Lab 1

Elizabeth McHugh

11:59PM February 18, 2021

You should have RStudio installed to edit this file. You will write code in places marked "TO-DO" to complete the problems. Some of this will be a pure programming assignment. The tools for the solutions to these problems can be found in the class practice lectures. I want you to use the methods I taught you, not for you to google and come up with whatever works. You won't learn that way.

To "hand in" the homework, you should compile or publish this file into a PDF that includes output of your code. Once it's done, push by the deadline to your repository in a directory called "labs".

- Print out the numerical constant pi with ten digits after the decimal point using the internal constant `pi`.

```
options(digits = 11)
x <- pi
x
```

```
[1] 3.1415926536
```

- Sum up the first 103 terms of the series $1 + 1/2 + 1/4 + 1/8 + \ldots$

```
sum(1 / (2^(0:102)) )
```

```
[1] 2
```

- Find the product of the first 37 terms in the sequence $1/3, 1/6, 1/9 \ldots$

```
prod(1 / (3 * (1:37)) )
```

```
[1] 1.61e-61
```

```
prod(1/seq(from = 3, by = 3, length.out = 37))
```

```
[1] 1.613528728e-61
```

- Find the product of the first 387 terms of `1 * 1/2 * 1/4 * 1/8 *` $\ldots$

```
prod(1/(2^(0:386)))
```

```
[1] 0
```

Is this answer *exactly* correct?

No. Since the sequence 1/2^(n) approaches 0 as n approaches infinity, the product would not be exactly 0 were we to stop before reaching an infinite number of terms. However, once we choose a large enough 'n', that is "enough" terms, the resulting fraction is so small that, when added in the product, the computer stops computing decimal places.

- Figure out a means to express the answer more exactly. Not compute exactly, but express more exactly.

```
sum(log(1/(2^(0:386))))
```

```
[1] -51771.856063
```

```
-log(2)*sum(0:386)
```

```
[1] -51771.856063
```

- Create the sequence `x = [Inf, 20, 18, ..., -20]`.

```
c(Inf, seq(from = 20, to = -20, by = -2))
```

```
 [1] Inf  20  18  16  14  12  10   8   6   4   2   0  -2  -4  -6  -8 -10 -12 -14 -16 -18 -20
```

Create the sequence `x = [log_3(Inf), log_3(100), log_3(98), ... log_3(-20)]`.

```
x = c(Inf, seq(from = 100, to = -20, by = -2))

log(x, base = 3)
```

```
NaNs produced
```

```
 [1]           Inf 4.19180654858 4.17341725189 4.15464876786 4.13548512895 4.11590933734 4.09590327429 4
[10] 4.03310325630 4.01116871959 3.98869253500 3.96564727304 3.94200336639 3.91772888179 3.89278926071 3
[19] 3.81358809222 3.78557852143 3.75667961083 3.72683302786 3.69597450568 3.66403300988 3.63092975357 3
[28] 3.52371901429 3.48497958377 3.44451784579 3.40217350273 3.35776278143 3.31107361282 3.26185950714 3
[37] 3.09590327429 3.03310325630 2.96564727304 2.89278926071 2.81358809222 2.72683302786 2.63092975357 2
[46] 2.26185950714 2.09590327429 1.89278926071 1.63092975357 1.26185950714 0.63092975357          -Inf
[55]           NaN           NaN           NaN           NaN           NaN           NaN           NaN
```

Comment on the appropriateness of the non-numeric values.

x[1]: Inf is appropriate since the limit approaching infinity of log is infinity.

x[52]: -Inf is appropriate since the limit approaching 0 of log is negative infinity.

x[53: 62]: NaN is appropriate, as log is undefined for all non-negative integers.

- Create a vector of booleans where the entry is true if `x[i]` is positive and finite.

```
y = c(!is.nan(x) & !is.infinite(x) & x > 0)
y
```

```
 [1] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TR
[22]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TR
[43]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
```

- Locate the indices of the non-real numbers in this vector. Hint: use the `which` function. Don't hesitate to use the documentation via `?which`.

```
which(!y)
```

```
 [1]  1 52 53 54 55 56 57 58 59 60 61 62
```

- Locate the indices of the infinite quantities in this vector.

```
which(is.infinite(x))
```

```
[1] 1
```

- Locate the indices of the min and max in this vector. Hint: use the `which.min` and `which.max` functions.

```
which.min(x)
```

```
[1] 62
```

```
which.max(x)
```

```
[1] 1
```

- Count the number of unique values in `x`.

```
length(unique(x))
```

```
[1] 62
```

