BITS Pilani
Pilani Campus

# Multivariate Analytics


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.

# 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 it is legitimate.

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

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

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

$= \frac{7}{640}$

# 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

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

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

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

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

$$\therefore c = 1/210$$

# Joint cumulative distribution function

- Suppose X and Y are jointly-distributed random variables.

- We will use the notation ‘$X \leq x, Y \leq y$’ to mean the event ‘$X \leq x$ and $Y \leq y$’.

- The joint cumulative distribution function (joint cdf) is defined as $F(x, y) = P(X \leq x, Y \leq y)$

# Joint cumulative distribution function

- If X and Y are continuous random variables with joint density f(x, y) over the range [a, b] × [c, d] then the joint cdf is given by the double integral

$$F(x,y) = \int_c^y \int_a^x f(u,v)\,du\,dv.$$

- If X and Y are discrete random variables with joint pmf $p(x_i, y_j)$ then the joint cdf is give by the double sum

$$F(x,y) = \sum_{x_i \le x} \sum_{y_j \le y} p(x_i, y_j).$$

# Properties of the joint cdf

The joint cdf $F(x, y)$ of $X$ and $Y$ must satisfy several properties:

1. $F(x, y)$ is non-decreasing: i.e. if $x$ or $y$ increase then $F(x, y)$ must stay constant or increase.

2. $F(x, y) = 0$ at the lower-left of the joint range.
   If the lower left is $(-\infty, -\infty)$ then this means $\lim_{(x,y)\to(-\infty,-\infty)} F(x, y) = 0.$

3. $F(x, y) = 1$ at the upper-right of the joint range.
   If the upper-right is $(\infty, \infty)$ then this means $\lim_{(x,y)\to(\infty,\infty)} F(x, y) = 1.$

# Exercise

A nut company markets cans of deluxe mixed nuts containing almonds, cashews, and peanuts. Suppose the net weight of each can is exactly 1 lb, but the weight contribution of each type of nut is random. Because the three weights sum to 1, a joint probability model for any two gives all necessary information about the weight of the third type. Let $X$ = the weight of almonds in a selected can and $Y$ = the weight of cashews. Then the region of positive density is $D = \{(x, y): 0 \le x \le 1, 0 \le y \le 1, x + y \le 1\}$, the shaded region pictured in Figure 5.2.

Now let the joint pdf for $(X, Y)$ be

$$f(x, y) = \begin{cases} 24xy & 0 \le x \le 1, 0 \le y \le 1, x + y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

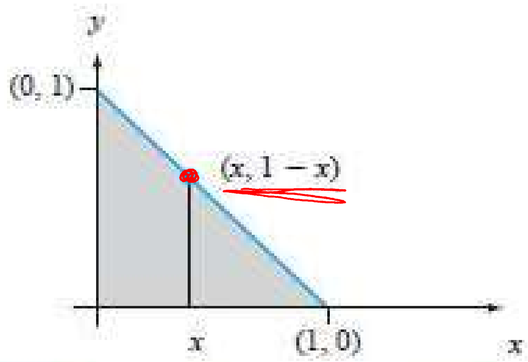


Figure 5.2 Region of positive density for Example

# Solution

$P(x+y=1)$

$$\int_0^1\left\{\int_0^{1-x} 24xy\,dy\right\}dx = \int_0^1 24x\left[\frac{y^2}{2}\right]_0^{1-x}dx$$

$$= \int_0^1 12x(1-x)^2\,dx = \underline{\underline{1}}$$

To compute the probability that the two types of nuts together make up at most 50% of the can, let $A = \{(x, y): 0 \le x \le 1,\ 0 \le y \le 1, \text{ and } x + y \le .5\}$, as shown in Figure 5.3. Then

$$P((X, Y) \in A) = \iint_A f(x, y)\,dx\,dy = \int_0^{.5}\int_0^{.5-x} 24xy\,dy\,dx = .0625$$

