CK-12 Probability and Statistics - Advanced (Second Edition)

Ellen Lawsky
Larry Ottman
Raja Almukkahal
Brenda Meery
Danielle DeLancey

Chapter 12

Non-Parametric Statistics

AUTHORS
Ellen Lawsky
Larry Ottman
Raja Almukkahal
Brenda Meery
Danielle DeLancey

To access a customizable version of this book, as well as other interactive content, visit www.ck12.org

CK-12 Foundation is a non-profit organization with a mission to reduce the cost of textbook materials for the K-12 market both in the U.S. and worldwide. Using an open-content, web-based collaborative model termed the **FlexBook®**, CK-12 intends to pioneer the generation and distribution of high-quality educational content that will serve both as core text as well as provide an adaptive environment for learning, powered through the **FlexBook Platform®**.

Printed: December 17, 2014

flexbook
next generation textbooks

CHAPTER 12 Non-Parametric Statistics

Chapter Outline

12.1 INTRODUCTION TO NON-PARAMETRIC STATISTICS
12.2 THE RANK SUM TEST AND RANK CORRELATION
12.3 THE KRUSKAL-WALLIS TEST AND THE RUNS TEST

12.1 Introduction to Non-Parametric Statistics

Learning Objectives

- Understand situations in which non-parametric analytical methods should be used and the advantages and disadvantages of each of these methods.
- Understand situations in which the sign test can be used and calculate z-scores for evaluating a hypothesis using matched pair data sets.
- Use the sign test to evaluate a hypothesis about the median of a population.
- Examine a categorical data set to evaluate a hypothesis using the sign test.
- Understand the signed-ranks test as a more precise alternative to the sign test when evaluating a hypothesis.

Introduction

In previous lessons, we discussed the use of the normal distribution, Student's t-distribution, and the F-distribution in testing various hypotheses. With each of these distributions, we made certain assumptions about the populations from which our samples were drawn. Specifically, we made assumptions that the underlying populations were normally distributed and that there was homogeneity of variance within the populations. But what do we do when we have data that are not normally distributed or not homogeneous with respect to variance? In these situations, we use something called non-parametric tests.

These tests include tests such as the sign test, the sign-ranks test, the ranks-sum test, the Kruskal-Wallis test, and the runs test. While parametric tests are preferred, since they are more powerful, they are not always applicable. The following sections will examine situations in which we would use non-parametric methods and the advantages and disadvantages of using these methods.

Situations Where We Use Non-Parametric Tests

If *non-parametric tests* have fewer assumptions and can be used with a broader range of data types, why don't we use them all the time? The reason is because there are several advantages of using parametric tests. They are more robust and have greater power, which means that they have a greater chance of rejecting the null hypothesis relative to the sample size when the null hypothesis is false.

However, parametric tests demand that the data meet stringent requirements, such as normality and homogeneity of variance. For example, a one-sample t-test requires that the sample be drawn from a normally distributed population. When testing two independent samples, not only is it required that both samples be drawn from normally distributed populations, but it is also required that the standard deviations of the populations be equal. If either of these conditions is not met, our results are not valid.

As mentioned, an advantage of non-parametric tests is that they do not require the data to be normally distributed. In addition, although they test the same concepts, non-parametric tests sometimes have fewer calculations than their parametric counterparts. Non-parametric tests are often used to test different types of questions and allow us to perform analysis with categorical and rank data. The table below lists the parametric tests, their non-parametric counterparts, and the purpose of each test.

Commonly Used Parametric and Non-parametric Tests

TABLE 12.1:

Parametric Test (Normal Distributions)	Non-parametric Test (Non-normal Distributions)	Purpose of Test
t-test for independent samples	Rank sum test	Compares means of two independent samples
Paired t-test	Sign test	Examines a set of differences of means
Pearson correlation coefficient	Rank correlation test	Assesses the linear association between two variables.
One-way analysis of variance (F-test)	Kruskal-Wallis test	Compares three or more groups
Two-way analysis of variance	Runs test	Compares groups classified by two different factors

The Sign Test

One of the simplest non-parametric tests is the *sign test*. The sign test examines the difference in the medians of matched data sets. It is important to note that we use the sign test only when testing if there is a difference between the matched pairs of observations. This test does not measure the magnitude of the relationship-it simply tests whether the differences between the observations in the matched pairs are equally likely to be positive or negative. Many times, this test is used in place of a paired t-test.

