作業 1　高斯分佈抽樣

一、單變數高斯分佈

1. 實驗目的

透過模擬單變數高斯分佈的隨機抽樣，使用最大概似估計（MLE）方法來估計母數平均值與變異數，並觀察估計值是否隨抽樣次數增加而趨近真實母數。

2. 實驗方法

- 母數：平均數 = 2 ，標準差 = 2
- 每輪抽樣數：每輪從 N（2, 2^2）抽取 n = 50 個樣本
- 使用公式計算 MLE：

$$\mu_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \sigma^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{ML})^2$$

- 總共重複 R = 100 輪
- 觀察單輪抽樣的直方圖與理論分布比較
- 觀察多輪抽樣後，μ_{ML}與σ^2_{ML}的收斂情形

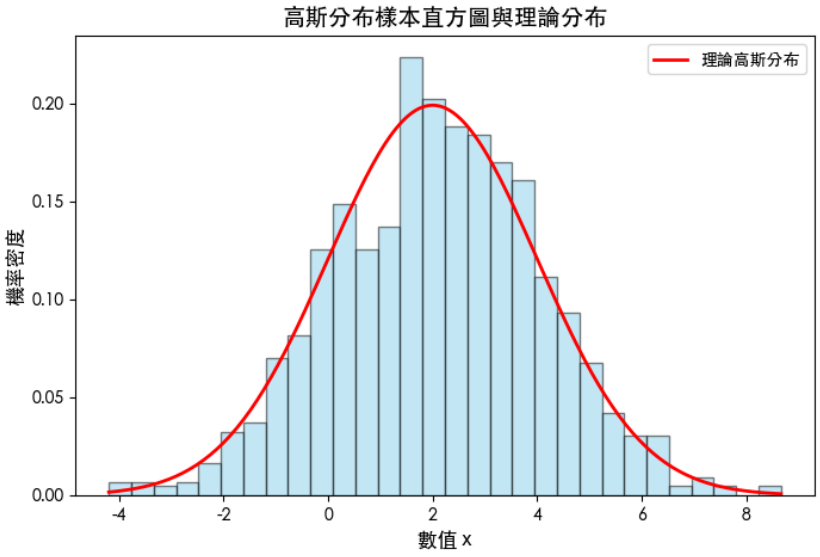
3. 程式碼執行截圖

- 終端輸出文字：

```
母數 μ=2, σ²=4
估計均值 μ_ML 平均=2.002
估計變異數 σ²_ML 平均=3.890
```

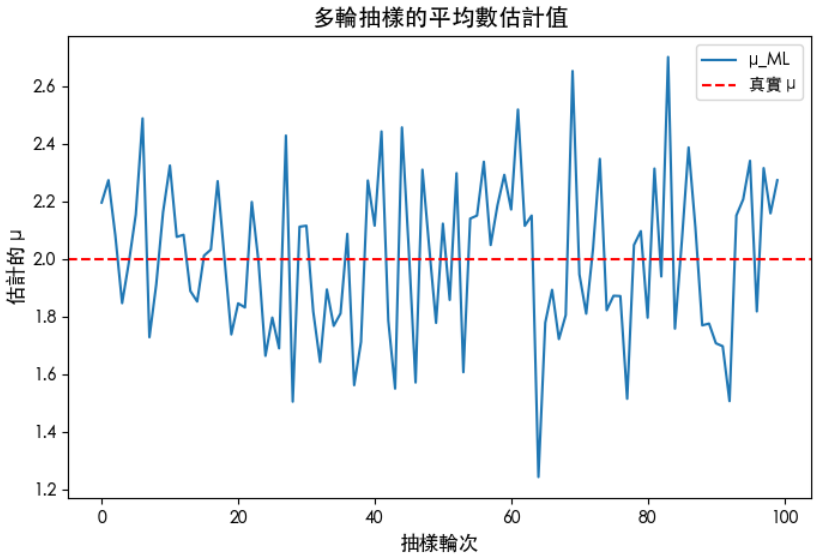
從文字輸出可以觀察到平均數的估計值在 100 輪抽樣後的平均約為 2.002，與真實母數$\mu = 2$ 十分接近，顯示最大概似估計在平均數方面具有無偏性，而變異數的估計平均為 3.890 ，略小於真實值 4，說明最大概似估計對變異數具有負偏差，這與其公式中分母為 n 有關（相較於不偏估計的 n - 1 分母）

