

7. Control Structures

Z620: Quantitative Biodiversity, Indiana University

OVERVIEW

In Dante's satire of hell each layer contains a new iterative task for the tortured to endure for eternity. Much of the power of computing stems from the efficiency of computers in performing small tasks in an iterative manner. This lesson introduces the concepts of control structures and random numbers. We will use these tools to understand the diversity of jellybeans flavors, but these basic tools are useful for a numbers of tasks in computing and will help you automate many of your future work flows.

1) INTRODUCTION TO CONTROL STRUCTURES

Control structures allow you to control the flow of execution of a script. Today, we will focus on the "for loop", but other common control structures include: "if, else" "while" and "next" Have a look at `?Control` for documentation.

A quick tutorial of common control structures in R can be found at <https://ramnathv.github.io/pycon2014-r/learn/controls.html>

```
rm(list = ls()) # clear working directory
```

A. 'If' and 'Else' Logical Tests

Short intro: when and why to use 1. The general syntax for an "if, else" control stucture in the R language.

```
if (condition) {  
  # do something  
} else {  
  # do something else  
}
```

2. An example of the an "if, else" control stucture.

```
x <- 1:15  
if (sample(x, 1) <= 10) {  
  print(paste("x is",x ,"and is less than 10"))  
} else {  
  print(paste("x is",x,"and is greater than 10"))  
}
```

3. "if" accepts logical values; can accept numbers but not strings. 4. "if" can be complemented by an "else"

5. ifelse as shorthand 6. if-else can be strung to create hirearchy of tests

B. For Loops

Let's introduce the "for loop" as a control structure.

Syntax The specific syntax of for loops will look different in other language but basics syntax remains the same. 1. Start a loop with for 2. Declare variable and number of iteration 3. Tasks to perform in code block

1. 2. 3.

In R this looks like: `for(i in 1:10){ print(i) }`

```
for(i in 1:10){  
  print(i)  
}
```

In this structure `i` is a variable that will hold each integer from 1 to 10 during each successive pass through the loop.

You can use this structure to move through each item of vector and perform any number of tasks to each element.

Like printing out a list of fruits.

```
x <- c("apples", "oranges", "bananas", "strawberries")

for (i in seq(x)) {
  print(x[i])
}
```

```
## [1] "apples"
## [1] "oranges"
## [1] "bananas"
## [1] "strawberries"
```

Note that a for loop can iterate over any type of vector: logical, numerical, string. You can also iterate over vectors using indices (R counts from 1 and not from 0 like some other languages)

```
for (i in x) {
  print(i)
}
```

```
## [1] "apples"
## [1] "oranges"
## [1] "bananas"
## [1] "strawberries"
```

Or get the number of characters in each item of the same list.

```
for (i in seq(x)) {
  len <- nchar(x[i])
  print(len)
}
```

```
## [1] 6
## [1] 7
## [1] 7
## [1] 12
```

Note on the use of for loops When possible try to avoid them. They take a long time to execute and are somewhat clumsy. Consider using apply type functions.

The `sapply` function can be used to get the same results as the loop above. Check out `?apply` and `?lapply`.

```
sapply(x, nchar)
```

```
##      apples      oranges      bananas strawberries
##          6           7           7             12
```

C. Combining for and if

e.g. print only numbers that can be divided by 3

```
for(i in 1:10){
  if(i %% 3 == 0){
    print(i)
  }
}
```

```
}
}
```

```
## [1] 3
## [1] 6
## [1] 9
```

D. While loops

Another loop control structure is the “while loop”.

An example of the while loop.

```
i <- 1
while (i < 10) {
  print(i)
  i <- i + 1
}
```

```
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
## [1] 9
```

Notice that a while loop needs an specific break point.

What would happen if the code block was changed to subtract 1 from i in each iteration of the loop?

2) Revisiting the jelly-bean community with control structure

The control structures discussed above have unlimited applications in the realm of quantitative methods. In this section you will learn how to apply these structures in conducting the analyses scheme you carried out for the jelly bean community. But before we can do that we need to get familiar with another powerful family of functions in R: built-in probability distributions.

A. Statistical distributions in R

Many groups of quantities can be described as frequency distributions. You are probably familiar with the Normal distribution and its hallmark bell curve. R has built-in functions describing many statistical distributions. You can find them listed in the R Documentation by typing `?Distributions` in the console.

Choose your favourite distribution and look at its specific documentation. Each distribution in R has 4 functions associated with it: **density**, **probability**, **quantiles** and **random sample**. For example the uniform distribution (an equal probability for all values) is called *unif* in R and has the following 4 functions.

```
dunif(x, min = 0, max = 1, log = FALSE)      # Density
punif(q, min = 0, max = 1, lower.tail = TRUE, log.p = FALSE) # probability
qunif(p, min = 0, max = 1, lower.tail = TRUE, log.p = FALSE) # quantile
runif(n, min = 0, max = 1)                   # random sample
```

A quick note on random sampling by computers Computers can't generate truly random number, but rather pseudorandom numbers. This is achieved by a function that generates sequences of numbers that behave

like random numbers, but are completely determined by the functions input known as a *seed*. Put in the same seed twice and you will get the same sequence of numbers.

