Non-parametric Tests and some data from aphasic speakers

Vasiliki Koukoulioti

Seminar Methodology and Statistics
19th March 2008

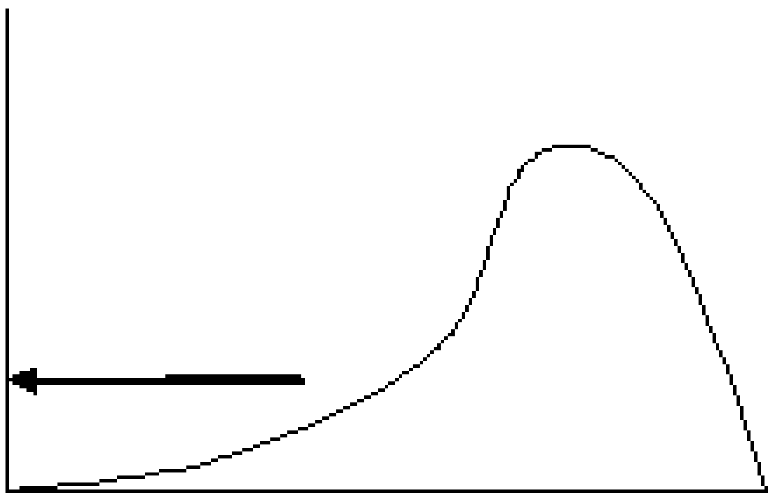
Some facts about non-parametric tests

- When to use non-parametric tests?
- What do they measure?
- What assumptions do they make?

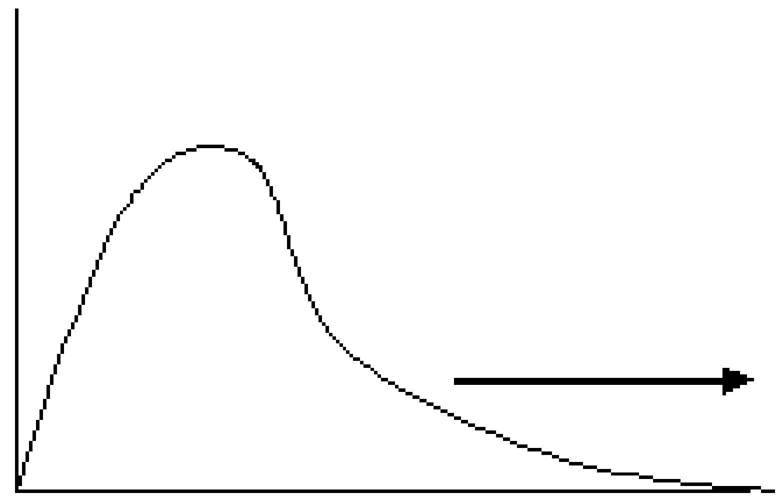
When to use non-parametric tests?

- When the normality conditions are not met (Moore & McCabe)
  - ✓ When the distribution of (at least) one variable is not normal
  - ✓ When the number of observations (N) is too small to assess normality adequately
  - ✓ When the distributions do not have the same shape

Compare:

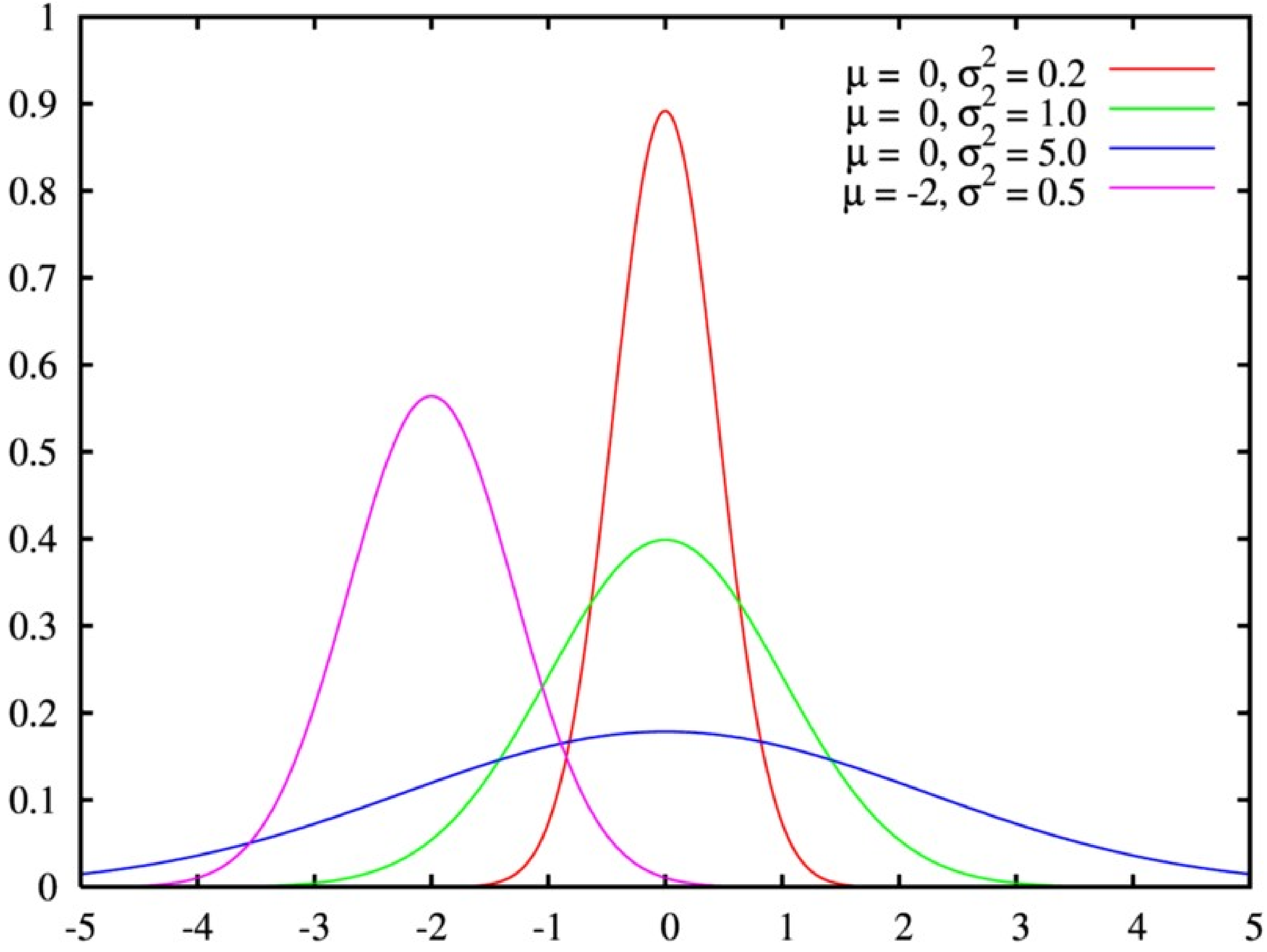


Negative Skew
Elongated tail at the **left**
More data in the left tail than would be expected in a normal distribution



Positive Skew
Elongated tail at the **right**
More data in the right tail than would be expected in a normal distribution

Compare:


μ = 0, σ² = 0.2
μ = 0, σ² = 1.0
μ = 0, σ² = 5.0
μ = -2, σ² = 0.5
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5

Setting	Normal test	Rank test
One sample	One-sample *t* test Section 7.1	Wilcoxon signed rank test Section 15.2
Matched pairs	Apply one-sample test to differences within pairs	
Two independent samples	Two-sample *t* test Section 7.2	Wilcoxon rank sum test Section 15.1
Several independent samples	One-way ANOVA *F* test Section 12	Kruskal-Wallis test Section 15.3

FIGURE 15.1 Comparison of tests based on normal distributions with nonparametric tests for similar settings.

Moore & McCabe Chapter 14, 5th Edition

What do non-parametric tests measure?

- Parametric tests make inferences about the mean of a sample
- When a distribution is strongly skewed→the center of the population is better represented by the median

