Sankhyā: The Indian Journal of Statistics
1995, Volume 57, Series B, Pt. 2, pp. 200-222

RANDOMIZATION TESTS TO COMPARE MEANS WITH UNEQUAL VARIATION

By BRYAN F. J. MANLY
University of Otago

SUMMARY. A randomization framework for testing for significant differences between the means of two or more samples is proposed for the situation where the samples may be from distributions with different variances. This framework is based on the concept that the observed data arise from a random allocation of a fixed set of values to the samples, followed by linear transformations that are not necessarily the same for each sample. The null hypothesis is that, with respect to the distributions generated by the random allocation, the expected values of the sample means are equal but the expected values of the sample variances may or may not be equal.

The model leads in an obvious way to a randomization test that is exact if the parameters for the linear transformations are known. When these parameters are not known (as is usually the case) three algorithms for approximate randomization tests are proposed. The properties of these algorithms have been studied by simulation in comparison with Welch's test, the usual randomization F-test, and the usual F-test using tables. It has been found that two of the three algorithms for approximate randomization tests have better properties than the other four tests when the null hypothesis is true, for data from uniform and normal distributions. None of the tests perform well with data from exponential distributions, but one of the approximate randomization tests is superior to all of the other tests under most of the conditions simulated.

1. Introduction

The usual randomization test for a significant difference between the mean values of two samples involves comparing the observed mean difference with the distribution generated by randomly allocating the data values to the two samples $R-1$ times. The observed mean difference is then declared to be significant at the $100\alpha\%$ level on a two-sided test if the absolute value of this difference is among the largest $100\alpha\%$ of the R absolute differences consisting of that from the observed data and those obtained by randomization (Fisher, 1936; Manly, 1991, p. 49).

AMS (1990) *subject classification.* 62G09.
Key words and phrases. Computer intensive statistics, permutation test, Behrens–Fisher problem, comparison of means, analysis of variance.

The null hypothesis for this test is that the data values observed were randomly allocated to the two samples, either as a result of an experimental manipulation or because the two samples came from the same source, and the test statistic is chosen to be sensitive to alternative hypotheses that lead to one sample tending to contain larger values than the other. With small samples the observed test statistic can be compared with the full randomization distribution. This becomes impractical with moderate to large sample sizes but an exact test is still possible if the observed test statistic is compared with a sample from the randomization distribution (Dwass, 1957). 'Exact' here is in the sense of having the correct size when the null hypothesis is true, although the outcome of a test will depend on the finite set of randomizations used.

Randomization methods can also be used to compare the means of more than two samples. One way to do this is to compute the F-value from a one factor analysis of variance, and compare this with $R-1$ alternative F-values, each of which is obtained by randomly allocating the observed data values to the samples. The F-value for the sample data is then significant at the $100\alpha\%$ level if it is among the largest $100\alpha\%$ of the set of F-values consisting of itself and the $R-1$ randomized values. Again this is an exact test because it has the correct size when the null hypothesis is true (Manly, 1991, p. 64).

Randomization tests have much to commend them as alternatives to standard parametric and non-parametric tests because they (a) do not require any particular assumptions about the distribution of the variable being tested; (b) use the original data values rather than just their ranks; and (c) have no difficulty with handling tied data values (Edgington, 1987; Manly, 1991). Furthermore, the primitive argument behind a randomization test is easily understood and convincing to those not familiar with statistical methods. For this reason, randomization tests should find favour in environmental applications where lack of assumptions and simplicity are key considerations.

However, a source of concern for some users of randomization tests to compare the means of two or more samples is the fact that the probability of a significant result on a test at the $100\alpha\%$ level may not equal α if the samples being considered are drawn from distributions with the same mean values but differing in some other respect. In particular it is known that randomization tests are not necessarily robust if the samples come from sources with the same mean values but different amounts of variation (Boik, 1987).

The comparison of the means of two samples that may come from populations with unequal variances has a long history under the name of the 'Behrens-Fisher problem'. Many solutions have been proposed over the years, including those of Behrens (1929) and Fisher (1939) themselves, and the alternatives of Welch (1937), James (1954), Barnard (1984), Cressie and Whitford (1986), Beran (1988), Schemper (1989), and Asiribo and Gurland (1989). The problem of comparing several means has also received much attention, with solutions being proposed by James (1951, 1954), Welch (1951), Brownie *et al.* (1990), Fisher and Hall (1990), and Shiraishi (1993), among others.

In this note a variation of the usual randomization test to compare two or more sample means is proposed that overcomes the problem of unequal population variances to a large extent. It involves assuming that the observations in the samples come from a random allocation of a fixed initial set of values to the samples, followed by unknown linear transformations of the observations in the different samples. The null hypothesis tested is that for the initial set of values the linear transformations do not change the expected values of the sample means with respect to the randomization distribution, although they may change the sample variances.

To be more precise, it is assumed that when the null hypothesis is true the observations that are available arose from a mechanism that is effectively as follows:

(a) fixed values $u_1, u_2, \ldots, u_n$ with mean $\mu = \Sigma u_i/n$ and variance $\mathrm{Var}(U) = \Sigma(u_i - \mu)^2/n = 1$ are assigned at random to s samples with sizes $n_1, n_2, \ldots, n_s$, such that $n_1 + n_2 + \cdots + n_s = n$; and

(b) the observations $x_{i1}, x_{i2}, \ldots, x_{in_i}$ in the i-th sample are obtained by a linear transformation $X = A_i + B_i U$ of the U values assigned to this sample that is not expected to change the sample mean but may be expected to change the sample variance.

Assuming that this is the case, the essential idea for a randomization test is that a test statistic should be used that makes an allowance for the possibility of variance changes, and the observed values of this statistic should be compared with the distribution that is generated by the random assignment at step (a). Because there is not enough information to do this exactly, some approximations to this ideal procedure are proposed and investigated. However, before considering these approximations it is useful to consider an exact test.

2. An exact randomization test

In order for the transformation $X = A_i + B_i U$ that is applied for the i-th sample to not be expected to change the mean with respect to the randomization distribution it is necessary that $E(X) = \mu = A_i + B_i\mu$. Hence

$$A_i = \mu(1 - B_i). \qquad \ldots(2.1)$$

The transformation from U to X for the i-th sample must therefore have the form

$$X = \mu + B_i(U - \mu), \qquad \ldots(2.2)$$

and the reverse transformation must have the form

$$U = \mu + (X - \mu)/B_i. \qquad \ldots(2.3)$$

Furthermore, using the definition $\mu = \Sigma u_i/n$ together with the last equation shows that

$$\mu = \sum_{i=1}^{s}(n_i\overline{x}_i/B_i)/\sum_{i=1}^{s}(n_i/B_i) \qquad \ldots(2.4)$$

holds exactly, where $\overline{x}_i$ is the observed mean for sample i.

