A package for simulation,statistics,plotting and curve fitting.
bill.bardsley@simfit.org.uk, University of Manchester,UK.
https://simfit.uk
https://simfit.org.uk
https://simfit.silverfrost.com

Contents

1 Introduction

This document is a simple introduction to the principles that form the basis for all data analysis and is intended as a preparation for using SIMFIT for both simple routine statistics and graph plotting as well as advanced analysis in such areas as nonlinear model fitting, calibration, multivariate methods, smoothing, survival analysis, generalized linear modeling, etc. It presumes that users are not prepared to use data analysis uncritically by using recommended techniques and blindly accepting the results reported, but is intended for those who are not content until they understand the theoretical principles on which the technique they are using are based.

Data analysis involves using a computer program to calculate parameters from a data set in order to summarise the contents of that set, usually to accept or reject the validity of a hypothesis or to compare alternative mathematical models that could be used to explain the data. The SIMFIT package provides a large number of data analysis techniques and this document only deals with the basic knowledge required to interpret data sets intelligently. This is how to use SIMFIT.

1. Select the program to be used from the appropriate drop down menu or the [A/Z] option listing all programs in alphabetical order.

2. Read the option to provide more information about use of the program chosen.

3. When asked to provide data the user–friendly programs provide a selected test data set so you can see what happens with correctly formatted data before inputting your own data. The advanced programs such as **qnfit** , **deqsol** , or **usermod** require more knowledge from users, but all programs offer appropriate selected test data sets by choosing the [Demo] button on the file open dialog.

4. All the results that are calculated are added to a results file that can be accessed retrospectively for saving or copying tables for inclusion into documents.

5. The [Tutorials] button offers a list of worked example, the [Examples] button collects all the tutorials together with an index, while the [Manual] button opens the SIMFIT reference manual which contains mathematical and statistical details for all SIMFIT procedures.

1.1 Types of data

There are two distinct types of data, either integers resulting from a counting process or floating point numbers resulting from measurements, but many data sets will consist of a mixture of types. However, all types will have associated integers such as the number of observations in the set. For instance a data set could consist of of observations of the room temperature measured hourly throughout a day which would consist of 24 values that could be plotted to summarise the variation.

1.2 Representation of numbers by computers

Integers, say n, occur somewhere in a finite set with a minimum and maximum value, e.g.

$$i \ldots < n-2 < n-1 < n < n+1 < n+2 \ldots < j$$

where the numbers are stored exactly within the limits i and j set by the operating system. If two integers are multiplied together to generate a number greater than k there will be a computer error unless the programmer has included code to trap that error. If two numbers are divided in such a way that a non-integer will result then either truncation will occur or a floating point number may result, e.g.

$$6/3 = 2 \quad \text{but} \quad 1/2 = 0 \quad \text{or} \quad 1/2 = 0.5$$

depending on how the program was compiled.

Floating point numbers, say x, contain a decimal point and, although there is an upper and lower limit, there is the problem that between any two non–identical decimal numbers there exist an infinite number between them. So the problem of truncating floating numbers to limit the number of significant digits has to be considered given that numbers have to be represented to the user in a form to be easily understood and incorporated in documents.

To surmount this problem computer programs were designed to output numbers in scientific, i.e, exponential format where the symbol E is used to indicate multiplication by a scaling factor in powers of ten. For instance

```
          123456.123456  =  1.23456123456E+06
0.000000123456123456  =  1.23456123456E-06
```

where orders of magnitude can be seen at a glance and tables would be aligned at the decimal points. Using the convention that the number of significant figures would be the number of non-zero digits in the scientific notation after deleting trailing zeros both of these numbers would have 12 significant figures which is approaching the maximum allowed depending on the operating system and the maximum realistically required to represent experimental data.

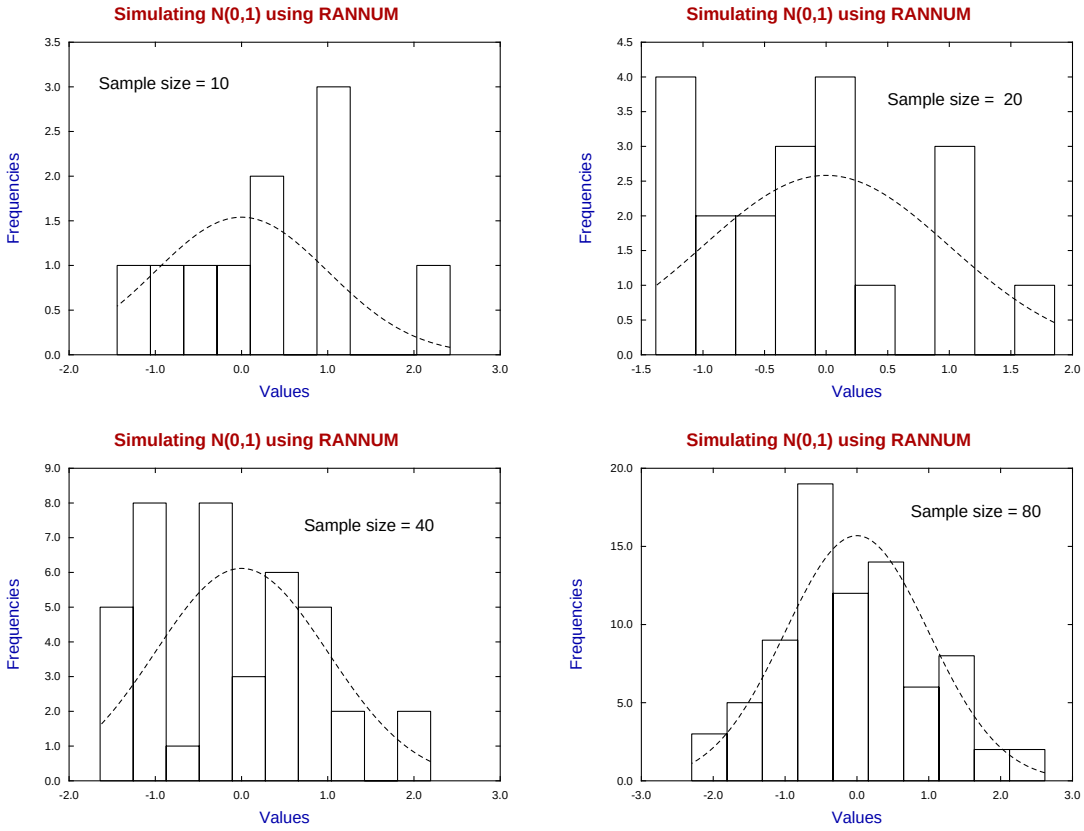
Now calculations on computers would normally be computed in up to 15 significant figures but the parameters calculated should have no more significant figures than the data. So the default in SimFiT is 6 significant digits for numbers in standard notation but swapping over to scientific notation for very large or very small numbers. However users can alter the number of significant digits or choose six significant figures in scientific notation.

