AN ASYMPTOTICALLY OPTIMAL HISTOGRAM
SELECTION RULE

BY

CHARLES J. STONE

TECHNICAL REPORT NO. 34
JUNE 1984

RESEARCH PARTIALLY SUPPORTED BY
NATIONAL SCIENCE FOUNDATION GRANT MCS83-01257

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA
AN ASYMPTOTICALLY OPTIMAL HISTOGRAM SELECTION RULE

by Charles J. Stone
University of California, Berkeley
July, 1983; Revised June, 1984

Abstract. A random sample is available from a multivariate distribution having a bounded density, which is assumed to satisfy a mild additional condition. A finite collection of histogram estimates of the unknown density is constructed, whose cardinality increases algebraically fast with respect to the size of the random sample. A histogram selection rule is introduced, which is shown to be asymptotically optimal relative to integrated squared error loss.

This research was supported in part by National Science Foundation Grant MCS83-01257.


Key words and phrases. Density estimation, histogram, selection rule, asymptotic efficiency.
1. Statement of the main result. Let $X_1, X_2, \ldots$ be independent $\mathbb{R}^d$-valued random variables having common absolutely continuous distribution $P$ with bounded density $p$. Let $P_n$ denote the empirical distribution of $X_1, \ldots, X_n$, defined by

$$P_n(A) = \frac{1}{n} \# \{i: 1 \leq i \leq n \text{ and } X_i \in A\}.$$ 

Let $\mathbb{R}_+^d$ denote the collection of $d$-tuples of positive numbers. Choose $a = (a_1, \ldots, a_d) \in \mathbb{R}^d$ and $b = (b_1, \ldots, b_d) \in \mathbb{R}_+^d$; set $h = (a, b)$. Consider the histogram estimate $p_{nh}$ of $p$ defined as follows: Let $\lambda = (\lambda_1, \ldots, \lambda_d)$ denote an arbitrary $d$-tuple of integers. Set

$$I_{h\lambda} = \prod_{j=1}^d [a_j + (\lambda_j - 1)b_j, a_j + \lambda_j b_j].$$

Each $d$-dimensional interval $I_{h\lambda}$ has volume $v_h = \prod_{j=1}^d b_j$; the collection of all such intervals forms a partition of $\mathbb{R}^d$. Finally, set

$$p_{nh} = \frac{P_n(I_{h\lambda})}{v_h} \text{ on } I_{h\lambda}.$$ 

(See page 21 of Kendall and Stuart, 1977, for a picture of a bivariate histogram based on a sample of size $n = 9,440$.) The integrated squared error loss of $p_{nh}$ as an estimate of $p$ is given by

$$L_{nh} = \int (p_{nh} - p)^2 = \frac{1}{v_h} \sum_{\lambda} P_n^2(I_{h\lambda}) - \frac{2}{v_h} \sum_{\lambda} P_n(I_{h\lambda})P(I_{h\lambda}) + \int p^2.$$

Let $H_n$ denote a finite subset of $\mathbb{R}^d \times \mathbb{R}_+^d$ whose cardinality increases algebraically fast with $n$; that is, $\lim_{n \to \infty} n^{-c}\#(H_n) = 0$ for some $c > 0$. A histogram selection rule $h_n$ is an $H_n$-valued function of $X_1, \ldots, X_n$. Clearly

$$\frac{L_{nh}}{\min_{h \in H_n} L_{nh}} \geq 1;$$
here it is understood that \( h \in H_n \). The selection rule \( h_n \) is said to be asymptotically optimal if

\[
\lim_{n} \left[ \frac{L_{nh}}{\min_{h} L_{nh}} \right] = 1 \quad \text{with probability one .}
\]

Set

\[
K_{nh} = \frac{1}{v_h} \left( \frac{2}{n} \sum_{k} p^2(I_{kh}) \right)
\]

(see Section 2 for motivation). Let \( \hat{h}_n \) be a value of \( h \) that minimizes \( K_{nh} \). It will be shown below, under a mild condition on \( p \), that the histogram selection rule \( \hat{h}_n \) is asymptotically optimal.

