BIOMETRICS 43, 641–655
September 1987

Two-Sample Rank Tests for Detecting Changes That Occur in a Small Proportion of the Treated Population

Richard A. Johnson

Department of Statistics, University of Wisconsin, 1210 West Dayton, Madison, Wisconsin 53706, U.S.A.

Steve Verrill

Lawrence Livermore National Laboratory, Livermore, California 94550, U.S.A.

and

Dan H. Moore II

Lawrence Livermore National Laboratory, Livermore, California 94550, U.S.A.

SUMMARY

In the course of studying a biological phenomenon thought to be a precursor to chromosome breakage, researchers have found that treatments sometimes produce a higher proportion of "outliers" than do controls. Our examples pertain to smokers and patients undergoing chemotherapy, although the statistical methods developed here would apply to subjects exposed to any other health hazard. We formulate the problem in a nonparametric setting. Locally most powerful rank tests are obtained for mixture alternatives. In one instance, the approximate scores test has the simple form of counting the number of treatment responses above a combined sample percentile.

Our test statistics are compared to the Wilcoxon and normal scores tests using empirical power studies and asymptotic efficiencies.

1. Introduction

We often suppose that the effect of a biological treatment is to shift a response. It sometimes happens, however, that the primary effect of a treatment is not to shift the complete population but to increase the proportion of large responses. This paper is devoted to the study of nonparametric procedures for use in this case.

In this paper, we obtain rank tests that are locally most powerful with respect to changes in the mixing proportion, ε, in the distribution

$$F_\varepsilon(x) = (1 - \varepsilon)F_1(x) + \varepsilon F_2(x), \tag{1.1}$$

where $F_2(x) \leq F_1(x)$ for all x and $F_1 \neq F_2$. We are primarily concerned with testing the null hypothesis $\varepsilon = 0$ versus the alternative $\varepsilon > 0$, but we will also investigate the nonzero null case.

The alternative of an increase in the mixing parameter typifies situations where most units are unaffected by the treatment but a small proportion have an unusually large response. This alternative was motivated by data collected by Carrano and Moore (1982) concerning the number of sister chromatid exchanges (SCEs) in the chromosomes of smokers. Figure 1 illustrates a representative set of data.

Key words: Chromosomes; Mixtures; Rank tests.

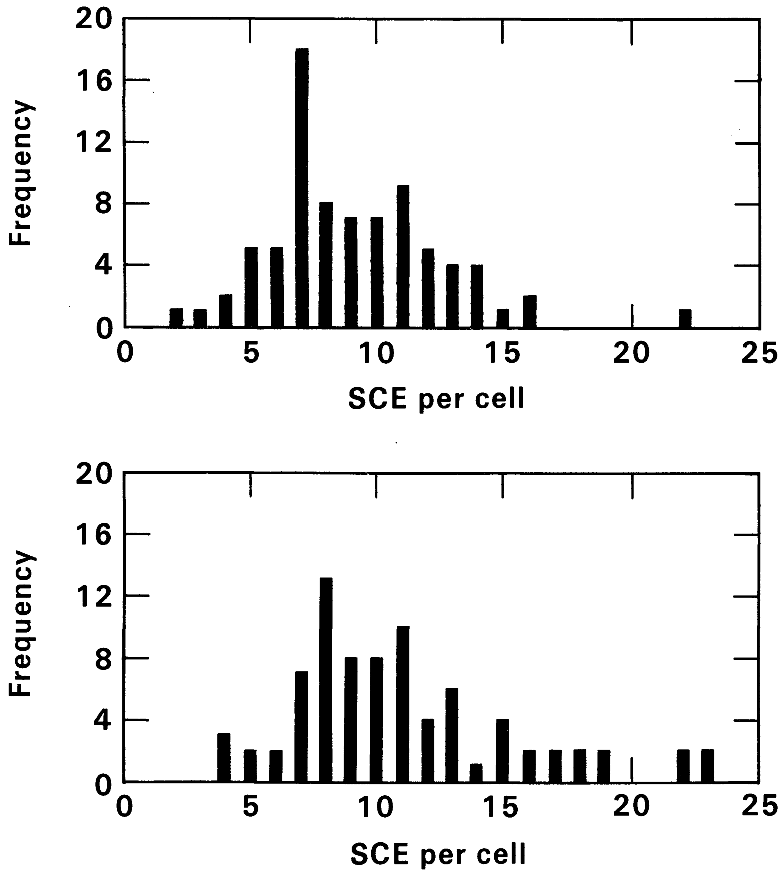

Figure 1. Distribution of sister chromatid exchanges (SCEs) in lymphocytes of a nonsmoker (top) and a smoker (bottom). Each distribution is based on scoring 80 cells.

Carrano and Moore demonstrated that exposure to toxic substances can lead to an increase in the frequency of SCEs in certain types of circulating blood cells (lymphocytes). However, only a small fraction (denoted by ε in the model) of the cells show persistent effects in the form of elevated SCE rates. Smokers appear to have a greater proportion of cells showing persistent damage, in the form of extra SCEs, than do nonsmokers. Although there is no direct evidence, biologists believe that SCEs may be precursors to more serious genetic damage (such as chromosome breakage) that may eventually lead to cancer. Thus, SCE counts obtained from subjects exposed to a particular substance might serve as a measure of the carcinogenicity of that substance.

The observations on SCEs are obtained through a time-consuming and technically difficult process of drawing blood samples, growing the cells of interest in a medium, staining them with a fluorescent dye, and then scoring them under a microscope. Typically, then, sample sizes will be moderate with 20–80 cells per person. Carrano and Moore showed that the distribution of SCEs varied from person to person and that no single parametric family of distributions could be fit to the observed data. These properties suggest that nonparametric procedures will be more useful in testing for effects of exposures than parametric procedures for which it is difficult to maintain the significance level.

Carrano and Moore (1982) proposed an ad hoc statistical test that is based on the count of observations above a fixed level. In Section 2, we derive the locally most powerful test for $\varepsilon = 0$ versus $\varepsilon > 0$ and show that one test, the quantile test, is asymptotically equivalent to the test proposed by Carrano and Moore. The most striking feature of our score functions is that they are the antitheses of those applied for purposes of robustness. Large responses are accentuated while other values are played down. An example is given illustrating the calculation of the quantile test using some chemotherapy data.

Large-sample normal approximations to the critical values are obtained in Section 3. We also present asymptotic relative efficiency calculations in that section. A power study is given in Section 4. The efficiency and power calculations confirm that there are a variety of cases in which the nonstandard rank tests proposed here surpass the performance of the standard normal scores and Wilcoxon rank tests. Finally, in Section 5, we extend our results to cases where the mixing proportion is nonzero under the null hypothesis.

