中央極限定理的『形象』

last modified July 22, 2008

中央極限定理是統計學應用上很重要的基礎, 許多理論也都以中央極限定理作爲假設的依據, 在學理及應用上都佔有一席之地。本章透過對中央極限定理的描述, 進一步以程式繪圖去驗證, 盼藉此徹底了解中央極限定理的眞正意涵。

要利用程式與繪圖來解釋或證實定理需要經過相當程度的訓練, 本章是一個不錯的開始。隨著練習的步驟一步步展開, 耐心的操作與細心的觀察, 非但程式寫作技巧會有很大的進步, 對抽象或艱澀的數學也比較不懼怕。

本章將學到關於程式設計

MATLAB的程式結構、程式的邏輯概念與程式檔案的管理。

⟨本章關於 MATLAB 的指令與語法⟩

指令: input, isempty, if, histfit

語法: 複合指令 (指令中使用另一個指令)

1 背景介紹

中央極限定理的基礎是大數法則, 其定理如下

大數法則:([1]第六章)
從同一母體隨機抽樣出n個樣本, 當n很大時, 樣本平均數$\bar{x}$會很接近母體平均數μ, 也就是,

$$\bar{x} \xrightarrow{n\to\infty} \mu$$

中央極限定理:([1]第六章)
從一個母體中抽樣一組n筆資料, 算出樣本平均數$\bar{x}$, 如果n很大(通常要求$n \geq 30$), 則$\bar{x}$的分配會接近常態分配, 而且$\bar{x}$的平均數仍爲原母體的平均數μ, 但$\bar{x}$的標準差變小, 爲$\frac{\sigma}{\sqrt{n}}$(σ爲原母體的標準差), 這個常態分配寫成,

$$\bar{x} \overset{n\to\infty}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

圖1說明了中央極限定理。

2 練習

定理是可以透過嚴謹的數學過程推理證明, 也可以透過實驗驗證。以下的練習試圖透過電腦的模擬, 將中央極限定理「物化、」「感官化。」雖非理論的證明, 但透過電腦圖形的展現, 讓一個有點抽象的定理, 活跳跳的呈現在眼前。另一方面, 模擬的實證練習, 可加強對定理的瞭解, 若非對定理瞭解透徹, 做出來的結果常讓自己都莫名其妙。

範例1: 先來驗證大數法則。找幾個不同分配的母體: 如normal,binomial,chi2,F...等, 分別自不同的母體抽取樣本, 樣本數由小逐漸變大 (例如,$5, 10, 50, 100, \cdots$), 觀

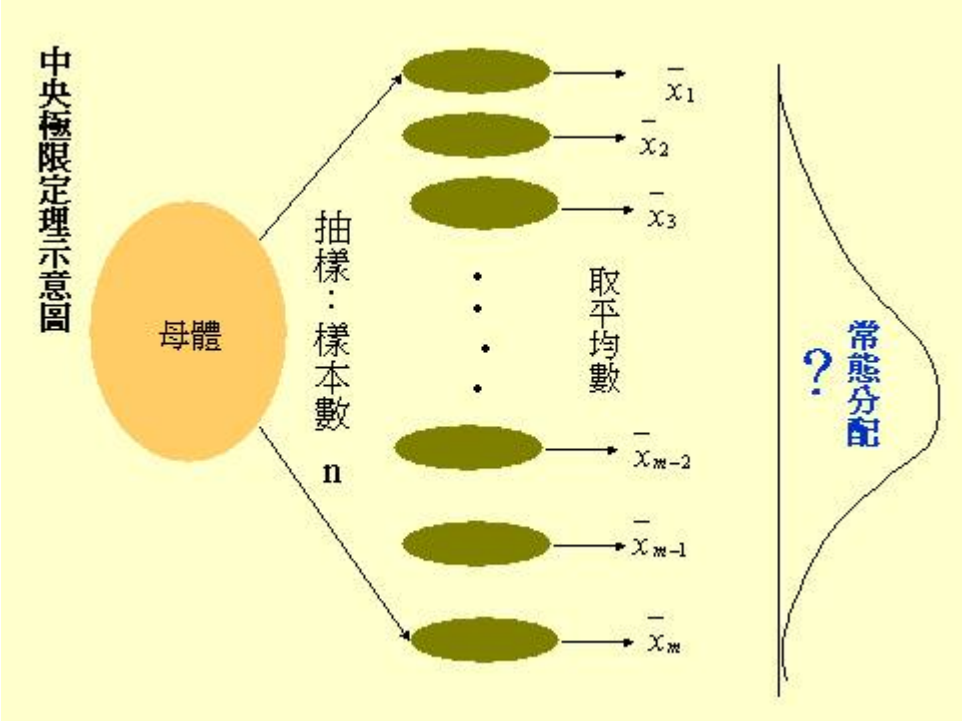

圖 1: 中央極限定理示意圖

察樣本平均數與母體平均數的關係。

在這個實驗裡, 我們需要不斷的改變樣本數來觀察其樣本平均數, 在程式的執行上有個指令:input, 可以在程式執行的過程中產生互動式的效果, 動態的輸入樣本數, 而不需每次修改程式。關於「input」的用法參照 [help input]或也可以用 [doc input]看較詳盡的說明。請注意, 計算出來的樣本平均數必須與母體平均數做比較。除常態分配外, 母體平均數也需要先計算取得。

