24

Non-Parametric Statistics for Ordinal Differences

The Rationale of Testing Ordinal Differences
Wilcoxin Rank-Sum Test
Mann-Whitney U Test
Wilcoxin Matched-Pairs Test
Kruskal-Wallis H Test

Chapters 16-21 covered major **parametric procedures** for testing hypotheses of difference (z, t, F). Here we will look at several **non-parametric procedures** used to test hypotheses of difference when the data is *ordinal* -- rankings. The most common application of these tests is with small group testing, where interval/ratio data is converted to ranks. These non-parametric tests are not constrained by the same mathematical restrictions as parametric tests, and so give better results for small n..

These procedures include the **Wilcoxin Rank-Sum** test, the **Mann-Whitney U** test, the **Wilcoxin Matched-Pairs** test, and the **Kruskal-Wallis H** test.

Dr. Gail Linam studied the Bible reading comprehension of children, grades 4-6, across three translations of Scripture: the King James (KJV), the New International (NIV), and New Century (NCV).[1] The children's reading comprehension was measured by two different instruments on a story from the Old and New Testaments. The first was the retelling method (OTR, NTR), and the second was the Cloze method (OTC, NTC). She also averaged the two stories into a single Bible comprehension score (BIBR, BIBC).

Ninety-two (92) children were tested. Scores were ranked without regard to group membership of the child, and then sums of ranks were computed for each group (KJV, NIV, NCV). The results are shown in the following computer printout[2]:

```
KRUSKAL-WALLIS ONE-WAY ANALYSIS OF VARIANCE FOR 92 CASES[2]
  DEPENDENT VARIABLE IS   OTR
  GROUPING VARIABLE IS    VER
```

[1]Gail Linam, "A Study of the Reading Comprehension of Older Children Using Selected Bible Translations," (Ed.D. diss., Southwestern Baptist Theological Seminary, 1993)

[2]Ibid., 204

```
GROUP          COUNT          RANK SUM
1.000          30              887.500
2.000          31             1603.500
3.000          31             1787.000

KRUSKAL-WALLIS TEST STATISTIC =            18.649
PROBABILITY IS    0.000 ASSUMING CHI-SQUARE DISTRIBUTION WITH 2 DF
```

The Kruskal-Wallis shows a significant difference ("p=0.000") in Old Testment Retelling scores (OTR) across the three translations (VER). Notice that the sum of ranks for Group 1 (KJV) is much smaller than Groups 2 and 3. This reveals much lower reading comprehension among children in grades 4-6 for the King James. She found the same results with the NTR, OTC, NTC, BIBR, and BIBC tests. In every case, children understood much less of the King James English than either the New International or the New Century versions.[3]

Each of the ordinal tests have parametric counterparts, with which you are already familiar.

When Groups are Small	***When Groups are Large***
Wilcoxin Rank-Sum Mann-Whitney U	Independent-samples t-Test
Both procedures test two independent samples for significant difference	
Wilcoxin Matched-Pairs	Correlated-samples t-Test
Both procedures test two matched samples for significant difference	
Kruskal-Wallis H	One-way ANOVA
Both procedures test three or more independent samples for significant difference	

The Rationale of Testing Ordinal Differences

Since ranked data is non-parametric, ordinal procedures are not limited by parametric restrictions which reduce power with small n. These ordinal procedures provide greater power when testing small samples.

All four of these ordinal tests follow the same rationale of testing. Rank subjects from lowest score (1) to highest score (n) without regard to group. Then separate rankings by their group. Sum the ranks within each group. These sums of ranks (**ΣR**) are used by the tests to determine whether groups are significantly different. Since low scores produce low ranks, groups that score systematically lower produce a smaller

[3] Ibid., 205-206

sum of ranks; groups that score systematically higher produce a larger sum of ranks. If the difference between the **ΣR** terms is large enough, it will be declared significant.

Wilcoxin Rank-Sum Test (W_s)

The Wilcoxin Rank-Sum test is one of the most common and best known distribution-free tests. The W_s computes **ΣR** for two groups of scores, then uses the **ΣR** of the **smaller group** to test the null hypothesis. If the groups are equal in size, then use the smaller of the two **ΣR**'s. Compare this value with the critical value in the Wilcoxin Table to test the null hypothesis.