2 Visual analysis of data

The visual analysis of data is widely used and the reason for this is the belief that patterns can be discerned from plotting the distribution of relatively small sample sizes. This is especially so with life science studies or where the sample sizes are small due to expense or the inability to obtain sufficient data. The next two examples are intended to warn against relying on conclusions resulting from the visualisation of small samples.

2.1 Simulating a normal distribution

A sample of pseudo–random numbers can be generated using program **rannum** after selecting the distribution, parameters required, and sample size. For instance, selecting [Simulate] from the main SIMFIT menu, then choosing to simulate a normal distribution with $\mu = 0$ and $\sigma^2 = 1$ for sample sizes of 10, 20, 40, and 80 created four vector files. These were then analyzed by program **normal** leading to the histograms shown below.



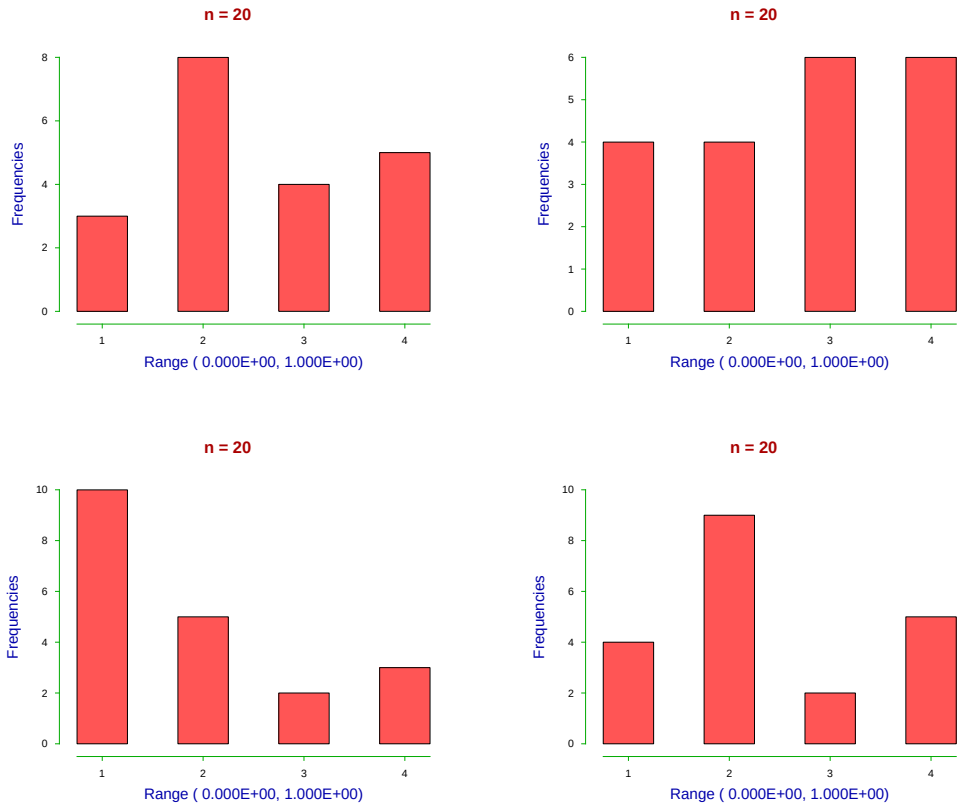
In these histograms we can see that a sample size of ten generates a sample that bears little comparison to the theoretical distribution shown as a dotted curve. Even a sample size of twenty does not look convincingly like the theoretical distribution and a sample size of eighty is required before the histogram does really suggest the theoretical distribution. Of course this is merely a warning not to the use the shape of a histogram to suggest the parent distribution unless the sample size is sufficiently large, and in any case the shape of histograms depend on the number of bins used and are therefore to a degree arbitrary. Of course program **normal** provides many graphical and statistical techniques to determine if a sample can be regarded as coming from a normal distribution.

2.2 Minimum sample size

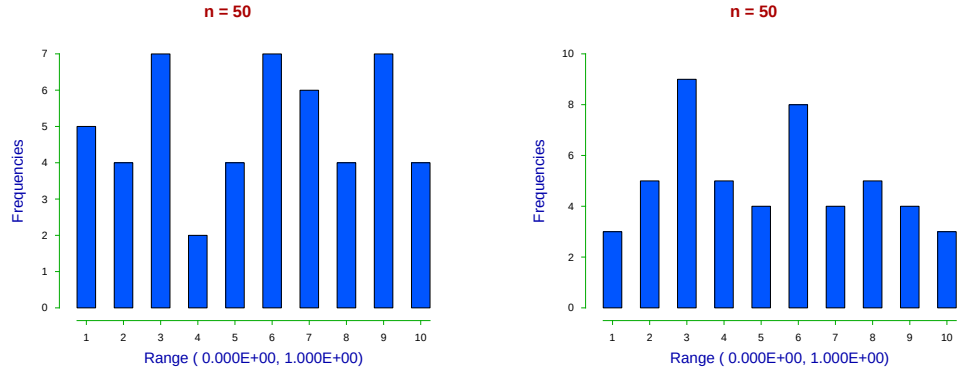
It is often advised that a minimum sample size of $n = 20$ is required to test if a sample is consistent with an assumed distribution and, although statistical tests like the above can be employed, decisions are more often made by visual inspection of a histogram. Now in the limit of very large samples with many bins histograms do converge in shape to the population distribution. However the next examples are intended to demonstrate that, in reality, sample

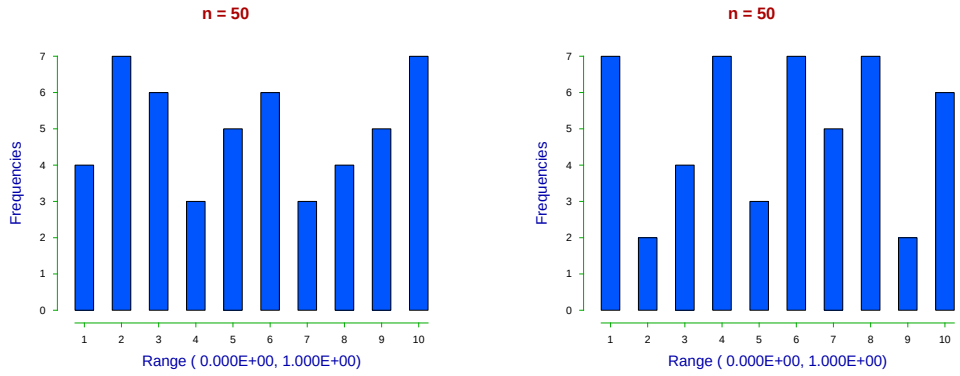
sizes much greater than $n = 20$ are required to carry conviction. The $U(0, 1)$ distribution is particularly suited for this purpose as the histogram should have every bin frequency of approximately the same size, since the probability density function is a horizontal line.

