Probability Density Estimation

Associated with each random variable, X, measured during an experiment is the cumulative distribution function (cdf), $F(x)$. The value of the cdf at $X = x$ is the probability that $X \leq x$. Features in a set of random data may be computed analytically if the cumulative distribution function is known. The probability that a random sample falls in an interval (a, b) may be computed as $F(b) - F(a)$. This formula assumes the values of X are continuous. Discrete random variables taking on only a few values are easy to handle, thus, this article focuses on continuous random variables with at least an interval scale.

The same information may be obtained from the probability density function (pdf), $f(x)$, which equals $\Pr(\mathrm{X} = \mathrm{x})$ if X takes on discrete values, or the derivative of $F(x)$ if X is continuous. If either $F(x)$ or $f(x)$ is known, then any quantity of interest such as the variance may be computed analytically or numerically. Otherwise, such quantities must be estimated using a random sample $x_1, x_2, \ldots, x_n$. The relative merits of estimating $F(x)$ versus $f(x)$ are an important part of this article.

Density estimation takes two distinct forms—parametric and nonparametric—depending on prior knowledge of the parametric form of the density. If the parametric form $f(x|\vec{\theta})$ is known up to the p parameters $\vec{\theta} = (\theta_1, \ldots, \theta_p)$, then the parameters may be estimated efficiently by maximum likelihood or Bayesian algorithms. When little is known of the parametric form, nonparametric techniques are appropriate. The most commonly used estimators are the histogram and kernel methods, as discussed below.

1 Histograms and CDF's

Grouping and counting the relative frequency of data to form a histogram may be taught at the beginning of every statistics course, but the concept seems to have emerged only in 1662 when John Graunt studied the Bills of Mortality and constructed a crude life table of the age of death during those times of plague (Westergaard 1968). Today the histogram and its close relative, the stem-and-leaf plot, provide the most important tools for exploring and representing structure in a set of data. Studying the shape of a well-constructed histogram can verify an assumption of normality, or reveal departures from normality such as skewness, kurtosis, or multiple modes. The recent study of the theoretical properties of histograms has led to concrete methods for computing well-constructed figures as discussed below.

With only a random sample of data, $\{x_1, x_2, \ldots, x_n\}$, the cdf may be estimated by the empirical cumulative distribution function (ecdf), denoted by $F_n(x)$. The value of the ecdf at x is the fraction of the sample with values less than or equal to x, that is, $\{\#x_i \leq x\}$. Now the probability any one of the x_i is less than x is given by the true cdf at x. Hence, on average, the value of $F_n(x)$ is exactly $F(x)$, and $F_n(x)$ is an unbiased estimator of $F(x)$. The inequality $x_i \leq x$ is a Binomial random variable, so that the variance of $F_n(x)$ may be computed as $F(x)(1 - F(x))/n$. In fact, there is no other unbiased estimator with smaller variance.

The graph of the empirical cdf looks like a staircase function. Figure 1 displays the ecdf of the treatment times 33 physicians estimated when presented with a hypothetical patient suffering from extreme migrane headaches (Hebl 2000). The dotted line is the cdf of the fitted normal density $N(31.3, 9.9^2)$, which appears to fit reasonably well. Figure 2 displays a histogram of the same data, which does not appear to be normal after all. Social scientists often are taught to favor the ecdf over the histogram, but in fact aside from computing probabilities, the ecdf is not well-suited for easily understanding data.

The probability density function (pdf), $f(x)$, carries the same information as the cdf. Since the derivative of the empirical cdf $F_n(x)$ is not well-defined, estimators that rely on grouping have been devised. A convenient interval, (a, b), containing the data is chosen, and subdivided into m "bins" of equal width $h = (b - a)/m$. The value h is called the bin width (and is often called the smoothing parameter). A "well-constructed" histogram must specify appropriate values for a, b, h, and m. Given

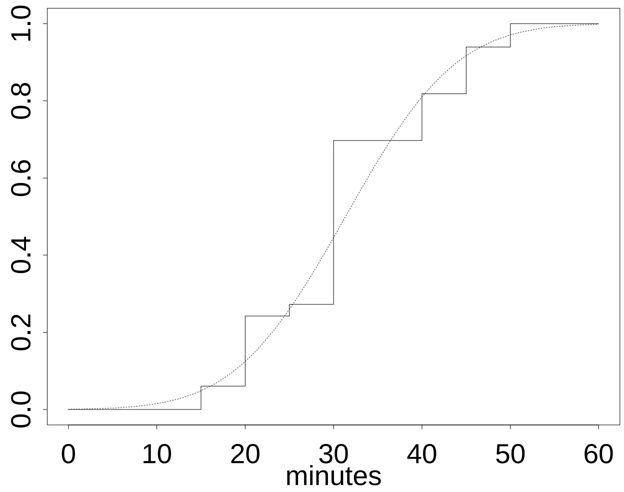

Figure 1: Empirical cumulative distribution function of physicians' estimated treatment times for a patient suffering migrane headaches. The dotted line is the cdf of the normal fit.

