Archives of Disease in Childhood 1993; **69:** 260–264

STATISTICS FROM THE INSIDE

11. Data transformations

M J R Healy

Additive and multiplicative effects

In the statistical analyses for comparing two groups of continuous observations which I have so far considered, certain assumptions have been made about the data being analysed. One of these is Normality of distribution; in both the paired and the unpaired situation, the mathematical theory underlying the significance probabilities attached to different values of t is based on the assumption that the observations are drawn from Normal distributions. In the unpaired situation, we make the further assumption that the distributions in the two groups which are being compared have equal standard deviations – this assumption allows us to simplify the analysis and to gain a certain amount of power by utilising a single pooled estimate of variance. It is necessary to stress that the importance of these assumptions is often exaggerated. Although they form part of the mathematical framework, departures from them have very little effect upon the outcome of the analyses unless they are particularly marked.

A third assumption is a good deal less obvious and a good deal more important. This is the assumption that the effect of the factor which defines the two groups is *additive*, meaning by this that it produces (apart from error) a constant *difference* between the readings in the two groups. Consider for example a paired situation, as when we compare readings before and after a treatment on the same subjects. For the purposes of analysis, we form the differences between the before and after readings for each subject, and the variability between these differences is regarded as error. Put another way, we assume that we can derive the after readings on the subjects (error apart) by *adding* a constant quantity to the corresponding before readings. Exactly the same is true of the unpaired situation, as when we compare independent samples of treated and control patients. Here the assumption is that we can derive the distribution of patient readings from that of control readings by shifting the latter bodily along the axis, and this again amounts to *adding* a constant amount to each of the control variate values (fig 1).

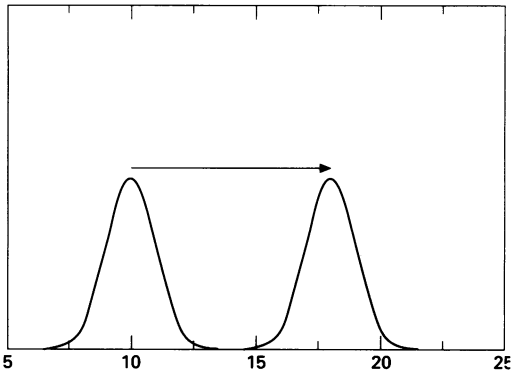

Figure 1 Two distributions differing by a constant amount.

This is not the only way in which two groups of readings can be related in practice. Suppose I asked you to guess the size of the effect of some treatment for (say) increasing forced expiratory volume in one second in asthmatic children. Your reply might well be 'Oh, perhaps +20%'. This implies a *multiplicative* effect of the treatment. A child with an initial level of 2·01 would be expected to increase this by 0·41, one with an initial level of 3·01 would be expected to increase this by 0·61. The difference between the 'before' and 'after' readings varies *systematically* with the level of response. If data following this pattern were to be analysed using the assumption of an additive treatment effect, this kind of systematic discrepancy between the (after−before) differences would be treated as if it was random and interpreted as being due to error.

Logarithmic transformation of data

A multiplicative treatment effect of this kind can be converted to an additive effect by transforming all the original readings to *logarithms* before doing the analysis. Your memories of logarithms may be of a rather outdated aid to arithmetic, but their usefulness in statistical analysis goes far beyond this. For statistical purposes, the only thing you need to remember about logarithms is that the logarithm of the *product* of two numbers is the *sum* of the logarithms of the numbers, while the logarithm of the *ratio* of two numbers is the *difference* between the logarithms. In symbols

$$\log(a \times b) = \log(a) + \log(b)$$
$$\log(a/b) = \log(a) - \log(b)$$

There are various kinds of logarithm, depending upon what number is chosen to have a logarithm of 1·0. So-called 'common' logarithms are based on $\log(10) = 1{\cdot}0$ (and consequently $\log(100) = 2{\cdot}0$, $\log(1000) = 3{\cdot}0$, $\log(0{\cdot}1) = -1{\cdot}0$ and so on, using the rules above). Mathematicians often use logarithms

23 Coleridge Court, Milton Road, Harpenden, Herts AL5 5LD

Correspondence to:
Professor Healy.
No reprints available.

based on log(e)=1·0, where e is a magic number equal to 2·7818 ... and rather presumptuously call these 'natural' logarithms. The numbers that describe the successive titres in a series of twofold dilutions are actually 'binary' logarithms based on log(2)=1·0. The numbers 10, e, and 2 are called the *bases* of the logarithms. Note that log(1)=0·0, no matter what the base is (you can prove this if you like, using the second of the above rules). It is usually unimportant for statistical purposes what base is used, since the different kinds of logarithm differ by no more than a scale factor. The 'natural' logarithm of a number, for example, is simply the 'common' logarithm multiplied by 2·3026. However, if you have to quote logarithmic values in the text or tables of a paper, it is essential to specify the base of the logarithms so as to avoid ambiguity – you can write $\log_{10}$, $\log_e$, or $\log_2$ as required. I personally use logarithmic values so often that I keep in my head a two figure table of common logarithms –

No	1	2	3	4	5	6	7	8	9	10
log	0·00	0·30	0·48	0·60	0·70	0·78	0·85	0·90	0·95	1·00

You may like to satisfy yourself that, for example, log(24)=1·38, log(0·15)=−0·82.

The treatment effect of +20% that we spoke of amounts to assuming that (apart from error) the treatment multiplies all the 'before' readings by the constant *factor* 1·20. When the observations are transformed to logarithms, the treatment effect correspondingly increases all the 'before' log observations by the constant *amount* 0·079, which is the logarithm (to the base 10) of 1·20.

A simple example, using paired data and logs to the base 10, is shown in the table. Here you can see quite a clear tendency for the difference within each pair of original readings to increase with the average level. The ratios show no such systematic tendency, suggesting that the data are closer to showing a constant ratio rather than a constant difference, in other words, that the effect of treatment is closer to being multiplicative than additive.

If you calculate the paired *t* value from the differences, you will get *t*=0·80/0·093=8·60.

Logarithmic transformation

Before	*After*	*Difference*	*Ratio*
	Original data		
2·0	2·4	+0·4	1·20
2·2	2·9	+0·7	1·32
2·5	3·2	+0·7	1·28
2·8	3·6	+0·8	1·29
3·2	4·3	+1·1	1·34
3·7	4·5	+0·8	1·22
4·3	5·4	+1·1	1·26
Mean (SE)		+0·80 (0·093)	
	Logged data		
0·301	0·380	+0·079	
0·342	0·462	+0·120	
0·398	0·505	+0·107	
0·447	0·556	+0·109	
0·505	0·633	+0·128	
0·568	0·653	+0·085	
0·633	0·732	+0·099	
Mean (SE)		+0·104 (0·0067)	

If instead you analyse the differences of the logs of the original readings, that is the logs of the ratios, you will get *t*=0·104/0·0067=15·52, a considerably larger value. Much more to the point, the mean and standard error for the differences are +0·80 and 0·093, so that the 95% confidence interval for the true mean difference (using the 5% significance value of *t* with 6 degrees of freedom) is 0·80±2·447×0·093=+0·57 to +1·03. For the logarithms of the ratios the mean and standard error are 0·104 and 0·0067, giving 95% limits for the true mean difference of 0·104±2·447×0·0067=0·088 to 0·120. This interval relates to the differences between the logarithms of the original data; taking the antilogs of these figures, the 95% confidence interval for the true ratio in the original data is from 1·22 to 1·32, a much more incisive result.