It follows that if $B_1, B_2, \ldots, B_s$ are known then an exact randomization test can be carried out by using equations (2.3) and (2.4) to convert the observed X values back to U values, and applying the test to the untransformed data. There are various ways in which this might be done (Manly, 1991, Chapter 4). An obvious possibility is to compare the F value from a one factor analysis of variance on the U values with the distribution of F values obtained by randomly allocating the same U values to the s samples.

3. THREE APPROXIMATE RANDOMIZATION TESTS WHEN B_i VALUES ARE UNKNOWN

In reality B_i will usually not be known. However, one way to estimate it is by using the variance of the i-th sample, so that

$$\widehat{B}_i^2 = \sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2/(n_i - 1). \qquad \ldots(3.1)$$

The test based on estimating B_i from this equation, untransforming the observed data using equation (2.3), and then carrying out a randomization F-test on the untransformed data will be referred to as Test 1.

Another method for estimating the B_i values is based on the condition that $Var(U) = 1$. This gives the exact equation

$$\sum_{i=1}^{s}\sum_{j=1}^{n_i}(u_{ij} - \mu)^2/n = 1$$

where $u_{ij} = \mu + (x_{ij} - \mu)/B_i$ is the j-th U value that is randomly assigned to sample i. This implies that

$$\sum_{i=1}^{s}\sum_{j=1}^{n_i}(x_{ij} - \mu)^2/B_i^2 = n$$

exactly, and that with respect to the randomization distribution

$$E\{\sum_{j=1}^{n_i}(x_{ij} - \mu)^2/B_i^2\} = n_i.$$

It follows that

$$B_i^2 = E\{\sum_{j=1}^{n_i}(x_{ij} - \mu)^2/n_i\}$$

which shows that an estimate $\widehat{B}_i$ of B_i can be obtained by solving the set of equations

$$\widehat{B}_i = \sqrt{\{\sum_{j=1}^{n_i}(x_{ij} - \widehat{\mu})^2/n_i\}}, \tag{3.2}$$

for $i = 1, 2, \ldots, s$, together with

$$\widehat{\mu} = \sum_{i=1}^{S}(n_i\bar{x}_i/\widehat{B}_i)/\sum_{i=1}^{S}(n_i/\widehat{B}_i), \tag{3.3}$$

based on equation (2.4).

Equations (3.2) and (3.3) can be solved iteratively by starting with $\widehat{B}_i = 1$ and successively applying equation (3.3), followed by equations (3.2), followed by equation (3.3), etc. This process has always converged with simulated and real data.

On the basis of these equations the following approximate randomization procedure can be applied:

(a) estimate $B_1, B_2, \ldots, B_s$ and μ using equations (3.2) and (3.3);

(b) calculate estimates of the U values using $\widehat{u}_{ij} = \widehat{\mu} + (x_{ij} - \widehat{\mu})/\widehat{B}_i$;

(c) carry out a one factor analysis of variance to find F_1, the ratio of the between sample mean square to the within sample mean square for the estimated U values;

(d) randomly reallocate the estimated U values to the samples and calculate F_2, the test statistic for the modified data;

(e) repeat step (d) $R-1$ times to generate further values $F_3, F_4, \ldots, F_R$ from the randomization distribution of the test statistic; and

(f) declare F_1 to be significantly large at the $100\alpha\%$ level if it is larger than 100 (1 - α)% of the set of values $F_1, F_2, \ldots, F_R$.

The hope is that this approximate test will give a probability of approximately α of yielding a significant result when in fact the U values for the real data are randomly allocated to samples. It will be referred to as Test 2.

An alternative test that will be called Test 3 also suggests itself. This is still based on equations (3.2) and (3.3), but includes the following modified version of step (d) in the algorithm:

(d') randomly reallocate the estimated U values to the samples, transform them to X values using equation (2.2) with the estimated values for $B_1, B_2, \ldots, B_s$ and μ obtained from the real data, solve equations (3.2) and (3.3) for these X values, untransform the X values using the new estimates of

$B_1, B_2, \ldots, B_s$ and μ to produce new U values, and calculate the test statistic for these U values.

This makes the algorithm more complicated and computer intensive, but it does have the merit of repeating on the randomized data exactly the same analysis as was carried out on the original data. It appears, therefore, that it may better capture the variation in F-values due to the process of estimating the B and μ values.

Note that when there are two samples of the same size a little algebra shows that equations (3.1) and (3.2) give exactly the same estimated B_i values. Thus Test 1 and Test 2 are the same under these conditions.

4. Simulation experiment 1

The small sample performance of Tests 1 to 3 has been examined by a simulation study. At the same time the opportunity has been taken of comparing the properties of these tests with three other tests for a difference between two sample means: Test 4 is Welch's (1951) test; Test 5 involves assessing the observed F-ratio by comparison with the distribution obtained from randomly reallocating the observations to samples (Manly, 1991, p. 64); and Test 6 involves assessing the significance of the observed F-ratio using tables. Test 4 makes an allowance for unequal sample variances, but Test 5 and Test 6 assume that this does not occur. The randomization tests do not make the assumption that data are normally distributed. The other tests do make this assumption.

The simulation was in two parts. The first part, which is considered in this section, involved data with two samples only. The second part, which is considered in the next section, involved four samples.

For the two sample simulations a four factor experimental design was used with the following factor levels :

(A) Sample sizes (n_1, n_2) were (12,4), (8,8), (24,8), or (16,16).

(B) Data were generated independently for the two samples from the uniform distribution, the normal distribution, or the exponential distribution, scaled in each case so that the mean and variance were 0 and 1. The n_2 values in the second sample were then subjected to a linear transformation to give the mean and variance with respect to the randomization distribution that were required by the levels of factors C and D.

(C) The mean of the second sample with respect to the randomization distribution was 0, 0.5, or 1.

(D) The standard deviation of the second sample with respect to the randomization distribution (B_2) was 0.125, 0.25, 0.5, 1, 2, 4 or 8.

For each combination of the factor levels 500 independent sets of data were generated. Tests to compare the two sample means were then carried out at the 5% level for each of the six tests being considered. For the randomization tests, 999 randomizations were carried out.

In order to generate simulated sets of data, values from the uniform distribution between 0 and 1 were generated using Wichmann and Hill's (1982) algorithm. These were then used directly for the first level of factor B, sums of 12 values were used for generating normal variables, and the transformation $U = -\log_e(R)$ was used for generating exponential variables.

The percentages of significant results obtained from the simulations are recorded in Table 1. For help in comparing the six tests, cases where the results can be considered to be satisfactory are outlined. For example, when the sample sizes were $n_1 = 12$ and $n_2 = 4$, with data from uniform distributions, only Test 2 and Test 3 are outlined as being satisfactory.

