

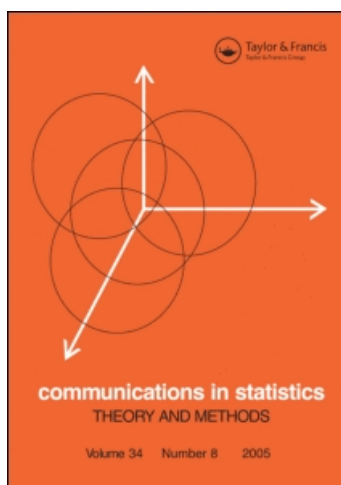
This article was downloaded by: [Duke University]

On: 12 November 2008

Access details: Access Details: [subscription number 731831809]

Publisher Taylor & Francis

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Communications in Statistics - Theory and Methods

Publication details, including instructions for authors and subscription information:

<http://www.informaworld.com/smpp/title~content=t713597238>

An optimal variable cell histogram

Yuichiro Kanazawa^a

^a Department of Statistics, Yale University, New Haven, CT

Online Publication Date: 01 January 1988

To cite this Article Kanazawa, Yuichiro(1988)'An optimal variable cell histogram',Communications in Statistics - Theory and Methods,17:5,1401 — 1422

To link to this Article: DOI: 10.1080/03610928808829688

URL: <http://dx.doi.org/10.1080/03610928808829688>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

An Optimal Variable Cell Histogram

Yuichiro Kanazawa

Department of Statistics
Yale University
New Haven, CT 06520-2179

Key Words and Phrases: density estimation; Hellinger distance; histogram.

ABSTRACT

A simple procedure for specifying a histogram with variable cell sizes is proposed. The procedure chooses a set of cutpoints that maximizes a criterion function based on the sample spacings. Under some conditions, this estimated set of cutpoints is shown to converge in probability to the theoretical set of cutpoints for the histogram estimate that minimizes the Hellinger distance to the underlying density. An algorithm for finding the set of cutpoints that numerically maximizes the criterion function is presented along with an example. Performance for finite sample sizes is evaluated by simulations.

1. INTRODUCTION

Procedures for choosing a variable cell histogram have received relatively little attention. Let X_1, \dots, X_n be independent random variables having common unknown density function $f(x)$ with a support I . Scott (1979) showed that the cell size,

$$h_n = 6^{1/3} \left[\int_I f'(x)^2 dx \right]^{-1/3} n^{-1/3},$$

minimizes the integrated mean squared error (*IMSE*) asymptotically for $f(x)$. This size, however, depends on $\int_I f'(x)^2 dx$, which in general is unknown. He proposed $3.49sn^{-1/3}$ as a cell size where s is an estimate of the standard deviation. This cell size is derived from solving h_n for the normal distribution with variance σ^2 . The difference in shape among densities is only reflected through s . Freedman and Diaconis (1981), having observed the results of numerical computations, suggested that the cell size of $2 \times (\text{interquartile range}) \times n^{-1/3}$ give reasonable results.

Rudemo (1982) proposed a more complicated rule for choosing the cell size of a histogram. Let a be the leftmost cutpoint and b be the cell size. A histogram requires choice of the pair $h = (a, b)$. For integer j , the cell $I_{h,j} = [a + (j-1)b, a + jb)$ has length b . Let $P_n(I_{h,j})$ denote the empirical distribution defined by $P_n(I_{h,j}) = \sum_{i=1}^n \{X_i \in I_{h,j}\} / n$. He proposed that we choose the cell selection rule h that minimizes the criterion function

$$K_n(h) = \frac{1}{b} \left[\frac{2}{n-1} - \frac{n+1}{n-1} \sum_j P_n(I_{h,j})^2 \right].$$

Stone (1984) extended this cell selection rule for a d -dimensional density. He showed that the cell selection rule is asymptotically optimal with respect to *IMSE* for the density with a finite support I satisfying a mild additional condition.

Kogure (1986) strengthened the result derived by Stone in the sense that the faster rate of convergence of $IMSE(n, \hat{k})/IMSE(n, k^*)$ to 1 is obtained where k^* and \hat{k} are the respective minimizer for $IMSE(n, k)$ and $\overline{IMSE}(n, k)$, an unbiased estimate of the $IMSE(n, k)$. He also proved that the global minimum of $IMSE$ is achieved by choosing the cell that minimizes $\overline{IMSE}(n, \cdot)$ over the class of cells Q dividing the interval I ,

$$Q = \left(a + (i-1) \frac{|I|}{M} + (j-1) \frac{|I|}{Mk_i}, \quad a + (i-1) \frac{|I|}{M} + j \frac{|I|}{Mk_i} \right], \\ 1 \leq j \leq k_i, \quad 1 \leq i \leq M.$$

Hence different cell sizes are allowed to some extent.

In this paper, the asymptotic behavior of a simple method for constructing a variable cell-size histogram with a finite number of cells is studied. The method is as follows. Suppose that X_1, \dots, X_n is a random sample, $X_{(1)}, \dots, X_{(n)}$ are corresponding order statistics, and i -th sample spacing T_i is defined as $X_{(i)} - X_{(i-1)}$. The location of the k -cutpoints as well as the heights of the $k-1$ cells are necessary to construct a variable cell-size histogram with $k-1$ cells. We restrict the location of the k -cutpoints:

- a. The k -cutpoints are chosen from $X_{(1)}, \dots, X_{(n)}$.
- b. The leftmost and rightmost cutpoints are restricted to the smallest and largest order statistics respectively.

From the restriction a, the k -cutpoints are written as $(X_{(n_1)}, \dots, X_{(n_k)})$ and $\mathbf{n} = (n_1, \dots, n_k)$ are the indices of $(X_{(n_1)}, \dots, X_{(n_k)})$. From the restriction b, $X_{(n_1)} = X_{(1)}$ and $X_{(n_k)} = X_{(n)}$. Under these restrictions, a variable cell-size histogram with $k-1$ cells is constructed as follows:

Step 1. Find the k -cutpoints $(X_{(n_1^*)}, \dots, X_{(n_k^*)})$ that maximizes the criterion function,

$$C(f_n, \mathbf{n}) = \frac{1}{n+1} \sum_{j=1}^{k-1} \frac{\left[\sum_{i=1+n_j}^{n_{j+1}} T_i^{1/2} \right]^2}{\sum_{i=1+n_j}^{n_{j+1}} T_i}. \quad (1.1)$$

Step 2. Compute the height a_j of the j -th cell by applying formula below to the k -cutpoints $(X_{(n_1^\circ)}, \dots, X_{(n_k^\circ)})$ found in **Step 1**,

$$a_j = \left[\frac{\sum_{i=1+n_j^\circ}^{n_{j+1}^\circ} T_i^{1/2}}{\sum_{i=1+n_j^\circ}^{n_{j+1}^\circ} T_i} \right]^2 / \sum_{j=1}^{k-1} \frac{\left[\sum_{i=1+n_j^\circ}^{n_{j+1}^\circ} T_i^{1/2} \right]^2}{\sum_{i=1+n_j^\circ}^{n_{j+1}^\circ} T_i}. \quad (1.2)$$

The purposes of this paper are:

- To show that $\mathbf{n}^\circ = (n_1^\circ, \dots, n_k^\circ)$ is asymptotically optimal in the sense that the probability limit \mathbf{p}° of $\mathbf{n}^\circ / (n+1)$ corresponds to the set of k -cutpoints of the theoretical histogram estimate $g(\mathbf{x})$ that minimizes the Hellinger distance to the underlying density $f(x)$.
- To present a simple algorithm for numerically finding the k -cutpoints $(X_{(n_1^\circ)}, \dots, X_{(n_k^\circ)})$ that maximizes the criterion function $C(f_n, \mathbf{n})$.

