

巨量資料管理學院碩士在職專班

# 統計分析

2022/10/7

陳光宏

# 軟體操作

# EXCEL樞紐分析表

- 請參考資料檔：[Week 3檔案.xlsx](#)
- 請製作銷售地區與商品名稱的列聯表
- 請問銷售地區與銷售的商品種類是否有相關？

訂單編號	交易日期	客戶編號	商品名稱	銷售數量	業務姓名	銷售地區	成交單價
P80105	2019/1/5	C81001	電腦	2	Alex Wang	東區	\$21,000
P80106	2019/1/6	C81006	螢幕	2	Grace Fang	中區	\$3,000
P80107	2019/1/7	C81003	印表機	7	Eddy Chen	北區	\$2,500
P80108	2019/1/17	C81002	螢幕	5	Eddy Chen	北區	\$3,000
P80109	2019/1/11	C81005	螢幕	16	Bob Lee	中區	\$3,000
P80110	2019/1/14	C81007	印表機	9	Frank Hsio	東區	\$2,000
P80111	2019/1/15	C81004	電腦	2	Chris Chang	南區	\$21,000
P80112	2019/1/9	C81004	電腦	3	Hans Lin	南區	\$21,000
P80113	2019/1/18	C81009	螢幕	8	Chris Chang	南區	\$3,000
P80114	2019/1/10	C81000	螢幕	4	Hans Lin	南區	\$2,000

檔案 常用 **插入** 頁面配置 公式 資料 校閱 檢視 開發人員 說明 Data Streamer Power Pivot

樞紐分析表 建議的樞紐分析表 表格 圖片 圖示 3D 模型

建立樞紐分析表

選擇您要分析的資料

☒ 選取表格或範圍(S)

表格/範圍(I): 工作表1!\$A\$1:\$H\$41

☐ 使用外部資料來源(U)

選擇連線(C)...

連線名稱:

☐ 使用此活頁簿的資料模型(D)

選擇您要放置樞紐分析表的位置

☒ 新工作表(N)

☐ 已經存在的工作表(E)

位置(L):

選擇您是否要分析多個表格

☐ 新增此資料至資料模型(M)

確定 取消

	A	B	
1	訂單編號	交易日期	客戶
2	P80105	2019/1/5	C8
3	P80106	2019/1/6	C8
4	P80107	2019/1/7	C8
5	P80108	2019/1/17	C8
6	P80109	2019/1/11	C8

G	H
售地區	成交單價
區	\$21,000
區	\$3,000
區	\$2,500
區	\$3,000
區	\$3,000

加總 - 銷售數量	欄標籤				
列標籤	印表機	電腦	螢幕	總計	
中區	17	4	41	62	
北區	28	37	18	83	
東區	35	8	29	72	
南區	49	15	28	92	
總計	129	64	116	309	

## 樞紐分析表欄位

選擇要新增到報表的欄位:

搜尋

- ☐ 訂單編號
- ☐ 交易日期
- ☐ 客戶編號
- ☒ 商品名稱
- ☒ 銷售數量
- ☐ 業務姓名
- ☒ 銷售地區
- ☐ 成交單價

其他表格...

在下列區域之間拖曳欄位:

篩選

欄

商品名稱

列

值

銷售地區

加總 - 銷售數量

加總 - 銷售數量 欄標籤  
列標籤 印表機 電腦

	印表機	電腦
中區	17	4
北區	28	37
東區	35	8
南區	49	15
總計	129	64

新細明體 12 A<sup>+</sup> A<sup>-</sup> \$ % ,

B I

- 複製(C)
- 儲存格式(E)...
- 數字格式(D)...
- 重新整理(R)
- 排序(S) >
- ✕ 移除 "加總 - 銷售數量"(V)
- 摘要值方式(M) >
- 值的顯示方式(A) >
- 顯示詳細資料(E)
- 值欄位設定(N)...
- 樞紐分析表選項(O)...
- 隱藏欄位清單(D)

- ✓ 加總(S)
- 項目個數(C)
- 平均值(A)
- 最大值(M)
- 最小值(I)
- 乘積(P)
- 更多選項(O)...

新細明體 12 A A \$ % ,

B I 格式 顏色 背景 減半 增加 0.00 0.00

計數 - 銷售數量	欄標籤	欄標籤
列標籤	印表機	電腦
中區	2	2
北區	3	3
東區	3	3
南區	4	4
總計	12	12

- 複製(C)
- 儲存格式(E)...
- 數字格式(I)...
- 重新整理(R)
- 排序(S) >
- ✗ 移除 "計數 - 銷售數量"(V)
- 摘要值方式(M) >
- 值的顯示方式(A) >
- + 顯示詳細資料(E)
- 值欄位設定(N)...
- 樞紐分析表選項(O)...
- 隱藏欄位清單(D)

