The binomial distribution and proportions

Erik-Jan Smits
& Eleonora Rossi

Sampling distributions

Moore and McCabe (2003:367) :

> “Nature of sampling distribution depends on both the nature of the population distribution and the way we collect the data from the population’’

Sampling distributions for counts and proportions?

> e.g. percentage of women customers for an e-business site; male vs. female and therefore ***categorical data***

Sampling distributions: counts and proportions

Question:

Is the percentage of women visiting e-business sites significantly greater at one sort of site (e.g. films) as opposed to another (e.g. music).

Two approaches:

1. Proportions may be viewed as numerical data. Use *t*-test.
   see http://home.clara.net/sisa/
2. Use the **binomial distribution**

Binomial distribution

- A random variable *X*: a ***count*** of the occurrences of some outcome in a fixed number of observations (*n*)
  - Each observation falls into one of just *two* categories: success vs. failure, male vs. female, child vs. adult
  - The *n* observations are all independent
  - The probability of success, call it *p*, is the same for each observation

In sum:

X is B(*n,p)*

Binomial distribution

X is B(*n,p)*

Example:

tossing a coin *n* times; each toss gives either heads or tails. Call heads a success; *p* is the probability of a head. The number of heads we count is a random variable X.

Sample proportions

Note:

Distinguish *the proportion p* from the *count* X! The distribution of the *count X* has a binomial distribution, the *proportion p* does NOT have a binominal distribution.

BUT:

If we want to do probability calculations about p*:*

Restate them in terms of count X and use binomial methods.

Proportions and the sign test

Example
17 teachers attend a summerschool to improve their French listening skills. They were given a pretest and a posttest; 16 teachers improved, 1 did more poorly. Question: did participation improve their performance on the listening tests?

- Assumption: the population distribution does not have any specific form, such as normal

↓

- Use distribution free procedures (also called 'non-parametric procedures') – uses probability calculations that are correct for a wide range of population distributions – e.g. the **sign test for matched pairs.**

Sign test

	Pair										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Pretest	32	29	31	10	30	33	22	32	24	20	30
Posttest	34	35	31	16	33	36	24	26	24	26	36
	+	+	0	+	+	+	+	-	0	+	+

data: M&M

Sign test for matched pairs
Ignore pairs with the difference 0; the number of trials *n* is the count of the remaining pairs. The test statistic is the count *X* of pairs with a positive difference. *P*-values for X are based on the binomial *B*(*n,* 1 / 2) distribution

Sign test

Example

17 teachers attend a summerschool to improve their French listening skills. 16 teachers improved, 1 did more poorly. Question: did participation improve their performance on the listening tests?

- Null-hypothesis of “no effect’’ is:
  $H_0 = p = 1 / 2$
  $H_a = p > 1 / 2$
- X (count) has the B(17,1/2) distribution
- $P(X = k) = (n/k)\, p^k (1 - p)^{n-k}$
- $P(X \geq 16) = P(X = 16) + P(X = 17)$
  $= 0.00014$
- Conclusion: there is an effect, reject H_o

Real data: the acquisition of the weak-strong distinction

- Weak vs. strong quantifiers:
  - There are *many* PhD students in the room (weak)
  - *There is/are *every/all* students in the room (strong)

- The Dutch quantifier *allemaal*:
  - **Er vliegen allemaal papegaaien**
    There flying [allemaal] parrots
    “There are flying many parrots”
  - **De papegaaien vliegen allemaal**
    The parrots flying [allemaal]
    “The parrots are all flying”

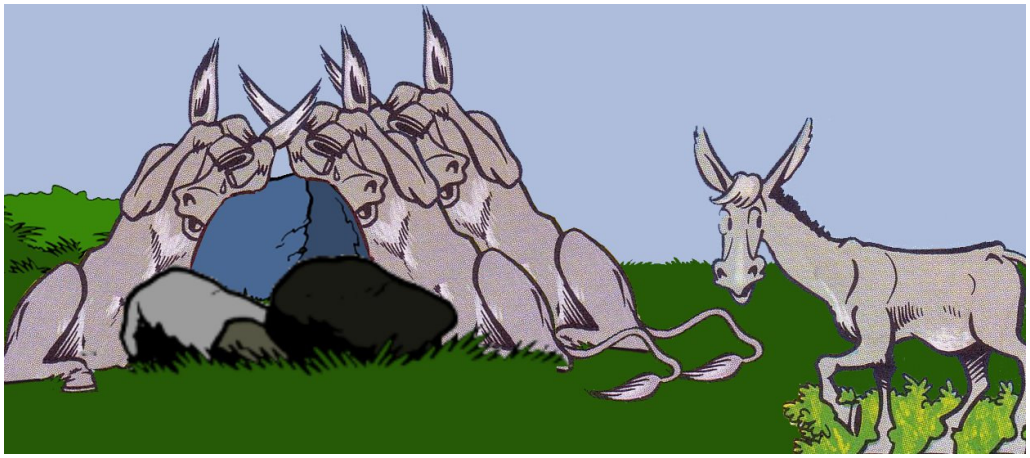
Experimental design

- **Question**: Is the interpretation of a weak quantified sentence (i.e. an existential sentence containing “allemaal”) of a child similar to the interpretation of an adult?

