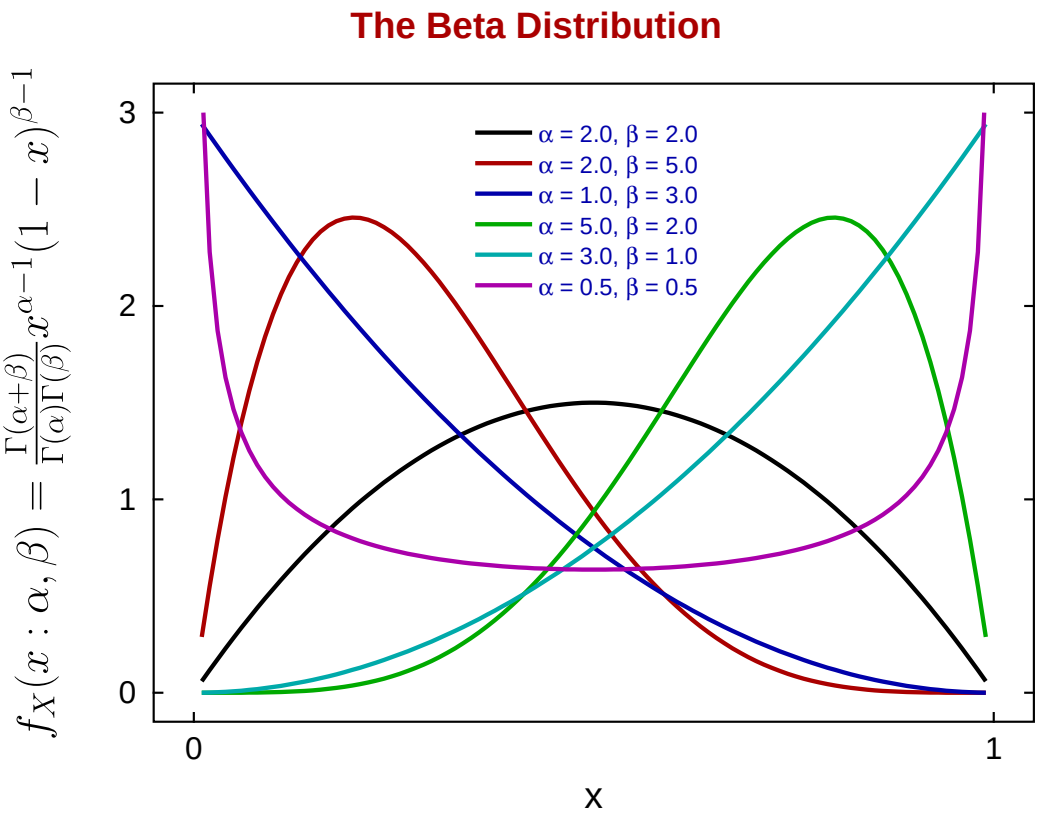
Tutorials and worked examples for simulation, curve fitting, statistical analysis, and plotting.
https://simfit.org.uk

A random variable X $(0 \le x \le 1)$ with the following pdf $f_X(x : \alpha, \beta)$ and cdf $F_X(x : \alpha, \beta)$

$$f_X(x : \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$
$$F_X(x : \alpha, \beta) = \int_0^x f_X(t : \alpha, \beta)\, dt$$
$$= I_x(\alpha, \beta)$$

with parameters $\alpha > 0$ and $\beta > 0$, where $I_x(\alpha, \beta)$ is the regularized incomplete beta distribution, is referred to as a beta random variable. The widespread use of this distribution in data analysis arises not because many experimental observations do actually arise from a beta distribution, but because it is often a convenient unimodal distribution that serves well as an approximation in many situations, such as those involving the estimation of proportions. Some idea of the variation in the profile of a beta distribution as a function of the shape parameters α and β will be clear for the next figure.

