Introduction to Data Science With Probability and Statistics

Lecture 18 & 19: Joint Distributions & Covariance

CSCI 3022 - Summer 2020
Sourav Chakraborty
Dept. of Computer Science
University of Colorado Boulder

Joint Distributions of Discrete Random Variables

The joint distribution of two discrete random variables X and Y, defined on the *same* sample space Ω, is given by prescribing the probabilities of all possible values of the pair *(X, Y)*.

DEFINITION. The *joint probability mass function* p of two discrete random variables X and Y is the function $p : \mathbb{R}^2 \to [0, 1]$, defined by

$$p(a, b) = \mathrm{P}(X = a, Y = b) \quad \text{for } -\infty < a, b < \infty.$$

Joint Distributions of Discrete Random Variables

List the elements of the event $\{S = 7, M = 4\}$ and compute its probability, where S and M being the sum and the maximum of two independent throws of a die.

Joint Distributions of Discrete Random Variables

List the elements of the event $\{S = 7, M = 4\}$ and compute its probability, where S and M being the sum and the maximum of two independent throws of a die.

$(3,4)$ and $(4,3)$

$$P(S=7, M=4) = \frac{2}{36}$$

Joint Distributions of Discrete Random Variables

Joint probability mass function $p(a, b) = P(S = a, M = b)$

	b					
a	1	2	3	4	5	6
2	1/36	0	0	0	0	0
3	0	2/36	0	0	0	0
4	0	1/36	2/36	0	0	0
5	0	0	2/36	2/36	0	0
6	0	0	1/36	2/36	2/36	0
7	0	0	0	2/36	2/36	2/36
8	0	0	0	1/36	2/36	2/36
9	0	0	0	0	2/36	2/36
10	0	0	0	0	1/36	2/36
11	0	0	0	0	0	2/36
12	0	0	0	0	0	1/36

what about $P(S=a)$?

$$P(S=5) = P(S=5, M=1)$$
$$+ P(S=5, M=2)$$
$$+ P(S=5, M=3)$$
$$+ \ldots$$
$$P(S=5, M=6)$$

Joint Distributions of Discrete Random Variables

Joint distribution and marginal distributions of *S* and *M*

a	b: 1	2	3	4	5	6	$p_S(a)$
2	1/36	0	0	0	0	0	1/36
3	0	2/36	0	0	0	0	2/36
4	0	1/36	2/36	0	0	0	3/36
5	0	0	2/36	2/36	0	0	4/36
6	0	0	1/36	2/36	2/36	0	5/36
7	0	0	0	2/36	2/36	2/36	6/36
8	0	0	0	1/36	2/36	2/36	5/36
9	0	0	0	0	2/36	2/36	4/36
10	0	0	0	0	1/36	2/36	3/36
11	0	0	0	0	0	2/36	2/36
12	0	0	0	0	0	1/36	1/36
$p_M(b)$	1/36	3/36	5/36	7/36	9/36	11/36	1

- The marginal distribution of a subset of a collection of random variables is the probability distribution of the variables contained in the subset.
- Summing over the columns of the table yields the marginal distribution of *M*

Joint Distributions of Continuous Random Variables

- We know that the probability that a single continuous random variable X lies in an interval $[a, b]$, is equal to the area under the probability density function f of X over the interval.

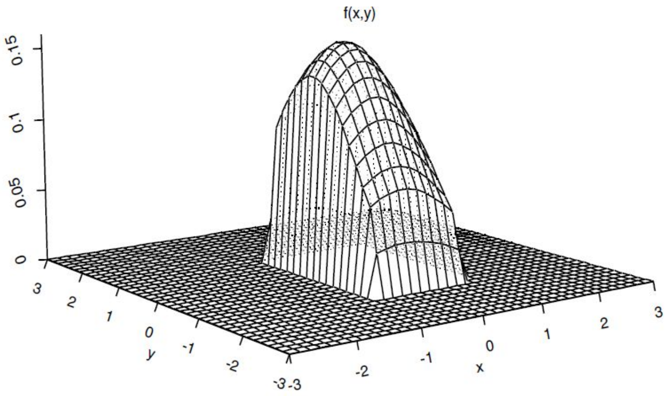
- For the joint distribution of continuous random variables X and Y, the probability that the pair (X, Y) falls in the rectangle [a1, b1]×[a2, b2] is equal to the volume under the joint probability density function $f(x, y)$ of (X, Y) over the rectangle.

