Flames courses in statistics

organized in co-operation with ICES

in 2012-2013

Framework

These courses are initiatives organized in the framework of Flames (**Fla**nders Training Network for **Me**thodology and **S**tatistics).

More information about Flames is available on the [Flames website](#).

1. [Database Management and SQL](#)

2. [Introduction to R](#)

3. [Software development in R](#)

Database Management and SQL

Dates

January 17, 24 and 31, February 7, 2013, from 5.30 pm to 9 pm.

Venue

Faculty of Science, <u>Campus Sterre</u>, Krijgslaan 281, Building S9, Ghent. PC room K.Zuse, 3rd floor.

Description

Data is abundant, actionable insights are scarce. Nowadays, the digital revolution allows companies and researchers to capture, store and manage massive volumes of data from a wide range of internal and external sources. Depending on the research discipline, researchers can often get access to the data in scientific data repositories. Some of the available data is structured (e.g. drug prescription data, purchase transaction data, protein database, etc.) whereas other data is unstructured (e.g. social media data, location-services data, website text data, etc.). But without the proper systems and processes to turn these massive data volumes into value, companies and researchers find themselves drowning in the data deluge.The key question is no longer how to gather more data, but how to turn the data that already exists into meaningful insights.

This course will give you the tools to manage structured data as stored in a Relational DataBase Management System (RDBMS). It will enable you to understand and analyze the structure of relational databases. You will learn how to draw insights from the data stored in such a RDBMS by writing appropriate and efficient queries in SQL (structured query language) programming language (in casu: Oracle SQL). You will have ample opportunities to put theory into practice during numerous exercises on PC querying various relational databases with Oracle SQL Developer.

Session 1

Discussion of the importance of database management systems. Explanation of the difference between structured and unstructured data and its implications for data management (relational database management systems vs unrelational database management systems like Hadoop File system HDFS). Which principles are at the basis of building a good database? How to express the relational structure by defining an appropriate Entity Relationship Diagram? How to read an Entity Relationship Diagram?

Session 2

How to query a relational database? Structured Query Language and SQL programming dialects. Core SQL objects and data types. Basic select statement. Difference between "where" statement and "having" statement in combination with "group by" statement. SQL functions allowing to find the minimum, average or maximum value, the first or last occurrence, etc.

Session 3

Advanced select statement: inner and outer joins, correlated queries, with statement. These advanced statements allow you to gain rich insights by combining data from multiple tables.

Session 4

How to increase programming efficiency? Using views and sorting indexes. How to update a table? How to define a stored procedure? An integrated exercise will allow you to test your ability to gain insights from a relational database by defining appropriate and efficient SQL queries.

Target audience

This introductory course targets participants who are interested in how to turn structured data as stored in a relational database management system into valuable insights.

Course prerequisites

The course is open to all.

Teacher

 Dr. **Anita Prinzie** obtained a PhD in Applied Economics from the University of Ghent specializing in data mining and analytical Customer Relationship Management (aCRM). Anita also worked as a post-doctoral researcher at Ghent University and Manchester Business School. Triggered by the challenge of translating analysis insights into value generating business recommendations, she joined Solutions-2 in 2010 as senior consultant. She is now innovation manager responsible for managing the innovation process. Anita is also a visiting professor at the Department of Marketing at Gent University. Her main expertise is datamining and predictive modelling with applications in analytical Customer Relationship Management (aCRM). She is also interested in the development of decision-support systems and web personalization. Most recently she is interested in social media mining and natural language processing.

Course material

Copies of lecture notes.

Recommended handbook (optional):

Beginning Oracle SQL (Expert's Voice in Oracle), L. deHaan, K. Morton, T. Gorman, I. Jørgensen, D. Fink (2009), Apress, ISBN: 978-14302-719-70.

The Full SQL Oracle Developer User Manual is available online.

Introduction to R

Dates

15, 18, 22 en 25 april 2013, telkens van 17u30 tot 21u.

