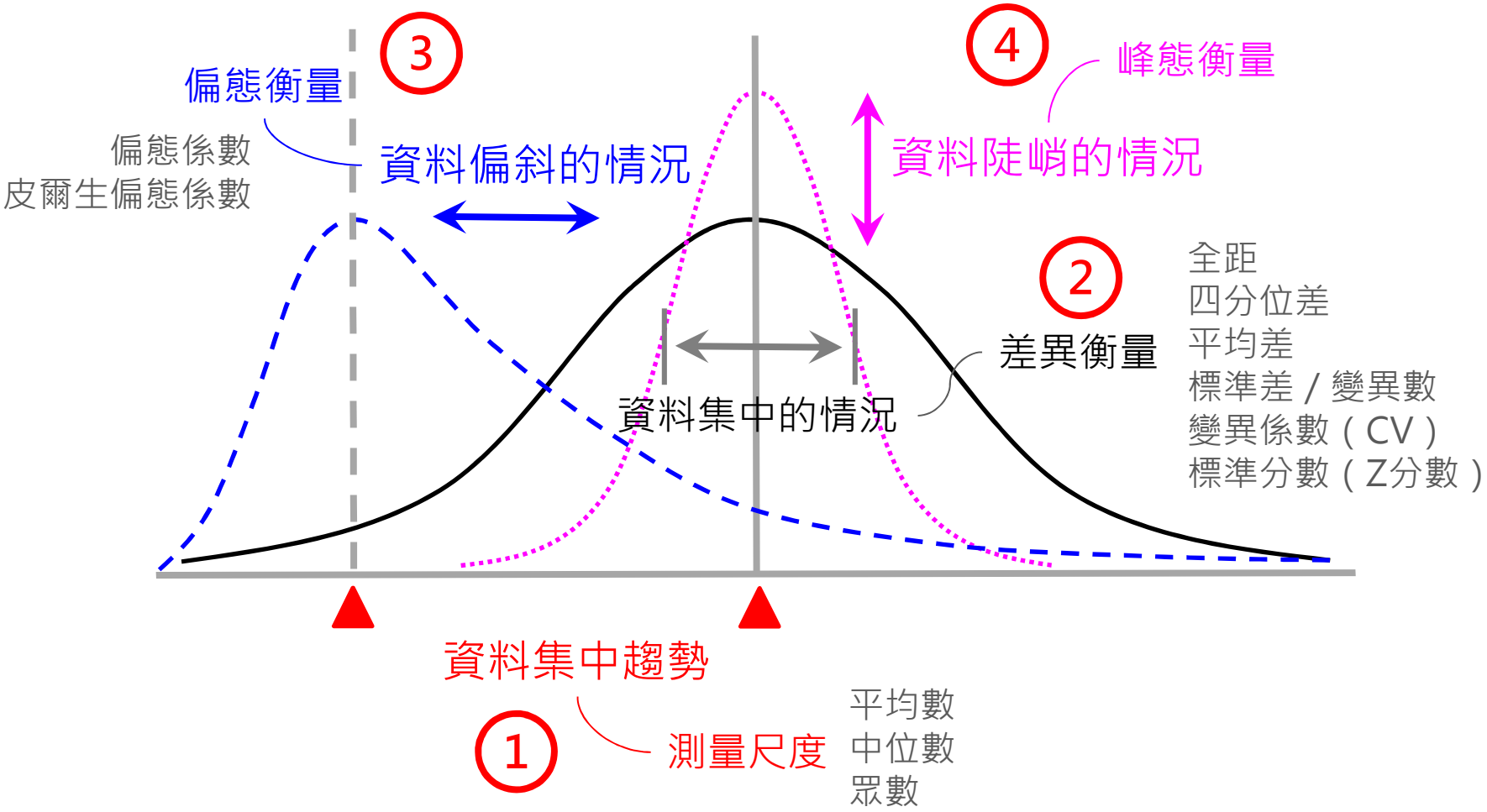
資料的整理與表現

統計測量數

主要統計測量數類型


3
偏態衡量
偏態係數
皮爾生偏態係數
資料偏斜的情況
4
峰態衡量
資料陡峭的情況
2
差異衡量
全距
四分位差
平均差
標準差 / 變異數
變異係數（CV）
標準分數（Z分數）
資料集中的情況
資料集中趨勢
1
測量尺度
平均數
中位數
眾數

