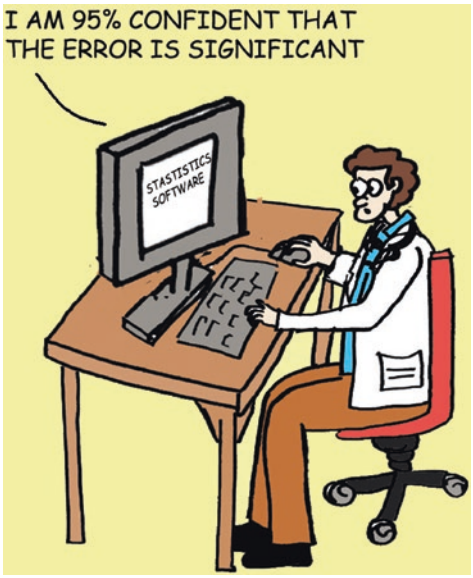
Statistical Packages for Data Analysis

N. Sreekumaran Nair, K. T. Harichandrakumar,
and N. Ravishankar

A statistical analysis, properly conducted, is a delicate dissection of uncertainties, a surgery of suppositions.
—MJ Moroney



N. S. Nair (✉)
Professor and Head, Department of Biostatistics, Jawaharlal Institute of Postgraduate Medical Education and Research, Pondicherry, India

K. T. Harichandrakumar
Assistant Professor, Department of Biostatistics, Jawaharlal Institute of Postgraduate Medical Education and Research, Pondicherry, India

N. Ravishankar
Assistant Professor, Department of Statistics, Manipal University, Manipal, India

© Springer Nature Singapore Pte Ltd. 2018

S. C. Parija, V. Kate (eds.), *Thesis Writing for Master's and Ph.D. Program*,
https://doi.org/10.1007/978-981-13-0890-1_26

Key Points

- This chapter provides a brief overview of the commonly used popular statistical packages for thesis or research project data analysis and report preparation.
- A comprehensive discussion on the strength, weakness and utility of packages namely Microsoft Excel, SPSS, PS, nMaster, Stata, Epi-info and EZR has been provided.
- The steps for installing data analysis add-in in MS Excel, using Master and gadgets of Epi-info have been described.

Background

This session deals with commonly available software packages for statistical analysis of health science research data. The main objective of this session is to sensitize the readers about various statistical packages which are regularly used for data analysis. This session introduces common packages, discuss various options available in each package and give an orientation to the difficulty level. However, this is not a manual to train you how to analyse data using these packages. For that purpose, you are advised to refer the manual of corresponding packages. Main packages included in the discussion are Microsoft Excel, SPSS, PS, nMaster, Stata, Epiinfo and EZR.

Microsoft Excel–MS-Excel [1]

MS Excel is the most common tool used for data entry and management. It is present in all the computers which possess Microsoft office. It features calculation, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications. It is used to store and retrieve numerical data in a grid format of columns and rows.

Excel can also be used for data analysis. The data analysis options are available as an add-in package in Excel and it has to be installed. The add-in package of data analysis in excel can be installed using the following steps. Figure 1 shows the steps for installing data analysis add-in in MS Excel.

The installed data analysis add-in can be accessed by clicking on '***Data***' in the toolbar.

The data analysis tools available in excel include descriptive statistics, t-test, Z-test ANOVA, Correlation, Regression, Covariance, moving average and random number generation. For a small set of data and simple analysis, this is a good package. However, the package does not have flexibility for more advanced options.

