Introduction to statistical software

Alberto Castellini
Introduction

- There exist several tools for statistics.

- In the following we briefly analyze and compare few of them:
  - SAS JMP
  - SPSS
  - Weka
  - Stata
  - R
  - Excel / OpenOffice Calc
  - Matlab
Introduction

• There exist several tools for statistics.

• In the following we briefly analyze and compare few of them:
  – SAS JMP
  – SPSS
  – Weka (--projects--)
  – Stata
  – R
  – Excel / OpenOffice Calc (--lab, projects--)
  – Matlab (--lab, projects--)
Introduction

• **Notice:** This is also an alternative (and applicative) way to look at the statistical methods learnt in the previous lessons (learning statistical software is also a good way to learn statistics).

• **After this lesson you should know some basic functionalities** of the software examined and you should be able to **choose the right tool** depending on the kind of analysis you are performing.

“The real voyage of discovery consists not in seeking new landscapes, but in having new eyes.”

Marcel Proust
SAS JMP
SAS JMP

• Features
  – “JMP is a software for interactive statistical graphics” (from the JMP introductory guide)
  – Good graphical interface to display and analyze data
  – Commercial

• References
SAS JMP

• Overview
  – Introduction to JMP
  – Summarizing data
  – Looking at distributions
  – Regression and curve fitting
  – Exploring data
  – Multiple regression
IBM SPSS
IBM SPSS

• Features
  – SPSS solutions:
    • IBM SPSS Statistics Family
    • IBM SPSS Modeler
    • IBM SPSS Data Collection Family
    • IBM SPSS Deployment Family
  – “SPSS Statistics is a comprehensive system for analyzing data. SPSS Statistics can take data from almost any type of file and use them to generate tabulated reports, charts, and plots of distributions and trends, descriptive statistics, and complex statistical analyses... Simple menus and dialog box selections make it possible to perform complex analyses without typing a single line of command syntax. The Data Editor offers a simple and efficient spreadsheet-like facility for entering data and browsing the working data file”
  – Commercial
IBM SPSS

- Features
  - SPSS Statistics add-on:
    - Regression
    - Advanced statistics
    - Custom tables
    - Forecasting
    - Categories
    - Conjoint
    - Exact tests
    - Missing values
    - Complex samples
    - Decision trees
    - Data preparation
    - Neural networks
    - EZ RFM
    - Amos

- Commercial
IBM SPSS

• Features
  – SPSS Statistics add-on:
    • Regression
    • Advanced statistics
    • Custom tables
    • Forecasting
    • Categories
    • Conjoint
    • Exact tests
    • Missing values
    • Complex samples
    • Decision trees
    • Data preparation
    • Neural networks
    • EZ RFM
    • Amos
  – Commercial
IBM SPSS

• References
  – http://www.spss.com/ : SPSS official website
IBM SPSS

• Overview
  – Introduction to SPSS
    • Opening a data file
    • Running an analysis
    • View results
    • Creating charts
  – Reading data
  – Using the data editor (hints)
  – Examining summary statistics for individual variables
Weka
• Features
  – Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code.
  – Weka contains tools for:
    • data pre-processing,
    • classification,
    • regression,
    • clustering,
    • association rules
    • visualization.
  – It is also well-suited for developing new machine learning schemes.

– Free, open source
– Java implementation
Weka

Features
- Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code.
- Weka contains tools for:
  - data pre-processing,
  - classification,
  - regression,
  - clustering,
  - association rules
  - visualization.
- It is also well-suited for developing new machine learning schemes.
- Free, open source
- Java implementation
Weka

• References
  – Book (in library): Data Mining – Practical machine learning tools and techniques. Ian H. Witten & Eibe Frank (University of Waikato)
    • Chapter 1: What’s it all about?
    • Part II The Weka machine learning workbench
Data mining and machine learning (hints)

• We are overwhelmed with data (computers, multigigabyte disks, ...)
• Growing gap between the generation of data and our understanding of it
• Looking for patterns in data, discover the mechanisms that govern systems (biological, economical, social,...), predicting what will happen in new situations
• **Data mining** is defined as the process of discovering patterns, automatically or semiautomatically in large quantities of data
• We are interested in techniques for finding and describing (learning) structural patterns from set of samples (input-output)
Weka

• Overview
  – Introduction to Weka
  – The Explorer
  – The Knowledge Flow Interface
  – The Experimenter
STATA

• Features
  – Data analysis and statistical software
  – “Stata is a complete, integrated statistical package that provides everything you need for data analysis, data management, and graphics” (from Stata website)
  – Commercial

• References
  – http://www.stata.com/ : Stata official website
R

• Features
  – R is a language and environment for statistical computing and graphics. It provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.
  – Free, open source

