2. 敘述統計學
Descriptive Statistics

- 次數分配(frequency distribution)。
- 統計特徵量數(measures of statistical characteristics)。
  - 中央趨勢量數
  - 位置量數
  - 分散趨勢量數
  - 分配形狀量數
- 探測性資料分析(exploratory data analysis, EDA)。

2.0 集合分類

- 集合(set)可分為
  - Finite set
  - Countable infinite set
  - Uncountable infinite set ➡

- 什麼是可數(countable)?

- 簡要數論
  - 自然數(正整數, N)
    - 無限還是有限? 可數還是不可數?
  - 整數(Z)
    - 無限還是有限? 可數還是不可數?
    - 自然數多還是整數多? 可否與自然數一一對應?
  - 有理數(Q)
    - 定義:
    - 包含哪些數:

2.0 集合分類(contd.)

- 簡要數論(contd.)
  - 有理數(Q)(contd.)
    - 無限還是有限? 可數還是不可數?
    - 自然數多還是有理數多? 可否與自然數一一對應?
    - $0.\bar{9}$ 與1哪個大?
  - 實數(R)
    - 包含哪些數:
    - 無理數的定義:
    - 0~1有幾個實數?
    - 無限還是有限? 可數還是不可數?
    - 自然數多還是實數多? 可否與自然數一一對應?
    - 不可數無限是什麼概念?
    - 連續的意思是?
    - Runner's paradox
  - 複數(C)
    - 包含:

2.1 次數分配

- 次數分配(frequency distribution)：變數的數據資料在各個分組中實際出現次數(frequency)的分配(distribution，或譯分佈)情形，亦即，將雜亂無章的資料整理為清晰且易於觀察、分析與解釋的形式，通常可將次數分配繪成次數分配表(frequency table)。

- 屬質(qualitative)與屬量(quantitative)兩種不同的變數資料其性質不相同，適用的方法與適用的圖形也不相同。

2.1.1 屬質資料次數分配表

- 編製次數分配表的步驟：
  1) 直接按不同的類別分組，太小的組可合併。
  2) 以正字或柵欄符號劃記次數。
  3) 計算各組次數。
  4) 以表格或圖形陳示。

- 適用的圖形：
  - 長條圖(bar chart or bar graph)
    - 僅能用於離散型的變數資料；組與組之間要留空隙，不可併在一起。
  - 圓餅圖(pie chart ，或稱圓形比例圖)
  - 柏雷托圖(Pareto chart)
    - 是一種經過特殊變化的長條圖，它進一步依據次數的多寡來排序，並會同時顯示出相對累加次數，以突顯出次數較多的類別。

範例2.1

- 以下為我國前20大企業CEO之座車廠牌資料：

  BMW, Benz, Benz, Volvo, SAAB, Nissan, Toyota, BMW, BMW, Volvo, Volvo, Benz, Benz, BMW, Volvo, Volvo, Opel, BMW, SAAB, BMW

  試編製其次數分配表，並以適當的圖形陳示之。

2.1.2 屬量資料次數分配表

- 屬量資料以連續型資料較為常見，所以在此只討論連續型的資料。
- 編製次數分配表的步驟：
  1) 將資料以「互斥(mutually exclusive)與週延(exhaustive)」的原則分組。
  2) 排序，求全距(range，R)：R = 最大值 − 最小值
  3) 決定組數(number of classes，k)與組距(class width，h)。How?
  4) 決定組限(class limits)與組界(class boundaries)。How?

2.1.2 屬量資料次數分配表(contd.)

