Critical Reviews in Biochemistry and Molecular Biology, 35(5):339–358 (2000)

Computer-Intensive Statistical Procedures

John Ludbrook

Table of Contents

I. INTRODUCTION

It might be thought that computer-intensive procedures were invented after the introduction of mainframe computers in the early 1950s or desktop computers in the early 1980s. For the most part, this is not so. Most of the statistical procedures to be discussed were invented in the 1930s and only one — bootstrapping — in the last 50 years. For instance, R.A. Fisher described his exact test for 2 × 2 tables in the first edition of his *The Design of Experiments*, published in 1935, in connection with the now-famous "thought" experiment about the lady who claimed she could tell whether milk or tea had been added first to a cup.[1] This test requires the tedious calculation of factorials, but tables of these for numbers up to

1040-9238/00/$.50

300 were provided in 1938 by Fisher and Yates.[2] Fisher also described in the first edition of *The Design of Experiments* an exact permutation test for the difference between paired means. This involved calculating no fewer than 32,768 possible permutations of the difference. Fisher did not get the outcome exactly right in the first edition, but in the second edition, published two years later, he did. An achievement at least as remarkable was made by Eden and Yates in 1933, when they analyzed a randomized block experiment by permutation.[3] The number of possible permutations was 24,[8] but Eden and Yates overcame the computational challenge by an ingenious Monte Carlo pseudo-random sampling technique in which they used cards. Fisher also introduced the notion of maximum-likelihood estimation in the 1930s. Even bootstrapping, a concept described formally by Efron in 1979 that is perhaps the most computer-intensive procedure imaginable,[4] could be executed by hand, though less than optimally.

The foregoing supports the point that the near-universal availability of fast computers on desktops and laps is a convenience rather than a necessity. However, it must be admitted that they are very, very convenient. For instance, it took Fisher two years to correctly execute his permutation test on the same set of data it took me 0.33 seconds to execute on my desktop!

This leads to my issuing a word of warning. Almost every biomedical investigator has access to very powerful computers and can afford sophisticated statistical software. However, this has led to a reliance on the software manuals and "Help" files for selecting the appropriate statistical procedures to analyze one's data, and to faith that the software will execute these procedures correctly. Neither of these assumptions is justified. Biomedical investigators continue to choose inappropriate statistical procedures to analyze their data. There is ample evidence that software manuals (and even some well-known statistical texts) can be an unreliable source of advice when choosing the appropriate test, and the statistical software may sometimes execute tests incorrectly.[5]

The computer-intensive tests to be discussed reflect to some extent the preferences and prejudices of the author. The list of tests is also restricted to those that might be of interest to pharmacologists, whether those whose research is conducted in the laboratory or those whose field is clinical pharmacology. A final point is that biomedical investigators in general, and pharmacologists in particular, employ very small groups in their experiments.[6] This fact, plus the rarity with which random sampling is used compared with randomized designs in biomedical research, provide strong arguments in favour of using exact tests of significance rather than those based on assumptions about population distributions. Computer-intensive procedures, including those described as exact, compile an empirical sampling distribution that is specific to the particular dataset and against which the observed outcome of the experiment, or the observed test statistic, is matched.

II. A NOTE ON INFERENCE

A. The Classical Population Model

This has been the main basis for statistical theory over the past 70 years.[7,8] It depends on the assumption that the experimental groups have been acquired by random sampling of defined (though not necessarily accessible) populations. The sample sizes

are very small compared with the size of the populations. Significance testing proceeds as follows. A population parameter of interest is selected, such as the mean, or the difference between means; the variance, or the variance ratio; or the difference between, or ratio of, proportions. The corresponding sample statistic is then calculated. A further assumption is then made about the distribution of the sample statistic in the population (for instance, according to the normal, *t*, *F*, or chi-squared distributions). A null hypothesis is then formulated. In its general form, this hypothesis states that the two or more random samples have been drawn from the same population. Somewhat more specifically, it may propose that the samples have been taken from populations with equal means, or equal variances, or equal proportions. The next step is to conduct an appropriate test of significance for the null hypothesis, such as Student's *t* test, Fisher's *F* test (analysis of variance), or one or another formulation of Pearson's χ^2 test. It is also necessary to set the desired level of significance. Traditionally, this is the 5% level ($P \leq 0.05$). What the 5% level of significance is intended to mean is that if a large (infinite) number of random samples were to be taken from the population(s) actually sampled in the experiment, with replacement, then in not more than one sample in 20 would the statistic of interest be equaled or exceeded. Hence the convention of expressing the level of significance as $P \leq 0.05$.

This lengthy and complex description of the classical, population, model of significance is given so as to be able to point out the difficulties with it in the setting of biomedical and pharmacological research. The fact is that investigators in these fields never, or almost never, fulfill even the first assumption. That is, they do not take random samples from defined populations.[6]

B. The Randomization Model of Inference

Fisher first introduced the notion of randomization in experimental design in 1936.[9] However, the randomization model of inference has not been nearly as well defined and debated by statisticians as the classical, population, model. It is hard to avoid concluding that this is because mathematical statisticians find that there is little opportunity to exercise their skills on this model. Nevertheless, there have been statisticians who have spoken out in favor of it, especially Otto Kempthorne and George Box.[10-13] There have also been biomedical experimenters who have written in favor of it.[6,14,15] Under this model, it is supposed that a nonrandom sample of experimental units (humans, animals, tissues, or cells) has been acquired. The members of this sample are then allocated at random to two or more conditions or treatment regimens. "Condition" could mean, for instance, rest or exercise, normovolaemia or hypovolaemia; and "treatment" could mean active treatment or placebo, a new treatment or an established treatment. These options are familiar to biomedical and pharmacological researchers. The experiment is then performed, and the results analyzed. The null hypothesis is no more than that the conditions or treatments have no differential effects on the groups with respect to a statistic such as the mean, proportion, or odds. There is no reference to a population and therefore no assumption is, or can be, made about the frequency-distribution of the population parameter corresponding to the statistic of interest. Instead, an empirical and unique sampling distribution of the statistic of interest is compiled by techniques such as permutation (see later). The *P* value is the frequency with which values of the statistic are equal to, or more extreme than, the experimentally de-

termined value occurring in the empirical distribution. The statistical inference refers only to the nonrandom sample that was randomized, and not to any theoretical more general population. Biological inferences beyond those made from the experiment can be made only by nonstatistical arguments about how representative the original nonrandom sample was.