- Condition: syntactic position of the quantifier (prenominal or floated)
- 39 subjects (aged 4 - 6)
- 7 adults (control group)
- Method: Truth Value Judgment Task
- Total of test sentences: 18 (12 test items, 3 no-fillers, 3 yes-fillers)

Testitems

- De ezels huilen allemaal
  (The donkeys are all crying)



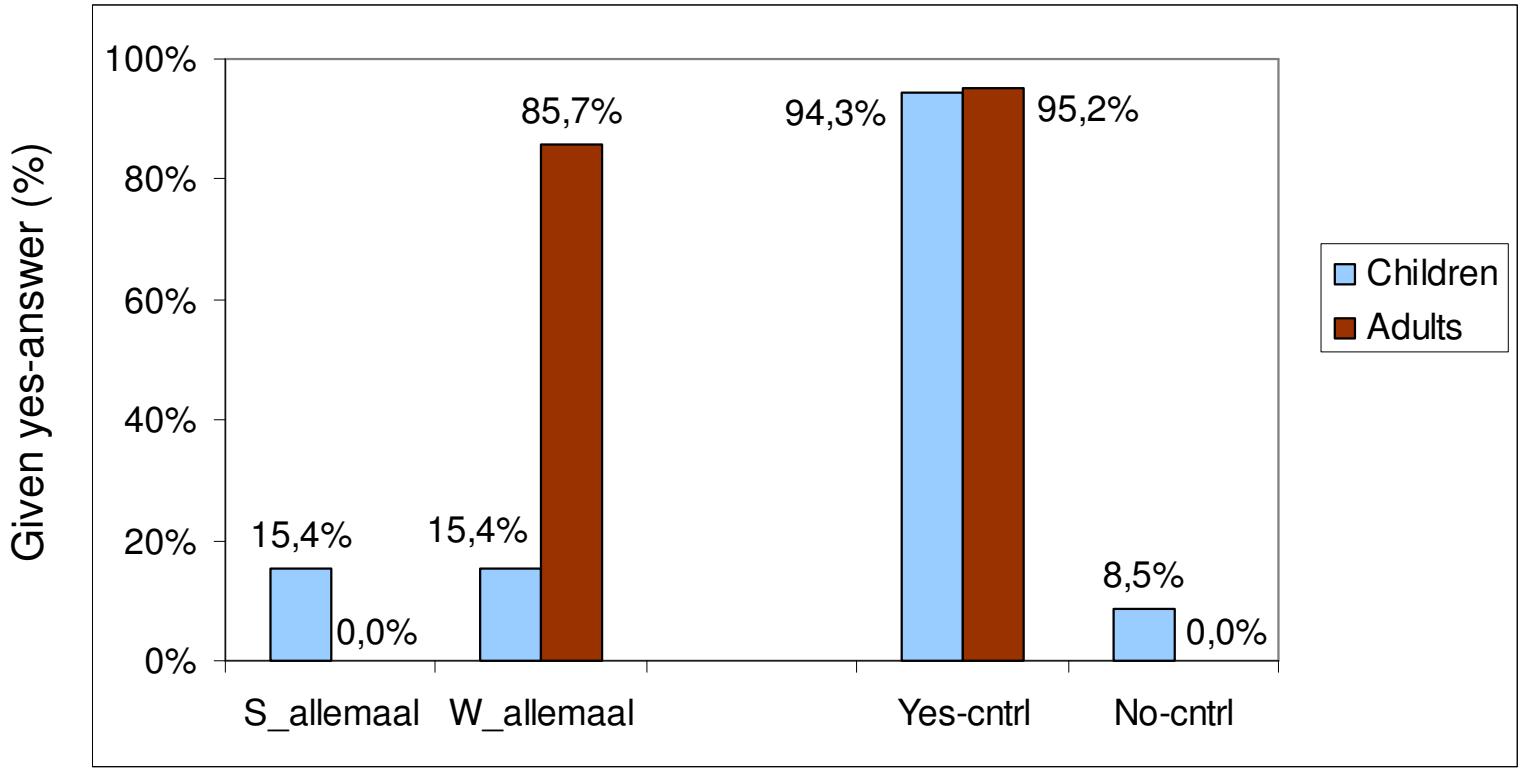
- Adult answer: no

Testitems

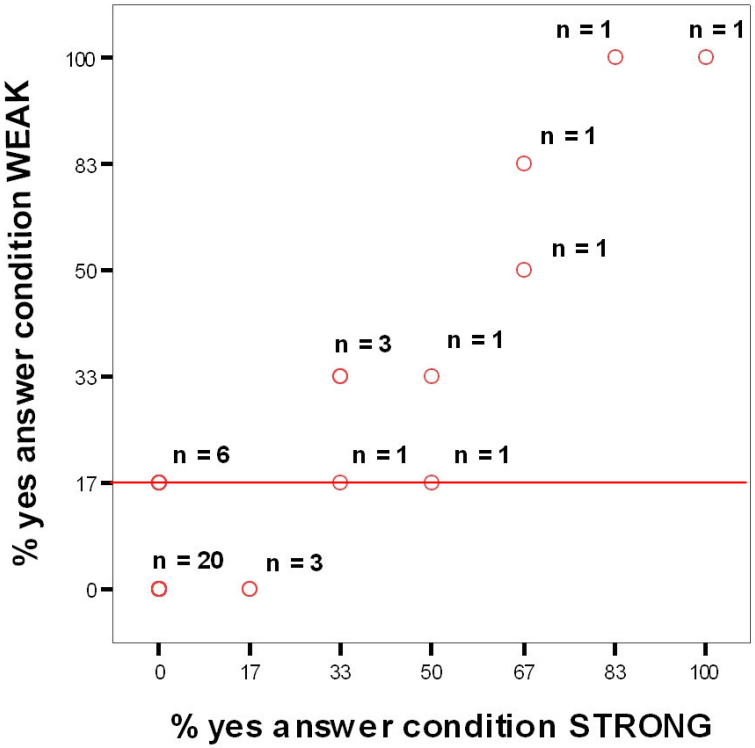
- Er dansen allemaal meisjes
  (There are dancing many girls)

- Adult answer: yes

Results


Given yes-answer (%)
100%
80%
60%
40%
20%
0%
15,4%
0,0%
15,4%
85,7%
94,3%
95,2%
8,5%
0,0%
S_allemaal
W_allemaal
Yes-cntrl
No-cntrl
Children
Adults

Results (2)



- Er dansen allemaal meisjes (There are dancing many girls)
- Adult answer: yes
- Child answer: no

Binomial test

	Pair										
	1	**2**	**3**	**…**	**5**	**6**	**7**	**…**	**9**	**…**	**11**
Yes	6	5	4		0	1	2		3		5
No	0	1	2		6	5	4		3		1
	+	+	+		-	-	-		0		+

Results:

+	3 children
0	1 child
-	35 children

Analysis weak quantified sentence

Case:

39 children are asked to analyze the quantifier *allemaal* as either strong or weak. Null hypothesis: no difference between adults and children.

