Introduction to R

Notes

Using basic functions in R

Inspecting function documentation

There are many functions in base R that you can use (in a later tutorial, we will discuss how to create your own functions!). In order to view documentation for a function in R, you can type `?function_name` or search for the function name in the help tab. As practice, let's explore the `matrix()` and `mean()` functions in R by typing `?matrix` and `?mean` into the console. Uncomment (by deleting the #s) the code below and run it in the console to inspect the documentation.

```
#?matrix
#?mean
```

Under usage for the matrix() function, you should see the following:
matrix(data = NA, nrow = 1, ncol = 1, byrow = FALSE, dimnames = NULL) The matrix function has 5 distinct parameters. All of them have default values. For example, if you don't put any data into the function, the resulting matrix will be made up of NA values (this is how missing data is generally coded in R). Additionally, the function will have 1 row and 1 column by default and data will be filled by columns because byrow=FALSE by default. The matrix will not have any row or column names because dimnames=NULL'= by default.

Because all parameters of the matrix() function have defaults, we could call matrix() with no inputs and we would get a 1x1 matrix with NA values and no dimension names. We can also pick and choose whatever parameters we do want to fill in and ignore anything that we want to leave as defaults.

```
# Call matrix() with no inputs
matrix()
```

```
##      [,1]
## [1,]   NA
```

```
# Make a 2x3 matrix of NAs
matrix(nrow=2, ncol=3)
```

```
##      [,1] [,2] [,3]
## [1,]   NA   NA   NA
## [2,]   NA   NA   NA
```

In contrast, the mean() function has some required parameters. When you type `?mean` into the console, you should see the following:
mean(x, trim = 0, na.rm = FALSE, ...)
The parameter x has no default value; you must specify the values that you want to take the mean of. The other paramters are given defaults: trim=0 and na.rm=FALSE. If you read the descriptions for these parameters, you will see that trim allows you to calculate a trimmed mean (i.e., eliminate some proportion of extreme values before calculating the mean) and na.rm allows you to remove missing (NA) values before calculating the mean. By default, R will not do any trimming and NAs will not be removed. This can cause issues (see below):

```
# Create some data and save it as data1 and data2
data1 <- c(1,2,3,4,7,NA)
data2 <- c(1,2,3,4,7)

# Calculate the mean of data1 and data2
```

```r
# Note that data1 has a mean of NA because there was an NA value that was not removed
mean(data1)
```

```
## [1] NA
```

```r
mean(data2)
```

```
## [1] 3.4
```

```r
# Now explicitly set na.rm=TRUE and recalculate the mean of data1. Now we get the same as data2
mean(data1, na.rm=TRUE)
```

```
## [1] 3.4
```

```r
# Now just to see what happens when you fail to provide a parameter, call mean()
# -- you need to get used to error messages and glean what you can from them; they often are very preci

# Uncomment the following code to see the error messages. This needs to be commented to be able to knit
# mean()
```

Note that many functions that take vector inputs (i.e., max(), min(), sum(), etc.) have an na.rm parameter, which is set to FALSE by default. So, if you try to use one of these functions and get NA as the result, then you might have to add na.rm=TRUE to the function call.

Calling functions in R

When using functions in R, you can make it clear which inputs refer to which parameters by either a) using the order of the parameters specified in the usage or b) naming them explicitly.

```r
# Note that these two calls to the mean() function will return the same output
mean(data1, .2, TRUE)
```

```
## [1] 3
```

```r
mean(x=data1, trim=.2, na.rm=TRUE)
```

```
## [1] 3
```

```r
# However, the following code will give an error because "TRUE" is not a valid input to trim,
# Which is the second parameter listed in the usage
# Mean(data1,TRUE)

# To specify na.rm=TRUE but leave the trim=0 default as is, we simply do the following:
mean(data1, na.rm=TRUE)
```

```
## [1] 3.4
```

```r
# (Note that, since data1 still matches to x, which is the first parameter in usage,
#  we do not need to specify x=data1)
```

Data types and structures

This section covers:
-Four basic data types in R: characters, numerics, integers, and logicals
-Four basic ways to store data in R: vectors, matrices, data frames, and lists

You can learn about these data types below

Numerics and integers

```r
# To store data in some variable name, use either = or <-
# To save the number 5 as "number":
number <- 5
number = 5 #does the same thing

# Print number:
print(number)
```

```
## [1] 5
```

```r
# Find out the class of number:
class(number)
```

```
## [1] "numeric"
```

```r
# Change number to an integer and re-save it as number2:
number2 <- as.integer(number)

# Another way to save a value as an integer:
number3 <- 5L
class(number3)
```

```
## [1] "integer"
```

```r
# Inspect the class of number2 to see that it is an integer:
class(number2)
```

```
## [1] "integer"
```

