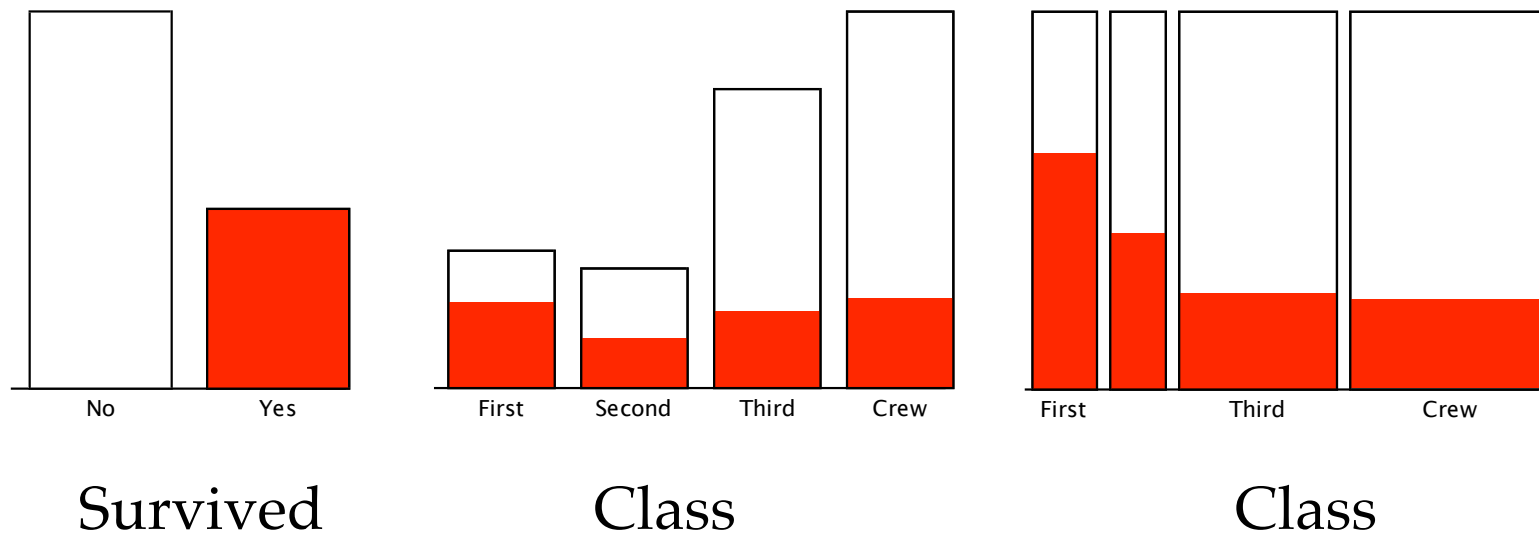# Exploratory Statistical Software

Antony Unwin

Augsburg University

Spineplots

Survival numbers and rates on the Titanic
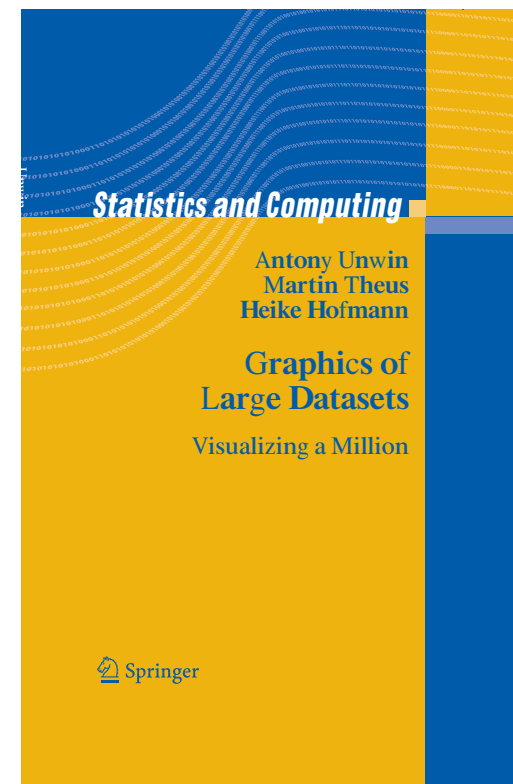
Other underused methods?

- Bayesian models

- Design of experiments

- Latent class methods

- Nonlinear models

  … everything?

Data Visualization

Data Visualization is a big part of exploratory data analysis …

Plan of the talk

- What is exploratory statistics?

- What tools are involved?

- What software do you need for this?

- What software is available?

- What should be developed in the future?

Statistics is … ?

- Design study

- Collect data

- Prepare tables and summaries

- Draw graphical displays

- Build models and test hypotheses

R.A. Fisher

"The elaborate mechanism built on the theory of infinitely large samples is not accurate enough for simple laboratory data."

Statistical Methods for Research Workers (1925)

might be restated to-day as

"The elaborate mechanisms of classical statistics for analysing small samples of simple laboratory data are not enough for large, complex datasets."

(Now)

Exploratory statistics is

- Tukey 1962 "The future of data analysis"

- EDA
  - to generate ideas, not to test hypotheses
  - informative rather than descriptive

- Exploratory Modelling Analysis
  - search in model space
  - model ensembles

… (and)

- Checking missing values

- Assessing data quality

- Data cleaning

- Data filtering

- Reformatting data

- Restructuring data

Exploratory tools

- Stem and leaf plots, median polishing, pencil and paper — NO

- Boxplots, Tukey smoothers, personal computers — MAYBE …

- Interactive graphics, model ensembles, laptops — YES!

An exploratory interlude

Web words data for marketing

- — keyword
- — date
- — # page impressions, #clicks
- — # sales and turnover
- — cost
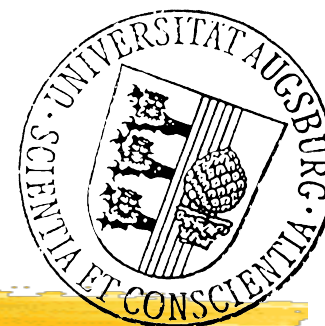
13675 records 1st Jan to 31st August, 2006

Exploratory Analyses

(Some data preparation has been done in advance.

– Mondrian, TextWrangler, R, Data Desk)

- ## EDA with interactive graphics

  – Mondrian

- ## EMA with linear models

  – R and JGR

ES Software properties

- Fast

  – in computing and in use

- Flexible

  – in functionality and in access

- eFFective

  – in results and in reporting

ES Software style

- Interactive

  – primarily direct manipulation

- Consistent

- Intuitive

  – familiarity and frequency of use

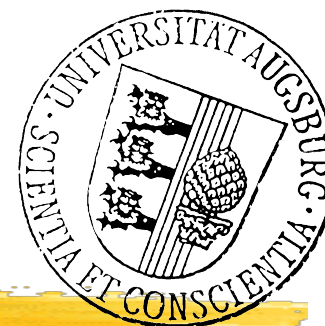 … can it then have lots of functions?

Buxton's Law

$$\text{Usefulness} \sim 1/\text{Functionality}_n$$

where n is the number of functions

Statistical Software

- Big commercial packages
    - SAS or SPSS (Minitab, Statistica?)

- Commercial languages
    - Stata or S-Plus

- Smaller commercial
    - Data Desk or JMP

- Open source language:

Remember the Vasa

The Vasa sank two hours into its maiden voyage in 1628.

Its design had been extended and its guns were too heavy.

quotations

- "R is being used for different purposes than the original interactive EDA environment."

  – Duncan Temple Lang (DSC Auckland 2007)

- "It's not whether you can do it in R, it's how."

  – r-help mailing list

- "It is all too easy for statistical thinking to be swamped by programming tasks."

  – Brian Ripley (R Conference Vienna 2001)

JGR

- A cross-platform GUI for R
  - easier dataset loading, viewing, saving
  - integrated editor with syntax support
  - command copying and transfer
  - object storing and browsing
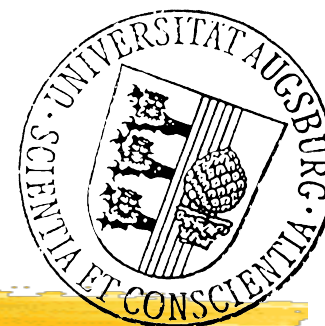 - more direct access to help

    … a form of secretarial support

JGR++?

- Additional requirements
  - more consistency within R
  - better window control
  - "Undo" and return to previous states
  - object history visualization
  - command overviews and thesaurus
  - parallel, multiple analyses

Software

- Mondrian (Martin Theus)

  – interactive graphics

  – crossplatform, links to R via Rserve

- JGR

  – GUI for R (available as R package)

  www.stats.math.uni-augsburg.de

What are the alternatives?

- General language with
  - an unencumbered interface (if we knew how)
  - empathetic abilities (to counteract our errors)
  - a productive blend of approximation/precision
    … probably impossible for now
- Many different tools (like in a kitchen?)