STATISTICAL COMPUTING AND GRAPHICS

Data-Based Choice of Histogram Bin Width

M. P. WAND

The most important parameter of a histogram is the bin width because it controls the tradeoff between presenting a picture with too much detail ("undersmoothing") or too little detail ("oversmoothing") with respect to the true distribution. Despite this importance there has been surprisingly little research into estimation of the "optimal" bin width. Default bin widths in most common statistical packages are, at least for large samples, quite far from the optimal bin width. Rules proposed by, for example, Scott lead to better large sample performance of the histogram, but are not consistent themselves. In this paper we extend the bin width rules of Scott to those that achieve root-n rates of convergence to the L_2-optimal bin width, thereby providing firm scientific justification for their use. Moreover, the proposed rules are simple, easy and fast to compute, and perform well in simulations.

KEY WORDS: Binning; Data analysis; Density estimation; Kernel functional estimator; Smoothing parameter selection.

1. INTRODUCTION

The histogram is the oldest and most popular tool for graphical display of a univariate set of data. It is taught in virtually all elementary data analysis courses, and is available in most statistical computing packages.

An important parameter that needs to be specified when constructing a histogram is the bin width. This is simply the length of the subintervals of the real line, sometimes called "bins," on which the histogram is based. It is not very difficult to see that the choice of the bin width has an enormous effect on the appearance of the resulting histogram. The choice of a very small bin width results in a jagged histogram, with a separate block for each distinct observation. A very large bin width results in a histogram with a single block. Intermediate bin widths lead to a variety of histogram shapes between these two extremes. Ideally, the bin width should be chosen so that the histogram displays the essential structure of the data, without giving too much credence to the data set at hand.

Scott (1992, p. 48) gave an interesting historical account of bin width selection. The earliest published rule for selecting the bin width appears to be that of Sturges (1926). As Scott points out, Sturges's proposal is more of a number-of-bins rule rather than a bin width rule itself, but essentially amounts to choosing the bin width

$$\hat{h} = \frac{\text{range of data}}{1 + \log_2 n},$$

where n is the sample size. Well-established theory (e.g., Scott 1992) shows that this bin width leads to an oversmoothed histogram, especially for large samples. However, Sturges's rule, or variations of it such as that proposed by Doane (1976), is often used in statistical packages as a default. The default bin width used by the popular languages S and S-PLUS is a modification of Sturges's rule that ensures nice break points between the bins.

It could be argued that this situation is somewhat unfortunate because inexperienced data analysts might miss important features in their data sets (an example is given in Section 5.2). Acceptance of this viewpoint implies that default bin widths should be "more scientific," driven by some sort of optimal estimation theory. At the same time one should not lose sight of the simplicity of the histogram, and the advantages of having the choice of the bin width kept relatively simple as well. The purpose of this paper is to develop rules that aim to achieve a balance between good mathematical and practical performance and simplicity.

It has been known for some time (e.g., Smirnov 1950) that, with respect to L_p norms, the optimal rate of decay of the bin width is $n^{-1/3}$. However, it was not until the relatively recent work of Scott (1979) and Freedman and Diaconis (1981) that the asymptotic effect of the bin width on the mean L_2 error, or mean integrated squared error (MISE), was fully understood. This theory has lead to the proposal of several rules of the form

$$\hat{h} = \hat{C} n^{-1/3}$$

for some statistic $\hat{C}$. Worthy of special mention is Scott's (1979) normal reference rule

$$\hat{h} = 3.49 \hat{\sigma} n^{-1/3},$$

M. P. Wand is Senior Lecturer, Australian Graduate School of Management, University of New South Wales, Sydney 2052, Australia. The author thanks Prof. Dennis Cox for suggesting the problem, and Prof. Steve Marron for a helpful comment.

where $\hat{\sigma}$ is an estimate of the standard deviation, so named because it is based on calibration with the normal distribution with variance σ^2. Modifications of this idea to allow for varying degrees of skewness and kurtosis have also been developed by Scott, and were presented and studied in section 3.2.3 of Scott (1992).

Although each of these $n^{-1/3}$ rules provides about the right number of bins for each situation, it must be recognized that they are only rough approximations to the MISE-optimal bin width and with no large sample consistency properties. It is interesting to note that, although there is a huge body of theory devoted to optimal estimation of common parameters, such as those based on moments, very little exists for estimation of the MISE-optimal bin width (and even less for bin widths that are optimal for other criteria). For example, is it possible to estimate the optimal bin width with root-n consistency, just as with regular parameters?

