1 Nonparametric Statistics

When finding confidence intervals or conducting tests so far, we always described the population with a model, which includes a set of parameters. Then we could make decisions about the values of the parameters with the help of inferential statistics. These methods are referred to as Parametric Statistics.
In nonparametric statistics we will not built such a model, and will not have to assume that the population is normally distributed. This is very relevant, because data in many cases can not be assumed to come from a normal distribution.
Another advantage of nonparametric test is that they are applicable, even when the data is non-numerical but can be ranked (ordinal data, with many categories), like Likert type scales, often used in psychology.
Many of these methods are based on the ranks of the measurements in the sample and are therefore called rank statistics or rank tests.
We will introduce rank tests for one sample, comparing two samples (independent and paired), and more than two samples). We will find competing methods for one sample -, two sample -, paired t, and One-Way ANOVA.

1.1 The Sign Test

The sign test can be used to test hypotheses about the central tendency of nonnormal distributions. In order to measure the central tendency the mean would not be an appropriate measure for the center, because distribution might be skewed or have outliers. Instead we will use the median as a measure for the center of the distribution.
Remember: The median of a distribution is the value, so that the probability to fall below equals 0.5, or 50% of the measurements in the population fall below.
The median of a sample, is the value so that 50% of the sample data fall below the median.

The sign test for a population median η

1. Hypotheses:

test type	hypotheses
upper tail	$H_0 : \eta \leq \eta_0$ vs. $H_a : \eta > \eta_0$
lower tail	$H_0 : \eta \geq \eta_0$ vs. $H_a : \eta < \eta_0$
two tail	$H_0 : \eta = \eta_0$ vs. $H_a : \eta \neq \eta_0$

   Choose α.

2. Assumptions: Random sample.

3. Test statistic: S_0

test type	test statistic
upper tail	number of measurements greater than η_0
lower tail	number of measurements less than η_0
two tail	max(S_1,S_2), where S_1= number of measurements greater than η_0 and S_2= number of measurements less than η_0

   Note: Ignore all measurement that are equal to η_0.

4. P-value:

test type	P-value
upper tail	$P(x \geq S_0)$
lower tail	$P(x \geq S_0)$
two tail	$2P(x \geq S_0))$

where x has a binomial distribution with parameters n=sample size and $p = 0.5$.

5. Decision: If $P - value < \alpha$ reject H_0, otherwise do not reject H_0.

6. Put into context

The guiding idea for this test statistic is based on the following: Assume the median equals η_0, then about 50% of the sample data should fall above η_0.
But if the median is smaller than η_0, then less than 50% of the data is expected to fall above η_0.
That is when testing $H_0 : \eta \leq \eta_0$, we would expect about or less than 50% of the data to fall above η_0. In fact the distribution of S_0, if H_0 is true, is a binomial distribution with parameters n and $p = 0.5$.
The P-value for the test statistic, S_0, now finds how likely this number of measurements falls above η_0, if H_0 is true.

Example 1
The ammonia level (in ppm) is studied near an exit ramp of a San Francisco highway tunnel.
The daily ammonia concentration on eight randomly chosen days during afternoon drive-time are reproduced below:
1.37, 1.41, 1.42, 1.48, 1.50, 1.51, 1.53, 1.55
Do the the data provide sufficient evidence that the median daily ammonia concentration exceeds 1.5ppm. To answer this question conduct a sign test.

1. Hypotheses:

   $H_0 : \eta \leq 1.5$ vs. $H_a : \eta > 1.5$

   Choose $\alpha = 0.05$.

2. Assumptions: Random sample. The days the measurements were taken were chosen randomly.

3. Test statistic: (uppertail test) S_0 = number of measurements greater than $1.5ppm = 3$

4. P-value $= P(x \geq S_0) = P(x \geq 3) = 1 - P(x < 3) = 1 - P(x \leq 2) = 1 - 0.209 = 0.791$ $(k = 2, n = 8, p = 0.5)$

5. Decision: Since the $P - value = 0.791 > \alpha = 0.05$ do not reject H_0.

6. The data do not provide sufficient evidence that the daily ammonia concentration at this location exceeds 1.5ppm on 50% of days, or most of the time, or on the majority of days.