'Satisfactory' in this context is with respect to the behaviour of the tests when the null hypothesis $\mu_1 = \mu_2$ is true. With 500 sets of simulated data, binomial sampling errors are such that if the probability of a significant result is 0.05 for each set then the overall percentage of significant results from a test will be approximately normally distributed with a mean of 5% and a standard error 1%. Therefore, with a probabilty of about 0.99 the observed percentage of significant results will lie within the range 5 ± 2.58 (1), or about 2.4% to 7.6%. On this basis it is reasonable to consider the range from 2% to 8% to be the maximum that is acceptable from simulation results, with percentages rounded to the nearest integer. It is for this reason that Test 2 and Test 3 are outlined with samples sizes of 12 and 4 and uniform data: for both tests the percentage of significant results ranges from 4% to 7% with the null hypothesis true. The other tests give percentages that are either less than 2% or more than 8%. For example, Test 6 gives from 0% to 25% significant results, which is completely unsatisfactory.

Table 1 shows that Test 2 and Test 3 have given performances with uniform and normal data, but the same cannot be said of the other four tests. With exponential data none of the tests has been very good: with unequal sample sizes Test 2 and Test 3 give the best performance, but Tests 4, 5 and 6 are better with equal sizes.

The results overall can be summarised as follows:

(a) If sample sizes are quite different, then Test 2 or Test 3 should be used, although these will tend to give too many significant results with exponential types of distributions.

(b) If sample sizes are equal then the best test depends on the nature of the data. With uniform or normal data Tests 1, 2 and 3 are best, but with exponential data Tests 4, 5 and 6 are better.

5. Simulation experiment 2

In order to examine the performance of the approximate randomization tests in situations with more than two samples, a second simulation experiment with a factorial design has been carried out, involving the comparison of the means of four samples. The factors considered were:

(A) Sample sizes were (8,6,4,2), (5,5,5,5), (12,10,8,6), (9,9,9,9), (16,14,12,10) or (13,13,13,13), for samples 1 to 4, in order.

(B) Data values were generated from the uniform, normal or exponential distribution.

(C) Sample means were (0,0,0,0), (0,0.25,0.5,1), or (0,0.5,1,2), for samples 1 to 4, in order.

(D) Sample standard deviations were (1,0.5,0.25,0.125), (1,1,0.5,0.25), (1,1,1,1), (1,1,2,4) or (1,2,4,8) for samples 1 to 4, in order, where these standard deviations correspond to the values B_1 to B_4 for the randomization model that has been described in Section 1.

For each of the $6\times3\times3\times5 = 270$ combinations of the levels of the four factors, 500 sets of four sample data were generated. Each set was then tested for a significant difference between the sample means using Tests 1 to 6 as with the first simulation experiment. In all cases the 5% level of significance was used. For the randomization tests 999 randomizations were carried out.

The results of the simulation experiment are summarised in Table 2, which shows the percentage of significant results obtained for each of the six tests, for each of the combinations of factor levels. In addition, for each of the six levels of sample sizes and each of the three distributions the tests with satisfactory behaviour when the null hypothesis is true (between 2% and 8% significant results, inclusive) are indicated by being outlined in the table, in the same way as was done with Table 1.

With uniformly distributed data Test 1 was never satisfactory, Test 2 and Test 3 were always satisfactory, Test 4 was satisfactory except with the smallest sample sizes, Test 5 was never satisfactory, and Test 6 was only satisfactory with sample sizes of (13,13,13,13). Test 2 and Test 3 had quite similar behaviour but, if anything, Test 3 has shown slightly more power when the null hypothesis was not true.

None of the tests performed very well with exponentially distributed data. However, inspection of the results shows that Test 2 and Test 3 generally performed better than the other tests when sample sizes were equal. Both Test 2 and Test 3 have a tendency to produce too many significant results when the null hypothesis is true but, if anything, Test 3 gives slightly better results in this respect.

6. Example

The simulations suggest that Test 3 is the most reliable of those that have been compared in the situation where samples may have unequal variances and observations are from an unknown distribution that is likely to be quite non-normal.

As an example of the practical use of this test, consider the example used by Welch (1951). This involves three samples, with sizes $n_1 = 20$, $n_2 = 10$ and $n_3 = 10$. The sample means are $\overline{x}_1 = 27.8$, $\overline{x}_2 = 24.1$ and $\overline{x}_3 = 22.2$, and the sample variances are $s_1^2 = 60.1$, $s_2^2 = 6.3$ and $s_3^2 = 15.4$. Welch shows that his test (Test 4 in the simulations) leads to the comparison of the statistic 3.35 with the F-distribution table with 2 and 22.6 degrees of freedom, corresponding to a p-value of 0.053. Hence, the test is not quite significant at the 5% level.

Test 3 can only be carried out when the individual observations in the data are known. However, these were not provided by Welch in his paper. To demonstrate Test 3 the following data were therefore generated on a computer from a normal distribution and scaled to give the sample means and standard deviations quoted by Welch:

Sample 1. 32.0 29.6 23.5 29.5 22.8 25.3 26.7 18.0 33.0 17.7 24.8 24.3 21.6 15.2 25.5 40.9 27.6 34.0 41.0 43.1

Sample 2. 25.5 25.4 26.9 19.7 28.0 21.9 22.2 24.1 24.6 22.8

Sample 3. 24.7 22.7 28.4 20.9 13.4 20.8 23.8 21.5 20.6 25.1

Solving equations (3.2) and (3.3) with these data gives the estimates $\widehat{B}_1 = 8.18$, $\widehat{B}_2 = 2.44$, and $\widehat{B}_3 = 4.49$, with $\widehat{\mu} = 24.65$. The estimated untransformed U values based on equation (2.3) are then as shown below:

Sample 1. 25.55 25.26 24.51 25.24 24.43 24.73 24.90 23.84 25.67 23.80 24.67 24.61 24.28 23.50 24.76 26.64 25.01 25.79 26.65 26.91

Sample 2. 25.00 24.96 25.57 22.62 26.02 23.52 23.65 24.43 24.63 23.89

Sample 3. 24.66 24.22 25.49 23.82 22.15 23.79 24.46 23.95 23.75 24.75

An analysis of variance gives an F-ratio of $F_1 = 3.58$.

Carrying out 99,999 randomizations of these estimated U values for Test 3 resulted in 3.0% of values greater than or equal to 3.58 in the set of 100,000 F-values comprising the one for the real data and those from the randomized sets of data. The test therefore gives a result that is significant at the 3.0% level, and there is slightly more evidence for a difference between the means for this test than is obtained from Welch's test.

Of course, there is no reason why Welch's test and the randomization test should give quite the same level of significance. There is therefore nothing that needs to be said about the difference of about 2%. If anything, the simulation results suggest that with the unequal and moderate sample sizes being considered the two tests both have reasonable performance except with highly skewed distributions like the exponential.

A question of some interest is the extent to which the distribution of observations affects the significance level for the randomization test for given sample means and standard deviations. This has been investigated briefly in the context of the example just considered, by generating data from a range of different distributions as well as normal. The significance levels obtained from Test 3 with 99,999 randomizations were found to be as follows: uniform, 3.6%, triangular, 3.3%, normal, 3.0%, chi-squared with 4 degrees of freedom, 3.3%; exponential, 3.8%; chi-squared with 1 degree of freedom, 4.6%. Sampling errors (99% confidence limits) are in each case within the range of the obtained percentage $\pm$ 0.1%. It does therefore appear that the distribution of the data has a small impact on the significance level.

