PSPP A FREE AND OPEN SOURCE TOOL FOR DATA ANALYSIS

Jignasu Yagnik
Associate Sr. Faculty, EDI of India

Voice of Research
Vol. 2, Issue 4,
March 2014
ISSN 2277-7733

Abstract

Besides carrying out research for academic purpose, it is becoming increasingly important for educational institutions to provide hands-on training to the students in use of statistical software for making informed decisions. Many proprietary solutions are available for the data analysis but initial licensing and subsequent upgrading prices of these solutions are beyond reach of majority of academic institutions, and small organisations with limited resources. PSPP has evolved as a valuable resource for educational institutions, MSMEs, non-government organisations and others requiring a free and easy to learn software for data analysis. Important features of PSPP are described in this paper.

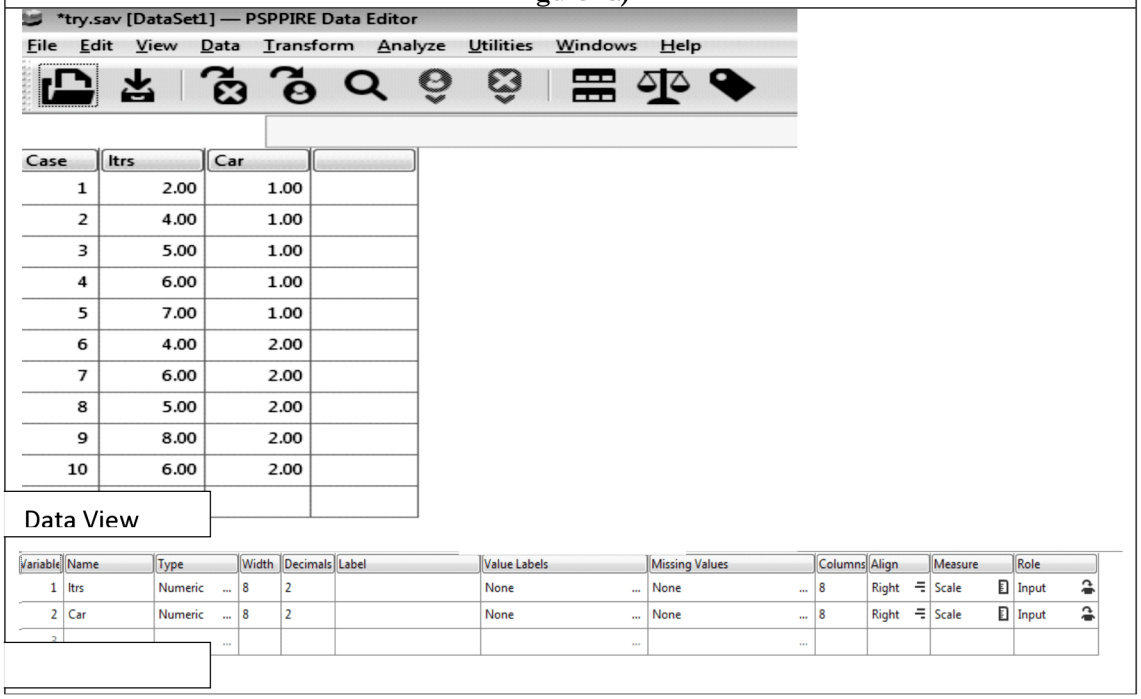
Keywords *: PSPP, Data Analysis, Free Software for Statistical Analysis*