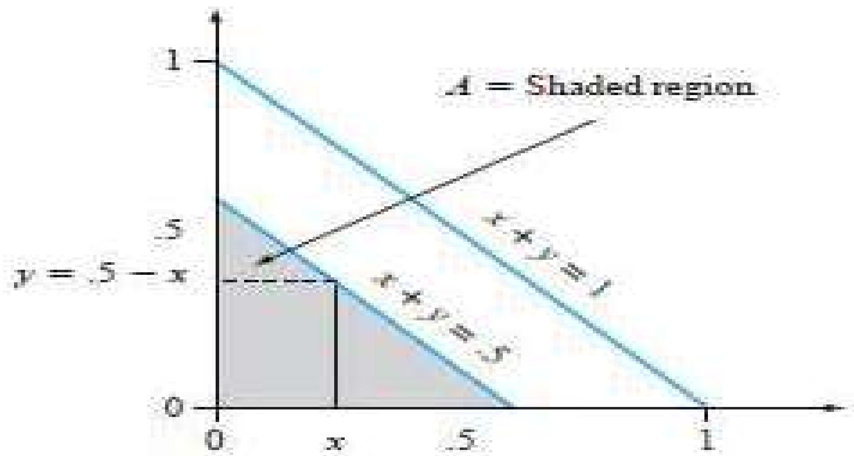


**Figure 5.3** Computing $P[(X, Y) \in A]$ for Example 5.5

# Independence

- Events A and B are independent if $P(A \cap B) = P(A)P(B)$.

- Jointly-distributed random variables X and Y are independent if their joint cdf is the product of the marginal cdf's $F(X, Y) = F_X(x)F_Y(y)$.

- For discrete variables this is equivalent to the joint pmf being the product of the marginal pmf's $p(x_i, y_j) = p_X(x_i)p_Y(y_j)$.

- For continuous variables this is equivalent to the joint pdf being the product of the marginal pdf's $f(x, y) = f_X(x)f_Y(y)$.

# Exercise

- Consider two random variables X and Y with joint PMF given in Table

| | $Y=2$ | $Y=4$ | $Y=5$ |
|---|---|---|---|
| $X=1$ | $\frac{1}{12}$ | $\frac{1}{24}$ | $\frac{1}{24}$ |
| $X=2$ | $\frac{1}{6}$ | $\frac{1}{12}$ | $\frac{1}{8}$ |
| $X=3$ | $\frac{1}{4}$ | $\frac{1}{8}$ | $\frac{1}{12}$ |

- Find $P(X \leq 2, Y \leq 4)$.
- Find the marginal PMFs of X and Y.
- Find $P(Y=2|X=1)$.
- Are X and Y independent?

$$P(Y|X) = \frac{P(Y \cap X)}{P(X)}$$

Ref: https://www.probabilitycourse.com

# Solution

innovate achieve lead

To find $P(X \leq 2, Y \leq 4)$, we can write

$$P(X \leq 2, Y \leq 4) = P_{XY}(1,2) + P_{XY}(1,4) + P_{XY}(2,2) + P_{XY}(2,4)$$
$$= \frac{1}{12} + \frac{1}{24} + \frac{1}{6} + \frac{1}{12} = \frac{3}{8}.$$

$$P_X(x) = \begin{cases} \frac{1}{6} & x = 1 \\ \frac{3}{8} & x = 2 \\ \frac{11}{24} & x = 3 \\ 0 & \text{otherwise} \end{cases}$$

$$P_Y(y) = \begin{cases} \frac{1}{2} & y = 2 \\ \frac{1}{4} & y = 4 \\ \frac{1}{4} & y = 5 \\ 0 & \text{otherwise} \end{cases}$$

$$P(Y = 2|X = 1) = \frac{P(X = 1, Y = 2)}{P(X = 1)}$$
$$= \frac{P_{XY}(1,2)}{P_X(1)}$$
$$= \frac{\frac{1}{12}}{\frac{1}{6}} = \frac{1}{2}.$$

$$P(X = 2, Y = 2) = \frac{1}{6} \neq P(X = 2)P(Y = 2) = \frac{3}{16}.$$

Thus, we conclude that $X$ and $Y$ are not independent.

