Statistical Science
2004, Vol. 19, No. 4, 676–685
DOI 10.1214/088342304000000396

Permutation Methods: A Basis for Exact Inference

Michael D. Ernst

Abstract. The use of permutation methods for exact inference dates back to Fisher in 1935. Since then, the practicality of such methods has increased steadily with computing power. They can now easily be employed in many situations without concern for computing difficulties. We discuss the reasoning behind these methods and describe situations when they are exact and distribution-free. We illustrate their use in several examples.

Key words and phrases: Distribution-free, Monte Carlo, nonparametric, permutation tests, randomization tests.

1. INTRODUCTION

Textbooks commonly used for a first course in nonparametrics have rarely included permutation methods in any depth. The books by Bradley (1968) and Lehmann (1975) contain excellent introductions to the ideas of permutation methods, but both are limited by the computing power of the day and the former is no longer in print. A more recent book by Higgins (2004) incorporates permutation methods nicely.

The basic idea behind permutation methods is to generate a reference distribution by recalculating a statistic for many permutations of the data. Fisher (1936) wrote that "the statistician does not carry out this very simple and very tedious process, but his conclusions have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method." Today, with fast computers, there is little reason for the statistician not to carry out this "very tedious process."

The probability basis for statistical inference in these procedures depends on the situation, but can largely be grouped into two probability models: where available subjects are randomly assigned to treatments and where subjects are randomly sampled from some population(s). Lehmann (1975) called the former situation the *randomization model* and the latter the *population model*. Edgington (1995) pointed out that the procedures used under the randomization model are commonly called *randomization tests* or *randomization intervals*, while the same procedures used under the population model are called *permutation tests* or *permutation intervals*. Unfortunately, this distinction is often overlooked, and the terms "randomization" and "permutation" are often used interchangeably. While the distinction in terminology may not be that important, understanding the underlying probability model is important. In addition to the terminology for each probability model, we also use the term "permutation methods" to generically refer to the methods under either model, as we have in the title of this paper.

In Section 2, we describe the idea of exact inference and argue why it is desirable. In Sections 3 and 4, we describe the rationale behind permutation methods under the randomization model and the population model, and show several examples of their use. In Section 5 we briefly discuss more complicated designs and give an example.

2. EXACT INFERENCE

The possibility of making an error is inherent to any statistical inference procedure. These procedures are generally constructed to control and quantify the probability of an error through significance levels and confidence coefficients. Approximate inference methods succeed at controlling errors with varying degrees of success. Exact inference methods guarantee control of the relevant errors. These methods are at their most useful when they can control the errors under relatively

Michael D. Ernst is Assistant Professor, Department of Mathematical Sciences, Indiana University–Purdue University Indianapolis, Indianapolis, Indiana 46202-3216, USA (e-mail: mernst@math.iupui.edu).

broad assumptions. We illustrate this idea with the one-sample Student's t-test and interval, which are exact only under very restrictive assumptions.

2.1 Hypothesis Tests

Hypothesis tests involve Type I errors (rejecting a true null hypothesis) and Type II errors (failing to reject a false null hypothesis). We try to control the probability of a Type I error by choosing an appropriate significance level. What is important to understand is that the probability of a Type I error does not necessarily equal the chosen significance level. For example, suppose that $Y_1, Y_2, \ldots, Y_n$ are a random sample from a normal distribution with mean μ and we wish to test the null hypothesis $H_0: \mu = \mu_0$ versus the alternative hypothesis $H_1: \mu \neq \mu_0$. If we use a significance level of α, we reject H_0 when

$$|T| = \left| \frac{\bar{Y} - \mu_0}{S_Y/\sqrt{n}} \right| \geq t\left(1 - \frac{\alpha}{2}, n-1\right),$$

where $\bar{Y} = \sum_{i=1}^{n} Y_i/n$, $S_Y^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2/(n-1)$ and $t(1-\alpha/2, n-1)$ is the $1-\alpha/2$ percentile of the t distribution with $n-1$ degrees of freedom. Since T has a t distribution with $n-1$ degrees of freedom under H_0, it follows that

$$P(\text{Type I error}) = P\big(|T| \geq t(1-\alpha/2, n-1) | H_0\big) = \alpha.$$

In other words, the probability of a Type I error is exactly equal to the significance level we have chosen. However, if the random sample does not come from a normal distribution, then T does not have a t distribution and $P(\text{Type I error})$ can differ from α, sometimes substantially, and so we really do not have as much control of $P(\text{Type I error})$ as the chosen value of α leads us to believe. A hypothesis test for which $P(\text{Type I error}) = \alpha$ is called an *exact test*. The one-sample t-test is guaranteed to be an exact test only when the data come from a normal distribution.

2.2 Confidence Intervals

For confidence intervals, we specify the confidence coefficient $1-\alpha$ with the desire that it accurately reflect the true coverage probability. In our previous example, the coverage probability of the one-sample t interval is

$$P\left(\bar{Y} - t\left(1 - \frac{\alpha}{2}, n-1\right)\frac{S_Y}{\sqrt{n}} \leq \mu \leq \bar{Y} + t\left(1 - \frac{\alpha}{2}, n-1\right)\frac{S_Y}{\sqrt{n}}\right) = 1-\alpha$$

only when the data come from a normal distribution. If the data do not come from a normal distribution, then the true coverage probability can be different from the chosen confidence coefficient. A confidence interval is called an *exact confidence interval* if the true coverage probability is equal to the confidence coefficient. The one-sample t interval is guaranteed to be exact only when the data come from a normal distribution.

