一些樣本資料

NOTE

本章所用的 Excel 工作表「Concessions.xlsx」，可於以下網址下載：
http://books.gotop.com.tw/download/ACD014500。

假裝一下，你的生活一團糟，現在你是個成年魯蛇阿宅，在高中母校有棒球比賽時，出去擺攤賺點錢（我發誓，這不完全是我的自傳）。

你有一張昨晚銷售的試算表，如圖 1-1。

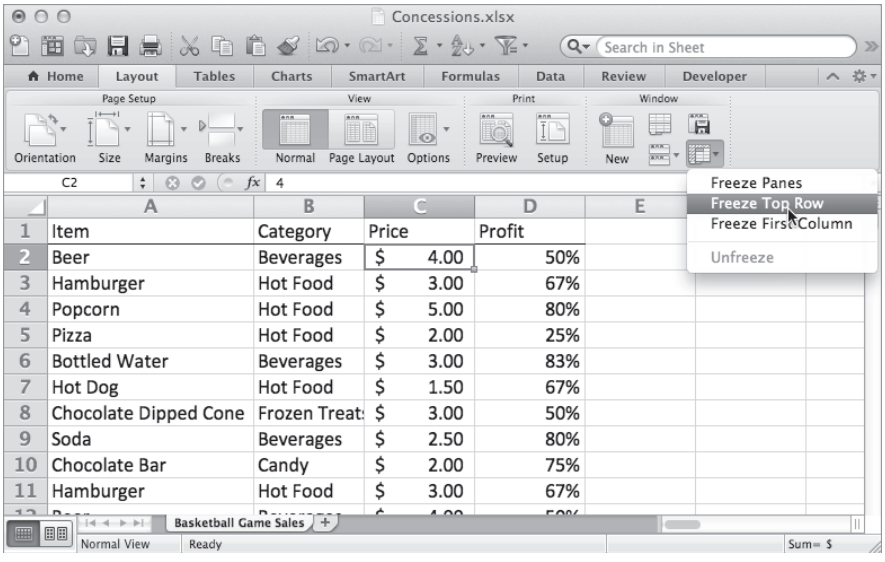
	A	B	C	D
1	Item	Category	Price	Profit
2	Beer	Beverages	$ 4.00	50%
3	Hamburger	Hot Food	$ 3.00	67%
4	Popcorn	Hot Food	$ 5.00	80%
5	Pizza	Hot Food	$ 2.00	25%
6	Bottled Water	Beverages	$ 3.00	83%
7	Hot Dog	Hot Food	$ 1.50	67%
8	Chocolate Dipped Cone	Frozen Treat	$ 3.00	50%
9	Soda	Beverages	$ 2.50	80%
10	Chocolate Bar	Candy	$ 2.00	75%
11	Hamburger	Hot Food	$ 3.00	67%
12	Beer	Beverages	$ 4.00	50%
13	Hot Dog	Hot Food	$ 1.50	67%
14	Licorice Rope	Candy	$ 2.00	50%
15	Chocolate Dipped Cone	Frozen Treat	$ 3.00	50%

圖 1-1：攤位銷售

圖 1-1 之中列出了每一筆銷售、賣出什麼商品、賣出哪一種食物或飲料、價格，以及獲利的百分比。

用控制按鈕快速移動

若你想瀏覽記錄，可以用滑鼠滾軸、觸控板或下方向鍵，把試算表往下捲動。捲動時，可以讓試算表頂端固定顯示首列，以便你記住每一欄的意義。為此，選擇「檢視」標籤中的「凍結窗格」（Freeze Panes）或「凍結頂端列」（Freeze Top Row），Mac 2011 版則是在「版面配置」下，如圖 1-2。



	A	B	C	D	E
1	Item	Category	Price	Profit	
2	Beer	Beverages	$ 4.00	50%	
3	Hamburger	Hot Food	$ 3.00	67%	
4	Popcorn	Hot Food	$ 5.00	80%	
5	Pizza	Hot Food	$ 2.00	25%	
6	Bottled Water	Beverages	$ 3.00	83%	
7	Hot Dog	Hot Food	$ 1.50	67%	
8	Chocolate Dipped Cone	Frozen Treat	$ 3.00	50%	
9	Soda	Beverages	$ 2.50	80%	
10	Chocolate Bar	Candy	$ 2.00	75%	
11	Hamburger	Hot Food	$ 3.00	67%	