7. Conclusion

Overall the simulation experiments indicates that Test 3 (as defined in Section 3) tends to have properties that are as good as or better than the other five tests for situations where data come from an unknown distribution and the expected values of sample variances may differ in unpredictable directions. However, with exponentially distributed data this test has a clear tendency to give too many significant results when the null hypothesis is true.

In essence, the simulation results suggest that for normally distributed data Test 3 solves the problem of comparing the means of two or more samples with possibly different variances, even with sample sizes that are quite small. This test has some robustness against non-normality but it appears that the probability of a significant result will be somewhat too high when the null hypothesis is true but the data are from highly skewed distributions like the exponential.

Further simulation studies are needed to examine the performance of Test 3 compared to alternatives under a wider range of conditions. These are planned in the near future.

TABLE 1. RESULTS FROM SIMULATION STUDY 1 TO COMPARE SIX TESTS FOR A SIGNIFICANT DIFFERENCE BETWEEN TWO SAMPLE MEANS
THE OUTLINES SHOW THE TESTS WITH SATISFACTORY BEHAVIOUR WHEN THE NULL HYPOTHESIS IS TRUE

Sample size		Sample mean		Sample SD		Uniform data % Significant results from Tests					
1	2	1	2	1	2	1	2	3	4	5	6
12	4	0.0	0.0	1	0.125	5	5	5	1	0	0
				1	0.25	8	5	6	2	0	0
				1	0.5	7	4	4	4	0	1
				1	1	12	7	7	9	7	7
				1	2	10	6	6	7	16	15
				1	4	11	5	4	5	21	21
				1	8	10	5	5	5	27	25
		0.0	0.5	1	0.125	53	46	47	22	1	1
				1	0.25	38	25	28	22	4	3
				1	0.5	22	14	14	18	4	5
				1	1	16	9	8	12	13	13
				1	2	11	6	6	8	17	17
				1	4	14	8	8	8	28	27
				1	8	15	8	8	8	32	30
		0.0	1.0	1	0.125	100	100	100	95	27	28
				1	0.25	98	91	92	88	31	30
				1	0.5	68	43	42	59	33	34
				1	1	35	17	17	25	33	33
				1	2	20	9	9	11	30	30
				1	4	14	6	6	6	28	27
				1	8	15	7	6	6	30	28
8	8	0.0	0.0	1	0.125	7	7	7	3	4	4
				1	0.25	6	6	6	4	5	4
				1	0.5	7	7	6	5	5	6
				1	1	7	7	6	6	6	6
				1	2	7	7	7	5	5	6
				1	4	6	6	6	3	4	4
				1	8	4	4	3	1	2	1
		0.0	0.0	1	0.125	34	34	31	15	20	20
				1	0.25	31	31	30	18	24	23
				1	0.5	20	20	19	18	19	19
				1	1	14	14	13	13	13	14
				1	2	10	10	9	7	8	8
				1	4	8	8	7	4	5	5
				1	8	5	5	5	1	2	2
		0.0	0.0	1	0.125	90	90	87	66	75	75
				1	0.25	84	84	82	65	70	72
				1	0.5	67	67	65	58	62	61
				1	1	37	37	36	36	35	37
				1	2	19	19	18	15	16	17
				1	4	11	11	9	6	8	8
				1	8	9	9	8	3	3	3

Table 1 (Continued)

Sample size		Sample mean		Sample SD		Uniform data % Significant results from Tests					
1	2	1	2	1	2	1	2	3	4	5	6
24	8	0.0	0.0	1	0.125	6	6	6	1	0	0
				1	0.25	6	5	4	1	0	0
				1	0.5	4	3	3	2	0	0
				1	1	4	3	3	4	3	3
				1	2	8	5	5	5	15	15
				1	4	8	5	5	3	22	22
				1	8	9	6	5	4	24	23
		0.0	0.5	1	0.125	96	96	96	67	5	5
				1	0.25	74	71	72	53	8	9
				1	0.5	40	38	38	37	14	14
				1	1	24	18	18	20	22	22
				1	2	13	9	8	8	23	23
				1	4	8	6	5	4	24	24
				1	8	9	7	6	5	26	25
		0.0	1.0	1	0.125	100	100	100	100	96	97
				1	0.25	100	100	100	100	93	93
				1	0.5	97	96	96	96	77	78
				1	1	66	58	57	61	64	65
				1	2	26	19	17	16	45	44
				1	4	12	8	8	6	29	28
				1	8	11	7	6	4	26	25
16	16	0.0	0.0	1	0.125	4	4	4	1	2	2
				1	0.25	5	5	5	2	2	2
				1	0.5	6	6	5	4	4	4
				1	1	6	6	7	7	6	7
				1	2	4	4	4	2	3	3
				1	4	7	7	6	2	3	3
				1	8	6	6	6	1	2	2
		0.0	0.5	1	0.125	65	65	63	37	42	42
				1	0.25	59	59	59	43	44	46
				1	0.5	43	43	42	37	37	38
				1	1	25	25	24	24	24	24
				1	2	15	15	15	12	13	13
				1	4	9	9	9	6	6	6
				1	8	6	6	6	1	1	1
		0.0	1.0	1	0.125	100	100	100	99	100	100
				1	0.25	100	100	100	99	99	99
				1	0.5	96	96	96	94	95	95
				1	1	71	71	71	71	71	71
				1	2	41	41	41	36	37	37
				1	4	17	17	17	10	11	11
				1	8	10	10	9	3	4	4

Table 1 (Continued)

Sample size		Sample mean		Sample SD		Normal data % Significant results from Tests					
1	2	1	2	1	2	1	2	3	4	5	6
12	4	0.0	0.0	1	0.125	5	4	4	1	0	0
				1	0.25	6	3	4	1	0	0
				1	0.5	7	3	4	4	1	1
				1	1	8	4	5	5	5	5
				1	2	11	4	6	6	16	16
				1	4	11	6	7	7	23	22
				1	8	11	4	5	4	30	30
		0.0	0.5	1	0.125	62	54	60	21	2	1
				1	0.25	43	30	36	22	2	2
				1	0.5	26	15	19	18	6	6
				1	1	14	7	8	9	10	10
				1	2	11	5	6	7	17	18
				1	4	12	6	7	7	24	23
				1	8	14	6	7	5	33	31
		0.0	1.0	1	0.125	100	100	100	99	28	28
				1	0.25	98	88	92	88	26	27
				1	0.5	72	50	58	62	35	36
				1	1	38	20	24	28	33	33
				1	2	20	9	10	10	30	30
				1	4	15	6	7	7	30	29
				1	8	12	5	6	5	31	29
8	8	0.0	0.0	1	0.125	8	8	7	3	5	4
				1	0.25	6	6	5	3	4	3
				1	0.5	5	5	5	3	4	4
				1	1	5	5	5	5	5	5
				1	2	7	7	6	4	5	5
				1	4	7	7	7	3	3	4
				1	8	6	6	5	2	2	3
		0.0	0.5	1	0.125	33	33	30	13	19	17
				1	0.25	30	30	27	17	22	21
				1	0.5	23	23	23	19	21	20
				1	1	15	15	15	14	15	15
				1	2	9	9	9	7	8	8
				1	4	8	8	7	3	5	5
				1	8	6	6	6	2	3	3
		0.0	1.0	1	0.125	89	89	88	69	76	76
				1	0.25	82	82	80	63	70	71
				1	0.5	65	65	63	56	61	60
				1	1	39	39	38	38	38	39
				1	2	22	22	21	18	20	20
				1	4	10	10	9	5	6	6
				1	8	9	9	9	4	5	5

