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

1 Statistical distributions in data analysis

Data analysis will usually consist of assuming a statistical distribution and comparing a sample, or a test statistic derived from it, to possible values from the assumed distribution. If the test statistic proves to have a rather extreme value when referred to the assumed distribution it may be taken to suggest that the assumed distribution may not be correct. So statistical testing will often consist of a null hypothesis, denoted as H_0, and there may be an alternative hypothesis or several alternative hypotheses, say H_A.

The situation can be summarized by the following sequence.

1. Collect data.
   An example could be a sample of sizes, times, weights, distances, etc.

2. Calculate a test statistic.
   An example could be calculating the sample mean or standard distribution.

3. Assume a theoretical null distribution, denoted by H_0.
   An example H_0 could be assuming a normal distribution with mean of 6 and standard deviation of 4.

4. Assume a possible alternative distribution, denoted by H_A.
   For instance H_A might be a normal distribution with a mean of 7 and a standard deviation of 4.

5. Check if the test statistics would be extreme if coming from the assumed distribution.
   For instance, to do this we could see if the sample estimates for mean and standard deviation are more consistent with H_0 rather than H_A. This would lead to one of two possible courses of action.

   - Consider the possibility that H_0 is likely to be correct.
   - If no satisfactory conclusion can be reached then accumulate more data or assume a new distribution, or the same distribution with different parameters.

Obviously, if the assumed distribution is incorrect, any conclusions drawn from this procedure will be of questionable value. Now almost no scientific experiment ever leads to data that follows a known distribution exactly, so what happens in practice is that a number of standard distributions are chosen in the hope that one of these will be sufficiently close to the distribution of the test statistic, or that the data can be transformed into an alternative form that is closer to an assumed distribution.

In actuality, only a limited number of standard distributions, such as the following, are encountered in data analysis.

a) Normal

b) t

c) chi-square

d) F

e) Binomial

f) Poisson

g) Uniform

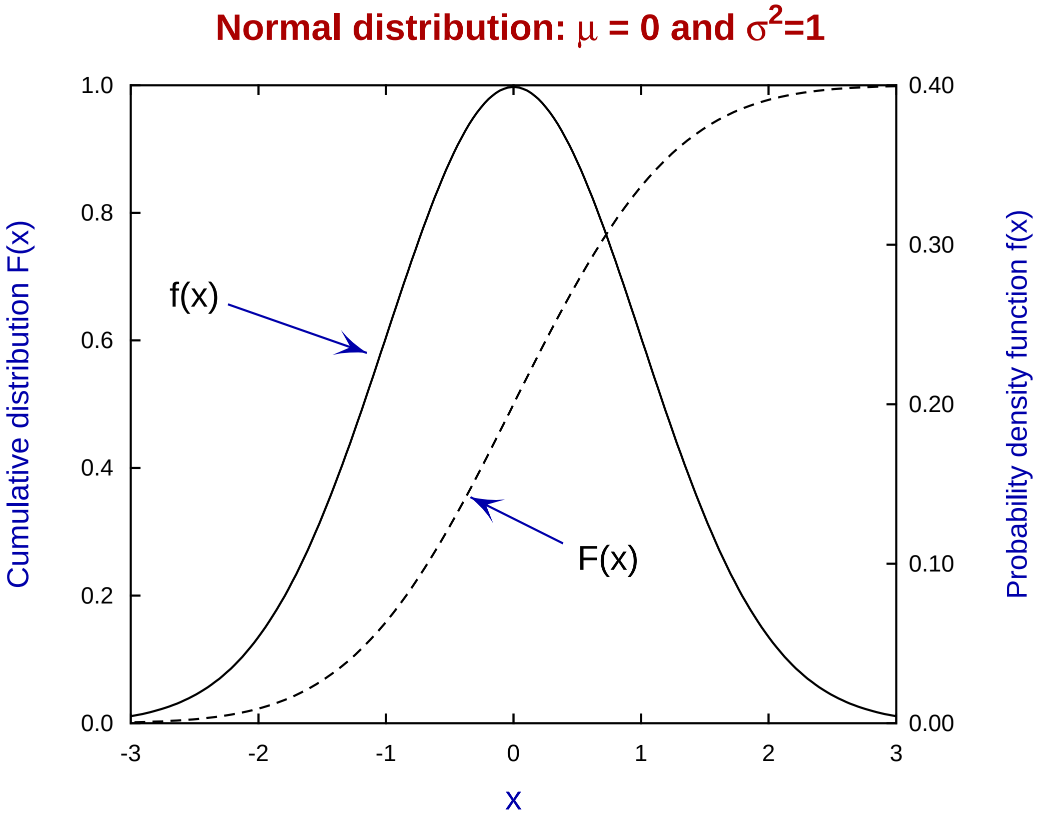
Even so-called nonparametric tests often finish up by relying on some standard distribution, and frequent use is made in statistical theory of the Gauss central limit theorem. This shows that sums of suitably normalized values will tend, in the limit of large sample size n, to a normal distribution. However n may often be very large before such convergence is achieved. Because of all this uncertainty it is often stated that statistical analysis can prove nothing, or alternatively anything. Nevertheless this is all we have so it is useful to sum up some unifying concepts that will be assumed in subsequent SIMFIT tutorials.

