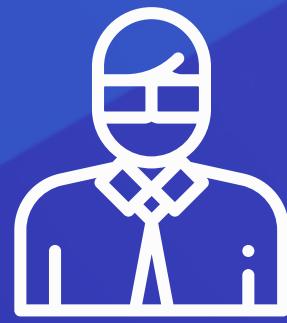




Day 23

## 特徵工程

# 數值型特徵-去除偏態



陳明佑

出題教練

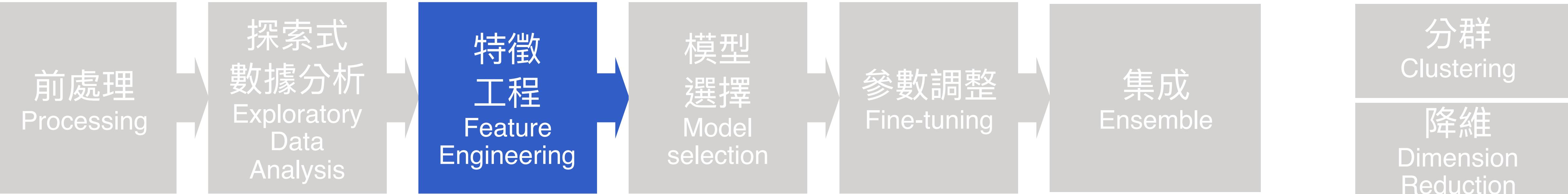


# 知識地圖 特徵工程 數值型特徵 - 去除偏態

## 特徵工程

監督式學習

**Supervised Learning**



非監督式學習  
**Unsupervised Learning**



## 特徵工程 Feature Engineering