圖 1-2：凍結頂端列

若要快速移動到試算表底部，看看有多少筆交易紀錄，可以點選其中一個欄位值，然後按 Ctrl+T。（Mac 為 Command+T。）此時會立即移動到此欄的最後一格。在本試算表中，最後一列為 200。此外，使用 Ctrl/Command 在試算表中左右跳動，作法也相同。

若想察看夜間的平均銷售價格，可在 C 欄，也就是價格欄之下，添加以下公式：

```
=AVERAGE(C2:C200)
```

平均是 $2.83，所以別想靠這個賺退休金。或者可以選擇此欄中的最後一格，C200，按住 Shift+Ctrl+↑，選取整欄，然後選擇試算表底端平均值（Average），察看簡單的統計數值（圖 1-3）。在 Windows，如果沒看到平均值，你需要在狀態列按右鍵，選擇平均值。在 Mac，如果狀態列並未顯示，按「檢視」選單，選擇「狀態列」即可開啟。

女生和女生跳舞，男生一邊涼快

K-means 叢集的目標，是在空間中找到一些點，放進 K 個群組之中（K 是你想區分的群組數量）。這 K 個群組，各自以其中心點為定義，類似於在月球上插根旗幟，喊「這裡是我的群組。如果你離這面旗子最近，就過來我這裡。」這個群組中心（正式名稱為叢集中心 cluster centroid），是一個平均值。

拿中學舞會來舉例。如果這讓你回想起中學舞會上的慘痛經驗，我在此致歉。

參加麥肯中學這場「海底世界」舞會的學生，分散在舞池之中，如圖 2-1。我還合成了一些地板花紋，幫助你融入情境。

如果你想順便配點音樂，以下是這場舞會可能播放的歌單：

- Styx: Come Sail Away
- Everything But the Girl: Missing
- Ace of Bass: All that She Wants
- Soft Cell: Tainted Love
- Montell Jordan: This is How We Do It
- Eiffel 65: Blue

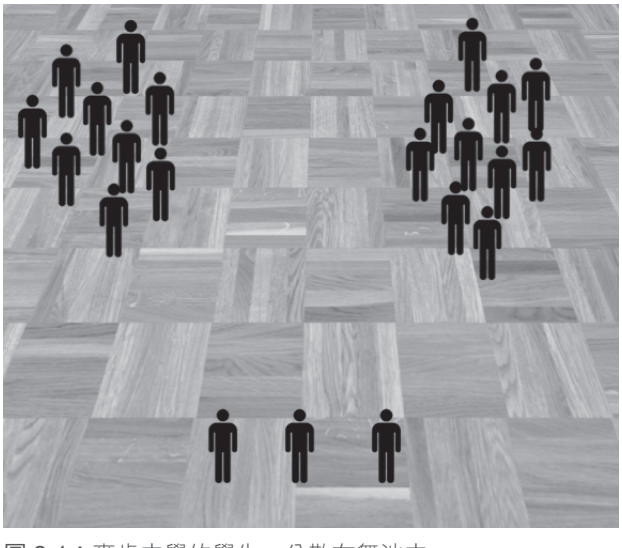
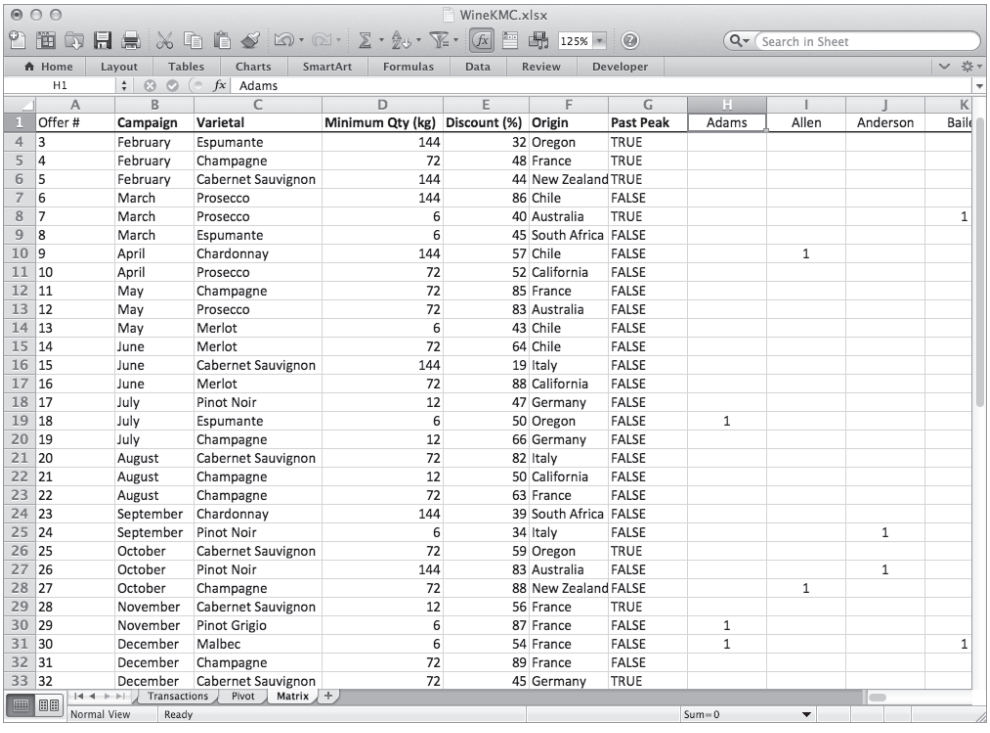


