LISA Short Course Series
Basics of R

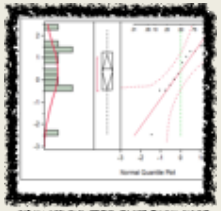
Olawale Awe

Summer 2014

LISA

Laboratory for Interdisciplinary Statistical Analysis

LISA helps VT researchers benefit from the use of **Statistics**



Collaboration:

Visit our website to request personalized statistical advice and assistance with:

Experimental Design • Data Analysis • Interpreting Results
Grant Proposals • Software (R, SAS, JMP, SPSS...)

LISA statistical collaborators aim to explain concepts in ways useful for your research.

Great advice right now: *Meet with LISA **before** collecting your data.*

LISA also offers:

Educational Short Courses: Designed to help graduate students apply statistics in their research
Walk-In Consulting: M-F 1-3 PM, GLC Video Conference Room & 11-1PM at OSB for questions requiring <30 mins

All services are **FREE** for VT researchers. We assist with research—not class projects or homework.

Outline

1. What is R?
2. Why use R?
3. Installing R into your computer
4. Introduction to R studio
5. Data Structures, Manipulation and Sampling
6. Data Import
7. Exploratory Data Analysis
8. Loops
9. If/Else Statements
10. Data Export

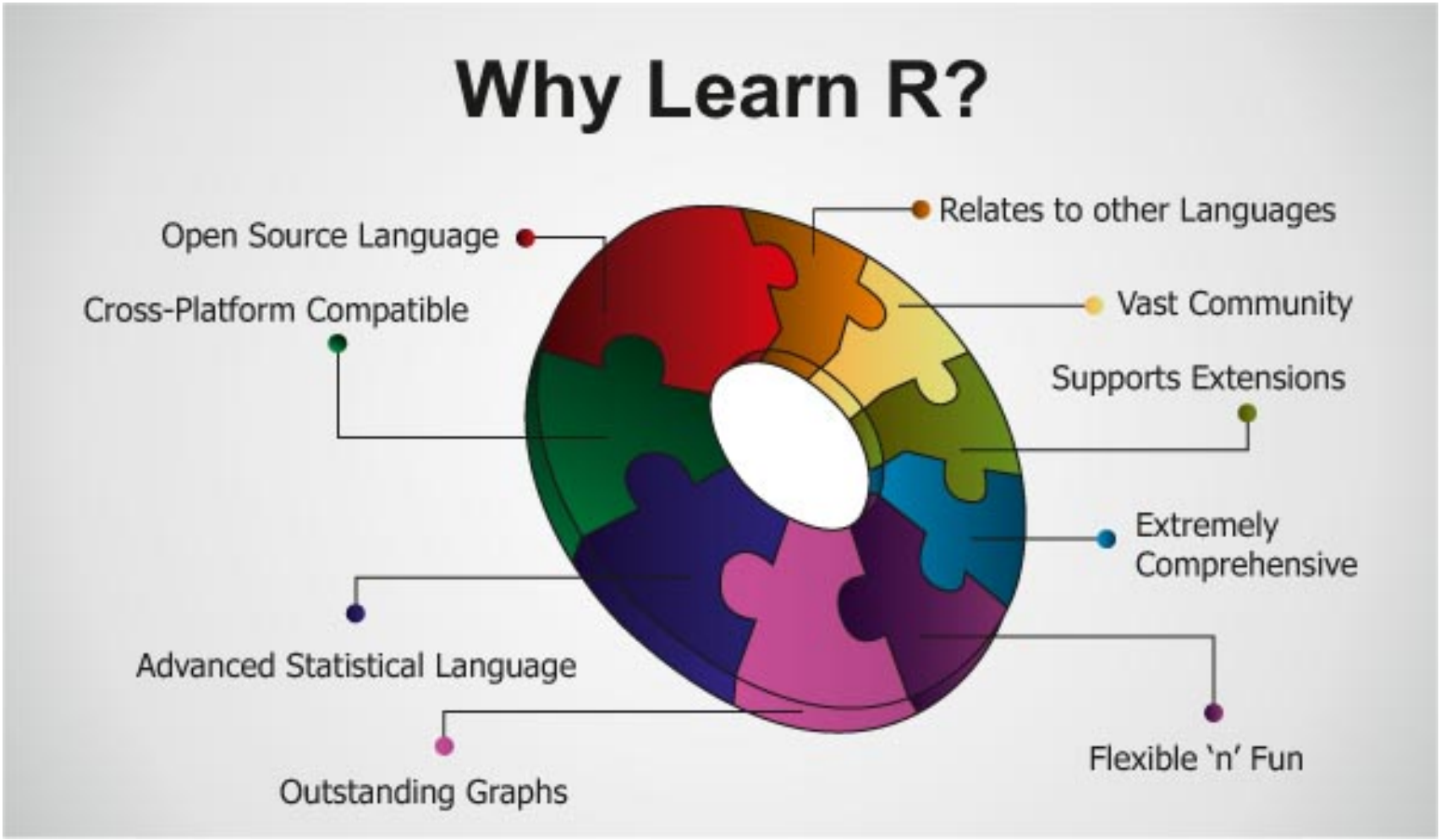
What is R?