Transformation of data and its effects

In some circumstances an additive effect is rather implausible, so that it should be considered whether transforming the data to logarithms would be advantageous. One such situation arises when the data values are necessarily positive (as with, for example, the concentrations of some chemical in the blood) but are close to a definite zero point, in the sense that zero is within one or two standard deviations of the mean so that the coefficient of variation (the ratio of the standard deviation over the mean) is 50% or more. Suppose that this describes the distribution of the control values in a trial; then with an additive treatment effect, the treated values would have to start abruptly at some non-zero value, an unlikely state of affairs. A multiplicative effect as shown in fig 2 is a more realistic model for data of this kind. A great many measurements, both physiological and biochemical, exhibit this behaviour.

But it is worth noticing that this situation, marked by a large ratio of standard deviation to mean, is also unlikely to meet the other two assumptions I have mentioned, of Normality and equal variability. With a mean close to zero in the above sense, a Normal distribution would necessarily imply an appreciable probability for the impossible negative values. Figure 2 also shows that, with a fixed terminus at zero for both distributions, the

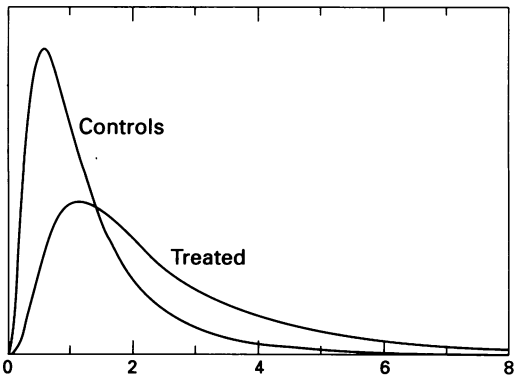

Figure 2 Two distributions differing by a multiplicative factor.

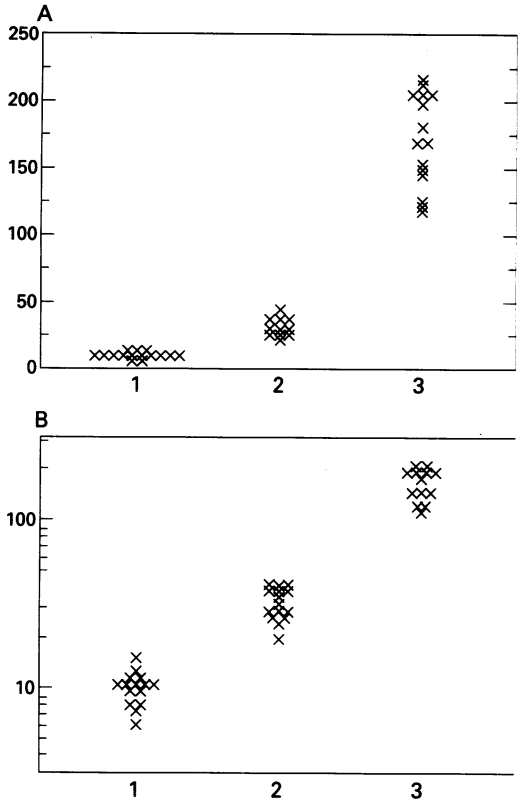

Figure 3 Three samples with similar coefficients of variation. (A) arithmetic scale; (B) logarithmic scale.

