Biometrika (1979), **66**, 3, *pp.* 605–10
Printed in Great Britain

On optimal and data-based histograms

BY DAVID W. SCOTT
Department of Mathematical Sciences, Rice University, Houston, Texas

SUMMARY

In this paper the formula for the optimal histogram bin width is derived which asymptotically minimizes the integrated mean squared error. Monte Carlo methods are used to verify the usefulness of this formula for small samples. A data-based procedure for choosing the bin width parameter is proposed, which assumes a Gaussian reference standard and requires only the sample size and an estimate of the standard deviation. The sensitivity of the procedure is investigated using several probability models which violate the Gaussian assumption.

Some key words: Frequency distribution; Histogram; Nonparametric density estimation; Optimal bin width.

1. INTRODUCTION

The histogram is the classical nonparametric density estimator, probably dating from the mortality studies of John Graunt in 1662 (Westergaard, 1968, p. 22). Today the histogram remains an important statistical tool for displaying and summarizing data. In addition it provides a consistent estimate of the true underlying probability density function. Present guidelines for constructing histograms do not directly address the issues of estimation bias and variance. Rather, they draw heavily on the investigator's intuition and past experience. In this paper we propose new guidelines that reduce the subjectivity involved in histogram construction by considering a mean squared error criterion.

2. BACKGROUND

We consider only histograms defined on an equally spaced mesh $\{t_{ni};\ -\infty < i < \infty\}$ with bin width $h_n = t_{n(i+1)} - t_{ni}$, where n denotes the sample size and emphasizes the dependence of the mesh and bin width on the sample size. For a fixed point x, the mean squared error of a histogram estimate, $\hat{f}(x)$, of the true density value, $f(x)$, is defined by

$$\text{MSE}\,(x) = E\{\hat{f}(x) - f(x)\}^2.$$

For a random sample of size n from f, Čencov (1962) proved that MSE (x) asymptotically converges to zero at a rate proportional to $n^{-2/3}$, that is, MSE $(x) = O(n^{-2/3})$. This rate is fairly close to the Cramér–Rao lower bound of $O(n^{-1})$. The integrated mean squared error represents a global error measure of a histogram estimate and is defined by

$$\text{IMSE} = \int E\{\hat{f}(x) - f(x)\}^2\,dx.$$

Since it is the shape of the density that is of most interest, the IMSE is more relevant than the mean squared error of the density height. The IMSE of a histogram also converges to zero as $O(n^{-2/3})$.

To achieve these rates of convergence requires proper choice of the two parameters of the histogram, the bin width h_n and the relative position of the mesh. The latter is determined by

any particular mesh point, say t_{n0}. Statistical texts suggest various methods for choosing these two parameters. First the bin width is determined indirectly by choosing an appropriate number of bins over the sample range. Most authors advise that 5–20 bins are usually adequate for real data sets (Haber & Runyon, 1969, p. 33; Guttman & Wilks, 1965, p. 59). Larson (1975, p. 15) suggests using $1+2{\cdot}2\log_{10} n$ bins as a first choice, similar to a formula proposed by Sturges in 1926. The final choice for h_n is a convenient whole number or fraction, often related to the accuracy with which the data are measured. Next, t_{n0} is picked so that the data do not fall on the bin boundaries. If we assume that the data are measured to infinite accuracy, then the choice of t_{n0} becomes less important as the sample size increases. Since we are focusing on consistency, we shall assume $t_{n0} = 0$ in the sequel. However, the choice of h_n is quite important. If h_n is too small, then the histogram will be too rough; on the other hand, if h_n is too large, then the histogram will be too smooth, equivalent statistically to large variance and large bias, respectively. The proper choice for h_n should balance the bias and variance by minimizing, for example, the integrated mean squared error.

In the past 20 years new nonparametric density estimators have been proposed and investigated (Tapia & Thompson, 1978; Wegman, 1972). The most extensively treated of these new estimators is the kernel probability density estimator developed by Rosenblatt (1956) and Parzen (1962). The kernel estimator is also consistent but with IMSE $= O(n^{-4/5})$, an improvement over the histogram. In spite of these advances, the histogram will almost surely retain important roles in data representation and density estimation, since it is simple to compute and easily understood. Fortunately, by using techniques employed in kernel density estimation consistency proofs, it is now possible to derive the optimal choice for the bin width h_n of a histogram.