測量尺度

資料中央趨勢、資料的代表

中位數（median）

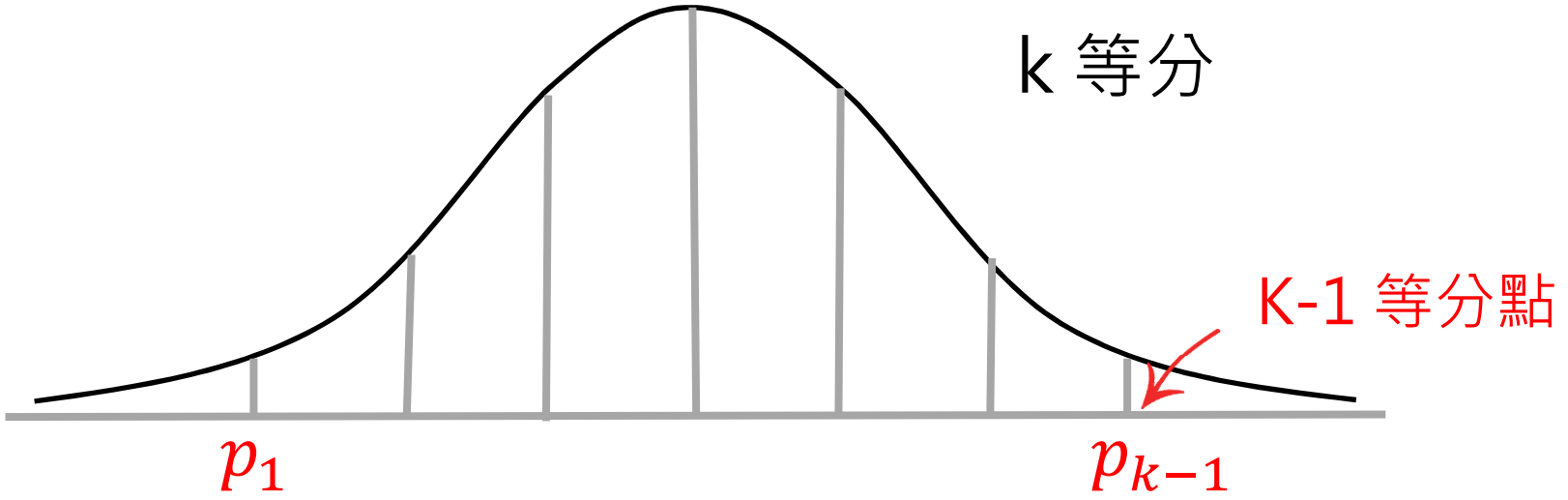
中位數是指將數據按大小順序排列起來，形成一個數列，居於數列中間位置的那個數據。
中位數用 Me 表示。

$$M_e = \begin{cases} \dfrac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}, & n\text{為偶數} \\ x_{(\frac{n+1}{2})}, & n\text{為奇數} \end{cases}$$

