

1 Objectif

Description de quelques fonctions du logiciel PSPP, comparaison des résultats avec ceux de Tanagra, R et OpenStat.

Tout le monde l'aura compris, je passe énormément de temps à analyser les logiciels de statistique et de data mining gratuits découverts ici ou là sur le web. Je suis toujours enthousiasmé à l'idée de découvrir les dispositifs imaginés par les uns et les autres pour proposer aux utilisateurs, nous, des solutions de traitement de données. Au fil des années, j'en suis arrivé à la conclusion qu'il n'existe pas de mauvais logiciels. Il y a simplement des outils plus ou moins adaptés à des contextes d'utilisation qu'il nous appartient de cerner, en tenant compte de nos objectifs, des caractéristiques de nos données, de notre mode opératoire, de nos affinités, etc. On ne gagnera jamais le Paris-Dakar avec une Formule Un ; Sébastien Loeb, aussi fort soit-il, ne peut pas gagner un rallye avec une semi-remorque (j'imagine hein, avec lui on ne sait jamais). C'est l'une des raisons pour lesquelles je parle énormément des autres logiciels, autres que ceux que je développe moi-même. Plus nous en verrons, plus nous saurons nous détacher de l'outil pour nous concentrer sur les finalités, les techniques, l'exploitation des résultats. C'est ce qui importe en définitive.

Dans ce tutoriel, nous décrivons le logiciel PSPP. Ses promoteurs la positionnent comme une alternative à SPSS (« *PSPP is a program for statistical analysis of sampled data. It is a free replacement for the proprietary program SPSS, and appears very similar to it with a few exceptions.* »^{1,2}). Plutôt que de procéder à une analyse exhaustive de ses fonctionnalités, ce qui est déjà très bien fait par ailleurs³, avec en particulier le document en français de Julie Séguéla (« [Introduction au logiciel PSPP – Version 0.4.0](#) », 2006 ; 119 pages⁴), nous préférons décrire quelques procédures statistiques en mettant en miroir les résultats fournis par **Tanagra, R 2.13.2** et **OpenStat (build 24/02/2012)**. C'est une manière de les valider mutuellement. Plus que les plantages, les erreurs de calculs sont la hantise des informaticiens. Obtenir des résultats identiques pour les mêmes traitements avec plusieurs logiciels n'est pas un gage d'exactitude. En revanche, en cas de disparités, il y a clairement un problème. L'affaire devient diablement compliquée lorsque ces disparités ne surviennent que dans des situations que l'on a du mal à identifier.

2 Données

Nous utilisons une variante du fichier « Automobile Dataset » du serveur UCI⁵. Il répertorie les caractéristiques de 205 véhicules : marque, poids, puissance, consommation, etc.

¹ <http://www.gnu.org/software/pspp/pspp.html>

² Pourquoi pas après tout ? Si on n'est pas ambitieux pour soi-même, qui le sera à notre place ? Le logiciel R a du commencer tout petit un jour. En voyant ce qu'il est devenu aujourd'hui, on ne peut que s'en réjouir.

³ <http://www.gnu.org/software/pspp/documentation.html>

⁴ Ce document est d'autant plus intéressant qu'il décrit de manière détaillée les procédures de manipulation de données (ajout - suppression de variables, transformations, sélection des observations) avec le langage de commande de PSPP. Ces tâches, très répétitives et fastidieuses, sont primordiales dans l'utilisation quotidienne du logiciel. Pouvoir les programmer dans un fichier script est un atout essentiel.

⁵ <http://archive.ics.uci.edu/ml/datasets/Automobile>

Selon les techniques que nous présenterons, nous utiliserons telle ou telle variable de la base. Qu'importe l'analyse des résultats dans ce tutoriel. L'important pour nous est de décrire la mise en œuvre des différentes méthodes statistiques sous PSPP pour que tout un chacun puisse reproduire la démarche sur son propre fichier.

3 Le logiciel PSPP

3.1 Charger et installer le logiciel

Le logiciel PSPP est accessible sur son site web (<http://www.gnu.org/software/pspp/>).

The screenshot shows the GNU PSPP website. The browser's address bar contains the URL <http://www.gnu.org/software/pspp/>. The page has a blue header with navigation links: 'Get PSPP', 'FAQ', 'Documentation', 'Contribute', and 'Quick Tour'. The main content area includes a PSPP logo, a description of the software as a free replacement for SPSS, and a list of features such as supporting over 1 billion cases and variables, and being cross-platform. A 'USEFUL LINKS' sidebar is also visible.

Nous avons récupéré la version **0.7.8**, datée du **11 novembre 2011**. PSPP a besoin de l'environnement MINGW pour fonctionner⁶. Son installation est transparente pour nous. Nous n'avons pas à configurer manuellement des bibliothèques additionnelles. C'est appréciable. Les utilisateurs sont souvent rebutés par la nécessité d'effectuer une série de manipulations systèmes avant de pouvoir lancer un logiciel.

Le processus d'installation sous Windows n'amène pas de commentaires particuliers. Le logiciel est accessible via le menu DEMARRER de Windows. Il est également possible d'intégrer des raccourcis sur le bureau ou dans la barre de lancement rapide de Windows.

⁶ <http://www.mingw.org/>

3.2 Fonctionnement en ligne de commande

PSPP peut travailler en ligne de commande. Nous décrivons les instructions dans un fichier script à l'aide d'un éditeur de texte, puis nous le transmettons à l'exécutable. Les résultats peuvent être affichées dans la console ou collectées dans un fichier texte.

Le langage est compatible avec celui de SPSS. Sa syntaxe est exhaustivement décrite sur le site web de l'éditeur (<http://www.gnu.org/software/pspp/manual/pspp.html#Language>). En apprenant à programmer avec PSPP, nous saurons le faire avec SPSS. Voilà un autre atout très intéressant.

A titre d'exemple, nous avons souhaité comparer la puissance (horsepower) des véhicules selon le type de carburant utilisé (fuel_type). Nous avons rédigé les instructions ci-dessous à l'aide d'un éditeur de texte. Puis nous l'avons sauvegardé dans le fichier « test.syn ».

```
GET FILE="D:\dataset\pspp\autos.sav".

T-TEST /VARIABLES= horsepower
      /GROUPS=fuel_type("gas","diesel") /MISSING=ANALYSIS
      /CRITERIA=CIN(0.95) .
```

Les résultats sont affichés dans la console MSDOS de Windows lorsque nous le transmettons à l'exécutable PSPP.EXE.

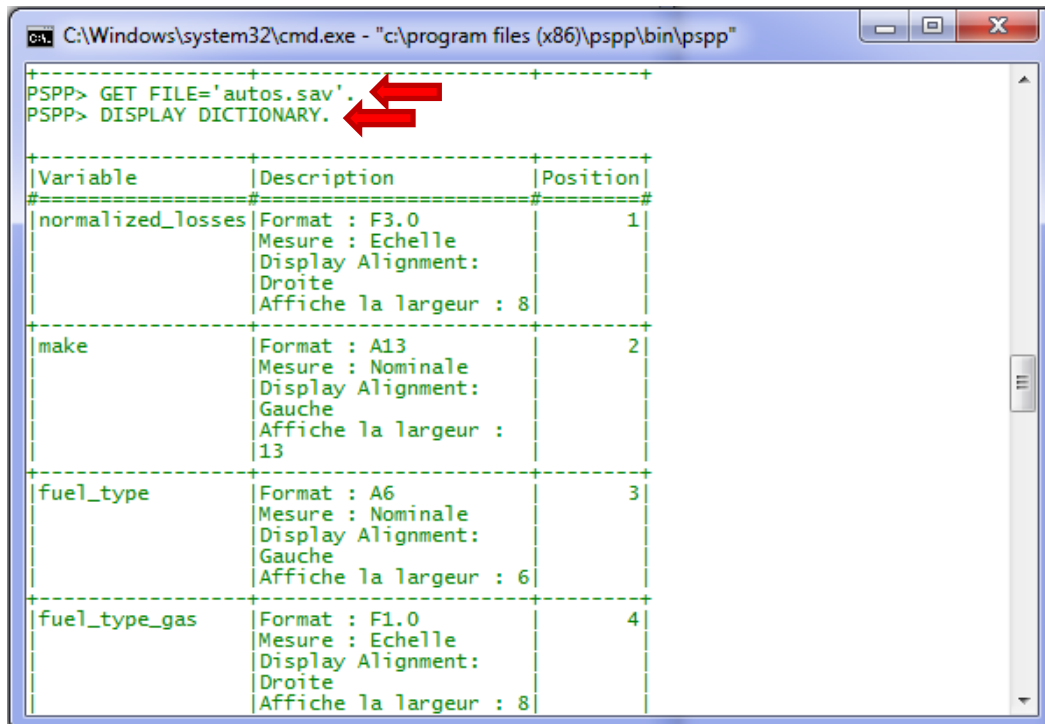
```
C:\Windows\system32\cmd.exe
D:\DataMining\Databases_for_mining\logiciels_dataset\pspp>c:\program files (x86)\pspp\bin\pspp test.syn
Group Statistics
-----#-----#-----#-----#-----#
#           fuel_type| N |Moyenne|Std. Deviation|S.E. Mean#
#-----#-----#-----#-----#
#horsepower gas      |185| 106.39|    40.18|     2.95#
#                                     #
#           diesel   | 20|  84.45|    25.96|     5.80#
#-----#-----#-----#-----#

Independent Samples Test
-----#-----#-----#-----#
#           # Levene's #           t-test for Equality of Means
#           # Test for #
#           # Equality #
#           # of #
#           # Variances#
#           #-----#-----#-----#-----#
#           # F      Sig.  t      df      Sig.  Mean  Std. Error
#           #           #           #           #           (2-  Difference|Difference
#           #           #           #           #           tailed) #
#-----#-----#-----#-----#
#horsepowerEqual # 1.92  .17| 2.39|203.00| .02|  21.94|  6.51
#   variances#
#   assumed #
#   Equal #
#   variances#
#   not #
#   assumed #
#-----#-----#-----#-----#
```

PSPP compare tout d'abord les variances conditionnelles à l'aide du test de Levene. Puis il effectue la comparaison des moyennes avec et sans l'hypothèse d'homoscédasticité.

