Key references

Carter HB, Pearson JD. PSA velocity for the diagnosis of early prostate cancer. *Urol Clin North Am* 1993;**20**:665-70

De Geeter P. The case against balloon dilatation of the prostate. *European Urology Update Series* 1993;**2**:146-51

Denis LJ. Diagnosing benign prostatic hyperplasia versus prostatic cancer. *Br J Urol* 1995;**76**(suppl 1):17-23

Partin AW, Oesterling JE. The clinical usefulness of prostatic antigen: update 1994. *J Urol* 1994;**152**:1358-68

Seaman E, Whang M, Olsson CA, Katz A, Coner WH, Benson MC. PSA density (PSAD). Role in patient evaluation and management. *Urol Clin North Am* 1993;**20**:653-63

Urethral stricture

Urethral stricture is also a cause of bladder outflow obstruction and is suggested by a history of urethral trauma, previous catheterisation, or sexually transmitted disease (although it may occur de novo in the absence of any of these factors). Uroflometry typically shows a trace with a plateau and a prolonged voiding cycle. Treatment initially is by urethrotomy, although complex or recurrent cases may need urethroplasty.

The two cystoscopic views were provided by Mr H N Blackford of the Edith Cavell Hospital, Peterborough.

The ABC of Urology is edited by Chris Dawson, a senior registrar in urology at the Edith Cavell Hospital, Peterborough, and Hugh Whitfield, a consultant urologist at the Central Middlesex Hospital and the Institute of Urology and Nephrology, London.

Statistics Notes

Transforming data

J Martin Bland, Douglas G Altman

This is the 17th in a series of occasional notes on medical statistics

We often transform data by taking the logarithm, square root, reciprocal, or some other function of the data. We then analyse the transformed data rather than the untransformed or raw data. We do this because many statistical techniques, such as *t* tests, regression, and analysis of variance, require that data follow a distribution of a particular kind. The observations themselves must come from a population which follows a normal distribution,[1] and different groups of observations must come from populations which have the same variance or standard deviation. We need this uniform variance because we estimate the variance within the groups, and we can do this well only if we can assume it to be the same in each group. Many biological variables do follow a normal distribution with uniform variance. Many of those which do not can be made to do so by a suitable transformation. Fortunately, a transformation which makes data follow a normal distribution often makes the variance uniform as well, and vice versa. In this note we shall try to explain why this is the case.

Firstly, the normal distribution and uniform variance go together. It can be shown mathematically that if we take random samples from a population the means and standard deviations of these samples will be independent (and thus uncorrelated) if the population has a normal distribution. In other words, the standard deviation of the samples will not be related to the mean. Furthermore, if the mean and standard deviation are independent the distribution must be normal. This is harder to credit, but it is true.

Secondly, if we add together many variables we usually get a normal distribution. For example, the central limit theorem shows that the means of large samples will follow a normal distribution, whatever the distribution of the observations themselves.[1] Similarly, if a biological variable is the result of the sum of many influences, it will follow a normal distribution. Human height is an example. Many biological measurements are not like this, however, but are the product of several factors. Substances in blood, for example, may be removed at a rate depending on the level of some other substance, which in turn is produced at a rate which depends on something else, and so on. We have the product of several influences multiplied together, rather than the sum. If we take the logarithm of the product of several variables, we get the sum of their logarithms.[2] So a variable which is the product of several factors has a logarithm which is the sum of several factors and so will follow a normal distribution.

Thirdly, any relation between variance and mean over several groups is usually fairly simple. The variance may be proportional to the group mean, the mean squared, the mean to the fourth power, etc. For such relations simple transformations can be found which will make the variance independent of the mean. If the variance is proportional to the mean we can use the square root transformation. This is often the case for data which are counts of things or events—for example, the number of cells of a particular type in a given volume of blood or number of deaths from AIDS in a geographical area over one year. Such data tend to follow a Poisson distribution, which has its variance equal to its mean. If the variance is proportional to the mean squared—that is, the standard deviation is proportional to the mean—we use the logarithmic transformation. This is the most frequent case in practice, suitable for variables such as serum cholesterol. If the variance is proportional to the mean to the fourth power—that is, the standard deviation is proportional to the mean squared—we use a reciprocal transformation, used for highly variable quantities such as serum creatinine. Thus we can transform the data to make the variance unrelated to the mean, in which case the data are likely to follow a normal distribution.

Some people ask whether the use of a transformation is cheating. There is no reason why the "natural" scale should be the only, or indeed the best, way to present measurements. pH, for example, is always presented as a logarithmic measure, $pH = -\log_{10}(H+)$, where H+ is the concentration of hydrogen ions in moles per cubic decimetre. Thus the "natural" scale is 10^{-pH}. This natural scale is very awkward to use, and the logarithm is always used instead.

If we can transform data to follow a normal distribution with variance independent of the mean, valid analyses can be carried out on this transformed scale. There is one drawback, however, as confidence intervals on the transformed scale may be difficult to interpret. We shall deal with this in a subsequent note.

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

ICRF Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, PO Box 777, Oxford OX3 7LF
Douglas G Altman, *head*

Correspondence to:
Professor Bland.

BMJ 1996;312:770

1 Altman DG, Bland JM. The normal distribution. *BMJ* 1995;310:298.
2 Bland JM, Altman DG. Logarithms. *BMJ* 1996;312:700.