**Nonparametric Statistics[[1]](#footnote-1)©**

 I shall compare the Wilcoxon rank-sum statistic with the independent samples *t*-test to illustrate the differences between typical nonparametric tests and their parametric “equivalents.”

|  |  |  |
| --- | --- | --- |
| Independent Samples *t* test |  | **Wilcoxon Rank-Sum Test** |
| H∅: μ1 = μ2  |  | H∅: Population 1 = Population 2 |
| Assumptions: |  | None for general test, but often assume: |
|  Normal populations |  | Equal shapes |
|  Homogeneity of variance |  | Equal dispersions |
|  (but not for separate variances test) |  |  |

 Both tests are appropriate for determining whether or not there is a significant association between a dichotomous variable and a continuous variable with independent samples data. Note that with the independent samples *t* test the null hypothesis focuses on the population means. If you have used the general form of the nonparametric hypothesis (without assuming that the populations have equal shapes and equal dispersions), rejection of that null hypothesis simply means that you are confident that the two populations differ on one or more of location, shape, or dispersion. If, however, we are willing to assume that the two populations have identical shapes and dispersions, then we can interpret rejection of the nonparametric null hypothesis as indicating that the populations differ in location. With these equal shapes and dispersions assumptions the nonparametric test is quite similar to the parametric test. In many ways the nonparametric tests we shall study are little more than parametric tests on rank-transformed data. The nonparametric tests we shall study are especially sensitive to differences in medians.

 If your data indicate that the populations are not normally distributed, then a nonparametric test may be a good alternative, especially if the populations do appear to be of the same non-normal shape. If, however, the populations are approximately normal but heterogeneous in variance, I would recommend a separate variances *t*-test over a nonparametric test. If you cannot assume equal dispersions with the nonparametric test, then you cannot interpret rejection of the nonparametric null hypothesis as due solely to differences in location.

# Conducting the Wilcoxon Rank-Sum Test

 Rank the data from lowest to highest. If you have tied scores, assign all of them the mean of the ranks for which they are tied. Find the sum of the ranks for each group. If *n1* = *n2*, then the test statistic, *WS,* is the smaller of the two sums of ranks. Go to the table (starts on page 709 of Howell) and obtain the one-tailed (lower tailed) *p*. For a two-tailed test (nondirectional hypotheses), double the *p*. If *n1* ≠ *n2*, obtain both *WS* and *WS′* : *WS* is the sum of the ranks for the group with the smaller *n*,  (see the rightmost column in the table), the sum of the ranks that would have been obtained for the smaller group if we had ranked from high to low rather than low to high. The test statistic is the smaller of *WS* and *WS′*. If you have directional hypothesis, to reject the null hypothesis not only must the one-tailed *p* be less than or equal to the criterion, but also the mean rank for the sample predicted (in H1) to come from the population with the smaller median must be less than the mean rank in the other sample (otherwise the exact *p* = one minus the *p* that would have been obtained were the direction correctly predicted).

 If you have large sample sizes, you can use the normal approximation procedures illustrated on page 672 of Howell. Computer programs generally do use such an approximation, but they may also make a correction for continuity (reducing the absolute value of the numerator by .5) and they may obtain the probability from a *t*-distribution rather than from a *z*-distribution. Please note that the rank-sum statistic is essentially identical to the (better know to psychologists) Mann-Whitney *U* statistic. but the Wilcoxon is easier to compute. If someone insists on having *U*, you can always transform your *W* to *U* (see page 673 in Howell).

 Here is a summary statement for the problem on page 672 of Howell (I obtained an exact *p* from SAS rather than using a normal approximation): A Wilcoxon rank-sum test indicated that babies whose mothers started prenatal care in the first trimester weighed significantly more (*N* = 8, *M* = 3259 g, *Mdn* = 3015 g, *s* = 692 g) than did those whose mothers started prenatal care in the third trimester (*N* = 10, *M* = 2576 g, *Mdn* = 2769 g, *s* = 757 g), *W* = 52, *p* = .034.

**Power of the Wilcoxon Rank Sums Test**

 You already know that the majority of statisticians reject the notion that parametric tests require interval data and thus ordinal data need be analyzed with nonparametric methods (Gaito, 1980). There are more recent simulation studies that also lead one to the conclusion that scale of measurement (interval versus ordinal) should not be considered when choosing between parametric and nonparametric procedures (see the references on page 57 of Nanna & Sawilowsky, 1998). There are, however, other factors that could lead one to prefer nonparametric analysis with certain types of ordinal data. Nanna and Sawilowsky (1998) addressed the issue of Likert scale data. Such data typically violate the normality assumption and often the homogeneity of variance assumption made when conducting traditional parametric analysis. Although many have demonstrated that the parametric methods are so robust to these violations that this is not usually a serious problem with respect to holding alpha at its stated level (but can be, as you know from reading [Bradley's articles](http://core.ecu.edu/psyc/wuenschk/read30.htm) in the *Bulletin of the Psychonomic Society*), one should also consider the power characteristics of parametric versus nonparametric procedures.

