工程統計

李水彬Shuipin

2023/10/02

敘述統計(II)統計量

學習重點

- 利用統計量數描述資料的分配。
- 知道平均數, 中位數, 眾數, 變異數, 標準差, 百分位數, 四分位數， 四分位距和變異係數等統計量數的定義與計算方式。
- 知道如何利用統計量數分辨資料的偏態。
- 如何使用經驗法則和謝比雪夫不等式說明一定範圍內的比例。

統計量數

- 集中趨勢: 描述大部分資料座落的位置。
- 分散趨勢: 描述一組資料內部的分散程度, 表現同一群體不同個體間的差異程度。
- 形狀參數 : 描述資料的『山勢』。
  - 偏態:
  - 峰態:

集中趨勢-平均數

- 假設有一組母體資料為 $x_1, \cdots, x_N$, 母體平均值以 μ 表之, 定義為

$$\mu = \frac{x_1 + \cdots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

N: 母體總數

- 樣本($x_1, \cdots, x_n$)平均值的定義為

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

n: 樣本大小

範例- 母體平均值

20 位慧而園幼稚園小班幼兒的體重為

18, 17, 18, 16, 17, 15, 15, 16, 18, 15, 14, 16, 12, 14, 13, 16, 15, 19, 16, 16

慧而園幼稚園(全體)小班幼兒的平均體重:

$$\mu = \frac{18 + 17 + \cdots + 16 + 16}{20} = 15.8$$

範例- 平均經濟成長率

台灣1990至2014年的經濟成長率:

5.65	8.36	8.29	6.8	7.49	6.5	6.18	6.11	4.21	6.72
6.42	-1.26	5.57	4.12	6.51	5.42	5.62	6.52	0.7	-1.57
10.63	3.8	2.06	2.23	3.77					

```
## [1] "總和為 126.85"
```

這段時間的平均經濟成長率:

$$\mu = \frac{5.65 + 8.36 + \cdots + 3.77}{2014 - 1990 + 1} = \frac{126.85}{25} = 5.074$$

範例- 樣本平均值

- 隨機抽出 5 位慧而園幼稚園小班兒童的體重 (x) 分別為 15, 18, 16, 19, 18

$$\bar{x} = \frac{15+18+16+19+18}{5} = \frac{86}{5} = 17.2$$

- 再隨機抽出 5 位, 測量其體重 (y) 分別為 15, 14, 15, 18, 12

$$\bar{y} = \frac{15+14+15+18+12}{5} = \frac{74}{5} = 14.8$$

樣本平均值會隨著會應選取的樣本不同而不同。

平均數的代數性質

假設 A 班 50 位同學品管平均成績為 78, B 班 60 位同學品管平均成績為 80, 則兩班平均成績為多少?

$$\text{平均成績} = \frac{\text{總分數}}{\text{總人數}} = \frac{\text{甲班總分數} + \text{乙班總分數}}{\text{甲班人數} + \text{乙班人數}}$$

平均數的性質- 離差和為零

離差: 觀察值與平均值之差。 第 i 個觀察值的離差為 $x_i - \mu$。

$$\sum_{i=1}^{N}(x_i - \mu) = 0$$

若資料為樣本:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

範例- 離差

計算 5 位小班兒童體重 15, 18, 16, 19, 18與平均值的離差($\bar{x} = 17.2$)

$$x_1 - \bar{x} = 15 - 17.2 = -2.2$$

$$x_2 - \bar{x} = 18 - 17.2 = 0.8$$

$$x_3 - \bar{x} = 16 - 17.2 = -1.2$$

$$x_4 - \bar{x} = 19 - 17.2 = 1.8$$

$$x_5 - \bar{x} = 18 - 17.2 = 0.8$$

平均數的性質- 離差平方和最小

(省略)

平均數的性質

- 離差和為零
- 離差平方和最小
- 每個數的權重相同

中位數

中位數位於 數據排序上的中間位置, 即有至少一半的數據大於或等於此數, 也有至少一半的數據小於或等於此數。

這組數據的中位數

15, 18, 16, 19, 18

範例- 中位數

- 8, 9, 11, 12, 13
- 1, 2, 3, 4
- 3, 3, 6, 7,1, 2, 8, 4

範例- 中位數

壽司店近48天銷售盤數排序後如下:

493	833	1004	1500	1703	1927	2491	3003
522	861	1077	1507	1714	1936	2504	3193
522	901	1094	1511	1729	2111	2526	3229
545	917	1221	1511	1740	2173	2555	3676
627	939	1383	1571	1837	2359	2650	4878
689	947	1390	1590	1880	2437	2662	5060

