DENSITY ESTIMATION

David W. Scott
Department of Statistics
Rice University
6100 Main Street
Houston, TX 77005-1892 USA
scottdw@rice.edu
January 31, 1997

INTRODUCTION

Density estimation is the fitting of a probability density function, $f(x)$, to data. We have the choice of performing a parametric or nonparametric fit. In common usage, the phrase *density estimation* usually refers to the nonparametric methodology, which is the focus of this article. We introduce the classic nonparametric estimator, the histogram, and outline its theoretical properties as well as good practice. We demonstrate how to improve the histogram, leading to our discussion of popular kernel methods. We conclude with a bivariate example, a way of choosing smoothing parameters, and new directions that promise further improvements.

Why choose nonparametric over parametric density estimation? Parametric density estimation requires both proper specification of the form of the underlying sampling density, $f_\theta(x)$, and estimation of the parameter vector θ. Parametric modeling entails two risks of bias: in estimation of θ and incorrect specification of f_θ. Nonparametric density estimation provides a consistent algorithm for nearly any continuous density and avoids the specification step. Although the cumulative distribution and probability density functions carry the same information, densities are more easily interpreted than distributions, especially in more than one dimension, so our focus on the density is appropriate.

Density estimation is broadly applicable for exploring data relationships, presenting data summaries, and constructing sophisticated nonparametric models of biostatistical data. Graphical representation of data is a powerful tool for summarization. Three simple exploratory graphical summaries are the box-and-whiskers plot (or boxplot), the stem-and-leaf plot, and the histogram. Consider the cholesterol levels of 320 males with diagnosed coronary artery disease (Scott et al., 1978). Figure 1 displays a boxplot of these data. The data appear symmetric with a few outliers. The various percentiles displayed in the boxplot do not hint of any unusual feature such as we see in Figure 2 in the right histogram, which shows mild evidence of bimodality; however, even with 320 observations, the weight of evidence is probably not strong. Observe that the two histograms have the same bin width, but their meshes are shifted. The stem-and-leaf plot (not shown) indicates specific data values but otherwise has no frequency information beyond the histogram.

Figure 1 about here.

Figure 1. Boxplot of $\log_{10}$ cholesterol data ($n = 320$).

Figure 2 about here.

Figure 2. Two histograms of the cholesterol data with the same bin width but shifted meshes. The first appears Gaussian; the second appears bimodal.

HISTOGRAM

A histogram is the simplest density estimator and is one example of a frequency curve, using tabulation of data in bins. Frequency curves have had an important role to play since their introduction by John Graunt, who searched for patterns of death in the Bills of Mortality collected during the London plague of the Seventeenth century. Graunt performed a primitive survival analysis by grouping age of death in five-year-wide intervals. Here we highlight the theoretical properties of a histogram.

Given a sample $x_1, x_2, \ldots, x_n$ contained in an interval (a, b), the histogram is constructed over a partition $\{t_k\}$ of (a, b) into M intervals, $B_k = [t_{k-1}, t_k)$, such that $a = t_0 < t_1 < \cdots < t_M = b$. Let the bin width of B_k be denoted by $h_k = t_k - t_{k-1}$. Let the bin count be denoted by ν_k, so that $\sum_{k=1}^{M} \nu_k = n$. Then the histogram estimate of the density, $f(x)$, is given by

$$\hat{f}_H(x) = \frac{\nu_k}{nh_k} = \frac{\nu_k}{n\,(t_k - t_{k-1})} \qquad x \in B_k\,,$$

and zero outside $[a,b)$. Observe that $\hat{f}_H$ satisfies both conditions of a density function as $\hat{f}_H \geq 0$ and $\int \hat{f}_H = 1$.

Usually all the bin widths are chosen to be equal, $h_k = h$, as in the stem-and-leaf plot. While any choice of the bin width will produce an informative diagram, the notion of an optimal bin width has been studied extensively (Scott, 1979; Freedman and Diaconis, 1981). At each point x, $\hat{f}_H(x)$ is generally biased, so that the mean squared error (MSE) is an appropriate criterion. A Taylor's series analysis reveals that the sum of the pointwise variance and squared bias decomposition is given by

$$\mathrm{MSE}[\hat{f}_H(x)] = \frac{f(x)}{nh} + \frac{1}{12}h^2 f'(x)^2\,.$$

(Terms omitted are of lower order n^{-1}.) If h is too large, then the histogram has too few bins and is "oversmoothed" — exhibiting low variance but high bias. On the other hand, if h is too small, then $\hat{f}_H(x)$ has too many bins and is "undersmoothed" — suffering high variance.