III. ANALYSIS OF TABLES OF FREQUENCY BY EXACT METHODS

This section is concerned with computer-intensive or "exact" techniques for analyzing tables of frequencies. An excellent elementary account of some of these is given by Siegel and Castellan,[16] and a more advanced account by Agresti.[17] This seems a good place at which to start any essay on computer-intensive statistical methods because the concept of an exact test is relatively simple. The older and more conventional methods for analyzing tables of frequencies depend on tests based on the chi-squared or normal distributions. These are approximate or asymptotic tests of significance that are reasonably accurate for 2×2 tables in which the cell frequencies are large. They are notably inaccurate when the cell frequencies are small. They have been, or should be, replaced by tests that are based on the exact sampling distributions for specific tables.

A. Analysis of 2 × 2 Tables

Two-by-two tables (that is, 2 columns, 2 rows, sometimes called four-fold tables), are the simplest way of tabulating frequencies. A stereotype 2×2 table is shown in Table 1. The dataset that will be used as an example is in Table 2.[18]

The asymptotic Pearson $\chi 2$ test. This is the oldest method for analyzing tables of frequencies. Karl Pearson described the chi-squared distribution in 1900,[19] though it was not until R.A. Fisher described the correct degrees of freedom in 1922 that the $\chi 2$ statistic could be used as a practical test of significance.[20] It is usually described as testing for goodness of fit: that is, for identity of observed and expected frequencies. However, one should, whether fairly or not, regard Pearson's $\chi 2$ test as of historical interest only. It is sufficient to cite Yates' recognition of its inaccuracy in his paper of 1934, when he prescribed his correction for continuity and controlled its efficacy by comparing outcomes with those of the Fisher exact test,[21] or to cite Cochran's paper of 1954 in which he suggested complex rules based on expected cell frequencies for defining the limits of cell frequencies for 2×2 and $r \times c$ tables below which the Pearson $\chi 2$ test is unsafe.[22] Yet some biomedical investigators still analyze their results with this test. However, the Pearson $\chi 2$ statistic can also be used in an exact test (see later).

The Fisher exact test. As indicated earlier, Fisher described this technique for analyzing 2×2 tables in 1935 in his *The Design of Experiments*.[1] It depends on constructing the hypergeometric distribution for all possible tables having the same marginal totals as the observed table. For any given table, and using the notation of Table 1, the probability of occurrence of the entries in the table is given by the formula:

$$\Pr = \frac{(a+b)(c+d)(a+c)(b+d)}{N!\,a!\,b!\,c!\,d!}$$

where ! indicates factorial.

All possible probabilities of occurrence for the example in Table 2 are given in Table 3.

Table 1
Stereotype 2 × 2 Table of Frequencies

	Column 1	Column 2	Total
Row 1	a	b	(a + b)
Row 2	c	d	(c + d)
Total	(a + c)	(b + d)	N

Note: The convention is to describe Rows as r_1, r_2, and so forth, and Columns as c_1, c_2, and so forth. It is usual to regard Columns as treatment groups, Rows as the outcomes of treatment.

Table 2
Example of a 2 × 2 Table after Davis[18]

	Group A	Group B	Total
Dead	3	2	5
Alive	6	19	25
Total	9	21	30

Note: For the purpose of discussion, assume that the investigators have acquired a nonrandom sample of 30 experimental units (humans, animals, tissues, cells). By a process of restricted randomization, they have assigned 9 to treatment Group A, 21 to treatment Group B. The outcomes of treatment are given by the Rows.

Table 3
List of All Possible 2 × 2 Tables Resulting from Example in Table 6.2, Given Fixed Marginal Totals

Tables					Test Statistics		
r_1c_1	r_2c_1	r_1c_2	r_2c_2	Hypergeometric Pr	Exact χ^2	Exact G^2	Exact lnOR
5	4	0	21	0.00008842*	14.00*	14.67*	∝*
4	5	1	20	0.018568*	7.14*	6.63*	2.77*
3	**6**	**2**	**19**	**0.12378***	**2.57***	**2.37***	**1.56***
2	7	3	18	0.33599	0.29	0.27	0.54
1	8	4	17	0.37798	0.29	0.30	0.63
0	9	5	16	0.14279	2.57*	3.98*	∝*
		Total Pr for table:		**1.0000**			

Note: The example is from Davis18. The observed table is in bold. * indicates a value that is ≥ that observed in either direction. For the Fisher test, two-sided *P* is the sum of hypergeometric probabilities (Pr) that are equal to or less than that observed in both directions. For the exact χ^2, G^2, and lnOR tests, two-sided *P* is the sum of the hypergeometric probabilities corresponding to tables in which the test statistic is equal to or greater than that for the observed table, in both directions.

Table 4
Outcomes of Significance (Hypothesis) Testing on the Example of Table 2

Test	Exact *P* Values		Asymptotic *P* Values	
	1-Sided	2-Sided	1-Sided	2-Sided
Pearson χ^2	0.143	0.286	0.0544	0.109
Yates χ^2	NA	NA	0.143	0.285
Fisher exact	0.143	0.143	NA	NA
Odds ratio	0.143	0.286	0.0544	0.109
Likelihood ratio	0.143	0.286	0.0619	0.124
Relative risk (ratio of proportions)	NA	1.000	NA 0.109*	0.115*
Difference of proportions	NA	0.260	NA	0.161
Permutation test on proportions (difference between proportions)	0.143	0.286	NA	NA

Note: *P* values from StatXact 3.1 (Cytel Software Corporation, Cambridge MA). * using two different formulae. NA, not applicable.

The null hypothesis for the Fisher test is rather vague. It is effectively that the cell frequencies in the observed table are disposed at random. The hypothesis is tested by summing the observed hypergeometric probabilities (Pr) with all more extreme (smaller) values of Pr (Table 3). If the summation is done unidirectionally, the test is one-sided. If a two-sided test is preferred (as it almost always should be), both tails of the distribution of Pr are summed to arrive at *P* values (Tables 3 and 4).

The Fisher test can be executed by means of a handheld calculator, but even for the example of Table 2 it is a tedious process. By means of a computer, the test can be performed on even the largest 2×2 table in a matter of milliseconds.

The odds ratio (OR). This is defined as:

$$\mathrm{OR} = \frac{a / c}{b / d}$$

The null hypothesis is that OR = 1, a relatively specific one.

Its asymptotic form makes use of the fact that lnOR (the natural logarithm of OR) is distributed approximately as the normal distribution. However, as with the Pearson test, this approximation is poor when cell sizes are small. There is an exact form of the test that depends on calculating lnOR for all 2×2 tables with the same marginal totals as those observed. *P* for the null hypothesis is obtained by summing the probabilities from the hypergeometric distribution that correspond to values for lnOR equal to or greater than that observed (Tables 3, 4). One-sided values for *P* are obtained by summing in one direction, the preferred two-sided P by summing both tails.

