1.5 抽樣與統計推論

三年____班　座號：____　姓名：

重點 1：抽樣方法

1.前言：統計是一種經由蒐集具代表性的資料、分析整理數據，以便用於估計和推論母體性質的科學方法。

2.意義：蒐集資料的方法依照所調查的對象是全體或部分，分為**普查**(census)及**抽樣調查**(sampling)兩種

3.普查：是指將我們想研究的全體對象(稱為母群體或母體 population)做全面性的調查

(1)特性：想要了解每一份子的資訊，最精確的方法就是普查

(2)缺點：費時、費成本、費人力等，也無法及時獲取訊息

4.抽樣調查：從母群體中抽取具代表性的一部分，並據以對整個母群體的某一特性作出估計和推論的一種調查方法。

其中具代表性的一部分稱為樣本，抽出樣本的過程就稱為抽樣

(1)特性：只抽取母群體的一部分，以估計母體中所有個體的特徵，不受母群體的環境或大小等因素影響

(2)缺點：受樣本取得方式、方法等影響甚大

註：方便性樣本：直接選擇容易取得的樣本做觀測，這種樣本可稱為「方便性樣本」

自發性回應樣本：由收集自動回應者做為樣本，稱為「自發性回應樣本」

例 1.1：試判斷下列蒐集資料的方法最適宜使用**普查**或**抽樣調查**？

(1)身高、體重的測量　(2)工廠所產燈泡的耐用時數　(3)水果的農藥殘留量

(4)縣市首長的滿意度　(5)各選舉候選人的支持度　(6)全國工商的調查

普查：

抽樣調查：

例 1.2：試指出下列各調查的母群體、樣本為何？

(1)調查本校學生每天上網的時數？

(2)消基會想調查湯圓所含防腐劑是否過量？

重點 2：亂數表 (或稱隨機號碼表)

亂數表：是一連串隨機出現的阿拉伯數字，經專家(或隨機)編製，為了讓這個表容易讀，數字每 5 個排在一起。

由於亂數表上的每個位置出現哪一個號碼的機率都相等，因此利用亂數表來抽取樣本，同樣保證每個個體被選取的機會均等

註：專家編製指利用人工取值完成的亂數表，特徵是耗工、費時，但不易重複

隨機編製指利用機器(如電腦等)取值完成的亂數表，特徵是迅速省時，機動性強，但容易重複

1	29280	39655	18902	92531	90374	07109	26627	59587	84340	98351
2	20123	82082	55477	22059	43168	12903	13436	25523	21090	73449
3	66405	35287	33248	67657	07702	01474	66068	01125	59258	30138
4	97299	83419	13069	17826	76984	48906	10567	17829	00723	46700
5	83923	92076	98880	33942	46841	58731	36513	16681	88722	61984
6	11258	92175	94894	97606	11134	51941	43733	00514	06694	27706
7	08522	48468	60789	47178	85587	78410	67050	41286	16545	22061
8	02114	89744	10115	39603	61089	79392	38945	77699	59054	07742

重點 3：抽樣方法

1.意義：為了確保從樣本所獲得的資訊足以代表母體，以得到可靠的統計結論，抽樣過程應該保證樣本選取的隨機性，設計抽樣的方法時，也必須注意能夠讓個體分布均勻並使每個個體被選取的機會相等

2.常用抽樣方法：簡單隨機抽樣、系統抽樣、分層抽樣、部落抽樣等四種

3.簡單隨機抽樣方法：

從元素個數為 N 的母體中選取 n 個作為樣本，則稱這種抽樣方法為簡單隨機抽樣。

(1)特性：是各種機率抽樣的基礎；沒有摻雜任何人為因素，且每一元素被抽取的機會均等；所抽取的樣本的確代表母群體，即樣本可視為是母群體的縮影

(2)使用時機：對母群體沒有任何資訊，亦或有資訊但是模糊不清或無法使用

(3)簡單隨機抽樣常使用抽籤、亂數表兩種方法取樣本(取碼)

A 抽籤：設取 n 個樣本

步驟 1：先對母體所有個體編號。假設母體元素個數為 N，則從 1 編號到 N (可以用學號、座號等)