- 編製次數分配表的步驟(contd.)：
  5) 以正字或柵欄符號劃記次數。
  6) 計算各組次數。
  7) 以表格或圖形陳示。

- 適用的圖形：
  - 直方圖(histogram)
    - 以長條的長度來詮釋每一個不同組別次數的大小。
    - 組與組間要緊密並排相連，以反應資料的連續性。因此，直方圖是特別被使用於屬量連續型的變數資料，能讓我們看出資料的分配(distribution，或分佈)情況。
  - 圓餅圖(pie chart ，或稱圓形比例圖)
  - 次數多邊圖(frequency polygon，或稱折線圖)
    - 用以看出整筆資料的分配情形。
    - 在每一組組中點上方，依該組之次數大小標出一個點，再將每一組的點以直線連結，同時，我們會往前與往後各延伸一組，讓折線起於橫軸，也終於橫軸。

2.1.3 累加次數與相對次數

- 累加次數分配(cumulative frequency distribution)：
  - 累加次數直方圖(cumulative frequency histogram)
  - 肩形圖(ogive，或稱累加次數折線圖，cumulative frequency polygon)
- 相對次數分配(relative frequency distribution)：
  - 相對次數? 相對次數可以類比什麼概念?
  - 相對次數直方圖(relative frequency histogram)
  - 相對次數折線圖(relative frequency polygon)
- 相對累加次數分配(relative cumulative frequency distribution)：
  - 相對累加次數直方圖(relative cumulative frequency histogram)
  - 相對累加次數折線圖(relative cumulative frequency polygon，或稱相對次數肩形圖，relative frequency ogive)

範例2.2

- 下面是中山大學某系的統計學期中考成績，該班共有30個同學，試編製次數分配表，並以適當的圖形陳示之：

  42 45 54 38 50 74 53 59 47 49 54 58 52 64 33

  44 39 62 67 52 49 53 52 68 47 44 52 63 55 79

2.2 中央趨勢量數

- 中央趨勢量數(measures of central tendency)，是摘要(summarize)出一筆資料「中央在哪裡」之資訊的量數(measure)。

2.2.1 算術平均數

定義 2.1 算術平均數 (arithmetic mean or mean)

(1)母體算術平均數：有一組母體資料$X_1, X_2, \ldots, X_N$，定義母體算術平均數為

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i$$

(2)樣本算術平均數：有一組樣本資料$X_1, X_2, \ldots, X_n$，定義樣本算術平均數為

2.2.1 算術平均數(contd.)

- 分組資料(grouped data)求算術平均數：採用「集中分配假設」，假設每一組的所有資料點均落在組中點(midpoint) 之上。
  - 母體資料：

$$\mu \doteqdot \frac{1}{N}\sum_{i=1}^{k} \mathrm{mp}_i f_i \text{ , } k\text{表組數, mp}_i\text{表第}i\text{組之組中點, } f_i\text{表第}i\text{組之次數(frequency)}$$

  - 樣本資料：

- 算術平均數之數學性質
  - 母體資料：
    - $\sum_{i=1}^{N} X_i = N\mu$
    - $\sum_{i=1}^{N}(X_i - \mu) = 0$
    - $\sum_{i=1}^{N}(X_i - \mu)^2 \leq \sum_{i=1}^{N}(X_i - a)^2 , \forall a$
      How to prove it?

2.2.1 算術平均數(contd.)

- 算術平均數之數學性質(contd.)
  - 母體資料(contd.)：
    - $Y_i = X_i + b, i = 1,2,\dots,N$, 其中$b$為任意常數 $\Rightarrow$ $\mu_Y = \mu_X + b$ How to prove it?
    - $Y_i = aX_i, i = 1,2,\dots,N$, 其中$a$為任意常數 $\Rightarrow$ $\mu_Y = a\mu_X$ How to prove it?
  - 樣本資料：

2.2.2 中位數

定義 2.2 中位數 (median)

中位數是一筆數據資料(母體資料或樣本資料皆可)中，由小排到大之後，最中央的數，換言之，有一半的數據要小於或等於中位數，有另外一半的數據要大於或等於中位數，一般用 η 來表示母體中位數，用 m_e 或 m_d 來表示樣本中位數。

- 有一組樣本資料$X_1, X_2, \ldots, X_n$，求樣本中位數：
  - 先排序：$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$
  - $$m_e = \begin{cases} \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2}, & \text{當} n \text{為偶數} \\ X_{\left(\frac{n+1}{2}\right)}, & \text{當} n \text{為奇數} \end{cases}$$

- 分組資料求樣本中位數：
  - 先找出中位數組(median class)：位在累加次數 $\frac{n}{2}$ 處
  - $$m_e \doteqdot L + \left(\frac{\frac{n}{2} - F_{-1}}{f_{m_e}}\right) \times h$$

    L表中位數組下界, f_{m_e}表中位數組之次數, F_{-1}表中位數組前一組為止之累加次數, h表組距

2.2.2 中位數(contd.)