For an entire histogram, the pointwise mean squared error criterion may be integrated to give a global criterion, the integrated mean squared error (IMSE):

$$\mathrm{IMSE}(\hat{f}_H) = \int_{-\infty}^{\infty} \mathrm{MSE}[\hat{f}_H(x)]\,dx = \frac{1}{nh} + \frac{1}{12}h^2 R(f')\,,$$

where $R(f') = \int_{-\infty}^{\infty} f'(x)^2\,dx$ is referred to as the "roughness" of the unknown density function. Ordinary calculus reveals the optimal bin width

$$h_H^*(f) = 6^{1/3} R(f')^{-1/3} n^{-1/3}.$$

The optimal IMSE decreases to zero at the rate $n^{-2/3}$, far short of the usual parametric rate of n^{-1}. Only one function of the unknown density is relevant to the optimal bin width.

Using a normal density, $\phi = N(\mu, \sigma^2)$, as a reference, $R(\phi') = 1/(4\sqrt{\pi}\sigma^3)$, so that

$$h_H^*(\phi) = 3.5\,\sigma\,n^{-1/3}\,.$$

In practice, the unknown standard deviation, σ, is replaced by the sample standard deviation or a robust estimate. This formula is more general and useful than its motivation might suggest. By considering *all* possible densities in $h_H^*(f)$, it has been found that $h_H^*(\phi)$ is within 7% of a theoretical upper bound (Terrell and Scott, 1987). Thus for real data, use of $h_H^*(\phi)$ will almost always results in mild oversmoothing of the data. In no case should a wider bin width be selected. More refined choices will be discussed in Choosing Smoothing Parameters below.

IMPROVEMENTS ON HISTOGRAMS

Frequency Polygon

A continuous version of the histogram is the frequency polygon (FP), which is formed by interpolating the midpoints of a histogram. The theoretical properties of the frequency polygon are superior to the histogram. Scott (1985a) showed that its IMSE decreased at the much faster rate of $n^{-4/5}$, and that $h_{FP}^*(\phi) = 2.15\,\sigma\,n^{-1/5}$. (This is within 8% of the oversmoothed upper bound bin width.) The optimal frequency polygon uses wider bins than the optimal histogram. (The optimal histogram requires more and narrower bins to try to approximate the density where the slope is greatest.)

Averaged Shifted Histogram

Both the histogram and frequency polygon share a second design parameter that can have a large visual impact, especially for small sample sizes, as demonstrated in Figure 2. This parameter is the bin origin, t_0. For a fixed bin width, there are an unlimited number of possible choices for t_0 in the interval $(a-h, a]$. In many situations, t_0 may be viewed as a nuisance parameter. Scott (1985b) proposed averaging over shifted meshes to eliminate the bin edge effect. To be specific, form a finer (narrower) mesh of width $\delta = h/m$ for some positive integer m, and let B_k and ν_k refer to this new, finer set of bins. Then for $x \in B_k$, there are m different histograms with bin width $h = m\delta$ that cover (include) bin B_k. The bin counts for these m shifted histograms are $(v_{k-m+1} + \cdots + \nu_k)$ to $(\nu_k + \cdots + \nu_{k+m-1})$. A little algebra reveals the mean or averaged shifted histogram (ASH) as

$$\begin{aligned} \hat{f}_A(x) &= \frac{1}{m} \sum_{j=1-m}^{m-1} \frac{(m - |j|)\, \nu_{k+j}}{nh} \\ &= \frac{1}{nh} \sum_{j=1-m}^{m-1} \left(1 - \frac{|j|}{m}\right) \nu_{k+j} \qquad x \in B_k\,. \end{aligned} \tag{1}$$

Figure 3 displays an example for $m = 14$ for the cholesterol data as in Figure 2. The estimate is not only visually smoother and more appealing than the histogram, but also shares the improved theoretical properties of the frequency polygon, while, in fact, being about 20% more efficient. The two modes are more clearly uncovered.

Figure 3 about here.

Figure 3. Weighted average of 14 shifted histograms of the cholesterol data. The weights $\{w_m(j)\}$ used were derived from the kernel $K(t) = 315/256\,(1-t^2)^4$ according to equation (3).

