1 Topic

PSPP, an alternative to SPSS.

I spend a lot of time to analyze the available free statistical and data mining tools. There is not bad software, but some tools are more appropriate for some tasks. Thus, we must identify the one which is the best suited to our configuration. For that, we must know a large number of tools.

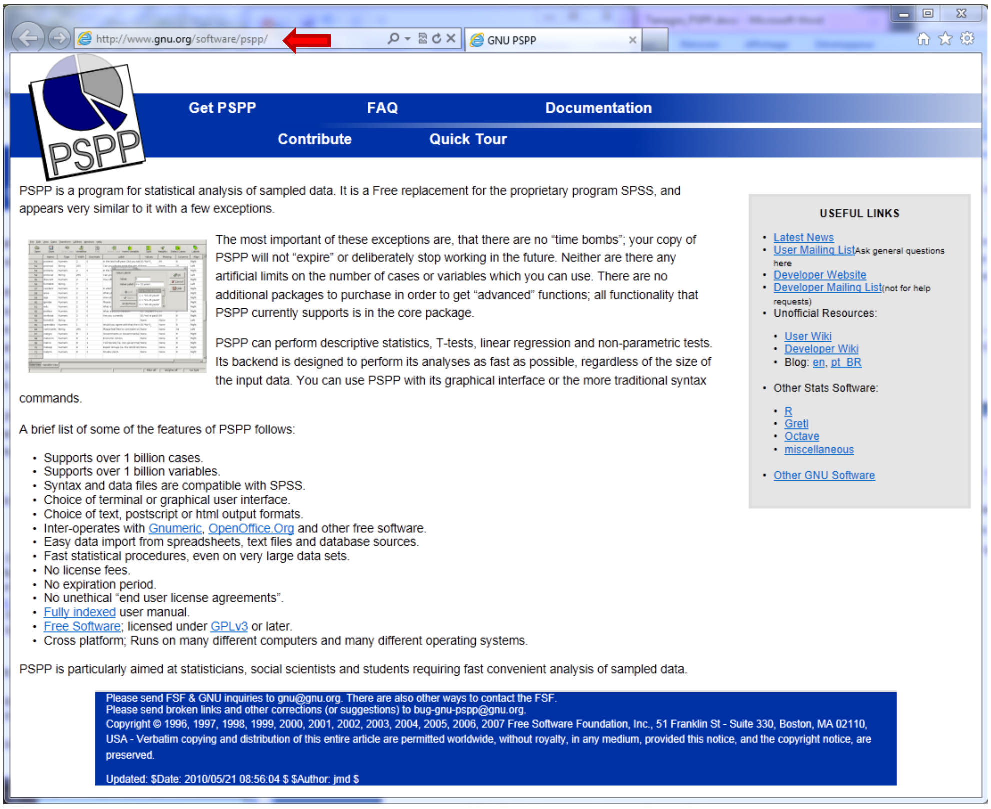
In this tutorial, we describe PSPP. It is presented as an alternative to the well-known SPSS: "*PSPP is a program for statistical analysis of sampled data. It is a free replacement for the proprietary program SPSS, and appears very similar to it with a few exceptions*[1]". Instead of to describe in detail each feature, the documentation is available on the website, we present some statistical techniques. We compare the results with those of **Tanagra**, **R 2.13.2** and **OpenStat (build 24/02/2012)**. This is also a way to validate them. If they provide different results, it means that there is a problem.

2 Dataset

We use a version of the "Automobile" dataset from the UCI server[2]. According to the statistical method that we analyze, we use some variables of the dataset. The most important here is to show how to perform the various analyses with PSPP.

3 PSPP

3.1 Loading and installing PSPP



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.

The most important of these exceptions are, that there are no "time bombs"; your copy of PSPP will not "expire" or deliberately stop working in the future. Neither are there any artificial limits on the number of cases or variables which you can use. There are no additional packages to purchase in order to get "advanced" functions; all functionality that PSPP currently supports is in the core package.

PSPP can perform descriptive statistics, T-tests, linear regression and non-parametric tests. Its backend is designed to perform its analyses as fast as possible, regardless of the size of the input data. You can use PSPP with its graphical interface or the more traditional syntax commands.

A brief list of some of the features of PSPP follows:

- Supports over 1 billion cases.
- Supports over 1 billion variables.
- Syntax and data files are compatible with SPSS.
- Choice of terminal or graphical user interface.
- Choice of text, postscript or html output formats.
- Inter-operates with Gnumeric, OpenOffice.Org and other free software.
- Easy data import from spreadsheets, text files and database sources.
- Fast statistical procedures, even on very large data sets.
- No license fees.
- No expiration period.
- No unethical "end user license agreements".
- Fully indexed user manual.
- Free Software; licensed under GPLv3 or later.
- Cross platform; Runs on many different computers and many different operating systems.

PSPP is particularly aimed at statisticians, social scientists and students requiring fast convenient analysis of sampled data.

USEFUL LINKS

- Latest News
- User Mailing ListAsk general questions here
- Developer Website
- Developer Mailing List(not for help requests)
- Unofficial Resources:
  - User Wiki
  - Developer Wiki
  - Blog: en, pt_BR
- Other Stats Software:
  - R
  - Gretl
  - Octave
  - miscellaneous
- Other GNU Software

Please send FSF & GNU inquiries to gnu@gnu.org. There are also other ways to contact the FSF.
Please send broken links and other corrections (or suggestions) to bug-gnu-pspp@gnu.org.
Copyright © 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007 Free Software Foundation, Inc., 51 Franklin St - Suite 330, Boston, MA 02110, USA - Verbatim copying and distribution of this entire article are permitted worldwide, without royalty, in any medium, provided this notice, and the copyright notice, are preserved.

Updated: $Date: 2010/05/21 08:56:04 $ $Author: jmd $

[1] http://www.gnu.org/software/pspp/pspp.html

[2] http://archive.ics.uci.edu/ml/datasets/Automobile

PSPP is downloadable online (http://www.gnu.org/software/pspp/). We use the 0.7.8 (2011/11/11) version. PSPP needs the MINGW[3] environment, but fortunately, the installation of this last one is done automatically. Thus, the installation process under Windows is performed easily.

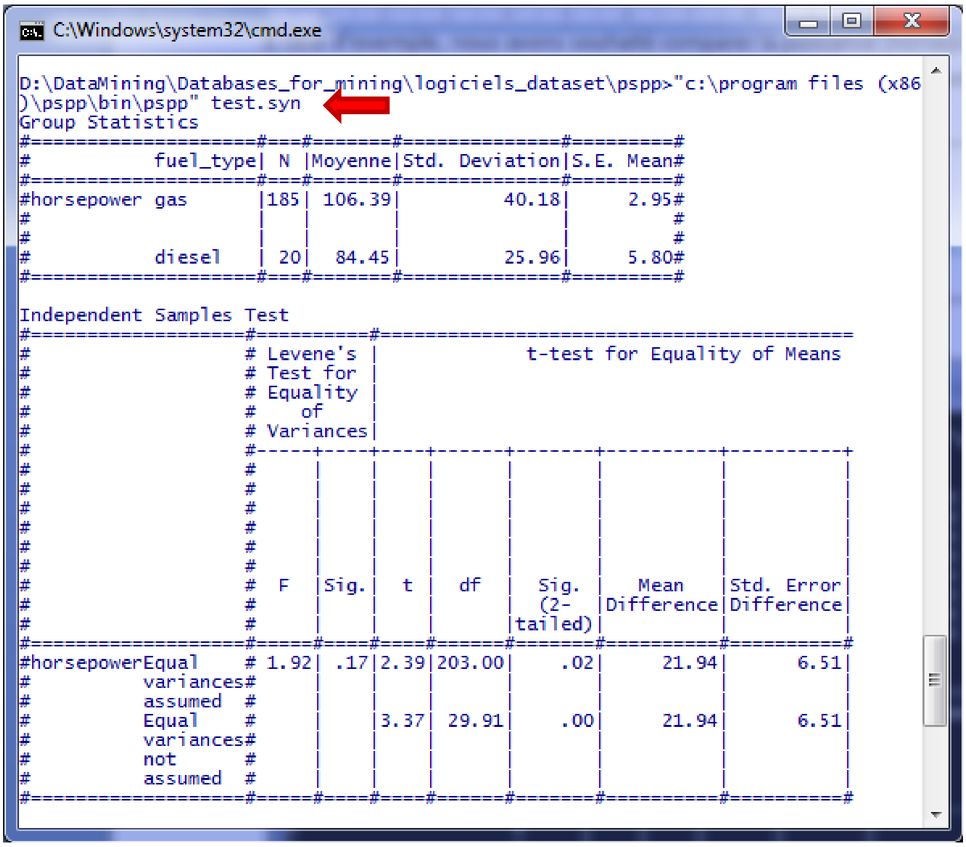
3.2 Command line mode

We can set instructions in a program file (e.g. with a standard text editor) and send this one to the executable file PSPP.EXE. The syntax of the commands is the same to the one of SPSS. By learning to program with PSPP, we will know to do this with SPSS. This is really interesting.

For instance, we want to compare the horsepower of the cars according to their fuel type using a t-test. We set the following commands into the "**test.syn**" program file.

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

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

The results are displayed into the MSDOS console.



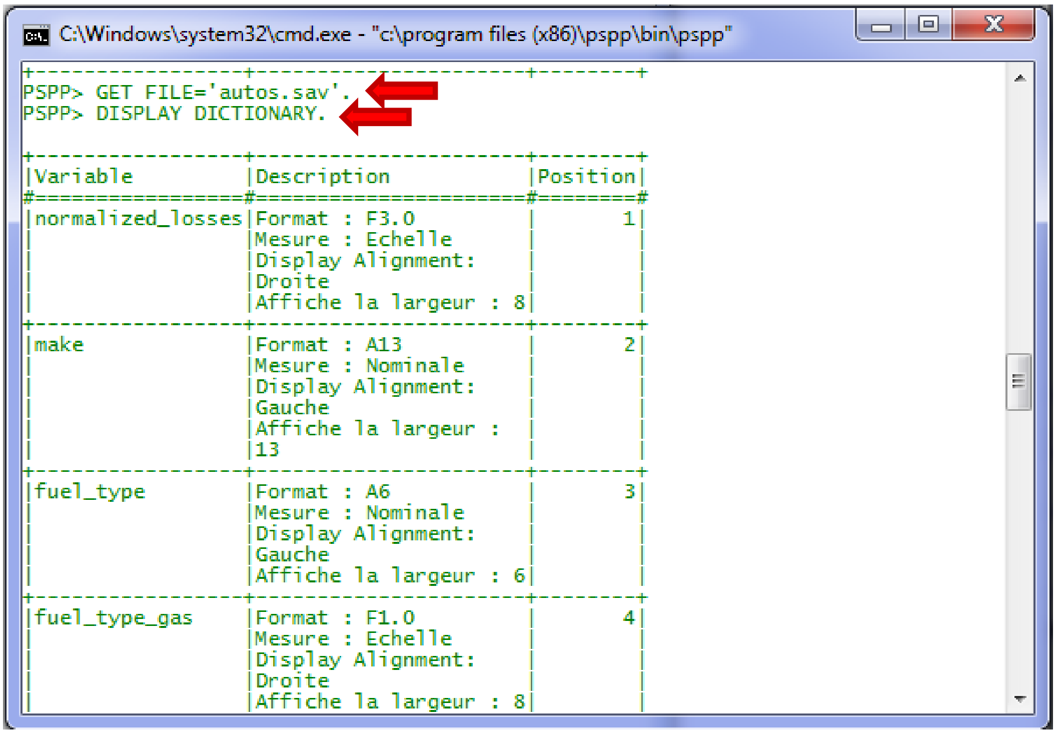
```
D:\DataMining\Databases_for_mining\logiciels_dataset\pspp>"c:\program files (x86
)\pspp\bin\pspp" test.syn
Group Statistics
#====================#===#=======#==============#=========#
#           fuel_type| N |Moyenne|Std. Deviation|S.E. Mean#
#====================#===#=======#==============#=========#
#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    #     |    |3.37| 29.91|    .00|     21.94|      6.51|
#          variances#     |    |    |      |       |          |          |
#          not      #     |    |    |      |       |          |          |
#          assumed  #     |    |    |      |       |          |          |
#===================#=====#====#====#======#=======#==========#==========#
```

PSPP compares first the conditional variances with the Levene's test. Then, it compares the means with and without the homoscedasticity assumption.