步驟 2：抽取 n 個

註：對母體元素個數不多時，適宜採用這個方法

B 亂數表：設隨機取 n 個樣本

步驟 1：先對母體所有個體編號。假設母體元素個數為 N(二位數)，則從 01 編號到 N (編號要使用同樣的位數)

步驟 2：訂亂數表之起始行、列與讀碼方向(向上、向下、向左、向右、向對角線等)

步驟 3：取樣本原則為在編碼範圍內的數取出，超過範圍碼不取、刪除重複出現碼、取足樣本數 n 個

註：對母體元素個數大時，適宜採用這個方法。實際上大量的抽樣是透過電腦產生的隨機號碼來選樣本

例 3.1：某校高三共有忠孝仁愛四班，各班的學生人數如下表：

班級	忠	孝	仁	愛
人數	30	40	50	60

欲從中抽選 8 名學生接受數學能力檢定，在下列各種抽樣方法中，求忠班的阿雄被抽到的機率？

(1)以班級為單位，每班抽出 2 名

(2)先隨機抽出一個班級，再從該班學生中抽出 8 名

(3)先將 180 名學生加以編號，再隨機抽出 8 名

例 3.2：以簡單隨機抽樣法在全班 50 名同學中進行抽樣：

(1)母體是什麼？　　(2)利用亂數表抽出 5 個樣本

重點 4：模擬隨機試驗

4.利用來模擬隨機試驗：

(1)模擬擲一枚均勻的硬幣：

預估均勻硬幣出現正反面的機率都是 0.5，所以可設定若取得奇數代表出現正面，取得偶數代表出現反面

(2)模擬擲一枚正面出現機率為 0.25 的硬幣：

正面出現機率為 0.25，所以可設定若取得 00 到 24 代表出現正面，取得 25 到 99 代表出現反面

例 4.1：利用所附的亂數表及下列指定查表方式，以 0、1、2、3 代表硬幣正面，4、5、6、7、8、9 代表硬幣反面，模擬擲一枚正面出現機率為 0.4 的不均勻硬幣多次，並由所得數據估計該硬幣出現正面的機率。

(1)從第 1 列第 25 行開始，向下取樣 10 次

(2)從第 5 列第 16 行開始，向右取樣 20 次

1	29280	39655	18902	92531	90374	07109	26627	59587	84340	98351
2	20123	82082	55477	22059	43168	12903	13436	25523	21090	73449
3	66405	35287	33248	67657	07702	01474	66068	01125	59258	30138
4	97299	83419	13069	17826	76984	48906	10567	17829	00723	46700
5	83923	92076	98880	33942	46841	58731	36513	16681	88722	61984
6	11258	92175	94894	97606	11134	51941	43733	00514	06694	27706
7	08522	48468	60789	47178	85587	78410	67050	41286	16545	22061
8	02114	89744	10115	39603	61089	79392	38945	77699	59054	07742
9	24580	05775	54677	04171	97815	35557	92626	29756	35289	97756
10	23937	25079	12306	23125	50842	51015	57436	71349	79397	06095

重點 5：常態分布(normal distribution)

1.意義：由模擬隨機試驗結果顯現，抽樣的結果與真正的機率會有出入，但是如果抽樣的次數增加。如 100 次、200 次甚至 1000 次時，所得出現正面的比率接近實際比率的機率較大(誤差會減少)，此現象稱為大數法則

2.常態分布：

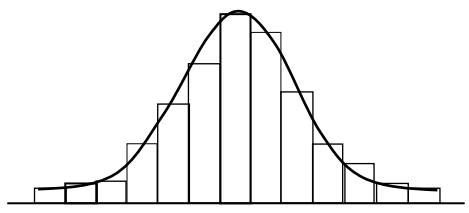
(1)常將蒐集之數據利用長條圖或直方圖繪製成(相對)次數分配統計圖，其各組的組中點所連成的次數分配折線圖，在數據夠多時，稱為次數分布曲線

