DATA ANALYTICS

Class # 12

Non-parametric tests II

Dr. Sreeja S R

Assistant Professor

Indian Institute of Information Technology

IIIT Sri City

QUOTE OF THE DAY..

Try not to become a person of success, but rather try to become a person of value.

ALBERT EINSTEIN, Theoretical physicist

EXAMPLE 2 FOR THE MANN WHITNEY U TEST

- A new approach to prenatal care is proposed for pregnant women living in a rural community. The new program involves in-home visits during the course of pregnancy in addition to the usual or regularly scheduled visits. A pilot randomized trial with 15 pregnant women is designed to evaluate whether women who participate in the program deliver healthier babies than women receiving usual care. The outcome is the APGAR score measured 5 minutes after birth. APGAR scores range from 0 to 10 with scores of 7 or higher considered normal (healthy), 4-6 low and 0-3 critically low. The data are shown below.

Usual Care	8	7	6	2	5	8	7	3
New Program	9	9	7	8	10	9	6	

Is there statistical evidence of a difference in APGAR scores in women receiving the new and enhanced versus usual prenatal care?

EXAMPLE 2

- **Step 1.** Set up hypotheses and determine level of significance.

  H_0: The two populations are equal

  H_1: The two populations are not equal. $\alpha = 0.05$

- **Step 2.** Select the appropriate test statistic.
  - Because APGAR scores are not normally distributed and the samples are small ($n_1=8$ and $n_2=7$), we can use the Mann Whitney U test.
  - The test statistic is U, the smaller of U_1 and U_2

- **Step 3.** Set up decision rule.
- The appropriate critical value can be found from the table. To determine the appropriate critical value we need sample sizes ($n_1=8$ and $n_2=7$) and our two-sided level of significance ($\alpha=0.05$).
- The critical value for this test with $n_1=8$, $n_2=7$ and $\alpha = 0.05$ is 10 and the decision rule is as follows: **Reject H_0 if $U \leq 10$.**

EXAMPLE 2

- **Step 4.** Compute the test statistic.
- The first step is to assign ranks of 1 through 15 to the smallest through largest values in the total sample, as follows:

		Total Sample (Ordered Smallest to Largest)		Ranks	
Usual Care	New Program	Usual Care	New Program	Usual Care	New Program
8	9	2		1	
7	8	3		2	
6	7	5		3	
2	8	6	6	4.5	4.5
5	10	7	7	7	7
8	9	7		7	
7	6	8	8	10.5	10.5
3		8	8	10.5	10.5
			9		13.5
			9		13.5
			10		15
				R_1=45.5	R_2=74.5

EXAMPLE 2

- Next, we sum the ranks in each group. In the usual care group, the sum of the ranks is R_1=45.5 and in the new program group, the sum of the ranks is R_2=74.5. Recall that the sum of the ranks will always equal n(n+1)/2. As a check on our assignment of ranks, we have n(n+1)/2 = 15(16)/2=120 which is equal to 45.5+74.5 = 120.
- We now compute U_1 and U_2, as follows:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 8(7) + \frac{8(9)}{2} - 45.5 = 46.5$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = 8(7) + \frac{7(8)}{2} - 74.5 = 9.5$$

Thus, the test statistic is U=9.5.

- **Step 5.** Conclusion:

We reject H_0 because 9.5 $\leq$ 10. We have statistically significant evidence at α =0.05 to show that the populations of APGAR scores are not equal in women receiving usual prenatal care as compared to the new program of prenatal care.

TESTS WITH MATCHED SAMPLES

- This section describes nonparametric tests to compare two groups with respect to a continuous outcome when the data are collected on matched or paired samples.
- This section describes procedures that should be used when the outcome cannot be assumed to follow a normal distribution. There are two popular nonparametric tests to compare outcomes between two matched or paired groups. The first is called the **Sign Test** and the second the **Wilcoxon Signed Rank Test**.
- When data are matched or paired, we compute difference scores for each individual and analyze difference scores. The same approach is followed in nonparametric tests. In parametric tests, the null hypothesis is that the mean difference (μ_d) is zero. In nonparametric tests, the null hypothesis is that the median difference is zero.