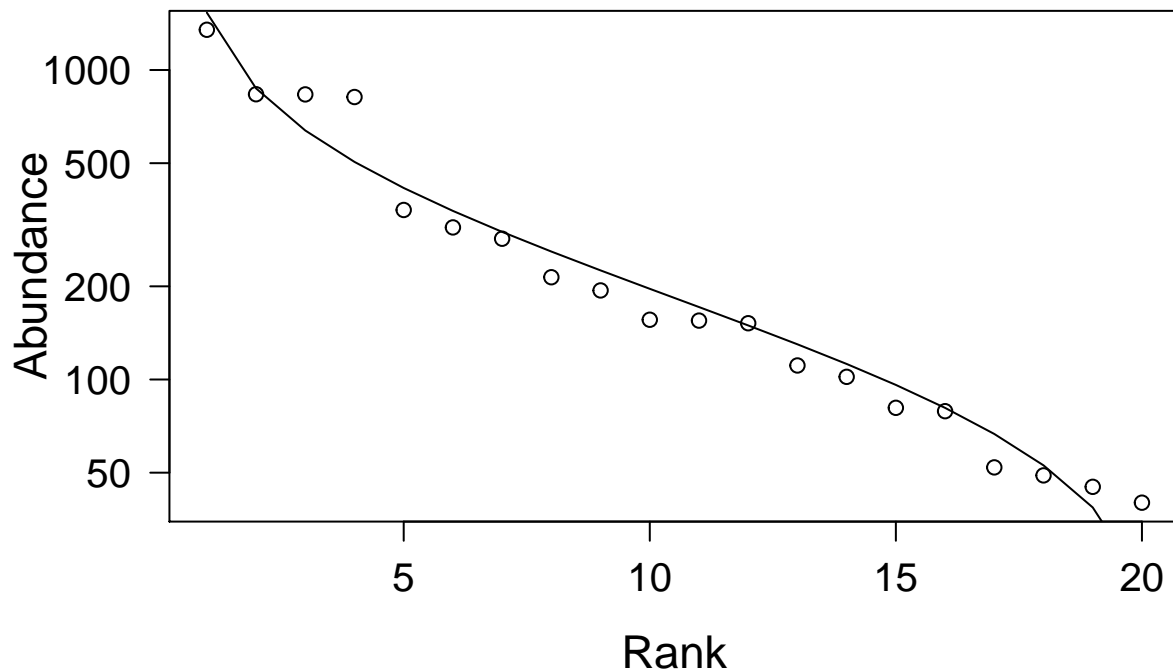
B. generating the source community

In the jelly bean exercise you had a jar of jelly beans simulationg a biological community. To generate our digital community we will recall that the species abundances in biological communities often follows the log-normal distribution. To simulate a community we will draw a seunce of random numbers from the log-normal dittribution. Each of theses numbers will represent the abundance of a single species in our simulated community.

```
# for reproducible results we will set the seed
set.seed(6)
# draw 20 random numbers from log-normal distribution with a mean of 5 (log scale)
jelly.jar <- rlnorm (n = 20, meanlog = 5 )
# force numers to integers
jelly.jar <- as.integer (jelly.jar)
```

Let's plot the rank-abundance curve of our community using the `rad` function from the `vegan` package

```
require (vegan)
RACresults <- rad.lognormal(jelly.jar)
plot(RACresults, las = 1, cex.lab = 1.4, cex.axis = 1.25)
```



B. Sampling the Source Community

Next we will write a function that will draw a random sample from our `jelly.jar` community. Take a moment to break down the process sampling the jelly bean jar into a sequence of actions.

One such sequence may be: 1. draw random jelly bean 2. classify jelly bean color 3. add 1 to list of colors by color sampled 4. repeat steps 1-3 N times, where N is your sample size

This sequence of actions is an algorithm. Here is how it looks like in R code:

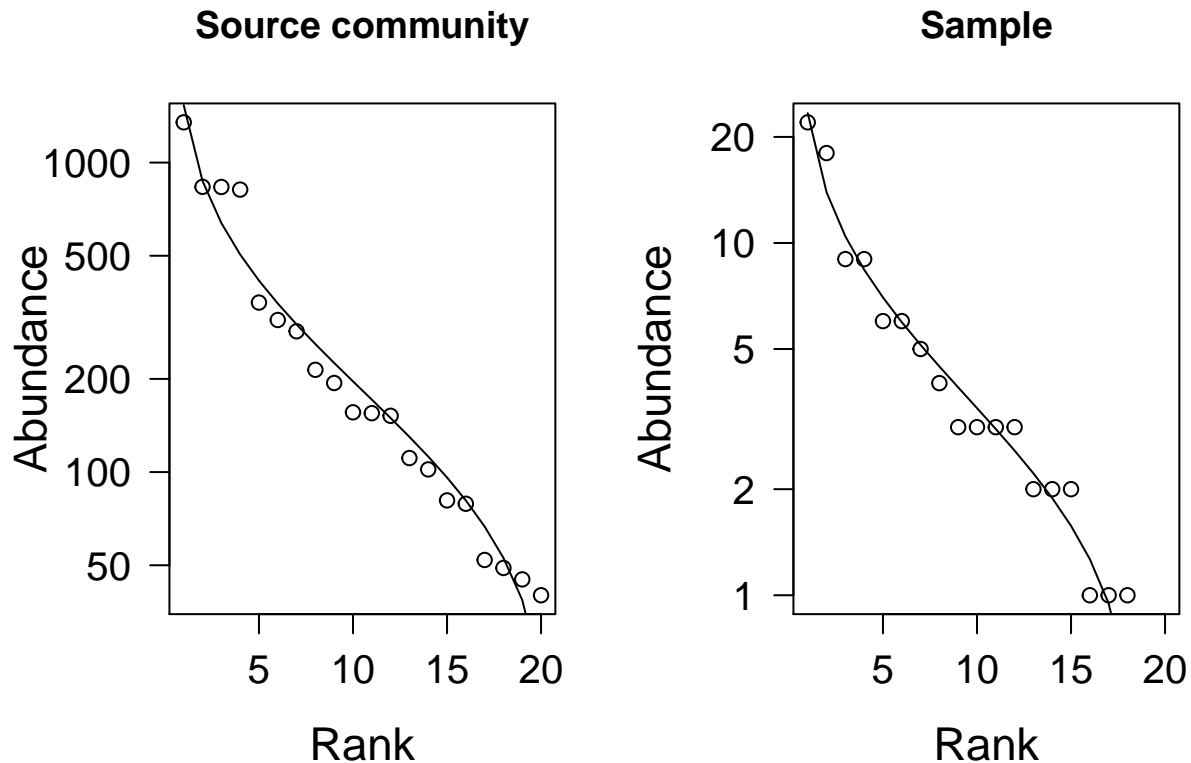
Function to generate a single sample * explain the rationale: sample and the summarize sample * highlight the use of for loop in binning * mention the sample is iterative as well, but built in function ("sample") negates the use of loop.

```
sample.community <- function(x,n){  
  # write out an explicit vector of the community  
  all.individuals <- rep(seq(x), x)  
  # take n samples from the community  
  survey <- sample(all.individuals,n)  
  # prepare vector of species with 0 in each bin  
  survey.sum <- rep(0, length(x))  
  # add 1 to species bin each time an individual of that species was sampled  
  for ( i in survey){  
    survey.sum[i] <- survey.sum[i] + 1  
  }  
  return(survey.sum)  
}
```

Note that in this implementation of the algorithm above we drew multiple samples using the `sample` function. This function iterated over the salping step n times. Next we wrote a for loop to iterate over those samples and bin them into our species list.

Let's compare the rank abundance curves of a sample to our source community

```
# for reproducible results we will set the seed  
set.seed(3)  
# sample using our function  
sample.jelly <- sample.community(x=jelly.jar,n=100)  
# generate RAC using vegan  
RACsample <- rad.lognormal(sample.jelly)  
plot.new()  
par(mfrow=c(1,2))  
plot(RACresults, las = 1, cex.lab = 1.4, cex.axis = 1.25, xlim=c(1,20), main="Source community")  
plot(RACsample, las = 1, cex.lab = 1.4, cex.axis = 1.25, xlim=c(1,20), main = "Sample")
```



How similar or different is the sample from the source community?

C. Estimating Richness in Repeated Samples

We will next use for loops to explore how well richness indices succeed in estimating the true richness of the source community. We will use the function `estimateR` of the `vegan` package to calculate diversity indices.

```
S <- estimateR(sample.jelly)
S

##      S.obs   S.chao1 se.chao1   S.ACE   se.ACE
## 18.000000 18.750000  1.417910 19.680642  1.722197
```

How well these 3 indices estimate the true diversity? Using our simulated community we can test that! We will draw repeated samples from our source community and estimate S for each of them. Lastly we will look at the distribution of the results compared to the true S .

