Exploratory data analysis

From Wikipedia, the free encyclopedia

In statistics, **exploratory data analysis (EDA)** is an approach to analysing data sets to summarize their main characteristics in easy-to-understand form, often with visual graphs, without using a statistical model or having formulated a hypothesis. Exploratory data analysis was promoted by John Tukey to encourage statisticians visually to examine their data sets, to formulate hypotheses that could be tested on new data-sets.

Tukey's championing of EDA encouraged the development of statistical computing packages, especially *S* at Bell Labs: The *S* programming language inspired the systems 'S'-PLUS and *R*. This family of statistical-computing environments featured vastly improved dynamic visualization capabilities, which allowed statisticians to identify outliers and patterns in data that merited further study.

Tukey's EDA was related to two other developments in statistical theory: Robust statistics and nonparametric statistics, both of which tried to reduce the sensitivity of statistical inferences to errors in formulating statistical models. Tukey promoted the use of five number summary of numerical data—the two extremes (maximum and minimum), the median, and the quartiles—because these median and quartiles, being functions of the empirical distribution are defined for all distributions, unlike the mean and standard deviation; moreover, the quartiles and median are more robust to skewed or heavy-tailed distributions than traditional summaries (the mean and standard deviation). The packages *S*, *S*-PLUS, and *R* included routines using resampling statistics, such as Quenouille and Tukey's jacknife and Efron's bootstrap, that were nonparametric and robust (for many problems).

Exploratory data analysis, robust statistics, nonparametric statistics, and the development of statistical programming languages facilitated statistician's work on scientific and engineering problems, such as on the fabrication of semiconductors and the understanding of communications networks, which concerned Bell Labs. These statistical developments, all championed by Tukey, were designed to complement the analytic theory of testing statistical hypotheses, particularly the Laplacian tradition's emphasis on exponential families.[1]

Contents

EDA development

Tukey held that too much emphasis in statistics was placed on statistical hypothesis testing (confirmatory data analysis); more emphasis needed to be placed on using data to suggest hypotheses to test. In particular, he held that confusing the two types of analyses and employing them on the same set of data can lead to systematic bias owing to the issues inherent in testing hypotheses suggested by the data.

The objectives of EDA are to:

- Suggest hypotheses about the causes of observed phenomena
- Assess assumptions on which statistical inference will be based
- Support the selection of appropriate statistical tools and techniques
- Provide a basis for further data collection through surveys or experiments

Many **EDA** techniques have been adopted into data mining and are being taught to young students as a way to introduce them to statistical thinking.[2]

Techniques

There are a number of tools that are useful for EDA, but EDA is characterized more by the attitude taken than by particular techniques.[3]

Typical graphical techniques used in EDA are:

- Box plot
- Histogram
- Multi-vari chart
- Run chart
- Pareto chart
- Scatter plot
- Stem-and-leaf plot
- Odds ratio
- Chi-square
- Multidimensional scaling
- Targeted projection pursuit

Typical quantitative techniques are:

- Median polish
- the Trimean
- Letter values
- Resistant line
- Resistant smooth
- Rootogram
- Ordination

History

Many EDA ideas can be traced back to earlier authors, for example:

- Francis Galton emphasized order statistics and quantiles.
- Arthur Bowley used precursors of the stemplot and five-number summary (Bowley actually used a "seven-figure summary", including the extremes, deciles and quartiles, along with the median - see his *Elementary Manual of Statistics* (3rd edn., 1920), p. 62 – he defines "the maximum and minimum, median, quartiles and two deciles" as the "seven positions").
- Andrew Ehrenberg articulated a philosophy of data reduction (see his book of the same name).

The Open University course *Statistics in Society* (MDST 242), took the above ideas and merged them with Gottfried Noether's work, which introduced statistical inference via coin-tossing and the median test.

Software

- GGobi is a free software for interactive Data visualization
- Mondrian is a free software for interactive Data visualization
- OpenSHAPA (http://www.openshapa.org) (modern open source successor to MacSHAPA), permits analysis of various media files (e.g. video, sound).
- CMU-DAP (Carnegie-Mellon University Data Analysis Package, FORTRAN source for EDA tools with English-style command syntax, 1977).
- Data Applied, a comprehensive web-based data visualization and data mining environment.
- Fathom (for high-school and intro college courses).
- JMP, an EDA package from SAS Institute.
- KNIME Konstanz Information Miner – Open-Source data exploration platform based on Eclipse.
- LiveGraph (open source real-time data series plotter).
- Orange, an open-source data mining software suite.
- SOCR provides a large number of free Internet-accessible.
- DASS-GUI – data mining framework written in C++ and Qt.
- TinkerPlots (for upper elementary and middle school students).
- Weka an open source data mining package that includes visualisation and EDA tools such as targeted projection pursuit

See also

- Anscombe's quartet, on importance of exploration
- Predictive analytics
- Structured data analysis (statistics)
- Configural frequency analysis

References

1. ^ "Conversation with John W. Tukey and Elizabeth Tukey, Luisa T. Fernholz and Stephan Morgenthaler, *Statistical Science*, Volume 15, Number 1 (2000), 79–94.
2. ^ Konold, C. (1999). Statistics goes to school. *Contemporary Psychology*, *44(1)*, 81–82.

3. ^ "We need both exploratory and confirmatory" John W. Tukey *The American Statistician*, *34(1)*, (Feb., 1980), pp. 23–25.

Bibliography

- Andrienko, N & Andrienko, G (2005) *Exploratory Analysis of Spatial and Temporal Data. A Systematic Approach*. Springer. ISBN 3-540-25994-5
- Hoaglin, D C; Mosteller, F & Tukey, John Wilder (Eds) (1985). *Exploring Data Tables, Trends and Shapes*. ISBN 0-471-09776-4.
- Hoaglin, D C; Mosteller, F & Tukey, John Wilder (Eds) (1983). *Understanding Robust and Exploratory Data Analysis*. ISBN 0-471-09777-2.
- Leinhardt, G., Leinhardt, S., *Exploratory Data Analysis: New Tools for the Analysis of Empirical Data*, Review of Research in Education, Vol. 8, 1980 (1980), pp. 85–157.
- Theus, M., Urbanek, S. (2008), Interactive Graphics for Data Analysis: Principles and Examples, CRC Press, Boca Raton, FL, ISBN 978-1-58488-594-8
- Tucker, L; MacCallum, R. (1993). *Exploratory Factor Analysis*.
- Tukey, John Wilder (1977). *Exploratory Data Analysis*. Addison-Wesley. ISBN 0-201-07616-0.
- Velleman, P F & Hoaglin, D C (1981) *Applications, Basics and Computing of Exploratory Data Analysis* ISBN 0-87150-409-X
- Young, F. W. Valero-Mora, P. and Friendly M. (2006) *Visual Statistics: Seeing your data with Dynamic Interactive Graphics* (http://www.uv.es/visualstats/Book) . Wiley ISBN 978-0-471-68160-1

External links

- Carnegie Mellon University – free online course on EDA (http://oli.web.cmu.edu/openlearning/forstudents/freecourses/statistics)

Categories: Exploratory data analysis | Data analysis