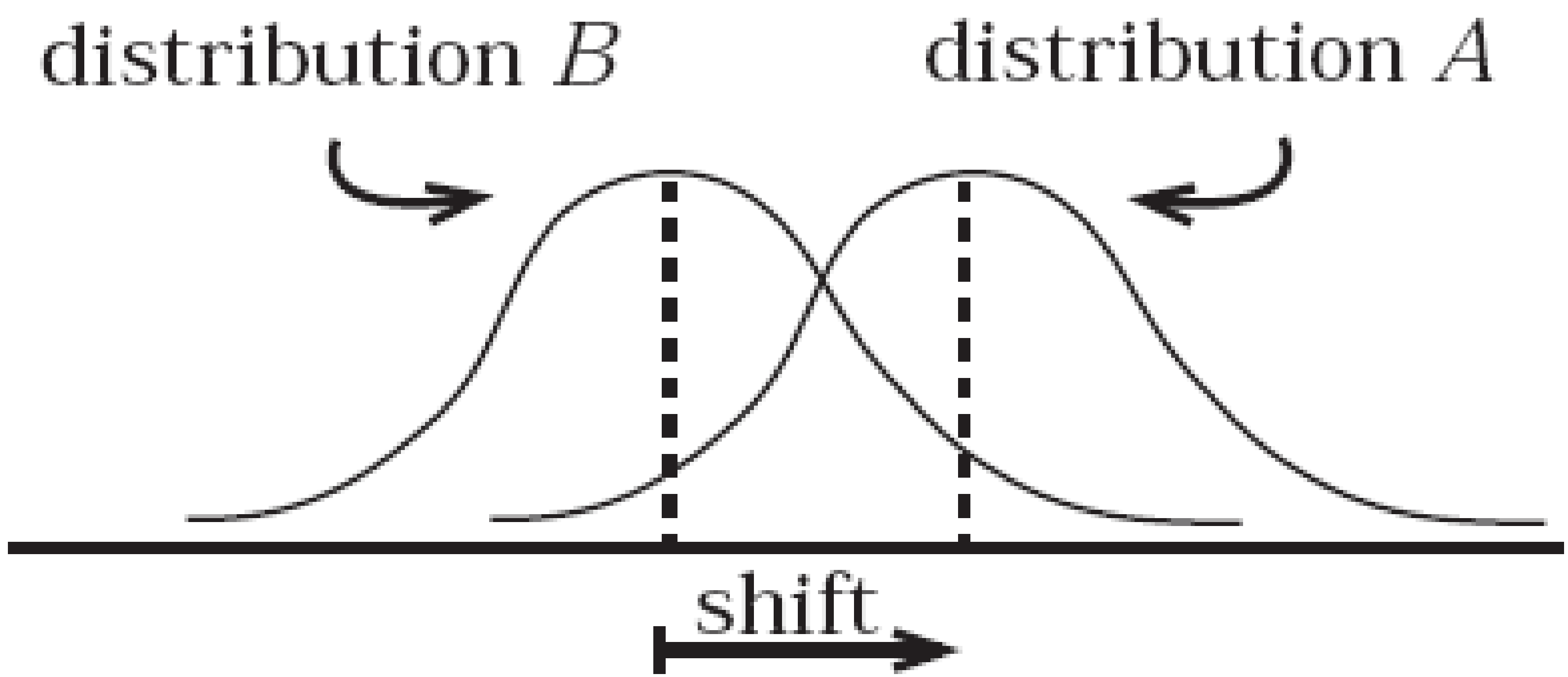
→ Non-parametric tests make hypotheses about the median instead of the mean

Recall:

- Mean $\mu=\sum xi/n$

- Median is the midpoint of a distribution, the number such that half the observations are smaller and the other half are larger.

