© Grace Scientific Publishing

Journal of
Statistical Theory and Practice
Volume 3, No. 3, September 2009

Partially Parametric Testing

J. C. W. Rayner, *School of Mathematical and Physical Sciences, University of Newcastle, NSW 2308, Australia.*
Email: John.Rayner@newcastle.edu.au

A. M. Carolan, *Westpac Banking Corporation, Sydney, Australia.*
Email: acarolan@westpac.com.au

Received: February 27, 2008 *Revised:* November 10, 2009

Abstract

Suppose a smooth test of goodness of fit has been applied to assess the validity of a parametric analysis, such as an analysis of variance. If the smooth test rejects the distributional assumption, the original parametric model can be replaced by the order k smooth alternative that is the basis of the smooth test. Here we demonstrate that basing the analysis on such an alternative, when it is consistent with the data, may result not only in a valid analysis, but also in a test with greater power than for the original parametric analysis. Several examples are given, including the one sample t-test, the one-way analysis of variance and randomised complete block designs.

AMS Subject Classification: 62F03; 62G07; 62G10.

Key-words: Smooth alternative distributions; Smooth tests of goodness of fit; Wald test.

1. The concept of partially parametric testing

When confronted with parametric data analysis a smooth test of goodness of fit may be applied to assess the parametric assumptions. If the null hypothesis that the parametric model is consistent with the data is accepted, the analysis can proceed. Typically rejection means a nonparametric analysis must be applied, or a completely different technique employed. Here we propose a different and more immediate approach: that inference be based on a model determined after the model assessment. The principal point of the paper is that by doing so, not only is the inference valid, it can be very much more powerful than both the original parametric analysis and competitor nonparametric analyses.

In order to smoothly assess if the random sample $X_1, \ldots, X_n$ comes from a distribution with probability density function $f(x;\beta)$ where β is a $q \times 1$ vector of nuisance parameters,

* 1559-8608/09-3/$5 + $1pp – see inside front cover

$f(x;\beta)$ is nested in an alternative of order k:

$$g_k(x;\theta,\beta) = C(\theta,\beta)\exp\left\{\sum_{i=1}^{k}\theta_i h_i(x;\beta)\right\} f(x;\beta).$$

We call this a *Neyman* model. Here $C(\theta,\beta)$ is a normalising constant that we assume exists and $\{h_i(x;\beta)\}$ is a set of orthonormal functions on $f(x;\beta)$ with $h_0(x;\beta) = 1$ for all x. We test for $f(x;\beta)$ by testing if all the θ_i are consistent with zero: that is, by testing $H_0 : \theta = (\theta_i) = 0$ against $K : \theta \neq 0$.

Rayner et al. (2009a, b) describes the derivation of both smooth and generalised smooth tests of H_0 against K. The smooth tests use maximum likelihood estimation while the generalised smooth tests use M-estimators that include both maximum likelihood and method of moments estimators. For both tests it may be first necessary, as in testing for a distribution from an exponential family, to remove any redundancy in the model by modifying $g_k(x;\theta,\beta)$ so that it involves the k $h_i(x;\beta)$ after the first q:

$$g_k(x;\theta,\beta) = C(\theta,\beta)\exp\left\{\sum_{i=q+1}^{q+k}\theta_i h_i(x;\beta)\right\} f(x;\beta).$$

In essence the β_i and the first q θ_i are playing similar roles, and having both in the model results in the test statistic being undefined due to the asymptotic covariance matrix being singular. In Rayner et al. (2009a, b) we construct smooth and generalised smooth test statistics that are sums of squares of components that are asymptotically independent and asymptotically standard normal.

If the null hypothesis is rejected, the revised $g_k(x;\theta,\beta)$ is a significantly better model for the data than $f(x;\beta)$. The fully parametric analysis assumes the model is $f(x;\beta)$. As the order k increases the alternative model $g_k(x;\theta,\beta)$ becomes richer and richer. If $k = n-1$ there is no restriction on the model assumed. This can be thought of as virtually a nonparametric environment. The parametric and nonparametric models occupy the extremes of this continuum; anywhere in between could reasonably be described as *partially parametric*. Clearly semi-parametric has a quite different meaning. It turns out that for the models we consider here, an advantage is obtained for k small - usually one or two. These methods could thus be described as *nearly parametric*.

The following sections briefly describe some of the results from Carolan (2000), some of which have been published in Carolan and Rayner (2000a,b; 2001) and Rayner (2001). Section 2 considers the one sample t-test, while sections 3 and 4 focus on completely randomised designs analysed by the one-way analysis of variance. Section 3 considers the simple two-sample case while section 4 demonstrates the sorts of gains that can be made in multi-sample situations. Finally randomised complete block designs are considered in section 5. For these simple but important designs the partially parametric tests exhibit large power gains. However, as the number of treatments and blocks in the model increases, we find the effectiveness of the partially parametric techniques is limited somewhat by computational considerations and the large sample sizes needed to obtain accurate maximum likelihood estimates.

In all cases in this paper the fully parametric model is normality. In this situation maximum likelihood and method of moments estimation coincide and to mark this all estimators will be written with hats: thus $\hat{\theta}$. Since σ^2 and θ_2 fill the same role, which one is removed

from the model is a matter of choice. Here σ^2 was usually the parameter removed. The test applied is usually a Wald test; see, for example, Rayner (1997). The Wald test statistic requires the asymptotic information matrix. This is given in, for example, Rayner and Best (1989, Section 3.3). After standardising the orthonormal functions $\{h_i(x;\beta)\}$ are the normalised Hermite-Chebyshev polynomials; see, for example, Abramowitz and Stegun (1972, 22.2.15). Because of their complexity we will only give the simpler test statistics.

When the data is *a priori* consistent with symmetry we will make that assumption as considerable efficiencies can be gained by removing all odd order terms in $g_k(x;\theta,\beta)$. This is consistent with the point we are consistently trying to demonstrate: that by working with appropriate models it is possible for considerable gains in power to be achieved. For the same reason we focus on models with less kurtosis than the normal; more kurtosis leads to difficulties in the tails of the adjusted models, and here we prefer to simply avoid these difficulties.

Rayner et al. (2009a, b) consider smooth testing for goodness of fit after model selection, based on a Barton model:

$$g_k(x;\theta,\beta) = \left\{1 + \sum_{i=q+1}^{q+k} \theta_i h_i(x;\beta)\right\} f(x;\beta).$$

For smooth testing this leads to the same tests as the Neyman models described previously. In fact, when testing for normality the normalising constant $C(\theta,\beta)$ does not exist, but because of the duality of the Neyman and Barton models, the smooth and generalised smooth tests are still valid.

