The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://amstat.tandfonline.com/loi/utas20

Why Permutation Tests are Superior to t and F Tests in Biomedical Research

John Ludbrook & Hugh Dudley

To cite this article: John Ludbrook & Hugh Dudley (1998) Why Permutation Tests are Superior to t and F Tests in Biomedical Research, The American Statistician, 52:2, 127-132, DOI: 10.1080/00031305.1998.10480551

To link to this article: http://dx.doi.org/10.1080/00031305.1998.10480551

Why Permutation Tests are Superior to t and F Tests in Biomedical Research

John LUDBROOK and Hugh DUDLEY

A survey of 252 prospective, comparative studies reported in five, frequently cited biomedical journals revealed that experimental groups were constructed by randomization in 96% of cases and by random sampling in only 4%. The median group sizes ranged from 4 to 12. In the randomized studies in which measurements were made on a continuous scale, comparisons of location were made by t or F tests in 84% of cases, and by nonparametric, rank-order, tests in the remainder. Because randomization rather than random sampling is the norm in biomedical research and because group sizes are usually small, exact permutation or randomization tests for differences in location should be preferred to t or F tests.

KEY WORDS: Biomedical research; Fisher; Permutation tests; Randomization tests.

1. INTRODUCTION

This review draws attention to a serious misunderstanding between statisticians (especially teachers and consultants) and biomedical scientists who employ statistical analyses. It can be summed up as follows. Statisticians appear to believe that biomedical researchers do most experiments by taking random samples, and therefore recommend statistical procedures that are valid under the population model of inference. In fact, randomization of a nonrandom sample, not random sampling, is more usual. Given this, it is our thesis that the randomization rather than population model applies, and that the statistical procedures best adapted to this model are those based on permutation.

2. DESIGN AND ANALYSIS OF BIOMEDICAL EXPERIMENTS

We have reviewed the design of experiments in five, frequently cited life-sciences journals in our own fields, published on either side of the Atlantic (Tab. 1). Of 252 prospective and comparative studies, only 4% used random sampling of defined populations to construct experimental groups and all these employed inbred strains of animals. In the remainder, experimental groups were constructed by randomization of nonrandom samples of humans, animals, tissues, or cells to two or more conditions or treatments. Furthermore, the group sizes used in the randomized studies were small. Overall, the median group size varied from 6 to 12 (Tab. 1). In 216 randomized-design experiments performed in a laboratory setting, the median group size was 6 (range 2–77). In 25 randomized clinical trials, the median size was 25 (range 4–345).

We also reviewed the statistical procedures used to test for differences in location when measurements had been made on a continuous scale (Tab. 2). In randomized-design experiments, the results were analyzed by t or F tests in 84% of cases. In the remainder, nonparametric tests based on rank-order were used, with one exception. In the last case the difference beween means was tested by permutation, though the test was done incorrectly.

3. TWO MODELS OF STATISTICAL INFERENCE

3.1 The Population Model

Neyman and Pearson formally proposed this model in 1928. It assumes there has been random sampling of a population or populations. Under it, the level of statistical significance (P) that results from applying a statistical test to the results of an experiment corresponds to the frequency with which the null hypothesis would be rejected in repeated random samplings of those populations. Because it would be inconvenient or impossible to sample real populations repeatedly, it is stipulated that the sampling distribution of those populations conforms to a theoretical frequency-distribution (for instance, the t or F distributions). Neyman and Pearson (1928) also formalized the concept of two sources of error in statistical inference. The first (Type I) refers to the risk of falsely rejecting the null hypothesis; in ideal conditions, this coincides with the P value resulting from a test of that hypothesis. The second (Type II) is false acceptance of the null hypothesis, which leads on to the concept of the power of tests to reject the null hypothesis and the influence of sample size in this context. In a later article they also addressed the trade-off between controlling Type I and Type II error rates, arguing in favor of the latter in scientific research (Neyman and Pearson 1933). However, in biomedical research we have a strong preference for controlling Type I error, because the penalty for false-positive inference may be the introduction of a valueless new therapy. Neyman (1934) was also the first to propose the use of confidence intervals as an alternative to hypothesis-testing, a practice that has recently become popular in clinical research.

