Stochastic Process Final Project

Testing robustness of centrel limit theroy

Wu, Tung Cheng

一、簡介

中央極限定理有許多不同的版本，根據 wikipedia 的說明：「大量互相獨立的隨機變數之均值，經適當的標準化後，依分布收斂於常態。」[1]。這個定義本身有許多可以延伸的問題，例如：

1. 何謂「適當的標準化」?
2. 任意的隨機變數都可以成立嗎？
3. 隨機變數的均質可以成立，那其他的描述統計量 (最大值、標準差…) 可以成立嗎？
4. 是否有一個定值，可以稱為大量？

本文將藉由統計語言 R 實作，嘗試回答上述問題，並以圖像呈現觀察到的結果。

二、背景回顧

最初始的中央極限定理版本，是法國數學家 Abraham de Moivre 於 1738 年在其著作 The Doctrine of Chances 中所提出 [2]。說明進行無限次獨立白努利測試 (Bernoulli trials) 所形成的二項分布，最終其機率質量函數 (probability mass function) 會近似於平均數為 np、標準差為 $\sqrt{np(1-p)}$ 的常態分布，這個發現於 1810 年由 Pierre-Simon Laplace 使用特徵函數的方式證明 [3]。其後的數學家們，如：Poisson, Cauchy, Dirichlet, Lindeberg 也都對這個理論，提出擴展與相關的證明，使得後世的統計學得以對樣本有常態如此方便的假設。

三、驗證問題

(一) 常態檢測工具

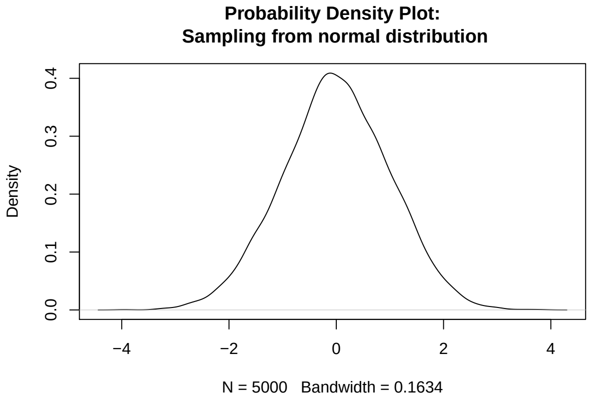
我們要檢驗一個分布是否符合常態分布，因此需要檢驗的方法。在統計檢定上有許多檢測常態的方法，如 Kolmogorov-Smirnov test、The Anderson-Darling test、Shapiro-Wilks test。在 Asghar Ghasemi 與 Saleh Zahediasl 的研究中 [4]，指出 Shapiro-Wilk test 是最穩定的方法，因此本文之常態檢定，將使用 Shapiro-Wilk test 進行。我們可以藉由檢定結果中的 p-value 判斷樣本是否為常態，若 $\text{p-value} > 0.05$ ，表示樣本為常態分布。下面提供一個範例：

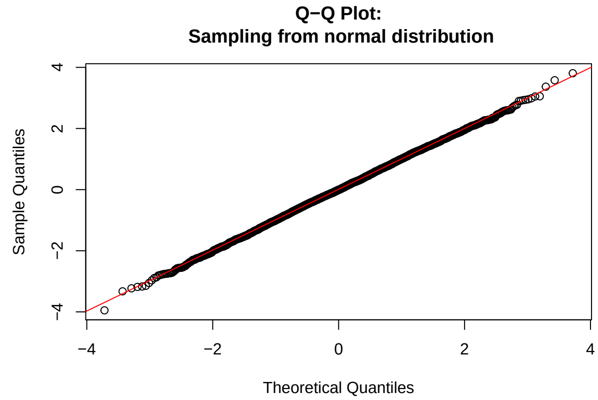
```
##
##  Shapiro-Wilk normality test
##
## data:  norm_dta
## W = 0.99973, p-value = 0.7858
```

再提供一個樣本為 Poisson 的範例 (即不符合常態分布)

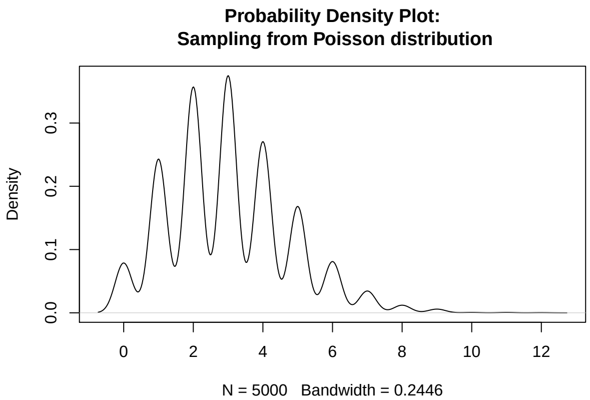
```
##
##  Shapiro-Wilk normality test
##
## data:  poi_dta
## W = 0.9492, p-value < 2.2e-16
```

