Nonparametric Tests

Introduction to Statistics
Carl von Ossietzky Universität Oldenburg
Fakultät III - Sprach- und Kulturwissenschaften

Introduction

- **Nonparametric procedure** or **distribution-free** inference procedures are not summerized by parameters to distributions: N(0,1), t(18), F(3,36), B(10,0.3).
- Nonparametric procedures are sometimes less powerful than tests designed for use with a specific distribution.
- It is better to use a test that is powerful when we believe that our assumptions are approximately satisfied than a less powerful test with fewer assumptions.
- Nonparametric tests use the order (or "rank") of the values rather than the actual values themselves.

Overview

Nonparametric	Parametric
Wilcoxon signed-rank	paired-samples t test
Mann-Whitney / Wilcoxon rank sum	independent-samples t test
Kruskal-Wallis	independent-samples one-way ANOVA
Friedman / Cochran's Q	repeated measures one-way ANOVA

Wilcoxon signed-rank test

- Can be used as an alternative to the paired-samples t-test when the population cannot be assumed to be normally distributed.
- Assumptions:
  - Data are paired and the differences come from the same population.
  - Each pair is chosen randomly and independent.
  - The data are measured at least on an ordinal scale.
  - The data need not be normally distributed, but the distribution of the differences should be symmetric around the median.

Wilcoxon signed-rank test

- **Example** (from Moore & McCabe): A study of early childhood education asked nursery school pupils to retell a fairy tale that had been read to them earlier in the week.
- Each child told two stories. The first had been read to them, and the second had been read but also illustrated with pictures.
- There were 5 low-progress readers.
- An expert listened to a recording of the children and assigned a score for certain uses of languages.
- Data:

Child	1	2	3	4	5
Story 2	0.77	0.49	0.66	0.28	0.38
Story 1	0.40	0.72	0.00	0.36	0.55
Difference	0.37	-0.23	0.66	-0.08	-0.17

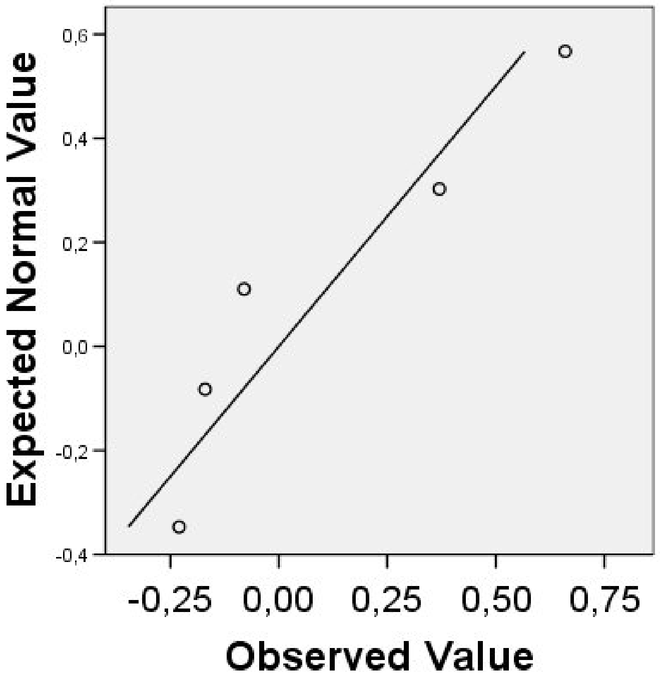
Wilcoxon signed-rank test

- Are the story 2 scores significantly higher than the story 1 scores?
- Hypotheses:

  H_0: the story 1 scores and the story 2 scores have identical distributions

  H_a: the story 2 scores are systematically higher than the story 1 scores
- Hypotheses are sometimes defines in terms of population medians η, but it is ok to express the hypothesis in words in this case.
- The Wilcoxon signed-rank test assumes that the population distribution of the paired differences is symmetric.

Wilcoxon signed-rank test



Normal quantile plots of the story 2 - story 1 differences.

