show that the Bayes rule is completely determined by t_n, the proportion classified g, we first need some notation. Let $Y_i = g(X_i)/h(X_i)$, $i = 1, \cdots, n$, and let $W_1, \cdots, W_n$ be the order statistics associated with $Y_1, \cdots, Y_n$. Define a new classification vector $\mathbf{Z} = (Z_1, \cdots, Z_n)$ by $Z_i = 1$ if that X_j corresponding to W_i is from g, $Z_i = 0$ if that X_j corresponding to W_i is from h. Define action $\mathbf{a}$ and rule $\mathbf{d}$ analogously. Also, let $\tilde{F}_\theta$, $\tilde{G}$, $\tilde{H}$ be the distribution of Y when X is distributed F_θ, G, H, respectively. Let $\tilde{f}_\theta$ be the density associated with $\tilde{F}_\theta$.

Let $[c]$ denote the greatest integer less than or equal to c. For t in $[0, 1]$, define

$$\phi_n(t, \theta) = \theta W_{n-[nt]} / \{\theta W_{n-[nt]} + 1 - \theta\},$$

where $W_0 = \inf g(x)/h(x)$. Let $\phi_n(t) = E[\phi_n(t, \theta) \mid \mathbf{x}]$. We now get the following theorem.

THEOREM 1. *The Bayes rule has the form* $\mathbf{d} = (0, \cdots, 0, 1, \cdots, 1)$ *where the number of* 1*'s is equal to* nt_n *and*

$$t_n = \inf \{t: \phi_n(t) \leqq \tfrac{1}{2}\}$$
$$= 1 \quad \text{if} \quad \phi_n(t) > \tfrac{1}{2} \quad \text{for all} \quad t.$$

PROOF. From Bayes Theorem we get

$$P(Z_i^* = 1 \mid \theta, \mathbf{x}) = \frac{\theta g(x_i)}{\theta g(x_i) + (1 - \theta)h(x_i)}$$

and

$$P(Z_i = 1 \mid \mathbf{x}) = \frac{\theta w_i}{\theta w_i + (1 - \theta)}.$$

Thus, $\phi_n(t, \theta) = P(Z_{n-[nt]} = 1 \mid \theta, \mathbf{x})$ for $t < 1$ and we have $\phi_n(t) = P(Z_{n-[nt]} = 1 \mid \mathbf{x})$. Also, $\phi_n(t, \theta)$ is decreasing in t for θ and $\mathbf{x}$ fixed and thus $\phi_n(t)$ is decreasing in t for fixed $\mathbf{x}$. This last property implies that $P(Z_i = 1 \mid \mathbf{x})$ is increasing in i which gives the form of the rule above. □

A classification rule of the form given in Theorem 1 is said to have the cut-type property since it makes exactly one cut in the order statistics $W_1, \cdots, W_n$, and classifies to g any X_j whose corresponding W_i is above that cut. This property allows us to find the limiting form of the rule by finding the limit of t_n.

To derive the cut-type property of the Bayes rule we needed no regularity assumptions on g and h or on the prior distribution. To derive the limiting properties of the rule we will need several conditions. We list them below. For the remainder of this paper we assume θ_0 is fixed in (0, 1).

A1. The functions $\log g(x)$ and $\log h(x)$ are integrable with respect to the measures induced by distributions G and H.

A2. The prior distribution Λ has density λ with respect to Lebesgue measure.

A3. Density λ is strictly positive in a neighborhood of θ_0.

A4. Density λ has two continuous derivatives in a neighborhood of θ_0 and finite second moment.

A5. Density λ has four continuous derivatives in a neighborhood of θ_0.

A6. The support of $\tilde{F}_{\theta_0}$ is an interval and $\tilde{F}_{\theta_0}$ is continuous.

A7. Density $\tilde{f}_{\theta_0}$ is continuous and strictly positive in a neighorhood of $(1 - \theta_0)/\theta_0$.

To find the limit of t_n we will need conditions A1—A3 and A6. To prove the asymptotic normality of $n^{\frac{1}{2}}(t_n - t_0)$, where t_0 is the a.s. limit of t_n, we will need A1—A4 and A6—A7. To derive an approximator of t_n, we will need A1—A3 and A5—A7.

3. The limiting form of the rule. In the last section we characterized t_n in

terms of the function $\phi_n(t)$. To find the limit of t_n, we will find the limit of $\phi_n(t)$, prove that this limit is uniform in t, and deduce the limit of t_n.

The main result we need to find, the limit of $\phi_n(t)$, is given in Lemma 1 below.

LEMMA 1. *Suppose conditions* A1—A3 *are satisfied. Then for any* $\delta > 0$, $P(|\theta - \theta_0| \geqq \delta \mid \mathbf{x})$ *tends to* 0 a.s. *given* θ_0.

The lemma follows easily from Lemma 2.3 of Johnson (1970) and a version of the uniform strong law given by Rubin (1956). The regularity conditions required for Johnson's lemma are conditions which imply the strong consistency and asymptotic normality of the maximum likelihood estimator $\hat{\theta}$ of θ. These are given by Johnson as Assumptions 1—9. We remark here that if g and h satisfy condition A1, then the class of mixtures, $f(x, \theta) = \theta g(x) + (1 - \theta)h(x)$, $0 < \theta < 1$, satisfies Johnson's Assumptions 1—9. In particular, the integrable functions of x which locally dominate the first and second order partial derivatives of $\log f(x, \theta)$ are

$$G_1(x) = (\theta_0 - \varepsilon)^{-1} + (1 - \theta_0 - \varepsilon)^{-1},$$
$$G_2(x) = (\theta_0 - \varepsilon)^{-2} + (1 - \theta_0 - \varepsilon)^{-2},$$

for θ in $(\theta_0 - \varepsilon, \theta_0 + \varepsilon)$ and $0 < \varepsilon < \min\{\theta_0, 1 - \theta_0\}$.

Let $\tilde{\xi}_t = \tilde{F}_\theta^{-1}(t)$ and define

$$\phi(t, \theta) = \theta\tilde{\xi}_{1-t}/\{\theta\tilde{\xi}_{1-t} + 1 - \theta\}.$$

THEOREM 2. *Suppose that conditions* A1—A3 *and* A6 *are satisfied. Then given* θ_0,

$$\phi_n(t) \to \phi(t, \theta_0) \qquad \text{a.s. } \textit{and uniformly in } t,$$

and thus,

$$t_n \to t_0 = 1 - \tilde{F}_{\theta_0}\left(\frac{1 - \theta_0}{\theta_0}\right) \qquad \text{a.s. } \textit{given } \theta_0.$$

PROOF. $\phi_n(t)$ is a decreasing bounded function of t for each $\mathbf{x}$ and by A6, $\phi(t, \theta_0)$ is continuous in t. Thus, the pointwise convergence of $\phi_n(t)$ to $\phi(t, \theta_0)$ will imply the uniform convergence by the same argument as in the nonrandom case. See Breiman (1968, page 160). Hence, it suffices to prove pointwise convergence.

Let $\lambda_n(\cdot)$ denote the posterior density of θ. Since ϕ_n and ϕ are bounded by 0 and 1, we have

$$|\phi_n(t) - \phi(t, \theta_0)| \leqq \int_{|\theta-\theta_0|\leq\delta} |\phi_n(t, \theta) - \phi(t, \theta_0)|\lambda_n(\theta)\, d\theta + P(|\theta - \theta_0| \geqq \delta \mid \mathbf{x}).$$

The second term above tends to zero a.s. given θ_0 by Lemma 1. The first term is less than or equal to

$$K_n(\mathbf{x}) \int_{|\theta-\theta_0|\leq\delta} |\theta - \theta_0|\lambda_n(\theta)\, d\theta$$

where $K_n(\mathbf{x})$ converges a.s. This term can be made arbitrarily small by the choice of δ. □

In the case of a degenerate prior the limit of t_n is the same as that given in Theorem 2.

4. Asymptotic normality results. Assuming that t_0 is in (0, 1), we will prove that under conditions A1—A4 and A6—A7, $n^{\frac{1}{2}}(t_n - t_0)$ is asymptotically normal.

Since $t_0 \neq 1$, eventually $t_n = \inf\{t: \phi_n(t) \leqq \frac{1}{2}\}$ a.s. Thus, for fixed real u,

$$P(n^{\frac{1}{2}}(t_n - t_0) \leqq u) = P(n^{\frac{1}{2}}[\phi_n(t_0 + un^{-\frac{1}{2}}) - \tfrac{1}{2}] \leqq 0)\,.$$

To prove the asymptotic normality of $n^{\frac{1}{2}}(t_n - t_0)$ we prove that $n^{\frac{1}{2}}(\phi_n(t_0 + un^{-\frac{1}{2}}) - \frac{1}{2})$, is asymptotically normal where "u" enters linearly in the mean of the asymptotic distribution. We first prove a lemma which allows us to consider $\phi_n(t, \hat{\theta})$ in place of $\phi_n(t)$, where $\hat{\theta}$ is the maximum likelihood estimator of θ. Let $W_n(u) = W_{n-[nt_0+un^{\frac{1}{2}}]}$.

