

2

進階資料處理

本章概念

- 2-1 資料整理與整合 P.38
大數據、資料與資訊、
資料前處理、資料處理實作
- 2-2 資料轉換 P.56
開放文件格式、加密、
文字與語音轉換

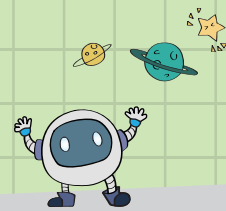




- 你是否曾發現，在網路上看到的廣告，似乎都是自己感興趣的產品？在網路上聆聽音樂時，推薦的歌曲也都符合自己的喜好？這些都是資料分析後產生的結果。
- 隨著網路、電腦技術進步，人們所產生的數位足跡都被一一記錄，形成「大數據」。
- 藉由電腦強大的運算能力，許多公司得以透過資料科學技術，將這些龐大且包羅萬象的資料，整理成有用的資訊，藉此進行預測分析、找出潛在商機。
-



資料整理與整合



1 大數據

隨著資訊科技發達，網路搜尋紀錄、網站點閱率、交易紀錄、通話紀錄等資料被蒐集起來，累積成龐大的資料，稱為「大數據（big data）」（又稱為巨量資料）。

大數據除了資料量龐大之外，還有多樣、即時、真實、具有價值等特性（表 1-2-1）。透過大數據的分析，能有效預測未來趨勢，使得這項技術被廣泛應用在各領域（圖 1-2-1）；然而，隱私權、資訊安全等議題，則成為大數據應用時所面臨的重大挑戰。

表 1-2-1 大數據的 5V 特性。

	特點
資料量大 Volume	大數據要處理的資料量很大，常需一次處理幾 TB（ 10^{12} 位元組）或幾 PB（ 10^{15} 位元組）以上的資料。 註 1TB 的數據需要 2 臺 500GB 的硬碟才能儲存，而 1PB 的數據則要動用 2000 臺 500GB 的硬碟才能容納。
資料多樣性 Variety	資料來源複雜、型態多樣，包含數值、文字、影音、圖片等不同類型，因此對資料處理能力要求更高。
資料即時性 Velocity	隨著網際網路普及，資料快速累積，因此越能即時處理，越能發揮資料的價值。
資料真實性 Veracity	資料必須經過分析、過濾，以排除異常的資料，防止有問題的數據影響到整個資料系統的完整性。
資料價值性 Value	資料多寡並非大數據重點，如何從數據中發掘有意義的特徵，進而產出有價值的結果，才是大數據的核心意義。

圖 1-2-1 大數據應用實例



情境 當你查詢過「高中會考」的資料，網站就會推薦你「考古題解析」、「名校生筆記」等書籍的廣告。

說明 網路業者經由蒐集「查詢過相關訊息的人」之基本資料、瀏覽紀錄，並在分析、判斷後，篩選與推薦最符合使用者需求的廣告。



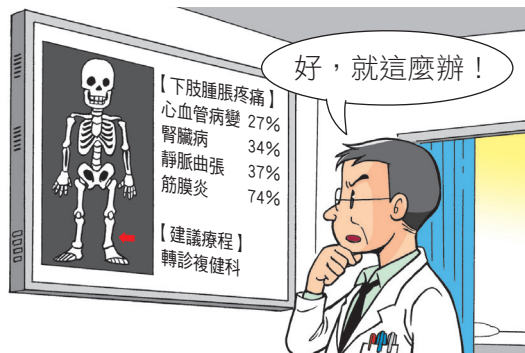
情境 當你利用手機 app 拍照，app 會自動標示地點、分辨景物等資訊。

說明 app 透過雲端數據分析和歸納，判斷拍攝內容是人、動物或景物，並自動標註地點、時間、人名等資訊。



情境 政府單位依據大數據分析結果，在犯罪率高的地區派駐警力，以降低犯罪事件的發生、提高破案率。

說明 利用大數據分析犯罪資料，在容易發生犯罪的時間、地點派駐警力，除了節省人力，還能提高事前預判的功效，有效減少犯罪發生。



情境 利用大數據輔助醫生進行疾病診斷，提供治療方案的參考。

說明 醫師能接觸的病患個案有限，但經過大數據的案例對比，能找出相關資訊，協助醫師作出診斷，並提出合宜的治療方案。

2 資料與資訊

當我們將真實世界發生的事情記錄下來，就形成了「資料」，然而，我們往往很難直接參考龐大的資料來作為判斷或行動的依據。

舉例來說，當一家公司想要根據去年的銷售狀況，制定今年的網路宣傳策略，就必須先將銷售資料（data）經過有計畫、有目的、有系統的處理，才能獲得所需的資訊（information），作為決策時的參考依據（圖 1-2-2）。

▼ 圖 1-2-2 資料與資訊的差異。

資料 (data)

北區銷售紀錄				
消費日	姓名	性別	生日	居住地
1/1	陳瑞娥		1970.04.02	台北
1/1	張紹華	男	1998.10.12	台北
1/1	盧韻筑	女	1990.11.23	桃園
1/2	李曉青	女	2001.03.22	基隆
1/2	黃維庭	男	2004.12.03	台北
1/2	李世芬		1984.09.13	新竹

南區銷售紀錄					
消費日	姓名	生日	性別	居住地	
1/1	趙政和	06/02/26	男	高雄	
1/1	王瑋	00/01/05	女	台南	
1/2	黃宜煊	92/02/16	女	高雄	
1/2	沈心潔	85/12/07		屏東	
1/2	吳慧如	99/02/07	女	台南	
1/2	張羿凱	77/09/16	男	台南	
1/3	賴元志	85/09/06	男	高雄	
1/3	張瑞鴻	93/05/19	男	高雄	
1/3	陳智瑋	02/06/15	男	台南	

雖然收集了所有分公司的銷售紀錄，但是很難直接從中作出決策。

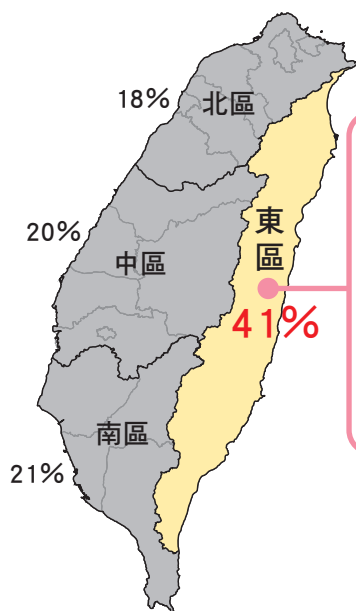
??

「資料」只是事實的紀錄，在未經處理前，對決策沒有幫助。

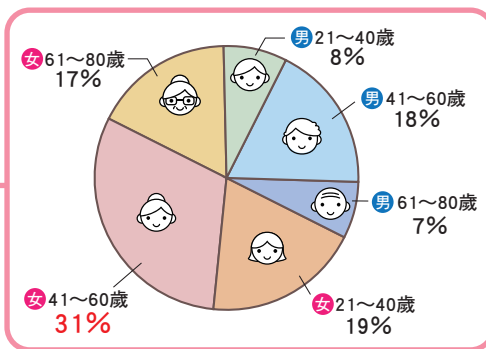
「資料」和「資訊」兩個詞看起來很像，但具有不同的內涵。資料是指未經處理的文字、數字、檔案等對事實的紀錄（例如：成績紀錄、產品訂購單、問卷調查表等），而將資料經過系統分析處理（分類、歸納、解釋等），才能成為具有意義的資訊。

資訊 (information)

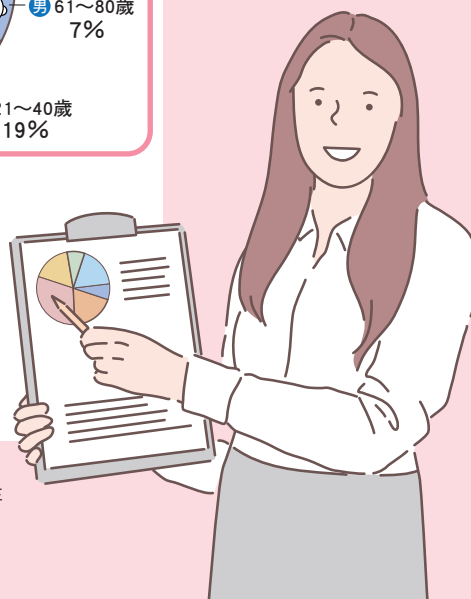
全國銷售比例



東區消費族群分析



從統計與分析的結果來看，可以針對東部地區、41～60歲的女性展開行銷宣傳活動。



⚠ 資料經過處理成為「資訊」，可以提供決策進行時的判斷依據。

3 資料前處理

一般而言，要將資料轉化為資訊，必須將各種不同來源的資料進行整合，再進行清理、轉換等前處理工作，最後再以統計、歸納等方法進行資料分析與解釋（圖 1-2-3）。

▼ 圖 1-2-3 資料處理流程。

1 資料整合

不同來源的資料分散於不同的檔案，而且每個檔案的欄位可能不一致，因此必須調整後匯集在同一檔案中，以利後續工作。

消費日	姓名	性別	生日
1/1	陳瑞娥		1970.04.02
1/1	張紹華	男	1998.10.12
1/1	盧韻筑	女	1990.11.23

消費日	姓名	生日	性別
1月1日	林志強	1985/5/21	男
1月1日	黃美英	1998/7/22	女
1月1日	陳宇鴻	1877/2/28	男



消費日	姓名	性別	生日
1/1	陳瑞娥		1970.04.02
1/1	張紹華	男	1998.10.12
1/1	盧韻筑	女	1990.11.23
1月1日	林志強	男	1985/5/21
1月1日	黃美英	女	1998/7/22
1月1日	陳宇鴻	男	1877/2/28