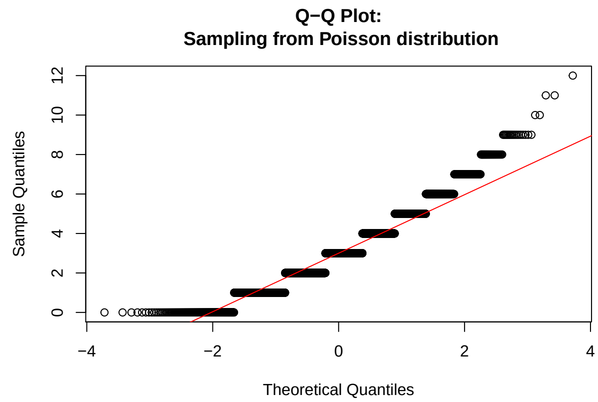
除了 Shapiro-Wilk test 外，在實務上除了繪製機率密度函數圖，也常繪製 Q–Q (quantile-quantile) plot，將手上之樣本與常態資料比較。因此在呈現 Shapiro-Wilk 檢定結果的同時，本文也將呈現 Q-Q plot。若 Q-Q plot 上的資料點落在紅線上，則表示樣本符合常態分布，上述符合常態配分之樣本的 Q-Q plot 如下：





樣本為 Poisson 的 Q-Q plot 如下：





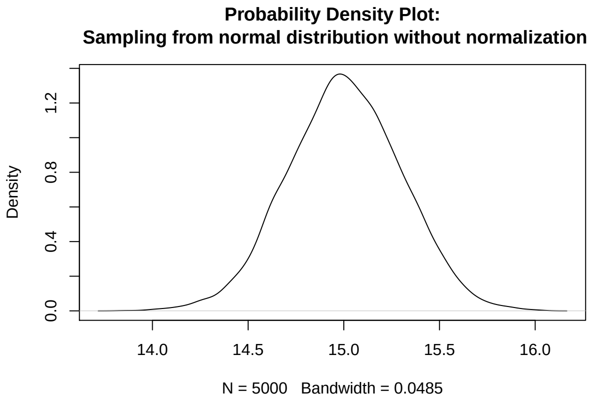
(二) 適當的標準化

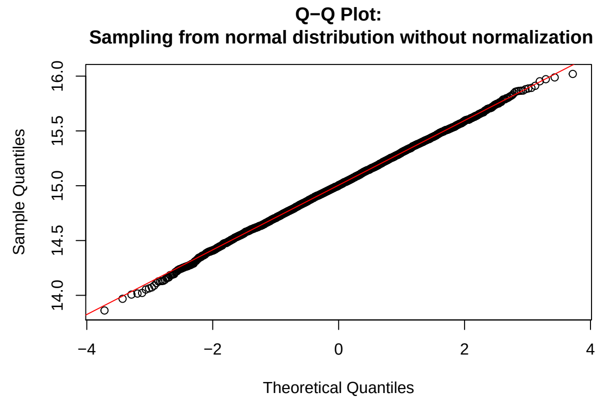
首先，我想先討論「適當的標準化」，因為會影響後續的實驗流程。測試：

1. 不標準化
2. $\frac{(x-\mu_x)}{\sigma_x}$

我們使用 R 的 function 製作 $N(15,21)$ 的常態分布 5000 個，並取其各自的平均數。

```
##
##  Shapiro-Wilk normality test
##
## data:  mean_dta
## W = 0.9997, p-value = 0.6973
```

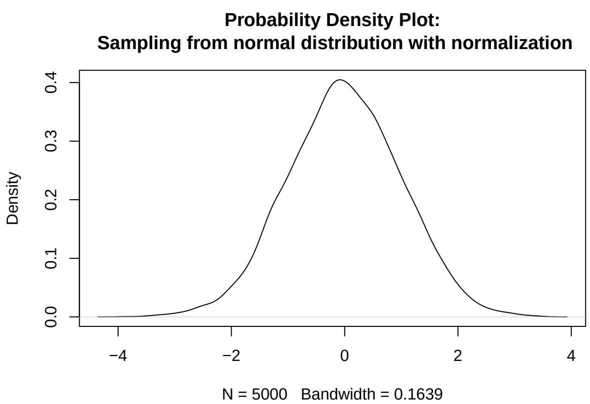


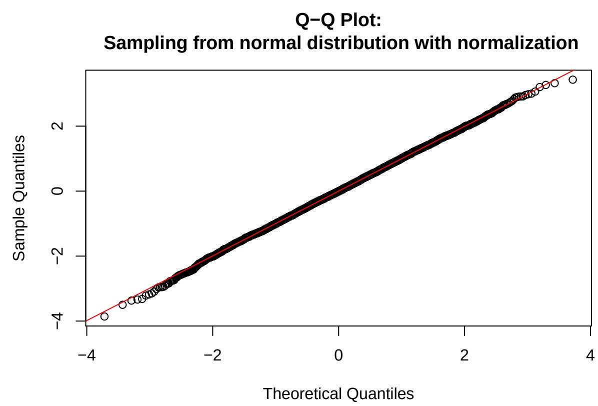


我們針對尚未標準化的樣本進行檢驗，從 Shapiro-Wilk test 的結果，我們可以看到 $p > 0.05$。從 Q-Q plot 中也可以觀察到，數據點幾乎都在紅線上，表示與常態分布相似。這與 wiki 上的說明不相符，但根據老師上課的說明，經過標準化的樣本是符合標準常態分布。

接著我們來看，經過 $\frac{(x-\mu_x)}{\sigma_x}$ 標準化的結果

```
##
##  Shapiro-Wilk normality test
##
## data:  dta.scale
## W = 0.9997, p-value = 0.6973
```