The Neyman–Pearson population model of inference is adopted, explicitly or implicitly, in almost every explanatory textbook on general statistics. Under it, statistical tests

John Ludbrook is a Professorial Research Fellow in the University of Melbourne Department of Surgery, Royal Melbourne Hospital, Parkville, Victoria 3050, Australia (E-mail: JohnLudbrook@Bigpond.com). Hugh Dudley is Professor Emeritus at London University, Department of Surgery, St. Mary's Hospital Medical School (Imperial College), Norfolk Place, Paddington, London W2 1PG, United Kingdom. John Ludbrook is supported by the National Health and Medical Research Council of Australia.

Table 1. Experimental Design and Group Sizes in Comparative Studies Published in Five Biomedical Journals in 1993—1994

Journal	Total no. articles (lab:clin ratio)	Randomization		Random sampling	
		No. studies (lab:clin ratio)	Median group size (range)	No. studies (lab:clin ratio)	Median group size (range)
American Journal of Physiology (Heart Circ. Physiol.)	101 (95:6)	61 (59:2)	7 (3–36)	3 (3:0)	6 (4–10)
British Journal of Pharmacology	92 (92:0)	76 (76:0)	6 (3–77)	3 (3:0)	9 (4–13)
Circulation Research	107 (104:3)	50 (49:1)	6 (3–77)	3 (3:0)	8 (5–12)
Annals of Surgery	100 (22:78)	20 (13:7)	10 (2–38)	1 (1:0)	4
British Journal of Surgery	132 (21:111)	34 (19:15)	12 (2–345)	1 (1:0)	7

NOTES: Data on original articles from consecutive issues. Lab:clin ratio, ratio of articles on laboratory experiments versus clinical trials. Discrepancy between total number of articles reviewed and those categorized as of randomized or random sampling design is accounted for by articles that fit neither classification because they were descriptive or retrospective.

that depend on sampling distributions such as the normal, t, or F are valid.

3.2 The Randomization Model

In the early 1930s, R.A. Fisher proposed that randomization should be the basis for experimental design and statistical inference (Fisher 1936, 1971). The premise behind this model is that a sample of experimental units, however acquired, is divided randomly into two or more groups. These are then exposed to different conditions or treatments. The null hypothesis is merely that the conditions or treatments have no differential effects on the groups with respect to a selected statistic such as the mean. There is no reference to a population and therefore no requirement that it should conform to a mathematically definable frequency distribution. Instead, for each experiment the unique sampling distribution of the test statistic is compiled exactly by permutation.

The randomization model has had relatively few supporters. Kempthorne and Box have been strong advocates of it (Kempthorne 1955; Box and Anderson 1955; Kempthorne and Doerfler 1969; Box, Hunter, and Hunter 1978). Other distinguished statisticians have described the model (for instance, Welch 1937a; Scheffé 1959, p. 313; Lehmann 1975, pp. 55–57), but they seem hesitant to recommend it because of the restrictive nature of inferences under it. Indeed, a contemporary theoretical statistician wrote to one of us (JL): "I do wonder, though, why one would call this *inference*?" (his emphasis). Of those who use statistical procedures to analyze their experimental results, only a few have been outspoken in favor of the randomization model (for instance, Feinstein 1973; Edgington 1995; Ludbrook 1994; Ludbrook and Dudley 1994).

3.3 The Two Models Compared

Under the population model, a statistical inference has

Table 2. Methods of Statistical Analysis Used for Continuous Data in Comparative Studies Published in Five Biomedical Journals in 1993—1994

Experimental design	Analytical procedures			
	Classical	Rank permutation	General permutation	Total
American Journal of Physiology (Heart Circ Physiol)				
Random sampling	3	0	0	3
Randomization	56	5	0	61
British Journal of Pharmacology				
Random sampling	2	1	0	3
Randomization	65	8	0	73
Circulation Research				
Random sampling	3	0	0	3
Randomization	48	1	0	49
Annals of Surgery				
Random sampling	0	0	1	1
Randomization	15	2	0	17
British Journal of Surgery				
Random sampling	0	1	0	1
Randomization	5	20	0	25
Total	197	38	1	236

