The Annals of Statistics
1987, Vol. 15, No. 3, 1023–1030

ASYMPTOTICALLY OPTIMAL CELLS FOR A HISTOGRAM

BY ATSUYUKI KOGURE

Fukushima University

The purpose of this paper is to examine the properties of the histogram when the cells are allowed to be arbitrary. Given a random sample from an unknown probability density f on I, we wish to construct a histogram. Any partition of I can be used as cells. The optimal partition minimizes the mean integrated squared error (MISE) of the histogram from f. An expression is found for the infimum of MISE over all partitions. It is proved that the infimum is attained asymptotically by minimizing MISE over a class of partitions of locally equisized cells.

1. Introduction. Let f be a continuous probability density on an interval I in the real line. I includes its endpoints if it is finite. Given a random sample $X_1, X_2, \ldots, X_n$ of size n from f, we wish to construct a histogram. Prior to the actual drawing of a histogram we need to choose cells into which the X_i's are grouped. Any partition of I can be used as cells. (We call a collection $\{C_j\}$ of intervals of I a partition of I if $C_i \cap C_j = \varnothing$ whenever $i \neq j$ and if $I = \bigcup_j C_j$.) When a partition $Q = \{C_j\}$ is chosen, the histogram on C_j is

$$H_n(x|Q) := F_n(C_j)/|C_j|,$$

where $|C_j|$ denotes the width of cell C_j and

$$F_n(C_j) := n^{-1} \sum_{i=1}^{n} \{X_i \in C_j\}.$$

Considering the histogram an estimate of f, we adopt the mean integrated squared error,

$$\text{MISE}(n, Q) := E\left[\int_I \left(H_n(x|Q) - f(x)\right)^2 dx\right], \tag{1.1}$$

as the risk is using Q. Here E denotes the expectation with respect to the X_i's. The risk is decomposed into two components as

$$\text{MISE}(n, Q) := \frac{1}{n} \sum_j \frac{F(C_j)\left(1 - F(C_j)\right)}{|C_j|} + \int_I \left(f_Q(x) - f(x)\right)^2 dx, \tag{1.2}$$

where $f_Q(x) := E[H_n(x|Q)]$. The first component on the right of (1.2) represents the sampling variability and the second the bias. Shrinking cell sizes decreases the bias at the cost of increasing the sampling variability. The optimal partition is to minimize MISE by compromising the conflict between the two components.

Received April 1986; revised January 1987.
AMS 1980 *subject classifications.* Primary 62G05; secondary 62E20.
Key words and phrases. Histogram, cells, partition, mean integrated squared error.

In this paper we examine the properties of the histogram when the cells are allowed to be arbitrary. We ask how close, in terms of the MISE, the histogram comes to f. Several authors—including Tapia and Thompson (1978), Scott (1979) and Freedman and Diaconis (1981)—have investigated this problem. However, they deal with the subject under the restrictive setting that all cells in a partition have a common cell width, i.e., for all j, $|C_j| = h$ for some $h > 0$. We pursue this problem with no restrictions on cells. We derive an asymptotic expression for the infimum of MISE over all partitions and show that the infimum can be achieved by a "partition of locally equisized cells."

We analyze the problem under the following conditions on f:

(C.1) f is twice continuously differentiable;
(C.2) f, f' and f'' all belong to L_2;
(C.3) $\int |f'(x)f(x)|^{2/3}\,dx > 0$.

An important class of densities excluded by the conditions is that of step functions. For any such density, the optimal histogram is the density itself.

