

# Circulation

JOURNAL OF THE AMERICAN HEART ASSOCIATION



## Rank Score Tests

Lisa M. LaVange and Gary G. Koch

*Circulation* 2006;114:2528-2533

DOI: 10.1161/CIRCULATIONAHA.106.613638

Circulation is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75214

Copyright © 2006 American Heart Association. All rights reserved. Print ISSN: 0009-7322. Online ISSN: 1524-4539

The online version of this article, along with updated information and services, is located on the World Wide Web at:

<http://circ.ahajournals.org/cgi/content/full/114/23/2528>

Data Supplement (unedited) at:

<http://circ.ahajournals.org/cgi/content/full/114/23/2528/DC1>

Subscriptions: Information about subscribing to *Circulation* is online at  
<http://circ.ahajournals.org/subscriptions/>

Permissions: Permissions & Rights Desk, Lippincott Williams & Wilkins, a division of Wolters Kluwer Health, 351 West Camden Street, Baltimore, MD 21202-2436. Phone: 410-528-4050. Fax: 410-528-8550. E-mail:  
[journalpermissions@lww.com](mailto:journalpermissions@lww.com)

Reprints: Information about reprints can be found online at  
<http://www.lww.com/reprints>

## Rank Score Tests

Lisa M. LaVange, PhD; Gary G. Koch, PhD

Nonparametric statistical methods are useful tools for data analysis when there is reason to believe that the outcome variables of interest may fail certain distributional assumptions required for parametric methods. Variables may be ordered categories in nature and thereby not suitable for analysis methods that assume normally distributed variables, such as *t* tests or analyses of variance and covariance. Variables may also be metric or continuous but subject to excessive variability or the presence of outliers. When the research hypothesis involves comparing a sample of subjects under 2 conditions or at 2 time points or comparing 2 samples of subjects with respect to an outcome variable of interest, then univariate nonparametric methods based on rank score tests can be invoked. A study design feature such as random assignment of conditions or treatments is typically all that is required for these methods to be valid. Furthermore, the methods can be quite powerful under a number of alternatives, particularly those involving shifts in the median.

The Wilcoxon signed rank test, the Spearman rank correlation coefficient, and the Wilcoxon rank sum test are among the most commonly used nonparametric tests and cover a variety of research questions. These tests are described here. Although the focus is on hypothesis testing, related methods for estimation of confidence intervals are also presented. Extensions of nonparametric methods to handle stratification and covariate adjustment are also described. Scenarios in which nonparametric methods may be most useful and the power they can be expected to yield are discussed. The methods are illustrated with data from a clinical trial assessing the impact of exposure to low levels of carbon monoxide on exercise capacity in patients with ischemic heart disease.

### Wilcoxon Signed Rank Test

When the response variable of interest is a metric measurement that follows a symmetrical distribution with substantial variability or outliers, then the Wilcoxon signed rank test<sup>1,2</sup> is a useful test for differences between paired samples of subjects or paired conditions on a sample of subjects. For a sample of size *n*, let *i*=1, . . . ,*n* identify the subjects in the sample, let *X<sub>i</sub>* and *Y<sub>i</sub>* denote the observed values of the outcome variable under the 2 conditions or at the 2 time points of measurement for the *i*th subject, and let *d<sub>i</sub>* denote the difference (*X<sub>i</sub>*−*Y<sub>i</sub>*). If  $\Delta$  represents the median of the distribution

of difference values {*d<sub>i</sub>*}, then the null hypothesis is that  $\Delta=0$ . That is, the null hypothesis is the hypothesis that the true, underlying difference between the 2 conditions is zero. The following steps describe the calculation of the test statistic to assess the null hypothesis against the alternative of a shift in location associated with one of the conditions or time points:

1. Differences {*d<sub>i</sub>*} are formed for each pair of observations on each subject.
2. The ranks of the absolute values of the nonzero differences are computed, with ties assigned the average of the applicable ranks (called midranks); zero differences are ignored.
3. The signs of the differences are computed as −1 or 1.
4. Signed ranks are then computed by multiplying the sign of the difference by the corresponding rank.
5. The test statistic is calculated by dividing the sum of the signed ranks *U* by the square root of the sum of squares of the signed ranks *S* to form *U/S*; *S* is the standard deviation of *U*.

