第2章資料分析的基礎

Basics of Data Analysis

1. 基本統計學（Basic Statistics）
2. Excel圖表製作（一）（Graphing/Tabling Using Excel）

第2.1節 基本統計學
Basic Statistics

2.1-1 機率與機率密度（Probability/Probability Density）

2.1-2 平均值與標準偏差（Average/Standard Deviation）

2.1-3 自由度（Degrees of Freedom）

2.1-4 常態分佈（Normal Distribution）

2.1-5 常態分佈曲線的應用：不良率（Defect Rate）

2.1-6 標準常態分佈（Standard Normal Distribution）

狀況描述（Scenario）

表2.1-1 量測數據（單位μm）

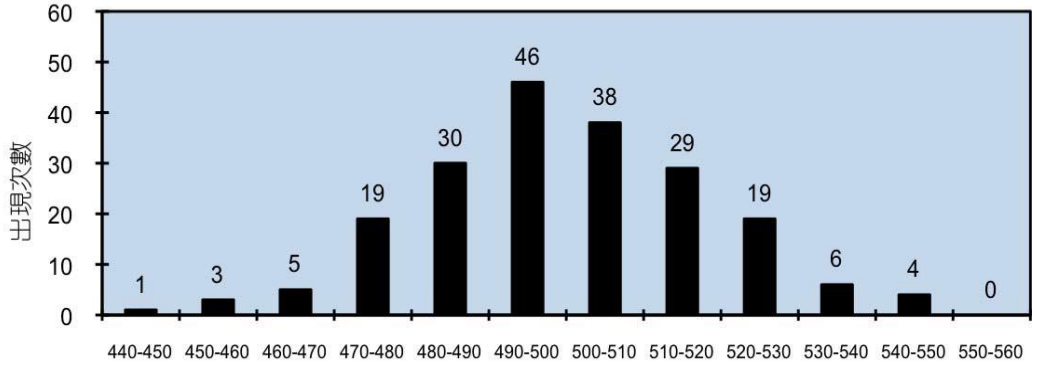
487	508	540	478	516	532	489	478	484	493
526	520	448	496	512	486	489	509	514	520
527	485	467	540	527	483	492	488	525	478
479	497	510	544	498	499	499	491	501	517
452	508	492	509	511	510	524	461	503	498
485	500	476	493	503	507	500	503	472	493
463	463	490	508	517	482	482	492	497	518
500	524	498	491	498	488	518	507	527	543
515	470	522	506	501	516	476	476	506	494
480	495	506	482	493	494	479	495	508	506
482	487	480	481	505	495	499	500	527	524
510	515	524	503	499	476	488	503	476	513
482	527	505	504	526	493	503	519	514	498
534	481	497	500	492	493	485	530	500	502
474	524	471	541	490	489	502	542	485	490
511	458	506	512	530	500	519	527	498	489
476	510	489	498	485	515	483	523	474	471
503	517	490	498	520	493	510	511	501	518
488	510	521	515	492	492	457	536	513	506
474	490	496	497	492	478	517	504	509	504
Average = 500.0 μm, Standard Deviation = 18.6 μm									

Go to 2.1-2

2.1-1 機率與機率密度
(Probability/Probability Density)