圖 2-1：麥肯中學的學生，分散在舞池中



WineKMC.xlsx

Home | Layout | Tables | Charts | SmartArt | Formulas | Data | Review | Developer

H1 | Adams

	A	B	C	D	E	F	G	H	I	J	K
1	Offer #	Campaign	Varietal	Minimum Qty (kg)	Discount (%)	Origin	Past Peak	Adams	Allen	Anderson	Baile
4	3	February	Espumante	144	32	Oregon	TRUE				
5	4	February	Champagne	72	48	France	TRUE				
6	5	February	Cabernet Sauvignon	144	44	New Zealand	TRUE				
7	6	March	Prosecco	144	86	Chile	FALSE				
8	7	March	Prosecco	6	40	Australia	TRUE				1
9	8	March	Espumante	6	45	South Africa	FALSE				
10	9	April	Chardonnay	144	57	Chile	FALSE		1		
11	10	April	Prosecco	72	52	California	FALSE				
12	11	May	Champagne	72	85	France	FALSE				
13	12	May	Prosecco	72	83	Australia	FALSE				
14	13	May	Merlot	6	43	Chile	FALSE				
15	14	June	Merlot	72	64	Chile	FALSE				
16	15	June	Cabernet Sauvignon	144	19	Italy	FALSE				
17	16	June	Merlot	72	88	California	FALSE				
18	17	July	Pinot Noir	12	47	Germany	FALSE				
19	18	July	Espumante	6	50	Oregon	FALSE	1			
20	19	July	Champagne	12	66	Germany	FALSE				
21	20	August	Cabernet Sauvignon	72	82	Italy	FALSE				
22	21	August	Champagne	12	50	California	FALSE				
23	22	August	Champagne	72	63	France	FALSE				
24	23	September	Chardonnay	144	39	South Africa	FALSE				
25	24	September	Pinot Noir	6	34	Italy	FALSE			1	
26	25	October	Cabernet Sauvignon	72	59	Oregon	TRUE				
27	26	October	Pinot Noir	144	83	Australia	FALSE			1	
28	27	October	Champagne	72	88	New Zealand	FALSE		1		
29	28	November	Cabernet Sauvignon	12	56	France	TRUE				
30	29	November	Pinot Grigio	6	87	France	FALSE	1			
31	30	December	Malbec	6	54	France	FALSE	1			1
32	31	December	Champagne	72	89	France	FALSE				
33	32	December	Cabernet Sauvignon	72	45	Germany	TRUE				

Transactions | Pivot | Matrix

Normal View | Ready | Sum=0

圖 2-10：把商品說明與購買資料合併成一個矩陣表格

資料標準化

在本章，資料的每一個維度，都是相同類型的二進位購買資料，不是 0 就是 1。然而許多叢集問題並非如此。試想這樣的情境，人們要依據身高、體重與薪資來劃分叢集。這三種資料的範圍都不同。身高可能從 150 公分到 200 公分，體重可能從 45 公斤到 150 公斤。

如此一來，測量顧客之間的距離（類似於舞池中的學生）就更顯困難。所已經常需要標準化處理每個資料欄，減去平均值，然後除以測量範圍，這會在第 4 章詳述，稱為標準差。如此可使每個欄的範圍相同，中心接近 0。

雖然我們在第 2 章所用的資料不需要標準化，但你可以在第 9 章的異常偵測中看到其作法。

如果在我的叢集裡，鄰居離我很近，但距離隔壁叢集的人就比較遠，那現在這些鄰居就是我的好鄰居，對吧？然而如果隔壁叢集的人，和我自己叢集的鄰居，距離差不多近呢？這樣一來，叢集分配就不恰當了，對吧？