3. THE RANDOMIZATION MODEL

The sole basis for inference in the randomization model is the random assignment of available subjects to treatment groups. It is not necessary to have random sampling from some population with a specified distribution. Strictly speaking, normal theory methods are not appropriate since their distribution theory depends on random sampling. The consequence of this is that any inferences in the randomization model are limited to the subjects in the study.

3.1 The Randomization Distribution

Suppose that a new treatment for postsurgical recovery is being compared to a standard treatment by observing the recovery times (in days) of the patients on each treatment. Of the N subjects available for the study, n are randomly assigned to receive the new treatment, while the remaining $m = N - n$ receive the standard treatment. The null and alternative hypotheses of interest are

H_0: There is no difference between the treatments,

H_1: The new treatment decreases recovery times.

Denote the recovery times for the standard and new treatments by $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$, respectively. To measure the difference between the treatments, we might calculate the difference in mean recovery times between the two groups, $T = \bar{Y} - \bar{X}$. We wish to determine if this difference is extreme enough in some reference distribution to suggest that the new treatment decreases recovery times.

To illustrate, consider the first row of Table 1, which displays the recovery times for $n = 4$ subjects on the new treatment and $m = 3$ subjects on the standard treatment. The difference in their mean recovery times is $\bar{Y} - \bar{X} = -9$ days, indicating a possible decrease in recovery times.

If the null hypothesis is true and there is no difference between the treatments, then the recovery time for each subject will be the same *regardless of which*

TABLE 1
All possible randomizations of seven recovery times (days) to two treatment groups of sizes $n = 4$ and $m = 3$

No.	Randomization: New treatment				Standard treatment			Difference in means	Sum of new	Sum of standard	Difference in medians
▲ 1	19	22	25	26	23	33	40	−9.00	92	96	−9.5
2	22	23	25	26	19	33	40	−6.67	96	92	−9.0
3	22	33	25	26	19	23	40	−0.83	106	82	2.5
4	22	25	26	40	19	23	33	3.25	113	75	2.5
5	19	23	25	26	22	33	40	−8.42	93	95	−9.0
6	19	25	26	33	22	23	40	−2.58	103	85	2.5
7	19	25	26	40	22	23	33	1.50	110	78	2.5
8	19	22	23	26	25	33	40	−10.17	90	98	−10.5
9	19	22	26	33	23	25	40	−4.33	100	88	−1.0
10	19	22	26	40	23	25	33	−0.25	107	81	−1.0
11	19	22	23	25	26	33	40	−10.75	89	99	−10.5
12	19	22	25	33	23	26	40	−4.92	99	89	−2.5
13	19	22	25	40	23	26	33	−0.83	106	82	−2.5
14	23	25	26	33	19	22	40	−0.25	107	81	3.5
15	22	23	26	33	19	25	40	−2.00	104	84	−0.5
16	22	23	25	33	19	26	40	−2.58	103	85	−2.0
17	19	23	26	33	22	25	40	−3.75	101	87	−0.5
18	19	23	25	33	22	26	40	−4.33	100	88	−2.0
19	19	22	23	33	25	26	40	−6.08	97	91	−3.5
20	23	25	26	40	19	22	33	3.83	114	74	3.5
21	22	23	26	40	19	25	33	2.08	111	77	−0.5
22	22	23	25	40	19	26	33	1.50	110	78	−2.0
23	19	23	26	40	22	25	33	0.33	108	80	−0.5
24	19	23	25	40	22	26	33	−0.25	107	81	−2.0
25	19	22	23	40	25	26	33	−2.00	104	84	−3.5
26	25	26	33	40	19	22	23	9.67	124	64	7.5
27	22	26	33	40	19	23	25	7.92	121	67	6.5
28	22	25	33	40	19	23	26	7.33	120	68	6.0
29	19	26	33	40	22	23	25	6.17	118	70	6.5
30	19	25	33	40	22	23	26	5.58	117	71	6.0
31	19	22	33	40	23	25	26	3.83	114	74	2.5
32	23	26	33	40	19	22	25	8.50	122	66	7.5
33	23	25	33	40	19	22	26	7.92	121	67	7.0
34	22	23	33	40	19	25	26	6.17	118	70	3.0
35	19	23	33	40	22	25	26	4.42	115	73	3.0

treatment is received. For example, the subject in Table 1 who recovered in 19 days on the new treatment would have recovered in the same amount of time on the standard treatment if there is no treatment effect.

So the recovery times are not random (because the subjects were not chosen randomly); only their assignment to the treatments is random. Therefore, the basis for building a probability distribution for $\bar{Y} - \bar{X}$ comes from the randomization of the available subjects to the treatments. This randomization results in n subjects getting the new treatment and m subjects getting the standard treatment, but this is just one of $\binom{N}{n}$ equally likely randomizations that could have occurred. A probability distribution for $\bar{Y} - \bar{X}$, called the *randomization distribution*, can be constructed by calculating $\bar{Y} - \bar{X}$ for each of the possible randomizations. Table 1 lists all $\binom{7}{4} = 35$ possible randomizations of the seven observed recovery times into groups of sizes $n = 4$ and $m = 3$, and the difference in means for each randomization. The probability under H_0 of any one of these randomizations is $1/35$. Figure 1 displays the randomization distribution of the difference in means, with the observed difference marked with a solid triangle (▲).

