The Annals of Statistics
1978, Vol. 6, No. 1. 117-131

ASYMPTOTIC DISTRIBUTIONS FOR CLUSTERING CRITERIA

BY J. A. HARTIGAN[1]

Yale University

A set of observations is partitioned into k clusters by optimizing a clustering criterion W. The asymptotic distribution of this clustering criterion may be determined simply in certain cases where the optimal sample partition differs negligibly from the optimal population partition. Detailed proofs are given in the one-dimensional case when the clustering criterion to be minimized is within cluster sum of squares. The asymptotic distributions are used to compute approximate significance levels of tests for the presence of clusters, and of tests for bimodality.

1. Introduction. Let $x_1, x_2, \cdots, x_n$ be observations from some distribution function F. Suppose the observations are divided into two groups to maximize the F-ratio for differences between the groups, or equivalently to minimize the within group sum of squares. The division will be specified by a split point s_n, such that observations less than s_n are in one group and observations not less than s_n are in the other group. For very general F, the asymptotic distribution of the maximum F-ratio, $F_{\max}$, and the optimal split point is shown to be normal. Finding the asymptotic distribution of $F_{\max}$ is much simplified by noting that the F-ratio changes negligibly if the sample is split at s_0, where s_0 is the optimal split point for the population. The results are generalised to optimal division of n observations into k groups.

More generally in p dimensions, observations $x_1, \cdots, x_n$ are divided into k groups with means $y_1, y_2, \cdots, y_k$, so as to minimize the within group sum of squares

$$W_n = \sum_i \inf_{1 \leqq j \leqq k} ||x_i - y_j||^2 .$$

By analogy with the one-dimensional case, it frequently happens that W_n differs negligibly from

$$W_0 = \sum_i \inf_{1 \leqq j \leqq k} ||x_i - \mu_j||^2$$

where μ_j are the group means in the optimal partition of the population. Unfortunately, this simplification does not occur for spherically symmetric distributions. Some examples are given for $k = 2$.

The above technique for division of a sample into k groups is known in the cluster analysis literature as k-means; there are a variety of techniques for finding approximations to the optimum partition; Fisher (1958) uses a dynamic programming technique for finding the exact optimum in $O(n^2k)$ steps for $p = 1$,

Received March 1976; revised July 1977.

[1] This research was sponsored by the National Science Foundation under Contract No. DCR75-08374.

AMS 1970 *subject classification.* Primary 62H30.

Key words and phrases. Clustering criteria, tests for bimodality, asymptotic distributions.

and for $k = 2$, the optimal hyperplane dividing the sample into 2 groups may be found in $O(n^p)$ steps. MacQueen (1967) has the following asymptotic result: beginning with arbitrary cluster means $y_1, y_2, \cdots, y_k$, each observation x_i is assigned to whichever of $y_1, y_2, \cdots, y_k$ it is closest, and the chosen cluster centre is modified to be the mean of observations assigned to it. The quantity $\sum ||x_i - y_{j_i}||^2/n$ converges with probability one to

$$\int \inf_{1 \leqq j \leqq k} ||x - \mu_j||^2 \, dF ,$$

where x_i is assigned to y_{j_i} and $\{\mu_j\}$ minimizes the integral.

The one-dimensional results have direct application to clustering in p dimensions. Given the clusters, an analysis of variance may be performed in each dimension; the asymptotic distributions in the one-dimensional case then provide a conservative test for significance of the F-ratios of each of these analyses of variance. See also Scott and Knott (1974) who use conjectured one-dimensional asymptotics for grouping means in analysis of variance. For $k = 2$, the set of n observations may be orthogonally projected onto the line between the two cluster means; the maximal F-ratio for these projected values offers a test for bimodality, using the asymptotic normal as a reference distribution. This test is used in Henderson et al. (1977). Note, however, Day (1969) who shows that the projection technique may be misleading for relatively few observations in many dimensions. The histogram of the projected values is helpful in suggesting the presence of two clusters, and similarly for $k = 3$, the projection onto the plane containing the three cluster centres may suggest the presence or absence of three clusters. For this reason, asymptotic distributions for 3 clusters in 2 dimensions should be the next step.

2. Terminology. The *quantile function* X for the distribution function F is defined by

$$X(p) = \sup \{x \mid F(x) \leqq p\} , \qquad 0 \leqq p \leqq 1 .$$

The function X is nondecreasing and right continuous. If U is uniformly distributed over (0, 1), then $X(U)$ has distribution function F, so that X may be regarded as a canonical random variable with distribution function F.

Define the *lower and upper means of X at p by*

$$\underline{X}(p) = \int_{q \leqq p} X(q) \, dq/p , \qquad 0 < p \leqq 1 ,$$
$$\bar{X}(p) = \int_{q > p} X(q) \, dq/(1 - p) , \qquad 0 \leqq p < 1 .$$

The *split function of X at p* is

$$B(X, p) = p\underline{X}^2(p) + (1 - p)\bar{X}^2(p) - \left(\int_0^1 X(q) \, dq\right)^2 , \qquad 0 < p < 1 .$$

For the case of two clusters in one dimension it is easier to work with B, which corresponds to a between cluster sum of squares, rather than with W used in the introduction, which corresponds to a within cluster sum of squares.

If $X^-(p) = \lim_{q \uparrow p} X(q)$, $X^+(p) = X(p) = \lim_{q \downarrow p} X(q)$, then the upper and

lower derivatives of B at p are given by

$$\frac{d}{dp} B^{\pm}(X, p) = \{\bar{X}(p) - \underline{X}(p)\}\{\bar{X}(p) + \underline{X}(p) - 2X^{\pm}(p)\} . \tag{1}$$

A value p_0 which maximizes $B(X, p)$ is called a *split point*. If X has finite variance,

$$\lim_{p\to 0} B(X, p) = \lim_{p\to 1} B(X, p) = 0$$

and a split point p_0, $0 < p_0 < 1$ exists, and satisfies

$$X^{-}(p_0) \leqq [\underline{X}(p_0) + \bar{X}(p_0)]/2 \leqq X^{+}(p_0) .$$

Thus for an optimal split into two clusters, the boundary of the clusters must lie half way between the cluster means; otherwise the within cluster sum of squares could be reduced by shifting the boundary towards the further mean. The distribution function $F(x) = \frac{1}{2} + \frac{1}{2}x/(1 + x^2)^{\frac{1}{2}}$ has the interesting property that $B(X, p) = 1$ all p, so that all values of p, $0 < p < 1$, are split points. If x_0 is an atom of X, so that $X(p) = x_0$ in some interval $[p_1, p_2)$, then p_0 is not an interior point of the interval (p_1, p_2) because $\underline{X}(p) + \bar{X}(p) - 2X^{\pm}(p) = \underline{X}(p) + \bar{X}(p) - 2x_0$ is strictly increasing in the interval, whereas it must be nonincreasing at a maximum.