[3] http://www.mingw.org/

3.3 Terminal mode

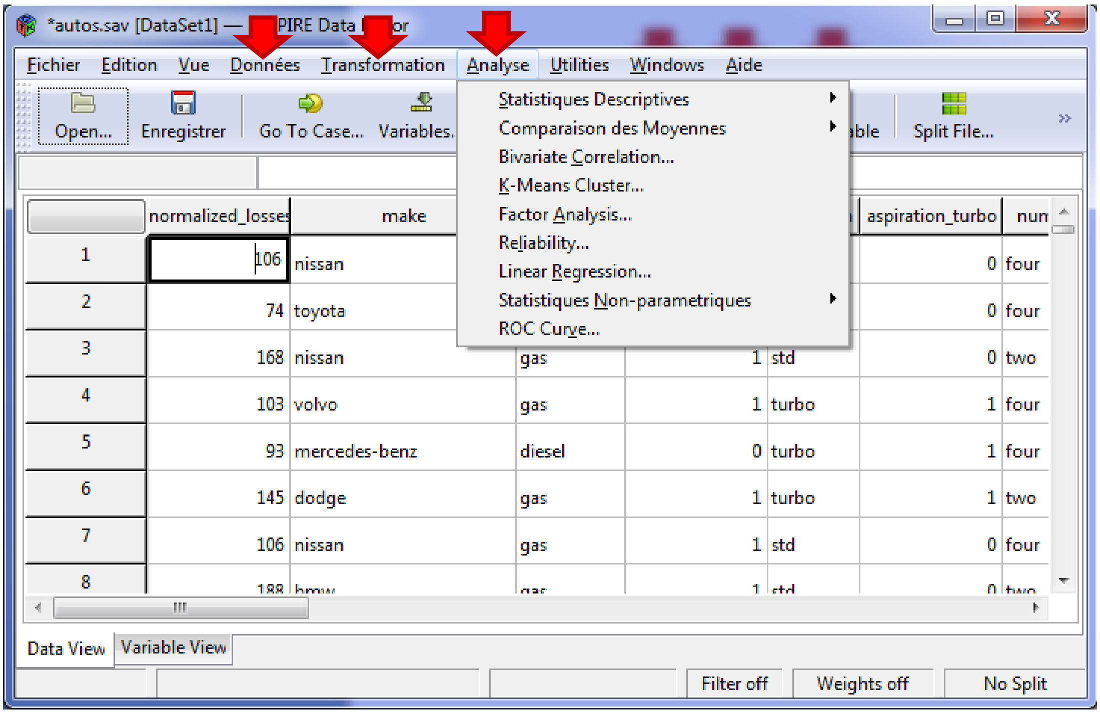
We can use also PSPP in a terminal mode. After we launch PSPP.EXE, a command interpreter allows us to set the instructions. The results are displayed into the same window. Here, we load the “autos.sav” data file, and we display the dataset dictionary.



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 Menu-driven mode

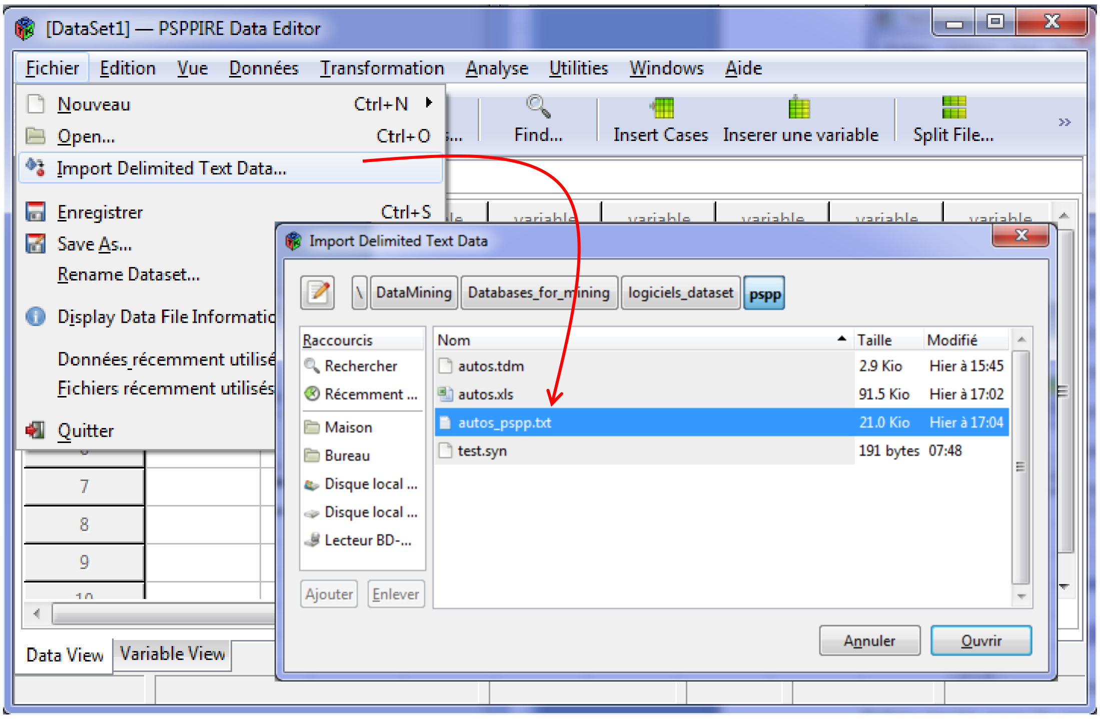


	normalized_losses	make				aspiration_turbo	num
1	106	nissan				0	four
2	74	toyota				0	four
3	168	nissan	gas	1	std	0	two
4	103	volvo	gas	1	turbo	1	four
5	93	mercedes-benz	diesel	0	turbo	1	four
6	145	dodge	gas	1	turbo	1	two
7	106	nissan	gas	1	std	0	four
8	188	bmw	gas	1	std	0	two

The easiest way to handle PSPP is the menu-guided mode. We use it in this tutorial. The features are grouped in menus “données” (dataset), “transformation”, and “analyse” (analysis). *Curiously,* *although that I have installed the English version, some menu-items are in French on my computer. I do not know if we have the same phenomenon with an English language operating system*.

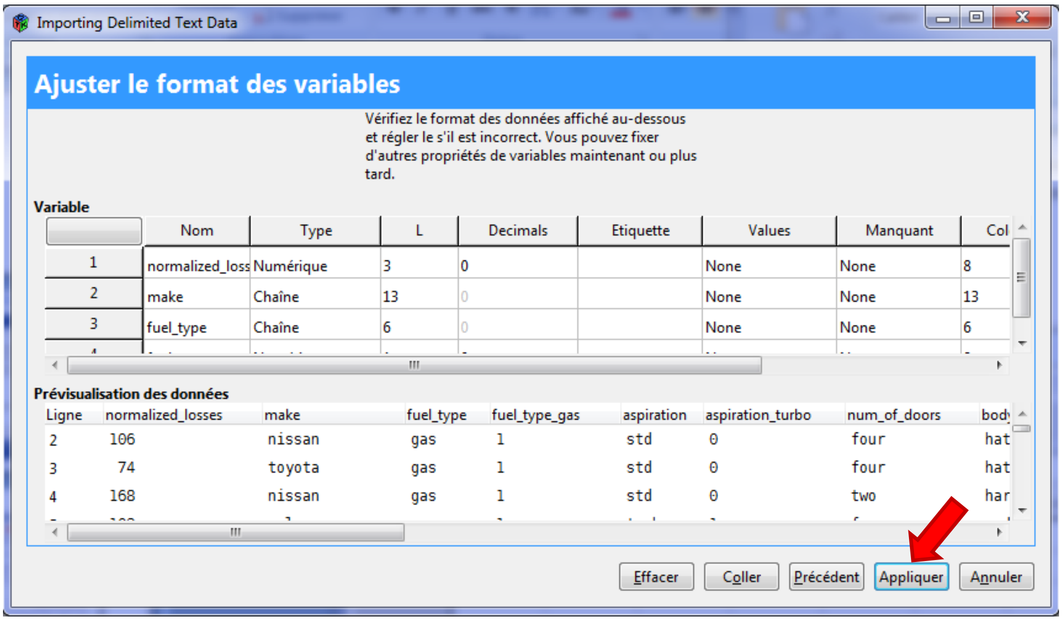
4 Importing the data file

First, we must import the « **autos_pspp.txt** » tab delimited text file. We click on the FICHIER / IMPORT DELIMITED TEXT DATA menu. We pick the file into the dialog box.



Nom	Taille	Modifié
autos.tdm	2.9 Kio	Hier à 15:45
autos.xls	91.5 Kio	Hier à 17:02
autos_pspp.txt	21.0 Kio	Hier à 17:04
test.syn	191 bytes	07:48

A wizard appears. We must: (1) import all the instances; (2) the first row corresponds to the variable name; (3) the TAB character is the column separator; (4) the values are string (chaîne) or numeric (numérique). We click on the APPLIQUER button.


Importing Delimited Text Data
Ajuster le format des variables
Effacer
Coller
Précédent
Appliquer
Annuler

Vérifiez le format des données affiché au-dessous et régler le s'il est incorrect. Vous pouvez fixer d'autres propriétés de variables maintenant ou plus tard.

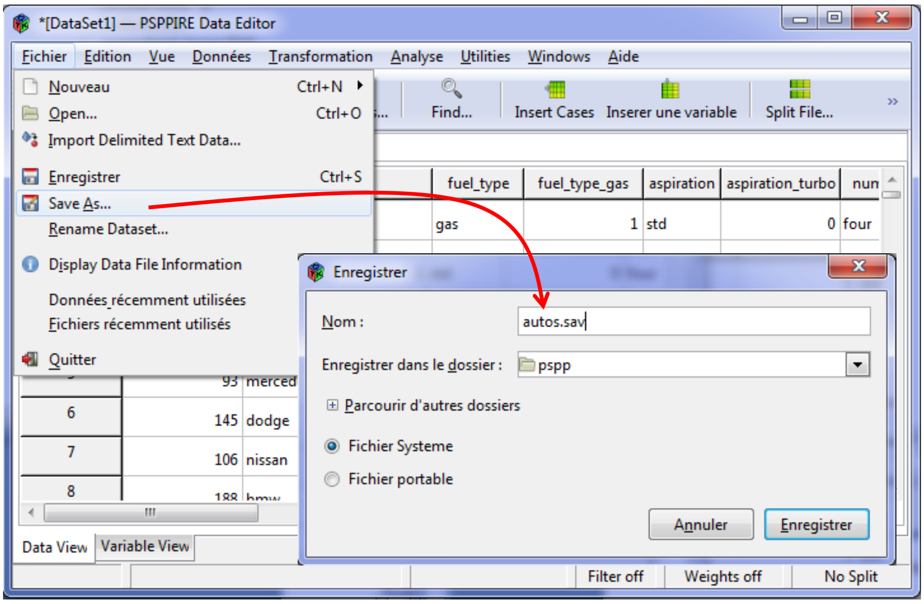
Variable

	Nom	Type	L	Decimals	Etiquette	Values	Manquant	Col
1	normalized_loss	Numérique	3	0		None	None	8
2	make	Chaîne	13	0		None	None	13
3	fuel_type	Chaîne	6	0		None	None	6

Prévisualisation des données

Ligne	normalized_losses	make	fuel_type	fuel_type_gas	aspiration	aspiration_turbo	num_of_doors	bod
2	106	nissan	gas	1	std	0	four	hat
3	74	toyota	gas	1	std	0	four	hat
4	168	nissan	gas	1	std	0	two	har

We save the loaded dataset into the PSPP native format ("**autos.sav**"). This is the same one as SPSS.