3. Derivation of the optimal histogram bin width

Suppose that $x_1, \ldots, x_n$ is a random sample from a continuous probability density function f with two continuous and bounded derivatives. We shall need to identify the bin interval that contains a fixed point x as n varies. Let $I_n(x)$ be that interval and let $t_n(x)$ denote the left-hand endpoint of $I_n(x)$. Define the bin probability

$$p_n(x) = \int_{t_n(x)}^{t_n(x)+h_n} f(y)\,dy.$$

For y in $I_n(x)$ we have, using Taylor's expansion, $f(y) = f(x)+f'(x)(y-x)+O(h_n^2)$. Therefore

$$\begin{aligned} p_n(x) &= \int_{t_n(x)}^{t_n(x)+h_n} \{f(x)+f'(x)(y-x)+O(h_n^2)\}\,dy \\ &= h_n f(x)+\tfrac{1}{2}f'(x)[h_n^2-2h_n\{x-t_n(x)\}]+O(h_n^3). \end{aligned}$$

Let $v_n(x)$ be the number of values falling in $I_n(x)$. Then $v_n(x)$ has a binomial distribution $B\{n, p_n(x)\}$. The histogram estimate is given by the random variable

$$\hat{f}(x) = v_n(x)/(nh_n),$$

with expectation

$$\begin{aligned} E\{\hat{f}(x)\} &= p_n(x)/h_n \\ &= f(x)+\tfrac{1}{2}h_n f'(x)-f'(x)\{x-t_n(x)\}+O(h_n^2). \end{aligned}$$

Therefore the bias is

$$\tfrac{1}{2}h_n f'(x)-f'(x)\{x-t_n(x)\}+O(h_n^2).$$

Now the variance of the histogram estimate at x is given by

$$\begin{aligned}\operatorname{var}\{\hat{f}(x)\} &= p_n(x)\{1-p_n(x)\}/(nh_n^2)\\ &= \{h_n f(x)+O(h_n^2)\}\{1-O(h_n)\}/(nh_n^2)\\ &= f(x)/(nh_n)+O(1/n).\end{aligned}$$

Combining, we have that

$$\text{MSE}\,(x) = f(x)/(nh_n)+\tfrac{1}{4}h_n^2 f'(x)^2+f'(x)^2\{x-t_n(x)\}^2-h_n f'(x)^2\{x-t_n(x)\}+O(1/n+h_n^3). \quad (1)$$

Integration of equation (1) over the real line implies that

$$\text{IMSE} = 1/(nh_n)+\tfrac{1}{4}h_n^2\int f'(x)^2\,dx+\int f'(x)^2\{x-t_n(x)\}^2\,dx$$

$$-h_n\int f'(x)^2\{x-t_n(x)\}\,dx+O(1/n+h_n^3). \quad (2)$$

Recall that $\{t_{ni}\}$ denotes the mesh. Then the third term in equation (2) may be written as

$$\sum_{i=-\infty}^{\infty}\int_{t_{ni}}^{t_{ni}+h_n} f'(x)^2(x-t_{ni})^2\,dx = \sum_{i=-\infty}^{\infty}\int_0^{h_n} f'(t_{ni}+y)^2y^2\,dy \quad (3)$$

by a change of variables. Now $f'(t_{ni}+y) = f'(t_{ni})+O(h_n)$, so that (3) becomes

$$\sum_{i=-\infty}^{\infty}\int_0^{h_n}\{f'(t_{ni})^2+O(h_n)\}y^2\,dy = \Sigma\,\tfrac{1}{3}h_n^3 f'(t_{ni})^2+O(h_n^4) = \tfrac{1}{3}h_n^2\int_{-\infty}^{\infty} f'(x)^2\,dx+O(h_n^3),$$

by standard numerical integration approximations. A similar analysis for the fourth term in (2) yields

$$-\tfrac{1}{2}h_n^2\int_{-\infty}^{\infty} f'(x)^2\,dx+O(h_n^3).$$

Therefore

$$\text{IMSE} = 1/(nh_n)+\tfrac{1}{12}h_n^2\int_{-\infty}^{\infty} f'(x)^2\,dx+O(1/n+h_n^3). \quad (4)$$

Minimizing the first two terms in (4), we obtain

$$h_n^* = \left\{6\Big/\int_{-\infty}^{\infty} f'(x)^2\,dx\right\}^{1/3} n^{-1/3}, \quad (5)$$

which, asymptotically, is the optimal choice for h_n.

We can estimate how the IMSE changes for poor choices of the bin width by using (4). For any density and any positive constant c, the IMSE using the bin width ch_n^* is larger than the minimum IMSE by the factor $(c^3+2)/(3c)$. Thus a bin width 50% too small implies an IMSE 42% too large. We remark that a change of scale in the density function results in a similar scaling of the optimal h_n since $y = x/\sigma$ leads to $\int f'(y)^2\,dy = \sigma^3\int f'(x)^2\,dx$. For Gaussian data, $h_n^* = 2\times 3^{1/3}\pi^{1/6}\sigma n^{-1/3}$.

