Probability and statistics

Lecture notes from March 13.

Definition 1. *Let (X,Y) be a two-dimensional random variable with the density function $f(x,y)$ and $f_1(x), f_2(y)$ be its marginal densities. Then the conditional probability density function is defined as:*

$$f_{X|Y}(x) = \frac{f(x,y)}{f_2(y)} \quad \text{and} \quad f_{Y|X}(y) = \frac{f(x,y)}{f_1(x)}. \tag{1}$$

In case of discrete distribution:

$$p_{i\leftarrow j} \equiv p_{x_i|y_j} = \frac{p_{ij}}{p_{\bullet j}} \quad \text{oraz} \quad p_{i\mapsto j} \equiv p_{y_j|x_i} = \frac{p_{ij}}{p_{i\bullet}}. \tag{2}$$

Example:
We return to the (slightly changed) example from the previous note.
(i) Let the two-dimensional density and the marginal densities be as follows.

$$(X,Y) = \begin{array}{r|ccc|c} X/Y & 2 & 3 & 5 & p_{1\bullet} \\ \hline -2 & 0.10 & 0.05 & 0.07 & 0.22 \\ 0 & 0.05 & 0.03 & 0.08 & 0.16 \\ 1 & 0.01 & 0.07 & 0.15 & 0.23 \\ 3 & 0.38 & 0.00 & 0.01 & 0.39 \\ \hline p_{\bullet j} & 0.54 & 0.15 & 0.31 & 1.00 \end{array}$$

Therefore, the variables X, Y have the marginal distributions:

$$X = \begin{array}{c|cccc} x_i & -2 & 0 & 1 & 3 \\ \hline p_{i\bullet} & 0.22 & 0.16 & 0.23 & 0.39 \end{array},$$
$$Y = \begin{array}{c|ccc} y_j & 2 & 3 & 5 \\ \hline p_{\bullet j} & 0.54 & 0.15 & 0.31 \end{array}$$

(ii) The conditional densities $p_{X|Y}$ are as follows. For each of the columns of the table above, we calculate $p_{x_i|y_j} = \dfrac{p_{ij}}{p_{\bullet j}}$ so that the values in the columns add up to 1. We say that the j-th column contains the probability values of x_i provided $Y = y_j$.

$$(X|Y = y_j) = \begin{array}{c|ccc|} X/Y & 2 & 3 & 5 \\ \hline -2 & 10/54 & 5/15 & 7/31 \\ 0 & 5/54 & 3/15 & 8/31 \\ 1 & 1/54 & 7/15 & 15/31 \\ 3 & 38/54 & 0 & 1/31 \\ \hline & 1 & 1 & 1 \end{array}$$

The same is true for the conditional densities $p_{Y|X}$. For each row of the table above, we calculate $p_{x_i|y_j} = \dfrac{p_{ij}}{p_{\ i\bullet}}$ so that the values in the rows add up to 1. We say that the i-th row contains the probability values of y_j provided $X = x_i$.

$$
(Y|X=x_i) = \begin{array}{c|ccc|c}
X/Y & 2 & 3 & 5 & \\
\hline
-2 & 10/22 & 5/22 & 7/22 & 1 \\
0 & 5/16 & 3/16 & 8/16 & 1 \\
1 & 1/23 & 7/23 & 15/23 & 1 \\
3 & 38/39 & 0 & 1/39 & 1 \\
\hline
\end{array}
$$

Because the columns (*respectively* rows) of the above tables describe random variables, it makes sense to use the phrase "conditional expected value", for example

$$
\mathrm{E}(X|Y=2) = -2\cdot 10/54 + 0\cdot 5/54 + 1\cdot 1/54 + 3\cdot 38/54 = \frac{95}{54}.
$$

(iii) Let's now determine the distribution of the random variable $Z = X+Y$. In the upper left corner of each element of the table, there is the value of the variable Z, and below the probability of such value of Z.

$$
Z = X+Y = \begin{array}{c|ccc|}
X/Y & 2 & 3 & 5 \\
\hline
-2 & 0/0.10 & 1/0.05 & 3/0.07 \\
0 & 2/0.05 & 3/0.03 & 5/0.08 \\
1 & 3/0.01 & 4/0.07 & 6/0.15 \\
3 & 5/0.38 & 6/0.00 & 8/0.01 \\
\hline
\end{array}
$$

After the ordering, we get the following distribution of the variable Z:

$$
Z = \begin{array}{c|cccccccc}
z_i & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 8 \\
\hline
p_i & 0.10 & 0.05 & 0.05 & 0.11 & 0.07 & 0.46 & 0.15 & 0.01
\end{array}.
$$

At the end of the example, we note that $4.04 = \mathrm{E}(X+Y) = 0.96 + 3.08 = \mathrm{E}(X) + \mathrm{E}(Y)$.