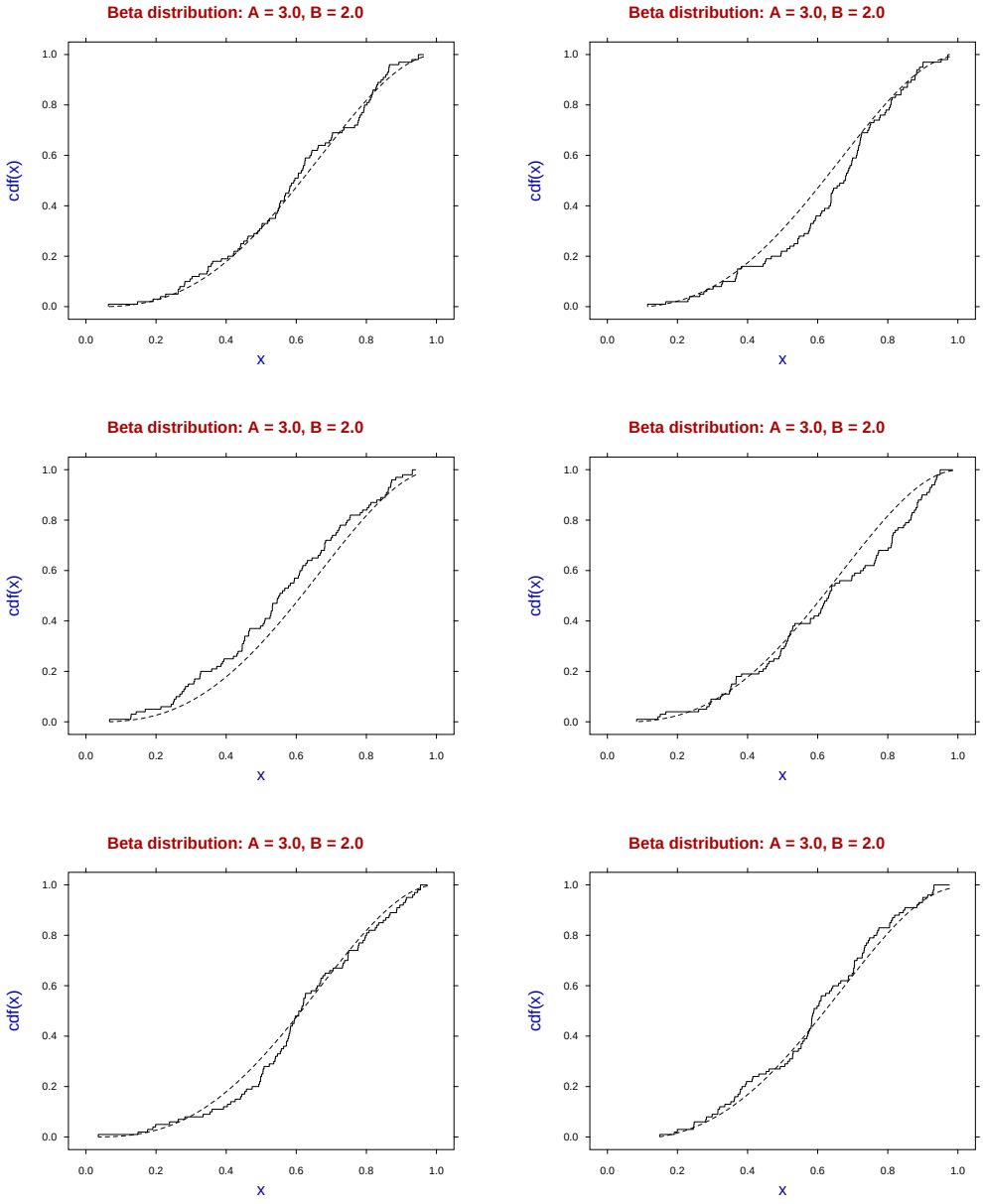


The wide variation in shape that is possible is what makes this a valuable empirical model for fitting arbitrary data that can be projected into the interval (0,1) in order to estimate and visualize skew and kurtosis. In addition the inversion of shape leading to poles at the extremes is often useful in some situations. This document explains how to use SIMFIT to simulate pseudo–random beta variables and fit a beta distribution to observations using constrained weighted least squares in order to estimate goodness of fit.

