EXERCISE SET 1

1. Exercises

Exercise 1 (Working with probability distributions and modeling)**:** The first two problems are inspired by those in section 2 of [1]. You should look there for more practice.

(a) Suppose that

$$Y \sim \text{Bernoulli}(q)$$

and let $Z = 1/(1+Y) + Y$. What is the sample space of Z and what is the probability function of Z? You can express the probability distribution either as a piecewise function or by specifying each probability, i.e., $P(Z = z) = \cdots$.

(b) Suppose a coin is flipped. If the coin is heads, we write down 0. If the coin is tails, we roll a dice and write down the number. Let Y be the number we write down. What is the sample space and the probability distribution for Y?

(c) For the previous problem, conditioned on the dice rolling a 4, what is the probability we write down 0? Conditioned on the coin being tails, what is the probability the dice rolls a 3?

(d) Consider the geometric distribution discussed in lecture. Provide three real-world examples of variables where the geometric distribution is a good model. Explain your reasoning.

Exercise 2 (Working with nested for loops)**:** Consider the following code:

```
> for i in range(5):
>   for j in range(i+1):
>     print(i,end='')
>   print("")
```

prints out

```
> 0
> 11
> 222
> 3333
> 44444
```

Modify this code to print

```
> 0
> 01
> 012
> 0123
> 01234
> 012345
```

Exercise 3 (Working with more complex data – **ChatGPT**)**:** Using ChatGPT, write python code to plot a map of the world with Hanover indicated by a red star. Then examine this code and answer the following questions:

- How was the data loaded and what variable was it stored in? Where is the information about the geometric shape of each country stored? Can you print out some of this information? Is there other information that is not used in the plot?

- Where is the information about the location of Hanover, NH stored?

- **Without using ChatGPT**, plot another point at Salt Lake City, UT with a green dot? (you can look up the coordinates).

Exercise 4 (Washington post data)**:** Below I load some data on homocide victims in US from the washington post. Don't worry about how I process it, all you need to work with is the DataFrame "data" on the very last line.

```
> data = pd.read_csv("https://raw.githubusercontent.com/washingtonpost
> /data-homicides/master/homicide-data.csv",encoding = "ISO-8859-1")
> data["victim_age"] = pd.to_numeric(data["victim_age"],errors="coerce")
```

(a) For each age $a = 1, \cdots, 100$ determine the number of victims $n(a)$ with an age $< a$ and put these values in a list. You can ignore the effects of those entries with missing ages.

(b) Now think for a moment about what you expect a plot of $n(a) vs. a$ to look like, then make a plot of $n(a)$ vs. a. Does it look like as expected?

(c) Divide the data into groups of white and non-white victims and repeat part (a) for each group. Then, for each group, make the plot from part (a). Comment on what you find.

Exercise 5 (Getting a sequence of wins)**:** Let J denote a random variable representing the number of times a fair coin is flipped before two heads appear in a row. As we saw in class, the following code generates simulations of J:

```
> def flip_until_two():
>   num_heads = 0
>   total_flips = 0
>   while num_heads <2:
>     y = np.random.choice([0,1])
>     if y == 0:
>       num_heads = 0
>     else:
>       num_heads = num_heads + 1
>     total_flips = total_flips + 1
>   return total_flips
```

(a) By modifying the above code, write a function rolluntil(n) that rolls a dice until we get n ones in a row. You should change the variable names accordingly. We will call this random variable R_n.

(b) Make a DataFrame where each column represents a value of n from 1 to 6 and each row is a simulation from the model R_n. There should be 100 rows.

(c) Create a plot comparing the maximum and minimum values of R_n as a function of n. You might notice one of these increases much faster than the other – why?

Exercise 6 (Joint distribution)**:** Consider the probability model defined by

$$\mathbb{P}(Y_A, Y_B) = \begin{cases} 1/3 & \text{if } Y_A = 0 \text{ and } Y_B = 0 \\ 1/3 & \text{if } Y_A = 0 \text{ and } Y_B = 1 \\ 1/6 & \text{if } Y_A = 1 \text{ and } Y_B = 0 \\ 1/6 & \text{if } Y_A = 1 \text{ and } Y_B = 1 \end{cases}$$

(a) What are the marginal distributions of Y_A and Y_B?

(b) Are Y_A and Y_B independent?

(c) Confirm your answer with simulations (ChatGPT allowed).

Exercise 7 (Verifying variance formula for Bernoulli variable)**:** Verify the formula for the variance

$$\text{Var}(Y) = q(1-q)$$

Remember, you can use pointwise arithmetic on numpy arrays e.g.

```
> q_range*q_range
```

makes a list where every element is the corresponding element of qrange squared. You should experiment to ensure you are using enough samples.

Exercise 8 (Working with Washington Post Data)**:** This a continuation of Exercise 4 Consider the quantities

$$P(age < z)$$
$$P(age < z|\text{white})$$
$$P(age < z|\text{not white}).$$

(a) Explain who each of these are related to the plot you made in Exercise 4.

(b) Make plots of them and comment of the difference between the plot in Exercise 4. Do you think age and race are independent based on these plots.

(c) Using the data, approximate (for this dataset)

$$P(\text{white}|10 < age < 60)$$

Hint: One way to do this is to use Bayes' rule

Exercise 9 (Covid modeling)**:** Suppose we are interested in modeling how likely we are to contract SARS CoV-3, a new more dangerous version of Covid, after a night out. To do this, we make the following assumptions:

- You interact with at most exactly N people in sequence, meaning no repeated interactions with the same person.
- Each person either has covid or does not (we do not distinguish between their viral load or how long they have had the disease).
- 10% of people in the student population have Covid.
- Given that someone has the Covid AND we interact with them, there is a 50% chance you contract the virus.

(a) Fill in the question marks in the following function so that it simulates whether or not you got covid from the night out; that is, so it returns 1 if you got covid and 0 if you didn't.

```
> def sim_covid(N):
>   got_covid = 0
>   for k in range(N):
>       got_covid_interaction = ????
>       if got_covid_interaction ==1:
>         got_covid =1
>   return got_covid
```

(b) The probability of getting covid form the entire night has the form

$$P(\text{get covid}) = 1 - (1 - x)^N \tag{1}$$

where $x \in [0, 1]$. Provide some justification of this formula and determine the value of x.

(c) Confirm Equation 1 using Monte Carlo simulations simulations. You should make a plot of this probability vs. n, similar to what we did for the Bernoulli distribution in the class notebook.

References

[1] Michael J Evans and Jeffrey S Rosenthal. *Probability and statistics: The science of uncertainty*. Macmillan, 2004.