LEMMA 2. *Suppose conditions* A1—A4 *and* A6—A7 *are satisfied. Then for fixed* u,

$$\phi_n(t_0 + un^{-\frac{1}{2}}) = \phi_n(t_0 + un^{-\frac{1}{2}}, \hat{\theta}) + O_{P_{\theta_0}}(n^{-1})\,.$$

PROOF. Expanding $\phi_n(t, \theta)$ in a Taylor series around $\hat{\theta}$, replacing t by $t_0 + un^{-\frac{1}{2}}$, and taking expectation conditional on $\mathbf{x}$, we get

$$\begin{aligned}\phi_n(t_0 + un^{-\frac{1}{2}}) = \phi_n(t_0 + un^{-\frac{1}{2}}, \hat{\theta}) + \phi_n'(t_0 + un^{-\frac{1}{2}}, \hat{\theta})E[\theta - \hat{\theta} \mid \mathbf{x}] \\ + E[\phi_n''(t_0 + un^{-\frac{1}{2}}, \theta^*)(\theta - \hat{\theta})^2 \mid \mathbf{x}]\,,\end{aligned}$$

where θ^* is between θ and $\hat{\theta}$ and ϕ_n' and ϕ_n'' are the first and second order partial derivatives of ϕ_n with respect to θ. We will show the last two terms above are $O_P(n^{-1})$.

A simple computation with order statistics shows that $\phi_n'(t_0 + un^{-\frac{1}{2}}, \hat{\theta})$ converges in probability given θ_0. Also, Theorem 3.1 of Johnson (1970) and conditions A1—A4 imply $E[\theta - \hat{\theta} \mid \mathbf{x}]$ is $O_P(n^{-1})$. Thus, the first term has the proper order.

For the second term,

$$|\phi_n''(t_0 + un^{-\frac{1}{2}}, \theta^*)| \leqq 2W_n(u)(1 + W_n(u))[\min\{\hat{\theta}^3 W_n(u)^3, (1 - \hat{\theta})^3\}]^{-1}$$

which converges in probability. Again, Theorem 3.1 of Johnson (1970) and conditions A1—A4 imply that $E[(\theta - \hat{\theta})^2 \mid \mathbf{x}]$ is $O_P(n^{-1})$. Thus, the last term has the proper order. □

By definition of t_0, $\frac{1}{2} = \theta_0\xi_{1-t_0}/\{\theta_0\xi_{1-t_0} + 1 - \theta_0\}$. Thus, if 0 is a continuity point of the limiting distribution of the first expression below, Lemma 2 implies

$$\begin{aligned}&\lim_{n\to\infty} P_{\theta_0}[n^{\frac{1}{2}}(\phi_n(t_0 + un^{-\frac{1}{2}}) - \tfrac{1}{2}) \leqq 0] \\ &\quad = \lim_{n\to\infty} P_{\theta_0}[n^{\frac{1}{2}}\{(1 - \theta_0)\hat{\theta}[W_n(u) - \xi_{1-t_0}] + \xi_{1-t_0}(\hat{\theta} - \theta_0)\} \leqq 0]\,.\end{aligned}$$

Each component in the sum in the probability above is marginally asymptotically normal. To prove the asymptotic normality of the sum we need the components jointly asymptotically normal. This follows from the fact that both components can be represented as sums of i.i.d. random variables.

In particular, from the usual argument used to prove that $n^{\frac{1}{2}}(\hat{\theta} - \theta_0)$ is asymptotically normal we get

$$(4.1) \qquad n^{\frac{1}{2}}(\hat{\theta} - \theta_0) = n^{-\frac{1}{2}} I(\theta_0)^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(x_i, \theta)|_{\theta=\theta_0} + o_{P_{\theta_0}}(1),$$

where $I(\theta_0)$ is the Fisher information at θ_0. A stronger result in Bahadur (1966) implies

$$(4.2) \qquad n^{\frac{1}{2}}(W_{n-[nt_0]} - \tilde{\xi}_{1-t_0}) = n^{-\frac{1}{2}} \tilde{f}_{\theta_0}(\tilde{\xi}_{1-t_0})^{-1} \sum_{i=1}^{n} \{I\{\tilde{\xi}_{1-t_0} - y_i\} - (1 - t_0)\} + o_{P_{\theta_0}}(1).$$

A simple computation with order statistics shows that $n^{\frac{1}{2}}(W_n(u) - W_{n-[nt_0]})$ tends in probability to $-u/\tilde{f}_{\theta_0}(\tilde{\xi}_{1-t_0})$. Thus, the component involving $W_n(u)$ is equivalent to a sum and the "u" will enter linearly in the mean of the asymptotic distribution.

We have now proved the following theorem.

THEOREM 3. *Suppose conditions* A1—A4 *and* A6—A7 *are satisfied. Then given* θ_0,

$$n^{\frac{1}{2}}(t_n - t_0) \to V \quad \textit{in distribution},$$

where V *is normal with mean* 0 *and variance*

$$t_0(1 - t_0) + \frac{\tilde{f}_0^2}{\theta_0^4 I(\theta_0)} + 2\frac{\tilde{f}_0}{\theta_0^2 I(\theta_0)}\left[\tilde{G}\left(\frac{1 - \theta_0}{\theta_0}\right) - \tilde{H}\left(\frac{1 - \theta_0}{\theta_0}\right)\right],$$

where $\tilde{f}_0 = \tilde{f}_{\theta_0}(\tilde{\xi}_{1-t_0})$.

