The Wilcoxon Rank-Sum Test by Chris Wild, University of Auckland

The Wilcoxon rank-sum test is a nonparametric alternative to the two-sample t-test which is based solely on the order in which the observations from the two samples fall. We will use the following as a running example.

Example 1 In a genetic inheritance study discussed by Margolin [1988], samples of individuals from several ethnic groups were taken. Blood samples were collected from each individual and several variables measured. For a detailed discussion of the study and a definition of the variable, see Exercises 10.1.3 in the text. We shall compare the groups labeled "Native American" and "Caucasian" with respect to the variable MSCE (mean sister chromatid exchange). The data is as follows:

Native American:	8.50	9.48	8.65	8.16	8.83	7.76	8.63		
Caucasian:	8.27	8.20	8.25	8.14	9.00	8.10	7.20	8.32	7.70

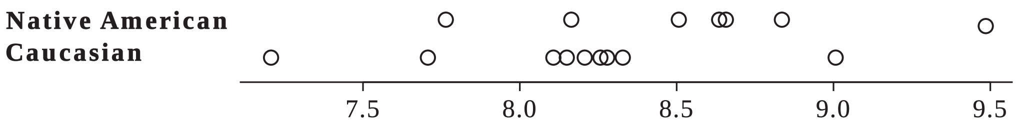


Figure 1 : Comparing MSCE measurements.

Looking at the dot plots for the two groups, several questions come to mind. Firstly, do the data come from Normal distributions? Unfortunately we can't say much about the distributions as the samples are too small. However there does not seem to be any clear lack of symmetry. Secondly, are the two distributions similar in shape? Again it is hard to say much with such small samples, though the Caucasian data seems to have longer tails. Finally, is there any difference in the centers of location? The plots suggest a difference with Native American values being larger on average. We shall now put this type of problem in a more general context and come back to this example later.

Suppose, more generally, that we have samples of observations from each of two populations A and B containing n_A and n_B observations respectively. We wish to test the hypothesis that the distribution of X-measurements in population A is the same as that in B, which we will write symbolically as $\boldsymbol{H_0 : A = B}$. The departures from H_0 that the Wilcoxon test tries to detect are location shifts. If we expect to detect that the distribution of A is shifted to the right of distribution B as in Fig. 2(b), we will write this as $\boldsymbol{H_1 : A > B}$. The other two possibilities are $H_1 : A < B$ (A is shifted to the left of B), and the two sided-alternative, which we will write as $H_1 : A \neq B$, for situations in which we have no strong prior reason for expecting a shift in a particular direction.

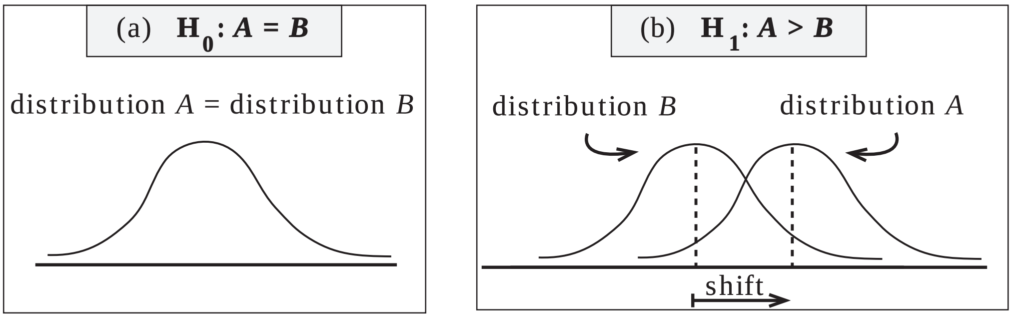


Figure 2 : Illustration of $H_0 : A = B$ versus $H_1 : A > B$.

The Wilcoxon test is based upon ranking the $n_A + n_B$ observations of the combined sample. Each observation has a ***rank***: the smallest has rank 1, the 2nd smallest rank 2, and so on. The Wilcoxon rank-sum test statistic is the sum of the ranks for observations from one of the samples. Let us use sample A here and use w_A to denote the observed rank sum and W_A to represent the corresponding random variable.

w_A = sum of the ranks for observations from *A*.

