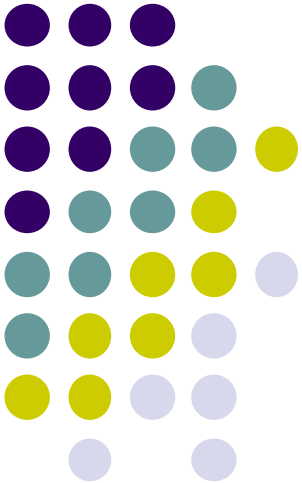
# Ch 4 實習

Numerical Descriptive Techniques



# Agenda

- 了解資料的 分佈型態
  - Mean, Median, Mode
  - Standard Deviation, Variance
  - Chebysheff's Theorem
- 了解資料的相對位子意涵
  - Percentiles
  - Quartiles & Box plot
- 如何知道兩變數的線性關係
  - Covariance, Correlation
- 作業演練

# 一、了解資料的分佈型態

- 林書豪近年來的失誤分數如下（求中位數）：

| 失誤次數 | 0 | 2 | 3 | 0 | 1 | 5 | 4 | 2 |
|---|---|---|---|---|---|---|---|---|

| 排序 | 0 | 0 | 1 | 2 | 2 | 3 | 3 | 5 |
|---|---|---|---|---|---|---|---|---|

**Step 1:** 先排序
**Step 2:** 因為有8個數值，所以是取4和5位子的數值
**Step 3:** 中位數 (2+2)/2=2

# 一、了解資料的分佈型態

- 林書豪近今年來的失誤分數如下（求中位數）：

| 失誤次數 | 0 | 2 | 3 | 0 | 1 | 5 | 4 | 2 | 3 |
|---|---|---|---|---|---|---|---|---|---|

| 排序 | 0 | 0 | 1 | 2 | 2 | 3 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|---|---|

**Step 1:** 先排序
**Step 2:** 因為有9個數值，所以是取5位子的數值
**Step 3:** 中位數 **2**

# 一、了解資料的分佈型態

- 黃蜂隊近年來的新進球員身高如下：

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 身高 | 185 | 190 | 188 | 173 | 170 | 173 | 188 | 185 | 190 |
| 排序 | 170 | 173 | 173 | 185 | 185 | 188 | 188 | 190 | 190 |

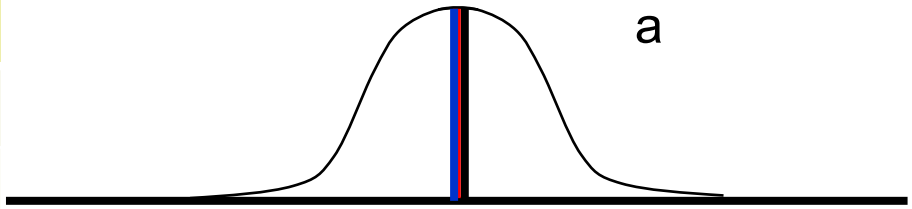
- Mean (平均數)
- Median (中位數)
- Mode（眾數)

- =average(range)=183.3
- 中位數=(9+1)/2=5 (看第五個值,185)
- 173,185,188,190

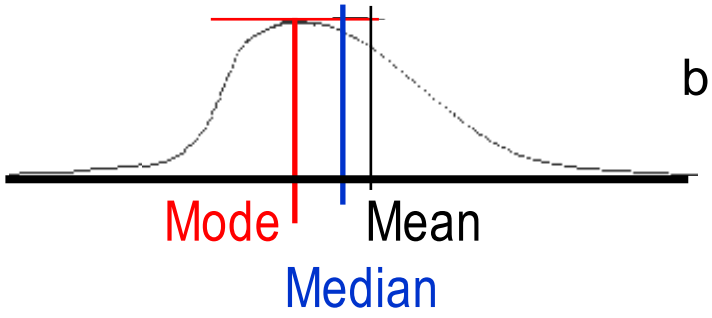
# 一、Relationship among Mean, Median, and Mode

下列三筆資料，應該是哪個圖？Why?

| | mean | median | mode |
|---|---|---|---|
| a | 5 | 5 | 5 |
| b | 5 | 4 | 2 |
| c | 5 | 4 | 7 |