**數值型特徵**

**類別型特徵**

**時間型特徵**

**填補缺值**

去離群值

**概論**

去偏態

特徵縮放

類別型特徵處理

時間型特徵處理

特徵組合  
特徵篩選

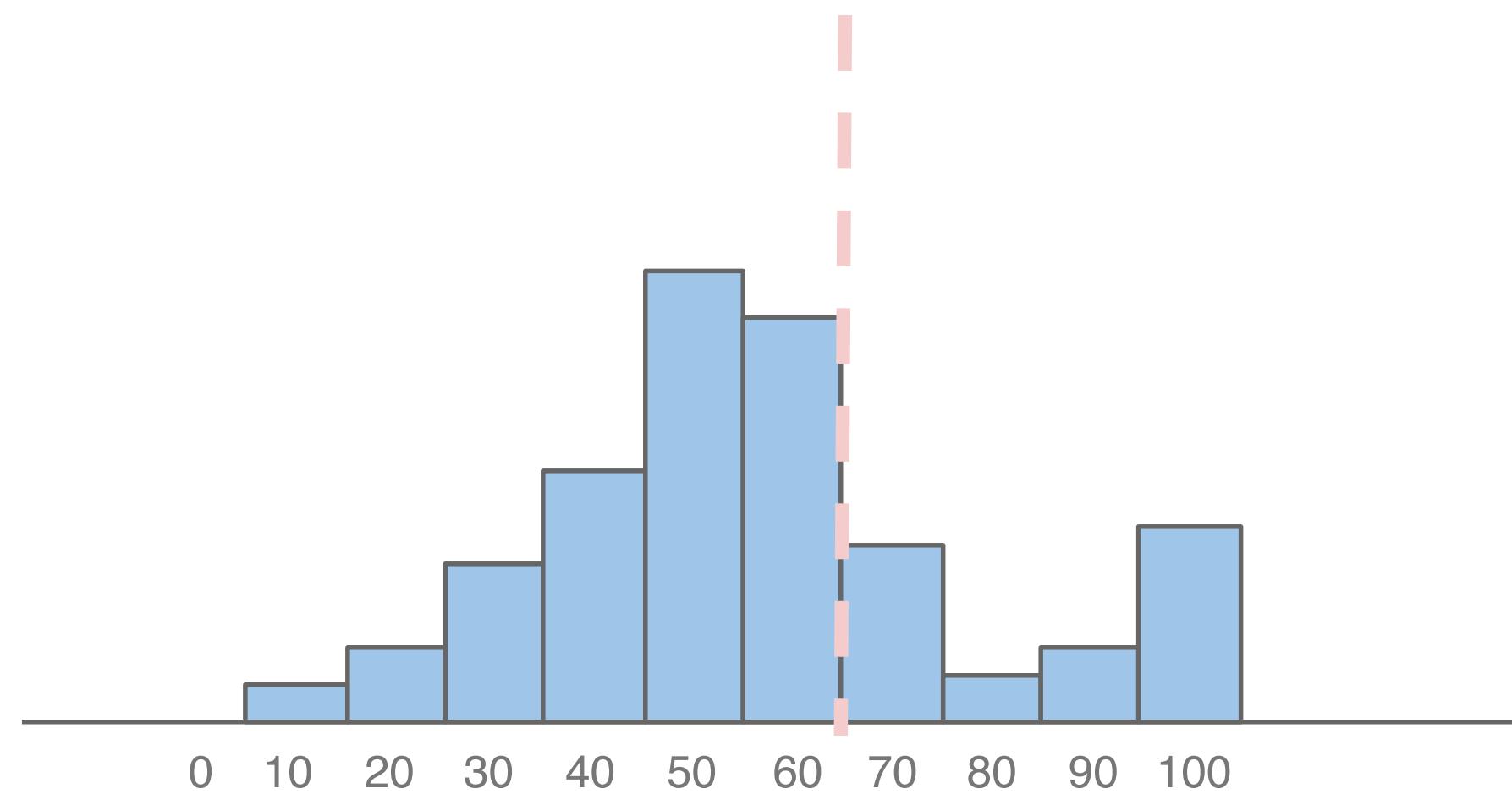
特徵評估

# 本日知識點目標

- 在哪些情況下，需要對資料去偏態
- 去除偏態有哪幾種方式？
- 使用 box-cox 去除偏態時，該注意什麼細節？

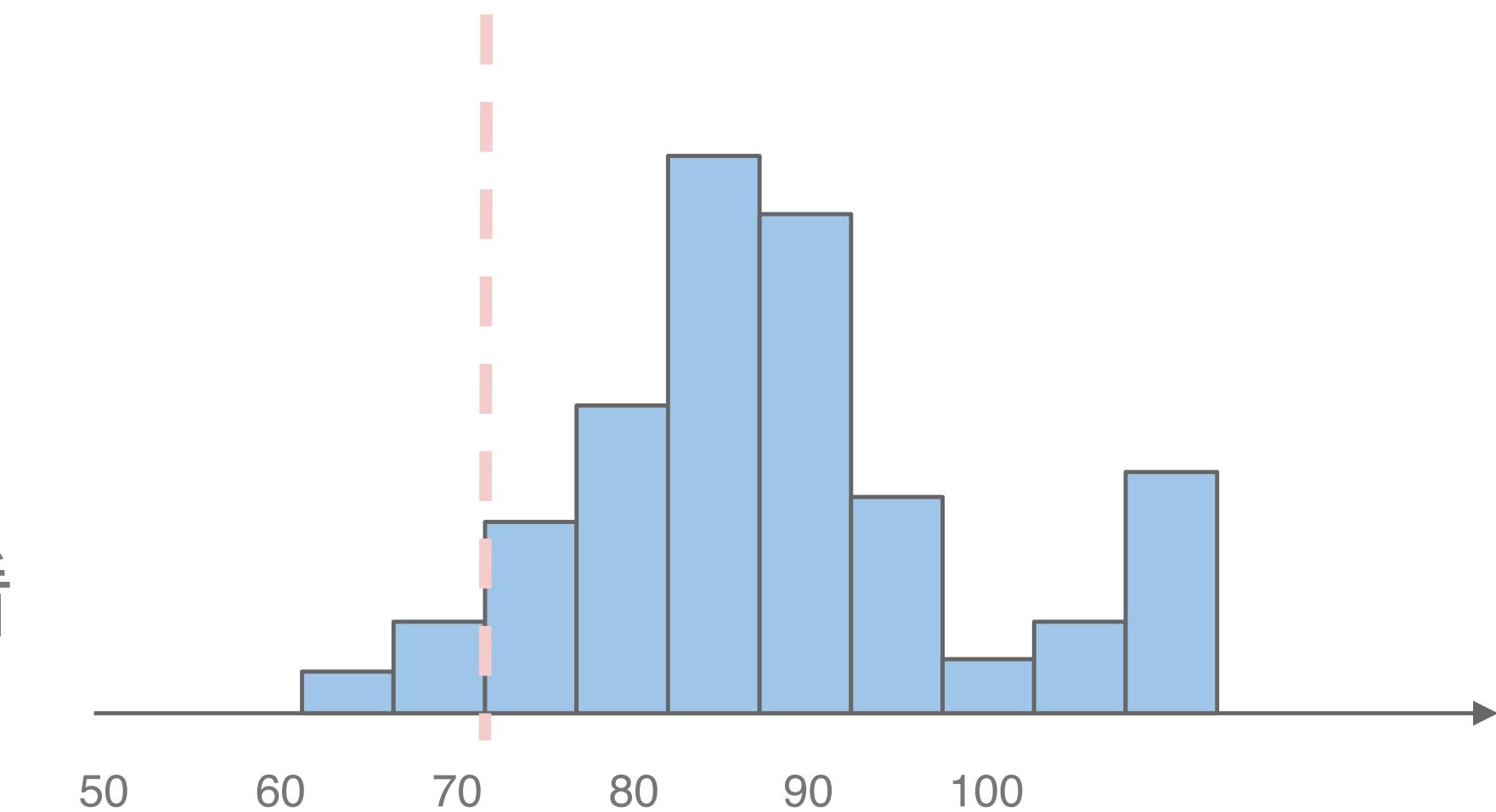
# 去除偏態 (1 / 3)

