分佈模型 - 圖形解讀的重要性

周芊妤

January 9, 2024

本章深入探討了在統計學中不可或缺的分佈模型，這些模型扮演著描述隨機變量的關鍵角色，區分為離散型和連續型的分佈模型用於描述不同取值方式的隨機變量，並著重於透過其圖形來深入理解其特性，如峰度和偏度。這些圖形為我們提供直觀、清晰的信息，使我們能更好地理解研究的隨機變量行為，強調了模型的圖形分佈對於統計學的重要性，並強調這些圖形在進行統計推斷和預測時的關鍵作用，只有深刻理解並善用這些圖形，我們才能在統計學領域取得更加優異的成果。

1 分佈模型的介紹

分佈模型在統計學中扮演著不可或缺的角色，它們是描述隨機變量的機率模型。這些模型被區分為離散型和連續型，分別用於描述具有離散和連續取值的隨機變量。這些分佈模型的圖形更是不可或缺的工具，透過這些圖形，我們可以深入探討其特性，包括峰度和偏度，這些特性能告訴我們關於分佈的形狀是右偏、左偏，還是對稱。換句話說，模型的圖形可以提供直觀且清晰的信息，讓我們更好地理解所研究的隨機變量行為，重視模型的圖形分佈是至關重要的。

在分佈中，分成了離散型分佈 (Discrete distribution) 以及連續型分佈 (Continuous distribution) 兩種。

1. 離散型分佈：

   適用於描述具有可數個可能取值的隨機變量，例如次數、計數等。其機率質量函數（PMF）描繪了每一個可能取值的機率，即隨機變量等於特定值的機率。

2. 連續型分佈：

   適用於描述具有連續可能取值範圍的隨機變量，即可能的取值是一個連續的區間，例如長度、時間、重量等。

   其機率密度函數（PDF）來描述變量的機率密度，非具體取值的機率。

2 離散型分佈 (Discrete distribution)

離散分佈即分佈函數的值域是離散的，比如只取整數值的隨機變數即屬於離散分佈。$F(x)$ 表示隨機變數 $X \leq x$ 的機率值，如果 X 的取值只有 $x_1 < x_2 < \cdots < x_n$ ，則：

1. $F_X(x_i) = \sum_{j=1}^{i} P(x_j)$
2. $\sum_{k=1}^{n} P(x_k) = 1$

2.1 Bernoulli distribution

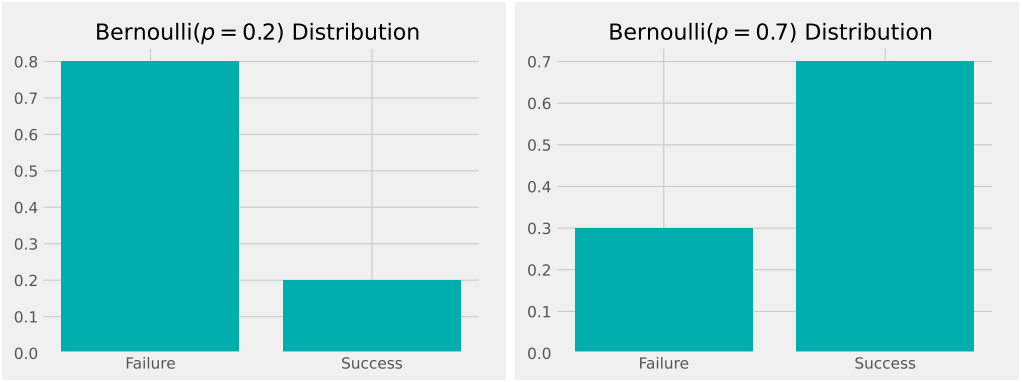
柏努力分佈 (Bernoulli distribution) 是一種描述在單次隨機試驗中成功或失敗的機率分佈。它具有以下特色 ：

1. 每次試驗只有兩種可能的結果 : 成功（success）和失敗（failure）。
2. 成功的機率為 p ，失敗的機率則是 $1-p$ 。
3. 其機率質量函數（PMF）為

$$f_x(x) = p^x(1-p)^{1-x},\ x = 0, 1$$

4. 期望值 $E(X) = p$ 、變異數 $Var(X) = p(1-p)$ 。

在圖 **??** 中我們可以看到 Bernoulli($p = 0.2$) 以及 Bernoulli($p = 0.7$) 的 PMF 差異，它們分別在在成功與失敗中的機率為 $0.2, 0.8$ 以及 $0.7, 0.3$ ，兩張圖形兩邊高度 (機率) 各不相同。



(a) Bernoulli($p = 0.2$)　　(b) Bernoulli($p = 0.7$)

圖 1: 柏努力分佈 : $p < 0.5$ 以及 $p > 0.5$ 的比較

2.2 Binomial distribution

二項分佈（Binomial Distribution）是描述在一系列相互獨立、相同機率的伯努利試驗中成功次數的機率分佈。以下是二項分佈的特性 ：

1. 進行一系列 Bernoulli 試驗，每次試驗都具有相同的成功機率 p ，且 n 次試驗之間相互獨立。
2. 每次試驗只有兩個可能的結果，成功（success）和失敗（failure）。
3. 隨機變數 $X = 0,\ 1,\ 2, \cdots$ 表示在 n 次試驗中成功的次數。
4. 其機率質量函數（PMF）為 :

$$f_x(x) = \begin{pmatrix} n \\ x \end{pmatrix} p^x (1-p)^{n-x},\ x = 0,\ 1,\ 2, \cdots$$

5. 期望值 $E(X) = np$ 、變異數 $Var(X) = np(1-p)$ 。

圖 **??** (a) 為 Binomial$(100, 0.5)$ 的 PMF 圖，我們可以看出此圖形具有以下特性 :

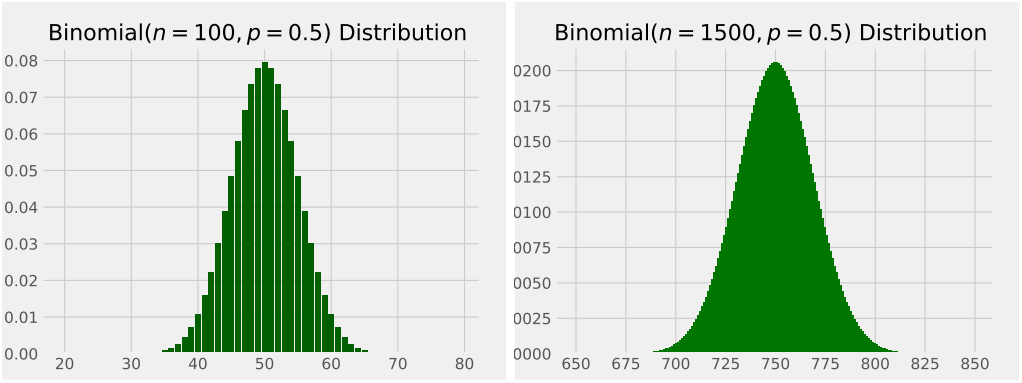
1. 圖形為鍾形。
2. 單峰性 ： 當成功機率 p 不等於 0 或 1 時，PMF 呈現單峰分佈，即在對稱中心 $\mu = np$ 附近具有最大的機率。
3. 對稱性 ： 當成功機率 $p = 0.5$ 時，二項分佈呈現對稱分佈，即以試驗次數的一半為中心對稱。

當我們在 p 不動的情況下增加樣本數 n 時，可以看到圖 **??** (b) 為 Binomial$(1500, 0.5)$ 的 PMF 圖，此圖形更加趨近於常態分配，我們可以看出此圖形具有以下特性 :

1. 隨著樣本數 n 的增加，二項分佈可以近似為常態分佈的常態近似效果更好。
2. 隨著樣本數 n 的增加，PMF 將變得更加窄銳，並趨向於常態分佈。

在圖 **??** 中我們可以看出當我們在固定樣本數 $n = 100$ 的情況下比較不同機率 $p = 0, 0.1, 0.2, ..., 1$ 時，我們會發現此圖形具有以下特性 :

1. 隨著 p 的變化，PMF 會改變 :
   當成功機率 p 變化時，PMF 的形狀也會相應地變化，這是因為圖形會對稱於中心 $\mu = np$ ，較大的 p 將使分佈向右偏移 (因為 $\mu = np$ 會更大)，較小的 p 則會使分佈向左偏移 (因為 $\mu = np$ 會更小)。



(a) Binomial$(n = 100, p = 0.5)$ (b) Binomial$(n = 1500, p = 0.5)$

圖 2: 二項分佈：調整樣本數 n 的大小做觀察

2. 當 p 為 0 或 1 時，PMF 在該處的面積即為 1 。

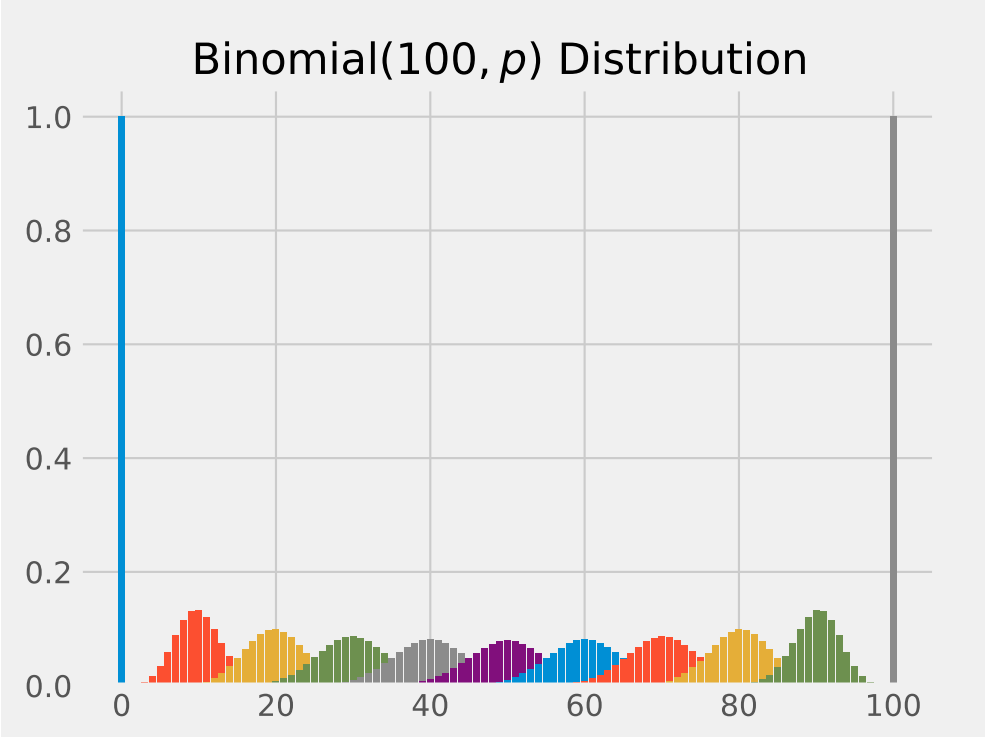


圖 3: 二項分佈：固定樣本數 $n = 100$ ，$p = 0, 0.1, 0.2, ..., 1$

2.3 Geometric distribution

幾何分佈（Geometric Distribution）是描述在一系列獨立伯努利試驗中，首次取得成功所需的試驗次數的機率分佈，以下是幾何分佈的特點 ：

1. 進行一系列 Bernoulli 試驗，每次試驗都具有相同的成功機率 p ，且試驗之間相互獨立。

2. 每次試驗只有兩個可能的結果，成功（success）和失敗（failure）。

3. 隨機變數 $X = 1,\ 2, \cdots$ 表示在 n 次試驗中成功的次數。

4. 其機率質量函數（PMF）為：

$$f_x(x) = P(X = x) = (1-p)^{x-1}p,\ x = 1,\ 2, \cdots$$

5. 期望值 $E(X) = \frac{1}{p}$ 、變異數 $Var(X) = \frac{1-p}{p^2}$ 。

6. 遺失記憶性：

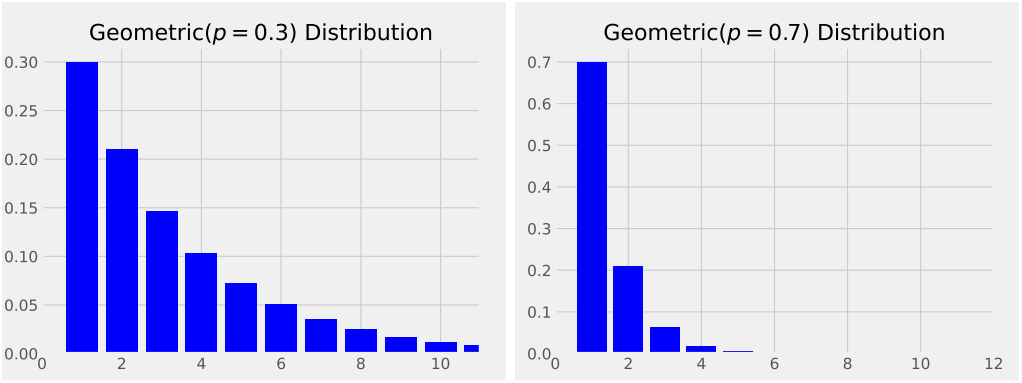
   對於任意正整數 m 和 n ， $P(X > m+n \mid X > m) = P(X > n)$。