觀察大數法則的實驗結果時, 如果只是一個個均數的看, 或是一次列出幾百個, 常因數字的紊亂看不出所以然, 只能看個大概。若能以圖形呈現結果, 或許會比較清楚。圖2比較了在三種不同的樣本數下, 其平均數間的比較。母體來自標準常態, 每個樣本數各抽100次, 得到100個均數, 並將之呈現在圖上。透過圖形的觀察可以很清楚的比較出來, 樣本數大小對於其平均數的影響及與母體均數的關係。試著畫出如圖2的圖, 一方面加強畫圖的能力, 一方面學習表達資訊的能力。當然也可以進一步的計算每種樣

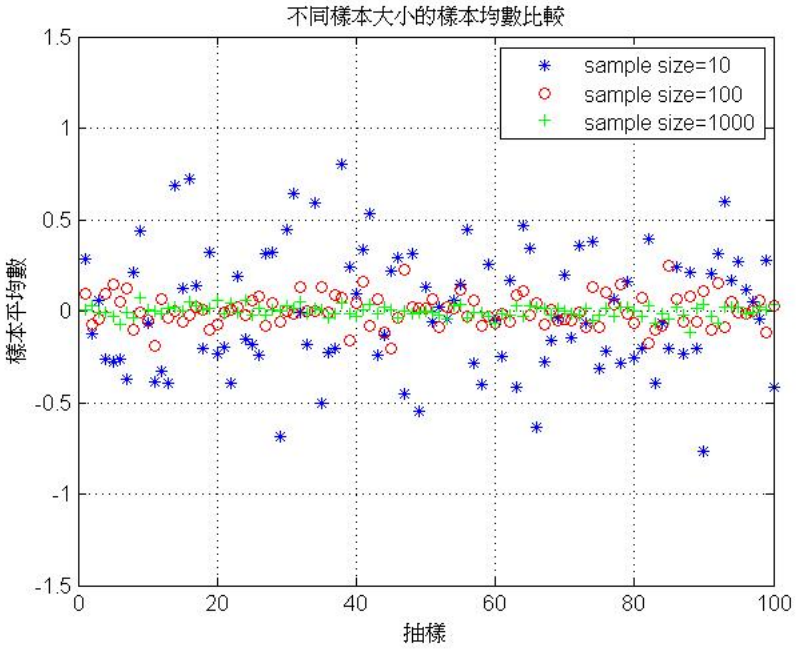

圖 2: 不同樣本大小的樣本均數比較

本數的平均「均數」及變異數。

範例2: 假設母體爲常態分配(μ與σ自訂), 進行100次抽樣, 每次抽30 個樣本, 並計算其均數, 共得100個均數。觀察這100個均數, 有沒有呈現想像中的常態分布呢?(你可以試著先畫直方圖來觀察)。

圖3呈現其中的兩次結果, 從直方圖看來似乎不像常態分配, 是這些均數不具常態分配的特性, 還是抽樣次數太少不足以穩定的出現常態的直方圖, 亦或是抽樣的樣本數太少, 尚不符中央極限定理的假設所致? 如果從直方圖看不出母體的分配, 不妨多做幾次看看, 也試著改變每次抽取的樣本數、或抽樣的次數。

多次的抽樣, 在 MATLAB 裡可以採矩陣的方式產生, 有別於其他語言慣於採迴圈的方式, 以下的程式片段提供這種 MATLAB 式的技巧:

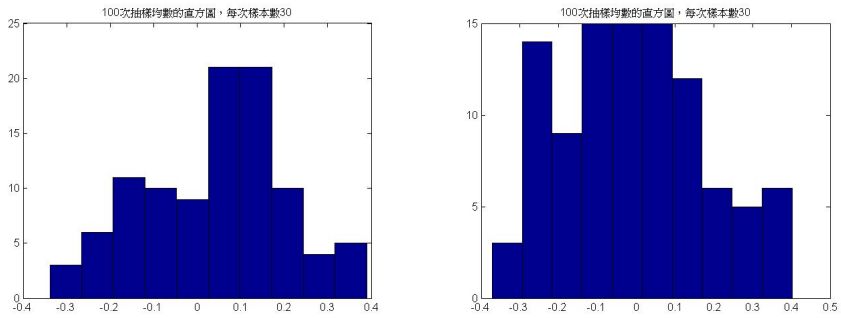

圖 3: 100次抽樣均數的直方圖, 母體爲標準常態

「進行100次抽樣, 每次抽30 個樣本」的三種作法

做法一:

```
X=normrnd( mu, sigma, 30,100);          %產生100x30的亂數矩陣
x=mean(X)          %產生100個平均值, 注意 mean 針對矩陣的作法
```