# Expected Value

Let $X$ and $Y$ be jointly distributed rv's with pmf $p(x, y)$ or pdf $f(x, y)$ according to whether the variables are discrete or continuous. Then the expected value of a function $h(X, Y)$, denoted by $E[h(X, Y)]$ or $\mu_{h(X,Y)}$, is given by

$$E[h(X, Y)] = \begin{cases} \sum_x \sum_y h(x, y) \cdot p(x, y) & \text{if } X \text{ and } Y \text{ are discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) \cdot f(x, y)\, dx\, dy & \text{if } X \text{ and } Y \text{ are continuous} \end{cases}$$

# Example

- Five friends have purchased tickets to a certain concert. If the tickets are for seats 1–5 in a particular row and the tickets are randomly distributed among the five, what is the expected number of seats separating any particular two of the five? Let *X* and *Y* denote the seat numbers of the first and second individuals, respectively. Possible (*X*, *Y*) pairs are {(1, 2), (1, 3), . . . , (5, 4)}, and the joint pmf of (*X*, *Y*) is

$$p(x, y) = \begin{cases} \frac{1}{20} & x = 1, \ldots, 5;\ y = 1, \ldots, 5;\ x \neq y \\ 0 & \text{otherwise} \end{cases}$$

The number of seats separating the two individuals is $h(X, Y) = |X - Y| - 1$. The accompanying table gives $h(x, y)$ for each possible $(x, y)$ pair.

# Solution

innovate achieve lead

$|X-Y|-1$ $\qquad$ $x \neq y$

| $h(x, y)$ | | $x$ = 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| $y$ | 1 | — | 0 | 1 | 2 | 3 |
| | 2 | 0 | — | 0 | 1 | 2 |
| | 3 | 1 | 0 | — | 0 | 1 |
| | 4 | 2 | 1 | 0 | — | 0 |
| | 5 | 3 | 2 | 1 | 0 | — |

$$\sum_{(x,y)}\sum h(x,y).p(x,y)$$

$$\frac{1}{20}(1+2+3+1+2+1+1+2+1+3+2+1)$$

$$=1$$

# Exercise (HW)

- Suppose that 2 batteries are randomly chosen without replacement from the following group of 12 batteries:
  - 3 new
  - 4 used (working)
  - 5 defective
- Let X denote the number of new batteries chosen. Let Y denote the number of used batteries chosen.
- Find $f_{XY}(x, y)$

# Solution

- Though X can take on values 0, 1, and 2, and Y can take on values 0, 1, and 2, when we consider them jointly, X + Y ≤ 2. So, not all combinations of (X, Y ) are possible.

There are 6 possible cases...

CASE: no new, no used (so all defective)

$$f_{XY}(0,0) = \frac{\binom{5}{2}}{\binom{12}{2}} = 10/66$$

CASE: no new, 1 used

$$f_{XY}(0,1) = \frac{\binom{4}{1}\binom{5}{1}}{\binom{12}{2}} = 20/66$$

CASE: no new, 2 used

$$f_{XY}(0,2) = \frac{\binom{4}{2}}{\binom{12}{2}} = 6/66$$

CASE: 2 new, no used

$$f_{XY}(2,0) = \frac{\binom{3}{2}}{\binom{12}{2}} = 3/66$$

CASE: 1 new, no used

$$f_{XY}(1,0) = \frac{\binom{3}{1}\binom{5}{1}}{\binom{12}{2}} = 15/66$$

CASE: 1 new, 1 used

$$f_{XY}(1,1) = \frac{\binom{3}{1}\binom{4}{1}}{\binom{12}{2}} = 12/66$$

x= number of *new* chosen

| | | 0 | 1 | 2 |
|---|---|---|---|---|
| y=number of | 0 | 10/66 | 15/66 | 3/66 |
| *used* | 1 | 20/66 | 12/66 | |
| chosen | 2 | 6/66 | | |

There are 6 possible $(X, Y)$ pairs.
And, $\sum_x \sum_y f_{XY}(x, y) = 1$.

# Exercise

Compute E(X) and E(Y ).

Compute E(XY ).

| $x$ | $y$ | $P(X=x, Y=y)$ |
|---|---|---|
| 0 | 1 | 0.2 |
| 0 | 2 | 0.1 |
| 1 | 1 | 0.0 |
| 1 | 2 | 0.2 |
| 2 | 1 | 0.3 |
| 2 | 2 | 0.2 |

$E(X) = 0 \times 0.2 + 0 \times 0.1 + 1 \times 0.0 + 1 \times 0.2 + 2 \times (0.5)$

$= 1.2$

# Solution

$$E(Y) = 1 \times (0.2 + 0.0 + 0.3) + 2 \times (0.1 + 0.2 + 0.2) = 1.5$$

$$E(XY) = 0 \times 1 \times 0.2 + 0 \times 2 \times 0.1 + 1 \times 1 \times 0.0 + 1 \times 2 \times 0.2 + 2 \times 1 \times 0.3 + 2 \times 2 \times 0.2$$

$$= 1.8$$

# Covariance

- When two random variables $X$ and $Y$ are not independent, it is frequently of interest to assess how strongly they are related to one another.

