COMPARING TWO UNRELATEDSAMPLES:

THE MANN–WHITNEY U-TEST &

KOLMOGOROV–SMIRNOV TWO-SAMPLE TEST

OBJECTIVE

In this lecture, you will learn the following items:

- How to perform the Mann−Whitney *U*-test.
- How to construct a median confidence interval based on the difference between two independent samples.
- How to perform the Kolmogorov−Smirnov two-sample test.

INTRODUCTION

Suppose a teacher wants to know if his first-period's early class time has been reducing student performance. To test his idea, he compares the final exam scores of students in his first-period class with those in his fourth-period class.

In this example, each score from one class period is independent, or unrelated, to the other class period.

The Mann−Whitney U-test and the Kolmogorov−Smirnov two-sample test
are nonparametric statistical procedures for comparing two samples that are independent, or not related. The parametric equivalent to these tests is the t-test for independent samples.

In this lecture, we will describe how to perform and interpret a **Mann–Whitney U-test** and a **Kolmogorov –Smirnov two-sample test**. We will demonstrate the small samples for each test.

COMPUTING THE MANN–WHITNEY U-TEST STATISTIC

The Mann−Whitney U-test is used to compare two unrelated, or independent, samples.

The two samples are combined and rank ordered together. The strategy is to determine if the values from the two samples are randomly mixed in the rank ordering or if they are clustered at opposite ends when combined.

A random rank ordered would mean that the two samples are not different, while a cluster of one sample's values would indicate a difference between them.

In Figure 1, two sample comparisons illustrate this concept.

The scores in Comparison 1 are rank ordered in clusters at opposite ends. This suggests that treatment X might be higher than treatment O.	COMPARISON 1 X X X O X X X X O O O O 1 2 3 4 5 6 7 8 9 10 11 12
The scores in Comparison 2 are spread along the entire distribution. This suggests that there is no clear difference between treatments.	COMPARISON 2 X O O X X O X O X O X X 1 2 3 4 5 6 7 8 9 10 11 12

FIGURE 1

Use Formula 1 to determine a Mann−Whitney U-test statistic for each of the two samples. The smaller of the two U statistics is the obtained value:

$$U_i = n_1 n_2 + \frac{n_i(n_i + 1)}{2} - \sum R_i \qquad (1)$$

where U_i is the test statistic for the sample of interest, n_i is the number of values from the sample of interest, n_1 is the number of values from the first sample, n_2 is the number of values from the second sample, and ΣR_i is the sum of the ranks from the sample of interest.

After the U statistic is computed, it must be examined for significance. We may use a table of critical values (Table B.4).

However, if the numbers of values in each sample, n_i, exceeds those available from the table, then a large sample approximation may be performed.

At this point, the analysis is limited to identifying the presence or absence of a significant difference between the groups and does not describe the strength of the treatment. We can consider the effect size (ES) to determine the degree of association between the groups. We use Formula 2 to calculate the ES:

$$ES = \frac{|z|}{\sqrt{n}} \quad (2)$$

where $|z|$ is the absolute value of the z-score and n is the total number of observations.

The ES ranges from 0 to 1. Cohen (1988) defined the conventions for ES as small = 0.10, medium = 0.30, and large = 0.50. (Correlation coefficient and ES are both measures of association).

Example

Mann–Whitney U-Test

The following data were collected from a study comparing two methods being used to teach reading recovery in the 4th grade.

Method 1 was a pull-out program in which the children were taken out of the classroom for 30 min a day, 4 days a week.

Method 2 was a small group program in which children were taught in groups of four or five for 45 min a day in the classroom, 4 days a week. The students were tested using a reading comprehension test after 4 weeks of the program.

The test results are shown in Table 1.

TABLE 1

Method 1	Method 2
48	14
40	18
39	20
50	10
41	12
38	102
53	17

1 State the Null and Research Hypotheses

The null hypothesis states that there is no tendency of the ranks of one method to be systematically higher or lower than the other. The hypothesis is stated in terms of comparison of distributions, not means. The research hypothesis states that the ranks of one method are systematically higher or lower than the other. Our research hypothesis is a two-tailed, non-directional hypothesis because it indicates a difference, but in no particular direction.

The null hypothesis is

H_O: There is no tendency for ranks of one method to be significantly higher (or lower) than the other.

The research hypothesis is

H_A: The ranks of one method are systematically higher (or lower) than the other.

2. Set the Level of Risk (or the Level of Significance) Associated with the Null Hypothesis

The level of risk, also called an alpha (), is frequently set at 0.05. We will use $\alpha = 0.05$ in our example. In other words, there is a 95% chance that any observed statistical difference will be real and not due to chance.