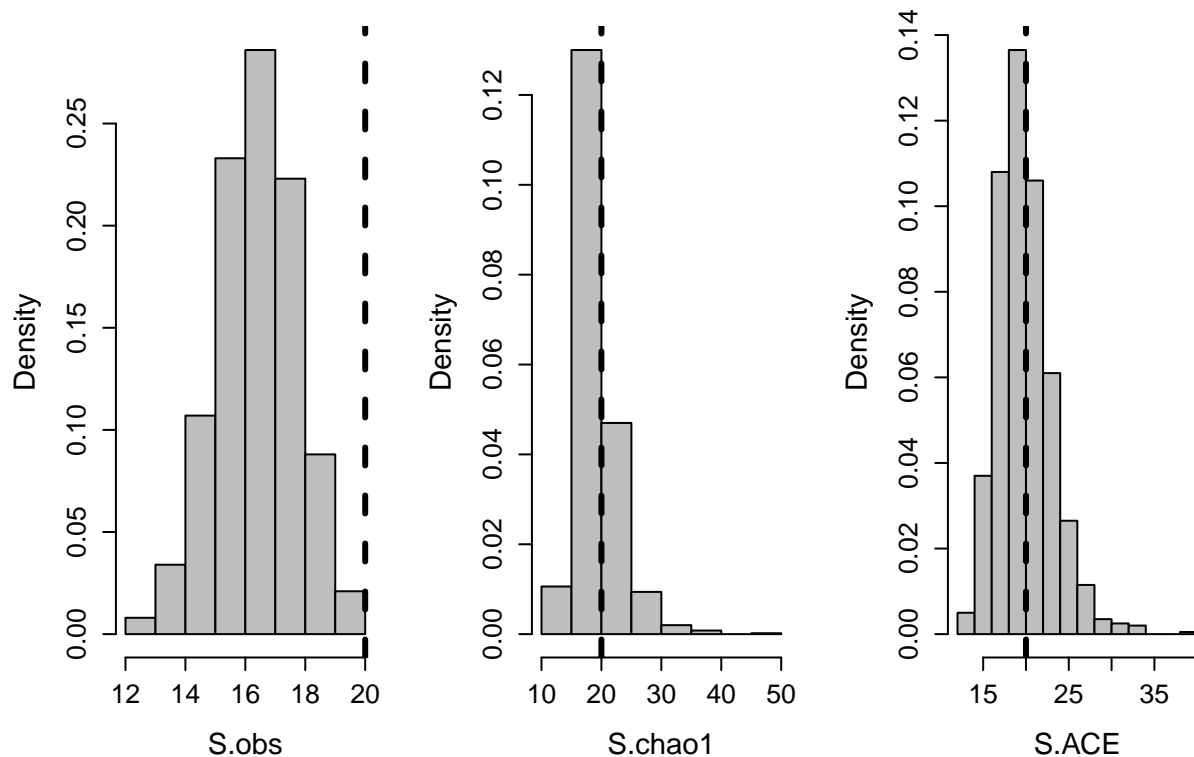
```
# Number of individuals per sample
n <- 100
# Number of repeated samples
N <- 1000
# initialize matrix to store results
repeated.S <- matrix(NA, nrow = N, ncol = length(S))
colnames(repeated.S) <- names(S)
# resample and store results
for (i in seq(N)){
  sample.jelly <- sample.community(x=jelly.jar,n=n)
  repeated.S[i,] <- as.vector(estimateR(sample.jelly))
}
```

```

}

#plot distribution of results as histogram
plot.new()
par(mfrow=c(1,3))
for (i in c("S.obs", "S.chao1", "S.ACE")){
  hist(repeated.S[,i],prob=TRUE, cex.lab = 1.4, cex.axis = 1.25, xlab=i, main = NULL, col="grey")
  #add dashed line to mark true S
  abline(v=length(jelly.jar), lty=2, lwd=3)
}

```



Play around with the sample size (by changing the value of `n`). How does sample size affect observed and estimated richness?

E. Collector's Curve

`S.observed` as a function of `N`. First take a single sample of size `N` from the source population. Subsample you sample with sample sizes increasing from 1 to `N` and note the observed richness for each subsample.

```

#first take a single sample of size N from the source population
N<-1000
sample.dj <- sample.community(jelly.jar,n = N)

#generate vector to store results
S.collectors <- rep(NA, N)
#loop through 1:N and check S.obs
for (i in 1:N){

```

```

    current <- sample.community(sample.dj,n = i)
    S.collectors[i] <- S.obs(t(current))
}

plot(1:N, S.collectors, xlab = "Sample size",ylab = "S.obs")

```

What is the smallest sample size to indicate the true richness? What is the largest sample size to underestimate richness? What sample size would you recommend for the next survey of this community?

Shannon diversity

- show how to translate mathematical notation of summation into loop, i.e. how to transfer the logic between math and code

** consider whether to add this next section** ### Beta diversity * nested for loops *how to avoid them