2. One sample tests of location

Testing for location in the one sample case is probably one of the simplest but also most important problems encountered in statistical analysis. Given a sample $X_1, X_2, \ldots, X_n$ we wish to test the hypothesis that a well-defined location parameter takes a given value against the hypothesis that it takes some other value. For data assumed to be normally distributed with mean μ and variance σ^2 we wish to test $H_0 : \mu = \mu_0$ against $K : \mu \neq \mu_0$. The standard parametric approach is to use the t test with the test statistic,

$$T = \frac{(\overline{X} - \mu_0)\sqrt{n}}{S}$$

in which $\overline{X}$ is the sample mean and S the sample standard deviation. This statistic has the t_{n-1} distribution under the null hypothesis.

For non-normal symmetric data there are several nonparametric alternatives. These include the sign test, and the more generally favoured Wilcoxon signed-rank test, on which we will focus here. Subsequently we will often assume symmetry, in which case the median and the mean coincide and both the t and Wilcoxon tests are testing the same hypotheses.

The t test is the optimal test for normally distributed data and even when the data are not normal, the actual test size is close to the nominal size for quite large symmetric deviations from normality. On the other hand, the power of the t test diminishes as data becomes progressively more non-normal.

In this situation Carolan (2000) derived the score test statistic assuming the alternative

$$g_{\mathrm{sym},k}(x;\theta,\beta) = C(\theta)\exp\left\{\sum_{i=1}^{k}\theta_{2i}h_{2i}(x;\mu,1)\right\}f(x;\mu,1), \quad -\infty < x < \infty,$$

that is symmetric as it includes only θ of even order. As above, θ_2 takes the place of σ. The test statistic is

$$W = -(\hat{\mu}-\mu_0)^2\left\{\sum_{i=1}^{k}\hat{\theta}_{2i}\sqrt{2i(2i-1)}\sum_{j=1}^{n}h_{2i-2}(x_j;\hat{\mu})\right\}\Big/ n$$

with the parameter estimates being found by numerical optimisation. The corresponding tests retain the good power properties of the t test but clearly are valid for a wider variety of distributions.

Compared with both the t test and the Wilcoxon test, the test based on W may have substantially greater power when the true distribution is symmetric and non-normal. However this test has poor size properties that can be alleviated somewhat using a *Bartlett adjustment* or avoided all together by taking a bootstrap approach. In calculating bootstrap estimates of the quantiles of the test statistic, significant computational gains can be achieved by replacing the normalising constant by a polynomial approximation, thus avoiding the evaluation of the constant by numerical integration at each bootstrap iteration. For more details, including an outline of the derivation of W, see Carolan and Rayner (2000b).

Head size in brothers data. Frets (1921) used a data set that gave head measurements (length and breadth) for each of the first two adult sons in 25 families. These data are available in Hand, Daly, Lunn, McConway and Ostrowsky (1994). Here we will test whether or not there is a difference in head length between first and second sons. The paired nature of the data means that testing the hypothesis $H_0 : \mu_1 = \mu_2$ against $K : \mu_1 \neq \mu_2$ is equivalent to the one sample test of zero mean difference, that is $H_D : \mu_D = \mu_1 - \mu_2 = 0$ against $K_D : \mu_D \neq 0$. Typically if the assumption of normality appears to be satisfied the paired t test would be used; otherwise a nonparametric test such as the Wilcoxon signed rank test would be applied.

The differences have a mean of 1.88mm and a standard deviation of 7.54mm. In testing goodness of fit we take as our model $g_k(x;\theta,\beta)$ including only θ_2 and θ_4; the only nuisance parameter is μ. This is denoted by $g_{\mathrm{sym},2}$. The estimate of μ under $g_{\mathrm{sym},2}$ is $\hat{\mu} = 2.50$. This distribution along with the fitted normal distribution and a histogram of the data are plotted in Figure 2.1. Applying the t test yields a p-value of 0.22 and the Wilcoxon test a p-value of 0.28. Neither of these suggests any evidence of a difference in means. The Wald test on the other hand gives a parametric bootstrap p-value of 0.07; weak evidence of a difference in head lengths between first and second sons.

3. Two sample tests of location for symmetric data

The approach outlined in the preceding section can be extended to multi-sample problems. In this section we discuss two sample tests of location, and extend that discussion to multi-sample testing in the following section. Thus here we compare the parametric pooled

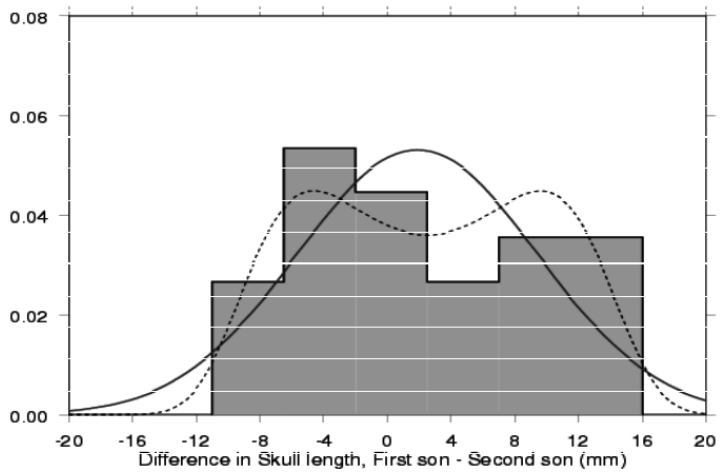

Figure 2.1. Scaled histogram and fitted probability density functions for the differences in skull length between first and second sons. Fitted normal distribution - solid, fitted $g_{\text{sym},2}$ - dashed

t test, the nonparametric Wilcoxon rank-sum test and a two sample partially parametric tests. We continue, as in the one sample case, to use Wald tests and partially parametric alternatives including θ_2, although we report σ because of its familiarity and ease of interpretation.

Egyptian skull measurements data. The data considered here are of ancient Egyptian skull measurements originally published by Thomson and Randall-Maciver (1905), used as an example in Manly (1986) and available in Hand, Daly, Lunn, McConway and Ostrowsky (1994). This data set contains measurements of various skull dimensions (maximum breadth, basibregmatic height, basialveolar length and nasal height) for samples of 30 Egyptian skulls from five different periods between 4000 BC and AD 150. One obvious question is whether or not these dimensions have changed over time. To fully answer this question a far more complex analysis taking into account the multivariate nature of the data and the ordering of the samples over time could be used. However here we will limit ourselves to the simple question of whether or not the mean basibregmatic height of skulls varies between a sample of skulls from circa 3330 BC and a sample from circa AD 150.