3.3 Fonctionnement en mode terminal

PSPP peut aussi fonctionner en mode interactif. Après avoir lancé PSPP.EXE, un terminal de commandes s'affiche. Nous pouvons saisir les instructions et visualiser directement les résultats. Dans la copie d'écran ci-dessous, après avoir chargé le fichier « autos.sav », nous faisons afficher le dictionnaire des données.



```

C:\Windows\system32\cmd.exe - "c:\program files (x86)\pspp\bin\pspp"
PSPP> GET FILE='autos.sav'.
PSPP> DISPLAY DICTIONARY.
+-----+-----+-----+
|Variable|Description|Position|
+-----+-----+-----+
|normalized_losses|Format : F3.0|1|
| |Mesure : Echelle| |
| |Display Alignment:| |
| |Droite| |
| |Affiche la largeur : 8| |
+-----+-----+-----+
|make|Format : A13|2|
| |Mesure : Nominale| |
| |Display Alignment:| |
| |Gauche| |
| |Affiche la largeur :| |
| |13| |
+-----+-----+-----+
|fuel_type|Format : A6|3|
| |Mesure : Nominale| |
| |Display Alignment:| |
| |Gauche| |
| |Affiche la largeur : 6| |
+-----+-----+-----+
|fuel_type_gas|Format : F1.0|4|
| |Mesure : Echelle| |
| |Display Alignment:| |
| |Droite| |
| |Affiche la largeur : 8| |
+-----+-----+-----+

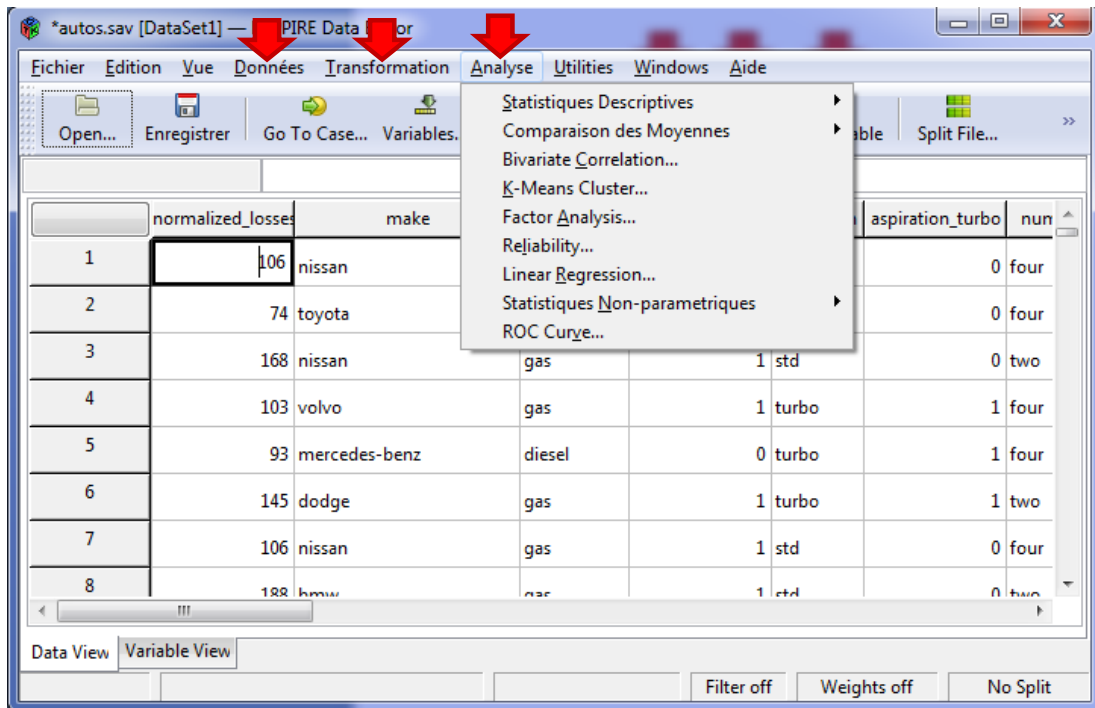
```

3.4 Fonctionnement en mode graphique

Nous devons faire l'apprentissage du langage PSPP pour exploiter les deux modes opératoires ci-dessus. Nous n'en avons pas toujours le temps ni l'envie. Pour les personnes peu enclines à programmer, PSPP propose le pilotage interactif par menu. Sans nul doute que la grande majorité des utilisateurs travaillerons sous ce mode.

C'est l'approche que nous privilégierons dans ce tutoriel. Au démarrage du logiciel via le raccourci Windows, nous obtenons la fenêtre suivante. Elle n'est pas sans rappeler celle de SPSS. Les commandes sont accessibles dans des menus regroupés par thèmes⁷ dont les principaux sont : « **données** » pour la manipulation des observations ; « **transformation** » pour la manipulation des variables ; « **analyse** » pour les traitements statistiques.

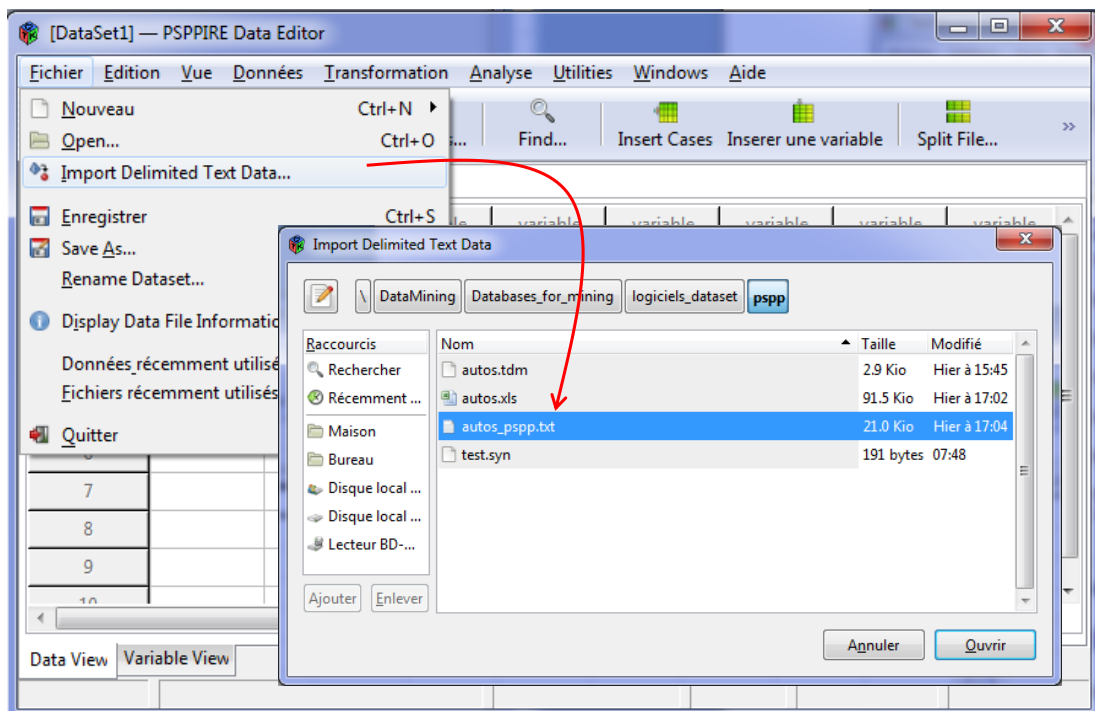
⁷ Curieusement, certains menus sont en français, d'autres en anglais. Je ne sais pas si c'est dû à ma configuration machine (Windows en français), alors que je n'ai demandé à aucun moment de disposer de la version française durant l'installation, ou si nous avons les mêmes menus sur les autres systèmes.



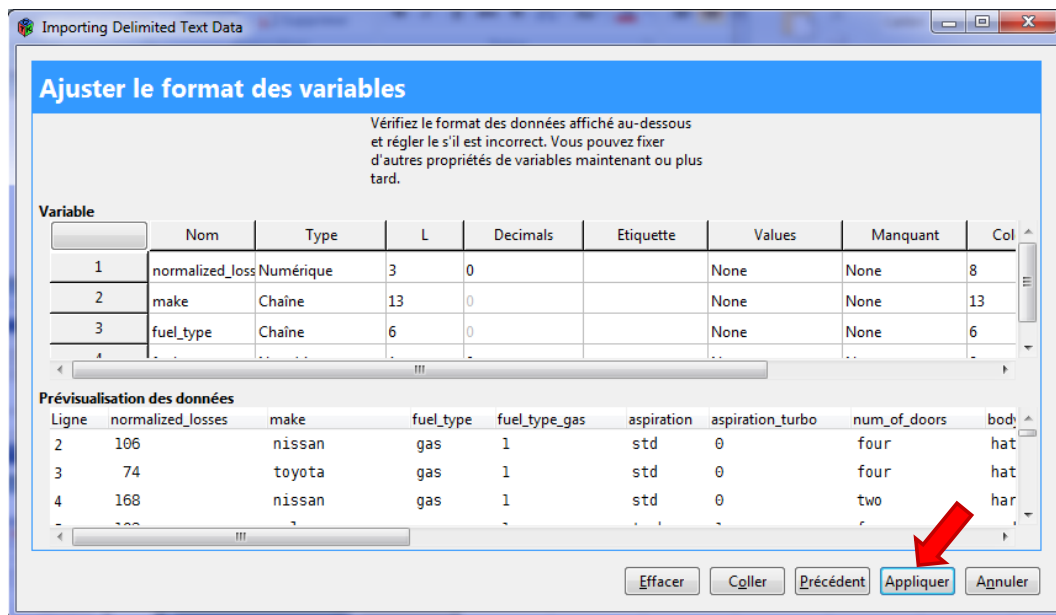
4 Importation des données

Première étape, nous devons importer les données contenues dans « [autos_psppt.txt](#) ». Il s'agit d'un fichier texte avec le caractère tabulation comme séparateur de colonnes, format reconnu par la quasi-totalité des logiciels de statistique et de data mining.

