AVERAGED SHIFTED HISTOGRAM

UNIVARIATE DATA

For a random sample of size n, the classic histogram takes the value

$$\hat{f}(x) = \frac{\text{bin count}}{nh},$$

where h is the width of a bin. Important details of proper histogram construction are usually overlooked. In practice, continuous data have a finite number of significant digits, with resultant accuracy $\pm\delta/2$. In this view, the raw data have been rounded and assume values at the midpoints of a finer histogram mesh with bin width δ. Let the kth such bin be denoted by $B_k = [t_k, t_{k+1})$, $-\infty < k < \infty$, where the bin edges $\{t_k\}$ satisfy $t_{k+1} - t_k = \delta$, for all k. If ν_k is the bin count for B_k, then $\sum_k \nu_k = n$. Thus the classic histogram should obey two constraints: first, the bin width should be a multiple of δ, that is, $h = m\delta$, for some integer m; and second, the bins of the histogram should conform to the finer mesh, for example, $[t_k, t_{k+m})$. The bin count for this wider bin, $[t_k, t_{k+m})$, would be computed as $\sum_{i=0}^{m-1} \nu_{k+i}$.

If the point of estimation, x, falls in the (narrower) bin, B_k, precisely which histogram with bin width h should be selected? Clearly, there are exactly m such *shifted histograms*, with explicit bin intervals ranging from $[t_{k-m+1}, t_{k+1})$ to $[t_k, t_{k+m})$. The ordinary *averaged shifted histogram* (ASH) estimates the density at $x \in B_k$ as the arithmetic mean of these m shifted histogram estimates. A simple calculation shows that the ordinary ASH is given by

$$\hat{f}(x) = \frac{1}{m} \sum_{i=1-m}^{m-1} \frac{(m - |i|)\,\nu_{k+i}}{nh}$$

$$= \frac{1}{nh} \sum_{i=1-m}^{m-1} \left(1 - \frac{|i|}{m}\right) \nu_{k+i} \quad x \in B_k\,.$$

As $m \to \infty$, Scott [3] showed that $\hat{f}(x)$ converges to the kernel density estimator

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

with the triangle kernel, $K(t) = 1 - |t|$ on $(-1, 1)$. The ASH may be generalized to mimic any kernel $K(t)$ defined on $(-1, 1)$ such that $K(\pm 1) = 0$ by

$$\hat{f}(x) = \frac{1}{n\delta} \sum_{i=1-m}^{m-1} w_m(i) \nu_{k+i} \qquad x \in B_k$$

by defining the weights

$$w_m(i) = \frac{K\left(\frac{i}{m}\right)}{\sum_{j=1-m}^{m-1} K\left(\frac{j}{m}\right)}.$$

In this form, the ASH is seen as a discrete convolution of adjacent bin counts. Silverman [6] provided a Fast Fourier Transform procedure for the Normal kernel, which does not have finite support. Fan and Marron [1] compare many algorithms.

MULTIVARIATE ASH

Binning seems an effective device in dimensions up to four or five. For example, with bivariate data, (x_1, x_2), construct a fine mesh of size $\delta_1 \times \delta_2$. Then construct bivariate histograms with bins of size $h_1 \times h_2$, where $h_1 = m_1\delta_1$ and $h_2 = m_2\delta_2$. Then the bivariate ASH is the average of the $m_1 \times m_2$ shifted histograms, and for $(x_1, x_2) \in B_{k\ell}$,

$$\hat{f}(x_1, x_2) = \frac{1}{n\delta_1\delta_2} \sum_{i=1-m_1}^{m_1-1} \sum_{j=1-m_2}^{m_2-1} w_{m_1}(i) w_{m_2}(j) \nu_{k+i,\ell+j}$$

with obvious extension to more dimensions. Scott [4] discusses visualization of trivariate and quadravariate densities with applications in clustering and discrimination.

COMPUTATION AND ASYMPTOTICS

The computational advantage of the ASH derives from the fact that the smoothing is applied to perhaps 50-500 bin counts, rather than to the raw data themselves. ASH software is available from *Statlib* or from ftp.stat.rice.edu by anonymous ftp.

