Chapter 12
Permutation and Rank Tests

12.1 Introduction

In the early 1930s R. A. Fisher discovered a very general exact method of testing hypotheses based on permuting the data in ways that do not change its distribution under the null hypothesis. This *permutation* method does not require standard parametric assumptions such as normality of the data. It does require, however, certain invariance properties under the null hypothesis that restricts application to fairly simple designs. But in such situations, the method results in exact tests with level α under very weak distributional assumptions. Moreover, the method is *statistic-inclusive* in the sense that any test statistic can be used and inherits the level-α property, although some statistics are much more powerful than others.

Tests based on this method are called *permutation tests* or *randomization tests* depending on whether the data can be viewed as samples from populations or not. That is, when sampling from populations, "permutation tests" refer to use of the permutation method to obtain level α tests under weak distributional assumptions. In Fisher's words (1935, Sec. 21), these are tests of a "wider" null hypothesis (as compared to assuming normal distributions, for example).

However, experiments may be performed on units that cannot be viewed as arising from random sampling of any population. In such situations "randomization inference" refers to inference drawn based only on the physical randomization of the units to different treatments, and on the test statistic calculated at all possible randomizations of the data. The same test that we called a permutation test in random sampling contexts is now called a randomization test. Of course one needs to qualify all statements of significance about such experiments with the disclaimer that randomization inference only applies to the units used in the experiment.

Permutation tests are the foundation of classical nonparametric statistics (also called *distribution-free* statistics), which itself is often identified with rank tests. Rank tests are actually a special subclass of permutation tests with three distinct advantages:

D.D. Boos and L.A. Stefanski, *Essential Statistical Inference: Theory and Methods*,
Springer Texts in Statistics, DOI 10.1007/978-1-4614-4818-1_12,

1. For data without ties, the conditional permutation distribution of a rank test is actually unconditional (does not change from sample to sample) because the ranks of a continuous data set are the same for every sample. Thus, the distribution of an important rank statistic like the Wilcoxon Rank Sum statistic can be tabulated or programmed. However, this computing advantage is less important today, and when there are ties in the data (a very common occurrence), the tabulated values are not appropriate, and the conditional permutation distribution is required for exact inference.
2. The key philosophical foundation of rank tests arises from the theory of invariant tests as described in Lehmann (1986, Ch. 5). The idea with invariant tests is to reduce the class of tests considered to those that are naturally invariant with respect to a group of transformations G on the sample space of the data. Given G, a maximal invariant is a statistic $M(\boldsymbol{x})$ with the property that any invariant test with respect to G must be a function of $\boldsymbol{x}$ only through $M(\boldsymbol{x})$. Now consider the two-sample problem with $H_0 : F_{\mathrm{X}}(x) = F_{\mathrm{Y}}(x)$ versus the alternative "F_{Y} is stochastically larger than F_{X}," that is, $H_a : 1 - F_{\mathrm{Y}}(x) \geq 1 - F_{\mathrm{X}}(x)$ for all x with strict inequality for at least one x. This alternative is more general than the usual shift alternative, $F_{\mathrm{Y}}(x) = F_{\mathrm{X}}(x - \Delta)$, but it certainly includes the shift alternative as a special case. Let G be the group of transformations such that each $g \in G$ is continuous and strictly increasing. For this testing problem and group G, the set of ranks of the combined X and Y samples is the maximal invariant statistic. Thus, any invariant test must be a function of the ranks. Does it make sense to require tests to be invariant with respect to monotone transformations? Whenever data are ordinal or we do not trust the measurement scale, then invariance certainly makes sense, and rank tests are the obvious choice.
3. Rank tests may be preferred in many situations because of their Type II error robustness. That is, for an appropriate data generation model, the permutation method can make any statistic Type I error robust (level α), but because rank tests are a function of the data only through the ranks, the influence of outliers is automatically limited. Thus, rank tests are power robust in outlier-prone situation. The key example is the Wilcoxon Rank Sum test that is powerful in the face of a wide variety of distributional shapes. In fact, Hodges and Lehmann (1956) showed that the asymptotic relative efficiency (ARE) of the Wilcoxon Rank Sum test to the t test satisfies the following:

   a) ARE= .955 for normal shift alternatives, and thus the Wilcoxon Rank Sum test loses little in comparison to the t where the t is best;
   b) and ARE $\geq$ .864 for any continuous unimodal shift alternative with finite variance, and thus the Wilcoxon Rank Sum test can never be much worse than the t-test but possibly much better.

   Optimality for permutation and rank procedures is discussed in more detail later.

Although the term "nonparametric" was classically associated with permutation and rank procedures, in recent times it is more commonly used for nonparametric density and regression estimation methods based on smoothing. Thus, when describing rank or permutation procedures, it is best to use the specific names “rank” or “permutation” rather than “nonparametric.” Although permutation tests are inherently defined in terms of randomization, they overlap with a variety of conditional procedures and uniformly most powerful unbiased (UMPU) “Neyman structure similar” tests based on exponential family theory (the most well known is Fisher’s Exact Test).

Permutation procedures are very computationally intensive. These extensive computations prevented widespread use of the method until the 1990’s. Thus, asymptotic approximations were dominant until the 1990’s, although exact small-sample distributions were tabled for a number of important rank test statistics.

The asymptotic approximations are basically of three kinds: normal approximations based on the Central Limit Theorem, F or *beta* approximations based on matching permutation moments with normal theory moments, and Edgeworth expansions that improve on the normal approximations. The normal approximations have been used the most due to their simplicity. However, the F approximations initiated by Pitman (1937a,b) and Welch (1937) in the 1930s and updated by Box and Andersen (1955) are generally better for situations where they apply. The Edgeworth approximations are very good for the Wilcoxon Rank Sum and Wilcoxon Signed Rank statistics, but are somewhat more complicated for other statistics and seem not to be in general usage. Thus, we emphasize the F approximations rather than the normal or Edgeworth approximations. In fact these F approximations appear to be underused in general, but the work of Conover and Iman (1981) may have rekindled their use. Asymptotic normal theory remains important for comparing different methods according to asymptotic power, rather than for finding critical values. We give an overview of these results and then a few technical details in an appendix. There are excellent texts such as Hajek and Sidak (1967) and Randles and Wolfe (1979) that carefully explain asymptotic normality proof techniques for rank statistics. We add that most nonparametric texts of the last forty years are mainly about rank statistics, although Lehmann (1975) and Pratt and Gibbons (1981) have portions devoted to permutation tests. Puri and Sen (1971) emphasize the theory of permutation tests in multivariate settings.

In our current situation of extensive computing power, Monte Carlo approximations are the most important alternative to exact calculations. By Monte Carlo approximation we mean random sampling from the set of all permutations. This method can be used for any statistic in a situation where permutation methods are appropriate. Moreover, the error of approximation can be reduced by just adding more replications. This sampling (or resampling) in the “permutation world” is very similar to sampling in the bootstrap world; the main difference is that bootstrap p-values are typically approximate, even using the limit as the number of resamples B goes to ∞. In contrast, the limiting p-value in the permutation world is exact, and even the finite B estimated p-value has an exact interpretation.

Thus, our treatment of nonparametric methods is quite a bit different from most texts written in the last half of the twentieth century, which have emphasized rank tests and asymptotic normal approximations. We believe the basic permutation approach is the most important idea because it provides Type I error robustness for any statistic. Monte Carlo approximations can handle any problem for which the exact permutation distribution is too difficult to compute. Rank methods are still very important, but now because they provide Type II error robustness (good power in the face of outliers), not because they are easy to use or their distributions are tabled.

We start first with the two-sample problem to illustrate the basic permutation test approach. We then give some general theory for permutation tests along with approximations and discuss optimality results. Then we review results for the most important designs admitting permutation tests, their use in contingency tables, and estimators and confidence procedures derived from inverting permutation and rank tests.

12.2 A Simple Example: The Two-Sample Location Problem

We illustrate here the basic permutation approach with a simple two treatment experiment.

A clever middle school student believes that she has discovered a new method for teaching fractions to third graders. To test her hypothesis, she selects six students from her father's third grade class and randomly assigns four to learn the new method and two to use the standard method. After training both groups, they are given twenty test problems. The scores for the standard method group are $x_1 = 6$, $x_2 = 8$ and for the new method group are $y_1 = 7$, $y_2 = 18$, $y_3 = 11$, $y_4 = 9$. The results look promising for the new method, but how shall we assess statistical significance?

One possible test statistic is the standard two-sample t,

$$t(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\overline{Y} - \overline{X}}{\sqrt{s_p^2 \left(\frac{1}{m} + \frac{1}{n}\right)}}, \tag{12.1}$$

where $s_p^2 = \{\sum(X_i - \overline{X})^2 + \sum(Y_j - \overline{Y})^2\}/(m + n - 2)$. If t is large, then one might be convinced that the new method is better than the standard one.

Another commonly used statistic is W = the sum of the ranks of the Y values when both X and Y samples are thrown together and ranked from smallest to largest. Let $\boldsymbol{Z}$ denote the joint sample of both $\boldsymbol{X}$ and $\boldsymbol{Y}$ together: $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{Y})$ with observed values here $(6, 8, 7, 18, 11, 9)$. The ranks of these observed values are then $(1, 3, 2, 6, 5, 4)$ and $W = 2 + 6 + 5 + 4 = 17$, the sum of the Y ranks. If the new teaching method is better, then on average we would expect W to be large. Assuming that either t or W are reasonable statistics for our testing problem, we still need to agree on what is a proper reference distribution for each. A simple but very general approach is to recognize that there were actually $\binom{6}{2} = 15$ different ways that two students could have been selected from the original six to go in the X sample (with the remaining four assigned to the Y sample). Table 12.1 is a listing of the possible samples and the values of t and W for both.

Table 12.1 All Possible Permutations for Example Data

	X Sample		Y Sample				$\sum Y_i$	t	W
1.	6	8	7	18	11	9	45	1.17	17
2.	7	8	6	18	11	9	44	0.91	16
3.	18	8	7	6	11	9	33	−1.36	12
4.	11	8	7	18	6	9	40	0.12	13
5.	9	8	7	18	11	6	42	0.49	14
6.	6	7	8	18	11	9	46	1.47	18
7.	6	18	7	8	11	9	35	−0.84	14
8.	6	11	7	18	8	9	42	0.49	15
9.	6	9	7	18	11	8	44	0.91	16
10.	7	18	6	8	11	9	34	−1.08	13
11.	18	11	7	6	8	9	30	−2.98	10
12.	11	9	7	18	6	8	39	−0.06	12
13.	7	11	6	18	8	9	41	0.30	14
14.	7	9	6	18	11	8	43	0.69	15
15.	18	9	7	6	11	8	32	−1.72	11

Table 12.2 Permutation Distribution of t

t	−2.98	−1.72	−1.36	−1.08	−0.84	−0.06	0.12
$P(t)$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$
t	0.30	0.49	0.69	0.91	1.17	1.47	
$P(t)$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	

If the treatments produce identical results, then the outcomes for each student would have been exactly the same for any of the 15 possible randomizations. Thus, a suitable reference distribution for t or W is just the possible 15 values of t or W along with the probability 1/15 of each. This reference distribution for t, called the permutation distribution, is in Table 12.2.

Note that the permutation distribution of t is discrete even when sampling from a continuous distribution. (Here the distribution of the data is also discrete because the possible test scores are 0, 1, . . . , 20).

Using the distribution in Table 12.2, a conditional test for this experiment with $\alpha = 1/15$ would be to reject if $t \geq 1.47$. A one-sided p-value for the observed value of $t = 1.17$ is 2/15. Similarly a conditional $\alpha = 1/15$ level test based on the rank sum W would reject if $W \geq 18$, and the one-sided p-value is 2/15.

In general, the tests based on t and W would not give exactly the same results. For example, suppose the original data had been the 14th permutation, (7,9,6,18,11,8). Then the permutation p-value for t would be $5/15 = .33$, whereas the permutation

p-value for W would be $6/15 = .40$. Note, however, the column in Table 12.1 (p. 453) for the sum of the Y values. Comparing the $\sum Y_i$ and t values, one can see that the permutation p-values from $\sum Y_i$ and t are identical if the original data had been any of the 15 permutations. In such a case, we say that the two statistics are permutationally equivalent because they give exactly the same testing results.

In Problem 12.1 (p. 523) we ask for the permutation distribution of W from Table 12.1 (p. 453). A unique feature of rank statistics when there are no ties in the data is that the permutation distribution is the same for every such data set. That is, although the data values would change for every data set, as long as there are no ties in the 6 data points, the ranks would always be (1,2,3,4,5,6). Thus, the results for W in Table 12.1 (p. 453) would be exactly the same except in a different order, and therefore the distribution would be the same. This is one reason that rank statistics gained popularity: without ties, the exact distribution does not change and can then be tabled for easy lookup.

For simplicity we purposely started with a data set having no ties. However, ties occur frequently in real data even in continuous data settings due to rounding or inaccurate measurement. The standard way to rank data with ties is to assign the average rank to each of a set of tied values. For example, suppose our second X data point had been 7 instead of 8. Then the Z vector would have been (6,7,7,18,11,9), and instead of (1,3,2,6,5,4) for the ranks we would have (1,2.5,2.5,6,5,4). These are now called the *midranks*. We have taken the values 7 and 7 that would have occupied ranks 2 and 3 and replaced them by $(2 + 3)/2 = 2.5$. If the first X data point had also been a 7, then the midrank vector would have been (2,2,2,6,5,4), where we have used $(1 + 2 + 3)/3 = 2$ for the first three midranks. The use of midranks has no effect on the general permutation approach, but tabling distributions as mentioned in the previous paragraph is no longer possible since every configuration of tied values has a different permutation distribution.

12.3 The General Two-Sample Setting

The two-sample problem assumes that N experimental units (rats, for example) are available to compare two treatments A and B. First, m units are randomly assigned to receive treatment A, and the $n = N - m$ remaining units are assigned to receive treatment B. After the experiment is run, we obtain realizations of some measurement $X_1, \ldots, X_m$ for treatment A and $Y_1, \ldots, Y_n$ for treatment B. The null hypothesis H_0 is that both treatments are the same or have identical effects on the rats. In other words, if the third rat in group A whose measurement is X_3 had been assigned to group B instead, the X_3 would still have been the result under H_0 for that rat, but now it would have a Y label. In fact, we can think of all possible $\binom{N}{m}$ random assignments of m rats to group A and n rats to group B, and assume that under H_0 the individual results would be the same regardless of group assignment.

We might then formulate a test procedure as follows.

1. Randomly assign m units to A and n units to B.
2. Run the experiment to obtain $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$.
3. Think of the collection $\boldsymbol{Z} = (X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ as fixed and order the $M_N = \binom{N}{m}$ values of some statistic T calculated for each $\boldsymbol{Z}^*$ obtained by permuting $\boldsymbol{Z}$ to have different sets of m first coordinates. Call these ordered values $T_{(1)} \leq T_{(2)} \leq \ldots \leq T_{(M_N)}$, and let $T_0 = T(\boldsymbol{X}, \boldsymbol{Y})$ be the statistic calculated for the original data.
4. Reject H_0 if $T_0 > T_{(k)}$.

This test, conditional on $\boldsymbol{Z}$, has conditional α-level

$$1 - \frac{k}{M_N}$$

if $T_{(k)} < T_{(k+1)}$ (not tied) since $M_N - k$ values of T are larger than $T_{(k)}$. The exact conditional p-value is the proportion of values greater than or equal to T_0,

$$\frac{[\# T_{(i)} \geq T_0]}{M_N}. \tag{12.2}$$

When T is the t statistic in (12.1, p. 452), the above two-sample permutation procedure was proposed by Pitman (1937a). The credit for the permutation approach, however, goes to R. A. Fisher who had earlier introduced the permutation approach in the fifth edition of *Statistical methods for Research Workers* (2 × 2 table example) published in 1934 and in the first edition of *The Design of Experiments* (one-sample t example) in 1935.

Besides computational problems, the main drawback of the procedure described in points 1.–4. outlined above is that:

a) the results pertain to the N units obtained and not to a larger population;
b) computations of test power are difficult.

Thus, it is often useful to assume a population sampling model of the usual form

$$X_1, \ldots, X_m \quad \text{iid} \quad F_{\mathrm{X}}(x) = P(X_1 \leq x),$$

$$Y_1, \ldots, Y_n \quad \text{iid} \quad F_{\mathrm{Y}}(x) = P(Y_1 \leq x),$$

with $H_0 : F_{\mathrm{X}}(x) = F_{\mathrm{Y}}(x)$. Under this model we can show that the conditional permutation test actually has exact size α unconditionally, i.e.,

$$P(\text{rejection} \mid H_0) = \alpha.$$

The permutation approach has the advantage that no assumption regarding distributions of random variables is required. Moreover, one can often show using permutational Central Limit Theorems (e.g., Theorem 12.2, p. 465) that the conditional distribution of $T(\boldsymbol{X}, \boldsymbol{Y})$ properly standardized converges to a standard normal as $\min(m,n) \to \infty$. Thus, in large samples one can use normal critical values rather than list all M_N possible values of T. Alternatively, one can randomly sample B of the possible permutations and base a test on the ordered values of $T_1, \ldots, T_B$. First we give the general theory of permutation tests and then discuss these approximations as well as the Box-Andersen F approximation.

12.4 Theory of Permutation Tests

12.4.1 Size α Property of Permutation Tests

In this subsection we show that permutation tests used in random sampling contexts can have exact size α when randomizing on rejection region boundaries, and otherwise has level α when the test is carried out without such randomization. Recall that a size α test is one for which $\sup_{H_0} P(\text{reject} H_0) = \alpha$ and level α means $\sup_{H_0} P(\text{reject} H_0) \le \alpha$. The reference to *randomization* merely refers to flipping a biased coin for sample points on the boundary between the rejection and acceptance region in order to obtain size α and has nothing to do with the randomization used in the definition of a permutation test.

To prove size-α results rigorously, we need some additional notation. Two useful sources are Hoeffding (1952) and Puri and Sen (1971). Let $\boldsymbol{Z} = (Z_1, \ldots, Z_N)^T$ have joint distribution function $F_{\boldsymbol{Z}}(\boldsymbol{z})$ and sample space S. Let G be a group of M_N transformations of S onto S such that under H_0 the distribution of each $g_i(\boldsymbol{Z})$, $g_i \in G, i = 1, \ldots, M_N$, is exactly the same as the distribution of $\boldsymbol{Z}$. Two examples of such groups are as follows.

Permutations: G consists of all $N!$ permutations of $\boldsymbol{Z}$. If $\boldsymbol{Z}$ is exchangeable or iid, then $g_i(\boldsymbol{Z}) \overset{d}{=} \boldsymbol{Z}$. Although, in the two-sample problem (two independent samples), we usually consider only the $\binom{N}{m}$ partitions into two groups since the statistics used do not change by permuting elements within each sample. In the k-sample problem (k independent samples), we consider only the

$$\binom{N}{n_1 n_2 \ldots n_k} = \frac{N!}{n_1! \cdots n_k!}$$

partitions into k groups, where $n_1 + n_2 + \cdots + n_k = N$. The group of $N!$ permutations is relevant for the two-sample, k-sample, and correlation problems.

Sign Changes: G consists of all 2^N sign change transformations, $g_1(\boldsymbol{Z}) = (Z_1, Z_2, \ldots, Z_N), g_2(\boldsymbol{Z}) = (-Z_1, Z_2, \ldots, Z_N), g_3(\boldsymbol{Z}) = (Z_1, -Z_2, Z_3, \ldots, Z_N)$, etc. If the Z_i's are independently (but not necessarily identically) distributed, where each Z_i is symmetrically distributed about 0, then $g_i(\boldsymbol{Z}) \overset{d}{=} \boldsymbol{Z}$. The sign change group is relevant for the paired two-sample problem and the one-sample symmetry problem.

The following development is due to Hoeffding (1952). Because the permutation distribution is discrete, it is not possible to achieve arbitrarily chosen α-levels like $\alpha = .05$ without using a randomized testing procedure. This makes the details seem harder than they really are.

Let $T(z)$ be a real-valued function on S such that for each $z \in S$

$$T_{(1)}(z) \leq T_{(2)}(z) \leq \cdots T_{(M_N)}(z)$$

are the ordered values of $T(g_i(z)), i = 1, \ldots, M_N$. Given α, $0 < \alpha < 1$, let k be defined by

$$k = M_N - [M_N \alpha],$$

where $[\cdot]$ is the greatest integer function. Let $M_N^+(z)$ and $M_N^0(z)$ be the numbers of $T_{(j)}(z), j = 1, \ldots, M_N$, which are greater than $T_{(k)}(z)$ and equal to $T_{(k)}(z)$, respectively. Define

$$a(z) = \frac{M_N \alpha - M_N^+(z)}{M_N^0(z)}.$$

Then define the test function $\phi(z)$ by

$$\phi(z) = \begin{cases} 1, & \text{if } T(z) > T_{(k)}(z); \\ a(z), & \text{if } T(z) = T_{(k)}(z); \\ 0, & \text{if } T(z) < T_{(k)}(z). \end{cases}$$

Note that for a test function, $\phi(z) = 1$ means rejection of H_0, $\phi(z) = 0$ means acceptance of H_0, and $\phi(z) = \pi$ means to randomly reject H_0 with probability π. The test defined by ϕ is an exact conditional level α test by construction. The following theorem tells us that under $g_i(\boldsymbol{Z}) \overset{d}{=} \boldsymbol{Z}$ for each $g_i \in G$, the test is unconditionally a size-α test.

Theorem 12.1. *(Hoeffding). Let the data $\boldsymbol{Z} = (Z_1, \ldots, Z_N)$ and the group G of transformations be such that $g_i(\boldsymbol{Z}) \overset{d}{=} \boldsymbol{Z}$ for each $g_i \in G$ under H_0. Then the test defined above by $\phi(\boldsymbol{Z})$ has size α.*

Proof. First note that by the definition of $a(z)$ and ϕ, we have for each $z \in S$

$$\frac{1}{M_N} \sum_{i=1}^{M_N} \phi(g_i(z)) = \frac{M_N^+ + a(z) M_N^0(z)}{M_N} = \alpha.$$

Now since $g_i(\boldsymbol{Z}) \overset{d}{=} \boldsymbol{Z}$ and G is a group, $\mathrm{E}_{H_0}\phi(\boldsymbol{Z}) = \mathrm{E}_{H_0}\phi(g_i(\boldsymbol{Z}))$ for each i, and

$$\begin{aligned} P_{H_0}(\text{rejection}) = E_{H_0}\phi(\boldsymbol{Z}) &= \frac{1}{M_N} \sum_{i=1}^{M_N} \mathrm{E}_{H_0}\phi(g_i(\boldsymbol{Z})) \\ &= \mathrm{E}_{H_0}\left[\frac{1}{M_N} \sum_{i=1}^{M_N} \phi(g_i(\boldsymbol{Z}))\right] = \alpha. \end{aligned}$$

■

The above proof is deceptively simple. The key fact that makes it work is that $\mathrm{E}_{H_0}\phi(g_i(\boldsymbol{Z}))$ is the same for each g_i including $g(\boldsymbol{Z}) = \boldsymbol{Z}$. This fact rests on the identical distribution of $g_i(\boldsymbol{Z})$ for each i and on the group nature of G. The identical distribution requirement is intuitive, but why do we need G to be a group? Recall that the test procedure consists of computing T for each member of G and then rejecting if $T(\boldsymbol{Z})$ is larger than an order statistic of the $T(g_i(\boldsymbol{Z}))$ values. Now $\phi(g_i(\boldsymbol{Z}))$ is the test that computes $T(g_j(g_i(\boldsymbol{Z})))$, $j = 1, \ldots, M_N$, orders all of them, and rejects if $T(g_i(\boldsymbol{Z}))$ is larger than one of the ordered values. If G is not a group, then the set of ordered values will not be the same for each test $\phi(g_i(\boldsymbol{Z}))$ because $g_j(g_i)$ will not be in G for some i and j. Since the sets of ordered values could be different, there would be no basis for believing that a test based on $g_i(\boldsymbol{Z})$ would have the same expectation as that based on $\boldsymbol{Z}$.

Note also that the use of $a(z)$ in $\phi(z)$ is a way of randomizing to get an exact size-α test. In practice we might just define $\phi(z)$ to be one if $t(z) > t_{(k)}(z)$ and zero otherwise. The resulting unconditional level is a weighted average of the discrete levels less than or equal to α and will usually be less than α.

The conditional test procedure described in 1) $-$ 4) may be used for any test statistic, but the rejection region in Step 4) should be modified to correspond to the situation. For example, the alternative hypothesis might be that the mean of A is less than that of B. We would then look for small values of t. Or the test could be two-sided and we would reject if $t < t_{(k)}$ or if $t > t_{(m)}$.

12.4.2 Permutation Moments of Linear Statistics

The exact permutation distribution may be difficult to compute. For certain linear statistics, though, we can calculate the moments of the permutation distribution quite easily. These moments are then used in the various normal and F approximations found in later sections.

We consider general results for situations associated with the group of transformations consisting of all permutations. These situations include the two-sample and k-sample situations, and bivariate data $(X_1, Y_1), \ldots, (X_N, Y_N)$ where correlation and regression of Y on X are of interest. Let $\boldsymbol{a} = (a_1, \ldots, a_N)$ and $\boldsymbol{c} = (c_1, \ldots, c_N)$ be two vectors of real constants. We select a random permutation of the a values, call them $A_1, \ldots, A_N$, and form the statistic

$$T = \sum_{i=1}^{N} c_i A_i. \tag{12.3}$$

In applications $\boldsymbol{a}$ is actually the observed vector $\boldsymbol{Z}$ (or a function of $\boldsymbol{Z}$ such as the rank vector), and $\boldsymbol{c}$ is chosen for the particular problem at hand. For example, in the two-sample problem, with $\boldsymbol{a} = \boldsymbol{Z}$ and $c_i = 0$ for $i = 1, \ldots, m$ and 1 otherwise, the observed value of T for the original data is $\sum_{i=1}^{n} Y_i$, and here $T = \sum_{i=m+1}^{N} A_i$ is a sum of the last n elements of a random permutation of $\boldsymbol{Z}$. A very important subclass of (12.3) are the linear rank statistics given in the next section.

