台塑企業
FORMOSA PLASTICS GROUP

AI基礎訓練初級班

二、指標衡量方法
（包含正確率分析及機率統計分析）

制定部門：總管理處技訓中心
編定日期：2020年12月30日編印
版次: R1

台塑企業
FORMOSA PLASTICS GROUP

台塑企業
FORMOSA PLASTICS GROUP

AI初級班課程項目：

一、人工智慧概論

二、指標衡量方法

三、資料探勘簡介

四、資料視覺化原理

台塑企業
FORMOSA PLASTICS GROUP

課 程 目 的

完成本課程後，您將能夠:

1. 對於“正確率分析”及“機率統計分析”有基本的概念

2. 對於正確率、標準差、均方根誤差、變異數、變異係數、主成分分析、最小平方法…等有基本的認識。

台塑企業
FORMOSA PLASTICS GROUP

課 程 大 綱

(一)統計量數及分析資料工具（基本知識及應用知識）

(二)正確率分析（基本知識）

(三)單因子變異數分析（應用知識）

(四)迴歸分析(應用知識)

(五)主成分分析（應用知識）

台塑企業
FORMOSA PLASTICS GROUP

目 錄

(一)統計量數的概念

台塑企業
FORMOSA PLASTICS GROUP

目 錄

(二)正確率分析的概念

台塑企業
FORMOSA PLASTICS GROUP

目 錄

(三)單因子變異數分析

台塑企業
FORMOSA PLASTICS GROUP

目 錄

(四)複迴歸分析

台塑企業
FORMOSA PLASTICS GROUP

目 錄

(五)主成分分析（應用知識）

台塑企業
FORMOSA PLASTICS GROUP

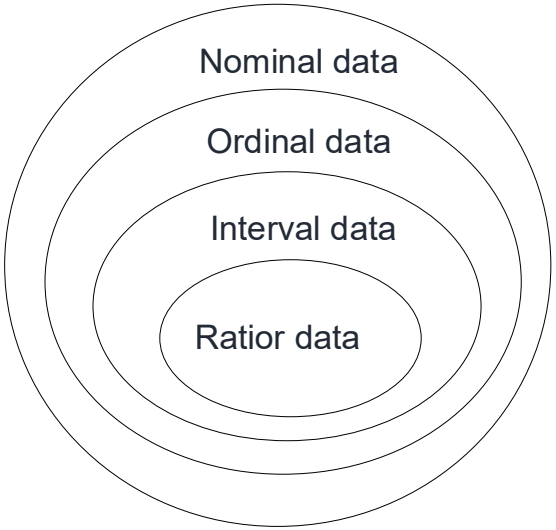
(一)統計量數的概念

該ppt的內容來源來自以下教科書

Neil A. Weiss. Introductory Statistics 10th ed., Pearson, Addison Wesley, 2017.

俞洪亮、蔡義清、莊懿妃，2018，商管研究資料分析：SPSS的應用，修訂三版，台北：華泰文化

1. 變數資料之類型



變數資料之類型細分為:

(1)名目資料(nominal data)，如：婚姻狀況、血型、居住縣市

(2)順序資料(ordinal data)，如：礦物的硬度、名次

(3)區間資料(interval data)，如：溫度

(4)比例資料(ratio data)，如：長度、金錢

1. 變數資料之類型

名目資料、順序資料、區間資料、與比例資料

類別		運算方式	行銷變數範例
名目資料 (Nominal)	非計量	= ≠	性別(男、女)、職業(農、工、商...)、語言(中、英、日...)、居住地區(台北、台中...)...等
順序資料 (Ordinal)	非計量	= ≠ > <	教育程度(國小、國中、高中、大學、研究所)...等
區間資料 (Interval)	計量	= ≠ > < + -	顧客滿意度 (非常滿意、滿意、普通、不滿意、非常不滿意)...等(注：行銷學裡假設滿意度裡不同尺度之間皆為等距)、溫度
比例資料 (Ratio)	計量	= ≠ > < + - X ÷	價格、年齡、所得...等

https://medium.com/marketingdatascience/%E6%B7%BA%E8%AB%87%E8%B3%87%E6%96%99%E9%A1%9E%E5%9E%8B-%E7%A0%94%E7%A9%B6%E8%B3%87%E6%96%99-d9bb456c2fef

2. 集中趨勢和相對量數
(central tendency and percentile value)

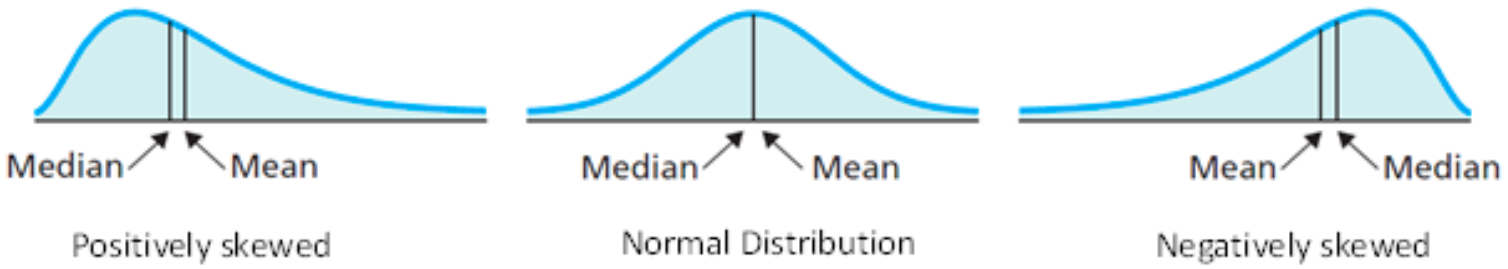
(1)集中趨勢

A.衆數 Mode (一組數據中出現次數最多的數據值)

B.平均值 Mean $\bar{x} = \frac{x_1+x_2+\cdots+x_n}{n}$

C.中位數 Median (所有資料排序後，正中間的資料。

如果資料有偶數個，通常取最中間的兩個資料的)



https://www.differencebetween.com/difference-between-mean-and-vs-median-and-vs-mode/

台塑企業
FORMOSA PLASTICS GROUP

2. 集中趨勢和相對量數
(central tendency and percentile value)

(2)相對量數

A.百分比位數 Percentiles (P_k)

P_k表示至少有k%的資料小於或等於這個數

B.四分位數 Quartiles (Q_n, n=1~3)

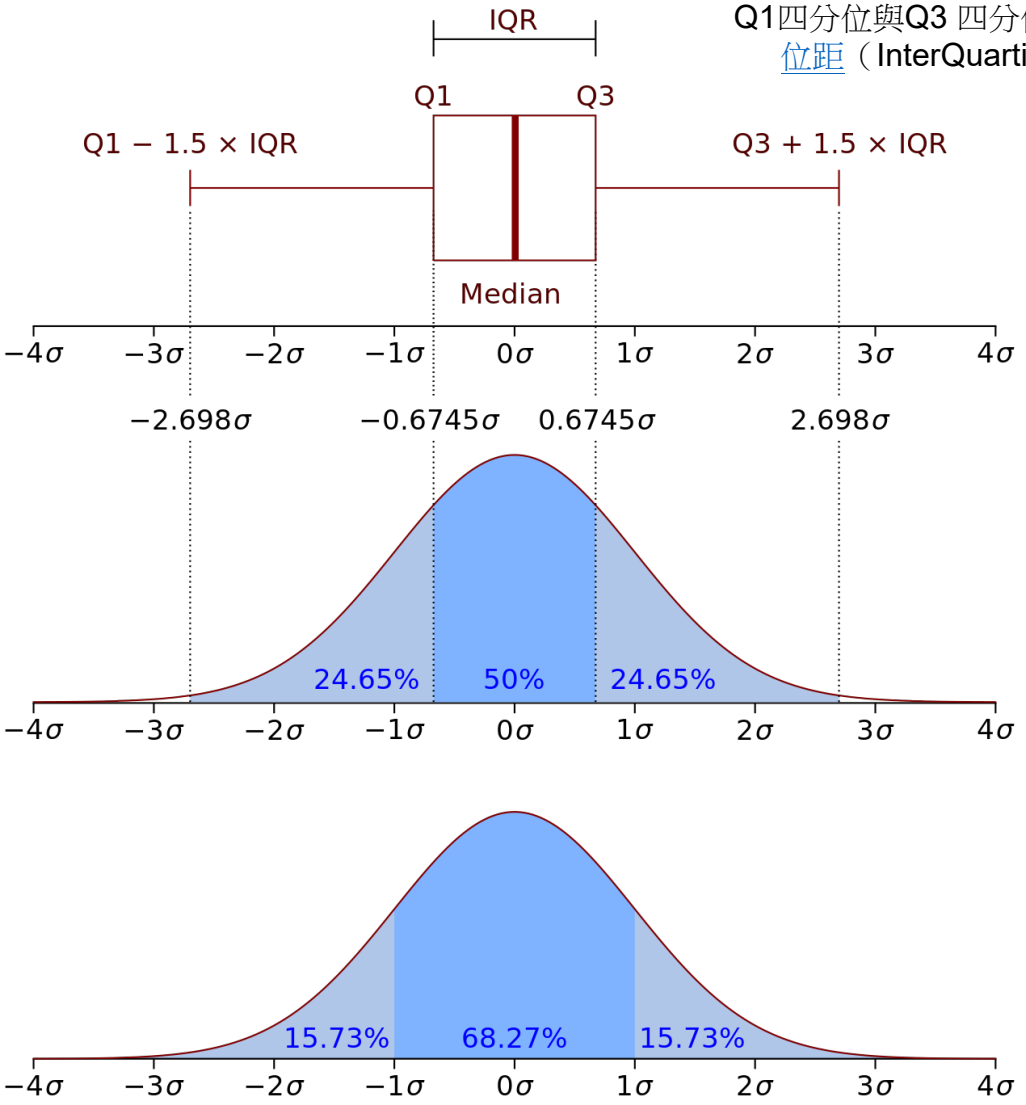
Q_n表示至少有n×25%的資料小於或等於這個數

假設有下列資料: 14.0、15.0、17.0、16.0、15.0

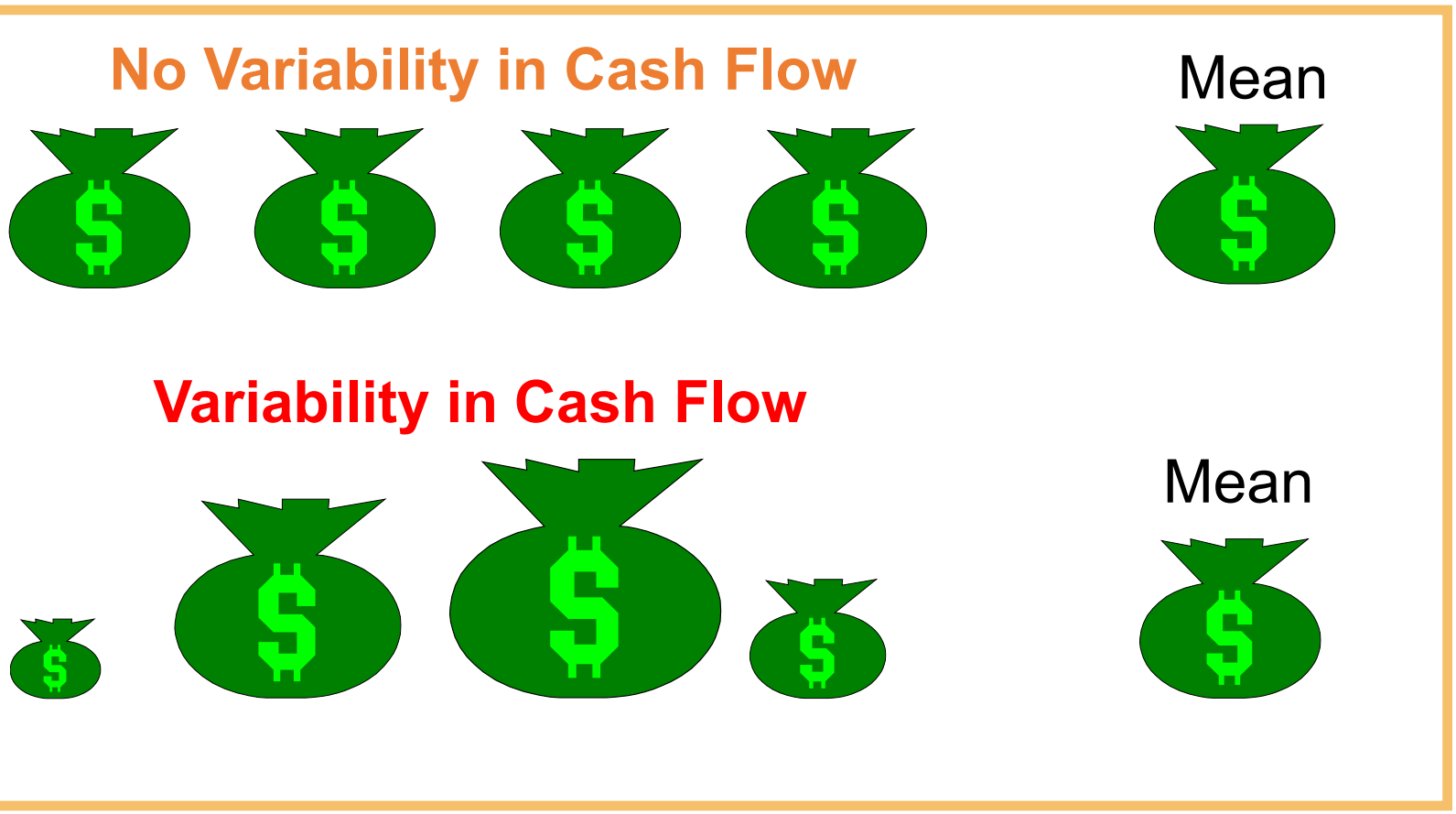
- 眾數: 15.0
- 中位數: 15.0 (也是P_{50}或 Q_2)
- 平均值: 15.4

2. 集中趨勢和相對量數
(central tendency and percentile value)

Q1四分位與Q3 四分位數的差距又稱四分位距（InterQuartile Range, IQR）



3. 差異量數或離散量數 (dispersion)



台塑企業
FORMOSA PLASTICS GROUP

3. 差異量數或離散量數 (dispersion)

(1)差異量數或離散量數(dispersion)

衡量資料中各觀測值之差異或離散程度。

(2)假設有下列資料: 14.0、15.0、17.0、16.0、15.0

A.全距(range) = 17.0 - 14.0 = 3.0 。

B.樣本的變異數與標準差

(sample variance and standard deviation)。

C.變異係數(coefficient of variation, CV)。

D.偏態的測定數(skewness)

4. 樣本的變異數與標準差
(sample variance and standard deviation)

樣本的標準差: 母體標準差是通過隨機抽取一定量的樣本並計算樣本標準差估計的。從一大組數值(n < N)當中取出一樣本數值組合

$X = \{x_1, x_2, x_3, \cdots, x_n\}$，常定義其樣本變異數與標準差：

- 樣本變異數公式:

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

X_1=14.0, X_2=15.0, X_3=17.0, X_4=16.0, X_5=15.0, $\bar{x} = 15.4$

- 樣本標準差公式:

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

維基百科，自由的百科全書

5. 變異係數
(coefficient of variation,CV)

變異係數，是機率分布離散程度的一個正規(Normalized)量度，其定義為標準差與平均值之比。

公式:

$$CV = \frac{\text{標準差}}{\text{平均值}} \times 100\%$$

$$= \frac{s}{\bar{x}} \times 100\% \left(\text{或 } \frac{\sigma}{\mu} \times 100\% \right)$$

- 優點:變異係數是一個比例尺度，因此在比較兩組因尺度不同或均值不同的數據時，應該用變異係數來作為比較的參考。
- 缺陷:當平均值接近於0的時候，微小的擾動也會對變異係數產生巨大影響，因此造成精確度不足。變異係數無法發展出類似於均值的置信區間的工具。

維基百科，自由的百科全書

6. 變異係數範例

假設五位學生之身高及體重如下，試比較其分散程度。

身高：172、168、164、170、176(公分)

體重：62、57、58、64、64(公斤)

身高平均為170公分，標準差為4.47公分

身高的變異係數為4.47/170*100% = 2.63%

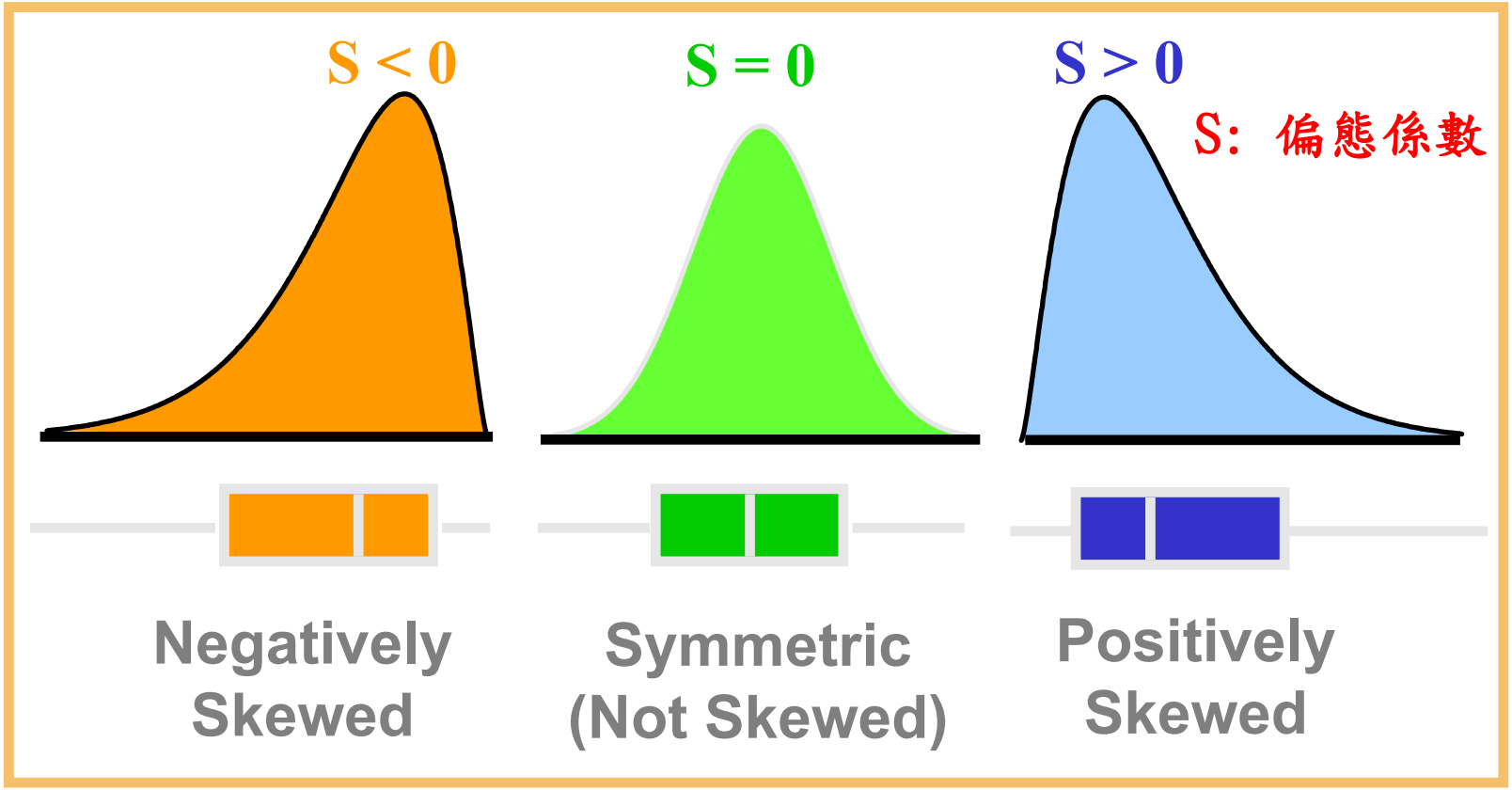
體重平均為61公斤，標準差為3.31公斤。

體重的變異係數為3.31/61*100% = 5.4%

因為體重的變異係數較大，所以體重的分散程度較大。

台塑企業
FORMOSA PLASTICS GROUP

7. 偏態的測定數
Measures of skewness



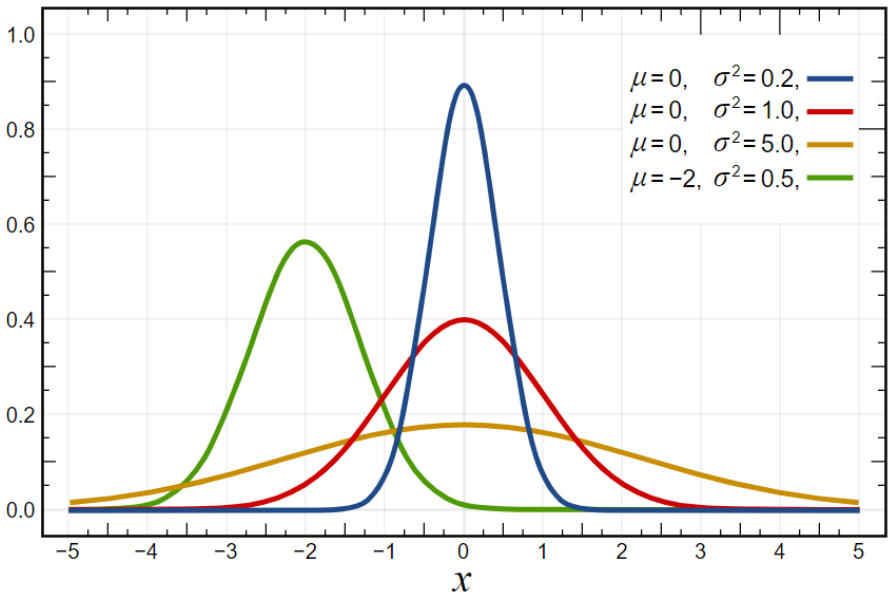
Skewness: Box and Whisker Plots, and Coefficient of Skewness

8. 常態分布
（normal distribution）

又名高斯分布（Gaussian distribution），是一個非常常見的連續機率分布。常態分布在統計學上十分重要，經常用在自然和社會科學來代表一個不明的隨機變量。

若隨機變量 x 服從一個位置參數為 μ 、尺度參數為σ 的常態分布，記為：

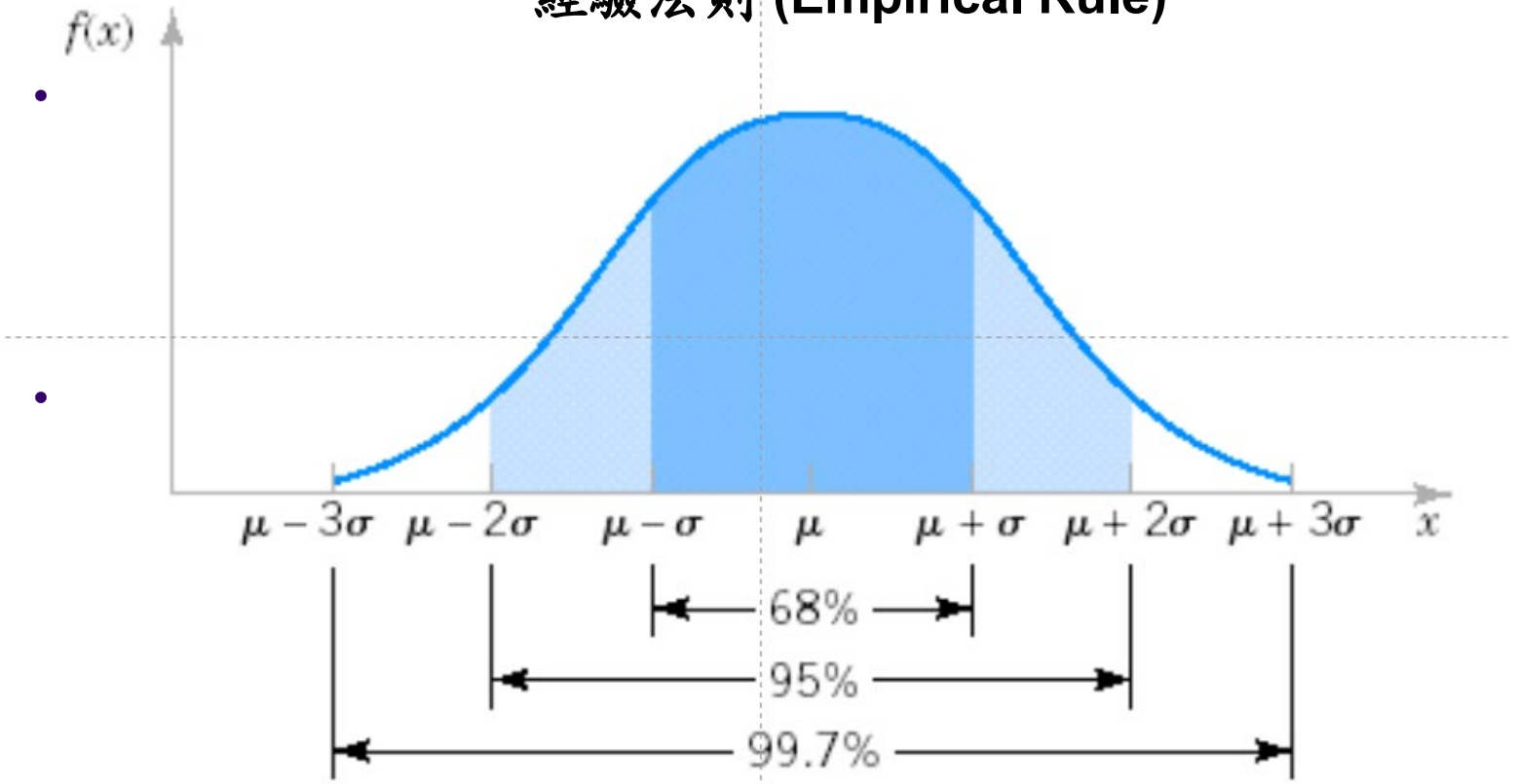
$$x \sim N\left(\mu, \sigma^2\right)$$



維基百科，自由的百科全書

9. 常態分佈線下的區域

經驗法則 (Empirical Rule)