2

2 資料清理

資料整合後，接著要預先處理資料中常見的問題，以確保後續統計、分析的正確性。

常見 資料問題

Q

1. 資料不完整：

資料漏失，缺少分析時需要的屬性資料。

例 漏填了性別、生日、年齡等資料。

2. 資料有雜訊：

資料錯誤造成的不合理或不正常。

例 身分證字號多 1 碼、電話多 1 碼、年齡 200 歲等。



資料前處理是確保產出正確分析結果的重要環節，例如：在數字的欄位中混雜文字資料，可能造成計算錯誤；而過大或過小的異常數據，則會使最終的統計結果失真。

3 資料轉換

將資料轉換成相同單位、格式，避免分析時產生錯誤。

例 日期格式「1900/1/1」：西元年、「/」分隔、個位數前不加0。
行動電話格式「0912345678」：10碼，不含「-」分隔。

消費日	姓名	性別	生日
1/1	陳瑞娥	女	1970.04.02
1/1	張紹華	男	1998.10.12
1/1	盧韻筑	女	1990.11.23
1月1日	林志強	男	1985/5/21
1月1日	黃美英	女	1998/7/22
1月1日	陳宇鴻	男	1977/2/28

3

消費日	姓名	性別	生日
1/1	陳瑞娥	女	1970/4/2
1/1	張紹華	男	1998/10/12
1/1	盧韻筑	女	1990/11/23
1/1	林志強	男	1985/5/21
1/1	黃美英	女	1998/7/22
1/1	陳宇鴻	男	1977/2/28

常用 清理方式

A

1. 直接刪除法：

當資料量很龐大，而遺缺資料只佔少數，則可直接刪除此筆資料。

2. 填補法：

以人工方式加以確認、修正資料，或以平均值、中位數等數值取代遺缺資料。若遺缺資料過多，則會浪費許多時間和人力。

4 資料處理實作

任務說明

某國中舉辦「9 年級模擬考」後，國文老師想要找出試卷中的「難題」，作為加強解說的依據。請協助國文老師進行資料處理。

題號	1	2	3	4	5	
國文答案	C	C	D	A	B	A
答對人數	157	154	173	102	136	
答錯人數	67	70	51	122	88	
答錯率	29.9%	31.3%	22.8%	54.5%	39.3%	25

↓ 檔案 2-1 下載方式請見 P.3

【資料說明】

1. 考生作答的讀卡資料存放於「國文作答資料.csv」中，各科模擬考答案存放於「模擬考答案.csv」中。
2. 全校 9 年級共 230 名學生，出席考試者以「1」表示，缺考者以「0」表示。
3. 試題答錯率達 50% 以上者即為難題。（以實際考試人數計算）

【任務目標】

請統計各題的答錯率，並將其數值超過 50% 的儲存格標示出來。



延伸學習

資料交換格式

不同試算表或統計軟體有其特定的檔案儲存格式，為了方便資料的傳遞與運用，我們可以將檔案存成 CSV、XML 等格式，供不同軟體讀取應用。

CSV

(Comma Separated Values)

CSV 是一種儲存純文字形式的資料交換格式，使用逗號分隔資料欄位，資料結構較單純，檔案也相對較小。

姓名, 座號, 國文, 英文
小明, 1, 80, 80
小美, 2, 60, 50
小花, 3, 90, 70
小忠, 4, 70, 90

XML

(Extensible Markup Language)

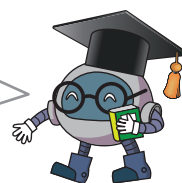
XML 是一種可延伸標記式語言，檔案中包含資料結構、儲存資訊、傳送資訊，適合存放結構較複雜的資料，但檔案相對較為龐大。

```
<Worksheet ss:Name="第一次段考">
<Table ss:ExpandedColumnCount="1024" ss:Expanded
x:FullColumns="1" x:FullRows="1" ss:DefaultColumnW
ss:DefaultRowHeight="15.75">
<Column ss:AutoFitWidth="0" ss:Width="48.75" ss:Sp
<Column ss:Index="5" ss:AutoFitWidth="0" ss:Width=
<Row ss:AutoFitHeight="0">
<Cell ss:StyleID="s65"><Data ss:Type="String">姓名<
<Cell ss:StyleID="s65"><Data ss:Type="String">座號<
<Cell ss:StyleID="s65"><Data ss:Type="String">國文<
```

1 資料匯入

我們將使用 Google 試算表來進行數據的統計處理，因此必須利用「匯入」的方式，讀入 CSV 格式資料。

註 你也可以使用其他試算表軟體進行操作。



1. 新增 Google 試算表

- ① 登入雲端硬碟。
- ② 點擊 **新增**。
- ③ 點擊 **Google 試算表**。



2. 將檔案匯入試算表中

- ① 點擊 **檔案**。
- ② 點擊 **匯入**。
- ③ 點擊 **上傳**，並選擇檔案。

註 請記得為試算表命名。



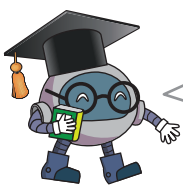
3. 設定匯入方式

- ① 點擊展開下拉式選單。
- ② 選擇 **插入新工作表**，其餘設定依預設即可。
- ③ 點擊 **匯入資料**。

註 請匯入以下兩個檔案：

- 國文作答資料.csv
- 模擬考答案.csv

2 資料清理



難題的計算是以實際出席考試者的答錯率來計算，因此我們可以利用「篩選」功能，將缺考者的資料刪除，以確定出席人數。但要注意的是，資料可能有誤，因此我們要重複檢查，以確保資料的正確性。

年級	班級	座號	性別
1	9	1	1
2	9	1	2
3	9	1	3
4	9	1	4
5	9	1	5
6	9	1	6
7	9	1	7
8	9	1	8
9	9	1	9
10	9	1	10
11	9	1	11

1. 依出席欄篩選資料

- ① 選擇工作表「國文作答資料」。
- ② 點擊欄、列的左上角，全選表格。
- ③ 點擊「▼」開啟篩選器。
註 若 ▼ 未出現在工具列上，請點擊「...」(更多)，展開折疊的工具列。
- ④ 第一列資料旁會出現篩選器符號。

班級	座號	性別	出席	作答
1	1	1	1	0
2	1	2	1	11
3	1	3	1	0
4	1	4	1	11
5	1	5	1	0
6	1	6	1	11
7	1	7	1	0
8	1	8	1	11
9	1	9	1	0
10	1	10	1	11
11	1	11	1	0

2. 篩選設定

- ① 點擊「出席欄」的篩選器。
- ② 選擇「依值篩選」。
- ③ 將「1」取消勾選。
→不要出現「有出席」的資料。
- ④ 點擊「確定」。

年級	班級	座號	性別	出席	作答
55	9	3	10	1	0
71	9	4	2	1	11
128	9	6	9	1	0
175	9	8	8	2	11
188	9	8	21	1	0
205	9	9	13	1	0
217	9	10	9	1	10
226	9	10	18	1	0
232					

3. 檢視篩選結果、進行清理

- ① 第 71、175、188、226 列，含有作答資料，出席應改為「1」。
- ② 第 55、128、205、217 列，沒有作答資料，故刪除以上 4 列資料。

註 刪除方式：選取列號／右鍵／刪除選取的列。

! 實際情形中，資料有誤時，應調閱考生作答卡進行核對，以維護考生權益。

	A	B	C	D	E	F	G
1	年級	班級	座號	性別	出席	作答	
70					1	D,C,D,A,D,C,C,B,B,A,C	
173					1	B,C,D,A,C,A,B,C,D,B,D,A,A	
186					1	C,C,D,B,C,B,B,A,D,B,D,C,D	
222					1	A,B,C,D,D,C,B,A,A,B,C,D,D	
228							
229							