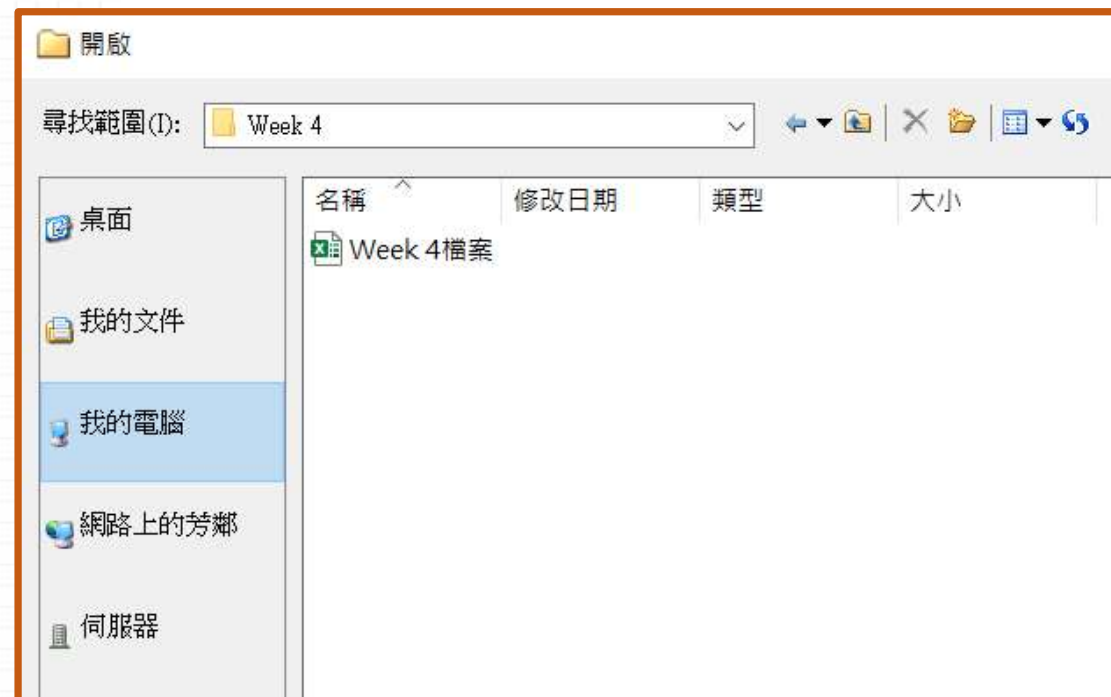
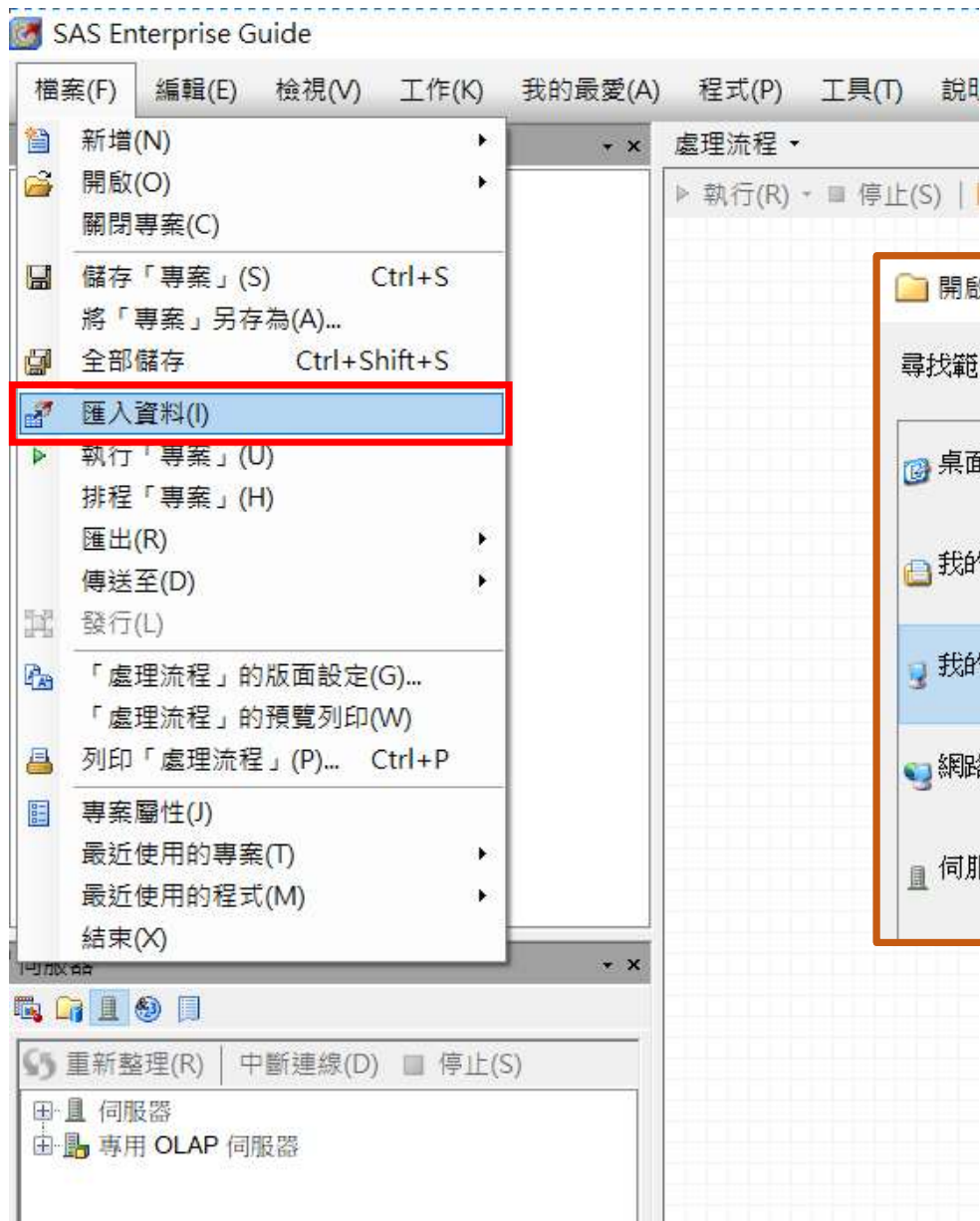
- ✓ 無計算(N)
- 總計百分比(G)
- 欄總和百分比(C)
- 列總和百分比(R)
- 百分比(O)...
- 父項列總和百分比(P)
- 父項欄總和百分比(A)
- 父項總和百分比(E)...
- 差異(D)...
- 差異百分比(E)...
- 計算加總至(I)...
- 計算加總至百分比(U)...
- 最小到最大排列(S)...
- 最大到最小排列(L)...

計數 - 銷售數量	欄標籤				
列標籤		印表機	電腦	螢幕	總計
中區		20.00%	20.00%	60.00%	100.00%
北區		30.00%	40.00%	30.00%	100.00%
東區		37.50%	25.00%	37.50%	100.00%
南區		33.33%	33.33%	33.33%	100.00%
總計		30.00%	30.00%	40.00%	100.00%



# 製作列聯表 – SAS EG

Table of 銷售地區 by 商品名稱					
		商品名稱			總計
		印表機	電腦	螢幕	
銷售地區					
中區	次數	2	2	6	10
	列百分比	20.00	20.00	60.00	
北區	次數	3	4	3	10
	列百分比	30.00	40.00	30.00	
東區	次數	3	2	3	8
	列百分比	37.50	25.00	37.50	
南區	次數	4	4	4	12
	列百分比	33.33	33.33	33.33	
總計	次數	12	12	16	40



從 Week 4檔案.xlsx 匯入的資料 ▾

篩選和排序(L) 查詢產生器(Q) 位置(W)   資料(D) ▾ 描述(E) ▾ 圖形(G) ▾ 分析(Z) ▾ 匯出(X) ▾ 傳送至(N) ▾									
	⚠ 訂單編號	📅 交易日期	⚠ 客戶編號	⚠	清單資料(L)...		⚠ 業務姓名	⚠	
1	P80105	05JAN2019	C81001	電腦	Σ	摘要統計精靈(M)...	2	Alex Wang	東區
2	P80106	06JAN2019	C81006	螢幕	Σ	摘要統計(S)...	2	Grace Fang	中區
3	P80107	07JAN2019	C81003	印表機	📊	摘要表精靈(B)...	7	Eddy Chen	北區
4	P80108	17JAN2019	C81002	螢幕	📊	摘要表(T)...	5	Eddy Chen	北區
5	P80109	11JAN2019	C81005	螢幕	📊	清單報表精靈...	6	Bob Lee	中區
6	P80110	14JAN2019	C81007	印表機	📊	特徵化資料(H)...	9	Frank Hsio	東區
7	P80111	15JAN2019	C81004	電腦	📊	分配分析(D)...	2	Chris Chang	南區
8	P80112	09JAN2019	C81004	電腦	📊	單因子次數(O)...	3	Hans Lin	南區
9	P80113	18JAN2019	C81009	螢幕	📊	表格分析(A)...	8	Chris Chang	南區
10	P80114	19JAN2019	C81009	螢幕			4	Hans Lin	南區

資料

表格

儲存格統計值

表格統計值

關聯性

一致性

排序差異

趨勢檢定

計算選項

結果

儲存格統計值結果

表格統計值結果

標題

屬性

資料

資料來源: Local:WORK.WEEK 4檔案

工作篩選: 無

要指派的變數(A):

名稱

△ 訂單編號

交易日期

△ 客戶編號

△ 商品名稱

● 銷售數量

△ 業務姓名

▲ 銷售地區

成交單價

工作角色(T):

次數計數 (限制: 1)

分析群組依據

表格變數

△ 銷售地區

△ 商品名稱



資料

**表格**

儲存格統計值

表格統計值

關聯性

一致性

排序差異

趨勢檢定

計算選項

結果

儲存格統計值結果

表格統計值結果

標題

屬性

表格

表格中允許的變數(V):

⚠ 商品名稱

⚠ 銷售地區

預覽:

<將變數拖放至此處>


⚠ 商品名稱

⚠ 銷售地區

➡

⬅

資料

表格

儲存格統計值

表格統計值

關聯性

一致性

排序差異

趨勢檢定

計算選項

結果

儲存格統計值結果

表格統計值結果

標題

屬性

### 儲存格統計值

可用的統計值

☐ 累積欄百分比(M)