Theorem 1. *Let (X,Y) be a 2-dimensional random variable. Then the expected value of the sum of the random variables X,Y is equal to the sum of the expected values of these random variables: $E(X+Y) = E(X) + E(Y)$.*

Proof. Let (X,Y) be a discrete random variable. Then:

$$
\begin{aligned}
\mathrm{E}(X+Y) &= \sum_i \sum_j (x_i + y_j)\cdot p_{ij} = \sum_i \sum_j x_i p_{ij} + \sum_j \sum_i y_j p_{ij} = \\
&= \sum_i \left(x_i \sum_j p_{ij} \right) + \sum_j \left(y_j \sum_i p_{ij} \right) = \sum_i x_i \cdot p_{i\bullet} + \sum_j y_j \cdot p_{\bullet j} = \mathrm{E}(X) + \mathrm{E}(Y)
\end{aligned}
\tag{3}
$$

In the case of a continuous random variable:

$$
\begin{aligned}
&\mathrm{E}(X+Y) = \int_{\mathbb{R}} \int_{\mathbb{R}} (x+y) f(x,y)\, dy\, dx = \int_{\mathbb{R}} \int_{\mathbb{R}} x f(x,y)\, dy\, dx + \int_{\mathbb{R}} \int_{\mathbb{R}} y f(x,y)\, dx\, dy = \\
&= \int_{\mathbb{R}} \left(x \int_{\mathbb{R}} f(x,y)\, dy \right) dx + \int_{\mathbb{R}} \left(y \int_{\mathbb{R}} f(x,y)\, dx \right) dy = \int_{\mathbb{R}} x f_1(x)\, dx + \int_{\mathbb{R}} y f_2(y)\, dy = \mathrm{E}(X) + \mathrm{E}(Y).
\end{aligned}
\tag{4}
$$

□

Recall that the covariance of the variables X,Y is the value of the expression $\mu_{11} \equiv \mathrm{Cov}(X,Y) = \mathrm{E}\left[(X - \mathrm{E}X)\cdot(Y - \mathrm{E}Y)\right]$. For discrete variables $\mathrm{Cov}(X,Y) = \sum_i \sum_j (x_i - \mathrm{E}X)\cdot(y_j - \mathrm{E}Y)\, p_{ij}$, for continuous variables $\int_{\mathbb{R}} \int_{\mathbb{R}} (x - \mathrm{E}X)(y - \mathrm{E}Y)\, dy\, dx$. The following theorem gives the relationship between the independence of the random variables and their covariance.

Theorem 2. *Let (X,Y) be a two-dimensional random variable, whose marginal variables X, Y are independent. Then $Cov(X,Y) = 0$.*

Proof. (For the discrete type random variables).

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= \sum_i \sum_j (x_i - \mathrm{E}X)\cdot(y_j - \mathrm{E}Y)\, p_{ij} = \\
&= \sum_i \sum_j x_i y_j p_{ij} - \mathrm{E}Y \sum_i \sum_j x_i p_{ij} - \mathrm{E}X \sum_i \sum_j y_j p_{ij} + \mathrm{E}X \cdot \mathrm{E}Y \sum_i \sum_j p_{ij} = \\
&= \sum_i x_i \left(\sum_j y_j p_{i\bullet} p_{\bullet j} \right) - \mathrm{E}Y \sum_i \left(x_i \sum_j p_{ij} \right) - \mathrm{E}X \left(\sum_j y_j \sum_i p_{ij} \right) + \mathrm{E}X \cdot \mathrm{E}Y = \\
&= \left(\sum_i x_i p_{i\bullet} \right) \left(\sum_j y_j p_{\bullet j} \right) - \mathrm{E}Y \sum_i x_i p_{i\bullet} - \mathrm{E}X \sum_j y_j p_{\bullet j} + \mathrm{E}X \cdot \mathrm{E}Y = 0.
\end{aligned}$$

□

COMMENTS:

1. If the variables are independent, $\mathrm{E}(X \cdot Y) = \mathrm{E}X \cdot \mathrm{E}Y$ (see task 1.7b).
2. The converse of Theorem (2) is not true.
3. Task 1.7a is a special case of the theorem $\mathrm{V}(X) = \mathrm{E}(X^2) - (\mathrm{E}X)^2$

Example:

(Continuation of the example on page 2 of note 3.)

We consider the density function $f(x,y) = \dfrac{3xy}{16}$ defined in the area bounded by straight lines $y = 0$, $x = 2$ and the curve $y = x^2$.

The 2-dimensional cumulative distribution function is $F(s,t) \equiv F_{XY}(x,y) = \int_{-\infty}^{s} \int_{-\infty}^{t} f(x,y)\, dy\, dx$.

Unfortunately, when calculating the cumulative distribution function, we should precisely define the integration intervals.

(i) Let A denotes (in geometric terminology equivalent to high school) II, III and IV "quadrant" of the plane. If $(s,t) \in A$ then $F(s,t) \equiv F_{XY}(x,y) = \int_{-\infty}^{s} \int_{-\infty}^{t} f(x,y)\, dy\, dx = 0$.