2. Locally Most Powerful Rank Tests

Most of the previous literature on rank tests pertains to location, scale, or regression alternatives. Exceptions include Bickel and Doksum (1969) and Bhattacharyya and Johnson (1973), who considered changes in the shape parameter of Weibull and gamma distributions. Good (1979) proposed a permutation test for the mixture model where the second component is a mean shift of the first component. Lehmann (1953) studied mixtures of the form

$$(1 - \varepsilon)F_1 + \varepsilon F_1^2$$

and showed that the Wilcoxon test is locally most powerful. He also indicated a procedure for specifying a second component, of the form $\varepsilon h(F_1)$, so that the median test is optimal. We consider a mixture model where the second component is not an explicit function of the first. In the process, we obtain a general quantile test including the median test as the approximate scores version of a locally most powerful test.

Theorem 2.1 Let $X_1, \ldots, X_m$ be a random sample from $F_1(x)$ and $Y_1, \ldots, Y_n$ be an independent random sample from the $F_\varepsilon(x)$ distribution given by (1.1). Let the support of $F_2(x)$ be contained within the support of $F_1(x)$, and let F_1 and F_2 have densities f_1 and f_2. The locally most powerful rank test for testing H_0: $\varepsilon = 0$ versus H_1: $\varepsilon > 0$ rejects H_0 for large values of

$$\sum_{i=1}^{n} \mathrm{E}\left[\frac{f_2(V^{(r_i)})}{f_1(V^{(r_i)})} - 1\right],$$

where $V^{(1)} \leq \cdots \leq V^{(m+n)}$ is an ordered sample from $F_1(x)$ and $r_1 \leq \cdots \leq r_n$ are the ranks of $Y_1, \ldots, Y_n$ in the combined sample.

Proof See Appendix A.1.

Quite often the locally most powerful rank test is approximated by replacing the expectation in the statement of Theorem 2.1 by the *score*

$$J\left(\frac{r_i}{m+n+1}\right) = \frac{f_2\{F_1^{-1}[r_i/(m+n+1)]\}}{f_1\{F_1^{-1}[r_i/(m+n+1)]\}} - 1, \tag{2.1}$$

where $J(u)$, the *score function*, is defined on the interval $0 < u < 1$ by

$$J(u) = \frac{f_2[F_1^{-1}(u)]}{f_1[F_1^{-1}(u)]} - 1. \tag{2.2}$$

We consider two special cases of our tests.

Mixture of normal distributions Let $F_1(x)$ be $N(\mu_1, \sigma^2)$ and $F_2(x)$ be $N(\mu_2, \sigma^2)$ with $\mu_2 > \mu_1$ and set $\Delta = (\mu_2 - \mu_1)/\sigma$. The locally most powerful test depends on

$$\frac{f_2(x)}{f_1(x)} - 1 = \exp\left[\Delta\left(x - \frac{\mu_1 + \mu_2}{2}\right)\bigg/\sigma\right] - 1. \tag{2.3}$$

Since $F_1^{-1}(u) = \mu_1 + \sigma\Phi^{-1}(u)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function, the mixed normal (MN) score function is

$$J_{MN}(u) = \frac{f_2[F_1^{-1}(u)]}{f_1[F_1^{-1}(u)]} - 1 = e^{-\Delta^2/2}e^{\Delta\Phi^{-1}(u)} - 1, \quad 0 < u < 1. \tag{2.4}$$

This function is bounded below by -1 and is unbounded as u approaches 1 (see Figure 2).

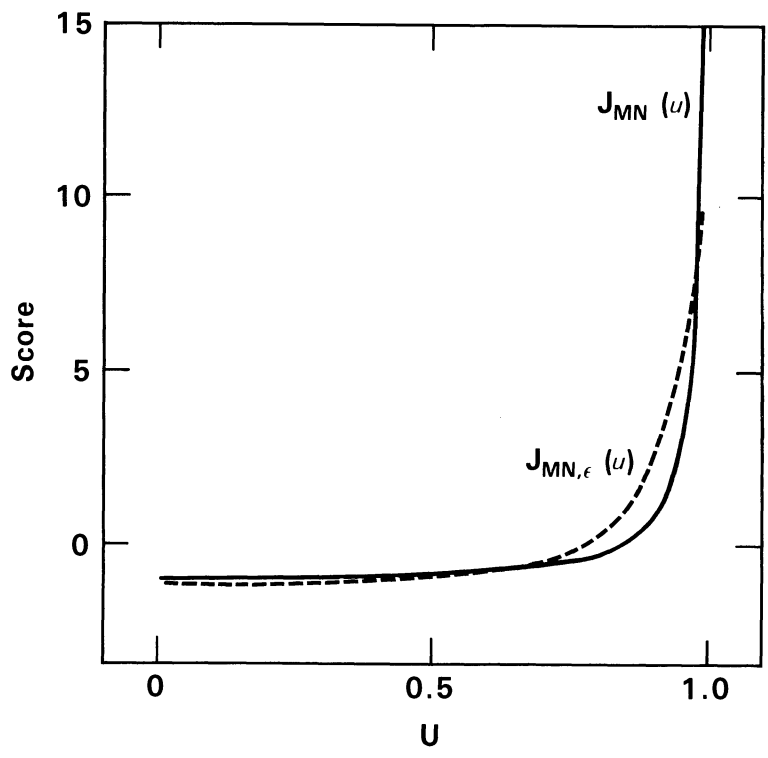

Figure 2. The score functions $J_{MN}(u)$ and $J_{MN,\varepsilon}(u)$ with $\Delta = 2$ and $\varepsilon = .1$.

Bi-uniform In order to obtain a simple step-shaped score function, we consider densities that assume only two positive values. Let

$$f_1(x) = \begin{cases} \dfrac{b_1}{c - a}, & a < x \leq c \\ \dfrac{1 - b_1}{d - c}, & c < x \leq d \end{cases} \qquad f_2(x) = \begin{cases} \dfrac{b_2}{c - a}, & a < x \leq c \\ \dfrac{1 - b_2}{d - c}, & c < x \leq d \end{cases} \tag{2.5}$$

where $1 > b_1 > b_2$ so that $F_2(x) \leqslant F_1(x)$ for all x. Then

$$\frac{f_2(x)}{f_1(x)} - 1 = \begin{cases} \dfrac{b_2 - b_1}{b_1}, & a < x \leqslant c \\ \dfrac{b_1 - b_2}{1 - b_1}, & c < x < d \end{cases}$$

and since $F_1^{-1}(u) \in (a, c]$ for $u \in (0, b_1]$ and $F_1^{-1}(u) \in (c, d]$ for $u \in (b_1, 1]$,

$$\frac{f_2[F_1^{-1}(u)]}{f_1[F_1^{-1}(u)]} - 1 = \begin{cases} \dfrac{b_2 - b_1}{b_1}, & 0 < u \leqslant b_1 \\ \dfrac{b_1 - b_2}{1 - b_1}, & b_1 < u < 1 \end{cases}.$$