如果你是老師，某次成績分布如左圖  
希望當掉的同學不要太多( 讓大部分同學都過 )，你該怎麼做？

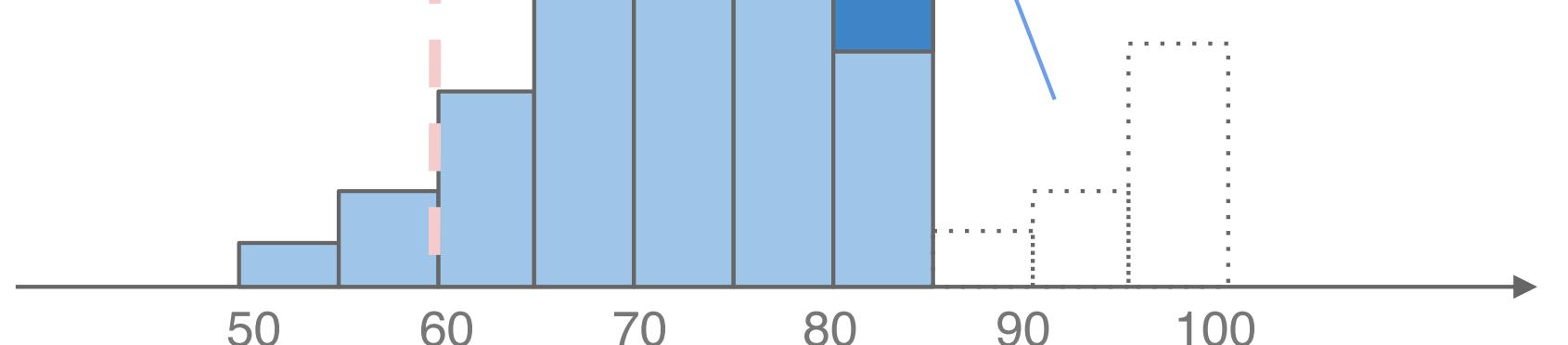


太多人當掉了，QQ!!

