Two permutation tests of equality of variance

Article *in* Statistics and Computing · December 1995

DOI: 10.1007/BF00162501

CITATIONS
9

READS
174

1 author:

Rose Dawn Baker
University of Salford
205 PUBLICATIONS **2,978** CITATIONS

SEE PROFILE

Statistics and Computing (1995) **5**, 289–296

Two permutation tests of equality of variances

ROSE D. BAKER

Centre for OR and Applied Statistics, Department of Mathematics and Computer Science, University of Salford, Salford M5 4WT, UK

Received and accepted November 1994

The F-ratio test for equality of dispersion in two samples is by no means robust, while non-parametric tests either assume a common median, or are not very powerful. Two new permutation tests are presented, which do not suffer from either of these problems. Algorithms for Monte Carlo calculation of P values and confidence intervals are given, and the performance of the tests are studied and compared using Monte Carlo simulations for a range of distributional types. The methods used to speed up Monte Carlo calculations, e.g. stratification, are of wider applicability.

Keywords: Permutation test, F-ratio test, scale problem, Monte Carlo, stratification

1. Introduction

The parametric test for equality of variances from two samples is the F-ratio test (Fisher, 1924), and the corresponding multisample test is Bartlett's modification of a test due to Pearson and Neyman (Bartlett, 1937). For two samples, the Bartlett test reduces to a two-sided F-ratio test. These tests assume that under H_0 the distributions do not differ in variance, i.e. for the two-sample test $\sigma_1^2 = \sigma_2^2$, but their means may differ, $\mu_1 \neq \mu_2$.

Unfortunately, when the random variates are not normally distributed, the F-test is not robust (Box, 1953), and neither is Bartlett's test (Rivest, 1986). Unlike the t-test, which is robust, the F-ratio test is very sensitive to non-normality. For short-tailed distributions the test is conservative, while more seriously, for long-tailed distributions the probability of a type 1 error can be much larger than the nominal size of the test (the test is liberal). Figure 1 shows the unacceptability of the parametric F-ratio test for a long-tailed (double exponential) distribution.

There are many non-parametric tests, which do not suffer from this problem. Duran (1976) gives a review. However, these tests usually assume a common centre of location for the two distributions. The only two-sample test which does not make this assumption is the Moses test (Moses, 1963), which lacks power.

Hence it seems that no test currently available is completely satisfactory. It is natural to turn to resampling tests, either bootstrap tests or permutation tests, which are powerful and distribution-free, but which require extensive computation.

When sample sizes are equal, Good (1994) proposes an ingenious permutation test that does not assume a common centre of location for the two distributions. In Good's test, squared deviations about sample medians are permuted. When sample sizes are equal, such transformed observations, although not independent, are claimed by Good to be exchangeable.

In this paper, the properties of Good's test are studied, and it is extended to cope with unequal sample sizes. A mean-based permutation test is also proposed, and the performance of the two tests compared.

2. Resampling tests

It is possible to develop both two-sample and multisample tests. This paper is concerned with two-sample tests only, but the extension to multisample tests is straightforward. Assume without loss of generality that sample sizes are n_1, $n_2 \geq n_1$, and let $n = n_1 + n_2$. Permutation tests in general take a test statistic T used for a parametric test, or one derived intuitively, and generate a reference distribution for the statistic by permuting observations between the two groups in all $\binom{n}{n_1}$ possible ways. If the groups are

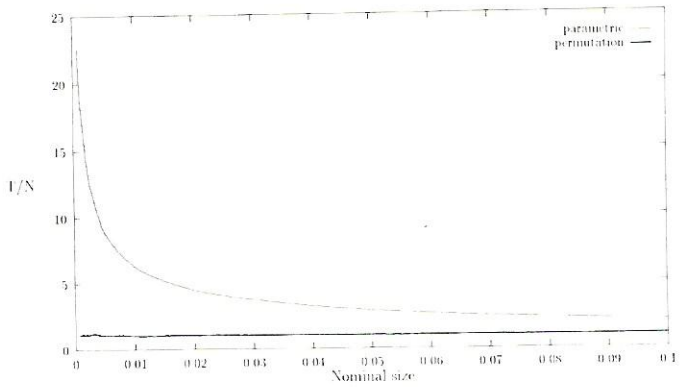

Fig. 1. *Ratio of true size of test to nominal size (T/N) plotted against nominal size, for two samples of sample size 50 from a double exponential (long-tailed) distribution, under H_0. The parametric (F-ratio) test is very 'liberal', but the permutation test (test 1) is exact*