The exact χ^2 test. The logical basis for this is similar to that of the odds ratio test. The χ^2 statistic is calculated for all 2×2 tables with the same marginal totals as those observed. *P* for the null hypothesis is obtained by summing the probabilities from the hypergeometric distribution that correspond to values for χ^2 equal to or greater than that observed (Table 3). One-sided values for *P* are obtained by summing in one direction, the preferred two-sided *P* by summing both tails.

Likelihood ratio. The log-likelihood statistic, sometimes called G or G^2, is described well by Sokal and Rohlf.[23] The G^2 statistic is distributed asymptotically and approximately according to the chi-squared distribution. It is usually used for log-linear mod-

eling, and for testing the independence of three-way or multiway tables of frequency.[23] However, it can be used to evaluate single 2×2 tables under a null hypothesis involving goodness of fit that is as vague as that for the Pearson test. There is an exact form of the log-likelihood ratio test. The G^2 statistic is calculated for all 2×2 tables with the same marginal totals as those observed. *P* for the null hypothesis is obtained by summing the probabilities from the hypergeometric distribution that correspond to values for G^2 equal to or greater than that observed (Tables 3, 4). One-sided values for *P* are obtained by summing in one direction, the preferred two-sided *P* by summing both tails.

Ratio of, or differences between, proportions. At first sight it is appealing to use the ratio of proportions (often called relative risk or RR), or the arithmetic difference between proportions ($\rho_1 - \rho_2$), to test null hypotheses such as RR = 1, or $\rho_1 = \rho_2$. These are concepts that biomedical investigators feel comfortable with. There are, however, considerable difficulties in doing this, whether by asymptotic or exact techniques. There are algorithms for doing this asymptotically,[17] but they are not regarded as altogether satisfactory. There are also algorithms for conducting these tests exactly,[17] but these are not merely computer-intensive, but also difficult to interpret. In short, there is not much enthusiasm for using proportions.

There is, however, another way of looking at the difference between proportions that seems to get round the difficulties referred to above. First, set out the experimental data as a rectangular file, identifying groups (1, 2) and the outcomes as, for instance, 1 = dead, 2 = alive. Then the proportion for each group coincides with the mean value. Then a permutation test for equality of group means (see below) is identical to a test for equality of proportions. The outcome of such a test is given in Table 4.

Summary. The following points can be made about the analysis of 2×2 tables:

1. Nowadays, it is almost impossible to make a case for using anything other than exact, computer-intensive techniques for analysing 2×2 tables. Asymptotic methods that involve approximations to the chi-squared or normal distributions are almost invariably too liberal: that is, the resultant *P* value is too small and the Type 1 error rate inflated beyond that specified. The only possible exception is to use Yates' continuity correction to the Pearson asymptotic χ^2 test. This usually gives a somewhat more conservative outcome than the Fisher exact test, which it was designed to simulate.[21]
2. Which of the exact techniques is to be preferred? The exact form of Pearson's χ^2 test, the Fisher exact test, and the exact log-likelihood ratio have the weakness that the null hypothesis is rather vague. The tests have been described in terms of "goodness of fit" or "independence of rows and columns." They do not test for equality of proportions, though this is a not uncommon misapprehension. There are exact tests for equality of proportions: the relative risk (RR, or ratio of proportions) and the difference between proportions, the null hypotheses being quite specific. That is, they are that RR = 1 and $\rho_1 - \rho_2 = 0$. However, the algorithms for executing these tests exactly are unconditional: that is, it is assumed that neither the column nor the row marginal totals are fixed in advance. This can lead to some rather curious results (Table 4). The author believes that the best option is the test on odds ratios where the null hypothesis is OR = 1. The algorithm is a conditional

one, the specific condition being that the column marginal totals are fixed in advance by the design of the experiment. When columns indicate groups and rows indicate outcomes, this conditioning is in accord with the usual experimental design, when group sizes (column totals) are determined in advance by the investigators. A neglected, but apparently valid way of testing for equality of proportions is to use a permutation test.

3. In the author's view, two-sided tests should be the norm in biomedical research in general, and pharmacological research in particular. However, it is worth noting that whereas the Fisher exact test, the Pearson exact test, and the exact tests on the odds ratio and log-likelihood ratio always result in identical one-sided values of *P*, the two-sided P value from the Fisher test is often less than that for the others (Tables 3, 4).[18]

B. 2 × c Tables of Frequencies

The advantages of exact, computer-intensive tests over asymptotic tests are rather less as the number of rows and columns increases. Nevertheless, exact tests are always more accurate, especially if the expected frequencies in some cells are small.[17] The example used in this section was described by Cochran.[22]

Pearson's χ^2 and log-likelihood tests for independence of rows and columns. These can be executed in the familiar, asymptotic form or in an exact form. Outcomes from the example are given in Table 5, and are very similar.

Fisher-Freeman-Halton exact test. This extends the Fisher exact test to 2 × c tables.[24] The outcome for the example is given in Table 5.

Cochran-Armitage test for ordered 2 ¥ c tables. Almost simultaneously, Cochran and Armitage described an asymptotic method of partitioning the χ^2 statistic in the case of 2 × c tables when the columns represent a variable that increases (or reduces) in a linear fashion.[22,25] Genuine linearity is not necessary, and rarely occurs with categorical variables, but monotonicity is essential. The test can be executed in either an asymptotic or exact form. Its value is illustrated by the example (Table 5). If the columns are regarded as in no particular order so that the null hypothesis is merely one of independence of rows and columns, *P* from the above tests ranges between 0.139 and 0.144. However, if the columns represent a monotonically ascending order (as was the case in Cochran's example), *P* from the asymptotic or exact forms of the Cochran-Armitage test are 0.010 and 0.011, respectively. The null hypothesis is, of course, different: expressed in a two-sided fashion it is that there is no monotonic change in the proportions according to columns.

Summary. The following points can be made about the analysis of 2 × c tables:

1. For unordered 2 × c tables, the Fisher-Freeman-Halton exact test is to be preferred.
2. For ordered 2 × c tables, the exact form of the Cochran-Armitage is the procedure of choice.

C. Unordered r × c Tables of Frequencies

When there are more than two rows in an r × c table, the null hypothesis is of necessity quite nonspecific. It is merely that there is no association (interaction) between rows and columns. In tables such as these, the advantage of exact over asymptotic tests

Table 5
Example of a 2 × c Table of Frequencies from Cochran[22]

Skin Infiltration	Col 1	Col 2	Col 3	Col 4	Col 5	Total
Extensive	1	13	16	15	7	52
Slight	11	53	42	27	11	144
Total	12	66	58	42	18	196

Note: The table summarizes observations on lepers undergoing treatment with sulphones and streptomycin. The rows indicate the severity of skin infiltration before treatment. The columns indicate the clinical responses to treatment.

