

Deciding the statistical significance of nonparametric tests with large sample sizes:

The Wilcoxon test:

When you perform a Wilcoxon test, you obtain a value of the test statistic W , which reflects the size of the difference between the ranks for the two conditions.

You normally compare your obtained value of W to a critical value of W that is obtained from a table. However, when the number of scores is high (over about 20-25 or so), then the distribution of the values of W becomes increasingly like a normal distribution.

Hence we can convert W into a z-score, and then see how likely it is to obtain a z-score this big merely by chance. Remember that the bigger the z-score (either positive or negative), the less likely it is to have occurred by chance. (There are a number of ways of converting W to z ; the method below is just one of them).

Note that SPSS turns W into a z-score regardless of sample size, which is why it's shown as "Z" in its output for the Wilcoxon test.

A step by step example:

Suppose 30 people rate Ant and Dec for attractiveness. Column A shows the ratings for Ant, and column B shows the same people's ratings for Dec. The table below shows the rating scores, out of 20. The mean rating for Ant is 5.73 (s.d. = 1.80) and the mean rating for Dec is 7.53 (s.d. = 3.99). It looks like people find Dec more attractive than Ant. But does this represent a *significant* difference between Ant and Dec in terms of attractiveness? These are ratings, and each person has participated in both conditions (i.e., rating both Ant and Dec) and so the appropriate test is the Wilcoxon test.

Step 1:

Calculate W as usual. Find the difference between each pair of scores. Ignoring any zero differences (such as in the case of participants 1 and 21 here) rank the differences in size, ignoring the sign of the difference. Then add up the positive differences and the negative differences separately. W is the smaller sum of ranks. Here the sum of the positive differences is 100.5, and the sum of the negative differences is 305.5. So $W = 100.5$. Are the ratings for Ant and Dec significantly different? In order to decide, we need to know how likely it is that we would get a W this large merely by chance.

Ant	Dec	Difference	rank
3	3	0	ignore
4	8	-4	20.5
5	9	-4	20.5
6	7	-1	4
5	14	-9	28
6	7	-1	4
7	3	4	20.5
7	14	-7	25
8	11	-3	17
6	3	3	17
8	10	-2	11.5
9	8	1	4
5	6	-1	4
6	9	-3	17
7	8	-1	4
6	5	1	4
7	5	2	11.5
4	6	-2	11.5
3	4	-1	4
6	8	-2	11.5
5	5	0	ignore
8	14	-6	23.5
7	5	2	11.5
6	8	-2	11.5
3	1	2	11.5
1	3	-2	11.5
5	13	-8	26.5
6	2	4	20.5
5	11	-6	23.5
8	16	-8	26.5

Step 2:

Our original n is 30, but we have two ties (where the participant gave the same rating for both celebrities), so we ignore these in the calculations: our n now becomes $30-2 = 28$.

Step 3:

We need to work out the mean and standard deviation for the distribution of W, for our particular sample size. (Remember that the frequency distribution of scores can be specified completely if we know the mean, s.d. and shape of the distribution. We know that with a large number of scores, W is roughly normally distributed. Therefore we merely need to know the mean and s.d. of the distribution in order to be able to estimate how often various values of W will occur by chance, for any particular sample size).

The **mean** of the W distribution =

$$\mu = \frac{n(n+1)}{4} = \mu = \frac{28 * (28+1)}{4}$$

$$= 812 / 4 = \mathbf{203}.$$

The **standard deviation** of the W distribution is

$$\sigma = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sigma = \sqrt{\frac{28 * (29) * (56+1)}{24}} = \sigma = \sqrt{\frac{812 * 57}{24}}$$

$$= \mathbf{43.915}$$

(The denominator for the mean is always 4, and for the s.d. is always 24. Sorry, I don't know why - it just is).

Step 4:

$$z = (W - \mu) / \sigma \quad z = (100.5 - 203) / 43.915$$

$$z = \mathbf{-2.33}$$

Step 5:

Find the "area beyond z", using a table of areas under the normal curve (such as the one on my website). The area beyond z = 0.0099. In other words, values of W as large as ours are likely to occur by chance with a probability of 0.0099. This is the one-tailed probability of obtaining W however; we were merely seeing whether there was a *difference* in the ratings of Ant and Dec (a two-tailed, or non-directional, hypothesis), so we need to multiply this probability by 2, to get $p = .019$, which rounds to $p = .02$.

Compare this to the output from SPSS for a Wilcoxon test on these data, and you'll see the values are fairly similar:

SPSS says $z = -2.35$, with a 2-tailed $p = 0.018$, which rounds to $p = .02$.

The Mann-Whitney test:

As with the Wilcoxon test, the distribution of Mann-Whitney's U approximates a normal distribution if the sample size is large enough.

The **mean** of the distribution of U values is:

$$\mu = 0.5 * N1 * N2$$

where $N1$ is the number of participants in condition 1, and $N2$ is the number of participants in condition 2.

The **standard deviation** of the distribution is:

$$\sigma = \sqrt{\frac{(N1 * N2) * (N + 1)}{12}}$$

where $N1$ is the number of participants in condition 1, $N2$ is the number of participants in condition 2, and N is the total number of participants overall (i.e. $N1+N2$).

The s.d. formula gets more complicated if you have lots of ties. In that case, use the following formula.

For each tied value, t = the number of ties at that value: work out $(t^3 - t)$. Then find the total sum of the ties, $T_c = \sum (t^3 - t)$.

$$\sigma = \sqrt{\frac{(N1 * N2) * (N^3 - N - T_c)}{12N * (N - 1)}}$$

Calculate the critical value of U , for a given significance level:

$$U \text{ critical} = \mu - z * \sigma - 0.5$$

You plug into this formula whatever value of z corresponds to your desired significance level. Thus for a 0.05 significance level, use $z = 1.96$ for a two-tailed test and 1.64 for a one-tailed test. For a 0.01 significance level, $z = 2.58$ for a two-tailed test and 2.33 for a one-tailed test.

Your obtained U is significant if it is less than or equal to this $U \text{ critical}$.