The main result is described in Section 2 and is proved in Section 3. Section 4 has the algorithm for finding the set of k -cutpoints numerically along with an example. Simulation results for finite sample sizes follow in Section 5 and a discussion is presented in Section 6.

2. MAIN RESULT

Theorem Suppose that X_1, \dots, X_n is a random sample, $X_{(1)}, \dots, X_{(n)}$ are corresponding order statistics, and i -th sample spacing T_i is defined as $X_{(i)} - X_{(i-1)}$ from a density function $f(x)$ with finite support $I = [L, M]$. The following conditions on the $f(x)$ are satisfied:

- The inverse of the cumulative distribution function $H(u) = F^{-1}(u)$ exists on $[0, 1]$.
- $H(u)$ is twice differentiable and its derivatives are continuous on $[0, 1]$.
- $0 < m_1 \leq H^{(1)}(u) \leq M_1, \quad 0 \leq u \leq 1.$
- $|H^{(2)}(u)| \leq M_2, \quad 0 \leq u \leq 1.$

A.5. There is a unique set of k -cutpoints $\mathbf{p}^\circ = (p_1^\circ, \dots, p_k^\circ)$ that maximizes

$$P(H, \mathbf{p}) = (\pi/4) \sum_{j=1}^{k-1} \frac{[\int_{p_j}^{p_{j+1}} H^{(1)}(u)^{1/2} du]^2}{\int_{p_j}^{p_{j+1}} H^{(1)}(u) du}$$

subject to the constraint that

$$0 < c \leq p_{j+1}^\circ - p_j^\circ, \quad j = 1, \dots, k-1, \text{ some constant } c.$$

A variable cell-size histogram with $k-1$ cells $f_n(x)$ is constructed by choosing the k -cutpoints from $X_{(1)}, \dots, X_{(n)}$. The leftmost and rightmost cutpoints are restricted to the smallest and largest order statistics respectively. The following condition on the k -cutpoints of $f_n(x)$ is satisfied:

A.6. There is a unique set of k -cutpoints $(X_{(n_1^\circ)}, \dots, X_{(n_k^\circ)})$ that maximizes

$$C(f_n, \mathbf{n}) = \frac{1}{n+1} \sum_{j=1}^{k-1} \frac{[\sum_{i=1+n_j}^{n_{j+1}} T_i^{1/2}]^2}{\sum_{i=1+n_j}^{n_{j+1}} T_i}$$

subject to the constraint that

$$0 < c \leq (n_{j+1}^\circ - n_j^\circ)/(n+1) \leq 1, \quad j = 1, \dots, k-1, \text{ some constant } c.$$

Then:

(1) For the indices $\mathbf{n}^\circ = (n_1^\circ, \dots, n_k^\circ)$ of the k -cutpoints $(X_{(n_1^\circ)}, \dots, X_{(n_k^\circ)})$,

$$\max_{i=n_1^\circ < n_2^\circ < \dots < n_k^\circ = n} |C(f_n, \mathbf{n}^\circ) - P(H, \mathbf{n}^\circ/(n+1))| = O_p(n^{-1/2}).$$

(2) The indices $\mathbf{n}^\circ = (n_1^\circ, \dots, n_k^\circ)$ of the k -cutpoints $(X_{(n_1^\circ)}, \dots, X_{(n_k^\circ)})$ that maximizes $C(f_n, \mathbf{n})$ converge to the k -cutpoints $\mathbf{p}^\circ = (p_1^\circ, \dots, p_k^\circ)$ that maximizes $P(H, \mathbf{p})$ in the sense that $\mathbf{n}^\circ/(n+1) \rightarrow \mathbf{p}^\circ$ in probability as $n \rightarrow \infty$.

(3) Let $g(x)$ be a histogram estimate with k -cutpoints $(L = q_1^\circ, \dots, q_k^\circ = M)$ that minimizes the Hellinger distance, $\int_I [f(x)^{1/2} - g(x)^{1/2}]^2 dx$, to the underlying density $f(x)$ with finite support I . Then for $1 \leq j \leq k$,

$$\int_L^{q_j^\circ} f(x) dx = p_j^\circ.$$

3. PROOF OF THE THEOREM

Proof The proof is given in the following order.

1. Expand the criterion function using Taylor's theorem and express it in terms of exponential random variables E_i .
2. Evaluate the order of magnitude for maximal errors when we approximate $\sum_{i=1+n}^{n_j+1} H^{(1)}\left(\frac{i-1}{n+1}\right)^\alpha / (n+1)$ by $\int_{n_j^\circ/(n+1)}^{n_j^\circ+1/(n+1)} H^{(1)}(u)^\alpha du$ for $\alpha = 1/2, 1$, and 2 .
3. Evaluate the order of magnitude for the differences between principal terms $\sum_{i=1+n}^{n_j+1} E_i^\alpha H^{(1)}\left(\frac{i-1}{n+1}\right)^\alpha / (n+1)$ and their expected values for $\alpha = 1/2$, and 1 . Then evaluate order of magnitude for the remainder terms $\sum_{i=1+n}^{n_j+1} E_i^\alpha R_{\beta i} / (n+1)$ for $\alpha = 1/2, \beta = 1$, and $\alpha = 1, \beta = 2$ using bounds that are independent of the choice of the k -cutpoints $(X_{(n_1^\circ)}, \dots, X_{(n_k^\circ)})$.
4. Evaluate the maximum bound for $|C(f_n, \mathbf{n}^\circ - P(H, \mathbf{n}^\circ / (n+1)))|$.
5. Show that the indices $\mathbf{n}^\circ = (n_1^\circ, \dots, n_k^\circ)$ of the k -cutpoints that maximizes $C(f_n, \mathbf{n})$ converge in probability to $\mathbf{p}^\circ = (p_1^\circ, \dots, p_k^\circ)$ that maximizes $P(H, \mathbf{p})$.
6. Show that the set of k -cutpoints of the histogram estimate $g(x)$ that minimizes the Hellinger distance to the underlying density corresponds to $\mathbf{p}^\circ = (p_1^\circ, \dots, p_k^\circ)$.