We propose and describe an extension of the rules of Scott that does have good consistency properties. In fact, we are able to present a relatively simple rule that achieves the best possible asymptotic performance, in terms of both rates of convergence and constant coefficient, to the MISE-optimal bin width. Although this rule is unbeatable from a theoretical standpoint, it does not necessarily follow that it is the best to use in practice because of the delicate asymptotics involved, so simpler rules with only slightly inferior theoretical performance are also considered.

Apart from exhibiting good convergence properties, our class of rules has the advantage of being simple and fast to compute, which is in keeping with the "rough-and-ready" nature of the histogram itself.

Full specification of a histogram also requires the positioning of the bins once a bin width has been decided upon. This choice can also have a marked effect on the resulting histogram; see, for example, figure 1.3 of Wand and Jones (1995). There does not seem to be a sensible data-based way of choosing the positioning the bins, so in practice it is recommended that the user look at several "shifted" histograms with the same chosen bin width.

The proposed bin width rules require considerably more computational effort than that required for a histogram, with the amount of work roughly equivalent to the construction of a kernel density estimate. As one of the reviewers has pointed out, an obvious question that arises is: Why would one go to all that computational trouble to use a histogram instead of a kernel density estimate? It is certainly true that more experienced statisticians might prefer to bypass the histogram completely and use a more sophisticated density estimator, such as one based on a kernel. Nevertheless, it must be recognized that the overwhelming majority of density estimates obtained in practice are of the histogram type, usually by a novice or part-time statistician using a computer package with a default bin width. Users such as these usually understand the construction of the histogram itself, but the bin width choice is almost always treated as a "black box." Clearly, it is imperative that this black box produces as good an answer as possible. It is hoped that this research will provide a better black box than those currently used in packages, and perhaps inspire the development of even better bin width selectors.

Section 2 contains the basic methodology, followed by a study of theoretical performance in Section 3. Implementation details are given in Section 4, numerical examples are given in Section 5, and concluding remarks are given in Section 6. Proofs are deferred to the Appendix.

2. METHODOLOGY

Let $X_1, \ldots, X_n$ be a real-valued random sample having density f. The histogram estimator of $f(x)$, based on a partition of the real-line into bins B_j of width h, is given by

$$\hat{f}(x; h) = \frac{\text{number of } X_i\text{'s in } B_j}{nh}, \qquad x \in B_j.$$

Before we can set out to select the bin width h scientifically we need some mathematical optimum at which to aim. For reasons of tractability and simplicity we will take the optimal bin width to be h_{MISE}, the bin width that minimizes

$$\text{MISE}\{\hat{f}(\cdot\,; h)\} = E \int_{-\infty}^{\infty} \{\hat{f}(x; h) - f(x)\}^2 \, dx.$$

If f is sufficiently regular, then as $h = h_n \to 0$ and $nh \to \infty$, we have

$$h_{\text{MISE}} \sim \left(\frac{6}{-\psi_2 n}\right)^{1/3} \qquad \text{as } n \to \infty \tag{2.1}$$

(Scott 1979; Freedman and Diaconis 1981) where

$$\psi_r \equiv E\{f^{(r)}(X)\} = \int_{-\infty}^{\infty} f^{(r)}(x) f(x) \, dx, \qquad r \text{ even.}$$

This does not solve our problem for estimation of h_{MISE}, but helps us because of the fact that high-quality kernel-type estimators of ψ_r exist. If L is a sufficiently smooth symmetric, unimodal probability density function, then a consistent estimator for ψ_r is

$$\hat{\psi}_r(g) = n^{-2} g^{-r-1} \sum_{i=1}^{n} \sum_{j=1}^{n} L^{(r)}\{(X_i - X_j)/g\} \tag{2.2}$$

(Hall and Marron 1987; Jones and Sheather 1991) for some positive parameter g, usually referred to as a "bandwidth." Therefore, in view of (2.1) h_{MISE} can be estimated by

$$\hat{h} = \left\{\frac{6}{-\hat{\psi}_2(g) n}\right\}^{1/3}.$$

However, we now have a new problem: the choice of g. With respect to $E[\{\hat{\psi}_2(g) - \psi_2\}]^2$, the asymptotically (as $g \to 0$ and $n^{-1} g^{-5} \to 0$) optimal g is

$$g_{\text{MSE},2} = \left\{\frac{2L^{(2)}(0)}{-\mu_2(L)\psi_4 n}\right\}^{1/5}$$

