

Advanced Data Analysis with R

Statistical reasoning, data preparation and performance awareness

FA-0677 · 2 days

Data & Analytics

Introduction

An intensive two-day course for participants who already program in R. Explore data sourcing and preparation, visual analysis, statistical inference, resampling and introductory Bayesian workflows through selected prepared examples.

The detailed source primarily teaches statistical analysis rather than a full machine-learning algorithm curriculum. Practice prioritises a small reproducible analysis; Bayesian modelling, JAGS and Rcpp/performance techniques are scoped demonstrations or extensions.

Learning outcomes

- Source, inspect, clean and reshape a selected dataset in R.
- Explore univariate and multivariate relationships with suitable summaries and graphics.
- Interpret selected hypothesis tests and confidence intervals with their assumptions.
- Explain bootstrap and introductory Bayesian approaches and inspect prepared examples.
- Identify missing-data and performance issues and choose reasonable next steps.

Prerequisites

- Fundamentals of R programming and file/data handling.
- Basic probability, statistics and linear mathematics; advanced mathematics is helpful but not required.

- Use the arranged R/package environment; JAGS and a C++ toolchain are needed only for the corresponding demonstrations.

Training outline

1. Day 1 — Data Preparation, Exploration and Statistical Foundations

Sources and Quality

- Relational databases
- Using JSON
- XML
- Other data formats
- Online repositories
- Exercises

Analysis with missing data

- Visualizing missing data
- Types of missing data
- Compare simple missing-data treatments and their assumptions
- Discuss suitable imputation approaches and validation limits
- Exercises

Dealing with Messy Data

- Checking unsanitized data
- Regular expressions
- Other tools for messy data
- Exercises

Compare missingness assumptions and treatment options. For predictive evaluation, fit learned imputation/preprocessing only on training data, not held-out evaluation data.

Summaries and Relationships

The Distribution of Data

- Univariate Data
- Frequency Distribution
- Central Tendency

- Spread
- Probability Distribution
- Visualization
- Exercises

Relationship between data

- Multivariate data
- Relationships between a categorical and continuous variable
- Relationships between two categorical variables
- The relationship between two continuous variables
- Visualization methods
- Exercises

Use dplyr for selected transformations and ggplot2 for suitable graphics; distinguish association from causation.

Probability and Estimation

- Basic probability
- Sampling from distributions
- The normal distribution
- Exercises
- Using Data To Reason About The World
- Estimating means
- The sampling distribution
- Interval estimation
- Smaller samples
- Exercises

Interpret confidence intervals as a procedure with repeated-sampling coverage, not a probability that a fixed parameter lies in one computed interval.

2. Day 2 – Inference, Resampling and Advanced Demonstrations

Hypothesis Testing

- The null hypothesis significance testing framework
- Testing the mean of one sample

- Testing two means
- Testing more than two means
- Testing independence of proportions
- What if my assumptions are unfounded?
- Exercises

Check assumptions and distinguish statistical significance, effect size and practical importance; a p-value is not the probability that the null hypothesis is true.

Bootstrap and Bayesian Examples

- Performing the bootstrap in R
- Confidence intervals
- A one-sample test of means
- Bootstrapping statistics other than the mean
- Exercises
- The big idea behind Bayesian analysis
- Choosing a prior
- Using MCMC
- Using JAGS and runjags
- Fitting distributions the Bayesian way
- The Bayesian independent samples t-test
- Exercises

Inspect one prepared model and sampling diagnostics; priors, model assumptions and convergence affect interpretation.

Larger-data and Performance Choices

- Wait to optimize
- Using a bigger and faster machine
- Be smart about your code
- Using optimized packages
- Evaluate compatible alternative implementations or data backends where justified
- Using parallelization
- Using Rcpp
- Being smarter about your code
- Exercises

Profile a bottleneck before changing hardware or implementation. Parallelism and Rcpp introduce setup and coordination costs and do not guarantee a speedup.

Analysis Review

Summarise findings, assumptions, uncertainty, reproducibility and limits of the selected exercise.

© 2026 Faysal Aziz · Kuala Lumpur

<https://faysalaziz.com/courses/advanced-data-analysis-with-r/>