**統計學**

1. 統計步驟
2. 收集樣本資料
3. 整理資料
4. 展示資料(chart)
5. 分析資料(估計、檢定、預測、指標)
6. 解釋結果
7. 推論群體做決策(bias、reliability)
8. 統計名詞
9. 母體population

-平均數(mean)µ

-變異數(variation)σ2

-標準差(Standard Deviation) σ

-比率(properbility)p

2.樣本Sample

-參數(parameter):由樣本計算出的指標

-平均數(mean) x bar

-變異數(variation)s2

-標準差(Standard Deviation) s

-比率(properbility)p hat

1. 隨機變異數(RV,Random Variance)
2. Quality(定性)/Category(類別) Data:只能表示類別的資料

e.g.性別、國籍//又稱低階資料

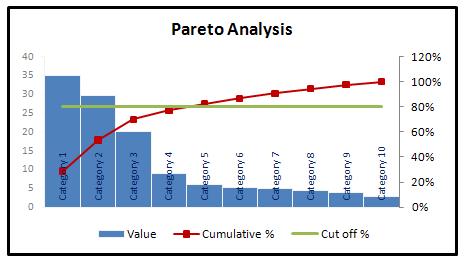
1. Quantity(定量) Data:

-Discrete(離散型):可數的資料、在既定範圍內是有限的

e.g.瑕疵個數、一棵樹樹葉的數量//又稱中階資料

-Continuous(連續型):量測出來的資料

e.g.身高、體重、溫度//又稱高階資料

四.Chart

1.Quality Data

關鍵少數問題

80%處畫一條線

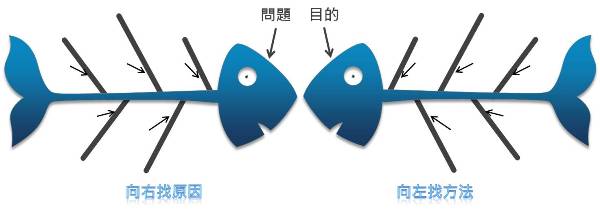
-pie chart

-bar chart

-pareto chart

-魚骨圖

各類別錯誤次數長條圖與累積次數



1. Continous Data

-Stem & Leafe🡪資料分布

-Hist🡪資料分布(單、雙峰?符合常態分配??)

-Scatter(有無離群值、兩項目呈正or負相關、線性關係)

五.蒐集資料方法

1.普查:調查母體資料

2.抽樣

-Simple Random

-System Random:隨意編號並隨機抽一個A，下一個為A+(總個數/要抽出數)

-Stratified Random:分成多個同性質群體，每群體抽一樣數量或按群體大小比例抽

-Cluster Random:分成多個群體，每個群體組成方式類似母體，抽一個群體出來

六.趨勢指標

1.Center Tendency(集中趨勢)

-mean

-medium

-mode

2.Dispersion Tendency(分散趨勢)

-Range(全距)

-Standard Deviation(標準差)

-Variation(變異數)

Population:(平方和-和平方/N)/N

Sample:(平方和-和平方/N)/(N-1)

母體絕對離差(AD):Σ|(xi-µ)|/N

-Coefficient of Variation(CV,變異係數)

Population: σ2/µ

Sample:s2/x bar

七.Skewness(偏態)🡪分布情形

左偏 中間 右偏

0

八.Kurtosis(峰值)

低擴峰 常態峰 高峽峰

0

九.非中趨勢指標

1.Quartile(四分數)

-Q1:第0.25\*(n+1)個值，在兩個數之間則按比例做加權

2.Percentiles(百分位數)

十.運用指標方法

1.經驗法則

在常態分配的情況下，一個標準差內大約包含68%的資料，兩個95%，三個99.75%

--可將三個標準差以外的資料當作離群值去除

2.柴比雪夫定理

在任何分配型態下，至少有1-(1/K2)的資料在K個標準差中

十一.判斷兩組資料的關係

1. 共變異數(COV) (Σ(xi-µx)(yi-µy)- (Σx)(Σy)/N)/N
2. 相關係數(r) COV/σx\*σy —去除單位

r=0🡪無線性關係(可能有曲線關係)

0<r<0.5無明顯正相關

0.5<r<0.7有正相關

0.7<r<0.9顯著正相關

0.9<r 極度顯著正相關

**機率論**

1. 名詞定義
2. Sample Space(樣本空間)-所有可能發生的事件集合
3. Event(事件)-做一次實驗的結果

-Simple Event(簡單事件):S只有一個元素e.g.骰出1的事件{1}

-Composed Event(複合事件):S含有一個以上的元素e.g.骰出奇數的事件{1,3,5}

二.機率三原理

1.0<=P(A)<=1

P(A):發生A事件的機率🡪#(A)/#(S)

2.P(φ)=0 / P(S)=1

Φ:該Event在S中找不到

3.若A1~AK為互斥事件，(A1 U A2 U …AK)等於A1+A2+A3…AK

三.機率計算方法

1.階:!