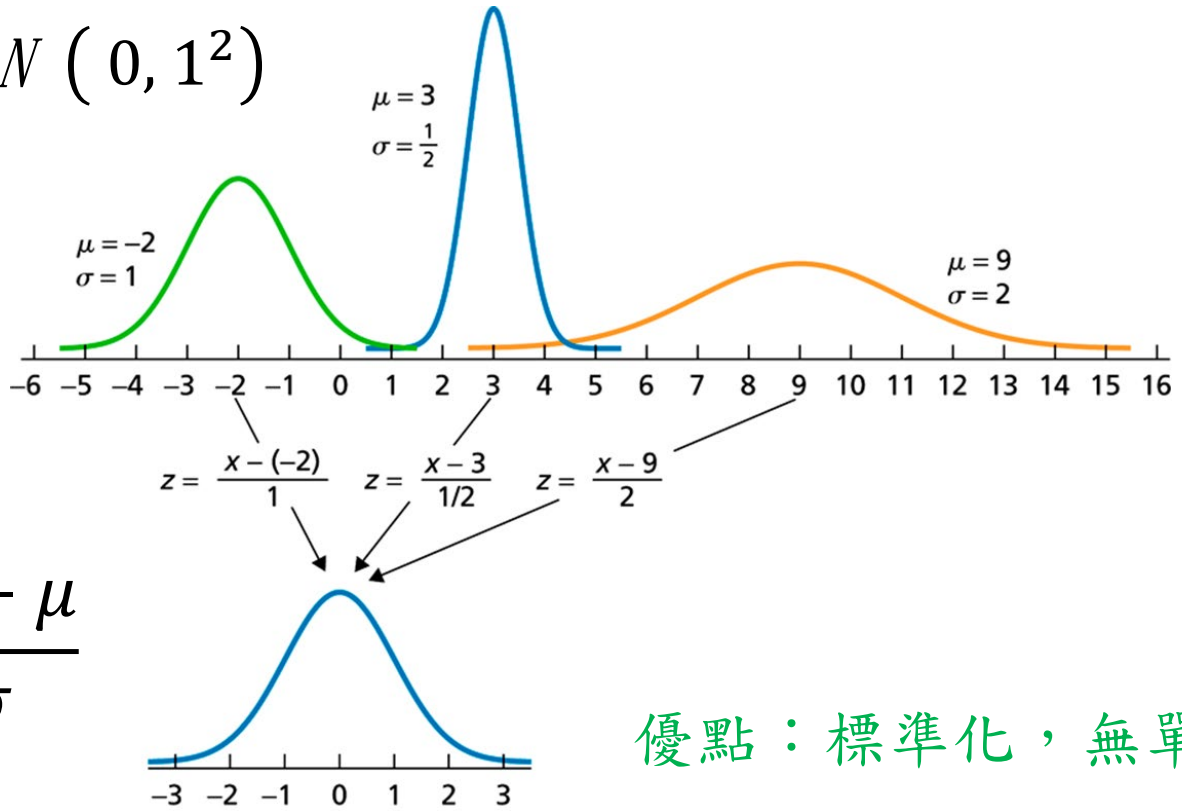
若資料為常態分佈，則大約有68%的資料會落在 ±s 的範圍內，大約有95%的資料會落在 ±2s 的範圍內則，大約有99%的資料會落在 ±3s 的範圍內。(s: **標準差**)

10. 標準常態分配:Z分布
(Standard Normal Distribution)

將任一常態分布經過Z轉換後所得之分佈(仍為常態分布)。

即取平均數為 0 與標準差為 1 之常態分佈。

$$z \sim N\left(0, 1^2\right)$$



$$z = \frac{x - \mu}{\sigma}$$

優點：標準化，無單位

台塑企業
FORMOSA PLASTICS GROUP

(二)正確率分析的概念

正確率分析在很大程度上取決於所使用的方法，但是採用哪種方法取決於自變數和依變數所需的資料衡量尺度。在以下各章節中，我們將介紹各種不同的數據類型的方法所會用的各種評價指標。

分析資料工具	自變數	依變數
分類演算法	類別/數值資料	類別資料
變異數分析	類別資料	數值資料
迴歸分析	數值資料	數值資料

y=f(x)
x:自變數
y:依變數

1. 正確率(Accuracy)高一定好嗎？

(1)正確率確實是一個很好很直觀的評價指標，但是有時候正確率高並不能代表一個演算法就必然是好的。

- 例如地震的預測，類別只有兩個(發生或不發生)。假設一個地震預測系統，每一次都預測不會發生地震，那麼它可能達到99%的正確率，但真的地震來臨時，這個系統去無法示警，導致帶來的損失是巨大的。

(2)為什麼99%的正確率的分類器卻不是我們想要的，因為這裡資料分佈不均衡，某一種類別的資料太少，即使忽略這種分類依然可以達到很高的正確率，卻忽視了我們關注的東西。

台塑企業
FORMOSA PLASTICS GROUP

2. 分類演算法的評價指標

(1)分類演算法有很多，這裏用簡單的二元分類模型展示。

(2)二元分類模型就是輸出結果只有兩種類別的模型，例如☹陽性／陰性）（有病／沒病）（垃圾郵件／非垃圾郵件）。

(3)就二元分類模型而言，評估分類結果的指標很多，這些指標皆源自混淆矩陣（Confusion Matrix），在二元分類模型中這是 2 x 2 的表格。

(4)當訊號偵測（或變數測量）的結果是一個連續值時，類與類的邊界必須用一個臨界值（threshold）來界定。

3. 分類演算法的評價指標(實例)

舉例來說，用血壓值來檢測一個人是否有高血壓，測出的血壓值是連續的實數（從0~200都有可能），以收縮壓140／舒張壓90為閾值，閾值以上便診斷為有高血壓，閾值未滿者診斷為無高血壓。

二元分類模型的個案預測有四種結局：

(1)真陽性(True Positive，TP)：**診斷為有，實際上也有**高血壓。

(2)偽陽性(False Positive，FP)：**診斷為有，實際卻沒有**高血壓。

(3)真陰性(True Negative，TN)：**診斷為沒有，實際上也沒有**高血壓。

(4)偽陰性(False Negative，FN)：**診斷為沒有，實際卻有**高血壓。

3. 分類演算法的評價指標(實例)

		預測類別		總計
		YES	NO	
實際類別	YES	TP (真陽性)	FN (偽陰性)	P(實際為YES)
	NO	FP (偽陽性)	TN (真陰性)	N(實際為NO)
總計		P' (被分為為YES)	N' (被分為為NO)	總合S (= P+N orP'+N')

(1)正確率(accuracy)

accuracy =(TP + TN)/ S，正確率是最常見的評價指標。很容易理解，這個就是被分對的樣本數除以所有的樣本數。

(2)錯誤率(error rate)

error rate = (FP + FN)/ S 描述被分錯的樣本數除以所有的樣本數。錯誤率則與正確率互斥，所以**accuracy = 1 - error rate**。

正確率=(100+80)/(100+10+20+80)=0.8571

錯誤率=(20+10)/(100+10+20+80)=0.1429=1-0.8571

ex.

TP=100	FN=10
FP=20	TN=80

3. 分類演算法的評價指標(實例)

		預測類別		總計
		YES	NO	
實際類別	YES	TP (真陽性)	FN (偽陰性)	P(實際為YES)
	NO	FP (偽陽性)	TN (真陰性)	N(實際為NO)
總計		P' (被分為為YES)	N' (被分為為NO)	總合S (= P+N orP'+N')

(3)靈敏度(Sensitivity) or (真陽性率, true positive rate)

sensitivity = (TP / P)，表示的是樣本中實際符合某特定條件的，被正確診斷為符合那個特定條件結果的比率。

(4)專一性 or 特異度 (Specificity) or (真陰性率, true negative rate)

specificity = (TN / N)，表示的是樣本中實際不符合某特定條件的，被正確診斷為不符合那個特定條件的比率。

靈敏度(sensitivity)=100/(100+10)=0.9091
專一性(specificity)=80/(20+80)=0.8000

ex.

TP=100	FN=10
FP=20	TN=80

3. 分類演算法的評價指標(實例)

		預測類別		總計
		YES	NO	
實際類別	YES	TP（真陽性）	FN（偽陰性）	P(實際為YES)
	NO	FP（偽陽性）	TN（真陰性）	N(實際為NO)
總計		P' (被分為為YES)	N' (被分為為NO)	總合S (= P+N orP'+N')

(5)假(偽)陽性率 (False Positive Rate)（第一類錯誤）

false positive rate = (FP / N)，表示的是樣本中實際不符合某特定條件，但根據診斷被識別為符合那個特定條件的比率。所以也稱**誤診率**= **1–specificity** (TN / N)

假陽性率(False Positive Rate) =20/(20+80)=0.2000
=1-80/(20+80)
=0.2000

ex.

