Subject

The algorithms, which are implemented in TANAGRA, come from the famous Siegel & Castellan's book **(Sidney SIEGEL and John CASTELLAN, « Nonparametric Statistics for the Behavioral Sciences », McGraw-Hill, 1988)**. We show how to use theses statistical tests on a dataset.

Dataset

Our dataset (NONPARAMETRIC_STATISTICS_DATASET.XLS) describes 300 households with 5 attributes (wage of man, wage of woman, household income [wage man + wage woman], housing and have a house with a garden).

The examples are more or less funny, the most important is how to use the soft and how to read the results.

Nonparametric Statistics

Download the dataset

First set, we must create a new diagram and import the dataset (FILE / NEW).

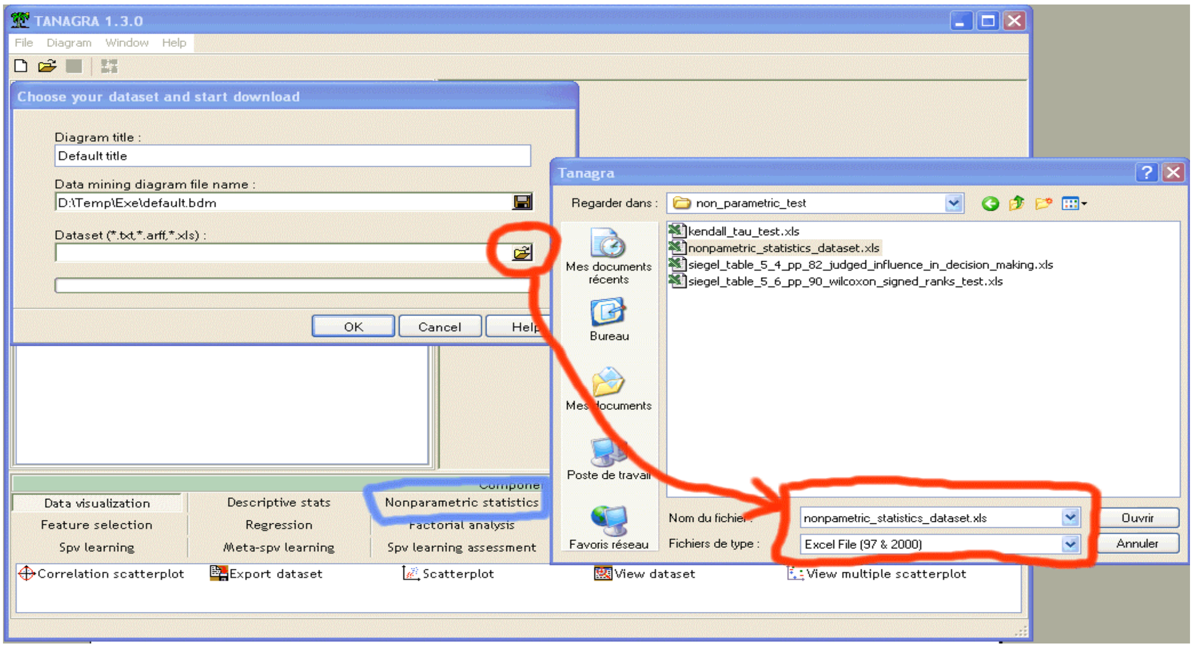


Diagram title : Default title

Data mining diagram file name : D:\Temp\Exe\default.bdm

Dataset (*.txt,*.arff,*.xls) :

Regarder dans : non_parametric_test

kendall_tau_test.xls
nonpametric_statistics_dataset.xls
siegel_table_5_4_pp_82_judged_influence_in_decision_making.xls
siegel_table_5_6_pp_90_wilcoxon_signed_ranks_test.xls

