Lab 1

Hanlin Wang

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)
pi
```

```
## [1] 3.1415926536
```

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

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

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

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

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

```
## [1] 1.613528728e-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 *` ...

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

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

```
prod(1 /(seq(from = 2, by = 3, length.out = 387 )))
```

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

```
.Machine$double.xmin
```

```
## [1] 2.2250738585e-308
```

Is this answer *exactly* correct? It's not correct because the answer numerically underflow

#TO-DO

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

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

```
## [1] -51771.856063
```

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

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

```
##  [1] Inf  20  18  16  14  12  10   8   6   4   2   0  -2  -4  -6  -8 -10 -12 -14
## [20] -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))
X = log(X, base = 3)
```

```
## Warning: NaNs produced
```

Comment on the appropriateness of the non-numeric values.

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

```
Y = is.nan(X) & is.finite(X) & X > 0
```

- 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
```

```
## starting httpd help server ... done
```

```
which(Y == FALSE)
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
## [26] 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
## [51] 51 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 52
```

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

```
which.min(X)
```

```
## [1] 52
```

```
which.max(X)
```

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

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

```
length(unique(X))
```

```
## [1] 53
```

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

```
as.factor(X)
```

```
##  [1] Inf               4.19180654857877  4.1734172518943   4.15464876785729
##  [5] 4.13548512895119  4.11590933734319  4.09590327428938  4.07544759935851
##  [9] 4.05452163806914  4.03310325630434  4.01116871959141  3.98869253500376
## [13] 3.96564727304425  3.94200336638929  3.91772888178973  3.89278926071437
## [17] 3.86714702345081  3.84076143030548  3.81358809221559  3.78557852142874
## [21] 3.75667961082847  3.72683302786084  3.69597450568212  3.66403300987579
## [25] 3.63092975357146  3.59657702661571  3.56087679500731  3.52371901428583
## [29] 3.48497958377173  3.44451784578705  3.40217350273288  3.3577627814323
## [33] 3.31107361281783  3.26185950714291  3.20983167673402  3.15464876785729
## [37] 3.09590327428938  3.03310325630434  2.96564727304425  2.89278926071437
## [41] 2.8135880922156   2.72683302786084  2.63092975357146  2.52371901428583
## [45] 2.40217350273288  2.26185950714291  2.09590327428938  1.89278926071437
## [49] 1.63092975357146  1.26185950714291  0.630929753571457 -Inf
## [53] NaN               NaN               NaN               NaN
## [57] NaN               NaN               NaN               NaN
## [61] NaN               NaN
## 53 Levels: -Inf 0.630929753571457 1.26185950714291 ... NaN
```

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

```
as.integer(X)
```

```
## Warning: NAs introduced by coercion to integer range
```

```
##  [1] NA  4  4  4  4  4  4  4  4  4  4  3  3  3  3  3  3  3  3  3  3  3  3  3
## [26]  3  3  3  3  3  3  3  3  3  3  3  3  2  2  2  2  2  2  2  2  2  1  1  1
## [51]  0 NA NA NA NA NA NA NA NA NA NA NA
```

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

```
y = X[(!(is.infinite(X) & is.nan(X) & X < 0))]
y
```

```
##  [1]           Inf 4.19180654858 4.17341725189 4.15464876786 4.13548512895
##  [6] 4.11590933734 4.09590327429 4.07544759936 4.05452163807 4.03310325630
## [11] 4.01116871959 3.98869253500 3.96564727304 3.94200336639 3.91772888179
## [16] 3.89278926071 3.86714702345 3.84076143031 3.81358809222 3.78557852143
## [21] 3.75667961083 3.72683302786 3.69597450568 3.66403300988 3.63092975357
## [26] 3.59657702662 3.56087679501 3.52371901429 3.48497958377 3.44451784579
## [31] 3.40217350273 3.35776278143 3.31107361282 3.26185950714 3.20983167673
## [36] 3.15464876786 3.09590327429 3.03310325630 2.96564727304 2.89278926071
## [41] 2.81358809222 2.72683302786 2.63092975357 2.52371901429 2.40217350273
## [46] 2.26185950714 2.09590327429 1.89278926071 1.63092975357 1.26185950714
## [51] 0.63092975357          -Inf           NaN           NaN           NaN
## [56]           NaN           NaN           NaN           NaN           NaN
## [61]           NaN           NaN
```

- 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.

