MATLAB 的機率分配

last modified September 18, 2011

機率相關的問題對於統計學門不僅是基礎的理論, 更是應用上不可或缺的工具。對多數初學機率的學生而言, 機率是有點抽象的, 需要帶點想像的, 學習的障礙不小, 不過有了電腦工具 (如 MATLAB) 的輔助, 那些想像與抽象的部分會變的比較具體些。本章旨在熟悉 MATLAB 處理機率問題的指令及其應用, 包括分配函數的繪製與亂數的產生。

本章將學到關於程式設計

圖形的表現技巧與變化、線上使用手册的運用。

⟨本章關於 MATLAB 的指令與語法⟩

指令:hold, normpdf, binopdf, stem, stairs, bar, normrnd, hist, subplot, copularnd, normplot,qqplot,cdfplot, ecdf, normspec,boxplot, type

1 練習

各種分配函數的「長相」最好能深植腦海, 當面對不同的資料時, 才能迅速的找到對應的母體。MATLAB 提供哪些 pdf 及 cdf 函式呢? 你可不可從 MATLAB 提供的查詢功能查到呢? 你必須練習找找看。

範例 **1.** 先來畫常態分配 $N(\mu, \sigma^2)$ 的 *pdf* 及 *cdf* 圖, 其中的參數 $\mu = 0, \sigma = 1$。請注意調整 x 軸的範圍, 方便看到最完整的分配圖形。計算常態分配的機率密度函數的指令爲 *normpdf,* 請自行以 *help normpdf* 或 *doc normpdf* 查詢其使用方式。其他分配的相關指令也可以在使用手册中查詢到。

- 練習給予不同的 μ 及 σ 值, 看看圖形的變化。
- 固定 μ 值, 改變 σ 值, 練習將每一張圖都疊在畫面上(如圖 *1* 所示)。
- 試試看其他的分配? 找出相對的指令來, 把圖畫出來, 並使用不同的參數。

請注意, 只要是繪圖就是描點法, 要將每一個點的 x, y 值都事先給定或計算好。MTALAB 中 pdf 及 cdf 函式都是用來計算 y 值的。另外本練習爲觀察參數對於一個分配函數的影響, 要求將不同參數的分配圖疊在一起, 方便觀察。MATLAB 利用 hold on 指令 將目前的圖形保留, 隨後畫上的圖會直接疊在上面, 再利用 hold off 取消保留的設定。譬如將三張圖疊在一起的作法:

```
plot(x1,y1)
hold on
plot(x2,y2)
plot(x3,y3)
hold off
```

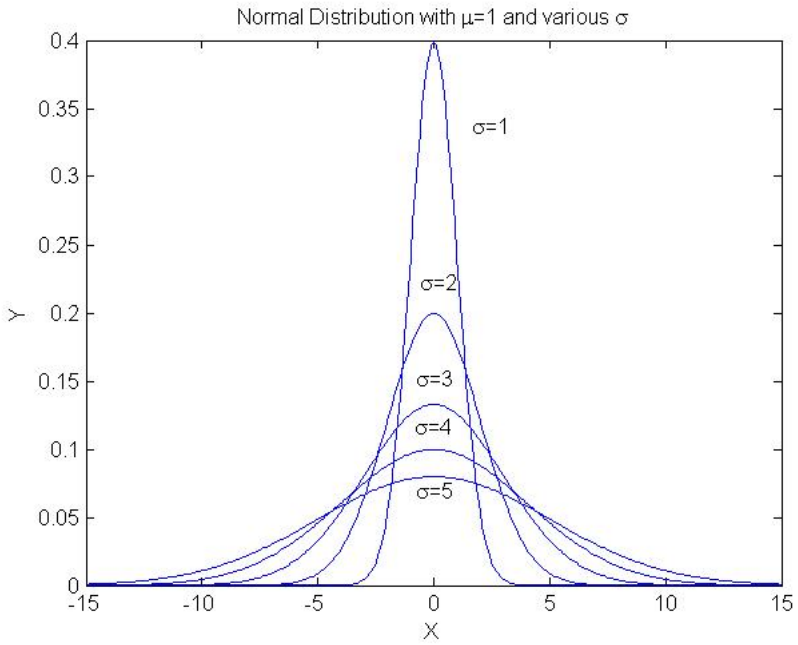


圖 1: 常態分配的機率密度函數(pdf)

