CST 363 - Intro to Data Science

Dr. Glenn Bruns

**Lab: conditional probabilities - review**

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

|  |  |  |  |
| --- | --- | --- | --- |
|  | parent used | parent didn’t | total |
| student uses | 125 | 94 | 219 |
| student doesn’t | 85 | 141 | 226 |
| total | 210 | 235 |  |

Fill in the margins of the table by taking totals.

1. Assuming the results of this study are valid for 2016 (I doubt they are), what is the probability that 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?

To get the prob. that the student uses given the student’s parents used, look at the case where the student’s parents used. We are interested in the prob. that a student used for the values in this column, which is 125/(125 + 85) = 0.595.

The probability that a student uses given the parents didn’t use is 94/(94 + 141) = 0.4.

The probability that the parent’s didn’t use given the student does use is 94/(94 + 125) = 0.429.

1. 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.

|  |  |  |  |
| --- | --- | --- | --- |
|  | parent used | parent didn’t | total |
| student uses | 0.281 | 0.211 | 0.492 |
| student doesn’t | 0.191 | 0.317 | 0.508 |
| total | 0.472 | 0.528 |  |

Fill in the margins of the table by taking totals.

1. What is the probability that a student uses and the student’s parents used?

0.281

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

For the first, 0.492. For the second, 0.472.

1. 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'.

SEE BELOW FOR THE CODE.

1. Create another array, child2, in the same way.
2. 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.
3. Similarly to problem 8, write code to assign to variable ‘both\_girls’ the number of families with two girls.
4. Using one\_girl and two\_girls, estimate the conditional probability of a family having two girls if it has at least one girl.
5. 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.
6. If you have time, run your code a bunch of times to see how much your results vary, time by time.

import numpy as np

import seaborn as sns

# 6

child1 = np.random.randint(1, 2+1, size=10000)

# 7

child2 = np.random.randint(1, 2+1, size=10000)

# 8

one\_girl = np.sum(child1 + child2 >= 3)

# 9

both\_girls = np.sum(child1 + child2 == 4)

# 10

# compute P(both\_girls | one\_girl)

#

# P(both\_girls | one\_girl)

# = P(both\_girls, one\_girl) / P(one\_girl)

# = P(both\_girls) / P(one\_girl) # since both\_girls => one\_girl

# (we can use counts rather than probs because to estimate

# the probs we would divide counts by 10000 in both cases)

both\_girls / one\_girl

# 11

# P(both\_girls | elder\_girl)

# = P(both\_girls) / P(elder\_girl)

elder\_girl = np.sum(child1 == 2)

both\_girls / elder\_girl

print(both\_girls / elder\_girl)

# print(both\_girls / np.sum(child1 == 2))

# 12

def p\_both\_elder(n):

child1 = np.random.randint(1, 2+1, size=n)

child2 = np.random.randint(1, 2+1, size=n)

elder\_girl = np.sum(child2 == 2)

both\_girls = np.sum(child1 + child2 == 4)

return both\_girls / elder\_girl

# estimate the conditional probability repeatedly

sns.distplot(np.array([ p\_both\_elder(100) for \_ in range(1000) ]))

plt.show()