表2.1-2 數據的分佈

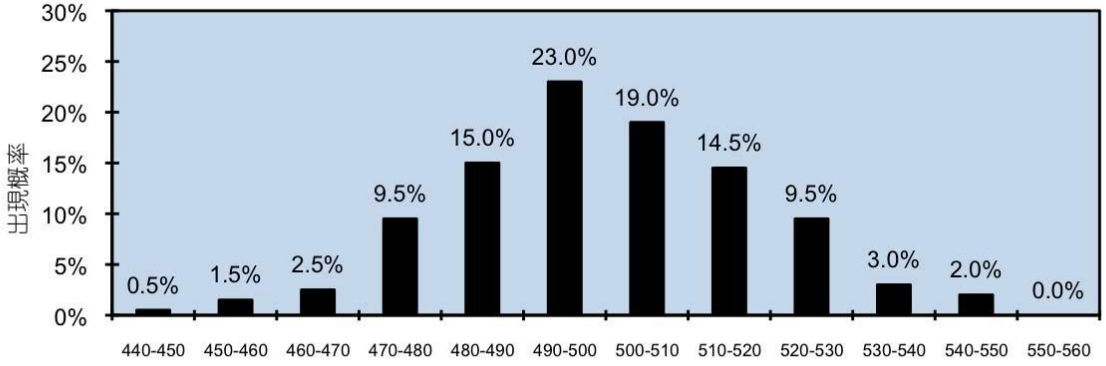
440-450	450-460	460-470	470-480	480-490	490-500	500-510	510-520	520-530	530-540	540-550	550-560
1	3	5	19	30	46	38	29	19	6	4	0



機率與機率密度

表2.1-2 數據的分佈

440-450	450-460	460-470	470-480	480-490	490-500	500-510	510-520	520-530	530-540	540-550	550-560
1	3	5	19	30	46	38	29	19	6	4	0

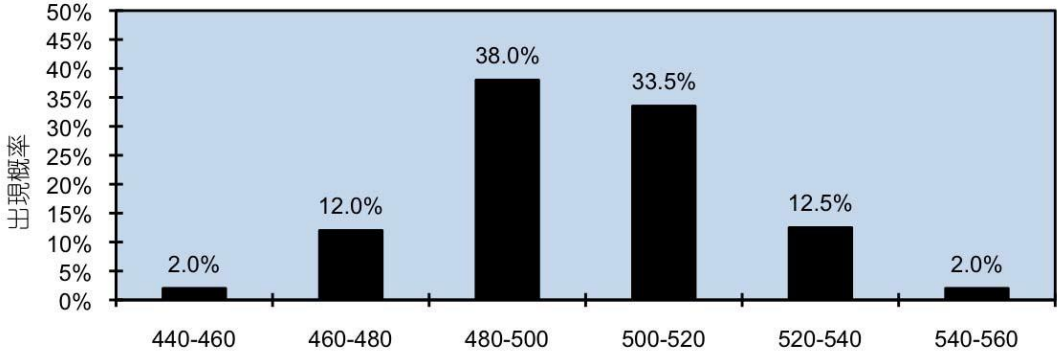


機率與機率密度

表2.1-2 數據的分佈

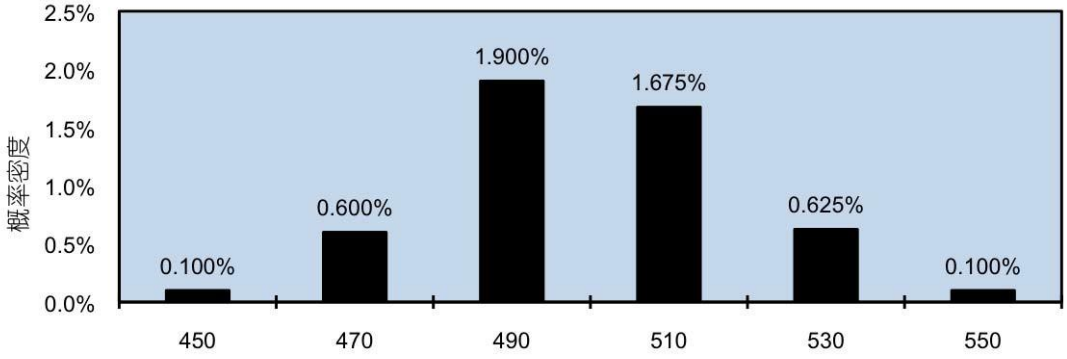
440-450	450-460	460-470	470-480	480-490	490-500	500-510	510-520	520-530	530-540	540-550	550-560
1	3	5	19	30	46	38	29	19	6	4	0

- 每一條方塊的縱軸值（出現機率）與橫軸所取的間隔有相關嗎?
- 譬如，如果間隔取20 μm來統計，則其結果為?