If the clinical responses are in no particular order, the outcomes of appropriate tests in terms of *P* values are:

Asymptotic Pearson χ^2 test	0.142
Exact Pearson χ^2 test	0.142
Asymptotic log-likelihood ratio test	0.123
Exact log-likelihood ratio test	0.139
Fisher-Freeman-Halton exact test	0.144

However, in reality the columns are in monotonic ascending order in terms of clinical response: 1 = Worse. 2 = Unchanged. 3= Slight improvement. 4 = Moderate improvement. 5 = Marked improvement. If the goal is to test for an association between the initial skin condition and the effect of treatment, then the Cochran-Armitage test is the appropriate one. The outcome is:

Asymptotic Cochran-Armitage test	0.010
Exact Cochran-Armitage test	0.011

is less obvious, unless there are several cells in the table with low expected frequencies. The tests available are as follows.

Pearson χ^2 test, asymptotic or exact versions. Log-likelihood ratio test, asymptotic or exact. The properties of these have already been described under 2 × c tables.

Fisher-Freeman-Halton exact test. This has the same properties as for 2 × c tables.

Summary. For unordered r × c tables, there is really little choice but to use the exact Fisher-Freeman-Halton test.

IV. PERMUTATION (RANDOMIZATION) TESTS

As indicated earlier, the idea of permutation tests goes back to Fisher, and Eden and Yates in the 1930s. There are now excellent books on this topic, in particular Edgington[15] and Manly.[26] These two books deal with permutation tests in the context of the randomization model of inference, and are directed towards investigators in the biological and biomedical sciences. Other books deal with the topic of computer-intensive tests in a more pragmatic fashion, without being based on explicit models of inference.[27,28] There is also an article that deals with these tests in the setting of biomedical research.[6]

The logic that underlies these tests is remarkably simple, and falls within the framework of the randomization model of inference. That is, the null hypothesis is that there is no differential effect of treatments on the randomized groups. The goal is to compile the permutation (randomization)

Table 6
The Process of Exact Permutation for Two Independent or Two Related Groups

Two Independent Groups

Formula for calculating all possible permutations for two independent groups of size n_1 and n_2:

$$\frac{(n_1 + n_2)!}{(n_1)!(n_2)!}$$

The 10 possible permutations for $n_1 = 2$, $n_2 = 3$. Observed values for n1 were A, B, and for n_2 were C, D, E.

A	C	A	B	A	B	A	B	B	A
B	D	C	D	D	C	E	C	C	D
E		E		E		D		E	
B	A	B	A	C	A	C	A	D	A
D	C	E	C	D	B	E	B	E	B
E		D		E		D		C	

Two Related Groups (Matched Pairs)

Formula for calculating all possible permutations of two related groups of size n is 2^n.

The 8 possible permutations for group size n. Observed values for the pairs were A1 & A2, B1 & B2, C1 & C2.

A1	A2	A2	A1	A1	A2	A1	A2
B1	B2	B1	B2	B2	B1	B1	B2
C1	C2	C1	C2	C1	C2	C2	C1
A2	A1	A2	A1	A1	A2	A2	A1
B2	B1	B1	B2	B2	B1	B2	B1
C1	C2	C2	C1	C2	C1	C2	C1

distribution for each set of experimental data. The only assumption is that there has been a randomized experimental design. All possible permutations of the experimental data are then compiled, preserving the actual group numbers and sizes. Then for each permutation the statistic of interest is calculated. The most commonly used statistic, and the easiest to understand, is the difference between two means. However, when there are more than two groups, the *F* statistic (or a simplified version of it) can be calculated as for a one-way or two-way analysis of variance. The values of the statistic associated with the permutations are then placed in rank order. The *P* value is calculated as all values of the selected statistic that are equal to or greater than that observed. This is done for both tails of the permutation distribution for a two-sided *P*, and for one tail to obtain a one-sided *P*.

As will become clear, the number of possible permutations increases steeply with the number of groups and the group sizes. Even with a very fast computer it may be quite impractical to list all possible permutations. In this case, a Monte Carlo pseudorandom sample of all the possible permutations is taken, usually 10,000. The calculation of *P* is then done on this Monte Carlo sample (and confidence limits for *P* can be provided).

There are several computer programs for executing permutation tests, listed in Section VI.

Two independent groups. How all possible permutations are compiled is illustrated in Table 6. The formula for calculating the

number of possible permutations is $(n_1 + n_2)!/(n_1)!(n_2)!$ where n_1 and n_2 are the group sizes, and ! indicates factorial. The rate of increase of permutations increases steeply with group size (Table 7), so that with groups larger than about 12, the Monte Carlo approach described above has to be used. The exact permutation distribution for the difference between two independent means is displayed in Figure 1. It should be noted that its bimodal appearance is very different from that of the t distribution.

Two related groups (matched pairs). How all possible permutations are compiled is shown in Table 6. The formula for the number of possible permutations is 2^n. The rate of increase with group size is much less than that for two independent groups (Table 7).

Rank permutation tests. Wilcoxon described these in 1945,[29] and Mann and Whitney described exactly equivalent tests in 1947[30] producing the Wilcoxon signed rank-sum test for matched pairs and the Wilcoxon-Mann-Whitney test for two independent groups. Wilcoxon made it clear that he strongly supported Fisher's idea of permutation tests,[29] but in 1945 there were no computers with which to execute them, so they were altogether impractical. He recognized that if ranks were substituted for continuous variables, the computational problem was much simplified. The number of possible permutations remained the same, but the number of values assumed by the test statistic was very much smaller. However, it is the author's view that the rank permutation tests should no longer be used. The arguments in support of this view are given in Ludbrook and Dudley[6] and Bergmann et al.[5] Perhaps the strongest argument is that these are tests for differences in mean ranks, and under the ranking system used, this does not coincide with differences between medians (though this is a common misapprehension).

Permutation tests for multiple randomized groups. These correspond to one-way or two-way analysis of variance (ANOVA). The principle of the test is very similar to that for two groups, except that the statistic of interest is the F statistic (or, rather, a simplified version of it).

Eden and Yates were the first to describe a statistic equivalent to *F*,[3] and Edgington gives a simple explanation of this.[15] The equivalent test statistic is:

$$\sum(T_i^2 / n_i)$$

where *T* is the sum of all values in a particular group, *n* the group size. T^2/n is summed across all groups under consideration, for each permutation of group values. If all groups are of the same size, the expression can be reduced to:

$$\sum(T_i)^2$$

These statistics are equivalent to *F* in the sense that across all the permutations,

Table 7
Number of Possible Permutations for Two Independent or Two Related Groups

Two Independent Groups		Two Related Groups (Matched Pairs)	
Size: $n_1 = n_2 =$	Permutations	Size: $n =$	Permutations
5	252	5	32
10	184,756	10	1,024
15	155,117,520	15	32,768

Note: Number of permutations calculated from the following formulae. Two independent groups: $(n_1 + n_2)!/(n_1)!(n_2)!$, where ! indicates factorial. Two related groups (matched pairs): 2^n.

