



# 機器與深度學習常用的數學 相關機率論與統計學

黃志勝 (Tommy Huang)

義隆電子 人工智慧研發部

國立陽明交通大學 AI學院 合聘助理教授

國立台北科技大學 電資學院合聘助理教授





# Outline

1. 機器與深度學習常用的數學 (Basic linear algebra)
2. 相關機率論與統計學 (Basic statistics)



教材： [機器學習的統計基礎：深度學習背後的核心技術](#)



# Why?

- Both are data dependence.

- **Statistical Learning**

1. operates on assumptions.
2. most of idea is from sample, population, and hypothesis.
3. math intensive, and requires a good understanding of data.

- **Machine(Deep) Learning**

1. not as assumptions dependent.
2. emphasizes predictions, supervised learning, unsupervised learning, and semi-supervised learning.
3. identifies patterns from the dataset which require a way less of human effort.





# What is statistical learning and machine learning?

In this course, we don't exploit the difference between them in details.

This course will introduce the common methods, you may use in not only data science, but also AI.

We quickly recall some basic conceptions for learning algorithm in this week.



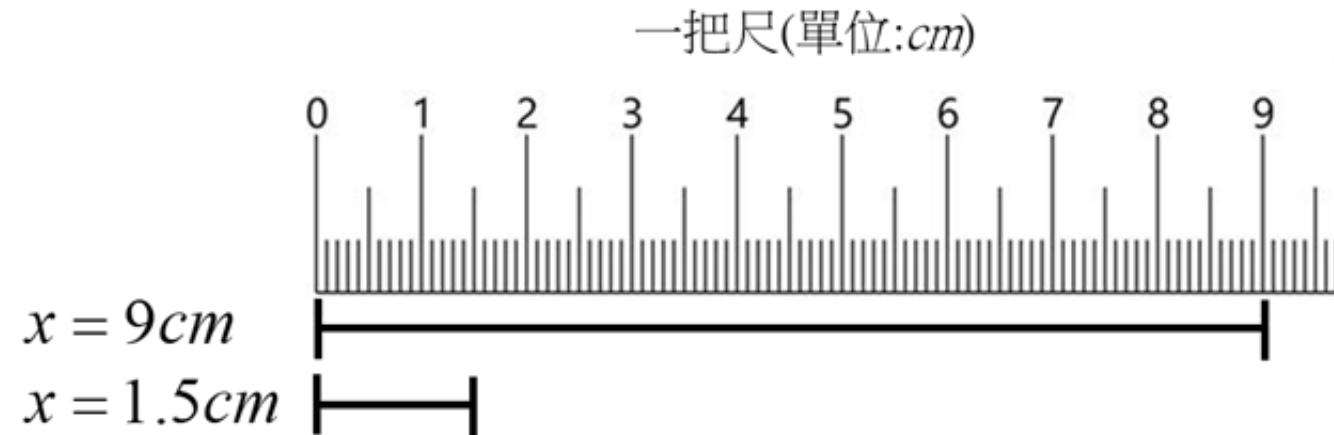
# Basic linear algebra for Learning algorithms

- Scalar
- Vector
- Matrix
- Tensor
- Matrix Computation
- Matrix Transpose
- Inverse matrix



# 純量(Scalar)

- 純量只有大小關係，例如：質量(*mass*)、長度(*length*)、速率(*speed*)，在同一個尺度下可以比較大小。
- 假設有一個純量  $x \in \mathbb{R}$  ( $\mathbb{R}$  代表實數空間)，這個純量  $x$  就只表示一個值。



$$9\text{cm} > 1.5\text{cm}$$

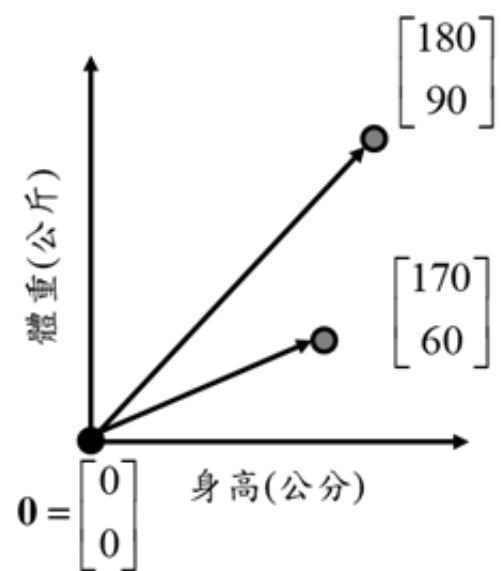


# 向量 (vector)

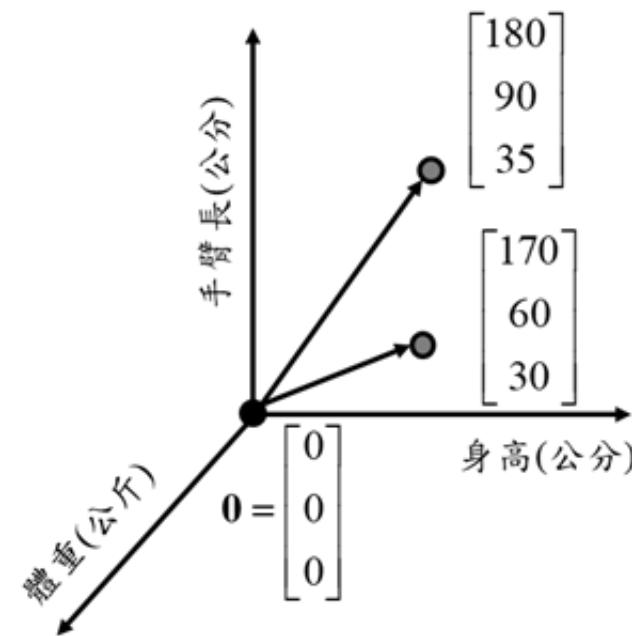
● 假設有一個向量  $\mathbf{x} \in \mathbb{R}^d$  ( $d$  維的實數空間) ,

- 當  $d = 1$  為純量；當  $d = 2$  為 2D 空間；當  $d = 3$  為 3D 空間
- 當  $d > 3$  為則為所謂的高維度空間

2D空間



3D空間



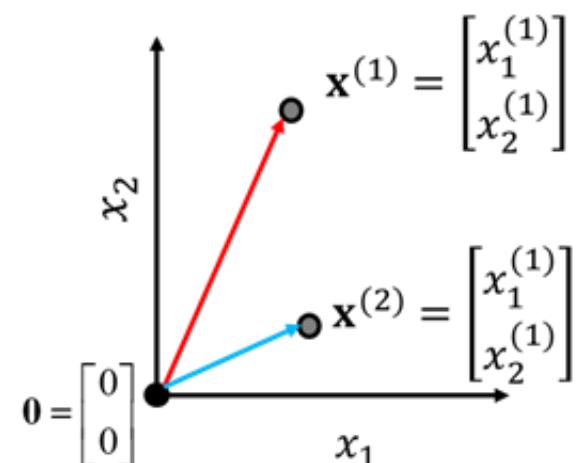
# 向量 (vector)

● 向量具有**大小**及**方向性**。

- 向量的**大小**則是以此點座標到原點的歐幾里得距離(*Euclidean Distance*)來表示，數學寫作 $\|\mathbf{x}\|$ ，當 $\mathbf{x} \in \mathbb{R}^d$

$$\|\mathbf{x}\| = \sqrt{\sum_{d=1}^d x_i^2}$$

- 向量的**方向**性是原點到此點座標的方向，



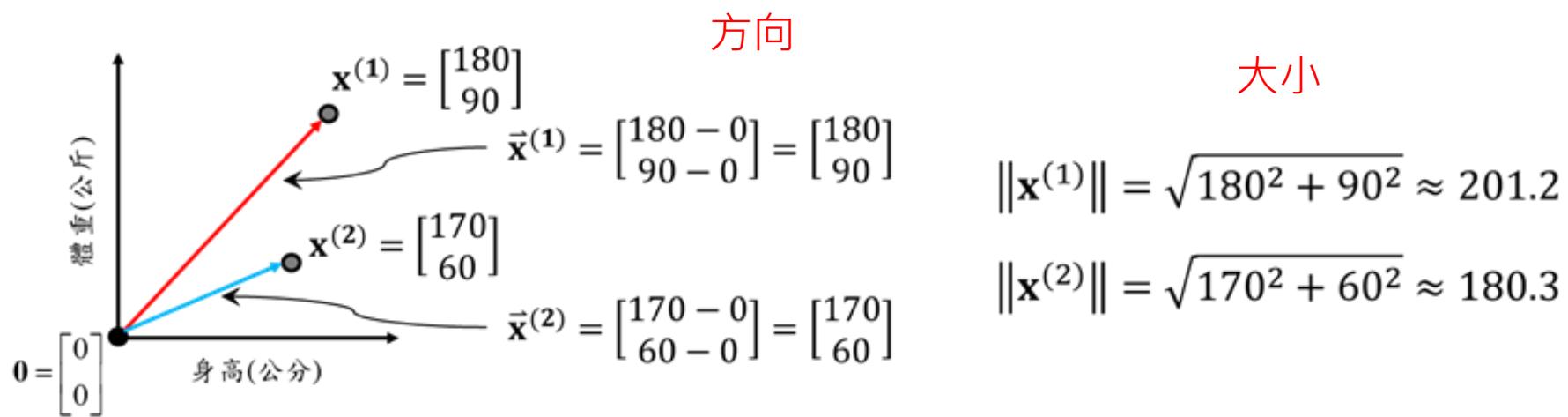
# 向量 (vector)

## 向量範例:

- 假設有兩個特徵值(身高和體重)組成的向量空間

$$\mathbf{x} = \begin{bmatrix} \text{身高} \\ \text{體重} \end{bmatrix} \in \mathbb{R}^d$$

其中有兩個向量分別為  $\mathbf{x}^{(1)} = \begin{bmatrix} \text{身高} = 180 \\ \text{體重} = 90 \end{bmatrix}$ ,  $\mathbf{x}^{(2)} = \begin{bmatrix} \text{身高} = 170 \\ \text{體重} = 60 \end{bmatrix}$



# 矩陣(Matrix)

$$X = [x_1 \quad x_2 \quad \cdots \quad x_m] = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{d1} & x_{d2} & \cdots & x_{dm} \end{bmatrix} \in \mathbb{R}^{d \times m}$$



$$x_1 = \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{d1} \end{bmatrix}, x_2 = \begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{d2} \end{bmatrix}, \dots, x_m = \begin{bmatrix} x_{1m} \\ x_{2m} \\ \vdots \\ x_{dm} \end{bmatrix}$$



# Tensor

## Multidimensional array

0-dimensional tensor: scalar

1-dimensional tensor: vector

0.5
-----

0.1	2	1
-3	1.2	5.1
-0.2	1	0

3x3

2-dimensional tensor: matrix

3-dimensional tensor

0.1
1
-3
-0.5

4x1

0.1	2	1
-----	---	---

1x3

0.1	2	1
-3	1.2	5.1

4x2x3



# 矩陣的秩

● 矩陣的秩(*Rank*)，矩陣 $X$ 的秩記為 $\text{rank}(X)$ 。

- 矩陣的秩指的是  
「行向量的線性獨立性→行秩)」  
「列向量的線性獨立性→列秩)」
- 概念可以用「描述矩陣最大的記憶空間」。
- 「行/列向量的線性獨立性」的個數為這個矩陣行/列的基底向量(*basis vector*)個數，也就是矩陣的秩。



# 矩陣的秩

範例來講述矩陣的秩

$$X = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 5 & 7 & 1 \end{bmatrix}$$

從列進行

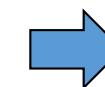
高斯消去法

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 5 & 7 & 1 \end{bmatrix} \begin{array}{l} \text{row}(2) = -2 \times \text{row}(1) + \text{row}(2) \\ \text{row}(3) = -5 \times \text{row}(1) + \text{row}(3) \end{array}$$

$$\Rightarrow \begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & 0 \\ 0 & -3 & -14 \end{bmatrix} \begin{array}{l} \text{row}(3) = \frac{-1}{3} \times \text{row}(3) \end{array}$$

$$\Rightarrow \begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & 0 \\ 0 & 1 & \frac{14}{3} \end{bmatrix}$$

由左邊算來第 1 個非 0 值為 1 叫做帶頭為 1



矩陣  $X$  的列基底向量

$$\begin{bmatrix} 1 & 2 & 3 \\ 5 & 7 & 1 \end{bmatrix}$$

所以矩陣  $X$  的列秩

$$\text{rank}_{\text{row}}(X) = 2$$



# 矩陣的秩

範例來講述矩陣的秩

$$X = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 5 & 7 & 1 \end{bmatrix}$$

從行進行  
高斯消去法

矩陣 $X$ 的行基底向量

$$\begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}, \begin{bmatrix} 2 \\ 4 \\ 7 \end{bmatrix}$$

所以矩陣 $X$ 的行秩  
 $rank_{column}(X) = 2$

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 5 & 7 & 1 \end{bmatrix}$$

$column(2) = -2 \times column(1) + column(2)$   
 $column(3) = -3 \times column(1) + column(3)$

$$\Rightarrow \begin{bmatrix} 1 & 0 & 0 \\ 2 & 0 & 0 \\ 5 & -3 & -14 \end{bmatrix}$$

$column(3) = -\frac{14}{3} \times column(2) + column(3)$

$$\Rightarrow \begin{bmatrix} 1 & 0 & 0 \\ 2 & 0 & 0 \\ 5 & -3 & 0 \end{bmatrix}$$

$column(2) = -\frac{1}{3} \times column(2)$

$$\Rightarrow \begin{bmatrix} 1 & 0 & 0 \\ 2 & 0 & 0 \\ 5 & 1 & 0 \end{bmatrix}$$



# 矩陣的秩

$$X = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 5 & 7 & 1 \end{bmatrix}$$

矩陣 $X$ 的列基底向量

$$\begin{bmatrix} 1 & 2 & 3 \\ 5 & 7 & 1 \end{bmatrix}$$

所以矩陣 $X$ 的列秩

$$\text{rank}_{\text{row}}(X) = 2$$

矩陣 $X$ 的行基底向量

$$\begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}, \begin{bmatrix} 2 \\ 4 \\ 7 \end{bmatrix}$$

所以矩陣 $X$ 的行秩

$$\text{rank}_{\text{column}}(X) = 2$$

- 列秩和行秩的確是相同的。
- 所以我們通常只會記矩陣的秩，不會特別說是列秩或是行秩。
- 當矩陣為  $\mathbf{X} \in \mathbb{R}^{d \times d}$ ,

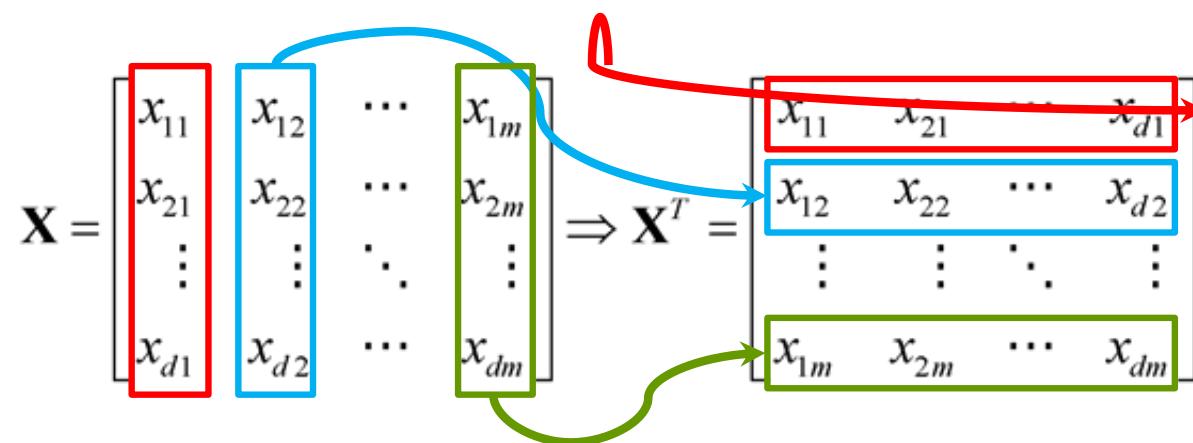
$\text{rank}(\mathbf{X}) = d$  稱為滿矩(full rank)。