For example, we would use the sign test when assessing if a certain drug or treatment had an impact on a population or if a certain program made a difference in behavior. We first determine whether there is a positive or negative difference between each of the matched pairs. To determine this, we arrange the data in such a way that it is easy to identify what type of difference that we have. Let's take a look at an example to help clarify this concept.

Example: Suppose we have a school psychologist who is interested in whether or not a behavior intervention program is working. He examines 8 middle school classrooms and records the number of referrals written per month both before and after the intervention program. Below are his observations:

TABLE 12.2:

Observation Number	Referrals Before Program	Referrals After Program
1	8	5
2	10	8
3	2	3
4	4	1
5	6	4
6	4	1
7	5	7
8	9	6

Since we need to determine the number of observations where there is a positive difference and the number of observations where there is a negative difference, it is helpful to add an additional column to the table to classify each observation as such (see below). We ignore all zero or equal observations.

TABLE 12.3:

Observation Number	Referrals Before Program	Referrals After Program	Change
1	8	5	−
2	10	8	−
3	2	3	+
4	4	1	−
5	6	4	−
6	4	1	−
7	5	7	+
8	9	6	−

The test statistic we use is $\frac{|\text{number of positive changes} - \text{number of negative changes}|-1}{\sqrt{n}}$.

If the sample has fewer than 30 observations, we use the t-distribution to determine a critical value and make a decision. If the sample has more than 30 observations, we use the normal distribution.

Our example has only 8 observations, so we calculate our t-score as shown below:

$$t = \frac{|2-6|-1}{\sqrt{8}} = 1.06$$

Similar to other hypothesis tests using standard scores, we establish null and alternative hypotheses about the population and use the test statistic to assess these hypotheses. As mentioned, this test is used with paired data and examines whether the medians of the two data sets are equal. When we conduct a pre-test and a post-test using matched data, our null hypothesis is that the difference between the data sets will be zero. In other words, under our null hypothesis, we would expect there to be some fluctuations between the pre-test and post-test, but nothing of significance. Therefore, our null and alternative hypotheses would be as follows:

$$H_0 : m = 0$$
$$H_a : m \neq 0$$

With the sign test, we set criterion for rejecting the null hypothesis in the same way as we did when we were testing hypotheses using parametric tests. For the example above, if we set $\alpha = 0.05$, we would have critical values at 2.36 standard scores above and below the mean. Since our standard score of 1.06 is less than the critical value of 2.36, we would fail to reject the null hypothesis and cannot conclude that there is a significant difference between the pre-test and post-test scores.

When we use the sign test to evaluate a hypothesis about the median of a population, we are estimating the likelihood, or the probability, that the number of successes would occur by chance if there was no difference between pre-test and post-test data. When working with small samples, the sign test is actually the binomial test, with the null hypothesis being that the proportion of successes will equal 0.5.

Example: Suppose a physical education teacher is interested in the effect of a certain weight-training program on students' strength. She measures the number of times students are able to lift a dumbbell of a certain weight before the program and then again after the program. Below are her results:

TABLE 12.4:

Before Program	After Program	Change
12	21	+
9	16	+
11	14	+
21	36	+
17	28	+
22	20	−
18	29	+
11	22	+

If the program had no effect, then the proportion of students with increased strength would equal 0.5. Looking at the data above, we see that 7 of the 8 students had increased strength after the program. But is this statistically significant? To answer this question, we use the binomial formula, which is as follows:

$$P(r) = \frac{N!}{r!(N-r)!} p^r (1-p)^{N-r}$$

Using this formula, we need to determine the probability of having either 7 or 8 successes as shown below:

$$P(7) = \frac{8!}{7!(8-7)!} 0.5^7 (1-0.5)^{8-7} = (8)(00391) = 0.03125$$
$$P(8) = \frac{8!}{8!(8-8)!} 0.5^8 (1-0.5)^{8-8} = 0.00391$$

To determine the probability of having either 7 or 8 successes, we add the two probabilities together and get $0.03125 + 0.00391 = 0.0352$. This means that if the program had no effect on the matched data set, we have a 0.0352 likelihood of obtaining the number of successes that we did by chance.

Using the Sign Test to Examine Categorical Data

We can also use the sign test to examine differences and evaluate hypotheses with categorical data sets. Recall that we typically use the chi-square distribution to assess categorical data. We could use the sign test when determining if one categorical variable is really more than another. For example, we could use this test if we were interested in determining if there were equal numbers of students with brown eyes and blue eyes. In addition, we could use this test to determine if equal numbers of males and females get accepted to a four-year college.