- R is an extensive, object-oriented programming language.
- An integrated suite of software facilities for data manipulation, simulation, calculation and sophisticated graphical displays.
- Started by Robert Gentleman and Ross Ihaka (hence "R") in 1995
  -as a free, independent, open-source implementation of the S programming.
- It handles and analyzes data very effectively and it contains a suite of operators for calculations on arrays and matrices.
- Currently maintained by the R Core development team – an international group of hard-working volunteer developers.

`http://www.r-project.org`

Why use/Learn R?

- R is FREE!
- It contains many built-in functions and installable packages that will cover nearly every possible need you have.
- R is flexible-you are not restricted to built-in functions. You can write your own codes.
- Interactive console makes testing and debugging easy.



Source: http://www.edureka.in/blog

How to Install R on your computer?

Windows:
http://cran.r-project.org/bin/windows/base/

MacOs X:
http://cran.r-project.org/bin/macosx/

Introduction to R- Studio

R

www.r-project.org

The engine

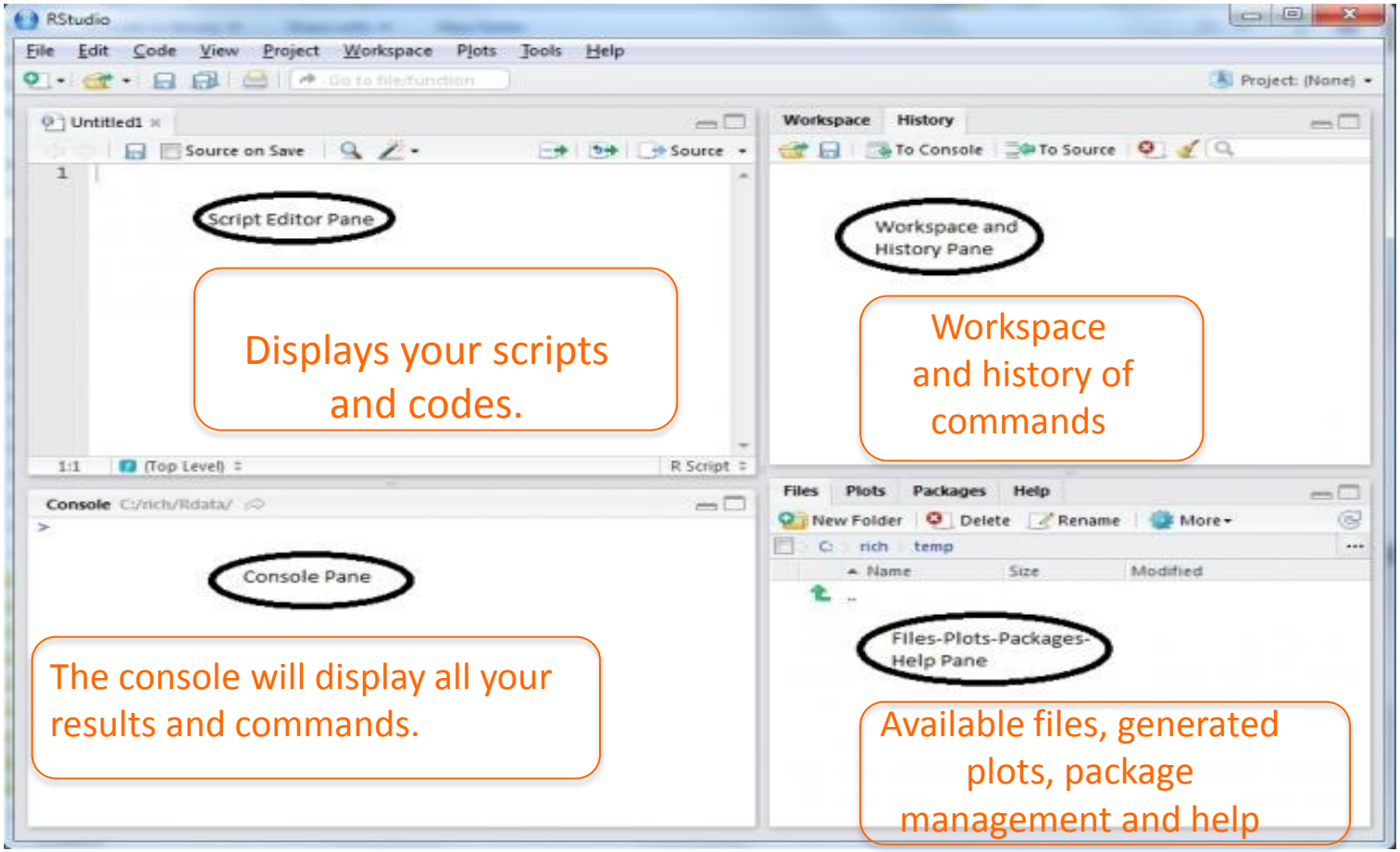
Rstudio

www.rstudio.org

The pretty face

An interface that makes working with R easier!

R Studio: Has 4 windows(panes)



Data Structures and Manipulation

1. Object Creation
   Expression: When you to write, evaluate and print your result on the screen.
   Assignment: Storing the results of expressions by assigning it to a variable.

2. Vectors:
   The basic data structure in R. (Scalars are vectors of dimension 1).
   a. Creating sequences:
      - : command. Creates a sequence incrementing/decrementing by 1
      - seq() command.
   b. Vector of numbers with no pattern. c() function.
   c. Vectors of characters. Also use c() function with the help of “”
   d. Repeating values. rep() function.
   e. Arithmetic with vectors: All basic operations can be performed with vectors. Round(),sort(),length().
   f. Subsets: The basic syntax for subsetting vectors is: vector[index]
   g. Sampling from vectors/distributions.

Random Sampling

- We can draw a random sample from a vector of numbers/or characters using the sample command:
- sample(vector,n,replace=T)-with replacement
- Sample(vector,n,replace=F)-without replacement.
- Sample(letters,n)-sample n letters from the 26 English alphabets.

Sampling from a Distribution

- You can generate random samples from various distributions, e.g :
- runif(n,min,max) : Uniform Distribution
- rbinom(n,size,prob.) :Binomial Distribution
- rexp(n,rate): Exponential Distribution
- rpois(n,rate): Poisson Distribution
- rnorm(n,mean,sd):Normal Distribution
- etc.

Data Structures and Manipulation

3. Matrices: Objects in two dimensions.
   a. Creating Matrices
      Command: matrix(data, nrow, ncol, byrow).
      Where data= list of elements that will fill the matrix.
      nrow, ncol: number of elements in the rows and the columns respectively.
      byrow: filling the matrix by row. The default is FALSE.

   b. Some Matrix Functions
      - dim(): Lists the dimensions of the matrix.
      - cbind: Creating matrix by putting columns together.
      - rbind: Creating matrix by putting rows together.
      - diag(d): Creates identity matrix of dimension d.

Data Structures and Manipulation

c. Some Matrix computations
   - Addition.
   - Subtraction
   - Inverse: function solve()
   - Transpose: function t()
   - Element-wise multiplication: *
   - Matrix multiplication: %*%

d. Subsets
   - Referencig a cell: matrix[r,c], where r represents the row and c represents the column.
   - Referencing a row: matrix[r,]
   - Referencing a column: matrix[,c]

Exercise 1: Prices Data Set (prices.csv)

The data are a random sample of records of resales of houses from Feb 15 to Apr 30, 1993 from the files maintained by the Albuquerque Board of Realtors.
Number of cases: 65

Variable Names:

PRICE = Selling price ($hundreds)
SQFT = Square feet of living space
AGE = Age of house (years)
NE = Located in northeast sector of city (1) or not (0)

Data Import

We need to set the working directory. For this we use the function setwd:

setwd("location")

1. Comma Separated Values:
   Use the function read.table
   mydatacsv<- read.table('prices.csv', sep=',', header=T)

2. Text File:
   Use the function read.table:
   mydatatxt<- read.table('prices.txt', sep='\t', header=T)

Practice 1.

