巨量資料管理學院碩士在職專班

統計分析

2022/10/21

陳光宏

機率分布

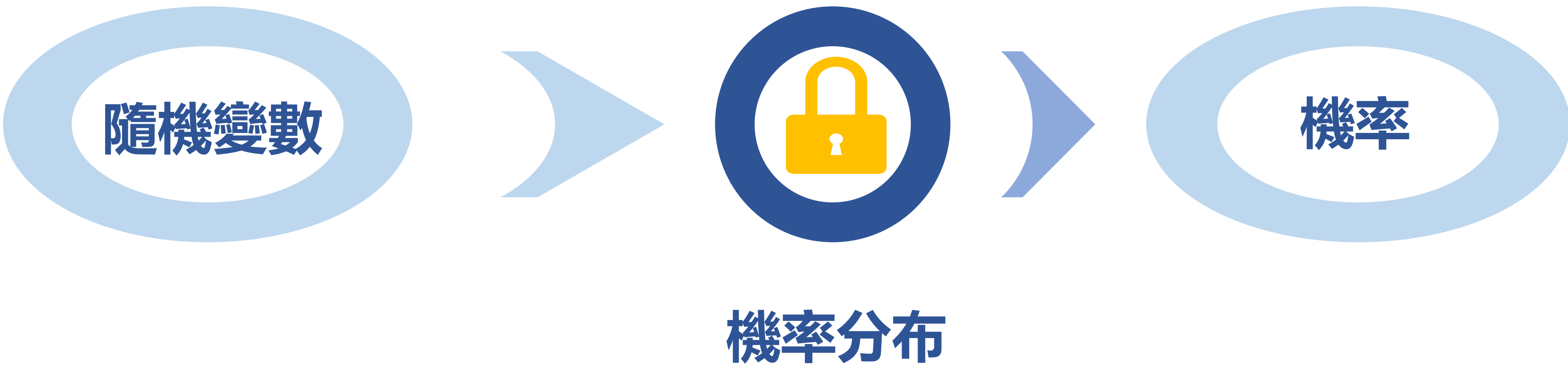
隨機變數 (Random variables)