Table 1 (Continued)

Sample size		Sample mean		Sample SD		Normal data % Significant results from Tests					
1	2	1	2	1	2	1	2	3	4	5	6
24	8	0.0	0.0	1	0.125	6	5	6	1	0	0
				1	0.25	6	5	5	2	0	0
				1	0.5	6	5	5	4	1	1
				1	1	6	5	5	5	5	4
				1	2	7	5	4	4	14	14
				1	4	7	5	5	4	21	20
				1	8	7	5	4	3	21	22
		0.0	0.5	1	0.125	94	93	93	65	7	7
				1	0.25	79	77	77	58	12	12
				1	0.5	48	44	45	42	18	17
				1	1	22	19	19	19	22	22
				1	2	17	12	13	11	27	27
				1	4	12	8	8	6	27	26
				1	8	8	7	6	4	24	24
		0.0	1.0	1	0.125	100	100	100	100	97	97
				1	0.25	100	100	100	100	91	91
				1	0.5	96	96	95	95	79	79
				1	1	65	59	60	61	63	63
				1	2	28	23	22	22	45	46
				1	4	15	10	10	7	33	32
				1	8	11	7	7	6	27	26
16	16	0.0	0.0	1	0.125	7	7	6	3	3	3
				1	0.25	5	5	4	2	3	3
				1	0.5	6	6	6	5	5	6
				1	1	7	7	7	6	6	6
				1	2	4	4	4	4	3	4
				1	4	6	6	6	3	4	4
				1	8	4	4	4	1	2	2
		0.0	0.5	1	0.125	62	62	59	36	42	41
				1	0.25	55	55	54	40	43	42
				1	0.5	42	42	41	36	38	37
				1	1	25	25	25	24	25	24
				1	2	14	14	14	11	12	11
				1	4	8	8	8	4	4	5
				1	8	8	8	8	3	3	3
		0.0	1.0	1	0.125	100	100	100	100	100	100
				1	0.25	100	100	100	99	99	99
				1	0.5	97	97	97	95	95	96
				1	1	75	75	75	75	75	75
				1	2	41	41	41	35	38	36
				1	4	18	18	17	11	13	12
				1	8	9	9	9	4	5	5

Sample size		Sample mean		Sample SD		Exponential data % Significant results from Tests					
1	2	1	2	1	2	1	2	3	4	5	6
12	4	0.0	0.0	1	0.125	6	4	3	2	0	1
				1	0.25	10	7	8	3	2	2
				1	0.5	7	5	5	2	2	2
				1	1	13	9	11	8	6	6
				1	2	15	10	11	12	14	14
				1	4	15	11	11	12	24	23
				1	8	17	10	10	9	31	29
		0.0	0.5	1	0.125	61	49	50	28	9	10
				1	0.25	43	24	22	28	13	13
				1	0.5	25	10	8	13	13	13
				1	1	9	2	2	3	12	12
				1	2	13	8	9	9	18	15
				1	4	12	6	5	7	22	22
				1	8	14	8	7	8	33	32
		0.0	1.0	1	0.125	100	100	100	96	30	32
				1	0.25	100	99	95	85	34	34
				1	0.5	80	47	40	56	31	31
				1	1	35	13	11	19	35	34
				1	2	12	4	3	4	25	22
				1	4	11	6	6	7	25	22
				1	8	12	6	6	6	27	26
8	8	0.0	0.0	1	0.125	11	11	10	6	8	7
				1	0.25	9	9	8	7	7	7
				1	0.5	8	8	7	5	7	5
				1	1	7	7	6	3	5	4
				1	2	8	8	8	7	8	7
				1	4	9	9	9	8	9	8
				1	8	9	9	8	6	7	7
		0.0	0.5	1	0.125	39	39	35	27	29	29
				1	0.25	36	36	35	30	32	32
				1	0.5	28	28	27	26	28	27
				1	1	13	13	12	10	13	11
				1	2	9	9	8	5	8	6
				1	4	6	6	5	4	4	4
				1	8	9	9	8	4	6	5
		0.0	1.0	1	0.125	83	84	79	61	68	67
				1	0.25	72	72	66	55	58	58
				1	0.5	63	63	59	54	55	56
				1	1	44	44	43	41	45	43
				1	2	22	22	19	11	19	13
				1	4	7	7	6	2	3	3
				1	8	6	6	6	3	4	4

Table 1 (Continued)

Sample size		Sample mean		Sample SD		Exponential data % Significant results from Tests					
1	2	1	2	1	2	1	2	3	4	5	6
24	8	0.0	0.0	1	0.125	5	5	4	1	0	0
				1	0.25	5	4	4	3	1	1
				1	0.5	7	6	6	3	2	2
				1	1	9	7	7	7	5	4
				1	2	9	6	6	7	14	15
				1	4	12	10	8	9	26	26
				1	8	11	10	8	8	23	23
		0.0	0.5	1	0.125	97	97	96	58	17	16
				1	0.25	80	78	78	54	18	18
				1	0.5	49	41	37	39	22	22
				1	1	19	13	9	12	22	22
				1	2	9	6	5	5	21	20
				1	4	8	5	4	4	20	19
				1	8	11	9	7	6	27	26
		0.0	1.0	1	0.125	100	100	100	100	92	94
				1	0.25	100	100	100	100	87	87
				1	0.5	100	100	100	99	76	76
				1	1	72	63	55	63	61	61
				1	2	23	15	8	11	43	43
				1	4	9	6	4	4	30	30
				1	8	8	6	4	4	23	24
16	16	0.0	0.0	1	0.125	7	7	6	4	4	4
				1	0.25	7	7	6	5	6	6
				1	0.5	5	5	4	4	4	4
				1	1	5	5	5	4	4	4
				1	2	5	5	4	3	4	4
				1	4	6	6	6	5	5	5
				1	8	10	10	9	7	8	8
		0.0	0.5	1	0.125	64	64	58	48	51	49
				1	0.25	55	55	51	43	45	46
				1	0.5	46	46	43	42	43	42
				1	1	31	31	29	29	31	29
				1	2	13	13	11	8	11	9
				1	4	6	6	5	2	3	3
				1	8	5	5	4	2	2	2
		0.0	1.0	1	0.125	100	100	100	96	97	97
				1	0.25	100	100	99	95	96	96
				1	0.5	95	95	94	90	91	91
				1	1	76	76	73	74	75	74
				1	2	41	41	36	33	37	35
				1	4	20	20	16	6	10	8
				1	8	7	7	5	1	2	2

TABLE 2. RESULTS FROM SIMULATION STUDY 2 TO COMPARE SIX TESTS FOR A SIGNIFICANT DIFFERENCE BETWEEN FOUR SAMPLE MEANS
THE OUTLINES SHOW THE TESTS WITH SATISFACTORY BEHAVIOUR WHEN THE NULL HYPOTHESIS IS TRUE.