Wilcoxon signed-rank test

- We consider absolute values of the differences:

Sign	+	-	+	-	-
Absolute value	0.37	0.23	0.66	0.08	0.17

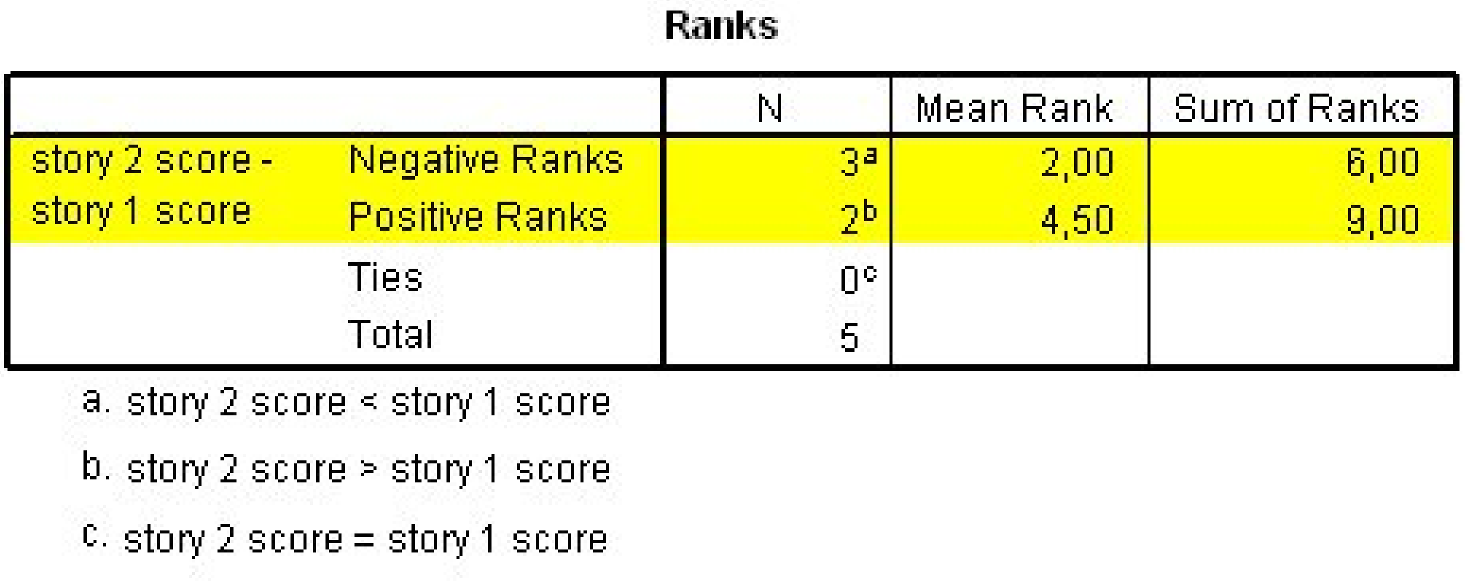
- Arrange the absolute differences in increasing order and assign ranks, keeping track of which values were originally positive and which were originally negative.
- Tied values receive the average of their ranks. If there are zero differences, discard them before ranking:

Sign	-	-	-	+	+
Absolute value	0.08	0.17	0.23	0.37	0.66
Rank	1	2	3	4	5

Wilcoxon signed-rank test

- The Wilcoxon W^+ statistic is the sum of the ranks of the positive differences: 4 $+$ 5 $=$ 9.
- The Wilcoxon W^- statistic is the sum of the ranks of the negative differences: 1 $+$ 2 $+$ 3 $=$ 6.
- S is the smaller of W^+ and W^-. Given S=6 and $n_{difference} = 5$, the software finds the p-value.

SPSS results



Ranks

		N	Mean Rank	Sum of Ranks
story 2 score - story 1 score	Negative Ranks	3[a]	2,00	6,00
	Positive Ranks	2[b]	4,50	9,00
	Ties	0[c]		
	Total	5		