Example 1 cont. We have sorted the combined data set into ascending order and used vertical displacement as well as ethnic group labels to make very clear which sample an observation comes from ("NA" for the Native American group and "Ca" for the Caucasian group). The rank of an observation in the combined sample appears immediately below the label.

	7.20	7.70	7.76	8.10	8.14	8.16	8.20	8.25	8.27	8.32	8.50	8.63	8.65	8.83	9.00	9.48
Race	Ca	Ca	NA	Ca	Ca	NA	Ca	Ca	Ca	Ca	NA	NA	NA	NA	Ca	NA
Rank:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

The sum of the ranks for the Native American group is

$$w_{NA} = 3 + 6 + 11 + 12 + 13 + 14 + 16 = 75. \qquad \blacksquare$$

How do we obtain the *P-value* corresponding to the rank-sum test statistic w_A? To answer this question we must first consider how rank sums behave under H_0, and how they behave under H_1. Fig. 3 depicts two situations using samples of size $n_A = n_B = 5$ and plotting sample A observations with a "•" and sample B observations with an "o".

Suppose that $H_0 : A = B$ is true. In this case, all $n = n_A + n_B$ observations are being drawn from the same distribution and we might expect behavior somewhat like Fig. 3(a) in which the pattern of black and white circles is random. The set of ranks for n observations are the numbers $1, 2, \ldots, n$.

When n_A of our n observations from a distribution are labeled A and n_B observations from the same distribution are labeled B, then as far as the behavior of the ranks (and thus w_A) is concerned, it is just as if we randomly labeled n_A of the numbers $1, 2, \ldots, n$ with A's and the rest with B's. The distribution of a rank sum, W_A, under such conditions has been worked out and computer programs and sets of Tables are available for this distribution.

	●	o	o	●	o	●	●	o	●	o		o	o	o	o	●	o	●	●	●	●
Rank	1	2	3	4	5	6	7	8	9	10		1	2	3	4	5	6	7	8	9	10
					(a)											(b)					

Figure 3 : Behaviour of ranks.

Suppose that $H_1 : A > B$ is true: In this case we would expect behavior more like that in Fig. 3(b) which results in sample A containing more of the larger ranks. Evidence against H_0 which confirms $H_1 : A > B$ is thus provided by an observed rank sum w_A which is unusually large according to the distribution of rank sums when H_0 is true. Thus the *P-value* for the test is

$$(H_1 : A > B) \qquad\qquad P\text{-}value\ = \mathrm{pr}(W_A \geq w_A),$$

where the probability is calculated using the distribution that W_A would have if H_0 was true. Suppose, on the other hand, that the alternative $H_1 : A < B$ is true. In this case we would expect the A observations to tend to be smaller than the B observations, resulting in a small rank sum w_A. The *P-value* for the alternative $H_1 : A < B$ is therefore

$$(H_1 : A < B) \qquad\qquad P\text{-}value\ = \mathrm{pr}(W_A \leq w_A).$$

Note that in testing one-sided alternatives, the direction of the inequality used in the calculation of the *P-value* is the same as the direction defining the alternative, e.g. $A > B$ and $W_A \geq w_A$.

For the two-sided test, i.e. testing $H_0 : A = B$ versus the alternative $H_1 : A \neq B$, a rank sum that is either too big or too small provides evidence against H_0. We then calculate the probability of falling into the tail of the distribution closest to w_A and double it. Thus if w_A is in the lower tail then *P-value* $= 2\ \mathrm{pr}(W_A \leq w_A)$, whereas if w_A is in the upper tail then *P-value* $= 2\ \mathrm{pr}(W_A \geq w_A)$.

