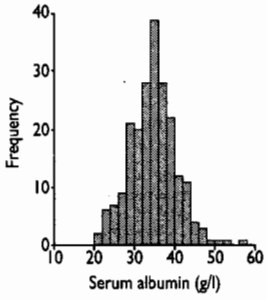
The normal distribution

Douglas G Altman, J Martin Bland

...s is the 11th in a series of occasional notes on medical statistics

When we measure a quantity in a large number of individuals we call the pattern of values obtained a distribution. For example, figure 1 shows the distribution of serum albumin concentration in a sample of adults displayed as a histogram. This is an empirical distribution. There are also theoretical distributions, of which the best known is the normal distribution (sometimes called the Gaussian distribution), which is shown in figure 2. Although widely referred to in statistics, the normal distribution remains a mysterious concept to many. Here we try to explain what it is and why it is important.



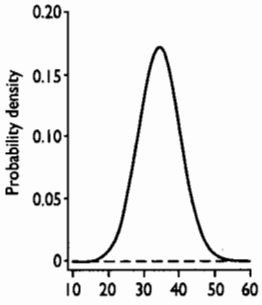


FIG 1 (left)—*Serum albumin values in 248 adults*
FIG 2 (right)—*Normal distribution with the same mean and standard deviation as the serum albumin values*

In this context the name "normal" causes much confusion. In statistics it is just a name; statisticians often use a capital N to emphasise this and to clarify that Normality does not necessarily imply normality. Indeed, in some medical specialties normal distributions are rare.

Various methods of analysis make assumptions about normality, including correlation, regression, *t* tests, and analysis of variance. It is not in fact necessary for the distribution of the observed data to be normal, but rather the sample values should be compatible with the population (which they represent) having a normal distribution. Indeed, samples from a population in which the true distribution is normal will not necessarily look normal themselves, especially if the sample is small. Figure 3 shows the distributions of samples of different sizes drawn at random from normal distributions—few of the small samples look like a normal distribution, but the similarity increases as the sample size increases.

Although some statistical methods, such as the *t* tests, are not sensitive to moderate departures from normality, it is generally preferable not to rely on this feature. Visual inspection of the distribution may suggest whether the assumption of normality is reasonable but, as figure 3 suggests, this approach is unreliable. Significance tests and normal plots can be used to assess formally whether sample data are a plausible sample from a normal population.[1] When data do not have a normal distribution we can either transform the data (for example, by taking logarithms) or use a method that does not require the data to be normally distributed. We consider these topics in future notes.

The normal distribution has another essential place in statistics. Just as separate samples selected at random from the same population will differ (fig 3), so will calculated statistics such as the mean blood pressure. We can think of the means from many samples as themselves also having a distribution. A key theoretical result, called the central limit theorem, underpins many methods of analysis. It states that the means of random samples from any distribution will themselves have a normal distribution. As a consequence, when we have samples of hundreds of observations we can often ignore the distribution of the data. Nevertheless, because most clinical studies are of a modest size, it is usually advisable to transform non-normal data, especially when they have a skewed distribution.

We can consider binary attributes in the same way. For example, the proportions of individuals with asthma will vary from sample to sample. If having asthma is represented by the value 1 and not having asthma by the value 0 then the mean of these values in the sample is the proportion of individuals with asthma. Thus a proportion is also a mean and will follow a normal distribution. These methods are not valid in small samples—some "exact" methods can be used.[2] Similar comments apply to some other statistics, such as regression coefficients or standardised mortality ratios, but for mortality ratios the sample size may have to be very large indeed.

One of the most important applications of these results is in calculating confidence intervals. The general method is based on the idea that the statistic of interest (such as the difference between two means or proportions) would have a normal distribution in repeated samples.[3]

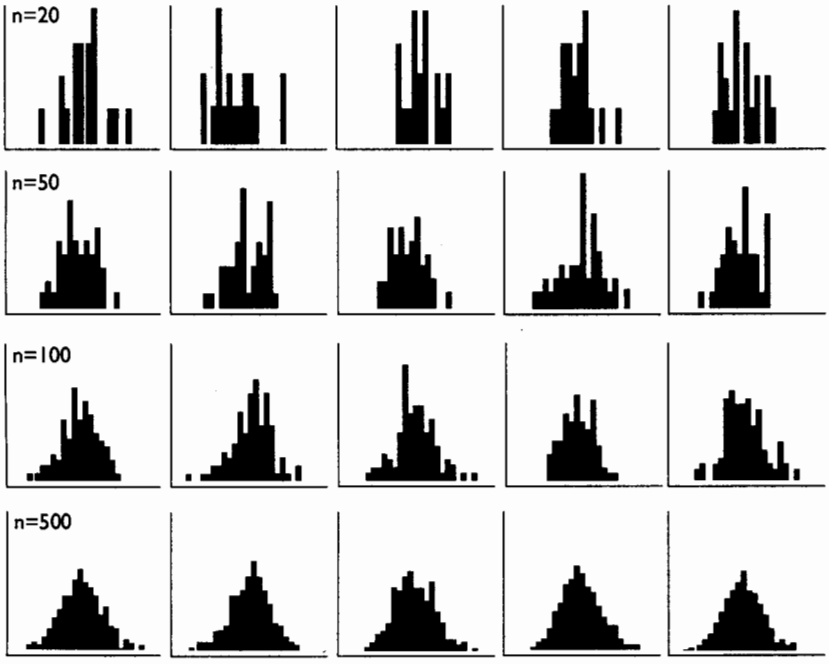


FIG 3—*Random samples from normal distributions—five samples of size 20, 50, 100, and 500*

Medical Statistics Laboratory, Imperial Cancer Research Fund, PO Box 123, London WC2A 3PX
Douglas G Altman, *Head*

Department of Public Health Sciences, St George's Hospital Medical School, London SW17 0RE
J Martin Bland, *reader in medical statistics*

Correspondence to:
Mr Altman.

BMJ 1995;310:298

1 Altman DG. *Practical statistics for medical research.* London: Chapman and Hall, 1991:132-45.
2 Gardner MJ, Altman DG. Calculating confidence intervals for proportions and their differences. In: Gardner MJ, Altman DG, eds. *Statistics with confidence.* London: British Medical Journal, 1989:28-33
3 Gardner MJ, Altman DG, eds. *Statistics with confidence.* London: British Medical Journal, 1989:17.