Assuming that each permutation of $\boldsymbol{A}$ is equally likely and thus has probability $1/N!$, it is easy to see that

$$P(A_i = a_s) = \frac{1}{N} \quad \text{for } s = 1, \ldots, N,$$

and

$$P(A_i = a_s, A_j = a_t) = \frac{1}{N(N-1)} \quad \text{for } s \neq t = 1, \ldots, N.$$

Then, using those two results, we get

$$\mathrm{E}(A_i) = \frac{1}{N} \sum_{i=1}^{N} a_i \equiv \overline{a}, \quad \text{for } i = 1, \ldots, N,$$

$$\mathrm{Var}(A_i) = \frac{1}{N} \sum_{i=1}^{N} (a_i - \overline{a})^2, \quad \text{for } i = 1, \ldots, N,$$

and

$$\mathrm{Cov}(A_i, A_j) = \frac{-1}{N(N-1)} \sum_{i=1}^{N} (a_i - \overline{a})^2, \quad \text{for } i \neq j = 1, \ldots, N.$$

Finally, putting these last three results together, we get

$$\mathrm{E}(T) = N\overline{c}\,\overline{a},$$

and

$$\mathrm{Var}(T) = \frac{1}{N-1} \sum_{i=1}^{N} (c_i - \overline{c})^2 \sum_{j=1}^{N} (a_j - \overline{a})^2, \tag{12.4}$$

where $\overline{a}$ and $\overline{c}$ are the averages of the a's and c's, respectively. These first two moments of T are sufficient for normal approximations based on the asymptotic normality of T as $N \to \infty$. In some cases it may be of value to use more complex approximations involving the third and fourth moments of T. Thus, the central third moment is

$$\mathrm{E}\{T - \mathrm{E}(T)\}^3 = \frac{N}{(N-1)(N-2)} \sum_{i=1}^{N} (c_i - \overline{c})^3 \sum_{j=1}^{N} (a_j - \overline{a})^3,$$

and the standardized third moment (skewness coefficient) is

$$\mathrm{Skew}(T) = \frac{\mathrm{E}\{T - \mathrm{E}(T)\}^3}{\{\mathrm{Var}(T)\}^{3/2}} = \frac{(N-1)^{1/2}}{(N-2)} \frac{\mu_3(\boldsymbol{c})\mu_3(\boldsymbol{a})}{\{\mu_2(\boldsymbol{c})\mu_2(\boldsymbol{a})\}^{3/2}},$$

where we have introduced the notation $\mu_q(\boldsymbol{c}) = N^{-1}\sum_{i=1}^N (c_i - \overline{c})^q$ for $q \geq 2$. Similarly the standardized central fourth moment (kurtosis coefficient) is

$$\begin{aligned}\mathrm{Kurt}(T) = \frac{\mathrm{E}\{T - \mathrm{E}(T)\}^4}{\{\mathrm{Var}(T)\}^2} &= \frac{(N+1)(N-1)}{N(N-2)(N-3)} \frac{\mu_4(\boldsymbol{c})\mu_4(\boldsymbol{a})}{\{\mu_2(\boldsymbol{c})\mu_2(\boldsymbol{a})\}^2} \\ &- \frac{3(N-1)^2}{N(N-2)(N-3)}\left[\frac{\mu_4(\boldsymbol{c})}{\{\mu_2(c)\}^2} + \frac{\mu_4(\boldsymbol{a})}{\{\mu_2(a)\}^2}\right] \\ &+ \frac{3(N^2 - 3N + 3)(N-1)}{N(N-2)(N-3)}.\end{aligned}$$

12.4.3 Linear Rank Tests

Many popular rank tests have the general form

$$T = \sum_{i=1}^{N} c(i)a(R_i) \tag{12.5}$$

of a *linear rank statistic*, where $c(1), \ldots, c(N)$ are called the *regression constants* and $a(1), \ldots, a(N)$ are called the *scores*, and $\boldsymbol{R}$ is the vector of ranks (possibly midranks due to ties) of some data vector $\boldsymbol{Z}$. There is a room for confusion here in the use of the notation for $\boldsymbol{c}$ and $\boldsymbol{a}$, because in the general notation of the last section, $(c_1, \ldots, c_N)$ and $(a_1, \ldots, a_N)$ are vectors of real numbers, but here $c(\cdot)$ and $a(\cdot)$ are functions so that $c_1 = c(1), \ldots, c_N = c(N)$ and $a_1 = a(1), \ldots, a_N = a(N)$. This function notation just makes it easier to work with rank statistics. In particular, the score functions $a(\cdot)$ are typically derived from *scores generating functions* ϕ via $a(i) = \phi(i/(N+1))$. In tied rank situations, $a(\cdot)$ needs to be defined for non-integer values.

The simplest setting is the two-sample problem where $\boldsymbol{Z}^T = (X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ and the c values are all zeroes for the Xs and ones for the Ys or vice-versa. A different situation covered by T, though, is for trend alternatives, where $c(1), \ldots, c(N)$ are the integers $1, \ldots, N$ and $T = \sum_{i=1}^N iR_i$ will tend to be large when Z_{i+1} tends to be larger than Z_i. A related problem is for N independent pairs $(X_1, Y_1), \ldots, (X_N, Y_N)$. Here, tests based on Spearman's Correlation (Section 12.7, p. 487) are equivalent to ones having the same null distribution as $T = \sum_{i=1}^N iR_i$.

Clearly T in (12.5) is a subclass of the linear permutation statistics given in (12.3, p. 458). Thus results for that class are inherited by T. For example, if $\boldsymbol{R}$ is uniformly distributed on the permutations of $1, \ldots, N$ (no tied ranks), then

$$\mathrm{E}(T) = N\overline{c}\,\overline{a},$$

and

$$\text{Var}(T) = \frac{1}{N-1}\sum_{i=1}^{N}(c(i)-\overline{c})^2\sum_{j=1}^{N}(a(j)-\overline{a})^2,$$

where of course $\overline{c}$ and $\overline{a}$ are the means of the c and a values, respectively. For a tied rank situation with observed vector of midranks $\boldsymbol{R}$, the expressions above still hold but with $a(j)$ replaced by $a(R_j)$.

For deciding on a score function in a given problem, we first select a parametric family and then derive an optimal score function for that family. An overview of how to do this is given in Section 12.5 (p. 473). The most important linear rank statistic is the Wilcoxon Rank Sum. So we give a few more details about it in the next section.

12.4.4 Wilcoxon-Mann-Whitney Two-Sample Statistic

For two independent samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, Wilcoxon (1945) introduced the linear rank statistic

$$W = \sum_{i=m+1}^{N} R_i, \tag{12.6}$$

where $R_1, \ldots, R_N$ are the joint rankings of $\boldsymbol{Z} = (X_1, \ldots, X_m, Y_1, \ldots, Y_n)^T$, $N = m + n$. The Wilcoxon Rank Sum test has a number of optimal properties that are mentioned in Section 12.5 (p. 473). Along with the Wilcoxon Signed Rank test for paired data (Section 12.8.3, 494), it is the simplest and most important rank test.

Independently, Mann and Whitney (1947) proposed the equivalent statistic

$$W_{\text{YX}} = \sum_{i=1}^{m}\sum_{j=1}^{n} I(Y_j < X_i), \tag{12.7}$$

where $I(\cdot)$ is the indicator function. In the absence of ties $W_{\text{YX}} = mn + n(n+1)/2 - W$. Another equivalent version is

$$W_{\text{XY}} = \sum_{i=1}^{m}\sum_{j=1}^{n} I(Y_j > X_i), \tag{12.8}$$

with $W_{\text{XY}} = W - n(n+1)/2$. We prefer this latter version and define the U-statistic estimator of $\theta_{\text{XY}} = P(Y_1 > X_1)$

$$\widehat{\theta}_{\text{XY}} = \frac{W_{\text{XY}}}{mn} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} I(Y_j > X_i). \tag{12.9}$$

In a clinical trial, θ_{XY} can be viewed as the probability of a more favorable response for a randomly selected patient getting Treatment 2 compared to another patient getting Treatment 1. For screening tests where a "positive" is declared if $Y > c$ for a diseased subject or if $X > c$ for a non-diseased subject, then θ_{XY} is the area under the receiver operating characteristic (ROC) curve. This interpretation is developed in Problem 12.8 (p. 525).

For hand computations, W is much easier to handle than these U-statistic versions. The null moments follow easily from Section 12.4.2 (p. 458) after noting that $c(1) = \cdots = c(m) = 0$ and $c(m+1) = \cdots = c(N) = 1$ lead to $\overline{c} = n/N$ and $\sum_{i=1}^{N}(c(i) - \overline{c})^2 = mn/N$. The null mean is $n(N+1)/2$ whether there are ties or not. The variance follows from (12.4, p. 459). With no ties, we have

$$\mathrm{Var}(W) = \frac{mn(N+1)}{12}. \tag{12.10}$$

With ties so that $(R_1, \ldots, R_N)$ are the tied ranks, we have

$$\mathrm{Var}(W) = \frac{mn}{N(N-1)} \left\{ \sum_{i=1}^{N} R_i^2 - \frac{N(N+1)^2}{4} \right\}. \tag{12.11}$$

Lehmann (1975, p. 20) gives a different expression for the variance of W in the face of ties,

$$\mathrm{Var}(W) = \frac{mn(N+1)}{12} - \frac{mn\sum_{i=1}^{e}(d_i^3 - d_i)}{12N(N-1)}, \tag{12.12}$$

where e are the number of tied groups, and d_i is the number of tied observations in each group. For example, with the simple example data modified to $(\{6, 7\}, \{7, 18, 11, 9\})$, the midranks are $(1, 2.5, 2.5, 6, 5, 4)$ and $e = 1$, $d_1 = 2$; so $\mathrm{Var}(W) = (2)(4)(6+1)/12 - (2)(4)[2^3 - 2]/[12(6)(5)] = 4.53$. Expression (12.12) may be easier to use by hand than (12.11), but its main value may be to show that the variance of W for tied data is always smaller than (12.10) for untied data.

The U-statistic versions in (12.7)–(12.9) are useful for easy calculation of moments and derivation of asymptotic normality under non-null distributions. For example, using equation (3.4.7, p. 91) of Randles and Wolfe (1979) for the variance of a two-sample U-statistic from independent iid samples, we have that

$$\mathrm{Var}(\widehat{\theta}_{\mathrm{XY}}) = \frac{1}{mn}\left\{(m-1)(\gamma_{0,1} - \theta_{\mathrm{XY}}^2) + (n-1)(\gamma_{1,0} - \theta_{\mathrm{XY}}^2) + \gamma_{1,1} - \theta_{\mathrm{XY}}^2\right\}, \tag{12.13}$$

where in the absence of ties $\gamma_{0,1} = P(Y_1 > X_1, Y_1 > X_2)$, $\gamma_{1,0} = P(Y_1 > X_1, Y_2 > X_1)$, and $\gamma_{1,1} = \theta_{\mathrm{XY}} = P(Y_1 > X_1)$. If the X and Y have identical continuous distributions, then it is easy to show that $\gamma_{0,1} = \gamma_{1,0} = 1/3$ and $\gamma_{1,1} = \theta_{\mathrm{XY}} = 1/2$ and (12.13) reduces to (12.10).

In the presence of ties, the U-statistic quantities need to be modified by adding $I(Y_j = X_i)/2$ to the indicators in the sums. For example,

$$\widehat{\theta}_{\text{XY}} = \frac{W_{\text{XY}}}{mn} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left\{I(Y_j > X_i) + I(Y_j = X_i)/2\right\}. \tag{12.14}$$

The relationships $W_{\text{YX}} = mn + n(n+1)/2 - W$ and $W_{\text{XY}} = W - n(n+1)/2$ then continue to hold. The definitions of $\gamma_{0,1}$, $\gamma_{1,0}$, and $\gamma_{1,1}$ for use in (12.13) have to be modified in the face of ties; see, for example, Boos and Brownie (1992, p. 72). In the next section we give the basic asymptotic normal results for linear statistics under the null hypothesis of identical populations. Those general results are useful for approximate critical regions for permutation and rank statistics. However, the Wilcoxon statistics are special because they are related to the U-statistic $\widehat{\theta}_{\text{XY}}$ for which a large body of theory exists. In particular, $\widehat{\theta}_{\text{XY}}$ is $\text{AN}\left\{\theta_{\text{XY}}, \text{Var}(\widehat{\theta}_{\text{XY}})\right\}$, and this follows from basic U-statistic theory with no assumptions except that $X_1, \ldots, X_m$ are iid with any distribution function $F(x)$, and $Y_1, \ldots, Y_n$ are iid with any distribution function $G(x)$. Because this asymptotic result is not just for null situations, it helps us think about i) the form of the alternative hypothesis, ii) the classes of distribution functions for which the Wilcoxon Rank Sum is consistent, in other words, rejects with probability converging to 1, and iii) asymptotic power and sample size determination. We now discuss these ideas.

In general, the null hypothesis of interest is

$$H_0 : F(x) = G(x), \text{ each } x \in (-\infty, \infty).$$

However, the alternative hypothesis can be formulated in several ways. The most common way is to assume the shift model $G(x) = F(x - \Delta)$, and then the alternative hypothesis is purely in terms of Δ, for example

$$H_1 : \Delta > 0.$$

Another popular, more nonparametric, way to phrase the alternative is

$$H_2 : F(x) \geq G(x), \text{ each } x \in (-\infty, \infty),$$

and with strict inequality for at least one x. Here, G is said to be *stochastically larger* than F. Clearly, H_2 is a larger class of alternatives since $(F, G) \in H_1$ implies $(F, G) \in H_2$. Lastly, the natural alternative when thinking in terms of $\widehat{\theta}_{\text{XY}}$ is

$$H_3 : \theta_{\text{XY}} > \frac{1}{2}.$$

Now if F and G are continuous distribution functions and $(F,G) \in H_2$, then $(F,G) \in H_3$. This follows from

$$\theta_{\mathrm{XY}} = P(Y_1 > X_1) = \int\int I(y > x)\,dF(x)\,dG(y) = \int\{1 - G(x)\}\,dF(x),$$

after noting that if continuous distribution functions satisfy $F(x) > G(x)$ for at least one x, then this strict inequality must hold for an interval of x values, and $\int F(x)\,dF(x) = 1/2$. Assuming that H_3 holds, then the Wilcoxon Rank Sum test is consistent because of the general asymptotic normality result mentioned above. This also means that it is also consistent under alternatives H_1 and H_2.

Lastly, following Noether (1987), the approximate power of a one-sided α level test when $\theta_{\mathrm{XY}} > \frac{1}{2}$ is given by

$$1 - \Phi\left\{\frac{1/2 - \theta_{\mathrm{XY}}}{\rho\sigma_0} + \frac{\Phi^{-1}(1-\alpha)}{\rho}\right\}, \tag{12.15}$$

where σ_0 is the square root of the null variance of W (12.10, p. 462), ρ is the ratio of the square root of the non-null variance of W (m^2n^2 times eq. 12.13, p. 462) to σ_0, and Φ is the standard normal distribution function. Typically, ρ is close to 1. Letting $\rho = 1$ and $m = \lambda N$, the total sample size N required to have power $1 - \beta$ for alternative θ_{XY} is given by Noether (1987) to be

$$N = \frac{\left\{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\right\}^2}{12\lambda(1-\lambda)(\theta_{\mathrm{XY}} - 1/2)^2}. \tag{12.16}$$

This is a fairly simple formula, but it might be preferable to state power and sample size in terms of the shift model. Plugging in $G(x) = F(x - \Delta)$, we have

$$\theta_{\mathrm{XY}} = P(Y_1 > X_1) = \int\{1 - F(x - \Delta)\}\,dF(x).$$

For example, if we wanted shifts of size Δ/σ in a normal(μ, σ^2) population, then a simple R program to get θ_{XY} using the midpoint rule is

```
theta.xy<-function(delta,n=10000){
# u-stat parameter for normal shift delta/sigma
# for sigma=1
# n is the number of points for midpoint rule
    points<-(2*(1:n)-1)/(2*n)
    mean(1-pnorm(qnorm(points)-delta))
}
```

If $\Delta/\sigma = .5$, then

```
> theta.xy(.5,10000)
[1] 0.6381632
```

so that $\theta_{\mathrm{XY}} = .638$. Choosing $\alpha = .05$, $\beta = .80$, and $\lambda = 1/2$, we find $N = 108$ or $m = n = 54$.

12.4.5 Asymptotic Normal Approximation

Approximate normal distributions for linear statistics have been the most popular approximation to permutation distributions, especially for rank statistics. Here we use the following permutation Central Limit Theorem for $T = \sum_{i=1}^{N} c_i A_i$, introduced in (12.3, p. 458), directly from Puri and Sen (1971, p. 73) who give credit to Wald and Wolfowitz (1944), Noether (1949), and Hoeffding (1951). The notation $\mu_q(\boldsymbol{c})$ is for the qth central moment $N^{-1}\sum_{i=1}^{N}(c_i - \overline{c})^q$.

Theorem 12.2 (Wald-Wolfowitz-Noether-Hoeffding). *If for $N \to \infty$*

(i)

$$\frac{\mu_q(\boldsymbol{c})}{\mu_2(\boldsymbol{c})^{q/2}} = O(1) \quad \textit{for all } q = 3, 4, \ldots$$

(ii)

$$\frac{\mu_q(\boldsymbol{a})}{\mu_2(\boldsymbol{a})^{q/2}} = o(N^{r/2-1}) \quad \textit{for all } q = 3, 4, \ldots,$$

then

$$\frac{T - E(T)}{\sqrt{Var(T)}} \xrightarrow{d} N(0, 1).$$

In a particular problem either or both of the vectors $\boldsymbol{c}$ and $\boldsymbol{a}$ may be random, that is, calculated from the data $\boldsymbol{Z}$. In such cases we would need to show that the appropriate conditions (i) and/or (ii) hold $wp1$ with respect to the random vector $\boldsymbol{Z}$. Moreover, the conclusion of Theorem 12.2 is that the permutation distribution of the standardized T converges to a standard normal distribution with probability one with respect to $\boldsymbol{Z}$.

In the case of linear rank statistics without ties, we can give a much simpler theorem due to Hajek (1961). We follow the exposition given in Randles and Wolfe (1979, Ch. 8) and state their version of Hajek's theorem.

Theorem 12.3 (Hajek). *Let $T = \sum_{i=1}^{N} c(i)a(R_i)$ be the linear rank statistic, where the rank vector $\boldsymbol{R}$ comes from data vector $\boldsymbol{Z}$ that is continuous (no ties with probability one) and exchangeable, the constants $c(1), \ldots, c(N)$ satisfy the Noether condition*

$$\frac{\sum_{i=1}^{N}(c(i) - \overline{c})^2}{\max_{1 \le i \le N}(c(i) - \overline{c})^2} \to \infty \quad \textit{as } N \to \infty, \tag{12.17}$$

and the scores have the form $a(i) = \phi(i/(N+1))$, where ϕ can be written as the difference of two nondecreasing functions and $0 < \int_0^1 \phi(t)^2 dt < \infty$ and $\int_0^1 |\phi(t)| dt < \infty$. Then T is $AN\{N\overline{c}\,\overline{a}, Var(T)\}$ as $N \to \infty$.

It has been customary to use the normal approximation with rank statistics, often with a continuity correction. For example, in the two-sample problem, consider the Wilcoxon Rank Sum W of (12.6, p. 461). Note that for application of Theorem 12.3 above, $\phi(u) = u$, and the theorem actually applies directly to $W/(N+1)$. For the simple example of Section 1.2 where $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y}) = (6, 8, 7, 18, 11, 9)$ with ranks $\boldsymbol{R} = (1, 3, 2, 6, 5, 4)$, we find $W = 17$, $\text{E}(W) = 4(6+1)/2 = 14$, $\text{Var}(W) = (2)(4)(6+1)/12 = 14/3$ (from 12.10, p. 462), and the normal approximation p-value is

$$p \approx P\left(N(0,1) \geq \frac{17-14}{\sqrt{14/3}}\right) = P(N(0,1) \geq 1.39) = 0.08.$$

With continuity correction the normal approximation p-value is

$$p \approx P\left(N(0,1) \geq \frac{17-14-1/2}{\sqrt{14/3}}\right) = P(N(0,1) \geq 1.16) = 0.12.$$

Lehmann (1975, p. 16) cites Kruskal and Wallis (1952, p. 591) with the recommendation that the continuity correction be used when the probability is above 0.02. Recall that the exact null distribution of W can be obtained from Table 12.1 leading to the usual p-value $P(W \geq 17) = 2/15 = 0.13$ which is closer to the continuity corrected value.

When there are tied values, we can still use the normal approximation with W, but we must be sure to use the null variance from (12.11, p. 462) or (12.12, p. 462) and not from (12.10, p. 462). Lehmann (1975, p. 20) does not use the continuity correction in the presence of ties.

We can also look at approximations to the permutation p-value of $T = \sum_{i=1}^{n} Y_i$ which is permutationally equivalent to the two-sample t statistic. For the simple example $\boldsymbol{c} = (0, 0, 1, 1, 1, 1)$ and $\boldsymbol{a} = \boldsymbol{z} = (6, 8, 7, 18, 11, 9)$. Thus, $\text{E}(T) = (6)(4/6)(59/6) = 39.33$, $\text{Var}(T) = 25.23$, and the normal approximation p-value is

$$p \approx P\left(N(0,1) \geq \frac{45-39.33}{\sqrt{25.23}}\right) = P(N(0,1) \geq 1.13) = 0.13.$$

This seems almost too good an approximation to the true permutation p-value of $2/15 = 0.13$. Usually the t approximation p-value is more accurate, but here it is $P(t_4 \geq 1.17) = 0.15$.

12.4.6 Edgeworth Approximation

Edgeworth approximations were mentioned briefly in Ch. 3 (5.6, p. 219) and Ch. 9 (11.7, p. 426). Basically, an Edgeworth expansion is an approximation to the distribution function of an asymptotically normal statistic. It is based on

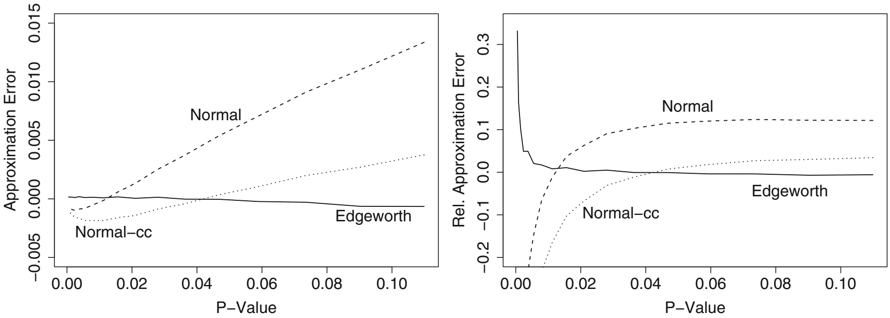

Fig. 12.1 Error (Left Panel) and relative error (Right Panel) of approximations to Wilcoxon Rank Sum p-values for $m = 10$, $n = 6$: normal approximation, normal approximation with continuity correction, and the Edgeworth approximation in (12.18, p. 467)

estimation of Skew and/or Kurt and other higher moments of the statistic. Rigorous development of Edgeworth expansions for general permutation statistics under the null hypothesis may be found in Bickel (1974), Bickel and van Zwet (1978), and Robinson (1980). However, it has not proved of much practical use for obtaining critical values or p-values of permutation statistics except in the special case of the Wilcoxon Rank Sum W and of the one-sample Wilcoxon signed rank statistic.

Here we give the approximation for W originally due to Fix and Hodges (1955). For $W = \sum_{i=1}^{n} R_i$,

$$P(W \geq w) \approx 1 - \Phi(t) - \left\{ \frac{m^2 + n^2 + mn + m + n}{20mn(m + n + 1)} \right\} (t^3 - 3t)\phi(t), \quad (12.18)$$

where ϕ and Φ are the standard normal density and distribution function, respectively, and $t = \{w - \mathrm{E}(W) - 1/2\}/\sqrt{\mathrm{Var}(W)}$, $\mathrm{E}(W) = n(N + 1)/2$, $\mathrm{Var}(W) = mn(N + 1)/12$.

Figure 12.1 gives the error = true p-value − (12.18) and the relative error = [true p-value − (12.18)]/(true p-value) of (12.18) compared to the true p-value and similar quantities for the normal approximations. The range of the p-values is most of the right tail of the distribution function of W plotted in reverse order, that is, 0.0005 to 0.11. The Edgeworth approximation is excellent for p-values larger than 0.0024, but then deteriorates as the p-value gets very small. For example, when the true p-value is 0.00087, the Edgeworth approximation is 0.00073, and at 0.00025 it is 0.00009. The right panel of Figure 12.1 is especially helpful for illuminating what happens at small p-values. The normal approximation is much cruder, and below 0.02 we can see that the continuity correction is no longer useful.

Figure 12.1 suggests that (12.18) can be used for most values of W, thus essentially replacing tabled values of the distribution of W. However, when there are ties in the data, (12.18) as well as tabled values are no longer correct, and the exact permutation distribution (or a Monte Carlo approximation) is required.

12.4.7 Box-Andersen Approximation

Pitman (1937a,b) and Welch (1937) pioneered an approximation to permutation distributions that was modernized by Box and Andersen (1955) and Box and Watson (1962). These later authors mainly used the approach to show the Type I error robustness of F statistics for tests comparing means and the nonrobustness of tests comparing variances. However, we follow the Box and Andersen (1955) formulation since it is the most straightforward.

The basic idea of the approximation is to get F statistics into their equivalent "beta" version, then match the first two permutation moments of this beta version to what one gets from the first two moments of a beta distribution with degrees of freedom multiplied by a constant d. Solving for d leads to the approximation of the permutation distribution of the F statistics by an F distribution with usual degrees of freedom multiplied by d. We develop the approximation here for the two-sample problem and later give it for one-way and two-way ANOVA situations.

The square of the t statistic in (12.1, p. 452) may be written in the one-way ANOVA F form

$$t^2 = \frac{m(\overline{X}-\overline{Z})^2 + n(\overline{Y}-\overline{Z})^2}{s_p^2} = \frac{\text{SSTR}}{\text{SSE}/(N-2)}, \qquad (12.19)$$

where recall we use the Z's to denote all the X and Y values thrown together, and SSTR and SSE are sums of squares for treatments and error, respectively. Using the fact that $\sum_{i=1}^{N}(Z_i-\overline{Z})^2 = \text{SSTR} + \text{SSE}$, we have for the beta version of the F statistic

$$b(t^2) = \frac{t^2}{t^2+N-2} = \frac{\text{SSTR}}{\sum_{i=1}^{N}(Z_i-\overline{Z})^2}.$$

Note that for normal data under the null hypothesis, $b(t^2)$ has a beta$(1/2, (N-2)/2)$ distribution. Originally $b(t^2)$ was used with the beta critical values rather than t^2 with $F(1, N-2)$ critical values. Although, t^2 and $b(t^2)$ are equivalent test statistics, for permutation analysis $b(t^2)$ is much simpler because the denominator is constant over permutations. Thus, the first permutation moment is

$$\text{E}_\text{P}\{b(t^2)\} = \frac{m\text{Var}_\text{P}(\overline{X}) + n\text{Var}_\text{P}(\overline{Y})}{\sum_{i=1}^{N}(Z_i-\overline{Z})^2} = \frac{1}{N-1},$$

where we have used (12.4, p. 459) to get

$$\text{Var}_\text{P}(\overline{X}) = \frac{n\sum_{i=1}^{N}(Z_i-\overline{Z})^2}{mN(N-1)} \qquad \text{Var}_\text{P}(\overline{Y}) = \frac{m\sum_{i=1}^{N}(Z_i-\overline{Z})^2}{nN(N-1)}.$$

Note also that under normal theory $\mathrm{E}\{b(t^2)\} = 1/2/(1/2+(N-2)/2) = 1/(N-1)$ from the beta distribution. Thus, the normal theory and permutation first moments of $b(t^2)$ are both $1/(N-1)$. The next step is to calculate the permutation variance of $b(t^2)$ (involving fourth moments), equate it to the variance of a beta$(d/2, d(N-2)/2)$ distribution, $2(N-2)/[d(N-1)(N+3)]$, and solve for d. Box and Andersen (1955, p. 13) give d for the general one-way ANOVA situation with k groups and sample sizes $n_1, n_2, \ldots, n_k$:

$$d = 1 + \left(\frac{N+1}{N-1}\right) \frac{c_2}{(N^{-1}+A)^{-1} - c_2}, \tag{12.20}$$

where

$$A = \frac{N+1}{2(k-1)(N-k)} \left(\frac{k^2}{N} - \sum_{i=1}^{k} \frac{1}{n_i}\right),$$

$c_2 = k_4/k_2^2$,

$$k_2 = \frac{1}{N-1} \sum_{i=1}^{N} (Z_i - \overline{Z})^2, \tag{12.21}$$

$$k_4 = \frac{N(N+1)\sum_{i=1}^{N}(Z_i - \overline{Z})^4 - 3(N-1)\left\{\sum_{i=1}^{N}(Z_i - \overline{Z})^2\right\}^2}{(N-1)(N-2)(N-3)}. \tag{12.22}$$

The statistics k_2 and k_4 are unbiased estimators of the population cumulants introduced in Chapter 1.