- 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



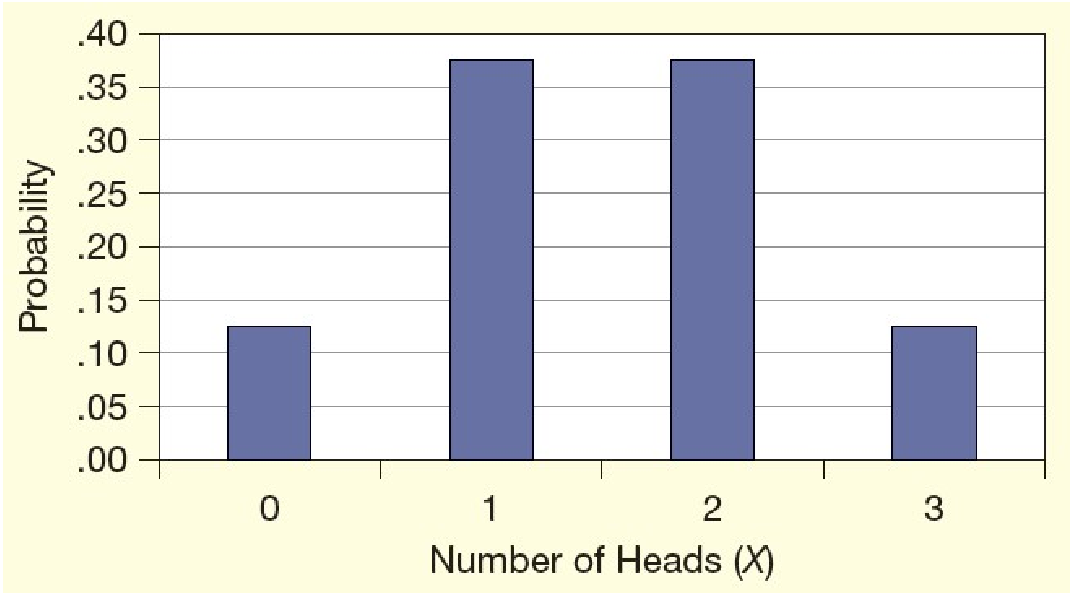
練習1

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？

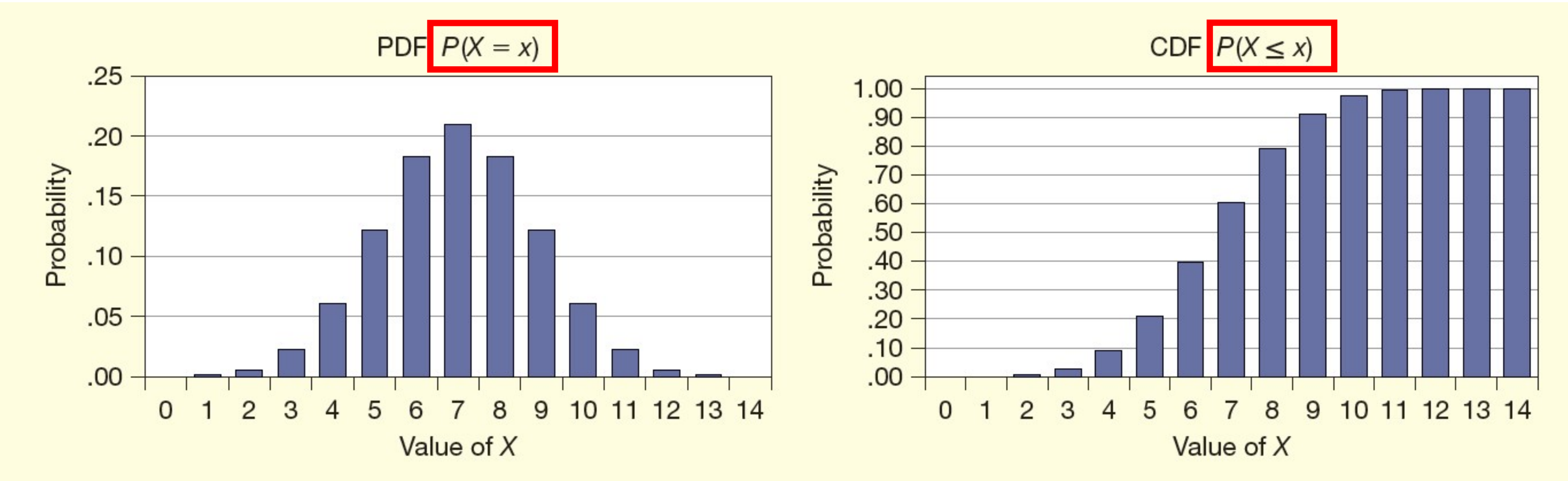
可能的情況		機率
1	---	
2	---	
3	---	
---	---	

Possible Events	x	P(x)
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)



期望值與變異數

- 期望值 (Expected value)
  - 加權平均的概念

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

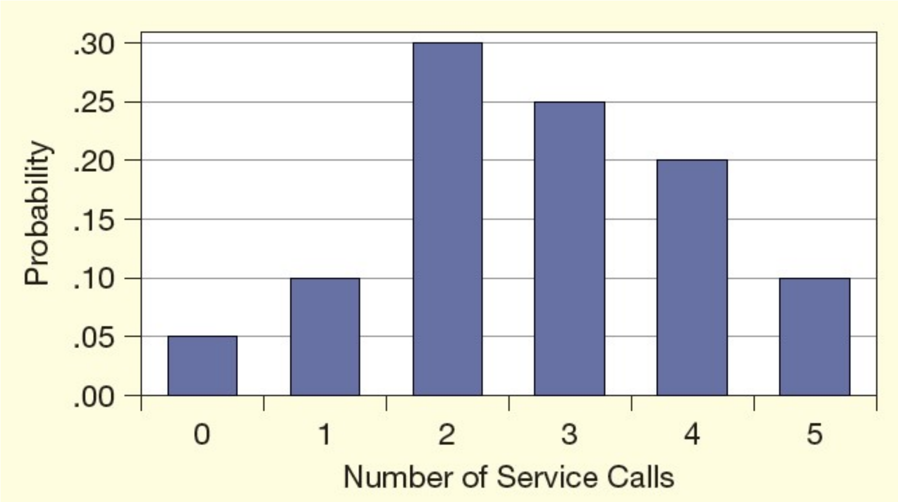
- 變異數 (Variance)
  - 離平均有多遠

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

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	$P(x)$	$xP(x)$	$x-\mu$	$[x-\mu]^2$	$[x-\mu]^2P(x)$
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)

Bernoulli Experiment	*Possible Outcomes*	*Probability of "Success"*
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

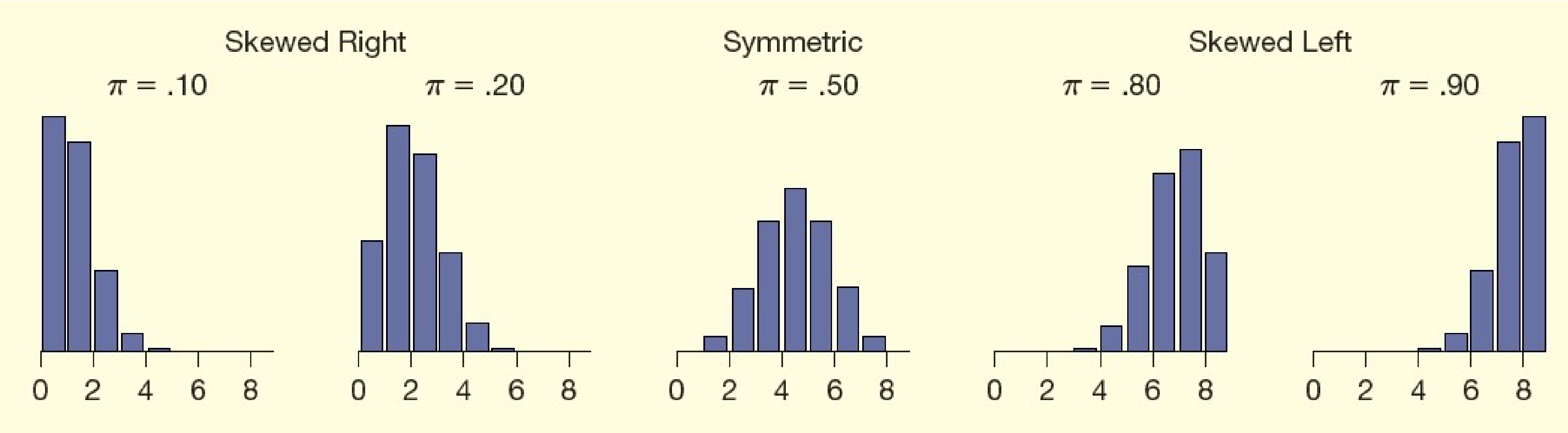
q = 1 – p

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

Parameters	n = number of trials π = probability of success
PDF	$P(X = x) = \dfrac{n!}{x!(n - x)!}\pi^x(1 - \pi)^{n-x}$
Excel* PDF	=BINOM.DIST(x, n, π, 0)
Excel* CDF	=BINOM.DIST(x, n, π, 1)
Domain	$x = 0, 1, 2, \ldots, n$
Mean	$n\pi$
Standard deviation	$\sqrt{n\pi(1 - \pi)}$
Random data generation in Excel	=BINOM.INV(n, π, 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



