QUANTITATIVE METHODS
BIOLOGY FINAL HONOUR SCHOOL
NON-PARAMETRIC TESTS

This booklet contains lecture notes for the nonparametric work in the QM course.

This booklet may be online at http://users.ox.ac.uk/~grafen/QMnotes/index.html.

ALAN GRAFEN

Nonparametric Tests I

(Watt discusses the Mann-Whitney test in section 9.2)

Nonparametric tests form an important topic in statistics for biologists, for two quite different reasons. The first is that they are frequently used in the literature. They form a parallel set of tests, so that with many simple datasets the question will arise whether to use a parametric (i.e. GLM or GLM-like test) or a nonparametric test. You need to know what the commonest nonparametric tests are, and when they are applicable. Then you will recognise those datasets where the choice arises. You also need to know how to choose between the two types of test.

The second reason nonparametric tests are useful for biologists is that although based on the same fundamental statistical logic as parametric tests, they are much simpler and their workings are more transparent. Learning nonparametric tests is therefore a chance to refresh your understanding of the basic logic of statistical inference.

This week we look at what nonparametric tests Minitab will do, and their GLM counterparts, and we discuss when nonparametric tests are preferable. Next week we will look carefully at the logic of some specific tests.

You are recommended to read the section from Watt given above, so that when you read the rest of this handout you have a specific nonparametric test well established in your mind.

What is a nonparametric test?

When you were introduced to any parametric test, there came a stage when you were asked to assume that some distribution was Normal. Normality of some distribution is an assumption of parametric tests. You might reasonably rebel, and decide you want to do a test that does not depend on Normality. In some tests, there is also a further assumption, for example linear regression assumes that x and y are related linearly and the test is invalid if

they are not. Again you might rebel and demand a test that worked whether or not x and y are linearly related. The set of tests loosely called nonparametric are tests based on minimal assumptions, and so do their best to be valid in as wide a range of circumstances as possible.

There is a threefold cost for 'saving on assumptions'. First, only simple kinds of tests can be performed without making assumptions about shapes of distributions and the nature of relationships (Normality and linearity in simple cases). In the table of equivalences below, notice that all the GLM equivalents have one or two x-variables, never three. You can have one continuous x-variable, or one or two categorical variables, but no more. Mixing continuous and categorical variables is not possible, at least within the tests Minitab will perform. So nonparametric tests are available only for a restricted types of dataset, though this does include the most common ones.

The nonparametric tests Minitab provides are the common ones, and they are quite restricted in the types of dataset they will handle. Statisticians do suggest nonparametric tests that will apply to wider types of dataset. However, these tests are of interest to biologists only once they become accepted as sensible by statisticians generally, and available in statistical packages. You probably could find a nonparametric test that would analyse a four-way analysis of variance with covariate, but unless and until it is accepted and available it is probably better *not* to find it!

Second, GLM always provides hypothesis tests (p-values) and estimation (confidence intervals, for example). Nonparametric tests always provide hypothesis tests, but only sometimes estimation.

A table of nonparametric tests available using Minitab, and their parametric equivalents, with the associated Minitab commands.

Nonparametric test(s)	Equivalent parametric tests
Sign Test (STEST, SINTERVAL), Wilcoxon Signed Ranks Test (WTEST, WINTERVAL)	One-sample t-test (TTEST, TINT or GLM)
Mann-Whitney Test (MANN-WHITNEY)	Two-sample t-test (TWOS, TWOT or GLM)
Spearman Rank Correlation (RANK each variable and use CORRELATE on the ranks for the value of the coefficient. Use a GLM between the ranks to get the p-value.)	Pearson Product-Moment Correlation, Bivariate Regression (CORRELATE or GLM with one continuous x-variable)
Kruskal-Wallis Test (KRUSKAL-WALLIS)	One-way Analysis of Variance (ANOVA or GLM)
Friedman Test (FRIEDMAN)	One-way Blocked Analysis of Variance (ANOVA or GLM)

Third, GLM tests are *efficient*. This is a technical term meaning they make the best possible use of the information. So, provided the assumptions of the GLM are true, the GLM is more efficient than any other test, including the corresponding nonparametric test. This loss of efficiency can be measured in the loss of power. That is, suppose the null hypothesis is false. The GLM will reject the null hypothesis more frequently than the nonparametric test, which therefore 'wastes information'. Of course if the assumptions of the GLM are not upheld, then the GLM may well be quite wrong, while the nonparametric test remains valid.

One further contrast is that the GLM framework combines a large group of parametric tests into a single scheme, while nonparametric tests are disparate with no 'grand scheme' to unite them.

Essential point. Nonparametric tests rely on fewer assumptions than GLM, but this does not mean they make no assumptions at all. Specifically, the assumption of independence is just as important for nonparametric as for parametric tests. Nonparametrics are not cure-alls for failures of assumptions.

Rank tests, distribution free tests and nonparametric tests

These are three names for three slightly different but largely overlapping sets of statistical tests. Rank tests are so called because they involve first ranking the data. Distribution free tests are so called because they make no assumptions about the distribution of the population from which the samples are drawn. Nonparametric tests are so called because the hypotheses tested do not involve the parameters of the population distributions (for example that the mean of population one equals the mean of population two). Instead the null hypothesis is simply that the two population distributions are the same. Notice that this implies that the means are the same, the median is the same, the variance is the same and so on. So the point of being 'nonparametric' is not to test a weaker hypothesis, it is to avoid assuming anything more than necessary. Parameters are always part of a model that restricts the range of applicability of the test, by specifying the shape of a relationship or the shape of a distribution.

A further property that tests can have is 'scale-invariance'. A test is scale-invariant if it gives exactly the same answer when you transform the variables, for example take logarithms. When we have an arbitrary scale (say of subjective preferences), scale-invariance can be an advantage.

Of the tests shown in the table, the Mann-Whitney test is a rank test, and is distribution free and scale invariant. The sign test is distribution free, and scale invariant - but it is not a rank test because the data are never ranked. The Wilcoxon signed ranks test is not a rank test because the original data are not ranked, they are subtracted from each other. It is not scale invariant because logging the original data will affect the result (try it!). It is distribution free, because it is valid whatever the shape of the original distribution.

Which type of test to use

Whenever we apply a test, we should check the assumptions on which it is based, and these checks (called 'model criticism') were discussed in Chapter 9 of Grafen and Hails. Given a dataset and a hypothesis that can be tested with both a parametric and a nonparametric method, an obvious question is: are the assumptions of the parametric test upheld?

If we are sure the assumptions are upheld, then we should use the more efficient parametric test. Because the assumptions refer to the population and not to the sample, it is hard to be sure the assumptions are upheld unless the data are created artificially. The paradox of Normality testing is that checking Normality of data is often unsatisfactory. With small datasets, we have too little data to decide whether a distribution is Normal or not. With large datasets, the Central Limit Theorem assures us that Normality is not very important! (You encountered the Central Limit Theorem in Mods - it says that the sampling distribution of the mean attains Normality as the sample size increases.) Added to which, making a formal statistical test for Normality of data is not simple even in principle. So the decision between the methods is not as clearcut as might be thought.

An extra complication is that if the assumptions of a parametric test do fail, then nonparametric testing is not the only option. Specifically, transforming the data (logging y and x; or square-rooting y) is another important possibility that was discussed in Chapter 9 of Grafen and Hails.

The following guidelines represent a sensible approach.

(1) With a small dataset and a simple question, do both types of test. Usually they will give the same result (i.e. both significant or both non-significant) and you need think no more about it. Where the p-values differ a little, its likely the parametric test is reasonable as the difference between the results can be put down to a difference in efficiency. Where they differ a lot, you would be relying strongly on untested assumptions to accept the parametric over the nonparametric test.

(2) If you need estimation and not just hypothesis testing, then you should probably be doing parametric tests.

(3) With a large dataset, in which conformity to the assumptions can be checked, it is likely that other ways of avoiding failed assumptions (e.g. transformations and polynomial regressions) will be preferable to nonparametric tests.

(4) When a number of tests are being conducted, for example as part of a project, consistency is desirable. Don't switch from one type to another, as the type of test used might then explain some of the results. How can consistency be obtained if we need parametric methods for some of the tests, and yet the datasets are all small so we would rather do nonparametrics where possible? The answer lies in the links between the datasets. By considering a number of small datasets at the same time, we may be able to make as convincing a case for Normality as if we had a single larger dataset. Or we may be able to persuade ourselves that a transformation applied to a variable in all datasets brings about Normality, in a way that would be impossible with a single small dataset. Then we can apply parametric tests throughout, relying on the links between datasets for assurance about the assumptions.

(5) Don't do over-simple tests just to be able to use nonparametric methods. We saw earlier in the course how important controlling for third variables can be. This is often impossible with nonparametric tests. Controlling for the right things is a substantive point that dominates the more technical problem of meeting assumptions. So if you need to control for other variables, find another way of solving failed assumptions.

In conclusion, nonparametric test are important in small datasets where simple hypotheses are to be tested, when the dataset is not linked to other tests where more complicated methods are required. The meaning of 'small' is not hard and fast. The essential question is whether we can or can't check assumptions adequately from the data. If we can't, the dataset is small. If we can, the dataset is large, and we can safely apply other methods of solving broken assumptions. Two datasets are 'linked' if we can assume that the same variable is Normal in both or not Normal in both, as this allows us to check the assumptions of Normality by considering both datasets at the same time.