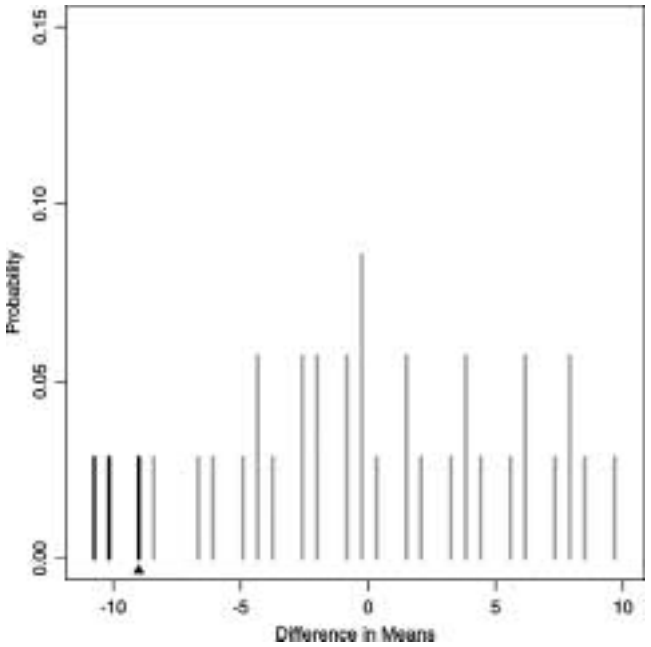

FIG. 1. *The randomization distribution of the difference in means from Table* 1.

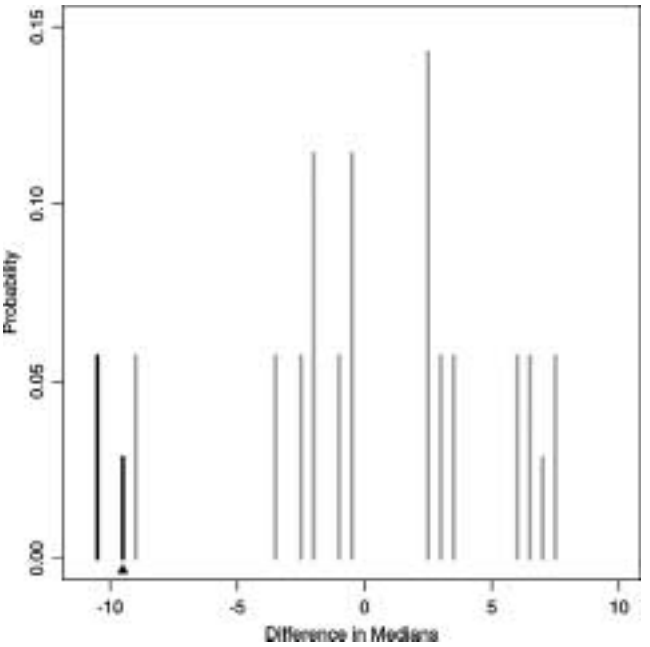

FIG. 2. *The randomization distribution of the difference in medians from Table* 1.

3.2 The Randomization p Value

The p value of the *randomization test* of H_0 can be calculated as the probability of getting a test statistic as extreme as, or more extreme than (in favor of H_1), the observed test statistic t^*. Since all of the $\binom{N}{n}$ randomizations are equally likely under H_0, the p value is

$$p = P(T \le t^* | H_0) = \frac{\sum_{i=1}^{\binom{N}{n}} I(t_i \le t^*)}{\binom{N}{n}},$$

where t_i is the value of the test statistic $T = \bar{Y} - \bar{X}$ for the ith randomization and $I(\cdot)$ is the indicator function. In Table 1, the observed test statistic is $t^* = -9$, so the p value is $p = P(\bar{Y} - \bar{X} \le -9) = 3/35 \approx 0.0857$. This is represented in Figure 2 by the darker portion of the distribution. There is only moderate evidence of a treatment effect.

It is clear from the discreteness of the randomization distribution that the p value must be a multiple of $1/\binom{N}{n}$, although not every multiple is possible. In our example, we can see from Figure 1 that the achievable p values are $k/35$, where $k = 1, 2, 3, 4, 5, 6, 7, 9, 10, 12, 14, 16, 19, 20, 22, 23, 24, 26, 27, 28, 30, 31, 33, 34, 35$. If we choose a significance level of $\alpha = k/\binom{N}{n}$ that is one of the achievable p values, then

$$\begin{aligned} P(\text{Type I error}) &= P(p \le \alpha | H_0) \\ &= P\left(\sum_{i=1}^{\binom{N}{n}} I(t_i \le t^*) \le k \,\middle|\, H_0 \right) \\ &= \frac{k}{\binom{N}{n}} = \alpha. \end{aligned}$$

That is, the randomization test of H_0 is an exact test. If α is not chosen as one of the achievable p values, but $k/\binom{N}{n}$ is the largest achievable p value less than α, then $P(\text{Type I error}) = k/\binom{N}{n} < \alpha$ and the randomization test is conservative. Either way, the test is guaranteed to control the probability of a Type I error under very minimal conditions: randomization of the subjects to treatments.

3.3 Other Test Statistics

We could, of course, choose a test statistic other than $T = \bar{Y} - \bar{X}$ to measure the effectiveness of the new treatment, such as the difference in group medians or trimmed means. The randomization distribution of the difference in group medians is shown in Figure 2 [with the observed value marked with a solid triangle (▲)] and results in the same p value of $3/35$.

