Data Analysis and Statistical Methods
Statistics 651

http://www.stat.tamu.edu/~suhasini/teaching.html

Lecture 20 (MWF) The Wilcoxon Sum Rank test (Mann-Whitney test)

Suhasini Subba Rao

Example

Roopa is conducting research on influence that diet has on absorption of iron gained (in particular a Vitamin C rich diet and a Calcium rich diet) She randomly allocated 20 volunteers into two groups (ten people in each group). Group 1 she put on a high vitamin C diet, the amount of iron gained or lost is given in the first row below (if the number is positive this is a gain, if the number is negative it is a loss). Group 0 she put on a high calcium diet, the iron gained or lost is given in the second row of the table below.

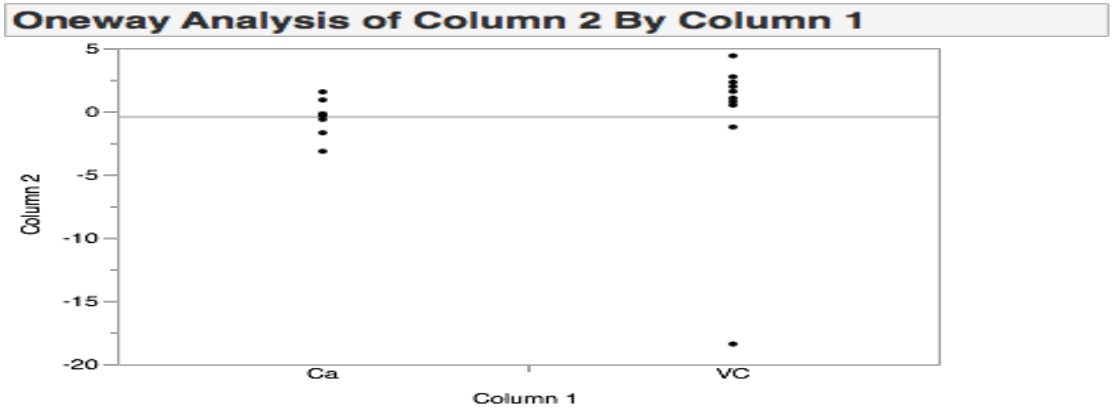
group										
VitC	0.51	2.75	0.79	4.41	-1.23	1.06	1.98	2.32	1.59	-18.41
Calc	-0.18	0.92	-0.25	1.56	-0.38	-0.21	-0.62	-1.68	-3.15	-0.33

`http://www.stat.tamu.edu/~suhasini/teaching651/irons_levels_calcium.dat`

She wants to investigate whether the mean absorption of iron of people on a high Vitamin C is more than the absorption of those on high calcium.

Solution

Roopa's research hypothesis is that the average amount of iron absorbed is higher for vitamin rich diets. Let μ_{VitC} denote the mean amount of iron gained on the vitamin C diet and μ_{Calc} denote the mean amount of iron gained on the high calcium diet. Roopa's null and alternative hypotheses are $H_0 : \mu_{VitC} - \mu_{Calc} \leq 0$ against $H_A : \mu_{VitC} - \mu_{Calc} > 0$.



Result of Roopa's t-test

- From the output using the one-sided test we see that the p-value is 48.65%, this p-value is a lot larger than the 5% significance level. hence we cannot reject the null. Why the negative result?
  - May be a small sample size to detect a difference even if there is one?
  - There is no difference between absorption of iron when a person takes either calcium of vitamin C.

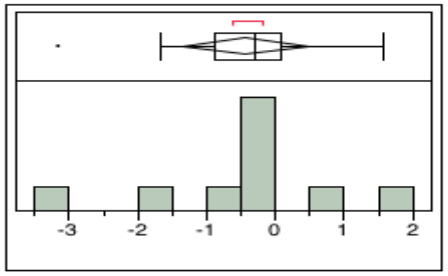
- Aside Negative results can also be of academic interest see the interesting article

  w.economist.com/news/leaders/21588069-scientific-research-has-changed-world-now-it-needs-change-itself-how-science-goes-wrong, one aspect of this argument is that negative results need to be published.

Histogram of iron data

Distributions Column 1=0

Column 2

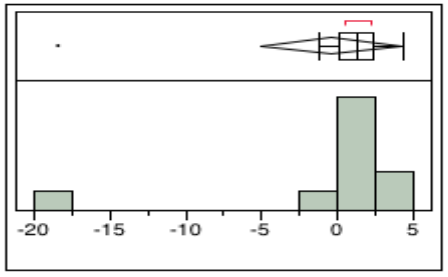


Summary Statistics

Mean	-0.432
Std Dev	1.2882098
Std Err Mean	0.4073677
Upper 95% Mean	0.4895298
Lower 95% Mean	-1.35353
N	10

Distributions Column 1=1

Column 2



Summary Statistics

Mean	-0.423
Std Dev	6.4935952
Std Err Mean	2.0534551
Upper 95% Mean	4.2222382
Lower 95% Mean	-5.068238
N	10

Observe the large outlier in the second plot.

- Besides the sample size, there can be two reasons that we may not be unable to reject the null, even if the alternative is true. Both these reasons are related to the outlier -18.41 we see in the plot.
  - This outlier makes a difference. It pulls down the sample mean for the iron level in Diet 0.
  - The large outlier also makes the sample standard deviation very large.

- Both these factors contribute to our being less likely to reject the null (by making the difference in the sample means smaller and the non-rejection regions wider).

- The above data set illustrates why we may, in certain cases, want to use a test which is more robust to outliers. The outlier carries too much weight, and is too influential.