When using the sign test to examine a categorical data set and evaluate a hypothesis, we use the same formulas and methods as if we were using nominal data. The only major difference is that instead of labeling the observations as positives or negatives, we would label the observations with whatever dichotomy we want to use (male/female, brown/blue, etc.) and calculate the test statistic, or probability, accordingly. Again, we would not count zero or equal observations.

Example: The UC admissions committee is interested in determining if the numbers of males and females who are accepted into four-year colleges differ significantly. They take a random sample of 200 graduating high school seniors who have been accepted to four-year colleges. Out of these 200 students, they find that there are 134 females and 66 males. Do the numbers of males and females accepted into colleges differ significantly? Since we have a large sample, calculate the z-score and use $\alpha = 0.05$.

To answer this question using the sign test, we would first establish our null and alternative hypotheses:

$$H_0 : m = 0$$
$$H_a : m \neq 0$$

This null hypothesis states that the median numbers of males and females accepted into UC schools are equal.

Next, we use $\alpha = 0.05$ to establish our critical values. Using the normal distribution table, we find that our critical values are equal to 1.96 standard scores above and below the mean.

To calculate our test statistic, we use the following formula:

$$\frac{|\text{number of positive changes } - \text{number of negative changes}| - 1}{\sqrt{n}}$$

However, instead of the numbers of positive and negative observations, we substitute the number of females and the number of males. Because we are calculating the absolute value of the difference, the order of the variables does not matter. Therefore, our z-score can be calculated as shown:

$$z = \frac{|134 - 66| - 1}{\sqrt{200}} = 4.74$$

With a calculated test statistic of 4.74, we can reject the null hypothesis and conclude that there is a difference between the number of graduating males and the number of graduating females accepted into the UC schools.

The Benefit of Using the Sign Rank Test

As previously mentioned, the sign test is a quick and easy way to test if there is a difference between pre-test and post-test matched data. When we use the sign test, we simply analyze the number of observations in which there is a difference. However, the sign test does not assess the magnitude of these differences.

A more useful test that assesses the difference in size between the observations in a matched pair is the *sign rank test*. The sign rank test (also known as the *Wilcoxon sign rank test*) resembles the sign test, but it is much more sensitive. Similar to the sign test, the sign rank test is also a nonparametric alternative to the paired Student's t-test. When we perform this test with large samples, it is almost as sensitive as Student's t-test, and when we perform this test with small samples, it is actually more sensitive than Student's t-test.

The main difference with the sign rank test is that under this test, the hypothesis states that the difference between observations in each data pair (pre-test and post-test) is equal to zero. Essentially, the null hypothesis states that the two variables have identical distributions. The sign rank test is much more sensitive than the sign test, since it measures the difference between matched data sets. Therefore, it is important to note that the results from the sign and the sign rank test could be different for the same data set.

To conduct the sign rank test, we first rank the differences between the observations in each matched pair, without regard to the sign of the difference. After this initial ranking, we affix the original sign to the rank numbers. All equal observations get the same rank and are ranked with the mean of the rank numbers that would have been assigned if they had varied. After this ranking, we sum the ranks in each sample and then determine the total number of observations. Finally, the one sample z-statistic is calculated from the signed ranks. For large samples, the z-statistic is compared to percentiles of the standard normal distribution.

It is important to remember that the sign rank test is more precise and sensitive than the sign test. However, since we are ranking the nominal differences between variables, we are not able to use the sign rank test to examine the differences between categorical variables. In addition, this test can be a bit more time consuming to conduct, since the figures cannot be calculated directly in Excel or with a calculator.

Lesson Summary

We use non-parametric tests when the assumptions of normality and homogeneity of variance are not met.

There are several different non-parametric tests that we can use in lieu of their parametric counterparts. These tests include the sign test, the sign rank test, the rank-sum test, the Kruskal-Wallis test, and the runs test.

The sign test examines the difference in the medians of matched data sets. When testing hypotheses using the sign test, we can calculate the standard z-score when working with large samples or use the binomial formula when working with small samples.

We can also use the sign test to examine differences and evaluate hypotheses with categorical data sets.

A more precise test that assesses the difference in size between the observations in a matched pair is the sign rank test.

