*\*Please save this file as “LAST NAME\_Assignment 1.docx”*

***Open-Ended Responses***

1. **I.1** If you are defining an object called **vec.x**, you can do so by typing it into a script file first and then executing or by typing it directly into the console. Which way is better for reproducibility and why?

**Type and then execute because of reproducibility**

|  |
| --- |
|  |

1. **I.5** Explain how R came up with the following result:

x <- 1:10

y <- 1:3

x-y

[1] 0 0 0 3 3 3 6 6 6 9

Warning message:

In x - y : longer object length is not a multiple of shorter object length

It subtracts the elements one by one, and it starts over in a repetition pattern. E.g. 1-1, 2-2,3-3, 4-1,5-2,6-3,7-1,8-2,9-3, and 10-1.

|  |
| --- |
|  |

1. **I.6** Explain the behavior of the **round()** function observed below where 0.5 is rounded down, but 1.5 is rounded up.

round(.5)

[1] 0

round(1.5)

[1] 2

It rounds to even number

|  |
| --- |
|  |

1. **I.7** A researcher wants to create a data set by sampling 100 integers ranging from 50 to 75 with replacement, center those data (subtract the mean from each data point), and then calculate the centered mean divided by the centered standard deviation. Spot the silent error in the following code written to do this:

1 data <- sample(50:75, 100, replace = TRUE) # sample the data

2 data.cen <- data - mean(data) # center data on mean

3 mean(data) / sd(data) # calculate mean / sd

The researcher did not divide the mean and the sd from the center data

|  |
| --- |
|  |

1. **I.7** Why does the following generate an error? Special note: if you copy/paste from this word document, it brings what are called “smart quotes” which R can’t recognize. You might need to type the quotes manually.

x <- c(1, 5, 3, 4, “3”) # runs fine

sum(x) # error generated

Because it evaluates it as “character”. Typeof(x) “claracter”

|  |
| --- |
|  |

1. **1.8** Below are two sections of code that accomplish the same thing. Which one would you think is the “best” way? There isn’t necessarily a right answer here, I just want to hear your rationale.

x <- sample(-50:50)

# Code Section 1

x.fil <- x[x > 0] # filter out all negative values

x.fil.sq <- x.fil^2 # square results

mean(x.fil.sq) # compute mean

# Code Section 1

mean((x[x > 0])^2) # filter, square, compute mean

I would go with the first code because you can easily spot the error and check your work in detailed compared to the second one. In other word code 1 is more paying attention to all details of the code.

|  |
| --- |
|  |

***Coding Section***

To complete this section, start a new script file with the following layout:

# YOUR NAME

# Assignment 1 Introduction to R

# #1 ---------------- (new section: CTRL + SHIFT + R)

here’s my code # with adequate commenting

# #2 ---------------- (new section: CTRL + SHIFT + R)

here’s my code # with adequate commenting

1. **I.2** Suppose I wrote 2 lines of code and then said the following: “Here, I defined an object x as a numeric vector that contains 5 numeric elements, the numbers 1 through 5. Then I told R to add 1 to each element and print the result.” What are the two lines of code?
2. **I.3** Cohen’s d is a metric that computes the effect size in a comparisons test (if you don’t know what I’m talking about, it’s okay). The formula is:

Graphical user interface

Description automatically generated with medium confidence

, where...

* d is Cohen’s d
* ME, MC are the means of experimental (E) and control (C) groups
* Sample SD pooled is the pooled standard deviation:

Diagram

Description automatically generated

Calculate Cohen’s d in R comparing the two simulated groups below. I would recommend using mean() and sd() to compute means and standard deviations (as opposed to doing them “by hand”). Note: there are functions in other packages that compute Cohen’s d for you, but do not use one of these functions here.

