Variations on the histogram

Lorraine Denby and Colin Mallows
Avaya Labs, Basking Ridge, NJ

January 19, 2007

Abstract

It is usual to choose to make the bins in a histogram all have the same width. One could also choose to make them all have the same area. These two options have complementary strengths and weaknesses; the equal-width histogram oversmooths in regions of high density, and is poor at identifying sharp peaks; the equal-area histogram oversmooths in regions of low density, and so does not identify outliers. We describe a compromise approach which avoids both of these defects. We argue that relying on asymptotics of the Integrated Mean Square Error leads to inappropriate recommendations.

Keywords: Diagonally-cut histogram, equal-area histogram, asymptotics, IMSE.

1 The usual histogram

The standard equal-bin-width histogram is a venerable tool. Starting with Sturges (1926), many authors have proposed rules for choosing the bin-width, usually called h, or equivalently the number of bins, *nbins*, always regarding the histogram as an estimator of an underlying density, and relying on large-n asymptotics. Histograms constructed according to these rules have been compared with other density estimators, notably kernel and spline estimators. We give a selection of references to this literature.

We regard the histogram primarily not as an estimator, but as a data-analytic tool, suitable for a first inspection of the data. We believe that reliance on the asymptotic formulas for integrated mean-square error (IMSE) is actively misleading; see Section 5 below. Properties that we think important include ease of use (e.g. effective default settings for design parameters) and effectiveness in identifying unexpected structure, for example, gaps, outliers, and spikes. Kernel estimators and splines are not good at finding these artifacts in data. The usual histogram shows outliers well, but oversmooths in regions of high density, and does not respond well to spikes in the data. Adaptive-kernel smoothers have been studied, but these are complicated to use and to study. The empirical cumulative distribution function (ecdf), in which the i-th largest x is plotted against $(i - 1/2)/n$ (or $i/(n + 1)$), in principle presents all the information in the data, but we have found that it is often hard to identify spikes in an ecdf.

2 The equal-area histogram

An equal-area histogram (e-a hist) consists of some number *nbins* of bins, each containing (as nearly as possible) the same number of data-values. Given data $x = (x_1, ..., x_n)$ we choose cell-boundaries $\xi = (\xi_0, ..., \xi_k)$ spanning the range of the data, so that the number of data-values between ξ_{i-1} and ξ_i is the same for each i. Clearly the bins will tend to be narrow where the population density is large, and wide where is it small. We need a rule for choosing the number of bins, or equivalently the number of values in each bin. We believe that the usual rules for choosing the number of bins in a standard histogram are adequate here also.
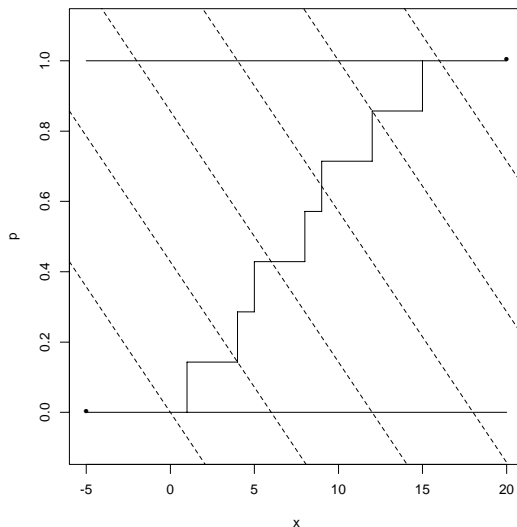
Figure 1: Construction of a dhist. The ecdf is cut by equally-spaced diagonal lines.

The e-a hist seems not to have attracted much attention. Scott (1992) calls the e-a hist the result of using a "percentile mesh" for defining the bins. Given a vector of data, the e-a hist is as easy to use as the usual histogram; and when the population density is smooth it gives a visual impression that is very similar to that of the standard histogram, so that human perceptions do not need to be recalibrated. An advantage the usual histogram has over the e-a hist is that the bins can be set up before the data are collected, so that only counts need to be recorded, not detailed data-values. But if this is done, it will usually not be possible to ensure in advance (in an initial exploratory investigation) that the bins have been chosen well. A two-stage strategy should work well; and this is available for the e-a hist also.

