Simfit

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

Often samples consist of a mixture of distributions. For instance a sample of heights of subjects drawn from a homogeneous population, i.e. of the same age and medical condition, could appear to be be approximately normally distributed but would actually consist of two sub-populations, male and female. In reality, special techniques exist for analyzing certain cases where populations cannot be physically separated into sub-groups but can be resolved into supposed sub-populations using the method of maximum likelihood. However, the curve fitting approach will be discussed in this tutorial because, in principle, it can be used for arbitrary mixtures of any any distributions, not just normal distributions.

For instance, to explain how to use SimFIT program **qnfit** for this purpose, consider the simplest case of a sample arising from a mixture of two normally distributed sub populations, so that a sample partitioned into histogram bins could be approximately modeled by the expression

$$f(x) = \frac{t}{\sigma_1\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left\{\frac{x-\mu_1}{\sigma_1}\right\}^2\right) + \frac{1-t}{\sigma_2\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left\{\frac{x-\mu_2}{\sigma_2}\right\}^2\right)$$

where $0 \le t \le 1$ and the parameters $t, \mu_1, \mu_2, \sigma_1, \sigma_2$ must be estimated by fitting to the histogram bins. The serious problem with this approach is that the shape of the histogram, and therefore the best-fit parameters, will depend on the number of bins chosen. It should therefore be obvious that a very large sample will be necessary, and meaningful parameter estimates can only be expected when μ_1 and μ_2 are widely separate, σ_1 and σ_2 are similar and less than the difference between μ_1 and μ_2, and the partitioning parameter t must obey $t \approx 0.5$. Of course the constraints $\sigma_1 > 0, \sigma_2 > 0$ also must be imposed.

Example 6: Fitting histogram data

The data file `qnfit_data.tf6` can be selected from **qnfit** by clicking on the [Demo] button, and it is listed below after extracting as a table using the [Results] button on the main SimFIT menu.

```
Data file qnfit_data.tf4
 10          3
-3.6    0.0375     1.0
-2.8    0.0625     1.0
-2.0    0.2000     1.0
-1.2    0.2000     1.0
-0.4    0.1000     1.0
 0.4    0.1250     1.0
 1.2    0.1250     1.0
 2.0    0.2500     1.0
 2.8    0.1250     1.0
 3.6    0.0250     1.0
begin{limits}
-5.0       -1.0    0.0
 0.1        0.8    5.0
 0.1        0.4    0.9
 0.0        1.0    5.0
 0.1        1.2    5.0
end{limits}
```

This file was created by reading a mixed sample of 50 $N(-1.5, 1)$ numbers and 50 $N(1.5, 1)$ numbers from program **rannum** into the exhaustive analysis of a vector routine available under [Data exploration] from the

[Statistics] option on the main SimFIT menu. This indicates that the mixed sample is not consistent with a single normal distribution and this step should be taken before fitting any data set because, if the sample is consistent with a single normal distribution, there is little point in trying to fit a sum of two non–identical distributions. This procedure also gives the option of plotting a histogram and then, having chosen the number of bins required, it can create a curve fitting file either unweighted or weighted by the square root of the bin size. Unless a very large sample is under investigation and there are no empty bins an unweighted file should be created. Note in particular that, as the best-fit curve integrates to unity over the data range $(-\infty, \infty)$, the option to normalize the histogram to area 1 must also be chosen.

After reading in the data file `qnfit_data.tf6` the model file `qnfit_model.tf6` should be selected, and this contains the following definition for a sum of two normal distributions.

```
%
Sum of two normal pdfs
A = -(1/2)[(x - p(1))/p(2)]^2
B = -(1/2)[(x - p(4))/p(5)]^2
f(x) = {[1 - p(3)]exp(A)/p(2) + p(3)exp(B)/p(5)}/sqrt{2*pi)
%
1 equation
1 variable
5 parameters
%
begin{expression}
A = -0.5*[(x - p(1))/p(2)]^2
B = -0.5*[(x - p(4))/p(5)]^2
C = [1.0 - p(3)]*exp(a)/p(2) + p(3)*exp(b)/p(5)
f(1) = C/root2pi
end{expression}
%
```