For our two-sample t^2, $k = 2$, $n_1 = m$, $n_2 = n$, $m + n = N$, and the Pitman-Welch-Box-Andersen approximation is to compare t^2 to an $F(d, d(m+n-2))$ distribution. Box and Andersen (1955) show that $\mathrm{E}(d) \approx 1 + (\mathrm{Kurt} - 3)/N$ under the null hypothesis of sampling from equal populations with kurtosis Kurt. Thus, t^2 with the usual $F(1,(m+n-2))$ is quite Type I error robust to nonnormality since the correction d is relatively small for moderate size N. Also, for long-tailed distributions with thicker tails than the normal distribution, Kurt >3 and thus $d > 1$, so that using the $F(1,(m+n-2))$ critical values results in conservative tests, that is, true test levels less than the nominal α values. For example, with Laplace data, Kurt $= 6$ and $d \approx 1+3/N$; at $m = n = 10$ $d \approx 1.15$, and a nominal $\alpha = .05$ level test would actually have true level approximately .043. For continuous uniform data, Kurt $= 1.8$; at $m = n = 10$ $d \approx .94$ and a nominal $\alpha = .05$ level test would have true level approximately .053. Since these deviations from α are small, common practice is to just use the standard $F(1,(m+n-2))$ reference distribution with the t^2 statistic rather than the permutation distribution or an approximation to it.

Although t^2 is Type I error robust in the face of outliers, it loses power because outliers inflate the variance estimate in the denominator of t^2. Thus t^2 is not Type II error robust when sampling from distributions heavier-tailed than the normal. In contrast, as we mentioned in the Chapter introduction, the Wilcoxon Rank Sum statistic W is Type II error robust, and later we use asymptotic power calculations to verify its superiority to t^2. But for the moment, we note that W is related to t^2 applied to the ranks of the data, and therefore inherits robustness to outliers because the ranks themselves are resistant to the effects of outliers. This relationship also allows us to use the above approximation for the permutation distribution of W.

Define the standardized Wilcoxon Rank Sum statistic by

$$W_{\mathrm{S}} = \frac{W - \mathrm{E}(W)}{\{\mathrm{Var}(W)\}^{1/2}}.$$

Then, t^2 applied to the ranks of the observations, that is, the X ranks $R_1, \ldots, R_m$ replacing $X_1, \ldots, X_m$, and the Y ranks $R_{m+1}, \ldots, R_N$ replacing $Y_1, \ldots, Y_n$, results in

$$t_{\mathrm{R}}^2 = \frac{(N-2)W_{\mathrm{S}}^2}{N - 1 - W_{\mathrm{S}}^2}.$$

Thus t_{R}^2 and W are equivalent test statistics and we can apply the Box-Andersen approximation to t_{R}^2 using $d \approx 1 + (1.8 - 3)/N$ because the ranks are a uniform distribution on the integers 1 to N and thus have Kurt $\approx$ 1.8, the kurtosis of a continuous uniform distribution. For example, in the case of $m = 10$ and $n = 6$ given in Figure 12.1 (p. 467), the Box-Andersen approximation along with the continuity correction gives results that are considerably better than the normal approximation with continuity correction but not quite as good as the Edgeworth approximation. In later sections we see that the Box-Andersen approximation is very good in one-way and two-way ANOVA situations when the number of treatments is greater than two.

12.4.8 Monte Carlo Approximation

In the previous sections, approximations to permutation distributions were given for statistics based on linear forms, and essentially rely on the Central Limit Theorem and its extensions. However, the simplest and most important approximation to a permutation distribution is to randomly sample from the set of all possible permutations, and directly estimate the permutation distribution. This approach can be used for any statistic T, and its accuracy is determined simply by the number B of random permutations used. This resampling of permutations is very similar to resampling in the bootstrap world, and we suggest sampling with replacement because of simplicity although sampling without replacement could be used.

Suppose that T calculated on all permutations has distinct values $t_1, \ldots, t_k$. For example, in Table 12.1 (p. 453) the t statistic has $k = 13$ distinct values –2.98, –1.72, –1.36, –1.08, –0.84, –0.06, 0.12, 0.30, 0.49, 0.69, 0.91, 1.17, 1.47, corresponding to the 15 permutations (0.49 and 0.91 appeared twice). The Monte Carlo approach is to randomly select B times from the 15 possible permutations, calculate the statistic for each random selection, say $T_1^*, \ldots T_B^*$, and let the number of T^*s equal to t_i be denoted N_i, $i = 1, \ldots, k$. If we select permutations with replacement, then $(N_1, \ldots, N_k)$ is multinomial$(B; p_1, \ldots, p_k)$, where p_i is the permutation distribution probability of obtaining t_i. The estimates N_i/B have binomial variances $p_i(1 - p_i)/B$. Thus, if we were trying to estimate the probabilities in Table 12.2 (p. 453), most of the estimates would have variance $(1/15)(14/15)/B$ although two of them would have variance $(2/15)(13/15)/B$ because of the duplication of values 0.49 and 0.91.

In typical applications, we are not interested in the whole permutation distribution, but merely want to estimate the p-value given in (12.2, p. 455) using

$$\widehat{p} = \frac{\left\{\#T_i^* \geq T_0\right\}}{B},$$

where T_0 is the value of the statistic for the original data. In the simple example, $T_0 = 1.17$. Recall that in this case the true permutation p-value is $2/15 = .13$. Thus, $B = 1000$ would yield an estimate with standard deviation $\{(.13)(.87)/1000\}^{1/2} = .01$ that would be adequate for most purposes. However, if the p-value were smaller, say .005, then we would want to take B larger so that the standard deviation of the estimate would be a small fraction of the p-value, say not more than 10–20%. For example, setting $.001 = \{(.005)(.995)/B\}^{1/2}$ would suggest $B = 4975$. When the estimated p-value is to be used with rejection rules like "reject H_0 if $\widehat{p} \leq \alpha$," then it is wise to choose B so that $(B + 1)\alpha$ is an integer as was discussed in the bootstrap Section 11.6.2 (p. 440) as the"99 rule". Mainly this would be used in Monte Carlo simulation studies where $B = 99$ or $B = 199$ might be used to save computing time. However, in situations where computations of the test statistic are extremely expensive, one may view the random partitions as part of the test itself, and the procedure "reject H_0 if $\widehat{p} \leq \alpha$" is called a Monte Carlo test, not just an approximation to the permutation test. This approach was first introduced by Barnard (1963) and later studied by Hope (1968), Jöckel and Jockel (1986), and Hall and Titterington (1989).

12.4.9 Comparing the Approximations in a Study of Two Drugs

A new drug regimen (B) was given to 16 subjects, and one week later each subject's status was assessed. A second independent group of 13 subjects received the standard drug regimen (A). Both sets of measurements were compared to baseline measurements taken before the treatment period began. The difference from baseline data is given in Figure 12.2. This is real data but the actual details are confidential. The drug company wanted to prove that regimen B involving their new drug had larger differences from baseline than the standard. In terms of means of the differences, the testing situation is $H_0 : \mu_B = \mu_A$ versus $H_a : \mu_B > \mu_A$.

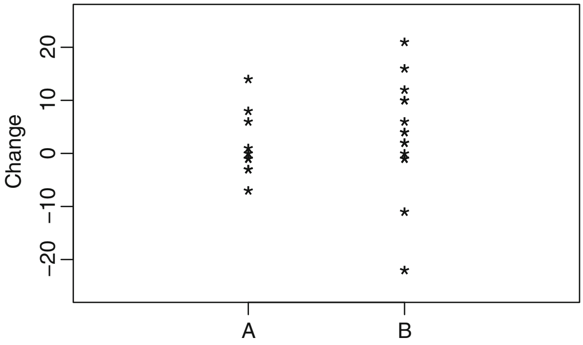

Fig. 12.2 Change from Baseline for Drugs A and B

The sample means and standard deviations are $\overline{X} = .92, \overline{Y} = 3.19, s_X = 5.45, s_Y = 10.21$. The standard pooled t from (12.1, p. 452) is .72 with one-sided p-value .24 from the t distribution. The exact permutation t p-value is 0.249, but with a large p-value like this, the t distribution approximation is adequate and agrees with the Type I error robustness mentioned previously. The Box-Andersen $d = 1.074$ leading to an adjusted t p-value of .245.

However, Figure 12.2 reveals that most of the Drug B subjects have positive changes from baseline whereas the Drug A changes are more centered around 0. The two large negative values −22 and −11 have a strong effect on the t statistic. The Wilcoxon Rank Sum statistic W is less affected by outliers, and might paint a different picture. First we compute the midranks and list them with the data ordered within samples.

A:	−7	−3	−3	−1	−1	−1	−1	0	0	1	6	8	14
Rank:	3	4.5	4.5	9	9	9	9	14	14	16	21.5	23	27

B:	−22	−11	−1	−1	−1	0	2	2	4	4	6	10	10
Rank:	1	2	9	9	9	14.0	17.5	17.5	19.5	19.5	21.5	24.5	24.5
B:	12	16	21										
Rank:	26	28	29										

Then $W = 1 + 2 + \ldots + 28 + 29 = 271.5$. The null mean of W is $(16)(16 + 13 + 1)/2 = 240$. To compute the null variance using the formula for ties, (12.12, p. 462), note that there are $e = 16$ distinct values and 2 values tied at −3, 7 tied at −1, 3 tied at 0, 2 tied at 2, 2 tied at 4, 2 tied at 6, and 2 tied at 10. Thus the null variance is

$$\frac{(16)(13)(16+13+1)}{12} - \frac{(16)(13)}{(12)(29)(29-1)}\left[(7^3-7)+(3^3-3)+5(2^3-2)\right]$$
$$= 520 - 8.325 = 511.675.$$

The approximate normal statistic is $(271.5 - 240)/\sqrt{511.675} = 1.39$ with p-value .082. The t statistic on the ranks is 1.42 with p-value .084. The Box and Andersen (1955) degrees of freedom approximation with $d = (1 - 1.2/29) = 0.96$ does not change that latter p-value until the fourth decimal. The Edgeworth approximation p-value is .084 without continuity correction and .087 with continuity correction.

Unfortunately, because of the ties we cannot trust the exact tables or a continuity correction or the Edgeworth approximation. Thus, it seems wise to either calculate the exact permutation p-value or estimate it by Monte Carlo methods. With $B = 10,000$ we got $\widehat{p} = .085$ with 95% confidence interval (.080,.090). Rather than make B larger, in this case it is fairly easy to get the exact p-value = .0849 with existing software. Summarizing the one-sided p-values, we have

Statistic	Method	P-value
t	Exact Permutation	0.2490
	$t(m+n-2)$	0.239
	Box-Andersen	0.245
W	Exact Permutation	0.0849
	Normal	0.082
	$t(m+n-2)$	0.084
	Box-Andersen	0.084
	Edgeworth	0.084
	Edgeworth (with cc)	0.087
	Monte Carlo (B=10,000)	0.085

So this is a situation where the Wilcoxon Rank Sum statistic might be preferred to the t because of its robustness to outliers. Here it apparently downweighted the outliers -22 and -11 enough to have a much lower p-value than the t statistic. The normal and t approximations to the W p-value are quite reasonable here, but we would not know that without getting the exact p-value = .0849 or by estimating it fairly accurately.

12.5 Optimality Properties of Rank and Permutation Tests

There are actually very few results available on the optimality properties of permutation tests. The main source is Lehmann and Stein (1949), see also Lehmann (1986, Ch. 5), who give the form of the most powerful permutation test for shift alternatives and note that it depends on a variety of unknown quantities including the form of the distribution. In the particular case of normal data with common unknown variance, they show that the most powerful permutation statistic is $\overline{Y}$ or equivalently $\overline{Y} - \overline{X}$ or the pooled two sample t statistic. Thus general optimality results are not available, but a general approach is clear: derive an (asymptotically) optimal parametric test statistic under a specific parametric family assumption (your best guess), and use the permutation approach for critical values. The resulting permutation test is valid under the null hypothesis for any distribution as long as the conditions of Theorem 12.1 (p. 457) hold, and is close to optimal if the distribution of the data is close to the one used to derive the test statistic.

For rank statistics there are two main bodies of results: locally most powerful rank tests and asymptotically most powerful rank tests based on Pitman Asymptotic Relative Efficiency (ARE). Here we briefly give the flavor of these approaches and main results leaving technical details for the Appendix.

12.5.1 Locally Most Powerful Rank Tests

For simplicity we focus on the two-sample shift model where $X_1, \ldots, X_m$ are iid with distribution function F, and $Y_1, \ldots, Y_n$ are iid with distribution $G(y) = F(y - \Delta)$. We assume that F is continuous with density f. Consider

$$H_0 : \Delta = 0 \quad \text{versus} \quad H_a : \Delta > 0.$$

If there exists a rank test that is uniformly most powerful of level α for some $\epsilon > 0$ in the restricted testing problem

$$H_0 : \Delta = 0 \quad \text{versus} \quad H_{a,\epsilon} : 0 < \Delta < \epsilon,$$

then we say that the test is the *locally most powerful rank test* for the original testing problem.

The basic approach to finding a locally most powerful rank test is to take a Taylor expansion of the probability of the rank vector as a function of Δ and maximize its derivative at $\Delta = 0$. For sufficiently small Δ, the values of the rank vector that are ordered by its probability under the alternative Δ are the same as those ordered by its derivative at $\Delta = 0$. Thus, we need only obtain an expression for the derivative and maximize it. These details are left for the Appendix.

For the two-sample shift problem, the locally most powerful rank test rejects for large values of

$$T = \sum_{i=m+1}^{N} a(R_i),$$

where $a(i) = \mathrm{E}\{\phi(U_{(i)}, f)\}$,

$$\phi(u, f) = -\frac{f'(F^{-1}(u))}{f(F^{-1}(u))} \tag{12.23}$$

is called the optimal score function, and $U_{(1)} \le U_{(2)} \le \cdots \le U_{(N)}$ are the order statistics from a uniform (0,1) distribution. Recall that $R_{m+1}, \ldots, R_N$ are the ranks of the Y values in the joint ranking of all the X's and Y's together. We see in the next section that a closely related statistic, $\sum_{i=m+1}^{N} \phi(R_i/(N+1), f)$, is asymptotically equivalent and comes naturally from asymptotic relative efficiency considerations.

If F is the logistic distribution, then we are led to the Wilcoxon Rank Sum as the locally most powerful rank test for shift alternatives because $-f'(x)/f(x) = 2F(x) - 1$ and $\mathrm{E}\{U_{(i)}\} = i/(N+1)$. When F is a normal distribution, then the optimal score function is $\phi(u, f) = \Phi^{-1}(u)$, and the locally most powerful test is based on the *normal scores*

$$a(i) = \mathrm{E}\{\Phi^{-1}(U_{(i)})\} = \mathrm{E}\{Z_{(i)}\},$$

where $Z_{(i)}$ is a standard normal order statistic. For shifts in the scale of an exponential distribution, $F(x;\sigma) = 1 - \exp(-x/\sigma)$, we can turn it into a shift in location of the negative of an extreme value distribution, $F(x) = 1 - \exp\{-\exp(x)\}$, by taking the natural logarithm of the exponential data. The resulting optimal test has score

$$a(i) + 1 = \sum_{j=N+1-i}^{N} \frac{1}{j},$$

where the latter sum is the expected value of the ith order statistic from a standard exponential distribution. These are called *Savage* scores from Savage (1956). In censored data situations, the analogous test is called the logrank test.

Lehmann (1953) studied alternatives of the form

$$F_\Delta(x) = (1-\Delta)F(x) + \Delta F^2(x),$$

and showed that the Wilcoxon Rank Sum is the locally most powerful rank test for these alternatives. In general, alternatives of the form $F_\Delta(x) = h_\Delta(F(x))$ for some function $h_\Delta(u)$, are called *Lehmann alternatives*. They have the property that two-sample rank tests have the same distribution under an alternative Δ for all continuous F.

Johnson et al. (1987) consider locally most powerful rank tests using Lehmann alternatives for the nonresponder problem where only a fraction of subjects respond to treatment. Conover and Salsburg (1988) consider other locally most powerful rank tests for the nonresponder problem. Additional situations where locally most powerful rank tests are considered include Doksum and Bickel (1969) and Bhattacharyya and Johnson (1973).

The optimal score functions (12.23, p. 475) appear in the k-sample problem, Section 12.6 (p. 480), and in the correlation problem, Section 12.7 (p. 487).

Analogous results are also available in the one-sample location or matched pairs problem, Section 12.7 (p. 487), and are mentioned there.

Theoretical development and rigorous theorems on locally most powerful rank tests may be found in Hajek and Sidak (1967, Ch. 2), Conover (1973), and Randles and Wolfe (1979, Chs. 4 and 9).

12.5.2 Pitman Asymptotic Relative Efficiency

Perhaps the most useful way to evaluate and compare rank tests is due to Pitman (1948) and further developed by Noether (1955) and others. The basic idea is that Pitman Asymptotic Relative Efficiency (ARE) is the ratio of sample sizes for two different tests to have the same power at a sequence of alternatives converging to the null hypothesis.

Let S and T be two test statistics for $H : \theta = \theta_0$ where θ_k is a sequence of alternatives converging to θ_0 as $k \to \infty$. If we can choose sample sizes N_{S_k} and N_{T_k} and critical values c_{S_k} and c_{T_k} for S and T, respectively, such that $S > c_{S_k}$ and $T > c_{T_k}$ have levels that converge to α and their powers under θ_k converge to β, $\alpha < \beta < 1$, then the Pitman asymptotic relative efficiency of S to T is given by

$$\text{ARE}(S, T) = \lim_{k\to\infty} \frac{N_{T_k}}{N_{S_k}}.$$

Note that if $\text{ARE}(S, T) > 1$, then S is preferred to T because it takes fewer observations (N_{S_k} is less than N_{T_k}) to achieve the same power. Technical conditions in the Appendix and $P(S_k > c_{S_k}) \to \beta < 1$ require that the alternatives have a specific form: for some $\delta > 0$

$$\theta_k = \theta_0 + \frac{\delta}{\sqrt{N_{S_k}}} + o\left(\frac{1}{\sqrt{N_{S_k}}}\right) \quad \text{as} \quad k \to \infty. \tag{12.24}$$

Such sequences of alternatives are called *Pitman alternatives*. Another important quantity arising from the technical details is the *efficacy* of a test statistic S,

$$\text{eff}(S) = \lim_{k\to\infty} \frac{\mu'_{S_k}(\theta_0)}{\sqrt{N_{S_k}\sigma^2_{S_k}(\theta_0)}},$$

where $\mu_{S_k}(\theta_0)$ and $\sigma_{S_k}(\theta_0)$ are the asymptotic mean of S and standard deviation of S. Thus, the efficacy of a test is the rate of change of its asymptotic mean at the null hypothesis relative to its asymptotic standard deviation (the factor $1/\sqrt{N_{S_k}}$ is introduced in the derivative because of 12.24). A powerful test in the Pitman sense is one that is able to detect changes in the parameter value near the null hypothesis. The ARE of S to T turns out to be

$$\text{ARE}(S, T) = \left\{\frac{\text{eff}(S)}{\text{eff}(T)}\right\}^2.$$

Table 12.3 ARE(W, t) for the Two-Sample Shift Model

Distribution	ARE(W, t)
Lower Bound	0.864
Normal	0.955
Uniform	1.00
Logistic	1.10
Laplace	1.50
t_6	1.16
t_3	1.90
t_1 (Cauchy)	∞
Exponential	3.00

The Pitman ARE is both a limiting ratio of sample sizes required to give the same power and the square of the ratio of the test efficacies. High efficacies lead to high ARE's.

In the Appendix we give details for finding efficacies in the one-sample problem, but here we use similar standard results on efficacies for the two-sample problem from Randles and Wolfe (1979, Chs. 5 and 9). The most important comparison is between the two-sample t test and the Wilcoxon Rank Sum test. The efficacy of the t test is

$$\mathrm{eff}(t) = \frac{\sqrt{\lambda(1-\lambda)}}{\sigma},$$

where σ is the standard deviation of the X distribution function $F(x)$ and of the Y distribution function $G(y) = F(x - \Delta)$, and $\lambda = \lim_{\min(m,n)\to\infty} m/(m+n)$. For the Wilcoxon Rank Sum statistic W we have

$$\mathrm{eff}(W) = \sqrt{12\lambda(1-\lambda)} \int_{-\infty}^{\infty} f^2(x)\, dx,$$

where f is the density of $F(x)$, and the integral is assumed to exist. Putting these efficacies together, we have that the Pitman ARE of W to t is

$$\mathrm{ARE}(W, t) = 12\sigma^2 \left\{ \int_{-\infty}^{\infty} f^2(x)\, dx \right\}^2. \tag{12.25}$$

We put ARE(W, t) into Table 12.3 for a number of distributions. Remember that ARE$(W, t) > 1$ means that the Wilcoxon Rank Sum test is preferred to the t test. The first number is the lower bound 0.864 derived by Hodges and Lehmann (1956) which shows that the Wilcoxon Rank Sum cannot do much worse than the t test for any continuous unimodal distribution. The second number 0.955 is for the normal distribution and shows that the Wilcoxon loses very little efficiency at the normal distribution where the t test is optimal. At the uniform distribution, the tests perform equivalently, and at the remaining examples in Table 12.3, the Wilcoxon is preferred.

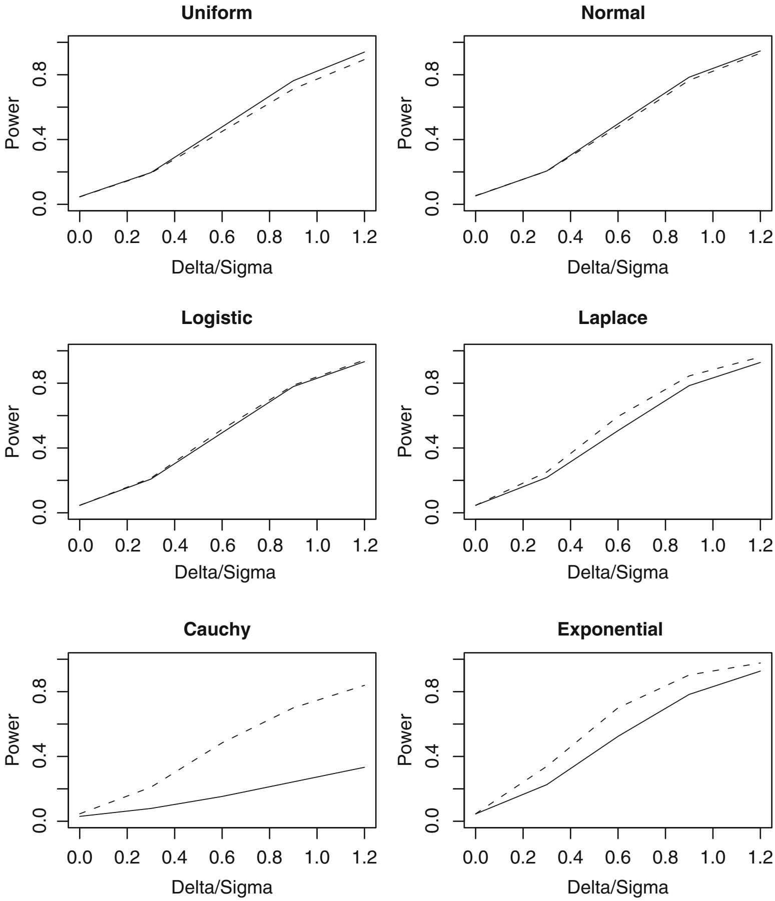

Fig. 12.3 *Power of Wilcoxon Rank Sum* $(\cdots)$ *and t* (______) *for* $m = n = 15$ *from Table 4.1.10 of Randles and Wolfe (1979)*

One might think that these ARE results are just asymptotic and may not relate to small sample results. To supplement the ARE results, in Figure 12.3 we plot power results for $m = n = 15$ taken from Table 4.1.10 of Randles and Wolfe (1979, p. 118–119). They simulated the power of the t and Wilcoxon using 1000 replications. Here we see good correspondence between small sample power and the ARE results of Table 12.3. For the normal, uniform, and logistic distributions, there is little power difference as one might expect from ARE values of .955, 1.00, and 1.10, respectively. For the Laplace, the Wilcoxon has a significant power advantage, perhaps not quite as large at the ARE$(W, t) = 1.5$ would imply. The t_1 (Cauchy) and exponential power results strongly favor the Wilcoxon and are consistent with the large ARE values.

We should mention that the Laplace distribution with density $f(x) = (1/2)\exp(-|x|)$ has been used quite a bit in the rank literature as a model for data, especially for ARE comparisons and simulation studies. But it may not be very useful as a model for real data, and ARE results for it are not as consistent with simulation results in small samples as with other densities. The optimal rank test for the Laplace uses scores $a(i) = 1$ for $i > (N+1)/2$ and 0 otherwise, and is called the two-sample median test. However, its power performance in small samples, even when simulating from the Laplace distribution, is poor. Freidlin and Gastwirth (2000) show by simulation that the Wilcoxon Rank Sum test outperforms the median test at the Laplace distribution for samples sizes $m = n$ less than or equal to 25. They recommend that the median test "be retired" from general usage, and we agree.

It turns out that in the scale problem mentioned briefly in Section 12.6.6 (p. 486), ARE values are overly optimistic when compared to small sample power results. This may reflect the fact that measuring scale (standard deviation) is an inherently harder problem that is not as well suited to rank statistics. Klotz (1962) pointed out this discrepancy between small sample power and ARE results. Fortunately, ARE results have been used mainly in location comparisons where they yield good intuition about the qualitative behavior of tests.