2. Asymptotic behaviour of MISE. How small can MISE be made as the sample size tends to infinity? Let $\mathscr{P}$ denote the class of all partitions of I. We have

THEOREM 1. *Let* (C.1) *and* (C.2) *be satisfied. Then*

$$\inf_{Q\in\mathscr{P}} \mathrm{MISE}(n, Q) = \left[\left(\tfrac{3}{2}\right)\left(\tfrac{1}{6}\right)^{1/3} \int_I |f'(x)f(x)|^{2/3}\,dx\right] n^{-2/3} + o(n^{-2/3}). \tag{2.1}$$

Under the restrictive setting that all cells in a partition have a common width h, Freedman and Diaconis (1981) obtained a relation parallel to (2.1):

$$\inf_{h>0} \mathrm{MISE}(n, h) = \left[\left(\tfrac{3}{2}\right)\left(\tfrac{1}{6}\right)^{1/3} \left\{\int_I (f'(x))^2\,dx\right\}^{1/3}\right] n^{-2/3} + o(n^{-2/3}), \tag{2.2}$$

where $\mathrm{MISE}(n, h)$ is the MISE of the histogram with common cell width h. See also Scott (1979). The ratio of the leading term of (2.1) to that of (2.2) is $\int_I |f'(x)f(x)|^{2/3}\,dx / \{\int_I (f'(x))^2\,dx\}^{1/3}$, which is seen to be no more than 1 by Hölder's inequality. It is approximately 0.89 for the normal density and 0.24 for the Cauchy density.

Theorem 1 will be derived by proving the following two propositions. The first of these gives an asymptotic lower bound for the MISE.

PROPOSITION 1. *Under* (C.1) *and* (C.2) *we have*

$$\liminf_{n\to\infty} n^{2/3} \inf_{Q\in\mathscr{P}} \{\mathrm{MISE}(n, Q)\} \geq \left(\tfrac{3}{2}\right) 6^{-1/3} \int_I |f'(x)f(x)|^{2/3}\,dx. \tag{2.3}$$

All we need then is to find a partition whose MISE behaves like the lower bound as n increases. Our scheme for getting such a partition is based on three

observations:

(i) on a small interval f is well approximated by a linear function;

(ii) the bias term $\int(f_Q(x) - f(x))^2\,dx$ of MISE(n, Q) is minimized for a partition of equisized cells if f is linear; and

(iii) if f does not change much over the region, then the magnitude of the sampling variation $(1/n)\Sigma_j F(C_j)(1 - F(C_j))/|C_j|$ is mostly determined by the number of cells, regardless of the cell sizes being equal or unequal.

With these observations we might expect that the desired partition can be found in a class of partitions of locally equisized cells (POLEC). A POLEC is constructed by "dividing I twice," a process separated into two steps:

Step 1. Choose a reference point $x_0 \in I$ and a mesh width $w > 0$ to obtain an equally spaced net $\{\Delta_i | i \in Z\}$, where Z is a set of integers and

$$\Delta_i := [x_0 + (i-1)w, x_0 + iw), \qquad \bigcup_{i \in Z} \Delta_i = I. \tag{2.4}$$

Step 2. For each $i \in Z$ choose a positive integer k_i and divide Δ_i into k_i cells, allowing different k_i's for different Δ_i's. Put $\mathbf{k} := \{k_i | i \in Z\}$ and let $Q(w, \mathbf{k})$ denote the resultant POLEC.

Let $\{w_n\}$ and $\{U_n\}$ be sequences of positive numbers such that

$$n^{-1/3} \ll w_n \ll n^{-2/9}, \qquad U_n \ll n^{1/3}, \qquad (w_n/U_n) \ll n^{-1/3}, \tag{2.5}$$

where $a_n \ll b_n$ means $\limsup_{n\to\infty}(a_n/b_n) = 0$. Define a class of POLECs as

$$\mathscr{P}(w_n, U_n) := \{Q(w_n, \mathbf{k}) | k_i \le U_n \text{ for all } i \in Z\}. \tag{2.6}$$

Then we have

PROPOSITION 2. *Under conditions* (C.1) *and* (C.2) *we have*

$$\inf\{\text{MISE}(n, Q): Q \in \mathscr{P}(w_n, U_n)\} = \left(\tfrac{3}{2}\right)6^{-1/3}\int_I |f'(x)f(x)|^{2/3}\,dx\,n^{-2/3} + o(n^{-2/3}). \tag{2.7}$$

Clearly, Propositions 1 and 2 together imply Theorem 1. The proofs of the propositions occupy the next section.

