巨量資料管理學院碩士在職專班

統計分析

2022/9/30

陳光宏

機率概論

入門

- 樣本空間 (Sample space)

書籍與音樂類
S = {Books, DVD, VHS, Magazines, Newspapers, Music, Textbooks}

- 事件 (Event)

電子媒體類　　A = {Music, DVD, VHS}
紙本期刊類　　B = {Newspapers, Magazines}

機率的定義

- A事件的機率：**銷售電子媒體類的書籍或音樂的機率/比例**

$$P(A) = A/S$$

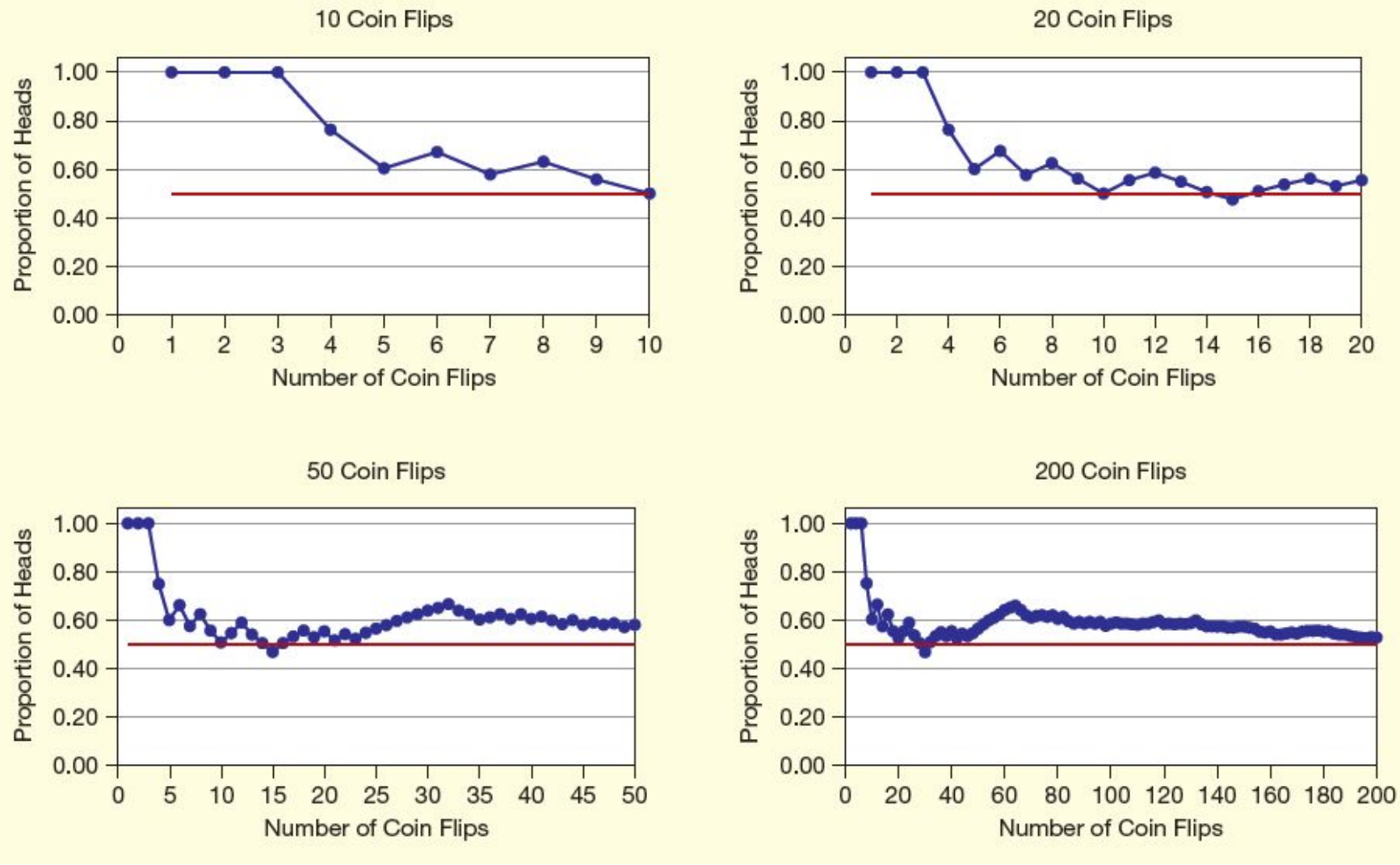
- 任何一件事情發生的機率介於0到1

$$0 \leq P(A) \leq 1$$

- 所有可能發生的事件，機率總和是1

$$P(S) = P(E_1) + P(E_2) + \cdots + P(E_n) = 1$$

大數法則 (Law of large numbers)