Characters

```r
# Save a character string in a variable named msg
#NOTE: we did not use "message" as the variable name as that would "mask" a funciton by the same name i
# in general, it is not advised to name variables: T, TRUE, F, FALSE, t, warnings, message, print, cat,
# reserved names.  When in doubt, type the name on the command line to see if it calls up a pre-defined
msg <- "welcome"

# Print msg:
print(msg)
```

```
## [1] "welcome"
```

```r
# Inspect the class of msg:
class(msg)
```

```
## [1] "character"
```

Logicals

Logical values are either TRUE or FALSE
In R, these reserved named variables function as if TRUE=1 and FALSE=0 (as.numeric(TRUE)=1)

```r
# Save the logical TRUE as a variable called outcome:
outcome <- TRUE

# Print outcome
print(outcome)
```

```
## [1] TRUE
```

```r
# Inspect class of outcome
class(outcome)
```

```
## [1] "logical"
```

```r
# Note that, based on the implicit conversion to numeric caused by calling the "+" operation on the log
outcome+outcome
```

```
## [1] 2
```

In R, you can test a statement to see if it is TRUE or FALSE. Note that R allows you to make comparisons across variable types: integers may be compared to numerics and logicals may be compared to integers/numerics (through a conversion process). For characters, comparatives are assessed using alphabetical order (letters earlier in the alphabet are "smaller"):
1) == means "is equal to"
2) != means "is not equal to"
3) > means "greater than"; >= means "greater than or equal to"
4) < means "less than; <= means"less than or equal to"
5) & means "and"; | means "or" (NOTE: & and && are not identical; nor are | and ||. The first operates on vectors; the second on scalars.)

```r
# Is 5 equal to 3?
5==3
```

```
## [1] FALSE
```

```r
# Is 5 not equal to 3?
5!=3
```

```
## [1] TRUE
```

```r
# Is 5 less than 3?
5<3
```

```
## [1] FALSE
```

```r
# Is 5 greater than 3?
5>3
```

```
## [1] TRUE
```

```r
# Is 5 greater than 3 AND less than 7?
5>3 & 5<7
```

```
## [1] TRUE
```

```r
# Is 5 less than 3 OR less than 7?
5<3 | 5<7
```

```
## [1] TRUE
```

```r
# Is 5 greater than 5?
5>5
```

```
## [1] FALSE
```

```r
# Is 5 greater than or equal to 5?
5>=5
```

```
## [1] TRUE
```

```r
# Is 5 equal to 5?
5==5
```

```
## [1] TRUE
```

```
# Is "hello" equal to "hello"?
"hello" == "hello"
```

```
## [1] TRUE
```

```
# Is "hello" equal to "goodbye"?
"hello" == "goodbye"
```

```
## [1] FALSE
```

```
# Is "hello" greater than "goodbye"? (in other words is "hello" after "goodbye" alphabetically?)
"hello">"goodbye"
```

```
## [1] TRUE
```

```
# Is TRUE == 1?
TRUE==1
```

```
## [1] TRUE
```

```
# Is FALSE==0?
FALSE==0
```

```
## [1] TRUE
```

Vectors

```
# The easiest way to create a vector is by using the c() function
vec1 <- c(2,3,4,5)
print(vec1)
```

```
## [1] 2 3 4 5
```

```
# Note that, if you include multiple data types in a vector, R will change all values to the same type
vec2 <- c(7,3)
vec2
```

```
## [1] 7 3
```

```
class(vec2)
```

```
## [1] "numeric"
```

```
vec3 <- c(7,3,"hello")
vec3
```

```
## [1] "7"     "3"     "hello"
```

```
class(vec3)
```

```
## [1] "character"
```

```
# For values in a row, we can also use a colon:
vec3 <- 2:5
print(vec3)
```

```
## [1] 2 3 4 5
```

```
# We can use the c() function to combine pre-saved vectors:
vec4 <- c(vec1,vec3)
print(vec4)
```

```
## [1] 2 3 4 5 2 3 4 5
```

```
# Use the length function to find the length of a vector
length(vec4)
```

```
## [1] 8
```