則矩陣  $\mathbf{X}$  可逆 → 反矩陣存在。



# 轉置矩陣(Matrix Transpose)

- 假設有一矩陣  $\mathbf{X} \in \mathbb{R}^{d \times m}$ ，則  $\mathbf{X}^T \in \mathbb{R}^{m \times d}$  為  $\mathbf{X}$  的轉置矩陣(Matrix Transpose)，



# 轉置矩陣的特性

設  $\mathbf{X} \in \mathbb{R}^{d \times m}$

- 1.  $(\mathbf{X}^T)^T = \mathbf{X}$
- 2. 當  $d = m$  ,  $\det(\mathbf{X}) = \det(\mathbf{X}^T)$   
 $\det$ 為矩陣的行列式(Determinant)
- 3.  $(\mathbf{X} + \mathbf{Y})^T = \mathbf{X}^T + \mathbf{Y}^T$  ( $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{d \times m}$ )
- 4.  $(\mathbf{XY})^T = \mathbf{Y}^T \mathbf{X}^T$  ( $\mathbf{X} \in \mathbb{R}^{d \times m}, \mathbf{Y} \in \mathbb{R}^{m \times d}$ )



# Matrix Computation



- Scalar multiplication:  $a\mathbf{x} = a \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} = \begin{bmatrix} ax_1 \\ ax_2 \\ \vdots \\ ax_d \end{bmatrix}$
- Vector multiplication:  $\mathbf{xy}$ , but  $(\mathbf{x} \in \mathbb{R}^{d \times 1}, \mathbf{y} \in \mathbb{R}^{1 \times d})$  or  $(\mathbf{x} \in \mathbb{R}^{1 \times d}, \mathbf{y} \in \mathbb{R}^{d \times 1})$
- $(\mathbf{x} \in \mathbb{R}^{d \times 1}, \mathbf{y} \in \mathbb{R}^{1 \times d})$ :  $\mathbf{xy} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} [y_1 \quad y_2 \quad \dots \quad y_d] = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \dots & x_1 y_d \\ x_2 y_1 & x_2 y_2 & \dots & x_2 y_d \\ \vdots & \vdots & \ddots & \vdots \\ x_d y_1 & x_d y_2 & \dots & x_d y_d \end{bmatrix}$
- $(\mathbf{x} \in \mathbb{R}^{1 \times d}, \mathbf{y} \in \mathbb{R}^{d \times 1})$ :  $\mathbf{xy} = [x_1 \quad x_2 \quad \dots \quad x_d] \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_d \end{bmatrix} = [x_1 y_1 + x_2 y_2 + \dots + x_d y_d] = \sum_{i=1}^d x_i y_i$



# Matrix Computation

●  $\mathbf{X} \in \mathbb{R}^{d \times m}, \mathbf{Y} \in \mathbb{R}^{m \times n}$

1.  $\mathbf{X}$ 乘上 $\mathbf{Y}$ ，成立  $\rightarrow \mathbf{XY} \in \mathbb{R}^{m \times n}$

2.  $\mathbf{Y}$ 乘上 $\mathbf{X}$ ，不成立

$$\mathbf{Z}_{d \times n} = \mathbf{X}_{d \times m} \mathbf{Y}_{m \times n} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1n} \\ z_{21} & z_{22} & \cdots & z_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ z_{d1} & z_{d2} & \cdots & z_{dn} \end{bmatrix}_{d \times n} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{d1} & x_{d2} & \cdots & x_{dm} \end{bmatrix}_{d \times m} \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{m1} & y_{m2} & \cdots & y_{mn} \end{bmatrix}_{m \times n}$$

↓

$$z_{22} = [x_{21} \quad x_{22} \quad \cdots \quad x_{2m}] \begin{bmatrix} y_{12} \\ y_{22} \\ \vdots \\ y_{m2} \end{bmatrix} = \sum_{k=1}^m x_{2k} y_{k2}$$



# Matrix Computation

● 手算矩陣相乘範例:

$$\mathbf{X}_{2 \times 3} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}, \mathbf{Y}_{3 \times 2} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}, \mathbf{Z} = \mathbf{XY} = ?$$

答案:

$$\begin{aligned}\mathbf{Z}_{2 \times 2} &= \mathbf{X}_{2 \times 3} \mathbf{Y}_{3 \times 2} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \\ &= \begin{bmatrix} 1 \times 1 + 2 \times 3 + 3 \times 5 & 1 \times 2 + 2 \times 4 + 3 \times 6 \\ 4 \times 1 + 5 \times 3 + 6 \times 5 & 4 \times 2 + 5 \times 4 + 6 \times 6 \end{bmatrix} = \begin{bmatrix} 22 & 28 \\ 49 & 64 \end{bmatrix}\end{aligned}$$



# Hadamard乘積

- Hadamard乘積(Hadamard product)是兩個矩陣點對點之間的元素相乘(通常用 $\odot$ 來表示)，所以兩個矩陣需要一樣大。

$$\mathbf{Z}_{m \times n} = \mathbf{X}_{m \times n} \odot \mathbf{Y}_{m \times n} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \odot \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{m1} & y_{m2} & \cdots & y_{mn} \end{bmatrix} = \begin{bmatrix} x_{11}y_{11} & x_{12}y_{12} & \cdots & x_{1n}y_{1n} \\ x_{21}y_{21} & x_{22}y_{22} & \cdots & x_{2n}y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1}y_{m1} & x_{m2}y_{m2} & \cdots & x_{mn}y_{mn} \end{bmatrix}$$



# Hadamard乘積-範例

$$\mathbf{X}_{2 \times 3} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}, \quad \mathbf{Y}_{2 \times 3} = \begin{bmatrix} 7 & 8 & 9 \\ 10 & 11 & 12 \end{bmatrix}$$

$$\mathbf{Z}_{2 \times 1} = \mathbf{X}_{2 \times 3} \odot \mathbf{Y}_{2 \times 3} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \odot \begin{bmatrix} 7 & 8 & 9 \\ 10 & 11 & 12 \end{bmatrix} = \begin{bmatrix} 1 \times 7 & 2 \times 8 & 3 \times 9 \\ 4 \times 10 & 5 \times 11 & 6 \times 12 \end{bmatrix} = \begin{bmatrix} 7 & 16 & 27 \\ 40 & 55 & 72 \end{bmatrix}$$



# 逆矩陣(反矩陣) Inverse matrix

- 純量在數學上，一個數 $x$ 的倒數(Reciprocal)，也寫作 $\frac{1}{x} = x^{-1}$ ，其此數和倒數的乘積為1

$$x \times \frac{1}{x} = x \times x^{-1} = 1$$

- 矩陣和向量可執行”加”、“減”和”乘”，  
**但不能直接進行”除”的動作**，因此在矩陣上需要引入倒數的概念來進行矩陣的乘的動作。
- 矩陣的倒數稱為逆矩陣(也稱為反矩陣， *Inverse matrix*)。



# 逆矩陣(反矩陣) Inverse matrix

嚴格定義: 矩陣須為方陣才能進行逆矩陣(Inverse matrix)，或稱為反矩陣)的運算。

- $\mathbf{A} \in \mathbb{R}^{d \times d}$

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

$\mathbf{A}^{-1}$ 為 $\mathbf{A}$ 的可逆矩陣(Invertible Matrix)。

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{d \times d}$$

Note:

1. 矩陣 $\mathbf{A}$ 滿秩 $\Leftrightarrow$ 矩陣 $\mathbf{A}$ 可逆
2. 矩陣的行列式(determinant)不等於0則矩陣可逆

$\det(\mathbf{A}) \neq 0$  若且為若矩陣 $\mathbf{A}$ 可逆



# 矩陣分解

- 矩陣分解(Matrix decomposition/factorization)是將矩陣拆解成多個矩陣的矩陣相乘，在非常多領域都很常被用到，最常用在資料壓縮。
- 假設矩陣**A**大小為 $1000 \times 1000$ ，如果可以將矩陣拆成兩個矩陣**B**(大小為 $1000 \times 10$ )和**C**( $10 \times 1000$ )相乘，理想狀態: 無損(Lossless)壓縮還原 $\rightarrow A_{1000 \times 1000} = B_{1000 \times 10} C_{10 \times 1000}$ 。

<b>A</b>	1,000,000
<b>B</b>	10,000
<b>C</b>	10,000

$$A = 1,000,000 > B + C = 10,000 + 10,000 = 20,000$$

$$\frac{1,000,000}{20,000} = 500 \quad \rightarrow \text{壓縮500倍的資料量}$$

此部分我們介紹最常使用的

- 特徵分解(Eigenvalue decomposition)
- 奇異值分解(Singular value decomposition)。



# 特徵分解

- 特徵分解(Eigenvalue decomposition)為統計方法或機器學習最常用的矩陣方式，後續的章節介紹的主成分分析經過假設和推導後就是在做特徵分解。
- 假設矩陣 $A$  為 $n \times n$ 的方陣，若存在一個 $n \times 1$ 的非零向量 $v$ 與純量 $\lambda$ ，可使得

$$Av = \lambda v \text{ 或 } (A - \lambda I)v = \boxed{0}$$

元素階為0的 $n$ 維向量

則 $\lambda$ 為矩陣 $A$ 的特徵值(Eigenvalue)， $v$ 為此特徵值 $\lambda$ 對應的特徵向量(Eigenvector)。



# 特徵分解

- 矩陣 $A$ 為 $n \times n$ 的方陣，求的特徵解的方式為

$$f(\lambda) = \det(A - \lambda I) = 0$$

也就是求矩陣 $A$ 的特徵多項式。

- $\lambda$ 為矩陣 $A$ 的特徵值若且唯若 $\det(A - \lambda I) = 0$ 。



# 矩陣分解-特徵分解

- 特徵值與特徵向量解法

- 範例：

$A = \begin{bmatrix} 2 & 1 \\ 5 & 6 \end{bmatrix}$  求此矩陣的特徵值與特徵值矩陣？

解：

$$f(\lambda) = \det(A - \lambda I) = 0$$

$$\Rightarrow \det \begin{pmatrix} 2-\lambda & 1 \\ 5 & 6-\lambda \end{pmatrix} = 0$$

$$\Rightarrow \lambda^2 - 8\lambda + 7 = 0$$

$$\Rightarrow (\lambda - 7)(\lambda - 1) = 0$$

$$\Rightarrow \lambda = 1, 7$$

當  $\lambda = 1$

$$(A - \lambda I)v = 0$$

$$\Rightarrow \begin{bmatrix} 2-\lambda & 1 \\ 5 & 6-\lambda \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\Rightarrow \begin{cases} v_1 + v_2 = 0 \\ 5v_1 + 5v_2 = 0 \end{cases}$$

$$\Rightarrow v_1 = -v_2$$

$$\Rightarrow v = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

當  $\lambda = 7$

$$(A - \lambda I)v = 0$$

$$\Rightarrow \begin{bmatrix} 2-\lambda & 1 \\ 5 & 6-\lambda \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\Rightarrow \begin{cases} -5v_1 + v_2 = 0 \\ 5v_1 - v_2 = 0 \end{cases}$$

$$\Rightarrow v_2 = 5v_1$$

$$\Rightarrow v = \begin{bmatrix} 1 \\ 5 \end{bmatrix}$$

可得到矩陣A的特徵值  $\lambda = 1$  和  $\lambda = 7$ ，對應的特徵向量分別為  $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$  和  $\begin{bmatrix} 1 \\ 5 \end{bmatrix}$



# 矩陣分解-特徵分解

- 利用特徵值與特徵向量做矩陣分解。

$$\begin{aligned} \mathbf{AV} &= \mathbf{V}\Lambda \\ \Rightarrow \mathbf{A} &= \mathbf{V}\Lambda\mathbf{V}^{-1} \end{aligned}$$

範例:  $\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 5 & 6 \end{bmatrix}$ , 利用特徵值與特徵向量做矩陣分解?

解:

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 7 \end{bmatrix} \quad \mathbf{V} = \begin{bmatrix} \mathbf{v}_{\lambda=1} & \mathbf{v}_{\lambda=7} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -1 & 5 \end{bmatrix}$$

要用兩個 $2 \times 2$ 矩陣( $\Lambda$ 和 $\mathbf{V}$ )來表示一個 $2 \times 2$ 矩陣( $\mathbf{A}$ )。

這邊可能有會有人覺得這樣的方式都沒有壓縮到(省到任何儲存空間)。

原因是此矩陣是滿秩，此方式的特徵拆解並不會節省到任何空間，在非滿秩的狀態則可以節省一些儲存空間。



# 矩陣分解-特徵分解

故意設計非滿秩矩陣

- 範例:  $A = \begin{bmatrix} 2 & 1 \\ 4 & 2 \end{bmatrix}$ , 利用特徵值與特徵向量做矩陣分解?

$$A = V\Lambda V^{-1}$$

$$f(\lambda) = \det(A - \lambda I) = 0$$

$$\Rightarrow \det \begin{pmatrix} 2-\lambda & 1 \\ 4 & 2-\lambda \end{pmatrix} = 0 \quad \Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 4 \end{bmatrix} \rightarrow V = [v_{\lambda=0} \quad v_{\lambda=4}] = \begin{bmatrix} -1 & -1 \\ 2 & 2 \end{bmatrix}$$

$$\Rightarrow \lambda^2 - 4\lambda = 0$$

$$\Rightarrow \lambda = 0, 4$$

$$\lambda = 0 \Rightarrow v = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

$$\lambda = 4 \Rightarrow v = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

這邊看起來還是用兩個 $2 \times 2$ 矩陣( $\Lambda$ 和 $V$ )來表示一個 $2 \times 2$ 矩陣( $A$ )。

\*因為只有一個特徵值，兩個特徵向量是一樣的，所以只有一個特徵值和一個特徵向量3個數字有意義，會比原本4個數字少。

\*現實世界的資料絕大部分都不可能滿秩，所以在採用特徵分解的矩陣分解大多可以節省一些空間。



# 矩陣分解-特徵分解

## ●無損(Lossless)壓縮還原

$$A = V\Lambda V^{-1}$$

- 假設A真的可以分解得很好，但A可能需要100G的儲存空間，我們也可以考慮在特徵分解下，將一些比較小的特徵轉成0，這樣的還原就稱為有損(lossy)壓縮。

$$\widehat{\Lambda} \rightarrow \Lambda$$
$$A \approx V\widehat{\Lambda}V^{-1}$$

在主成分分析的時候會再說明。

非方陣拆解請看教科書奇異值分解(SVD)的內容。





# Outline

1. 機器與深度學習常用的數學 (Basic linear algebra)
2. 相關機率論與統計學 (Basic statistics)



教材： [機器學習的統計基礎：深度學習背後的核心技術](#)