If the sample size is too large the calculation of the binomial probabilities becomes very time consuming (even for a computer).
Then we can use a test statistic which is based on the normal approximation of the binomial distribution: Replace steps 3 and 4 above by the following steps

3. Test statistic

$$z_0 = \frac{(S_0 - 0.5) - (0.5)n}{(0.5)\sqrt{n}}$$

4. P-value:

test type	P-value
upper tail	$P(z > z_0)$
lower tail	$P(z > z_0)$
two tail	$2P(z > z_0))$

Use this test statistic for $n \geq 10$.

Example 2
The above experiment is repeated with a higher sample size of 42. Out of the 42 days the ammonia concentration exceeded 1.5ppm on 24 days.

3. Test statistic

$$z_0 = \frac{(24 - 0.5) - (0.5)42}{(0.5)\sqrt{42}} = 0.77$$

4. P-value = $P(z > 0.77)$=1-(0.7794)= 0.2206 (z-table)

5. The P-value is greater than 0.05, do not reject H_0.

5. At significance level of 0.05 the data do not provide sufficient evidence, that at this location half of the days the ammonia concentration exceeds 1.5ppm.

Comment: Handling Ties
When two measurements are equal, we say there is a tie. For tied measurements we will assign the mean rank that the measurements would have received.

measurements:	1.2	2.4	2.4	3.1	4.7	9.5
rank:	1	2	3	4	5	6
rank(ties):	1	2.5	2.5	4	5	6

1.2 Rank Sum Test – Wilcoxon Test

Suppose measures of central tendency of two populations shall be compared but it is not save to assume that the populations are normally distributed. If the sample sizes are not large it is not appropriate to use the t-test.
In this case we will use the Wilcoxon Rank Sum Test.
Instead of using the original measurements, all sample data is ranked from smallest to largest and according to its position in the combined data set the rank is assigned.

Example 3
Six economists who work for the federal government and seven university economists are randomly selected, and each asked to predict next year's percentage change in cost of living as compared with this year's figure.
The objective is to compare the government economists' prediction to those of the university economists.

The data:

Government	University
3.1	4.4
4.8	5.8
2.3	3.9
5.6	8.7
0.0	6.3
2.9	10.5
	10.8

Experience has shown that such predictions tend to be skewed, and because of the small sample sizes a t-test is not appropriate.
To check if the data provide sufficient evidence that the center of the distributions of predictions are not the same for government and university economists, instead of using the raw scores, ranks are used.
Combine the two samples, sort from smallest to largest and assign ranks according to the position:

Government		University	
Prediction	Rank	Prediction	Rank
3.1	4	4.4	6
4.8	7	5.8	9
2.3	2	3.9	5
5.6	8	8.7	11
0.0	1	6.3	10
2.9	3	10.5	12
		10.8	13

The total ranks for the two groups are good indicators if the centres are different. If the medians for the two populations would be the same the average rank for the two groups should be close. If the mean rank for one group is "much smaller" than for the other this indicates that the median for the first group is most likely smaller than for the second (more of the small numbers came from group 1).

$$T_G = 4 + 7 + 2 + 8 + 1 + 3 = 25$$
$$T_U = 6 + 9 + 5 + 11 + 10 + 12 + 13 = 66$$

The total of the ranks is always equal to the total of the numbers from 1 to $n = n_1 + n_2$, $n(n+1)/2$. Since the total is fixed,the total for group 1 can serve as a test statistic and the average is not required.

Example 4
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. We shall compare "Native American" and "Caucasian" with respect to the variable MSCE (mean sister chromatid exchange).
Sister Chromatic Exchange occurs normally in cells during or cell division, but when a cells DNA is damaged by genotoxic agents, the rate of SCE increases. It is thought that SCE is an attempt

by the cell to fix the DNA damage caused by genotoxic agents; therefore, you would expect a more potent genotoxic agent to generate a higher rate of SCE.
The data is as follows:

Native American: 8.50 9.48 8.65 8.16 8.83 7.76 8.63 8.34 9.01 7.23
Caucasian: 8.27 8.20 8.25 8.14 9.00 8.10 7.20 8.32 7.70 8.43

	7.20	7.23	7.70	7.76	8.10	8.14	8.16	8.20	8.25	8.27	8.32	8.34	8.43	8.50
Race:	C	N	C	N	C	C	N	C	C	C	C	N	C	N
Rank:	1	2	3	4	5	6	7	8	9	10	11	12	13	14