Excel is ideal to prepare different graphs and diagrams to summarize the data. It produces Line graph, Pie, Bar, Area, Scatter, Stack, Surface, Radar and Combo charts. The graphs and diagrams in excel are of high quality. Another advantage of the excel graph is that it can be edited as per the requirement. This is a good package for a beginner who does not have much knowledge about other statistical packages

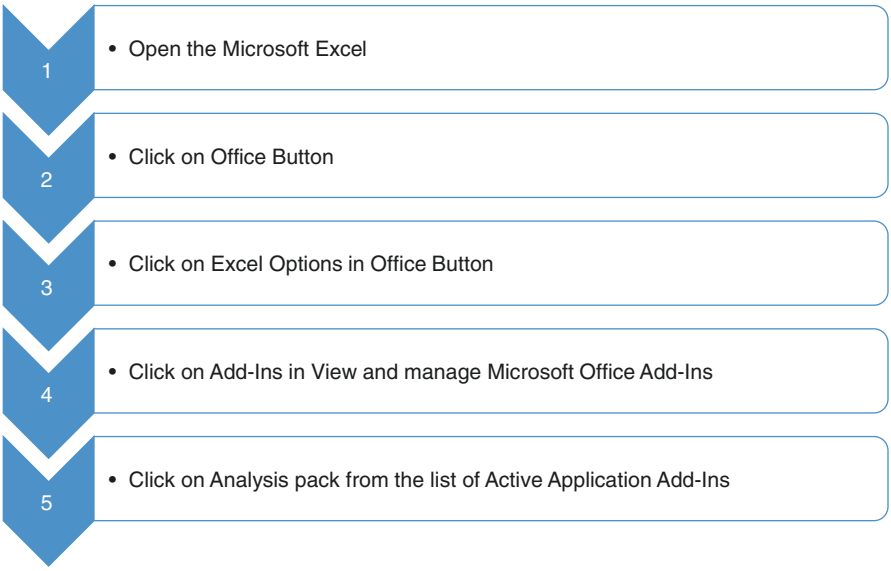


Fig. 1 Steps for installing data analysis add-in in MS excel

and required to do simple descriptive statistics, graphs and tests of significance. Data base prepared in Excel is compatible with many of the advanced statistical packages.

Statistical Packages for Social Sciences (SPSS) [2]

Statistical Packages for Social Sciences (SPSS) is one of the most commonly used statistical packages for data analysis by the health science researchers and Residents. The simple menu-driven characteristics, availability of most of the statistical methods and compatibility of the worksheet with other packages like excel make SPSS popular among medical researchers.

The SPSS basically consists of three windows namely Data editor/window, Output window and Syntax editor. The Data Editor is similar to Microsoft Excel and it consists of two sub-windows namely data view and variable view. Data view provides the complete view of the data set and variable view provides the characteristics of the variables. The variable characteristics such as type of variables, required width and decimal points, label of the variable name etc can be specified in the variable view. The coding for the categorical variables can also be defined through the values option in the variable view window. The characteristics of the variable should be clearly defined in variable view before entering the data. The Database for analysing the data in SPSS can be either created in SPSS or can be created in the worksheets of packages namely Excel, Systat, Stata, SAS etc. The database created in other formats can also be imported into SPSS for the analysis.

The spreadsheet in SPSS allows the user to split files, select cases based on specific characteristics of a particular variable, combine file, redefine variables, filter variables based on specific characteristic etc.

The different statistical analysis procedures available in SPSS are given under the drop-down menu '***Analysis***' in the data view. Descriptive Statistics for the data can be obtained through the drop-down '***Descriptive Statistics***' and '***Reports***' options respectively. The parametric tests such as one sample t-test, Independent Students t-test, Paired t-test and one-way Analysis of Variance (ANOVA) is available in '***Compare Means***' in the drop-down menu "*Analysis"*.

Non-parametric tests are available under the menu '***Nonparametric Test***'. Correlation and regression analysis is given under '***Correlation***' and '***regression***' options respectively in '***Analysis***'. The analysis of time to even data such as Life Tables, Kaplan Meier estimates, Log Rank Test, Cox regression etc. are provided in '***Survival Analysis***'**.** The advanced statistical analysis is available and is given in respective options in the *Analysis* menu.

The output of the analysis is displayed in output window. The results will be produced in table format with extensive details. The SPSS output file will be saved in '***save***' format and it can be exported into other formats such as word document, excel etc. as per the requirement through '***Export***' option in '*File*' in the '*output*' window.

Overall SPSS is a user-friendly comprehensive programme used for statistical analysis. The resultant output requires editing and it requires a moderate level of statistical knowledge to choose items from the output.