- 高斯分布樣本直方圖與理論分布：



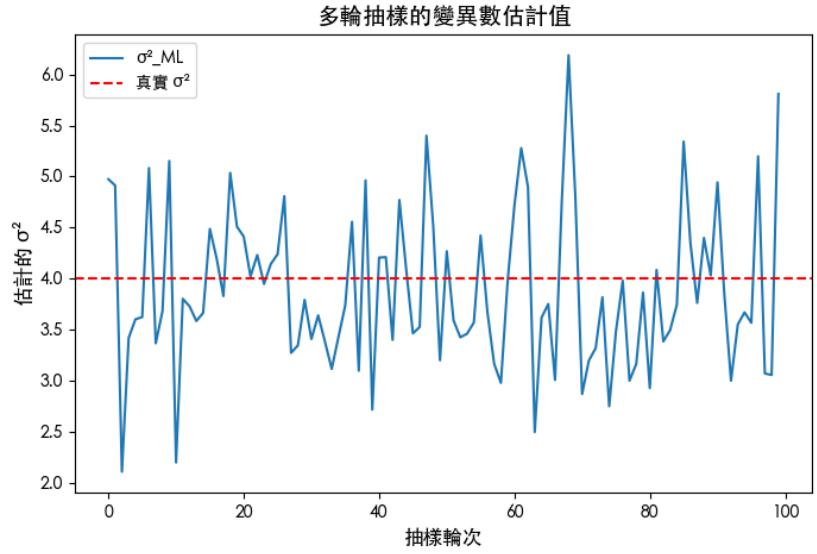
此圖顯示從真實常態分佈抽樣的結果與理論分佈的比較，藍色直方圖代表從 N（2, 2）抽出的 1000 筆樣本所形成的分佈，紅色曲線則為理論上的常態分佈密度函數，從圖中可以觀察到，實際樣本分佈大致貼近理論分佈，說明了隨機抽樣可以合理反映母體分佈形態。

• 多輪抽樣的平均數估計值：



此圖顯示每輪抽樣後所得到的平均數估計值變化情形，可以觀察到估計值在不同輪次間雖有小幅波動，但大致集中在真實值 $\mu = 2$ 附近，顯示平均數的最大概似估計在多次抽樣下表現穩定，具有一致性與無偏性。

• 多輪抽樣的變異數估計值：



此圖顯示每輪抽樣後所得到的變異數估計值變化情形，與平均數相比，變異數的估計值波動稍大，且普遍低於真實值 $\sigma^2 = 4$ ，這反映出最大概似估計在有限的樣本

下對變異數估計具有負偏誤，這是因為 MLE 使用的變異數分母是 n ，不是 n - 1 ，若使用不偏估計，則估計的結果會更接近真實值。

每一次輸出的結果會稍微有點不同，這是由於程式中每輪皆使用隨機抽樣的方式從常態分布中取樣（np.random.normal(mu, sigma, n)）。隨機樣本具有不確定性，就算是使用相同的母體還有樣本數，不同輪次抽出的樣本仍可能略有差異，進而影響每輪所算出的平均數和變異數，因此整體模擬的平均結果也會有小幅波動，是統計模擬中很常見的現象。

4. 實驗結果

- 抽樣的直方圖大致呈現鐘型，與理論高斯分佈曲線接近
- μ_{ML} 在多輪後的平均約等於真實母數 $\mu = 2$
- σ^2_{ML} 在多輪中略低於真實值 $\sigma^2 = 4$ ，但整體波動仍在合理範圍內
- 隨著輪數增加，估計值的波動逐漸減少，符合最大概似估計的一致性

5. 結論

本實驗透過模擬從單變數常態分佈 N（2, 2^2）中進行重複抽樣，利用最大概似估計法來估計平均數和變異數，實驗結果顯示，在總共進行 100 輪、每輪抽樣 50 筆資料的情況下，平均數估計值的平均為 2.002 ，變異數估計值的平均為 3.890 ，這個結果提供了幾點值得探討的觀察。

首先，平均數的估計結果很接近真實母數 $\mu = 2$ ，誤差只有 0.002 ，幾乎可以視為無偏，這證實了最大概似估計對於平均數具有良好的估計性質，在樣本數適當時，MLE 對平均數的估計不僅無偏，且具一致性與漸近正態性，可以穩定收斂至真實值。

