How does the **Wilcoxon-signed rank test** work?

by Juliana Vega Lacorte

1. Non-parametric tests

Each statistical test has its own prerequisites and assumptions. The commonly used t-test assumes a normal distribution of data. If we know that our data is not normally distributed but we cannot say exactly what type of distribution applies, we might consider using a non-parametric test. Non-parametric tests do not make specific assumptions about the distribution of a population.

Non-parametric tests are also called "distribution-free" tests. Just don't let the name mislead you into thinking they make absolutely no assumptions about the distribution. Truth is, non-parametric tests do make assumptions about the distribution but in a more general way, hence less restrictive. For example, the Wilcoxon-signed rank test assumes the distribution is symmetric. But the parametric t-test assumes a very specific form: a normal distribution. A large number of distributions would satisfy being "symmetric", but less would fall within the category "normally distributed".

Here is an overview of the main non-parametric tests:

Non-parametric	Parametric version	
Wilcoxon-Signed Rank Test	paired t-test	Also categorized as a one sample test.
Mann-Whitney U test	two independent samples t-test	Also known as Wilcoxon rank rum test or Wilcoxon-Mann-Whitney test.
Kolmogorov-Smirnov test	two independent samples t-test	
Sign test	one sample/ paired t-test	It is the simplest nonparametric test for paired data, but the Wilcoxon-Signed rank test is preferred because it uses more information.
Kruskal-Wallis	multiple samples	

Since you need a test is for paired data (same sample of individuals before and after training), I will focus on the Wilcoxon-signed rank test and explain how it works in the following pages.

2. Wilcoxon-signed rank test (WSR)

Procedure

1. Taking Differences

The WSR test procedure starts by taking the differences between each pair of observations. In this case, the difference between before and after training. In the treatment literature these differences are sometimes called „responses".

Let's say we have data on scores before and after training for twelve individuals. The Table on the right has some fake data to illustrate the process. The values in **Column (3)** are the **differences**: After - Before.

Table 1. Scores before-after training

	(1)	(2)	**(3)**
id	Before	After	**Diff.**
1	2	1	-1
2	4	3	-1
3	3	2	-1
4	3	5	2
5	1	6	5
6	4	7	3
7	1	7	6
8	7	3	-4
9	6	1	-5
10	2	4	2
11	7	4	-3
12	4	6	2

2. Assigning Ranks

Working with the differences, we need to **order** these values **from lowest to highest**, ignoring the sign. The sign is ignored because we want to assign ranks based on absolute differences. The ranks are just numbers to specify the position each value occupies in the ordered series. A rank of 1 is assigned to the smallest value, rank 2 to the next, and so on.

Column (4) shows the **absolute differences**. These are the values that need to be ordered. And that's exactly what is shown in column (5).

Table 2. Absolute differences in ascending order

(3)	**(4)**		**(5)**
Diff.	|Diff|		Ordered |Diff|
-1	1		1
-1	1		1
-1	1		1
2	2	→	2
5	5		2
3	3		2
6	6		3
-4	4		3
-5	5		4
2	2		5
-3	3		5
2	2		6

Assigning the ranks is straightforward, except for cases in which we have two or more values that are equal. **Repeated values** are called a „tie“ or „**tied ranks**“. And the rule is to assign them the mid-rank, that is, the average of the ranks the values would have gotten if they were not repeated values.

Column (6) shows the **ranks** that correspond to the ordered values in column (5). As you can see, the first three values of 1 have been assigned the rank of 2. This rank comes from taking the average of one, two and three, because the three values of 1 occupy the first three positions. Following the same rule, the values of 2 are assigned a rank of 5, the values of 3 have rank 7.5, and so on. I have included the column *Position* just to keep track of the numbers that are averaged.

Table 3. Ranks

(5) |Diff|	(6) **Rank**		Position
1	2	$\frac{1+2+3}{3}$	1
1	2		2
1	2		3
2	5	$\frac{3+4+5}{3}$	4
2	5		5
2	5		6
3	7.5		7
3	7.5		8
4	9		9
5	10.5		10
5	10.5		11
6	12		12

3. Putting back the signs

The signs were eliminated just to be able to order the data and assign ranks based on the magnitude of the differences, irrespective of the direction. But at the end, the test does need to distinguish what **ranks** refer to **positive** differences and which ones refer to **negative** differences. This is the reason why the signs are re-attached to the ranks and why the test is called signed-rank.