We could also use the sum of the responses in one of the treatment groups as a test statistic. Small values of

the new treatment sum or large values of the standard treatment sum would indicate improvement by the new treatment. These are shown in Table 1. By rewriting $\bar{Y} - \bar{X}$ as

$$\bar{Y} - \bar{X} = \frac{m+n}{mn} \sum_{j=1}^{n} Y_j - \frac{1}{m} \left(\sum_{i=1}^{m} X_i + \sum_{j=1}^{n} Y_j \right)$$

and noting that $\sum_{i=1}^{m} X_i + \sum_{j=1}^{n} Y_j$ is fixed over all $\binom{N}{n}$ randomizations, it can be seen that $\bar{Y} - \bar{X}$ is a monotone function of $\sum_{j=1}^{n} Y_j$ and the ordering of the $\binom{N}{n}$ test statistics is preserved. Therefore, $\sum_{j=1}^{n} Y_j$ is an equivalent test statistic to $\bar{Y} - \bar{X}$. Similarly, $\sum_{i=1}^{m} X_i$ is also an equivalent test statistic. It can also be shown that the two-sample pooled t statistic is a monotone function of $\bar{Y} - \bar{X}$ and, therefore, is another equivalent test statistic.

Because of their equivalence, the sum of the responses from one treatment group is often used rather than the t statistic or the difference in means since it is computationally more efficient. This becomes important since the number of randomizations $\binom{N}{n}$ quickly becomes large as m and n increase.

Another test statistic that could be used is the sum of the ranks of the responses in one group after the responses from both groups are ranked from 1 to N. Of course, this is just the Wilcoxon rank-sum test. This illustrates how rank tests are just one type of permutation method.

3.4 Estimation of the Treatment Effect

If we are willing to assume that the new treatment has a constant additive effect above and beyond the standard treatment, we can estimate this treatment effect by inverting the randomization test. Suppose that the new treatment increases the recovery times by Δ. Then if we shift the responses in the new treatment group by Δ, these shifted responses, $Y_1 - \Delta$, $Y_2 - \Delta, \ldots, Y_n - \Delta$, should be similar in magnitude to the standard treatment responses, $X_1, X_2, \ldots, X_m$, and the randomization test on these two sets of responses should not reject H_0.

An interval estimate of Δ can be constructed by considering all values of Δ for which the randomization test does not reject H_0. For our one-sided alternative hypothesis, this will be a one-sided confidence interval. Let T_Δ be the test statistic calculated using the new treatment responses that are shifted by Δ and let t_Δ^* be its observed value. Furthermore, let $p_1(\Delta) = P(T_\Delta \le t_\Delta^*)$ and $p_2(\Delta) = P(T_\Delta \ge t_\Delta^*)$ be the areas in the left and right tails of the randomization distribution of T_Δ, respectively, and let $\Delta_L = \min_{p_2(\Delta)>\alpha} \Delta$ and $\Delta_U = \max_{p_1(\Delta)>\alpha} \Delta$. In other words, Δ_L and Δ_U are the amounts of shift that would make the right and left tail p values for T_Δ approximately equal to α. Then $(-\infty, \Delta_U)$ and (Δ_L, ∞) are each one-sided $(1-\alpha)100\%$ *randomization confidence intervals* for Δ, and (Δ_L, Δ_U) is a two-sided $(1-2\alpha)100\%$ randomization confidence interval for Δ.

If α is chosen as one of the achievable p values in the randomization distribution of T, then these confidence intervals are exact. Otherwise, they are conservative intervals where the true coverage probability is at least as large as the chosen confidence coefficient.

Finding the endpoints of these intervals involves a tedious and laborious search that requires the recomputation of the randomization distribution of T_Δ for each value of Δ that is tried. Garthwaite (1996) described an efficient method for constructing confidence intervals from randomization tests, but this method is not implemented in any commercial software.

In our example, a $(1 - 2/35)100\% \approx 94.29\%$ one-sided confidence interval for the effect of the new treatment that is consistent with our one-sided alternative hypothesis is $(-\infty, 2)$. In other words, we are about 94.29% confident that the new treatment either decreases recovery time or increases it by at most 2 days. Notice that this interval includes zero and therefore agrees with the results of the randomization test where the p value was $p = 3/35 > \alpha$.

4. THE POPULATION MODEL

Inferences in the population model are based on random samples from populations. Random sampling is generally more difficult to accomplish than random assignment of subjects, but the advantage is that the conclusions can be generalized to the populations.

4.1 Permutation Tests and Intervals

Consider the situation in which we have two independent random samples from two populations. Let $X_1, X_2, \ldots, X_m$ be a random sample from a population with c.d.f. F and let $Y_1, Y_2, \ldots, Y_n$ be a random sample from a second population with c.d.f. G. We wish to test the hypotheses

$$H_0 : F = G,$$
$$H_1 : F \neq G.$$

The mechanics of constructing a *permutation test* for the population model are identical to those of the ran-

domization test in Section 3, although the reasoning behind it, first discussed by Pitman (1937a), is different. If H_0 is true, then all $N = m + n$ of the random variables could have reasonably come from either of the identical populations, and so the sample we obtained could have been any one of the $\binom{N}{n}$ possible divisions of these random variables into two groups of sizes m and n. Conditional on the observed values of the random variables, each of the $\binom{N}{n}$ divisions of these random variables is equally likely. We calculate an appropriate test statistic T for each of these rearrangements of the observed values to obtain the *permutation distribution* of T, which is used to calculate a p value for the test.

The permutation test in the population model is a conditional test since it generates the permutation distribution conditional on the observed values of the random variables (unlike the randomization model where the observed values were not random, only their treatment assignments). The test is also conditionally distribution-free since, conditional on the observed data, the permutation distribution of T does not depend on the population distributions F and G. Finally, the permutation test is conditionally exact for the same reasons that the randomization test is exact, but it is also unconditionally exact since the probability of a Type I error is controlled for all possible samples from F and G.

