Problem Set #2

For this problem set, you will need to submit your responses in a form on Blackboard. In addition, there will be a blank at the end for you to fill in the name of students you collaborated with (up to TWO). You will also need to submit your code via a separate link. Note that answers that are not supported by code where necessary will not receive full credit.

Part 1: Slot Machine Simulation

In the file slot_machine.py, you have been given code for three functions. Use this code to complete the tasks and answer the questions below. You may also wish to use functions/objects we wrote in class or from itertools. (42 pts total)

a) The function slot_machine() is a simulation of a game of chance. The slot machine contains 4 wheels which spin independently (that is, the outcome of one wheel does not influence any other wheel). Each wheel has 30 symbols on it: 4 each of the symbols for lemon, grapes, sevens, oranges, gold bar, cherries, and cash, and 2 of the word WIN. When the slot machine is played, the four wheels spin and each wheel eventually comes to rest with a single symbol displayed to the player. A player wins or loses depending on the combination of symbols on the wheels (e.g. they might win a certain amount, like double their bet, if all of the symbols are the same).

   Based on the code given in the function slot_machine(), figure out the rules of the game - that is, what do you have to spin to win each prize value? The syntax 10X indicates that a player wins 10 times their original bet, the X indicates multiplication. Fill the description of what the wheels must show into the table provided in the Wheel Result column (the table is meant to help you organize your answers but is not submitted).

   Here are some examples (they do not correspond to this problem):

   Prize: Win 5X Wheel Result: Lemon, Lemon, Any symbol but lemon, Same as symbol3
   (e.g. Lemon Lemon Cherries Cherries)
   Prize: Win 15X Wheel Result: Any symbol, Any symbol except symbol1, WIN, Same as symbol1
   (e.g. Gold bar, Sevens, WIN, Gold bar)

b) Calculate the number of ways the symbols can be combined such that you get the wheel result for that prize, i.e., the number of outcomes that result in you winning that prize. It is OK if you do these calculations with a calculator or with other means than Python. Fill these into the table in the Count of possible outcomes column.

   Here is an example for a different problem of what your answers should look like:

   You are drawing marbles out of a bag that contains 10 blue marbles and 15 red marbles. After you draw each marble, you return it to the bag (that is, you are sampling with replacement). Suppose that you win a prize if you draw 4 blue marbles in a row. The number of outcomes where you draw 4 blue marbles in a row are:

   10*10*10*10 or 10**4 = 10000

   Suppose that you also win a prize if you first draw 2 red marbles and then draw 2 blue marbles. The number of outcomes where you first draw 2 red marbles and then draw 2 blue marbles are:

   15*15*10*10 or 15**2*10**2 = 22500

   You would fill 10,000 and 22,500 into the table, and the Blackboard form.

c) Calculate the total number of possible outcomes for spinning four independent wheels.

d) Using your results from parts b and c calculate the probability of the outcomes Win 50X, Win 20X, Win 10X, Win 5X, Win 2X, and Lose. Note that the only probability you are required to submit is that for Lose (give your answer to 4 decimal places), but you will need the others to answer e) and f), so you should fill them into the table.

e) Based on the probabilities you calculated, would you re-assign any of the win amounts? That is, have any of the prize amounts have been assigned unfairly or incorrectly? Explain why or why not in 2 sentences or less.

f) Use the function test_probs given in the slot_machine.py file, run simulations where the slot machine is played 10,000 times. Record in the table the number of times each outcome occurred in your simulation results, along with the number of times (rounded up) you would have expected it. You should notice that the observed values did not match the expected values exactly. Why did this occur? Explain in 2-3 sentences using statistical terminology.

   If you repeated the simulation with only 100 plays of the machine, would you expect (theoretically) the observed frequencies to be closer to, about the same distance from, or further from the expected frequencies? (You do not need to run the simulation.)

g) Re-run the simulation using the slot_machine2 function. Again, you should notice that the observed values do not match what you expected. This occurred for a different reason than in part f) above. What is going on? Why did this occur? Explain in 2-3 sentences what you observed and why it happened, the latter using statistical terminology.

