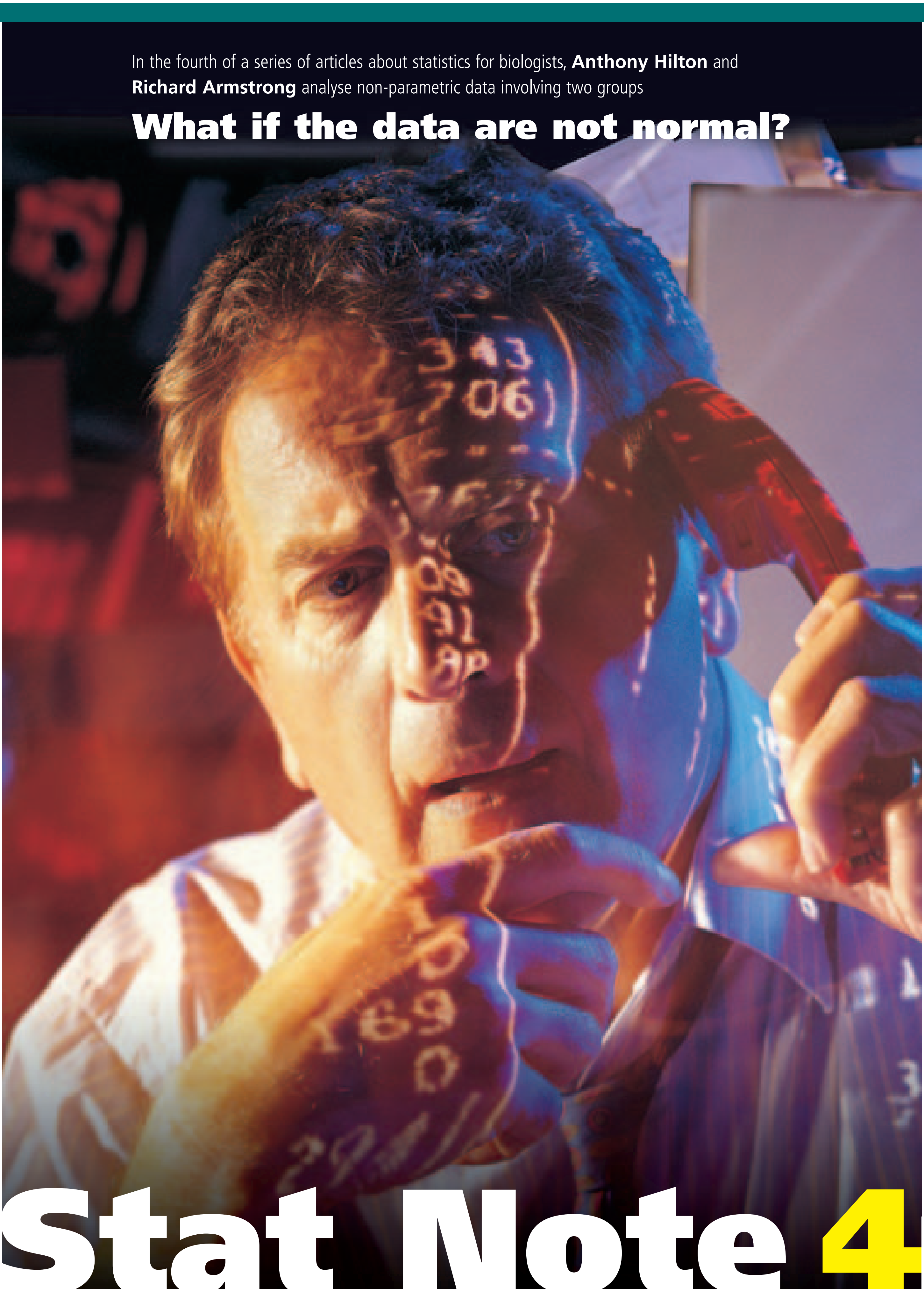
In the fourth of a series of articles about statistics for biologists, **Anthony Hilton** and **Richard Armstrong** analyse non-parametric data involving two groups

What if the data are not normal?



Stat Note 4

THE STATISTICAL tests described in previous Statnotes (see *Microbiologist* September and December 2005) make a number of assumptions about the experimental data.

The most important of these assumptions is that the quantity analysed, whether an individual measurement, treatment mean, or difference between two means, must be a parametric variable, i.e., a member of a normally distributed population. When this assumption is met, the 'z' and 't' distributions can be used to make statistical inferences from the data. In some circumstances, however, a variable may not be normally distributed and this Statnote is concerned with the analysis of non-parametric data involving two groups.

How do we know if the data are not normally distributed?