其次，變異數估計值的平均為 3.890 ，略低於真實變異數 $\sigma^2 = 4$，誤差約為 - 0.110，這樣的偏小現象符合理論預期，因為最大概似估計在變異數的計算中使用的分母為 n ，非不偏估計所使用的 n - 1，因此，MLE 對變異數本質上是有偏估計，特別在樣本數不大的情況下偏誤較明顯，不過在本實驗中，由於每輪樣本數為 50 ，屬於中等規模，偏差已經非常有限，顯示此偏誤在實務上是可以接受的。

此外透過觀察多輪抽樣的估計趨勢圖可以發現，無論是平均數還是變異數的估計值，隨著抽樣輪數增加皆呈現出穩定與集中趨勢，進一步證明最大概似估計在重複抽樣下具有良好的收斂性。同時，單輪抽樣所生成的直方圖與理論高斯分佈曲線高度吻合，說明實際樣本分佈亦能良好反映母體特性。

總而言之，最大概似估計法在本實驗條件下，對平均數的估計準確穩定，對變異數雖理論上有偏，但偏誤極小且收斂性良好，整體估計效果良好。此實驗不僅驗證了統計理論中 MLE 的估計特性，也顯示在實際應用中，MLE 能作為有效估計方法，尤其在樣本數適中以上的情況下具有良好的表現。

二、雙變數高斯分佈

1. 實驗目的

透過電腦模擬雙變數高斯分佈，驗證最大概似估計，在抽樣過程中的收斂特性。我們將自行設定母數，包括平均向量和相關係數，並進行多輪抽樣。在每一輪抽樣後，分別計算樣本的均值與變異數，觀察其與母數之間的差距。透過多輪實驗，希望

可以看到隨著抽樣的次數增加，估計值會逐漸趨近於母數，說明 MLE 在大樣本下具備一致性。

2. 實驗方法

• 母數設定：$\sigma_1 = 1.5$、$\sigma_2 = 1.0$、$\rho = 0.6$、$\mu = (2, -1)$

協方差矩陣：

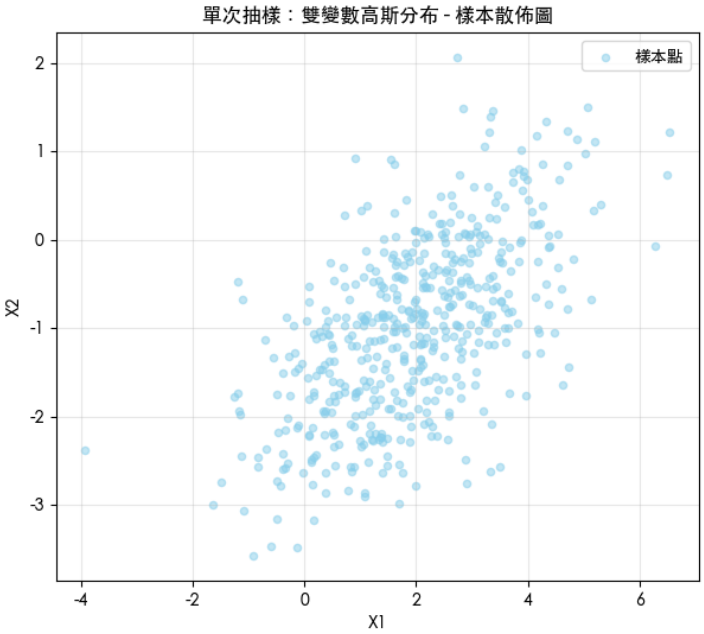
$$\Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix} \begin{bmatrix} 2.25 & 0.9 \\ 0.9 & 1.0 \end{bmatrix}$$

