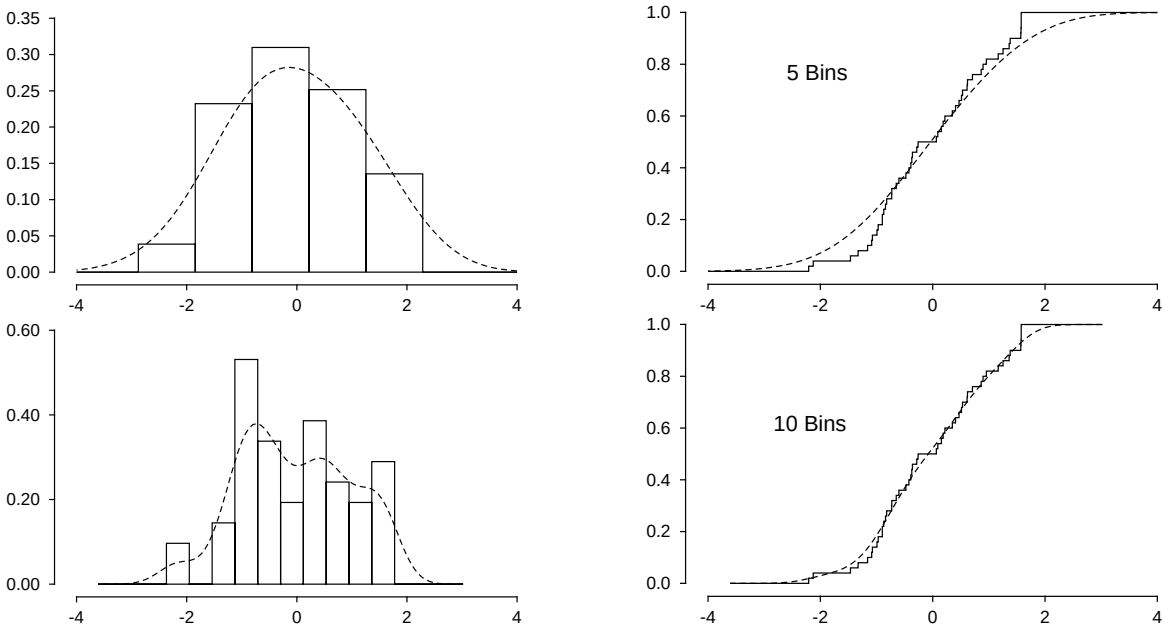
Tutorials and worked examples for simulation, curve fitting, statistical analysis, and plotting.
http://www.simfit.org.uk

Kernel density estimation is a technique used to create a numerical approximation to a density function given a random sample of observations for which there is no known density.

To use this method choose [Statistics] from the main SIMFIT menu then [Statistical calculations] and select [Kernel density estimation].

At this stage it is necessary to input a sample of observations and the following figure illustrates the results when this was done with a data set simulated from a normal distribution with $\mu = 0$ and $\sigma^2 = 1$, using 5 bins for the histogram in the top row of figures, but using 10 bins for the histogram in the bottom row.



The parameters used for the method are adjusted until a satisfactory fit has been obtained when it is then possible to save the best-fit kernel to be used retrospectively as a representation of the data set. However it should be noted that users have to exert considerable control over the parameters chosen as these will greatly affect the kernel estimated. Understanding of the meaning of the parameters selected helps, but naturally a visual display of the fit of the kernel estimate using a reasonable number of bins is recommended.

For instance, in this example changing the number of bins k alters the density estimate since, given a sample of n observations $x_1, x_2, \ldots, x_n$ with $A \le x_i \le B$, the Gaussian kernel density estimate $\hat{f}(x)$ is defined as

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

$$\text{where } K(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)$$

$$\text{and in this case } h = (B - A)/(k - 2).$$

Clearly, a window width h similar to the bin width, as in the top row, can generate an unrealistic over-smoothed density estimate, while using narrower many bins, as in the second row, can lead to over-fitting.

Details are as follows.

- The calculation involves four steps.

  1. From the n data points x_i choose a lower limit a, an upper limit b, and m equally spaced points t_i where
  $$a = A - 3h \leq t_i \leq B + 3h = b,$$
  and m is power of 2. The value of m can be altered interactively from the default value of 128 if necessary for better representation of multi-modal profiles. Data are discretized by binning the x_i at points t_i to generate weights ξ_l.
  2. Compute FFT of the weights, ξ_l to give Y_l.
  3. Compute $\xi_l = Y_l \exp(h^2 s_l^2/2)$ where $s_l = 2\pi l/(b-a)$
  4. Find the inverse FFT of ξ_l to give $\hat{f}(x)$.

- The histograms shown on the left use k bins to contain the sample, and the height of each bin is the fraction of sample values in the bin. The value of k can be changed interactively, and the dotted curves are the density estimates for the m values of t. The program generates additional empty bins for the FFT outside the range set by the data to allow for tails. However, the total area under the histogram is one, and the density estimate integrates to one between $-\infty$ and ∞.

- In addition to the definition of the smoothing parameter h depending on the number of bins chosen for display in the above figure the default setting, which is
  $$h = 1.06\hat{\sigma} n^{-1/5},$$
  uses the sample standard deviation and sample size, as recommended for a normal distribution. Users can also set arbitrary smoothing parameters and, with these two options, the histograms plotted simply illustrate the fit of the kernel density estimate to the data and do not alter the smoothing parameter h.

- The sample cumulative distributions shown on the right have a vertical step of $1/n$ at each sample value, and so they increase stepwise from zero to one. The density estimates are integrated numerically to generate the theoretical cdf functions, which are shown as dashed curves. They will attain an asymptote of one if the number of points m is sufficiently large to allow accurate integration, say ≥ 100.

- The density estimates are unique given the data, h and m, but they will only be meaningful if the sample size is fairly large, say ≥ 50 and preferably much more. Further, the histogram bins will only be representative of the data if they have a reasonable content, say $n/k \geq 10$.

- The histogram, sample distribution, pdf estimate and cdf estimate can be saved to file by selecting the [Advanced] option then creating ASCII text coordinate files.