3.1 Taylor Expansion of the Criterion Function and Its Expression as Exponentials

Let $U_{(1)}, \dots, U_{(n)}$ be order statistics from the uniform $[0, 1]$; then $U_{(i-1)}$ is Beta($i-1, n-i+2$) with expected value $E(U_{(i-1)}) = (i-1)/(n+1)$. From A.1, $H(u) = F^{-1}(u)$ exists on $[0, 1]$. Set $X_{(i)} = H(U_{(i)})$ for $2 \leq i \leq n$.

It follows from the inverse of the probability integral transformation that $X_{(1)}, \dots, X_{(n)}$ are the order statistics of a sample of n independent random variables with the distribution function $F(x)$. Hence the spacings T_i between $i-1$ th and i th order statistics, $X_{(i-1)}$ and $X_{(i)}$, are

$$T_i = H(U_{(i)}) - H(U_{(i-1)}), \quad 2 \leq i \leq n. \quad (3.1)$$

Since $H(u)$ is twice differentiable and its derivatives are continuous on $[0, 1]$ from A.2, we obtain

$$H(U_{(i)}) - H(U_{(i-1)}) = (U_{(i)} - U_{(i-1)})H^{(1)}(U_{(i-1)}) + \frac{(U_{(i)} - U_{(i-1)})^2}{2}H^{(2)}(c_1), \quad (3.2)$$

for some c_1 between $U_{(i-1)}$ and $U_{(i)}$; c_1 is a random variable because it depends on $U_{(i-1)}$ and $U_{(i)}$. Also

$$H^{(1)}(U_{(i-1)}) = H^{(1)}\left(\frac{i-1}{n+1}\right) + \left(U_{(i-1)} - \frac{i-1}{n+1}\right)H^{(2)}(c_2), \quad (3.3)$$

for some c_2 between $U_{(i-1)}$ and $E(U_{(i-1)}) = (i-1)/(n+1)$; c_2 is also a random variable because it depends on $U_{(i-1)}$. From (3.2) and (3.3), we obtain

$$H(U_{(i)}) - H(U_{(i-1)}) = (U_{(i)} - U_{(i-1)})\left[H^{(1)}\left(\frac{i-1}{n+1}\right) + R_{2i}\right], \quad (3.4)$$

where

$$R_{2i} = \left(U_{(i-1)} - \frac{i-1}{n+1}\right)H^{(2)}(c_2) + \frac{U_{(i)} - U_{(i-1)}}{2}H^{(2)}(c_1).$$

From (3.4), we obtain

$$\begin{aligned} & [H(U_{(i)}) - H(U_{(i-1)})]^{1/2} \\ &= (U_{(i)} - U_{(i-1)})^{1/2} \left[H^{(1)}\left(\frac{i-1}{n+1}\right)^{1/2} + R_{1i} \right], \end{aligned} \quad (3.5)$$

where

$$R_{1i} = \frac{R_{2i}}{\left[H^{(1)}\left(\frac{i-1}{n+1}\right) + R_{2i} \right]^{1/2} + H^{(1)}\left(\frac{i-1}{n+1}\right)^{1/2}}$$

Substituting (3.1),(3.4), and (3.5) for the criterion function $C(f_n, n^\circ)$ gives

$$C(f_n, n^\circ) = \frac{1}{n+1} \sum_{j=1}^{k-1} \frac{\left[\sum_{i=1+n_j^\circ}^{n_{j+1}^\circ} (U_{(i)} - U_{(i-1)})^{1/2} \left[H^{(1)}\left(\frac{i-1}{n+1}\right)^{1/2} + R_{1i} \right] \right]^2}{\sum_{i=1+n_j^\circ}^{n_{j+1}^\circ} (U_{(i)} - U_{(i-1)}) \left[H^{(1)}\left(\frac{i-1}{n+1}\right) + R_{2i} \right]}$$

Let E_1, \dots, E_n be independent exponential random variables with expected value 1. Set $S_{n+1} = \sum_{i=1}^{n+1} E_i$ and $D_i = E_i/S_{n+1}$; then (D_1, \dots, D_{n+1}) is distributed as the set of $n+1$ spacings determined by n independent uniform random variables (Pyke 1965). Thus the criterion function is now

$$C(f_n, n^\circ) = \sum_{j=1}^{k-1} \frac{\left[\frac{1}{n+1} \sum_{i=1+n_j^\circ}^{n_{j+1}^\circ} E_i^{1/2} \left[H^{(1)}\left(\frac{i-1}{n+1}\right)^{1/2} + R_{1i} \right] \right]^2}{\frac{1}{n+1} \sum_{i=1+n_j^\circ}^{n_{j+1}^\circ} E_i \left[H^{(1)}\left(\frac{i-1}{n+1}\right) + R_{2i} \right]} \tag{3.6}$$

3.2 Evaluation of Maximum Bounds for Approximation Error of Sums by Integrals

For a bounded function $G(u)$ defined on $[n_j^\circ/(n+1), n_{j+1}^\circ/(n+1)]$

$$\left| \frac{1}{n+1} \sum_{i=1+n_j^\circ}^{n_{j+1}^\circ} G\left(\frac{i-1}{n+1}\right) - \int_{n_j^\circ/(n+1)}^{n_{j+1}^\circ/(n+1)} G(u) du \right| \leq \frac{n_{j+1}^\circ - n_j^\circ}{2(n+1)^2} \sup |G^{(1)}(u)|.$$

For $\alpha = 1/2, 1,$ and $2,$ set $G(u) = H^{(1)}(u)^\alpha$. Then from A.3, A.4,

$$\sup |H^{(1)}(u)^\alpha| \leq \alpha [m_1^{\alpha-1} \{\alpha = 1/2\} + M_1^{\alpha-1} \{\alpha = 1, 2\}] M_2.$$

Since this bound does not depend on the choice of the k -cutpoints, the error ERR^α of approximating the sum $\sum_{i=1+n_j^\circ}^{n_{j+1}^\circ} H^{(1)}((i-1)/(n+1))^\alpha/(n+1)$ by the integral $\int_{n_j^\circ/(n+1)}^{n_{j+1}^\circ/(n+1)} H^{(1)}(u)^\alpha du$ is from A.6

$$ERR^\alpha \leq O(n^{-1}), \quad \alpha = 1/2, 1, \text{ and } 2. \tag{3.7}$$

3.3 Evaluation of Bounds for the Principal and the Remainder

Terms

Let Z_{1j} , Z_{2j} , W_{1j} , and W_{2j} be as follows.