這個值的正式描述為：

（我到最接近叢集中每個人的平均距離 - 我到所屬叢集中每個人的平均距離）/ 兩個平均值中較大者

由於算式中的分母，這個值會介於 -1 到 1 之間。

現在來思考這個公式。當隔壁叢集的人，距離我越來越遠（越來越不適合我），這個值就會趨近 1。如果兩個平均距離差不多呢？值會趨近 0。

將每一位顧客的此項計算結果平均後，就可以獲得側影值。如果側影值為 1，表示完美。如果為 0，表示極差。要是低於 0，表示大部分顧客應該換到其他叢集，這就太糟糕了。

對於不同的 K 值，你可以比較側影值，判斷是否有改善。

為了進一步瞭解這個概念，我們回到中學舞會的範例。圖 2-26 是構成側影的距離計算圖解。其中有一位工作人員，與另外兩位工作人員之間的距離，會拿來和次接近叢集的距離作比較，而次接近叢集就是一群高中男生。

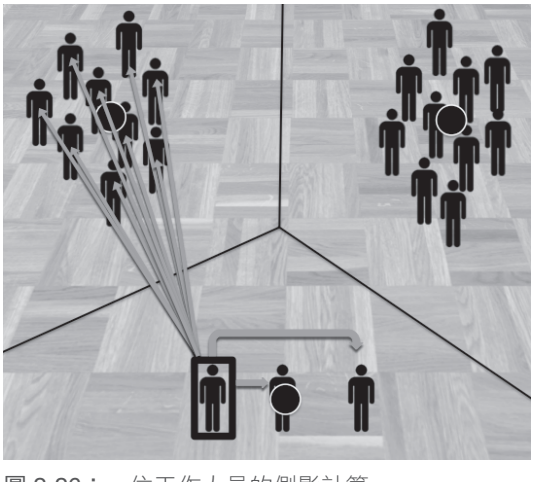


圖 2-26：一位工作人員的側影計算

計算 5-Means 叢集的側影值

你可能會想，五叢集是否優於四叢集。一眼望去，差異似乎不大。我們來計算五叢集的側影值，看看電腦有何想法。

先複製 4MC Silhouette，改名為 5MC Silhouette。然後右鍵點選 G 欄，插入一欄，命名為 Distance From People in 5。把 F2 的公式拖到 G2，將要檢查的叢集從 4 改成 5，然後雙擊儲存格，向下複製。

與前一節作法相同，你要尋找 '4MC'!$L$39:$DG$39，取代成 '5MC'!$M$40:$DH$40。

在儲存格 H2、I2 與 J2，你應當在算式中加上到 Cluster 5 成員的距離，所以原本只到 F2 的範圍，要擴展到 G2。然後選取 H2:J2 範圍，雙擊右下角，把更新後的算式向下複製。

最後，你還需要從 5MC 的 40 列，將叢集分配值選擇性貼上 至 5MC Silhouette 標籤的 B 欄。換言之，你在選擇性貼上視窗中，必須勾選「轉置」。

試算表修改完成後，應當如圖 2-40。

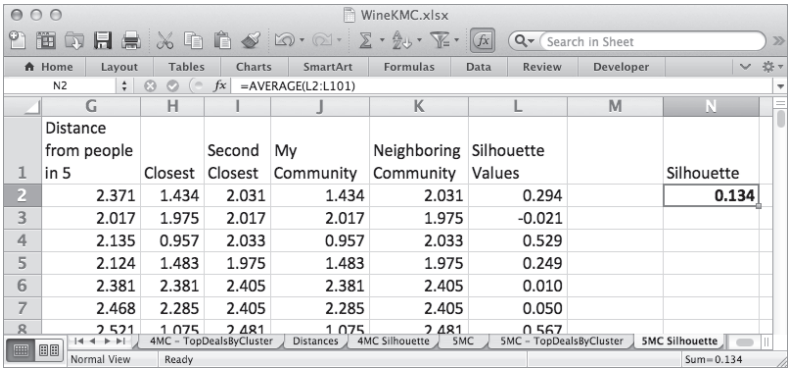


圖 2-40：5-means 叢集的側影值

有點糟，是不是？側影值沒什麼變化。事實上，0.134，變糟了！然而觀察叢集後，並不意外。兩種叢集分配，其中都有三個叢集是有道理的。其他叢集就不知所云。也許你應該換個方向，試試看 K=3？如果你願意嘗試，就交給你自己練習啦。