• References
  – http://www.statmethods.net/index.html : QuickR, a website for experienced users of popular statistical packages such as SAS, SPSS, Stata, who want to quickly access the R language.
OpenOffice Calc / Microsoft Excel

• Features
  – Spreadsheet
  – Simplicity
  – Free (Calc), commercial (Excel)
  – Calc 3 can store a maximum of 65,536 rows with 1024 columns in each sheet, with a maximum of 256 sheets.
  – Excel 2007 can store a maximum of 1,048,576 rows with 16,384 columns in each sheet, with a maximum of sheets depending on the available memory.
OpenOffice Calc / Microsoft Excel

• References:
  – [http://it.openoffice.org/](http://it.openoffice.org/) : OpenOffice website (download, documentation, etc.)
  Notice: all functions are listed along with their formula.
OpenOffice Calc / Microsoft Excel

• Overview:
  – Introducing Calc
  – Entering, Editing, and Formatting Data
  – Creating Charts and Graphs
  – Using Formulas and Functions
Overview:

– Introducing Calc
– Entering, Editing, and Formatting Data
– Creating Charts and Graphs
– Using Formulas and Functions
Exercise 1

Generate a spreadsheet file named “barChart.ods” containing values 0.3 for U.S.’s sales, 0.29 for Japan’s sales, 0.08 for Britain’s sales, 0.25 for Europe’s sales and 0.08 for Other’s sales. Then generate a bar chart for displaying the sales data.
Exercise 2

Generate a spreadsheet file named “BasicStatistics.ods” (p 29 Aczel) containing:

• a named range (cells K5:K44) called “Data” (values: 9, 6, 12, 10, 13, 15, 16, 14, 14, 16, 17, 16, 24, 21, 22, 18, 19, 18, 20, 17);
• measures of central tendency for Data: mean, median, mode;
• measures of dispersion for Data:
  – variance and standard deviation if the data is i) of a sample and ii) of a population; range; interquartile range;
• skewness and kurtosis of Data;
• percentiles (50th, 80th, 90th) and percentile ranks (for 16.0, 19.2, 21.1) of Data;
• quartiles (1st, 2nd, 3rd) of Data;
• sum, size, min, max of Data;
• Chebyshev’s theorem observations:
  – data points within 1.5 stdev from mean
  – minimum predicted by Chebyshev’s Theorem
  – minimum predicted by Empirical Rule
Exercise 3

Generate a spreadsheet file named “TestingPopulationMean.ods” (p 298 Aczel) in which a sample mean is tested for equality and inequality to 2000 (Z-test if stdev is known, T-test if stdev is unknown):

• n=20
• Sample mean=1999.6
• Population stdev= 1.3 (or unknown)
Exercise 4

Generate a spreadsheet file named “TestingPopulationVariance.ods” (p 302 Aczel) in which a sample variance (normal population) is tested for equality and inequality to 1 (Chi-square test):

• Sample size=31
• Sample variance=1.62
Exercise 5

Generate a spreadsheet file named “TestingEqualityOfVariances.ods” (p 358 Aczel) in which the variances of samples in file “TestingEqualityOfVariancesData.ods” are tested for equality and inequality by means of Fisher’s test.
Exercise 6

Generate a spreadsheet file named “SimpleRegression.ods” (p 447 Aczel) in which data of file “SimpleRegressionData.ods” are used for simple linear regression (method of least squares):

• X (quality): independent variable
• Y (Mkt share): dependent variable

Compute the following parameters:
• Estimators b0 and b1 of equation Y = b0 + b1X;
• Standard errors s(b0) and s(b1) of estimators b0 and b1 respectively;
• Coefficient of determination $r^2$, coefficient of correlation r
• Standard error of prediction $s$
• Residuals, SSE, SSR, SST, df(SSE), df(SSR), df(SST)
• T-test for the existence of a linear relationship between X and Y (H0: $\beta_1 = 0$)
• F-test for the existence of linear relationship between Y and X (H0: $\beta_1 = 0$)

Display a scatter plot of data with regression (trend) line and its equation.
Other exercises

7. Generalize the spreadsheet “SimpleRegression.ods” in order to generate a multiple regression model for data in file “MultipleRegressionData.ods” (p 513 Aczel).

8. Open and analyze spreadsheet file “Trend+SeasonForcating.ods” (p 599 Aczel) wherein seasonal indexes are computed by the ratio-to-moving average method.