Question 2: Crash probabilities

This question makes use of the data in the file Police_Crash_Reports_2016.csv. This is a subset of the data publicly available at: https://data.vbgov.com/Public-Safety/Police-Traffic-Crash-Reports/rx8z-sq53 which only includes information on crash reports for 2016. Use the data provided to calculate the following probabilities which are related to the day of the week a crash occurred and/or the Police precinct (district) in which it occurred. Include the code that shows how you read the data in and how you generated your answers. (12 pts total)

For all calculated probabilities, give your answer to 4 decimal places.

a) What is the probability that a randomly selected crash did not occur in the 3rd precinct?

b) What is the probability that a randomly selected crash occurred on the weekend (Saturday or Sunday) in the 2nd precinct?

c) You know that a crash occurred on a Tuesday. What is the probability it occurred in the 4th precinct?

d) Are the events: 'crash occurred in the 3rd precinct' and 'crash occurred on Friday' disjoint? Justify your answer in one sentence including the appropriate probabilities.

Part 3: Disease probabilities

According to the CDC, the prevalence of HIV in the US, that is the probability an individual will have HIV is approximately 0.0034. The prevalence of TB is 3 per 100,000 individuals. As of 2015, 6% of individuals known to have TB were co-infected with HIV. (If you just used a calculator, that is fine, you don't have to show code here.) (16 pts total)

a) Are having HIV and having TB independent events? Justify your answer in one sentence including the appropriate probabilities.

b) Out of 100,000 individuals in the US population, how many would you expect to have HIV or TB? Round your answer up to the nearest whole number.

c) What proportion of individuals known to have TB are HIV negative? Give your answer to 2 decimal places.

d) What proportion of individuals known to be HIV positive do not have TB? Give your answer to 2 decimal places.

e) Suppose that you look at diagnostic test results for individuals tested for HIV. You find out that of the individuals who truly had the disease, 60% tested negative. You conclude that (select any answers that apply):

1. The prevalence of HIV is much higher than the CDC claims.
2. The specificity of the test is 40%.
3. It is more rare for an individual with HIV to test negative than to have TB.
4. The positive predictive value is 60%.
5. Individuals in the US population who have HIV and test positive make up less than 1% of the population.
6. The test data is not normally distributed, which skews the results.
7. The negative predictive value is 40%.
8. The sensitivity is 60%.

Part 4: Distributions

a) A traffic camera monitors an intersection to identify individuals who drive through red lights. Each day on average, 5 cars drive through a red light. Let X = the number of cars who run the red light in a day. (5 pts)

What distribution best describes X?

Would it be unusual for more than 10 cars to run the red light? Support your answer by calculating and reporting a probability. Give your answer to 4 decimal places. Provide code that shows your calculation.

b) A survey of European mitochondrial DNA variation has found that the most common haplotype, known as "H", occurs in 40% of people. Suppose we randomly sample 400 individuals of European descent and determine whether they have haplotype "H" or not. Let X=the number of individuals with the H haplotype. (7 pts)

What distribution best describes X?

How many individuals in this sample would you expect to have the 'H' haplotype?

What is the probability that at least 150 of the 400 individuals will have haplotype "H"? Give your answer to 4 decimal places. Provide code to support your calculation.

c) The .csv file congress_terms.csv contains data derived from the data set freely available at https://github.com/fivethirtyeight/data/blob/master/congress-age/congress-terms.csv. Subset the data so that it only includes entries from congress 105. Provide code to support your answers below.(18 pts total)

Plot the distribution of the variable age in congress 105. We will model this data with a normal distribution.

What are the parameters (mu and sigma) of this distribution (the model)? Give your answers to two decimal places.

Using the distribution you described above:

-What is the probability that a randomly selected individual from congress 105 is older than 75? Give your answer to 4 decimal places.

-What proportion of individuals from congress 105 are between the ages of 40 and 50? Give your answer to 4 decimal places.

-If an individual in congress 105 is younger than 76% of their fellow congressmen, how old is this individual? Give your answer to 2 decimal places.

Using the actual data (not the model):

-What is the probability that a randomly selected individual from congress 105 is older than 75? Give your answer to 4 decimal places.

-What proportion of individuals from congress 105 are between the ages of 40 and 50 (consider this as inclusive)? Give your answer to 4 decimal places.

-If an individual in congress 105 is younger than 76% of their fellow congressmen, how old is this individual? Give your answer to 2 decimal places.

- How do these values compare to what you calculated above? Note any similarities or differences and explain why they may occur in 1-2 sentences.

What do your results (whether using the model or not) tell you about ages in congress 105? Explain in 1-2 sentences.

Prize	Wheel Result	Count of possible outcomes	Expected probability	Observed count 10000 (Expected)
Win 50X!				
Win 20X!				
Win 10X!				
Win 5X!				
Win 2X!				