An investigator may know in advance from previous studies whether or not a quantity comes from a normal distribution. In other circumstances, data may have been collected to specifically test whether the data come from a normal distribution, a procedure that was described in Statnote 1 (*Microbiologist*, June 2005). In many experimental situations, however, there may be insufficient data available to carry out a test of normality and to obtain such data may be either too expensive or time-consuming. In situations such as these, the following points should be considered. First, many measurements in the biosciences made to at least three significant figures have a reasonable chance of being normally distributed. Second, the distribution of sample means taken from a population is more likely to be normal than the individual measurements and therefore, inferences about means are less susceptible to this problem. Third, moderate departures from normality do not significantly affect the validity of parametric tests. Consideration of these points may lead to the conclusion that despite some reservations, the data may not depart radically enough from normality to question the validity of a parametric analysis. In other circumstances, however, it may be clear that the data depart significantly from normality and a different approach required.

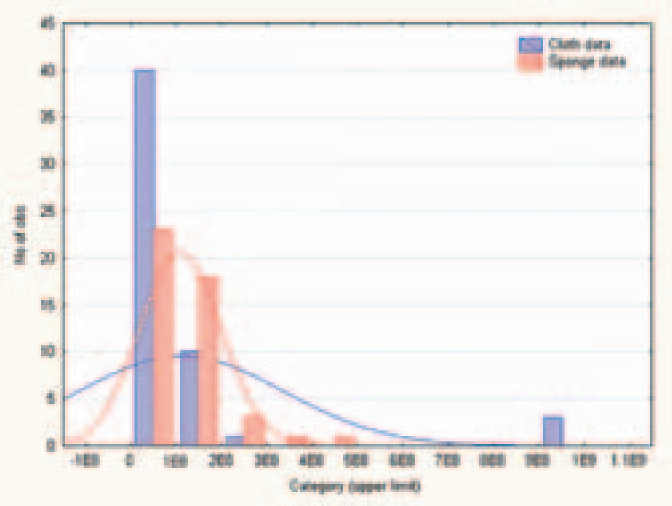


Fig. 1*. Frequency distribution of the bacterial counts on cloths and sponges. Fitted curves are those of the normal distribution.*

Deviations from a normal distribution

The two most common ways in which a distribution may deviate from normality are called *skew* and *kurtosis*. Most statistical software will provide tests of these properties and tables of significance (Table A20) of the relevant statistics are given by Snedecor and Cochran (1980). It is important to note that some distributions may deviate from normal in more complex ways and therefore, absence of skew and kurtosis does not guarantee that a distribution is normal. A skewed distribution is asymmetrical and the mean is displaced either to the left (positive skew) or to the right (negative skew). By contrast, distributions that exhibit kurtosis are either more 'flat-topped' (negative kurtosis) or have longer tails than normal (positive kurtosis). Fig. 1 shows the frequency distribution of the bacterial counts on 54 sponges and 46 cloths introduced in Statnote 1 (*Microbiologist*, June 2005). In both cases the distributions are clearly asymmetrical with the means located to the left of the histogram and therefore exhibit a degree of positive skew. As a result, the arithmetic mean is no longer a good description of the *central tendency* of such a distribution. There are two additional statistics that can be used to describe the central tendency of a skewed distribution. First, is the *mode*, the value of the variable 'x' with the highest frequency, i.e., the maximum point of the curve. Second, is the *median*, the middle value of 'x', i.e., if all the values of 'x' were listed in ascending or descending order, the median would be the middle value of the array. Little progress has been made in devising statistical tests based on the mode but there are tests, to be described later, that essentially test the differences between the medians of two groups.

An important property of non-normal distributions is that the standard deviation (SD) is no longer an accurate descriptor of the spread of a distribution with a given mean. Hence, 'z' and 't' tables cannot be used to predict the proportion of observations that fall a given distance from the mean. On reporting frequency distributions from large samples that are not normally distributed, investigators often quote the *percentiles* of the distribution, e.g., the 90% percentile of a distribution is the score such that 90% of the observations fall short of and 10% exceed the score (Snedecor and Cochran, 1980).

What is data transformation?

One method of analysing non-normal data is to convert or *transform* the original measurements so that they are expressed on a new scale that is more likely to be normally distributed than the original. The usual parametric 't' tests can then be carried out on the transformed values. There are three common circumstances in which such a transformation should be considered. First, if the data are percentages and especially if the majority of the observations are close to zero or 100%. Percentage data can be transformed to an *angular* or *arcsin* scale defined as follows:

Angular measurement = $\sin^{-1} \sqrt{\%/100}$

Statistical software will often provide this transformation or see Table X in Fisher and Yates (1963). Percentage data can be significantly skewed when the mean is small or large and consequently, the effect of the transformation is that percentages near 0% or 100% are spread out to a greater degree than those near the mean so as to increase their variance. A paired or unpaired ‘t’ test can then be carried out using the transformed values as described previously (*Microbiologist*, December 2005). Second, data that comprise small whole numbers or quantities assessed using a score that has a limited scale, e.g., if bacterial abundance was scored from 0 to 5, are unlikely to be normally distributed. In this case, a transformation to $\surd\ x$ (or $\surd\ x + 1$ if many zeroes are present) may make the scores more normally distributed. Third, the ‘t’ test described in Statnote 3 also assumes *homogeneity of variance*, i.e., that the degree of variability is similar for both groups of observations. It is not unusual, however, for results from a ‘control’ group to be more consistent than values from an experimentally treated group. In this case, a transformation of the original measurements to a logarithmic scale may equalise the variance and in addition, may also improve the degree of normality of the data.

