大數據資料處理實務

李水彬

2023-09-01

Chapter 06 統計量-集中趨勢量數

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: 對稱、左偏、右偏
    - 峰態: 高狹或平闊

集中趨勢-平均數

- 數據 $x_1, \cdots, x_n$ 平均值的定義為

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

n: 樣本大小

範例- 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

慧而園幼稚園幼兒的平均體重:

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

```
#mean()函數計算平均值
weight<-c(18, 17, 18, 16, 17, 15, 15, 16, 18, 15, 14, 16, 12, 14, 13, 16, 15, 19, 16, 16)
weight.mu<-mean(weight)
```

經濟成長率

台灣1981至2024年的經濟成長率(23,24年為預測值):

7.11	4.8	9.04	10.05	4.81	11.52	12.7	8.02	8.75	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
3.67	2.22	2.48	4.72	1.47	2.17	3.31	2.79	3.06	3.39
6.62	2.59	1.42	3.35						

平均經濟成長率

```
rate<-c(7.11, 4.80, 9.04,10.05,4.81,11.52,12.70,8.02,8.75, 5.65, 8.36, 8.29,6.80,
7.49,6.50,6.18,6.11,4.21,6.72,6.42,-1.26,5.57,4.12,6.51,5.42,5.62,6.52,0.70,-1.57,
10.63, 3.67,2.22,2.48,4.72,1.47,2.17,3.31,2.79,3.06,3.39,6.62,2.59,1.42,3.35)#單位%
year.count<-length(rate)#year.count=44
rate.mu<-mean(rate)
```

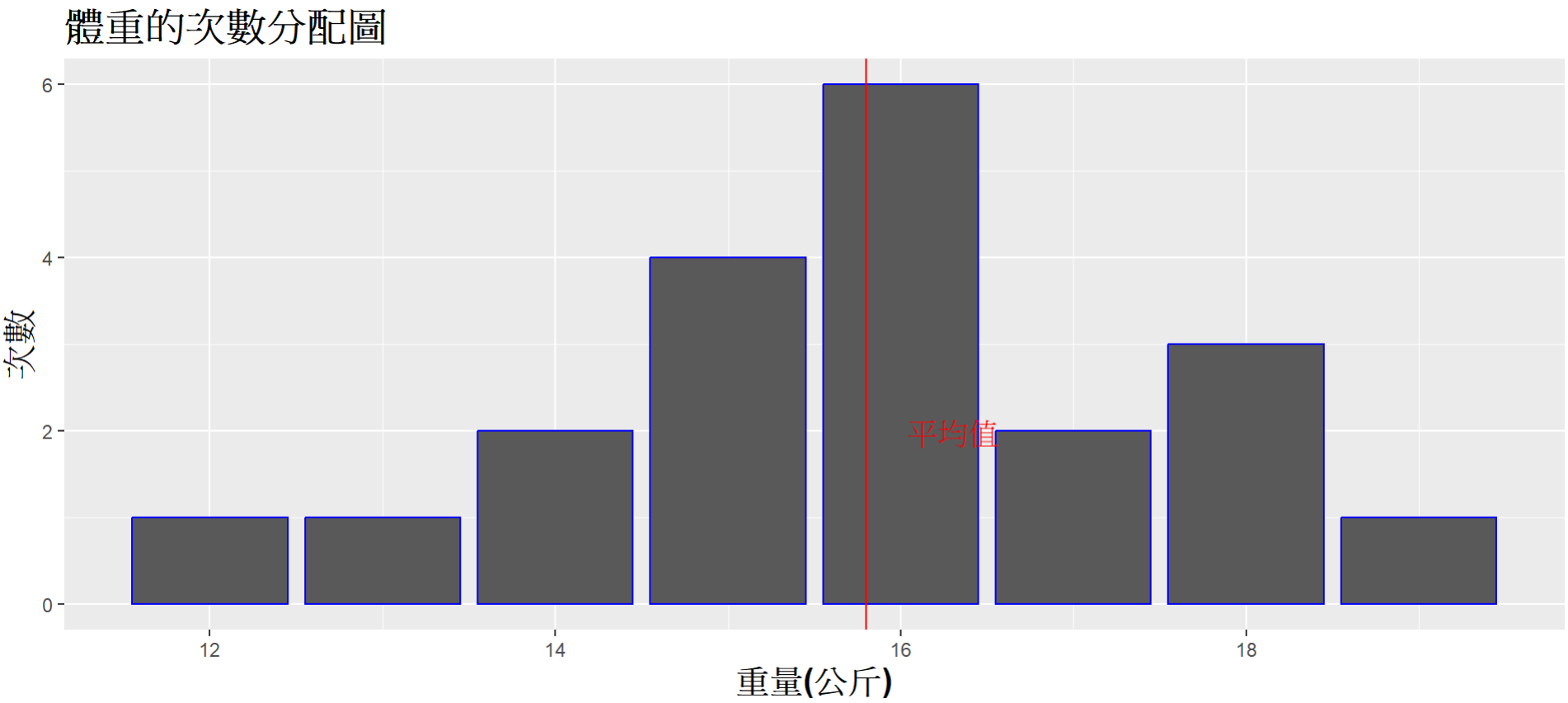
```
year.count;rate.mu
```

```
## [1] 44
```