The null hypothesis that the median difference is zero is assessed by comparing the test statistic *U/S* to critical values of the standard normal distribution for large sample sizes (eg, *n*≥20) or by tabulating the exact critical region for small sample sizes. Computation of *P* values from critical values for both the normal approximation and the exact distribution is available through commercial statistical software packages (eg, SAS Proc Univariate<sup>3</sup> and StatXact<sup>4</sup>).

When there are no ties among the observed differences and no differences equal to zero, then the signed rank test statistic simplifies to a commonly used form. If *T* denotes the sum of the positive ranks, then  $U=2T-n(n+1)/2$  and

$$S = \sqrt{n(n+1)(2n+1)/6}.$$

The significance of *U/S* is determined by comparison with a standard normal distribution or by computation of the exact critical region, as above.

If it is assumed that the distribution of the {*d<sub>i</sub>*} is continuous and symmetrical, a point estimate of the median difference  $\Delta$  is given by

$$\hat{\Delta} = \text{Median}[f_{ii} = (d_i + d_{i'})/2] \quad \text{for } 1 \leq i \leq i' \leq n.$$

From the Department of Biostatistics, School of Public Health, University of North Carolina, Chapel Hill.

The online-only Data Supplement, consisting of a technical appendix, is available with this article at <http://circ.ahajournals.org/cgi/content/full/114/23/2528/DC1>.

Correspondence to Lisa M. LaVange, PhD, Collaborative Studies Coordinating Center, Department of Biostatistics, CB 8030, School of Public Health, University of North Carolina, Chapel Hill, NC 27514-4145. E-mail [lisa\\_lavange@unc.edu](mailto:lisa_lavange@unc.edu)

(*Circulation*. 2006;114:2528-2533.)

© 2006 American Heart Association, Inc.

*Circulation* is available at <http://www.circulationaha.org>

DOI: 10.1161/CIRCULATIONAHA.106.613638

The  $n(n+1)/2$  quantities involved here are the  $n$  differences and their  $n(n-1)/2$  pairwise averages. Furthermore, a confidence interval can be constructed about  $\Delta$  via the methods of Hodges and Lehmann on the basis of the exact distribution of  $T$ .<sup>1,5,6</sup>

The asymptotic relative efficiency (ARE) is a useful way to compare a nonparametric test with its parametric counterpart. Briefly, the ARE can be defined as the ratio of sample sizes required by the 2 statistics to achieve the same power under a certain distributional assumption.<sup>7</sup> For a test of paired samples or paired conditions on a sample of subjects, the paired  $t$  test would be the parametric test of choice. The ARE of the Wilcoxon signed rank test relative to the paired  $t$  test is at least 0.864 in the entire class of continuous symmetrical distributions and at least 0.955 when the differences  $\{d_i\}$  follow a normal distribution.<sup>5</sup>

**Example Dataset**

The methods described above are illustrated with an example from cardiovascular research. Example data are from a clinical trial designed to assess the impact of low levels of exposure to carbon monoxide (CO) on exercise tolerance in patients with ischemic heart disease.<sup>8</sup> A total of 42 nonsmoking patients with documented obstructive coronary artery disease and a history of exercise-induced ischemia were enrolled in this 2-period crossover study. The study period consisted of 3 days, a training day and 2 exposure days during which patients were exposed to either air or CO in an environmentally controlled chamber. On all 3 days, patients followed a bicycle exercise protocol in which exercise was conducted at increasing work loads until angina, fatigue, or hypertension occurred. On the first or training day of the study, patients became familiar with the environmentally controlled chamber and conducted a training exercise on the bicycle. Patients demonstrating exercise-induced ischemia during training were then randomly assigned to one of 2 exposure sequences, exposure to air followed by carbon monoxide (Air:CO) and the reverse (CO:Air). Cardiac function and exercise capacity were measured on each day after exposure in the chamber.