The following histograms display four consecutive simulations using program **rannum** , and note that the usual advice is to have an expected value of at least 5, and preferably an observed value of the same order, for each bin. Of course, a major failing of analysis based on histograms is that the visual appearance and results from statistical analysis depend on the number of bins chosen.

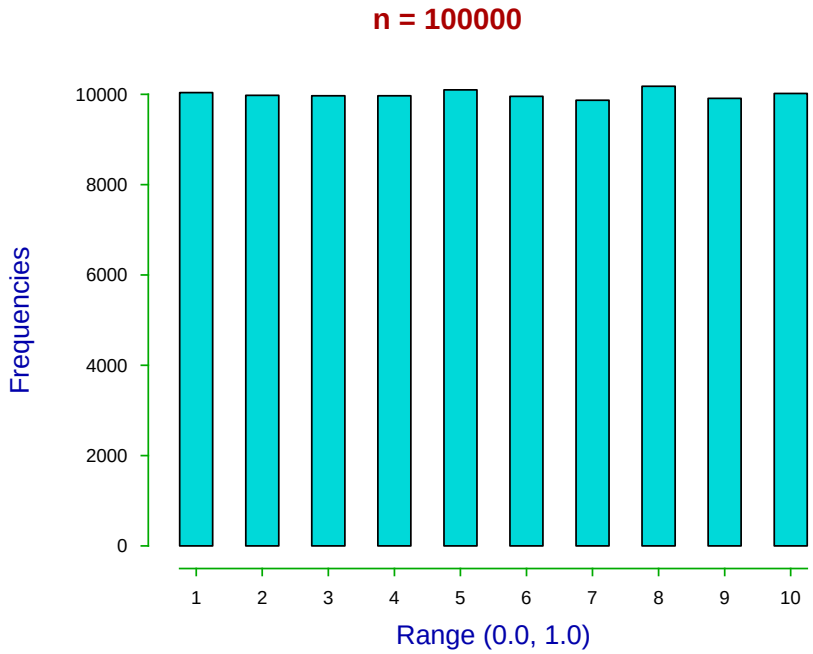


It will be clear from these results that a sample size of $n = 20$ is insufficient and could easily lead to false conclusions, as the histogram can suggest almost any shape for the population distribution. Increasing the sample size to $n = 50$ can still appear to be rather low as will be clear from the next four successive simulations.





Actually numerical results concerning the sample size required can be obtained from the SIMFIT section on power as a function of sample size. Meanwhile here is the sort of convincing result obtained with large samples, in the next case $n = 10000$.



In special cases the power as a function of sample size can be calculated by SIMFIT if the distribution and parameter values is known.

3 Statistical analysis of data

It is often the case that data samples are analysed by statistical methods in order to be less subjective but, in order to understand how to do this, an understanding of probability and statistical theory is required. It is essential to understand that statistical analysis assumes that a sample can be regarded as a set of random numbers from a defined distribution. If an incorrect distribution is assumed then application of statistical analysis is likely to be misleading.

A distribution defines a probability p where $0 \leq p \leq 1$ with an expectation E and variance V and where the sum of all possible p is 1. Examples to illustrate this follow, and a list of all the standard distributions used in data analysis is contained in the SimFIT reference manual `w_manual.pdf`.

3.1 Discrete distributions

A discrete random variable X can have one of n possible values $x_1, x_2, \ldots, x_n$ and has a mass function $f_X \geq 0$ and cumulative distribution function $0 \leq F_X \leq 1$ that define probability, expectation, and variance by

$$\begin{aligned}
P(X = x_j) &= f_X(x_j), \text{ for } j = 1, 2, \ldots, n \\
&= 0 \text{ otherwise} \\
P(X \leq x_j) &= \sum_{i=1}^{j} f_X(x_i) \\
1 &= \sum_{i=1}^{n} f_X(x_i) \\
E(g(X)) &= \sum_{i=1}^{n} g(x_i) f_X(x_i) \\
E(X) &= \sum_{i=1}^{n} x_i f(x_i) \\
V(X) &= \sum_{i=1}^{n} (x_i - E(X))^2 f_X(x_i) \\
&= E(X^2) - E(X)^2.
\end{aligned}$$

3.1.1 Bernoulli distribution

A Bernoulli trial has only two possible outcomes, $X = 1$ or $X = 0$ with probabilities p and $q = 1 - p$.

$$\begin{aligned}
P(X = k) &= p^k q^{1-k} \text{ for } k = 0 \text{ or } k = 1 \\
E(X) &= p \\
V(X) &= pq
\end{aligned}$$

3.1.2 Binomial distribution

This models the case of n independent Bernoulli trials with probability of success (i.e. $x_i = 1$) equal to p and failure (i.e. $x_i = 0$) equal to $q = 1 - p$. The random binomial variable S_n is defined as the sum of the n values of x_i without regard to order, i.e. the number of

successes in n trials.

$$
\begin{aligned}
S_n &= \sum_{i=1}^{n} x_i \\
P(S_n = k) &= \binom{n}{k} p^k (1-p)^{n-k}, \text{ for } k = 0, 1, 2, \ldots, n \\
E(S_n) &= np \\
V(S_n) &= np(1-p)
\end{aligned}
$$

The run test, sign test, analysis of proportions, and many methods for analyzing experiments with only two possible outcomes are based on the binomial distribution.

3.1.3 Example: application of a binomial distribution

In the case where tossing a coin has only two possible outcomes, say heads when $x_i = 1$ and tails when $x_i = 0$ then we can calculate the probability of all possible outcomes in a given number of trials, say $n = 6$ and $p = 0.5$ using program **binomial** .

```
Current binomial parameters: N = 6, p = 0.5
        x      pmf(x)
        0    0.015625
        1    0.093750
        2    0.234375
        3    0.312500
        4    0.234375
        5    0.093750
        6    0.015625
```

From this we see that extreme values of 0 or 6 could not provide support for rejecting $H_0 : p = 0.5$ at the 1% level. Experimenters should perhaps consider this when using sample sizes of say ≤ 6 rats or cultures are often used in pilot biological studies.