Generating random samples

When fitting a specified probability distribution to a sample of observations it is valuable first to simulate random samples for the distribution, then observe how the values for random observations change as the parameters vary. Random samples can then be plotted as histograms or cumulative distributions to get a feel as to how well your data can be modeled by the distribution and what are likely to be reasonable parameters.

From the main SIMFIT menu choose [Simulate] then [Generate random numbers and walks] which opens up program **rannum**, and then select to generate sequences of random numbers for the stated distribution and parameters. As an example, consider the following six samples with 100 observations using a beta distribution with $\alpha = 3$ and $\beta = 2$ which demonstrates some rather surprising issues.



It will be seen that, even with a fairly large sample, it is possible to get seemingly large systematic deviations (from a Kolomogorov–Smirnov perspective) between the sample cumulative distribution (solid step curve) and the theoretical distribution (dotted curves) due to the unavoidable pseudo serial correlation in the sample

cumulative. This must be kept in mind when assessing the goodness of fit by subjective graphical inspection of this type.

Of course things are no better with displaying the theoretical PDF overlayed on a histograms, as histogram shape depends on the number of bins chosen. At this point it should be noted that the data exploration option in program **simstat** allows users to examine such PDF and CDF overlays for chosen distributions using any sample of observations.

Parameter estimation for statistical distributions

Before describing methods to estimate parameters for selected statistical distributions, such as the beta distribution from samples of observations, three points should be considered.

1. Experimental observations do not often follow statistical distributions exactly, rather distributions are assumed for convenience. For instance, the distribution of biological variables such as height, weight, blood pressure, etc., in populations are often analyzed as if the data followed a Gaussian distribution, which may appear reasonable in practise but is impossible mathematically, because the Gaussian distribution assumes $-\infty \leq X \leq \infty$.
2. Mathematical statistics is based on such precisely defined variables but everything that is measured experimentally has unavoidable observational error in addition to natural variation.
3. Many methods for parameter estimation depend on sample moments and it is well know that, apart from perhaps the first moment in some situations, higher moments are themselves parameter estimates with large variances, that is, are very innacurate.

For such reasons there is something to be said for estimating parameters by constrained nonlinear regression which offers the possibility of calculating parameter confidence limits and assessing goodness of fit by residuals analysis. That is, arranging a single sample of observations into a form suitable for fitting a statistical distribution as if it were a model for constrained weighted least squares fitting. Usually this means fitting a PDF to a histogram, or a CDF to a sample cumulative. If a very large sample is available, or observations are only available as already partitioned into bins, then fitting a histogram could be considered, as long as it is realized that the result will depend on the number of bins chosen.

Preparing samples of observations for curve fitting

Data must be available as vector, that is, a single column of values with no labels or missing values, and then this is input into the SIMFIT program for exhaustive analysis of a sample. This can be opened from the main SIMFIT menu by choosing [Data exploration]. There are then two options that will prove useful, and both are available using program **rannum**.

1. **Exhaustive analysis of an arbitrary vector**

   This allows you to create a PDF file for fitting by choosing the number of bins required then creating a PDF curve fitting file where the histogram area is scaled to one. Alternatively you can create a CDF curve fitting file. From this procedure you can also calculate the sample moments if these are needed to estimate starting estimates.

2. **Comparing data with a known distribution**

   A distribution is chosen then the parameters are varied until a reasonable fit is apparent when the data are displayed as a PDF–histogram or CDF–cumulative plot. The values chosen can then be used as starting estimates.

Another issue that is often considered is the minimum sample size that is required to begin to justify concluding that a specified distribution with the estimate parameters does reasonably represent the data. Rules of thumb such as . . . *at least ten times the number of parameters* . . . or similar are often suggested which would mean 20 for the beta distribution, but experience indicates a minimum sample size of about 100. So we now turn to a worked example using a beta distribution with a sample size of 100 and $\alpha = 3$ and $\beta = 2$, where the mode is shifted slightly to the right.