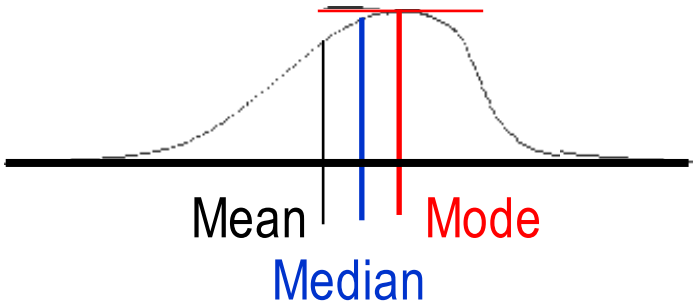



**A positively skewed distribution**
**(“skewed to the right”)**




C

**A negatively skewed distribution**
**(“skewed to the left”)**



# 一、Variance...

- The variance of a **population** is:

population mean

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

population size

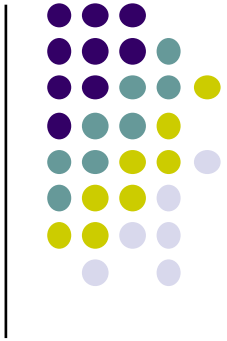
- The variance of a **sample** is:

sample mean

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

Note! the denominator is sample size (n) minus one !

# 一、Variance...



- As you can see, you have to calculate the sample mean (x-bar) in order to calculate the sample variance.
- Alternatively, there is a short-cut formulation to calculate sample variance directly from the data without the intermediate step of calculating the mean. Its given by:0

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]$$

# Proof

$$s^2 = \frac{1}{n-1}\left[\sum (x_i - \bar{x})^2\right]$$

$$= \frac{1}{n-1}\left[\sum (x_i^2 - 2x_i\bar{x} + \bar{x}^2)\right]$$

$$= \frac{1}{n-1}\left[\sum (x_i^2) - 2\bar{x}\sum x_i + n\bar{x}^2\right]$$

$$\bar{x} = \frac{\sum x_i}{n}$$

$$= \frac{1}{n-1}\left[\sum (x_i^2) - 2\frac{(\sum x_i)^2}{n} + \frac{(\sum x_i)^2}{n}\right]$$

$$= \frac{1}{n-1}\left[\sum (x_i^2) - \frac{(\sum x_i)^2}{n}\right]$$

# 一、Variance的應用

- 今天你的好姐妹Lily跟你訴說她的煩惱。有兩個高富帥的男生，同時追她，但是其中一位可能是股票超盤手，但是lily很討厭風險的生活。
- 聰明的你，能否從他們的收入數據，猜測哪位可能是股票超盤手？？他的收入，大概是在哪個範圍？

| | Mean | sd |
|---|---|---|
| Jack | 10 | 8 |
| Steven | 10 | 1 |

Jack可能是賭徒,薪資位於2~18萬間

10-1* sd=10-8=2
10+1* sd=10+8=18

# 一、Coefficient of Variation (CV)

- The ***coefficient of variation*** of a set of observations is the standard deviation of the observations divided by their mean, that is:
  - ✓ Population coefficient of variation = CV = $\sigma / \mu$
  - ✓ Sample coefficient of variation = cv = $s / \bar{x}$

- CV是相對離勢量數（measure of relative dispersion）
  - ✓ 比較幾組資料單位不同的差異情形。(公斤與體重無法比)
  - ✓ 比較幾組資料單位相同，但平均數相差懸殊之差異情形。

$$CV1 = \frac{5\ \cancel{cm}}{170\ \cancel{cm}} \qquad CV2 = \frac{10\ \cancel{kg}}{60\ \cancel{kg}}$$

# 一、Coefficient of Variation (CV)

- 有兩位籃球選手，其得分成績如下，
- 如果你是教練，你會偏好簽下哪個選手？Why?