→ Mean is more sensitive to outliers than the median



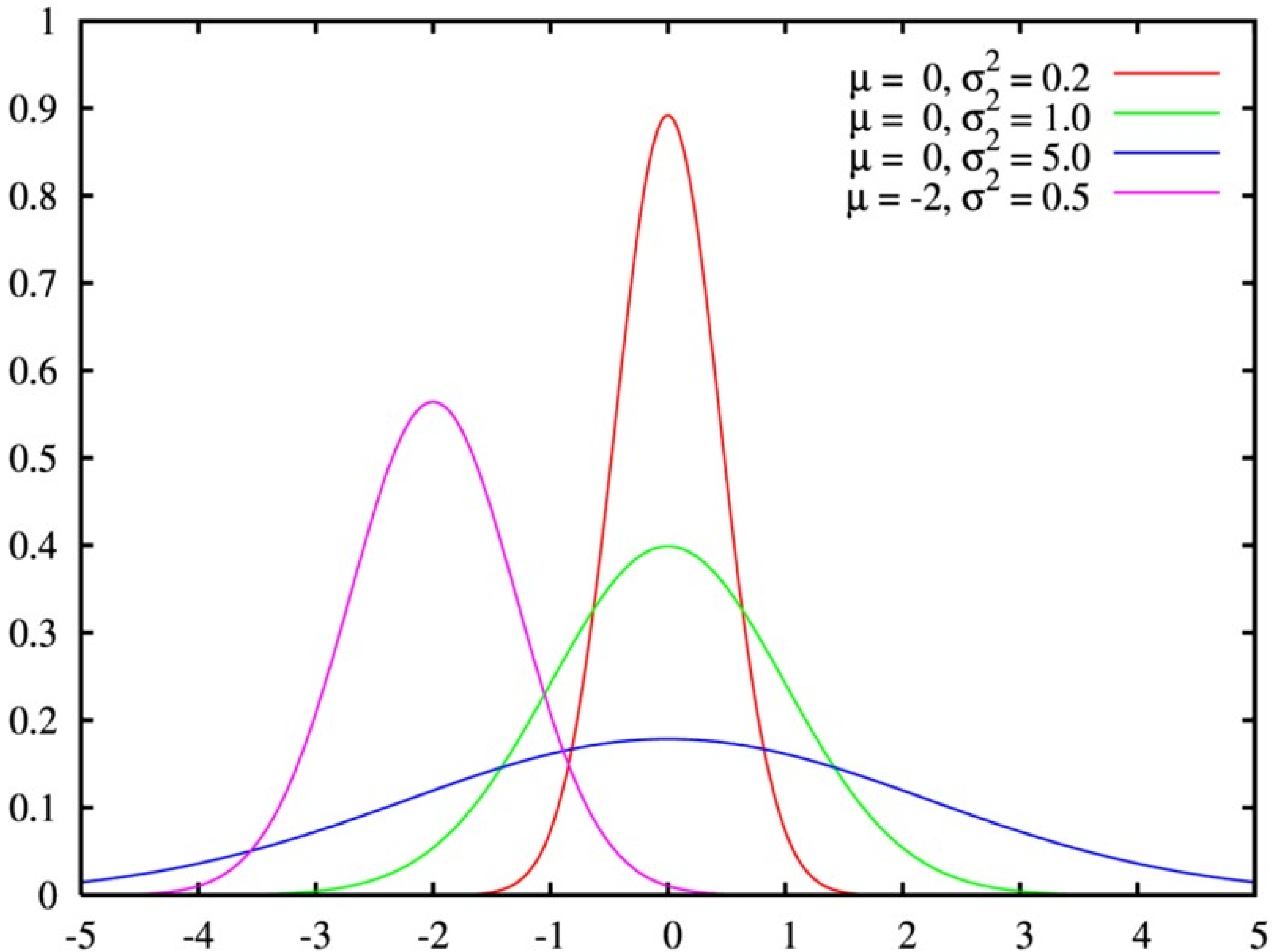
distribution B
distribution A
shift

But:

- This is so only if the (two or more) distributions have the same shape (practically impossible)

- Actually non-p tests measure whether the values of one distribution are systematically different than the values of the other distribution

Compare:



$\mu = 0, \sigma^2 = 0.2$
$\mu = 0, \sigma^2 = 1.0$
$\mu = 0, \sigma^2 = 5.0$
$\mu = -2, \sigma^2 = 0.5$
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5

Hypotheses with non-parametric tests

- One-tailed Hypothesis

✓H_0→ The two distributions are the same

✓H_a→ One distribution has values that are systematically larger

- Two-tailed Hypothesis

✓H_0→ The two distributions are the same

✓H_a→ One distribution has values that are systematically different (larger or smaller) than the other

What assumptions do non-parametric tests make?

- They are NOT totally assumption-free tests

- The variables must be continuous →

They can take any possible value within a given range

(very often violated assumption!!!)

Tests to be introduced:

- Wilcoxon Rank-Sum test (Mann-Whitney test)
- Wilcoxon Signed-Rank test
- Friedman Anova (x^2)

Wilcoxon Rank-Sum test (Mann-Whitney test)-an example

We want to see if weeds have an influence on the amount of yields of corn

Weeds per meter	Yield (bu/acre)			
0	166.7	172.2	165.0	176.9
3	158.6	176.4	153.1	156.0

Moore & Mc Cabe

Our Hypotheses:

H_0→ There is no difference in yields between plots with weed and weed free plots

H_a→ Plots with weed produce systematically fewer yields than weed-free plots

How to perform Wilcoxon Rank Sum test by hand

1) Rank the values

RANKS

To rank observations, first arrange them in order from smallest to largest. The **rank** of each observation is its position in this ordered list, starting with rank 1 for the smallest observation.

2) Keep track of which sample each value belongs to

Yield	153.1	156.0	158.6	**165.0**	**166.7**	**172.2**	176.4	**176.9**
Rank	1	2	3	**4**	**5**	**6**	7	**8**

3) Sum the ranks for each sample

Treatment	Sum of ranks
No weeds	23
Weeds	13

If H_0 is true the sum of ranks for each sample should be exactly the same!

