

The Wilcoxon test:

Use this when the *same* participants perform *both* conditions of your study: i.e., it is appropriate for analysing the data from a repeated-measures design with two conditions. Use it when the data do not meet the requirements for a parametric test (i.e. if the data are not normally distributed; if the variances for the two conditions are markedly different; or if the data are measurements on an ordinal scale). Otherwise, if the data meet the requirements for a parametric test, it is better to use a repeated-measures *t*-test (also known as a "dependent means" or "matched pairs" *t*-test).

The logic behind the Wilcoxon test is quite simple. The data are ranked to produce two rank totals, one for each condition. If there is a systematic difference between the two conditions, then most of the high ranks will belong to one condition and most of the low ranks will belong to the other one. As a result, the rank totals will be quite different and one of the rank totals will be quite small. On the other hand, if the two conditions are similar, then high and low ranks will be distributed fairly evenly between the two conditions and the rank totals will be fairly similar and quite large. The Wilcoxon test statistic "W" is simply the smaller of the rank totals. The SMALLER it is (taking into account how many participants you have) then the less likely it is to have occurred by chance. A table of critical values of W shows you how likely it is to obtain your particular value of W purely by chance. Note that the Wilcoxon test is unusual in this respect: normally, the BIGGER the test statistic, the less likely it is to have occurred by chance).

This handout deals with using Wilcoxon with small sample sizes. If you have a large number of participants, you can convert W into a *z*-score and look this up instead. The same is true for the Mann-Whitney test. There is a handout on my website that explains how to do this, for both tests.

Step by step example of the Wilcoxon test:

Suppose we wanted to know if people's ability to report words accurately was affected by which ear they heard them in. To investigate this, we performed a dichotic listening task. Each participant heard a series of words, presented randomly to either their left or right ear, and reported the words if they could. Each participant

thus provided two scores: the number of words that they reported correctly from their left ear, and the number reported correctly from their right ear. Do participants report more words from one ear than the other? Although the data are measurements on a ratio scale ("number correct" is a measurement on a ratio scale), the data were found to be positively skewed (i.e. not normally distributed) and so we use the Wilcoxon test.

Here are the data. It looks like, on average, more words are reported if they are presented to the right ear. However it's not a big difference, and not all participants show it. Therefore we'll use a Wilcoxon test to assess whether the difference between the ears could have occurred merely by chance.

Number of words reported:

Participant	Left ear	Right ear
1	25	32
2	29	30
3	10	7
4	31	36
5	27	20
6	24	32
7	27	26
8	29	33
9	30	32
10	32	32
11	20	30
12	5	32
median:	24.08	32.00

- a) Find the difference between each pair of scores.
- b) Rank these differences, ignoring any "0" differences and ignoring the sign of the difference (i.e. whether it is a positive or negative difference).

Participant	Left ear	Right ear	Difference (d)
1	25	32	-7
2	29	30	-1
3	10	7	3
4	31	36	-5
5	27	20	7
6	24	32	-8
7	27	26	1
8	29	33	-4
9	30	32	-2
10	32	32	0
11	20	30	-10
12	5	32	-27

To rank the differences:

Give the lowest rank to the smallest difference-score, ignoring whether it's a positive or negative difference. If two or more difference-scores are the same, this is a "tie": tied scores get the average of the ranks that those scores would have obtained, had they been different from each other.

Here, ignoring the sign of the difference, the lowest difference is -1. However there are two instances of this score (one positive and one negative). Therefore we add up the ranks that these scores *would* have had, if they had been different from each other (the ranks of 1 and 2), and then divide the sum of these ($1+2 = 3$) by the number of ranks involved (2). This gives us an "average" rank, 1.5, that we allocate to both of these two scores. The next lowest difference-score is -2. We have now used up the ranks of 1 and 2, so this difference-score gets the ranks of 3. After that, ranking is straightforward until we get to the two difference scores of -7 and 7. These would have got the ranks of 7 and 8, but instead get the average rank of 7.5 ($7+8 = 15$; $15/2 = 7.5$). This "uses up" the ranks of 7 and 8, so the next highest difference-score (-8) gets the rank of 9.

Participant	Left ear	Right ear	Difference (d)	ranked difference
1	25	32	-7	7.5
2	29	30	-1	1.5
3	10	7	3	4
4	31	36	-5	6
5	27	20	7	7.5
6	24	32	-8	9
7	27	26	1	1.5
8	29	33	-4	5
9	30	32	-2	3
10	32	32	0	ignore
11	20	30	-10	10
12	5	32	-27	11

(c) Add together the ranks belonging to scores with a positive sign (shaded in the table above):

$$4 + 7.5 + 1.5 = 13$$

(d) Add together the ranks belonging to scores with a negative sign (unshaded in the table above):

$$7.5 + 1.5 + 6 + 9 + 5 + 3 + 10 + 11 = 53$$

(e) Whichever of these sums is the smaller, is our value of W. So, $W = 13$.

(f) N is the number of differences (omitting "0" differences). We have $12 - 1 = 11$ differences. (NB: this is NOT the same as degrees of freedom. We only use $N-1$ here because we have one difference which equals zero. if we had two zero differences, we would use $N-2$, and so on).

(g) Use the table of critical Wilcoxon values (e.g. the one on my website, which is reproduced below). With an N of 11, the critical value for a two-tailed test at the 0.05 significance level is 11. (I've shaded the relevant row in the table below, to make it easy to find).

Critical values for the Wilcoxon test:

Table of critical values for the Wilcoxon test:

To use this table: compare your obtained value of Wilcoxon's test statistic to the critical value in the table (taking into account N, the number of subjects). Your obtained value is statistically significant if it is equal to or **SMALLER** than the value in the table.

e.g.: suppose my obtained value is 22, and I had 15 participants. The critical value in the table is 25: my obtained value is *smaller* than this, and so I would conclude that the difference between the two conditions in my study was unlikely to occur by chance (p<.05 two-tailed test, or p<.025, one-tailed test).

	One Tailed Significance levels:		
	0.025	0.01	0.005
	Two Tailed significance levels:		
N	0.05	0.02	0.01
6	0	-	-
7	2	0	-
8	4	2	0
9	6	3	2
10	8	5	3
11	11	7	5
12	14	10	7
13	17	13	10
14	21	16	13
15	25	20	16
16	30	24	20
17	35	28	23
18	40	33	28
19	46	38	32
20	52	43	38
21	59	49	43
22	66	56	49
23	73	62	55
24	81	69	61
25	89	77	68

With the Wilcoxon test, an obtained W is significant if it is LESS than or EQUAL to the critical value.

Our obtained value of 13 is *larger* than 11, and so we can conclude that there is no significant difference between the number of words recalled from the right ear and the number of words recalled from the left ear. We would write this as follows.

"A Wilcoxon test showed that the number of words reported correctly was not significantly affected by which ear they were presented to ($W(11) = 13, p > .05$, two-tailed test)."

This was a two-tailed test, because we were merely predicting that there would be some kind of difference between the two ears. Had we been able to make a more specific prediction *in advance of collecting the data*, e.g. that the right ear would be better than the left, then we could have used a "directional" or "one-tailed" test. This is more likely to be statistically significant, but only if the difference goes in the predicted direction: note how in the table, for an N of 11, a value of 10 for our obtained W would be significant at the .05 level for a two-tailed test, but at the .025 level for a one-tailed test.