Computing the Wilcoxin W

Researchers tabulated the number of stressful events reported by two groups of patients in a local hospital. The first group were cardiac patients and the second were a control group of orthopedic patients.[5] Here's the data:

	Cardiac Patients						Orthopedic Patients				
Raw Scores	12	8	7	9	5	0	1	2	2	3	6
Ranks	***11***	***9***	***8***	***10***	***6***	***1***	***2***	***3.5***	***3.5***	***5***	***7***
	ΣR = **45**						ΣR = **21**				

The lowest score (0) receives the rank of 1, and the highest count (12) receives the rank of 11. Two scores have the same count of 2. They are assigned the tied ranks of 3.5, 3.5 in place of "3, 4" (there is no "4" rank). These rankings are then summed by group, yielding sums of **45** and **21**. Since Group 2 (n=5) is **smaller** than Group 1 (n=6), use **ΣR** of Group 2: **21**.

The Wilcoxin W Table

The Wilcoxin Rank-Sum table is located on page **A3-6**. The column labelled $\mathbf{N_1}$ refers to the ***size (n)*** of the smaller group. The column labelled $\mathbf{N_2}$ refers to the **n** of the second group.

Locate the segment headed $\mathbf{N_1 = 5}$. Move down the left side of the segment to row $\mathbf{N_2=6}$. The two-tailed 0.05 value, move across to the column headed by "**2.5**" (2.5% = 0.025 = 0.05/2). The critical value is ***18***.

In order for the W_s statistic to be declared significant, it must be ***less than the critical value*** [no, this isn't a typo!].

Our value of W_s=21 is greater than W_{cv}=18, so we retain the null hypothesis. Interpreting the statistical result in light of the study, we declare that there is no difference in the number of reported stressful events between cardiac patients and orthopedic patients.

The Mann-Whitney U Test

Another popular non-parametric equivalent to the independent samples t-test is the Mann-Whitney U. It is equivalent to the Wilcoxin Rank-Sum Test, but is included here because of its popularity in social science research. The Mann-Whitney U computes two ***U values*** with the following formulas:

[5]David C. Howell, *Statistical Methods for Psychology*, (Boston: Duxbury Press, 1982), 500

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

where n_1 = number of observations in group 1, n_2 = number of observations in 2, ΣR_1 = sum of ranks assigned to group 1, and ΣR_2 = sum of ranks assigned to 2.

Computing the Mann-Whitney U

The ***smaller*** of U_1 and U_2 is the **U Test statistic**, and is compared to U_{cv} to determine whether to reject the null hypothesis. ***The computed U statistic must be less than the critical value in order to reject the null hypothesis.*** Using the **ΣR** values of **45** and **21** from the cardiac example, we'll compute U_1 and U_2 .

$$U_1 = 6 \cdot 5 + \frac{6 \cdot 7}{2} - 45 = 6 \qquad U_2 = 6 \cdot 5 + \frac{5 \cdot 6}{2} - 21 = 24$$

The U_1 term (6) is smaller than the U_2 term (24), so the ***U statistic is 6***.

The Mann-Whitney U Table

The critical value for U is found in the U-distribution table on page A3-9. N_1 is the size of the smaller group, in our case, **5**. N_2 is the size of the larger group, in our case, **6**. Since we are conducting a 2-tail test at α=0.05, we'll use the upper table on A3-9. Locate the column labelled **5** and move down to the row labelled **6**. The critical value is **3**.

The computed value ***must be less*** than the critical value[6] in order to reject the null hypothesis. Since in this case it is not less, we retain the null. This is the same result we obtained with the Wilcoxin Rank-Sum test.

Wilcoxin Matched-Pairs Test (T)

The Wilcoxin Matched-Pairs test statistic (T) is computed in the same straightforward manner as W_s except that it is used with ***matched scores.*** "After" scores are subtracted from "Before" scores to yield a "difference."

Computing the Wilcoxin T

These differences are ranked from low to high without regard to sign (+,-). Then the sign of the difference (+,-) is applied to the ranks.

All +ranks are summed to yield T+ and all -ranks are summed to yield T-. The **smaller** of the two sums (T+, T-) is taken, regardless of sign, as the statistic **T**.

If computed T is *smaller than the critical value* (one-tail) or *outside the range of critical values* (two-tail) found in the Wilcoxin T Table (A3-7), then reject the null hypothesis.

