Multivariate Analytics

BITS Pilani
Pilani Campus

Akanksha Bharadwaj
Asst. professor, CS/IS Department

innovate achieve lead

BITS Pilani
Pilani Campus

SS ZG536, ADV STAT TECHNIQUES FOR ANALYTICS
Contact Session 12

Introduction

- In science and in real life, we are often interested in two (or more) random variables at the same time.

- For example, we might measure the IQ and birthweight of children, or the level of air pollution and rate of respiratory illness in cities.

innovate achieve lead

BITS Pilani

Pilani Campus

Joint Distribution

Joint PMF

- Suppose X and Y are two discrete random variables and that X takes values $\{x_1, x_2, . . . , x_n\}$ and Y takes values $\{y_1, y_2, . . . , y_m\}$.

- The ordered pair (X, Y) take values in the product $\{(x_1, y_1),(x_1, y_2), . . .(x_n, y_m)\}$.

- If X and Y are discrete, this distribution can be described with a **joint probability mass function.**

- The joint probability mass function (joint pmf) of X and Y is the function $p(x_i , y_j)$ giving the probability of the joint outcome $X = x_i , Y = y_j$.

Joint probability table

$X \backslash Y$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_m
x_1	$p(x_1, y_1)$	$p(x_1, y_2)$	$\cdots$	$p(x_1, y_j)$	$\cdots$	$p(x_1, y_m)$
x_2	$p(x_2, y_1)$	$p(x_2, y_2)$	$\cdots$	$p(x_2, y_j)$	$\cdots$	$p(x_2, y_m)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_i	$p(x_i, y_1)$	$p(x_i, y_2)$	$\cdots$	$p(x_i, y_j)$	$\cdots$	$p(x_i, y_m)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	
x_n	$p(x_n, y_1)$	$p(x_n, y_2)$	$\cdots$	$p(x_n, y_j)$	$\cdots$	$p(x_n, y_m)$

Properties of joint probability mass function

A joint probability mass function must satisfy two properties:
1. $0 \leq p(x_i, y_j) \leq 1$
2. The total probability is 1. We can express this as

$$\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) = 1$$

Example

- Two balls are selected at random from a bag containing three green, two blue and four red balls.

green blue

- If X and Y are respectively the numbers of green and blue balls included among the two balls drawn from the bag, find the probabilities associated with all possible pairs of value of X and Y.

Solution

(X, Y) green Blue

- Here the possible pairs are (0, 0), (0, 1), (1, 0), (1, 1), (0, 2), (2, 0).
- To obtain the probability associated with (1, 0), we see that we are dealing with the event of getting one of the three green balls, no blue ball and hence, one of the red ball is the number of ways in which we get this event

$$^{3}C_{1} \times {}^{2}C_{0} \times {}^{4}C_{1} = 12$$

- total number of ways in which two ball are drawn out of nine

$$^{9}C_{2} = 36$$

- probability of the event associated with (1, 0) is $12/36 = 1/3$

Similarly,

y	X		
	0	1	2
0	1/6	1/3	1/12
1	2/9	1/6	0
2	1/36	0	0

Marginal Probability?

$P(X=0) = \frac{1}{6} + \frac{2}{9} + \frac{1}{36}$ $\quad P(X=1) = 1/3 + 1/6$ $\quad P(X=2) = 1/12$

$P(Y=0) = \frac{1}{6} + \frac{1}{3} + 1/2$ $\quad P(Y=1) = 2/9 + 1/6$ $\quad P(Y=2) = 1/36$

Joint PDF

- If X takes values in [a, b] and Y takes values in [c, d] then the pair (X, Y) takes values in the product [a, b] × [c, d].

- If X and Y are continuous, this distribution can be described with a **joint probability density function**

- The joint probability density function (joint pdf) of X and Y is a function f(x, y) giving the probability density at (x, y).