做法二: 利用迴圈

```
x=[ ];          %變數 x 將儲存最後的100個平均數, 必須先定義爲空矩陣
for i=1:100
        smean=mean(normrnd(mu, sigma, 1,30));
        x=[x smean]          %將平均數一一置入變數 x 的向量中
end
```

做法三: 利用迴圈

```
x=zeros(1,100);          %變數 x 將儲存最後的100個平均數, 先設定爲0
for i=1:100
        x(i)=mean(normrnd(mu, sigma, 1,30));
                                %將平均數逐一放入向量中的特定位置
end
```

範例3: 中央極限定理並不論『原母體是什麼分配』, 不管是連續型或離散型, 對稱或不對稱, 右偏或左偏都無所謂。只要n夠大,$\bar{x}$的分配就會趨向鐘型的常態分配。這個結論在應用上非常重要(譬如: 區間估計與假設檢定), 因爲即使不知道原母體的分配, 也可以從$\bar{x}$的常態分配計算出機率相關的一些數值。在這個練習中, 我們要來驗證這個定理: 請依下列步驟進行:

1. 假設母體爲卡方分配。

2. 先畫出卡方分配的 pdf 圖, 以取得對該分配的認識以及其自由度 k 的改變對 pdf 圖的影響。

3. 進行抽樣; 你要決定兩個數字:(1) n: 每次抽幾個樣本 (2) N: 總共要抽幾次。想想看這裡的 n 與 N 在中央極限定裡的角色是什麼?

4. 搞清楚 n 與 N 你才能進行適當的計算並利用直方圖畫出$\bar{x}$的分配(爲什麼要畫直方圖呢? 回憶一下)。

5. 除了畫出$\bar{x}$的直方圖外, 也請畫出一個$\bar{x}$應該在理論上趨近的常態分配圖(也就是符合中央極限定理時的圖)。當然你得先算出必要的μ與σ。這兩張圖最好疊在一起, 方便觀察, 如圖 5 所示。
   註: 參考本章最後的補充說明, 關於直方圖與機率密度圖的疊合。另外,MATLAB 也提供相對應的指令 histfit, 用法同 hist。

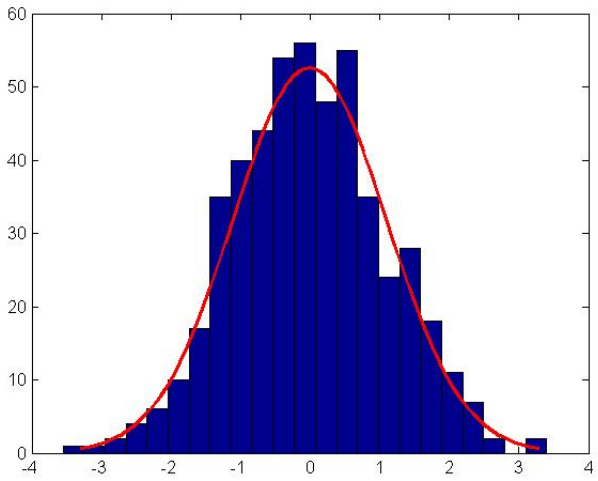

圖 4: 直方圖與其分配

6. 讓 n 與 N 在程式裡面保持彈性, 隨時可以調整。你可能需要不斷的調整這兩個數字, 來觀察中央極限定理所賦予的意義。爲方便程式的進行, 利用指令:input,

在程式進行中手動輸入適當數值, 並配合 isempty 指令讓程式更爲完整與方便。參考以下範例:

```
n=input('輸入樣本數或按 Enter 令 n=30:');
if isempty(n)                                  %檢查是否輸入 n 值, 如果
                                               沒有 (n is empty), 表示按
                                               了 Enter 鍵
        n=30;                                  % n=30爲預設值
end
```

3 觀察

1. 如果畫出來的$\bar{x}$分配圖怪怪的, 有可能你對中央極限定理有所誤解或者程式有問題。爲釐清問題所在, 不妨先試試看母體爲常態的情況, 這時不管這麼畫都應該是常態囉!?

2. 樣本數n要夠大, 但多大才算是『夠大』? 這似乎與原母體的分配有關, 你的觀察是什麼呢? 一般所謂的『大樣本』(n大於等於25或30), 是否適用於所有的分配?