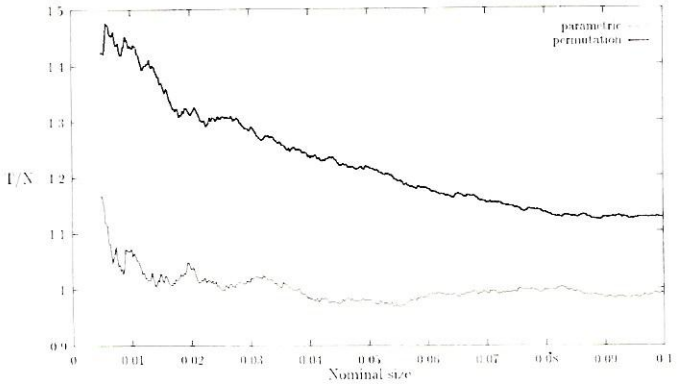

Fig. 2. *Ratio of true size of test to nominal size (T/N) plotted against nominal size, for two samples of sample size 10 from a normal distribution, under H_0. The parametric (F-ratio) test is exact, but the 'invalid' permutation test which permutes differences of observations from their group mean is liberal*

identical under H_0, this operation is justified. The P value (the 'significance probability' or 'associated probability') for a test is just the fraction of permuted test statistics T' which are as large or larger than T, if large values of T denote disagreement with H_0. If complete enumeration of all permutations is not feasible, Monte Carlo simulation provides an adequate substitute.

In general, the true size of a permutation test equals its nominal size (the test is exact). This is because one can rank the $\binom{n}{n_1}$ values of T', and under H_0 the observed value, T, is equally likely to be any one of the T'. Hence under H_0 the P value of the test is uniformly distributed among the fractions $1/\binom{n}{n_1}$, $2/\binom{n}{n_1} \cdots 1$. For all but very small samples the P value can be thought of as a random variate from the continuous uniform distribution.

To test $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_2^2 > \sigma_1^2$ the statistic used is $F = s_2^2/s_1^2$, the ratio of sample variances for the two groups. As the two groups may differ in mean, and μ_1, μ_2 are unknown, it is now not possible to swap observations between groups, and recalculate $F' = s_2'^2/s_1'^2$ in the usual way, where F' is a value of the test statistic obtained by randomly assigning observations to the two groups. Figure 2 shows the failure of a 'gung-ho' permutation test, in which one simply subtracts the appropriate sample mean from each observation before permutation, and ignores the fact that observations are not now exchangeable. Simulations show this crude permutation test to be still liberal, albeit less so than is the parametric test.

An alternative approach is a bootstrap resampling test. Here one technique is that random samples are generated with replacement from each of the two groups separately. This gives a distribution for the test statistic F from which 95% confidence limits, for example, can be read off. H_0 is rejected at the 5% level of significance if the lower confidence limit exceeds unity. The P value will only be exact asymptotically, when the empirical cumulative distribution function approaches the true distribution function, and one is then effectively sampling from the true population. For small sample sizes the test is not exact. Good (1994) presents results based on a small number (500) of simulations which suggest that the bootstrap test is at least as powerful as his permutation test. It is however possible that this extra power may be obtained at the cost of rejecting H_0 too often when it is true (a liberal test).

Bootstrap methods are surveyed in Hinkley (1988) and the construction of bootstrap confidence intervals described in detail in Diciccio and Romano (1988). These papers are followed by a discussion.

In this paper the permutational method is adopted. To obtain a valid permutation test, the F-ratio statistic must be modified somehow. A mean-based test, which modifies the F statistic slightly, for each of the $\binom{n}{n_1}$ permutations envisaged is described first (test 1), and then an extension to Good's median-based test (test 2).

2.1. *A mean-based permutation test*

The first step in obtaining test 1 is to reformulate the inequality $F' \geq F$. For a particular permutation, there are four classes of observation. These classes are:

1. $n_1 - m$ group 1 observations that remain in group 1 with sum of squares $SS_1(\mu_1)$
2. m group 1 observations that move to group 2, with sum of squares $SS_1'(\mu_1)$
3. $n_2 - m$ group 2 observations that remain in group 2, with sum of squares $SS_2(\mu_2)$
4. m group 2 observations that move to group 1, with sum of squares $SS_2'(\mu_2)$.

Here $m \leq n_1$ and varies from permutation to permutation, and sums of squares are taken about group population means. The corresponding sums of squares taken about the sample means of the observations in the four classes are written as SS_1, SS_1', SS_2, SS_2' respectively.

The F ratio is

$$F = \frac{(SS_2(\mu_2) + SS_2'(\mu_2))/(n_2 - 1)}{(SS_1(\mu_1) + SS_1'(\mu_1))/(n_1 - 1)}. \quad (1)$$

On making the permutation, the new value of F is

$$F' = \frac{(SS_2(\mu_2) + SS_1'(\mu_1))/(n_2 - 1)}{(SS_1(\mu_1) + SS_2'(\mu_2))/(n_1 - 1)}. \quad (2)$$

The condition $F' \geq F$ obtains iff $SS_1'(\mu_1) \geq SS_2'(\mu_2)$; this result is intuitive, and the proof, which is omitted, is trivial.