• 每一輪抽取 n = 100 筆資料，共抽取 T = 30 輪

• 每一輪計算：樣本均值向量與樣本協方差矩陣

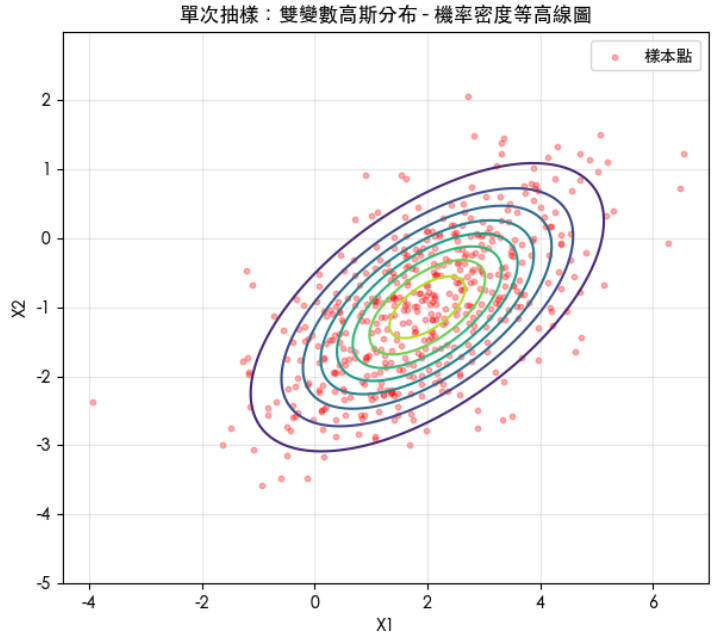
3. 程式碼執行截圖

• 單次抽樣的樣本散佈圖：



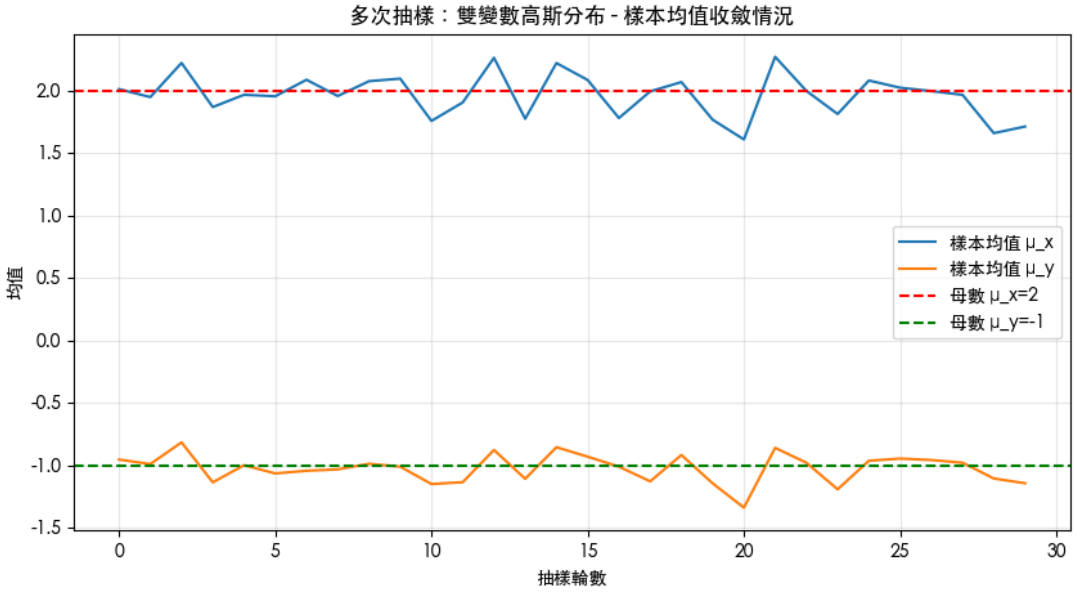
此圖顯示從設定的雙變數高斯分佈（$\rho = 0.6$、$\mu = [2, -1]$）中抽取 500 筆樣本所形成的散佈圖，樣本點大致集中在 [2, -1] 附近，並呈現右上至左下斜向橢圓形分佈，反映變數之間具有正相關，這與我們設定的協方差矩陣一致，說明抽樣結果符合期望的統計特性。

• 雙變數高斯分佈的理論等高線圖：



圖中為根據理論機率密度函數所繪製的等高線圖，並重疊顯示紅色樣本點，可以觀察到多數樣本集中於機率密度高的區域，即等高線中心附近，樣本分佈輪廓也跟橢圓形等高線近似，進一步驗證樣本確實來自設定之雙變數常態分布。