**CONDITION 1.** There are positive constants \( \alpha \) and \( \beta \) such that

\[
\int (p_0 - p)^2 \geq \alpha (v_h^\beta \wedge 1) \quad \text{for} \quad n \geq 1 \quad \text{and} \quad h \in H_n.
\]

Here \( s \wedge t = \min(s,t) \). Condition 1 is satisfied if, say, there is some nonempty open subset of \( \mathbb{R}^d \) on which the derivative of \( p \) exists and is continuous and nonzero. For an alternative set of assumptions which guarantees that this condition is satisfied, at least when \( d = 1 \), see Freedman and Diaconis (1981).

**THEOREM 1.** If Condition 1 holds, then \( \hat{h}_n \) is asymptotically optimal.

For other theoretical results on the selection of a histogram see Freedman and Diaconis (1981); Chow, Geman and Wu (1981, 1983); and Burman (1984). For an analogous result on kernel density estimates see Stone (1984). The latter two papers were written after the original version of this paper.
2. Motivation for $K_n$. Ideally, $h$ should be chosen to minimize

$$L_{nh} - \int_{\mathbb{R}} p^2 = \frac{1}{v_h} \sum_{i \neq i} p^2(I_{nh}^i) - \frac{2}{v_h} \sum_{i \neq i} \hat{p}_n(I_{nh}^i)p(I_{nh}^i),$$

but the quantity $p(I_{nh}^i)$ is unknown. The estimate $\hat{p}_n(I_{nh}^i)$ of $p(I_{nh}^i)$ leads to the biased estimate $\hat{p}_n^2(I_{nh}^i)$ of $\hat{p}_n(I_{nh}^i)p(I_{nh}^i)$. It is easily checked that

$$n \frac{\hat{p}_n^2(I_{nh}^i)}{n-1} - \frac{\hat{p}_n(I_{nh}^i)}{n-1}$$

is an unbiased estimate of $\hat{p}_n(I_{nh}^i)p(I_{nh}^i)$; that is,

$$E \left[ n \frac{\hat{p}_n^2(I_{nh}^i)}{n-1} - \frac{\hat{p}_n(I_{nh}^i)}{n-1} \right] = E[p_n(I_{nh}^i)p(I_{nh}^i)] = p^2(I_{nh}^i).$$

This leads to the following histogram selection rule: choose $h$ to minimize

$$K'_nh = \frac{1}{v_h} \sum_{i \neq i} \hat{p}_n^2(I_{nh}^i) - \frac{2}{v_h} \sum_{i \neq i} \left[ \frac{n-1}{n-1} \hat{p}_n^2(I_{nh}^i) - \frac{\hat{p}_n(I_{nh}^i)}{n-1} \right]$$

$$= \frac{1}{v_h} \left( \frac{2}{n-1} - \sum_{i \neq i} \hat{p}_n^2(I_{nh}^i) \right).$$

An inessential simplifying approximation leads to the formula for $K_{nh}$ given in Section 1. For an alternative motivation in terms of cross-validation see Rudemo (1982).

3. Proof of Theorem 1. Recall that $p$ is assumed to be bounded and that the cardinality of $H_n$ increases algebraically fast with $n$. Define the density $p_h$ on $\mathbb{R}^d$ by $p_h(x) = p(I_{nh}^x)/v_h$ for $x \in I_{nh}$. Set

$$G_{nh} = \frac{1}{n} \sum_{i=1}^{n} p_h(x_i) - E_p(X),$$

$$G_n = \frac{1}{n} \sum_{i=1}^{n} p(x_i) - E_p(X),$$
\[ J_{nh} = \int (p_{nh} - p)^2 + \frac{1}{nv_h}, \]

and

\[ J_{nhr} = v_h^r \wedge 1 + \frac{1}{nv_h} \quad \text{for } r > 0. \]

**Lemma 1.** If Condition 1 holds, then 
\[ \lim_{n} \max_{h} \frac{|G_{nh} - G_n|}{J_{nh}} = 0 \]
with probability one.

**Lemma 2.** For all \( r > 0 \)
\[ \lim_{n} \max_{h} \frac{1}{J_{nhr}} \left| \int (p_{nh} - p)^2 - \frac{1}{nv_h} \right| = 0 \]
with probability one.

The proofs of these two lemmas will be given at the end of the paper.