特點：
(1)不受分佈數列的極大或極小值影響
(2)當次數分佈偏態時，中位數的代表性會受到影響。
(3)缺乏敏感性。

k 分位數

k 分位數是中位數的延伸，是指將一組數據按大小順序排列後，分成 k 等分形成一個數列。



1.百分位數(percentile)：將資料分割成100等分；通常以P_i表示

2.十分位數(deciles)：將資料分割成10等分；通常以D_i表示

3.四分位數(quartile)：將資料分割成4等分；通常以Q_i表示

4分位數

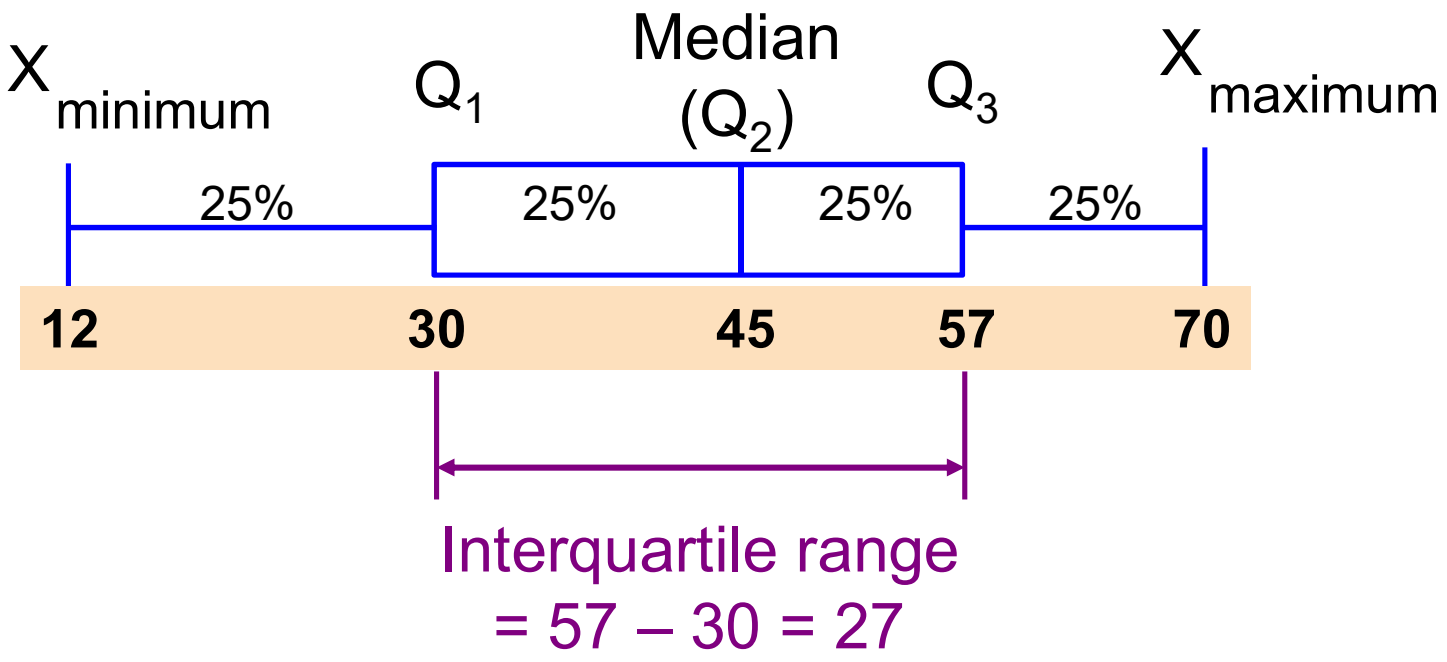
$$Q_1 = \begin{cases} \dfrac{x_{(\frac{n}{4})} + x_{(\frac{n}{4}+1)}}{2}, & \dfrac{n}{4} \text{ 為整數} \\ x_{(\frac{n}{4}+1)}, & \dfrac{n}{4} \text{ 不為整數} \end{cases}$$

$$Q_2 = M_e$$

$$Q_3 = \begin{cases} \dfrac{x_{(\frac{3n}{4})} + x_{(\frac{3n}{4}+1)}}{2}, & \dfrac{3n}{4} \text{ 為整數} \\ x_{(\frac{3n}{4}+1)}, & \dfrac{3n}{4} \text{ 不為整數} \end{cases}$$

Calculating The Interquartile Range

Example:



眾數（mode）

眾數是指一組數據中出現次數最多的那個數據 M_o
一組數據可以有多個眾數，也可以沒有眾數。

特性：

(1)不受分佈數列的極大或極小值的影響
(2)缺乏敏感性。
(3)一組數據中的眾數可能不存在。
(4)眾數粗糙，但眾數不受個別數據的影響，可在數據缺陷較大或需要快速而粗略地尋求一組數據的代表值時用。

算術平均數（arithmetic mean）

$$\bar{X} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_1 = \frac{\sum x}{n}$$

$$\bar{X} = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_n f_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum xf}{\sum f}$$