If $U_1, \cdots, U_n, \cdots$ is a sequence of independent uniforms, then $X(U_1), \cdots, X(U_n), \cdots$ is a sequence of observations from F where X is the quantile function of F. The *empirical random variable* X_n is the quantile function of the empirical distribution function F_n of the sample $X(U_1), \cdots, X(U_n)$. If $U_{(1)}, \cdots, U_{(n)}$ denote the order statistics of $U_1, \cdots, U_n$ then

$$X_n(p) = X[U_{(i)}] \qquad \text{for} \quad \frac{i-1}{n} \leqq p < \frac{i}{n} , \qquad 1 \leqq i \leqq n .$$

Note that X_n is a function of the random variables $U_1, \cdots, U_n$.

The *sample split function* $B(X_n, p)$ has its maximum at the *sample split point* p_n. Since X_n is carried by the n atoms $X(U_1), \cdots, X(U_n)$, so that $X_n(p)$ is constant in $[i - 1/n, i/n)$, the only possible values of p_n are the jump-points of X_n, namely i/n, $1 \leqq i \leqq n$. Thus p_n may be determined by n computations of $B(X_n, p)$. The quantity $nB(X_n, i/n)$ is the between group sum of squares for the first i observations against the remaining $(n - i)$ observations. Similarly $(n - 2)R(X_n, i/n)$ is the F-ratio for the difference between the means of the first i observations against the remaining $(n - i)$. The principal results give the asymptotic behaviour of p_n, $B(X_n, p_n)$ and $R(X_n, p_n)$.

All asymptotic calculations will use the underlying probability space of the sequence of independent uniforms $U_1, U_2, \cdots, U_n, \cdots$. The notation $[P, n]$ will be used as a shorthand for "in probability as n approaches ∞." Thus $Y_n \leqq 1[P, n]$ means

$$\lim_{n\to\infty} P[Y_n \leqq 1] = 1 .$$

The notation $\Rightarrow$ denotes convergence in distribution. The notation $Y_n = \Delta$

means that Y_n is a function of $U_1, \cdots, U_n$ and p which converges to zero in probability, uniformly over p in a neighbourhood of a fixed p_0: that is, for each $\varepsilon > 0$, $\exists\, \delta$, N such that

$$P[\sup_{|p-p_0|<\delta} |Y_n(p)| > \varepsilon] < \varepsilon \qquad \text{for} \quad n > N. \tag{2}$$

(Here P is outer measure whenever $\sup_{|p-p_0|<\delta} |Y_n(p)| > \varepsilon$ is not measurable.)

For a partition into k cells, specified by cutpoints $0 = p^0 \leqq p^1 \leqq \cdots \leqq p^k = 1$, the *partition function* is

$$B(X, \mathbf{p}) = \sum_{i=1}^{k} (\int_{p^{i-1}}^{p^i} X(p)\, dp)^2/(p^i - p^{i-1}) - (\int_0^1 X(p)\, dp)^2 .$$

If $B(X, \mathbf{p})$ is maximized at $\mathbf{p} = \mathbf{p}_0$, and $E_i = (p_0^{i-1}, p_0^i]$, $1 \leqq i \leqq k$, set

$$\begin{aligned} q_i &= \int_{E_i} dp \\ \mu_i &= \mathscr{E}[X \mid E_i] = \int_{E_i} X(p)\, dp/q_i \\ \sigma_i^2 &= \mathscr{E}[(X - \mu_i)^2 \mid E_i] \\ \nu_i &= \mathscr{E}[(X - \mu_i)^3 \mid E_i] \\ \tau_i &= \mathscr{E}[(X - \mu_i)^4 \mid E_i] . \end{aligned} \tag{3}$$

Define the *ratio function*

$$R(X, \mathbf{p}) = B(X, \mathbf{p})/\{\int X^2(p)\, dp - B(X, \mathbf{p}) - (\int_0^1 X(p)\, dp)^2\} .$$

The following formulae arise in the asymptotic distributions of the sample versions of B and R, with $\mu = \int X(p)\, dp$.

$$\begin{aligned} \mu_B &= \sum q_i(\mu_i - \mu)^2 \\ \mu_R &= \sum q_i(\mu_i - \mu)^2/\sum q_i\sigma_i^2 \\ \sigma_B^2 &= \sum q_i(\mu_i - \mu)^4 - \mu_B^2 + 4 \sum q_i(\mu_i - \mu)^2\sigma_i^2 \\ \sigma_R^2 &= \sum q_i[(\mu_i - \mu)^2 - \mu_R\sigma_i^2]^2\mu_R^2/\mu_B^2 + \sum 4q_i(\mu_i - \mu)^2\sigma_i^2\mu_R^2/\mu_B^2 \\ &\quad - \sum 4q_i(\mu_i - \mu)\nu_i\mu_R^3/\mu_B^2 + \sum q_i(\tau_i - \sigma_i^4)\mu_R^4/\mu_B^2 . \end{aligned} \tag{4}$$

3. Asymptotic behaviour of $B(X_n, p)$.

THEOREM 1. *Suppose X has finite variance and a unique split point p_0. Let p_n be a sequence of sample split points. Then $p_n \to p_0[P, n]$.*

PROOF. The quantities $B(X_n, p)$ will be shown to be uniformly small for p near 0 and 1, and uniformly continuous in probability in intervals not including 0 or 1. For each fixed p, $B(X_n, p)$ converges to $B(X, p)$ in probability. The convergence of p_n to p_0 then follows by a standard compactness argument.

In order to connect the empirical random variable to the empirical distribution function, it is convenient to define $d(U) = 1$ if $U \leqq p$, $d(U) = 0$ otherwise, $r_n = \sum d(U_i)/n$. If $U_{(1)}, \cdots, U_{(n)}$ are the ordered U_i, $U_{(nr_n)} \leqq p < U_{(nr_n+1)}$. Since X is nondecreasing, $X_n(r_n - 1/n) = X(U_{(nr_n)}) \leqq X(p)$ and $X_n(r_n) = X(U_{(nr_n+1)}) \geqq X(p)$. By the law of large numbers, $p - \varepsilon \leqq r_n \leqq p + \varepsilon[P, n]$ for

each $\varepsilon > 0$. Therefore for each $\varepsilon > 0$,

$$X_n(p - \varepsilon) \leqq X(p) \leqq X_n(p + \varepsilon)[P, n]\,. \tag{5}$$

It follows that $X_n(p) \to X(p)[P, n]$ at points of continuity of X, as is well known (Rao (1973), page 423). Much more detailed analyses of the deviations $X_n(p) - X(p)$ are given by Kiefer (1967).

To show convergence of $\underset{\sim}{X}_n(p)$, using (5),

$$\begin{aligned}
\textstyle\int_0^p X_n(q)\,dq &= \textstyle\int_0^{r_n} X_n(q)\,dq + \int_{r_n}^p X_n(q)\,dq \\
|\textstyle\int_{r_n}^p X_n(q)\,dq| &\leqq (|X_n(p)| + |X_n(r_n)|)|p - r_n| \leqq (|X(p+\varepsilon)| + |X(p-\varepsilon)|)\varepsilon[P, n] \\
\textstyle\int_0^{r_n} X_n(q)\,dq &= \textstyle\sum X(U_i)\,d(U_i)/n \to \int_{q \leq p} X(q)\,dq[P, n]\,.
\end{aligned}$$

Thus $\underset{\sim}{X}_n(p) \to \underset{\sim}{X}(p)[P, n]$ whenever it exists.