Power and Sample Size Programme [3]

Power and Sample Size programme is abbreviated as PS, is an interactive program for performing power and sample size calculations. It is a free software and can be downloaded using the link (http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize). PS software can be used for estimating the sample size and power for the studies with dichotomous, continuous, or survival response measures.

The PS software provides the sample size and power for the studies which involve independent and dependent groups with continuous outcome measure, for the studies related to survival times and hazard ratio, case-control and cohort studies, studies related to correlation and regression analysis and the studies involving independent and dependent groups with dichotomous outcomes. In PS, the explanation is provided for each of the items in the '***input***' menu.

Another merit of this package is that the output displays the sample size/power with a clear explanation of the calculation which will be helpful for users for writing the thesis or project report.

nMaster [4]

nMaster is another package which is exclusively meant for computation of sample size. This package is developed and marketed by Christian Medical College (CMC) Vellore, India. It provides a sample size for estimation and testing of hypothesis problems.

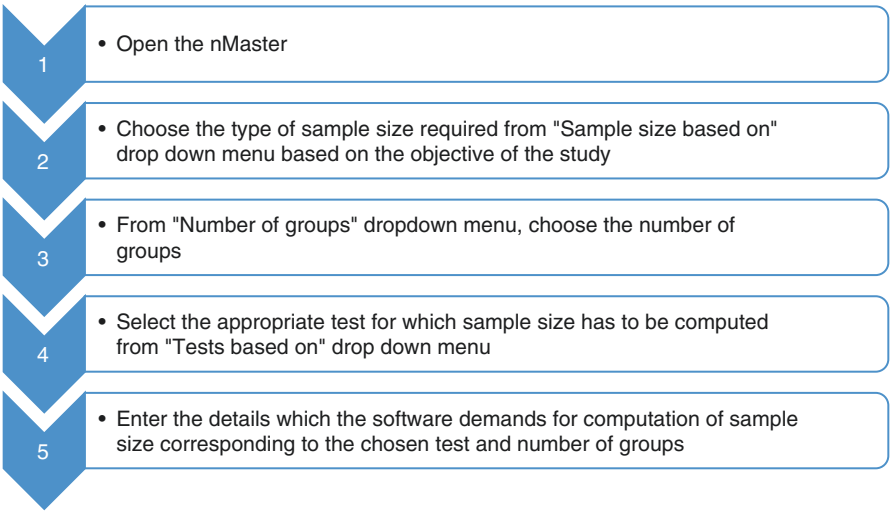


Fig. 2 Steps for using nMaster

Estimating and testing arithmetic means and Proportions. Further, sample size computations are available for Diagnostic test, Regression methods, Survival analysis, Cluster design, Clinical Trials, Epidemiologic Methods and Non-parametric methods.

Figure 2 provides the steps to compute sample size using nMaster.

In case of doubt, the users can click on the '***help***' button, which displays the assumptions of the concerned statistical test, the sample size formula along with an explanation for the terms in the formula and also an example.

Output from nMaster can be saved in different formats/printed.

STATA [5]

Stata is developed and marketed by "Stata Corp". It is the preferred statistical package for public health professionals and epidemiologists. Stata is a powerful statistical package with smart data-management facilities, a wide array of up-to-date statistical techniques and an excellent system for producing high-quality graphs. Stata is fast and easy to use package. It has both menu driven as well as command driven options.

The main window of stata consists of five sub-win dows. The "variables" menu displays the name and label of all the variables in the data-set. "Properties" displays the properties of variables as well as the dataset. Commands are typed in "Commands" window. The executed commands are displayed in "Review" window. Upon clicking a command in review window, it again appears in the command window, which can be re-executed by pressing the enter button to get the output.

Upon clicking the data editor in the main window (exactly below the "Statistics"), opens the data editor sheet, which provides access to the data.

Data can be imported into stata from different sources. The most unique feature of stata is that the value labels assigned to categorical variables can be saved as 'do files' and can be imported and used whenever required (Files ➔Do➔Save).