範例3

到某醫院急診的病人中，大約有20%沒有額外買保險，現在觀察5位病人

1. 請問隨機變數是什麼？
2. 是否符合二項式分布？
3. 使用什麼參數 (Parameter)？
4. 請問這5位病人中，有3位沒有買保險的機率是多少？

判斷是否為二項式分布

1. 試驗有兩種結果：是/否，成功/失敗，...
2. 有 n 次相同的試驗 (experiments/trials)
3. 每次試驗成功的機率為 p
4. 每次試驗都是獨立的
5. 我們有興趣的隨機變數為”n次試驗中，成功的次數”

範例3解法

方法一

代公式

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

方法二

查表

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

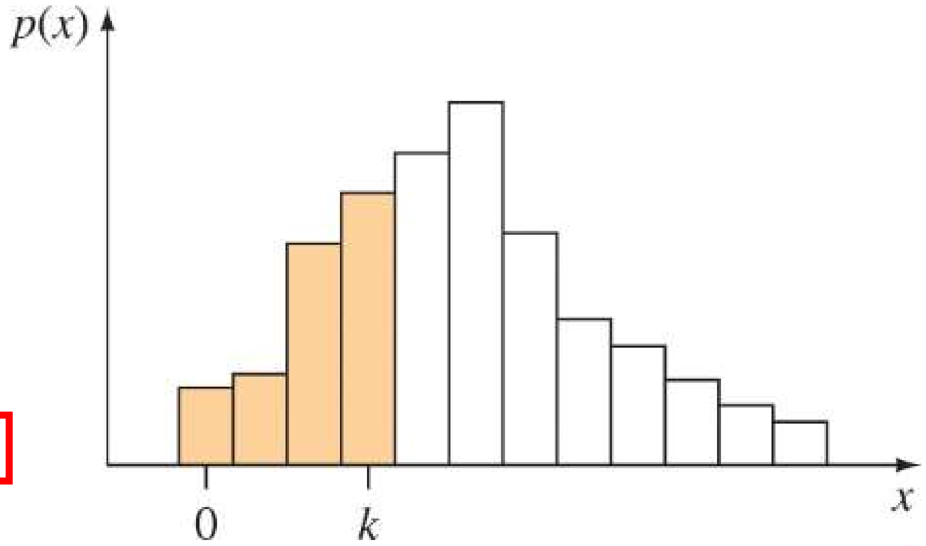
n	*k*	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
	18	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	19	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
20	0	.3585	.1216	.0388	.0115	.0032	.0008	.0002	.0000	.0000	.0000
	1	.3774	.2702	.1368	.0576	.0211	.0068	.0020	.0005	.0001	.0000
	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

範例3解答 – 查表



■ Table 1 **Cumulative Binomial Probabilities**

Tabulated values are $P(x \leq k) = p(0) + p(1) + \cdots + p(k)$.

(Computations are rounded to the third decimal place.)

$n = 2$

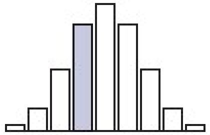
	p													
k	**.01**	**.05**	**.10**	**.20**	**.30**	**.40**	**.50**	**.60**	**.70**	**.80**	**.90**	**.95**	**.99**	**k**
0	.980	.902	.810	.640	.490	.360	.250	.160	.090	.040	.010	.002	.000	0
1	1.000	.998	.990	.960	.910	.840	.750	.640	.510	.360	.190	.098	.020	1
2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	2

n = 5

	p													
k	.01	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95	.99	*k*
0	.951	.774	.590	.328	.168	.078	.031	.010	.002	.000	.000	.000	.000	0
1	.999	.977	.919	.737	.528	.337	.188	.087	.031	.007	.000	.000	.000	1
2	1.000	.999	.991	.942	.837	.683	.500	.317	.163	.058	.009	.001	.000	2
3	1.000	1.000	1.000	.993	.969	.913	.812	.663	.472	.263	.081	.023	.001	3
4	1.000	1.000	1.000	1.000	.998	.990	.969	.922	.832	.672	.410	.226	.049	4
5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	5

BINOMIAL PROBABILITIES

Example: $P(X = 3 \mid n = 8, \pi = .50) = .2188$



This table shows $P(X = x)$.

		π																
n	x	.01	.02	.05	.10	.15	.20	.30	.40	.50	.60	.70	.80	.85	.90	.95	.98	.99
2	0	.9801	.9604	.9025	.8100	.7225	.6400	.4900	.3600	.2500	.1600	.0900	.0400	.0225	.0100	.0025	.0004	.0001
	1	.0198	.0392	.0950	.1800	.2550	.3200	.4200	.4800	.5000	.4800	.4200	.3200	.2550	.1800	.0950	.0392	.0198
	2	.0001	.0004	.0025	.0100	.0225	.0400	.0900	.1600	.2500	.3600	.4900	.6400	.7225	.8100	.9025	.9604	.9801
3	0	.9703	.9412	.8574	.7290	.6141	.5120	.3430	.2160	.1250	.0640	.0270	.0080	.0034	.0010	.0001	—	—
	1	.0294	.0576	.1354	.2430	.3251	.3840	.4410	.4320	.3750	.2880	.1890	.0960	.0574	.0270	.0071	.0012	.0003
	2	.0003	.0012	.0071	.0270	.0574	.0960	.1890	.2880	.3750	.4320	.4410	.3840	.3251	.2430	.1354	.0576	.0294
	3	—	—	.0001	.0010	.0034	.0080	.0270	.0640	.1250	.2160	.3430	.5120	.6141	.7290	.8574	.9412	.9703
4	0	.9606	.9224	.8145	.6561	.5220	.4096	.2401	.1296	.0625	.0256	.0081	.0016	.0005	.0001	—	—	—
	1	.0388	.0753	.1715	.2916	.3685	.4096	.4116	.3456	.2500	.1536	.0756	.0256	.0115	.0036	.0005	—	—
	2	.0006	.0023	.0135	.0486	.0975	.1536	.2646	.3456	.3750	.3456	.2646	.1536	.0975	.0486	.0135	.0023	.0006
	3	—	—	.0005	.0036	.0115	.0256	.0756	.1536	.2500	.3456	.4116	.4096	.3685	.2916	.1715	.0753	.0388
	4	—	—	—	.0001	.0005	.0016	.0081	.0256	.0625	.1296	.2401	.4096	.5220	.6561	.8145	.9224	.9606
5	0	.9510	.9039	.7738	.5905	.4437	.3277	.1681	.0778	.0313	.0102	.0024	.0003	.0001	—	—	—	—
	1	.0480	.0922	.2036	.3281	.3915	.4096	.3602	.2592	.1563	.0768	.0284	.0064	.0022	.0005	—	—	—
	2	.0010	.0038	.0214	.0729	.1382	.2048	.3087	.3456	.3125	.2304	.1323	.0512	.0244	.0081	.0011	.0001	—
	3	—	.0001	.0011	.0081	.0244	.0512	.1323	.2304	.3125	.3456	.3087	.2048	.1382	.0729	.0214	.0038	.0010
	4	—	—	—	.0005	.0022	.0064	.0284	.0768	.1563	.2592	.3602	.4096	.3915	.3281	.2036	.0922	.0480
	5	—	—	—	—	.0001	.0003	.0024	.0102	.0313	.0778	.1681	.3277	.4437	.5905	.7738	.9039	.9510