• 多輪抽樣下樣本均值的收斂圖：



此圖顯示 30 輪獨立抽樣中，每輪樣本均值（μ_x 與 μ_Y）的變化情況，雖然樣本均值在初期會有隨機波動，但隨著抽樣輪數增加，其估計值逐漸穩定並收斂至紅色與綠色需線索代表的真實母數 $\mu_x = 2$、$\mu_Y = 1$，此現象說明樣本均值為一致估計量，即樣本數越多時，估計結果越趨近於真實值。

4. 實驗結果

- 樣本大致分布在平均向量 [2, -1] 附近，整體分佈呈橢圓形，斜率方向與設定的正相關係數 $\rho = 0.6$ 一致
- 理論機率密度函數的等高線圖與樣本點分佈高度吻合
- 多數樣本點落在機率密度較高的中心區域，證實模擬分佈正確
- 使用了最大概似法計算出的樣本均值在 30 輪抽樣中波動逐漸變小
- 每輪樣本的協方差矩陣雖有些微波動，但整體趨近於理論協方差
- 樣本均值呈現無偏性、一致性，協方差估計亦穩定逼近母數，符合統計理論

5. 結論

這個實驗透過模擬雙變數高斯分佈的抽樣與參數估計，成功驗證了樣本均值和樣本協方差矩陣作為統計估計量的有效性和收斂性。

首先，平均向量的估計採用最大概似法，透過計算每輪樣本的平均值來近似母數，結果顯示即使單輪樣本數僅為 100 ，所得到的樣本均值在 30 輪抽樣後仍穩定收斂於母體參數，並隨抽樣輪數增加而減少波動，這證實了樣本均值為一個無偏且一致的估計量，擁有良好的統計性質。

接著，協方差矩陣的估計則使用無偏估計（將樣本變異數除以 n - 1 而非 n），這樣的做法可以消除由樣本平均帶來的估計偏誤，讓協方差估計更接近母體的真實值，儘管每輪抽樣下估計出的協方差略有變動，但整體仍明顯圍繞在母體協方差矩陣附近，顯示樣本協方差矩陣亦為穩定有效的估計方式。

在圖形部分，散佈圖與理論等高線圖顯示樣本點主要集中在高機率密度區域，且整體分佈形狀與母體的協方差關係相符，進一步驗證了樣本來自所設定的雙變數常態分布，此外，樣本均值收斂圖視覺化呈現了估計值如何隨輪數穩定趨近母體參數，提供直觀的統計解釋。

綜合而言，實驗成果成功展現了雙變數常態分布下統計估計的正確性與實用性，樣本均值與樣本協方差這兩個常見的估計量，在模擬中都展現出良好的統計性質，包括無偏性、一致性、逼近真實參數的能力，這些方法不僅適用於基礎統計分析，也為後續在機器學習、資料建模等應用中，提供了堅實的理論基礎與實作依據。