Results:

35 children analyze a weak quantifier as a strong one, 3 children behave adult-like (i.e. say yes)

Question:

Do children analyze weak quantifiers significantly different as adults (accept in 0.86% of the cases the non-exhaustive picture as describing a weak quantified sentence)?

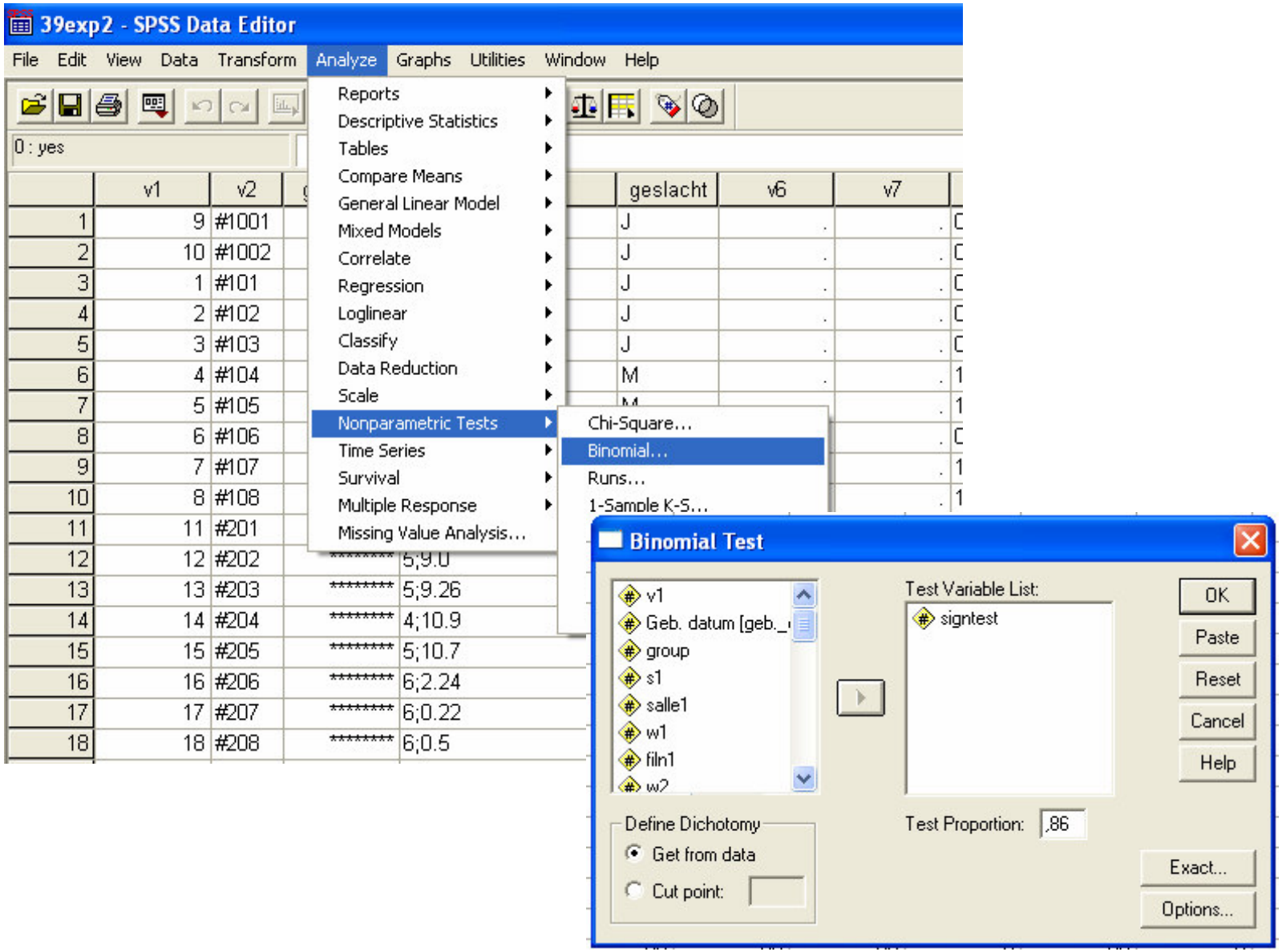
Binomial test

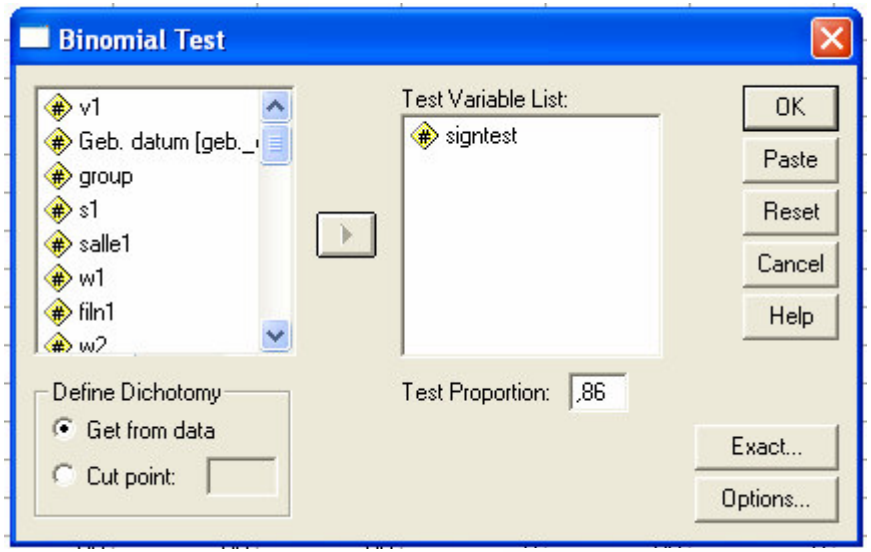
- Null-hypothesis of “no difference between adults and children’’ is:

  $H_0 = p = 0.86$

  $H_a = p > 0.86$
- X (3) has the B(3,0.86) distribution
- $P(X = k) = (n/k)\, p^k (1 - p)^{n-k}$
- $P(X = 3) = (38/3)(0.86)^3 (0.14)^{35}$

  $= 0{,}000$
- Conclusion: there is an effect, reject H_o


39exp2 - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Chi-Square...
Binomial...
Runs...
1-Sample K-S...
0 : yes
v1 v2 geslacht v6 v7
Binomial Test
v1
Geb. datum [geb._
group
s1
salle1
w1
filn1
w2
Test Variable List:
signtest
OK
Paste
Reset
Cancel
Help
Define Dichotomy
Get from data
Cut point:
Test Proportion: ,86
Exact...
Options...



NPar Tests

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
SIGNTEST	39	,0769	,26995	,00	1,00

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (1-tailed)
SIGNTEST	Group 1	ja	3	,08	,86	,000[a,]
	Group 2	nee	36	,92		
	Total		39	1,00		

a. Alternative hypothesis states that the proportion of cases in the first group < ,86.

b. Based on Z Approximation.

CLITIC PRODUCTION IN ITALIAN

ACCUSATIVE CLITICS: THE OPTIONAL CONDITION

Maria vuole mangiare *la mela*$_{NP}$

Maria vuole mangiar***la***

Maria ***la*** vuole mangiare

Maria wants to eat *the apple*$_{NP}$

Maria wants to eat ***it***

Maria ***it*** wants to eat

Research questions:

- Research question 1:
  - Will agrammatic patients produce less object clitics than normal controls?

- Research question 2:
  - In the optional condition will agrammatic subjects prefer to leave clitics at the original site, i.e. in the place where they are originated, or will they prefer to move them before the verbal complex?
  - Will this pattern differ from the one that normal controls will show?

Task:

- Sentence completion task

Maria la vuole mangiare, invece Gianni…

- …non *la* vuole mangiare
- …non vuole mangiar*la*

} Possible outcomes

Maria non vuole leggerlo, invece Gianni…

- … vuole legger*lo*
- …*lo* vuole leggere

} Possible outcomes

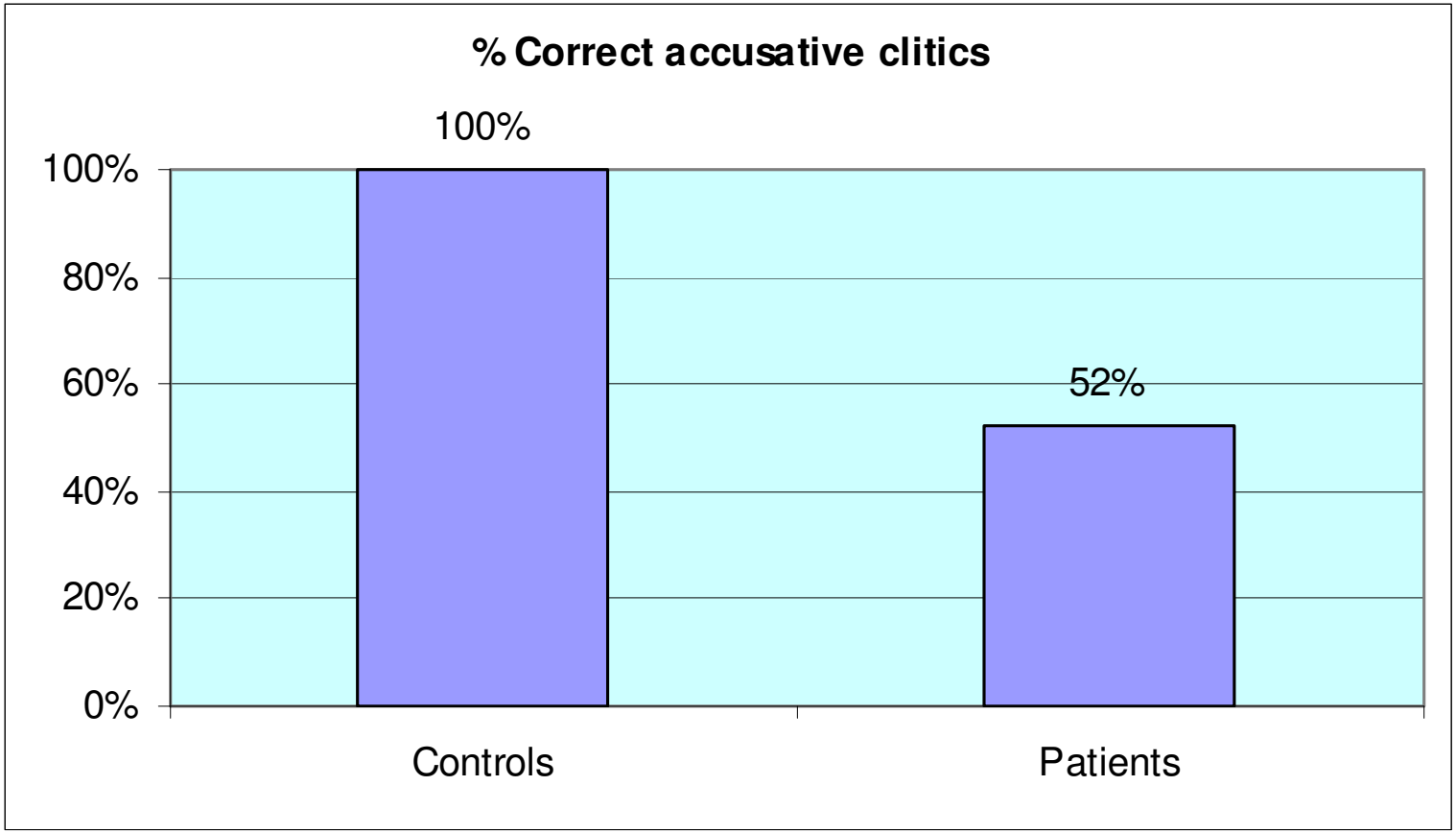
Design of the experiment:

- For each test there are 30 stimuli sentences: 15 with the clitic moved and 15 with the clitic at the base position.

Subjects:

- Two Italian agrammatic speakers
- Three Italian non brain damaged speakers.

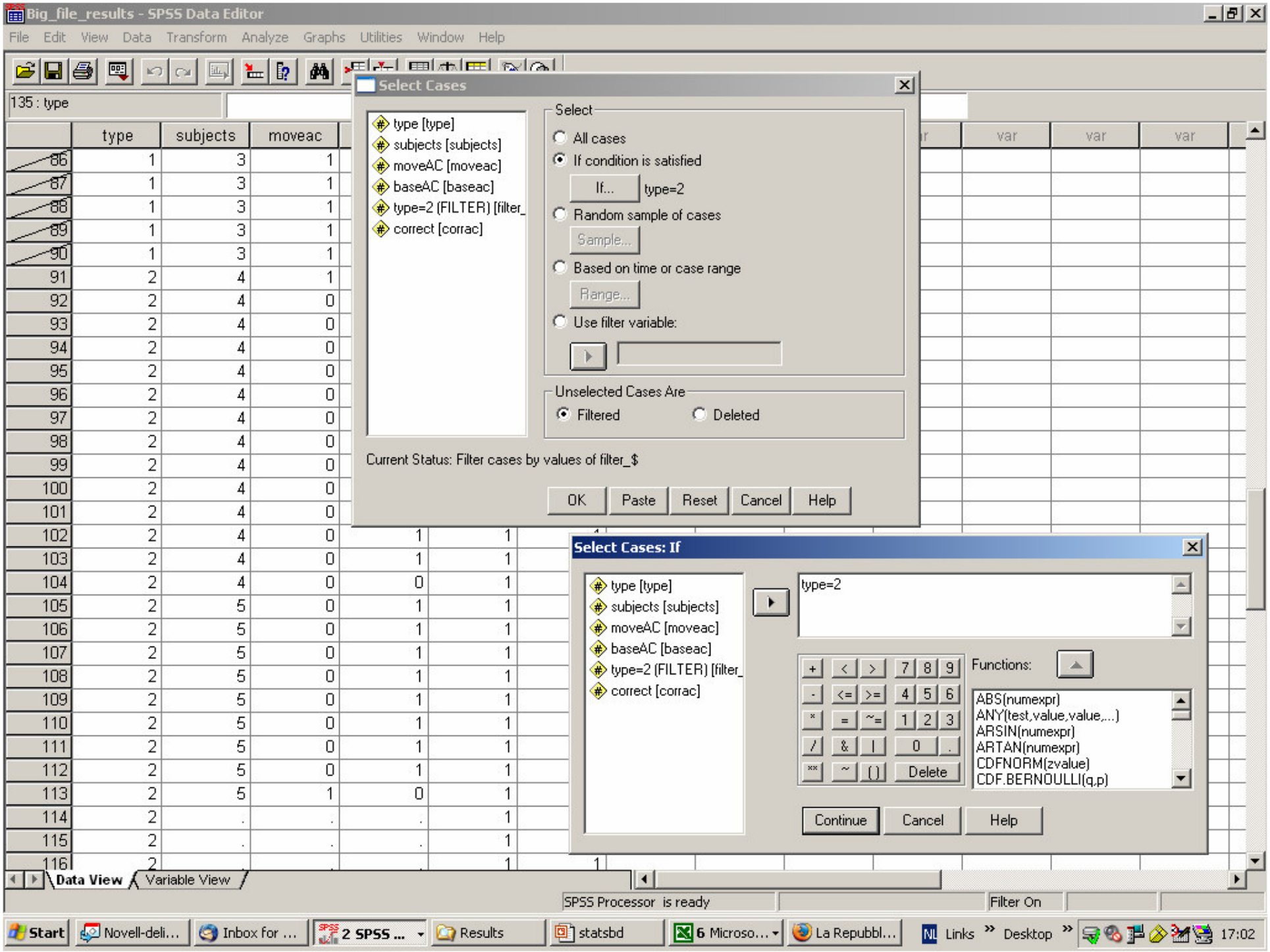
Results 1: Correctly produced clitics

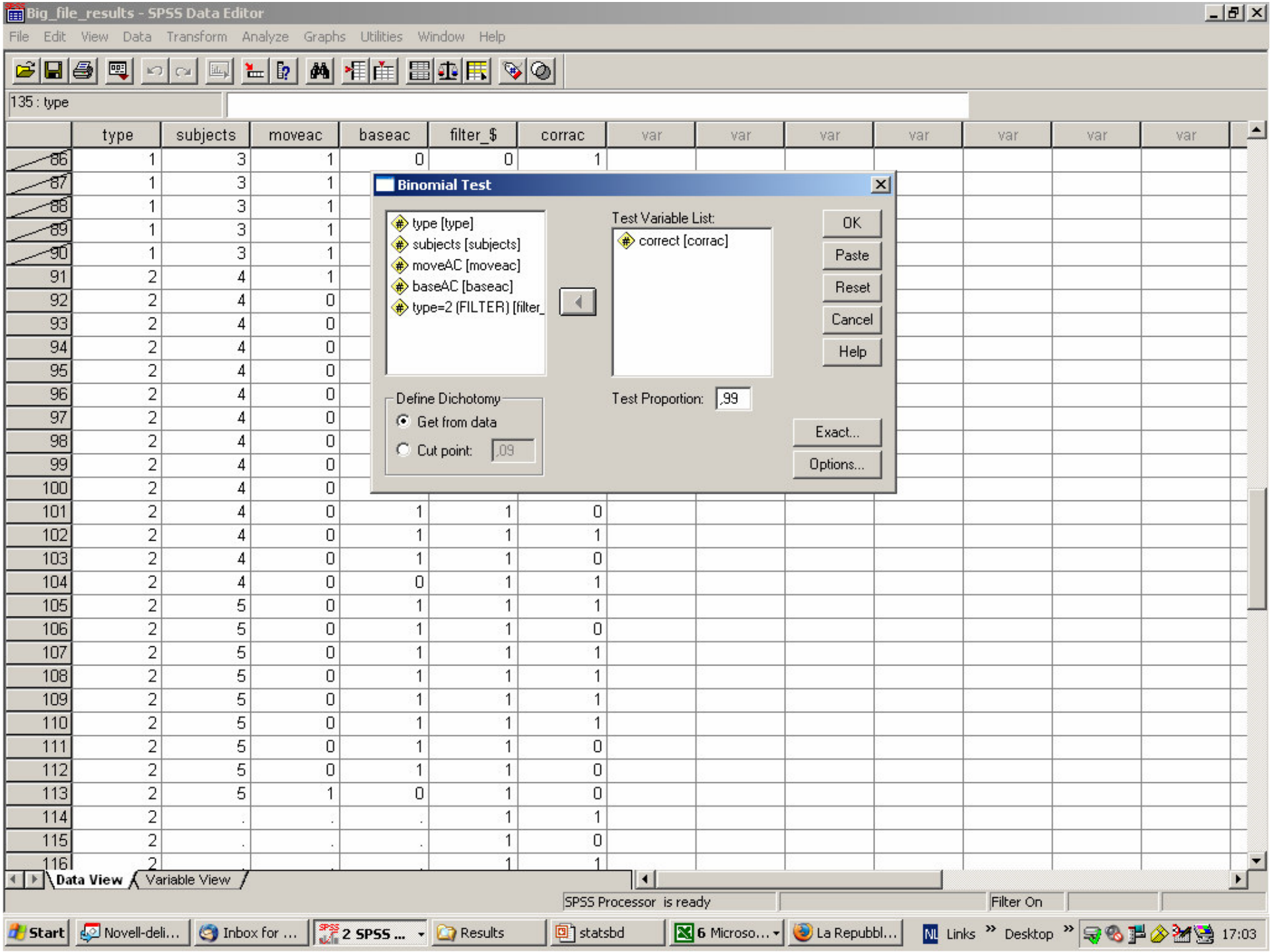


Significance test

- In this case we will use a ***Binomial Test***
- We know from the descriptive statistics that normal controls score at ceiling, i.e. (100%)
- We have then to contrast the performance of our patients against the performance of the controls.

$$H_0 = Correct_P = 1$$
$$H_a = Correct_P < 1$$


Big_file_results - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
135 : type
type subjects moveac var var var
Select Cases
type [type]
subjects [subjects]
moveAC [moveac]
baseAC [baseac]
type=2 (FILTER) [filter_
correct [corrac]
Select
All cases
If condition is satisfied
If... type=2
Random sample of cases
Sample...
Based on time or case range
Range...
Use filter variable:
Unselected Cases Are
Filtered
Deleted
Current Status: Filter cases by values of filter_$
OK Paste Reset Cancel Help
Select Cases: If
type [type]
subjects [subjects]
moveAC [moveac]
baseAC [baseac]
type=2 (FILTER) [filter_
correct [corrac]
type=2
Functions:
ABS(numexpr)
ANY(test,value,value,...)
ARSIN(numexpr)
ARTAN(numexpr)
CDFNORM(zvalue)
CDF.BERNOULLI(q,p)
Delete
Continue Cancel Help