Figure 3.1 shows histograms of the measurements in mm, for the two samples. Also shown are fitted normal distributions, $g_{\text{sym},2}^{AD150}$ and $g_{\text{sym},2}^{3300BC}$, fitted to each sample separately and, $g_{\text{sym},2}^{\text{pool}}$, the distribution obtained when estimates of θ_4 are pooled across samples. As before, these models have $\theta_3 = 0, \mu$ fitted from the data and so are two parameter smooth alternatives to normality. Looking at the histograms it would appear the assumption of normality is questionable for the 150 AD data. An omnibus smooth goodness of fit test applied to the AD 150 data yields a statistic of $\hat{S}_4 = \hat{V}_3^2 + \ldots + \hat{V}_6^2 = 4.15$, which is significant at the 5% level. The components $\hat{V}_4^2$ and $\hat{V}_6^2$ make up 94% of this statistic suggesting that the deviation from normality is symmetric. The 3300 BC data gives an insignificant smooth goodness of fit test statistic of 0.55. Initial observations would also suggest that the AD 150 distribution is located slightly to the left of the distribution for 3300 BC: $\bar{x}_{3300BC} = 132.7$ and $\bar{x}_{AD150} = 130.3$. The pooled t test returns a p-value of 0.062, quite small but not significant at the oft used 5% level. The Wilcoxon rank-sum test returns a slightly larger p-value of 0.088.

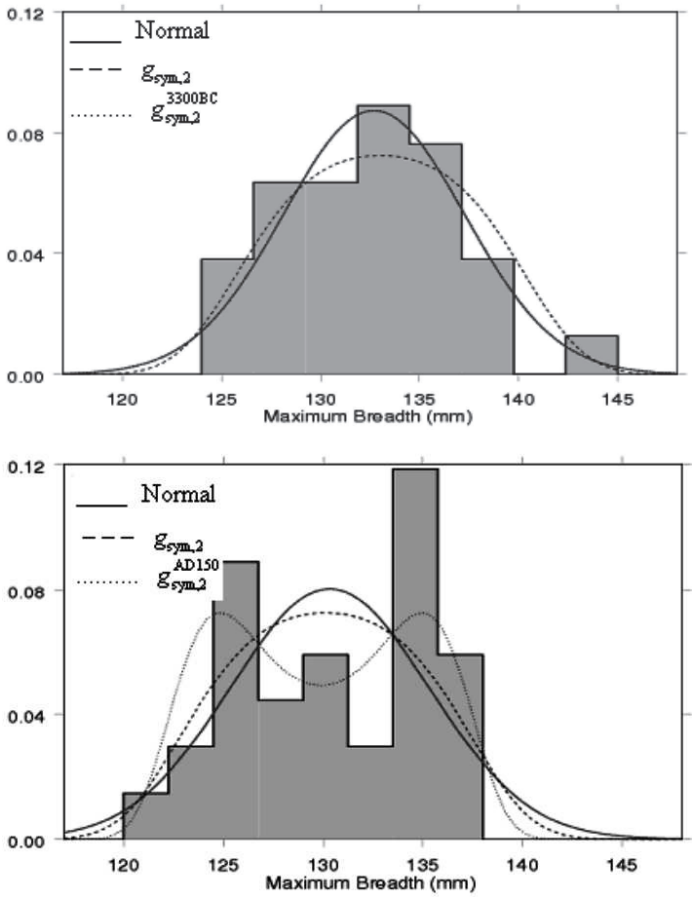

Figure 3.1. Histograms and fitted probability density functions for Egyptian skulls data. Note that for the 3300 BC data the normal distribution and $g_{\text{sym},2}^{3300BC}$ are identical

So both the standard tests suggest there is weak evidence of a change in location. If a $g_{\text{sym},2}$ distribution is fitted to each sample separately then parameter estimates $\hat{\mu} = 132.7, \hat{\sigma} = 4.47$ and $\hat{\theta}_4 = 0$ are obtained for 3300 BC (so the best fitting $g_{\text{sym},2}$ distribution is in fact the normal distribution) and $\hat{\mu} = 129.9, \hat{\sigma} = 4.12$ and $\hat{\theta}_4 = -0.82$ for 150 AD. The partially parametric test based on these estimates and using a bootstrap sampling size of 2500 yields a parametric bootstrap p-value of 0.02. In this case to make the assumption that σ and θ_4 can be pooled across samples appears somewhat dubious, although less dubious than the assumptions of normality and equal variance. For demonstration purposes this partially parametric test was applied with pooled estimates. Now our pooled estimates of σ and θ_4 are 4.51 and - 0.29 respectively, with estimates of μ_{3300BC} and μ_{AD150} being 133.03 and 130.83. The parametric bootstrap p-value for the test based on this model is 0.03. It appears all tests are suggesting some evidence of a shift in location with the partially parametric tests being most critical of the null hypothesis.

Having demonstrated the usefulness of the partially parametric tests on a particular data set, we now report on a simulation study to demonstrate their effectiveness in general. As

suggested in the example above, there are many partially parametric tests that can be applied in the multi-sample cases. The notation used in reporting the results are as follows. Subsequently the partially parametric tests are generically denoted by W because they are Wald tests; W_k means θ up to order k are included in the model, although there are other caveats as indicated. The number of samples compared is denoted by T. Of course in this section $T = 2$.

We now compare the pooled t test, the two sample Wilcoxon rank sum test and two partially parametric tests. The first test is denoted by W_k and involves fitting symmetric models to each sample separately. This may be impractical for small sample sizes or it may be that the assumption that the only difference between samples is a shift in location is valid, in which case the test statistic $W_{\text{pool},k}$, that assumes a common θ_2 (or σ^2) across symmetric populations, is appropriate. Here we take $k = 4$ and sample sizes $n_1 = n_2 = 20$. For T samples the test statistic W_k is given by

$$W_k = \sum_{t=1}^{T} n_t E\left[\hat{R}_t^2\right] \hat{\tau}_t^2 - \frac{\left(\sum_{t=1}^{T} n_t E\left[\hat{R}_t^2\right] \hat{\tau}_t\right)^2}{\sum_{t=1}^{T} n_t E\left[\hat{R}_t^2\right]}$$

in which $R(x_{tj}; \mu_t, \theta_t) = (x_{tj} - \mu_t) - \sum_{i=1}^{k} \theta_{2i}\sqrt{2i}\, h_{2i-1}(x_{ij}; \mu), \hat{R}_t = R(x; \hat{\mu}_t, \hat{\theta}_t)$ and the required expected values in W_k are evaluated numerically or replaced by the corresponding component of the observed information from the adjusted smooth model.