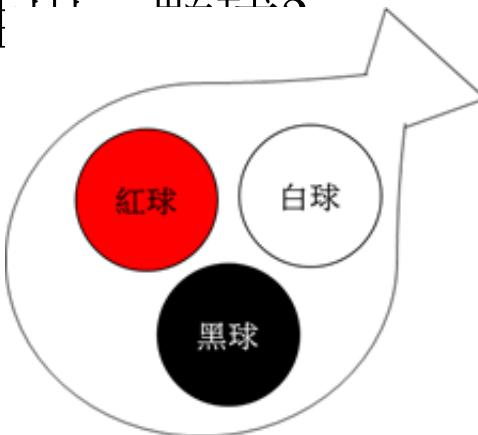
# 機率

- 假設在一個有限的樣本空間(  $\Omega$  )下，對於某個事件發生的機率( $p(S)$ )定義為  
：

$$p(S) = \frac{n(S)}{n(\Omega)}$$

$n(S)$ 和 $n(\Omega)$ 分別代表事件( $S$ )和樣本空間( $\Omega$ )的樣本點個數。

範例：一個公正的袋子裡面放了三顆一樣大小但顏色不同的球，要從這個袋子  
抽



$$p(S_R) = \frac{n(S_R)}{n(\Omega)} = \frac{n(\{R\})}{n(\{B, W, R\})} = \frac{1}{3}$$



# 事件機率三大公理

- 必然事件( $S = \Omega$ ) 的機率依據上述公式可得到

$$p(S=\Omega) = \frac{n(S=\Omega)}{n(\Omega)} = 1$$

- 不可能事件( $\emptyset$ ) 的機率則為：

$$p(S=\emptyset) = \frac{n(S=\emptyset)}{n(\Omega)} = \frac{0}{n(\Omega)} = 0$$

- 因此所有事件的機率範圍為：

$$0 = p(\emptyset) \leq p(S) \leq p(\Omega) = 1$$



# 事件機率三大公理

- **第一公理(非負性質)**：機率必須大於等於0，對任意一個事件 $x$ ，  
 $p(x) \geq 0$
- **第二公理(全機率為1)**： $\Omega$ 為所有可能結果的集合，其總集合機率  
為 1，  $p(\Omega) = 1$ 。

此公理會廣泛被用在機器學習和深度學習中，例如深度學習常用的 *softmax* 函數就是為了達到機率第一和第二公理。

- **第三公理(機率的可加性)**：任意兩兩不相交事件 $(E_1, E_2, \dots)$  滿足

$$p(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} p(E_i)$$



# 範例

●投擲公平骰子兩次，點數和的所有可能分別的機率為何？

投擲公平骰子兩次的樣本空間為

$$\Omega = \left\{ (1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), \right. \\ \left. (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), \right. \\ \left. (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), \right. \\ \left. (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), \right. \\ \left. (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), \right. \\ \left. (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6), \right\} \quad (1, 1, \dots, 6)$$



# 範例

## ●投擲公平骰子兩次，點數和的所有可能分別的機率為何？

點數和的所有可能事件分別為2、3  
 、4、...、12，我們將全部的事件  
 空間依據可能發生的事件，定義為

$$S = \left\{ S_1, S_2, S_3, S_4, S_5, \right. \\ \left. S_6, S_7, S_8, S_9, S_{10}, S_{11} \right\}$$

投擲公正骰子兩次 點數和為 $x$	事件 ( $S_i$ )	$n(S_i)$	$P(S_i) = \frac{n(S_i)}{n(\Omega)}$
$x=2$	$S_1 = \{(1, 1)\}$	1	$\frac{1}{36}$
$x=3$	$S_2 = \{(1, 2), (2, 1)\}$	2	$\frac{2}{36}$
$x=4$	$S_3 = \{(1, 3), (2, 2), (3, 1)\}$	3	$\frac{3}{36}$
$x=5$	$S_4 = \{(1, 4), (2, 3), (3, 2), (4, 1)\}$	4	$\frac{4}{36}$
$x=6$	$S_5 = \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}$	5	$\frac{5}{36}$
$x=7$	$S_6 = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$	6	$\frac{6}{36}$
$x=8$	$S_7 = \{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\}$	5	$\frac{5}{36}$
$x=9$	$S_8 = \{(3, 6), (4, 5), (5, 4), (6, 3)\}$	4	$\frac{4}{36}$
$x=10$	$S_9 = \{(4, 6), (5, 5), (6, 4)\}$	3	$\frac{3}{36}$
$x=11$	$S_{10} = \{(5, 6), (6, 5)\}$	2	$\frac{2}{36}$
$x=12$	$S_{11} = \{(6, 6)\}$	1	$\frac{1}{36}$



# 範例

●投擲公平骰子兩次，點數和的所有可能分別的機率為何？

所以「點數和的所有可能的事件」的機率和就是全機率

$$\frac{1}{36} + \frac{2}{36} + \frac{3}{36} + \frac{4}{36} + \frac{5}{36} + \frac{6}{36} + \frac{5}{36} + \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36} = 1$$

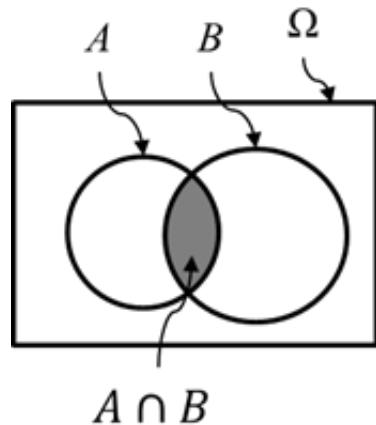
●從事件的角度來看全機率

$$\begin{aligned} & p(S_1(\square \square - 2)) + p(S_2(\square \square - 3)) + \dots + p(S_{11}(\square \square - 12)) \\ &= \frac{n(S_1)}{n(\Omega)} + \frac{n(S_2)}{n(\Omega)} + \dots + \frac{n(S_{12})}{n(\Omega)} = \frac{\sum_{i=1}^{11} n(S_i)}{n(\Omega)} \\ &= \frac{n(\{(1, 1)\}) + n(\{(1, 2), (2, 1)\}) + \dots + n(\{(6, 6)\})}{n(\Omega)} = \frac{n(\Omega)}{n(\Omega)} = 1 \end{aligned}$$



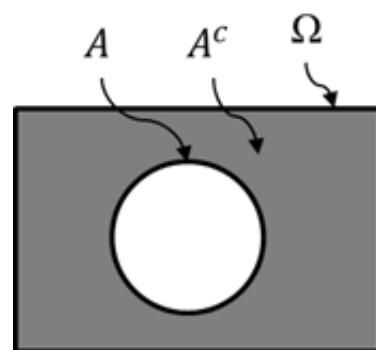
# 機率-運算規則

- 若事件 $A$  和事件 $B$  都包含於樣本空間內( $A, B \subset \Omega$ )

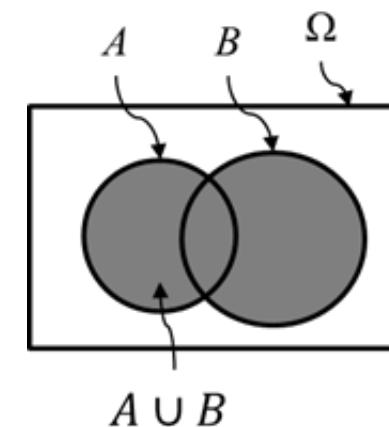


同時發生事件 $A$ 和事件 $B$ 的事件稱為交事件(*intersection event*，或稱為積事件)  
記作  $A \cap B$ 。

交事件的機率( $p(A \cap B)$ )則稱為事件 $A$ 和事件 $B$ 的聯合機率(*joint probability*)。



事件 $A$ 和事件 $B$ 的事件的聯集稱為和事件(*sum event*)，即事件 $A$ 和事件 $B$ 至少有一事件會發生的事件，數學記作 $A \cup B$



事件 $A \subset \Omega$ ，不在事件 $A$  中的事件稱為餘事件(*complement event*)，通常用 $A^c$ 表示，事件 $A$ 與 $A^c$ 互斥

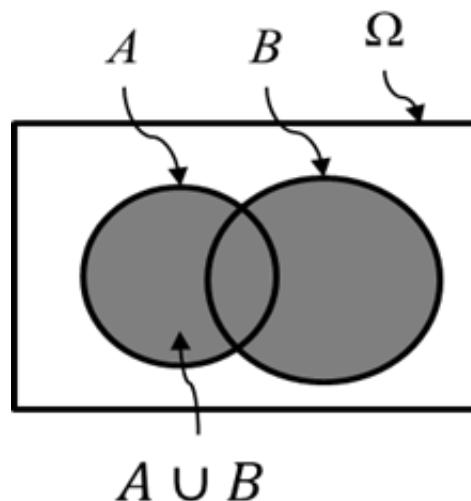


# 機率-運算規則

概率加法法則 (**addition rule**) :

若事件 $A$ 和事件 $B$ 都包含於樣本空間內( $A, B \subset \Omega$ )，則：

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$



# 機率-運算規則-範例

範例：假設有一公平六面骰子如下圖：



請問投擲這個骰子一次，發生點數小於4或點數為紅色的機率為何？

投擲這個骰子一次： $\Omega = \{1, 2, 3, 4, 5, 6\}$

事件A為投出點數小於4： $S_A = \{1, 2, 3\}$       事件B為出現點數為紅色： $S_B = \{1, 4\}$

問題轉換成數學寫法為 $p(S_A \cup S_B)$

$$S_A \cup S_B = \{1, 2, 3, 4\}, \quad S_A \cap S_B = \{1\}$$

解法1：

$$p(S_A \cup S_B) = \frac{n(S_A \cup S_B)}{n(\Omega)} = \frac{4}{6} = \frac{2}{3}$$

解法2：

$$\begin{aligned} p(S_A \cup S_B) &= p(S_A) + p(S_B) - p(S_A \cap S_B) \\ &= \frac{n(S_A)}{n(\Omega)} + \frac{n(S_B)}{n(\Omega)} - \frac{n(S_A \cap S_B)}{n(\Omega)} = \frac{3}{6} + \frac{2}{6} - \frac{1}{6} = \frac{2}{3} \end{aligned}$$



# 條件機率與貝氏定理

● 條件機率(conditional probability)為貝氏機率(Bayesian probability)的基礎，且貝氏機率為機器學習的技術發展基礎，也被廣泛使用在強化式學習(Reinforcement Learning)上。

- 條件機率定義：

事件 $x$ 和事件 $y$ 為樣本空間的兩個事件，且 $y \neq \emptyset \Rightarrow p(y) \neq 0$ ，則給定事件 $y$ 發生的條件下，發生事件 $x$ 的機率稱為條件機率，數學定義為：

$$p(x|y) = \frac{p(x \cap y)}{p(y)} = \frac{p(x, y)}{p(y)}$$



# 條件機率-範例

●某大學經由課程調查得知，70%學生從來不翹機率課，從調查資料發現90%不翹機率課的學生會通過機率課程考試，而只有40%有翹過機率課的學生會通過機率課程考試。

Q: 「那如果有個學生通過機率課，他不翹課的機率為何？」。



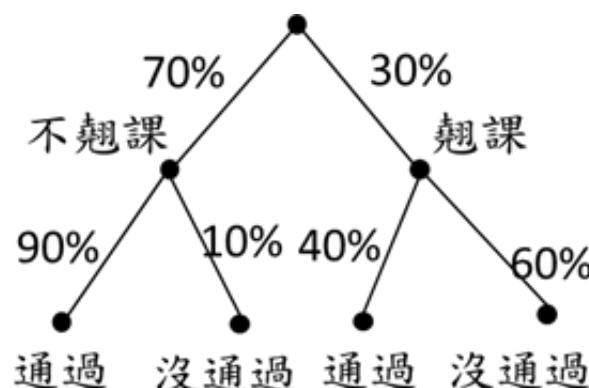
# 條件機率-範例

- 某大學經由課程調查得知，70%學生從來不翹機率課，從調查資料發現90%不翹機率課的學生會通過機率課程考試，而只有40%有翹過機率課的學生會通過機率課程考試。

Q: 「那如果有個學生通過機率課，他不翹課的機率為何？」。

我們將題目拆開來觀察：

1. 70%學生從來不翹機率課→30%學生有翹課過。
2. 不翹機率課的學生有90%會通過機率課程考試→不翹機率課的學生有10%沒有通過考試。
3. 有翹過機率課的學生有40%會通過機率課程考試→有翹過機率課的學生有60%沒有通過考試。



# 條件機率-範例

●某大學經由課程調查得知，70%學生從來不翹機率課，從調查資料發現90%不翹機率課的學生會通過機率課程考試，而只有40%有翹過機率課的學生會通過機率課程考試。

Q: 「那如果有個學生通過機率課，他不翹課的機率為何？」。

「學生通過機率課，他不翹課的機率」

換成

條件機率的說法(「在給定通過機率課( $y$ )下，他不翹課( $x$ )的機率」)

$$p(x = \text{不翹課}|y = \text{通過})$$

這時候的事件有： $x = \{\text{翹課}, \text{不翹課}\}$ ， $y = \{\text{通過}, \text{沒通過}\}$ 。

$$p(x = \text{不翹課}|y = \text{通過}) = \frac{p(x = \text{不翹課}, y = \text{通過})}{p(y = \text{通過})}$$

我們要求的是 $p(x = \text{不翹課}, y = \text{通過})$ 和 $p(y = \text{通過})$



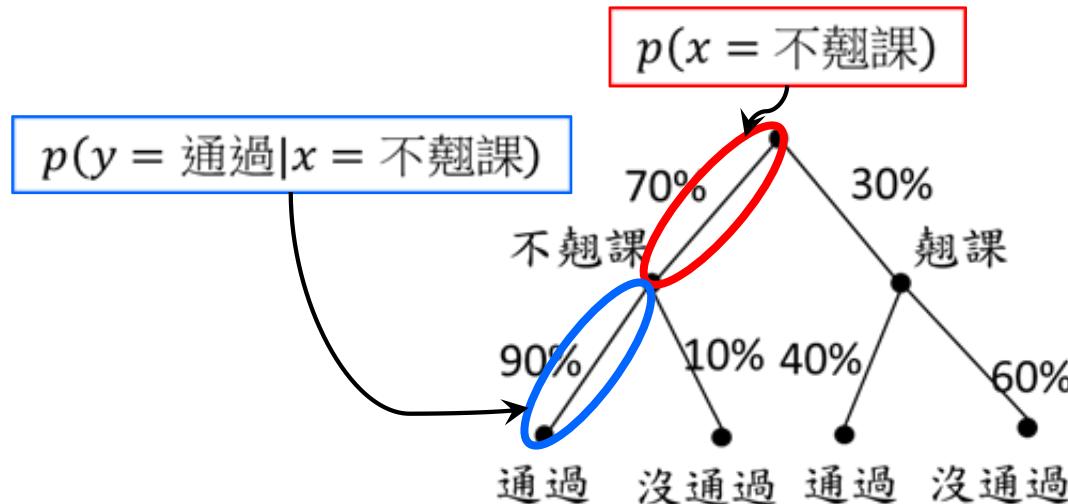
# 條件機率-範例

- 某大學經由課程調查得知，70%學生從來不翹機率課，從調查資料發現90%不翹機率課的學生會通過機率課程考試，而只有40%有翹過機率課的學生會通過機率課程考試。

Q: 「那如果有個學生通過機率課，他不翹課的機率為何？」。

我們要求的是1.  $p(x = \text{不翹課}, y = \text{通過})$ 和2.  $p(y = \text{通過})$

$$1. p(x = \text{不翹課}, y = \text{通過}) = p(x = \text{不翹課})p(y = \text{通過}|x = \text{不翹課}) = 0.7 \times 0.9 = 0.63$$



