CHAPTER

8

Nonparametric Statistics

Stephen W. Scheff

University of Kentucky Sanders–Brown Center on Aging, Lexington, KY, USA

Statistics is exciting because you get to play with others' data while telling them their research is crap. ***Stephen Senn***

OUTLINE

http://dx.doi.org/10.1016/B978-0-12-804753-8.00008-7

Another name for nonparametric statistics is "**distribution-free statistical procedures**." What this means is that there are no assumptions made about the distribution of the data, unlike the parametric tests that "assume" a normal distribution. In short, there is no assumption about the variability of the data (forget about homogeneity of variance) and the form of the distribution. These statistics are designed to work with data that have a lot of variance either in one group or in several groups, or with data that are primarily ordinal and nominal.

Major advantages of nonparametric statistics:

1. Can be used with nominal, ordinal, interval, or ratio data
2. Are not restrictive about assumptions concerning distribution and variance
3. These tests are not affected by extreme outliers in the data
4. Can sometimes detect differences between groups that parametric statistics do not
5. Can be used with very small sample sizes
6. Can be used even when the data are skewed
7. Often very easy to calculate
8. Quite often they are very easy to understand

Major disadvantages of nonparametric statistics:

1. Less powerful than parametric statistics
   a. Parametric statistics are more powerful only when the assumptions underlying their use are valid (e.g., homogeneity of variance)
2. Can only be used with relatively "simple" experimental designs
3. Do not take advantage of all the information about a sample distribution
4. Analyze ranks rather than actual experimental values
5. Conclusions are more general because hypotheses tested are less specific

Why are nonparametric statistics less powerful than parametric procedures? The very simple reason is that the parametric statistics use all

TABLE 8.1 Rank Ordering of Numbers

Rank	1	2	3	4	5	6	7	8
	717	672	534	314	298	197	123	111

of the information and nonparametric statistics do not. For example, if you were to count the number of astrocytes in a given region of the thalamus and got the following data from eight rats:

$$197, 717, 298, 123, 672, 534, 111, 314$$

you would subsequently rank the data (for nonparametric analysis) from highest value to lowest (Table 8.1).

The distance between 534 and 314 (ranks 3 and 4) would be the same as between 123 and 111 (ranks 7 and 8). The information concerning the magnitude of the scores is lost when converted into ranks.

For some nonparametric statistics, it is assumed that no two values are the same. While this may sound rather strange, for these statistics, ties that do occur are simply eliminated from consideration. Some statistics software programs actually attempt to make a correction for ties. Still in other statistics, ties are given an average score. It is also important to remember that when reporting results involving nonparametric statistics, it is **inappropriate** to report the mean and standard deviation. Unfortunately, this mistake is made very often in journal articles. The appropriate descriptive statistic to report is the median and the range for each group either in the text or in a table.

There are so many different distribution-free procedures and individual statistical tests that it is beyond the scope of this book to present them all. The reader is advised to consult several excellent texts on the topic (Siegel and Castellan[1]; Hollanders and Wolfe[2]; Corder and Foreman[3]). Most of the commercially available statistical packages have a limited number of nonparametric statistical tests available.

SIGN TEST

This nonparametric statistic is used for ordinal data and essentially measures the relative ordering of different categories of a variable. At the end of a study evaluating the effectiveness of an antioxidant on recovery from traumatic brain injury you note a particular cluster of glial cells in the anterior thalamus contralateral to the injury that you have never observed before. It does not appear in all of the animals but you want to know if there is any evidence (at the $\alpha = 0.05$) that the antioxidant

treatment is related to this "anatomical anomaly." Although you are not sure what this anatomical feature means, it might first be wise to test whether or not it is worth pursuing. The H_o is that antioxidant treatment has no relationship to the anatomical clustering of cells in the anterior thalamus. To test this you obtain histological sections from 12 animals that have had both brain injury and the antioxidant therapy. You then simply find out how many of these animals show the anatomical clustering and assign those animals a plus sign (+) and those that do not a minus sign (−). In the example in Table 8.2, there is a clustering of cells in 9 of the 12 animals. If chance alone is working, the probability of a plus is equal to the probability of a minus. Since there are only two alternatives (yes or no), $P = 0.50$. By simply consulting a binomial distribution table, we find that for $n = 12$, with nine pluses and $P = 0.50$, the resulting p value is 0.0737, which does not quite reach significance and thus the H_o is not rejected.

The sign test is sometimes used in a repeated measure type design. For example you might want to know if lowering the water temperature in your Morris water maze makes the rats swim faster. You record the swim speed of the subjects with the temperature at two different temperatures (27 and 24 °C). If the speed is faster at the lower temperature it is a plus and if it is slower it is a minus. If there is no difference between the two

TABLE 8.2 Sign Test

Animal	Yes	No
1	+	
2	+	
3		−
4	+	
5	+	
6	+	
7		−
8		−
9	+	
10	+	
11	+	
12	+	
Total	9	3

scores the values are ignored and the subject is not included in the analysis. Consider the data in Table 8.3.