由圖 **??** 中我們可以比較兩張圖，並看出此圖的特性：

1. 單峰性：PMF 呈現單峰分佈。

2. 指數下降：隨著試驗次數的增加，PMF 值呈指數下降的趨勢。

3. p 越大時下降越快，這是由於在 $f_x(x) = (1-p)^{x-1}p$ 中，如果 p 越大， $(1-p)$ 就會越小，隨之 $(1-p)^{x-1}$ 就會更快速的下降。



(a) Geometric($p = 0.3$)　　(b) Geometric($p = 0.7$)

圖 4: 幾何分佈：比較 $p = 0.3$ 與 $p = 0.7$ 之不同

2.4 Hypergeometric distribution

超幾何分佈（Hypergeometric Distribution）描述了在不放回的抽樣過程中，成功和失敗的隨機變數。以下是超幾何分佈的特點 ：

1. 試驗總體包含了 N 個元素，其中有 M 個成功和 $N - M$ 個失敗。

2. 每次抽樣 (都是相互獨立的) 後，抽取的元素不放回總體中。

3. 隨機變數 $X = 0,\ 1,\ 2, ...,\ n$ 表示在抽取 n 次後成功的次數。

4. 其機率質量函數（PMF）為：

$$P(X = x) = \frac{\begin{pmatrix} M \\ x \end{pmatrix} \begin{pmatrix} N - M \\ n - x \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}},\ x = 0,\ 1,\ 2, ...,\ n$$

5. 期望值 $E(X) = \frac{nM}{N}$ 、變異數 $Var(X) = \frac{nM(N-M)(N-n)}{N^2(N-1)}$ 。

圖 **??** 是 Hypergeometric$(150, 80, 35)$ 的圖形。

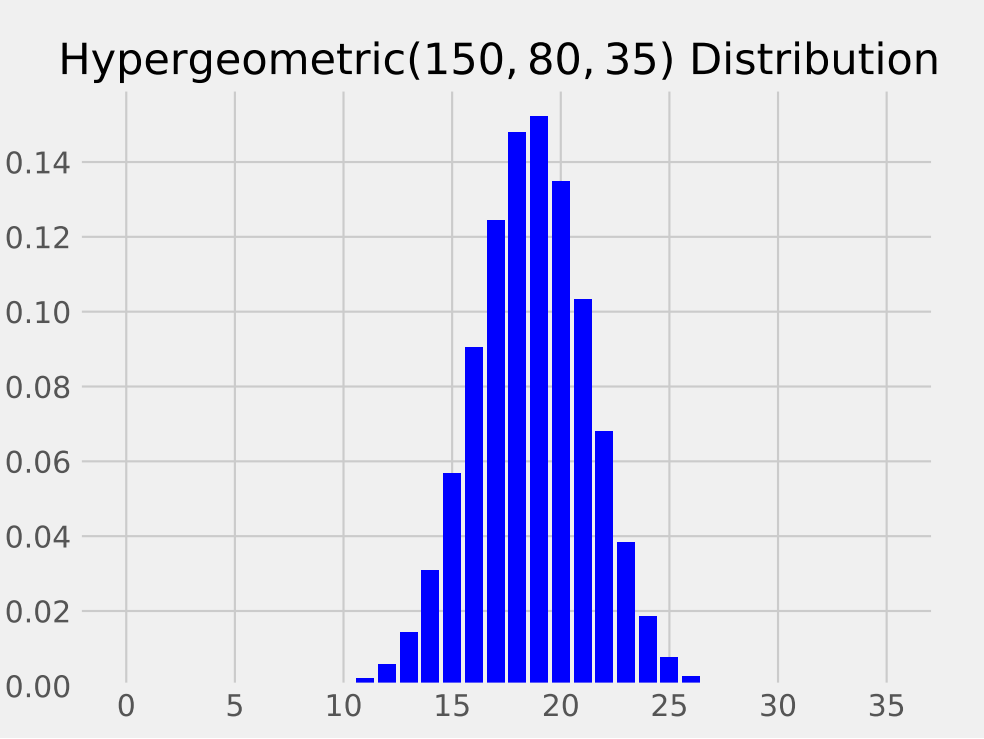


圖 5: Hypergeometric$(150, 80, 35)$

2.5 Nagative Binomial distribution

負二項分佈 (Nagative Binomial distribution) 描述了在進行一系列二元試驗（成功和失敗）中，直到達到指定數量的成功之前，所需的失敗次數。以下是負二項分佈的特點：

1. 每次試驗都是相互獨立的，且試驗只有兩種可能的結果，即成功或失敗。

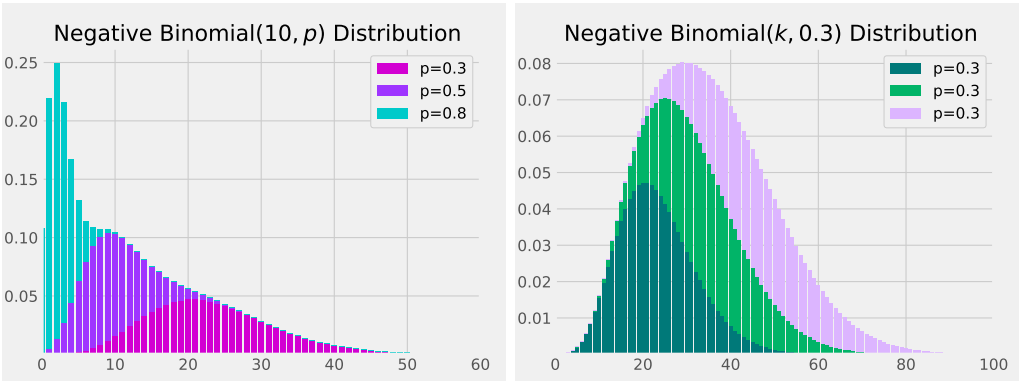
2. 隨機變數 $X = r,\ r + 1, \cdots,\ \infty$ 表示在達到 r 次成功之前所需的失敗次數。

3. 其機率質量函數（PMF）為：

$$f_x(x) = \begin{pmatrix} x-1 \\ r-1 \end{pmatrix} p^r (1-p)^{r-1}, \ x = r, \ r+1, \cdots, \ \infty$$

圖 **??** (a) 為 Negative Binomial$(10, p)$ ， $p = 0.3, 0.5, 0.8$ 的圖形。我們可以看出來，隨著 p 的增大，分佈將變得更為尖峭，即失敗次數少但成功次數多的情況更有可能發生。

圖 **??** (b) 為 Negative Binomial$(k, 0.3)$ ， $k = 10, 15, 20$ 的圖形。隨著 k 的增大，分佈的峰值位置會向右偏移，且變異性增加，表示需要更多的失敗次數才能達到 k 次成功。



(a) Negative Binomial$(10, p)$ (b) Negative Binomial$(k, 0.3)$

圖 6: Negative Binomial(k, p) 的比較

2.6 Poisson distribution

卜瓦松分佈 (Poisson distribution) 是一種描述在單位時間內隨機事件發生次數的機率分佈，它具有以下幾個特性 ：

1. 定義域 ： Poisson 分佈的定義域是非負整數集合， $X = 0, 1, 2, \cdots$ ，因為它描述的是事件發生的次數。

2. 參數 λ ：Poisson 分佈由一個參數 λ 來描述，表示在單位時間內事件的平均發生率。

3. 機率質量函數（PMF） ：

$$f(x; \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}, \ X = 0, 1, 2, \cdots$$

4. 平均值 $E(X) = \lambda$ 、變異數 $Var(X) = \lambda$ 。

由圖 **??** 中我們可以看出隨著參數 λ 越大時會有以下特性 :

1. 分佈向右延伸：隨著 λ 的增大，整個分佈會向右延伸，表示事件發生的次數有更大的範圍。
2. 變異數增加：隨著 λ 的增大，分佈的變異數也會增加，表示事件發生次數的變異性增大。
3. 因為變異數增加，因此峰值會隨之降低。

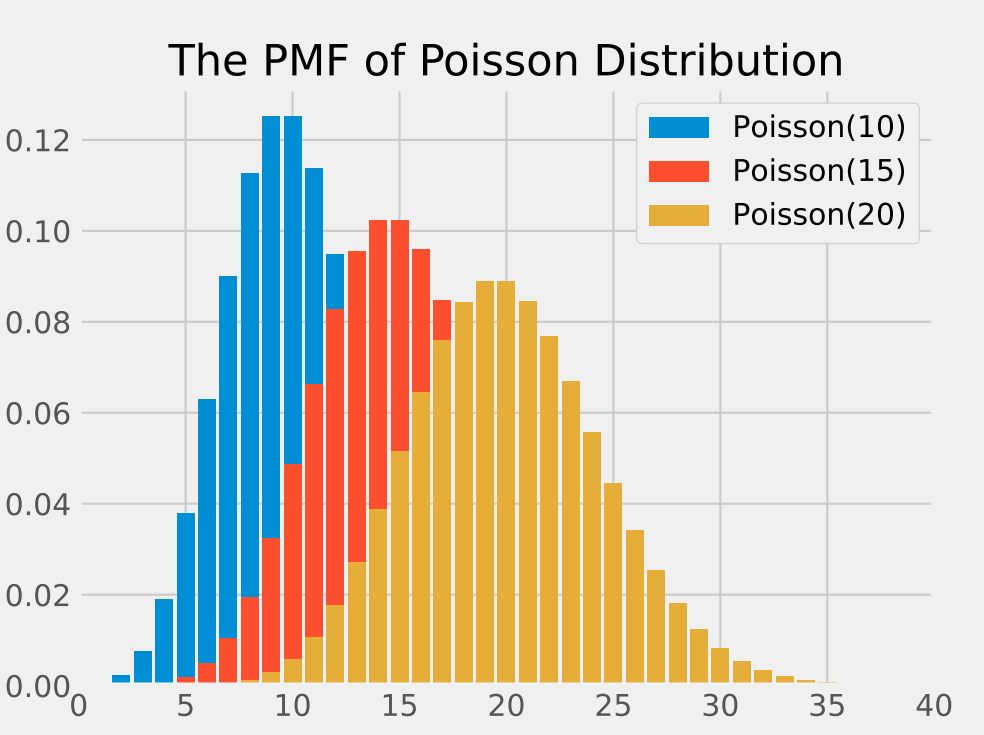
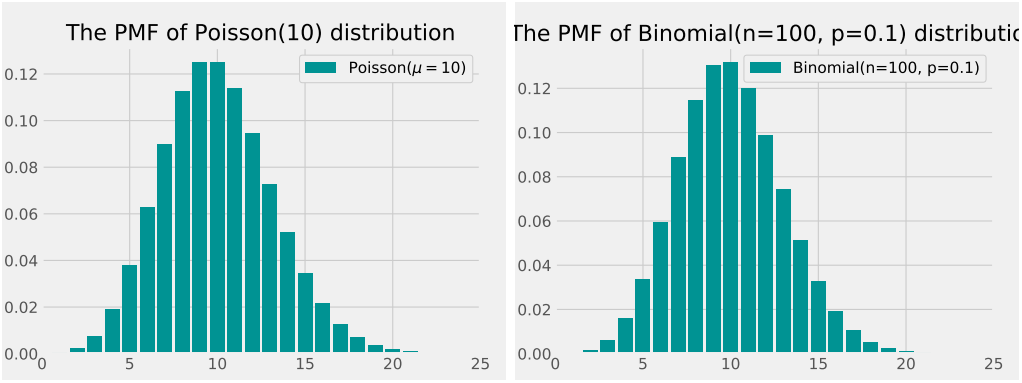