- 若有一組母體資料$X_1, X_2, \ldots, X_N$，如何求母體中位數?
- 中位數的數學性質：
  - $\sum_{i=1}^{N}|X_i - \eta| \leq \sum_{i=1}^{N}|X_i - a|, \forall a$
  - $Y_i = X_i + b, i = 1,2,\ldots,N$, 其中$b$為任意常數 $\Rightarrow$
  - $Y_i = aX_i, i = 1,2,\ldots,N$, 其中$a$為任意常數 $\Rightarrow$

2.2.3 眾數

定義 2.3 眾數 (mode)

眾數是一筆數據資料(母體資料或樣本資料皆可)中，出現頻率(次數，frequency)最多者，一般用 m_o 來表示眾數。

- 眾數可以有幾個? 單峰分配(unimodal distribution)? 雙峰分配(bimodal distribution)? 多峰分配(multimodal distribution)?
- 對稱分配(symmetric distribution)? 非對稱分配(asymmetric distribution，或稱偏斜分配，skewed distribution)?
- 分組資料求眾數：
  - King's method
    - 先找出眾數組(modal class)
    - $m_o \doteqdot L + \left(\frac{f_{+1}}{f_{-1}+f_{+1}}\right) \times h$

      L表眾數組下界, f_{-1}表眾數組的上一組之次數, f_{+1}表眾數組的下一組之次數, h表組距

2.2.3 眾數(contd.)

- 分組資料求眾數(contd.)：
  - Czuber's method
    - 先找出眾數組
    - $m_o \doteqdot L + \left(\frac{\Delta_{-1}}{\Delta_{-1}+\Delta_{+1}}\right) \times h, \quad \Delta_{-1} = f_{m_o} - f_{-1}, \quad \Delta_{+1} = f_{m_o} - f_{+1}$
      L表眾數組下界, f_{m_o}表眾數組次數, h表組距
  - Pearson's method (it is an empirical rule)
    - $|\mu - m_o| \doteqdot 3|\mu - \eta|$

2.2.4 截尾平均數(Trimmed mean)

- 有一組樣本資料$X_1, X_2, \ldots, X_n$，求截尾平均數：
  - 先排序：$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$
  - $m_T = \frac{X_{([pn]+1)} + X_{([pn]+2)} + \cdots + X_{(n-[pn])}}{n - 2[pn]}$, $[pn]$表高斯符號

範例2.3

- Determine the estimated mean, median, and mode (by three different methods) of the following frequency distribution of a sample with size $n = 30$.

Class No.	Class	Midpoint	Frequency	Cumulative Frequency
1	20-30	25	2	2
2	30-40	35	8	10
3	40-50	45	11	21
4	50-60	55	6	27
5	60-70	65	3	30

2.2.5 其他中央趨勢量數

定義 2.4 幾何平均數 (geometric mean)

有一筆資料$X_1, X_2, \ldots, X_N$，定義幾何平均數為 $m_G = \sqrt[N]{X_1 \cdot X_2 \cdots X_N}$

- 適用時機：
  - 計算連續數期之比率、變化率、報酬率、成長率等資料的中央趨勢時，採用幾何平均數較有代表性。

定義 2.5 調和平均數 (harmonic mean)

有一筆資料$X_1, X_2, \ldots, X_N$，定義調和平均數為 $m_H = \frac{N}{\sum_{i=1}^{N} \frac{1}{X_i}}$

- 適用時機：
  - 每一段行走距離皆相同時(但速度不見得相同)，求各段之總平均速度。
  - 每一次購買之總金額相同時(但價格不見得相同)，求各次購買之總平均價格。

範例2.4

- A fund manager reported that his previous performance of past three years are 25%, -45%, and 90%, respectively, and the average performance is 23%, which is better than 20% of market. Do you agree of his assertion? Explain.