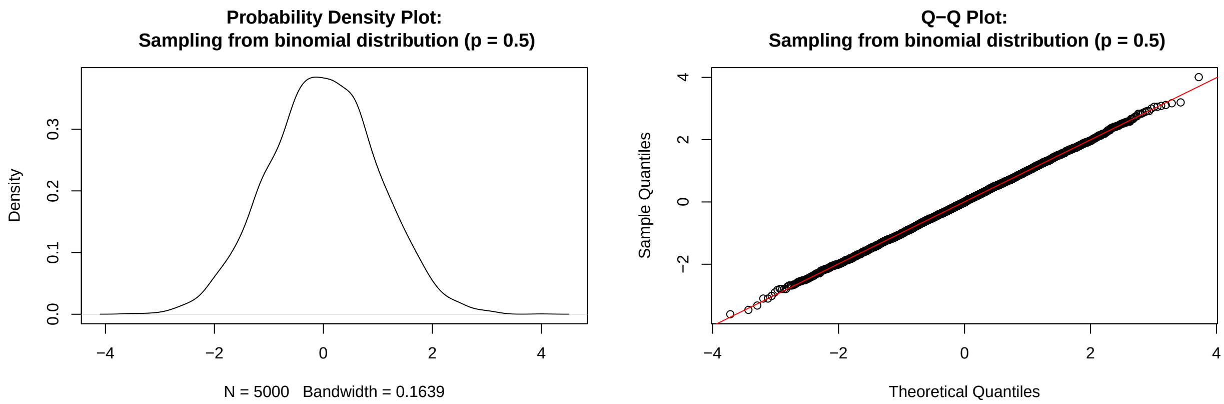
從檢定的結果上來說並沒有差異，但觀察兩張圖的縱軸，可以發現數值與刻度寬度不同，表示兩樣本的平均與標準差不同。總結來說，就我們的實驗結果顯示，不標準化不影響是否為常態分布，但影響最後收斂於哪個常態。為避免有本實驗未觀察到的情況，往後實驗仍會進行標準化。

(三) 任意的隨機變數

接著我們討論不同隨機變數的平均數，是否也會收斂於常態分布，以下測試一些上課提過的分布：

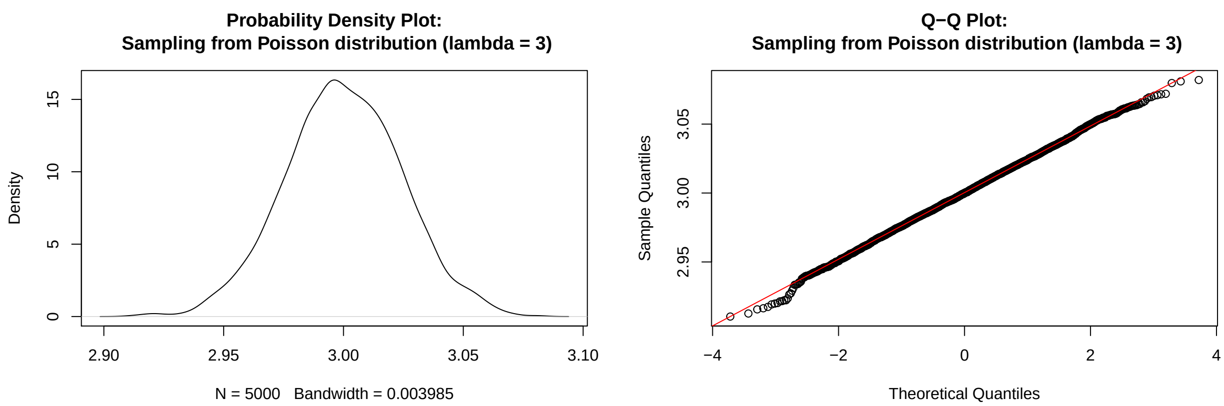
1. Binomial distribution (set probability = 0.5)

```
##
##  Shapiro-Wilk normality test
##
## data:  mean_dta.binom
## W = 0.99977, p-value = 0.8821
```



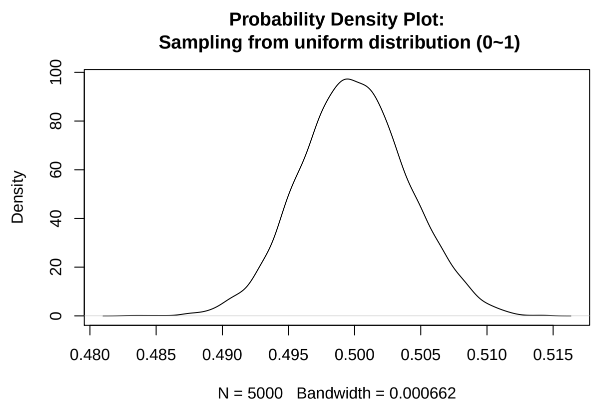
2. Poisson distribution (set $\lambda = 3$)

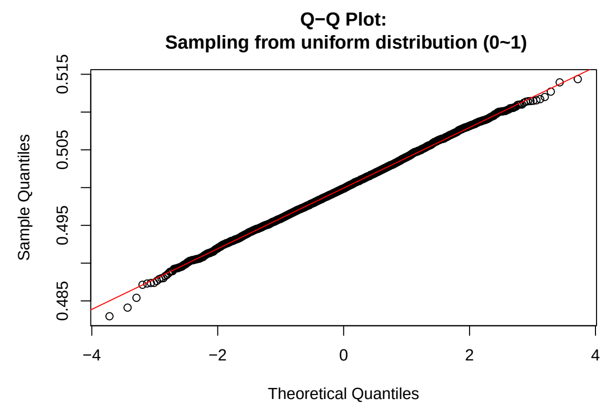
```
##
##  Shapiro-Wilk normality test
##
## data:  mean_dta.poi
## W = 0.99947, p-value = 0.1668
```