☒ 列百分比(W)

☐ 欄百分比(U)

☒ 儲存格次數(F)

☐ 儲存格百分比(P)

☐ 遺漏值次數(V)

☐ 儲存格對 Pearson 卡方的貢獻(L)


☐ 儲存格次數與預期值的離差(D)

☐ 預測的儲存格次數(E)

☐ 總次數的百分比(T)

☐ 在資料集內包含百分比(N)

選取窗格可讓您為工作選擇不同的選項集。

 預覽程式碼(C)

執行(R)

# 機率分布

# 隨機變數 (Random variables)

- 欲描述的事件
- 有多種可能的情況
- 每一種情況有特定的發生機率
- 互斥與collectively exhaustive
- 例如
  - 某人買了五樣商品，想了解這五樣商品屬於書籍類的機率
  - 丟兩個骰子，想了解出現點數總和的機率



# 機率分布 (Probability distribution)

- 描述隨機變數的行為
- 用數學來描述
- 透過機率分布，計算隨機變數每種情況下的機率



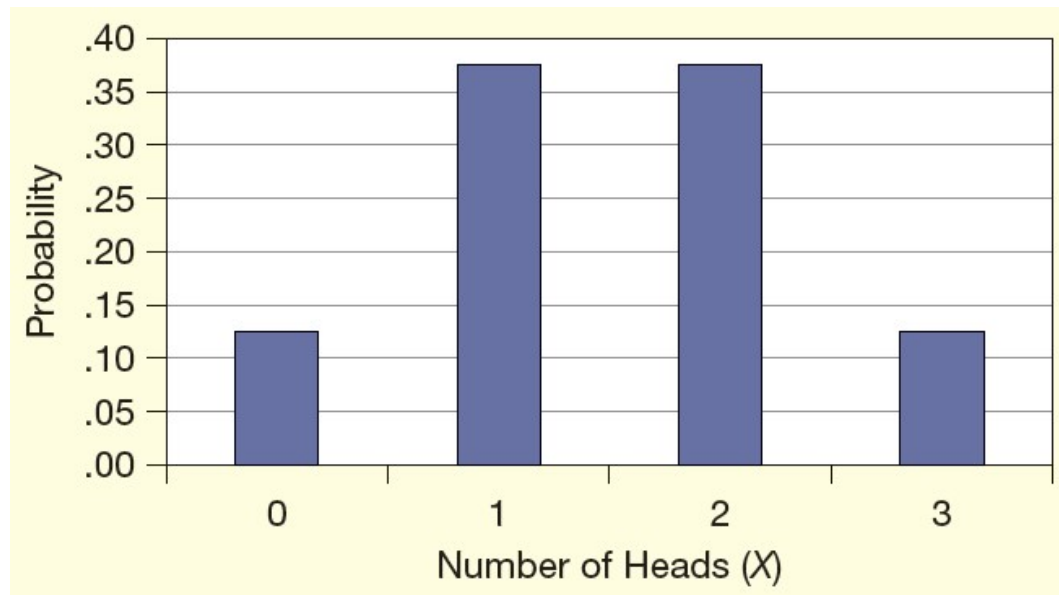
# 練習1

假設同時丟三個硬幣

1. 請寫出總共有幾種可能的情況？
2. 請列出每種情況的機率？
3. 請問隨機變數是什麼？

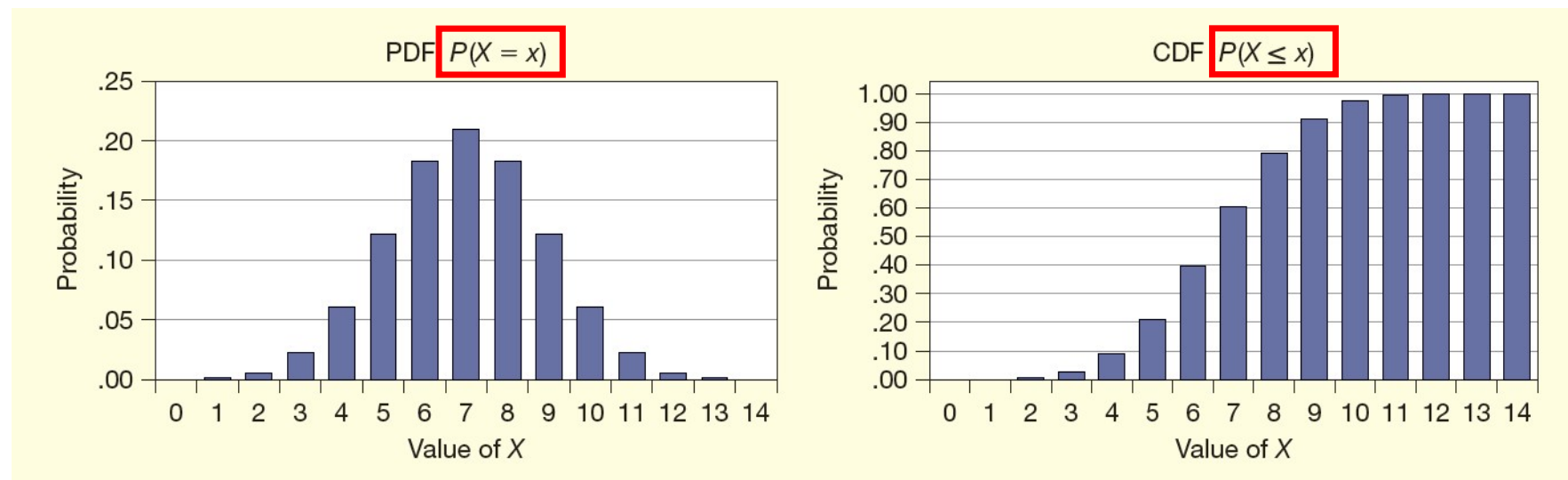
可能的情況	機率
1	---
2	---
3	---
---	---

<i>Possible Events</i>	<i>x</i>	<i>P(x)</i>
TTT	0	1/8
HTT, THT, TTH	1	3/8
HHT, HTH, THH	2	3/8
HHH	3	1/8
Total		1



# 描述機率分布

- 機率密度函數 (Probability density function, PDF)
  - 隨機變數為X軸，對應的機率值為Y軸
  - Probability mass function (PMF)
- 累積機率分布函數 (Cumulative distribution function, CDF)



## 練習2

- 請問下列三個例子裡，哪些不能說是機率分布？

Example A		Example B		Example C	
$x$	$P(x)$	$x$	$P(x)$	$x$	$P(x)$
0	.80	1	.05	50	.30
1	.20	2	.15	60	.60
		3	.25	70	.40
		4	.40		
		5	.10		

# 期望值與變異數

- 期望值 (Expected value)

- 加權平均的概念

$$E(X) = \mu = \sum_{i=1}^N x_i P(x_i)$$

- 變異數 (Variance)

