

Objectif

Le test d'adéquation sert à vérifier que la distribution empirique des observations est compatible avec une distribution théorique spécifique. Dans ce didacticiel, nous montrons le mode d'utilisation du composant NORMALITY TEST, il intègre plusieurs tests visant à vérifier l'adéquation des données avec une la distribution normale (loi de Laplace-Gauss).

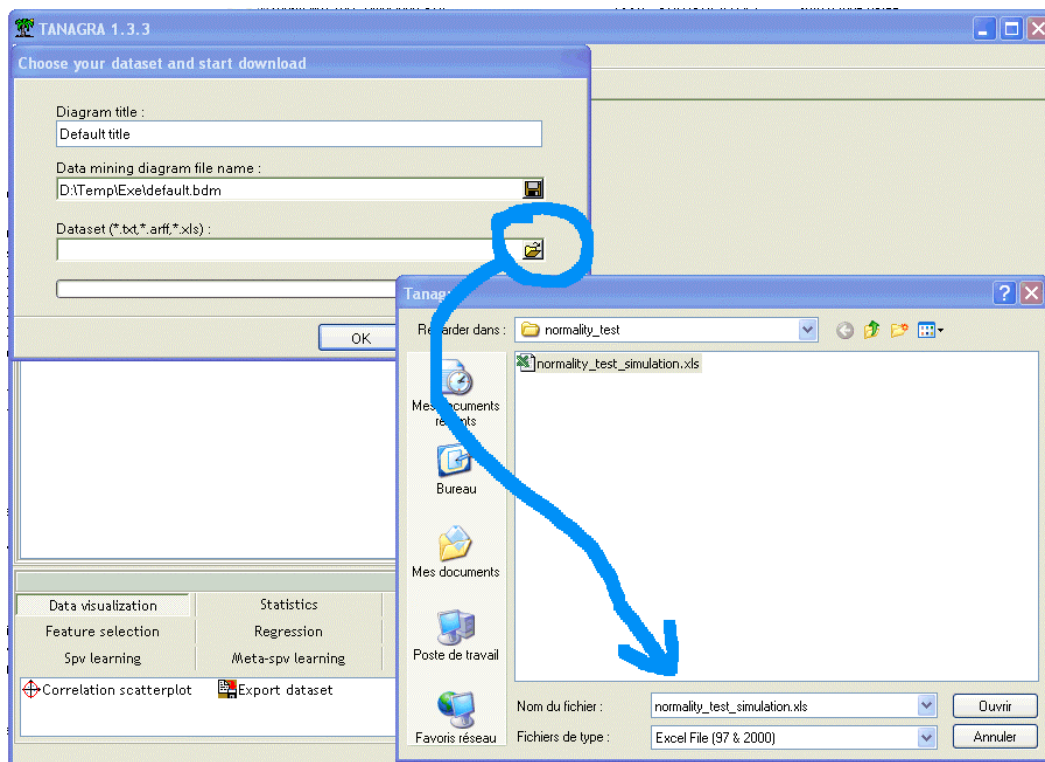
Fichier

Le fichier NORMALITY_TEST_SIMULATION.XLS contient 500 observations : les données ont été générées à l'aide du générateur de nombre aléatoire du tableur EXCEL, les 3 variables générées correspondent aux distributions uniformes, normales et log-normales. A priori, les tests que nous présentons ci-dessous doivent mettre en exergue des résultats cohérents avec le mode de génération de ces données.

Test d'adéquation à la loi normale

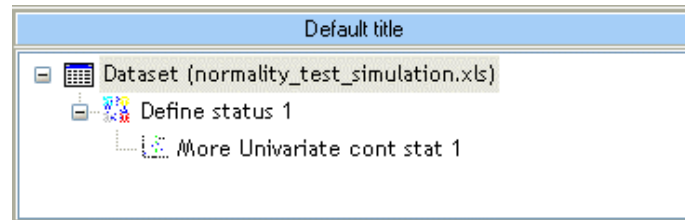
Charger le fichier de données

La première étape consiste à importer les données dans TANAGRA. Nous activons le menu FILE/NEW pour créer un nouveau diagramme.



Statistiques descriptives

Les statistiques descriptives permettent de se donner une idée de la forme des distributions. Pour ce faire, nous ajoutons dans le diagramme un composant DEFINE STATUS, nous plaçons les trois variables en INPUT ; puis, nous insérons le composant MORE UNIVARIATE CONT STAT à partir de la palette STATISTICS.



Les résultats appellent quelques commentaires.