Sample size				Sample mean				Sample SD				Uniform data % Significant results from Tests					
1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	5	6
8	6	4	2	0	0	0	0	1	0.5	0.25	0.125	24	4	6	4	2	2
								1	1	0.5	0.25	28	3	6	6	3	2
								1	1	1	1	28	3	6	13	5	5
								1	1	2	4	31	3	4	16	32	31
								1	2	4	8	36	6	7	17	37	37
				0	0.25	0.5	1	1	0.5	0.25	0.125	100	16	21	83	9	10
								1	1	0.5	0.25	70	10	14	35	7	7
								1	1	1	1	34	6	9	17	12	13
								1	1	2	4	35	15	7	15	39	38
								1	2	4	8	36	7	9	18	39	39
				0	0.5	1	2	1	0.5	0.25	0.125	100	56	71	100	82	84
								1	1	0.5	0.25	100	55	45	94	51	52
								1	1	1	1	59	14	21	33	43	43
								1	1	2	4	40	8	11	18	41	41
								1	2	4	8	56	7	9	18	39	39
5	5	5	5	0	0	0	0	1	0.5	0.25	0.125	18	4	5	9	11	10
								1	1	0.5	0.25	15	3	4	7	7	7
								1	1	1	1	15	5	6	8	5	6
								1	1	2	4	16	4	6	7	10	9
								1	2	4	8	17	3	5	6	11	10
				0	0.25	0.5	1	1	0.5	0.25	0.125	100	53	60	98	47	47
								1	1	0.5	0.25	77	26	34	42	26	27
								1	1	1	1	29	10	14	16	20	21
								1	1	2	4	15	4	6	7	11	10
								1	2	4	8	16	5	5	7	11	10
				0	0.5	1	2	1	0.5	0.25	0.125	100	90	94	100	100	100
								1	1	0.5	0.25	100	76	83	100	95	96
								1	1	1	1	72	34	42	52	66	58
								1	1	2	4	27	9	11	13	20	19
								1	2	4	8	20	5	8	6	9	9
12	10	8	6	0	0	0	0	1	0.5	0.25	0.125	8	4	4	4	3	3
								1	1	0.5	0.25	11	5	7	6	5	5
								1	1	1	1	9	4	4	6	5	4
								1	1	2	4	10	5	5	4	15	14
								1	2	4	8	11	5	5	6	16	16
				0	0.25	0.5	1	1	0.5	0.25	0.125	100	88	88	100	64	64
								1	1	0.5	0.25	99	61	64	93	35	37
								1	1	1	1	37	20	21	28	32	33
								1	1	2	4	18	8	10	9	24	23
								1	2	4	8	13	5	6	7	22	21
				0	0.5	1	2	1	0.5	0.25	0.125	100	100	100	100	100	100
								1	1	0.5	0.25	100	98	99	100	100	100
								1	1	1	1	92	74	78	87	92	92
								1	1	2	4	35	24	26	21	37	36
								1	2	4	8	22	13	13	10	21	21

Table 2 (Continued)

Sample size				Sample mean				Sample SD				Uniform data % Significant results from Tests					
1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	5	6
9	9	9	9	0	0	0	0	1	0.5	0.25	0.125	9	4	4	3	9	9
								1	1	0.5	0.25	9	6	6	5	7	7
								1	1	1	1	9	5	5	5	6	6
								1	1	2	4	10	6	7	7	9	8
								1	2	4	8	9	5	5	3	9	8
				0	0.25	0.5	1	1	0.5	0.25	0.125	100	100	100	100	92	91
								1	1	0.5	0.25	99	97	97	92	64	63
								1	1	1	1	37	22	23	26	34	34
								1	1	2	4	17	9	10	9	11	11
								1	2	4	8	12	8	8	6	11	10
				0	0.5	1	2	1	0.5	0.25	0.125	100	100	100	100	100	100
								1	1	0.5	0.25	100	100	100	100	100	100
								1	1	1	1	96	90	92	93	96	96
								1	1	2	4	37	26	27	23	29	29
								1	2	4	8	22	15	16	10	13	13
16	14	12	10	0	0	0	0	1	0.5	0.25	0.125	6	4	4	2	4	4
								1	1	0.5	0.25	9	5	6	4	4	4
								1	1	1	1	8	5	6	6	5	5
								1	2	4	8	8	4	4	5	12	12
				0	0.25	0.5		1 1	0.5	0.25	0.125	100	100	100	100	97	97
								1	1	0.5	0.125	100	100	100	100	77	77
								1	1	1	1	54	43	43	46	50	51
								1	1	2	4	17	12	12	12	20	20
								1	2	4	8	12	8	8	7	17	17
				0	0.5	1		2	0.5	0.25	0.125	100	100	100	100	100	100
								1	1	0.5	0.25	100	100	100	100	100	100
								1	1	1	1	100	97	97	99	99	99
								1	1	2	4	47	39	39	32	43	43
								1	2	4	8	26	22	21	15	18	18
13	13	13	13		0 0	0		0	0.5	0.25	0.125	8	7	7	6	9	8
								1	1	0.5	0.25	8	5	5	4	5	5
								1	1	1	1	5	4	3	4	4	4
								1	1	2	4	9	6	6	6	7	7
								1	2	4	8	8	5	5	5	8	8
				0	0.25	0.5		1	0.5	0.25	0.125	100	100	100	100	100	100
								1	1	0.5	0.25	100	100	100	99	86	86
								1	1	1	1	58	50	50	53	55	56
								1	1	2	4	16	11	12	10	13	13
								1	2	4	8	14	11	11	6	10	9
					0.5	1	2	1	0.5	0.25	0.125	100	100	100	100	100	100
								1	1	0.5	0.25	100	100	100	100	100	100
								1	1	1	1	100	99	99	100	100	100
								1	1	2	4	53	45	44	36	36	36
								1	2	4	8	29	24	25	15	13	13

Table 2 (Continued)