Figure 3.2 gives power curves for both partially parametric approaches along with the t test and Wilcoxon test, for $n_1 = n_2 = 20$, in a variety of situations covering cases where the assumption of fixed shape is both valid and invalid. We make some observations.

- In cases where the assumption of fixed shape across treatments is satisfied, the pooled test works just as well and in some cases (for example, for normally distributed data) slightly better than the test based on W_4. These improvements would be even more evident in cases where there are more treatments with fewer replicates.
- When both samples have the same non-normal distribution W_4 and $W_{\text{pool},4}$ are both considerably more powerful than the pooled t test and the Wilcoxon test.
- The behaviour of the power curves when the samples are taken from different populations is somewhat mixed. When the first sample is uniformly distributed and the second comes from $g_{\text{sym},2}$ with $\theta_4 = -0.8$ we find, as should be expected, that W_4, which allows different distributions for each treatment group, clearly performs better than the t test and $W_{\text{pool},4}$, both of which assume that the two samples are from identical distributions. In the other cases where populations differ between samples there is no clear pattern. When the first sample is normally distributed and the second $g_{\text{sym},2}$ with $\theta_4 = -0.8$, the t test and W_4 are very similar and both out perform $W_{\text{pool},4}$. On the other hand, when one sample is uniformly distributed and the other normally distributed, W_4 is inferior to the other tests.
- As previously, whether or not to use a partially parametric test, and if so, which one, depends on the sample size as well as degree of non-normality. This can be seen in Figure 3.3 which shows power curves for samples of size 50 from the normal distribution and $g_{\text{sym},2}$ with $\theta_4 = -0.8$. Comparison with the corresponding graph for samples of size 20 in Figure 3.2 shows the increase in sample size has improved both of the partially parametric tests relative to the t and Wilcoxon tests.

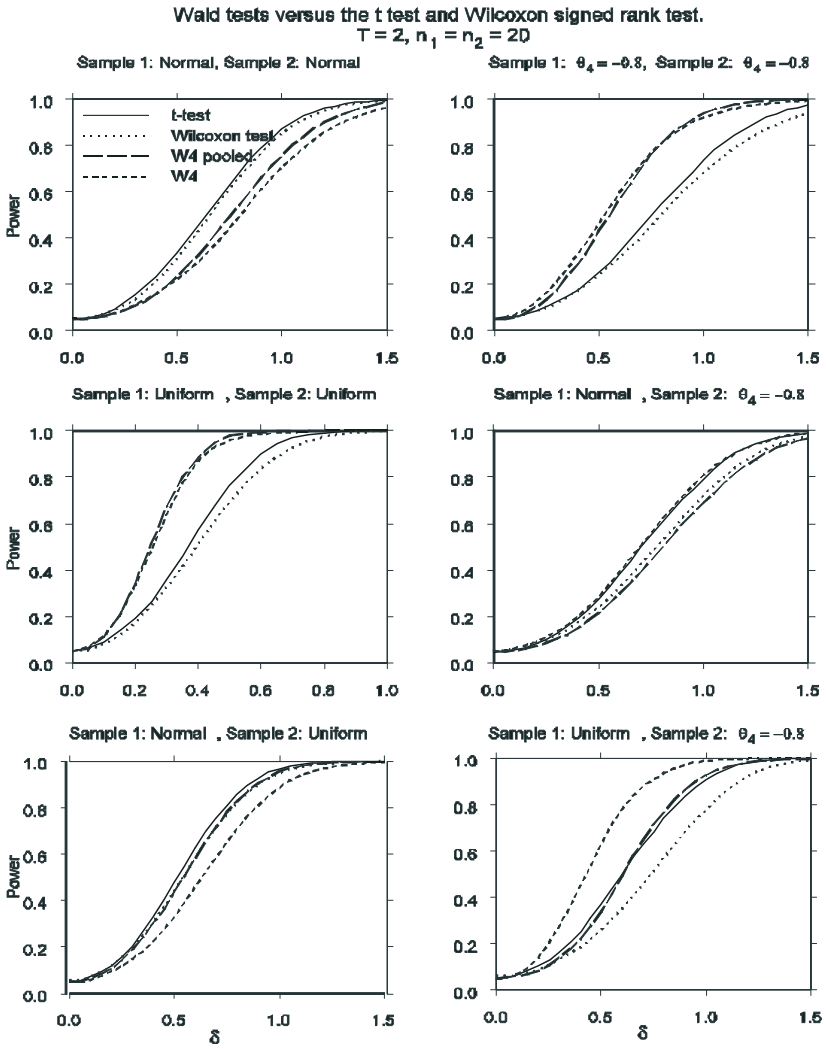

Figure 3.2. Comparison of power curves for the two sample t test, the Wilcoxon signed rank test, and the partially parametric tests based on W_4 and $W_{\text{pool},4}$ for testing $H_0 : \mu_1 = \mu_2$ against $K : \mu_1 \neq \mu_2$. Data come from the normal distribution, the uniform distribution and $g_{\text{sym},2}$ with $\theta_4 = -0.8$, and $\delta = \mu_1 - \mu_2$ is the true difference in means. Both samples contain 50 observations

4. Multi-sample tests of location for symmetric data

The results of the small power study reported here follow much the same pattern as in the previous section. When the data are significantly non-normal the pooled partially parametric test, $W_{\text{pool},4}$, can be considerably more powerful than the standard test, the one-way analysis of variance F test and the fully non-parametric competitor, the Kruskal-Wallis test. Figures 4.1, 4.2 and 4.3 give some examples.

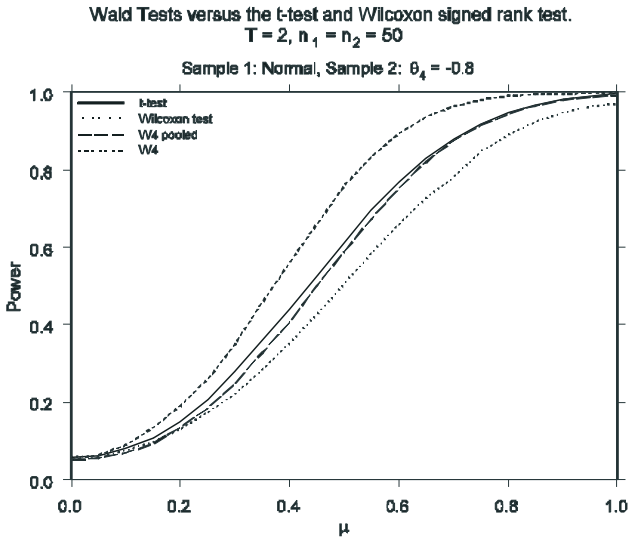

Figure 3.3. Comparison of power curves for the two sample t test, the Wilcoxon signed rank test, W_4 and $W_{\text{pool},4}$ for testing $H_0 : \mu_1 = \mu_2$, against $K : \mu_1 \neq \mu_2$. Data come from the normal distribution, the uniform distribution and $g_{\text{sym},2}$ with $\theta_4 = -0.8; \delta = \mu_1 - \mu_2$, the true difference in means. Both samples contain 20 observations