12.2 The Rank Sum Test and Rank Correlation

Learning Objectives

- Understand the conditions for use of the rank sum test to evaluate a hypothesis about non-paired data.
- Calculate the mean and the standard deviation of rank from two non-paired samples and use these values to calculate a z-score.
- Determine the correlation between two variables using the rank correlation test for situations that meet the appropriate criteria, using the appropriate test statistic formula.

Introduction

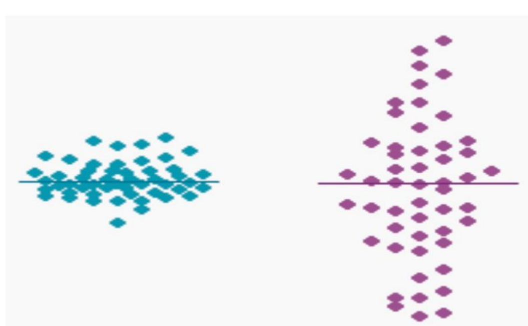
In the previous lesson, we explored the concept of nonparametric tests. We explored two tests-the sign test and the sign rank test. We use these tests when analyzing matched data pairs or categorical data samples. In both of these tests, our null hypothesis states that there is no difference between the medians of these variables. As mentioned, the sign rank test is a more precise test of this question, but the test statistic can be more difficult to calculate.

But what happens if we want to test if two samples come from the same non-normal distribution? For this type of question, we use the rank sum test (also known as the *Mann-Whitney v-test*). This test is sensitive to both the median and the distribution of the sample and population.

In this section, we will learn how to conduct hypothesis tests using the Mann-Whitney v-test and the situations in which it is appropriate to do so. In addition, we will also explore how to determine the correlation between two variables from non-normal distributions using the rank correlation test for situations that meet the appropriate criteria.

Conditions for Use of the Rank Sum Test to Evaluate Hypotheses about Non-Paired Data

The *rank sum test* tests the hypothesis that two independent samples are drawn from the same population. Recall that we use this test when we are not sure if the assumptions of normality or homogeneity of variance are met. Essentially, this test compares the medians and the distributions of the two independent samples. This test is considered stronger than other nonparametric tests that simply assess median values. For example, in the image below, we see that the two samples have the same median, but very different distributions. If we were assessing just the median value, we would not realize that these samples actually have distributions that are very distinct.



When performing the rank sum test, there are several different conditions that need to be met. These include the following:

- Although the populations need not be normally distributed or have homogeneity of variance, the observations must be continuously distributed.
- The samples drawn from the population must be independent of one another.
- The samples must have 5 or more observations. The samples do not need to have the same number of observations.
- The observations must be on a numeric or ordinal scale. They cannot be categorical variables.

Since the rank sum test evaluates both the medians and the distributions of two independent samples, we establish two null hypotheses. Our null hypotheses state that the two medians and the two standard deviations of the independent samples are equal. Symbolically, we could say $H_0 : m_1 = m_2$ and $\sigma_1 = \sigma_2$. The alternative hypotheses state that there is a difference in the medians and the standard deviations of the samples.

Calculating the Mean and the Standard Deviation of Rank to Calculate a

When performing the rank sum test, we need to calculate a figure known as the *U-statistic*. This statistic takes both the median and the total distribution of the two samples into account. The U-statistic actually has its own distribution, which we use when working with small samples. (In this test, a small sample is defined as a sample less than 20 observations.) This distribution is used in the same way that we would use the t-distribution and the chi-square distribution. Similar to the t-distribution, the *U-distribution* approaches the normal distribution as the sizes of both samples grow. When we have samples of 20 or more, we do not use the U-distribution. Instead, we use the U-statistic to calculate the standard z-score.

To calculate the U-statistic, we must first arrange and rank the data from our two independent samples. First, we must rank all values from both samples from low to high, without regard to which sample each value belongs to. If two values are the same, then they both get the average of the two ranks for which they tie. The smallest number gets a rank of 1, and the largest number gets a rank of n, where n is the total number of values in the two groups. After we arrange and rank the data in each of the samples, we sum the ranks assigned to the observations. We record both the sum of these ranks and the number of observations in each of the samples. After we have this information, we can use the following formulas to determine the U-statistic:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$
$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

where:

n_1 is the number of observations in sample 1.

n_2 is the number of observations in sample 2.

R_1 is the sum of the ranks assigned to sample 1.

R_2 is the sum of the ranks assigned to sample 2.

We use the smaller of the two calculated test statistics (i.e., the lesser of U_1 and U_2) to evaluate our hypotheses in smaller samples or to calculate the z-score when working with larger samples.