Because animals No. 2 and No. 3 had the same value, they are not included in the analysis. In this situation the n = the total number of +'s and −'s, which is 10. In the present example the H_o is that a lower water temperature does not increase swim speed. This is a one-tailed test because it is asserting direction. By consulting the binomial distribution table, for n = 10, with seven pluses and three minuses and $P = 0.50$, the resulting p value is 0.1719. In this case the H_o is supported and we conclude that the lower water temperature did not significantly increase the overall swim speed. What is of importance in the sign test is the lower number, whether it is minuses or pluses. In Table 8.3 the minus signs have a lower number than the plus signs. Very often the results of a sign test are given by a simple p value such as $p > 0.1$. For somewhat large samples, some investigators will calculate a Z value. In those situations the reporting of the sign test in a journal would be the Z value (e.g., $Z = 3.68$) along with the p value (e.g., $p < 0.001$).

The sign test can only indicate the probability of obtaining a specific score that is in the opposite direction of the other cumulated value. It cannot indicate cause and effect. If the results are statistically significant then it indicates that the two groups were derived from different populations. What is absolutely critical is that the scores are paired if it is a repeated measure design.

TABLE 8.3 Sign Test: Water Temperature

Animal	27 °C	24 °C	+/−
1	20	23	+
2	**34**	**34**	**na**
3	**32**	**32**	**na**
4	27	39	+
5	31	33	+
6	22	20	−
7	26	29	+
8	24	28	+
9	30	32	+
10	28	31	+
11	27	26	−
12	29	28	−

WILCOXON MATCHED PAIRS SIGNED RANK TEST (WILCOXON SIGNED RANK TEST)

This particular test is also called the **Wilcoxon matched pairs test** or the **Wilcoxon signed rank test**. It is very appropriate for a repeated measure design where the same subjects are evaluated under two different conditions such as with the water maze temperature experiment in Table 8.3. It is the nonparametric equivalent of the parametric paired *t*-test. This is not the same as the **Wilcoxon rank sum test**, which compares two nonpaired groups and is equivalent to the parametric unpaired *t*-test. The Wilcoxon signed rank is more powerful than the sign test. This statistic differs from the sign test in that it considers the magnitude of the difference while the sign test does not. It uses more information from the sets of scores than the simple sign test. Because it uses more information it is considered to be more precise than the sign test. Look at the swim speed data (cm/s) in Table 8.4 and at the result of the three different statistics in Table 8.5.

If a pair of scores are equal (the same value) then they are considered tied and dropped from the analysis and the sample size is reduced. In the data below, there are two tied scores (pair No. 2 and No. 3) and three pairs of scores where the swim speed was slower in the colder water (Nos. 6, 11, 12), thus the n = 10 for this nonparametric test. What is absolutely critical in using this test is that the pairs of scores under consideration are related

TABLE 8.4 Wilcoxon Signed Rank Test

Animal	27 °C	24 °C	Difference
1	20	23	+3
2	**34**	**34**	**na**
3	**32**	**32**	**na**
4	27	39	+12
5	31	33	+2
6	22	20	−2
7	26	29	+3
8	24	28	+4
9	30	32	+2
10	28	31	+3
11	27	26	−1
12	29	28	−1

TABLE 8.5 Results: Wilcoxon Signed Rank Test

Student *t*-test	Sign test	Wilcoxon signed rank test
$p = 0.0756$	$p = 0.1719$	$p = 0.0367$

and that they are at least ordinal scale. It is unclear why this test is not used more especially in behavioral neuroscience where much of the data do not follow a normal distribution. Many statistical software programs include this statistical test.

Because of the variance in the scores, the Student *t*-test says there is no significant difference. The Wilcoxon signed rank test, which is more sensitive than the sign test, shows a very different outcome and supports the alternative hypothesis (H_A).

When To Use the Wilcoxon Test versus the Sign Test

Whenever you have data that are composed of definite scores, the Wilcoxon signed rank test is preferred. When the data are not a definite score, or if the data are observational, such as "more aggressive" versus "less aggressive" then the sign test is the appropriate statistic. Whenever there is a difference in a particular direction but the absolute quantity of that difference is not precise, and the scores are paired, then the sign test is the statistic to use.

Reporting the results in a journal article requires reporting the observed Z value, the number of observations, and the significance. Often some investigators will report the number of instances with no difference. In the above experiment, one would write: "A Wilcoxon signed rank test revealed a significant difference in the swim speeds between the two water temperatures, $n = 10$, $Z = 2.09$, $p < 0.05$. There were two pairs that showed no difference." Sometimes a T value is reported instead of the Z value. Typically the data are not graphed since it is a repeated measure analysis.

MEDIAN TEST

This is a very simple statistic that can be used to quickly determine if there is a difference between two independent samples even with unequal sample size. Like the Sign test above, it simply counts how many observations occur in a given category regardless of the magnitude of the difference.

TABLE 8.6 Median Test: Enriched Environment

No toys	Toys
19	7
16	8
16	8
20	6
8	7
9	11
7	12
11	14
10	6
10	6

Let us say that you found out that the workers in the animal care facility have been placing novel objects ("toys") in the cages of some of the rats immediately after weaning and you have been using these animals in a learning experiment. This may constitute an enriched environment and could have an effect on some of your experimental studies especially concerning recovery from brain injury. You review your laboratory notes and randomly select a group of animals that had the novel objects and 10 that did not (Table 8.6). The H_o is that the placement of novel objects during weaning had no effect on learning. You analyze errors to criterion in the learning phase of the task. The important factor here is that the animals are randomly selected from the population. In this case, they would have to be randomly selected from the entire group of animals that had the novel objects and then from animals that did not.

