Teaching R and statistics to higher degree research students and industry professionals

Alethea Rea

10/07/2019

University of Western Australia

The University of Western Australia (UWA):

- is a public research university in the Australian state of Western Australia
- has its main campus is in Perth, the state capital
- has total student body of ~24,000 students with ~4,000 higher degree research (HDR) students
- has ~1,400 academic staff and ~1,100 general staff (full time equavilent)

Centre for Applied Statistics

Our three main roles are to:

- to develop and hold short courses for students, staff and industry partners
- to provide support and education to postgraduates
- provide consultancy services to the University community and outside organisations and companies

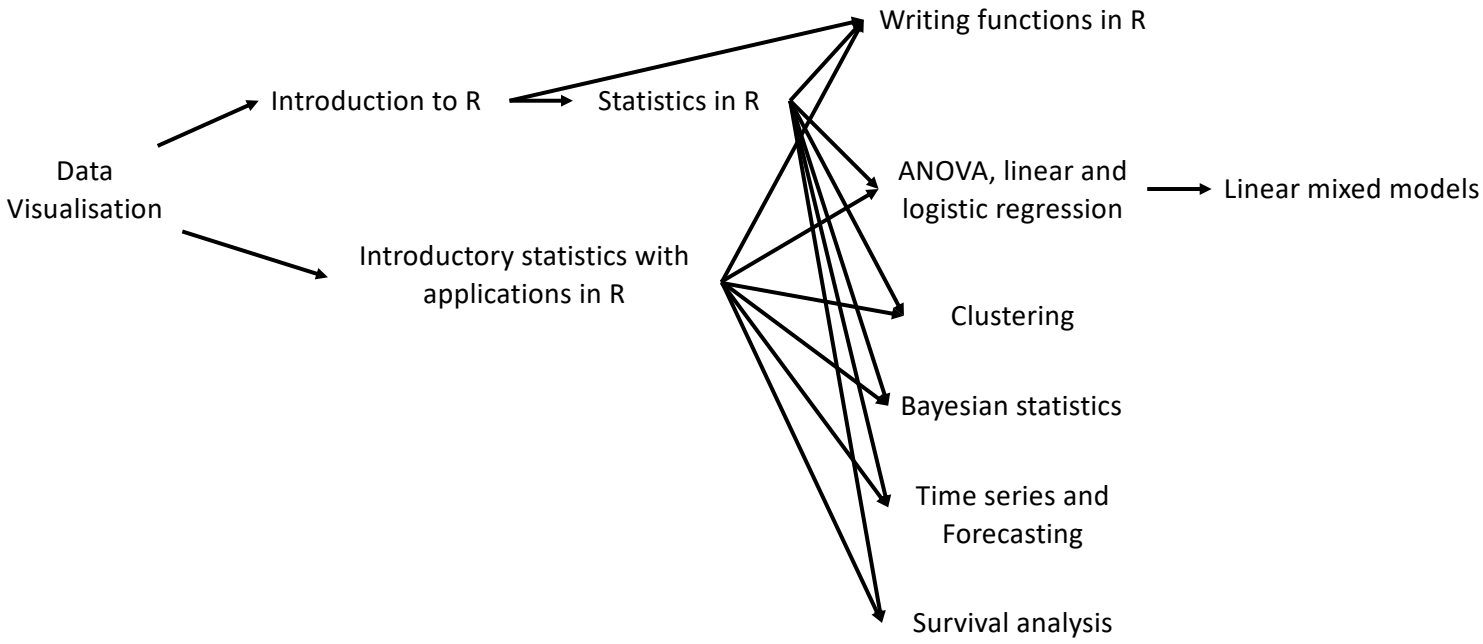

Figure 1: Course overview

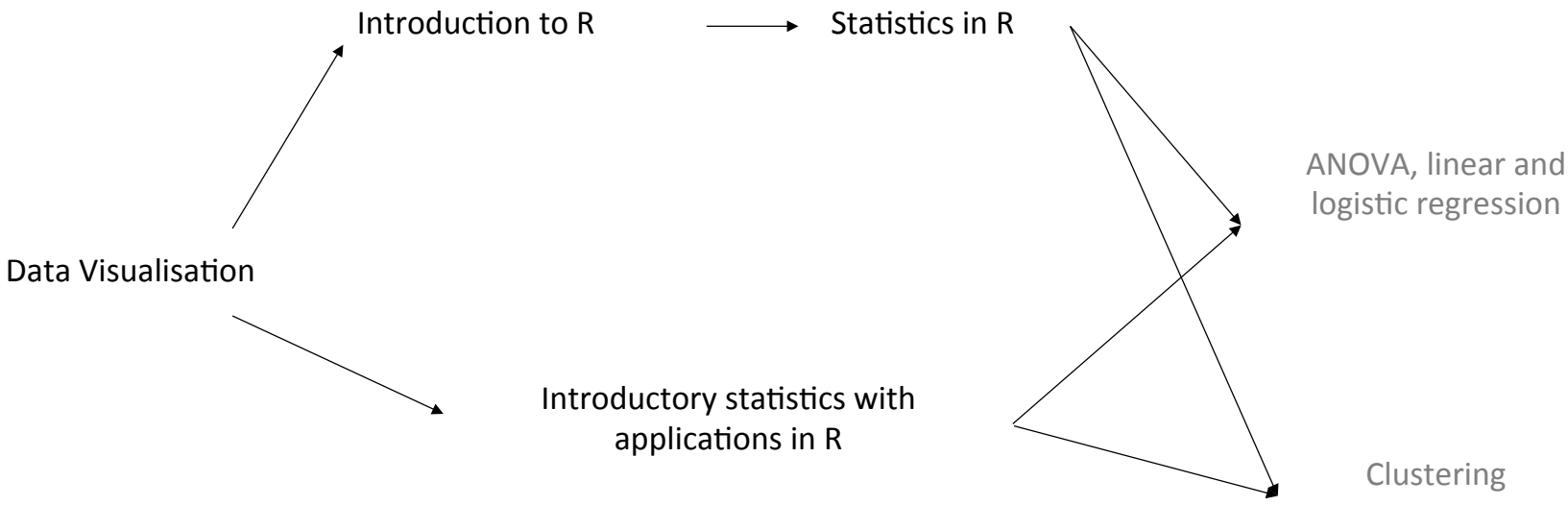

Figure 2: Courses that exist

Who comes?


Intro To R (Feb)
Stat in R (Feb)
Intro To R (Apr)
Data Vis (Jun)
Intro To R (Jun)
Stat in R (Jun)
Intro Stat (Jun)
Participants
15
10
5
0
HDR
UWA
Other

Data Visualisation

This course covers:

- Presenting data for a single variable: Including an introduction to histograms, box plots, and bar graphs
- Visualisation of two or more variables: Including an introduction to scatterplots, pairs plots, parallel coordinate plots and variable-width stack bar charts
- Other plots and maps: Including a brief introduction to plots for time series, bubble plots and more
- Data Ink: Essential parts of a graphic, Tufte's Data-Ink ratio and how to increase it
- Colour and perception: Colour palettes, preattentive features

Target audience: People beginning to work with data who want to understand the basics and how to go about visualising data

Use of R: None, the course is 'theory'

Length: One day

Introduction to R: Overview

This course covers:

- Introduction to R and RStudio: Installing R, RStudio and the tidyverse library
- Importing data: Importing the data and checking the quality