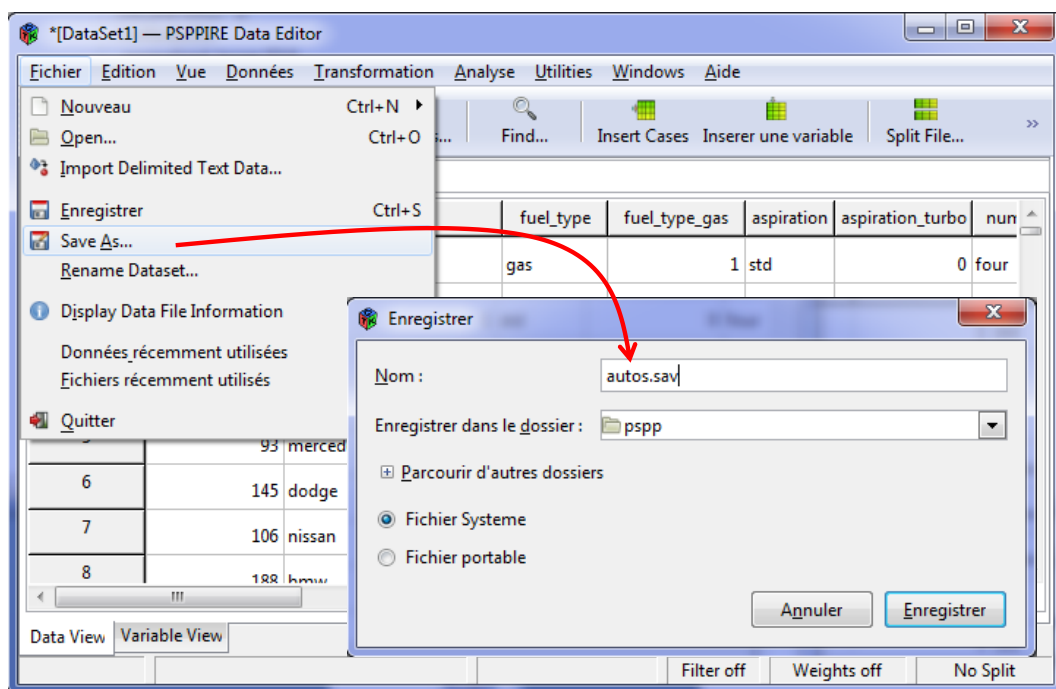
Nous actionnons le menu FICHIER / IMPORT DELIMITED TEXT DATA dans PSPP. Nous sélectionnons notre fichier dans la boîte de dialogue.



En cliquant sur le bouton OUVRIER, un guide (wizard) apparaît. Il nous aide à paramétrer l'importation. En résumé, il nous faut : (1) importer toutes les observations ; (2) la première ligne correspond au nom des variables ; (3) le caractère TAB est le séparateur de colonnes ; (4) pour notre fichier, les variables sont soit « numériques » (variables quantitatives) soit chaînes (variables qualitatives). Après cette dernière étape, il ne nous reste plus qu'à cliquer sur le bouton APPLIQUER.



Les données sont chargées. Nous nous empressons de le sauvegarder dans le format natif de SPSS.



Nous nommons le fichier « **autos.sav** ». Ce format est reconnu par le logiciel SPSS.

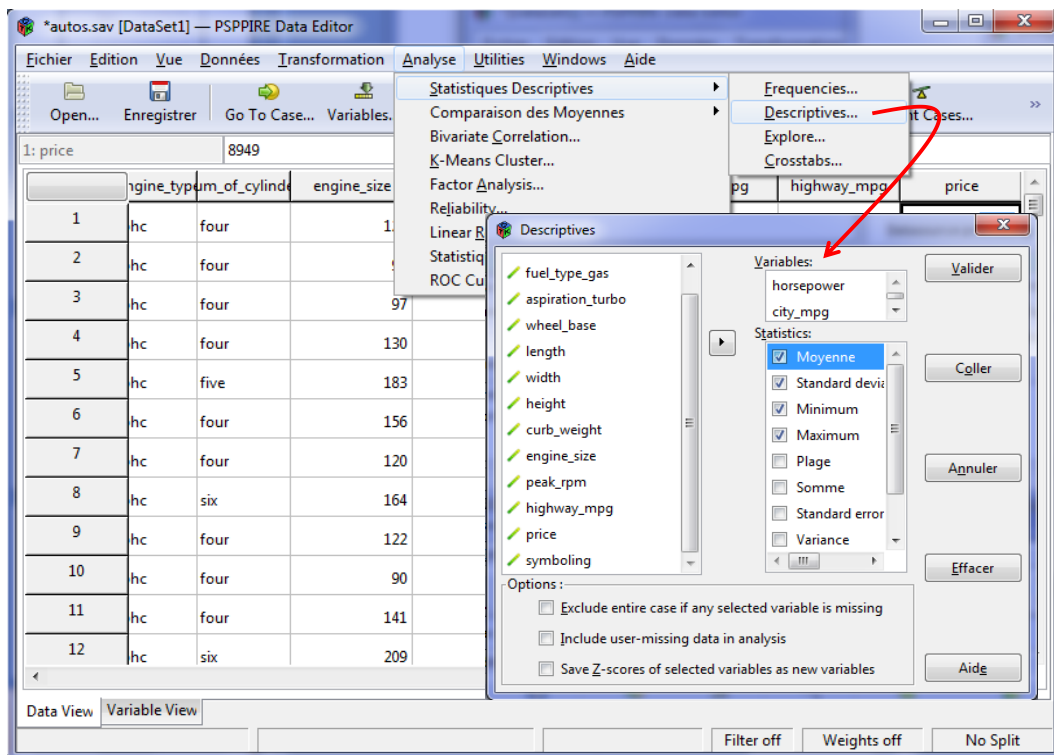
5 Quelques techniques statistiques avec SPSS

Dans cette section, nous décrivons le paramétrage et la lecture des résultats fournis par quelques techniques disponibles dans SPSS. Lorsque la comparaison est possible, soit parce qu'elles sont

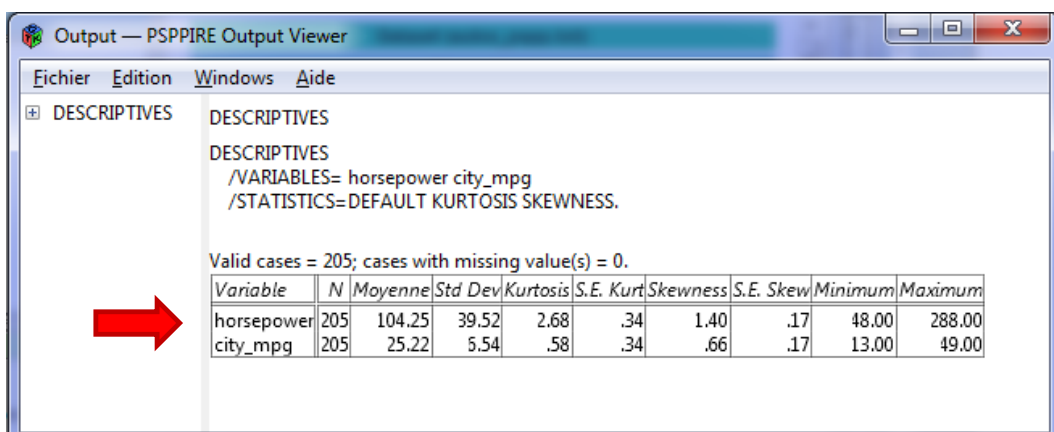
présentes dans les deux logiciels, soit parce que notre fichier de données se prête à l'analyse, nous mettons en miroir les sorties de Tanagra. Les méthodes étant basées sur des algorithmes exacts, les valeurs obtenues sont identiques. Seule la présentation peut différer parfois.

5.1 Statistiques descriptives – Variables quantitatives

Pour calculer les statistiques descriptives des variables « horsepower » (puissance) et « city_mpg » (consommation en ville⁸). Nous actionnons le menu ANALYSE / STATISTIQUES DESCRIPTIVES / DESCRIPTIVES. Dans la boîte de paramétrage, nous sélectionnons les deux variables et nous spécifions les indicateurs à calculer.



Les résultats sont affichés dans une fenêtre de sortie distincte.



Le composant MORE UNIVARIATE CONST STAT de **Tanagra** fournit les valeurs suivantes. Les valeurs, pour celles qui sont présentées en tous les cas, concordent en tous points.

⁸ L'unité est MPG, miles parcourus avec un gallon de carburant. Une valeur élevée indique un véhicule sobre.