3.1.4 The Poisson distribution

This is the limiting form of the binomial distribution for large n and small p but finite $np = \lambda > 0$.

$$
\begin{aligned}
P(X = k) &= \frac{\lambda^k}{k!} \exp(-\lambda), \text{ for } x = 0, 1, 2, \ldots, \\
E(X) &= \lambda \\
V(X) &= \lambda
\end{aligned}
$$

The limiting result, for fixed $np > 0$, that

$$
\lim_{n \to \infty} \binom{n}{k} p^k (1-p)^{n-k} = \frac{(np)^k}{k!} \exp(-np)
$$

can be used to support the hypothesis that counting is a Poisson process, as in the distribution of bacteria in an sample, so that the error is of the order of the square root of the mean. The

Poisson distribution also arises from Poisson processes, like radioactive decay, where the probability of k events which occur at a rate λ per unit time is

$$P(k \text{ events in } (0,t)) = \frac{(\lambda t)^k}{k!} \exp(-\lambda t).$$

The Poisson distribution has the additive property that, given n independent Poisson variables X_i with parameters λ_i, the sum $Y = \sum_{i=1}^{n} X_i$ has a Poisson distribution with parameter $\lambda_y = \sum_{i=1}^{n} \lambda_i$.

3.2 Continuous distributions

A continuous random variable X is defined over some range by a probability density function $f_X \geq 0$ and cumulative distribution function $0 \leq F_X \leq 1$ that define probability, expectation, and variance by

$$\begin{aligned}
F_X(x) &= \int_{-\infty}^{x} f_X(t)\, dt \\
P(A \leq x \leq B) &= F_x(B) - F_X(A) \\
&= \int_{A}^{B} f_X(t)\, dt \\
1 &= \int_{-\infty}^{\infty} f_X(t)\, dt \\
E(g(X)) &= \int_{-\infty}^{\infty} g(t) f_X(t)\, dt \\
E(X) &= \int_{-\infty}^{\infty} t f_X(t)\, dt \\
V(X) &= \int_{-\infty}^{\infty} (t - E(X))^2 f_X(t)\, dt.
\end{aligned}$$

In the context of survival analysis, the random survival time $X \geq 0$, with density $f(x)$, cumulative distribution function $F(x)$, survivor function $S(x)$, hazard function $h(x)$, and integrated hazard function $H(x)$ are defined by

$$\begin{aligned}
S(x) &= 1 - F(x) \\
h(x) &= f(x)/S(x) \\
H(x) &= \int_{0}^{x} h(u)\, du \\
f(x) &= h(x) \exp\{-H(x)\}.
\end{aligned}$$

3.2.1 Uniform distribution

This assumes that every value is equally likely for $A \leq X \leq B$, so that

$$\begin{aligned}
f_X(x) &= 1/(B - A) \\
E(X) &= (A + B)/2 \\
V(X) &= (A + B)^2/12.
\end{aligned}$$

3.2.2 Normal (or Gaussian) distribution

This has mean μ and variance σ^2 as follows.

$$\begin{aligned} f_X(x) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\ E(X) &= \mu \\ V(X) &= \sigma^2 \\ \Phi(z) &= F_X(z) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp(-t^2/2)\, dt. \end{aligned}$$

It is widely used in statistical modeling, e.g., the assumption of normally distributed dosage tolerance leads to a probit regression model for the relationship between the probability of death and dose. There are several important results concerning the normal distribution which are heavily used in hypothesis testing.

3.2.3 Standardised normal variate

The standardised normal variate Z defined as

$$Z = \frac{X - \mu}{\sigma}$$

where X is a normal variate with mean μ and variance σ^2 so that Z has a normal distribution with mean 0 and variance 1 is very important in data analysis, e.g, consider the following derived distributions.

3.2.4 Chi-square distribution

The χ^2 distribution with ν degrees of freedom results from adding together the squares of ν independent Z variables.

$$\begin{aligned} \chi^2(\nu) &= \sum_{i=1}^{\nu} z_i^2 \text{ or, setting } X = \chi^2(\nu), \\ f_X(x) &= \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} \exp(-x/2) \\ E(X) &= \nu \\ V(X) &= 2\nu. \end{aligned}$$

It is the distribution of the sample variance from a normal distribution, and is widely used in goodness of fit testing since, if n frequencies E_i are expected and n frequencies O_i are observed, then

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = \chi^2(\nu).$$

Here the degrees of freedom ν is just $n-1$ minus the number of extra parameters estimated from the data to define the expected frequencies. Cochran's theorem is another result of

considerable importance in several areas of data analysis, e.g., the analysis of variance, and this considers the situation where $Z_1, Z_2, \ldots, Z_n$ are independent standard normal variables that can be written in the form

$$\sum_{i=1}^{n} Z_i^2 = Q_1 + Q_2 + \cdots + Q_k$$

where each Q_i is a sum of squares of linear combinations of the Z_i. If the rank of each Q_i is r_i and

$$n = r_1 + r_2 + \cdots + r_k$$

then the Q_i have independent chi-square distributions, each with r_i degrees of freedom.

3.2.5 F distribution

The F distribution arises when a chi-square variable with ν_1 degrees of freedom (divided by ν_1) is divided by another independent chi-square variable with ν_2 degrees of freedom (divided by ν_2).

$$\begin{aligned}
F(\nu_1, \nu_2) &= \frac{\chi^2(\nu_1)/\nu_1}{\chi^2(\nu_2)/\nu_2} \text{ or, setting } X = F(\nu_1, \nu_2), \\
f_X(x) &= \frac{\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}\Gamma((\nu_1+\nu_2)/2)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} x^{(\nu_1-2)/2}(\nu_1 x + \nu_2)^{-(\nu_1+\nu_2)/2} \\
E(X) &= \nu_2/(\nu_2 - 2) \text{ for } \nu_2 > 2.
\end{aligned}$$

The F distribution is used in the variance ratio tests and analysis of variance, where sums of squares are partitioned into independent chi-square variables whose normalized ratios, as described above, are tested for equality, etc. as variance ratios.

3.2.6 t distribution

The t distribution arises naturally as the distribution of the ratio of a normalized normal variate Z divided by the square root of a chi-square variable χ^2 divided by its degrees of freedom ν.

$$\begin{aligned}
t(\nu) &= \frac{Z}{\sqrt{\chi^2(\nu)/\nu}}, \text{ or setting } X = t(\nu) \\
f_X(x) &= \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2} \\
E(X) &= 0 \\
V(X) &= \nu/(\nu - 2) \text{ for } \nu > 2.
\end{aligned}$$

The use of the t test for testing for equality of means with two normal samples X_1, and X_2 with sizes n_1 and n_2 and the same variance, uses the fact that the sample means are normally

distributed, while the sample variances are chi-square distributed, so that under H_0,

$$\begin{aligned} Z &= \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{1/n_1 + 1/n_2}} \\ U &= \frac{n_1 s_1^2 + n_2 s_2^2}{\sigma^2(n_1 + n_2 - 2)} \\ T &= Z/\sqrt{U} \\ &\sim t(n_1 + n_2 - 2) \\ T^2 &\sim F(1, n_1 + n_2 - 2). \end{aligned}$$

