MSc in Official Statistics
Statistical Computing: Database Methods

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

Database Methods

- A **Database** is:
  - » An organised collection of related information
- Different Models exist for
  - » The structures of information that can be stored
  - » Operations that can be performed on the information
  - » How the collection is organised
- Examples of Database Models
  - » Relational
  - » Object-relational
  - » XML
  - » ...

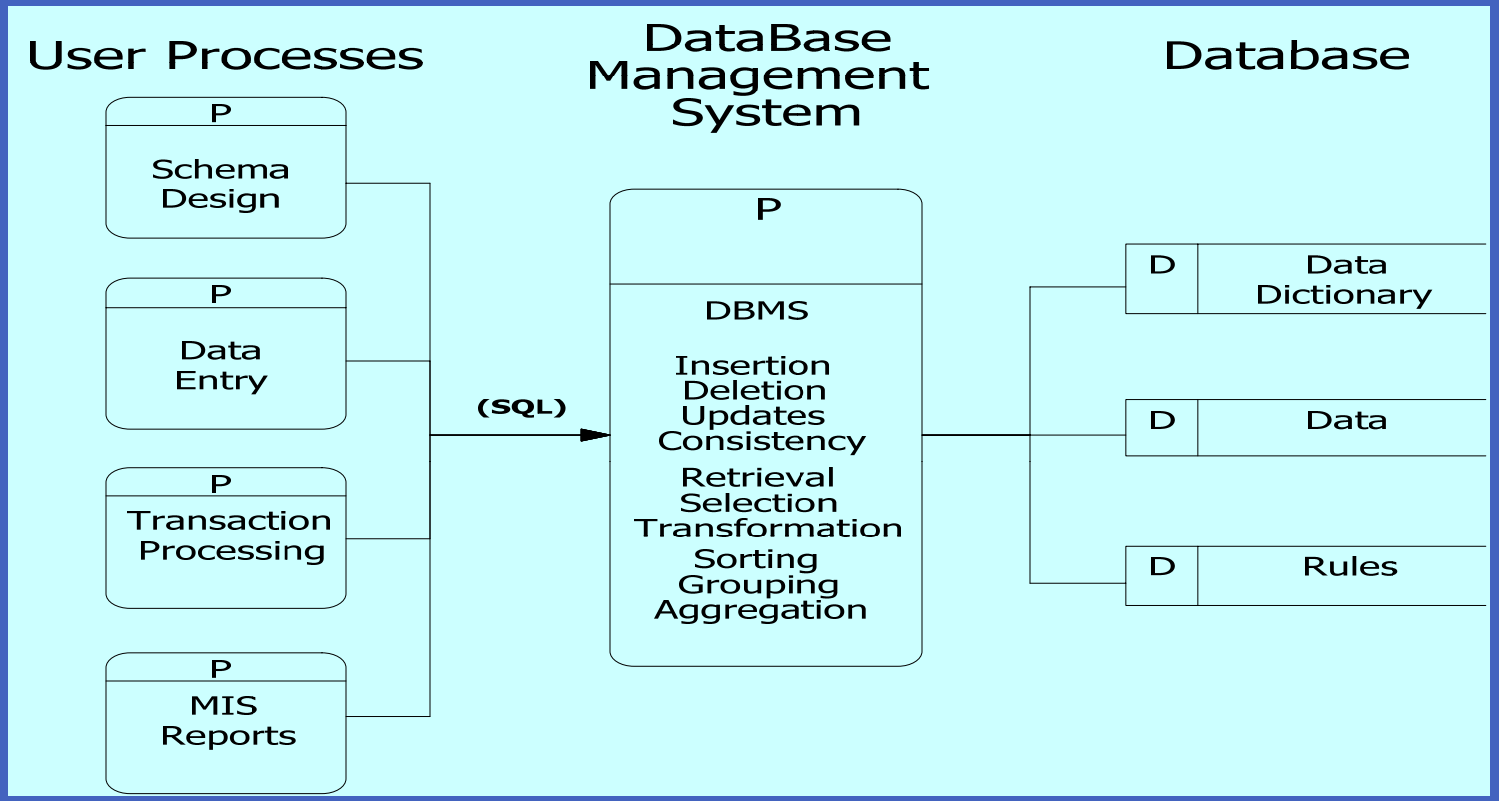
What is a Model?