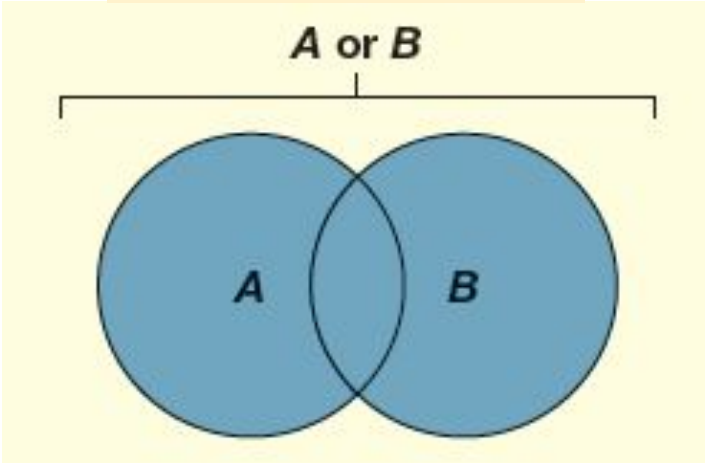


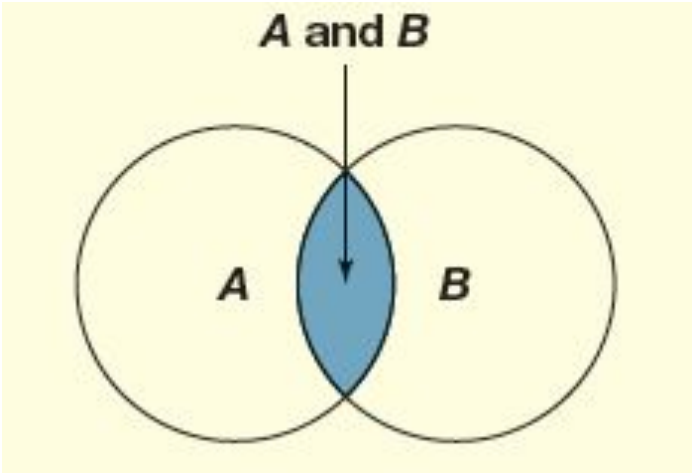
名詞與概念

- 聯集：發生A事件或B事件
  - $A \cup B$
- 交集：發生A事件且B事件
  - $A \cap B$

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

Venn diagram





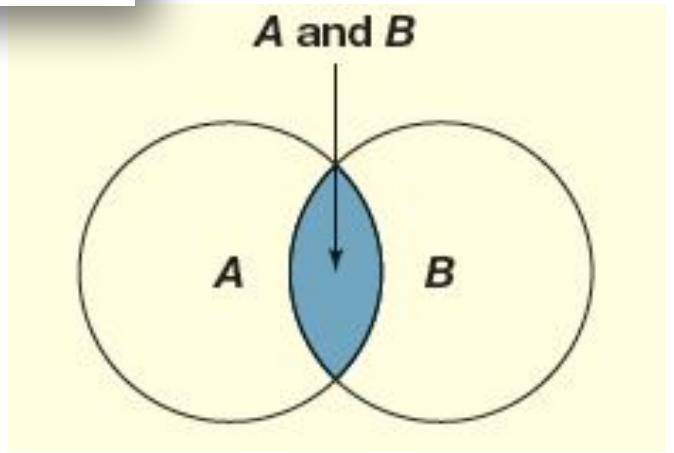
範例1

用60秒思考一下

請問一副撲克牌中，抽中紅色或是Queen的機率是多少？

$$\begin{aligned}
\text{Queen: } P(Q) &= 4/52 && \text{(there are 4 queens in a deck)}\\
\text{Red: } P(R) &= 26/52 && \text{(there are 26 red cards in a deck)}\\
\text{Queen and Red: } P(Q \cap \mathrm{R}) &= 2/52 && \text{(there are 2 red queens in a deck)}
\end{aligned}$$

$$\begin{aligned}
P(Q \cup R) &= P(Q) + P(R) - P(Q \cap R)\\
&= 4/52 + 26/52 - 2/52\\
&= 28/52 = .5385, \text{ or a } 53.85\% \text{ chance}
\end{aligned}$$