NOTES: Classical tests: based on t or F distributions. Rank permutation tests: Wilcoxon–Mann–Whitney, Wilcoxon matched pairs, Kruskal–Wallis, Friedman. General permutation tests: based on permutation distribution of differences between means. Not included in the table are seven cases in which no analytical test was used.

implications for the future even if these are unstated. If the outcome of a t or F test under random sampling is $P = .05$, it is implied that if repeated random samples of the same size were taken from the same population(s), 19 out of 20 would yield a difference between sample means of the same size or greater than that originally observed. But if $P = .05$ results from randomization and the application of a permutation test, there is no statistical promise for the future. The statistical inference refers only to the actual experiment that has been performed and the P value indicates the probability that the way in which this experiment turned out was merely a matter of chance. However, this need not deter experimenters from inferring that their results are applicable to similar patients, animals, tissues, or cells, though their arguments must be verbal rather than statistical.

Though the notion of Type I error can be applied to inferences under the randomization model, some of the other elements of the population model cannot. For instance, the relationship between power to reject the null hypothesis and sample size is a continuous one under random sampling, whereas it is discontinuous under randomization (see later). And confidence intervals cannot be used as an alternative to hypothesis-testing under the randomization model, for there are no true parent populations to which they could be referred.

3.4 A Hybrid Model

What if the sample that is subjected to randomization were taken randomly from a defined population? In this circumstance, it is reasonable to suppose that the statistical inference or confidence intervals could be extended to refer to other hypothetical experiments in which the same population was randomly sampled and then divided by randomization into subsamples.

4. PERMUTATION TESTS

4.1 History

The first edition of R.A. Fisher's *The Design of Experiments* was published just over 60 years ago. Its popularity can be judged from the fact that the eighth edition (1966) was reprinted in 1971 and again in 1990 (Fisher 1971). The book is nowadays read chiefly for its exposition of the principles and practice of randomized experimental designs. However, it also contains first descriptions of two tests of significance that depend on permutation. One, Fisher's exact test for analyzing categorical data set out as a 2×2 table of frequencies, has entered the mainstream of statistical practice. The other was a permutation test for the difference between means.

Fisher took Charles Darwin's data on the height of cross- and self-fertilized *Zea mays* plants and analyzed them by a permutation test for the difference between the two means (Fisher 1971, p. 30). This involved the compilation by hand of the 32,768 possible permutations of the data, with the proviso that the permutations be divided into groups of the same size as those in the experiment. In 1726 of the permutations the difference between the group means was equal to or greater than that observed. The two-sided value for P is thus $1{,}726/32{,}768 = .05267$. He had earlier used Student's pooled-variance t test to analyze the same data, which gave $P = .0497$. His conclusion was that permutation tests provide "the possibility of an independent check on the more expeditious [classical] methods in common use" (Fisher 1971, p. 45). He did confess, however, that "the one flaw in Darwin's procedure was the absence of randomisation" (Fisher 1971, p. 44). At about the same time, Eden and Yates (1933) used permutation to analyze an agricultural experiment based on a randomized-block design with replicates. Their purpose was to defend analysis of variance against charges that it is very sensitive to departures from normality. Fisher's z (now the F) statistic was permuted by taking a random sample ($n = 1{,}000$) of the 24^8 possible permutations of the data. From the close correspondence between the permutation and theoretical distributions of F they concluded that "the z test may safely be applied to [skewed] data of this type."

It appears that Fisher then lost interest in the permutation method for analyzing continuous data and he did not mention it again in his publications or scientific correspondence (Bennett 1990). Part of the reason may have been the exhausting process of compiling permutations by hand, so that as Bradley wrote 30 years later: "[permutation tests were] little more than curiosities ... almost never quick ... seldom practical, and often ... not even feasible" (Bradley 1968, p. 84).