我們在此來多想一想，到底是什麼問題，造成這樣難以理解的叢集。

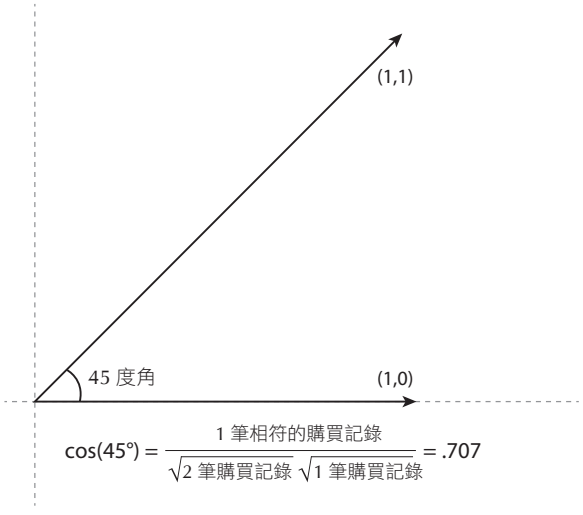


圖 2-41：兩個二進位購買向量的餘弦相似度

兩者之間的餘弦相似度，可以說是 cos(45 度) = 0.707。但這是為什麼？其實，兩個二進位購買向量之間的角度，餘弦值等於：

兩個向量中相符的購買資料相加，除以第一個向量的購買數量平方根乘以第二個向量的購買數量平方根。

在此範例中，兩個向量 (1,1) 與 (1,0)，各有一筆相符的購買資料，所以算式是 1 除以 2 的平方根（購買商品數量為 2），乘以 1（購買商品數量為 1）。答案是 0.707 (見圖 2-41)。

這個算式好在哪裡？

有三個原因：

- 算式中的分子，僅計算相符的購買次數，所以這是一個非對稱的測量值，正是你所需要的。
- 將兩個向量中的購買商品數，取平方根後做為分母，你可以計算出每一件商品被購買的向量值，稱為交互購買向量（promiscuous purchase vector），讓僅購買部分商品的向量值距離拉遠。你要比對找出品味相同的向量，而非找出符合其他向量品味的向量。

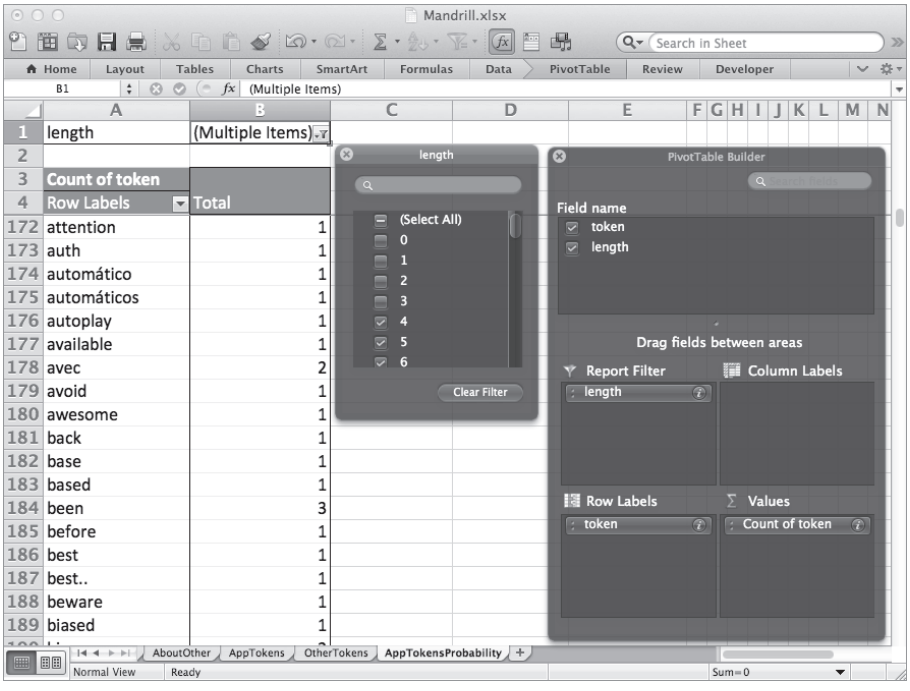


圖 3-9：計算符號數量的樞紐分析表工具設定

既然所有數值都加 1，你也需要一個新的符號計數。所以在表格底端（AppTokensProbability 標籤的 828 列），設定一個儲存格，對上方累計加總。老問題，如果你用 Windows，所有範圍都要往上挪一列（加總範圍為 C4:C826）：