To show that $B(X_n, p)$ is negligible for p near 0 or 1,

$$\begin{aligned}
B(X_n, p) &= p[\underset{\sim}{X}_n(p) - \bar{X}_n(0)]^2 + (1 - p)[\bar{X}_n(p) - \bar{X}_n(0)]^2 \\
&\leqq \textstyle\int_0^p (X_n(q) - \bar{X}_n(0))^2\,dq + [\bar{X}_n(p) - \bar{X}_n(0)]^2\,.
\end{aligned}$$

The quantity on the right is nondecreasing as a function of p, so

$$\begin{aligned}
\sup_{p \leq r} B(X_n, p) &\leqq \textstyle\int_0^r [X_n(p) - \bar{X}_n(0)]^2\,dp + (\bar{X}_n(r) - \bar{X}_n(0))^2 \\
&\to \textstyle\int_0^r [X(p) - \bar{X}(0)]^2\,dp + (\bar{X}(r) - \bar{X}(0))^2[P, n]\,.
\end{aligned}$$

Since X has finite variance, the quantity on the right approaches 0 as $r \to 0$, so for each $\varepsilon > 0$,

$$\sup_{p \leq r} B(X_n, p) < \varepsilon[P, n] \quad \text{and} \quad \sup_{p \geq 1-r} B(X_n, p) < \varepsilon[P, n] \quad \text{for some} \quad r\,.$$

To show that $B(X_n, p)$ is uniformly continuous, using (1),

$$|B(X, p) - B(X, q)| \leqq \delta K(r) \qquad \text{for} \quad r < p, q < 1 - r, |p - q| < \delta$$

where

$$K(r) = 4(\bar{X}(1 - r) - \underset{\sim}{X}(r))(|\underset{\sim}{X}(r)| + |\bar{X}(1 - r)| + |X(r)| + |X(1 - r)|)\,.$$

Using the convergence of $X_n(p)$, $\underset{\sim}{X}_n(p)$, $\bar{X}_n(p)$, if r and $(1 - r)$ are points of continuity of X,

$$|B(X_n, p) - B(X_n, q)| \leqq 2\delta K(r) \qquad \text{for} \quad r < p, q < 1 - r, |p - q| < \delta[P, n]\,.$$

For each $\varepsilon > 0$, choose r, δ such that

$$\begin{aligned}
B(X, p), B(X_n, p) &< \varepsilon \qquad \text{for} \quad p \leqq r, \quad \text{or} \quad p \geqq 1 - r[P, n]\,; \\
|B(X_n, p) - B(X_n, q)| &< \varepsilon \qquad \text{for} \quad r < p, q < 1 - r, |p - q| \leqq \delta[P, n]\,; \\
|B(X, p) - B(X, q)| &< \varepsilon \qquad \text{for} \quad r < p, q < 1 - r, |p - q| \leqq \delta[P, n]\,; \\
B(X_n, p) &< B(X, p) + \varepsilon \qquad \text{for} \quad p = i\delta, 1 \leqq i \leqq 1/\delta[P, n]\,.
\end{aligned}$$

It follows that $B(X_n, p) < B(X, p) + 3\varepsilon$ for $0 \leqq p \leqq 1[P, n]$. Since $B(X, p)$ is continuous and has a unique maximum at $p = p_0$, for each $\delta_0 > 0$, $\exists\, \varepsilon > 0$ such that $B(X, p) < B(X, p_0) - 4\varepsilon$ for $|p - p_0| \geqq \delta_0$. Thus $B(X_n, p) < B(X, p_0) - \varepsilon$

for $|p - p_0| \geqq \delta_0[P, n]$. Since $B(X_n, p_0) > B(X, p_0) - \varepsilon[P, n]$, $B(X_n, p) < B(X_n, p_0)[P, n]$ for $|p - p_0| \geqq \delta_0$. It follows that $|p_n - p_0| < \delta_0[P, n]$ for each $\delta_0 > 0$, proving Theorem 1. □

THEOREM 2. *Suppose that X has finite fourth moment, and that B has a unique maximum p_0. Assume that X has a continuous derivative in the neighbourhood of p_0, and that $B_0 = [(d^2/dp^2)B(X, p)]_{p=p_0} < 0$. Then*

$$n^{\frac{1}{2}}(p_n - p_0) \Rightarrow N\Big(0, p_0(1 - p_0) + (\sigma_1^2/q_1 + \sigma_2^2/q^2)\Big/\Big[\frac{\mu_2 - \mu_1}{2q_1q_2} - 2X'(p_0)\Big]^2\Big) \tag{6}$$

$$n^{\frac{1}{2}}(B(X_n, p_n) - B(X, p_0)) \Rightarrow N(0, \sigma_B^2)$$

$$n^{\frac{1}{2}}(R(X_n, p_n) - R(X, p_0)) \Rightarrow N(0, \sigma_R^2)$$

where μ_1, μ_2, σ_1, σ_2, q_1, q_2, σ_B^2, σ_R^2 are given in (3) *and* (4).

PROOF. The method of proof approximates $B(X_n, p)$ by a parabola in the neighbourhood of p_0. The value of p which maximizes this parabola differs negligibly from p_n. The value of B changes negligibly for p within $O_p(n^{-\frac{1}{2}})$ of p_0, and a particular value of p is selected which allows simple asymptotic calculations for B.

Define $Y_n(p) = U_{(i)}$ for $p = (i - 1)/n$, $Y_n(1) = 1$, and let $Y_n(p)$ be linear for $(i - 1)/n \leqq p \leqq i/n$. Then $n^{\frac{1}{2}}(Y_n(p) - p)$ converges weakly to $Z(p) - pZ(1)$ where Z is a Brownian motion; see for example, Billingsley (1968), page 105. For each continuous function W on [0, 1] define

$$H_\delta(W) = \sup_{|p-p_0|<\delta} |W(p) - W(p_0)| .$$

Since H_δ is continuous in the uniform metric on continuous functions W,

$$H_\delta(n^{\frac{1}{2}}(Y_n(p) - p)) \to H_\delta(Z(p) - pZ(1))$$

in distribution. Since the sample paths of Z are continuous with probability one,

$$H_\delta(Z(p) - pZ(1)) \to 0 \qquad \text{in probability as} \quad \delta \to 0 .$$

Thus $H_\delta(n^{\frac{1}{2}}(Y_n(p) - p)) \to 0$ in probability as $\delta \to 0$, $n \to \infty$. Now $|Y_n(p) - U_n(p)| \leqq \max_i |U_{i+1} - U_i| = O_p(\log n/n)$. Thus

$$U_n(p) = U_n(p_0) \dot{+} p - p_0 + \Delta/n^{\frac{1}{2}} ,$$

where Δ denotes uniform convergence to zero near p_0, as defined in (2).