A total of 30 patients (8 women and 22 men) successfully completed training and were randomized to exposure sequence. The outcome variable used for illustration here is duration of exercise (seconds) after the exposure condition, provided in Table 1 for each patient. This outcome variable typically has a somewhat skewed distribution and can be subject to outliers, and therefore the use of nonparametric methods for hypothesis testing is particularly appealing. The order variable groups the patients according to order of exposure (1=CO first and 2=Air first). Therefore, 16 patients were exposed to CO on the first day and exposed to Air on the second day, and 14 patients were exposed to Air on the first day, followed by CO on the second day. The baseline measure corresponds to the duration of exercise recorded on the training day before randomization.

The Wilcoxon signed rank test can be used to test for differences between duration of exercise under the 2 exposures. First, differences are formed between the 2 exercise times, and the absolute values of the nonzero differences are

**TABLE 1. Example Data From Ischemic Heart Disease Clinical Trial: Duration of Exercise in Seconds After Exposure to Air and CO**

Patient	Order	Baseline, s	Duration After Exposure to Air, s	Duration After Exposure to CO, s
1	1	240	270	285
2	1	270	220	220
3	1	450	406	370
4	1	540	540	480
5	1	1020	1020	1020
6	1	560	540	540
7	1	240	180	180
8	1	810	780	780
9	1	780	780	780
10	1	750	868	840
11	1	300	505	330
12	1	190	540	360
13	1	360	510	322
14	1	1002	1020	1020
15	1	475	462	540
16	1	900	798	720
17	2	390	370	300
18	2	405	325	300
19	2	285	280	280
20	2	640	795	720
21	2	400	345	320
22	2	1100	1280	1170
23	2	840	720	490
24	2	780	740	720
25	2	510	780	780
26	2	540	510	550
27	2	395	1020	780
28	2	855	614	750
29	2	1200	840	840
30	2	540	720	759

ranked across the 30 patients. Differences of zero are ignored, and midranks are used in the case of ties. Signs are then applied to indicate which differences are  $<0$  or  $>0$ , corresponding to a decrease and an increase in exercise time, respectively. Table 2 provides the rank matrix for the example data. In this example, the sum of the positive ranks is  $T=166.5$ , the sum of the signed ranks is  $U=123$ , the square root of the sum of squares of the signed ranks (the standard deviation of  $U$ ) is  $S=53.56$ , and the test statistic is  $U/S=2.296$ . The exact  $P=0.0198$ , and the approximate  $P=0.0217$ , both indicating that the null hypothesis of a zero median for the difference between exposure conditions in exercise times is rejected in favor of a significant difference. Subjects were able to exercise for significantly longer periods of time after exposure to Air than after exposure to CO. The Hodges-Lehmann point estimate for the median difference is 54.0 seconds with a 95% confidence interval of 15.5 to 110.0.

**TABLE 2. Rank Matrix for the Wilcoxon Signed Rank Test**

Patient	Order	Difference Air vs CO, s	Signed Ranks*
2	1	0	...
5	1	0	...
6	1	0	...
7	1	0	...
8	1	0	...
9	1	0	...
14	1	0	...
19	2	0	...
25	2	0	...
29	2	0	...
1	1	-15	-1.0
24	2	20	2.0
18	2	25	3.5
21	2	25	3.5
10	1	28	5.0
3	1	36	6.0
30	2	-39	-7.0
26	2	-40	-8.0
4	1	60	9.0
17	2	70	10.0
20	2	75	11.0
15	1	-78	-12.5
16	1	78	12.5
22	2	110	14.0
28	2	-136	-15.0
11	1	175	16.0
12	1	180	17.0
13	1	188	18.0
23	2	230	19.0
27	2	240	20.0

\*Signed ranks reflect ranks of absolute values of differences, with zeroes ignored and midranks used in the case of ties.