$$Z_{1j} = \frac{1}{n+1} \sum_{i=1+n_j^{\circ}}^{n_{j+1}^{\circ}} E_i^{1/2} H^{(1)}\left(\frac{i-1}{n+1}\right)^{1/2},$$

$$Z_{2j} = \frac{1}{n+1} \sum_{i=1+n_j^{\circ}}^{n_{j+1}^{\circ}} E_i H^{(1)}\left(\frac{i-1}{n+1}\right),$$

$$W_{1j} = \frac{1}{n+1} \sum_{i=1+n_j^{\circ}}^{n_{j+1}^{\circ}} E_i^{1/2} R_{1i},$$

$$W_{2j} = \frac{1}{n+1} \sum_{i=1+n_j^{\circ}}^{n_{j+1}^{\circ}} E_i R_{2i}.$$

Then Z_{1j} and Z_{2j} are the principal terms and W_{1j} and W_{2j} are the remainder terms. First we evaluate the differences between the principal terms and their corresponding expected values. The expected values of Z_{1j} and Z_{2j} are

$$\begin{aligned} E(Z_{1j}) &= \frac{1}{n+1} \sum_{i=1+n_j^{\circ}}^{n_{j+1}^{\circ}} \left[\int_0^{\infty} x^{3/2-1} \exp(-x) dx \right] H^{(1)}\left(\frac{i-1}{n+1}\right)^{1/2} \\ &= \frac{\pi^{1/2}}{2} \frac{1}{n+1} \sum_{i=1+n_j^{\circ}}^{n_{j+1}^{\circ}} H^{(1)}\left(\frac{i-1}{n+1}\right)^{1/2}, \end{aligned}$$

$$E(Z_{2j}) = \frac{1}{n+1} \sum_{i=1+n_j^{\circ}}^{n_{j+1}^{\circ}} H^{(1)}\left(\frac{i-1}{n+1}\right).$$

Since the E_i are independent, the variances of Z_{1j} and Z_{2j} are

$$V(Z_{1j}) = \frac{1}{(n+1)^2} \left(1 - \frac{\pi}{4}\right) \sum_{i=1+n_j^{\circ}}^{n_{j+1}^{\circ}} H^{(1)}\left(\frac{i-1}{n+1}\right),$$

$$V(Z_{2j}) = \frac{1}{(n+1)^2} \sum_{i=1+n_j^{\circ}}^{n_{j+1}^{\circ}} H^{(1)}\left(\frac{i-1}{n+1}\right)^2.$$

For $\alpha = 1/2, 1,$ and $2,$ we have from A.3 and A.6

$$\frac{1}{n+1} \sum_{i=1+n_j^0}^{n_j^0+1} H^{(1)}\left(\frac{i-1}{n+1}\right)^\alpha = O(1).$$

Thus

$$V(Z_{1j}) = O(n^{-1}), \tag{3.8}$$

$$V(Z_{2j}) = O(n^{-1}). \tag{3.9}$$

The Kolmogorov inequality gives

$$\max_{1 \leq i \leq n} |S_i - E(S_i)| = O_p(V(S_n)^{1/2}),$$

where S_i is the sum of i independent random variables Y_1, \dots, Y_i if each of Y_i has a finite variance. Since the E_i are independent, applying the Kolmogorov inequality to (3.8) and (3.9) gives

$$\max_{1 \leq n_j^0 \leq n_j^0+1 \leq n} |Z_{1j} - E(Z_{1j})| = O_p(n^{-1/2}), \tag{3.10}$$

$$\max_{1 \leq n_j^0 \leq n_j^0+1 \leq n} |Z_{2j} - E(Z_{2j})| = O_p(n^{-1/2}). \tag{3.11}$$

Next we evaluate the remainder terms. Observe that $U_{(i)} - U_{(i-1)} = E_i/S_{n+1}$ where E_i are exponential random variables with expected value 1 and $S_{n+1} = \sum_{i=1}^{n+1} E_i$. Let $E_{(n)} = \max_{1 \leq i \leq n} E_i$, then

$$\Pr(E_{(n)} \leq 2 \ln n) = [1 - \exp(-2 \ln n)]^n = (1 - n^{-2})^n \rightarrow 1 \text{ as } n \rightarrow \infty.$$

Thus $\max_{1 \leq i \leq n} E_i = O_p(\ln n)$. Since S_n is of order n , we obtain from A.4

$$\begin{aligned} \max_{1 \leq i \leq n} \left| \frac{U_{(i)} - U_{(i-1)}}{2} H^{(2)}(c_1) \right| &\leq \max_{1 \leq i \leq n} \left| \frac{U_{(i)} - U_{(i-1)}}{2} \right| M_2 \\ &= O_p\left(\frac{\ln n}{n}\right). \end{aligned} \tag{3.12}$$

Since $U_{(i-1)} = S_{i-1}/S_{n+1}$, we have

$$\begin{aligned} \left| U_{(i-1)} - \frac{i-1}{n+1} \right| &= \left| \frac{S_{i-1}}{S_{n+1}} - \frac{i-1}{n+1} \right| \\ &= \left| \frac{S_{i-1} - (i-1) - (i-1)/(n+1)[S_{n+1} - (n+1)]}{S_{n+1}} \right| \\ &\leq \left| \frac{S_{i-1} - (i-1)}{S_{n+1}} \right| + \frac{i-1}{n+1} \left| \frac{S_{n+1} - (n+1)}{S_{n+1}} \right|. \end{aligned}$$

Hence the Kolmogorov inequality gives

$$\begin{aligned} \max_{1 \leq i \leq n} \left| U_{(i-1)} - \frac{i-1}{n+1} \right| &\leq \max_{1 \leq i \leq n} \left| \frac{S_{i-1} - (i-1)}{S_{n+1}} \right| \\ &\quad + \max_{1 \leq i \leq n} \frac{i-1}{n+1} \left| \frac{S_{n+1} - (n+1)}{S_{n+1}} \right| \\ &= O_p(n^{-1/2}). \end{aligned}$$

Thus from A.4, we obtain

$$\begin{aligned} \max_{1 \leq i \leq n} \left| \left(U_{(i-1)} - \frac{i-1}{n+1} \right) H^{(2)}(c_2) \right| &\leq \max_{1 \leq i \leq n} \left| U_{(i-1)} - \frac{i-1}{n+1} \right| M_2 \\ &= O_p(n^{-1/2}). \end{aligned} \tag{3.13}$$

From (3.12) and (3.13)

$$\begin{aligned} &\max_{1 \leq i \leq n} |R_{2i}| \\ &= \max_{1 \leq i \leq n} \left| \left(U_{(i-1)} - \frac{i-1}{n+1} \right) H^{(2)}(c_2) + \frac{U_{(i)} - U_{(i-1)}}{2} H^{(2)}(c_1) \right| \\ &= O_p(n^{-1/2}), \end{aligned} \tag{3.14}$$

and this bound is independent of the choice of the k -cutpoints. From (3.14) the order of magnitude of the remainder term W_{2j} is

$$\begin{aligned} W_{2j} &\leq \frac{1}{n+1} \sum_{i=1+n_j^*}^{n_j^*+1} E_i O_p(n^{-1/2}) \\ &\leq \frac{1}{n+1} \sum_{i=1}^{n+1} E_i O_p(n^{-1/2}) \\ &= O_p(n^{-1/2}). \end{aligned} \tag{3.15}$$