2 Continuous variables

A continuous random variable X is a number that can take all values in a range, say $-\infty \le X \le \infty$ but is subject to certain constraints. Typical continuous variables would be time, size, blood pressure, etc., which like so many measured variables happen to be necessarily non-negative. In particular, there will be a non-negative probability distribution function $f(x) \ge 0$ and a cumulative distribution function $F(x)$ such that that the probability that X has a particular value x in the range $A \le X \le B$ will be

$$\begin{aligned} P(A \le X \le B) &= \int_A^B f(t)dt \\ &= F(B) - F(A). \end{aligned}$$

Here, for example are $f(x)$ and $F(x)$ for a normal distribution with mean zero and variance one.



Evidently values of X less than -3 or greater than 3 would be very unlikely for this distribution and could indicate a mean differing from zero and/or a variance differing from 1. A statistical test using the sample mean and sample variance could be constructed by such reasoning.

Note also that, because the variable is continuous, it makes no sense to assign a probability of the random variable having a definite value, but only the probability of it taking a value in an interval $A \leq X \leq B$. However, the integration of $f(x)$ over the possible range, say $-\infty \leq X \leq \infty$ would be one, i.e.

$$\int_{-\infty}^{\infty} f(t)dt = 1.$$

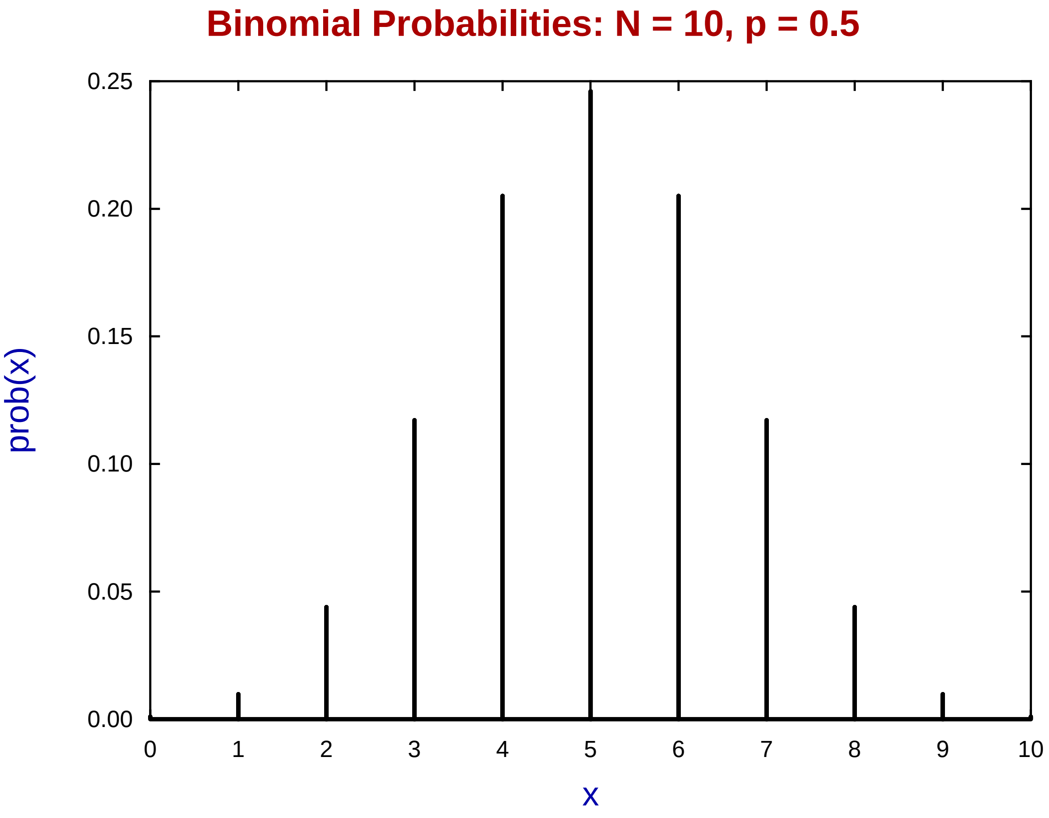
3 Discrete variables

A discrete random variable X is an integer that can only take a limited number of values. Examples would be the number of heads resulting from a fixed number of coin tossings, or the number of eggs hatching as males from a clutch of eggs.

In particular, there will be a non-negative probability mass function $p(x) \geq 0$ which would describe the probability of X having a particular integer value, that is $P(X = k) = p(k)$. Obviously, if there are n possible values that X can have, say $k_1, k_2, \ldots k_n$ then

$$\sum_{i=1}^{n} p(k_i) = 1.$$

Here, for example is the plot of probabilities for a binomial distribution with $N = 10$ and $p = 0.5$ such as would result, for instance, by adding up the number of times a head would occur in ten throws of a coin.



Evidently numbers of heads of 0, 1, 9, or 10 would be very unlikely for this distribution and could be taken to indicate a biased coin. A statistical test using the sample mean and sample variance could be constructed by such reasoning.