What the median test does is set up a contingency table (Table 8.7) and then applies a chi-square test or X^2. In this case the chi-square test shows a value of 3.20 with a p value of 0.0736, supporting the H_o that the addition

TABLE 8.7 Results: Median Test—Enriched Environment

Groups	<Median	>Median	Totals
No toys	3	7	10
Toys	7	3	10
Totals	10	10	20

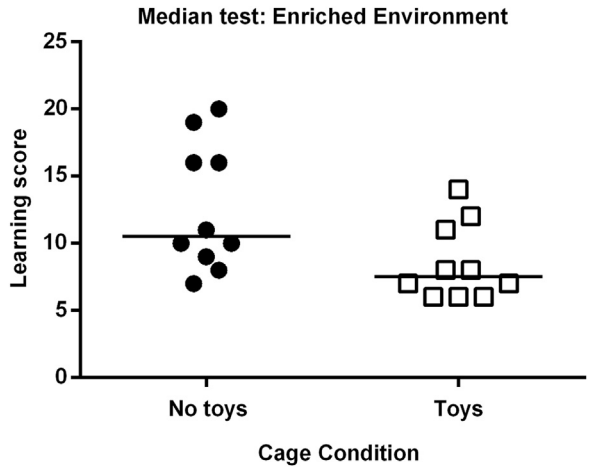

FIGURE 8.1 It is quite apparent that the two groups overlap and the medians are not that different. Each circle or square is the learning score from an individual subject.

of toys did not influence the learning. However, this particular test is not very sensitive to the Type I error, but it is a very quick and easy test to do.

These results are sometimes graphed as shown in Figure 8.1.

WILCOXON RANK SUM TEST (MANN–WHITNEY U TEST)

The Wilcoxon rank sum statistic can also be used when comparing two independent groups with the added advantage that it is applicable whether the groups have equal sample size or not. This is the equivalent of the parametric unpaired *t*-test. It is sometimes also referred to as the **Wilcoxon two-sample test**, the **Wilcoxon test**, the **Wilcoxon–Mann–Whitney test**, or more commonly the **Mann–Whitney U test**. Technically the Mann–Whitney U test uses a slightly different formula but results in the same outcome. This statistic is more sensitive than the median test. As the name implies, one sums the ranking of data from a particular group and compares it to the sum of the ranks of the other group. The idea is to determine if the combined scores are randomly mixed or do the scores from one group cluster toward one end when ranked from lowest to highest. Overall this is a very simple test to run. The only stipulations are that the two samples being investigated must be independent. The variable under consideration must be continuous and the two groups have approximately equal variance. The two groups do not need to have an equal number of scores. Basically, you pool the data

and rank it in ascending order, keeping track of which values belong to which group. For example, if we take the data concerning the novel objects (toys) (Table 8.6) and rank all of the data we would get Table 8.8.

Note that when three scores are identical they have identical ranking (e.g., the three 8's represent ranks 7, 8, 9 and thus each gets the rank 8). The ranks are then separated back into the two groups and one computes the U statistic. Most statistical programs will list the Mann–Whitney U test and provide a U score, a Z value, and a p value.

As can be gleaned from summary Table 8.9, the Mann–Whitney U test supports rejection of the H_o and supports the idea that the addition of the

TABLE 8.8 Mann–Whitney U Test—Enriched Environment

Scores	Rank	No toys	Toys
6	2		2
6	2		2
6	2		2
7	5		5
7	5		5
7	5	5	
8	8	8	
8	8		8
8	8		8
9	10	10	
10	11.5	11.5	
10	11.5	11.5	
11	13.5	13.5	
11	13.5		13.5
12	15		15
14	16		16
16	17.5	17.5	
16	17.5	17.5	
19	19	19	
20	20	20	
Sum	**210**	**133.5**	**76.5**

TABLE 8.9 Results: Mann–Whitney U Test—Enriched Environment

U	Z value	*p* value
21.5	−2.154	0.0312

novel objects had a significant influence on the learning. So why do the results differ between the **median test** and the **Mann–Whitney U test**? The simple answer to this is that the median test only looks at the difference between the medians of the two groups while the Mann–Whitney U looks at the difference in the shape and spread of the scores. The Mann–Whitney U is not really a test of medians. Because it can test a difference in the spread of scores, even when the medians are very similar, it is a more powerful statistic. The Mann–Whitney U test uses more information than the median test.

Reporting the results in a journal article requires reporting the U statistic (e.g., U = 31), the n's per group, and the significance. In the above experiment, one would write: "The Mann–Whitney *U* test revealed that the group exposed to toys (n = 10) was less likely to make errors in the maze acquisition than animals without toys (n = 10), U = 30, $p < 0.05$." Some statisticians suggest that as the groups become large (15–20), then the *z* value is reported instead of U.

The Mann–Whitney U test is really great as long as the n/group is less than 20. When the n is >20, the distribution begins to approach that of a parametric *t*-test and then the *t*-test has more power. However, if the n/group is small and there is large variance in the data then the Mann–Whitney U and *t*-test will give vastly different results. For example in a study looking at CYP-labeled neurons in the paraventricular nucleus of the hypothalamus, the following data were collected from a group of male and female rats (Table 8.10).

There is difference in the means for each group (Male, 138.7; Female, 223.3) and the question becomes: Are these two groups significantly different? If a parametric *t*-test is applied to these data one obtains the results in Table 8.11.

