Advanced Statistics

Dr. Syed Faisal Bukhari

Associate Professor

Department of Data Science

Faculty of Computing and Information Technology

University of the Punjab

Textbooks

- **Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer

Conditional probability distribution from a joint probability distribution

Let X and Y be two random variables, discrete or continuous. The **conditional distribution** of the random variable Y given that $X = x$ is

$$f(y|x) = \frac{P(X = x, Y = y)}{P(X = x)}, \text{ provided } P(X = x) > 0$$

Or

$$f(y|x) = \frac{f(x, y)}{g(x)}, \text{ provided } g(x) > 0$$

Conditional probability distribution from a joint probability distribution

Similarly, the conditional distribution of X given that $Y = y$ is

$f(x|y) = \frac{P(X=x, Y=y)}{P(Y=y)}$, provided $P(Y = y) > 0$

Or

$f(x|y) = \frac{f(x,y)}{h(y)}$, provided h(y) > 0

.

Checking Independence

Two random variables X and Y are independent if:

$P(X = x, Y = y) = P(X = x) \times P(Y = y)$

(for discrete variables)

$f(x, y) = f(x)_X \times f(y)_Y$

(for continuous variables)

Statistically Independent

Let X and Y be two **random variables**, **discrete or continuous**, with **joint probability distribution** $f(x, y)$ and **marginal distributions** $g(x)$ and $h(y)$, respectively.

The random variables X and Y are said to be **statistically independent** if and only if

$$f(x, y) = g(x)h(y)$$

for all (x, y) within their range.

Key Points on Joint Probability and Statistically Independent

- It is possible for the **product of the marginal distributions** to equal the **joint probability distribution for some but not all combinations of (x, y).**
- This is a crucial point when determining the statistical independence of variables.
- If you can find any point (x, y) for which f(x, y) is defined such that:
- **$f(x, y) \neq g(x)h(y)$**, the discrete variables X and Y are not **statistically independent**.

Example : Two ballpoint pens are selected at random from a box that contains **3 blue pens**, **2 red pens**, and **3 green pens**. If ***X*** is the **number of blue pens** selected and ***Y*** is the **number of red pens** selected,

(a) Find the joint probability function $f(x, y)$,

(b) Check the random variables X and Y are said to be **statistically independent**

(c) Check the point **(0, 1)** is statistically independent

(d) find the conditional distribution of X, given that $Y = 1$, and use it to determine $P(X = 0 \mid Y = 1)$.

Solution

a) The possible pairs of values (x, y) are (0, 0), (0, 1), (1, 0), (1, 1), (0, 2), and (2, 0).

The joint probability distribution of

$$f(x, y) = \frac{({}_3C_x)\,({}_2C_y)({}_3C_{2-x-y})}{{}_8C_2},$$ for $x = 0, 1, 2$; $y = 0, 1, 2$; and $0 \le x + y \le 2$.

$$f(0, 0) = \frac{({}_3C_0)({}_2C_0)({}_3C_{2-0-0})}{{}_8C_2} = \frac{3}{28}$$

$$f(0, 1) = \frac{({}_3C_0)({}_2C_1)({}_3C_{2-0-1})}{{}_8C_2} = \frac{6}{28}$$

$$f(1, 0) = \frac{({}_3C_1)({}_2C_0)({}_3C_{2-1-0})}{{}_8C_2} = \frac{9}{28}$$

$$f(1, 1) = \frac{({}_3C_1)({}_2C_1)({}_3C_{2-1-1})}{{}_8C_2} = \frac{6}{28}$$

$$f(0, 2) = \frac{({}_3C_0)({}_2C_2)({}_3C_{2-0-2})}{{}_8C_2} = \frac{1}{28}$$

$$f(2, 0) = \frac{({}_3C_2)({}_2C_0)({}_3C_{2-2-0})}{{}_8C_2} = \frac{3}{28}$$

f(x, y)		x			Row totals
		0	1	2	
y	0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
	1	$\frac{6}{28}$	$\frac{6}{28}$	0	$\frac{12}{28}$
	2	$\frac{1}{28}$	0	0	$\frac{1}{28}$
Column totals		$\frac{10}{28}$	$\frac{15}{28}$	$\frac{3}{28}$	$\frac{28}{28} = 1$

