22S:166
Computing in Statistics

Intro to relational database concepts

Lecture 15
Oct. 12, 2009

Kate Cowles
374 SH, 335-0727
kcowles@stat.uiowa.edu

Introduction to relational database concepts

- database: a system for storing data
- *relational* database model has become the de-facto standard for the design of databases both large and small
- storage of data for use in statistical analysis

ideally should follow this model

- today's lecture deals with two related topics
  - efficient storage of data (applies to setting up datafiles for use by SAS or any other analysis system)
  - some aspects of relational database software (such as Microsoft Access)

Material drawn in part from
`www.citilink.com/~jgarrick/vbasic/database/rdbms.html` and `http://www.citilink.com/~jgarrick/vbasic/database/fundamentals.html`

What is a relational database?

- relational database stores all its data in "tables"
- table is a set of rows and columns
  - set has no predefined sort order for its elements
  - "record" is database terminology for a row or observation
  - "field" or "attribute" is database terminology for a column or variable

Basic concepts

- Primary and Foreign Keys
- Queries
- Referential Integrity
- Normalization

Flat files (how not to store complex data)

- simplest model for a database
- a single table which includes fields for each element you need to store
- you have probably worked with flat file databases, at least in the form of spreadsheets
- waste storage space and are problematic to maintain

Example: customer order entry system

- You're managing the data for a company with a number of customers, each of which will be placing multiple orders.
- Each order can have one or more items

Data that we wish to record for each component of the application

- Customers
  - Customer Number
  - Company Name
  - Address
  - City, State, ZIP Code
  - Phone Number
- Orders
  - Order Number
  - Order Date
  - PO Number
- Order Line Items
  - Item Number
  - Description
  - Quantity
  - Price

Problems with a flat file for representing this data

- Each time an order is placed, you'll need to repeat the customer information, including the Customer Number, Company Name, etc.
- What's worse is that for each item, you not only need to repeat the order information such as the Order Number and Order Date, but you also need to continue repeating the customer information as well.
- Let's say there's one customer who has placed two orders, each with four line items. To maintain this tiny amount of information, you need to enter the Customer Number and Company Name eight times.
- What if the company should send you a change of address?

- unacceptable aspects of flat file storage
  - effort required to maintain the data
  - likelihood of data entry errors causing inconsistency in customer address between records

Solution: use a relational model for the data

- each order entered is related to a customer record
- each line item is related to an order record
- relational database management system (RDBMS) is a piece of software that manages groups of records which are related to one another

Break flat file into three tables

- Customers
  - CustID
  - CustName
  - CustAddress
  - CustCity
  - CustState
  - CustZIP
  - CustPhone
- Orders
  - OrdID
  - OrdCustID (new field)
  - OrdDate
  - OrdPONumber

- OrderDetails
  - ODID
  - ODOrdID (new field)
  - ODDescription
  - ODQty
  - ODPrice