範例- 中位數

有5位男同學的期中考成績 64, 52, 70, 74, 68, 和 6 位女同學的期中考成績 55, 48, 90, 83, 77, 89。

- 請問這 5 位男同學成績的中位數。
- 請問這 6 位女同學成績的中位數。
- 請問這 11 位同學成績的中位數。

範例- 比較平均數與中位數

計算下列兩組數據的中位數與平均數。

- 甲: 14, 13, 17, 18, 20
- 乙: 14, 13, 17, 18, 200

衆數

一組數據中出現次數最多, 重複最多的數稱為眾數。

擲骰子12次, 結果為 4, 2, 1, 6, 5, 3, 3, 2, 4, 3, 6, 3。 衆數為

範例- 眾數

求以下三組數據的眾數。

- 甲: 15, 17, 15, 12, 14,
- 乙: 14, 14, 2, 4, 2, 3, 1, 9
- 丙: 1, 2, 3, 4, 5

百分位數

將資料分割成100等份, 我們稱這些分割點為百分位數。

令 P_i 代表一組樣本的第i 個百分位數, 在這組樣本中至少有 $(100-i)\%$ 的數據大於或等於此數, 也至少有 $i\%$ 的數據小於或等於此數”。

百分位數的計算方法

- 排序: $x_{(1)}, \cdots, x_{(n)}$ 。
- 求出位置指標 k 。

$$k = \frac{i}{100} \times n$$

  - 若k為整數, 則百分位數 $P_i = \frac{x_{(k)} + x_{(k+1)}}{2}$
  - 若k為非整數, 假設k 進一位的整數為k^*, 則百分位數$P_i = x_{(k^*)}$ 。

十分位數

十分位數是利用 9 個分割點將資料分割成 10 等分。 這9個分割點以 $D_1, D_2, \cdots, D_9$ 表示之。根據定義, 可以用百分位數求十分位數，即

$$D_1 = P_{10}, D_2 = P_{20}, \cdots, D_9 = P_{90}$$

四分位數

四分位數是利用 3 個分割點將資料分割成 4 等分。這 3 個分割點為 Q_1, Q_2, Q_3。根據定義,可以用百分位數求四分位數，即

$$Q_1 = P_{25}, Q_2 = P_{50}, Q_3 = P_{75}$$

範例-百分位數

求晶圓厚度 3 個四分位數與 P_{30}。

1379	1428	1459	1462	1472	1484	1501	1513	1543	1566
1385	1448	1460	1465	1474	1488	1505	1516	1546	1568
1412	1449	1460	1469	1482	1495	1509	1517	1547	1569
1425	1451	1461	1470	1483	1500	1512	1520	1548	1601
1427	1455	1461	1470	1483	1500	1512	1533	1554	1614

範例-百分位數

求新生兒體中的 3 個四分位數與 P_{45} 。

2.03	2.51	2.62	2.69	2.80	2.90	3.02	3.08	3.17	3.27	3.40	3.53	3.70
2.21	2.54	2.63	2.70	2.82	2.92	3.03	3.12	3.21	3.30	3.45	3.55	3.71
2.29	2.54	2.64	2.71	2.87	2.93	3.04	3.15	3.22	3.33	3.46	3.56	3.72
2.43	2.55	2.64	2.72	2.88	2.96	3.04	3.16	3.23	3.37	3.46	3.61	3.88
2.44	2.61	2.66	2.80	2.88	3.02	3.05	3.16	3.24	3.38	3.49	3.68	4.20

分散趨勢量數

描述這組資料內容的 差異程度。

- 方法一: 每一個數與集中趨勢量數 (中位數或平均值) 的距離。
  - 變異數 Variance、標準差 Standard deviation、平均絕對離差 Mean absolute deviation
- 方法二: 位置參數間的距離代表全體資料的差異程度。
  - 全距、四分位距、四分位差。