Another implication of the two propositions is

$$\inf\{\text{MISE}(n, Q): Q \in \mathscr{P}(w_n, U_n)\} = \inf_{Q \in \mathscr{P}}\{\text{MISE}(n, Q)\} + o(n^{-2/3}).$$

This means that there is a POLEC whose associated MISE is nearly as small as that of the optimal partition as the sample size grows large. This may call forth some interest in developing a data-driven rule for choosing a POLEC. Some discussion on this can be found in Kogure (1986). For the related subject of the data-based choice of a cell width, see Scott (1979), Rudemo (1982), Chow, Geman and Wu (1983) and Stone (1985).

3. Proofs.

PROOF OF PROPOSITION 1. For each interval C let

$$\gamma(C) := \left(\tfrac{3}{2}\right)6^{-1/3}\int_C |f'(x)f(x)|^{2/3}\,dx$$

and for each $Q = \{C_j\} \in \mathscr{P}$ let $\alpha_n(Q) := n^{2/3}\mathrm{MISE}(n, Q) - \gamma(I)$.

Put $h_n := (n^{2/9}\log n)^{-1}$ and define a subclass $\mathscr{P}_n$ of $\mathscr{P}$ as

$$\mathscr{P}_n := \left\{Q = \{C_j\} \mid 0 < |C_j| \le h_n \text{ for all } j\right\}.$$

For each interval C, $0 < |C| < \infty$, let

$$\delta_n(C) := \left(\tfrac{1}{12}\right)\left(n^{1/3}|C|\right)^2 + \int_C \left(f'(x)\right)^2 dx + \left(n^{1/3}|C|\right)^{-1}F(C)$$

and for each $Q = \{C_j\} \in \mathscr{P}_n$ let $\beta_n(Q) := \sum_j(\delta_n(C_j) - \gamma(C_j))$. Observe that

$$\begin{aligned}\inf_{Q\in\mathscr{P}} \alpha_n(Q) \ge{}& \inf_{Q\in\mathscr{P}_n} \beta_n(Q) + \inf_{Q\in\mathscr{P}_n}\left(\alpha_n(Q) - \beta_n(Q)\right)\\ &+\left(\inf_{Q\in\mathscr{P}} \alpha_n(Q) - \inf_{Q\in\mathscr{P}_n} \alpha_n(Q)\right).\end{aligned} \tag{3.1}$$

Then it suffices to show that each term on the right of (3.1) has a nonnegative lower limit. First, we will show

$$\liminf_{n\to\infty} \inf_{Q\in\mathscr{P}_n} \beta_n(Q) \ge 0. \tag{3.2}$$

Fix $Q = \{C_j\} \in \mathscr{P}_n$. For each $C_j \in Q$

$$\begin{aligned}\delta_n(C_j) &\ge \left(\tfrac{3}{2}\right)6^{-1/3}\left\{\int_{C_j}\left(f'(x)\right)^2 dx\right\}^{1/3}\left(F(C_j)\right)^{2/3}\\ &\ge \left(\tfrac{3}{2}\right)6^{-1/3}\int_{C_j}|f'(x)f(x)|^{2/3}\,dx \quad \text{(by Hölder's inequality)}\\ &= \gamma(C_j).\end{aligned}$$

Thus, $\beta_n(Q) \ge 0$, which implies (3.2).

Second, we will show

$$\liminf_{n\to\infty} \inf_{Q\in\mathscr{P}_n} \left(\alpha_n(Q) - \beta_n(Q)\right) \ge 0. \tag{3.3}$$

Fix $Q = \{C_j\} \in \mathscr{P}_n$. Then

$$\begin{aligned}\alpha_n(Q) - \beta_n(Q) &= n^{2/3}\mathrm{MISE}(n, Q) - \sum_j \delta_n(C_j)\\ &= n^{2/3}\left\{\int_I \left(f_Q(x) - f(x)\right)^2 dx - \left(\tfrac{1}{12}\right)\sum_j |C_j|^2\int_{C_j}\left(f'(x)\right)^2 dx\right\}\\ &\quad - n^{1/3}\sum_j \left(F(C_j)\right)^2/|C_j|.\end{aligned} \tag{3.4}$$