TP=100	FN=10
FP=20	TN=80

3. 分類演算法的評價指標(實例)

		預測類別		總計
		YES	NO	
實際類別	YES	TP（真陽性）	FN（偽陰性）	P(實際為YES)
	NO	FP（偽陽性）	TN（真陰性）	N(實際為NO)
總計		P' (被分為為YES)	N' (被分為為NO)	總合S (= P+N orP'+N')

(6)假(偽)陰性率 (False Negative Rate)（第二類錯誤）

false negative rate = (FN / P)，表示的是樣本中實際符合某特定條件，但根據診斷被識別為不符合那個特定條件的比率。所以也稱**漏診率**= 1 – sensitivity (TP / P)

假陰性率(False Negative Rate) =10/(100+10)=0.0910
=1-100/(100+10)
=0.0910

ex.

TP=100	FN=10
FP=20	TN=80

3. 分類演算法的評價指標(實例)

		預測類別		總計
		YES	NO	
實際類別	YES	TP (真陽性)	FN (偽陰性)	P(實際為YES)
	NO	FP (偽陽性)	TN (真陰性)	N(實際為NO)
總計		P' (被分為為YES)	N' (被分為為NO)	總合S (= P+N orP'+N')

(7)精確率(Precision)

precision=TP/(TP+FP) 是精確性的度量，表示被分為陽性(正例)的樣本中實際為陽性(正例)的比例。

精確率(precision) =100/(100+20)=0.8333

正確率(accuracy) =(100+80)/210=0.8571

ex.

TP=100	FN=10
FP=20	TN=80

4. 分類演算法的評價指標(實例2)

對於分類器而言，不可能同時提高所有的指標

（它們可能具有逆關係，例如，當A變大時B會變小）。

比如我們開頭說的地震預測，沒有誰能準確預測地震的發生，但我們能容忍一定程度的誤報，假設1000次預測中，有5次預測為發生地震，其中一次真的發生了地震，而其他4次為誤報，那麼正確率從原來的99.9%(999/1000)下降到99.6%（996/1000），但靈敏度從0%(0/1)上升為100%（1/1），這樣雖然謊報了幾次地震，但真的地震來臨時，我們沒有錯過，這樣的分類器才是我們想要的:在一定正確率的前提下，我們要求分類器的靈敏度盡可能的高，但它們是會互相影響的，所以在實際中常常需要根據具體情況做出取捨。

5. ROC曲線
（接收者操作特徵曲線
receiver operating characteristic curve）

(1)ROC曲線是一種坐標圖式的分析工具，用於

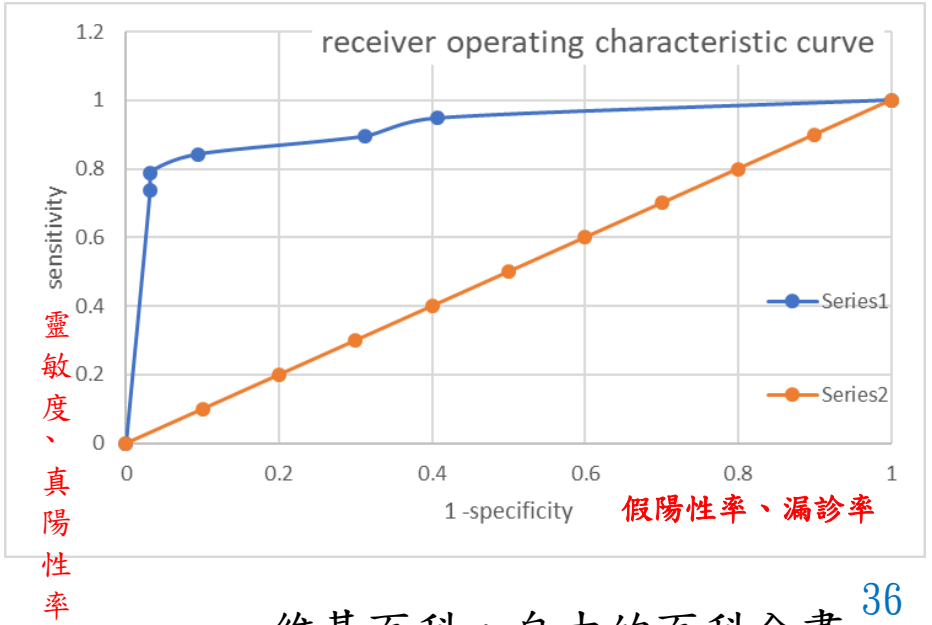
A.選擇最佳的信號偵測模型、捨棄次佳的模型。

B.在同一模型中設定最佳閾值。

(2)在做決策時，ROC分析能不受成本／效益的影響，給出客觀中立的建議。

ROC空間將1-specificity定義為X軸，sensitivity定義為Y軸。

		預測類別	
		YES	NO
實際類別	YES	TP	FN
	NO	FP	TN



6. AUC數值一般的判別規則

ROC曲線下的面積(Area Under Curve; AUC)來判別ROC曲線的鑑別力，AUC數值的範圍從0到1，數值愈大愈好：

(1)AUC = 1，是完美分類器，採用這個預測模型時，存在至少一個閾值能得出完美預測。絕大多數預測的場合，不存在完美分類器。

(2)0.5 < AUC < 1，優於隨機猜測。這個分類器（模型）妥善設定閾值的話，能有預測價值:

A. 0.9 ≦ AUC < 1.0 極佳的鑑別力

B. 0.8 ≦ AUC < 0.9 優良的鑑別力

C. 0.7 ≦ AUC < 0.8 可接受的鑑別力

(3)AUC = 0.5，ROC剛好是對角線，跟隨機猜測一樣（例：丟銅板），模型沒有預測價值。

(4)AUC < 0.5，比隨機猜測還差；但只要總是反預測而行，就優於隨機猜測。

7. 均方根誤差、共變異數

在判斷一個預測模型是否準確時，我們常用均方根誤差(root mean square error, RMSE)來評量預測模型的預測值之偏差程度。一般來說，預測模型的均方根誤差常被定義為“預測值”(依據模型所計算)和“實際值”(觀測值)之差的方均根值，是最常用的評價指標之一。其計算式如下：

$$RMSE = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{Y}_i - Y_i\right)^2}$$

實際值	200	135	195	180	310
預測值	190	145	201	198	290

$$RMSE = \sqrt{\frac{1}{4}((200-190)^2 + (135-145)^2 + (195-201)^2 + (180-198)^2 + (310-290)^2)} = 15.49153$$

共變異數(covariance)在機率論和 統計學中用於衡量兩個變數的母體誤差。而變異數是共變異數的一種特殊情況，即當兩個變數是相同的情況

$$Covar(x,y) = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$

7. 均方根誤差、共變異數

共變異數可以用來表示二個變項間關聯的強度及方向

(1)當二個變項之相關為零時，其共變異數為零；

(2)若相關為正，則共變異數大於零；

(3)若相關為負，則共變異數小於零。

但共變數的大小會因測量單位的不同而有所差異，二個相同的變項，以不同測量單位加以表示時，其共變數就會產生極大的差異，因為共變數此一特性，所以在表示變項間關係時，習慣上常用相關係數(correlation coefficient)加以表示。而將共變數分別除以二個變項之標準差即可求得二變項之相關係數，即

$$r_{x,y} = \frac{Covar(x,y)}{S_x S_y}$$

台塑企業
FORMOSA PLASTICS GROUP

(三)單因子變異數分析
1-ANOVA
(Analysis of Variance)