Lets review some of the matrix commands we learned previously by applying them to our new dataset.

1. What is the dimension of our dataset?
2. Assign the value of the cell [2,3] to the new variable var1
3. Assign the value of the cell [10,4] to the new variable var2
4. Output the value of each column separately.
5. Assign the values of SQFT to a new variable SQFT.
6. Output the value of row 15.

Exploratory Data Analysis: Summaries

Quantitative summary of variable SQFT. We will calculate the minimum, maximum, mean, variance, median for that variable.

```
mean(SQFT)
var(SQFT)
min(SQFT)
max(SQFT)
median(SQFT)
```

You can obtain the 5 number summary for the variable by using the command:

```
summary(SQFT)
```

Exploratory Data Analysis: Graphs

1. Histogram of SQFT

   hist(SQFT, main="Histogram of square feet of living space", col="dodgerblue", breaks=10)

2. Boxplot of SQFT

   Boxplot(SQFT, main="Boxplot of square feet of living space", col="khaki1", ylab="SQFT")

3. Boxplot of SQFT by NE .

   boxplot(PW~mydatacsv[,4])

4. Normal Quantile-Quantile Plot

   qqnorm(SQFT, main="Normal QQ Plot SQFT")

http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf

For Loops

This statement allows for code to be executed repeatedly.

```
for(i in 1:n){
    statement
}
```

While Loops

This statement allows for code to be executed repeatedly while a condition holds true.

```
while(condition){
   statement
}
```

If/Else Statement

if statement – use this statement to execute some code only if a specified condition is true:

```
if (condition) {
    statement
}
```

If/Else Statement

if...else statement – use this statement to execute some code if the condition is true and another code if the condition is false.

```
if ( condition )
 statement
else
statement2
```

If/Else Statement

if...else if....else statement – use this statement to select one of many blocks of code to be executed

```
if (condition){
statement1
} else{
if (condition2){
statement2
} else {
Statement3
}
}
```

Data Export: csv

If you have modified your dataset in R you can export it as a .csv file using the following code:

```
write.csv(mydatacsv,file="mydatacsv.csv")
```

Can also export vectors or other objects that you have created to .csv file:

```
write.csv(vec2,file="vec2.csv")
```

Data Export: txt

If you have modified your dataset in R you can export it as a space delimited .txt file using the following code:

```
write.table(mydatacsv,file="mydatatxt.txt", sep=" ")
```

You can export it as a tab delimited .txt file using the following code:

```
write.table(mydatacsv,file="mydatatxt2.txt",
sep="\t")
```

Excercise 2:National Longitudinal Mortality Study Dataset

The variable content for each record on the file includes demographic and socioeconomic variables from the current population survey combined with the underlying cause of death mortality outcome and the follow-up time until death for records of the deceased or 11 years of follow-up for those not deceased.

The manual of the data and a complete variable description is attached in the course materials on our website.

Practice 2 a.

1. Read into R the dataset pubfile.csv.
2. Determine the dimensions of the dataset
3. Extract the variable povpct, income as percent of poverty level (column 35) as a new variable.
4. Extract the variable ms, marital status (column 5) as a new variable.
5. Obtain the minimum, maximum, mean, variance, median for the variable povpct and store them in separate variables.
6. Create a vector with the stored values from 4.
7. Create a histogram of povpct of a different color with 20 breaks.

Practice 2 b.

1. Create a boxplot of povpct of a different color.
2. Create a boxplot of povpct by ms with the same color for all boxes.
3. Create a boxplot of povpct by ms with the same color for the first three boxes and another color for the remaining three boxes.
4. Create a normal Q-Q plot for Sepal Length.
5. Using for loops count how many observations are there in a metropolitan area (smsast=1) (col 20) with an age lower than 15 (col 2).
6. Export your extracted variables as a .csv file and the dataset as a tab delimited .txt file.

Please don't forget to fill the sign in sheet and to complete the survey that will be sent to you by email.

Thanks for coming!