Example 1 cont. Here, we want to test a null hypothesis H_0 which says that the MSCE distribution for Native Americans is the same as that for Caucasians. Although the Native American MSCE values in the data tend to be higher, there was no prior theory to lead us to expect this so we should be doing a two-sided test. The rank sum for for the Native American group was $w_{NA} = 75$. We know from the plot of the data that this will be in the upper tail of the distribution. The *P-value* is thus

$$P\text{-}value = 2\ \mathrm{pr}(W_{NA} \geq 75) = 0.114 \text{ (computer)}.$$

The evidence against H_0 which suggests that median MSCE measurements are higher for Native Americans than for Caucasians is, at best, weak. In fact we can't be sure that this evidence points to a difference in the shapes of the two distributions rather than a difference in the centers of location. [We note that, for this data set, a two-sample t-test (Welch) for no difference in means gives almost exactly the same *P-value* .] ∎

Treatment of ties

Consider the data to follow. It has two observations tied with value 4, three more tied with value 6 and a set of four observations tied with value 11.

Data:	0	1	2	2	5	6	6	6	7	8	10	11	11	11	11	14	....
From:	A	B	A	B	A	A	B	B	B	A	B	A	A	B	B	A	......
1 to n:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	....
Ranks:	1	2	3.5	3.5	5	7	7	7	9	10	11	13.5	13.5	13.5	13.5	16	....

We listed the data in ascending order together with the labels telling us which sample each data point came from. Next we wrote down the numbers 1 to n as we have done before to construct ranks. Both sample A and sample B have an observation with value 2. In the initial allocation of ranks, the sample A observation has arbitrarily been assigned a 3 and the sample B observation a 4. Since we cannot distinguish order between these two observations, they should be treated in the same way. What values should we use as the ranks to construct a rank sum when we have ties?

We assign each observation in a tie its average rank.

The rule in the box tells us to assign both observations tied values[1] at value 2 the average of the ranks from the first pass through the data, namely 3.5, the average of 3 and 4. Similarly, each of the three observations with value 6 should be assigned rank 7 (the average of 7, 8, and 9), whereas the four observations tied with value 11 should be given rank 13.5 (the average of $12, 13, 14$ and 15).

[1]Strictly, the distribution of W_A should also be modified when ties are present (see Bhattacharyya and Johnson [1977, page 515]).

Example 2 We will also use the Wilcoxon rank-sum test to compare the Native American (NA) and Caucasian (Ca) groups with respect to another variable called DISPERSION (see Table 10.3.3 in the text for the original data set). The ordered data, ranks etc. are as follows:

NA:			0.45	0.50						0.61	0.63		0.75		0.85	0.93
Ca:	0.44	0.45			0.52	0.53	0.56	0.58	0.58			0.65		0.79		
Race:	Ca	Ca	NA	NA	Ca	Ca	Ca	Ca	Ca	NA	NA	Ca	NA	Ca	NA	NA
Rank:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
		2.5	2.5					8.5	8.5							

The sum of the ranks for the Native American group is

$$w_{NA} = 2.5 + 4 + 10 + 11 + 13 + 15 + 16 = 71.5.$$

Our null hypothesis H_0 says that the distribution of DISPERSION measurements is the same for both ethnic groups. Since we have no prior theory, we will perform a 2-sided test. There is a slight, but noticeable, tendency for Native American values to be higher, so that this rank sum will be in the upper tail of the distribution. The *P-value* is thus

$$P\text{-}value = 2\ \mathrm{pr}(W_{NA} \geq 71.5) > 2\ \mathrm{pr}(W_{NA} \geq 72) = 0.210 \text{ (computer)}.$$

Thus, the Wilcoxon test gives no evidence against the null hypothesis of identical distributions for DISPERSION. [We note that a two-sided two-sample t-test of $H_0 : \mu_{NA} = \mu_{Ca}$ has *P-value* 0.19 (Welch) , 0.15 (pooled) showing no evidence of a difference in the means.]

P-values from Tables

When one performs a Wilcoxon test by hand, Tables are required to find *P-values* . Readers who are not interested in this level of detail should proceed to the Notes at the end of the subsection. For small sample sizes, tables for Wilcoxon rank-sum test are given. We supplement this with a Normal approximation for use with larger samples. All probabilities discussed relate to the distribution of W_A when H_0 is true.

