The Annals of Statistics
1992, Vol. 20, No. 1, 291–304

AN OPTIMAL VARIABLE CELL HISTOGRAM BASED ON THE SAMPLE SPACINGS[1]

BY YUICHIRO KANAZAWA

University of Tsukuba

Suppose we wish to construct a variable k-cell histogram based on an independent identically distributed sample of size $n - 1$ from an unknown density f on the interval of finite length. A variable cell histogram requires cutpoints and heights of all of its cells to be specified. We propose the following procedure: (i) choose from the order statistics corresponding to the sample a set of $k + 1$ cutpoints that maximize a criterion, a function of the sample spacings; (ii) compute heights of the k cells according to a formula. The resulting histogram estimates a k-cell *theoretical* histogram that stays constant within a cell and that minimizes the Hellinger distance to the density f. The histogram tends to estimate low density regions accurately and is easy to compute. We find the number of cells of order $n^{1/3}$ minimizes the mean Hellinger distance between the density f and a class of histograms whose cutpoints are chosen from the order statistics.

1. Introduction. Suppose we construct a histogram based on an independent and identically distributed sample $X_1, \ldots, X_n$ of size n from an unknown density f on the interval I of finite length. A histogram $\hat{f}$ with an identical cell width requires choice of $(a, |C|)$, where a is the leftmost cutpoint and $|C|$ is the common cell width. Heights of the cells are then determined by counting observations that fall in the cells. The cell width $|C|$ is customarily chosen to minimize the mean integrated squared error $\rho(\hat{f}, f) = E[\int_I (\widehat{f(x)} - f(x))^2\, dx]$ to the density f over the sample. Scott (1979) showed that the cell width

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

asymptotically minimizes $\rho(\hat{f}, f)$. The cell width, however, depends on the unknown $\int_I f'(x)^2\, dx$. For a normal density f, it is $3.49 s n^{-1/3}$, where s is the sample standard deviation. Freedman and Diaconis (1981) suggested that the cell width of $2 \times$ interquartile range $\times n^{-1/3}$. Rudemo (1982) proposed the cell selection rule $(a, |C|)$ that minimizes

$$\frac{1}{|C|}\left[\frac{2}{n-1} - \frac{n+1}{n-1}\sum_j \left[\frac{\sum_{i=1}^n \{X_i \in C_j\}}{n}\right]^2\right],$$

whose expected value is $\rho(\hat{f}, f) - \int f(x)^2\, dx$. (Here C_j is the jth cell $[a +$

Received September 1989; revised April 1991.
[1]Research supported in part by NSF Grant DMS-86-17919.
AMS 1980 *subject classifications*. Primary 62G05; secondary 62E20.
Key words and phrases. Density estimation, Hellinger distance, histogram, order statistics, spacing.

$(j-1)|C|, a+j|C|]$.) Stone (1984) extended this cell selection rule to multidimensional density. He showed that the cell selection rule is asymptotically close to the minimum of $\rho(\hat{f}, f)$ for a density on the interval of finite length.

Procedures for choosing a histogram with variable cell widths have received relatively little attention, however. Kogure (1987) extends Rudemo's rule to a histogram with locally equisized cells. In this paper we propose a variable k-cell histogram f_{n^0} that (i) accurately estimates low density regions where lack of the observations makes estimation difficult; and (ii) is simple to compute. We then study the proposed histogram.

To derive a variable cell histogram with these properties, we shall proceed as follows: (i) compute a k-cell *theoretical histogram* g_{p^0} that minimizes the Hellinger distance $\mathrm{HD}(g_p, f) = \int_I [g_p(x)^{1/2} - f(x)^{1/2}]^2\,dx$ to a density f over a class of k-cell histogram-type densities g_p that stay constant within a cell; (ii) derive a histogram f_{n^0} based on the sample that estimates g_{p^0}. The theoretical histogram g_{p^0} depends on f but not on the sample. Examples 1.1 and 1.2 will illustrate how we actually compute the theoretical histogram for a particular density. Instead of the integrated squared error (ISE) adopted by Scott, Rudemo, Stone and Kogure, we measure error by the Hellinger distance. The Hellinger distance allocates larger weight to regions with low density relative to ones with high density than the ISE. This enables our histogram to estimate low density regions more accurately than the one based on the ISE.

We shall describe the two steps heuristically but with more details. First we note the height h_j,

$$h_j = \left[\frac{\int_{I_j} f(x)^{1/2}\,dx}{|I_j|}\right]^2 \Bigg/ \sum_{i=1}^{k} \frac{\left[\int_{I_i} f(x)^{1/2}\,dx\right]^2}{|I_i|}, \qquad x \in I_j,$$

of the jth cell I_j of a class of k-cell histogram-type density g_p

$$\mathscr{L}(k) = \left\{ g_p \colon g_p(x) = \sum_{j=1}^{k} h_j\{x \in I_j\},\ \sum_{j=1}^{k} h_j|I_j| = 1,\ h_j \geq 0 \right\},$$

where $|I_j|$ is the width of the jth cell, minimizes the Hellinger distance to the density f,

$$\mathrm{HD}(g_p, f) = \sum_{j=1}^{k} \int_{I_j} \left[h_j^{1/2} - f(x)^{1/2}\right]^2 dx.$$

This may be done by the method of Lagrange multipliers. The resulting Hellinger distance $\mathrm{HD}(g_p, f)$ is

$$\mathrm{HD}(g_p, f) = 2 - 2\left[\sum_{j=1}^{k} \frac{\left[\int_{I_j} f(x)^{1/2}\,dx\right]^2}{|I_j|}\right]^{1/2}.$$