Solution : For the random variable X, we see that

$$g(x) = \sum_{y} f(x, y)$$

$$g(0) = f(0, 0) + f(0, 1) + f(0, 2)$$

$$= \frac{3}{28} + \frac{6}{28} + \frac{1}{28} = \frac{10}{28} = \frac{5}{14}$$

$$g(1) = f(1, 0) + f(1, 1) + f(1, 2)$$

$$= \frac{9}{28} + \frac{6}{28} + 0 = \frac{15}{28}$$

$$g(2) = f(2, 0) + f(2, 1) + f(2, 2)$$

$$= \frac{3}{28} + 0 + 0 = \frac{3}{28}$$

Marginal Distribution of x

x = 0	0	1	2	Total
g(x)	$\frac{10}{28}$	$\frac{15}{28}$	$\frac{3}{28}$	$\frac{28}{28} = 1$

For the random variable *y*, we see that

$$h(y) = \sum_{x} f(x, y)$$

$$h(0) = f(0, 0) + f(1, 0) + f(2, 0)$$

$$= \frac{3}{28} + \frac{9}{28} + \frac{3}{28} = \frac{15}{28}$$

$$h(1) = f(0, 1) + f(1, 1) + f(2, 1)$$

$$= \frac{6}{28} + \frac{6}{28} + 0 = \frac{12}{28}$$

$$h(2) = f(0, 2) + f(1, 2) + f(2, 2)$$

$$= \frac{1}{28} + 0 + 0 = \frac{1}{28}$$

Marginal Distribution of y

y = 0	0	1	2	Total
h(y)	$\frac{15}{28}$	$\frac{12}{28}$	$\frac{1}{28}$	$\frac{28}{28} = 1$

Check for Independence

To check if $f(x, y) = g(x) \times h(y)$ for all x and y:

$$f(0, 0) = \frac{3}{28} = 0.1071$$

$$g(0) \times h(0) = \frac{10}{28} \times \frac{15}{28} = \frac{75}{392} = 0.1913$$

Since **$f(0, 0) \neq g(0) \times h(0)$,** X and Y are NOT independent. X (number of blue pens) and Y (number of red pens) are **dependent variables**.

Check for Independence

Check for Independence at (0, 1):

$$f(0, 1) = \frac{6}{28} = \frac{3}{14} = 0.2143$$

$$g(0) \times h(1) = \frac{5}{14} \times \frac{3}{7} = \frac{15}{98} = 0.1531$$

Since $\frac{3}{14} \neq \frac{15}{98}$

X and Y are NOT independent at point (0, 1).

(d) find the conditional distribution of *X*, given that *Y* = 1, and use it to determine $P(X = 0 \mid Y = 1)$.

$$f(x|y) = \frac{f(x,y)}{h(y)}, \text{ provided } h(y) > 0$$

$$f(x|1) = \frac{f(x,1)}{h(1)}, x = 0, 1, 2$$

$$\mathbf{h(y)} = \sum_{x} f(x, y)$$

$$\Rightarrow \mathbf{h(y)} = \sum_{x=0}^{2} f(x, 1)$$

$$\mathbf{h(1)} = f(0, 1) + f(1, 1) + f(2, 1)$$

$$= \frac{6}{28} + \frac{6}{28} + 0 = \frac{12}{28} = \frac{3}{7}$$

$$f(x|1) = \frac{f(x,1)}{h(1)}, x = 0, 1, 2$$

$$f(0|1) = \frac{f(0,1)}{h(1)} = \frac{6/28}{\mathbf{12/28}} = \frac{1}{2}$$

$$f(1|1) = \frac{f(1,1)}{h(1)} = \frac{6/28}{\mathbf{12/28}} = \frac{1}{2}$$

$$f(2|1) = \frac{f(2,1)}{h(1)} = \frac{0}{\mathbf{12/28}} = 0$$

x	f(x | 1)
0	$\frac{1}{2}$
1	$\frac{1}{2}$
2	0

(d) $P(X = 0 \mid Y = 1)$

$$= \frac{f(0,1)}{h(1)}$$

$$= \frac{1}{2}$$

Example: The joint density for the random variables (X, Y), where ***X* is the unit temperature change** and ***Y* is the proportion of spectrum shift** that a certain atomic particle produces, is

$$f(x, y) = \begin{cases} 10xy^2, 0 < x < y < 1, \\ 0, eslewhere \end{cases}$$

(a) Find the marginal densities $g(x)$, $h(y)$, and the conditional density $f(y|x)$.