Next consider $X_n(p)$; since X is continuously differentiable near p_0,

$$X_n(p) = X(U_n(p)) = X(U_n(p_0)) + X'(p_n^*)(p - p_0 + \Delta/n^{\frac{1}{2}}) .$$

where p_n^* lies between $U_n(p)$ and $U_n(p_0)$. Since $U_n(p_0) \to p_0[P, n]$, and $U_n(p) = U_n(p_0) + \Delta$, $p_n^* = p_0 + \Delta$. Therefore

$$X_n(p) = X_n(p_0) + \frac{\Delta}{n^{\frac{1}{2}}} + (X'(p_0) + \Delta)(p - p_0) . \tag{7}$$

Integration of this equation, justified by the uniform convergence of Δ, gives

$$\begin{aligned}
\underline{X}_n(p) &= \underline{X}_n(p_0) + \{(X_n(p_0) - \underline{X}_n(p_0))/p_0 + \Delta\}(p - p_0) \\
\bar{X}_n(p) &= \bar{X}_n(p_0) + \{(\bar{X}_n(p_0) - X_n(p_0))/(1 - p_0) + \Delta\}(p - p_0) \\
\frac{dB^{\pm}(X_n, p)}{dp} &= (\bar{X}_n(p) - \underline{X}_n(p))(\bar{X}_n(p) + \underline{X}_n(p) - 2X_n(p)) \\
&= \frac{dB^{\pm}(X_n, p_0)}{dp_0} + \frac{\Delta}{n^{\frac{1}{2}}} \\
&\quad + (p - p_0)(\bar{X}_n(p_0) - \underline{X}_n(p_0))\{(X_n(p_0) - \underline{X}_n(p_0))/p_0 \\
&\quad - (X_n(p_0) - \bar{X}_n(p_0))/(1 - p_0) - 2X'(p_0) + \Delta\} \\
&\quad + (p - p_0)\{-(X_n(p_0) - \underline{X}_n(p_0))/p_0 - (X_n(p_0) \\
&\quad - \bar{X}_n(p_0))/(1 - p_0) + \Delta\} \\
&\quad \times (\bar{X}_n(p_0) + \underline{X}_n(p_0) - 2X_n(p_0)) .
\end{aligned}$$

The term in $(p - p_0)^2$ has been absorbed in one of the Δ's. Using the convergence of $\underline{X}_n(p_0)$, $\bar{X}_n(p_0)$, $X_n(p_0)$, and noting that $\bar{X}(p_0) - \underline{X}(p_0) = 2X(p_0)$ since p_0 maximizes B, reduces the expression to

$$\frac{dB^{\pm}(X_n, p)}{dp} = \frac{dB^{\pm}(X_n, p_0)}{dp_0} + \frac{\Delta}{n^{\frac{1}{2}}} + (p - p_0)(B_0 + \Delta) \tag{8}$$

where

$$\begin{aligned}
B_0 &= \frac{d^2B(X, p_0)}{dp_0^2} \\
&= [\bar{X}(p_0) - \underline{X}(p_0)]\left[\frac{X(p_0) - \underline{X}(p_0)}{p_0} - \frac{X(p_0) - \bar{X}(p_0)}{1 - p_0} - 2X'(p_0)\right],
\end{aligned}$$

$$B_0 = [\bar{X}(p_0) - \underline{X}(p_0)]\left[\frac{\bar{X}(p_0) - \underline{X}(p_0)}{2p_0(1 - p_0)} - 2X'(p_0)\right]. \tag{9}$$

Integration of (8) gives an approximation of B by a parabola near p_0,

$$\begin{aligned}
B(X_n, p) &= B(X_n, p_0) + \left[\frac{dB^{\pm}(X_n, p_0)}{dp_0} + \frac{\Delta}{n^{\frac{1}{2}}}\right](p - p_0) \\
&\quad + \tfrac{1}{2}(B_0 + \Delta)(p - p_0)^2 .
\end{aligned} \tag{10}$$

The sample split point p_n, maximizing $B(X_n, p)$ is therefore the solution of

$$\frac{dB^{\pm}(X_n, p)}{dp} = 0 , \qquad p_n = p_0 - \frac{dB^{\pm}(X_n, p_0)}{dp_0} \Big/ (B_0 + \Delta) + \frac{\Delta}{n^{\frac{1}{2}}} .$$

The asymptotic properties of p_n, $B(X_n, p_n)$ and $R(X_n, p_n)$ are best studied by expansions not about p_0, but about r_n, the proportion of $U_1, \cdots, U_n \leqq p_0$. Let $d(U) = 1$ if $U \leqq p_0$, $d(U) = 0$ if $U > p_0$, so that $r_n = \sum d(U_i)/n$, $\underline{X}_n(r_n) = \sum X(U_i)\, d(U_i)/(nr_n)$, $\bar{X}_n(r_n) = \sum X(U_i)(1 - d(U_i))/(nr_n)$. Since $r_n = 0$ or 1 with positive probability, define $\underline{X}_n(0) = \underline{X}(p_0)$, $\bar{X}_n(1) = \bar{X}(p_0)$. Then

$\mathscr{E}[\underline{X}_n(r_n) \mid r_n] = \mathscr{E}[\sum d(U_i)\underline{X}(p_0) \mid r_n] = \underline{X}(p_0)$, and similarly $\mathscr{E}[\bar{X}_n(r_n) \mid r_n] = \bar{X}(p_0)$, $\mathscr{E}[\bar{X}_n(r_n)\underline{X}_n(r_n) \mid r_n] = \underline{X}(p_0)\bar{X}(p_0)$. Thus r_n, $\underline{X}_n(r_n)$ and $\bar{X}_n(r_n)$ are uncorrelated, which makes them convenient quantities for asymptotic calculations. Setting

$$\begin{aligned} \underline{X}_n(r_n) &= \underline{X}(p_0) + \sum [X(U_i) - \underline{X}(p_0)]\, d(U_i)/np_0 + \Delta/n^{\frac{1}{2}} \\ \bar{X}_n(r_n) &= \bar{X}(p_0) + \sum [X(U_i) - \bar{X}(p_0)][1 - d(U_i)]/n(1 - p_0) + \Delta/n^{\frac{1}{2}}, \end{aligned} \tag{11}$$

since $d(U)$, $[X(U) - \underline{X}(p_0)]\, d(U)$, and $[X(U) - \bar{X}(p_0)][1 - d(U)]$ are uncorrelated, the variables r_n, $\underline{X}_n(r_n)$, $\bar{X}_n(r_n)$ are asymptotically normal with means p_0, $\underline{X}(p_0)$, $\bar{X}(p_0)$ and variances $p_0(1 - p_0)/n$, σ_1^2/np_0, $\sigma_2^2/n(1 - p_0)$ and zero covariances.

The sample split point p_n is the solution of

$$\frac{dB^+(X_n, p)}{dp} \leqq 0 \leqq \frac{dB^-(X_n, p)}{dp},$$

which is obtained from (7) with p_0 replaced by r_n, as

$$p_n = r_n + \frac{\Delta}{n^{\frac{1}{2}}} - \frac{1}{B_0}[\bar{X}_n(r_n) - \underline{X}_n(r_n)][\bar{X}_n(r_n) + \underline{X}_n(r_n) - 2X_n(r_n)]\,. \tag{12}$$

Using (7), since $X_n(r_n - 1/n) \leqq X(p_0) \leqq X_n(r_n)$, $X_n(r_n) = X(p_0) + \Delta/n^{\frac{1}{2}}$. From (11), $\bar{X}_n(r_n) + \underline{X}_n(r_n) - 2X(p_0)$ is asymptotically normal with mean 0 and variance $O(n^{-1})$. Therefore

$$p_n = r_n + \frac{\Delta}{n^{\frac{1}{2}}} - \frac{1}{B_0}[\bar{X}(p_0) - \underline{X}(p_0)][\bar{X}_n(r_n) + \underline{X}_n(r_n) - 2X_n(r_n)]\,.$$

From (11) and (9), p_n is asymptotically normal with mean p_0 and the variance given in (6).