*[DataSet1] — PSPPIRE Data Editor
Fichier Edition Vue Données Transformation Analyse Utilities Windows Aide
Nouveau Ctrl+N
Open... Ctrl+O
Import Delimited Text Data...
Enregistrer Ctrl+S
Save As...
Rename Dataset...
Display Data File Information
Données récemment utilisées
Fichiers récemment utilisés
Quitter
Find... Insert Cases Inserer une variable Split File...
fuel_type fuel_type_gas aspiration aspiration_turbo nun
gas 1 std 0 four
Enregistrer
Parcourir d'autres dossiers
Annuler Enregistrer
Data View Variable View
Filter off Weights off No Split

Nom : autos.sav

Enregistrer dans le dossier : pspp

● Fichier Systeme

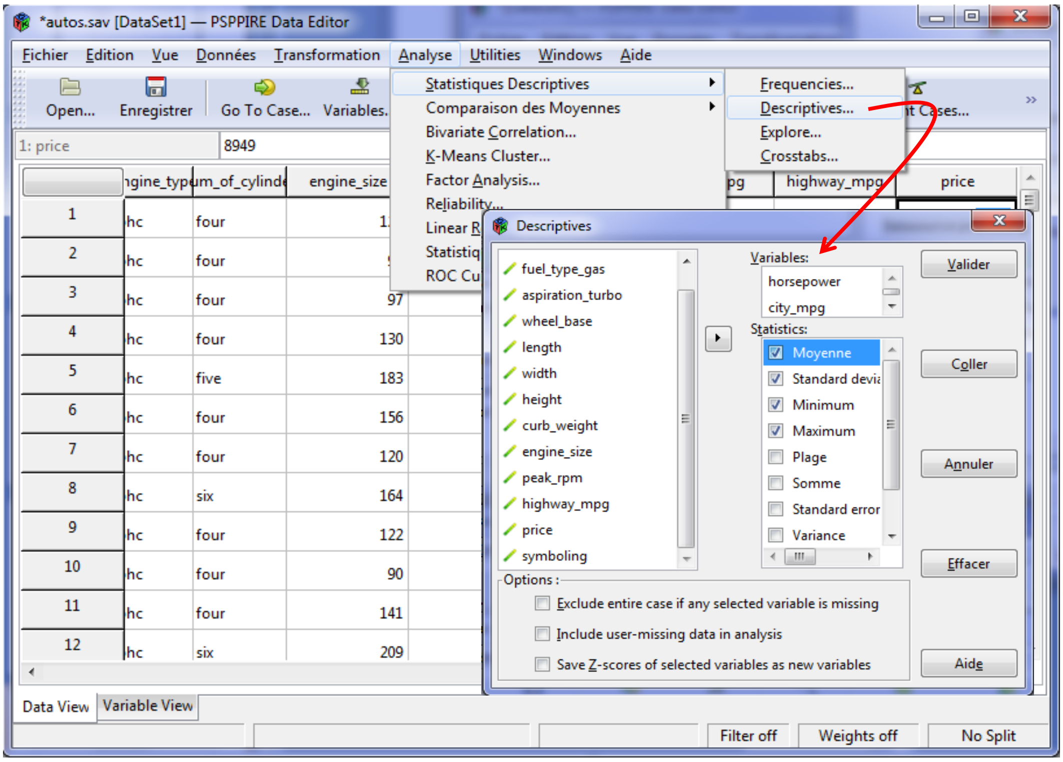
○ Fichier portable

5 A few statistical methods with PSPP

In this section, we present some statistical techniques available in PSPP. When this is possible, we compare the results with those of Tanagra.

5.1 Descriptive statistics – Numeric variables

To obtain the description for horsepower and city mpg, we click on the ANALYSE / STATISTIQUES DESCRIPTIVES / DESCRIPTIVES menu. Into the dialog settings, we select the indicators to compute.



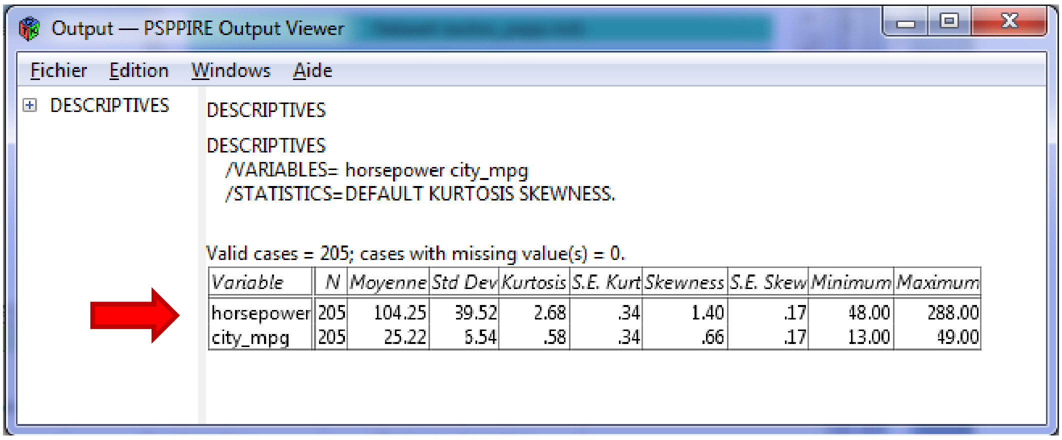
	ngine_typ	um_of_cylind	engine_size
1	hc	four	1
2	hc	four	
3	hc	four	97
4	hc	four	130
5	hc	five	183
6	hc	four	156
7	hc	four	120
8	hc	six	164
9	hc	four	122
10	hc	four	90
11	hc	four	141
12	hc	six	209

Variables:
horsepower
city_mpg

Statistics:
☑ Moyenne
☑ Standard devia
☑ Minimum
☑ Maximum
□ Plage
□ Somme
□ Standard error
□ Variance

Options :
□ Exclude entire case if any selected variable is missing
□ Include user-missing data in analysis
□ Save Z-scores of selected variables as new variables

The results are displayed into a new window ("Output viewer").



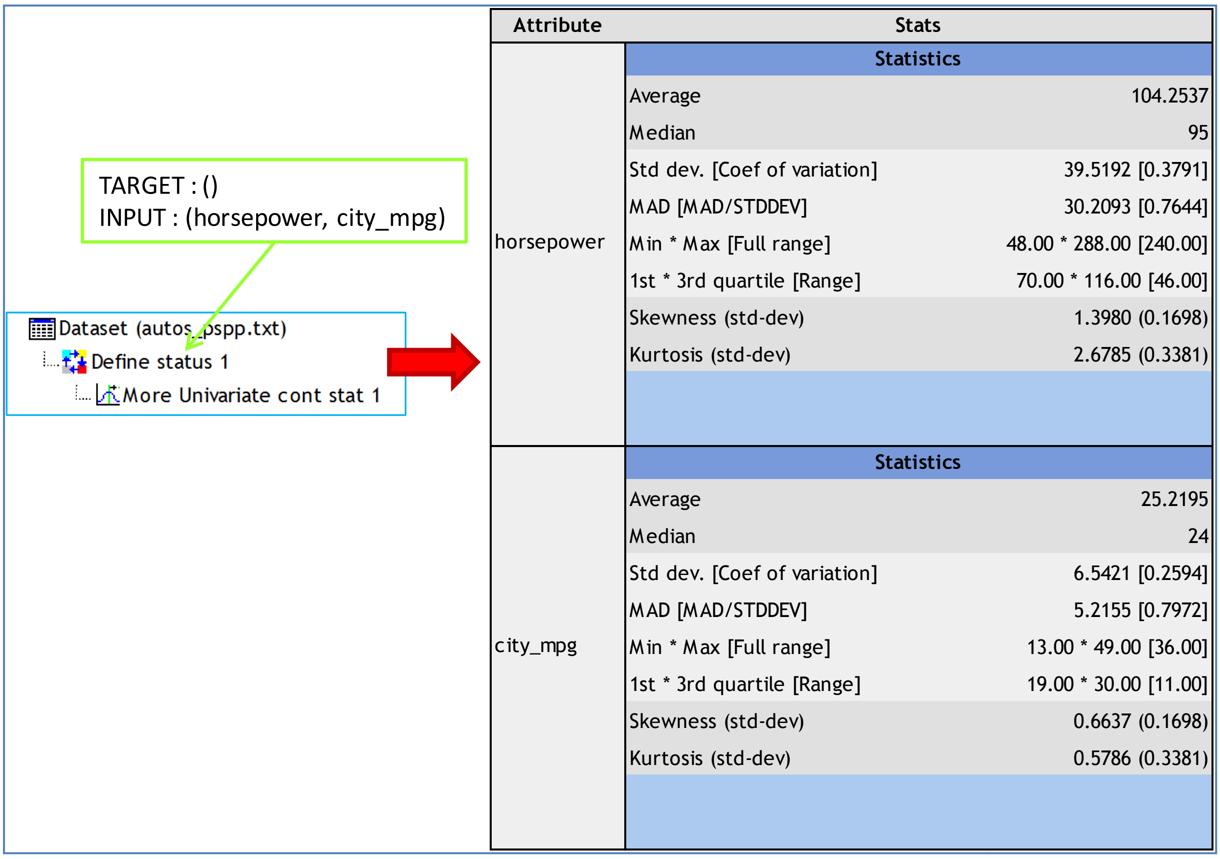
DESCRIPTIVES

```
DESCRIPTIVES
  /VARIABLES= horsepower city_mpg
  /STATISTICS=DEFAULT KURTOSIS SKEWNESS.
```

Valid cases = 205; cases with missing value(s) = 0.

Variable	N	Moyenne	Std Dev	Kurtosis	S.E. Kurt	Skewness	S.E. Skew	Minimum	Maximum
horsepower	205	104.25	39.52	2.68	.34	1.40	.17	48.00	288.00
city_mpg	205	25.22	6.54	.58	.34	.66	.17	13.00	49.00

The MORE UNIVARIATE CONT STAT of **Tanagra** provides the same values.

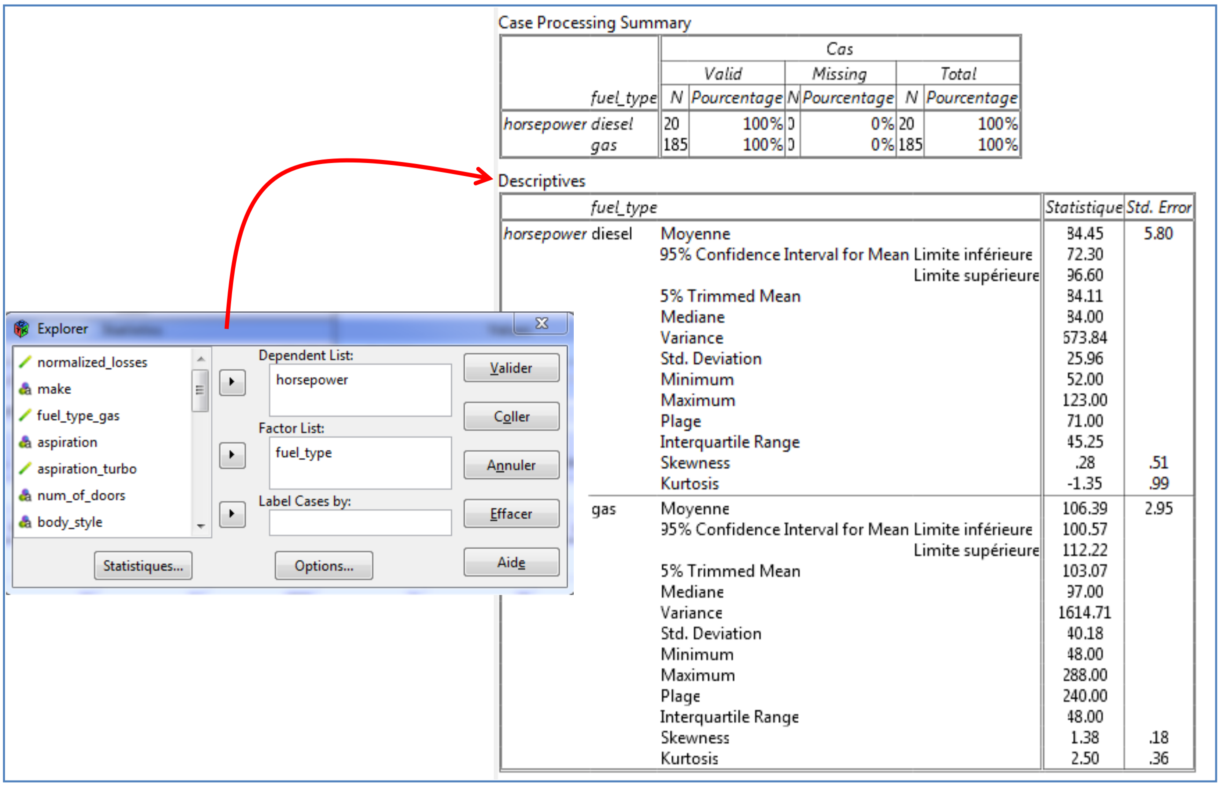


Attribute	Stats	
horsepower	Statistics	
	Average	104.2537
	Median	95
	Std dev. [Coef of variation]	39.5192 [0.3791]
	MAD [MAD/STDDEV]	30.2093 [0.7644]
	Min * Max [Full range]	48.00 * 288.00 [240.00]
	1st * 3rd quartile [Range]	70.00 * 116.00 [46.00]
	Skewness (std-dev)	1.3980 (0.1698)
	Kurtosis (std-dev)	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 Conditional descriptive statistics

We want calculate the descriptive statistics for horsepower according to the values of fuel type (“gas” or “diesel”).