There has been a remarkable adoption of Information and Communication Technology (ICT) during the last two decades, but concerns regarding appropriateness of technology adopted and optimal usage of the same persist (Brzyckiand Dudt, 2005). One ICT component requiring thorough evaluation prior to selection, is software.Rise in adoption of computers due to drastic decline in prices of hardware has created a huge market for software applications to accurately and expeditiously carry out voluminous and complex tasks.In order to facilitate decision making, use of software for statistical analysis has gained prominence among corporates, small businesses, consultancy firms, teaching and training institutions etc. Besides carrying out research for academic purpose, it is becoming increasingly important for educational institutions to provide hands-on training to the students in use of statistical software for making informed decisions.There are many proprietary solutions available in the market to cater to this demand prevailing across different verticals. Players like SAS, SPSS, Minitab, and Stata etc. offer excellent software suites for data analysis and business intelligence requirements but these suites are beyond reach of majority of organisations and individuals. Also, as these suites are proprietary and not based open licensing they lead to vendor lock-in in the long run. Piracyremains the only choice for those who cannot purchase these costly products.In order to overcome limitations posed by high initial cost and subsequent costs for support and upgrade; based on philosophy of Free Software Foundation, open source initiatives likeR and PSPP were promoted in early nineties so that the statistical software could be made available to users without any cost and with no restriction on copying and distribution of the same.R has carved a niche for itself as a software environment for statistical computing and has become centre of attraction for users who are capable of handling complexities of coding and are comfortable with command based interface. PSPP is software that has evolved substantially in the open source domain under General Public License published by free software foundation. It has an interface which resembles that of SPSS and has potential to meet requirements of a large majority users in search of a free and user-friendly option for statistical analysis that does not involve complex modelling. Being a clone of SPSS which is widely recognised as a user-friendly and 'easy to learn' software, PSPP appears to be emerging as a viable Free and Open Source alternative of SPSS for many organisations and individuals. It is worth reiterating that the software is free and also can be copied and distributed along with its source code without violating copyrights laws.Like SPSS, PSPP also provides a menu-based interface and also permits use of syntax for doing analysis. Moreover, it can seamlessly open files created in SPSS on a click of mouse. Features like computing new variables, recoding variables have been incorporated to facilitate the user. Besides basic tools like Frequencies, Test of Normality, Crosstabs & chi-square analysis, T-tests, One-way ANOVA, Correlation and Regression, Reliability estimation and non-parametric tests; advanced statistical tools like Cluster analysis Factor analysis and Logistic Regressions are also available in this version. Some statistical tests like Mann-Whitney, Friedman's etc. are not included in the menu but they can be performed using a simple syntax explained in the user manual. The new version with more features is due for release. PSPP can be freely downloaded from http://www.gnu.org/software/pspp/.Hereafter, we discuss the features of PSPP (Version 0.8.1) as valuable resource for MSMEs, educational Institutions, non-government organisations and others requiring a free and easy to learn software for data analysis. After discussing the main views that appear at the outset, eachmenu of the software has been focussed upon to give an overview of its features and functionality.

PSPP Gateway

Figure-1a depicts the primary interface in which a data editor is opened in data view of PSPP. Shown in Figure 1b) is the variable view for defining variables, entering variable labels, value labels, defining missing values, measurement scale and role of variables. One can switch between the two views by selecting tabs given in the left bottom corner.

Figure 1a)



Users of SPSS will agree that theprimary interface and views are to a large extent resembling SPSS. Features covered under different menus are briefly discussed below

File Menu: Figure-2 depicts a snapshot of the File menu of PSPP. As in SPSS, under the "File->New" option, it is possible to open a new syntax file and a new data file as well. Also shown encircled in Figure-3 is the option "**import data"** to fetch data entered in text, spreadsheets and other formats to facilitate the users. It is worth noting that PSPP opens SPSS files directly on a click of mouse.