Kernel Estimator

As $m \to \infty$, the ASH takes on an equivalent and widely studied form. Since ν_k is either 0 or 1 in the limit (excluding ties), the sum in equation (1) can be re-expressed as a sum over the n data points:

$$\hat{f}_K(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{2}$$

with $K(t) = [1 - |t|]_+$, where $[x]_+$ is the positive part of x, or zero. This so-called kernel estimator was extensively studied by Rosenblatt (1956) and Parzen (1962), although first proposed in a technical report by Fix and Hodges (1951). Similar ideas were more developed in spectral density estimation at that time. The parameter h is no longer a bin width, *per se*, and is called a smoothing parameter.

The kernel density estimator turns out to be quite general. Any probability density that is square integrable can be selected for the kernel. The usual requirements are that $\int K = 1$ and $\int xK = 0$. Picking a symmetric probability density satisfies both. Even higher order rates of convergence such as $n^{-8/9}$ are possible if $\int x^2 K = 0$, but such kernels must take on negative values.

Apparently, the kernel estimator is a mixture of n densities, each centered on a data point. Computationally, kernel estimation can be quite expensive for large samples. Pre-binning the data is an accepted technique for speeding the evaluation. This is equivalent to the ASH with weight function $w_m(j) = 1 - |j|/m$ in equation (1) replaced by

$$w_m(j) = \frac{mK(j/m)}{\sum_{i=1-m}^{m-1} K(i/m)} \tag{3}$$

for kernels defined on $[-1, 1]$. Fast Fourier Transformations may be used if the kernel is the Gaussian pdf (Silverman, 1982).

There are many other techniques based on filtering or orthonormal estimation, but these may be shown to be equivalent to the use of a particular "equivalent" kernel function. Recent interest in the use of wavelets as an orthonormal basis illustrates the generality of kernel methodology.

MULTIVARIATE DENSITY ESTIMATION

The kernel estimator has a simple extension to two dimensions (and similarly for more dimensions) by using a bivariate probability density, $K(x, y)$, as the kernel:

$$\hat{f}_K(x, y) = \frac{1}{h_x h_y} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_x}, \frac{y - y_i}{h_y}\right), \tag{4}$$

where each coordinate direction has its own smoothing parameter. Some authors (Wand and Jones, 1993), advocate the use of the correlation coefficient as an additional smoothing parameter. Similarly, the averaged shifted histogram may be defined by averaging over histograms shifted along the both x and y axes. The latter is illustrated in Figure 4 on the coronary artery disease data. (The bivariate kernel was taken as the product of two univariate kernels used in Figure 3.) The bivariate ASH clearly reveals the non-Normality of the data, suggesting an extra mode or two much more strongly.

Figure 4 about here.

Figure 4. Bivariate averaged shifted histogram of cholesterol and triglyceride blood concentrations for 320 patients in earlier figures. At least two patient clusters are suggested.

Examples in three and four dimensions, including visualization, with applications to clustering, discrimination, and regression are presented in Scott (1992).

CHOOSING SMOOTHING PARAMETERS

A good deal of research has appeared on improved and automatic algorithms for choosing h in equations 1, 2, and 4. Some focus on plug-in estimates of quantities such as $R(f')$, while others rely on modification of maximum likelihood or information measures.

We mention only one method, least-squares cross-validation, which if not the most efficient procedure, is clearly the most general and widely applicable. The IMSE criterion is the average of the integrated squared error (ISE) between $\hat{f}_h$ and f:

$$\begin{aligned}
\mathrm{ISE}(h) &= \int \left[\hat{f}_h(x) - f(x)\right]^2 dx \\
&= \int \hat{f}_h(x)^2 - 2\int \hat{f}_h(x) f(x) dx + \int f(x)^2 dx \\
&= R(\hat{f}_h) - 2E\hat{f}_h(X) + R(f).
\end{aligned}$$

Rudemo (1982) and Bowman (1984) observed that the third term, $R(f)$, is constant for all choices of h and can be ignored. The first integral is directly computable as h varies.