Example 1: Fitting a beta pdf

A random sample contained in the file `beta32_data.tf1` was generated by program **rannum** then transformed into a pdf–fitting histogram file with area one by the option to perform exhaustive analysis of a vector, leading to the curve–fitting data file `beta32_pdf.tf1` shown below.

```
beta pdf fitting file generated by RANNUM: A = 3, B = 2
10    3
1.8144613E-01    6.0634121E-01    1
2.6390795E-01    8.4887769E-01    1
3.4636977E-01    7.2760945E-01    1
4.2883159E-01    1.8190236E+00    1
5.1129342E-01    1.8190236E+00    1
5.9375524E-01    1.2126824E+00    1
6.7621706E-01    1.3339507E+00    1
7.5867888E-01    1.3339507E+00    1
8.4114070E-01    1.6977554E+00    1
9.2360252E-01    7.2760945E-01    1
begin{limits}
1    1    5
1    1    5
end{limits}
```

The first column contains the centers of the ten histogram bins, and the second column contains the scaled frequencies, while the third column (with weights equal to one) indicates that unweighted fitting is to be used.

The section starting with the token `begin{limits}` and ending with the token `end{limits}` gives the lower limits, the starting estimates, then the upper limits to be used by program **qnfit** for constrained nonlinear regression in the EXPERT mode, i.e. where such estimates are appended to the data file.

The results from fitting by the SIMFIT quasi–Newton constrained optimization technique are shown next.

Number	Low-Limit	High-Limit	Value	Std.Error	Lower95%cl	Upper95%cl	p
1	1	5	2.26957	0.454304	1.22195	3.31720	0.0011
2	1	5	1.70249	0.301995	1.00609	2.39889	0.0005
3	1	1	1.00000	0.000000	1.00000	1.00000	fixed

For 50,90,95,99% con. lim. using [parameter value +/- t(alpha/2)*std.err.]
t(.25) = 0.706, t(.05) = 1.860, t(.025) = 2.306, t(.005) = 3.355

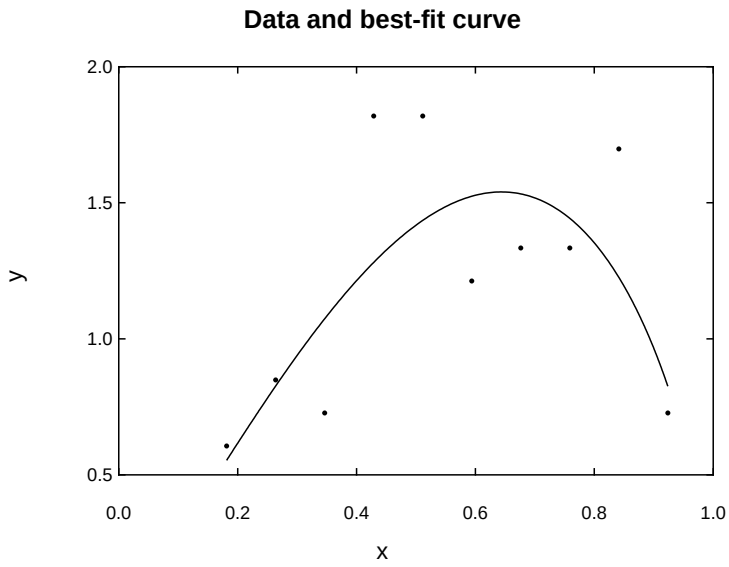
Note that the model used by SIMFIT has the following parameter definitions.