Adding $(b_1 - b_2)/b_1$ and then multiplying by $b_1(1 - b_1)/(b_1 - b_2)$, we obtain the equivalent bi-uniform (BU) score function

$$J_{\mathrm{BU}}(u) = \begin{cases} 0, & 0 < u \leqslant b_1 \\ 1, & b_1 < u < 1 \end{cases}.$$

If the approximate score statistic

$$\sum_{i=1}^{n} J_{\mathrm{BU}}\left(\frac{r_i}{m + n + 1}\right) \tag{2.6}$$

is used, only observations above the combined sample $100b_1$th percentile receive weight. Moreover, the largest order statistics are all given the same weight so the test statistic reduces to a count of the number of Y_i among the extreme order statistics. For instance, if $b_1 = .8$ with $m = 10$ and $n = 10$, 17 is the smallest rank for which $rank/(10 + 10 + 1) > .8$ so only the upper 4 order staistics are considered. For $b_1 = .5$, we obtain the median test. We name the test corresponding to (2.6) a *quantile test.*

Most importantly, the null distribution of statistic (2.6) is the hypergeometric. It can be calculated exactly. Let $c = m + n - [(m + n + 1)b_1]$ be the number of ranks receiving weight 1. Then, under the null hypothesis,

$$\Pr\left[\sum_{i=1}^{n} J_{\mathrm{BU}}\left(\frac{r_i}{m + n + 1}\right) = x\right] = \frac{\binom{m + n - c}{n - x}\binom{c}{x}}{\binom{m + n}{n}}. \tag{2.7}$$

Asymptotically, the test (2.7) is equivalent to the test where only observations above the fixed level $c = F_1^{-1}(b_1)$ are counted. This latter count is precisely the one used by Carrano and Moore (1982) in a two-sided test. Their apparent success provided the motivation for the current study.

Example We illustrate the calculation of a quantile rank statistic and a mixed normal rank statistic. The data are counts of the SCEs found in sets of ten cells obtained from a cancer patient. The first set of cells was obtained before the patient underwent chemotherapy. The second set was obtained just after the drug treatment was completed. We are interested in whether the chemotherapy increases SCE counts in some fraction of the cells. The counts for the pretreatment cells are

9, 9, 10, 10, 14, 14, 14, 15, 16, 20

and those for the posttreatment cells are

6, 10, 13, 15, 18, 21, 22, 23, 30, 37.

The combined order statistics are then

$$\underline{6},\ 9,\ 9,\ 10,\ 10,\ \underline{10},\ \underline{13},\ 14,\ 14,\ 14,\ 15,\ \underline{15},$$
$$16,\ \underline{18},\ 20,\ \underline{21},\ \underline{22},\ \underline{23},\ \underline{30},\ \underline{37}$$

and the r_i's are 1, 5, 7, 11.5, 14, 16, 17, 18, 19, and 20. Here we use *average ranks* in the presence of ties. We decided to use $b = .8$ for the bi-uniform statistic so only the top $10 + 10 - [(10 + 10 + 1).8] = 4$ observations receive weight 1.

Since $J_{\mathrm{BU}}[r_i/(m + n + 1)] = 0$ if $r_i \leqslant 16$,

$$\sum_{j=1}^{10} J_{\mathrm{BU}}\left(\frac{r_i}{m+n+1}\right) = 0 + 0 + 0 + 0 + 0 + 0 + 1 + 1 + 1 + 1 = 4.$$

This is the most extreme case and, by (2.7), H_0 would be rejected at the

$$\binom{16}{6}\binom{4}{4}\bigg/\binom{20}{10} = .043$$

significance level.

To calculate a mixed normal score, we must select a value for Δ. We take $\Delta = 1$ so

$$J_{\mathrm{MN}}(u) = e^{-\Delta^2/2}e^{\Delta\Phi^{-1}(u)} - 1 = e^{\Phi^{-1}(u)-1/2} - 1.$$

The smallest treatment observation, 6, has rank 1 so it receives score

$$\exp\left[\Phi^{-1}\left(\frac{r_i}{m+n+1}\right) - \frac{1}{2}\right] - 1 = \exp\left[\Phi^{-1}\left(\frac{1}{21}\right) - \frac{1}{2}\right] - 1 = -.8856.$$

The largest treatment observation, 37, is assigned score

$$\exp\left[\Phi^{-1}\left(\frac{20}{21}\right) - \frac{1}{2}\right] - 1 = 2.2168.$$

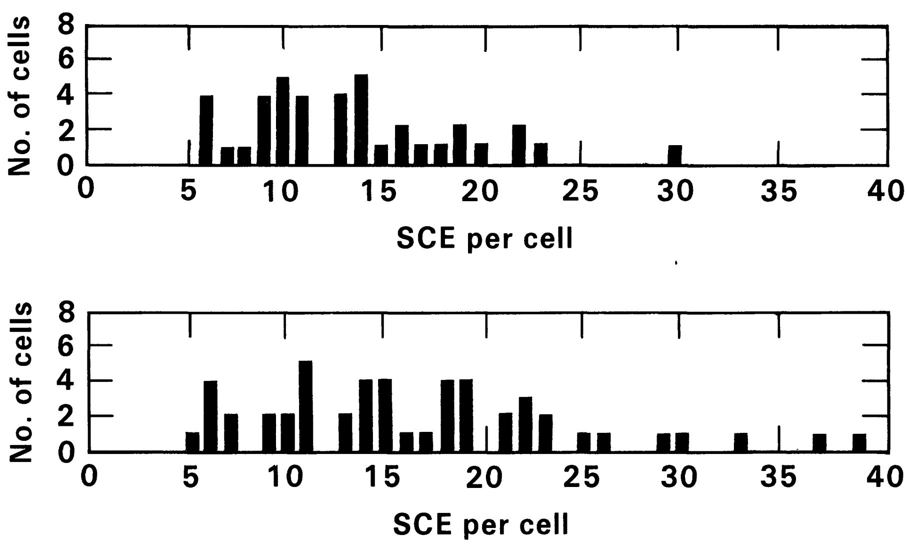

Figure 3. Distribution of SCEs prior to treatment (top) and after treatment (bottom).