3. Uniform distribution (set Uniform distribution $\in [0, 1]$)

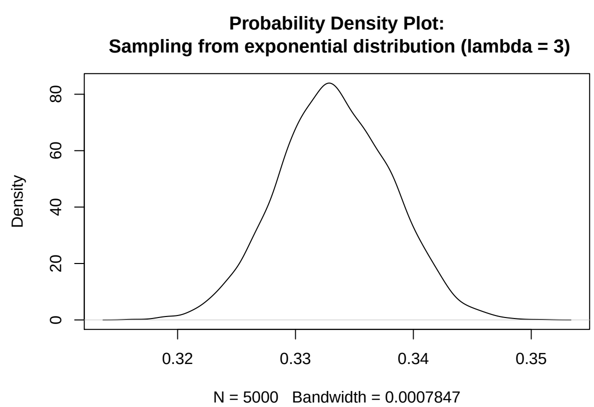
```
##
##  Shapiro-Wilk normality test
##
## data:  mean_dta.uni
## W = 0.9995, p-value = 0.2066
```

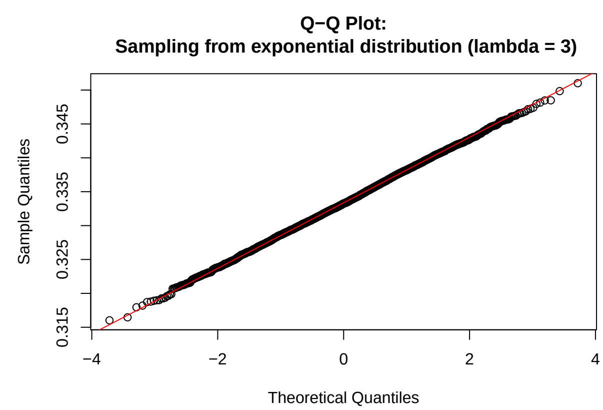




4. Exponential distribution (set $\lambda = 3$)

```
##
##  Shapiro-Wilk normality test
##
## data:  mean_dta.exp
## W = 0.99972, p-value = 0.7763
```





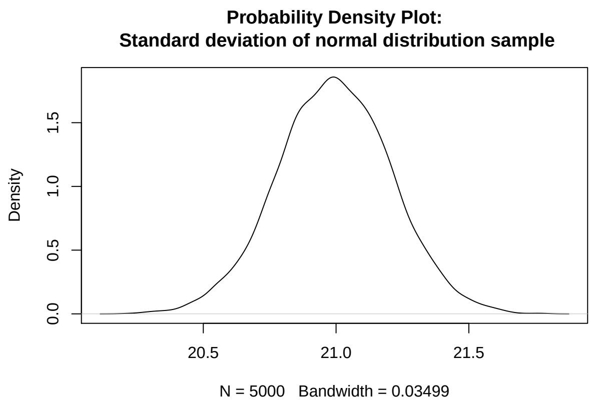
由上述的檢定與 Q-Q plot 顯示，隨機變數的種類，不影響其平均值是否收斂於常態分布。總結來說，在上課提到的四個隨機變數，經過測試後發現 5000 個樣本平均值，會收斂於常態分布。

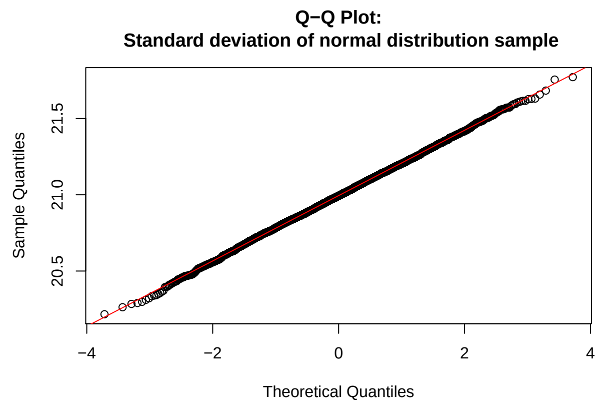
(四) 其他的描述統計量

Wikipedia 說明的中央極限定理，隨機變數的平均數會收斂於常態分布，以下測試其他的描述統計量 (ex: 標準差、最大值、最小值、偏態 (skewness, 三階動差)、峰度 (kurtosis, 四階動差))，使用前述的 $N(15, 21)$ 分布進行測試。

1. 標準差

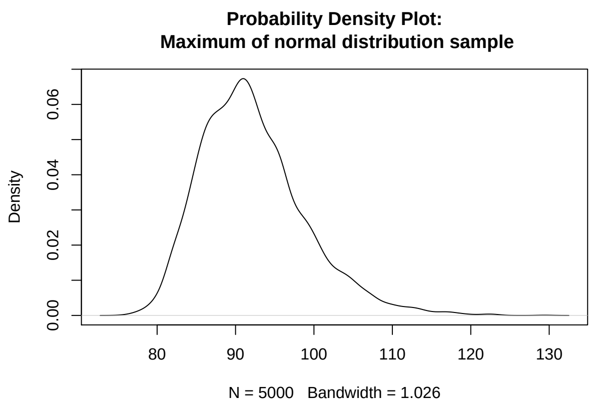
```
##
##  Shapiro-Wilk normality test
##
## data:  sd_dta
## W = 0.99974, p-value = 0.8158
```