Does running reduce blood pressure? The "Before" scores show patients' blood pressure before the running program, and the "After" scores show blood pressure after six weeks of running.[6]

[6]Howell, p. 505

Before		After		Change	Rank	Signed Rank
130	-	120	=	10	5	+5
170		163		7	4	+4
125		120		5	2	+2
160		135		25	7	+7
143		130		13	6	+6
130		136		-6	3	-3
145		144		1	1	+1
160		120		40	8	+8

T+ =Σ(+ranks) = +33 T- = Σ(-ranks) = -3

The Rank column shows the **Change** values ranked low to high without regard to sign (score 1 = rank 1; score 40 = rank 8).

The Signed Rank column applies the sign (-,+) of Change to Rank. Add together all positive ranks for T+ and all negative ranks for T-.

The T statistic equals the ***smaller of the two T values***. Since T- (-3) is smaller than and T+ (33), ***T = 3***.

The Wilcoxin T Table

The critical value is found in the Wilcoxin T table (A3-7). We are testing at ***α=0.05*** with ***N=8*** pairs of scores. Read down the left side of the table to **8**. Looking under the 0.05 column, we read a critical value of **5** . Since the computed T =3 is smaller than the critical values, we **reject the null** hypothesis. ***The running program did significantly lower the subjects' blood pressure levels.***

Kruskal-Wallis H Test

The Kruskal-Wallis H Test is a generalization of the Wilcoxin Rank-Sum test to the case where we have three or more independent groups. As such it is the distribution-free counterpart to the one-way analysis of variance test. Using the following equation, we test whether the **ΣR**'s for all groups are equal:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \left[\frac{\Sigma R_i^2}{n_i} \right] - 3(N+1)$$

where k = the number of groups, n_i = the number of observations in group i, ΣR_i = the sum of ranks in group i, and N = the total sample size.

Computing the Kruskal-Wallis H

A researcher designs an experiment to measure the effect of a depressant and a stimulant on the rate of performance in solving simple arithmetic problems. He also includes a control condition of a placebo. The scores below equal the number of problems solved in one hour.[8]

[8] Ibid., p. 507

Depressant		Stimulant		Placebo	
Score	**Rank**	**Score**	**Rank**	**Score**	**Rank**
55	9	73	15	61	11
23	2	82	18	54	8
40	3	51	7	80	17
17	1	63	12	47	5
50	6	74	16		
60	10	85	19		
44	4	66	13		
		69	14		
ΣR_i:	$\Sigma R_1 = 35$		$\Sigma R_2 = 114$		$\Sigma R_3 = 41$

Substituting the **ΣR** values into the H Test formula, we have the following:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \left[\frac{\Sigma R_i^2}{n_i} \right] - 3(N+1)$$

$$H = \frac{12}{19 \cdot 20} \cdot \left[\frac{35^2}{7} + \frac{114^2}{8} + \frac{41^2}{4} \right] - 3(20)$$

$$H = \frac{12}{380} \cdot \left[2219.75 \right] - 60 = 70.10 - 60 = 10.10$$

Using the Chi-Square Table with Kruskal-Wallis H

The **critical value** for the **Kruskal-Wallis H test** comes from the Chi-square Table (A3-3) with ***k-1 degrees of freedom***, where **k** is the number of groups being tested.

Since the critical value is taken from the Chi-Square Table, the **computed value must be *larger than* the critical value in order to reject the null hypothesis.** The critical value for this example is 5.99 (0.05, df=2).

Since **H = 10.10** is larger than the critical value of **5.99**, we ***reject the null hypothesis***. *The three drugs lead to different rates of performance. The performance of the "stimulant" group suggests that this group did better than either the depressant or control groups.*

Summary

In this chapter we have investigated the more popular and powerful of the distribution-free ordinal tests of difference. We analyzed the Wilcoxin Rank-Sum test, the Mann-Whitney U test, the Wilcoxin Matched-Pairs Signed-Ranks test, and the Kruskal-Wallis H test.

The value of these tests is their ability to handle smaller groups of subjects than their comparable parametric counterparts. This is particularly helpful in the kinds of studies designed in the context of Christian education, administration, counseling and social work.