4. Small sample properties

The formula for h_n^* is based on an asymptotic expression. To investigate the small sample properties of the IMSE, we undertook a fairly extensive Monte Carlo study of standardized Gaussian data. For a range of values of h, the integrated squared error was computed exactly

for each of 1000 generated samples and then averaged over the number of repetitions to obtain an estimate of the IMSE. The optimal bin widths predicted by equation (5) were quite close to the empirically observed optimal bin widths for the Monte Carlo study even for samples as small as 25. The estimated IMSE also increased as $(c^3+2)/(3c)$ for bin widths differing from the empirically optimal bin width by the factor c.

5. DATA-BASED HISTOGRAMS

The optimal choice for h_n requires knowledge of the true underlying density f. This knowledge is rare. In another context Tukey (1977, p. 623) has suggested using the Gaussian density as a reference standard, to be used cautiously but frequently. Therefore, we propose the data-based choice for the bin width

$$h_n = 3{\cdot}49sn^{-1/3}, \tag{6}$$

where s is an estimate of the standard deviation. Although the Gaussian density forms the basis of (6), this assumption is not so strong as a parametric Gaussian assumption, i.e. use of equation (6) on non-Gaussian data will not result in a histogram that looks Gaussian. For density functions with equal variances, the data-based choice (6) results in the same bin width. To show that (6) is useful for a large class of densities, we considered Gaussian and non-Gaussian densities with equal variances and observed how their theoretically optimal bin widths (5) differed. In particular we considered three models of non-Gaussian behaviour: skewed, heavy-tailed and bimodal densities.

As a model of skewed data, we used a log normal density with variance equal to $w^2(w^2-1)\exp(-2\gamma/\delta)$ and skewness $(w^2+2)(w^2-1)^{\frac{1}{2}}$, where $w = \exp(\frac{1}{2}\delta^{-2})$. In Fig. 1(a) we plot the ratio of the optimal bin width for the log normal density to the optimal bin width for a Gaussian density with the same variance as a function of the log normal skewness. This ratio does not depend on the sample size. We see that using the Gaussian h_n^* oversmoothes a log normal density; however, for skewnesses as great as one, the difference is less than 30%. A similar plot results when using a gamma probability density model.

We used Student's t_r density to model heavy-tailed data. The variance and kurtosis are $r/(r-2)$ and $6/(r-4)$, respectively. In Fig. 1(b) we plot the ratio of bin widths as a function of the kurtosis, connecting the discrete points by a solid line for convenience. The insensitivity of the data-based choice for h_n for any moderate kurtosis is apparent.

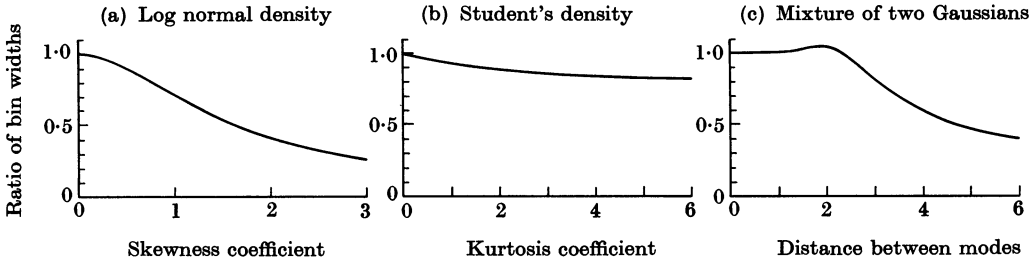

Fig. 1. Ratio of theoretical bin width for several non-Gaussian probability densities to the theoretical bin width for a Gaussian density with the same variance.

Finally as a model of bimodal data, we used a mixture of Gaussian distributions, $\frac{1}{2}N(-\mu, 1)+\frac{1}{2}N(\mu, 1)$, with variance $1+\mu^2$. In Fig. 1(c) we give a similar ratio of theoretical bin widths as a function of 2μ, the distance between the two modes. For strongly bimodal data ($\mu > 1{\cdot}5$), the ratio falls below 0·8, corresponding to oversmoothing of the bimodal data. When distinctly bimodal data are encountered, the data-based histogram may be inadequate.

Thus the data-based algorithm leads to h_n's that are generally too big for all our models of non-Gaussian data. A correction factor may be applied to the data-based h_n by computing the sample skewness or kurtosis and reading the correction factor from Fig. 1. A histogram does not exhibit sensitivity to small changes from the optimal h_n as is evident from the discussion after (5). We do not advocate using exactly the h_n suggested by (6), but rather a convenient choice either slightly larger or smaller.

6. Examples

In Fig. 2 we display three histograms of a Monte Carlo $N(0, 1)$ sample of size 1000 which has a sample standard deviation equal to 1·011 with $h = 0{\cdot}176$, 0·353 and 0·706, the second choice obtained from (6). Many statisticians prefer a smaller bin width and a rougher histogram, leaving the final smoothing to be done by eye.