$$p(1) = \alpha$$
$$p(2) = \beta$$
$$p(3) = \Delta$$

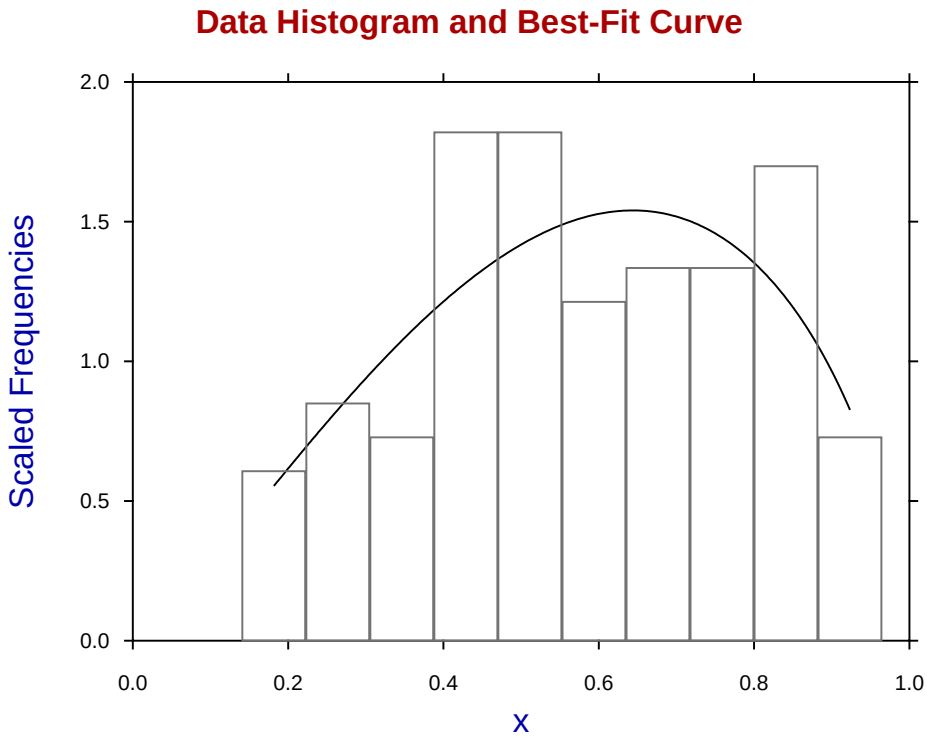
where Δ is a scaling factor than can be used if the area under the histogram is not one. For this fit $p(3)$ was not varied but was fixed, i.e., $\Delta = 1$.

After listing all the goodness of fit results program **qnfit** first shows a default graph where the tops of the histogram bins are shown as dots and the best–fit curve is displayed as a smooth curve ranging between the the centers of the first and last histogram bins.

This default graph from program **qnfit** is shown next.



Here is the default graph after editing to replace the dots by outline type histogram bars width 1.47, and other obvious changes, to give the next graph.



It is possible to create such graphs with many more possible options by saving the best–fit curve parameters, then reading the data into the Data Exploration option of program **simstat** to create the histogram overlayed by the pdf for a beta distribution with the best-fit parameters over the full range, etc.