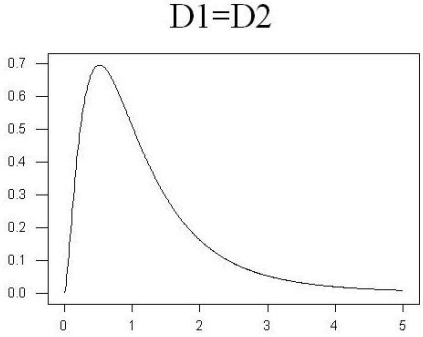
	8.63	8.65	8.83	9.00	9.01	9.48
Race:	N	N	N	C	N	N
Rank:	15	16	17	18	19	20

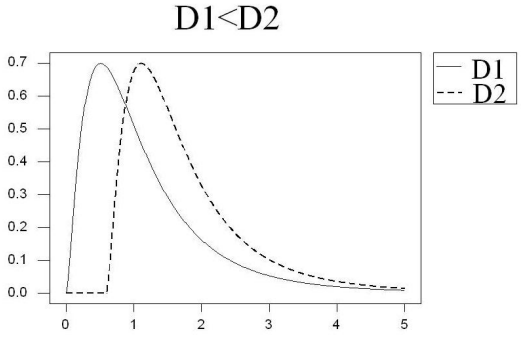
Then the sum of ranks for the Native American sample is
$T_N = 2 + 4 + 7 + 12 + 14 + 15 + 16 + 17 + 19 + 20 = 126$ and the sum of ranks for the Caucasian sample is
$T_C = 1 + 3 + 5 + 6 + 8 + 9 + 10 + 11 + 13 + 18 = 84$
The total of all ranks $= 1 + 2 + \ldots + 19 + 20 = 20(20 + 1)/2 = 210$.

Since $T_1 + T_2$ is the same for given sample sizes, it means that a small T_1 indicates a large T_2 and viceversa.

The rank sum of sample 1 is used as a test statistic to conduct tests for comparing the population distributions or medians.

Let D_1 and D_2 denote the distributions of population 1 and population 2, respectively. We will use the Wilcoxon test to detect location shifts.





In the first diagram, the density curves of the 2 distribution match each other, there is no shift ($D_1 = D_2$). In this case we would expect so see a random pattern in the ranks of the measurements from the two samples, which would result in the mean rank for the samples to be about the same. The distribution of the rank sum under the assumption that $D_1 = D_2$ has been worked out and computer programs are available to obtain probabilities for this random variable.

In the second diagram, the density curve of population 2 has the same shape as the one for population 1, but shifted to the right, denote ($D_1 < D_2$). In such a situation it should be expected that measurements from population one have lower ranks and measurements from population 2 have higher ranks. The mean rank for sample one should be expected to be smaller than for the sample from distribution 2.

In the case that $D_1 > D_2$, for similar reasons it is expected that the mean rank for sample 1 is large.

In conclusion: The mean rank (or the rank sum) of sample 1 reflects the location shift of distributions and is suitable to be used as a test statistic.

For paper and pencil applications of this test statistic, finding the p-value is cumbersome. In this case use the normal approximation of the rank sum as test statistic. The approximation is based on the following result.

Let T_1 be the rank sum of sample 1 out of two random samples. If $D_1 = D_2$ and the sample sizes are large, then T_1 is approximately normally distributed with

- mean
$$\mu_1 = \frac{n_1(n_1 + n_2 + 1)}{2}$$
- and standard deviation
$$\sigma_1 = \sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}}.$$

The Wilcoxon Rank Sum Test

1. Hypotheses: Let D_1 and D_2 denote the distributions of population 1 and population 2, respectively.

test type	hypotheses
upper tail	$H_0 : D_1 \leq D_2$ vs. $H_a : D_1 > D_2$
lower tail	$H_0 : D_1 \geq D_2$ vs. $H_a : D_1 < D_2$
two tail	$H_0 : D_1 = D_2$ vs. $H_a : D_1 \neq D_2$

   Choose α.

2. Assumptions: random samples from continuous distribution (not many ties)

3. Test statistic: Both sample sizes are at least 10.

   T_1=sum of ranks for sample 1.

$$z_0 = \frac{T_1 - \frac{n_1(n_1+n_2+1)}{2}}{\sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}},$$

4. P-value:

test type	P-value
upper tail	$P(z > z_0)$
lower tail	$P(z < z_0)$
two tail	$2P(z > abs(z_0))$

5. Decision: If $P - value < \alpha$ reject H_0, otherwise do not reject H_0.

6. Put into context

Continue example (Economists): Test if the center of the government economists is smaller than the center for the university economists:

1. Hypotheses: $H_0 : D(G) \geq D(U)$ vs. $H_a : D(G) < D(U)$ Let $\alpha = 0.05$.

2. Assumptions: random samples have to be from continuous distribution, the percentage is continuous.

3. Test statistic: Sample sizes are too small to use normal approximation (but we will do it anyway for demonstration purpose).