Table 1. Comparison of the number of bacteria on 10 cloths and sponges (two independent groups, Mann-Whitney test)

Clothes (A)		Sponges (B)	
Count	Rank	Count	Rank
1.8×10^6	4	1.1×10^8	13
1.8×10^7	6	2.2×10^8	20
2.0×10^7	7	4.6×10^6	5
5.9×10^7	10	9.8×10^7	11.5
1.6×10^8	19	1.3×10^8	15.5
2.0×10^5	2	1.3×10^8	15.5
9.8×10^7	11.5	1.5×10^8	18
1.1×10^6	3	4.7×10^7	9
6.9×10^4	1	1.4×10^8	17
3.0×10^7	8	1.2×10^8	14

1. Add up the ranks for each group: $R_A = 71.5$, $R_B = 138.5$
2. $U_A = \{n_A(n_A + 1)/2 + (n_A n_B)\} – R_A = 83.5$ where n_A and n_B are the number of observations in each group
3. $U_B = \{n_B(n_B + 1)/2 + (n_A n_B)\} – R_B = 16.5$
4. The smaller U (in this case 16.5) is the test statistic
5. Lesser U must be ≤ Wilcoxon’s tabulated U for significant difference
6. For larger samples: $Z = (\phi\mu – T\phi –1/2)\sigma$ where $\sigma = \sqrt{}n_B\ \mu/6$ and $\mu = n_A(n_A + n_B + 1)/2$

How are non-parametric tests done?

An alternative approach to transformation in the analysis of non-normal data is to use a non-parametric test. As an illustration, we return to the scenario described in Statnote 1 (*Microbiologist*, June 2005). To recapitulate, given the intrinsic structural and compositional differences between cloths and sponges, a study was envisaged to investigate if one material provided a more favourable environment for bacterial survival than the other. A total of 54 ‘in-use’ dishcloths and 46 sponges were collected from domestic kitchens and the aerobic colony count of each determined in the laboratory. The frequency distributions of the counts from both materials are shown in Fig. 1. In Statnote 1, these distributions were tested for normality and it was concluded that the cloth data exhibited a marked deviation from normal whereas the sponge data were closer to a normal distribution. However, it may be prudent to conclude that the data as a whole do not conform closely enough to a normal distribution to use the parametric ‘t’ tests described in Statnote 3. An alternative approach is to use a *distribution-free or non-parametric test*. These tests can be used regardless of the shape of the underlying distribution as long as the samples being compared can be assumed to come from distributions of the same general shape.

The Mann-Whitney U-Test (for unpaired data)

To illustrate this test and to simplify the calculations we will use data from a sample of 10 cloths and 10 sponges only. The Mann-Whitney U-test can be carried out on two independent groups of data (A,B) and is the non-parametric equivalent of the unpaired ‘t’ test (Statnote 3). Although most statistical software will carry out this test, it is still useful to understand its ‘mechanics’ (Table 1). First, ranks 1, 2, 3, ... are assigned to the whole set of observations, regardless of group. A rank of 1 is given to the lowest count, 2 to the next lowest etc. with repeated values, called ‘ties’, given the mean of the ranks within that run. The ranks of each group are then added together separately (R_A, R_B). The quantities U_A and U_B are then calculated as shown in Table 1. Whichever is the smaller of U_A and U_B, is taken to the table of Wilcoxon’s U to judge the significance of the difference between cloths and sponges (Snedecor and Cochran, 1980; Table A10). The lesser U has to be equal to or *less* than the tabulated value for significance, i.e., low values of U indicate a significant difference between the groups. In the present example, a value of $U = 16.5$ was obtained which is less than the value tabulated at $P = 0.05$. Hence, there is evidence that the sponges harbour considerably more bacteria than the cloths. For larger samples, outside the range of the statistical table, the data may approach a normal distribution more closely and a value of Z can be calculated (Table 1), the statistic being referred to tables of the normal distribution.