 While it is generally agreed that parametric procedures are a little more powerful than nonparametric procedures when the assumptions of the parametric procedures are met, what about the case of data for which those assumptions are not met, for example, the typical Likert scale data? Nanna and Sawilowsky demonstrated that with typical Likert scale data, the Wilcoxon rank sum test has a considerable power advantage over the parametric *t* test. The Wilcoxon procedure had a power advantage with both small and large samples, with the advantage actually increasing with sample size.

**Wilcoxon’s Signed-Ranks Test**

 This test is appropriate for matched pairs data, that is, for testing the significance of the relationship between a dichotomous variable and a continuous variable with related samples. It does assume that the difference scores are rankable, which is certain if the original data are interval scale. The parametric equivalent is the correlated *t*-test, and another nonparametric is the binomial sign test. To conduct this test you compute a difference score for each pair, rank the absolute values of the difference scores, and then obtain two sums of ranks: The sum of the ranks of the difference scores which were positive and the sum of the ranks of the difference scores which were negative. The test statistic, *T*, is the smaller of these two sums for a nondirectional test (for a directional test it is the sum which you predicted would be smaller). [Difference scores of zero](http://core.ecu.edu/psyc/wuenschk/docs30/WSRT_Ties.docx) are usually discarded from the analysis (prior to ranking), but it should be recognized that this biases the test against the null hypothesis. A more conservative procedure would be to rank the zero difference scores and count them as being included in the sum which would otherwise be the smaller sum of ranks. Refer to the table that starts on page 703 of Howell to get the exact one-tailed (lower-tailed) *p*, doubling it for a nondirectional test. Normal approximation procedures are illustrated on page 677 of Howell. Again, computer software may use a correction for continuity and may use *t* rather than *z*.

 Only a few people know that the signed-ranks test assumes that the difference scores are [symmetrically distributed](https://www.quality-control-plan.com/StatGuide/srank_paired_ass_viol.htm#Skewness) – not necessarily normally distributed, but symmetrically. Accordingly, if the difference scores are distinctly skewed, the test is not valid. One alternative is the binomial sign test, but I would prefer a resampling test.

 Here is an example summary statement using the data on page 677 of Howell: A Wilcoxon signed-ranks test indicated that participants who consumed glucose had significantly better recall (*M* = 7.62, *Mdn* = 8.5, *s* = 3.69) than did subjects who consumed saccharine (*M* = 5.81, *Mdn* = 6, *s* = 2.86), *T*(*N* = 16) = 14.5, *p* = .004.

**Kruskal-Wallis ANOVA**

 This test is appropriate to test the significance of the association between a categorical variable (*k* ≥ 2 groups) and a continuous variable when the data are from independent samples. Although it could be used with 2 groups, the Wilcoxon rank-sum test would usually be used with two groups. To conduct this test you rank the data from low to high and for each group obtain the sum of ranks. These sums of ranks are substituted into the formula on page 678 of Howell. The test statistic is *H*, and the *p* is obtained as an upper-tailed area under a chi-square distribution on *k*-1 degrees of freedom. Do note that this one-tailed *p* is appropriately used for a nondirectional test. If you had a directional test (for example, predicting that Population 1 < Population 2 < Population 3), and the medians were ordered as predicted, you would divide that one-tailed *p* by *k* ! before comparing it to the criterion.

 The null hypothesis here is: Population 1 = Population 2 = ......... = Population *k*. If you reject that null hypothesis you probably will still want to make “pairwise comparisons,” such as group 1 versus group 2, group 1 versus group 3, group 2 versus group 3, etc. This topic is addressed in detail in Chapter 12 of Howell. One may need to be concerned about inflating the “**familywise alpha**,” the probability of making one or more Type I errors in a family of *c* comparisons. If *k* = 3, one can control this familywise error rate by using **Fisher’s procedure** (also known as “a protected test”): Conduct the omnibus test (the Kruskal-Wallis) with the promise not to make any pairwise comparisons unless that omnibus test is significant. If the omnibus test is not significant, you stop. If the omnibus test is significant, then you are free to make the three pairwise comparisons with Wilcoxon’s rank-sum test. If *k* > 3 Fisher’s procedure does not adequately control the familywise alpha. One fairly conservative procedure is the **Bonferroni procedure**. With this procedure one uses an adjusted criterion of significance, . This procedure does not require that you first conduct the omnibus test, and should you first conduct the omnibus test, you may make the Bonferroni comparisons whether or not that omnibus test is significant. Suppose that *k* = 4 and you wish to make all 6 pairwise comparisons (1-2, 1-3, 1-4, 2-3, 2-4, 3-4) with a maximum familywise alpha of .05. Your adjusted criterion is .05 divided by 6, .0083. For each pairwise comparison you obtain an exact *p*, and if that exact *p* is less than or equal to the adjusted criterion, you declare that difference to be significant. Do note that the cost of such a procedure is a great reduction in power (you are trading an increased risk of Type II error for a reduced risk of Type I error).