As an example, consider the data in Table 2, which displays measurements on the wing length and antennae length of two species of small flies that belong to an insect family termed "biting midges" that were discovered in the jungles of Central and South America. These very similar species were discovered by biologists Grogan and Wirth (1981), who dubbed them *A. fasciata* (Af) and *A. pseudofasciata* (Apf). The biologists took detailed measurements of important morphological features of the midges in an attempt to distinguish between the two species.

TABLE 2
Wing length (*WL*) *and antennae length* (*AL*), *in millimeters, of* 9 *Amerohelea fasciata* (*Af*) *midges and* 6 *Amerohelea pseudofasciata* (*Apf*) *midges*

Af		Apf	
WL	**AL**	**WL**	**AL**
1.72	1.24	1.78	1.14
1.64	1.38	1.86	1.20
1.74	1.36	1.96	1.30
1.70	1.40	2.00	1.26
1.82	1.38	2.00	1.28
1.82	1.48	1.96	1.18
1.90	1.38		
1.82	1.54		
2.08	1.56		

If we consider the data in Table 2 as random samples of sizes 9 and 6 from the two populations of midges, then constructing the permutation distribution for either measurement will consist of calculating the test statistic (T = sum of the measurements of the Af midges) for all $\binom{15}{6} = 5005$ possible divisions of the measurements into two samples of sizes 9 and 6. Figures 3 and 4 show the permutation distributions for the wing and antennae measurements, respectively, with the observed value of the test statistic, t^*, marked with a solid triangle (▲).

If we are using a two-sided alternative, the question arises about how to calculate the p value. These distributions are not symmetric and so there is no justification for doubling the probability in one tail. If we let

$$\bar{t} = \frac{1}{\binom{N}{n}} \sum_{i=1}^{\binom{N}{n}} t_i$$

be the mean of the permutation distribution, then we

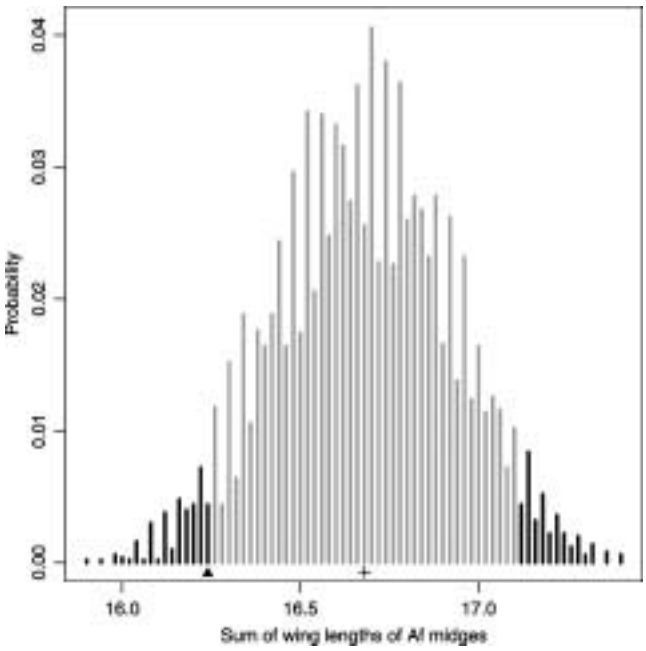

FIG. 3. *The permutation distribution of the sum of wing lengths of the Af midges in Table* 2.

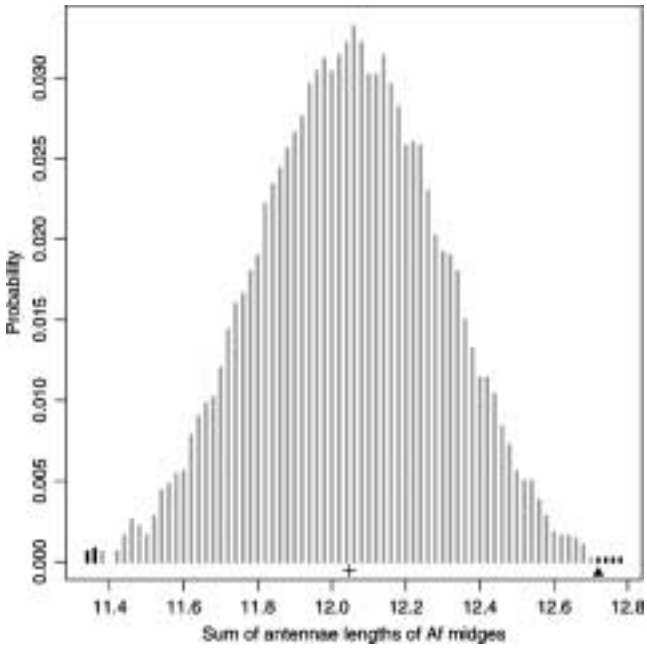

FIG. 4. *The permutation distribution of the sum of antennae lengths of the Af midges in Table* 2.

can define the two-sided p value as

$$p = P\big(|T - \bar{t}| \geq |t^* - \bar{t}| \big| H_0\big)$$
$$= \frac{1}{\binom{N}{n}} \sum_{i=1}^{\binom{N}{n}} I\left(|t_i - \bar{t}| \geq |t^* - \bar{t}|\right).$$

This is the probability of getting a test statistic as far, or farther, from the mean of the permutation distribution as the observed one. In Figures 3 and 4, $\bar{t}$ is marked with a plus $(+)$ and the darker portions of the distribution represent the p value. For the wing measurements, the p value is 0.0719, and for the antennae measurements, it is 0.0022. These species differ significantly in mean antennae length and marginally in mean wing length.