# 條件機率-範例

- 某大學經由課程調查得知，70%學生從來不翹機率課，從調查資料發現90%不翹機率課的學生會通過機率課程考試，而只有40%有翹過機率課的學生會通過機率課程考試。

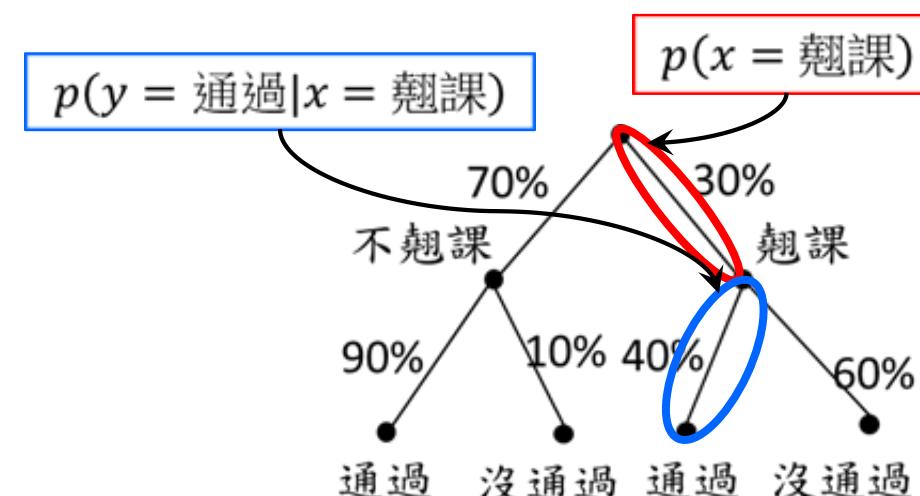
Q: 「那如果有個學生通過機率課，他不翹課的機率為何？」。

我們要求的是1.  $p(x = \text{不翹課}, y = \text{通過})$ 和2.  $p(y = \text{通過})$

2.  $p(y = \text{通過})$

$$= p(x = \text{不翹課})p(y = \text{通過}|x = \text{不翹課}) + p(x = \text{翹課})p(y = \text{通過}|x = \text{翹課})$$

$$= 0.7 \times 0.9 + 0.3 \times 0.4 = 0.75$$



# 條件機率-範例

- 某大學經由課程調查得知，70%學生從來不翹機率課，從調查資料發現90%不翹機率課的學生會通過機率課程考試，而只有40%有翹過機率課的學生會通過機率課程考試。

Q: 「那如果有個學生通過機率課，他不翹課的機率為何？」。

$$p(x = \text{不翹課}|y = \text{通過}) = \frac{p(x = \text{不翹課}, y = \text{通過})}{p(y = \text{通過})} = \frac{0.63}{0.75} = 84\%$$

有個學生通過機率課，他不翹課的機率為84%



# 貝氏定理

- 貝氏定理是樣本空間的任意兩事件 $x$ 和 $y$ 的條件機率。若 $p(x) > 0, p(y) > 0$ ，則

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

看起來跟條件機率一樣，但貝氏定理會將公式衍生為

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



# 貝氏定理

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

- 在貝氏定理上，「給定在事件 $y$ 下，事件 $x$ 發生的機率」 $p(x|y)$ 稱為事件 $x$ 的**後驗機率**  
*(posterior probability)*，或稱事後機率。

「事件 $x$ 發生的機率」 $p(x)$ 稱為事件 $x$ 的**先驗機率***(prior probability)*，也稱為事前機率或邊際機率*(marginal probability)*；

「事件 $y$ 發生的機率」 $p(y)$ 稱為事件 $y$ 的先驗機率或邊際機率。





# 範例：用貝氏定理算出是否得到流感的條件機率

假設從資料調查發現目前每 10 萬人有 1 萬人得過流感，也就是隨機選一個人且他得流感的機率為 0.1 (稱為盛行率 prevalence rate)。

對健康的人，流感快篩檢查正確率為 0.99

(對健康的人來說，正確率就是篩檢結果為沒得流感)。

對得流感的人，流感快篩檢查正確率為 0.95

(對得流感的人而言，正確率就是篩檢的結果為得流感)。

如果要做流感快篩檢查，我們要問以下三個問題：

問題1: 檢查結果為沒有得流感的機率為何?

問題2: 檢查結果為有得流感，但實際上沒流感的機率為何?

問題3: 檢查結果為有得流感，但實際上得流感的機率為何?



# 範例：用貝氏定理算出是否得到流感的條件機率

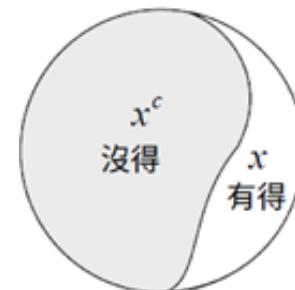
假設事件 $x$ 表示為實際得流感， $x^c$ 為實際沒得流感事件

得流感與否的樣本空間為  $\Omega_x = x^c \cup x$

$$p(x^c \cup x) = p(x^c) + p(x) = p(\Omega_x) = 1$$

由題目所述知道目前流感的盛行率  $p(x) = 0.1$

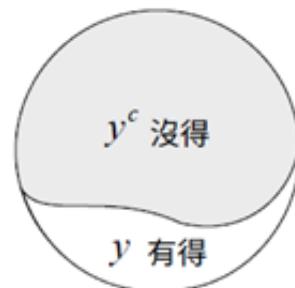
所以  $p(x^c) = 1 - p(x) = 1 - 0.1 = 0.9$



假設事件 $y$ 表示快篩檢查為得到流感， $y^c$ 為快篩檢查為沒有得流感

快篩檢查的樣本空間為  $\Omega_y = y^c \cup y$

$$p(y^c \cup y) = p(y^c) + p(y) = p(\Omega_y) = 1$$



# 範例：用貝氏定理算出是否得到流感的條件機率

●「對健康的人，在流感快篩檢查下正確檢出沒得流感的機率為0.99」，代表在沒得流感( $x^c$ )的條件下，快篩檢查沒得流感( $y^c$ )的機率為

$$p(y^c|x^c) = 0.99$$

- 因此「對健康的人，在流感快篩檢查下不正確的機率」，也就是代表沒得流感( $x^c$ )下，但快篩檢查為得流感( $y$ )的機率為

$$p(x^c) = p(x^c, y) + p(x^c, y^c)$$

$$p(y|x^c) = \frac{p(y, x^c)}{p(x^c)} = \frac{p(x^c) - p(x^c, y^c)}{p(x^c)} = 1 - p(y^c|x^c) = 0.01$$

- 「對得流感的人，在流感快篩檢查正確檢出的機率為0.95」，代表在得流感( $x$ )的條件下，且快篩檢查得流感( $y$ )的機率為

$$p(y|x) = 0.95$$

- 因此「對得流感的人，在流感快篩檢查不正確檢查的機率」，也就是代表在得流感( $y$ )下，但快篩檢查為沒流感( $x^c$ )，機率為

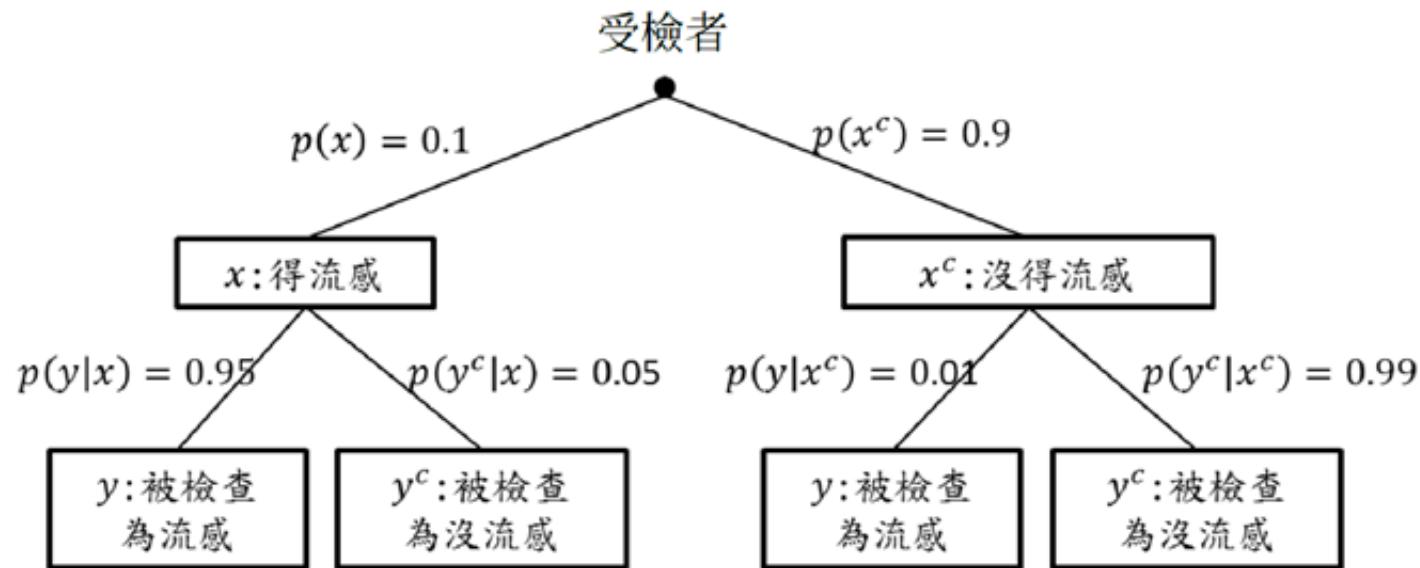
$$p(x) = p(x, y) + p(x, y^c)$$

$$p(y^c|x) = \frac{p(x, y^c)}{p(x)} = \frac{p(x) - p(x, y)}{p(x)} = 1 - p(y|x) = 0.05$$



# 範例：用貝氏定理算出是否得到流感的條件機率

- 「對健康的人，在流感快篩檢查下正確檢出沒得流感的機率為0.99」
- 「對健康的人，在流感快篩檢查下不正確的機率為0.01」
- 「對得流感的人，在流感快篩檢查正確檢出的機率為0.95」
- 「對得流感的人，在流感快篩檢查不正確檢查的機率為0.05」

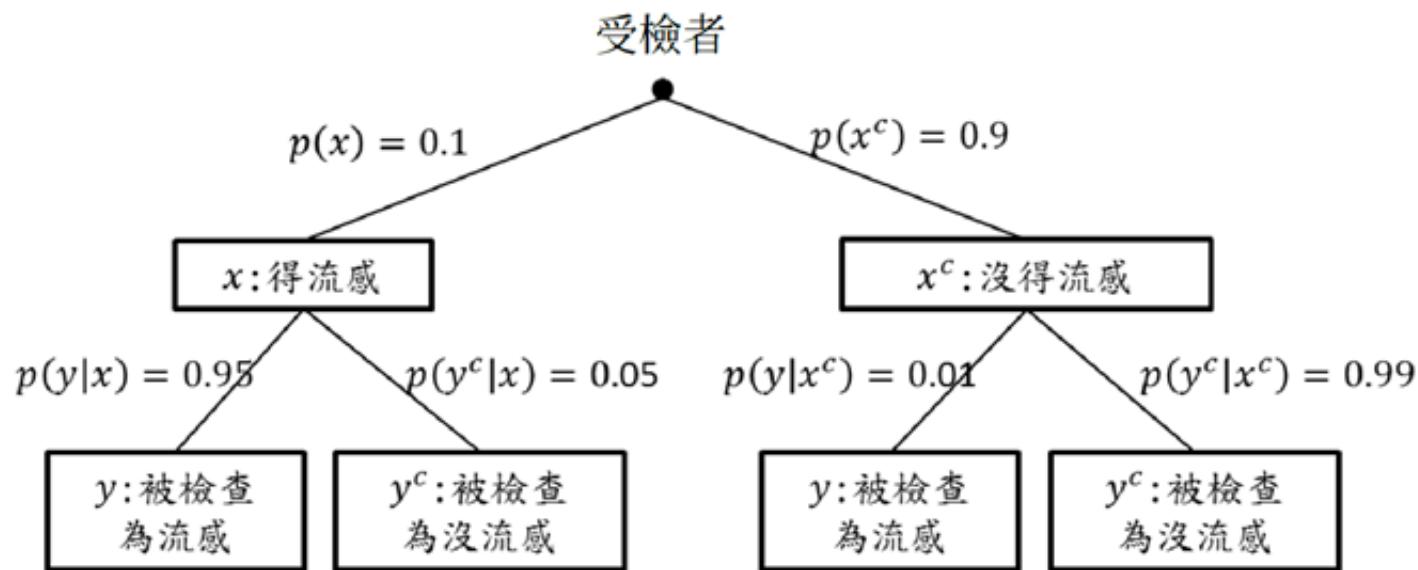


# 範例：用貝氏定理算出是否得到流感的條件機率

## 問題1：檢查結果為沒有得流感的機率為何？

檢查結果為沒有得流感( $y^c$ )的機率( $p(y^c)$ )為

$$\begin{aligned}
 p(y^c) &= p(x, y^c) + p(x^c, y^c) = p(x)p(y^c|x) + p(x^c)p(y^c|x^c) = 0.1 \times 0.05 + 0.9 \times 0.99 \\
 &= 0.896
 \end{aligned}$$

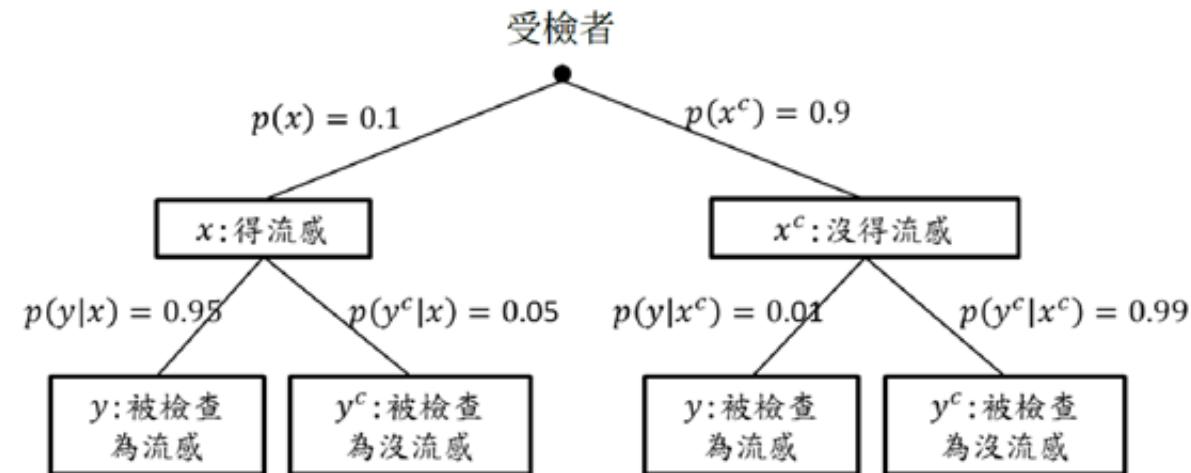


# 範例：用貝氏定理算出是否得到流感的條件機率

問題2：檢查結果為有得流感，但實際上沒流感的機率為何？

檢查結果為有得流感( $y$ )，但實際上沒流感( $x^c$ )的機率的機率( $p(x^c|y)$ )為

$$p(x^c|y) = \frac{p(x^c)p(y|x^c)}{p(y)} = \frac{p(x^c)p(y|x^c)}{p(x)p(y|x) + p(x^c)p(y|x^c)} = \frac{0.9 \times 0.01}{0.1 \times 0.95 + 0.9 \times 0.01} = 0.0865$$