The test statistic W

- W is the sum of the ranks of the one sample
- In this case the sum of ranks for corns with weeds is 23

THE WILCOXON RANK SUM TEST

Draw an SRS of size n_1 from one population and draw an independent SRS of size n_2 from a second population. There are N observations in all, where $N = n_1 + n_2$. Rank all N observations. The sum W of the ranks for the first sample is the **Wilcoxon rank sum statistic.** If the two populations have the same continuous distribution, then W has mean

$$\mu_W = \frac{n_1(N+1)}{2}$$

and standard deviation

$$\sigma_W = \sqrt{\frac{n_1 n_2(N+1)}{12}}$$

The **Wilcoxon rank sum test** rejects the hypothesis that the two populations have identical distributions when the rank sum W is far from its mean.*

Moore & McCabe

In this case:

$$\mu_W = \frac{n_1(N+1)}{2}$$

$$= \frac{(4)(9)}{2} = 18$$

and standard deviation

$$\sigma_W = \sqrt{\frac{n_1 n_2 (N+1)}{12}}$$

$$= \sqrt{\frac{(4)(4)(9)}{12}} = \sqrt{12} = 3.464$$

Is it significant?

- $W=23$ and $\mu W=18$, and $\sigma W=3.64$
- $W>\mu W$ but only 1.4 SDs [(23-18)/3.64]

✂→ probably not significant difference

✓We can calculate it

  - ✓By the tables
  - ✓ By the normal approximation (with continuity correction!!)

		Lower Tail						Upper Tail					
		prob						***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

Normal approximation-z-score

$$z = \frac{W - \mu_W}{\sigma_W} = \frac{23 - 18}{3.464} = 1.44$$

P(Z≥1.44)=1-0.9251=0.0749 from the tables of the normal curve

Continuity correction!

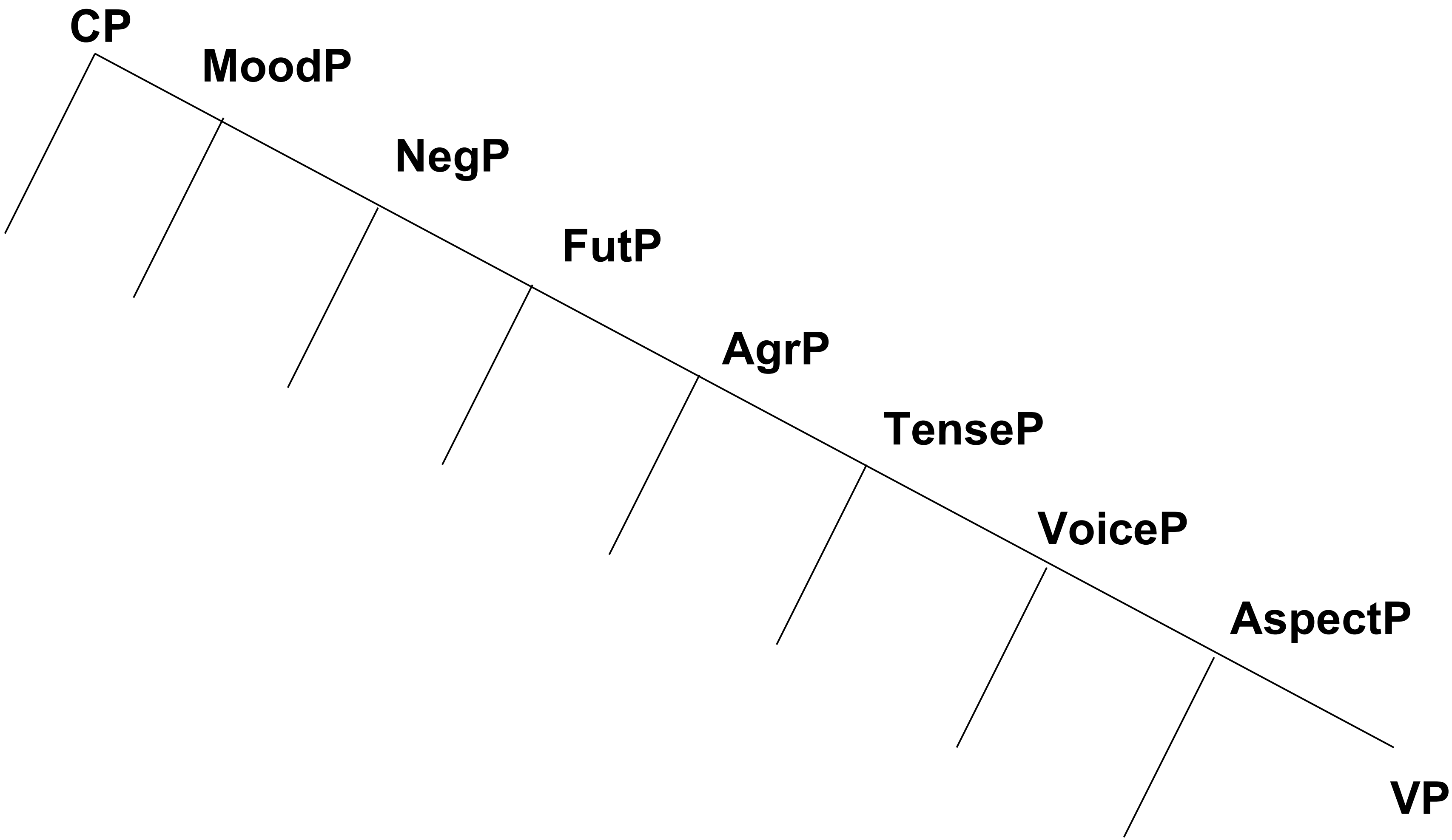
- Continuity correction assumes that X=23 includes all the values from 22.5 to 23.5
- So here we will calculate the z-score of 22.5 since we want to find P(W≥23)