另外, 不連續型分配 (如二項分配) 的繪製要特別謹慎, 要注意 x 軸範圍的限制與間距的特性, 不能以一般繪製連續函數的方式大剌剌的一筆帶過, 畫出來的圖必須合乎學理, 不是畫出圖來便是對的。以二項分配爲例,x 座標的選定與母體的選擇息息相關, 假設母體爲$B(20, 0.2)$, 其 pdf 的繪製可以這樣做

$x = 0 : 20;$
$y = binopdf(x, 20, 0.2);$
$stem(x, y)$

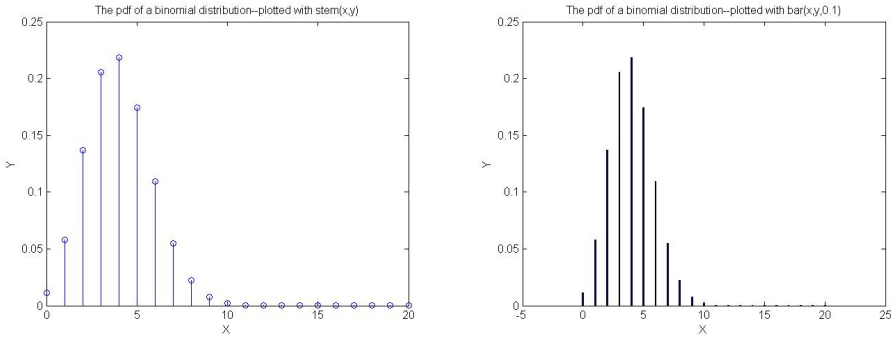
其結果如圖 2(a) 所示,x 的範圍與間距準確的呈現出母體的所有可能性, 太大的範圍沒意義 (如 $x = 0 : 50$), 太小的範圍 (如 $x = 0 : 15$) 與非整數的間距 (如 $x = 0 : 0.1 : 20$) 則是對二項分配的不瞭解, 都應該很警覺的避免。MATLAB指令 stem 可以畫出漂亮的間距圖形, 也有些變化可以應用, 如

$stem(x, y,' filled')$

試看看畫出什麼不一樣的圖形。stem 不是唯一的選擇, 圖2(b) 利用寬度比較窄的長條圖 bar, 也可以畫出類似的效果, 甚至變化更豐富, 譬如指令的第三個參數用來指定長條的寬度

$$bar(x, y, 0.5, 'r')$$

其中第四個參數決定顏色。至於 cdf 圖, 可以採用 stairs 的指令, 畫出如階梯般的機率累積圖。staris指令怎麼用呢? 其實繪圖指令用多了, 猜猜看往往八九不離十, 學習程式語言要有此本事才會越學越輕鬆愉快。



(a) (b)

圖 2: 二項分配的機率密度函數 (pdf), 以指令 (a) stem (b) bar 繪製的不連續函數圖。

範例 **2.** 如何從 *MATLAB* 產生亂數 (*random numbers*)*?* 照例可以從 *Help* 裡面查到亂數相關的指令及其使用方式。其指令參數的給定不外乎是該分配的參數與欲產生亂數的個數。譬如下列指令產生 $M \times N$ 個具常態分配的亂數 (樣本), 放在矩陣 A 裡*:*

```
A=normrnd(mu,sigma,M,N);
```

第三及第四個參數決定亂數的個數與排列的方式 (依實際需要排成矩陣或向量), 不妨多試幾個不同的數字便一目了然。

- 隨意產生 *100*個具常態分配的亂數。你如何確定這些亂數值具備常態分配的特性？ 用 *plot* 畫出這些值, 看看長什麼樣子？ 再試試用直方圖去畫。

```
x=normrnd(0,1,1,100);%標準常態
hist(x)
```