We remark here that if we replace θ_0 with an estimator $\hat{\theta}^*$ of θ, which is equivalent to a sum of i.i.d. random variables with finite second moment, we get $n^{\frac{1}{2}}(t_n - t_0(\hat{\theta}^*))$ asymptotically normal. In particular, if $\hat{\theta}^* = \hat{\theta}$, the maximum likelihood estimator, then the asymptotic variance is

$$t_0(1 - t_0) + 3I(\theta_0)^{-1}\left[\tilde{G}\left(\frac{1 - \theta_0}{\theta_0}\right) - \tilde{H}\left(\frac{1 - \theta_0}{\theta_0}\right)\right]^2.$$

5. Approximation of t_n. To compute t_n, we must find $P(\mathbf{Z} = \mathbf{z} \mid \mathbf{x})$ for all classifications $\mathbf{z}$, then add the probabilities with $z_i = 1$ to get $P(Z_i = 1 \mid \mathbf{x})$. Since the number of computations involved is greater than 2^n, it is of interest to approximate t_n for n large. In the last section we remarked that for certain estimators $\hat{\theta}^*$ of θ, $n^{\frac{1}{2}}(t_n - t_0(\hat{\theta}^*))$ is asymptotically normal. Thus, we can approximate t_n by $t_0(\hat{\theta}^*)$. In this section we present a different type of approximator of t_n and prove that the probability of nt_n and the approximated nt_n differing by more than one observation tends to zero.

To define our approximator of t_n we need the results and methods of Johnson (1970) on the asymptotic expansion of the posterior distribution.

Specifically, let $\Psi(w, \theta) = w\theta/\{\theta w + 1 - \theta\}$. We need an asymptotic expansion of $E[\Psi(w, \theta) \mid \mathbf{x}]$ which is uniform in w for w bounded away from 0 and ∞. This is given in the following lemma.

LEMMA 3. *Let a and c be given such that $0 < a < c < \infty$. Suppose conditions* A1—A3 *are satisfied and let K be an integer. If the prior density λ has $K + 1$ continuous derivatives in a neighborhood of θ_0 then there exist functions $\gamma_j(w, \mathbf{x})$ and a constant D depending on a and c such that for* a.e. $\mathbf{x}$

$$|E[\Psi(w, \theta) \mid \mathbf{x}] - \textstyle\sum_{j=0}^{K} \gamma_j(w, \mathbf{x}) n^{-j/2}| \leqq D n^{-(K+1)/2}$$

for $n \geqq N_{\mathbf{x}}$ and w in $[a, c]$. The odd terms in the sum are zero.

PROOF. For fixed $w > 0$, the expansion above follows immediately from Johnson's methods where the integrable functions which locally dominate $\partial^k / \partial \theta^k \log f(x, \theta)$ are

$$G_k(x) = (k - 1)! \left((\theta_0 - \varepsilon)^{-k} + (1 - \theta_0 - \varepsilon)^{-k}\right).$$

To prove uniformity in w, it is easy to show that the constants involved in deriving the expansion depend only on a and c. □

We will use the expansion above with $K = 3$. In this case, the coefficients are

$$\gamma_0 = \gamma_0(w, \mathbf{x}) = \Psi(w, \hat{\theta}), \qquad \gamma_1(w, \mathbf{x}) = 0 = \gamma_3(w, \mathbf{x})$$

$$\gamma_2 = \gamma_2(w, \mathbf{x}) = b^{-2}\Psi''(w, \hat{\theta}) + 2b^{-2}\Psi'(w, \hat{\theta})\lambda'(\hat{\theta})\lambda(\hat{\theta})^{-1} + b^{-4}a_{3n}(\hat{\theta})\Psi'(w, \hat{\theta}),$$

where the primes denote partial derivatives with respect to θ and

$$a_{3n}(\theta) = n^{-1} \textstyle\sum_{i=1}^{n} \dfrac{\partial^3}{\partial \theta^3} \log f(x_i, \theta)/3!.$$

We first approximate $\phi_n(t)$ by the first two nonzero terms of the asymptotic expansion of $E[\Psi(w, \theta) \mid \mathbf{x}]$ evaluated at $w = W_{n-[nt]}$. We then define $\hat{t}_n$ as the "inverse" of the approximated $\phi_n(t)$ evaluated at $\frac{1}{2}$.

Specifically, define

$$\begin{aligned}\hat{\phi}_n(t) &= \gamma_0(W_{n-[nt]}, \mathbf{x}) + \gamma_2(W_{n-[nt]}, \mathbf{x}) n^{-1} \\ &= \gamma_0(t) + \gamma_2(t) n^{-1}.\end{aligned}$$

Let

$$\begin{aligned}\hat{t}_n &= \inf\{t : \hat{\phi}_n(t) \leqq \tfrac{1}{2}\} \\ &= 1 \qquad \text{if} \quad \hat{\phi}_n(t) > \tfrac{1}{2} \quad \text{for all} \quad t.\end{aligned}$$

Note that $\gamma_0(t) = \phi_n(t, \hat{\theta})$.