Stata performs a very wide range of statistical techniques (basic techniques to most advanced techniques like structural equation modelling, multi-level modelling, network meta-analysis etc.). Upon clicking the '***Statistics***' option on the main window, a drop-down of statistical techniques appears.

The options namely (Statistics ➔Summaries, tables and tests ➔ Other tables) compact table of summary statistics, flexible table of summary statistics, table of means, standard deviation and frequencies make stata an ideal choice for analysis of descriptive studies. Presence of calculators; CI calculator (Statistics ➔Summaries, tables and tests➔Summary and descriptive statistics), t-test calculator and effect size calculator (Statistics➔Summaries, tables and tests ➔Classical tests of hypotheses), odds ratio calculator, risk ratio calculator and matched case-control calculator (Statistics➔Epidemiology and related➔ Table for epidemiologists) are highlights of stata. Stata computes sample size for most of the statistical procedures (Statistics➔Power and sample size). It produces sophisticated graphs. Stata also provides example datasets for practice (File ➔ Example datasets), thereby facilitates self-learning of the users.

As stata is also command driven, it has the flexibility of providing customized outputs by executing appropriate commands. Results can be exported to word document/text document/excel.

Stata is a package with a lot of flexibility to customize the analysis based on individual requirement. However, it is not very user-friendly.

Epi Info [6]

Epi Info, as the name suggests it is meant for the analysis of data from epidemiological investigations. It is a freeware which works only in Windows and is developed by Center for Disease Control and Prevention (CDC), Atlanta. Epi Info is a user-friendly package that comes handy with key applications; (1) Creation of data entry forms (2) Data entry (3) Computation of sample size (4) Analysis and (5) Creation of maps. The most remarkable feature of Epi Info is that it is also available as a mobile application that works on smartphones and tablets, which makes it an ideal software for use during outbreak investigations and public health emergencies. The mobile application is endowed with "cloud services" which helps the investigators to collect the data from multiple sites and pool the collected data on a common platform for a combined analysis. An additional exiting feature of Epi Info is "Cloud Data Analytics" that facilitates handling and analysis of large-scale datasets.

Upon opening the Epi Info, it displays a set of six gadgets, which the users can choose based on their requirements. Figure 3 shows the gadgets of Epi Info.

Epi Info enables the creation of customized data entry forms (choose "Create forms"). It has provision for creating a facility for entering text, number, date and time, check-box, yes or no and choosing one among many options. Created data entry forms can be easily edited/modified. Data entry forms can be created only in Windows version. However, the created user forms can also be used in the mobile

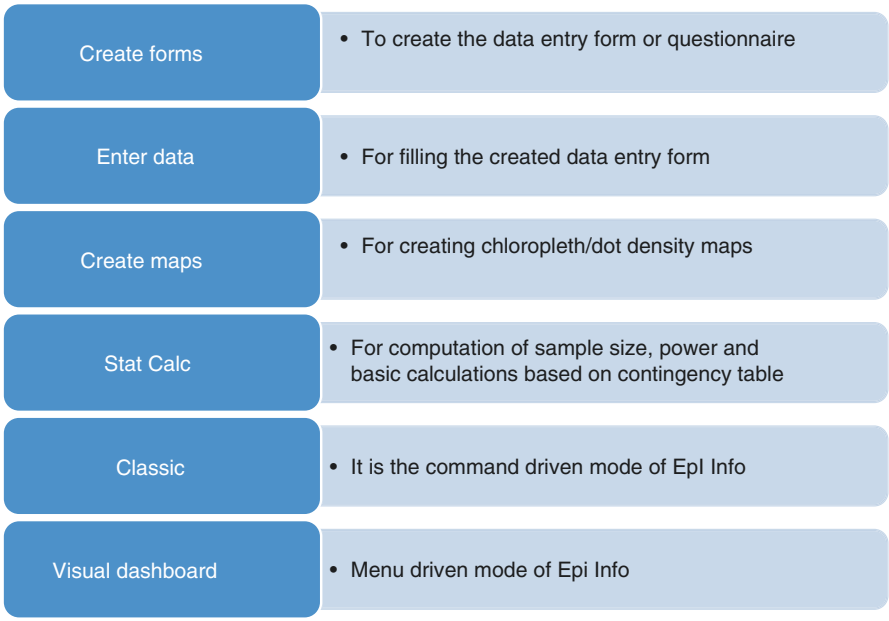


Fig. 3 Gadgets of Epi Info

version. Data can be entered either in Windows version or mobile version. The entered data gets saved in the database and can also be extracted in a Microsoft Excel data sheet.