TESTS WITH MATCHED SAMPLES – PARAMETRIC TESTS

- The two comparison groups are said to be ***dependent,*** and the data can arise from a single sample of participants where each participant is measured twice (possibly before and after an intervention) or from two samples that are matched on specific characteristics (e.g., siblings).
- When the samples are dependent, we focus on ***difference scores*** in each participant or between members of a pair and the test of hypothesis is based on the mean difference, μ_d. The null hypothesis again reflects "no difference" and is stated as H_0: $\mu_d = 0$.
- Note that there are some instances where it is of interest to test whether there is a difference of a particular magnitude (e.g., $\mu_d = 5$) but in most instances the null hypothesis reflects no difference (i.e., $\mu_d = 0$).

TESTS WITH MATCHED SAMPLES – PARAMETRIC TESTS

Example:

- A new drug is proposed to lower total cholesterol and a study is designed to evaluate the efficacy of the drug in lowering cholesterol. Fifteen patients agree to participate in the study and each is asked to take the new drug for 6 weeks. However, before starting the treatment, each patient's total cholesterol level is measured. The initial measurement is a pre-treatment or baseline value. After taking the drug for 6 weeks, each patient's total cholesterol level is measured again and the data are shown below. The rightmost column contains difference scores for each patient, computed by subtracting the 6 week cholesterol level from the baseline level. The differences represent the reduction in total cholesterol over 4 weeks. (The differences could have been computed by subtracting the baseline total cholesterol level from the level measured at 6 weeks. The way in which the differences are computed does not affect the outcome of the analysis only the interpretation.)

TESTS WITH MATCHED SAMPLES – PARAMETRIC TESTS

Subject Identification Number	Baseline	6 Weeks	Difference
1	215	205	10
2	190	156	34
3	230	190	40
4	220	180	40
5	214	201	13
6	240	227	13
7	210	197	13
8	193	173	20
9	210	204	6
10	230	217	13
11	180	142	38
12	260	262	-2
13	210	207	3
14	190	184	6
15	200	193	7

Because the differences are computed by subtracting the cholesterols measured at 6 weeks from the baseline values, positive differences indicate reductions and negative differences indicate increases (e.g., participant 12 increases by 2 units over 6 weeks). The goal here is to test whether there is a statistically significant reduction in cholesterol.

TESTS WITH MATCHED SAMPLES – PARAMETRIC TESTS

In order to conduct the test, we need to summarize the differences. In this sample, we have

Subject Identification Number	Difference	Difference^2
1	10	100
2	34	1156
3	40	1600
4	40	1600
5	13	169
6	13	169
7	13	169
8	20	400
9	6	36
10	13	169
11	38	1444
12	-2	4
13	3	9
14	6	36
15	7	49
Totals	**254**	**7110**

$$s_d = \sqrt{\frac{\Sigma \text{Differences}^2 - (\Sigma \text{Differences})^2 / n}{n-1}}$$

$$s_d = \sqrt{\frac{7110 - (254)^2/15}{14}} = \sqrt{\frac{2808.93}{14}} = \sqrt{200.64} = 14.2$$

n =15,

$\bar{x}_d = 16.9,$

$S_d = 14.2$

TESTS WITH MATCHED SAMPLES – PARAMETRIC TESTS

Is there statistical evidence of a reduction in mean total cholesterol in patients after using the new medication for 6 weeks? We will run the test using the five-step approach.

- **Step 1.** Set up hypotheses and determine level of significance

$$H_0: \mu_d = 0$$

$$H_1: \mu_d > 0 \qquad \alpha=0.05$$

NOTE: If we had computed differences by subtracting the baseline level from the level measured at 6 weeks then negative differences would have reflected reductions and the research hypothesis would have been $H_1: \mu_d < 0$.

- **Step 2**. Select the appropriate test statistic.

Because the sample size is small (n<30) the appropriate test statistic is

$$t = \frac{\bar{X}_d - \mu_d}{s_d/\sqrt{n}}$$

- **Step 3.** Set up decision rule.