First, in the four sample case we compare the powers of the test based on $W_{\text{pool},4}$ (using resampling critical values), the standard one-way ANOVA F test and the Kruskal-Wallis test. Figure 4.1 shows the power of these tests for alternatives where $\mu_1 = \mu_2, \mu_3 = \mu_4$ but there may be a difference between μ_1 and μ_3. We use four samples of size 25 from the $g_{\text{sym},2}$ distribution with $\theta_4 = -0.8$. The pattern is much the same as we previously encountered. The partially parametric test is by far the most powerful with there being little difference between the other two tests.

For the three samples case we show, in Figure 4.2 and Figure 4.3, plots of the difference in power between Wpool,4 and the ANOVA F test. These simulations are for the case where one of the three sample means, μ_2, is fixed and the other two, μ_1 and μ_3, vary. The partially parametric test is more powerful than the ANOVA F test when the difference is greater than zero. Once again, as non-normality increases (data comes from $\theta_4 = -0.4, -0.8$ and -1.2 and the uniform distribution) the advantages of the partially parametric test over ANOVA become increasingly obvious.

5. Randomised complete block designs

Another important and frequently applied experimental design is the randomised complete block design. In this section we show how the partially parametric approach can be extended to this design. Again the Wald test statistics needed for the partially parametric tests discussed will not be made explicit here; they are given in Carolan (2000). We begin with an example.

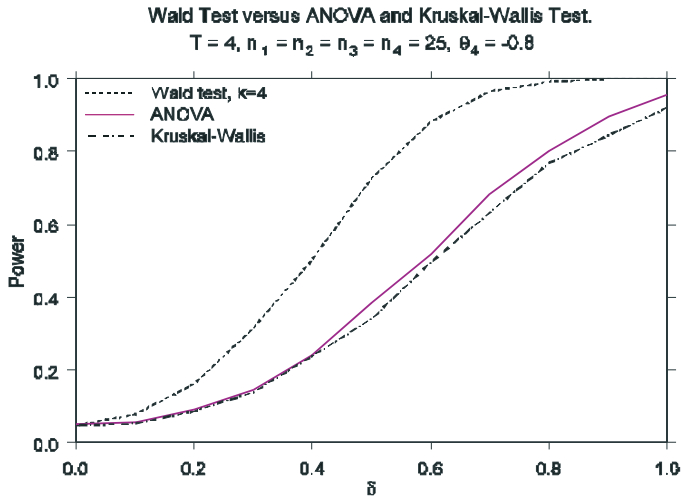

Figure 4.1. Power curves for the ANOVA F test, the Kruskal-Wallis test and the pooled partially parametric test based on $W_{\text{pool},4}$, for testing the hypothesis H_0 : $\mu_1 = \mu_2 = \mu_3 = \mu_4$ against K: at least one μ_i is different from the others in the four sample case with $n_1 = n_2 = n_3 = n_4 = 25$. Data comes from $g_{\text{sym},2}$ with $\theta_2 = 0$ and $\theta_4 = -0.8$ and $\mu_1 = \mu_2, \mu_3 = \mu_4$ with $\delta = \mu_1 - \mu_3$

Limpits data. Sokal and Rohlf (1995) considered a data set concerning the oxygen consumption of two different species of limpets, *Acmaea scabra* and *Acmaea digitalis*, in three concentrations of salt water: 50%, 75% and 100%. Figure 5.1 shows histograms of the data. While there are only eight observations in each cell it would appear that there is a consistent bimodal pattern. This may be due to male and female limpets differing in oxygen consumption or there being two distinct different age groups of limpets or some other such factor which has not been controlled in this experiment. In any case the distributions appear roughly symmetric but of dubious normality. A standard two way analysis of variance as performed by Sokal and Rolf (1995) yields p-values of less than 0.01 for salinity effects and 0.19 for species effects. Based on this analysis it would be reasonable to conclude that while oxygen consumption differs with salinity there is little evidence to suggest any difference between species. A nonparametric alternative is to apply ANOVA to the ranks of the data. For this data set ANOVA applied to rank transform of the data (ranked within blocks) yields a p-value of 0.17.

We now use a partially parametric approach here to test for a species effect; concentration is a factor of no interest here. Because there are only eight observations in each species by concentration combination it is impractical to estimate the θ_i separately for each combination. Furthermore the histograms do not suggest great differences in shape between the three concentrations or the two species so it would seem reasonable to pool estimates of the shape parameters, θ_i across treatments and blocks.

Figure 5.2 shows histograms for each species with the block (salinity) effects removed. At first glance it would appear that *Acmaea scabra* tends to have higher oxygen consumption than *Acmaea digitalis*. Furthermore the fitted curves suggest that the $g_{\text{sym},2}$ distribution provides a better estimate of the true distribution than the normal distribution in this case.