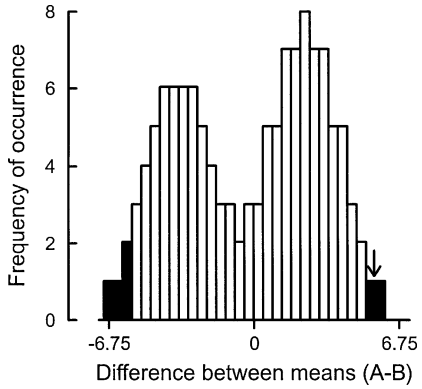


FIGURE 1. Frequency-histogram of the exact permutation distribution for the difference between two independent means. Values for group A: 20, 10, 9, 8, 6. Values for group B: 7, 5, 4, 3. There are 126 possible permutations. ■ the 6 permutations for which the difference between means equals or exceeds, in either direction, the observed value of 5.85, so that two-sided $P = 6/126 = 0.048$. ▾ observed value. From Ludbrook[41] with kind permission of Blackwell Science Asia.

the rank order of $\sum(T_i^2/n_i)$ is identical to that of F. Obviously, the equivalent test statistic requires much less computation than F.

There are no problems with executing an exact or Monte Carlo version of one-way ANOVA. However, two-way ANOVA and factorial analysis by permutation present some problems. The main effects can be extracted without difficulty. But more often than not, investigators use interactions within two-way ANOVA to test their main hypothesis. Interactions do present a problem of execution by permutation. There is no consensus about the best way to go about extracting these,[15,26-28] and the outcomes using different techniques can be very different. Repeated measures ANOVA, commonly used in biomedical laboratory research, seem to have no corresponding permutation version, especially for within-group interactions.[31] This is scarcely surprising, since measurements made serially cannot be regarded as randomized. However, there may be no cause for despair. I know of no study that proves the case, but it is my strong impression from analyzing experimental data that the greater the number of randomized groups, regardless of their size, the less the advantage of permutation tests compared with classical F tests.

The use of permutation is not confined to comparing two or more group means. For instance, it can be used to test for equality of group variances. Its use in many other experimental designs has been described, such as for regression and correlation, time series, spatial analysis, repeated measures analysis, multivariate analyses, and clinical experiments of the $n = 1$ type.[15,26] However, the experiments used as examples for the above applications do not always fulfill the requirement of a randomized design. In the author's opinion, if this should be the case, then permutation techniques should not be used.

Summary. The following points can be made about the use of permutation (randomization) tests:

1. These should be used only when the experiment is a prospective one in which there has been genuine randomization of the experimental units (humans, animals, tissues, cells) to "treatment" groups.
2. If the circumstances are as above, and randomization has been to only two "treatments," a very strong and almost irrefutable case can be made for *never* using Student's t test, but only using a permutation test for equality

of means (independent groups or matched pairs).

3. If the experimental design is one of randomization, then the best test for equality of variances is a permutation test.
4. The case for a permutation test is weaker if there are more than two randomized experimental groups, and the design corresponds to two-way or multiway analysis of variance or factorial analysis. There are two reasons for this view: (a) More often than not investigators are interested chiefly in interactions, such as a dose × treatment interaction when doses are given in random order and treatments are, for instance, exposure to one or more antagonists and a control. It is not at all clear what the best approach is for testing interactions by permutation. (b) There is little or no empirical evidence that, in complex designs such as these, permutation analysis provides greater accuracy in hypothesis-testing than conventional forms of analysis of variance. In other words, the greater the number of experimental groups and the greater the degrees of freedom attached to the Residual Mean Square in ANOVA, the closer the outcomes. This is especially so if the investigators (as they always should) incorporate an analysis of residuals into their analyses of variance and, if indicated, undertake data transformation or weighting.[32]
5. There seems to be no case at all for using permutation techniques to analyze data that have been acquired neither by randomization nor by random sampling. A very obvious example is when serial observations have been made in time, dose, or space when conventional repeated-measures analysis of variance would be used with appropriate correction for serial autocorrelation.[31]

V. MONTE CARLO METHODS

The term "Monte Carlo" originates from the gaming tables of Monaco. It refers to repeated random sampling of mathematically defined populations, with replacement. Thus, repeated tosses of a coin, repeated throws of a die, or repeated spins of a roulette wheel are forms of repeated random sampling with replacement, in which the sample size is 1. If one wishes to make inferences from Monte Carlo random sampling, they would clearly be under the population model of inference (see Section II).

If Monte Carlo random sampling is applied to the results of biomedical experiments, it is likely that the sample size will be a good deal greater than 1. However, in other respects, the situation is not dissimilar to that of the games of chance referred to above:

1. The mathematical form of the parent population is known or postulated by the investigators.
2. The biological population is large compared with the sample size(s).
3. The experimental design involves genuine random sampling of the biological population.

Usually, hypothesis-testing or the construction of confidence intervals would be undertaken by parametric tests, for instance, those in which the t or F statistics are employed to analyze continuous data sampled from a population that is regarded as conforming to the normal distribution. So how can Monte Carlo methods be applied usefully in biomedical work?

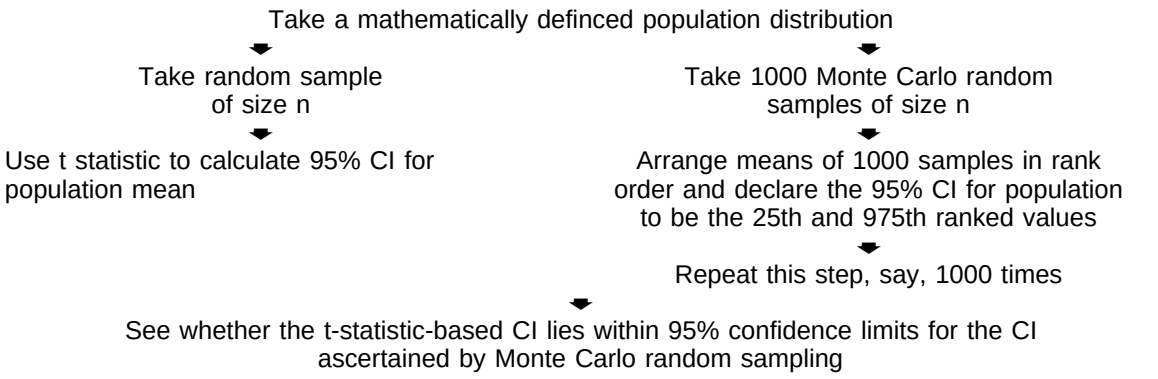
A better question is how Monte Carlo methods have been applied. A search of Medline for the keyword "Monte Carlo" over the two-year period 1998–99 resulted in nearly 1000 references. Most of these referred to organ imaging via nuclear medicine techniques, computerized tomography (CT), magnetic resonance (MR) imaging, and positron emission tomography (PET). Many were concerned with radiotherapy. A smaller number were in the field of genetics. There were some concerned with clinical trials, but only a handful dealt with pharmacology, mostly pharmacokinetics. An exception was the paper by Christopoulos.33

