Problem set 3

Your name here

Due 10/15/2021 at 5pm

NOTE1*: Start with the file* `ps3_2021.Rmd` *(available from the github repository at https://github.com/UChicago-pol-methods/IntroQSS-F21/tree/main/assignments). Modify that file to include your answers. Make sure you can "knit" the file (e.g. in RStudio by clicking on the* `Knit` *button). Submit both the Rmd file and the knitted PDF via Canvas*

NOTE2*: You will need to have a working LaTeX installation to compile your code.*

Question 1:

Consider the random process of flipping a fair coin three times.

(1a) Describe the sample space, Ω.

$$\Omega = \{TTT, TTH, THT, HTT, THH, HTH, HHT, HHH\}$$

(1b) The random variable X that we're interested is the number of heads that we get from our random process. Write out the probability mass function for this random variable. *Hint: the coin is fair, so each of the events in the sample space above occurs with equal probability. Note how many heads we get in each event. Then look at the proportion of times we get no heads, one head, etc. These proportions are equal to the probability. List the number of heads under the x column. List the associated probabilities under the $P(X = x)$ column.*

$$f(x) = \begin{cases} 1/8 & x = 0 \\ 3/8 & x = 1 \\ 3/8 & x = 2 \\ 1/8 & x = 3 \\ 0 & \text{otherwise} \end{cases}$$

OR

x	$P(X = x)$
0	1/8
1	3/8
2	3/8
3	1/8

(1c) Calculate the mean of this random variable. Please show your work.

$$
\begin{aligned}
\mathrm{E}[X] &= \sum_{x} x \times \mathrm{P}[X = x] \\
&= 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} \\
&= \frac{12}{8} \\
&= 1.5
\end{aligned}
$$

(1c) Write out code to simulate this random process, where the output is a single realization of the random variable (i.e., a number that represents the number of heads in your coin flips).

Note3*: I set a random seed here, so that every time you recompile your assignment, you'll get the same number. For analyses that involve sampling or random processes, it is really important to set a random seed so that you can get reproducible results. Feel free to change the seed number to anything you want. In general you only need to set your random seed ONCE per script.*

```r
set.seed(60637)

X <- c(0, 1, 2, 3)
probs <- c(1/8, 3/8, 3/8, 1/8)

sample(x = X,
       size = 1,
       prob = probs)
```

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

(1d) Now run your random process so you sample from it 10,000 times [PLEASE DON'T OUTPUT ALL 10,000 OBSERVATIONS IN YOUR HOMEWORK, just save it to an R object]. What is the average number of heads across these 10k observations? This is the sample mean for a given sample.

```r
X <- c(0, 1, 2, 3)
probs <- c(1/8, 3/8, 3/8, 1/8)

result_n <- sample(x = X,
                   size = 10000,
                   prob = probs,
                   replace = TRUE)

mean(result_n)
```

```
## [1] 1.5062
```

(1e): Write your own function called `mymean()` to calculate the sample mean from a vector. Apply your function to your size 10k sample that you saved in the last problem. (Don't use `mean()` inside your function.)

```r
mymean <- function(x){
  sum(x)/length(x)
}

mymean(result_n)
```

```
## [1] 1.5062
```

Question 2:

Using the same random process of flipping three fair coins, code the random variable Y as 1 if we get three heads, and 0 otherwise.

(2a) Write out the probability mass function for this random variable Y.

y	$P(Y=y)$
0	7/8
1	1/8

(2a) Write out the joint probability mass function for the joint distribution of X and Y.

$$f(x,y) = \begin{cases} 1/8 & x=0, y=0 \\ 3/8 & x=1, y=0 \\ 3/8 & x=2, y=0 \\ 1/8 & x=3, y=1 \\ 0 & \text{otherwise} \end{cases}$$

OR

x	y	$P(X=x, Y=y)$
0	0	1/8
1	0	3/8
2	0	3/8
3	1	1/8

(2b) Write out the probability mass function for this random variable X *conditional* on Y.

$$f(x|y) = \begin{cases} 1/7 & x=0|y=0 \\ 3/7 & x=1|y=0 \\ 3/7 & x=2|y=0 \\ 1 & x=3|y=1 \\ 0 & \text{otherwise} \end{cases}$$

OR

x	y	$P(X=x|Y=y)$
0	0	1/7
1	0	3/7
2	0	3/7
3	1	1

Question 3:

(3a) Load the data set that you selected for your independent project. If your data set is not already in tibble format, transform it into a tibble. Print the data set so that we can see the

top few observations and the column names and types.

```
mydf <- cars %>%
  as_tibble()
mydf
```

```
## # A tibble: 50 x 2
##    speed  dist
##    <dbl> <dbl>
##  1     4     2
##  2     4    10
##  3     7     4
##  4     7    22
##  5     8    16
##  6     9    10
##  7    10    18
##  8    10    26
##  9    10    34
## 10    11    17
## # ... with 40 more rows
```

(3b) What do you think is the appropriate unit of observation in your data? Is your data set already formatted so that each row describes a unique unit of observation? If not, what does each row describe?

[**Extra credit:** if your data set will need to be reshaped using `pivot_longer()` or `pivot_wider()`, try reshaping it now. If it doesn't need to be reshaped, you can try reshaping it anyhow. Give your reshaped columns informative names. Explain what the unit of observation is in your reshaped data set.]