### Spearman Rank Correlation Coefficient

The Spearman rank correlation coefficient is a measure of association between 2 variables that is particularly useful when 1 or both variables are either ordered categorical or are continuous but from a highly skewed distribution. The coefficient can be used to test the null hypothesis of no association between the 2 variables versus the alternative hypothesis of an association. To compute the Spearman rank correlation coefficient, the values of each of the 2 variables are first ranked, with the use of midranks in the case of ties, and the standard Pearson correlation coefficient is then computed with the use of these ranks. Let  $i=1, \dots, n$  identify the  $n$  subjects in the sample, and let  $X_i$  and  $Y_i$  denote the variable values for the  $i$ th subject. Let  $R_i$  and  $S_i$  denote the ranks of  $X_i$  and  $Y_i$ , respectively. Then the Spearman rank correlation coefficient is given by the following equation:

$$r_s = \frac{\sum(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum(R_i - \bar{R})^2 \sum(S_i - \bar{S})^2}}$$

A test of significance for the association between the 2 variables of interest ( $X$  and  $Y$ ) is given by  $(n-1)r_s^2$ , which is approximately  $\chi^2$  distributed with 1 degree of freedom, when the 2 variables are independent and thereby have no association (ie, the null hypothesis is true).<sup>9</sup>

The Spearman rank correlation coefficient is appropriate for both ordered categorical and continuous variables. The computations are valid with the use of midranks, and therefore ties with respect to either variable can be accommodated. Critical values of the test statistic can be computed with the large-sample  $\chi^2$  approximation when sample sizes are large (eg,  $n \geq 40$ ) and through tabulation of the critical regions of the exact distribution, when sample sizes are small. The statistical procedures SAS Proc FREQ<sup>3</sup> and StatXact<sup>4</sup> both provide exact probability levels for the Spearman rank correlation test.

### Example Dataset, Continued

In the example study, it may be of interest to determine whether the differences in exercise times under the 2 exposure conditions vary with baseline values. If baseline appears to be correlated with differences in exercise times, then an analysis that adjusts for baseline differences among subjects may be warranted. The Spearman rank correlation coefficient can be applied to assess this correlation. To compute the correlation coefficient, the differences in exercise times between the 2 conditions require ranking, irrespective of zero values. Midranks are again used in the case of ties. The ranked values are provided in Table 3.

The Spearman rank correlation coefficient for baseline by differences in exercise times is  $-0.1093$  with  $P=0.5564$ , indicating no significant association between these 2 variables. A logical follow-up question is whether baseline is correlated with exercise times after either exposure condition. The Spearman rank correlation coefficient between the baseline value and exercise time after exposure to Air is  $0.7843$  ( $P<0.0001$ ) and between baseline and exercise time after exposure to CO is  $0.8234$  ( $P<0.0001$ ). Baseline values are therefore strongly associated with postexposure exercise times, regardless of the condition. The difference between conditions with respect to exercise duration does not, however, appear to vary with baseline.

### Wilcoxon Rank Sum Test

When the research question of interest involves comparing 2 samples or groups of subjects with respect to a response variable, such as comparing disease outcomes among patients randomized to receive a test versus control treatment, the Wilcoxon rank sum test has utility. The null hypothesis in this setting is that of no association between the 2 groups of subjects and the response variable, and the alternative hypothesis is that of a location shift for the population represented by one group versus the other, eg, relatively more higher values of the response variable as a result of the test treatment. The Wilcoxon rank sum test can be applied regardless of whether the response variable is metric or ordered categorical, and, as is the case for the signed rank test, the methods of Hodges and Lehmann can be applied to