特性：

(1)計算簡單

(2)較沒有考慮到近期的變動趨勢，因而預測值與實際值往往會發生較大的誤差。

(3)通常適用於預測銷售比較穩定的產品。

加權平均數

$$\text{加權平均數} = \frac{\sum_{i=1}^{k} w_i x_i}{\sum_{i=1}^{k} w_i}$$

w_i ：權重（weighting）

特性：

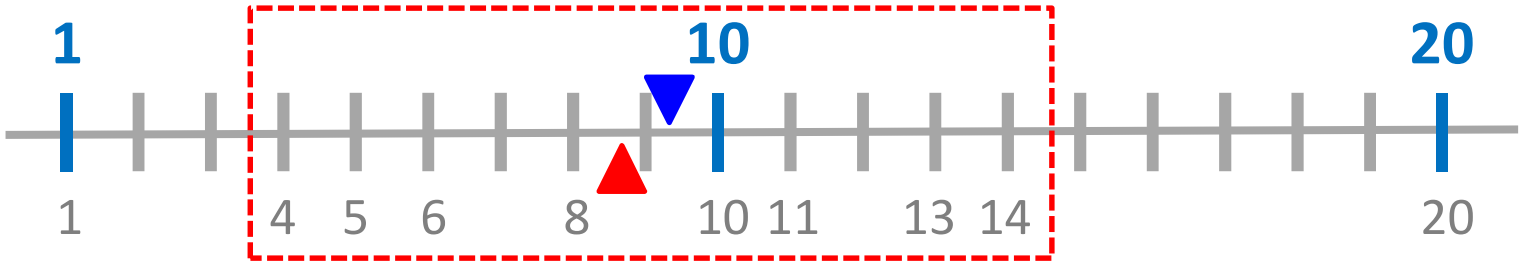
(1)為算術平均數的延伸

(2)計算原理與算術平均數相同

(3)常用於物價指數、股票加權指數、學術平均成績等

剪尾平均數(trimmed mean) $\overline{x}_a$

假設有一組10個數字，若用直線表示，其各數據的值對應如下。求這10組數字的平均數與10%（a=0.1）的剪尾平均數（$\overline{x}_{0.1}$）：



平均數： $\bar{x}$ =(1+4+5+6+8+10+11+13+14+20)/10 = 9.2

10%的剪尾平均數：

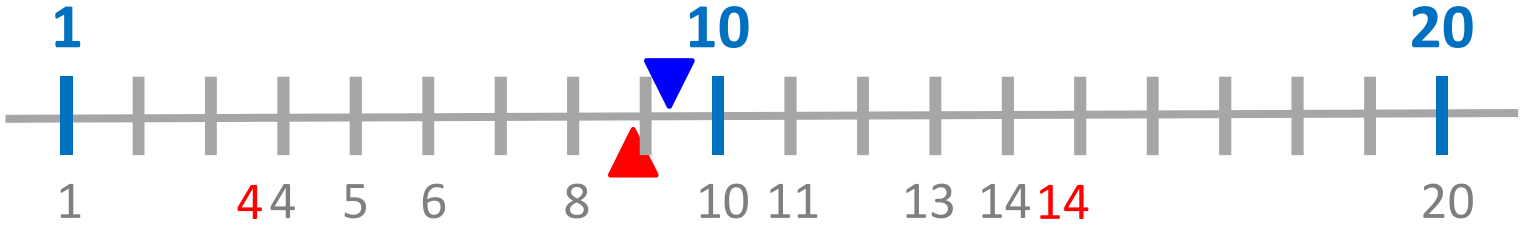
Step 1 先將數據由小到大排列（如上數線所列）

Step 2 10%的剪尾平均數
→ 10(組)×10% = 1 (前後將去掉的數值量)
→ 去掉最小值「1」與最大值「20」各 1 個

Step 3 $\overline{x}_{0.1}$ = (4+5+6+8+10+11+13+14)/8 = 8.875

截尾平均數(truncated mean) $\overline{x}_a$

假設有一組10個數字，若用直線表示，其各數據的值對應如下。求這10組數字的平均數與10%（a=0.1）的截尾平均數（$\overline{x}_{0.1}$）：



平均數： $\bar{x}$ =(1+4+5+6+8+10+11+13+14+20)/10 = 9.2

10%的截尾平均數：

Step 1 先將數據由小到大排列（如上數線所列）

Step 2 10%的截尾平均數
→ 10(組)×10% = 1 (前後將去掉的數值量)
→ 去掉最小值「1」，補上最左邊值「4」
→ 去掉最大值「20」，補上最右邊值「14」

Step 3 $\overline{x}_{0.1}$ = (4+4+5+6+8+10+11+13+14+14)/10 = 8.9

幾何平均數（geometric mean）

$$G = \sqrt[n]{X_1 \times X_2 \times \ldots \times X_n} = \sqrt[n]{\prod_{i=1}^{N} X_i}$$

$$G = \sqrt[\sum f]{X_1^{f_1} \times X_2^{f_2} \times \ldots \times X_n^{f_n}} = \sqrt[\sum_{i=1}^{n} f]{\prod_{i=1}^{N} X_i^{f_i}}$$