| | Mean | sd |
|---|---|---|
| Jack | 300 | 20 |
| Steven | 300 | 25 |

a) 選jack, 因為sd小

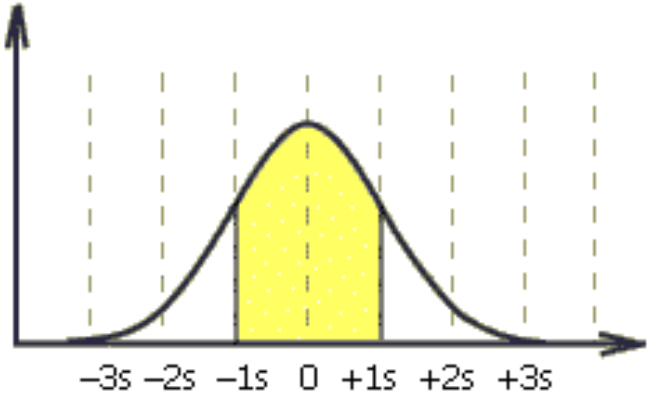
| | Mean | sd |
|---|---|---|
| Jack | 300 | 20 |
| Steven | 250 | 20 |

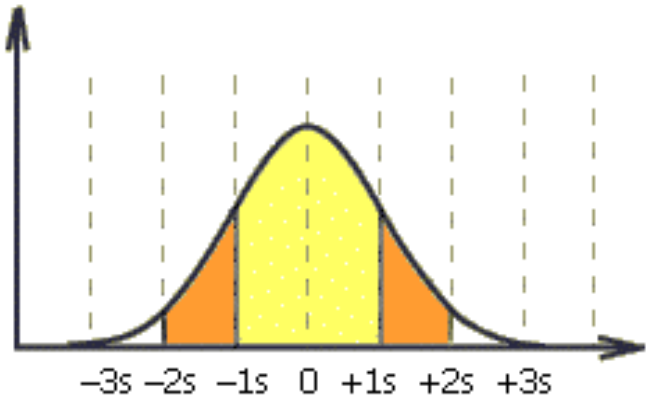
b) 選jack, 因為 mean大 or (CV小)

Jack's CV=20/300
Steven's CV=20/250

# 一、The Empirical Rule(常態分配)

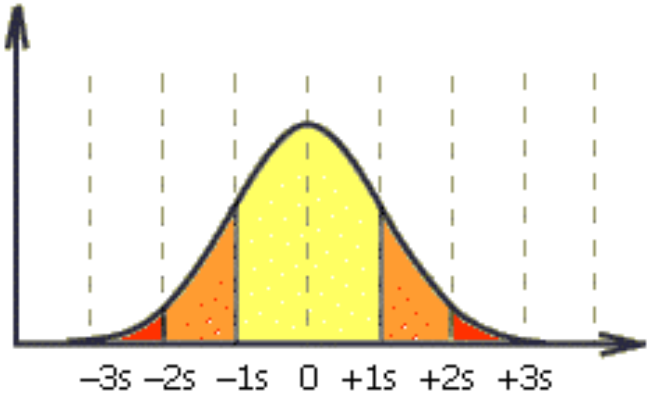
Approximately 68% of all observations fall within **one** standard deviation of the mean.



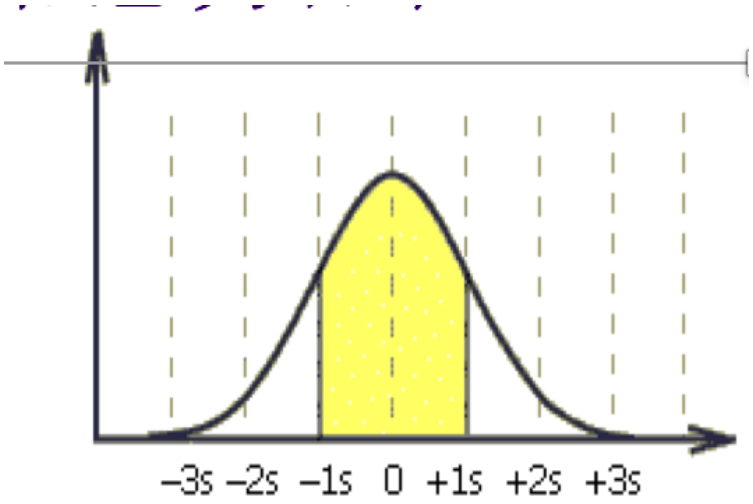

Approximately 95% of all observations fall within **two** standard deviations of the mean.