範例2.5

- 某人以 5公里/小時、10公里/小時、20 公里/小時三種速度各行走 1公里, 則他總共行走 3公里的平均行走速度是多少?

2.3 位置量數

- 有時候我們感興趣的不只是「中央在哪裡」而已，例如，我們想要知道考試成績排在前20%的分數是多少分，處理這類的問題，我們就需要位置量數(measures of location)。
- 位置量數跟中位數的關係是什麼?

2.3.1 百分位數

定義 2.6 百分位數 (percentile)

針對一筆數據資料(母體資料或樣本資料皆可), 其第 r百分位數(the rth percentile, 記為 P_r)的定義是，有 $r\%$的數據要小於或等於 P_r，有另外$(100-r)\%$的數據要大於或等於 P_r，其中 $r = 1,2,\dots,99$。

- 有一組樣本資料$X_1, X_2, \dots, X_n$，求 P_r：
  - 先排序：$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$
  - 計算 P_r的位置： $i = n \times \frac{r}{100}$
  - $P_r = \begin{cases} \frac{X_{(i)}+X_{(i+1)}}{2}, \text{當} i \text{為整數} \\ X_{([i]+1)}, \text{當} i \text{不為整數} \end{cases}$, $[i]$表高斯符號
- 分組資料求 P_r：
  - 計算 P_r的位置： $i = n \times \frac{r}{100}$
  - $P_r \div L + \left(\frac{i-F_{-1}}{f_{P_r}}\right) \times h$

    L表第r百分位數組下界, f_{P_r}表第r百分位數組之次數,
    F_{-1}表第r百分位數組前一組為止之累加次數, h表組距

2.3.2 十分位數(decile)與四分位數(quartile)

- $\begin{cases} D_1 = \\ \vdots \\ D_5 = \\ \vdots \\ D_9 = \end{cases}$
- $\begin{cases} Q_1 = \\ Q_2 = \\ Q_3 = \end{cases}$

範例2.6

- For the grouped data in Example 2.3, determine the estimated 63th percentile, the estimated lower quartile, and the estimated upper quartile.

2.4 分散趨勢量數

- 分散趨勢量數(measures of dispersion)，是摘要出一筆資料的「分散性(dispersion」或「變異性(variability)」之資訊的量數(measure)。

2.4.1 變異數

定義 2.7 變異數 (variance)

(1) 母體變異數：有一組母體資料$X_1, X_2, \ldots, X_N$，定義母體變異數為

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2 =$$

(2) 樣本變異數：有一組樣本資料$X_1, X_2, \ldots, X_n$，定義樣本變異數為

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 =$$

Why (n-1)?

2.4.1 變異數(contd.)

- 分組資料求變異數：採用集中分配假設，假設每一組的所有資料點均落在組中點(midpoint)上。
  - 母體資料：$\sigma^2 \doteqdot \frac{1}{N}\sum_{i=1}^{k}(\mathrm{mp}_i - \mu)^2 f_i = \frac{1}{N}\left(\sum_{i=1}^{k}\mathrm{mp}_i^2 f_i - \frac{\left(\sum_{i=1}^{k}\mathrm{mp}_i f_i\right)^2}{N}\right)$
  - 樣本資料：
- 變異數之數學性質：
  - 母體資料：
    - $\sigma^2 \geq 0$ When does the equality hold?
    - $N\sigma^2 = \sum_{i=1}^{N}(X_i - \mu)^2 = \sum_{i=1}^{N}X_i^2 - \frac{\left(\sum_{i=1}^{N}X_i\right)^2}{N}$
    - $\sum_{i=1}^{N}X_i^2 = N\sigma^2 + N\mu^2$
    - $Y_i = X_i + b, i = 1,2,\ldots,N$, 其中$b$為任意常數 $\Rightarrow$ $\sigma_Y^2 = \sigma_X^2$ How to prove it?
    - $Y_i = aX_i, i = 1,2,\ldots,N$, 其中$a$為任意常數 $\Rightarrow$ $\sigma_Y^2 = a^2\sigma_X^2$ How to prove it?