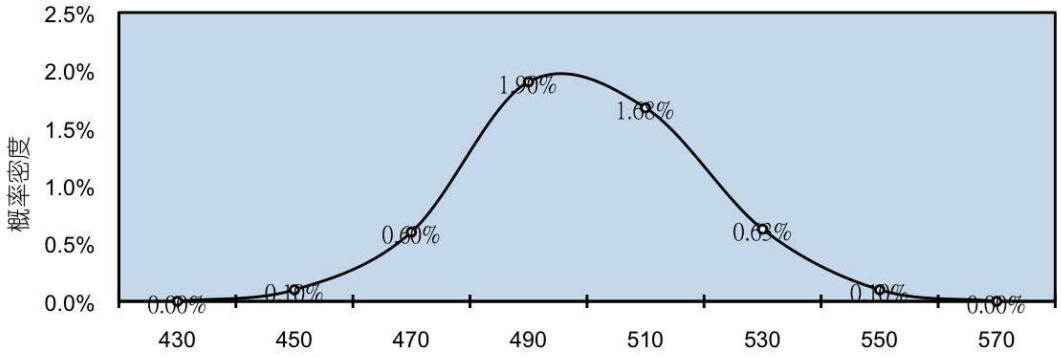
機率與機率密度

- 為了讓縱軸的值不因橫軸的取樣間隔而改變，該怎麼辦?
- 以「平均出現在一個橫軸單位範圍內的機率」（稱為機率密度，probability density）來表示縱軸的值。



機率與機率密度

- 用平滑曲線來表示，如下圖所示。
- 這個平滑曲線稱為「機率密度分佈曲線」。
- 機率密度分佈曲線下的面積代表機率。
- 整個機率密度分佈曲線下的面積為?



2.1-2 平均值與標準偏差 (Average/Standard Deviation)

- 如何以最少的數字來代表一群數據？
- 最常使用的方法是利用平均值及標準偏差來代表一群數據。
- 平均值代表? 資料的重心位置，亦即分佈曲線下面積的重心位置。
- 標準偏差代表? 資料的散佈情形，亦即分佈曲線的扁平程度。

平均值與標準偏差

- 假設y_i $(i = 1,2, \dots, n)$ 代表 n 個獨立的數據（獨立的意思是它們之間沒有任何相依關係）。
- 它們有幾個自由度?
- 平均值定義為

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \qquad \text{（2.1-1式）}$$

平均值與標準偏差

- 每一個資料偏離此平均值的量是

$$(y_i - \bar{y}), i = 1,2, \dots, n$$

- 此 n 個偏離量是完全獨立的嗎? 理由是?

$$\sum_{i=1}^{n} (y_i - \bar{y}) = 0 \qquad (2.1\text{-}2式)$$

- 所以，此 n 個偏離量有 n-1 個自由度。

平均值與標準偏差

- 此 n 個偏離量的平方和、除以它們的自由度，稱為這 n 個資料的變異數（variance）

$$V = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} \qquad (2.1\text{-}3式)$$

- 標準偏差則定義為變異數的平方根

$$S = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} \qquad (2.1\text{-}4式)$$

- 標準偏差和資料量測值的單位相同嗎? 理由是?