(Jones and Sheather 1991), where $\mu_k(L) = \int u^k L(u) \, du$. But because $g_{\text{MSE},2}$ depends on ψ_4, we have to estimate this functional to obtain a practical rule. In general, the optimal bandwidth $g_{\text{MSE},r}$ for estimation of ψ_r depends on ψ_{r+2}, so this process can be continued indefinitely, but eventually

a "rough" estimate of a higher order ψ_r will be required. A common approach to this problem is the replacement of f by the normal density with variance $\hat{\sigma}^2$ where $\hat{\sigma}$ is an appropriate estimate of scale (see, e.g., Sheather and Jones 1991). Such an estimate of ψ_r is called a normal scale estimator, and we will denote this by $\hat{\psi}_r^{\mathrm{NS}}$. A useful result is

$$\hat{\psi}_r^{\mathrm{NS}} = \frac{(-1)^{r/2} r!}{(2\hat{\sigma})^{r+1}(r/2)!\pi^{1/2}}.$$

A popular choice for $\hat{\sigma}$ is

$$\hat{\sigma} = \min\{s,\ \mathrm{IQR}/1.349\}, \tag{2.3}$$

where s is the sample standard deviation and IQR is the interquartile range. The factor 1.349 ensures that $\hat{\sigma}$ is consistent for σ when the data are normal. More sophisticated choices for $\hat{\sigma}$ are possible: see, for example, Janssen, Marron, Veraverbeke, and Sarle (1995). Alternatively, one could obtain an initial estimate of ψ_r using the skewness-kurtosis idea of Scott (1992, pp. 56–57).

The number of stages l of functional estimation before a rough estimate is used is another variable that needs to be specified. This means that we actually have a family of plug-in rules indexed by l. Let $\hat{h}_l$ denote the l-stage plug-in rule with initial estimates found using a normal scale rule. Examples of $\hat{h}_l$ are the following:

The zero-stage rule $\hat{h}_0$:

$$\hat{h}_0 = \left(\frac{6}{-\hat{\psi}_2^{\mathrm{NS}} n}\right)^{1/3} = (24\pi^{1/2}/n)^{1/3}\hat{\sigma} \simeq 3.49\hat{\sigma}n^{-1/3}.$$

Note that $\hat{h}_0$ is simply the normal scale bin width selection rule of Scott (1979).

The one-stage rule $\hat{h}_1$:

$$\hat{h}_1 = \left\{\frac{6}{-\hat{\psi}_2(g_{11})n}\right\}^{1/3},$$

where

$$g_{11} = [-2L^{(2)}(0)/\{\mu_2(L)\hat{\psi}_4^{\mathrm{NS}} n\}]^{1/5}.$$

The two-stage rule $\hat{h}_2$:

$$\hat{h}_2 = \left\{\frac{6}{-\hat{\psi}_2(g_{21})n}\right\}^{1/3},$$

where

$$g_{21} = [-2L^{(2)}(0)/\{\mu_2(L)\hat{\psi}_4(g_{22})n\}]^{1/5}$$

and

$$g_{22} = [-2L^{(4)}(0)/\{\mu_4(L)\hat{\psi}_6^{\mathrm{NS}} n\}]^{1/7}.$$

These rules belong to what is usually called the "plug-in" family of smoothing parameter selectors because they are based on plugging in estimates of unknown quantities. For the bandwidth selection problem in kernel density estimation this idea dates back to Woodroofe (1970), and recently has been popularized by several authors including Park and Marron (1990), Sheather and Jones (1991), Chiu (1991), Hall, Sheather, Jones, and Marron (1991), and Engel, Herrmann, and Gasser (1995).

3. THEORETICAL PERFORMANCE

A reward for our "scientific" approach to the selection of h is very good theoretical performance, at least if l is sufficiently high. Obviously, $\hat{h}_0$ does not converge to h_{MISE} because it does not use a consistent estimate of ψ_2. However, for $l \geq 1$ we have the following.

Theorem. If (A.1) and (A.2) in the Appendix are satisfied, then

$$n^{2/5}(\hat{h}_1/h_{\mathrm{MISE}} - 1) \to_p D, \tag{3.1}$$

where

$$D = \frac{1}{3}\,\psi_2^{-1}\left\{\frac{\mu_2(L)^3 L^{(2)}(0)^2}{8}\right\}^{1/5} \times \left\{\left(\frac{3\sigma^5}{8\pi^{1/2}}\right)^{3/5} - \psi_4\left(\frac{3\sigma^5}{8\pi^{1/2}}\right)^{-2/5}\right\},$$

and for $l \geq 2$,

$$n^{1/2}(\hat{h}_l/h_{\mathrm{MISE}} - 1) \to_D N(0, \sigma_{\mathrm{PI}}^2), \tag{3.2}$$

where

$$\sigma_{\mathrm{PI}}^2 = \frac{1}{9}\,\psi_2^{-2}\left[-\left\{\int (L'')^2\right\} \times \mu_2(L)L^{(2)}(0)^{-1}\psi_0\psi_4 + 4\ \mathrm{var}\{f''(X)\}\right].$$

Therefore, provided that at least two stages of functional estimation are used, our rules achieve a root-n relative rate of convergence.