When working with larger samples, we need to calculate two additional pieces of information: the mean of the sampling distribution, μ_U, and the standard deviation of the sampling distribution, σ_U. These calculations are relatively straightforward when we know the numbers of observations in each of the samples. To calculate these figures, we use the following formulas:

$$\mu_U = \frac{n_1 n_2}{2} \text{ and } \sigma_U = \sqrt{\frac{n_1(n_2)(n_1+n_2+1)}{12}}$$

Finally, we use the general formula for the test statistic to test our null hypothesis:

$$z = \frac{U - \mu_U}{\sigma_U}$$

Example: Suppose we are interested in determining the attitudes on the current status of the economy from women who work outside the home and from women who do not work outside the home. We take a sample of 20 women who work outside the home (sample 1) and a sample of 20 women who do not work outside the home (sample 2) and administer a questionnaire that measures their attitudes about the economy. These data are found in the tables below:

TABLE 12.5:

Women Working Outside the Home	Women Working Outside the Home
Score	Rank
9	1
12	3
13	4
19	8
21	9
27	13
31	16
33	17
34	18
35	19
39	21
40	22
44	25
46	26
49	29
58	33
61	34
63	35
64	36
70	39
	$R_1 = 408$

TABLE 12.6:

Women Not Working Outside the Home	Women Not Working Outside the Home
Score	Rank
10	2
15	5
17	6
18	7
23	10

TABLE 12.6: (continued)

Women Not Working Outside the Home	**Women Not Working Outside the Home**
24	11
25	12
28	14
30	15
37	20
41	23
42	24
47	27
48	28
52	30
55	31
56	32
65	37
69	38
71	40
	$R_2 = 412$

Do these two groups of women have significantly different views on the issue?

Since each of our samples has 20 observations, we need to calculate the standard z-score to test the hypothesis that these independent samples came from the same population. To calculate the z-score, we need to first calculate U, μ_U, and σ_U. The U-statistic for each of the samples is calculated as follows:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = (20)(20) + \frac{(20)(20+1)}{2} - 408 = 202$$
$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = (20)(20) + \frac{(20)(20+1)}{2} - 412 = 198$$

Since we use the smaller of the two U-statistics, we set $U = 198$. When calculating the other two figures, we find the following:

$$\mu_U = \frac{n_1 n_2}{2} = \frac{(20)(20)}{2} = 200$$

and

$$\sigma_U = \sqrt{\frac{n_1(n_2)(n_1+n_2+1)}{12}} = \sqrt{\frac{(20)(20)(20+20+1)}{12}} = \sqrt{\frac{(400)(41)}{12}} = 36.97$$

Thus, we calculate the z-statistic as shown below:

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{198-200}{36.97} = -0.05$$

If we set $\alpha = 0.05$, we would find that the calculated test statistic does not exceed the critical value of -1.96. Therefore, we fail to reject the null hypothesis and conclude that these two samples come from the same population.

We can use this z-score to evaluate our hypotheses just like we would with any other hypothesis test. When interpreting the results from the rank sum test, it is important to remember that we are really asking whether or not the populations have the same median and variance. In addition, we are assessing the chance that random sampling would result in medians and variances as far apart (or as close together) as observed in the test. If the z-score is large (meaning that we would have a small P-value), we can reject the idea that the difference is a coincidence. If the z-score is small, like in the example above (meaning that we would have a large P-value), we do not have any reason to conclude that the medians of the populations differ and, therefore, conclude that the samples likely came from the same population.

Determining the Correlation between Two Variables Using the Rank Correlation Test

It is possible to determine the correlation between two variables by calculating the Pearson product-moment correlation coefficient (more commonly known as the linear correlation coefficient, or r). The correlation coefficient helps us determine the strength, magnitude, and direction of the relationship between two variables with normal distributions.

We also use the *Spearman rank correlation coefficient* (also known simply as the *rank correlation coefficient*, ρ, or 'rho') to measure the strength, magnitude, and direction of the relationship between two variables. This test statistic is the nonparametric alternative to the correlation coefficient, and we use it when the data do not meet the assumptions of normality. The Spearman rank correlation coefficient, used as part of the *rank correlation test*, can also be used when one or both of the variables consist of ranks. The Spearman rank correlation coefficient is defined by the following formula:

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

where d is the difference in statistical rank of corresponding observations.