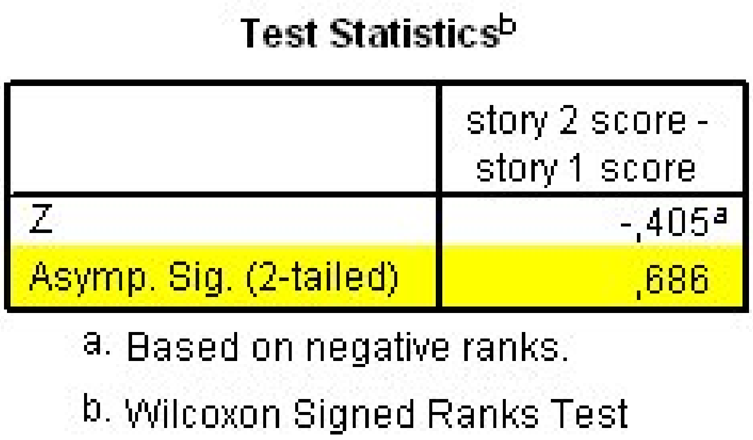
a. story 2 score < story 1 score

b. story 2 score > story 1 score

c. story 2 score = story 1 score

The sum of positive ranks (W^+) is higher than the sum of negative ranks (W^-) which suggests that the story 2 scores are higher than the story 1 scores. Therefore SPSS considers the 'story 2 score - story 1 score'. The corresponding one-sided alternative hypothesis H_a will be that the story 2 scores are systematically higher than the story 1 scores.

SPSS results



Test Statistics[b]

	story 2 score - story 1 score
Z	-,405[a]
Asymp. Sig. (2-tailed)	,686

a. Based on negative ranks.

b. Wilcoxon Signed Ranks Test

The p-value is based on a normal approximation. Since we test one-sided, our p-value is 0.686/2=0.343. Since this is higher than α (if α=0.05), we accept H_0 and conclude that there is no evidence that the the story 2 scores are systematically higher than the story 1 scores. When using a t test, the one-sided p-value would be 0.280 which leads to the same conclusion.

Effect size

- Is calulated as follows:

$$r = \frac{z}{\sqrt{N}}$$

where N is the total number of observations on which z is based.

Mann-Whitney test

- The MannWhitney U test is an alternative for the independent-samples t test in case of non-normality.
- Assumptions:
  - Mutual independence between samples.
  - Observations within a sample are independent of each other.
  - The data are measured at least on an ordinal scale.
  - The data need not be normally distributed, but the distributions of the samples should have the same shape.

Mann-Whitney test

- **Example** (from Moore & McCabe): A study of early childhood education asked nursery school pupils to retell a fairy tale that had been read to them earlier in the week.
- There were 5 high-progress readers and 5 low-progress readers.
- An expert listened to a recording of the children and assigned a score for certain uses of languages.
- Data:

high:	0.55	0.57	0.72	0.70	0.84
low:	0.40	0.72	0.00	0.36	0.55

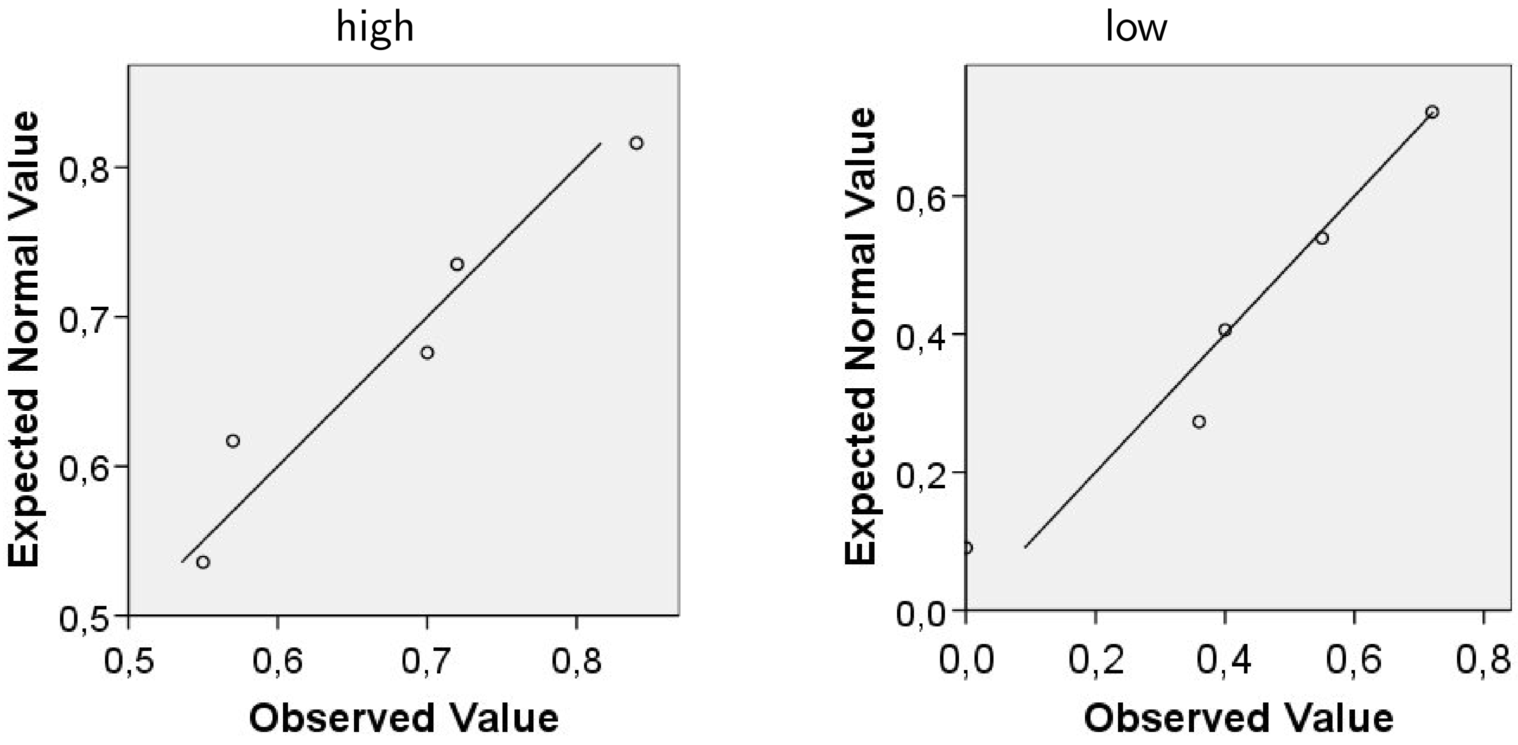
Mann-Whitney test

- Are the scores of the high progress group significantly higher than those of the low progress groups?
- Hypotheses:

  H_0: the high progress scores and the low progress scores have identical distributions

  H_a: the high progress scores are systematically higher than the low progress scores
- Hypotheses are sometimes defines in terms of population medians η, but can also be expressed in words.
- The Mann-Whitney test requires that the two tested samples be similar in shape.

Mann-Whitney test



Normal quantile plots for each of the two groups.

Mann-Whitney test

- **Step 1**: Taking each observation in low progress sample, count the number of observations in the high progress sample that are smaller than it (count a half for any that are equal to it):

scores low-progress readers	lower scores high-progress readers	number of lower scores
0.00		0
0.36		0
0.40		0
0.55	0.55	0.5
0.72	0.55 0.57 0.70 0.72	3.5
		4

Mann-Whitney test

- **Step 2**: Taking each observation in high progress sample, count the number of observations in the low progress sample that are smaller than it (count a half for any that are equal to it):

scores high progress readers	lower scores low-progress readers	number of lower scores
0.55	0.00 0.36 0.40 0.55	3.5
0.57	0.00 0.36 0.40 0.55	4
0.70	0.00 0.36 0.40 0.55	4
0.72	0.00 0.36 0.40 0.55 0.72	4.5
0.84	0.00 0.36 0.40 0.55 0.72	5
		21

- We found a total of 4 at step 1, and a total of 21 at step 2. The Mann-Whitney U statistic is the smaller of these two numbers: 4. Given $n_{high} = 5$ and n_{low}=5, the software finds the p-value.

Wilcoxon Rank Sum Test

- The **Wilcoxon rank sum test** is actually the same as the **Mann-Whitney test**.
- Rank all 10 observations. Arrange them in order from smallest to largest.
- The rank of each observation is its position in this ordered list, starting with rank 1 from the smallest observation.
- Assign all tied values the average of the ranks they occupy.

Wilcoxon Rank Sum Test

- Ranking:

progress	score	rank
low	0.00	1
low	0.36	2
low	0.40	3
high	0.55	4.5
low	0.55	4.5
high	0.57	6
high	0.70	7
high	0.72	8.5
low	0.72	8.5
high	0.84	10