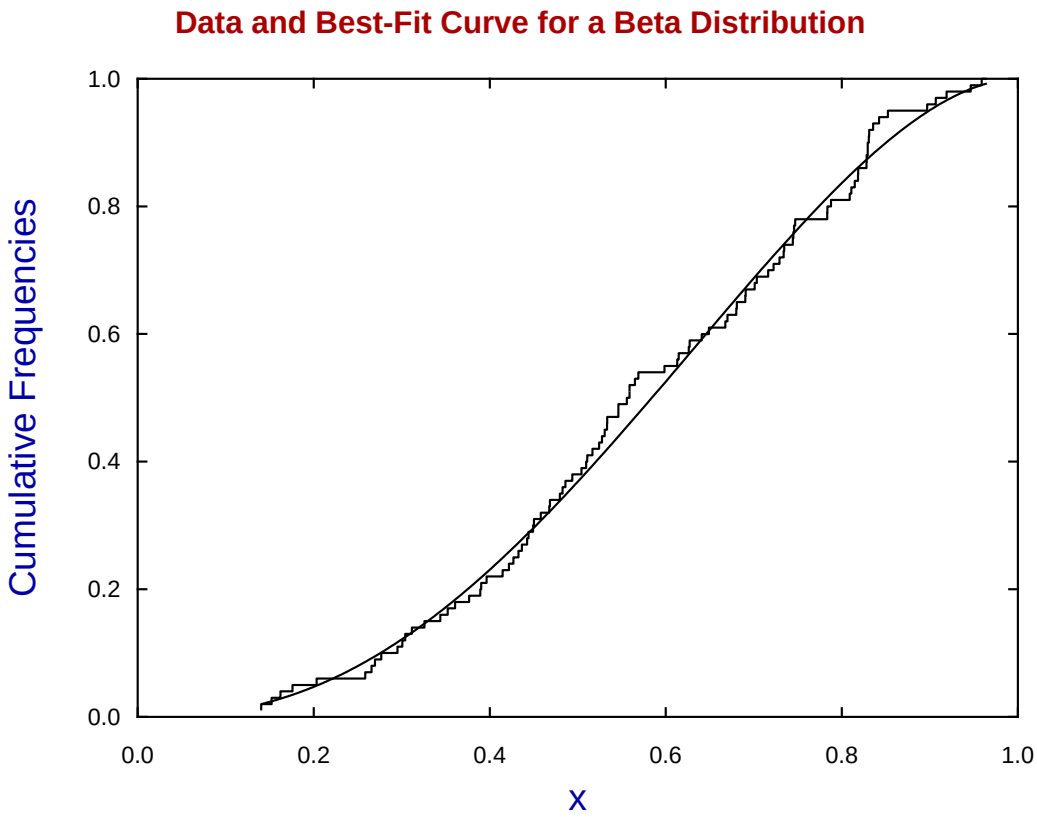
Example 2: Fitting a beta cdf

Proceeding as before then fitting a beta cdf to the data file `beta32_cdf.tf1` using program **qnfit** yields these parameter estimates.

Number	Low-Limit	High-Limit	Value	Std.Error	Lower95%cl	Upper95%cl	p
1	1	5	2.52779E+00	5.57879E-02	2.41708E+00	2.63850E+00	0.0000
2	1	5	1.88851E+00	4.03438E-02	1.80845E+00	1.96857E+00	0.0000
3	1	1	1.00000E+00	0.00000E+00	1.00000E+00	1.00000E+00	fixed

For 50,90,95,99% con. lim. using [parameter value +/- t(alpha/2)*std.err.]
t(.25) = 0.677, t(.05) = 1.661, t(.025) = 1.984, t(.005) = 2.627

The following best-fit curve was edited by simply replacing the default plotting symbols (dots with no lines) for the data by no symbols but a cdf-type step curve.



It is clear that fitting such simple models with just two varied parameters gives well–defined parameter estimates ($p = 0$) but fitting the cdf using all 100 points gives better estimates than fitting to a histogram which only fits ten points.

To quantify this observation, the procedure of data generation by program **rannum** followed by fitting using program **qnfit** was repeated, and the Euclidean distance D between the estimates $(\hat{\alpha}, \hat{\beta})$ and the actual parameter values (α, β) was calculated, where D is defined as follows,

$$D = \sqrt{(\alpha - \hat{\alpha})^2 + (\beta - \hat{\beta})^2}.$$

$\hat{\alpha}$	$\hat{\beta}$	D	Type	Best Fit
2.38518	1.84594	0.633828	pdf: $\alpha = 3, \beta = 2$	
2.52170	2.06080	0.482149	cdf: $\alpha = 3, \beta = 2$	cdf
2.60811	1.65327	0.523259	pdf: $\alpha = 3, \beta = 2$	
2.63019	1.76255	0.439479	cdf: $\alpha = 3, \beta = 2$	cdf
2.26957	1.70249	0.788695	pdf: $\alpha = 3, \beta = 2$	
2.52799	1.88851	0.484998	cdf: $\alpha = 3, \beta = 2$	cdf
1.94617	3.97530	0.059226	pdf: $\alpha = 2, \beta = 4$	
1.85620	3.86457	0.197534	cdf: $\alpha = 2, \beta = 4$	pdf
1.64910	3.79293	0.407442	pdf: $\alpha = 2, \beta = 4$	
1.67319	4.03689	0.328885	cdf: $\alpha = 2, \beta = 4$	cdf
2.45517	4.91705	1.023797	pdf: $\alpha = 2, \beta = 4$	
2.37186	4.82599	0.905836	cdf: $\alpha = 2, \beta = 4$	cdf
2.13494	7.54346	0.476065	pdf: $\alpha = 2, \beta = 8$	
2.12449	8.17192	0.212260	cdf: $\alpha = 2, \beta = 8$	cdf

From this table, where both the pdf and cdf were fitted to the same data set as both histograms (observations pooled into 10 bins) and cumulative frequencies (all 100 observations) for a total of seven separate simulations, a number of tentative conclusions can be drawn.