圖 7: Poisson(10)、Poisson(15)、Poisson(20)

當 n 足夠大時，母數 $\lambda = np$ 的 Poisson(λ) 分佈會近似於 Binomial(n, p) 分佈。
試著比較 Poisson$(\lambda = np = 10)$ 分佈以及 Binomial$(n = 100, p = 0.1)$ 分佈。
在圖 **??** 中，在此處為 $n = 100, p = 0.1$ ，我們可以發現，當 n 足夠大時，母數為 $\lambda = np = 10$ 的 Poisson(10) 分佈會近似於 Binomial$(n = 100, p = 0.1)$ 分佈。

3 連續型分佈 (Continuous distribution)

設 X 是具有分佈函數 F 的連續隨機變數，且 F 的一階導數處處存在，則其導函數 $f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}$ 稱為 X 的機率密度函數。



(a) Poisson($\lambda = np = 10$) (b) Binomial($n = 100, p = 0.1$)

圖 8: Poisson 分佈近似 Binomial 分佈

每個機率密度函數都有性質：

1. $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$
2. $\int_{a}^{b} f(x)\,\mathrm{d}x = \mathrm{P}(a \leq X \leq b) = F(b) - F(a)$

第一個性質表明，機率密度函數與 X 軸形成的區域的面積等於 1 ，第二個性質表明，連續隨機變數在區間 $[a, b]$ 的機率值等於密度函數在區間 $[a, b]$ 上的積分，也即是與 X 軸在 $[a, b]$ 內形成的區域的面積。因為 $0 \leq F(x) \leq 1$ ，且 $f(x)$ 是 $F(x)$ 的導數，因此按照積分原理不難推出上面兩個公式。

3.1 Chi-square distribution

以下是卡方分佈 (Chi-square distribution) 的特色 :

1. 定義域：卡方分佈的定義域為非負實數集合，即 $x \geq 0$ 。
2. 形態：卡方分佈的形態取決於自由度（ df ）。自由度是一個影響分佈形態的參數，通常用於描述相關變量之間的獨立性或者進行假設檢驗。
3. 機率密度函數：卡方分佈的機率密度函數（PDF）為：

$$f(x \mid k) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})},\ x \geq 0$$

   其中， k 是自由度， $\Gamma(\cdot)$ 是伽瑪函數。

4. 期望值與方差：卡方分佈的期望值為 $E(X) = k$ ，方差為 $Var(X) = 2k$

圖 **??** 為卡方分佈自由度從 4 每跨 2 單位直到 30 的圖形。

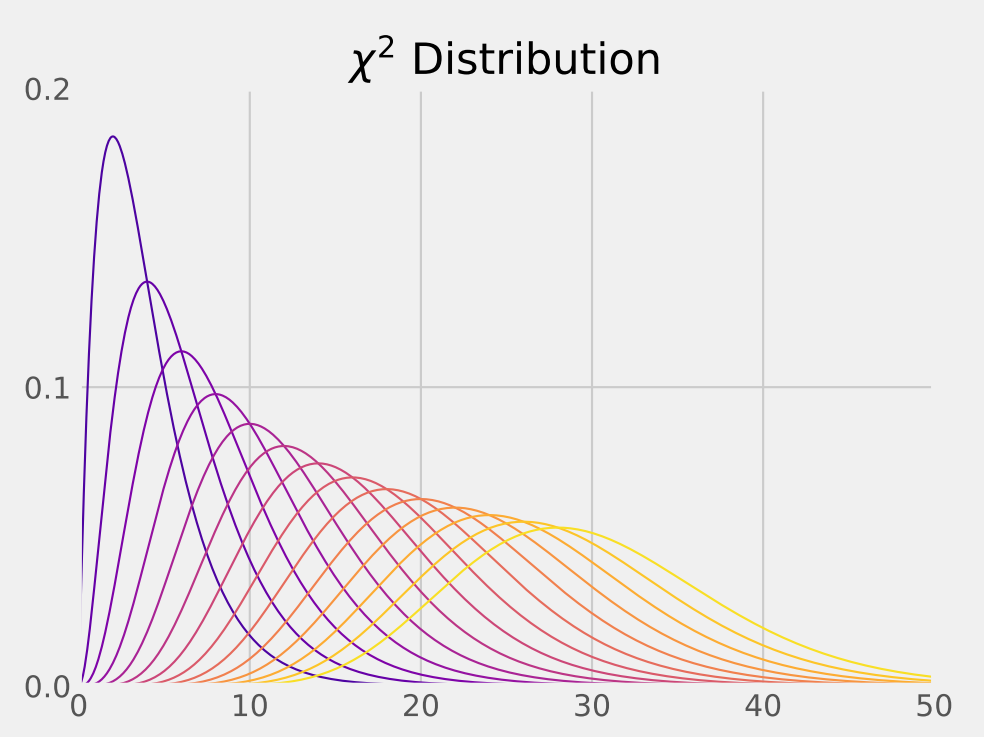


圖 9: 卡方分佈的 PDF

當卡方分佈的自由度 K 越大時，其機率密度函數 PDF 的主要特徵如下：

1. 形狀趨於對稱：隨著自由度的增加，卡方分佈的形狀會趨於對稱，且呈現出明顯的尖峰。
2. 集中度增加：隨著自由度的增加，卡方分佈的機率質量集中在較大的區域內，即隨著自由度增加，變異性減小。
3. 右偏態減弱：隨著自由度的增加，卡方分佈的右偏態（右側尾部相對較長）逐漸減弱，分佈更接近對稱。
4. 當自由度較小時，卡方分佈呈現右偏態，即右側尾部較長，大部分機率密度集中在左側；隨著自由度的增加，卡方分佈會趨近於對稱，尤其當自由度很大時，它會逐漸接近常態分佈。

3.2 Exponential distribution

指數分佈（Exponential Distribution）是連續型隨機變量的一種，可以用來表示獨立隨機事件發生的時間間隔 (等候一事件發生所需的等候時間)，其特性為：

1. 定義域：指數分佈的定義域是所有非負實數 ($X \geq 0$) 。

2. 機率密度函數（PDF）:

$$f(x;\lambda) = \lambda e^{-\lambda x},\ X \geq 0,\ \lambda > 0$$

3. 平均值 : $\frac{1}{\lambda}$ 、變異數 : $\frac{1}{\lambda^2}$ 。

4. 遺失記憶性 : 如果一個隨機變量服從指數分佈，則

$$P(X > s + t \mid X > s) = P(X > t)$$

5. 生存函數（Survival Function）: $S(x) = P(X > x) = e^{-\lambda x}$ 。

6. 累積分佈函數（CDF） : $F(x) = 1 - e^{-\lambda x}$ 。

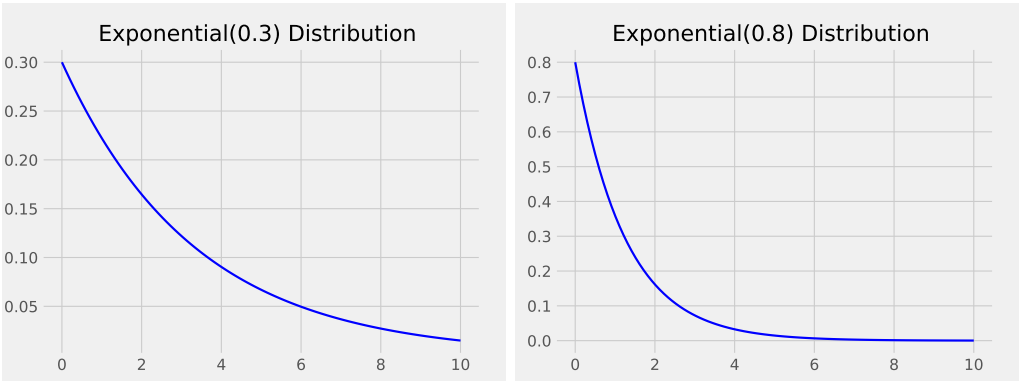
在某些書中，定義 θ 平均等候時間，而它的 PDF 則表示成下列形式 :

1. 機率密度函數（PDF）:

$$f(x;\theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}},\ X \geq 0,\ \theta > 0$$

2. 期望值 : θ 、變異數 : θ^2 。

圖 **??** 是 Exponential Distribution 在不同 λ 的比較。其中，圖 **??** (a) 為 $\lambda = 0.3$ 時，我們可以看出其 PDF 的下降趨勢較為緩慢，而圖 **??** (b) 為 $\lambda = 0.8$ 時，我們可以看出其 PDF 的下降趨勢非常快速，這是因為 $e^{-\lambda x}$ 在 λ 越大時會越小。速率參數 λ 可以被視為成功事件發生的平均速率，因此越大表示成功事件發生的速度越快。



(a) Exponential(0.3) Distribution (b) Exponential(0.8) Distribution

圖 10: 不同 λ 的比較

3.3 Gamma distribution

伽瑪分佈（Gamma Distribution）是一種連續型的機率分佈，通常用於描述一系列獨立且相同分佈的隨機變量的和， X 是用來表示等候第 α 事件所需要的等候時間。以下是伽瑪分佈的一些特性 ：

1. 定義域：伽瑪分佈的定義域為正實數軸，即隨機變數 $X > 0$ 。
2. 伽瑪分佈通常由兩個參數 : 形狀參數（shape parameter） α 和尺度參數（scale parameter） λ 來描述。
3. 機率密度函數（PDF） ：

$$f(x; k, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x},\ x > 0$$

4. 期望值 : $E(X) = \frac{\alpha}{\lambda}$ 、變異數 : $Var(X) = \frac{\alpha}{\lambda^2}$ 。

在某些書中，定義 θ 平均等候時間，而它的 PDF 則表示成下列形式 :

1. 機率密度函數（PDF） ：

$$f(x; \alpha, \theta) = \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} e^{-\frac{x}{\theta}}, x > 0$$

2. 期望值 : $E(X) = \alpha\theta$ 、變異數 : $Var(X) = \alpha\theta^2$ 。
3. 指數分佈是一種伽瑪分佈的特例，當 $\alpha = 1$ 時，

$$X \sim Gamma(1, \theta) \equiv Exponential(\theta)$$

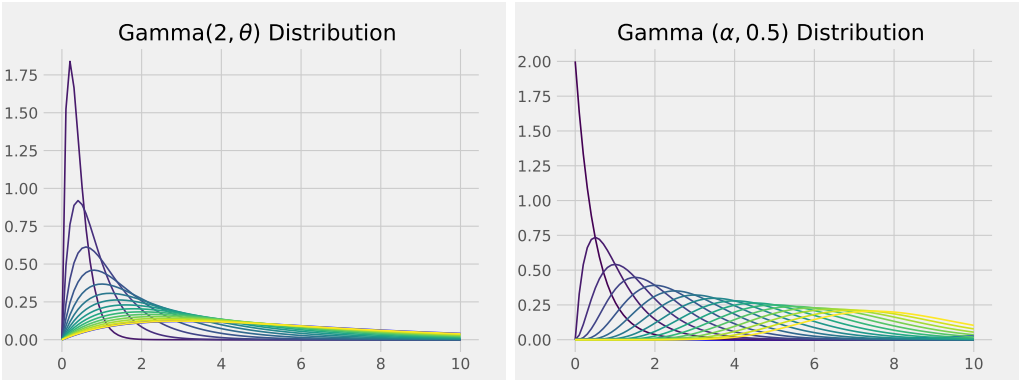
補充 1 : Gamma function : $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx,\ t > 0$

1. $\Gamma(1) = \int_0^\infty e^{-x} dx = 1$ 。
2. $\Gamma(t) = (t-1)\Gamma(t-1) = (t-1)!,\ t > 0$ 。
3. $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$ 。

補充 2 : 廣義伽瑪積分 : $\int_0^\infty x^{\alpha-1} e^{-\lambda x} dx,\ t < \lambda$

圖 **??** (a) 為伽瑪分佈在固定 $\alpha = 2$ 時， $\theta = 0$ 每隔 0.2 加到 $\theta = 3$ 的圖，我們可以發現當 θ 越大時，分佈變得更加平均，峰值變得更低。
圖 **??** (b) 為伽瑪分佈在固定 $\theta = 0.5$ 時， $\alpha = 1$ 每隔 1 加到 $\alpha = 20$ 的圖，我們可以發