This is an one-tailed test, using a t statistic and a 5% level of significance. The appropriate critical value can be found in the t Table at the right, with df=15-1=14. The critical value for an upper-tailed test with df=14 and α=0.05 is 2.145 and the decision rule is Reject H_0 if $t \geq 2.145$.

TESTS WITH MATCHED SAMPLES – PARAMETRIC TESTS

- **Step 4.** Compute the test statistic.
- We now substitute the sample data into the formula for the test statistic identified in Step 2.

$$t = \frac{\bar{X}_d - \mu_d}{s_d/\sqrt{n}} = \frac{16.9 - 0}{14.2/\sqrt{15}} = 4.61$$

- **Step 5.** Conclusion.
- We reject H_0 because 4.61 ≥ 2.145. We have statistically significant evidence at α=0.05 to show that there is a reduction in cholesterol levels over 6 weeks.

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TESTS

Example:

- Consider a clinical investigation to assess the effectiveness of a new drug designed to reduce repetitive behaviors in children affected with autism. If the drug is effective, children will exhibit fewer repetitive behaviors on treatment as compared to when they are untreated. A total of 8 children with autism enroll in the study. Each child is observed by the study psychologist for a period of 3 hours both before treatment and then again after taking the new drug for 1 week. The time that each child is engaged in repetitive behavior during each 3 hour observation period is measured. Repetitive behavior is scored on a scale of 0 to 100 and scores represent the percent of the observation time in which the child is engaged in repetitive behavior. For example, a score of 0 indicates that during the entire observation period the child did not engage in repetitive behavior while a score of 100 indicates that the child was constantly engaged in repetitive behavior. The data are shown below.

Child	Before Treatment	After 1 Week of Treatment
1	85	75
2	70	50
3	40	50
4	65	40
5	80	20
6	75	65
7	55	40
8	20	25

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TESTS

- Looking at the data, it appears that some children improve (e.g., Child 5 scored 80 before treatment and 20 after treatment), but some got worse (e.g., Child 3 scored 40 before treatment and 50 after treatment). Is there statistically significant improvement in repetitive behavior after 1 week of treatment?.
- Because the before and after treatment measures are paired, we compute difference scores for each child. In this example, we subtract the assessment of repetitive behaviors after treatment from that measured before treatment so that difference scores represent improvement in repetitive behavior. The question of interest is whether there is significant improvement after treatment.

Child	Before Treatment	After 1 Week of Treatment	Difference (Before-After)
1	85	75	10
2	70	50	20
3	40	50	-10
4	65	40	25
5	80	20	60
6	75	65	10
7	55	40	15
8	20	25	-5

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TESTS

- In this small sample, the observed difference (or improvement) scores vary widely and are subject to extremes (e.g., the observed difference of 60 is an outlier). Thus, a nonparametric test is appropriate to test whether there is significant improvement in repetitive behavior before versus after treatment. The hypotheses are given below.

  H_0: The median difference is zero

  H_1: The median difference is positive

- In this example, the null hypothesis is that there is no difference in scores before versus after treatment. If the null hypothesis is true, we expect to see some positive differences (improvement) and some negative differences (worsening). If the research hypothesis is true, we expect to see more positive differences after treatment as compared to before.

How to solve this: Sign test and Wilcoxon signed Rank test

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TESTS

SIGN TEST:

- The Sign Test is the simplest nonparametric test for matched or paired data. The approach is to analyze only the signs of the difference scores, as shown below:

Child	Before Treatment	After 1 Week of Treatment	Difference (Before-After)	Sign
1	85	75	10	+
2	70	50	20	+
3	40	50	-10	-
4	65	40	25	+
5	80	20	60	+
6	75	65	10	+
7	55	40	15	+
8	20	25	-5	-

If the null hypothesis is true (i.e., if the median difference is zero) then we expect to see approximately half of the differences as positive and half of the differences as negative. If the research hypothesis is true, we expect to see more positive differences.