From (10), expanded about r_n,

$$\begin{aligned} B(X_n, p_n) &= B(X_n, r_n) + \left(\frac{dB^{\pm}(X_n, r_n)}{dr_n} + \frac{\Delta}{n^{\frac{1}{2}}}\right)(p_n - r_n) \\ &\quad + \tfrac{1}{2}(B_0 + \Delta)(p_n - r_n)^2 \\ &= B(X_n, r_n) + \frac{\Delta}{n^{\frac{1}{2}}} \end{aligned}$$

$$\begin{aligned} B(X_n, r_n) &= r_n(1 - r_n)(\bar{X}_n(r_n) - \underline{X}_n(r_n))^2 \\ &= B(X, p_0) + (1 - 2p_0)(\bar{X}(p_0) - \underline{X}(p_0))^2(r_n - p_0) \\ &\quad + 2p_0(1 - p_0)(\bar{X}_n(r_n) - \bar{X}(p_0) - \underline{X}_n(r_n) \\ &\quad + \underline{X}(p_0))(\bar{X}(p_0) - \underline{X}(p_0)) + \frac{\Delta}{n^{\frac{1}{2}}}\,. \end{aligned} \tag{13}$$

Thus $B(X_n, r_n)$ is asymptotically normal with mean $B(X, p_0)$ and variance, from (11),

$$\begin{aligned} &[(1 - 2p_0)^2 p_0(1 - p_0)(\bar{X}(p_0) - \underline{X}(p_0))^4 \\ &\qquad + 4p_0^2(1 - p_0)^2(\bar{X}(p_0) - \underline{X}(p_0))^2(\sigma_1^2/p_0 + \sigma_2^2/(1 - p_0))]/n \end{aligned}$$

which reduces after some algebra to σ_B^2/n, given in (3).

To obtain the asymptotic distribution of $R(X_n, p_n)$, define

$$n_1 = nr_n\,, \qquad n_2 = n(1 - r_n)$$

$$\bar{x}_1 = \underline{X}_n(r_n)\,, \qquad \bar{x}_2 = \bar{X}_n(r_n)$$

$$s_1^2 = \sum (X(U_i) - \bar{x}_1)^2\, d(U_i)/nr_n\,, \qquad s_2^2 = \sum (X(U_i) - \bar{x}_2)^2(1 - d(U_i))/n(1 - r_n)\,.$$

$$\begin{aligned} R(X_n, r_n) &= \sum n_i(\bar{x}_i - \mu)^2/\sum n_i s_i^2 + \Delta/n^{\frac{1}{2}} = \sum q_i(\mu_i - \mu)^2/\sum q_i\sigma_i^2 \\ &\quad + \sum (n_i - nq_i)[(\mu_i - \mu)^2 - \sigma_i^2 \sum q_i(\mu_i - \mu)^2/\sum q_i\sigma_i^2]/n \sum q_i\sigma_i^2 \\ &\quad + 2\sum q_i(\mu_i - \mu)(\bar{x}_i - \mu_i)/\sum q_i\sigma_i^2 \\ &\quad - 2\sum q_i(s_i^2 - \sigma_i^2) \sum q_i(\mu_i - \mu)^2/(\sum q_i\sigma_i^2)^2 + \Delta/n^{\frac{1}{2}}\,. \end{aligned}$$

The quantities n_i, $\bar{x}_i$ and s_i^2 are asymptotically normal; n_1 and n_2 have correlation -1, and $\bar{x}_i$ and s_i^2 have covariance ν_i/nq_i but other pairs of variables are asymptotically uncorrelated. Thus $R(X_n, p_n)$ is asymptotically normal with variance σ_R^2/n as given in (4). □

NOTES. The essential result of this theorem is that $B(X_n, p_n)$ is adequately approximated by $B(X_n, r_n)$, the between cluster sum of squares using the population cutpoint $X(p_0)$, which has simple asymptotic behaviour. An interesting novelty is the approximation of $B(X_n, p)$ near p_0 by a parabola, so that $B(X_n, p)$ has a "Taylor series" expansion about p_0 which is valid for p not too close and not too far from p_0; the notion of uniform convergence near p_0, represented by the symbol Δ, is necessary for easy manipulation of the expansions. Using the expansion (13), the error $B(X_n, p_n) - B(X_n, r_n)$ is seen to be a X_1^2 variable with expectation

$$\frac{[\sigma_1^2/q_1 + \sigma_2^2/q_2]q_1q_2}{[4X'(p_0)q_1q_2/(\mu_2 - \mu_1) - 1]n}\,,$$

ignoring terms $o_p(n^{-1})$.

A standard appoximate optimisation technique acts as follows when $p = 1$, $k = 2$. Begin with initial split point p_1, move to p_2 satisfying $X_n(p_2) = \frac{1}{2}\underline{X}_n(p_1) + \frac{1}{2}\bar{X}_n(p_1)$, move to p_3 satisfying $X_n(p_3) = \frac{1}{2}\underline{X}_n(p_2) + \frac{1}{2}\bar{X}_n(p_2)$, and so on until a local maximum $\hat{p}$ satisfies $X_n(\hat{p}) = \frac{1}{2}\underline{X}_n(\hat{p}) + \frac{1}{2}\bar{X}_n(\hat{p})$. It follows from (8) that $\hat{p} - p_n = o_p(n^{-\frac{1}{2}})$, and so $B(X_n, \hat{p}) = B(X_n, p_n) + o_p(n^{-1})$. The local and global maxima thus differ negligibly. From (7), assuming $p_i - p_0 = O_p(n^{-\frac{1}{2}})$,

$$[p_2 - \hat{p}] = \frac{\bar{X}(p_0) - \underline{X}(p_0)}{4p_0(1 - p_0)X'(p_0)}[p_1 - \hat{p}] + o_p(n^{-\frac{1}{2}})\,.$$

The coefficient of $(p_1 - \hat{p})$ is less than 1 because $B_0 < 0$. The convergence of p_i to $\hat{p}$ is thus linear, and $O(\log n)$ iterations will be required to go from the assumed initial accuracy $O_p(n^{-\frac{1}{2}})$ to the sufficient final accuracy $O_p(n^{-1})$.

