Statistical softwares: introduction

Maarten Jansen

Team and contact

1. Maarten Jansen and Toufik Zahaf
2. Practical information available on
   `https://maarten.jansen.web.ulb.be/teaching/STAT-F-413/index.html`
3. For general questions about the project, always send a mail to **both** teachers!! (Mails sent to one of them only will be ignored)

Objectives

- Retrieve and analyse your own real data
- Use at least two different software systems and two different types of analyses (typically ANOVA and regression, but others are equally welcome: principle component analysis etc.)
- Find your data
  1. at a company, hospital, banks, insurance company: this option is by far the best. If you get data, then also try to get to know what sort of business questions the company/organization is trying to answer: use the data to respond to the questions.
  2. Otherwise (but less preferable) on the internet, e.g.: government data (such as `statbel.gov.be`) This option has the drawback that it is harder to be original and harder to focus on specific business questions.
     The data should be original, in the sense that they must not be popular in scientific papers or textbooks as illustration of a method.
     - Number of births per communality
     - Macro-economical data; per country, european, regional, provinces etc.
     - Socio-economical data

Forbidden data

Not allowed:

- Time series: time dependence of your data is allowed (longitudinal), but time must not be the dominant explanatory variable
- Birth weights of babies

Why not time-series

Example: unemployement rate of a country (dependent variable) with covariates: economic and social indicators. Measurements: annual averages.

Problem: the successive observations of unemployement rate are highly correlated, these are no independent observations, even not after regression for the covariates. Time is an important factor, and has to be modeled not as a mere covariate but in a way that expresses autocorrelation (autoregression – cross-correlation).

These models are very specific and subject of other classes.

But: number of car accidents, with time one of the covariates should be fine (car accidents are not auto-correlated)

Note on the data size: large enough...

- The number of observations should be **large enough**
  - typically, but dependent on the specific nature of your subject and the characteristics of the data, some 50 at least
  - A minimum sample size is preferred because small samples require very outspoken trends to be confirmed in testing and also many of the estimation and testing methods (ANOVA, Student's t, Fisher F, regression) rely on the central limit theorem.
  - An alternative for small samples can be exact nonparametric tests, such as permutation or randomization tests for the comparison of two populations. Such tests are often too limited as basis for a sound statistical study.

... but not too large

- The number of observations should **not** be **too large**
  - Say, at most a couple of 100 observations, not more than 1000.
  - In many situations (assurance companies,...) there is access to thousands of data, but then it is better to start with a random subsample
  - Indeed, with thousands of data, it is easy to reveal many statistically **significant** trends, even if these trends are physically **irrelevant**
  - Thousands of data are a subject for **data mining**, which often uses a subset of the data for **training** with statistical techniques, and the rest of the data for **validation**

Evaluation

The homework consists of the analysis of your own real data set. Each work is individual, and at least two software packages should be used.

A report should have (typically) **10 to 20 pages**, figures, tables and references included.

Your report should have **title** revealing the topic(s) of your study (so not just "Project Statistical Software, STAT-F-413": that can be used as sub-title)

Evaluation (1)

1. Originality of the subject(s), discussion of the problem; 3/20

- All reports must start with a formulation of problem and objectives, in non-statistical terms.
- The work should not discuss time series or longitudinal data, as they are subject of more specialised courses.
- The data should allow two types of methods for analysis (see 3 and 4 below). If this is not possible, it is allowed to discuss two different data sets.
- The origin of the data should be clearly mentioned in the report (including, if applicable, references).

Evaluation (2 and 3)

2. Exploratory analysis, quality of the graphical summaries; 3/20

- Focus on relations between two or more variables and comparison of two or more subsamples (subpopulations)
  (See details in section on descriptive statistics)
- See section on Descriptive statistics for guidelines

3. Selection of methods for statistical analysis and inference and correct usage of these methods; 6/20

- analysis of variance, co-variance analysis, etc...
- multiple linear regression, principle component analysis, etc...
- All final models (resulting from regression, ANOVA) should be **validated**: that means, after estimation and testing, (for instance) normality, homoscedasticity, independence of the residuals should be verified (graphically/by testing), etc.
  Valid conclusions require validated models
- Well established methods, other than those discussed in this class, are also welcome to be used.

Evaluation (4 and 5)

4. Quality of report, conclusions; 4/20

- This includes writing style and spelling.
- The language of the report may be French or English.
- Good reports are complete and concise: quality is more important than quantity. The ratio length/information will be used as quality measure.

5. Statistical competence 4/20

- Measured by the level of detail
- and by the answers to questions during the oral presentation

Deadlines

- Submission of project and analysis: Th, 2 May 2024
- Oral discussion: We, 15 May 2024
- The oral discussion will be short and strictly time-limited (max. 20 minutes = 13 minutes presentation + 7 minutes discussion)

An overview of statistical software packages

1. Spreadsheets such as Microsoft Excel
   - Example: Microsoft Excel
   - Limited possibilities
   - XLSTAT: data analysis with Excel
2. Relational database management systems such as Microsoft Access also perform poorly as statistical software
3. Statistical software packages
   - SPSS (Statistical Package for Social Sciences), now know as PASW
   - SAS
   - Statistica
   - Minitab
   - STATIab
   - SPAD
   - Systat/Mystat (free)
   - Statgraphics
4. Econometrical software packages

   - TSP
   - E-Views
   - RATS
5. Scientific (Software) Libraries
   - IMSL (International Math & Stat Library)
   - NAG (Numerical Algorithm Group)
6. Interactive programming environments with statistical software libraries
   - Gauss
   - R (R started off as a free clone of S-plus)
   - Matlab, with stat-toolbox; free (open source) alternatives: GNU Octave, Scilab
   - General high-level programming languages (open source): Python, Julia

A few links

- **R-CRAN = Comprehensive R Archive Network**
  `http://cran.r-project.org/`
  `http://www.r-project.org/`

- **Free introduction to R** (requires creation of account)
  `https://www.datacamp.com/courses/free-introduction-to-r`
  Free introduction to python
  `https://www.datacamp.com/courses/intro-to-python-for-data-scie`
- **Campus licences** for ULB students (also for ULB teachers and researchers)
  `https://sisc.ulb.ac.be/shop/`
  includes free versions of SPSS, Matlab, Stata
- Free alternatives to Matlab
  - **SciLab** `https://www.scilab.org`
  - **GNU Octave** `https://www.octave.org`