The **covariance** between two rv's $X$ and $Y$ is

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

$$= \begin{cases} \displaystyle\sum_x \sum_y (x - \mu_X)(y - \mu_Y)\, p(x, y) & X, Y \text{ discrete} \\ \displaystyle\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y)\, f(x, y)\, dx\, dy & X, Y \text{ continuous} \end{cases}$$

# Example

| $p(x, y)$ | | $y$: 0 | 100 | 200 |
|---|---|---|---|---|
| $x$ | 100 | .20 | .10 | .20 |
| | 250 | .05 | .15 | .30 |

| $x$ | 100 | 250 |
|---|---|---|
| $p_X(x)$ | .5 | .5 |

| $y$ | 0 | 100 | 200 |
|---|---|---|---|
| $p_Y(y)$ | .25 | .25 | .5 |

$$\mu_y = 0 \times 0.25 + 100 \times 0.25 + 200 \times 0.5 = 125$$

from which $\mu_X = \Sigma x p_X(x) = 175$ and $\mu_Y = 125$. Therefore,

Covariance???

$$\text{Cov}(X,Y) = \sum_{(x,y)}\sum (x-175)(y-125)\, p(x,y)$$

$$= (100-175)(0-125)(0.2) + \ldots\ldots + \ldots$$

$$+ (250-175)(200-125)(0.3)$$

The following shortcut formula for $\text{Cov}(X, Y)$ simplifies the computations.

$$\boxed{\text{Cov}(X, Y) = E(XY) - \mu_X \cdot \mu_Y}$$

# Covariance

Cov(X, Y)= E(XY) - E(X). E(Y) where

$$E(X) = \int_{-\infty}^{\infty} x\, f_X(x)\, dx$$

$$E(Y) = \int_{-\infty}^{\infty} y\, f_Y(y)\, dy$$

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f(x, y)\, dx\mathrm{dy}$$

$$V(X) = E(X^2) - [E(X)]^2$$

$$V(Y) = E(Y^2) - [E(Y)]^2$$

# Correlation

The **correlation coefficient** of $X$ and $Y$, denoted by Corr$(X, Y)$, $\rho_{X,Y}$, or just $\rho$, is defined by

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

- If $X$ and $Y$ are independent, then correlation is 0, but if correlation is 0 that does not imply independence.

# Exercise

A large insurance agency services a number of customers who have purchased both a homeowner's policy and an automobile policy from the agency. For each type of policy, a deductible amount must be specified. For an automobile policy, the choices are \$100 and \$250, whereas for a homeowner's policy, the choices are 0, \$100, and \$200. Suppose an individual with both types of policy is selected at random from the agency's files. Let $X$ the deductible amount on the auto policy and $Y$ the deductible amount on the homeowner's policy

Suppose the joint pmf is given by

| $p(x, y)$ | | $y$: 0 | 100 | 200 |
|---|---|---|---|---|
| $x$ | 100 | .20 | .10 | .20 |
| | 250 | .05 | .15 | .30 |

Find

(i) Marginal probabilities of X and Y.

(ii) $\mathrm{P}(Y \geq 100)$

$P_X(100) = 0.2 + 0.1 + 0.2$
$= 0.5$

$P_X(250) = 0.05 + 0.15 + 0.30$
$= 0.5$

# Solution

$$p_X(x) = \begin{cases} 0.5 & x = 100, 250 \\ 0 & \text{otherwise} \end{cases}$$

$$p_Y(y) = \begin{cases} 0.25 & y = 0, 100 \\ 0.5 & y = 200 \\ 0 & \text{otherwise} \end{cases}$$