標準化(平移)  
高低分群體還是  
分得太明顯，不好看

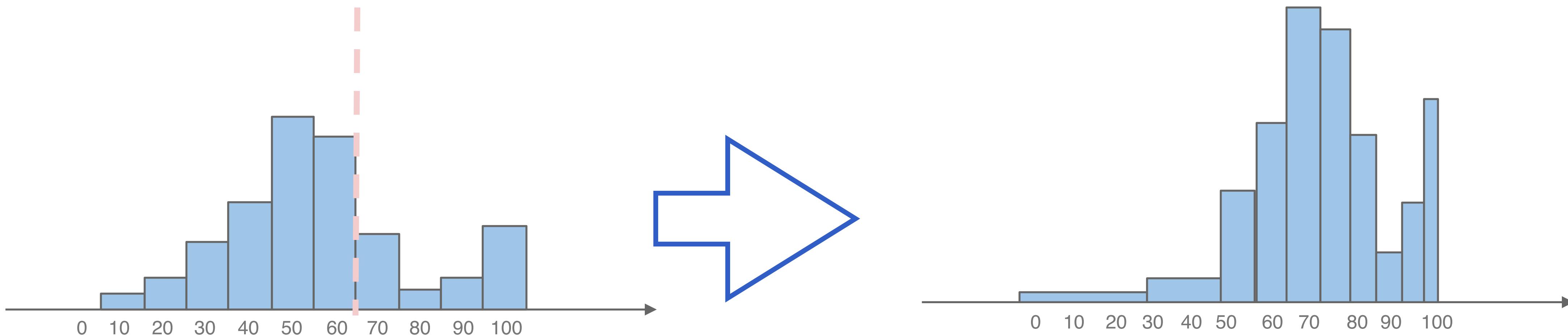


去離群值  
高分群的努力  
都白費了，不公平



# 去除偏態 (2 / 3)

去除偏態：開根號乘以 10



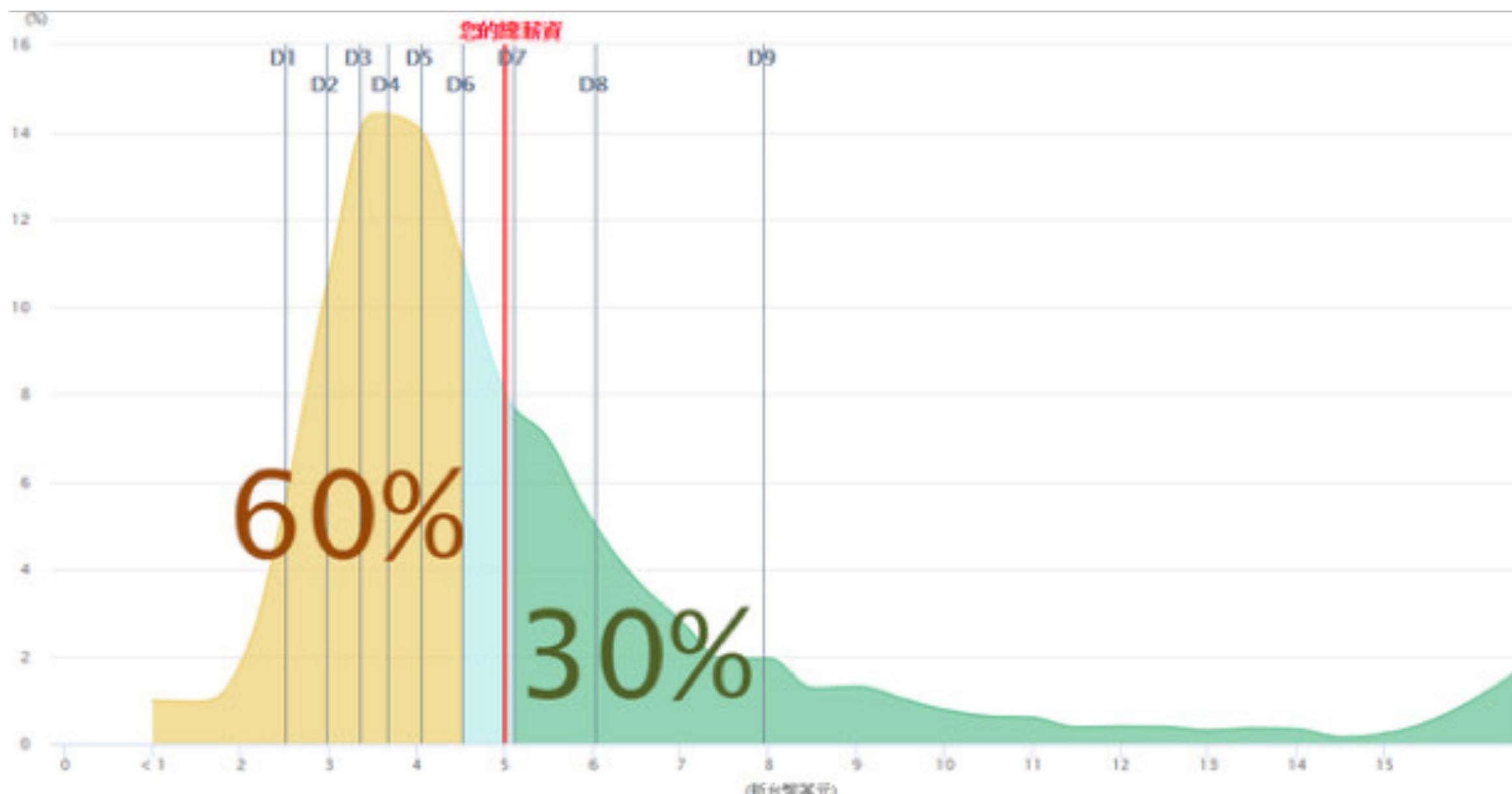
感覺上：考試成績分布越集中在中央，似乎越理想

(其實我們所謂的感覺更合理，意思就是越接近常態分布)

去除偏態的目標在於讓數值更接近常態分布(左右對稱，集中點在中央)，讓平均值更具有代表性。

# 去除偏態 (3 / 3)

平均值更具有**代表性**又是什麼意思？



例：台灣整體薪資分布  
圖源：行政院主計處

**平均值**(紅線處) 不具有**代表性**  
**中位數**(D5線) 反而較具**代表性**

薪資分布中，高薪群的長尾分布造成平均值不具**代表性**  
但是對數去偏後的新分布，平均值就比較具有**代表性** (請見下頁)