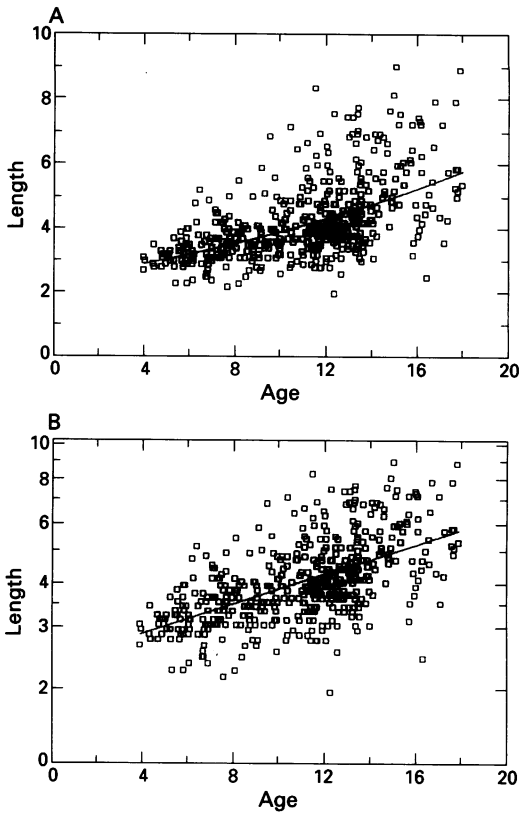

Figure 4 Ovary length versus age. (A) Arithmetic scale; (B) logarithmic scale.

distribution with the higher mean is likely also to have the larger variability. It is important to realise that transformation of the original data values to logarithms is likely in practice to be helpful in both these directions. The transformation has two effects on the shape of the distribution; it pulls in the long right hand tail and it extends the left hand tail, sending the barrier at zero off to infinity. In this way it can bring very skew distributions close to Normality. It can also bring the unequal standard deviations closer to agreement. In fact, it can be shown that, if the original data have equal coefficients of variation rather than equal standard deviations, then the logarithms of the data will themselves have equal standard deviations. An example of all this, with three independent samples, is shown in fig 3. Once again, if asked how variable a particular kind of data was, most people would reply 'Oh, ±15% or so'. This implies a constant coefficient of variation, not a constant standard deviation as the standard analyses assume.

Transformation of the data can also be useful in a regression context. It is remarkable how often a curvilinear relationship can be converted at least approximately to a straight line by transforming either the y's or the x's to logarithms. An example is shown in fig 4, where ultrasound measurements of the size of the ovary are plotted against age (I am grateful to Dr P Hindmarsh for access to these data). On the original scale the relationship is curved, and the scatter of the points increases with age. It will be seen that a log transformation of the y axis straightens out the relationship, and also tends to equalise the scatter.

The idea of transforming the data before doing a statistical analysis is unappealing to some people, who suspect an element of statistical cookery. The suspicion is quite unwarranted. It is to the contrary a manoeuvre necessitated by the fact that many datasets do not support the assumptions which underlie the usual methods of simple statistical analyses, but can be made to do so by re-expressing them on a transformed scale. The use of a logarithmic scale cannot be claimed to be unfamiliar by anyone who is used to such scales as pH and decibels.

Quite apart from the technicalities of statistical analysis, the usefulness of logarithmic scales for graphical data presentation should also be emphasised. Data such as those plotted in fig 5A (note the off scale points) are much clearer and more informative on the logarithmic scale of fig 5B.

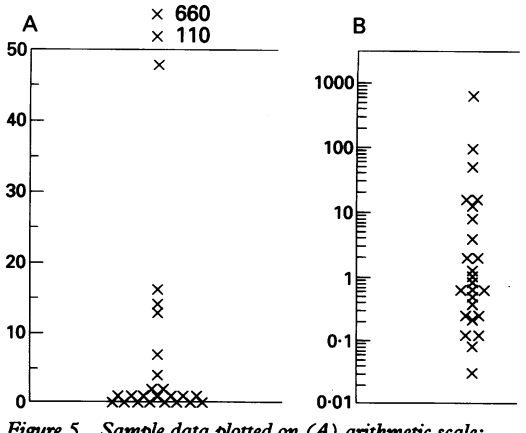

Figure 5 Sample data plotted on (A) arithmetic scale; (B) logarithmic scale.

A logarithmic scale is particularly appropriate when the quantity plotted is a ratio or percentage. Consider a scale of (treatment−control) differences. The null value is 0 and differences of +1 and −1 can be interchanged by taking the differences the other way round. With a plot of (treatment/control) ratios, the null value is 1 and ratios of 2 and ½ are interchangeable by taking the ratios the other way up. Plotting the ratios on a logarithmic scale produces a more intuitive display.