Nom du fichier : nonpametric_statistics_dataset.xls

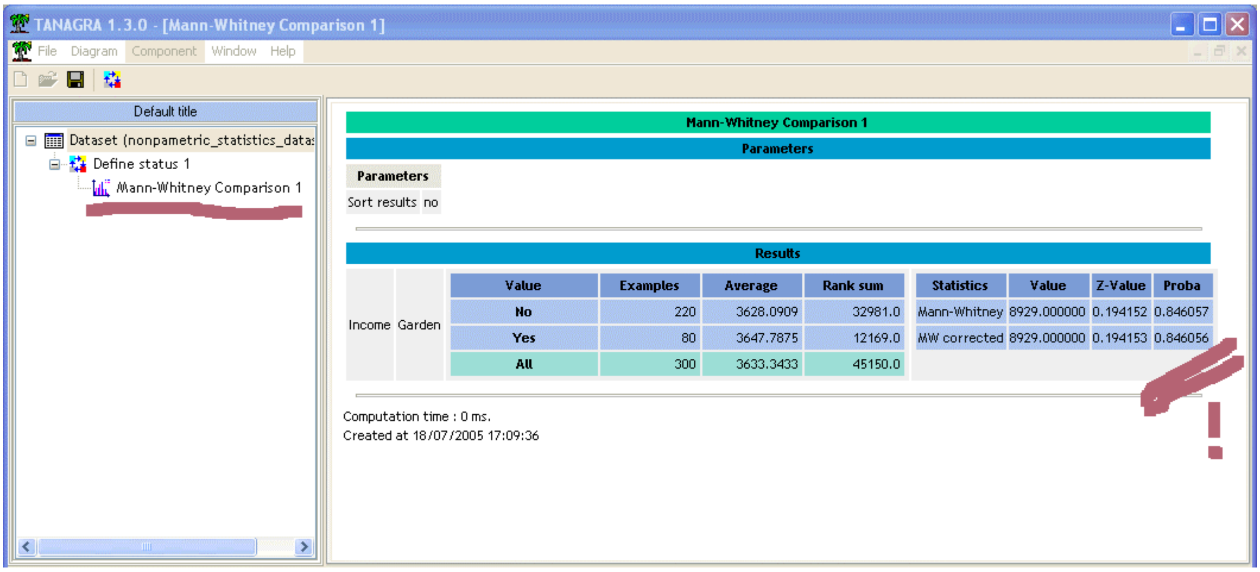
Fichiers de type : Excel File (97 & 2000)

Studying differences between two independent groups

We want to compare the total income of households (INCOME) according to whether they occupy a house with garden or not (GARDEN - 2 values).

We add a DEFINE STATUS component in the diagram; we set INCOME as TARGET and GARDEN as INPUT. Various methods can be used, we give preference to methods that work only on comparison with two independent samples in this section: MANN &WHITNEY and WALD & WOLFOWITZ.

MANN & WHITNEY test uses the sum of ranks (Siegel & Castellan, pages 128 to 137), the Z-value is asymptotically normally distributed, and the corrected statistic takes into account ties. In our dataset, we see that the data do not give evidence, which justify the rejecting H0 (the INCOME of households is the same whatever the value of GARDEN) for a level of significance at 5%.


TANAGRA 1.3.0 - [Mann-Whitney Comparison 1]
File
Diagram
Component
Window
Help
Default title
Dataset (nonparametric_statistics_datas
Define status 1
Mann-Whitney Comparison 1

Mann-Whitney Comparison 1

Parameters

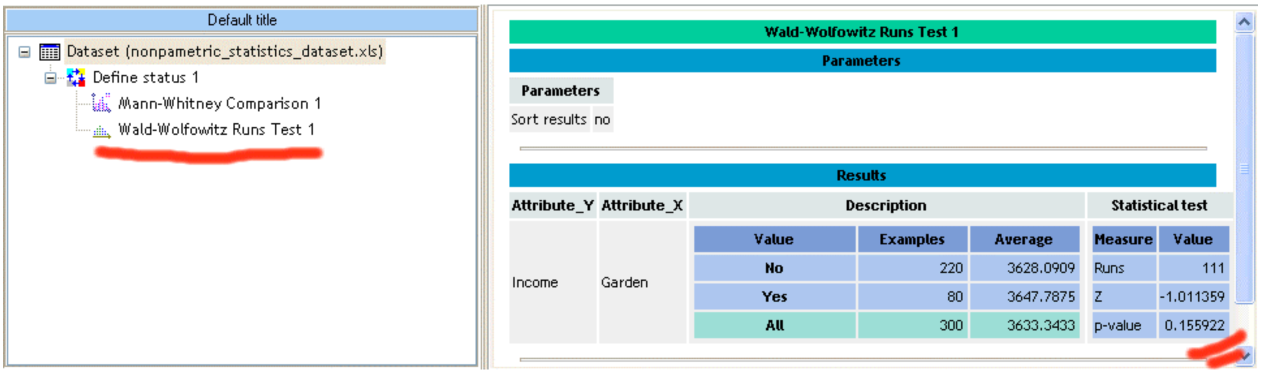
Parameters	
Sort results	no

Results

		Value	Examples	Average	Rank sum	Statistics	Value	Z-Value	Proba
Income	Garden	No	220	3628.0909	32981.0	Mann-Whitney	8929.000000	0.194152	0.846057
		Yes	80	3647.7875	12169.0	MW corrected	8929.000000	0.194153	0.846056
		All	300	3633.3433	45150.0				