```
sample( c(0,1), size = 500, replace = TRUE)
```

```
##   [1] 1 0 1 0 1 1 0 0 1 0 1 1 1 0 1 1 1 0 0 0 1 0 1 0 0 1 0 1 1 0 0 0 1 1 0 0 0
##  [38] 1 0 1 1 0 0 0 1 1 0 0 1 1 0 1 1 0 0 0 0 0 1 0 0 1 0 0 1 1 1 0 0 1 0 0 1 0
##  [75] 1 1 0 0 1 1 1 0 0 1 0 1 1 1 0 1 0 1 0 1 0 0 0 0 1 0 0 0 1 1 0 1 1 0 1 0 1
## [112] 0 0 0 1 1 1 1 0 0 1 0 1 1 0 1 0 0 1 1 1 0 1 1 0 0 1 1 0 0 0 0 0 1 1 1 0 0
## [149] 1 1 0 0 1 1 1 0 0 1 1 0 1 0 0 0 1 1 0 1 1 1 1 1 0 0 0 0 0 1 0 0 0 1 0 1 1
## [186] 0 0 0 0 1 0 1 0 1 0 0 1 1 0 1 1 1 1 1 0 0 1 1 1 1 1 1 0 0 1 1 1 1 1 0 1 0
## [223] 0 1 0 1 1 1 1 1 0 0 1 1 1 0 0 0 1 0 1 0 1 0 1 0 0 1 0 0 1 0 1 1 1 1 1 0 0
## [260] 1 0 0 1 0 0 1 0 0 0 1 0 0 1 1 1 1 1 0 0 1 1 1 0 0 1 1 0 1 1 0 1 1 1 1 1 0
## [297] 0 0 0 1 0 1 1 1 0 1 0 1 1 1 1 0 1 0 0 1 1 0 0 1 1 0 0 0 1 0 1 1 0 0 0 0 0
## [334] 0 0 0 1 0 1 0 0 0 0 0 0 0 1 1 0 1 1 0 1 1 1 1 1 1 1 1 0 1 0 0 1 1 1 1 0 1
## [371] 1 1 0 0 0 1 1 0 0 0 1 1 1 0 1 0 1 0 1 0 1 1 0 1 0 0 1 0 1 1 1 0 1 1 1 0 1
## [408] 1 1 0 1 1 1 1 0 0 0 1 1 0 0 1 0 1 0 0 1 1 1 0 1 0 0 1 1 1 1 0 0 1 1 1 1 0
## [445] 1 0 1 1 1 0 0 1 0 1 0 0 1 0 0 1 0 1 1 0 0 1 1 1 0 0 1 0 1 1 1 0 1 0 0 1 0
## [482] 1 0 0 0 1 1 1 0 0 1 1 1 1 0 1 0 1 0 0 1
```

- 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 = 100, replace = TRUE, prob = c(0.9, 0.1)))
```

```
## [1] 0.11
```

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

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

```
## [1] 0.896
```

- 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", "misdimeanor" , "felony"), size = 100, replace = TRUE))
X_3
```

```
##   [1] misdimeanor misdimeanor misdimeanor felony      infraction  none
##   [7] felony      infraction  infraction  misdimeanor misdimeanor none
##  [13] none        none        misdimeanor infraction  misdimeanor felony
##  [19] felony      infraction  felony      felony      felony      infraction
##  [25] infraction  felony      felony      none        none        infraction
##  [31] none        none        infraction  misdimeanor misdimeanor felony
##  [37] none        felony      infraction  felony      none        misdimeanor
##  [43] none        misdimeanor none        infraction  felony      none
##  [49] misdimeanor felony      infraction  felony      infraction  infraction
##  [55] infraction  misdimeanor infraction  infraction  misdimeanor infraction
##  [61] felony      misdimeanor misdimeanor misdimeanor felony      misdimeanor
##  [67] misdimeanor infraction  felony      infraction  misdimeanor felony
##  [73] felony      misdimeanor none        felony      felony      none
##  [79] none        none        infraction  none        felony      infraction
##  [85] felony      none        none        felony      infraction  misdimeanor
##  [91] misdimeanor felony      infraction  misdimeanor misdimeanor infraction
##  [97] misdimeanor infraction  infraction  misdimeanor
## Levels: felony infraction misdimeanor none
```

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

```
X_3_bin = X_3 != "none"
```

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

```
factor(X_3, levels = c("none", "infraction", "misdimeanor" , "felony"), order = TRUE)
```

```
##   [1] misdimeanor misdimeanor misdimeanor felony      infraction  none
##   [7] felony      infraction  infraction  misdimeanor misdimeanor none
##  [13] none        none        misdimeanor infraction  misdimeanor felony
##  [19] felony      infraction  felony      felony      felony      infraction
##  [25] infraction  felony      felony      none        none        infraction
##  [31] none        none        infraction  misdimeanor misdimeanor felony
##  [37] none        felony      infraction  felony      none        misdimeanor
##  [43] none        misdimeanor none        infraction  felony      none
##  [49] misdimeanor felony      infraction  felony      infraction  infraction
##  [55] infraction  misdimeanor infraction  infraction  misdimeanor infraction
##  [61] felony      misdimeanor misdimeanor misdimeanor felony      misdimeanor
##  [67] misdimeanor infraction  felony      infraction  misdimeanor felony
##  [73] felony      misdimeanor none        felony      felony      none
##  [79] none        none        infraction  none        felony      infraction
##  [85] felony      none        none        felony      infraction  misdimeanor
##  [91] misdimeanor felony      infraction  misdimeanor misdimeanor infraction
##  [97] misdimeanor infraction  infraction  misdimeanor
## Levels: none < infraction < misdimeanor < felony
```

```
?factor
```

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

```
#TO-DO
```

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

#TO-DO

Verify that.

```
#TO-DO
```

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

#TO-DO

Verify that.

```
#TO-DO
```

- 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"
)
```

- 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.

```
#TO-DO
```

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

```
#TO-DO
```

Print out a summary of the whole dataframe.

```
#TO-DO
```

- 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.

```
#TO-DO
```

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

```
#TO-DO
```

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

```
#TO-DO
```

- 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`.

```
#TO-DO
```

- 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
```