We recommend making the counts in the bins exactly equal, by counting a data-value that lies on the boundary between two bins as contributing partially to each of those bins. Note that we do not need to evaluate these fractions explicitly. We need merely find the proper values for the breaks, and draw the equal-area histogram so that each bin has the same area, namely $n/nbins$.

A major strength of the e-a hist is that it shows spikes well. Its main defect is that it oversmooths in the tails. It does not show individual outliers, but smears them into wide bins.

3   A new proposal

Consider the ecdf. This is a plot of p against the ordered x-values. The usual histogram defines bins by cutting the x-scale at equally-spaced points; the e-a hist defines bins by cutting the p-scale at equally-spaced points. Our new idea is to cut the ecdf plot diagonally. This new display, which we term "dhist" (for diagonally-cut histogram) preserves the desirable features of both the equal-width hist and the equal-area hist. It will show tall narrow bins like the e-a hist when there are spikes in the data, and will show isolated outliers just like the usual histogram.

In more detail, choose two parameters a and h, or equivalently a and $nbins$, and cut the ecdf by lines $x + ap = kh$ for integer k. If the kth line cuts the ecdf at the point (x_k, p_k), the kth bin for the dhist is (x_k, x_{k+1}), and it contains $n_k = n(p_{k+1} - p_k)$ points. Thus the height of the kth bin is $n_k/(x_{k+1} - x_k)$.

Note that $a = 0$ gives the usual histogram; $a \to \infty$ approaches the e-a hist. We suggest choosing $a = 5 * IQR(x)$, where IQR is the inter-quartile range of the data. This rule seems to work well for data whose center is uniform or approximately Normal. Some examples are given below.

A detail that must be addressed is that in this construction we need to consider four ways a cut-line can intersect the ecdf, as illustrated in Figure 1, which shows one example of each type. The

cut can be (1) at a lower corner of the ecdf, (2) on a flat section, (3) at a riser, or (4) at an upper corner. The only case that causes any difficulty is (3), where we should allocate the weight $(1/n)$ for that value of x between the two adjacent bins. Thus in Figure 1 the five bins contain respectively $1, 2, 1.5, 1.5, 1$ observations.

A presentation detail concerns what happens when the bin-width is very small; it can even be zero. We have found it necessary to artificially bound the height of the histogram bars, since otherwise the display is dominated by one or more very tall bins and all detail in the rest of the display is not visible. Our suggested resolution of this dilemma is to truncate the heights of the tallest bins (which can be infinite) at some multiple, say 2, of the height of the next-highest non-zero width bin, and to present the count associated with a very narrow bin by drawing the corresponding area as a "flag" attached to the top of the dhist bin. The area of this flag can be compared visually with the areas of the other conventional bins.

The dhist can be implemented (to a good approximation when n is large) in Splus and R using two calls to the `hist` function. Given a sorted data-vector `x` with length `n`, we form a vector of p-values `p = (1:n)/n`. We choose some number a. We call `hist(x+a*p, plot=FALSE)` to find the counts for `x+a*p`; these are the counts we require for the dhist, and we can determine the breaks for `x` from them. We can then call `hist` again (with the argument `plot=TRUE`) to draw the dhist. This method is only approximate because it replaces the ecdf by the set of "upper corner" points. An R function which fully implements the idea in Figure 1 is available from the authors. It uses four calls to the `cut` function. We hope that software providers such as SAS and SPSS will see fit to provide a dhist facility in future releases.

4   Examples

We present four examples. In all cases, the number of bins is determined by Sturges's rule. The first is from the study that motivated this work, a Six Sigma quality improvement project. The purpose of the study was to reduce the time from trouble report to restored service. The particular repairs of interest were those that needed a technician dispatched and could not be fixed remotely. One important aspect in the analysis was the identification of features in the distribution of the variable of interest (either time-to-restore or subintervals of time-to-restore) for various subsets of the data. The data was not only long- tailed but also had many non-smooth features such as multimodality, spikes and outliers. We started by producing numerous ordinary histograms using the default number of bins in Splus. Upon inspection and comparison of a few of these we quickly discovered that these histograms often masked the salient features that we were seeking. This led to the work on dhist. The histograms produced via the dhist method using the Sturges default number of bins and $a = 5 * IQR$ worked well in displaying the features of each set of data. We found little need for further refining the choice of the number of bins.