Under conditions A1—A3 and A5, Lemma 3 holds with $K = 3$. Thus we get

$$\sup_{\alpha \leqq t \leqq \beta} |\phi_n(t) - \hat{\phi}_n(t)| = O_{P_{\theta_0}}(n^{-2})$$

for any α, β such that $\alpha > 0$ and $\tilde{F}_{\theta_0}^{-1}(1 - \beta) > 0$. The main theorem relating t_n and $\hat{t}_n$ is given below.

THEOREM 4. *Suppose conditions* A1—A3 *and* A5—A7 *are satisfied and that t_0 is in* $(0, 1)$. *Then*

$$P_{\theta_0}[|t_n - \hat{t}_n| \leqq 1/n] \to 1.$$

The main idea of the proof is this. In a neighborhood of t_0, ϕ_n and $\hat{\phi}_n$ get

close faster than the smallest jump of ϕ_n or of $\hat{\phi}_n$ decreases to zero. Thus, t_n and $\hat{t}_n$ can differ by at most one jump point as n gets large.

Fix $\nu < \frac{1}{2}$, and define the set

$$A_n = \{t: |t - t_0| \leqq n^{-\nu}\}.$$

Define

$$B_n = \{\sup_{t \in A_n} |\phi_n(t) - \hat{\phi}_n(t)| < \inf_{t \in A_n} |\hat{\phi}_n(t) - \hat{\phi}_n(t \pm n^{-1})|\},$$
$$C_n = \{\sup_{t \in A_n} |\phi_n(t) - \hat{\phi}_n(t)| < \inf_{t \in A_n} |\phi_n(t) - \phi_n(t \pm n^{-1})|\}.$$

It is then easy to show that for fixed n, if t_n and $\hat{t}_n$ are in A_n, then B_n and C_n imply that $|t_n - \hat{t}_n| \leqq 1/n$. Thus, to prove Theorem 4 we must show that $P(\hat{t}_n, t_n \text{ in } A_n)$, $P(B_n)$, and $P(C_n)$ all tend to 1.

The asymptotic normality of $n^{\frac{1}{2}}(t_n - t_0)$ implies that $t_n - t_0$ is $O_P(n^{-\frac{1}{2}})$. Therefore, $P(\hat{t}_n, t_n \text{ in } A_n)$ tends to 1 for $\nu < \frac{1}{2}$ if we can show $\hat{t}_n - t_0$ is $O_P(n^{-\frac{1}{2}})$. We first note that $\hat{t}_n$ tends to t_0 a.s. given θ_0 from the same argument used to prove $t_n \to t_0$. The next lemma gives the appropriate order.

LEMMA 4. *Suppose conditions* A1 — A3, A5—A7 *are satisfied. If* t_0 *is in* (0, 1), *then* $\hat{t}_n - t_0$ *is* $O_{P_{\theta_0}}(n^{-\frac{1}{2}})$.

PROOF. $\hat{\phi}_n(t) = \phi_n(t, \hat{\theta}) + \gamma_2(t)n^{-1}$, where we recall that $\phi_n(t, \hat{\theta})$ is decreasing in t. Choose α, β such that $0 < \alpha < t_0 < \beta < 1$. Then it can be shown that

$$\sup_{\alpha \leqq t \leqq \beta} |\gamma_2(t)| = O_P(1).$$

Thus,

$$\sup_{\alpha \leqq t \leqq \beta} |\hat{\phi}_n(t) - \phi_n(t, \hat{\theta})| = O_P(n^{-1}).$$

Define $t_{1n}^* = \inf\{t: \phi_n(t, \hat{\theta}) - n^{-1} \leqq \frac{1}{2}\}$ and $t_{2n}^* = \inf\{t: \phi_n(t, \hat{\theta}) + n^{-1} \leqq \frac{1}{2}\}$. By an argument similar to that used to prove the asymptotic normality of $n^{\frac{1}{2}}(t_n - t_0)$, we can prove that $n^{\frac{1}{2}}(t_{1n}^* - t_0)$ and $n^{\frac{1}{2}}(t_{2n}^* - t_0)$ are each asymptotically normal. Thus, $t_{1n}^* - t_{2n}^*$ is $O_P(n^{-\frac{1}{2}})$.

Since $\hat{t}_n$ tends to t_0 in (0, 1), we have $\hat{t}_n$ eventually in (α, β) a.s. Thus, $\hat{t}_n$ is eventually between t_{1n}^* and t_{2n}^* with probability 1, and we get

$$|\hat{t}_n - t_0| \leqq |t_{1n}^* - t_{2n}^*| + |t_{2n}^* - t_0| = O_P(n^{-\frac{1}{2}}). \qquad \square$$

To prove $P(B_n)$ and $P(C_n)$ tend to 1, we need a lemma concerning order statistics.

