DUG

Introduction to R – Towards Reproducible Research

Getting to know your workspace (RStudio)

- **EDITOR** Window – keep your code here and save it as a script
  - File > New > R Script
  - File > Save as Note: R Script files have the extension .R
- **CONSOLE** Window – test your code here
- **WORKSPACE** Window – Workspace and History a list of the objects you have created
- **CONTROL** Window – Files, Plots, Packages and Help

Starting to use R

- A **CALCULATOR**: 2+2 4*4
  Follows the order of operations – use brackets carefully
- Assigning values: use the = or <-
- An **ALGEBRAIC CALCULATOR** to solve equations: x = a+b

  a = 2, b=4 x
- Showing results: print(x)

Loading Data

WHERE IS THE DATA?

```
getwd()          # get the current working directory – this is where the files are expected
                  to be located
setwd("<new path>")  # set the working directory to "<new path>"  - file path should
                  use / and not the \
```

or: Session > Set Working Directory

CSV DATA: easiest to load:

```
Dataframe <- read.csv("C:/file.csv")      #reads the entire file
```

RStudio: Import Dataset > From Text File

EXCEL DATA: requires the packge gdata

```
Dataframe = read.xls"C:/file.xls          #reads the first sheet
```

SPSS DATA: requires the package foreign

```
Dataframe = read.spss("C:/file.spss", to.data.frame=TRUE)
```

When the data is loaded, it becomes an object that is labelled. The default label for your object is the file name that was loaded. This name is often long and cumbersome so it is good practice

Learning a new language

- **OBJECT**: anything that can be assigned to a variable
    - **MODE**: how the object is stored (numeric, complex, character and logical). An object has only 1 mode
    - **CLASS**: information on how it should be handled. If Class hasn't been defined, it will by default be the same as the mode.
    - **WHAT IS THE DIFFERENCE BETWEEN THE TWO**? You could have a set of numbers that have a "numeric" mode but are organized as a class "matrix". However, maybe these numbers aren't values but identifiers like an OEN. In that case you might want to work with the numbers as characters. You can change the mode from numeric to character (this is called coercion) while still keeping the format of the information in a matrix (the class).
    - **WHAT IS THE MODE OR CLASS OF AN OBJECT?**

      mode(object)
      class(object)

    - **WHY SHOULD I CARE?:** Some of the functions that we will want to use (creating graphs, summarizing data into tables, performing statistical analyses) can only be done with data that is formatted in a particular way (class). Some graphs will only work if the data is in a matrix, other graphs will only work if the data is a factor (categories). If you are running something and it doesn't work, double check that the data is in the class required by the function
- **DATA FRAME**: a table with columns that are variables and rows that are records or observations
    - Variables can be different types (numeric, character, logic etc.)
- **DATA STRUCTURES**
    - **VECTOR:** a row of data where all elements are of the same type
    - **MATRIX:** a table of data that where all elements are of the same type
    - **ARRAY:** a cube of data where all elements are of the same type
    - **DATA FRAME:** a table of data where the columns (variables) can be of different types
    - **LIST:** a row of data where all elements can be of different types
- **FACTORS:** data that is nominal or ordinal. Stored as a vector of values (numeric or character) with a corresponding set of character values that serve as labels (called levels).
    - i.e. 1=male, 2=female - Which is the value and which is the level?
        - 1 could be both a value and a level
        - Using male and female as the level, makes it easier to recognize and understand the data in that variable. Factor levels can be renamed.
        - Factor levels have an order. The default order is alphabetical. In this case, the factor levels male and female would actually be female as the first factor and male as the second factor. Factor levels can be re-ordered.
        - **WHAT ARE THE LEVELS OF A FACTOR**? levels(factor)

Functions

A function performs a specific action or task (a calculation, a plot, a statistical analysis). How the task is performed will depend on the criteria/information (arguments) you assign to it. 6 functions have already been presented:

- print()
- getwd()
- setwd()
- read.csv()
- mode()
- class()
- typeof()

Any word that is followed by () is a function. To get more information about a function use ?:

- ?getwd()
- Help(getwd)

Functions can be used within functions.