Another result from Randles and Wolfe (1979, p. 307) is that under suitable regularity results on the score functions, the efficacy of any linear rank test $S = \sum_{i=m+1}^{N} \phi(R_i/(N+1))$ in the two-sample shift model is given by

$$\text{eff}(S) = \sqrt{\lambda(1-\lambda)} \frac{\displaystyle\int_0^1 \phi(u)\phi(u,f)\,du}{\left[\displaystyle\int_0^1 \{\phi(u) - \overline{\phi}\}^2\,du\right]^{1/2}}, \tag{12.26}$$

where $\phi(u, f)$ is given in (12.23, p. 475). Expression (12.26) now justifies the name *optimal score function* since the efficacy in (12.26) is optimized by choosing $\phi(u) = \phi(u, f)$. This can be seen by noting that

$$\int_0^1 \phi^2(u,f)\,du = \int_{-\infty}^{\infty} \left\{\frac{f'(x)}{f(x)}\right\}^2 f(x)\,dx = I(f),$$

where $I(f)$ is the Fisher information for the model $f(x;\theta) = f(x-\theta)$. Now, noting that $\int_0^1 \phi(u,f)\,du = 0$, (12.26) can be reexpressed as

$$\text{eff}(S) = \sqrt{\lambda(1-\lambda)I(f)}\,\text{Corr}(\phi(U), \phi(U,f)), \tag{12.27}$$

where U is a uniform random variable and Corr is the correlation. Clearly, the correlation is maximized by choosing $\phi(u) = \phi(u, f)$. Moreover, it can also be shown that $\sqrt{\lambda(1-\lambda)I(f)}$ is not only the largest possible efficacy among linear rank tests but also among all α-level tests. Thus, optimal linear rank tests are asymptotically equivalent in terms of Pitman ARE to the best possible tests, say likelihood ratio or score or Wald tests for the shift model in a parametric framework. Of course, this optimality in either the rank test or the parametric test requires that the assumed family is correct.

In the next sections we consider i) the k-sample problem that is a generalization of the two-sample problem to $k > 2$ samples; ii) the correlation or regression problem; and then iii) the matched pairs or one-sample symmetry problem. The Pitman ARE analysis has to be adjusted to handle each situation, but the numbers found in Table 12.3 (p. 477) continue to hold for these situations as well. Thus Wilcoxon procedures, in other words rank methods using scores $a(i) = i$, tend to give very good results across a wide range of distributions in each of these situations.

12.6 The k-sample Problem, One-way ANOVA

The extension of the two-sample case to k samples or treatments is straightforward. Suppose that we have available k independent random samples $\{Y_{i1}, \ldots, Y_{in_i};\ i = 1, \ldots, k\}$, where in each sample the Y_{ij} $(j = 1, \ldots, n_i)$ are iid with distribution function $F_i(x)$, and $N = n_1 + \cdots + n_k$. The linear model representation is

$$Y_{ij} = \mu + \alpha_i + e_{ij}. \qquad (12.28)$$

If the errors e_{ij} all come from the same distribution, then (12.28) is an extension of the shift model for two-sample data.

For example, the following are data on the ratio of Assessed Value to Sale Price for single family dwellings ($n_1 = 27$), two-family dwellings ($n_2 = 22$), three-family dwellings ($n_3 = 17$), and four or more family dwellings ($n_4 = 14$) in Fitchburg, Massachusetts, in 1979.

1 Family			2 Family			3 Family		4 or More	
46	74	87	55	85	129	51	100	22	119
60	75	87	60	86	150	64	107	44	120
65	75	87	67	90	203	73	111	71	129
67	77	89	73	94	730	82	112	85	143
68	78	92	76	96		83	126	89	487
69	81	95	77	97		85	134	90	
70	82	95	80	98		89	140	98	
71	84	100	80	100		95	195	102	
73	85	121	82	113		100		113	

The null hypothesis of interest is of identical distribution functions,

$$H_0 : F_1(y) = F_2(y) = \cdots = F_k(y), \qquad (12.29)$$

which arises most naturally if we randomly assigned N experimental units to k treatment groups with sample sizes $n_1, n_2, \ldots, n_k$. (The above data are not of this type.) There are

$$M_N = \binom{N}{n_1 n_2 \cdots n_k} = \frac{N!}{n_1! n_2! \cdots n_k!}$$

possible assignments, which of course is the relevant number of permutations even if the data do not come from a randomized experiment. Pitman (1938) proposed the permutation approach for the ANOVA F statistic

$$F = \frac{\dfrac{1}{k-1}\displaystyle\sum_{i=1}^{k} n_i (\overline{Y}_{i.} - \overline{Y}_{..})^2}{\dfrac{1}{N-k}\displaystyle\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2}, \tag{12.30}$$

where $\overline{Y}_{i.} = n_i^{-1} \sum_{j=1}^{n_i} Y_{ij}$, and $\overline{Y}_{..} = N^{-1} \sum_{i=1}^{k} n_i \overline{Y}_{i}$. The number of permutations M_N gets large very fast. For example, with $k = 3, N = 15, n_1 = n_2 = n_3 = 5$, we get $M_N = \binom{15}{5\,5\,5} = 756,756$. Thus Monte Carlo or asymptotic approximations are more important than in the two-sample case. For the above housing data, the ANOVA F in (12.30) is $F = 1.24$ with p-value = .30 from the $F(3, 75)$ distribution. The exact permutation p-value is obtained by computing F for each of the 1.9×10^{44} distinct allocations of $\{Y_{i1}, \ldots, Y_{in_i}; i = 1, \ldots, 4\}$ to samples of size $n_1 = 27$, $n_2 = 22$, $n_3 = 17$, and $n_4 = 14$, and finding the proportion of these greater to or equal to $F = 1.24$. A Monte Carlo estimate of the exact permutation p-value is .267 based on 100,000 resamples with standard error = .0014. Because the housing ratios are quite skewed with a number of large observations, it is not surprising that F is small. Now we turn to rank methods that naturally limit the effect of outliers.

12.6.1 Rank Methods for the k-Sample Location Problem

Kruskal and Wallis (1952) proposed the rank extension of the Wilcoxon Rank Sum statistic to the k-sample situation. The rank approach is to put all N observations together and rank them; let R_{ij} be the rank of Y_{ij} in the combined sample. Further define the sample sums

$$S_i = \sum_{j=1}^{n_i} a(R_{ij}),$$

where the scores $a(i)$ could be of any form for permutational analysis, but for asymptotic results we assume $a(i) = \phi(i/(N+1))$ and ϕ is a scores generating function as in Theorem 12.3 (p. 465). The Kruskal-Wallis statistic uses $a(i) = i$ or equivalently $a(i) = i/(N+1)$. Note that S_i is just a two-sample linear rank statistic for comparing the ith population to all the others combined. The general linear rank statistic form for comparing the k populations is then

$$Q = \sum_{i=1}^{k} \frac{1}{s_a^2 n_i}(S_i - n_i\overline{a})^2 = \sum_{i=1}^{k} \left(\frac{N-n_i}{N}\right) \frac{(S_i - \mathrm{E}S_i)^2}{\mathrm{Var}(S_i)}, \tag{12.31}$$

where $s_a^2 = (N-1)^{-1}\sum_{i=1}^{N}\{a(i) - \overline{a}\}^2$, $\overline{a} = \sum_{i=1}^{N} a(i)$, and $\mathrm{Var}(S_i)$ is given by (12.4, p. 459) with the constants c_i in that expression equal to 1 for n_i of them and 0 otherwise. The reason for giving the second form in (12.31) is that it is then clear that $\mathrm{E}(Q) = k - 1$ under the null hypothesis of equal populations. The Kruskal-Wallis statistic that allows for ties is explicitly given by

$$H = \frac{(N-1)\left\{\sum_{i=1}^{k} n_i \left(\overline{R}_{i.} - \frac{N+1}{2}\right)^2\right\}}{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} R_{ij}^2\right) - N(N+1)^2/4},$$

where $\overline{R}_{i.} = n_i^{-1}\sum_{j=1}^{n_i} R_{ij}$. If there are no ties in the data, then

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} R_{ij}^2 = N(N+1)(2N+1)/6,$$

and H reduces to the more familiar form

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left(\overline{R}_{i.} - \frac{N+1}{2}\right)^2.$$

Under the null hypothesis (12.29, p. 480), standard asymptotic theory similar to Theorem 12.3 (p. 465) yields that $Q \xrightarrow{d} \chi_{k-1}^2$ as $\min\{n_1, \ldots, n_k\} \to \infty$. The χ_{k-1}^2 approximation is not very good in small samples, but fortunately the F statistic on the scores $a(R_{ij})$ is a monotone function of Q,

$$F_{\mathrm{R}} = \left(\frac{N-k}{k-1}\right)\left(\frac{Q}{N-1-Q}\right),$$

and using $F(k-1, N-k)$ as a reference distribution or the Box-Andersen adjusted $F(d(k-1), d(N-k))$ distribution yields excellent results. For the housing data above, $H = 9.8856$ with p-value = 0.020 from the χ_3^2 distribution. $F_R = 3.6283$

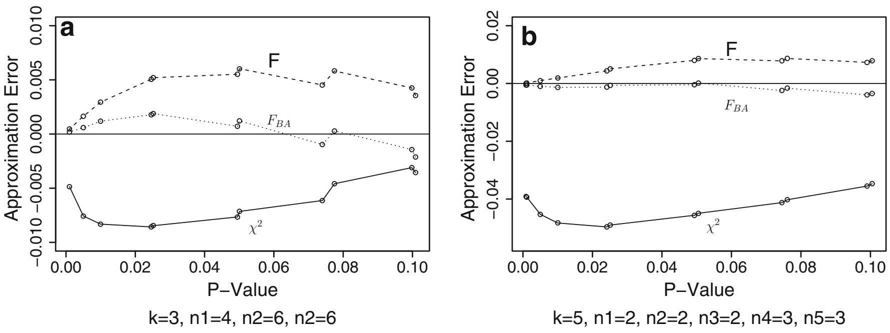

Fig. 12.4 (Exact P-Values − Approximate P-Values) versus Exact P-Values for Kruskal-Wallis Statistic. $F = F(k-1, N-k)$, $F_{BA} = F(d(k-1), d(N-k))$, and $\chi^2 = \chi^2_{k-1}$

with p-value 0.017 from the $F(3, 75)$ distribution. The Box-Andersen d=0.9876, and so the adjustment is very minor, only in the fourth decimal place. A Monte Carlo approximation to the exact p-value is .017 based on 100,000 samples with standard error .0004. So here the F distribution approximation is right on target to 3 decimals, but the χ^2 approximation is not bad due to the fairly large samples.

In Figure 12.4 we look at much smaller sample sizes for $k = 3$ and $k = 5$. Figure 12.4 shows the difference between the exact permutation p-value and each approximation versus the exact p-value for the Kruskal-Wallis statistic. Note that the left panel is more expanded in the vertical scale than the right panel and actually has less error. Nevertheless, the Box-Andersen approximation is the best in both plots and is generally very good for $k > 2$. The χ^2_{k-1} approximation gets more conservative as k gets larger. This can be explained by the following large-k asymptotic results.

12.6.2 Large-k Asymptotics for the ANOVA F Statistic

Brownie and Boos (1994) show under the null hypothesis of equal populations that

$$\sqrt{k}(F_{\mathrm{R}} - 1) \xrightarrow{d} \mathrm{N}\left(0, \frac{2n}{n-1}\right), \qquad (12.32)$$

for equal sample sizes $n_1 = n_2 = \cdots = n_k = n$ and $k \to \infty$ with n fixed. Note that the usual result with $n \to \infty$ and k fixed is $(k-1)F_{\mathrm{R}} \xrightarrow{d} \chi^2_{k-1}$, similar to the result for Q. The "large k" asymptotic result (12.32) implies that

$$\sqrt{k}\left(\frac{Q}{k-1} - 1\right) \xrightarrow{d} \mathrm{N}\left(0, \frac{2(n-1)}{n}\right), \qquad (12.33)$$

as $k \to \infty$ with n fixed, using

$$Q = \frac{(N-1)F_R}{(N-k)/(k-1) + F_R} \tag{12.34}$$

(see Problem 12.17, p. 527). Note that comparing Q to a χ^2_{k-1} is asymptotically ($k \to \infty$) like comparing $Q/(k-1)$ to a $N\{1, 2/(k-1)\}$ because a χ^2_{k-1} random variable obeys the Central Limit Theorem (it is a sum of χ^2_1 random variables). However, (12.33) says that $Q/(k-1)$ should be compared to a $N\{1, 2(n-1)/(kn)\}$ distribution. Because $2(n-1)/(kn) < 2/(k-1)$, using the χ^2_{k-1} distribution with Q results in conservative true levels. For example, if $k = 5$ and $n = 5$, then the large sample 95th percentile from $N\{1, 2/(k-1)\}$ is $1 + (2/4)^{1/2}1.645 = 2.16$, and the approximate true level of a nominal $\alpha = .05$ test is

$$P(Q \geq \chi^2_4(.95)) \approx P(1 + (8/25)^{1/2}Z \geq 2.16) = P(Z \geq 2.05) = .02.$$

In contrast, use of F_R with an $F(k-1, N-k)$ reference distribution is supported by (12.32) under $k \to \infty$ and by the usual asymptotics $(k-1)F_R \xrightarrow{d} \chi^2_{k-1}$ when $n \to \infty$ with k fixed. We leave those details for Problem 12.18 (p. 527). Thus, it is not surprising that the F approximations in Figure 12.4 are much better than the χ^2_{k-1} ones.

12.6.3 Comparison of Approximate *P*-Values – Data on Cadmium in Rat Diet

Nation et al. (1984) studied the effect of diets containing cadmium (Cd) on the neurobehavior of adult rats. The data consists of the number of platform descents during a passive-avoidance training scheme for 27 rats randomly assigned to three groups:

										$\overline{Y}$	s_{n-1}
Control:	82	80	77	75	72	68	59	47	42	67	14
Cd1:	86	66	60	51	44	41	38	29	10	47	22
Cd5:	81	67	38	36	32	29	20	17	14	37	23

The control group had no Cd in the diet, and Cd1 and Cd5 refer to daily diets containing 1 milligram and 5 milligrams, respectively, of Cd per kilogram of body weight. The usual one-way ANOVA $F = 5.10$, and the permutation p-value F statistic is $\widehat{p} = 0.016$ based on 100,000 random permutations. The $F(2, 24)$ distribution gives p-value = .014, and the Box-Andersen correction factor is $d = .954$ leading to p-value = .016. The Kruskal-Wallis rank statistic is $Q = 8.18$ with permutation p-value $\widehat{p} = .012$ based on 100,000 random permutations. The χ^2_2

approximation gives p-value = .017. The associated F statistic is $F_{\mathrm{R}} = 5.51$ with p-value = .011. The Box-Andersen correction factor is $d = 1 - 1.2/24 = .95$ leading to p-value = .012. A summary is as follows:

Statistic	Method	P-value
F	Monte Carlo (B=100,000)	0.016
	$F(2, 24)$	0.014
	Box-Andersen	0.016
KW	Monte Carlo (B=100,000)	0.012
	χ^2_2	0.017
	$F(2, 24)$	0.011
	Box-Andersen	0.012

As expected the F approximations give excellent p-values.

12.6.4 Other Types of Alternative Hypotheses

The k-sample F statistic and Kruskal-Wallis statistic are used to compare the centers or locations of the k populations. Other statistics could be used for that purpose, perhaps ones more suited to long-tailed or skewed populations. The logrank or Savage scores, for example, are asymptotically optimal for detecting shifts in the scale parameter of exponential populations (or the shift parameter of extreme value distributions).

Other types of alternatives may also be of interest. For example, there may be an implied order in the populations, say increasing doses, and there may be interest in trends in location. There might also be interest in comparing the spread of the populations or even the skewness.

These latter alternatives present a problem to permutation and rank methods because the null hypothesis of interest may not be the one of identical populations. For comparing spread, the usual null hypothesis of interest would be equal spread rather than identical populations. In such a situation, use of the permutation approach would require subtraction of unknown location parameters. We first discuss ordered alternatives in location.

12.6.5 Ordered Means or Location Parameters

Recall Section 3.6.1a (p. 151) where we discussed likelihood-based methods for ordered alternatives. Here we discuss permutation methods with simple statistics in the context of a Phase I toxicology study where there seems to be trends in both the means and variances with dose:

Dose					$\overline{Y}$	s_{n-1}
0	1.44	1.63	1.40	1.59	1.52	0.11
1	1.27	1.50	1.45	1.57	1.45	0.13
2	1.26	1.07	1.38	1.75	1.37	0.29
3	1.04	1.14	1.46	1.06	1.18	0.19
4	1.37	0.79	1.32	1.42	1.23	0.29

The F statistic for comparing means is $F = 1.77$, and the usual $F(4, 16)$ distribution and the Box-Andersen approximation give p-value = 0.19. Similarly, a Monte Carlo estimated p-value based on 10,000 random permutations gives $\widehat{p} = 0.19$. The Kruskal-Wallis statistic is $H = 6.73$ with χ^2_4 p-value = 0.15. The F approximation from $F_{\text{R}} = 2.06$ and the Box-Andersen approximation both give p-value = 0.14. A Monte Carlo estimated p-value based on 10,000 random permutations gives $\widehat{p} = 0.14$. So the global comparison of location is not significant at usual levels.

Suppose that we consider H_0 : identical populations versus H_a : means are decreasing. The permutation approach with $M_N = \binom{20}{44444}$ permutations may be used with the t statistic from a regression of the observations on dose or equivalently Pearson's correlation coefficient (see also the next section). Pearson's correlation coefficient is $r = -0.53$ with Monte Carlo estimated p-value $\widehat{p} = 0.007$ based on 10,000 random permutations. Spearman's correlation coefficient is –0.56 with $\widehat{p} = 0.005$. Another statistic that could have been used is the likelihood ratio statistic for decreasing means assuming the data are normally distributed (see Section 3.6.1a, p. 151). In addition to Spearman's correlation coefficient, the standard rank-based statistic is the Jonckheere-Terpstra statistic based on summing pairwise Wilcoxon Rank Sum statistics in increasing order, $\sum_{i<j} W_{ij}$, where W_{ij} is the Wilcoxon Rank Sum for comparing dose group i with dose group j (see Lehmann 1975, p. 233). Its value here is –2.458 with exact permutation p-value = 0.0069. So we can be pretty confident that there is a downward trend in means or other location measures.

12.6.6 Scale or Variance Comparisons

Motivated by the apparent increase in variances for the dose-response data above, we now discuss hypotheses about variances or scale parameters. Unfortunately, there is a philosophical dilemma for using permutation procedures here. Usually, the typical set of hypotheses when testing for unequal variances is for a semiparametric model, $P(Y_{ij} \leq y) = F_0((y - \mu_i)/\sigma_i)$, $j = 1, \ldots, n_i$; $i = 1, \ldots, k$, where F_0 is an unknown distribution function. Note that if $F_0(x)$ has mean 0 and variance 1, then μ_i is the ith population mean, and σ_i^2 is the ith population variance. In any case, under this semiparametric model, the ith standard deviation is $c\sigma_i$ for some constant c, and we can always refer to σ_i as a scale parameter. The hypotheses for increasing scale are then $H_0 : \sigma_1 = \cdots = \sigma_k$ versus $H_a : \sigma_1 \leq \cdots \leq \sigma_k$ with at least one inequality. The reason for this hypothesis formulation is that we often know that the means are different; therefore it makes little sense to assume identical populations when testing for variance differences. Basically, we usually want to test for variance differences in the presence of location differences.

Unfortunately, the permutation argument requires that the null hypothesis be one of identical populations. It makes intuitive sense to center the data first by subtracting means, but these residuals $Y_{ij} - \overline{Y}_i$ no longer satisfy exchangeability required for using Theorem 12.1 (p. 457). The permutation distribution is correct asymptotically, but the exact level-α property no longer holds. An overview of the scale testing problem is given in Boos and Brownie (2004). The best method that has emerged for comparing scales is to use t or F statistics on the data Y_{ij} replaced by $|Y_{ij} - M_i|$, where M_i is the ith sample median.

One way to avoid the centering problem for the dose-response data is to reduce the data to the sample standard deviations (or some other scale estimator) and then calculate an appropriate statistic for the $5! = 120$ permutations possible. For the correlation between dose and standard deviation we get $r = 0.79$ and p-value $= 7/120 = .058$. If we use the likelihood ratio test for increasing variances for normal distributions, we get p-value = 5/120=.042. There is a loss of information when the number of permutations get reduced so much, from $M_N = \binom{20}{44444}$ to $M_N = 120$; perhaps the loss of information is just a discreteness problem caused by having too few permutations. This can be seen more clearly by calculating the exact permutation test on the data reduced to the five means; the correlation is higher than when using all the data, but the p-value = 2/120 = .017 is much larger than the .007 value we obtained previously with the whole data set.

We note that the use of rank statistics for scale comparisons has not been very successful. The subtraction of means or medians ruins the permutation argument as mentioned above. However, rank statistics for scale based on centered data are asymptotically distribution free if the samples are symmetrically distributed. The larger problem is that rank tests for scale tend to have low power in small samples. Although rank tests for location perform well in small samples and are consistent with asymptotic relative efficiency comparisons, the opposite is true for rank tests for scale. The latter statistics are not as powerful in small samples as would be expected from asymptotic relative efficiency calculations.

12.7 Testing Independence and Regression Relationships

Regression methods are among the most important tools of statistics. Unfortunately, permutation methods can really be applied in only the simplest setting of (X, Y) pairs; that is, correlation or simple regression (not necessarily linear). Here we discuss that simple situation and mention at the end of the section why permutation methods cannot handle the more interesting case of multiple explanatory variables.

Suppose that we have iid random pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ and permute each coordinate independently to get $n!$ different pairings. In reality, we need only permute one of the coordinates to obtain all the different pairings. For example, suppose that $n = 3$ with pairs $(1, 2.5), (2, 3.7), (3, 6.4)$. Then the 6 possible permutations are

1	2	3	4	5	6
(1,2.5)	(1,3.7)	(1,6.4)	(1,2.5)	(1,3.7)	(1,6.4)
(2,3.7)	(2,2.5)	(2,3.7)	(2,6.4)	(2,6.4)	(2,2.5)
(3,6.4)	(3,6.4)	(3,2.5)	(3,3.7)	(3,2.5)	(3,3.7)

Pitman (1937b) suggested that a test for independence of X and Y based on the sample correlation

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\left[\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2\right]^{1/2}}$$

use this permutation distribution for critical values. A permutationally equivalent statistic is the least squares slope estimate $\widehat{\beta} = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) / \sum_{i=1}^{n}(X_i - \overline{X})^2$. Other popular measures that could be used to test independence are Kendall's rank correlation and Spearman's rank correlation. Spearman's estimated correlation coefficient r_{S} is simply to replace X_i by its rank among $X_1, \ldots, X_n$ and Y_i by its rank among $Y_1, \ldots, Y_n$, and compute the Pearson correlation r between these pairs of ranks. It is important to keep in mind that the null hypothesis is independence of X and Y and not zero correlation. Independence is needed for the $n!$ different pairings to have the same distribution and thus for Theorem 12.1 (p. 457) to apply.

Typical approximations to the permutation distribution of r (and similarly of r_{S}) are to compare $(n-1)^{1/2} r$ to a standard normal distribution or $(n-2)^{1/2} r/(1-r^2)^{1/2}$ to a $t(n-2)$ distribution. Pitman (1937b) gave the first two permutation moments of r^2, $\text{E}_{\text{P}}(r^2) = 1/(n-1)$, and

$$\text{E}_{\text{P}}(r^4) = \frac{3}{(n-1)(n+1)} + \frac{(n-2)(n-3)}{n(n+1)(n-1)^3} \left\{ \frac{k_4(X)}{k_2(X)^2} \right\} \left\{ \frac{k_4(Y)}{k_2(Y)^2} \right\},$$

where the sample cumulants k_2 and k_4 were given in (12.21, p. 469) and (12.22, p. 469), respectively. Note that these moments are straightforward from the results in Section 12.4.2 (p. 458) since the numerator of r has the form (12.3, p. 458) of a linear statistic, and the denominator is constant over permutations. If the pairs are iid with a bivariate normal distribution, then r^2 has a beta$(1/2, n/2-1)$ distribution with $\text{E}(r^2) = 1/(n-1)$ and $\text{E}(r^4) = 3/(n-1)(n+1)$. Because the permutation moments and normal theory moments are so close, Pitman (1937b) suggested using the beta approximation, which is equivalent to comparing $(n-2)r^2/(1-r^2)$ to an $F(1, n-2)$ distribution. Box and Watson (1962) generalized these results to the full p regressor case for the test that all regressors are independent of Y. They derived the adjusted F approximation (see Box and Watson 1962, p. 100), which for the $p = 1$ case here is to compare $(n-2)r^2/(1-r^2)$ to an $F(d, d(n-2))$ distribution, where

$$\frac{1}{d} = 1 + \frac{(n+1)\alpha_1}{n-1-2\alpha_1}, \quad \alpha_1 = \frac{n-3}{2n(n-1)} \left\{ \frac{k_4(X)}{k_2(X)^2} \right\} \left\{ \frac{k_4(Y)}{k_2(Y)^2} \right\}.$$

In large samples, $d \approx 1 + \{\text{Kurt}(X) - 3\}\{\text{Kurt}(Y) - 3\}/2n$, revealing a double Type I error robustness to nonnormality: if either X or Y is approximately normally distributed, then the usual F approximation is very good. To numerically illustrate, recall $r = -.53$ from the dose-response data (p. 486) where the Monte Carlo estimated one-sided p-value was $\widehat{p} = .007$. Taking half of the $F(1, 18)$ p-value approximation for $18r^2/(1-r^2) = 7.03$, we get p-value = .008. Similarly, for Spearman's $r_S = -.56$ we obtained previously $\widehat{p} = .005$. Using one half of the $F(1, 18)$ p-value for $18r_S^2/(1-r_S^2) = 8.22$ yields p-value = .005.

Now let us move to the more complicated situation of the linear model,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i, \quad i + 1, \ldots, n,$$

where we assume $e_1, \ldots, e_n$ are iid from some distribution and independent of all the X_{ij}. As mentioned above, permuting the Y's under the assumption $H_0 : \beta_1 = \beta_2 = 0$ yields a suitable permutation distribution for testing independence of Y and (X_1, X_2). Unfortunately, we are usually much more interested in testing $H_0 : \beta_2 = 0$ with β_0 and β_1 unrestricted. Without knowledge of β_1, however, an exact permutation procedure for $H_0 : \beta_2 = 0$ is not possible. (Actually, it is possible to take the maximum over permutation p-values for each value of β_1 in a confidence interval under H_0 as described in Berger and Boos (1994), but the loss in power is typically not worth the gain in exactness.) Anderson and Robinson (2001) review a number of different proposals that use residuals from first fitting the reduced model, and show that they are asymptotically correct but do not satisfy the assumptions of Theorem 12.1 (p. 457). Fortunately, standard linear model and rank-based linear model testing procedures have good Type I error robustness properties in general. The rank-based linear model methods given in Ch. 5 of Hettmansperger (1984) have good Type II error robustness properties as well. Similarly, the M-estimation regression methods discussed in Ch. 5 also have good robustness properties.