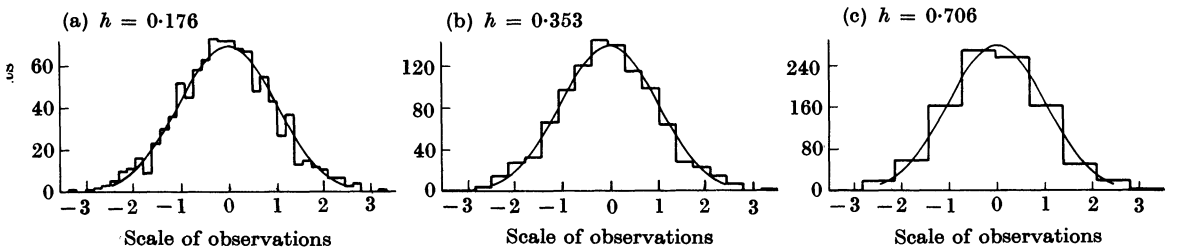

Fig. 2. Histograms of 1000 pseudorandom Gaussian numbers for three bin widths: the data-based choice and that choice perturbed by a factor of 2.

To illustrate extremely large sample sizes, Kendall & Stuart (1969, p. 8) consider a histogram of the ages of 301,785 Australian bridegrooms (1907–14) with a bin width of 3 years. The sample standard deviation and skewness for these data are 7·97 and 1·93, respectively. Thus the data-based choice for h is 0·41 years. Applying a skewness correction factor of 0·43 using Fig. 1(a), the final data-based choice is 0·18 years. Thus the sample is of sufficient size to use a bin width of 1 year or even 3 months if the data were recorded to sufficient accuracy.

7. Discussion

We have considered the optimal construction of histograms given either knowledge of the true underlying density or, more commonly, given only the data. Waterman & Whiteman (1978) have recently carried out a similar attack for Rosenblatt's kernel estimator. Kernel estimates converge faster than histograms to the true density, and therefore integrated mean squared error is more sensitive to the choice of the smoothing parameter; see also Silverman (1978). Furthermore, kernel estimates require the entire data set for evaluation. Thus in some modern automated data collectors, it is often more economical to summarize sequentially relatively more samples, calibrating the histogram using a small training sample.

Some recently developed nonparametric techniques for density estimation start with a histogram and then smooth it; see, for example, Boneva, Kendall & Stefanov (1971). Our procedures could be used to construct the required histogram directly from the data. We remark that our analysis extends easily to histograms in higher dimensions.

It should be possible to further reduce the integrated mean squared error by using an unequally spaced mesh. However, the algorithms required would surely be iterative and would require the entire data set. It is easier to discount rougher estimates in the tails or to construct a rootgram as suggested by Tukey (1977, p. 543).

This research was supported in part by the National Heart, Lung, and Blood Institute, the National Institutes of Health, the Department of Health, Education and Welfare. The author would like to thank a referee for helpful comments.

REFERENCES

BONEVA, L. I., KENDALL, D. G. & STEFANOV, I. (1971). Spline transformations: Three new diagnostic aids for the statistical data-analyst (with discussion). *J. R. Statist. Soc.* B **33**, 1–70.

ČENCOV, N. N. (1962). Estimation of an unknown distribution density from observations. *Soviet Math.* **3**, 1559–62.

GUTTMAN, I. & WILKS, S. S. (1965). *Introductory Engineering Statistics*. New York: Wiley.

HABER, A. & RUNYON, R. P. (1969). *General Statistics*. Reading, Mass: Addison-Wesley.

KENDALL, M. G. & STUART, A. (1969). *The Advanced Theory of Statistics*, Vol. 1. London: Griffin.

LARSON, H. J. (1975). *Statistics: An Introduction*. New York: Wiley.

PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33**, 1065–76.

ROSENBLATT, M. (1956). Remarks on some nonparamteric estimates of a density function. *Ann. Math. Statist.* **27**, 832–7.

SILVERMAN, B. W. (1978). Choosing the window width when estimating a density. *Biometrika* **65**, 1–11.

STURGES, H. A. (1926). The choice of a class interval. *J. Am. Statist. Assoc.* **21**, 65–6.

TAPIA, R. A. & THOMPSON, J. R. (1978). *Nonparametric Probability Density Estimation*. Baltimore: Johns Hopkins.

TUKEY, J. W. (1977). *Exploratory Data Analysis*. Reading, Mass: Addison-Wesley.

WATERMAN, M. S. & WHITEMAN, D. E. (1978). Estimation of probability densities by empirical density functions. *Int. J. Math. Educ. Sci. Technol.* **9**, 127–37.

WEGMAN, E. J. (1972). Nonparametric probability density estimation: I. A summary of available methods. *Technometrics* **14**, 533–46.

WESTERGAARD, H. (1968). *Contributions to the History of Statistics*. New York: Agathon.

[*Received February* 1979. *Revised May* 1979]