Data View Variable View
SPSS Processor is ready
Filter On
Start Novell-deli... Inbox for ... 2 SPSS ... Results statsbd 6 Microso... La Repubbl... Links Desktop 17:02


Big_file_results - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
135 : type
type subjects moveac baseac filter_$ corrac var var var var var var var
86 1 3 1 0 0 1
87 1 3 1
88 1 3 1
89 1 3 1
90 1 3 1
91 2 4 1
92 2 4 0
93 2 4 0
94 2 4 0
95 2 4 0
96 2 4 0
97 2 4 0
98 2 4 0
99 2 4 0
100 2 4 0
101 2 4 0 1 1 0
102 2 4 0 1 1 1
103 2 4 0 1 1 0
104 2 4 0 0 1 1
105 2 5 0 1 1 1
106 2 5 0 1 1 0
107 2 5 0 1 1 1
108 2 5 0 1 1 1
109 2 5 0 1 1 1
110 2 5 0 1 1 1
111 2 5 0 1 1 0
112 2 5 0 1 1 0
113 2 5 1 0 1 0
114 2 . . . 1 1
115 2 . . . 1 0
116 2 . . . 1 1
Binomial Test
type [type]
subjects [subjects]
moveAC [moveac]
baseAC [baseac]
type=2 (FILTER) [filter_
Test Variable List:
correct [corrac]
OK
Paste
Reset
Cancel
Help
Define Dichotomy
Get from data
Cut point: ,09
Test Proportion: ,99
Exact...
Options...
Data View
Variable View
SPSS Processor is ready
Filter On
Start
Novell-deli...
Inbox for ...
2 SPSS ...
Results
statsbd
6 Microso...
La Repubbl...
Links
Desktop
17:03

Result

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (1-tailed)
correct	Group 1	correct	23	,52	,99	,000[a,b]
	Group 2	wrong	21	,48		
	Total		44	1,00		

a. Alternative hypothesis states that the proportion of cases in the first group < ,99.

b. Based on Z Approximation.