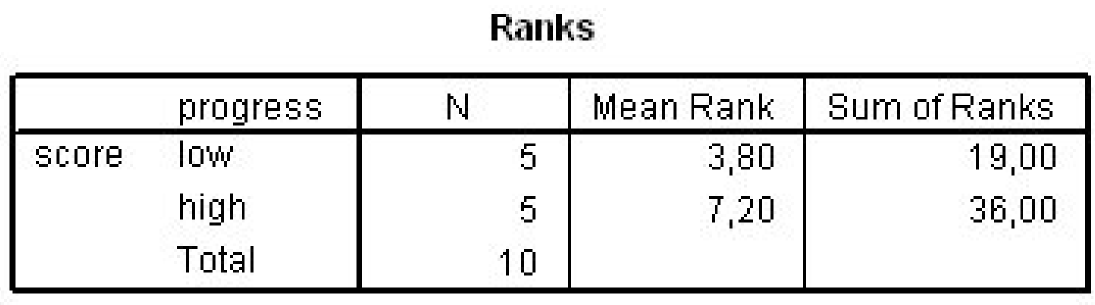
Wilcoxon Rank Sum Test

- We calculate the sum of ranks per group:

progress	sum of ranks
high	19
low	36

- The Wilcoxon W statistic is the smaller of these two numbers: 19. Given $n_{high} = 5$ and n_{low}=5, the software finds the p-value.

SPSS results

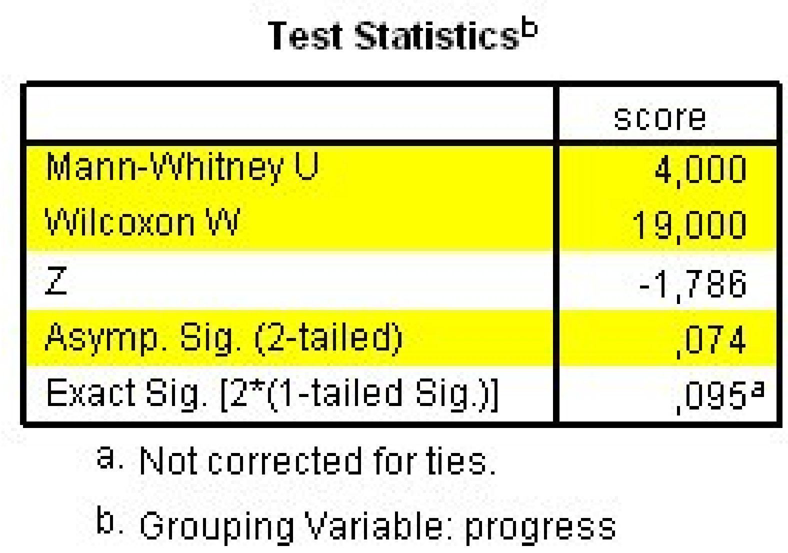


Ranks

	progress	N	Mean Rank	Sum of Ranks
score	low	5	3,80	19,00
	high	5	7,20	36,00
	Total	10		

The mean rank of the low progress group is lower than the mean rank of the high progress group. The number of observations in each group is the same. Therefore the corresponding one-sided alternative hypothesis H_a will be that the high progress scores are systematically higher than the low progress scores.

SPSS results



Test Statistics[b]

	score
Mann-Whitney U	4,000
Wilcoxon W	19,000
Z	-1,786
Asymp. Sig. (2-tailed)	,074
Exact Sig. [2*(1-tailed Sig.)]	,095[a]

[a] Not corrected for ties.

[b] Grouping Variable: progress

The Mann-Whitney U statistic and the Wilcoxon W statistic are the same as we found manually. Since we test one-sided, our p-value is 0.074/2=0.037. Since this is lower than α (if α=0.05), we reject H_0 and conclude that the high progress scores are systemetically higher than the low progress scores. When using a t test, the left-sided p-value would be 0.0445 which leads to the same conclusion.

Effect size

- Is calulated as follows:

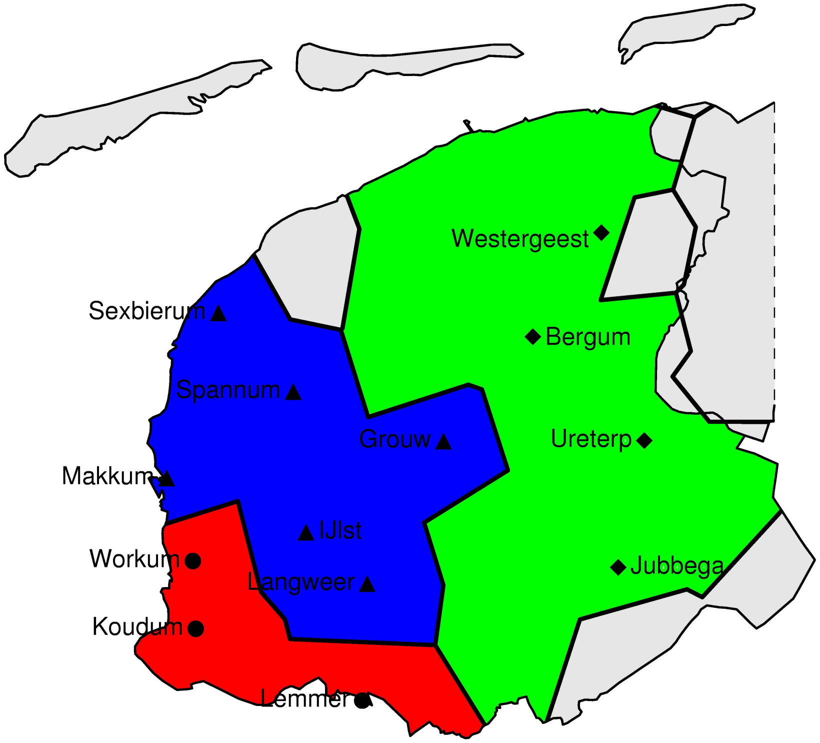
$$r = \frac{z}{\sqrt{N}}$$

where N is the total number of observations on which z is based.

Kruskall-Wallis test

- Used for comparing more than two samples that are independent. It is an extension of the Mann-Whitney U test to 3 or more groups.
- The test is an alternative for the one-way analysis of variance (ANOVA) when the data are not normally distributed.
- When the Kruskal-Wallis test gives significant results, then at least one of the samples is different from the other samples.
- Assumptions:
  - Mutual independence between samples.
  - Observations within a sample are independent of each other.
  - The data are measured at least on an ordinal scale.
  - The data need not be normally distributed, but the distributions of the samples should have the same shape.

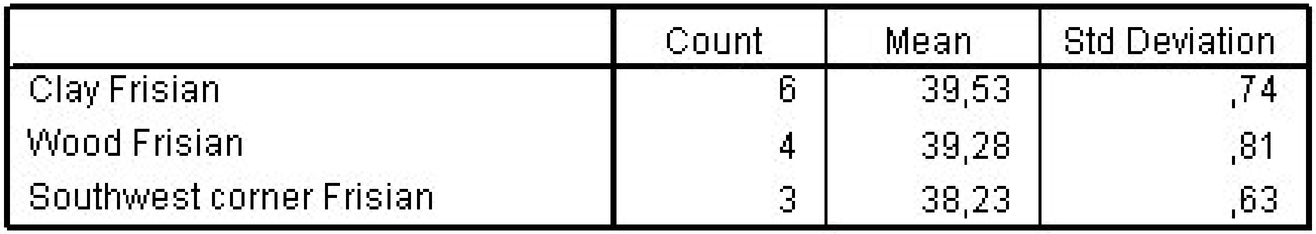
Kruskall-Wallis test



Frisian dialects are known to be distant to standard Dutch. They are traditionally divided in Clay Frisian (blue), Wood Frisian (green) and Soutwest corner Frisian (red).