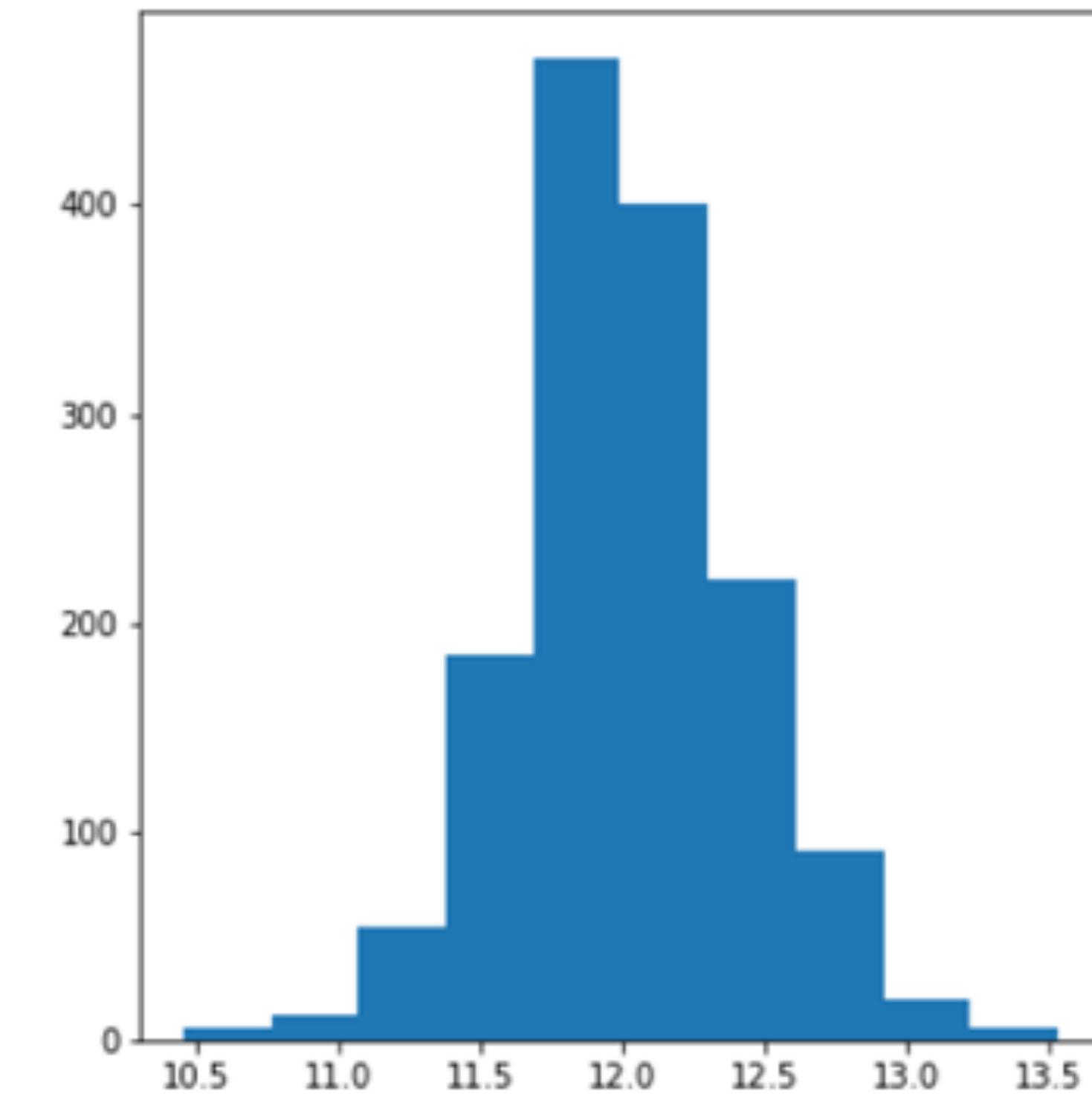
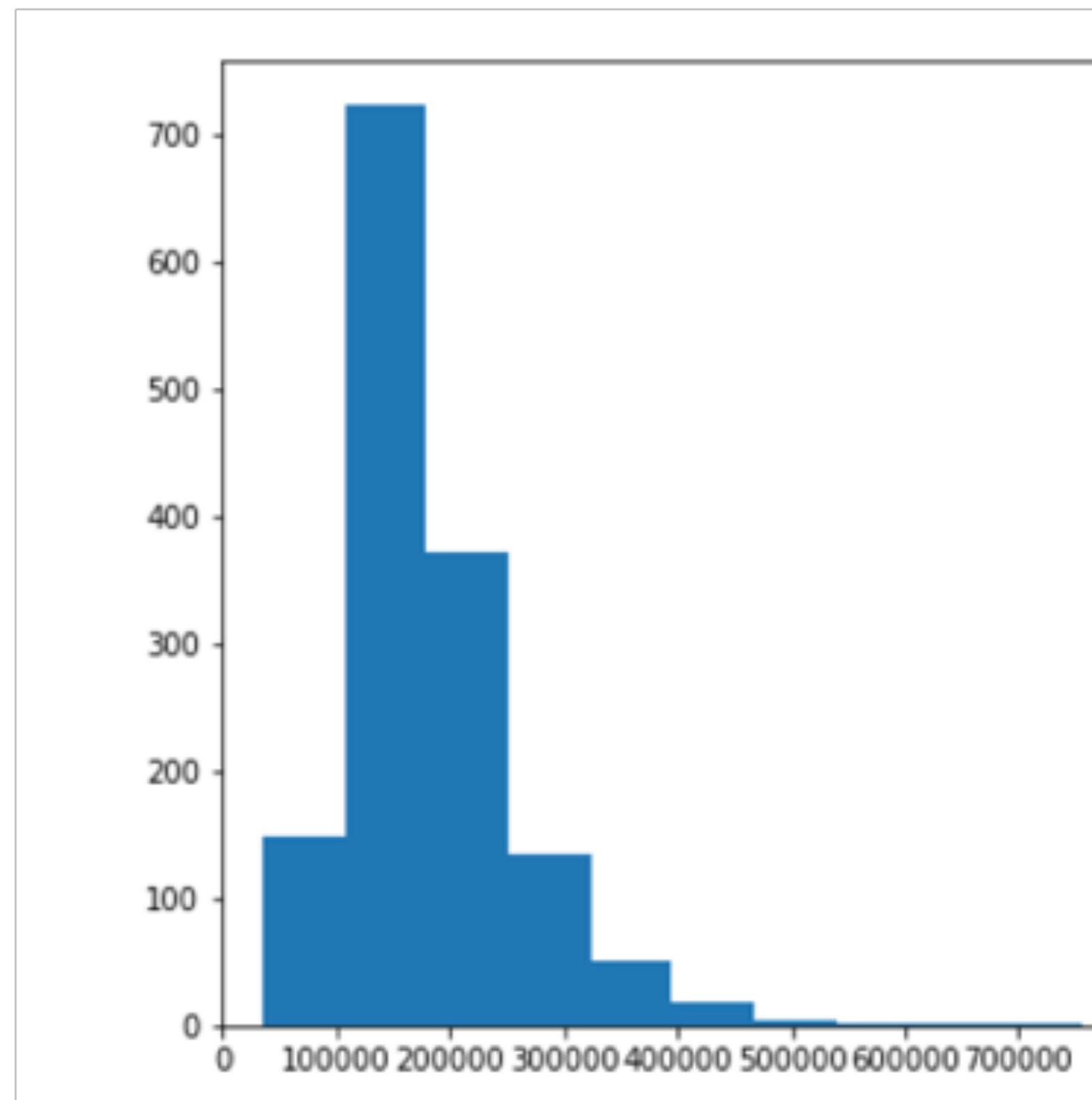
# 複習：對數去偏(log1p)