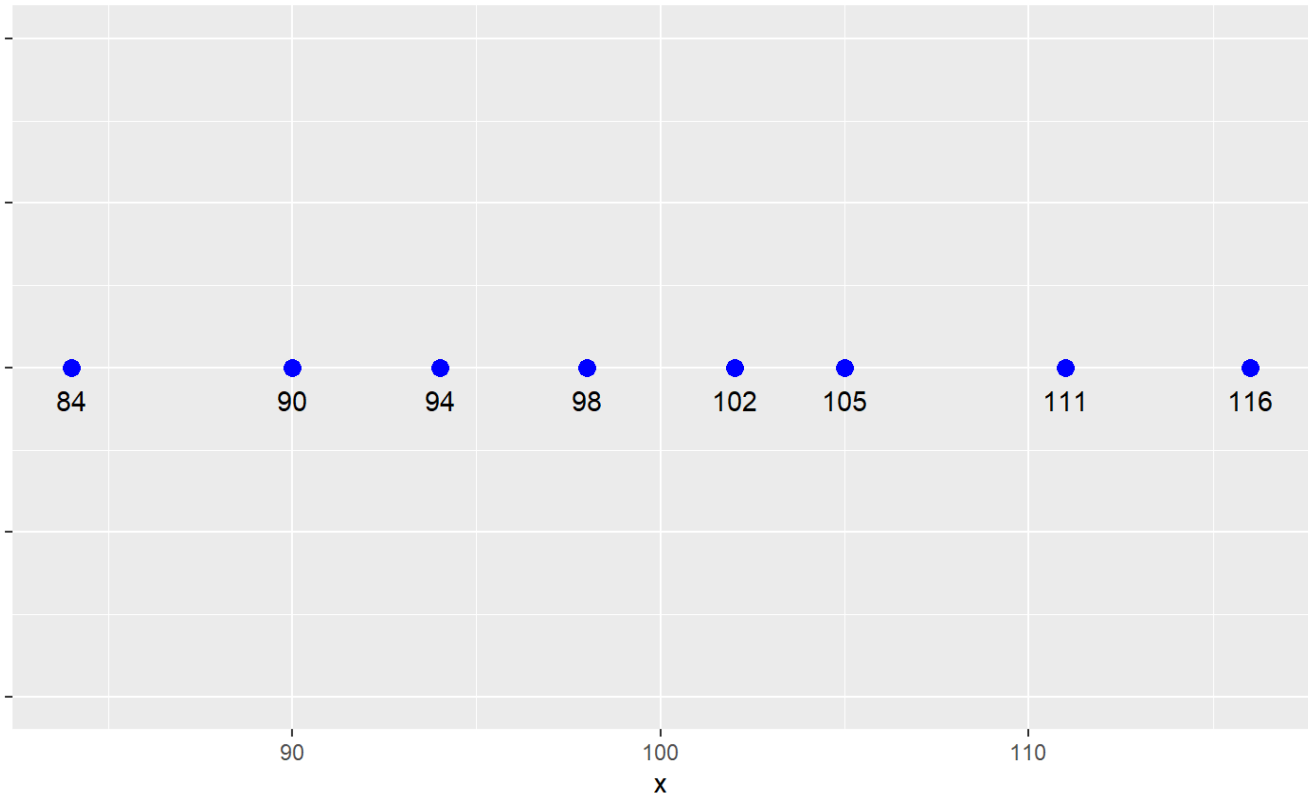
方法一

資料: 84, 105, 98, 102, 94, 116, 111, 90

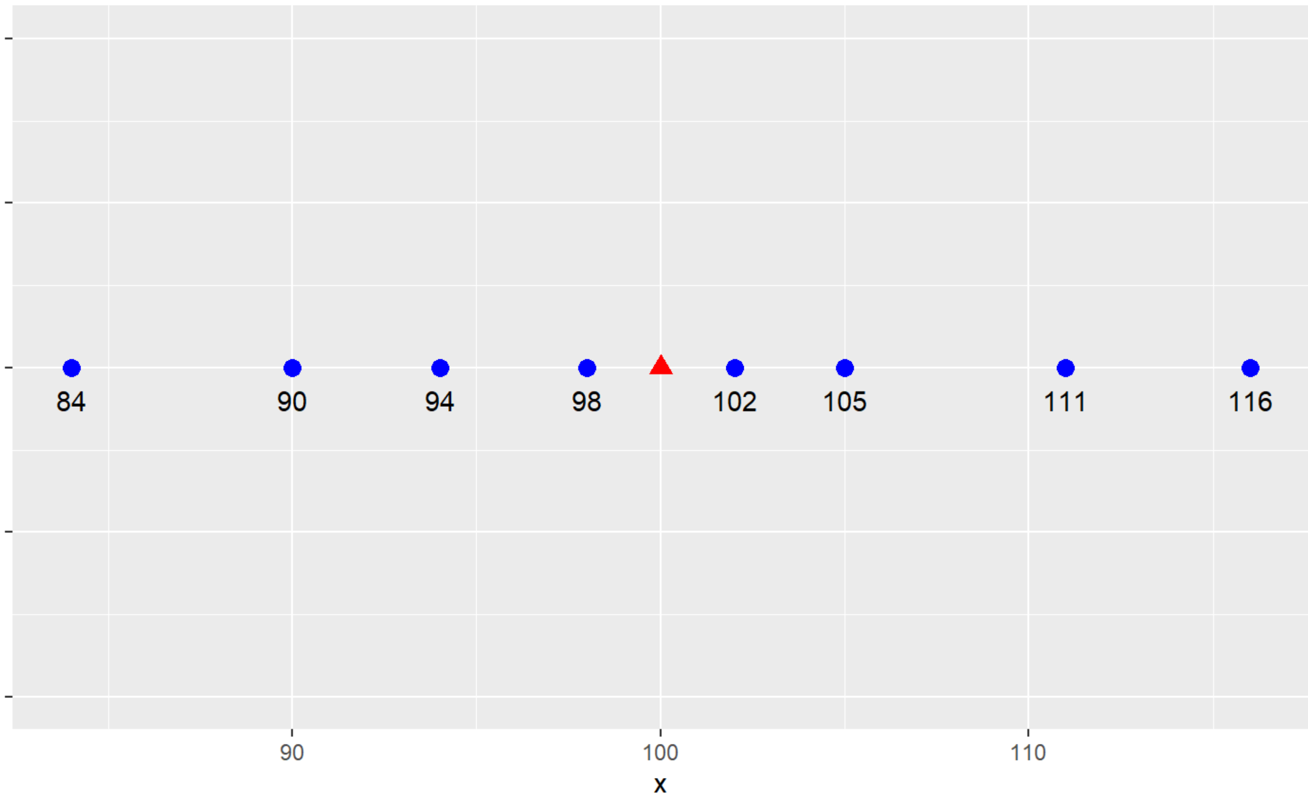
排序: 84, 90, 94, 98, 102, 105, 111, 116

樣本平均值: $\frac{\sum_{i=1}^{8} x_i}{8} = \frac{800}{8} = 100$

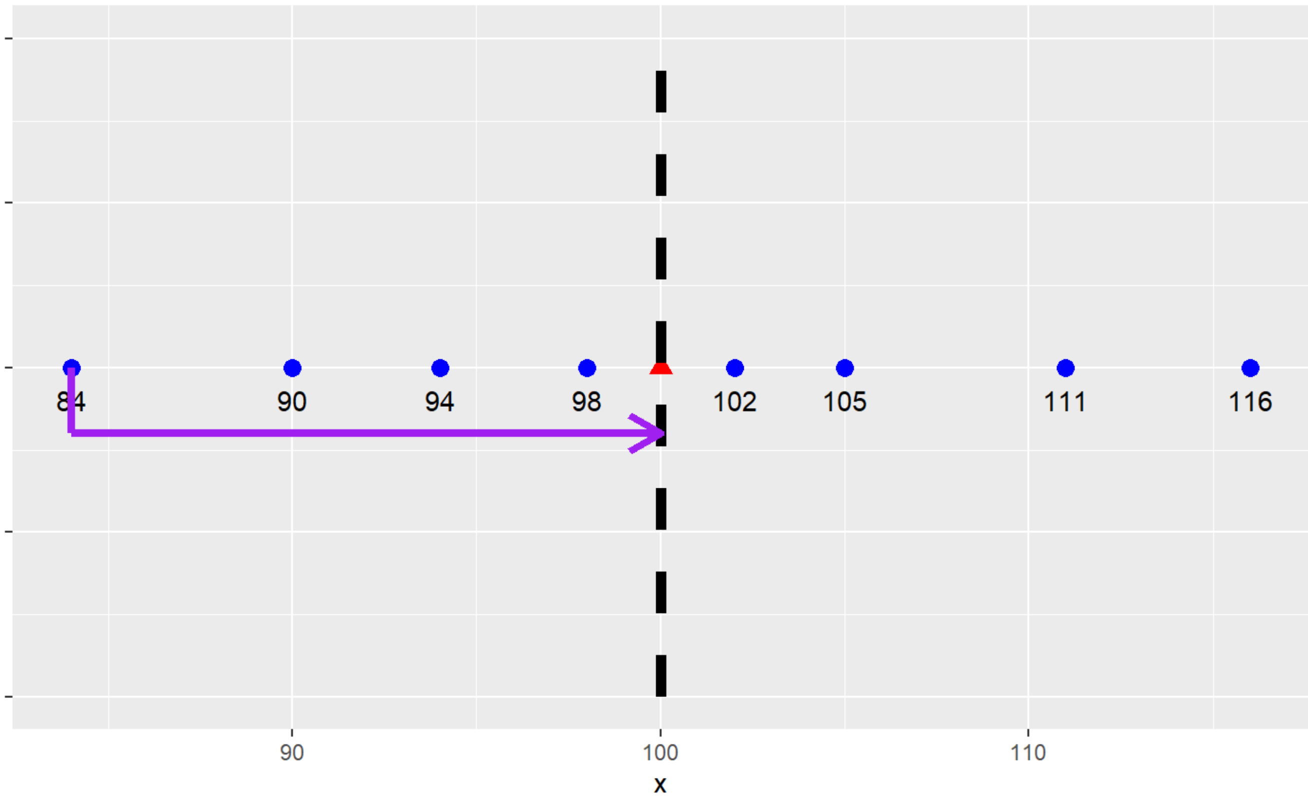
方法一



方法一

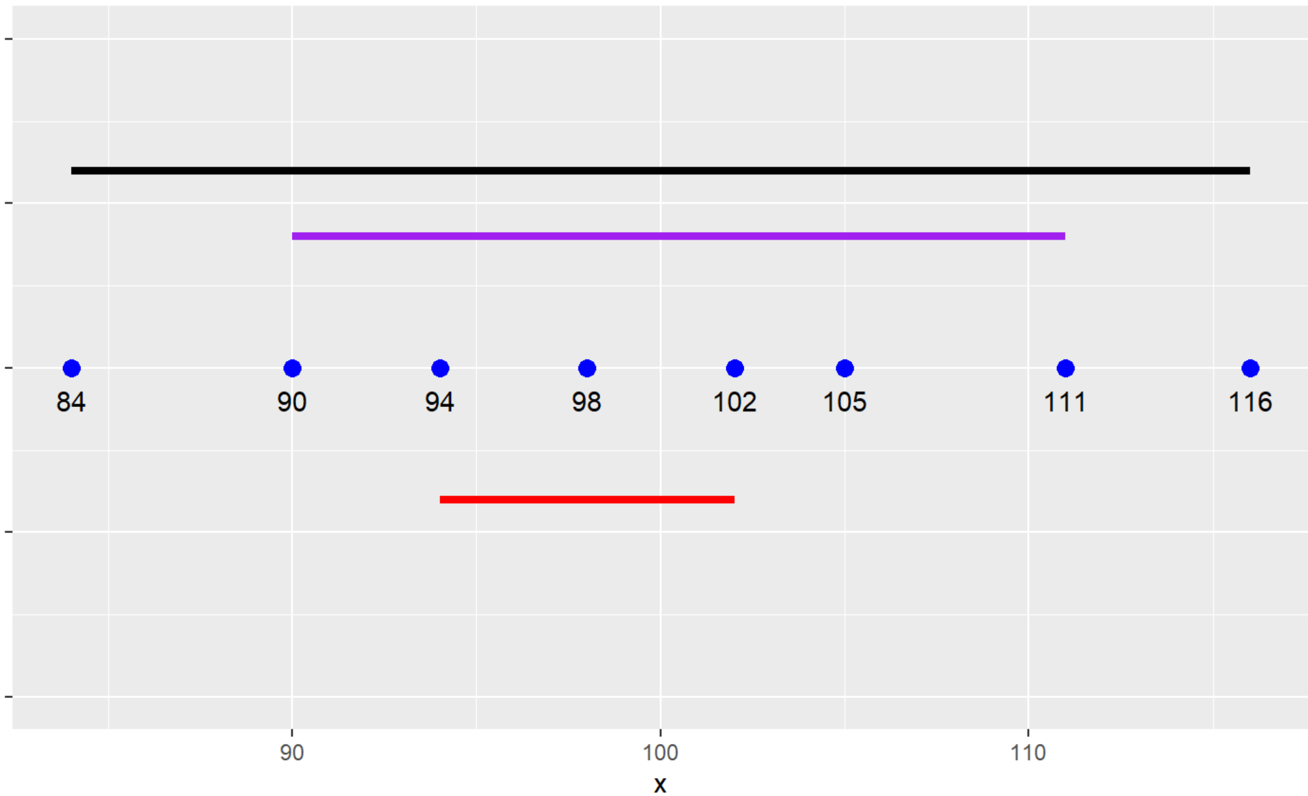


方法一



方法二

位置參數間的距離代表全體資料的差異程度。



變異數- 母體

母體資料為 $x_1, \cdots, x_N$, 母體平均值用 μ 表示。

母體變異數為

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