(2)次數分布曲線，與具單一高峰、且呈現左、右側對稱的平滑曲線，形狀似鐘形曲線的常態分布曲線近似，因此又叫做鐘形曲線，其機率分配稱為常態分配

註：最早將常態分布曲線用在描述分布的是大數學家高斯，所以常態分布又叫做「高斯分布」

註：高斯用函數 $f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ 來描述常態分布曲線。

其中 μ ， σ ，分別為母群體的平均數和標準差，e＝2.71828…



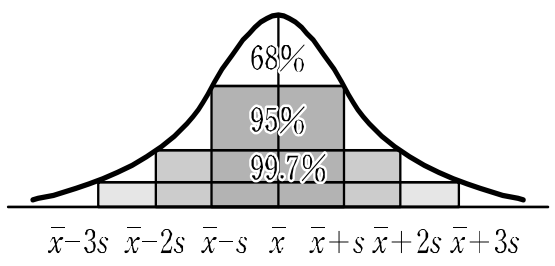
3.常態分布曲線的性質：

(1)對稱中心(尖峰點)為平均數 μ 所在位置

(2)離散的程度則可以用標準差 σ 來描述

(3)標準差較小的分布，其散布的範圍比較小，尖峰也比較陡

(4)當 $\mu=0$ ， $\sigma=1$ 時，稱為標準常態分布，記作 N(0，1)



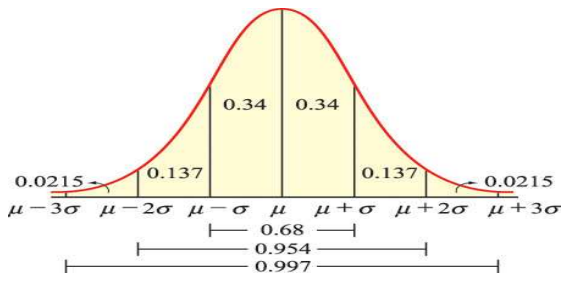
4.常態分布規則(也稱 68–95–99.7 規則)

在任何平均數為 μ，標準差為 σ 的常態分布曲線當中，大約有

68%的數據落在距平均數 1 個標準差的區間 $[\mu-\sigma,\mu+\sigma]$ 內

95%的數據落在距平均數 2 個標準差的區間 $[\mu-2\sigma,\mu+2\sigma]$ 內

99.7%的數據落在距平均數 3 個標準差的區間 $[\mu-3\sigma,\mu+3\sigma]$ 內

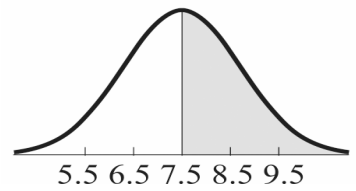


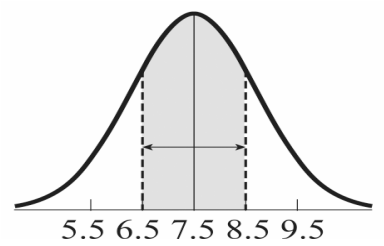
例 5.1：從實驗室的數據證實，人的睡眠時數呈現常態分布，其平均數為 7.5 小時，標準差 1 小時。根據此睡眠分布，試估計下列各項所占的人數比例：

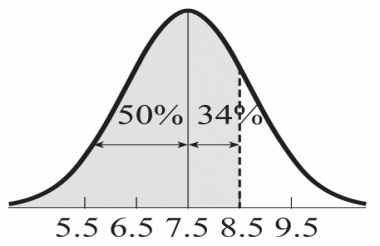
(1)睡眠時數超過 7.5 小時者

(2)睡眠時數介於 6.5 到 8.5 小時者

(3)睡眠時數不到 8.5 小時者







重點 6：二項分布與常態分布

1.二項分布：將重複做成功機率是 p 的伯努利試驗 n 次的機率分布，稱為參數是(n，p)的二項分布

設隨機變數 X 表示成功的次數，即 X~B(n，p)，則：

(1) X 的期望值 E(X)＝np

(2) X 的標準差 σ(X)＝$\sqrt{np(1-p)}$

2.常態分布：考慮 $Y=\frac{X}{n}$ ，即 Y 表示成功的比率，得知 Y 也是隨機變數，則：

(1) Y 的期望值 E(Y)＝p

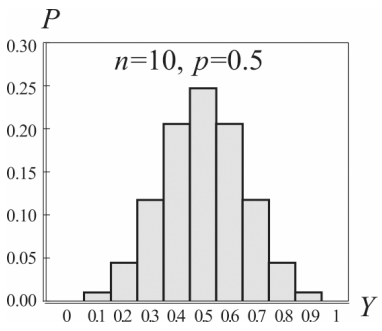
(2) Y 的標準差 σ(X)＝$\sqrt{\frac{p(1-p)}{n}}$