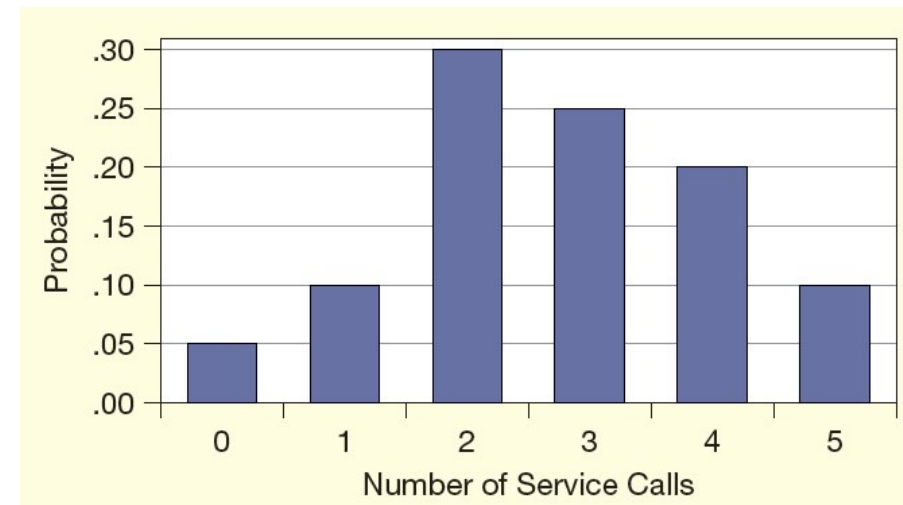
- 離平均有多遠

$$\text{Var}(X) = \sigma^2 = \sum_{i=1}^N [x_i - \mu]^2 P(x_i)$$

# 範例1

- 某公司周日緊急客服電話的PDF如下
- 請計算來電次數的期望值

$x$	$P(x)$
0	.05
1	.10
2	.30
3	.25
4	.20
5	.10
Total	1.00



## 範例2

- 某個小旅館有7間房間，二月是旅遊旺季，老闆想了解二月份佔房的狀況
- 計算平均佔房數，及其變異數

$x$	$P(x)$	$xP(x)$	$x - \mu$	$[x - \mu]^2$	$[x - \mu]^2 P(x)$
0	.05				
1	.05				
2	.06				
3	.10				
4	.13				
5	.20				
6	.15				
7	.26				
Total	1.00	$\mu =$			$\sigma^2 =$



# 解開黑盒子

- 二項式分布 (Binomial distribution)
- 卜瓦松分布 (Poisson distribution)
- 常態分布 (Normal distribution)

# 二項式分布 (Binomial distribution)

<i>Bernoulli Experiment</i>	<i>Possible Outcomes</i>	<i>Probability of "Success"</i>
Flip a coin	1 = heads 0 = tails	$\pi = .50$
Inspect a jet turbine blade	1 = crack found 0 = no crack found	$\pi = .001$
Purchase a tank of gas	1 = pay by credit card 0 = do not pay by credit card	$\pi = .78$
Do a mammogram test	1 = positive test 0 = negative test	$\pi = .0004$

# 二項式分布 (Binomial distribution)

$k$  = 成功次數

$n$  = 總數

$p$  = 成功機率

$q = 1 - p$

$$Pr(X = k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, \dots, n$$

Parameters

$n$  = number of trials  
 $\pi$  = probability of success

PDF

$$P(X = x) = \frac{n!}{x!(n-x)!} \pi^x (1 - \pi)^{n-x}$$

Excel\* PDF

=BINOM.DIST( $x$ ,  $n$ ,  $\pi$ , 0)

Excel\* CDF

=BINOM.DIST( $x$ ,  $n$ ,  $\pi$ , 1)

Domain

$x = 0, 1, 2, \dots, n$

Mean

$n\pi$

Standard deviation

$$\sqrt{n\pi(1 - \pi)}$$

Random data  
generation in Excel

=BINOM.INV( $n$ ,  $\pi$ , RAND())  
or use Excel's Data Analysis Tools

Comments

Skewed right if  $\pi < .50$ , skewed left if  $\pi > .50$ , and symmetric if  $\pi = .50$ .

## Binomial Shape

$$\pi < .50$$

skewed right

$$\pi = .50$$

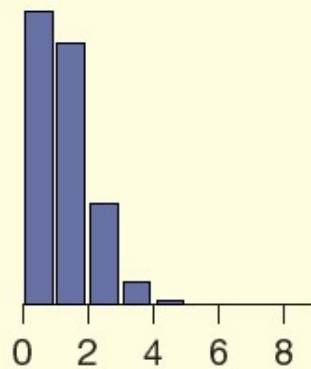
symmetric

$$\pi > .50$$

skewed left

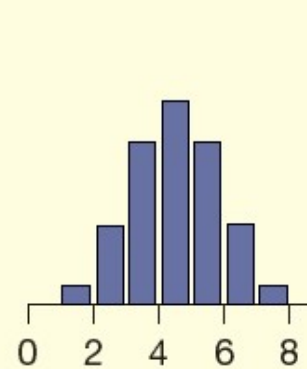
Skewed Right

$$\pi = .10$$



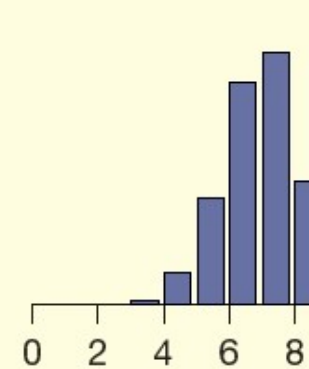
Symmetric

$$\pi = .50$$

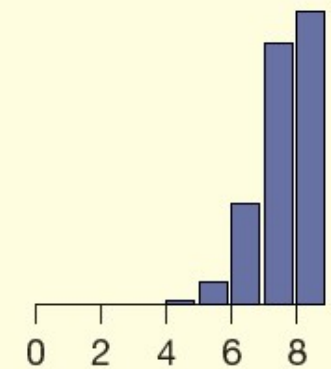


Skewed Left

$$\pi = .80$$



$$\pi = .90$$



# 範例

到某醫院急診的病人中，大約有20%沒有額外買保險

1. 隨機選取5個病人，請問至少有3個病人沒買保險的機率是多少  $\Pr(X \geq 3)$ ？
2. 承上，請問這5個病人中，預期會有多少人沒有額外買保險？

# 範例解答

方法一  
代公式

$$Pr(X = k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, \dots, n$$

方法二  
查表

Exact binomial probabilities  $Pr(X = k) = \binom{n}{k} p^k q^{n-k}$  (continued)

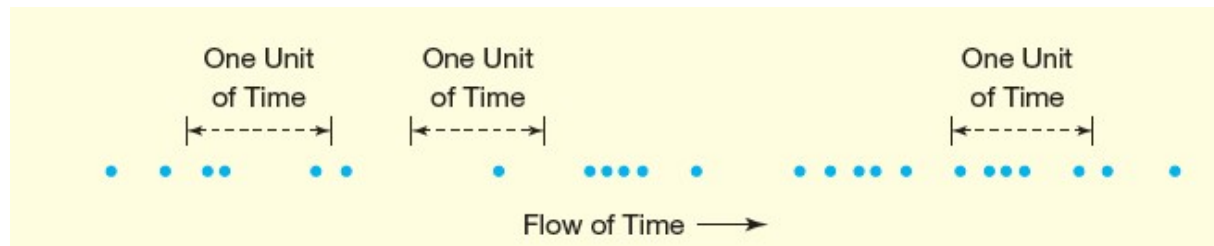
n	k	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
18		.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
19		.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
20	0	.3585	.1216	.0388	.0115	.0032	.0008	.0002	.0000	.0000	.0000
	1	.3774	.2702	.1368	.0576	.0211	.0068	.0020	.0005	.0001	.0000
	2	.1887	.2852	.2293	.1369	.0669	.0278	.0100	.0031	.0008	.0002
	3	.0596	.1901	.2428	.2054	.1339	.0716	.0323	.0123	.0040	.0011
	4	.0133	.0898	.1821	.2182	.1897	.1304	.0738	.0350	.0139	.0046
	5	.0022	.0319	.1028	.1746	.2023	.1789	.1272	.0746	.0365	.0148
	6	.0003	.0089	.0454	.1091	.1686	.1916	.1712	.1244	.0746	.0370
	7	.0000	.0020	.0160	.0546	.1124	.1643	.1844	.1659	.1221	.0739
	8	.0000	.0004	.0046	.0222	.0609	.1144	.1614	.1797	.1623	.1201

方法三  
利用excel

x	P(x)
0	0.3277
1	0.4096
2	0.2048
3	0.0512
4	0.0064
5	0.0003

# 卜瓦松分布 (Poisson distribution)

- 可視為二項式分布的一種極端例子 (稀有事件)
  - $n$  很大、 $p$  很小的時候
- 某一段時間內，發生某事件的個案數



$X$  = number of customers arriving at a bank ATM in a given minute.