A. Parametric Applications of Monte Carlo Methods

Monte Carlo simulation studies. One of the most common applications of Monte Carlo methods is in simulation studies.[34] These have been used extensively in the computer era to decide how accurate conventional tests such as t and F tests are when the randomly sampled population is not normal, or when the populations sampled have different variances, or both.[6] This process can be most easily explained if, for instance, the accuracy of 95% confidence intervals (95% CI) for a population mean are estimated by the conventional method of using the t statistic. It is illustrated in Table 8.

There is nothing magical about the figure 1000 for the number of Monte Carlo random samples (Table 8), though this is commonly used. Nor is there anything magical in repeating the process 1000 times (Table 8), though this, too, is a common choice. The accuracy of using the t statistic to estimate the 95% CI from a single sample can then be compared with the mean value from Monte Carlo sampling (and its 95% range). It should be apparent that the Monte Carlo method is precisely the theoretical process that underlies inferences made under the population model (see Section II). The process illustrated in Table 8 can be repeated with different sample sizes (for instance, $n = 10$, 20, or 30). In the context of testing the "robustness" of 95% CIs for the population mean estimated by using the t statistic, populations with different mathematically defined frequency distributions can be sampled. These might be, for in-

Table 8
Flow Chart to Demonstrate the Process of Monte Carlo Simulation, to Test the Accuracy of Parametric *t*-Statistic-Based 95% Confidence Intervals (CI)



Note: The process can be repeated with a variety of mathematically defined frequency distributions and with a range of sample sizes.

stance, distributions such as the normal, log-normal, Cauchy, beta, triangular, and so forth (see Reference 35 for definitions and illustrations). At the end of the simulation study, it should be possible to reach conclusions about the accuracy of the *t* statistic technique under different forms of nonnormality and different size samples. If the population distribution is genuinely normal, the t statistic approach is not likely to be subject to systematic error. Conversely, other population distributions and other patterns of sample size or variance may be associated with systematic errors in the estimation of the 95% CI.

The example given above is a rather trivial one, but Monte Carlo simulation studies with which the author is familiar have been used to test the accuracy and power of hypothesis-testing by classical *t* tests and analysis of variance (ANOVA),[6] the relative accuracy of different corrections for autocorrelation in repeated-measures ANOVA,[31] and the accuracy of control of the familywise Type I error rate by a wide variety of multiple comparison procedures.[36] From these simulations, it has been possible to make certain generalizations. For instance, the independent-sample *t* test is relatively robust against nonnormality of the population, but is very sensitive to inequality of population variance, in the direction that if the smaller sample comes from the population with the greater variance, the Type I error rate is inflated; whereas if the converse is the case, the statistical inferences are too conservative and the Type II error rate is inflated. Monte Carlo simulations have also been used to test the accuracy of bootstrapping (see below).

I have glossed over just how Monte Carlo random sampling is done, but it is described in relatively simple terms by Morgan.[34] For those in the know, it is a relatively simple matter to write computer programs (or to use programs written by others) to undertake the process described above.

Hypothesis-testing by Monte Carlo methods. It should be clear from the flowchart given in Table 8, that all parametric tests of significance could be replaced by Monte Carlo tests. That is, if investigators must postulate that their biological populations conform to the normal distribution and would usually use the *t* or *F* statistic to test hypotheses, then multiple (Monte Carlo) random sampling of the theoretical normal distribution could be used to test the hypotheses, and this approach would be entirely in line with the theory of the population model of inference. However, this would be like using a piledriver to crack a nut. Not only would the process be highly computer-intensive, but ultimately its validity depends on knowledge of the frequency distribution of the biological population. This is impossible to discover, except by studying multiple random samples from that population. Furthermore, it requires that the biological population has been randomly sampled, a condition that is excessively rare in practice.[6] Nevertheless, the Monte Carlo approach has been supported both by biostatisticians and by pharmacologists. For instance, Manly gives examples of applying the method to the spatial distribution of plants and to mandible length in jackals.[26] Concato and Feinstein indicate applications in clinical research.[37] More immediately relevant, Christopoulos has applied the method to the analysis of ligand-receptor interactions.[33]

Nonparametric applications of Monte Carlo methods. One of the few revolutionary ideas in statistics in the computer era (post-1950) was the notion of bootstrapping introduced by Efron in 1979.[4] The word "bootstrapping" originates from one of the extravagent tales of Baron Munchausen, in which the hero extricates himself from a perilous situation by pulling himself up by his

bootstraps. Bootstrapping has more or less eclipsed an earlier application of Monte Carlo methods known as "jacknifing," because of the greater generalizability of the former.

A Medline search for 1998–99 turned up only 46 articles in which bootstrapping was used. Only two of these were in the field of pharmacology. An earlier paper by Lew and Angus used bootstrapping in the context of nonlinear regression.[38]

The concept that underlies bootstrapping is relatively simple. It will be recalled that, under the population model of inference (Section II), repeated, relatively small, random samples taken with replacement from a defined population allow the distribution of that population to be accurately constructed. However, biomedical populations are rarely, if ever accessible and so cannot be repeatedly randomly sampled. Efron proposed that if the actual, experimental, random sample were repeatedly randomly sampled with replacement, then an reasonably close approximation of the population distribution could be deduced.[4] Authoritative accounts of bootstrapping are given by Efron,[4] Efron and Tibshirani,[39] Mooney and Duval,[40] Manly,[26] and an elementary account by Ludbrook.[41]

The main thrust of Efron's proposal was to be able to construct bootstrap standard errors and confidence intervals without having to make assumptions about the population distribution. The simplest way of doing this is to compile, say, 1000 resamples of the original sample by random sampling with replacement. These are then placed in rank order in terms of the statistic of interest, for instance the mean, or the difference between two means. This is the so-called bootstrap percentile method. In Table 9 an example is given of the first 10 out of 1000 resamples of a set of data, and Figure 2 shows the bootstrap frequency distribution of the difference between two means for the same set of data as for Figure 1. It should be noted that there is no more than a hint of the bimodality that is evident in the permutation distribution of Figure 1. However, it is now clear on theoretical grounds and as a result of Monte Carlo simulation studies that there can be serious errors in confidence intervals constructed in this way from small samples (see Reference 41). There may be bias, in that the bootstrap statistic does not correspond to the population parameter (for instance, mean or difference between means). More importantly, the con-

Table 9
The Process of Bootstrap Resampling for Two Independent Samples

For the actual experimental samples, the values were A, B, C, D, E and F, G, H, I, respectively. The values in the first 10 bootstrap random resamples were as below.

