Diagnostic Problem Set

Katja Della Libera

September 2020

1 Discrete Probabilities[1]

Consider the following joint distribution $p(x, y)$, over two discrete random variables X and Y.

y_1	0.01	0.02	0.03	0.1	0.1
y_2	0.05	0.1	0.05	0.07	0.2
y_3	0.1	0.05	0.03	0.05	0.04
	x_1	x_2	x_3	x_4	x_5

1. Compute the marginal distributions $p(x)$ and $p(y)$

   We add the columns in the table to get $p(x)$...

x	x_1	x_2	x_3	x_4	x_5
$p(x)$	0.26	0.17	0.11	0.22	0.34

   ... and the rows to get $p(y)$

y	y_1	y_2	y_3
$p(y)$	0.26	0.47	0.27

2. Compute the conditional distribution $p(x|Y = y_1)$

   To get the conditional probability given $Y = y_1$, we only consider the first row in the original table where this condition is met. Then we look at the $p(y)$ marginal distribution and find the probability that y_1 occurs (0.26). The original row now needs to be divided by this number to get the conditional distribution $p(x|Y = y_1)$

x	x_1	x_2	x_3	x_4	x_5
$p(x\|Y = y_1)$	0.038	0.077	0.115	0.385	0.385

3. Compute the conditional distribution $p(y|X = x_3)$

   We repeat the process in 3. and find the probability that $X = x_3$ to be 0.11. We then divide every item in column 3 of the original table by 0.11 and get the following table.

y	y_1	y_2	y_3
$p(y\|X = x_3)$	0.273	0.455	0.273

   Notice the probabilities don't add up to exactly 1 because of rounding errors.

2 Calculating Probabilities

1. How many students are in your CS146 class, including yourself? (Take a guess if you're not entirely sure.)

   17 at the time I wrote this.

2. Assuming you know nothing about your classmates –

   a. What is the probability that you were born after all of them

      If I know nothing at all, I can assume all of us to have equal probability of being the oldest student, which would mean 1/17 chance for me to be born after all of them.

[1] **#professionalism**: I have been meaning to learn LaTex and took this problem set as the opportunity to try it out. I hope you enjoy my very first LaTex document and think it looks professional.

b. What is the probability that you were born before all of them?

   This is identical to the previous question if I make no assumptions and also equal to 1/17.

c. What is the probability that you were born after at least half of the other students?

   In this case, I have to be older than at least 8 of them. If I think of the student sorted by age as a sequence whose ordering doesn't matter, there are 9 spots in this sequence that would fulfill this requirement and all 9 of them are equally likely at 1/17. So the total probability is 9/17.

3 Normal distribution

If x is distributed according to the normal distribution with mean μ and standard deviation σ, and if $f(x) = x^3 + 2x + 1$.

1. Calculate the expected value of $f(x)$.

   With x distributed normally around μ, the expected value is

   $$\int f(x)PDF(x)dx$$

   where PDF is the probability density function of the normal distribution with mean μ and standard deviation σ and the integral is taken over the domain of this pdf (in this case $(-\infty, \infty)$. This works out to

   $$\mu^3 + 3\mu\sigma^2 + 2\mu + 1$$

2. Calculate the probability $P(f(x) > 1)$.

   The probability again depends on σ and μ and is equal to

   $$\int boolean(f(x) > 1) * PDF(x)dx$$

   integrated over the domain of the pdf of our normal distribution. This works out to be

   $$1 - \frac{1}{2}erfc(\frac{\mu}{\sqrt{2}\sigma})$$

   where erfc(x) is the compliment of the error function.

   Evaluating this function for a given distribution, we can get a numerical value. For example $\mu = 1$ and $\sigma = 1$ works out to 0.841345 or $\mu = 0$ and $\sigma = 2$ works out to 0.5.

3. Write a python script to confirm your answer to question 2. Generate a lot of random numbers from a normal distribution with a particular mean and standard deviation. Calculate $f(x)$ for each of these random numbers. How many of them are greater than 1? Does that match the probability calculated in question 2?

```
from numpy import random

def f(x):
    return x**3+2*x+1
```

```
def prob(mu,sigma):
    samples  = random.normal(loc=mu,  scale=sigma,
    size=10000)
    dist = [f(x) for x in samples]
    prob = sum([1 for f in dist if f>1])/len(dist)
    return prob
```

The result of running *prob* will vary slightly due to the random component. With my sample size I got 0.8406 for $\mu = 1$ and $\sigma = 1$ and 0.5027 for $\mu = 0$ and $\sigma = 2$.[2]

4 Double-headed coin

A bag contains 1000 coins, where 999 are normal(they land on heads 50% of the time) and 1 coin is double-headed (lands on heads 100% of the time). A coin is chosen uniformly at random and flipped 10 times. It lands on heads every time-so, 10 heads in a row.

Question 1 Given this information, what is the probability that the chosen coin is the double-headed coin?

Without considering the result of the flips, the coin has a $\frac{1}{1000}$ chance of being all heads in which case its chance of coming up heads 10 times is 1. The chance of it coming up heads 10 times if it is normal is $(\frac{1}{2})^10 = \frac{1}{1024} = 0.0009765625$ and the chance of being normal from the random draw is $\frac{999}{1000}$. $\frac{1}{2}^10 * \frac{999}{1000} = 0.00097558593$ and $\frac{1}{1000} = 0.001$ which gives them a ratio of approximately 1:1 or a likelihood of $\frac{1}{1}$ that the coin is double-headed.

Question 2 You should find that the answer to Question 1 is approximately 1/2. I find this surprising. Explain as straightforwardly as you can (without using complicated math) why we should, in fact, expect the result to be approximately 1/2

The surprise may come from the fact that given a fair coin, it is incredibly unlikely to get 10 heads (less than $\frac{1}{1000}$ as we calculated above). However, this neglects to consider that getting the all-heads coin is almost as unlikely at exactly $\frac{1}{1000}$. One of these unlikely events is happening from our observation and to find their relative likelihood we just need to look at the ratio between the two which is 1:1.[3]

5 Smokers

According to the CDC (Centers for Disease Control and Prevention) in the USA, men who smoke are 23 times more likely to develop lung cancer than men who don't smoke. Also according to the CDC, 21.6% of men in the USA smoke. What is the probability that a man in the USA is a smoker, given that he develops lung cancer?

This is similar to the case of our coin in that we're finding the ratio between two probabilities, the one of being a smoking male and having lung cancer and the one of being a non-smoking man

[2] #distributions: Throughout question 3, I show my understanding of how to make calculations around probability and expected values for a given distribution and even use general formulas. I explain variations from random samples and how to interpret them.

[3] #audience: Throughout the problem set I use a mix of mathematical terms and simple and intuitive language to explain my thinking, adjusting to the task at hand in specific questions such as this one that is one of the least mathematical ones.

and having lung cancer. Since we don't know the actual probability of developing lung cancer in either or the total population, we will work with a variable x indicating the probability of a non-smoker of developing lung cancer.

The probability of being a smoker and having lung cancer is

$$p(s_1 \cap l_1) = 0.216 * 23 * x = 4.968 * x$$

where S describes the smoking status with s_1 being smokers and s_2 being non-smokers, and L is the lung cancer status with l_1 being lung cancer patients and l_2 being healthy individuals.

The probability of being a non-smoker and having lung-cancer is

$$p(s_2 \cap l_1) = 0.784 * x$$

The probability of a lung cancer patient to be a smoker is therefore.

$$\frac{4.968x}{4.968x + 0.784x} \approx 0.864$$

or around 86%.