- Cast `x` to a factor. Do the number of levels make sense?

```
factor(x)
```

```
 [1] Inf 100 98  96  94  92  90  88  86  84  82  80  78  76  74  72  70  68  66  64  62  60  58  56  54
[33] 38  36  34  32  30  28  26  24  22  20  18  16  14  12  10  8   6   4   2   0   -2  -4  -6  -8  -1
62 Levels: -20 -18 -16 -14 -12 -10 -8 -6 -4 -2 0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 4
```

- Cast `x` to integers. What do we learn about R's infinity representation in the integer data type?

```
as.integer(x)
```

```
NAs introduced by coercion to integer range
```

```
 [1]  NA 100  98  96  94  92  90  88  86  84  82  80  78  76  74  72  70  68  66  64  62  60  58  56  54
[33]  38  36  34  32  30  28  26  24  22  20  18  16  14  12  10   8   6   4   2   0  -2  -4  -6  -8 -10
```

- Use `x` to create a new vector `y` containing only the real numbers in x.

```
y = x[!is.nan(x) & is.finite(x) ]
y
```

```
 [1] 100  98  96  94  92  90  88  86  84  82  80  78  76  74  72  70  68  66  64  62  60  58  56  54  52
[33]  36  34  32  30  28  26  24  22  20  18  16  14  12  10   8   6   4   2   0  -2  -4  -6  -8 -10 -12
```

- Use the left rectangle method to numerically integrate x^2 from 0 to 1 with rectangle width size 1e-6.

```
sum(seq(from = 0, to = 1 - 1e-6, by = 1e-6)^2) * 1e-6
```

```
[1] 0.33333283333
```

- Calculate the average of 100 realizations of standard Bernoullis in one line using the `sample` function.

```
sum( sample(c(0, 1), size = 100, replace = TRUE) ) / 100
```

```
[1] 0.47
```

- Calculate the average of 500 realizations of Bernoullis with p = 0.9 in one line using the `sample` and `mean` functions.

```
mean(sample(c(0, 1), size = 500, replace = TRUE, prob = c(0.1, 0.9)))
```

```
[1] 0.906
```

- Calculate the average of 1000 realizations of Bernoullis with p = 0.9 in one line using `rbinom`.

```
mean(rbinom(n = 1000, size = 1, prob = 0.9))
```

```
[1] 0.916
```

- In class we considered a variable `x_3` which measured "criminality". We imagined L = 4 levels "none", "infraction", "misdimeanor" and "felony". Create a variable `x_3` here with 100 random elements (equally probable). Create it as a nominal (i.e. unordered) factor.

```
x_3 = as.factor(sample(c("none", "infraction", "misdemeanor", "felony"), size = 100, replace = TRUE))

x_3
```

```
  [1] felony      infraction  none        misdemeanor felony      felony      infraction  felony      fe
 [11] felony      infraction  felony      felony      none        none        infraction  infraction  fe
 [21] infraction  misdemeanor none        felony      felony      felony      infraction  misdemeanor in
 [31] infraction  infraction  felony      none        none        none        misdemeanor felony      fe
 [41] felony      none        felony      misdemeanor felony      felony      felony      misdemeanor mi
 [51] misdemeanor felony      none        infraction  misdemeanor misdemeanor felony      felony      no
 [61] misdemeanor misdemeanor none        felony      misdemeanor felony      felony      none        mi
 [71] infraction  none        misdemeanor misdemeanor none        none        none        felony      no
 [81] felony      infraction  none        felony      infraction  felony      infraction  felony      no
 [91] infraction  infraction  infraction  felony      felony      infraction  infraction  none        fe
Levels: felony infraction misdemeanor none
```

- Use `x_3` to create `x_3_bin`, a binary feature where 0 is no crime and 1 is any crime.

```
as.numeric(x_3_bin)
```

```
  [1] 1 1 0 1 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 0 0 0 1 1 1 1 1 0 1 1 1 1 1 1 1
 [55] 1 1 1 1 0 1 1 1 0 1 1 1 1 0 1 1 1 0 1 1 0 0 0 1 0 1 1 1 0 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1 1
```

- Use `x_3` to create `x_3_ord`, an ordered factor variable. Ensure the proper ordinal ordering.

```
x_3_ord = factor(x_3, levels = c("none", "infraction", "misdemeanor", "felony"), ordered = TRUE)

x_3_ord
```

```
  [1] felony      infraction  none        misdemeanor felony      felony      infraction  felony      fe
 [11] felony      infraction  felony      felony      none        none        infraction  infraction  fe
 [21] infraction  misdemeanor none        felony      felony      felony      infraction  misdemeanor in
 [31] infraction  infraction  felony      none        none        none        misdemeanor felony      fe
 [41] felony      none        felony      misdemeanor felony      felony      felony      misdemeanor mi
 [51] misdemeanor felony      none        infraction  misdemeanor misdemeanor felony      felony      no
 [61] misdemeanor misdemeanor none        felony      misdemeanor felony      felony      none        mi
 [71] infraction  none        misdemeanor misdemeanor none        none        none        felony      no
 [81] felony      infraction  none        felony      infraction  felony      infraction  felony      no
 [91] infraction  infraction  infraction  felony      felony      infraction  infraction  none        fe
Levels: none < infraction < misdemeanor < felony
```