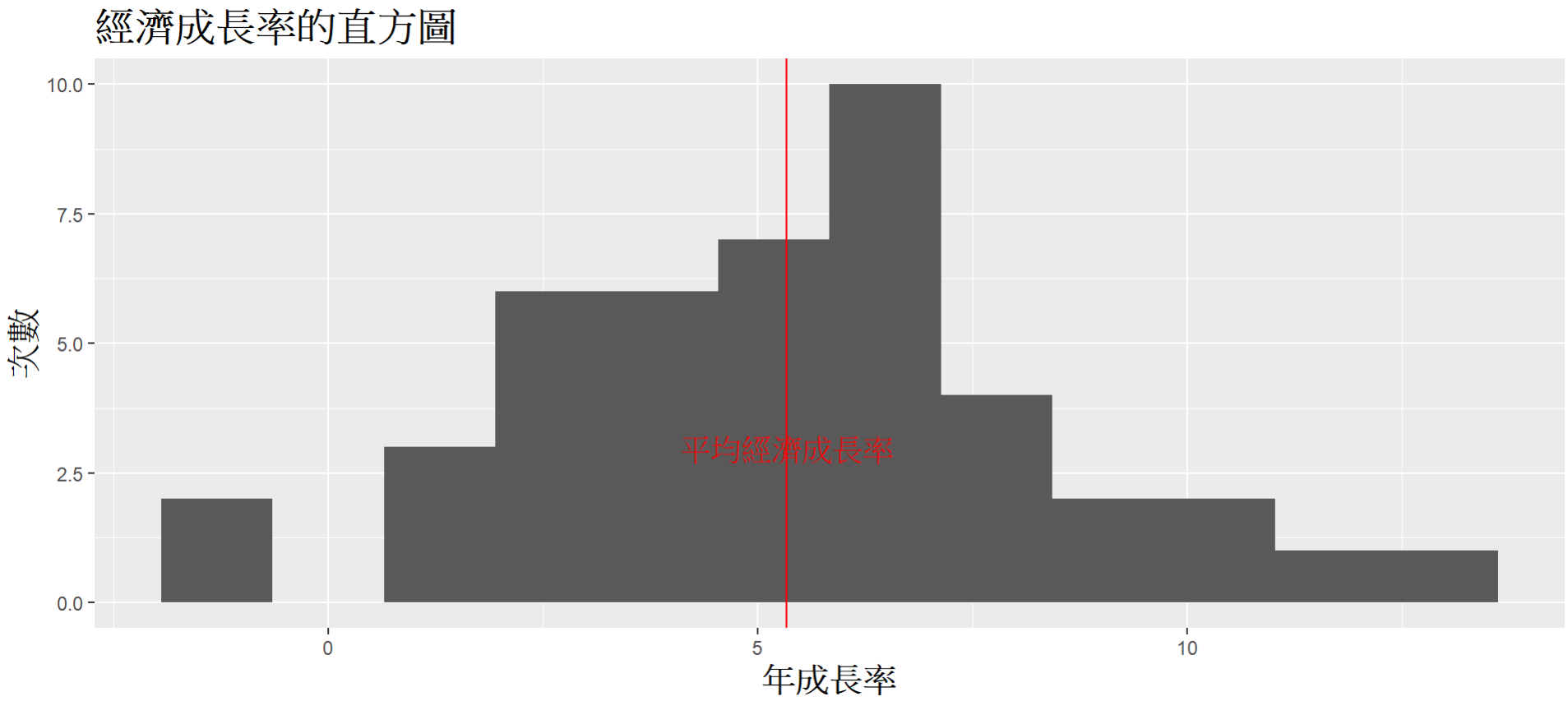
```
## [1] 5.342045
```

1981~2024共44個年份，這段時間的平均經濟成長率為5.34%。

平均數的幾何意義



平均數的幾何意義



中位數

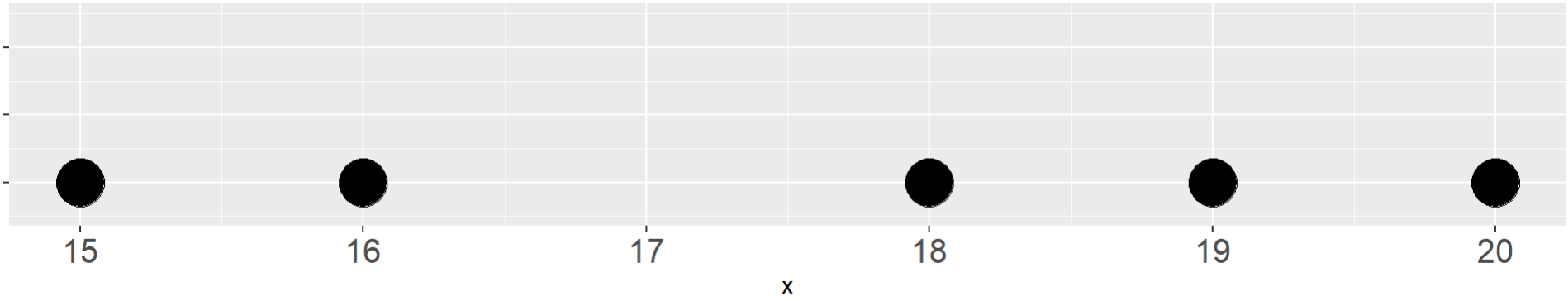
中位數位於 數據排序上的中間位置, 即有至少一半的數據大於或等於此數, 也有至少一半的數據小於或等於此數。

這組數據的中位數

15, 18, 16, 19, 20

```
x<-c(15, 18, 16, 19, 20)
median(x)# median() 函數求中位數
```

```
## [1] 18
```



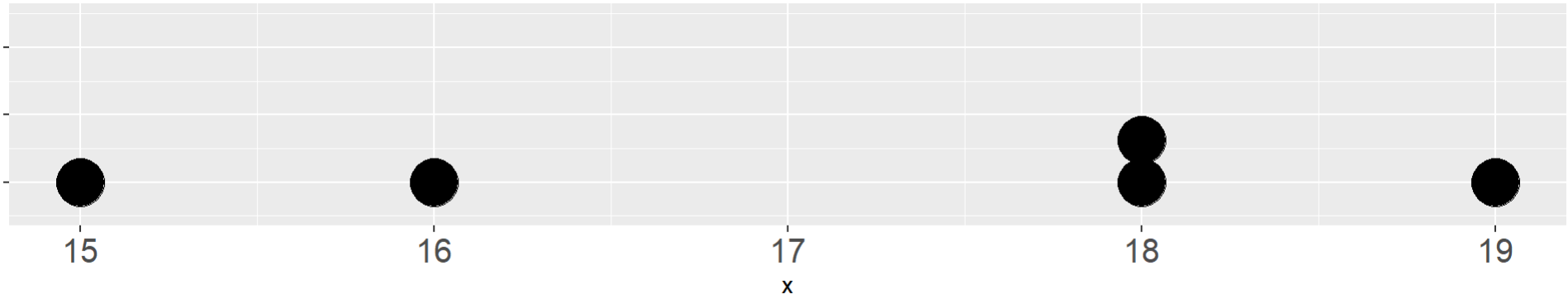
中位數

這組數據的中位數

15, 18, 16, 19, 18

```
x<-c(15, 18, 16, 19, 18)
median(x)# median() 函數求中位數
```

```
## [1] 18
```



中位數

這組數據的中位數

14,15, 18, 16, 19, 20

```
x<-c(14,15, 18, 16, 19, 20)
sort(x)
```

```
## [1] 14 15 16 18 19 20
```

```
median(x)# median() 函數求中位數
```

```
## [1] 17
```

壽司店近48天銷售盤數

壽司店近48天銷售盤數排序後如下:

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

中位數

```
x
```

```
##  [1]  493  522  522  545  627  689  833  861  901  917  939  947 1004 1077 1094
## [16] 1221 1383 1390 1500 1507 1511 1511 1571 1590 1703 1714 1729 1740 1837 1880
## [31] 1927 1936 2111 2173 2359 2437 2491 2504 2526 2555 2650 2662 3003 3193 3229
## [46] 3676 4878 5060
```

```
x[c(24,25)]
```

```
## [1] 1590 1703
```

```
median(x)
```

```
## [1] 1646.5
```