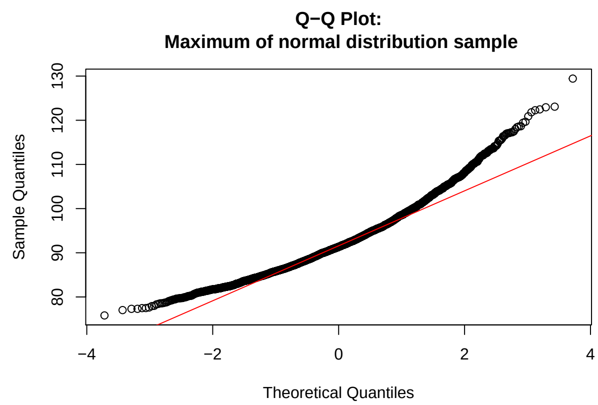




2. 最大值

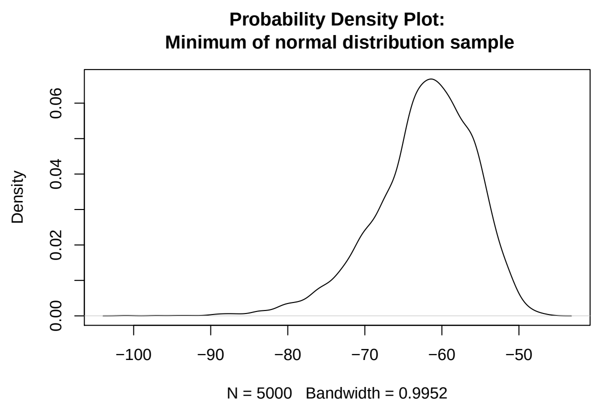
```
##
##  Shapiro-Wilk normality test
##
## data:  max_dta
## W = 0.9623, p-value < 2.2e-16
```

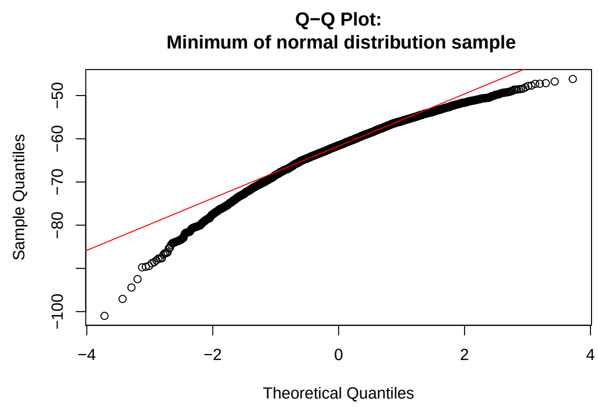




3. 最小值

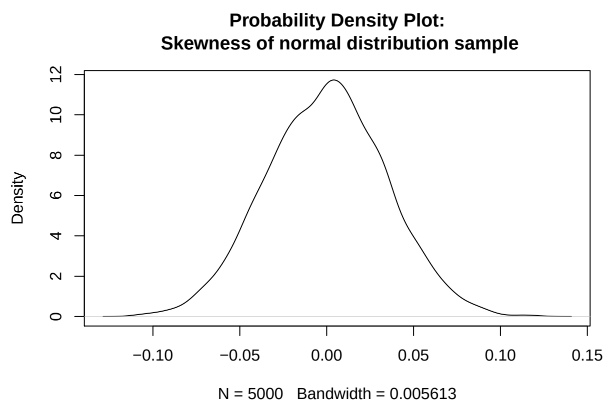
```
##
##  Shapiro-Wilk normality test
##
## data:  min_dta
## W = 0.96835, p-value < 2.2e-16
```

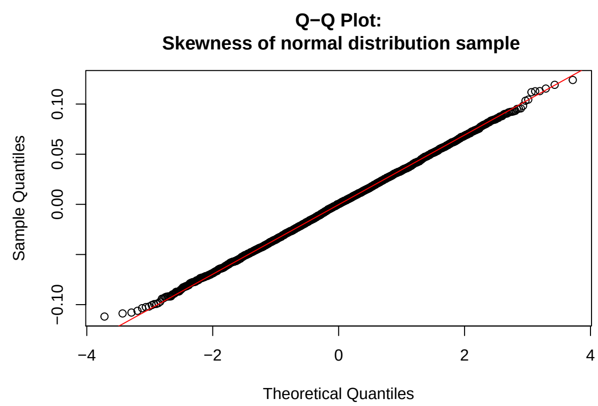




4. 偏態

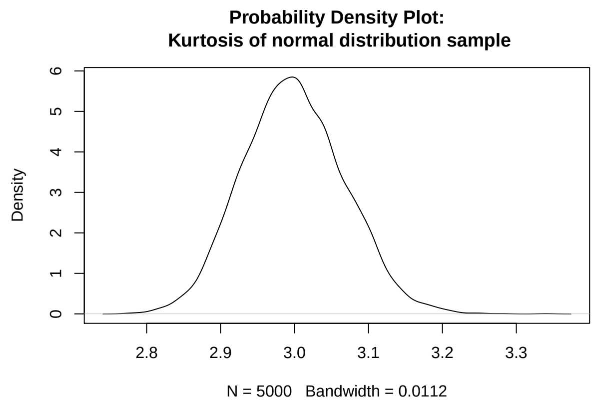
```
##
##  Shapiro-Wilk normality test
##
## data:  skew_dta
## W = 0.99971, p-value = 0.7493
```