- Durbin
  - » All models are wrong, but some are useful
- Statistical Models
- Database Models
- IT Models and modelling
  - » Structural, conceptual, logical
  - » Modelling is the process of refining your understanding of a system
- Means different things in different contexts

The Relational Model

- A logical specification of the content and behaviour of a database management system, including
  - The types of structure that can be present in a database
  - The properties of elements that can be stored in these structures
  - The operations that can be performed on these structures and their behaviour
  - Facilities that must be present in the database management system
  - The general nature of the interactions between the database and its users and administrators.
  - » Codd's Rules specify properties that a Relational DataBase Management System (RDBMS) must possess
- SQL
  - » A standard language with which to interact with a RDBMS

Relational Database System Structure



User Processes	***Data Modelling***	SQL	Relational Model	Storage & Access Methods
Applications & Tools, include functionality, semantics appropriate to application	Analysis of the data structures and flows needed to produce the objectives of the system - the **logical model**	Standard interface to an RDBMS, syntax and embedding	Specification of functionality, behaviour and scope	Implementation issue, affects performance

Commercial Relational DataBase Management Systems

- Good implementation of the relational model and SQL
  - » Structure, Organization, Manipulation, Description, Storage, Integrity, Security
    - of Data, but NOT Interpretation
  - » Concern for practical problems of data access and manipulation
  - » Optimized for commercial applications, transaction processing
- Consists of DBMS and a set of tools
  - » Data entry, Reporting, Application development
- Support for Client-Server architecture
  - » i.e. separation of DBMS and programs which use data
  - » allows independent suppliers for tools
  - » allows data use by other applications
- Useful functionality for statistical data management

Objectives of DBMS

- The DBMS layer between the Data Store and the User Processes should mean that
  - » Redundancy can be reduced
  - » Inconsistency can be avoided
  - » The data can be shared
  - » Standards can be enforced
  - » Security restrictions can be applied
  - » Integrity can be maintained
  - » Conflicting requirements can be balanced
  - » Data Independence can be achieved

RDBMS Strengths

- Data Modelling
  - » Useful tools for understanding data structures and flows
- Relational Model
  - » Precise, formal mathematical specification of structure and behaviour
- SQL
  - » International Standard (SQL2, 1992), widely implemented
  - » Various extensions since then – latest in 2008
- Current Implementations
  - » Widely available, well supported, good implementations, integration with other products, add-on market for tools

Relational Model

- Components
  - » Tables, Keys, Integrity, Domains, Nulls, Joins, Security
- Data Independence
  - » Separate processes from information which is not essential for them, e.g. physical aspects of storage
  - » Cf. Statistical Packages
- Views
  - » User processes (and people) see data (dynamically) in the form and structure they need, not as it is decomposed in the database
- Universality
  - » Everything is data, and is processed in the same way (subject to permissions)
- Flexibility of access
  - » Data linking determined at run-time, based on data values
  - » SQL commands can be constructed at run-time

Illustration

- Structure of Pakistan Fertility and Family Planning Survey
- PFFPS in MS Access

HH and HHM Sample Records

Microsoft Access

File Edit View Insert Format Records Tools Window Help Adobe PDF

Type a question for help

HH : Table

	RECORD_	PROVI	AREA	CLUSTE	H_NO	DISTRIC	VISIT	F_DAY	F_MON	F_YEA	INTER	SUPER	DEO	RESULT	Q16T	Q16M	Q16F	Q17
▶	HID	5	1	57	1	1	1	26	1	97	10	20	4	1	5	3	2	1
	HID	5	1	57	12	1	1	27	1	97	12	20	4	1	7	3	4	1
	HID	5	3	58	6	1	1	6	2	97	10	20	4	1	5	3	2	1
*																		