Approximately 99.7% of all observations fall within **three** standard deviations of the mean.



# 一、The Empirical Rule(常態分配)

- 民調顯示，目前小英總統的聲望，約為 mean=70分, sd=5。假設畫出來的histogram為鐘型分配，請問68%民眾的分數，大概介於幾分到幾分間？

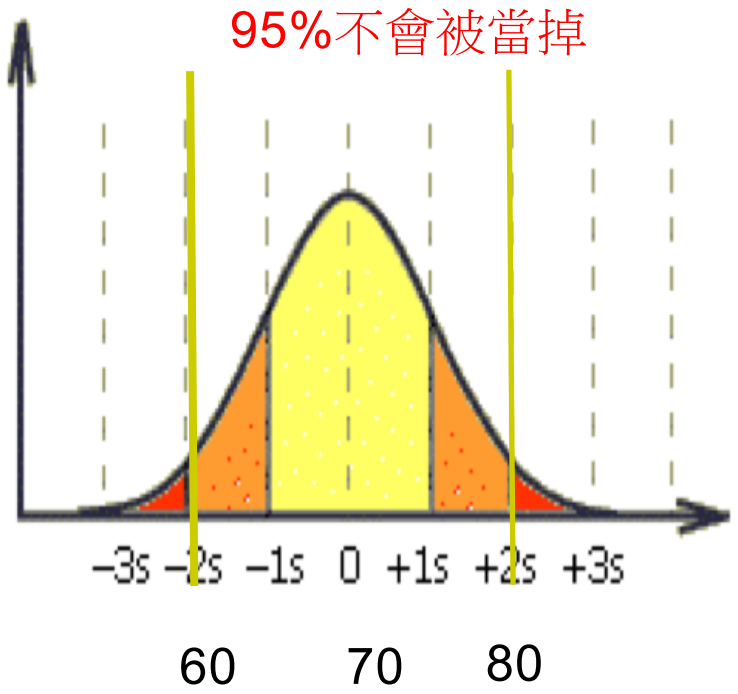



- Step1: 先看為幾倍的標準差
  - ✓ 68%約為一個標準差
- Step2:分數位於 65~75
- (70-1*5=65, 70+1*5=75)

# 一、The Empirical Rule(常態分配)

- 第一次統計期中考，全班分數為 mean=70分, sd=5。假設畫出來的histogram為鐘型分配，請問大概有多少% 學生會被當掉？（提示：先算多少人會過！）

- Step1: 先算出是幾倍標準差
- Ex: (mean-60)/sd= (70-60)/5=2
- Step2:2倍標準差，約為95%
- Step3: (1-95%)/2=不會過的比例
- EX:約有2.5%會被當



# 一、Chebysheff's Theorem（非常態分配）

A more general interpretation of the standard deviation is derived from ***Chebysheff's Theorem***, which applies to all shapes of histograms (not just bell shaped).

The proportion of observations in any sample that lie within **k** standard deviations of the mean is *at least:*

- 知道幾倍標準差，反求有多少比例
- k=2, 則有75%若在此區間
- k=3, 則有90%若在此區間

$$1-\frac{1}{k^2} \quad for \quad k>1$$

For k=2 (say), the theorem states that *at least* 3/4 of all observations lie within 2 standard deviations of the mean. This is a "lower bound" compared to Empirical Rule's approximation (95%).

# 一、Chebysheff's Theorem（非常態分配）

- 班上同學的第一次考試成績平均為60分，標準差為11分，全班共有100人，成績分配不為常態分配，請問班上至少有多少人位於82分和38分之間?