There was, in addition, a theoretical difficulty in accepting permutation tests. Fisher was often enigmatic in the views he expressed about statistical inference. Neyman and Pearson were not. Their theoretical expositions on inference from random sampling were formal and precise, and their population model of inference became widely accepted from the mid-1930s onward. As a result, most of the further theoretical and empirical development of permutation tests was done, somewhat illogically, under the Neyman–Pearson population model. Beginning with Pitman (1937), a number of statisticians showed how the scope of permutation tests could be extended beyond those envisaged by Fisher, Eden, and Yates. Others suggested approximate versions of exact permutation tests (for instance, Pitman 1937; Box and Anderson 1955) or developed further Eden and Yates's (1933) notion of sampled permutation tests, sometimes called randomization tests (see Edgington 1995; Manly 1997). There have been a great many studies designed to confirm the asymptotic, large-sample, equivalence of permutation and classical tests (see Ludbrook 1994). Others have pointed out discrepancies between permutation and classical tests when population variances are unequal and especially when samples are small (for instance, Boik 1987; Romano 1990). For completeness, we refer also to the nonparametric tests whose theoretical basis is the permutation of differences between mean-ranks, invented by Wilcoxon in 1945 as a way of overcoming the computational difficulties of permuting differences between means. It was later demonstrated that if rank-order tests are used to test specifically for equality of location, under the population model it is required that the shapes and dispersions of the randomly sampled pop-

ulations be identical (see Bradley 1968, p. 109; Ludbrook 1996). It should also be noted that the null hypothesis—equality of mean-ranks—does not correspond to equality of medians, because of the method of ranking employed in these test procedures.

4.2 Principles

Lucid descriptions of how permutation tests are conducted can be found in successive editions of the specialized monographs by Edgington (1995) and Manly (1997), and in Siegel and Castellan (1988) (see also Fig. 1). Their underlying rationale is rarely mentioned in texts of general statistics (a notable exception being that of Box et al. 1978). It is referred to in texts on nonparametric statistics, though the focus is usually on rank-order tests rather than on permutation tests in general (exceptions are Bradley 1968; Siegel and Castellan 1988). There have been occasional attempts to explain permutation tests to biomedical investigators (for instance, Feinstein 1973; Ludbrook 1994).

In the case of independent groups, these tests depend on compiling all possible permutations of the values that result from an experiment, provided the permutations are divided into sets of the same size as the randomized experimental groups. When repeated measurements are made on the same set of experimental units, all possible interchanges of the values attached to each experimental unit are compiled. The null hypothesis is that there is no differential effect of experimental treatments or conditions on the statistic of interest. The probability attached to the null hypothesis is then

$$\frac{\text{Number of the same or more extreme outcomes as that observed}}{\text{Total number of possible outcomes}}.$$

The outcome can be defined in many different ways: for instance, as the difference between arithmetic means, geometric means, medians, mid-ranges, mean-ranks, proportions, or variances. In the case of multiple randomized groups, an F-like statistic can be permuted.

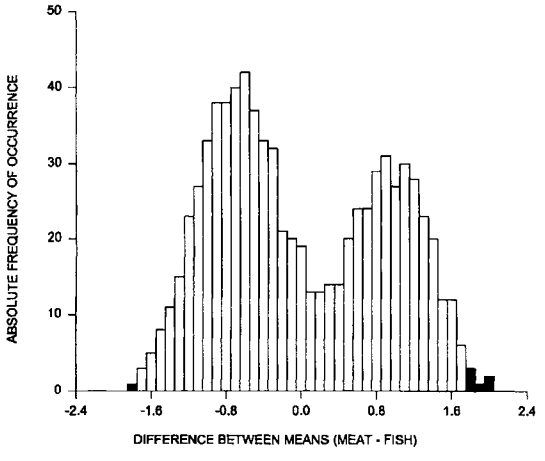
One of the practical difficulties with exact permutation tests is the number of permutations required. For instance, for k independent groups of size n_x, the general formula for calculating the number of possible permutations is

$$(n_1 + n_2 + n_3 + \cdots + n_k)!/\,(n_1)!(n_2)!(n_3)!\ldots(n_k)!\,.$$

For k repeated measurements on the same group of size n, the general formula is k^n. A little work with a handheld calculator will show that with just two independent groups, where $n_1 = n_2 = 15$ the number of possible permutations is about 1.5×10^8; and for two repeated measurements made on a group of size $n = 15$ the number is 32,768.

Even with a fast microcomputer, compilation of the exact permutation distribution may take too long. The solution is to construct the distribution by taking a Monte Carlo random sample from all the possible permutations, in much the same way as Eden and Yates did by hand in 1933. The result is a P value that is only marginally less exact and to which confidence limits can be attached. The Monte Carlo solution is sometimes called a randomization test (Edgington 1995; Manly 1997).