The test then consists of generating the $\binom{n}{n_1}$ permutations of the sample, and counting the fraction of occurrences of $F' \geq F$, or equivalently of $SS_1'(\mu_1) \geq SS_2'(\mu_2)$. Since μ_1, μ_2 are unknown, it is natural to replace them in each of the $\binom{n}{n_1}$ comparisons by the sample means of the m observations concerned in the comparison, i.e. to take the P value as the fraction of occurrences of $SS_1' \geq SS_2'$.

In terms of the original condition $F' \geq F$, in addition to F' changing from permutation to permutation, the value of

$$F = \frac{(SS_2 + SS_2')/(n_2 - 1)}{(SS_1 + SS_1')/(n_1 - 1)}$$

itself now changes slightly with each permutation. The argument leading to test exactness no longer applies. However, the test has been seen from many simulations to be always conservative and never liberal. Thus it is a valid test, and the only problem is the loss of power caused by taking sums of squares about class means. This means that the test is in general conservative, and when $\sigma_2^2/\sigma_1^2 \simeq 1$ it can even be biased, so that the test power $\beta < \alpha$, where α is the nominal size of the test.

The conservative property of the test can be made plausible by seeing that the new condition $SS_1' \geq SS_2'$ is 'noisier' than the old, so that small P values will tend to increase. When H_0 is true, this makes the test biased. Consider a large number of simulations of samples of size n, from two distributions such that $\sigma_2 > \sigma_1$, so that the condition $SS_1'(\mu_1) \geq SS_2'(\mu_2)$ holds with probability $p_0 < 1/2$. $SS_1'(\mu_1) - SS_2'(\mu_2)$ will be positive when the sample 2 observations cluster round their mean μ_2 and the sample 1 observations do not cluster round μ_1. However, $SS_1' - SS_2'$ will be positive whenever the m sample 2 observations are clustered round any value (so that SS_2' is small) and sample 1 observations are not. The probability of this event will on average be larger, giving a P value less than p_0.

Asymptotically, the numerator and denominator approach those of the original F statistic, and thus the permutation test becomes identical to the parametric test for large samples and normal distributions. In particular, it has asymptotic relative efficiency (ARE) of unity, and is asymptotically unbiased.

The multisample test is based upon Bartlett's k-group statistic

$$T = \prod_{i=1}^{k} (s_i^2/s^2)^{(n_i - 1)/2},$$

where s_i is the sample standard deviation of the ith group, and s^2 the pooled variance,

$$s^2 = \sum_{i=1}^{k} (n_i - 1)s_i^2 / \left(\sum_{i=1}^{k} n_i - k \right).$$

Under a random permutation, the n_i observations in the ith group are replaced by m_{ij} observations from the jth of the k groups, so that $\sum_{j=1}^{k} m_{ij} = n_i$. As before, each of the k^2 classes of m_{ij} observations is allowed its own sample mean, so that the s_i^2 are replaced by the corresponding pooled variances. In the spirit of Bartlett's approach, the powers of $n_i - 1$ should be replaced by $n_i - k$. Clearly in the multisample situation more degrees of freedom must be sacrificed, and the computational simplicity of the two-sample case is lost.

2.2. *An extension of Good's test*

Test 2 is based upon Good's test. The equal sample-size test described in Good (1994) is discussed first, and then an extension to $n_2 \neq n_1$ is given.

Since μ_1, μ_2 are unknown, Good's solution is to permute deviations from the corresponding sample medians, the sample median itself being discarded. When the number of observations is even, the median is taken as the mean of the two central order statistics as usual, and both these two are discarded. Let the remaining number of observations be $n_1' = n_1 - 1$ (odd n_1) or $n_1 - 2$ (even n_1).

Good (1994) argues that the squared deviations are exchangeable. It is clear that under H_0 they all have the same univariate distribution, all being squared deviations from the median which do not depend on the population mean. However, the operation of subtracting the sample median from each observation induces a weak dependency among the squared deviations from one group, which is not preserved on permuting squared deviations between groups. A Monte Carlo simulation of P values when $n_1 = n_2 = 5$ (odd) showed that the test was definitely slightly conservative, especially at low P values. For $n_1 = n_2 = 6$ (even) the test was definitely liberal. Good has not observed an exact test, but his use of the median has greatly reduced the departure from exactness shown in Fig. 2.