The *t*-statistic would support the H_o that there is no difference between the groups. If these same data were analyzed using the **Mann–Whitney U test**, the opposite conclusion would be obtained, **supporting the H_A** that the two groups were significantly different (Table 8.12).

The reason for the difference is due to the magnitude of the difference in the variance in each of the groups. The SD for the Male group = 38.5 while the Female group = 105.5. Homogeneity of variance is an important factor in using parametric statistics such as a *t*-test. These results can be subsequently graphed as shown in Figure 8.2.

TABLE 8.10 Mann–Whitney U Test: Neurons in Paraventricular Nucleus

Gender	Neurons
Male	128
Male	127
Male	105
Male	214
Male	138
Female	249
Female	296
Female	137
Female	387
Female	142
Female	129

TABLE 8.11 *t*-Test: Neurons in Paraventricular Nucleus

Mean difference	df	*t* value	*p* value
84.7	10	1.846	0.0946

TABLE 8.12 Mann–Whitney U Test—Neurons in Paraventricular Nucleus

U	*Z* value	*p* value
30	−2.082	0.0374

KOLMOGOROV–SMIRNOV TWO-SAMPLE TEST

The Kolmogorov–Smirnov two-sample test (K–S test) is a nonparametric statistic that tests the hypothesis that two independent samples have been drawn from the same population. It analyzes the data very differently from the Mann–Whitney U test in that rather than evaluating the location of the median it looks at a single maximum difference between the two distributions of scores. This test is often used when the question of equal variance between groups is very uncertain. It is based

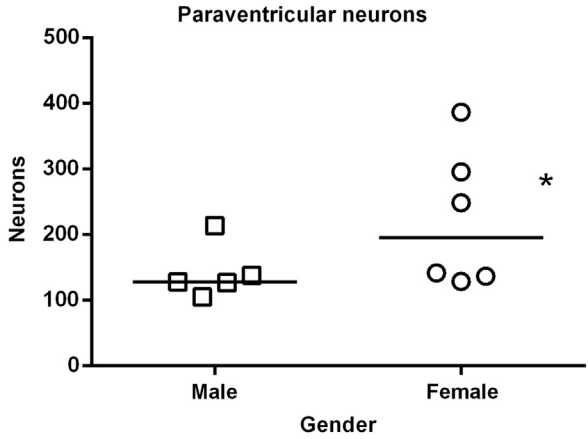

FIGURE 8.2 The Mann–Whitney U statistic is ideal when comparing two groups that have heterogeneity of variance. Note that all of the scores of the female group are above the median of the male group. Open squares and open circles represent individual subject data points. *$p < 0.05$.

on the chi-square distribution and relies on the relative expected frequencies. In the above example (Table 8.10) looking at gender-related neuron number in the paraventricular nucleus, the K–S test would report a p value of 0.14 and support the H_o. One important difference between this statistics and the Mann–Whitney U test is that the Mann–Whitney U test can handle tied scores quite well while the K–S does not. However, the K–S statistic can also be used in tests of goodness of fit. Most statisticians agree that the Mann–Whitney U is a better nonparametric statistic when you have two independent samples.

CHI-SQUARE

This is a rather interesting statistic that deals with the probability of a certain number of occurrences in a specific category. What this statistic does is test whether or not some observed frequency of events is statistically different from what one might expect to be the frequency by chance. If you were to ask a group of scientists for their opinion about the fairness of funding for basic research you would expect there to be a fair amount of disagreement. However, the level of disagreement may be related to a specific discipline in science such as neuroscience funding versus funding for cancer or cardiovascular research. If the variables of interest are categorical, such as scientific discipline and fairness of funding, then a chi-square is the way to go. In many neuroscience paradigms categorical variables could be gender or type of drug treatment.

The key here is that the variables are categorical in nature and not quantitative. What the researcher does is count the number of times a particular event or condition occurs and tests whether or not the values observed are different from what would be expected. This statistic can be used on nominal and ordinal data.

Chi-Square Goodness-of-Fit Test

This is a use of a chi-square test to evaluate whether or not a given frequency distribution fits a predicted distribution. In a recent study it was determined that the size of synapses (based on post synaptic density size) in the frontal cortex, Brodmann area 9, has a unique distribution in older rats. The laboratory is testing a compound that can mimic a disease state that is believed to alter the distribution of synapses in the frontal cortex by eliminating some of the small and very small synapses and augmenting the number of larger contacts. The compound was given to 15 rats and the synaptic contact size was evaluated for a random sample of 100 synapses in each subject (Table 8.13). The H_o simply states that the actual observed frequency of the distribution synapse size in the drug-treated animals is the same as that previously observed (expected) in the nontreated animals.

If we run a simple chi-square test on the data in Table 8.13 we get $X^2 = 9.92$. Checking a table of critical values for chi-square with $k - 1$ degrees of freedom (**k** is the total number of categories) this value is significant at the 0.05 level, indicating that the H_o is rejected and the hypothesis that the compound alters synapse size is supported.

When reporting these results in a journal article it is important to report the actual X^2 value along with the degrees of freedom and the p values. This would look like: "A chi-square goodness-of-fit test showed that the compound had a significant effect on the synapse size ($p < 0.05$)."