Record: 1 of 3

HHM : Table

	RECORD_	PROV_	AREA_	CLUST_	H_NO_	Q01	Q03	Q04	Q05	Q06	Q07	Q08	Q09	Q10	Q11	Q12	Q13	Q14	Q15
▶	HHM	5	1	57	1	1	1	1	1	1	35	1	1	12	7	1	0	1	0
	HHM	5	1	57	1	2	2	1	1	2	33	1	1	14	7	1	0	2	97
	HHM	5	1	57	1	3	3	1	1	1	5	6	1	0	1	1	2	1	1
	HHM	5	1	57	1	4	3	1	1	2	3	6	3	97	7	1	2	1	1
	HHM	5	1	57	1	5	3	1	1	1	1	6	3	97	7	1	2	1	1
	HHM	5	1	57	12	1	1	1	1	1	39	1	1	6	7	2	97	1	0
	HHM	5	1	57	12	2	2	1	1	2	38	1	1	5	7	1	0	1	0
	HHM	5	1	57	12	3	3	1	1	2	18	6	1	10	1	1	2	1	1
	HHM	5	1	57	12	4	3	1	1	2	14	6	1	9	1	1	2	1	1
	HHM	5	1	57	12	5	3	1	1	1	12	6	1	5	1	1	2	1	1
	HHM	5	1	57	12	6	3	1	1	2	8	6	1	4	1	1	2	1	1
	HHM	5	1	57	12	7	3	1	1	1	6	6	1	2	1	1	2	1	1
	HHM	5	3	58	6	1	1	1	1	1	34	1	1	14	7	1	0	2	97
	HHM	5	3	58	6	2	2	1	1	2	25	1	3	97	7	1	0	2	97
	HHM	5	3	58	6	3	3	1	1	1	4	6	3	97	7	1	2	1	1
	HHM	5	3	58	6	4	3	1	1	2	3	6	3	97	7	1	2	1	1
	HHM	5	3	58	6	5	3	1	1	1	2	6	3	97	7	1	2	1	1
*																			

Record: 1 of 17

Datasheet View NUM

Components of a SQL database

- Data type
  - » Integer, Real, String, Date, Memo, etc
- Field
  - » defined over a data type, has a name, cf. variable.
  - » NULL values supported. Can have constraints
- Record
  - » a set of values, one associated with each field
- Table
  - » defined over a set of fields, has a name, consists of a set of records, can have keys and indexes
- SQL DataBase
  - » a set of tables, can have other properties, including relationships and implementation details
- PFFPS example

SQL

- International Standard, actively revised
  - » SQL2 (1992) has major improvements related to Domains and User Integrity Rules
  - » Later versions (1999, 2003, 2008) offer minor changes, not widely implemented
- Widely available in good RDBMS software
- Text (script) language, used by programs and people
  - » Stored or constructed at run-time
  - » Easy for simple tasks, but limited in scope
- Designed to support tools which are independent of the DBMS
  - » cf. Client-Server architecture
- User and Programmer skills portable across products and sites
- Has sections for
  - » Defining database structure (DDL),
  - » Manipulating database content (DML)
  - » Ensuring database integrity, and
  - » Managing database security

Views

- Stored definition about how to select and manipulate data from the database
  - » Important idea, with wide implications
  - » Implemented as Queries (SQL Select statement)
- Result looks like a table
- Can be used like a table in many contexts
  - » Viewing data in the form needed by the user
    - Can sometimes be used for data entry, but depends on the form of query
- Dynamic evaluation
  - » Ensures that the viewed information is up to date
    - May be inefficient if the information does not change

Current Implementations

- Stable, Mature products
  - » Major products easily scaleable across wide range of hardware. Oracle, MS SQL Server
  - » Good PC products now available, particularly Access, MySQL
- Useful Tool kits provided
  - » Data Entry and retrieval screens, report writers
  - » Active market in add-on products
- Client-Server facilities
  - » Many packages can act as clients, e.g. SAS, SPSS
  - » Efforts towards standardization of Client-Server communications, ODBC, ODAPI, XML
- Design tools
  - » Various systems for Entity-Relationship models, and accompanying code development

Summary

- Relational databases are ubiquitous, and are useful for large-scale data collections
- Some manipulation and aggregation operations can be done more easily than in statistical packages
- Relational model is a useful way of thinking about data structures
- Implementations do not address issues of importance to Statisticians
- IT staff and Statisticians have different ways of thinking about data – we both have things to learn
- MS Access is a useful tool for manipulating moderate amounts of data with more complex structure
- Not a replacement for statistical packages for statistical analysis