Applied statistics and statistical software

Lasse Engbo Christiansen

DTU Applied Mathematics and Computer Science
Technical University of Denmark

January 5, 2015

Outline of the lecture

- Course overview
- Course contents
- Introduction to R
- Hands-on exercises

Organization of the course

Week 1

- Lectures on topics in applied statistics
- Lectures on applied statistics using R
- Hands-on exercises
- Introduction to case studies

Weeks 2 and 3

- Statistical analysis – project work - report
- Oral presentation of project work (Last 2 days)

Week 1

Every day we'll cycle through

- Introduction to applied statistics using R.
- Short lectures on different statistical methods of data analysis will be given,
- followed by hands-on exercises by the course participants.
- The solution of the hands-on exercises will then be discussed. The course participants are required to participate in this discussion by showing how they have chosen to solve the exercise.

Week 2 and 3

- Project work in small groups (2-3 students).
- The course participants will be given 2 cases (dataset + description of problem), provided by Danish companies.
- The cases should then be analyzed using statistical methods and R.
- Write a report, describing and discussing the results of the analysis.
- At the end of the course an oral presentation of the work must be given. And constructive feedback given to an other group.

The course homepage

The course homepage:

`http://www.compute.dtu.dk/courses/02441`

contains material such as:

- data/cases for week 1
- slides for week 1
- links/material for R

CampusNet will be used for messages etc.

Methods covered in the course

- Descriptive statistics — Day 1
- Comparing treatment means (t-test and non-parametric tests) — Day 1
- Multiple regression analysis — Day 2
- Analysis of variance — Day 3
- Analysis of proportions and counts — Day 4
- The general linear model — Day 5

General Types of Statistical Procedures

- Descriptive Statistics - concerned with
  - presentation and
  - display of data
- Inferential Statistics - concerned with
  - collection of data
  - to draw general conclusions
  - by applying appropriate statistical models

Statistics is concerned with analysis of Data

- Data is a set of *observations*, where each observation contains values of an arbitrary number of *variables*

Gender	Age	Noise	Hearing level
Male	59.3	45	I
Female	67.2	55	III
Female	64.8	55	IV
...	...	...	...

- Data can be recorded in different ways, e.g. by
  Measuring,
  counting,
  sorting etc.

Data is most conveniently analyzed by using statistical software ...

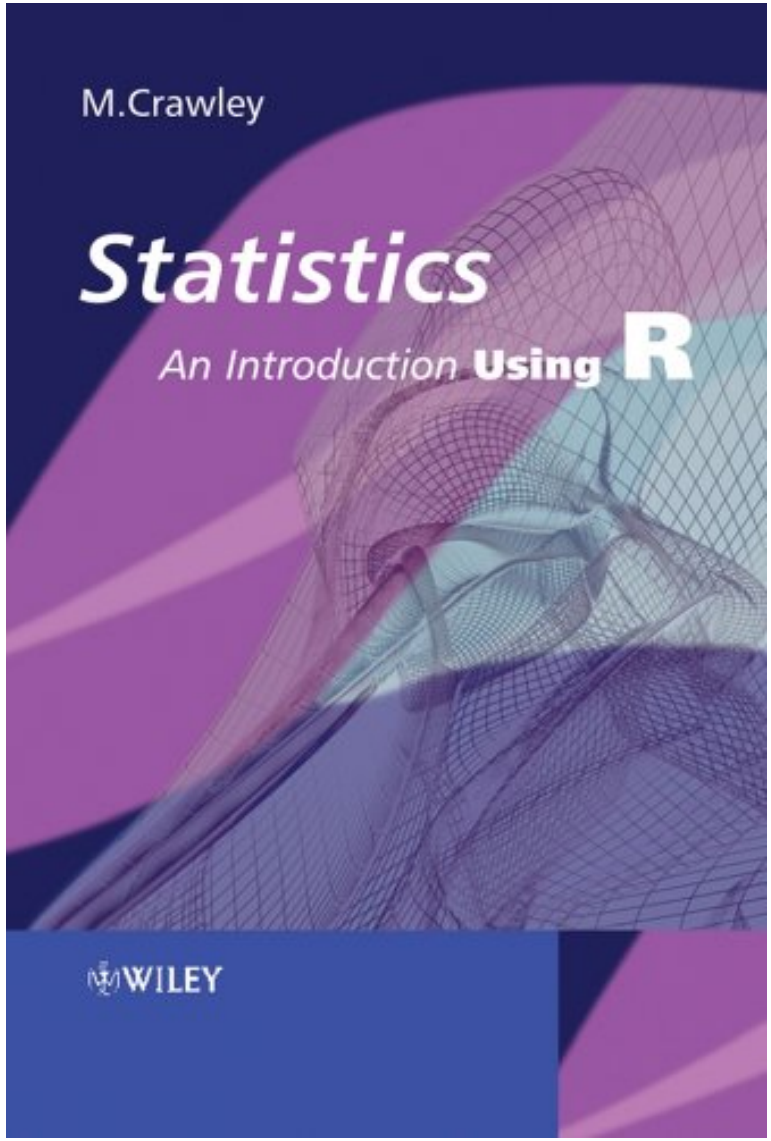
Introduction to R

- R is freely available
- R is one of the most widespread and popular software tools for statistical analysis
- R runs on most platforms
- R has several interfaces (We'll use Rstudio)

`http://www.r-project.org`

`http://www.rstudio.com`

Statistics: An introduction Using R



`http://findit.dtu.dk`

Getting started

- Starting R
- Getting data into R
- Some basic commands
- Understanding objects and statistical output
- Writing script files
- Saving your work and quitting R

Hands-on exercise 1: Descriptive statistics using R

After Exercise 1 you should be able to

- Import data to R
- Perform descriptive statistics in terms of summary statistics (mean, variance, tables etc) as well as figures
- Present your work in a report (suitable for other to read and understand)

Comparing treatment means

- A common problem in statistics concerns the comparison of treatment means
- The t-test can be applied to compare (1 or 2) treatment means
- The t-test assumes that data in each group can be regarded as being normally distributed with the same variance.

$$X_{i,j} \sim N\left(\mu_j, \sigma^2\right)$$

Comparing treatment means

- Comparing 2 means corresponds to testing a statistical hypothesis:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

- The p-value of the analysis is used as a criterion in order to reject the null hypothesis

Based on the p-value we will either
fail to reject H_0 or reject H_0

Checking if data is normally distributed

- The t-test assumes that data in each group can be regarded as being normally distributed
- The assumption about normality can be checked using graphical methods, e.g. histogram or Q-Q plot
- Statistical tests can be applied

Example: comparing 2 means

- In an 1898 biology lecture, Hermon Bumpus reminded his audience of natural selection. As evidence, he presented an example of house sparrows brought to Brown University after a severe winter storm. Some of these birds had survived and some had perished. Could this be because of characteristics?
- The humerus (arm bone) lengths for 24 adult male sparrows that perished and for the 35 adult males that survived were measured.
- Do humerus lengths tend to be different for survivors than for those that perished?
- If so, how large is the difference?

If data is not normally distributed?

- Transform the data so that normality can be assumed.
- Use a test that doesn't require normally distributed data.
- E.g.:
  - Wilcoxon Rank Sum Test
  - Kolmogorov-Smirnov Test

Example: Looking at assumptions

- In a controlled experiment, seeding of clouds were tested for possible effect on rainfall
- Graphical comparisons are important in order to check for assumptions about normality and detection of outliers