We click on the ANALYSE / STATISTIQUES DESCRIPTIVES / EXPLORE menu. We set “horsepower” into DEPENDENT LIST, “fuel type” into FACTOR LIST. By clicking the STATISTICS button, we can specify the indicators to compute. We validate.



Dependent List: horsepower

Factor List: fuel_type

Label Cases by:

Case Processing Summary

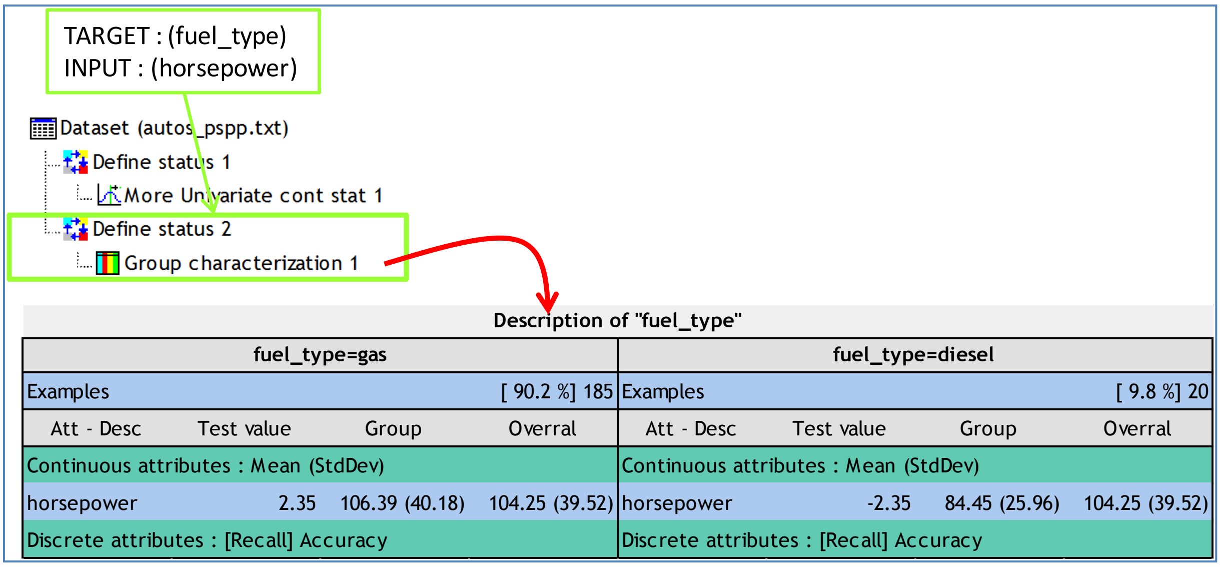
		Cas					
		Valid		Missing		Total	
	fuel_type	N	Pourcentage	N	Pourcentage	N	Pourcentage
horsepower	diesel	20	100%	0	0%	20	100%
	gas	185	100%	0	0%	185	100%

Descriptives

	fuel_type			Statistique	Std. Error
horsepower	diesel	Moyenne		84.45	5.80
		95% Confidence Interval for Mean	Limite inférieure	72.30	
			Limite supérieure	96.60	
		5% Trimmed Mean		84.11	
		Mediane		84.00	
		Variance		573.84	
		Std. Deviation		25.96	
		Minimum		52.00	
		Maximum		123.00	
		Plage		71.00	
		Interquartile Range		45.25	
		Skewness		.28	.51
		Kurtosis		-1.35	.99
	gas	Moyenne		106.39	2.95
		95% Confidence Interval for Mean	Limite inférieure	100.57	
			Limite supérieure	112.22	
		5% Trimmed Mean		103.07	
		Mediane		97.00	
		Variance		1614.71	
		Std. Deviation		40.18	
		Minimum		48.00	
		Maximum		288.00	
		Plage		240.00	
		Interquartile Range		48.00	
		Skewness		1.38	.18
		Kurtosis		2.50	.36

For instance, we observe that the mean of horsepower is 84.45 for the cars corresponding to "fuel type = gas", and 106.39 for "fuel type = diesel". We obtain very detailed results.

The GROUP CHARACTERIZATION component of **Tanagra** provides the same values, but the results are definitely less detailed.



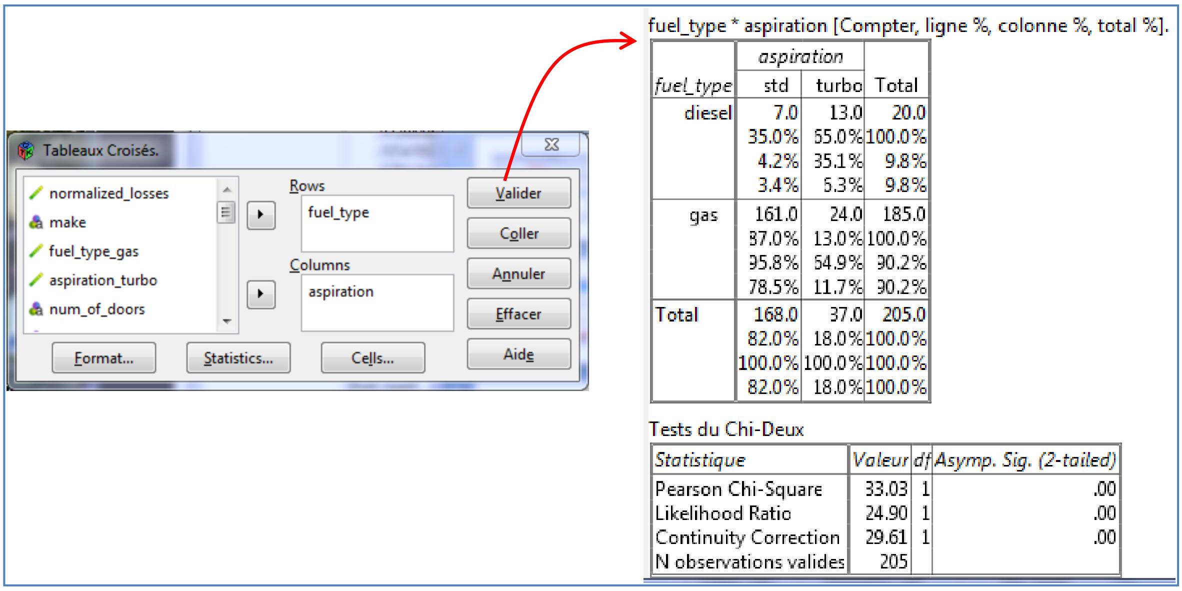
TARGET : (fuel_type)
INPUT : (horsepower)

Description of "fuel_type"

fuel_type=gas				fuel_type=diesel			
Examples			[90.2 %] 185	Examples			[9.8 %] 20
Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral
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 Contingency table

We want to analyze the relation between fuel type and aspiration ("turbo" or "standard") with a cross tab. We click on the ANALYSE / STATISTIQUES DESCRIPTIVES / CROSSTABS menu. We set the first variable into ROWS list, the second one into the COLUMNS.



Rows: fuel_type

Columns: aspiration

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

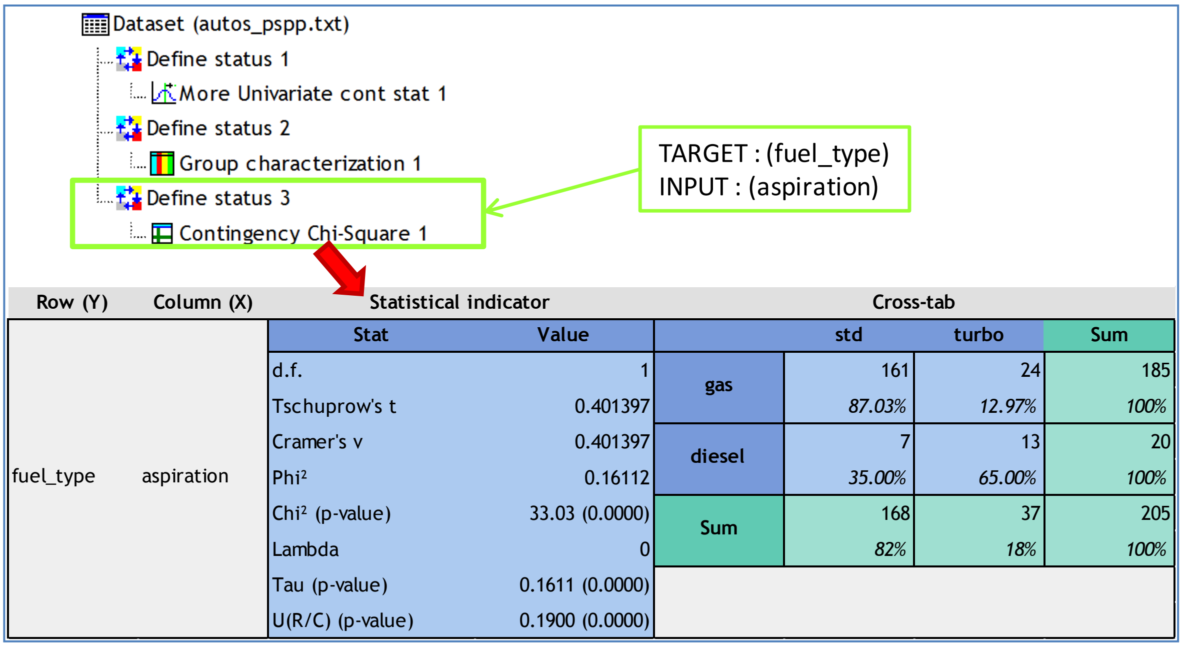
	aspiration		
fuel_type	std	turbo	Total
diesel	7.0	13.0	20.0
	35.0%	65.0%	100.0%
	4.2%	35.1%	9.8%
	3.4%	6.3%	9.8%
gas	161.0	24.0	185.0
	87.0%	13.0%	100.0%
	95.8%	64.9%	90.2%
	78.5%	11.7%	90.2%
Total	168.0	37.0	205.0
	82.0%	18.0%	100.0%
	100.0%	100.0%	100.0%
	82.0%	18.0%	100.0%

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		

Here also, PSPP can provide very detailed results. Some measures of association (Theil's U, Cohen's Kappa, etc.) and the various percentages are also displayed.

The same results are available under **Tanagra** with the CONTINGENCY CHI-SQUARE component. But the organization is a little different.



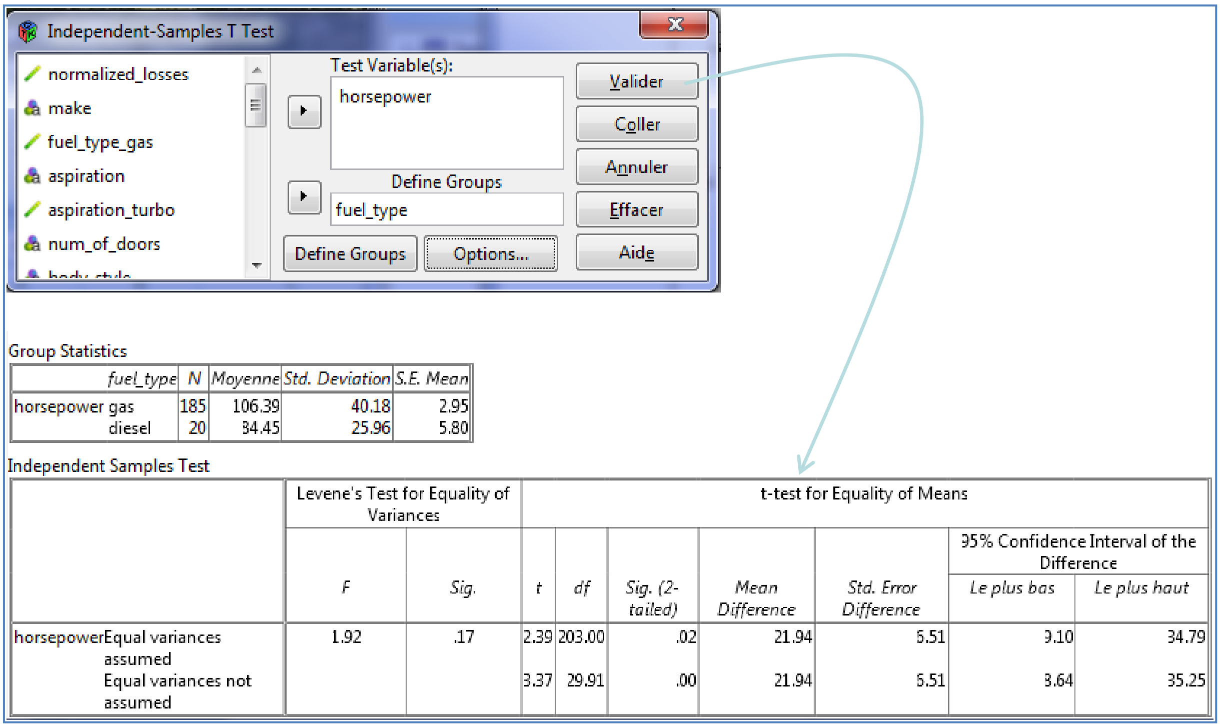
TARGET : (fuel_type)
INPUT : (aspiration)

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

5.4 Comparison of two means – Independent samples

Beyond the conditional descriptive statistics, we can test if the means are significantly different. We click on the ANALYSE / COMPARAISON DES MOYENNES / INDEPENDENT SAMPLES T TEST, we set horsepower into TEST VARIABLE list, fuel type into DEFINE GROUPS.