It is easy to see that for each j there is a point x_j inside c_j such that

$$\int_{C_j}\left(f_Q(x) - f(x)\right)^2 dx \ge \left(\tfrac{1}{12}\right)\left(f'(x_j)\right)^2|C_j|^3.$$

Then, by Lemma 2.22 of Freedman and Diaconis (1981), we have

$$\left| \left(f'(x_j)\right)^2 |C_j| - \int_{C_j} \left(f'(x)\right)^2 dx \right| \le 2|C_j| \int_{C_j} |f'(x) f''(x)|\, dx.$$

Thus,

$$\begin{aligned} &\int_I \left(f_Q(x) - f(x)\right)^2 dx - \left(\tfrac{1}{12}\right) \sum_j |C_j|^2 \int_{C_j} \left(f'(x)\right)^2 dx \\ &\qquad \ge -2h_n^3 \int_I |f'(x) f''(x)|\, dx. \end{aligned} \tag{3.5}$$

Note that

$$\sum_j \left(F(C_j)\right)^2 / |C_j| = \int_I \left(f_Q(x)\right)^2 dx \le \int_I \left(f(x)\right)^2 dx. \tag{3.6}$$

Combining (3.4), (3.5) and (3.6) and recalling that $h_n = (n^{2/9} \log n)^{-1}$ we have (3.3).

Lastly, we will show

$$\liminf_{n \to \infty} \left\{ \inf_{Q \in \mathscr{P}} \alpha_n(Q) - \inf_{Q \in \mathscr{P}_n} \alpha_n(Q) \right\} \ge 0. \tag{3.7}$$

It would suffice to show that for each $Q \in \mathscr{P}$, there is $Q^0 \in \mathscr{P}_n$ such that for all sufficiently large n

$$\alpha_n(Q) - \alpha_n(Q^0) \ge r_n, \tag{3.8}$$

where $\{r_n\}$ is a sequence of real numbers such that $\liminf_{n \to \infty} r_n = 0$. Because then for all sufficiently large n: $\inf_{Q \in \mathscr{P}} \alpha_n(Q) - \inf_{Q \in \mathscr{P}_n} \alpha_n(Q) \ge r_n - n^{-1}$. We construct such Q^0 as follows. Fix $C \in Q$. If $|C| \le h_n$, then let C be included in Q^0. If $|C| > h_n$, then divide C equally into subintervals of width h_n^*. Set h_n^* equal to h_n if $|C| = \infty$. If $|C| < \infty$, then there is a positive integer $m \ge 2$ such that $h_n(m-1) \le |C| < h_n m$. Set h_n^* equal to $|C|/m$ in this case. Let the subintervals thus made be included in Q^0. Repeat this for each $C \in Q$. Then we have

$$\begin{aligned} &\alpha_n(Q) - \alpha_n(Q^0) \\ &\quad = n^{2/3} \Bigg\{ \int_I \left(f_Q(x) - f(x)\right)^2 dx - \int_I \left(f_{Q^0}(x) - f(x)\right)^2 dx \\ &\qquad + (1/n) \Bigg(\sum_{C \in Q,\, C \in Q^0} F(C)(1 - F(C))/|C| \\ &\qquad\qquad - \sum_{C \notin Q,\, C \in Q^0} F(C)(1 - F(C))/|C| \Bigg) \Bigg\}. \end{aligned} \tag{3.9}$$

Note that Q^0 is finer than Q in the sense that for each $C \in Q$ there is $C^0 \in Q^0$ such that $C^0 \subset C$. Thus

$$\int_I \left(f_Q(x) - f(x)\right)^2 dx \ge \int_I \left(f_{Q^0}(x) - f(x)\right)^2 dx. \tag{3.10}$$

Observe that

$$\sum_{C\notin Q,\, C\in Q^0} F(c)/|C| \le \Big(\inf_{C\in Q^0}|C|\Big)^{-1} \le Qh_n^{-1}. \tag{3.11}$$