4. 關閉篩選

- ① 點擊 **出席欄** 的 **篩選器**。
- ② 點選 **全部選取**。
→所有資料都出現，也就是取消篩選功能。
- ③ 點擊 **確定**。

	A	B	C	D	E	F	G
1	年級	班級	座號	性別	出席	作答	
2	9					A,B,D,C,A,C,A,B,D,C,D,A	
3	9					A,D,D,C,C,C,C,B,D,A,C,C,B	
4	9					B,D,C,C,D,B,A,C,D,B,B,A,A	
5	9					B,D,D,A,B,A,D,A,D,C,C,A,D	
6	9					C,C,D,C,B,A,B,C,D,B,B,C,D	
7	9					C,C,C,D,B,A,A,C,D,B,A,C,D	
8	9					C,C,D,A,B,A,D,C,A,B,B,A,D	
9	9					A,C,D,A,C,A,D,C,D,B,C,C,D	
10	9					C,D,B,B,C,D,D,A,B,A,C,B	

5. 依作答資料篩選

- ① 點擊 **作答欄** 的 **篩選器**。
- ② 點擊 **清除**，取消所有勾選。
- ③ 點選 **空白**。
→我們要檢查是否還有「完全沒作答（未出席）」的資料。
- ④ 點擊 **確定**。

	A	B	C	D	E	F	G
1	年級	班級	座號	性別	出席	作答	
37	9	2	15	1	1		
154	9	7	12	1	1		
228							

6. 檢視篩選結果、進行清理

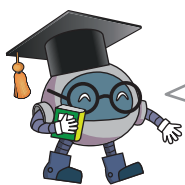
第 37、154 列，雖然標示出席，但沒有作答，因此將這兩筆資料刪除。

國文模擬考作答分析							
檔案 編輯 查看 插入 格式 資料 工具 外掛程式 說明 上一步...							
100% NT\$ % .0 .00 123 預設 (Arial) 10 B I U A ...							
G20	A	B	C	D	E	F	G
1	年級	班級	座號	性別	出席	作答	
2	9	1	1	1	1	1 A,B,D,C,A,C,A,B,D,C,D,A,B,B,C,A,D,B,A,C,B,D,D,C,A,C,D,	
3	9	1	2	1	1	1 A,D,D,C,C,C,C,B,D,A,C,C,B,A,C,C,D,D,D,A,D,B,A,B,C,C,C,C,B	
4	9	1	3	1	1	1 B,D,C,C,D,B,A,C,D,B,B,A,A,B,C,B,C,D,B,C,D,B,A,D,A,C,B,A,B,	
5	9	1	4	1	1	1 B,D,D,A,B,A,D,A,D,C,C,A,D,D,C,C,A,C,B,B,B,A,C,A,D,A,B,C,A,D,	
...	...	...	...	...	...	...	...
222	9	10	20	2	1	1 C,C,C,C,D,A,D,C,C,C,C,A,A,A,C,C,C,B,B,A,C,C,C,C,C,C,C,C,C,	
223	9	10	21	1	1	1 C,	
224	9	10	22	2	1	1 C,C,D,A,B,A,D,C,D,B,A,A,D,B,C,C,B,D,A,B,B,D,D,A,C,D,C,C,D,B,	
225	9	10	23	1	1	1 C,B,A,D,C,A,C,B,B,B,A,B,D,A,D,C,B,C,B,A,A,C,B,B,D,A,C,B,B,A,	
226							

7. 完成清理、關閉篩選

- ① 點擊工具列上的 **篩選器**，關閉篩選功能（呈現空心漏斗 ）。
- 註 資料清理後，共有 224 筆作答資料（含標題共 225 列）。

3 資料整理



統計試題答錯率時，必須逐題比對作答資料，但讀卡資料為一長串以逗號分隔的文字，不利於統計分析；因此我們要先將每個作答資料分隔成獨立欄位，再進行後續工作。



1. 將作答資料分隔成獨立欄位

- ① 點擊欄位 F，選取一整欄。
- ② 點擊資料，開啟選單。
- ③ 點擊選單中的將文字分隔成不同欄。

E	F	G	H	AZ	BA
出席	作答				
1	A	B	D	B	C
1	A	D	D	B	A
1	B	D	C	D	A
1	B	D	D	C	A
1	C	C	D	A	A
1	C	C	C	C	A
1	C	C	D	A	B

2. 完成數據分隔

註 數據分隔完成後，最後一欄資料是「BA 欄」。

	A	B	C	D	E	F	G	H
1	年級	班級	座號	性別	出席	作答		
2	9	1	1	1	1	A	B	D
3	9	1	2	1	1	A	D	D
4	9	1	3	1	1	B	D	C
5	9	1	4	1	1	B	D	D
6	9	1	5	2	1	C	C	D
7	9	1	6	2	1	C	C	C

	A	B	C	D	E	F	G	H	I	J	K	L	M	P	Q	R	S	T	U		
1	年級	班級	座號	性別	出席	作答															
2	9	1	1	1	1	A	B	D	C	A	C	A	A	B	D	C	D	A	D	B	B
3	9	1	2	1	1	A	D	D	C	C	C	C	B	D	A	C	C	B	A	C	C
4	9	1	3	1	1	B	D	C	C	D	B	A	C	D	B	B	A	A	A	B	C
5	9	1	4	1	1	B	D	D	A	B	A	D	A	D	C	C	A	D	D	C	C
6	9	1	5	2	1	C	C	D	C	B	A	B	C	D	B	B	C	D	A	A	C
7	9	1	6	2	1	C	C	C	D	B	A	A	C	D	B	A	C	D	A	A	C
8	9	1	7	2	1	C	C	D	A	B	A	D	C	A	B	B	A	D	A	A	C
9	9	1	8	1	1	A	C	D	A	C	A	D	C	D	B	C	C	D	C	D	D

3. 快速調整欄寬

表格欄寬太大時不易操作，因此先進行此調整。

- ① 點選欄、列的最左上角處，全選表格。
- ② 將滑鼠游標移到 A、B 欄的分隔線之上，呈現↔符號時，雙擊滑鼠左鍵。
- ③ 欄寬調整完成。

The diagram illustrates the process of selecting a seat. A red arrow points down to a table with columns E through M. The first row is highlighted in blue, and the number '1' is in the 'F' column. A red circle with the number '3' is next to the 'F' column. To the right, another table with columns AX through BA is shown, with the first row highlighted in blue and the number '48' in the 'BA' column. A red circle with the number '4' is next to the 'BA' column.

註 總共有 48 題。



選取 1 項
→複製內容

選取 2 項
→ 形成等差數列

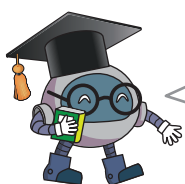
選取 1 項
→ 形成等差數列

選取多項
→循環複製

選取多項
→循環複製

→循環複製

4 資料分析



現在，我們要計算出各題的答對數量（利用 COUNTIF 函式計算各題的作答中，有幾個答案和正確答案相同），然後再計算出每題的答錯數、答錯率。



試算表 COUNTIF 函式

語法：COUNTIF（範圍, 條件）

用途：在指定的範圍內，逐一比對儲存格內容與條件，計算出符合條件的儲存格數量。

註 第 2 引數「條件」可以是數字、文字、表示式，或是指向其他儲存格的位址。

【範例】

	A	B	C	D
1	姓名	性別	年齡	身高
2	小美	女	11	152
3	小華	男	12	144
4	小真	女	11	140
5	小全	男	12	130
6	小可	男	11	155

統計條件	公式	結果
①年齡為 11 歲	=COUNTIF(C2:C6, 11)	3
②女生	=COUNTIF(B2:B6, "女")	2
③身高 140 公分以上	=COUNTIF(D2:D6, ">=140")	4
④和小華相同性別	=COUNTIF(B2:B6, B3)	3



知識快遞

1. 若函式的引數為文字，必須使用雙引號" " 將文字引起來，如例②的「"女"」。
2. 若函式的參數為數字、位址，則可直接寫，如例①的「11」為數字，例④的「B3」以及各例中的範圍均為位址。
3. COUNTIF 第二引數的表示式參考：
 - 大於 5：">5"
 - 不等於 5："<>5"
 - 大於或等於 5：">=5"

Step1 設定「統計答對數」公式

	A	B	C	D	E	F	G	H	I	J	K
1	題號	1	2	3	4	5	6	7	8	9	10
2	國文答案	C	C	D	A	B	A	D	C	D	B
3	答對人數										
4	答錯人數										
5	答錯率										
6											

1 模擬考答案

1. 準備統計表

- ① 點選工作表 **模擬考答案**。
- ② 刪除原本的英、數、自、社答案資料，改為「答對人數」、「答錯人數」、「答錯率」三列。

B3	A	B	C	D	E	F	G	H	I
1	題號	1	2	3	4	5	6	7	8
2	國文答案	C	C	D	A	B	A	D	C
3	答對人數	=COUNTIF(範圍, 條件)							
4	答錯人數								
5	答錯率								
6									