現當 α 增加時，分佈的峰值變得更加銳利，也就是說，分佈變得更加集中在平均值附近。伽瑪分佈的偏度為正，所以伽瑪分佈不管 α, θ 為多少，都會是右偏的。



(a) Gamma$(2, \theta)$ Distribution　　(b) Gamma$(\alpha, 0.5)$ Distribution

圖 11: Gamma distribution 不同 α、θ 下的比較

3.4 Normal distribution

常態分佈 :

將一連續變項之觀察值發生機率以圖呈現其分佈情形，並且具有以下特性:

1. 隨機變數 (random variable) X 具有以下 pdf :

$$f_x(x) = \frac{1}{2\pi\sigma} exp(-\frac{(x-\mu)^2}{2\sigma^2}),\ x \in \mathbb{R}$$

2. 隨機變數 (random variable) x 的範圍為 $(-\infty, \infty)$ 。

3. 若隨機變數 X 為常態分佈，則稱 $X \sim N(\mu,\ \sigma^2)$

$$E(X)\ =\ \mu$$

$$Var(X)\ =\ \sigma^2$$

$$mgf\ : M_x(t)\ =\ e^{\mu t + \frac{\sigma^2 t^2}{2}}$$

4. 常態分佈的圖形是以平均數 μ 為中線，構成左右對稱之單峰、鐘型曲線分佈。

5. 若我們把方程式 $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-3)^2}{2}}$ 在 x 的區間範圍內積分起來，會得到圖下的面積為 1 ，也就是說 $\int_{-\infty}^{\infty} \frac{1}{2\pi\sigma} exp(-\frac{(x-\mu)^2}{2\sigma^2}) = 1$ 。

6. 變項之平均數、中位數和眾數為同一數值。

7. 中央極限定理：當從任何分佈中抽取大量獨立且相同分佈的隨機變量時，這些隨機變量的和會趨近於常態分佈。

8. 標準偏差 (standard deviation)(參考圖 **??**)：

   68.3% 的數值，落在平均數 ± 1 個標準差間；

   95.4% 的數值，落在平均數 ± 2 個標準差間；

   99.7% 的數值，落在平均數 ± 3 個標準差間。

標準常態分佈：

1. 隨機變數 (random variable) Z 具有以下 pdf :

$$f_z(z) = \frac{1}{2\pi} exp(-\frac{z^2}{2}),\ z \in \mathbb{R}$$

2. 隨機變數 (random variable) z 的範圍為 $(-\infty, \infty)$ 。

3. 若隨機變數 Z 為標準常態分佈，則 $Z = \frac{X-\mu}{\sigma} \sim N(0,\ 1)$

   $E(Z) = 0$

   $Var(Z) = 1$

   $mgf : M_z(t) = e^{\frac{t^2}{2}}$

4. 圖 **??** 為標準常態分佈的圖形，很明顯的，我們可以看出圖形是以綠色虛線 ($\mu = 0$) 為中線，構成左右對稱之單峰、鐘型曲線分佈。

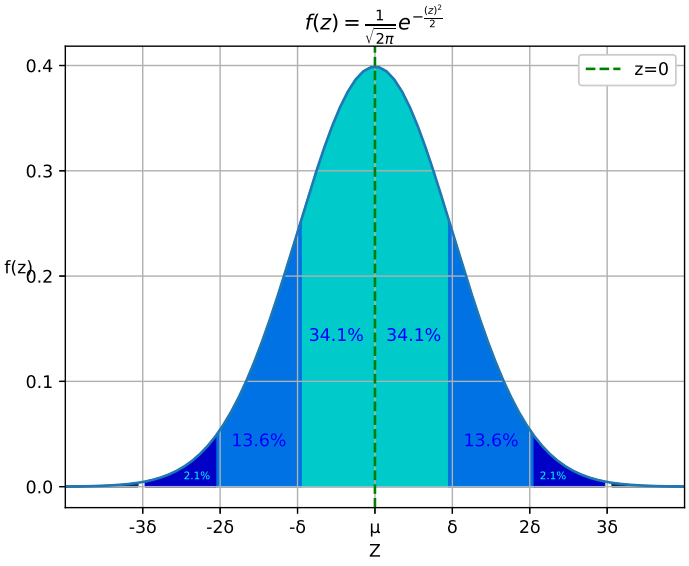


圖 12: $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(z)^2}{2}}$

5. 若我們把方程式 $f_z(z) = \frac{1}{2\pi}exp(-\frac{z^2}{2})$ 在 x 的區間範圍內積分起來，會得到 $f(z)$ 圖形下的面積 (藍色區域) 為 1 ，也就是說 $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z)^2}{2}} = 1$ 。

改變 μ 以及 σ 的差異

在圖 **??** 中我們看到了常態分配在不同 μ 以及 σ 的差異。

1. 圖 **??** 上為在固定 $\sigma = 1$ 的情況下， μ 分別帶入 $0, 1, 2, 3, 4$ 所得的結果。可以看出他們因為 σ 相同，因此高度、寬度都相同，而對稱的中心則分別依序為 $\mu = 0, 1, 2, 3, 4$ 。

2. 圖 **??** 下為在固定 $\mu = 0$ 的情況下， σ 分別帶入 $1, 2, 3, 4, 5$ 所得的結果。可以看出他們都對稱於中心點 $\mu = 0$ ，並且隨著 σ 越大，圖形會更加平緩也越矮，而隨著 σ 越小，圖形尖峰會更加集中也越高。

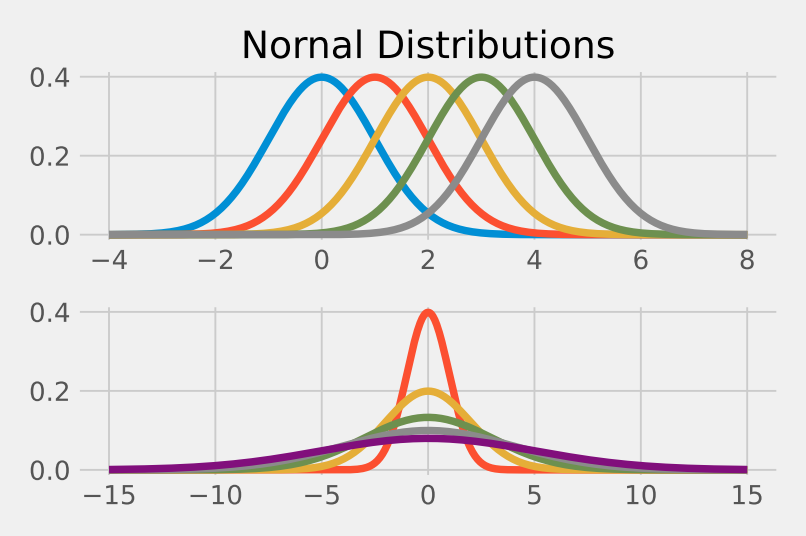


圖 13: 改變 μ 以及 σ

3.5 Uniform distribution

均勻分佈（Uniform Distribution）是一種機率分佈，它假設在一個有限的區間內，每個可能的數值都是等可能的，且在這個區間外的數值的機率為零。

1. 其機率密度函數（PDF）為 ：

$$\begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & otherwise \end{cases}$$

2. 期望值 $E(X) = \frac{a+b}{2}$ 、變異數 $Var(X) = \frac{(b-a)^2}{12}$ 。

3.6 T distribution

t 分配（Student's t 分配）是統計學中常用的機率分佈之一，用於估計樣本均值的分佈。以下是 t 分配的特性：

1. 對稱性： t 分配是一個對稱的分佈，其機率密度函數在均值處達到最大值，並隨著離均值越遠而下降。
2. 厚尾性：相較於常態分佈， t 分配具有更厚的尾部，表示在分佈的兩側有較高的機會出現極端值。
3. 依賴自由度： t 分配的形狀取決於其自由度（ df ，通常以 ν 表示）。當自由度增加時， t 分配會趨近於標準常態分佈。
4. 隨著自由度增加， t 分配的機率密度函數逐漸接近常態分佈的機率密度函數。當自由度足夠大時（通常大於 30 ）， t 分配可以很好地近似常態分佈。
5. 應用於小樣本估計：當樣本量較小時，樣本均值的分佈會因為樣本數的不足而偏離常態分佈。此時， t 分配提供了一個更合適的估計。

圖 **??** 是 t 分配的 PDF 圖 (紫色線條)，其中自由度 ν 從 0.1 到 1 每次增加 0.1 延續到從 3 到 30 每次增加 3 ，我們可以看出當自由度增加時， t 分配會趨近於標準常態分佈 (藍色線條)。

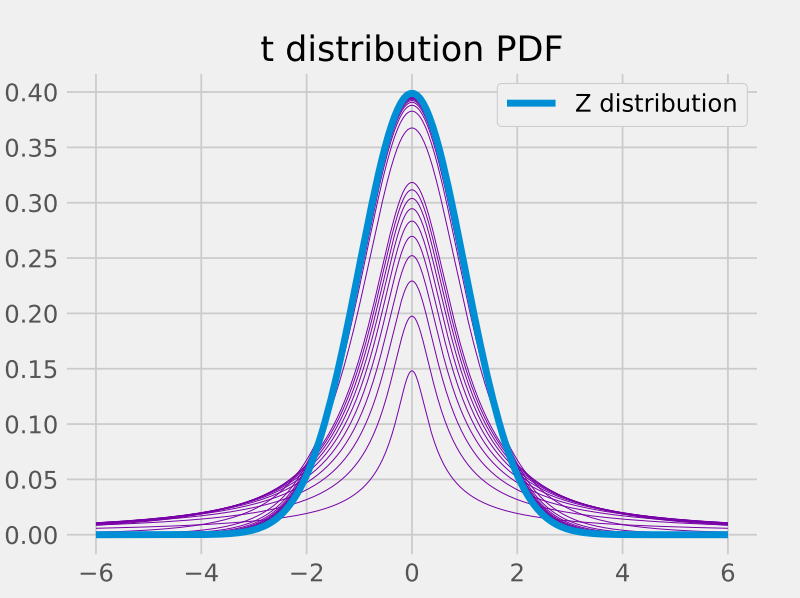


圖 14: T distribution

3.7 Beta distribution

Beta 分配是一種連續型機率分佈，通常用來描述在一個有界範圍內的隨機變量。這個分佈的機率密度函數（PDF）為：

$$f(x;a,b) = \frac{x^{a-1}(1-x)^{b-1}}{\beta(a,b)},\ 0 \le x \le 1,\ a > 0,\ b > 0$$

這個分佈的特點在於它的取值範圍固定在 $[0,1]$ 之間，且形狀受到 a 和 b 兩個參數的影響。

接下來，因為兩個參數 (a, b) 影響 β 分佈的形狀，在圖 **??** 中，我們試著嘗試各種可能的組合，看看能得到多少種不同特性 :

1. 左上代表了在 β 分配中，固定 $a = 9$ ， $1 \le b \le 30$ 的圖形變化。
2. 右上代表了在 β 分配中，固定 $b = 9$ ， $1 \le a \le 30$ 的圖形變化，由此圖我們可以看出 β 分配是可以左偏也可以右偏的。
3. 左下代表了在 β 分配中，固定 $a = b$ ，並且讓他們為 $1 \le a = b \le 9$ 的圖形變化，我們可以看出當 $a = b$ 時，圖形會對稱。
4. 右下代表了在 β 分配中，讓 $1 \le a \le 5$ ， $1 \le b \le 5$ 的圖形變化。