Let H and $H^{(1)} = 1/f$ be the inverse of the distribution function F of the unknown density f and its first derivative, respectively. If we denote the

endpoints of $F(I_j)$ by $[p_j, p_{j+1}]$, elementary calculation shows that

$$P(H,\mathbf{p}) = \frac{\pi}{4}\sum_{j=1}^{k}\frac{\left[\int_{p_j}^{p_{j+1}}H^{(1)}(u)^{1/2}\,du\right]^2}{\int_{p_j}^{p_{j+1}}H^{(1)}(u)\,du} = \frac{\pi}{4}\sum_{j=1}^{k}\frac{\left[\int_{I_j}f(x)^{1/2}\,dx\right]^2}{|I_j|}.$$

Let $K = k + 1$ be the number of cutpoints of the histogram-type density g_p. If a set of K cutpoints $\mathbf{p}^0 = (p_1^0, \ldots, p_K^0)$ maximizes $P(H,\mathbf{p})$, then it minimizes $\mathrm{HD}(g_p, f)$. For this set of cutpoints $\mathbf{p}^0$, we obtain a set of k-heights $\mathbf{h}^0 = (h_1^0, \ldots, h_k^0)$ by substituting $\mathbf{p}^0$ in a formula of h_j above. Thus a pair $(\mathbf{p}^0, \mathbf{h}^0)$ determines the theoretical histogram g_{p^0}.

Second, to find the histogram f_{n^0} that best estimates g_{p^0}, we shall proceed as follows: (i) construct sample-based analogs $C(X_{(1)}, \ldots, X_{(n-1)}, \mathbf{n})$ and h_{nj} of $P(H,\mathbf{p})$ and h_j, respectively; (ii) find a set of cutpoints that maximizes $C(X_{(1)}, \ldots, X_{(n-1)}, \mathbf{n})$; (iii) compute h_{nj} for the set of cutpoints. We expect the histogram f_{n^0} constructed this way to converge to the theoretical histogram g_{p^0}. Now we need a sample-based analog $C(X_{(1)}, \ldots, X_{(n-1)}, \mathbf{n})$ of $P(H,\mathbf{p})$. Define the ith spacing as $T_i = X_{(i+1)} - X_{(i)}$, $i = 1, \ldots, n-2$, where $X_{(i)}$ is an ith order statistic of an independent and identically distributed sample $X_1, \ldots, X_{n-1}$ of size $n-1$. The inverse of the probability integral transformation gives $X_{(i)} = H(U_{(i)})$, where $U_{(i)}$ is the ith order statistic from an independent and identically distributed sample of size $n-1$ from the uniform $[0,1]$. Let $e_1, \ldots, e_n$ be independent exponentials with mean 1 and set $s_n = \sum_{i=1}^{n} e_i$. From the well-known relation between the uniform spacings and standardized exponentials e_i/s_n, we have

$$T_i \approx (U_{(i+1)} - U_{(i)})H^{(1)}(i/n) = e_i H^{(1)}(i/n)/s_n.$$

Then for a criterion $C(X_{(1)}, \ldots, X_{(n-1)}, \mathbf{n})$ below, we obtain the following:

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

$$\approx \sum_{j=1}^{k}\frac{\left[n^{-1}\sum_{i=n_j}^{-1+n_{j+1}}e_i^{1/2}H^{(1)}(i/n)^{1/2}\right]^2}{n^{-1}\sum_{i=n_j}^{-1+n_{j+1}}e_i H^{(1)}(i/n)} \approx P(H,\mathbf{p}).$$

We can construct a sample-based analog h_{nj} of h_j similarly.

We summarize the procedure to construct a variable k-cell histogram f_{n^0}:

STEP 1. Find a set of K cutpoints $(X_{(n_1^0)}, \ldots, X_{(n_K^0)})$ that maximizes

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

STEP 2. Compute the optimal height h^0_{nj} from the cutpoints in Step 1 by

$$h_{nj} = \left[\frac{\sum_{i=n_j}^{-1+n_{j+1}} T_i^{1/2}}{\sum_{i=n_j}^{-1+n_{j+1}} T_i} \right]^2 \Bigg/ \sum_{m=1}^{k} \frac{\left[\sum_{i=n_m}^{-1+n_{m+1}} T_i^{1/2}\right]^2}{\sum_{i=n_m}^{-1+n_{m+1}} T_i}. \tag{2}$$

We note the criterion requires a class of k-cell empirical histograms f_n with the jth cell I_{nj}

$$\mathscr{L}_n(k) = \left\{ f_n \colon f_n(x) = \sum_{j=1}^{k} h_{nj}\{x \in I_{nj}\}, \sum_{j=1}^{k} h_{nj}|I_{nj}| = 1, h_{nj} \geq 0 \right\},$$

from which our histogram f_{n^0} is chosen to have: (i) variable cell widths for the cutpoints are chosen from $X_{(1)}, \ldots, X_{(n-1)}$; (ii) a domain $[X_{(1)}, X_{(n-1)}]$.

The intuition behind the choice of criterion (1) is that the information on the density is reflected in the spacings in such a way that regions with narrow spacings tend to have high density, while ones with wide spacings tend to have low density. Maximizing $C(X_{(1)}, \ldots, X_{(n-1)}, \mathbf{n})$ is computationally simple through the *dynamic programming algorithm*. For a description and an example of the algorithm applied to $C(X_{(1)}, \ldots, X_{(n-1)}, \mathbf{n})$, see Kanazawa (1988a, b).