- The parameters were estimated rather better using the data in cumulative distribution format.
- There is a tendency to underestimate the parameters.

Although not shown, there is an improvement in parameter estimates when the additional normalizing parameter is allowed to vary, more so with histograms of course. However there are other ways to decide which technique to use.

Example 3: Plotting a combined graph

Often beta distributions are plotted simply to estimate the extent to which the mode is skewed away from the central position, and this is most convincingly seen in histograms as long as the number of bins is not too large.

So, as there are only two, or rarely three parameters to be fitted and the beta distribution is robust as an empirical model and easy to fit to a sample of observations, there seems no reason why both should not be fitted at the same time.

For a combined graph from such a fitting procedure there are two considerations.

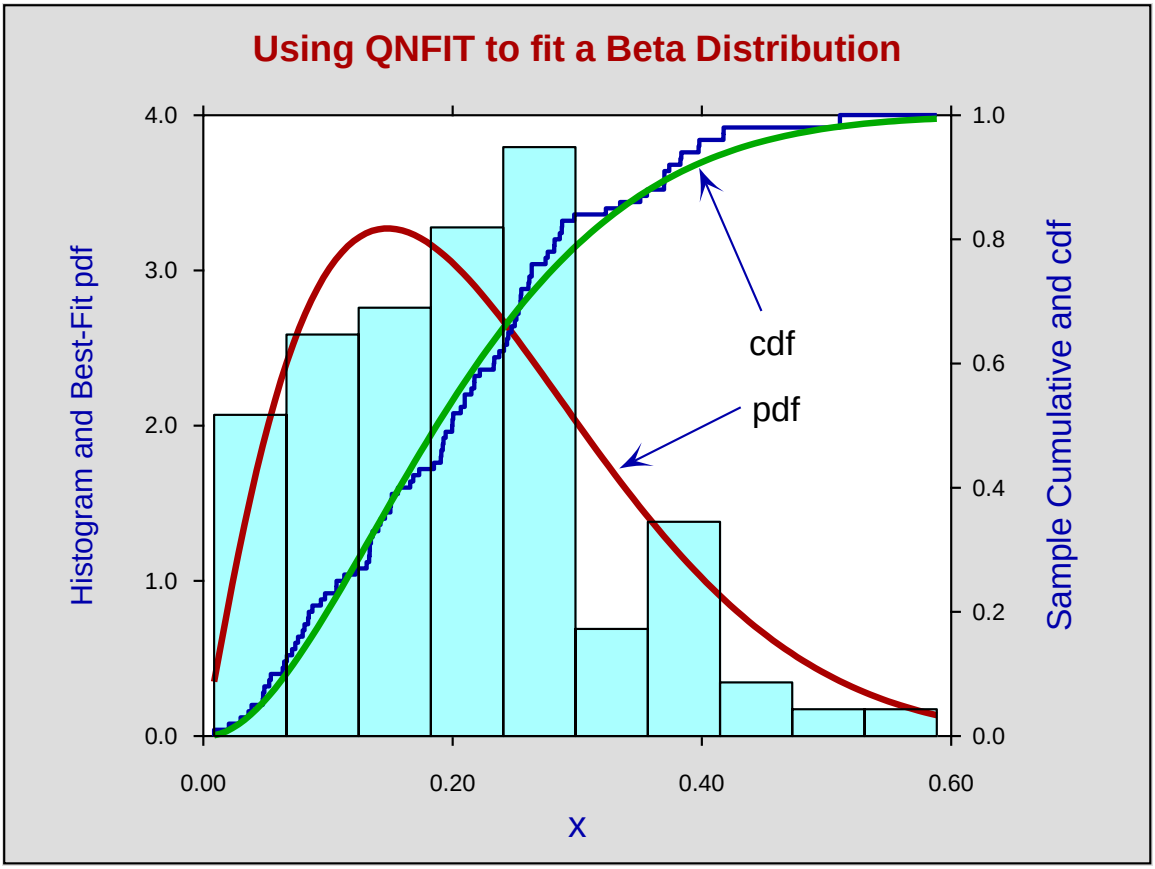
1. Four files of coordinates are required, that is:
   - File 1: coordinates for the histogram;
   - File 2: coordinates for the best–fit pdf;
   - File 3: coordinates for the sample cumulatives; and
   - File 4: coordinates for the best-fit cdf.
2. Two scales are required for the vertical axes, such as:
   - plotting the histogram and pdf using a left–hand scale; and
   - plotting the sample cumulative and cdf using a right–hand scale.