TARGET : ()
INPUT : (horsepower, city_mpg)

Dataset (autos_psp.ppt.txt)
 Define status 1
 More Univariate cont stat 1

Attribute	Stats	
horsepower	Statistics	
	Average	104.2537
	Median	95
	Std dev. [Coef of variation]	39.5192 [0.3791]
	MAD [MAD/STDDEV]	30.2093 [0.7644]
	Min * Max [Full range]	48.00 * 288.00 [240.00]
	1st * 3rd quartile [Range]	70.00 * 116.00 [46.00]
	Skewness (std-dev)	1.3980 (0.1698)
	Kurtosis (std-dev)	2.6785 (0.3381)
	city_mpg	Statistics
Average		25.2195
Median		24
Std dev. [Coef of variation]		6.5421 [0.2594]
MAD [MAD/STDDEV]		5.2155 [0.7972]
Min * Max [Full range]		13.00 * 49.00 [36.00]
1st * 3rd quartile [Range]		19.00 * 30.00 [11.00]
Skewness (std-dev)		0.6637 (0.1698)
Kurtosis (std-dev)		0.5786 (0.3381)

5.2 Statistiques descriptives conditionnelles

Nous souhaitons maintenant calculer les statistiques descriptives de « horsepower » en fonction des valeurs prises par « fuel_type » (type de carburant utilisé, « gas » [essence] ou « diesel » [gazole]). Nous actionnons le menu ANALYSE / STATISTIQUES DESCRIPTIVES / EXPLORE. Nous plaçons « horsepower » en DEPENDENT LIST, « fuel_type » en FACTOR LIST. Via le bouton STATISTIQUES, nous demandons les statistiques descriptives. Il ne reste plus qu'à valider.

	Cas					
	Valid		Missing		Total	
fuel_type	N	Pourcentage	N	Pourcentage	N	Pourcentage
horsepower diesel	20	100%	0%	0%	20	100%
gas	185	100%	0%	0%	185	100%

fuel_type	Statistique	Std. Error
horsepower diesel	Moyenne	34.45
	95% Confidence Interval for Mean	72.30
	Limite inférieure	36.60
	Limite supérieure	34.11
	5% Trimmed Mean	34.00
	Mediane	573.84
	Variance	25.96
	Std. Deviation	52.00
	Minimum	123.00
	Maximum	71.00
gas	Plage	45.25
	Interquartile Range	-.28
	Skewness	.51
	Kurtosis	-.99
	Moyenne	106.39
	95% Confidence Interval for Mean	100.57
	Limite inférieure	112.22
	Limite supérieure	103.07
	5% Trimmed Mean	97.00
	Mediane	1614.71
Variance	40.18	
Std. Deviation	48.00	
Minimum	288.00	
Maximum	240.00	
Plage	48.00	
Interquartile Range	1.38	
Skewness	.18	
Kurtosis	.36	

Explorer
X

- normalized_losses
- make
- fuel_type_gas
- aspiration
- aspiration_turbo
- num_of_doors
- body_style

Dependent List: horsepower

Factor List: fuel_type

Label Cases by:

Statistiques...
Options...
Aider

Nous nous intéressons à quelques valeurs du tableau de comparaison : la moyenne de horsepower est de 84.45 (resp. 106.39) chez les « gas » (resp. chez les « diesel ») ; les écarts-type sont respectivement de 25.96 et 40.18 ; etc. Notons que les sorties sont très complètes.

Le composant GROUP CHARACTERIZATION de **Tanagra** effectue des calculs similaires, sans toutefois prétendre à une telle exhaustivité. Il affiche essentiellement les moyennes et les écarts-type.

TARGET : (fuel_type)
 INPUT : (horsepower)

Dataset (autos_psppt.txt)

- Define status 1
- More Univariate cont stat 1
- Define status 2
- Group characterization 1

Description of "fuel_type"

fuel_type=gas				fuel_type=diesel			
Examples [90.2 %] 185				Examples [9.8 %] 20			
Att - Desc	Test value	Group	Overall	Att - Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)			
horsepower	2.35	106.39 (40.18)	104.25 (39.52)	horsepower	-2.35	84.45 (25.96)	104.25 (39.52)
Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy			

5.3 Tableaux de contingence

Nous croisons les variables « fuel-type » (type de carburant) et « aspiration » (turbo ou standard) dans un tableau de contingence. Nous actionnons le menu ANALYSE / STATISTIQUES DESCRIPTIVES / CROSSTABS. Nous plaçons la première variable en ligne, la seconde en colonne. L'outil est très riche, nous pouvons sélectionner un grand nombre d'indicateurs avec l'option STATISTICS (D de Sommers, Coefficient d'incertitude de Theil, Lambda et Tau de Goodman-Kruskal, Kappa de Cohen, etc.⁹).

Tableaux Croisés.

Rows: fuel_type

Columns: aspiration

Valider

Coller

Annuler

Effacer

Format...

Statistics...

Cells...

Aide

fuel_type * aspiration [Compter, ligne %, colonne %, total %].

fuel_type	aspiration		Total
	std	turbo	
diesel	7.0	13.0	20.0
	35.0%	55.0%	100.0%
	4.2%	35.1%	9.8%
	3.4%	5.3%	9.8%
gas	161.0	24.0	185.0
	37.0%	13.0%	100.0%
	95.8%	54.9%	90.2%
	78.5%	11.7%	90.2%
Total	168.0	37.0	205.0
	82.0%	18.0%	100.0%
	100.0%	100.0%	100.0%
	82.0%	18.0%	100.0%

Tests du Chi-Deux

Statistique	Valeur	df	Asymp. Sig. (2-tailed)
Pearson Chi-Square	33.03	1	.00
Likelihood Ratio	24.90	1	.00
Continuity Correction	29.61	1	.00
N observations valides	205		

⁹ Voir R. Rakotomalala, « [Etude des dépendances – Variables qualitatives](#). Tableau de contingence et mesures d'association », Version 2.0, Mars 2011.

Nous disposons des différents profils (pourcentages). PSPP propose, entre autres, le KHI-2 de Pearson, le test du rapport de vraisemblance, le KHI-2 avec la correction de continuité.

Les mêmes outils existent dans **Tanagra**. Mais l'organisation est différente. Tanagra fournit dans la foulée le Lambda et le Tau de Goodman et Kruskal, ainsi que le coefficient d'incertitude de Theil.

Row (Y)	Column (X)	Statistical indicator		Cross-tab			
		Stat	Value	std	turbo	Sum	
fuel_type	aspiration	d.f.	1	gas	161 87.03%	24 12.97%	185 100%
		Tschuprow's t	0.401397	diesel	7 35.00%	13 65.00%	20 100%
		Cramer's v	0.401397	Sum	168 82%	37 18%	205 100%
		Phi ²	0.16112				
		Chi ² (p-value)	33.03 (0.0000)				
		Lambda	0				
		Tau (p-value)	0.1611 (0.0000)				
		U(R/C) (p-value)	0.1900 (0.0000)				

5.4 Comparaison de 2 moyennes – Echantillons indépendants

Au-delà des statistiques descriptives conditionnelles, nous souhaitons comparer les moyennes conditionnelles de « horsepower » selon « fuel-type ». Dans le menu ANALYSE / COMPARAISON DES MOYENNES / INDEPENDENT SAMPLES T TEST, nous plaçons « horsepower » en TEST VARIABLE et « fuel-type » en DEFINE GROUPS (gas vs. diesel).

Group Statistics

fuel_type	N	Moyenne	Std. Deviation	S.E. Mean
horsepower gas	185	106.39	40.18	2.95
horsepower diesel	20	84.45	25.96	5.80

Independent Samples Test

	Levene's Test for Equality of Variances		t-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Le plus bas	Le plus haut
horsepower	1.92	.17	2.39	203.00	.02	21.94	5.51	9.10	34.79
Equal variances assumed			3.37	29.91	.00	21.94	5.51	8.64	35.25
Equal variances not assumed									

PSPP effectue un test de comparaison de variances au préalable (test de Levene). Nous constatons qu'elles ne sont pas significativement différentes à 5 % (p -value = 0.17). Néanmoins, il propose le résultat du test de comparaison de moyennes avec et sans l'hypothèse d'homoscédasticité. Dans les deux cas, nous constatons que la puissance diffère selon le type de carburant utilisé.

Trois composants sont nécessaires dans **Tanagra** pour ces mêmes résultats. En revanche, ils partagent la même définition du rôle des variables c.-à-d. ils sont situés en aval du même DEFINE STATUS. Il n'y a pas de manipulations répétitives.

TARGET : (horsepower)
INPUT : (fuel-type)

Define status 4
Levene's test 1
T-Test 1
T-Test Unequal Variance 1

Attribute_Y	Attribute_X	Description				Statistical test			
		Value	Examples	Average	Std-dev	Test			
Levene	horsepower fuel_type	gas	185	106.3946	40.1834	Levene's W	1.924219		
		diesel	20	84.45	25.9584			df	1/203
		All	205	104.2537	39.5192			p-value	0.166913

Attribute_Y	Attribute_X	Description				Statistical test			
		Value	Examples	Average	Std-dev	T			
T-Test	horsepower fuel_type	gas	185	106.3946	40.1834	T	21.9446 / 9.1970 = 2.386065		
		diesel	20	84.45	25.9584			d.f.	203
		All	205	104.2537	39.5192			p-value	0.017949