Good's test will not work for unequal sample sizes, as even the univariate distribution of deviations from the sample median then differs between the samples. Squared deviations are on average larger for smaller samples because of the sampling error on the median. The test would tend to find that smaller samples had larger

variance. The following procedure is a possible solution, in which the group 2 median is randomly chosen from a subset of n_1 group 2 observations:

1. Find the n_1' squared deviations of group 1 observations from their group median.
2. Select a random sample of n_1 observations from group 2.
3. Find the n_1' squared deviations from the median of this random sample.
4. Count the fraction of randomizations in which the sum of squared deviations for a random sample of n_1' observations from the $2n_1'$ observations from the combined group 1 sample and group 2 subsample is at least as large as the observed group 1 sum of squared deviations.

In this procedure, like is compared with like. There are $\binom{2n_1'}{n_1'}$ $\binom{n_2}{n_1}$ possible permuted group 1 sums of squared deviations.

3. Monte Carlo calculation of *P* values

The calculation of P values for test 1 is given first, then that for test 2, and finally a stratified method applicable to both tests.

3.1. *Calculation of P values for test 1*

The most direct method is to choose n_1 observations randomly from the total to form the permuted group 1, while retaining their original group labels. F and F' can then be found because all required sums of observations and sums of squares can be found directly or by subtraction from the grand total over all $n_1 + n_2$ observations.

An alternative and slightly faster method results from the direct use of the condition $SS_1' \geq SS_2'$. This method starts with the generation of a random value of m. The probability of a value of m is

$$P_m = \binom{n_1}{m}\binom{n_2}{m} \Big/ \binom{n_1+n_2}{n_1},$$

so that the distribution of m is a special case of the hypergeometric distribution. Random values of m are generated using the probability integral method.

When $m \leq n_2/2$, m observations are randomly selected from group 2 and used to calculate SS_2'. When $m > n_2/2$, the $n_2 - m$ unpermuted observations are randomly selected, and the required sum and sum of squared values needed to calculate SS_2' are found by subtraction from the corresponding totals for group 2. Similarly, when $m \leq n_1/2$, m observations are randomly selected from group 1 and used to calculate SS_1'. When $m > n_1/2$, the $n_1 - m$ unpermuted observations are randomly selected, and the required sum and sum of squared values needed to calculate SS_1' found by subtraction from the corresponding totals for group 1. When m is 0 or 1, $SS_1' - SS_2'$ is zero, and so the condition $SS_1' \geq SS_2'$ is automatically satisfied.

This method is slightly faster than the naive method, as it needs $\min(m, n_1 - m) + \min(m, n_2 - m)$ permutations, which is less than n_1 permutations needed previously. For equal sample sizes, when $n_2 = n_1$, on approximating the hypergeometric distribution which has mean $\mu = n_1 n_2/(n_1 + n_2)$, $\sigma_2^2 = n_1^2 n_2^2/(n_1 + n_2)^2(n_1 + n_2 - 1)$ by a normal distribution, the number of permutations can be shown to be $n_1 - n_1^{1/2}/\sqrt{\pi}$. As sample sizes increase, the random variate m clusters ever more closely around its mean value of $n_1/2$, and the percentage improvement over the naive method, $100 n_1^{-1/2}/\sqrt{\pi}$, tends to zero.

3.2. *Calculation of P values for test 2*

A simple algorithm follows. Here the number of observations in group 1 after removing the median is n_1'. The data structure required is simply an array of group 1 observation values, and similarly for group 2. No extra storage is required. The steps are:

1. Find the median of the group 1 array. Subtract the median from each observation and square, and squeeze the array up to remove the median observations from the array. The following steps are carried out *nsims* times.
2. Obtain an array of n_1 observations by randomly permuting the group 2 array by exchanging the first, second, ... element with a random element from the remainder. If $n_2 - n_1 < n_1$ however, it is faster to choose the $n_2 - n_1$ elements *not* to be selected, and to use the remainder of the array in what follows.
3. Find the median of the n_1 elements selected from group 2, and reorder this section of the group 2 observations array so that the one or two median values fall at the end, by swapping it (them) with the last one (two) observations. The n_1' observations remaining, excluding the median(s) are the 'group 2 subset'.
4. Randomly choose m, the number of group 2 observations in the permuted group 1, where m runs from 0 to n_1'.
5. Accumulate the group 1 contribution to the permuted group 1 sum of squares, by permuting transformed group 1 observations and summing the $n_1' - m$ elements selected.
6. Accumulate the group 2 contribution to the permuted group 1 sum of squares, by permuting m group 2 observations within the group 2 subset. The sums of squared deviations from the subset median for these m observations are added to the permuted group 1 sum of squares.
7. Increment a counter *exceed* if the permuted group 1 sum of squares is at least as large as the observed value, or if $m = 0$.
8. Finally, the P value is *exceed/nsims*.