LEMMA 5. *Fix* $\nu < \frac{1}{2}$ *and* t_0 *in* (0, 1). *Suppose* $Y_1, \cdots, Y_n$ *are* i.i.d. F, *where the support of* F *is an interval and* F *is continuous. Let* $W_1, \cdots, W_n$ *be the corresponding order statistics. Assume* F *has density* f *which is continuous and strictly positive in a neighborhood of* $F^{-1}(1 - t_0)$. *Then for* $\delta > 1 - \nu$ *and any* $M > 0$,

$$P[n^{\delta+1} \inf_{t \in A_n} |W_{n-[nt]} - W_{n-[nt]\pm 1}| \geqq M] \to 1.$$

PROOF. We prove the above with $n - [nt] - 1$. The proof is the same for $n - [nt] + 1$. Also, without loss of generality we assume that $F(x) = 1 - e^{-x}$, the standard exponential distribution. Then

$$n^{\delta+1} \inf_{t \in A_n} W_{n-[nt]} - W_{n-[nt]-1}$$

is equal in distribution to

$$\inf_{t\in A_n} n^{\delta+1}\frac{1}{[nt]+1}Y_{n-[nt]} \geqq \frac{n}{[nt_0+n^{1-\nu}]+1}n^{\delta}\inf_{t\in A_n} Y_{n-[nt]}\,.$$

Fix $M>0$. Then

$$P(n^{\delta}\inf_{t\in A_n} Y_{n-[nt]} \geqq M) \simeq \exp[-2Mn^{1-\delta-\nu}]$$

which tends to 1 for $\delta > 1-\nu$. □

PROOF OF THEOREM 4. Lemma 4 implies $P(t_n, \hat{t}_n \text{ in } A_n)$ tends to 1. Thus, to complete the proof we must show that $P(B_n)$ and $P(C_n)$ tend to 1. Fix δ, ν such that $1-\nu<\delta<1$ and $0<\nu<\frac{1}{2}$.

Using the particular form of $\phi_n(t)$, and methods similar to those used in deriving the limit of $\phi_n(t)$, we get

$$\inf_{t\in A_n} n^{1+\delta}|\phi_n(t)-\phi_n(t\pm n^{-1})| \geqq \inf_{t\in A_n} n^{1+\delta}|W_{n-[nt]}-W_{n-[nt]\pm 1}|R_n^{*}(\mathbf{x})\,,$$

where $R_n^{*}(\mathbf{x})$ is bounded away from 0. Lemma 5 implies this tends to infinity in probability.

From the form of $\hat{\phi}_n(t)$, we get

$$\begin{aligned}\inf_{t\in A_n} n^{1+\delta}&|\hat{\phi}_n(t)-\hat{\phi}_n(t\pm n^{-1})| \\ &= n^{1+\delta}\inf_{t\in A_n}|\phi_n(t,\hat{\theta})-\phi_n(t\pm n^{-1},\hat{\theta})+n^{-1}[\gamma_2(t)-\gamma_2(t\pm n^{-1})]| \\ &\geqq \inf_{t\in A_n} n^{1+\delta}|W_{n-[nt]}-W_{n-[nt]\pm 1}|R_n^{**}(\mathbf{x})\,,\end{aligned}$$

where $R_n^{**}(\mathbf{x})$ is bounded away from 0. Thus, Lemma 5 implies this tends to infinity in probability.

We can write $P(B_n)$ as

$$P(n^{1+\delta}\inf_{t\in A_n}|\phi_n(t)-\hat{\phi}_n(t)| < n^{1+\delta}\inf_{t\in A_n}|\hat{\phi}_n(t)-\hat{\phi}_n(t\pm n^{-1})|)\,.$$

The first term in the probability tends to zero in probability from Lemmas 3 and 4. The second term tends to infinity in probability from the remarks above. Thus, $P(B_n)$ tends to 1. In like manner, we can prove $P(C_n)$ tends to 1. □

REMARK. The "n^{-1}" in Theorem 4 cannot be improved since $\phi_n(t)$ and $\hat{\phi}_n(t)$ are step functions.

6. A second approximator of t_n. In the last section, we proposed an approximator for t_n using $\hat{\phi}_n(t)$, an approximation of $\phi_n(t)$. Since $\hat{\phi}_n(t)$ involves $\hat{\theta}$, there is some difficulty in computing it. In this section we give another approximation of $\phi_n(t)$ by replacing $\hat{\theta}$ with another estimator $\hat{\theta}^{*}$ of θ. We show (Theorem 5) that if $\hat{\theta}^{*}-\hat{\theta}$ is $O_P(n^{-2})$ then the resulting approximation of t_n will have the property given in Theorem 4. We then describe briefly how estimators of θ with the proper order can be found.