Small sample Tables

Tables for the Wilcoxon rank-sum test are given in the Appendix at the end of this module. A segment is printed as Table 1. When the two samples have different sizes, the tables are set up for use with the rank sum for the smaller of the two samples, so that we define sample A to be the smaller of the two samples. One chooses the row of the table corresponding to the combination of sample sizes, n_A and n_B, that one has.

For given n_A, n_B and *prob*, the tabulated value for the **lower** *prob*-tail is the largest value of w_A for which[2] $\text{pr}(W_A \le w_A) \le prob$. For example, when $n_A = 7$ and $n_B = 9$, the tabulated value for $prob = 0.2$ is $w_A = 50$. Thus,

$$\text{pr}(W_A \le 50) \le 0.2, \text{ but } \text{pr}(W_A \le 51) > 0.2.$$

In other words, the values in the lower 0.2 (or 20%) tail are those ≤ 50.

The tabulated value for the **upper** *prob*-tail is the smallest value of w_A for which $\text{pr}(W_A \ge w_A) \le prob$. When $n_A = 7$ and $n_B = 9$, the tabulated value for the upper tail and $prob = 0.2$ is $w_A = 69$. Thus,

$$\text{pr}(W_A \ge 69) \le 0.2 \text{ but } \text{pr}(W_A \ge 68) > 0.2.$$

In other words, the values in the upper 0.2 (or 20%) tail are those ≥ 69.[3]

Table 1 : Segment of Wilcoxon Rank-Sum Table

		Lower Tail						Upper Tail					
		prob						***prob***					
n_A	n_B	.005	.01	.025	.05	.10	.20	.20	.10	.05	.025	.01	.005
7	7	32	34	36	39	41	45	60	64	66	69	71	73
	8	34	35	38	41	44	48	64	68	71	74	77	78
	9	35	37	40	43	46	50	69	73	76	79	82	84
	10	37	39	42	45	49	53	73	77	81	84	87	89
	11	38	40	44	47	51	56	77	82	86	89	93	95
	12	40	42	46	49	54	59	81	86	91	94	98	100

These tables can be used to bracket tail probabilities in the same way as the tables for Student's t-distribution (described in Section 7.6.3 in the text). For example, suppose that $n_A = 7$ and $n_B = 9$ and $w_A = 71$. We see that 71, being bigger than 69, is in the upper 20% tail but is not large enough to be in the upper 10% tail. Thus $\text{pr}(W_A \ge 71)$ is between 0.1 and 0.2. Similarly, $\text{pr}(W_A \ge 78)$ is between 0.025 and 0.05, and $\text{pr}(W_A \ge 88) \le 0.005$.

Using $n_A = n_B = 7$ (and the lower tail of the distribution) we see that $\text{pr}(W_A \le 38)$ is between 0.025 and 0.05, and $\text{pr}(W_A \le 47)$ is bigger than 0.20.

When $n_A = 7, n_B = 10$, $\text{pr}(W_A \le 35)$ is smaller than 0.005, $\text{pr}(W_A \le 41)$ is between 0.01 and 0.025, $\text{pr}(W_A \le 50)$ is greater than 0.2, and $\text{pr}(W_A \ge 82)$ is between 0.025 and 0.05.

[2]The more complicated description of the entries of the Wilcoxon tables stems from the fact that the distribution of rank sums is discrete. For the Student(df) distribution, and given *prob*, one can find a value of t so that $\text{pr}(T \ge t) = prob$, exactly. This cannot be done in general for discrete distributions.

[3]Similarly, the lower 10% tail consists of values ≤ 46, the upper 10% tail consists of values ≥ 73, and the upper 5% tail consists of values ≥ 76.

Remember to double the probabilities for a two-sided test.