Figure 2 addresses 506 durations of one step of the process from trouble report to service restoration. We show the ecdf, a standard histogram, an equal-area histogram, and a diagonally-cut histogram. All three histograms have the same number of bins, namely 14, chosen by Sturges's rule, and we have taken $a = 5 * IQR$. We see that the standard histogram does not show the spike in the data, and the e-a hist smears out the outliers; the dhist shows both effects.

Figure 3 relates to data from Harrison and Rubinfeld (1978). This dataset contains information collected by the U.S Census Service concerning housing in the area of Boston Massachusetts. It contains 506 observations and 14 variables.

For most of the variables the dhist looks very like the ordinary histogram; but for the PT variable (pupil-teacher ratio by town) the dhist shows that the distribution contains a sharp spike at $x = 20.2$. This may not mean anything, but certainly merits investigation. We feel that it is useful to have a technique that displays such effects when they are present in the data.

Figure 4 shows an artificial example, exhibiting both a spike and outliers. It is a mixture of three distributions: 775 points drawn from a standard Normal distribution, 150 points at $x=7$, and 75 points distributed uniformly over the interval (0,10). We show the histogram, a dhist with $a=5*IQR$, and the e-a hist. We see that the histogram fails to show the extreme sharpness of the spike, while
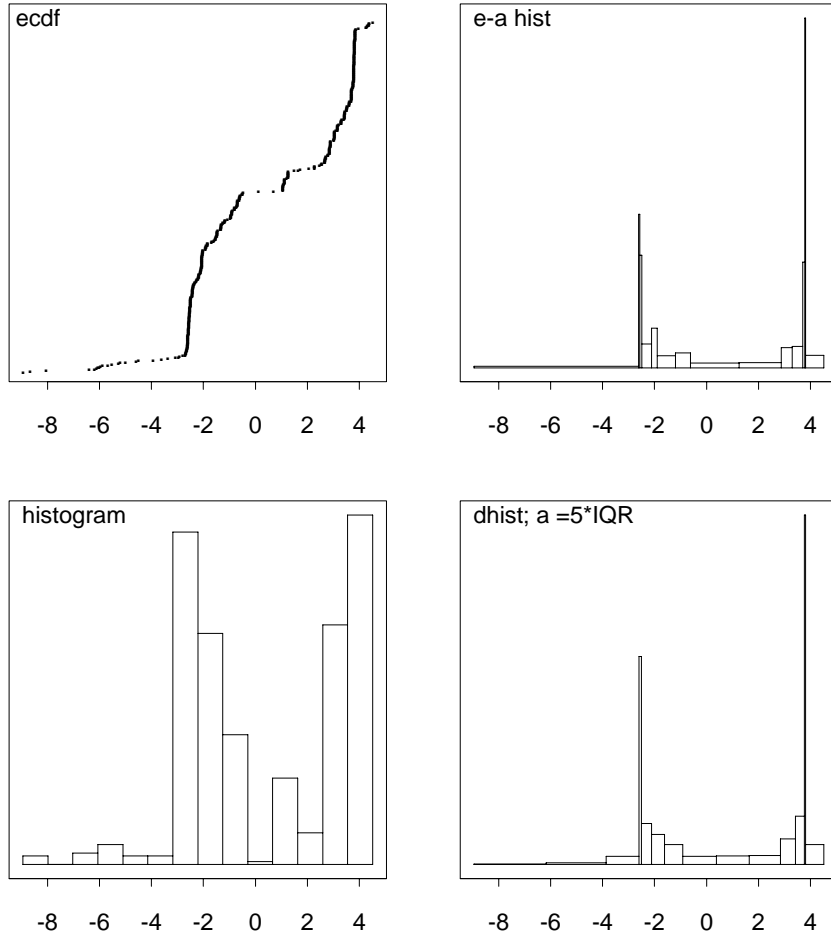
Figure 2: ECDF for Time Interval, with an ordinary histogram, an e-a hist, and a dhist with $a = 5 * IQR$. N = 506. x-axis is log base 2 of time.
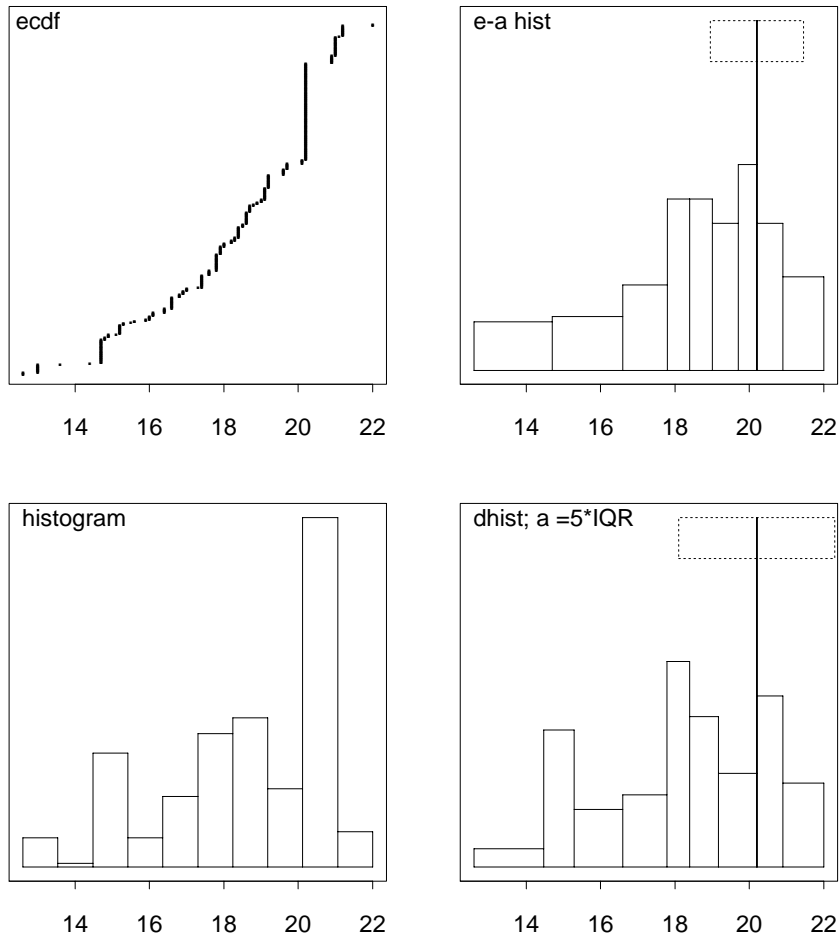
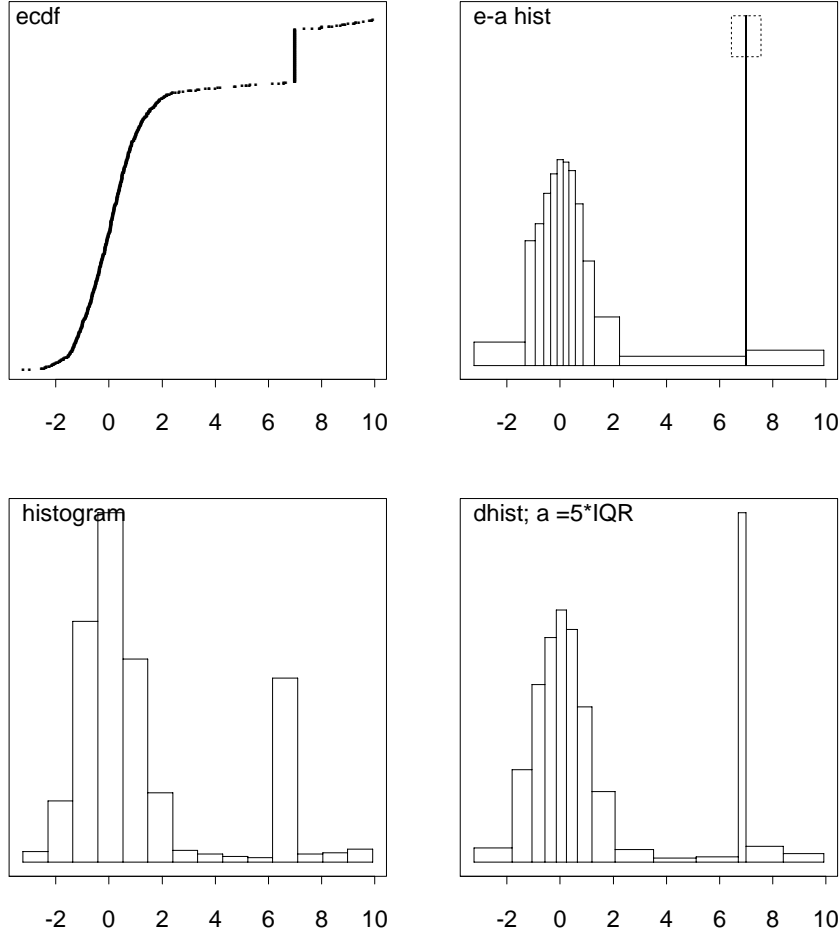Figure 3: ECDF for pupil/teacher ratio, with an ordinary histogram, an e-a hist, and a dhist. N = 506.

