edX Microsoft : DAT209x Programming with R for Data Science

Notebook shortcuts

'ESC' to get to command prompt
A to insert a cell above the current cell
B to insert a cell below the current cell
M to set the cell Markdown
enter to enter into the cell
Add two spaces at the end to make a line break
'#' largest heading
'####' smaller heading

Introduction

In [8]:
```
set.seed(476)
```

In [9]:
```
x <- rnorm(100)
```

In [10]:
```
head(x)
```

-0.259658827115081 -0.543928748781071 -0.397645938992977
-0.805136600320548 -0.885429804502682 -0.131783413719148

In [11]:
```
mean(x)
sd(x)
min(x)
max(x)
```

0.0461312650105662

1.01318515246851

-3.20791594624788

2.55092278500848

Basic Operations

In [12]:
```
2+2
7*17
sqrt(9)
3^3
log(7)
log10(7)
```

4

119

3

27

1.94591014905531

0.845098040014257

Precision

In [15]:
```
sin(pi/2)
pi
options(digits=22)
pi
```

1

3.14159265358979

3.14159265358979

Infinity or not defined, and missings

In [16]:
```
1/0
2*Inf
-1/0
0/0
c(1,2,3,NA,5)
mean(c(1,2,3,NA,5))
```

Inf

Inf

-Inf

NaN

1 2 3 NA 5

[1] NA

Assignments to varaibles

In [17]:
```
rm(list=ls())
options(digits=7)
x <- 5
x
#x=5 can be used; not recormmended
x * x
y <- x+5
ls()
rm(x)
ls()
```

```
5

25

    'x'  'y'

'y'
```

Internal help function

In [18]:
```
?mean   # shorthand for help(mean)
example(mean)
??"fitting linear model"  # shorthand for help.search("fitting linear model")
 manuals
help.start()
```

```
mean> x <- c(0:10, 50)

mean> xm <- mean(x)

mean> c(xm, mean(x, trim = 0.10))
[1] 8.75 5.50

starting httpd help server ... done

If nothing happens, you should open
'http://127.0.0.1:28098/doc/html/index.html' yourself
```

Exercise 1.1

Let's try to use R to solve a simple mathematical equation. X is normally distributed with some values for mean and variance, say, X~N (μ, σ2). Which value is the mean of Y = e^X ?

a) e^(μ+σ2/2)

b) e^(μ−σ2/2)

Use the rnorm() function and the other functions from the first slide in the introduction session to figure it out. HINT: The mean μ and variance σ2 of the standard normal distribution are 0 and 1 respectively.

Question 1

Which option is closer to the mean of Y? a) or b)? : a b neither both

The standard rnorm() function corresponds to mean μ = 0 and variance σ^2 = 1. Using the mean() function, we can simulate the mean of Y with the following code:

a) e^(μ+σ^2/2) # μ = 0 and variance σ^2 = 1, exp(0 + 1/2)=0.6065307

b) e^(μ−σ2/2) # exp(0-1/2) = 1.648721

mean(exp(rnorm(1000)))

In [22]:
```
exp(0-1/2)
exp(0+1/2)
x <- rnorm(1000)
y <- exp(x)
mean(y)      # Which returns a value of about 1.6-1.7, bigger than 1. Therefor
e, the solution is a
```

0.606530659712633

1.64872127070013

1.59788807352942

Discussion

What the heck is exp()?

I fell at the first hurdle! :) It has been a REALLY long time since I did formal Statistics and Mathematics and although I would like to think I am reasonably good as grasping the concepts I think I might struggle with the actual maths, especially the symbolic representations. To save someone else the headache I thought I would post a comment here. I didn't understand why exp(0-1/2) gives 0.6 and exp(0+1/2) gives 1.6. exp() is a computer notation for the e^x where e is acutally Eulers constant (approximately 2.718). the application of the larger the factor, the more distance between e and the factor result which effectively drags up the mean of the factored results so the results will gravitate towards the e^0.5 rather than towards e^-0.5 I hope I have that right, if not I am sure someone will correct me.