$X$  = number of file server virus infections at a data center during a 24-hour period.

$X$  = number of asthma patient arrivals in a given hour at a walk-in clinic.



# 卜瓦松分布 (Poisson distribution)

$x$  = 發生個數或次數

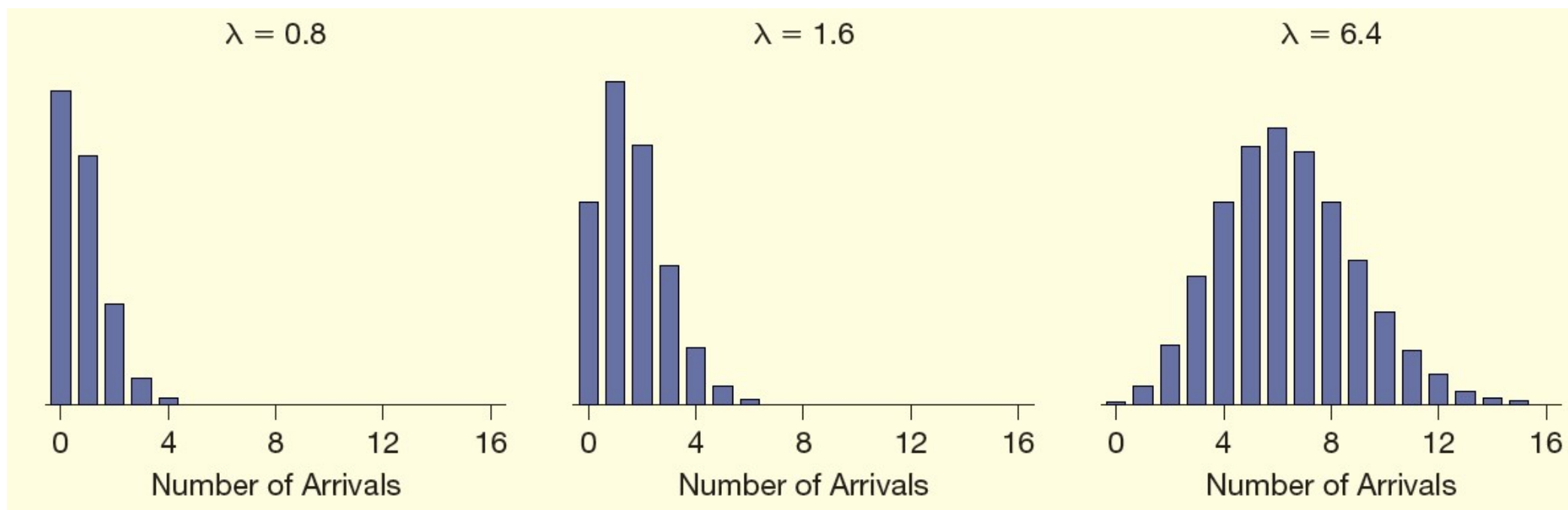
$\lambda$  = 期望發生個數或次數

$e = 2.71828.....$

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Parameter	$\lambda$ = mean arrivals per unit of time or space
PDF	$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$
Excel* PDF	=POISSON.DIST(x, $\lambda$ , 0)
Excel* CDF	=POISSON.DIST(x, $\lambda$ , 1)
Domain	$x = 0, 1, 2, \dots$ (no obvious upper limit)
Mean	$\lambda$
Standard deviation	$\sqrt{\lambda}$
Comments	Always right-skewed, but less so for larger $\lambda$ .

# 卜瓦松分布的PDF



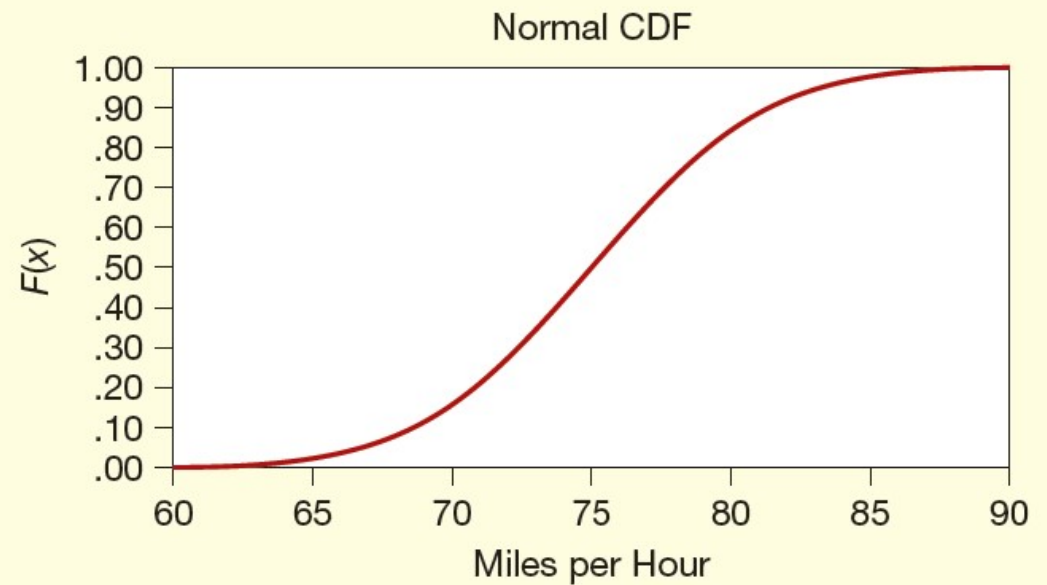
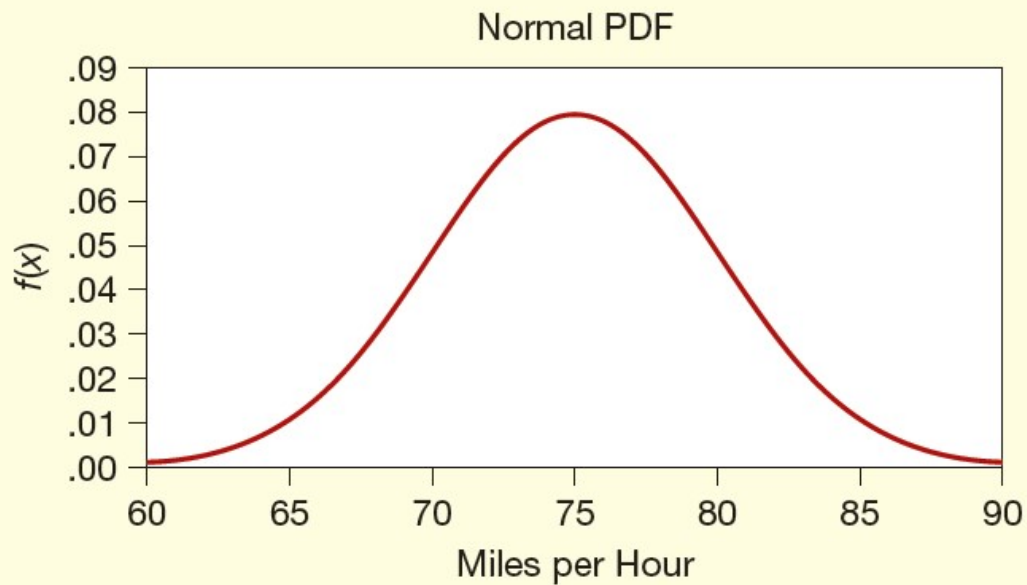
## 範例