可以寫成

$$\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2$$

變異數- 樣本

假設樣本資料為 $x_1, \cdots, x_n$, 樣本平均值為 $\bar{x}$, 則 樣本變異數定義為

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

可以寫成

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

範例- 樣本變異數

5 位男同學期中考成績:

64, 52, 70, 74, 68

樣本變異數為

- 總和= 328
- 樣本平均值=65.6
- 離差= -1.6, -13.6, 4.4, 8.4, 2.4
- 離差平方和=283.2=2.56+184.96+19.36+70.56+5.76
- 樣本變異數=$\frac{283.2}{5-1}$=70.8

範例- 樣本變異數

6 位女同學期中考成績:

55, 48, 90, 83, 77, 89

樣本變異數為

- 總和= 442
- 樣本平均值=73.67
- 離差= -18.67, -25.67, 16.33, 9.33, 3.33, 15.33
- 離差平方和=1607.34=348.57+658.95+266.67+87.05+11.09
- 樣本變異數=$\frac{1607.34}{6-1}$=$\mathbf{321.47}$

練習

5 位同的的體重:

75, 77, 73, 63, 75

樣本變異數為

- 總和=
- 樣本平均值=
- 離差=
- 離差平方和=
- 樣本變異數=

練習

5 位同的的體重:

75, 77, 73, 63, 75

- 總和= 363
- 樣本平均值=72.6
- 離差= 2.4, 4.4, 0.4, -9.6, 2.4
- 離差平方和=123.2=5.76+19.36+0.16+92.16+5.76
- 樣本變異數=$\frac{123.2}{5-1}$=30.8