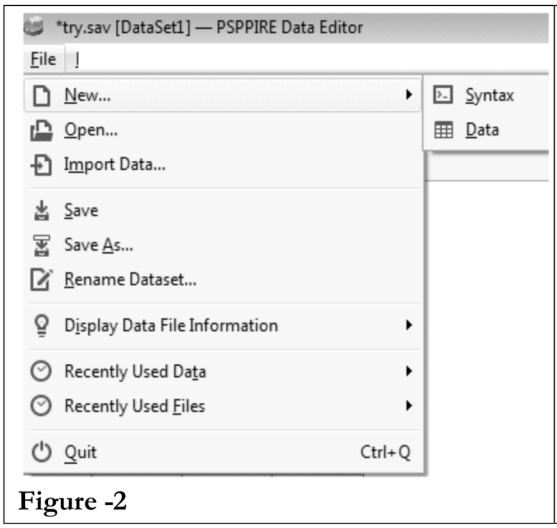


Figure -2

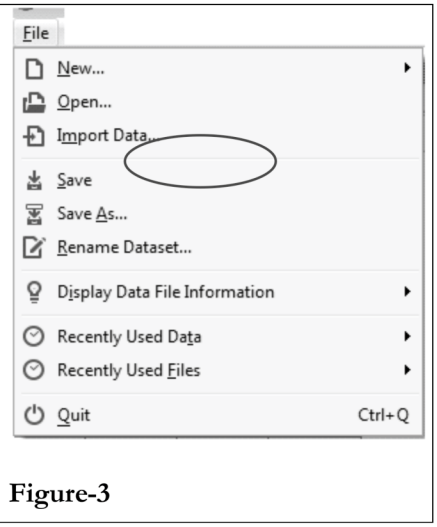


Figure-3

Edit Menu: This menu provides option to edit and manage variables and cases in PSPP. Besides option to copy or move content, one can insert as well as delete/clear variables and cases in the data set. A snapshot of Edit menu is given in Figure-4.

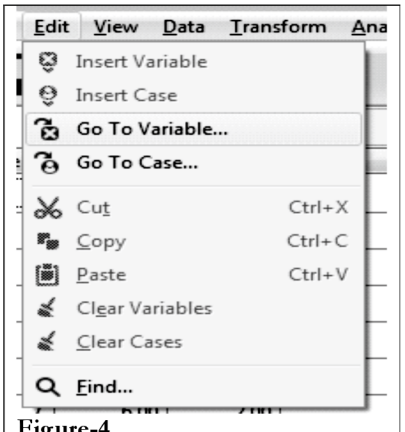


Figure-4

View Menu: The options under the View menu are depicted in Figure-5. Besides options to set font size and type,options to switch between values and value labels is incorporated here. One can also switch between data and variable views by selecting appropriate option listed at the bottom of this menu.

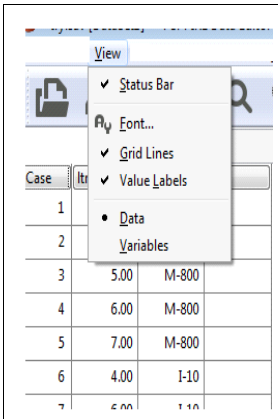


Figure-5

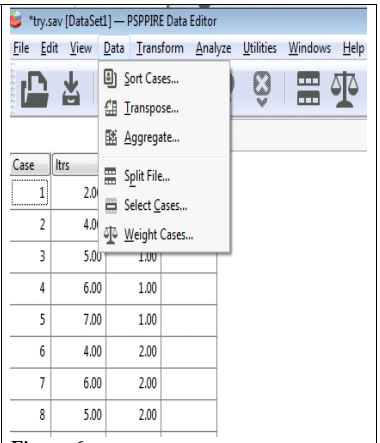


Figure-6

Data Menu: Options under data menu are shown in Figure-6. It includes features for organising the data in ascending or descending order, converting rows to columns and columns to rows, creating new files by aggregating data on selected variables. Furthermore, it provides options for splitting the data into subgroups so that separate analysis can be conducted for each subgroup. Option for selecting cases based on some condition is also part of this menu. The last option allows you to select a weight variable that can be used as multiplier during analysis.