Here are some other useful shortcuts for creating vectors in R:

```
# Use the rep() function to repeat values. Inspect the following to see how it works!
rep(x=0, times=5) #create a vector of 5 zeros
```

```
## [1] 0 0 0 0 0
```

```
rep(x=c(1,2,3), times=5) #create a vector with five repeats of 1,2,3
```

```
##  [1] 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3
```

```
rep(c(1,2,3), each=2, times=5) #repeat each value in 1,2,3 twice, then repeat that 5 times
```

```
##  [1] 1 1 2 2 3 3 1 1 2 2 3 3 1 1 2 2 3 3 1 1 2 2 3 3 1 1 2 2 3 3
```

```
# Use the seq function to create a sequence of values
seq(from=1, to=5, by=.5) #create a vector with values from 1-5, incementing by .5
```

```
## [1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
```

```
seq(from=1, to=5, length.out=17) #create a vector with 17 equally spaced values going from 1 to 5
```

```
##  [1] 1.00 1.25 1.50 1.75 2.00 2.25 2.50 2.75 3.00 3.25 3.50 3.75 4.00 4.25
## [15] 4.50 4.75 5.00
```

Matrices

As shown above, an m by n matrix can be created in R using the matrix() function
Here are some examples

```
A <- matrix(2, nrow=3, ncol=3)
print(A)
```

```
##      [,1] [,2] [,3]
## [1,]    2    2    2
## [2,]    2    2    2
## [3,]    2    2    2
```

```
B <- matrix(c(1,2,5,3,4,0,2,1,5), nrow=3, ncol=3, byrow=TRUE)
print(B)
```

```
##      [,1] [,2] [,3]
## [1,]    1    2    5
## [2,]    3    4    0
## [3,]    2    1    5
```

```
# Print the dimensions of a matrix (number of rows followed by number of columns)
dim(A)
```

```
## [1] 3 3
```

```
# Matrix multiplication
A %*% B
```

```
##      [,1] [,2] [,3]
## [1,]   12   14   20
```

```
## [2,]   12   14   20
## [3,]   12   14   20
```

```
# Element-wise multiplication
A * B
```

```
##      [,1] [,2] [,3]
## [1,]    2    4   10
## [2,]    6    8    0
## [3,]    4    2   10
```

```
# Element-wise addition
A+B
```

```
##      [,1] [,2] [,3]
## [1,]    3    4    7
## [2,]    5    6    2
## [3,]    4    3    7
```

```
# Transpose of a matrix
t(B)
```

```
##      [,1] [,2] [,3]
## [1,]    1    3    2
## [2,]    2    4    1
## [3,]    5    0    5
```

```
# Inverse of a matrix
solve(B)
```

```
##            [,1]        [,2]        [,3]
## [1,] -0.5714286  0.14285714  0.57142857
## [2,]  0.4285714  0.14285714 -0.42857143
## [3,]  0.1428571 -0.08571429  0.05714286
```

Data Frames

Data frames are generally used to store tabular data and are composed of same-length vectors; these vectors can be of differing data types. In general, when you read a .csv data file into R, it will be saved as a data frame.
We can create a data frame in R as follows:

```
# Create a fake dataset called example_data
example_data <- data.frame(ID_Num = c(1:10),
                           Age = rep(24:28, each=2),
                           State = c(rep("New Jersey", 5), rep("New York", 5)))

# Change row names of the data frame (some made up names)
#NOTE: rownames must be unique; they will be coerced to be unique in some instances
rownames(example_data) <- c("Sarah", "Mike", "Drew", "Eric", "Maria",
                            "Lindsey", "Mark", "Jenny", "Sophie", "Paul")

# Print the data frame
example_data
```

```
##         ID_Num Age      State
## Sarah        1  24 New Jersey
## Mike         2  24 New Jersey
## Drew         3  25 New Jersey
```

```
## Eric         4  25 New Jersey
## Maria        5  26 New Jersey
## Lindsey      6  26   New York
## Mark         7  27   New York
## Jenny        8  27   New York
## Sophie       9  28   New York
## Paul        10  28   New York
```

The following R code outlines a few ways to inspect data in a data frame.

```
# Get dimensions (same as matrices)
dim(example_data)
```

```
## [1] 10  3
```

```
# Get number of columns
ncol(example_data)
```

```
## [1] 3
```

```
# Get number of rows
nrow(example_data)
```

```
## [1] 10
```

```
# Get summaries of the columns
summary(example_data)
```

```
##      ID_Num           Age             State
##  Min.   : 1.00   Min.   :24   New Jersey:5
##  1st Qu.: 3.25   1st Qu.:25   New York  :5
##  Median : 5.50   Median :26
##  Mean   : 5.50   Mean   :26
##  3rd Qu.: 7.75   3rd Qu.:27
##  Max.   :10.00   Max.   :28
```

```
# Access a single column of the data frame using $
example_data$Age
```

```
##  [1] 24 24 25 25 26 26 27 27 28 28
```

```
# Inspect row names
rownames(example_data)
```

```
##  [1] "Sarah"   "Mike"    "Drew"    "Eric"    "Maria"   "Lindsey" "Mark"
##  [8] "Jenny"   "Sophie"  "Paul"
```

```
# Inspect column names
colnames(example_data)
```

```
## [1] "ID_Num" "Age"    "State"
```