1. 變異數分析的概念

(1)檢定三組以上的資料之平均值是否相等時，可採用單因子變異數分析(analysis of variance; ANOVA)。

(2)單因子變異數分析可用來來預測變量

$\mathbf{H_0}$ ：$\mu_1 = \mu_2 = \cdots = \mu_k$

$\mathbf{H_1}$ ：並非所有的μ_i皆相等

進行ANOVA分析時，均必須符合以下之假定：

A.各組資料呈常態分配。

B.變異數同質：各組資料的變異數σ^2都相等。

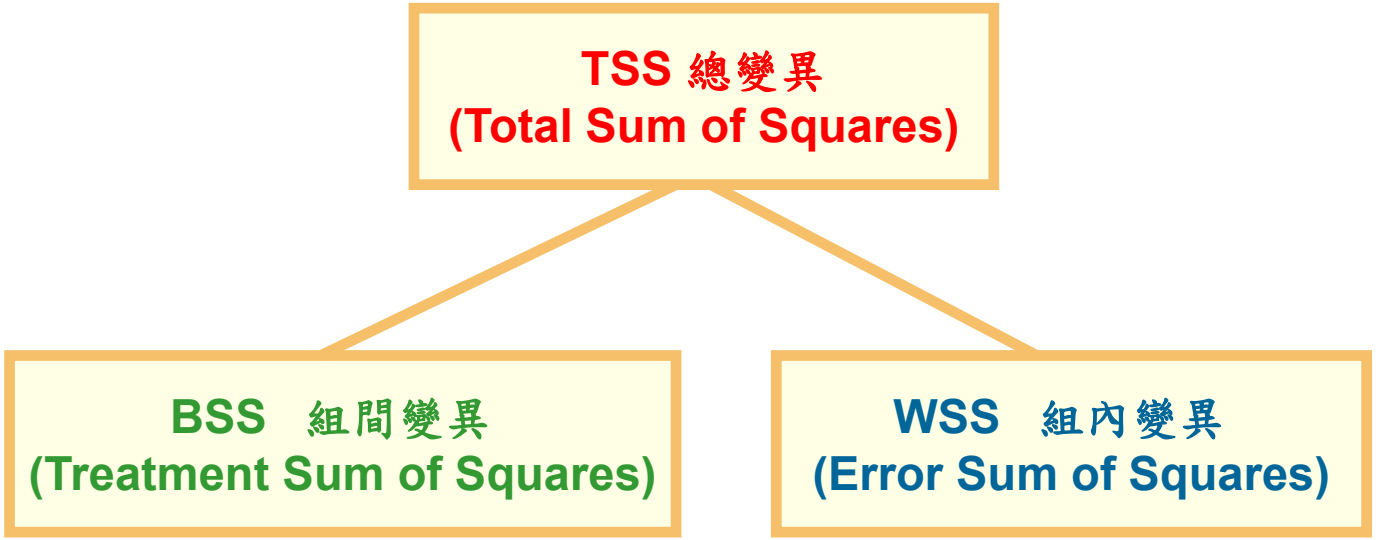
C.所有樣本都是隨機抽樣，而且彼此獨立，可以進行累積與加減。

D.對極端值應有足夠的敏感性。

2. 變異數分析原理說明

總變異可分為兩部分，即組間變異與組內變異來說明。每個觀察值與總平均差異的來源，可分為兩大部分：一為來自分組或方法別所造成的差異（組間變異）；另一為來自觀察值本身的個別差異（組內變異）。

變異數分析的檢定統計量乃用*F*-值來進行。*F*檢定值越大，代表組間變異量越大；隱含各組存在差異。



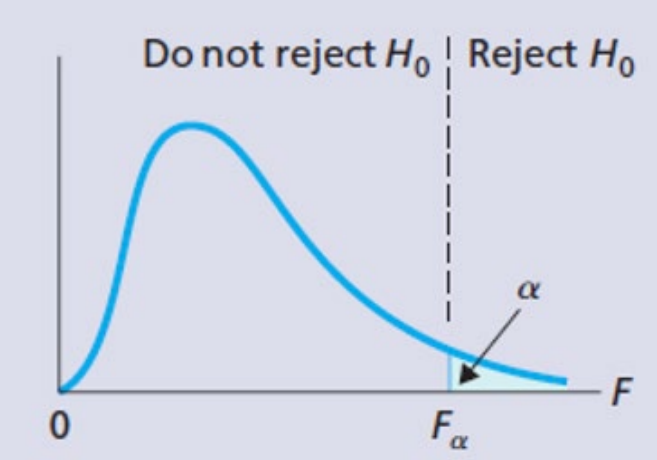
2. 變異數分析原理說明

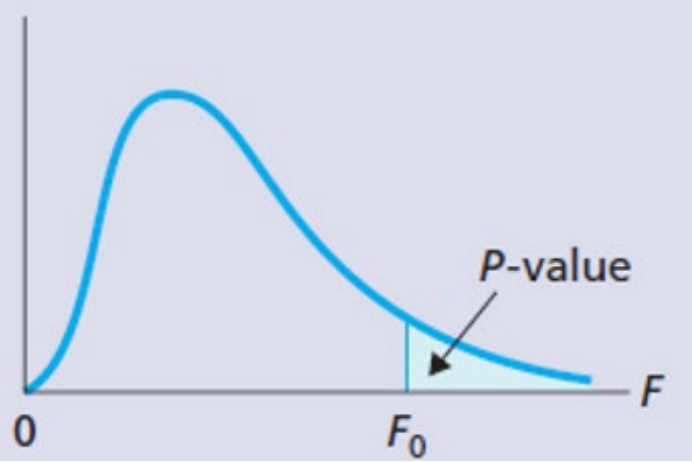
變異數分析摘要表

	離均差平方和(SS)	自由度(DF)	均方和(MS)	F（檢定）	顯著性
組間	BSS（組間變異）	DFB=(#組別-1)=K-1	MSB	F0 =MSB/MSW	p-值
組內	WSS（組內變異）	DFW=(N-1)-(K-1)=N-K	MSW		
全體	TSS（總變異）	DFT=(#樣本數-1)=(N-1)			

如果 $F_0 \geq F_\alpha$（p-值很小），即組間變異顯著，代表所檢定的組別中，最少有一組之平均數是與其他組有顯著差異的，因此拒絕H_0

如果 $F_0 < F_\alpha$（p-值很不夠小），即組間變異不顯著(在α水準下)，無法拒絕H_0





3. 完全隨機設計範例：

範例：若是某製造廠想要驗證其不同生產線或不同生產方法或不同機器/不同個人在生產績效上是否有差異？

本測試統計假設檢驗設定0.05顯著水準(α)，來判斷有無足夠證據接受6條生產線的生產績效不會顯著的有差異？(常見設定α臨界值為0.1、0.05與0.01)

3. 完全隨機設計範例：

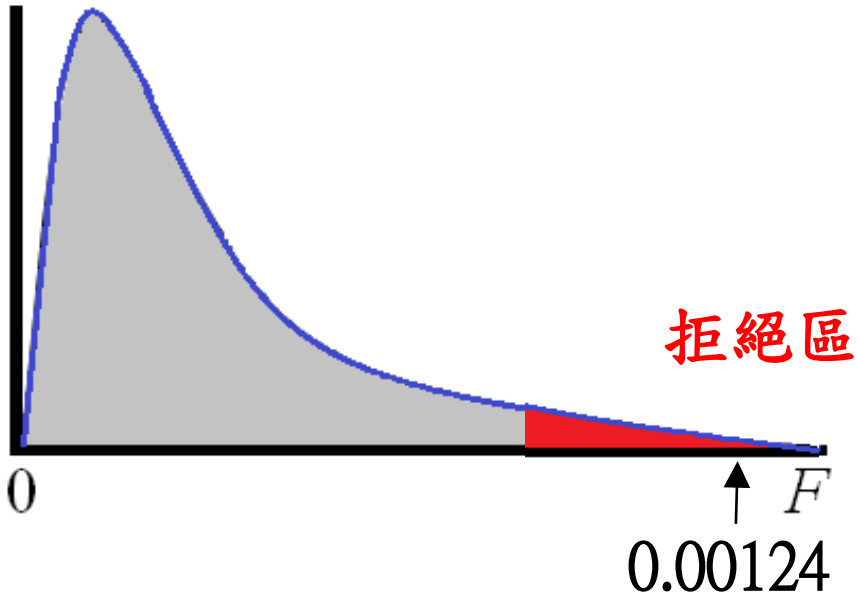
針對此一範例我們所發展的假設如下：

(1)H_0：$\mu_1=\mu_2=\ldots=\mu_6$ (6條生產線的平均產量皆相等)

(2)H_1：至少有一生產線的平均產量與其他生產線的平均產量不同

組	個數	總和	平均	變異數
L1	13	415	31.923	24.577
L2	12	373	31.083	32.083
L3	10	358	35.800	28.178
L4	10	380	38.000	43.556
L5	12	351	29.250	36.568
L6	11	314	28.545	28.073

4. 單因子變異數分結果



拒絕區域：0.05

（顯著水準，α）

ANOVA					
變源	SS	自由度	MS	F	P-值
組間	733.27	5	146.65	4.60055	0.00124
組內	1976.42	62	31.88		
總和	2709.69	67			

5. 範例異數分析結論

求出之顯著性(p-value)= 0.00124 < 0.05，代表在5%的顯著水準下，拒絕H_0的假設。換言之，我們有足夠證據足以顯示至少有一條生產線的平均產量與其他生產線的平均產量有所不同。

(1)判斷方法:顯著性 (p-value)= 0.00124 < 0.05，

故拒絕 **H_0**

(2)結論：生產線的平均產量有顯著差異(但我們不知道哪一條與哪一條不同)。

台塑企業
FORMOSA PLASTICS GROUP

(四)複迴歸分析
(Multiple Regression Analysis)

該ppt的內容來源來自以下教科書

Neil A. Weiss. Introductory Statistics 10th ed., Pearson, Addison Wesley, 2017.