C	H	B	G	D	H	E	G	E	F
C	G	D	G	B	F	D	G	C	F
B	H	D	G	B	F	B	H	B	H
E	H	E	F	A	F	C	I	D	I
D	A	A	D	E					
E	G	A	G	E	H	A	G	C	I
A	I	B	I	A	G	A	H	C	I
B	I	C	F	E	G	A	F	E	F
C	H	E	F	C	F	D	H	A	I
E	D	C	A	A					

Note: Because bootstrap resampling is done with replacement, the same values can appear more than once in each group. This is in contrast to compiling permutations (Table 6.6), where random resampling is done without replacement.

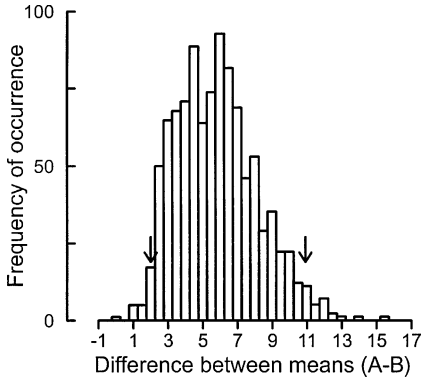


FIGURE 2. Frequency-histogram of the percentile bootstrap distribution for the difference between two independent means, based on 1000 resamples. Observed values as in Figure 1. Observed difference between means 5.85. ◆ indicate 95% bootstrap percentile confidence limits. From Ludbrook[41] with kind permission of Blackwell Science Asia.

fidence interval is likely to be too narrow. Several ways of overcoming these deficiencies have been proposed. These include the percentile-*t* method,[39,42] and the bias-corrected and accelerated (BC_a) method.[39] The BC_a method is more widely applicable, but the percentile-t method is much simpler to execute (see Reference 41). Several other much more complex techniques for improving the accuracy of bootstrap resampling have been proposed. These include balanced random sampling,[39] double-bootstrapping,[40] and applicable only to very small samples, complete enumeration of all possible bootstrap resamples.[44] These proposals have yet to be fully evaluated.

Bootstrapping can also be used for hypothesis testing. In its crudest form, for testing for equality of population means, the bootstrap distribution is used in much the same sort of way as the permutation distribution (see Section IV). That is, the bootstrap differences between means (say 1000) are placed in rank order, and *P* is calculated as:

$$\frac{\text{Number of bootstrap mean differences} \geq \text{that observed}}{\text{Total number of bootstrap differences}}$$

However, this simple approach is susceptible to serious errors, especially when the original samples are small. Several solutions to this have been proposed, including the use of *t*- or *F*-like statistics,[41] but none is entirely convincing. In particular, a simulation study suggests that if samples are small (15 to 25) the percentile-*t* method is associated with an inflated risk of Type 1 error.[43]

The percentile-^{t} bootstrap distribution for the dataset of Figures 1 and 2 is shown in Figure 3. It should be noted that it resembles the permutation distribution of Figure 1 in being bimodal.

Summary. The folllowing points can be made about Monte Carlo methods:

1. Strictly, these are applicable only to experimental situations in which genuine random samples are taken from defined populations. This is uncommon in biomedical research.
2. Monte Carlo simulation studies have played an important part in defining the conditions under which parametric or nonparametric methods for hypothesis testing are accurate (or, more importantly, when they are potentially inaccurate).
3. Parametric Monte Carlo methods for testing hypotheses suffer much the same defects as more traditional methods. That is, both depend on assumptions about random sampling, and

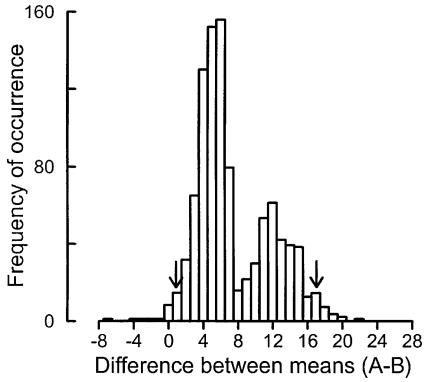


FIGURE 3. Frequency-histogram of the percentile-*t* bootstrap distribution for the difference between two independent means, based on 1000 resamples. Observed values as in Figure 1. Observed difference between means 5.85. ▾ indicate 95% bootstrap percentile-t confidence limits. Note the difference between this frequency-histogram of the percentile-t bootstrap distribution and that of the percentile distribution in Figure 2. From Ludbrook[41] with kind permission of Blackwell Science Asia.

about the frequency distributions of the populations that have been randomly sampled.

4. Nonparametric Monte Carlo methods in general, and bootstrapping in particular, can be grossly inaccurate in estimating the dispersion of populations and in testing hypotheses when the experimental sample sizes are small. "Small" appears to refer to samples of less than 30.

VI. PROGRAMS FOR COMPUTER-INTENSIVE TESTS

This should not be regarded as a comprehensive list, since it includes only programs of which the author is aware.

A. Categorical Data

A very comprehensive set of routines is found in StatXact (Cytel Software Corporation, Cambridge, MA). Some of these same routines are included as special modules in SPSS (SPSS Inc., Chicago) and SAS (SAS Institute Inc., Cary, NC).

B. Permutation Tests

DOS routines are provided by Edgington in his book.[15] Manly has a set of routines in the low-cost DOS program RT, described in his book,[26] and available from the Centre for Applications of Statistics and Mathematics, University of Otago, P.O. Box 56, Dunedin, New Zealand. The commercial program StatXact (Cytel Software Corporation, Cambridge, MA) has powerful routines for permutation tests on continuous and ranked data on a Windows platform.

C. Monte Carlo Methods

Routines for Monte Carlo simulations are usually written by the authors of the papers in which the results are described.

Routines for bootstrapping are described by the authors of several monographs.[39,40,45] They can also be constructed by those who are adept at using statistics programming

languages such as S-PLUS (MathSoft Inc., Seattle, see Reference 44) and SAS (SAS Institute Inc., Cary, NC), and those who are adept at writing macros in programs such as Minitab (Minitab, State College, PA). There are some built-in routines in the most recent version of SYSTAT (SPSS Inc., Chicago). However, the occasional bootstrapper (like the author) is advised to use a spreadsheet. These carry out random resampling with replacement with impressive speed.

