PART 4

Wilcoxon

Imagine you give an attitude test to a small group of people. After you deliver some treatment, say a daily vitamin supplement for a few weeks, you give the same group of people another attitude test. You then compare the two measures to see if there is any meaningful difference between the two sets of scores.

The characteristic to note is that every test score is paired. Since everyone is tested twice, for each initial test score, there is another test score for that same person after the treatment. This forms a pair for each test score[1]. The parametric equivalent to these tests is called the Student's t-test, t-test for matched pairs, t-test for paired samples, or t-test for dependent samples.

Note that the procedures for the Wilcoxon signed-rank and sign tests change depending on whether the samples are small or large.

[1] Assuming no attrition, of course

4.1

RANKING DATA

Many nonparametric procedures involve ranking data values. To rank all values from the first

Students who ate breakfast	Students who skipped breakfast
87	93
96	83
92	79
84	73

Table 1 Healthy breakfast or not

table, we simply place them all in order in a new table from smallest to largest. The table below shows how to do this with our breakfast example, keeping the values for the students who ate breakfast bolded. On the surface, it seems that the students who ate breakfast scored higher.

Students who ate breakfast	Students who skipped breakfast
73	1
79	2
83	3
84	4
87	5
92	6
93	7
96	8

Table 2 Ranking data

However, if you wish to claim statistical significance, some type of procedure or test is required.

What if we have ties? Imagine we replaced the score for the student with a rank of 4 (scoring 84 points) with 83 points. This student is now tied with the student with a rank of 3. We simply give the students the average of their rank values. See the table below for how this would work. The rows in bold are now the tied values instead of values for those who ate breakfast. Most nonparametric statistical tests require a different formula when a sample of data contains ties.

Students who ate breakfast	Students who skipped breakfast
73	1
79	2
83	**3.5**
83	**3.5**
87	5
92	6
93	7
96	8

Table 3 Ties in data

4.2

WILCOXON SIGNED-RANK/WILCOXON MATCHED-PAIRS

The formula for computing the Wilcoxon T for small samples is:

$$T = \text{smaller of } \sum \mathbb{R}_+ \text{ and } \sum \mathbb{R}_- \tag{11}$$

where $\sum \mathbb{R}_+$ is the sum of the ranks with positive differences and $\sum \mathbb{R}_-$ is the sum of the ranks with negative differences.

We then examine the T statistic using the relevant table of critical values. However, if the number of pairs n exceeds those available from the table, then a large sample approximation may be performed. Simply put, we compute a z-score and use a table with the normal distribution to obtain a critical region of z-scores. First calculate:

$$\bar{x}_T = \frac{n\,(n+1)}{4} \tag{12}$$

where $\bar{x}_T$ is the mean and n is the number of matched pairs included in the analysis,

$$s_T = \sqrt{\frac{n\,(n+1)\,(2n+1)}{24}} \tag{13}$$

where s_T is the standard deviation. The z-score approximation is therefore:

$$z^* = \frac{T - \bar{x}_T}{s_T}. \tag{14}$$

Let's work through a sample matched-pairs test with small data samples. The counseling staff of Clear Creek County School District has implemented a new program this year to attempt reducing bullying at primary schools. To evaluate the effectiveness of our anti-bullying program, the school district has decided to compare the percentage of successful interventions last year before the program began with the percentage of successful interventions this year with the program in

place[1]. The next table shows 12 elementary school counselors and their reported percentage of successful interventions last year and this. Since the samples are small, we need a nonparametric

Participant number	Last year	This year
1	31	31
2	14	14
3	53	50
4	18	30
5	21	28
6	44	48
7	12	35
8	36	32
9	29	34
10	29	34
11	17	27
12	40	42

Table 4 Percentage of successful interventions

procedure. Since we are comparing two related, or paired, samples, we will use the Wilcoxon matched-pairs test. We will use the steps laid out in the previous section.

State the null and alternative hypotheses. The null hypothesis states that the counselors reported no difference in the percentages last year and this year. The alternative states that the counselors observed some differences between this year and last year. Note that our alternative hypothesis does not claim that these differences are supposed to be positive ones, so our alternative is a two-tailed, nondirectional hypothesis. They can be written as

$$H_0 : \mu_D = 0$$
$$H_A : \mu_D \neq 0.$$

Set the level of risk. It is standard to choose $\alpha = 0.05$, which states that we will make type I errors 5% of the time.

Choose the appropriate test statistic. We have already shown that the Wilcoxon matched-pairs test is appropriate.