Nonparametric Tests II

This week we interpret nonparametric tests as randomisation tests. There are two purposes. First, it is possible to understand how nonparametric tests work from first principles in this way. With GLMs, the calculation of a p-value from an F-ratio is a bit mysterious at the level of this course. But nonparametric tests can be understood completely. The second purpose is that many modern statistical techniques use randomisation, and it is useful for you to have an idea of how randomisation tests work.

Background: Nonparametric tests as randomisation tests

The sign test

The sign test applies to a set of values whose median, according to the null hypothesis, is zero. Usually, each value will be the difference between two observations (for example, a measure of illness before and after treatment; a measure of size for the female and male in a pair of starlings). If the median really is zero, then the number of positive differences should on average be half the total number of differences. Indeed, the number of positive differences should have a binomial distribution, with n=total number of differences and p=0.5.

To find whether the observed number of positive differences is different from what would be expected just by chance, we could simulate the tossing of coins, and use a large number of such observations to estimate how likely the actual value is to arise by chance. If there are ten differences, we would toss a set of ten coins repeatedly, recording each time how many out of ten were heads.

But in this case, we know the exact distribution, and so we can appeal to the binomial distribution directly.

The Wilcoxon signed ranks test

The sign test discussed above may seem rather wasteful. A difference that is just negative counts the same as a really large negative difference. Is there any way to give greater weight to larger differences, without going so far as parametric methods? There is. The procedure is to work out the rank of each difference irrespective of sign. So the smallest difference (positive or negative) is given rank 1. The next smallest (positive or negative) is given rank 2, and so on. Instead of just counting how many negative differences there are, we weight each difference by its rank. We therefore add up all the ranks of the negative differences. This is to be the test statistic.

But how can we work out how unusual our observed value of the test statistic is? We turn to a randomisation procedure, based on the supposition that the distribution of each difference is symmetrical. On the null hypothesis, each difference, no matter what its rank, is therefore equally likely to be positive or negative. If there are ten differences, a suitable randomisation procedure is to mark ten coins 1 to 10, and toss each of them. We add up the scores of all the coins that fall tails. This parallels the adding of ranks used to get the test statistic. The distribution of the test statistic obtained from repetition of the randomisation procedure can then be compared with the actual value.

If the observed value falls in the top 2.5%, or the bottom 2.5%, then we reject the null hypothesis at the 5% level on a two-tailed test.

The Mann-Whitney test

The Mann-Whitney test tests for a difference between two groups of observations, and looks specifically for a difference in *location*, that is whether one group tends to be have larger values than the other. If the groups are two samples from the same underlying distribution, then any differences between the samples is due to chance. We could think of drawing each observation from the one underlying distribution, and then afterwards allocating them at random to the two groups. This would be a way of working out, for a given set of observations, how likely differences of any magnitude between two samples are to arise by chance.

The randomisation test then involves taking as fixed the set of values observed, and how many belong to each group; but randomising which values belong to which group. The test statistic used to assess the location of a group is the sum of the ranks. If there are 5 in each group, then a complete separation of the two groups gives a summed rank of 15 for the higher group and a summed rank of 40 for the lower. If there is no systematic difference between the groups, then on average the summed rank of a group is 27.5.

In the randomisation test, we create random divisions of the datapoints into two groups, and record the mean rank of group A. We do this lots of times, and build up a distribution that shows how often the mean rank would take different values if there were in fact no real differences between the groups. The single observed mean rank of group A is then compared to this null distribution. If it is in the top 2.5% or the bottom 2.5% of the null distribution, then we reject the null hypothesis at the 5% level on a two-tailed test.

General remarks on randomisation tests

The general idea of randomisation tests is a very powerful one, and it can be invoked to cope with difficult statistical situations in which the assumptions of standard methods are not met. There are two elements: the test statistic and the randomisation procedure. The validity of the test is guaranteed by the randomisation procedure alone, and so there is wide flexibility in choosing a test statistic.

The randomisation procedure divides the information in the dataset into two sets: the fixed and the unfixed. The table lists what is considered fixed and unfixed in each of the three tests described above. In statistical terms, we condition on the fixed information and conduct the test on the unfixed information. Any information that we do not want to use should be included in the fixed information. Conversely, any information we do want to use must be included in the unfixed information.

Test	Fixed Information	Unfixed Information	Test Statistic
Sign Test	The absolute values of the differences	Whether each difference is + or -	Number of +
Wilcoxon signed ranks	The absolute values of the differences	Whether each difference is + or -	Sum of ranks of negative differences
Mann-Whitney, or Rank Sum	The set of all values; the number in each group.	Which precise values are in which group	Sum of the ranks in the first group