俞洪亮、蔡義清、莊懿妃，2018，商管研究資料分析：SPSS的應用，修訂三版，台北：華泰文化

邱皓政, 量化研究法（二）統計原理與分析技術, 2007, 雙葉書廊

台塑企業
FORMOSA PLASTICS GROUP

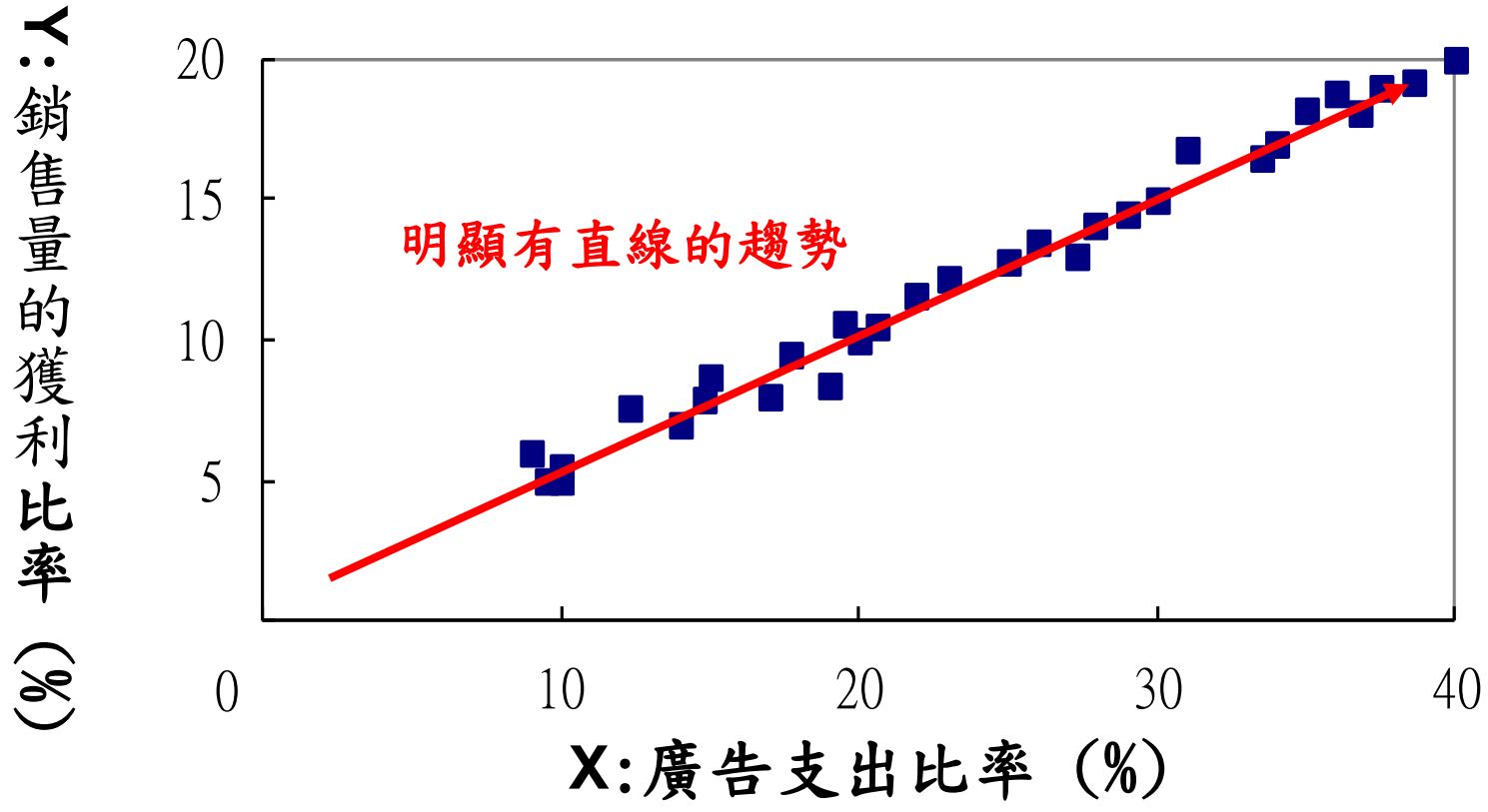
1. 迴歸原理

迴歸分析是運用一個或多個變項來預測另一個變項的統計技術總稱，被預測的變項稱為依變項，預測變項也可稱為自變項。只根據一個自變項來預測依變項的迴歸分析稱為「簡單迴歸」（simple regression），若自變項為兩個或兩個以上則稱為「多元迴歸」（multiple regression）。

迴歸分析的原理是找出最適切的直線數學方程式來表示自變項和依變項之間的關係，此式稱為迴歸方程式，若假定自變項和依變項間的函數關係為線性，稱為線性迴歸（linear regression），否則稱為非線性迴歸（nonlinear regression）。

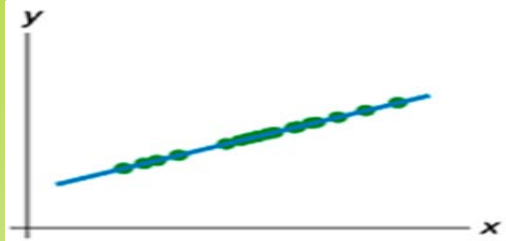
2. 迴歸分析的基本概念

在許多研究問題中，變數與變數之間有時會呈現**線性相關**。
例如廣告支出金額與銷售量之間的關係，如圖：

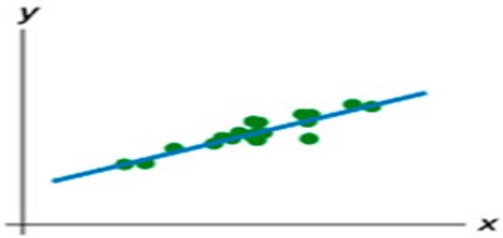


廣告支出金額與銷售量(XY)的散佈圖

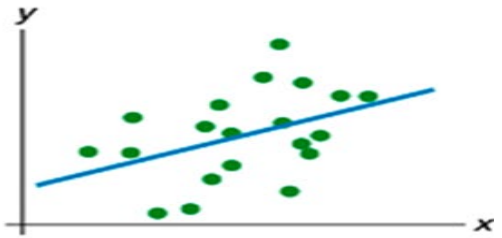
3. 不同的相關情形圖示



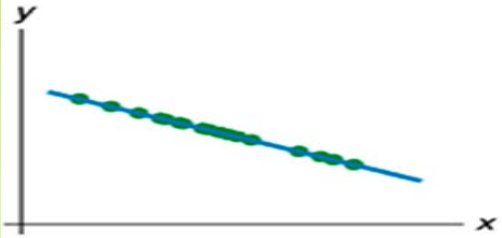
完全正相關　r＝1



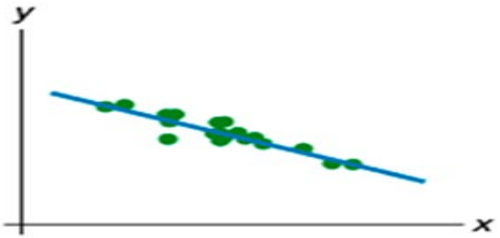
強的正相關 r＝0.9



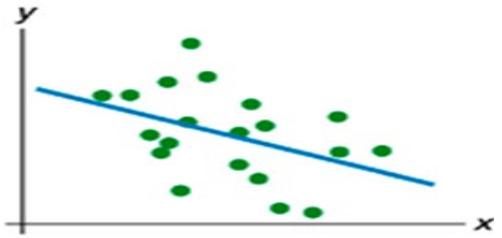
弱的正相關r＝0.4



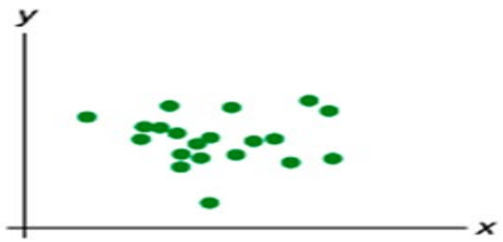
完全負相關 r＝- 1



強的負相關 r＝- 0.9



弱的負相關r＝- 0.4



零相關 r＝0

線性關係可以用XY散佈圖的方式來表現

4. 相關分析的基本概念

當X增加時，Y也會跟著增加，即是代表X與Y之間有很高的相關(正相關)，通常我們用皮爾森(Pearson)相關係數來表示兩個變數間之相關係數，計算公式如下：

$$r = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{\sqrt{\Sigma(X - \overline{X})^2 (Y - \overline{Y})^2}}$$

相關係數為一標準化之數字，其值不受變項特性的影響，其數值是介於－1至＋1之間。

- $r \rightarrow +1$: 表示這個關係趨向正相關的關係
- $r \rightarrow -1$: 表示這個關係趨向負相關的關係

相關係數的絕對值越接近1，關係越強

5. 相關係數的強度大小與意義

相關係數範圍(絕對值)	變項關聯程度
1.00	完全相關
.70 至.99	高度相關
.40 至.69	中度相關
.10 至.39	低度相關
.10 以下	微弱或無相關

