Archives of Disease in Childhood 1994; **70:** 158–163

STATISTICS FROM THE INSIDE

12. Non-Normal data

M J R Healy

The Normality assumption

When we consider the analysis of continuous data (measurements rather than counts), we may think first of describing our samples by way of means and standard deviations and then of calculating significance levels and confidence intervals by way of Student's *t* distribution. These latter manoeuvres, mathematically speaking, are based on the assumption that the data are drawn from Normal distributions. This assumption can itself be tested statistically. The numerous tests that are available are not particularly powerful (they will often give 'non-significant' results when applied to samples of moderate size from definitely non-Normal distributions), and for this very reason the truth of the Normality assumption may frequently be questioned. Many users of statistical methods consequently feel the need for analytical techniques which do not depend upon the Normality of the data.

It may certainly be doubted whether a genuinely Normal distribution ever existed in the real world. It has been well said that practitioners believe in the Normal distribution because they think that the mathematicians have proved its existence, while the mathematicians believe in it because they think that the practitioners have discovered it in their data. Nobody should pretend that *exact* Normality is the normal state of affairs. How come then that methods based upon the Normality assumption are so widely used?

One reason (which I have emphasised in previous articles in this series) is that the Normality assumption is not a very important one. No doubt, the significance probabilities found in the usual tables and computer programs for *t* tests and the like will be wrong if the data to which they are applied come from non-Normal distributions; but they will not be far wrong unless the degree of non-Normality is extreme. The fact that these methods are optimal for genuinely Normal data suggests that they are unlikely to be greatly improved upon when the data are only approximately Normal, and the clarity and flexibility of Normal-theory methods are enough to justify their widespread use.

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

Correspondence to:
Professor Healy.
No reprints available.

Non-Normality in practice

What departures from Normality actually occur in practice? In principle there are unlimited ways in which data distributions can depart from Normality, but most non-Normal datasets, at least in a medical context, exhibit one of two types of pattern. First, we often encounter distributions which are not symmetric but *skew*, with a long tail of high values. Such distributions usually start at or near zero; they are characteristic of a wide variety of physiological measurements, ranging from skinfold thicknesses to enzyme concentrations. In clinical and laboratory research, these skew distributions are if anything more common than the Normal distribution of the textbooks.

The other type of non-Normal data which commonly occurs in practice consists of a central part coinciding fairly closely with Normality, plus one or a few extreme values on the high or low side. Such outlying values are a source of considerable difficulty in statistical analysis. It is orthodox to take the view that all data values obtained should be presented and included in the statistical analysis. Yet outlying values, by definition, are different from the main body of the data and it may simply obscure the message that the data are trying to impart to analyse them all together. In theory, such outlying values may be a source of new discovery – if one or two subjects out of 40 or so behave quite differently from the rest, maybe they belong to an unrecognised subcategory and this finding could lead to new scientific discoveries. Long and bitter experience, however, leads me to suggest that the vast majority of outlying values represent no more than errors of recording or transcription. Only professional statisticians, and experienced ones at that, are fully aware of human frailty (including their own) in the practical handling of numerical material.

Checking on non-Normality

How should one decide whether one's sample is non-Normal to an extent which requires attention? It is tempting to suggest, as above, that one of many possible significance tests should be used. However, the use of a significance test in this context is logically inappropriate. In the first place, a non-significant result can never be interpreted as proving the truth of the null hypothesis; a test of Normality which yields a result which is non-significant (at some conventional level or another) cannot establish the fact that the sample comes from a Normal distribution. In fact, as mentioned above, we can be fairly sure in advance that the distribution from which our sample is drawn is not rigorously Normal. Secondly, a significant

Arch Dis Child: first published as 10.1136/adc.70.2.158 on 1 February 1994. Downloaded from http://adc.bmj.com/ on January 30, 2023 by guest. Protected by copyright.

result (again, at some conventional level) may be taken to establish non-Normality, but with a sample of moderate size and an appropriate test the degree of non-Normality established may be small enough to be safely ignored. Many of the available tests are unsatisfactory in that they do not indicate whether or not steps can be taken to counteract the non-Normality in ways which I describe later.