If we can assume that the distributions F and G differ by a shift in location, that is, $G(x) = F(x - \Delta)$, then we can estimate Δ with a *permutation confidence interval* by inverting the permutation test in the same manner as in Section 3.4. A 94.90% confidence interval for the mean difference (Af $-$ Apf) in wing length is $(-0.250, 0.010)$ and a 94.94% confidence interval for the mean difference in antennae length is $(0.087, 0.286)$. Notice that even with these rather small sample sizes, the discreteness of the permutation distribution has diminished enough to get achievable values of the confidence coefficient that are quite close to rather arbitrary values such as 95%.

4.2 Monte Carlo Sampling

Computation of the permutation distribution of a test statistic involves careful enumeration of all $\binom{N}{n}$ divisions of the observations. This poses two computational challenges. First, the sheer number of calculations required becomes very large as the samples become only moderate in size. There are over 155 million ways to divide 30 observations into two groups of size 15, and over 5.5 trillion ways to divide them into three groups of size 10. Second, enumerating each unique division of the data is not easily programmed and requires specialized software. The distributions in Figures 3 and 4 were produced with StatXact (Cytel Software Corporation, 2003), which uses efficient algorithms to calculate the permutation distribution of a variety of test statistics in many situations.

One easy and very practical solution to both these problems is to use Monte Carlo sampling from the permutation distribution to estimate the exact p value. Since the p value is simply the proportion of test statistic as extreme or more extreme than the observed value, we can naturally estimate this by randomly choosing test statistics from the permutation distribution and calculating the sample proportion that are as extreme or more extreme than the observed value. This is easily accomplished by repeatedly and randomly dividing the N observations into groups of size m and n and calculating the test statistic. A few thousand test statistics from the permutation distribution usually are sufficient to get an accurate estimate of the exact p value and sampling can be done with or without replacement (although with replacement is much easier). If M test statistics, t_i, $i = 1, \ldots, M$, are randomly sampled from the permutation distribution, a one-sided Monte Carlo p value for a test that rejects for large values of t is

$$\hat{p} = \frac{1 + \sum_{i=1}^{M} I\left(t_i \geq t^*\right)}{M + 1}.$$

Including the observed value t^*, there are a total of $M + 1$ test statistics. Since the observed value will always be "as extreme" as itself, the Monte Carlo p value will be no smaller than $1/(M + 1)$. This is consistent with the exact p value, which must be at least $1/\binom{N}{n}$. The idea of sampling from the permutation distribution was first proposed by Dwass (1957). The test remains exact and conditionally distribution-free, the only penalty being a small loss of efficiency. Jöckel (1986) showed that this loss of efficiency decreases as M increases and he gave a lower bound for the loss of efficiency as a function of M.

To illustrate, consider the data in Table 2 again. If we consider the two measurements on each midge, wing length and antennae length, as a bivariate observation, then we have two samples of vectors, $\mathbf{X}_1, \ldots, \mathbf{X}_9$, and $\mathbf{Y}_1, \ldots, \mathbf{Y}_6$. We could test whether the mean vectors of the two groups differ significantly, rather than doing individual tests for wing length and antennae length. One measure of the difference in mean vectors is the two-sample Hotelling T^2 statistic

$$T^2 = (\bar{\mathbf{X}} - \bar{\mathbf{Y}})' \left[\frac{\mathbf{S_X}}{9} + \frac{\mathbf{S_Y}}{6} \right]^{-1} (\bar{\mathbf{X}} - \bar{\mathbf{Y}}),$$

where $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$ are the sample mean vectors and $\mathbf{S_X}$ and $\mathbf{S_Y}$ are the sample covariance matrices. StatXact does not currently calculate T^2, so we will use Monte Carlo sampling. This can be programmed easily in many software packages and was done here in R (Ihaka and Gentleman, 1996). We repeatedly divide the 15 bivariate vectors randomly into groups of 9 and 6 and calculate T^2. Figure 5 shows a histogram of a random sample of $M = 999$ values of T^2 from its permutation distribution. The observed value of $T^2 = 72.4$ is larger than any of these $M = 999$ values, so the Monte Carlo p value is $\hat{p} = 1/(999 + 1) = 0.001$.

Of course $\hat{p}$ will vary depending on the Monte Carlo sample. It is simply the proportion of successes in $M + 1$ independent trials where the true probability of success is p. Therefore, $\hat{p}$ is an unbiased estimator of the exact p value and we can use it to construct a confidence interval for p, either with a large sample approximation or directly from the binomial distribution. A 99.9% confidence interval for the exact p value for Hotelling's T^2 based on the Monte Carlo results above is $(0.0000005, 0.0099538)$, which was calculated directly from the binomial distribution. Clearly, there is a significant difference in the mean vectors of wing and antennae measurements for the Af and Apf midges. If the confidence interval for the exact p value had been more ambiguous and included our significance level, we could simply take a larger Monte Carlo sample to get a more accurate estimate of the exact p value. For all practical purposes, we can get as accurate an estimate of the exact p value as desired.

