Subject

A goodness-of-fit test is used to decide if a sample comes from a population with specific distribution. TANAGRA has a new component, which uses several tests in order to check the normality assumption.

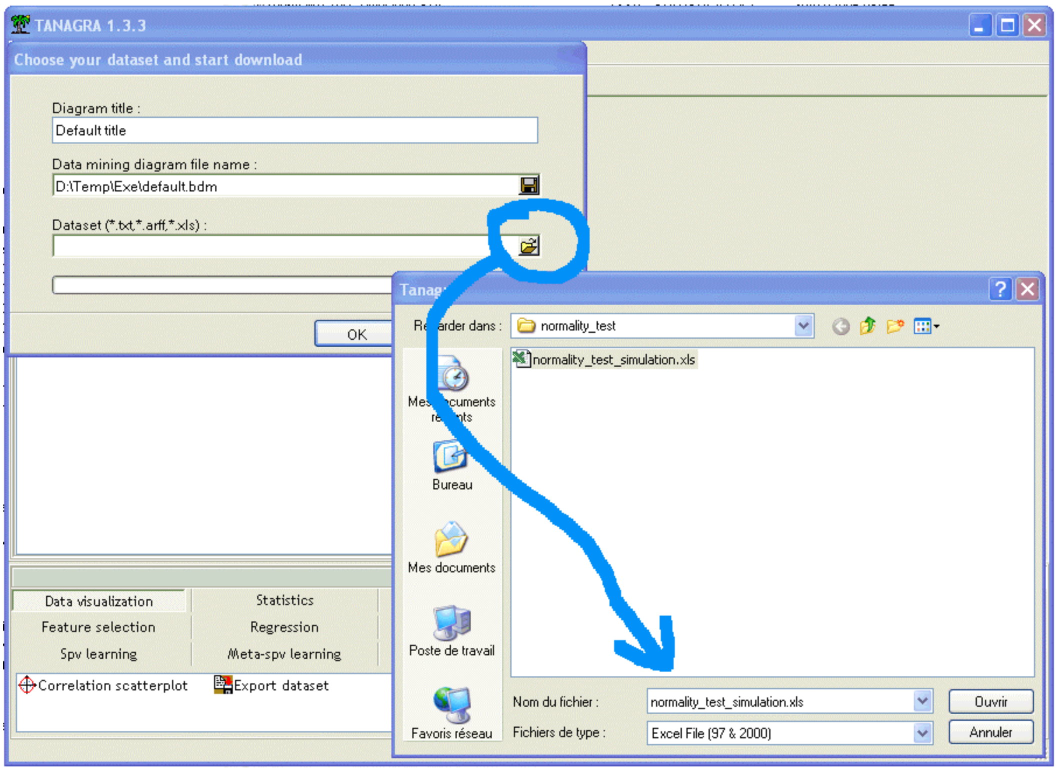
Dataset

The dataset NORMALITY_TEST_SIMULATION.XLS contains 500 examples, this is an artificial dataset generated from 3 distributions: uniform, normal and lognormal.

Test for normality

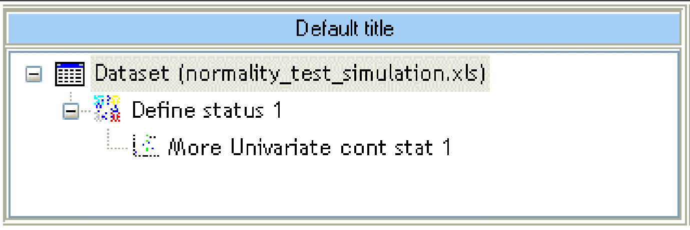
Download the dataset

First of all, create a new diagram and import the dataset (FILE / NEW).



Descriptive statistics

In the first time, we compute descriptive statistics in order to visualize the shape of the distribution. We add a DEFINE STATUS component in the diagram, and set all attributes as INPUT. Then, we add the MORE UNIVARIATE CONT STAT component from STATISTICS.



We obtain the following results.

Attribute	Stats		Histogram			
	Statistics		**Values**	**Count**	**Percent**	**Histogram**
UNIFORM	Average	0.5048	x_<_0.1018	40	8.00%	
	Median	0.5069	0.1018_=<_x_<_0.2016	57	11.40%	
	Std dev. [Coef of variation]	0.2838 [0.5622]	0.2016_=<_x_<_0.3014	55	11.00%	
	MAD [MAD/STDDEV]	0.2470 [0.8704]	0.3014_=<_x_<_0.4011	48	9.60%	
	Min * Max [Full range]	0.00 * 1.00 [1.00]	0.4011_=<_x_<_0.5009	46	9.20%	
	1st * 3rd quartile [Range]	0.26 * 0.75 [0.49]	0.5009_=<_x_<_0.6007	49	9.80%	
	Skewness	0.0193	0.6007_=<_x_<_0.7004	57	11.40%	
	Kurtosis	-1.2164	0.7004_=<_x_<_0.8002	49	9.80%	
			0.8002_=<_x_<_0.8999	55	11.00%	
			x>=_0.8999	44	8.80%	
	Statistics		**Values**	**Count**	**Percent**	**Histogram**
NORMAL	Average	0.0301	x_<_-2.2363	9	1.80%	
	Median	0.0174	-2.2363_=<_x_<_-1.6062	6	1.20%	
	Std dev. [Coef of variation]	0.9786 [32.4771]	-1.6062_=<_x_<_-0.9762	59	11.80%	
	MAD [MAD/STDDEV]	0.7801 [0.7971]	-0.9762_=<_x_<_-0.3461	115	23.00%	
	Min * Max [Full range]	-2.87 * 3.43 [6.30]	-0.3461_=<_x_<_0.2839	113	22.60%	
	1st * 3rd quartile [Range]	-0.64 * 0.68 [1.32]	0.2839_=<_x_<_0.9140	107	21.40%	
	Skewness	0.1735	0.9140_=<_x_<_1.5441	59	11.80%	
	Kurtosis	0.3045	1.5441_=<_x_<_2.1741	23	4.60%	
			2.1741_=<_x_<_2.8042	6	1.20%	
			x>=_2.8042	3	0.60%	
	Statistics		**Values**	**Count**	**Percent**	**Histogram**
LOGNORMAL	Average	1.7253	x_<_3.1520	443	88.60%	
	Median	1.0175	3.1520_=<_x_<_6.2472	37	7.40%	
	Std dev. [Coef of variation]	2.5887 [1.5005]	6.2472_=<_x_<_9.3423	14	2.80%	
	MAD [MAD/STDDEV]	1.3500 [0.5215]	9.3423_=<_x_<_12.4374	1	0.20%	
	Min * Max [Full range]	0.06 * 31.01 [30.95]	12.4374_=<_x_<_15.5326	2	0.40%	
	1st * 3rd quartile [Range]	0.53 * 1.97 [1.45]	15.5326_=<_x_<_18.6277	0	0.00%	
	Skewness	6.0455	18.6277_=<_x_<_21.7228	1	0.20%	
	Kurtosis	51.8795	21.7228_=<_x_<_24.8179	1	0.20%	
			24.8179_=<_x_<_27.9131	0	0.00%	
			x>=_27.9131	1	0.20%	