TABLE 8.13 Chi-Square Goodness of Fit: Synapse Size

Synapse size	Expected percentage (%)	Observed percentage (%)
Very large	5	8
Large	20	28
Medium	40	42
Small	30	19
Very small	5	3

Chi-Square Test of Independence

The chi-square statistic can be used to test for the independence of variables. For example, you are trying to determine which therapy might enhance recovery from a moderate spinal cord injury in rats. You have an antioxidant compound that improves recovery but believe the recovery can be enhanced if combined with locomotor training or no locomotor training. To show improvement, the animals must reach a specific criterion on a mesh grid. If the animals do not reach criterion, then there is no improvement but if they do reach criterion, then there is improvement. A total of 30 animals are pretested on a mesh walking grid and subjected to a spinal cord contusion. Two days post injury they are evaluated for their locomotor ability and subsequently given the compound with or without the added locomotor training for the next 14 days. At 16 days post injury they are tested on the walking grid and compared to their scores on day 2 post trauma. Each rat is then categorized as either showing improvement or no improvement. The combined data are shown in Table 8.14.

The H_o is that there is no difference in the recovery improvement with the addition of the locomotor training. If we run a simple chi-square test on these data we get $X^2 = \mathbf{3.33}$. In testing independence in any contingency table the degrees of freedom are given by (rows -1) (columns -1). Checking a table of critical values for chi-square with 1 degree of freedom we find that a value of **3.84** is necessary for significance at the 0.05 level. Our obtained value is close to reaching this value but does not. Consequently the H_o cannot be rejected. What this essentially says is that locomotor training and recovery improvement are independent and do not have a relationship even though you might think it does. Perhaps if there were a larger number of subjects per group it would be significant. The closer the numbers in each cell are to each other, the greater the probability that they are independent. It is important to note that the degrees of freedom for the chi-square statistic differ depending on its use as a test of independence and goodness of fit. What is interesting to note is the fact that the chi-square distribution is very asymmetrical especially when the degrees of freedom are very small. As the degrees of freedom become larger the distribution becomes more symmetrical.

TABLE 8.14 Chi-Square Test of Independence: Locomotor Training

	Improvement	**No improvement**	**Totals**
Training	10	5	**15**
No training	5	10	**15**
Totals	**15**	**15**	**30**

In the above example (Table 8.14) a 2 × 2 contingency table was used. This statistic will also work for 3 × 3 or 2 × 5 contingency tables. It is not necessary that there be the same number of subjects in each group. For example, one might have 20 subjects in the training group and only 13 in the no training group. One just does not want the differences in sample size to be extreme such as 20 in one group and only 5 in another.

When reporting these results in a journal article it is important to report the actual X^2 value along with the degrees of freedom and the p values. This would look like: "A 2 × 2 chi-square test revealed that the relationship between training and locomotor activity was not significant ($p > 0.05$)."

FISHER'S EXACT TEST

This statistic is like the chi-square test, but is very useful when the number of frequencies per cell is small (i.e., one of the cells has an expected count of less than 5) and it is a 2 × 2 contingency table. Like the chi-square test, the assumption is that there is no relationship between the two variables being tested. This statistic calculates all possible outcomes by rearrangements of the observations and comparing the number of unusual rearrangements to the observed counts under the assumption that there is no association between the two variables. This statistic calculates what is known as a phi (Φ) coefficient, which is some indication of the effect size. For example, if in a set of animal experiments there is an unexpected level of death in one of the experimental conditions, and you want to determine if it may be related to one of the experimental conditions, then Fisher's exact test is appropriate. The H_o would be that there is no difference in the proportion of deaths in the two groups. You cannot use the chi-square because the expected frequencies are less than 5.

When reporting these results in a journal article, it is important to show the frequency table. It is also important to report the phi (Φ) coefficient (e.g., $\Phi = 0.648$) along with the X^2 value and the p value.

KRUSKAL–WALLIS ONE-WAY ANALYSIS OF VARIANCE

If there are only two independent groups being tested and the data probably do not follow a normal distribution then the Mann–Whitney U test (Wilcoxon rank sum test) is usually the most appropriate. However, if there are multiple groups being compared from possibly different populations, then the Kruskal–Wallis (KW) test would be the statistic of choice. This is equivalent to the parametric one-way analysis of variance.

Essentially what this test does is determine if a set of independent samples are from simply random samples from the same population or from different populations. The H_o states that there is no difference between the groups, while the H_A states that at least one of the groups will be different from the others. Because this test is nonparametric it makes no prediction about the different population means but rather compares the medians. The sample size does not have to be the same for each group but should not differ greatly. However, this test works for small sample sizes such as 5 or 6.

During a lunch break, one of your colleagues mentions that she has discovered that several different natural compounds improve learning and believes it is due to an increase in neuronal firing rate in the CA1 region of the hippocampus. You decide to test whether or not any of these compounds can increase the firing rate and if they are different from each other. You both collaborate and collect the following neurophysiology data showing percent change in baseline firing rate using a slice preparation (Table 8.15). Different animals are used to test each compound and the values indicate the mean percent increase in firing rate for each animal. In other words, each group of compounds was tested on five different animals with a total $n = 25$.

Because there is wide variance in the scores, you do not think they follow a normal distribution and decide to apply the KW statistic. With the assistance of your statistical software program you determine that the KW value = 10.3 (Table 8.16). The degrees of freedom for the KW are calculated as $k - 1$, with k being the number of groups. To determine the level of significance you check the chi-square distribution table. For $df = 4$, a value of 9.49 is necessary to reach significance at the 0.05 level. The H_o is not supported and you feel that the H_A is.