The test works by converting each of the observations to ranks, just like we learned about with the rank sum test. Therefore, if we were doing a rank correlation of scores on a final exam versus SAT scores, the lowest final exam score would get a rank of 1, the second lowest a rank of 2, and so on. Likewise, the lowest SAT score would get a rank of 1, the second lowest a rank of 2, and so on. Similar to the rank sum test, if two observations are equal, the average rank is used for both of the observations. Once the observations are converted to ranks, a correlation analysis is performed on the ranks. (Note: This analysis is not performed on the observations themselves.) The Spearman correlation coefficient is then calculated from the columns of ranks. However, because the distributions are non-normal, a regression line is rarely used, and we do not calculate a non-parametric equivalent of the regression line. It is easy to use a statistical programming package, such as SAS or SPSS, to calculate the Spearman rank correlation coefficient. However, for the purposes of this example, we will perform this test by hand as shown in the example below.

Example: The head of a math department is interested in the correlation between scores on a final math exam and math SAT scores. She took a random sample of 15 students and recorded each student's final exam score and math SAT score. Since SAT scores are designed to be normally distributed, the Spearman rank correlation test may be an especially effective tool for this comparison. Use the Spearman rank correlation test to determine the correlation coefficient. The data for this example are recorded below:

TABLE 12.7:

Math SAT Score	Final Exam Score
595	68
520	55
715	65

TABLE 12.7: (continued)

Math SAT Score	Final Exam Score
405	42
680	64
490	45
565	56
580	59
615	56
435	42
440	38
515	50
380	37
510	42
565	53

To calculate the Spearman rank correlation coefficient, we determine the ranks of each of the variables in the data set, calculate the difference for each of these ranks, and then calculate the squared difference.

TABLE 12.8:

Math SAT Score (X)	Final Exam Score (Y)	X Rank	Y Rank	d	d^2
595	68	4	1	3	9
520	55	8	7	1	1
715	65	1	2	-1	1
405	42	14	12	2	4
680	64	2	3	-1	1
490	45	11	10	1	1
565	56	6.5	5.5	1	1
580	59	5	4	1	1
615	56	3	5.5	-2.5	6.25
435	42	13	12	1	1
440	38	12	14	-2	4
515	50	9	9	0	0
380	37	15	15	0	0
510	42	10	12	-2	4
565	53	6.5	8	-1.5	2.25
Sum				0	36.50

Using the formula for the Spearman correlation coefficient, we find the following:

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{(6)(36.50)}{(15)(225-1)} = 0.9348$$

We interpret this rank correlation coefficient in the same way as we interpret the linear correlation coefficient. This coefficient states that there is a strong, positive correlation between the two variables.

Lesson Summary

We use the rank sum test (also known as the Mann-Whitney v-test) to assess whether two samples come from the same distribution. This test is sensitive to both the median and the distribution of the samples.

When performing the rank sum test, there are several different conditions that need to be met, including the population not being normally distributed, continuously distributed observations, independence of samples, the samples having greater than 5 observations, and the observations being on a numeric or ordinal scale.

When performing the rank sum test, we need to calculate a figure known as the U-statistic. This statistic takes both the median and the total distribution of both samples into account and is derived from the ranks of the observations in both samples.

When performing our hypotheses tests, we calculate the standard score, which is defined as follows:

$$z = \frac{U - \mu_U}{\sigma_U}$$

We use the Spearman rank correlation coefficient (also known simply as the rank correlation coefficient) to measure the strength, magnitude, and direction of the relationship between two variables from non-normal distributions. This coefficient is calculated as shown:

$$\rho = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

12.3 The Kruskal-Wallis Test and the Runs Test

Learning Objectives

- Evaluate a hypothesis for several populations that are not normally distributed using multiple randomly selected independent samples with the Kruskal-Wallis Test.
- Determine the randomness of a sample using the runs test to access the number of data sequences and compute a test statistic using the appropriate formula.

Introduction

In the previous sections, we learned how to conduct nonparametric tests, including the sign test, the sign rank test, the rank sum test, and the rank correlation test. These tests allowed us to test hypotheses using data that did not meet the assumptions of being normally distributed or having homogeneity with respect to variance. In addition, each of these non-parametric tests had parametric counterparts.

In this last section, we will examine another nonparametric test-the Kruskal-Wallis one-way analysis of variance (also known simply as the Kruskal-Wallis test). This test is similar to the ANOVA test, and the calculation of the test statistic is similar to that of the rank sum test. In addition, we will also explore something known as the runs test, which can be used to help decide if sequences observed within a data set are random.