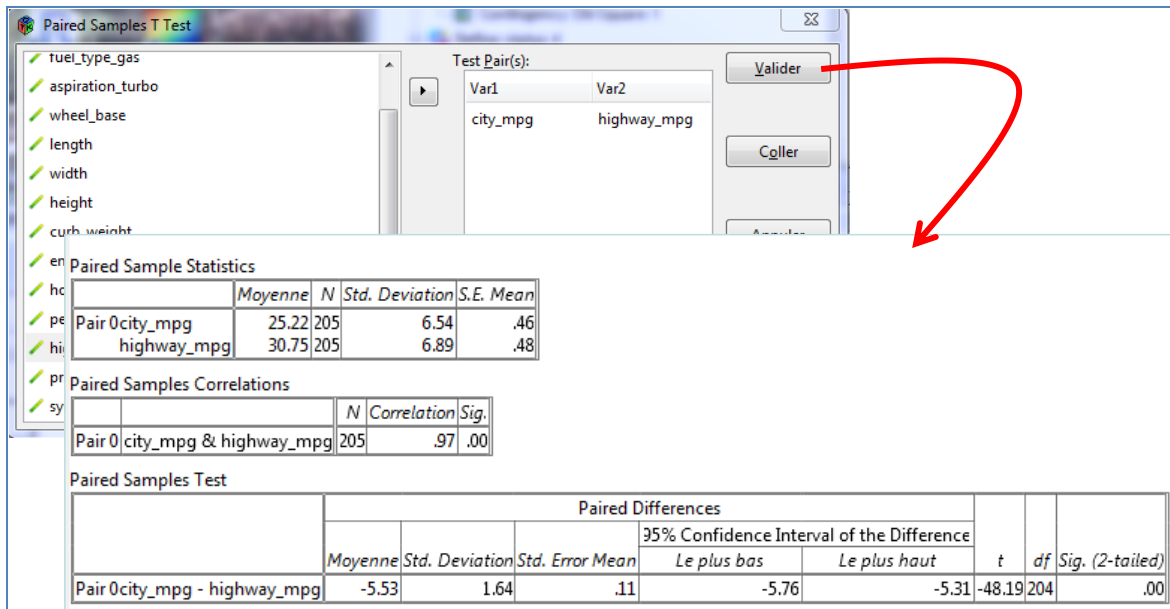
Attribute_Y	Attribute_X	Description				Statistical test			
		Value	Examples	Average	Std-dev	T			
T-Test Uneq. variance	horsepower fuel_type	gas	185	106.3946	40.1834	T	21.9446 / 6.5131 = 3.369315		
		diesel	20	84.45	25.9584			d.f.	29.91
		All	205	104.2537	39.5192			p-value	0.002085

Bien évidemment, les résultats sont strictement identiques à ceux de PSPP.

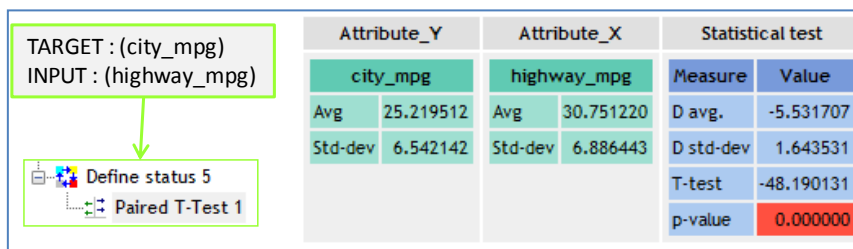
5.5 Comparaison de 2 moyennes – Echantillons appariés

Nous comparons la consommation en ville (*city_mpg*) et sur autoroute (*highway_mpg*). A priori, à voiture égale, on consomme plus en ville (c.-à-d. *la valeur de MPG est plus faible*). Est-ce que nos données confirment cette hypothèse ?

Via le menu ANALYSE / COMPARAISON DES MOYENNES / PAIRED SAMPLES T TEST, nous plaçons les paires de variables « *city_mpg* » et « *highway_mpg* ». PSPP affiche les moyennes (25.2 pour *city_mpg*, 30.75 pour *highway_mpg*) et écarts-type de chaque variable, leur corrélation (0.97, les deux variables sont fortement liées), et le tableau décrivant les résultats du test. Manifestement, on consomme plus en ville que sur autoroute (on parcourt moins de miles avec le même gallon de carburant).

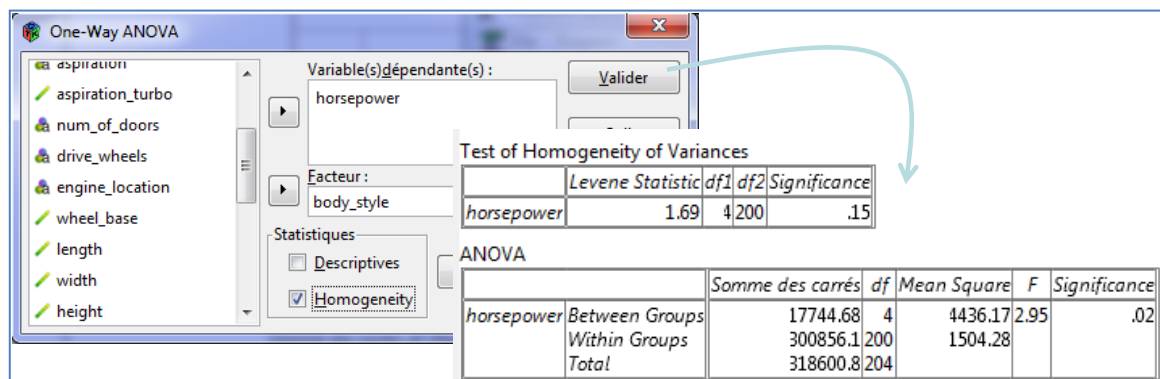


Le composant PAIRED T-TEST de **Tanagra** indique exactement la même chose.



5.6 Comparaison de K moyennes – Analyse de variance (ANOVA)

L'analyse de variance est la généralisation de la comparaison de moyennes à K groupes. Dans notre fichier, nous souhaitons comparer les moyennes de « horsepower » selon le style des véhicules. Avec le menu ANALYSE / COMPARAISON DES MOYENNES / ONE WAY ANOVA, nous plaçons « horsepower » en VARIABLE DEPENDANTE et « body_style » en FACTEUR.



PSPP vérifie l'hypothèse d'homoscédasticité à l'aide du test de Levene. A 5%, on ne peut pas rejeter l'hypothèse nulle ici (p-value = 0.15). Puis il effectue le test de comparaison de moyennes. Le F de Fisher est de 2.95 avec un p-value de 0.02. Au risque 5%, l'hypothèse d'égalité des moyennes dans les groupes est rejetée. Nous pouvons obtenir les moyennes et écarts-type conditionnels si nous le souhaitons.

Tanagra nous annonce la même chose, les tests sont dissociés dans deux composants.

Attribute_Y	Attribute_X	Description				Statistical test	
horsepower	body_style	Value	Examples	Average	Std-dev	Test	
		hatchback	70	101.3714	42.3728	Levene's W	1.690393
		hardtop	8	142.2500	50.6127	df	4/200
		sedan	96	103.1042	37.1641	p-value	0.153627
		wagon	25	98.0000	27.9672		
		convertible	6	131.6667	42.5566		
		All	205	104.2537	39.5192		

Attribute_Y	Attribute_X	Description				Statistical test		
horsepower	body_style	Value	Examples	Average	Std-dev	Variance decomposition		
		hatchback	70	101.3714	42.3728	Source	Sum of square	d.f.
		hardtop	8	142.2500	50.6127	BSS	17744.6752	4
		sedan	96	103.1042	37.1641	WSS	300856.1345	200
		wagon	25	98.0000	27.9672	TSS	318600.8098	204
		convertible	6	131.6667	42.5566	Significance level		
		All	205	104.2537	39.5192	Statistics	Value	Proba
				Fisher's F	2.949030	0.021317		

5.7 Régression linéaire

Nous cherchons à expliquer la consommation en ville (city_mpg) à l'aide du type de carburant utilisé (l'indicatrice de « fuel_type = gas »), du mode d'aspiration (indicatrice de « aspiration = turbo »), du poids du véhicule (curb_weight) et de sa puissance (horsepower). Nous indiquons les paramètres idoines dans la boîte de paramétrage accessible via le menu ANALYSE / LINEAR REGRESSION.

The screenshot shows the 'Regression' dialog box in Tanagra. The dependent variable is 'city_mpg'. Independent variables are 'fuel_type_gas', 'aspiration_turbo', 'curb_weight', and 'horsepower'. A red arrow points from the 'Valider' button to the 'Model Summary' table.

R	R Square	Adjusted R Square	Std. Error of the Estimate
.89	.79	.79	3.02

	Somme des carrés	df	Mean Square	F	Significance
Regression	5906.23	4	1726.56	189.22	.00
Residuel	1824.89	200	9.12		
Total	3731.12	204			

	B	Std. Error	Beta	t	Significance
(Constant)	57.64	1.68	.00	34.28	.00
fuel_type_gas	-8.60	.93	-.39	-9.25	.00
aspiration_turbo	-1.64	.64	-.10	-2.58	.01
curb_weight	-.01	.00	-.63	-10.92	.00
horsepower	-.04	.01	-.24	-4.18	.00

PSPP fournit le coefficient de corrélation multiple $R = 0.89$, le coefficient de détermination $R^2 = 0.79$, le tableau d'analyse de variance, et le tableau des coefficients. Le modèle est globalement significatif ($F = 189.22$, $p\text{-value} = 0.00$). Nous constatons que tous les coefficients sont significatifs au risque 5%. Toutes les variables incluses dans le modèle influent sur la consommation, le poids et le type de carburant ayant l'impact le plus élevé.

Avec le composant MULTIPLE LINEAR REGRESSION, Tanagra fournit exactement les mêmes résultats.