Computation time : 0 ms.
Created at 18/07/2005 17:09:36

WALD & WOLFOWITZ test uses the number of runs (Siegel & Castellan, pages 58 to 64), the Z-value is asymptotically normally distributed. The results are coherent with the previous test. Note that we use a one-tailed test for groups' comparison.


Default title
Dataset (nonparametric_statistics_dataset.xls)
Define status 1
Mann-Whitney Comparison 1
Wald-Wolfowitz Runs Test 1

Wald-Wolfowitz Runs Test 1

Parameters

Parameters	
Sort results	no

Results

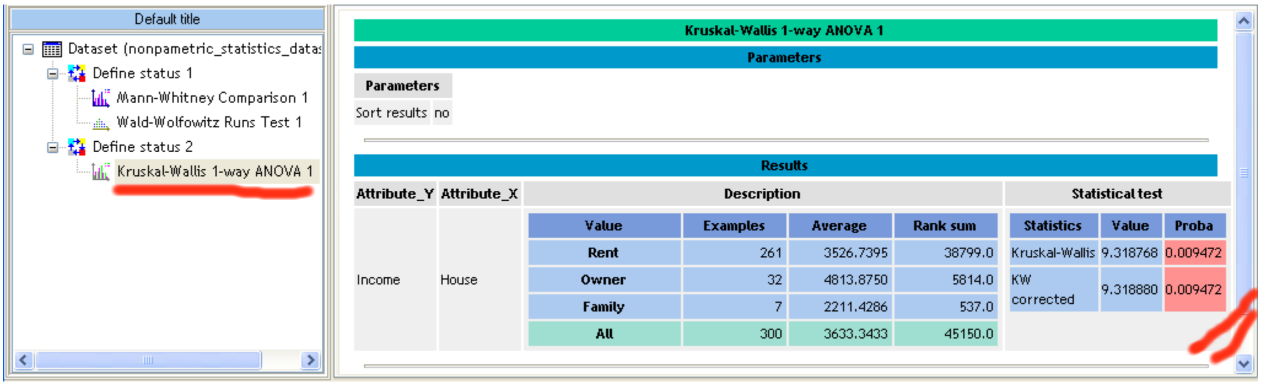
Attribute_Y	Attribute_X	Description			Statistical test	
		Value	Examples	Average	Measure	Value
Income	Garden	No	220	3628.0909	Runs	111
		Yes	80	3647.7875	Z	-1.011359
		All	300	3633.3433	p-value	0.155922

Differences between K independent samples

For K populations (K > 2), the previous test cannot be used; the KRUSKAL & WALLIS test is more appropriate.

We add a new DEFINE STATUS component at the root of the diagram; we set INCOME as TARGET and HOUSE (3 values) as INPUT. We want to determine if the INCOME of households is different according of the status of their house (“owner”, “rent”, “family”).

KRUSKALL & WALLIS can be seen as a generalization of MANN & WHITNEY, the statistic is approximated by the CHI-2 distribution with degree of freedom = K-1 (Siegel & Castellan, pages 206 to 212). We can introduce a correction when ties occur between two or more examples.


Default title
Dataset (nonparametric_statistics_data:
Define status 1
Mann-Whitney Comparison 1
Wald-Wolfowitz Runs Test 1
Define status 2
Kruskal-Wallis 1-way ANOVA 1