- Because p=0.000 we can reject H_0 and confirm H_a
- Patients produce less correct object clitic than normal controls do.

H_0= $Correct_P$=1

H_a= $Correct_P$<1

The position of the clitics

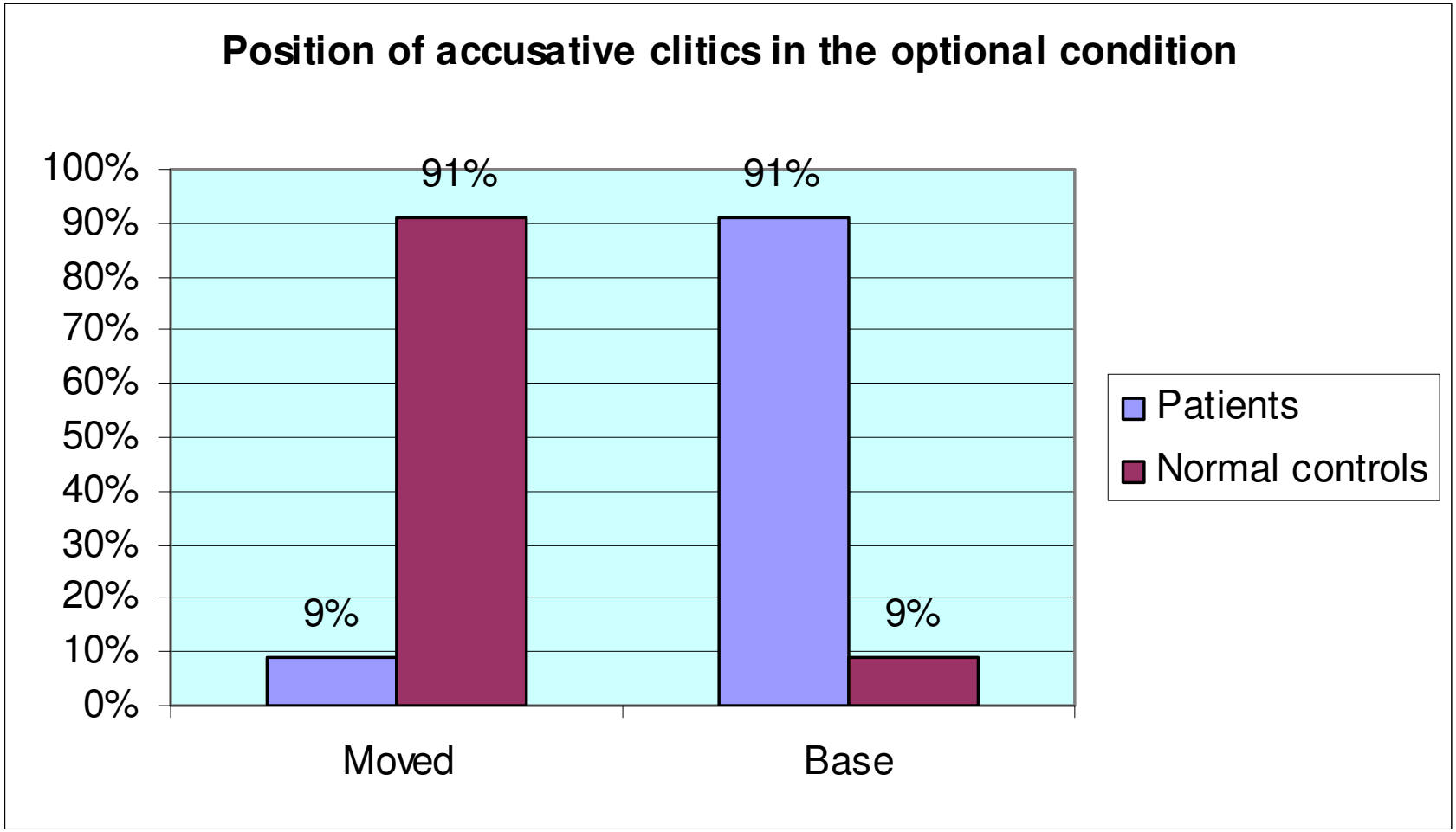
The optional condition

Rationale behind it!
The stimuli in the test

- 50% stimuli:
- Maria ***it*** wants to eat but Gianni…
  - Not ***it*** wants to eat
  - Not wants to eat ***it***
- 50% stimuli:
- Maria not wants to eat ***it*** but Gianni..
  - ***It*** wants to eat
  - Wants to eat ***it***

- ***If the position of the clitics would follow the stimuli, I expect a 50% 50% distribution in the procuction of the clitics:***
  - ***50% moved (when the stimuli prompted that structure)***
  - ***50% based (when the stimuli prompted that structure)***

Results 2: Position of the accusative clitic



Remember the research questions!

In the optional condition will agrammatic subjects prefer to leave clitics at the original site, i.e. in the place where they are originated, or will they prefer to move them before the verbal complex?

Will this pattern differ from the one that normal controls will show?