For the case of unequal variances the Welch approximation is used, where the above test statistic T and degrees of freedom ν calculated using a pooled variance estimate, are replaced by

$$\begin{aligned} T &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \\ \nu &= \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}. \end{aligned}$$

The paired t test uses the differences $d_i = x_i - y_i$ between correlated variables X and Y and only assumes that the differences are normally distributed, so that the test statistic for the null hypothesis is

$$\begin{aligned} \bar{d} &= \sum_{i=1}^{n} d_i/n \\ s_d^2 &= \sum_{i=1}^{n} (d_i - \bar{d})^2/(n-1) \\ T &= \bar{d}\,/\sqrt{s_{\bar{d}}^2/n}. \end{aligned}$$

3.2.7 Example 1. Sums of normal variables

Given n independent random variables $X_i \sim N(\mu_i, \sigma_i^2)$, then the linear combination $Y = \sum_{i=1}^{n} a_i X_i$ is normally distributed with parameters $\mu_y = \sum_{i=1}^{n} a_i \mu_i$ and $\sigma_y^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2$.

3.2.8 Example 2. Convergence of a binomial to a normal distribution

If S_n is the sum of n Bernoulli variables that can be 1 with probability p, and 0 with probability $1 - p$, then S_n is binomially distributed and, by the central limit theorem, it is asymptotically normal in the sense that

$$\lim_{n\to\infty} P\left(\frac{S_n - np}{\sqrt{np(1-p)}} \leq z\right) = \Phi(z).$$

The argument that experimental error is the sum of many errors that are equally likely to be positive or negative can be used, along with the above result, to support the view that experimental error is often approximately normally distributed.

3.2.9 Example 3. Distribution of a normal sample mean and variance

If $X \sim N(\mu, \sigma^2)$ and from a sample of size n the sample mean

$$\bar{x} = \sum_{i=1}^{n} x_i/n$$

and the sample variance

$$S^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2/n$$

are calculated, then

(a) $\bar{X} \sim N(\mu, \sigma^2/n)$;
(b) $nS^2/\sigma^2 \sim \chi^2(n-1)$, $E(S^2) = (n-1)\sigma^2/n$, $V(S^2) = 2(n-1)\sigma^4/n^2$; and
(c) $\bar{X}$ and S^2 are stochastically independent.

3.2.10 Example 4. The Gauss central limit theorem

If independent random variables X_i have mean μ and variance σ^2 from some distribution, then the sum $S_n = \sum_{i=1}^{n} X_i$, suitably normalized, is asymptotically normal, that is

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq z\right) = \Phi(z), \text{ or}$$

$$P(X_1 + X_2 + \cdots + X_n \leq y) \approx \Phi\left(\frac{y - n\mu}{\sigma\sqrt{n}}\right).$$

Under appropriate restrictions, even the need for identical distributions can be relaxed.

3.2.11 The use of a normal distribution in statistics

Because the normal distributions includes the possibility $-\infty \leq x \leq \infty$ then no possible sample of observations can actually be normally distributed. For example, blood pressure, height and weight cannot be negative or infinite. The value arises because many observations obey a normal distribution sufficiently closely or can be transformed into new variables that are more approximately normal.

4 Estimation of parameters

Given a sample $X = x_1 + x_2 + \ldots + x_n$ of size n then it is useful to be able to calculate a small subset of derived numbers in order to summarise the properties of the data set. For instance the average $\bar{x}$ or sample mean can always be calculate using

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

or, using the more convenient mathematical notation for a sum,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

As calculated this expression for $\bar{x}$ has the unambiguous interpretation of meaning that the result of adding the n individual numbers x_i together is the same as multiplying the average by n. However, using this value as an estimate for the value of the mean, i.e., the expected value of a probability distribution, i.e. the theoretical mean, requires further assumptions.

Similarly the sample variance S or sample standard deviation i.e. standard error s can always be calculated using

$$S = \frac{1}{n-1}\sum_{i=1}^{n-1}(\bar{x} - x_i)^2$$
$$s = \sqrt{S}$$

and this clearly represents a measure of the spread of the data around the mean value. Note that in this expression for S the factor $1/(n-1)$ is used for an unbiased estimate because one degree of freedom has been used as the calculation involves the previously calculated $\bar{x}$ in case the value is used for inferences about the population variance and population standard deviation.

4.1 Confidence limits for a binomial parameter

For k successes in n trials, the binomial parameter estimate $\hat{p}$ is k/n and three methods are used to calculate confidence limits p_1 and p_2 so that

$$\sum_{x=k}^{n}\binom{n}{x}p_1^x(1-p_1)^{n-x} = \alpha/2,$$
$$\text{and } \sum_{x=0}^{k}\binom{n}{x}p_2^x(1-p_2)^{n-x} = \alpha/2.$$

- If $\max(k, n-k) < 10^6$, the lower tail probabilities of the beta distribution are used as follows

$$p_1 = \beta_{k,n-k+1,\alpha/2},$$
$$\text{and } p_2 = \beta_{k+1,n-k,1-\alpha/2}.$$

- If $\max(k, n-k) \geq 10^6$ and $\min(k, n-k) \leq 1000$, the Poisson approximation with $\lambda = np$ and the chi-square distribution are used, leading to

$$p_1 = \frac{1}{2n}\chi^2_{2k,\alpha/2},$$
$$\text{and } p_2 = \frac{1}{2n}\chi^2_{2k+2,1-\alpha/2}.$$

- If $\max(k, n-k) > 10^6$ and $\min(k, n-k) > 1000$, the normal approximation with mean np and variance $np(1-p)$ is used, along with the lower tail normal deviates

$Z_{1-\alpha/2}$ and $Z_{\alpha/2}$, to obtain approximate confidence limits by solving

$$\frac{k - np_1}{\sqrt{np_1(1-p_1)}} = Z_{1-\alpha/2},$$
$$\text{and } \frac{k - np_2}{\sqrt{np_2(1-p_2)}} = Z_{\alpha/2}.$$

The following very approximate rule-of-thumb can be used in counting situations to get a quick idea of the range of a binomial mean np when n is large and p small by exploiting the fact that the binomial variance variance equals $np(1-p)$ while the Poisson variance equals the mean

$$P(x - 2\sqrt{x} \leq np \leq x + 2\sqrt{x}) \approx 0.95$$

where $0 \leq x \leq n$ is a binomial sample such as the number of successes in n trials.

Example:In a study the number of deaths among pensioners in a six year period were as follows.

	Sample size	Deaths	Probability	95% confidence interval
Non-smokers	1067	117	0.109653	$0.091533 \leq p \leq 0.129957$
Smokers	402	54	0.134328	$0.102548 \leq p \leq 0.171609$

Again, note the noncentral 95% confidence intervals for the probability estimates $\hat{p}$ as summarized below.