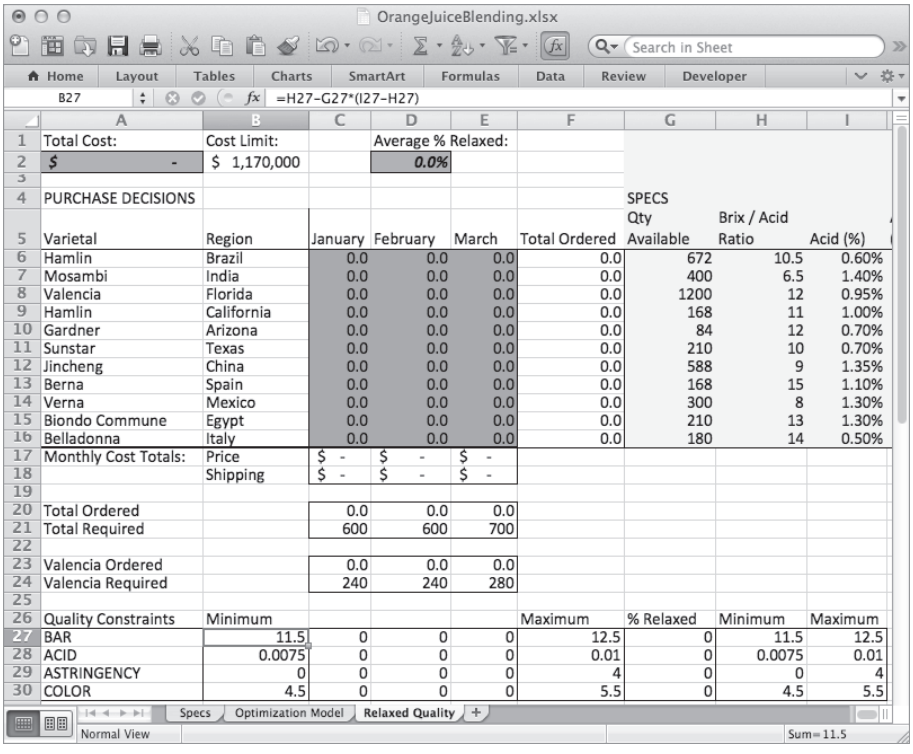
```
=SUM(C5:C827)
```

在 D 欄，你可以計算每一個符號的機率，把 C 欄計算的出現次數，除以全部的符號出現次數。D 欄標記為 P（Token|App）。第一個符號位於 D5（Windows 版為 D4），機率為：

```
=C5/C$828
```

請注意符號總出現次數的絕對連結。你可以雙擊此公式，沿 D 欄向下複製。然後在 E 欄（標記為 LN(P)），對於 D5 中的機率取自然對數：

```
=LN(D5)
```



	A	B	C	D	E	F	G	H	I
1	Total Cost:	Cost Limit:		Average % Relaxed:					
2	$ -	$ 1,170,000		0.0%					
3									
4	PURCHASE DECISIONS						SPECS		
5	Varietal	Region	January	February	March	Total Ordered	Qty Available	Brix / Acid Ratio	Acid (%)
6	Hamlin	Brazil	0.0	0.0	0.0	0.0	672	10.5	0.60%
7	Mosambi	India	0.0	0.0	0.0	0.0	400	6.5	1.40%
8	Valencia	Florida	0.0	0.0	0.0	0.0	1200	12	0.95%
9	Hamlin	California	0.0	0.0	0.0	0.0	168	11	1.00%
10	Gardner	Arizona	0.0	0.0	0.0	0.0	84	12	0.70%
11	Sunstar	Texas	0.0	0.0	0.0	0.0	210	10	0.70%
12	Jincheng	China	0.0	0.0	0.0	0.0	588	9	1.35%
13	Berna	Spain	0.0	0.0	0.0	0.0	168	15	1.10%
14	Verna	Mexico	0.0	0.0	0.0	0.0	300	8	1.30%
15	Biondo Commune	Egypt	0.0	0.0	0.0	0.0	210	13	1.30%
16	Belladonna	Italy	0.0	0.0	0.0	0.0	180	14	0.50%
17	Monthly Cost Totals:	Price	$ -	$ -	$ -				
18		Shipping	$ -	$ -	$ -				
19									
20	Total Ordered		0.0	0.0	0.0				
21	Total Required		600	600	700				
22									
23	Valencia Ordered		0.0	0.0	0.0				
24	Valencia Required		240	240	280				
25									
26	Quality Constraints	Minimum				Maximum	% Relaxed	Minimum	Maximum
27	BAR	11.5	0	0	0	12.5	0	11.5	12.5
28	ACID	0.0075	0	0	0	0.01	0	0.0075	0.01
29	ASTRINGENCY	0	0	0	0	4	0	0	4
30	COLOR	4.5	0	0	0	5.5	0	4.5	5.5