- That is, the probability that (X, Y) is in a small rectangle of width dx and height dy around (x, y) is f(x, y) dx dy.

Properties of joint probability distribution function

A joint probability density function must satisfy two properties:

1. $0 \leq f(x, y)$
2. The total probability is 1. We now express this as

$$\int_c^d \int_a^b f(x, y)\, dx\, dy = 1$$

Exercise

- A bank operates both a drive-up facility and a walk-up window. On a randomly selected day, let X = the proportion of time that the drive-up facility is in use (at least one customer is being served or waiting to be served) and Y = the proportion of time that the walk-up window is in use. Then the set of possible values for (X, Y) is the rectangle D {(x, y): 0 <=x<=1, 0<=y<=1}. Suppose the joint pdf of (X, Y) is given by

$$f(x, y) = \begin{cases} \frac{6}{5}(x + y^2) & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

1. Verify that this is a legitimate pdf
2. The probability that neither facility is busy more than one-quarter of the time

Solution

(1) $$\int_0^1\int_0^1 \frac{6}{5}(x+y^2)\,dx\,dy = \int_0^1\int_0^1 \frac{6}{5}x\,dx\,dy + \int_0^1\int_0^1 \frac{6}{5}y^2\,dx\,dy$$

$$= \int_0^1 \frac{6}{5}x\,dx\,[y]_0^1 + \int_0^1 \frac{6}{5}y^2\,dy\,[x]_0^1 = \int_0^1 \frac{6}{5}x\,dx + \int_0^1 \frac{6}{5}y^2\,dy$$

$$= \frac{6}{5}\left[\frac{x^2}{2}\right]_0^1 + \frac{6}{5}\left[\frac{y^3}{3}\right]_0^1 = \frac{6}{10} + \frac{6}{15} = 1$$

$\therefore$ It is legitimate.

(2) $P(0 \leq X \leq 1/4,\ 0 \leq Y \leq 1/4)$

$$= \frac{6}{5}\int_0^{1/4}\int_0^{1/4} x\,dx\,dy + \frac{6}{5}\int_0^{1/4}\int_0^{1/4} y^2\,dx\,dy = \frac{6}{5}\int_0^{1/4}\left[\frac{x^2}{2}\right]_0^{1/4} dy + \frac{6}{5}\int_0^{1/4}\left[\frac{y^3}{3}\right]_0^{1/4} dx$$

$$= \frac{6}{5}\times\frac{1}{32}\int_0^{1/4} dy + \frac{\cancel{6}^{\,2}}{5}\times\frac{1}{\underset{32}{\cancel{64\times 3}}}\int_0^{1/4} dx = \frac{6}{5}\times\frac{1}{32}\times\frac{1}{4} + \frac{\cancel{6}^{\,1}}{5}\times\frac{1}{32}\times\frac{1}{4}$$

$$= \frac{7}{640}$$

Exercise

Suppose the random variables X and Y have the joint density function defined by

$$f(x,y) = \begin{cases} c(2x+y) & 2 < x < 6, \quad 0 < y < 5 \\ 0 & \text{otherwise} \end{cases}.$$

Find value of c.

Solution

$$1 = \int_2^6 \int_0^5 c(2x+y)\, dy\, dx = \int_2^6 c\left[2xy + \frac{y^2}{2}\right]_0^5 dx$$

$$1 = \int_2^6 c\left(10x + \frac{25}{2}\right)dx = c\left(10 \times \frac{x^2}{2} + \frac{25}{2}x\right)_2^6$$

$$1 = c\left[\frac{10 \times \overset{3}{\cancel{6}} \times 6}{\cancel{2}} + \frac{25}{\cancel{2}} \times \overset{3}{\cancel{6}} - \frac{10 \times 2 \times \cancel{2}}{\cancel{2}} - \frac{25 \times \cancel{2}}{\cancel{2}}\right]$$

$$1 = c\,[180 + 75 - 20 - 25]$$

$$\therefore c = 1/210$$