   $T(G)$=25, $n_G = 5, n_U = 6$.

$$z_0 = \frac{25 - \frac{6(6+7+1)}{2}}{\sqrt{\frac{6\times 7(6+7+1)}{12}}} = \frac{25-42}{\sqrt{49}} = -2.43,$$

4. P-value:

   lower tail P-value=$P(z < z_0) = P(z < -2.43) = 0.0075$

5. Decision: Since $P - value < 0.05 = \alpha$ reject H_0.

6. At significance level of 0.05 the data do provide sufficient evidence that the median predictions of next year's change in cost of living by government economists is less than the median predictions by university economists. Government economists are more optimistic about the future than university economists.

Continue example (genetic inheritance study):

1. Hypotheses: $H_0 : D(N) = D(C)$ vs. $H_a : D(N) \neq D(C)$ Let $\alpha = 0.05$.
2. Assumptions: random samples have to be from continuous distribution, the MSCE is continuous.
3. Test statistic: Sample sizes are at large enough to use normal approximation.
   $T(N)$=126, $n_N = n_C = 10$.
$$z_0 = \frac{126 - \frac{10(10+10+1)}{2}}{\sqrt{\frac{10\times 10(10+10+1)}{12}}} = \frac{126 - 105}{\sqrt{2100/12}} = 1.587,$$
4. P-value:
   two tail P-value=$2P(z > abs(z_0)) = 2(1 - 0.9429) = 0.1142$
5. Decision: Since $P - value > 0.05 = \alpha$ do not reject H_0.
6. At significance level of 0.05 the data do not provide sufficient evidence that there is a difference in the distribution of MSCE for the Native American and Caucasian population.

Comment: Handling Ties
For tied measurements we will assign the mean rank that the measurements would have received (like for the sign test).

measurements:	1.2	2.4	2.4	3.1	4.7	9.5
rank:	1	2	3	4	5	6
rank(ties):	1	2.5	2.5	4	5	6

1.3 Signed Rank Test

For comparing paired samples one should use the signed rank test. It is based on the pair differences like the paired t-test, and on a rank sum similar to the Wilcoxon rank sum test.

Example 5
Mileage tests are conducted to compare a new versus a conventional spark plug. A sample of 12 cars ranging form subcompacts to full-sized sedans are included in the study. Cars were driven with new and the conventional plug and the mileage was recorded. Based on the data we want to decide if the mileage increases with the new plug, that is test $H_0 : D_N \geq D_C$ versus $H_a : D_N > D_C$
The data:

Car Number	New A	Conventional B	Difference A-B	Rank
1	26.4	24.3	2.1	12+
2	10.3	9.8	0.5	4+
3	15.8	16.9	-1.1	8-
4	16.5	17.2	-0.7	5-
5	32.5	30.5	2.0	11+
6	8.3	7.9	0.4	3+
7	22.1	22.4	-0.3	2-
8	30.1	28.6	1.5	9+
9	12.9	13.1	-0.2	1-
10	12.6	11.6	1.0	7+
11	27.3	25.5	1.8	10+
12	9.4	8.6	0.8	6+

Now the data is ranked according to the ABSOLUTE value of the difference. Assign labels for positive and negative measurements.
The test statistic to use will be the sum of the ranks with a labelled "+". Here:

$$T_+ = 12 + 4 + 11 + 3 + 9 + 7 + 10 + 6 = 62$$

T_+ and T_- will for all samples of size 12 add up to 1+2+3+...+12= 12(13)/2=78.
We will have to judge T_+ if it indicates a shift in one of the distributions.
The argument follows the same line, as for the Wilcoxon rank sum test.
When $D_1 = D2$, then the rank sums should be about the same,
when $D_1 < D_2$, we have to expect more of the differences to be negative, and T_+ to be small, and
when $D_1 > D_2$, we should expect more positive differences and T_+ to be large.
Again the distribution of T_+ was worked out, and computer programs can work out the P-value for the Statistic.
For pencil and paper applications, we will use the normal approximation of that distribution, which is appropriate to use for sample sizes of at least 25.
When $D_1 = D_2$ and the sample size is large and the distribution of differences is symmetric, then T_+ is approximately normally distributed with

- mean
$$\mu_+ = \frac{n(n+1)}{4}$$
- and standard deviation
$$\sigma_+ = \sqrt{\frac{n(n+1)(2n+1)}{24}}.$$

Symmetry of the pairwise differences follows if for example the distributions for the first and second population have the same shape.