The number of possible permutations is important in another way. If it is < 20, then $P \leq .05$ can never be attained. For two independent groups, this critical number of 20 permutations is attained when $n_1 = n_2 \geq 3$, and for two related groups when $n \geq 5$. Thus, though it is difficult to construct a theoretical statement about power to reject the null hypothesis under randomization, it is easy to arrive empirically at minimum group sizes that will allow the null hypothesis to be rejected at a chosen level of significance.



Permutation no:	1	787*	788	789	790	791	792
Fish (F)	11.5	5.42	5.42	5.42	5.42	5.42	5.42
	7.84	5.86	5.86	5.86	5.86	5.86	5.86
	7.61	6.16	6.16	6.16	6.16	6.16	6.16
	7.56	6.55	6.51	6.51	6.51	6.51	6.51
	7.11	6.80	6.55	6.80	6.55	6.55	6.55
	7.00	7.00	6.80	7.00	7.00	6.80	6.80
	6.80	7.11	7.56	7.11	7.11	7.11	7.00
Meat (M)	5.42	6.51	7.00	6.55	6.80	7.00	7.11
	5.86	7.56	7.11	7.56	7.56	7.56	7.56
	6.16	7.61	7.61	7.61	7.61	7.61	7.61
	6.51	7.84	7.84	7.84	7.84	7.84	7.84
	6.55	11.5	11.5	11.5	11.5	11.5	11.5
Difference between means (M - F)	-1.82	1.79*	1.80	1.80	1.89	1.96	2.00

*Figure 1. Outcome of Permutation Test for Difference Between Group Means. Above: Permutation distribution of absolute frequency of occurrence according to difference between means (meat eaters-fish eaters). All possible permutations: 792. Filled columns: difference between means equals or exceeds observed value in either direction. Below: Permutations of values in which difference between means (meat eaters-fish eaters) is equal to or exceeds in either direction the observed difference of 1.79 (indicated by *). Under the randomization model, the two- sided probability that diet has a differential effect on plasma cholesterol concentration is 7/792 = .0088.*

4.3 Example of a Permutation Test for the Difference Between Means of Two Independent Groups

This example is a hypothetical one. Twelve men are recruited from among those attending a fitness clinic and asked to participate in an experiment to establish whether eating fish (but not meat) results in lower plasma cholesterol concentrations than eating meat (but not fish). The subjects are randomly allocated to the fish-eating ($n_1 = 7$) and meat-eating regimens ($n_2 = 5$). At the end of one year their plasma cholesterol concentrations are measured. These

are:

Fish eaters: 5.42, 5.86, 6.16, 6.55, 6.80, 7.00, 7.11
Meat eaters: 6.51, 7.56, 7.61, 7.84, 11.50

An exact permutation test is performed to test whether the treatments had a differential effect on plasma cholesterol concentration. There are 792 possible permutations of the data, which are set out as a frequency-histogram for the differences between group means in Figure 1. The bimodal and asymmetric shape of the permutation distribution bears no resemblance to the symmetrical t distribution. In seven of these permutations the difference between group means is equal to or exceeds in one or other direction the observed difference of 1.79 (Fig. 1). This corresponds to a two-sided value for $P = 7/792 = .0088$. Under the randomization model, the experimenters can therefore reject the null hypothesis with some confidence, and infer that diet did have a differential effect on plasma cholesterol concentration in their nonrandom sample of men. But they could only extend this inference to other men attending fitness clinics, or to men in general, by verbal argument.

4.4 Other Ways of Analyzing the Same Data Set

Most statisticians would very likely take a different approach and proceed under the population model, and our survey certainly suggests that biomedical investigators and their statistical advisors would do so (Tab. 2). They would note the outlying value in the meat- eating group and confirm that it did not result from an error of measurement or transcription. Given that the sample variance ratio is 7.16 and that the larger variance is associated with the smaller sample, they would be wary of using the pooled-variance t test because of the risk that the Type I error rate may exceed that nominated. Some would use Welch's (1937b) separate-variance version of the t test, which minimizes this risk (Fligner and Policello 1981). Others might log-transform the data before performing a t or Welch test, though this reduces the variance ratio only to 4.47. Biomedical investigators—and surgical ones in particular (Tab. 2)—appear to favor using the Wilcoxon–Mann–Whitney procedure, though they do not appreciate the restrictions to its use under the population model and the difficulty in interpreting the mean-rank as an index of location (see Sec. 4.1).