PSPP compares the conditional variances first by using the Levene's test. We observe that they are not significantly different at the 5% level (p-value = 0.17). However, it displays the comparison of means with and without the homoscedasticity assumption. In both cases, we see that the means are significantly different.



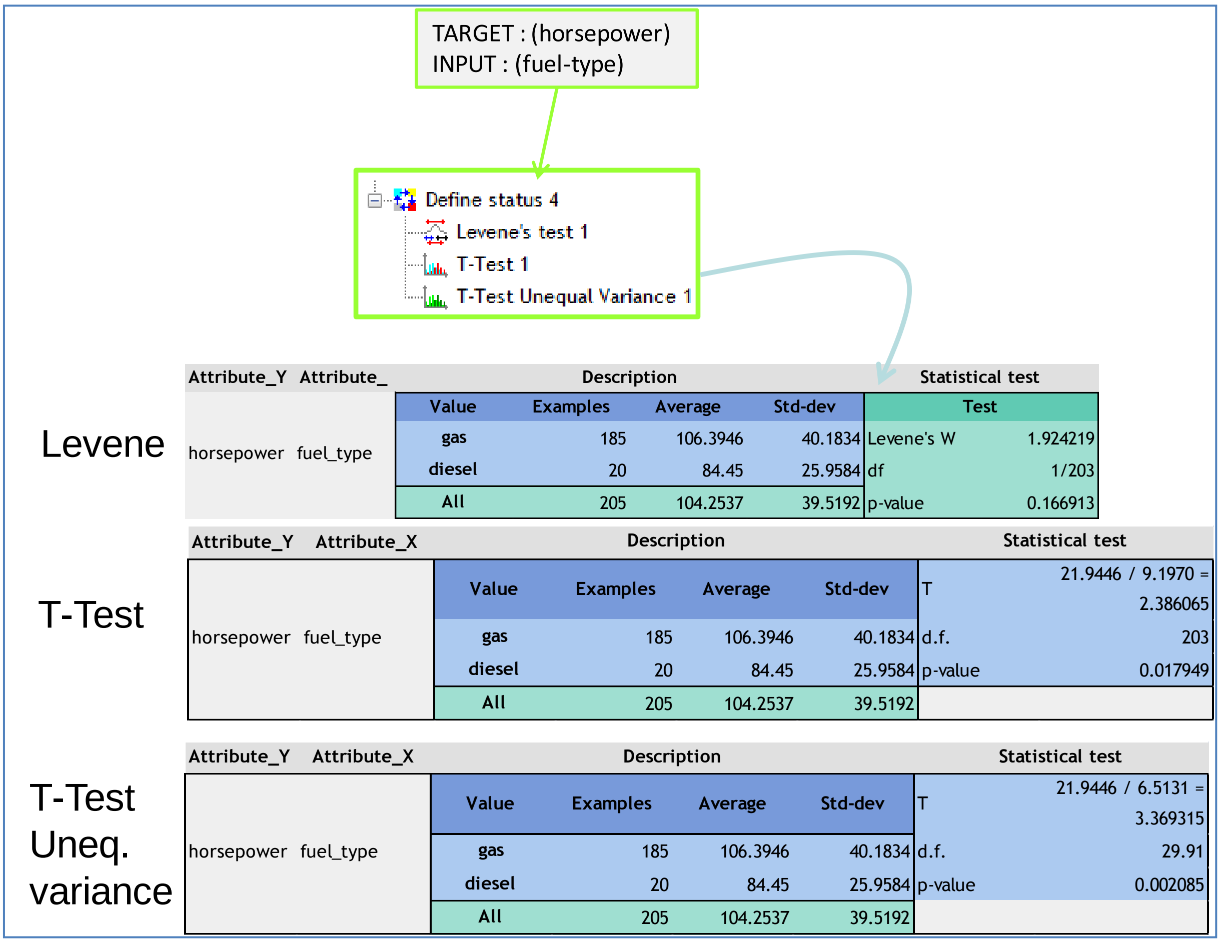
Group Statistics

	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 Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Le plus bas	Le plus haut
horsepower	Equal variances assumed	1.92	.17	2.39	203.00	.02	21.94	6.51	9.10	34.79
	Equal variances not assumed			3.37	29.91	.00	21.94	6.51	8.64	35.25

We use three components under **Tanagra**. But they are plugged after the same DEFINE STATUS, which specify the role of the variables, into the diagram. There is not repetitive handling.



Levene

Attribute_Y	Attribute_	Description				Statistical test	
		Value	Examples	Average	Std-dev	Test	
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

T-Test

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

T-Test Uneq. variance

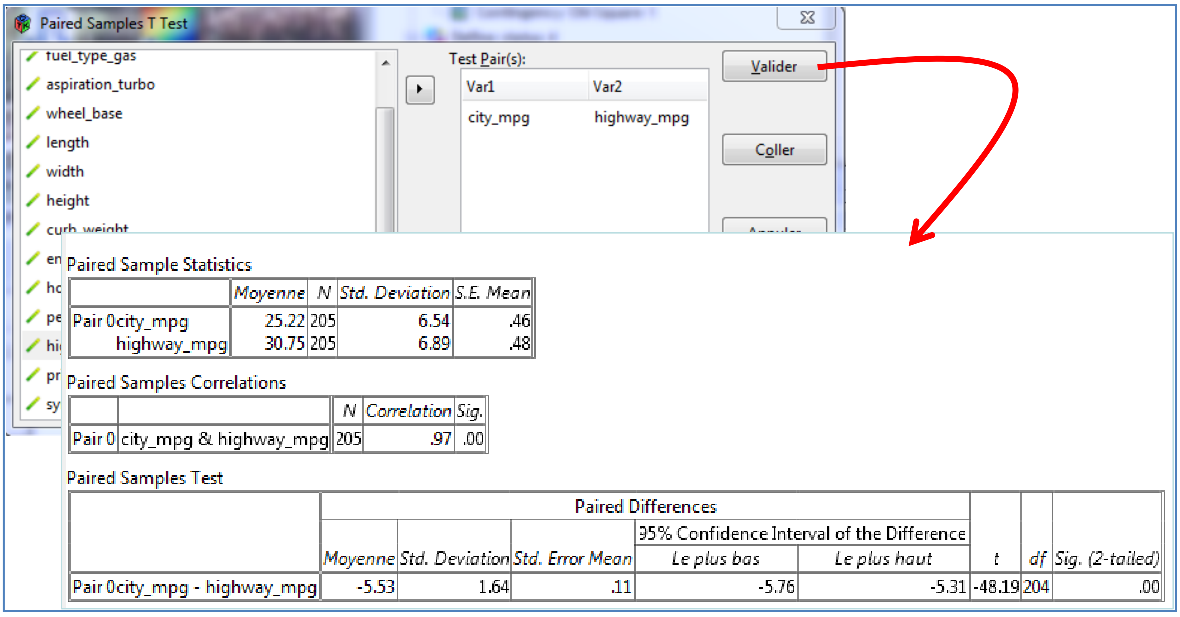
Attribute_Y	Attribute_X	Description				Statistical test	
horsepower	fuel_type	Value	Examples	Average	Std-dev	T	21.9446 / 6.5131 = 3.369315
		gas	185	106.3946	40.1834	d.f.	29.91
		diesel	20	84.45	25.9584	p-value	0.002085
		All	205	104.2537	39.5192		

Of course, the results are identical to those of PSPP.

5.5 Comparison of two means – Paired samples

Now, we want to compare the city mpg and the highway mpg. The difference has to be computed for each car i.e. we have paired samples. We click on the ANALYSE / COMPARAISON DES MOYENNES / PAIRED SAMPLES T TEST menu. We select the pair of variables “city mpg” and “highway mpg”.

PSPP provides the mean for each variable (25.2 city mpg, 30.75 highway mpg), the correlation between them, and the details for the statistical test. Patently, the consumption is higher in city (the miles that we can cover with one gallon of fuel is significantly lower).



Test Pair(s):

Var1	Var2
city_mpg	highway_mpg

Paired Sample Statistics

		Moyenne	N	Std. Deviation	S.E. Mean
Pair 0	city_mpg	25.22	205	6.54	.46
	highway_mpg	30.75	205	6.89	.48

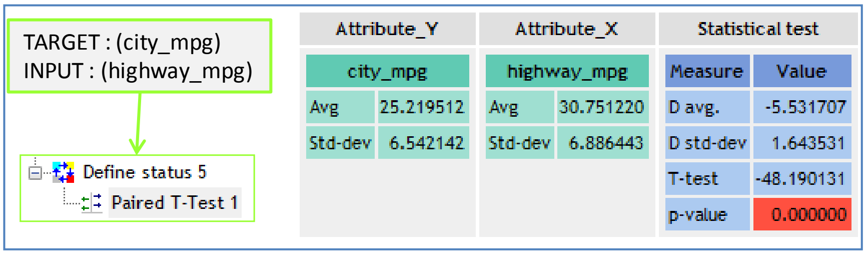
Paired Samples Correlations

		N	Correlation	Sig.
Pair 0	city_mpg & highway_mpg	205	.97	.00

Paired Samples Test

		Paired Differences							
					95% Confidence Interval of the Difference				
		Moyenne	Std. Deviation	Std. Error Mean	Le plus bas	Le plus haut	t	df	Sig. (2-tailed)
Pair 0	city_mpg - highway_mpg	-5.53	1.64	.11	-5.76	-5.31	-48.19	204	.00

Le PAIRED T-TEST component of **Tanagra** provides the same values.

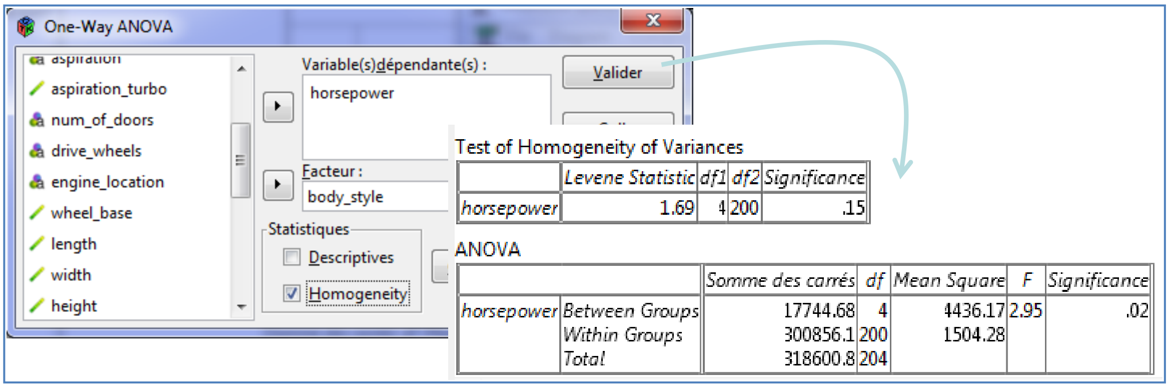


TARGET : (city_mpg)
INPUT : (highway_mpg)

Attribute_Y		Attribute_X		Statistical test	
city_mpg		highway_mpg		Measure	Value
Avg	25.219512	Avg	30.751220	D avg.	-5.531707
Std-dev	6.542142	Std-dev	6.886443	D std-dev	1.643531
				T-test	-48.190131
				p-value	0.000000

5.6 Comparison of K means – Analysis of variance (ANOVA)

We want now to compare the means of several (> 2) groups. We click on the ANALYSE / COMPARAISON DES MOYENNES / ONE WAY ANOVA menu.