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%		
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%		
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%		

Les histogrammes donnent une idée assez précise des distributions : UNIFORM et NORMAL sont symétriques, LOGNORMAL est très dissymétrique. Les indicateurs numériques confirment l'impression visuelle : pour les deux premières variables, la moyenne est assez proche de la médiane ; l'écart est nettement plus conséquent la dernière variable. Autre indicateur intéressant, la valeur du coefficient d'asymétrie (Skewness) vient étayer ces indications, il est proche de zéro pour les deux premières variables. A partir de ces premières impressions, nous pouvons d'ores et déjà écarter l'hypothèse de normalité pour la variable LOGNORMAL.

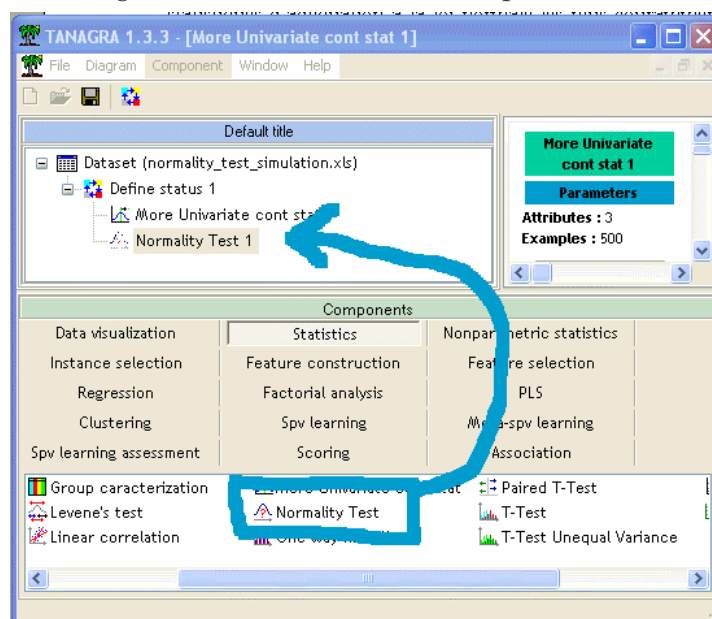
Nous constatons que UNIFORM présente un histogramme plat, à la différence de NORMAL. Conséquence, la valeur du coefficient d'aplatissement s'écarte assez fortement de zéro : dans ce cas, il paraît également douteux que cette variable provienne d'une loi de distribution normale.

Il nous reste donc la variable NORMAL. La compatibilité avec la loi normale ne semble pas farfelue, petite indication qui nous conforte dans cette idée, nous constatons que le rapport entre l'écart absolu moyen (MAD) et l'écart type (STDDEV) est proche de $4/5$, ce qui est caractéristique de la loi normale (« La Statistique », André Vessereau, Collection QSJ, 1996 -- Page 39).

Bien que très instructive, cette première évaluation reste purement empirique. Pour la confirmer ou l'infirmer, nous utilisons dans la section suivante les tests statistiques d'adéquation à la loi normale les plus couramment cités et présents dans les logiciels.

Tester l'adéquation à une distribution normale

Nous complétons notre diagramme avec le nouveau composant NORMALITY TEST.



Quatre tests sont automatiquement calculés : le test de SHAPIRO WILK – utilisable uniquement pour un effectif inférieur à 5000 observations ; le test de LILLIEFORS qui s'appuie sur la statistique de KOLMOGOROV et SMIRNOV mais tabule différemment les valeurs critiques sachant que nous testons l'adéquation à la loi normale et que les paramètres de la distribution sont estimés à partir de l'échantillon, ce test ne fournit pas directement la p-value, elle positionne la statistique par rapport aux seuils critiques calculés sur les niveaux de significations couramment utilisés dans la pratique (10%, 5%, etc.) ; le test d'ANDERSON et DARLING, qui est dérivé du test de KOLMOGOROV, est assez proche dans son esprit du test précédent, il tient mieux compte des queues de distribution ; enfin, le test de D'AGOSTINO calcule la statistique de test à partir des indicateurs d'asymétrie et d'aplatissement, elle suit une loi du CHI-2 à deux degrés de liberté.

Les principales références utilisées, sur les méthodes et leur programmation, sont disponibles sur notre site (<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)

Par rapport à un niveau de signification de 5%, les tests conduisant au rejet de l'hypothèse de normalité sont signalés en rouge, les résultats sont sur fond vert dans le cas contraire.

Dans notre exemple, nous constatons (1) que tous les tests sont cohérents, (2) seule la variable NORMAL est compatible avec une distribution théorique normale.

Ce résultat est assez heureux compte tenu du fait que ces variables ont été générées artificiellement en utilisant des générateurs de nombres aléatoires avec les caractéristiques voulues.