眾數

眾數(mode)：數據中重複次數最多的數。

眾數

- 類別變數或計數型資料，根據定義求算。
- 使用table()函數統計各種項目出現的次數。

```
PhoneBrand<-c("HTC", "iPhone", "HTC", "iPhone", "Samsung", "HTC", "Samsung",
              "iPhone", "Samsung", "Samsung", "Sony", "Samsung", "Sony",
              "iPhone", "Nokia", "iPhone", "Sony", "Asus", "HTC", "Huawei",
              "HTC", "iPhone", "Blackberry", "Blackberry", "Sony", "HTC",
              "Huawei", "Samsung", "Samsung", "iPhone") #手機品牌的調查數據
tt<-table(PhoneBrand) #次數分配表
```

	Asus	Blackberry	HTC	Huawei	iPhone	Nokia	Samsung	Sony
次數	1	2	6	2	7	1	7	4

- 調查數據中次數最高的品牌為Apple和Samsung。

眾數- 計數型資料

- 每批產品的不良數

```
BadCounts<-c(6, 4, 3, 3, 6, 1, 12, 8, 2, 2, 9, 2, 8, 6, 9, 1, 5, 7, 4, 7, 5, 4,
             7, 9, 8, 6, 4, 5, 7, 5, 8, 4, 4, 13, 4, 3, 1, 3, 4, 3, 6, 5, 6, 7,
             4, 5, 6, 5, 3, 5, 5, 7, 6, 10, 4, 10, 7, 6, 4, 7, 7, 1, 9, 6, 6, 8,
             4, 7, 8, 2, 6, 2, 6, 6, 4, 8, 3, 6, 4, 7, 5, 7, 3, 4, 6, 10, 6, 4,
             9, 10, 3, 7, 4, 6, 3, 11, 2, 9, 11, 9)#不良數
length(BadCounts)#共有100批
```

```
## [1] 100
```

```
tt<-table(BadCounts)
```

眾數- 計數型資料

	1	2	3	4	5	6	7	8	9	10	11	12	13
次數	4	6	10	17	10	18	13	7	7	4	2	1	1

眾數為6，共有18 次。

眾數- 計量型(連續型)資料

- 使用卡爾皮爾森經驗法則:平均數至眾數的距離是平均數至中位數距離的3倍。。

$$眾數 = 平均數 - 3(平均數 - 中位數)$$

```
weight<-c(18, 17, 18, 16, 17, 15, 15, 16, 18, 15, 14, 16, 12, 14, 13, 16, 15, 19, 16, 16)
mu<-mean(weight);md<-median(weight)
mu;md
```

```
## [1] 15.8
```

```
## [1] 16
```

- 體重的眾數

```
## [1] 16.4
```

課堂練習 13

求新生兒體重的平均值、中位數和眾數。

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

課堂練習 14

以下是早餐點某日飲料的銷售紀錄，請問哪種飲料銷售杯數最高？

```
drinks<-c("美式咖啡", "米漿", "奶茶", "美式咖啡", "豆漿", "美式咖啡", "奶茶", "拿鐵",
  "拿鐵", "拿鐵", "拿鐵", "紅茶", "拿鐵", "鹹豆漿", "美式咖啡", "奶茶",
  "美式咖啡", "拿鐵", "豆漿", "米漿", "奶茶", "美式咖啡", "豆漿", "豆漿", "米漿",
  "豆漿", "鹹豆漿", "奶茶", "拿鐵", "紅茶", "奶茶", "豆漿", "豆漿", "奶茶",
  "美式咖啡", "拿鐵", "拿鐵", "米漿", "美式咖啡", "紅茶", "米漿", "美式咖啡",
  "拿鐵", "拿鐵", "拿鐵", "美式咖啡", "美式咖啡", "奶茶", "奶茶", "紅茶", "奶茶",
  "拿鐵", "奶茶", "鹹豆漿", "豆漿", "美式咖啡", "紅茶", "米漿", "拿鐵", "拿鐵",
  "奶茶", "豆漿", "紅茶", "美式咖啡", "鹹豆漿", "美式咖啡", "美式咖啡", "奶茶",
  "豆漿", "美式咖啡", "拿鐵", "拿鐵", "拿鐵", "拿鐵", "紅茶", "豆漿", "拿鐵",
  "奶茶", "豆漿", "豆漿", "拿鐵", "拿鐵", "鹹豆漿", "奶茶", "豆漿", "紅茶", "拿鐵",
  "美式咖啡", "紅茶", "美式咖啡", "美式咖啡", "拿鐵", "拿鐵", "美式咖啡",
  "米漿", "豆漿", "奶茶", "豆漿", "美式咖啡", "美式咖啡", "豆漿", "豆漿", "奶茶")
```