In the table you can see the signs corresponding to each rank (column 7), based on the original differences (the ones from Column 3).

Table 4. Signed-ranks

(6) Rank	(7) Sign	(8) **Signed** Ranks
2	-	-2
2	-	-2
2	-	-2
5	+	5
5	+	5
5	+	5
7.5	+	7.5
7.5	-	-7.5
9	-	-9
10.5	+	10.5
10.5	-	-10.5
12	+	12

4. Sums of ranks

Now that the ranks are being classified according to signs, their respective **sums** can be calculated. For our data, the sum of ranks for the positive signs (W^+) equals 45; the sum of ranks for the negative signs (W^-) equals 33. With this information, the test statistic for the **Wilcoxon Signed-Rank** test is computed. Either W^+ or W^- can be taken as the test statistic (Stata uses W^+).The test statistic is then compared to a critical value to decide whether or not to reject the null.

Table 5. Sum of Ranks

(6) Rank	(7) Sign	(8) Signed Ranks
2	-	-2
2	-	-2
2	-	-2
5	+	5
5	+	5
5	+	5
7.5	+	7.5
7.5	-	-7.5
9	-	-9
10.5	+	10.5
10.5	-	-10.5
12	+	12

Sum **Positive** ranks	W^+	45
Sum **Negative** ranks	W^-	33

The logic behind the test

The Wilcoxon Signed-Rank test is testing the null hypothesis of no treatment effect. In statistical terms, the null hypothesis is that the distribution of differences has a median of zero(H_0: Md = 0). The intuition behind is that if there is no effect in either direction, the sum of positive ranks and the sum of negative ranks will be approximately equal. On the contrary, if a positive (negative) effect is present, it will show itself as a higher sum of positive (negative) ranks; higher enough for us to believe that this is not just due to chance.

In terms of our example, from the twelve individuals that are in the sample, six experienced a positive difference, which means their scores increased after the training. We also have the same number of individuals with a negative difference. But the magnitude of the positive differences was slightly higher than the negative ones, and so the positive differences got higher ranks. This is reflected in the slightly higher sum of positive ranks (45). Actually, from Table 1, we see there's someone (id 7) with a difference of +6. Since this is the highest difference in our sample of twelve, it gets the rank of 12 (see Table 3). Looking at the sums of ranks, the positive and negative sums differ exactly by 12. The question we will try to answer with the test is: does the difference in the sums of ranks is high enough to suggest that the scores increased after the training?

3. Implementation in Stata

The command you need in order to run the test in Stata is **signrank**

The syntax is: `signrank [var1] = [var2]`

For our example, we would write: `signrank scores_after = scores_before`

The output we get is:

```
Wilcoxon signed-rank test

        sign |      obs    sum ranks    expected
-------------+---------------------------------
    positive |        6           45          39
    negative |        6           33          39
        zero |        0            0           0
-------------+---------------------------------
         all |       12           78          78

unadjusted variance      162.50
adjustment for ties       -1.25
adjustment for zeros       0.00
                     ----------
adjusted variance        161.25

Ho: scores_after = scores_before
             z =   0.472
    Prob > |z| =   0.6366
```

As you can see, we get the same sum ranks we calculated in the steps before. Now, to make a conclusion based on this test we need to look at the p-value, shown as Prob > |z| = 0.6366. The decision rule, like with any other statistical test, is to reject the null if the p-value is less than the chosen significance level (e.g. if p-value < 0.05 if we choose a significance level of 5%). Given the results obtained, the p-value of 0.6366 indicates that we cannot reject the null of no effect. In other words, there is not enough evidence to suggest that the training had an effect on the scores. Going back to our sample, we already knew there was no clear predominance going in one direction and this is confirmed by the test. Negative and positive responses were almost equal except for one outlier (person with id 7). And one outlier is not enough to suggest that the effect exists.

Assumptions

→ Values can be compared so their differences make sense. We can say that one value is greater, equal, or less than the other.
→ The distribution of the differences is symmetric.