常用分配與亂數產生

林哲兆

January 10, 2024

本文旨在使用 python 繪製不同的分配函數圖形，並驗證抽樣分配中，理想與實際狀況的差異。相較於過去在數理統計課本的探討的純數學算式，本文將著重討論不同分配實際的形狀、參數對函數的影響與驗證抽樣分配理論，希望透過視覺化的圖形，讓讀者更了解分配的特性。

1 離散型機率分配的圖形

以下將會繪製數個離散型分配的機率質量函數圖形。

1.1 伯努力分配

伯努力分配是離散型分配中最簡單的分配，只有一個參數 p，樣本觀察值也只有 $0, 1$，以下是的機率質量函數與圖形：

$$P(X = x) = p^x(1 - p)^{1-x}$$

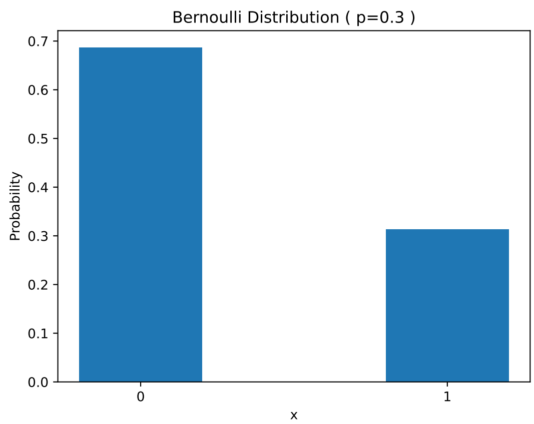

圖 1: 伯努力分配 ($p = 0.3$)

1.2 二項分配

二項分配為伯努力分配的相加，因此參數除了 p 之外，還多了樣本數 n，而 x 代表的則是成功次數，以下是它的機率質量函數與不同參數下的圖形：

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$$

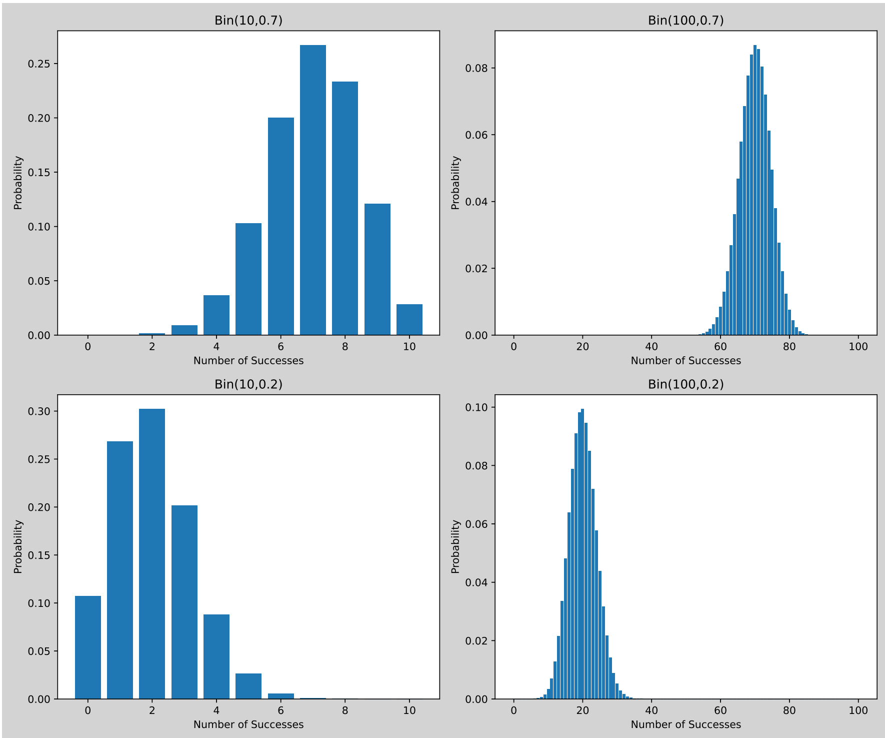

圖 2: 不同參數下二項分配的圖形

從圖 2 的上左圖與上右圖可以發現，改變 n 的話分配形狀不變，但整體圖形會右移，而且將樣本增加 10 倍，可以看出圖形原本中心點的中心點會一致原本中心點 10 倍的地方，驗證了數理統計上 $\mu = np$ 的性質。