Suppose $\hat{\theta}^{*}$ is an estimator of θ. Define

$$\hat{\phi}_n^{*}(t) = \phi_n(t,\hat{\theta}^{*}) + \gamma_2^{*}(t)n^{-1}\,,$$

where $\gamma_2^{*}(t)$ indicates that $\hat{\theta}$ is replaced by $\hat{\theta}^{*}$ in the expression for $\gamma_2(t)$ given

in the last section. Define

$$\hat{t}_n^* = \inf\{t: \hat{\phi}_n^*(t) \leqq \tfrac{1}{2}\}$$
$$= 1 \quad \text{if} \quad \hat{\phi}_n^*(t) > \tfrac{1}{2} \quad \text{for all} \quad t.$$

THEOREM 5. *Suppose t_0 is in (0, 1) and conditions* A1—A3 *and* A5—A7 *are satisfied. If $\hat{\theta}^*$ is an estimator of θ such that $\hat{\theta}^* - \hat{\theta}$ is $O_{P_{\theta_0}}(n^{-2})$ then*

(i) *for any α, β such that $0 < \alpha < \beta < 1$,*

$$\sup_{\alpha \leqq t \leqq \beta} |\hat{\phi}_n^*(t) - \hat{\phi}_n(t)| = O_{P_{\theta_0}}(n^{-2}),$$

(ii) $P_{\theta_0}[|\hat{t}_n^* - t_n| \leqq 1/n] \to 1$.

PROOF. (i)

$$\sup_{\alpha \leqq t \leqq \beta} |\hat{\phi}_n^*(t) - \hat{\phi}_n(t)| \leqq \sup_{\alpha \leqq t \leqq \beta} \frac{|\hat{\theta} - \hat{\theta}^*|}{(\hat{\theta}^* W_{n-[nt]} + 1 - \hat{\theta}^*)(\hat{\theta} W_{n-[nt]} + 1 - \hat{\theta})}$$
$$+ \sup_{\alpha \leqq t \leqq \beta} n^{-1} |\gamma_2(t) - \gamma_2^*(t)| .$$

The first sup is

$$\leqq \frac{|\hat{\theta} - \hat{\theta}^*|}{(\hat{\theta}^* W_{n-[n\alpha]} + 1 - \hat{\theta}^*)(\hat{\theta} W_{n-[n\alpha]} + 1 - \hat{\theta})} = O_P(n^{-2}) .$$

A lengthy computation gives the second sup

$$\leqq n^{-1} R_n(\mathbf{x}) |\hat{\theta} - \hat{\theta}^*| ,$$

where $R_n(\mathbf{x})$ is $O_P(1)$. Thus, the second sup is $O_P(n^{-2})$.

(ii) follows from the rate in (i) and noting that the proof of Theorem 4 remains valid if $\hat{t}_n$ is replaced by $\hat{t}_n^*$. □

To obtain estimators of the order required in the hypotheses of Theorem 5, we start with an initial estimator of θ and apply the Newton–Raphson iteration method given in Scarborough (1950). We describe the iteration for our special case below.

Suppose $\hat{\theta}^*$ is a strongly consistent estimator of θ. Let

$$l_n(\theta) = \textstyle\sum_{i=1}^n \log f(x_i, \theta)$$

and let $l_n'(\theta)$, $l_n''(\theta)$, and $l_n'''(\theta)$ denote the first, second, and third order partial derivatives of $l_n(\theta)$ with respect to θ. Define

$$T_n(\theta) = \theta - l_n'(\theta)/l_n''(\theta) .$$

We define the ith iteration of $\hat{\theta}^*$ as

$$\hat{\theta}_{in}^* = T_n^i(\hat{\theta}^*) ,$$

where T_n^i indicates T_n composed with itself i times.

The following proposition, which easily follows from several applications of the mean value theorem, shows how this iteration method can be used to get estimators of the rates needed to apply Theorem 5.

PROPOSITION 1. *Suppose condition* A1 *is satisfied. Let* $\hat{\theta}^*$ *be a strongly consistent estimator of* θ *such that* $\hat{\theta}^* - \hat{\theta}$ *is* $O_{P_{\theta_0}}(n^{-\nu})$ *for some* $\nu > 0$. *Then*

$$\hat{\theta}^*_{in} - \hat{\theta} = O_{P_{\theta_0}}(n^{-2i\nu}), \qquad i = 1, 2, \cdots.$$

We now give a class of initial estimators of θ which satisfy the proposition above with $\nu = \frac{1}{2}$. These estimators are considered by Boes (1966).