TABLE 1 Exact binomial probabilities $Pr(X = k) = \binom{n}{k} p^k q^{n-k}$

n	k	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
2	0	.9025	.8100	.7225	.6400	.5625	.4900	.4225	.3600	.3025	.2500
	1	.0950	.1800	.2550	.3200	.3750	.4200	.4550	.4800	.4950	.5000
	2	.0025	.0100	.0225	.0400	.0625	.0900	.1225	.1600	.2025	.2500
3	0	.8574	.7290	.6141	.5120	.4219	.3430	.2746	.2160	.1664	.1250
	1	.1354	.2430	.3251	.3840	.4219	.4410	.4436	.4320	.4084	.3750
	2	.0071	.0270	.0574	.0960	.1406	.1890	.2389	.2880	.3341	.3750
	3	.0001	.0010	.0034	.0080	.0156	.0270	.0429	.0640	.0911	.1250
4	0	.8145	.6561	.5220	.4096	.3164	.2401	.1785	.1296	.0915	.0625
	1	.1715	.2916	.3685	.4096	.4219	.4116	.3845	.3456	.2995	.2500
	2	.0135	.0486	.0975	.1536	.2109	.2646	.3105	.3456	.3675	.3750
	3	.0005	.0036	.0115	.0256	.0469	.0756	.1115	.1536	.2005	.2500
	4	.0000	.0001	.0005	.0016	.0039	.0081	.0150	.0256	.0410	.0625
5	0	.7738	.5905	.4437	.3277	.2373	.1681	.1160	.0778	.0503	.0313
	1	.2036	.3280	.3915	.4096	.3955	.3602	.3124	.2592	.2059	.1563
	2	.0214	.0729	.1382	.2048	.2637	.3087	.3364	.3456	.3369	.3125
	3	.0011	.0081	.0244	.0512	.0879	.1323	.1811	.2304	.2757	.3125
	4	.0000	.0004	.0022	.0064	.0146	.0283	.0488	.0768	.1128	.1563
	5	.0000	.0000	.0001	.0003	.0010	.0024	.0053	.0102	.0185	.0313

範例3解答 – 利用excel計算

Excel* PDF	=BINOM.DIST(x, n, π, 0)
Excel* CDF	=BINOM.DIST(x, n, π, 1)