The most useful ways of investigating possible non-Normality are graphical. Suppose we take a sample of size 20 (say) from a genuine Normal distribution and sort the values into ascending order. It can be shown mathematically that the smallest sample value will lie on average 1·87 standard deviations below the mean, the second smallest 1·41 standard deviations below the mean, the third smallest 1·13 standard deviations below the mean, and so on. The numbers −1·87, −1·41, −1·13, ... are called *Normal scores* for a sample size of 20. They are available for any sample size in several collections of tables and computer programs; the score for the i'th value in a sample of size n is well approximated by the Normal equivalent deviate of the fraction $(i-\frac{1}{2})/n$.

Suppose now that we plot the observed values in the sample against their Normal scores. On average, this will produce a straight line whose slope is one over the standard deviation. Systematic departures from a straight line are evidence of non-Normality – skewness, for example, produces a quadratic-like curve – and outliers show up a conspicuous departure from the line determined by the bulk of the sample. This graphical display is called a *Normal plot*. Some experience is needed to assess it judiciously, but it is an extremely useful and simple procedure and I strongly recommend that it should routinely precede most statistical analyses. Some examples appear later on in this article.

Data description

What problems, then, do non-Normal data present? First, we should consider questions of data description. It may need stating that there is nothing wrong with calculating and presenting the mean and standard deviation of a non-Normal sample as descriptive properties. The trouble is that the two quantities describe the sample less effectively when Normality cannot be assumed – it is unsafe to suppose that the mean is in the middle of the sample, and more so to reckon that the bulk of the sample will lie within 2SD on either side of the mean.

There are various alternatives which are commonly used. The sample *median* is the observation which divides the sample values into equal halves when they are arranged in ascending order. It is by definition the central value and it is less influenced than the sample mean by a few extreme readings such as are liable to occur in the long tail of a skew distribution. Describing variability is more difficult. The extreme sample values (which define the sample *range*) are easy to determine and their meaning is easy to appreciate. The principal defect of the range as a measure of variability is that it is not independent of sample size; a large sample is likely to have extreme values which are farther apart than those of a smaller sample from the same population. Rather than quoting the extreme values, there is a case for giving the *quartiles*, the observations which divide the sorted sample into four equal parts, but it must be admitted that these are essentially less easy to comprehend (in a Normal sample, the expected positions of the quartiles are at 0·675 standard deviations below and above the mean). A skew sample of 28 values with its median and quartiles is shown in fig 1A. A graphical display of the same sample known as a *boxplot* is shown in fig 1B. Here the box is delimited by the quartiles, with the median marked by an asterisk, and the more extreme sample values are shown individually. The mean of the sample and the points 1 SD below and above it are also marked in fig 1A. Note that the quantity (mean −2 SD), sometimes taken as the 'lower limit of normality', is negative. Figure 1C shows a Normal plot of the sample data; the curvature is obvious.

Another Normal plot is shown in fig 2A (a boxplot (fig 2B) provides the same information in a different form). This is based on a small clinical trial of aspirin prophylaxis against migraine (B P Neill and J D Mann. Aspirin prophylaxis in migraine. *Lancet* 1978; ii: 1178–81). The data are shown in table 1 and the numbers plotted are the differences between the two columns of the table. It is fairly clear from the figure that one of the patients has behaved quite differently from the others, her 15 attacks per three months having

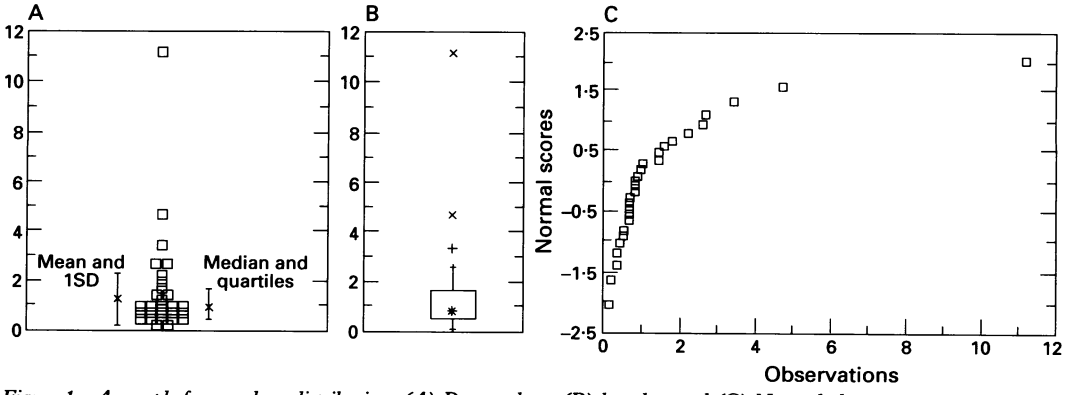

Figure 1 A sample from a skew distribution. (A) Data values, (B) boxplot, and (C) Normal plot.