對數去偏就是使用**自然對數**去除偏態

常見於計數 / 價格這類非負且可能為 0 的欄位

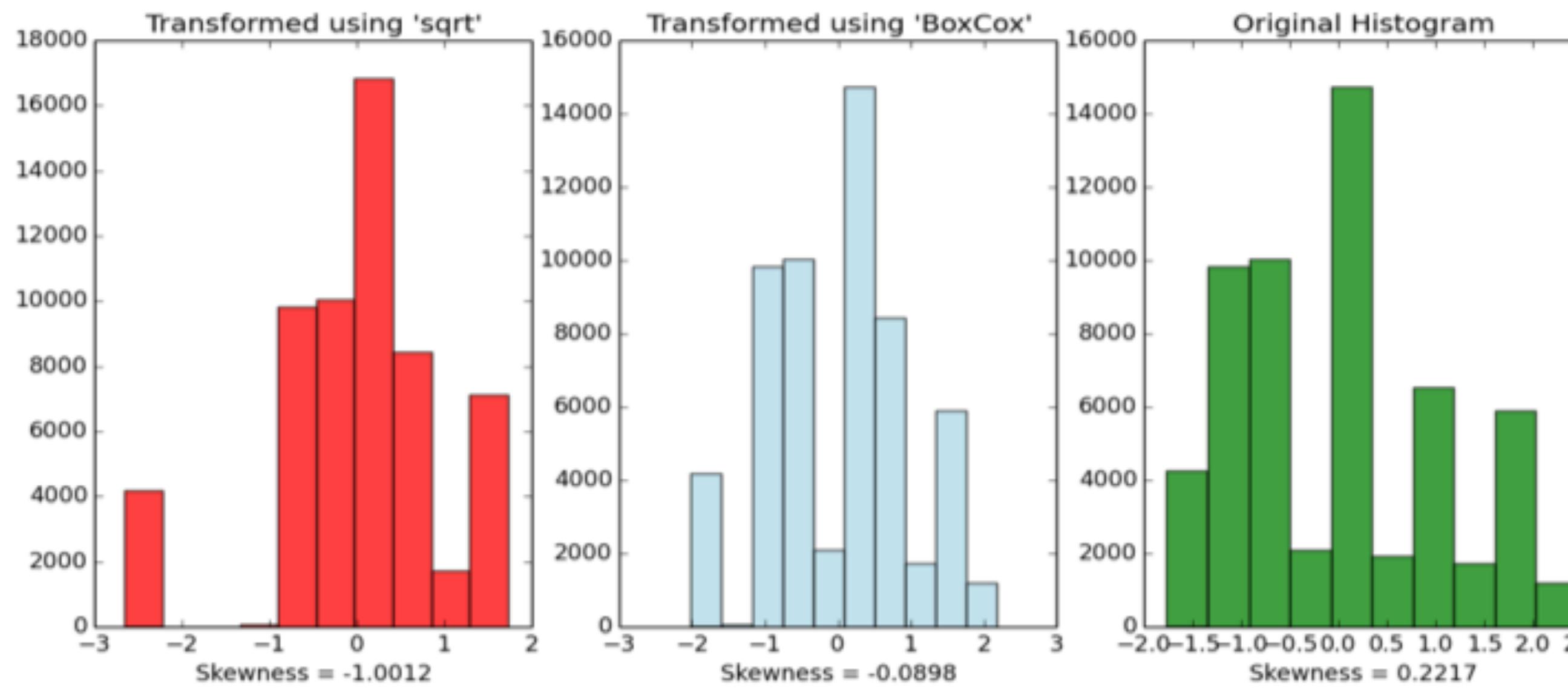
因為需要將 0 對應到 0，所以先加一 (plus one) 再取對數 (log)

還原時使用 expm1，也就是先取指數 (exp) 後再減一 (minus one)