範例4

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？

Mean	$n\pi$
Standard deviation	$\sqrt{n\pi(1-\pi)}$

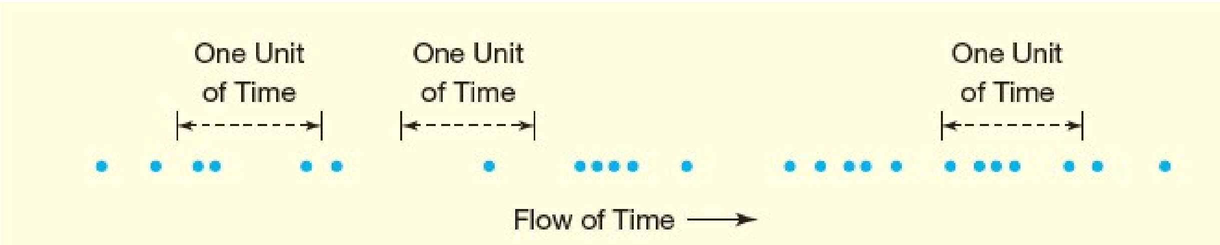
練習2

假設發生肺炎的機率是1%，現觀察1000位病人

1. 請問有10位發生肺炎的機率是多少？
2. 期望會有多少人發生肺炎？

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

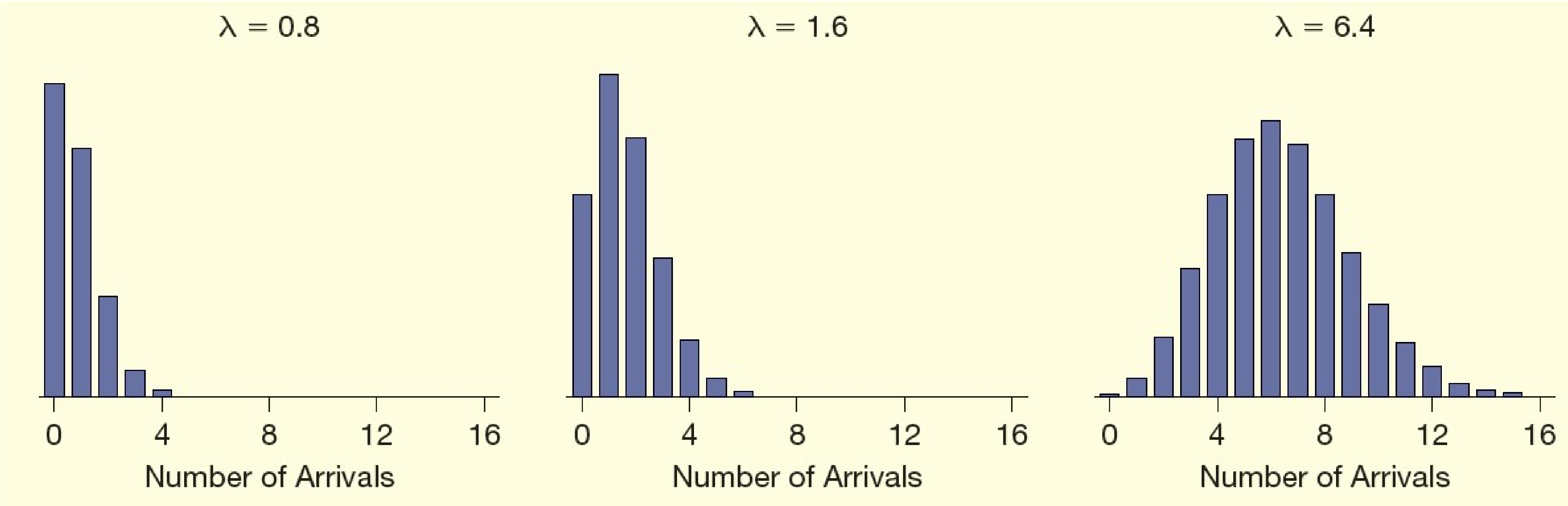
λ = 期望發生個數或次數

e = 2.71828.....

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

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

卜瓦松分布的PDF



λ = 0.8
λ = 1.6
λ = 6.4
0
4
8
12
16
Number of Arrivals
Number of Arrivals
Number of Arrivals

範例4

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？

範例4解答

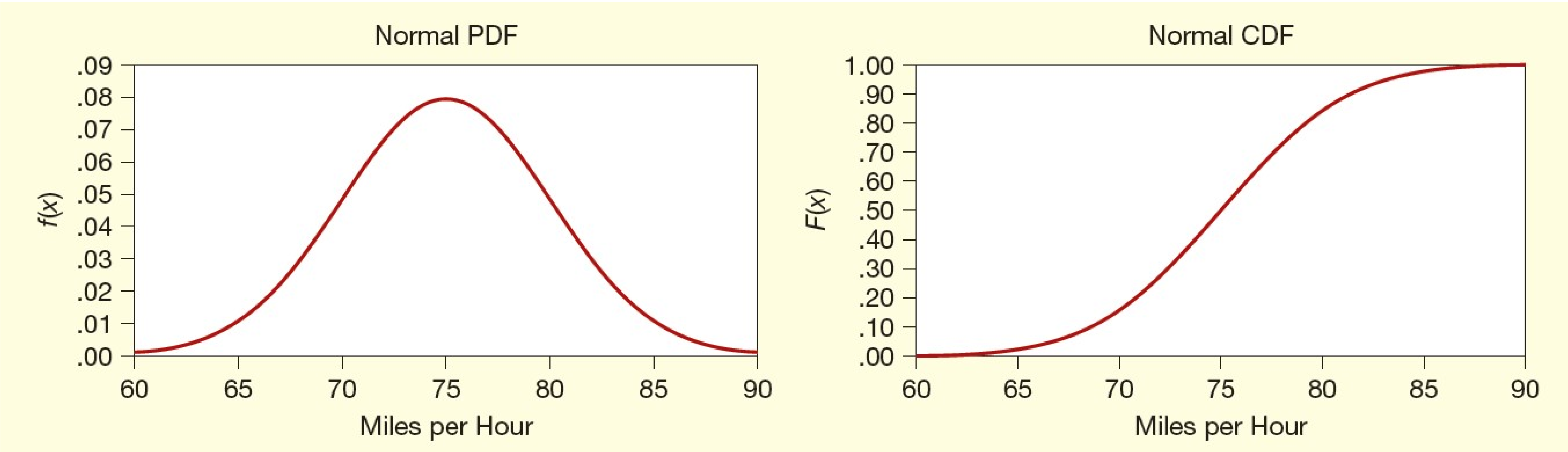
x	$P(X = x)$	$P(X \le x)$
0	.1827	.1827
1	.3106	.4932
2	.2640	.7572
3	.1496	.9068
4	.0636	.9704
5	.0216	.9920
6	.0061	.9981
7	.0015	.9996
8	.0003	.9999
9	.0001	1.0000