標準差

$$\text{標準差} = \sqrt{\text{變異數}}$$

- 母體標準差

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

- 樣本標準差

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

變異數的單位為資料單位的平方, 標準差為變異數開根號, 單位與觀察值的單位相同。

範例- 標準差

- 已知 5 位男同學成績的變異數為 70.8, 其標準差為

$$s = \sqrt{70.8} = 8.41$$

- 已知 6 位女同學成績的變異數為 321.47, 其標準差為

$$s = \sqrt{321.47} = 17.94$$

平均絕對離差

(略)

全距

一組資料的最大值減去其最小值。

範例- 全距

5 位男同學期中考成績的全距。

64, 52, 70, 74, 68

四分位距

四分位距為第 3 個四分位數減去第1 個四分位數, 即

$$IQR = Q_3 - Q_1$$

範例-四分位距

1379	1428	1459	1462	1472	1484	1501	1513	1543	1566
1385	1448	1460	1465	1474	1488	1505	1516	1546	1568
1412	1449	1460	1469	1482	1495	1509	1517	1547	1569
1425	1451	1461	1470	1483	1500	1512	1520	1548	1601
1427	1455	1461	1470	1483	1500	1512	1533	1554	1614

範例-四分位距

2.03	2.51	2.62	2.69	2.80	2.90	3.02	3.08	3.17	3.27	3.40	3.53	3.70
2.21	2.54	2.63	2.70	2.82	2.92	3.03	3.12	3.21	3.30	3.45	3.55	3.71
2.29	2.54	2.64	2.71	2.87	2.93	3.04	3.15	3.22	3.33	3.46	3.56	3.72
2.43	2.55	2.64	2.72	2.88	2.96	3.04	3.16	3.23	3.37	3.46	3.61	3.88
2.44	2.61	2.66	2.80	2.88	3.02	3.05	3.16	3.24	3.38	3.49	3.68	4.20

綜合練習

6件產品的厚度(cm) :

2.1, 1.9, 1.7, 1.8, 2.3, 2.2

求以下統計量

- 總和=
- 樣本平均值=
- 中位數=
- 離差=
- Q_1=
- Q_3=
- IQR=
- 離差平方和=
- 樣本變異數=
- 標準差=
- 全距=

範例- 偏態係數

5 位男同學的成績為64, 52, 70, 74, 68，成績的皮爾生偏態係數:

因為 $\bar{x} = 65.6, Me = 68$, 變異數為

$$s^2 = \frac{64^2 + 52^2 + 70^2 + 74^2 + 68^2 - 5 \times 65.6^2}{4} = 70.8$$

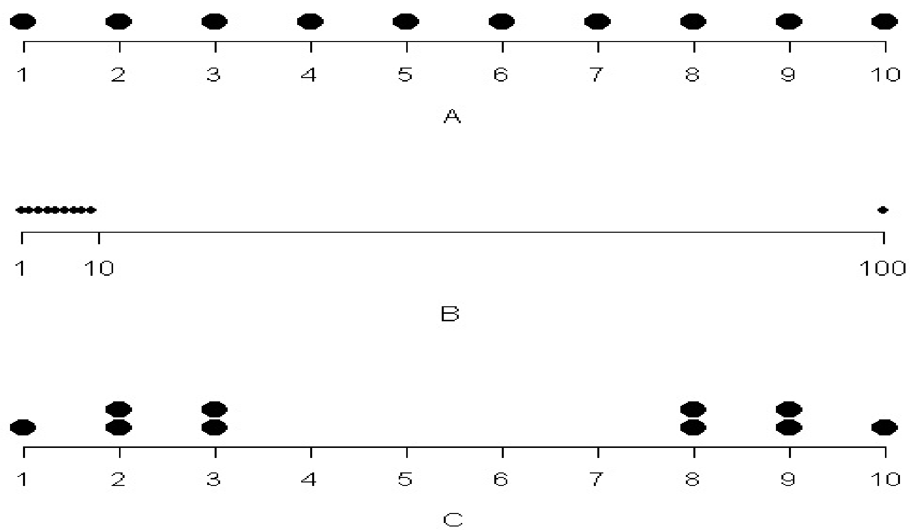
標準差 $s = \sqrt{70.8} = 8.41$, 所以

$$KP_s = \frac{3(\bar{x} - Me)}{s} = \frac{3(65.6 - 68)}{8.41} = -0.856$$

範例- 比較

分別計算下列三組資料的平均值, 中位數, 變異數, 標準差, 平均絕對離差、全距和四分位距。

- A組: 1,2,3,4,5,6,7,8,9,10
- B組: 1,2,3,4,5,6,7,8,9,100
- C組: 1,2,2,3,3,8,8,9,9,10