 Here is a summary statement for the problem on page 679 of Howell: Kruskal-Wallis ANOVA indicated that type of drug significantly affected the number of problems solved, *H*(2, *N* = 19) = 10.36, *p* = .006. Pairwise comparisons made with Wilcoxon’s rank-sum test revealed that ......... Basic descriptive statistics (means, medians, standard deviations, sample sizes) would be presented in a table.

**Friedman’s ANOVA**

 This test is appropriate to test the significance of the association between a categorical variable (*k* ≥ 2) and a continuous variable with randomized blocks data (related samples). While Friedman’s test could be employed with *k* = 2, usually Wilcoxon’s signed-ranks test would be employed if there were only two groups. Subjects have been matched (blocked) on some variable or variables thought to be correlated with the continuous variable of primary interest. Within each block the continuous variable scores are ranked. Within each condition (level of the categorical variable) you sum the ranks and substitute in the formula on page 680 of Howell. As with the Kruskal-Wallis, obtain *p* from chi-square on *k*−1 degrees of freedom, using an upper-tailed *p* for nondirectional hypotheses, adjusting it with *k*! for directional hypotheses. Pairwise comparisons could be accomplished employing Wilcoxon signed-ranks tests, with Fisher’s or Bonferroni’s procedure to guard against inflated familywise alpha.

 Friedman’s ANOVA is closely related to Kendall’s coefficient of concordance. For the example on page 680 of Howell, the Friedman tests asks whether the rankings are the same for the three levels of visual aids. Kendall’s coefficient of concordance, *W*, would measure the extent to which the blocks agree in their rankings. .

 Here is a sample summary statement for the problem on page 680 of Howell: Friedman’s ANOVA indicated that judgments of the quality of the lectures were s**ignifica**ntly affected by the number of visual aids employed, (2, *n* = 17) = 10.94, *p* = .004. Pairwise comparisons with Wilcoxon signed-ranks tests indicated that ....................... Basic descriptive statistics would be presented in a table.

**Power**

 It is commonly opined that the primary disadvantage of the nonparametric procedures is that they have less power than does the corresponding parametric test. The reduction in power is not, however, great, and if the assumptions of the parametric test are violated, then the nonparametric test may be more powerful.

**Everything You Ever Wanted to Know About Six But Were Afraid to Ask**

 You may have noticed that the numbers 2, 3, 4, 6, 12, and 24 commonly appear as constants in the formulas for nonparametric test statistics. This results from the fact that the sum of the integers from 1 to *n* is equal to *n*(*n* + 1) / 2.

**Effect Size Estimation**

 Please read my document [Nonparametric Effect Size Estimators](http://core.ecu.edu/psyc/wuenschk/docs30/Nonparametric-EffectSize.pdf) .

# Using SAS to Compute Nonparametric Statistics

 Run the program Nonpar.sas from my [SAS programs page](http://core.ecu.edu/psyc/wuenschk/SAS/SAS-Programs.htm). Print the output and the program file.

The first analysis is a **Wilcoxon Rank Sum Test**, using the birthweight data also used by Howell (page 672) to illustrate this procedure. SAS gives us the sum of scores for each group. That sum for the smaller group is the statistic which Howell calls *WS* (100). Note that SAS does not report the W′S statistic (52), but it is easily computed by hand -- . Please remember that the test statistic which psychologists report is the smaller of W and W′ SAS does report both a normal approximation (*z* = 2.088, *p* = .037) and an exact (not approximated) *p* = .034. The *z* differs slightly from that reported by Howell because SAS employs a correction for continuity (reducing by .5 the absolute value of the denominator of the *z* ratio).

The next analysis is a **Wilcoxon Matched Pairs Signed-Ranks Test** using the data from page 677 of Howell. Glucose-Saccharine difference scores are computed and then fed to Proc Univariate. Among the many other statistics reported with Proc Univariate, there is the Wilcoxon Signed-Ranks Test. For the data employed here, you will see that SAS reports “*S* = 53.5, *p* = .004.” S, the signed-rank statistic, is the absolute value of , where T is the sum of the positive ranks or the negative ranks.