Sample size				Sample mean				Sample SD				Normal data % Significant results from Tests					
1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	5	6
8	6	4	2	0	0	0	0	1	0.5	0.25	0.125	17	3	5	3	2	2
								1	1	0.5	0.25	20	3	5	5	3	3
								1	1	1	1	24	3	5	9	6	6
								1	1	2	4	26	4	6	12	31	34
								1	2	4	8	27	3	4	11	36	36
				0	0.25	0.5	1	1	0.5	0.25	0.125	98	19	29	86	15	12
								1	1	0.5	0.25	77	11	16	41	7	7
								1	1	1	1	39	5	10	16	14	14
								1	1	2	4	32	4	7	13	33	32
								1	2	4	8	27	3	6	14	34	35
				0	0.5	1	2	1	0.5	0.25	0.125	100	62	77	100	77	79
								1	1	0.5	0.25	100	38	52	92	53	51
								1	1	1	1	64	16	24	35	43	42
								1	1	2	4	38	8	13	16	42	44
								1	2	4	8	31	5	8	13	34	34
5	5	5	5	0	0	0	0	1	0.5	0.25	0.125	13	4	7	5	12	10
								1	1	0.5	0.25	13	2	4	5	7	7
								1	1	1	1	12	2	5	5	6	6
								1	1	2	4	13	2	6	5	9	8
								1	2	4	8	12	3	4	5	9	9
				0	0.25	0.5	1	1	0.5	0.25	0.125	100	61	70	99	51	49
								1	1	0.5	0.25	83	36	46	50	32	31
								1	1	1	1	26	7	11	16	16	17
								1	1	2	4	13	5	6	6	12	11
								1	2	4	8	12	3	4	5	8	7
				0	0.5	1	2	1	0.5	0.25	0.125	100	94	94	100	100	100
								1	1	0.5	0.25	100	77	84	100	96	96
								1	1	1	1	72	38	48	53	64	64
								1	1	2	4	24	8	12	12	18	17
								1	2	4	8	22	6	9	8	14	11
12	10	8	6	0	0	0	0	1	0.5	0.25	0.125	10	5	6	5	3	3
								1	1	0.5	0.25	11	5	6	6	5	4
								1	1	1	1	10	4	6	4	6	5
								1	1	2	4	7	3	3	3	18	17
								1	2	4	8	10	5	5	4	18	17
				0	0.25	0.5	1	1	0.5	0.25	0.125	100	89	92	100	71	69
								1	1	0.5	0.25	99	62	68	93	35	36
								1	1	1	1	40	23	27	30	31	31
								1	1	2	4	17	9	10	10	23	22
								1	2	4	8	14	6	7	6	20	19
				0	0.5	1	2	1	0.5	0.25	0.125	100	100	100	100	100	100
								1	1	0.5	0.25	100	99	99	100	100	100
								1	1	1	1	92	71	75	84	92	92
								1	1	2	4	38	25	27	23	38	36
								1	2	4	8	21	13	14	9	25	23

Table 2 (Continued)

Sample size				Sample mean				Sample SD				Normal data % Significant results from Tests					
1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	5	6
9	9	9	9	0	0	0	0	1	0.5	0.25	0.125	9	5	5	5	7	7
								1	1	0.5	0.25	16	4	5	4	5	5
								1	1	1	1	8	5	5	5	5	5
								1	1	2	4	10	5	5	4	5	5
								1	2	4	8	7	4	5	3	8	7
				0	0.25	0.5	1	1	0.5	0.25	0.125	100	100	100	100	92	91
								1	1	0.5	0.25	99	96	96	92	66	65
								1	1	1	1	45	32	36	37	39	39
								1	1	2	4	15	9	10	10	13	12
								1	2	4	8	12	7	8	5	9	9
				0	0.5	1	2	1	0.5	0.25	0.125	100	100	100	100	100	100
								1	1	0.5	0.25	100	100	100	100	100	100
								1	1	1	1	95	88	90	93	95	95
								1	1	2	4	38	27	29	25	28	26
								1	2	4	8	20	13	14	8	12	11
16	14	12	10	0	0	0	0	1	0.5	0.25	0.125	6	4	4	3	4	4
								1	1	0.5	0.25	7	5	5	4	5	5
								1	1	1	1	7	4	5	4	5	5
								1	2	4	8	8	5	5	4	13	12
				0	0.25	0.5		1	0.5	0.25	0.125	100	100	100	100	96	96
								1	1	0.5	0.125	100	100	100	100	75	74
								1	1	1	1	55	42	43	48	51	52
								1	1	2	4	16	12	12	11	20	19
								1	2	4	8	14	9	10	7	18	16
				0	0.5	1		2 1	0.5	0.25	0.125	100	100	100	100	100	100
								1	1	0.5	0.25	100	100	100	100	100	100
								1	1	1	1	100	98	99	99	100	100
								1	1	2	4	48	41	42	37	47	46
								1	2	4	8	27	21	22	13	19	18
13	13	13	13		0 0	0		0 1	0.5	0.25	0.125	7	5	5	3	10	9
								1	1	0.5	0.25	7	4	5	4	8	8
								1	1	1	1	6	4	4	4	4	5
								1	1	2	4	8	5	5	5	11	10
								1	2	4	8	7	4	5	3	5	5
				0	0.25	0.5		1	0.5	0.25	0.125	100	100	100	100	100	100
								1	1	0.5	0.25	100	100	100	100	89	89
								1	1	1	1	58	51	52	53	55	57
								1	1	2	4	19	14	14	12	15	14
								1	2	4	8	13	10	10	7	8	8
					0.5	1		2 1	0.5	0.25	0.125	100	100	100	100	100	100
								1	1	0.5	0.25	100	100	100	100	100	100
								1	1	1	1	100	99	100	99	100	100
								1	1	2	4	56	46	47	40	40	40
								1	2	4	8	31	25	25	15	14	14

Table 2 (Continued)

Sample size				Sample mean				Sample SD				Exponential data % Significant results from Tests					
1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	5	6
8	6	4	2	0	0	0	0	1	0.5	0.25	0.125	24	6	9	7	5	5
								1	1	0.5	0.25	22	4	5	6	3	2
								1	1	1	1	33	4	7	12	7	6
								1	1	2	4	38	4	6	22	32	34
								1	2	4	8	39	5	8	20	35	35
				0	0.25	0.5	1	1	0.5	0.25	0.125	100	27	36	91	24	25
								1	1	0.5	0.25	92	19	27	55	13	15
								1	1	1	1	35	8	12	11	12	12
								1	1	2	4	30	5	6	14	28	28
								1	2	4	8	38	3	6	17	29	30
				0	0.5	1	2	1	0.5	0.25	0.125	100	73	83	100	65	64
								1	1	0.5	0.25	100	44	54	97	48	48
								1	1	1	1	71	20	29	37	42	41
								1	1	2	4	33	11	16	12	29	30
								1	2	4	8	35	5	7	17	29	30
5	5	5	5	0	0	0	0	1	0.5	0.25	0.125	23	5	8	13	14	14
								1	1	0.5	0.25	18	3	5	9	6	4
								1	1	1	1	20	5	7	5	5	5
								1	1	2	4	18	3	6	9	12	11
								1	2	4	8	22	4	6	12	9	9
				0	0.25	0.5	1	1	0.5	0.25	0.125	100	74	79	99	53	54
								1	1	0.5	0.25	92	54	62	69	36	35
								1	1	1	1	41	18	25	22	21	18
								1	1	2	4	18	4	6	5	10	8
								1	2	4	8	20	4	6	9	13	10
				0	0.5	1	2	1	0.5	0.25	0.125	100	92	95	100	97	97
								1	1	0.5	0.25	100	78	83	100	84	86
								1	1	1	1	82	50	57	64	64	65
								1	1	2	4	32	11	14	10	14	11
								1	2	4	8	21	7	9	5	10	8
12	10	8	6	0	0	0	0	1	0.5	0.25	0.125	14	8	8	8	5	5
								1	1	0.5	0.25	14	8	8	8	4	4
								1	1	1	1	15	7	7	6	4	5
								1	1	2	4	19	9	9	12	22	21
								1	2	4	8	21	9	7	13	17	16
				0	0.25	0.5	1	1	0.5	0.25	0.125	100	91	89	100	59	59
								1	1	0.5	0.25	100	80	79	98	47	45
								1	1	1	1	42	26	27	26	30	29
								1	1	2	4	18	9	9	5	18	16
								1	2	4	8	15	8	7	8	17	16
				0	0.5	1	2	1	0.5	0.25	0.125	100	100	100	100	100	100
								1	1	0.5	0.25	100	99	98	100	100	100
								1	1	1	1	99	77	79	95	93	93
								1	1	2	4	32	24	23	14	34	31
								1	2	4	8	19	10	10	6	23	19

