Programming

John M. Drake & Ana I. Bento

Learning outcomes

1. Writing scripts
2. Writing functions
3. Writing loops

Introduction

This exercise is about writing *scripts*, purpose-built computer programs for performing some analysis. Our scripts will be written using the *RStudio Editor* and compiled using Rstudio. In this exercise we review many basic numerical operations and R functions, programming style, the development of custom functions, pipes, flow of control and loops.

Data

West Nile virus (WNV) is a positive-sense single-stranded RNA virus transmitted by mosquitoes to a range of vertebrate hosts. WNV was first identified in Uganda in 1937 and found in parts of Europe, Asia, and Australia during the 1950s and 1960s. In 1999, WNV was first reported in the Americas in association with dieoffs of captive and wild birds. This outbreak initiated widespread epidemic that swept across North America and is now spreading in Central and South America. Humans are "dead end" hosts (humans do not achieve sufficiently high viremia to be infectious to mosquitoes). The majority of human cases are asymptomatic, but a small fraction of cases result in meningitis, encephalitis, and/or death. State-level data on the number of reported cases, meningitis/encephalitis, and fatalities are compiled and reported by the CDC and US Geological survey at https://diseasemaps.usgs.gov/. The file `wnv.csv` contains tabular data on the number of reported cases (mostly febrile cases), neuroinvasive cases (meningitis/encephalitis), and fatalities for all continental US states from 1999-2007. Additional data are the latitude and longitude of the centroid of each state.

Scripts

Exercise. Write a script to load the West Nile virus data and use ggplot to create a histogram for the total number of cases in each state in each year. Follow the format of the *prototypical script* advocated in the presentation: Header, Load Packages, Declare Functions, Load Data, Perform Analysis.

With each of the following exercises, extend your script so that at the end of the unit you have one script that performs the entire analysis.

Exercise. The state-level and case burden is evidently highly skewed. Plot a histogram for the logarithm of the number of cases. Do this two different ways.

Exercise. Use arithmetic operators to calculate the raw case fatality rate (CFR) in each state in each year. Plot a histogram of the calcated CFRs.

Exercise. Use arithmetic operators, logical operators, and the function `sum` to verify that the variable `Total` is simply the sum of the number of febrile cases, neuroinvasive cases, and other cases.

Exercise. Use modular arithmetic to provide an annual case count for each state rounded (down) to the nearest dozen. Use modular arithmetic to extract the rounding errors associated with this calculate, then add the errors to obtain the total error.

Functions

Let us call the ratio of meningitis/encephalitis cases to the total number of cases the *neuroinvasive disease rate*.

Exercise. Write a function to calculate the mean and standard error (standard deviation divided by the square root of the sample size) of the neuroinvasive disease rate for all the states in a given list and given set of years. Follow the Google R style and remember to place the function near the top of your script. Use your function to calculate the average severe disease rate in California, Colorado, and New York.

Exercise. Use ggplot to show the neurovinvasive disease rate for these states as a bar graph with error bars to show the standard deviation.

Exercise. Use your function and ggplot to show the neurovinvasive disease rate for all states.

Pipes

Exercise. Use pipes to produce the same plots without using your function.

Control of flow

Conditional execution. There is evidence that the WNV case fatality rate differs between the Eastern part of the Unoted States and the Western part, probably due to a difference in the mosquito vector species responsible for transmission in different parts of the country. We can use our variable `Longitude` (the longitude of the state centroid) to study this pattern.

Exercise. Choose a longitude to designate the "center" of the country. Use the function `ifelse` to assign each state to an "Eastern" region or a "Western" region.

Exercise. Analyse your data to compare case fatality rates in the Eastern vs. Weestern United States.

Exercise. Is there evidence for a *latitudinal gradient* in case fatality rate?

Loops. One useful task for looping is to select parts of a data set for analysis following an algorith. For instance, analysis can be performed over a series of time points or regions using a script with the following architecture.

```
times <- seq(1:10)

some.algorithm <- function(t){
  y <- t*10  # a silly example
}

output <- c() # Question: What is this line doing?
for(t in times){
  output <- c(output, some.algorithm(t))
}

plot(times, output, type='p')
```

Exercise. Loop over all the years in the WNV data set (1999-2007) and compute the following statistics: Total number of states reporting cases, total number of reported cases, total number of fatalities, and case fatality rate. Produce some plots to explore how these quantities change over and with respect to each other. Explain what you have learned or suspect based on these plots.

Exercise. How does your choice of longitudinal breakpoint matter to the evidence for a geographic difference in case fatality rate? Combine conditional execution and looping to study the difference in case fatality rate over a range of breakpoints. What is the "best" longitude for separating case fatality rate in the East vs. West?

Using help

Using the help functions, i.e., reading the argument list, methods, usage, etc. and being example to apply or extend examples to a new case is an important programming skill (albeit one that doesn't involve any programming!). In this section we practice reading and using R help.

Exercise. We may interpret raw case fatality rate (i.e. ratio of the number of deaths, x, to number of infections, n) as a realization from a binomial process with n trials and x "successes" generated by an unknown rate parameter p. This p may be the quantity truly of interest (for instance, if we wish to ask if the case fatality rate in California is significantly different from the case fatality rate in Colorado. In R, the *estimated rate* and its *confidence interval* can be obtained using the function `prop.test` for testing equal proportions. Use the help to determine the proper usage of `prop.test` and calculate confidence intervals for the case fatality rates in all states for which there have been reported cases of WNV.

Exercise. The "See Also" section of the help for `prop.test` states that a different function might be useful for an exact test of our hypotheses. Use the help to identify what this function is, learn how to use it, and compare the differences.