We conclude this section with an example that illustrates how easy it is to use Monte Carlo approximation in an autocorrelation setting.

Example 12.1 (Raleigh snowfall). Is the total snowfall in one year independent of the total snowfall in other years? The left panel of Figure 12.5 plots Raleigh, NC, annual snowfall for 1962–1991 versus year. The right panel plots each year's

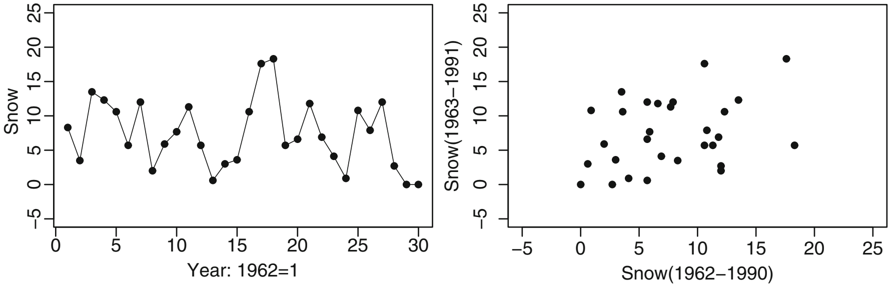

Fig. 12.5 Annual snowfall in Raleigh, NC, 1962–1991 (left panel) and annual snowfall versus annual snowfall of previous year (right panel)

snowfall versus the previous year's snowfall. The sample correlation from the right panel is $r = .32$. Does that suggest nonzero autocorrelation? The null hypothesis for a permutation approach is that the sequence of yearly snowfalls is iid or at least exchangeable. Below we give R code for sampling B permutations from the set of 30! possible permutations, computing the lag-1 sample correlation for each, and estimating the one-sided $p$-value for a positive autocorrelation. Using $B = 10,000$, we get $\widehat{p} = .027$ with standard error .0016. Thus there is good evidence of a positive autocorrelation. The main point here is to illustrate how easy it is to carry out the permutation test.

```
r.auto<-function(x){
     n<-length(x)
     cor(x[1:(n-1)],x[2:n])
}
perm1<-function(b, x, stat, ...){
  # Gives est. permutation $p$-value for vector x.
  # Assumes test rejects for large values of stat.
        call <- match.call()
        n <- length(x)
        t0 <- stat(x)
        res <- numeric(b)
        for(i in 1:b) {
                perm.xx <- sample(x)
                res[i] <- stat(perm.xx)
        }
        pvalue <- sum(res >= t0)/b
        se<-sqrt(pvalue*(1-pvalue)/b)
        return(list(call=call,results=data.frame
          (nperm=b, stat0=round(t0,4),pvalue=pvalue,
          se=round(se,5))))
}
> set.seed(2458)
> perm1(10000,raleigh.snow$snow,r.auto)
  nperm  stat0 pvalue      se
1 10000 0.3245 0.0269 0.00162
```

♦

12.8 One-Sample Test for Symmetry about θ_0 or Matched Pairs Problem

Fisher (1935) introduced the permutation approach for the matched-pairs problem in a discussion of Darwin's data on self-fertilized and cross-fertilized plants. There were 15 pairs of plants, and the differences

$$49, -67, 8, 16, 6, 23, 28, 41, 14, 29, 56, 24, 75, 60, -48$$

have mean $\overline{D} = 20.933$, $s = 37.744$, and $t = 2.148$ for testing $H_0 : \mu_{\mathrm{D}} = 0$ versus $H_a : \mu_{\mathrm{D}} \neq 0$, where μ_{D} is the population mean difference. The two-sided p-value is .0497 from the t table with 14 degrees of freedom. Alternatively, consider Fisher's permutation argument. There were 2^{15} possible random assignments of types of seeds to the 15 blocks of size 2. Thus, Fisher considered all 2^{15} sums $\sum_{i=1}^{15} D_i$, where D_i is the ith difference, and found only 835+28 = 863 which are greater than or equal to the observed sum = 314. The two-sided p-value is (2)(863)/32,768 = .0527 (by symmetry there are 863 sums ≤ -314). Note that $t = \sqrt{n}\overline{D}/s$ is permutationally equivalent to $\sum_{i=1}^{15} D_i$ because t is a monotonic function of $\sum_{i=1}^{15} D_i$ that depends on $\sum_{i=1}^{15} D_i^2$, which is constant over all 2^{15} permutations.

Let us consider the theory behind Fisher's approach. The population null model is that the differences $D_1, \ldots, D_n$ are independent, each with a symmetric distribution about some θ_0; often $\theta_0 = 0$. The distributions do not need to be the same, merely symmetric about θ_0. Thus

$$H_0 : D_i - \theta_0 \overset{d}{=} \theta_0 - D_i, \quad i = 1, \ldots, n. \tag{12.35}$$

The group of transformations to be used with Theorem 12.1 (p. 457) is the set of 2^n sign changes applied to the data with θ_0 subtracted. For notational simplicity, let $D_{i0} = D_i - \theta_0$, $i = 1, \ldots, n$. Then, for example, if $n = 4$, one such transformation is $(-, +, +, -)$. It would transform

$$(D_{10}, D_{20}, D_{30}, D_{40}) \tag{12.36}$$

into

$$(-D_{10}, D_{20}, D_{30}, -D_{40}). \tag{12.37}$$

Because of (12.35) and independence, all 2^n transformations of the original data have the same distribution. That is, under (12.35) and independence, the joint distribution of (12.36) is the same as (12.37), etc. Thus, the conditions of Theorem 12.1 (p. 457) apply with the group of sign changes, and Fisher's original method is a valid permutation approach.

12.8.1 Moments and Normal Approximation

Now let us abstract the above situation slightly in order to compute moments and approximations. Suppose that $d_1, \ldots, d_n$ is a sequence of real constants, playing the role of the observed $D_i - \theta_0$ above. Let $c_1, \ldots, c_n$ be iid random variables with $P(c_i = 1) = P(c_i = -1) = 1/2$; these play the role of making the sign changes. Now consider the linear statistic $T = \sum_{i=1}^n c_i d_i$. Note that the c_i are symmetrically distributed around 0 so that all odd moments of c_i are 0 and all even moments equal to 1. Then T is also symmetrically distributed about 0 with odd moments 0 and $\mathrm{E}(T^2) = \mathrm{Var}(T) = \sum_{i=1}^n d_i^2$ and $\mathrm{E}(T^4) = 3(\sum_{i=1}^n d_i^2)^2 - 2\sum_{i=1}^n d_i^4$. Now we give a Central Limit Theorem for T. A more general version and proof are given in Hettmansperger (1984, p. 302–303).

Theorem 12.4. . *Suppose that $d_1, \ldots, d_n$ and $c_1, \ldots, c_n$ are defined as above and*

$$\frac{1}{n}\sum_{i=1}^n d_i^2 \longrightarrow \sigma^2 < \infty \qquad as \;\; n \to \infty.$$

Then

$$\frac{T}{\sqrt{Var(T)}} = \frac{\sum_{i=1}^n c_i d_i}{\left(\sum_{i=1}^n d_i^2\right)^{1/2}} \xrightarrow{d} N(0,1) \qquad as \;\; n \to \infty.$$

Now we apply this theorem to the permutation distribution of $\sum_{i=1}^n D_i$ when sampling from a population.

Theorem 12.5. *Suppose that $D_1, \ldots, D_n$ are iid random variables satisfying (12.35) and with variance $\sigma^2 < \infty$. Then the permutation distribution function of $\sum_{i=1}^n (D_i - \theta_0)$ under the group of sign changes satisfies*

$$P^*\left\{\sum_{i=1}^n (D_i - \theta_0)/\sqrt{n}\sigma\right\} \xrightarrow{wp1} N(0,1) \qquad as \;\; n \to \infty.$$

We have used the notation P^* to emphasize that the probability is taken with respect to the permutation distribution holding $D_1, \ldots, D_n$ fixed. An alternative statement of the result is that the permutation distribution of $\sum_{i=1}^{n}(D_i - \theta_0)/\sqrt{n}\sigma$ converges in distribution to a standard normal distribution with probability 1. Note also that we could just as well have put $\{\sum_{i=1}^{n}(D_i - \theta_0)^2\}^{1/2}$ in place of $\sqrt{n}\sigma$ in the conclusion, giving

$$\frac{\sum_{i=1}^{n}(D_i - \theta_0)}{\left\{\sum_{i=1}^{n}(D_i - \theta_0)^2\right\}^{1/2}} \xrightarrow{d^*} \mathrm{N}(0,1) \qquad \text{as } n \to \infty \quad wp1. \tag{12.38}$$

The result follows from Theorem 12.4 because for each infinite sequence $D_1(\omega)$, $D_2(\omega), \ldots$ where $\omega \in \Omega$ with $P(\Omega) = 1$,

$$\frac{1}{n}\sum_{i=1}^{n}(D_i(\omega) - \theta_0)^2 \longrightarrow \sigma^2 \qquad \text{as } n \to \infty$$

by the Strong Law of Large Numbers. For each of these sequences, Theorem 12.4 holds, and thus the convergence in distribution holds with probability 1.

12.8.2 Box-Andersen Approximation

The Box-Andersen adjusted F approximation to the permutation distribution of $\sum_{i=1}^{n}(D_i - \theta_0)$ uses the $beta$ version of $t^2 = n(\overline{D} - \theta_0)^2/s^2$,

$$b(t^2) = \frac{t^2}{n - 1 + t^2} = \frac{n(\overline{D} - \theta_0)^2}{\sum_{i=1}^{n}(D_i - \theta_0)^2}.$$

Under an iid normal distribution assumption for $D_1, \ldots, D_n$, $b(t^2)$ has a $beta(1/2, (n-1)/2)$ distribution with mean $1/n$ and variance $2(n-1)/\{n^2(n+2)\}$. Using the results in the previous section for $T = \sum_{i=1}^{n} c_i d_i$, where $d_i = (D_i - \theta_0)/n$, the permutation moments of $b(t^2)$ are $\mathrm{E_P}\{b(t^2)\} = 1/n$ and

$$\mathrm{Var_P}\{b(t^2)\} = \frac{2(n-1)}{n^2(n+2)}\left(1 - \frac{f_2 - 3}{n-1}\right), \tag{12.39}$$

where $f_2 = (n+2)\sum_{i=1}^{n}(D_i - \theta_0)^4/\{\sum_{i=1}^{n}(D_i - \theta_0)^2\}^2$. Equating the permutation moments to those of a $beta(d/2, d(n-1)/2)$ distribution leads to

$$d = 1 + \frac{f_2 - 3}{n\{1 - f_2/(n+2)\}}. \tag{12.40}$$

In the above derivation we have followed the notation in Box and Andersen (1955, p. 9), but their W is $1 - b(t^2)$, and we relabeled their b_2 as f_2. Note that f_2 is close to the sample kurtosis of the $D_i - \theta_0$, and thus $d \approx 1 + \{\text{Kurt}(D) - 3\}/n$.

For the Darwin data, $d = .94$ and the F adjusted two-sided p-value is .053. Recall from previous analysis that the exact two-sided permutation p-value is .0527. The normal approximation here is $Z = 1.9282$ with two-sided p-value$= .054$. Thus, the normal approximation is surprisingly good here, better than the $F = t^2$ approximation that Fisher gave (.0497), but the Box-Andersen adjustment has made the F approximation slightly better than the normal approximation.

12.8.3 Signed Rank Methods

Now we turn to signed rank methods. Here again for simplicity we use the notation D_{i0} for $D_i - \theta_0$. Let R_i be the rank of $|D_{i0}|$ among $|D_{10}|, \ldots, |D_{n0}|$. Let the sign function be defined by $\text{sign}(x) = I(x > 0) - I(x < 0)$ if x is nonzero and $\text{sign}(0) = 0$. Then the signed rank of D_{i0} is $\text{sign}(D_{i0})R_i$ although some authors use $I(D_{i0} > 0)R_i$ as the definition of the signed rank. We illustrate with a simple data set from Wilcoxon (1945) on the difference between wheat yields in two treatments in 8 blocks:

D_{i0}	58	32	30	5	−7	6	11	10
R_i	8	7	6	1	3	2	5	4
$sign(D_{i0})R_i$	8	7	6	1	−3	2	5	4
$I(D_{i0} > 0)R_i$	8	7	6	1	0	2	5	4

Then define $W^+ = \sum_{i=1}^n I(D_{i0} > 0)R_i$, $W^- = \sum_{i=1}^n I(D_{i0} < 0)R_i$ and $W = \sum_{i=1}^n \text{sign}(D_{i0})R_i$. As long as there are no ties in the data, then all three of these are equivalent and $W = W^+ - W^-$. For the above sample we have $W^+ = 33$, $W^- = 3$, and $W = 30$. It is perhaps more standard to call W^+ the Wilcoxon Signed Rank statistic. Under (12.35, p. 491) and continuity of the data (implying no ties with probability 1), the basic facts are that:

1. $\text{sign}(D_{10}), \ldots, \text{sign}(D_{n0})$ and $I(D_{10} > 0), \ldots, I(D_{n0} > 0)$ are independent of $|D_{10}|, \ldots, |D_{n0}|$ and thus also independent of $R_1, \ldots, R_n$;
2. $W^+ \overset{d}{=} W^- \overset{d}{=} \sum_{i=1}^n I(D_{i0} > 0)i$, and $I(D_{10} > 0), \ldots, I(D_{n0} > 0)$ are independent Bernoulli(1/2) random variables;
3. $W \overset{d}{=} \sum_{i=1}^n \text{sign}(D_{i0})i$, and $\text{sign}(D_{10}), \ldots \text{sign}(D_{n0})$ are iid with $P(\text{sign}(D_{i0}) = 1) = 1/2$;
4.

$$\text{E}(W^+) = \frac{1}{2}\sum_{i=1}^n i = \frac{n(n+1)}{4}, \quad \text{Var}(W^+) = \frac{1}{4}\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{24};$$

5.
$$\mathrm{E}(W) = 0, \quad \mathrm{Var}(W) = \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}.$$

For the simple example above with $n = 8$, we have $\mathrm{E}(W^+) = (8)(9)/4 = 18$ and $\mathrm{Var}(W^+) = (8)(9)(17)/24 = 51$ leading to the standardized value $(33 - 18)/\sqrt{51} = 2.1$, which is clearly the same for W^- and W as well. From a normal table, we get the right-tailed p-value .018, whereas the exact permutation p-value for the signed rank statistics is $5/256 = .01953$.

Although the Wilcoxon Signed Rank is by far the most important of the signed rank procedures, the general signed rank procedures are $T^+ = \sum_{i=1}^n I(D_{i0} > 0)a(R_i)$, $T^- = \sum_{i=1}^n I(D_{i0} < 0)a(R_i)$, and

$$T = \sum_{i=1}^{n} \mathrm{sign}(D_{i0})a(R_i), \tag{12.41}$$

where the scores $a(i)$ could be of any form. The analogues of the above properties for W hold for the general signed rank statistics. In particular $T \overset{d}{=} \sum_{i=1}^n \mathrm{sign}(D_{i0})a(i)$ simplifies the distribution and moment calculations in the case of no ties. In the case of ties, the permutation variance of T, given the midranks $R_1, \ldots, R_n$, is $\sum_{i=1}^n \{a(R_i)\}^2$. Thus, for the normal approximation, it is simplest to use the form

$$Z = \sum_{i=1}^{n} \mathrm{sign}(D_{i0})a(R_i) \Big/ \left[\sum_{i=1}^{n} \{a(R_i)\}^2\right]^{1/2}, \tag{12.42}$$

that automatically adjusts for ties (see Section 12.8.6, p. 497, for a discussion of ties).

The most well-known score functions are $a(i) = i$ for the Wilcoxon, the quantile normal scores $a(i) = \Phi^{-1}(1/2 + i/[2(n+1)])$, and the sign test $a(i) = 1$. These are asymptotically optimal for shifts in the center of symmetry D_0 of the logistic distribution, the normal distribution, and the Laplace distribution, respectively. For asymptotic analysis we assume $a(i) = \phi^+(i/(n+1))$, where $\phi^+(u)$ is nonnegative and nonincreasing and $\int_0^1 [\phi^+(u)]^2 du < \infty$. The asymptotically optimal general form for data with density $f(x - \theta_0)$ and $f(x) = f(-x)$ is

$$\phi^+(u) = -\frac{f'\left\{F^{-1}\left(\frac{1}{2} + \frac{u}{2}\right)\right\}}{f\left\{F^{-1}\left(\frac{1}{2} + \frac{u}{2}\right)\right\}}.$$

Asymptotic normality is similar to Theorem 12.5 (p. 492) (see for example, Theorem 10.2.5, p. 333 of Randles and Wolfe, 1979). The Edgeworth expansion for W^+ and T^+ may be found on p. 37 and p. 89, respectively, of Hettmansperger (1984).

Table 12.4 Pitman ARE's for the One-Sample Symmetry Problem

Distribution	ARE(S, t)	ARE(S, W^+)	ARE(W^+, t)
Normal	0.64	0.67	0.955
Uniform	0.33	0.33	1.00
Logistic	0.82	0.75	1.10
Laplace	2.00	1.33	1.50
t_6	0.88	0.76	1.16
t_3	1.62	0.85	1.90
t_1 (Cauchy)	∞	1.33	∞

12.8.4 Sign Test

The sign test mentioned in the last section as (12.41) with $a(i) = 1$ is usually given in the form $T^+ = \sum_{i=1}^n I(D_{i0} > 0)$, the number of positive differences. Under the null hypothesis (12.35, p. 491), T^+ has a binomial$(n, 1/2)$ distribution and is extremely easy to use. Because of this simple distribution, T^+ is often given early in a nonparametric course to illustrate exact null distributions.

The sign test does not require symmetry of the distributions to be valid. It can be used as a test of H_0 : median of $D_i - \theta_0 = 0$, where it is assumed only that $D_1, \ldots, D_n$ are independent, each with the same median. Thus, the test is often used in skewed distributions to test that the median has value θ_0. This generality, though, comes with a price because typically the sign test is not as powerful as the signed rank or t test in situations where all three are valid. If there are zeroes in $D_1, \ldots, D_n$, the standard approach is remove them before applying the sign test.

12.8.5 Pitman ARE for the One-Sample Symmetry Problem

In the Appendix, we give some details for finding expressions for the efficacy and Pitman efficiency of tests for the one-sample symmetry problem. Here we just report some Pitman ARE's in Table 12.4 for the sign test, the t test, and the Wilcoxon signed rank. The comparison of the signed rank and the t are very similar to those given in Table 12.3 (p. 477) for the two-sample problem. The only difference is that skewed distributions are allowed in the shift problem but not here.

The general message from Table 12.4 is that the tails of the distribution must be very heavy compared to the normal distribution in order for the sign test to be preferred. This is a little unfair to the sign test because symmetry of f is not required for the sign test to be valid, whereas symmetry is required for the Wilcoxon signed rank test. In fact Hettmansperger (1984, p. 10–12) shows that the sign test is uniformly most powerful among size-α tests if no shape assumptions are made about the density of f. Moreover, in the matched pairs situation where symmetry is justified by differencing, the uniform distribution is not possible, and that is where the sign test performs so poorly.

Monte Carlo power estimates in Randles and Wolfe (1979, p. 116) show that generally the ARE results in Table 12.4 correspond qualitatively to power comparisons. For example, at $n = 10$ and normal alternative $(\theta_0 + .4)/\sigma$, the Wilcoxon signed rank has power .330 compared to .263 for the sign test. The ratio $.263/.330 = .80$ is not too far from ARE$= .64$. The estimated power ratio at $n = 20$ is $.417/.546 = .76$. The Laplace distribution AREs in Table 12.4 are not as consistent. For example, at $n = 20$ for a similar alternative, the ratio is $.644/.571 = 1.13$, not all that close to ARE$= 2.00$.

The Wilcoxon signed rank test is seen to have good power relative to the sign test and to the t test. The Hodges and Lehmann (1956) result that ARE$(W^+, t) \geq .864$ also holds here for all symmetric unimodal densities. Coupled with the fact that there is little loss of power relative to the t test at the normal distribution (ARE$(W^+, t) = 0.955$), W^+ should be the statistic of choice in many situations.

12.8.6 *Treatment of Ties*

The general permutation approach is not usually bothered by ties in the data, although rank methods typically require some thought about how to handle the definition of ranks in the case of ties. For the original situation of n pairs of data and a well-defined statistic like the paired t statistic, the 2^n permutations of the data merely yield redundance if members of a pair are equal. For example, consider $n = 3$ and the following data with all 8 permutations (1 is the original data pairing):

1	2	3	4	5	6	7	8
3,5	5,3	3,5	5,3	3,5	5,3	3,5	5,3
2,2	2,2	2,2	2,2	2,2	2,2	2,2	2,2
7,4	7,4	4,7	4,7	7,4	7,4	4,7	4,7

Permutations 1–4 are exactly the same as permutations 5–8 because permuting the 2nd pair has no effect. Thus, a permutation p-value defined from just permutations 1–4 is exactly the same as for using the full set 1–8. After taking differences between members of each pair, the 2^n sign changes work in the same way by using sign$(0) = 0$; that is, there is the same kind of redundancy in that there are really just 2^{n-n_0} unique permutations, where n_0 is the number of zero differences.

For signed rank statistics, there are two kinds of ties to consider after converting to differences, multiple zeros and multiple non-zero values. For the non-zero multiple values, we just use mid-ranks (average ranks) as before. For the multiple zeros, there are basically two recommended approaches:

Method 1: Remove the differences that are zero and proceed with the reduced sample in the usual fashion. This is the simplest approach and the most powerful for the sign statistic (see Lehmann 1975, p. 144). Pratt and Gibbons (1981, p. 169) discuss anomalies when using this procedure with W^+.

Method 2: First rank all $|D_{10}|, \ldots, |D_{n0}|$. Then remove the ranks associated with the zero values before getting the permutation distribution of the rank statistic, *but do not change the ranks associated with the non-zero values*. However, as above, since the permutation distribution is the same with and without the redundancy, it really just makes the computing easier to remove the ranks associated with the zero values. The normal approximation in (12.42, p. 495) automatically eliminates the ranks associated with the zero values because $\text{sign}(0) = 0$. For the Box-Andersen approximation, the degrees of freedom are different depending on whether the reduced set is used or not. It appears best to use the reduced set for the Box-Andersen approximation although a few zero values make little difference.

Example 12.2 (Fault rates of telephone lines). Welch (1987) gives the difference (times 10^5) of a transformation of telephone line fault rates for 14 matched areas. We modify the data by dividing by 10 and rounding to 2 digits leading to

D_{i0}	−99	31	27	23	20	20	19	−14	11	9	8	−8	6	0
$\text{sign}(D_{i0})R_i$	−14	13	12	11	9.5	9.5	8	−7	6	5	3.5	−3.5	2	0

Notice that there two ties in the absolute values 20 and 8 for which the midranks are given. The exact right-tailed permutation p-value based on the t statistic is .38, whereas the t tables gives .33 and the Box-Andersen approximation is .40. The large outlier −99 essentially kills the power of the t statistic. The sign test first removes the 0 value and then the binomial probability of getting 10 or more positives out of 13 is .046. Welch (1987) used the sample median as a statistic and for these data we get exact p-value .062. Note that the mean and sum and t statistic are all permutationally equivalent, but the median is not permutationally equivalent to using a Wald statistic based on the median. So, the properties of using the median as a test statistic are not totally clear.

For the Wilcoxon Signed Rank, no tables can be used because of the ties and the 0. However, it is straightforward to get the permutation after choosing one of the methods above for dealing with the 0 difference.

Method 1: First remove the 0, then rank. The remaining data are

D_{i0}	−99	31	27	23	20	20	19	−14	11	9	8	−8	6
$\text{sign}(D_{i0})R_i$	−14	13	12	11	9.5	9.5	8	−7	6	5	3.5	−3.5	2

The exact p-value based on the $\text{sign}(D_{i0})R_i$ values above (for example, just insert the signed ranks into the R program below) is.048, the normal approximation is .047, and the Box-Andersen approximation is .049.

Method 2: Rank the data first, then throw away the signed rank associated with the 0. The exact $p$-value is .044 Recall, for the permutation $p$-value, it does not matter whether we drop the 0 or not after ranking. Similarly, the normal approximation $p$-value .042 based on (12.42, p. 495) automatically handles the 0 value. For the Box-Andersen approximation, we get .0437 based on all 14 signed ranks and .0441 after throwing out the 0; so it matters very little whether we include the 0 or not. ♦

For problems with $n \le 20$, the following R code modified from Venables and Ripley (1997, p. 189-190) gives the exact permutation $p$-value for signed statistics:

```
perm.sign<-function(d,stat,pr=FALSE, ...){
 # Exact perm. $p$-value for one-sample problem.
 # Assumes test rejects for large values of stat.
 # Looks at all 2^n sign change samples.
 # Use only for small n.
 # Need the following obscure function
bi<-function(x,digits=if(x>0)1+
             floor(log(x,base=2)) else 1){
  ans<-0:(digits-1)
  (x %/% 2^ans) %% 2
  }        # note %/% and %% are different
# The main program
  t0<-stat(d, ...)
  digits<-length(d)
  b <- 2^digits
  res <- numeric(b)
  for(i in 1:b){
    x <- d*2*(bi(i,digits=digits) - 0.5)
    res[i] <- stat(x, ...)
    if(pr)cat(i,x,res[i],fill=T) # prints
  }
  pvalue <- sum(res >= t0)/b
  sum(res==t0)->co
  return(data.frame(b=b,stat0=round(t0,4),
    eq.t0=co,rt.pvalue=pvalue,pv2=2*pvalue))
}
```

12.9 Randomized Complete Block Data—the Two-Way Design

Blocking is one of the most important techniques for reducing variation in experimental designs. The usual Randomized Complete Block design may be viewed as a generalization of the matched pairs to situations with more than two treatments. To use the permutation argument with blocked data, we do not need for the treatments to be assigned randomly, but it is most natural to discuss blocked data in that context. The key assumption required under H_0 is that the data are exchangeable within blocks.

Suppose that k treatments are to be assigned at random within each block of size k. For n blocks, there are $(k)^n$ possible permutations of the data corresponding to permuting independently among treatments within each block. In the following table there are $k = 4$ blocks with $n = 10$ treatments, thus $M_N = 24^{10} = 6.34\times 10^{13}$ possible permutations. These data are actually treatments 6–15 from an example of aphid infestation of crepe myrtle cultivars given in Table 1 of Brownie and Boos (1994). The response variable is the number of aphids on the three most heavily infested leaves plus the percent of foliage covered with sooty mold.

