Student Name: Raymond Shum
Due Date: 01-25-2022
CST 383 - Intro to Data Science
Dr. Glenn Bruns

Lab: Conditional probability

All coding parts of this assignment (starting on problem 6) are to be done in Python/NumPy. There are hints at the end of the lab, but use only after trying hard to solve without a hint.

1. The contingency table below is from a 1979 study of marijuana smoking among college students. The study looked at use of marijuana by students and whether the students' parents smoked marijuana.

	parents used	parents didn't use	total
student uses	125	94	219
student doesn't use	85	141	226
total	210	235	445

   Fill in the margins of the table by computing the totals.

2. Assuming the results of this study are valid for 2016 (I doubt they are), what is the probability a student uses, given the student's parents used? What is the probability that a student uses given the student's parents didn't use? What is the probability that the parents didn't use given the student does use?

   P("student uses" | "parents used") = 125/210 = .595

   P("student uses" | "parents didn't use") = 94 / 235 = .4

   P("parents didn't use" | "student uses") 94/219 = .43

3. Convert the table into a probability table by dividing all cells by the total number of samples in the study. Also update the values in the margins.

	parents used	parents didn't use	total
student uses	.28	.21	.49
student doesn't use	.19	.32	.51
total	.47	.53	1

4. What is the probability that a student uses and the student's parents used?

   P("student uses" ^ "parents used) = .28

5. What is the probability that a student uses? What is the probability that a parent used?

   P("student uses") = .49

   P("parents used") = .47

6. Let's look at the "elder girl" problem using a simulation. First, write code to compute an array of 10^4 samples of 1 (boy) or 2 (girl). Assign the array to variable 'child1'.

   child1 = np.random.randint(low=1,high=3,size=10**4)

7. Create another array, child2, in the same way.

   child2 = np.random.randint(low=1,high=3,size=10**4)

8. Think of child1[0] and child2[0] as the two children of a family, child1[1] and child2[1] as the two children of another family, etc. Write code to assign to variable 'one_girl' the number of families with at least one girl.

   one_girl = ((child1 == 2) | (child2 ==2)).sum()

9. Similarly to problem 8, write code to assign to variable 'both_girls' the number of families with two girls.

   two_girls = ((child1==2) & (child2==2)).sum()

10. Using one_girl and two_girls, estimate the conditional probability of a family having two girls if it has at least one girl.

    # one_girl and two_girls are both size == 10,000

    # P('two_girls | 'one_girl')

    two_girls / one_girl

11. Assign to 'elder_girl' the number of families in which the first child is a girl, then estimate the probability that the family has two girls given the family has an elder girl.

    elder_girl = (child1 == 2).sum()

    # P(two_girls | elder_girl)

    two_girls / elder_girl

12. If you have time, run your code a bunch of times to see how much your results vary, time by time.

13. If you still have time, download the free text Introductory Statistics with Randomization and Simulation, and find something that interests you in Chapter 1. Find a problem and see if you can solve it, with or without code.

Hints

1. -
2. For the first question, remember that we are going to think of the 'parents used' situation as the entire world. So for the first question you can focus on the column 'parents used'. In that column, look at the ratio of 'student uses' to the sum of the column. In other words, 125 / (125 + 85).
3. -
4. -
5. -
6. Use numpy.random.randint().
7. -
8. Remember that NumPy supports vectorized operations. No loop is needed.
9. -
10. Remember that the definition of P(A | B) = P(A & B) / P(B). In this case, we want P(two girls | at least one girl), which is equal to P(two girls & at least one girl) / P(at least one girl). (I wrote '&' to mean 'and'). Note that P(two girls & at least one girl) can be simplified.
11. -
12. You may find it helpful to to write a function prob_both_elder(n), which computes the probability that both are girls given the elder is a girl. The argument n indicates how many samples to take. (In other words, how many families are being simulated.)
13. -