$$P(W \geq 22.5) = P\left(\frac{W - \mu_W}{\sigma_W} \geq \frac{22.5 - 18}{3.464}\right)$$
$$= P(Z \geq 1.30)$$
$$= 0.0968$$

The experimental design

- 2 Groups
  - ✓ non-fluent patients (N=3)
  - ✓ healthy controls (N=4)
- 4 conditions
  - ✓ Indicative affirmative (24)
  - ✓ Indicative negative (24)
  - ✓ Subjunctive affirmative (24)
  - ✓ Subjunctive negative (24)

The Greek clause structure (Philippaki-Warburton, 1990;1998)



Wilcoxon Rank-Sum test (Mann-Whitney test)

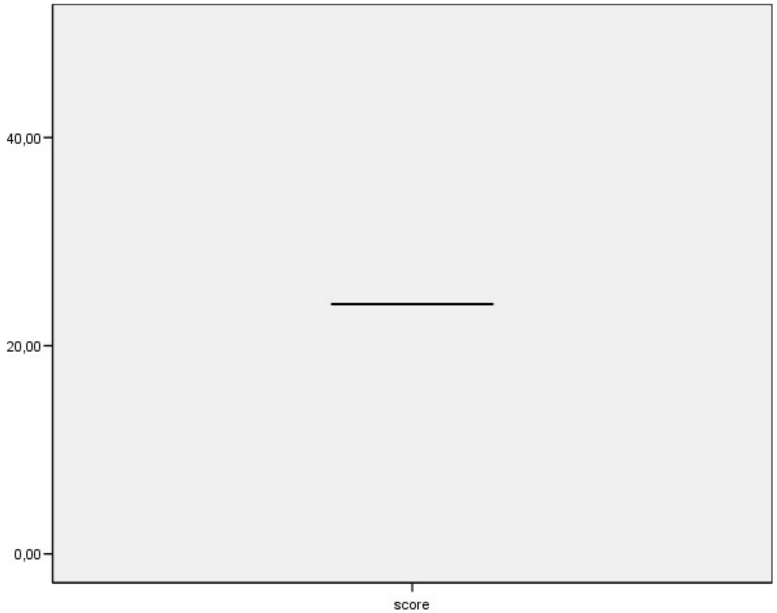
- Comparison between 2 independent samples – 1 condition (Indicative affirmative)