(b) Find the **probability that the spectrum shifts more than half of the total observations**, given that the temperature is increased by 0.25 unit.

(a) $\mathbf{g(x) = \int_{y=-\infty}^{y=+\infty} f(x,y)\,dy}$

$$g(x) = \int_{y=x}^{y=1} 10xy^2\,dy$$

$$= \left| \frac{10xy^3}{3} \right|_{y=x}^{y=1}$$

$$= \frac{10x(1^3 - x^3)}{3}$$

$$\mathbf{g(x) = \frac{10x(1-x^3)}{3},\ 0 < x < 1}$$

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

$$h(y) = \int_{x=0}^{x=y} 10xy^2\,dx$$

$$= \left| \frac{10x^2y^2}{2} \right|_{x=0}^{x=y}$$

$$= \frac{10y^2(y^2-0^2)}{2}$$

$$h(y) = 5y^4,\ 0 < y < 1$$

$$f(y|x) = \frac{f(x,y)}{g(x)}, \text{ provided } g(x) > 0$$

$$f(y|x) = \frac{10xy^2}{\frac{10}{3}x(1-x^3)}$$

$$f(y|x) = \frac{3y^2}{(1-x^3)}, 0 < x < y < 1$$

$$\boldsymbol{f(y|x) = \frac{3y^2}{(1-x^3)}, 0 < x < y < 1}$$

$$f(y|0.25) = \frac{3y^2}{(1-0.25^3)}, 0 < x < y < 1$$

$$f(y|0.25) = \frac{3y^2}{(1-0.25^3)}, 0 < x < y < 1$$

$$\boldsymbol{f(y|0.25) = \frac{3y^2}{0.9844}, 0 < x < y < 1}$$

$$P(Y > 0.5 \mid X = 0.25) = \int_{y=1/2}^{y=1} \frac{3y^2}{0.9844} dy$$

$$= \frac{3}{0.9844} \int_{y=1/2}^{y=1} y^2 dy$$

$$= \frac{3}{0.9844} \left| \frac{y^3}{3} \right|_{y=1/2}^{y=1}$$

$$= \frac{1}{0.9844} \left| y^3 \right|_{y=1/2}^{y=1}$$

$$= \frac{1}{0.9844} \{1^3 - (\frac{1}{2})^3\} = \frac{1}{0.9844} (\frac{7}{8})$$

$$= 0.889$$

Given the joint density function

$$f(x, y) = \begin{cases} \frac{x(1+3y^2)}{4}, 0 < x < 2, 0 < y < 1, \\ 0, eslewhere \end{cases}$$

find ***g*(*x*), *h*(*y*), *f*(x|*y*),** and evaluate **P($\frac{1}{4}$ < X < $\frac{1}{2}$ | Y = $\frac{1}{3}$).**

(a) $\mathbf{g(x)} = \int_{y=-\infty}^{y=+\infty} \mathbf{f(x,y)dy}$

$$g(x) = \int_{y=0}^{y=1} \frac{x(1+3y^2)}{4} dy$$

$$= \left| \frac{xy}{4} + \frac{xy^3}{4} \right|_{y=0}^{y=1}$$

$$= \frac{x(1)}{4} + \frac{x(1)^3}{4} - \frac{x(0)}{4} - \frac{x(0)^3}{4}$$

$\mathbf{g(x)} = \frac{\mathbf{x}}{\mathbf{2}}, \quad \mathbf{0 < x < 2}$

(b) $\mathbf{h(y)} = \int_{\mathbf{x}=-\infty}^{\mathbf{x}=+\infty} \mathbf{f(x,y)dx}$

$$h(y) = \int_{x=0}^{x=2} \frac{x(1+3y^2)}{4}\, dx$$

$$= \Big| \frac{x^2}{8} + \frac{3x^2y^2}{8} \Big|_{x=0}^{x=2}$$

$$= \frac{2^2}{8} + \frac{3(2^2)y^2}{8} - \frac{0}{8} - \frac{0}{8}$$

$$= \frac{1}{2} + \frac{3y^2}{2}$$

$$\boldsymbol{h(y) = \frac{1+3y^2}{2},\ 0 < y < 1}$$

$$\boldsymbol{f(x|y) = \frac{f(x,y)}{h(y)}}$$, **provided h(y) > 0**

$$f(x|y) = \frac{x(1+3y^2)/4}{(1+3y^2)/2}$$

$$f(x|y) = \frac{x}{2}$$

$$\Rightarrow \boldsymbol{f\left(x \mid \frac{1}{3}\right) = \frac{x}{2}}$$

$$P\left(\frac{1}{4} < X < \frac{1}{2} \mid Y = \frac{1}{3}\right) = \int_{x=1/4}^{x=1/2} \frac{x}{2}\, dx$$

