Introduction to Scientific Computing for Ecologists I

John M. Drake & Andrew Park

Introduction

Scientific programming is the development of special purpose computer programs for science. Scientific programming may be used to accomplish a wide range of scientific tasks including *numerical analysis*, *data management*, *data analysis*, and *visualization*. Scientific programming is an essential skill for research. This course, *Population & Community Ecology* (ECOL 4000/6000), will use the programming language R both to learn basic scientific programming techniques, perform data analysis, and understand ecological theory.

R is the name of the programming language as well as the basic software needed to create programs in that language. More information about R is available at https://www.r-project.org/. To facilitate working in R, we will use the RStudio *Integrated Development Environment*. More information about RStudio is available at https://www.rstudio.com/. RStudio will also allow us to produce *statistically literate programs* using the RMarkdown language for integrating R code and commentary in ordinary langauge. More information about RMarkdown is available at http://rmarkdown.rstudio.com/.

Today's demonstration will introduce:

- The three R's: R itself, RStudio, and RMarkdown
- How to create a new project
- Statistically literate programming: the idea of *code chunks*
- How to read data into R from a comma separated file
- Simple data summaries
- Data visualization

Three R's

Key concepts

- There is a native R user-interface, but we never use it
- Navigating RStudio: *console*, *editor*, *workspace*, *browser* (for figures, files, help, *etc.*)

How to create a new project

- Working directory
- The parts of a new RMarkdown document
  - Header
  - Words
  - Code
  - Comments

Example

Forest regeneration in degraded landscapes depends on the dispersal of seeds from intact forest fragments. As part of a study of plant succession in abandonded agricultural lands in Kibale National Park, Uganda, (R. S. Duncan and Duncan 2000) performed an experiment to measure the predation of seeds in grassy areas at distances of 10 and 25 meters from a forest fragment. Seeds of several species were placed at stations (a 10cm by 10cm tray made of wire screen) at regular intervals the relevant distance from the forest fragment and checked after one day, four days and then weekly (until the end of the study). Seeds that disappeared from the trays were presumed to have been predated by rodents, which results in seed death. Interpreted properly, these data provide useful information about the limitations to seedling recruitment in this ecosystem. We will use R to ask some questions about this data set and produce a basic report. These data are also discussed in the textbook *Ecological Models and Data in R* (Bolker 2008), where they are used to illustrate basic statistical concepts.

The data are contained in the csv file `seeds.csv`. We can view this file using any spreadsheet or text editing program. However, to analyze the data in R we will need to load the data from with RStudio. First we will work in *interactive mode*. We can use the function `read.csv`. Notice that we have to store our data in an R object using the *assignment operator* `<-`. We have now have several ways we can view these data (*e.g.*, print to screen, using the function `head`, using the RStudio environment browser).

One of the first questions we might ask of a data set is what variables are there? These data consist of nine variables.

1. `station`: station number where the observation was made
2. `dist`: distance from forest edge in meters
3. `species`: plant species designation
4. `date`: sample date
5. `seeds`: number of seeds present at observation
6. `tcum`: cumulative time elapsed since the seeds were initially placed
7. `tint`: time elapsed since the last observation
8. `taken`: seeds removed since the last observation
9. `available`: number of seeds available to be taken at the start of the observation interval

Given this information, there are a number of questions we might ask:

- How many records are there in the data set?
- How many different distances were tested?
- How many species were studied?
- What was the longest time elapsed?
- What time of year was the study conducted?
- How many seeds were included in the study altogether?

What other questions might we ask of these data? What would an answer look like? Together we can create an RMarkdown report to examing some of these questions.

References

Bolker, B.M. 2008. *Ecological Models and Data in R.* Princeton University Press.

Duncan, R. Scot, and Virginia E. Duncan. 2000. "Forest Succession and Distance from Forest Edge in an Afro-Tropical Grassland1." *Biotropica* 32 (1). Blackwell Publishing Ltd: 33–41. doi:10.1111/j.1744-7429.2000.tb00445.x.