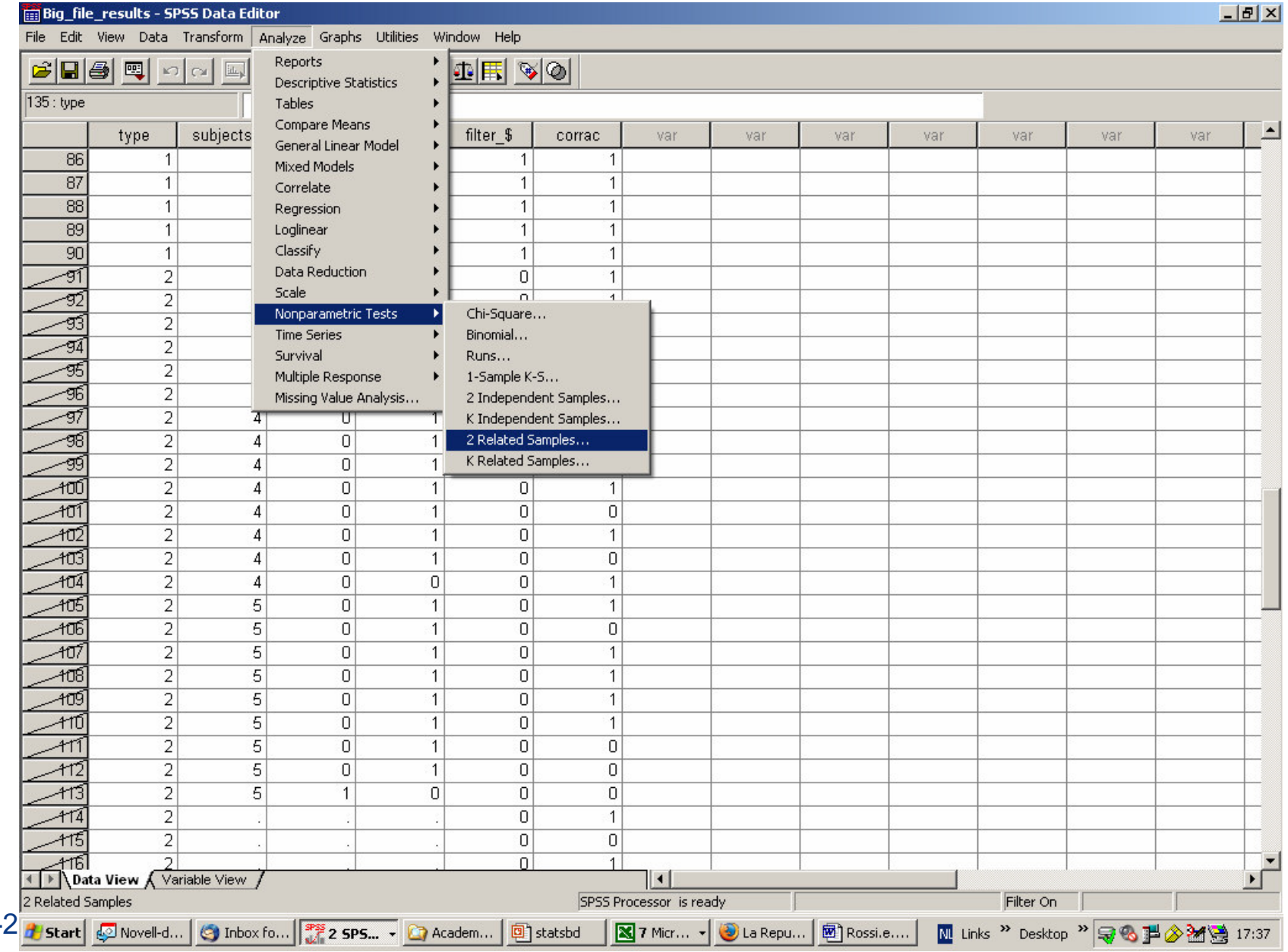
Analysis 1

In this case a *Sign Test* is suitable for our analysis

We will compare the distribution of two related samples for each group (controls and patients):

MOVED vs. BASED

Normal controls



Frequencies

		N
baseAC - moveAC	Negative Differences[a]	82
	Positive Differences[b]	8
	Ties[c]	0
	Total	90

a. baseAC < moveAC

b. baseAC > moveAC

c. baseAC = moveAC

Normal controls produce more moved accusative clitics than not based clitics. (see a.)
Our a priori distribution is not confirmed.

Test Statistics[a]

	baseAC - moveAC
Z	-7,695
Asymp. Sig. (2-tailed)	,000

a. Sign Test

Patients

Frequencies

		N
baseAC - moveAC	Negative Differences[a]	1
	Positive Differences[b]	20
	Ties[c]	2
	Total	23

a. baseAC < moveAC

b. baseAC > moveAC

c. baseAC = moveAC

Patients produce more based accusative clitics than not moved clitics. (see b.)
Our a priori distribution is not confirmed.

Test Statistics[b]

	baseAC - moveAC
Exact Sig. (2-tailed)	,000[a]

a. Binomial distribution used.

b. Sign Test

Are patients behaving differently than normals? Significance test

- In this case we will use a ***Binomial Test***
- We know from the descriptive statistics that normal controls move the accusative clitics in 91% of the cases. (0.91 will be our test proportion)
- We have then to contrast the performance of our patients against the performance of the controls.

$$H_0 = Move_P = 0.91$$
$$H_a = Move_P < 0.91$$

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (1-tailed)
baseAC	Group 1	<= ,91	2	,09	,91	,000[a]
	Group 2	> ,91	21	,91		
	Total		23	1,00		

a. Alternative hypothesis states that the proportion of cases in the first group < ,91.

P=0.000

We can reject H_o and confirm H_a

H_0= $Move_P$=0.91

H_a= $Move_P$<0.91