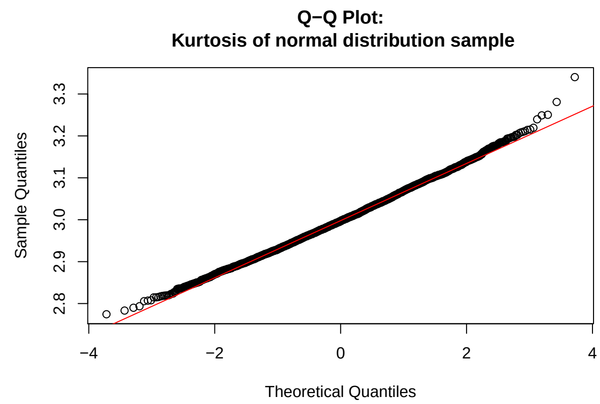




5. 峰度

```
##
##  Shapiro-Wilk normality test
##
## data:  kurt_dta
## W = 0.99798, p-value = 4.202e-06
```





透過上面的測試，我們發現改成其他描述統計量時，只有標準差與偏態 (即二階與三階動差) 會收斂於常態。提出以下兩種可能性：

1. 所有的描述統計量都會收斂於常態分布，只是目前 5000 個隨機變數的描述統計量，可能不夠「大量」，因此沒有收斂於常態分布
2. 只有動差類的描述統計量，會收斂於常態分布，隨著階層越高，需要的「大量」也越大，因此本次測試的峰度並沒有收斂。

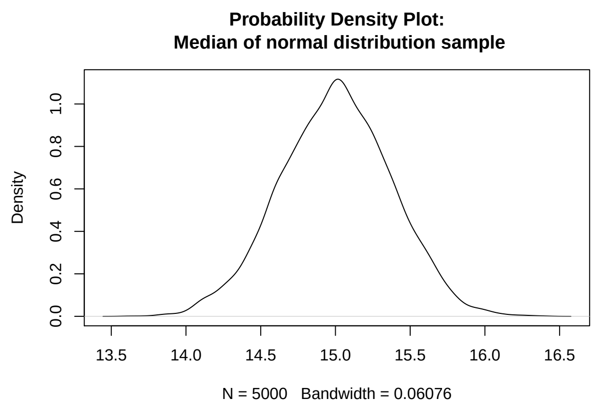
上述的可能性可以歸納成兩個測試項目：

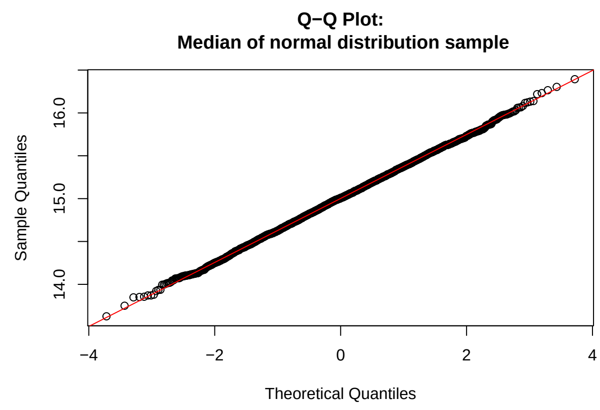
1. 增加隨機變數的量

然而本文中使用的 Shapiro-Wilk 最多接受 5000 筆資料。但更換檢定方法，無法確定效果來自「更換檢定方法」還是「增加樣本量」。由於時間因素，本文增加樣本數繼續測試。

2. 使用其他描述統計量測試 (下方使用中位數測試)

```
##
##  Shapiro-Wilk normality test
##
## data:  median_dta
## W = 0.99972, p-value = 0.7681
```





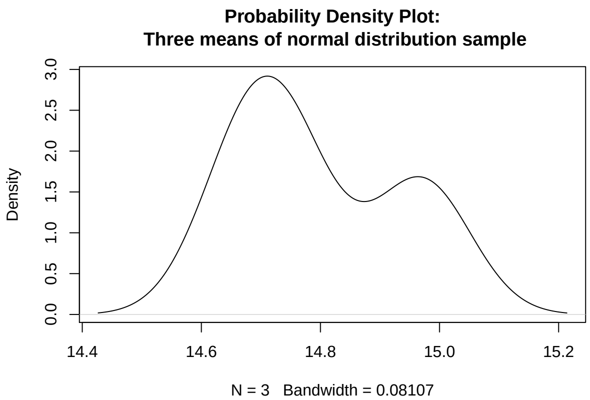
使用中位數測試後發現，並非只有動差類的描述統計量，會收斂於常態分布。整體來說，更改樣本的描述統計值時，經過測試結果發現「隨機變數的中位數、標準差、偏態，會收斂於常態分布」。其他的描述統計量，由於 Shapiro-Wilk test 的限制，不確定是不會收斂，還是樣本不夠大量，因此沒有收斂於常態分配。此處的隨機變數是常態分布，並未測試其他分布的情況。

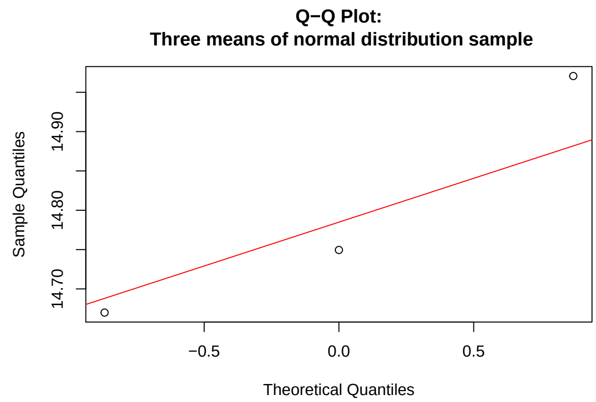
(五) 大量有多大

最後測試需要多少樣本數才能稱為大量，由於在極小樣本時，p-value 的參考價值有限，故本小節主要觀察機率密度函數圖與 Q-Q plot。以下測試不同數量隨機變數的平均數，會收斂於常態分布。