	Treatments									
Block	1	2	3	4	5	6	7	8	9	10
1	0	0	93	78	5	1	0	21	1	1
2	0	24	0	3	2	180	0	0	3	9
3	0	2	10	0	0	3	2	3	3	140
4	0	4	2	2	0	0	1	47	1	52

The linear model representation is

$$Y_{ij} = \mu + \beta_i + \alpha_j + e_{ij}, \tag{12.43}$$

where $\alpha_1, \ldots, \alpha_k$ are the treatment effects, and $\beta_1, \ldots \beta_n$ are the block effects. Note that we have switched subscripts on Y_{ij} compared to the one-way model (12.28, p. 480) so that the blocks can be the rows. Often the block effects are assumed random, but the nonparametric literature typically considers them fixed effects.

The usual ANOVA F statistic could be used with these data:

$$F = \frac{\dfrac{1}{k-1}\displaystyle\sum_{j=1}^{k} n(\overline{Y}_{.j} - \overline{Y}_{..})^2}{\dfrac{1}{(k-1)(n-1)}\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{k}(Y_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y}_{..})^2}, \tag{12.44}$$

where $\overline{Y}_{i.} = k^{-1}\sum_{j=1}^{k} Y_{ij}$, $\overline{Y}_{.j} = n^{-1}\sum_{i=1}^{n} Y_{ij}$, and $\overline{Y}_{..} = n^{-1}\sum_{i=1}^{n}\overline{Y}_{i.}$. For the above data $F = 0.80$ with p-value = 0.62 from an F distribution with 9 and 27 degrees of freedom. Since the F distribution approximates the permutation distribution, the value 0.62 should be satisfactory. A Monte Carlo approximation to the exact permutation p-value based on 10,000 samples gave .60 with standard error .005, thus confirming the Type I error robustness of the usual F procedure. However, the nonnormality of the response variable is cause for concern because the F statistic is not Type II error robust in the face of outliers. Transformations are an obvious approach, and F on $\log(Y_{ij} + 1)$ resulted in p-value = .29. Fortunately, with rank procedures we do not have to guess the correct transformation.

12.9.1 Friedman's Rank Test

The standard rank procedure was introduced by Friedman (1937). For the untied case, it has the form

$$T = \frac{12n}{k(k+1)} \sum_{j=1}^{k} \left(\overline{R}_{.j} - \frac{k+1}{2} \right)^2, \tag{12.45}$$

where R_{ij} is the rank of Y_{ij} within the ith row, and $\overline{R}_{.j} = n^{-1} \sum_{i=1}^{n} R_{ij}$ is the jth treatment mean rank. Note that $(k+1)/2$ is $\overline{R}_{..}$ since the average of the integers 1 to k is $(k+1)/2$. The within-row ranks R_{ij} for the above table are

	Treatments									
Block	1	2	3	4	5	6	7	8	9	10
1	2	2	10	9	7	5	2	8	5	5
2	2.5	9	2.5	6.5	5	10	2.5	2.5	6.5	8
3	2	4.5	9	2	2	7	4.5	7	7	10
4	2	8	6.5	6.5	2	2	4.5	9	4.5	10

We see immediately that there are numerous ties in the data. The form of the Friedman statistic that accommodates ties is (see, for example, Conover and Iman, 1981, p. 126)

$$T = \frac{(k-1)n^2 \sum_{j=1}^{k} \left(\overline{R}_{.j} - \frac{k+1}{2} \right)^2}{\left(\sum_{i=1}^{n} \sum_{j=1}^{k} R_{ij}^2 \right) - \frac{nk(k+1)^2}{4}}. \tag{12.46}$$

Under the null hypothesis of identical treatments, T converges to a χ^2_{k-1} distribution as $n \to \infty$ and k remains fixed. For the above data, $T = 13.7732$, and comparing to a χ^2_9 distribution gives p-value = .13. However, as in the one-way design, the χ^2 approximation becomes increasingly conservative as the number of treatments gets large relative to the number of blocks. F distribution p-values provide much better approximations and can be justified by either asymptotic theory or the Box-Andersen permutation moment approximations.

12.9.2 *F* Approximations

Friedman (1937, pp. 694–695) conjectured that the Friedman statistic is asymptotically normal as $k \to \infty$ with mean $k-1$ and variance $2(n-1)(k-1)/n$ (a proof may be found in Lemma 4 of Brownie and Boos, 1994). Similar to the one-way design, this asymptotic normal result is consistent with applying the F statistic (12.44, p. 500) to the within-row Friedman ranks and then using the $F(k-1, (k-1)(n-1))$ distribution for p-values. This argument is to be fleshed out in Problem 12.22 (p. 528). Of course, the F distribution should be used in practice; the asymptotic normal result just supports use of the F distribution.

From Box and Andersen (1955, p. 14-15), we may approximate the permutation distribution of F of (12.44, p. 500) or of the same F applied to the within-row Friedman ranks by a $F(d(k-1), d(k-1)(n-1))$ distribution, where

$$d = 1 + \frac{(nk - n + 2)V_2 - 2n}{n(k-1)(n - V_2)},$$

$$V_2 = \frac{1}{n-1}\sum_{i=1}^{n}(s_i^2 - \overline{s}^2)^2/(\overline{s}^2)^2,$$

and the s_i^2 are the within-row variances, and $\overline{s}^2 = n^{-1}\sum_{i=1}^{n} s_i^2$. In the case of the Friedman ranks with no ties in the data, $d = 1 - 2/\{n(k-1)\}$. For the Crepe Myrtle data this latter expression is $d = .944$, the same (to three decimals) as the actual d value from the tied ranks. We summarize the various approximations in the following table:

Approximate P-Values
for the Crepe Myrtle Data

	Monte Carlo	$F(9,27)$	Box-And. $F(9d, 27d)$	χ^2_9
Friedman	.10			.13
F_R	.10	.10	.11	
F on Y	.60	.62	.63	
F on $\log(Y+1)$	.29	.29	.30	

The Monte Carlo estimates are based on 10,000 random permutations and have standard error bounded by .005. The F approximations are good, but the Box-Andersen adjustments do not help here. Interestingly, $d = 1.08$ for the usual F (row 3), but the p-value is adjusted upwards because the $F = .80$ is so small. Typically, a d value greater than 1 lowers the p-value from the F approximation.

Table 12.5 Pitman ARE of the Friedman Test to the F Test

	k = Number of Treatments					
Distribution	2	3	4	5	10	∞
Normal	0.64	0.72	0.76	0.80	0.87	0.955
Uniform	0.67	0.75	0.80	0.83	0.91	1.000
t_3	1.27	1.42	1.52	1.58	1.73	1.900

12.9.3 Pitman ARE for Blocked Data

From van Elteren and Noether (1959) we find the surprising result that the Pitman asymptotic relative efficiency of the Friedman test to the ANOVA F depends on the number of treatments k,

$$\text{ARE(Friedman, } F) = \left\{ \frac{k}{k+1} \right\} 12\sigma^2 \left\{ \int_{-\infty}^{\infty} f^2(x)\, dx \right\}^2, \qquad (12.47)$$

where σ^2 is the variance of the observations. Expression (12.47) is just $k/(k+1)$ times the ARE(W, t) in (12.25, p. 477). Table 12.5 gives a few values of (12.47) for several distributions.

The value .64 at $k = 2$ for the normal distribution is the same as the ARE of the sign test to the t in Table 12.4 (p. 496). That is no accident. It turns out that for $k = 2$, the Friedman test is equivalent to the sign test. (The other values in Table 12.4, p. 496, do not correspond to the $k = 2$ values in Table 12.5 because Table 12.4 refers to the distribution after taking differences, whereas Table 12.5 is for the distribution of the individual treatment results, not the difference of treatment results. For the normal distribution, the difference of normal random variables is also normally distributed; so for the normal the results are the same in both tables.)

The reason for the low efficiency in Table 12.5 is that ranking within rows (intrablock ranking) takes no advantage of between block (interblock) information. For the $k = 2$ case, the Wilcoxon signed rank statistic uses interblock information by ranking the absolute differences (note the improved efficiencies in Table 12.4, p. 496, for the signed rank test compared to the sign test). In the next section we discuss some rank approaches that use interblock information.

12.9.4 Aligned Ranks and the Rank Transform

Many approaches have been used to remedy the low efficiency in Table 12.5 for small values of k. Perhaps the earliest approach (and still one of the best) is the aligned rank method due to Hodges and Lehmann (1962). The aligned rank approach is to first subtract the block mean (or any other location measure such as the median) from each observation Y_{ij}, then rank all the resulting nk residuals together. These latter ranks on the residuals, denoted $\widehat{R}_{ij}$, are called *aligned ranks*. We suggest using F of (12.44, p. 500) on these aligned ranks.

Actually, Sen (1968) and Lehmann (1975, p. 272) use

$$\widehat{Q} = \frac{n^2(k-1)\sum_{j=1}^{k}\left(\overline{\widehat{R}}_{.j} - \frac{nk+1}{2}\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{k}\left(\widehat{R}_{ij} - \overline{\widehat{R}}_{i.}\right)^2}, \tag{12.48}$$

a statistic that is asymptotically χ^2_{k-1} under H_0. The justification for the form (12.48) comes from noting that the permutation mean of $\overline{\widehat{R}}_{.j}$ is $(nk+1)/2$, and the permutation covariance matrix of $(\overline{\widehat{R}}_{.1}, \ldots, \overline{\widehat{R}}_{.k})$ is

$$\frac{\sigma^2 k}{k-1}\text{diag}\left(\boldsymbol{I}_k - \frac{\mathbf{1}_k\mathbf{1}_k^T}{k}\right), \tag{12.49}$$

where $\boldsymbol{I}_k$ is the k-dimensional identity matrix, $\mathbf{1}_k$ is a vector of ones, and

$$\sigma^2 = \frac{1}{n^2k}\sum_{i=1}^{n}\sum_{j=1}^{k}(\widehat{R}_{ij} - \overline{\widehat{R}}_{i.})^2 \tag{12.50}$$

is the permutation variance of $\overline{\widehat{R}}_{.j}$. $\widehat{Q}$ in (12.48) is the appropriate quadratic form in $(\overline{\widehat{R}}_{.1}, \ldots, \overline{\widehat{R}}_{.k})$ upon noting that $(k-1)\boldsymbol{I}_k/(k\sigma^2)$ is a generalized inverse of the covariance matrix (12.49).

Other authors (Fawcett and Salter, 1984, and O'Gorman, 2001) use a one-way ANOVA F on the aligned ranks, but we prefer the two-way F of (12.44, p. 500) because the Box-Andersen adjustment is readily available. All three statistics, $\widehat{Q}$ and the two F statistics on the aligned ranks, are permutationally equivalent to the numerator of $\widehat{Q}$; so if exact or Monte Carlo approximations are used, it does not matter which of the three statistics is chosen. Clearly, either of the two Fs gives better approximate p-values than $\widehat{Q}$ with χ^2_{k-1} p-values.

Mehra and Sarangi (1967) give somewhat complicated formulas for the Pitman ARE of the aligned rank approach to the usual F and to Friedman's statistic, but the bottom line is that the AREs of the aligned rank procedure to the usual F are close to the last column of Table 12.5 (p. 503). Thus, the aligned rank approach is able to recover most of the interblock information.

Another approach to recovering the interblock information is to just rank all the observations together and apply F of (12.44, p. 500) on the resulting ranks. This *rank transform* approach, due to Conover and Iman (1981) works well as long as the block effects are not strong. When the block effects are strong, then this approach is similar to Friedman's test. Hora and Iman (1988) give Pitman ARE results for this approach.

There is an extensive literature on rank methods in block models. Mahfoud and Randles (2005) and Kepner and Wackerly (1996) are several places that briefly review many of the approaches. The latter also gives extensions to incomplete blocks.

12.9.5 Replications within Blocks

In the preceding discussion we have been talking about cases where there is just one observation per cell, nk total observations for n blocks and k treatments, and no block by treatment interaction. Consider the $k = 2$ case and n blocks where there are m_i Xs for the first treatment in block i and n_i Ys for the second treatment, $i = 1, \ldots, n$. These type data arise naturally in clinical trials at n centers or sites. The sites might be hospitals or clinics or individual doctors. The usual rank approach is the van Elteren statistic (van Elteren, 1960, or Lehmann 1975, p. 145), a weighted sum of individual Wilcoxon rank sum statistics W_i within each block,

$$W_{\mathrm{VE}} = \sum_{i=1}^{n} \frac{W_i}{m_i + n_i + 1}.$$

van Elteren (1960) showed that the weights $1/(m_i + n_i + 1)$ are asymptotically optimal among all linear combinations of the W_i. This optimality makes sense if we write the standardized version of W_{VE} as

$$\sum_{i=1}^{n} \frac{1}{\sigma_0^2(\widehat{\theta}_i)} \left(\widehat{\theta}_i - \frac{1}{2}\right) \Big/ \left\{\sum_{i=1}^{n} \frac{1}{\sigma_0^2(\widehat{\theta}_i)}\right\}^{1/2}, \tag{12.51}$$

where $\widehat{\theta}_i$ is the Mann-Whitney estimator of $\theta_i = P(Y_{i1} > X_{i1}) + (1/2)P(Y_{i1} = X_{i1})$ given in (12.14, p. 463) (here we have dropped the XY subscript for simplicity), and $\sigma_0^2(\widehat{\theta}_i)$ is the variance of $\widehat{\theta}_i$ under the null hypothesis of identical X and Y populations. In the completely nonparametric case (in the absence of the shift model), θ_i is the underlying parameter of interest for Wilcoxon statistics. For continuous data (no ties), $\sigma_0^2(\widehat{\theta}_i) = (m_i + n_i + 1)/(12 m_i n_i)$. Thus, the numerator of the standardized version of W_{VE} is a weighted average of $\widehat{\theta}_i - 1/2$, where the weights are inversely proportional to null variances.

The analogous t procedure is based on standardizing

$$\sum_{i=1}^{n} \frac{m_i n_i}{m_i + n_i} (\overline{Y}_i - \overline{X}_i). \tag{12.52}$$

Thus, the t procedure uses a weighted linear combination of the difference of sample means, where the weights are inversely proportional to $\mathrm{Var}\left(\overline{Y}_i - \overline{X}_i\right) = \sigma^2(1/m_i + 1/n_i)$.

The standard permutation approach is to consider all possible

$$M_N = \prod_{i=1}^{n} \binom{m_i + n_i}{n_i}$$

independent permutations within sites. The normal approximation for W_{VE} should be very good if $\sum_{i=1}^{n} m_i$ and $\sum_{i=1}^{n} n_i$ are reasonably large and therefore is widely used in practice. In the case that $\sum_{i=1}^{n} m_i$ and $\sum_{i=1}^{n} n_i$ converge to ∞, Hodges and Lehmann (1962) give the Pitman ARE of (12.51) to (12.52) for normal data as

$$.955 \sum_{i=1}^{n} \frac{m_i n_i}{m_i + n_i + 1} \Bigg/ \sum_{i=1}^{n} \frac{m_i n_i}{m_i + n_i} .$$

Thus, if $m_i + n_i$ is reasonably large, then the ARE is close to the best value .955. For example, if $m_i + n_i = 10$ for each site, then the ARE is .955(10/11).

For the case that there are small numbers of replications per block (site), we are led back to the procedures of the previous section, aligned ranks and possibly the rank transform. With replications within blocks, however, we now have the ability to test for block by treatment interactions. Unfortunately, standard permutation procedures are not available for testing the no interaction hypothesis in the face of main effects. A large literature exists evaluating and criticizing the rank transform approach for testing interactions. See, for example, Akritas (1990, 1991) and Thompson (1991). In general, for more complicated fixed effects models with interaction, to achieve robustness via rank methods, we feel it is better to use the general R-estimation linear model approach mentioned at the end of Section 12.7 (p. 487).

Boos and Brownie (1992) argue that a mixed model approach is usually more appropriate, allowing inferences to be made to a larger population, but the mixed model leads away from van Eltern's statistic (12.51, p. 505) and permutation inference.

12.10 Contingency Tables

12.10.1 2 x 2 Table – Fisher's Exact Test

The first use of the permutation method was given by Fisher (1934a, *Statistical Methods for Research Workers*, fifth edition) in an analysis of 2 × 2 tables. Fisher's example was of 13 identical twins and 17 fraternal twins (of the same sex) who had at least one of the pair convicted of a crime. Of the 13 identical twins only 3 had a twin free of conviction. Of the 17 fraternal twins 15 had a twin free of conviction. Thus the table is as follows,

	Both Convicted	One Convicted	Total
Identical	10	3	13
Fraternal	2	15	17
Total	12	18	30

To fix notation, a general 2 × 2 table is,

	Category 1	Category 2	Total
Group 1	N_{11}	N_{12}	$N_{1.}$
Group 2	N_{21}	N_{22}	$N_{2.}$
Total	$N_{.1}$	$N_{.2}$	N

A standard analysis of these data assumes that N_{11} is binomial $(N_{1.}, p_1)$ and independent of N_{21} assumed to be binomial $(N_{2.}, p_2)$. The usual statistic for testing $H_0 : p_1 = p_2$ is the pooled Z, the square root of the score statistic found in Section 3.2.9 (p. 142),

$$Z = \frac{\widehat{p}_1 - \widehat{p}_2}{\left\{\frac{\widetilde{p}(1-\widetilde{p})}{N_{1.}} + \frac{\widetilde{p}(1-\widetilde{p})}{N_{2.}}\right\}^{1/2}},$$

where $\widehat{p}_1 = N_{11}/N_{1.}$, $\widehat{p}_2 = N_{21}/N_{2.}$, and $\widetilde{p} = N_{.1}/N$. To test $H_a : p_1 > p_2$, the standard approach would be to compare Z to z_α, the $1-\alpha$ quantile of the standard normal.

Instead of this approximate procedure, Fisher noted that conditional on the margins $N_{.1}$ and $N_{.2}$ held fixed in addition to $N_{1.}$ and $N_{2.}$, that a given table has hypergeometric probability of $(n_{11}, n_{12}, n_{21}, n_{22})$ given by

$$\frac{\binom{N_{1.}}{n_{11}}\binom{N_{2.}}{n_{21}}}{\binom{N}{N_{.1}}} = \frac{N_{1.}!N_{2.}!N_{.1}!N_{.2}!}{N!n_{11}!n_{12}!n_{21}!n_{22}!}.$$

This hypergeometric probability is easily obtained if one thinks about an urn with $N_{.1}$ balls of type 1 and $N_{.2}$ of type 2. If we draw out $N_{1.}$ balls without replacement, then the above probability is the probability of getting n_{11} of type 1 and n_{21} of type 2.

One can also think of the above table arising in the two-sample problem where the data consists of just 1's and 0's. Although there are $\binom{N}{N_{1.}}$ permutations of interest, many of them yield the same table. The numerator of the above hypergeometric probability just gives the number of permutations which lead a given table.

Now a variety of statistics can be used to order the possible tables from supporting H_0 to strongly rejecting H_0 and to calculate a p-value. Or one can just use intuition for the ordering: most people would agree that for testing $H_a : p_1 > p_2$, the table below is more extreme than the original.

	Category 1	Category 2	Total
Group 1	$N_{11} + 1$	$N_{12} - 1$	$N_{1.}$
Group 2	$N_{21} - 1$	$N_{22} + 1$	$N_{2.}$
Total	$N_{.1}$	$N_{.2}$	N

Thus, a one-tailed p-value would be obtained by summing up the hypergeometric probabilities of those tables as extreme or more extreme than the original table $(N_{11}, N_{12}, N_{21}, N_{22})$. A number of seemingly different ways of ordering the tables lead to the same definition of "more extreme" and are called Fisher's Exact Test. The simplest way to order is either the intuitive notion above or to order via the pooled Z statistic.

For the twins data, Fisher noted that the two more extreme tables have $N_{11} = 11$, $N_{12} = 2$, $N_{21} = 1$, $N_{22} = 16$ and $N_{11} = 12$, $N_{12} = 1$, $N_{21} = 0$, $N_{22} = 17$. Thus the p-value is the probability of the original table plus the probability of these two more extreme tables:

$$\frac{13!17!12!18!}{30!}\left\{\frac{1}{10!3!2!15!} + \frac{1}{11!2!1!16!} + \frac{1}{12!1!0!17!}\right\} = \frac{619}{1330665} = .000465.$$

The definition of a two-sided p-value is not so clear, but the usual practice is to add in the probabilities of tables as extreme or more extreme in the other direction (having probabilities less than or equal to the probability of the observed table). In the above example we would need to add the probabilities of tables with $N_{11} = 0$, $N_{12} = 13$, $N_{21} = 12$, $b_{22} = 5$ and $N_{11} = 1$, $N_{12} = 12$, $N_{21} = 11$, $N_{22} = 6$ but not $N_{11} = 2$, $N_{12} = 11$, $N_{21} = 10$, $N_{22} = 7$ since it has higher probability than the original table.

When accompanied by a randomization rule to yield exact α levels, Fisher's Exact Test is uniformly most powerful unbiased as discussed in Lehmann (1986, Ch. 4). But many people have noted how conservative it is when p-values are used with the rule: reject H_0 when p-value $\leq \alpha$. In this case the discreteness of the permutation distribution does prove costly in terms of power.

Barnard (1945, 1947), Boschloo (1970), and Suissa and Shuster (1985) proposed unconditional tests in the 2 x 2 table that are typically more powerful than the

Fisher Exact Test without randomization. See Berger (1996) for details and power comparisons.

We have given Fisher's Exact Test in the context of two independent binomials and $H_0 : p_1 = p_2$. It also applies in the context of multinomial data where the data consists of a pair of binary variables (X, Y) with values x_1 and x_2 and y_1 and y_2, respectively:

		Y		
		y_1	y_2	Total
X	x_1	N_{11}	N_{12}	$N_{1.}$
	x_2	N_{21}	N_{22}	$N_{2.}$
	Total	$N_{.1}$	$N_{.2}$	N

The entries $(N_{11}, N_{12}, N_{21}, N_{22})$ are multinomial$(N; p_{11}, p_{12}, p_{21}, p_{22})$ with associated parameters

		Y		
		y_1	y_2	Total
X	x_1	p_{11}	p_{12}	$p_{1.}$
	x_2	p_{21}	p_{22}	$p_{2.}$
	Total	$p_{.1}$	$p_{.2}$	1

In this paired variable context, the null hypothesis for Fisher's Exact Test is independence of X and Y,

$$H_0 : p_{ij} = p_{i.}p_{.j}, \quad i = 1, 2; j = 1, 2. \tag{12.53}$$

Of course, if $p_{11} = p_{1.}p_{.1}$, then all the other equalities such as $p_{12} = p_{1.2}p_{.2}$ hold as well.

12.10.2 Paired Binary Data – McNemar's Test

In the context of paired binary data introduced in the last section, we might expect association between X and Y, but our main interest could be in their marginal probabilities. In particular, the null hypothesis is often

$$H_0 : p_{1.} = p_{.1}. \tag{12.54}$$

A typical application is in matched pair studies such as the following well-known case-control data from Miller (1980),

		Sibling (Control)		
		Tons.	No Tons.	Total
Hodgkin's Patient	Tons.	26	15	41
	No Tons.	7	37	44
	Total	33	52	85

where Hodgkin's patients were paired with a sibling and it was determined whether they each had a tonsillectomy or not. If the marginal estimates $\widehat{p}_{1.} = N_{1.}/N = 41/85$ and $\widehat{p}_{.1} = N_{.1}/N = 33/85$ differ significantly, then incidence of tonsillectomies may be associated with contracting Hodgkin's disease. Noting that $\widehat{p}_{1.} - \widehat{p}_{.1} = N_{12}/N - N_{21}/N$ has multinomial variance $\{p_{12} + p_{21} - (p_{12} - p_{21})^2\}/N = (p_{12} + p_{21})/N$ under H_0, the score statistic is

$$Z = \frac{N_{12} - N_{21}}{(N_{12} + N_{21})^{1/2}}.$$

Exact inference follows by noting that under (12.54, p. 509), $N_{12}|N_{12} + N_{21}$ has a binomial$(N_{12} + N_{21}, 1/2)$ distribution. Thus, $Z = 1.71$ has approximate normal one-sided p-value $= .044$, but $P(\text{binomial}(22, 1/2) \geq 15) = .067$. These procedures are generally referred to as McNemar's test.

What do these tests have to do with permutation and rank statistics? Let $X = 1$ denote that a Hodgkin's patient had a tonsillectomy, and $X = 0$ denote that he/she did not, and similarly $Y = 1$ and $Y = 0$ for the sibling control. Then the paired data and their differences are

	Hodgkin's	Sibling	
Pair	Patient	(Control)	Diff.
1	1	1	0
.	.	.	.
.	.	.	.
26	1	1	0
27	1	0	1
.	.	.	.
.	.	.	.
41	1	0	1
42	0	1	−1
.	.	.	.
.	.	.	.
48	0	0	0
49	0	0	0
.	.	.	.
.	.	.	.
85	0	0	0

Note that there are $N_{12} = 15$ positive differences out of $N_{12} + N_{21} = 22$ nonzero differences. Thus, the exact binomial procedure above is just the sign test for the differences, and Z is exactly (12.42, p. 495) for $a(i) = 1$. In fact, since all the nonzero absolute differences are identically 1, the exact signed rank test (assuming zeroes are deleted) yields the same binomial procedure, and Z is also (12.42, p. 495) with $a(i) = i$.

12.10.3 *I by J Tables*

We now consider the general I by J contingency table

				Y			
		y_1	.	.	.	y_J	Total
X	x_1	N_{11}	.	.	.	N_{1J}	$N_{1.}$
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	x_I	N_{I1}	.	.	.	N_{IJ}	$N_{J.}$
	Total	$N_{.1}$	.	.	.	$N_{.J}$	N

The distribution of these data could be a full multinomial with IJ cells or I independent rows of multinomial data. In either case, exact permutation analysis is achieved by conditioning on the marginal totals resulting in a multiple hypergeometric for the joint distribution of the entries N_{ij} having probability $P(N_{ij} = n_{ij}, i = 1, \ldots, I; j = 1, \ldots, J \mid N_{1.}, \ldots, N_{I.}, N_{.1}, \ldots, N_{.J})$ given by

$$\frac{\left(\prod_{i=1}^{I} N_{i.}!\right)\left(\prod_{j=1}^{J} N_{.j}!\right)}{N! \prod_{i=1}^{I} \prod_{j=1}^{J} n_{ij}!}.$$

The question remains as to what statistic should be used. If both X and Y have nominal categories, then the chi-squared goodness-of-fit statistic is natural, but not very interesting. If X and Y have numerical scores or are at least ordered, then some type of association or correlation statistic should be used. For example, one might use Pearson's r or Spearman's rank correlation. If X has nominal categories and Y has numerical categories, then ANOVA type comparisons among the row means makes sense. If X has nominal categories and Y has ordered categories, then the Kruskal-Wallis test might be a good choice of statistic. Moreover, all these situations can be generalized to multi-way tables, say I by J by K tables, usually viewed as stratified comparisons of X and Y.