Lists

Lists enable multiple data types or data sets to be stored in a single object. For example, a list could have a data frame as its first element, a vector as its second element, and a character string as its third element.

```
# Save vector vec1, matrix A, and vector vec2 in a list called example list
example_list <- list(vec1, A, vec2)
```

```
# Print example_list
example_list
```

```
## [[1]]
## [1] 2 3 4 5
##
## [[2]]
##      [,1] [,2] [,3]
## [1,]    2    2    2
## [2,]    2    2    2
## [3,]    2    2    2
##
## [[3]]
## [1] 7 3
```

Indexing

In R, indices start at 1, not 0 as in some other languages. For example, the index of the 3rd element in a vector is 3.

Using indices to extract elements in a vector

We can use indices enclosed in square brackets in order to extract data from a vector as follows:

```
# This R chunk uses vector vec4 from above
# Re-print vec4
vec4
```

```
## [1] 2 3 4 5 2 3 4 5
```

```
# Extract the 3rd element in vec4
vec4[3]
```

```
## [1] 4
```

```
# Extract the 3rd through 5th elements in vec4
vec4[3:5]
```

```
## [1] 4 5 2
```

```
# Extract the 1st, 3rd, and 7th elements in vec4
vec4[c(1,3,7)]
```

```
## [1] 2 4 4
```

```
# Remove the 2nd element from vec4
vec4[-2]
```

```
## [1] 2 4 5 2 3 4 5
```

```
# Remove the 2nd, 4th, and 5th elements from vec4
vec4[-c(2,4,5)]
```

```
## [1] 2 4 3 4 5
```

We can also use the following functions to either a) get a logical vector indicating which values in the vector meet some criterion or b) get indices of values in a vector that meet some criterion.

```
# Logical vector of the same length as vec4
# TRUE wherever elements equal 2; FALSE elsewhere
vec4==2
```

```
## [1]  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
```

```
# Get indices of all values in vec4 that are equal to 2
which(vec4==2)
```

```
## [1] 1 5
```

```
# Get index of the maximum value in vec4
# If the maximum occurs more than once, this returns the first location by default
which.max(vec4)
```

```
## [1] 4
```

```
# Return logical vector indicating which elements of vec4 are either equal to 2 or 4
vec4 %in% c(2,4)
```

```
## [1]  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE
```

```
# Another way to do the same. Note that | means "or" and & means "and"
vec4==2 | vec4==4
```

```
## [1]  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE
```

By enclosing the output from the above functions in square brackets, we can extract elements meeting particular criteria from a vector. For example:

```
# Extract elements of vec4 that are equal to 2
vec4[vec4==2]
```

```
## [1] 2 2
```

```
# Extract elements of vec4 that are equal to 2 or 4
vec4[vec4==2|vec4==4]
```

```
## [1] 2 4 2 4
```

```
#or:
vec4[vec4 %in% c(2,4)]
```

```
## [1] 2 4 2 4
```

Using indices to extract elements in a matrix

In a similar way, we can use square brackets to extract elements from a matrix. However, we now need both row and column indices to specify a particular element. See examples below:

```
# Re-print matrix B for reference
B
```

```
##      [,1] [,2] [,3]
## [1,]    1    2    5
## [2,]    3    4    0
## [3,]    2    1    5
```

```
# Extract the element of matrix B that is located in row 3, column 2
B[3,2]
```

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

```
# Extract all elements of B that are greater than 2 and less than 5
B[B>2 & B<5]
```

```
## [1] 3 4
```

We can also use indexing to extract particular rows or columns of a matrix. Note that, in general, we extract elements from a matrix by using [row_index,column_index]. If we just want to specify row indices, but not column indicies, we can leave the column index blank; similarly, if we just want to specify column indices, we can leave the row index blank. For example:

```
# Extract the 2nd row of matrix B
B[2,]
```

```
## [1] 3 4 0
```

```
# Extract the 1st and 3rd rows of matrix B
B[c(1,3),]
```

```
##      [,1] [,2] [,3]
## [1,]    1    2    5
## [2,]    2    1    5
```

```
# Extract the 2nd column of matrix B
# Note that, because these values are in the same column, they are returned as a vector, not matrix
B[,2]
```

```
## [1] 2 4 1
```

```
# Extract the values that are in the 1st and 3rd rows and 2nd and 3rd columns of matrix B
B[c(1,3),2:3]
```

```
##      [,1] [,2]
## [1,]    2    5
## [2,]    1    5
```