Evaluating Hypotheses Using the Kruskal-Wallis Test

The *Kruskal-Wallis test* is the analog of the one-way ANOVA and is used when our data set does not meet the assumptions of normality or homogeneity of variance. However, this test has its own requirements: it is essential that the data set has identically shaped and scaled distributions for each group.

As we learned in Chapter 11, when performing the one-way ANOVA test, we establish the null hypothesis that there is no difference between the means of the populations from which our samples were selected. However, we express the null hypothesis in more general terms when using the Kruskal-Wallis test. In this test, we state that there is no difference in the distributions of scores of the populations. Another way of stating this null hypothesis is that the average of the ranks of the random samples is expected to be the same.

The test statistic for this test is the non-parametric alternative to the F-statistic. This test statistic is defined by the following formula:

$$H = \frac{12}{N(N+1)} \sum_{k=1}^{m} \frac{R_k^2}{n_k} - 3(N+1)$$

where:

$N = \sum n_k$.

n_k is number of observations in the k^{th} sample.

R_k is the sum of the ranks in the k^{th} sample.

m is the number of samples.

Like most nonparametric tests, the Kruskal-Wallis test relies on the use of ranked data to calculate a test statistic. In this test, the measurement observations from all the samples are converted to their ranks in the overall data set. The smallest observation is assigned a rank of 1, the next smallest is assigned a rank of 2, and so on. Similar to this procedure in the rank sum test, if two observations have the same value, we assign both of them the same rank.

Once the observations in all of the samples are converted to ranks, we calculate the test statistic, H, using the ranks and not the observations themselves. Similar to the other parametric and non-parametric tests, we use the test statistic to evaluate our hypothesis. For this test, the sampling distribution for H is the chi-square distribution with $m-1$ degrees of freedom, where m is the number of samples.

It is easy to use Microsoft Excel or a statistical programming package, such as SAS or SPSS, to calculate this test statistic and evaluate our hypothesis. However, for the purposes of this example, we will perform this test by hand.

Example: Suppose that a principal is interested in the differences among final exam scores from Mr. Red, Ms. White, and Mrs. Blue's algebra classes. The principal takes random samples of students from each of these classes and records their final exam scores as shown:

TABLE 12.9:

Mr. Red	Ms. White	Mrs. Blue
52	66	63
46	49	65
62	64	58
48	53	70
57	68	71
54		73

Determine if there is a difference between the final exam scores of the three teachers.

Our hypothesis for the Kruskal-Wallis test is that there is no difference in the distributions of the scores of these three populations. Our alternative hypothesis is that at least two of the three populations differ. For this example, we will set our level of significance at $\alpha = 0.05$.

To test this hypothesis, we need to calculate our test statistic. To calculate this statistic, it is necessary to assign and sum the ranks for each of the scores in the table above as follows:

TABLE 12.10:

Mr. Red	Overall Rank	Ms. White	Overall Rank	Mrs. Blue	Overall Rank
52	4	66	13	63	10
46	1	49	3	65	12
62	9	64	11	58	8
48	2	53	5	70	15
57	7	68	14	71	16
54	6			73	17
Rank Sum	29		46		78

Using this information, we can calculate our test statistic as shown:

$$H = \frac{12}{N(N+1)} \sum_{k=1}^{m} \frac{R_k^2}{n_k} - 3(N+1) = \frac{12}{(17)(18)} \left(\frac{29^2}{6} + \frac{46^2}{5} + \frac{78^2}{6} \right) - (3)(17+1) = 7.86$$

Using the chi-square distribution, we determine that with $3-1=2$ degrees of freedom, our critical value at $\alpha=0.05$ is 5.991. Since our test statistic of 7.86 exceeds the critical value, we can reject the null hypothesis that stated there is no difference in the final exam scores among students from the three different classes.

Determining the Randomness of a Sample Using the Runs Test

The *runs test* (also known as the *Wald-Wolfowitz test*) is another nonparametric test that is used to test the hypothesis that the samples taken from a population are independent of one another. We also say that the runs test checks the randomness of data when we are working with two variables. A *run* is essentially a grouping or a pattern of observations. For example, the sequence $++--++--++--$ has six runs. Three of these runs are designated by two positive signs, and three of the runs are designated by two negative signs.

We often use the runs test in studies where measurements are made according to a ranking in either time or space. In these types of scenarios, one of the questions we are trying to answer is whether or not the average value of the measurement is different at different points in the sequence. For example, suppose that we are conducting a longitudinal study on the number of referrals that different teachers give throughout the year. After several months, we notice that the number of referrals appears to increase around the time that standardized tests are given. We could formally test this observation using the runs test.