All this test has told you so far is that there is a difference somewhere between the different groups but not which ones (Table 8.17). To determine this, a subsequent series of tests need to be applied to the data.

There is little agreement as to which subsequent test to use and many statisticians recommend either the Dunn's test or the Mann–Whitney U

TABLE 8.15 Kruskal–Wallis ANOVA: Hippocampal Firing

KrO	SJW	CC	PC	HCE
9	8	50	29	35
5	42	45	10	48
14	33	51	35	53
13	20	15	17	26
17	30	38	19	12

TABLE 8.16 Results: Kruskal–Wallis ANOVA—Hippocampal Firing

df	4
No. groups	5
No. ties	2
H value	10.3
p value	0.0367

TABLE 8.17 Results: Kruskal–Wallis ANOVA—Hippocampal Firing

Group	n	Mean rank
KrO	5	5.3
SJW	5	13.0
CC	5	19.0
PC	5	11.2
HCE	5	16.5

statistic. Many statistical packages have these options following the KW analysis. The Dunn's test (Chapter 6) shows that only KrO and CC differ significantly. This is a rather conservative test but recommended by many statisticians. If you decide to use the Mann–Whitney U statistic then you simply have to pair each of the groups, rank the data from smallest to highest, and calculate the U score. This is repeated for each pair of scores (e.g., KrO vs SJW; KrO vs CC; PC vs HCE) and one then draws conclusions based on the outcomes (Table 8.18). As one might suspect, this can be very time-consuming if there are many groups used in the KW analysis. An alternative multiple comparison method is given by Siegel and Castellan.[1] Your friendly biostatistician can help you with this.

Often these results are graphed as shown in Figure 8.3. Since the analysis used nonparametric techniques one must plot the median and not the mean such as in this scatterplot.

When reporting the results of the KW test one must include the KW or H score, the degrees of freedom, and the *p* value. In the journal the text for the above experiment would be: The KW test indicated a significant effect ($KW = 10.23$, $df = 4$, $p < 0.05$) or ($H_{(4)} = 10.23$, $p < 0.05$). In the methods section it is important to state what statistic was used to contrast the different groups, i.e., Mann–Whitney U test.

TABLE 8.18 Multiple Comparisons Following Kruskal–Wallis ANOVA

Contrast	Dunn's	Mann–Whitney U
KrO-SJW	ns	ns
KrO-CC	$p < 0.05$	$p < 0.02$
KrO-PC	ns	ns
KrO-HCE	ns	$p < 0.05$
SJW-CC	ns	ns
SJW-PC	ns	ns
SWJ-HCE	ns	ns
CC-PC	ns	ns
CC-HCE	ns	ns
PC-HCE	ns	ns

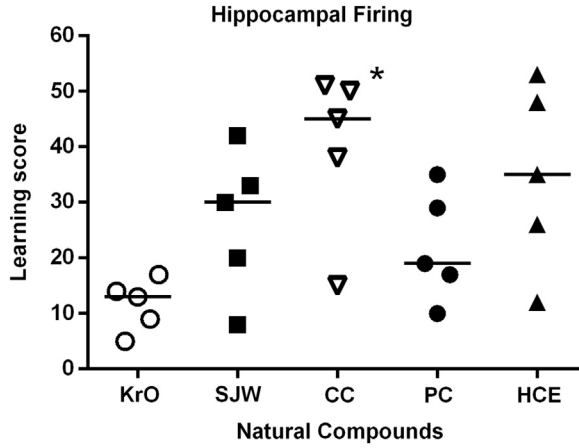

FIGURE 8.3 Note the differences in variance between the groups. Because of this heterogeneity a nonparametric ANOVA would provide a very false analysis of the data. Each symbol represents a single subject's learning score. The line represents the group median. *$p < 0.05$ compared to KrO.

FRIEDMAN ONE-WAY REPEATED MEASURE ANALYSIS OF VARIANCE BY RANKS

This nonparametric test is used to compare three or more **matched** groups. It is sometimes simply called the Friedman test and often cited as Friedman's two-way ANOVA, although it is really a one-way ANOVA. **There is not a true nonparametric two-way ANOVA**. This Friedman's test is an ideal statistic to use for a repeated measures type of experiment to determine if a particular factor has an effect. As an example look at the swim speed data again in a Morris water maze (Table 8.19).

Here is a group of six different rats with their swim speeds on different days as a result of a change in the water temperature. The question is whether or not changing the water temperature in the maze affects the swimming speed. The H_o is that the swim speed will be the same regardless of the water temperature. Essentially what this test does is rank the swim speeds for each subject across the various water temperatures and in this way each animal is compared to itself. The magnitude of the differences with each subject is not important. This test results in a Friedman statistic (F_r) and the possible significance of this value is then looked up in a specific table literally called the critical values for the Friedman two-way analysis of variance by ranks statistic, which is available on the web. In this experiment there are four treatments with a total of six subjects and the $F_r = 14.90$. In this table a value of 7.6 is necessary for $\alpha < 0.05$ and 10.0 for significance at the 0.01 level. Thus the H_A is supported. A Dunn's test is used to subsequently compare the different treatments following the Friedman analysis. In this particular example, the 20 °C temperature was significantly different from the 27 and 32 °C temperatures but not the 24 °C.