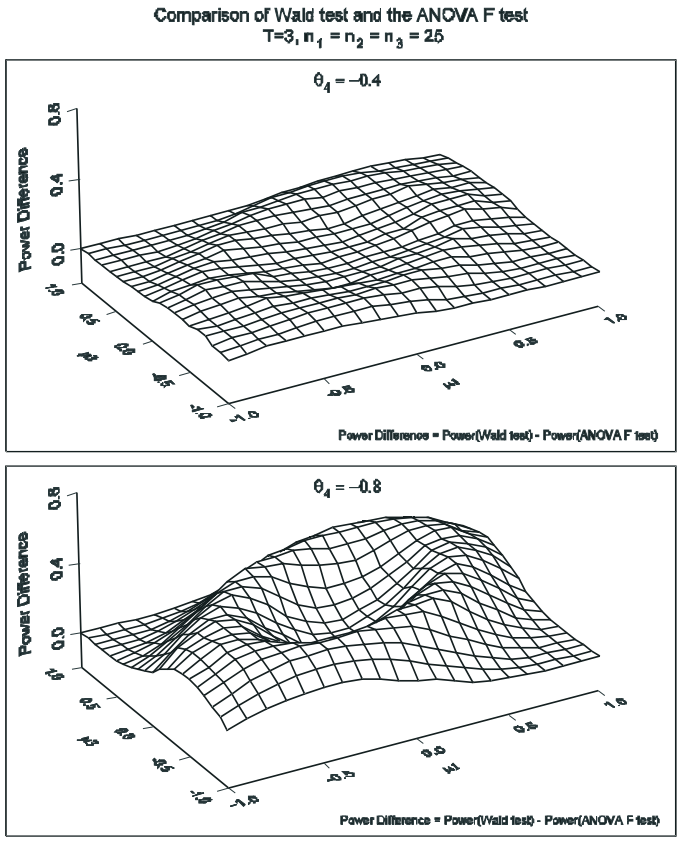

Figure 4.2. The difference in power between the partially parametric (Wald) test and the ANOVA F test, testing $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ against K: at least one μ_i is different from the others in the three sample case. Data comes from $g_{\text{sym},2}$ with $\mu_2 = 0, \theta_2 = 0$ and $\theta_4 = -0.4, -0.8$. Values greater than 0 indicate the Wald test is more powerful than the ANOVA F test

Applying the partially parametric Wald test as detailed in the previous section with $k = 4$ and using 500 bootstrap simulations we obtain a p-value of less than 1% for the test of $H_0 : \mu_{A.scabra} = \mu_{A.digitalis}$ against $K : \mu_{A.scabra} \neq \mu_{A.digitalis.}$ This p-value is more in line with our 'by eye' observations than the ANOVA p-value of 19%.

Having again demonstrated the usefulness of the partially parametric tests, we now report on simulation study to demonstrate their effectiveness in general. We first note that the ANOVA F test is based on a normal model, which, for T treatments and B blocks, has $BT + 1$ parameters: one for the residual variance and BT for location. These location parameters are partitioned into one for overall mean, $B - 1$ for blocks, $T - 1$ for treatments and $(B - 1)(T - 1)$ for interaction. We test if the $T - 1$ treatment parameters are zero with the remaining $BT - T + 2$ parameters entering the problem as nuisance parameters. To permit more flexibility in the model, the normal model can be replaced by, say, an order k alternative to normality. This will add $(k - 1)$ BT additional nuisance parameters.

As the number of parameters that must be estimated by maximum likelihood increases for

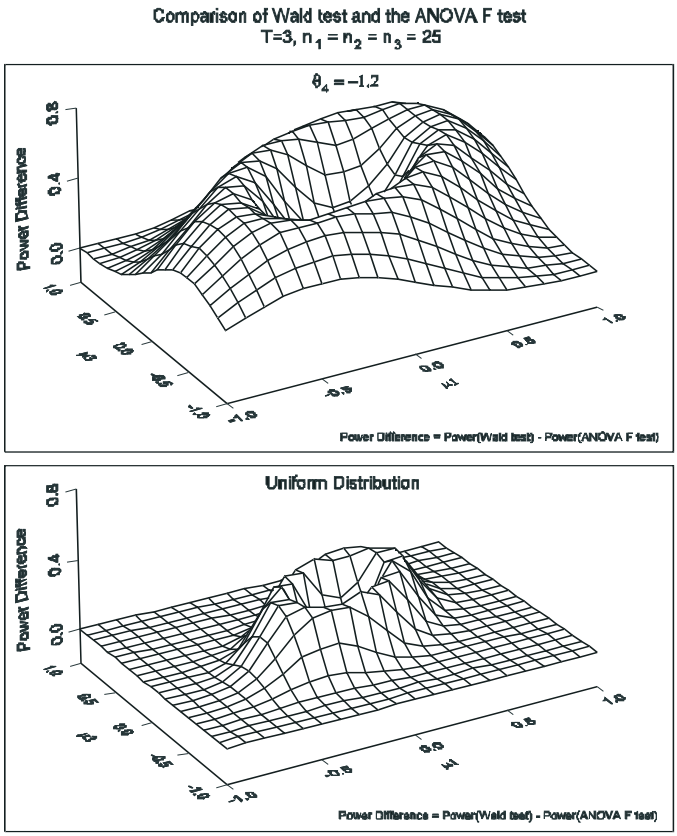

Figure 4.3. The difference in power between the partially parametric (Wald) test and the ANOVA F test, testing $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ against K: at least one μ_i is different from the others in the three sample case. Data come from $g_{\text{sym},2}$ with $\mu_2 = 0, \theta_2 = 0$ and $\theta_4 = -1.2$, and from the uniform (-1, 1) distribution. Values greater than 0 indicate the partially parametric test is more powerful than the ANOVA F test

a given sample size, so do the computational difficulties. Rather than itemise these difficulties, we will address the measures undertaken to proceed. We now consider three partially parametric tests when there are two treatments and two blocks. As before deviations from normality of order greater than four will not be considered, and symmetry will be assumed. In WRCB1 the smooth alternatives may vary between blocks but not between treatments. In WRCB2 the smooth alternatives are not permitted to vary between blocks or treatments. In WRCBLS to reduce computation even more the location block effects are replaced by their least squares estimates and again, the smooth alternatives are not permitted to vary between blocks or treatments.

In Figure 5.3 there appears to be little difference between the two pooled tests WRCB2 and WRCBLS. The only case where this difference appears to be significant is in the last graph where WRCBLS is actually more powerful (both WRCB2 and WRCBLS being less powerful than WRCB1). This is not entirely surprising as in this case data in one of

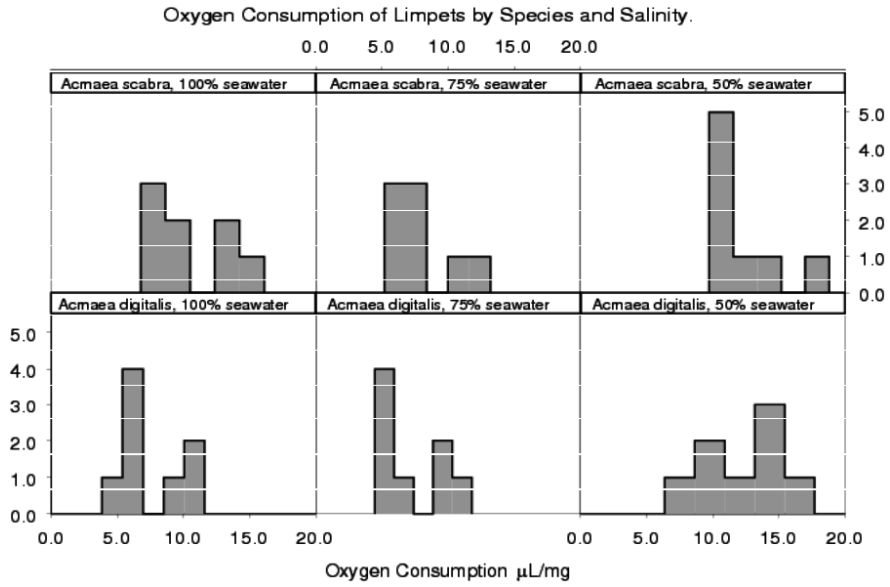

Figure 5.1. Oxygen consumption for two species of limpets in three different concentrations of seawater. Each cell has 8 observations