Extracting elements of a data frame

Extracting values from a data frame works in much the same way as above; however, it is also possible to specify rows and columns of a data frame by name (or by using a dollar sign for columns). See below:

```
# Print the 3rd row of example_data
example_data[3,]
```

```
##      ID_Num Age      State
## Drew      3  25 New Jersey
```

```
# Print the 2nd and 3rd row of example_data
example_data[2:3,]
```

```
##      ID_Num Age      State
## Mike      2  24 New Jersey
## Drew      3  25 New Jersey
```

```
# Print Sarah's Age
example_data["Sarah","Age"]
```

```
## [1] 24
```

```
### THREE DIFFERENT WAYS TO GET THE 2ND COLUMN (Age)
# Using the column index
example_data[,2]
```

```
##  [1] 24 24 25 25 26 26 27 27 28 28
```

```
# Using a $
example_data$Age
```

```
##  [1] 24 24 25 25 26 26 27 27 28 28
```

```
# Using square brackets and the column name
example_data[,"Age"]
```

```
##  [1] 24 24 25 25 26 26 27 27 28 28
```

Some more examples of using logicals to extract specific data from a data frame:

```
# Extract the ages of people who are from New York
example_data$Age[example_data$State=="New York"]
```

```
## [1] 26 27 27 28 28
```

```
# Extract only the rows of example_data where Age is equal to 24
example_data[example_data$Age==24,]
```

```
##       ID_Num Age      State
## Sarah      1  24 New Jersey
## Mike       2  24 New Jersey
```

```
# Extract only the rows of example_data where Age is 24 and State is New York
example_data[example_data$Age==24 & example_data$State=="New York",]
```

```
## [1] ID_Num Age    State
## <0 rows> (or 0-length row.names)
```

```
# Extract the row names (i.e., names) of those who are 24 from New York
rownames(example_data)[example_data$Age==24 & example_data$State=="New Jersey"]
```

```
## [1] "Sarah" "Mike"
```

Extracting elements of a list

Extracting values from a list requires two steps: first you'll need to extract the element of the list you are interested in using double square brackets: [[]]. Then, you can use regular square brackets to extract values from each element as described in the above sections. See below:

```
# Re-print example_list
example_list
```

```
## [[1]]
## [1] 2 3 4 5
##
## [[2]]
##      [,1] [,2] [,3]
## [1,]    2    2    2
## [2,]    2    2    2
## [3,]    2    2    2
##
## [[3]]
## [1] 7 3
```

```
# Extract the first element in the list (which is a vector)
example_list[[1]]
```

```
## [1] 2 3 4 5
```

```
# Extract the 2nd value in that vector
example_list[[1]][2]
```

```
## [1] 3
```

```
# Extract the value in the 1st row, 3rd column of the second element of the list
example_list[[2]][1,3]
```

```
## [1] 2
```

Factor variables and levels

One other data structure that you'll see is a factor variable with levels. This is often used for categorical data. For example, suppose you collect some data where you ask people to rate their agreement with the statement, "I like coffee." Each person responds on a scale from 1-5 where 1=strongly disagree, 2=disagree, 3=no opinion, 4=agree, and 5=strongly agree. When you collect this data, you might want the numerical responses to be linked with their descriptions. Also, you probably don't want R to treat this as a continuous variable when you run a model (because a value of 2.3 is not possible). (note: factor variables can also be used for non-ordered categorical variables). The code below shows an example:

```
# Create some fake data
fake_data <- sample(c("strongly disagree", "disagree", "no opinon", "agree", "strongly agree"),
                    size=100, replace=TRUE, prob=c(.2,.3,.1,.3,.2))

# Make fake_data into a factor variable
fake_data <- as.factor(fake_data)

# Look at the first 3 elements of the fake data
fake_data[1:3]
```

```
## [1] strongly disagree strongly agree    strongly agree
## Levels: agree disagree no opinon strongly agree strongly disagree
```

```
# Inspect class of fake_data
class(fake_data)
```

```
## [1] "factor"
```

```
# Inspect the structure of fake_data
str(fake_data)
```

```
##  Factor w/ 5 levels "agree","disagree",..: 5 4 4 3 2 2 1 1 4 1 ...
```

```
# Inspect the levels of fake_data
levels(fake_data)
```

```
## [1] "agree"             "disagree"          "no opinon"
## [4] "strongly agree"    "strongly disagree"
```

Note that we have created a factor variable with 5 levels. Even though each element in the data is a character string, the class is "factor", not "character". This tells R that the variable is categorical. You can control the order of the levels in factor variables when you create them with additional parameters.