Variable(s)dépendante(s) : horsepower

Facteur : body_style

Statistiques
□ Descriptives
☑ Homogeneity

Test of Homogeneity of Variances

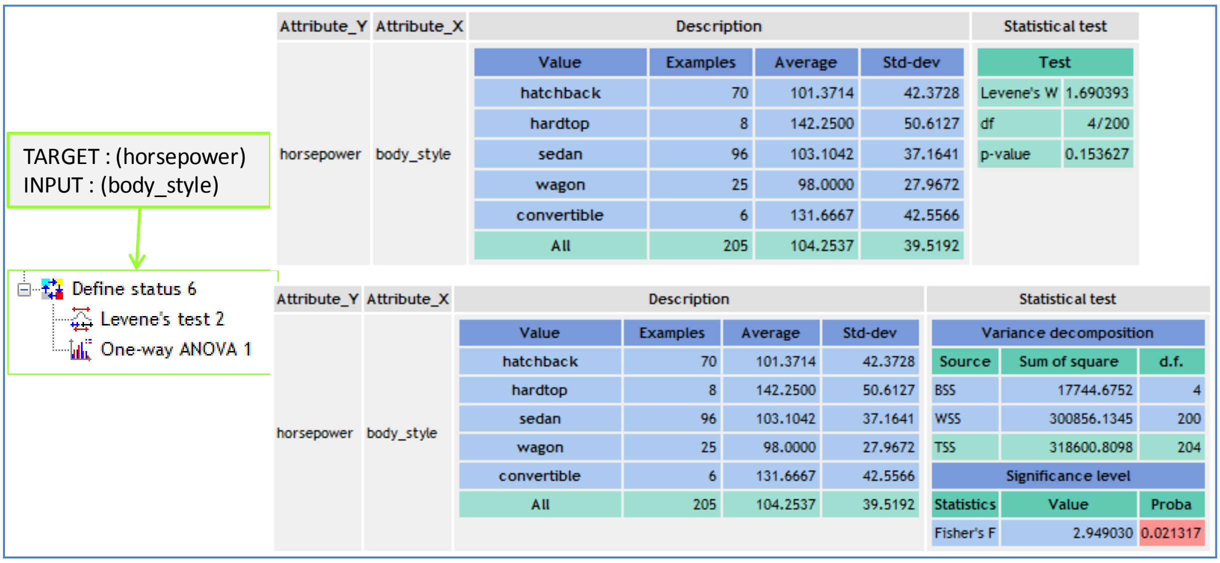
	Levene Statistic	df1	df2	Significance
horsepower	1.69	4	200	.15

ANOVA

		Somme des carrés	df	Mean Square	F	Significance
horsepower	Between Groups	17744.68	4	4436.17	2.95	.02
	Within Groups	300856.1	200	1504.28		
	Total	318600.8	204			

We set “horsepower” as DEPENDENT VARIABLE, and “body style” as FACTOR. PSPP assess the homoscedasticity assumption with the Levene's test. Then, it performs the comparison of means. At the 5% level, we reject the null hypothesis.

Tanagra provides the same results. But we use two components.



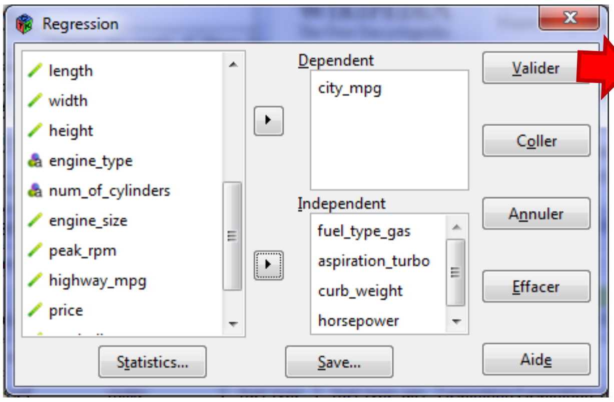
TARGET : (horsepower)
INPUT : (body_style)

Attribute_Y	Attribute_X	Description				Statistical test	
		Value	Examples	Average	Std-dev	Test	
horsepower	body_style	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		
		Value	Examples	Average	Std-dev	Variance decomposition		
horsepower	body_style	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

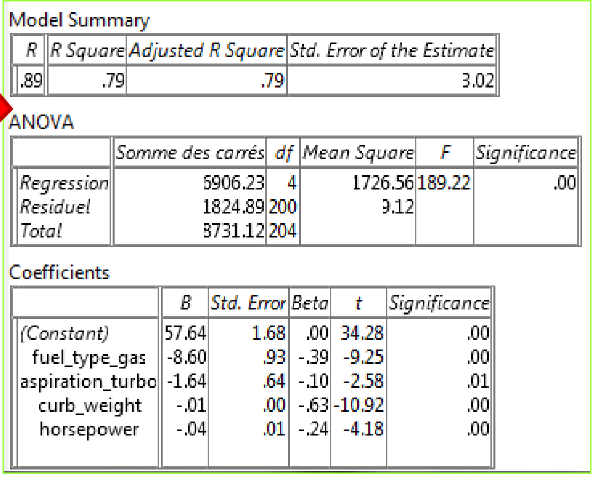
5.7 Multiple regression

We want to explain the consumption (city mpg) from the fuel type (dummy variable), the aspiration, the curb weight and the horsepower. We set these parameters into the dialog box of ANALYSE / LINEAR REGRESSION menu.



Dependent
city_mpg

Independent
fuel_type_gas
aspiration_turbo
curb_weight
horsepower



Model Summary

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

ANOVA

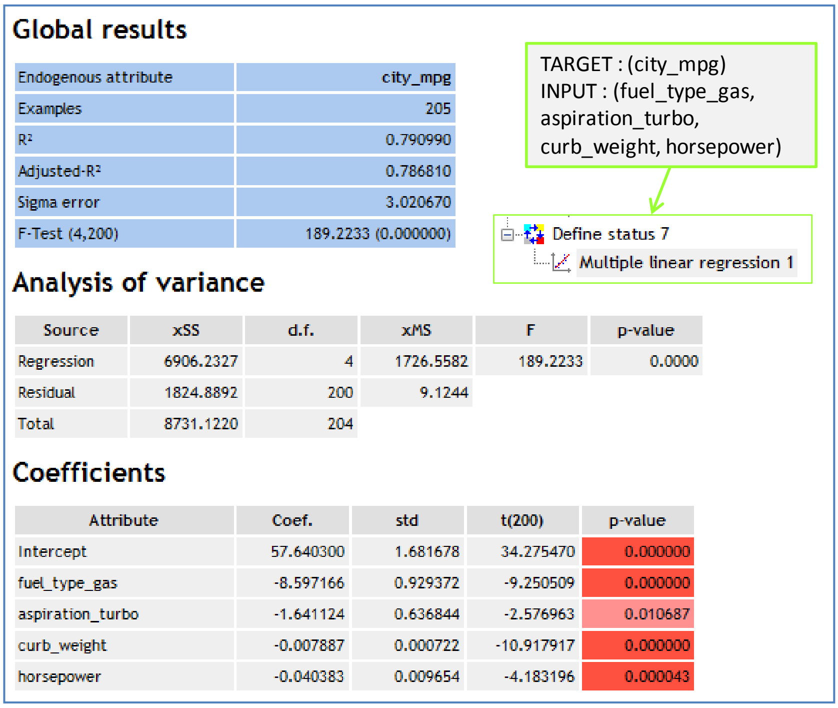
	Somme des carrés	df	Mean Square	F	Significance
Regression	6906.23	4	1726.56	189.22	.00
Residuel	1824.89	200	9.12		
Total	8731.12	204			

Coefficients

	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 provides the multiple correlation coefficient R; the coefficient of determination R-squared; the ANOVA table for the regression, the table of coefficients. The model is globally significant at the 5% level. And all the coefficients are significant. “Weight” and “fuel type” have the highest influence on the consumption.

The MULTIPLE LINEAR REGRESSION of **Tanagra** provides the same results.



Global results

Endogenous attribute	city_mpg
Examples	205
R²	0.790990
Adjusted-R²	0.786810
Sigma error	3.020670
F-Test (4,200)	189.2233 (0.000000)

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

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			

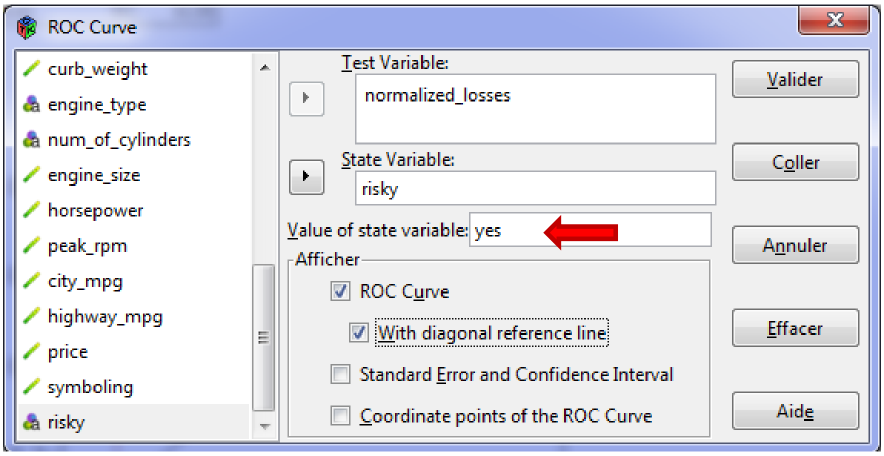
Coefficients

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

5.8 ROC curve

Cars are initially assigned a risk factor symbol associated with its price. Then, if it is more risky (or less), this symbol is adjusted by moving it up (or down) the scale. Insurance companies call this process "symboling"[4]. From this variable, we define a new column "risky". Its value is “yes” if “symboling > 0”, “no” otherwise. At the same time, the “normalized losses” is defined as the relative average loss payment per insured vehicle year. This value is normalized for all autos within a particular size classification (two-door small, station wagons, sports/specialty, etc...).

We were wondering if the “normalized losses” allows to distinguish risky cars from non-risky ones. We want to use the ROC curve for answering to this question. We click on the ANALYSE / ROC CURVE menu. We set the following settings.



ROC Curve

Test Variable: normalized_losses

State Variable: risky

Value of state variable: yes

Afficher

☑ ROC Curve

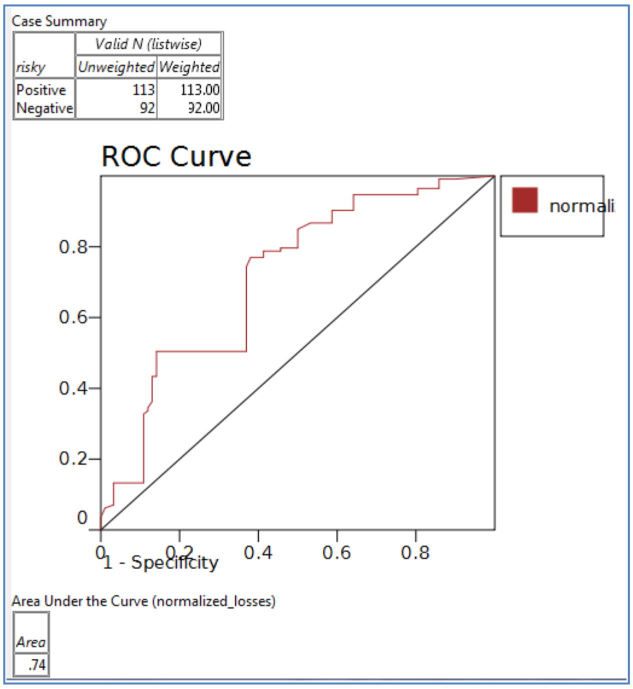
☑ With diagonal reference line

☐ Standard Error and Confidence Interval

☐ Coordinate points of the ROC Curve

The “positive” instances corresponds to “risky = yes” values. PSPP calculates automatically the AUC criterion (area under curve).