Step 1: 先求幾倍標準差
Ex: (82-38)/11=4

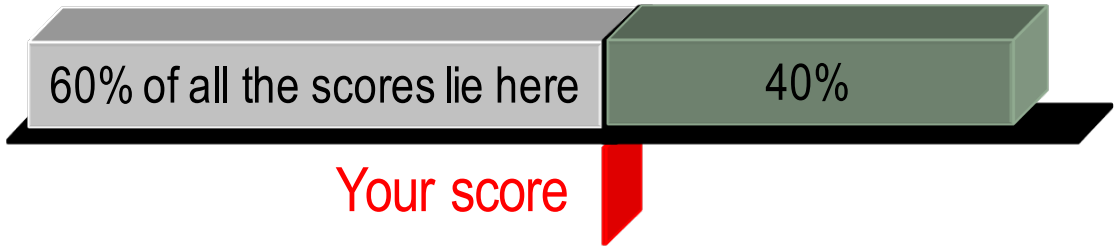
Step2:求出k^2
EX: 表示$k^2$ =4, k=2

Step3: 代入下列公式，可以求出75%
Step4: 75%*100=75人

$$1-\frac{1}{k^2} \quad for \quad k>1$$

# 二、了解資料的相對位子意涵

- **Percentile**
  - The *p*th percentile of a set of measurements is the value for which
    - p percent of the observations are less than that value
    - 100(1-p) percent of all the observations are greater than that value.
  - **Example**
    - Suppose your score is the 60% percentile of a SAT test. Then



# 二、了解資料的相對位子意涵

Quartiles：

- Commonly used percentiles
  - ✓ First (lower) decile = 10th percentile
  - ✓ First (lower) quartile, $\mathbf{Q_1}$, = 25th percentile
  - ✓ Second (middle)quartile, $\mathbf{Q_2}$, = 50th percentile
  - ✓ Third quartile, $\mathbf{Q_3}$, = 75th percentile
  - ✓ Ninth (upper) decile = 90th percentile

為什麼Q1是25%?

因為是四分位數

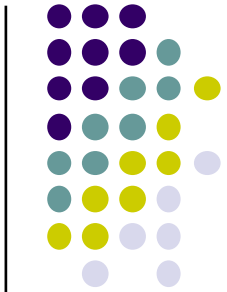
# Location of Percentiles

- Find the location of any percentile using the formula
- 算出在資料中的相對位子
- 有些統計課本，公式會不同

$$L_P = (n+1)\frac{P}{100}$$

where $L_P$ is the location of the $P^{th}$ percentile

# Interquartile Range (IQR)



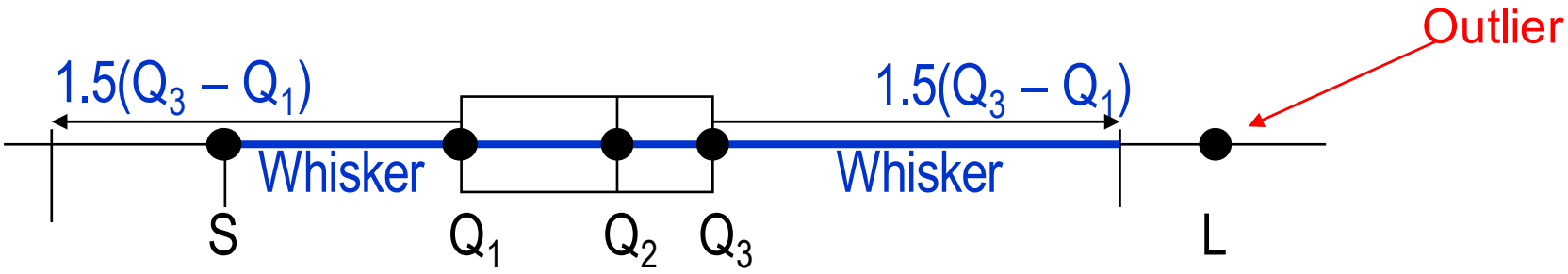
- This is a measure of the spread of the middle 50% of the observations
- Large value indicates a large spread of the observations

$$\text{Interquartile range} = Q_3 - Q_1$$

- Outlier =1.5* IQR

# Box Plot (盒鬚圖)

- This is a pictorial display that provides the main descriptive measures of the data set:
  - L - the largest observation
  - $Q_3$ - The upper quartile
  - $Q_2$ - The median
  - $Q_1$ - The lower quartile
  - S - The smallest observation
  - IQR=Q3-Q1



