Meeting 3

Anh Le

September 12, 2014

1 Readings

Read Sec. 1.4, Functions, operators, and loops.

This is where you start thinking like a programmer. Up until now, we have only learned names of commands, what are their options, etc. Basically, instead of point-and-click in Excel, we now type into RStudio.

On the other hand, functions and loops are the real "logic" of programming and will make you very powerful. We don't point-and-click anymore, but write guidances for the computer to point-and-click for us.

Consider the following function that is trivial to write but very useful in real (grad) life. Suppose your advisor told you to calculate the summary statistics of ALL the variable in the dataset.

```
data <- data.frame(var1=rnorm(10, mean=0, sd=1),
                   var2=rnorm(10, mean=2, sd=5),
                   var3=runif(10, min=1, max=6))
data
```

```
##        var1    var2  var3
## 1  -0.17069  2.5044 5.273
## 2   0.38687 -8.3671 2.219
## 3   0.72805 12.4770 5.809
## 4  -0.72370 -0.7633 2.312
## 5  -0.40455  2.8970 2.024
## 6  -0.09673  1.7595 4.119
## 7   0.71740 -1.6797 5.212
## 8   1.17321 -2.7695 1.522
## 9   0.04649  1.2906 5.887
## 10 -1.12565 11.7990 2.329
```

Now, if we know some R but still in the "point-and-click" stage, we will write something like this:

```r
mean(data$var1)
```

```
## [1] 0.05307
```

```r
sd(data$var1)
```

```
## [1] 0.7101
```

```r
mean(data$var2)
```

```
## [1] 1.915
```

```r
sd(data$var2)
```

```
## [1] 6.313
```

```r
mean(data$var3)
```

```
## [1] 3.67
```

```r
sd(data$var3)
```

```
## [1] 1.755
```

So instead of "point-and-click" 6 times in Excel, we wrote 6 lines in R. Still do-able.

But God forbid, what if the dataset has 100 variables? What if we need more summary statistic than just `mean` and `sd`? Functions to the rescue.

```r
my_summary_statistic_function <- function(data_vector) {
  mean_of_data <- mean(data_vector, na.rm=TRUE)
  median_of_data <- median(data_vector, na.rm=TRUE)
  sd_of_data <- sd(data_vector, na.rm=TRUE)

  result <- c(mean_of_data, median_of_data, sd_of_data)
  names(result) <- c("mean", "median", "standard deviation")

  return(result)
}

my_summary_statistic_function(data$var1)
```

```
##               mean             median standard deviation
##            0.05307           -0.02512            0.71007
```

```r
my_summary_statistic_function(data$var2)
```

```
##               mean             median standard deviation
##              1.915              1.525              6.313
```

```
my_summary_statistic_function(data$var3)
```

```
##               mean             median standard deviation
##              3.670              3.224              1.755
```

We have just created a **re-usable** function that can take in any vector of data and spits out our desirable statistic. If the advisor at some point wants us to add a statistic, we only have to add that statistic to our function.

And lastly, instead of using our `my_summary_statistic_function` on each column by hand, we can also tell R to do it as follows:

```
apply(data, MARGIN=2, FUN=my_summary_statistic_function)
```

```
##                        var1  var2  var3
## mean                0.05307 1.915 3.670
## median             -0.02512 1.525 3.224
## standard deviation  0.71007 6.313 1.755
```

```
# Equivalently, we can use loop
for (i in 1:ncol(data)) {
  print(my_summary_statistic_function(data[ , i]))
}
```

```
##               mean             median standard deviation
##            0.05307           -0.02512            0.71007
##               mean             median standard deviation
##              1.915              1.525              6.313
##               mean             median standard deviation
##              3.670              3.224              1.755
```

2 In-class Exercises

1. See our website.
2. Trivial: `print("Hello World!")` 100 times.
3. Less trivial: `print("Hello friend #1!")`, `print("Hello friend #2!")`, etc. 100 times
4. Decidedly not trivial: Prove the Central limit theorem with simulation.

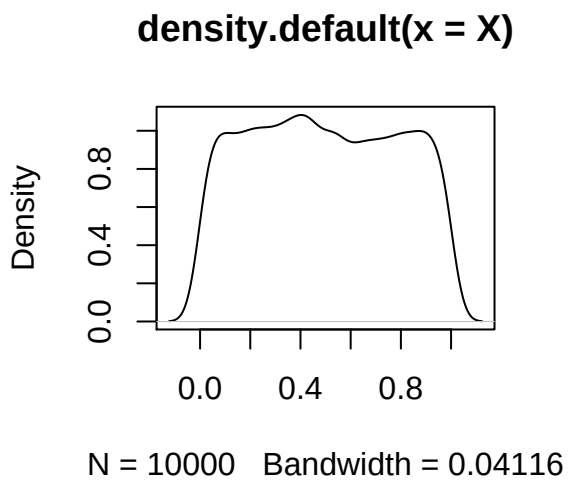
   First, the Central Limit Theorem (CLT)–the most important theorem in statistics–states that the mean of a random variable is normally distributed, **regardless of the distribution of the random variable itself**

   For example, let's say that X is uniformly distributed between 0 and 1

```r
X <- runif(10000, min=0, max=1)
mean(X)

## [1] 0.4951

plot(density(X))
```



Clearly, the distribution of X is not normal. If we generate X many times (says, n times), and each time calculate its mean, then we will get a vector containing n means. The CLT states that these n means will be normally distributed. (It's amazing! Where does the normal come from??)

Prove it using for loop. Hint: `for (i in 1:n) { do stuff in here }`