The Wilcoxon signed rank test (for paired data)

If the data in the two groups are paired (*Microbiologist*, December 2005), then the appropriate non-parametric test is the Wilcoxon signed rank test. To illustrate this test (Table 2), we collected data on the number of bacteria on a single pair of cloths and sponges on 10 separate occasions. Hence, we do not have two independent samples as in the previous example. In this case, there is a link between a particular cloth and sponge in that the data for each pair were collected on a specific occasion. Essentially, the data are subtracted for each pair of observations (A - B). Omitting zero differences, ranks (r) are applied to all of the remaining values of A - B regardless of whether the difference is positive or negative. If ties occur between positive and negative columns, the ranks are amended in any such run of ties to the mean rank within the run. The positive and negative signs are restored to the ranks and the positive and negative ranks added up. R is the smaller of the two sums of ranks and is taken to the table of the Wilcoxon signed rank statistic T to obtain a P-value (Snedecor and Cochran 1980;

Table A9). The value of R has to be equal to or LESS than the value of T in the P = 0.05 column to demonstrate a significant difference between the two groups. In this case, our value of R = 1 was less than the tabulated value indicating that sponges harbour more bacteria than the cloths. With larger numbers of observations, a value of Z can be calculated and referred to tables of the normal distribution.

Comparison of the parametric and non-parametric tests

It is reasonable to ask what is the relative sensitivity of parametric and non-parametric tests and what happens if they are used incorrectly? If a t-test is used on non-normal data, the significance probabilities are changed and the sensitivity or power of the test is altered and this can result in erroneous conclusions especially if treatment effects are of borderline significance. With non-parametric tests, the significance levels remain the same for any continuous distribution with the exception that they are affected by the number of zeros and tied values in the Wilcoxon signed rank test (Snedecor and Cochran 1980). With large normal samples, the efficiency of the non-parametric tests is about 95% compared with the t-test. With non-normal data from a continuous distribution, however, the efficiency of the non-parametric tests relative to 't' never falls below 86% in large samples and may be greater than 100% for distributions that are highly skewed.

Conclusions

When testing the difference between two groups, if previous data indicate non-normality, then either transform the data if they comprise percentages, integers or scores or use a non-parametric test. If there is uncertainty whether the data are normally distributed, then deviations from normality are likely to be small if the data are measurements to three significant figures. Unless there is clear evidence that the distribution is non-normal, it is more efficient to use the conventional t-tests. It is poor statistical practice to carry out both the parametric and non-parametric tests on a set of data and then choose the result that is most convenient to the investigator!

Table 2. Comparison of bacteria on pairs of cloths and sponges sampled on 10 occasions (two dependent groups, Wilcoxon signed rank test)

Occasion	Cloth (A)	Sponge (B)	A - B	Rank
1	1 x 10^4	4.6 x 10^6	-4.5 x 10^6	-2
2	3.3 x 10^7	9.8 x 10^7	-6.5 x 10^7	-6
3	5.7 x 10^7	1.3 x 10^8	-7.3 x 10^7	-7
4	1.9 x 10^7	1.3 x 10^8	1.11 x 10^8	-9
5	1.2 x 10^4	6.0 x 10^2	+1.1 x 10^4	+1
6	8.8 x 10^2	4.7 x 10^7	-4.7 x 10^7	-5
7	2.6 x 10^6	1.4 x 10^8	-1.14 x 10^7	-3
8	3.3 x 10^7	1.2 x 10^8	-8.7 x 10^7	-8
9	8.7 x 10^6	2.1 x 10^8	-2.0 x 10^8	-10
10	7.6 x 10^7	1.1 x 10^8	-3.4 x 10^7	-4

1. Subtract each pair of counts A - B
2. Assign ranks (r) to differences ignoring the sign of the difference
3. Restore the signs and add up the positive and negative ranks
4. Compare the lesser R (in this case +R = 1) with the tabulated Wilcoxon's signed rank statistic T, $R \leq T$ for significance
5. For larger samples $Z = (\mu - T - 1/2)\sigma$ where T is the smaller rank sum and $\sigma = \sqrt{}(2n + 1)\mu/6$ where n = number of pairs and $\mu = n(n + 1)/4$

References

- Fisher RA and Yates F (1963) Statistical tables. Longman, London.
- Hilton A & Armstrong RA (2005a) Are the data normal? *Microbiologist* June 2005. **6:** 2, pp34-35
- Hilton A & Armstrong RA (2005b) Describing the normal. *Microbiologist* September 2005. **6:** 3, pp30-33.
- Hilton A & Armstrong RA (2005c) Testing the difference between two groups. *Microbiologist* December 2005. **6:** 4, pp30-33.
- Snedecor GW and Cochran WG (1980) Statistical methods, 7th Ed. Iowa State University Press, Ames Iowa. Chapters **5,6.**

Dr Anthony* Hilton and Dr Richard Armstrong**
*Pharmaceutical Sciences and **Vision Sciences, Aston University, Birmingham, UK