註：如果X和Y具有非線性關係，則相關係數可能表示X和Y是沒有關係

6. 迴歸模式之判定係數

(1)根據現有的資料建立一個迴歸模式時，必須檢定此模式與資料的符合程度。檢定適合度最常用的量數是 R^2 (R-square)，或稱判定係數(coefficient of determination)。

(2)判定係數（R^2）是依變數（Y）的變量之總變化的比例（%），可由自變數（X）的之變化解釋。它 的範圍從0%到100%。

(3)在學術研究上最直接的觀念是R^2愈接近1.0(100%)愈好。

台塑企業
FORMOSA PLASTICS GROUP

7. 迴歸模型範例

假如在某製造過程中，有一些因素可能會影響製程結果，如某原料的使用量、程序參數以及加工時間的長短。對於製造商而言，找出哪些因素會影響結果以及它們如何影響生產可能至關重要。

下表列了產量(Y), 原料A的使用量(X_1), 程序參數1(X_2), 程序參數2 (X_3), 程序參數3 (X_4), 和加工時間(X_5)。找出線性相關係數並進行迴歸分析

	產量	原料A的使用量	程序參數1	程序參數2	程序參數3	加工時間
產量	1.0000					
原料A的使用量	0.6775	1.0000				
程序參數1	0.6199	0.3589	1.0000			
程序參數2	-0.2896	-0.5084	0.0608	1.0000		
程序參數3	0.6263	0.5006	0.6521	-0.2986	1.0000	
加工時間	0.6745	0.7342	0.5550	-0.2208	0.5236	1.0000

8. 範例的迴歸分析(刪除之後)

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	210.7795	51.0845	4.1261	0.0001	108.0108	313.5483
原料A的使用量	0.6255	0.1151	5.4347	0.0000	0.3939	0.8570
程序參數1	0.3097	0.0688	4.4999	0.0000	0.1712	0.4481

產量

= 210.78 + 0.625(原料A的使用量) + 0.3097(程序參數1)

+ 殘差

= 預測值 + 殘差

台塑企業
FORMOSA PLASTICS GROUP

（四）主成分分析
(Principal Component Analysis)

該ppt的內容來源來自以下教科書

Multivariate Data Analysis – A global perspective, 7/e, Prentice Hall International, J.F. Hair, W.C. Black, B.J. Babin and R.E. Anderson.

台塑企業
FORMOSA PLASTICS GROUP

1. 主成分分析
(Principal Component Analysis)

主成分分析(principal components analysis, PCA)是一種多變量的統計分析的方法。利用原有的變數組合成新的變數，以達到簡化數據集的目的；用最精簡的主成份特徵來解釋目標變數的最大變異，避免共線性與過度配適等問題，但卻能夠保留住數據本身所提供的重要資訊。由於主成分分析主要依賴數據提供的訊息，所以數據的準確性對分析結果影響很大。

主成分分析屬於非監督式學習法；它是用來處理一組自變數($X_1, X_2, \ldots, X_n$)，沒有依變數Y。

維基百科，自由的百科全書

台塑企業
FORMOSA PLASTICS GROUP

2. 主成分分析的概念

主成分分析(Principal components Analysis)假設:

(1)各共同因素間彼此均無關聯，即相關係數為零，也不考慮變數分數中的獨特因素

(2)目的在使每一個成分應用觀察變項的線性整合，來代表最的觀察變異量

(3)第一個主成分能夠反應最大的變異量，依序發展各主成分

(4)比較其它因素分析(Factor Analysis)，主成分分析可以得到最大的解釋變異量

優點：所得之共同因素彼此無相關

缺點：忽略獨特性，可能損失有用信息，故共同性有高估的危險

3. 主成分分析之主要的功能

(1)發掘多變量資料中各變數間複雜的組合型式。

(2)進行探索性的研究，能找出資料中潛在的共同特徵，供未來實驗之用。

(3)減少多變量資料的維度。

(4)因素分析的兩個主要目的：

—減少維度 (dimension reduction)

—歸納變數 (summarization)

4. 範例:客戶對製造商的看法(HBAT)

HBAT公司的管理階層想要瞭解及掌握客戶的購買記錄，並預期能夠洞悉客戶的消費行為。管理階層建立一個有18個獨立變量的資料庫；此資料庫收集了HBAT公司的100個客戶的記錄。變量可以分為三種主要類型：

分類客戶(Customer Classification)的變量(名目變數):

1. 客戶類型
2. 產業類型
3. 公司規模
4. 區域
5. 分發系統

績效感知(Performance Perceptions)的變量(範圍從優=10分到差=0分):

6. 產品品質
7. 電子商務活動
8. 技術支援
9. 解決顧客投訴
10. 廣告
11. 生產線
12. 銷售_營業形象
13. 競爭價格
14. 產品保固聲明
15. 新產品
16. 訂購與帳務處理
17. 彈性價格
18. 交貨速度

推薦(Recommendation)的變量: 剩餘變量(X19 to X23)

4. 範例:客戶對製造商的看法(HBAT)

HBAT的數據庫變量的描述

	Variable Description	Variable Type
Data Warehouse Classification Variables		
X1	Customer Type	nonmetric
X2	Industry Type	nonmetric
X3	Firm Size	nonmetric
X4	Region	nonmetric
X5	Distribution System	nonmetric
Performance Perceptions Variables		
X6	Product Quality	metric
X7	E-Commerce Activities/Website	metric
X8	Technical Support	metric
X9	Complaint Resolution	metric
X10	Advertising	metric
X11	Product Line	metric
X12	Salesforce Image	metric
X13	Competitive Pricing	metric
X14	Warranty & Claims	metric
X15	New Products	metric
X16	Ordering & Billing	metric
X17	Price Flexibility	metric
X18	Delivery Speed	metric
Outcome/Relationship Measures		
X19	Satisfaction	metric
X20	Likelihood of Recommendation	metric
X21	Likelihood of Future Purchase	metric
X22	Current Purchase/Usage Level	metric
X23	Consider Strategic Alliance/Partnership in Future	nonmetric

5. 因素分析的目標及設計

(1)基於管理階層想對HBAT客戶的作市場區隔研究。但公司的管理人員想知道可以從該數據集中導出哪些方面和因素，以便更好地瞭解公司對該公司的客戶的看法。

(2)出於以上原因，可以使用主成分分析(PCA) 將彼此相關的變數，轉化成為少數獨力並有概念化意義的因素。

(3)X_6至X_{18}這13個量性的變項適合用於此主成分分析

(4)主成分分析將瞭解這些看法是否可以“分組”；能夠實現構面縮減的目標。

(5)減少的數據變項集可使潛在因素更易於觀察，並且可能更容易和準確地解釋感興趣的現象

(6)將13個變項被另一組較少的變項取代，可協助歸納市場區隔後的客戶群體的不同。

台塑企業
FORMOSA PLASTICS GROUP

描述性統計資料

	平均數	標準偏差	分析 N
產品品質	7.810	1.3963	100
電子商務活動	3.672	.7005	100
技術支援	5.365	1.5305	100
解決顧客投訴措施	5.442	1.2084	100
廣告	4.010	1.1269	100
生產線	5.805	1.3153	100
銷售_營業形象	5.123	1.0723	100
競爭價格	6.974	1.5451	100
產品保固聲明	6.043	.8197	100
新產品	5.150	1.4930	100
訂購與帳務處理	4.278	.9288	100
彈性價格	4.610	1.2060	100
交貨速度	3.886	.7344	100

6. 因素的轉軸

顧客服務

行銷

售後服務及保固

產品的性價比

旋轉元件矩陣[a]

	元件			
	1	2	3	4
解決顧客投訴措施	.933	.105	.057	.076
交貨速度	.931	.167	.005	.005
訂購與帳務處理	.886	.098	.087	.069
銷售_營業形象	.138	.898	.076	-.168
電子商務活動	.057	.868	.049	-.141
廣告	.156	.743	-.085	.043
技術支援	.017	-.025	.940	.097
產品保固聲明	.103	.054	.933	.082
產品品質	.029	-.014	-.022	.892
競爭價格	-.104	.228	-.255	-.730

擷取方法：主體元件分析。
轉軸方法：具有 Kaiser 正規化的最大變異法。

a. 在 5 疊代中收斂循環。

台塑企業
FORMOSA PLASTICS GROUP

7. 因素的解釋

(1)第一個因子與X_9（投訴解決方案），X_{18}（交貨速度）和X_{16}（訂購與帳務處理：訂單和開票）具有高度正相關。因此，我們可以將其命名為“客戶服務”。

(2)第二個因子與X_{12}（銷售人員形象），X_7（電子商務活動）和X_{10}（廣告）具有高度正相關。 因此，我們可以將其命名為“行銷”。

(3)第三個因子與X_8（技術支持）和X_{14}（產品保固聲明）具有高度正相關。 因此，我們可以將其命名為“售後服務及保固”。

(4)第四個因子與X_6（產品品質）高度相關，而與X_{13}（彈性價格）高度負相關。 因此，我們可以將其命名為“產品之C/P值”或“產品的性價比”。