Mean: $\lambda = 1.7$

Standard deviation: $\sigma = \sqrt{\lambda} = \sqrt{1.7} = 1.304$

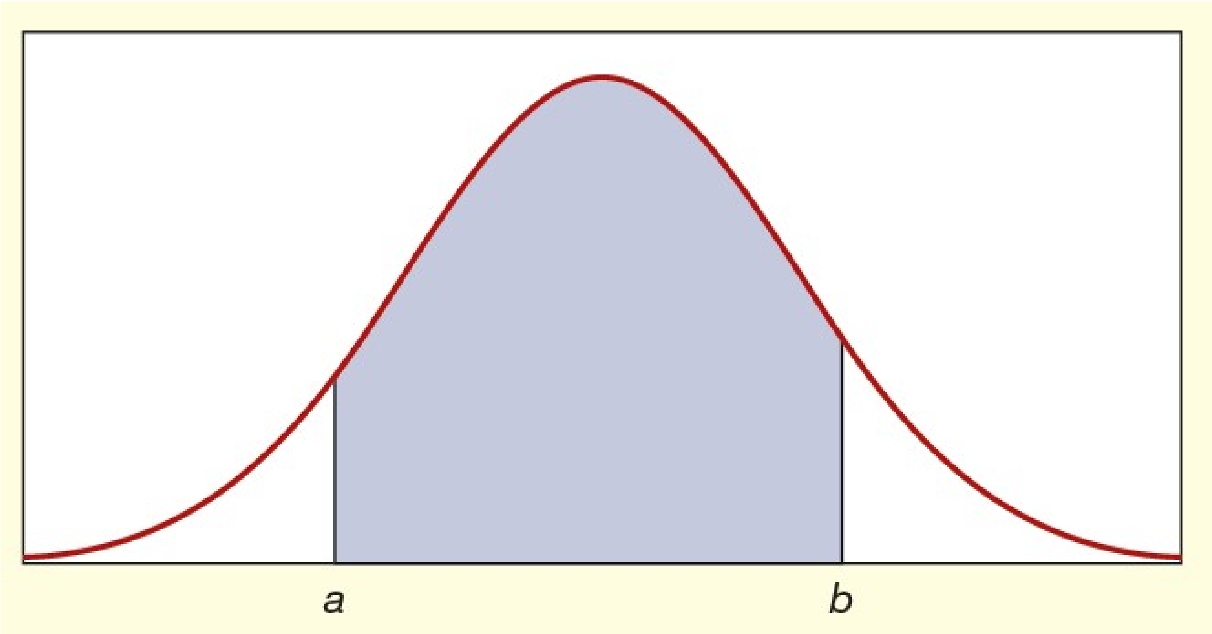
連續變項的機率分布

- 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

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

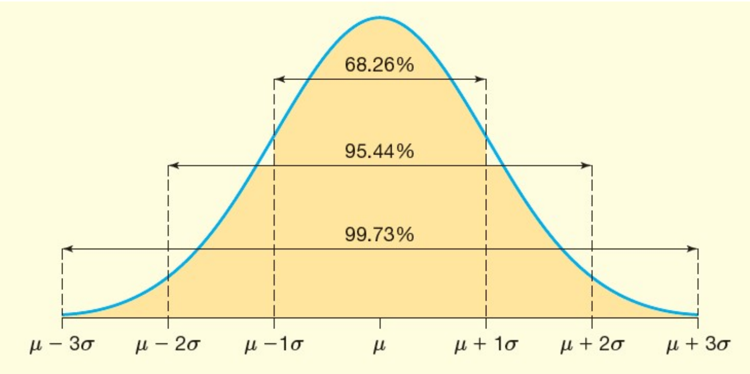
$$\text{Variance} \qquad \text{Var}(X) = \sigma^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x)\, dx$$

期望值與變異數

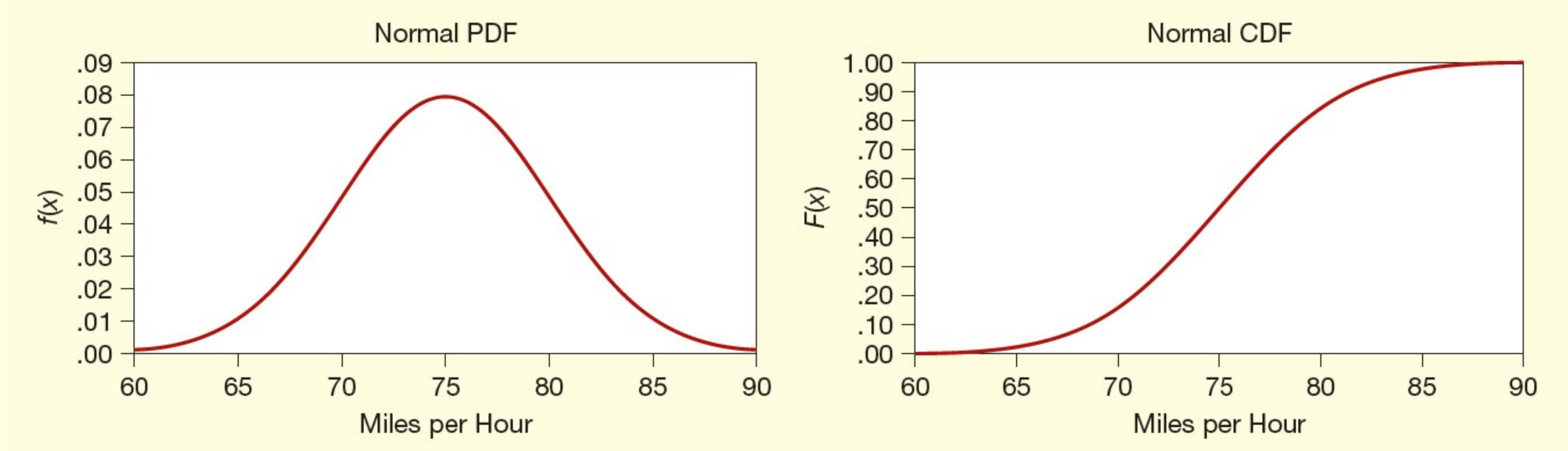
	Continuous Random Variable	*Discrete Random Variable*
Mean	$E(X) = \mu = \int_{-\infty}^{+\infty} x f(x)\, dx$	$E(X) = \mu = \sum_{\text{all } x} x P(x)$
Variance	$\mathrm{Var}(X) = \sigma^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x)\, dx$	$\mathrm{Var}(X) = \sigma^2 = \sum_{\text{all } x} [x - \mu]^2 P(x)$

常態分布 (Normal distribution)