# 方根去偏(sqrt) / 分布去偏(boxcox)

- 方根去偏(sqrt) 就是將數值減去最小值後開根號，最大值有限時適用 (例：成績轉換)
- 分布去偏(boxcox) 是採用boxcox轉換函數(詳見右表)，函數的 lambda( $\lambda$ ) 參數為 0 時等於 log 函數，lambda( $\lambda$ ) 為 0.5 時等於開根號 (即sqrt)，因此可藉由參數的調整更靈活地轉換數值，但要特別注意Y的輸入數值必須要為正 (不可為0)

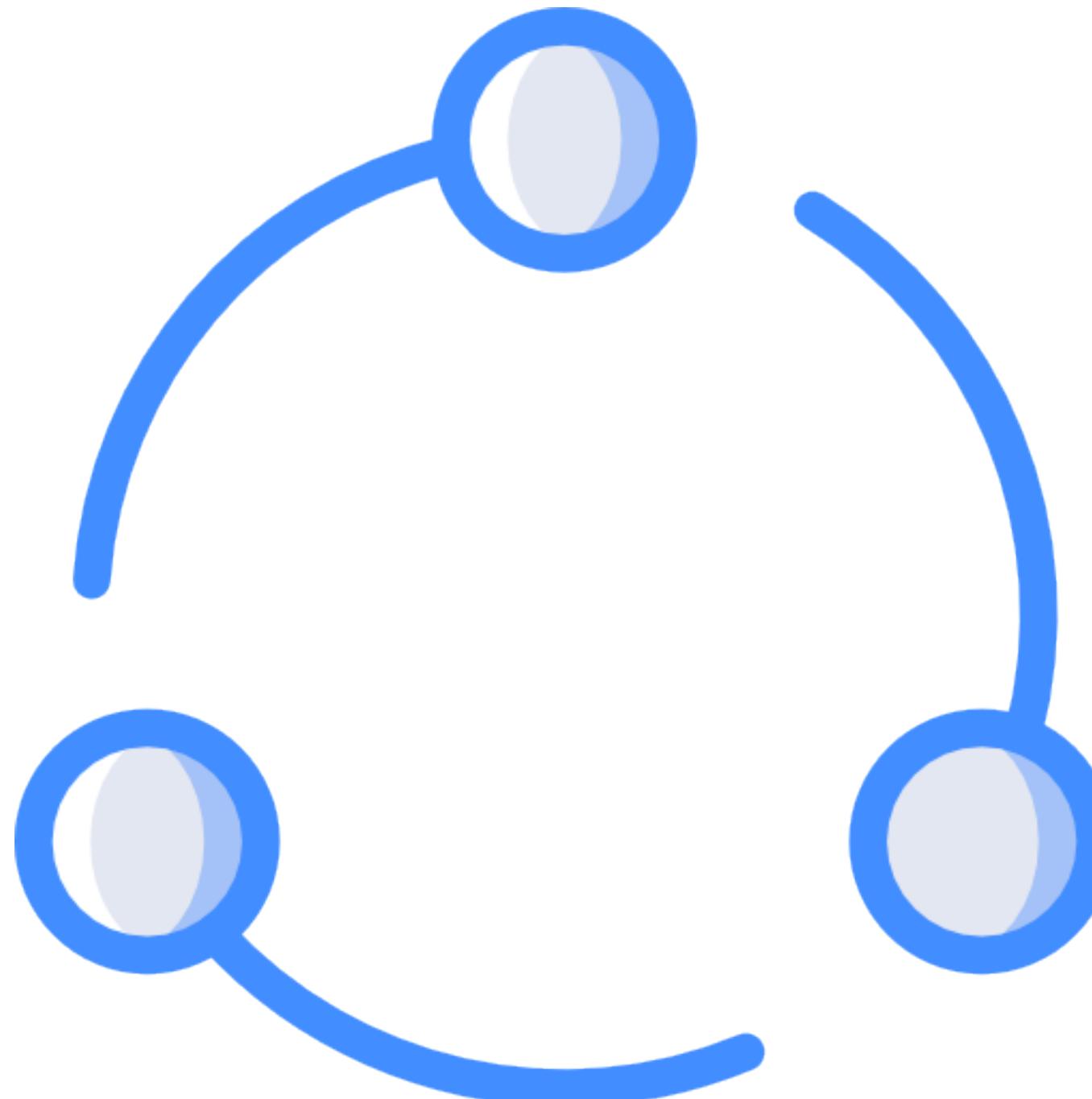


boxcox 參數對照表

commonly used exponents	
$\lambda$	$Y$
-2	$\frac{1}{Y^2}$
-1	$\frac{1}{Y}$
-0.5	$\frac{1}{\sqrt{Y}}$
0	$\log Y$
0.5	$\sqrt{Y}$
1	$Y$
2	$Y^2$

Legend: logarithmic transformation, inverse transformation, square root transformation, no transformation, quadratic transformation.