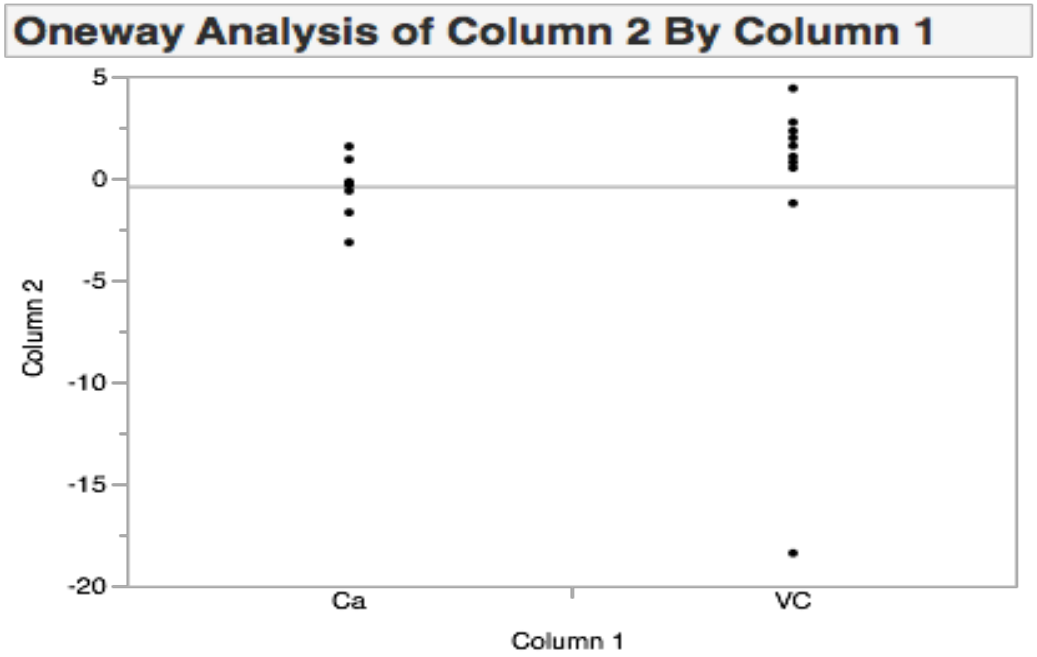
- In this example, the outlier was was making the sample means closer.
- In other examples, one outlier can make the difference appear greater and give the appearance of a difference.
- When outliers exist do not remove them!
- Instead we use tests that gives less weight to individual outliers.

  One solution is to use ranks rather than raw observations.

When it is best not to use the independent t-test

- In the case that n or m is small and has several outliers or appears skewed (non-symmetric) the t-test may give unreliable results:
  - Outliers can wrongly shift the sample mean to high or low, making a comparison difficult, they can also make the standard errors larger than what they should be.
  - If the observations come from a distribution which is skewed, and the sample size is not sufficiently large. Then the central limit theorem may not hold, in which case the t-test is inappropropiate.
  - In a nutshell, the assumption that the difference in the sample means is normally distributed may not hold (the original data set is too non-normal for the central limit theorem to have kicked in).

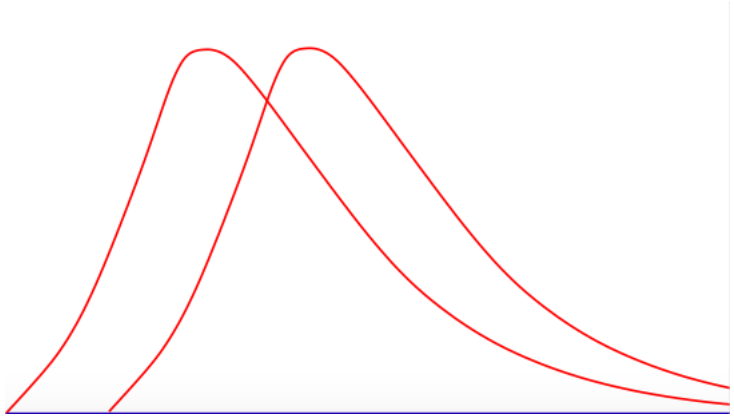
A second look at the iron data



There appears to be a 'separation' in the two data, that could be difficult to explain by random chance (ie, if both data sets came from the same distribution how likely can be get a sample that looks this separated?). How to quantify this separation?

The Wilcoxon Rank sum test/ Mann-Whitney U statistic

In the hypothesis test we do not assume normality, but we do require the two distributions are the same except for a possible shift. In other words both distributions have the same "shape".



- We do not characterise the hypothesis test in terms of the mean. Instead we do the characterisation in terms of the 'location' of the distribution.
  - H_0: The populations have identical distributions.
  - H_A: One population is a shift of the other (as in the plot above).

- The independence assumption still holds.

- The test is often called 'distribution free' (or nonparametric) meaning that it does not require any assumptions on the distribution of the observations.

Ranking Data

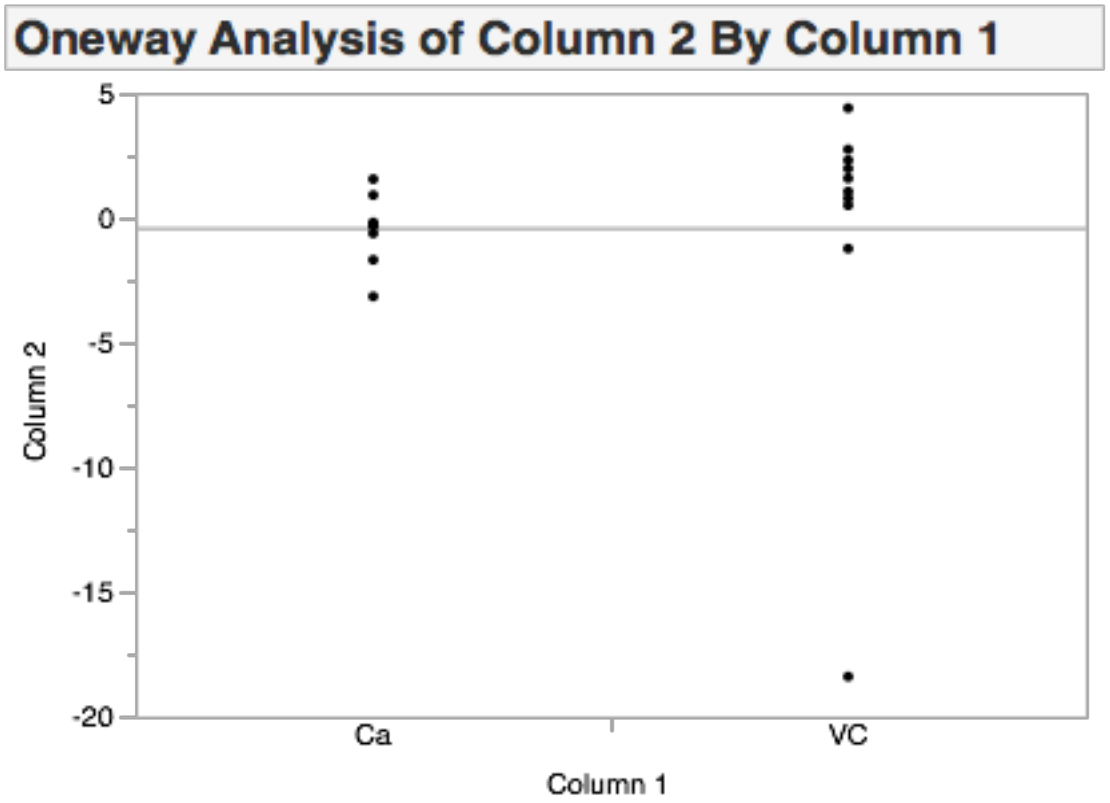
- The test is not sensitive to outliers and just as the median is based on ranking the data, so to is the Wilcoxon sum rank test.