Figure 4: Artificial example. Ordinary histogram, e-a hist and dhist. N = 1000.

the e-a hist shows no detail in the outliers. The dhist shows the spike well, and also shows the outliers.

The visual effect of a histogram display may change as we change the position of the left-most bin boundary. Our final example is drawn from a study of round trip time of Voice over IP (VoIP) packets between two devices on the data network. Each observation is the median round trip time of 100 packets sent 20 milliseconds apart from a particular source and destination device on the network. One thousand sets of packets were sent, of which 69 were lost.

Figure 5 shows the ecdf, the e-a hist, the histogram, and three versions of the dhist, with a taken to iequal (a) the IQR, (b) 5*IQR, (c) 11*IQR. We also show what happens when the histogram bins are shifted by 23 units to the right (this is half the width of a bin), and when the dhists are shifted by the same amount. We see that neither of the hists shows the extreme sharpness of the spikes. The e-a hist shows three distinct spikes, while only the dhist with a=11 and offset = 23 shows all three. The e-a hist conceals all the detail in the upper tail. We feel that the compromise value a = 5*IQR is a reasonable default setting.
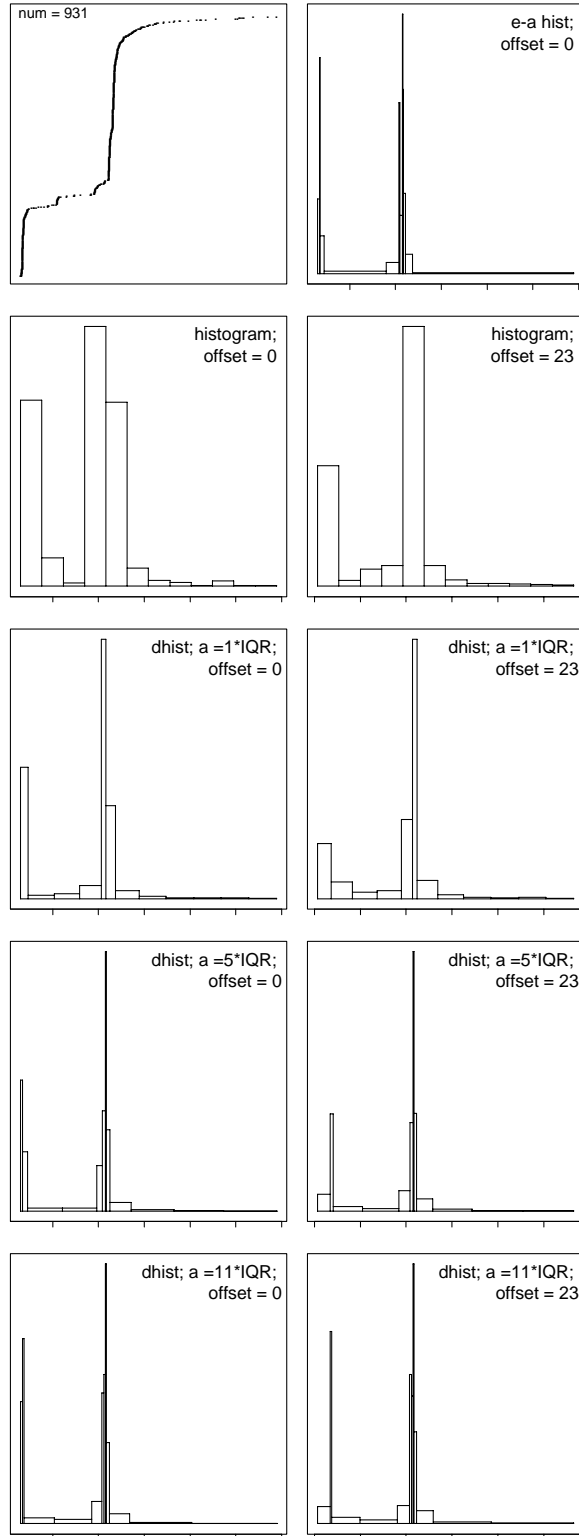
Figure 5: Comparing ordinary histogram, e-a hist and dhist for VoIP packet round trip time. N = 931