範例
COUNTIF(A1:A10, ">20")
說明
傳回範圍內符合條件的項目數量。

2. 設定公式

- ① 在 B3 格輸入「=COUNTIF(」。
 - ② 輸入完左括號之後，會出現函數說明的浮動視窗，提示接下來要輸入的是「範圍」。
- 註**「等號」和「左括號」都必須是半型文字。

F1:BA1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	年級	班級	座號	性別	出席	1	2	3	4	5	6	7	8	9	10	11
2	9		模擬考答案!B3	1	A	B	D	C	A	C	A	A	B	D	C	
3	9		=COUNTIF('國文作答資料'!F:F													
4	9	1	3	1	1	B	D	C	C	B	A	C	D	B	B	
5	9	1	4	1	1	B	D	D	A	B	A	D	A	D	C	
6	9	1	5	2	1	C	C	D	C	B	A	B	C	D	B	
7	9	1	6	2	1	C	C	C	D	B	A	A	C	D	B	
8	9	1	7	2	1	C	C	D	A	B	A	D	C	A	B	
9	9	1	8	1	1	A	C	D	A	C	A	D	C	D	B	
10	9	1	9	1	1	C	C	D	B	B	C	D	D	A	B	

1 國文作答資料

3. 選取「要統計的範圍」

第 1 題的作答資料在「國文作答資料」工作表中的 F 欄。

- ① 點擊工作表 **國文作答資料**。
- ② 點選 F 欄的欄號，全選 F 欄。

註我們也可以直接輸入：
'國文作答資料'!F:F

	A	B	C	D	E	F	G	H	I	J	K	L
1	年級	班級	座號	性別	出席	1	2	3	4	5	6	7
2	9		模擬考答案'B3			0		B	D	C		A
3	9		=COUNTIF('國文作答資料'!F:F,)									
4	9		COUNTIF(範圍, 條件)									
5	9											

4. 完成範圍選取

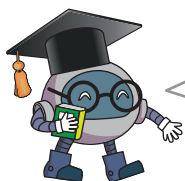
- ① 範圍選擇好之後，請以鍵盤輸入一個半型逗號。
→代表完成「範圍」的輸入。
- ② 此時浮動視窗會提示，接下來要輸入「條件」。

B3	=COUNTIF('國文作答資料'!F:F, B2)											
	A	B	C	D	E	F	G	H	I	J	K	L
1	題號	1	2	3	4	5	6	7	8	9	10	11
2	國文答案	C	C	D	A	B	A	D	C	D	B	B
3	答對人數	157										
4	答錯人數											
5	答錯率											
6												

5. 輸入「比對條件」

比對的條件就是第 1 題的正確答案。

- ① 直接輸入「B2」，然後按 ENTER。
- ② 統計結果為 157 人。



步驟 5 的公式說明

=COUNTIF('國文作答資料'!F:F, B2)

範圍 條件

計算「國文作答資料表的 F 欄」裡，有幾個和「B2 格」相同的資料。



「引用位址」的表示方式

在公式或函式中引用資料時，會依據「是否在同一工作表中」有不同寫法。

例 我們要在「工作表 1」中，計算不同範圍的總和：

引用工作表 引用範圍	引用「工作表 1」的資料 同工作表，直接表示位址	引用「工作表 2」的資料 跨工作表，加註工作表名稱
一格：F5	=SUM(F5)	=SUM('工作表 2'!F5)
多格：F5~H6	=SUM(F5:H6)	=SUM('工作表 2'!F5:H6)
整欄：F 欄	=SUM(F:F)	=SUM('工作表 2'!F:F)

	A	B	C	D
1	題號	1	2	3
2	國文答案	C	C	D
3	答對人數	157		
4	答錯人數	=224-B3		
5	答錯率			

	A	B	C	D	E
1	題號	1	2	3	4
2	國文答案	C	C	D	A
3	答對人數	157			
4	答錯人數	67			
5	答錯率				

6. 計算「答錯人數」

已知考試人數有 224 人，B3 格為答對人數。

① B4 答錯人數：

$$= 224 - B3$$

② 完成計算。

	A	B	C	D
1	題號	1	2	3
2	國文答案	C	C	D
3	答對人數	157		
4	答錯人數	67		
5	答錯率	=B4/224		

	A	B	C	D	
1	題號		1	2	3
2	國文答案	C		C	D
3	答對人數		157		
4	答錯人數		67		
5	答錯率	0.2991071429			

7. 計算「答錯率」

$$\text{答錯率} = \frac{\text{答錯人數}}{\text{考試人數}}$$

① B5 答錯率：

$$= B4 / 224$$

② 完成計算。

		100%		NT\$ %		.0		.00	
B5	fx	=B4/224		格式化為百分比					
	A	B	C	D	E	F	G		
1	題號		1	2	3	4	5	6	
2	國文答案	C		C	D	A	B	A	
3	答對人數		157						
4	答錯人數		67						
5	答錯率	0.2991071429							

		100%		NT\$ %		.0		.00	
B5	fx	=B4/224		減少小數位數					
	A	B	C	D	E	F	G	H	
1	題號		1	2	3	4	5	6	7
2	國文答案	C		C	D	A	B	A	D
3	答對人數		157						
4	答錯人數		67						
5	答錯率	29.91%							

100% NT\$ %							
B5	fx =B4/224						
	A	B	C	D	E	F	
1	題號		1	2	3	4	5
2	國文答案	C		C	D	A	B
3	答對人數		157				
4	答錯人數		67				
5	答錯率		29.9%				

8. 設定「百分比%」格式

將 B5 格設為百分比（小數 1 位）。

① 選取 B5 格。

② 點擊「格式化為百分比」。

③ 點擊「減少小數位數」。

④ 完成設定。

	A	B	C	D	E	F	G	H	I	J	K
1	題號	1	2	3	4	5	6	7	8	9	10
2	國文答案	C	C	D	A	B	A	D	C	D	B
3	答對人數	157									
4	答錯人數	67									
5	答錯率	29.9%									
6											

結果：

	A	B	C	D	E	F	G	H	I	J	K
1	題號	1	2	3	4	5	6	7	8	9	10
2	國文答案	C	C	D	A	B	A	D	C	D	B
3	答對人數	157	154	173	102	136	167	123	155	140	157
4	答錯人數	67	70	51	122	88	57	101	69	84	67
5	答錯率	29.9%	31.3%	22.8%	54.5%	39.3%	25.4%	45.1%	30.8%	37.5%	29.9%
6											

10. 快速複製公式到其他格

- ① 框選 B3、B4、B5 三格。
- ② 將滑鼠游標停在 B5 格的右下角，形成「+」符號。
- ③ 按住「滑鼠左鍵」向右拖曳。
- ④ 放開滑鼠，完成設定。

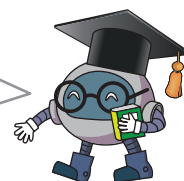


手腦並用

試試看，統計出下列資訊：

1. 每一題中，ABCD 選項分別有多少人選答？
2. 每一題中，ABCD 選項的選答率分別為何？（選答率＝選答人數／考試人數）
3. 本試卷共有幾題難題？

	A	B	C	D	E	F	G	H	I	J	K
1	題號	1	2	3	4	5	6	7	8	9	10
2	國文答案	C	C	D	A	B	A	D	C	D	B
3	答對人數	157	154	173	102	136	167	123	155	140	157
4	答錯人數	67	70	51	122	88	57	101	69	84	67
5	答錯率	29.9%	31.3%	22.8%	54.5%	39.3%	25.4%	45.1%	30.8%	37.5%	29.9%
6											
7	選A人數	28	22	14	102	20	167	22	35	18	21
8	A選答率	12.5%	9.8%	6.3%	45.5%	8.9%	74.6%	9.8%	15.6%	8.0%	9.4%
9	選B人數	31	33	13	34	136	16	50	20	22	157
10	B選答率	13.8%	14.7%	5.8%	15.2%	60.7%	7.1%	22.3%	8.9%	9.8%	70.1%
11	選C人數	157	154	24	37	45	31	29	155	44	26
12	C選答率	70.1%	68.8%	10.7%	16.5%	20.1%	13.8%	12.9%	69.2%	19.6%	11.6%
13	選D人數	8	15	173	51	23	10	123	14	140	20
14	D選答率	3.6%	6.7%	77.2%	22.8%	10.3%	4.5%	54.9%	6.3%	62.5%	8.9%