For the intermediate values, average ranks are assigned in the presence of ties. Then

$$\sum_{j=1}^{10} J_{\mathrm{MN}}\left(\frac{r_i}{m+n+1}\right) = -.8856 + \cdots + 2.2168 = 2.3430.$$

When several ties are present, it is prudent to calculate the null distribution conditional on the tie structure. In our case, the observed statistic value, 2.3430, is greater than 182,113 of the $\binom{20}{10} = 184{,}756$ possible statistic values [the $\binom{20}{10}$ sums obtained by selecting and adding 10 of the scores $J_{\mathrm{MN}}(\frac{1}{21})$, $J_{\mathrm{MN}}(\frac{2.5}{21})$, $J_{\mathrm{MN}}(\frac{2.5}{21})$, $J_{\mathrm{MN}}(\frac{5}{21})$, $J_{\mathrm{MN}}(\frac{5}{21})$, $J_{\mathrm{MN}}(\frac{5}{21})$, $J_{\mathrm{MN}}(\frac{7}{21})$, $\ldots$, $J_{\mathrm{MN}}(\frac{19}{21})$, $J_{\mathrm{MN}}(\frac{20}{21})$]. Thus, the P-value is .014. Using the Nijenhuis and Wilf (1975) NEXKSB algorithm to do the subset selection, a Cray 1 computer generated all $\binom{20}{10}$ sums in 2 seconds.

In the study from which this data set was selected, samples of size 50 are now being collected in order for the procedure to give reasonable power for monitoring damage introduced by the rigorous drug treatment. Figure 3 illustrates a representative set of data.

3. Large-Sample Results

3.1 *Distribution*

In a large-sample context, there is no difference between the test statistic based on exact scores,

$$\sum_{i=1}^{n} \mathrm{E}\left[\frac{f_2(V^{(r_i)})}{f_1(V^{(r_i)})} - 1\right] \tag{3.1}$$

and that based on the approximate scores

$$\sum_{i=1}^{n} J\left(\frac{r_i}{m+n+1}\right), \tag{3.2}$$

where

$$J(u) = \frac{f_2[F_1^{-1}(u)]}{f_1[F_1^{-1}(u)]} - 1.$$

Theorem 3.1 gives the asymptotic distribution of both statistics.

Theorem 3.1 Let $J(u)$ be monotone and $\int_{-\infty}^{\infty} [f_2^2(x)/f_1(x)]\, dx < \infty$. If $\min(m, n) \to \infty$, under H_0, both

$$\sum_{i=1}^{n} \mathrm{E}\left[\frac{f_2(V^{(r_i)})}{f_1(V^{(r_i)})} - 1\right]$$

and

$$S_{m+n} = \sum_{i=1}^{n} J\left(\frac{r_i}{m+n+1}\right) \tag{3.3}$$

are asymptotically normal with mean 0 and variance

$$\sigma_{mn}^2 = \frac{mn}{m+n}\int_0^1 J^2(u)\, du = \frac{mn}{m+n}\int_{-\infty}^{\infty} \frac{[f_2(x) - f_1(x)]^2}{f_1(x)}\, dx.$$

Proof See Appendix A.2.

3.2 *Efficiency*

The locally most powerful rank test depends on $f_1(x)$ and $f_2(x)$. However, as these densities are likely to be difficult to specify accurately, it is of interest to study the loss of efficiency due to inaccurate specifications.

We denote the statistic actually used by

$$T_{m+n} = \sum_{i=1}^{n} J_T\left(\frac{r_i}{m+n+1}\right). \tag{3.4}$$

In Theorem 3.2 we compare the asymptotic behavior of T_{m+n} with that of S_{m+n} [which employs the optimal score function given $f_1(x)$ and $f_2(x)$].

Theorem 3.2 If $\int_0^1 J_T^2(u)\,du < \infty$, $\int_{-\infty}^{\infty} [f_2^2(x)/f_1(x)]\,dx < \infty$, and S_{m+n} is given by (3.3), the asymptotic relative efficiency of T_{m+n} is given by

$$e_{T:S} = \frac{\left(\int_0^1 J_T(u)\left\{\frac{f_2[F_1^{-1}(u)]}{f_1[F_1^{-1}(u)]} - 1\right\} du\right)^2}{\int_0^1 [J_T(u) - \bar{J}_T]^2\,du \int_0^1 \left\{\frac{f_2[F_1^{-1}(u)]}{f_1[F_1^{-1}(u)]} - 1\right\}^2 du},$$

where

$$\bar{J}_T = \int_0^1 J_T(u)\,du.$$

Proof See Appendix A.3.

In Tables 1–3 we present some calculations based on Theorem 3.2. We use the notation W: Wilcoxon, MN: Mixed normal, NS: Normal scores, BU: Bi-uniform.

From the numerical values in Table 1 we see that the standard tests for location, the Wilcoxon and normal scores, do not perform well under mixture alternatives. When the underlying population is the mixed normal with $\Delta = (\mu_2 - \mu_1)/\sigma \geq 1.5$, all of the efficiencies are below 26.5%. Under local changes in the bi-uniform, the asymptotic relative efficiencies of both the Wilcoxon and normal scores tests are below 75%. They are much lower for alternatives with large b.

Table 2 gives the asymptotic relative efficiency of the mixed normal statistic test relative to the bi-uniform. The efficiency is symmetric in the two distributions. Therefore, if the population is bi-uniform with $b = .7$, the mixed normal statistic with $\Delta = .5$ has asymptotic relative efficiency .607. The same efficiency applies to the bi-uniform statistic under a mixed normal population. From Table 2 we can see that the mixed normal test is quite different from the bi-uniform.

Table 1
Asymptotic relative efficiency of the Wilcoxon and normal scores tests

Mixed normal population			Bi-uniform population		
Δ	$e_{\mathrm{W:MN}}$	$e_{\mathrm{NS:MN}}$	b	$e_{\mathrm{W:BU}}$	$e_{\mathrm{NS:BU}}$
.5	.736	.880	.5	.750	.637
1.0	.394	.592	.6	.720	.622
1.5	.112	.265	.7	.630	.576
2.0	.015	.075	.8	.480	.490
2.5	.001	.012	.9	.270	.342
3.0	.000	.001	.95	.143	.224

Table 2
Asymptotic relative efficiency of the mixed normal scores test relative to the bi-uniform scores test

		Bi-uniform b				
	Δ	.1	.3	.5	.7	.9
Mixed normal	.5	.153	.363	.516	.607	.538
	1.0	.051	.155	.271	.406	.541
	1.5	.012	.044	.088	.161	.310
	2.0	.002	.008	.017	.035	.091

Table 3
Asymptotic relative efficiency of the mixed normal test

		Mixed normal statistic Δ_1				
	Δ_2	.5	1.0	1.5	2.0	2.5
Mixed normal population	.5		.860	.518	.194	.042
	1.0			.831	.443	.141
	1.5				.801	.393
	2.0					.784

Finally, suppose that a mixed normal test is correctly specified but that an incorrect choice is made for Δ. Table 3 shows that if Δ can be specified within .5 unit, the loss of efficiency is not too great.

