Lab 5: Probability Foundations

Answer the following questions in a Quarto document (.qmd). Render the completed document as a PDF, and turn it into Gradescope.

Question 1

This lab will focus on questions involving randomness. Write a line of code that will ensure that you receive the same output each time you render your Quarto document.

Question 2

There are many times in statistics when computing true probabilities is hard, but simulating is easy (and gets you close enough). We're going to practice that here. Flip to the backside of *PS: Probability Foundations.* Instead of explicitly calculating the probabilities as you did in class, you're going to *simulate rolls* to approximate the true probabilities.

part a

Create a vector which contains the spots on the 7-sided die as mentioned on the second page of the problem set and save it into the object `seven_sided`.

part b

Create a vector which contains the results of 3,000 rolls of the `seven_sided` die, and save it into the object `q2_rolls`.

part c

Calculate the approximate probabilities of each of the following events *using `q2_rolls` and the appropriate comparison operators and/or logical operators.* **Hint**: see the *Conditioning* Notes from last week.

For each event, you must **show your code** to receive credit.

subpart I: $P(A)$

subpart II: $P(B)$

subpart III: $P(C)$

subpart IV: $P(B \cap C)$

subpart V: $P\left(A \cap B^C\right)$

part d

Insert a Table into your Quarto document using the Table button in your Editor. Recreate and fill out the table below.

- In the second column: write down the pen-and-paper answers you calculated on the Problem Set.
- In the third column: write the approximate probabilities of each event from **part c**. (You can round where necessary since we will be checking your answers via the code).

Subpart	P.S. Calculation (Theoretical)	Lab Calculation (Simulation)
I		
II		
III		
IV		
V		

part e

Describe the results you see in the table. Do the values in columns two and three match up exactly? Should we expect them to?

Question 3

Consider a random experiment in which we:

- roll 2 six-sided dice
- find the sum of the spots across both of the dice.

part a

Write R code to create a vector which contains 3,000 simulations of this experiment (so each element of the vector is a sum from two dice). Save the vector into `sums`. *Hint: can you simulate 3,000 rolls of the two individual dice first?*

part b

With your simulated rolls obtained in **part a**, approximate the probability of getting *a sum of 8*.

part c

With your simulated rolls obtained in **part a**, approximate the probability of getting *a sum of less than or equal to 5*.

part d

With your simulated rolls obtained in **part a)**, approximate the probability of getting a sum *that is less than 3 or greater than 7*.

part e

With your simulated rolls obtained in **part a)**, approximate the probability of getting a sum *that is an odd number*.

Question 4

part a

Write R code to create a vector containing the integers 1, 2, 3, ..., 3,000 and save it into the vector `simulation_number` .

part b

Write R code to create a vector containing 3,000 simulated rolls of a standard, *six-sided* die. Save the vector into an object called `q4_rolls`.

part c

Create a data frame called `q4` with two columns: `simulation_number` and `q4_rolls` .

part d

Add a new column called `less_than_3` to the `q` data frame that contains 3,000 elements that are either `TRUE` if a die roll from the `q4_rolls` is *less than 3* or `FALSE` if the roll is *3 or greater*. Save the new data frame back into the object `q4`.

part e

Explore the *cumulative sum function* `cumsum()` in R. This function works on vectors, and therefore, can be used when mutating columns in a data frame. Call `cumsum()` on these two vectors, and, based on the results, describe what the function is doing in general.

```
my_numbers <- seq(from = 1, by = 1, to = 10)
my_logicals <- c(FALSE,TRUE, TRUE, FALSE, TRUE)
```

part f

Add a new column to the `q4` data frame called `cumulative_proportion`. Here are what the first three entries in the column should be:

- *First* entry: after *one roll*, the proportion of die rolls that are less than 3
- *Second* entry: after *two rolls*, the proportion of die rolls that are less than 3
- *Third* entry: after *three rolls*, the proportion of die rolls that are less than 3

The rest of the entries should also follow the pattern. Once you're finished, save the new data frame back into the object `q4`.

part g

Create a line plot using `ggplot()` to visualize how the cumulative proportion changes as the simulation number increases. Label your axes.

part h

The `geom_hline()` layer allows you to add a horizontal line on top of your visualization. Tack it onto the code of your previous plot to add a horizontal line which sits at the theoretical probability of rolling a number less than 3.

part i

Describe the pattern in the line plot you see. How does it relate to the theoretical probability of rolling a number less than 3?