The $n^{-1/2}$ convergence rate of $\hat{h}_l/h_{\mathrm{MISE}} - 1$, for $l \geq 2$, is the best possible (Hall and Marron 1991), but does not quite reach the lower bound on the asymptotic variance

$$\sigma_{\mathrm{LB}}^2 = \frac{4}{9}\,\mathrm{var}\{f''(X)\}/[E\{f''(X)\}]^2 \tag{3.3}$$

(by extension of Fan and Marron 1992). In fact, one can achieve this best possible variance by simply replacing L by a fourth-order kernel $L^{[4]}$ (i.e., satisfying $\int u^2 L^{[4]}(u)\,du = 0$) and $l \geq 2$. For example, the appropriate fourth-order version of $\hat{h}_2$, which we will denote by $\hat{h}_2^{[4]}$, is given by

$$\hat{h}_2^{[4]} = \left\{\frac{6}{-\hat{\psi}_2(g_{21}^{[4]})n}\right\}^{-1/3},$$

where

$$g_{21}^{[4]} = [-24L_{[4]}^{(2)}(0)/\{\mu_4(L_{[4]})\hat{\psi}_6(g_{22}^{[4]})n\}]^{1/7}$$

and

$$g_{22}^{[4]} = [-24L_{[4]}^{(6)}(0)/\{\mu_8(L_{[4]})\hat{\psi}_{10}^{\mathrm{NS}} n\}]^{1/11}.$$

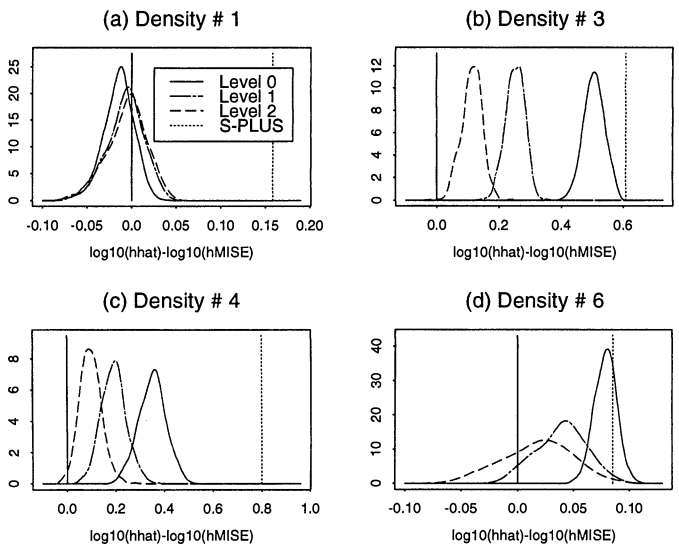

Figure 1. Graphical Summary of Simulation Results for n = 500. Kernel density estimates of 500 $\log_{10}(\hat{h}_l) - \log_{10}(h_{MISE})$ values; l = 0, 1, 2. The average bin width chosen by S-PLUS *is shown by a vertical dotted line.*

For this rule it can be shown that, under sufficient smoothness assumptions on f,

$$n^{1/2}(\hat{h}_2^{[4]}/h_{\mathrm{MISE}} - 1) \to_D N(0, \sigma_{\mathrm{LB}}^2).$$

This means that $\hat{h}_2^{[4]}$ has the best asymptotic performance among all possible bin width selectors (Fan and Marron 1992). However, the delicateness of the asymptotic arguments by which this result is obtained means that "best" practical performance is not guaranteed. Moreover, the extra complexity and loss of interpretability of higher order kernels makes $\hat{h}_2^{[4]}$ less attractive than the rules $\hat{h}_l$ based on nonnegative kernels.