3. Choose the Appropriate Test Statistic

The data are obtained from two independent, or unrelated, samples of 4th-grade children being taught reading. Both the small sample sizes and an existing outlier in the second sample violate our assumptions of normality. Since we are comparing two unrelated, or independent, samples, we will use the Mann−Whitney U-test.

4. Compute the Test Statistic

First, combine and rank both data samples together (Table 2).

Next, compute the sum of ranks for each method. Method 1 is ΣR_1 and method 2 is ΣR_2. Using Table 2,

$$\sum R_1 = 7+8+9+10+11+12+13$$

$$\sum R_1 = 70$$

and

$$\sum R_2 = 1+2+3+4+5+6+14$$

$$\sum R_2 = 35$$

TABLE 2

Ordered scores

Rank	Score	Sample
1	10	Method 2
2	12	Method 2
3	14	Method 2
4	17	Method 2
5	18	Method 2
6	20	Method 2
7	38	Method 1
8	39	Method 1
9	40	Method 1
10	41	Method 1
11	48	Method 1
12	50	Method 1
13	53	Method 1
14	102	Method 2

Now, compute the U-value for each sample. For sample 1,

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum R_1 = 7(7) + \frac{7(7+1)}{2} - 70 = 49 + 28 - 70$$

$$U_1 = 7$$

and for sample 2,

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum R_2 = 7(7) + \frac{7(7+1)}{2} - 35 = 49 + 28 - 35$$

$$U_2 = 42$$

The Mann–Whitney U-test statistic is the smaller of U_1 and U_2. Therefore, $U = 7$.

5. Determine the Value Needed for Rejection of the Null Hypothesis Using the Appropriate Table of Critical Values for the Particular Statistic

Since the sample sizes are small ($n < 20$), we use Table B.4 in Appendix B, which lists the critical values for the Mann−Whitney U.

The critical values are found on the table at the point for $n_1 = 7$ and $n_2 = 7$.

We set = 0.05. The critical value for the Mann−Whitney U is 8.

A calculated value that is less than or equal to 8 will lead us to reject our null hypothesis.

6. Compare the Obtained Value with the Critical Value

The critical value for rejecting the null hypothesis is 8 and the obtained value is U = 7. If the critical value equals or exceeds the obtained value, we must reject the null hypothesis.

If instead, the critical value is less than the obtained value, we must not reject the null hypothesis.

Since the critical value exceeds the obtained value, we must reject the null hypothesis.

7. Interpret the Results

We rejected the null hypothesis, suggesting that a real difference exists between the two methods. In addition, since the sum of the ranks for method 1 ($\sum R_1$) was larger than method 2 ($\sum R_2$), we see that method 1 had significantly higher scores.

8. Reporting the Results

The reporting of results for the Mann−Whitney U-test should include such information as the sample sizes for each group, the U statistic, the p-value's relation to , and the sums of ranks for each group.

For this example, two methods were used to provide students with reading instruction.

Method 1 involved a pull-out program and method 2 involved a small group program. Using the ranked reading comprehension test scores, the results indicated a significant difference between the two methods ($U = 7$, $n_1 = 7$, $n_2 = 7$, $p < 0.05$).

The sum of ranks for method 1 ($\sum R_1 = 70$) was larger than the sum of ranks for method 2 ($\sum R_2 = 35$). Therefore, we can state that the data support the pull-out program as a more effective reading program for teaching comprehension to 4th-grade children at this school.

Confidence Interval for the Difference between Two Location Parameters

The American Psychological Association (2001) has suggested that researchers report the *confidence interval* for research data. A confidence interval is an inference to a population in terms of an estimation of sampling error. More specifically, it provides a range of values that fall within the population with a level of confidence of $100(1 - \alpha)\%$.

A median confidence interval can be constructed based on the difference between two independent samples. It consists of possible values of differences for which we do not reject the null hypothesis at a defined significance level of α.

The test depends on the following assumptions:

1. Data consist of two independent random samples: $X_1, X_2, \ldots, X_n$ from one population and $Y_1, Y_2, \ldots, Y_n$ from the second population.
2. The distribution functions of the two populations are identical except for possible location parameters.

To perform the analysis, set up a table that identifies all possible differences for each possible sample pair such that $D_{ij} = X_i - Y_j$ for (X_i, Y_j). Placing the values for X from smallest to largest across the top and the values for Y from smallest to largest down the side will eliminate the need to order the values of D_{ij} later.

The sample procedure to be presented later is based on the data from Table 2 near the beginning of this chapter.

The values from Table 2 are arranged in Table 3 so that the method 1 (X) scores are placed in order across the top and the method 2 (Y) scores are placed in order down the side. Then, the n_1n_2 differences are calculated by subtracting each Y value from each X value.