All these options for statistics in two-way and multiway tables come under the general purview of *Generalized Cochran-Mantel-Haenszel statistics*. Expositions of these statistics may be found in Landis et al. (1978) and Agresti (2002, Section 7.5.3) and implementation is found in `SAS PROC FREQ`.

12.11 Confidence Intervals and R-Estimators

Confidence intervals can be obtained from permutation and rank test statistics in the same way as for other types of statistics: choose values of θ appearing in a null hypothesis such that the statistic $T(\theta)$ viewed as a function of θ does not reject the null hypothesis (see 3.19, p. 144). We often refer to this approach as "inverting a test statistic." For example, in the one-sample problem with data $D_1, \ldots, D_n$ assumed to be symmetrically distributed about θ_0, a two-sided permutation t test could just as well be based on $T(\theta_0) = |\sum_{i=1}^{n}(D_i - \theta_0)|$. The permutation distribution depends on the 2^n sign change configurations of $D_i - \theta_0, \ldots, D_n - \theta_0$; we reject if $T(\theta_0)$ is larger than the largest α of the 2^n values of $T(\theta_0)$ computed on those permutations. So the $1 - \alpha$ confidence interval can be found by trial and error, but it would seem to be a pretty laborious task because the permutation distribution changes with each θ_0. A somewhat easier computing method is suggested in Lehmann (1986, p. 263), but in general, the usual t interval is close enough to the permutation interval that it is mostly used in practice.

Inverting the signed rank statistic W^+ leads to an interval $[W_{(k_1)}, W_{(k_2)}]$, where $W_{(1)} \le W_{(2)} \cdots \le W_{(n(n+1)/2)}$ are the ordered values of the *Walsh averages*

$$W_{ij} = \frac{D_i + D_j}{2}, \qquad 1 \le i \le j \le n. \tag{12.55}$$

The order number k_2 is such that $P(W^+ \ge k_2) \le \alpha/2$, and $k_1 = n(n+1)/2+1-k_2$. We have specified a closed interval so that the probability of coverage is at least $1-\alpha$ for tied data situations (see Randles and Wolfe, 1979, p. 181-183). For example, at $n = 7$ with continuous data and $\alpha = .05$, $P(W^+ \ge 26) = P(W^+ \le 2) = .0234$, and thus the interval $[W_{(3)}, W_{(26)}]$ has exact confidence level $1-.0468 = .9532$. Often k_1 and k_2 are taken from the normal approximation to the permutation distribution of W^+. For example, $k_1 = q + 1$ and $k_2 = n(n + 1)/2 - q$, where q is the closest integer to

$$\frac{n(n+1)}{4} - z_{\alpha/2}\left\{\frac{1}{4}\sum_{i=1}^{n} R_i^2\right\}^{1/2}.$$

In the $n = 7$ example above, this latter calculation gives 2.4, and thus $q = 2$, $k_1 = 3$, and $k_2 = 28 - 2 = 26$ as before. For the sample -1.11, 2.23, 3.35, 4.67, 5.34, 6.17, 7.44, the interval is $[W_{(3)}, W_{(26)}] = [1.12, 6.39]$.

Inverting the sign test leads to an interval of order statistics

$$(D_{(k)}, D_{(n-k+1)}), \quad 1 \le k \le n - k + 1.$$

This interval has exact coverage probability $C_n(k) = 1 - (1/2)^{n-1} \sum_{i=0}^{k-1} \binom{n}{i}$ for the population median from any continuous, not necessarily symmetric distribution. To obtain at least the same coverage for any discrete distribution, we need to again change to the closed interval $[D_{(k)}, D_{(n-k+1)}]$. An interesting addendum to these intervals due to Guilbaud (1979) is that the average of two such intervals,

$$\left[\frac{D_{(k)} + D_{(k+t)}}{2}, \frac{D_{(n-k-t+1)} + D_{(n-k+1)}}{2}\right], \quad k + t \le n - k - t + 1,$$

has guaranteed coverage $\{C_n(k) + C_n(k + t)\}/2$ for any distribution. This latter interval is useful for small n because it give more options for the confidence level than given by $C_n(k)$ alone. A more practical solution is given byHettmansperger and Sheather (1986), who interpolate between adjacent order statistics to get an interval with approximately the specified confidence, say 95%. The intervals are no longer distribution-free, but the confidence is close to the specified value.

Moving to the two-sample problem, the permutation interval based on the two-sample t is hard to compute, similar to the one-sample interval, and the usual t interval is mostly used in practice. Inversion of the Wilcoxon Rank Sum statistic for the shift model $G(x) = F(x - \Delta)$ leads to a confidence interval for Δ of the form $[U_{(k_1)}, U_{(k_2)}]$, where $U_{(1)} \le U_{(2)} \cdots \le U_{(mn)}$ are the ordered values of the pairwise differences

$$U_{ij} = Y_j - X_i, \quad i = 1, \ldots, m; j = 1, \ldots, n. \tag{12.56}$$

Similar to the one-sample case, k_2 is chosen so that $P(W \ge k_2 + n(n+1)/2) = \alpha/2$ and $k_1 = mn + 1 - k_2$. In practice, one often uses the normal approximation interval with $k_1 = q + 1$ and $k_2 = mn - q$, where q is the integer closest to

$$\frac{mm}{2} - z_{\alpha/2} \{\mathrm{Var}(W)\}^{1/2},$$

where Var(W) is given by (12.10, p. 462) or (12.11, p. 462).

Point estimators obtained from rank test statistics were introduced by Hodges and Lehmann (1963). These *R-estimators* inherit some of the natural robustness properties of rank methods; see, for example Huber (1981) and Serfling (1980, Ch. 9), Randles and Wolfe (1979, Ch. 7), and Hettmansperger (1984, Ch. 5). The most well known are: i) the one-sample center of symmetry estimator $\widehat{\theta} = \mathrm{median}\{W_{ij}\}$, where the W_{ij} are in (12.55, p. 512); and ii) the two-sample shift estimator $\widehat{\Delta} = \mathrm{median}\{U_{ij}\}$, where the U_{ij} are in (12.56, p. 513). Asymptotic relative efficiency comparisons for confidence intervals and estimators derived from rank tests are exactly the same as for the associated rank tests.

12.12 Appendix – Technical Topics for Rank Tests

12.12.1 Locally Most Powerful Rank Tests

Recall from Section 12.5.1 (p. 474) that for $H_0 : \Delta = 0$ versus $H_a : \Delta > 0$, if there exists a rank test that is uniformly most powerful of level α for some $\epsilon > 0$ in the restricted testing problem $H_0 : \Delta = 0$ versus $H_{a,\epsilon} : 0 < \Delta < \epsilon$, we say that the test is the *locally most powerful rank test* for the original testing problem. By using a Taylor expansion of the probability of the rank vector $\boldsymbol{R}$ as a function of Δ, $L_{\boldsymbol{r}}(\Delta) \equiv P_\Delta(\boldsymbol{R} = \boldsymbol{r})$, we need only obtain an expression for the derivative of $L_{\boldsymbol{r}}(\Delta)$ and maximize it.

To see this consider the Taylor expansion

$$L_{\boldsymbol{r}}(\Delta) = L_{\boldsymbol{r}}(0) + L'_{\boldsymbol{r}}(0)\Delta + o(|\Delta|),$$

and a rank test with $\alpha = k/N!$ based on maximizing $L'_{\boldsymbol{r}}(0)$. Let $\boldsymbol{r}^{(1)}$ be the rank configuration that makes $L'_{\boldsymbol{r}}(0)$ largest among all $N!$ rank configurations, $\boldsymbol{r}^{(2)}$ makes $L'_{\boldsymbol{r}}(0)$ second largest among all $N!$ rank configurations, etc. Such a rank test has power

$$\beta(\Delta) = \sum_{j=1}^{k} L_{\boldsymbol{r}^{(j)}}(\Delta) = \sum_{j=1}^{k} \left[\frac{1}{N!} + L'_{\boldsymbol{r}^{(j)}}(0)\Delta + o(|\Delta|) \right].$$

For each rank configuration $\boldsymbol{r}^{(j)}$, we can choose Δ_j small enough so that $L_{\boldsymbol{r}^{(j)}}(\Delta)$ is also the jth largest among $L_{\boldsymbol{r}^{(1)}}(\Delta), \ldots, L_{\boldsymbol{r}^{(N!)}}(\Delta)$ for all $0 < \Delta < \Delta_j$. Now take ϵ to be smaller than all of the Δ_j. This shows that for $0 < \Delta < \epsilon$, the power of the test that places points in the rejection region as ordered by $L'_{\boldsymbol{r}}(0)$ also puts points in the rejection as ordered by $P_\Delta(\boldsymbol{R} = \boldsymbol{r}) = L_{\boldsymbol{r}}(\Delta)$; in other words, it is the locally most powerful rank test.

Let us now consider the two-sample problem where $X_1, \ldots, X_m$ are iid with distribution function $F(x)$, and $Y_1, \ldots, Y_n$ are iid with distribution function $G(x)$. Suppose that F and G have densities $f(x)$ and $g(x)$, respectively, whose support is contained in that of a density $h(x)$. This means that $h(x)$ is positive whenever $f(x)$ and $g(x)$ are positive; for example, when all three densities have support on $(-\infty, \infty)$. From Theorem 12.6, (p. 515), we have

$$P(\boldsymbol{R} = \boldsymbol{r}) = \frac{1}{N!}\mathrm{E}\left[\frac{\prod_{i=1}^{m} f(V_{(r_i)}) \prod_{i=m+1}^{N} g(V_{(r_i)})}{\prod_{i=1}^{m} h(V_{(r_i)}) \prod_{i=m+1}^{N} h(V_{(r_i)})} \right],$$

where $V_{(1)} < \cdots < V_{(N)}$ are the order statistics of an iid sample of size N from $h(x)$.

Shift alternatives have the form $g(x) = f(x - \Delta)$ so that the X distribution has the same shape as the Y distribution but shifted Δ to the right of it. If $f(x)$ has support on $(-\infty, \infty)$, then we may take $h(x) = f(x)$ and obtain

$$P_{\Delta}(\boldsymbol{R} = \boldsymbol{r}) = \frac{1}{N!}\mathrm{E}\left[\frac{\prod_{i=m+1}^{N} f(V_{(r_i)} - \Delta)}{\prod_{i=m+1}^{N} f(V_{(r_i)})}\right], \tag{12.57}$$

where now $V_{(1)} < \cdots < V_{(N)}$ are order statistics for a random sample from f. Now suppose that $f(x)$ is differentiable and that we can take the derivative inside the expectation in (12.57) . Then,

$$L'_{\boldsymbol{r}}(0) = \frac{\partial}{\partial \Delta} P_{\Delta}(\boldsymbol{R} = \boldsymbol{r})\bigg|_{\Delta=0} = \frac{1}{N!}\sum_{i=m+1}^{N} \mathrm{E}\left[\frac{-f'(V_{(r_i)})}{f(V_{(r_i)})}\right]. \tag{12.58}$$

The locally most powerful rank test places points in the rejection region according to large values of this latter expression.

If we let $V_{(1)} < \cdots < V_{(N)}$ be replaced by $F^{-1}(U_{(1)}) < \cdots < F^{-1}(U_{(N)})$ where the $U_{(i)}$ are uniform order statistics from an iid sample $U_1, \ldots, U_N$, then the locally most powerful rank test rejects for large values of

$$T = \sum_{i=m+1}^{N} a(R_i),$$

where $a(i) = \mathrm{E}\phi(U_{(i)}, f)$, and $\phi(u, f) = -f'(F^{-1}(u))/f(F^{-1}(u))$ is given in (12.23, p. 475) and called the optimal score function.

12.12.2 Distribution of the Rank Vector under Alternatives

A version of the following result first appeared in Hoeffding (1951).

Theorem 12.6. *Suppose that $Z_1, \ldots Z_N$ are independent continuous random variables with respective densities $f_1, \ldots, f_N$. Let $\boldsymbol{R} = (R_1, \ldots, R_N)^T$ be the corresponding rank vector. If h is the density of a continuous random variable whose support contains the support of each of $f_1, \ldots, f_N$, then*

$$P(\boldsymbol{R} = \boldsymbol{r}) = \frac{1}{N!}E\left[\frac{\prod_{i=1}^{N} f_i(V_{(r_i)})}{\prod_{i=1}^{N} h(V_{(r_i)})}\right],$$

where $V_{(1)} < \cdots < V_{(N)}$ are the order statistics of an iid sample from h.

Proof. Let $C = \{\boldsymbol{t} : t_i \text{ has rank } r_i\}$. Then by definition

$$P(\boldsymbol{R} = \boldsymbol{r}) = \int \cdots \int I(\boldsymbol{t} \in C)\left\{\prod_{i=1}^{N} f_i(t_i)\right\} dt_1 dt_2 \cdots dt_N.$$

Now let $v_{(r_i)} = t_i$ so that $v_{(1)} < \cdots < v_{(N)}$. On the set C this is just a 1-to-1 change of variable, but its implications are important. For a given vector $\boldsymbol{t}$ suppose that t_1 has rank $r_1 = 3$; that is, t_1 is third from the bottom when the components of $\boldsymbol{t}$ are ranked. Then $v_{(r_1)} = v_{(3)} = t_1$. If t_2 has rank $r_2 = 9$, then $v_{(r_2)} = v_{(9)} = t_2$. Now we make the change of variable, and multiply and divide by $N!\prod_{i=1}^{N} h(v_{(r_i)})$ to get

$$P(\boldsymbol{R} = \boldsymbol{r}) = \frac{1}{N!}\int \cdots \int \left[\frac{\prod_{i=1}^{N} f_i(v_{(r_i)})}{\prod_{i=1}^{N} h(v_{(r_i)})}\right] I(v_{(1)} < \cdots < v_{(N)})N!$$

$$\times \left\{\prod_{i=1}^{N} h(v_{(i)})\right\} dv_{(1)}dv_{(2)}\cdots dv_{(N)}.$$

The result follows by noticing that $I(v_{(1)} < \cdots < v_{(N)})N!\prod_{i=1}^{N} h(v_{(i)})$ is the density of the order statistic vector from h. ■

12.12.3 Pitman Efficiency

Recall from Section (12.5.2, p. 476) that the Pitman asymptotic relative efficiency of test S to test T is given by

$$\text{ARE}(S, T) = \lim_{k\to\infty} \frac{N_k'}{N_k},$$

where N_k and N_k' are the sample sizes required for the two tests to have the same limiting level α and power β under the sequence of alternatives

$$\theta_k = \theta_0 + \frac{\delta}{\sqrt{N_k}} + o\left(\frac{1}{\sqrt{N_k}}\right) \quad \text{as} \quad k \to \infty. \tag{12.59}$$

These sequences of alternatives are called *Pitman alternatives*, and the basic approach is due to Pitman (1948) and Noether (1955). In the following we have drawn heavily from the accounts in Lehmann (1975) and Randles and Wolfe (1979).

We assume in Theorem 12.7 below that both test statistics satisfy 1–7 below. For simplicity we state the conditions for just S and then give a result on asymptotic power before giving the main theorem.

In the following $\mu_{S_k}(\theta)$ and $\sigma_{S_k}(\theta)$ refer to sequences of constants associated with S_k under θ. They might be the means and standard deviations, but need not be.

1.

$$\theta_k \to \theta_0 \quad \text{as} \quad k \to \infty.$$

2.
$$N_k \to \infty \text{ as } k \to \infty.$$

3. Under $\theta = \theta_0$
$$\frac{S_k - \mu_{S_k}(\theta_0)}{\sigma_{S_k}(\theta_0)} \xrightarrow{d} \mathrm{N}(0,1) \text{ as } k \to \infty.$$

4. Under $\theta = \theta_k$
$$\frac{S_k - \mu_{S_k}(\theta_k)}{\sigma_{S_k}(\theta_k)} \xrightarrow{d} \mathrm{N}(0,1) \text{ as } k \to \infty.$$

5. The derivative $\mu'_{S_k}(\theta)$ exists in a neighborhood of $\theta = \theta_0$ with $\mu'_{S_k}(\theta_0) > 0$ and
$$\frac{\mu'_{S_k}(\theta_k^*)}{\mu'_{S_k}(\theta_0)} \to 1 \text{ for all } \theta_k^* \to \theta_0 \text{ as } k \to \infty.$$

6.
$$\frac{\sigma_{S_k}(\theta_k)}{\sigma_{S_k}(\theta_0)} \to 1 \text{ as } k \to \infty.$$

7. There exists a positive constant c such that
$$c = \lim_{k\to\infty} \frac{\mu'_{S_k}(\theta_0)}{\sqrt{N_k \sigma^2_{S_k}(\theta_0)}}.$$

This constant c is called the efficacy of S and denoted eff(S). Based on these conditions we first give a result on asymptotic power. The result shows that the higher the efficacy of a test, the more power it has. The result also gives a way to approximate the power of a test based on S. Let Z be a standard normal random variable, and let z_α be its upper $1-\alpha$ quantile.

Theorem 12.7. *Suppose that the test that rejects for $S_k > c_k$ has level $\alpha_k \to \alpha$ as $k \to \infty$ under $H_0 : \theta = \theta_0$.*

a) If Conditions 1–7 and (12.59, p. 516) hold, then
$$\beta_k = P(S_k > c_k) \to P(Z > z_\alpha - c\delta) \text{ as } k \to \infty, \tag{12.60}$$
where δ is given in (12.59, p. 516).

b) If Conditions 1–7 and (12.60) *hold, then (12.24, p. 476) holds.*

Proof. Note first that if Condition 3. holds, then since $\alpha_k \to \alpha$
$$\frac{c_k - \mu_{S_k}(\theta_0)}{\sigma_{S_k}(\theta_0)} \to z_\alpha \text{ as } k \to \infty.$$

Now $P(S_k > c_k)$ is given by

$$P\left(\frac{S_k - \mu_{S_k}(\theta_k)}{\sigma_{S_k}(\theta_k)} > \left[\frac{c_k - \mu_{S_k}(\theta_0)}{\sigma_{S_k}(\theta_0)} - \frac{\mu_{S_k}(\theta_k) - \mu_{S_k}(\theta_0)}{\sigma_{S_k}(\theta_0)}\right]\frac{\sigma_{S_k}(\theta_0)}{\sigma_{S_k}(\theta_k)}\right)$$

$$\to P(Z > z_\alpha - c\delta) \quad \text{as} \quad k \to \infty.$$

To see this last step, note that by the mean value theorem there exists a θ_k^* such that

$$\begin{aligned}\frac{\mu_{S_k}(\theta_k) - \mu_{S_k}(\theta_0)}{\sigma_{S_k}(\theta_0)} &= \frac{\mu'_{S_k}(\theta_k^*)(\theta_k - \theta_0)}{\sigma_{S_k}(\theta_0)} \\ &= \frac{\mu'_{S_k}(\theta_k^*)}{\mu'_{S_k}(\theta_0)}\frac{\mu'_{S_k}(\theta_0)}{\sqrt{N_k\sigma^2_{S_k}(\theta_0)}}\sqrt{N_k}(\theta_k - \theta_0) \to c\delta.\end{aligned}$$

For part b) we just work backwards and note that (12.60) and Conditions 1–7 force the convergence to $c\delta$ which means that $\sqrt{N_k}(\theta_k - \theta_0) \to \delta$ which is equivalent to (12.59, p. 516). ■

Now we give the main Pitman ARE theorem.

Theorem 12.8. *Suppose that the tests that reject for $S_k > c_k$ and $T_k > c'_k$ based on sample sizes N_k and N'_k, respectively, have levels α_k and α'_k that converge to α under $H : \theta = \theta_0$ and their powers under θ_k both converge to β, $\alpha < \beta < 1$. If conditions 1–7 hold and their efficacies are c =eff(S) and c'=eff(T), respectively, then the Pitman asymptotic relative efficiency of S to T is given by*

$$\text{ARE} = \left\{\frac{\text{eff}(S)}{\text{eff}(T)}\right\}^2.$$

Proof. By Theorem 12.7 (p. 517) b), $\beta = P(Z > z_\alpha - c\delta) = P(Z > z_\alpha - c'\delta')$. Thus $c\delta = c'\delta'$ and

$$\begin{aligned}\text{ARE}(S,T) &= \lim_{k\to\infty}\frac{N'_k}{N_k} \\ &= \lim_{k\to\infty}\left(\frac{\sqrt{N'_k}(\theta_k - \theta_0)}{\sqrt{N_k}(\theta_k - \theta_0)}\right)^2 \\ &= \left(\frac{\delta'}{\delta}\right)^2 = \left(\frac{c}{c'}\right)^2.\end{aligned}$$

■

To apply Theorem 12.8 it would appear that we have to verify Conditions 3–6 above for arbitrary subsequences θ_k converging to θ_0 and then compute the efficacy in 7 for such sequences. However, if Conditions 1–7 and (12.60, p. 517) hold, we know by Theorem 12.7 (p. 517) that (12.24, p. 476) holds. Thus, we really only need to assume Condition 2 and verify Conditions 3–6 for alternatives of the form (12.59, p. 516). Moreover, the efficacy need only be computed for a simple sequence N converging to ∞ since the numerator and denominator in Condition 7 only involve θ_0.

12.12.4 Pitman ARE for the One-Sample Location Problem

Using the notation of Section 12.8 (p. 491) let $D_1, \ldots, D_N$ be iid from $F(x-\theta)$, where $F(x)$ has density $f(x)$ that is symmetric about 0, $f(x) = f(-x)$. Thus D_i has density $f(x-\theta)$ that is symmetric about θ. The testing problem is $H_0 : \theta = \theta_0$ versus $H_a : \theta = \theta_k$, where θ_k is given by (12.59).

12.12.4a Efficacy for the One-Sample *t*

The one-sample t statistic is

$$t = \frac{\sqrt{N}(\overline{D} - \theta_0)}{s},$$

where s is the $n-1$ version of the sample standard deviation. The simplest choice of standardizing constants are

$$\mu_{t_k}(\theta_k) = \frac{\sqrt{N_k}(\theta_k - \theta_0)}{\sigma}$$

and $\sigma_{t_k}(\theta_k) = 1$, where σ is the standard deviation of D_1 (under both $\theta = \theta_0$ and $\theta = \theta_k$). To verify Conditions 3 and 4 (p. 517), we have

$$\begin{aligned}\frac{t_k - \mu_{t_k}(\theta_0)}{\sigma_{t_k}(\theta_0)} &= \frac{\sqrt{N_k}(\overline{D} - \theta_0)}{s} - \frac{\sqrt{N_k}(\theta_k - \theta_0)}{\sigma} \\ &= \frac{\sqrt{N_k}(\overline{D} - \theta_k)}{\sigma}\left(\frac{s}{\sigma}\right) + \sqrt{N_k}(\theta_k - \theta_0)\left(\frac{1}{s} - \frac{1}{\sigma}\right).\end{aligned}$$

Under both $\theta = \theta_0$ and $\theta = \theta_k$, s has the same distribution and converges in probability to σ if D has a finite variance. Thus, under $\theta = \theta_k$ the last term in the latter display converges to 0 in probability since (12.59) forces $\sqrt{N_k}(\theta_k - \theta_0)$ to converge to δ. Of course under $\theta = \theta_0$ this last term is identically 0. The standardized means converge to standard normals under both $\theta = \theta_0$ and $\theta = \theta_k$ by Theorem 5.33 (p. 262). Two applications of Slutsky's Theorem then gives

Conditions 3 and 4 (p. 517). Since the derivative of $\mu_{t_k}(\theta)$ is $\mu'_{t_k}(\theta) = \sqrt{N_k}/\sigma$ for all θ, Condition 5 (p. 517) is satisfied. Since $\sigma_{t_k}(\theta_k) = 1$, Condition 6 (p. 517) is satisfied. Finally, dividing $\mu'_{t_k}(\theta_0) = \sqrt{N_k}/\sigma$ by $\sqrt{N_k}$ yields

$$\text{eff}(t) = \frac{1}{\sigma}.$$

It should be pointed out that this efficacy expression also holds true for the permutation version of the t test because the permutation distribution of the t statistic also converges to a standard normal under $\theta = \theta_0$.

12.12.4b Efficacy for the Sign Test

The sign test statistic is the number of observations above θ_0,

$$S = \sum_{i=1}^{N} I(D_i > \theta_0).$$

S has a binomial$(N, 1/2)$ distribution under $\theta = \theta_0$ and a binomial$(N, 1-F(\theta_0-\theta))$ distribution under general θ. Let $\mu_{S_k}(\theta) = N[1 - F(\theta_0 - \theta)]$ and $\sigma^2_{S_k}(\theta) = N[1 - F(\theta_0 - \theta)]F(\theta_0 - \theta)$. Conditions 3. and 4. (p. 517) follow again by Theorem 5.33 (p. 262), and $\mu'_{S_k}(\theta) = Nf(\theta_0 - \theta)$. Since F is continuous, Condition 6 (p. 517)is satisfied, and if f is continuous, then Condition 5 (p. 517) is satisfied, and the efficacy is

$$\text{eff}(S) = \lim_{N\to\infty} \frac{Nf(0)}{\sqrt{N^2/4}} = 2f(0).$$

Now we are able to compute the Pitman ARE of the sign test to the t test:

$$\text{ARE}(S, t) = 4\sigma^2 f^2(0).$$

Table 12.4 (p. 496) gives values of ARE(S, t) for some standard distributions.

12.12.4c Efficacy for the Wilcoxon Signed Rank Test

Recall that the signed rank statistic is

$$W^+ = \sum_{i=1}^{N} I(D_i > \theta_0)R_i^+,$$

where R_i^+ is the rank of $|D_i - \theta_0|$ among $|D_1 - \theta_0|, \ldots, |D_N - \theta_0|$. The asymptotic distribution of W^+ under θ_k requires more theory than we have developed so far, but Olshen (1967) showed that the efficacy of W^+ is

$$\sqrt{12} \int_{-\infty}^{\infty} f^2(x) dx$$

under the condition that $\int_{-\infty}^{\infty} f^2(x) dx < \infty$. Thus the Pitman asymptotic relative efficiency of the sign test to the Wilcoxon Signed Rank test is

$$\text{ARE}(S, W^+) = \frac{f^2(0)}{3 \left(\int_{-\infty}^{\infty} f^2(x) dx\right)^2}.$$

Similarly, the Pitman asymptotic relative efficiency of the Wilcoxon Signed Rank test to the t test is

$$\text{ARE}(W^+, t) = 12\sigma^2 \left(\int_{-\infty}^{\infty} f^2(x) dx\right)^2.$$

Table 12.4 (p. 496) displays these AREs for a number of distributions.

12.12.4d Power approximations for the One-Sample Location problem

Theorem 12.7 (p. 517) gives the asymptotic power approximation

$$P(Z > z_\alpha - c\delta) = 1 - \Phi\left(z_\alpha - c\sqrt{N}(\theta - \theta_0)\right)$$

based on setting $\delta = \sqrt{N}(\theta - \theta_0)$ in (12.60, p. 517), where θ is the alternative of interest at sample size N.