This bound is independent of the choice of the k -cutpoints. From A.3 the denominator of R_{1i} is

$$0 < \left[H^{(1)}\left(\frac{i-1}{n+1}\right) + R_{2i} \right]^{1/2} + H^{(1)}\left(\frac{i-1}{n+1}\right)^{1/2} = O(1). \tag{3.16}$$

Thus we obtain from (3.14) and (3.16)

$$\begin{aligned} & \max_{1 \leq i \leq n} |R_{1i}| \\ &= \max_{1 \leq i \leq n} \left| R_{2i} / \left[H^{(1)}\left(\frac{i-1}{n+1}\right) + R_{2i} \right]^{1/2} + H^{(1)}\left(\frac{i-1}{n+1}\right)^{1/2} \right| \\ &= O_p(n^{-1/2}). \end{aligned} \tag{3.17}$$

From (3.17) the order of magnitude of the remainder term W_{1j} is

$$\begin{aligned} W_{1j} &\leq \frac{1}{n+1} \sum_{i=1+n_j^*}^{n_{j+1}^*} E_i^{1/2} O_p(n^{-1/2}) \\ &\leq \frac{1}{n+1} \sum_{i=1}^{n+1} E_i^{1/2} O_p(n^{-1/2}) \\ &= O_p(n^{-1/2}). \end{aligned} \tag{3.18}$$

This bound is independent of the choice of the k -cutpoints $(X_{(n_1^*)}, \dots, X_{(n_k^*)})$.

3.4 Evaluation of the Difference between the Empirical Criterion Function and the Theoretical Criterion Function

We know that for $|a| \leq \frac{1}{2}|x|$, $|b| \leq \frac{1}{2}|y|$, $x \neq 0$, and $y \neq 0$,

$$\left| \frac{(x+a)^2}{y+b} - \frac{x^2}{y} \right| \leq \frac{x^2}{y} \left[6 \left| \frac{a}{x} \right| + 2 \left| \frac{b}{y} \right| \right]. \tag{3.19}$$

Let

$$\frac{(Z_{1j} + W_{1j})^2}{Z_{2j} + W_{2j}} = \frac{[(\pi^{1/2}/2) \int_{n_j^*/(n+1)}^{n_{j+1}^*/(n+1)} H^{(1)}(u)^{1/2} du]^2}{\int_{n_j^*/(n+1)}^{n_{j+1}^*/(n+1)} H^{(1)}(u) du} + R_j.$$

where Z_{1j} and Z_{2j} are the principal terms and W_{1j} and W_{2j} are the remainder terms in subsection 3.3. We wish to evaluate the order of magnitude of R_j . Let r_{1j} , and r_{2j} be as follows.

$$\begin{aligned}
 r_{1j} &= \frac{Z_{1j} + W_{1j} - (\pi^{1/2}/2) \int_{n_j^{\circ}/(n+1)}^{n_{j+1}^{\circ}/(n+1)} H^{(1)}(u)^{1/2} du}{(\pi^{1/2}/2) \int_{n_j^{\circ}/(n+1)}^{n_{j+1}^{\circ}/(n+1)} H^{(1)}(u)^{1/2} du} \\
 &= \frac{W_{1j} + [Z_{1j} - E(Z_{1j})] + [E(Z_{1j}) - (\pi^{1/2}/2) \int_{n_j^{\circ}/(n+1)}^{n_{j+1}^{\circ}/(n+1)} H^{(1)}(u)^{1/2} du]}{(\pi^{1/2}/2) \int_{n_j^{\circ}/(n+1)}^{n_{j+1}^{\circ}/(n+1)} H^{(1)}(u)^{1/2} du}, \\
 r_{2j} &= \frac{Z_{2j} + W_{2j} - \int_{n_j^{\circ}/(n+1)}^{n_{j+1}^{\circ}/(n+1)} H^{(1)}(u) du}{\int_{n_j^{\circ}/(n+1)}^{n_{j+1}^{\circ}/(n+1)} H^{(1)}(u) du} \\
 &= \frac{W_{2j} + [Z_{2j} - E(Z_{2j})] + [E(Z_{2j}) - \int_{n_j^{\circ}/(n+1)}^{n_{j+1}^{\circ}/(n+1)} H^{(1)}(u) du]}{\int_{n_j^{\circ}/(n+1)}^{n_{j+1}^{\circ}/(n+1)} H^{(1)}(u) du}.
 \end{aligned}$$

From A.3 and A.6,

$$\frac{\pi^{1/2}}{2} \int_{n_j^{\circ}/(n+1)}^{n_{j+1}^{\circ}/(n+1)} H^{(1)}(u)^{1/2} du \geq \frac{c(\pi m_1)^{1/2}}{2}, \tag{3.20}$$

$$\int_{n_j^{\circ}/(n+1)}^{n_{j+1}^{\circ}/(n+1)} H^{(1)}(u) du \geq cm_1. \tag{3.21}$$

The order of magnitude of r_{1j} is $O_p(n^{-1/2})$ from (3.7), (3.10), (3.18), and (3.20), and that of r_{2j} is $O_p(n^{-1/2})$ from (3.7), (3.11), (3.15), and (3.21).

Thus for $i = 1, 2$, we obtain

$$\Pr \left\{ \max |r_{ij}| \leq \frac{1}{2} \right\} \rightarrow 1 \quad \text{as } n \rightarrow \infty.$$

Notice also that $Z_{2j} + W_{2j}$ is greater than zero because $Z_{2j} + W_{2j}$ is the sum of $(n_{j+1}^{\circ} - n_j^{\circ})$ spacings of adjacent order statistics and this must be greater

than zero from A.3. Hence (3.19), (3.20), and (3.21) gives

$$\begin{aligned} & \max_{1 \leq n_j^\circ \leq n_{j+1}^\circ \leq n} |R_j| \\ &= \max_{1 \leq n_j^\circ \leq n_{j+1}^\circ \leq n} \left| \frac{(Z_{1j} + W_{1j})^2}{Z_{2j} + W_{2j}} - \frac{[(\pi^{1/2}/2) \int_{n_j^\circ/(n+1)}^{n_{j+1}^\circ/(n+1)} H^{(1)}(u)^{1/2} du]^2}{\int_{n_j^\circ/(n+1)}^{n_{j+1}^\circ/(n+1)} H^{(1)}(u) du} \right| \\ &\leq \left| \frac{[(\pi^{1/2}/2) \int_{n_j^\circ/(n+1)}^{n_{j+1}^\circ/(n+1)} H^{(1)}(u)^{1/2} du]^2}{\int_{n_j^\circ/(n+1)}^{n_{j+1}^\circ/(n+1)} H^{(1)}(u) du} \right| [6|r_{1j}| + 2|r_{2j}|] \\ &\quad \text{if } \max |r_{ij}| \leq \frac{1}{2}, \quad i = 1, 2 \\ &\leq O_p(n^{-1/2}). \end{aligned}$$

From (3.6), $C(f_n, \mathbf{n}^\circ) = \sum_{j=1}^{k-1} (Z_{1j} + W_{1j})^2 / (Z_{2j} + W_{2j})$ where k is finite.