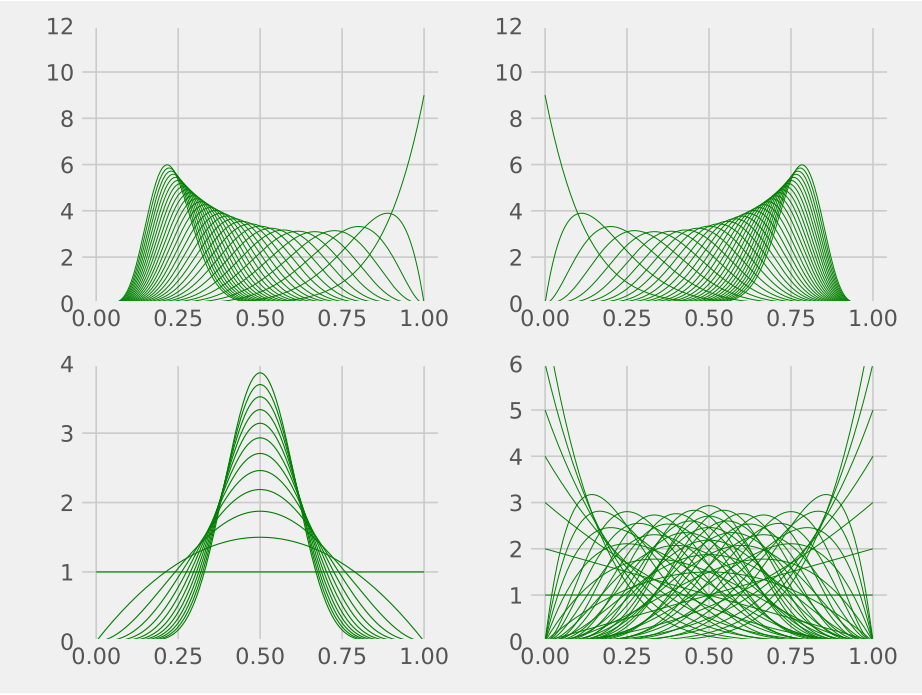


圖 15: $\beta(a,b)$ distribution

3.8 F distribution

同樣有兩個參數左右分佈的形狀：

在圖 **??** 中，圖形為 F 分佈時，固定 $n_2 = 60$ ， $10 \leq n_1 \leq 34$ 的變化。並且因為 F 分佈的偏度無論 n_1 以及 n_2 取值如何皆為正，所以 F 分配的圖形皆為右偏，不會出現左偏，但隨著自由度增大時，圖形會趨於對稱。

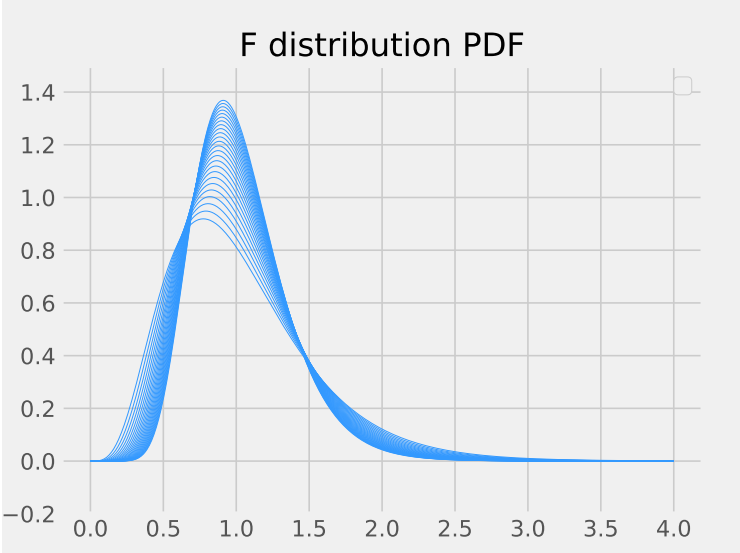
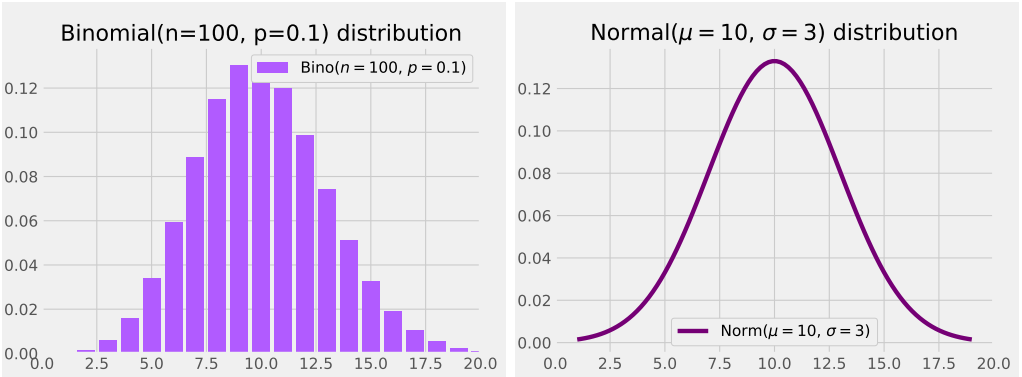


圖 16: F distribution

4 二項分佈趨近常態分佈

圖 **??** (a) 為 Bino$(n = 100, p = 0.1)$ 的圖形，圖 **??** (b) 為 Norm$(\mu = 10, \sigma = 3)$ 的圖形。在圖 **??** 中我們可以看到，因為在這邊的樣本數 $n = 100$ 足夠大，所以使得 Bino$(n = 100, p = 0.1)$ 會趨近於 Norm$(\mu = 10, \sigma = 3)$ 。



(a) Bino$(n = 100, p = 0.1)$ (b) Norm$(\mu = 10, \sigma = 3)$

圖 17: Bino$(n = 100, p = 0.1)$ and Norm$(\mu = 10, \sigma = 3)$ 分別呈現

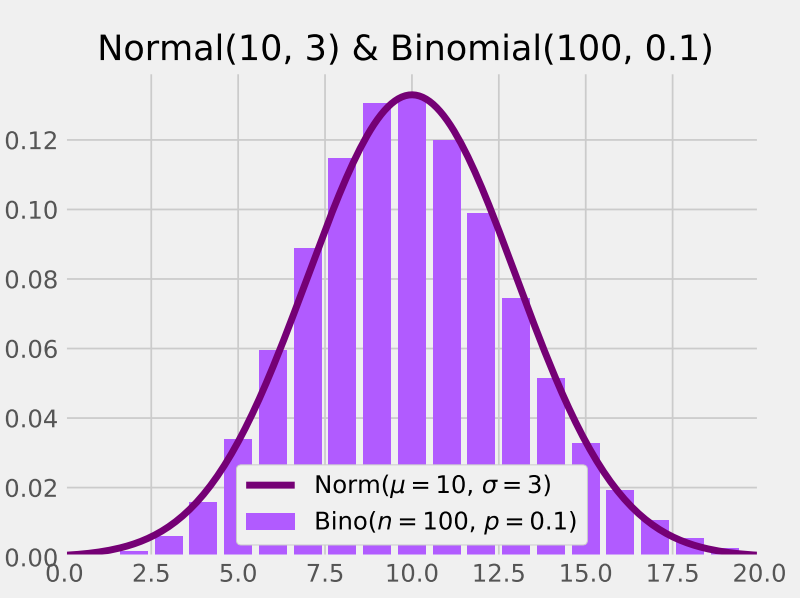
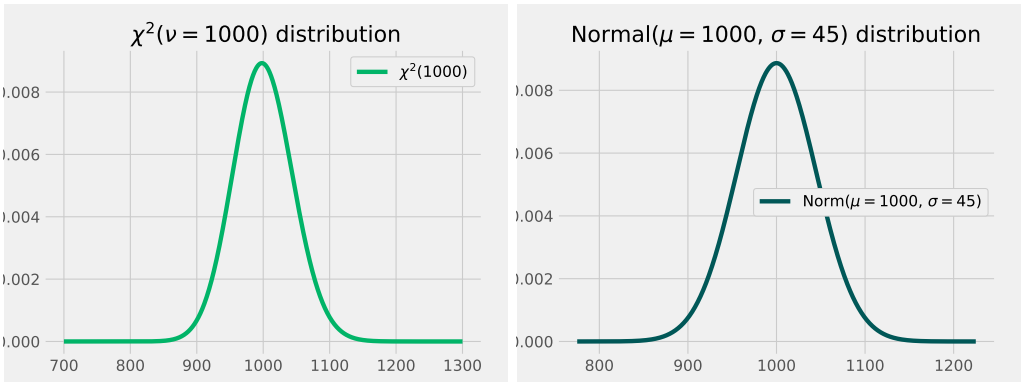


圖 18: Bino$(100, 0.1)$ and Norm$(10, 3)$ 放在一起

5 卡方分佈趨近常態分佈

圖 **??** (a) 為 $\chi^2(1000)$ 的圖形，圖 **??** (b) 為 Norm$(1000, 45)$ 的圖形。在圖 **??** 中我們可以看到，因為卡方分佈在這邊的自由度 $\nu = 1000$ 足夠大，所以使得 $\chi^2(1000)$ 會趨近於 Norm$(1000, 45)$ 。

卡方分佈 (Chi-Square Distribution) 在自由度 (degrees of freedom；df) 足夠大的情況下，會呈現出類似於常態分佈分佈的特性。這個現象被稱為中央極限定理 (Central Limit Theorem) 的一個特例。



(a) $\chi^2(1000)$ (b) Norm$(1000, 45)$

圖 19: $\chi^2(1000)$ and Norm$(1000, 45)$ 分別呈現

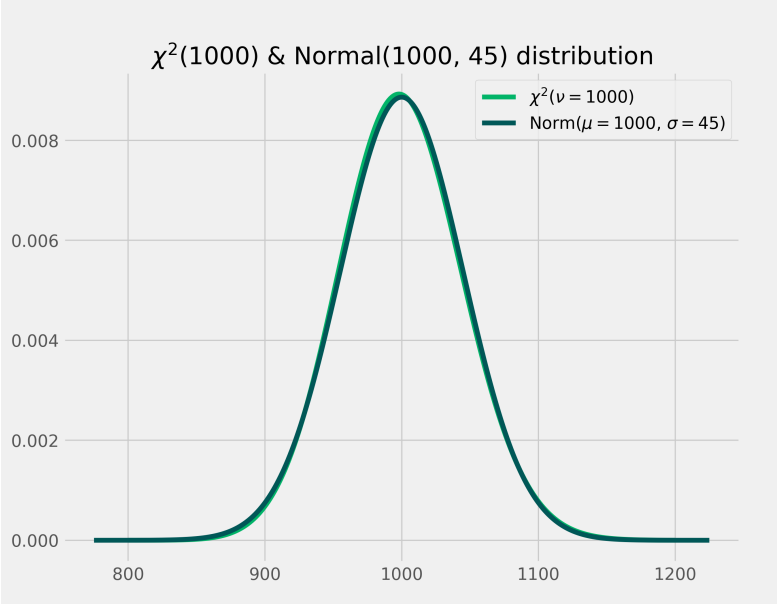


圖 20: $\chi^2(\nu)$ and Norm(μ, σ) 放在一起

6 亂數產生與相關圖形

接著我們隨機產生 100 個亂數，再去繪製圖 **??** : Histogram、 圖 **??** : Boxplot、圖 **??** : Probability plot、 圖 **??** : Empirical CDF 的圖表分析。

選擇以下不同分配 :

1. Normal distribution
2. Chi-square distribution
3. Exponential distribution
4. Uniform distribution

圖形分析使用方式 :

1. Histogram（直方圖）:

   直方圖將數據分成不同的區間，並且計算每個區間內有多少個數據點，它可以展示數據的分佈情況，特別是在不同區間的數據密度。

2. Boxplot（箱形圖）:

   箱形圖展示了數據集的五個摘要統計量：最小值、第一四分位數、中位數、第三四分位數和最大值，它提供了關於數據的集中趨勢、離散程度和可能的異常值情況的信息。

3. Probability Plot（機率圖）：概率圖用於檢驗數據是否符合某種特定的概率分佈，它將樣本數據的排序值與理論分佈的對應分位點進行比較，如果數據點與理論分佈的線性關係近似，則表明數據可能符合該理論分佈。

4. Empirical Cumulative Distribution Function (ECDF)（經驗累積分佈函數）：ECDF 展示了累積概率分佈，即給定數據集中的每個值，可以用來比較不同數據集之間的分佈情況，以及數據集中數據的集中度和離散度。