4. IMPLEMENTATION

One of the main obstacles with practical implementation of the bin width selectors proposed in the previous section is that they depend upon estimators of the form (2.2) which, for even moderate sample sizes, can be very expensive to compute. One simple way around this is to replace $\hat{\psi}_r(g)$ by its binned approximation. Let

$$\min(X_i) = G_1 \le G_2 \le \cdots \le G_M = \max(X_i) \quad (4.1)$$

be an equally spaced grid of width $\delta = (G_M - G_1)/(M-1)$ over the range of the data. Then the linear binned approximation to $\hat{\psi}_r(g)$ is

$$\tilde{\psi}_r(g) = n^{-2} \sum_{j=1}^{M} \left(\sum_{j'=1}^{M} c_{j'} \kappa_{j'-j}^{(r)} \right) c_j. \quad (4.2)$$

Here

$$c_j = \sum_{i=1}^{n} (1 - |\delta^{-1} X_i - j|)_+ \quad (4.3)$$

[with $x_+ = \max(0, x)$] is the count at grid point G_j and

$$\kappa_j^{(r)} = g^{-r-1} L^{(r)}(\delta j / g), \qquad |j| = 0, \ldots, M. \quad (4.4)$$

The advantage of $\tilde{\psi}_r(g)$ is that it involves only $O(M)$ evaluations of $L^{(r)}$, as opposed to the $O(n^2)$ kernel evaluations required by $\hat{\psi}_r(g)$. This makes it much faster to compute. Results of Gonzalez-Manteiga, Sanchéz-Sellero, and Wand (1996) indicate that $M \simeq 400$ leads to a very accurate approximation of $\hat{\psi}_r(g)$ by $\tilde{\psi}_r(g)$ for a wide range of density shapes.

Because of its smoothness and simplicity we will describe the full binned implementation of $\hat{h}_1$ and $\hat{h}_2$ for the normal kernel $L(x) = (2\pi)^{-1/2} e^{-x^2/2}$. We will also take $\hat{\sigma} = \min\{s, \mathrm{IQR}/1.349\}$ and $M = 400$. Other choices of $L, \hat{\sigma}$, and $M \ge 400$ might be considered, but these should make little difference to the answer for most applications.

Let $\tilde{\psi}_r(g)$ be defined by (4.1)–(4.4) with $M = 400$ and L equal to the standard normal kernel.

Binned implementation of $\hat{h}_1$ with normal kernel:

$$\tilde{h}_1 = \left\{ \frac{6}{-\tilde{\psi}_2(g_{11}) n} \right\}^{1/3}$$

where

$$g_{11} = \{2/(3n)\}^{1/5} 2^{1/2} \hat{\sigma}.$$

Binned implementation of $\hat{h}_2$ with normal kernel:

$$\tilde{h}_2 = \left\{ \frac{6}{-\tilde{\psi}_2(g_{21}) n} \right\}^{1/3}$$

where

$$g_{21} = [2/\{(2\pi)^{1/2} \tilde{\psi}_4(g_{22})^{1/5} n\}]^{1/5} 2^{1/2} \hat{\sigma}$$

and

$$g_{22} = \{2/(5n)\}^{1/7} 2^{1/2} \hat{\sigma}.$$

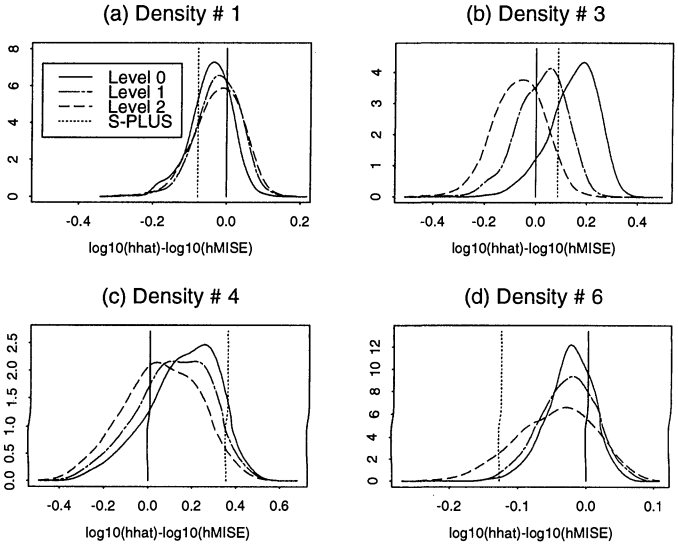

Figure 2. Graphical Summary of Simulation Results for n = 50. Kernel density estimates of 500 $\log_{10}(\hat{h}_l) - \log_{10}(h_{MISE})$ values; l = 0, 1, 2. The average bin width chosen by S-PLUS *is shown by a vertical dotted line.*

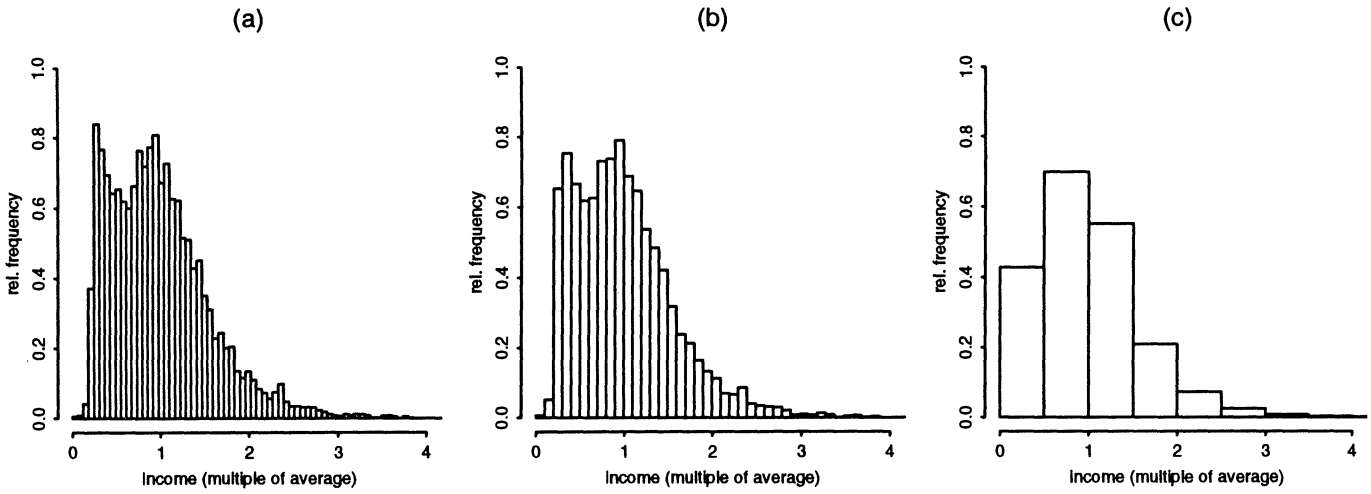

Figure 3. Histograms of the British Incomes Data Based on (a) the Bin Width $\hat{h}_2$, (b) the Bin Width $\hat{h}_0$, and (c) the S-PLUS Default Bin Width.

It should be pointed out that a fast algorithm based on integer division exists for computing the c_j. See Fan and Marron (1994) for its description. Also, the inner summation of (4.2) is a discrete convolution, and so can be computed quickly using the fast Fourier transform (see, e.g., Wand 1994) or an economical direct approach (see Scott 1992, p. 118).

5. NUMERICAL RESULTS

5.1 Simulation Results

To test the efficacy of our proposal a small simulation study was carried out with 500 replications, $n = 500$, and the densities #1 (standard normal), #3 (strongly skewed), #4 (kurtotic), and #6 (bimodal) from Marron and Wand (1992).

The results are graphically summarized in Figure 1. Figure 1 shows kernel density estimates of $\log_{10}(\hat{h}_l) - \log_{10}(h_{\mathrm{MISE}})$ for $l = 0, 1, 2$. The dashed vertical line shows the position of the average default bin width used by S-PLUS. Because S-PLUS "rounds off" its default bin width there is not enough variability among the sample of bin widths to use a density estimate. For example, for the normal density S-PLUS chose a bin width of .5 for 361 of the samples and a bin width of 1.0 for the remaining 139 samples.

Among the $\hat{h}_l$ there is a definite pattern of higher values of l leading to a rule with lower bias, but with a possible increase in variance (especially apparent for density #6). This can be explained by the fact that larger l corresponds to less dependence on the normal scale rule, but more estimation steps. Overall, $\hat{h}_2$ is the rule that is consistently closest to the optimal bin width.

The S-PLUS default bin width is always, to varying degrees, too large compared to h_{MISE}. This can be partially explained by the rounding off procedure. Nevertheless, this tendency to oversmooth is disconcerting because it can give misleading results when the distribution has some interesting structure.

We then ran the study simulation for $n = 50$ to check that the selectors did not rely on the sample size being large. The results are summarized in Figure 2. They indicate that the higher level bin width selectors maintain their good performance.

Finally, we ran the simulation for $n = 10$. In this case the optimal bin width is not very well defined so the main goal was to make sure that the rules were not susceptible to complete breakdown. Analysis of the simulation output showed that all of the bandwidth samples were relatively well behaved, without any strong outliers.

5.2 An Example

Figure 3 shows the results of applying (a) $\hat{h}_2$, (b) Scott's (1979) normal reference rule $\hat{h}_0$, and (c) the S-PLUS default bin width to a real data set. The data represent 7,201 British incomes for the year 1975, and have been divided by their sample average. The source of these data is the ESCR Data Archive at the University of Essex, Family Expenditure Survey, Annual Tapes, 1968–1983, Department of Employment, Statistics Division, Her Majesty's Stationery Office, London.

The bin width choice $\hat{h}_2$ leads to a histogram that clearly shows the interesting bimodal structure in the data. The selector $\hat{h}_0$ also shows the bimodal structure, but not quite as sharply. The S-PLUS default bandwidth leads to a grossly oversmoothed histogram that does not detect the bimodality. An inexperienced data analyst using a package like S-PLUS would probably miss this important feature in the data due to its ad hoc choice of bin width.

6. CONCLUSION

We have shown that simple plug-in ideas lead to bin width rules with very good theoretical properties, as well as good to very good practical performance. The rule $\hat{h}_2$ gives the most satisfactory practical performance, and has near-optimal theoretical properties. This contrasts with the logarithmic varying bin width rules used by many statistical packages, which tend to hide detail in the data.

An S-PLUS function for computation of $\hat{h}_l$ is available by request from the author (e-mail: wand@agsm.unsw.edu.au).

APPENDIX: PROOF OF THEOREM

Conditions sufficient for the theorem to hold are as follows.

- (A.1) The kernel L has all moments finite and two continuous derivatives.
- (A.2) The density f has three continuous, ultimately decreasing derivatives.

Weaker conditions are certainly possible, but probably not worth pursuing.

Let $M(h) \equiv \text{MISE}\{\hat{f}(\cdot\,; h)\}$. An extension of the results of Scott (1979) and Freedman and Diaconis (1981) leads to

$$M(h) = n^{-1}h^{-1} - \frac{1}{6}\, h^2\psi_2 + \frac{1}{30}\, h^4\psi_4 + o(n^{-1}h^{-1} + h^4)$$

from which it follows that

$$M'(h) = -n^{-1}h^{-2} - \frac{1}{3}\, h\psi_2 + \frac{2}{15}\, h^3\psi_4 + o(n^{-1}h^{-2} + h^3)$$

and

$$M''(h) = 2n^{-1}h^{-3} - \frac{1}{3}\, \psi_2 + \frac{2}{5}\, h^2\psi_4 + o(n^{-1}h^{-3} + h^2).$$

Now

$$0 = M'(h_{\text{MISE}}) = M'(h_{\text{AMISE}}) + (h_{\text{MISE}} - h_{\text{AMISE}})M''(\bar{h}),$$

where $\bar{h} = O(n^{-1/3})$ is between h_{AMISE} and h_{MISE}. Therefore,

$$h_{\text{MISE}} = h_{\text{AMISE}} - \frac{M'(h_{\text{AMISE}})}{M''(\bar{h})} = h_{\text{AMISE}} + O(n^{-1}).$$

It follows that the relative error of $\hat{h}_l$ is given by

$$\begin{aligned} &\hat{h}_l/h_{\text{MISE}} - 1 \\ &\quad = \psi_2^{-1/3}\{\hat{\psi}_2(g)^{-1/3} - \psi_2^{-1/3}\} + O_P(n^{-2/3}) \\ &\quad \simeq -\frac{1}{3}\, \psi_2^{-1/3}\{\hat{\psi}_2(g) - \psi_2\} + \cdots + O_P(n^{-2/3}). \quad \text{(A.1)} \end{aligned}$$

Results for the asymptotic distribution of degenerate (Hall 1984) and nondegenerate (Serfling 1980) U-statistics lead to

$$\hat{\psi}_2(g) - \psi_2 = [\text{var}\{\hat{\psi}_2(g)\}]^{1/2} Z_n + E\hat{\psi}_2(g) - \psi_2, \quad \text{(A.2)}$$

where Z_n is asymptotically $N(0,1)$. Note that, for $g = Cn^{-1/5}$, $\text{var}\{\hat{\psi}_2(g)\}^{1/2} = O(n^{-1/2})$.