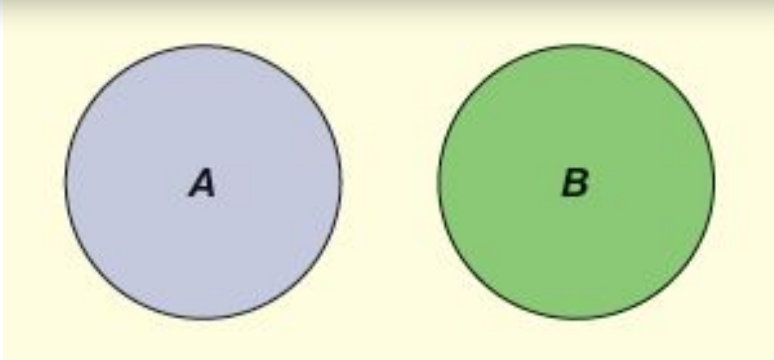


互斥事件 (Mutually exclusive events)

Customer age: A = under 21, B = over 65

Purebred dog breed: A = border collie, B = golden retriever

Business form: A = corporation, B = sole proprietorship



$$P(A \cup B) = P(A) + P(B)$$

Collective exhaustive events

書籍與音樂類
S = {Books, DVD, VHS, Magazines, Newspapers, Music, Textbooks}

電子媒體類　　A = {Music, DVD, VHS}
紙本期刊類　　B = {Newspapers, Magazines}
其他　　C = {Books, Textbooks}

當所有事件的聯集就是整個樣本空間時，則稱為
Collective exhaustive events

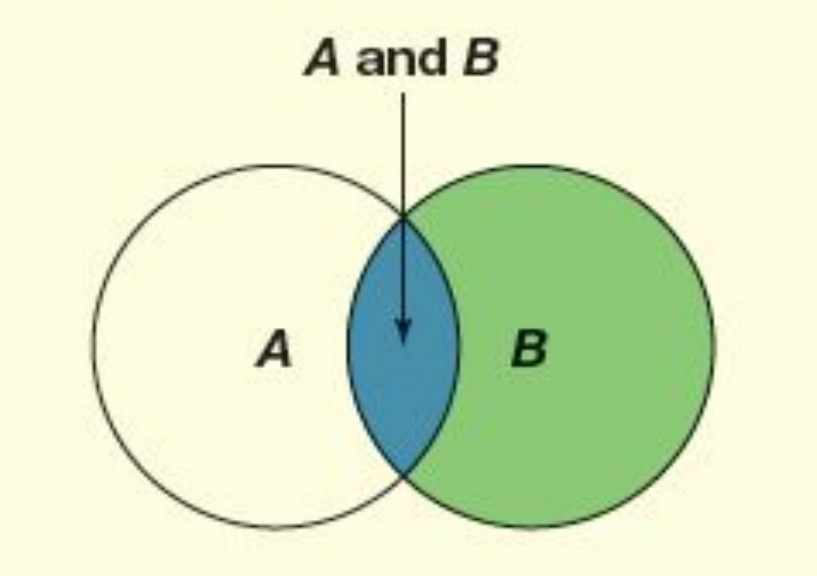
練習1

請舉一個例子，可以說明

互斥且collectively exhaustive

條件機率 (Conditional probability)

在事件B發生的情況下，A事件發生的機率 (事件B變成樣本空間)