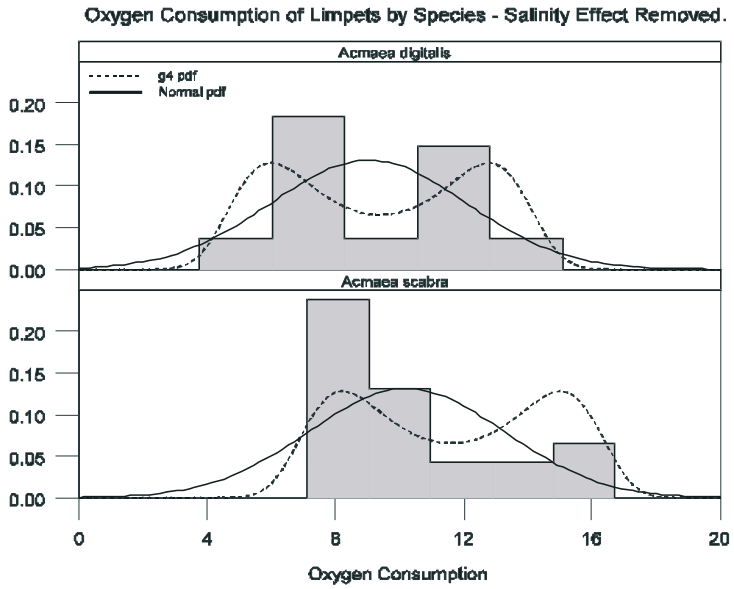

Figure 5.2. Oxygen consumption for two species of limpets with the salinity effect removed. Each species has 24 observations. Also shown is the fitted normal probability density function (with pooled estimate of standard deviation) and the fitted $g_{\text{sym},2}$ distribution with estimates of σ and θ_4 pooled across species

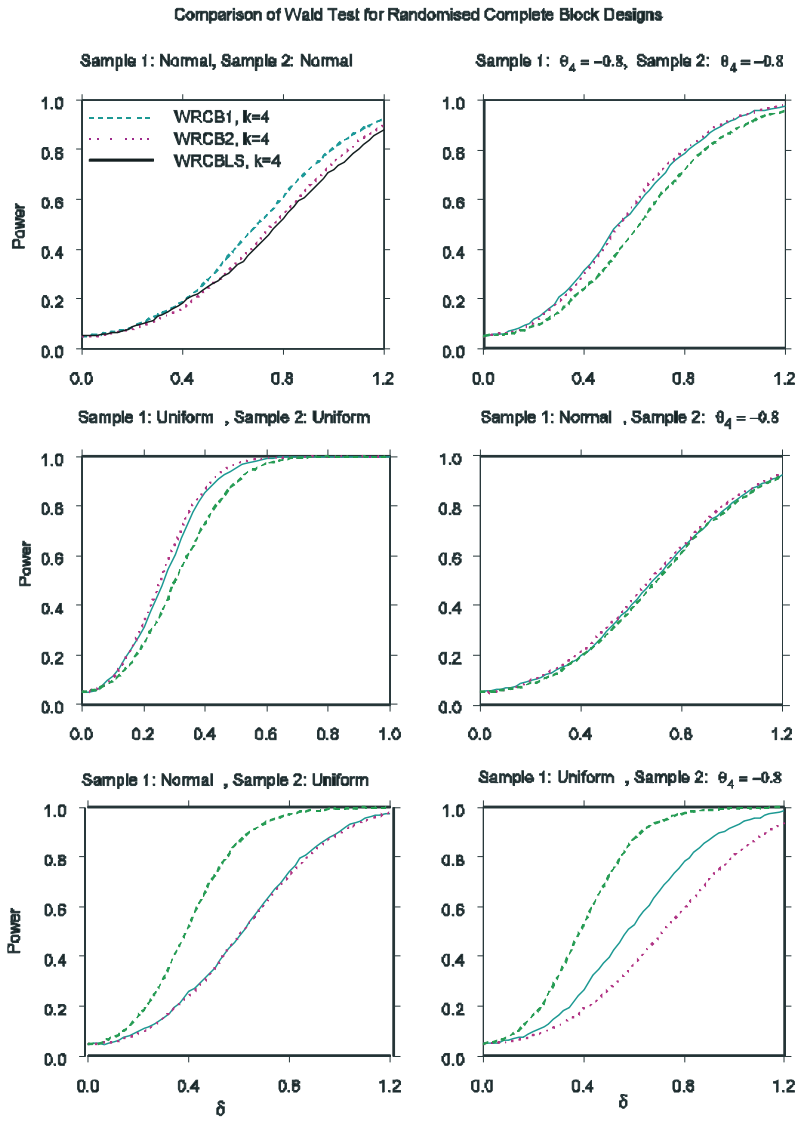

Figure 5.3. Power curves for the different versions of the Wald test for randomised complete block designs, testing $H_0 : \mu_1 = \mu_2$ against $K : \mu_1 \neq \mu_2$ for two treatments and two blocks with a total sample size of 40 observations

the blocks are uniformly distributed and so do not come from a $g_{\text{sym},k}$ distribution; furthermore both these tests are assuming that the shape, reflected by the θ_{tb} parameters, is fixed across blocks which is not the case. So we do not expect the test based on WRCB2 to be optimal and it is quite reasonable for the least squares version to perform better in this particular case. The 'full' model test WRCB1 performs much as expected: when there is no difference in shape between blocks it is no better and in some cases slightly worse than the tests that assume the shape is fixed across blocks. When there is a significant difference in shape between blocks WRCB1 performs best.