Thus we obtain

$$\begin{aligned} & \max_{1=n_1^\circ < n_2^\circ < \dots < n_k^\circ = n} \left| C(f_n, \mathbf{n}^\circ) - \frac{\pi}{4} \sum_{j=1}^{k-1} \frac{[\int_{n_j^\circ/(n+1)}^{n_{j+1}^\circ/(n+1)} H^{(1)}(u)^{1/2} du]^2}{\int_{n_j^\circ/(n+1)}^{n_{j+1}^\circ/(n+1)} H^{(1)}(u) du} \right| \\ &= O_p(n^{-1/2}), \end{aligned}$$

or

$$\max_{1=n_1^\circ < n_2^\circ < \dots < n_k^\circ = n} \left| C(f_n, \mathbf{n}^\circ) - P\left(H, \frac{\mathbf{n}^\circ}{n+1}\right) \right| = O_p(n^{-1/2}). \tag{3.22}$$

3.5 Consistency of Cell Selection Rule

First we will show that $P(H, \mathbf{p})$ is continuous in the compact set $A_{c,j} = \{[p_j, p_{j+1}]; 0 < c \leq p_{j+1} - p_j, j = 1, \dots, k-1\}$. It suffices to show that $\int_{p_j}^{p_{j+1}} H^{(1)}(u)^{1/2} du / \int_{p_j}^{p_{j+1}} H^{(1)}(u) du$ are continuous because $P(H, \mathbf{p})$ is a finite sum of these terms. Both $\int_{p_j}^{p_{j+1}} H^{(1)}(u)^{1/2} du$ and $\int_{p_j}^{p_{j+1}} H^{(1)}(u) du$ are differentiable at p_j and p_{j+1} and $\int_{p_j}^{p_{j+1}} H^{(1)}(u) du > 0$ whenever $p_j < p_{j+1}$. Thus $\int_{p_j}^{p_{j+1}} H^{(1)}(u)^{1/2} du / \int_{p_j}^{p_{j+1}} H^{(1)}(u) du$ is continuous on the set $A_{c,j}$. Hence $P(H, \mathbf{p})$ is continuous on the set $\cup_{j=1}^{k-1} A_{c,j} = A_c$. Since the sets

of k -cutpoints that maximize $C(f_n, \mathbf{n})$ and $P(H, \mathbf{p})$ are $(X_{(n_1^\circ)}, \dots, X_{(n_k^\circ)})$ and $\mathbf{p}^\circ = (p_1^\circ, \dots, p_k^\circ)$ respectively, we have from (3.22)

$$\begin{aligned} \left| C(f_n, \mathbf{n}^\circ) - P(H, \frac{\mathbf{n}^\circ}{n+1}) \right| &\leq O_p(n^{-1/2}), \\ \left| C(f_n, (n+1)\mathbf{p}^\circ) - P(H, \mathbf{p}^\circ) \right| &\leq O_p(n^{-1/2}), \end{aligned}$$

Since $C(f_n, \mathbf{n}^\circ) \geq C(f_n, (n+1)\mathbf{p}^\circ)$ and $P(H, \mathbf{n}^\circ/(n+1)) \leq P(H, \mathbf{p}^\circ)$, we have

$$\left| P(H, \frac{\mathbf{n}^\circ}{n+1}) - P(H, \mathbf{p}^\circ) \right| \leq O_p(n^{-1/2}).$$

Thus for each $\epsilon > 0$, we have

$$\Pr \left\{ \left| P(H, \frac{\mathbf{n}^\circ}{n+1}) - P(H, \mathbf{p}^\circ) \right| < \epsilon \right\} \rightarrow 1, \quad \text{as } n \rightarrow \infty. \quad (3.23)$$

Since $\mathbf{p}^\circ = (p_1^\circ, \dots, p_k^\circ)$ is unique from A.5 and $P(H, \mathbf{p})$ is continuous on the set A_c , for each $\delta > 0$, there is a $\epsilon_\delta > 0$ such that

$$|\mathbf{p} - \mathbf{p}^\circ| \geq \delta \text{ implies } |P(H, \mathbf{p}) - P(H, \mathbf{p}^\circ)| \geq \epsilon_\delta.$$

Otherwise there would be a set of k -cutpoints \mathbf{p}' with $|\mathbf{p}' - \mathbf{p}^\circ| > 0$ such that $P(H, \mathbf{p}') = P(H, \mathbf{p}^\circ)$ by the compactness of A_c , contradicting the uniqueness of \mathbf{p}° . Thus for each $\delta > 0$, there is a $\epsilon_\delta > 0$ such that

$$|P(H, \mathbf{p}) - P(H, \mathbf{p}^\circ)| < \epsilon_\delta \text{ implies } |\mathbf{p} - \mathbf{p}^\circ| < \delta. \quad (3.24)$$

From (3.23) and (3.24), we obtain for each δ

$$\Pr \left\{ \left| \frac{\mathbf{n}^\circ}{n+1} - \mathbf{p}^\circ \right| < \delta \right\} \rightarrow 1, \quad \text{as } n \rightarrow \infty.$$

Therefore

$$\frac{\mathbf{n}^\circ}{n+1} \rightarrow \mathbf{p}^\circ \text{ in probability.}$$

3.6 Relation to the Hellinger Distance

Since $H(u) = x$, we have $H^{(1)}(u) = 1/f(x)$ where $f(x)$ is the density function of the underlying distribution. Then

$$\begin{aligned}
 P(H, \mathbf{p}^\circ) &= \frac{\pi}{4} \sum_{j=1}^{k-1} \frac{[\int_{p_j^\circ}^{p_{j+1}^\circ} H^{(1)}(u)^{1/2} du]^2}{\int_{p_j^\circ}^{p_{j+1}^\circ} H^{(1)}(u) du} \\
 &= \frac{\pi}{4} \sum_{j=1}^{k-1} \frac{[\int_{q_j^\circ}^{q_{j+1}^\circ} f(x)^{1/2} dx]^2}{q_{j+1}^\circ - q_j^\circ},
 \end{aligned}$$

where $q_j^\circ = H(p_j^\circ)$ for $j = 1, \dots, k - 1$. Suppose that we choose a histogram $g(x)$ where $g(x) = b_j$ on $q_j^\circ \leq x \leq q_{j+1}^\circ$. The Hellinger distance between $f(x)$ and $g(x)$, $\sum_{j=1}^{k-1} \int_{q_j^\circ}^{q_{j+1}^\circ} [f(x)^{1/2} - b_j^{1/2}]^2 dx$, is minimal when the height b_j of the histogram is

$$b_j = \left[\frac{\int_{q_j^\circ}^{q_{j+1}^\circ} f(x)^{1/2} dx}{q_{j+1}^\circ - q_j^\circ} \right]^2 \bigg/ \sum_{j=1}^{k-1} \frac{[\int_{q_j^\circ}^{q_{j+1}^\circ} f(x)^{1/2} dx]^2}{q_{j+1}^\circ - q_j^\circ}, \quad q_j^\circ \leq x \leq q_{j+1}^\circ. \tag{3.25}$$