From the point of view of statistical efficiency, the ordinary ASH represents a compromise between the histogram and triangle kernel estimator. For the univariate ASH, the asymptotic mean integrated squared error (AMISE) is

$$\text{AMISE}(h, m) = \frac{2}{3nh}\left(1 + \frac{1}{2m^2}\right) + \frac{h^2}{12m^2} R(f') + \frac{h^4}{144}\left(1 - \frac{2}{m^2} + \frac{3}{5m^2}\right) R(f''),$$

where $R(\phi) = \int \phi(x)^2 dx$. When $m = 1$, the bias has the usual $O(h^2)$ of the histogram, whereas the bias is of order $O(h^4)$ as $m \to \infty$ as for the positive kernel estimator. If a piecewise linear interpolant of the ASH midpoints is used, then the bias is $O(h^4)$ for all m. Thus, linear interpolation is almost always recommended for the ASH.

The exact optimal choices of h and m depend on the unknown density through $R(f'')$. However, Terrell and Scott [7] provide useful upper bounds for h using any reasonable estimate of the standard deviation: for the ordinary ASH with $m = 1$, $h < 3.73\sigma n^{-1/3}$; for the ASH with $m = \infty$, $h < 2.78\sigma n^{-1/5}$. These estimates converge at the rates $O(n^{-2/3})$ and $O(n^{-4/5})$, respectively. However, if the linear interpolant is used, the bound for the ASH with $m = 1$ is $h < 2.33\sigma n^{-1/5}$; the $m = \infty$ bound is unchanged. If δ is fixed, these formulae give upper bounds for m. A discussion of algorithms for choosing h may be found in Scott and Terrell [5] and Scott [4]. A simple table of factors to obtain smoothing parameters for use with other kernel weights is given on page 142 in Scott [4].

ASH REGRESSION

An efficient nonparametric regression estimate may be constructed by using the multivariate ASH. For example, with data (x, y), construct the bivariate ASH and then compute the conditional mean of the estimate. The result is particularly simple as $\delta_y \to 0$:

$$\hat{m}(x) = \frac{\sum_{i=1-m_1}^{m_1-1} w_{m_1}(i) v_{k+i} \bar{y}_{k+i}}{\sum_{i=1-m_1}^{m_1-1} w_{m_1}(i) v_{k+i}} \qquad x \in B_k,$$

where $\bar{y}_k$ is the average of the y_i's for those points (x_i, y_i) where $x_i \in B_k$ and ν_k is the univariate bin count. This estimator was introduced by Härdle and Scott [2]; two and three dimensional versions with application to spatial estimation from sample surveys of agriculture data are given in Whittaker and Scott [8].

EXAMPLES

Bradford Brown (see [4], Appendix B.4) measured the thickness of 90 U.S. pennies to the nearest tenth of a mil. Two shifted histograms plus an ASH are shown in Figure 1, all with smoothing parameter $h = 1.5$ mils. The visual impression of the two histograms is markedly different. The ASH essentially removes the effect of the choice of bin origin, which is a nuisance parameter. With a smaller smoothing parameter ($h = 0.9$), two extra modes appear (rather than only one) at 55.9 and 57.1 mils.

A fuller view of these data may be obtained by plotting them as a time series as in Figure 2. Contours of the bivariate ASH, together with the regression ASH, are also displayed. Apparently, the thickness of pennies has changed more than once since World War II. As an exercise, try to spot the two extra clusters mentioned in the previous paragraph.

References

[1] Fan, J. and Marron, J.S. (1994). *J. Comp. Graph. Stat.*, **3**, 35-56.

[2] Härdle, W. and Scott, D. W. (1992). *Comp. Stat.*, **7**, 97-128.

[3] Scott, D. W. (1985). *Ann. Statist.*, **13**, 1024-1040.

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

[5] Scott, D. W. and Terrell, G. R. (1987). *J. Amer. Statist. Ass.*, **82**, 1131-1146.