Exercise 1.2

The general polynomial equation AX^2 + BX + C = 0 has the solutions -B + sqrt(B^2 -4AC)/2A or -B - sqrt(B^2 -4AC)/2A

Now, let say you have the following polynomial equation: X^2 + 3X + 1 = 0

1. Construct a vector of length 2 that contains the solutions to the equation (a), and display it on the screen with 1 decimal point.

2.Work out how much error you make (in percent) by referring to the solutions with just one decimal. The result will be different for the two solutions.

Question : Which are the possible solutions for task 1? -0.4, 0.4, 2.6, -2.6

In [20]:
```
# hint : In our case, we have A=1, B=3 and C=1. We can thus construct the vect
or of the solutions as follows:
# A <- 1, B <- 3, C <- 1
# my.vector<-c((-B+sqrt(B^2-4*A*C))/(2*A),(-B-sqrt(B^2-4*A*C))/(2*A))
# Typing the identifier in the R command prompt give the result: my.vector [1]
 -0.381966 -2.618034
```

In [21]:
```
A <- 1
B <- 3
C <- 1
my.vector<-c((-B+sqrt(B^2-4*A*C))/(2*A),(-B-sqrt(B^2-4*A*C))/(2*A))
my.vector
```

-0.381966011250105 -2.61803398874989

Exercise 1.3

Construct the object x as follows:

x <- rnorm(100, mean=.5, sd=.3)

Perform the following task (check out the help system as needed):

1.Calculate mean and standard deviation of x.

2.Plot a histogram of x.

3.Put the second axis on the right side of the histogram plot instead of on the left.

Question

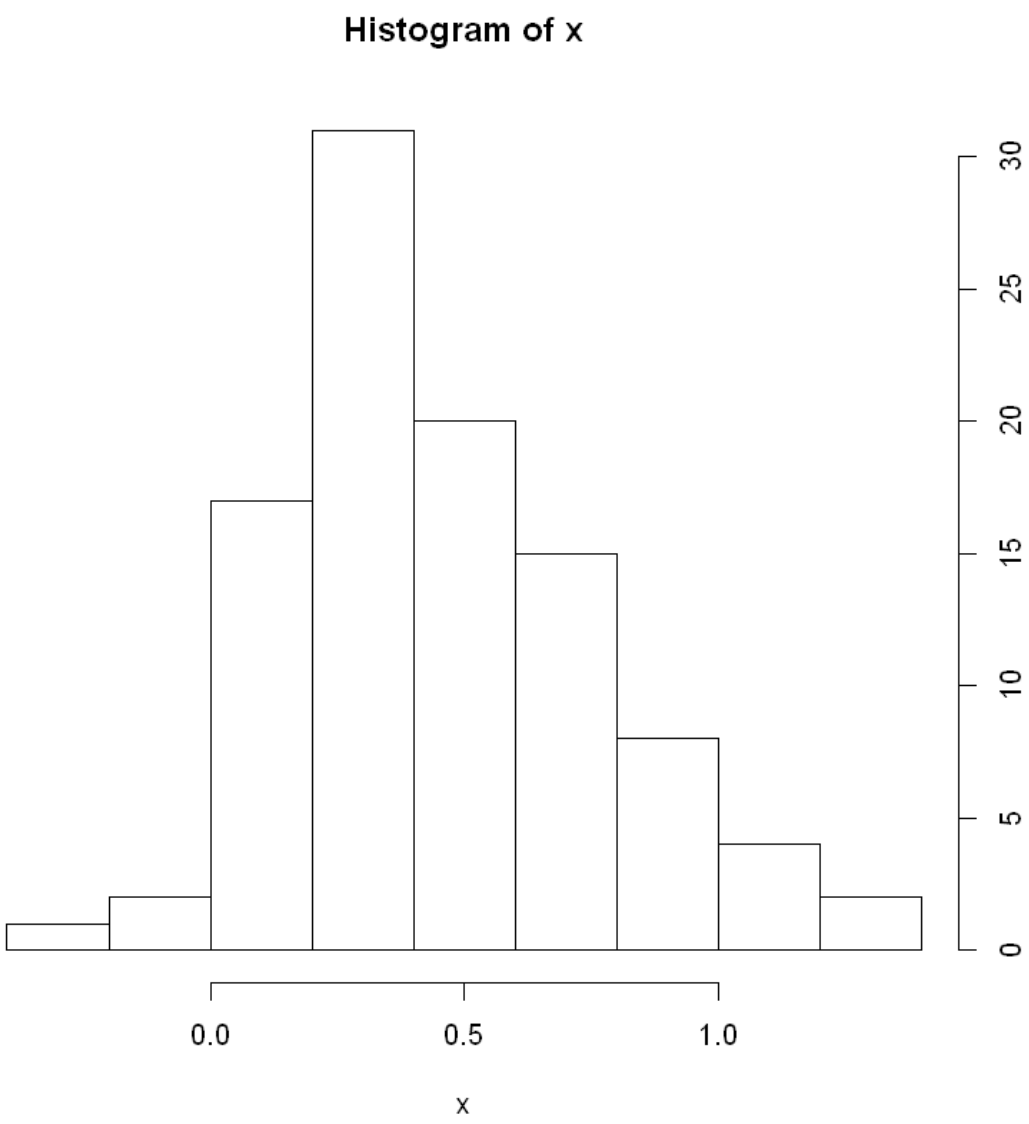
If you set the seed as 1234, run set.seed(1234), prior to assigning the variables to x, which options represent the mean and standard deviation for task 1? (one option is the mean, the other is the standard deviation).