Kruskal-Wallis 1-way ANOVA 1

Parameters

Parameters	
Sort results	no

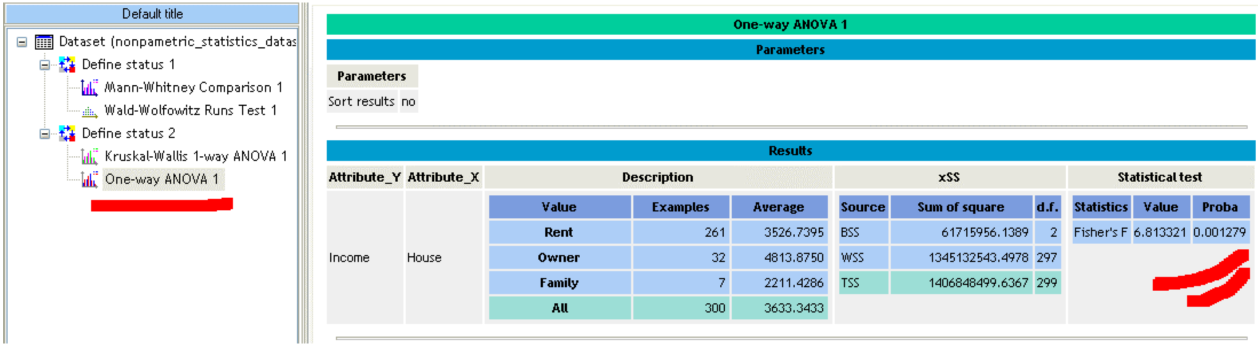
Results

Attribute_Y	Attribute_X	Description				Statistical test		
		Value	Examples	Average	Rank sum	Statistics	Value	Proba
Income	House	Rent	261	3526.7395	38799.0	Kruskal-Wallis	9.318768	0.009472
		Owner	32	4813.8750	5814.0	KW corrected	9.318880	0.009472
		Family	7	2211.4286	537.0			
		All	300	3633.3433	45150.0			

For a level of significance at 5%, we reject the null hypothesis: the hypothesis of no difference in INCOME according to HOUSE status is rejected. The computed average shows that the OWNER of their houses have a higher income than the others (OWNER # 4813, RENT # 3526, FAMILY # 2211).

There is a generalization of runs test (A.MOOD, « The distribution theory of runs », Ann. of Math. Stat., 11, pp. 367-392, 1940), but this method being implemented nowhere (???), it was not possible to make comparisons, we add it into TANAGRA but the results are to be taken with cautions.

ONE WAY ANOVA is a parametric alternative of the previous tests. We obtain the same conclusion.


Default title
Dataset (nonparametric_statistics_datas
Define status 1
Mann-Whitney Comparison 1
Wald-Wolfowitz Runs Test 1
Define status 2
Kruskal-Wallis 1-way ANOVA 1
One-way ANOVA 1

One-way ANOVA 1

Parameters

Parameters	
Sort results	no

Results

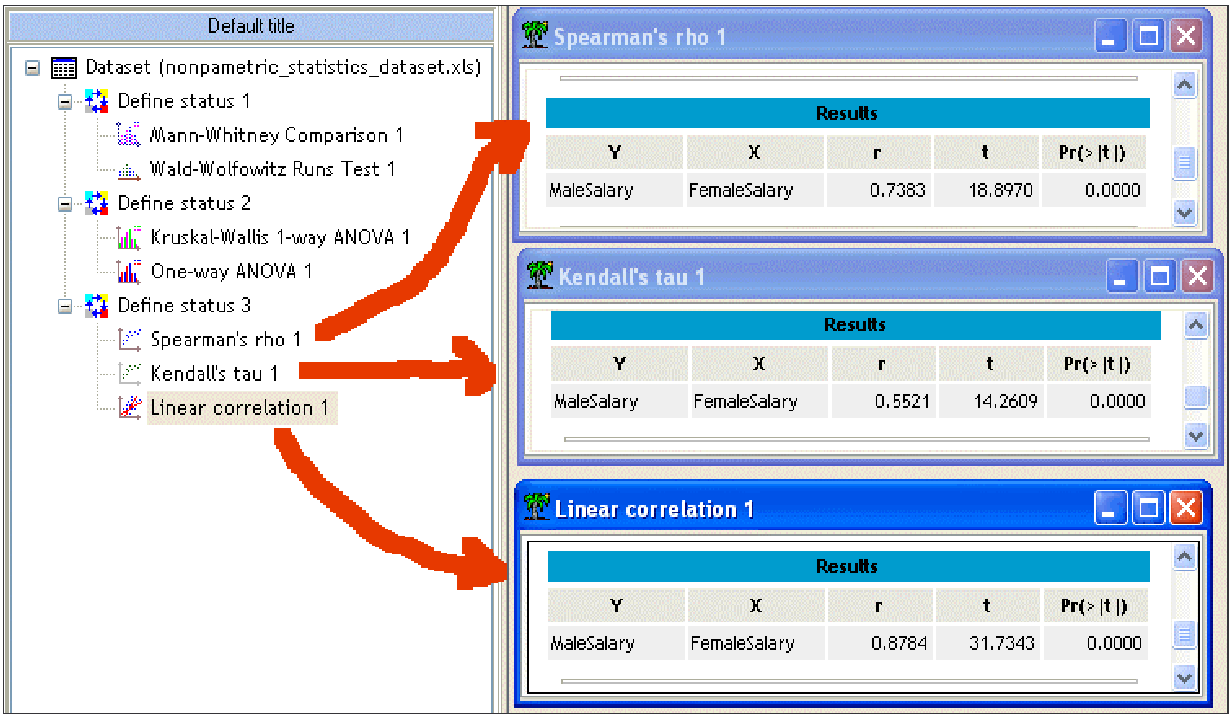
Attribute_Y	Attribute_X	Description			xSS			Statistical test		
		Value	Examples	Average	Source	Sum of square	d.f.	Statistics	Value	Proba
Income	House	Rent	261	3526.7395	BSS	61715956.1389	2	Fisher's F	6.813321	0.001279
		Owner	32	4813.8750	WSS	1345132543.4978	297			
		Family	7	2211.4286	TSS	1406848499.6367	299			
		All	300	3633.3433						