Epi Info computes sample size for basic epidemiological study designs namely surveys, cross-sectional studies, cohort studies and unmatched case- control studies (choose "Stat Calc"). It is an easy procedure which requires entering the desired information in the checkboxes and in the fraction of a second, Epi Info displays sample sizes for different combinations of confidence levels/cluster sizes depending on the situations.

Epi info has the facility of calculators (choose "Stat Calc"). It has got a calculator for odds ratio, risk ratio, chi-square, Poisson probability, Binomial probability and matched case-control odds ratio.

Epi info has statistical methods that are sufficient for analysis of common epidemiological investigations (choose "Classic" or "Visual dashboard"). It includes descriptive statistics, linear regression, logistic regression and conditional logistic regression. Kaplan-Meier and Cox-proportional Hazards are available only in the classic (command-driven) version. Among graphical illustrations, Epi Info produces basic charts, aberration detection chart, Pareto chart and Epi curve. Additionally, Epi Info produces growth charts according to WHO and CDC standards for different anthropometric measurements. The Epi Info software permits importing of data from other databases; Microsoft access, excel, CSV files and My SQL. Output is displayed as ready to use tables. Outputs can be easily be exported to Microsoft word and excel.

Creation of maps (Choose "create maps")—Epi Info utilizes Geographic Information System (GIS) for creation of "Choropleth maps" and "Dot density maps" which are most essential for preparation of reports during outbreak investigations and public health emergencies

EZR [7]

EZR, which is also referred to as "Easy R" is a menu driven and a user-friendly version of R package. It is a free downloadable statistical package which has provision for statistical techniques such as descriptive statistics, parametric and non-parametric tests, linear regression, logistic regression, survival analysis, meta-analysis, meta-regression, sample size computation etc.

EZR consists of toolbar, command window where the executed commands are displayed and output window to display the output. A user can perform the statistical functions in EZR either by menu driven options or by means of typing commands.

Data from different sources; text, SPSS, Stata, Minitab and excel can be imported into EZR by File➔Import. Users can choose the statistical techniques by clicking on '***Statistical Analysis***'.

Output from EZR can be exported to text document by File➔Save output as.

Figure 4 shows the statistical packages and their utility.

Conclusion

Appropriate statistical analysis and interpretation is very crucial in any research. Numerous computer packages are available for statistical analysis of data. The choice of the statistical package for the preparation of data-base and data analysis depends on the ease, comfort, availability and flexibility. The knowledge in handling a particular package and the type of statistical analysis required also determine the selection. Many of the statistical packages require official licence to use. The researcher should have an idea about the computer packages going to be used for the preparation of data- base and data analysis and accordingly they can plan the format of the data- base.

Case Scenarios

1. You are a Resident in a Medical college and either you are in the stage of preparation of research proposal to be submitted to research monitoring committee in which sample size has to be computed or you have finished collecting the data and about to begin entry and analysis of the collected data.
   (a) Which will be your preferred statistical package for computing sample size?
   (b) Which package would you prefer for data entry?