There are many other ways of transforming continuous measurement data as well as the logarithmic transformation, but none of these is equal to it in practical importance. Most of the common transformations involve raising the data values to a power, such as the square root. These have the drawback that the results may be hard to interpret – a treatment effect which is constant on a square root scale is at best unfamiliar. An exception is the reciprocal transformation, where a value y becomes 1/y (or perhaps −1/y, so as to avoid the scale being inverted). This is commonly used for analysing successive creatinine concentrations in cases of kidney failure, on the grounds that reciprocal creatinine often decreases with time at a more or less constant rate. It can also be useful when analysing the times taken to reach some event, such as death or recovery. In this context, the transformed values can sometimes be interpreted as the speed with which the event is being attained.

More esoteric transformations such as x^{λ} or $\log(x+c)$ can be useful in specialised circumstances when Normality of distribution is genuinely important. T J Cole for example has recommended the use of such transformations in the construction of age related centile charts (*European Journal of Clinical Nutrition* 1990; **44:** 45–60).

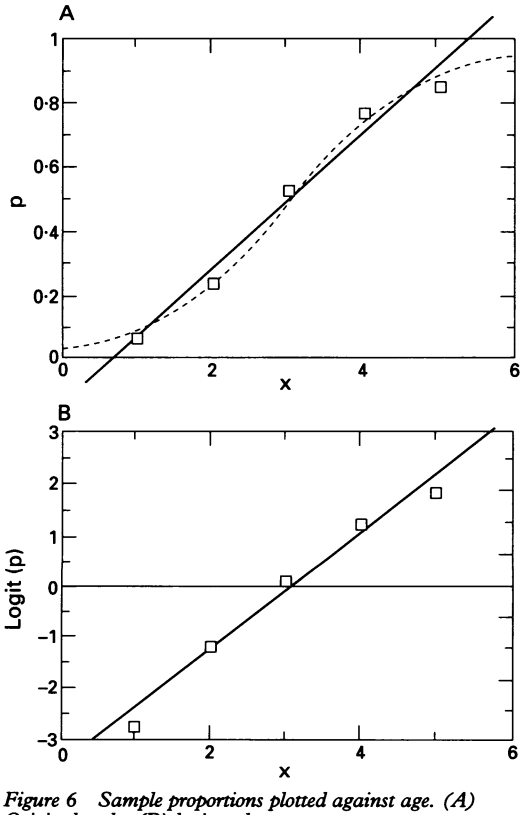

Figure 6 Sample proportions plotted against age. (A) Original scale; (B) logit scale.

Transformations for discrete data

A quite different set of data transformations are applicable to discrete data in the form of counted proportions. A particular difficulty arises when such proportions occur as the y variate in a regression context – we might, for example, wish to relate the proportion of girls who have reached menarche to age. The problem here is that the true proportions cannot go outside the range from 0 to 1. If a simple regression of proportion on age is fitted, then at extreme ages the predicted 'proportions' will be impossible. We need a transformation which removes the barriers at 0 and 1 by sending them off to infinity in both directions. There are several possibilities, one of the commonest being the *logit* or *log odds* transformation which transforms a proportion p to a value $z=\log(p/(1-p))$ using natural rather than common logarithms. This is illustrated in fig 6 which shows a set of proportions plotted against an x variable and the same data after transforming the proportions to logits. Real life data surprisingly often are well approximated by a straight line after a logit transformation; a log transformation of the *x* scale is sometimes needed as well.