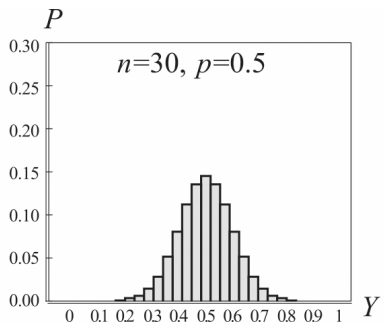
說明：期望值 $E(Y)=E\left(\frac{X}{n}\right)=\frac{1}{n}E(X)=\frac{1}{n}(np)=p$

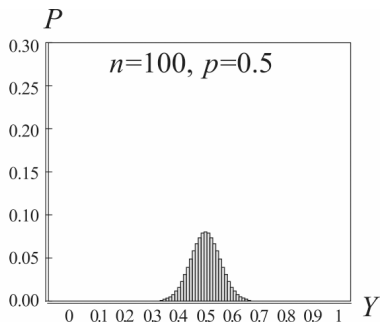
變異數 $Var(Y)=Var\left(\frac{X}{n}\right)=\frac{1}{n^2}Var(X)=\frac{1}{n^2}\cdot np(1-p)=\frac{p(1-p)}{n}$ ，∴標準差 $\sigma(Y)=\sqrt{\frac{p(1-p)}{n}}$

註：Y 的標準差會隨著 n 愈大而變得愈小，其機率分布圖就愈集中

即當 n 夠大時，Y 接近母體平均數的機率很大，這就是大數法則







3.中央極限定理：

在參數是(n，p)的二項分布中，即隨機變數 Y~B(n，p)，則當試驗的次數 n 足夠大時，

成功比率 p 經標準化後的機率分布會近似於標準常態分布($\mu=0$ ，$\sigma=1$)，此性質稱做中央極限定理

4.性質：

在參數是(n，p)的二項分布中，即隨機變數 Y 表示成功的比率，即 Y~B(n，p)，則當試驗的次數 n 足夠大時，

約有 95%的成功比率 Y 會落在區間 $[\mu-2\sigma,\mu+2\sigma]=\left[p-2\sqrt{\frac{p(1-p)}{n}},p+2\sqrt{\frac{p(1-p)}{n}}\right]$ 內，$\mu=p$，$\sigma=\sqrt{\frac{p(1-p)}{n}}$

例 6.1：投擲一枚均勻的硬幣 100 次，設隨機變數 Y 表示正面出現的比率，試估計約 95%的 Y 所在的區間。

例 6.2：已知袋中有 5 個球，其中 2 個是紅色球，從袋中每次取出一球，取完均放回，連取 24 次。
設隨機變數 Y 表示取出紅球的比率。
(1)求 Y 的期望值與標準差　　(2)重複此試驗多次，估計約 95%的紅球比率 Y 所在的區間

重點 7：信賴區間與信心水準的解讀

1.信賴區間：使母體平均數 μ 落在某一範圍，並以區間表示，稱為信賴區間，通常為估計值 ± 抽樣誤差
即[估計值－抽樣誤差，估計值＋抽樣誤差]

註：抽樣誤差：母體比率 p 與樣本比率 $\hat{p}$ 的差＝$p-\hat{p}$ 稱為抽樣誤差

2.信賴區間之計算：母體比率 p 為真實值，抽樣時，以樣本比率 $\hat{p}$ 代替 p，即 $p=\hat{p}$

(1) 68%、95%、99.7%的誤差值分別為 σ 、2σ 、3σ

(2) 68%信賴區間：$[\mu-\sigma\ ,\ \mu+\sigma]=[p-\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ,\ p+\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$