The histograms show than UNIFORM and NORMAL variables seems symmetric, the descriptive statistics confirm this result: mean and median are very close for the 2 first attributes and the SKEWNESS is roughly equal to zero. At the opposite LOGNORMAL has a "skewed-left" distribution, and the SKEWNESS seems "significantly" upper than zero.

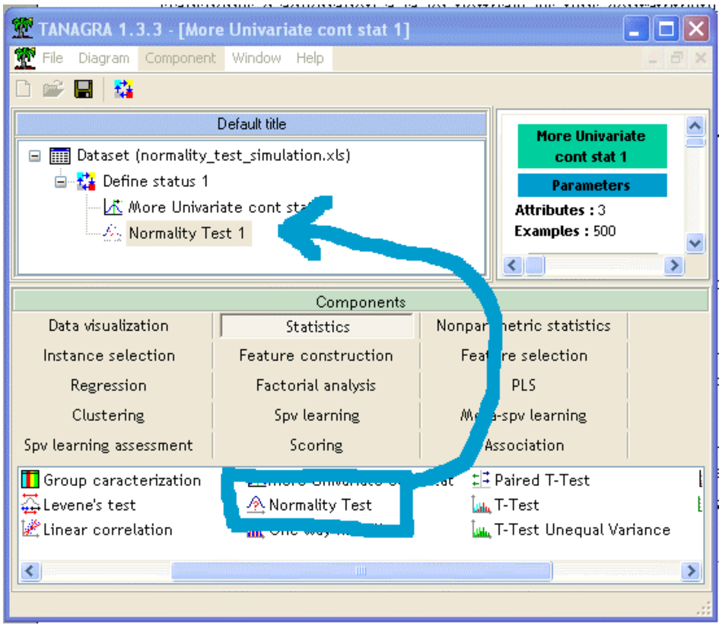
UNIFORM has a flattened shape; the KURTOSIS seems significantly lower than zero, in this case, it appears also doubtful that this variable comes from a normal distribution.

There remains the NORMAL attribute. Compatibility with the normal distribution does not seem eccentric. Some indications consolidate this idea (Skewness and Kurtosis), we note also that the relationship between the mean absolute deviation (MAD) and the standard deviation is near to 4/5, which is a characteristic of normal distribution.

Nevertheless, it is necessary to confirm these impressions with more rigorous statistical tests.

Goodness-of-fit tests for the hypothesis of normality

We add the new component NORMALITY TEST in the diagram.



Four tests are computed: SHAPIRO-WILK that can be used only if the sample size is lower than 5000 examples; LILLIEFORS which is a modification of the KOLMOGOROV-SMIRNOV test; ANDERSON-DARLING which is also a modification of KOLMOGOROV-SMIRNOV test; D'AGOISTINO test which is based on the SKEWNESS and the KURTOSIS measured on the dataset.

The references about these methods are available on the website of TANAGRA (http://chirouble.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html -- Section « Releases »).

Normality Test 1

Parameters

Attributes : 3
Examples : 500

Results

Attribute	Mu ; Sigma	Shapiro-Wilk (p-value)	Lilliefors D = max[D-,D+] (p-value)	Anderson-Darling (p-value)	d'Agostino (p-value)
UNIFORM	0.5048 ; 0.2838	0.954528 *(0.0000)*	0.0740 = max [0.0650,0.0740] *(p < 0.01)*	6.084513 *(p < 0.01)*	0.1781 ^2 + 3.7850 ^2 = 14.3577 *(0.0008)*
NORMAL	0.0301 ; 0.9786	0.994937 *(0.1003)*	0.0304 = max [0.0207,0.0304] *(p >= 0.20)*	0.498039 *(p >= 0.10)*	1.5903 ^2 + 1.3453 ^2 = 4.3389 *(0.1142)*
LOGNORMAL	1.7253 ; 2.5887	0.494280 *(0.0000)*	0.2596 = max [0.2596,0.2162] *(p < 0.01)*	62.993379 *(p < 0.01)*	20.6785 ^2 + 13.8530 ^2 = 619.5065 *(0.0000)*

For a significance level of 5%, we see that the hypothesis of normality cannot be rejected for the NORMAL attribute, contrary the other attributes.

When the results are statistically significant, the cells are colored in red.