# Box example

2. A random sample of Boston Marathon runners was drawn and the times to complete the race were recorded.

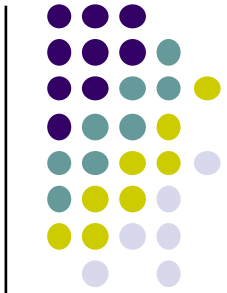
a. Draw the box plot.
b. What are the quartiles?
c. Identify outliers.
d. What information does the box plot deliver?

a) 如何畫圖：data → data analysis plus → box plot

b) Q1, Q2, Q3, IQR, S,L

c) 1.5* IQR

# Example 5 （類似第七題）



- Determine the first, second, and third quartiles of the following data
- Draw the box plot

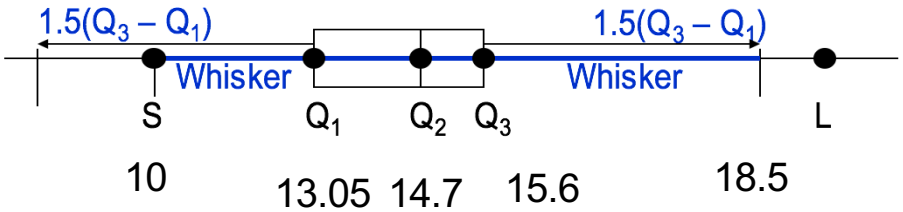
10.5 14.7 15.3 17.7 15.9 12.2 10.0 14.1 13.9
18.5 13.9 15.1 14.7

| 編號 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 分數 | 10 | 10.5 | 12.2 | 13.9 | 13.9 | 14.1 | 14.7 | 14.7 | 15.1 | 15.3 | 15.9 | 17.7 | 18.5 |

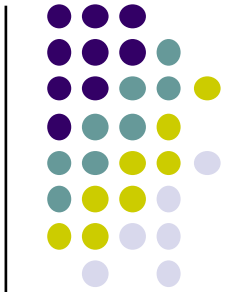
- Step 1:排序後數列
- Step2:求出下列數值
  - ✓ Q1: L25=(13+1)*25/100=3.5
    - the first quartile is 13.05
  - ✓ Q2: L50=(13+1)*50/100=7
    - the first quartile is 14.7
  - ✓ Q3: L75=(13+1)*75/100=10.5
    - the first quartile is 15.6
  - ✓ IQR= 15.6-13.05=2.55
    - 1.5*2.55=3.825
    - 13. 05-3.825=9.225 （左）
    - 15.6+3.825=19.425 （右）
  - ✓ S:10 (左邊鬍鬚畫到10就好)
  - ✓ L:18.5 (右邊鬍鬚畫到18.5)
  - ✓ 此資料顯示，所有數值都介於1.5*IQR(沒有outlier)

3位子的數值：12.2
0.5表示需要3和4位子間，0.5位子的數值:（13.9-12.2)*0.5=0.85
Q1:12.2+0.85=13.05

10位子的數值：15.3
0.5表示需要10和11位子間，0.5位子的數值:（15.9-15.3)*0.5=0.3
Q1:15.3+0.3=15.6



# 三、如何知道兩變數的線性關係



- We now present two numerical measures of linear relationship that provide information as to the **strength & direction** of a linear relationship between two variables (if one exists).
- They are the ***covariance*** and the ***coefficient of correlation***.
- ***Covariance*** - is there any **pattern** to the way two variables move together?
- ***Coefficient of correlation*** - how **strong** is the linear relationship between two variables?

# Covariance…

population mean of variable X, variable Y

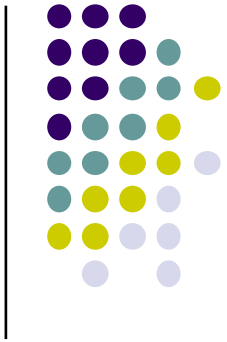
$$\text{Population covariance} = \sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

sample mean of variable X, variable Y

$$\text{Sample covariance} = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Note: divisor is n-1, not n as you may expect.

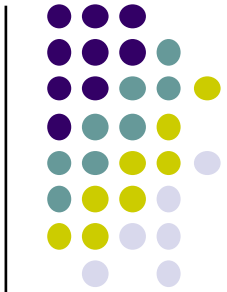
# Covariance...