To prove that \( \hat{h}_n \) is asymptotically optimal it suffices to show that
\[ \lim_{n} \max_{h, h'} \frac{|L_{nh} - L_{nh'} - (K_{nh} - K_{nh'})|}{L_{nh} + L_{nh'}} = 0 \]
with probability one. \( (1) \)

To verify (1) it suffices to show that
\[ \inf \min_{n, h} \frac{L_{nh}}{J_{nh}} > 0 \]
with probability one. \( (2) \)

and
\[ \lim_{n} \max_{h, h'} \frac{|L_{nh} - L_{nh'} - (K_{nh} - K_{nh'})|}{J_{nh} + J_{nh'}} = 0 \]
with probability one. \( (3) \)

Observe that
\[ L_{nh} = \int (p_{nh} - p)^2 = \int (p_{nh} - p_h)^2 + \int (p_h - p)^2. \]

It now follows easily from Condition 1 and Lemma 2 that \( (2) \) holds.

By elementary algebra
\[ L_{nh} - K_n - 2G_n = \int p^2 = 2(G_{nh} - G_n) + 2\int (p_{nh} - p_h)^2 - \frac{2}{nv_h}. \]
It now follows easily from Lemma 1 and Lemma 2 that (3) holds.

Thus the proof of Theorem 1 is complete once the two lemmas are verified.

To prove Lemma 1 write

\[ G_{nh} - G_n = \frac{1}{n} \sum_{i} Z_{ih} = \overline{Z}_{nh}, \]

where

\[ Z_{ih} = p_h(X_i) - p(X_i) - E(p_h(X_i) - p(X_i)). \]

Then \( Z_{ih}, \ i \geq 1, \) are independent and identically distributed random variables having mean zero. Since \( p \) is bounded, there is a positive constant \( c \) independent of \( h \) such that \( |Z_{ih}| \leq c \) and \( \text{Var}(Z_{ih}) \leq cu_h^2, \) where

\[ u_h^2 = \int (p_h - p)^2. \]

By Bernstein's inequality (see Hoeffding, 1963)

\[ \Pr(|\overline{Z}_{nh}| > t) \leq 2 \exp[-\tau \lambda / 2(1+\lambda/3)], \]

where \( 0 \leq \lambda \leq t/u_h^2 \) and \( \tau = nt/c. \)

Choose \( \varepsilon > 0. \) Suppose that \( u_h \geq n^{\varepsilon-1/2}. \) Set \( t = n^{\varepsilon-1/2}u_h \) and \( \lambda = n^{\varepsilon-1/2}/u_h \leq 1. \) Then \( \lambda \tau = n^{2\varepsilon}/c. \) Suppose instead that \( u_h < n^{\varepsilon-1/2}. \) Set \( t = n^{2\varepsilon-1} \) and \( \lambda = 1. \) Again, \( \lambda \tau = n^{2\varepsilon}/c. \) Thus in either case it follows from Bernstein's inequality that

\[ \Pr(|\overline{Z}_{nh}| > t) \leq 2 \exp(-n^{2\varepsilon}/3c). \]

Consequently

\[ \lim_{n} \Pr(|\overline{Z}_{nh}| > n^{\varepsilon-1/2}u_h + n^{2\varepsilon-1} \text{ for some } h \in H_n) = 0. \]

Thus to verify Lemma 1 it is enough to show that for some \( \varepsilon > 0 \)

\[ \lim_{n} \max_{u > 0} \frac{n^{\varepsilon-1/2}u + n^{2\varepsilon-1}}{u^2 + 1/nu^2/\beta} = 0, \]

where \( \beta \) is from Condition 1. For \( 0 < \varepsilon < 1/2(1+\beta), \) this result is easily shown by considering separately: \( 0 < u \leq n^{\varepsilon-1/2}, n^{\varepsilon-1/2} < u < n^{-\beta/2(1+\beta)}, \) and
The simplest way to prove Lemma 2 is by means of the technique called "Poissonization." It was used by Rosenblatt (1975) in a related context.