# 範例：用貝氏定理算出是否得到流感的條件機率

## 題3：檢查結果為有得流感，但實際上得流感的機率為何？

檢查結果為有得流感( $y$ )，但實際上得流感( $x$ )的機率的機率( $p(x|y)$ )為

$$p(x|y) = \frac{p(x)p(y|x)}{p(y)} = \frac{p(x)p(y|x)}{p(x)p(y|x) + p(x^c)p(y|x^c)} = \frac{0.1 \times 0.95}{0.1 \times 0.95 + 0.9 \times 0.01} = 0.9135$$

- 上面的問題(2) 與(3) 都是基於「先檢查出得到流感」之後才去推估判斷實際上有沒有得流感的機率。
- $p(x|y)$ 稱為事件 $x$ 的後驗機率(事後機率)。
- 此例中的  $p(x|y)$  為實際上有得流感( $x$ )的後驗機率
- $p(x^c|y)$ 為實際上沒得流感( $x^c$ )的後驗機率。



# 貝氏定理-統計獨立

- 統計獨立(statistically independent)也稱為事件獨立，事件獨立是指這件事的發生不會影響到另外一件事的發生，比如說事件「機車闖紅燈」和「飛機成功飛上天」這兩件事情是完全不相關的，也就是兩個事件發生的機率是獨立的。
- 統計獨立定義：

兩個事件 $A$ 和 $B$ 是統計獨立

若且唯若(if and only if)

$$p(A, B) = p(A)p(B)$$



# 貝氏定理-統計獨立

● 條件機率與統計獨立，假設A和B統計獨立：

$$\cdot p(A|B) = \frac{p(A,B)}{p(B)} = \frac{p(A)p(B)}{p(B)} = p(A)$$



# 貝氏定理-統計獨立：範例

- 我們擲一個公正六面骰子兩次，事件 $x$ 為第一次擲出點數2，事件 $y$ 為第二次擲出點數3，如果我們沒有作弊，可以認為連續兩次擲骰子得到的點數結果是相互獨立。

事件 $x$ 為第一次擲出點數2

$$S_x = \{(2,1), (2,2), (2,3), (2,4), (2,5), (2,6)\}, \quad p(x) = \frac{6}{36} = \frac{1}{6}$$

事件 $y$ 第二次擲出點數3

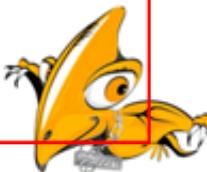
$$S_y = \{(1,3), (2,3), (3,3), (4,3), (5,3), (6,3)\}, \quad p(y) = \frac{6}{36} = \frac{1}{6}$$

事件 $x$ 和事件 $y$ 同時發生的可能結果為：

$$S_{x \cap y} = \{(2,3)\}$$

事件 $x$ 和事件 $y$ 同時發生的機率為  $p(x,y) = \frac{1}{36}$

因為  $p(x,y) = p(x)p(y)$ ，所以事件 $x$ 和事件 $y$ 是事件獨立或稱統計獨立。

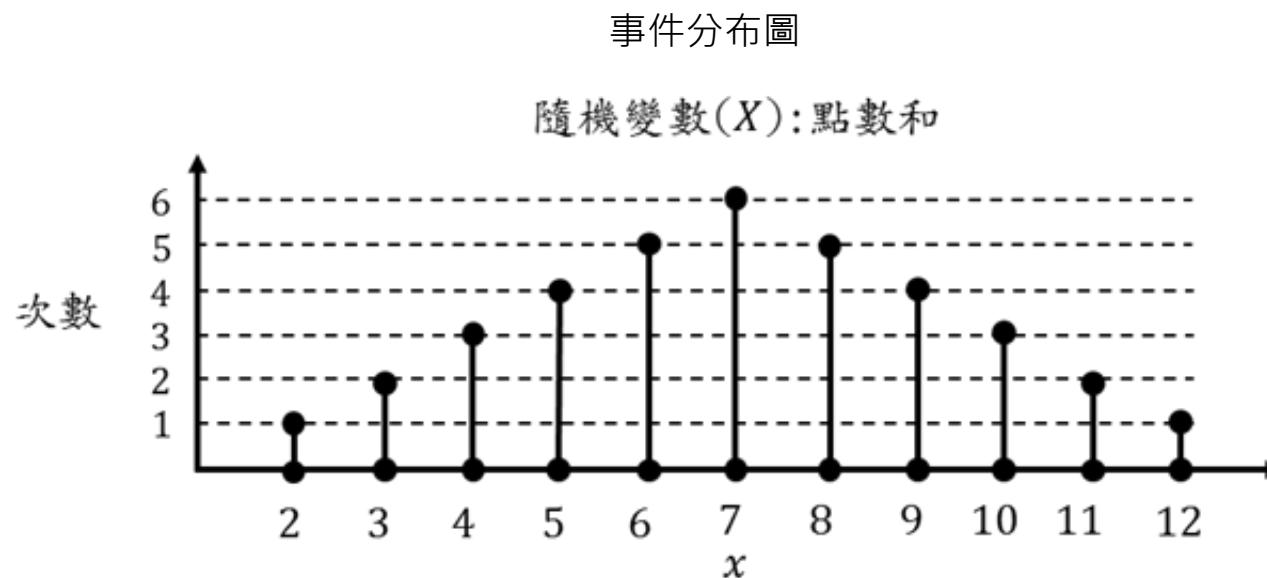


# 機率分布

- 投擲公正骰子兩次，假設我們今天要觀察的隨機試驗是這兩顆骰子的點數和，用隨機變數( $X$ ) 表示，其值 $x$ 有下面幾種可能：

$$x = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12$$

隨機變數 $X = x$ 的發生次數，事件分布圖



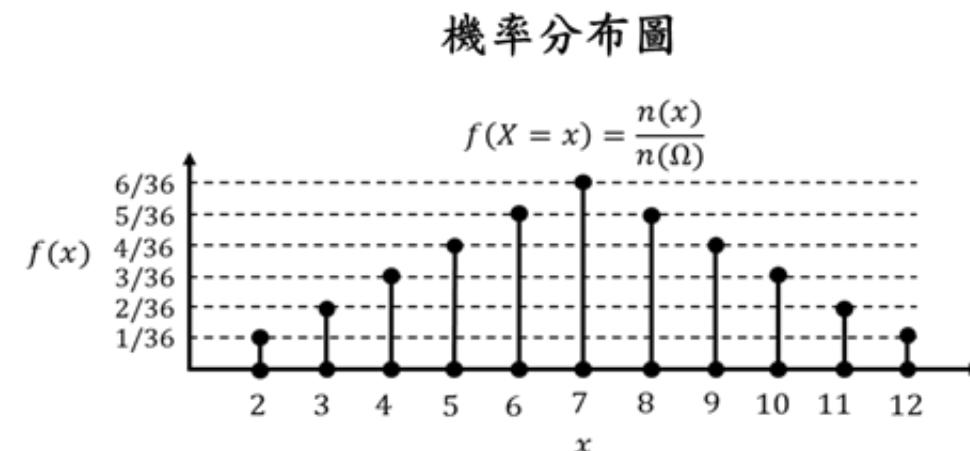
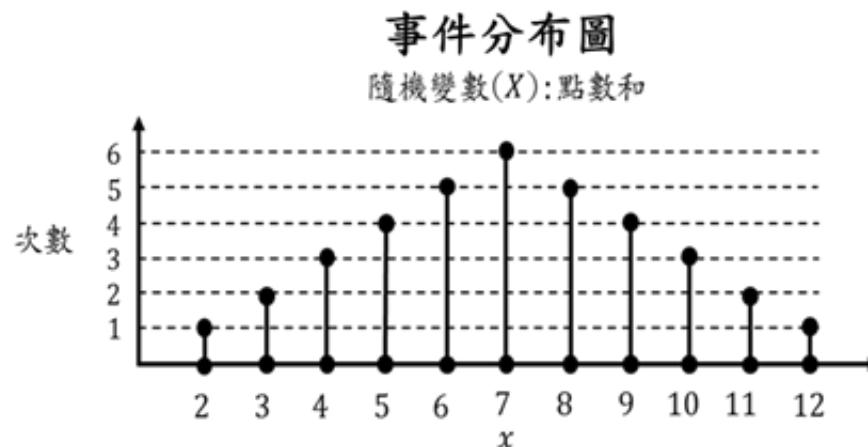
# 機率分布

隨機變數X的每個事件的機率函數可寫為

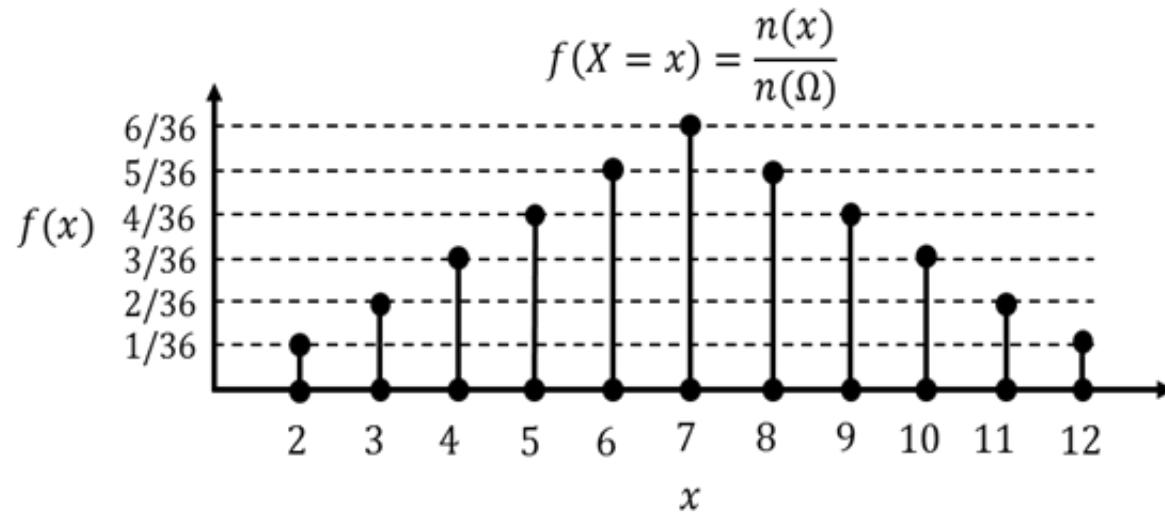
$$\text{機率計算 recall: } p(S) = \frac{n(S)}{n(\Omega)}$$

$$f(x) = \frac{n(x)}{n(\Omega)}$$

- 因此不同事件發生的機率可以用各事件發生次數除以全部樣本數，即可得到下面的機率分布圖



# 機率分布



● 隨機變數的函數輸出值服從機率定理，

$$f(\Omega) = \sum_{x=2}^{12} f(X = x)$$

✓  $f(x) \geq 0$ ，每個輸出值必大於或等於 0。

$$= \frac{1}{36} + \frac{2}{36} + \frac{3}{36} + \frac{4}{36} + \frac{5}{36} + \frac{6}{36}$$

✓ 機率加總需等於 1。

$$+ \frac{5}{36} + \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36}$$

✓ 不相交(互斥) 事件的機率可相加

$$= 1$$



# 機率分布-樣本空間與母體

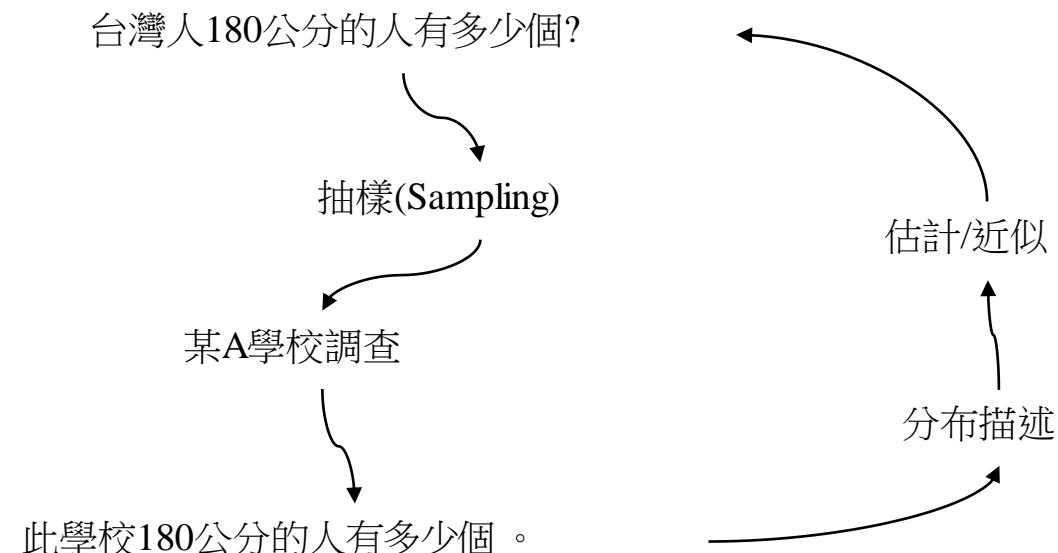
- 樣本空間為機率論的用語，在統計用語上稱作母體。  
至此介紹的範例都可以完整描述所有樣本空間(母體)。
- 例如：台灣人身高的分布，如果找一人出來，他身高為180公分的機率為何？
  - (1) 全台灣人個數→台灣戶政資料找的到。
  - (2) 找出樣本空間(母體)=180公分的人有多少個。

幾乎是不可能將數據都收集齊全的。



# 分布怎麼來的

範例: 台灣人身高的分布，如果找一人出來，他身高為180公分的機率為何？



Q:某A學校是否足以代表全台灣人?

