Data Extraction using conditionals

Gordon Erlebacher

Review

- Vectors, lists are both containers (also called objetcts)
- Functions take arguments, list instructions, return one or more objects
- Most of R operates through functions
- Create vectors using 3:6, seq(10,23,3), rep(...) (?rep)
- Extract data from vectors and lists using [..] and [[..]]
- Extract data from vectors and lists using conditionals (covered today)

Random numbers

- Random numbers selected from a normal distribution: *rnorm(n)*
- Random numbers from a uniform distribution: *runif(n)*
- Histograms: *hist(vector)* # *?hist*

```
hist(rnorm(2000), bins=30)
hist(runif(2000), bins=30)
```

Random Integers

?sample

- sample(x, size, replace = FALSE, prob = NULL)
- Use the function sample(....) to generate random integers
- sample(10) # permutation of numbers (change their order)
- sample(10,20)
  sample(10,20,rep=T)
- sample(c(66:80),20,rep=T) # 15 numbers to choose from
- sample(c(66,80),20,rep=T) # only two numbers to choose from

Example

- One has a list of 65 respondents to a questionaire
  - Their heights range from 60 to 100 inches
  - Their weights range from 140 to 190 pounds
  - Their eye colors are either brown, blue or green
- Create vectors of weights, heights and eye colors (using the sample() function)
- Calculate the mean weight of all respondents whose heights are between 70 and 85 inches and have blue eyes.
- Repeat the calculation for respondents with brown eyes.

Calculate the mean weight of all respondents whose heights are between 70 and 85 inches and have blue eyes.

1. Construct a logical vector with **T** matching the conditions above.

2. Extract the required weight elements

3. Calculate the mean value

Height, weight, eye vectors

```
weights = sample(c(60:100),65,rep=T)

heights  = sample(c(140:190),65,rep=T)

eye.color = sample(c("blue","green", "brown", 65, rep=T)
```

Repeat the calculation for respondents with brown eyes

What should you do?

```
cond = heights >= 70  &&  heights <= 85 and eye.color == “blue”

weights.subset = weights[cond]

mean.weight = mean(weights.subset)
```

Create a function!!!

```
function = color.mean(wgts, hghts, eyes, eye.color) {
   cond = hghts >= 70  &&  hgths <= 85 and eyes == eye.color
   weights.subset = wghts[cond]
   mean.weight = mean(weights.subset)
   return(mean.weight)
}
```

```
print(color.mean(weights, heights, eyes, “blue”))
print(color.mean(weights, heights, eyes, “brown”))
```

Data Frames

- collection of columns
- every column is a vector (elements of the same type)
- the different vectors are elements of a list

- Function to create data frames:

  data.frame(.....)

Before data.frames
some useful functions

- printing: print()
- concatenating: paste(str1, str2, sep=",")
- structure: str(...)
- object type: typeof(...)
- object class: class()
- read csv file: read.csv(...)
- write csv file: write.csv(...)
- built-in datasets: data(...)

Datasets

- data() returns a list of all datasets that are integrated into R. Each package/library has its own datasets.
- One of these datasets is Orange
- Find out more about Orange: ?Orange
- Explore the examples using Orange:
  - example(Orange)

Orange

- Orange has three columns: Tree, age, circumference
- Find out more about Orange: examine its structure: str(Orange)

Seatbelts

- class(Seatbelts) # returns "mts" "ts"
- sb = as.data.frame(Seatbelts) # new function (converts object to data.frame)

  class(sb) # returns data.frame

- compare str(Seatbelts) to str(sb)
  Different types of objects display differently when their names are typed

sb : data.frame

- sb # prints too many lines. Can no longer see column headers

- head(sb) # prints top 6 lines
  head(sb,15) # prints top 15 lines

data.frame: sb

- What is the average number of drivers killed with when the seatbelt law was inactive versus active?
- What is the average "fraction" of drivers killed (compared to the number of drivers killed or injured). Did the enactment of seatbelt laws increase or decrease the average fraction of drivers killed? Note that we are not measuring the percentage of killed drivers relative to all drivers (we do not have the total number of drivers)
- We will discuss regression, correlation, hypothesis tests in future lessons.

Solution to this example

- Demonstration in class with help of students

Orange

- Questions one might ask:
  - what is the mean circumference of all oranges
  - what is the mean trunk circumference of each orange tree (there are 4 types of trees)
  -