Using the laws of probability, it is possible to estimate the number of runs that one would expect by chance, given the proportion of the population in each of the categories and the sample size. Since we are dealing with proportions and probabilities between discrete variables, we consider the binomial distribution as the foundation of this test. When conducting a runs test, we establish the null hypothesis that the data samples are independent of one another and are random. On the contrary, our alternative hypothesis states that the data samples are not random and/or not independent of one another.

The runs test can be used with either nominal or categorical data. When working with nominal data, the first step in conducting the test is to compute the mean of the data and then designate each observation as being either above the mean (i.e., $+$) or below the mean (i.e., -). Next, regardless of whether or not we are working with nominal or categorical data, we compute the number of runs within the data set. As mentioned, a run is a grouping of the variables. For example, in the following sequence, we would have 5 runs. We could also say that the sequence of the data switched five times.

$$++----+++-+$$

After determining the number of runs, we also need to record each time a certain variable occurs and the total number of observations. In the example above, we have 11 observations in total, with 6 positives $(n_1=6)$ and 5 negatives $(n_2=5)$. With this information, we are able to calculate our test statistic using the following formulas:

$$z=\frac{\text{number of observed runs}-\mu}{\sigma}$$
$$\mu=\text{expected number of runs}=1+\frac{2n_1n_2}{n_1+n_2}$$
$$\sigma^2=\text{variance of the number of runs}=\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}$$

When conducting the runs test, we calculate the standard z-score and evaluate our hypotheses, just like we do with other parametric and non-parametric tests.

Example: A teacher is interested in assessing if the seating arrangement of males and females in his classroom is random. He observes the seating pattern of his students and records the following sequence:

MFMMFFFFMMMFMFMMMMFFMFFMFFFF

Is the seating arrangement random? Use $\alpha = 0.05$.

To answer this question, we first generate the null hypothesis that the seating arrangement is random and independent. Our alternative hypothesis states that the seating arrangement is not random or independent. With $\alpha = 0.05$, we set our critical values at 1.96 standard scores above and below the mean.

To calculate the test statistic, we first record the number of runs and the number of each type of observation as shown:

$$R = 14 \quad M : n_1 = 13 \quad F : n_2 = 15$$

With these data, we can easily compute the test statistic as follows:

$$\begin{aligned}
\mu &= \text{expected number of runs} = 1 + \frac{(2)(13)(15)}{13+15} = 1 + \frac{390}{28} = 14.9 \\
\sigma^2 &= \text{variance of the number of runs} = \frac{(2)(13)(15)[(2)(13)(15) - 13 - 15]}{(13+15)^2(13+15-1)} = \frac{(390)(362)}{(784)(27)} = 6.67 \\
\sigma &= 2.58 \\
z &= \frac{\text{number of observed runs} - \mu}{\sigma} = \frac{14 - 14.9}{2.58} = -0.35
\end{aligned}$$

Since the calculated test statistic is not less than $z = -1.96$, our critical value, we fail to reject the null hypothesis and conclude that the seating arrangement of males and females is random.

Lesson Summary

The Kruskal-Wallis test is used when we are assessing the one-way variance of a specific variable in non-normal distributions.

The test statistic for the Kruskal-Wallis test is the non-parametric alternative to the F-statistic. This test statistic is defined by the following formula:

$$H = \frac{12}{N(N+1)} \sum_{k=1}^{m} \frac{R_k^2}{n_k} - 3(N+1)$$

The runs test (also known as the Wald-Wolfowitz test) is another non-parametric test that is used to test the hypothesis that the samples taken from a population are independent of one another. We use the z-statistic to evaluate this hypothesis.

On the Web

http://tinyurl.com/334e5to Good explanations of and examples of different nonparametric tests.

http://tinyurl.com/33s4h3o Allows you to enter data and then performs the Wilcoxon sign rank test.

http://tinyurl.com/33s4h3o Allows you to enter data and performs the Mann Whitney Test.

Keywords

Kruskal-Wallis test

Mann-Whitney *v*-test

Non-parametric tests

Rank correlation coefficient

Rank correlation test

Rank sum test

Run

Runs test

Sign rank test

Sign test

Spearman rank correlation coefficient

U-distribution

U-statistic

Wald-Wolfowitz test

Wilcoxon sign rank test