Let $E_G(\cdot)$ and $E_H(\cdot)$ denote expectation with respect to distributions G and H, let $E_\theta(\cdot)$ denote expectation with respect to F_θ. Suppose $Q(\cdot)$ is a measurable function of x such that $E_G Q(X)^2$ and $E_H Q(X)^2$ are finite. Further suppose that $E_G Q(X) \neq E_H Q(X)$. Then

$$E_\theta Q(X) = \theta E_G Q(X) + (1 - \theta) E_H Q(X).$$

Solving this equation for θ, we get

$$\theta = [E_\theta Q(X) - E_H Q(X)]/[E_G Q(X) - E_H Q(X)].$$

Thus, an obvious estimator of θ is

$$\hat{\theta}_Q = [n^{-1} \textstyle\sum_{i=1}^n Q(X_i) - E_H Q(X)]/[E_G Q(X) - E_H Q(X)].$$

Each estimator derived from some function Q above is strongly consistent for θ and $\hat{\theta}_Q - \hat{\theta}$ is $O_P(n^{-\frac{1}{2}})$. Thus, the second iteration can be used to define $\hat{\phi}_n{}^*(t)$ and the resulting $\hat{t}_n{}^*$.

7. Final remarks. If the prior distribution is degenerate at θ_0 then the Bayes classification rule is

$$a_i{}^* = 1 \quad \text{if} \quad g(x_i)/h(x_i) > \frac{1 - \theta_0}{\theta_0}, \qquad i = 1, \cdots, n.$$

If θ is unknown, a natural classification rule is to use the form of the rule above with θ_0 replaced by an estimate $\hat{\theta}^*$. Rules of this form are considered by Hannan and Robbins (1955). This type of rule makes one cut in the order statistics $W_1, \cdots, W_n$ and classifies to g any X_j whose corresponding W_i is above $(1 - \hat{\theta}^*)/\hat{\theta}^*$. The rule is determined by t_n, the proportion classified to g, where

$$t_n = \inf\left\{t : W_{n-[nt]} \leqq \frac{1 - \hat{\theta}^*}{\hat{\theta}^*}\right\}$$

$$= 1 \quad \text{if} \quad W_{n-[nt]} > \frac{1 - \hat{\theta}^*}{\hat{\theta}^*} \quad \text{for all} \quad t.$$

Using the same techniques as in Sections 3 and 4, we can prove under regularity conditions on g, h, and $\hat{\theta}^*$ that t_n tends to t_0 a.s. given θ_0 and that $n^{\frac{1}{2}}(t_n - t_0)$ is asymptotically normal.

Our techniques in Sections 3 and 4 can also be applied to unknown g and h in a special case. Let D be an interval on the real line and suppose $\{f_\alpha\}_{\alpha \in D}$ is a family of univariate densities with monotone likelihood ratio in x. We observe random variables $X_1, \cdots, X_n$, which conditional on (α, β, θ) are i.i.d. with density $f(x \mid \alpha, \beta, \theta) = \theta f_\alpha(x) + (1 - \theta) f_\beta(x)$. We assume (α, β, θ) is a random vector

with prior distribution Λ. If we assume that (α, β) and θ are independent and that $P_\Lambda(\alpha < \beta) = 1$ then the Bayes rule will have the cut-type property with respect to the order statistics from $X_1, \cdots, X_n$. That is, the rule makes exactly one cut in the ordered values of $X_1, \cdots, X_n$, and is completely determined by the proportion of observations classified to f_α.

Under additional assumptions including an identifiability assumption on the 2-component mixtures over the family $\{f_\alpha\}$, the limiting form of the rule can be found. Under very strong conditions which imply the consistency and asymptotic normality of the joint maximum likelihood estimator of (α, β, θ), the asymptotic normality of $n^{\frac{1}{2}}(t_n - t_0)$ can be proved from the results of Chao (1970).

8. Acknowledgment. The author wishes to thank Professor Michael Woodroofe for his guidance and encouragement during this research.

REFERENCES

[1] ANDERSON, T. W. (1958). *An Introduction to Multivariate Analysis.* Wiley, New York.
[2] BAHADUR. R. R. (1966). A note on quantiles in large samples. *Ann. Math. Statist.* **37** 577–580.
[3] BOES, D. C. (1966). On the estimation of mixing distributions. *Ann. Math. Statist.* **37** 177–188.
[4] BREIMAN, L. (1968). *Probability.* Addison-Wesley, Reading.
[5] CHAO, M. T. (1970). The asymptotic behavior of Bayes estimators. *Ann. Math. Statist.* **41** 601–608.
[6] HANNAN, J. and ROBBINS, H. (1955). Asymptotic solutions of the compound decision problem for two completely specified distributions. *Ann. Math. Statist.* **26** 37–51.
[7] JOHNSON, R. A. (1970). Asymptotic expansions associated with posterior distributions. *Ann. Math. Statist.* **41** 851–864.
[8] RUBIN, H. (1956). Uniform convergence of random functions with applications to statistics. *Ann. Math. Statist.* **27** 200–203.
[9] SCARBOROUGH, J. B. (1950). *Numerical Mathematical Analysis.* The Johns Hopkins Press, Baltimore.

DEPARTMENT OF STATISTICS & PROBABILITY
MICHIGAN STATE UNIVERSITY
EAST LANSING, MICHIGAN 48823