$$= \left| \frac{x^2}{4} \right|_{x=1/4}^{x=1/2}$$

$$= \frac{3}{64}$$

Discrete Case

Let X and Y be random variables with joint probability distribution f(x, y).

The **mean, or expected value**, of the random variable g(X, Y) is:

$$\mu_g(X,Y) = E[g(X, Y)] = \sum_x \sum_y g(x, y)f(x, y)$$

if **X and Y are discrete**.

Continuous Case

Let X and Y be continuous random variables with **joint probability distribution f(x, y)**.

The **mean, or expected value**, of the random variable g(X, Y) is:

$$\mu_g(X,Y) = E[g(X, Y)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x, y) f(x, y)\,dx\,dy$$

if **X and Y are continuous**.

Example Let X and Y be the random variables with **joint probability distribution** given below. Find the **expected value of $g(X, Y) = XY$** .

f(x, y)		x			Row totals
		0	1	2	
y	0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
	1	$\frac{6}{28}$	$\frac{6}{28}$	0	$\frac{12}{28}$
	2	$\frac{1}{28}$	0	0	$\frac{1}{28}$
Column totals		$\frac{10}{28}$	$\frac{15}{28}$	$\frac{3}{28}$	$\frac{28}{28} = 1$

$\mu_g(X,Y) = E[g(X, Y)] = \sum_x \sum_y g(x, y)f(x, y)$

$\mu_g(X,Y) = \sum_{x=0}^{x=2} \sum_{y=0}^{y=2} xyf(x,y)$

E(XY) = (0)(0) × f(0,0) + (0)(1) × f(0,1) + (0)(2) × f(0,2) + (1)(0) × f(1,0) + **(1)(1) × f(1,1)** + (1)(2) × f(1,2) + (2)(0) × f(2,0) + (2)(1) × f(2,1) + (2)(2) × f(2,2)

E(XY) = $(0)(0) \times \frac{3}{28} + (0)(1) \times \frac{6}{28} + (0)(2) \times \frac{1}{28} + (1)(0) \times \frac{9}{28}$
$+ (1)(1) \times \frac{6}{28} + (1)(2) \times 0 + (2)(0) \times \frac{3}{28}$
$+ (2)(1) \times 0 + (2)(2) \times 0 = \frac{6}{28}$

E(XY) = $\frac{3}{14}$

Expected Value of Conditional Probability for Joint Prob Distribution

E(Y|X) = $\int_{y=-\infty}^{y=+\infty} yf(y|x)dy$

where

$$f(y|x) = \frac{f(x,y)}{g(x)}$$

E(X|Y) = $\int_{x=-\infty}^{x=+\infty} xf(x|y)dx$

where

$$f(x|y) = \frac{f(x,y)}{h(y)}$$

Example: Find E(Y|X) for the density function

$$f(x, y) = \begin{cases} \frac{x(1+3y^2)}{4}, 0 < x < 2, 0 < y < 1, \\ 0, eslewhere \end{cases}$$

$$\mathbf{E(Y|X)} = \int_{y=-\infty}^{y=+\infty} y f(y|x)\, dy$$

$$f(y|x) = \frac{f(x,y)}{g(x)}$$

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

$$g(x) = \int_{y=0}^{y=1} \frac{x(1+3y^2)}{4}\, dy$$

$$= \left| \frac{xy}{4} + \frac{xy^3}{4} \right|_{y=0}^{y=1}$$

$$= \frac{x(1)}{4} + \frac{x(1)^3}{4} - \frac{x(0)}{4} - \frac{x(0)^3}{4}$$

$$\mathbf{g(x)} = \frac{x}{2}, \quad \mathbf{0 < x < 2}$$

$$\boldsymbol{f(y|x) = \frac{f(x,y)}{g(x)}}\text{, } \textbf{provided g(x) > 0}$$