Combining (3.6), (3.9), (3.10) and (3.11) we have

$$\alpha_n(Q) - \alpha_n(Q^0) \ge -n^{-1/3}\Big(2h_n^{-1} + \int_I (f(x))^2\,dx\Big).$$

This in turn implies the existence of $\{r_n\}$ in (3.8). The proof of the proposition is now complete. □

PROOF OF PROPOSITION 2. The mean value theorem implies that for each $i \in Z$ there is $x_i \in \Delta_i$ such that

$$\int_{\Delta_i} |f'(x)f(x)|^{2/3}\,dx = w_n |f'(x_i)f(x_i)|^{2/3}, \tag{3.12}$$

where Δ_i is defined as at (2.4). For each $i \in Z$ define a real-valued function $H_i(\cdot)$ on $(0, \infty)$ as

$$H_i(t) := \left(\tfrac{1}{12}\right)w_n^3 (f'(x_i))^2/t^2 + f(x_i)t/n. \tag{3.13}$$

Put $Z^0 := \{i \in Z | f(x_i) \neq 0 \text{ and } f'(x_i) \neq 0\}$. For each $i \in Z^0$ let

$$t_i^* := 6^{-1/3}\Big\{(f'(x_i))^2/f(x_i)\Big\}^{1/3} n^{1/3} w_n.$$

t_i^* minimizes $H_i(t)$ and $H_i(t_i^*) = (\frac{3}{2})6^{-1/3}\int_{\Delta_i}|f'(x)f(x)|^{2/3}\,dx\, n^{-2/3}$. By (3.12) $\int_{\Delta_i}|f'(x)f(x)|^{2/3}\,dx = 0$ for $i \in Z^0$. Thus,

$$\sum_{i\in Z^0} H_i(t_i^*) = \left(\tfrac{3}{2}\right)6^{-1/3}\int_I |f'(x)f(x)|^{2/3}\,dx\, n^{-2/3}. \tag{3.14}$$

Let $Q(w_n, \mathbf{k})$ be a partition in $\mathscr{P}(w_n, U_n)$, where $\mathscr{P}(w_n, U_n)$ is defined as (2.6). It is easy to see that

$$\mathrm{MISE}(n, Q(w_n, \mathbf{k})) = \sum_{i\in Z} H_i(k_i) + o(n^{-2/3}), \tag{3.15}$$

where the $o(n^{-2/3})$ term is bounded by a quantity which converges to zero faster than $n^{-2/3}$, uniformly in both $\{x_i\}$ and $\mathbf{k}$. In the light of (3.14) and (3.15) the conclusion follows if we verify the existence of a sequence of positive integers $\{k_i^* | i \in Z,\ k_i^* \le U_n\}$ such that

$$\sum_{i\in Z} H_i(k_i^*) - \sum_{i\in Z^0} H_i(t_i^*) = o(n^{-2/3}). \tag{3.16}$$

If both $f'(x_i)$ and $f(x_i)$ are zero, then $H_i(t) \equiv 0$. Thus, assume that at least one of $f'(x_i)$ and $f(x_i)$ is not zero. For any $i \in Z^0$, let k_i^0 be the smallest positive

integer such that $H_i(k_i^0)$ is closest to $H_i(t_i^*)$. For each $i \in Z$ define k_i^* as

$$k_i^* = \begin{cases} k_i^0 & \text{if } i \in Z \text{ and } k_i^0 \le U_n, \\ U_n - 1 & \text{if } i \in Z \text{ and } k_i^0 > U_n, \\ U_n & \text{if } f'(x_i) \ne 0 \text{ and } f(x_i) = 0, \\ 1 & \text{if } f'(x_i) = 0 \text{ and } f(x_i) \ne 0. \end{cases}$$

Now consider the four possible cases and the respective inequalities below. For the first two cases we utilize the following relation: If $i \in Z^0$, then

$$2\left(\tfrac{1}{12}\right)\left(f'(x_i)\right)^2 w_n^3 t^{-2} \gtreqless \left(f(x_i)/n\right)t, \quad \text{iff } t_i^* \gtreqless t. \tag{3.17}$$