set.seed(42)

exp <- rnorm(100, 1.1, .1) # experimental

set.seed(42)

con <- rnorm(100, 1, .1) # control

1. **I.5** Create an object called **data** and define it as a numeric sequence that starts at 0 and goes to 200 in increments of 0.5 (*i.e.* 0, 0.5, 1, 1.5... ...199, 199.5, 200). Then, take a random sample of 50 points without replacement (cannot sample the same set of points more than once) and assign it to an object called **dat.sample**. Set the seed to 42 prior to sampling so we get the same result. Calculate a 5-number summary of **dat.sample** (minimum, 1st quartile, median, 3rd quartile, and maximum).
2. **I.8** Binning numbers is a pretty common task in research that entails taking a numeric vector and binning them into categories. The following will simulate this process commonly found in educational testing.
   1. Simulate a dataset that contains 200 students’ scores that follow a normal distribution (?rnorm) that have a class average of 80% and a standard deviation of 20%. Use a seed of 42 so we get the same data. You will notice that many of the sampled scored go above 100, which is not possible. Replace any number over 100 with 100.
   2. What is the average and standard deviation of just the students in the top third of the class?
   3. Create a vector that bins scores according to a common grade distribution (below) and tally how many of each grade were given
      * A >= 90%
      * B >= 80%, < 90%
      * C >= 70%, < 80%
      * D >= 60%, < 70%
      * F < 60%

In case the script file did not work,

# Yousif Muten

# Assignment 1 Introduction to R

# Question (1) 1.2 --------------------------------------------------------------

x<-c(1:5)+1 # creat a vector contain elements and add 1 to each element

print(x) # print the vector's values resulting from the previous command

# Question (2) 1.3 ------------------------------------------------------------

set.seed(42) #set the starting number to get the same data

exp<-rnorm(100, 1.1, .1) # stimulate group that follow normal distribution

set.seed(42) # set the starting number to get the same data

con<-rnorm(100, 1, .1) ## stimulate group that follow normal distribution

mean(exp)# calculate the mean of exp

sd(exp) # calculate the sd of exp

mean(con) # calculate the mean of con

sd(con) # calculate the sd of con

SDpooled <- sqrt((sd(exp)^2+sd(con)^2)/2) # manually calculating the SDpooled

print(SDpooled) # print

d<-((mean(exp)-mean(con))/SDpooled) # manually calculating the cohen's d

# Question (3) 1.5 --------------------------------------------------------

data<-c(seq(0,200, by=0.5)) # create an object 0:200 in increments of .5

set.seed(42) # set the starting number

dat.sample<-sample(data, 50) # take and assign a random number of 50 points without replacement to an object called dat.sample

summary(dat.sample) # to compute summary statistics of the data from the object

# Question (4) 1.8 --------------------------------------------------------

set.seed(42) #set the starting number to get the same data

dataset<-rnorm(200, 80, 20) # sti,ulate a dataset contains 200 students with a mean of 80, and sd of 20

dataset<-replace(dataset, dataset>100, 100) # replace any value over 100 with 100

table(dataset>100) # to confirm than there are no data larger than 100

quantile(dataset, probs = seq(0, 1 , 1/3)) # to find the third top

mean(dataset[dataset>=quantile(dataset, 0.6666667, na.rm = T)], na.rm=T) # to calculate the mean using the third top quantile from 0.6666667-1 which correspond to the top third

sd(dataset[dataset>=quantile(dataset, 0.6666667, na.rm = T)], na.rm=T) # to calclate the sd using the third top quantile from 0.6666667-1 which correspond to the top third

dataset\_scores<-c(dataset)# create new vector as a second version of the previous vector but different name

dataset\_grade<-c(case\_when(dataset\_scores >= 90 ~ 'A', dataset\_scores >=80 ~ 'B', dataset\_scores >=70 ~ 'C', dataset\_scores >= 60 ~ 'D', dataset\_scores <60 ~ 'F'))# use case\_when()on the vector and use multiple cases using the following value

df<-data.frame(dataset\_scores, dataset\_grade)# create dataframe to bin scores to grade (cut function was not easy to use and do not have grade packages

table(dataset\_grade) # count each grade in the vector