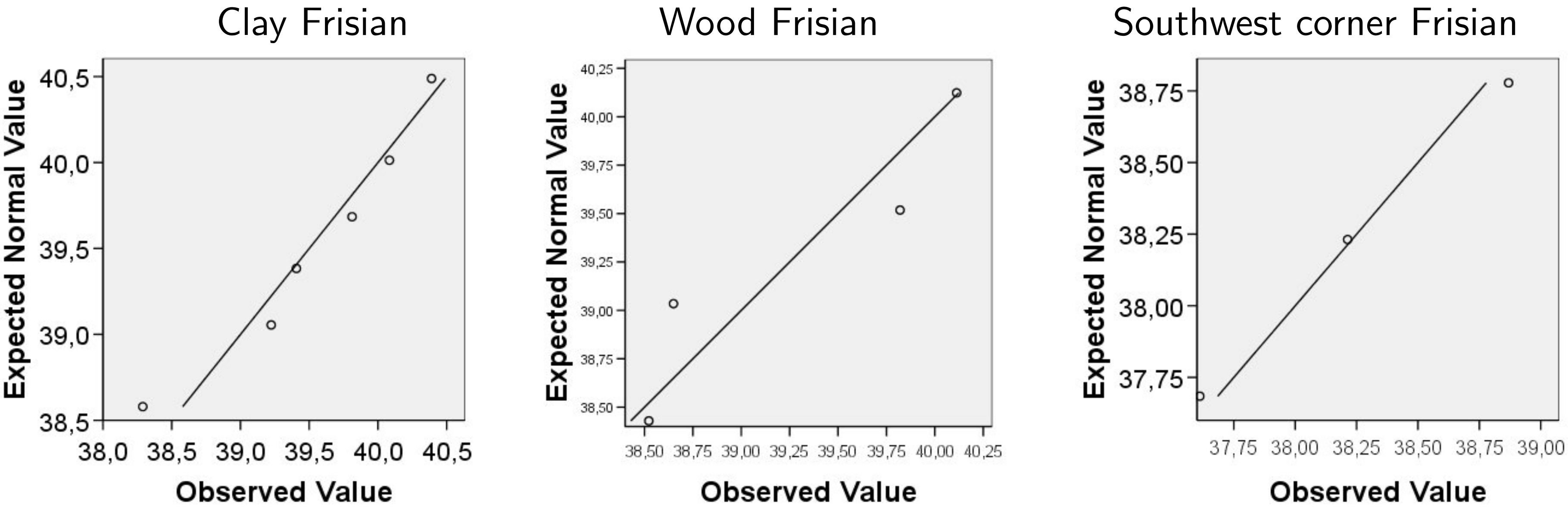
Kruskall-Wallis test

- Do Clay Frisian, Wood Frisian and Southwest corner Frisian have the same distance to standard Dutch?
- On the basis of 125 words we measure pronunciation distances between standard Dutch and Clay Frisian, Wood Frisian and Southwest corner Frisian dialects.
- Pronunciation distances are measured in a scale from 0% to 100%:
- Results:



	Count	Mean	Std Deviation
Clay Frisian	6	39,53	,74
Wood Frisian	4	39,28	,81
Southwest corner Frisian	3	38,23	,63

Kruskall-Wallis test



Clay Frisian

Wood Frisian

Southwest corner Frisian

Normal quantile plots for each of the three groups.

Kruskall-Wallis test

- Do Clay Frisian, Wood Frisian and Southwest corner Frisian have the same distance to standard Dutch?
- Hypotheses:

  H_0: The three Frisian populations have the same median

  H_a: Not all of the three population medians are the same.
- Rank all data from all groups together. Arrange them in order from smallest to largest.
- Assign all tied values the average of the ranks they occupy.

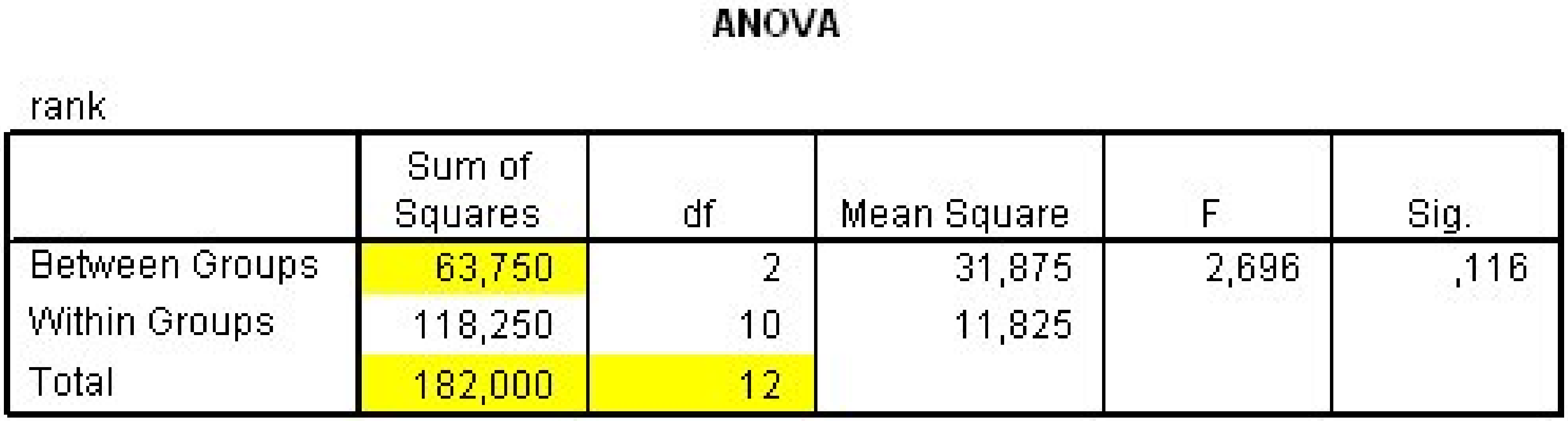
Kruskall-Wallis test

- Results:

group	dialect	distance	rank
3	Lemmer	37,61%	1
3	Workum	38,21%	2
1	Langweer	38,29%	3
2	Ureterp	38,52%	4
2	Jubbega	38,65%	5
3	Koudum	38,87%	6
1	Makkum	39,22%	7
1	Spannum	39,41%	8
1	Sexbierum	39,81%	9
2	Westergeest	39,82%	10
1	Grouw	40,08%	11
2	Bergum	40,11%	12
1	IJlst	40,39%	13

Kruskall-Wallis test

- Calculate SSG (sum of squares group) and SST (sum of squares total) on the basis of the **ranks**:



ANOVA

rank

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	63,750	2	31,875	2,696	,116
Within Groups	118,250	10	11,825		
Total	182,000	12			

- The Kruskal-Wallis statistic is:

$$H = (N-1)\frac{SSG}{SST} = (13-1)\frac{63.750}{182.000} = 4.203$$

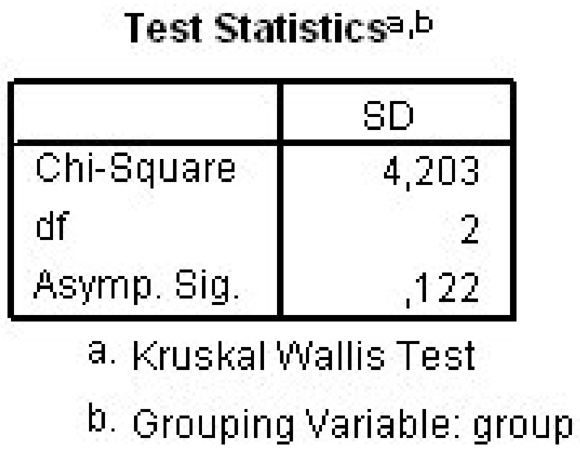
Kruskall-Wallis test

- When the sample sizes n_i are large (at least 5 observations per sample) and all I populations have the same continuous distribution, H has approximately the chi-square distribution with $I - 1$ degrees of freedom.
- The Kruskall-Wallis test rejects the null hypothesis that all populations have the same distribution when H is large.
- Given $H = 4.203$ and 3-1=2 degrees of freedom, go to:

  `http://www.vassarstats.net/`

  and choose: Distributions, Chi-Square Distributions, and enter the degrees of freedom.
- 4.203 is not found in the table, closest values are 4.0 and 4.25. We conclude that the p-value lies in the interval $0.1194 < P < 0.1353$.

SPSS results



Test Statistics[a,b]

	SD
Chi-Square	4,203
df	2
Asymp. Sig.	,122

a. Kruskal Wallis Test

b. Grouping Variable: group

The Kruskal-Wallis H statistic is the same as we found manually. The p-value is lower than α (if α=0.05), so we accept H_0 and conclude that there is no evidence that pronunciation distances are systematically higher in some groups than in others. When using ANOVA (on the basis of the real observations), the p-value would be 0.087 which leads to the same conclusion.

Effect size

- There is no easy way to find the effect size.

Friedman test

- This test is an alternative to the repeated measures ANOVA, when the assumption of normality or equality of variance is not met.
- The Friedman test is an extension of the Wilcoxon signed-rank test.
- Assumptions:
  - All observations are mutually independent
  - The rows are mutually independent. That is, the results in one row do not affect the results within other rows.
  - The data can be meaningfully ranked. The data are measured at least on an ordinal scale.
- Hypotheses:

  H_0: The different samples are drawn from distributions with the same median.

  H_a: At least one median is different from the rest.

Friedman test

- If this test gives a significant result, multiple comparisons are made by pairwise comparisons of the variables with the Wilcoxon signed-rank test, with the Bonferroni correction.
- The **Cochran's Q test** can be used for nominal data.

Effect size

- There is no easy way to find the effect size.
- The effect size per pair can easily be obtained: perform a Wilcoxon signed rank test per pair, and derived the effect size from the z test statistic.

Other options

- Remove outliers.
- Sometimes a right skewed distribution can be transformed in a normal distribution by performing a logarithmic transformation. This transformation tends to pull the righ tail of a distribution.
- Transforming the data also affects your hypotheses!