Correlation

We study another problem in this section: is the man and the woman, within a household, have similar salary? We use a test of correlation to evaluate this assumption.

We add a new DEFINE STATUS component at the root of the diagram. Set SALAIRE HOMME (man wage) as TARGET, and SALAIRE FEMME (woman wage) as INPUT. We use two components: the “SPEARMAN Rank Order Correlation Coefficient” and the “KENDALL Rank Order Correlation Coefficient” (Siegel & Castellan, pages 235 to 254).

We see that the results of both methods are very similar, with the level of significance at 5%, the man and the woman within a household have similar wages.


Default title
Dataset (nonparametric_statistics_dataset.xls)
Define status 1
Mann-Whitney Comparison 1
Wald-Wolfowitz Runs Test 1
Define status 2
Kruskal-Wallis 1-way ANOVA 1
One-way ANOVA 1
Define status 3
Spearman's rho 1
Kendall's tau 1
Linear correlation 1

Spearman's rho 1

Results

Y	X	r	t	Pr(> |t |)
MaleSalary	FemaleSalary	0.7383	18.8970	0.0000

Kendall's tau 1

Results

Y	X	r	t	Pr(> |t |)
MaleSalary	FemaleSalary	0.5521	14.2609	0.0000

Linear correlation 1

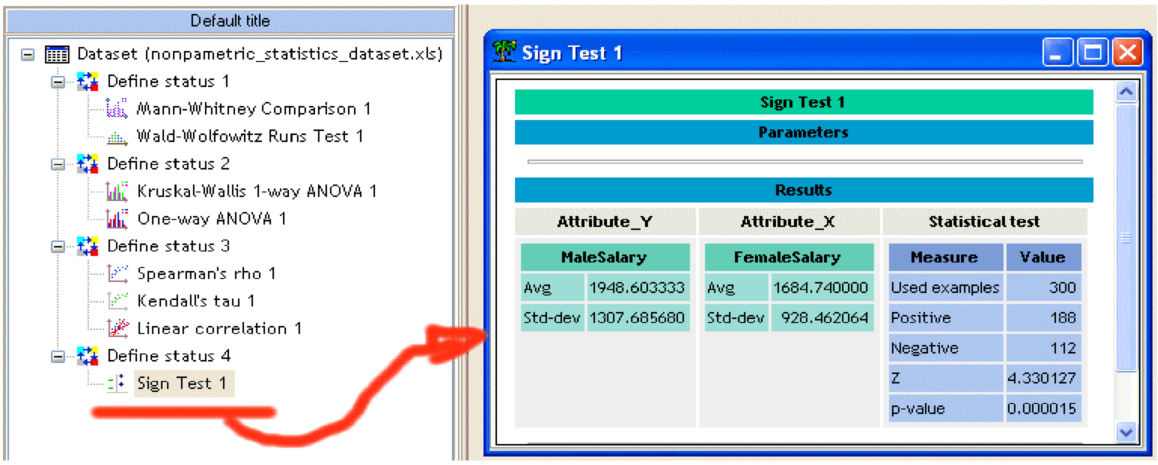
Results

Y	X	r	t	Pr(> |t |)
MaleSalary	FemaleSalary	0.8784	31.7343	0.0000

Comparisons on two related samples (Paired Replicates)

We want now to compare the salary of the man and the woman within a household. We have two measures on each example, two methods are available in TANAGRA: SIGN TEST (Siegel et Castellan, pages 80 to 87) and WILCOXON SIGNED RANK TEST (Siegel & Castellan, pages 87 to 95).