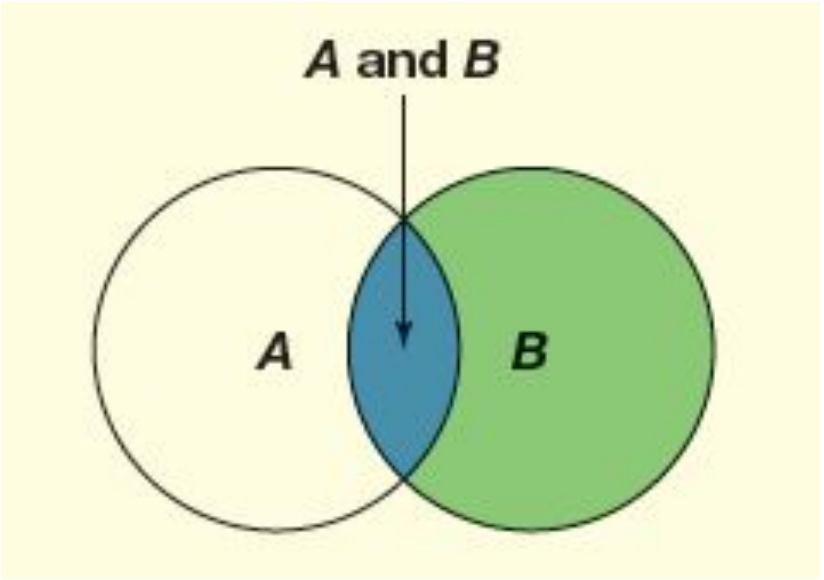
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad \text{for } P(B) > 0$$

$$P(A \cap B) = P(A \mid B)P(B)$$

範例2

在一個16-21歲未就讀大學的樣本中 (n=10,000)，13.50%未就業, 29.05%是高中輟學, 5.32%是高中輟學且未就業, 請問高中輟學的人當中, 有多少比例是未就業的？

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad \text{for } P(B) > 0$$



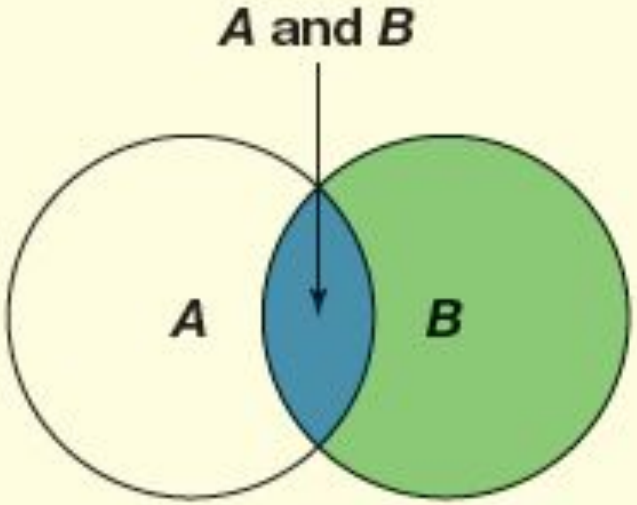
範例2 解答

U = the event that the person is unemployed

D = the event that the person is a high school dropout

$P(U) = .1350 \qquad P(D) = .2905 \qquad P(U \cap D) = .0532$

$$P(U \mid D) = \frac{P(U \cap D)}{P(D)} = \frac{.0532}{.2905} = .1831, \text{ or } 18.31\%$$



獨立事件 (Independent events)

A事件發生與B事件沒有關係

兩個事件的相關性

$$P(A \mid B) = P(A).$$

$$P(A \cap B) = P(A)P(B)$$

探討兩件事的相關性

一般情況下

$$P(A \cap B) = P(A \mid B)P(B)$$

當兩件事互相獨立時

$$P(A \cap B) = P(A)P(B)$$

如果兩個數值相等，表示兩事件沒有相關，是獨立事件，即A事件的發生不會影響B事件

範例3

在一個16-21歲未就讀大學的樣本中 (n=10,000)，13.50%未就業，29.05%是高中輟學，5.32%是高中輟學且未就業，請問高中輟學和是否就業是否相關？

範例3 解答

方法一

$P(U|D) = P(U)$

方法二

$P(U \cap D) = P(U) \times P(D)$

U = the event that the person is unemployed

D = the event that the person is a high school dropout

$P(U) = .1350 \qquad P(D) = .2905 \qquad P(U \cap D) = .0532$

$$P(U \mid D) = \frac{P(U \cap D)}{P(D)} = \frac{.0532}{.2905} = .1831, \text{ or } 18.31\%$$