**Lemma 3.** Let $N_\xi$ be independent Poisson random variables with mean $\lambda_\xi$ such that $0 < \lambda = \sum \lambda_\xi < \infty$. Set $N = \sum N_\xi$, $P_\xi = \lambda_\xi / \lambda$ and $\bar{P} = \max_\xi P_\xi$. For each positive integer $k$ there is a finite positive universal constant $c_k$ such that

$$E[(\sum_{\xi} (N_\xi - \bar{N}P_\xi)^2 - N)^{2k}] \leq c_k (\lambda + \lambda^k + \lambda^{2k}\bar{P})^k.$$  

This lemma follows in a straightforward manner from properties of cumulants summarized in Gnedenko and Kolmogorov (1954) or Kendall and Stuart (1977). (Observe that $E[(N-\lambda)^{2k}]$ is a polynomial in $\lambda$ of degree $k$ with zero constant term. The next step is to prove the desired conclusion with $N$ replaced by $\lambda$.)

Set $\tau = \sup p$ and $N_n(I_{h\lambda}) = nP_n(I_{h\lambda}).$

**Lemma 4.** For each positive integer $k$ there is a universal constant $c'_k$ such that $E[(\sum_{\xi} (N_n(I_{h\lambda}) - nP(I_{h\lambda}))^2 - n)^{2k}] \leq c'_k n^k (1 + (n\pi \nu_h)^k)$.

**Proof.** Let $\mu_n$ denote the $2k$th moment of

$$Z = \sum_{\xi} (N_n(I_{h\lambda}) - nP(I_{h\lambda}))^2 - n$$

and set $\mu_0 = 0$. Let $R(\lambda)$ denote the $2k$th moment of the random variable obtained through replacing $n$ in the definition of $Z$ by a Poisson number $N$ having mean $\lambda$. Then

$$R(\lambda) = \sum_{n} \Pr(N=n)\mu_n = \sum_{n} \frac{\lambda^n}{n!} e^{-\lambda} \mu_n.$$
According to Lemma 3 and the well-known connection between multinomial and independent Poisson random variables, $R(\lambda)$ is a polynomial of degree $2k$ in $\lambda$ and

$$0 \leq \sum_{j=1}^{2k} \frac{R^{(j)}(0)}{j!} \lambda^j = R(\lambda) \leq c_k(\lambda + \lambda^k + \lambda^2(\tau v_h)^k)$$

for $\lambda \geq 0$.

Thus there is a finite positive universal constant $c''_k$ such that

$$\sum_{j=1}^{2k} \frac{|R^{(j)}(0)|}{j!} \lambda^j \leq c''_k(\lambda + \lambda^k + \lambda^2(\tau v_h)^k)$$

for $\lambda \geq 0$.

(For suppose otherwise and note that for each fixed $c > 0$, if $\lambda > 0$ and

$$\sum_{j=1}^{2k} \frac{|R^{(j)}(0)|}{j!} \lambda^j \geq c_k(\lambda + \lambda^k + \lambda^2(\tau v_h)^k)$$

then

$$\sum_{j=1}^{2k} \frac{|R^{(j)}(0)|}{j!} (c\lambda)^j \geq c_k(\lambda + \lambda^k + \lambda^2(\tau v_h)^k),$$

by a compactness argument, there would then be a nonzero polynomial in $c$ of degree $2k$ that equals zero at more than $2k$ distinct points.)

Consequently,

$$\nu_n = \sum_{j=1}^{2k} \frac{n!R^{(j)}(0)}{(n-j)!j!} \leq \sum_{j=1}^{2k} \frac{|R^{(j)}(0)|}{j!} \lambda^j \leq c''_k(n + n^k + n^2(\tau v_h)^k)$$

for $n \geq 0$,

which yields the desired result.

To prove Lemma 2, observe that by Lemma 4 and Chebyshev's inequality,

$$\lim_{n \to \infty} \max_{h \in H} \frac{|\sum_{v=1}^{n} (P_n(I_{h,v}) - P(I_{h,v}))^2 - \frac{1}{n}|}{n^{\frac{1}{2}}(v^{\frac{1}{2}} + n^{-\frac{1}{2}})} = 0$$

for $\varepsilon > 0$.

Let $r > 0$ be fixed. It is easily seen that for sufficiently small $\varepsilon > 0$,

$$\lim_{n \to \infty} \max_{v \in V} \frac{n^{\varepsilon - 1}(v^{\frac{1}{2}} + n^{-\frac{1}{2}})}{v^{r} + \frac{1}{nv}} = 0.$$
The desired conclusion follows from these two observations.

REFERENCES