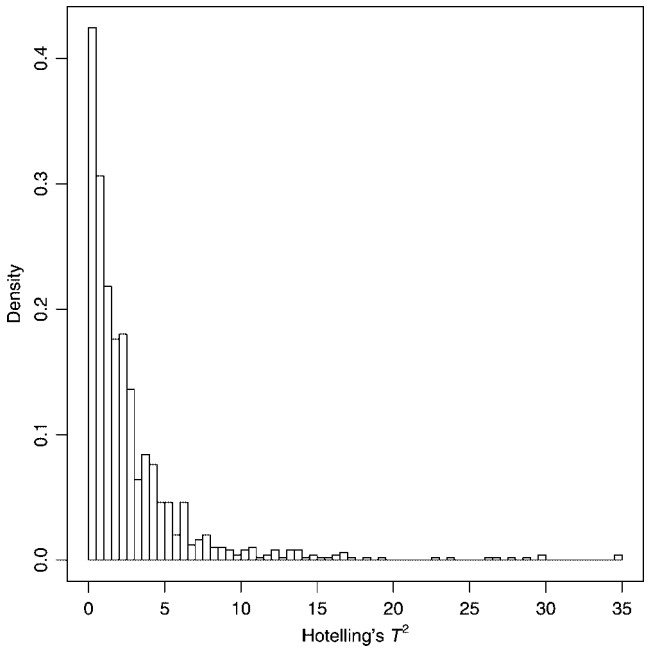

FIG. 5. *A histogram of a random sample of $M = 999$ test statistics from the permutation distribution of Hotelling's T^2 for the data in Table 2.*

5. MORE COMPLICATED DESIGNS

We have described in detail the rationale and mechanics involved in permutation methods, in particular for the two group problem. While the rationale differs for the randomization and population models, the mechanics are generally the same. Permutation methods are very flexible and may be applied in many other situations simply by identifying rearrangements of the data that are equally likely under the null hypothesis (or unequally likely, but with known probabilities). The test statistic can also be chosen to suit the situation, depending on what kind of alternative is of interest. We leave it to the reader to investigate how permutation methods might be applied in other situations, but we note a few references that consider some common situations: paired data (Fisher, 1935, Section 21; Ernst and Schucany, 1999), the correlation coefficient (Pitman, 1937b), simple linear regression (Manly, 1997, Section 8.1), multiple linear regression (Kennedy and Cade, 1996) and randomized complete block designs (Pitman, 1938).

5.1 The One-Way Layout

We conclude by considering the one-way layout with k independent groups and we discuss, in particular, an approach to multiple comparisons that controls the Type I error rate. Suppose that N subjects are randomized to k treatment groups of size n_i, $i = 1, 2, \ldots, k$, and their responses Y_{ij}, $j = 1, 2, \ldots, n_i$, are recorded. We wish to test the hypotheses

H_0: There is no difference between the treatments,

H_1: At least one treatment differs from the others.

TABLE 3
Reading speeds for 14 *subjects randomly assigned to three typeface styles*

	Typeface style		
	1	**2**	**3**
	135	175	105
	91	130	147
	111	514	159
	87	283	107
	122		194
$\bar{Y}_i$	109.2	275.5	142.4

There are $\binom{N}{n_1\, n_2 \cdots n_k}$ possible randomizations of the N subjects to the k groups and these are equally likely under H_0. For each of these randomizations we calculate a test statistic, like the analysis of variance F statistic, or equivalently $T = \sum_{i=1}^{k} n_i \bar{Y}_i^2$, where $\bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i$ is the mean of the ith group. Large values of T favor the alternative hypothesis, so the p value is the proportion of test statistics greater than or equal to the observed value.

To illustrate, we consider an example from Bradley (1968), who investigated whether reading speed is affected by the typeface style of the text. Fifteen subjects were randomly assigned to one of three different typeface styles and their reading speeds are recorded in Table 3 along with the group means. One subject was unable to complete the exercise for reasons unrelated to the experiment, so one group has four subjects. The data are displayed in Figure 6. There are $\binom{14}{5\,4\,5} = 252{,}252$ test statistics in the randomization distribution of T. Figure 7 shows a histogram of a random sample of $M = 9999$ test statistics from this distribution with the observed value $t^* = 464{,}613$ marked with a solid triangle (▲). With 114 test statistics greater than t^*, the Monte Carlo p value is $\hat{p} = (1 + 114)/(9999 + 1) = 0.0115$. A 99.9% confidence interval for the exact p value is $(0.0083, 0.0154)$.

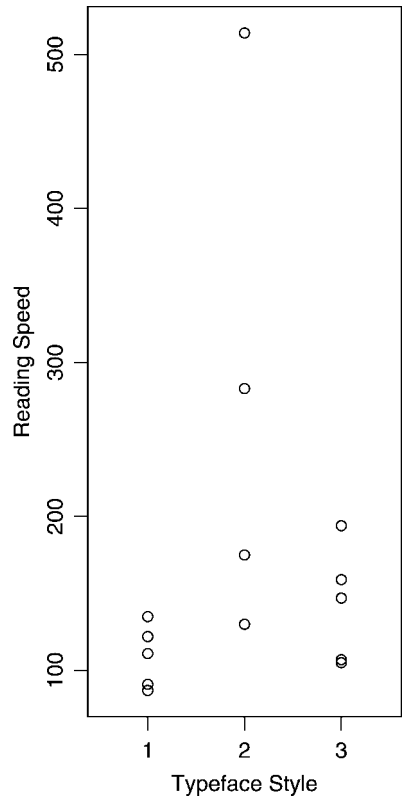

FIG. 6. *Reading speeds for the* 14 *subjects in Table* 3.