Note that finding the median of an array of n observations can be carried out in $O(n)$ steps: see for example Press *et al.*

(1992). In the faster method they quote, as a side-effect of the selection the array is permuted, which does not matter here. The number of random numbers required is $\min(2n_1, n_2)$, neglecting terms of order 1, and hence the computing time required is of this order.

3.3. *A stratified method*

Stratification offers a way of speeding up Monte Carlo calculations, and the following stratified method works for either test.

Under H_0, the probability p_m that $SS'_1 \geq SS'_2$ will not depend on m, is always $\frac{1}{2}$. However under H_1 this probability will depend on m, which suggests that stratified Monte Carlo simulation might be quicker than naive Monte Carlo. The strata would be values or ranges of m.

Under a stratified simulation, let the probability that a random sample falls in the ith of s strata be f_i. Let there be N_i samples in the stratum, of which $\hat{p}_i N_i$ satisfy the condition $SS'_1 \geq SS'_2$. The P value of the test is estimated as

$$p = \sum_{i=1}^{s} f_i \hat{p}_i,$$

and its variance is

$$\sigma_p^2 = \sum_{i=1}^{s} f_i^2 p_i(1-p_i)/N_i.$$

There is a (very) small immediate gain in stratification, because known probabilities f_i replace the estimated probabilities $N_i / \sum_{j=1}^{s} N_j$ which are implicitly used if the P value is calculated ignoring stratification. However, there is a much greater gain if the N_i are chosen to minimize σ_p^2, so that $N_i \propto f_i \sqrt{p_i(1-p_i)}$.

Simulation studies showed that the benefit from stratifying on m was small. A variable which correlates more strongly with the F-ratio is needed, so that the p_i will vary strongly across strata. To exploit such a variable, it is necessary to be able to compute its distribution, the f_i, and to generate random samples from a specific stratum. A variable that satisfies these conditions is the median of the group 1 Siegel–Tukey ranks, and the stratification procedure described in Appendix A is faster than the naive method, especially when the P value is small.

4. Calculation of confidence intervals on the ratio of variances

Test 1 is discussed first, then test 2. The ratio of population variances is $\sigma_2^2/\sigma_1^2 = r$, and a confidence interval for r can be found by exploiting the fact that the $100(1-\alpha)\%$ confidence interval for r is the set of values of r_0 for which a test with H_0 that $r = r_0$ against H_1 that $r \neq r_0$ has P value not less than α.

The 'brute force' method is to scale up group 2 observations by some factor $\sqrt{r_0}$, and to increase r_0 until the P value of a 1-sided test permutation test with H_1 that $r > r_0$ decreases to $\alpha/2$. This gives the upper $100(1-\alpha/2)\%$ confidence limit, and r_0 is then lowered until a test with H_1 that $r < r_0$ has P value of $\alpha/2$. This requires one significance test to be carried out for each trial value of r_0.

A much quicker method, which is based on the use of pivotal statistics, is given for tests of location in Gabriel and Hall (1983), and in Tritchler (1984). Here, it is only necessary for the ratios SS'_2/SS'_1 from each permutation to be stored and sorted into ascending order.

When under H_0 $r \neq 1$, to carry out a significance test the sums of squares in equations (1) and (2) must be scaled as follows:

$$SS_1 \to SS_1/\sigma_1^2 \qquad SS'_1 \to SS'_1/\sigma_1^2$$
$$SS_2 \to SS_2/\sigma_2^2 \qquad SS'_2 \to SS'_2/\sigma_2^2,$$

and the condition that $F' \leq F$ is now equivalent to $SS'_1/\sigma_1^2 \leq SS'_2/\sigma_2^2$, or $r \leq SS'_2/SS'_1$. Hence if $\Pr(F' \leq F) = \alpha$, then $\Pr(r \leq SS'_2/SS'_1) = \alpha$. Hence the upper $100\alpha\%$ confidence limit of r is the $100\alpha\%$ percentage point of the distribution of SS'_2/SS'_1, e.g. the limits of the 95% confidence interval are the 2.5% and 97.5% points of this distribution. Cases where m is 0 or 1 and SS'_2/SS'_1 is undefined are excluded.

Taking sums of squares about class means rather than population means leads to a confidence interval that is slightly too wide, but asymptotically the ratio of its width to the width of the true interval will approach unity.