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TESTS

Test Statistic for the Sign Test

- The test statistic for the Sign Test is the number of positive signs or number of negative signs, whichever is smaller. In this example, we observe 2 negative and 6 positive signs. Is this evidence of significant improvement or simply due to chance?
- Determining whether the observed test statistic supports the null or research hypothesis is done following the same approach used in parametric testing. Specifically, we determine a critical value such that if the smaller of the number of positive or negative signs is less than or equal to that critical value, then we reject H_0 in favor of H_1 and if the smaller of the number of positive or negative signs is greater than the critical value, then we do not reject H_0. Notice that this is a one-sided decision rule corresponding to our one-sided research hypothesis.

Critical Values for the Sign Test

- To determine the appropriate critical value we need the sample size, which is equal to the number of matched pairs (n=8) and our one-sided level of significance α=0.05. For this example, the critical value is 5, and the decision rule is to reject H0 if the smaller of the number of positive or negative signs < 5.

Conclusion:

- **We reject H0 because 2 < 5.** We have sufficient evidence at α=0.05 to show that there is improvement in repetitive behavior after taking the drug as compared to before. In essence, we could use the critical value to decide whether to reject the null hypothesis.

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TESTS

When Difference Scores are Zero

- There is a special circumstance that needs attention when implementing the Sign Test which arises when one or more participants have difference scores of zero (i.e., their paired measurements are identical). If there is just one difference score of zero, some investigators drop that observation and reduce the sample size by 1 (i.e., the sample size would be n-1). This is a reasonable approach if there is just one zero. However, if there are two or more zeros, an alternative approach is preferred.
- If there is an even number of zeros, we randomly assign them positive or negative signs.
- If there is an odd number of zeros, we randomly drop one and reduce the sample size by 1, and then randomly assign the remaining observations positive or negative signs.

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TESTS

Example:

- A new chemotherapy treatment is proposed for patients with breast cancer. Investigators are concerned with patient's ability to tolerate the treatment and assess their quality of life both before and after receiving the new chemotherapy treatment. Quality of life (QOL) is measured on an ordinal scale and for analysis purposes, numbers are assigned to each response category as follows: 1=Poor, 2= Fair, 3=Good, 4= Very Good, 5 = Excellent. The data are shown below.

Patient	QOL Before Chemotherapy Treatment	QOL After Chemotherapy Treatment
1	3	2
2	2	3
3	3	4
4	2	4
5	1	1
6	3	4
7	2	4
8	3	3
9	2	1
10	1	3
11	3	4
12	2	3

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TESTS

The question of interest is whether there is a difference in QOL after chemotherapy treatment as compared to before.

- **Step 1.** Set up hypotheses and determine level of significance.

  H_0: The median difference is zero

  H_1: The median difference is not zero (α=0.05)

- **Step 2.** Select the appropriate test statistic.

The test statistic for the Sign Test is the smaller of the number of positive or negative signs.

- **Step 3.** Set up the decision rule.

The appropriate critical value for the Sign Test can be found in the table of critical values for the Sign Test. To determine the appropriate critical value we need the sample size (or number of matched pairs, n=12), and our two-sided level of significance α=0.05.

The critical value for this two-sided test with n=12 and a =0.05 is 13, and the decision rule is as follows: Reject H_0 if the smaller of the number of positive or negative signs $\leq$ 13.

- **Step 4.** Compute the test statistic.

Because the before and after treatment measures are paired, we compute difference scores for each patient. In this example, we subtract the QOL measured before treatment from that measured after.

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TESTS

We now capture the signs of the difference scores and because there are two zeros, we randomly assign one negative sign (i.e., "-" to patient 5) and one positive sign (i.e., "+" to patient 8), as follows:

Patient	QOL Before Chemotherapy Treatment	QOL After Chemotherapy Treatment	Difference (After-Before)	Sign
1	3	2	-1	-
2	2	3	1	+
3	3	4	1	+
4	2	4	2	+
5	1	1	0	-
6	3	4	1	+
7	2	4	2	+
8	3	3	0	+
9	2	1	-1	-
10	1	3	2	+
11	3	4	1	+
12	2	3	1	+

The test statistic is the number of negative signs which is equal to 3.