Table 3. Outcome of Tests of Significance on the Data From the Hypothetical Experiment Comparing Plasma Cholesterol Concentrations After Fish- or Meat-Eating

Test procedure	Untransformed values P	Log-transformed values P
Student's t test	.041	.029
Welch test	.105	.070
Exact WMW test	.030	.030
Exact permutation test	.009	.013

NOTES: For original values see text and Figure 1. Student's t test: pooled variance at df = 10. Welch test: separate variances, adjusted df. Exact WMW: Wilcoxon–Mann–Whitney test by exact permutation. Exact Permutation Test: all 792 possible permutations of differences between group means listed.

The outcome of these several approaches is in Table 3. Our main point is that there are marked discrepancies between P values that result from permutation and those from parametric tests. Though we have referred earlier to the ample theoretical and empirical evidence that classical tests provide good large-sample approximations to the corresponding permutation tests, it is clear that this need not be so when group sizes are small. It should also be noted that the two permutation tests give different outcomes. The Wilcoxon–Mann–Whitney test is unaffected by log-transformation, because this does not alter the order of the difference between mean ranks. It is also less sensitive than the permutation test for a difference between group means, because information is discarded when the original values are transformed into ranks.

5. CONCLUSION

Randomized rather than random-sampling designs are used in most comparative biomedical experiments. On the basis of pure theory, statistical inferences from the experiments are valid only under the randomization model of inference. Why, then, do biomedical investigators not employ exact or sampled permutation tests to analyze their results?

A trivial reason is that the editors of biomedical journals might not understand permutation tests and their statistical advisers might not accept the arguments we have put forward. Our personal experience is that it is much easier to get a manuscript published if one stays with classical tests under the population model.

There is also an important practical point. There are plenty of microcomputer statistical software packages with which to perform classical or modified t and F tests, but a dearth of software for performing permutation tests for differences between means. Those packages that we know to be available are listed in Appendix A.

Because the sets of continuous data acquired by biomedical investigators are rarely published, we cannot tell how often their statistical inferences are seriously flawed. But the small group sizes they use in their experiments (Tab. 1), their propensity for employing nonparametric rank-order tests if they have doubts that the assumptions for t and F tests are fulfilled (Tab. 2), and the discrepancy between the results of classical and permutation tests that can be demonstrated in individual examples (Fig. 1), leads us to suspect that the problem is not a trivial one.

Kempthorne (1955, p. 966) wrote: "When one considers the whole problem of experimental inference, that is of tests of significance, estimation of treatment differences and estimation of the errors of estimated differences, there seems little point in the present state of knowledge in using [a] method of inference other than randomization analysis." Those words summarize the substance of this review. We hope that biomedical scientists, statistical teachers and consultants, and the editors of biomedical journals and their statistical advisors will take note of them.

APPENDIX: MICROCOMPUTER SOFTWARE FOR PERMUTATION TESTS

We are aware of only three (and a half) pieces of software that cater for differences between means. Edgington's (1995) RANDIBM program provides exact permutation tests for the t and F statistics up to 30,000 permutations, and sampled permutation (randomization) tests thereafter. Manly's (1997) RT 2.1 software provides a similar range. The only fully commercial microcomputer software package that we know of is STATXACT 3.0 for Windows (Cytel Software Corp., Cambridge, MA). Its permutation algorithm is very efficient (up to 6×10^5 permutations per second). For large data sets, Monte Carlo sampled permutation tests can be done and confidence intervals for P are given. It caters for one-way designs with k groups, but for only two related groups. SAS (SAS Institute Inc., Cary, NC) offers permutation tests, but only for making multiple pairwise contrasts between means.

[Received June 1995. Revised March 1996.]

REFERENCES

Bennett, J.H. (ed) (1990), *Statistical Inference and Analysis: Selected Correspondence of R.A. Fisher*, Oxford: Clarendon Press.

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.