2.Permutation(排列):P(a,b)=a!/(a-b)!

3.Combination(組合):C(a,b)=a!/b!(a-b)!

四.事件關係

1.Dependency(相依)

2.Independency(獨立)🡪A、B一方發生不影響另一方發生機率

3.Mutually exclusive(互斥)🡪A有B就沒有

4.Union(聯集) 🡪P(A∪B)

5.Intersection(交集)🡪P(A∩B)

6.Complementary Event(互補)🡪P(A)的互補=1-P(A)

五.條件機率

1.若A、B為相依事件P(A|B)= P(A∩B)/P(B)🡪在B發生的條件下發生A的機率

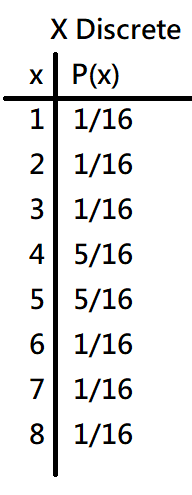
2.若A、B為獨立事件，P(A|B)=P(A)，因為P(A)與P(B)沒有關係

3.若A、B為獨立事件，P(A∩B)=P(A)\*P(B)

**離散型資料**

一.繪製機率分布表

1.X為隨機變數定義(Discrete or Continuous)

 2.x為可能發生的事件

3.P(xi)表x發生的機率

4. 0<P(xi)<1 AND ΣP(xi)=1

5.建立機率分配表

6.繪製機率分配圖

7. µx=x\*P(x)

8.σ2=x2P(x)-µx2

9.r=0表示兩者無線性關係(可能有曲線關係)

1. Poisson(普瓦松)分配🡪針對稀有事件(短時間內發生不到一次)

λ:單位時間發生次數

t:單位時間

µ:平均發生次數

e:自然常數

e.g.每分鐘平均有0.75個人來開戶，下六分鐘有2人來開戶的機率

λ: 0.75人/分

t:6分鐘

µ: 4.5人

e:自然常數

🡪P(x)=µx\*e-µ/x!

1. Jointly Distribution RV聯合分配隨機變數

1.cov= ΣxyP(xy)-µxµy

2.µ=xP(x)

e.g.投擲三次硬幣，第一次為正面與投三次的正面數做機率比較

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 投擲三次的正面次數 | | | | | |
|  | 0 | 1 | 2 | 3 |  |
| 第一丟為正面的次數 | 0 | 1/8 | 2/8 | 1/8 | 0 | 4/8 |
| 1 | 0 | 1/8 | 3/8 | 1/8 | 4/8 |
|  |  | 1/8 | 3/8 | 4/8 | 1/8 | 1 |

**連續型隨機變數**

一.基本性質

1.機率分布函數(PDF:probability distribution function)

2.機率為pdf的積分🡪曲線下的面積

3.累積分配函數(cdf)

4.E(2x+1)=2E(x)+1

5.VAR(2x+1)=4VAR(x)

二.機率分配

1.Uniform Probability Distribution齊一分布

-X平均分配於一定範圍

e.g.介於0-10，P(x<3)=0.3

-E(x)=(α+β)/2

-VAR(x)=1/12\*(β-α)2

2.Normal Probability Distribution常態分佈

-z值:常態分配下(µ=0 σ=1) µ位於-∞~z或依表訂範圍的機率🡪查表

-標準化: (x-µ)/σ= z-score x:樣本值 µ:母體平均值

e.g.從x-bar=40,S=15,分數=50換成x-bar=60,S=10🡪(50-40)/15\*10+60

-繪製直方圖，N筆資料的組數為log2N較合適

-二項分配的N很大，P不接近1或0🡪Normal

連續性正規化:下限-0.5,上限+0.5

µ=np && σ=sqrt(npq)

-二項分配的N很大，P接近1或0🡪Poisson

3.Expotential distribution指數分布🡪事件與事件之間時間

-當f(x)=1/β\*e-(1/β)x=0 && x,β>0🡪Expotential distribution

-E(x)=µ=β

-VAR(x)=β2

-λ:單位時間發生次數=1/β

-事件間時間<T個期間的機率P(x)=1-e-λ\*T

4.連續型jointly distribution

-