

# **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

## Target Journals

Some of the journals that are relevant to health care and the medical field, based on computer science.

1. **Medical Decision Making**, JCR Impact Factor (2017-18) = 2.793
2. **Health Informatics Journal**, JCR Impact Factor (2017-18) = 2.297
3. **Informatics for Health and Social Care**, JCR Impact Factor (2017-18) = 1.137
4. **Health Care Analysis**, , JCR Impact Factor (2017-18) = 1.043
5. **International Journal of Health Care Quality Assurance**, JCR Impact Factor (2017-18) = 1.218

## Target Journals

Some of the journals that are relevant to education, based on computer science.

1. **Computers & Education**, JCR Impact Factor (2017-18) = 5.627
2. **Computer Applications in Engineering Education**, JCR Impact Factor (2017-18) = 1.435
3. **Journal of Computing in Higher Education**, JCR Impact Factor (2017-18) = 1.870
4. **Acm Transactions on Computing Education**, JCR Impact Factor (2017-18) = 1.356
5. **Assessment & Evaluation In Higher Education**, JCR Impact Factor (2017-18) = 2.473
6. **Educational Assessment Evaluation and Accountability**, JCR Impact Factor (2017-18) = 1.772
7. **Computer Applications in Engineering Education** = Impact Factor: 1.435

# What is Bayes' Law?

Formula:  $P(A | B) = \frac{P(B | A)P(A)}{P(B)}$

Bayes' Law is a formula for determining **conditional probabilities**.

## Components:

**P(A | B)**: Posterior probability

**P(B | A)**: Likelihood

**P(A)**: Prior probability

**P(B)**: Marginal likelihood

# Scenario: Classifying Emails as Spam or Not Spam

You are tasked with classifying an email as spam or not spam using Bayes' Law.

Given:

**Prior:**  $P(\text{Spam}) = 0.2$

**Likelihood:**  $P(\text{Word} = \text{"win"} \mid \text{Spam}) = 0.8$

**Marginal likelihood:**  $P(\text{Word} = \text{"win"}) = 0.3$

# Objective

Calculate the **posterior probability** of the **email being spam**, given that the word "**win**" appears in the email.

**Calculate:**  $P(\text{Spam} \mid \text{Word} = \text{"win"})$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(\text{Spam} \mid \text{Word} = \text{"win"}) = \frac{P(\text{Word} = \text{"win"} \mid \text{Spam}) \times P(\text{Spam})}{P(\text{Word} = \text{"win"})}$$

$$P(\text{Spam} \mid \text{Word} = \text{'win'}) = \frac{0.80 \times 0.20}{0.30}$$

$$= 0.533$$



# Joint Probability Distributions

- Given random variables  $X, Y, \dots$ , that are defined on a probability space, the **joint probability distribution** for  $X, Y, \dots$ , is a **probability distribution** that gives the probability that each of  $X, Y, \dots$ , falls in any particular range or discrete set of values specified for that variable.
- In the case of only **two random variables**, this is called a **bivariate distribution**, but the concept generalizes to **any number of random variables**, giving a **multivariate distribution**.

# Joint Probability Distributions

- The **joint probability distribution** can be expressed either in terms of a **joint cumulative distribution function** or in terms of a **joint probability density function** (in the case of continuous variables) or **joint probability mass function** (in the case of discrete variables).

# Joint Probability Distributions cont.

- These in turn can be used to find two other types of distributions: the **marginal distribution** giving the probabilities for any one of the variables with **no reference to any specific ranges of values for the other variables**, and the **conditional probability distribution** giving the probabilities for any subset of the variables conditional on particular values of the remaining variables

# Joint Probability Distributions

- ❑ Our study of random variables and their probability distributions in previous lectures is restricted to **one-dimensional sample spaces**, in that we recorded outcomes of an experiment as values assumed by a **single random variable**.
- ❑ There will be situations, however, where we may find it desirable to record the **simultaneous outcomes** of **several random variables**.

# Joint Probability Distributions

- ❑ **For example**, we might measure the amount of precipitate  $P$  and volume  $V$  of gas released from a controlled chemical experiment, giving rise to a **two-dimensional sample space** consisting of the **outcomes  $(p, v)$** , or
- ❑ we might be interested in the **hardness  $H$**  and **tensile strength  $T$**  of cold-drawn copper, resulting in the **outcomes  $(h, t)$** .

# Joint Probability Distributions

- ❑ In a study to determine the likelihood of **success in college** based on high school data, we might use a **three dimensional sample space** and record for **each individual** his or her **aptitude test score**, **high school class rank**, and **grade-point average** at the end of freshman year in college.

# Joint Probability Distributions

- If  $X$  and  $Y$  are two discrete random variables, the probability distribution for their simultaneous occurrence can be represented by a function with values  $f(x, y)$  for any pair of values  $(x, y)$  within the range of the random variables  $X$  and  $Y$ .
- It is customary to refer to this function as the joint probability distribution of  $X$  and  $Y$ . Hence, in the discrete case,  
$$f(x, y) = P(X = x, Y = y);$$
that is, the values  $f(x, y)$  give the probability that outcomes  $x$  and  $y$  occur at the same time.