# 重要知識點複習



- 當離群資料**比例太高**，或者**平均值沒有代表性**時，可以考慮去除偏態
- 去除偏態包含：對數去偏、方根去偏以及分布去偏
- 使用 box-cox 分布去偏時，除了注意入參數要介於 **0 到 0.5 之間**，並且要注意轉換前的數值**不可小於等於 0**



# 延伸 閱讀

除了每日知識點的基礎之外，推薦的延伸閱讀能補足學員們對該知識點的了解程度，建議您解完每日題目後，若有  
多餘時間，可再補充延伸閱讀文章內容。

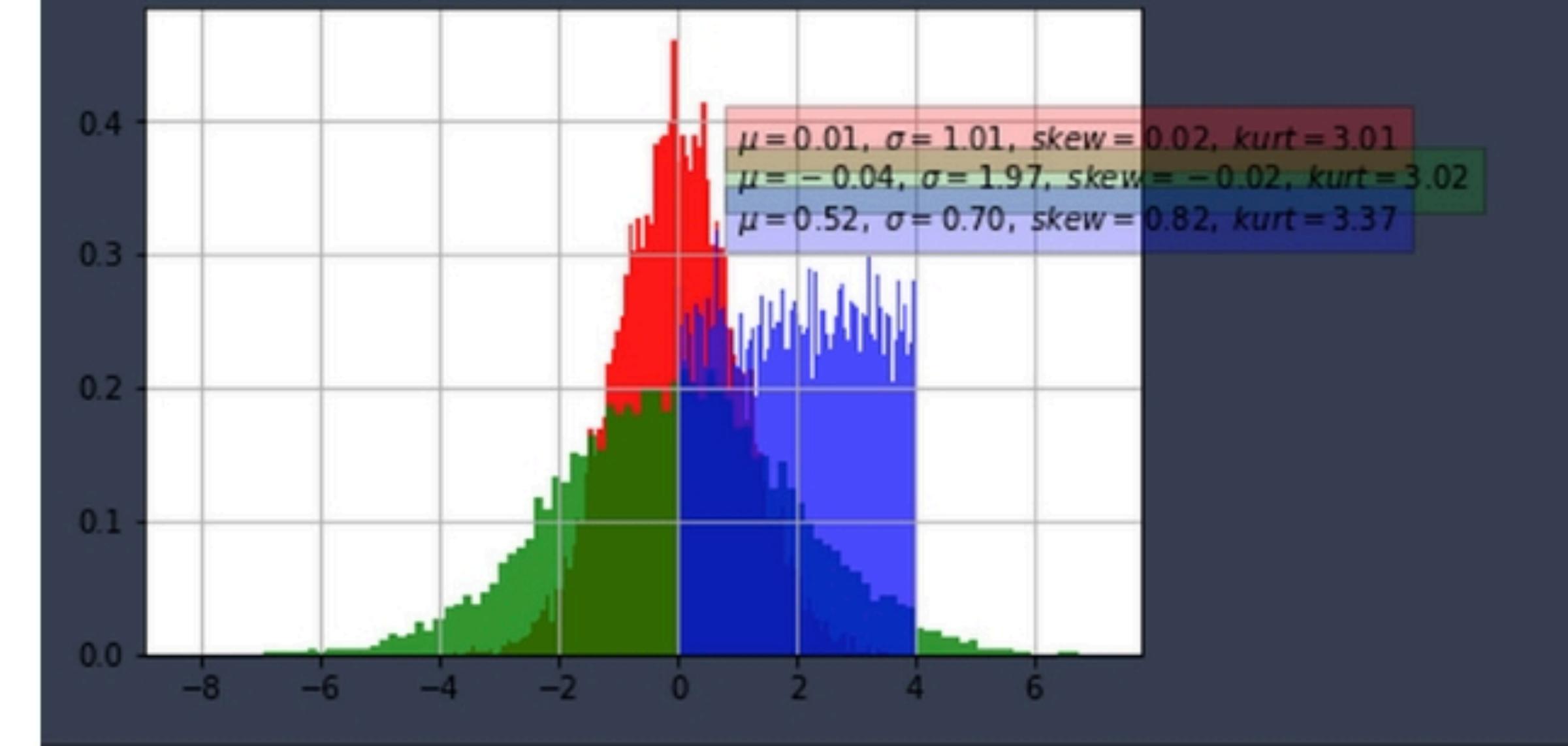
# 推薦延伸閱讀

## 機器學習數學|偏度與峰度及其 python 實現 ()

程式前沿 網頁連結

- 本文提到了峰度與偏度的定義與統計上的意義之外，也提供了Python 上的實現，同學可以參考一下這些圖片，而把其對應的程式碼當作工具參考。

```
0.00586577633351 1.0056869931818935 0.0224896050811 3.01444134137  
0.0285758241356 1.9697642778503606 -0.0160539192756 3.02387565662  
0.498240177117 0.7048476393783777 0.823117334737 3.37402538925  
2.01426010165 1.156579726239703 -0.0174587688069 1.79713470735
```





解題時間

It's Your Turn

請跳出PDF至官網Sample Code & 作業  
開始解題