Setting up your workspace/global environment

Reading in data

There are a number of ways to read data into R, but the most common is to use the read.csv function, which can read in a .csv file and convert it to a data frame in your working environment. In order for R to read a file, you must either include the full path to the file or you must set your working directory to the location of the file. If you are planning to knit to PDF, the call to setwd() or the full file name must be included in an R chunk (instead of just running it in the Console). Another option is to use an Rproj file to automatically set your working directory to a particular location every time you open it.

Note that an easy way to get a file path is to right click on the file in your finder window and then click "get info". The info page should include a file path that you can copy and paste into R. Note that file names and path names need to be in quotes.

```
# Use setwd() to set your working directory to a particular location
# setwd("/Users/sophiesommer/Desktop/Grad School/A3SR-Welcome-Package/Tutorials")

# Now that I've set my working directory to my "Assignments" folder,
# I can read any .csv file in that location by simply writing the name in quotes:
data <- read.csv("height_sex.csv")
```

Some commonly used data sets are already available to you in base R (and others are available once you download particular packages). For example, I can load the iris dataset by simply using the data() function:

```
#load iris data set into my working environment
data(iris)

#view the structure of the iris dataset
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Finally, if you have data in other formats, you might need some other functions in R. Probably the most relevant is the read_dta function (in the "haven" package) or read.dta (in the "foreign" package). Both functions read .dta files, which is how Stata files are generally saved.

Downloading new packages and functions

To download new packages in R, you can use the install.packages() function and then use library() or require() to load the package in your workspace. Note that you will only have to use install.packages() once (unless you update R and have to delete packages for some reason); whereas you will need to re-load packages with library() or require() every time you close and re-open R Studio (assuming that you cleared your working environment). You also must include the library() or require() call within Rmd files if you are planning to knit to PDF (but you can write {r, echo=FALSE, message=FALSE} at the beginning of the chunk to surpress the output if you don't want it to show). Package names must always be in quotes. NOTE: best practices are moving toward full reference of the package everytime you call it to avoid ambiguity. E.g., psych::describe(...). In development phase, one often requires multiple libraries; in production phase, full referencing is likely to reduce errors. Consider moving in that direction early in your career.

```
# Install the psych package
# Note: I have commented out the code so that it doesnt re-install on my computer
# install.packages("psych")

# Load psych package
library(psych)
#require() does the same thing but will first check if the package is already loaded,
#and only loads it if it is not already there:
require(psych)
```

Using R as a calculator

It is helpful to know how to do some basic calculations in R (see below)

```
# Add, subtract, multiply or divide values:
5+3 #add
```

```
## [1] 8
```

```
5-2 #subtract
```

```
## [1] 3
```

```
5*3 #multiply
```

```
## [1] 15
```

```
5/2 #divide
```

```
## [1] 2.5
```

```
# Exponents and logs
5^2 #5 squared
```

```
## [1] 25
```

```
log(5) #log base e of 5
```

```
## [1] 1.609438
```

```
# Modulo
5%%2 #gives remainder of 5/2
```

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

```
# Factorial
factorial(3) #gives 3 factorial
```

```
## [1] 6
```

Many operations can also be applied to an entire vector or matrix of values.

```
# Create a vector with the numbers from 1-10
vec <- 1:10
print(vec)
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

```
# If you add, subtract, multiply, divide, exponentiate, etc.. a vector by a constant
# Then that operation is applied to every element of the vector (same for matrices)

# Add 5 to every value in a vector
vec+5
```

```
##  [1]  6  7  8  9 10 11 12 13 14 15
```

```
# Multiply every value in a vector by 2
vec*2
```

```
##  [1]  2  4  6  8 10 12 14 16 18 20
```

```
# Cube every element in a vector
vec^3
```

```
##  [1]    1    8   27   64  125  216  343  512  729 1000
```

```
# Take the factorial of every value in a vector
factorial(vec)
```

```
##  [1]       1       2       6      24     120     720    5040   40320
##  [9]  362880 3628800
```

Practice Problems

1. Creating and manipulating data
   a. Create a vector with the following values: {1,2,3,4,5,1,2,3,4,5,1,2,3,4,5,100.5,NA} and save it as my_vec
   b. Calculate the mean of my_vec
   c. Multiply all of the elements in my_vec by 2 and take the sum of the doubled values
   d. Extract the 1st, 3rd, and 10th values in my_vec and save them as my_short_vec
   e. Calculate the median of my_short_vec
   f. Create the following matrix and save it as my_mat (hover your mouse over the code if viewing in the Rmd): $\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 4 & 4 & 0 \end{bmatrix}$
   g. Find the transpose of my_mat
   h. Find the inverse of my_mat
   i. Add 3 to every value of my_mat and save the result as my_mat2
   j. matrix multiply my_mat by my_mat2
   k. Create a data frame with two columns: the first column is called "name"; the second column is called "age". The data frame should have 5 rows, where each row corresponds to the following 5 people: James (age 12), Sara (age 25), Jen (age 50), Ellie (age 64), and Mike (age 30). Save the data frame as my_df
   l. Print the first two rows of my_df
   m. Print only the rows of my_df that correspond to people who are under the age of 30
   n. Create a list called my_list which contains 3 elements: my_vec, my_mat, and my_df (in that order)
   o. Use double brackets to extract the 3rd value in the first element of my_list