# Joint Probability Distributions

The **function  $f(x, y)$**  is a **joint probability distribution or probability mass function** of the **discrete random variables  $X$  and  $Y$**  if

1.  $f(x, y) \geq 0$  for all  $(x, y)$ ,
2.  $\sum_x \sum_y f(x, y) = 1$ ,
3.  $P(X = x, Y = y) = f(x, y)$ .

For any region  $A$  in the  $xy$  plane,

$$P[(X, Y) \in A] = \sum \sum_A f(x, y).$$



**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, find

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

(b)  $P[(X, Y) \in A]$ , where  $A$  is the region  $\{(x, y) / x + y \leq 1\}$ .

## 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

$$f(x, y) = \frac{({}_3C_x)({}_2C_y)({}_3C_{2-x-y})}{8C_2}, \text{ for } x = 0, 1, 2; y = 0, 1, 2; \text{ and } 0 \leq x + y \leq 2.$$

$$f(0, 0) = \frac{({}_3C_0)({}_2C_0)({}_3C_{2-0-0})}{8C_2} = \frac{3}{28}$$

$$f(0, 1) = \frac{({}_3C_0)({}_2C_1)({}_3C_{2-0-1})}{8C_2} = \frac{6}{28}$$

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

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

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

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

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

(b)  $P[(X, Y) \in A]$ , where  $A$  is the region  $\{(x, y) / x + y \leq 1\}$ .

The probability that  $(X, Y)$  fall in the region  $A$  is

$$\begin{aligned} P[(X, Y) \in A] &= P(X + Y \leq 1) \\ &= f(0, 0) + f(0, 1) + f(1, 0) \end{aligned}$$

$$= \frac{3}{28} + \frac{6}{28} + \frac{9}{28}$$

$$= \frac{18}{28}$$

$$= \frac{9}{14}$$

# Joint Density Function

The function  $f(x, y)$  is a **joint density function** of the **continuous random variables  $X$  and  $Y$**  if

1.  $f(x, y) \geq 0$ , for all  $(x, y)$ ,

2.  $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$ ,

3.  $P[(X, Y) \in A] = \int \int_A f(x, y) dx dy$ , for any region  $A$  in the  $xy$  plane.

**Example:** A privately owned business operates both a drive-in facility and a walk-in facility. On a randomly selected day, let  $X$  and  $Y$ , respectively, be the proportions of the time that the drive-in and the walk-in facilities are in use, and suppose that the joint density function of these random variables is

$$f(x, y) = \begin{cases} \frac{2}{5} (2x + 3y), & 0 \leq x \leq 1, 0 \leq y \leq 1. \\ 0, & \text{elsewhere} \end{cases}$$

(a) Verify  $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$ .

(b)  $P[(X, Y) \in A]$ , where  $A = \{(x, y) \mid 0 < x < \frac{1}{2}, \frac{1}{4} < y < \frac{1}{2}\}$



## Solution

$$\begin{aligned}\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy &= \int_{y=0}^1 \left\{ \int_{x=0}^1 \frac{2}{5} (2x + 3y) dx \right\} dy \\&= \frac{2}{5} \int_{y=0}^1 \left| \frac{(2)x^2}{2} + 3xy \right|_{x=0}^1 dy \\&= \frac{2}{5} \int_0^1 \{1^2 + 3(1)y\} dy - 0 \\&= \frac{2}{5} \left| y + 3 \frac{y^2}{2} \right|_0^1 \\&= \frac{2}{5} \left( 1 + \frac{3}{2} \right) - 0 \\&= 1\end{aligned}$$

$$P[(X, Y) \in A] = P(0 < x < \frac{1}{2}, \frac{1}{4} < y < \frac{1}{2})$$

$$= \int_{y=\frac{1}{4}}^{\frac{1}{2}} \left\{ \int_{x=0}^{\frac{1}{2}} \frac{2}{5} (2x + 3y) dx \right\} dy$$

$$= \frac{2}{5} \left\{ \int_{y=\frac{1}{4}}^{\frac{1}{2}} \left[ \frac{2x^2}{2} + 3xy \right]_{x=0}^{\frac{1}{2}} dy \right\}$$

$$= \frac{2}{5} \int_{y=\frac{1}{4}}^{\frac{1}{2}} (x^2 + 3xy)_{x=0}^{\frac{1}{2}} dy$$

$$\begin{aligned}
 P[(X, Y) \in A] &= \frac{2}{5} \int_{\frac{1}{4}}^{\frac{1}{2}} \left\{ \frac{1}{4} + 3 \left( \frac{1}{2} \right) y \right\} dy - 0 \\
 &= \frac{2}{5} \int_{\frac{1}{4}}^{\frac{1}{2}} \left\{ \frac{1}{4} + \frac{3}{2} y \right\} dy \\
 &= \frac{2}{5} \left| \frac{1}{4} y + \frac{3}{4} y^2 \right|_{\frac{1}{4}}^{\frac{1}{2}}
 \end{aligned}$$