Table 2 (Continued)

Sample size				Sample mean				Sample SD				Exponential data % Significant results from Tests					
1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	5	6
9	9	9	9	0	0	0	0	1	0.5	0.25	0.125	14	8	7	9	8	8
								1	1	0.5	0.25	10	10	9	10	7	7
								1	1	1	1	12	8	7	6	3	3
								1	1	2	4	13	10	9	8	9	9
								1	2	4	8	15	9	8	10	11	11
				0	0.25	0.5	1	1	0.5	0.25	0.125	100	100	100	100	77	77
								1	1	0.5	0.25	100	100	100	96	63	65
								1	1	1	1	58	45	43	47	41	39
								1	1	2	4	20	14	12	9	12	10
								1	2	4	8	14	10	8	7	9	8
				0	0.5	1	2	1	0.5	0.25	0.125	100	100	100	100	100	100
								1	1	0.5	0.25	100	100	100	100	100	100
								1	1	1	1	100	98	97	98	98	98
								1	1	2	4	39	32	29	23	26	21
								1	2	4	8	20	15	13	4	10	7
16	14	12	10	0	0	0	0	1	0.5	0.25	0.125	15	12	10	10	8	7
								1	1	0.5	0.25	14	10	8	9	5	5
								1	1	1	1	11	8	7	6	5	5
								1	2	4	8	15	10	7	10	15	15
				0	0.25	0.5		1	0.5	0.25	0.125	100	100	100	100	86	87
								1	1	0.5	0.125	100	100	100	100	72	71
								1	1	1	1	62	49	45	51	50	50
								1	1	2	4	15	11	9	7	16	14
								1	2	4	8	12	9	7	5	13	12
				0	0.5	1		2 1	0.5	0.25	0.125	100	100	100	100	100	100
								1	1	0.5	0.25	100	100	100	100	100	100
								1	1	1	1	100	100	100	100	100	100
								1	1	2	4	49	43	38	34	49	44
								1	2	4	8	23	21	16	7	19	17
13	13	13	13		0 0	0		0 1	0.5	0.25	0.125	14	10	8	9	12	11
								1	1	0.5	0.25	13	12	9	11	8	8
								1	1	1	1	11	7	7	5	5	4
								1	1	2	4	13	8	7	7	8	9
								1	2	4	8	12	8	7	7	7	7
				0	0.25	0.5		1	0.5	0.25	0.125	100	100	100	100	95	94
								1	1	0.5	0.25	100	100	100	100	79	77
								1	1	1	1	66	56	51	57	56	54
								1	1	2	4	16	13	11	7	9	8
								1	2	4	8	11	8	5	4	9	8
					0.5	1		2 1	0.5	0.25	0.125	100	100	100	100	100	100
								1	1	0.5	0.25	100	100	100	100	100	100
								1	1	1	1	100	100	100	100	100	100
								1	1	2	4	52	44	39	32	34	30
								1	2	4	8	27	23	18	7	13	10

References

Asiribo, O. and Gurland, J. (1989). Some simple approximate solutions to the Behrens-Fisher problem. *Communications in Statistics - Theory and Methods* **18**, 1201-1216.

Barnard, G. A. (1984). Comparing the means of two independent samples. *Applied Statistics* **33**, 266-271.

Behrens, W. V. (1929). Ein betrag zur fehlerberechnung bei weningen beobachtungen. *Landwirtschaft Jahrbücher* **68**, 807-837.

Beran, R. (1988). Prepivoting test statistics : a bootstrap view of asymptotic refinements. *J. Amer. Statist. Assoc.* **83**, 687-697.

Boik, R. J. (1987). The Fisher-Pitman permutation test: a non-robust alternative to the normal theory F test when variances are heterogeneous. *British Journal of Mathematical and Statistical Psychology* **40**, 26-42.

Brownie, C., Boos, D. D. and Hughes-Oliver, J. (1990). Modifying the t and ANOVA F tests when treatment is expected to increase variability relative to controls. *Biometrics* **46**, 259-266.

Cressie, N. A. C. and Whitford, H. J. (1986). How to use the two sample t-test. *Biometrical Journal* **28**, 131-148.

Dwass, M. (1957). Modified randomization tests for non-parametric hypotheses. *Ann. Math. Statist.* **28**, 181-187.

Edgington, E. S. (1987). *Randomization Tests, 2nd Ed.* Marcel Dekker, New York.

Fisher, N. I. and Hall, P. (1990). On bootstrap hypothesis testing. *Austral. J. Statist.* **32**, 177-190.

Fisher, R. A. (1936). The coefficient of racial likeness and the future of craniometry. *Journal of the Royal Anthropological Institute* **66**, 57-63.

– – – – (1939). The comparison of samples with possibly unequal variances. *Annals of Eugenics* **9**, 174-180.

James, G. S. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika* **38**, 324-329.

– – – – (1954). Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown. *Biometrika* **41**, 19-43.

Manly, B. F. J. (1991). *Randomization and Monte Carlo Methods in Biology.* Chapman and Hall, London.

Schemper, M. (1989). A closed-form jackknife solution for the Behrens-Fisher problem. *Biometrical Journal* **8**, 931-939.

Shiraishi, T. (1993). Statistical procedures based on signed ranks in k samples with unequal variances. *Ann. Inst. Statist. Math.* **45**, 265-278.

Welch, B. L. (1937). The significance of the difference between two means when the population variances are unequal. *Biometrika* **29**, 350-362.

– – – – (1951). On the comparison of several mean values: an alternative approach. *Biometrika* **38**, 330-336.

Wichmann, B. A. and Hill, I. D. (1982). Algorothm AS183: an efficient and portable pseudorandom number generator. *Applied Statistics* **31**, 188-190. (Correction in *Applied Statistics* **33**, 123, 1984).

University of Otago
P.O. Box 56
Dunedin
New Zealand