2. Inspecting a dataset
   a. Download the "carData" package and load it into your workspace
   b. Use the data() function to load the "TitanicSurvival" dataset from the "carData" package
   c. Inspect the structure of the TitanicSurvival dataset. How many rows and columns are there? What are the variable names and types?
   d. Print the first five rows of the dataset
   e. Print the survival status of people in the 5th, 8th, and 9th rows of the dataset
   f. Create a new data frame called "survival" which only includes age and survival status for the first 50 people in the dataset
   g. Print the first 5 rows of your new data frame ("survival")
   h. What is the mean age of passengers on the ship? How old was the oldest person on the ship? How old was the youngest person on the ship? Save your answers as mean_age, oldest_age, and youngest_age, respectively.
   i. Did the youngest person on the ship survive? Did the oldest person?
   j. How many males and females were on the ship?
   k. How many people on the ship meet the following conditions?: female AND older than 30
   l. How many people on the ship meet the following conditions?: male AND survived
   m. What is the mean age of females on the ship?

Answers

1. Creating and manipulating data

a. Create a vector with the following values: {1,2,3,4,5,1,2,3,4,5,1,2,3,4,5,100,NA} and save it as my_vec

```
my_vec <- c(rep(1:5, times=3), 100, NA)
```

b. Calculate the mean of my_vec

```
mean(my_vec, na.rm=TRUE)
```

```
## [1] 9.0625
```

c. Multiply all of the elements in my_vec by 2 and take the sum of the doubled values

```
sum(my_vec*2, na.rm=TRUE)
```

```
## [1] 290
```

d. Extract the 1st, 3rd, and 10th values in my_vec and save them as my_short_vec

```
my_short_vec <- my_vec[c(1,3,10)]
```

e. Calculate the median of my_short_vec

```
median(my_short_vec)
```

```
## [1] 3
```

f. Create the following matrix and save it as my_mat (hover your mouse over it to view it in the Rmd):

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 4 & 4 & 0 \end{bmatrix}$$

```
my_mat <- matrix(c(1:6,4,4,0),nrow=3, byrow=TRUE)
```

g. Find the transpose of my_mat

```
t(my_mat)
```

```
##      [,1] [,2] [,3]
## [1,]    1    4    4
## [2,]    2    5    4
## [3,]    3    6    0
```

h. Find the inverse of my_mat

```
solve(my_mat)
```

```
##            [,1]       [,2]  [,3]
## [1,] -2.0000000  1.0000000 -0.25
## [2,]  2.0000000 -1.0000000  0.50
## [3,] -0.3333333  0.3333333 -0.25
```

i. Add 3 to every value of my_mat and save the result as my_mat2

```
my_mat2 <- my_mat+3
```

j. matrix multiply my_mat by my_mat2

```
my_mat%*%my_mat2
```

```
##      [,1] [,2] [,3]
## [1,]   39   42   33
## [2,]   93  102   87
## [3,]   44   52   60
```

k. Create a data frame with two columns: the first column is called "name"; the second column is called "age". The data frame should have 5 rows, where each row corresponds to the following 5 people: James (age 12), Sara (age 25), Jen (age 50), Ellie (age 64), and Mike (age 30). Save the data frame as my_df

```
my_df <- data.frame(name=c("James", "Sara", "Jen", "Ellie", "Mike"),
                    age=c(12,25,50,64,30))
```

l. Print the first two rows of my_df

```
my_df[1:2,]
```

```
##    name age
## 1 James  12
## 2  Sara  25
```

m. Print only the rows of my_df that correspond to people who are under the age of 30

```
my_df[my_df$age<30,]
```

```
##    name age
## 1 James  12
## 2  Sara  25
```

n. Create a list called my_list which contains 3 elements: my_vec, my_mat, and my_df (in that order)

```
my_list <- list(my_vec, my_mat, my_df)
```

o. Use double brackets to extract the 3rd value in the first element of my_list

```
my_list[[1]][3]
```

```
## [1] 3
```