The density f has to be *smooth* to substantiate the heuristic argument that $C(X_{(1)}, \ldots, X_{(n-1)}, \mathbf{n})$ and $P(H, \mathbf{p})$ are close. Also the theoretical histogram g_{p^0} has to be unique to establish the histogram f_{n^0} converges to g_{p^0}. In Kanazawa (1988a, b), we showed f_{n^0} with the known number of cells converges in probability to g_{p^0} under the smoothness conditions A1–A3 on f and the uniqueness condition B1 on g_{p^0}:

A1. F is twice continuously differentiable except for finite points.
A2. f is bounded away from 0 and ∞.
A3. f' is bounded away from ∞.
B1. A unique choice of cells I_j, $j = 1, \ldots, k$, that maximizes

$$P(H, \mathbf{p}) = \sum_{j=1}^{k} \frac{\left[\int_{I_j} f(x)^{1/2}\, dx\right]^2}{|I_j|}.$$

Some densities f that satisfy these conditions and their corresponding theoretical histograms g_{p^0} are given below.

EXAMPLE 1.1. For a finite mixture of uniforms, the number and location of cells of g_{p^0} coincide with those of f. We obtain the minimal $\mathrm{HD}(g_p, f) = 0$.

EXAMPLE 1.2. Let f be a quadratic density $f = 3x^2/7$ on $I = [1, 2]$. For g_{p^0} with two cells, let the mid-cutpoint be x, where $1 < x < 2$. The resulting $P(H, \mathbf{p}) \propto [(x + 1)^2(x - 1) + (2 + x)^2(2 - x)]$ has its maximum at $x = 3/2$. Thus the two cells have equal width. In general for g_{p^0} with k cells, any two neighboring cells have the same width and so all cells are of the same width.

For a uniform density f on the interval I of finite length, a theoretical histogram g_{p^0} with $k > 1$ cells does not exist because any choice of cells that covers the interval produce an identical $P(H, \mathbf{p})$. This violates B1.

For a density whose theoretical histogram g_{p^0} has a *correct* and finite number of cells as in Example 1.1, consistency of f_{n^0} with the known number of cells to g_{p^0} in Kanazawa (1988a, b) in principle warrants the validity of the procedure, though the problem remains regarding how we actually identify the true number of cells for the density. For smoother densities such as the one in Example 1.2, however, there is no *correct* number of cells and we are forced to determine the number of cells. As a global measure of error between a class $\mathscr{L}_n(k)$ of k-cell histograms f_n whose cutpoints are chosen from the order statistics $X_{(1)}, \ldots, X_{(n-1)}$ and the unknown density f, we use the mean Hellinger distance between f_n and f

$$d(f_n, f) = E\left[\int_I \left(f_n(x)^{1/2} - f(x)^{1/2}\right)^2 dx\right],$$

where E denotes the expectation with respect to the sample $X_1, \ldots, X_{n-1}$. We note that the mean Hellinger distance to f is defined for f_n, and not for f_{n^0}, the k-cell histogram obtained by maximizing $C(X_{(1)}, \ldots, X_{(n-1)}, \mathbf{n})$. The mean Hellinger distance can be broken down into two components:

$$d(f_n, f) = \mathrm{HD}(g_p, f) - 2E\int_I \left(f_n(x)^{1/2} - g_p(x)^{1/2}\right) f(x)^{1/2}\, dx,$$

where the first component on the right is the bias, the second the sampling variation. Increasing the number of cells decreases the bias while it increases the sampling variation. Given the sample size $n - 1$, we wish to know if there is an optimal number of cells $\hat{k}$ for f_n as a function of the sample size that strikes the balance between these two components. We study this problem in Section 2 and present the proof in Section 3. We impose a condition that prevents a small number of cells from dominating the other cells on f_n. We also add a smoothness condition on f. Under these conditions we find the number of cells $\hat{k} = \lambda(f)n^{1/3}$, where $\lambda(f)$ is dependent only on f, asymptotically minimizes the mean Hellinger distance $d(f_n, f)$.

2. Optimal number of cells. Given the sample size $n - 1$, we wish to know if there is an optimal number of cells $\hat{k}$ for f_n as a function of the sample size that strikes the balance between the bias and the sampling variation. Let $N_j = n_j/n$ and $\Delta_j N = (n_{j+1} - n_j)/n$, where $[n_j, n_{j+1}]$ are the indices of the cutpoints $[X_{(n_j)}, X_{(n_{j+1})}]$ of the jth cell of f_n. We denote $\sum_i = \sum_{i=n_j}^{-1+n_{j+1}}$, $\sum_j = \sum_{j=1}^{k}$, $i/n = i_n^-$ and $(i+1)/n = i_n^+$. In addition to A1–A3, we need the following:

A4. f'' exists and is bounded away from ∞.

C1. For all $\Delta_j N$, where $j = 1, \ldots, k$, and some constants C_0 and C^0

$$0 < \frac{C_0}{k} \le \Delta_j N \le \frac{C^0}{k} < 1.$$

An additional smoothness condition on the density is in A4. A constraint that prevents a small number of cells from dominating the other cells on the histogram is in C1. We note that the cutpoints of f_n in C1 do not involve maximizing the criterion (1) and are denoted by $(X_{(n_1)}, \ldots, X_{(n_K)})$, while those of f_{n^0} obtained by maximizing (1) are denoted by $(X_{(n_1^0)}, \ldots, X_{(n_K^0)})$. We present the theorem in terms of H's.

THEOREM 1. *Let the following conditions and C1 be satisfied:*

A1′. *$H(u)$ is three times continuously differentiable except for finite points.*
A2′. $0 < m_1 \le H^{(1)}(u) \le M_1 < \infty$, $0 \le u \le 1$.
A3′. $|H^{(2)}(u)| \le M_2 < \infty$, $0 \le u \le 1$.
A4′. $|H^{(3)}(u)| \le M_3 < \infty$, $0 \le u \le 1$.