Arch Dis Child: first published as 10.1136/adc.70.2.158 on 1 February 1994. Downloaded from http://adc.bmj.com/ on January 30, 2023 by guest. Protected by copyright.

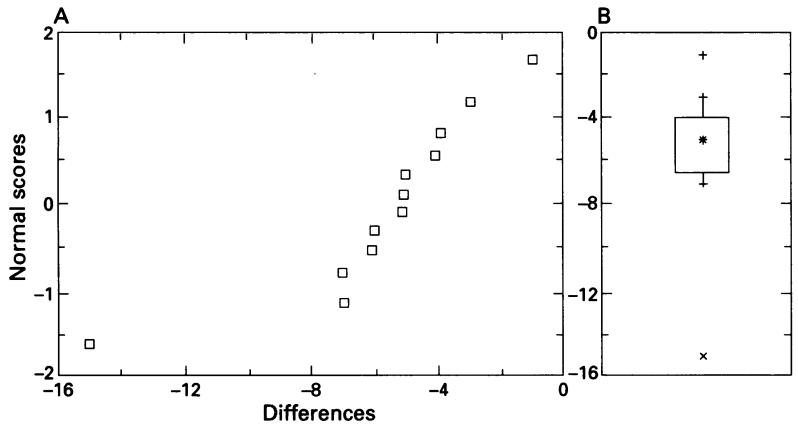

Figure 2 A sample containing an outlier. (A) Normal plot and (B) boxplot.

Table 1 Trial of aspirin in migraine prophylaxis; number of attacks per three months

Placebo (P)	*Aspirin (A)*	*(A−P)*
5	4	−1
6	1	−5
18	14	−4
6	1	−5
7	3	−4
15	0	−15
9	3	−6
9	2	−7
7	2	−5
8	2	−6
8	5	−3
9	2	−7

been totally abolished. She may have been someone on whom aspirin has a special effect and whose thrombotic and other mechanisms deserve further study; alternatively, she may have detected the active treatment and tried to be helpful, or she may simply have become bored reporting all those headaches. Either way, it does not make a lot of sense to report a mean decrease of 5·7 headaches per three months (SD 3·4) based on all 12 patients. It is much more informative to summarise the remaining 11 (mean 4·8, SD 1·8) along with the actual value of the outlier.

The median as a descriptive statistic has one drawback which is not commonly realised. Consider a set of paired measurements taken before and after treatment. We can take the mean of the before measurements and that of the after measurements, and form the difference between these to assess the change in the

Table 2 Paired data

	Before	*After*	*Difference*
	11	97	+86
	23	86	+43
	47	5	−42
	62	53	−9
	81	21	−60
Means	44·8	48·4	+3·6
Medians	47	53	−9

measurement due to treatment. Alternatively we can form the (after−before) differences for each subject and take the mean of these differences. These two methods of calculation give exactly the same answer. In a phrase, the mean of the differences is equal to the difference between the means. This is not in general true for the median – the median of the changes due to treatment is not necessarily equal to the difference between the before and after medians. This can lead to paradoxical results. Look at the miniature example in table 2. The median has increased after treatment from 47 to 53, but the median of the changes turns out to be negative.

