



Stat Note 1

In the first of a series of articles about statistics for biologists, **Anthony Hilton** and **Richard Armstrong** talk about the distribution of data — are they normal?

Is the data normal? Chi Squared and the Kolmogorov-Smirnov test

The first stage of any statistical analysis is to determine the degree to which, if at all, the data depart from normality. Having established the distribution of the data, parametric or non-

parametric statistical tests may be applied as appropriate. Microbiological data, especially from environmental sources, may have very large counts and associated standard deviations, and are unlikely to be normally distributed. In this StatNote we describe the application of two tests of normality.

The Scenario

The domestic kitchen is increasingly recognised as an important reservoir of pathogenic microorganisms, with dishcloths and sponges providing an ideal environment for their growth, survival and dissemination. Given the intrinsic structural and compositional differences

between these two material types, a study was envisaged to investigate if one provided a more favourable environment for bacterial survival than the other; the hypothesis being that there would be a quantitative difference between the number of microorganisms recovered from dishcloths compared to sponges. A total

Table 1. Observed and expected frequencies for the sponge data. (Tests of normality: chi-square (all categories) = 38.99 ($P < 0.01$); chi-square (adjusted for expected values < 5) = 4.80 ($P > 0.05$); Kolmogorov-Smirnov (KS) test = 0.0894 ($P > 0.05$))

| Category (upper limits) | Observed F | Expected F | O — E |
|-------------------------|------------|------------|--------|
| <=5000000 | 6 | 5.78 | 0.22 |
| 60000000 | 11 | 7.91 | 3.09 |
| 115000000 | 8 | 10.864 | -2.86 |
| 170000000 | 14 | 10.33 | 3.67 |
| 225000000 | 4 | 6.795 | -2.795 |
| 280000000 | 1 | 3.09 | -2.09 |
| 335000000 | 0 | 0.98 | -0.97 |
| 390000000 | 1 | 0.21 | 0.79 |
| 445000000 | 1 | 0.03 | 0.967 |
| < Infinity | 0 | 0.003 | -0.003 |

O = Observed frequency, E = Expected frequency

of 54 'in-use' dishcloths and 46 sponges were collected from domestic kitchens and the aerobic colony count of each determined in the laboratory.

How is the Test Done?

To fit the normal distribution, the variable (aerobic colony count on 46 sponges) is first divided into frequency classes describing the range of the variable in

the population. In the present case, ten classes were used for the sponge data (Table 1). The limits of these classes are then converted so that they are members of the standard normal distribution. To carry out this calculation, the mean and standard deviation of the observations are first calculated. The sample mean is then subtracted from each class limit and divided by the standard deviation, which

converts the original measurements to those of the standard normal distribution. Tables of the standard normal distribution are then used to determine the expected number of observations that should fall into each class if the data are normally distributed. The observed and expected values (Fig. 1) are then compared using either a chi-square 'goodness of fit' or a Kolmogorov-Smirnov (KS) test. This statistical analysis is available in many of the popular statistical analysis software packages such as Prism, Statview, SPSS or Statistica.

How do you Interpret the results?

The chi-square (χ^2) test (7DF, $\chi^2 = 38.99$, $P < 0.01$) for the sponge data is significant at the 1% level of probability suggesting that the distribution deviates significantly from the normal distribution. The Kolmogorov-Smirnov test (KS=0.089, $P > 0.05$), however, is not significant, a not uncommon result since this test is less sensitive than chi-square and only indicates gross deviations from the normal distribution (Pollard, 1977). The chi-square test also has limitations in this context as it is greatly affected by how many categories are selected to define the variable and how these categories are divided up. In addition, if a number of the categories have expected numbers of observations below five, adjacent categories should be combined until their expected values are greater than five. If this procedure is carried out using the present data, the value of chi-square is not significant. In cases like this, the general shape of the observed distribution is probably the best method of judging normality. Although this distribution (Fig. 1) exhibits a degree of skew, the deviations from normal

(supported by the KS test) suggest the deviations are not significant enough to warrant using a non-parametric test. However, a similar analysis carried out on the cloth data resulted in considerable deviations from a normal distribution on both tests ($\chi^2 = 3007.78$, $P < 0.001$; KS = 0.28, $P < 0.01$). Hence, in an analysis to compare the cloth and sponge data it may be prudent not to use a parametric unpaired t-test. In this case, we have two ways in which to proceed to compare the two groups: (1) transform the data to normality thus allowing the application of the parametric 'unpaired' t-test, or (2) employ the non-parametric equivalent, the Mann-Whitney test. Both procedures will be illustrated in future StatNotes.

Summary

Testing whether an observed distribution of observations deviates from normality is a common type of statistical test available in statistics software. Most software offer two ways of judging whether there are significant deviations of the observed from the expected distributions: chi-square and the KS test. These tests have different sensitivities and problems and often give conflicting results. The results of these tests together with observations of the shape of the observed distribution should be used to judge normality.

Reference

■ Pollard J H (1977) A Handbook of Numerical and Statistical techniques. Cambridge University Press, Cambridge

Dr Richard Armstrong and Dr Anthony Hilton
Life and Health Sciences, Aston University

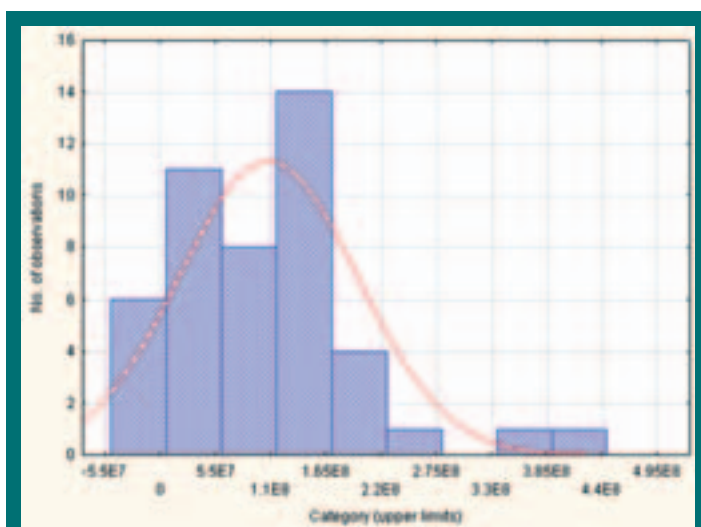


Figure 1. Histogram illustrating the observed distribution of values for the sponge data and the predicted normal distribution (continuous line). The chi-square goodness of fit and Kolmogorov-Smirnov tests test the difference between the observed and expected frequencies.