First suppose that $l = 1$. Then

$$E\hat{\psi}_2(g) - \psi_2 = \left\{ L^{(r)}(0) + \frac{1}{2}\, \mu_2(L)\psi_4^{\text{NS}} \right\} n^{-2/5}, \quad \text{(A.3)}$$

which dominates the first term of (A.2), so (3.1) follows from this result and (A.1). However, for $l \geq 2$ the coefficient of $n^{-2/5}$ in (A.3) cancels and we are left with

$$E\hat{\psi}_2(g) - \psi_2 = O(n^{-4/5}),$$

which means that the first term of (A.2) is dominant. Results of Hall and Marron (1987) for the variance of $\hat{\psi}_2(g)$ lead to (3.2).

[Received December 1994. Revised May 1996.]

REFERENCES

Chiu, S. T. (1991), "Bandwidth Selection for Kernel Density Estimation," *Annals of Statistics*, 19, 1883–1905.

Doane, D. P. (1976), "Aesthetic Frequency Classifications," *The American Statistician*, 30, 181–183.

Engel, J., Herrmann, E., and Gasser, T. (1995), "An Iterative Bandwidth Selector for Kernel Estimation of Densities and Their Derivatives," *Journal of Nonparametric Statistics*, 4, 21–34.

Fan, J., and Marron, J. S. (1992), "Best Possible Constant for Bandwidth Selection," *Annals of Statistics*, 20, 2057–2070.

——— (1994), "Fast Implementations of Nonparametric Curve Estimators," *Journal of Computational and Graphical Statistics*, 3, 35–56.

Freedman, D., and Diaconis, P. (1981), "On the Histogram as a Density Estimator: L_2 Theory," *Zeitschrift fuer Wahrsheinlichkeïtstheorie und Verwandte Gebeite*, 57, 453–476.

González-Manteiga, W., Sanchéz-Sellero, C., and Wand, M. P. (1996), "Accuracy of Binned Kernel Functional Approximations," *Computed Statistical Data Analysis*, 19.

Hall, P. (1984), "Central Limit Theorem for Integrated Square Error of Multivariate Nonparametric Density Estimators," *Journal of Multivariate Analysis*, 14, 1–16.

Hall, P., and Marron, J. S. (1987), "Estimation of Integrated Squared Density Derivatives," *Statistics and Probability Letters*, 6, 109–115.

——— (1991), "Lower Bounds for Bandwidth Selection in Density Estimation," *Probability Theory and Related Fields*, 90, 149–173.

Hall, P., Sheather, S. J., Jones, M. C., and Marron, J. S. (1991), "On Optimal Data-Based Bandwidth Selection in Kernel Density Estimation," *Biometrika*, 78, 263–269.

Janssen, P., Marron, J. S., Veraverbeke, N., and Sarle, W. (1995), "Scale Measures for Bandwidth Selection," *Journal of Nonparametric Statistics*, 5, 359–380.

Jones, M. C., and Sheather, S. J. (1991), "Using Nonstochastic Terms to Advantage in Kernel-Based Estimation of Integrated Squared Density Derivatives," *Statistics and Probability Letters*, 11, 511–514.

Marron, J. S., and Wand, M. P. (1992), "Exact Mean Integrated Squared Error," *Annals of Statistics*, 20, 712–736.

Park, B. U., and Marron, J. S. (1990), "Comparison of Data-Driven Bandwidth Selectors," *Journal of the American Statistical Association*, 85, 66–72.

——— (1992), "On the Use of Pilot Estimators in Bandwidth Selection," *Nonparametric Statistics*, 1, 231–240.

Scott, D. W. (1979), "On Optimal and Data-Based Histograms," *Biometrika*, 66, 605–610.

——— (1992), *Multivariate Density Estimation*, New York: John Wiley.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, New York: John Wiley.

Sheather, S. J., and Jones, M. C. (1991), "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation," *Journal of the Royal Statistical Society*, Ser. B, 53, 683–690.

Smirnov, N. V. (1950), "On the Construction of a Confidence Region for the Density of the Distribution of a Random Variable," *Akademija Nauk Ukraĭnskei SSR, Doklady, Serija A*, 74, 189–191.

Sturges, H. A. (1926), "The Choice of a Class Interval," *Journal of the American Statistical Association*, 21, 65–66.

Wand, M. P. (1994), "Fast Computation of Multivariate Kernel Estimators," *Journal of Computational and Graphical and Statistics*, 3, 433–445.

Wand, M. P., and Jones, M. C. (1995), *Kernel Smoothing*, London: Chapman and Hall.

Woodroofe, M. (1970), "On Choosing a Delta Sequence," *Annals of Mathematical Statistics*, 41, 1665–1671.