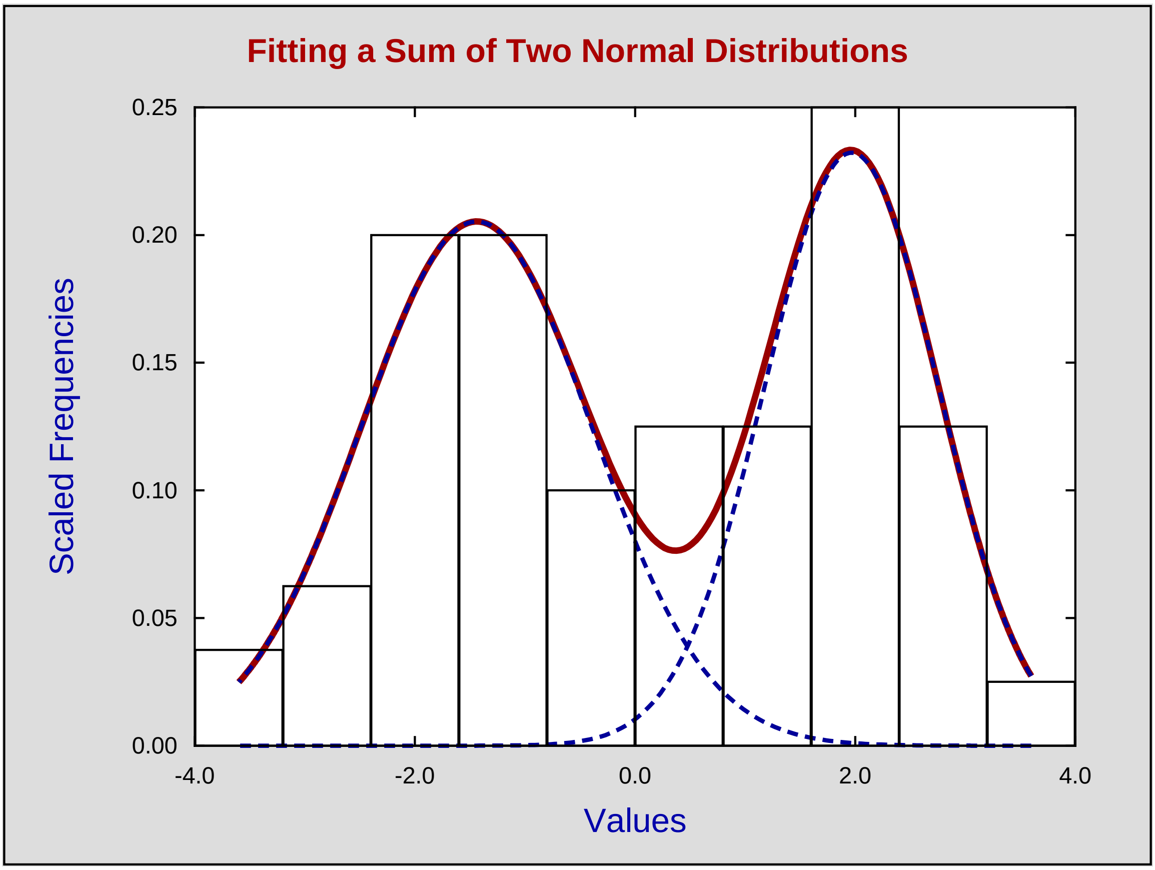
The best fit results table follows.

Number	Low-Limit	High-Limit	Value	Std.Error	Lower95%cl	Upper95%cl	*p*
1	-5.0	0.0	-1.44100	0.172367	-1.88408	-0.99792	0.0004
2	0.1	5.0	1.05066	0.181518	0.58405	1.51726	0.0022
3	0.1	0.9	0.45929	0.061382	0.30150	0.61707	0.0007
4	0.0	5.0	1.96743	0.133634	1.62392	2.31095	0.0000
5	0.1	5.0	0.78877	0.135524	0.44039	1.13714	0.0021

It might be required to plot the best-fit curve superimposed on the sample histogram and the following steps are required to do this.

1. Request a plot in the usual way then choose [Advanced] and transfer to advanced editing.
2. The plot displayed will have symbols for the mid–points of the histogram which need to be changed.
3. From the [Data] options choose to plot bars instead of symbols.
4. The bar type, fill–style, color, and width can be altered if required.

A typical plot resulting from this editing is shown next and clearly shows that, with such dense and well–separated accurate data, a reasonable fit has been achieved. The profile of the two contributing sub–groups was obtained by using the SimFIT library built–in equation instead of the model file and finally requesting graphical deconvolution