相關內容關鍵字為統計抽樣調查

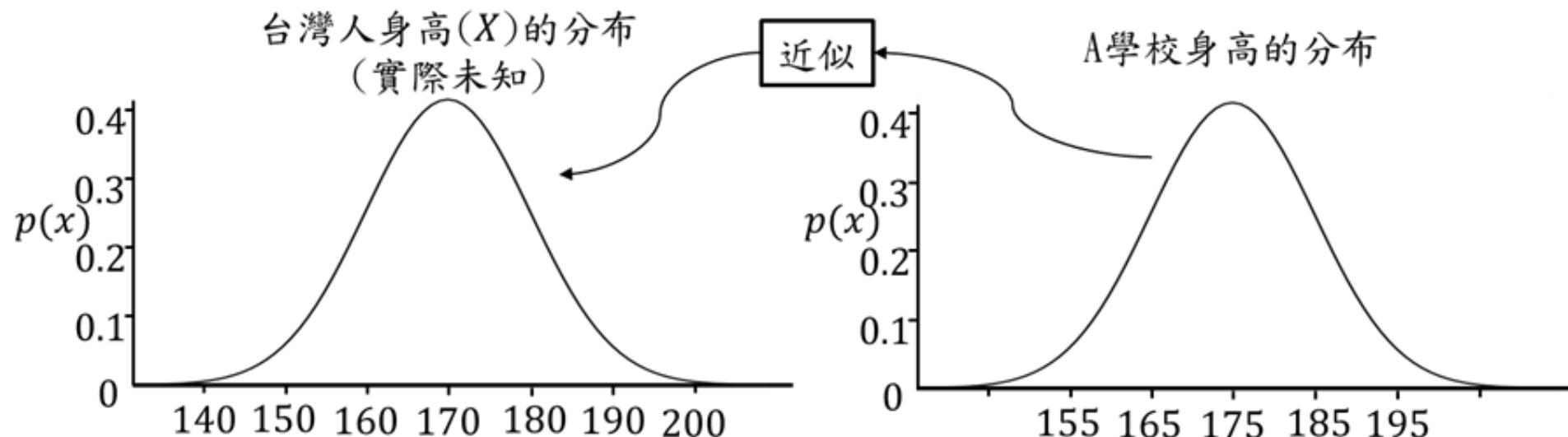


# 機率分布-樣本空間與母體

- 台灣人身高( $X$ )的分布，如果找一人出來，他身高為180公分的機率為何？

某A學校調查此學校180公分的人有300人，全校有1000人，則隨機找一人出來此人180公分的機率為

$$p(X = 180) = \frac{300}{1000} \times 100\% = 30\%$$



# 統計機率分布模型-均勻分布

- 均勻分布為最廣泛使用且最為簡單的一種分布，其意義為隨機變數中每一個事件出現的機率是相等的。
- 前面內容一直採用的「投擲一個公平骰子一次」即是一種均勻分布的型態。
- 假設有一隨機變數服從均勻分布 $X \sim U[a, b]$ ， $a$ 和 $b$ 代表均勻分布的下限和上限，其機率密度函數

$$f(x) = \begin{cases} \frac{1}{b-a+1} & x = a, a+1, \dots, b \\ 0 & O.W. \end{cases}$$

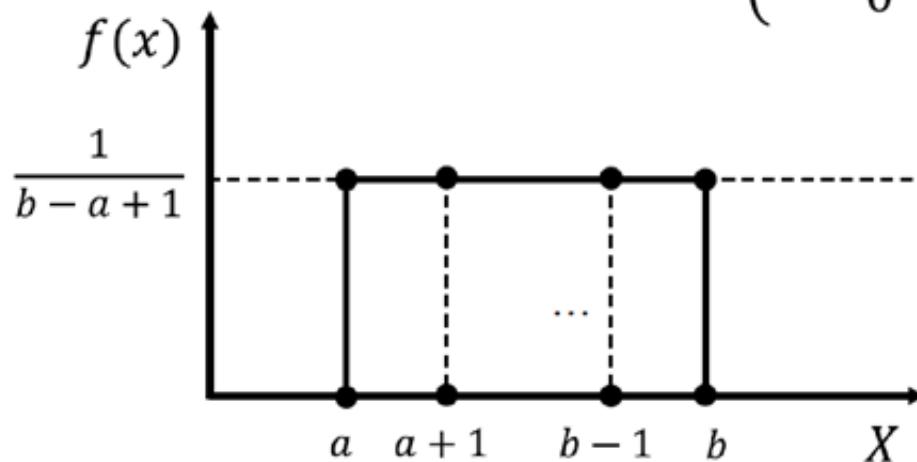


圖 2.16 均勻分布機率密度函數

一般在寫程式時會用到亂數產生器產生0~1之間的數字，如果沒有特別指定，基本上就是在均勻分布 $U[0, 1]$ 中去隨機產生亂數出來。



# 統計機率分布模型-均勻分布

- 隨機變數 $X$ 為投擲一個公正骰子一次出現之點數  
(均勻分布 $X \sim U[1, 6]$ )，則投擲一個公正骰子一次的機率分配為：

$$f(x) = \begin{cases} \frac{1}{6} & x = 1, 2, 3, 4, 5, 6 \\ 0 & O.W. \end{cases}$$

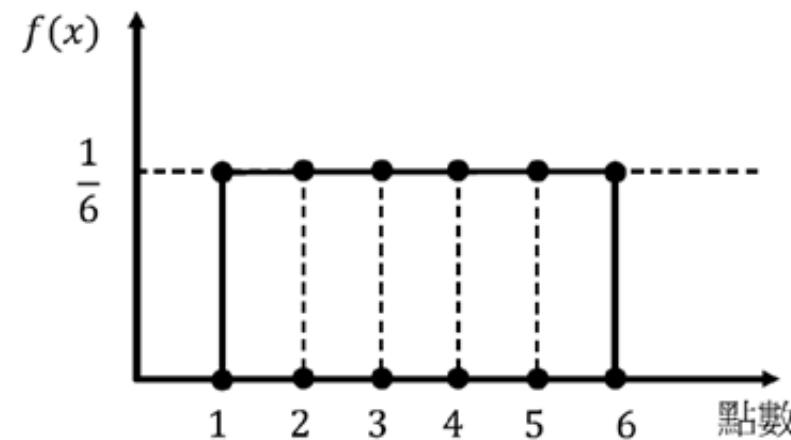


圖 2.17 投擲一個公平骰子的均勻分布機率密度函數



# 統計機率分布模型-常態分布

- 常態分布的機率密度函數平均數為 $\mu$ 、變異數為 $\sigma^2$  ( $\sigma$ 稱為標準差)，常態分布又稱為高斯分布(*Gaussian distribution*)。

假設有一隨機變數服從平均數 $\mu$ 、變異數為 $\sigma^2$ 的常態分布

$$X \sim N(\mu, \sigma^2)$$

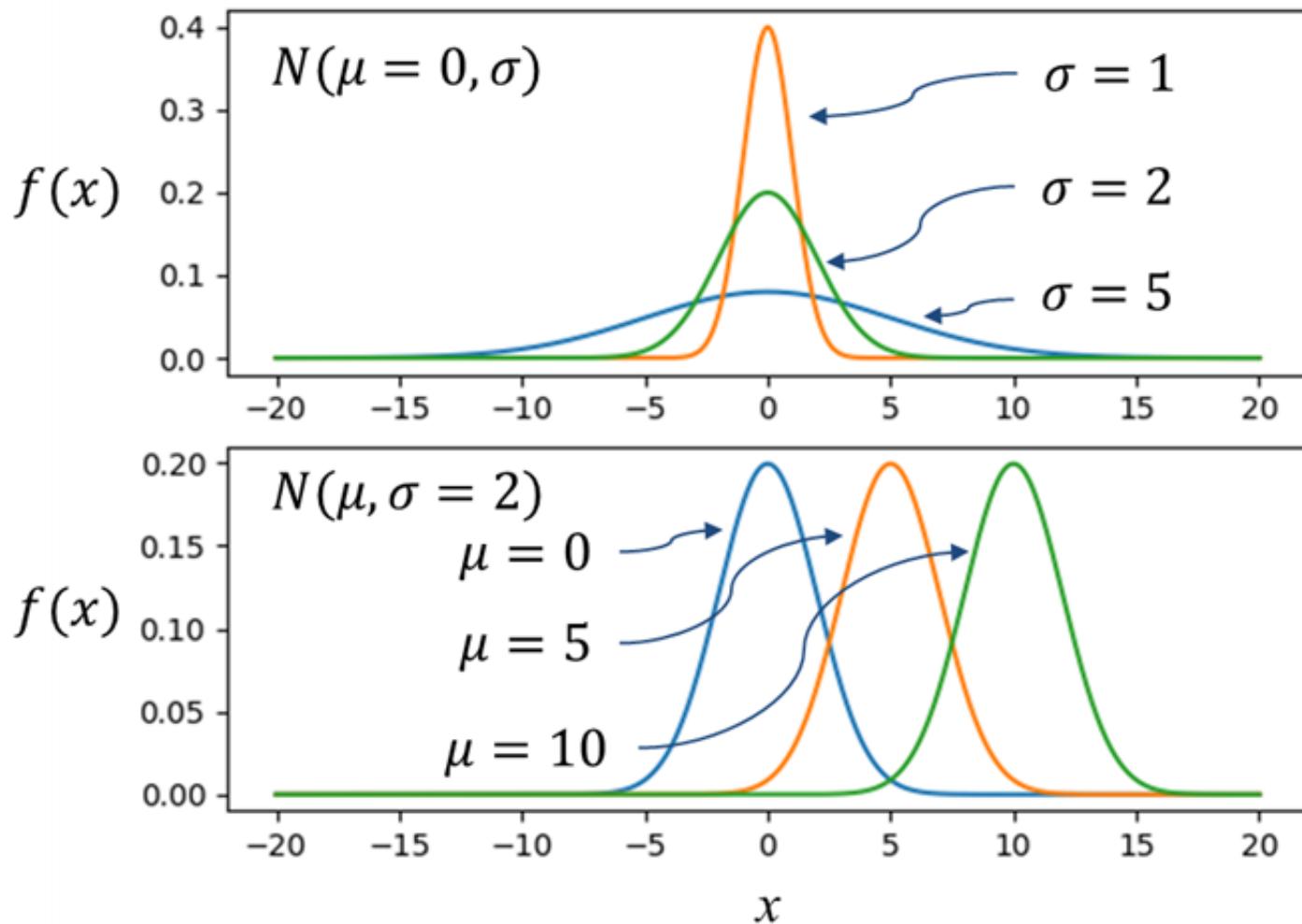
$\mu$ 和 $\sigma$ 為此常態分布的參數，其機率密度函數(也稱為高斯機率密度函數)為

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- 當 $\mu = 0$ 和 $\sigma = 1$ 的常態分布( $N(0,1)$ )我們稱為標準常態分布(*Standard normal distribution*)。



# 統計機率分布模型-常態分布



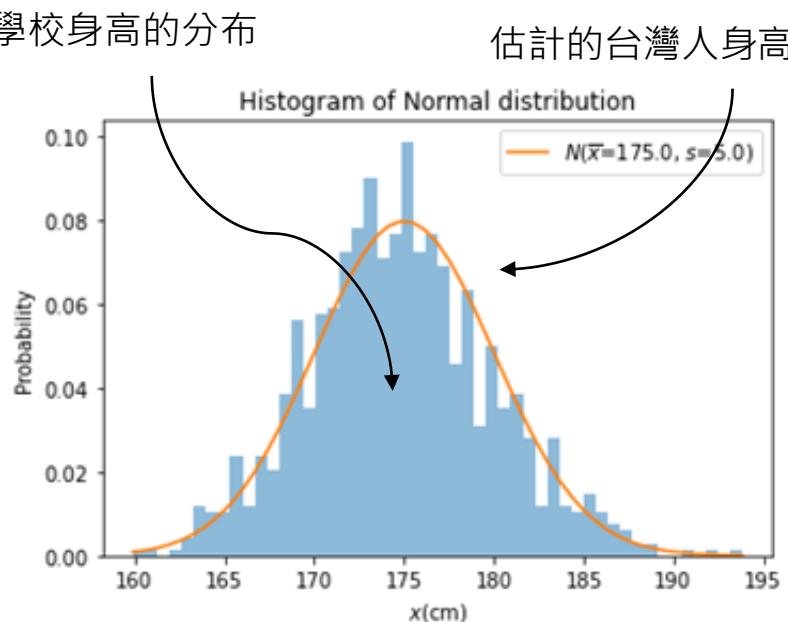
# 統計機率分布模型-常態分布-範例

- 台灣人身高( $X$ )的分布，如果找一人出來，他身高為180公分的機率為何？

如果假設台灣人的身高分布服從常態分布(  $X \sim N(\mu, \sigma^2)$  )，某A學校全校有1000人每個人身高經由統計平均數為175公分，標準數為5公分。

因為是抽樣的資料(非母體)，所以會用 $\bar{x}$ 表示(估計) $\mu$ ，會用 $s$ 表示(估計) $\sigma$ 。

A學校身高的分布



估計的台灣人身高分布

$$f(x; \mu, \sigma^2) \leftarrow f(x; \bar{x}, s^2) = \frac{1}{s\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\bar{x}}{\sigma}\right)^2} = \frac{1}{5\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-175}{5}\right)^2}$$

$$f(X=180; \bar{x}, s^2) = \frac{1}{5\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{180-175}{5}\right)^2} \approx 0.0484$$

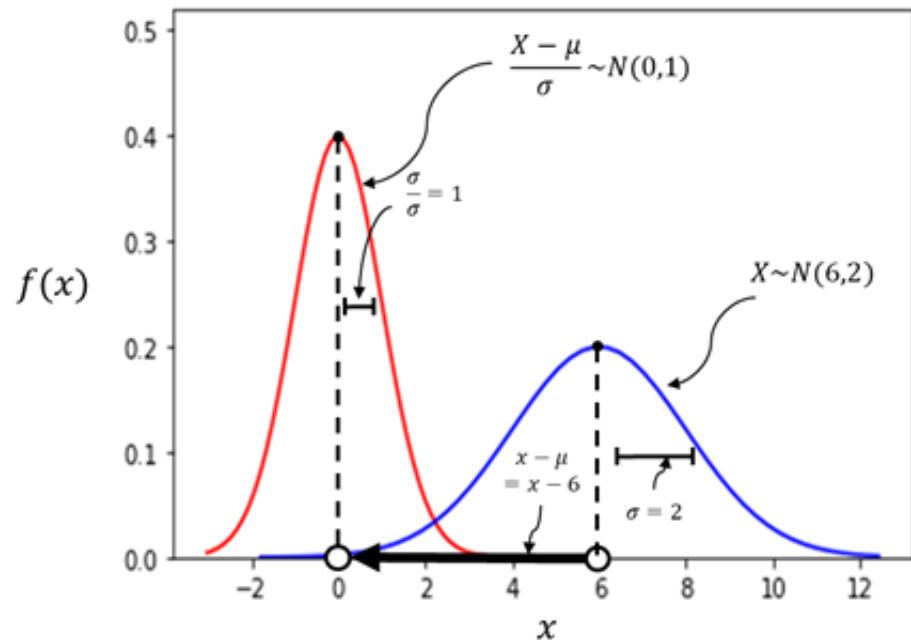
所以隨機從路上找到一個人身高剛好為180公分的機率為4.84%



# 統計機率分布模型-常態分布

- 由上圖此可知藉由平均數平移和改變變異數的範圍來改變資料的特性，而在統計學上最有名的方式即稱為z-score，

$$z\text{-score} = \frac{x - \mu}{\sigma}$$

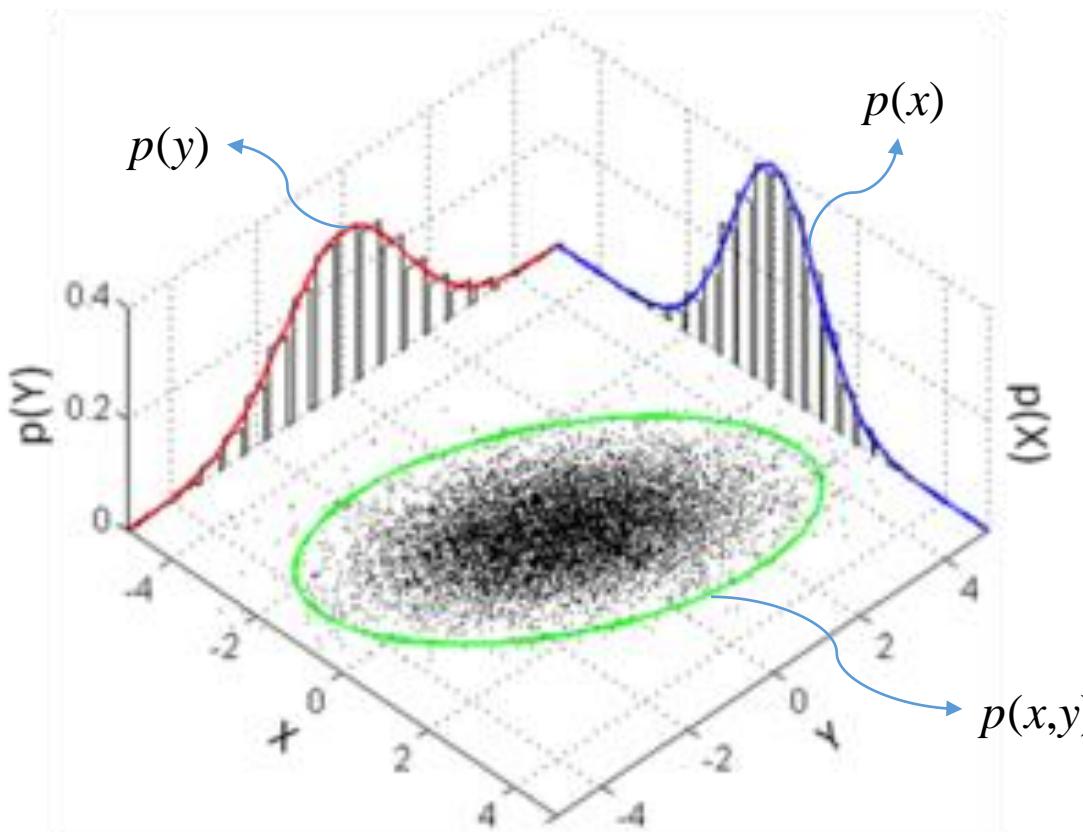


優點:

- 如果資料是常態分布，經由z-score標準化後，將資料的分布將轉換成標準常態分佈。(藍色轉換成紅色)
- 避免不同變數資料之間的變異數差異過大造成建模的影響。例如年收入(0~10,000,000元)和體重(0~200公斤)。



# Distribution Joint probability



[https://upload.wikimedia.org/wikipedia/commons/thumb/9/95/Multivariate\\_normal\\_sample.svg/300px-Multivariate\\_normal\\_sample.svg.png](https://upload.wikimedia.org/wikipedia/commons/thumb/9/95/Multivariate_normal_sample.svg/300px-Multivariate_normal_sample.svg.png)

Discrete: probability mass function.