平均值與標準偏差

- 本例中，平均值與標準偏差分別為

$$\bar{y} = 500\ \mu\text{m}, \qquad S = 18.6\ \mu\text{m}$$

- 使用表格、直條圖、分佈圖、上述兩個數字來代表玻璃厚度的統計資訊，哪一種比較簡潔?
- 平均值表示：資料的重心位置，亦即分佈曲線下面積的重心位置。
- 標準偏差表示：資料的散佈情形，亦即分佈曲線的扁平程度。

均方偏差及母體標準偏差

- 此外，我們定義另外一個量

$$S_p = \sqrt{\frac{\sum_{i=1}^{n}(y_i - m)^2}{n}} \qquad (2.1\text{-}5式)$$

- 當 m 為已知的目標值時，2.1-5式稱為根均方偏差（root mean square deviation, RMSD），用來表示?「平均偏離目標值的量」
- 2.1-5式根號內部份則稱為均方偏差（mean square deviation, MSD）。
- 當 m 為母體資料的平均值時，2.1-5式稱為母體標準偏差（population standard deviation）。

另一個標準偏差

- 2.1-4式中，如果將分母改為 n（而非 n-1），我們稱之為S_n：

$$S_n = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}} \qquad \text{（2.1-6式）}$$

- S_n 與 S_p 的關係： S_n可視為將 S_p中的母體平均值 m 以樣本平均值 $\bar{y}$ 取代而得。
- 當我們要表示 n 個數據的散佈程度時，使用2.1-4式或2.1-6式都是可以的，理由是? 使用任何一個所獲得的結論應該是一致的
- 工程實務上，分母是 n-1 或是 n，(1)常常不用太在乎，或(2)很在乎?

2.1-3 自由度（Degrees of Freedom）

- 一組數據的自由度是指這組數據最多的獨立資訊個數。
- 假設你量了A, B, C三個人的身高分別是178 cm, 171 cm, 167 cm。
- 這些數據可以導出其它許多資訊，譬如「A比B高7 cm」、「B比C高4 cm」、「三個人平均172 cm」、「A和B加起來比C高182 cm」等。
- 以上所有資訊中，獨立的有幾個? 理由是?
- 3個是獨立的，其它都可以由這3個獨立的資訊導出。
- 所以，上述提到的所有數據（7個數據）有3個自由度。
- 一般而言，如果獨立地量測了 n 個數據，則這 n 個數據的自由度是 n。

自由度

- 自由度的觀念常被用來解說直交表的橫列數目（實驗數據個數）與直行（所能提供的資訊個數）之間的關係。
- 表1.2-5最右行的18個S/N比，可以導出表1.2-6的一大堆資訊（包括23個反應值及15個因子效應）。
- 上述18個S/N比的自由度是?
- 表1.2-6最多可以有幾個獨立資訊?
- 表1.2-6中的非獨立資訊如何產生? 互相計算出來的

表1.2-6 *S/N*比的因子反應表

	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*
Level 1	43.1	40.5	40.5	40.3	44.5	41.1	40.4	39.9
Level 2	39.5	41.2	41.0	40.9	40.1	41.4	41.5	42.8
Level 3		42.2	42.5	42.7	39.3	41.4	42.0	41.2
$E^{1\to2}$	-3.6	0.7	0.5	0.6	-4.4	0.3	1.0	2.9
$E^{2\to3}$		0.9	1.5	1.8	-0.8	0.0	0.5	-1.6
Range	3.6	1.6	2.1	2.4	5.3	0.3	1.6	2.9
Rank	2	6	5	4	1	8	7	3
Significant?	yes	no	yes	yes	yes	no	no	yes

自由度

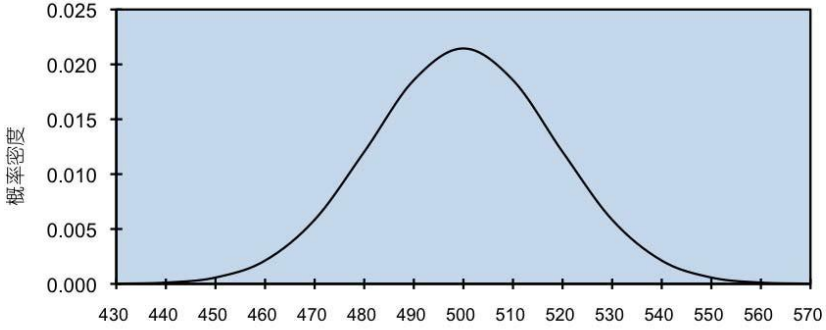
- 非陰影部份的數據與陰影部份的數據，其關係為何?
- 考慮陰影部份的23個數據及總平均值41.3，總共24個數據。
- A因子的兩個反應值有幾個自由度? 為什麼?
- B因子的三個反應值有幾個自由度? 為什麼?
- 這24個數據的總自由度是?
  - 1（A因子）+ 2 x 7（B, C, D, E, F, G, H）+ 1（總平均值41.3）= 16