2.4.1 變異數(contd.)

- 變異數之數學性質(contd.)：
  - 樣本資料：

2.4.2 標準差

定義 2.8 標準差 (standard deviation)

(1) 母體標準差：有一組母體資料$X_1, X_2, \ldots, X_N$，定義母體標準差為 $\sigma = \sqrt{\sigma^2}$

(2) 樣本標準差：有一組樣本資料$X_1, X_2, \ldots, X_n$，定義樣本標準差為 $S = \sqrt{S^2}$

- 標準差之數學性質：
  - 母體資料：
    - $\sigma \geq 0$　When does the equality hold?
    - $Y_i = X_i + b, i = 1,2, \ldots, N$, 其中$b$為任意常數　⇒　$\sigma_Y = \sigma_X$
    - $Y_i = aX_i, i = 1,2, \ldots, N$, 其中$a$為任意常數　⇒ $\sigma_Y = |a|\sigma_X$
  - 樣本資料：

範例2.7

- 某校財管系大三男同學共40人，平均身高為170公分，標準差為10公分；大三女同學共60人，平均身高為160 公分，標準差為8 公分；試求全體大三同學身高之平均數與標準差。

2.4.3 變異係數

定義 2.9 變異係數 (coefficient of variation，CV)

(1) 母體變異係數：有一組母體資料$X_1, X_2, \ldots, X_N$，定義 $CV = \frac{\sigma}{\mu} \times 100\%$

(2) 樣本變異係數：有一組樣本資料$X_1, X_2, \ldots, X_n$，定義 $CV = \frac{S}{\bar{X}} \times 100\%$

- 變異係數的單位是什麼?
- 一群大象平均體重200 kg，標準差50 kg；一群螞蟻平均體重1 g，標準差0.3 g。何者體重的分散程度較大?
- 變異係數用以比較兩筆資料的分散程度，特別適用於以下兩種情況：
  - 兩筆資料數值大小(scale) 差距過大時
  - 兩筆資料單位截然不同時

2.4.3 其他分散趨勢量數

- 全距(range)：$R = X_{(n)} - X_{(1)}$
- 四分位距(interquartile range，IQR)：$IQR = Q_3 - Q_1$
- 四分位差(quartile deviation，QD，或稱四分半距，semi-interquartile range)：$QD = \frac{Q_3 - Q_1}{2}$
- 平均絕對離差(mean absolute deviation，MAD，或簡稱平均離差，mean deviation)：

  $MAD = \frac{1}{N}\sum_{i=1}^{N}|X_i - \mu|$ 或 $MAD = \frac{1}{N}\sum_{i=1}^{N}|X_i - \eta|$
- 均互差(mean difference，MD)：$MD = \frac{1}{\binom{n}{2}}\sum_{i<j}|X_i - X_j|$

範例2.8

- For the grouped data in Example 2.3, determine the estimated variance, standard deviation, coefficient of variation, interquartile range, and quartile deviation.

2.5 其他量數與課題

定義 2.10 z分數 (z-score，或稱標準分數，standard score)

有一組母體資料$X_1, X_2, \ldots, X_N$或樣本資料$X_1, X_2, \ldots, X_n$，定義其中第i個資料點X_i的z分數為

$Z_i = \frac{X_i - \mu}{\sigma}$ 或 $Z_i = \frac{X_i - \bar{X}}{S}$

- 某個資料點的z-score等於2，這代表什麼意思?
- z-score適用於比較屬於不同資料集的兩個資料點的相對大小 。

鐘形分配的經驗法則

若手中的資料近似鐘形分配(bell-shaped distribution)，則有以下經驗法則：

- 約有68%的資料點落在　　的範圍之內。
- 約有95%的資料點落在　　的範圍之內。
- 約有99.7%的資料點落在　　的範圍之內。

- 哪些資料點可被視為離群值(outliers)?

範例2.9

- 在某次考試中，甲生的成績為統計66分，微積分82分，而統計的平均分數為50分，標準差為11分，微積分的平均分數為73分，標準差為9分，若參加兩科目的考生皆為同一批人，請問甲生在哪一科中的表現較為傑出?