Probability:

$$p(x) = \sum_i p(x_i) = 1$$

Joint probability:

$$p(x, y) = \sum_i \sum_j p(x_i, y_i) = 1$$

Continuous: probability density function.

Probability:

$$p(x) = \int p(x_i) dx = 1$$

Joint probability:

$$p(x, y) = \iint p(x_i, y_i) = 1$$



# Distribution (multivariate normal distribution)

Single-variate normal distribution:

$$x \sim N(\mu, \sigma^2)$$
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

multivariate normal distribution:

$$\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$f(\mathbf{x}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$



# 將統計量作為資料的特徵表徵

深度學習可以自動從非結構資料中學習特徵表徵(*Feature representation*)。

在機器學習或是資料科學則需要人為介入，尤其是資料科學需要探討資料背後隱含的特性或是挖掘眾多資料內的有效訊息。

一般通常會利用統計量(*statistics*)來作為機器學習和資料科學中的特徵表徵。

本節將介紹以下的統計方法：

- 1. 期望值
- 2. 各階中心動差
- 3. 相關係數與共變異數
- 4. 共變異數矩陣



# 1. 期望值

期望值(*expectation*) 是希望由過去蒐集的資料中，在事件未發生前，進行統計推論得到一個期望值(預期出現的值)。

期望值的定義是，假設隨機變數 $X$ 的機率分布是 $f(x)$ ，其期望值為：

$$E(X) = \sum_x xf(x)$$

從公式來看，期望值就是計算加權平均，其中 $x$ 是隨機變數的值， $f(x)$ 是隨機變數的權值 (*weight*) 或稱權重。



# 六面骰子點數的期望值

投擲一個公正的六面骰子來說，隨機變數 $X \in \{1,2,3,4,5,6\}$ ，期望值算法：

$$f(X = x) = \begin{cases} \frac{1}{6} & x = 1 \\ \frac{1}{6} & x = 2 \\ \frac{1}{6} & x = 3 \\ \frac{1}{6} & x = 4 \\ \frac{1}{6} & x = 5 \\ \frac{1}{6} & x = 6 \end{cases}$$
$$\begin{aligned} E(X) &= \sum_{x=1}^6 xf(x) \\ &= 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} \\ &\quad + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5 \end{aligned}$$



# 標準常態分布的隨機變數期望值

假設隨機變數 $X$ 服從標準常態分布 $N(0,1)$ ，其期望值為

$$\begin{aligned}
 E(X) &= \int_{-\infty}^{\infty} x \cdot f(x) dx \\
 &= \int_{-\infty}^{\infty} x \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx \\
 &= -\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \Big|_{-\infty}^{\infty} \\
 &= -\frac{1}{\sqrt{2\pi}} - \left( -\frac{1}{\sqrt{2\pi}} \right) = 0
 \end{aligned}$$

因為  $x \cdot e^{-\frac{1}{2}x^2}$  剛好是  $e^{-\frac{1}{2}x^2}$   
的微分乘以  $-1$



## 2. 各階中心動差

動差(*moment*)也是統計學常用的統計量，主要用來評估隨機變數與特定值 $a$ 之間差異的 $n$ 次方之期望值。

假設 $X$ 為一隨機變數，其與特定值 $a$ 的 $n$ 階動差定義為：

$$E[(X - a)^n]$$

變數 $X$ 的 $n$ 階中心動差(*Central moment*)，亦稱中央動差或主動差，也就是以平均數 $\mu$ 為中心的 $n$ 階動差。

$$E[(X - \mu)^n]$$



## 2. 各階中心動差- 0 階與 1 階中心動差

當  $n = 0$  為 0 階中心動差

$$E[(X - \mu)^0] = E[1] = 1$$

當  $n = 1$  為 1 階中心動差

$$E[(X - \mu)^1] = E[X] - \mu = 0$$



## 2. 各階中心動差- 2階中心動差：變異數

當  $n = 2$  為 2 階中心動差

$$E[(X - \mu)^2]$$

上式就是變異數(Variance)的定義，用  $Var(X)$  或  $\sigma_X^2$  表示。

變異數的平方根  $\sigma_X = \sqrt{Var(X)}$  稱為標準差(Standard deviation)

變異數就是隨機變數和其中心點(平均數) 差值平方的期望值，其使用意義是在估算隨機變數資料分布的分散程度，變異數越大代表資料分散程度越大。



## 2. 各階中心動差- 2階中心動差：變異數

2 階中心動差  $E[(X - \mu)^2]$  是變異數的一般式

當  $X$  有  $n$  個樣本點，且服從均勻分布  $X \sim U(1, n)$ ：

$$f(x) = \begin{cases} \frac{1}{n} & x = 1, 2, \dots, n \\ 0 & O.W. \end{cases}$$

$$\begin{aligned} E[(X - \mu)^2] \\ = \sum_{i=1}^n (x_i - \mu)^2 f(x_i) \end{aligned}$$

$$= \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

高中課本會看到的變異數公式



## 2. 各階中心動差- 3 階中心動差：偏度

當  $n = 3$  的 3 階標準中心動差也就是**偏度**(Skewness)的定義

$$Skewness = E \left[ \left( \frac{X - \mu}{\sigma} \right)^3 \right] = E[Z^3]$$

「偏度」為量化隨機變數機率的不對稱性，用來敘述資料分布的偏斜方向和大小程度的一種估算統計。

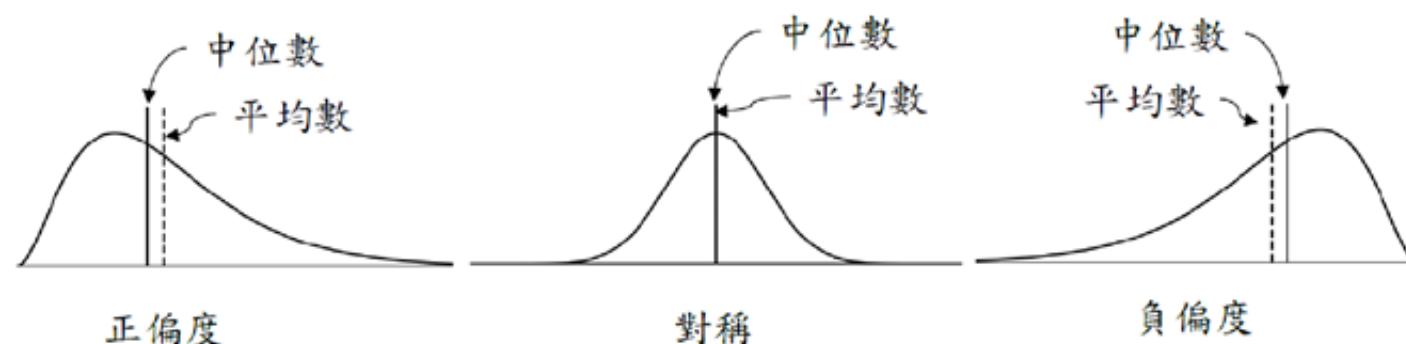


圖 3.5 資料分布的偏度



## 2. 各階中心動差- 4 階中心動差：峰度

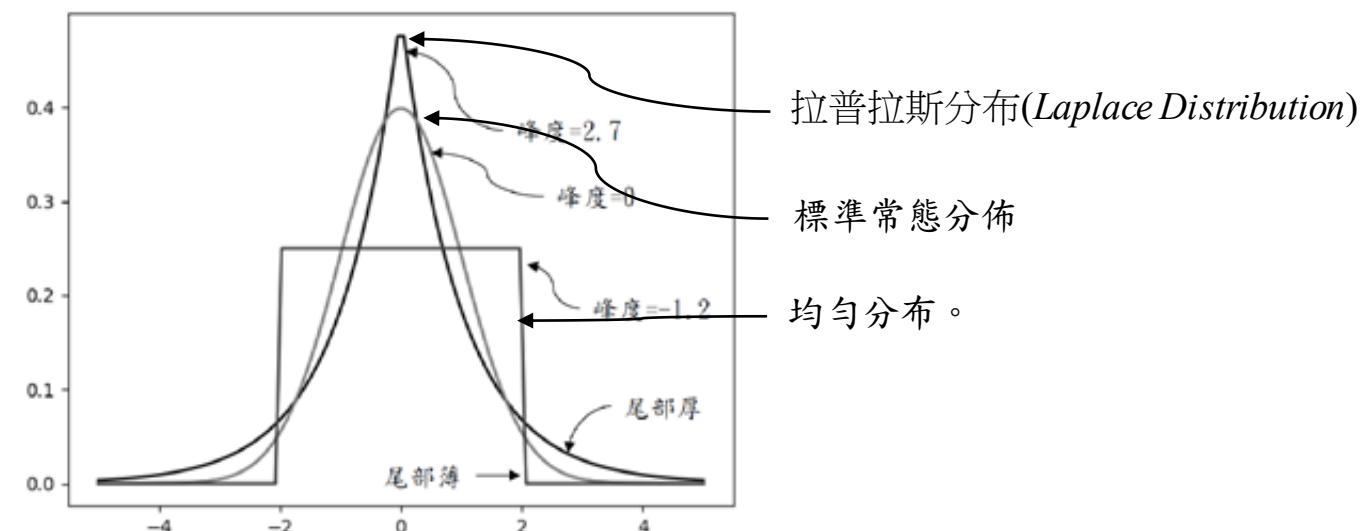
當  $n = 4$  的 4 階標準中心動差也就是峰度(Kurtosis)的定義

$$\text{Kurtosis} = E \left[ \left( \frac{X - \mu}{\sigma} \right)^4 \right] = E[Z^4]$$

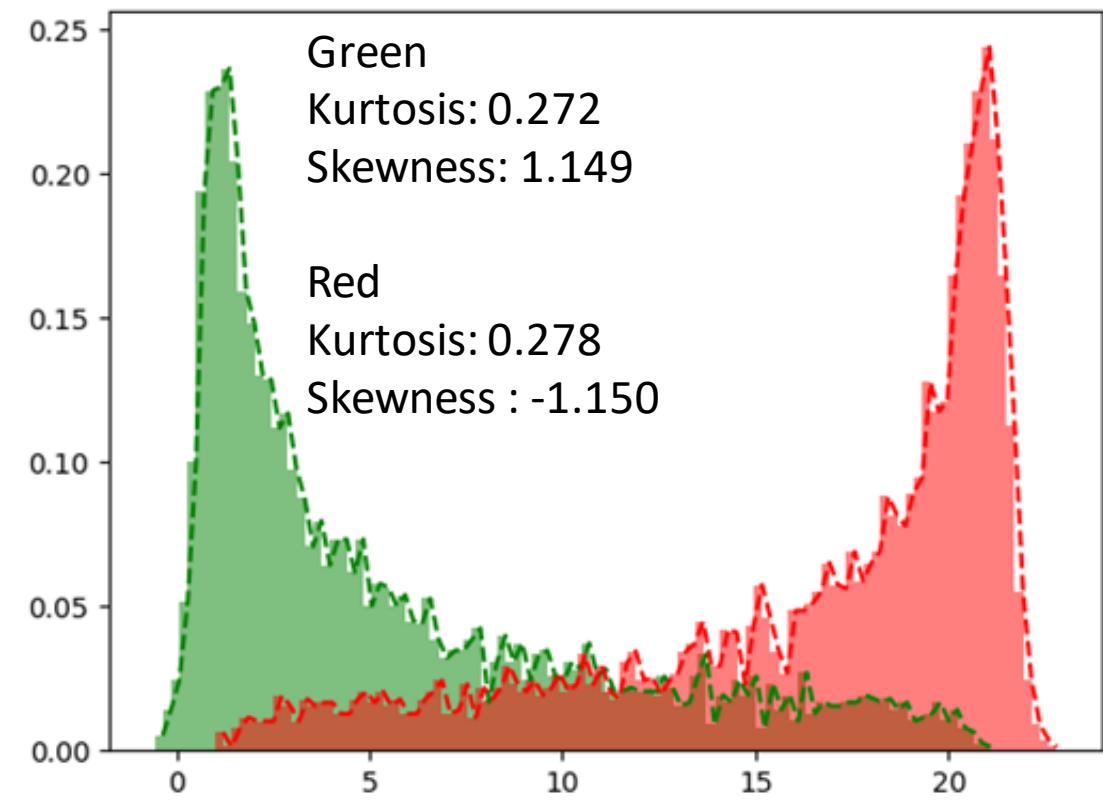
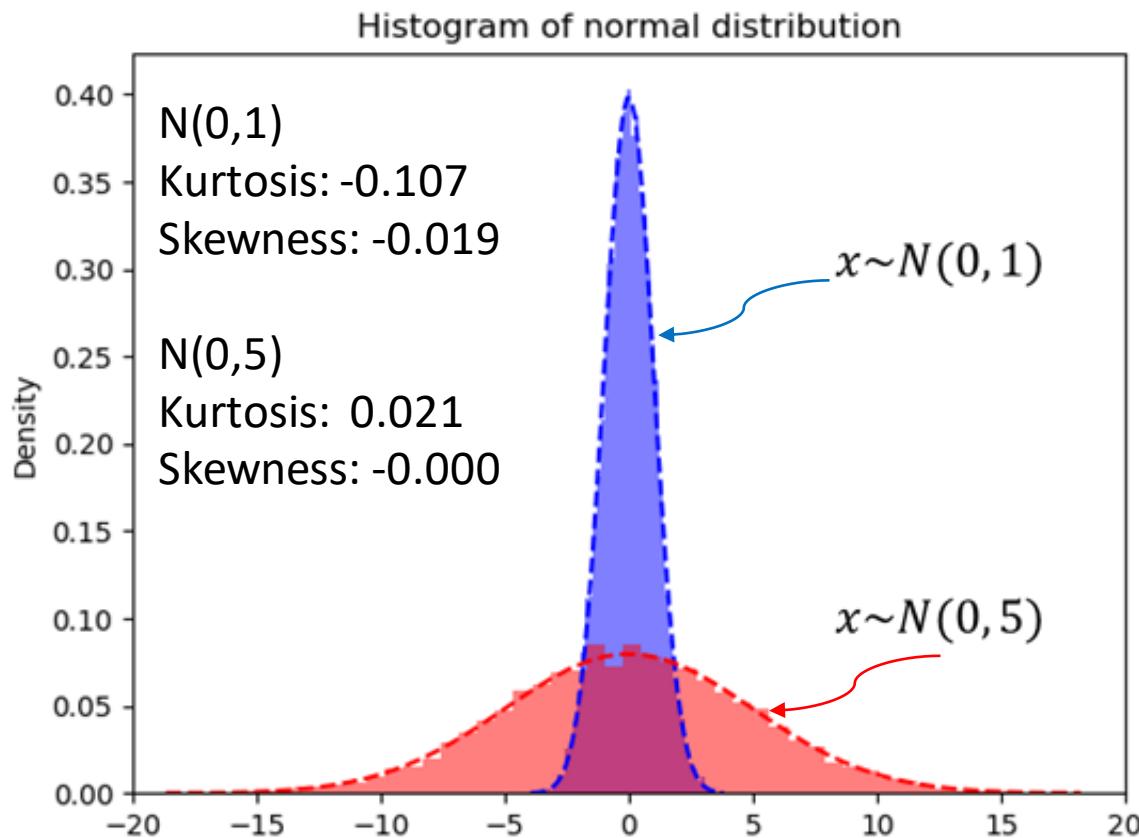
有時候公式會寫成

$$\text{Kurtosis} = E \left[ \left( \frac{X - \mu}{\sigma} \right)^4 \right] - 3$$