- Below is an example of a data set and its corresponding rank

Data	0.37	0.38	0.83	0.95	1.78	2.86	6.61
Rank	1	2	3	4	5	6	7

- The advantage of ranks over raw data is that the sample is uneffected by extremely small of large values. For example if 6.61 were changed to 10.8, the rank of 7 remains.

Back to the Iron vs Vitamin C data set



The Wilcoxon test is based on **collectively** ranking the data.

- We count the total ranks in each group. And we use the "rank sums" as the basis of the test.

- Effectively, if both groups are drawn from the same population and the sample sizes in both groups are the same we would expect the rank sums to be about the same too.

- As it is difficult to clearly rank the above data set. We start with a simpler example. And return to this data set later.

The Wilcoxon Rank sum test: example

In a study, 19 mold sensitive volunteers were exposed to mold. The objective of the study was to understand the influence antihistamines had on the allergic reaction to the mold. The 19 volunteers were placed into two treament groups, one of size 10 the other of size 9. Group 1 (size 10) was given the antihistamine, Group 2 (size 9) was given the placebo. The size of the allergic reaction is given below.

Antihistimine	0.90	0.37	1.63	0.15	0.95	0.78	0.05	0.61	0.51	0.20
Placebo	1.60	1.50	1.76	1.44	1.11	3.07	1.05	1.27	2.56	

- Data:

  `http://www.stat.tamu.edu/~suhasini/teaching651/antihistamine_placebo.dat`.

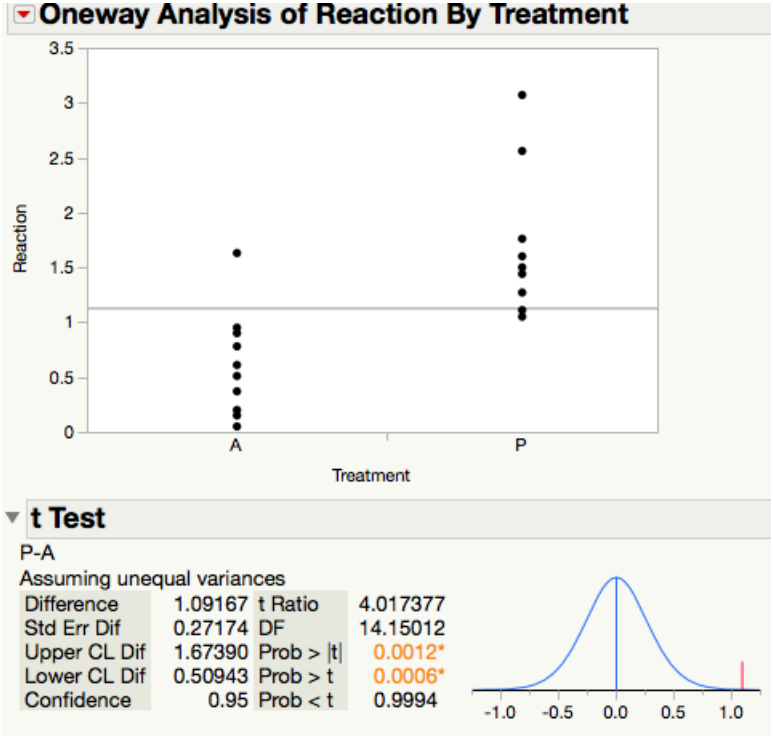
- Our aim is to see whether in general people with dust allergies have

smaller reaction when taking an antihistamine than those on a control (placebo). Our hypotheses are:

H_0: distribution of both the antihistamine and placebo populations are the same or antihistamine is a right shift of the placebo (view this as $H_0 : \mu_A \geq \mu_P$).

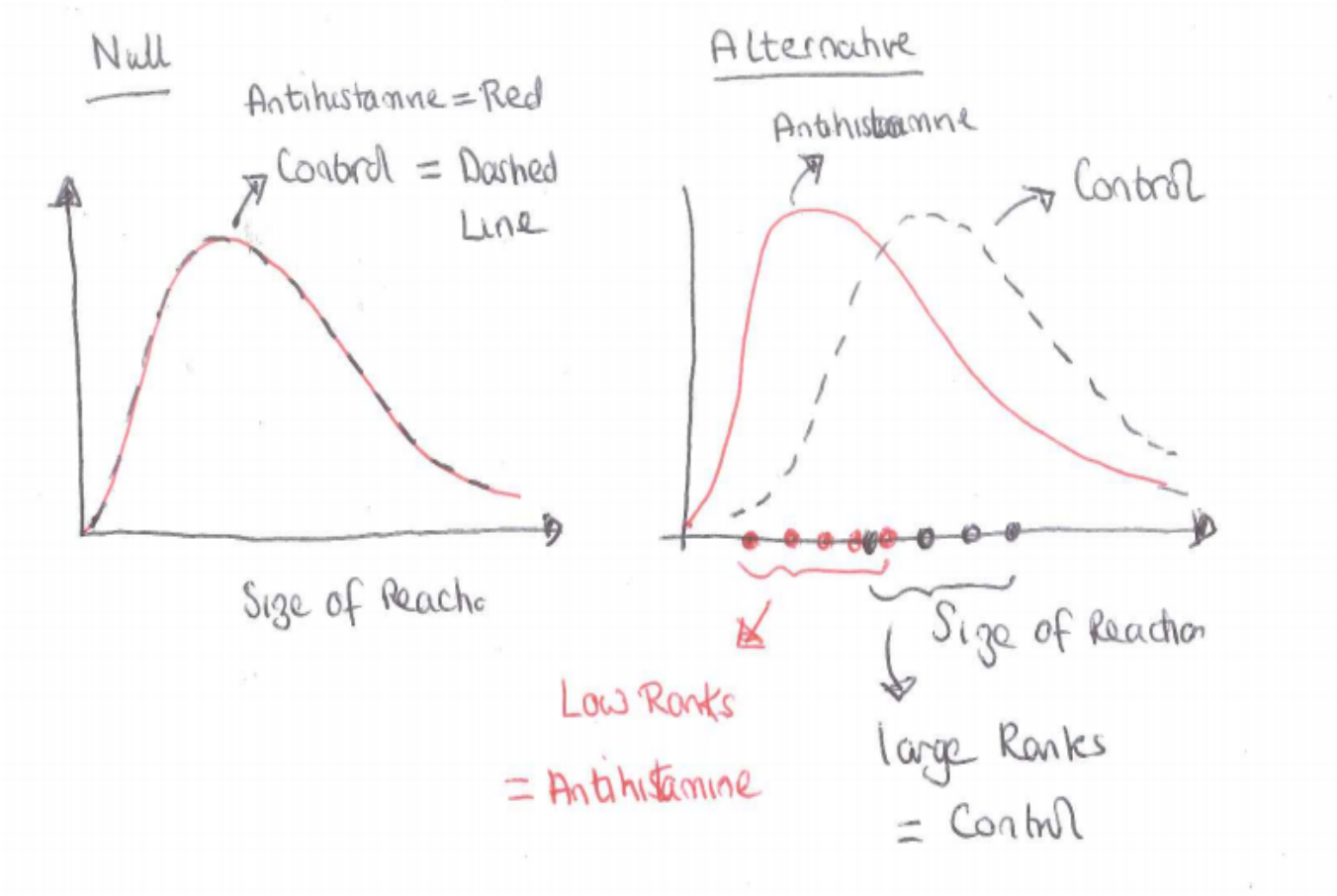
H_A: distribution of the antihistamine is a left shift of the placebo (view this as $H_A : \mu_A < \mu_P$).