範例2.10

- Given a symmetric, bell-shaped distribution of 100 values with a mean of 115.2 and standard deviation of 3.2. Roughly how many values should lie between 108.8 and 121.6?

2.6 分配形狀量數(measures of shape)

- 數據資料的分配形狀特徵主要分為兩類：
  - 偏態(skewness)：一筆資料的非對稱性(asymmetry)。
  - 峰態(kurtosis)：一筆資料在中央附近群聚的情形(或是尾端的厚實程度)。

2.6.1 動差體系(system of moments)

定義 2.11 樣本動差 (sample moment)

有一組樣本資料$X_1, X_2, \ldots, X_n$， 我們定義下列兩種動差：

(1)第r階樣本原動差(the rth sample moment about the origin)：$m'_r = \frac{1}{n}\sum_{i=1}^{n}(X_i - 0)^r$， $r = 1,2,3,\ldots$

(2)第r階樣本主動差(the rth sample moment about the mean)：$m_r = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^r$， $r = 1,2,3,\ldots$

- 分組資料求樣本動差：採用集中分配假設，假設每一組的所有資料點均落在組中點上。
  - $m'_r \doteqdot \frac{1}{n}\sum_{i=1}^{k} \mathrm{mp}_i^r \, f_i$ 與 $m_r \doteqdot \frac{1}{n}\sum_{i=1}^{k}(\mathrm{mp}_i - \bar{X})^r \, f_i$

  其中k表組數，mp_i表第 i組之中點，f_i表第 i組之次數(frequency)。

2.6.1 動差體系(contd.)

- 各階動差含有的資訊：
  - 一階動差(原動差)：含有中央趨勢的資訊
  - 二階動差(主動差)：含有分散趨勢的資訊
  - 三階動差(主動差)：含有偏態程度的資訊
  - 四階動差(主動差)：含有峰態程度的資訊
  - 統計學家尚未發現五階以上動差的具體用途

定義 2.12 以原動差轉主動差

給定一組樣本資料$X_1, X_2, \ldots, X_n$，其主動差與原動差的關係為：

(1) $m_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = m_2' - (m_1')^2$

(2) $m_3 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3 = m_3' - 3m_2'm_1' + 2(m_1')^3$

(3) $m_4 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4 = m_4' - 4m_3'm_1' + 6m_2'(m_1')^2 - 3(m_1')^4$

How to prove (2)?

2.6.2 樣本偏態係數

定義 2.13 (動差法)樣本偏態係數 (sample coefficient of skewness)

一組樣本資料$X_1, X_2, \ldots, X_n$，其(動差法)樣本偏態係數定義為：$\alpha_3 = \frac{m_3}{s^3}$

- $\begin{cases} \alpha_3 > 0, \text{ 右偏(skewed to the right, 又稱正偏, postively skewed)} \\ \alpha_3 = 0, \text{ 對稱分配(symmetric distribution)} \\ \alpha_3 < 0, \text{ 左偏(skewed to the left, 又稱負偏, negatively skewed)} \end{cases}$

定義 2.14 皮爾森法樣本偏態係數 (sample Pearson coefficient)

一組樣本資料$X_1, X_2, \ldots, X_n$，其皮爾森法樣本偏態係數定義為：$sk_P = \frac{3(\bar{X}-m_e)}{s}$ 或 $sk_P = \frac{\bar{X}-m_o}{s}$