From (8), a Newton–Raphson step is

$$p_2 - p_1 = \frac{\bar{X}_n(p_1) + \underline{X}_n(p_1) - 2X_n(p_1)}{2X'(p_1) - (\bar{X}_n(p_1) - \underline{X}_n(p_1)/2p_1(1 - p_1))}$$

which goes from $O_p(n^{-\frac{1}{2}})$ to $O_p(n^{-1})$ accuracy *in a single step*. Of course $X'(p_1)$

must be estimated by, say,

$$n^{\frac{1}{2}}\left[X_n\left(p_1+\frac{1}{n^{\frac{1}{2}}}\right)-X_n(p_1)\right].$$

For k clusters in p dimensions, a Newton-Raphson step requires estimation of a $kp \times kp$ matrix, the second derivative of the between cluster sum of squares as a function of the cluster centres; this could be a complex and risky task. In any case, there might be an advantage in using a relaxation technique in which the within cluster mean $\bar{x}_1$ at the first iteration, is replaced by the cluster centre $(1+\alpha)\bar{x}_2 - \alpha\bar{x}_1$ on the second iteration rather than by the within cluster mean $\bar{x}_2$. Here α, $0 \leqq \alpha \leqq 1$, is chosen to speed up convergence.

The conditions in the theorem could perhaps be weakened. The condition that X have finite variance is used only in showing that p_n is bounded away from 0 and 1, which reduces to showing that $pX_n^2(p)$ is uniformly small for p near 0 (with a similar problem for p near 1). It may be that

$$\left(\int_0^p X^\alpha(q)\,dq\right)p^{\alpha/2-1} \downarrow 0 \qquad \text{as} \quad p \downarrow 0,$$

for $1 < \alpha \leqq 2$ is enough to establish this uniform convergence, but I haven't been able to prove it. If X is discontinuous at p_0, the formulae in (6) apply with $X'(p_0)$ replaced by ∞. If $|X(p) - X(q)| < K|p-q|$ for p and q near p_0, the asymptotic normality of p_n may no longer hold but the formulae for R_n and B_n remain valid.

4. The normal case. The statistic $R(X_n, p_n)$ is the likelihood ratio statistic for the null hypothesis that the observations come from a normal distribution, against the alternative hypothesis that each observation comes from one of two normal distributions with different means but the same variance. The normal distribution is therefore a natural null distribution.

Define $\varphi(x) = \exp(-\frac{1}{2}x^2)/2\pi^{\frac{1}{2}}$, $\Phi(x) = \int_{-\infty}^{x}\varphi(u)\,du$, and $\Phi(X(p)) = p$. Then $\underline{X}(p) = -\varphi(X)/p$, $\bar{X}(p) = \varphi(X)/(1-p)$, and $B(X,p) = \varphi^2(X)/p(1-p)$ has a maximum when $2X - \varphi/(1-\Phi) + \varphi/\Phi = 0$. Since $\bar{X}(p) = \varphi(X)/(1-\Phi(X))$ increases with X, $\log P(X \geqq x+h \mid X \geqq x)$ [with derivative $-\varphi(x+h)/(1-\Phi(x+h)) + \varphi(x)/(1-\Phi(x))$] decreases with x whenever $h > 0$. Thus $E[X - x \mid X \geqq x]$ decreases with x. Similarly $E[X - x \mid X \leqq x]$ decreases with x. Therefore $2x - \varphi/(1-\Phi) + \varphi/\Phi = E[x - X \mid X \leqq x] + E[x - X \mid X \geqq x]$ increases with x and so has a unique zero at $x = 0$, $p = \frac{1}{2}$.

The quantities in (2) are $q_1 = q_2 = \frac{1}{2}$, $\mu_1 = -(2/\pi)^{\frac{1}{2}}$, $\mu_2 = (2/\pi)^{\frac{1}{2}}$, $\sigma_1^2 = \sigma_2^2 = (1-2/\pi)$, $\nu_1 = -\nu_2 = (2/\pi)^{\frac{1}{2}}(1-4/\pi)$, $\tau_1 = \tau_2 = 3 - 4/\pi - 12/\pi^2$. From (6),

$$p_n \sim N\left(.5,\ \frac{1}{4n} + \frac{1}{n(2\pi-4)}\right)$$

$$B(X_n, p_n) \sim N\left(\frac{2}{\pi},\ \frac{8}{\pi}\left(1-\frac{2}{\pi}\right)\Big/n\right)$$

$$R(X_n, p_n) \sim N\left(\frac{2}{\pi-2},\ \frac{8}{\pi}\left(1-\frac{3}{\pi}\right)\Big/\left(1-\frac{2}{\pi}\right)^4 n\right).$$

The last two formulae appear in Hartigan (1975), page 98, without proof. Some Monte Carlo evaluations of the distribution of R for small n are given in Engelman and Hartigan (1969), and the asymptotic normality of R is suggested there; more precisely, it is suggested that

$$\log(1 + R) \sim N\left(-\log\left(1 - \frac{2}{\pi}\right) + \frac{2.4}{n-2}, \frac{1}{n-2}\right)$$

by examination of the Monte Carlo results for small n.

As a check on the asymptotic formulae, samples of size $n = 10$ and $n = 100$ were drawn from the normal distribution, and the quantities p_n, $B(X_n, p_n)$, $R(X_n, p_n)$ were computed for 100 such samples. In Table 1 it is seen that the computed means and variances approximate the theoretical ones. Normal plots reveal that the quantities $B(X_n, p_n)$ and $R(X_n, p_n)$ are quite skew for small n, so that the normal approximation is not too safe for $n < 100$. However $B^{\frac{1}{4}}$ and $R^{-\frac{1}{4}}$ were found by experimentation to be quite normal even for $n = 5$.

TABLE 1

Moments of split statistics over 100 *samples from normal parent.*

	p_n		$B(X_n, p_n)$		$R(X_n, p_n)$	
	Mean	$n \times$ Variance	Mean	$n \times$ Variance	Mean	$n \times$ Variance
$n = 10$	.498	.358	.621	.733	2.26	36.7
$n = 100$	.490	.760	.629	.899	1.82	6.42
Asymptotic	.500	.688	.637	.924	1.75	6.74

5. The uniform spike case. If p_0 is fixed, $R(X, p_0)$ is maximal (∞) when X is concentrated on two points with weights p_0 and $1 - p_0$. If clusters are thought to be modelled by modes in the parent population, an appropriate null hypothesis is the unimodal parent population which maximizes $R(X, p_0)$. For $p_0 > \frac{1}{2}$, the optimal distribution has an atom for $p \in [\frac{1}{2}(1 - p_0), \frac{3}{2}p_0 - \frac{1}{2})$ and is uniform elsewhere

$$\begin{aligned} X(p) &= (p + p_0 - 1)/2(1 - p_0) && 0 \leqq p < \tfrac{1}{2}(1 - p_0) \\ X(p) &= -\tfrac{1}{4} && \tfrac{1}{2}(1 - p_0) \leqq p < \tfrac{3}{2}p_0 - \tfrac{1}{2} \\ X(p) &= (p - p_0)/2(1 - p_0) && \tfrac{3}{2}p_0 - \tfrac{1}{2} \leqq p < 1 . \end{aligned}$$

$EX = \frac{1}{4}(1 - 2p_0)$, $\mu_1 = -\frac{1}{2}(1 - p_0)$, $\mu_2 = \frac{1}{2}p_0$, $\sigma_1^2 = (1 - p_0)/48p_0$, $\sigma_2^2 = \frac{1}{48}$, $\nu_1 = \nu_2 = 0$, $\tau_1 = (1 - p_0)/1280p_0$, $\tau_2 = \frac{1}{1280}$, $B(X, p_0) = \frac{1}{4}p_0(1 - p_0)$, $R(X, p_0) = 6p_0$, $\sigma_B^2 = (1 - p_0)(1 + p_0 - 10p_0^2 + 12p_0^3)/48$, $\sigma_R^2 = 12(10 - 20p_0 + 17p_0^2 + 30p_0^3)/10(1 - p_0)$. These formulae hold only for $p_0 \geqq \frac{1}{2}$.