The resulting Hellinger distance is

$$2 - 2 \left[\sum_{j=1}^{k-1} \frac{[\int_{q_j^\circ}^{q_{j+1}^\circ} f(x)^{1/2} dx]^2}{q_{j+1}^\circ - q_j^\circ} \right]^{1/2} = 2 - 2 \left[\frac{4}{\pi} P(H, \mathbf{p}^\circ) \right]^{1/2}.$$

Hence choosing the set of k -cutpoints that maximizes $P(H, \mathbf{p})$ is equivalent to choosing the set of k -cutpoints that minimizes the Hellinger distance between an underlying density $f(x)$ and the histogram estimate $g(x)$ whose height b_j in $q_j^\circ \leq x \leq q_{j+1}^\circ$ is (3.25).

4. CELL SELECTION ALGORITHM

The set of k -cutpoints $(X_{(n_1^\circ)}, \dots, X_{(n_k^\circ)})$ that numerically maximizes the criterion function

$$C(f_n, \mathbf{n}) = \frac{1}{n+1} \sum_{j=1}^{k-1} \frac{[\sum_{i=1+n_j}^{n_{j+1}} T_i^{1/2}]^2}{\sum_{i=1+n_j}^{n_{j+1}} T_i}$$

can be found using a dynamic programming algorithm described below. The value of $C(f_n, \mathbf{n})$ depends both on the number and choice of the cutpoints. We make this dependence specific by introducing the sequence of functions C_{lm} defined for $l = 2, \dots, k$ and $l \leq m \leq n$ as follows.

$$C_{lm} = \max_{1=n_1 < n_2 < \dots < n_l=m} \frac{1}{n+1} \sum_{j=1}^{l-1} \frac{[\sum_{i=1+n_j}^{n_{j+1}} T_i^{1/2}]^2}{\sum_{i=1+n_j}^{n_{j+1}} T_i}$$

The function C_{lm} is larger than any other choice of l -cutpoints from the first m order statistics. The initial value C_{1m} is given by

$$C_{1m} = \frac{1}{n+1} \frac{[\sum_{i=2}^m T_i^{1/2}]^2}{\sum_{i=2}^m T_i} \tag{4.1}$$

To obtain the recurrence relation connecting C_{l-1j} and C_{lm} , we proceed as follows. When $l-1$ cutpoints are chosen from the first j order statistics $X_{(1)}, \dots, X_{(j)}$, the criterion function is, by definition, C_{l-1j} . The last cutpoint is $X_{(m)}$ and this adds $[\sum_{i=j+1}^m T_i^{1/2}]^2 / [(n+1) \sum_{i=j+1}^m T_i]$ to C_{l-1j} . The optimal choice of the l -cutpoints $(X_{(n_1^*)}, \dots, X_{(n_l^*)})$ from $X_{(1)}, \dots, X_{(m)}$ is therefore the one that maximizes the sum of these two terms over j , or

$$C_{lm} = \max_{l-1 \leq j \leq m-1} \left[C_{l-1j} + \frac{1}{n+1} \frac{[\sum_{i=j+1}^m T_i^{1/2}]^2}{\sum_{i=j+1}^m T_i} \right] \tag{4.2}$$

By applying (4.2) recursively for $l = 2, \dots, k$ and $l \leq m \leq n$ along with (4.1), we obtain the maximum of $C(f_n, \mathbf{n})$ over $X_{(1)}, \dots, X_{(n)}$.

Example Suppose that we have six order statistics $X_{(1)}, X_{(2)}, X_{(3)}, X_{(4)}, X_{(5)}, X_{(6)}$ and that we would like to choose four cutpoints. From the restrictions on the location of the cutpoints, the leftmost and rightmost cutpoints are $X_{(n_1)} = X_{(1)}$ and $X_{(n_4)} = X_{(6)}$ respectively. We need to choose two more cutpoints from $X_{(2)}, X_{(3)}, X_{(4)}, X_{(5)}$. The dynamic programming algorithm described above works as follows.

Stage 1. Choosing the best three cutpoints.

1. The best three cutpoints in $X_{(1)}, \dots, X_{(5)}$ are one of the following three whose criterion function $C(f_n, \mathbf{n})$ is maximal.
 - a. Three cutpoints $X_{(1)}, X_{(2)}, X_{(5)}$.
 - b. Three cutpoints $X_{(1)}, X_{(3)}, X_{(5)}$.
 - c. Three cutpoints $X_{(1)}, X_{(4)}, X_{(5)}$.
2. The best three cutpoints in $X_{(1)}, \dots, X_{(4)}$ are one of the following two whose criterion function $C(f_n, \mathbf{n})$ is maximal.
 - a. Three cutpoints $X_{(1)}, X_{(2)}, X_{(4)}$.
 - b. Three cutpoints $X_{(1)}, X_{(3)}, X_{(4)}$.
3. The best three cutpoints in $X_{(1)}, \dots, X_{(3)}$ are the following.
 - a. Three cutpoints $X_{(1)}, X_{(2)}, X_{(3)}$.

Stage 2. Choosing the best four cutpoints.

1. The best four cutpoints in $X_{(1)}, \dots, X_{(6)}$ are one of the following three whose criterion function $C(f_n, \mathbf{n})$ is maximal.
 - a. The best three cutpoints in $X_{(1)}, \dots, X_{(5)}$ and one cutpoint $X_{(6)}$.
 - b. The best three cutpoints in $X_{(1)}, \dots, X_{(4)}$ and one cutpoint $X_{(6)}$.
 - c. The best three cutpoints in $X_{(1)}, \dots, X_{(3)}$ and one cutpoint $X_{(6)}$.

Since the values of the criterion function $C(f_n, \mathbf{n})$ in each of the best three cutpoints selection are already computed at **Stage 1**, you only have to calculate the increment of $C(f_n, \mathbf{n})$ that is brought in by adding one cutpoint at **Stage 2**.