There are several methods by which this process can be done. Perhaps the most obvious is to save the coordinate files from the graphs of data and best fit graphs displayed by program **qnfit**, but this is not necessarily the best way.

Probably the best and easiest SimFIT technique do this, for instance, for a beta distribution like the one from the previous table with $\alpha = 2$, $\beta = 8$, is as follows.

1. Open the option in the SimFIT program **simstat** to compare a sample with an assumed distribution.
2. Read in the data and construct a histogram plotted against a beta distribution with best–fit parameters $\hat{\alpha} = 2.13494$, $\hat{\beta} = 7.54346$.
3. From this save the coordinates to File 1 and File 2.
4. Now construct a cumulative distribution stair-step type plot with added cdf with best–fit parameters $\hat{\alpha} = 2.12449$, $\hat{\beta} = 8.17192$.
5. From this save the coordinates to File 3 and File 4.
6. Open program **simplot** then choose two create a double axis plot and read in File 1 and File 2 to plot against the left–hand Y-axis, then File 3 and File 4 for the right–hand axis.

All that remains is fine tuning to create the following plot.



It should be noted that plotting symbols can be replaced by filled polygons, but if these are filled with color the histogram bin outlines will be lost. This can be overcome by using the same file (File 1) added interactively as an additional file (File 5) used to outline the resulting filled polygons. Alternatively, if this situation is anticipated, an additional copy of File 1 containing the histogram outlines can be added right from the start.

Practical issues

As the shape of the data will be evident before any computation of best–fit parameters, then visual inspection helps in the choice of starting estimates and limits.

Writing the beta probability density function, i.e. the PDF, in the following form

$$f_x(x : \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

emphasizes that, as the complete beta function itself, i.e., $B(\alpha, \beta)$ is a constant and not dependent on x, the graphical behaviour of this density function for $0 \leq x \leq 1$ depends only on the expression

$$x^{\alpha-1}(1-x)^{\beta-1}$$

so there is a single turning point for non-degenerate cases at the mode M where

$$M = \frac{\alpha - 1}{\alpha + \beta - 2}.$$

As $M = 0.5$ when $\alpha = \beta$, while $M > 0.5$ if $\alpha > \beta$, and $M < 0.5$ if $\alpha < \beta$, the displacement of the mode from 0.5 indicates the relative magnitude of α and β. Of course the degenerate case when $\alpha = \beta = 1$ corresponding to a uniform distribution, the complications due to vertical asymptotes when $x = 0$ for $\alpha < 1$ and $x = 1$ for $\beta < 1$, along with the general inversion of shape when $\alpha < 1$ and $\beta < 1$ must be considered. As the general shape would be indicated by the data then this means it is easy to decide on the lower limits of 1 when $\alpha > 1$ and $\beta > 1$ and upper limits of 1 when $\alpha < 1$ and $\beta < 1$.

There are two other practical issues to consider when fitting the beta distribution to observations.

1. **The range of x values**

   In the cases where $\alpha > 1$ and $\beta > 1$ then there is no restriction of range and observations can be anywhere between $x = 0$ and $x = 1$. However, if vertical asymptotes are anticipated, then values must be restricted near potential asymptotes so that computation does not lead to overflow.

2. **The parameter limits**

   As computation of best–fit parameters proceeds then, at every fixed value of x, the values of the internal estimates $\hat{\alpha}$ and $\hat{\beta}$ are perturbed by factors of the order of machine precision. So the upper and lower limits should normally be chosen such that singular cases are avoided.

   Another issue concerns the evaluation of the complete beta function for non–integer arguments. As α and β become larger then the time taken to evaluate the complete beta function increases very rapidly. Of course the computer code doing this is is optimized, but is still faced with such limitations. So it is recommended that the upper limits requested for parameter estimates should be selected conservatively with this in mind to avoid lengthy computations.