The differences are shown in Table 3. Notice that the values of Dij are ordered in the table from highest to lowest starting at the top right and ending at the bottom left.

TABLE 3

	X_i						
Y_j	38	39	40	41	48	50	53
10	28	29	30	31	38	40	43
12	26	27	28	29	36	38	41
14	24	25	26	27	34	36	39
17	21	22	23	24	31	33	36
18	20	21	22	23	30	32	35
20	18	19	20	21	28	30	33
102	−64	−63	−62	−61	−54	−52	−49

We use Table B.4 to find the lower limit of the confidence interval, L, and the upper limit U.

For a two-tailed test, L is the $w_{\alpha/2}$th smallest difference and U is the $w_{\alpha/2}$th largest difference that correspond to $\alpha/2$ for n_1 and n_2 for a confidence interval of $(1 - \alpha)$.

For our example, $n_1 = 7$ and $n_2 = 7$. For $\alpha/2 = 0.05/2 = 0.025$, Table B.4 returns $w_{\alpha/2} = 9$.

This means that the ninth values from the top and bottom mark the limits of the 95% confidence interval on both ends. Therefore, L = 19 and U = 36.

Based on these results, we are 95% certain that the difference in population median is between 18 and 36.

COMPUTING THE KOLMOGOROV–SMIRNOV TWO-SAMPLE TEST STATISTIC

In Lecture 2, we used the Kolmogorov–Smirnov one-sample test to compare a sample with the normal distribution. We can use the Kolmogorov–Smirnov two sample test to analyze two different data samples for independence. Our data must meet two assumptions.

1. Observations $X_1, \ldots, X_m$ are a random sample from a continuous population 1, where the X-values are mutually independent and identically distributed. Likewise, observations $Y_1, \ldots, Y_n$ are a random sample from a continuous population 2, where the Y-values are mutually independent and identically distributed.
2. The two samples are independent.

We begin by placing the data in a form that will permit us to compute the two-sided Kolmogorov–Smirnov test statistic Z. The first step in this procedure is to find the empirical distribution functions $F_m(t)$ and $G_n(t)$ for the samples of X and Y, respectively. Combine and rank order both sets of values. For every real number t, let

$$F_m(t) = \frac{\text{number of observed } X\text{'s} \leq t}{m}$$

and

$$G_n(t) = \frac{\text{number of observed } Y\text{'s} \leq t}{n}$$

where m is the sample size of X and n the sample size of Y.

Next, use Formula 3 to find each absolute value divergence D between the empirical distributions functions:

$$D = \left| F_m(t) - G_n(t) \right| \qquad (3)$$

Use the largest divergence D_{max} with Formula 4 to calculate the Kolmogorov–Smirnov test statistic Z:

$$Z = D_{\max} \sqrt{\frac{mn}{m+n}} \qquad (4)$$

Then, use the Kolmogorov–Smirnov test statistic, Z, and the Smirnov (1948) formula (Formula 5, Formula 6, Formula 7, Formula 8, Formula 9, and Formula 10) to find the two-tailed probability estimate p. This is the same procedure shown in Lecture 2 when we performed the Kolmogorov–Smirnov one-sample test:

$$\text{if } 0 \le Z < 0.27, \text{ then } p = 1 \qquad (5)$$

$$\text{if } 0.27 \le Z < 1, \text{ then } p = 1 - \frac{2.506628}{Z}(Q + Q^9 + Q^{25}) \qquad (6)$$

where

$$Q = e^{-1.233701Z^{-2}} \qquad (7)$$

$$\text{if } 1 \le Z < 3.1, \text{ then } p = 2(Q - Q^4 + Q^9 - Q^{16}) \qquad (8)$$

where

$$Q = e^{-2Z^2} \qquad (9)$$

$$\text{if } Z \ge 3.1, \text{ then } p = 0 \qquad (10)$$

Once we have our p-value, we can compare it against our level of risk α to determine if the two samples are significantly different.

Example

Kolmogorov–Smirnov Two-Sample Test

We have two methods.

Method 1 was a program in which children were taken out of the classroom for 30 min a day, Monday through Thursday each week.

Method 2 was a small group program in which the children were taught in groups of no more than five for 45 min a day in the classroom. These small classes were taught Monday through Thursday, also. The students were tested using a reading comprehension test after 4 weeks of instruction.

Table 2 recalls the data from the study involving reading recovery in the 4th grade.

TABLE 2.

Method 1	Method 2
48	14
40	18
39	20
50	10
41	12
38	102
53	17

1. State the Null and Alternate Hypotheses

Let $X_1, \ldots, X_m$, and $Y_1, \ldots, Y_n$ be independent random samples. The null hypothesis indicates that there is no difference between the reading groups X and Y. Our research hypothesis is a two-tailed, non-directional hypothesis because it indicates a difference, but in no particular direction.