When reporting the results of the Friedman test one must include the number of subjects, the F_r score, and the p value. Some journals suggest that the degree of freedom (here 4,6) should also be given. This might

TABLE 8.19 Friedman Repeated Measure ANOVA: Swim Speed

Animal	20 °C	24 °C	27 °C	32 °C
1	39	38	34	33
2	29	25	20	20
3	36	37	29	24
4	25	29	18	19
5	31	27	24	22
6	34	33	30	31

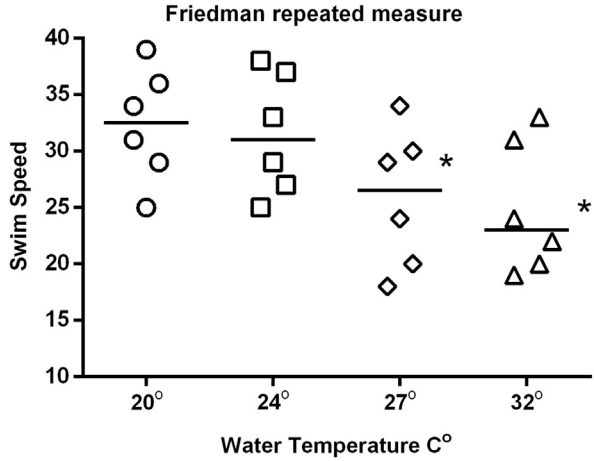

FIGURE 8.4 The nonparametric version of a repeated measures one-way ANOVA is the Friedman test. Each symbol represents an individual data point for each subject under different water temperature conditions. The horizontal line represents the group median. *$p < 0.05$ compared to the 20 °C cohort.

appear as: "a Friedman analysis of variance was applied and indicated that a change in water temperature significantly altered the swim speeds ($F_r = 14.90$, df 4,6, $p < 0.01$). A Dunn's test revealed that the swim speed in 20 °C temperature water was significantly increased compared to the 27 and 32 °C water temperatures." These data can be graphed as shown in Figure 8.4 using a scatterplot and the median.

SPEARMAN'S RANK ORDER CORRELATION

This nonparametric statistic, also known as Spearman's rho, is the equivalent of the parametric Pearson correlation coefficient. The Spearman's rho is named after Charles Spearman, an experimental psychologist at the University College of London, who invented the procedure in 1904. It is used when one or both of the variables are ordinal scaling. It is based on ranks of the data and not on the data itself and thus is resistant to outliers. This is a very important feature since outliers can significantly alter the outcome of the Pearson statistic (Chapter 5). The H_o is that the two variables are independent and have no relationship to each other. Like the Pearson statistic, the range of the correlation is +1 to −1 with a zero indicating no correlation between the two variables. An added characteristic of the Spearman's rho is that it does not assume that the correlation will be linear. Consequently, one can use this statistic even if the relationship is curved. The one basic assumption is that the

underlying relationship is **monotonic**, which means that as one variable becomes larger the other becomes consistently either larger or smaller. Statisticians might say that the two variables under investigation covary, which means simply that as one variable increases, the other variables will either increase or decrease. A **nonmonotonic** relationship would be one where one variable becomes larger while the other sometimes becomes larger and then sometimes becomes smaller. The other basic assumption for using the Spearman is that the observations are independent. The interpretation of Spearman's rho is the same as that for the Pearson statistic. Let us take the example of examining the possible association between synaptic numbers in a small part of the frontal cortex and an individual's global cognitive score (Table 8.20).

This particular set of scores includes both ordinal and ratio types of data. There is also a considerable range in the scores, making the use of

TABLE 8.20 Spearman's Rho Correlation: Synapses × Cognition

Subject	Synapses	Global cognition
1	380.1	26
2	415.2	24
3	381.1	16
4	548.5	27
5	700.5	26
6	354.1	13
7	418.5	27
8	525.0	25
9	532.3	23
10	497.3	21
11	169.8	9
12	306.5	24
13	378.6	28
14	121.6	11
15	296.8	15
16	431.7	28
17	387.2	23
18	494.6	20
19	433.2	16

Spearman's rank correlation ideal for these data. This statistic can also be used with interval scale data. What this test does is first rank all of the data from highest to lowest in terms of one of the variables. Most statisticians use the ordinal variable to rank the data (Table 8.21). The highest score is assigned a 1 and then a 2 to the next highest score and so on. When there are tied scores involved, these are all assigned the same average rank. For example, in this set of data, the first and second scores are a 28 so they both get 1.5. This rank of 1.5 would occupy both the first and the second places in the list of scores. It is important to note that if there are too many tied scores it will affect the size of rho (ρ), the end statistic.

Almost every statistical software program has the option of using this nonparametric statistic. The printout may differ but should contain information such as the "*r*" or rho value and the *p* value, usually a

TABLE 8.21 Spearman's Rho Correlation: Synapses × Cognition

Rank	Global cognition	Synapses
1.5	28	431.7
1.5	28	378.6
3.5	27	548.5
3.5	27	418.5
5.5	26	700.5
5.5	26	380.1
7	25	525.0
8.5	24	415.2
8.5	24	306.5
10.5	23	532.3
10.5	23	387.2
12	21	497.3
13	20	494.6
14.5	16	381.1
14.5	16	433.2
16	15	296.8
17	13	354.1
18	11	121.6
19	9	169.8

TABLE 8.22 Results: Spearman's Rho Correlation—Synapses × Cognition

Spearman's *r*	0.475
p value	0.05
Rho corrected for ties	0.473
Tied *p* value	0.05
No. of ties global cognition	6

two-tailed *p* value (Table 8.22). If there are ties, it is important to note the tied *r* value and the tied *p* value. For this example, what we can say is that the relationship between the number of synapses in the frontal cortex and the performance on a global cognition test for the 19 subjects is 0.473, which is significant at $p < 0.05$. If the H_o stated that there was no relationship then these results would support the H_A.