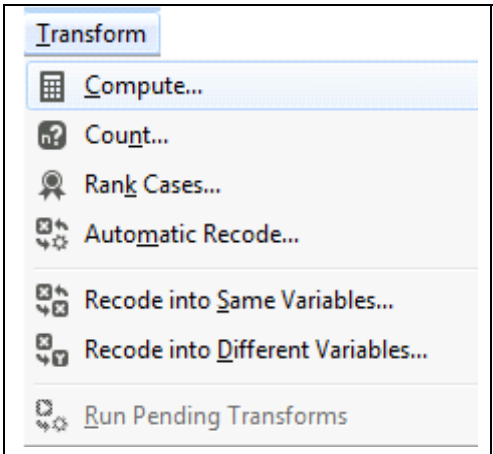


Figure-7

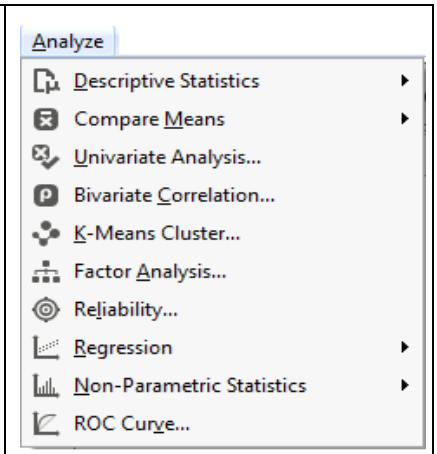


Figure-8

Transform Menu: Options in Transform menu are depicted in Figure-7. This menu deals with features related to creation variables and manipulation of data. The compute option enables creation of new variables using arithmetic operators, other operators and functions available in PSPP. Another important option is the count feature which counts number of cases based on specified criteria of variables. Rank cases option allows ranking of data on a chosen variable and also includes features to create percentiles. The Automatic recode feature enables conversion of Alphabetic data into numeric codes in a typical order (ex. in case of Gender Automatic recode can be used to create a new variable where Male will be coded as 1 and Female will be coded as 2). The recode function has utility in clubbing categories of variables (ex. converting age in years to groups 25-30, 31-40 and above 40). Recoding can be done into the same variable or a new variable can be created to keep the original values intact. It is advisable to use recode into different variables.

AnalyzeMenu: Figure-9 demonstrates the options in Analyze menu with sub-options under the "descriptive statistics" option. A list of other analysis options is also seen. Depending on the requirement the user has to select appropriate analysis option and provide details related to variables and other selections to generate required statistics. Each option of analysis menu is discussed in brief in the following paragraphs.

a) Descriptives Option: Under the descriptive statistics option PSPP provides a means to generate

univariatestatistics like counts and percentagesetc. for categorical variables using the option Frequencies.

summary statistics like mean, median mode standard deviation etc. using descriptivesprocedure

some important statistics used to examine the data like 95% confidence interval of mean, trimmed mean etc. besides other descriptive measures through the procedure Explore

bivariate or multivariate frequency tables along with statistical tests of association and agreement.

Seen in Figure-10 has been constructed by putting sub-option items beside group processes except for the descriptives group already discussed earlier. Other analysis procedures/optionsin theAnalyze menu are discussed below with reference to Figure-10.

b) Compare Means: This option under analysis menu helps the user to generateaverage and other summary measures for different categories of an independent variable . For example getting mean salary separately for males and females. conductT-test for single sample mean, T-test for comparing means of two independent samples and t-test to compare means of paired data. Conduct univariate ANOVA to compare means of dependent variable between more than two groups/samplesalong with post-hoc tests

c) Univariate Analysis:Univariate Analysis procedure meant to perform two-way analysis of variance is under testing for final implementation and is likely to be through in the next release of PSPP.

d) Bivariate Correlations: Correlation matrix between pairs of variables can be generated using this procedure in analyse menu. It also provides statistical significance of correlation coefficients.

e) K-means Cluster:Cluster analysis is widely used to identify

similar cases in data set for identifying segments in data. In PSPP one can specify number of clusters required from a group of variables included in the analysis. The hierarchical clustering procedure used to determine number of clusters to be generated has not been implemented yet.

f) Factor Analysis:Factor analysis is widely used for variable reduction through creation of fewer factors from a set of correlated variables. Factors represent underlying composite constructs generated from the variables. This procure performs factor analysis with principal component analysis as default method. Options for rotation of axis like varimax, quatrimax etc. can be to gain clarity about factors that are unclear in un-rotated solution.

g) Reliability: This procedure provides means for testing consistency of a measurement scales and its items. PSPP provides reliability estimates based on Cronbach's alpha and split-half models.

h) Regression:Regression group under analysis offers options for conducting linear and logistic regressions. It is possible to perform bivariate and multivariate linear regression. In case the dependent variable is binary logistic regression is the procedure of choice for predicting odds associated with categories of predictive variables.

i) Non-Parametric Procedures Group:This command group includes a set of non-parametric procedures (Figure -10) appropriate for conducting statistical testing when the number of observations is small or when the normality assumption for parametric tests is not met. Although the GUI enlists six sub-procedures other procedures for independent samples like Mann Whitney test, Median test etc. can be performed using a very simple coding as demonstrated in Figure-11 and Figure-12. Nonparametric procedures included in the GUI are

Chi-square - For testing similarity of proportions across different values of a categorical variable.