Checking Data

TO SEE THE DATA FRAME:

- double click on the object in the workspace and it will display in the editor window or,
- type the name of the object in the console and it will display in the console window

CHANGING COLUMN NAMES: colnames(dataframe)[column number] <- "new label"

[] indicates the position of a number in a list. [5] is the fifth number in a list of numbers

For tables, two numbers are used: [column,row]. [6,2] is the number located in the sixth column in the second row.

CHECKING HOW THE DATA HAS LOADED:

- dim(dataframe) – lists the number of records and the number of variables
- dimnames(dataframe) – list of variable names
- head(dataframe) – displays the list of the variables and 6 data examples
- str(dataframe) - displays the structure
- summary(dataframe) - displays summary statistics
- fix(dataframe) – opens an interactive table where you can make changes (good for small tables)

Missing Values

Missing or NULL values are coded as NA. When you load a data file, R will decide the class of each variable by the values that are included. If you have a variable with mixed data (numeric and character) you will need to decide whether you want the variable to be numeric or character. In the instance of the fictional data set, the assessment (Reading, Writing, Math) scores are mixed with numeric values for scores and character values for the test taking status (absent, deferred etc.). By setting the character values to NA, the variables will be loaded as numeric which will allow for the calculation of means, standard deviations, medians etc.

- object <- read.csv("pathway/file.csv", na.strings=c("character 1", "character 2"))

Some functions will automatically deal with NA values, others need to have NA's removed:

- na.rm =TRUE is the most often used to make a function ignore NA's
- NaN is an undefined mathematical function. An NaN is always NA but NA is not always NaN

Each of these functions needs to have the NA values excluded:

- sd(sample2$ROverallLevel, na.rm=TRUE)
- median(sample2$ROverallLevel, na.rm=TRUE)
- quantile(sample2$ROverallLevel,na.rm=TRUE)
- range(sample2$ROverallLevel, na.rm=TRUE)
- sum(sample2$ROverallLevel, na.rm=TRUE)
- min(sample2$ROveralLevel, na.rm=TRUE)
- max(sample2$ROverallLevel, na.rm=TRUE)

Coercion – Changing Data Classes

Sometimes you will want to use the data in ways that aren't possible with the data class using the as.dataclass functions. You can also check to see if a variable has a particular data class using the is.dataclass functions:

- as.numeric()
- as.character()
- as.vector()
- as.matrix()
- as.table()
- as.data.frame()
- as.factor

- is.numeric()
- is.character()
- is.vector()
- is.matrix()
- is.table()
- is.data.frame()
- is.factor

is.dataclass returns a "TRUE" or "FALSE" value. as.dataclass temporarily changes the data class for the function you are using it in. You can make it permanent by assigning it to the dataframe object:

dataframe <- as.numeric(sample$variable)

Useful Functions

FUNCTION	EXAMPLE
Table()	Table(dataframe$variable) - summary of counts for 1 variable Table(dataframe$variable1, dataframe$variable2) – summary of counts for 2 variables as a cross-tab
t-test(y~x)	t.test(sample2$ROverallLevel~sample2$SIF_IEP) – independent 2 group: y is numeric and x is binary
t.test(y1,y2)	Independent 2 group: y1 is numeric and y2 is numeric
t.test(y1.y2, paired=TRUE)	Paired t-test: y1 and y2 are numeric
lm(outcome ~ predictor)	lm(sample2$ROverallRawLevel_Dot ~ sample2$WOverallRawLevel_Dot)
Summary(object)	Table of summary statistics
Chisq.test(table)	Chisq.test(Reading.Table) - use this function with a table of frequencies
Prop.test()	Used to test the null that the proportions (probabilities of success) in several groups are the same.

Useful Packages

vcd – for visualizing categorical data

ggplot2 – grammar for graphics plotting system

plyr – package to solve a variety of common problems