圖 4-25：品質寬鬆模型

開啟規劃求解，改變目標式為儲存格 D2 的品質寬鬆平均值，取最小值。你也需要在決策變數表中加入 G27:G30，設定 A2 的成本為小於或等於 B2 的上限。新的公式如圖 4-26。你已經把先前的成本目標式，變成了限制式。你也把原本對於品質的限制，變成了可以用 G27:G30 調整的彈性限制。你的目標放在 D2，是使品質降低程度的平均值最小化。按下求解。

若上下兩端平均寬鬆 35%時，Excel 可以找到符合成本限制的解答，如圖 4-27。

- 如果開關與採購數量都是 0，乘積為 0。
- 如果你採購了一些果汁，但不打算去酸，則乘積還是 0。
- 如果你選擇去酸，乘積就是採購果汁數量。

在各種狀況中，可以去酸的果汁數量，受限於去酸開關變數乘以可採購的果汁總數。如果不去酸，這個上限就是零。如果要去酸，上限就會變成可購買的最大數量。這是一個「大 M」限制式，與前一節類似。

到了巴西 Hamlin 柳橙，「大 M」限制式的計算是把儲存格 C26 的開關，乘以儲存格 G6 的可購買數量，672,000 加侖。將此算式放在 G26 開關變數旁，然後將其複製到其他月份、其他產地。

現在的試算表如圖 4-35。

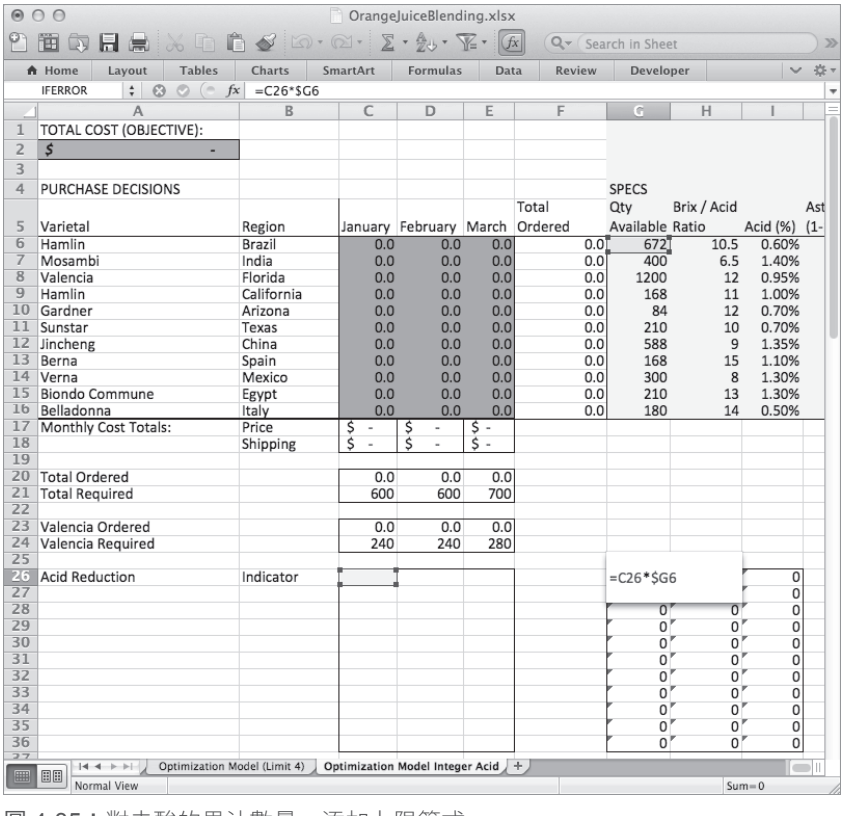


圖 4-35：對去酸的果汁數量，添加上限算式

很有可能，你從埃及訂購的 Biondo Commune 柳橙汁，事實上甜度 / 酸度比例並非恰好 13。這是你期待的數字，但實際上可能有所偏差。這個偏差空間，往往可以用**機率分布**來描述。