Case 1. $i \in Z^0$ and $k_i^0 \le U_n$. By (3.17)

$$H_i(t_i^* + 1) \le \left(\tfrac{3}{2}\right)\left(f(x_i)/n\right)(t_i^* + 1)$$

and

$$H_i(t_i^*) = \left(\tfrac{3}{2}\right)\left(f(x_i)/n\right)t_i^*.$$

Thus,

$$\begin{aligned} H_i(k_i^0) - H_i(t_i^*) &\le H_i(t_i^* + 1) - H_i(t_i^*) \\ &= \left(\tfrac{3}{2}\right)\left(f(x_i)/n\right). \end{aligned} \tag{3.18}$$

Case 2. $i \in Z^0$ and $k_i^0 > U_n$. Then $t_i^* + 1 \ge U_n$. Thus,

$$H_i(U_n - 1) \le \left(\tfrac{1}{4}\right)\left(f'(x_i)\right)^2 w_n^3 (U_n - 1)^{-1}. \tag{3.19}$$

Case 3. $f'(x_i) \ne 0$ and $f(x_t) = 0$. Then

$$H_i(U_n) = \left(\tfrac{1}{12}\right)\left(f'(x_i)\right)^2 w_n^3 U_n^{-2}. \tag{3.20}$$

Case 4. $f'(x_i) = 0$ and $f(x_i) \ne 0$. Then

$$H_i(1) = f(x_i)/n. \tag{3.21}$$

Combining (3.18)–(3.21) we have

$$\begin{aligned} &\sum_{i \in Z} H_i(k_i^*) - \sum_{i \in Z^0} H_i(t_i^*) \\ &\quad \le \left(\tfrac{3}{2}\right)\left(\sum_{i \in Z} f(x_i)\right)/n + \left(\tfrac{1}{4}\right)\left(\sum_{i \in Z}\left(f'(x_i)\right)^2\right) w_n^3 (U_n - 1)^{-2}. \end{aligned} \tag{3.22}$$

By Corollary 2.24 of Freedman and Diaconis (1981) we have

$$\left| \sum_{i \in Z} f(x_i) w_n - \int_I f(x)\, dx \right| \le w_n \int_I |f'(x)|\, dx \tag{3.23}$$

and

$$\left| \sum_{i \in Z}\left(f'(x_i)\right)^2 w_n - \int_I \left(f(x)\right)^2 dx \right| \le 2w_n \int_I |f'(x) f(x)|\, dx. \tag{3.24}$$

Equation (3.22) together with (3.23) and (3.24) implies

$$\left| \sum_{i \in Z} H_i(k_i^*) - \sum_{i \in Z^0} H_i(t_i^*) \right| = o\left((nw_n)^{-1} + (w_n/U_n)^2\right)$$

$$= o(n^{-2/3}).$$

The last relation holds because of (2.5). This completes the proof of Proposition 2. □

Acknowledgments. This work is part of the author's Ph.D. thesis at Yale University, written under the supervision of Professor J. A. Hartigan, whose guidance and suggestions are gratefully acknowledged and appreciated. Thanks are also due to Professors K. Takeuchi and Y. Hosoya for helpful comments.

REFERENCES

CHOW, Y.-S., GEMAN, S. and WU, L. D. (1983). Consistent crossvalidated density estimation. *Ann. Statist.* **11** 25–38.

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

KOGURE, A. (1986). Optimal cells for a histogram. Ph.D. thesis, Yale Univ.

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. (1985). An asymptotically optimal histogram selection rule. In *Proc. 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, Belmont, Calif.

TAPIA, R. A. and THOMPSON, J. R. (1978). *Nonparametric Probability Density Estimation*. Johns Hopkins Univ. Press, Baltimore, Md.

FACULTY OF ECONOMICS
FUKUSHIMA UNIVERSITY
MATSUKAWA-MACHI, FUKUSHIMA 960-12
JAPAN