- 重複上個練習, 但試著改變所產生的亂數個數 (變多或少), 與不同的分配。觀察畫出來的直方圖有何不同？ 跟你原本的認知有什麼差別？

直方圖常用來觀察資料的分佈情形 (頻率、落點分佈),MATLAB 對應的指令爲 hist。繪製直方圖需要很有「感覺」, 否則容易畫出連自己都不易看懂的圖, 其中樣本數的多寡牽涉到直方圖選定的組界數 (bins), 尤其困擾初學者, 不妨多練習、不斷的變更組界數並觀察圖形的變化, 在 hist 指令中, 組界數在第二個參數, 不指定時組界數內定爲10。圖 3 展示不同組界範圍的視覺效果。

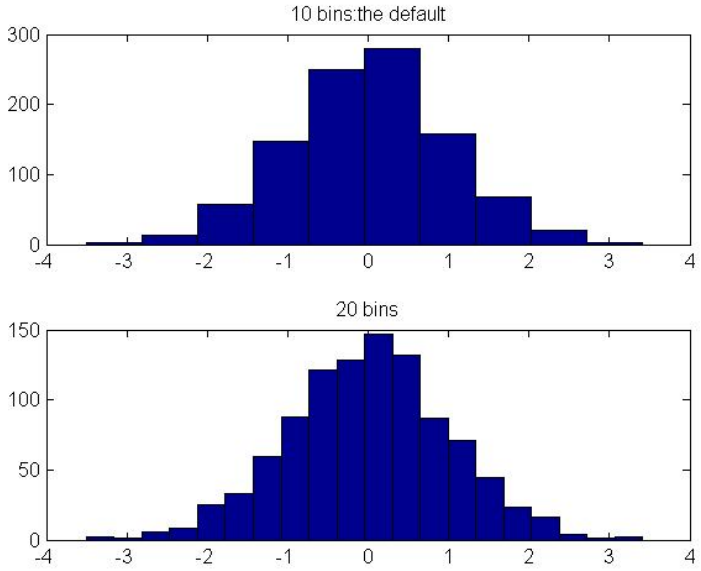


圖 3: 相同的 100 個樣本畫出不同組界數的直方圖, 上圖 hist(x), 下圖 hist(x,20)

從直方圖只能大概判斷樣本是否來自常態的母體, 若能爲直方圖疊上接近的常態分配的機率密度圖, 可以提供更精準的觀察。histfit 指令提供這樣的功能,

```
x=normrnd(0,1,1,1000);
histfit(x)
```

結果如圖 4 左圖所示。而右圖則展示 MATLAB 指令 cdfplot 畫出從樣本資料計算的 Empirical CDF(直線部分), 虛線是理論的 CDF 函數圖。兩者幾乎重疊, 幾可判斷樣本資料來自常態分配。cdfplot 指令若無法從 help 指令得到其使用方式, 表示使用的是早期的版本, 屬於未公開的指令, 但仍可透過在命令視窗輸入 type cdfplot 看到程式的原貌及並從裡面的結構與說明探知其使用方式與限制。另外一個繪製 Empirical CDF 的指令是 ecdf, 這個指令還可以取得一組累積機率值與對應的 X 值, 這一組資料可以用來以 X 值查詢該分配的累積機率值。