Joint Distributions of Continuous Random Variables

DEFINITION. Random variables X and Y have a *joint continuous distribution* if for some function $f : \mathbb{R}^2 \to \mathbb{R}$ and for all numbers a_1, a_2 and b_1, b_2 with $a_1 \le b_1$ and $a_2 \le b_2$,

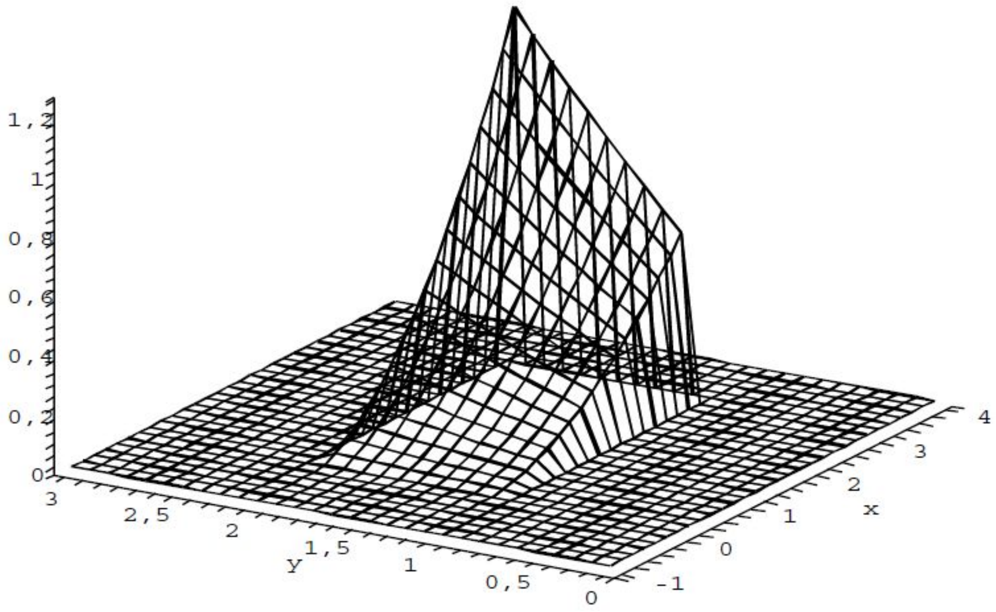
$$P(a_1 \le X \le b_1, a_2 \le Y \le b_2) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x, y)\, dx\, dy.$$

The function f has to satisfy $f(x, y) \ge 0$ for all x and y, and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$. We call f the *joint probability density function* of X and Y.



Bivariate Normal Probability Density Function

$$f(x,y) = \frac{2}{75}(2x^2y + xy^2) \quad \text{for } 0 \le x \le 3 \text{ and } 1 \le y \le 2, \text{ and } f(x,y) = 0 \text{ otherwise;}$$



Bivariate Normal Probability Density Function

$$P\left(1 \leq X \leq 2, \frac{4}{3} \leq Y \leq \frac{5}{3}\right) = ?$$

Bivariate Normal Probability Density Function

$$P\left(1 \leq X \leq 2, \frac{4}{3} \leq Y \leq \frac{5}{3}\right) = ? \quad \int_{1}^{2}\int_{4/3}^{5/3} f(x,y)\,dx\,dy$$

$$= \frac{2}{75}\int_{1}^{2}\left(\int_{4/3}^{5/3}(2x^2y + xy^2)\,dy\right)dx$$

$$= \frac{2}{75}\int_{1}^{2}\left(x^2 + \frac{61}{81}x\right)dx = \frac{187}{2025}$$

Bivariate Normal Probability Distribution Function

$$F(a,b) = P(x \leq a, y \leq b)$$

$$= \int_{-\infty}^{a} \int_{-\infty}^{b} f(x,y)\,dy\,dx$$

$$= \frac{2}{75} \int_{0}^{a} \left(\int_{1}^{b} (2x^2y + xy^2)\,dy \right) dx$$

$$= \frac{1}{225} \left(2a^3b^2 - 2a^3 + a^2b^3 - a^2 \right)$$

Joint Distributions of Continuous Random Variables

FROM JOINT TO MARGINAL PROBABILITY DENSITY FUNCTION. Let f be the joint probability density function of random variables X and Y. Then the *marginal* probability densities of X and Y can be found as follows:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x.$$

Joint Distributions of Continuous Random Variables

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$$

$$= \frac{2}{75}\int_{1}^{2}(2x^2y + xy^2)\,dy$$

$$= \frac{2}{225}(9x^2 + 7x)$$

Independent random variables

DEFINITION. The random variables X and Y, with joint distribution function F, are *independent* if

$$\mathrm{P}(X \le a, Y \le b) = \mathrm{P}(X \le a)\,\mathrm{P}(Y \le b)\,,$$

that is,

$$F(a,b) = F_X(a)F_Y(b) \tag{9.4}$$

for all possible values a and b. Random variables that are not independent are called *dependent*.