The Signed Rank Test

1. Hypotheses: Let D_1 and D_2 denote the distributions of population 1 and population 2, respectively.

test type	hypotheses
upper tail	$H_0 : D_1 \leq D_2$ vs. $H_a : D_1 > D_2$
lower tail	$H_0 : D_1 \geq D_2$ vs. $H_a : D_1 < D_2$
two tail	$H_0 : D_1 = D_2$ vs. $H_a : D_1 \neq D_2$

Choose α.

2. Assumptions: Paired random samples from continuous distribution (not many ties), with a symmetric distribution for the differences.

3. Test statistic: Sample size of pairs is at least 10.

   T_+=sum of ranks for positive differences.

$$z_0 = \frac{T_+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}},$$

4. P-value:

test type	P-value
upper tail	$P(z > z_0)$
lower tail	$P(z < z_0)$
two tail	$2P(z > abs(z_0))$

5. Decision: If $P - value < \alpha$ reject H_0, otherwise do not reject H_0.

6. Put into context

Continue Example:

1. Hypotheses: Let D_1 and D_2 denote the distributions of mileage using the new plug and using the conventional plug, respectively.

   upper tail $H_0 : D_1 \leq D_2$ vs. $H_a : D_1 > D_2$

   Choose $\alpha = 0.05$.

2. Assumptions: Random sample, Mileage is continuous we do not too have many ties. It can be assumed that the distributions have the same shape and therefore the differences are symmetrically distributed.

3. Test statistic: Sample sizes is $n = 12$.

   T_+=62

$$z_0 = \frac{T_+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{62 - \frac{12(13)}{4}}{\sqrt{\frac{12(13)(25)}{24}}} = 1.841$$

4. P-value: upper tail $P(z > 1.841) \approx 1 - 0.9671 = 0.0329$

5. Decision: $P - value < \alpha = 0.05$ reject H_0

6. At significance level of 5% the data provide sufficient evidence that the new plug increases the mileage.

1.4 Kruskal-Wallis H Test

We will introduce on more nonparametric Test. Here we will use ranks for the analysis of more than 2 populations when the response variable is quantitative. This will become an alternative for 1-way ANOVA.
This test is also based on the ranks assigned to the all measurements in the combined data set. The following example shall illustrate the test statistic.

Example 6
Three different methods for memorizing information are compared in an experiment. Eighteen participants are randomly assigned to three groups of 6 individuals. Each group uses a different method of memorization to memorize a list of 50 words. Then all participants are tested to determine how many of the words on the list they can resall.
The data:
Method 1: 23, 27, 30, 32, 33, 36
Method 2: 28, 30, 31, 35, 37, 40
Method 3: 16, 18, 22, 24, 27, 37

First we will assign ranks for each measurement, according to its placement in the combined data set. The ranks are:
Method 1: 4, 6.5, 9.5, 12, 13, 15
Method 2: 8, 9.5, 11, 14, 16.5, 18
Method 3: 1, 2, 3, 5, 6.5, 16.5

Now find the mean rank for each sample:
Method 1: $\bar{R}_1 = 60/6 = 10$
Method 2: $\bar{R}_2 = 77/6 = 12.83$
Method 3: $\bar{R}_3 = 34/6 = 5.67$
Then the mean of the mean ranks is always $\bar{R}$ = total of ranks)$n = \frac{n+1}{2} = 9.5$, because total of ranks=$n(n+1)/2$

The test statistic is comparing these mean ranks with $\bar{R}$.

$$H = \frac{12}{n(n+1)} \sum n_i (\bar{R}_i - \frac{n+1}{2})^2 = \frac{12}{18(19)}(6(10-9.5)^2 + 6(12.83-9.5)^2 + 6(5.67-9.5)^2) = 5.4753$$

If $D_1 = D_2 = D_3$, then the mean ranks would be close to $\bar{R}$, and H would be small, but if at least one of the distributions is (let's say) shifted to right, the that rank mean would tend to be larger than $\bar{R}$, and H would tend to be larger. This way H is sensitive to differences in the populations under investigation.