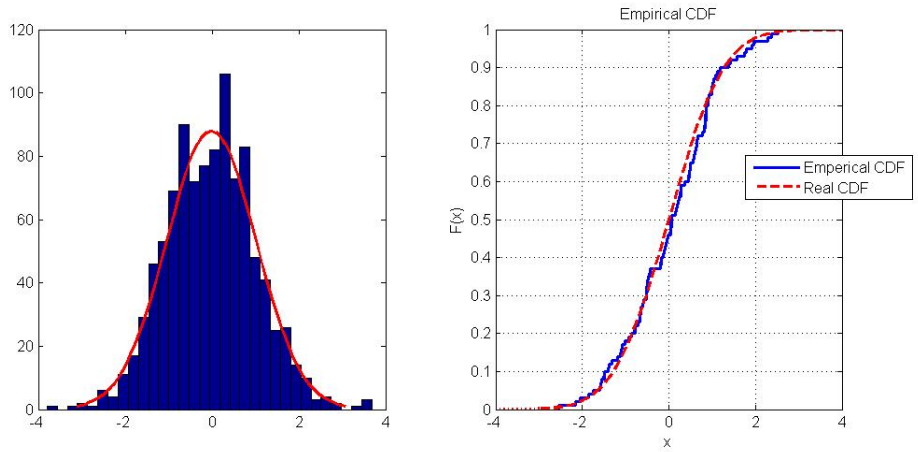


圖 4: 直方圖配上估計的常態分配。

判定樣本是否來自常態分配的母體, 還可以使用指令 normplot 繪製常態機率圖 (Normal Probability Plot), 方式如

```
x=normrnd(0,1,1,100);
normplot(x)
```

圖 5 顯示常態機率圖。當圖中的 '+' 號越趨近線性 (即與虛線越貼近) 時, 表示樣本來自常態的可能性越高。其繪圖原理請參考手册上的說明。讀者不妨試試其他分配的樣本, 觀察 '+' 的分佈情況。其 他指令如 qqplot 也提供類似的功能, 做法也相同。

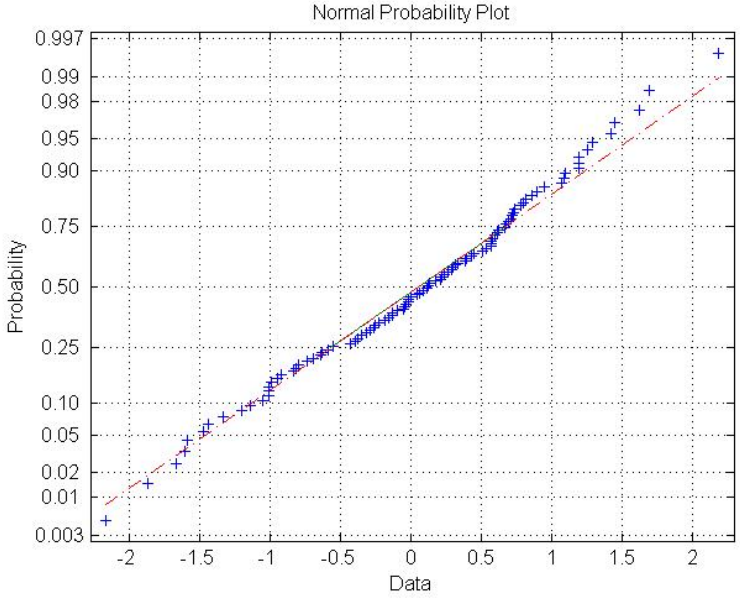


圖 5: 常態機率圖

範例 **3.** 執行下列步驟*:*

- 產生兩組資料, 每組各 *100*個樣本。這兩組資料來自兩個不相關的變數 (*uncorrelated variables*) y_1 及 y_2 (可以先假設變數都具常態分配)。畫一個散佈圖來呈現資料間的不相關性 y_1 *vs.* y_2。並且計算出兩者間的相關係數r。
- 同上, 但資料來自兩個完全相關 (*completely correlated*) 的變數並觀察相關係數, 例如*:* $y_1 = cy_2$,c 代表一個常數。
- 同上, 但資料來自兩個部分相關 (*partially correlated*) 的變數並觀察相關係數。如何模擬「部分相關」的兩組資料呢*?* 動動腦筋*!*

本範例假設學習者對於兩變數間的相關性及其散佈圖的模樣已經有相當的概念, 方能藉此逐步模擬出自己所認知的相關性, 特別是模擬部分相關的資料。此外, 當畫出的幾張圖形具有比較意義時, 可以利用 MATLAB 提供的畫面切割技術, 將幾

個圖分別呈現在畫面的不同位置, 這個指令是 subplot。用來作爲繪圖前位置的指定, 使用方式如下 (圖 3 展示了下列切圖的方式)

```
subplot(2,1,1),plot(x1,y1)
subplot(2,1,2),plot(x2,y2)
```

至於 subplot 的參數代表的意義, 在命令視窗上輸入 help subplot, 即可一目了然, 當然親自動手畫圖的時候便可以領會, 無須在此贅述。寫得太仔細會把學的人教笨了, 豈不罪過。圖 6 展示 subplot 的 2×2 切圖方式及不同相關性變數的散佈圖, 其中的 ρ 代表相關係數。