- Data manipulation: Basic data cleaning
- Data summaries: An introduction to creating summaries of the data
- Data visualisation: An introduction to creating visualisations using ggplot
- R markdown: The basics of creating reproducible reports in HTML

Target audience: Anyone wanting to learn the basics of working with data in R

Use of R: Entry level, no assumed knowledge

Length: One day or four evenings

Introduction to R: Base R compared to tidyverse

One thing that works well regardless of the approach is using Rstudio because:

- syntax highlighting: helps the instructors see where students have gone wrong
- seeing 'everything': can see code and console, plots appear beside code
- setting the working directory: can be done via the Session menu

Introduction to R: Base R compared to tidyverse (cont)

Advantages of base R:

- No package installations

Disadvantages of base R:

- base R graphics are increasingly rarely used, and not of a high standard

Introduction to R: Base R compared to tidyverse (cont)

Advantages of tidyverse:

- the quality of visuals
- ease of making custom summarises
- R Markdown

Disadvantages of tidyverse:

- the pipe seems unfamilar at first
- packages need to be installed
- teaching the conversion from wide to long format can be tricky

Statistics in R: Overview

This course covers how to run the following in R:

- T-tests: one and two sample t-tests
- One-way ANOVA
- Chi-squared tests
- Linear regression: Simple linear regression and multiple regression with interaction

Target audience: People who already know these statistical methods and just want to learn how to run them in R

Use of R: Nearly intermediate (one day course Introduction to R is sufficient)

Length: One day

Statistics in R: Base R compared to tidyverse (cont)

There is not a lot of difference between the two approaches but:

- with tidyverse it is easier to make custom summarises
- visuals are mixed even if teaching primarily tidyverse (for example I still prefer the four plot base R summary for looking at linear model diagnostics)