透過這些分析方式，即可以接著去探討亂數產生與相關圖形。

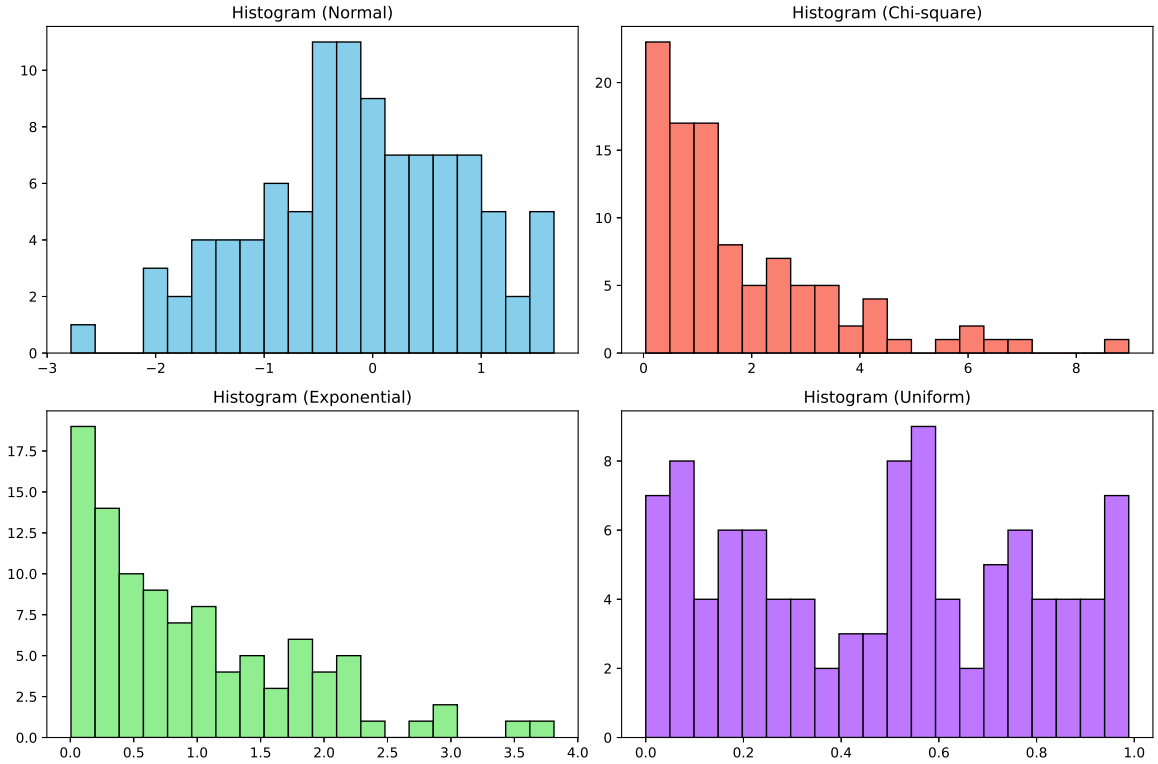


圖 21: Histogram（直方圖）

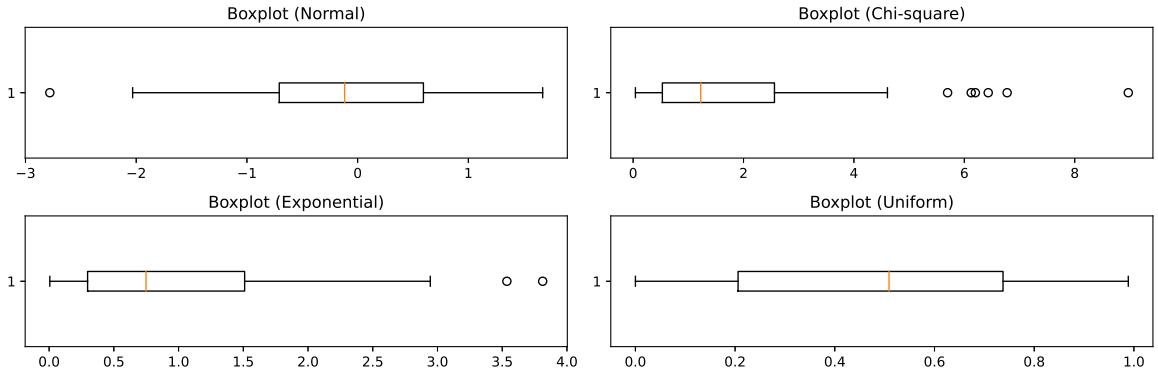


圖 22: Boxplot（箱形圖）

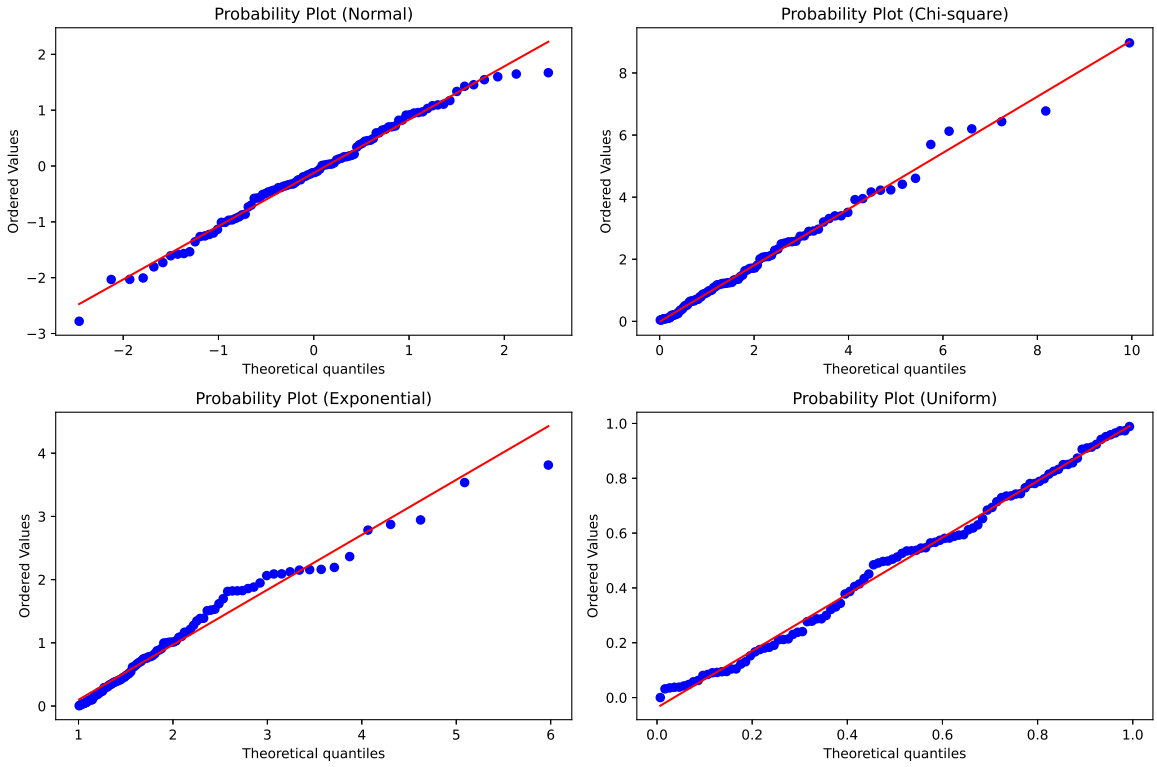


圖 23: Probability plot（機率圖）

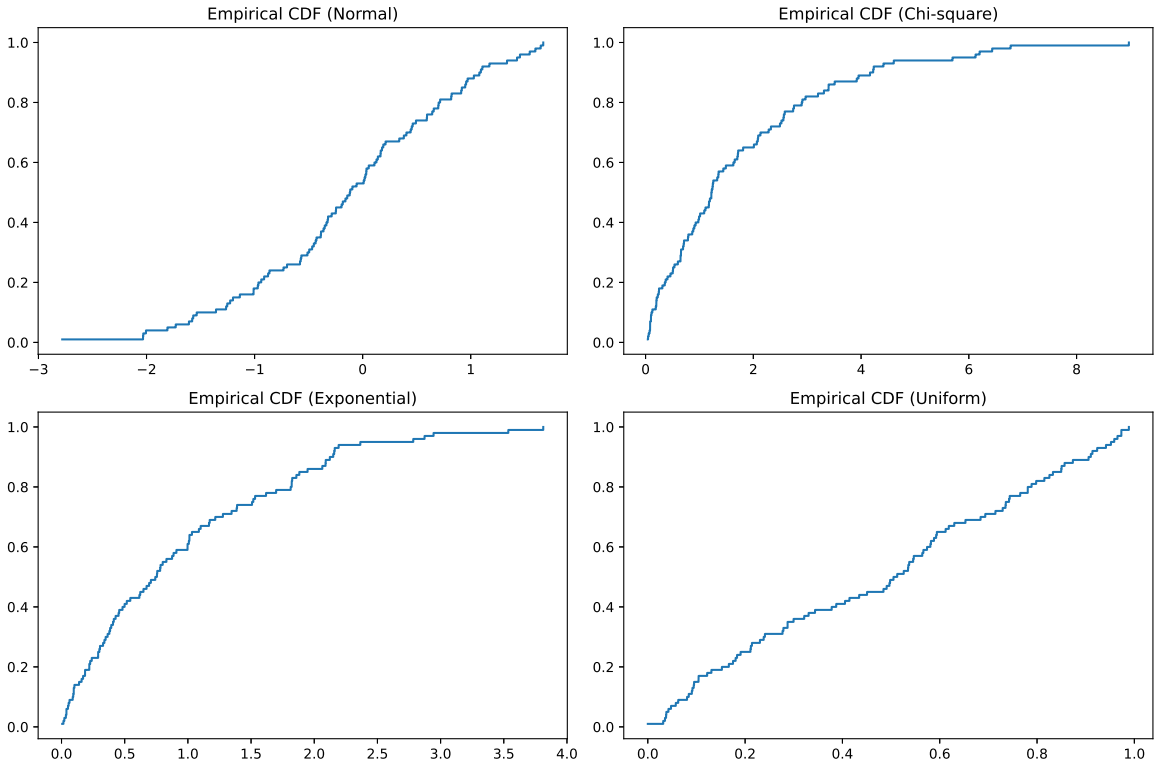


圖 24: Empirical CDF（經驗累積分佈函數）

7 抽樣分配

抽樣分配是統計學中一個重要的概念，指的是對樣本數據的統計量（例如平均值、變異數等）的分佈情況，這些統計量的分佈稱為抽樣分配。抽樣分配是在描述對單個樣本統計量的分佈，例如樣本平均值的分佈，當我們從一個母體中抽取多個樣本，每個樣本都有自己的平均值，這些樣本平均值的分佈就是抽樣分配，根據中心極限定理，當樣本容量足夠大時，這些樣本平均值的分佈將近似為常態分佈。同時，它也描述在不同樣本容量條件下統計量的分佈，例如平均值或者變異數，當樣本容量增加時，統計量的分佈也會改變，例如，隨著樣本容量增加，樣本平均值的分佈將更趨近於母體的真實平均值，並且變異性會減少。通過理解抽樣分配，我們可以進行對於母體參數的估計、檢驗統計假設等。

1. 設 $X_1, \cdots, X_n \overset{i.i.d.}{\sim} \text{Exponential}(\lambda)$ ，則 $X_1 + \cdots + X_n \sim \text{Gamma}(n, \lambda)$ 。
   圖 **??** 中我們代入 $\lambda = 5,\ n = 1000$ 來用圖形做驗證，我們也可以透過這幾張子圖來做判斷 :

   (a) 左上方子圖（Histogram 和 Gamma PDF）:
   直方圖展示了數據的分佈情況，其中包含了 5000 次指數分佈隨機變量相加的結果，曲線代表了 Gamma$(1000, 5)$ 的 PDF。X 軸表示隨機變量的值，Y 軸表示密度或者概率。

   (b) 右上方子圖（Boxplot）:
   箱形圖展示了單個 Gamma$(1000, 5)$ 分佈的隨機變量的統計摘要資訊，包括中位數、四分位範圍等。

   (c) 左下方子圖（Probability Plot）:
   用於比較實際觀察值與 gamma 分佈理論上的概率分佈情況，如果數據符合 gamma 分佈，觀察值將趨近於直線。

   (d) 右下方子圖（Empirical CDF and Real CDF）:
   這個子圖顯示了實際的經驗累積分佈函數和 gamma 分佈的理論累積分佈函數，藍色的步階狀線代表實際數據的 CDF，紅色虛線代表理論的 gamma 分佈 CDF，當兩者越接近，表示數據越符合 gamma 分佈。