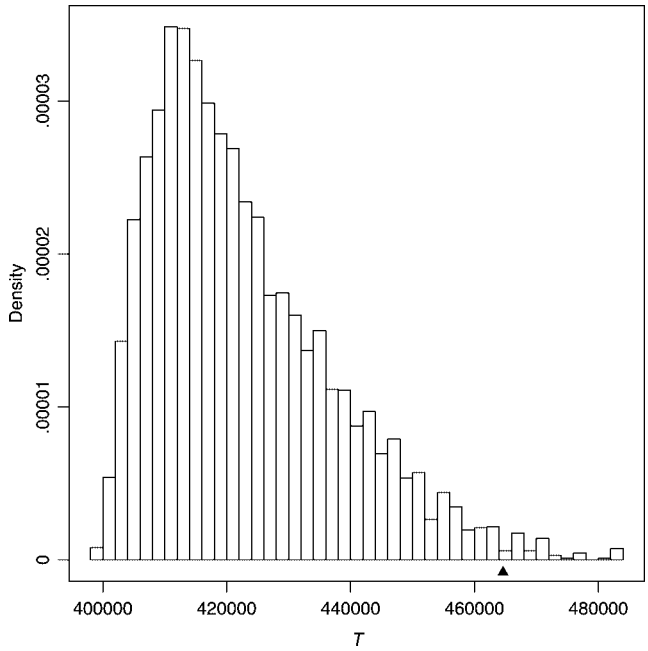

FIG. 7. *A histogram of a random sample of* $M = 9999$ *test statistics from the permutation distribution of* $T = \sum_{i=1}^{k} n_i \bar{Y}_i^2$ *for the data in Table* 3.

5.2 Multiple Comparisons

Clearly, there is a difference between the treatments, so we would like to know which treatments are significantly different while controlling the probability of making a Type I error. There are a total of $\binom{k}{2}$ treatment comparisons that can be made. Two treatments will be called significantly different if their means differ by more than the critical value C_α. We have committed a Type I error if we find any one of the comparisons to be significant when H_0 is true. To control this error, we

need to find C_α such that

$$P\left(\max_{1\le i<j\le k}|\bar{Y}_i-\bar{Y}_j|\ge C_\alpha\Big|H_0\right)=\alpha.$$

Under H_0, all $\binom{N}{n_1\,n_2\cdots n_k}$ possible randomizations are equally likely and so this probability is simply the proportion of the randomizations for which at least one difference in treatment means (the largest one) is considered significant. By choosing C_α so that this probability is α, we control the probability of a Type I error and obtain an exact multiple comparison procedure. We determine C_α by calculating the largest difference in treatment means for each randomization and finding the $1-\alpha$ quantile of this distribution.

In our example, there are three mean comparisons: $|\bar{Y}_1-\bar{Y}_2|=166.3$, $|\bar{Y}_1-\bar{Y}_3|=33.2$ and $|\bar{Y}_2-\bar{Y}_3|=133.1$. A Monte Carlo estimate of C_α can be obtained by calculating the largest difference in treatment means from a random sample of randomizations. Based on $M=9999$ randomizations, our estimate for $\alpha=0.05$ is $\hat{C}_{0.05}=142.5$, which indicates that treatments 1 and 2 are the only treatments that are significantly different.

This procedure can easily be modified if only certain comparisons are of interest or for one-sided comparisons. The key is to find C_α so that the probability under H_0 of finding a significant difference is α.

ACKNOWLEDGMENT

I thank Cliff Lunneborg for providing the R code for calculating the permutation/randomization confidence intervals.

REFERENCES

BRADLEY, J. V. (1968). *Distribution-Free Statistical Tests*. Prentice-Hall, Englewood Cliffs, NJ.

CYTEL SOFTWARE CORPORATION (2003). *StatXact 6*. Cytel Software Corporation, Cambridge, MA.

DWASS, M. (1957). Modified randomization tests for nonparametric hypotheses. *Ann. Math. Statist.* **28** 181–187.

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

ERNST, M. D. and SCHUCANY, W. R. (1999). A class of permutation tests of bivariate interchangeability. *J. Amer. Statist. Assoc.* **94** 273–284.

FISHER, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.

FISHER, R. A. (1936). "The coefficient of racial likeness" and the future of craniometry. *J. Royal Anthropological Institute of Great Britain and Ireland* **66** 57–63.

GARTHWAITE, P. H. (1996). Confidence intervals from randomization tests. *Biometrics* **52** 1387–1393.

GROGAN, W. L., JR. and WIRTH, W. W. (1981). A new American genus of predaceous midges related to *Palpomyia* and *Bezzia* (Diptera: Ceratopogonidae). *Proc. Biological Society of Washington* **94** 1279–1305.

HIGGINS, J. J. (2004). *An Introduction to Modern Nonparametric Statistics*. Brooks/Cole, Pacific Grove, CA.

IHAKA, R. and GENTLEMAN, R. (1996). R: A language for data analysis and graphics. *J. Comput. Graph. Statist.* **5** 299–314.

JÖCKEL, K.-H. (1986). Finite sample properties and asymptotic efficiency of Monte Carlo tests. *Ann. Statist.* **14** 336–347.

KENNEDY, P. E. and CADE, B. S. (1996). Randomization tests for multiple regression. *Comm. Statist. Simulation Comput.* **25** 923–936.

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

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

PITMAN, E. J. G. (1937a). Significance tests which may be applied to samples from any populations. *J. Roy. Statist. Soc. Suppl.* **4** 119–130.

PITMAN, E. J. G. (1937b). Significance tests which may be applied to samples from any populations. II. The correlation coefficient test. *J. Roy. Statist. Soc. Suppl.* **4** 225–232.

PITMAN, E. J. G. (1938). Significance tests which may be applied to samples from any populations. III. The analysis of variance test. *Biometrika* **29** 322–335.