SIGN TEST consists in counting the number of times where the wages of the men are higher than that of the women, then to compare this value with the theoretical value under the null hypothesis "they have the same salary". We add a new DEFINE STATUS component at the root of the diagram and set SALAIRE HOMME as TARGET, and SALAIRE FEMME as INPUT, we add also the SIGN TEST component. The results show that on 300 studied households, the man has a higher salary on 188 cases. The Z-value is asymptotically normally distributed. With the level of significance at 5%, the salary of man is higher than the salary of woman within a household.



Default title

- Dataset (nonparametric_statistics_dataset.xls)
  - Define status 1
    - Mann-Whitney Comparison 1
    - Wald-Wolfowitz Runs Test 1
  - Define status 2
    - Kruskal-Wallis 1-way ANOVA 1
    - One-way ANOVA 1
  - Define status 3
    - Spearman's rho 1
    - Kendall's tau 1
    - Linear correlation 1
  - Define status 4
    - Sign Test 1

Sign Test 1

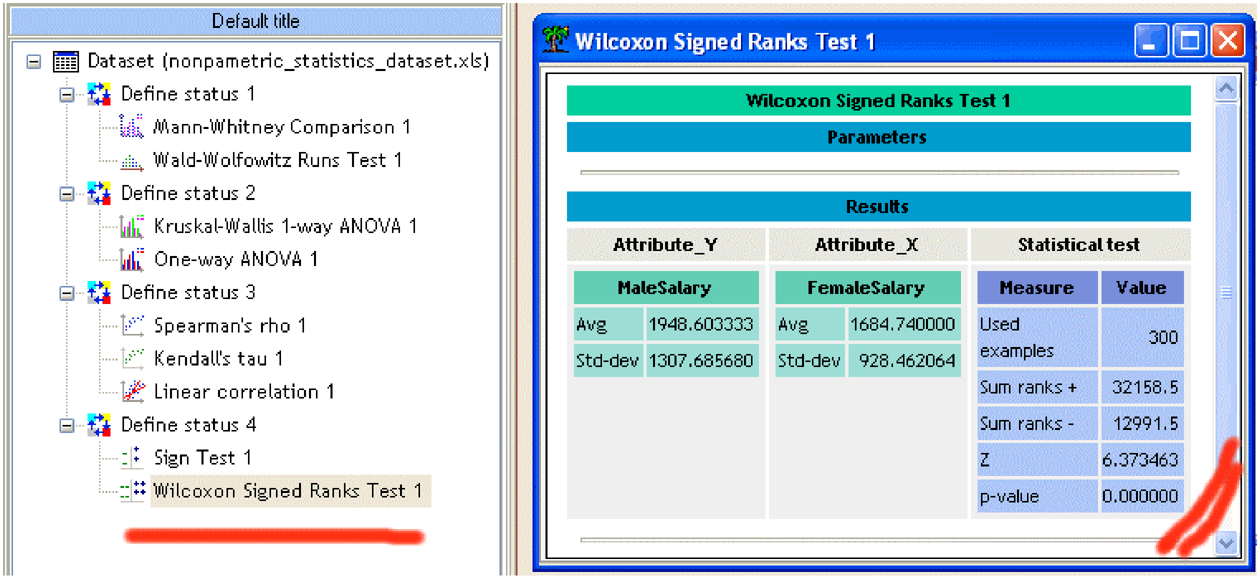
Sign Test 1

Parameters

Results

Attribute_Y		Attribute_X		Statistical test	
MaleSalary		FemaleSalary		Measure	Value
Avg	1948.603333	Avg	1684.740000	Used examples	300
Std-dev	1307.685680	Std-dev	928.462064	Positive	188
				Negative	112
				Z	4.330127
				p-value	0.000015

Sign test is very conservative; it does not take account of the magnitude of the difference. The WILCOXON SIGNED RANK TEST is more powerful. We add the component; we see that the previous result is strengthened.