Binomial - To compare observed of a binary variable with binomial distribution

Runs – to test randomness of a Variable

1-Sample K-S – To compare observed distribution of a variable with theoretical distribution (useful in testing whether a variable adheres to normal distribution). K-S stands for Kolmogorov- Smirnov

2 Related Samples – Wilcoxon matched pair test, a nonparametric alternative of Paired t-test (Computation are based on Ranks)

K related Samples- Friedman test a nonparametric alternative of repeated measures ANOVA comparing more than two related samples (Computation are based on Ranks)

Non-parametric Procedures not included in GUI but can be performed using syntax are

Mann-Whitney U Test – Nonparametric alternative of T-test of Independent samples based (Computation are based on Ranks), Figure- 11 shows the syntax and Figure-12 is the output generated on executing the syntax.

Kruskal-Wallis Test – Nonparametric alternative of ANOVA independent Samples.

Other procedures like Mcnemar test, Sign test, Median test are also available.

j) ROC Curve:This command is used to get the receiver operating characteristic curve and to estimate the area under the curve which has application in classification analysis of a contingency table with two dichotomous variables. It also provides sensitivity and specificityestimates

Conclusion

PSPP is a good alternative of SPSS. MSMEs, Consultancy firms, Educational Institutions can adopt this data analysis software that can meet the need of a majority of users. A lot of unwanted investment in software can be saved and piracy can be significantly reduced by adopting PSPP. The tools available in PSPP are sufficient for an intermediate level course on data analysis.

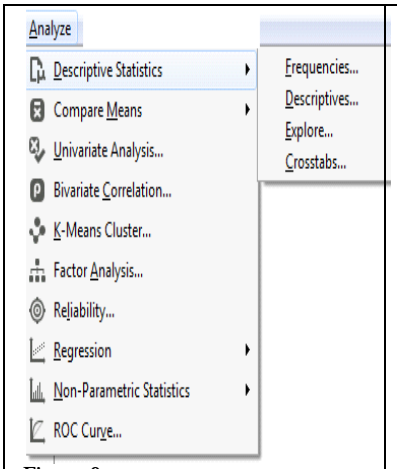


Figure-9

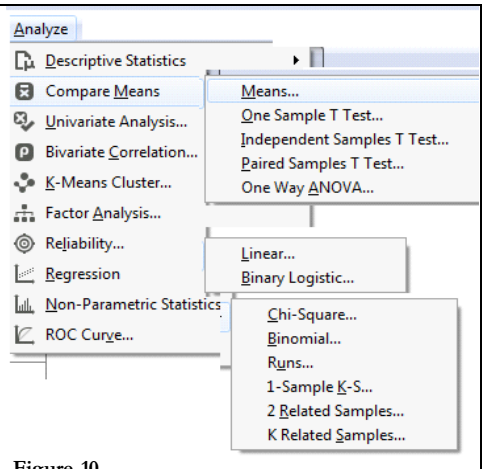


Figure-10

Syntax For Mann-Whitney Independent Samples Test

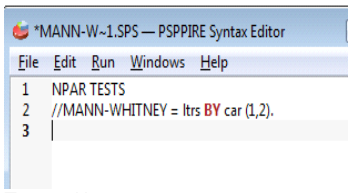


Figure-11

Output of Mann-Whitney test Generated with **Run all** on syntax

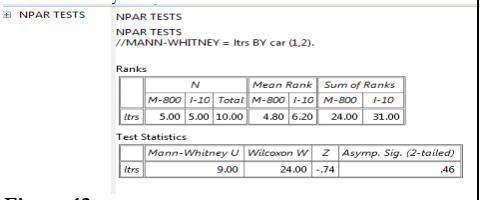


Figure-12

References

Brzycki,D., Dudt, D, 2005, Overcoming barriers to technology use in teacher preparation programs, *Journal of Technology and teacher education*, 13(4), pp 619-641

http://it.chass.ncsu.edu/training/pspp/

http://www.gnu.org/

http://www.gnu.org/software/pspp/

http://www.r-project.org/