計算完答錯率之後，我們要把答錯率高於 50% 的數據標示出來。

Step2 標示難題



1. 啟動「條件式格式設定」

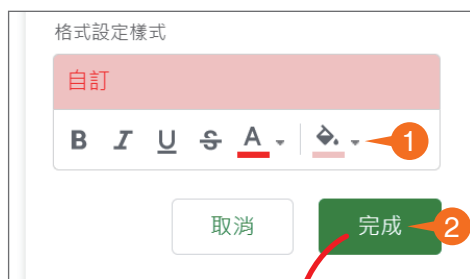
- ① 框選所有答錯率數據 (B5:AW5)。
- ② 點擊「格式」。
- ③ 點擊「條件式格式設定」。



2. 設定「規則」

我們要先設定一個條件，當儲存格的數據符合這個條件時，就會在範圍內套用格式設定。(此處條件為 $\geq 50\%$)

- ① 條件選擇「大於或等於」。
- ② 數值填寫「0.5」或「50%」。



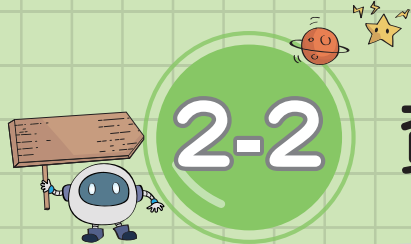
3. 設定「格式」

設定想要呈現的格式。

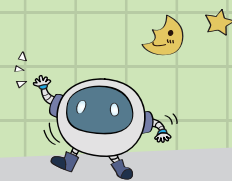
- ① 可以設定文字格式(字型、粗體、顏色等)，也可以設定儲存格的填滿顏色。
- ② 點擊完成。

結果：

題號	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
國文答案	C	C	D	A	B	A	D	C	D	B	C	D	A	A	
答對人數	157	154	173	102	136	167	123	155	140	157	69	86	132	131	57
答錯人數	67	70	51	122	88	57	101	69	84	67	155	138	92	93	167
答錯率	29.9%	31.3%	22.8%	54.5%	39.3%	25.4%	45.1%	30.8%	37.5%	29.9%	69.2%	61.6%	41.1%	41.5%	74.6%

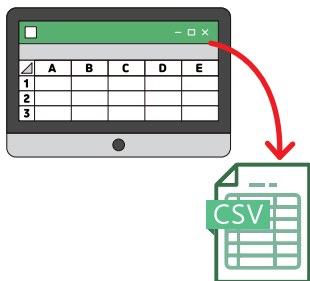


資料轉換

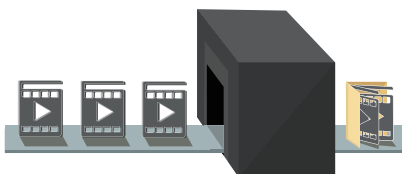


在資料的運用過程中，我們會為了不同的目的而改變資料的儲存格式或表示方式，這就是資料轉換。常見資料轉換分為「檔案格式轉換」和「資料內容轉換」兩類（圖 1-2-3）。

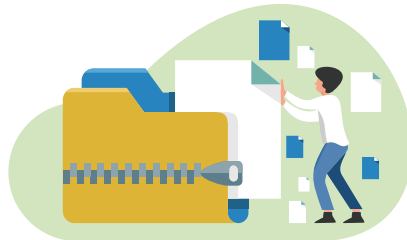
A. 檔案格式轉換



把試算表資料儲存成 CSV 或 XML 格式，方便不同軟體使用。

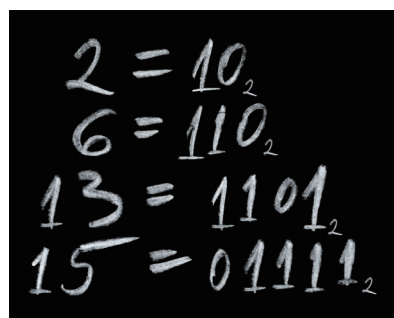


為了畫質、音質、傳輸便利性等不同目的，將影音檔案儲存成不同格式。

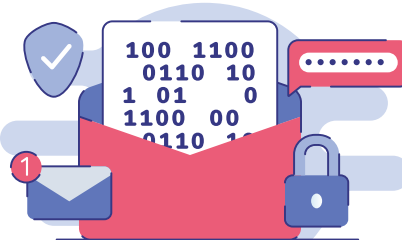


將資料檔案壓縮，減少儲存空間，以方便儲存或加快網路傳輸。

B. 資料內容轉換



把十進位數字轉換成二進位數字，方便電腦進行計算。



使用加密技術傳輸資料，保護資料的安全性及完整性。



將文字透過機器翻譯技術，轉換成另一種語言文字或語音。

▲ 圖 1-2-3 常見的資料轉換案例。

1 開放文件格式（ODF）

一般的文件檔案會以特定的格式儲存（如 .doc、.xls、.ppt 等），這些檔案必須使用特定的商用軟體才能開啟，造成文件流通的阻礙。

為了便於文件的流通，我們可以將檔案存成 ODF 開放文件格式，讓使用者能自由選擇文書處理軟體，並確保文件能被順利閱讀、編輯（表 1-2-2）。

▼ 表 1-2-2 開放文件格式與常見的編輯軟體。

ODF 開放文件格式

文書處理——副檔名.odt
試算表——副檔名.ods
簡報——副檔名.odp
繪圖——副檔名.odg
資料庫——副檔名.odb

常見 ODF 免費編輯軟體

1. Libre Office（支援 Windows、macOS、Linux）
2. Google Docs（支援 Windows、macOS、Linux）
3. Collabora Office（支援 Android、iOS）



知識快遞

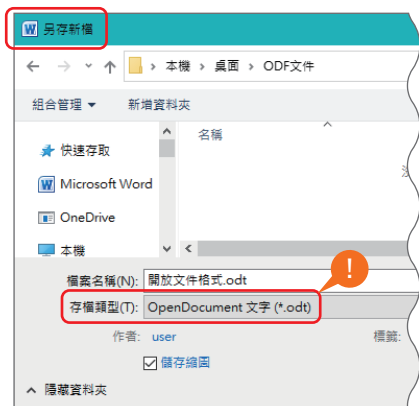
1. ODF（Open Document Format，開放文件格式）：

是一種基於 XML 為基礎的檔案格式規範，因應試算表、圖表、簡報和文書處理文件等電子文件而設置。

2. 將檔案存成 ODF 格式：

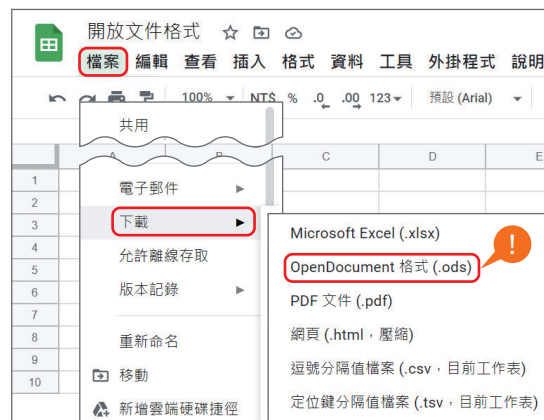
(1) 一般文書軟體：

只要將檔案以「另存新檔」的方式儲存，並選擇存檔類型為「Open Document」的相關格式（.odt、.ods 等）即可。



(2) Google 雲端工具：

執行「檔案」／「下載」，並選擇「OpenDocument 格式（.ods）」類型。





延伸學習

我國推動 ODF 歷程

由於 ODF 具有跨平台、跨應用程式、可避免版本升級衝突的優點，全球許多國家、企業都改用 ODF 做為標準格式，我國政府單位也逐年推廣，目前已全面採用 ODF 格式：

2006 年 ODF 成為 ISO 國際標準（ISO/IEC26300）。

2009 年 我國將 ODF 標準訂為國家標準 CNS15251。

2015 年 我國各機關網站提供下載檔案規範：

1. 可編輯文件：應支援 ODF 文書格式。
2. 非可編輯文件：採用 PDF 文書格式。

2016 年 我國各政府機關的資料交換，須支援 ODF 文書格式。

2017 年 我國全面推動各機關使用可編輯 ODF 文書軟體。



手腦並用

隨著網路的普及與科技的進步，為了促使國家機關資料流通，我國政府開放涵蓋食、醫、住、行、育樂、就業、文化、經濟發展等面向的統計資料，以 ODF 檔案供人民下載、應用。（政府資料開放平臺網址為 <https://data.gov.tw/>）