Compute the test statistic. Exclude data points where the difference between two matched pairs is zero, then rank the data again using the **absolute values**. With the new ranks, check that the ranks with positive differences are $9, 7, 4.5, 10, 1, 6, 8$, and 2. Adding the ranks with positive difference we get $\sum \mathbb{R}_+ = 47.5$. The ranks with negative differences are 3 and 4.5, summing to $\sum \mathbb{R}_- = 7.5$. We choose the smaller of the two rank sums and therefore the Wilcoxon is $T = 7.5$.

[1]You may realise that since we have not stated what kind of program this is nor defined clearly what "interventions" mean or what a "successful intervention" would look like, this is not a very good experiment. Hopefully in any experiments you run in real life, you will have defined criteria for success better than I have here but for now, we will assume that we have some well-defined concept of a "successful intervention".

Determine the critical value needed to reject the null hypothesis. To find the critical value of the Wilcoxon matched-pairs test, we need the sample size, level of significance, and type of hypothesis. Note that since we discarded the two participants with no reported differences between this year and the last, we now have a sample size of 10 instead of 12. The table in the course handbook says that for a significance level of 0.05 with a two-tailed hypothesis and $n = 10$, the critical value is $T = 8$. An obtained value less than or equal to 8 will lead us to reject our null hypothesis.

Compare obtained value with critical value. Since our obtained value $T = 7.5$ is less than the critical value, 8, we reject the null hypothesis.

Interpret and report results. By rejecting the null hypothesis, we suggest that a real difference exists between last year's percentages and this year's percentages. Reporting the results would go something like this.
For this experiment, the Wilcoxon signed rank (matched-pairs) test ($T = 7.5, n = 12, p < 0.05$) indicated that the difference in percentage of successful interventions was statistically significant. In addition, the sum of the positive difference ranks ($\sum \mathbb{R}_+ = 47.5$ was larger than the sum of the negative difference ranks ($\sum \mathbb{R}_- = 7.5$), showing a positive impact from the program. Therefore, our analysis supports that the new anti-bullying program provides positive benefits towards the improvement of the aspect of student behaviour that school counselors perceive.

Confidence Intervals

Now we aim to construct a confidence interval based on the Wilcoxon signed rank test for matched pairs. Such a confidence interval provides a range of values that fall within the population with a $100\,(1-\alpha)$ percent chance. So now instead of just answering yes or no, we can say that the difference in successful interventions due to the new anti-bullying program is some interval $[x, y]$.

First, let us introduce ourselves with some notation. Index each participant as some $i = 1, \ldots, n$. Each counselor i has matched pair $\left(x_1^i, x_2^i\right)$, where x_1^i is the i's reported percentage of successful interventions before the anti-bullying program was implemented and x_2^i the reported percentage after program implementation. Compute value $D_i = x_1^i - x_2^i$ for each participant i. Then compute u_{ij} as defined below:

$$u_{ij} = \frac{D_i + D_j}{2} \text{ for all} 1 \leq i \leq j \leq n.$$

So let's pause here to give some interpretation. All we have done here is said that the difference between the reported percentages of counselor i is represented by D_i so for each of our matched pair data points, we have another number D_i. Then we calculate u_{ij} for all unique combinations of individuals. For example:

$$u_{11} = \frac{D_1 + D_1}{2} = \frac{-3 + -3}{2} = -3$$
$$u_{12} = \frac{D_1 + D_2}{2} = \frac{-3 + 12}{2} = \frac{9}{2}.$$

You may notice that the number of u_{ij}s we have is a counting problem. If you wish to practice, note that $u_{12} = u_{21}$ so we have an unordered problem with replacement. Then order these averages from smallest to largest. The median of the ordered averages gives a point estimate of the population median difference.

Use the table in the handbook to find the endpoints of the confidence interval. First, determine T from the table that corresponds with the sample size and desired confidence such that $p = \alpha/2$. We seek to find a 95% confidence interval and we have $n = 10$ and $p = 0.05/2$. For single-tailed hypotheses, the table provides $T = 8$. Note that we are using a single-tailed hypothesis because we wish to provide a confidence interval around the point estimate of the population median. Therefore, we construct two single-tailed hypotheses around the median with a level of significance that adds up to $\alpha = 0.05$.

The endpoints of the confidence interval are the Kth smallest and Kth largest values of u_{ij}, where $K = T + 1$. So for our case, $K = 9$. The ninth smallest value is 0.5 and the ninth largest is 12. Therefore, we can say that we are 95% confident that the difference of successful interventions due to the new bullying programs lies between 0.5 and 12.

I can go through the same test with a large data sample and normal approximation if required.

4·3

WILCOXON SIGN TEST/RANK-SUM TEST

Related samples can be analyzed more efficiently if the data is reduced to, or already comes in, dichotomous (binary) results. This method is based on Gibbons and Chakraborti (2010).

We start by identifying the sign of the difference demonstrated by each set from the related data samples, then sum all positive and negative differences as n_p and n_n, respectively.