特性：

(1)幾何平均數受極端值的影響較算術平均數小。

(2)如果變數值有負值，計算出的幾何平均數就會成為負數或虛數。

(3)它僅適用於具有等比或近似等比關係的數據。

(4)幾何平均數的對數是各變數值對數的算術平均數。

(5)變數數列中任何一個變數值不能為0，一個為0，則幾何平均數為0。

(6)幾何平均法主要用於動態平均數的計算。

幾何平均數的一個重要用途

平均成長率：

$$\bar{r} = \sqrt[n]{(1+r_1)(1+r_2)\cdots(1+r_n)} - 1$$

其中 r_i = 1,2,3,.....,n 表示每年成長率

$$(1+\bar{r}) = \sqrt[n]{(1+r_1)(1+r_2)\cdots(1+r_n)}$$

幾何平均數

整體的代表〝平均值〞

第2年的實際成長值

調和平均數（harmonic mean）

$$H = \frac{1}{\frac{\sum \frac{1}{x}}{n}} = \frac{n}{\sum \frac{1}{x}} \qquad H = \frac{1}{\frac{\sum \frac{1}{x} f}{\sum f}} = \frac{\sum f}{\sum \frac{1}{x} f}$$

特性：

(1)調和平均數易受極端值的影響，且受極小值的影響比受極大值的影響更大。

(2)只要有一個變數值為零，就不能計算調和平均數。

(3)變數x的值不能為0。

(4)調和平均數易受極端值的影響。

(5)要註意其運用的條件。調和平均數多用於已知分子資料，缺分母資料時。

算術平均數、調和平均數與幾何平均數關連

算術平均數、調和平均數和幾何平均數三者間存在如下數量關係：

$$H \le G \le X$$

並且只有當所有變數值都相等時，這三種平均數才相等

二數的調和平均數

$$H = \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}} = \frac{2x_1x_2}{x_1 + x_2} \qquad H = \frac{G^2}{X}.$$

[經濟計算和經濟分析中的算術平均數和調和平均數](#) 2頁

差異衡量

資料集中的情況

絕對離差量數~全距(range)

$$R = x_{max} - x_{min}$$

特點：

(1)易受極端值影響

(2)無法測出中間值之間的差異狀況

(3)資料單位不同時，無法比較

絕對離差量數~四分位距與四分位差

(interquartile-rang, IQR) (quartile deviation, Q.D.)

$$IQR = Q_3 - Q_1$$

$$Q.D. = \frac{Q_3 - Q_1}{2}$$

特點：

(1)不容易受極端值影響

(2)只考慮第1與第3四分位數，忽略其他資料

(3)不具代數運算性質

絕對離差量數~平均差

(mean absolute deviation, MAD)

$$MAD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

特點：

(1)所有值都考量，較全距與四分位距敏感

(2)處理運算較為複雜

(3)容易受到極端值影響

絕對離差量數～變異數與標準差

	樣本	母體
變異數 (variance)	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$	$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$
	$= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2\right)$	$= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2$
標準差 (deviation)	$s = \sqrt{s^2}$	$\sigma = \sqrt{\sigma^2}$

自由度(degree of freedom)：
最大獨立變數的個數

絕對離差量數~變異數與標準差

特點：

(1)標準差越小 → 大部分數值越集中於平均數附近

(2)所有資料皆考量，感應靈敏

(3)具有代數運算特性，所以應用範圍最廣

(4)適用在資料單位相同時比較

對同一筆資料而言：

$$R > s > MAD > Q.D.$$

相對離差量數～變異係數

(coefficient of variation)

$$CV = \frac{\text{標準差}}{\text{平均數}} \times 100\%$$

相對離差量數：

利用去除單位的技巧來克服單位不同或平均數不同無法比較差異大小的情況

絕對離差量數～Z分數

(standardize value)