列聯表 (Contingence tables)

- 類別變項

- 交叉表 (cross-table)
- 樞紐分析表

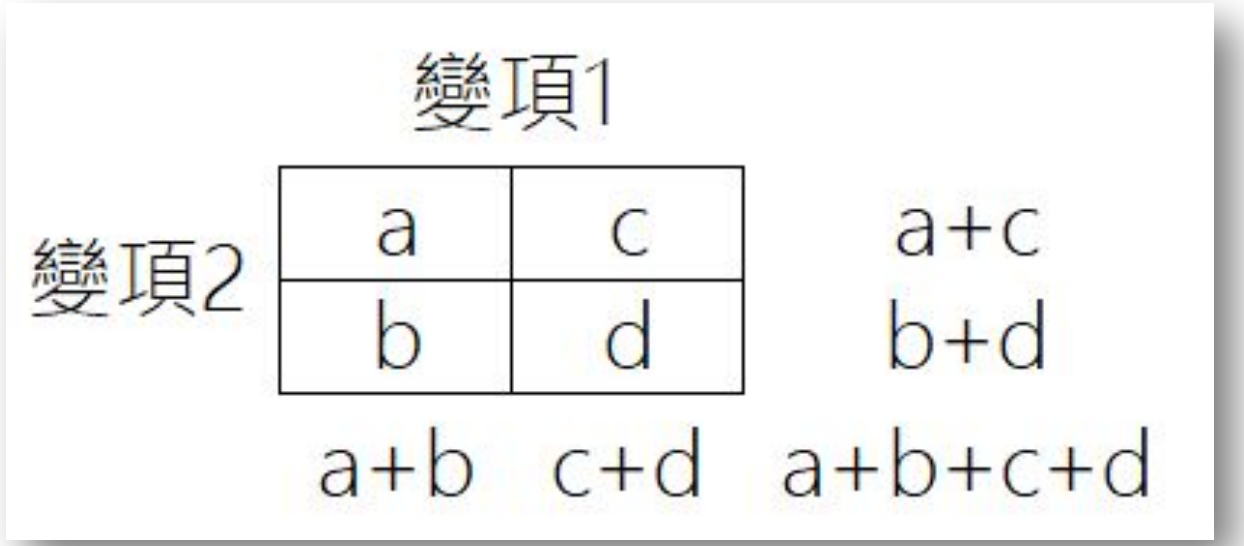
	Dog	Cat	Total
Male	42	10	52
Female	9	39	48
Total	51	49	100

educational level	*smoking status*			
	never smoked	currently smoke	former smoker	totals
did not finish high school	25	40	30	95
high school graduate	30	30	40	100
BS degree	50	10	60	120
totals	105	80	130	315