**TABLE 3. Rank Matrix for the Wilcoxon Rank Sum Test and Spearman Correlation**

Patient	Order	Difference Air vs CO, s	Ranks*
1	1	-15	5.0
2	1	0	10.5
3	1	36	20.0
4	1	60	21.0
5	1	0	10.5
6	1	0	10.5
7	1	0	10.5
8	1	0	10.5
9	1	0	10.5
10	1	28	19.0
11	1	175	26.0
12	1	180	27.0
13	1	188	28.0
14	1	0	10.5
15	1	-78	2.0
16	1	78	24.0
17	2	70	22.0
18	2	25	17.5
19	2	0	10.5
20	2	75	23.0
21	2	25	17.5
22	2	110	25.0
23	2	230	29.0
24	2	20	16.0
25	2	0	10.5
26	2	-40	3.0
27	2	240	30.0
28	2	-136	1.0
29	2	0	10.5
30	2	-39	4.0

\*Ranks of differences incorporate all values, including zeroes, and midranks are used in the case of ties.

compute point estimates and confidence intervals for the difference in medians between the 2 samples.

Let  $n_1$  denote the sample size in the first group and  $n_2$  denote the sample size in the second group. The total sample size is  $n=n_1+n_2$ . The Wilcoxon rank sum test is computed as follows:

1. Ranks are assigned to all observations of the response variable, pooling across groups of subjects and using midranks for ties.
2. The test statistic can most easily be expressed in the manner previously described for the Spearman rank correlation coefficient by letting  $S$  have the value 1 for the  $n_1$  subjects in group 1, and the value 0 for the  $n_2$  subjects in group 2 (where  $n=n_1+n_2$ ), and letting  $R$  correspond to the ranks of the response variable.
3. The significance level is calculated by comparing the test statistic to the  $\chi^2$  distribution with 1 degree of freedom,

when sample sizes are large (eg,  $\geq 20$  per group). When sample sizes are small and subjects are randomly allocated to groups (either by the design of the study or as implied by the null hypothesis), the significance level is calculated by comparing the test statistic to the critical region of the exact distribution.<sup>10</sup>

If there are no ties among the ranks, then the test statistic simplifies to a commonly used form. Let  $T$  be the sum of the ranks in group 1. Then the rank sum test statistic is given by

$$\frac{T - n_1(n+1)/2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1) / 12}}$$

The statistical procedures SAS Proc NPARIWAY<sup>3</sup> and StatXact<sup>4</sup> both provide exact probability levels for the Wilcoxon rank sum test.

If it is assumed that metric distributions for the 2 groups have the same shape and scale, Hodges-Lehmann estimates for the difference in medians between the 2 groups of patients,  $\Delta$ , and confidence limits about  $\Delta$  are available. The point estimate corresponds to the median of all pairwise differences between observations in one group versus those in the other group. There are  $n_1 n_2$  such differences.

The ARE for the Wilcoxon rank sum test relative to the  $t$  test for comparing 2 independent samples is at least 0.864 when the alternative hypothesis is a location shift in the distributions of the 2 samples and all continuous distributions are considered. When the distributions are normal, the ARE is at least 0.955. Note that when the distributions of the response variables are highly skewed, with long tails at either end, then the ARE can exceed 1.0, indicating that the Wilcoxon rank sum test will be more powerful than a  $t$  test in this instance.<sup>11</sup>

### Example Dataset, Continued

The association between order of exposure and difference in exposure times can be assessed by applying the Wilcoxon rank sum test because order of exposure defines 2 groups of patients for comparison. The sum of the ranks for subjects exposed to CO first is 245.5, and the sum of the ranks for patients exposed to Air first is 219.5. Expected values of these 2 quantities under the null hypothesis of no association are 248.0 and 217.0, respectively, where expected value is defined as the average value of the sum of the ranks across all possible randomizations of  $n_1$  subjects to the first group and  $n_2$  subjects to the second group. The Wilcoxon rank sum test statistic=219.5, and the exact  $P=0.9250$ , compatible with no difference between the 2 groups with respect to differences in exercise duration. The approximate  $P=0.9178$ . The Spearman rank correlation coefficient for order of exposure by differences in exercise times is  $-0.0197$ , and  $P=0.9178$ , identical to that associated with the Wilcoxon rank sum test, which is compatible with no correlation between these 2 variables.