When reporting the results of the Spearman's rho, it can be graphed in the same way as the Pearson correlation (Figure 8.5). In the text it is important to include the number of independent measures (here the $n = 19$), the names of the two variables being evaluated (global cognition; synapses), the value of rho ($r_s = 0.475$), and the *p* value ($p < 0.05$). Some journals also require stating the number of ties (6) in the analysis and the

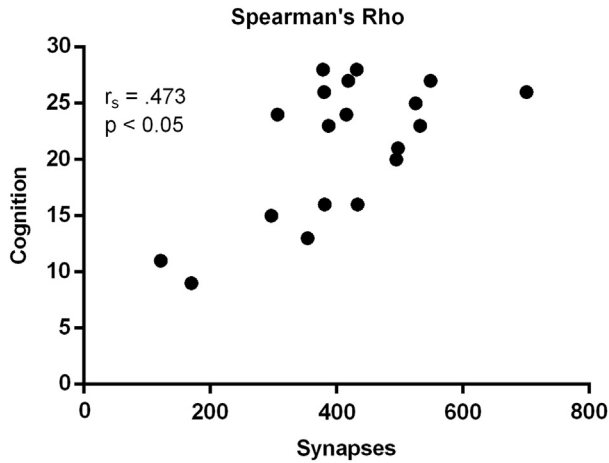

FIGURE 8.5 Spearman's rho (r_s) statistic is the equivalent of the parametric Person's product–moment correlation. Like the parametric correlation, it is improper to place a regression line on the graph. Each circle represents an individual data point.

degrees of freedom. For the Spearman's rho, the degrees of freedom are simply the number of pairs in the sample minus 2 (n − 2). For this example df = 17. Reports in the literature might include: "There was a significant correlation between the individual's global cognition score and the synaptic counts in the superior frontal cortex [r_s (17) = 0.475, $p < 0.05$]. There were six tied scores in the data."

One final word on reporting the results of the Spearman's rank order correlation—when graphing the results, it is inappropriate to place a regression line on a graph because this line indicates a cause and effect relationship, which may or may not be the case. This is the same as that discussed in Chapter 5 for the Pearson rho. Some authors place a dotted "regression" line and explain in the figure legend that this line merely indicates the direction of the association and does not imply cause and effect. Some journals allow this and others do not.

KENDALL RANK ORDER CORRELATION COEFFICIENT

This nonparametric statistic is very similar to Spearman's rho. It is sometimes referred to as Kendall's tau (T). It is calculated differently from Spearman's rho and consequently has different underlying scales resulting in different correlation values. If the data concerning global cognition and synaptic numbers are analyzed with Kendall's T one would get the results shown in Table 8.23.

The interpretation of Kendall's T is complicated and different from that of Spearman's rho. Spearman's rho can be interpreted the same as the parametric Pearson product−moment correlation in terms of the proportion of the variability accounted for by each variable, as described above. Kendall's T deals with the probability that the two variables **are** in a specific order as opposed to the probability that they **could** be in a different order. A more detailed explanation is beyond the scope of the present discussion. This statistic is not used very often and researchers

TABLE 8.23 Kendall Rank Order Correlation Coefficient

Kendall's T	0.368
p value	0.028
Rho corrected for ties	0.375
Tied *p* value	0.025
No. of ties global cognition	6

who wish to apply it to their data should first consult with a biostatistician. It is inappropriate to analyze some of the data with Kendall's T and some with Spearman's rho.

NONPARAMETRIC AND DISTRIBUTION-FREE ARE NOT REALLY THE SAME

The actual terms nonparametric and distribution-free are not completely synonymous but the popular statistical press has reinforced this idea. One must remember that the nonparametric tests, just like their counterparts such as the *t*-test and the *F*-test, are really only estimations of the population as characterized by the sample. When interval or ratio data are converted to ranks or frequencies and evaluated by nonparametric techniques they lose power. There is a greater chance of a Type II error. One also must remember that some of the information used when collecting these data has also been lost in the sense that it no longer plays the same influential role in the interpretation of the results.

SUMMARY

- Nonparametric statistics should be used when the data contain a large amount of variance in one or more groups.
- Nonparametric tests can be used on very small samples.
- For every parametric statistic, with the exception of the two-way ANOVA, there is an equivalent nonparametric test.
- One should never report the mean and standard deviation when using a nonparametric statistic.

References

1. Siegel S, Castellan NJ. *Nonparametric Statistics for the Behavioral Sciences*. Boston: McGraw Hill; 1988.
2. Hollanders M, Wolfe DA. *Nonparametric Statistical Methods*. New York: John Wiley & Sons; 1999.
3. Corder GW, Foreman DI. *Nonparametric Statistics for Non-statisticians*. New York: John Wiley & Sons; 2009.