0.4529715, 0.3013216, 0.5123424, 0.2995845

In [36]:
```
set.seed(1234)
x <- rnorm(100, mean=.5, sd=.3)
mean(x)
sd(x)
hist(x, axes=FALSE, ylab="")   # ??histogram
axis(4)     # axis is placed as follows 1=below, 2=left, 3-above, and 4=right
axis(1)
```

0.452971477267388

0.301321590916126



Discussion

Axes and axis are still confused in the instructions above

- Axes is is the plural form of the word axis
- Google says that axes is the only word in English that can be the plural of three different singular noun forms: ax, axe, and axis...

Quiz

Question 1 You are examining the following code

set.seed(1)

x <- rnorm(100)

What will be printed in the console when you run head(x)?

The first 5, or 6 of the elments of the vector x? or the last 5 or 6 ?

Answer: first 6

Question 2

You want to open the review the internal help documentation for the tail() function. Answer: help(tail), ?tail, ?"tail"

Question 3

You are examing the cose x <- 5:6 What will be returned when you show the value x in the console? Answer : 5 6

Question 4

You are examing the following code

```
x <- 5 + 6
y <- x + 3
z <- y - 10
```

What is the value of z after you run the code? Answer: 4

Question 5
Which function can be used to list all the R objects stored in the working memory?
lm(), lst(), list(), ls()

Answer: ls()

In [37]:
```
set.seed(1)
x <- rnorm(100)
head(x)
```

-0.626453810742332 0.183643324222082 -0.835628612410047
1.59528080213779 0.329507771815361 -0.820468384118015

In [41]:
```
# ?tail
# ?"tail"
# help(tail)
# internal(tail) # Incorrect
```

```
Error in eval(expr, envir, enclos): could not find function "internal"
Traceback:
```

In [43]:
```
x <- 5:6
x
```

5 6

In [44]:
```
x <- 5 + 6
y <- x + 3
z <- y - 10
z
```

4

Lab

In this very first assignment, you need to perform the following tasks:

1. Assign the first five positive odd numbers to a vector named A.
2. Assign the mean of vector A to variable B.
3. Assign the first five positive even numbers (zero excluded) to a vector named X.
4. Add vector A and X and assign the result to variable Z.

In [45]:
```
# A <- 1:5  wrong
A <- seq(1,10,2)
B <- mean(A)
X <- seq(2,10,2)
Z <- A + X
Z
```

3 6 9 12 15