4. Empirical Power Study

The asymptotic relative efficiencies described above indicate that there are mixture alternatives under which the Wilcoxon and normal scores tests do not fare as well as ours. In an attempt to confirm this result for small and moderate sample sizes, we performed an empirical power study, under the mixed normal alternative

$$H_0\colon F(x) = \Phi(x) \quad \text{vs} \quad H_1\colon F(x) = (1 - \varepsilon)\Phi(x) + \varepsilon\Phi(x - \Delta)$$

with $\varepsilon = .1(.1).3$ and $\Delta = 1.0(1.0)3.0$. Power depends only on ε and $\Delta = (\mu_2 - \mu_1)/\sigma$ so unit variance normal distributions were selected.

Eight test statistics were calculated:

W: $J(u) = u$	MN(.5): J_{MN} with $\Delta = .5$
NS: $J(u) = \Phi^{-1}(u)$	MN(1.0): J_{MN} with $\Delta = 1.0$
BU(.8): $J_{\mathrm{BU}}(u) = 1,\ u > .8$	MN(1.5): J_{MN} with $\Delta = 1.5$
BU(.9): $J_{\mathrm{BU}}(u) = 1,\ u > .9$	MN(2.0): J_{MN} with $\Delta = 2.0$

The results of our power study appear in Table 4.

For $m = n = 5$ and $m = n = 10$, critical values were obtained by calculating the exact null distribution. We selected significance levels close to $\alpha = .05$. Because of the discreteness of distributions, without randomization, the closest levels that could be attained for the bi-uniform when $m = n = 5$ were .222 and .500 [see (2.7)]. We can, however, attain $\alpha = .05$ for the BU(.8) test if we randomly reject the null hypothesis (.05/.222)100% = 22.5% of the time when the second sample has the top two ranks, 9 and 10. The corresponding power is obtained by multiplying the entries in the $m = n = 5$, BU(.8) row in Table 4 by .225. Similar calculations can be made for the BU(.9) test.

Table 4
Empirical power under mixed normal alternatives

Sample size	Test	H_0 ε 0	Alternative $\Delta = 1$ ε .1	.2	.3	$\Delta = 2$ ε .1	.2	.3	$\Delta = 3$ ε .1	.2	.3
$m = n = 5$											
	W	.048	.047	.056	.057	.093	.105	.145	.078	.123	.172
	NS	.048	.047	.056	.057	.093	.105	.145	.078	.123	.172
	BU(.8)	.222	.236	.271	.271	.328	.452	.545	.337	.499	.622
	BU(.9)	.500	.524	.555	.579	.624	.757	.816	.679	.815	.878
	MN(.5)	.048	.046	.059	.061	.096	.107	.153	.085	.134	.187
	MN(1.0)	.048	.047	.056	.061	.090	.117	.164	.083	.140	.203
	MN(1.5)	.048	.042	.048	.058	.085	.111	.148	.075	.126	.195
	MN(2.0)	.048	.047	.055	.061	.089	.123	.159	.085	.135	.214
$m = n = 10$											
	W	.053	.083	.129	.172	.090	.147	.253	.096	.172	.320
	NS	.050	.079	.128	.166	.092	.154	.247	.096	.178	.320
	BU(.8)	.043	.066	.104	.140	.102	.179	.297	.096	.252	.415
	BU(.9)	.237	.292	.375	.468	.416	.560	.685	.505	.725	.835
	MN(.5)	.050	.084	.125	.175	.095	.179	.291	.103	.214	.379
	MN(1.0)	.050	.085	.118	.167	.105	.196	.318	.108	.242	.422
	MN(1.5)	.050	.084	.119	.166	.113	.206	.326	.112	.252	.431
	MN(2.0)	.050	.084	.117	.165	.115	.200	.327	.104	.259	.435
$m = n = 20$											
	W	.055	.099	.159	.241	.112	.227	.431	.113	.253	.491
	NS	.055	.099	.163	.224	.134	.247	.458	.137	.292	.524
	BU(.8)	.021	.041	.071	.107	.070	.153	.367	.087	.234	.552
	BU(.9)	.056	.094	.148	.182	.180	.324	.545	.257	.581	.813
	MN(.5)	.056	.102	.172	.250	.155	.312	.553	.167	.398	.682
	MN(1.0)	.057	.103	.177	.245	.170	.343	.606	.215	.471	.755
	MN(1.5)	.053	.104	.163	.233	.181	.344	.609	.239	.525	.795
	MN(2.0)	.049	.097	.156	.209	.163	.337	.574	.234	.519	.804
$m = n = 40$											
	W	.058	.110	.202	.329	.170	.360	.661	.171	.460	.775
	NS	.056	.114	.201	.340	.194	.399	.689	.206	.524	.828
	BU(.8)	.086	.179	.266	.389	.268	.564	.821	.294	.703	.993
	BU(.9)	.032	.059	.130	.197	.194	.417	.699	.276	.749	.962
	MN(.5)	.057	.131	.233	.384	.247	.529	.813	.283	.697	.922
	MN(1.0)	.055	.131	.249	.392	.282	.609	.870	.368	.819	.969
	MN(1.5)	.053	.121	.242	.378	.305	.637	.872	.434	.866	.986
	MN(2.0)	.049	.108	.216	.344	.290	.626	.850	.461	.887	.988

With larger sample sizes, $m = n = 20$ and $m = n = 40$, we used the normal approximation (see Appendix A.3)

$$n\bar{J} + 1.96\left\{\frac{mn}{(m+n)(m+n+1)}\sum_{i=1}^{m+n}\left[J\left(\frac{i}{m+n+1}\right) - \bar{J}\right]^2\right\}^{1/2}$$

to the critical value, where

$$\bar{J} = \frac{\sum_{i=1}^{m+n} J[i/(m+n+1)]}{m+n}.$$

A total of 6000 trials were conducted to estimate the actual probabilities with which the test statistics exceeded these normal approximations to their critical values. The estimated probabilities for the BU(.8) and BU(.9) statistics were checked to be within sampling variation of the exact hypergeometric probabilities (2.7).

The estimated powers in Table 4 are based on 1000 trials. Each sample size took about 1 minute of CPU time on a Cray 1. From Table 4, it is apparent that moderate sample sizes are required in order to obtain reasonable power. The power is quite high for the mixed normal and, to some degree, the quantile statistic when $m = n \geqslant 20$. For $\Delta = 2$ or 3, the tests designed to detect mixing have more power than the Wilcoxon or normal scores tests.