As $n \to \infty$, the number of cells $\hat{k}$ that minimize the mean Hellinger distance between f_n and f, $d(f_n, f)$, satisfies

$$\frac{\hat{k}}{n^{1/3}} \to \left[\frac{\pi}{24(4-\pi)}\right]^{1/3} \int_0^1 \left[\frac{H^{(2)}(u)}{H^{(1)}(u)}\right]^{2/3} du.$$

As $n \to \infty$ the minimal $d(f_n, f)$ satisfies

$$\min d(f_n, f) n^{2/3} \to \left[\frac{3(4-\pi)}{8\pi}\right]^{2/3} \int_0^1 \left[\frac{H^{(2)}(u)}{H^{(1)}(u)}\right]^{2/3} du.$$

For a density in Example 1.2, $\hat{k}/n^{1/3}$ converges to 0.64 and $\min d(f_n, f)n^{2/3}$ converges to 0.26. Theorem 1 is derived from two propositions. The first of these gives the mean Hellinger distance between f_n and f.

PROPOSITION 1. *Under* A1′–A4′ *and* C1 *we have*

$$(3)\quad d(f_n, f) = \frac{1}{48} \sum_j (\Delta_j N)^3 \left[\frac{H^{(2)}(N_j)}{H^{(1)}(N_j)}\right]^2 + \left(\frac{4}{\pi} - 1\right)\frac{k}{n} + O(n^{-1}k^{1/2}).$$

Divide the domain $[0, 1)$ of $H(u)$ into $M = 1/\delta$ subintervals each of which has length δ. Assume that for each m, $0 \le m \le M - 1$, some $N_j = m\delta$. Let k_m be the number of N_j that falls into mth subinterval $[m\delta, (m+1)\delta)$. We shall minimize (3) in three stages in the second proposition: first with respect to $\Delta_j N$ subject to the constraint that there are k_m cutpoints in the mth subinterval; then with respect to k_m subject to the constraint that there are $K = k + 1$ cutpoints in the interval $[0, 1)$; finally with respect to k.

PROPOSITION 2. *Suppose that $d(f_n, f)$ is minimized with respect to $(\Delta_j N, k_m, k)$. Under* A1′–A4′ *and* C1 *the number of cutpoints k_m in the mth*

subinterval as $n \to \infty$ *satisfies*

$$\frac{k_m}{k} \to \frac{\left[H^{(2)}(m\delta)/H^{(1)}(m\delta)\right]^{2/3}}{\sum_{m=0}^{M-1}\left[H^{(2)}(m\delta)/H^{(1)}(m\delta)\right]^{2/3}}. \tag{4}$$

As $n \to \infty$ *and* $\delta \to 0$ *the number of cells* $\hat{k}$ *that minimize* $d(f_n, f)$ *satisfies*

$$\frac{\hat{k}}{n^{1/3}} \to \left[\frac{\pi}{24(4-\pi)}\right]^{1/3} \int_0^1 \left[\frac{H^{(2)}(u)}{H^{(1)}(u)}\right]^{2/3} du.$$

As $n \to \infty$ *the minimal* $d(f_n, f)$ *satisfies*

$$\min_{(\Delta_j N, k_m, k)} d(f_n, f) n^{2/3} \to \left[\frac{3(4-\pi)}{8\pi}\right]^{2/3} \int_0^1 \left[\frac{H^{(2)}(u)}{H^{(1)}(u)}\right]^{2/3} du.$$

Since $H^{(2)}/H^{(1)} = -f'/f^2$, (4) implies that the mean Hellinger distance is minimal if we take a small number of cells in a region with small change in f and a large number of cells in a region with big change in f, provided that the average values of f in the two regions are the same. The result of Proposition 2 holds without the constraint that the cutpoints include the set $\{m\delta\colon 0 \le m \le M-1\}$, since the change in $d(f_n, f)$ due to adding such cutpoints is $O(\delta k^{-2})$. Thus Propositions 1 and 2 imply Theorem 1.

3. Proofs. This section contains proofs of two propositions.

PROOF OF PROPOSITION 1. Applying the height h_{nj} of jth cell in (2) to $\mathrm{HD}(f_n, f) = 2 - 2\Sigma_j h_{nj}^{1/2} \int_{X_{(n_j)}}^{X_{(n_j+1)}} f(x)^{1/2}\, dx$ gives the Hellinger distance

$$\mathrm{HD}(f_n, f) = 2 - 2\left[\sum_j \frac{\Sigma_i T_i^{1/2} \int_{X_{(n_j)}}^{X_{(n_j+1)}} f(x)^{1/2}\, dx}{\Sigma_i T_i}\right] \Bigg/ \left[\sum_j \frac{\left[\Sigma_i T_i^{1/2}\right]^2}{\Sigma_i T_i}\right]^{1/2}. \tag{5}$$

Expand the components in (5) and express them in terms of independent exponentials e_i with mean 1 and $s_n = \sum_{i=1}^n e_i$. Then we have

$$\sum_i T_i^{1/2} = \frac{1}{s_n^{1/2}} \sum_i e_i^{1/2}\left[H^{(1)}(i_n)^{1/2} + R_{1i}\right] = \frac{n}{s_n^{1/2}}(Z_{1j} + W_{1j}),$$

$$\sum_i T_i = \frac{1}{s_n} \sum_i e_i\left[H^{(1)}(i_n) + R_{2i}\right] = \frac{n}{s_n}(Z_{2j} + W_{2j}),$$

$$\int_{X_{(n_j)}}^{X_{(n_j+1)}} f(x)^{1/2}\, dx = \frac{1}{s_n} \sum_i e_i\left[H^{(1)}(i_n)^{1/2} + R_{3i}\right] = \frac{n}{s_n}(Z_{3j} + W_{3j}),$$

where R_{1i}, R_{2i} and R_{3i} are, for $c_1 \in [U_{(i)}, U_{(i+1)}]$; c_2 between $U_{(i)}$ and