   因此透過圖 **??** 可以知道 $\lambda = 5,\ n = 1000$ 。

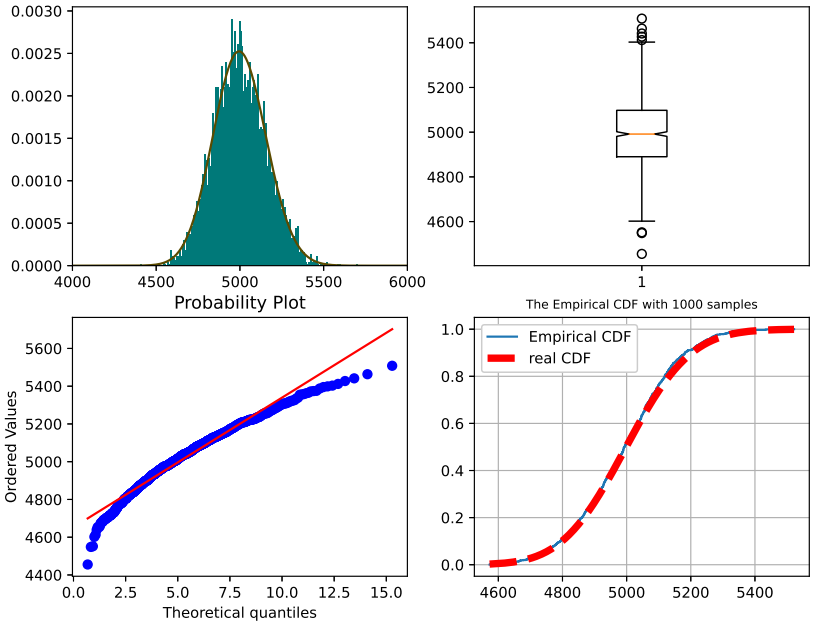


圖 25: $X_1 + \cdots + X_{1000} \sim \text{Gamma}(1000, 5)$

2. 設 $X_1 \sim \chi^2(\nu_1),\ X_2 \sim \chi^2(\nu_2)$ ，則 $\frac{X_1}{X_2} \sim \text{F}(\nu_1, \nu_2)$ 。

   圖 **??** 中我們代入 $\nu_1 = 8,\ \nu_2 = 15$ 來用圖形做驗證：

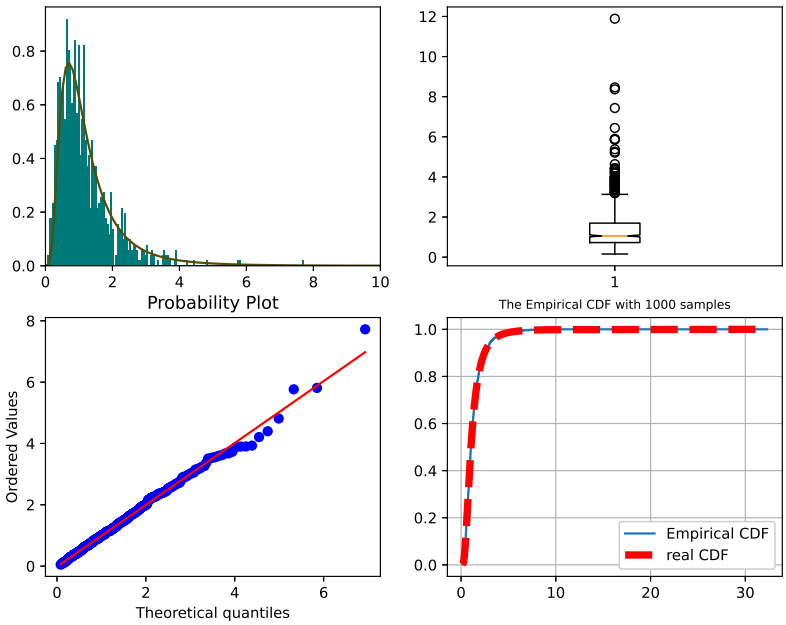


圖 26: $\frac{X_1}{X_2} \sim \text{F}(8, 15)$

3. 當 Binomial(n,p) 的 n 足夠大時， Binomial$(n,p) \approx$ Normal$(np, \sqrt{np(1-p)}$ 。
   圖 **??** 中我們代入 $p = 0.6$ 以及分別用 $n = 10,\ n = 100$ 來跟 Normal$(np, \sqrt{np(1-p)}$ 做比較 ，我們會發現當樣本數 n 越大時，二項分佈會越趨近於常態分佈:

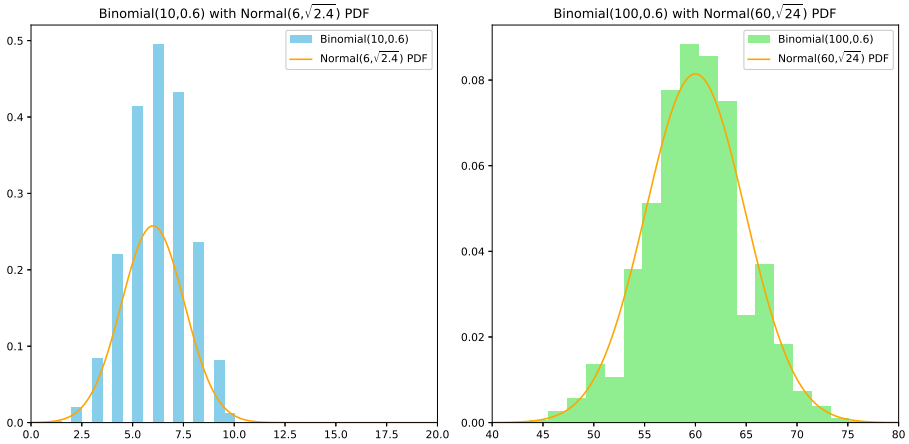


圖 27: Binomial$(n,p) \approx$ Normal$(np, \sqrt{np(1-p)})$

8 四個數字的隨機抽樣

圖 **??** 為給定四個數字 $(2,4,9,12)$ ，從這四個數字中隨機抽取四個數字（取後放回）並計算其平均數 Y 的 PMF。在此使用隨機抽樣的方式，估計出這些機率值。
抽樣過程如下 ：

1. 抽樣 :
   從數字 2、4、9 和 12 中進行有放回抽樣，抽取四個數字，並計算這四個數字的平均值 Y。

2. 概率質量函數（PMF）:
   重複進行這樣的抽樣過程多次，計算每次抽樣得到的平均數 Y。統計這些 Y 的值，並計算每個可能的平均數 Y 對應的發生概率。通過這種隨機抽樣的方式，我們可以估計從給定的數字集合中抽取四個數字後平均數 Y 的概率質量函數。

這種抽樣方法可以幫助我們 ：

1. 模擬隨機實驗：通過重複抽樣的過程，模擬並評估特定事件的概率。
2. 評估不確定性：了解樣本平均值的變化範圍，幫助估計樣本均值的不確定性。
3. 觀察分佈特徵：通過估計平均數的 PMF，了解給定數據集中平均數的可能取值及其對應的概率。

4. 探索抽樣誤差：通過重複抽樣可以觀察平均數 Y 的變化，評估抽樣誤差對結果的影響。

5. 驗證概率模型：可以用實際抽樣結果驗證或擬合概率模型，例如檢驗抽樣結果是否與理論概率分佈相符。

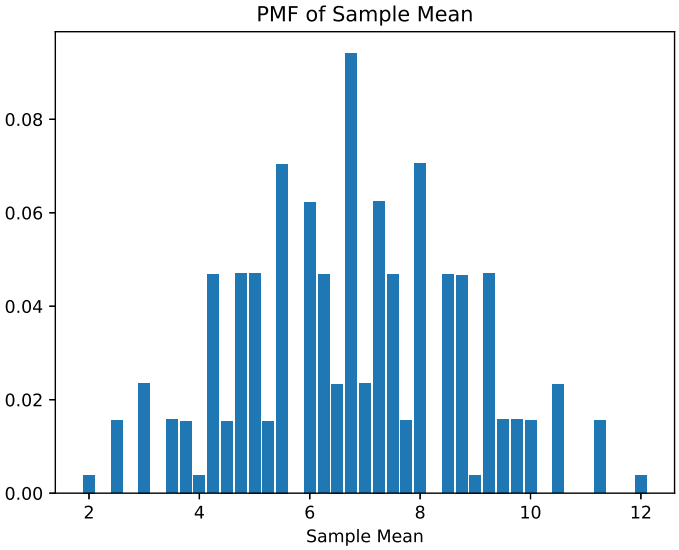


圖 28: 隨機抽樣

9 專題

蒙特卡洛模擬旨在驗證特定的統計量是否符合預期的分佈，在這個情境下，我們要驗證來自標準常態分佈的隨機樣本所計算的某個統計量是否符合特定的分佈，通過模擬生成不同數量的隨機樣本進行了模擬，我們可以觀察該統計量在不同樣本量情況下的變化，並對其分佈情況進行驗證，這種方法可以幫助我們了解該統計量在樣本量增加時的表現情況，以及在預期分佈下的接近程度。透過蒙特卡洛模擬，我們可以觀察統計量在不同樣本規模下的行為，以驗證其是否符合預期的分佈特徵，這有助於評估統計量的性質和穩定性。

令 $x_i, i = 1, \cdots, n$ 代表來自標準常態 N(0, 1) 的 n 個隨機樣本。利用蒙地卡羅模擬（Monte Carlo Simulation）驗證統計量是否服從我們要的分配。其中蒙地卡羅模擬的環境設定（scenarios）為 ：

1. 樣本數 $n = 10,\ 20,\ 30,\ 50,\ 100,\ 300,\ 500$ 。

2. 針對每個樣本數 n ，模擬次數皆為 $N = 10000$ 。

9.1 $G_1 = \sqrt{\frac{n}{6}}\hat{s}$

統計量 G_1 表示為 $G_1 = \sqrt{\frac{n}{6}}\hat{s}$ ，其中 $\hat{s}$ 為偏態係數（skewness）的估計值。利用蒙地卡羅模擬（Monte Carlo Simulation）驗證統計量 G_1 服從標準常態 $\mathrm{N}(0,1)$ 。

圖 **??** 為 $n = 10$ 的情況下 G_1 的圖形趨勢；圖 **??** 為 $n = 500$ 的情況下 G_1 的圖形趨勢，根據中心極限定理，當樣本數 n 足夠大時，樣本平均數的分佈將趨近於常態分佈，即使原始資料的分佈不一定是常態分佈。這表示無論原始資料來自什麼分佈，只要樣本數足夠大，樣本平均數的分佈就會接近常態分佈。因此我們比較圖 **??** 以及圖 **??** 各自左下角子圖的部分會發現當樣本數比較大時，也就是 $n = 500$ 的情況下，樣本的分佈會更加趨近於常態分佈。

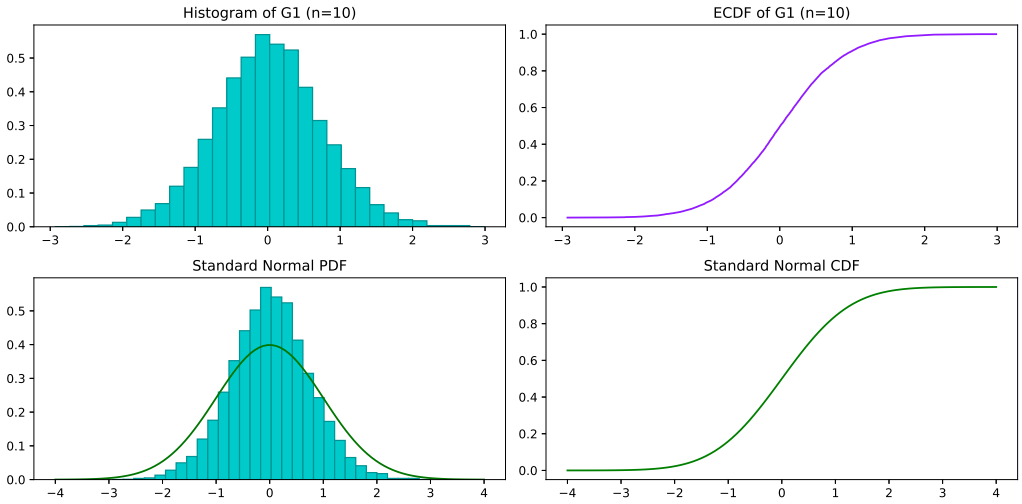


圖 29: 樣本數 = 10

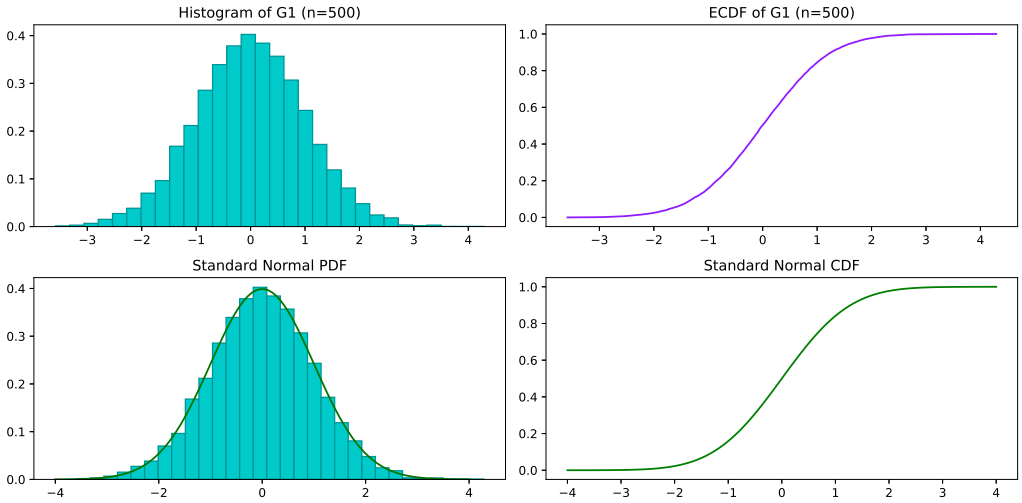


圖 30: 樣本數 = 500

9.2 $G_2 = \sqrt{\frac{n}{24}}(\hat{k} - 3)$

令統計量為 $G_2 = \sqrt{\frac{n}{24}}(\hat{k} - 3)$ ，其中 $\hat{k}$ 為峰態係數（Kurtosis）的估計值。同樣利用蒙地卡羅模擬，驗證統計量 G_2 服從標準常態 $\mathrm{N}(0, 1)$ 。