The null hypothesis is

$$H_O: [F(t) = G(t), \text{ for every } t]$$

The research hypothesis is

$$H_A: [F(t) \neq G(t) \text{ for at least one value of } t]$$

2. Set the Level of Risk (or the Level of Significance) Associated with the Null Hypothesis

We will use $\alpha = 0.05$ in our example. In other words, there is a 95% chance that any observed statistical difference will be real and not due to chance.

3. Choose the Appropriate Test Statistic

We are seeking to compare two random samples, X and Y. Each sample is mutually independent and identically distributed. The X's and Y's are mutually independent. The Kolmogorov–Smirnov two-sample test will provide this comparison.

4. Compute the Test Statistic

Begin by computing the empirical distribution functions for the X and Y samples:

$$F_m(t) = \frac{\text{number of observed } X\text{'s} \leq t}{m}$$

and

$$G_n(t) = \frac{\text{number of observed } Y\text{'s} \leq t}{n}$$

where $m = 7$ and $n = 7$.

We use the data in Table 2 and Formula 3 to find each divergence and
generate Table 3.

Next, we find the largest divergence D_{max}. Table 3 shows that $D_{max} = 6/7 = 0.86$.

Now, we use Formula 4 to calculate the Kolmogorov–Smirnov test statistic Z:

$$Z = D_{\max}\sqrt{\frac{mn}{m+n}} = (0.86)\cdot\sqrt{\frac{(7)(7)}{7+7}} = (0.86)\cdot\sqrt{3.5} = (0.86)(1.87)$$

$$Z = 1.604$$

TABLE 3

	Z_i	$F_7(Z_i)$	$G_7(Z_i)$	$\lvert F_7(Z_i) - G_7(Z_i)\rvert$
1	10	0/7	1/7	1/7
2	12	0/7	2/7	2/7
3	14	0/7	3/7	3/7
4	17	0/7	4/7	4/7
5	18	0/7	5/7	5/7
6	20	0/7	6/7	6/7
7	38	1/7	6/7	5/7
8	39	2/7	6/7	4/7
9	40	3/7	6/7	3/7
10	41	4/7	6/7	2/7
11	48	5/7	6/7	1/7
12	50	6/7	6/7	0/7
13	53	7/7	6/7	1/7
14	102	7/7	7/7	0/7

5. Determine the p-Value Associated with the Test Statistic

Now, we find the p-value using Formula 8 since they satisfy the condition that $1 \leq Z < 3.1$. We first need Q using Formula 9:

$$Q = e^{-2Z^2} = e^{(-2)(1.604)^2} = e^{-5.146}$$
$$Q = 0.0058$$

Now, we can use Formula 8:

$$p = 2(Q - Q^4 + Q^9 - Q^{16}) = (2)(0.0058 - 0.0058^4 + 0.0058^9 - 0.0058^{16})$$
$$= (2)(0.0058)$$
$$p = 0.012$$

6. Compare the Obtained Value with the Critical Value Needed for Rejection of the Null Hypothesis

The two-tailed probability, p = 0.012, was computed and is now compared with the level of risk specified earlier, α = 0.05. If is greater than the p-value, we must reject the null hypothesis. If α is less than the p-value, we must not reject the null hypothesis. Since is greater than the p-value (0.05 > 0.012), we reject the null hypothesis.

7. Interpret the Results

We rejected the null hypothesis, suggesting that the two methods for teaching reading recovery have significantly different effects on the learning of students. In studying the results, it appears that method 1 was more effective than method 2, in general.

8. Reporting the Results

When reporting the results from the Kolmogorov–Smirnov two-sample test, include such information as the sample sizes for each group, the D statistic, and the p-value's relation to .

For this example, two methods were used to provide students with reading instruction. Method 1 involved a pull-out program and method 2 involved a small group program. Both methods include seven participants. The results from the Kolmogorov–Smirnov two-sample test ($D = 0.857$, $p < 0.05$) indicate a significant difference between the two methods.

Therefore, we can state that the data support the pull-out program as a more effective reading program for teaching comprehension to 4th-grade children at this school.

SUMMARY

Two samples that are not related may be compared using a non-parametric procedure.

Examples include the Mann–Whitney U-test (or the Wilcoxon rank sum test) and the Kolmogorov–Smirnov two-sample test. The parametric equivalent to these tests is known as the t-test for independent samples.

In this lecture, we described how to perform and interpret the Mann–Whitney U-test and the Kolmogorov–Smirnov two-sample test. We demonstrated a small samples for each test.