[8]Linam, pp. 205-206

Example

Here are the remaining Kruskal-Wallis H test results from Dr. Linam's study:

```
KRUSKAL-WALLIS ONE-WAY ANALYSIS OF VARIANCE FOR 92 CASES[8]
  DEPENDENT VARIABLE IS   NTR (New Testament Retelling Test)
  GROUPING VARIABLE IS    VER

        GROUP       COUNT       RANK SUM
        1.000       30           888.000    (KJV)
        2.000       31          1679.500    (NIV)
        3.000       31          1710.000    (NCV)
KRUSKAL-WALLIS TEST STATISTIC =          17.884
PROBABILITY IS   0.000 ASSUMING CHI-SQUARE DISTRIBUTION WITH 2 DF
..................
KRUSKAL-WALLIS ONE-WAY ANALYSIS OF VARIANCE FOR 92 CASES[8]
  DEPENDENT VARIABLE IS   OTC (Old Testament Cloze Test)
  GROUPING VARIABLE IS    VER

        GROUP       COUNT       RANK SUM
        1.000       30           808.500
        2.000       31          1546.500
        3.000       31          1923.000
KRUSKAL-WALLIS TEST STATISTIC =          27.115
PROBABILITY IS   0.000 ASSUMING CHI-SQUARE DISTRIBUTION WITH 2 DF
...........
KRUSKAL-WALLIS ONE-WAY ANALYSIS OF VARIANCE FOR 92 CASES[8]
  DEPENDENT VARIABLE IS   NTC (New Testament Cloze Test)
  GROUPING VARIABLE IS    VER

        GROUP       COUNT       RANK SUM
        1.000       30           705.000
        2.000       31          1742.500
        3.000       31          1830.500
KRUSKAL-WALLIS TEST STATISTIC =          33.342
PROBABILITY IS   0.000 ASSUMING CHI-SQUARE DISTRIBUTION WITH 2 DF
...........

KRUSKAL-WALLIS ONE-WAY ANALYSIS OF VARIANCE FOR 92 CASES[8]
  DEPENDENT VARIABLE IS   BIBR (Average Bible Retelling Test)
  GROUPING VARIABLE IS    VER

        GROUP       COUNT       RANK SUM
        1.000       30           851.500
        2.000       31          1654.000
        3.000       31          1772.500
KRUSKAL-WALLIS TEST STATISTIC =          20.822
PROBABILITY IS   0.000 ASSUMING CHI-SQUARE DISTRIBUTION WITH 2 DF
............
KRUSKAL-WALLIS ONE-WAY ANALYSIS OF VARIANCE FOR 92 CASES[8]
  DEPENDENT VARIABLE IS   BIBC (Average Bible Cloze Test)
  GROUPING VARIABLE IS    VER

        GROUP       COUNT       RANK SUM
        1.000       30           711.000
        2.000       31          1664.500
        3.000       31          1902.500
KRUSKAL-WALLIS TEST STATISTIC =          33.765
PROBABILITY IS   0.000 ASSUMING CHI-SQUARE DISTRIBUTION WITH 2 DF
```

In every case, comprehension of the KJV was significantly lower than NIV or NCV

Vocabulary

Kruskal-Wallis H test	ordinal alternative to one-way ANOVA
Mann-Whitney U test	ordinal alternative to t-test for independent samples
sum of ranks	key concept in ordinal statistics (ΣR) - used to differentiate groups
Wilcoxin T test	ordinal alternative to matched samples t-test
Wilcoxin W_s test	ordinal alternative to the t-test for independent samples

Study Questions

1. Describe the rationale for using non-parametric tests.

2. Describe the appropriate way to handle tied ranks in the procedures discussed in this chapter.

3. Explain in your own words when to use the Kruskal-Wallis H, the Wilcoxin T, the Mann-Whitney U and the Wilcoxin Ws tests.

Sample Test Questions

1. The common term in the Wilcoxin W, the Mann-Whitney U, and the Kruskal-Wallis H is
   A. R^2
   B. X
   C. r^2
   D. ΣR

2. The appropriate test to use when testing the difference between 10 husbands and their wives in marital satisfaction scores is
   A. correlated samples t-test
   B. independent t-test
   C. Mann-Whitney U test
   D. Wilcoxin T test

3. The appropriate test to use when testing the difference among 10 third graders, 8 fourth graders and 9 fifth graders in problem-solving ability scores is
   A. one-way ANOVA
   B. multiple independent t-test
   C. Mann-Whitney U test
   D. Kruskal-Wallis H test

4. The non-parametric test equivalent to the Mann-Whitney U is
   A. Wilcoxin Ws
   B. Wilcoxin T
   C. Independent samples t-test
   D. Kruskal-Wallis H