三、多變數高斯分佈

1. 實驗目的

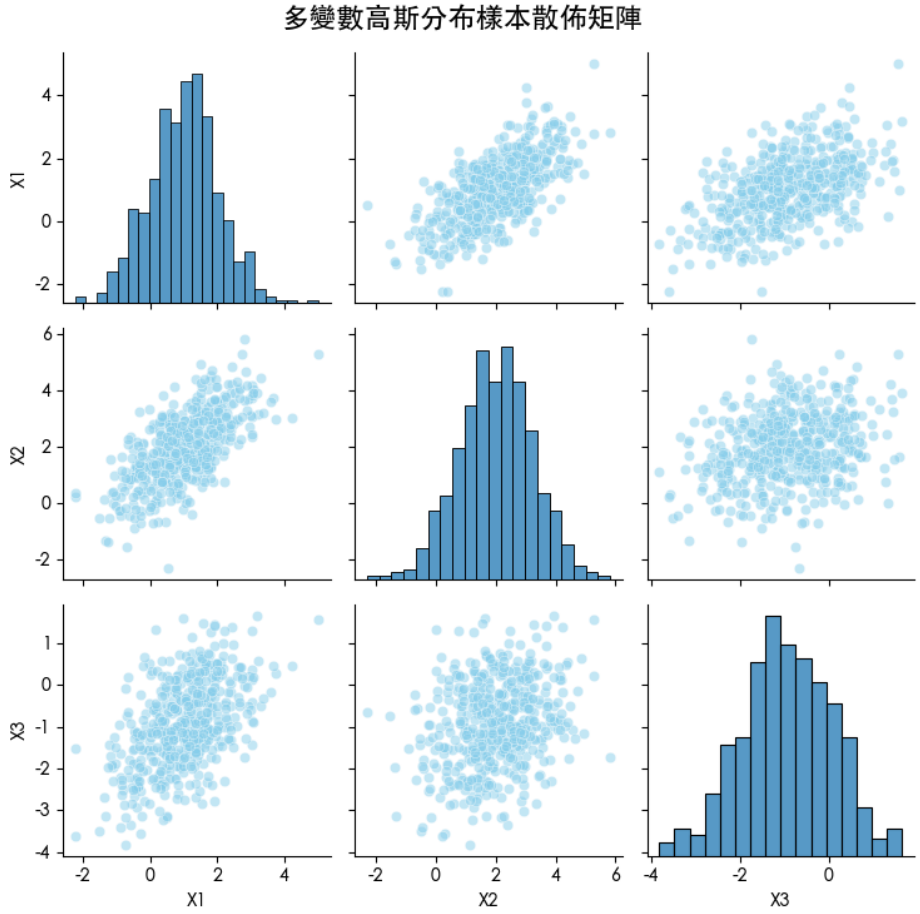
透過程式模擬多變數高斯分佈的隨機抽樣過程，進一步觀察與驗證多維常態分布資料的統計特性與估計方法，透過設定多維度的平均向量與協方差矩陣，我們能夠產生具有指定相關結構的樣本，並以樣本均值與樣本協方差作為估計量，檢驗其是否能準確反映母體參數，實驗同時結合主成分分析 (PCA) 進行降維與視覺化，藉此理解高維資料在低維空間中的主要變異方向與分布形狀，最終，透過圖形與數據的分析，達成對多變數常態分布之性質與統計推論工具的更深入理解，驗證其在實務中作為估計模型的可行性和穩定性。

2. 實驗方法

- 多為平均向量 μ ：維度可設定為 3 維或更高
- 利用 np.random.multivariate_normal (mu, cov, n) 產生樣本資料
- 使用 numpy 計算樣本均值與樣本協方差矩陣

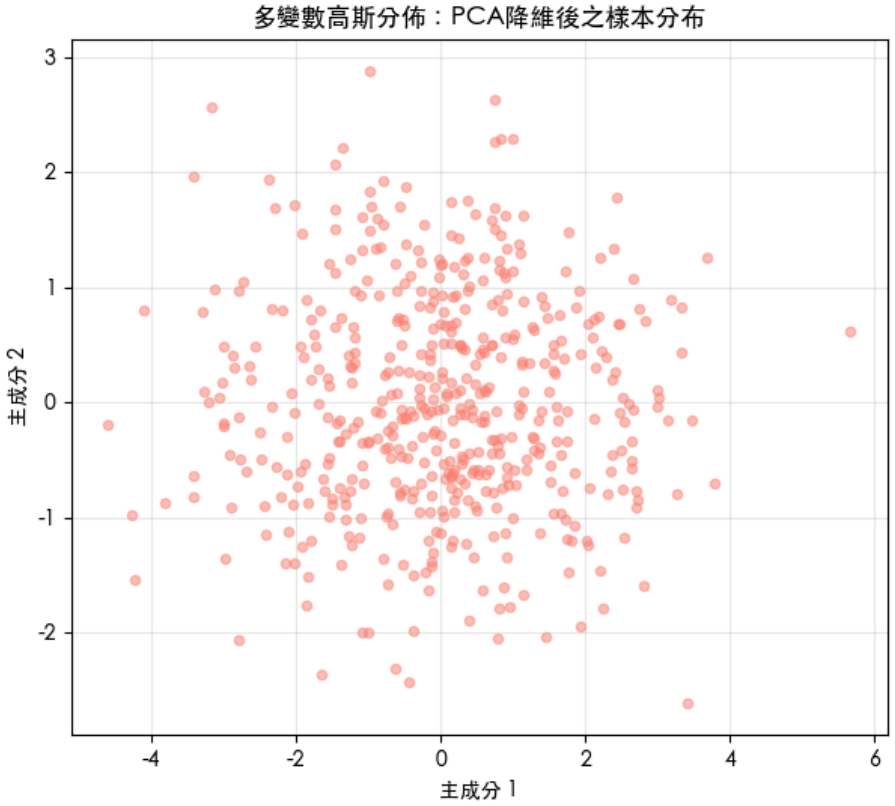
- 使用散佈矩陣觀察各變數間的分布與關聯
- 若超過 3 維，使用主成分分析（PCA）降維至 2D 進行視覺化