Independent two-sample t-test

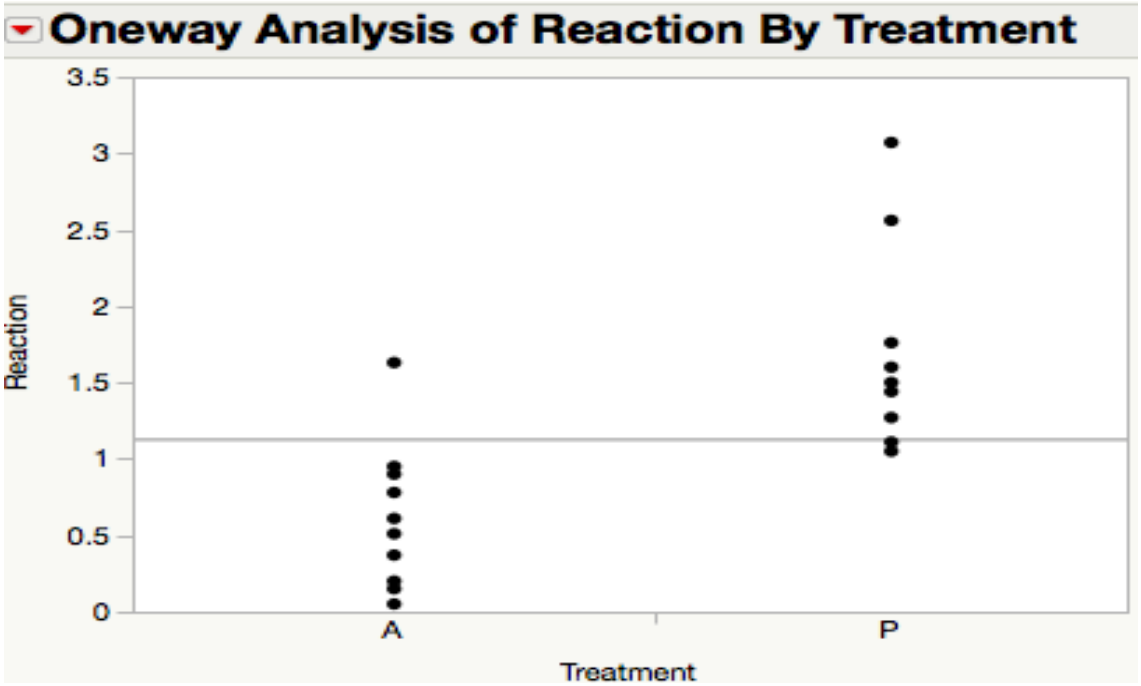


The hypothesis that is of interest to us is $H_0 : \mu_P - \mu_A \leq 0$ vs $H_A : \mu_P - \mu_A > 0$. The p-value is 0.6% and we reject the null at the 5% and 1% level. We now apply the Wilcoxon test to this data set.