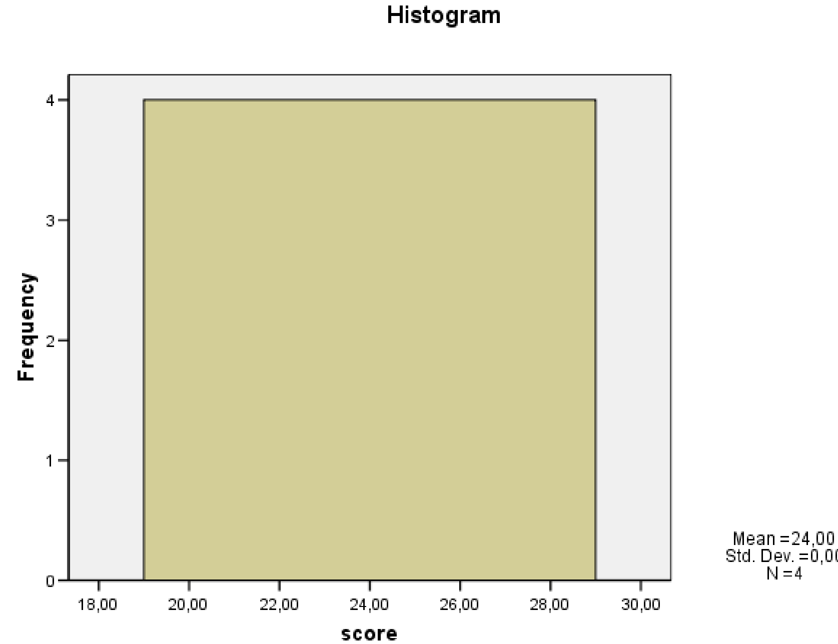
H_0 → Both groups perform equally

H_a → Controls perform better than patients

Data

Participant	Score
C1	24
C2	24
C3	24
C4	24
P1	22
P2	22
P3	23

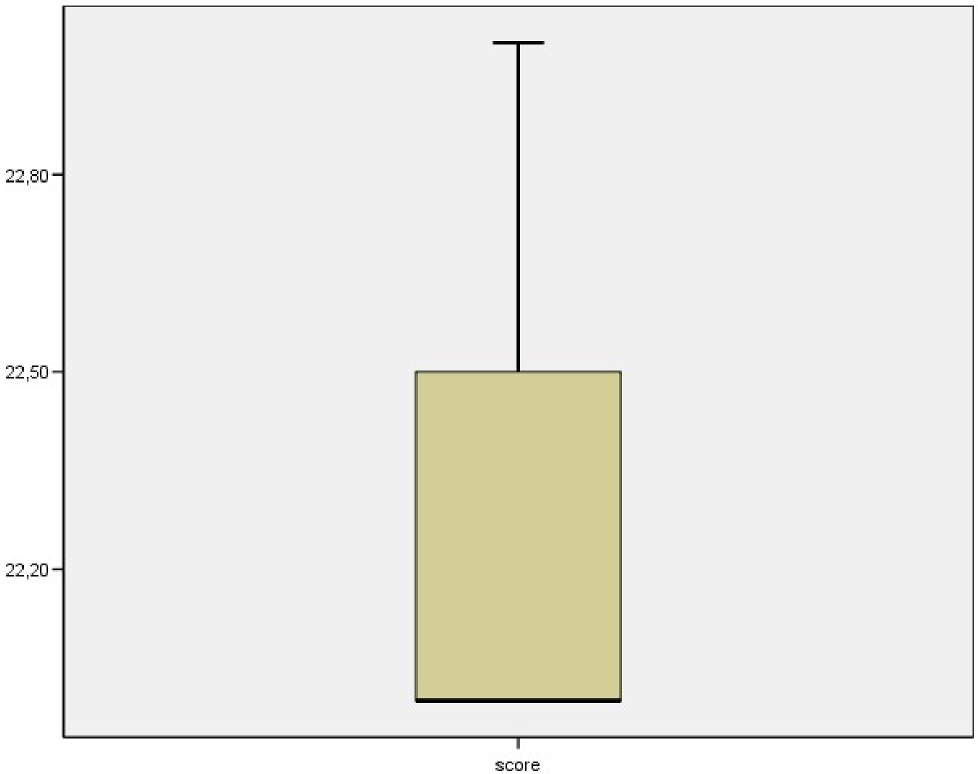




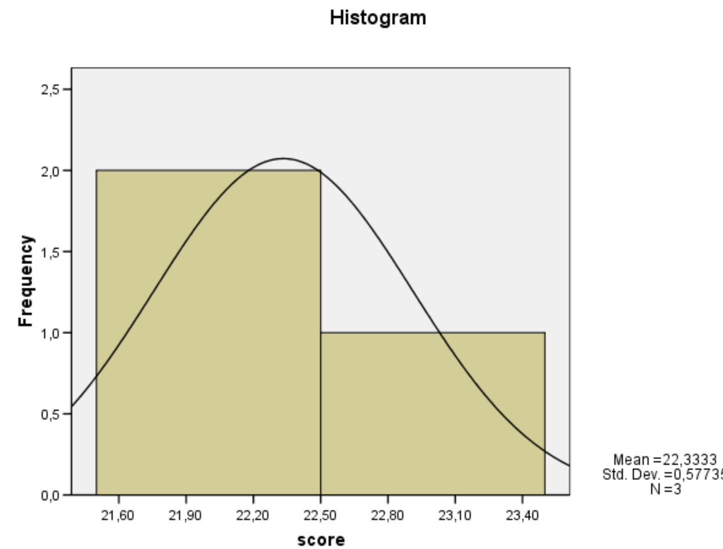
Boxplot

Histogram

Distribution of the controls'scores



Boxplot



Histogram

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
score	,385	3	.	,750	3	,000

a. Lilliefors Significance Correction

Distribution of the patients'scores

Ranking

1 22 23 24 24 24 24

1 2 3 4 5 6 7

1.5 1.5 3 5.5 5.5 5.5 5.5

- Because we have a lot of ties we must trust a statistics package!