7

5   Asymptotics

It is well known that if we regard the histogram as an estimate of a smooth density, and assess its effectiveness by the integrated mean-square error (IMSE)

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

then asymptotically the optimal choice for bin-width, is of order $n^{-1/3}$. Scott (1978) gave the formula $nbins = (6/nI)^{1/3}$ where $I = \int f'(x)^2 dx$. Below we report a similar result for the dhist. We have not succeeded in finding a similar result for the e-a hist, we conjecture that the optimal choice may be of a different order in n, possibly depending on the shape of the tails of the population.

We do not regard these results as being prescriptive for choosing a histogram technique, for the following reasons.

First, the asymptotic theory considers the limiting case as $n \to \infty$, with the number of bins increasing slower than n, so that in the limit there are very many observations in each bin, and the bins are narrow relative to changes in the density. Thus in this limit, in each bin the density can be approximated by at worst a linear function. These calculations do not respond adequately to the presence of spikes of width comparable to h, where the second derivative of the density cannot be ignored. It is possible that a more refined asymptotic theory could allow for this effect. Also, these asymptotic calculations do not respond adequately to the presence of outliers, which by definition have a small number of observations in each bin.

Second, we see no imperative reason to rely on the IMSE; a relative measure

$$IRMSE = \int (\hat{f}(x) - f(x))^2 / f(x) dx$$

has at least an equal claim to attention (compare Pearson's chi-square statistic and Tukey's "rootogram".)

If we allow the bin-widths to vary over the range of the data, Scott (1992, Section 3.2.1) claims that the asymptotic results imply that "the bin width should be relatively wider in regions of higher density to reduce the variance...". Also he claims (Section 3.2.8.2) that bins should be wider in the tails. Both of these recommendations would be reversed if the IRMSE to be used, and both conflict with our intuition regarding the use of a histogram as a data-analytic device. We think that where the data is plentiful, the bins should be narrow, to give more detail in the shape of the data, and showing spikes when they are present. Also, large bins in the tails will conceal outliers, which may be important features of the data. When there are very narrow spikes in the density, as in our examples and several others given by Marron and Wand (1992), the e-a hist and dhist perform much better than the usual histogram in identifying these phenomena; but the IMSE does not respond to them when their probability content is small. We find that among Marron and Wand's 15 examples, only for examples 3,4,11,13 is the optimal value of a (to minimize the IMSE) not at $a = 0$, though there are sharp spikes in several of the other cases.

With these caveats we derive in the Appendix an asymptotic formula for the IMSE of the dhist, derived following the method of Scott (1972). The result is that asymptotically as $n \to \infty$, with a held fixed and h growing slower than n,

$$IMSE \sim V/nh + Bh^2$$

where

$$V = \int f(x)(1 + af(x))dx$$

$$B = (1/12) \int (f'(x)/(1 + af(x)))^2 dx$$

When $a = 0$ we retrieve the result for the usual histogram; the limit as $a \to \infty$ (the e-a hist) cannot be relied on because it involves interchanging two limiting processes) and even after rescaling by putting $h' = h/a$ it gives the answer "infinity" for many standard densities, for example Normal.

The optimizing value for h is $h^* = (V/2nB)^{1/3}$ and the minimum value of the IMSE is $3(BV^2/4n^2)^{1/3}$. Using the wrong value mh^* instead of h^* increases the IMSE by a factor $(m^3 + 2)/(3m)$.

For any given density f, we can find the value of a that minimizes the optimal IMSE. For many smooth densities this is $a = 0$; i.e. the usual histogram is optimal. But for densities that exhibit "spikes" this is not so; for example for the density

$$f(x) = (n/2)(1 - |x|)^{n-1} - 1 < x < 1$$

the optimal a is not zero if $n > 3$.

