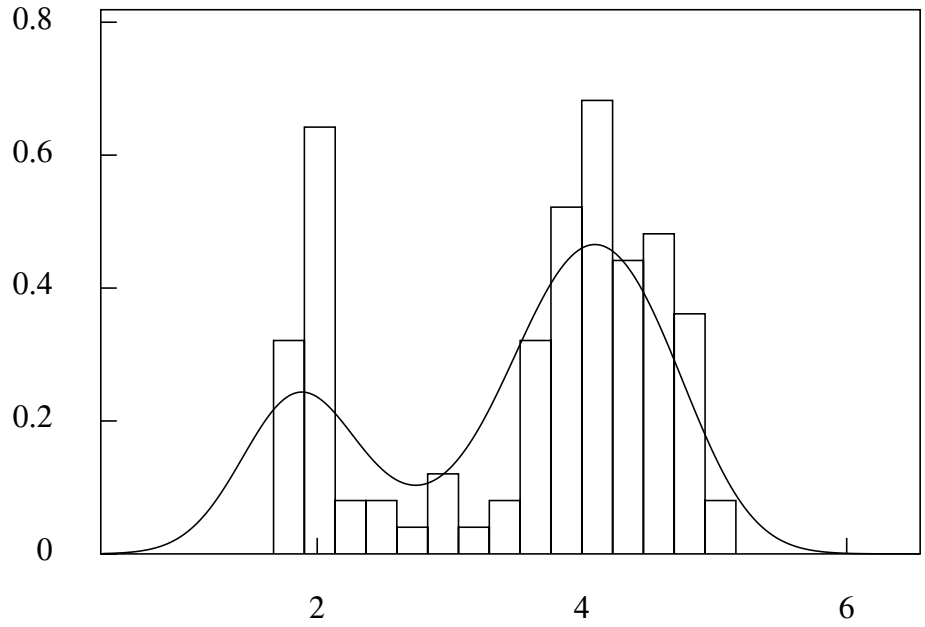

Figure 1. A histogram and an estimated density function based on 107 observations.

Density Function Estimation

The object of density-function estimation is create an accurate representation of a parent distribution on the basis of a random sample $\xi_1, \ldots, \xi_n$. A random sample, taken from a continuous probability distribution, can be represented by a set of points scattered along the real line. The task is to erect a smooth canopy over these points which has the essential characteristics of a density function.

Often, the first step in density-function estimation is to create a discrete version of the random sample. This entails replacing each of the sample points by the nearest point on a grid which is marked out at equal intervals. The grid points can be denoted by $\{x_j = \delta \times j; j = 0 \pm 1, \pm 2, \ldots\}$. A sample point ξ_i will be replaced by the grid point x_j if it falls in the interval $(x_j - 0.5\delta, x_j + 0.5\delta]$, which is the extent of one of a finite number of adjacent bins in which the random sample is collected.

If the number of sample elements which fall into the jth bin is denoted by n_j, then the value x_j will have a relative frequency of $r_j = n_j/n$, where $n = \sum_j n_j$. Since $\sum_j r_j = 1$, the relative frequencies resemble the probability masses of a discrete distribution, and the assemblage of points $\{(x_j, r_j)\}$ gives rise to a so-called bar chart which has a spike erected on each grid point x_j of a height that is proportional to r_j.

Closely related to the bar char is a so-called histogram in which the spike at x_j is replaced by a rectangle of area r_j erected of the interval $(x_j - 0.5\delta, x_j + 0.5\delta]$, which is the domain of the jth bin. If the height of the rectangle is r_j/δ, then the area of the histogram will be unity, as it should be if it is to represent a probability density function. However, the rectangles will combine to give the profile of the histogram a stepped appearance, which is liable to be at variance with the smooth profile of the parent distribution.

To derive a superior representation of the parent density function, we can replace each of the mass points (x_j, r_j) by a kernel function $r_j K(x - x_j, \sigma)$, which is centered on x_j. The kernel function, which also replaces the rectangles of the histogram, commonly takes the form of a probability density function scaled by the relative frequency. The parameter σ, which determines the width of the kernel, is akin to the standard deviation

of probability density function. Indeed, a common choice is to use the normal density function $N(x, \mu = x_j, \sigma^2) = K(x - x_j, \sigma)$ as a kernel.

In effect, the kernel function disperses the mass r_j smoothly over an interval centered on x_j or over the entire real line. The function

$$f(x) = \sum_j r_j K(x - x_j, \sigma),$$

which is the estimate of the probability density function, will inherit some of the characteristics of the constituent kernel functions. In particular, if the kernel functions integrate to unity, then so will the density estimate. Likewise, the continuity properties of the kernel functions will be inherited by the density estimate.

By increasing the width of the kernels via the dispersion parameter σ, the irregularities of the random sample can be smoothed out. However, as the degree of smoothing increases, there is an increasing danger that some genuine features of the parent distribution will be obscured. Although ways are available for determining the degree of smoothing automatically, it is judgment based on prior beliefs that must often be relied upon.

It should be noted that an estimate of the density function with desirable characteristics can be obtained, in principle, by associating a kernel function $n^{-1}K(x - \xi_i, \sigma_i)$ with each of the original data points $\xi_1, \ldots, \xi_n$ which are distributed over the real line. Also, it might be reasonable to vary the dispersion parameter σ_i of the kernels amongst the sample points, by increasing it where the points are sparse. However, as we shall see, the purpose of replacing the data points by points on a grid, and of using a kernels of constant width, is to facilitate some of the computations.

For computational purposes, it is usually appropriate to replace the kernel function by a discrete sequence of weights which represent the ordinates of the function at the grid points. Let $k_{i-j} = K(x_i - x_j, \sigma)$. Then

$$f(x_i) = \sum_j r_j k_{i-j}$$

will be the value of the density estimate at the grid point x_i; and this sum represents a simple convolution of the weighting sequence and the sequence of relative frequencies. Thus, the density function can be estimated via a weighted moving average of the relative frequencies.

There is an efficient method of computing the values $f_i = f(x_i)$ which depends upon the Fourier transform. Let $\{\rho_i\}$ and $\{\kappa_i\}$ be the sequences obtained by applying a discrete Fourier transform to $\{r_i\}$ and $\{k_i\}$ respectively. According to a well-known theory, the Fourier transform of a convolution of two sequences is the product of the transforms of the individual sequences. Thus, if $\{\phi_i\}$ is the transform of $\{f_i\}$, then

$$\{\phi_j\} = \{\rho_j \kappa_j\}.$$

The prescription for finding the values of $\{f_i\}$ is as follows. First transform the sequence of the weights $\{k_i\}$ and the sequence of the relative frequencies $\{r_i\}$. Then form the products $\{\phi_j\}$ of the elements of the transforms. Finally, apply an inverse transform to the product sequence to obtain the sequence of density estimates $\{f_i\}$.

Matters are further simplified if the sequence of the weights has a know Fourier transform, as in the case of a kernel function based on the normal distribution. Then, there is no need to transform the weight sequence, and only two applications of the Fourier transform are required—the direct transformation of the relative frequencies and the inverse transformation of the product sequence which leads to the density values.