Exercise 10

Ngày 30 tháng 1 năm 2021

1. Consider the following bivariate distribution $p(x, y)$ of two discrete random variables X and Y

Y \ X	x_1	x_2	x_3	x_4	x_5
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

Compute

(a) The marginal distributions $p(x)$ and $p(y)$

(b) The conditional distributions $p(x|Y = y_1)$ and $p(y|X = x_3)$

Solution:

(a) The marginal distributions are obtained by summing the probabilies over all the values of the variable being marginalized. Thus, to obtain $p(x)$ we sum over columns (i.e., over the values corresponding to different y):

$$p(x_1) = P(X = x_1) = P(X = x_1, Y = y_1) + P(X = x_1, Y = y_2) + P(X = x_1, Y = y_3) = \tag{1}$$

$$p(x_2) = P(X = x_2) = P(X = x_2, Y = y_1) + P(X = x_2, Y = y_2) + P(X = x_2, Y = y_3) = \tag{2}$$

$$p(x_3) = P(X = x_3) = P(X = x_3, Y = y_1) + P(X = x_3, Y = y_2) + P(X = x_3, Y = y_3) = \tag{3}$$

$$p(x_4) = P(X = x_4) = P(X = x_4, Y = y_1) + P(X = x_4, Y = y_2) + P(X = x_4, Y = y_3) = \tag{4}$$

$$p(x_5) = P(X = x_5) = P(X = x_5, Y = y_1) + P(X = x_5, Y = y_2) + P(X = x_5, Y = y_3) = \tag{5}$$

As a correctness check, note that this distribution satisfies the normalization condition, i.e. that sum of the probabilities is 1:

$$\sum_{i=1}^{5} p(x_i) = 1 \tag{6}$$

The marginal distribution $p(y)$ can be obtained in a similar way, by summing the matrix rows:

$$p(y_1) = P(Y = y_1) = \sum_{i=1}^{5} P(X = x_i, Y = y_1) = 0.01 + 0.02 + 0.03 + 0.1 + 0.1 = 0.26 \tag{7}$$

$$p(y_2) = P(Y = y_2) = \sum_{i=1}^{5} P(X = x_i, Y = y_2) = 0.05 + 0.1 + 0.05 + 0.07 + 0.2 = 0.47 \tag{8}$$

$$p(y_3) = P(Y = y_3) = \sum_{i=1}^{5} P(X = x_i, Y = y_3) = 0.1 + 0.05 + 0.03 + 0.05 + 0.04 = 0.27 \tag{9}$$

We can again check that the normalization condition is satisfied:

$$\sum_{i=1}^{3} p(y_i) = 1 \tag{10}$$

To determine conditional distributions we use the definition of the conditional probability:
$P(X = x, Y = y_1) = P(X = x|Y = y_1)P(Y = y_1) = p(x|Y = y_1)p(y_1)$. Thus,

$$\begin{aligned}
p(x_1|Y = y_1) &= \frac{P(X=x_1,Y=y_1)}{p(y_1)} = \frac{0.01}{0.26} \approx 0.038 \\
p(x_2|Y = y_1) &= \frac{P(X=x_2,Y=y_1)}{p(y_1)} = \frac{0.02}{0.26} \approx 0.077 \\
p(x_3|Y = y_1) &= \frac{P(X=x_3,Y=y_1)}{p(y_1)} = \frac{0.03}{0.26} \approx 0.115 \\
p(x_4|Y = y_1) &= \frac{P(X=x_4,Y=y_1)}{p(y_1)} = \frac{0.1}{0.26} \approx 0.385 \\
p(x_5|Y = y_1) &= \frac{P(X=x_5,Y=y_1)}{p(y_1)} = \frac{0.1}{0.26} \approx 0.385
\end{aligned}$$

Likewise the conditional distribution $p(y|X = x_3)$ is given by

$$\begin{aligned}
p(y_1|X = y_3) &= \frac{P(X=x_3,Y=y_1)}{p(x_3)} = \frac{0.03}{0.11} \approx 0.273 \\
p(y_2|X = y_3) &= \frac{P(X=x_3,Y=y_2)}{p(x_3)} = \frac{0.05}{0.11} \approx 0.454 \\
p(y_3|X = y_3) &= \frac{P(X=x_3,Y=y_3)}{p(x_3)} = \frac{0.03}{0.11} \approx 0.273
\end{aligned}$$

2. Consider two random variables x, y with joint distribution $p(x, y)$. Show that:

   $E_X[X] = E_Y[E_X[x|y]]$

   Here, $E_X[x|y]$ denotes the expected value of x under the conditional distribution $p(x, y)$

   Solution:

   The expectation value and the conditional expectation value are given by

   $E_X[x] = \int xp(x)dx,$

$E_Y[f(y)] = \int f(y)p(y)dy$,
$E_X[x|y] = \int xp(x|y)dx$

We then have

$E_Y\left[E_X[x|y]\right] = \int E_X[x|y]p(y)dy = \int \left[\int xp(x|y)dx\right] p(y)dy = \int\int xp(x|y)p(y)dxdy =$
$\int\int xp(x,y)dxdy = \int x\left[\int p(x,y)dy\right] dx = \int xp(x)dx = E_X[x]$,

where we used the definition fo the conditional probability density

$p(x|y)p(y) = p(x,y)$

3. Một cuộc điều tra cho thấy, ở 1 thành phố 20.7% dân số dùng sản phẩm X, 50% dùng loại sản phẩm Y và trong những người dùng Y thì 36.5% dùng X. Phỏng vấn ngẫu nhiên một người dân trong Thành phố đó, tính xác xuất đề người ấy:

   (a) Dùng cả X và Y.

   (b) Dùng Y, và biết rằng người đó không dùng X.

4. Ở một vụ án, ảnh sát xác định được mẫu DNA dài 100 base (ATCG...) của 1 tội phạm ở hiện trường. Một người bình thường có xác suất 10^-5 trùng đoạn 100 base như vậy. Cảnh sát khoanh vùng được tội phạm là dân cư trong thành phố có một triệu dân, trong đó có 10000 dân đã được thả khỏi tù trong vòng 10 năm trở lại đây và DNA họ vẫn được giữ lại trong tù. Xác suất của 1 người đã từng phạm tội gây án lần này là α, còn xác suất những người còn lại gây án là β. Khi phân tích DNA của các tù nhân, Jones là người duy nhất có DNA match với DNA còn lại ở hiện trường, xác suất Jones là thủ phạm là bao nhiêu?

   Solution:

   Gọi A là sự kiện J phạm tội $P(A) = \alpha(1)$
   An là sự kiện J không phạm tội $P(An) = 1 - P(A)(2)$
   B là sự kiện chỉ J có DNA match trong database

   $P(A|B) = \frac{P(B|A)*P(A)}{/P(B)} = \frac{P(B|A)*P(A)}{P(B|A)*P(A)+P(B|An)*P(An)}(3)$
   Cần tính $P(B|A)$ và $P(B|An)$

Dễ thấy, để chỉ J có DNA trong database match và điều điện J phạm tội thì tất cả những DNA còn lại trong database ko đc match vs DNA ở hiện trường:
$=> P(B|A) = (1 - 10^{-5})^{9999}(4)$
Lại có, nếu J vô tội và đảm bảo chỉ J có DNA match thì cần 3 điều kiện: DNA của J phải match, tất cả các người còn lại trong database cx phải vô tội và không ai trừ J trong database có DNA match.
$=> P(B|An) = 10^{-5} * \frac{1-10000\alpha}{1-\alpha}(1 - 10^{-5})^{9999}(5)$
Trong đó, $(1 - 10000 * \alpha)/(1 - \alpha)$ chính là xác suất điều kiện tất cả mọi người trong database vô tội nếu J vô tội. Thay (1), (2), (4) và (5) vào (3), ta được: $P(A|B) = 1/(0.9 + 10^{-5}\alpha)$

5. Prove the relationship: $V_X = E_X[x^2] - (E_X[x])^2$, which relates the standard definition of the variance to the raw-score expression for the variance

   Solution:

   The standard definition of variance is

   $V_X[x] = E_X[(x - \mu)^2]$,

   Where $\mu = E_X[x]$.

   Using the properties of average we can write:

   $V_X[x] = E_X[(x - \mu)^2] = E_X[x^2 - 2x\mu + \mu^2] = E_X[x^2] - E_X[2x\mu] + E_X[\mu^2] =$
   $E_X[x^2] - 2\mu E_X[x] + \mu^2 = E_X[x^2] - 2\mu^2 + \mu^2 = E_X[x^2] - \mu^2$

   By substituting to this equation the definition of μ, we obtain the desired equation

   $V_X[x] = E_X[(x - \mu)^2] = E_X[x^2] - (E_X[x])^2$

6. Cho

$$F_X(x) = \begin{cases} 0 & (x < 0) \quad (11a) \\ x/2 & (0 \le x \le 1) \quad (11b) \\ x/6 + 1/3 & (1 < x < 4) \quad (11c) \\ 1 & (x \ge 4) \quad (11d) \end{cases}$$

là hàm phân bố xác suất của biến ngẫu nhiên liên tục X

(a) Tính hàm mật độ của X

(b) Tìm phân vị mức 75% của X (tức là tìm $x_{0.75}$ sao cho $P(X < x_{0.75}) < 0.75$)

(c) Tính kỳ vọng của X

(d) Tính $E(1/X)$

(e) Ta định nghĩa

$$Y = \begin{cases} -1 & (X \le 1) \quad (12a) \\ 1 & (X > 1) \quad (12b) \end{cases}$$

i. Tìm $F_Y(0)$

ii. Tìm phương sai của Y