- Convert this variable into three binary variables without any information loss and put them into a data matrix.

```
x_3_infraction = as.integer( x_3_ord == "infraction" )
x_3_misdemeanor = as.integer( x_3_ord == "misdemeanor")
x_3_felony = as.integer( x_3_ord =="felony" )

X = cbind(x_3_infraction, x_3_misdemeanor, x_3_felony)
colnames(X) = list("Infraction", "Misdemeanor", "Felony")

head(X)
```

```
     Infraction Misdemeanor Felony
[1,]          0           0      1
[2,]          1           0      0
[3,]          0           0      0
[4,]          0           1      0
[5,]          0           0      1
[6,]          0           0      1
```

- What should the sum of each row be (in English)?

The sum of each row should be tell us whether or not each subject had been charged with a crime (0 for no crime, 1 for crime).

Verify that.

```
rowSums(X, na.rm = FALSE, dims = 1)
```

```
  [1] 1 1 0 1 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 0 0 0 1 1 1 1 1 0 1 1 1 1 1 1
 [55] 1 1 1 1 0 1 1 1 0 1 1 1 1 0 1 1 1 0 1 1 0 0 0 1 0 1 1 1 0 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1 1
```

- How should the column sum look (in English)?

The sum of each column should give the number of occurrences of each type of crime.

Verify that.

```
colSums (X, na.rm = FALSE, dims = 1)
```

```
 Infraction Misdemeanor      Felony
         23          18          38
```

- Generate a matrix with 100 rows where the first column is realization from a normal with mean 17 and variance 38, the second column is uniform between -10 and 10, the third column is poisson with mean 6, the fourth column in exponential with lambda of 9, the fifth column is binomial with n = 20 and p = 0.12 and the sixth column is a binary variable with exactly 24% 1's dispersed randomly. Name the rows the entries of the `fake_first_names` vector.

```
fake_first_names = c(
  "Sophia", "Emma", "Olivia", "Ava", "Mia", "Isabella", "Riley",
  "Aria", "Zoe", "Charlotte", "Lily", "Layla", "Amelia", "Emily",
  "Madelyn", "Aubrey", "Adalyn", "Madison", "Chloe", "Harper",
  "Abigail", "Aaliyah", "Avery", "Evelyn", "Kaylee", "Ella", "Ellie",
  "Scarlett", "Arianna", "Hailey", "Nora", "Addison", "Brooklyn",
  "Hannah", "Mila", "Leah", "Elizabeth", "Sarah", "Eliana", "Mackenzie",
  "Peyton", "Maria", "Grace", "Adeline", "Elena", "Anna", "Victoria",
  "Camilla", "Lillian", "Natalie", "Jackson", "Aiden", "Lucas",
  "Liam", "Noah", "Ethan", "Mason", "Caden", "Oliver", "Elijah",
  "Grayson", "Jacob", "Michael", "Benjamin", "Carter", "James",
  "Jayden", "Logan", "Alexander", "Caleb", "Ryan", "Luke", "Daniel",
  "Jack", "William", "Owen", "Gabriel", "Matthew", "Connor", "Jayce",
  "Isaac", "Sebastian", "Henry", "Muhammad", "Cameron", "Wyatt",
  "Dylan", "Nathan", "Nicholas", "Julian", "Eli", "Levi", "Isaiah",
```

```
  "Landon", "David", "Christian", "Andrew", "Brayden", "John",
  "Lincoln"
)

n = length(fake_first_names)

Names = data.frame(
  normal = rnorm(n, mean = 17, sd = 38),
  uniform = runif(n, min = -10, max = 10),
  poisson = rpois(n, 6),
  exponential = rexp(n, rate = 9),
  binomial = rbinom(n, size = 20, p = 0.12),
  binary =  rbinom(n, size = 1, p = 0.24)
)


head(Names)
```

```
tail(Names)
```

- Create a data frame of the same data as above except make the binary variable a factor "DOMESTIC" vs "FOREIGN" for 0 and 1 respectively. Use RStudio's `View` function to ensure this worked as desired.

```
n = length(fake_first_names)

Names = data.frame(
  normal = rnorm(n, mean = 17, sd = 38),
  uniform = runif(n, min = -10, max = 10),
  poisson = rpois(n, 6),
  exponential = rexp(n, rate = 9),
  binomial = rbinom(n, size = 20, p = 0.12),
  binary =  rbinom(n, size = 1, p = 0.24)
)

Names$binary = factor(Names$binary, labels = c("DOMESTIC", "FOREIGN"))

head(Names)
```

```
tail(Names)
```

```
NA
```

- Print out a table of the binary variable. Then print out the proportions of "DOMESTIC" vs "FOREIGN".

```
print((Names$binary)[2,2])
```

```
Error in `[.default`((Names$binary), 2, 2) :
  incorrect number of dimensions
```

Print out a summary of the whole dataframe.

```
summary(Names)
```

```
     normal                uniform
 Min.   :-82.9847884   Min.   :-9.80188198
 1st Qu.: -8.0377319   1st Qu.:-4.48766736
 Median : 11.9964172   Median :-1.20804997
 Mean   : 15.7718494   Mean   :-0.70935411
 3rd Qu.: 38.8386927   3rd Qu.: 2.65104750
 Max.   : 86.8490803   Max.   : 9.56918845
    poisson       exponential               binomial
 Min.   : 0.00   Min.   :0.00070692469   Min.   :0.00
 1st Qu.: 4.00   1st Qu.:0.03790873563   1st Qu.:1.00
 Median : 6.00   Median :0.08137113394   Median :2.00
 Mean   : 5.97   Mean   :0.11448977910   Mean   :2.29
 3rd Qu.: 8.00   3rd Qu.:0.15338476304   3rd Qu.:3.00
 Max.   :12.00   Max.   :0.54336554326   Max.   :6.00
      binary
 DOMESTIC:78
 FOREIGN :22
```

- Let `n = 50`. Create a n x n matrix `R` of exactly 50% entries 0's, 25% 1's 25% 2's. These values should be in random locations.

```
n = 50

R = matrix(
  sample(c(0,1,2), size = n*n, replace = TRUE, prob = c(0.5, .25, .25)),
  nrow = n,
  ncol = n)

table(R)/(n*n)  #This doesn't give EXACTLY the proportions desired, so I'm not doing something correctly
```

```
R
     0      1      2
0.5064 0.2388 0.2548
```

- Randomly punch holes (i.e. `NA`) values in this matrix so that each entry is missing with probability 30%.

```
table(R)
```

```
R
   0    1    2
1250  550  700
```

- Sort the rows in matrix `R` by the largest row sum to lowest. Be careful about the NA's!

```
R_row_sums = rowSums(R)
R_row_sums
```

```
 [1]   0  50   0   0  50   0  50 100   0   0   0   0  50
[14]  50  50   0   0 100   0 100 100   0 100 100   0   0
[27] 100   0   0 100   0 100   0   0   0   0   0   0   0
[40]   0   0 100   0 100   0   0   0   0  50   0
```

```
order(R_row_sums, decreasing = FALSE)
```

```
 [1]  1  3  4  6  9 10 11 12 16 17 19 22 25 26 28 29 31 33
[19] 34 35 36 37 38 39 40 41 43 45 46 47 48 50  2  5  7 13
[37] 14 15 49  8 18 20 21 23 24 27 30 32 42 44
```

- We will now learn the `apply` function. This is a handy function that saves writing for loops which should be eschewed in R. Use the apply function to compute a vector whose entries are the standard deviation of each row. Use the apply function to compute a vector whose entries are the standard deviation of each column. Be careful about the NA's! This should be one line.

```
#TO-DO
```

- Use the `apply` function to compute a vector whose entries are the count of entries that are 1 or 2 in each column. This should be one line.

```
#TO-DO
```

- Use the `split` function to create a list whose keys are the column number and values are the vector of the columns. Look at the last example in the documentation `?split`.

- In one statement, use the `lapply` function to create a list whose keys are the column number and values are themselves a list with keys: "min" whose value is the minimum of the column, "max" whose value is the maximum of the column, "pct_missing" is the proportion of missingness in the column and "first_NA" whose value is the row number of the first time the NA appears.

```
#TO-DO
```

- Set a seed and then create a vector `v` consisting of a sample of 1,000 iid normal realizations with mean -10 and variance 100.

```
#TO-DO
```

- Repeat this exercise by resetting the seed to ensure you obtain the same results.

```
#TO-DO
```

- Find the average of `v` and the standard error of `v`.

```
#TO-DO
```

- Find the 5%ile of `v` and use the `qnorm` function to compute what it theoretically should be. Is the estimate about what is expected by theory?

```
#TO-DO
```

- What is the percentile of `v` that corresponds to the value 0? What should it be theoretically? Is the estimate about what is expected by theory?

```
#TO-DO
```