We thank an anonymous reader for pointing out that in the usual asymptotic situation, the width of the dhist bin that contains the point x is inversely proportional to $1 + af(x)$. Thus when $a = 0$, and also when $f(x)$ is small (in the tails of the distribution) it is independent of $f(x)$, like the standard histogram; and when $af(x)$ is large it is inversely proportional to $f(x)$, like the e-a hist.

6   A comment on ASH

Scott (1992) has a section on "Average Shifted Histograms". In the simplest version of this, we choose a bin-width h and an integer m. We compute m histograms (with bin-width h) with the breaks displaced by $0, h/m, 2h/m, ...(m-1)h/m$. These m histograms are averaged to give the ASH. The result is equivalent to starting with a histogram with bin-width h/m and applying a "triangle" smoother to the counts, with weights proportional to $1, 2, ..., m - 1, m, m - 1, ..., 1$. In the limit as $m- > \infty$ we are applying a continuous triangle smoother to the data, with half-base h.

In principle we could adapt the ASH idea to take care of spikes (with width comparable to h/m), and outliers. To deal with spikes, observe that there will be an abrupt change in the count when a break passes over a spike, so that in principle we could identify these places and use a bin with width h/m there. Also each isolated outlier becomes a triangle of width $2h$ which could be recognized and replaced by a rectangular bin of width h or h/m. Outliers that are separated by less than $2h$ would be harder to deal with.

Implementing these refinements would require several delicate design decisions, and would necessarily involve several arbitrary parameters. The dhist takes care of both problems in a simple way, requiring us to specify just one parameter (the slope a) for which we have what we believe to be a satisfactory default value.

Appendix Asymptotics

Consider the dhist where the cuts are made at angle θ (anticlockwise from the vertical), so $\theta = 0$ for the usual hist, $\theta = \pi/2$ for the e-a hist. Set $\tan(\theta) = a$. First we approximate the resulting dhist by assuming that the bins are defined by cuts of the population cdf, i.e. the j-th bin is (ξ_j, ξ_{j+1}) where

$$\xi_j + a\pi_j = jh \quad j = ..., -1, 0, 1, ...$$

where $\pi_j = F(\xi_j)$. Then K_j, the number of observations (out of n) that fall in the j-th bin, is Binomial$(n, \pi_{j+1} - \pi_j)$.

We follow the method of Scott (1978), neglecting terms of small order. In the j-th bin, the estimate of the population density is

$$\hat{f}(x) = K_j/n(\xi_{j+1} - \xi_j)$$

which has expectation $(\pi_{j+1} - \pi_j)/(\xi_{j+1} - \xi_j)$. Expanding $p = F(x)$ about $x = \xi_j$ we get

$$E(\hat{f}(x)) = f_j + (1/2)(\xi_{j+1} - \xi_j))f_j' + \cdots$$

where $f_j = f(\xi_j), f_j' = f'(\xi_j)$. Hence the bias is approximately

$$E(\hat{f}(x)) - f(x) = ((1/2)(\xi_{j+1} + \xi_j) - x)f_j'.$$

Also the variance is approximately $((\pi_{j+1} - \pi_j)/n(\xi_{j+1} - \xi_j)^2$ which is approximately $f(\xi_j)/n(\xi_{j+1} - \xi_j)$.

We have $h = (\xi_{j+1} - \xi_j) + a(\pi_{j+1} - \pi_j)$ which is approximately $(\xi_{j+1} - \xi_j)(1 + af_j)$ so the integrated variance is approximately $(1/n)\Sigma f_j$ which is approximately V/nh where

$$V = \int f(x)(1 + af(x))dx$$

and the integrated squared bias is

$$\Sigma_j \int_{\xi_j}^{\xi_{j+1}} f_j'^2 (x - (\xi_j + \xi_{j+1})/2)^2 dx$$

which is approximately Bh^2 where

$$B = (1/12)\int (f'(x)/(1 + af(x))^2 dx.$$

The integrated mean square error (IMSE) is thus approximately $V/nh + Bh^2$ and is a minimum when $h = h^*$ where
$$h^* = (V/2nB)^{1/3}$$
The minimum value of the IMSE is asymptotically $3(BV^2/4n^2)^{1/3}$.