9. Open and analyze spreadsheet file “ExponentialSmoothing.ods” (p 606 Aczel) wherein the exponential smoothing method is implemented.
To get started, select MATLAB Help or Demos from the Help menu.
Matlab

• Features
  – Commercial
  – “MATLAB is a high-level technical computing language and interactive environment for algorithm development, data visualization, data analysis, and numeric computation. Using the MATLAB product, you can solve technical computing problems faster than with traditional programming languages, such as C, C++, and Fortran” (from the “Getting started guide of Matlab 7”)
  – High-level language for technical computing
  – Development environment for managing code, files, and data
  – Interactive tools for iterative exploration, design, and problem solving
  – Mathematical functions for linear algebra, statistics, Fourier analysis, filtering, optimization, and numerical integration
  – 2-D and 3-D graphics functions for visualizing data
  – Tools for building custom graphical user interfaces
References:

- [http://faculty.olin.edu/bstorey/Notes/matlab.pdf](http://faculty.olin.edu/bstorey/Notes/matlab.pdf) : Matlab introductory tutorial (by Brian D. Storey).
Matlab

Matlab system consists of the following main elements:

• Desktop Tools and Development Environment
• Mathematical Function Library
• The Language
• Graphics
• External Interfaces
Overview:

- Introduction ("Getting started guide", "MATLAB primer tutorial")
  - Documentation
  - Starting and quitting the MATLAB program
  - Calculator
- Matrices and arrays ("Getting started guide")
  - Entering matrices
  - Sum, transpose, and diag
  - Subscripts
  - The colon operator
  - The magic function
  - Expressions
  - Variables
  - Operators
  - Functions
  - Generating matrices
  - Concatenation
  - Basic operations on matrices and arrays
  - Multivariate data
  - Scalar expansion
  - Logical subscripting
  - The find function
  - Controlling command window input and output
Matlab

- Graphics ("Getting started guide")
  - Overview of plotting
  - Creating a graph
  - Graph components
  - Figure tools
  - Plotting tools
  - Arranging graphs within a figure
  - Choosing a type of graph to plot
  - Plot selector
  - Plot catalog
  - Editing plots
  - Setting object properties
  - Property editor
  - Property inspector
  - Plotting two variables with plotting tools
  - Changing the appearance of lines and markers
  - Adding more data to the graph
  - Changing the type of graph
  - Modifying the chart data source
  - Preparing graphs for presentation
  - Exporting graphs
  - Basic plotting functions
  - (Creating mesh and surface plots)
Matlab

• Programming (“Getting started guide”)
  – Flow control
  – Conditional control – if, else, switch, case
  – Loop control – for, while, continue, break
  – Return
  – Multidimensional arrays
  – Cell arrays
  – Characters and text
  – Structures
  – Scripts and functions
  – Types of functions
  – The eval function
  – Vectorization
  – Preallocation
Matlab

- Data analysis (“Getting started guide”, “Data analysis guide”)
  - Preprocessing data
  - Loading data
  - Importing/Exporting data
  - Missing data
  - Outliers
  - Smoothing and filtering
  - Summarizing data
  - Measures of location
  - Measures of scale
  - Shape of distribution
  - Visualizing data
  - 2-D scatter plots
  - Covariance and correlation
  - 3-D scatter plots
  - Scatter plot arrays
  - Exploring data in graphs – data cursor, data brushing, data linking
  - Modeling data
  - Polynomial regression
  - General linear regression
  - FFT
Exercise 1

• Implement a Matlab function called “firstStatistics” with the following features:
  – Input: a column array of numbers
  – Output: mean, median, mode, standard deviation and variance of the input dataset, and the list of outliers (and outliers positions in the input array), where a point is identified as an outlier if its “distance” from the mean is greater than 3 standard deviations.
Exercise 2

• Implement a Matlab function called “fibonacci” with the following features:
  – Input: a number $n$
  – Output: the first $n$ Fibonacci numbers in a row array
<table>
<thead>
<tr>
<th>Name</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Open source?</th>
<th>Typical users</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>Library support; visualization</td>
<td>Steep learning curve</td>
<td>Yes</td>
<td>Finance; Statistics</td>
</tr>
<tr>
<td>Matlab</td>
<td>Elegant matrix support; visualization</td>
<td>Expensive; incomplete statistics support</td>
<td>No</td>
<td>Engineering</td>
</tr>
<tr>
<td>SciPy/NumPy/Matplotlib</td>
<td>Python (general-purpose programming language)</td>
<td>Immature</td>
<td>Yes</td>
<td>Engineering</td>
</tr>
<tr>
<td>Excel</td>
<td>Easy; visual; flexible</td>
<td>Large datasets</td>
<td>No</td>
<td>Business</td>
</tr>
<tr>
<td>SAS</td>
<td>Large datasets</td>
<td>Expensive; outdated programming language</td>
<td>No</td>
<td>Business; Government</td>
</tr>
<tr>
<td>Stata</td>
<td>Easy statistical analysis</td>
<td></td>
<td>No</td>
<td>Science</td>
</tr>
<tr>
<td>SPSS</td>
<td>Like Stata but more expensive and worse</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>