MSc in Official Statistics Statistical Computing: Introduction

Andrew Westlake

Survey & Statistical Computing

63 Ridge Road, London N8 9NP, UK

+44 (0) 20 8374 4723

AJW@SaSC.co.uk (E-Mail)

www.SaSC.co.uk

Aims and objectives

- Some insight into the IT issues and choices that relate to Official Statistics
  - » Help you to contribute constructively
    - to discussions when new systems are being designed, and
    - to management of operational systems
- Cover some standard ideas from computer science and databases
  - » Database methods and Entity-Relationship design
  - » The Relational Database model
  - » Object-based system design using UML (the Unified Modelling Language)
- Application to statistical systems
  - » Statistical metadata and statistical dissemination systems

Outline

- Module is about Computing and IT issues that are of importance for Statisticians
  - » IT specialists are needed to support users and do implementations, but concepts and details are important to success, so Statisticians need to be well-informed
- Special characteristics of Official Statistics
  - » Recurrent processes, distant users, importance of precision (quality) information
- Importance of structure in data
  - » Micro, macro, meta data
- Contrast
  - » between statistical, computer science and various consumer views of data
  - » what is needed, what is important, what is usable, what is valid

Difficult Module

- About new Concepts
  - » Which take time to assimilate
- Assessment
  - » Apply concepts and methods to a system related to your own work
  - » Discuss topic individually
  - » Difficult to do really well
  - » Difficult to fail

Statistical Objectives

- Quality Data Collection
  - » accurate information, accurately described
- Quality Processing
  - » correct processes, valid analyses, accurate documentation, clear and accessible conclusions
- Quality Dissemination
  - » usability, accessibility, suitability, clarity, flexibility, functionality, usefulness
- Quality Use
  - » discovery, understanding, valid use

Overview (1)

- Computing
  - » DataBase ideas are important (see Date)
  - » RDBMS are the main type available (see Codd)
  - » Good implementations exist
  - » SQL is an important standard (see Date & Darwin)
  - » Object-Oriented ideas pervade programming
  - » Importance of Conceptual Models
    - Identify important structures, processes and functionality, without getting too involved in specific cases

- Project development
  - » Its not easy, statisticians need to get involved
  - » Tools and methods available to help structure ideas and processes, particularly UML

Overview (2)

- Statistics
  - » Statistical problems are different
  - » Important to take a broad, well-informed view
  - » Ideal solutions are elusive
- Standards and standardisation
  - » Aid to communication between different disciplines
  - » Conceptual structure is an aid to design process
  - » Need semantics in addition to structure

Concepts

- Databases
  - » Rich ways of thinking about data structures and access
- Functionality
  - » Not just tasks, but methods for more general classes of problem
- Usability
  - » Appropriateness to task and mindset of users
- Design
  - » A process with methodologies, not just an option
- Abstraction
  - » Need for generalisation
  - » Thinking about structure and requirements

Methods

- Structure
  - » Entity-Relationship design
  - » UML Class models
- Process
  - » Data flows, process control, development
  - » Requirements, activities, sequences, states, collaboration
- Metadata
  - » Not just the statistical numbers
  - » All the associated information needed to describe, control, explain, validate, discover, ...

Tools

- Relational DataBases
- Object-Oriented Paradigm
- UML – the Unified Modelling Language
- XML – the eXtensible Markup Language
- Internet
- Agile development methods
- Component-based development
- …

Formalities

- Safety instructions

- Coursework and Assessment

- Introductions

Case Studies

- PFFPS – Pakistan Fertility Study
  - » National sample survey
  - » Complex enough to demonstrate structural options
- BCS – British Crime Survey
- HAP – HIV and AIDS reporting system for PHLS
  - » Simplification of processes through re-design of underlying structure.
- Statistical metadata and dissemination
  - » More abstract exploration of requirements and principles

Sequence

- Database methods
  - » Relational databases
    - Entity-Relationship design
    - RDBMS functionality
  - » Why are these ideas useful for statistics
  - » Use PFFPS as example
- Statistical manipulation and Dissemination
  - » Structures and metadata
  - » Objectives and concepts
- Modelling for statistical processes
  - » Methodologies for development and implementation
  - » UML to express designs
  - » Use HAP as example

Abstraction, *generalisation*

- Seeing the bigger picture, not focussing too much on the immediate problems
  - » Top-down vs. Bottom-up approaches
- Not just What, but also How and Why
  - » Examples: Variables, Coding, Statistical functionality
- Idea of generalisation is itself an abstraction
  - » Different abstractions in different areas and between disciplines
  - » Many dimensions and many levels
  - » Statisticians good about methods and variables, but have fixed idea about data structures
  - » IT has rich approach to data structures, processes, functionality, but not about application areas

Levels of Abstraction

- Helpful to think at different levels
  - » Depends on the purpose and objective
  - » Related to 'top-down' vs 'bottom-up' views of systems
  - » Applies to multiple aspects
- Often choose 4 levels for modelling
  - » 0 The real thing, no abstraction
  - » 1 The specification (model) of the real thing
  - » 2 The specification of what is allowed in a model
  - » 3 The specification of what is allowed in a specification
    - (Level 3 is not used much)

Database Modelling

- 2 The Relational Database Model

- 1 The definitions for an real database

- 0 The actual database, containing real data

Variables in Statistics

- E.g Employment Measurements
  - » 0 - Instances of employment for respondents within data files
  - » 1 - Agreed coding standards for Employment
  - » 2 - Concepts: what is Employment?
- Different purposes
  - » 2 - Concepts for Discovery and understanding
  - » 1 - Standards for exchange and comparability, presentation and understanding
  - » 0 - Codes in data for analysis

OMG Architecture for Systems