A very similar diagram would illustrate the *probit* or *Normal equivalent deviate* transformation. This stems from the observation that the dotted curve in fig 6A (to a statistician's eye it resembles a letter S and is often called a *sigmoid* curve) is similar to the curve describing a Normal distribution in its cumulative form. Suppose that, given a proportion p, we transform it to a deviate z which cuts off a proportion p on the left of a Normal distribution (for instance if p=0·5, z=0·0; if p=0·975, z=1·96). The quantity z is known as the Normal Equivalent Deviate or NED of p (the probit of p is just the NED plus 5). Then if the proportions p lie on a Normal sigmoid, the transformed values z will lie on a straight line. Strangely, the probit and logit transformations are very closely similar and can only be distinguished in very large samples.

The probit transformation can be motivated by a simple model for the data. Suppose we observe the proportion of girls who have achieved menarche at a set of different ages and plot these proportions against age. Suppose too that the distribution of the actual ages at menarche is Normal. Then the plot we have made will be the cumulative version of this Normal distribution and a probit transformation will convert it into a straight line. The logit transformation can be linked in the same way to a distribution very like the Normal but with longer tails.

You should note that the objective of the logit and probit transformations is to cause the transformed data points to lie on a straight line. This corresponds to a form of additivity, with constant increments in the x variable leading to constant increments in the transformed proportions. The original data

will follow a set of binomial distributions and these will have different standard deviations, another departure from the usual simple regression model. The logit and probit transformations do not help with this and they call for a rather complicated form of weighted regression (this also has to cope with observed proportions of 0 of 100% for which the transformed values are infinite). There are good computer programs available for implementing this, notably the GLIM and GENSTAT packages.

Startle disease

I think of it as the stiff baby syndrome but it seems that the in-crowd are now calling it familial startle disease or hyperexplexia. A faint echo from a previous incarnation reminds me of the 'jumping Frenchmen of Maine'. It is an autosomal dominant condition presenting usually in the newborn with either muscle rigidity or episodes of stiffening and apnoea often misdiagnosed as epileptic seizures. An exaggerated response to glabellar tap is characteristic and the jerks and spasms are inhibited by tight swaddling or trunk flexion.

Two recent publications throw further light on the syndrome. A paper from America (Stephen G Ryan and colleagues, *Annals of Neurology* 1992; **31**: 663–8) describes 30 affected people in five generations of a single family. All were hypertonic at birth and feeding difficulty was common as were inguinal and umbilical hernias, presumably caused by raised intra-abdominal pressure. Four babies died from apnoea due to intense muscular spasm. Motor development was delayed in the early years but improved later as the stiffness regressed and most of the affected individuals were of normal intelligence. In adults the major complaint is of sudden falls caused by transient intense muscle spasm with inability to extend the arms, resulting in frequent head and face injuries. (Look for the scars on the parents' faces.) In this American series 16 patients were treated with clonazepam, all apparently with 'dramatic' improvement. Genetic linkage studies on this family put the gene on the long arm of chromosome 5 linked to a marker locus (colony stimulating factor receptor or CSFIR) at 5q 33-q 35. Several genes are known to be located in this region including one which encodes a subunit of the gamma-amino butyric acid (GABA) receptor.

From the Hammersmith Hospital in London Dr Lilly Dubowitz and her colleagues (*Lancet* 1992; **340**: 80–1) describe a baby with this condition who was at first thought to be suffering from neonatal epilepsy. They measured the concentration of free GABA in the cerebrospinal fluid at 14 days and found it to be low compared with previously published data. The baby improved on treatment with clonazepam and at nine weeks the GABA in the cerebrospinal fluid was within the normal range.

The basic disorder in startle disease is not understood and the neuropharmacology and neurochemistry are complex. Clonazepam is a serotonin antagonist and an excess of serotoninergic transmission in the medullary and pontine reticular formation could, apparently, explain the condition. GABA, of course, is an important inhibitory neurotransmitter and the finding of low cerebrospinal fluid GABA is an interesting clue. Dubowitz and colleagues suggest that clonazepam might work by increasing the sensitivity of the GABA receptor though they don't explain how it might do that.

Clearly the hunt is well and truly on for the explanation of this rare but fascinating disease.

ARCHIVIST