*S* is the difference between the expected and the obtained sums of ranks. You know that the sum of the ranks from 1 to *n* is . Under the null hypothesis, you expect the sum of the positive ranks to equal the sum of the negative ranks, so you expect each of those sums of ranks to be half of . For the data we analyzed here, the sum of the ranks 1 through 16 = 136, and half of that is 68. The observed sum of positive ranks is 121.5, and the observed sum of negative ranks is 14.5 The difference between 68 and 14.5 (or between 121.5 and 68) is 53.5, the value of S reported by SAS.

 To get *T* from *S*, just subtract the absolute value of *S* from the expected value for the sum of ranks, that is, . Alternatively, just report *S* instead of *T* and be prepared to explain what *S* is to the ignorant psychologists who review your manuscript.

If you needed to conduct several signed-ranks tests, you might not want to produce all of the output that you get by default with Proc Univariate. See my program WilcoxonSignedRanks.sas on my [SAS programs page](http://core.ecu.edu/psyc/wuenschk/SAS/SAS-Programs.htm) to see how to get just the statistics you want and nothing else.

Note that a **Binomial Sign Test** is also included in the output of Proc Univariate. SAS reports “*M* = 5, *p* = .0213.” *M* is the difference between the expected number of negative signs and the obtained number of negative signs. Since we have 16 pairs of scores, we expect, under the null, to get 8 negative signs. We got 3 negative signs, so M - 8 - 3 = 5. The *p* here is the probability of getting an event as or more unusual than 3 successes on 16 binomial trials when the probability of a success on each trial is .5. Another way to get this probability with SAS is: **Data** p; p = **2**\*PROBBNML(**.5**, **16**, **3**); **proc** **print**; **run**;

 Next is a **Kruskal-Wallis ANOVA**, using Howell’s data on effect of stimulants and depressants on problem solving (page 679). Do note that the sums and means reported by SAS are for the ranked data. Following the overall test, I conducted pairwise comparisons with Wilcoxon Rank Sum tests. Note how I used the subsetting IF statement to create the three subsets necessary to do the pairwise comparisons.

 The last analysis is **Friedman’s Rank Test for Correlated Samples**, using Howell’s data on the effect of visual aids on rated quality of lectures (page 680). Note that I first had to use Proc Rank to create a data set with ranked data. Proc Freq then provides the Friedman statistic as a Cochran-Mantel-Haenszel Statistic. One might want to follow the overall analysis with pairwise comparisons, but I have not done so here.

 I have also provided an alternative rank analysis for the data just analyzed with the Friedman procedure. Note that I simply conducted a factorial ANOVA on the rank data, treating the blocking variable as a second independent variable. One advantage of this approach is that it makes it easy to get the pairwise comparisons -- just include the LSMEANS command with the PDIFF option. The output from LSMEANS includes the mean ranks and a matrix of *p* values for tests comparing each group’s mean rank with each other group’s mean rank. [Read this.](https://www.r-bloggers.com/2012/02/beware-the-friedman-test/)

References

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin*, *87*, 564-567. doi:10.1037/0033-2909.87.3.564

Howell, D. C. (2013). [***Statistical methods for psychology***](http://www.cengagebrain.com/shop/isbn/9781111835484) (8th ed.). Belmont, CA: Cengage Wadsworth.

Nanna, M. J., & Sawilowsky, S. S. (1998). Analysis of Likert scale data in disability and medical rehabilitation research. *Psychological Methods*, *3*, 55–67. doi:10.1037/1082-989X.3.1.55

* [Annotated SAS Output](http://core.ecu.edu/psyc/wuenschk/docs30/Nonparm_Output.pdf)
* [Do it with SPSS](http://core.ecu.edu/psyc/wuenschk/SPSS/Nonparametrics.htm)
* Effect Size [Estimators](http://core.ecu.edu/psyc/wuenschk/docs30/Nonparametric-EffectSize.pdf)
* [How large must the sample sizes be to use the normal approximation?](http://www.amstat.org/publications/jse/v18n2/bellera.pdf)
* [Those Nonparametric Tests are Not Tests of Differences in Medians](http://core.ecu.edu/psyc/wuenschk/docs30/MannWhitney-Medians.docx)
* [Return to Wuensch’s Statistics Lessons Page](http://core.ecu.edu/psyc/wuenschk/StatsLessons.htm)

Copyright 2021, [Karl L. Wuensch](http://core.ecu.edu/psyc/wuenschk/klw.htm) - All rights reserved.

1. © Copyright 2021, Karl L. Wuensch - All rights reserved. [↑](#footnote-ref-1)