Independent random variables

1). $P(X \in A, Y \in B) = P(X \in A)\, P(Y \in B)$

True for both discreet and continuous.

2). $f(x,y) = f_X(x) \cdot f_Y(y)$; $f(\cdot)$: prob. density functions.

And the general version, with more than 2 R.Vs, also holds.

$$f(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) \cdots \cdot f_{X_n}(x_n)$$

Propagation of independence

PROPAGATION OF INDEPENDENCE. Let $X_1, X_2, \ldots, X_n$ be independent random variables. For each i, let $h_i : \mathbb{R} \to \mathbb{R}$ be a function and define the random variable

$$Y_i = h_i(X_i).$$

Then $Y_1, Y_2, \ldots, Y_n$ are also independent.

Propagation of independence

PROPAGATION OF INDEPENDENCE. Let $X_1, X_2, \ldots, X_n$ be independent random variables. For each i, let $h_i : \mathbb{R} \to \mathbb{R}$ be a function and define the random variable

$$Y_i = h_i(X_i).$$

Then $Y_1, Y_2, \ldots, Y_n$ are also independent.

For instance; if X, Y, Z are independent,

then $X^2, 2Y, 4Z+3$ are also independent.

Expectation and joint distributions

Example:- China vases of various shapes are produced in the Delftware factories in the old city of Delft. One particular simple cylindrical model has height *H* and radius *R* centimeters.Due to certain circumstances (manual error), *H* and *R* are not constants but are random variables.

The volume of a vase is equal to the random variable $V = \pi H R^2$.

$$\mathrm{E}[V] = \int_{-\infty}^{\infty} v f_V(v)\,\mathrm{d}v,\ f_V \text{ denotes the probability density of } V$$

$$\mathrm{E}[V] = \mathrm{E}\left[\pi H R^2\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \pi h r^2 f(h,r)\,\mathrm{d}h\,\mathrm{d}r$$

Expectation and joint distributions

$E[V] = ?$

Suppose, $H = U(25, 35)$ & $R = U(7.5, 12.5)$

Expectation and joint distributions

$$E[V] = \int_{25}^{35} \int_{7.5}^{12.5} \pi h r^2 \cdot \frac{1}{10} \cdot \frac{1}{5} \, dh \, dr$$

$$= \frac{\pi}{50} \int_{25}^{35} h \, dh \int_{7.5}^{12.5} r^2 \, dr$$

$$= 9621.127 \ cm^2$$

Expectation and joint distributions

TWO-DIMENSIONAL CHANGE-OF-VARIABLE FORMULA. Let X and Y be random variables, and let $g : \mathbb{R}^2 \to \mathbb{R}$ be a function.
If X and Y are *discrete* random variables with values $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$, respectively, then

$$\mathrm{E}[g(X,Y)] = \sum_i \sum_j g(a_i, b_j)\mathrm{P}(X = a_i, Y = b_j).$$

If X and Y are *continuous* random variables with joint probability density function f, then

$$\mathrm{E}[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

Expectation and joint distributions

Example: Take *g(x, y) = xy* for discrete random variables *X* and *Y* with the joint probability distribution given in table below. The expectation of XY is computed as follows: E[XY] = ?

b	a = 0	1	2
0	0	1/4	0
1	1/4	0	1/4
2	0	1/4	0

$$E[xy] = \sum_x \sum_y (xy) \cdot P(x = x, y = y)$$

$$= (0 \cdot 0) \cdot 0 + (1 \cdot 0) 1/4 + (2 \cdot 0) \cdot 0$$

$$+ (0 \cdot 1) 1/4 + (1 \cdot 1) \cdot 0 + (2 \cdot 1) \cdot 1/4$$

$$+ (0 \cdot 2) \cdot 0 + (1 \cdot 2) \cdot \frac{1}{4} + (2 \cdot 2) \cdot 0$$

$$= 1.$$

Covariance

We saw E[X + Y] = E[X] + E[Y]

Does such a simple relation also hold for the variance of the sum Var(X + Y) or for expectation of the product E[XY]?