Distribution under null and alternative



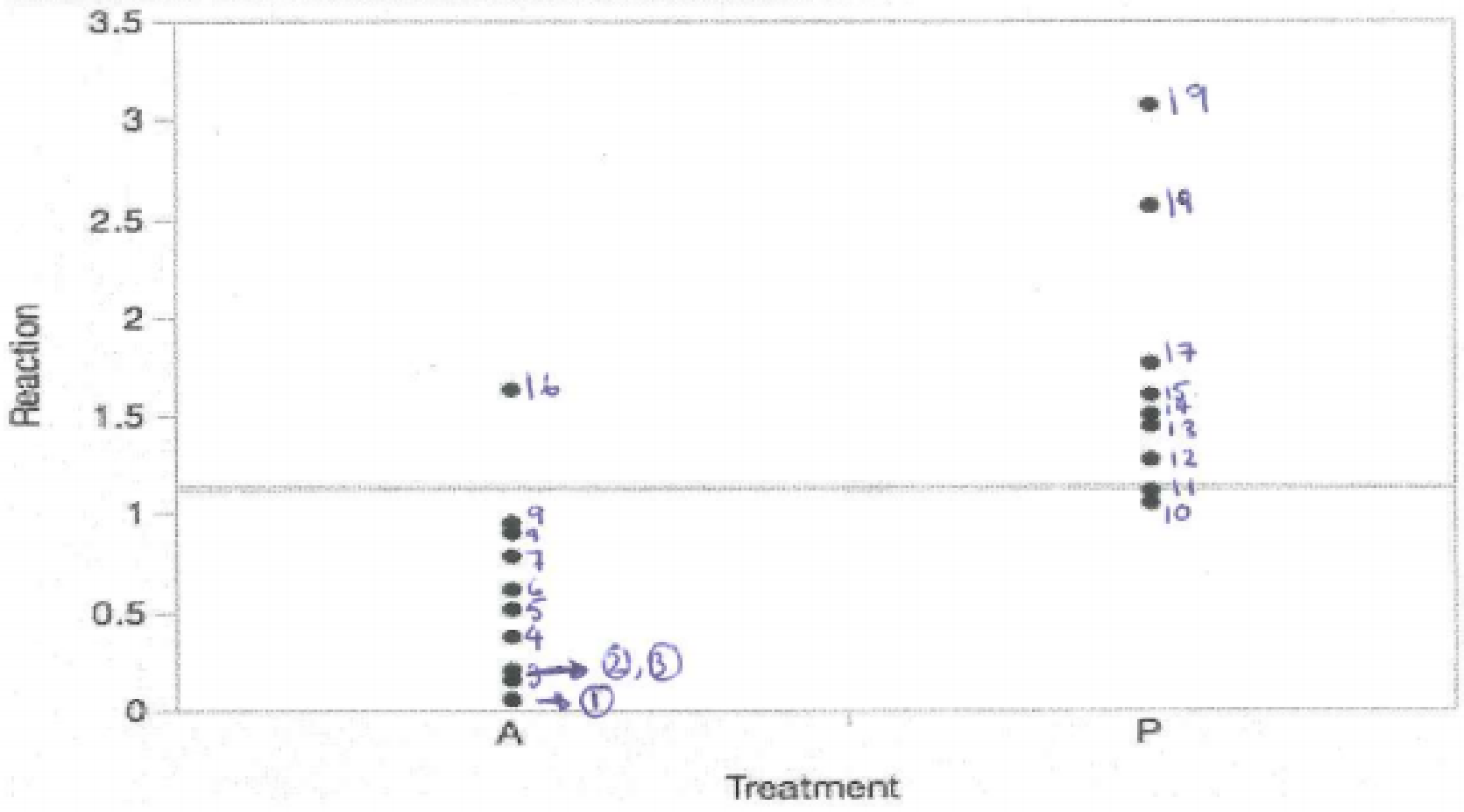
Dotplot of the data



Data plotted side by side. Visually there appears to be a right shift of the placebo group. Is this for real (something systematic in the population)? How likely is such a shift in data when really both the placebo and the antihistamine data came from the **same** distribution?

Basic idea of a Wilcoxon Rank sum test

- Order both samples separately from smallest to largest.
- In separate columns rank them collectively (both samples together) (from 1 the smallest to N the largest).
- If there is a tie in the numbers, add the two ranks together and divide by two.
- Identify the smallest sample size in the two groups. Add the ranks in this group. We denote this sum as T.


3.5
3
2.5
2
1.5
1
0.5
0
Reaction
A
P
Treatment

Example

Antihistamine	Rank	Control/Placebo	Rank
0.05	1	1.05	10
0.15	2	1.11	11
0.20	3	1.27	12
0.37	4	1.44	13
0.51	5	1.50	14
0.61	6	1.60	15
0.78	7	1.76	17
0.90	8	2.56	18
0.95	9	3.07	19
1.63	16		
T	61		129

Interpreting T

- The total sum of all the ranks is $19 \times 20/2 = 190$. This is a non-random quantity that only depends on the sample size.

- These following quantities are important.

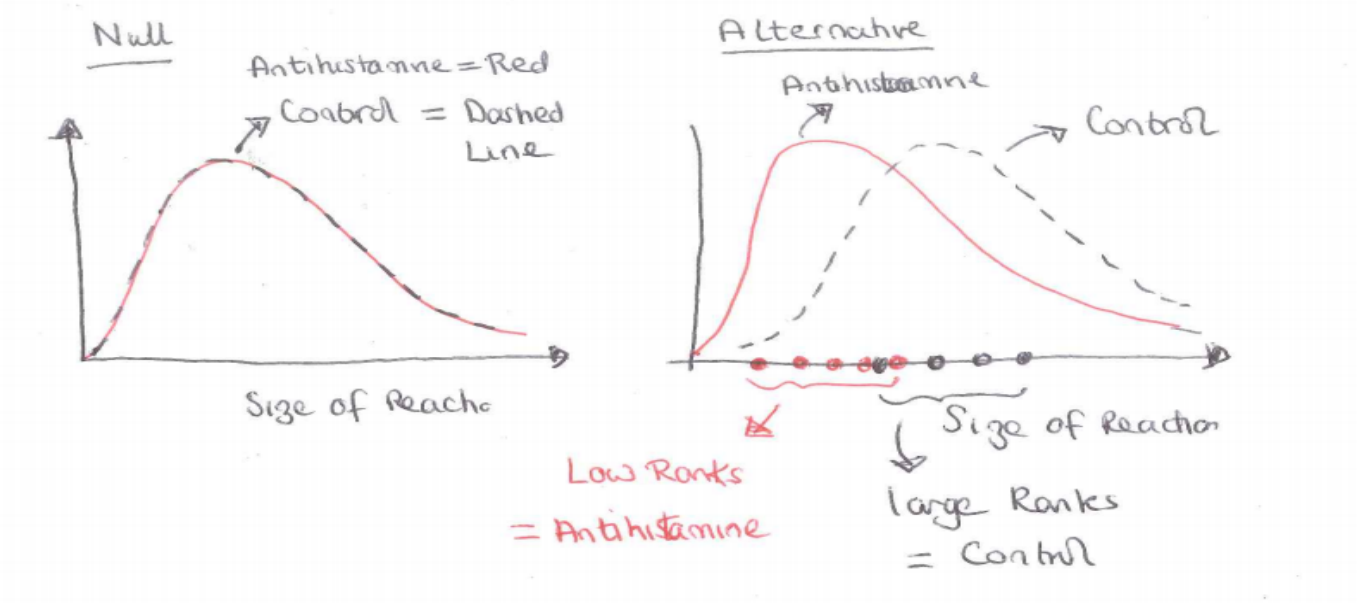
  The ranks corresponding to the smallest sample size is $T = 129$. Convention means we always focus on the rank corresponding to the smallest sample size.

  The ranks corresponding to the larger sample size is $T = 61$.

- The sum of ranks from the smallest sample size is $T = 129$ is random and depends on the the outcomes from the two groups.

  Compare $T = 129$ with the sum of the larger sample size 61. 129 is substantially larger than 61.

- Since T belongs the control/placebo group, if the alternative were true, we would expect the ranks in the placebo group to be large. Indeed the T in our sample does "appear" to be large. But is it large enough?



We need to use the critical region from a properly defined test to determine if T is enough for the null to be deemed implausible.

