Analytical Methods

RSC Publishing

AMC TECHNICAL BRIEFS

An introduction to non-parametric statistics

Analytical Methods Committee, AMCTB No 57

Cite this: *Anal. Methods*, 2013, **5**, 5373

Received 4th September 2013

DOI: 10.1039/c3ay90070c

www.rsc.org/methods

Non-parametric statistical methods, which make fewer assumptions about population error distributions, have perhaps been unjustly neglected in the analytical sciences. A major advantage is that some of them are so simple that they can be used "at the bench."

Parameters or no parameters?

Analytical scientists generally make replicate measurements and treat them as a random sample, from which estimates are made of the properties of the (hypothetically infinite) population of measurements. The population mean, confidence limits *etc.* are usually calculated using the assumption that the underlying distribution is normal (Gaussian), with mean μ and variance σ^2, *i.e.* it can be summarised as $N(\mu,\sigma^2)$. The two terms μ and σ are the parameters of the distribution. Similarly a binomial distribution is described as $B(n, p)$, where the parameters n and p are respectively the total number of measurements and the probability of one of the two possible outcomes.

This parameter-based approach to data handling is not essential, and may not always be appropriate. Sometimes it is known that a population distribution is not normal or even close to it, so deductions made on the assumption of normality might be unreliable. This is particularly true in cases where the same measurements are made on similar but non-identical sample materials of natural origin. The antibody levels in blood plasma samples from different human subjects are roughly log-normally distributed, with the addition of some subjects with exceptionally high levels in various disease states. Methods that do not make assumptions about the form of the population distribution are called non-parametric or distribution-free methods. In applying them the familiar approach to significance testing is still used. We set up a null hypothesis H_0 and find the probability of obtaining the actual or more extreme results if H_0 is true: if this probability is very low H_0 is rejected. But their simplicity makes non-parametric methods attractive even in situations where more familiar tests such as the *t*-test might otherwise be applied, as the examples below will show.

Some simple examples

Suppose that an analytical reagent is stated to have a purity of 99.5%, and that successive batches are found to have purity levels of 99.2%, 99.8%, 98.9%, 99.4%, 99.1%, 99.3%, and 99.0%. Is there evidence that the purity of the material is lower than it should be? Such results are unlikely to come from a normal population (after all, the maximum possible purity is 100%) so a *t*-test or other parametric approach could well be unsafe. A key statistic here is the median: the null hypothesis is that the data come from a population with a median purity level of 99.5%. To carry out the test we simply subtract this median from each of the experimental results, and note the sign of the result. This gives six minus signs and one positive sign, *i.e.* six of the seven results lie below the median. (Any result that equals the hypothetical median is ignored completely). The probability of getting six (or more) minus signs out of seven is provided by the binomial theorem, but the values are provided in statistical tables, and can be memorised if we always make the same number of measurements. Here the probability of getting 6 or more minus signs is 0.0625, a little higher than the probability level commonly used in significance testing ($p = 0.05$), so we retain the null hypothesis that the results could come from a population with a median purity of 99.5%. As always we have not proved that they do come from such a population: we have failed to disprove it. Note that this is a one-tailed test, as the question is whether the purity is lower than it should be. With seven measurements the null hypothesis would only be rejected at the $p = 0.05$ level if all seven results give minus signs when compared with the median value: this outcome has a probability of only $(1/2)^7 = 1/128$. This method is called the sign test, and it can be extended to other situations, such as comparing two sets of paired results, or studying a possible trend in a sequence of results.

