ConditionalProbabilityExercise

December 20, 2016

1 Conditional Probability Activity & Exercise

Below is some code to create some fake data on how much stuff people purchase given their age range.

It generates 100,000 random "people" and randomly assigns them as being in their 20's, 30's, 40's, 50's, 60's, or 70's.

It then assigns a lower probability for young people to buy stuff.

In the end, we have two Python dictionaries:

"totals" contains the total number of people in each age group. "purchases" contains the total number of things purchased by people in each age group. The grand total of purchases is in totalPurchases, and we know the total number of people is 100,000.

Let's run it and have a look:

```
In [16]: from numpy import random
         random.seed(0)

         totals = {20:0, 30:0, 40:0, 50:0, 60:0, 70:0}
         purchases = {20:0, 30:0, 40:0, 50:0, 60:0, 70:0}
         totalPurchases = 0
         for _ in range(100000):
             ageDecade = random.choice([20, 30, 40, 50, 60, 70])
             purchaseProbability = float(ageDecade) / 100.0
             totals[ageDecade] += 1
             if (random.random() < purchaseProbability):
                 totalPurchases += 1
                 purchases[ageDecade] += 1
         #https://www.tutorialspoint.com/python/python_dictionary.htm

In [17]: totals

Out[17]: {20: 16576, 30: 16619, 40: 16632, 50: 16805, 60: 16664, 70: 16704}

In [18]: purchases

Out[18]: {20: 3392, 30: 4974, 40: 6670, 50: 8319, 60: 9944, 70: 11713}

In [4]: totalPurchases
```

```
Out[4]: 45012
```

Let's play with conditional probability.

First let's compute P(E|F), where E is "purchase" and F is "you're in your 30's". The probability of someone in their 30's buying something is just the percentage of how many 30-year-olds bought something:

```
In [5]: PEF = float(purchases[30]) / float(totals[30])
        print('P(purchase | 30s): ' + str(PEF))
```

```
P(purchase | 30s): 0.29929598652145134
```

P(F) is just the probability of being 30 in this data set:

```
In [6]: PF = float(totals[30]) / 100000.0
        print("P(30's): " +  str(PF))
```

```
P(30's): 0.16619
```

And P(E) is the overall probability of buying something, regardless of your age:

```
In [7]: PE = float(totalPurchases) / 100000.0
        print("P(Purchase):" + str(PE))
```

```
P(Purchase):0.45012
```

If E and F were independent, then we would expect P(E | F) to be about the same as P(E). But they're not; PE is 0.45, and P(E|F) is 0.3. So, that tells us that E and F are dependent (which we know they are in this example.)

What is P(E)P(F)?

```
In [8]: print("P(30's)P(Purchase)" + str(PE * PF))
```

```
P(30's)P(Purchase)0.07480544280000001
```

P(E,F) is different from P(E|F). P(E,F) would be the probability of both being in your 30's and buying something, out of the total population - not just the population of people in their 30's:

```
In [11]: print("P(30's, Purchase)" + str(float(purchases[30]) / 100000.0))
```

```
P(30's, Purchase)0.04974
```

P(E,F) = P(E)P(F), and they are pretty close in this example. But because E and F are actually dependent on each other, and the randomness of the data we're working with, it's not quite the same.

We can also check that P(E|F) = P(E,F)/P(F) and sure enough, it is:

```
In [12]: print((purchases[30] / 100000.0) / PF)
```

```
0.29929598652145134
```

1.1 Your Assignment

Modify the code above such that the purchase probability does NOT vary with age, making E and F actually independent.

Then, confirm that P(E | F) is about the same as P(E), showing that the conditional probability of purchase for a given age is not any different than the a-priori probability of purchase regardless of age.

In [ ]: