

SHORT COURSE at the Student Conference on Conservation Science

Department of Zoology, University of Cambridge

26-27 March 2012

Getting started in R to analyse biological field data

Course outline:

This course is designed to continue general skills training in the analysis and appraisal of data. The course will allow practice of existing statistical skills (and like playing the piano – without practice, stats knowledge disappears), hopefully teach some new ones and most importantly provide an introduction to code based statistical analysis in R. This is one of the most efficient ways to analyse data. By the end of this course students should be a bit better at basic statistical analysis, and will know how to start analysing data efficiently and within an environment that provides the framework for any type of analysis in the future.

General aims of the course:

To start you off or make you more comfortable with General Linear Modelling in an R environment.

The course will consist of two days of teaching:

We will start immediately by importing a data file into R and introducing the main features of R. At all times you will be at your computer able to carry out the analysis we are all talking about to get direct practical experience of data analysis. The emphasis is on being able to do proper and efficient analysis – we will have some theory to support this – but the course is all about enabling you to get started on analysis immediately. Good biological statisticians get to be so by analysing lots of real data and to get this practice you have to be able to get started.

The first day will be an introduction to cover the ideas behind modelling and data analysis in biology. This will be in a practical example framework with an accompanying data file and R script file so you can follow the analysis yourself on your computer. We will cover/review:

1. Sample sizes and statistical power
 - a. What is a sampling unit
 - b. Repeated measures and pseudoreplication
 - c. Power
 - d. Fundamentally flawed studies
 - e. Experimental design
2. General linear modelling as good general approach
 - a. Data forms
 - b. The checklist for General Linear Modelling
 - i. What is your y variable?
 - ii. What is the distribution of your y variable?
 - iii. What is your x variable?
 - iv. Is your x variable a factor or a continuous variable (also called a covariate)?

- v. If x is a covariate does it have a linear relationship with y?
 - vi. What interactions should you consider
- 3. Dependent variables
 - a. Data distributions
- 4. Categorical and continuous data
 - a. The two types of data
 - b. Examples of Categorical or Factors
 - c. Examples of Continuous or Covariates
 - d. The blurring of the two categories
- 5. Bar charts and means
 - a. Bar charts as illustrations of factors
 - b. Box plots
- 6. Scatter plots and lines
 - a. How general linear modelling works - residuals
 - b. Scatter plots as illustrations of relationships
 - c. Linear relationships
 - d. Other relationships (quadratic and exponential)
- 7. Specifying models
 - a. Format in R
- 8. Running simple GLM with a factor and then a covariate and then with both variables
 - a. ANOVA tables –statistical significance
 - b. Summary tables – biological significance
 - c. Interpreting R output tables
 - d. Turning parameter estimates back into understandable biology
 - i. Factors to bar charts
 - ii. Covariates to scatter plots
 - e. Illustrating your results
 - f. Making predictions from your models
 - g. Checking model assumptions
- 9. Comparing and evaluating models
 - a. AIC
 - b. R^2
- 10. Interactions
 - a. Interactions between factors
 - b. Interactions involving covariates
 - c. Treatment of interactions in R
 - d. Getting the biology back out of interactions
- 11. Summary of what has been covered during the day

The second day will consist of two workshops that will explore two different data sets to allow you to practice General Linear Modelling. We will practice some common procedures in R in both workshops:

1. Importing data
2. Inspecting data
3. Computing new variables
4. Conditional computing of new variables
5. Selecting cases
6. Basic descriptive statistics
7. Plotting data
8. Basic GLM analyses to test hypotheses
9. Evaluating the importance and meaning of interactions

Workshop 1 analyses a dataset of counts of two monkey species in undisturbed and logged forest at different altitudes on a mountain. We test the hypothesis that abundance is affected

by habitat quality and altitude and whether these effects operate in the same way for both monkey species.

Workshop 2 analyses a real dataset of attack success of two species of birds of prey, peregrines and sparrowhawks, attacking a species of shorebird, the redshank, in order to examine if the two predators are affected by flock size in the same way, and so how selection for group size in redshank might depend on the relative abundance of the two predators at a site.

We repeat the same kinds of data exploration and data analysis in both workshops to give you more confidence in General Linear Modelling. The workshops shows how similar approaches in analysis are used for quite different data sets.

If we have time we will finish by outlining some of the potential within R to analyse other types of data sets. In particular, presence/absence data in logistic regression and mixed models.