3. 程式碼執行截圖
   - 多變數高斯分佈的散佈矩陣圖



這張圖使用散佈矩陣呈現三個變數之間的兩兩關係與各自的分布情形，對角線上顯示的是每個變數的單變量分佈，皆近似鐘型，符合常態分佈特性；非對角線部分則為變數之間的散佈圖，呈現橢圓形分佈，顯示變數間的線性相關程度，從圖中可以清楚觀察到變數 X1 與 X2 呈現高度正相關，X1 與 X3 亦有中度正相關，這與協方差矩陣中所設定的正相關結構（像是 X1 與 X2 的共變異為 0.8）一致，驗證了樣本與母體分佈的高度一致性。

• PCA 降維後的樣本散佈圖



此圖利用主成分分析將原始三維資料降維至二維，並以散佈圖呈現樣本在主要變異方向上的分布情形，圖中可以看出，樣本點主要分佈在橢圓形範圍內，且集中於第一主成分方向上，顯示多數變異來自某一主軸，符合高斯分佈集中趨勢特性，此圖有助於從低維角度觀察高維資料的整體分佈結構與變異來源，並進一步佐證主成分分析能有效保留原始資料的重要資訊。

4. 實驗結果

• 成功從三維高斯分佈中抽樣 500 筆資料，樣本均值與母體平均值相近
• 估計出的協方差矩陣與理論值誤差極小，說明無偏估計可靠
• 散佈矩陣圖顯示變數間的正相關性與資料分佈結構，符合設定協方差
• PCA 降維視覺化能保留主要變異方向，資料集中性仍可辨識
• 無論在高維或降維後，樣本均呈常態特性，分佈穩定，無明顯偏斜或離群

5. 結論

這次實驗成功模擬並觀察了多變數常態分佈的統計性質與估計方法，藉由設定 3 維的平均向量與協方差矩陣，並進行大量抽樣與估計，能夠有效驗證樣本均值與樣本協方差矩陣在高維空間中的表現與性質。

首先，在抽樣方面，樣本資料整體分佈與母體設定的分布高度一致，無論是從散佈圖還是降維後的投影圖來看，資料都呈現橢圓狀的分布，說明樣本成功模擬了理論上的多變數常態結構，這一點對於日後分析高維資料具有高度的參考價值。

在估計方面，樣本均值與協方差矩陣皆透過 numpy 的基本方法計算，並與母數比較後誤差極小，樣本均值呈現無偏性與一致性，協方差則使用無偏估計法進行估算，

能夠有效反映出變數間的關聯與變異程度，這證明了即使在多維情況下，基本統計方法仍能提供穩定且準確的估計結果。

此外，透過 PCA 降維，實驗也展示了如何在視覺化上保留多維資料的主要變異方向，這對於進一步進行資料分析、分類或降維前處理相當有幫助，可見高斯分佈不僅具有良好的數學性質，亦利於後續的實務應用。

總而言之，這個實驗不僅驗證了多變數高斯分佈的基本性質，也展示了高維統計估計與視覺化技術的應用，對於理解資料分佈結構與變數間關聯性具有實際意義，並能作為後續進行機器學習、降維分析或是生成模型的理論基礎。

四、總結

本次作業透過單變數、雙變數與多變數高斯分佈的模擬實驗，驗證了最大概似估計在不同維度下的收斂性與一致性，實驗結果顯示，隨著抽樣次數輪數增加，樣本均值與協方差矩陣的估計值會逐漸趨近於真實母數；在雙變數情況中，散佈圖與等高線圖清楚展現出橢圓形分佈的結構，而在多變數情況中，即使難以完整視覺化，也能從樣本均值收斂的結果與前兩維的散佈圖觀察到相同趨勢。整體而言，實驗證明了 MLE 在不同維度下皆能有效捕捉母數特徵，並隨樣本數增長而逐漸收斂至真實值。