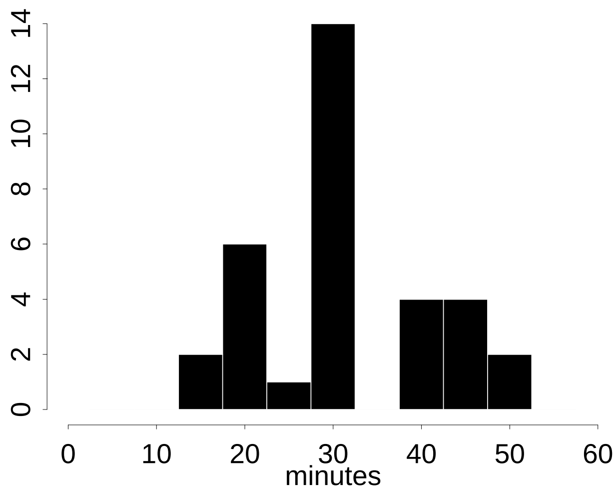

Figure 2: Frequency histogram of treatment times.

such parameters, the number of samples falling into each of the m bins is tallied. Label the m bin counts $\nu_1, \nu_2, \ldots, \nu_m$ from left to right. If the interval (a, b) contains all the samples, then $\sum_{k=1}^{m} \nu_k = n$. A *frequency histogram* is simply a plot of the bin counts, as in Figure 2. Although the values are all multiples of 5 minutes, the data may be treated as continuous. However, the bin width should only be a multiple of 5 and care should be exercised to choose a and h so that no data points fall where two bins adjoin. Here, $a = 2.5$, $b = 57.5$, and $h = 5$. The histogram suggests a departure from normality in the form of 3 modes. Re-examine the ecdf in Figure 1 and try to identify the same structure.

A true probability density is a nonnegative function that integrates (sums) to 1. Hence, the *density histogram* is a simple normalization of the frequency histogram in the k^{th} bin:

$$\hat{f}(x) = \frac{\nu_k}{nh}.$$

The normalization is important when the data are heavily skewed and unequal bin widths are constructed. In such cases, the interval (a, b) is partitioned into m bins of width $h_1, h_2, \ldots, h_m$. The bin counts $\nu_1, \nu_2, \ldots, \nu_m$ should *not* be plotted as they are not relatively proportional to the true density histogram, which is now given by $\hat{f}(x) = \nu_k/(nh_k)$ in the k^{th} bin. Figure 3 displays a density histogram of the distances of the 70 home runs hit by Mark McGwire during his record-setting 1998 American baseball season. A bin width of 20 feet was chosen, and then several adjacent bins were collapsed to make the histogram look unimodal. The long right tail reflects the booming home runs McGwire is known for.

2 Bin Width Construction

Rules for determining the number of bins were first considered by Sturges in 1926. Such rules are often designed for normal data, which occur often in practice. Sturges observed that the m^{th} row of Pascal's triangle, which contains the combinatorial coefficients $\binom{m-1}{0}, \binom{m-1}{1}, \ldots, \binom{m-1}{m-1}$, when plotted, appears as an ideal frequency histogram of normal data. (This is a consequence of the Central Limit Theorem for large m.) The sum of these "bin counts" is 2^{m-1} by the Binomial expansion of

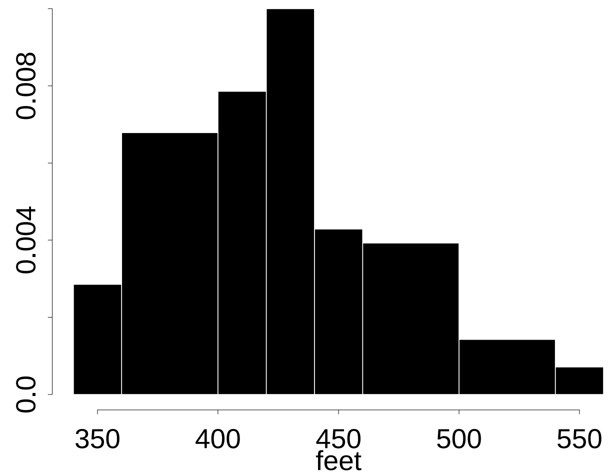

Figure 3: Density histogram of the distances of Mark McGwire's 70 home runs in the 1998 season.