$E(U_{(i)}) = i_n$; c_3 between i_n and u,

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

$$R_{2i} = (U_{(i)} - i_n)H^{(2)}(c_2) + \frac{U_{(i+1)} - U_{(i)}}{2}H^{(2)}(c_1),$$

$$R_{3i} = \frac{H^{(1)}(c_3)^{-1/2}H^{(2)}(c_3)}{4}\left[(U_{(i+1)} - i_n^+) + (U_{(i)} - i_n) + n^{-1}\right].$$

We rewrite $\mathrm{HD}(f_n, f)$ in (5) in terms of Z_{lj} and W_{lj} for $l = 1, 2, 3$ as

$$\mathrm{HD}(f_n, f) = 2 - 2\sum_j \frac{(Z_{1j} + W_{1j})(Z_{3j} + W_{3j})}{(Z_{2j} + W_{2j})} \Bigg/ \left(\frac{s_n}{n}\right)^{1/2}\left[\sum_j \frac{(Z_{1j} + W_{1j})^2}{Z_{2j} + W_{2j}}\right]^{1/2}. \tag{6}$$

From Lemma A.1 in the Appendix the maximal differences between Z_{lj} and their expected values $E(Z_{lj})$ are

$$\max_{1 \le n_j < n_{j+1} \le n-1} \left|Z_{lj} - E(Z_{lj})\right| = O_p(n^{\varepsilon - 1/2}k^{-\varepsilon/2 - 1/2}).$$

The maximal orders of magnitude for W_{lj} are easily obtained as

$$\max_{1 \le n_j < n_{j+1} \le n-1} W_{lj} = O_p(n^{-1/2}k^{-1}\log n).$$

We write $Z_{lj} + W_{lj}$ in (6) with random variables $\xi_{lj} = O_p(n^{\varepsilon}k^{-\varepsilon/2})$ as

$$Z_{lj} + W_{lj} = E(Z_{lj} + W_{lj}) + \xi_{lj}(nk)^{-1/2}, \tag{7}$$

$$E(\xi_{lj}) = 0. \tag{8}$$

Abbreviate $E_l = E(Z_{lj} + W_{lj})$ and use (7), then $\mathrm{HD}(f_n, f)$ in (6) is

$$\mathrm{HD}(f_n, f) = 2 - 2\sum_j N(j) \Bigg/ \left(\frac{s_n}{n}\right)^{1/2}\left[\sum_j D(j)\right]^{1/2}, \tag{9}$$

where

$$N(j) = \frac{\left[E_1 + \xi_{1j}(nk)^{-1/2}\right]\left[E_3 + \xi_{3j}(nk)^{-1/2}\right]}{E_2 + \xi_{2j}(nk)^{-1/2}},$$

$$D(j) = \frac{\left[E_1 + \xi_{1j}(nk)^{-1/2}\right]^2}{E_2 + \xi_{2j}(nk)^{-1/2}}.$$

We rewrite $\Sigma_j N(j)$ and $\Sigma_j D(j)$ in (9) in the following form:

$$\sum_j N(j) = E_N + \sum_j N_0(j) + \sum_j N_1(j) + O_p(n^{3\varepsilon-3/2}k^{-3\varepsilon/2+1/2}),$$

$$\sum_j D(j) = E_D + \sum_j D_0(j) + \sum_j D_1(j) + O_p(n^{3\varepsilon-3/2}k^{-3\varepsilon/2+1/2}),$$

where $E_N = \Sigma_j[E_1E_3/E_2]$ and $E_D = \Sigma_j[E_1^2/E_2]$ are $O(1)$; $N_0(j)$ and $D_0(j)$ with means 0 from (8); $N_1(j)$ and $D_1(j)$ with means to be estimated,

$$N_0(j) = \left[\frac{E_3}{E_2}\xi_{1j} - \frac{E_1E_3}{E_2^2}\xi_{2j} + \frac{E_1}{E_2}\xi_{3j}\right](nk)^{-1/2},$$

$$D_0(j) = \left[\frac{2E_1}{E_2}\xi_{1j} - \frac{E_1^2}{E_2^2}\xi_{2j}\right](nk)^{-1/2},$$

$$N_1(j) = \left[\frac{\xi_{1j}\xi_{3j}}{E_2} - \frac{E_3}{E_2^2}\xi_{1j}\xi_{2j} - \frac{E_1}{E_2^2}\xi_{2j}\xi_{3j} + \frac{E_1E_3}{E_2^3}\xi_{2j}^2\right](nk)^{-1},$$

$$D_1(j) = \left[\frac{\xi_{1j}^2}{E_2} - \frac{2E_1}{E_2^2}\xi_{1j}\xi_{2j} + \frac{E_1^2}{E_2^3}\xi_{2j}^2\right](nk)^{-1}.$$