Deaths/Subjects	$\hat{p}$	95% Confidence Interval	Group
117/1067	0.1097	0.1097 - 0.0182, 0.1097 + 0.0203	Non-smokers
54/402	0.1343	0.1343 - 0.0318, 0.1343 + 0.0373	Smokers

4.2 Confidence limits for a normal mean and variance

If the sample mean is $\bar{x}$, and the sample variance is s^2, with a sample of size n from a normal distribution having mean μ and variance σ^2, the confidence limits are defined by

$$P(\bar{x} - t_{\alpha/2,n-1}s/\sqrt{n} \leq \mu \leq \bar{x} + t_{\alpha/2,n-1}s/\sqrt{n}) = 1 - \alpha,$$
$$\text{and } P((n-1)s^2/\chi^2_{\alpha/2,n-1} \leq \sigma^2 \leq (n-1)s^2/\chi_{1-\alpha/2,n-1}) = 1 - \alpha$$

where the upper tail probabilities of the t and chi-square distribution are used.

Example: we can use the body temperature of 25 intertidal crabs recorded in ℃ as follows: 24.3, 25.8, 24.6, 26.1, 22.9, 25.1, 27.3, 24.0, 24.5, 23.9, 26.2, 24.3, 24.6, 23.3, 25.5, 28.1, 24.8, 23.5, 26.3, 25.4, 25.5, 23.9, 27.0, 24.8, 22.9, 25.4. The sample mean, variance and standard deviation were $\bar{x} = 25.03$, $s^2 = 1.8$, and $s = 1.3416408$ leading to the following central confidence intervals for the mean and unsymmetrical confidence limits for the variance.

Sample size	Level	Parameter	Estimate	Interval
25	95%	Mean	25.03	$24.4762 \leq \mu \leq 25.5838$
25	99%	Mean	25.03	$24.2795 \leq \mu \leq 25.7805$
25	95%	Variance	1.8	$1.09745 \leq \sigma^2 \leq 3.48355$
25	99%	Variance	1.8	$0.948231 \leq \sigma^2 \leq 4.36971$

Note that the mean and 95% confidence confidence can be written as $25.03(\pm 0.5538)$ but because of the unsymmetrical confidence this type of shorthand cannot be used for the variance. In fact, as just seen with the binomial distribution, the normal distribution is fairly unique in this respect and displaying the confidence limits in this way will usually only be an approximation.

4.3 Hypothesis testing and power

Hypothesis testing is based upon specifying a null hypothesis H_0 then testing to see if a statistic calculated from the data is sufficiently extreme to justify rejecting the null hypothesis. There are two possible errors.

- **Type I error**
  The null hypothesis is rejected when it is true and the probability of this happening is α.

- **Type II error**
  The null hypothesis is accepted when it is false and the probability of this happening is β.

The significance level is α while the power is $1 - \beta$, often expressed as a percentage. The situation can be summarized in the following table.

Decision	H_0 is true	H_0 is false
Reject H_0	Type I error α	Correct $1 - \beta$
Accept H_0	Correct $1 - \alpha$	Type II error β

4.4 Power and sample size

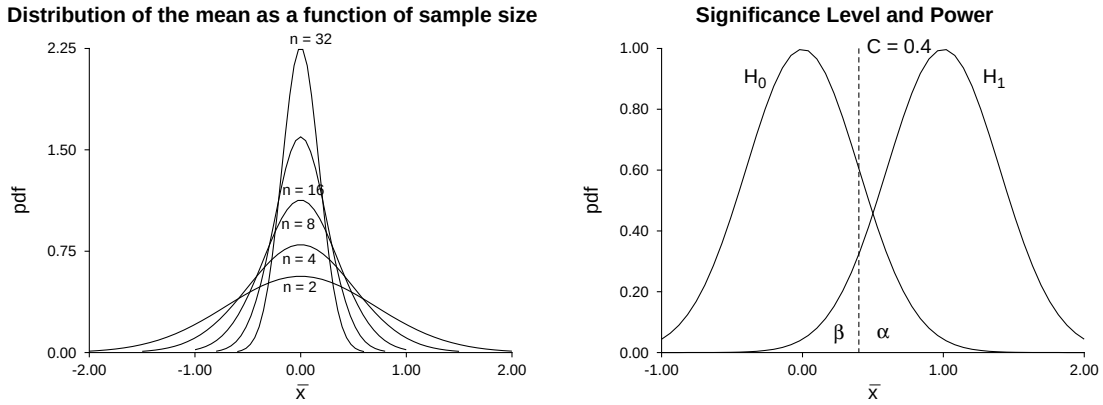
Calculations related to power as a function of sample size can be performed as long as the statistical distributions and parameters required for the null hypothesis are correct and specified. Unfortunately, while calculation of α is straightforward, calculation of β requires that an alternative hypothesis H_1 be specified and can be much more difficult.
If $f(x)$ is the density function for a random variable X, then the null and alternative hypotheses can sometimes be expressed as

$$H_0 : f(x) = f_0(x)$$
$$H_1 : f(x) = f_1(x)$$

while the error sizes, given a critical region C, are

$$\begin{aligned}\alpha &= P_{H_0}(\text{reject } H_0) \text{ (i.e., the Type I error)} \\ &= \int_C f_0(x)\,dx \\ \beta &= P_{H_1}(\text{accept } H_0) \text{ (i.e., the Type II error)} \\ &= 1 - \int_C f_1(x)\,dx.\end{aligned}$$

Usually α is referred to as the significance level, β is the operating characteristic, while $1-\beta$ is the power, frequently expressed as a percentage, i.e., $100(1-\beta)\%$, and these will both alter as the critical region is changed.



This figure illustrates the concepts of signal-to-noise ratio, significance level, and power. The family of curves on the left are the probability density functions for the distribution of the sample mean $\bar{x}$ from a normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$. The curves on the right illustrate the significance level α, and operating characteristic β for the null and alternative hypotheses

$$\begin{aligned}H_0 &: \mu = 0, \sigma^2 = 4 \\ H_1 &: \mu = 1, \sigma^2 = 4\end{aligned}$$

for a test using the sample mean from a sample of size $n = 25$ from a normal distribution, with a critical point $C = 0.4$. The significance level is the area under the curve for H_0 to the right of the critical point, while the operating characteristic is the area under the curve for H_1 to the left of the critical point. Clearly, increasing the critical value C will decrease α and increase β, while increasing the sample size n will decrease both α and β.