圖 **??** 為 $n = 10$ 的情況下 G_2 的圖形趨勢；圖 **??** 為 $n = 500$ 的情況下 G_2 的圖形趨勢，同樣地，根據中心極限定理，當樣本數 n 足夠大時，樣本平均數的分佈將趨近於常態分佈，即使原始資料的分佈不一定是常態分佈。這表示無論原始資料來自什麼分佈，只要樣本數足夠大，樣本平均數的分佈就會接近常態分佈。因此我們比較圖 **??** 以及圖 **??** 各自左下角子圖的部分會發現當樣本數比較大時，也就是 $n = 500$ 的情況下，樣本的分佈會更加趨近於常態分佈。

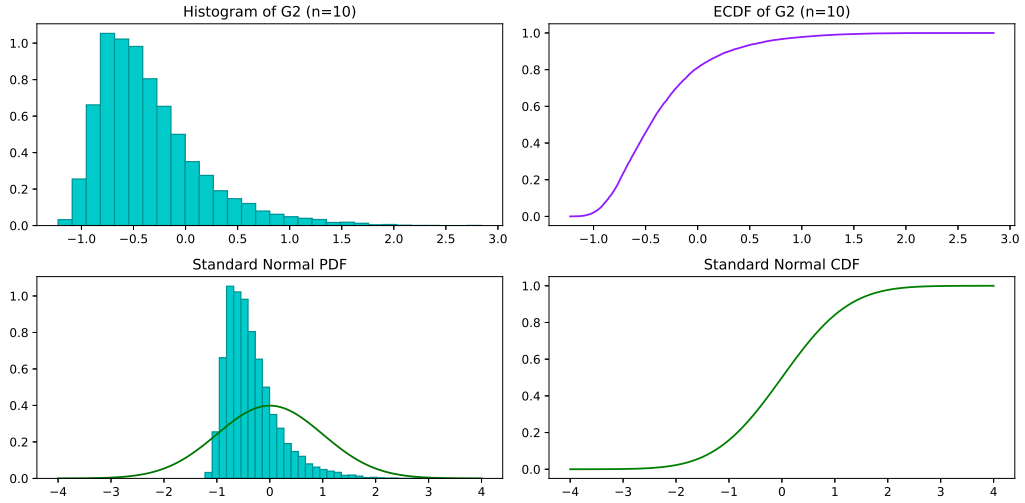


圖 31: 樣本數 = 10

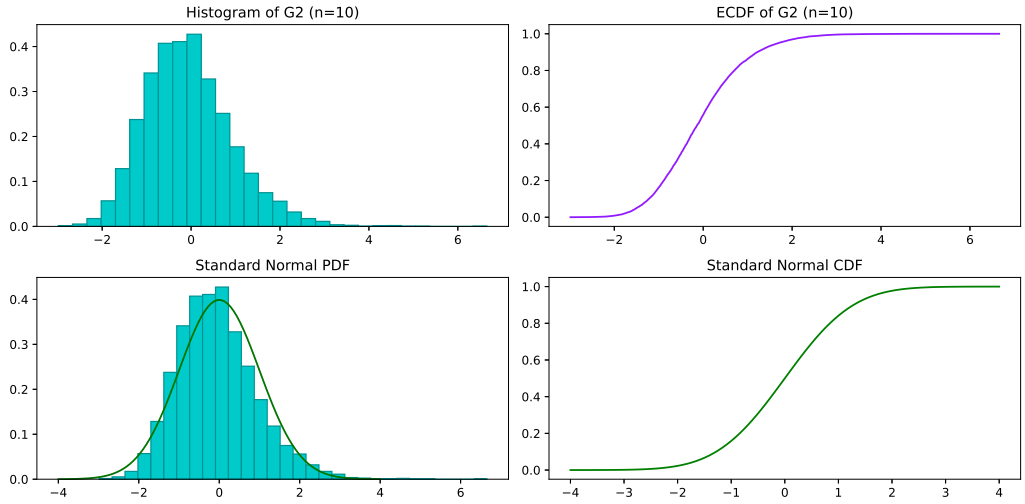


圖 32: 樣本數 = 500

9.3 $G_3 = G_1^2 + G_2^2 = \frac{n}{6}\left(\hat{s}^2 + \frac{(\hat{k}-3)^2}{4}\right)$

統計量為 $G_3 = G_1^2 + G_2^2 = \frac{n}{6}\left(\hat{s}^2 + \frac{(\hat{k}-3)^2}{4}\right)$ ，同樣利用蒙地卡羅模擬，驗證統計量 G_3 服從卡方分配 $\chi^2(2)$ 。

圖 **??** 為 $n = 10$ 的情況下 G_3 的圖形趨勢，圖 **??** 是將樣本數 $= 10$ 的 G_3、$\chi^2(2)$ 合在一起看，發現它們的分佈是非常相近的；圖 **??** 為 $n = 500$ 的情況下 G_3 的圖形趨勢。這是由於當兩個獨立標準常態分佈隨機變數的平方相加時，會趨近於自由度為 2 的卡方分佈。根據中心極限定理，當樣本數較大時，這種趨勢會變得更加明顯，使得結果更接近理論上的卡方分佈。

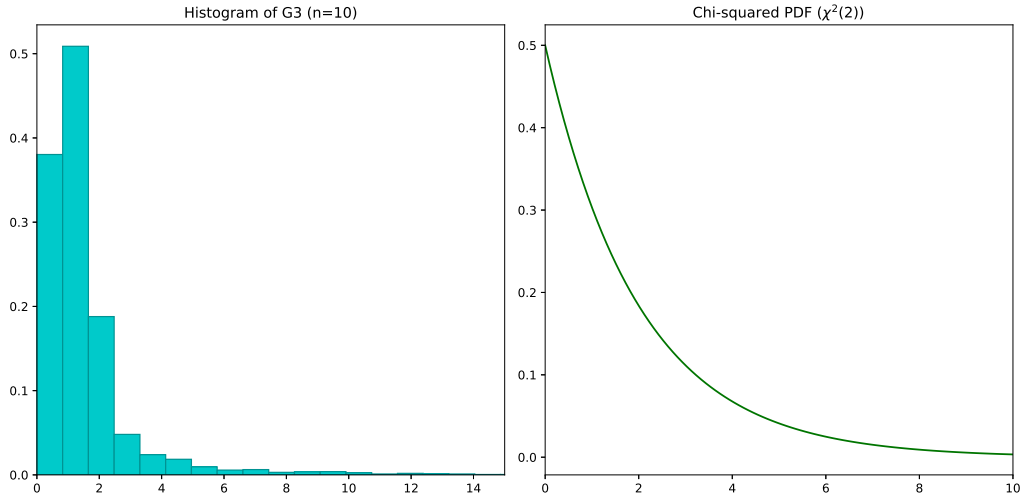


圖 33: 樣本數 $= 10$

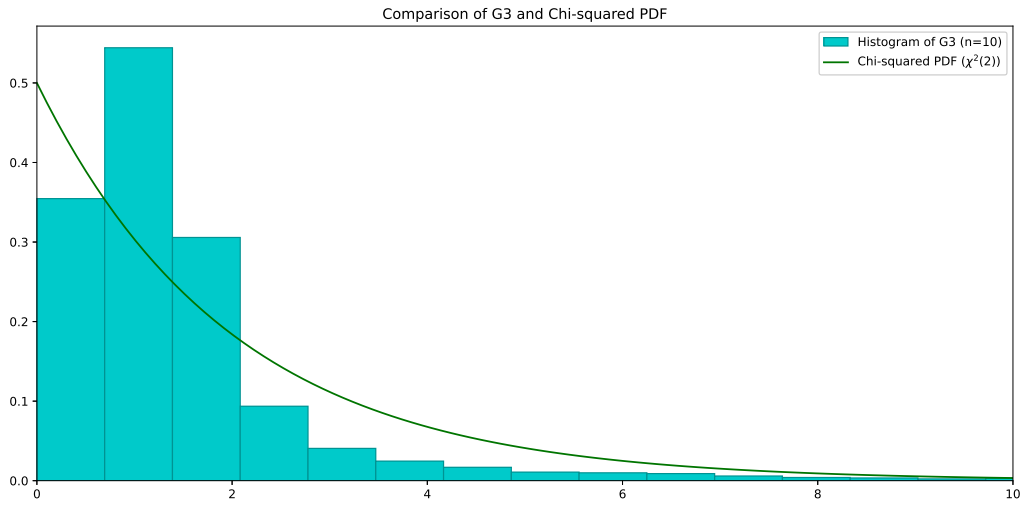


圖 34: 將樣本數 $= 10$ 的 G_3、$\chi^2(2)$ 合在一起看

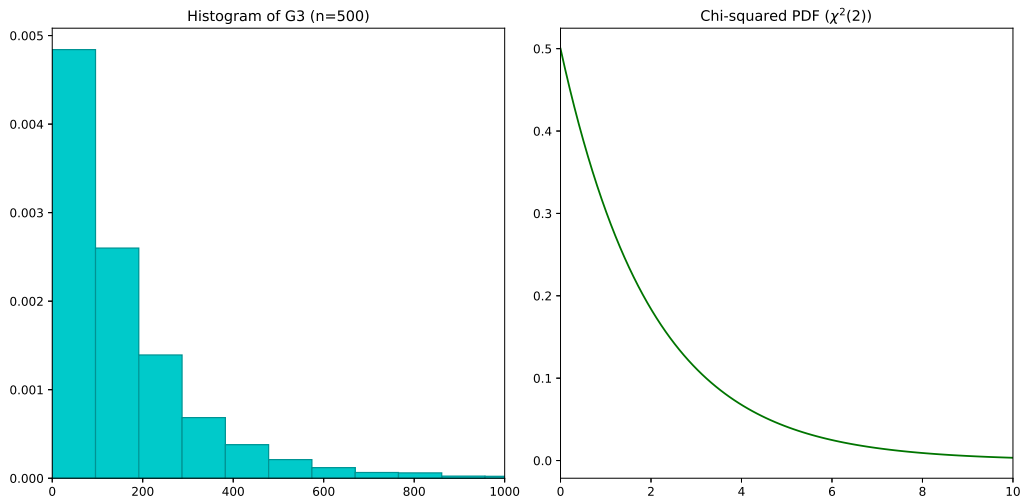


圖 35: 樣本數 = 500

10 結論

這次的實驗和模擬工作有助於驗證統計量是否符合特定分佈，強調了模型圖形對理解分佈特性的重要性。蒙特卡洛模擬驗證了偏態和峰態係數統計量在不同樣本量下的分佈情況，並確認了樣本數足夠大時中心極限定理的作用，同時，模擬也驗證了統計量 G_3 是否趨近於卡方分佈。隨機抽樣方法估計了特定數字集合平均數的機率質量函數，提供了了解隨機變量特性的重要途徑，這個研究突出了理解模型圖形在統計推斷中的重要性，為我們更深入理解統計學的應用提供了有價值的洞察。因此，理解模型的圖形分佈分佈是至關重要的，它們不僅提供了直觀、清晰的信息，也使我們能夠更準確地進行統計推斷，進而做出更具意義的預測，只有真正理解並善用這些圖形，我們才能在統計學中取得更加優異的成果。