From (6), at $p_0 = \frac{1}{2}$, $R(X_n, p_n)$ has asymptotic mean 3 and variance $19.2/n$ for a uniform parent, compared with asymptotic mean 1.75 and variance $6.6/n$ for a normal parent. Thus the "uniform-spike" based test is much less likely to reject the null hypothesis.

6. Many clusters in one dimension. The partition function $B(X, \mathbf{p})$, corresponding to cut points $0 = p^0 \leqq p^1 \leqq \cdots p^{k-1} \leqq p^k = 1$, is

$$\sum_{i=1}^{k} \left(\int_{p^{i-1}}^{p^i} X(p)\, dp\right)^2/(p^i - p^{i-1}) - \left(\int_0^1 X(p)\, dp\right)^2 .$$

The partition function is maximized by the *optimal cut* $\mathbf{p}_0$, and the sample partition function is maximized by the optimum sample cut $\mathbf{p}_n$.

The generalization of Theorem 1 states that $\mathbf{p}_n \to \mathbf{p}_0[P, n]$ if X has finite variance and unique optimal cut. This is proved analogously to Theorem 1. First if $\mathbf{p}_0$ is unique, then X must be carried by at least k distinct points; otherwise $X(p) = x_i$ for $p^{i-1} \leqq p < p^i$, $0 = p^0 < p^1 \leqq p^2 \cdots \leqq p^{k-1} = 1$. An optimal cut is $(p^0, p, p^1, \cdots, p^{k-1})$ for any p, $p^0 \leqq p \leqq p^1$ which contradicts uniqueness of $\mathbf{p}_0$. Secondly, $B(X, \mathbf{p}_0)$ is larger for k clusters than for j clusters, $j < k$, since any j-cut may be improved to a $(j+1)$-cut with larger B by splitting a cell $[p^{i-1}, p^i)$ on which X is not constant.

Now consider a cut $(p^0, \cdots, p^k)$ in which j_1 of the p^i are less than r and j_2 of the p^i are greater than $(1 - r)$. Since X has finite variance, if r is chosen sufficiently small, the contribution to $B(X_n, \mathbf{p})$ from the $j_1 + j_2$ cut points in the tails is less than $\varepsilon[P, n]$. As a function of the p's between r and $1 - r$, $B(X_n, \mathbf{p})$ satisfies a Lipschitz condition $[P, n]$. Thus $\sup_p B(X_n, \mathbf{p}) < B(X, \mathbf{p}_*) + \varepsilon[P, n]$ where the sup is taken over all $\mathbf{p}$ with j_1 less than r and j_2 greater than $1 - r$, and where $\mathbf{p}_*$ denotes the optimal cut for $(k - j_1 - j_2)$ cut points. Therefore $\sup_p B(X_n, \mathbf{p}) < B(X, \mathbf{p}_0) - \varepsilon[P, n]$, unless $j_1 = j_2 = 0$. If $j_1 = j_2 = 0$, use of the Lipschitz condition on $B(X_n, p)$ shows that $\sup_p B(X_n, \mathbf{p}) < B(X, \mathbf{p}_0) - \varepsilon[P, n]$ where the sup is now taken over all $\mathbf{p}$ outside a neighbourhood of $\mathbf{p}_0$. The concludes the proof.

The generalisation of Theorem 2 states that $\mathbf{p}_n$ and $B(X_n, \mathbf{p}_n)$ and $R(X_n, \mathbf{p}_n)$ are asymptotically normal provided that X has finite variance, that $\mathbf{p}_0$ is unique, that X is continuously differentiable at p_0 and that $\partial^2 B/\partial p_0^2$ is nonnegative definite. The basic technique is to show that B is approximated by a quadratic form, uniformly, close to $\mathbf{p}_0$. Thus

$$\bar{X}^i = \int_{p^{i-1}}^{p^i} X\, dp/(p^i - p^{i-1})$$

$$\frac{dB^{\pm}}{dp^i} = (\bar{X}^{i+1} - \bar{X}^i)(\bar{X}^{i+1} + \bar{X}^i - 2X^{\pm}(p^i)) .$$

At $\mathbf{p} = \mathbf{p}_0$,

$$B_{ii} = \frac{d^2B}{dp^i\, dp^i} = (\bar{X}^{i+1} - \bar{X}^i)\left(\frac{(\bar{X}^{i+1} - \bar{X}^i)(p^{i+1} - p^{i-1})}{2(p^{i+1} - p^i)(p^i - p^{i-1})} - 2X'(p^i)\right)$$

$$B_{ii+1} = \frac{d^2B}{dp^i\, dp^{i+1}} = (\bar{X}^{i+1} - \bar{X}^i)(X(p^{i+1}) - \bar{X}^{i+1})/(p^{i+1} - p^i)$$

$$B_{ij} = \frac{d^2B}{dp^i\, dp^j} = 0 \qquad \text{if} \quad |i - j| > 1 .$$

Straightforward calculation gives the analogues of (8) and (9),

$$\frac{dB^{\pm}(X_n, p^i)}{dp^i} = \frac{dB^{\pm}(X_n, p_0{}^i)}{dp_0{}^i} + \frac{\Delta}{n^{\frac{1}{2}}} + \sum (B_{ij} + \Delta)(p_j - p_j{}^0)\,,$$

(14)
$$1 \leqq i \leqq k-1\,.$$

$$B(X_n, \mathbf{p}) = B(X_n, \mathbf{p}_0) + \sum \left(\frac{dB(X_n, \mathbf{p}_0)}{dp_0{}^i} + \frac{\Delta}{n^{\frac{1}{2}}}\right)(p^i - p_0{}^i) + \tfrac{1}{2}\sum (B_{ij} + \Delta)(p^i - p_0{}^i)(p^j - p_0{}^j)\,.$$