5 Regression

The use of regression is based on the assumption that, when a series of n observations y_i are made as a function of some variables $X = x_1, x_2, \ldots, x_j$, then there is a function $F(X, \Theta)$ involving parameters $\Theta = \theta_1, \theta_2, \ldots, \theta_k$ such that there is a relationship

$$y_i = F(X_i, \Theta) + \epsilon_i$$

where ϵ_i is the experimental error or natural variation. Then the residuals r_i defined as

$$r_i = y_i - F(X_i, \Theta)$$

represent the errors and it is required to define a model equation $F(X, \Theta)$ then estimate the best–fit parameters $\hat{\Theta}$.

5.1 Linear and nonlinear models

A linear model is one where the parameters occur linearly. For instance

$$\begin{aligned} y &= mx + c \\ y &= b_1 x_1 + b_2 x_2 + \ldots + b_j x_j \\ y &= a_0 + a_1 x + a_2 x^2 + \ldots a_k x^k \end{aligned}$$

that is, a straight line, a multiliear equation or a polynomial. Such equations are easy to fit. Typical nonlinear models would be

$$\begin{aligned} v &= \frac{V_{max} S}{K_m + S} \\ F(t) &= A e^{-kt} \end{aligned}$$

and such models are difficult to fit so that, before the computer age, they could only be fitted using linearising transformations such as double reciprocal plots or logs.

5.2 Weighting

The idea of weighting is to give more importance to more accurate observations. Where the range of y_i is small it is reasonable to assume that s_i^2 the variance of y_i will be approximately constant and weighting is not required. However, in a sequence of values where there is a large range of y_i values it will be the case that large values will have a larger variance than small values and if this is not taken into account the regression will be dominated by large values leading to biased parameter estimates. It would require a very large sample of replicates to estimate the variances s_i^2 accurately so there are several options.

1. Undertake an independent estimate of the variance of y_i as a function of the magnitude of x_i leading to a predictive formula.

2. If the sample size at each replicate x_i is at least 5 the sample variance at that x_i value could be used as an attempt to smooth out the size differences.

3. Assume constant relative error and use an assumed fraction of the y_i values.

4. Use an assumed fraction of the the best–fit model at each fixed x_i, which does mean that the weights change with each iteration.

Given a sample of n observations y_i with anticipated error $\epsilon_i > 0$ with $k \geq 1$ independent, i.e. relatively error–free, variables $X_i = x_{1i}, x_{2i}, \ldots, x_{k_i}$ and a theoretical model $f(X)$ with m parameters $p_1, p_2, \ldots, p_m$, it is recommended to find the best–fit model parameters $\hat{p}_i$ in order to summarise a data set in terms of a small number of estimated parameters with confidence limits. The method of maximum likelihood suggests that parameter estimates should be chosen to minimise the weighted sum of squared residuals $WSSQ$ given by

$$WSSQ = \sum_{i=1}^{n} w_i (y_i - f(x_{1i}, x_{2i}, \ldots, x_{ni} \mid p_1, p_2, \ldots, p_m)^2$$

where the weights are $w_i = 1/s_i^2$. Actually most meaningful models are nonlinear, weights are seldom known with any accuracy and such minimisation requires non–linear optimisation and well–defined starting parameter estimates which can give non–unique results due to failure to converge to a unique minimum. Not surprisingly, because of such difficulties, this approach is seldom used and simpler but more unrealistic regression is resorted to.

5.3 Linear regression

From the main SIMFIT menu choose the [A/Z] option, open program **linfit** , choose [multilinear regression] using least squares, then browse the default test file `linfit.tf2` which contains the following data set.

x_1	x_2	x_3	x_4	y	s
7.00	26.0	6.00	60.0	78.50	1
1.00	29.0	15.0	52.0	74.30	1
11.0	56.0	8.00	20.0	104.3	1
11.0	31.0	8.00	47.0	87.60	1
7.00	52.0	6.00	33.0	95.90	1
11.0	55.0	9.00	22.0	109.2	1
3.00	71.0	17.0	6.00	102.7	1
1.00	31.0	22.0	44.0	72.50	1
2.00	54.0	18.0	22.0	93.10	1
21.0	47.0	4.00	26.0	115.9	1
1.00	40.0	23.0	34.0	83.80	1
11.0	66.0	9.00	12.0	113.3	1
10.0	68.0	8.00	12.0	109.4	1

Note that the weighting factor s must be supplied as the last column so that SIMFIT knows how many variables are present. It is usual to set all the values of s to one as in the above example, but if accurate estimates for the standard deviations of Y are known these could be used so that weighted least squares fitting can be done.

Analysis of these data then leads to the following table of results using this model

$$y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

where $x_0 = 1$ if β_0 is to be estimated and a constant term is required, or $\beta_0 = 0$ otherwise.

Parameter Estimates

Number of parameters: 5, Rank: 5, Number of points: 13, Degrees of freedom: 8
Residual-SSQ: 47.864, Mallows' C_p: 5.0, R^2: 0.9824

Parameter	Value	Lower95%cl	Upper95%cl	Std. Error	p	
β_0 (Constant)	62.405	-99.179	223.99	70.071	0.3991	***
β_1	1.5511	-0.16634	3.2685	0.74477	0.0708	*
β_2	0.51017	-1.1589	2.1792	0.72379	0.5009	***
β_3	0.10191	-1.6385	1.8423	0.75471	0.8959	***
β_4	-0.14406	-1.7791	1.4910	0.70905	0.8441	***

The stars shown against the parameter estimates in Table 1 are displayed when the parameter estimates are not significantly different from zero, so this table indicates that none of the five parameters were well determined. This will normally be followed by selecting sub–groups of variables until a set is obtained where all parameters are significantly different from zero

5.4 Nonlinear regression

This technique demands a knowledge of the principles of constrained nonlinear optimisation and should only be employed if users appreciate the need for several additional steps if a meaningful analysis is to be achieved.

1. The data should be very accurate and extend over the full range necessary to make sure all parameters are making contributions to the fit.

2. There must be good independent evidence that the model selected is a good choice for the type of data being fitted.

3. Parameter limits must be in order, i.e., lower limit ≤ starting estimate ≤ upper limit.

As an example consider using SIMFIT program **qnfit** to fit a mixture of three normal distributions with scaling factors and a possible background constant to the data set `gauss3.tf1` using the model

$$pdf_n(x) = \frac{p_1}{p_{2n+1}\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left\{\frac{x-p_{n+1}}{p_{2n+1}}\right\}^2\right) + \frac{p_2}{p_{2n+2}\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left\{\frac{x-p_{n+2}}{p_{2n+2}}\right\}^2\right) + \cdots$$
$$+ \frac{p_n}{p_{3n}\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left\{\frac{x-p_{2n}}{p_{3n}}\right\}^2\right) + p_{3n+1}.$$

For a sum of n such sub–models then up to $3n+1$ parameters have to be estimated, namely

- $p_1, p_2, \ldots, p_n$ are positive partitioning fractions, which may be constrained, e.g., $\sum_{i=1}^{n} p_i = 1$
- $p_{n+1}, p_{n+2}, \ldots, p_{2n}$ are means of arbitrary sign,
- $p_{2n+1}, p_{2n+2}, \ldots, p_{3n}$ are positive standard deviations, and
- p_{3n+1} is an arbitrary background correction factor that is sometimes required.