想一想，如果想知道 2002 年到 2018 年間，新北市國民中學學生裸視不良率的變化情形，你要如何獲取資料？取得資料後，你會如何處理分析，以呈現相關資訊？

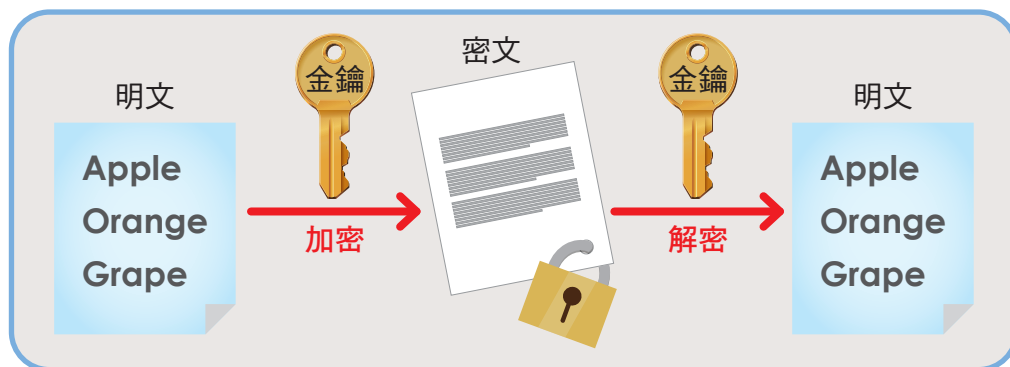


2 加密

古時候的情報訊息只能以人力或飛鴿等方式傳送，被敵人截獲的機率很高，於是人們想到將訊息「加密」，讓敵人就算看到了訊息，也無法得知訊息的真實意義，因而保障情報訊息不會外洩。

1 凱薩密碼（Caesar Cipher）

二千多年前，凱薩大帝發明了「字母替換」的加密方式，將原始訊息（明文）的字母，透過固定的偏移量（加密金鑰），替換成另一個字母（密文），達到為訊息加密的目的。收到密文的人必須再以相同的金鑰回推（解密），以獲得原始訊息（圖 1-2-4）。

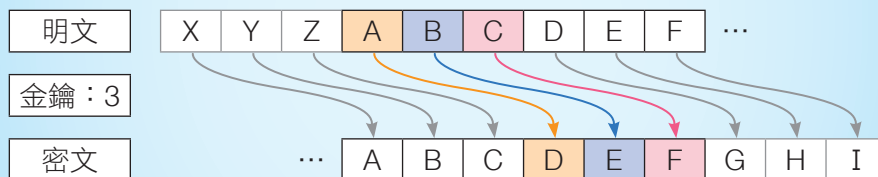


▲ 圖 1-2-4 加密／解密的過程。

【規則】將明文依金鑰的數字偏移。

加密 以「3」為金鑰，原本的明文為 A，密文就變成 D。

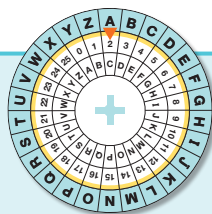
解密 依金鑰「3」往回推，看到的密文是 D，代表明文為 A。



手腦並用

請利用課本附件 1 製作「凱薩密碼盤」。

1. 金鑰 4，明文為 HAPPY，密文應寫作什麼？
2. 金鑰 7，密文為 ZJYHAJO，明文是什麼？



2 維吉尼亞密碼 (Vigenère Cipher)

西元 1586 年，法國外交官維吉尼亞 (Blaise de Vigenère) 改進凱撒密碼，利用維吉尼亞密碼表 (圖 1-2-5) 搭配不同的金鑰來提升加密的複雜度，也增加資料傳送的安全性。

接下來，我們就以「TAIWAN」為明文，介紹維吉尼亞密碼的應用。

明文 \ 金鑰	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
0	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A
2	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B
3	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C
4	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D
5	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E
6	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F
7	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G
8	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H
9	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I
10	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J
11	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K
12	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L
13	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M

▲ 圖 1-2-5 請搭配課本附件 2 維吉尼亞密碼表使用。

加密法 1 依照字母的序號加密。

明文	T	A	I	W	A	N
金鑰	1	2	3	4	5	6
密文	U	C	L			

第 1 個字母，密鑰就是 1，依此類推。

說明 以 TAIWAN 的 A 為例：A 為第 2 個字母，金鑰為 2。
由表中 A 欄，對應至金鑰 2，密文為 C。

加密法 2 依照約定的數字金鑰加密。(例如 12、5、8、23)

明文	T	A	I	W	A	N
金鑰	12	5	8	23	12	5
密文	F	F	Q			

密鑰依序循環。

說明 以 TAIWAN 的 T 為例：T 的金鑰為 12。
由表中 T 欄，對應至金鑰 12，密文為 F。

加密法 3 依照約定的字母金鑰加密。(例如 C、A、K、E)

明文	T	A	I	W	A	N
金鑰	C	A	K	E	C	A
密文	V	A	S			

密鑰依序循環。

說明 以 TAIWAN 的 I 為例：I 的金鑰為 K。
由表中 I 欄，對應金鑰第 K 列，密文為 S。



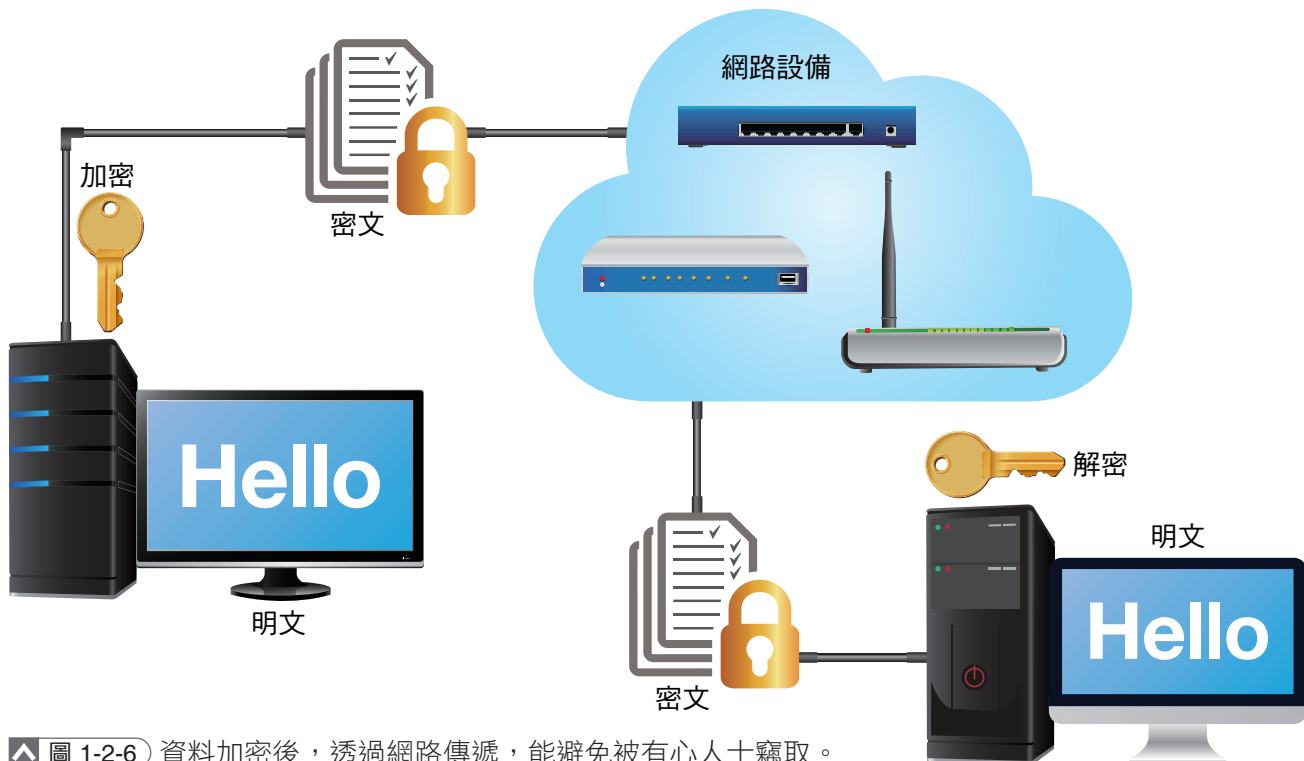
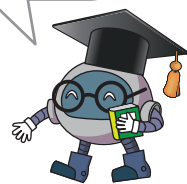
手腦並用

假設你收到一份密文，內容是「TWTMSYXHESM」，已知這份密文是使用維吉尼亞密碼來加密，而且加密金鑰是「APPLE」，則這份訊息的明文是什麼呢？



隨著科技發達，生活中越來越多資料要依賴網路傳送，過程中會經過無數的伺服器、路由器等網路設備，為了避免資料被有心人士竊取窺視，網址為「https://」開頭的網頁，就是運用加密方式（當然比上述介紹的加密方法更加複雜），來保障資訊安全（圖 1-2-6）。