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



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

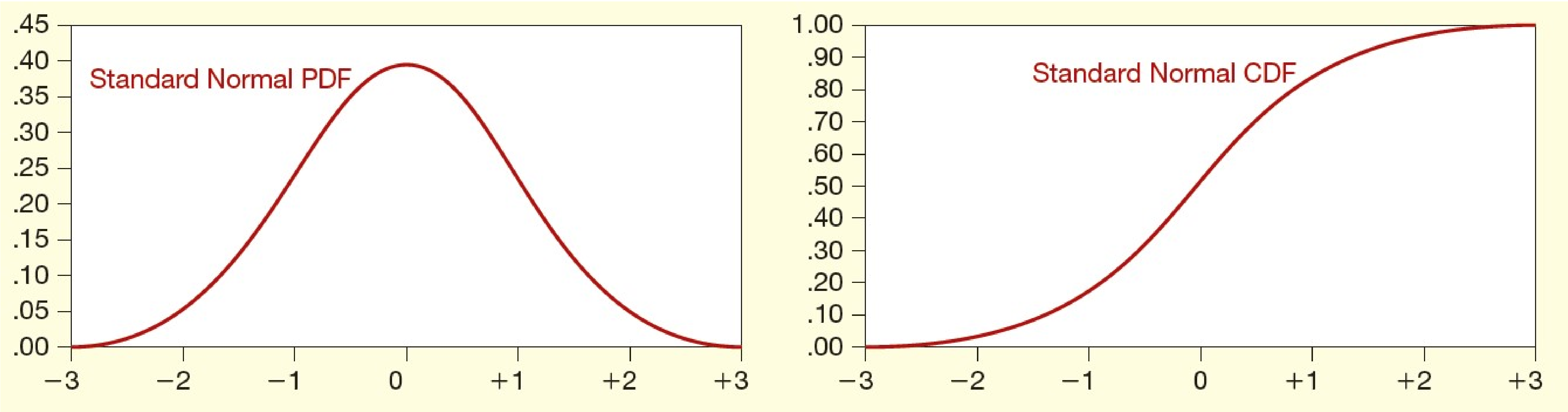


Parameters	μ = population mean σ = population standard deviation
PDF	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
Domain	$-\infty < x < +\infty$
Mean	μ
Std. Dev.	σ
Shape	Symmetric, mesokurtic, and bell-shaped.
PDF in Excel*	=NORM.DIST(x,μ,σ,0)
CDF in Excel*	=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

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

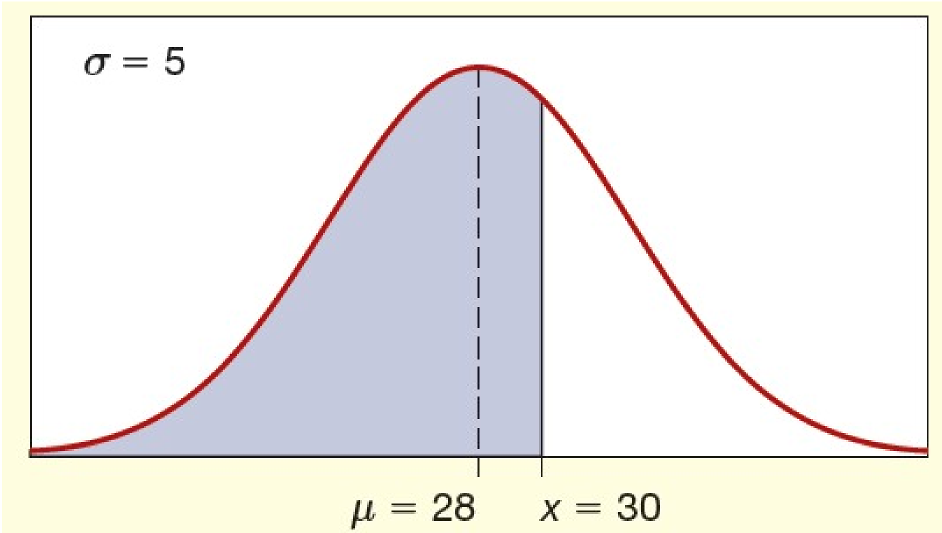


Parameters	μ = population mean σ = population standard deviation
PDF	$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ where $z = \frac{x-\mu}{\sigma}$
Domain	$-\infty < z < +\infty$
Mean	0
Standard deviation	1
Shape	Symmetric, mesokurtic, and bell-shaped.
CDF in Excel*	=NORM.S.DIST(z,1)
Random data in Excel	=NORM.S.INV(RAND())
Comment	There is no simple formula for a normal CDF, so we need normal tables or Excel to find areas.