The best–fit parameters are displayed next.

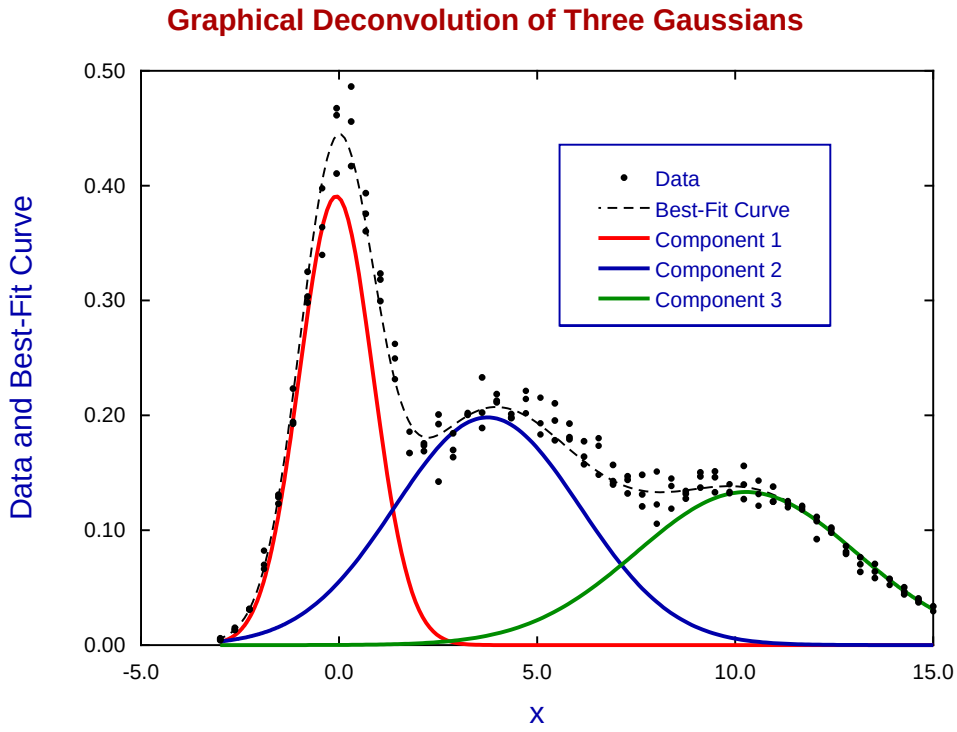
Best-fit parameters for 3 Gaussians using LBFGSB							
Number	Low-Limit	High-Limit	Value	Std.Error	Lower95%cl	Upper95%cl	p
1	0.0	2.0	0.90754	0.021624	0.86479	0.95029	0.0000
2	0.0	2.0	1.16433	0.042173	1.08096	1.24770	0.0000
3	0.0	2.0	0.92519	0.030130	0.86562	0.98475	0.0000
4	-2.0	2.0	-0.07298	0.015572	-0.10376	-0.04219	0.0000
5	2.0	6.0	3.74510	0.050816	3.64464	3.84556	0.0000
6	8.0	12	10.2774	0.096413	10.0868	10.4680	0.0000
7	0.1	2.0	0.92640	0.014331	0.89807	0.95474	0.0000
8	1.0	3.0	2.34330	0.070567	2.20380	2.48281	0.0000
9	2.0	4.0	2.76906	0.062637	2.64523	2.89289	0.0000

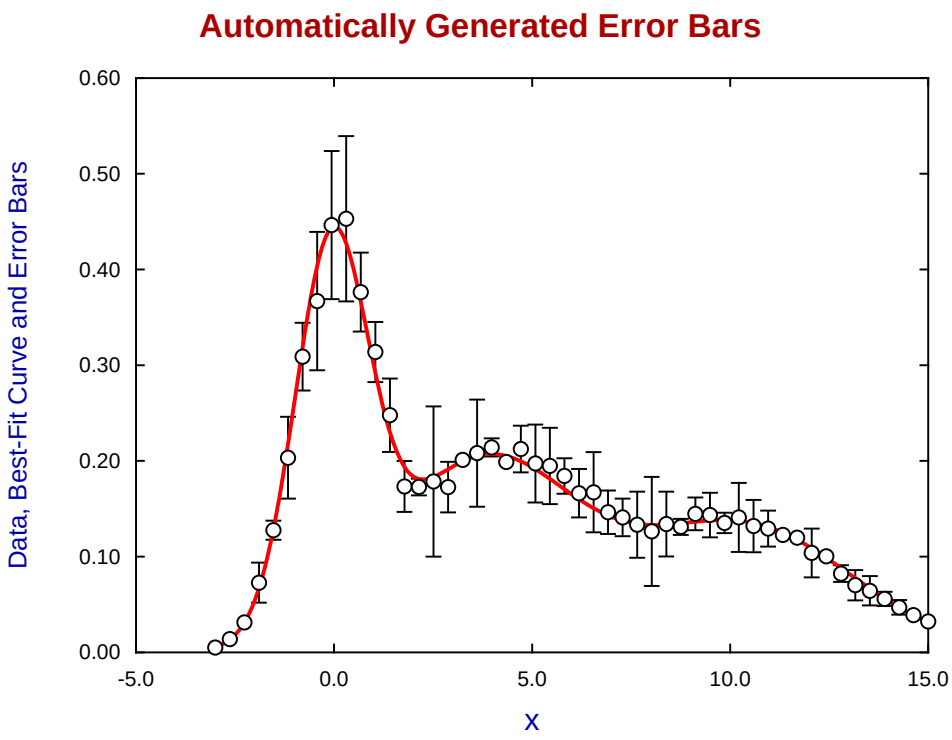
parameter(10) is the excluded constant term

For 50,90,95,99% con. lim. using [parameter value +/- t(α/2)*std.err.]

t(.25) = 0.676, t(.05) = 1.656, t(.025) = 1.977, t(.005) = 2.611

The next graph shows a graphical deconvolution of the fit, that is displaying the three Gaussians along with the overall fit, which is a good way to estimate the contributions of the separate components. This is followed by the replicates replaced by the means together with the calculated error bars which can be done by program **qnfit** .




Automatically Generated Error Bars
Data, Best-Fit Curve and Error Bars
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-5.0
0.0
5.0
10.0
15.0
X