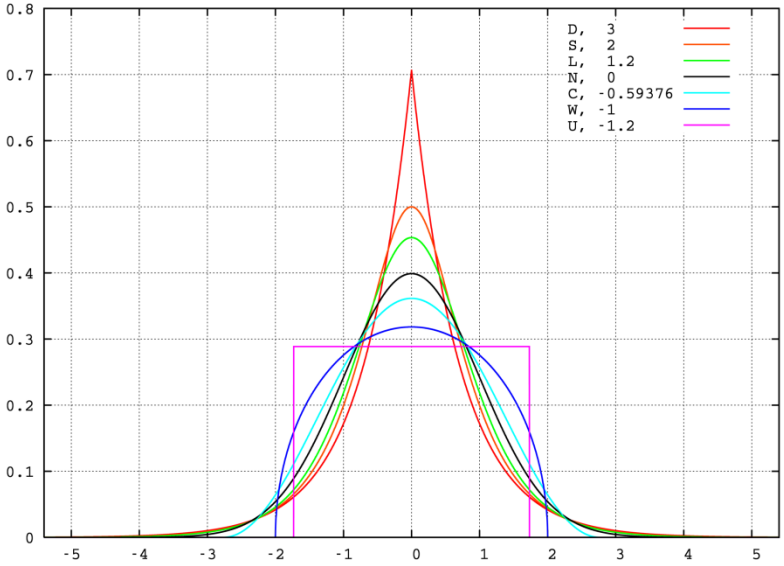
- $\begin{cases} sk_P > 0, \\ sk_P = 0, \\ sk_P < 0, \end{cases}$

2.6.3 樣本峰態係數

定義 2.15 樣本峰態係數 (sample coefficient of kurtosis)

一組樣本資料$X_1, X_2, \ldots, X_n$，其樣本峰態係數定義為：$\alpha_4 = \frac{m_4}{S^4}$

- $\begin{cases} \alpha_4 > 3, \text{ 高狹峰分配(leptokurtic distribution)} \\ \alpha_4 = 3, \text{ 常態峰分配(mesokurtic distribution)} \\ \alpha_4 < 3, \text{ 低闊峰分配(platykurtic distribution)} \end{cases}$



範例2.11

- For the grouped data in Example 2.3, determine the coefficient of skewness and the coefficient of kurtosis.

2.7 探測性資料分析

- 探測性資料分析(exploratory data analysis，EDA)，是以簡單的計算與簡易的圖形，快速地呈現出資料的兩大特性：
  - 分配的位置(location of distribution)：資料所在的位置。
  - 分配的形狀(shape of distribution)：分散程度、偏態與峰態的情況。

2.7.1 五數字摘要

- 五數字摘要(five-number summary)：
  - the smallest value：$X_{(1)}$
  - the 1st quartile：Q_1
  - the median：m_e
  - the 3rd quartile：Q_3
  - the largest value：$X_{(n)}$

  兩相鄰數字的數值接近，代表什麼意思?

2.7.2 盒鬚圖

- 盒鬚圖(box-and-whisker plot)繪製步驟：
  - 繪出盒子(box)，盒子兩邊為 Q_1 與 Q_3。
  - 在盒子內以直線繪出 m_e 的位置。
  - 由 Q_1 往下延伸$1.5 \times IQR$，且由 Q_3 往上延伸$1.5 \times IQR$，當作內籬(inner fence)的位置。
  - 由 Q_1 往下延伸$3 \times IQR$，且由 Q_3 往上延伸$3 \times IQR$，當作外籬(outer fence)的位置。
  - 從盒子的兩端，以虛線(代表鬚，whisker)延伸至內籬以內的最小值和最大值。
  - 溫和離群值以符號 $*$ 繪出，嚴重離群值以符號 $\circ$ 繪出。
  - 內籬與外籬之間的數據，稱之為溫和離群值(mild outliers)，外籬以外的數據，稱之為嚴重離群值(extreme outliers)。

2.7.3 莖葉圖

- 莖葉圖(stem-and-leaf display)很像直方圖，但它含有比直方圖更多的資訊，透過莖葉圖，我們不但可以看出資料分配的形狀，還可以看出原始資料的數值，以及原始資料排序後的結果，而且，莖葉圖比直方圖容易繪製。

範例2.12

- 欲研究管理學院畢業生的起始月薪(US$)，抽取一組樣本大小 n=12 的樣本如下：
  1650, 1750, 1850, 1680, 1555, 1510, 1690, 1930, 1740, 2125, 1720, 1680
  求這筆樣本資料之 five-number summary，並繪出盒鬚圖。
- 繪出範例2.2的資料之莖葉圖。