機率分布，簡單來說，是大致說明某種狀況的每一種可能結果，而所有的機率相加會等於 1。最出名、使用最廣泛的分布是常態分布，也稱為「鐘形曲線」。鐘形曲線經常被採用，是因為當你有大量獨立、複雜、真實的要素，造成隨機分布的資料時，資料**經常**是常態分布，或者鐘形曲線。這稱之為**中心極限定理**（central limit theorem）。

我們做個小實驗來看看。拿出你的手機，把每個聯絡人的電話號碼後四碼找出來。第一位數字，可能是**均勻分布**在 0 到 9 之間，換言之每一個數字出現的頻率都差不多。第二位、第三位與第四位也都是同樣狀況。

現在，我們把這四個「隨機變數」加總。可能的最低數字是 0 (0 + 0 + 0 + 0)。最高是 36 (9 + 9 + 9 + 9)。獲得 0 與 36，各只有一種狀況。獲得 1 與 35，各有四種狀況，獲得 20 的狀況就多到數不清了。所以如果你取的電話號碼夠多，將這些數字總和製作成圖表，就會獲得如圖 4-40 的鐘形曲線（我這張圖中用了 1,000 組電話號碼，你看我人緣多好）。

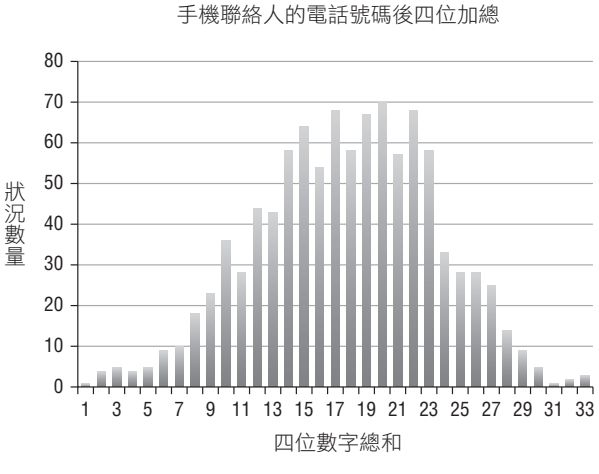


圖 4-40：組合獨立隨機變數，成為鐘形曲線

換言之，你從埃及拿到的這批果汁，很可能甜度 / 酸度比例為 13.5，但不太可能是 10。

計算樣本平均值與標準差

如果你以前沒有算過標準差，有興趣知道要怎麼計算，這再簡單不過。

圖 4-42 是埃及 Biondo Commune 柳橙汁的過去 11 筆訂單，每一批的甜度 / 酸度比例列於 B 欄。這些數值的平均值是 13，如原始試算表上的規格。

標準差，就只是與平均值之間的誤差，取平方根。所謂的「誤差」，是指每一筆訂單與期望值 13 之間的差距。

在圖 4-42 的 C 欄中，可以看到誤差算式，D 欄則是誤差的平方。平均誤差平方是 AVERAGE(D2:D12)，等於 0.77。把平均誤差平方開根號，得到 0.88。超簡單！

然而就現實而言，對少量訂單計算樣本標準差時，如果把誤差平方相加，除以總訂單數減去 1（此例中用 10 代替 11），會比較好。

如果改用此算法，標準差變成 0.92，如圖 4-42。

	A	B	C	D	E	F
1	Order	BAR	Error	Squared Error		MEAN
2	1	14	1	1		**13**
3	2	13	0	0		
4	3	13	0	0		Mean Squared Error
5	4	13.5	0.5	0.25		**0.77**
6	5	14	1	1		Standard Deviation
7	6	13	0	0		**0.88**
8	7	12.5	-0.5	0.25		
9	8	11	-2	4		Sum Squared Error / N-1
10	9	13	0	0		**0.85**
11	10	12	-1	1		Adjusted Standard Deviation
12	11	14	1	1		**0.92**

圖 4-42：標準差計算範例

在混合例題中，從標準差產生的情境

NOTE

如前一節所述，使用 Excel 2010 與 Excel 2013 的人，必須安裝 OpenSolver。只需正常設定，然後在求解時，改用「OpenSolver Solve」按鈕。詳見第 1 章關於 OpenSolver 的說明。