5. SIMULATION

Consider the step distribution whose density function $f(x)$ is as follows.

$$f(x) = \frac{3}{2}\{0 \leq x \leq \frac{1}{2}\} + \frac{1}{2}\{\frac{1}{2} < x \leq 1\}.$$

Then the distribution function $F(x)$ is given by

$$F(x) = \frac{3}{2}x\{0 \leq x \leq \frac{1}{2}\} + \frac{1}{2}(x+1)\{\frac{1}{2} < x \leq 1\}.$$

The conditions A.1 through A.5 are met as follows for $k = 3$.

A.1. The inverse of the cumulative distribution function $H(u) = F^{-1}(u)$ exists on $[0, 1]$ and is given as follows.

$$H(u) = \frac{2}{3}u\{0 \leq u \leq \frac{3}{4}\} + (2u - 1)\{\frac{3}{4} < u \leq 1\}.$$

A.2. $H(u)$ is twice differentiable and its derivatives are continuous on $[0, 1]$ except for one point as follows.

$$H^{(1)}(u) = \frac{2}{3}\{0 \leq u \leq \frac{3}{4}\} + 2\{\frac{3}{4} < u \leq 1\},$$

$$H^{(2)}(u) = 0.$$

A.3. $0 < m_1 = 2/3 \leq H^{(1)}(u) \leq M_1 = 2, \quad 0 \leq u \leq 1.$

A.4. $|H^{(2)}(u)| \leq M_2 = 0, \quad 0 < u \leq 1.$

A.5. There is a unique set of 3-cutpoints $\mathbf{p}^\circ = (0, 3/4, 1)$ that maximizes $P(H, \mathbf{p})$. Let the middle cutpoint be p . Then

$$\begin{aligned} P(H, \mathbf{p}) &= \frac{\pi}{4} \left[\frac{[\int_0^{3/4} (2/3)^{1/2} du]^2}{\int_0^{3/4} (2/3) du} + \frac{[\int_{3/4}^1 2^{1/2} du]^2}{\int_{3/4}^1 2 du} \right] \\ &= \frac{\pi}{4}, \quad \text{at } p = 3/4, \end{aligned}$$

while both

$$\begin{aligned} P(H, \mathbf{p}) &= \frac{\pi}{4} \left[\frac{[\int_0^{3/4} (2/3)^{1/2} du + \int_{3/4}^p 2^{1/2} du]^2}{\int_0^{3/4} (2/3) du + \int_{3/4}^p 2 du} \right] \\ &\quad + \frac{\pi}{4} \frac{[\int_p^1 2^{1/2} du]^2}{\int_p^1 2 du} \\ &= \frac{\pi}{4} \left[\frac{3^{1/2}p + (2 - 3^{3/2})/4}{2p - 1} \right], \quad 3/4 < p, \\ P(H, \mathbf{p}) &= \frac{\pi}{4} \frac{[\int_0^p (2/3)^{1/2} du]^2}{\int_0^p (2/3) du} \\ &\quad + \frac{\pi}{4} \left[\frac{[\int_p^{3/4} (2/3)^{1/2} du + \int_{3/4}^1 2^{1/2} du]^2}{\int_p^{3/4} (2/3) du + \int_{3/4}^1 2 du} \right] \\ &= \frac{\pi}{4} \left[\frac{-(1/3^{1/2})p + (2 + 3^{1/2})/4}{-\frac{2}{3}p + 1} \right], \quad p < 3/4, \end{aligned}$$

are decreasing as p moves away from $3/4$. Thus

$$0 < c = 1/4 \leq p_{j+1}^c - p_j^c, \quad j = 1, 2.$$

The simulation was done using this step distribution with the sample size $n = 9, 49, 99$ and 499 . For each sample size, random numbers from the uniform $[0, 1]$ were generated and converted to those corresponding to the step distribution using the inverse of the probability integral transformation. Then the center cutpoint that maximizes the empirical criterion function $C(f_n, \mathbf{n})$ was computed using the algorithm described in Section 4. The other two cutpoints were restricted to the smallest and the largest order statistics. For each sample, the proportion $N/(n+1)$ was computed where N is the index of the center cutpoint. This procedure was repeated a hundred times for each sample size and the mean and mean squared error (MSE) of $N/(n+1)$ to the theoretical center cutpoint $p = 3/4$ were computed. The portable random number generator used was written by ALAN M. GROSS in Bell Laboratories in FORTRAN and is widely available (Digital Signal Processing Committee 1979). The rest of the program was written by the author in FORTRAN. Convergence of $N/(n+1)$ in probability to $p = 3/4$ is observed from the simulation result in TABLE. I.

TABLE. I. Error Between Sample and Theoretical Cutpoint
in One Hundred Repetitions

Sample Size	Theoretical Mean	Sample Mean	MSE
9	0.750	0.579	0.0705
49	0.750	0.677	0.0400
99	0.750	0.729	0.0104
499	0.750	0.749	0.0005

6. DISCUSSION

The proposed procedure to construct a variable cell-size histogram can be applied to a wide variety of problems arising in applied statistics. It is simple not only conceptually but also computationally because it relies on a well established algorithm.

Further work needs to be done:

- a. To develop a method for choosing the number of cells.
- b. To develop asymptotics when the number of cells approaches to ∞ as $n \rightarrow \infty$.
- c. To quantify the rate of convergence of $\mathbf{n}^o / (n + 1)$ to \mathbf{p}^o .
- d. To generalize the theorem to distributions with infinite support.

ACKNOWLEDGEMENT

The author wishes to thank Professor John A. Hartigan for his helpful comments and encouragement.

BIBLIOGRAPHY

- Digital Signal Processing Committee (1979), *Programs for Digital Signal Processing*, S-11-S-12, New York, NY: The Institute of Electrical and Electronics Engineers, Inc..
- Freedman, D., and Diaconis, P. (1981), "On the Histogram as a Density Estimator: L_2 Theory," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57, 453-476.
- Kogure, A. (1986), "Optimal Cells for a Histogram," unpublished Ph.D. dissertation, Yale University, Dept. of Statistics.
- Kogure, A. (1987), "Asymptotically Optimal Cells for a Histogram," *The Annals of Statistics*, Vol. 15, No. 3, 1023-1030.

- Pyke, R. (1965), "Spacings," *J. R. Statist. Soc. B*, 27, 395-449.
- Rudemo, M. (1982), "Empirical Choice of Histograms and Kernel Density Estimators," *Scand. J. Statist.*, 9, 65-78.
- Scott, D. W. (1979), "On Optimal and Data-Based Histograms," *Biometrika*, 66, 605-610.
- Stone, C. J. (1984), "An Asymptotically Optimal Histogram Selection Rule," *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, eds. L. M. Le Cam and R. A. Olshen, Monterey, CA: Wadsworth, Inc., 2, 513-520.

Received by Editorial Board member August, 1987; Revised December, 1987.

Recommended by Emanuel Parzen, Department of Statistics, Texas A & M University.