Global results	
Endogenous attribute	city_mpg
Examples	205
R ²	0.790990
Adjusted-R ²	0.786810
Sigma error	3.020670
F-Test (4,200)	189.2233 (0.000000)

Analysis of variance					
Source	xSS	d.f.	xMS	F	p-value
Regression	6906.2327	4	1726.5582	189.2233	0.0000
Residual	1824.8892	200	9.1244		
Total	8731.1220	204			

Coefficients				
Attribute	Coef.	std	t(200)	p-value
Intercept	57.640300	1.681678	34.275470	0.000000
fuel_type_gas	-8.597166	0.929372	-9.250509	0.000000
aspiration_turbo	-1.641124	0.636844	-2.576963	0.010687
curb_weight	-0.007887	0.000722	-10.917917	0.000000
horsepower	-0.040383	0.009654	-4.183196	0.000043

TARGET : (city_mpg)
 INPUT : (fuel_type_gas,
 aspiration_turbo,
 curb_weight, horsepower)

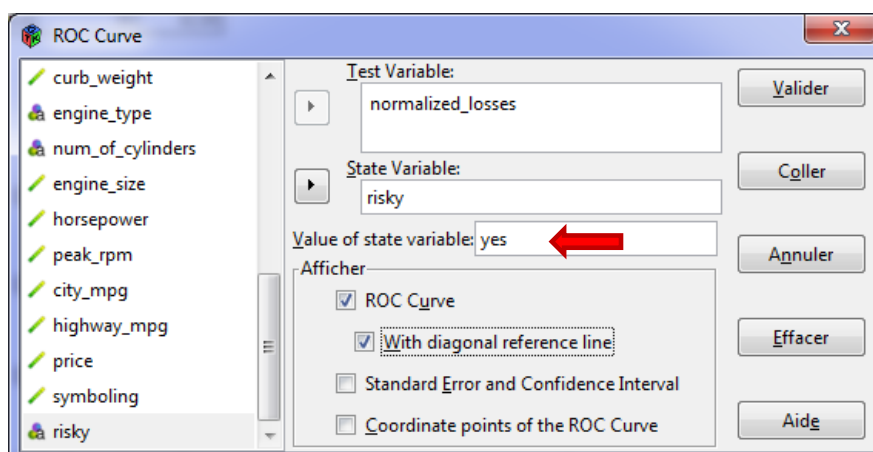
Define status 7
 Multiple linear regression 1

5.8 Courbe ROC

Les compagnies d'assurance indexent les véhicules selon leur niveau de risque, au-delà de leur valeur vénale (\approx le prix). Si la valeur de « symboling » est positive, cela veut dire que la voiture est plus risquée à assurer que ce qu'indique son prix. Nous avons déduit la variable « **risky** » prenant les valeurs « yes » si « symboling > 0 » et « no » si « symboling ≤ 0 ».

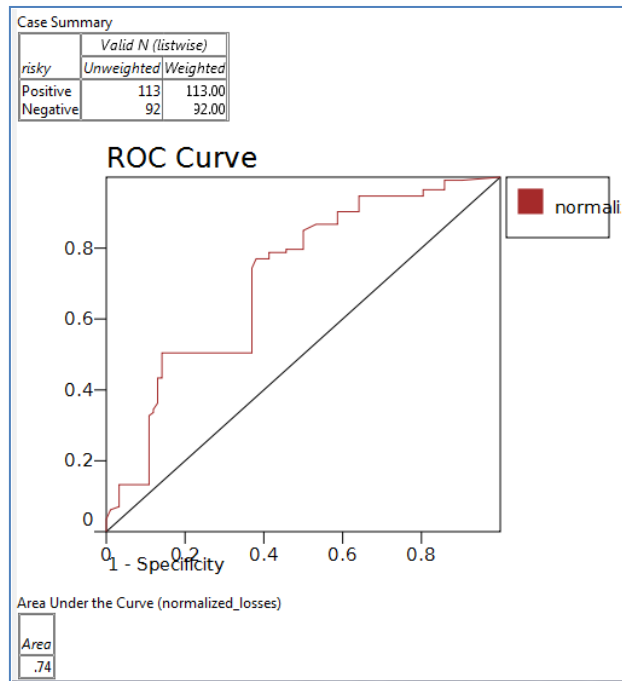
Dans le même temps, elles calculent la perte moyenne pour chaque véhicule (« **normalized_losses** »). La valeur est normalisée selon une classification interne¹⁰.

Nous posons la question suivante : l'indicateur « normalized_losses » permet-il de distinguer les véhicules risqués de ceux qui ne le sont pas ? Nous utilisons la courbe ROC pour y répondre. Via le menu ANALYSE / ROC CURVE, nous paramétrons le logiciel de la manière suivante.



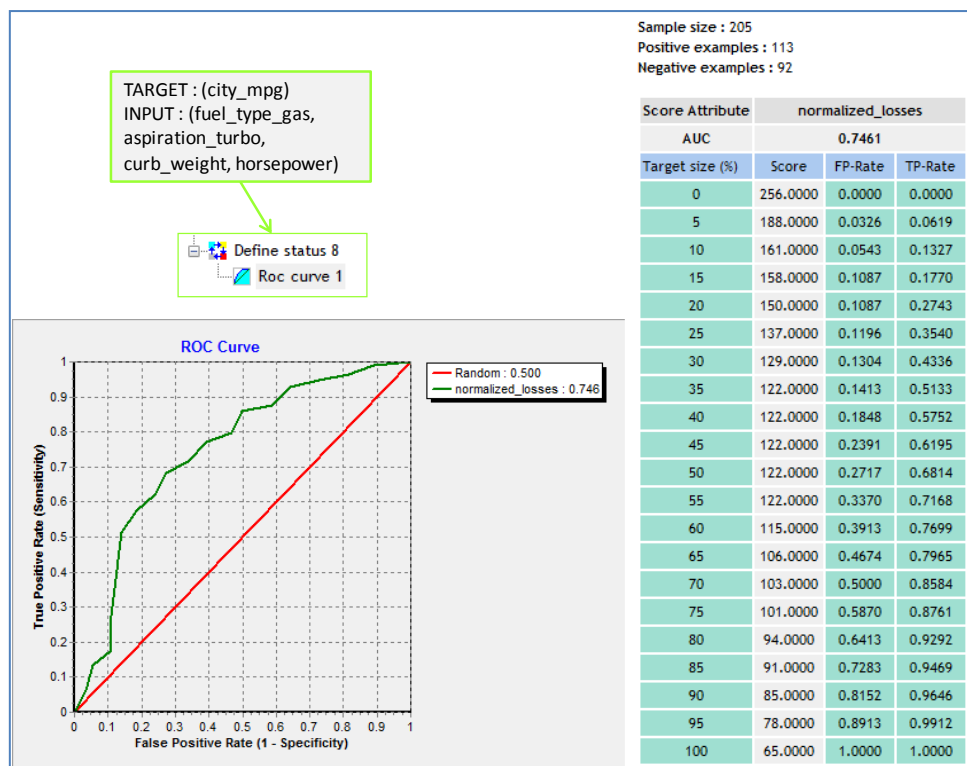
Les individus « positifs » correspondent à ceux portant la modalité « yes » de la variable « risky ». La précision est importante. SPSS calcule directement l'aire sous la courbe (AUC). Nous obtenons :

¹⁰ <http://archive.ics.uci.edu/ml/datasets/Automobile>



Il y a 113 positifs (« risky = yes ») dans le fichier. En les ordonnant les observations selon « normalized_losses », pour deux véhicules pris au hasard, il y a AUC = 74% de chances qu'un véhicule risqué soit placé devant un véhicule non risqué. « Normalized_losses » est plutôt un bon indicateur pour distinguer les véhicules qui seront classés risqués.

Pour une fois, **Tanagra** semble réagir différemment. En réalité, il n'en est rien. Il calcule le graphique en découpant la plage des valeurs en 20 intervalles de fréquences égales (cf. tableau des valeurs). **Ce qui produit une sorte de lissage** et évite le « creux » intempestif du graphique de PSPP. Mais fondamentalement, le résultat est le même, comme l'indique le critère AUC = 74.6%.