2. Inspecting a dataset

a. Download the "carData" package and load it into your workspace

```
#uncomment the line below and run in the Console to install the package
#install.packages("carData")
library(carData)
```

b. Use the data() function to load the "TitanicSurvival" dataset from the "carData" package

```
data(TitanicSurvival)
```

c. Inspect the structure of the TitanicSurvival dataset. How many rows and columns are there? What are the variable names and types? There are 1309 observations of 4 variables: survived, sex, age, and passengerClass. age is numeric and the other three variables are factors. survived has two levels ("yes" or "no"), sex has two levels (female" or "male"), and passengerClass has 3 levels ("1st", "2nd", or "3rd")

```
str(TitanicSurvival)
```

```
## 'data.frame':    1309 obs. of  4 variables:
##  $ survived      : Factor w/ 2 levels "no","yes": 2 2 1 1 1 2 2 1 2 1 ...
##  $ sex           : Factor w/ 2 levels "female","male": 1 2 1 2 1 2 1 2 1 2 ...
##  $ age           : num  29 0.917 2 30 25 ...
##  $ passengerClass: Factor w/ 3 levels "1st","2nd","3rd": 1 1 1 1 1 1 1 1 1 1 ...
```

d. Print the first five rows of the dataset

```
TitanicSurvival[1:5,]
```

```
##                                survived    sex     age passengerClass
## Allen, Miss. Elisabeth Walton       yes female 29.0000            1st
## Allison, Master. Hudson Trevor      yes   male  0.9167            1st
## Allison, Miss. Helen Loraine         no female  2.0000            1st
## Allison, Mr. Hudson Joshua Crei      no   male 30.0000            1st
## Allison, Mrs. Hudson J C (Bessi      no female 25.0000            1st
```

e. Print the survival status of people in the 5th, 8th, and 9th rows of the dataset

```
TitanicSurvival$survived[c(5,8,9)]
```

```
## [1] no  no  yes
## Levels: no yes
```

f. Create a new data frame called "survival" which only includes survival status and age for the first 50 people in the dataset

```
survival <- TitanicSurvival[1:50,c("survived", "age")]
survival <-TitanicSurvival[1:50,c(1,3)] #does the same thing as above
```

g. Print the first 5 rows of your new data frame ("survival")

```
survival[1:5,]
```

```
##                                survived     age
## Allen, Miss. Elisabeth Walton       yes 29.0000
## Allison, Master. Hudson Trevor      yes  0.9167
## Allison, Miss. Helen Loraine         no  2.0000
## Allison, Mr. Hudson Joshua Crei      no 30.0000
## Allison, Mrs. Hudson J C (Bessi      no 25.0000
```

h. What is the mean age of passengers on the ship? How old was the oldest person on the ship? How old was the youngest person on the ship? Save your answers as mean_age, oldest_age, and youngest_age, respectively.

```
mean_age <- mean(TitanicSurvival$age, na.rm=TRUE) #mean
oldest_age <- max(TitanicSurvival$age, na.rm=TRUE) #min
youngest_age <- min(TitanicSurvival$age, na.rm=TRUE) #max

#print the ages
mean_age
```

```
## [1] 29.88113
```

```
oldest_age
```

```
## [1] 80
```

```
youngest_age
```

```
## [1] 0.1667
```

i. Did the youngest person on the ship survive? Did the oldest person?

```
TitanicSurvival$survived[which(TitanicSurvival$age==youngest_age)] #yes!
```

```
## [1] yes
## Levels: no yes
```

```
TitanicSurvival$survived[which(TitanicSurvival$age==oldest_age)] #yes!
```

```
## [1] yes
## Levels: no yes
```

j. How many males and females were on the ship?

```
# We could take advantage of TRUE=1 and use the sum function
sum(TitanicSurvival$sex=="female", na.rm=TRUE)
```

```
## [1] 466
```

```
sum(TitanicSurvival$sex=="male", na.rm=TRUE)
```

```
## [1] 843
```

```
#Or, we could just use table()
table(TitanicSurvival$sex)
```

```
##
## female   male
##    466    843
```

k. How many people on the ship meet the following conditions?: female AND older than 30

```
nrow(TitanicSurvival[TitanicSurvival$sex=="female" & TitanicSurvival$age > 30,])
```

```
## [1] 231
```

l. How many people on the ship meet the following conditions?: male AND survived

```
nrow(TitanicSurvival[TitanicSurvival$sex=="male" & TitanicSurvival$survived=="yes",])
```

```
## [1] 161
```

m. What is the mean age of females on the ship?

```
mean(TitanicSurvival$age[TitanicSurvival$sex=="female"], na.rm=TRUE)
```

```
## [1] 28.68707
```