網址為 http:// 開頭的網頁，不具有加密的功能，傳遞訊息可能會有外洩的安全疑慮。



▲ 圖 1-2-6 資料加密後，透過網路傳遞，能避免被有心人士竊取。

3 文字、語音轉換

隨著科技進展，文字、語音轉換技術的應用，讓我們的生活、學習、工作都更加便利，以下介紹常用的相關工具與功能。

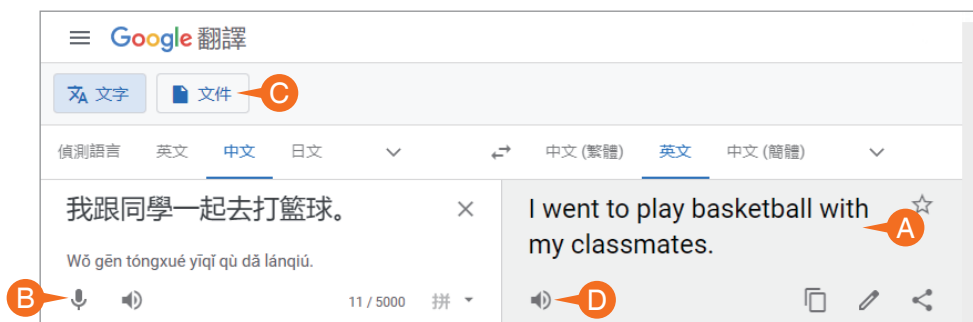
1 Google 翻譯

Google 翻譯（圖 1-2-7）是一款整合了文字轉文字、文字轉語音、語音轉文字功能的軟體，支援超過 100 種語言以上的翻譯功能。另外，還具有整份文件翻譯、語音輸入、語音輸出等功能，翻譯準確度也相當高。

► 圖 1-2-7

Google 翻譯的功能：

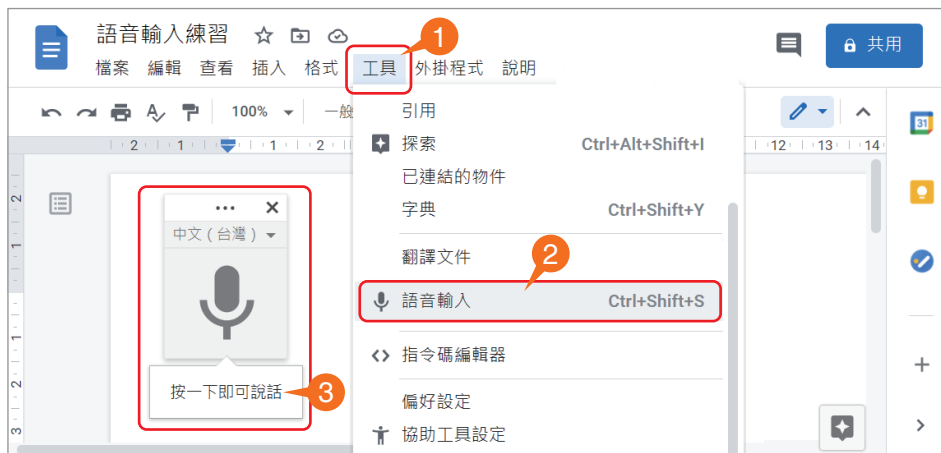
- A. 文字輸入翻譯
- B. 語音輸入翻譯
- C. 整份文件翻譯
- D. 語音輸出



2 Google 文件：語音輸入功能

若想在會議中進行文字記錄，往往因為打字跟不上說話速度，只好錄音下來、重複聆聽謄打。但現在，Google 文件內建了語音輸入功能（圖 1-2-8），只要經由麥克風收音，就可以自動產生文字，方便又快速。

► 圖 1-2-8 在 Google 文件中，只要啟動語音輸入功能，就可以利用說話的方式進行文字輸入，省去打字的麻煩。





延伸學習

文字轉語音、語音轉文字

1. 文字轉語音（Text To Speech，TTS）

這項技術要事先將大量字詞錄製成語音資料，使用時再取出錄音片段、拼接成句子；這種方式的缺點是要耗費大量人力來錄製語音，並且需要龐大的資料庫來儲存語音資料。

除了上述的拼接合成語音，也有利用數學模型統計建構合成聲音的波形，雖然節省了資料庫的空間及人力，但語音會較為僵硬、不自然。

2. 語音轉文字（Speech To Text，STT）

語音辨識系統是結合聲學、語言模型、辭典等技術所構成，將人類說話的語音內容轉換成相對應的文字。當我們使用這項功能時，程式會先將說話聲音的類比訊號轉為數位訊號，經過語音辨識程式對聲音樣本進行比對，最後再轉換成文字。

語音辨識應用相當廣泛，包括語音助理、語音撥號、導航、裝置控制、文件檢索、語音輸入等，都會使用此技術，讓我們的生活更加方便。



第2章 學習重點



2-1 資料整理與整合

1. 大數據 5V 特性：資料量大、資料多樣性、資料即時性、資料真實性、資料價值性。
2. 資料和資訊：資料是指未經處理的內容；資訊則是指經過系統分析處理的資料。
3. 資料處理流程：資料前處理（資料整合→資料清理→資料轉換）→資料分析與解釋。
 - (1) 資料整合：將不同來源的資料，調整、整合在同一檔案中。
 - (2) 資料清理：處理資料中不完整或有雜訊的部分，以確保後續統計、分析的正確性。
 - (3) 資料轉換：將資料轉換成相同單位、格式，避免數值分析時產生錯誤。

2-2 資料轉換

1. 資料轉換：檔案格式轉換、資料內容轉換。
2. 開放文件格式（ODF）：可跨平台、跨應用程式、避免版本升級衝突的開放格式。
3. 加密：利用金鑰轉換明文與密文。
4. 凱薩密碼：以固定偏移量的金鑰，轉換密碼。
5. 維吉尼亞密碼：改良凱薩密碼，金鑰可設定為約定好的數字、文字，並依序循環。

幫檔案瘦身

資料壓縮



隨著網路應用普及，檔案傳輸需求越來越多。然而儲存空間、網路頻寬有限，「將檔案變小」成為相當重要的課題。

我們在搬家時，會盡可能將物品收納成更小體積，檔案壓縮也是利用相似概念，透過各種演算法，將資料收納得更密集，以減少佔用的空間。一般壓縮分為兩種類型：

無失真壓縮法 Lossless Compression

又稱非破壞性壓縮，資料經過壓縮後，仍可還原成原始資料。此種方法的壓縮比率較低，檔案相對較大。

常見的無失真壓縮格式包含影像格式 TIFF、PNG，聲音格式 WAV、FLAC，或壓縮檔格式 RAR、ZIP 等。

例 以影像資料的記錄方式改變為例：

	1	2	3	4	5	6	7	8
1								
2								
3								
4								
5								
6								
7								

壓縮

逐格記錄 記錄 56 次

(1,1) 白色
(1,2) 粉紅色
(1,3) 粉紅色
...

可還原

區塊記錄 記錄 17 次

(1,1) 白色
(1,2 ~ 1,3) 粉紅色
(1,4 ~ 1,5) 白色
...

失真壓縮法 Lossy Compression

又稱破壞性壓縮，資料經過壓縮之後即無法還原。此方法壓縮比率高，檔案相對較小，但失真情形也會隨壓縮比率提高而變嚴重。

常見的失真壓縮格式如影像格式 JPEG、聲音格式 MP3 等。

例 以影像的取樣解析度改變為例：

以 20×16 取樣

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2																				
3																				
4																				
5																				
6																				
7																				
8																				
9																				
10																				
11																				
12																				
13																				
14																				
15																				
16																				

壓縮

以 10×8 取樣

	1	2	3	4	5	6	7	8	9	10
1										
2										
3										
4										
5										
6										
7										
8										

無法還原

在 20×16 的解析度中，(6,2) 格為深紅色，但在一半解析度下，遺失了深紅色的資訊。

常見的檔案壓縮／解壓縮軟體



WinZip

類型：共享軟體

壓縮檔格式：.zip



WinRAR

類型：共享軟體、商業軟體

壓縮檔格式：.rar、.zip



7-Zip

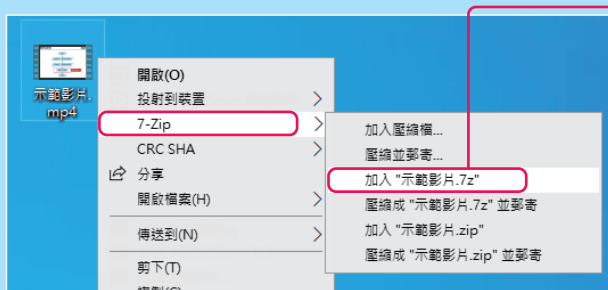
類型：自由軟體

壓縮檔格式：.7z、.zip

7-ZIP 操作簡介 下載位置：<https://www.developershome.com/7-zip/>

①檔案壓縮

選取要壓縮的檔案，按滑鼠右鍵，移至 **7-Zip** 上，再點選要壓縮的格式（.7z 或 .zip）。



②檔案解壓縮

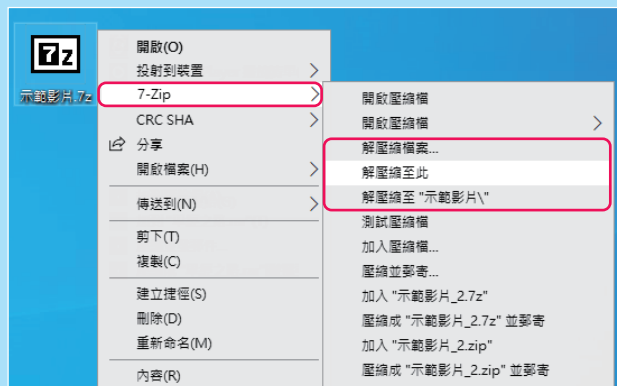
選取要解壓縮的壓縮檔，按滑鼠右鍵，移至 **7-Zip** 上，並依需求選擇 **解壓縮檔案**、**解壓縮至此** 等，即可進行解壓縮。