For this example dataset, the fact that the order of exposure conditions did not appear to be related to the response variable validates the use of a signed ranks analysis of exercise times under the 2 conditions, ignoring the order of exposure. Had order been related to response, then a proper

crossover analysis of the study data would be required that accounted for the order of exposure in assessing the impact of CO versus air on exercise times.<sup>12</sup>

### Extensions of the Rank Sum Test

The Wilcoxon rank sum test can be extended to allow for covariate adjustment in a nonparametric analog to ANCOVA. Rank ANCOVA<sup>13</sup> can be performed through the following steps:

1. Ranks are computed for the response variable, ignoring groups and using midranks in the case of ties.
2. Ranks are then computed for the covariate, ignoring groups and using midranks in the case of ties.
3. A linear regression model is fit, regressing the ranked response variable onto the ranked covariates (with groups ignored), and the residuals are output.
4. A test of the association between group and the response variable, adjusting for the covariate, is provided by applying computations like those described for the Spearman rank correlation coefficient (the formula for  $r_s$ ), but with  $R$  equal to the residuals,  $S=1$  for subjects in group 1, and  $S=0$  for subjects in group 2.<sup>9</sup>

The method in step 4 for comparing the residuals in the 2 groups is an extension of the Wilcoxon rank sum test to provide covariate adjustment. Rank ANCOVA can provide additional power through the variance reduction typically associated with a baseline covariate adjustment, even when the response variable does not follow a normal distribution.<sup>14</sup>

Stratification may also be an important aspect of the study design resulting from patients being sorted into subsets before the conduct of the study, (eg, male and female strata or strata consisting of patients from different clinical centers in a multicenter study). Extensions for the Wilcoxon rank sum test and the test for the Spearman rank correlation coefficient that account for stratification are available.<sup>9,10</sup> The stratified extension for the Wilcoxon rank sum test is referred to as the van Elteren statistic.<sup>15</sup> Computations for this method are shown in the technical appendix in the online-only Data Supplement and are available through SAS Proc FREQ.<sup>3</sup>

### Example Dataset, Continued

Rank ANCOVA can be applied to the example dataset to assess the impact of order of exposure on the differences in exercise times between exposure conditions while controlling for baseline values. Following the steps outlined above, the difference scores between air and CO are ranked as for the Wilcoxon rank sum test (Table 3). Baseline values are also ranked, and a linear regression model is fit with the use of SAS Proc GLM.<sup>3</sup> The residuals from the model are output, and the Pearson correlation coefficient between the residuals and the order variable is computed as  $-0.0372$ . The probability value from the exact test is 0.8437 and from the asymptotic test is 0.8412, both showing little association between order of exposure and difference scores, after adjustment for baseline values. These results are similar to the unadjusted test of association computed above. Because the Spearman rank correlation coefficient showed no evidence of association between baseline exercise times and the differ-

ence in exercise times, the fact that the covariate adjustment had little impact on the results of the unadjusted Wilcoxon rank sum test is not surprising.

To illustrate the extension of these methods for stratification, subjects were stratified according to whether their baseline value was below the median of 540 seconds versus equal to or above the median. A stratified Wilcoxon rank sum test (ie, van Elteren test) was then performed. The probability value for testing the null hypothesis of no association in all strata is 0.8810 and therefore is compatible with no association between order of exposure and differences in exercise duration after stratifying on baseline value (below versus above the median).

### Discussion

Nonparametric methods such as those described here are most useful in situations in which their use is prespecified before data analysis. When response variables are ordered categories or when it is known in advance that assumptions about metric response variables will likely fail, then such prespecification is possible at the time the statistical analysis plan is prepared. In these situations, nonparametric methods will have good power properties, with power as high as 93% of that expected for standard parametric methods (eg,  $t$  tests) applied under ideal circumstances.

The methods described here all address a shift in location as the alternative hypothesis, in which location corresponds to the median of the response variable distribution. If transformed data are expected to be normally distributed (eg, the response variable follows a log-normal distribution), then nonparametric methods will be at least 95% efficient, and their use precludes identifying the most optimal transformation.