Introductory Statistics with Applications in R: Overview

The overview for this course is:

- Day 1: Introduction to data exploration; sampling and inference
- Day 2: Introduction to research design and comparison of means; sampling distributions; one-way ANOVA
- Day 3: Simple linear regression; comparison and analysis of proportions and odds; contingency tables

Target audience: People who want to learn the basis of statistics or have a refresher

Use of R: Secondary to learning statistics

Length: Three days

ANOVA, Linear Regression and Logistic Regression: Overview

The overview for this course is:

- Day 1: Simple linear regression, multiple regression, polynomial regression, regression model diagnostics, model selection.
- Day 2: One way ANOVA, blocking, simple interactions, more complex interactions, analysis of covariance, ANOVA model diagnostics.
- Day 3: Introduction to logistic regression, odds and risk ratios, multiple logistic regression, model building in logistic regression, assessing goodness of fit and model diagnostics, ordinal logistic regression.

Target audience: People who want a follow on from introductory statistics

Use of R: Secondary to learning statistics

Length: Three days

Clustering in R: Overview

Topics covered are likely to include:

- k-means clustering
- Hierarchical clustering
- Principal component analysis
- Multidimensional scaling

Target audience: People who are interested in learning about some basics of clustering

Use of R: Secondary to learning about clustering

Length: One day

Summary

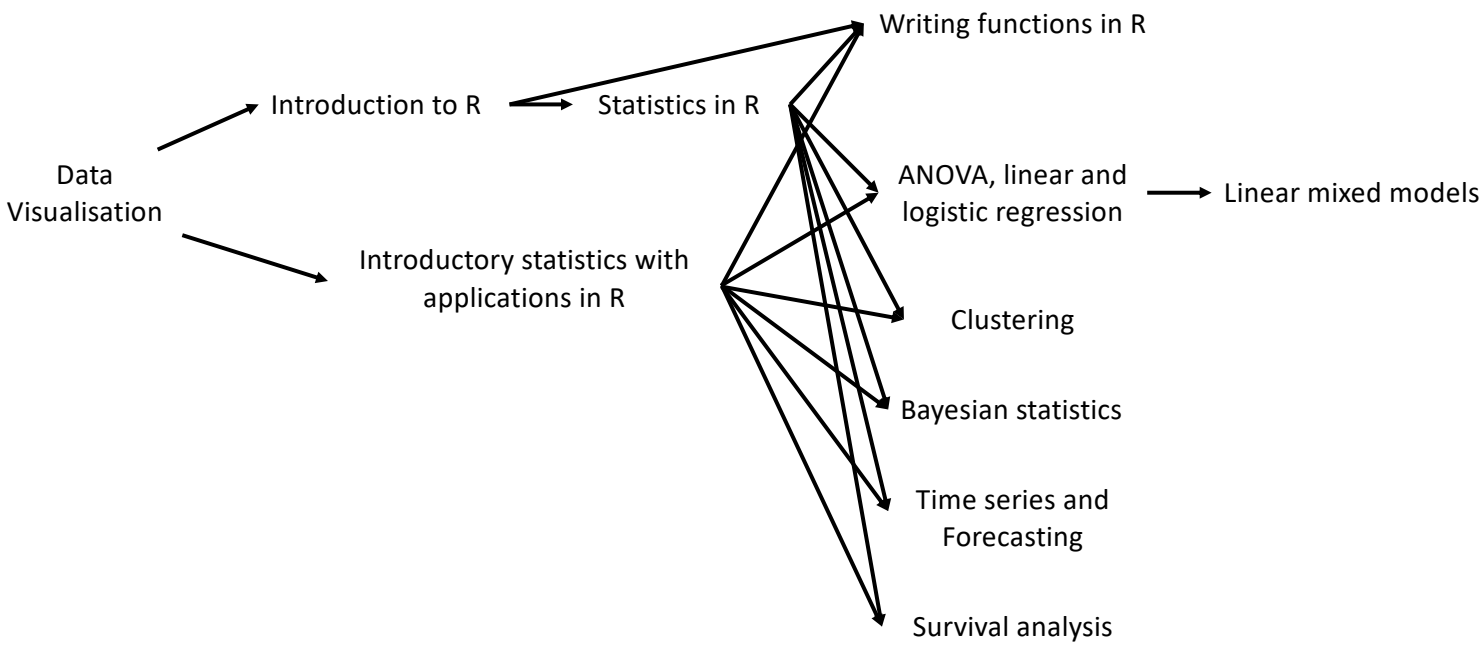

Figure 3: Course overview