挑戰

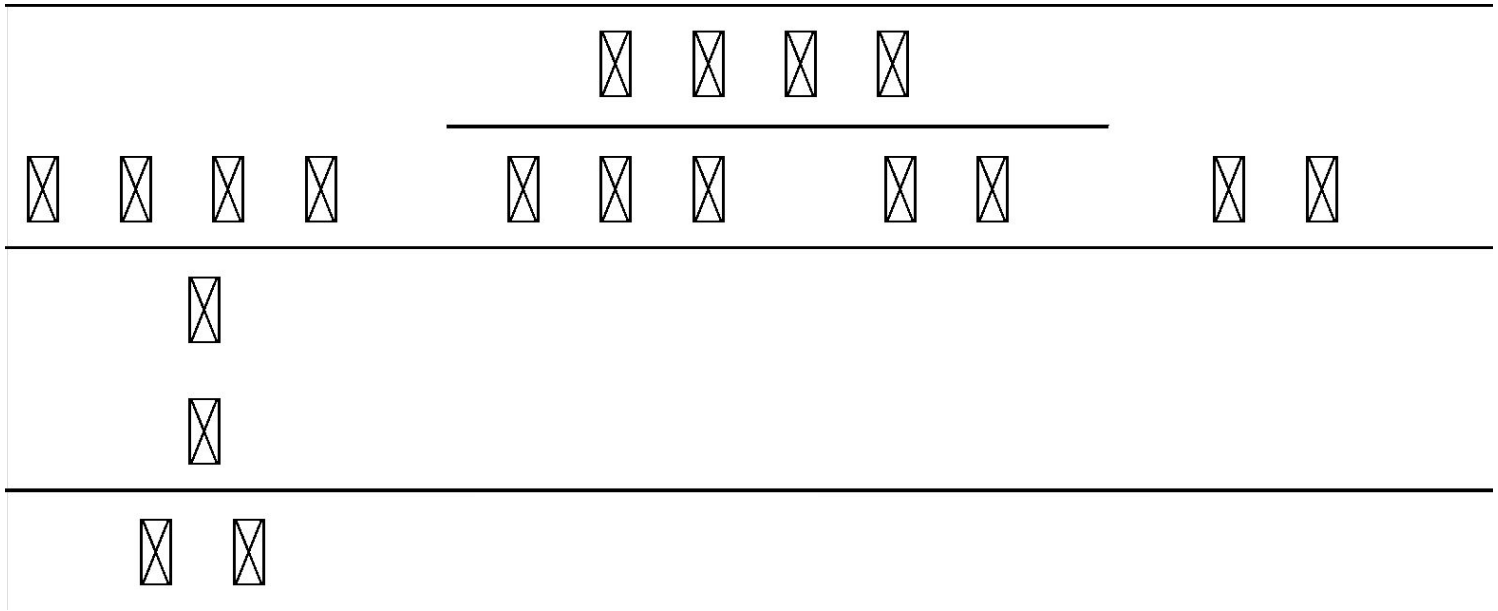
- 在一個16-21歲未就讀大學的樣本中，13.50%未就業，29.05%是高中輟學，5.32%是高中輟學且未就業

- 能否用這個例子做成列聯表？



列聯表範例

在一個16-21歲未就讀大學的樣本中，13.50%未就業，29.05%是高中輟學，5.32%是高中輟學且未就業



	5.32%	23.73%	29.05%
	8.18%	62.77%	70.95%
	13.50%	86.50%	100%

勝算 (Odds)

發生A事件是不發生A事件的幾倍

$$\frac{P(A)}{1 - P(A)}$$

練習2

- 在一個16-21歲未就讀大學的樣本中，13.50%未就業，29.05%是高中輟學，5.32%是高中輟學且未就業

- 1. 請問高中輟學的勝算是多少？
- 2. 請問未就業的人當中，高中輟學的勝算是多少？

範例4

商店老闆想要增加一個僅供現金交易專用的結帳櫃檯，以加速結帳速度，請根據此數據說明這個做法是否必要？

	Payment Method			
Number of Items Purchased	*Cash*	*Check*	*Credit/ Debit Card*	*Row Total*
5 or fewer	30	15	43	88
6 to 9	46	23	66	135
10 to 19	31	15	43	89
20 or more	19	10	27	56
Column Total	126	63	179	368

思考

- $$P(A \mid B) = P(A).$$

$$P(A \cap B) = P(A)P(B)$$

- $P(A|B_1) = P(A)$
- $P(A|B_2) = P(A)$
- $P(A|B_3) = P(A)$
- $P(A|B_4) = P(A)$

範例4 解答

$$P(A \mid B) = P(A).$$

$$P(A \cap B) = P(A)P(B)$$

- 現金結帳的機率

$$P(C) = \frac{126}{368} = .3424$$

- 各個條件機率

$$P(C \mid 5 \text{ or fewer}) = \frac{30}{88} = .3409 \qquad P(C \mid 6 \text{ to } 9) = \frac{46}{135} = .3407$$

$$P(C \mid 10 \text{ to } 19) = \frac{31}{89} = .3483 \qquad P(C \mid 20 \text{ or more}) = \frac{19}{56} = .3393$$

練習3

1. 請問高中輟學的人當中，有多少比例是未就業的？
2. 在16-21歲未就讀大學的族群裡，如果找到一個未就業的受訪者，他/她是高中輟學生的機率是多少？
3. 請問高中是否輟學和未就業與否的相關性是正相關或負相關？

	5.32%	23.73%	29.05%
	8.18%	62.77%	70.95%
	13.50%	86.50%	100%

EXCEL樞紐分析表

- 請參考資料檔：Week 4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

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)