- In much the same way there was a "shortcut" for calculating sample variance without having to calculate the sample mean, there is also a shortcut for calculating sample covariance without having to first calculate the mean:

$$s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right]$$

# Covariance... (Generally speaking)

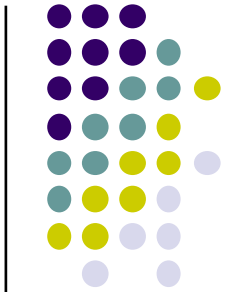


- When two variables move in the ***same direction*** (both increase or both decrease), the covariance will be a ***large positive number***.

- When two variables move in ***opposite directions***, the covariance is a ***large negative number***.

- When there is ***no particular pattern***, the covariance is a ***small number***.

# Coefficient of Correlation...



- The coefficient of correlation is defined as the covariance divided by the standard deviations of the variables:

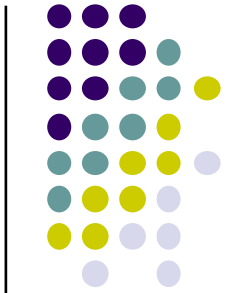
$$\text{Population coefficient of correlation: } \rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Greek letter "rho"

$$\text{Sample coefficient of correlation: } r = \frac{s_{xy}}{s_x s_y}$$

This coefficient answers the question:
How **strong** is the association between X and Y?

# Coefficient of Correlation...

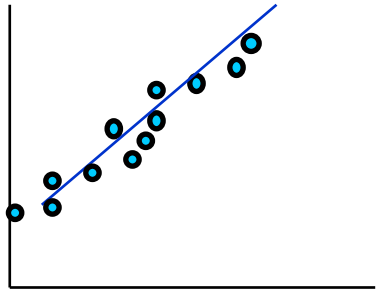


- The advantage of the coefficient of correlation over covariance is that it has fixed range from -1 to +1, thus:
- If the two variables are very strongly positively related, the coefficient value is close to +1 (strong positive linear relationship).
- If the two variables are very strongly negatively related, the coefficient value is close to -1 (strong negative linear relationship).
- No straight line relationship is indicated by a coefficient close to zero.

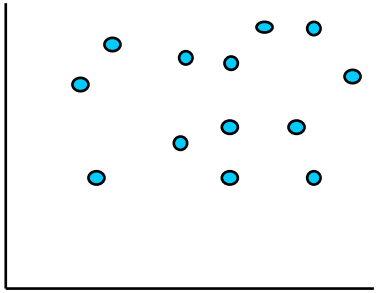
# Coefficient of Correlation...

$\rho$ or r =

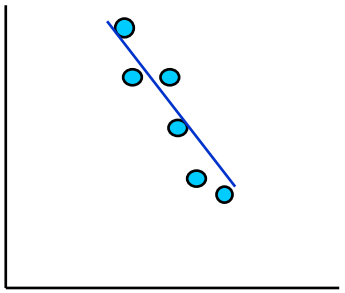
+1 Strong positive linear relationship



0 No linear relationship



-1 Strong negative linear relationship



# Example 6 (類似第八題)

- A retailer wanted to estimate the monthly fixed and variable selling expenses. As a first step she collected data from the past 8 months. The total selling expenses (in $thousands) and the total sales (in $thousands were recorded and listed below.

| Total sales | Selling Expenses |
|---|---|
| 20 | 14 |
| 40 | 16 |
| 60 | 18 |
| 50 | 17 |
| 50 | 18 |
| 55 | 18 |
| 60 | 18 |
| 70 | 20 |

- Compute the covariance and the coefficient of correlation

# Solution

| | $x_i$ | $y_i$ | $x_i^2$ | $y_i^2$ | $x_i y_i$ |
|---|---|---|---|---|---|
| | 20 | 14 | 400 | 196 | 280 |
| | 40 | 16 | 1600 | 256 | 640 |
| | 60 | 18 | 3600 | 324 | 1080 |
| | 50 | 17 | 2500 | 289 | 850 |
| | 50 | 18 | 2500 | 324 | 900 |
| | 55 | 18 | 3025 | 324 | 990 |
| | 60 | 18 | 3600 | 324 | 1080 |
| | 70 | 20 | 4900 | 400 | 1400 |
| Total | 405 | 139 | 22,125 | 2,437 | 7,220 |

$$\sum_{i=1}^{n} x_i = 405 \qquad \sum_{i=1}^{n} y_i = 139 \qquad \sum_{i=1}^{n} x_i^2 = 22{,}125 \qquad \sum_{i=1}^{n} y_i^2 = 2{,}437 \qquad \sum_{i=1}^{n} x_i y_i = 7{,}220$$

$$s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right] = \frac{1}{8-1}\left[7{,}220 - \frac{(405)(139)}{8}\right] = 26.16$$