95%信賴區間：$[\mu-2\sigma\ ,\ \mu+2\sigma]=[p-2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ,\ p+2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$

99.7%信賴區間：$[\mu-3\sigma\ ,\ \mu+3\sigma]=[p-3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ,\ p+3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$

3.信心水準：使正確值落在信賴區間內有 n%之信心，稱為 n%之信心水準

註：信心水準：在抽樣很多次中，每次所求得的信賴區間有 n%會涵蓋正確值
信心水準：正確值落在信賴區間之機率為 n%，稱為 n%之信心水準

4.解讀：以 95%為例

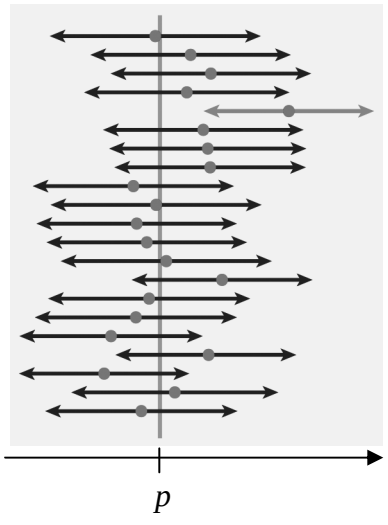
在 95%的信心水準下的信賴區間$[\mu-2\sigma\ ,\ \mu+2\sigma]=[p-2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ,\ p+2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$

稱做 p 的一個「95%的信賴區間」，其中 $2\sigma=2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ 稱做 95%信心水準下的抽樣誤差

註：右圖信賴區間的概念(以 95%為例)

1.圖中共有 21 條線段，每條線段代表一個在 95%信心水準下抽樣所得的信賴區間，圓點代表樣本的比例 $\hat{p}$ ，有些線段包含真正的 p (有 20 條)，有的線段並不包含真正的 p (只有一條)，這種偏差的抽樣結果也許會發生，但機率不大(約 1–95%=5%)

2.實際的情形只會進行一次抽樣調查，也就是說只會得到一個信賴區間，這個區間含或不含真正的 p 值我們並不知道。
我們只用「對此區間，我們有 95%的信心認為它將涵蓋真正的 p 值」來描述
即「若重複抽樣下，約有 95%的區間會涵蓋真正的 p 值」這個特性



註：通常將信心水準定為 95%，希望誤差不超過 3%，則只需抽樣數約 1000

例 7.1：某民調公司做總統大選支持度調查，成功訪問了 1100 位合格選民，其中有 605 位表示支持甲候選人。則：
(1)求這次調查中，甲候選人的支持度
(2)在 95%的信心水準下，這次調查的抽樣誤差是多少個百分點？
(3)計算 95%的信賴區間

例 7.2：某廠商委託民調機構在甲地調查聽過該品牌洗衣粉的居民占當地居民之百分比(以下簡稱為「知名度」)。結果在95%信心水準之下，該品牌洗衣粉在甲地的知名度之信賴區間為[0.608，0.672]。試問此次民調中：

(1)該品牌洗衣粉在甲地的知名度為多少？

(2)抽樣誤差為何？

(3)共成功訪問幾位甲地民眾？其中有多少人聽過該產品？

例 7.3：甲參選角逐某里的里長寶座，其競選團隊隨機抽樣 100 人，其中有 64 人對甲表示支持。則：

(1)求甲支持率 95%的信賴區間

(2)在相同的支持率與信心水準的條件下，欲使信賴區間長度縮短一半，需抽樣多少人？

例 7.4：設在不同的抽樣調查中，分別訪問 1200 人，得樣本滿意度比例為 $\hat{p}_1=0.3$， $\hat{p}_2=0.5$， $\hat{p}_3=0.8$，試求在 95%的信心水準下，何者的信賴區間最長？

※常態分布之信賴區間

例 7.5：為了驗證一枚古硬幣是否為均勻的硬幣，某人做了 600 次的投擲試驗，其中有 240 次出現正面。則：

(1)求此硬幣出現正面比率 95%的信賴區間

(2)求此硬幣出現正面比率 99.7%的信賴區間