Rather than about $\mathbf{p}_0$, expand about $\mathbf{r}_n$ where $nr_n{}^i$ is the number of observations not exceeding $X(p_0{}^i)$, $\bar{x}_i$ is the mean of observations between $X(p_0{}^{i-1})$ and $X(p_0{}^i)$, and $n_i := nr_n{}^i - nr_n{}^{i-1}$ is the number of observations between $X(p_0{}^{i-1})$ and $X(p_0{}^i)$. Then n_i is multinomial with parameters n, $q_1, \cdots, q_k$ and the $\bar{X}^i$ are asymptotically independent normal with means μ_i and variances $\sigma_i{}^2/nq_i$, as specified in (2). Then $\mathbf{p}_n$ is asymptotically normal with a complicated expression for the variance involving B^{-1}. Since $B(X_n, \mathbf{p}_n) = B(X_n, \mathbf{r}_n) + O(n^{-1})$, it is sufficient to consider only $B(X_n, \mathbf{r}_n)$ in determining the asymptotics of $B(X_n, \mathbf{p}_n)$ and $R(X_n, \mathbf{p}_n)$. Thus $B(X_n, \mathbf{r}_n) = 1/n \sum n_i \bar{X}_i{}^2 - (\sum n_i \bar{X}_i)^2/n$ is asymptotically normal with mean $B(X, \mathbf{p}_0) = \sum q_i(\mu_i - \mu)^2$ and variance $\sigma_B{}^2/n$ where $\sigma_B{}^2$ is given in (3). Letting $s_i{}^2$ denote the sample variance of observations between $X(p_0{}^{i-1})$ and $X(p_0{}^i)$, and noting that the $(k+1)$ sets of random variables $\{n_1, \cdots, n_k\}$, $\{\bar{x}_i, s_1{}^2\}, \cdots, \{\bar{x}_k, s_k{}^2\}$ are asymptotically independent, and that $\bar{x}_i$, $s_i{}^2$ have variances $\sigma_i{}^2/q_i n$, $\tau_i/q_i n$ and covariance $\nu_i/q_i n$, then

$$R(X_n, \mathbf{p}_n) = \sum n_i(\bar{X}_i - \bar{X})^2 / \sum n_i s_i{}^2$$

is asymptotically normal with mean $R(X, \mathbf{p}_0)$ and variance $\sigma_R{}^2/n$ where $\sigma_R{}^2$ is given by (3).

There is some interest in estimating the term $B(X_n, \mathbf{p}_n) - B(X_n, \mathbf{r}_n)$, the increase in B due to selecting the optimum sample cut, rather than dividing the sample by the population cutpoints $X(p_0{}^i)$. Choosing $\mathbf{p}_n$ to maximize (14) expanded about $\mathbf{r}_n$, the asymptotic expectation of $B(X_n, \mathbf{p}_n) - B(X_n, \mathbf{r}_n)$ is $-1/2n \operatorname{tr}(CD^{-1})$ where C and D are $k-1 \times k-1$ symmetric tridiagonal matrices with

$$\begin{aligned} c_{ii} &= \sigma_i{}^2/q_i + \sigma_{i+1}^2/q_{i+1}\,, \\ c_{ii+1} &= \sigma_{i+1}^2/q_{i+1} \\ d_{ii} &= \tfrac{1}{2}\left(\frac{1}{q_i} + \frac{1}{q_{i+1}}\right) - 2X'(p_0{}^i)/(\mu_{i+1} - \mu_i) \\ d_{ii+1} &= 1/(2q_{i+1})\,. \end{aligned}$$

For example, in the uniform case, $c_{ii} = 1/6k$, $c_{ii+1} = 1/12k$, $d_{ii} = -k$, $d_{ii+1} = k/2$ so that the expectation is $(1/n)\frac{1}{3}(k-1)(2k-1)/12k^2$, just half of the expression in Marriott (1970, 1971). (In Marriott's expansion analogous to (14), the second order term is neglected, although in the neighbourhood of the maximum of B it is half the size of the first order term and of opposite sign.)

7. Generalization to p dimensions. In p dimensions, observations $x_1, \cdots, x_n$ from a p-dimensional variable X are divided into k groups with means $y_1, \cdots, y_k$ to minimize the within group sum of squares

$$W_n(\mathbf{y}) = \sum_i \inf_{1 \leq j \leq k} ||x_i - y_j||^2 .$$

By analogy with the one-dimensional case, one expects that $W_n(\mathbf{y})$ would be closely approximated by $W_n(\boldsymbol{\mu})$ where $\boldsymbol{\mu}$ are the population means chosen to minimize

$$E(\inf_{1 \leq j \leq k} ||x - \mu_j||^2) .$$

Since $W_n(\boldsymbol{\mu})$ is the sum of i.i.d. random variables, it will be normal provided X has finite fourth moment. For example if $k = 2$, and if X is from a p-dimensional normal with mean 0 and diagonal covariance matrix with variances $\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2$, where $\sigma_1^2 > \sigma_2^2, \cdots, \sigma_p^2$, then it is conjectured that the optimal split takes place along the first dimension, and so asymptotically

$$W_n(\mathbf{y}) \sim N(\sum \sigma_i^2 - 2\sigma_1^2/\pi, 2 \sum \sigma_i^4/n - 16n\sigma_1^4/n\pi^2) .$$

Expressions for F-ratios are given in Hartigan (1975), page 100.

On the other hand, if the optimal $\boldsymbol{\mu}$ are not unique, as when sampling from a bivariate circular normal, $k = 2$, the approximation by $W_n(\boldsymbol{\mu})$ may no longer be valid; in this case $W_n(\mathbf{y})$ is conjectured to be asymptotically distributed as the minimum of the normal process $Z(\theta)$ on the circle $0 \leqq \theta < 2\pi$, where $Z(\theta)$ has mean $(2 - 2/\pi)$ and covariance with $Z(\varphi)$,

$$\left\{4 - \frac{8}{\pi} - \frac{16}{\pi^2} + \frac{16}{\pi^2}\left(\sin\alpha - \left(\alpha - \frac{\pi}{2}\right)\cos\alpha\right)\right\}\Big/n , \quad \alpha = |\theta - \varphi| \leqq \pi .$$

Much work remains to be done in p-dimensions.

Acknowledgments. I am indebted to David Pollard for suggesting the weak convergence argument in the proof of Theorem 2.

REFERENCES

[1] BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

[2] DAY, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika* **56** 463–474.

[3] ENGELMAN, L. and HARTIGAN, J. A. (1969). Percentage points of a test for clusters. *J. Amer. Statist. Assoc.* **64** 1647–1648.

[4] FISHER, W. D. (1958). On grouping for maximum homogeneity. *J. Amer. Statist. Assoc.* **53** 789–798.

[5] HARTIGAN, J. A. (1975). *Clustering Algorithms.* Wiley, New York.

[6] HENDERSON, A. S., HARTIGAN, J., DAVIDSON, J., LANCE, G. N., DUNCAN-JONES, P., KOLLER, K. M., RITCHIE, KAREN, MCAULEY, HELEN, WILLIAMS, C. L. and SLAGHUIS, W. (1977). A typology of parasuicide. *Brit. J. Psychiat.* **130**.

[7] KIEFER, J. (1967). On Bahadur's representation of sample quantiles. *Ann. Math. Statist.* **38** 1323–1341.

[8] MACQUEEN, J. (1967). Some methods for classification and analysis of multivariate observations. *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* **1** 281–297.

[9] MARRIOTT, F. H. C. (1970). A problem of optimum stratification. *Biometrics* **26** 845–847.

[10] MARRIOTT, F. H. C. (1971). Practical problems in a method of cluster analysis. *Biometrics* **27** 501–514.

[11] RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*. Wiley, New York.

[12] SCOTT, A. J. and KNOTT, M. (1974). A cluster analysis method for grouping means in analysis of variance. *Biometrics* **30** 507–512.

DEPARTMENT OF STATISTICS
YALE UNIVERSITY
NEW HAVEN, CONNECTICUT 06520