Now we argue that this result applies not only to the (non-operational) histogram defined by cutting the population cdf, but also to the dhist itself. Consider the event E that for all j, the j-th bin of the dhist contains K_j observations. Set $D_j = \Sigma_{k \leq j} K_k$. Then the j-th cut is at $x_j = jh - aD_j/n$ and the estimate of the density within the j-th bin is $\hat{f}_j = K_j/n(x_{j+1} - x_j)$. The probability of the event E is
$$\frac{n!}{\prod K_j!} \prod (F(x_{j+1}) - F(x_j))^{K_j}$$
The joint probability of D_j, D_{j+1} is thus

$$\log P = \frac{n!}{D_j!(D_{j+1} - D_j)!(n - D_{j+1})!} F(x_j)^{D_j} (F(x_{j+1}) - F(x_j))^{D_{j+1} - D_j} (1 - F(x_{j+1}))^{n - D_{j+1}}$$

Regarding D_j and D_{j+1} as continuous variables, we differentiate with respect to them, remembering that x_j is a function of D_j. If we approximate $d/dN \log N!$ by $\log N$, we find that $\log P$ is a maximum at $D_j = n\pi_j = nF(\xi_j)$. Differentiating again, and evaluating at the maximum point, we find

$$
\begin{aligned}
(\log P)_{D_j,D_j} &= -(1 + af_j)^2/n(\pi_{j+1} - \pi_j) \\
(\log P)_{Dj,D_{j+1}} &= (1 + af_j)(1 + af_{j+1})/n(\pi_{j+1} - \pi_j) \\
(\log P)_{D_{j+1},D_{j+1}} &= -(1 + af_{j+1})^2/n(\pi_{j+1} - \pi_j)
\end{aligned}
$$

where $f_j = f(\xi_j), f_{j+1} = f(\xi_{j+1})$. We equate these second derivatives to the coefficients in the exponent of a bivariate normal density, thus finding the asymptotic variances and covariance. The final result is that asymptotically the bias of the dhist estimate is

$$\hat{f}_j - f(x) = ((\xi_{j+1} + \xi_j)/2 - x)f_j'$$

and the variance is

$$\mathrm{Var}(K_j)\frac{(1 + af_j)^2}{n^2(\xi_{j+1} - \xi_j)^2}$$

Now $\mathrm{Var} K_j$ is asymptotically $n(\pi_{j+1} - \pi_j)/(1 + af_j)^2$ so the integrated variance is $\Sigma f_j/n$. These results agree with those for the non-operational form of the dhist.

Acknowledgements

References

[1] Birge, L. and Rozenholc, Y. (1999) How many bins should be put in a regular histogram? Preprint; also preprint of the LPMA, 2002.

[2] Diaconis, P. and Freedman, D (1981) On the histogram as a density estimator: L2 theory. *Zeit. Wahr. verw. Gebiete* **57** 453-476.

[3] Hall, P. and Hannan, E. J. (1988) On stochastic complexity and nonparametric density estimation *Biometrika* **75**, pp. 705-714.

[4] Harrison, D. and Rubinfeld, D.L. (1978) Hedonic prices and the demand for clean air *J. Environ. Economics & Management* **5**, pp. 81-102.

[5] Jones, M. C., Marron, J, S, and Sheather, S. J. (1996) A brief survey of bandwidth selection for density estimation *J. Amer. Statist. Assoc.* **91**, pp. 401-407.

[6] Marron, J. S. (1992) Bootstrap bandwidth selection, In *Exploring the Limits of Bootstrap* eds R. LePage and L. Billard, Wiley 249-262.

[7] Marron, J. S. and Wand, M. P. (1992) Exact mean integrated squared error. *Ann. Statist.* **20**, pp. 712-736.

[8] Scott, D. W. (1978) On optimal data-based histograms. *Biometrika* **66**, pp. 605-610.

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

[10] Silverman, B. W. (1986) *Density Estimation* London: Chapman and Hall.

[11] Tapia, R. A. and Thompson, J. R. (1978) *Nonparametric Density Estimation.* Baltimore: Johns Hopkins University Press.

[12] Taylor, C. C. (1987) Akaike's information criterion and the histogram. *Biometrika* **74**, pp. 636-639.

[13] Tukey, J. W. (1977) *Exploratory Data Analysis.* Reading, MA: Addison-Wesley.

[14] Wand, M. P. (1997) Data-based choice of histogram bin width. *American Statistician* **51**, pp. 59-64