Box, G.E.P., and Anderson, S.L. (1955), "Permutation Theory in the Derivation of Robust Criteria and the Study of Departures from Assumption," *Journal of the Royal Statistical Society*, Ser. B, 17, 1–26.

Box, G.E.P., Hunter, W.G., and Hunter, J.S. (1978), *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*, New York: Wiley.

Bradley, J.V. (1968), *Distribution-Free Statistical Tests* (2nd ed.), Eaglewood Cliffs, NJ: Prentice Hall.

Eden, T., and Yates, F. (1933), "On the Validity of Fisher's z Test When Applied to an Actual Example of Non-Normal Data," *Journal of Agricultural Science*, 23, 6–16.

Edgington, E. S. (1995), *Randomization Tests* (3rd ed.), New York: Marcel Dekker.

Feinstein, A.R. (1973), "Clinical Biostatistics. XIII. The Role of Randomization in Sampling, Testing, Allocation, and Credulous Idolatry (Part 2)," *Journal of Clinical Pharmacology and Therapeutics*, 14, 898–915.

Fisher, R.A. (1936), "'The Coefficient of Racial Likeness' and the Future of Craniometry," *Journal of the Royal Anthropological Society*, 66, 57–63.

——— (1971), *The Design of Experiments* (8th ed.), New York: Hafner Publishing. Reprinted in Bennett, J.H. (ed.) (1990), *Statistical Methods, Experimental Design, and Scientific Inference*, Oxford: Oxford University Press.

Fligner, M.A., and Policello, G.E. (1981) "Robust Rank Procedures for the Behrens-Fisher Problem," *Journal of the American Statistical Association*, 76, 162–168.

Kempthorne, O. (1955), "The Randomization Theory of Experimental Inference," *Journal of the American Statistical Association*, 50, 946–967.

Kempthorne, O., and Doerfler, T.E. (1969), "The Behavior of Some Significance Tests Under Experimental Randomization," *Biometrika*, 56, 231–248.

Lehmann, E.L. (1975), *Nonparametrics: Statistical Methods Based on Ranks*, San Francisco: Holden Day.

Ludbrook, J. (1994), "Advantages of Permutation (Randomization) Tests in Clinical and Experimental Pharmacology and Physiology," *Clinical and Experimental Pharmacology and Physiology*, 21, 673–686.

——— (1996), "The Wilcoxon–Mann–Whitney Test Condemned," (letter), *British Journal of Surgery*, 83, 136–137.

Ludbrook, J., and Dudley, H.A.F. (1994), "Issues in Biomedical Statistics: Statistical Inference," *Australian and New Zealand Journal of Surgery*, 64, 630–636.

Manly, B.F.J. (1997), *Randomization, Bootstrap and Monte Carlo Methods in Biology* (2nd ed.), London: Chapman & Hall.

Neyman, J. (1934), "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection," *Journal of the Royal Statistical Society*, 97, 558–606.

Neyman, J., and Pearson, E.S. (1928), "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference. Part I," *Biometrika*, 20A, 175–240.

Neyman, J., and Pearson, E.S. (1933), "The Testing of Statistical Hypotheses in Relation to Probabilities a priori," in *Proceedings of the Cambridge Philosophical Society*, 20, pp. 492–510.

Pitman, E.J.G. (1937), "Significance Tests Which May be Applied to Samples from Any Population," *Journal of the Royal Statistical Society*, Ser. B, 4, 119–130.

Romano, J.P. (1990), "On the Behavior of Randomization Tests without a Group Invariance Assumption," *Journal of the American Statistical Association*, 85, 686–692.

Scheffé, H. (1959), *The Analysis of Variance*, New York: Wiley.

Siegel, S., and Castellan, N.J. (1988), *Nonparametric Statistics for the Behavioral Sciences* (2nd ed.), New York: McGraw-Hill.

Welch, B.L. (1937a), "On the z-Test in Randomized Blocks and Latin Squares," *Biometrika*, 29, 21–52.

——— (1937b), "The Significance of the Difference Between Two Means When the Population Variances are Unequal," *Biometrika*, 29, 350–362.

Wilcoxon, F. (1945), "Individual Comparisons by Ranking Methods," *Biometrics*, 1, 80–83.