Another simple test with many applications is called Tukey's Quick Test (after John W. Tukey, a major figure in non-parametric statistics and initial data analysis) or the Tail Count Test, the latter being a good description of its operation. It is used to compare two independent data sets, which need not be of the same size. Suppose we obtain six values of the level of atmospheric NO_x ($\mu g\ m^{-3}$) at a roadside site: 128, 121, 117, 125, 131 and 119. At a nearby off-road site we make six more measurements using the same analytical method, obtaining the results 120, 108, 109, 112, 114 and 110 $\mu g\ m^{-3}$. Is there any evidence that the NO_x level is lower at the second site than at the first? These two sets of results could be compared using a (one-tailed) t-test, but the Tukey approach is simpler. We simply count the number of results in the first data set that are higher than all the values in the second set (there are 4 of them), and the number of values in the second set that are lower than all those in the first set (5 of them). If either of these counts is zero, the test ends at once with the null hypothesis (here, that moving away from the road does not affect the NO_x level) being accepted. Otherwise the two counts are added together to provide the test statistic T ($= 9$ here), and this is compared with the critical value. For a one-tailed test at $p = 0.05$, T must be greater than or equal to 6 if H_0 is to be rejected. So H_0 can be rejected here; the NO_x level at the off-road site does seem to be lower. The merit of the Tukey method is that if the total number of measurements is no more than $\sim$20, and if the two sample sizes are not greatly different (conditions often met in analytical practice), the critical T values are independent of sample size! For the rejection of the null hypothesis in a one-tailed test the value of T must be $\geq$6, 7, 10 and 14 at $p = 0.05$, 0.025, 0.005, and 0.0005 respectively. For a two-tailed test the corresponding critical values of T are 7, 8, 11 and 15 respectively. This remarkable feature of the method means that it can be carried out using mental arithmetic only.

What's not to like?

Many non-parametric methods have been developed, including tests analogous to the familiar t- and F-tests, analysis of variance, and calibration and regression methods, but despite their practical merits only a few have found favour in the analytical sciences. One possible reason for this is that most non-parametric methods need a sample of at least 6 measurements. Another reason is the growing popularity of robust methods (AMCTB 6, 50), which are well suited to the common situation where the error distribution is unimodal but not very different from Gaussian. Furthermore it is evident that in the two examples above the full numerical content of the data is not used. In the sign test only the signs of the differences are counted, not their magnitude; and in the Tukey method the test statistic is again a count rather than an exact reflection of the numerical results. We might thus expect that non-parametric methods would be poorer than methods such as the t-test at identifying significant differences in situations where the error distribution is Gaussian or nearly so. That is, non-parametric methods may have less power, *i.e.* less capacity to reject a false null hypothesis, than parametric ones in situations where the latter can legitimately be applied. If a one-tailed t-test is applied to the first example above, the outcome suggests that the mean of the seven purity values is (at $p = 0.05$) significantly lower than the quoted value of 99.5%. So the sign test and the t-test give contradictory results in this case, albeit in a situation where the t-test is of uncertain legitimacy. In the second example, where it would be reasonable to assume a Gaussian error distribution, the Tukey test suggests that the null hypothesis can be rejected at $p = 0.025$, but not at $p = 0.005$, whereas a one-tailed t-test gives the probability of the data under the null hypothesis as only 0.0016. Again it seems that the non-parametric method is the less powerful of the two. In comparing two or more tests with the same purpose we use the concept of efficiency. The efficiencies of two tests, 1 and 2, are compared using the ratio of the sample sizes, n_1/n_2, necessary to detect a defined but small departure from H_0. The efficiency of the sign test relative to the t-test when the latter is valid is only about 2/3. But in other cases where the t-test is not appropriate the sign test can be up to twice as efficient. The sign test is virtually assumption-free (except that the data must come from a continuous distribution), but other non-parametric methods do make additional assumptions: for example the Tukey test assumes a simple shift of location between two samples from otherwise identical distributions. So as with other statistical methods, non-parametric approaches must be used with care and discrimination, but their practical merits remain formidable.

Bibliography

1 P. Sprent and N. C. Smeeton, *Applied Nonparametric Statistical Methods*, 4/e, Chapman and Hall, 2007.

2 M. Hollander and D. A. Wolfe, *Nonparametric Statistical Methods*, Wiley, 1999 (a new edition is expected in December 2013).

CPD Certification I certify that I have studied this document as a contribution to Continuing Professional Development.

Name……………………………………………………

Signature…………………………….Date………….

Name of supervisor…………………………………..

Signature…………………………….Date…………