$(1+1)^{m-1}$. Solving the equation $n = 2^{m-1}$ for the number of bins, we obtain Sturges' rule

$$m = 1 + \log_2 n \,.$$

This rule is widely applied in computer software.

The modern study of histograms began in Scott (1979) and Freedman and Diaconis (1981). The optimization problem formulated was to find the bin width h to make the average value of the "distance" $\int [\hat{f}(x) - f(x)]^2 dx$ as small as possible. The solution for the best bin width is

$$h^* = \left[\frac{6}{\int f'(x)^2 dx} \right]^{1/3} n^{-1/3} \,.$$

While the integral of the derivative of the unknown density is required to use this formula, two useful applications are available. First, the *normal reference rule* gives

$$h^* = 3.5 \sigma n^{-1/3} \,,$$

since the value of the integral for a normal density is $(4\sqrt{\pi}\sigma^3)^{-1}$. The second application was given by Terrell and Scott (1985) who searched for the "easiest" smooth density (which excludes the uniform density) and found that the number of bins in a density should always exceed $\sqrt[3]{2n}$ or that the bin width should always be less than

$$h^* \leq 3.73 \sigma n^{-1/3} \,.$$

These choices are called *oversmoothed rules*, as the optimal choices will never be more extreme for other less smooth densities. Notice that the normal reference rule is very close to the upper bound given by the oversmoothed rule. For the McGwire data, the oversmoothed rule gives $h^* \leq 42'$.

More sophisticated algorithms for estimating a good bin width for a particular data set are described in Scott (1992) and Wand (1997). For example, a cross-validation algorithm (Rudemo 1982) that attempts to minimize the error distance directly leads to the following criterion:

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

The bin counts are computed for a selection of bin widths (and bin origins), the criterion computed, and the minimizer chosen, subject to the oversmoothed bound. The criterion tends to be quite noisy and Wand (1997) describes alternative plug-in formulae that can be more stable.

All of these rules give bin widths which shrink at the rate $n^{-1/3}$ and give many more bins than Sturges' rule. In particular, most computer programs give histograms which are *oversmoothed*. The default values should be overridden, especially for large datasets. Similarly, the error of the best histogram decreases to zero at the rate $n^{-2/3}$, which can be improved to $n^{-4/5}$ as described below. However, the histogram remains a powerful and intuitive choice for density estimation and presentation.

3 Advanced Algorithms

Density estimators with smaller errors are continuous functions themselves. The easiest such estimator is the frequency polygon, which connects the midpoints of a histogram. A separate error theory exists. For example, the oversmoothed bin width for a histogram from which the frequency polygon is drawn is given by

$$h^* \leq 2.33\sigma n^{-1/5} .$$

Note that this bin width is wider than the rules for the histogram alone.

Many other estimators exist, based on splines, wavelets, Fourier series, or local likelihood, but all are closely related to the kernel estimator of Rosenblatt (1956) and Parzen (1962), which is described here. The kernel can be any probability density function, but the normal density is a common and recommended choice. Here the smoothing parameter h is the standard deviation of the normal kernel. If the normal density $N(\mu, \sigma^2)$ at a point x is denoted by $\phi(x|\mu, \sigma)$, then the kernel estimator is

$$\hat{f}_K(x) = \frac{1}{n}\sum_{i=1}^{n} \phi(x|x_i, h) .$$

This form shows that the kernel estimator is a mixture of normal densities which are located at the data points. The oversmoothed smoothing parameter value is

$$h^* \leq 1.14\sigma n^{-1/5} .$$

An example is given below.

4 Visualization

The primary uses of density estimation are to display the structure in a set of data and to compare two sets of data. The histogram is adequate for the first task, but overlaying two histograms for comparison is difficult to interpret. The continuity and improved error properties of the kernel estimators excel for this purpose.

In the great home run race of 1998, Sammy Sosa finished second with 66 home runs. Figure 4 displays the empirical cdf's of the distances for the two home run kings (Keating and Scott 1999). This graph clearly shows that McGwire hits longer home runs than Sosa at every distance. In Figure 5, kernel estimates of the McGwire and Sosa data with $h = 10$ are overlaid. The possible structure in these data is reinforced by the similarity of the two densities. Each has two primary modes at 370 and 425 feet, with smaller feature around 475 feet. In addition, McGwire's density clearly shows the right tail of his longest homers. In contrast, the modes in the original histogram of Figure 3 did not appear convincing, and were eliminated by locally adjusting the bin widths. The continuity of the kernel and other estimators makes the assessment of modes much easier.