TABLE 5

Critical values of T_L and T_U for the Wilcoxon rank sum test: independent samples. Test statistic is rank sum associated with smaller sample (if equal sample sizes, either rank sum can be used).

a. $\alpha = .025$ one-tailed; $\alpha = .05$ two-tailed

n_2 \ n_1	3		4		5		6		7		8		9		10	
	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U
3	5	16	6	18	6	21	7	23	7	26	8	28	8	31	9	33
4	6	18	11	25	12	28	12	32	13	35	14	38	15	41	16	44
5	6	21	12	28	18	37	19	41	20	45	21	49	22	53	24	56
6	7	23	12	32	19	41	26	52	28	56	29	61	31	65	32	70
7	7	26	13	35	20	45	28	56	37	68	39	73	41	78	43	83
8	8	28	14	38	21	49	29	61	39	73	49	87	51	93	54	98
9	8	31	15	41	22	53	31	65	41	78	51	93	63	108	66	114
10	9	33	16	44	24	56	32	70	43	83	54	98	66	114	79	131

b. $\alpha = .05$ one-tailed; $\alpha = .10$ two-tailed

n_2 \ n_1	3		4		5		6		7		8		9		10	
	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U
3	6	15	7	17	7	20	8	22	9	24	9	27	10	29	11	31
4	7	17	12	24	13	27	14	30	15	33	16	36	17	39	18	42
5	7	20	13	27	19	36	20	40	22	43	24	46	25	50	26	54
6	8	22	14	30	20	40	28	50	30	54	32	58	33	63	35	67
7	9	24	15	33	22	43	30	54	39	66	41	71	43	76	46	80
8	9	27	16	36	24	46	32	58	41	71	52	84	54	90	57	95
9	10	29	17	39	25	50	33	63	43	76	54	90	66	105	69	111
10	11	31	18	42	26	54	35	67	46	80	57	95	69	111	83	127

Source: From F. Wilcoxon and R. A. Wilcox, *Some Rapid Approximate Statistical Procedures* (Pearl River, N.Y.

Critical region: One-sided test 5% pointing RIGHT

- Use $\alpha = 0.05$ and one-sided. Too prove the alternative at the 5% level we have to determine if T (which corresponds to the placebo group) is too large for the null to be plausible.

- Choose $n_1 = 9$ and $n_2 = 10$ (it does not matter which way round you choose 9 and 10) and find where they cross. Since we are conducting a one-sided test to determine if T is too large, use the largest value in $[69, 111]$ as the critical point.

- The critical region for the one-sided test is any value greater than 111.

- $T = 129 > 111$, therefore we reject the null and determine that using a antihistamine reduces the area of reaction (at the 5% level).

Critical region: One-sided test 5% pointing LEFT

- If there is reason to believe that Antihistamines may **increase** the size of the reaction. Then we should use the following one-sided test:

  H_0 : Placebo and Antihistamines have the same distribution or the placebo distribution is to left of the antihistamine distribution.

  H_A : The distribution of antihistamine population is to the right of placebo population. Or equivalently the distribution of the placebo is to left of the antihistamine group. Draw the plot:

- We use the same $T = 129$, but reject the null if it is too small.

- The critical region is any number less than 69 (using $n_1 = 10$ and $n_2 = 9$). Since $T = 129 > 69$ we cannot reject then null.

Critical region: Two-sided test 5% level

- If we want to test

  H_0 : Placebo and Antihistamines have the same distribution.

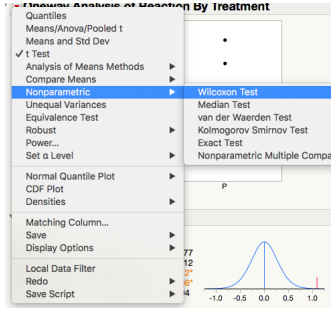
  H_A : The distribution of antihistamine population is a shift of the placebo population.

- Look up the Table 5, using the 5% level, two-sided test. The intersection of $n_1 = 10$ and $n_2 = 9$ gives the non-rejection region $[66, 114]$.

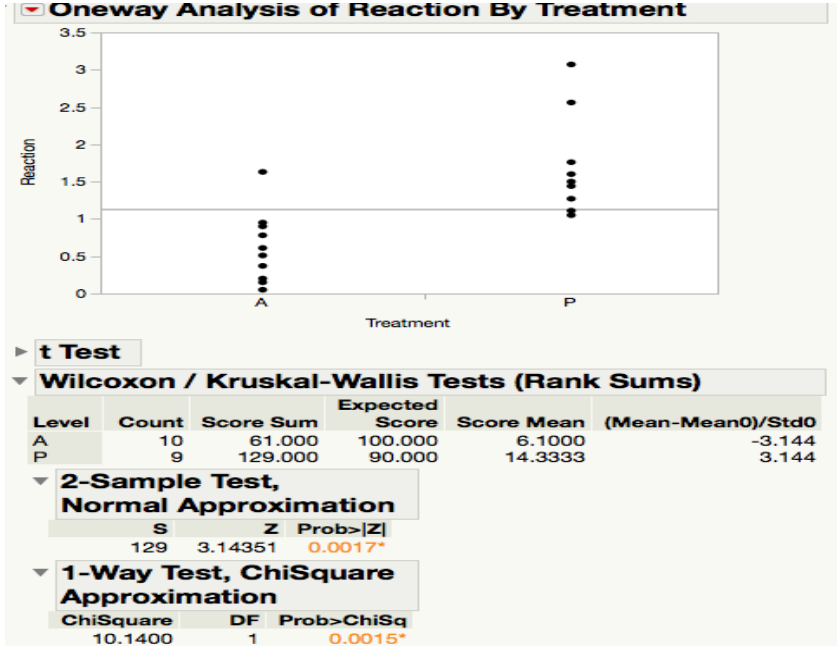
  If T lies outside this region we reject the null.

- Since $T = 129$ is not inside $[66, 114]$ we reject the null in the two-sided test and determine there is a difference. Which makes sense, if we reject the null in a one-sided test than we have to reject the null for a two-sided test.

The Wilcoxon-test in JMP



Use the usual two-sample instruction and then do this.



The sum of ranks is given in the output.
As are the results to two approximate two-sided tests.

Discussion of JMP output

- The JMP output gives the sum of the ranks; 129 for the small sample size and 61 for the large sample size.

- It also gives p-values, however, these are calculated using certain distributional approximations.

  Whereas the Wilcoxon tables give the exact critical regions.

- If the sample sizes are less than 10, use the Wilcoxon tables.

Example 2: Iron data using the Wilcoxon Rank test

group										
VitC	0.51	2.75	0.79	4.41	-1.23	1.06	1.98	2.32	1.59	-18.41
Calc	-0.18	0.92	-0.25	1.56	-0.38	-0.21	-0.62	-1.68	-3.15	-0.33

- H_0 : Vit C and Calcium have the same distribution or the Vit distribution is to left of the Calcium distribution. H_A : The distribution of Vitamin C population is to the right of Calcium population. Draw plot:

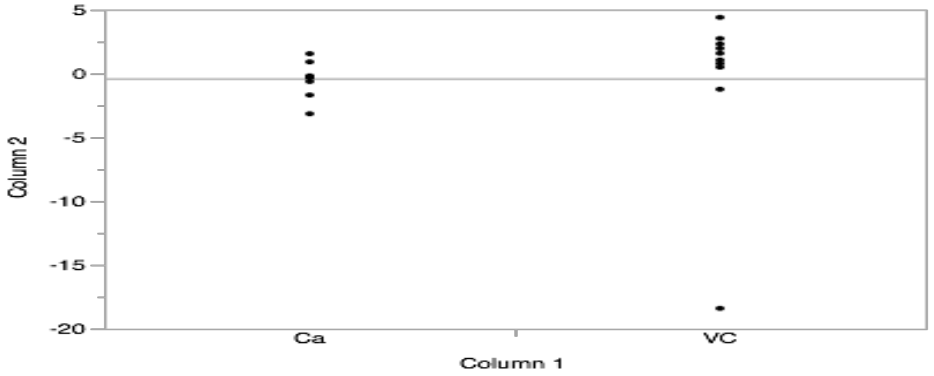
- The sample sizes for both groups is $n_1 = 10$ and $n_2 = 10$. This means we can pick any sample size for analysis.