Cov.

# Solution

$$s_x^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right] = \frac{1}{8-1}\left[22{,}125 - \frac{(405)^2}{8}\right] = 231.7$$

$$s_x = \sqrt{s_x^2} = \sqrt{231.7} = 15.22$$

$$s_y^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right] = \frac{1}{8-1}\left[2{,}437 - \frac{(139)^2}{8}\right] = 3.13$$

$$s_y = \sqrt{s_y^2} = \sqrt{3.13} = 1.77$$

$$r = \frac{s_{xy}}{s_x s_y} = \frac{26.16}{\sqrt{(15.22)(1.77)}} = ..9711$$

Cor.

# Excel內的Cov.是用母體去算

- Example 3:

數值為何不一樣？

**SPSS output**

Correlations

| | | Internet | Education |
|---|---|---|---|
| Internet | Pearson Correlation | 1 | .642** |
| | Sig. (2-tailed) | | .000 |
| | Sum of Squares and Cross-products | 22054.036 | 2888.000 |
| | Covariance | 88.570 | 11.598 |
| | N | 250 | 250 |
| Education | Pearson Correlation | .642* | 1 |
| | Sig. (2-tailed) | .000 | |
| | Sum of Squares and Cross-products | 2888.000 | 918.000 |
| | Covariance | 11.598 | 3.687 |
| | N | 250 | 250 |

**. Correlation is significant at the 0.01 level (2-tailed).

Covariance

| | Internet | Education |
|---|---|---|
| Internet | 88.21614 | |
| Education | 11.552 | 3.672 |

Coefficient of Correlation

| | Internet | Education |
|---|---|---|
| Internet | 1 | |
| Education | 0.641847 | 1 |

# Excel內的Cov用母體去算

- Example 3:

請將excel數值＊(n/n-1)

**SPSS output**

88.216*(250/249)=88.57

Correlations

| | | Internet | Education |
|---|---|---|---|
| Internet | Pearson Correlation | 1 | .642** |
| | Sig. (2-tailed) | | .000 |
| | Sum of Squares and Cross-products | 22054.036 | 2888.000 |
| | Covariance | 88.570 | 11.598 |
| | N | 250 | 250 |
| Education | Pearson Correlation | .642** | 1 |
| | Sig. (2-tailed) | .000 | |
| | Sum of Squares and Cross-products | 2888.000 | 918.000 |
| | Covariance | 11.598 | 3.687 |
| | N | 250 | 250 |

**. Correlation is significant at the 0.01 level (2-tailed).

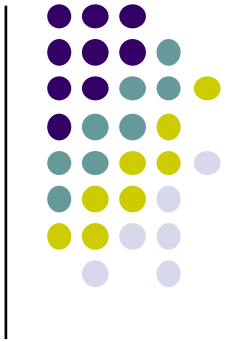
Covariance

| | Internet | Education |
|---|---|---|
| Internet | 88.21614 | |
| Education | 11.552 | 3.672 |

Coefficient of Correlation

| | Internet | Education |
|---|---|---|
| Internet | 1 | |
| Education | 0.641847 | 1 |

# 為什麼老師說， Var. 是 covariance 的 特例？



sample mean of variable X, variable Y

$$\text{Sample covariance} = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- 繳交時間：9:10
- 請繳交紙本（不要寄電子檔）
- 用A4大小，訂好來～
- 請算到小數點後兩位 （考試也是）
- 看到manually，請務必手寫
  - 1-3電腦題，4-8手算

- Office 去哪下載
- https://download.cc.ntu.edu.tw/index.php