(ii) Let B be the area bounded by the lines $y = 0$, $x = 2$ and the curve $y = x^2$. Then:

$$\begin{aligned}
F(s,t) &= \int_{-\infty}^{s} \int_{-\infty}^{t} f(x,y)\, dy\, dx = \\
&= \int_{0}^{\sqrt{t}} \int_{0}^{x^2} f(x,y)\, dy\, dx + \int_{\sqrt{t}}^{s} \int_{0}^{t} f(x,y)\, dy\, dx\,.
\end{aligned}$$

Intuition:

first we compute the area (integral, ppb) under the curve $y = x^2$, for $x \in (0, \sqrt{t})$ (area S_1) and next we add the area (integral, ppb) under the straight line $y = t$ for $x \in (\sqrt{t}, s)$ (area S_2). In the first integral, y changes (for the set x) from 0 to x^2, and in the second integral (also for the set x) from 0 to t.

The area under B will also be useful in (iii) and (iv).

(iii) Area $C = [0,2] \times [x^2, \infty)$. Then $F(s,t) \equiv F_{XY}(x,y) = \int_{-\infty}^{s} \int_{-\infty}^{t} f(x,y)\, dy\, dx = F(s, s^2)$.

Intuition: The area to the left and below the point $c_1 = (s,t)$ for the density function $f(x,y)$ is the same as the area left and below the point $c_2 = (s, s^2)$. It is therefore possible to refer to the formula in point (ii).

(iv) Area $D = [2, \infty) \times [0, 4]$. Here $F(s, t) = F(2, t)$. Please compare the intersection of set $(-\infty, s] \times (-\infty, t]$ with the area where $f(x, y)$ is not equal to 0. Graphically: instead of the point $d_1 = (s, t)$, the point $d_2 = (2, t)$ must be taken for the calculations. The formula from point (ii) applies here as well.

(v) Area $E = [2, \infty) \times [4, \infty)$. This, along with the area A , is the simplest case. Here $F(s, t) = 1$, because we integrate the density over the entire "non-zero" area. Ultimately, the formula for the cumulative distribution function is:

$$F(s,t) = \begin{cases} 0, & \text{for } (s,t) \in A, \\ \dfrac{3s^2t^2}{64} - \dfrac{t^3}{96}, & \text{for } (s,t) \in B, \\ \dfrac{s^6}{64}, & \text{for } (s,t) \in C, \\ \dfrac{3t^2}{16} - \dfrac{t^3}{96}, & \text{for } (s,t) \in D, \\ 1, & \text{for } (s,t) \in E. \end{cases}$$

Another example illustrating the computation of the sum of random variables.

Example:

The random variable (X, Y) has a distribution with the density $f(x, y) = 3x\sqrt{y}$ inside the area $[0, 1] \times [0, 1]$. Determine the distribution of the variable $Z = X + Y$.

We start with the transformation $(X, Y) \mapsto (Z, T)$. Let $Z = X + Y$, $T = Y$ (the formula for T may be different). First, we reverse the transformation and get $X = Z - T$, $Y = T$. We compute the Jacobian:

$$J = \begin{vmatrix} \dfrac{\partial x}{\partial z} & \dfrac{\partial x}{\partial t} \\ \dfrac{\partial y}{\partial z} & \dfrac{\partial y}{\partial t} \end{vmatrix} = \begin{vmatrix} 1 & -1 \\ 0 & 1 \end{vmatrix} = 1.$$

The next step is to substitute the variables in the density function $f(x, y)$ and multiply it by the modulus of the Jacobian: $g(z, t) = f(x(z, t), y(z, t)) \cdot |J| = 3(zt)\sqrt{t}$.

Important: the distribution of the variable Z is one of the marginal distributions of the 2-dimensional (Z, T) variable. Therefore, integrate the function $g(z, t)$ by the variable t. To find the value of $f_1(z)$ $(z \in [0, 2])$, the range of the variable t should be determined.

$$\begin{cases} 0 < x < 1 \\ 0 < y < 1 \end{cases} \quad \begin{cases} 0 < z - t < 1 \\ 0 < t < 1 \end{cases} \quad \begin{cases} z - 1 < t < z \\ 0 < t < 1 \end{cases}.$$

Integration interval for the variable t is $[\max\{0, z - 1\}, \min\{1, z\}]$. For $z \in [1, 2]$ we have $t \in [0, z]$; for $z \in [1, 2]$ we have $t \in [z - 1, 1]$.

We calculate the indefinite integral

$$\int g(z,t)\,dt = \int 3\,(z\sqrt{t} - t\sqrt{t})\,dt = t\sqrt{t}\,(2z - {}^6\!/_5\,t + C).$$

Finally

$$g_1(z) = \begin{cases} t\sqrt{t}\,\left(2z - {}^{6}\!/\!{}_{5}\, t\right)\Big|_{t=0}^{z}, & z \in [0,1], \\[2ex] t\sqrt{t}\,\left(2z - {}^{6}\!/\!{}_{5}\, t\right)\Big|_{t=z-1}^{1}, & z \in [1,2]. \end{cases}$$

↤

Z poważaniem,
Witold Karczewski