Venue

Faculty of Science, Campus Sterre, Krijgslaan 281, Building S9, Ghent

Description

R is an environment for statistical computing and graphics, which is becoming increasingly popular as a tool to get insight in often complex data. While somewhat similar to other programming languages (such as C, Java and Perl), R is particularly suited for data analysis because ready-made functions are available for a wide variety of graphical and statistical techniques (classical statistical tests, linear and nonlinear modeling, time-series analysis, classification, clustering, ...). The base R program can be extended with user-submitted packages, which means new techniques are often implemented in R prior to being available in other software. This is one of the reasons why R is becoming the de facto standard in certain fields such as bioinformatics (Bioconductor) and financial services.

This course introduces the use of the R environment for the implementation of data management, data exploration, basic statistical analysis and automation of procedures.

The course starts with a description of the R GUI, the use of the command line and an overview of basic data structures. The application of standard procedures to import data or to export results to external files will be illustrated. Creation of new variables, subsetting, merging and stacking of data sets will be covered in the data management section. Exploration of the data by histograms, box plots, scatter plots, summary numbers, correlation coefficients and cross-tabulations will be performed. Simple statistical procedures that will be covered are: comparison of observed group means (t-test, ANOVA and their non-parametric versions) and proportions, test for independence in 2-way cross tables and linear regression (focusing on the implementation in R of the statistical methods that are the subject of other modules of the statistics course). Finally, installing new packages and automation of analysis procedures will also be discussed.

Practical sessions and specific exercises will be provided to allow the participants to practice their R skills in interaction with the teachers.

Target audience and Doctoral Schools

This course targets participants with little to no R-programming experience who wish to start using R for their data manipulation, data exploration or statistical analysis.

This course can be incorporated in a UGent doctoral training program and is eligible for a refund under certain conditions.

Course prerequisites

The course is open to all. Knowledge of basic statistical concepts and experience with other programming languages are considered advantages, but not required for learning the R language.

Teacher

Dr. **Geert Silversmit** has a PhD in Physics from Ghent University, post-doctoral experience in analytical chemistry and a training in Statistical Data Analysis. He is currently working at Stat-Gent CRESCENDO, the Industrial Research Fund (IOF) valorization consortium with the mission to valorize UGent statistical expertise.

Course material

Copies of lecture notes.
The presented material is the result of an UA-UGent collaboration.

Software development in R

Dates

Thursday May 16 and 30, 2013, from 5.30 pm to 9 pm.

Venue

Faculty of Science, Campus Sterre, Krijgslaan 281, Building S9, Ghent.

Description

May 16 - Architect: making the most out of R

R has become the lingua franca of statistical computing and Architect is an integrated development environment (IDE) that has been designed to make the most out of R.

Architect is based on the award-winning Eclipse IDE and aims at increasing R users' pleasure, programming comfort and productivity.

The IDE is open source and freely available from [OpenAnalytics](#).

This session will primarily focus on the tool set Architect offers to R users.

May 30 - Building R packages with Architect

One of the reasons for the big succes of R is the possibility to share and distribute modules ('packages') among users in a standardized way.

In this hands-on training we start from an R function and go all the way to building an R package. Topics that will be dealt with are

- anatomy and physiology of an R package
- technical and functional documentation (Rd, Roxygen, vignettes)
- the nuts and bolds of namespaces
- the R package checker
- sharing your package with the world

Doelpubliek

16/05: R users.

30/05: Intermediate R users that have experience writing R functions and have had good exposure to R packages.

Toelatingsvoorwaarden

16/05: The course is open to all. Basic knowledge of R is required.

30/05: Knowledge of writing R functions is required; during the session the Architect IDE will be used to demonstrate the package development process.

Lesgever

Tobias Verbeke is a statistician by training and active member of the R community since 13 year. He heads a team of professional R developers at OpenAnalytics and has managed the design and development of the Architect IDE project since day one. He has authored 100+ R packages crossing the borders of many statistical disciplines and applied domains.

Lesmateriaal

Slides and exercise sheets will be provided by the instructor.