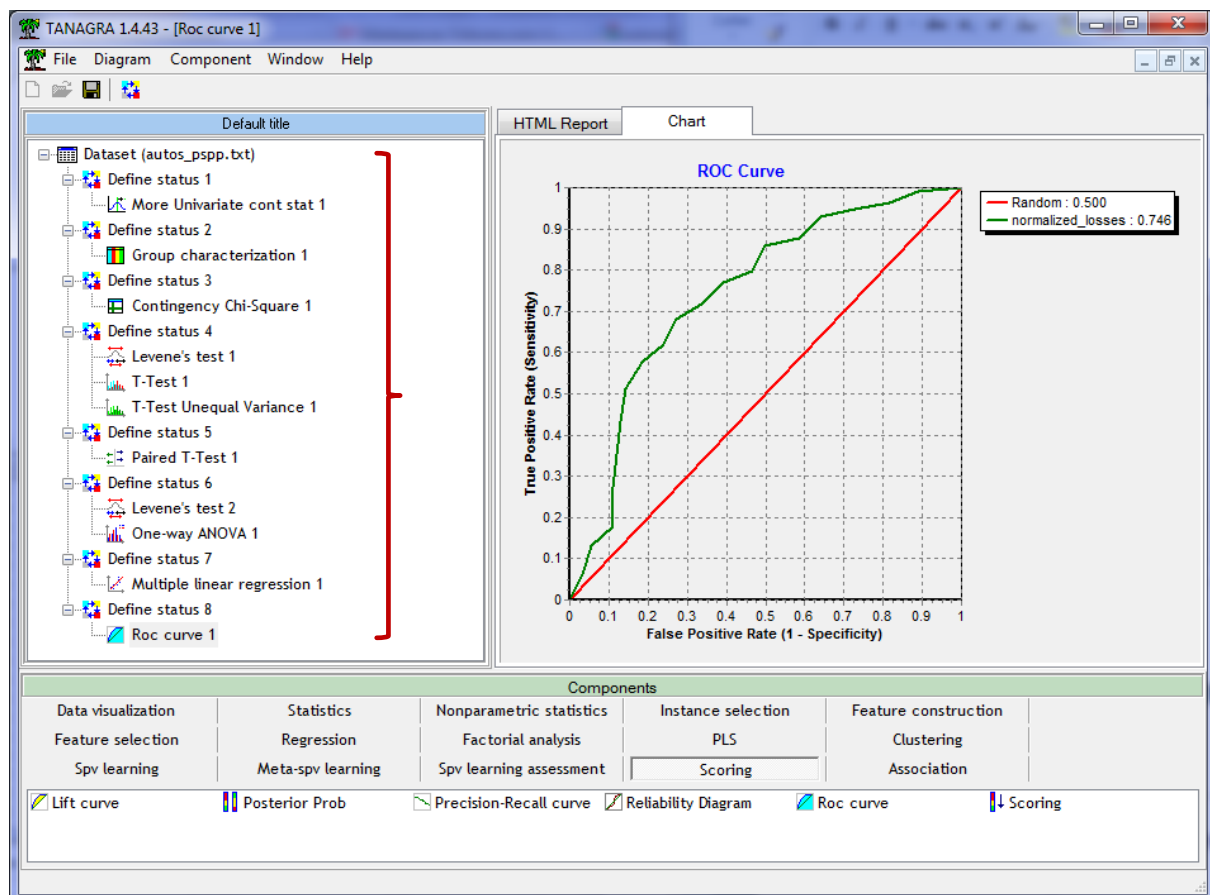


6 Traitements sous d'autres logiciels

Bien évidemment, les traitements réalisés dans PSPP peuvent l'être avec d'autres logiciels. Le choix nous appartient et c'est très bien ainsi. Pour illustrer cela, nous montrons ce que l'on peut faire avec les logiciels **Tanagra**, **R 2.13.2** et **OpenStat** (version du 24/02/2012) qui est très similaire à PSPP.

6.1 Traitements sous Tanagra

Certaines techniques présentes dans PSPP ne le sont pas dans Tanagra, et inversement. Pour celles que nous avons abordées dans ce document, voici le diagramme de traitements.



6.2 Traitements sous R

Mis à part la courbe ROC, nous reproduisons ici les commandes associées et les résultats obtenus avec le [logiciel R](#)¹¹. Dans certains cas, nous utilisons des packages spécifiques, repérés par leur chargement préalable à l'aide de la commande `library(.)`.

```
> #loading the dataset
> setwd("D:/DataMining/Databases_for_mining/logiciels_dataset/pspp")
> autos <- read.table(file="autos_pssp.txt",header=T,sep="\t",dec=".")
>
> #descriptive statistics
> print(summary(data.frame(autos$horsepower, autos$city_mpg)))
autos.horsepower autos.city_mpg
Min.      : 48.0    Min.      :13.00
```

¹¹ Les commandes sont en bleu, les commentaires en vert, les résultats en noir.


```

1st Qu.: 70.0    1st Qu.:19.00
Median : 95.0    Median :24.00
Mean   :104.3    Mean   :25.22
3rd Qu.:116.0    3rd Qu.:30.00
Max.   :288.0    Max.   :49.00
>
> #conditionnal descriptive statistics
>
print(tapply(autos$horsepower, autos$fuel_type, FUN=function(x) {c(m=mean(x), s=sd(x))}))
$diesel
      m      s
84.45000 25.95842

$gas
      m      s
106.39459 40.18342

> #crosstabs and test of independence
> library(gmodels)
>
print(CrossTable(autos$fuel_type, autos$aspiration, prop.r=F, prop.c=F, prop.t=F, chisq=T))

Cell Contents
|-----|
|                N |
| Chi-square contribution |
|-----|

Total Observations in Table:  205

autos$fuel_type | autos$aspiration
                |      std |      turbo | Row Total |
-----|-----|-----|-----|
      diesel    |         7 |         13 |         20 |
                |    5.380 |    24.427 |            |
-----|-----|-----|-----|
      gas       |       161 |         24 |        185 |
                |    0.582 |     2.641 |            |
-----|-----|-----|-----|
Column Total   |       168 |         37 |        205 |
-----|-----|-----|-----|

Statistics for All Table Factors

Pearson's Chi-squared test
-----
Chi^2 = 33.02955    d.f. = 1    p = 9.076896e-09

Pearson's Chi-squared test with Yates' continuity correction
-----
Chi^2 = 29.60576    d.f. = 1    p = 5.294738e-08

$t
      y
x      std turbo
diesel  7     13
gas     161   24

$prop.row

```

```

      y
x      std      turbo
diesel 0.3500000 0.6500000
gas    0.8702703 0.1297297

$prop.col
      y
x      std      turbo
diesel 0.04166667 0.35135135
gas    0.95833333 0.64864865

$prop.tbl
      y
x      std      turbo
diesel 0.03414634 0.06341463
gas    0.78536585 0.11707317

$chisq

      Pearson's Chi-squared test

data:  t
X-squared = 33.0295, df = 1, p-value = 9.077e-09

$chisq.corr

      Pearson's Chi-squared test with Yates' continuity correction

data:  t
X-squared = 29.6058, df = 1, p-value = 5.295e-08

> #Levene test for variance homogeneity
> library(lawstat)
> print(levene.test(autos$horsepower, autos$fuel_type, location="mean"))

      classical Levene's test based on the absolute deviations from the
mean ( none not applied
      because the location is not set to median )

data:  autos$horsepower
Test Statistic = 1.9242, p-value = 0.1669

> #t-test for independent samples
> print(t.test(autos$horsepower ~ autos$fuel_type, var.equal=T))

      Two Sample t-test

data:  autos$horsepower by autos$fuel_type
t = -2.3861, df = 203, p-value = 0.01795
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-40.078454 -3.810736
sample estimates:
mean in group diesel      mean in group gas
      84.4500              106.3946

> #Welch t-test for independent samples
> print(t.test(autos$horsepower ~ autos$fuel_type, var.equal=F))

      Welch Two Sample t-test

```

```

data: autos$horsepower by autos$fuel_type
t = -3.3693, df = 29.912, p-value = 0.00209
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -35.247706 -8.641483
sample estimates:
mean in group diesel      mean in group gas
           84.4500              106.3946

> #t-test for paired samples
> print(t.test(autos$city_mpg,autos$highway_mpg, paired=T))

      Paired t-test

data: autos$city_mpg and autos$highway_mpg
t = -48.1901, df = 204, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.758033 -5.305382
sample estimates:
mean of the differences
           -5.531707

> #Levene test for variance homogeneity
> print(levene.test(autos$horsepower,autos$body_style,location="mean"))

      classical Levene's test based on the absolute deviations from the
mean ( none not applied
      because the location is not set to median )

data: autos$horsepower
Test Statistic = 1.6904, p-value = 0.1536

> #analysis of variance
> print(aov(horsepower ~ body_style, data = autos))
Call:
  aov(formula = horsepower ~ body_style, data = autos)

Terms:
              body_style Residuals
Sum of Squares    17744.68 300856.13
Deg. of Freedom         4         200

Residual standard error: 38.78506
Estimated effects may be unbalanced
>
> #linear regression
> print(summary(lm(city_mpg
fuel_type_gas+aspiration_turbo+curb_weight+horsepower, data=autos)))

Call:
lm(formula = city_mpg ~ fuel_type_gas + aspiration_turbo + curb_weight +
    horsepower, data = autos)

Residuals:
    Min       1Q   Median       3Q      Max
-9.1931 -1.4955 -0.1292  0.8772 15.8097

Coefficients:
              Estimate Std. Error t value Pr(>|t|)

```

```
(Intercept)      57.6402999  1.6816779  34.275 < 2e-16 ***
fuel_type_gas   -8.5971662  0.9293722  -9.251 < 2e-16 ***
aspiration_turbo -1.6411239  0.6368442  -2.577  0.0107 *
curb_weight     -0.0078871  0.0007224 -10.918 < 2e-16 ***
horsepower      -0.0403830  0.0096536  -4.183  4.3e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

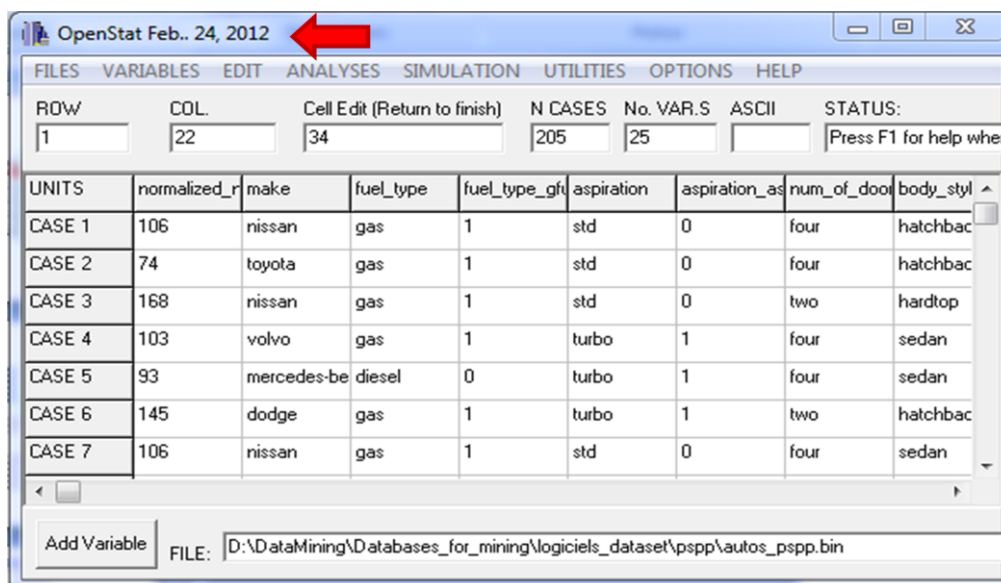
Residual standard error: 3.021 on 200 degrees of freedom
Multiple R-squared:  0.791,    Adjusted R-squared:  0.7868
F-statistic: 189.2 on 4 and 200 DF,  p-value: < 2.2e-16
```

L'écueil toujours dans R est d'identifier la commande adéquate pour les traitements. Mais ce n'est pas si pénalisant, on peut toujours trouver en ligne (*merci Google*) des indications plus ou moins pertinentes. Le site « [Quick-R](#) » par exemple est d'une très grande aide. Enfin, si avoir à saisir manuellement des instructions vous rebute vraiment, vous pouvez toujours passer par une surcouche qui permet de piloter R en passant par des menus. [R-Commander](#) semble très bien pour cela. Nous retrouvons dans ce document une grande partie des techniques décrites dans ce didacticiel.