We calculate $\mathrm{HD}(f_n, f)$ in (9) in terms of E_N, E_D, $N_m(j)$ and $D_m(j)$ as

$$\begin{aligned}\mathrm{HD}(f_n, f) = 2 - 2E_NE_D^{-1/2}\Bigg[1 &+ \frac{1}{E_N}\sum_j N_0(j) - \frac{1}{2E_D}\sum_j D_0(j) - \frac{1}{2}\eta_2 n^{-1/2} \\ &+ \frac{1}{E_N}\sum_j N_1(j) - \frac{1}{2E_D}\sum_j D_1(j) + \frac{3}{8E_D^2}\left[\sum_j D_0(j)\right]^2 \\ &- \frac{1}{2E_NE_D}\sum_j N_0(j)\sum_j D_0(j) - \frac{\eta_2 n^{-1/2}}{2E_N}\sum_j N_0(j) \\ &+ \frac{\eta_2 n^{-1/2}}{4E_D}\sum_j D_0(j) + O_p(n^{3\varepsilon-3/2}k^{-3\varepsilon/2+3/2})\Bigg],\end{aligned} \tag{10}$$

where η_2 is $\Sigma_{i=1}^n(e_i - 1)/n$; e_i are independent exponentials with mean 1. We shall first evaluate means of the components in (10) separately and later put them together to obtain the mean Hellinger distance. From Lemmas A.2–A.4 in the Appendix, the quantities E_N and E_D are computed as

$$E_N = \frac{\pi^{1/2}}{2}\left[1 - \frac{1}{48}\sum_j (\Delta_j N)^3\frac{H^{(2)}(N_j)^2}{H^{(1)}(N_j)^2}\right] + O(n^{-1}) + O(k^{-3}), \tag{11}$$

$$E_D = \frac{\pi}{4}\left[1 - \frac{1}{48}\sum_j (\Delta_j N)^3\frac{H^{(2)}(N_j)^2}{H^{(1)}(N_j)^2}\right] + O(n^{-1}) + O(k^{-3}). \tag{12}$$

From (11), (12), Lemmas A.2–A.4, we obtain

$$E_N E_D^{-1/2} = 1 - \frac{1}{96}\sum_j (\Delta_j N)^3 \frac{H^{(2)}(N_j)^2}{H^{(1)}(N_j)^2} + O(n^{-1}) + O(k^{-3}). \tag{13}$$

From (7) and (8), the expected value of $\xi_{lj}\xi_{mj}$ for $l \neq m$, $l, m = 1, 2, 3$ is

$$E(\xi_{lj}\xi_{mj}) = nk\,\mathrm{Cov}(Z_{lj} + W_{lj}, Z_{mj} + W_{mj}). \tag{14}$$

From conditions A1′–A4′, the means, variances and covariances of the remainder terms W_{lj} are easily obtained as

$$\begin{aligned}
&\max_{1\le n_j < n_{j+1} \le n-1} \left|E(W_{lj})\right| = O\big((nk)^{-1}\big),\\
&\max_{1\le n_j < n_{j+1} \le n-1} V(W_{lj}) = O(n^{-1}k^{-2}),\\
&\max_{1\le n_j < n_{j+1} \le n-1} \mathrm{Cov}(Z_{lj}, W_{mj}) \le O(n^{-1}k^{-3/2}),\\
&\max_{1\le n_j < n_{j+1} \le n-1} \mathrm{Cov}(W_{lj}, W_{mj}) \le O(n^{-1}k^{-2}).
\end{aligned} \tag{15}$$

From definition of $N_1(j)$ and (14) the expected value of $N_1(j)$ is

$$\begin{aligned}
E[N_1(j)] = {} & \frac{\mathrm{Cov}(Z_{1j}, Z_{3j})}{E(Z_{2j})} - \frac{E(Z_{3j})\mathrm{Cov}(Z_{1j}, Z_{2j})}{E^2(Z_{2j})}\\
& - \frac{E(Z_{1j})\mathrm{Cov}(Z_{2j}, Z_{3j})}{E^2(Z_{2j})} + \frac{E(Z_{1j})E(Z_{3j})V(Z_{2j})}{E^3(Z_{2j})},
\end{aligned}$$

neglecting terms in W_{lj} in view of (15), and Lemmas A.2 and A.3. Evaluating each component in $E[N_1(j)]$ separately and applying Lemma A.4 obtains

$$\begin{aligned}
E[N_1(j)] &= \frac{1}{n}\frac{\pi^{1/2}}{64}(\Delta_j N)^2 \frac{H^{(2)}(N_j)^2}{H^{(1)}(N_j)^2} + O(n^{-1}k^{-1/2})\\
&= O(n^{-1}k^{-1/2}).
\end{aligned} \tag{16}$$

From definition of $D_1(j)$ and (14) the expected value of $D_1(j)$ is

$$E[D_1(j)] = \frac{V(Z_{1j})}{E(Z_{2j})} - \frac{2E(Z_{1j})\mathrm{Cov}(Z_{1j}, Z_{2j})}{E^2(Z_{2j})} + \frac{E^2(Z_{1j})V(Z_{2j})}{E^3(Z_{2j})},$$

neglecting terms in W_{lj} again for the same reason as in $E[N_1(j)]$. Evaluating each component in $E[D_1(j)]$ separately and applying Lemma A.4 obtains

$$\begin{aligned}
E[D_1(j)] &= \frac{1}{n}\left[\left(1 - \frac{\pi}{4}\right) + \frac{\pi}{96}(\Delta_j N)^2 \frac{H^{(2)}(N_j)^2}{H^{(1)}(N_j)^2}\right] + O(n^{-1}k^{-1/2})\\
&= \frac{1}{n}\left(1 - \frac{\pi}{4}\right) + O(n^{-1}k^{-1/2}).
\end{aligned} \tag{17}$$

Expected values of the rest are easily obtained by

$$
\begin{aligned}
E\left[\sum_j D_0(j)\right]^2 &= O(n^{-1}),\\
E\left[\eta_2 n^{-1/2}\sum_j N_0(j)\right] &= O(n^{-3/2}),\\
E\left[\sum_j \frac{\xi_{2j}^3(nk)^{-3/2}}{E_2^3}\right] &= O(n^{-2}k^3).
\end{aligned}
\tag{18}
$$