Example 7: Fitting a cumulative frequency

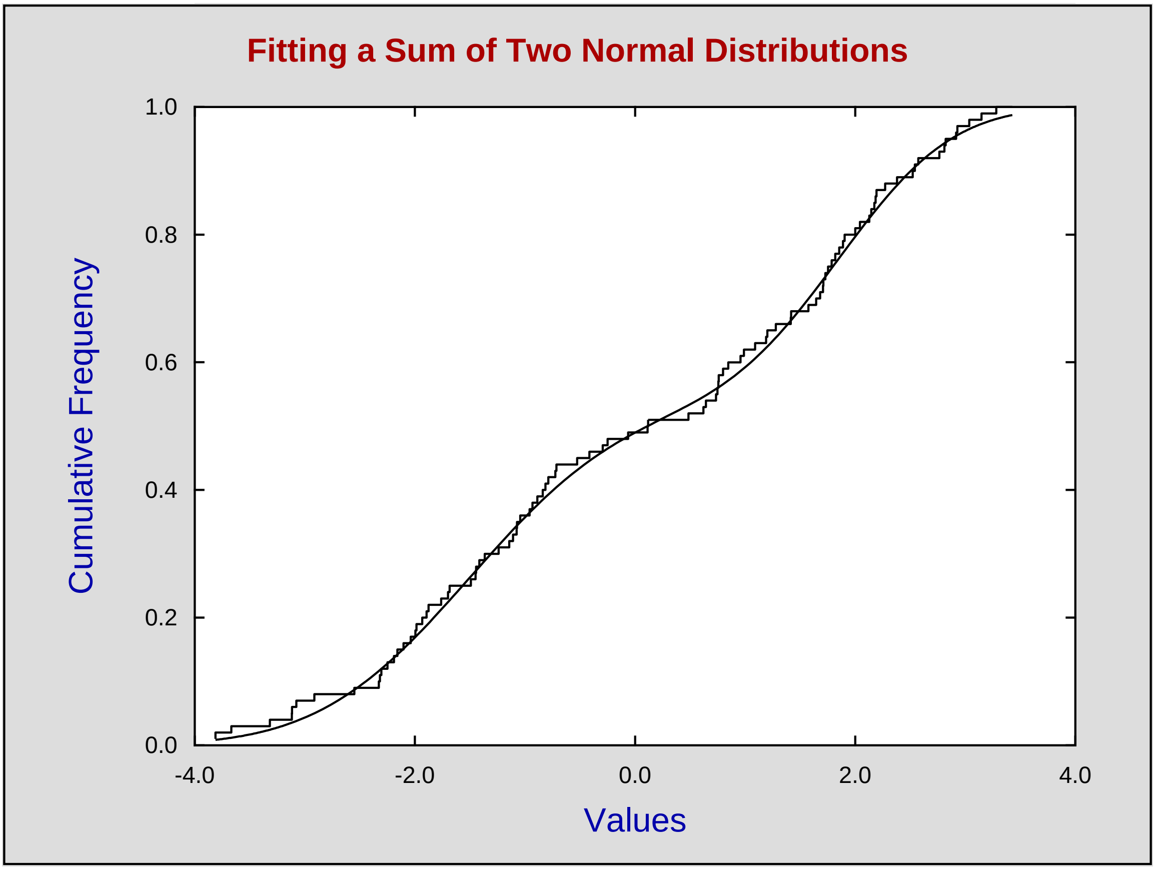
Often data are only available in partitioned form. For instance, counts from channels in flow cytometry are effectively in the form of histogram bins, so the analysis by fitting *pdfs* is all that is possible despite the fact that the results will depend on the number of bins. However, when a sample is available it is possible to fit a sum of two normal *cdfs* as discussed next, and this does not depend on partitioning into bins.

Read test file `normal.tf3` into the exhaustive analysis of a vector procedure exactly as with Example 6 but this time choose to export a *cdf* type curve fitting file. This test file is called `qnfit_data.tf7` and the model file `qnfit_model.tf7` created using SIMFIT **usermod** is as below.

```
%
Sum of two normal distributions
p(3)Phi((x - p(1))/p(2)) + (1 - p(3))Phi((x - p(4))/p(5))
%
1 equation
1 variable
5 parameters
%
begin{expression}
A = p(3)normalcdf((x - p(1))/p(2))
B = (1.0 - p(3))normalcdf((x - p(4))/p(5))
f(1) = A + B
end{expression}
%
```

The table of parameter estimates is displayed next followed by a plot of the data with best–fit curve.

Number	Low-Limit	High-Limit	Value	Std.Error	Lower95%cl	Upper95%cl	p
1	-5.0	0.0	-1.49012	0.034982	-1.55957	-1.42067	0.0000
2	0.1	5.0	1.08482	0.036551	1.01226	1.15738	0.0000
3	0.1	0.9	0.52956	0.010863	0.50799	0.55112	0.0000
4	0.0	5.0	1.85840	0.028793	1.80124	1.91556	0.0000
5	0.1	5.0	0.81238	0.030455	0.75192	0.87284	0.0000



To compare the results from fitting the *pdfs* and *cdfs* we can define the sums of squares SSQ between the parameter estimates $\hat{p}_i$ and the population parameters p_i as

$$SSQ = \sum_{i=1}^{5} (\hat{p}_i - p_i)^2$$

and note that

for the *pdfs*: $SSQ = 0.271$, and $\sqrt{SSQ} = 0.520$, while
for the *cdfs*: $SSQ = 0.171$, and $\sqrt{SSQ} = 0.414$

a slightly better result from fitting the *cdfs*.

In order to succeed in estimating convincing parameter estimates there must be a very large sample with well–separated means, similar variances that do not cause too much overlap, and approximately equally sized sub–groups. Then fitting a *cdf* will give a unique set of parameter estimates as opposed to the way that fitting *pdfs* is dependent on the number of bins, but a visual display of the contribution by sub–groups is perhaps easier judged by superimposing a best fit curve on a histogram.