6.3 Traitements sous OpenStat

[OpenStat](#) se positionne également comme une alternative à SPSS. Après avoir été payant pendant une certaine (très courte) période, il est de nouveau gratuit aujourd'hui. Il est beaucoup plus complet que PSPP. Nous l'avons déjà décrit dans un précédent tutoriel, ou plutôt son cousin [LazStats](#)¹². Il s'agissait alors de réaliser une régression linéaire multiple, avec et sans sélection de variables. Nous aurons encore l'occasion de revenir plusieurs fois sur ce logiciel tant il est riche.

Tout comme PSPP, il faut tout d'abord importer les données au format texte (FILES / IMPORT TAB FILE). La première correspond au nom des variables. Nous obtenons la grille des données.

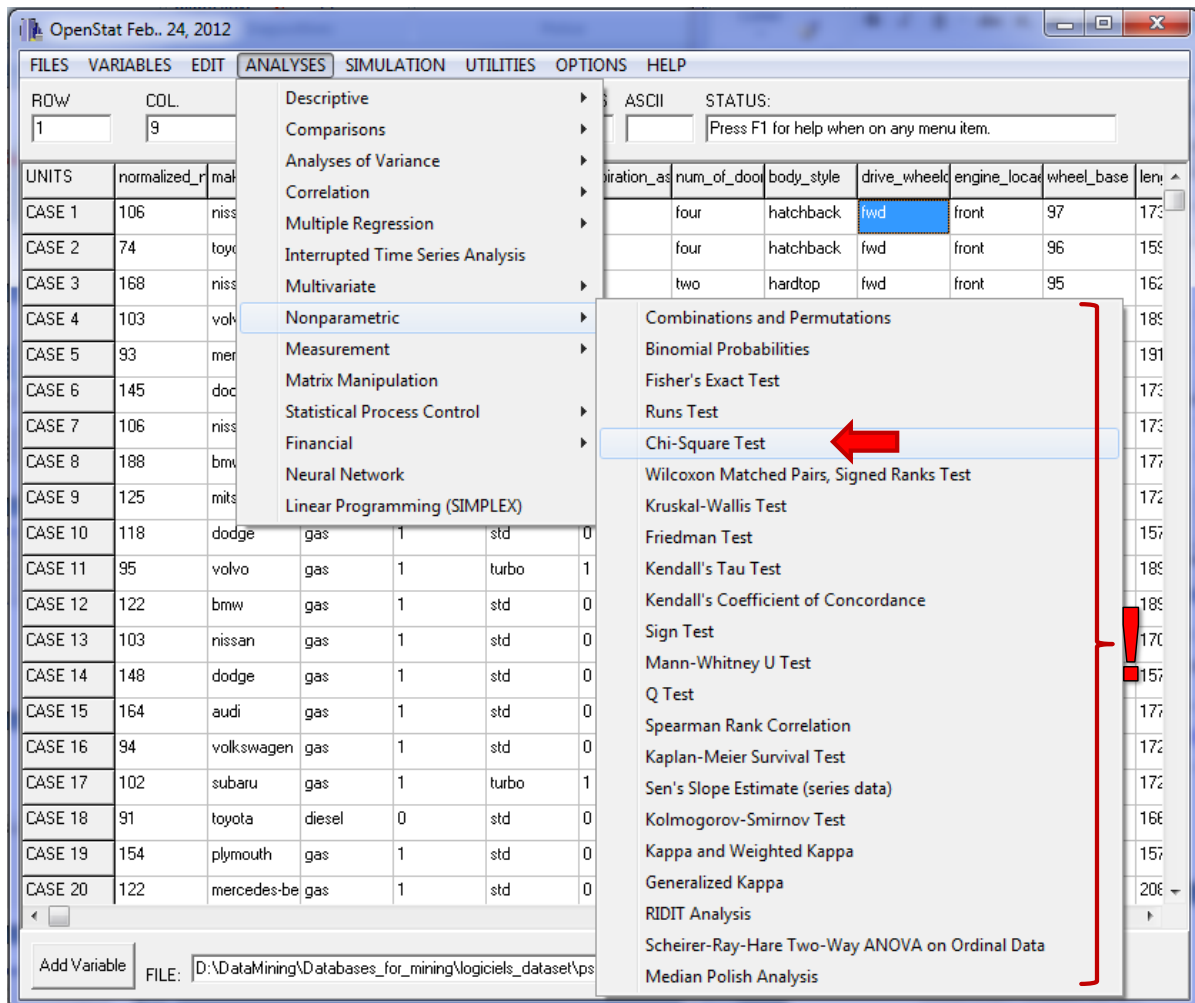


Nous n'énumérerons pas les traitements ci-dessus. Je l'ai fait par ailleurs, les résultats sont complètement conformes. Nous nous contenterons de reproduire le test d'indépendance du KHI-2

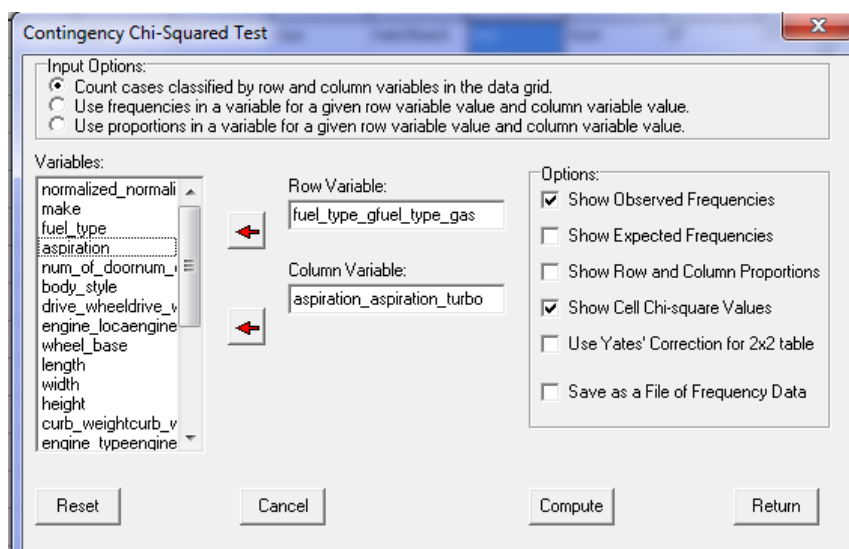
¹² <http://tutoriels-data-mining.blogspot.com/2011/05/regression-avec-le-logiciel-lazstats.html>

entre « fuel_type » et « aspiration ». **Attention, nous devons utiliser les indicatrices lors du paramétrage du logiciel.**

Comme PSPP, les techniques statistiques disponibles sont regroupées dans le menu ANALYSES. Nous sélectionnons l'item NONPARAMETRIC. La liste est longue !



Pour le test du KH-2, nous spécifions les paramètres suivants puis nous cliquons sur COMPUTE.



Une fenêtre de visualisation est affichée. Les résultats sont identiques à ceux de PSPP, Tanagra et R. OpenStat propose des indicateurs supplémentaires.

No. of Cases = 205

OBSERVED FREQUENCIES

	Frequencies		Total
	COL. 1	COL. 2	
Row 1	7	13	20
Row 2	161	24	185
Total	168	37	205

CHI-SQUARED VALUE FOR CELLS

	Chi-square Values	
	COL. 1	COL. 2
Row 1	5.380	24.427
Row 2	0.582	2.641

Chi-square = 33.030 with D.F. = 1. Prob. > value = 0.000

Likelihood Ratio = 24.904 with prob. > value = 0.0000

G statistic = 24.904 with prob. > value = 0.0000

phi correlation = 0.4014

Pearson Correlation r = -0.4014

Mantel-Haenszel Test of Linear Association = 32.868 with probability > value = 0.0000

The coefficient of contingency = 0.373

Cramer's V = 0.401

Contrairement à PSPP, OpenStat est piloté exclusivement par menu. Il n'est pas possible de conserver sur un support externe la séquence des commandes associées aux traitements réalisés. Reproduire l'analyse à l'identique sur une mise à jour du fichier (ex. de nouvelles observations ont été recueillies) n'est pas très aisé.

7 Conclusion

PSPP est un outil en devenir. La structure a manifestement bien été pensée. La possibilité de retranscrire dans la syntaxe PSPP les commandes définies par menu est un atout indéniable. C'est, entre autres, une excellente manière de faire l'apprentissage du langage de programmation. Je m'en suis d'ailleurs inspiré pour créer le fichier script décrit dans la section 3.2. Par exemple, pour l'analyse de variance (One Way Anova) de la section 5.6, PSPP génère la commande suivante.

```
ONEWAY /VARIABLES= horsepower BY body_style
/STATISTICS=HOMOGENEITY .
```

D'autres méthodes statistiques sont accessibles dans PSPP (statistiques non paramétriques – ex. test des rangs signés de Wilcoxon, test de Friedman ; classification automatique avec la méthode des K-Means ; analyse en composantes principales). Le logiciel couvre déjà une bonne partie des techniques usuelles. Il sera certainement complété au fil du temps. C'est un logiciel que je suivrai avec beaucoup d'intérêt.