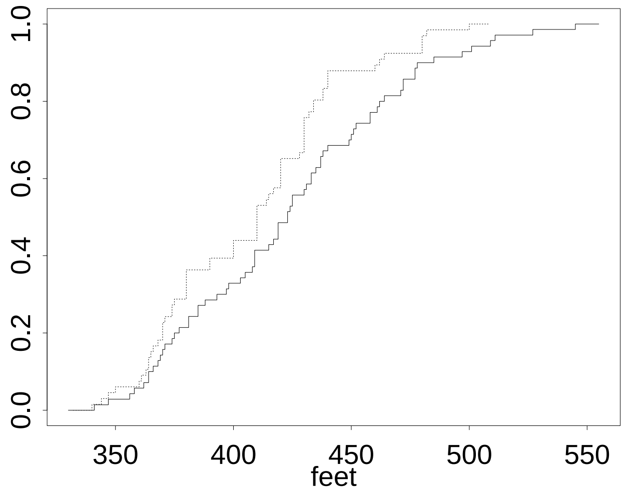

Figure 4: Empirical cumulative distribution functions of the distances of Mark McGwire's 70 (solid line) and Sammy Sosa's 66 (dotted line) home runs in 1998.

5 Multivariate Densities

Histograms can be constructed in the bivariate case and displayed in perspective, but cannot be used to draw contour plots. Kernel methods are well-suited for this purpose. Given a set of data $(x_1, y_1), \ldots, (x_n, y_n)$, the bivariate normal kernel estimator is given by

$$\hat{f}_K(x, y) = \frac{1}{n}\sum_{i=1}^{n} \phi(x|x_i, h)\phi(y|y_i, h) .$$

A different smoothing parameter can be used in each variable.

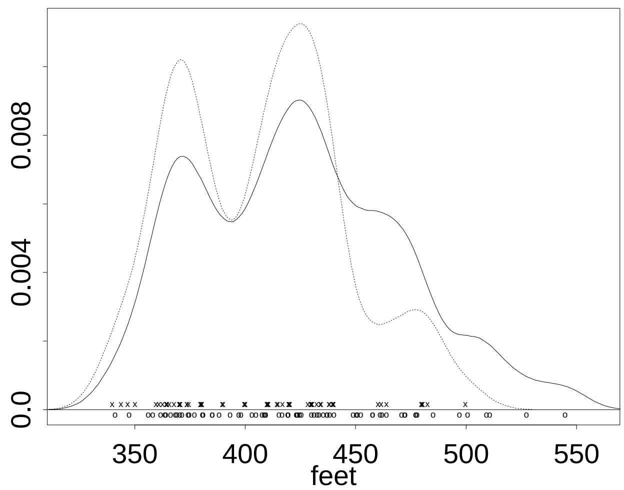

Figure 5: Kernel density estimates of the distances of Mark McGwire's 70 (solid line) and Sammy Sosa's 66 (dotted line) home runs in the 1998 season. McGwire's data are plotted below the horizontal line and Sosa's above.

For three- and four-dimensional data, the kernel estimator can be used to prepare advanced visualization renderings of the data. Scott (1992) provides an extension set of examples. A thorough survey of all nonparametric approaches by top researchers is contained in Schimek (2000).

Bibliography

Freedman D, Diaconis P 1981 On the histogram as a density estimator: L_2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 57:453-476

Hebl M, Xu J 2000 In Rice Virtual Lab in Statistics. http://www.ruf.rice.edu/~lane

Keating J P, Scott D W 1999 A primer on density estimation for the great home run race of '98. *STATS* 25:16-22

Parzen E 1962 On estimation of probability density function and mode. *Ann. Math. Statist.* 33:1065-1076

Rudemo M 1982 Empirical choice of histogram and kernel density estimators. *Scand. J. Stat.* 9:65-78

Rosenblatt M 1956 Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* 27:832-837

Schimek M G (ed.) 2000 *Smoothing and regression.* John Wiley & Sons, New York

Scott D W 1979 On optimal and data-based histograms. *Biometrika* 66:605-610

Scott D W 1992 *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley & Sons, New York

Terrell G R, Scott D W 1985 Oversmoothed nonparametric density estimates. *J. Amer. Statist. Assoc.* 80:209-214

Wand M P 1997 Data-based choice of histogram bin width histogram. *The American Statistician* 51:59-64

Westergaard H 1968 *Contributions to the history of statistics.* Agathon, New York

David W. Scott
Department of Statistics
Rice University
Houston, Texas USA