[4] http://archive.ics.uci.edu/ml/datasets/Automobile



Case Summary

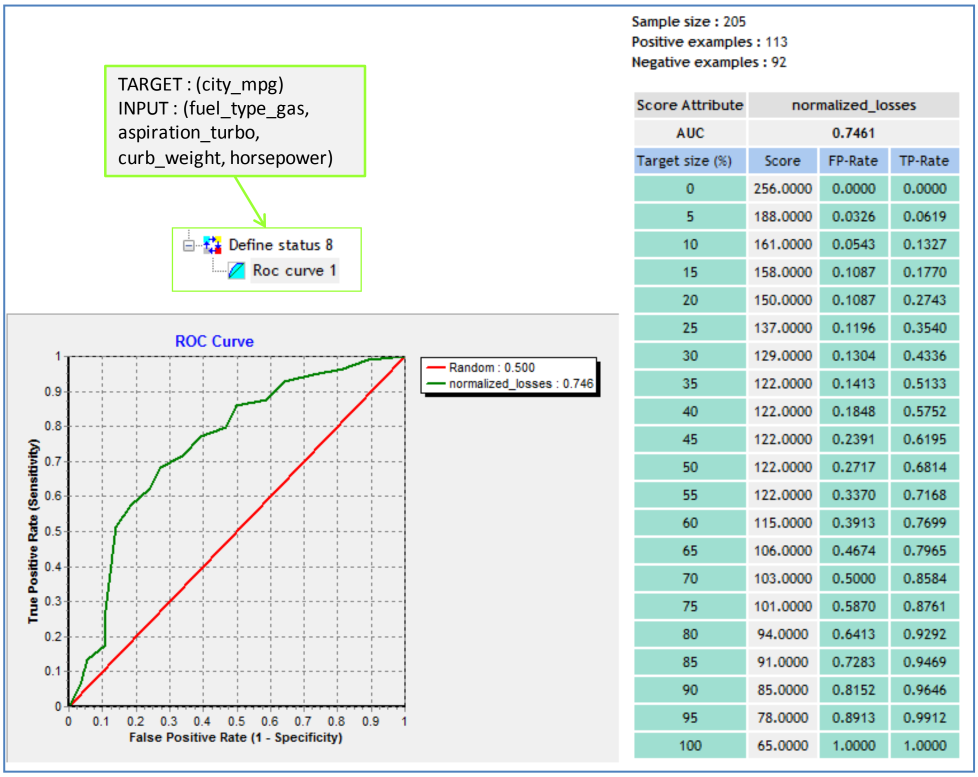
	Valid N (listwise)	
risky	Unweighted	Weighted
Positive	113	113.00
Negative	92	92.00

Area Under the Curve (normalized_losses)

Area
.74

There are 113 positive instances into the dataset. AUC = 74% i.e. the probability that a randomly chosen risky cars has higher normalized losses than a randomly chosen non-risky cars is 74%.

For this time, **Tanagra** seems not agree. The shape of the ROC curve is a little different. In effect, Tanagra cuts the score value in 20 intervals and build the ROC curve from the corresponding values. Thus, we obtain **a smooth curve**. But the underlying values used for the construction of the curve are the same. We obtain the same AUC criterion.



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

Sample size : 205
Positive examples : 113
Negative examples : 92

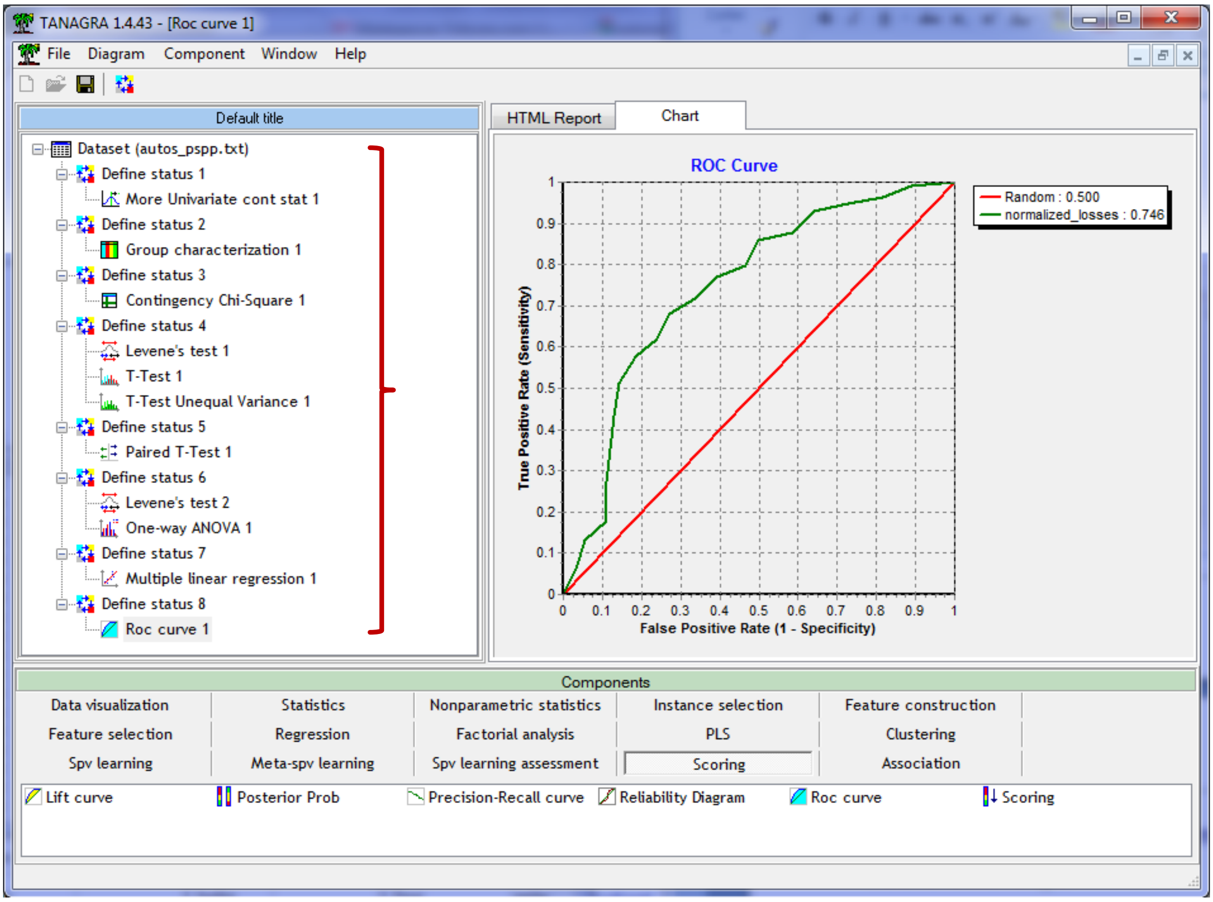
Score Attribute	normalized_losses		
AUC	0.7461		
Target size (%)	Score	FP-Rate	TP-Rate
0	256.0000	0.0000	0.0000
5	188.0000	0.0326	0.0619
10	161.0000	0.0543	0.1327
15	158.0000	0.1087	0.1770
20	150.0000	0.1087	0.2743
25	137.0000	0.1196	0.3540
30	129.0000	0.1304	0.4336
35	122.0000	0.1413	0.5133
40	122.0000	0.1848	0.5752
45	122.0000	0.2391	0.6195
50	122.0000	0.2717	0.6814
55	122.0000	0.3370	0.7168
60	115.0000	0.3913	0.7699
65	106.0000	0.4674	0.7965
70	103.0000	0.5000	0.8584
75	101.0000	0.5870	0.8761
80	94.0000	0.6413	0.9292
85	91.0000	0.7283	0.9469
90	85.0000	0.8152	0.9646
95	78.0000	0.8913	0.9912
100	65.0000	1.0000	1.0000

6 Perform the same analyses with other tools

Of course, we can implement the same statistical approaches with other tools. In this section, we show briefly the whole diagram under **Tanagra**, we details the commands under **R 2.13.2**, and we show the available features of **OpenStat** (which is really similar to PSPP).

6.1 Tanagra

Some statistical approaches available in PSPP are not available in Tanagra, and conversely. For the methods described in this tutorial, here is the whole diagram under Tanagra.



6.2 R software

Excluding the ROC curve, we detail here the commands and the corresponding outputs in R. In some circumstances, we need a specific package that we load with the **library(.)** instruction.

```
> #loading the dataset
> setwd("D:/DataMining/Databases_for_mining/logiciels_dataset/pspp")
> autos <- read.table(file="autos_pspp.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
 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 independance
> 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$aspiration 
autos$fuel_type |       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 ***
```

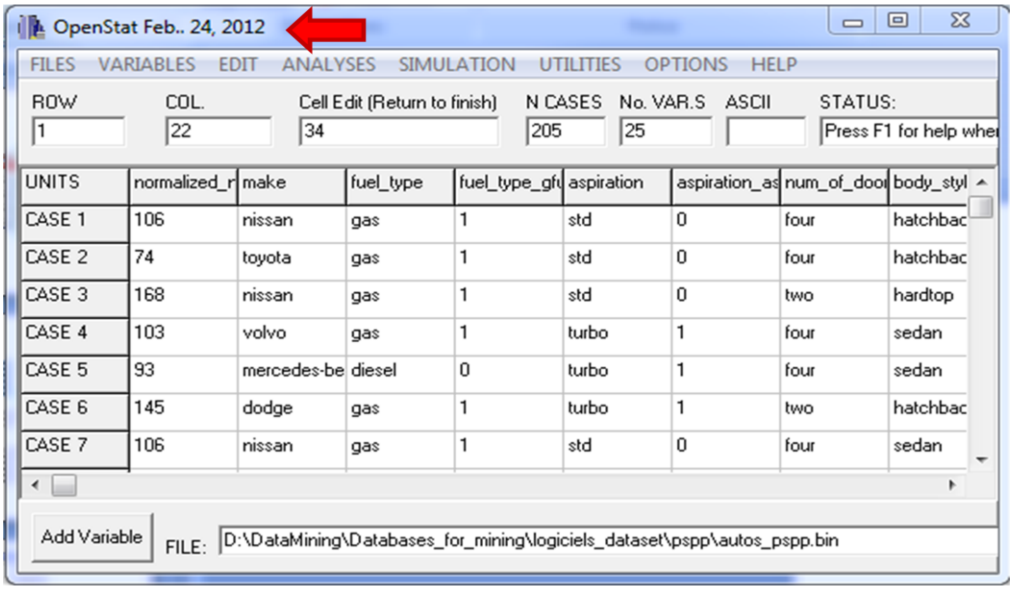
```
horsepower        -0.0403830  0.0096536  -4.183  4.3e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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
```

Knowing the appropriate commands for each method is the main difficulty under R. Fortunately, some websites provide a valuable assistance (e.g. Quick-R).

6.3 OpenStat

OpenStat is also a credible alternative to SPSS. In a previous tutorial, we have studied one of its variations (LazStat) in the regression analysis framework[5].

Like PSPP, we must import the data file first (FILES / IMPORT TAB FILE menu). The dataset is displayed into the grid.