$$P(Y \geq 100) = p(100, 100) + p(250, 100) + p(100, 200) + p(250, 200)$$

$$= 0.75$$

# Exercise (HW)

- When a certain method is used to collect a fixed volume of rock samples in a region, there are four resulting rock types. Let $X1$, $X2$, and $X3$ denote the proportion by volume of rock types 1, 2, and 3 in a randomly selected sample (the  proportion of rock type 4 is 1  $X1$  $X2$  $X3$, so a variable $X4$ would be redundant). If the joint pdf of $X1$, $X2$, $X3$ is

$$f(x_1, x_2, x_3) = \begin{cases} kx_1x_2(1 - x_3) & 0 \le x_1 \le 1, 0 \le x_2 \le 1, 0 \le x_3 \le 1, x_1 + x_2 + x_3 \le 1 \\ 0 & \text{otherwise} \end{cases}$$

then $k$ is determined by

- The probability that rocks of types 1 and 2 together account for at most 50% of the sample is

# Solution

$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2, x_3)\, dx_3\, dx_2\, dx_1$$

$$= \int_0^1\left\{\int_0^{1-x_1}\left[\int_0^{1-x_1-x_2} kx_1x_2(1-x_3)\, dx_3\right] dx_2\right\} dx_1$$

This iterated integral has value $k/144$, so $k$ 144.

$$P(X_1 + X_2 \le .5) = \iiint\limits_{\left\{\substack{0 \le x_i \le 1 \text{ for } i=1,2,3 \\ x_1+x_2+x_3 \le 1,\, x_1+x_2 \le .5}\right\}} f(x_1, x_2, x_3)\, dx_3\, dx_2\, dx_1$$

$$= \int_0^{.5}\left\{\int_0^{.5-x_1}\left[\int_0^{1-x_1-x_2} 144x_1x_2(1-x_3)\, dx_3\right] dx_2\right\} dx_1$$

$$= .6066$$

innovate achieve lead

BITS Pilani
Pilani Campus

# Multivariate normal distribution

# Univariate Normal Distribution

- The normal distribution , also known as the Gaussian distribution, is so called because its based on the Gaussian function .

- This distribution is defined by two parameters: the mean μ, which is the expected value of the distribution, and the standard deviation σ, which corresponds to the expected deviation from the mean.

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Given this mean and variance we can calculate the probability density function (pdf) of the normal distribution with the normalised Gaussian function. For a value x the density is:

# Multivariate normal distribution

- The multivariate normal distribution is a multidimensional generalisation of the one-dimensional normal distribution .
- “a random vector is said to be r-variate normally distributed if every linear combination of its r components has a univariate normal distribution”.
- A bivariate normal distribution is made up of two independent random variables. The two variables in a bivariate normal are both are normally distributed, and they have a normal distribution when both are added together.
- **Visually**, the bivariate normal distribution is a **three-dimensional bell curve.**

# PDF of the Bivariate Normal Distribution

- The bivariate normal distribution can be defined as the probability density function (PDF) of two variables X and Y that are linear functions of the same independent normal random variables:

$$P(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{z}{2(1-\rho^2)}\right],$$

- Where σ is standard deviation
- ρ correlation of x1 and x2.

$$\rho \equiv \operatorname{cor}(x_1, x_2) = \frac{V_{12}}{\sigma_1 \sigma_2}$$

$$z \equiv \frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2},$$

# Probability for Bivariate

Probabilities still relate to the area under the pdf:

$$P([a_x \leq X \leq b_x] \text{ and } [a_y \leq Y \leq b_y]) = \int_{a_x}^{b_x} \int_{a_y}^{b_y} f(x,y) dy dx \quad (6)$$

where $\int \int f(x,y) dy dx$ denotes the multiple integral of the pdf $f(x,y)$.

Defining $\mathbf{z} = (x,y)$, we can still define the cdf:

$$\begin{aligned} F(\mathbf{z}) &= P(X \leq x \text{ and } Y \leq y) \\ &= \int_{-\infty}^{x} \int_{-\infty}^{y} f(u,v) dv du \end{aligned} \quad (7)$$

# Conditional distribution

The conditional distribution of a variable $Y$ given $X = x$ is

$$f_{Y|X}(y|X = x) = \frac{f_{XY}(x, y)}{f_X(x)}$$

where

- $f_{XY}(x, y)$ is the joint pdf of $X$ and $Y$
- $f_X(x)$ is the marginal pdf of $X$

In the bivariate normal case, we have that

$$Y|X \sim \mathrm{N}(\mu_*, \sigma_*^2)$$

where $\mu_* = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x)$ and $\sigma_*^2 = \sigma_y^2(1 - \rho^2)$

# Statistical Independence

Two variables $X$ and $Y$ are statistically independent if

$$f_{XY}(x, y) = f_X(x) f_Y(y) \qquad (10)$$

where $f_{XY}(x, y)$ is joint pdf, and $f_X(x)$ and $f_Y(y)$ are marginals pdfs.