某家商店平常週四上午九點到十點，大約會有1.7個客人來店購物

1. 請問“大約會有1.7個客人來店購物”怎麼計算來的？
2. 在同日同個時段，會有三個客人來購物的機率是多少？
3. 會有三個以上的客人來購物的機率是多少？
4. 請問期望值和標準差分別是多少？

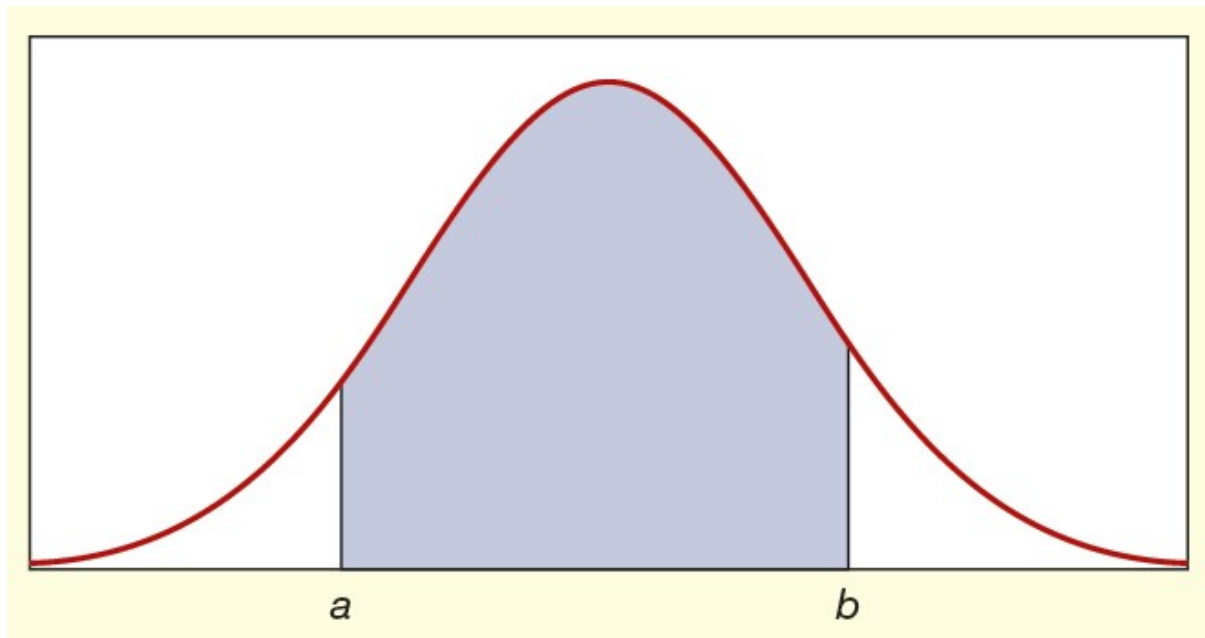
# 連續變項的機率分布

- PDF  $f(x)$  與 CDF  $F(x)$



# 機率 = PDF的面積

- 連續型的隨機變數介於 $a$ 和 $b$ 之間的機率  $P(a < X < b)$



# 期望值與變異數

Mean  $E(X) = \mu = \int_{-\infty}^{+\infty} x f(x) dx$

Variance  $\text{Var}(X) = \sigma^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx$

# 期望值與變異數

*Continuous Random Variable*

*Discrete Random Variable*

Mean

$$E(X) = \mu = \int_{-\infty}^{+\infty} x f(x) dx$$

$$E(X) = \mu = \sum_{\text{all } x} x P(x)$$

Variance

$$\text{Var}(X) = \sigma^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx$$

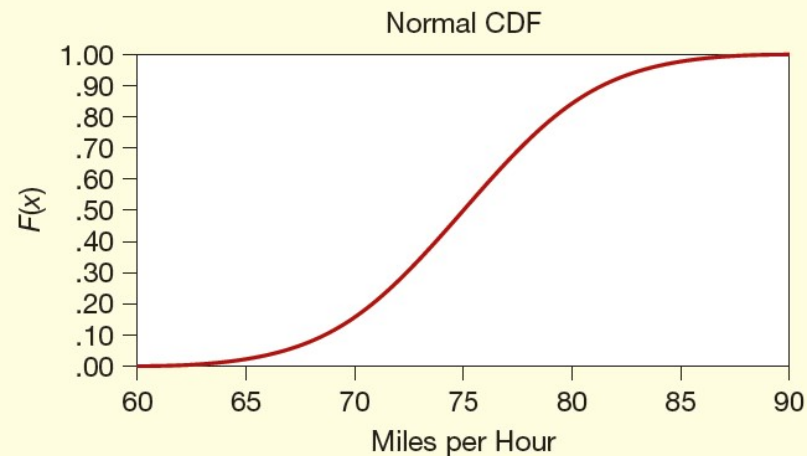
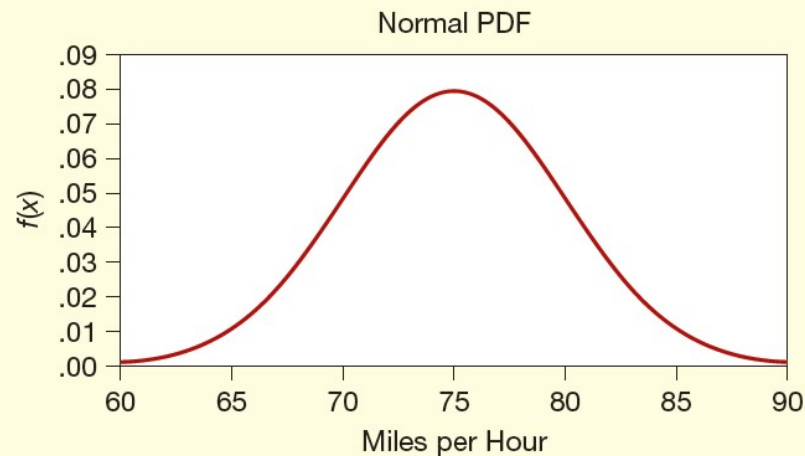
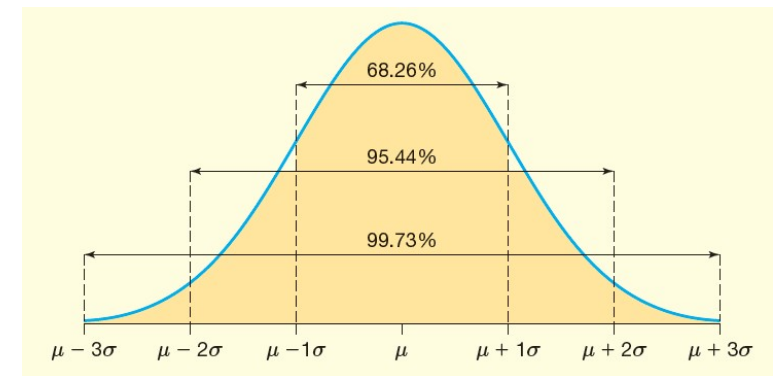
$$\text{Var}(X) = \sigma^2 = \sum_{\text{all } x} [x - \mu]^2 P(x)$$



# 常態分布 (Normal distribution)

- $N(\mu, \sigma^2)$
- $[\mu - 3\sigma, \mu + 3\sigma]$  包含幾乎所有數值

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



Parameters

$\mu$  = population mean

$\sigma$  = population standard deviation

PDF

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

Domain

$$-\infty < x < +\infty$$

Mean

$\mu$

Std. Dev.

