CS 133 - Introduction to Computational and Data Science

Instructor: Renzhi Cao
Computer Science Department
Pacific Lutheran University
Spring 2017

Announcement

- *Read book for R control structure and function.*
- *Final project*
- *Today we are going to learn R control structure and function.*

Selected looping command

R has some functions which implement looping in a compact form to make your life easier.

lapply(): Loop over a list and evaluate a function on each element:

>str(lapply)

example

>mylist <- list(a=1:10, b=20:100, c=30:50)

>lapply(mylist,mean)

Exercises

- Create PracticeR3.R and save today's work on that file.
- Create a list **mylist** with three elements: a, b, c, assign values to there three elements (you can decide what values to put).
- Create a function f with one parameter (a list), and evaluate the mean of each elements in the input parameter.

Useful statistics function

Function	Description
mean(*x*, trim=0, na.rm=FALSE)	mean of object x # trimmed mean, removing any missing values and # 5 percent of highest and lowest scores mx <- mean(x,trim=.05,na.rm=TRUE)
sd(*x*)	standard deviation of object(x). also look at var(x) for variance and mad(x) for median absolute deviation.
median(*x*)	median
quantile(*x*, *probs*)	quantiles where x is the numeric vector whose quantiles are desired and probs is a numeric vector with probabilities in [0,1]. # 30th and 84th percentiles of x y <- quantile(x, c(.3,.84))
range(*x*)	range
sum(*x*)	sum
diff(*x*, lag=*1*)	lagged differences, with lag indicating which lag to use
min(*x*)	minimum
max(*x*)	maximum
scale(*x*, center=TRUE, scale=TRUE)	column center or standardize a matrix.

Useful statistics function

Function	Description
seq(*from* , *to*, *by*)	generate a sequence indices <- seq(1,10,2) #indices is c(1, 3, 5, 7, 9)
rep(*x*, *ntimes*)	repeat *x n* times y <- rep(1:3, 2) # y is c(1, 2, 3, 1, 2, 3)

For final project:

Cor(x, y = NULL, use = "everything", method = c("pearson", "kendall",
"spearman"))

Calculate correlation between two vectors.

Exercises

- Use seq and rep function. First create vector v1 with odd numbers from 0 to 100. And then create vector v2 which repeats the vector v1 three times.

- Calculate the mean, standard deviation, median, sum, min, max and range of v3.

- Create two vectors: (1,2,3,4,5,6), (9,8,7,6,5,4), use Cor function to calculate the correlation between this two vectors. (This is very useful for your final project).

Learning R plotting by example

- R has very powerful plotting function.

Application of R

Variable	Description
Sales	Total unit sales of the grape juice in one week in a store
Price	Average unit price of the grape juice in the week
ad_type	The in-store advertisement type to promote the grape juice. ad_type = 0, the theme of the ad is natural production of the juice ad_type = 1, the theme of the ad is family health caring
price_apple	Average unit price of the apple juice in the same store in the week
price_cookies	Average unit price of the cookies in the same store in the week

http://www.dataapple.net/?p=19

Reading data from files

```
> data <- read.csv("http://cs.plu.edu/~caora/Rdata/grapeJuice.csv",
    header = T)
```

If you just want to have a look for this data, you can:

```
> initial <- read.csv("http://cs.plu.edu/~caora/Rdata/grapeJuice.csv",
    header = T, nrows=5)
> names(initial) <- c("name1","name2","name3","name4","name5")
> initial$name1
```

Simple analysis of the marketing data

```
> data <- read.csv("http://cs.plu.edu/~caora/Rdata/grapeJuice.csv", header =
    T)

> head(data)

> summary(data)
```

Simple analysis of the marketing data

```
> par(mfrow = c(1,2))  #set the 1 by 2 layout plot window

> boxplot(data$sales,horizontal = TRUE, xlab="sales") # boxplot to check if
      there are outliers

> hist(data$sales,main="",xlab="sales",prob=T) # histogram to explore the
      data distribution shape

> lines(density(data$sales),lty="dashed",lwd=2.5,col="red")
```

More analysis

The marketing team wants to find out the ad with better effectiveness for sales between the two types of ads, one is with natural production theme; the other is with family health caring theme.

```
> #divide the dataset into two sub dataset by ad_type
> sales_ad_nature = subset(data,ad_type==0)
> sales_ad_family = subset(data,ad_type==1)

> #calculate the mean of sales with different ad_type
> mean(sales_ad_nature$sales)
> mean(sales_ad_family$sales)

> # calculating the t test
> t.test(sales_ad_nature$sales,sales_ad_family$sales)
```

More analysis

The marketing team wants to find out the ad with better effectiveness for sales between the two types of ads, one is with natural production theme; the other is with family health caring theme.

```
> #set the 1 by 2 layout plot window
> par(mfrow = c(1,2))
>
> # histogram to explore the data distribution shapes
> hist(sales_ad_nature$sales,main="",xlab="sales with nature production theme ad",prob=T)
> lines(density(sales_ad_nature$sales),lty="dashed",lwd=2.5,col="red")
>
> hist(sales_ad_family$sales,main="",xlab="sales with family health caring theme ad",prob=T)
> lines(density(sales_ad_family$sales),lty="dashed",lwd=2.5,col="red")
```

Practice more plots

You can try all different kinds of plots on your data, and it's quite easy with the help of R

```
> # line charts
> plot(sales_ad_family$sales, sales_ad_nature$sales) #(type="o", col="blue")
> # Bar plot
> barplot(sales_ad_family$sales)
> # pie charts
> testData <- c(100,20,300,100,1)
> pie(testData, col=rainbow(length(testData)),labels=c("Mon","Tue","Wed","Thu","Fri"))
```

More examples:
http://www.harding.edu/fmccown/r/

Final best profit

Assume you want to get higher profit rather than just higher sales quantity, and you find out the relationship between sales and price is: ***Sales = 772.64 – 51.24*price***
Assume the cost per each juice is 5, you can now calculate the profit by:
Y = (price – 5) * Sales = – 51.24 * price^2 + 1028.84 * price – 3863.2

```
> f <- function(x) {
    profit = -51.24*x*x + 1028.84 * x - 3863.2
    return(profit)
}

> optimize(f,lower=0,upper=20,maximum=TRUE)
```

Practice

https://www.cs.plu.edu/~caora/cs133/Code/day24/IntroR.html

Do statistical analysis and draw pictures for your final project.