For example, let us first consider the t statistic with $c = 1/\sigma$ and $\theta_0 = 0$. The power approximation is then

$$1 - \Phi\left(z_\alpha - \sqrt{N}\theta/\sigma\right).$$

This is the exact power we get for the Z statistic $\sqrt{N}(\overline{X} - \theta_0)/\sigma$ when we know σ instead of estimating it. At $\theta/\sigma = .2$ and $N = 10$, we get power 0.16, which may be compared with the estimated exact power taken from the first four distributions in Randles and Wolfe (1979, p. 116): .14, .15, .16, .17. These latter estimates were based on 5000 simulations and have standard deviation around .005. At $\theta/\sigma = .4$ and $N = 10$, the approximate power is 0.35, and the estimated exact powers for those first four distributions in Randles and Wolfe (1979, p. 116) are .29, .33, .35, and .37, respectively. So here our asymptotic approximation may be viewed as substituting a Z for the t, and the approximation is quite good. Of course, for the normal distribution we could easily have used the noncentral t distribution to get the exact power.

For the sign test, the approximation is

$$1 - \Phi\left(z_\alpha - \sqrt{N} 2 f(0)\theta\right) = 1 - \Phi\left(z_\alpha - \sqrt{N} 2 f_0(0)\theta/\sigma\right),$$

where we have put f in the form of a location-scale model $f(x) = f_0((x - \theta)/\sigma)/\sigma$, where $f_0(x)$ has standard deviation 1, and thus σ is the standard deviation. For the uniform distribution, $f_0(x) = I(-\sqrt{3} < x < \sqrt{3})/\sqrt{12}$, so that $2 f_0(0) = 2/\sqrt{12}$. The approximate power at $\theta/\sigma = .2, .4, .6, .8$ and $N = 10$ is then .10, .18, .29, .43, respectively. The corresponding Randles and Wolfe (1979, p. 116) estimates are .10, .19, .30, and .45, respectively. Here of course we could calculate the power exactly using the binomial. The approximate power we have used is similar to the normal approximation to the binomial but not the same because our approximation has replaced the difference of $p = F(0) = 1/2$ and $p = F(\theta)$ by a derivative times θ (Taylor expansion) and also used the null variance. It is perhaps surprising how good the approximation is.

The most interesting case is the signed rank statistic because we do not have any standard way of calculating the power. The approximate power for an alternative θ when $\theta_0 = 0$ is

$$\begin{aligned} P(Z > z_\alpha - c\delta) &= 1 - \Phi\left(z_\alpha - \theta\sqrt{12N}\int_{-\infty}^{\infty} f^2(x)dx\right) \\ &= 1 - \Phi\left(z_\alpha - \frac{\theta}{\sigma}\sqrt{12N}\int_{-\infty}^{\infty} f_0^2(x)dx\right). \end{aligned}$$

Here again in the second part we have substituted so that σ is the standard deviation of $f(x)$. For example, at the standard normal $\int_{-\infty}^{\infty} f_0^2(x)dx = 1/\sqrt{4\pi}$, and the approximate power is

$$1 - \Phi\left(z_\alpha - \sqrt{\frac{3N}{\pi}}\frac{\theta}{\sigma}\right).$$

Plugging in $\theta/\sigma = .2$, .4, .6, and .8 at $N = 10$, we obtain .15, .34, .58, and .80, respectively. The estimates of the exact powers from Randles and Wolfe (1979, p. 116) are .14, .32, .53, and .74. Thus the asymptotic approximation is a bit too high, especially at the larger θ/σ values.

Although the approximation is a little high, it could easily be used for planning purposes. For example, suppose that a clinical trial is to be run with power $= .80$ at the $\alpha = .05$ level against alternatives expected to be around $\theta/\sigma = .5$. Since the FDA requires two-sided procedures, we use $z_{.025} = 1.96$ and solve $\Phi^{-1}(1 - .8) = 1.96 - \sqrt{3N/\pi}(.5)$ to get

$$N = \left[\frac{1.96 - \Phi^{-1}(.2)}{.5}\right]^2 \frac{\pi}{3} = 32.9.$$

Notice that if we invert the Z statistic power formula used above for approximating the power of the t statistic, the only difference from the last display is that the factor $\pi/3$ does not appear. Thus for the t the calculations result in 31.4 observations. Of course this ratio $3/\pi = 31.4/32.9$ is just the ARE efficiency of the signed rank test to the t test at the normal distribution.

12.13 Problems

12.1. For the permutations in Table 12.1 (p. 453), give the permutation distribution of the Wilcoxon Rank Sum statistic W.

12.2. For the two-sample problem with samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, show that the permutation test based on $\sum_{i=1}^{n} Y_i$ is equivalent to the permutation tests based on $\sum_{i=1}^{m} X_i$, $\sum_{i=1}^{n} Y_i - \sum_{i=1}^{m} X_i$, and $\overline{Y} - \overline{X}$.

12.3. A one-way ANOVA situation with $k = 3$ groups and two observations within each group ($n_1 = n_2 = n_3 = 2$) results in the following data. Group 1: 37, 24; Group 2: 12, 15; Group 3: 9, 16. The ANOVA $F = 5.41$ results in a p-value of .101 from the F table. If we exchange the 15 in Group 2 for the 9 in Group 3, then $F = 7.26$.

a. What are the total number of ways of grouping the data that are relevant to testing that the means are equal?
b. Without resorting to the computer, give reasons why the permutation p-value using the F statistic is 2/15.

12.4. In a one-sided testing problem with continuous test statistic T, the p-value is either $F_H(T_{\text{obs.}})$ or $1 - F_H(T_{\text{obs.}})$ depending on the direction of the hypotheses, where F_H is the distribution function of T under the null hypothesis H, and $T_{\text{obs.}}$ is the observed value of the test statistic. In either case, under the null hypothesis the p-value is a uniform random variable as seen from the probability integral transformation. Now consider the case where T has a discrete distribution with values $t_1, \ldots, t_k$ and probabilities $P(T = t_i) = p_i, i = 1, \ldots, k$ under the null hypothesis H_0. If we are rejecting H_0 for small values of T, then the p-value is $p = P(T \le T_{\text{obs.}}) = p_1 + \cdots + P(T = T_{\text{obs.}})$, and the mid-$p$ value is $p - (1/2)P(T = T_{\text{obs.}})$. Under the null hypothesis H_0, show that E(mid-p)=1/2 and thus that the expected value of the usual p-value must be greater than 1/2 (and thus greater than the expected value of the p-value in continuous cases).

12.5. Consider a finite population of values $a_1, \ldots, a_N$ and a set of constants $c_1, \ldots, c_N$. We select a random permutation of the a values, call them $A_1, \ldots, A_N$, and form the statistic

$$T = \sum_{i=1}^{N} c_i A_i.$$

The purpose of this problem is to derive the first two permutation moments T given in Section 12.4.2 (p. 458).

a. First show that

$$P(A_i = a_s) = \frac{1}{N} \quad \text{for } s = 1, \ldots, N,$$

and

$$P(A_i = a_s, A_j = a_t) = \frac{1}{N(N-1)} \quad \text{for } s \neq t = 1, \ldots, N.$$

(Hint: for the first result there are $(N-1)!$ permutations with a_s in the ith slot out of a total of $N!$ equally likely permutations.)

b. Using a. show that

$$\mathrm{E}(A_i) = \frac{1}{N}\sum_{i=1}^{N} a_i \equiv \overline{a}, \quad \mathrm{Var}(A_i) = \frac{1}{N}\sum_{i=1}^{N}(a_i - \overline{a})^2, \quad \text{for } i = 1, \ldots, N,$$

and

$$\mathrm{Cov}(A_i, A_j) = \frac{-1}{N(N-1)}\sum_{i=1}^{N}(a_i - \overline{a})^2, \quad \text{for } i \neq j = 1, \ldots, N.$$

c. Now use b. to show that

$$\mathrm{E}(T) = N\overline{c}\,\overline{a} \quad \text{and} \quad \mathrm{Var}(T) = \frac{1}{N-1}\sum_{i=1}^{N}(c_i - \overline{c})^2 \sum_{j=1}^{N}(a_j - \overline{a})^2,$$

where $\overline{a}$ and $\overline{c}$ are the averages of the a's and c's, respectively.

12.6. As an application of the previous problem, consider the Wilcoxon Rank Sum statistic W = sum of the ranks of the Y's in a two-sample problem where we assume continuous distributions so that there are no ties. The c values are 1 for $i = m + 1, \ldots, N = m+n$ and 0 otherwise. With no ties the a's are just the integers $1, \ldots, N$ corresponding to the ranks. Show that

$$\mathrm{E}(W) = \frac{n(m+n+1)}{2}$$

and

$$\mathrm{Var}(W) = \frac{mn(m+n+1)}{12}.$$

12.7. In Section 12.4.4 (p. 461), the integral

$$P(X_1 < X_2) = \mathrm{E}\{I(X_1 < X_2)\} = \int\int I(x_1 < x_2)\, dF(x_1)\, dF(x_2)$$
$$= \int F(x)\, dF(x)$$

arises, where X_1 and X_2 are independent with distribution function F. If F is continuous, argue that $P(X_1 < X_2) = 1/2$ since $X_1 < X_2$ and $X_1 > X_2$ are equally likely. Also use iterated expectations and the probability integral transformations to get the same result. Finally, let $u = F(x)$ in the final integral to get the result.

12.8. Suppose that X and Y represent some measurement that signals the presence of disease via a threshold to be used in screening for the disease. Assume that Y has distribution function $G(y)$ and represents a diseased population, and X has distribution function $F(x)$ and represents a disease-free population. A "positive" for a disease-free subject is declared if $X > c$ and has probability $1 - F(c)$, where $F(c)$ is called the *specificity* of the screening test. A "positive" for a diseased subject is declared if $Y > c$ and has probability $1 - G(c)$, called the *sensitivity* of the test. The receiver operating characteristic (ROC) curve is a plot of $1 - G(c_i)$ versus $1 - F(c_i)$ for a sequence of thresholds $c_1, \ldots, c_k$. Instead of a discrete set of points, we may let $t = 1 - F(c)$, solve to get $c = F^{-1}(1 - t)$, and plug into $1 - G(c)$ to get the ROC curve $R(t) = 1 - G(F^{-1}(1 - t))$. Show that

$$\int_0^1 R(t)\, dt = \int \{1 - G(u)\}\, dF(u) = \theta_{\mathrm{XY}}$$

for continuous F and G.

12.9. Use the asymptotic normality result for $\widehat{\theta}_{\mathrm{XY}}$ to derive (12.15, p. 464).

12.10. Use (12.15, p. 464) to prove that the power of the Wilcoxon Rank Sum Test goes to 1 as m and n go to ∞ and m/N converges to a number λ between 0 and 1. You may assume that the F and G are continuous.

12.11. Use (12.15, p. 464) to derive (12.16, p. 464).

12.12. Suppose that $\widehat{\theta}_{\mathrm{XY}}$ is .7 and $m = n$. How large should $m = n$ be in order to have approximately 80% power at $\alpha = .05$ with the Wilcoxon Rank Sum Test?

12.13. Suppose that two normal populations with the same standard deviation σ differ in means by $\Delta/\sigma = .7$. How large should $m = n$ be in order to have approximately 80% power at $\alpha = .05$ with the Wilcoxon Rank Sum Test?

12.14. The number of permutations needed to carry out a permutation test can be computationally overwhelming. Thus the typical use of a permutation test involves estimating the true permutation p-value by randomly selecting $B = 1{,}000$, $B = 10{,}000$, or even more of the possible permutations. If we use sampling with replacement, then $B\widehat{p}$ has a binomial distribution with the true p-value p being the probability in the binomial. Consider the following situation where an approach of questionable ethics is under consideration. A company has just run a clinical trial comparing a placebo to a new drug that they want to market, but unfortunately the estimated p-value based on $B = 1000$ shows a p-value of around $\widehat{p} = .10$. Everybody is upset because they "know" the drug is good. One clever doctor suggests that they run the simulation of $B = 1000$ over and over again until they get a $\widehat{p}$ less than .05. Are they likely to find a run for which $\widehat{p}$ is less than .05 if the true p-value is $p = .10$? Use the following calculation based on k separate (independent) runs resulting in $\widehat{p}_1, \ldots, \widehat{p}_k$:

$$\begin{aligned} P(\min_{1\le i\le k} \widehat{p}_i \le .05) &= 1 - P(\min_{1\le i\le k} \widehat{p}_i > .05) \\ &= 1 - [1 - P(\widehat{p}_1 \le .05)]^k \\ &= 1 - [1 - P(\text{Bin}(1000,.1) \le 50)]^k. \end{aligned}$$

Plug in some values of k to find out how large k would need to be to get a $\widehat{p}$ under .05 with reasonably high probability.

12.15. The above problem is for given data, and we were trying to estimate the true permutation p-value conditional on the data set and therefore conditional on the set of test statistics computed for every possible permutation. In the present problem we want to think in terms of the overall unconditional probability distribution of $B\widehat{p}$ where we have two stages: first the data is generated and then we randomly select $T_1^*, \ldots, T_B^*$ from the set of permutations. The calculation of importance for justifying Monte Carlo tests is the unconditional probability $P(\widehat{p} \le \alpha) = P(B\widehat{p} \le B\alpha)$ that takes both stages into account.

a. First we consider a simpler problem. Suppose that we get some data that seems to be normally distributed and decide to compute a t statistic, call it T_0. Then we discover that we have lost our t tables, but fortunately we have a computer. Thus we can generate normal data and compute $T_1^*, \ldots, T_B^*$ for each of B independent data sets. In this case $T_0, T_1^*, \ldots, T_B^*$ are iid from a continuous distribution so that there are no ties among them with probability one. Let $\widehat{p} = \sum_{i=1}^{B} I(T_i^* \ge T_0)/B$ and prove that $B\widehat{p}$ has a discrete uniform distribution on the integers $(0, 1, \ldots, B + 1)$. (Hint: just use the argument that each ordering has equal probability $1/((B + 1)!)$. For example, $B\widehat{p} = 0$ occurs when T_0 is the largest value. How many orderings have T_0 as the largest value?)
b. The above result also holds if $T_0, T_1^*, \ldots, T_B^*$ have no ties and are merely exchangeable. However, if we are sampling $T_1^*, \ldots, T_B^*$ with replacement from a finite set of permutations, then ties occur with probability greater than one. Think of a way to randomly break ties so that we can get the same discrete uniform distribution.
c. Assuming that $B\widehat{p}$ has a discrete uniform distribution on the integers $(0, 1, \ldots, B)$, show that $P(\widehat{p} \le \alpha) = \alpha$ as long as $(B + 1)\alpha$ is an integer.

12.16. From (12.20, p. 469), $d = .933$ for the Wilcoxon Rank Sum statistic for $m = 10$ and $n = 6$ and assuming no ties. This corresponds to $\boldsymbol{Z}$ being the integers 1 to 16. For no ties and $W = 67$, the exact p-value for a one-sided test is .0467. Show that the normal approximation p-value is .0413 and the Box-Andersen p-value is .0426. Also find the Box-Andersen p-values using the approximations $d = 1 + (1.8 - 3)/(m + n)$ and $d = 1$.

12.17. Show that the result "$Q/(k-1)$ of (12.31, p. 482) is AN$\{1, 2(n-1)/(kn)\}$ as $k \to \infty$ with n fixed" follows from (12.32, p. 483) and writing

$$\sqrt{k}\left(\frac{Q}{k-1} - \frac{nF_{\rm R}}{n-1+F_{\rm R}}\right) = \frac{\sqrt{k}\{(N-1)/(k-1) - n\}F_{\rm R}}{(n-1)\left(\dfrac{k}{k-1}\right) + F_{\rm R}}$$

$$+ \sqrt{k}(nF_{\rm R})\left(\frac{1}{(n-1)\left(\dfrac{k}{k-1}\right) + F_{\rm R}} - \frac{1}{n-1+F_{\rm R}}\right).$$

Then show that each of the above two pieces converges to 0 in probability and use the delta theorem on $nF_{\rm R}/(n-1+F_{\rm R})$. (Keep in mind that n is a fixed constant.)

12.18. Justify the statement: "use of $F_{\rm R}$ with an $F(k-1, N-k)$ reference distribution is supported by (12.32, p. 483) under $k \to \infty$ and by the usual asymptotics $(k-1)F_{\rm R} \xrightarrow{d} \chi^2_{k-1}$ when $n \to \infty$ with k fixed." Hint: for the $k \to \infty$ asymptotics, write an $F(k-1, N-k)$ random variable as an average of $k-1$ χ^2_1 random variables divided by an independent average of $k(n-1)$ χ^2_1 random variables. Then subtract 1, multiply by $\sqrt{k}$ and use the Central Limit Theorem and Slutsky's Theorem.

12.19. From Section 12.8.1 (p. 492), show that for $T = \sum_{i=1}^n c_i d_i$, $\mathrm{E}(T^4) = 3(\sum_{i=1}^n d_i^2)^2 - 2\sum_{i=1}^n d_i^4$. (Hint: first show that

$$\left(\sum c_i d_i\right)^4 = \sum c_i^4 d_i^4 + 6\sum_{i<j} c_i^2 d_i^2 c_j^2 d_j^2$$

plus sums of odd moments.)

12.20. Verify (12.39, p. 493) and (12.40, p. 493) for the Box-Andersen approximation in the matched pairs problem.

12.21. Using results in Section 12.4.2 (p. 458), show that $\mathrm{E}\{\overline{R}_{.j}\} = (k+1)/2$, $\mathrm{Var}\{\overline{R}_{.j}\} = (k^2-1)/(12n)$, and $\mathrm{Cov}\{\overline{R}_{.j}, \overline{R}_{.m}\} = -(k^2-1)/\{12n(k-1)\}$, where $R_{i1}, \ldots R_{ik}$ are Friedman ranks in the ith block randomly assigned to the integers 1 to k and independent of the ranks in the other blocks. Putting these results together, the covariance matrix of $\overline{\boldsymbol{R}} = (\overline{R}_{.1}, \ldots, \overline{R}_{.k})^T$ is $\{k(k+1)/(12n)\}\boldsymbol{C}_k$, where $\boldsymbol{C}_k =$

diag $\left(\boldsymbol{I}_k - \frac{\mathbf{1}_k \mathbf{1}_k^T}{k}\right)$. Using the fact that C_k is idempotent, find a generalized inverse of the covariance matrix of $\overline{\boldsymbol{R}}$, call it G, and show that (12.45, p. 501) is given by $\overline{\boldsymbol{R}}^T G \overline{\boldsymbol{R}}$.

12.22. Similar to Problem 12.18, explain why asymptotic normality of the Friedman statistic (12.45, p. 501) supports use of the F in (12.44, p. 500) on the within row Friedman ranks with an $F(k-1, (k-1)(n-1))$ reference distribution.

12.23. From Section 12.9.4 (p. 503) verify the permutation moments in (12.49, p. 504) and (12.50, p. 504). Use results from Section 12.4.2 (p. 458) under the assumption that permutations are independently carried out within rows.

12.24. From Section 12.10.1 (p. 506) consider the two independent binomial testing problem where $m = 12$ $(N_{11} + N_{12})$ for Group 1 and $n = 4$ $(N_{21} + N_{22})$ for Group 2, and we want to test $H_0 : p_1 = p_2$ versus $H_a : p_1 < p_2$, where p_1 and p_2 are the respective probabilities of falling in Category 1. Suppose that $T = 4$ $(N_{11} + N_{21})$ is observed. Write down the conditional probability distribution of $N_{11}|T = 4$ (just the hypergeometric probabilities for $n_{11} = 0, 1, 2, 3, 4$). Also, letting each of $0, 1, 2, 3, 4$ be considered observed values for N_{11}, list:

a. the Fisher Exact p-values
b. the Fisher Exact mid-p values.

12.25. For a multinomial vector $(N_{11}, N_{12}, N_{21}, N_{22})$, $N_{11}+N_{12}+N_{21}+N_{22} = N$, with associated probabilities $(p_{11}, p_{12}, p_{21}, p_{22})$, show that the variance of $N_{12}-N_{21}$ is $N\{p_{12} + p_{21} - (p_{12} - p_{21})^2\}$.

12.26. Show that (12.58, p. 515) follows from (12.57, p. 515) if the derivative can be taken inside the expectation.

12.27. Show why $\alpha_k \to \alpha$ and Condition 3. (p. 517) imply that

$$\frac{c_k - \mu_{S_k}(\theta_0)}{\sigma_{S_k}(\theta_0)} \to z_\alpha \quad \text{as} \quad k \to \infty.$$

(Hint: it helps to use Pólya's result on uniform convergence, Theorem 5.6, p. 222.)

12.28. Verify that Theorem 5.33 (p. 262) applies to $\overline{X}$ when $X_1^*, \ldots, X_{N_k}^*$ are iid from $F(x)$ having mean 0 and finite variance σ^2, and $X_i = X_i^* + \delta/\sqrt{N_k}, i = 1, \ldots, N_k$.

12.29. Verify that Theorem 5.33 (p. 262) applies to $S = \sum_{i=1}^{N} I(X_i > 0$ when $X_1^*, \ldots, X_{N_k}^*$ are iid from $F(x)$ having median 0 and $X_i = X_i^* + \delta/\sqrt{N_k}, i = 1, \ldots, N_k$.

12.30. The data are $Y_1, \ldots, Y_n$ iid with median θ. For $H_0 : \theta = 0$ versus $H_a : \theta > 0$, use the normal approximation to the binomial distribution to find a power approximation for the sign test and compare to the expression

$1 - \Phi\left(z_\alpha - \sqrt{N}2f(0)\theta_a\right)$ derived from Theorem 12.7 (p. 517), where θ_a is an alternative. Where are the differences?

12.31. For the Wilcoxon Signed Rank statistic, calculate an approximation to the power of a .05 level test for a sample of size $N = 20$ from the Laplace distribution with a shift of .6 in standard deviation units. Compare with the simulation estimate .63 from Randles and Wolfe (1979, p.116).

12.32. Consider the two-sample problem where $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are iid from $F(x)$ under H_0, but the Y's are shifted to the right by $\Delta_k = \delta/\sqrt{N_k}$ under a sequence of the Pitman alternatives. Verify Conditions 3.-6 (p. 517), making any assumptions necessary and show that the efficacy of the two-sample t test is given by $\text{eff}(t) = \sqrt{\lambda(1-\lambda)}/\sigma$, where σ is the standard deviation of F.

12.33. Consider a variable having a Likert scale with possible answers 1,2,3,4,5. Suppose that we are thinking of a situation where the treatment group has answers that tend to be spread toward 1 or 5 and away from the middle. Can we design a rank test to handle this? Here is one formulation. For the two-sample problem suppose that the base density is a beta density of the following form:

$$\frac{\Gamma(2(1-\theta))}{\Gamma(1-\theta)\Gamma(1-\theta)} x^{-\theta}(1-x)^{-\theta}, \quad 0 < x < 1, \quad \theta < 1.$$

A sketch of this density shows that it spreads towards the ends as θ gets large. Using the LMPRT theory, find the optimal score function for $H_0 : \theta = \theta_0$ versus $H_a : \theta > \theta_0$, where $0 \le \theta_0 < 1$. At $\theta_0 = 0$, the score function simplifies to $\phi(u) = -2 - \log[u(1-u)]$. Sketch this score function and comment on whether a linear rank statistic of the form $S = \sum_{i=1}^m \phi(R_i/(N+1))$ makes sense here.

12.34. For the two-sample problem with $G(x) = (1-\Delta)F(x) + \Delta F^2(x)$ and $H_0 : \Delta = 0$ versus $H_a : \Delta > 0$, show that the Wilcoxon Rank Sum test is the locally most powerful rank test. (You may take $h(x) = f(x)$ in the expression for $P(\boldsymbol{R} = \boldsymbol{r})$.)

12.35. In some two-sample situations (treatment and control), only a small proportion of the treatment group responds to the treatment. Johnson et al. (1987) were motivated by data on sister chromatid exchanges in the chromosomes of smokers where only a small number of units are affected by a treatment, that is, where the treatment group seemed to have a small but higher proportion of large values than the control group. For this two-sample problem, they proposed a mixture alternative,

$$G(x) = (1-\Delta)F(x) + \Delta K(x),$$

where $K(x)$ is stochastically larger than $F(x)$, i.e., $K(x) \le F(x)$ for all x, and Δ refers to the proportion of responders. For $H_0 : \Delta = 0$ versus $H_a : \Delta > 0$, verify that the locally most powerful rank test has optimal score function

$k(F^{-1}(u))/f(F^{-1}(u)) - 1$. Let $F(x)$ and $K(x)$ be normal distribution functions with means μ_1 and μ_2, respectively, $\mu_2 > \mu_1$, and variance σ^2. Show that the optimal score function is

$$\phi(u) = \exp(-\delta^2/2)\exp(\delta\Phi^{-1}(u)) - 1, \tag{12.61}$$

where $\delta = (\mu_2 - \mu_1)/\sigma$.

12.36. Related to the previous problem, Johnson et al. (1987) give the following example data:

```
X: 9  9 10 10 14 14 14 15 16 20
Y: 6 10 13 15 18 21 22 23 30 37
```

By sampling from the permutation distribution of the linear rank statistic $\sum_{i=m+1}^{m+n}\phi(R_i/(m+n+1))$ with score function in (12.61), estimate the one-sided permutation p-values with $\delta = 1$ and $\delta = 2$. For comparison, also give one-sided p-values for the Wilcoxon rank sum (exact) and pooled t-tests (from t table).

12.37. Similar in motivation to problem 12.35 (p. 529), Conover and Salsburg (1988) proposed the mixture alternative

$$G(x) = (1-\Delta)F(x) + \Delta\{F(x)\}^a.$$

Note that $\{F(x)\}^a$ is the distribution function of the maximum of a random variables with distribution function $F(x)$. For $H_0 : \Delta = 0$ versus $H_a : \Delta > 0$, verify that the locally most powerful rank test has optimal score function u^{a-1}.

12.38. For the data in Problem 12.36 (p. 530), by sampling from the permutation distribution of the linear rank statistic $\sum_{i=m+1}^{m+n}\phi(R_i/(m+n+1))$ with score function $\phi(u) = u^{a-1}$, estimate the one-sided permutation p-value with $a = 5$. For comparison, also give one-sided p-values for the Wilcoxon rank sum (exact) and pooled t-tests (from t table).

12.39. Conover and Salsburg (1988) gave the following example data set on changes from baseline of serum glutamic oxaloacetic transaminase (SGOT):

```
X:  -50  -17  -10  -3   4   7   8  12  26  37
Y: -116  -56   20  24  29  29  35  35  37  41
```

Plot the data and decide what type of test should be used to detect larger values in some or all of the Y's. Then, give the one-sided p-value for that test and for one other possible test.

12.40. Use `perm.sign` to get the exact one-sided p-value 0.044 for the data give in Example 12.2 (p. 498). Then by trial and error get an exact confidence interval for the center of the distribution with coverage at least 90%. Also give the exact confidence interval for the median based on the order statistics with coverage at least 90%.