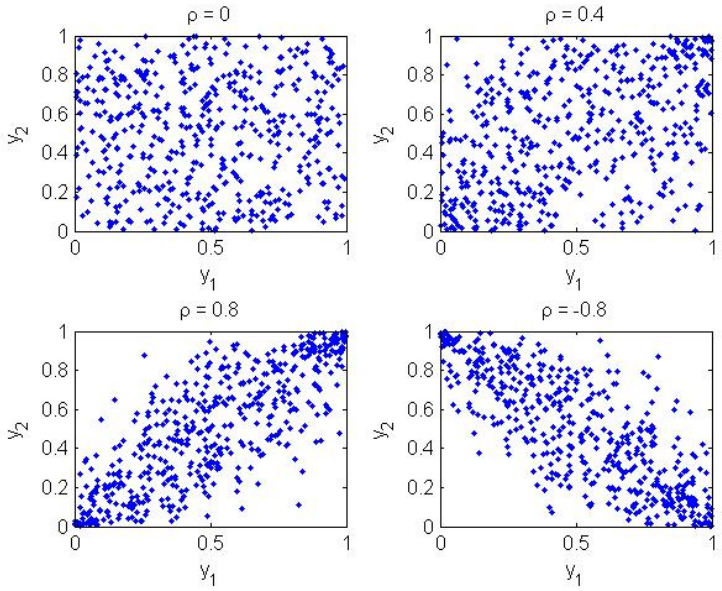


圖 6: subplot 的 2×2 切圖方式及不同相關性變數的散佈圖。

2 觀察

1. 本單元繪製機率密度函數時, 也可以嘗試使用 ezplot 指令, 譬如繪製標準常態分配的 pdf 圖,

```
ezplot('normpdf(x,0,1)')
```

建議經常使用, 相當方便。不過當需要疊圖時,ezplot 似乎不太好搞, 不妨試試看。

2. MATLAB 關於機率方面的指令不少, normspec 也是一個常用來表達機率概念的圖, 其使用方式如下, 結果如圖 7。

```
normspec([-5  5],0,3)
```

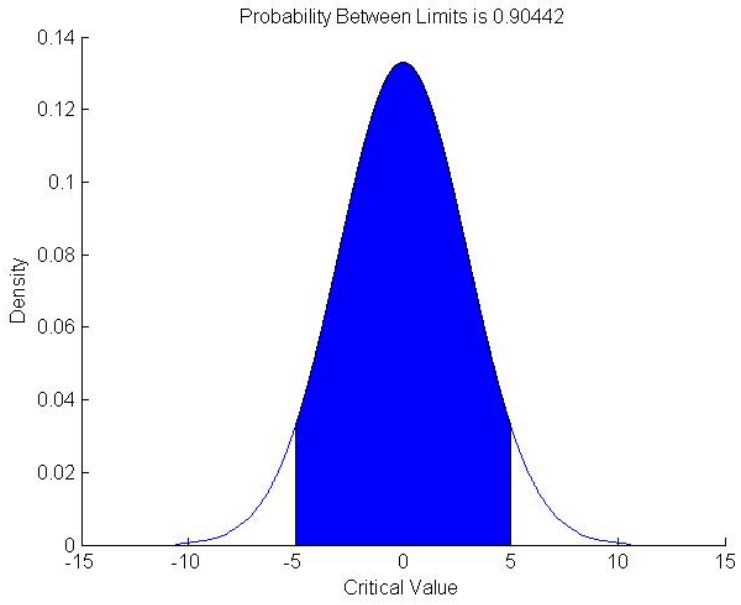


圖 7: 常態分配函數的部分機率圖。

其中第一個參數指出涵蓋的範圍, 第二與第三個參數則是常態分配的參數。normspec 指令除畫出覆蓋的面積圖之外, 也計算出這個面積並展示在 title。

3. 畫出來的 pdf 與 cdf 圖是否都符合你的預期呢? 如果不是, 去翻翻統計學、機率論的書驗證一下。

4. 釐清楚機率分配的 pdf 圖與依亂數值產生的直方圖。不要搞混了!兩者間有什麼異同?