The most striking feature of the empirical power study is that small sample sizes do not yield much power. When the mixing parameter, ε, is small, only a few observations are available from $F_2(x) = \Phi(x - \Delta)$ and detection is almost impossible. This must be kept in mind when selecting sample sizes for comparative tests under alternatives of the form (1.1).

For larger sample sizes Table 4 tends to confirm the general picture suggested by the asymptotic relative efficiencies.

Comparing Table 4 to Table 1 of Good (1979), we see that, for small samples, Good's statistic has power comparable to the asymptotically most powerful mixed normal scores rank tests.

5. Extension

Sometimes even the control population exhibits a small amount of mixing. In the context of SCE counts, the distribution of control measurements might have a few elevated values caused by extraneous factors. This indeed seems to occur in settings similar to those illustrated above. In this section we obtain the optimal rank tests for this more general setting. From the form of the test statistics derived here, we also gain some understanding of the sensitivity of the previous optimal rank tests to the presence of mixing under the null hypothesis.

We now suppose that the control population has the cumulative distribution function

$$F_\varepsilon(x) = (1 - \varepsilon)F_1(x) + \varepsilon F_2(x)$$

where $0 \leqslant \varepsilon < 1$ and $F_1(x)$ and $F_2(x)$ are absolutely continuous. In order to specify a direction we assume that $F_2(x)$ is stochastically larger than $F_1(x)$, i.e., $F_2(x) \leqslant F_1(x)$ for all x. The alternative is that $F_2(x)$ contributes more to the mixture. That is,

$$G_\varepsilon(x, \delta) = (1 - \varepsilon - \delta)F_1(x) + (\varepsilon + \delta)F_2(x) = F_\varepsilon(x) + \delta[F_2(x) - F_1(x)]. \quad (5.1)$$

Here ε indicates the mixing proportion under H_0 and $\varepsilon + \delta$ is the proportion under the alternative. These alternatives lie within the general class of stochastically larger distributions. However, because the parameter δ enters in a nonstandard manner, our test statistics will involve unusual scores.

Theorem 5.1 Let $X_1, \ldots, X_m$ be a random sample from $F_\varepsilon(x)$, and $Y_1, \ldots, Y_n$ be an independent random sample from $G_\varepsilon(x, \delta)$. Assume that F_1 and F_2 have densities f_1 and f_2. The locally most powerful rank test for testing H_0: $\delta = 0$ versus H_1: $\delta \neq 0$ rejects H_0 for large values of

$$\sum_{i=1}^{n} \mathrm{E}\left[\frac{f_2(V^{(r_i)}) - f_1(V^{(r_i)})}{(1 - \varepsilon)f_1(V^{(r_i)}) + \varepsilon f_2(V^{(r_i)})}\right],$$

where $V^{(1)} \leqslant \cdots \leqslant V^{(m+n)}$ is an ordered sample from $F_\varepsilon(x)$ and $r_1 \leqslant \cdots \leqslant r_n$ are the ranks of $Y_1, \ldots, Y_n$ in the combined sample.

Proof See Appendix A.1.

We consider two special cases of our test.

Bi-uniform With $f_1(x)$ and $f_2(x)$ defined by (2.5), the score function for the locally most powerful test becomes

$$\frac{\dfrac{f_2[F_\varepsilon^{-1}(u)]}{f_1[F_\varepsilon^{-1}(u)]}-1}{1+\varepsilon\left\{\dfrac{f_2[F_\varepsilon^{-1}(u)]}{f_1[F_\varepsilon^{-1}(u)]}-1\right\}}=\begin{cases}\dfrac{b_2-b_1}{b_1+\varepsilon(b_2-b_1)}, & 0<u\leqslant(1-\varepsilon)b_1+\varepsilon b_2\\ \dfrac{b_1-b_2}{1-b_1+\varepsilon(b_1-b_2)}, & (1-\varepsilon)b_1+\varepsilon b_2<u<1\end{cases}$$

Equivalently, we can take

$$J_{\mathrm{BU},\varepsilon}(u)=\begin{cases}0, & 0<u\leqslant(1-\varepsilon)b_1+\varepsilon b_2\\ 1, & (1-\varepsilon)b_1+\varepsilon b_2<u<1\end{cases}. \tag{5.2}$$

It is somewhat surprising that, for $\varepsilon>0$, the score function has the same structure as in the $\varepsilon=0$ case. That is, the test statistic is still the number of second sample observations above an upper quantile of the combined sample.

Mixture of normal distributions Let $f_1(x)$ be $\mathrm{N}(\mu_1,\sigma^2)$ and $f_2(x)$ be $\mathrm{N}(\mu_2,\sigma^2)$ with $\mu_2>\mu_1$. We first note that if X is distributed as $F_\varepsilon(x)$,

$$\Pr\left(\frac{X-(\mu_1+\mu_2)/2}{\sigma}\leqslant y\right)=(1-\varepsilon)\Phi\left(y+\frac{\Delta}{2}\right)+\varepsilon\Phi\left(y-\frac{\Delta}{2}\right)\equiv F_{\varepsilon,\Delta}(y), \tag{5.3}$$

where $\Delta=(\mu_2-\mu_1)/\sigma$ and $\Phi(\cdot)$ is the standard normal cdf. The locally most powerful test depends on

$$\frac{f_2(x)-f_1(x)}{(1-\varepsilon)f_1(x)+\varepsilon f_2(x)}=\frac{f_2(x)/f_1(x)-1}{1+\varepsilon[f_2(x)/f_1(x)-1]}.$$

From (5.3) we have $[F_\varepsilon^{-1}(u)-(\mu_1+\mu_2)/2]/\sigma=F_{\varepsilon,\Delta}^{-1}(u)$. So, using (2.3), we can write the score function as

$$J_{\mathrm{MN},\varepsilon}(u)=\frac{\exp[\Delta F_{\varepsilon,\Delta}^{-1}(u)]-1}{1+\varepsilon\{\exp[\Delta F_{\varepsilon,\Delta}^{-1}(u)]-1\}}. \tag{5.4}$$

This function (which depends on μ_1, μ_2, and σ only through Δ) is bounded below by $-1/(1-\varepsilon)$ and increases monotonically toward ε^{-1} (see Figure 2). Unlike the score (2.4), the assumed presence of some mixing under H_0 keeps $J_{\mathrm{MN},\varepsilon}$ bounded as u approaches 1.

We state the following large-sample results whose proofs are given in the Appendix.