表1.2-6 *S/N*比的因子反應表

	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*
Level 1	43.1	40.5	40.5	40.3	44.5	41.1	40.4	39.9
Level 2	39.5	41.2	41.0	40.9	40.1	41.4	41.5	42.8
Level 3		42.2	42.5	42.7	39.3	41.4	42.0	41.2
$E^{1\to2}$	-3.6	0.7	0.5	0.6	-4.4	0.3	1.0	2.9
$E^{2\to3}$		0.9	1.5	1.8	-0.8	0.0	0.5	-1.6
Range	3.6	1.6	2.1	2.4	5.3	0.3	1.6	2.9
Rank	2	6	5	4	1	8	7	3
Significant?	yes	no	yes	yes	yes	no	no	yes

2.1-4 常態分佈（Normal Distribution）

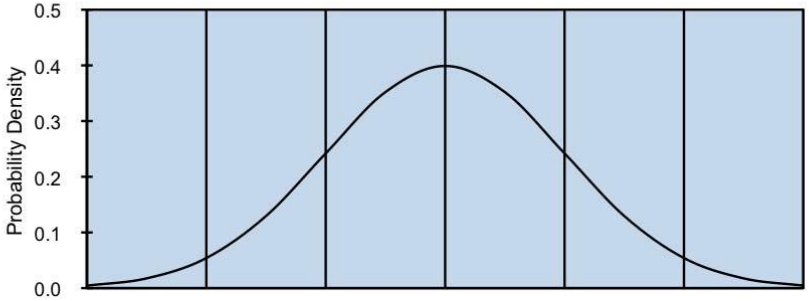
- 樣本夠多的話，則其機率密度分佈圖會呈現何種樣貌?
  - 呈現以平均值為中心的對稱鐘形分佈，左右各有一個反曲點，它們的位置大約在橫座標值等於 $\bar{y} \pm S$ 的地方。
  - 在這兩個反曲點之間的分佈曲線下面積大約是68%。Why?
  - 橫座標值等於 $\bar{y} \pm 2S$ 的之間的分佈曲線下面積大約是95%。Why?
  - 橫座標值等於 $\bar{y} \pm 3S$ 的之間的分佈曲線下面積非常接近100%。Why?



常態分佈

- 是否可能用一個數學函數來表示一常態機率密度分佈曲線，而此數學函數只含平均值及標準偏差兩個參數？
- Gaussian分佈（常態分佈）

$$f(y) = \frac{1}{\sqrt{2\pi}S} \exp\left(\frac{-(y-\bar{y})^2}{2S^2}\right) \qquad \text{（2.1-7式）}$$



2.1-5 常態分佈曲線的應用：不良率（Defect Rate）

- 在Excel中有兩個與常態分佈有關的函數

$$\mathrm{NORMDIST}(y, \bar{y}, S, Cumulative) \qquad (2.1\text{-}8式)$$

- NORMDIST函數會傳回常態分佈曲線下橫座標值小於 y 的面積（亦即機率）或 縱座標值，依 Cumulative 而定：
- TRUE時，傳回的是「分佈曲線下橫座標值小於 y 的面積」，亦即機率。
- FALSE時，傳回的是「座標值 y 所相對的縱座標值」，亦即機率密度。

$$\mathrm{NORMINV}(p, \bar{y}, S) \qquad (2.1\text{-}9式)$$

- NORMINV函數則是傳回面積（機率）是 p 時所相對的橫座標值。

常態分佈曲線的應用：不良率

- 本例中，分佈曲線下橫座標值介於440 μm與560 μm之間的面積如何計算?