Significance testing

If it is desired to draw conclusions from the data avoiding the assumption of Normality, what is to be done? In theory, one might attempt to describe the distribution of the data by some different mathematical formula and work out some equivalent of the *t* distribution, but the range of possibilities is far too wide for this to be a practical proposition in general terms. One very important possibility, though, is that of transforming the data to a scale on which they are more nearly Normally distributed. I have discussed this in a previous note; it is remarkable how often transforming readings to logarithms, for example, produces a set of figures which are a very plausible sample from a Normal distribution. Figure 3A shows the results of transforming the data from fig 1A in this way. Note that there is much more graphical information about the lower values, and that the mean and SD are very reasonable descriptive statistics for the sample. Figure 3B is a Normal plot of the logged sample, showing that, contrary to appearances, the largest observation cannot be considered to

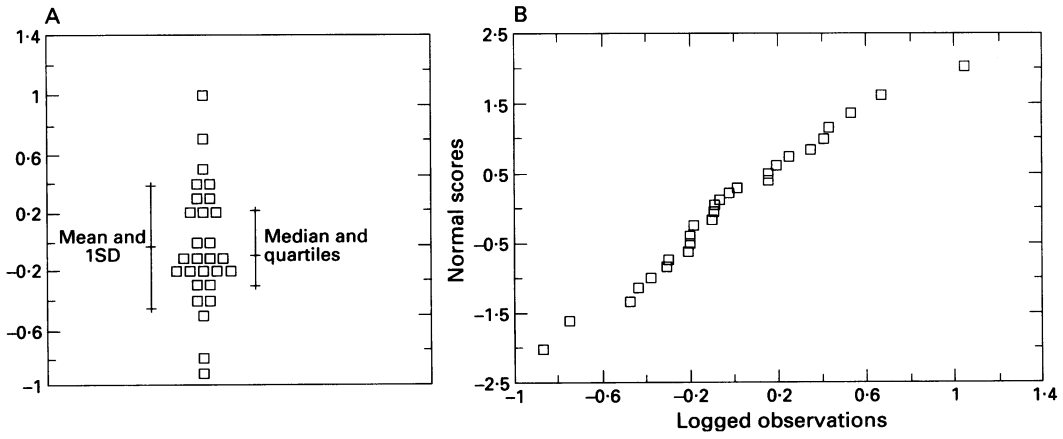

Figure 3 Data from fig 1 after log transformation. (A) Data values and (B) Normal plot.

Arch Dis Child: first published as 10.1136/adc.70.2.158 on 1 February 1994. Downloaded from http://adc.bmj.com/ on January 30, 2023 by guest. Protected by copyright.

be an outlier. When, as is so often the case, non-Normality takes the form of more or less extreme skewness to the right, the best way of handling the data will usually consist of a logarithmic transformation, with its effectiveness checked by means of a Normal plot.

Ranking methods

How far can we get without making any assumptions at all about the distribution of the population from which the data values are sampled? Remarkably, the answer is a very fair distance. Consider again the differences in table 1. There are 12 of them and they are all positive. On the null hypothesis of no treatment effect, is this 'significant' (that is surprising)? If the treatment has absolutely no effect, presumably each difference is equally likely to go either way, to be either positive or negative. The signs of the 12 differences will thus be a sample from a binomial distribution with $p=½$, and the probability of getting 12 positives will be $(½)^{12}=0{\cdot}00024$. This is a one tailed probability; doubling it (the binomial distribution is here symmetric), the significance level using only the signs of the data is 0·00048. This technique is called a *sign test*. Its null hypothesis is formally that the median of the underlying population is zero.

Table 3 Differences and ranks

−1	−4	5	10	−13	15	18	21	26	29
1	2	3	4	5	6	7	8	9	10

As might be expected, a price has to be paid for not using the Normality assumption when it is in fact true. Consider the less extreme situation in table 3. Here we have a sample of differences which do not appear to depart from Normality to any marked extent, so that the significance probability provided by the *t* distribution will be very close to the truth. The mean difference is 10·60, SE 4·34, $t=2{\cdot}44$ on 9 degrees of freedom, $p=0{\cdot}037$, a reasonably convincing level of significance. But 7 +'s out of 10, using the sign test, corresponds to a significance level of only 0·34, much less extreme. Ignoring the sizes of the differences has jettisoned a good deal of useful information.

We can do a good deal better than this by replacing the original data values, not simply by their signs, but by their *ranks*. The trick is to sort the data values ignoring their signs and label them from 1 to 10. Now add up the ranks of the negative differences and look up the result (8 in our example) in appropriate tables.