The last expectation in (18) is on the terms of order $O_p(n^{3\varepsilon-3/2}k^{-3\varepsilon/2+3/2})$ in (10). The terms $\Sigma_j D_0(j)\Sigma_j N_0(j)$ and $\eta_2 n^{-1/2}\Sigma_j D_0(j)$ can be handled similarly. From (11) and (12), dividing (13), (16), (17) and (18) by $E_N^\alpha E_D^\beta$ where $\alpha, \beta = 0, 1, 2$ does not alter their orders of magnitude. Proposition 1 follows by putting these equations together in (10). □

PROOF OF PROPOSITION 2. Since $H^{(2)}$ is continuous, $d(f_n, f)$ in (3) is

$$
\begin{aligned}
d(f_n, f) = \frac{1}{48}\sum_{m=0}^{M-1}\sum_{m\delta\le N_j<(m+1)\delta}(\Delta_j N)^3\frac{H^{(2)}(m\delta)^2}{H^{(1)}(m\delta)^2}\\
+\left(\frac{4}{\pi}-1\right)\frac{k}{n}+O(\delta k^{-2})+O(n^{-1}k^{1/2}).
\end{aligned}
\tag{19}
$$

The first term in (19) is of $O(k^{-2})$; the second term is $O(n^{-1}k)$; asymptotically these two terms are larger than the other terms, provided that δ is small. Let $d'(f_n, f)$ be the first two terms in (19). First we minimize $d'(f_n, f)$ with respect to $\Delta_j N$ subject to the constraint that there are k_m cutpoints in the mth subinterval $[m\delta,(m+1)\delta)$. Since $[H^{(2)}(m\delta)/H^{(1)}(m\delta)]^2 \ge 0$, we only minimize $\Sigma_{m\delta\le N_j<(m+1)\delta}(\Delta_j N)^3$ with respect $\Delta_j N$ subject to the constraint that $\Sigma_{m\delta\le N_j<(m+1)\delta}\Delta_j N = \delta$ to obtain

$$
\min_{(\Delta_j N)} d'(f_n, f) = \frac{1}{48}\sum_{m=0}^{M-1}\frac{\delta^3}{k_m^2}\frac{H^{(2)}(m\delta)^2}{H^{(1)}(m\delta)^2}+\left(\frac{4}{\pi}-1\right)\frac{k}{n}.
$$

Second, we minimize $\min_{(\Delta_j N)} d'(f_n, f)$ with respect to k_m subject to the constraint $\Sigma_{m=0}^{M-1}k_m = K$ to obtain

$$
\begin{aligned}
\min_{(\Delta_j N, k_m)} d'(f_n, f) &= \frac{1}{48}\left[\sum_{m=0}^{M-1}\delta\left[\frac{H^{(2)}(m\delta)}{H^{(1)}(m\delta)}\right]^{2/3}\right]^3 k^{-2}+\left(\frac{4}{\pi}-1\right)\frac{k}{n}\\
&= \frac{1}{48}\left[\int_0^1\left[\frac{H^{(2)}(u)}{H^{(1)}(u)}\right]^{2/3}du\right]^3 k^{-2}+\left(\frac{4}{\pi}-1\right)\frac{k}{n}+O(\delta k^{-2}).
\end{aligned}
$$

The minimum is achieved at

$$k_m = \frac{\left[H^{(2)}(m\delta)/H^{(1)}(m\delta)\right]^{2/3}}{\sum_{m=0}^{M-1}\left[H^{(2)}(m\delta)/H^{(1)}(m\delta)\right]^{2/3}}k.$$

Finally we minimize $\min_{(\Delta_j N, k_m)} d'(f_n, f)$ with respect to k to obtain

$$\hat{k} = \left[\frac{\pi}{24(4-\pi)}\right]^{1/3}\int_0^1\left[\frac{H^{(2)}(u)}{H^{(1)}(u)}\right]^{2/3} du\, n^{1/3}.$$

With this $\hat{k}$, we have

$$\min_{(\Delta_j N, k_m, k)} d'(f_n, f) = \left[\frac{3(4-\pi)}{8\pi}\right]^{2/3}\int_0^1\left[\frac{H^{(2)}(u)}{H^{(1)}(u)}\right]^{2/3} du\, n^{-2/3}.$$

Since we have neglected smaller order terms, the result follows as $n \to \infty$ and $\delta \to 0$. □

APPENDIX

LEMMA A.1. *For Z_{lj}, $l = 1, 2$ and 3,*

$$\max_{1 \le n_j < n_{j+1} \le n} \left|Z_{lj} - E(Z_{lj})\right| = O_p(n^{\varepsilon - 1/2}k^{-\varepsilon/2 - 1/2}).$$

PROOF. The Markov inequality for an integer M and a real $A > 0$ is

$$\Pr\left\{\left|Z_{2j} - E(Z_{2j})\right| \ge A\right\} \le \frac{1}{A^{2M}}E\left|Z_{2j} - E(Z_{2j})\right|^{2M}. \tag{20}$$

For independent random variables $Y_1, \ldots, Y_n$ with mean zero and for an integer $M > 0$, the following inequality holds [Whittle (1960)]:

$$E\left|\sum_{i=1}^n Y_i\right|^{2M} \le \frac{2^{3M}}{\pi^{1/2}}\Gamma\left(\frac{2M+1}{2}\right)\left[\sum_{i=1}^n \left(E|Y_i|^{2M}\right)^{1/M}\right]^M. \tag{21}$$