Let

$$J_\varepsilon(u)=\frac{f_2[F_\varepsilon^{-1}(u)]-f_1[F_\varepsilon^{-1}(u)]}{(1-\varepsilon)f_1[F_\varepsilon^{-1}(u)]+\varepsilon f_2[F_\varepsilon^{-1}(u)]}.$$

The test based on the exact scores

$$\sum_{i=1}^{n}\mathrm{E}\left[\frac{f_2(V^{(r_i)})-f_1(V^{(r_i)})}{(1-\varepsilon)f_1(V^{(r_i)})+\varepsilon f_2(V^{(r_i)})}\right] \tag{5.5}$$

and that based on the approximate scores

$$S_{m+n,\varepsilon}=\sum_{i=1}^{n}J_\varepsilon\left(\frac{r_i}{m+n+1}\right) \tag{5.6}$$

are asymptotically equivalent, as described by Theorem 5.2.

Theorem 5.2 Let $J_\varepsilon(u)$ be monotone. If $\min(m,n)\to\infty$, under H_0, both

$$\sum_{i=1}^{n}\mathrm{E}\left[\frac{f_2(V^{(r_i)})-f_1(V^{(r_i)})}{(1-\varepsilon)f_1(V^{(r_i)})+\varepsilon f_2(V^{(r_i)})}\right]$$

and $S_{m+n,\varepsilon}$ are asymptotically normal with mean 0 and variance

$$\sigma^2_{mn,\varepsilon} = \frac{mn}{m+n} \int_{-\infty}^{\infty} \frac{[f_2(x) - f_1(x)]^2}{(1-\varepsilon)f_1(x) + \varepsilon f_2(x)}\, dx.$$

The locally most powerful rank test depends on ε, $f_1(x)$, and $f_2(x)$. However, as these three quantities are likely to be difficult to specify accurately, it is again of interest to study the loss of efficiency due to inaccurate specification.

We denote the statistic actually used by

$$T_{m+n} = \sum_{i=1}^{n} J_T\left(\frac{r_i}{m+n+1}\right). \tag{5.7}$$

In Theorem 5.3, the asymptotic behavior of T_{m+n} is compared with that of $S_{m+n,\varepsilon}$, which employs the optimal score function given ε, $f_1(x)$, and $f_2(x)$.

Theorem 5.3 If $\int_0^1 J_T^2(u)\, du < \infty$ and $S_{m+n,\varepsilon}$ is given by (5.6), the asymptotic relative efficiency of T_{m+n} is given by

$$e_{T:S,\varepsilon} = \frac{\left[\int_0^1 J_T(u) J_\varepsilon(u)\, du\right]^2}{\int_0^1 [J_T(u) - \bar{J}_T]^2\, du \int_0^1 J_\varepsilon^2(u)\, du},$$

where

$$\bar{J}_T = \int_0^1 J_T(u)\, du.$$

Résumé

Dans l'étude d'un phénomène que l'on suppose précéder une rupture chromosomique, on a trouvé que, quelquefois, les traitements fournissent davantage d'observations suspectes que les contrôles. Nos exemples concernent des fumeurs et des malades soumis à la chimiothérapie, mais leur caractère est beaucoup plus général.

Le problème est formulé dans le cadre non paramétrique. Les tests de rang localement les plus puissants sont obtenus sous des alternatives de mélange. Dans un cas, le test des scores approché a une forme simple, qui consiste à décompter le nombre de réponses du traitement au dessus d'un percentile obtenu à partir des échantillons fusionnés.

On compare cette statistique de test à ceux de Wilcoxon et des scores Normaux, en utilisant des études empiriques de puissance et d'efficacités asymptotiques.

References

Bhattacharyya, G. and Johnson, R. A. (1973). On the robustness of rank tests in life testing. *Journal of Statistical Computation and Simulation* **2,** 159–169.

Bickel, P. and Doksum, K. (1969). Tests for monotone failure rate based on normalized spacings. *Annals of Mathematical Statistics* **40,** 1216–1235.

Carrano, A. and Moore, D. (1982). The rationale and methodology for quantifying sister chromatid exchange in humans. In *Mutagenicity: New Horizons in Genetic Toxicology*, J. A. Heddle (ed.), 268–304. New York: Academic Press.

Good, P. I. (1979). Detection of a treatment effect when not all experimental subjects will respond to treatment. *Biometrics* **35,** 483–489.

Hajek, J. and Šidák, Z. (1967). *Theory of Rank Tests.* New York: Academic Press.

Lehmann, E. L. (1953). The power of rank tests. *Annals of Mathematical Statistics* **24,** 23–43.

Nijenhuis, A. and Wilf, H. S. (1975). *Combinational Algorithms.* New York: Academic Press.

Received August 1985; *revised August* 1986 *and January* 1987.

APPENDIX

Derivations

The theory of rank tests is thoroughly explored in Hajek and Šidák (1967). We prove our results by verifying their conditions.

A.1 *Proof of Theorem* 2.1 *and Theorem* 5.1

We exploit the relationship between Theorem 2.1 and Theorem 5.1 to give a single proof. We prove Theorem 5.1 first.

Refer to the general conditions in Hajek and Šidák (1967, pp. 70–73). In their notation, we verify that, for all δ such that $|\delta| \leqslant \min(\varepsilon, 1-\varepsilon)$, (i) the density

$$d(x, \delta) = (1 - \varepsilon - \delta)f_1(x) + (\varepsilon + \delta)f_2(x)$$

is absolutely continuous in δ, for each x, with Radon–Nikodym derivative $\dot{d}(x, \delta)$; (ii) the limit

$$\lim_{\delta \to 0} \frac{d(x, \delta) - d(x, 0)}{\delta} = \dot{d}(x, 0) = f_2(x) - f_1(x)$$

exists for almost all x; and (iii)

$$\int_{-\infty}^{\infty} |\dot{d}(x, \delta)|\, dx = \int_{-\infty}^{\infty} |f_2(x) - f_1(x)|\, dx = \int_{-\infty}^{\infty} |\dot{d}(x, 0)|\, dx < \infty.$$

It follows that the locally most powerful test is based on the test statistic

$$\sum_{i=1}^{n} \mathrm{E}\left[\frac{\dot{d}(V^{(r_i)}, 0)}{d(V^{(r_i)}, 0)}\right] = \sum_{i=1}^{n} \mathrm{E}\left[\frac{f_2(V^{(r_i)}) - f_1(V^{(r_i)})}{(1 - \varepsilon)f_1(V^{(r_i)}) + \varepsilon f_2(V^{(r_i)})}\right],$$

where $V^{(1)} \leqslant \cdots \leqslant V^{(m+n)}$ is an ordered sample from $F_\varepsilon(x)$. (Here, the N in the statement of Hajek and Šidák's theorem is $m + n$, and the c_i's in the statement of their theorem are 0's for $X_1, \ldots, X_m$ and 1's for $X_{m+1}, \ldots, X_{m+n} = Y_1, \ldots, Y_n$.)