$$\begin{aligned}
&= \frac{2}{5} \left| \frac{1}{4} y + \frac{3}{4} y^2 \right|^{\frac{1}{2}} \Big|_{\frac{1}{4}}^{\frac{1}{2}} \\
&= \frac{2}{5} \left\{ \frac{1}{8} + \left(\frac{3}{4}\right) \left(\frac{1}{2}\right)^2 - \frac{1}{16} - \left(\frac{3}{4}\right) \left(\frac{1}{4}\right)^2 \right\} \\
&= \frac{2}{5} \left( \frac{1}{8} + \frac{3}{16} - \frac{1}{16} - \frac{3}{64} \right) \\
&= \frac{2}{5} \left( \frac{2+12-4+3}{64} \right) \\
&= \frac{(2)(13)}{(5)(64)} \\
&= \frac{13}{160}
\end{aligned}$$

# The marginal distributions of $X$ alone and of $Y$ alone are

- Given the **joint probability distribution**  $f(x, y)$  of the discrete random variables  $X$  and  $Y$ , the **probability distribution**  $g(x)$  of  $X$  alone is obtained by **summing**  $f(x, y)$  over the **values of  $Y$** .
- Similarly, the probability distribution  $h(y)$  of  $Y$  alone is obtained by **summing**  $f(x, y)$  over the **values of  $X$** . We define  $g(x)$  and  $h(y)$  to be the **marginal distributions** of  $X$  and  $Y$ , respectively.

# The marginal distributions of $X$ alone and of $Y$ alone are

- When  $X$  and  $Y$  are continuous random variables, summations are replaced by integrals.

# The marginal distributions of $X$ alone and of $Y$ alone are

$g(x) = \sum_y f(x, y)$  and  $h(y) = \sum_x f(x, y)$  for discrete case  
, and

$$g(x) = \int_{y=-\infty}^{\infty} f(x, y) dy \text{ and } h(y) = \int_{x=-\infty}^{\infty} f(x, y) dx$$

for the continuous case.

# The marginal distributions of $X$ alone and of $Y$ alone are

□ **Note:** The term *marginal* is used here because, in the **discrete case**, the values of  $g(x)$  and  $h(y)$  are just the **marginal totals** of the respective **columns** and **rows** when the values of  $f(x, y)$  are displayed in a **rectangular table**.



□ **Example** : Show that the column and row totals of the table in the coming slide give the **marginal distribution** of  **$X$**  alone and of  **$Y$**  alone.

f(x, y)		x			Row totals
		0	1	2	
y	0	<div>3</div> <div>—</div> <div>28</div>	<div>9</div> <div>—</div> <div>28</div>	<div>3</div> <div>—</div> <div>28</div>	<div>15</div> <div>—</div> <div>28</div>
	1	<div>6</div> <div>—</div> <div>28</div>	<div>6</div> <div>—</div> <div>28</div>	<div>0</div> <div>—</div> <div>28</div>	<div>12</div> <div>—</div> <div>28</div>
	2	<div>1</div> <div>—</div> <div>28</div>	<div>0</div> <div>—</div> <div>28</div>	<div>0</div> <div>—</div> <div>28</div>	<div>1</div> <div>—</div> <div>28</div>
Column totals		<div>10</div> <div>—</div> <div>28</div>	<div>15</div> <div>—</div> <div>28</div>	<div>3</div> <div>—</div> <div>28</div>	<div>28</div> <div>—</div> <div>28</div> <div>= 1</div>

**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
$h(y)$	$\frac{15}{28}$	$\frac{12}{28}$	$\frac{1}{28}$	$\frac{28}{28} = 1$

**Example** : Find  $g(x)$  and  $h(y)$  for the joint density function

$$f(x, y) = \begin{cases} \frac{2}{5} (2x + 3y), & 0 \leq x \leq 1, 0 \leq y \leq 1. \\ 0, & \text{elsewhere} \end{cases}$$

# Marginal Density of x

## Solution

$$\begin{aligned}g(x) &= \int_{y=-\infty}^{\infty} f(x, y) dy \\&= \int_{y=0}^1 \frac{2}{5} (2x + 3y) dy \\&= \left| \frac{2}{5} (2xy + \frac{3y^2}{2}) \right|_{y=0}^1 \\&= \frac{2}{5} \left\{ 2x(1) + \frac{3(1)^2}{2} \right\} - 0 \\&= \frac{2}{5} \left( \frac{4x + 3}{2} \right) \\&= \frac{4x + 3}{5}\end{aligned}$$



# Marginal Density of y

$$\begin{aligned}h(y) &= \int_{x=-\infty}^{\infty} f(x, y) dx \\&= \int_{x=0}^1 \frac{2}{5} (2x + 3y) dx \\&= \frac{2}{5} \left| \frac{2x^2}{2} + 3xy \right|_{x=0}^1 \\&= \frac{2}{5} \{1 + 3(1)y\} - 0 \\&= \frac{2}{5} (1 + 3y)\end{aligned}$$