$\sigma$

Shape

Symmetric, mesokurtic, and bell-shaped.

PDF in Excel\*

=NORM.DIST(x,μ,σ,0)

CDF in Excel\*

=NORM.DIST(x,μ,σ,1)

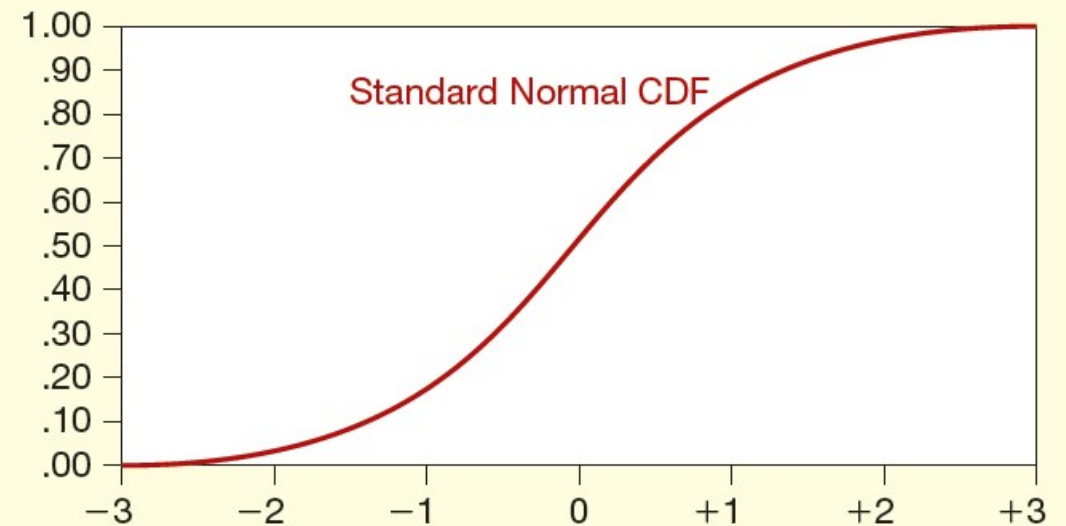
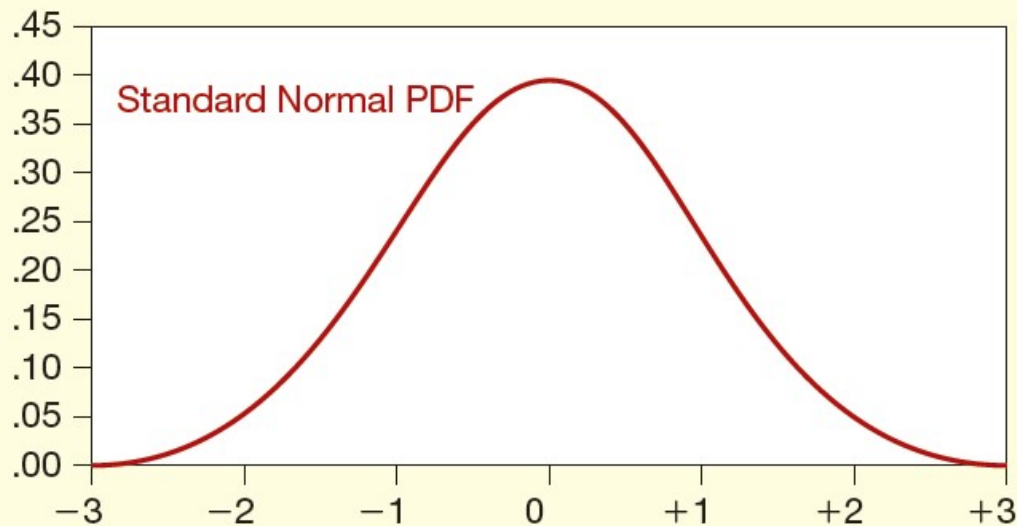
Random data in Excel

=NORM.INV(RAND(),μ,σ)

# 標準常態分布

- 把隨機變數 $x$ 標準化

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



Parameters

$\mu$  = population mean

$\sigma$  = population standard deviation

PDF

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \text{ where } z = \frac{x - \mu}{\sigma}$$

Domain

$$-\infty < z < +\infty$$

Mean

0

Standard deviation

1

Shape

Symmetric, mesokurtic, and bell-shaped.

CDF in Excel\*

=NORM.S.DIST(z,1)

Random data in Excel

=NORM.S.INV(RAND())

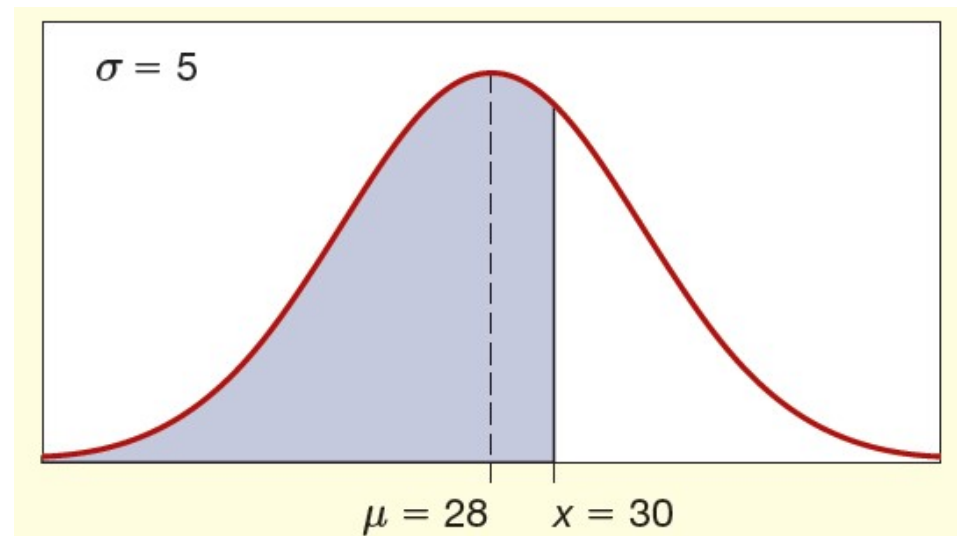
Comment

There is no simple formula for a normal CDF, so we need normal tables or Excel to find areas.