Finally, the second term has an unbiased estimator:

$$-\frac{2}{n}\sum_{i=1}^{n}\hat{f}_{h,-i}(x_i),$$

where $\hat{f}_{h,-i}(x_i)$ is the density estimate based on the $n-1$ points with x_i omitted. For the histogram, the least-squares cross-validation functional is

$$\mathrm{CV}(h) = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)n^2h}\sum_k \nu_k^2 \, .$$

The functional can be immediately extended to pick the bin origin, t_0 as well. The minimizer of $\mathrm{CV}(h)$ may be found by grid search or numerical methods. Several CV formulae for kernel estimates, including the multivariate case, are given in Sain, et al. (1994).

NEW DIRECTIONS AND RESOURCES

Locally adaptive density estimation will play an increasingly important role now that that technology has begun to appear. Promising approaches include plug-in (Schucany, 1989), log-spline (Kooperberg and Stone, 1991), wavelet (Donoho, et al., 1996), local likelihood (Loader, 1996; Hjort and Jones, 1996), and local CV bandwidths (Sain and Scott, 1996). The reader is cautioned about overfitting difficulties, as many degrees of freedom are consumed during the estimation of the adaptive smoothing parameters (either explicitly or implicitly). A more conservative approach is to follow some of the simple transformation ideas of Ruppert and Wand (1992).

Historical information and greater detail are available in the following selection of monographs: Tapia and Thompson (1978), Silverman (1986), Devroye (1987), Härdle (1991), Scott (1992), Tarter and Lock (1993), Wand and Jones (1995), Fan and Gijbels (1996), and Simonoff (1996). These references also discuss the wide array of closely related nonparametric techniques including regression, spectral densities, time series, cluster analysis, discrimination, and survival and hazard estimation. Software is available in packages such as SAS, Splus, and Systat, for example, or at internet sites such as statlib and netlib.

REFERENCES

Devroye, L. (1987). *A Course in Density Estimation.* Birkhäuser, Boston.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., and Picard, D. (1996). *Annals of Statistics* **24**, 508-539.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications.* Chapman and Hall, New York.

Fix, E. and Hodges, J.L. (1951). Reprinted by Silverman, B.W. and Jones, M.C. (1989). *International Statistical Review* **57**, 233-247.

Freedman, D. and Diaconis, P. (1981). *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **57**, 453-476.

Härdle, W. (1991). *Smoothing Techniques with Implementation in S.* Spring-Verlag, New York.

Hjort, N.L. and Jones, M.C. (1996) *Annals of Statistics* **24**, 1619-1647.

Kooperberg, C. and Stone, C.J. (1991). *Computational Statistics and Data Analysis* **12**, 327-347.

Loader, C.R. (1996) *Annals of Statistics* **24**, 1602-1618.

Parzen, E. (1962). *Annals of Mathematical Statistics* **33**, 1065-1076.

Rosenblatt, M. (1956). *Annals of Mathematical Statistics* **27**, 832-837.

Ruppert, D. and Wand, M.P. (1994). *Australian Journal of Statistics* **34**, 19-29.

Sain, S.R., Baggerly, K.A., and Scott, D.W. (1994). *Journal of the American Statistical Association* **89**, 807-817.

Sain, S.R. and Scott, D.W. (1996). *Journal of the American Statistical Association* **91**, 1525-1534.

Schucany, W.R. (1989). *Statistics and Probability Letters* **7**, 401-405.

Scott, D.W. (1979). *Biometrika* **66**, 605-610.

Scott, D.W. (1985a). *Journal of the American Statistical Association* **80**, 348-354.

Scott, D.W. (1985b). *Annals of Statistics* **13**, 1024-1040.

Scott, D.W. (1992). *Multivariate Density Estimation.* Wiley, New York.

Scott, D.W., Gotto, A.M., Cole, J.S., and Gorry, G.A. (1978). *Journal of Chronic Diseases* **31**, 337-345.

Silverman, B.W. (1982). *Applied Statistics* **31**, 93-99.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Simonoff, J.S. (1996). Smoothing Methods in Statistics. Springer, New York.

Tapia, R.A. and Thompson, J.R. (1978). *Nonparametric Probability Density Estimation.* Hopkins Press, Baltimore.

Tarter, M.E. and Lock, M.D. (1993). *Model-Free Curve Estimation.* Chapman and Hall, New York.

Terrell, G.R., and Scott, D.W. (1985). *Journal of the American Statistical Association* **80**, 209-214.

Wand, M.P. and Jones, M.C. (1993). *Journal of the American Statistical Association* **88**, 520-528.

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