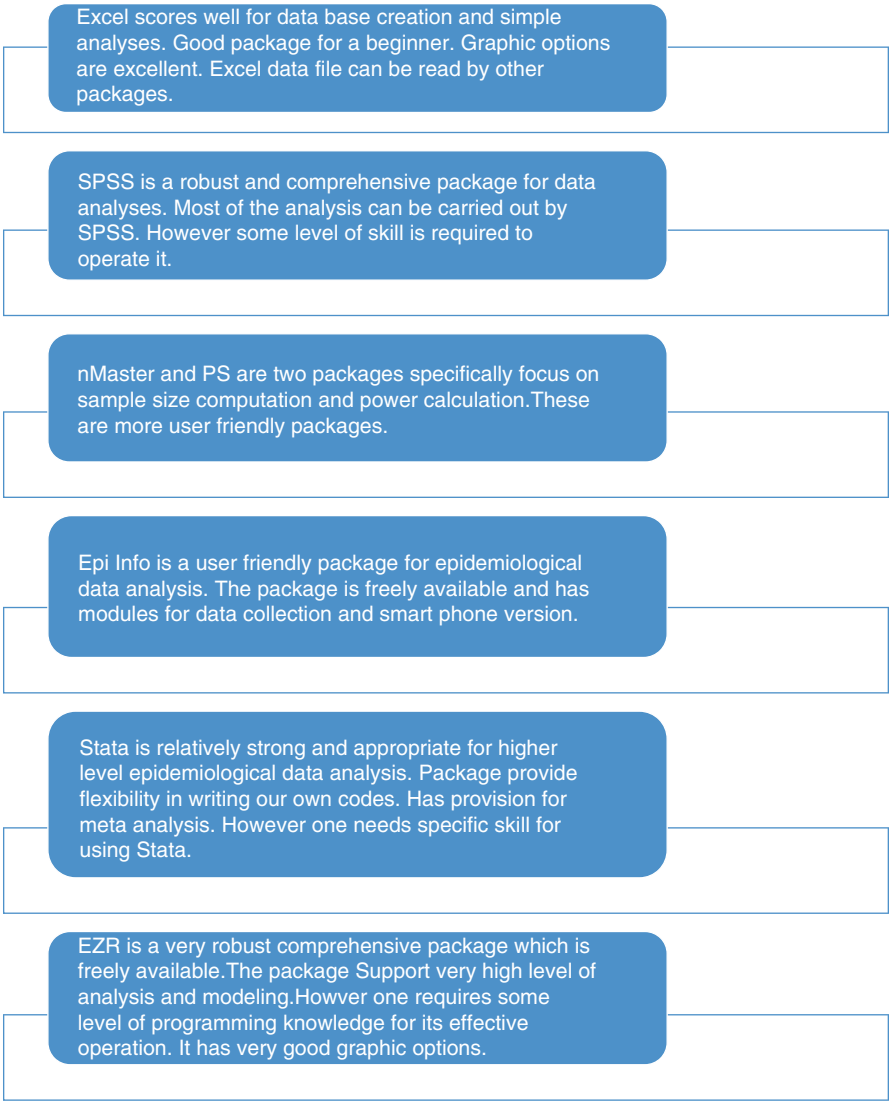


Fig. 4 Statistical packages and their utility

(c) Which statistical package suits your data analysis requirement?
(d) Which package would you use for producing graphs?

2. Your project is in evidence-based medicine and you have to finish meta-analysis as part of a systematic review.
   (a) Which are the statistical packages suits your meta-analysis requirement?
   (b) Which one you prefer among these and why?

References

1. About Microsoft Excel [Internet]. 2017 [cited 2017 Nov 12]. https://products.office.com/en-in/excel.
2. About SPSS [Internet]. 2017 [cited 2017 Nov 12]. https://www.ibm.com/analytics/us/en/technology/spss/.
3. About Power and Sample Size Programme [Internet]. 2017 [cited 2017 Nov 12]. http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize.
4. About nMaster [Internet]. 2017 [cited 2017 Nov 12]. http://www.nmaster.cmc-biostatistics.ac.in/.
5. About Stata [Internet]. 2017 [cited 2017 Nov 12]. http://www.stata.com/.
6. About Epi Info [Internet]. 2016 [cited 2017 Nov 12]. https://www.cdc.gov/epiinfo/index.html.
7. About EZR [Internet]. 2017 [cited 2017 Nov 12]. https://www.r-project.org/about.html.