3. 你知道爲什麼在統計的應用裡, 總喜歡用樣本平均數 來當作母體平均數的『估計量』(estimator) 嗎? 在統計推論的範疇裡, 對於抽樣誤差、假設檢定的分析, 中央極限定理扮演關鍵的角色。你可以說出個所以然?

4. 1773年棣美弗 (De Moivre) 提出二項分配, 並討論其極限爲常態分配 (頁6-13 [1])。試著從中央極限定理去探討這個問題。並從作業1的模擬當中去求證。看看能否發覺這個結論中的『極限』到底在哪裡? 在此時常態分配的機率值是否接近二項分配的機率值呢? 下圖描繪出兩者間的共通性。

4 作業

1. 如範例3, 但母體改爲 (1) Binomial (2) Exponential, 並記錄你的觀察。特別針對樣本數n要多大才符合常態的分配? 換句話說, 去感覺一下所謂『大樣本』該大到哪裡才算數?

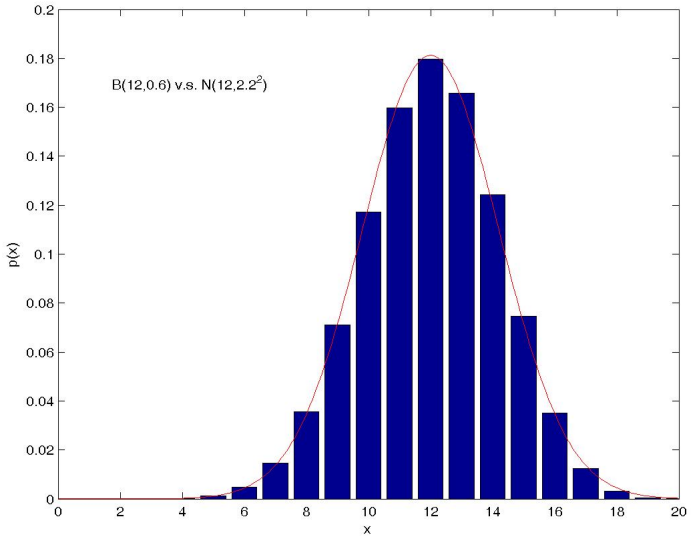

圖 5: 二項分配與常態分配

2. 民意調查的對象通常是從所有可能的人選中抽樣取得。譬如常見的選舉民調是從選民中抽取適當的人數做爲樣本。民調關心的是某位候選人的支持率。假設抽樣 1000 人, 其中有 600 人支持 1 號候選人。我們是否可以下結論:1 號候選人的支持率是

$$\hat{p} = \frac{600}{1000} = 60\%$$

呢? 這樣的說法似乎很難讓人信服。因爲如果同一時間再做一次民調的話, 得到的支持率可能是57%或甚至更慘的50%。如果可以做很多次, 不難想像每一次的結果可能都不一樣。有趣的是, 這麼多次的結果儘管數值不一樣, 似乎存在著某種『規律』。這個規律對民調結果的推論有加強的效果, 比較令人接受。這個練習便是要找出這個『混亂中的秩序』。至於那個比較能被接受的推論, 請參考有關『信賴區間』的單元。

先不管理論怎麼說, 我們已經學會怎麼叫電腦抽樣, 就來寫一支程式模擬民調的結果, 實際觀察那個「傳說」中的「規律」。叫電腦做 1000 次「民調」來觀察這

1000個$\hat{p}$。該怎麼開始呢?

- 抽樣的母體該用哪個分配才對呢? 相關的參數怎麼給呢? 需要什麼假設嗎? 在動手前, 這些問題要仔細想想。
- 你該怎麼來表示你觀察到的結果呢? 或者說, 你該觀察什麼呢? 計算哪些統計量或畫什麼圖可以幫你做判斷呢?
- 靜下心來想想, 對理論再推敲一下, 前面學過的東西有哪些可以幫上忙?

補充: 直方圖與機率圖

中央極限定理的實驗, 是從母體中抽樣、計算均數後, 畫直方圖來觀察其均數的分配。當樣本數夠大時, 其分配趨近常態。若能將該趨近的常態分配機率圖與直方圖畫在一起, 類似圖5, 那樣會更方便觀察樣本數在中央極限定理中的角色。不過因直方圖畫的是次數與機率密度圖的大小等級不同, 不能直接畫在同一張圖上, 需經過適當的轉換, 將次數換成頻率。以下的程式碼提供這樣的轉換參考:

```
[a,b]=hist(x,20);   % x 代表 m 個樣本均數
bar(b,a/m,1);       %利用 bar 做轉換, 其中
                    a 代表每一個範圍的絕對次數
                    b 代表範圍間距,1是直條的寬度
```

利用 bar 來畫直方圖, 效果一樣, 甚至 bar 的表現比較多樣。

參考文獻

[1] 陳順宇, "統計學," 華泰書局。