- We pick the the ranks corresponding to the vitamin C group.

  The choice is arbitrary. However, under the alternative we are seeking to determine if the vitamin C ranks are "too large" for the null to hold.

Example 2: Iron data using the Wilcoxon Rank test

Oneway Analysis of Column 2 By Column 1



t Test

VC-Ca

Assuming unequal variances

Difference	0.0090	t Ratio	0.004299
Std Err Dif	2.0935	DF	9.707299
Upper CL Dif	4.6927	Prob > |t|	0.9967
Lower CL Dif	-4.6747	Prob > t	0.4983
Confidence	0.95	Prob < t	0.5017

Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Expected Score	Score Mean	(Mean-Mean0)/Std0
Ca	10	78.000	105.000	7.8000	-2.003
VC	10	132.000	105.000	13.2000	2.003

2-Sample Test, Normal Approximation

S	Z	Prob>|Z|
132	2.00321	0.0452*

1-way Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
4.1657	1	0.0413*

Example 2: cont

- From the JMP output the rank corresponding to the Vitamin C group is 132.

- Looking at the alternative we reject the null if $T = 132$ is larger than what we expect under the null.

- We do the one-sided test at the 5% level $n_1 = 10$ and $n_2 = 10$ gives $83, 127$. Since we want to determine if T is too large. The critical region is any number larger than 127.

- Since $T = 132 > 127$ we reject the null at the 5% level.

- There is some evidence in the data to suggest that those who consume vitamin C with tend to increase iron absorption over those who consume Calcium with Iron.

Example 3: Diet example using Wilcoxon Sum Rank test

Recall the sample in the diet data is rather small.

Diet I	2.9	2.7	3.9	2.7	2.1	2.6	2.2	4.2	5.0	0.7
Diet II	3.5	2.5	3.8	8.1	3.6	2.5	5.0	2.9	2.3	3

- Test the hypothesis that the two diets are the same against the alternative that the two diets are different.

- Test the hypothesis that the two diets are the same against the alternative that diet II is better than diet I.

Solution 3: By hand

First order both data sets:

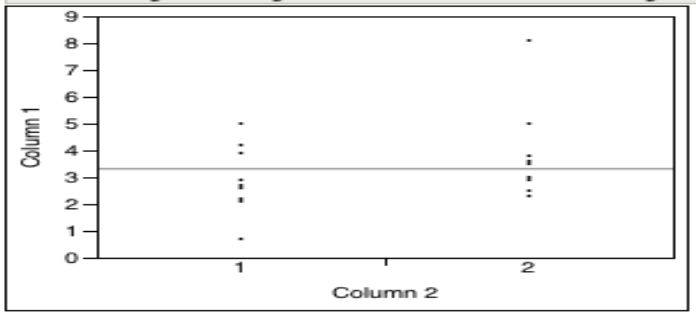
Diet I	2.9	2.7	3.9	2.7	2.1	2.6	2.2	4.2	5.0	0.7
Ordered	0.7	2.1	2.2	2.6	2.7	2.7	2.9	3.9	4.2	5.0
Diet II	3.5	2.5	3.8	8.1	3.6	2.5	5.0	2.9	2.3	3
Ordered	2.3	2.5	2.5	2.9	3.0	3.5	3.6	3.8	5.0	8.1
Rank diet I	1	2	3	7	8.5	8.5	10.5	16	17	18.5
Rank diet II	4	5.5	5.5	10.5	12	13	14	15	18.5	20

- Now add the ranks of either sample (because both have the same sample size - when the sample sizes are different, add the ranks in the smaller sample).

- Using software is a lot easier....

Diet data in JMP

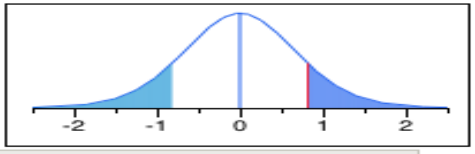
Oneway Analysis of Column 1 By Column 2



t Test

2-1
Assuming unequal variances

Difference	0.8200	t Ratio	1.22517
Std Err Dif	0.6693	DF	16.10337
Upper CL Dif	2.2381	Prob > |t|	0.2381
Lower CL Dif	-0.5981	Prob > t	0.1191
Confidence	0.95	Prob < t	0.8809



Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Expected Score	Score Mean	(Mean-Mean0)/Std0
1	10	92.000	105.000	9.2000	-0.946
2	10	118.000	105.000	11.8000	0.946

2-Sample Test, Normal Approximation

S	Z	Prob>|Z|
118	0.94634	0.3440

1-way Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
0.9686	1	0.3250

Determining the rejection region

- The sum of the ranks in Diet I is $T = 92$. The sum of ranks of Diet II is $T = 118$ (we can use either sums since the sample sizes of both samples are the same).

- We use these numbers to do the formal test.

Testing: H_0 : The distributions of both populations are both the same. H_A : One distribution is a shift of another (two sided test).

- Kook up Table 5, the first table (the column $n_2 = 10$ and $n_1 = 10$).

- Reading the table we see that the non-rejection region is

$$[79, 131].$$

- Since $T = 118$ lies in $[79, 131]$ there is not enough evidence to reject H_0 (noting if T were to lie outside this region there is enough evidence to reject H_0). $T = 118$ is neither too small or too big to suggest the null is implausible.

Warnings

- Nonparametric tests are conservative, meaning that even if the alternative is true, the test errs on the cautious side and tends not to reject the null (see the example on the next slide).

- We cannot test more complex hypothesis such as $H_0 : \mu_Y - \mu_X \leq 0.3$ vs $H_A : \mu_Y - \mu_X > 0.3$ using a nonparametric test.