Normal approximation for larger samples

Our Wilcoxon Tables cater for sample sizes up to $n_A = n_B = 12$. When both sample sizes are 10 or greater, we can treat the distribution of W_A as if it were Normal(μ_A, σ_A), where

$$\mu_A = \frac{n_A(n_A + n_B + 1)}{2} \quad \text{and} \quad \sigma_A = \sqrt{\frac{n_A n_B(n_A + n_B + 1)}{12}}.$$

More precisely,

$$\text{pr}(W_A \geq w_A) \approx \text{pr}(Z \geq z), \quad \text{where } z = \frac{w_A - \mu_A}{\sigma_A}$$

and $Z \sim \text{Normal}(0, 1)$. For example, suppose that $n_A = 10, n_B = 12$, and we want $\text{pr}(W_A \geq 145)$. Then, $\mu_A = 10 \times (10 + 12 + 1)/2 = 115$ and

$$\sigma_A = \sqrt{\frac{10 \times 12 \times (10 + 12 + 1)}{12}} = 15.16575$$

so that

$$\text{pr}(W_A \geq 145) \approx \text{pr}\left(Z \geq \frac{145 - 115}{15.16575}\right) = \text{pr}(Z \geq 1.978) = 0.024.$$

Notes

1. The Wilcoxon test is still valid for data from any distribution, whether Normal or not, and is much less sensitive to outliers than the two-sample t-test.

2. If one is primarily interested in differences in location between the two distributions, the Wilcoxon test has the disadvantage of also reacting to other differences between the distributions such as differences in shape.

3. When the assumptions of the two-sample t-test hold, the Wilcoxon test is somewhat less likely to detect a location shift than is the two-sample t-test. However, the losses in this regard are usually quite small.[4]

4. ***Nonparametric confidence intervals*** for $\theta = \widetilde{\mu}_A - \widetilde{\mu}_B$, the difference between the two population medians (or any other measure of location), can be obtained by inverting the Wilcoxon test, provided one is willing to assume that the two distributions differ only by a location shift.[5] In Example 10.2.5, a 95% confidence interval for the true difference in median MSCE between Native Americans and Caucasians is given by $[-0.2, 1.1]$.

[4] See Bhattacharyya and Johnson [1977, page 538].
[5] See Bhattacharyya and Johnson [1977, page 525].

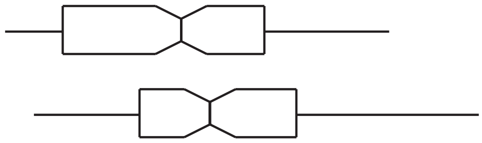
A 95% confidence interval (Welch) for the difference in means is given by $[-0.1, 1.0]$.

5. In a practical situation in which we are uneasy about the applicability of two-sample t methods, we use both them and the Wilcoxon and feel happiest when both lead to very similar conclusions.

6. The ***Mann-Whitney test*** is essentially identical to the Wilcoxon test, even though it uses a different test statistic.

7. ***Notched box plots.*** Many computer programs give the option of putting notches in box plots as shown in Fig. 10.2.10. If the sloping parts of the notches overlap there is no significant difference between the medians at the 5% level. If they do not overlap there is a significant difference.[6]

Figure 4 : Notched box plots.



8. Just as the Wilcoxon test is a nonparametric alternative to the two-sample t-test, the ***Kruskal-Wallis test*** is a nonparametric alternative to the one-way analysis of variance F-test.[7]

Quiz for Section 10.1

1. What test statistic is used by the Wilcoxon rank-sum test?
2. Verbally, how is that test statistic obtained?
3. What null hypothesis does it test? What are the possible alternative hypotheses?
4. What assumptions are made by the test?
5. Qualitatively, what should happen to the rank sum for sample A if distribution A is shifted to the right of distribution B? if it is shifted to the left?
6. Which test is less sensitive to outliers, the Wilcoxon test or the two-sample t-test? Which test is least sensitive to non-Normality? Why?

Exercises for Section 10.1

Consider the data described in Exercises 10.3 and given in Table 10.3.3 in the text. Use the Wilcoxon test to determine whether there is any evidence of a difference in the median DISPERSION between:

(a) the Asian group and the Native American group;

(b) the Native American group and the Black group. Make the comparison (b) in two ways, **(i)** with the outlier in the Black group included in the analysis and **(ii)** with it omitted from the analysis.