Similar arguments apply to test 2, where now SS'_1, SS'_2 are sums of squares about sample medians, and r is now not the ratio of variances, but the ratio of mean squared deviations from the respective population medians. However, under a scale shift where $X \to \sqrt{r}(X + \delta)$, the ratio of squared deviations from population medians is equal

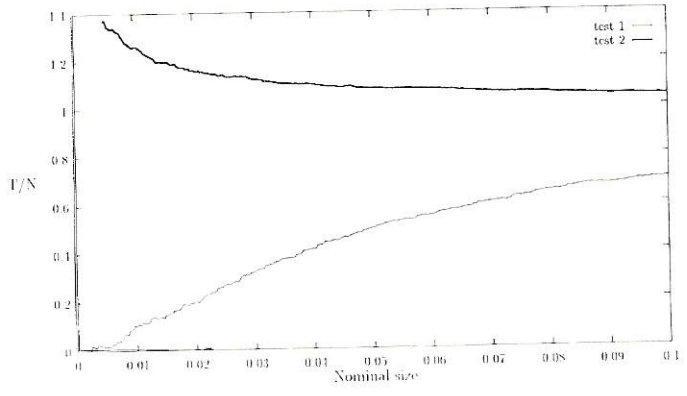

Fig. 3. *Ratio of true size of test to nominal size (T/N) plotted against nominal size, for two samples of sample size 10 from a normal distribution, under H_0. Test 1 is conservative, test 2 slightly liberal*

to the ratio of variances. Note that only the $m = 0$ case need now be excluded.

In general, the confidence intervals for r will be in error if the two distributions differ also in their higher moments.

5. Test performance

Monte Carlo simulations were used to study the tests' performance. In power calculations the P value of a test must be calculated many times to estimate the test power as the fraction of simulated datasets for which H_0 is rejected. Each P value calculation itself requires many simulations. The optimum division of effort between these inner and outer loops was taken from Oden (1991).

Figure 3 shows the ratio T/N of true size to nominal size for the permutation test with 10 observations per group. Test 1 is conservative and test 2 is slightly liberal.

Figure 4 shows the power of the permutation and parametric tests as a function of σ_2^2/σ_1^2, at sample size 10, and Fig. 5 shows the same power curves at sample size 40. At this sample size, already the permutation tests are nearly as powerful as the parametric test.

6. Conclusions

Permutation F-ratio tests offer powerful non-parametric tests of equality of variances which do not suffer from the drawback of assuming that both groups have the same centre of location. Unlike the permutation test cited by Good (1994), the tests proposed here do not require equal sample sizes. The tests are asymptotically exact and have ARE of unity. It is easy to calculate confidence intervals on the ratio of population variances. There is a multi-sample extension, but it lacks the computational simplicity of the two-sample test.

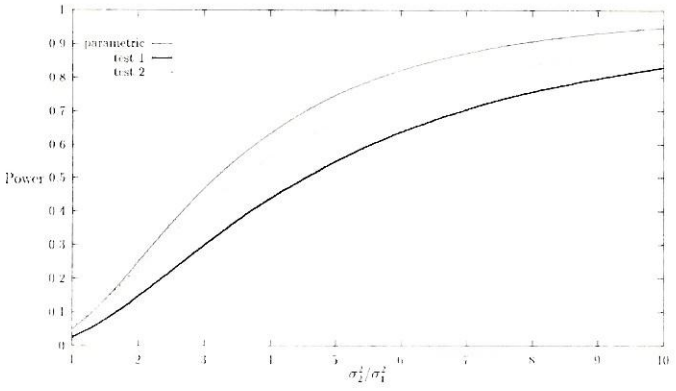

Fig. 4. *Power of both permutation tests and the parametric test for normal distributions, as a function of σ_2^2/σ_1^2, the ratio of group 1 to group 2 variances, for sample size 10 and a significance level of $\alpha = 0.05$. Test 2 is more powerful than test 1*

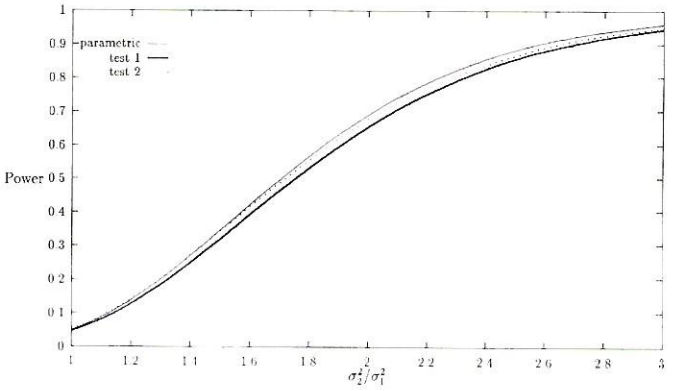

Fig. 5. *Power of permutation and parametric tests as a function of σ_2^2/σ_1^2, the ratio of group 1 to group 2 variances, for sample size 40 and a significance level of $\alpha = 0.05$. The permutation tests are now very nearly as powerful as the parametric test*

Simulation studies show that test 1 is always conservative, and so loses power for small sample sizes and for small significance levels α. Test 2 has greater power than test 1 for small sample sizes, but achieves this at the cost of sometimes being slightly liberal. Which test is preferable will depend on the application. When the P value must be above reproach (e.g. in a drug trial) test 1 is preferable, but for routine testing when power is important and it is acceptable to reject H_0 slightly too often, test 2 seems preferable. For large samples both tests perform similarly.