If $\varepsilon = 0$, we restrict δ to a one-sided interval $[0, \delta_0]$. An examination of Hajek and Šidák's proof reveals that it remains valid under this restriction if $d(x, 0) = 0$ implies $\dot{d}(x, 0) = 0$ almost surely. Since the support of F_2 is contained in the support of F_1, $d(x, 0) = f_1(x) = 0$ does imply $\dot{d}(x, 0) = f_2(x) - f_1(x) = 0$ almost surely. Consequently, the test statistic reduces to

$$\sum_{i=1}^{n} \mathrm{E}\left[\frac{f_2(V^{(r_i)})}{f_1(V^{(r_i)})} - 1\right],$$

where $V^{(1)} \leqslant \cdots \leqslant V^{(m+n)}$ is an ordered sample from $F_1(x)$.

A.2 *Proofs of Theorem* 3.1 *and Theorem* 5.2

We give the proof for Theorem 5.2 followed by the proof for Theorem 3.1.

We verify the conditions of Hajek and Šidák (1967, p. 161, Theorem b). Let $\varepsilon > 0$. The exact scores in (5.1) can be written as $\mathrm{E}[J_\varepsilon(U_1) \mid \mathrm{rank}(U_1) = i]$. Under the change of variable $x = F_\varepsilon^{-1}(u)$, we establish that

$$\bar{J}_\varepsilon = \int_0^1 J_\varepsilon(u)\, du = \int_{-\infty}^{\infty} [f_2(x) - f_1(x)]\, dx = 0,$$

and

$$\int_0^1 J_\varepsilon^2(u)\, du = \int_{-\infty}^{\infty} \frac{[f_2(x) - f_1(x)]^2}{(1 - \varepsilon)f_1(x) + \varepsilon f_2(x)}\, dx < \infty,$$

since the integrand is dominated by $(1 - \varepsilon)^{-1}f_1(x) + \varepsilon^{-1}f_2(x)$. Because $F_1(x) \neq F_2(x)$, this integral is positive. Asymptotic normality follows.

The equivalence of (5.5) and (5.6) follows for monotone $J_\varepsilon(u)$ by Hajek and Šidák (1967, pp. 163–165, Theorem a and Lemma a). For the case covered in Theorem 3.1 the exact scores can be written

$$\mathrm{E}[J_0(U_1) \mid U_1 = i] \quad \text{where} \quad J_0(u) = \frac{f_2[F_1^{-1}(u)]}{f_1[F_1^{-1}(u)]} - 1.$$

Under the transformation $x = F_1^{-1}(u)$ we verify

$$\int_0^1 J_0(u)\,du = \int_{-\infty}^{\infty}\left[\frac{f_2(x)}{f_1(x)} - 1\right] f_1(x)\,dx = 0,$$

and

$$\int_0^1 J_0^2(u)\,du = \int_{-\infty}^{\infty}\frac{[f_2(x) - f_1(x)]^2}{f_1^2(x)} f_1(x)\,dx = \int_{-\infty}^{\infty}\frac{f_2^2(x)}{f_1(x)}\,dx - 1 < \infty,$$

thus establishing asymptotic normality. The equivalence of (3.1) and (3.2) for monotone $J(u)$ follows as above.

A.3 *Proofs of Theorem* 3.2 *and Theorem* 5.3

We first prove Theorem 5.3 and then Theorem 3.2. We require limiting distributions under the sequence of local alternatives (5.1) with $\delta = h/\sqrt{m+n}$. Asymptotic normality will follow when we verify the three conditions of Hajek and Šidák (1967, pp. 238–239). Conditions (i) and (ii) were established in the proof of Theorem 5.1. In the same notation

$$\frac{[\dot{d}(x, \delta)]^2}{d(x, \delta)} \leqslant 2(1-\varepsilon)^{-1} f_1(x) + 2\varepsilon^{-1} f_2(x),$$

for $|\delta| < \min(\varepsilon, 1-\varepsilon)/2$, so by the dominated convergence theorem,

$$\lim_{\delta\to 0}\int_{-\infty}^{\infty}\frac{[\dot{d}(x, \delta)]^2}{d(x, \delta)}\,dx = \int_{-\infty}^{\infty}\frac{[f_2(x) - f_1(x)]^2}{(1-\varepsilon) f_1(x) + \varepsilon f_2(x)}\,dx,$$

thus verifying condition (iii). We conclude that $T_{m+n} - n\bar{J}_{m+n,T}$ is asymptotically normal with mean

$$\mu_T = \frac{mnh}{(m+n)^{3/2}}\int_0^1 J_T(u) J_\varepsilon(u)\,du$$

and variance

$$\sigma_T^2 = \frac{mn}{(m+n)}\int_0^1 (J_T - \bar{J}_T)^2(u)\,du,$$

where

$$\bar{J}_{m+n,T} = \frac{\sum_{i=1}^{m+n} J_T[i/(m+n+1)]}{m+n} \quad \text{and} \quad \bar{J}_T = \int_0^1 J_T(u)\,du.$$

We also conclude that $S_{m+n,\varepsilon} - n\bar{J}_{m+n,S}$ is asymptotically normal with mean

$$\mu_S = \frac{mnh}{(m+n)^{3/2}}\int_0^1 J_\varepsilon^2(u)\,du$$

and variance

$$\sigma_S^2 = \frac{mn}{(m+n)}\int_0^1 J_\varepsilon^2(u)\,du.$$

Consequently (see Hajek and Šidák, 1967, Chap. VII, §2.1), the efficiency expression is

$$e_{T:S,\varepsilon} = \left(\frac{\mu_T \sigma_S}{\mu_S \sigma_T}\right)^2 = \frac{\left[\int_0^1 J_T(u) J_\varepsilon(u)\,du\right]^2}{\int_0^1 [J_T(u) - \bar{J}]^2\,du \int_0^1 J_\varepsilon^2(u)\,du}.$$

For Theorem 3.2 where $\varepsilon = 0$, we can restrict δ to be nonnegative. Condition (iii) again follows by the dominated convergence theorem since, for $\delta < \frac{1}{2}$,

$$\frac{[\dot{d}(x, \delta)]^2}{d(x, \delta)} = \frac{[f_2(x) - f_1(x)]^2}{(1-\delta) f_1(x) + \delta f_2(x)} \leqslant \frac{2 f_2^2(x)}{f_1(x)} + 2 f_1(x).$$