Note that if $X$ and $Y$ are independent, then

$$f_{Y|X}(y|X = x) = \frac{f_{XY}(x, y)}{f_X(x)} = \frac{\cancel{f_X(x)} f_Y(y)}{\cancel{f_X(x)}} = f_Y(y) \qquad (11)$$

so conditioning on $X = x$ does not change the distribution of $Y$.

# Example

Let $X$ be the height of the father, $Y$ the height of the son, in a sample of father-son pairs. Assume $X$ and $Y$ bivariate normal, as found by Karl Pearson around 1900. Assume $E(X) = 68$ (inches), $E(Y) = 69$, $\sigma_X = \sigma_Y = 2, \rho = .5$. (We expect $\rho$ to be positive because on the average, the taller the father, the taller the son.)

mean ?

variance ?

# Solution

Given $X = 80$ (6 feet 8 inches), $Y$ is normal with mean

$$\mu_y + \frac{\rho \sigma_y}{\sigma_x}(x - \mu_x) = 69 + \frac{0.5 \times 2}{2}(80 - 68)$$

$$= 75$$

which is 6 feet 3 inches. The variance of $Y$ given $X = 80$ is

$$\sigma_y^2 (1 - \rho^2) = (2)^2 (1 - (0.5)^2)$$

$$= 4\left(1 - \frac{1}{4}\right) = 3$$

# The standard multivariate normal distribution

- The adjective "standard" is used to indicate that the mean of the distribution is equal to zero

- Its covariance matrix is equal to the identity matrix.

Standard MV-N random vectors are characterized as follows.

**Definition** Let $X$ be a $K \times 1$ continuous random vector. Let its support be the set of $K$-dimensional real vectors:

$$R_X = \mathbb{R}^K$$

We say that $X$ has a **standard multivariate normal distribution** if its joint probability density function is

transpose

$$f_X(x) = (2\pi)^{-K/2} \exp\left(-\frac{1}{2}x^\top x\right)$$

# Expected value

- The expected value of a standard MV-N random vector X is E[X]=0

**Proof**

- All the components of X are standard normal random variables and a standard normal random variable has mean 0.

# Covariance matrix

- Since the components of X are all standard normal random variables, their variances are all equal to 1, i.e.,

$$\mathrm{Var}[X_1] = \ldots = \mathrm{Var}[X_K] = 1$$

Furthermore, since the components of $X$ are mutually independent and independence implies zero-covariance, all the covariances are equal to 0, i.e.,

$$\mathrm{Cov}[X_i, X_j] = 0 \quad \forall i,j$$

Therefore,

$$\mathrm{Var}[X] = \begin{bmatrix} \mathrm{Var}[X_1] & \mathrm{Cov}[X_1,X_2] & \ldots & \mathrm{Cov}[X_1,X_K] \\ \mathrm{Cov}[X_1,X_2] & \mathrm{Var}[X_2] & \ldots & \mathrm{Cov}[X_2,X_K] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[X_1,X_K] & \mathrm{Cov}[X_2,X_K] & \ldots & \mathrm{Var}[X_K] \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix} = I$$

# The multivariate normal distribution in general

**Definition** Let $X$ be a $K \times 1$ continuous random vector. Let its support be the set of $K$-dimensional real vectors:

$$R_X = \mathbb{R}^K$$

Let $\mu$ be a $K \times 1$ vector and $V$ a $K \times K$ symmetric and positive definite matrix. We say that $X$ has a **multivariate normal distribution** with mean $\mu$ and covariance $V$ if its joint probability density function is

$$f_X(x) = (2\pi)^{-K/2} |\det(V)|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^\intercal V^{-1}(x-\mu)\right)$$

---

We indicate that $X$ has a multivariate normal distribution with mean $\mu$ and covariance $V$ by

$$X \sim N(\mu, V)$$

The $K$ random variables $X_1, \ldots, X_K$ constituting the vector $X$ are said to be **jointly normal**.