Let X and Y be two random variables with joint probability density as

$$f(x, y) = \frac{2}{75}(2x^2y + xy^2) \quad \text{for } 0 \leq x \leq 3 \text{ and } 1 \leq y \leq 2.$$

Var(X + Y) = $939/2000$ $\neq$

Var(X) +Var(Y) = $\frac{4747}{10000}$

Covariance

$$\mathrm{Var}(x+y) = E\left[(x+y - E[x+y])^2\right]$$

$$= E\left[\left((x-E[x]) + (y-E[y])\right)^2\right]$$

$$= \overbrace{E\left[(x-E[x])^2\right]}^{\mathrm{Var}(x)} + \overbrace{E\left[(y-E[y])^2\right]}^{\mathrm{Var}(y)}$$

$$+ 2E\left[(x-E[x])(y-E[y])\right]$$

$$\mathrm{Var}(x+y) = \mathrm{Var}(x) + \mathrm{Var}(y) + 2\underbrace{E\left[(x-E[x])(y-E[y])\right]}_{\text{Covariance}}$$

Covariance

DEFINITION. Let X and Y be two random variables. The *covariance* between X and Y is defined by

$$\mathrm{Cov}(X,Y) = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])] .$$

- If the covariance of X and Y is positive, we say that X and Y are positively correlated.
- If the covariance of X and Y is negative, we say that X and Y are negatively correlated.
- If the covariance of X and Y is zero, we say that X and Y are uncorrelated.

Covariance

AN ALTERNATIVE EXPRESSION FOR THE COVARIANCE. Let X and Y be two random variables, then

$$\mathrm{Cov}(X,Y) = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]\,.$$

Independent versus uncorrelated

Let *X* and *Y* be two independent random variables.

$$\begin{aligned}
\mathrm{E}[XY] &= \sum_i \sum_j a_i b_j \mathrm{P}(X = a_i, Y = b_j) \\
&= \sum_i \sum_j a_i b_j \mathrm{P}(X = a_i)\,\mathrm{P}(Y = b_j) \\
&= \left(\sum_i a_i \mathrm{P}(X = a_i) \right) \left(\sum_j b_j \mathrm{P}(Y = b_j) \right) \\
&= \mathrm{E}[X]\,\mathrm{E}[Y]\,.
\end{aligned}$$

Independent versus uncorrelated: If two random variables X and Y are independent, then X and Y are uncorrelated.

Covariance

Consider the random variables X and Y with the joint distribution given in below. Check that X and Y are dependent, but that also E[XY] = E[X]E[Y].

$$P(X=0, Y=0) \neq P(X=0)\, P(Y=0)$$

$$\quad 0 \qquad\qquad\qquad \tfrac{1}{4} \qquad \tfrac{1}{4}$$

	a			
b	0	1	2	$P(Y=b)$
0	0	1/4	0	1/4
1	1/4	0	1/4	1/2
2	0	1/4	0	1/4
$P(X=a)$	1/4	1/2	1/4	

$$E[XY] = 1 = E[X]\, E[Y]$$

$$0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$$

Covariance

VARIANCE OF THE SUM. Let X and Y be two random variables. Then always

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y).$$

If X and Y are *uncorrelated*,

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

We always have that $E[X + Y] = E[X]+E[Y]$, whereas $\mathrm{Var}(X + Y) = \mathrm{Var}(X)+\mathrm{Var}(Y)$ only holds for uncorrelated random variables (and hence for independent random variables!)

Covariance Change of Units

COVARIANCE UNDER CHANGE OF UNITS. Let X and Y be two random variables. Then

$$\mathrm{Cov}(rX + s, tY + u) = rt\,\mathrm{Cov}(X, Y)$$

for all numbers r, s, t, and u.

For instance;

X, Y be temperatures in F

$rX+s$, $tY+u$ are in °C.

Covariance Change of Units

For X and Y having joint probability distribution function as below:-

$$f(x,y) = \frac{2}{75}(2x^2y + xy^2) \quad \text{for } 0 \le x \le 3 \text{ and } 1 \le y \le 2, \text{ and } f(x,y) = 0 \text{ otherwise;}$$

Find Cov(−2X + 7, 5Y − 3)

Covariance Change of Units

For X and Y having joint probability distribution function as below:-

$$f(x, y) = \frac{2}{75}(2x^2y + xy^2) \quad \text{for } 0 \leq x \leq 3 \text{ and } 1 \leq y \leq 2, \text{ and } f(x, y) = 0 \text{ otherwise;}$$

The correlation coefficient

A standardized version of the covariance called the correlation coefficient of *X* and *Y*

DEFINITION. Let X and Y be two random variables. The *correlation coefficient* $\rho(X,Y)$ is defined to be 0 if $\mathrm{Var}(X)=0$ or $\mathrm{Var}(Y)=0$, and otherwise

$$\rho(X,Y)=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}.$$

Important Property:

The correlation coefficient

For X and Y in the previous problem, find $\rho(-2X + 7, 5Y - 3)$

The correlation coefficient

For X and Y in the previous problem, find $\rho(-2X + 7, 5Y - 3)$