The distribution of H if the distributions are equal is approximately χ^2 distributed with $df = k - 1$. Based on this result we get the

The Kruskal-Wallis Test

1. Hypotheses:

   $H_0 : D_1 = D_2 = \ldots = D_k$ vs. H_a : at least on distribution is shifted to the right or left.

   Choose α.

2. Assumptions: Random samples of independent measurements, numerical response variable

3. Test statistic:
$$H_0 = \frac{12}{n(n+1)} \sum n_i(\bar{R}_i - \frac{n+1}{2})^2, \quad df = k-1$$

4. P-value=$P(\chi^2 > H_0)$ (table VII)

5. Decision: If $P - value < \alpha$ reject H_0, otherwise do not reject H_0.

6. Put into context

Continue example:

1. Hypotheses:
   $H_0 : D_1 = D_2 = D_3$ vs. H_a : at least on distribution is shifted to the right or left.
   Choose $\alpha = 0.05$.

2. Assumptions: We base this on random samples, number of words recalled is numerical

3. Test statistic:
$$H_0 = 5.4753 \quad df = 2$$

4. P-value=$P(\chi^2 > 5.4753)$, from table VII find 0.05 <P-value< 0.1.

5. Decision: Since $P - value > \alpha = 0.05$ do not reject H_0.

6. The data do not provide sufficient evidence against the assumption that the distributions of recollected words are not the same for the three memorization methods.

Multiple Comparison

If an analysis returns a significant Kruskal-Wallis Test we conclude that at least on of the distribution is different from the others. In this case the same question comes like after a significant ANOVA test: Where are the differences?
To answer this question a multiple comparison should be conducted:

Bonferroni as post hoc analysis for Kruskal Wallis Test

Choose experiment wise error rate α.

1. When comparing k distributions let $m = k(k-1)/2$ and the comparison wise error rate $\alpha^* = \alpha/m$.

2. Use the Wilcoxon Rank Sum test to compare every pair of treatments at significance level α^*.

Then all significant results hold simultaneously at experiment wise error rate of α.

Continue Example:
If we would have used $\alpha = 0.1$ in the example above, we would have decided that the data provide sufficient evidence to conclude that the distribution are not all the same.

Let us continue that example for $\alpha = 0.1$, and conduct a multiple comparison of the three distributions.

$m = 3$ and $\alpha^* = 0.1/3 = 0.0333$

1. We test:

   $H_0 : D_1 = D_2$ vs. $H_a : D_1 \neq D_2$
   $H_0 : D_1 = D_3$ vs. $H_a : D_1 \neq D_3$
   $H_0 : D_2 = D_3$ vs. $H_a : D_2 \neq D_3$

   $\alpha = 0.0333$.

2. Assumptions: checked already

3. Test statistic: Use normal approximation for demonstration purpose.

Ranks for comparing 1 and 2			test statistic
Method 1:	1, 2 , 4.5, 7, 8, 11	$T_1 = 33.5$	$z_0 = -0.88$
Method 2:	3, 4.5, 6, 9, 10, 12	$T_2 = 44.5$	
Ranks for comparing 1 and 3			
Method 1:	4, 6.5, 8, 9, 10, 11	$T_1 = 48.5$	$z_0 = 1.52$
Method 3:	1, 2, 3, 5, 6.5, 12	$T_1 = 29.5$	
Ranks for comparing 2 and 3			
Method 2:	6, 7, 8, 9, 10.5, 12	$T_2 = 52.5$	$z_0 = 2.16$
Method 3:	1, 2, 3, 4, 5, 10.5	$T_3 = 25.5$	

$$z_0 = \frac{T - \frac{6(13)}{2}}{\sqrt{\frac{6(6)(13)}{12}}} = \frac{T - 39}{\sqrt{39}},$$

4. P-value: 2 tailed test

Comparison	P-value
1-2	2(1-0.8106)= 0.3788
1-3	2(1-0.9357)= 0.1286
2-3	2(1-0.9846)= 0.0308

5. Decision: Since only the p-value for the comparison of method 2 and 3 falls below $\alpha^* = 0.0333$, we only reject that $D_2 = D_3$.

6. At significance level of 0.1, we conclude that method 2 results in a significantly higher number of recollected words than method 3. Method 1 is neither significantly different from method 2 nor from method 3.

```
3      1     2
-------
       ------
```