# Multivariate Normal probabilities

Probabilities still relate to the area under the pdf:

$$P(a_j \le X_j \le b_j \ \forall j) = \int_{a_1}^{b_1} \cdots \int_{a_p}^{b_p} f(\mathbf{x}) \mathrm{d}x_p \cdots \mathrm{d}x_1 \quad (13)$$

where $\int \cdots \int f(\mathbf{x}) \mathrm{d}x_p \cdots \mathrm{d}x_1$ denotes the multiple integral $f(\mathbf{x})$.

We can still define the cdf of $\mathbf{x} = (x_1, \ldots, x_p)'$

$$\begin{aligned} F(\mathbf{x}) &= P(X_j \le x_j \ \forall j) \\ &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_p} f(\mathbf{u}) \mathrm{d}u_p \cdots \mathrm{d}u_1 \end{aligned} \quad (14)$$

# Multivariate conditional distribution

Given variables $\mathbf{x} = (x_1, \dots, x_p)'$ and $\mathbf{y} = (y_1, \dots, y_q)'$, we have

$$f_{Y|X}(\mathbf{y}|X = \mathbf{x}) = \frac{f_{XY}(\mathbf{x}, \mathbf{y})}{f_X(\mathbf{x})} \tag{15}$$

where

- $f_{Y|X}(\mathbf{y}|X = \mathbf{x})$ is the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$
- $f_{XY}(\mathbf{x}, \mathbf{y})$ is the joint pdf of $\mathbf{x}$ and $\mathbf{y}$
- $f_X(\mathbf{x})$ is the marginal pdf of $\mathbf{x}$

# Conditional Normal Multivariate

Suppose that $\mathbf{z} \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where

- $\mathbf{z} = (\mathbf{x}', \mathbf{y}')' = (x_1, \ldots, x_p, y_1, \ldots, y_q)'$
- $\boldsymbol{\mu} = (\boldsymbol{\mu}_x', \boldsymbol{\mu}_y')' = (\mu_{1x}, \ldots, \mu_{px}, \mu_{1y}, \ldots, \mu_{qy})'$
  Note: $\boldsymbol{\mu}_x$ is mean vector of $\mathbf{x}$, and $\boldsymbol{\mu}_y$ is mean vector of $\mathbf{y}$
- $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{xy}' & \boldsymbol{\Sigma}_{yy} \end{pmatrix}$ where $(\boldsymbol{\Sigma}_{xx})_{p\times p}$, $(\boldsymbol{\Sigma}_{yy})_{q\times q}$, and $(\boldsymbol{\Sigma}_{xy})_{p\times q}$,
  $\Sigma_{yx}$
  Note: $\boldsymbol{\Sigma}_{xx}$ is covariance matrix of $\mathbf{x}$, $\boldsymbol{\Sigma}_{yy}$ is covariance matrix of $\mathbf{y}$, and $\boldsymbol{\Sigma}_{xy}$ is covariance matrix of $\mathbf{x}$ and $\mathbf{y}$

In the multivariate normal case, we have that

$$\mathbf{y}|\mathbf{x} \sim \mathrm{N}(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*) \tag{16}$$

where $\boldsymbol{\mu}_* = \boldsymbol{\mu}_y + \boldsymbol{\Sigma}_{xy}'\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{x} - \boldsymbol{\mu}_x)$ and $\boldsymbol{\Sigma}_* = \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{xy}'\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$

# Statistical Independence

Using Equation (16), we have that

$$\mathbf{y}|\mathbf{x} \sim \mathrm{N}(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*) \equiv \mathrm{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_{yy}) \tag{17}$$

if and only if $\boldsymbol{\Sigma}_{xy} = \mathbf{0}_{p \times q}$ (a matrix of zeros).

Note that $\boldsymbol{\Sigma}_{xy} = \mathbf{0}_{p \times q}$ implies that the $p$ elements of $\mathbf{x}$ are uncorrelated with the $q$ elements of $\mathbf{y}$.