Table 4 Independent samples

	Controls			*Treated*		
	Data	*Ranks*	*Logs*	*Data*	*Ranks*	*Logs*
	1	1	0·000	3	3	0·477
	2	2	0·301	5	5	0·699
	4	4	0·602	6	6	0·778
	9	7	0·954	13	8	1·114
	17	9	1·230	25	10	1·398
				33	11	1·518
Means	6·60		0·617	14·17		0·997
SDs	6·58		0·492	12·24		0·413

This is *Wilcoxon's signed rank test*. In table 3 the significance level is 0·048, not too different from that given by the *t* test.

The same thing can be done in other circumstances. Take the data in table 4, consisting of two independent samples. The values look very skew (the standard deviations are much the same size as the means) and one standard deviation is almost twice the size of the other, so that the usual assumptions of the unpaired *t* test are rather dubious. For what it is worth, this test gives a difference in means of 7·57, SE 6·13, $t=1{\cdot}235$ on 9 degrees of freedom, $p=0{\cdot}25$. As an alternative, we can rank the 11 data values taken all together as in the table, add up the ranks in one of the samples and again refer the result to suitable tables. This is the *Mann-Whitney* test (it occurs in other flavours and under other names, including confusingly that of Wilcoxon). The significance probability for the data in table 4 is $p=0{\cdot}24$. As a third possibility we can convert the data values to logarithms. As the table shows, the figures are now less skew and the two standard deviations are nearly equal. The difference in means is 0·380, SE 0·273 giving $t=1{\cdot}393$ on 9 degrees of freedom, $p=0{\cdot}20$. Much more to the point, the technique provides a 95% confidence interval for the difference between the two mean logs of −0·996 to 0·237, corresponding to a ratio between 0·101 and 1·73 on the original scale.

Two other rank based tests of this kind are in common use. If we have several independent samples we can again rank all the data values together and do a one way analysis of variance on the ranks, obtaining an F test for differences between the groups. This is the *Kruskal-Wallis* test. With two variates x and y, we can rank the x's and the y's separately and calculate the correlation coefficient between the ranks. This is known as *Spearman's rank correlation coefficient*.

You will have noticed that these last two tests are implemented by using the usual statistical arithmetic applied to the ranks of the data values. In fact, the Wilcoxon and Mann-Whitney tests can be implemented in exactly the same way, by doing appropriate *t* tests on the ranks as if they were measurements. The usual *t* probabilities will not be exact but will be quite a good approximation.

Tests of the kind which make no assumptions about the underlying distributions from which the samples are drawn are called *distribution free*. There being no reference to distributions, the hypotheses tested cannot involve parameters so that the tests are also called *non-parametric*. It appears necessary to say that the phrase 'non-parametric data' is meaningless; 'non-Normal data' is what is usually intended.

Non-parametric methods, for and against

Non-parametric methods cover the same ground as the simpler statistical tests of hypotheses using the *t* distribution and the Normality assumption. They generate a warm glow of confidence in practitioners who are nervous about assuming Normality, perhaps

Arch Dis Child: first published as 10.1136/adc.70.2.158 on 1 February 1994. Downloaded from http://adc.bmj.com/ on January 30, 2023 by guest. Protected by copyright.

not realising the relative unimportance of the assumption in practice. Is there any reason why they should not be the usual methods of data analysis?

One accusation that can certainly not be raised against ranking tests is one of inefficiency, of wasting the information in the data. The relative efficiency of two alternative tests of significance can be measured by the ratio of the sample sizes needed to achieve the same power for a given significance level and a given departure from the null hypothesis – if one test needs twice as many observations as another, its relative efficiency is 50%. When the Normal assumption is true, the usual t and F tests can be shown to be the most efficient that can be devised so that under these circumstances using a non-parametric test must be equivalent to throwing away some of the observations. However, the cost is remarkably small, no more than around 5%. With non-Normal data non-parametric methods, apart from giving a correct significance level, may be more efficient than t tests. Many people feel that the small price is worth paying for peace of mind.

It is ironic that the high efficiency of ranking methods was not always realised by their originators, who offered the tests as quick-and-dirty alternatives aimed at people who did not possess the mechanical calculators of the day. In actual fact they are a fiddly nuisance to apply by hand and they have only become really popular since being incorporated into computer packages.