ROW: 1 | COL.: 22 | Cell Edit (Return to finish): 34 | N CASES: 205 | No. VAR.S: 25 | ASCII: | STATUS: Press F1 for help whe

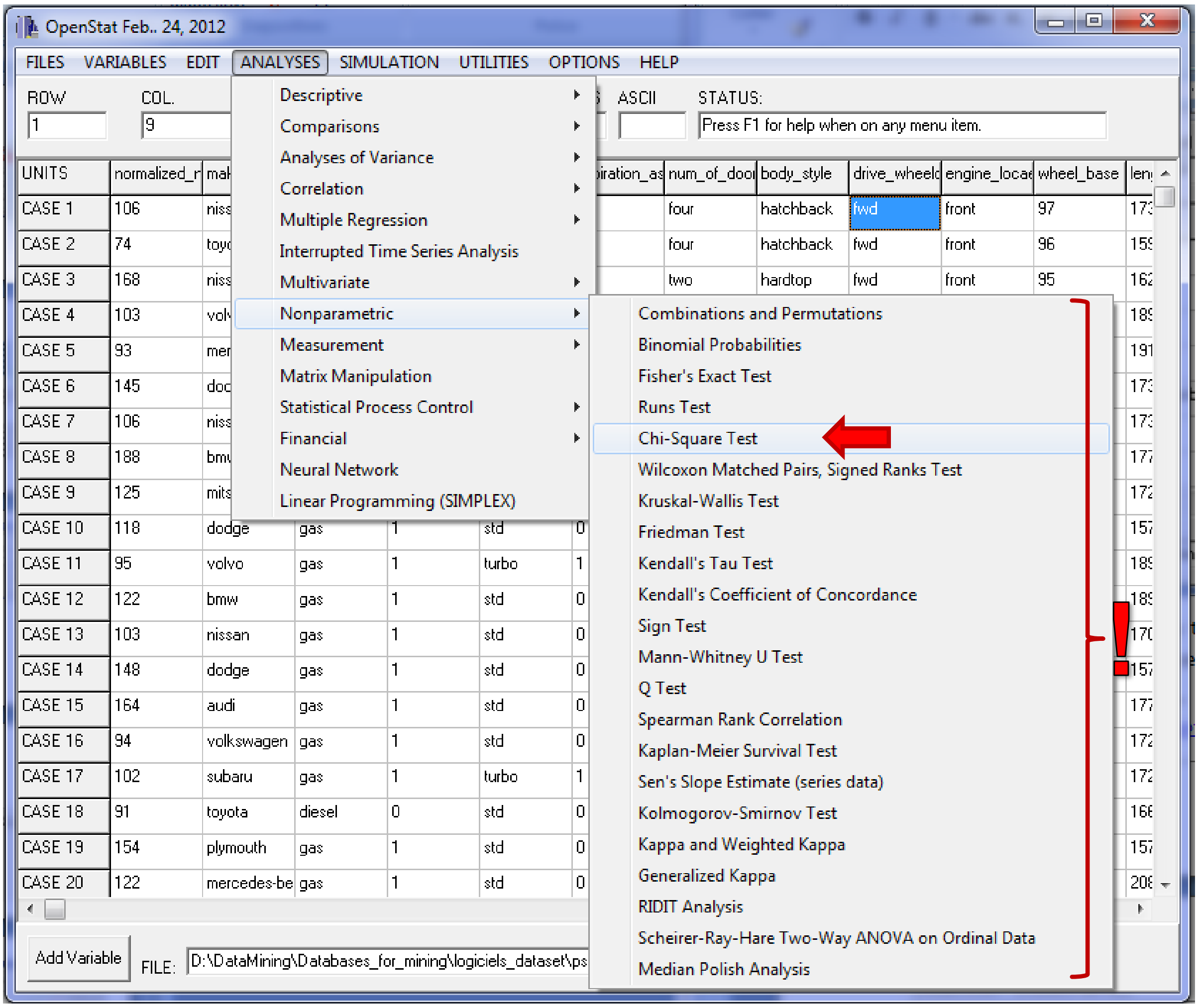
UNITS	normalized_r	make	fuel_type	fuel_type_gfu	aspiration	aspiration_as	num_of_door	body_styl
CASE 1	106	nissan	gas	1	std	0	four	hatchbac
CASE 2	74	toyota	gas	1	std	0	four	hatchbac
CASE 3	168	nissan	gas	1	std	0	two	hardtop
CASE 4	103	volvo	gas	1	turbo	1	four	sedan
CASE 5	93	mercedes-be	diesel	0	turbo	1	four	sedan
CASE 6	145	dodge	gas	1	turbo	1	two	hatchbac
CASE 7	106	nissan	gas	1	std	0	four	sedan

FILE: D:\DataMining\Databases_for_mining\logiciels_dataset\pspp\autos_pspp.bin

We do not detail the results for each of the methods studied above. I perform these analyses elsewhere, I note that the results are the same as PSPP (as Tanagra and R). We describe only the results for the contingency table. We note that we must code the categorical variables as dummy ones before performing the analysis.

The statistical methods are available into the ANALYSES menu. We select the NONPARAMETRIC item. The list of available techniques is really long. We will try to describe them in a forthcoming tutorial.

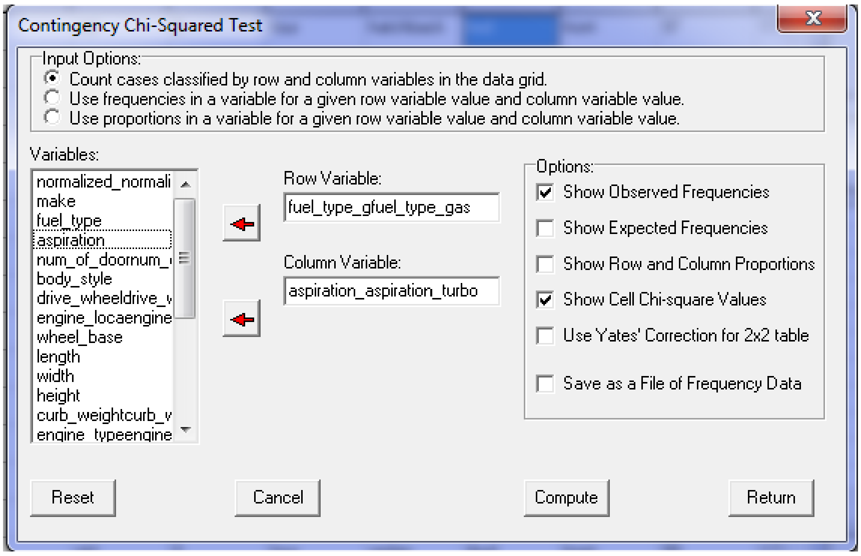
[5] http://data-mining-tutorials.blogspot.com/2012/03/regression-analysis-with-lazstats.html


OpenStat Feb.. 24, 2012
FILES VARIABLES EDIT ANALYSES SIMULATION UTILITIES OPTIONS HELP
ROW 1
COL. 9
ASCII
STATUS: Press F1 for help when on any menu item.
Descriptive
Comparisons
Analyses of Variance
Correlation
Multiple Regression
Interrupted Time Series Analysis
Multivariate
Nonparametric
Measurement
Matrix Manipulation
Statistical Process Control
Financial
Neural Network
Linear Programming (SIMPLEX)
Combinations and Permutations
Binomial Probabilities
Fisher's Exact Test
Runs Test
Chi-Square Test
Wilcoxon Matched Pairs, Signed Ranks Test
Kruskal-Wallis Test
Friedman Test
Kendall's Tau Test
Kendall's Coefficient of Concordance
Sign Test
Mann-Whitney U Test
Q Test
Spearman Rank Correlation
Kaplan-Meier Survival Test
Sen's Slope Estimate (series data)
Kolmogorov-Smirnov Test
Kappa and Weighted Kappa
Generalized Kappa
RIDIT Analysis
Scheirer-Ray-Hare Two-Way ANOVA on Ordinal Data
Median Polish Analysis
Add Variable
FILE: D:\DataMining\Databases_for_mining\logiciels_dataset\ps

UNITS	normalized_r	mak				
CASE 1	106	niss				
CASE 2	74	toyo				
CASE 3	168	niss				
CASE 4	103	volv				
CASE 5	93	mer				
CASE 6	145	doc				
CASE 7	106	niss				
CASE 8	188	bmw				
CASE 9	125	mits				
CASE 10	118	dodge	gas	1	std	0
CASE 11	95	volvo	gas	1	turbo	1
CASE 12	122	bmw	gas	1	std	0
CASE 13	103	nissan	gas	1	std	0
CASE 14	148	dodge	gas	1	std	0
CASE 15	164	audi	gas	1	std	0
CASE 16	94	volkswagen	gas	1	std	0
CASE 17	102	subaru	gas	1	turbo	1
CASE 18	91	toyota	diesel	0	std	0
CASE 19	154	plymouth	gas	1	std	0
CASE 20	122	mercedes-be	gas	1	std	0

UNITS	num_of_door	body_style	drive_wheel	engine_loca	wheel_base
CASE 1	four	hatchback	fwd	front	97
CASE 2	four	hatchback	fwd	front	96
CASE 3	two	hardtop	fwd	front	95

For the CHI-SQUARE TEST, we set the following settings and we click on COMPUTE.



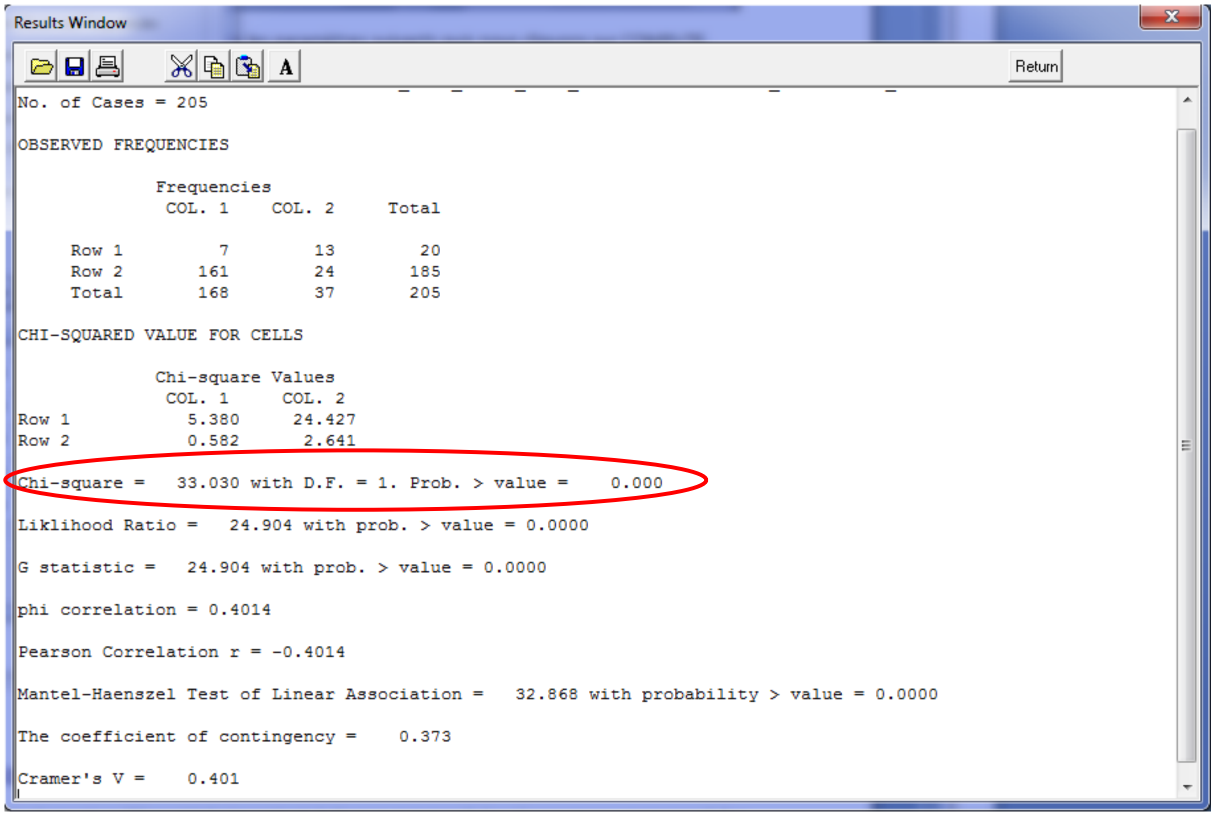
Input Options:
- ● Count cases classified by row and column variables in the data grid.
- ○ Use frequencies in a variable for a given row variable value and column variable value.
- ○ Use proportions in a variable for a given row variable value and column variable value.

Row Variable: fuel_type_gfuel_type_gas

Column Variable: aspiration_aspiration_turbo

Options:
- ☑ Show Observed Frequencies
- □ Show Expected Frequencies
- □ Show Row and Column Proportions
- ☑ Show Cell Chi-square Values
- □ Use Yates' Correction for 2x2 table
- □ Save as a File of Frequency Data

A results window appears. **OpenStat** provides additional statistics such as the Mantel-Haenszel test of linear association, the coefficient of contingency, etc.



```
No. of Cases = 205

OBSERVED FREQUENCIES

             Frequencies
              COL. 1    COL. 2     Total

     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

Liklihood 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
```

Unlike to PSPP, because **OpenStat** is only a menu-guided program, it is not possible to store a description of the treatments. So, it is not easy to perform the same sequence of analyses if we have a new version of the dataset for instance.

7 Conclusion

PSPP is a promising project. The structure of the software is really well thought out. Each menu action is translated into a PSPP instruction. Thus, we can save the sequence of commands into a script file. For instance, for the analysis of variance described above (section 5.6), PSPP generates the following instruction.

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

PSPP already proposes a large part of the statistical methods. It will be further complemented in the future. This is a tool that I will follow with interest.