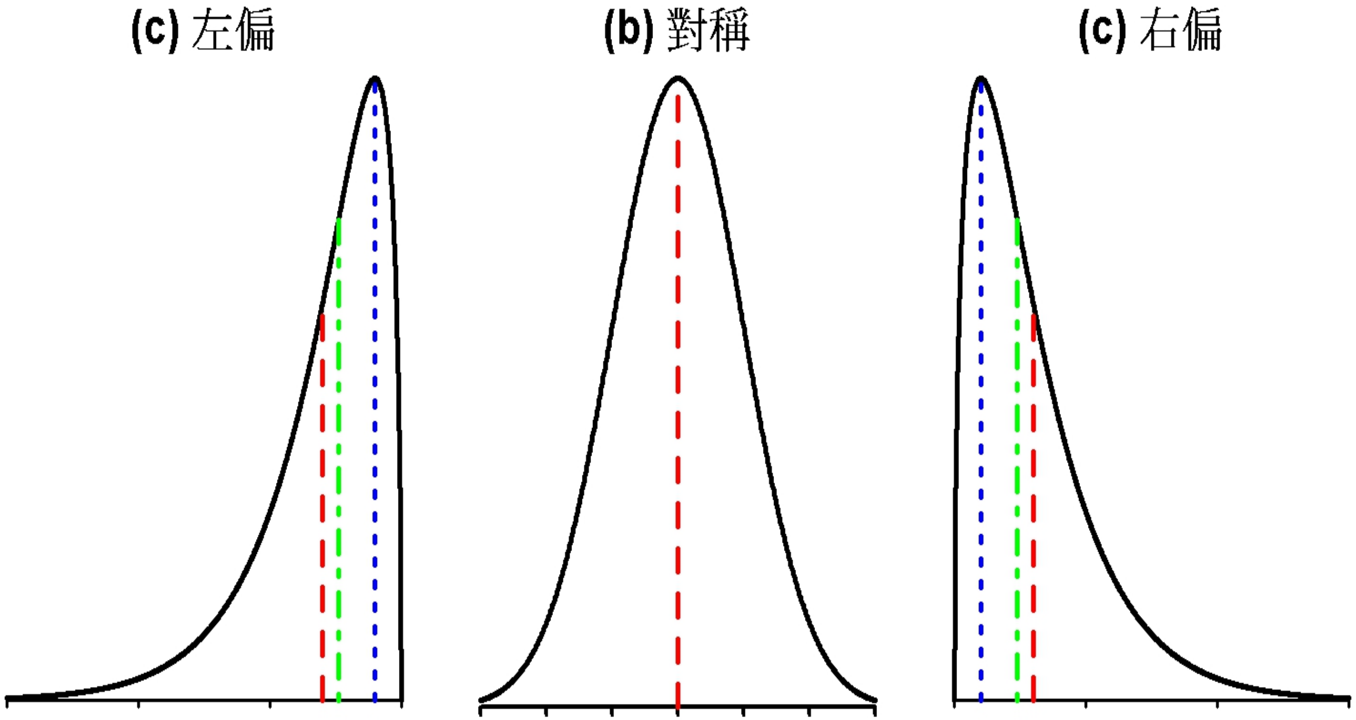


四分位差

四分位距的一半，即

$$QD = \frac{IQR}{2} = \frac{Q_3 - Q_1}{2}$$

偏態係數



偏態- 定性法

- 對稱: 三條線重疊。

$$平均值 = 中位數 = 眾數$$

- 右偏:

$$平均值 > 中位數 > 眾數$$

- 分配:

$$平均值 < 中位數 < 眾數$$

皮爾生偏態係數

皮爾生偏態係數

- 母體

$$KP_s = \frac{3(\mu - Me)}{\sigma}$$

- 樣本

$$KP_s = \frac{3(\bar{x} - Me)}{s}$$

謝比雪夫不等式與經驗法則

使用樣本平均數與標準差知悉分配的方法。

謝比雪夫不等式

- 不論資料的分配為何, 至少有 $1-\frac{1}{k^2}$ 的資料落在以平均數為中心的 $\boldsymbol{k}$ 個標準差之內。或者說
- 不論資料的分配為何, 最多有 $\frac{1}{k^2}$ 的資料落在以平均數為中心的 $\boldsymbol{k}$ 個標準差之外。