When distributions are very long tailed, the variance may be infinite, e.g. for the Cauchy distribution. These tests would not then be appropriate.

The stratification method of carrying out Monte Carlo estimation of test P values offers an alternative to methods based on importance sampling (Mehta and Patel, 1988). Importance sampling can only be used when there is a good normal-theory approximation to the P value, as is the case for two-sample tests of location. Stratification does not use this information, but relies on the existence of a stratifying variable correlating with the test statistic. There is scope for further work in applying this method to such statistical tests. With ingenuity, suitable stratifying variables closely related to the test statistic may be found.

References

Bartlett, M. S. (1937) Properties of sufficiency and statistical tests. *Proceedings of the Royal Society* Series A, **160,** 268–82.

Box, G. E. P. (1953) Non-normality and tests on variances. *Biometrika*, **40,** 318–35.

Diciccio, T. J. and Romano, J. P. (1988) A review of bootstrap confidence intervals (with discussion). *Journal of the Royal Statistical Society*, Series B, **50,** 338–70.

Duran, B. S. (1976) A survey of nonparametric tests for scale. *Communications in Statistics—Theory and Methods*, **A5,** 1287–312.

Fisher, R. A. (1924) On a distribution yielding the error functions of several well known statistics. *Proc. Int. Congress Math., Toronto*, **2**, 805–13.

Gabriel, K. R. and Hall, W. J. (1983) Rerandomisation inference on regression and shift effects: computationally feasible methods. *Journal of the American Statistical Association*, **78**, 827–36.

Gibbons, J. D. and Chakraborti, S. (1992) *Nonparametric Statistical Inference*, 3rd edn. Marcel Dekker, New York.

Good, P. (1994) *Permutation Tests*. Springer-Verlag, New York.

Hinkley, D. V. (1988) Bootstrap methods. *Journal of the Royal Statistical Society* Series B, **50**, 321–37.

Mehta, C. R., Patel, N. R., Senchaudhuri, P. (1988) Importance sampling for estimating exact probabilities in permutational inference. *Journal of the American Statistical Association*, **83**, 999–1005.

Moses, L. E. (1963) Rank tests for dispersion. *Annals of Mathematical Statistics*, **34**, 973–83.

Oden, N. L. (1991) Allocation of effort in Monte-Carlo simulation for power of permutation tests. *Journal of the American Statistical Association*, **86**, 1074–6.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (1992) *Numerical Recipes in C: the Art of Scientific Computing*, 2nd edn. Cambridge University Press, New York.

Rivest, L. P. (1986) Bartlett's, Cochran's and Hartley's tests on variances are liberal when the underlying distribution is long-tailed. *Journal of the American Statistical Association*, **81**, 124–8.

Tritchler, D. (1984) On inverting permutation tests. *Journal of the American Statistical Association*, **79**, 200–7.

Appendix A

This appendix describes a stratified method of Monte Carlo calculation based on the use of Siegel–Tukey ranks.

Siegel–Tukey ranks are assigned by subtracting the sample means from each group, and assigning ranks to the transformed observations by working progressively inwards towards the combined median (Gibbons and Chakraborti, 1992). In the Siegel–Tukey test, a scale test results from performing a Mann–Whitney U test on these ranks. As it is not clear how to generate random samples with a specific average group 1 Siegel–Tukey rank, the median of the group 1 ranks is used instead.

Appendix B describes the calculation of the distribution of median Siegel–Tukey rank and the generation of random variates of specific median rank. The remainder of this appendix is devoted to an explanation of the procedure for iteratively sampling from the selected strata. Such a procedure is needed because the probabilities p_i which determine stratum sample size are not known in advance. It is necessary to sample from strata in such a way as to avoid overmuch sampling from unimportant strata (with very low or very high values of p_i) in the early stages, whilst not ruling out important strata on the basis of early small-sample results which might be misleading. It is also necessary to limit the number of strata. For example, the median rank can vary from $n1 - \text{int}(n1/2)$ to $n1 + n2 - \text{int}(n1/2)$, so taking $n_2 + 1$ values. For large sample sizes, if the total number of simulations N remains fixed, not all strata can be sampled. There is also an overhead in computer time in sampling from a fresh group of strata, which is another reason why the total number of strata sampled must be limited to the most useful.

The algorithm adopted proceeds as follows: note that with the assignment of ranks adopted, low group 1 median rank i implies low probability p_i. Because the 'strata' used in the simulations comprised a range of values of the median, these ranges are called 'groups of strata' and the term 'stratum' is reserved for a possible value of the median.