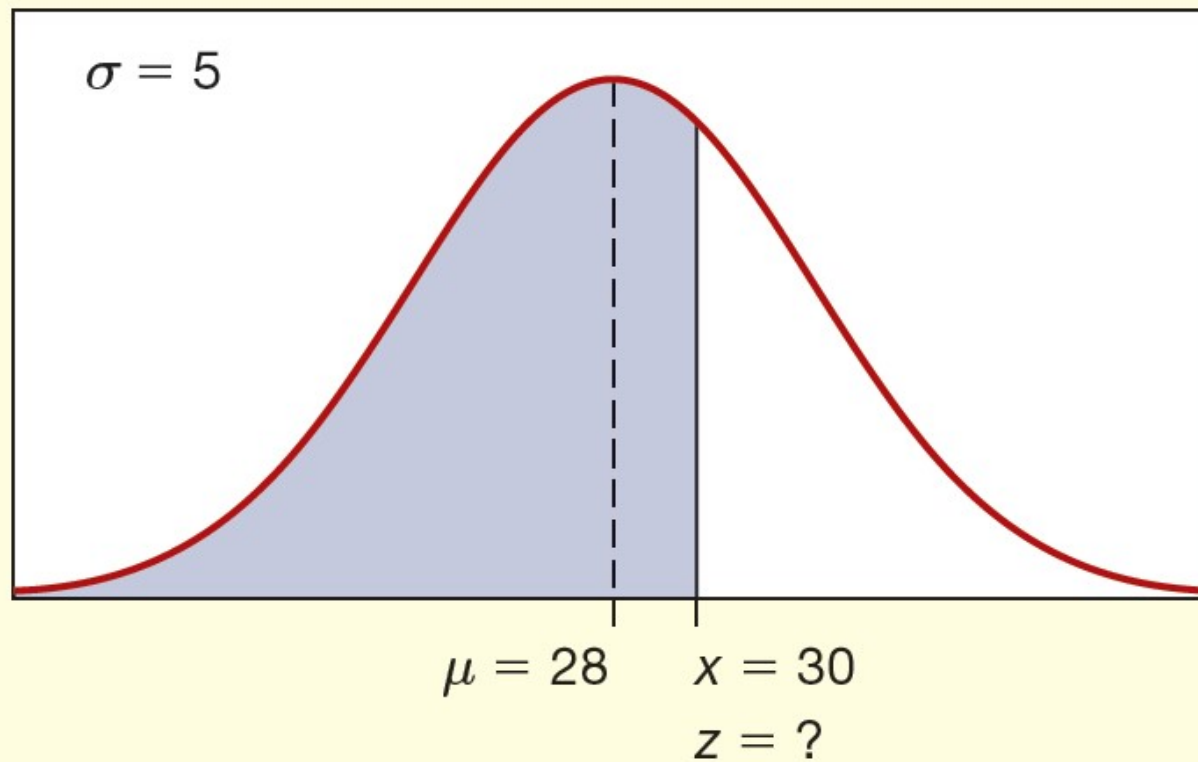
# 範例

某修車廠修車時間的PDF如右，平均時間是28分鐘，標準差5分鐘

1. 請問有多少比例的車，其修車時間會低於半小時？
2. 現在來了一台車，請問修這台車的時間超過40分鐘的機率是多少？
3. 老闆希望80%的車，修車時間不要超過半小時，請問平均修車時間必須是多少才能符合老闆的要求？



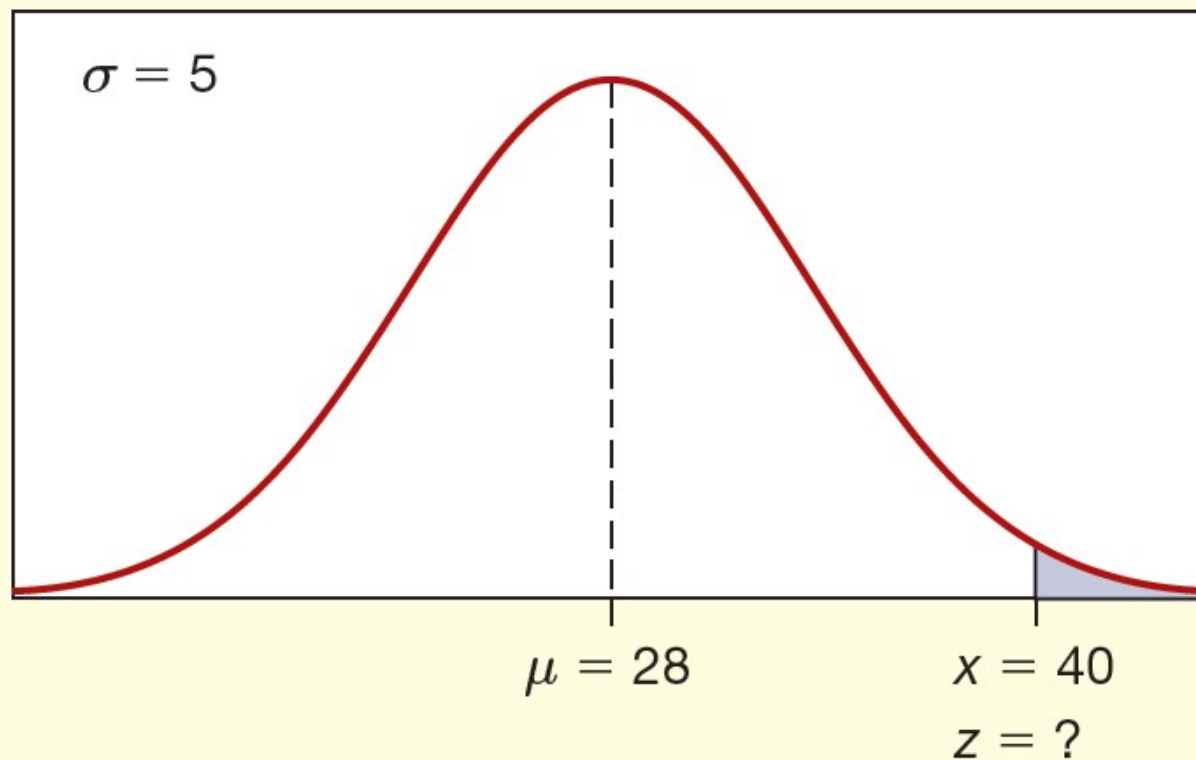
## 範例解答 1



Using Excel,  
=NORM.DIST(30,28,5,1)  
= .655422

$$z = \frac{30 - 28}{5} = 0.40$$

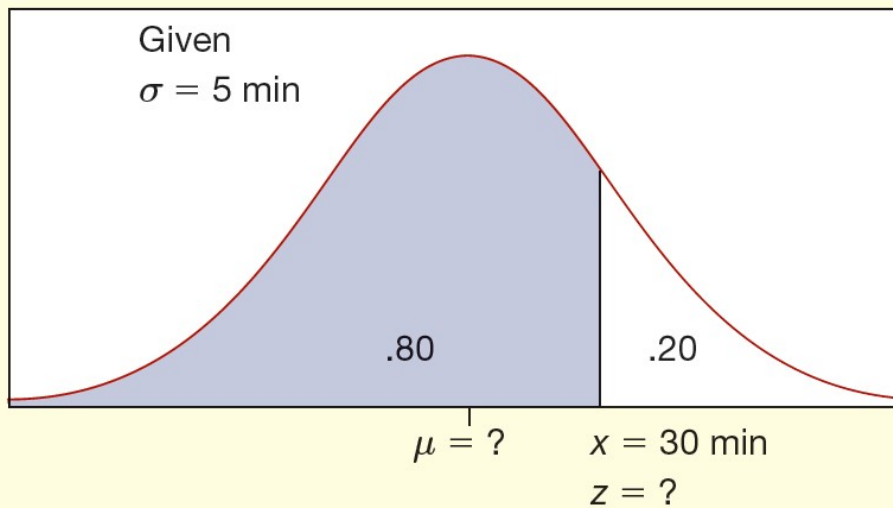
## 範例解答 2



Using Excel,  
 $= 1 - \text{NORM.DIST}(40, 28, 5, 1)$   
 $= .008198$

$$z = \frac{40 - 28}{5} = 2.4$$

## 範例解答 3



Using Excel,  
=NORM.S.INV(.80)  
=.841621

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

$$0.84 = \frac{30 - \mu}{5}$$

$$\mu = 30 - 0.84(5) = 25.8$$



# 課後作業

- 請具體寫出一個今天學習到的統計概念 (字數不限)