(1) 7z 格式檔案：

可設定為自動執行檔，直接點擊就可解壓縮，不需安裝解壓縮軟體。

(2) zip 格式檔案：

需要安裝解壓縮程式（如 WinZip、7-ZIP），才能進行解壓縮。



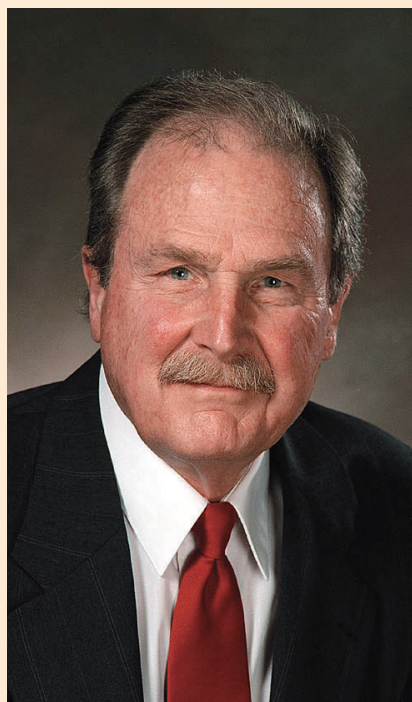
簡單快速的無失真壓縮法

霍夫曼編碼

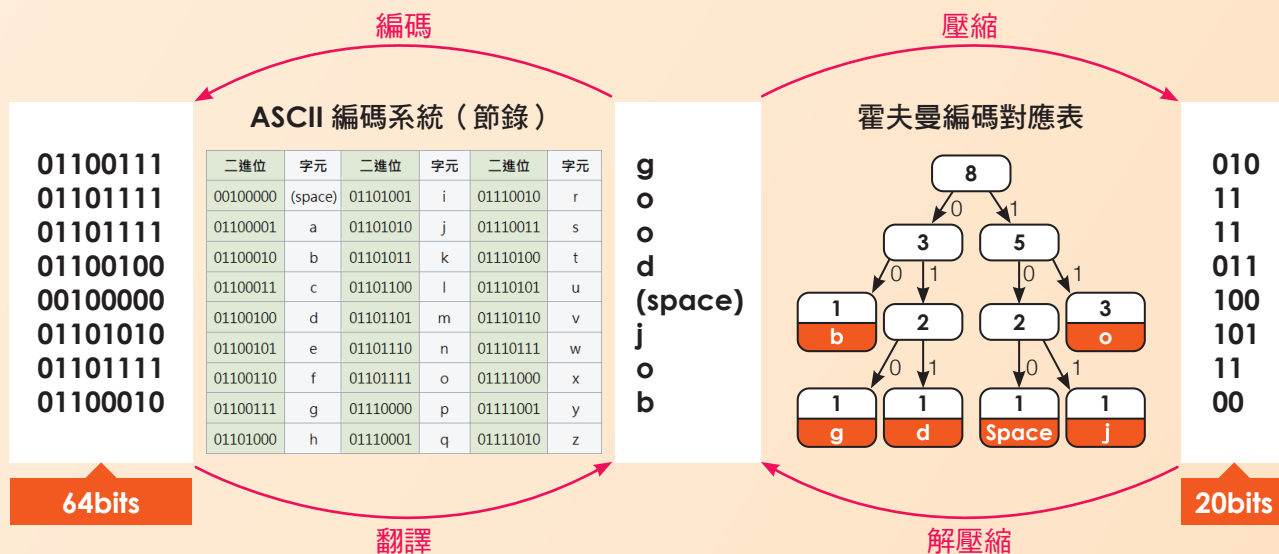
電腦中的所有資料都是以 0 與 1 的組合儲存，例如 ASCII 編碼中，每個字元都是使用 1 個位元組（8bits）的 0、1 組合來表示。西元 1951 年，大衛·霍夫曼提出新的壓縮觀念：將最常見、出現頻率最高的資料，用最小的空間（即最短的編碼）來轉存，以壓縮資料占用的空間。

以字串「good job」（包含空格共 8 個字元）為例，若利用 ASCII 編碼來儲存，會使用 $8 \times 8 = 64\text{bits}$ 的空間，而透過霍夫曼編碼壓縮後，只需要 20bits 便能儲存原字串，約為原本的 33%。

註 霍夫曼編碼需要額外空間來儲存編碼對應表，以利解壓縮還原成原始資料，而原始資料的重複內容越多，能壓縮的空間就更多，因此，雖然需要額外儲存對應表，此方法仍能大幅壓縮所需的儲存空間。



▲ 大衛·霍夫曼（David Albert Huffman，西元1925～1999年）

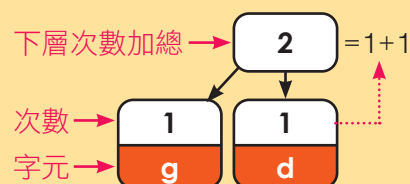


霍夫曼編碼 · 產生步驟

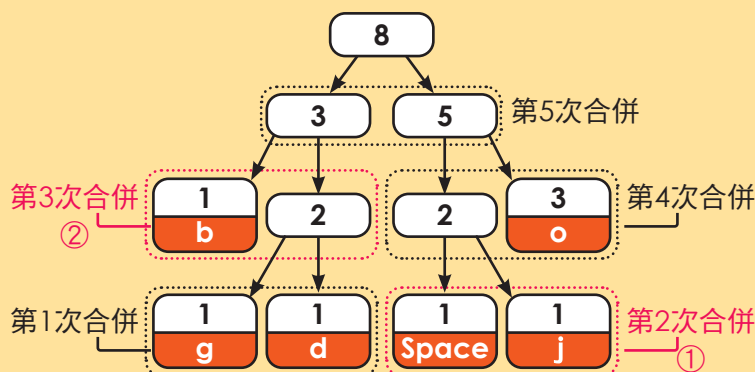
(1) 統計「good job」字元出現的次數。

字元	g	d	(space)	j	b	o
出現次數	1	1	1	1	1	3

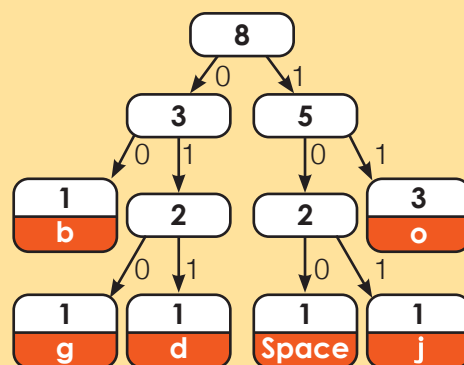
(2) 畫出兩個節點，將出現次數最小的兩個字元及次數填入節點中，並加總次數產生上層節點（如右圖）。



(3) 重複 (2)，優先選擇尚未填入圖中且出現次數最小的字元（下圖中①），或合併後數字最小的節點（下圖中②），依序兩兩填入圖中並合併。



(4) 將每層節點下方的左側箭頭標示 0、右側箭頭標示 1，即完成「霍夫曼二元樹」。



(5) 各字元對應的編碼，即為從最上層節點往下，順著箭頭走到對應字元的路徑標示，其結果如右表所示。

字元	g	d	(space)	j	b	o
對應編碼	010	011	100	101	00	11

手腦並用

(1) 右圖是某段文字壓縮後產生的霍夫曼二元樹，編碼後結果為「0001111011100011010」，想一想，這段文字本來寫了什麼？

(2) 請試著以霍夫曼編碼，將「kangaroo」進行編碼，並利用霍夫曼二元樹產生網站，查看結果是否與自己畫的相同。

網址：<https://www.csfieldguide.org.nz/en/interactives/huffman-tree/>