1. Divide the total number of samples N into G (say 50) groups. Allocate the first batch of N/G samples equally between the groups of strata below and equal to or above the observed median Siegel–Tukey rank. Generate random samples from each group, and record numbers of samples and 'hits' (samples with condition $SS'_1 \geq SS'_2$ satisfied) *per stratum*. Let $N_0 = N/G$ be the total number of samples made so far. The remaining steps are looped over G times.

2. Find the number of samples N_i and 'hits' H_i for each of the k groups. If this is the Gth cycle round this loop, calculate the estimated probability $p = \sum_{i=1}^{k} f_i H_i/N_i$ and its estimated standard error σ, where $\sigma^2 = \sum_{i=1}^{k} f_i^2 (H_i/N_i)(1 - H_i/N_i)/N_i$, and stop.

3. Calculate the 'Bayesian' probability $Pr_i = (H_i + 0.1)/(N_i + 0.2)$. Find $g_i = f_i\{Pr_i(1 - Pr_i)\}^{1/2}$. Allot $N_0 + N/G$ samples to the k groups, the ith group receiving a sample size N_i proportional to g_i.

4. Divide each group of strata into two if necessary. Groups will be divided if all the following conditions hold:

(a) there is more than one stratum per group

(b) strata in the group do *not* contain either no 'hits' or no 'misses', i.e. the value of p_i estimated from each stratum is not uniformly zero or unity

(c) $kg_i/\sum_{i=1}^{k} g_i > 0.005$.

Groups are divided by

(a) splitting off the set of all strata with $p_i = 0$ from the left, or, if there are none

(b) splitting off the set of all strata with $p_i = 1$ from the right, or, if there are none

(c) splitting the group into two as evenly as possible.

The allocation of samples to the original group is split evenly between the two new groups.

5. Count the number of samples already made in each new group (the sum of the number of samples recorded per component stratum). The shortfall of this from the number of samples allocated gives the number of samples S_i to be made per group.

6. Sample from each group S_i times.

The purpose of using the 'Bayesian' estimate of p_i in allocating sample sizes to groups is to avoid giving a zero sample size to a group currently with a zero estimated variance.

The method of splitting groups splits off unimportant left and right tails with very small g_i values, which are then not further subdivided.

Appendix B

The probability distribution of the median group 1 (Siegel–Tukey) rank is derived, and the generation of random samples with given median rank described.

When the number of observations n_1 is even, the 'median' rank M is taken as the rank of the $n_1/2$th observation. In general, consider an order statistic with n_a observations of higher rank and n_b of lower rank. The total of $n_a + n_b + 1 = n_1$ group 1 observations are to be assigned ranks from 1 to $n_1 + n_2 = N$. The probability that the median observation is assigned rank m is $1/N$. The probability that n_a observations take ranks p such that $m < p \leq N$ is

$$\frac{(N-m)}{(N-1)}\frac{(N-m-1)}{(N-2)}\cdots\frac{(N-m-n_a+1)}{(N-n_a)} = \frac{(N-m)!(N-n_a-1)!}{(N-m-n_a)!(N-1)!}.$$

The probability that n_b observations are assigned ranks $1 < p < m$ is

$$\frac{(m-1)}{(N-n_a-1)}\frac{(m-2)}{(N-n_a-2)}\cdots\frac{(m-n_b)}{(N-n_a-n_b)} = \frac{(m-1)!(N-n_a-n_b-1)!}{(m-n_b-1)!(N-n_a-1)!}.$$

The observations can be chosen in

$$\frac{(n_a+n_b+1)!}{n_a!n_b!}$$

ways, and so the probability of 'median' rank m is, on cancelling a common factor,

$$\Pr(M=m) = \frac{(N-m)!(m-1)!(n_a+n_b+1)!}{(N-m-n_a)!N!(m-n_b-1)!n_a!n_b!}.$$

The fact that $\sum_{m=n_b+1}^{n_1-n_a} P_m = 1$ for all n_a, n_b can be exploited to find the mean and variance of the group 1 median, by considering the factorial moments $E\{M-n_b-1\}$ and $E\{(M-n_b-1)(M-n_b-2)\}$ and reinterpreting the expressions obtained in terms of probabilities with n_b increased by 1 or 2 respectively. One thus obtains

$$E(M) = \frac{(n_b+1)(N+1)}{n_a+n_b+2}$$

and

$$\mathrm{Var}(M) = \frac{(n_a+1)(n_b+1)(N+1)(N-n_a-n_b-1)}{(n_a+n_b+2)(n_a+n_b+3)}.$$

Random sampling of a value of M within a group of strata, which latter is a range of values of M, can now be achieved by the probability-integral transform method. Once the 'median' or any order statistic has been selected, the n_b ranks above m may be chosen randomly from the $N-m$ ranks available, and the n_a ranks below m chosen randomly from the ranks $1 \cdots m-1$.