REFERENCES

1. Fisher, R.A., The design of experiments, in *Statistical Methods, Experimental Design, and Scientific Inference*, J.H. Bennett, Ed., Oxford University Press, Oxford, 1990.
2. Fisher R.A. and Yates, F., *Statistical Tables for Biological, Agricultural and Medical Research*, 6th edition, Harlow, Longman, London, 1963.
3. Eden, T. and Yates, F., On the validity of Fisher's z test when applied to an actual example of non-normal data, *J. of Agric. Sci.*, 23:6–16, 1993.
4. Efron, B., Bootstrap methods: another look at the jacknife, *Ann. of Stat.*, 7:1–26, 1979.
5. Bergmann, R., Ludbrook, J., and Spooren, W.P.J.M., Different outcomes of the Wilcoxon-Mann-Whitney test from different statistics packages, *The American Statistician*, 2000 (in press).
6. Ludbrook, J. and Dudley, H., Why permutation tests are superior to t or F tests in biomedical research, *The American Statistician*, 52:127–32, 1998.
7. Neyman, J. and Pearson, E.S., On the use and interpretation of certain test criteria for purposes of statistical inference. Part I, *Biometrika*, 20A:175–240, 1928.
8. Lehmann, E.L., *Nonparametrics; Statistical Methods Based on Ranks*, 1st edition revised. Prentice Hall, Upper Saddle River, NJ, 1998.
9. Fisher, R.A., 'The coefficient of racial likeness' and the future of craniometry, *J. of the Royal Anthropological Soc.*, 66:57–63, 1936.
10. Kempthorne, O., The randomization theory of experimental inference, *J. of the Am. Stat. Assoc.*, 50:946–67, 1955.
11. Kempthorne, O. and Doerfler, T.E., The behavior of some significance tests under experimental randomization, *Biometrika*, 56:231–48, 1969.
12. Box, G.E.P. and Anderson, S.L., Permutation theory in the derivation of robust criteria and the study of departures from assumption, *J. of the Royal Stat. Soc.*, B 17:1–26, 1955.
13. Box, G.E.P., Hunter, W.G., and Hunter, J.S., *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*, New York, Wiley, 1978.
14. Feinstein, A.R. Clinical biostatistics. XII. The role of randomization in sampling, testing, allocation, and credulous idolatry (Part 2), *J. of Clin. Pharmacol. and Ther.*, 14:898–915, 1973.
15. Edgington, E.S., *Randomization Tests*, 3rd edition, Marcel Dekker, New York, 1995.
16. Siegel, S. and Castellan, N.J., *Nonparametric Statistics for the Behavioral Sciences*, 2nd edition, McGraw-Hill, New York, 1988.
17. Agresti, A., *Categorical Data Analysis*, John Wiley and Sons, New York, 1990.
18. Davis, L.J., Exact tests for 2×2 contingency tables, *The Am. Stat.*, 40:139–41, 1986.
19. Pearson, K., On the criterion that a given system of deviations from the probable in the case of a correlated system of vari-

ables is such that it can reasonably be supposed to have arisen from random sampling, *Philosophical Magazine*, (5) 50, 157–75, 1900.

20. Fisher, R.A., On the interpretation of χ^2 from contingency tables and the calculation of P, *J. of the Royal Stat. Soc.*, 85:87–94, 1922.

21. Yates, F., Contingency tables involving small numbers and the χ^2 test, *J. of the Royal Stat. Soc., Suppl.* 1, 217–35, 1934.

22. Cochran, W.G., Some methods for strengthening the common χ^2 tests, *Biometrics*, 10:417–51, 1954.

23. Sokal, R.R. and Rohlf, F.J., *Biometry*, 2nd edition, WH Freeman & Co., New York, 1981.

24. Freeman, G.H. and Halton, J.H., Note on exact treatment of contingency, goodness of fit and other problems of significance, *Biometrika*, 38:141–9, 1951.

25. Armitage, P., Test for linear trend in proportions and frequencies, *Biometrics*, 11:375–86, 1955.

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

27. Good, P., *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, Springer-Verlag, New York, 1994.

28. Sprent, P., *Data Driven Statistical Methods*, Chapman & Hall, London, 1998.

29. Wilcoxon, F., Individual comparison by ranking methods, *Biometrics*, 1:80–3, 1945.

30. Mann, H.B. and Whitney, D.R., On a test of whether one of two random variables is stochastically larger than the other, *Ann. of Math. Stat.*, 18, 50–60, 1947.

31. Ludbrook, J., Repeated measurements and multiple comparisons in cardiovascular research, *Cardiovasc. Res.*, 28:303–11, 1994.

32. Neter, J., Wasserman, W., and Kutner, M.H., *Applied Linear Statistical Models*, 3rd edition, Richard D. Irwin, Homewood, IL, 1990.

33. Christopoulos, A., Assessing the distribution of parameters in models of ligand-receptor interaction: to log or not to log, *Trends in Pharmacological Science*, 19:351–7, 1998.

34. Morgan, B.J.T., *Elements of Simulation*, Chapman & Hall, London, 1986.

35. Everitt, B.S., *The Cambridge Dictionary of Statistics*, Cambridge University Press, Cambridge, 1998.

36. Ludbrook, J., Multiple comparison procedures updated, *Clin. and Exp. Pharmacol. and Physiol.*, 25:1032–7, 1998.

37. Concato, J. and Feinstein, A.R., Monte Carlo methods in clinical research: applications in multivariable analysis, *J. of Invest. Med.*, 45:394–400, 1997.

38. Lew, M.J. and Angus, J.A., Analysis of competitive agonist-antagonist interactions by nonlinear regression, *Trends in Pharmacological Science*, 16:328–37, 1995.

39. Efron, B. and Tibshirani, R.J., *An Introduction to the Bootstrap*, Chapman & Hall, New York, 1993.

40. Mooney, C.Z. and Duval, R.D., *Bootstrapping: A Nonparametric Approach to Statistical Inference*, Sage Publications, Newbury Park, 1993.

41. Ludbrook, J., Issues in biomedical statistics; comparing means by computer-intensive tests, *Australian and NZ Journal of Surgery*, 65:812–19, 1995.

42. Hall, P. and Martin, M., On the bootstrap and two-sample problems, *Australian Journal of Statistics*, 30A:179–92, 1988.

43. Fisher, N.I. and Hall, P., On bootstrap hypothesis testing, *Australian Journal of Statistics*, 32:177–90, 1990.

44. Fisher, N.I. and Hall, P., Bootstrap algorithms for small samples, *J. of Stat. Plann. and Inference*, 27:157–69, 1991.

45. Venables, W.N. and Ripley, B.D., *Modern Applied Statistics with S-PLUS*, 2nd edition, Springer, New York, 1997.