---

[6]We have also seen shading of part of the box being used instead of cutting notches to accomplish the same purpose. Here significant differences between groups correspond to no overlap of the shaded regions.

[7]See Bhattacharyya and Johnson [1977, page 533].

Appendix : Wilcoxon Rank-Sum Table

Probabilities relate to the distribution of W_A, the rank sum for group A when $H_0 : A = B$ is true. The tabulated value for the **lower tail** is the largest value of w_A for which $\text{pr}(W_A \leq w_A) \leq prob$. The tabulated value for the **upper tail** is the smallest value of w_A for which $\text{pr}(W_A \geq w_A) \leq prob$.

		Lower Tail *prob*						**Upper Tail** *prob*					
n_A	n_B	.005	.01	.025	.05	.10	.20	.20	.10	.05	.025	.01	.005
4	4			10	11	13	14	22	23	25	26		
	5		10	11	12	14	15	25	26	28	29	30	
	6	10	11	12	13	15	17	27	29	31	32	33	34
	7	10	11	13	14	16	18	30	32	34	35	37	38
	8	11	12	14	15	17	20	32	35	37	38	40	41
	9	11	13	14	16	19	21	35	37	40	42	43	45
	10	12	13	15	17	20	23	37	40	43	45	47	48
	11	12	14	16	18	21	24	40	43	46	48	50	52
	12	13	15	17	19	22	26	42	46	49	51	53	55
5	5	15	16	17	19	20	22	33	35	36	38	39	40
	6	16	17	18	20	22	24	36	38	40	42	43	44
	7	16	18	20	21	23	26	39	42	44	45	47	49
	8	17	19	21	23	25	28	42	45	47	49	51	53
	9	18	20	22	24	27	30	45	48	51	53	55	57
	10	19	21	23	26	28	32	48	52	54	57	59	61
	11	20	22	24	27	30	34	51	55	58	61	63	65
	12	21	23	26	28	32	36	54	58	62	64	67	69
6	6	23	24	26	28	30	33	45	48	50	52	54	55
	7	24	25	27	29	32	35	49	52	55	57	59	60
	8	25	27	29	31	34	37	53	56	59	61	63	65
	9	26	28	31	33	36	40	56	60	63	65	68	70
	10	27	29	32	35	38	42	60	64	67	70	73	75
	11	28	30	34	37	40	44	64	68	71	74	78	80
	12	30	32	35	38	42	47	67	72	76	79	82	84
7	7	32	34	36	39	41	45	60	64	66	69	71	73
	8	34	35	38	41	44	48	64	68	71	74	77	78
	9	35	37	40	43	46	50	69	73	76	79	82	84
	10	37	39	42	45	49	53	73	77	81	84	87	89
	11	38	40	44	47	51	56	77	82	86	89	93	95
	12	40	42	46	49	54	59	81	86	91	94	98	100
8	8	43	45	49	51	55	59	77	81	85	87	91	93
	9	45	47	51	54	58	62	82	86	90	93	97	99
	10	47	49	53	56	60	65	87	92	96	99	103	105
	11	49	51	55	59	63	69	91	97	101	105	109	111
	12	51	53	58	62	66	72	96	102	106	110	115	117
9	9	56	59	62	66	70	75	96	101	105	109	112	115
	10	58	61	65	69	73	78	102	107	111	115	119	122
	11	61	63	68	72	76	82	107	113	117	121	126	128
	12	63	66	71	75	80	86	112	118	123	127	132	135
10	10	71	74	78	82	87	93	117	123	128	132	136	139
	11	73	77	81	86	91	97	123	129	134	139	143	147
	12	76	79	84	89	94	101	129	136	141	146	151	154
11	11	87	91	96	100	106	112	141	147	153	157	162	166
	12	90	94	99	104	110	117	147	154	160	165	170	174
12	12	105	109	115	120	127	134	166	173	180	185	191	195

by Chris Wild
Department of Statistics
University of Auckland
1997