$$Z_i = \frac{x_i - \bar{x}}{s}$$

樣本

$$Z_i = \frac{x_i - \mu}{\sigma}$$

母體

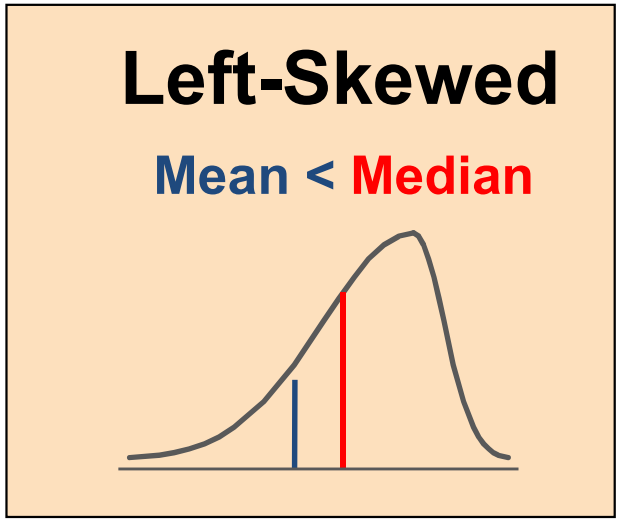
偏態衡量

資料偏斜的情況

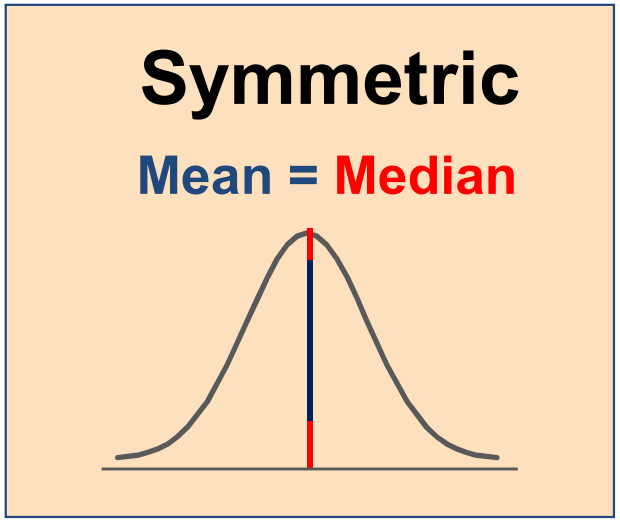
偏態(skewness)係數

$$\beta_1 = \frac{\frac{1}{N}\sum_{i=1}^{N}(\mathrm{x_i} - \mu)^3}{\sigma^3}$$

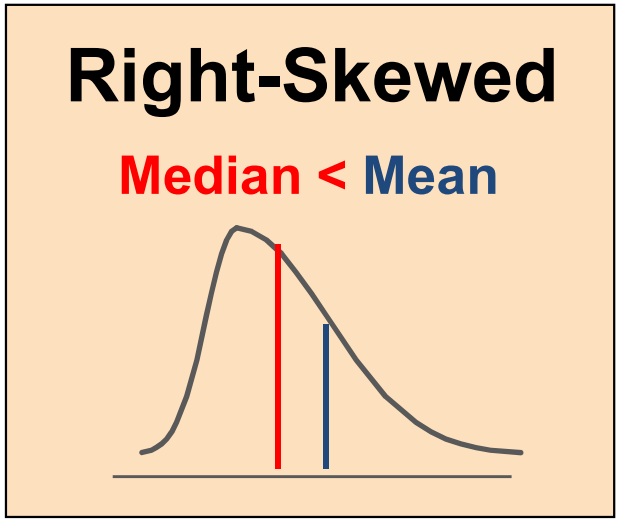
左偏分配
$\beta_1 < 0$



對稱分配
$\beta_1 = 0$



右偏分配
$\beta_1 > 0$

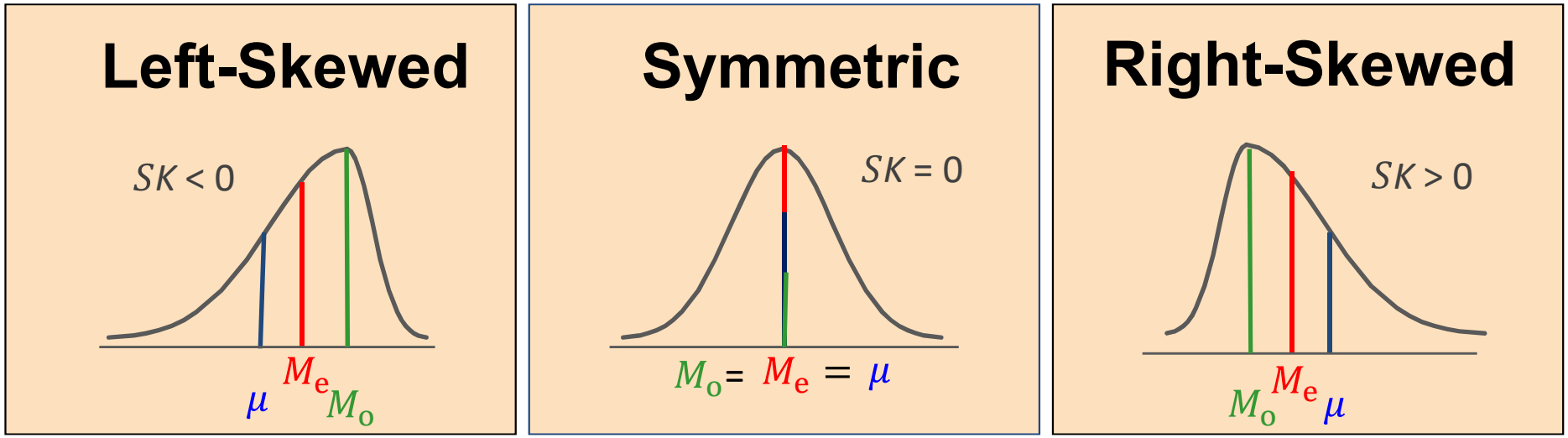


皮爾生(Pearson)偏態係數