NORMDIST(560, 500, 18.6, TRUE)
– NORMDIST(440, 500, 18.6, TRUE) = 99.874%

- 這個值代表良率（yield rate），不良率（defect rate）則是

1 – 99.874% = 0.126%

常態分佈曲線的應用：不良率

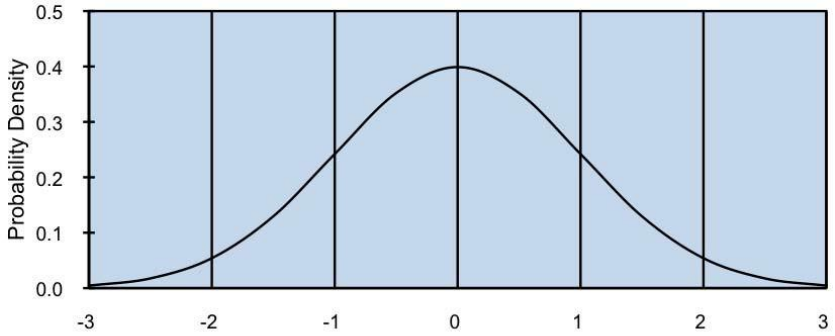
- 以上不良率的計算是基於下列幾個假設：
  - 玻璃厚度是呈常態分佈的。
  - 抽樣檢測是隨機的。
  - 這200個抽樣數據的標準偏差及平均值可以代表母體的標準偏差及平均值。Why?
- 這些假設對本例而言都是可以接受的。
- 當樣本不是那麼多時，第3個假設是值得懷疑的。

2.1-6 標準常態分佈 （Standard Normal Distribution）

- 當身邊沒有電腦、或沒有Excel這一類程式的時，要知道分佈曲線下的面積時，通常需要?
- 要決定分佈曲線下的面積，至少必須有3個參數：平均值、標準偏差、及此橫座標值 y。
- 如此，需要製作的表有多少? 可行嗎?
- 一個可行的做法是只製作一個標準常態分佈（standard normal distribution）的圖表供查閱，經適當的轉換，我們可以獲得所需的值。

標準常態分佈

- 所謂標準常態分佈是指平均值是0，而標準偏差是1的常態分佈。
- 表B.2-1列出標準常態分佈曲線下，橫座標值小於 Y 所相對的面積（機率p）。
- 表B.2-2則列出機率 p 時所相對的橫座標值 Y。
- 在標準常態分佈中，我們用 Y 來代表橫座標，而在一般常態分佈中，我們用 y 來代表橫座標。



Y 和 y 之間的轉換

- Y 和 y 可以經下列的簡單公式互相轉換：

$$Y = \frac{y - \bar{y}}{S} \qquad (2.1\text{-}10式)$$

或

$$y = \bar{y} + YS \qquad (2.1\text{-}11式)$$

- 要查閱某 y 值所相對的機率 p 時：
  - 只要轉換成 Y 值，再由表B.2-1即可查到 Y 所相對的機率 p。
- 若我們要查閱某 p 值所相對的橫座標 y 時：
  - 只要先由表B.2-2查到 p所相對的橫座標 Y 值，再轉換成 y 值。
- 必要的話，我們可以進行線性內插（linear interpolation）的計算。

實例：玻璃厚度不良率的計算

- 規格是500±60 μm，平均值是500 μm，標準偏差是18.6 μm。
- y = 440，其相對的 Y 值是?

- 由表B.2-1查得在 Y = -3.22時，相對的機率 p = ?；在 Y = -3.24時，相對的機率 p = ?。
- Y = -3.226時所相對的 p 值，如何計算?

- 因為玻璃厚度分佈曲線是對稱於目標值，所以不良率是?