When the response variable is so highly skewed that the distribution appears to have an "L" or "J" shape, the Wilcoxon tests will not have good power, and Savage (or log-rank) tests will be better.<sup>16</sup> With highly skewed distributions, both groups of subjects will tend to have ranked values on one side of the median, but only one group will have ranked values on the other or tail side of the distribution. In this case, only the rank values on the tail side are informative, and against this alternative, the Wilcoxon rank sum test will not be the most appropriate test. Because the Wilcoxon tests address shifts in location only, ranked values from both groups are expected to occur to the left and to the right of the median, and both are informative. Under the alternative hypothesis of a shift in location, one group will tend to have ranks on one side of the overall median, whereas the other will tend to have ranks on the opposite side. This is precisely the setting in which the Wilcoxon tests are most useful and nearly as powerful as parametric methods applied when all assumptions hold.

### Disclosures

None.

### References

1. Hollander M, Wolfe DA. *Nonparametric Statistical Methods*. New York, NY: John Wiley & Sons; 1973.
2. Conover WJ. *Practical Nonparametric Statistics*. New York, NY: John Wiley & Sons; 1971.

3. SAS Institute Inc. *SAS/STAT User's Guide, Version 9*. Cary, NC: SAS Institute Inc; 2004.
4. Cytel Software Corporation. *StatXact 7 On-line User Manual*. Cambridge, Mass: Cytel Software Corporation; 2005.
5. Woolson RF. Wilcoxon signed-rank test. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*. Vol 6. West Wessex, England: John Wiley & Sons Ltd; 1998:4739–4740.
6. Hodges JL, Lehmann EL. Rank methods for combination of independent experiments in analysis of variance. *Ann Math Stat*. 1962;33:482–497.
7. DasGupta A. *Encyclopedia of Biostatistics*. Vol 1. Armitage P, Colton T, eds. West Wessex, England: John Wiley & Sons Ltd; 1998:210–215.
8. Adams KF, Koch GG, Chatterjee B, Goldstein GM, O'Neil JJ, Bromberg PA, Sheps DS, McAllister S, Price CJ, Bissette J. Acute elevation of blood carboxyhemoglobin to 6% impairs exercise performance and aggravates symptoms in patients with ischemic heart disease. *J Am Coll Cardiol*. 1988;12:900–909.
9. Stokes ME, Davis CS, Koch GG. *Categorical Data Analysis Using the SAS System*. 2nd ed. Cary, NC: SAS Institute Inc; 2000.
10. Landis RJ, Sharp TJ, Kuritz SJ, Koch GG. Mantel-Haenszel methods. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*. Vol 3. West Essex, England: John Wiley & Sons Ltd; 1998:2378–2391.
11. Moses L. Wilcoxon-Mann-Whitney test. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*. Vol 6. West Wessex, England: John Wiley & Sons Ltd; 1998:4742–4745.
12. Tudor G, Koch GG. Review of non-parametric methods for the analysis of crossover studies. *Stat Methods Med Res*. 1994;3:345–381.
13. Koch GG, Carr GJ, Amara IA, Stokes ME, Uryniak TJ. Categorical data analysis. In: Berry DA, ed. *Statistical Methodology in the Pharmaceutical Sciences*. New York, NY: Marcel Dekker; 1990:389–473.
14. LaVange LM, Durham TA, Koch GG. Randomization-based nonparametric methods for the analysis of multicentre trials. *Stat Methods Med Res*. 2005;14:281–301.
15. Lehmann EL. *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco, Calif: Holden-Day; 1975.
16. Koch GG, Sen PK, Amara I. Log-rank scores, statistics, and tests. In: Kotz S, Johnson NL, eds. *Encyclopedia of Statistical Sciences*. Vol 5. New York, NY: John Wiley & Sons; 1985:136–142.

---

KEY WORDS: statistics, nonparametric ■ probability ■ variable distributions