5. 畫出來的直方圖若與自己認定的不符時, 請特別注意是畫圖技巧不好, 還是指令的操作錯誤或是觀念的錯誤。通常直方圖的畫圖技巧必須注意樣本數及範圍間距的選擇。

6. 透過這個單元的練習, 當給予一組隨機資料時, 你有多少把握知道其原始的分配是什麼? 這與資料量的多寡有關嗎? 除了畫直方圖之外, 還有沒有其他方式可以提供更多的參考訊息呢?

7. Boxplot 是一種常用的統計圖, 可以立刻看出其代表變數的位置及其分佈狀況, 也可以呈現出資料分佈的對稱性或偏斜情況。另外最大、最小值及 outliers 也都可以清楚的看出來, 方便迅速的比較並得到粗淺的認識。MATLAB 的 boxplot 指令的幾種做法如下:

```
%先模擬兩組資料
x1=normrnd(0,1,100,1);
x2=normrnd(1,2,100,1);
boxplot([x1 x2])
```

圖 8 顯示了兩組資料的 Boxplot, 每個「盒子」裡面有三條水平線, 由上而下分別代表 75,50,25 百分位的位置。「盒子」上下個有一條垂直虛線, 頂端分別代表這組資料 (扣除 outliers 後) 的最大值與最小值。「+v 代表 outliers, 垂直虛線的長度是 interquartile(75 百分位- 25之百分位的距離) 的 1.5 倍 (可以調整)。boxplot 指令爲方便比較, 有可以增加如下的選項:

```
%先模擬兩組資料
boxplot([x1 x2],'notch','on','labels',{'Group A','Group B'})
```

圖 9 將中位數的上下切出 95%的信賴區間的缺口, 方便兩組並列時比較中位數的差異。另外也加入自訂的組別名稱。圖 8 與圖 9 的資料來自對稱的

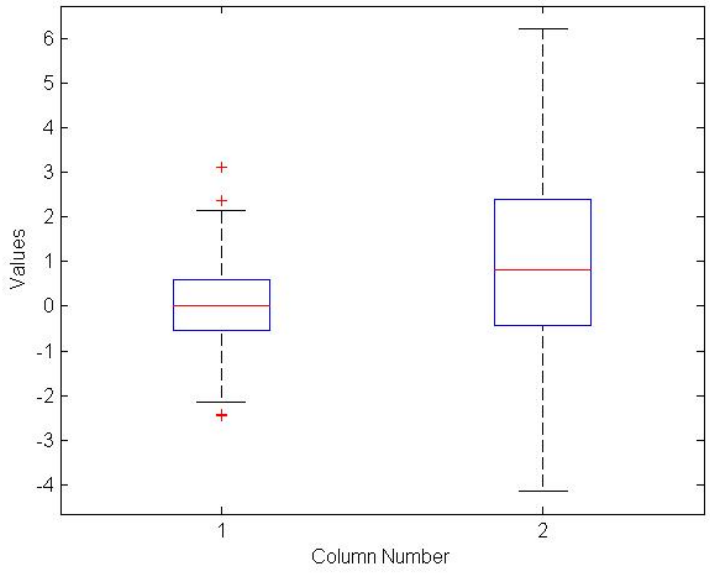


圖 8: 兩組資料的Boxplot 圖。

常態分配, 不妨試著拿偏斜分配的資料畫幾張 Boxplot, 才能對 Boxplot 的特質有所掌握。

8. 觀察兩組資料的相關性, 相當具實用價值。本練習進而去模擬產生具某種相關性的兩組實驗資料。產生模擬資料對研究工作而言, 往往是必備的基本動作。

9. 兩變數的相關性不一定是線性的, 這裡只是給予線性的訓練。你也可以試試產生非線性相關的資料, 再去計算 r, 看看發生什麼事了?