$$SK = \frac{\mu - M_o}{\sigma} = \frac{3(\mu - M_e)}{\delta}$$

Pearson經驗法則可知：

眾數到平均數的距離大約等於中位數到平均數距離的3倍



峰態衡量

資料陡峭的情況

峰度(kurtosis)

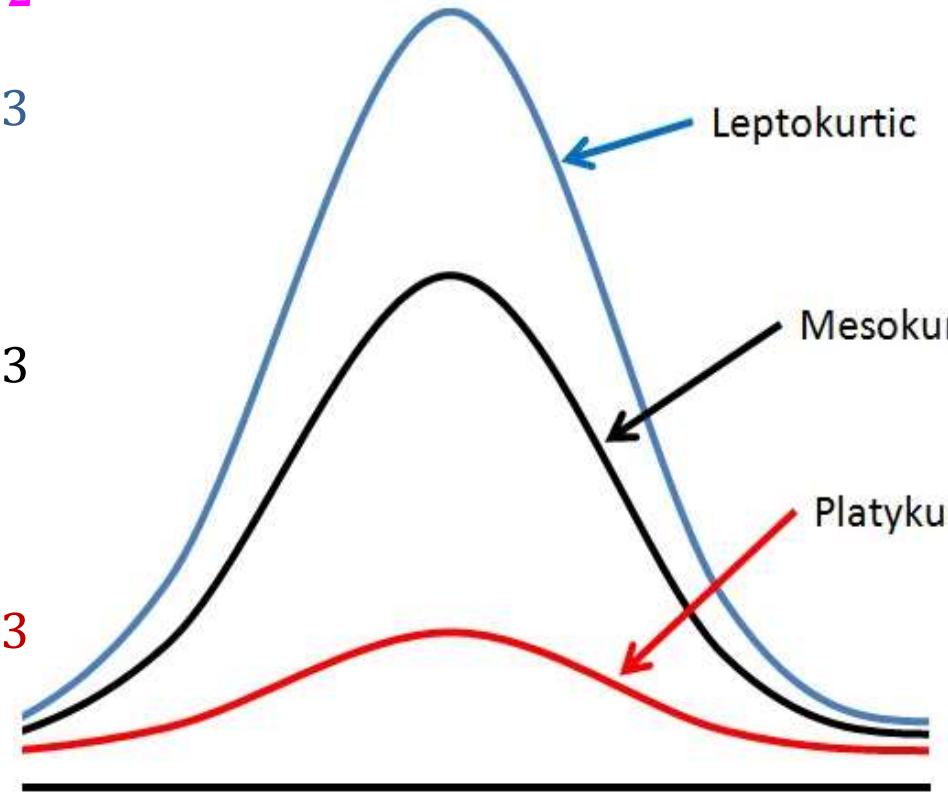
$$\beta_2 = \frac{\frac{1}{N}\sum_{i=1}^{N}(\mathrm{x_i} - \mu)^4}{\sigma^4} = \frac{\mathrm{M_4}}{\sigma^4}$$

常態分配的 β_2 為 3

高狹峰 $\beta_2 > 3$

常態峰 $\beta_2 = 3$

低闊峰 $\beta_2 < 3$



Sharper Peak Than Bell-Shaped (Kurtosis > 0)

Bell-Shaped (Kurtosis = 0)

Flatter Than Bell-Shaped (Kurtosis < 0)

The End