[6] Silverman, B. W. (1982). *J. R. Statist. Soc. B*, **31**, 93-97.

[7] Terrell, G. R. and Scott, D. W. (1985). *J. Amer. Statist. Ass.*, **80**, 209-214.

[8] Whittaker, G. and Scott, D. W. (1995). *Proc. 26th Interface Comp. Sci and Stat.*.

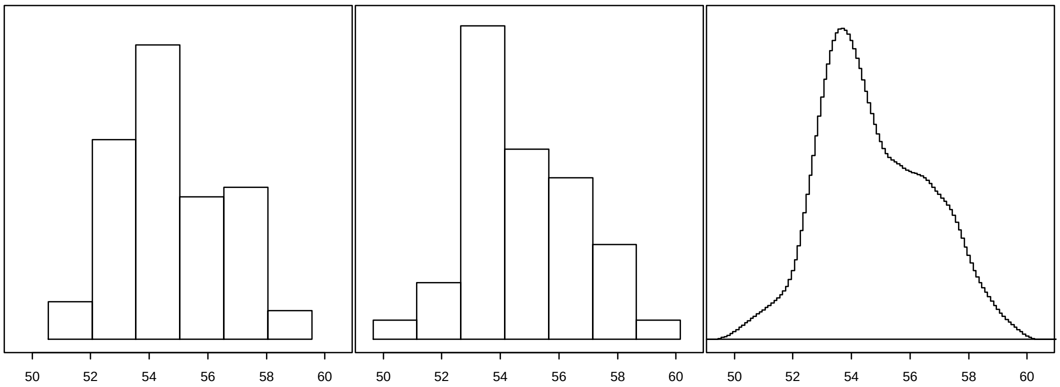

Figure 1: Two shifted histograms and ASH of the penny thickness (*mils*) data. The same smoothing parameter $h = 1.5$ *mils* was used for each. The bin origins for the histograms were 50.55 and 49.65. The triweight kernel was used for the ASH with $\delta = 0.1$ and $m = 15$.

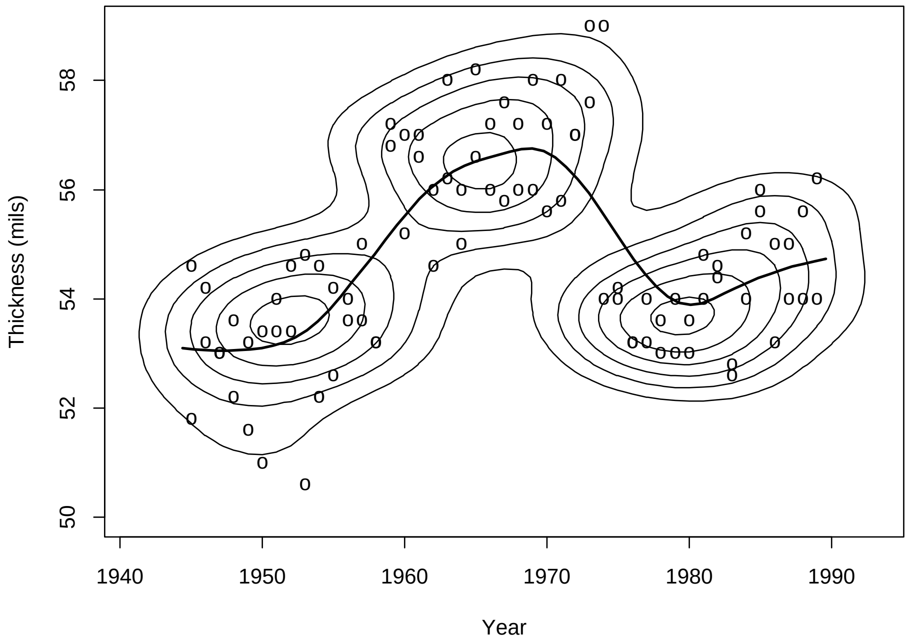

Figure 2: Bivariate ASH density and regression estimates of the penny data.