1. $N = 3$

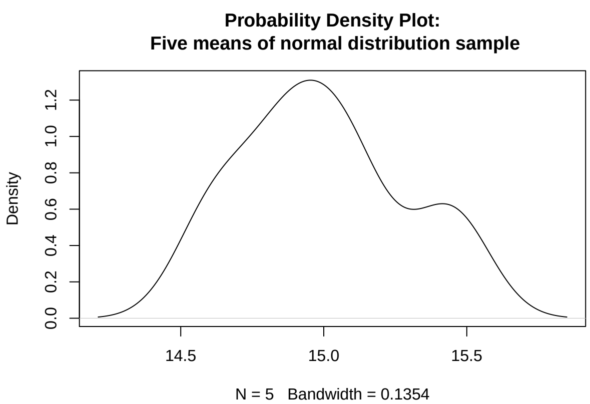
```
##
##   Shapiro-Wilk normality test
##
## data:  N3_dta
## W = 0.93123, p-value = 0.4932
```

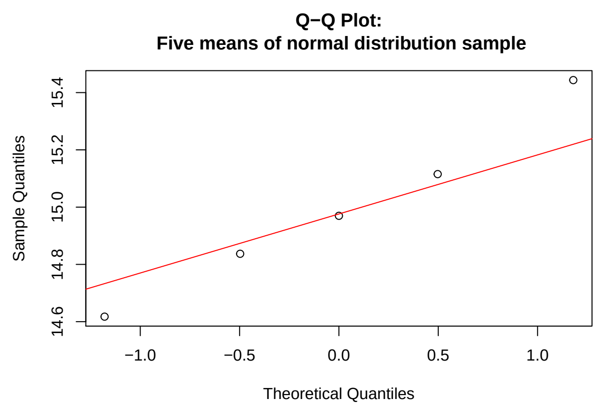




2. $N = 5$

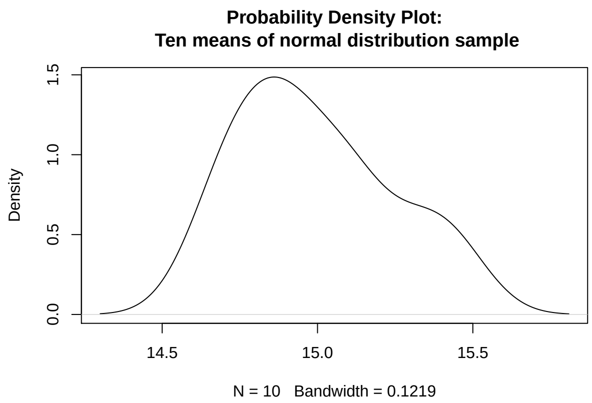
```
##
##   Shapiro-Wilk normality test
##
## data:  N5_dta
## W = 0.98938, p-value = 0.9775
```

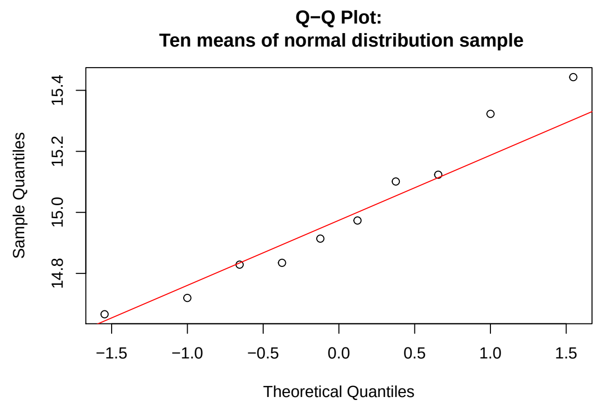




3. $N = 10$

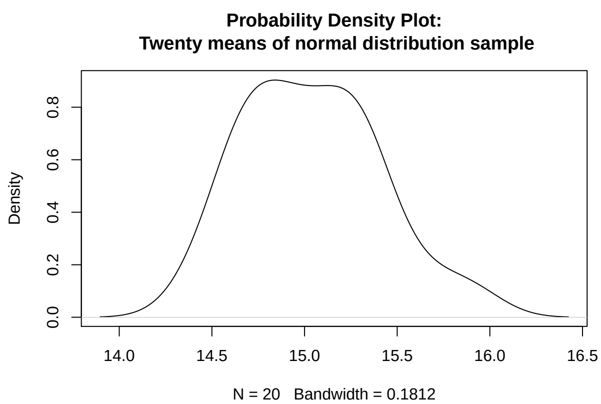
```
##
##   Shapiro-Wilk normality test
##
## data:  N10_dta
## W = 0.94986, p-value = 0.6669
```

