

Problem: You have some data that was gathered at different sites, as shown in the picture above. You want to know if the values at different sites are significantly different, but you don’t know what kind of stats model to run on these data.

Goal: Write a custom R function that will *bootstrap* these data and produce a P value based on the results. (You may want to consider coming back to this exercise if you’re unfamiliar with bootstrapping).

New Functions to know:

1. lm() & anova(). These run linear models and ANOVAs on data. lm() uses “formula notation” (see below).

2. hist(). Creates a histogram of data.

Steps to take:

Because this isn’t supposed to be a statistics workshop, we’ve written this function for you! However, we’ve left blanks in the code that we feel you should be able to fill in yourself. Fill in the blanks, then test to see if your function works. We’ve annotated each line of the function so you have some idea of what we’re looking for.

Code:

boot.strap.fnc = function(\_\_\_, \_\_\_, num.Reps=999) { #We'll need X and Y data, plus a number of repetitions to run for.

\_\_\_\_ = rep(0, times = num.Reps) #Create an empty vector to store our F statistics in.

lm.actual = lm(Y.vec ~ X.vec) #Run an ANOVA on our raw data.

f.stat.actual = anova(lm.actual)$"F value"[1] #Extract the F stat from that run.

for (\_\_\_\_ in \_\_\_\_) { #For each repetition we do...

X.vec.temp = \_\_\_\_\_ #Shuffle the X.vector column and save that shuffled version in a new object.

lm.temp = lm(\_\_\_\_\_\_\_\_\_\_) #Run an ANOVA on this shuffled-up data.

f.stats[\_\_\_\_] = anova(\_\_\_\_)\_\_\_\_\_ #Take the F-stat of that ANOVA run and save it.

} #Close for loop

hist(f.stats) #Plot a histogram of all of our F-stats

p.value = 1 - (length(which(f.stats <= f.stat.actual))/num.Reps) #Calculate what prop. of F.stats are above our real one, which is equivalent to a P value for a bootstrap.

return(list(f.stat.actual = f.stat.actual, p.value = p.value)) #Tell us our real F.stat and p-value.

}

Function testing:

Use the data sheet “bootstrap.data.csv” to test your function. Save the data sheet to your working directory (ask for help if you don’t know how to do this). Then, load the data sheet by running this code:

data1 = as.matrix(read.csv(“bootstrap.data.csv”))

You can view your data sheet before you run your function on it by using View()

View(data1)