範例- 謝比雪夫不等式

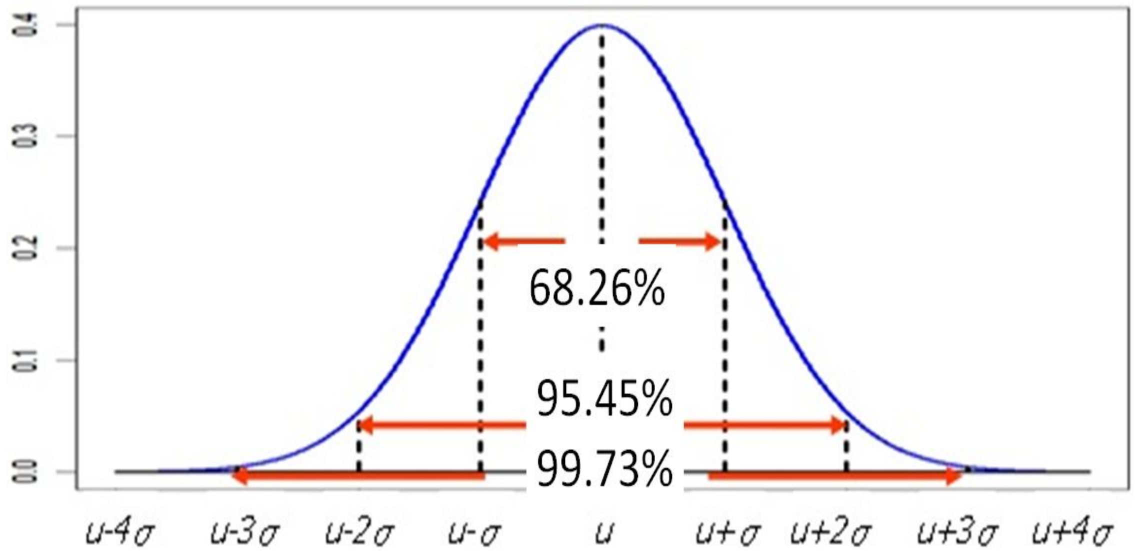
本年級男新生有 1000 人, 假設身高的平均值為 168 分, 標準差為 4 公分。請問:

- 身高介於 160 ～ 176 公分的男同學至少有多少人?
- 身高介於 158 ～ 178 公分的男同學至少有多少人?
- 身高介於 156 ～ 180 公分的男同學至少有多少人?

經驗法則(68-95-99.73法則)

若資料呈現鐘形分配，則

- 約有 68.26% 的資料落在 1 個標準差之內 。
- 約有 95.45% 的資料落在 2 個標準差之內 。
- 約有 99.73% 的資料落在 3 個標準差之內 。



範例- 經驗法則

本年級男新生有 1000 人, 假設身高為鐘型分配平均值為 168 分, 標準差為 4 公分。請問:

- 身高介於 160 ～ 176 公分的男同學約有多少人?
- 身高介於 156 ～ 180 公分的男同學約有多少人?

範例- 經驗法則

平均值為 $\bar{x} = 1491.02$, 標準差為 $s = 50.4$

1379	1428	1459	1462	1472	1484	1501	1513	1543	1566
1385	1448	1460	1465	1474	1488	1505	1516	1546	1568
1412	1449	1460	1469	1482	1495	1509	1517	1547	1569
1425	1451	1461	1470	1483	1500	1512	1520	1548	1601
1427	1455	1461	1470	1483	1500	1512	1533	1554	1614

探索式資料分析

- 對資料有全面性的瞭解。
- 能夠發現過去不知道的資料結構。
- 有無極端值或異常值存在。
- 檢定丨些現象是否仍然保有。
- 提供模型參數簡化的資訊。

探索式資料分析-枝葉圖

50 位同學考試的成績

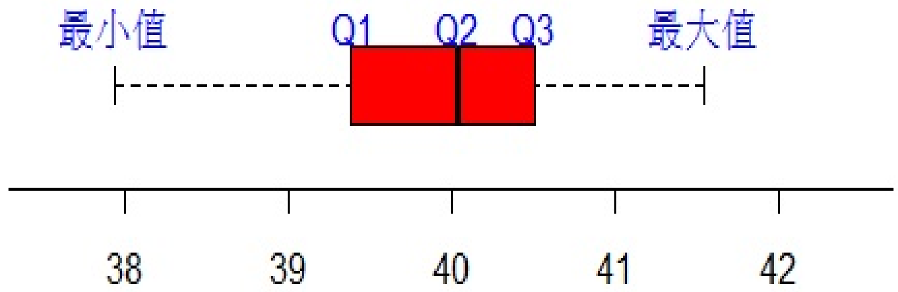
74	72	84	74	54	69	57	64	58	65
64	75	71	86	77	59	76	61	65	75
58	74	85	73	67	58	93	60	80	69
69	68	77	62	59	84	67	74	72	54
70	75	57	69	61	70	72	78	61	60

探索式資料分析-枝葉圖

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   5 | 44
##   5 | 7788899
##   6 | 00111244
##   6 | 557789999
##   7 | 00122234444
##   7 | 5556778
##   8 | 044
##   8 | 56
##   9 | 3
```

探索式資料分析-盒型圖 Box plot

- 資料的集中趨勢。
- 資料的分散趨勢。
- 資料的形狀。
-



是否有極端值。

範例- 盒型圖

1379	1428	1459	1462	1472	1484	1501	1513	1543	1566
1385	1448	1460	1465	1474	1488	1505	1516	1546	1568
1412	1449	1460	1469	1482	1495	1509	1517	1547	1569
1425	1451	1461	1470	1483	1500	1512	1520	1548	1601
1427	1455	1461	1470	1483	1500	1512	1533	1554	1614

範例- 盒型圖

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1379    1460    1484    1491    1517    1614
```