- For multivariate normal variables: uncorrelated $\rightarrow$ independent
- For non-normal variables: uncorrelated $\not\rightarrow$ independent

# Exercise

A statistics class takes two exams $X$ (Exam 1) and $Y$ (Exam 2) where the scores follow a bivariate normal distribution with parameters:

- $\mu_X = 70$ and $\mu_Y = 60$ are the marginal means
- $\sigma_X = 10$ and $\sigma_Y = 15$ are the marginal standard deviations
- $\rho = 0.6$ is the correlation coefficient

Suppose we select a student at random. What is the probability that...

(a) the student scores over 75 on Exam 2?

(b) the student scores over 75 on Exam 2, given that the student scored $X = 80$ on Exam 1?

(c) the sum of his/her Exam 1 and Exam 2 scores is over 150?

(d) the student did better on Exam 1 than Exam 2?

(e) $P(5X - 4Y > 150)$?

# Solution: a

Solution $P(Y > 75) = P\left(Z > \frac{75-60}{15}\right)$

$P(Z > 1)$

$= 1 - P(Z < 1) = 1 - \phi(1)$

$= 1 - 0.8413$

where $\Phi(x) = \int_{-\infty}^{x} f(z)dz$ with $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ denoting the standard normal pdf

# Solution: b

Note that $(Y|X = 80) \sim \mathrm{N}(\mu_*, \sigma_*^2)$ where

$$\mu_* = \mu_y + \rho \frac{\sigma_y}{\sigma_x}(x - \mu_x)$$

$$= 60 + 0.6 \times \frac{15}{10}(80-70) = 69$$

$$\sigma_*^2 = \sigma_y^2(1-\rho^2) = 15^2(1-0.6^2) = 144$$

If a student scored $X = 80$ on Exam 1, the probability that the student scores over 75 on Exam 2 is

$$P(Y > 75 \mid X = 80) = P\left(Z > \frac{75-69}{12}\right)$$

$$= P(Z > 0.5)$$

$$= 1 - P(Z < 0.5)$$

$$= 1 - 0.6914$$

# Solution: c

Note that $(X+Y) \sim \mathrm{N}(\mu_*, \sigma_*^2)$ where

$\mu_* = \mu_X + \mu_Y = 70 + 60 = 130$

$\sigma_*^2 = \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y = 10^2 + 15^2 + 2(0.6)(10)(15) = 505$

The probability that the sum of Exam 1 and Exam 2 is above 150 is

$$\begin{aligned}
P(X+Y > 150) &= P\left(Z > \frac{150-130}{\sqrt{505}}\right) \\
&= P(Z > 0.8899883) \\
&= 1 - \Phi(0.8899883) \\
&= 1 - 0.8132639 \\
&= 0.1867361
\end{aligned}$$

# Solution: d

Note that $(X - Y) \sim \mathrm{N}(\mu_*, \sigma_*^2)$ where

✓ $\mu_* = \mu_X - \mu_Y = 70 - 60 = 10$

✓ $\sigma_*^2 = \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y = 10^2 + 15^2 - 2(0.6)(10)(15) = 145$

The probability that the student did better on Exam 1 than Exam 2 is

$$\begin{aligned} P(X > Y) &= P(X - Y > 0) \\ &= P\left(Z > \frac{0 - 10}{\sqrt{145}}\right) \\ &= P(Z > -0.8304548) \\ &= 1 - \Phi(-0.8304548) \\ &= 1 - 0.2031408 \\ &= 0.7968592 \end{aligned}$$

# Solution: e

Note that $(5X - 4Y) \sim \mathrm{N}(\mu_*, \sigma_*^2)$ where
$\mu_* = 5\mu_X - 4\mu_Y = 5(70) - 4(60) = 110$
$\sigma_*^2 = 5^2\sigma_X^2 + (-4)^2\sigma_Y^2 + 2(5)(-4)\rho\sigma_X\sigma_Y =$
$25(10^2) + 16(15^2) - 2(20)(0.6)(10)(15) = 2500$

Thus, the needed probability can be obtained using

$$
\begin{aligned}
P(5X - 4Y > 150) &= P\left(Z > \frac{150 - 110}{\sqrt{2500}}\right) \\
&= P(Z > 0.8) \\
&= 1 - \Phi(0.8) \\
&= 1 - 0.7881446 \\
&= 0.2118554
\end{aligned}
$$