- Ties influence the exact distribution of the W and the SD of the W must be adjusted

Ranks

	group	N	Mean Rank	Sum of Ranks
score	controls	4	5,50	22,00
	patients	3	2,00	6,00
	Total	7		

Test Statistics[b]

	score
Mann-Whitney U	,000
Wilcoxon W	6,000
Z	-2,366
Asymp. Sig. (2-tailed)	,018
Exact Sig. [2*(1-tailed Sig.)]	,057[a]
Exact Sig. (2-tailed)	,029
Exact Sig. (1-tailed)	,029
Point Probability	,029

a. Not corrected for ties.

b. Grouping Variable: group

We should accept the Ha that the control group performed systematically better than the patient group

Friedman's ANOVA

- We want to compare the performance of the aphasic speakers in the 4 condition
- 1 group k conditions
- Hypotheses:

H_0→ Patients perform equally in all 4 condition

H_a→ There is a difference in the performance of patients across conditions

The data

	scores				ranks			
	i.a.	i.n.	s.a.	s.n.	i.a.	i.n.	s.a.	s.n.
P1	22	18	12	12	4	3	1.5	1.5
P2	22	18	11	1	4	3	2	1
P3	23	23	0	1	3.5	3.5	1	2
Sum of Ranks					11.5	9.5	4.5	4.5

The test statistic Fr

$$F_r = \left[\frac{12}{Nk(k+1)} \sum_{j=1}^{k} R_j^2 \right] - 3N(k+1)$$

N= sample size, k=number of conditions, Rj=sum of ranks for each condition

P-value from tables of chi-square distribution

Here we have

- $Fr = 7.6$, $p > 0.05$, we accept the H_0

Ranks

	Mean Rank
indicative affirmative	3,83
indicative negative	3,17
subjunctive affirmative	1,50
subjunctive negative	1,50

We should accept the Ha that the perfromance of the patients is different across conditions

Test Statistics[a]

N	3
Chi-Square	8,143
df	3
Asymp. Sig.	,043
Exact Sig.	,021
Point Probability	,014

a. Friedman Test

Post hoc

- There are differences but between which conditions and which direction do they have?
- Wilcoxon signed-rank test
- Bonferroni correction (α-level/ number of comparisons=0.05/6=0.008)

Theory of Wilcoxon‘s sign rank test

	i.a.	i.n.	Diff	sign	Rank	+	-
P1	22	18	4	+	1.5	1.5	
P2	22	18	4	+	1.5	1.5	
P3	23	23	0	excl			
Total						3	0

THE WILCOXON SIGNED RANK TEST FOR MATCHED PAIRS

Draw an SRS of size n from a population for a matched pairs study and take the differences in responses within pairs. Rank the absolute values of these differences. The sum W^+ of the ranks for the positive differences is the **Wilcoxon signed rank statistic.** If the distribution of the responses is not affected by the different treatments within pairs, then W^+ has mean

$$\mu_{W^+} = \frac{n(n+1)}{4}$$

and standard deviation

$$\sigma_{W^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

The **Wilcoxon signed rank test** rejects the hypothesis that there are no systematic differences within pairs when the rank sum W^+ is far from its mean.

Test Statistics[b]

	indneg - indaff
Z	-1,414[a]
Asymp. Sig. (2-tailed)	,157
Exact Sig. (2-tailed)	,500
Exact Sig. (1-tailed)	,250
Point Probability	,250

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

Test Statistics[b]

	subjaff - indaff
Z	-1,604[a]
Asymp. Sig. (2-tailed)	,109
Exact Sig. (2-tailed)	,250
Exact Sig. (1-tailed)	,125
Point Probability	,125

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

Test Statistics[b]

	subjneg - indaff
Z	-1,604[a]
Asymp. Sig. (2-tailed)	,109
Exact Sig. (2-tailed)	,250
Exact Sig. (1-tailed)	,125
Point Probability	,125

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

Test Statistics[b]

	subjaff - indneg
Z	-1,604[a]
Asymp. Sig. (2-tailed)	,109
Exact Sig. (2-tailed)	,250
Exact Sig. (1-tailed)	,125
Point Probability	,125

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

Test Statistics[b]

	subjneg - indneg
Z	-1,604[a]
Asymp. Sig. (2-tailed)	,109
Exact Sig. (2-tailed)	,250
Exact Sig. (1-tailed)	,125
Point Probability	,125

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

Test Statistics[b]

	subjneg - subjaff
Z	-,447[a]
Asymp. Sig. (2-tailed)	,655
Exact Sig. (2-tailed)	1,000
Exact Sig. (1-tailed)	,500
Point Probability	,250

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

No difference could be found between conditions! Recall that Friedman‘s ANOVA was marginally significant!