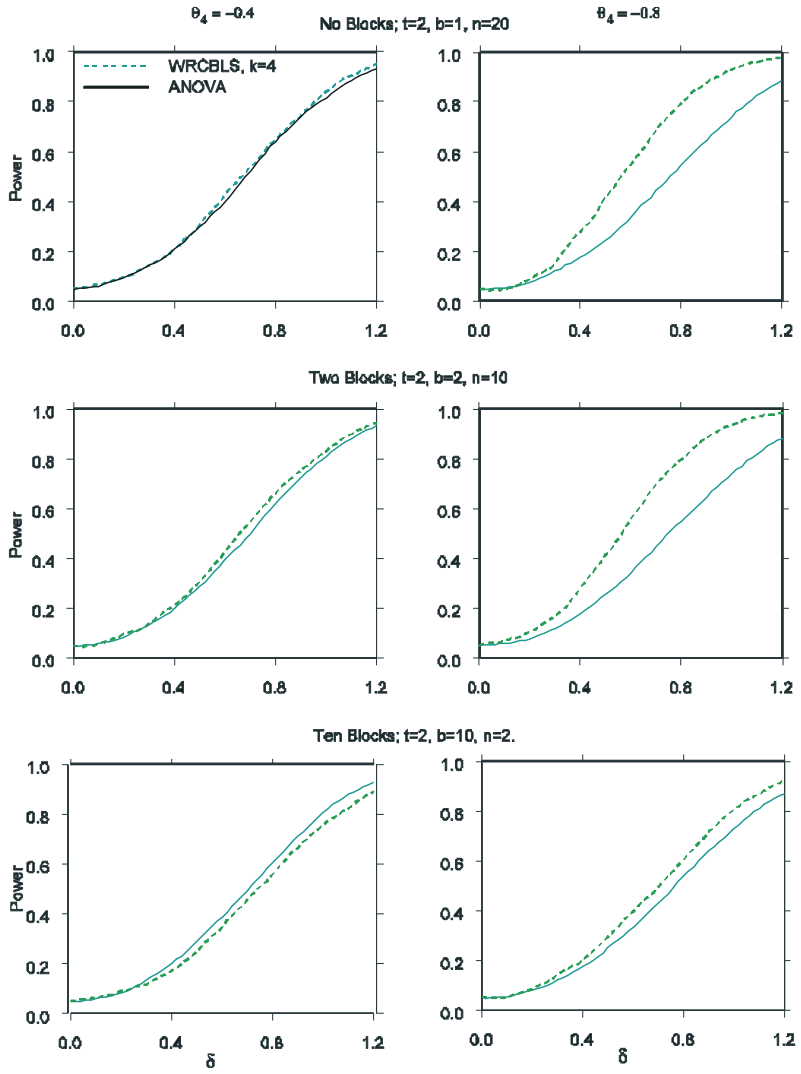

Figure 5.4. Power curves for the partially parametric test and ANOVA, testing H_0 : $\mu_1 = \mu_2$ against $K : \mu_1 \neq \mu_2$ for two treatments and various blocking scenarios with a total sample size of 40 observations. Data comes from $g_{\text{sym},2}$ distribution with $\theta_4 = -0.4$ and $\theta_4 = -0.8$

We now investigate the effect of an increasing number of blocks with fixed sample size. It is impractical to consider partially parametric tests other than WRCBLS. This is compared with the ANOVA F test. Results are summarised in Figure 5.4 in which we demonstrate the power of a test with two samples and two blocks with data from $g_{\text{sym},2}$ distribution with $\theta_4 = -0.4$ and $\theta_4 = -0.8$. In each case the total number of observations is 40 with the number of blocks either one (no blocking, and thus one-way analysis of variance), 2 or 10. In each case there are significant gains over the standard ANOVA for $\theta_4 = -0.8$, diminishing as the number of blocks increases. When $\theta_4 = -0.4$ the data are close to normal and ANOVA

actually begins to perform better than the partially parametric test as the number of blocks increases. It seems that the advantage of the partially parametric approach is greater for blocks containing more observations and for distinctly non-normal data.

6. Conclusion

In traditional parametric inference the statistician often checks the assumptions of the analysis, and hopes they are either valid or the deficiencies can be easily remedied. The alternative is often to use less powerful or more complicated methods of analysis. Of course, doing nothing is easiest, but then the inference is invalid. Here we have tried to demonstrate, by examples and power studies, that basing analysis on a valid model can have considerable benefits. The models we use here are perhaps somewhat simplistic, from the smooth goodness of fit analysis of the model initially proposed. In Rayner et al. (2009b) and elsewhere, more sophisticated model selection methods are discussed, but the message remains the same: far more powerful analysis is available if valid models are used.

References

Abramowitz, M., Stegun, I.A., 1972. *Handbook of Mathematical Functions*. Dover, New York.

Carolan, A.M., 2000. *Partially Parametric Testing*. Unpublished PhD thesis. School of Mathematics and Applied Statistics, University of Wollongong.

Carolan, A.M., Rayner, J.C.W., 2000a. One sample tests of location for nonnormal symmetric data. *Commun. Statist.-Theor. Meth.*, 29(7), 1569–1581.

Carolan, A.M., Rayner, J.C.W., 2000b. Wald tests of location for symmetric nonnormal data. *Biometrical Journal*, 42(6), 777–792.

Carolan, A.M., Rayner, J.C.W., 2001. One sample tests for the location of modes of nonnormal data. *Journal of Applied Mathematics and Decision Sciences*, 5(1), 1–19.

Frets, G.P. (1921), Heredity of head form in man. *Genetica*, 3, 193–384.

Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J., Ostrowski, E., 1994. *A Handbook of Small Data Sets*. Chapman and Hall, London.

Manly, B.F.J., 1986. *Multivariate Statistical Methods*. Chapman and Hall, New York.

Rayner, J.C.W., 1997. The Asymptotically Optimal Tests. *J.R.S.S., Series D* (The Statistician), 46(3), 337–346.

Rayner, J.C.W., 2001. Partially parametric testing. Chapter 32 in *Goodness-of-Fit Tests and Validity of Models*, Huber-Carol, C., Balakrishnan, N., Nikulin, M. S., Mesbah, M. (editors), Birkhauser, Boston, 423–431.

Rayner, J.C.W., Best, D.J., 1989. *Smooth Tests of Goodness of Fit*. Oxford University Press, New York.

Rayner, J.C.W., Best, D.J., Thas, O., 2009a. Generalised smooth tests of goodness of fit. *Journal of Statistical Theory and Practice*, 3(3), 665–679. Companion paper.

Rayner, J.C.W., Thas, O., Best, D.J., 2009b. *Smooth Tests of Goodness of Fit: Using R* (2nd ed.). Wiley, Singapore.

Sokal, R.R., Rohlf, F.J., 1995. *Biometry: the Principals and Practice of Statistics in Biological Research*. Freeman, New York.

Thomson, A., Randall-MacIver, R., 1905. *Ancient Races of the Thebaid*. Oxford University Press, Oxford.