$$f(y|x) = \frac{x(1+3y^2)/4}{x/2}$$

$$f(y|x) = \frac{(1+3y^2)}{2}$$

$$E(Y|X) = \int_{\boldsymbol{y}=-\infty}^{\boldsymbol{y}=+\infty} yf(y|x)dy$$

$$= \int_{\text{y}=0}^{\text{y}=1} y \times \frac{(1+3y^2)}{2} dy$$

$$= \frac{1}{2} \int_{\text{y}=0}^{\text{y}=1} (y + 3y^3)dy$$

$$= \frac{1}{2} \{/ \frac{y^2}{2} + \frac{3y^4}{4} |_{y=0}^{y=1}\}$$

$$= \frac{1}{2} \{\frac{1^2}{2} + \frac{3(1)^4}{4} - 0 - 0\}$$

$$= \frac{5}{8}$$

Expected Value of X

E(X) = $\sum_x \sum_y x\mathbf{f(x, y)}$ $= \sum_x xg(x)$ **(discrete case)**

E(X) = $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x\mathbf{f(x, y)dx\ dy} = \int_{-\infty}^{+\infty} xg(x)$ **(continuous case)**

where g(x) is the marginal distribution of X.

Therefore, in calculating $E(X)$ over a two-dimensional space, one may use either the joint probability distribution of X and Y or the marginal distribution of X

Expected Value of Y

E(Y) = $\sum_y \sum_x y$f(x, y) = $\sum_y yh(y)$ (discrete case)

E(Y) = $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x$f(x, y)dx dy =$\int_{-\infty}^{+\infty} yh(y)$ (continuous case)

where h(y) is the marginal distribution of Y.

Therefore, in calculating $E(Y)$ over a two-dimensional space, one may use either the joint probability distribution of X and Y or the marginal distribution of Y

Feature Expectation in Naive Bayes Classification

In **Naive Bayes classifiers**, **expected value** helps calculate the **likelihood of a feature given a class**.

$$E[X_1, X_2] = \Sigma\, \Sigma\, x_1\, x_2 \times P(X_1 = x_1, X_2 = x_2)$$

Example:

$P(X_1 = 25, X_2 = 3000) = 0.1$, $P(X_1 = 30, X_2 = 4000) = 0.2$

$$E[X_1, X_2] = 25 \times 3000 \times 0.1 + 30 \times 4000 \times 0.2 = 31{,}500$$

Example: Consider a dataset where we have two clusters, C_1 and C_2, and we need to assign a **data point $X=(1.2,2.4)$** to one of these clusters based on the following information:

The **mean of cluster C_1** is $\mu_1 = (1,2)$

The **mean of cluster C_2** is $\mu_2 = (3,4)$

The probability of X belonging to cluster C_1, denoted by **$P(C_1|X)$** is calculated based on Gaussian likelihood.

For this example, we assume the **variance for both clusters is 1**, and we calculate **$P(C_1|X)$** using the formula for the **Gaussian distribution**:

$$P(C_1|X) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{(X-\mu_1)^2}{2\sigma^2}\right)$$

Where:

σ^2 is the variance (assumed 1)

$X=(1.2,2.4)$

$\mu_1=(1,2)$

P(C_1|X) = $\frac{1}{2\pi\sigma^2}\exp(-\frac{(X-\mu_1)^2}{2\sigma^2})$

For X = (1.2, 2.4) and μ_1 = (1, 2):

P(C_1 | X) ≈ 0.144

P(C_2 | X) = P(C_1|X) = $\frac{1}{2\pi\sigma^2}\exp(-\frac{(X-\mu_2)^2}{2\sigma^2})$

For X = (1.2, 2.4) and μ_2 = (3, 4):

P(C_2 | X) ≈ 0.00875

$P(X) = P(C_1 \mid X) + P(C_2 \mid X)$

For $P(C_1 \mid X) \approx 0.144$ and **$P(C_2 \mid X) \approx 0.00875$:**

$P(X) \approx 0.15275$

$$E[X \mid C_1] = \frac{P(C_1 \mid X)}{P(X)}$$

For **$P(C_1 \mid X) \approx 0.144$** and **$P(X) \approx 0.15275$**:

$E[X \mid C_1] \approx 0.943$

Conclusion: The data point X = (1.2, 2.4) has a 94.3% probability of belonging to cluster C_1 based on the expected value calculation.