Step 5. Conclusion.

We reject H_0 because 3 < 13. We have statistically significant evidence at α=0.05 to show that there is a difference in QOL after chemotherapy treatment as compared to before.

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TEST- WILCOXON SIGNED RANK TEST

Another popular nonparametric test for matched or paired data is called the Wilcoxon Signed Rank Test. Like the Sign Test, it is based on difference scores, but in addition to analyzing the signs of the differences, it also takes into account the magnitude of the observed differences.

Child	Before Treatment	After 1 Week of Treatment
1	85	75
2	70	50
3	40	50
4	65	40
5	80	20
6	75	65
7	55	40
8	20	25

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TEST- WILCOXON SIGNED RANK TEST

Child	Before Treatment	After 1 Week of Treatment	Difference (Before-After)
1	85	75	10
2	70	50	20
3	40	50	-10
4	65	40	25
5	80	20	60
6	75	65	10
7	55	40	15
8	20	25	-5

The next step is to rank the difference scores. We first order the *absolute values of the difference scores* and assign rank from 1 through n to the smallest through largest absolute values of the difference scores, and assign the mean rank when there are ties in the absolute values of the difference scores.

Observed Differences		Ordered Absolute Values of Differences	Ranks
10		-5	1
20		10	3
-10		-10	3
25		10	3
60		15	5
10		20	6
15		25	7
-5		60	8

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TEST- WILCOXON SIGNED RANK TEST

The final step is to attach the signs ("+" or "-") of the observed differences to each rank as shown below.

Observed Differences		Ordered Absolute Values of Difference Scores	Ranks	Signed Ranks
10		-5	1	-1
20		10	3	3
-10		-10	3	-3
25		10	3	3
60		15	5	5
10		20	6	6
15		25	7	7
-5		60	8	8

Similar to the Sign Test, hypotheses for the Wilcoxon Signed Rank Test concern the population median of the difference scores. The research hypothesis can be one- or two-sided. Here we consider a one-sided test.

H_0: The median difference is zero
H_1: The median difference is positive

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TEST- WILCOXON SIGNED RANK TEST

Test Statistic for the Wilcoxon Signed Rank Test

The test statistic for the Wilcoxon Signed Rank Test is W, defined as the smaller of W+ (sum of the positive ranks) and W- (sum of the negative ranks). If the *null hypothesis* is true, we expect to see similar numbers of lower and higher ranks that are both positive and negative (i.e., W+ and W- would be similar). If the *research hypothesis* is true we expect to see more higher and positive ranks (in this example, more children with substantial improvement in repetitive behavior after treatment as compared to before, i.e., W+ much larger than W-).

In this example, W+ = 32 and W- = 4. Recall that the sum of the ranks (ignoring the signs) will always equal n(n+1)/2. As a check on our assignment of ranks, we have n(n+1)/2 = 8(9)/2 = 36 which is equal to 32+4. The test statistic is W = 4.

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TEST- WILCOXON SIGNED RANK TEST

Critical Values of W

To determine the appropriate one-sided critical value we need sample size (n=8) and our one-sided level of significance (α=0.05). For this example, the critical value of W is 5 and the decision rule is to reject H_0 if $W \leq 5$. Thus, we reject H_0, because $4 \leq 5$.

Conclusion:

We have statistically significant evidence at α =0.05, to show that the median difference is positive (i.e., that repetitive behavior improves.)

Any question?

REFERENCE

- The detail material related to this lecture can be found in

- D'Agostino RB and Stevens MA. Goodness of Fit Techniques.
- Apgar, Virginia (1953). " A proposal for a new method of evaluation of the newborn infant ". Curr. Res. Anesth. Analg. 32 (4): 260-267.
- Conover WJ. Practical Nonparametric Statistics, 2nd edition, New York: John Wiley and Sons.
- Siegel and Castellan. (1988). "Nonparametric Statistics for the Behavioral Sciences," 2nd edition, New York: McGraw-Hill.