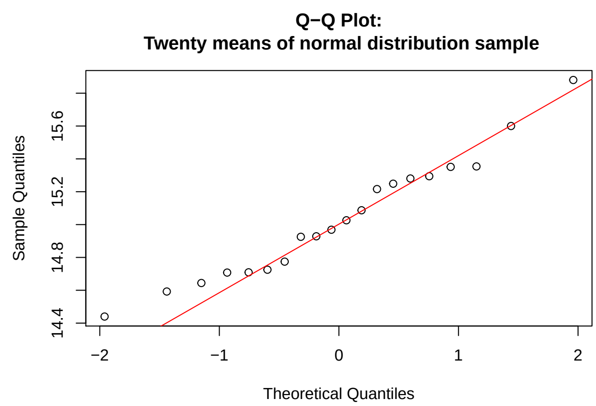




4. $N = 20$

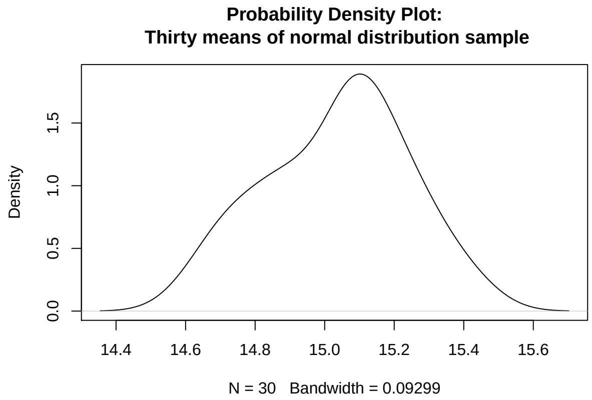
```
##
##  Shapiro-Wilk normality test
##
## data:  N20_dta
## W = 0.96685, p-value = 0.6874
```

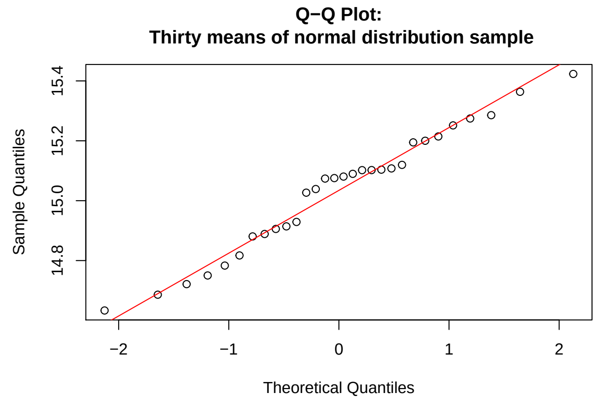




5. $N = 30$

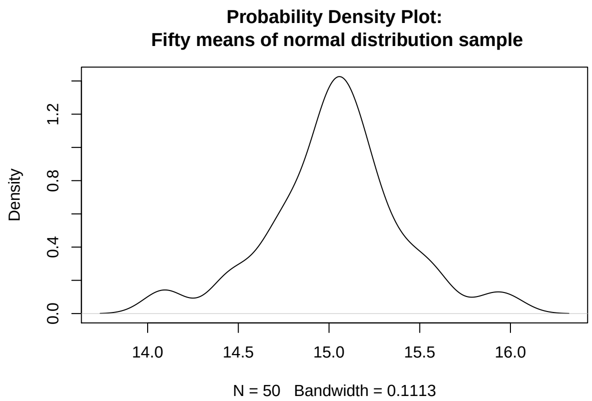
```
##
##  Shapiro-Wilk normality test
##
## data:  N30_dta
## W = 0.97183, p-value = 0.5903
```

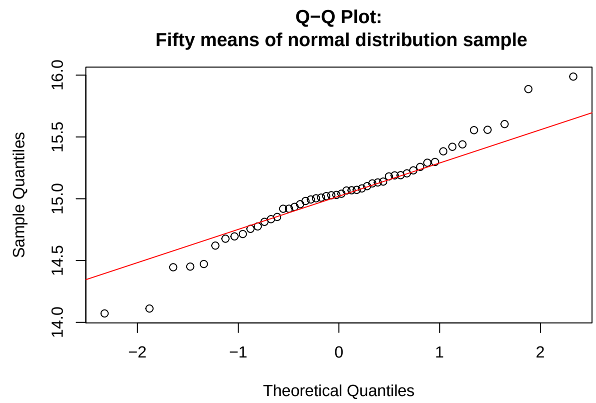




5. $N = 50$

```
##
##  Shapiro-Wilk normality test
##
## data:  N50_dta
## W = 0.97185, p-value = 0.2749
```





四、總結

經過本文的一系列實驗，關於「大量互相獨立的隨機變數之均值，經適當的標準化後，依分布收斂於常態。」我們提出了四個問題並發現，中央極限定理比想像中的還要穩定。不論是替換隨機變數，或是計算不同的隨機變數描述統計值，甚至是沒有標準化，觀察到結果依然收斂於常態分布。

不過本文也有些不完整之處，個人認為最明顯的問題在於檢定常態的方法。本文使用統計檢定的方式，並直接使用 p-value 來判斷結果，這其實是一個蠻危險的作法。p-value 只是眾多指標中的一個指標，由於使用簡單與教學方便，因此在許多檢定上常被使用，但不表示 p-value 可以當作唯一的指標。(ex: 重複 100 次實驗，可能會得到 94 次 $p\text{-value} < 0.05$ 的結果，當觀察到另外的 6 次時，直接下結論就錯了。)

Reference

[1] wikipedia, "中央極限定理 - wikipedia, the free encyclopedia." 2018.

[2] A. de Moivre, *The doctrine of chances*. Woodfall, 1738, pp. 235–243.

[3] H. Fischer, “The central limit theorem from laplace to cauchy: Changes in stochastic objectives and in analytical methods,” in *A history of the central limit theorem: From classical to modern probability theory*, New York, NY: Springer New York, 2011, pp. 17–74.

[4] A. Ghasemi and S. Zahediasl, “Normality tests for statistical analysis: A guide for non-statisticians,” *International journal of endocrinology and metabolism*, vol. 10, no. 2, p. 486, 2012.