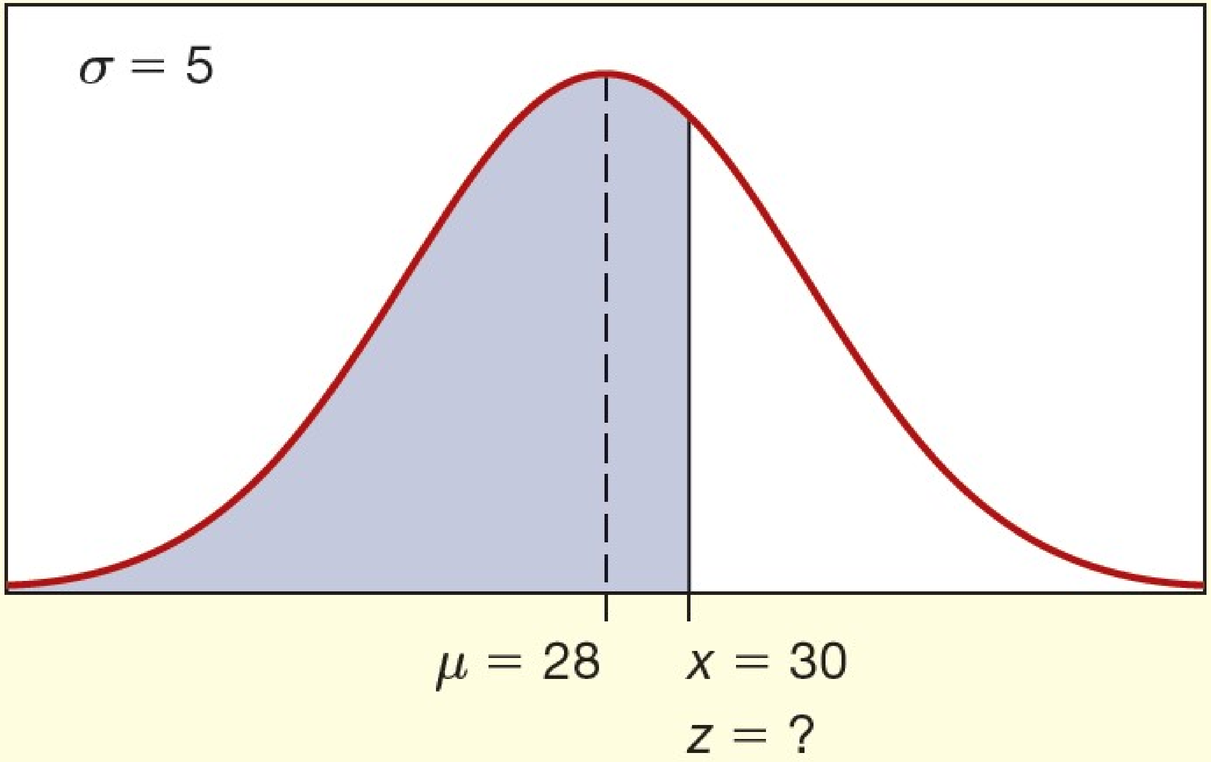
範例5

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？



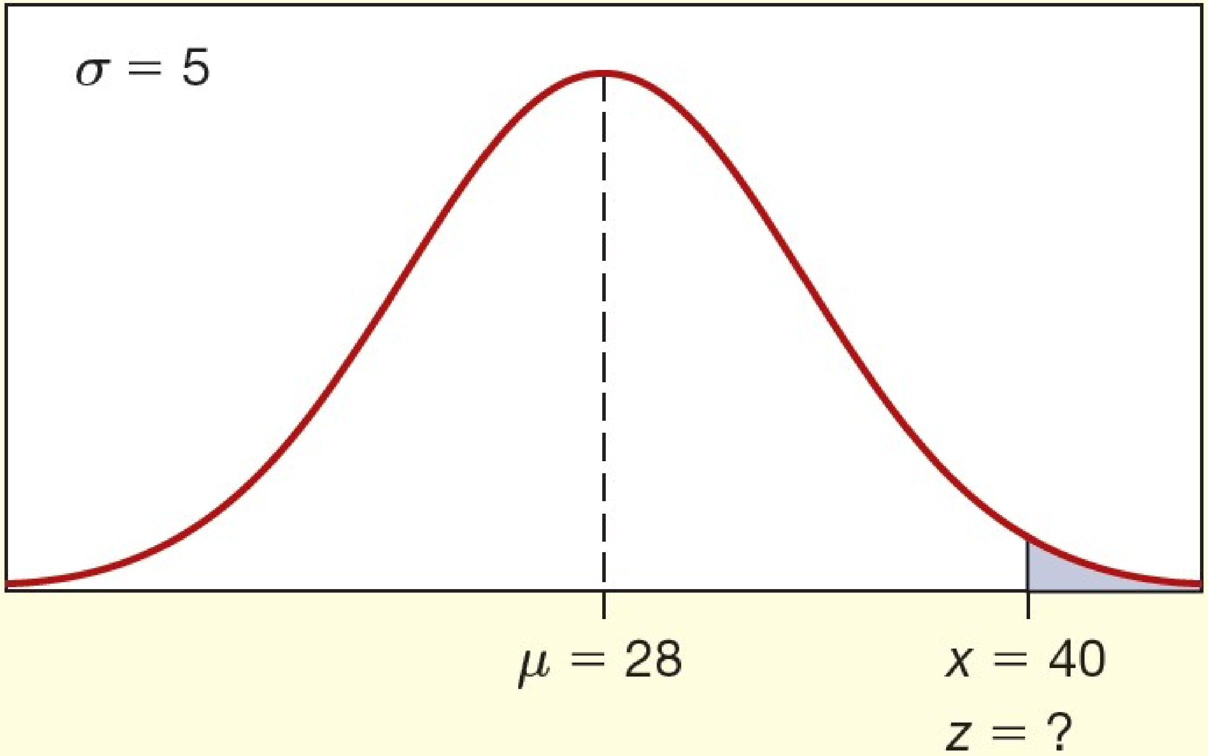
範例5解答 (1)



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

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

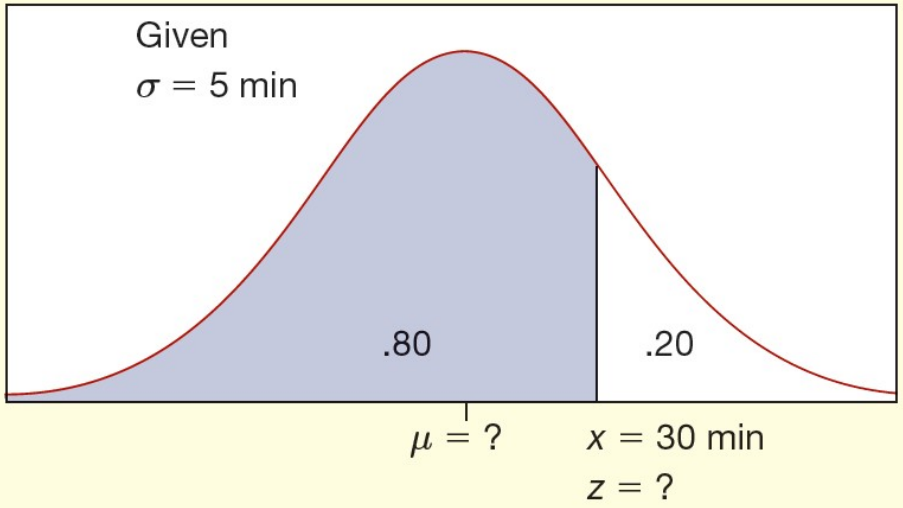
範例5解答 (2)



Using Excel,
= 1 − NORM.DIST(40,28,5,1)
= .008198

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

範例5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)