Apply (21) to $Z_{2j} - E(Z_{2j})$ with A2′ and C1, then (20) is

$$\Pr\left\{\left|Z_{2j} - E(Z_{2j})\right| \ge A\right\} < \frac{1}{A^{2M}}\kappa'(M)n^{-M}k^{-M},$$

where

$$\kappa'(M) = \frac{2^{3M}}{\pi^{1/2}}\Gamma\left(\frac{2M+1}{2}\right)M_1^{2M}\Gamma(2M+1)(C^0)^M.$$

We obtain $1 + [l^- C_0 n]/k \le n_l \le 1 + [l^- C^0 n]/k$ by applying the constraint in C1 repeatedly from $j = 1$ through $l^- = l - 1$. Hence for any l, there are at most $l^-(C^0 - C_0)n/k$ possible choices for n_l. Given this n_l, there are at most $(C^0 - C_0)n/k$ possible choices for n_{l+1}. Since $l = 2, \ldots, k$, there are at most

$(C^0 - C_0)^2 n^2/k$ possible choices for a pair (n_j, n_{j+1}) and

$$\Pr\left\{\max_{1\le n_j<n_{j+1}\le n}\left|Z_{2j}-E(Z_{2j})\right|\ge A\right\}\le\frac{(C^0-C_0)^2n^2}{k}\Pr\left\{\left|Z_{2j}-E(Z_{2j})\right|\ge A\right\}.$$

Hence for $\varepsilon = 1/M$,

$$\max_{1\le n_j<n_{j+1}\le n}\left|Z_{2j}-E(Z_{2j})\right| = O_p\left(n^{\varepsilon-1/2}k^{-\varepsilon/2-1/2}\right).$$

Apply the same procedure for $i = 1$ and 3. □

LEMMA A.2. *For* $|a| \le |x|/2$, $|b| \le |y|/2$, $|c| \le |z|/2$, $x \ne 0$, $y \ne 0$ *and* $z \ne 0$,

$$\left|\frac{(x+a)(y+b)}{(z+c)}-\frac{xy}{z}\right|\le\left|\frac{xy}{z}\right|3\left[\left|\frac{a}{x}\right|+\left|\frac{b}{y}\right|+\left|\frac{c}{z}\right|\right].$$

LEMMA A.3. *For* $|a| \le |x|/2$, $|b| \le |y|/2$, $x \ne 0$ *and* $y \ne 0$,

$$\left|\frac{x+a}{y+b}-\frac{x}{y}\right|\le\left|\frac{x}{y}\right|2\left[\left|\frac{a}{x}\right|+\left|\frac{b}{y}\right|\right].$$

PROOF. Algebra. □

LEMMA A.4. *For* $\alpha, \beta, \gamma, \delta \in \mathscr{R}$,

$$\frac{\left[\Sigma_i H^{(1)}(i_n)^{1/2}\right]^\alpha\left[\Sigma_i H^{(1)}(i_n)^{3/2}\right]^\beta\left[\Sigma_i H^{(1)}(i_n)^2\right]^\gamma}{\left[\Sigma_i H^{(1)}(i_n)\right]^\delta}$$

$$= H^{(1)}(N_j)^{(\alpha+3\beta+4\gamma-2\delta)/2}(n_j-n_{j+1})^{\alpha+\beta+\gamma-\delta}\left[1+\frac{\alpha+3\beta+4\gamma-2\delta}{4}\right.$$

$$\times\left[\left(\Delta_j N-\frac{1}{n}\right)\frac{H^{(2)}(N_j)}{H^{(1)}(N_j)}+\frac{1}{n_{j+1}-n_j}\sum_i(i_n-N_j)^2\frac{H^{(3)}(c_{ij})}{H^{(1)}(N_j)}\right]$$

$$\left.+\Theta(\alpha,\beta,\gamma,\delta)(\Delta_j N)^2\frac{H^{(2)}(N_j)^2}{H^{(1)}(N_j)^2}+O\left((nk)^{-1}\right)+O(k^{-3})\right],$$

where

$$\Theta(\alpha,\beta,\gamma,\delta)=\frac{\alpha(3\alpha-7)+3\beta(9\beta-5)+16\gamma(3\gamma-1)+12\delta(\delta+1)}{96}$$

$$+\frac{\alpha(3\beta-2\delta)+6\beta(2\gamma-\delta)+4\gamma(\alpha-2\delta)}{16}.$$

PROOF. Calculus. □

Acknowledgments. Major parts of this research were carried out at Yale University as part of the requirements for a doctorate degree in statistics. The author wishes to thank Professor John A. Hartigan for his insightful comments and continual support throughout the course of this research. He also wishes to thank a referee for his helpful comments on an earlier version of the paper.

REFERENCES

FREEDMAN, D. and DIACONIS, P. (1981). On the histogram as a density estimator: L_2 theory. *Z. Wahrsch. Verw. Gebiete* **57** 453–476.

KANAZAWA, Y. (1988a). An optimal variable cell histogram. *Comm. Statist. Theory Methods* **17** 1401–1422.

KANAZAWA, Y. (1988b). An optimal variable cell histogram based on the sample spacings. Ph.D. dissertation, Yale Univ.

KOGURE, A. (1987). Asymptotically optimal cells for a histogram. *Ann. Statist.* **15** 1023–1030.

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. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (L. M. Le Cam and R. A. Olshen, eds.) **2** 513–520. Wadsworth, Monterey, Calif.

WHITTLE, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory Probab. Appl.* **5** 302–305.

INSTITUTE OF SOCIO-ECONOMIC PLANNING
UNIVERSITY OF TSUKUBA
TSUKUBA, IBARAKI 305
JAPAN