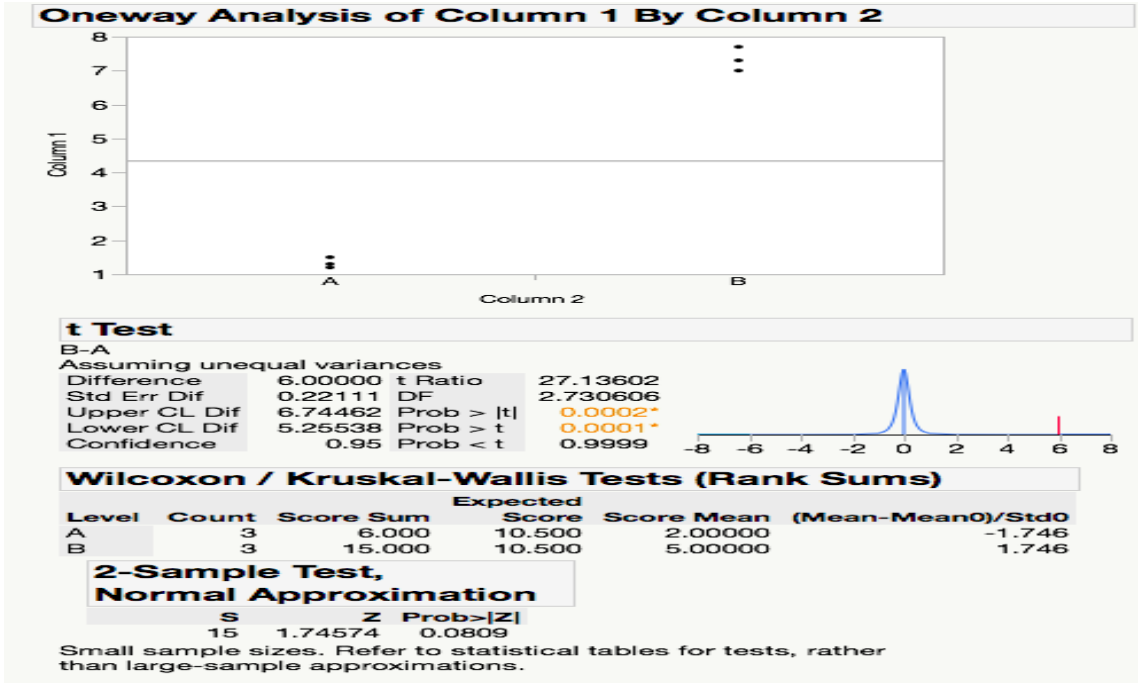
- We cannot make confidence intervals with nonparametric.

- One underlying assumption is that the shape of the distributions for populations is the same, we are only testing shifts of the distributions.

- If the shapes of the two populations are different, such as one distribution being narrower than another the test can yield incorrect results. This is

not an issue for the independent two sample t-test. The main assumption there is that the sample means are close to normal.

Why the Wilcoxon sum rank test is called conservative



- Observe that there are three observations in each of the two groups.
- There appears to be a clear difference between the two groups.

- There is a large separating between the two groups and the separation between the groups is greater than the variation/spread within the groups.

- Result of Independent sample t-test Despite the sample sizes being extremely small, the t-test test detects a difference and the p-value is extremely small (less than 0.02%). Hence there is evidence to suggest that this data has been take from two populations with two different means.

- Results of the Wilcoxon sum rank test The sum of the ranks is 6 or 15 (as the sample size in both groups is the same it does not matter which rank is chosen). The non-rejection region (for the two sided test) is $[5, 16]$. Since 6 (or 15) is within this region, we **cannot** reject the null using the Wilcoxon test at the 5% level. The p-value for this test is 5% of greater.

- The reason for the large discrepency between the p-values of the independent sample t-test and the Wilcoxon test is that Wilcoxon test does **not take into account the magntitude difference** between group A and B.

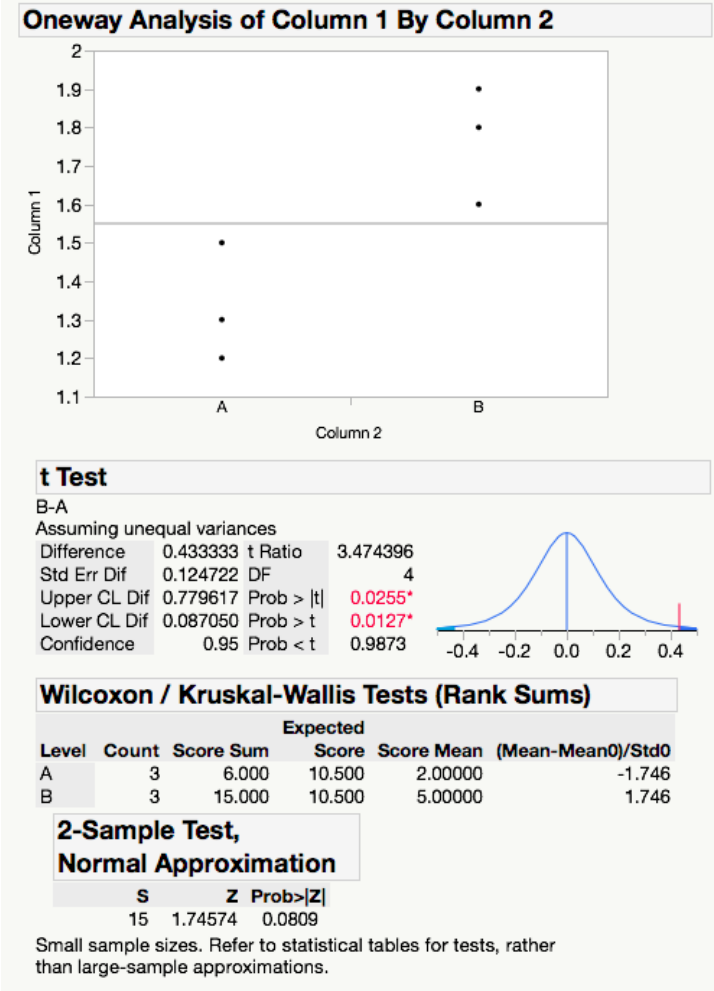
- One the other hand the independent sample t-test does account for the difference in magnitudes and spread.

- To illustrate this, consider the new data set on the next slide. The difference between the groups is less (and the spread greater).

- This leads to a larger p-value for the independent sample t-test.

- But for the Wilcoxon sum-rank test the ranks are identical.

Different Example: Same ranks different data



- **The results of the Wilcoxon test** Because the ranks of the test is **same** as in the previous example, the result of the Wilcoxon test is the same as in the previous example.

- **The results of the independent sample t-test** Because the magnitude of the difference between the two groups is **less** than the previous example, the p-value of the independent sample t-test is **greater** than the previous example $(2.5\%$ vs $0.02\%)$.