Default title

- Dataset (nonparametric_statistics_dataset.xls)
  - Define status 1
    - Mann-Whitney Comparison 1
    - Wald-Wolfowitz Runs Test 1
  - Define status 2
    - Kruskal-Wallis 1-way ANOVA 1
    - One-way ANOVA 1
  - Define status 3
    - Spearman's rho 1
    - Kendall's tau 1
    - Linear correlation 1
  - Define status 4
    - Sign Test 1
    - Wilcoxon Signed Ranks Test 1

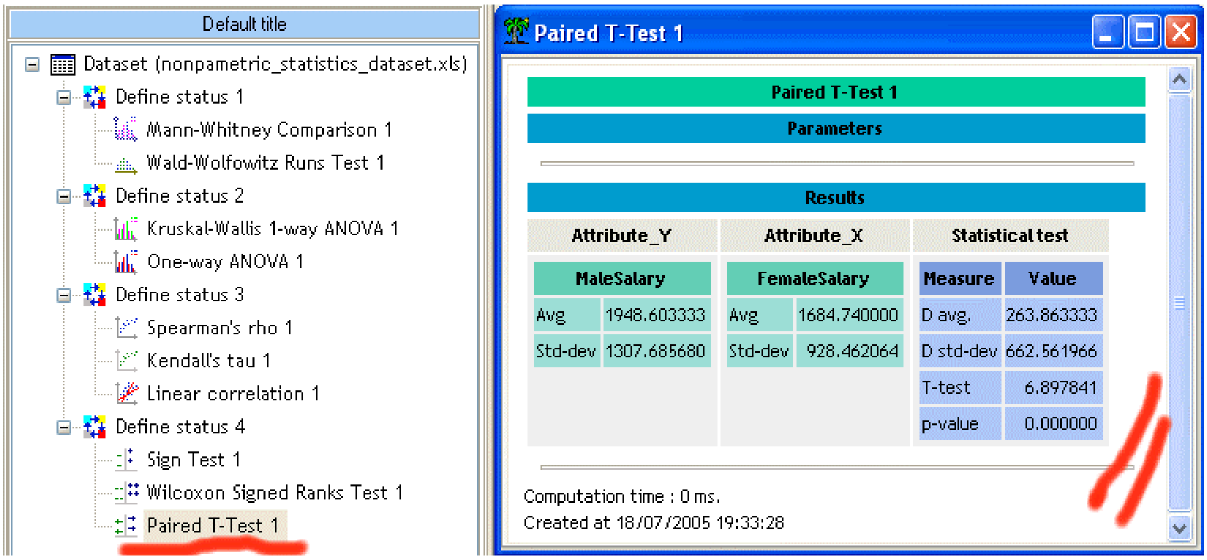
Wilcoxon Signed Ranks Test 1

Parameters

Results

Attribute_Y		Attribute_X		Statistical test	
MaleSalary		FemaleSalary		Measure	Value
Avg	1948.603333	Avg	1684.740000	Used examples	300
Std-dev	1307.685680	Std-dev	928.462064	Sum ranks +	32158.5
				Sum ranks -	12991.5
				Z	6.373463
				p-value	0.000000

There is a parametric alternative, the PAIRED STUDENT TEST that uses the value of the difference. We obtain the same conclusion.



Default title

- Dataset (nonparametric_statistics_dataset.xls)
  - Define status 1
    - Mann-Whitney Comparison 1
    - Wald-Wolfowitz Runs Test 1
  - Define status 2
    - Kruskal-Wallis 1-way ANOVA 1
    - One-way ANOVA 1
  - Define status 3
    - Spearman's rho 1
    - Kendall's tau 1
    - Linear correlation 1
  - Define status 4
    - Sign Test 1
    - Wilcoxon Signed Ranks Test 1
    - Paired T-Test 1

Paired T-Test 1

Parameters

Results

Attribute_Y		Attribute_X		Statistical test	
MaleSalary		FemaleSalary		Measure	Value
Avg	1948.603333	Avg	1684.740000	D avg.	263.863333
Std-dev	1307.685680	Std-dev	928.462064	D std-dev	662.561966
				T-test	6.897841
				p-value	0.000000

Computation time : 0 ms.
Created at 18/07/2005 19:33:28

All the methods implemented in TANAGRA were validated several manners: first of all we reproduced the examples described in our main reference (Siegel's book), that allowed to validate the detail of calculations; thereafter, we took several dataset and we compared our results with some popular commercial software.