因為標準常態分佈的峰度為3，減去3的用意是希望標準常態分佈的峰度為0。

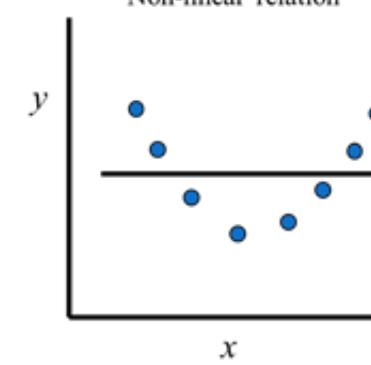
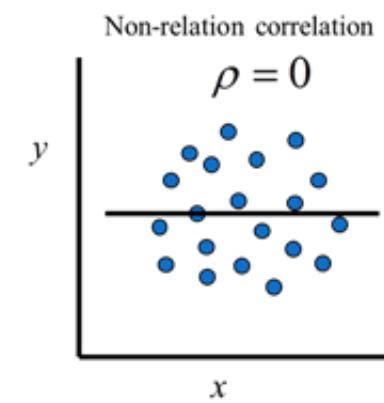
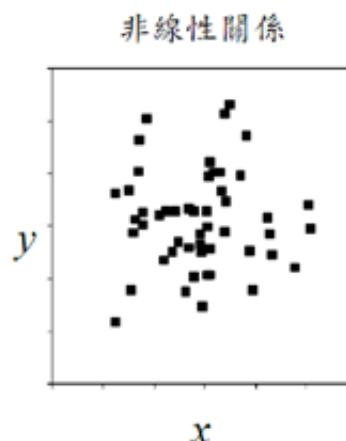
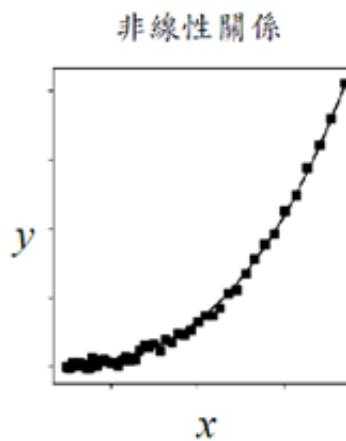
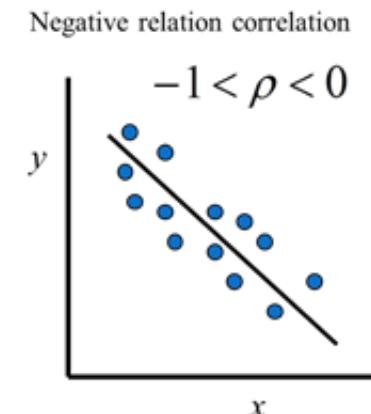
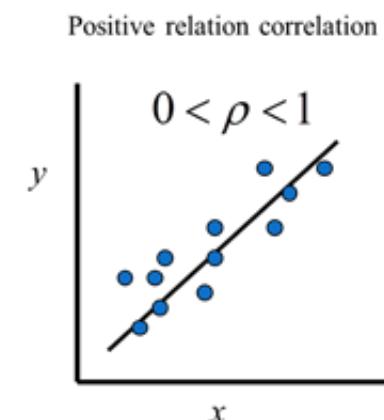
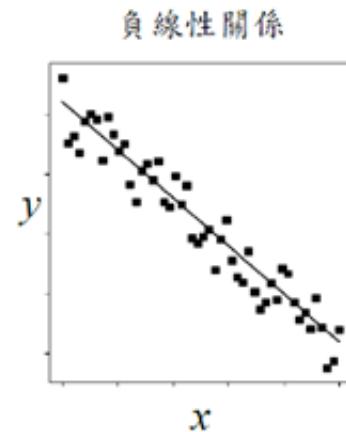
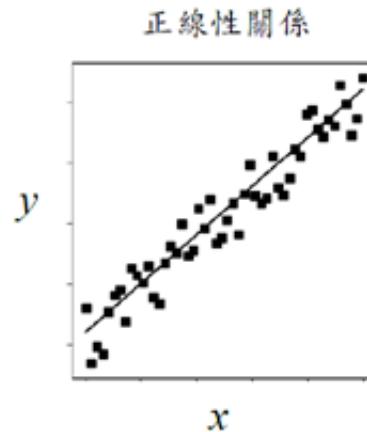


# Basic statistics



### 3. 相關係數與共變異數

- 共變異數(Covariance)就可用來檢視多維度變數之間的相關性。
- 相關係數(Correlation Coefficient)則是基於共變異數得到的統計量。



### 3. 相關係數與共變異數

共變異數(covariance)是用來計算相關係數的一個統計量，主要用來衡量兩個隨機變數共同變化的程度，也就是其線性關係。共變異數定義公式如下：

$$Cov(X, Y) = E((X - \mu_X)(Y - \mu_Y)) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)$$

而當 $X = Y$ ，則 $Cov(X, Y)$ 就等於計算隨機變數 $X$ 的變異數( $Var(X)$ )：

$$Cov(X, Y) = Cov(X, X) = E((X - \mu_X)(X - \mu_X)) = E((X - \mu_X)^2) = Var(X)$$



### 3. 相關係數與共變異數-範例

範例1：「男生的身高和體重的共變異數」

範例1 計算出的共變異數為：

範例 1 男生的身高和體重的共變異數				
身高 ( $X$ )	體重 ( $Y$ )	$(x_i - \mu_x)$	$(y_i - \mu_y)$	$(x_i - \mu_x)(y_i - \mu_y)$
180	80	10	12	120
175	70	5	2	10
170	70	0	2	0
165	70	-5	2	-10
160	50	-10	-18	180

男生的身高和體重的共變異數

$$\frac{1}{5} \sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y) = 300 / 5 = 60$$



### 3. 相關係數與共變異數

皮爾森相關係數評估線性相依性

假設有兩個隨機變數 $X$  和  $Y$ ，皮爾森相關係數 (*correlation coefficient*) 公式定義如下：

$$\rho = \frac{E((X - \mu_X)(Y - \mu_Y))}{\sqrt{E((X - \mu_X)^2)} \sqrt{E((Y - \mu_Y)^2)}} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

$\mu_X$  和  $\mu_Y$  分別為隨機變數  $X$  和  $Y$  的平均數， $\sigma_X$  和  $\sigma_Y$  分別為隨機變數  $X$  和  $Y$  的變異數，且  $-1 \leq \rho \leq 1$ 。



### 3. 相關係數與共變異數

相關係數回到比值的基準線

$$\rho = \frac{\text{cov}(x, y) \text{ (單位: 公分 * 公斤)}}{std(x) \text{ (單位: 公分)} \times std(y) \text{ (單位: 公斤)}}$$
$$= \frac{\text{cov}(x, y) \text{ (單位: 公分 * 公斤)}}{std(x) \text{ (單位: 公分)} \times std(y) \text{ (單位: 公斤)}}$$

值會落在正負1之間



# 相關係數範圍介於 -1 到 1

至於相關係數為什麼會介於 -1 到 1 之間。

做法一：利用柯西不等式(*Cauchy-Schwarz inequality*) 來證明。  
我們將相關係數取平方

$$\rho^2 = \left( \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \right)^2$$

分子項： $\text{Cov}^2(X, Y) = (E[(X - \mu_X)(Y - \mu_Y)])^2$   
分母項： $\sigma_X^2 \sigma_Y^2 = E[(X - \mu_X)^2]E[(Y - \mu_Y)^2]$

套用柯西不等式可得出相關係數介於 -1 到 1 之間：

$$\begin{aligned}\text{Cov}^2(X, Y) &= (E[(X - \mu_X)(Y - \mu_Y)])^2 \leq E[(X - \mu_X)^2]E[(Y - \mu_Y)^2] = \sigma_X^2 \sigma_Y^2 \\ \Rightarrow \frac{\text{Cov}^2(X, Y)}{\sigma_X^2 \sigma_Y^2} &= \rho^2 \leq 1 \\ \Rightarrow -1 \leq \rho &\leq 1\end{aligned}$$



## 4. 共變異數矩陣

共變異數是看兩個隨機變數之間的相關性。

當變數大於兩個以上，則可以用共變異數矩陣(*Covariance matrix*)來描述整批資料的分散量或分散性。

假設有 $d$ 個隨機變數和 $n$ 筆資料，整個資料的矩陣為 $\mathbf{X} \in \mathbb{R}^{d \times n}$

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{d1} & x_{d2} & \cdots & x_{dn} \end{bmatrix}_{d \times n}$$

第2筆資料

← 第2個隨機變數

← 第 $d$ 個隨機變數

第 $j$ 個變數和第 $k$ 個變數的共變異數為

$$\text{Cov}(X_j, X_k) = E[(X_j - E(X_j))(X_k - E(X_k))^T]$$



## 4. 共變異數矩陣

共變異數矩陣定義為

$$\Sigma = \begin{bmatrix} Cov(X_1, X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_d) \\ Cov(X_2, X_1) & Cov(X_2, X_2) & \cdots & Cov(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_d, X_1) & Cov(X_d, X_2) & \cdots & Cov(X_d, X_d) \end{bmatrix}$$

共變異數矩陣特性：

1.  $\Sigma$ 為方陣。
2.  $\Sigma$ 為對稱矩陣(*symmetric matrix*)， $\Sigma = \Sigma^T$ ， $Cov(X_i, X_j) = Cov(X_j, X_i)$
3.  $\Sigma$ 為半正定矩陣(*positive semi-definite matrix*)，也就是其行列式必須大於等於0  
 $det(\Sigma) \geq 0$ 。



## 4. 共變異數矩陣-範例

- 共變異數矩陣來敘述資料集的分散量
  - 用兩個來自常態分佈的隨機變數( $X_1, X_2$ )來進行生成資料，

$$X_1 \sim N(0,5), X_2 \sim N(0,2)$$

範例一： $X_1$  和  $X_2$  彼此獨立 ( $Cov(X_1, X_2) = 0$ )，其共變異數矩陣

$$\Sigma = \begin{bmatrix} 5 & 0 \\ 0 & 2 \end{bmatrix}.$$

範例二： $X_1$  和  $X_2$  彼此非獨立 ( $Cov(X_1, X_2) = 1$ )，其共變異數矩陣

$$\Sigma = \begin{bmatrix} 5 & 1 \\ 1 & 2 \end{bmatrix}.$$

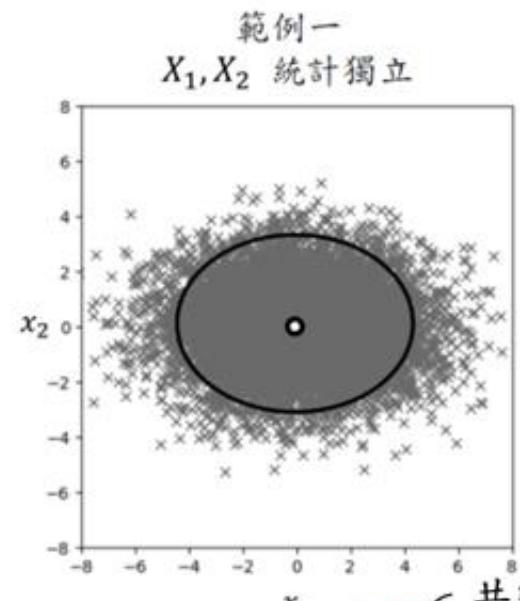
範例三： $X_1$  和  $X_2$  彼此非獨立 ( $Cov(X_1, X_2) = 3$ )，其共變異數矩陣

$$\Sigma = \begin{bmatrix} 5 & 3 \\ 3 & 2 \end{bmatrix}.$$

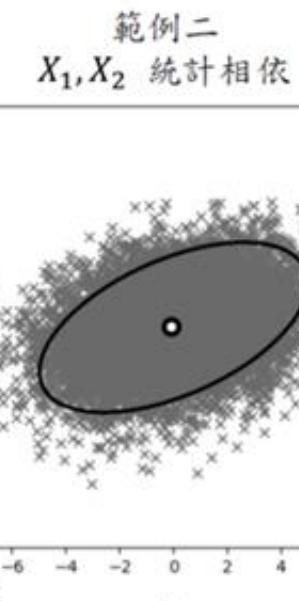


## 4. 共變異數矩陣-範例

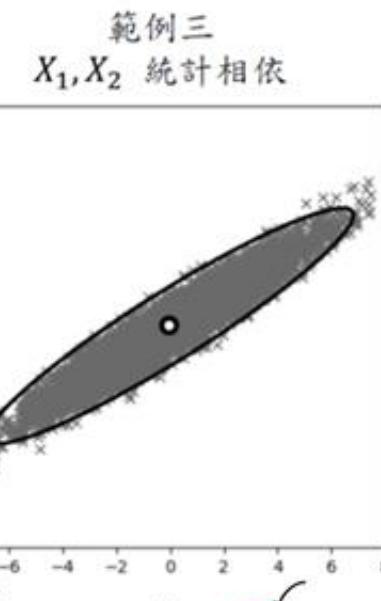
- 這三個範例參數從常態分佈生成各自一萬筆資料，觀察在同樣標準差下，不同共變異數不同的結果。



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 5 & 0 \\ 0 & 2 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 5 & 1 \\ 1 & 2 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 5 & 3 \\ 3 & 2 \end{bmatrix}$$