10. 請觀察變數資料的產生與關係的形成, 及最後散佈圖的樣子, 要牢牢的將這些東西連結在一起。加強對資料的感覺, 培養與資料間的感情, 即所謂的『資料感』。

3 作業

1. 畫出下列分配的 pdf 及 cdf 圖, 並觀察參數的改變與圖形的關係。將相關的圖疊在一起 (至少5張圖疊成一張)。所有的參數資料盡量寫在圖形的空白

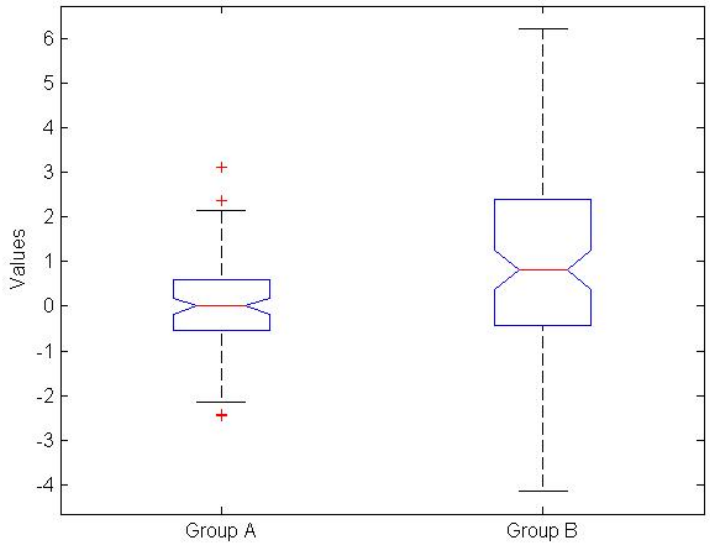


圖 9: 兩組資料的Boxplot 圖 (加選項)。

處。

- Normal Distribution:(1) 固定 μ 改變 σ,(2) 固定 σ 改變 μ。
- Chi Square Distribution: 觀察在不同自由度 ($\nu < 30$) 的變化情形。另外可以觀察當自由度很大時, 卡方分配的長相? 譬如, 畫一張圖將自由度 1000 的卡方分配 與 常態 $N(1000, 45)$ 的 pdf 畫在一起。
- Binomial Distribution: 自行調整參數。
- F Distribution: 當兩個自由度的參數 ν_1, ν_2 由小變大時,F 分配的樣子會如何改變呢? 有其極限嗎? 請仔細觀察。畫出有代表性的分配圖。
- T Distribution: 觀察當自由度由小變大時, 圖形的變化情形。當自由度很大時是否接近標準常態? 請以圖形表達出來。
- β Distribution: 這個分配很豐富, 千變萬化, 非常精采。當兩個參數 a, b 大小不同時, 分配的樣子會如何改變? 譬如, 觀察當 $b > a$ (固定 a, 調整 b) 時、當 $a > b$ (固定 b, 調整 a) 時及當 $a = b$ 時。

- Exponential Distribution: 自行調整參數。

其他如幾何分配、卜瓦松分配都可以嘗試。

2. 產生5組亂數 (來自5個不同的分配), 其個數與分配自行決定, 分別畫出直方圖。觀察畫出來的圖是否符合預期呢? 你必須確定這一點。不能畫了就算!

3. 分別產生具左偏與右偏分配的資料各一組, 畫出 Boxplot。

4. 假設 x 爲一服從標準常態分配的變數, 令變數 $y = x^2$。利用適當的亂數指令產生 1000 個變數 y 的樣本並繪製其直方圖。類似這樣的抽樣分配在機率課本找到很多, 不妨多做幾個。

5. MATLAB 的 statistics toolbox 也提供一組叫做「Copulas」的指令, 可依指定的相關係數產生兩個 (或以上) 變數的亂數, 其使用方式與範例可以在線上使用手册以「Copulas」爲關鍵字查詢到, 譬如

```
n = 500;
U = copularnd('Gaussian',[1 0.8; 0.8 1], n);
plot(U(:,1),U(:,2),'.');
```

其中第二個參數爲相關係數矩陣。這裡產生的亂數具均等分配, 其值在 $(0, 1)$ 之間。如果你使用的 MATLAB 找不到 copularnd 這個指令, 表示版本較早。