Even so, there is something a bit odd about ranking tests. One may ask, if the idea of drawing one's sample from a distribution is given up, how is the significance probability arrived at? With no parent distribution, how can there be a sampling distribution? The answer is an ingenious one. Take the Wilcoxon test as an example. In table 3 there were 10 differences which were ranked, and on the null hypothesis each difference was equally likely to have been positive or negative. We could set out all the 1024 possible patterns of + and − signs, apply them to the ranks and work out the Wilcoxon statistic for each pattern. On the null hypothesis all the patterns are equally likely so that each value of the statistic has a probability of 1/1024, and this provides the distribution we require; if the value actually observed lies in one or other of the tails, we say that it is significant. But it should be noted that the 10 differences remain fixed throughout this argument – there is no reference to what might have happened to other patients with other values of the difference. In particular, the method does not permit us to assess the probability of the treatment being effective on a new patient – the significance probability is quite distinct from this. Exactly the same is true, *mutatis mutandis*, for the other ranking tests. As a solution to the fundamental problem of statistics, that of arguing from the particular to the general, non-parametric methods must be subject to question.

Ranking methods do not cope with the problem of outliers, or only by sweeping it under the rug. Distribution free methods of their very nature do not allow us to cast suspicion on a reading which differs by what may seem to be an implausible amount from the bulk of the data. Such readings must be important, no matter what their origin; either they are genuine, in which case they deserve further investigation, or they are spurious (in my experience, 95% or so of the time) and have no business forming part of the analysis. The non-parametric approach encourages the throw-it-at-the-computer-and-stand-back attitude to statistical analysis which is no part of good scientific practice.

But the most important case against the widespread use of non-parametric methods is that they do not lend themselves at all readily to estimation. This is almost true by definition – if no parameters are brought into consideration, it is difficult to see how to describe and assess the *magnitude* of a treatment effect, always so much more important than mere statistical significance. It cannot be said often enough that simply establishing that an effect exists (at some conventional level of significance) is hardly ever a satisfactory goal for an illuminating statistical analysis.

It is actually possible to obtain confidence intervals in a non-parametric framework. On average just 50% of sample observations are expected to fall below the population median, so the number doing so in a particular sample will follow a binomial distribution with $p=0{\cdot}5$. This fact can be used to obtain a confidence interval for the population median (agreeable philosophical discussions can be had as to whether or not the median is a population parameter). An extension of the argument leads to confidence intervals for other quantiles, such as the 2½% point which is commonly used as a lower normal limit or reference value. It is worth noting that the price paid for abandoning the assumption of Normality (or logNormality) when it is in fact justified is a heavy one in this context – to achieve the same level of precision in estimating the 2½% point without assuming Normality requires about between 2 and 3 times as many observations.

Far more commonly, what is required is a confidence interval for the difference between two population medians, treated v control or after v before. Here again, methods are available for obtaining such intervals without assuming Normality of the populations (see chapter 8 of *Statistics with Confidence*, M J Gardner and D G Altman, eds. BMA, 1989). But these methods are by no means distribution free. They rest upon the assumption that the two distributions are identical in shape and differ only in location. This is precisely the sort of assumption that users of non-parametric methods are trying to avoid. It is almost certainly false when the distributions concerned share a common start, as is the case for many biochemical and endocrinological determinations. Happily, the fact that the methods concerned are not widely available in computer packages has prevented them from being extensively used.

On balance, the widespread use of non-parametric methods is in my view pernicious. They

Arch Dis Child: first published as 10.1136/adc.70.2.158 on 1 February 1994. Downloaded from http://adc.bmj.com/ on January 30, 2023 by guest. Protected by copyright.

encourage the two major sources of bad statistical practice: failure to look carefully at the data values, and concentration upon significance testing at the expense of estimation. If I were asked to name ways in which published medical statistical analyses might be improved, a decrease in the frequency of use of non-parametric methods would come high on my list.

All sample data should be plotted and examined before they are analysed – Normal plots and boxplots are useful graphical tools for this purpose. Outlying values call for special investigation. Apart from these, if the assumption of Normality seems questionable, a simple transformation of the data will very often remedy the situation.

Arch Dis Child: first published as 10.1136/adc.70.2.158 on 1 February 1994. Downloaded from http://adc.bmj.com/ on January 30, 2023 by guest. Protected by copyright.