再來，從圖 2 的上左圖與下左圖可以發現，改變 p 的話分配形狀會改變，大致呈現以 $x=5$ 為中線，左右對稱的情況。

最後我們看到圖 2 的上右圖與下右圖，兩者趨近於一個鐘型分配。為了更精確的驗證二項分配當樣本數夠大時，會趨近常態分配，所以在畫了一張圖來呈現此性質，圖如下：

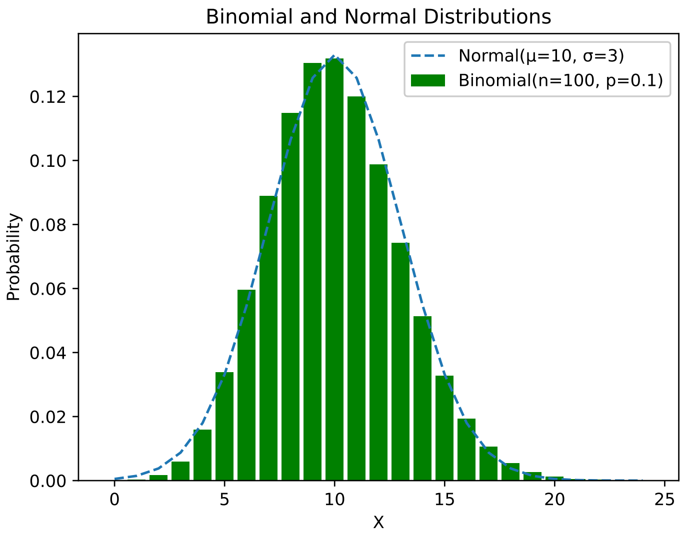

圖 3: 二項分配的圖形與近似的常態分配

從圖 3 可以看出二項分配的圖形與常態分配的函數圖形幾乎一致，成功驗證了二項分配當樣本數夠大時，會趨近常態分配的性質。

1.3 卜瓦松分配

卜瓦松分配的參數只有 λ 表示一段時間內事件發生的頻率，x 則表示一段時間內事件發生的次數，其機率質量函數與 $\lambda = 10$ 的圖形如下：

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

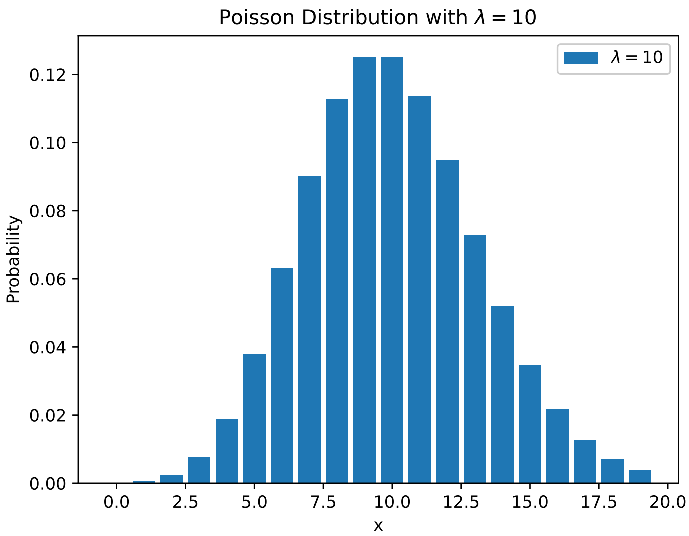

圖 4: 卜瓦松分配分配的圖形

1.4 幾何分配

幾何分配的參數只有 p，表示成功的機率，x 則表示成功一次所需次數，其機率質量函數與 $p = 0.3, p = 0.6$ 的圖形如下：

$$P(X = k) = (1-p)^{k-1}p$$

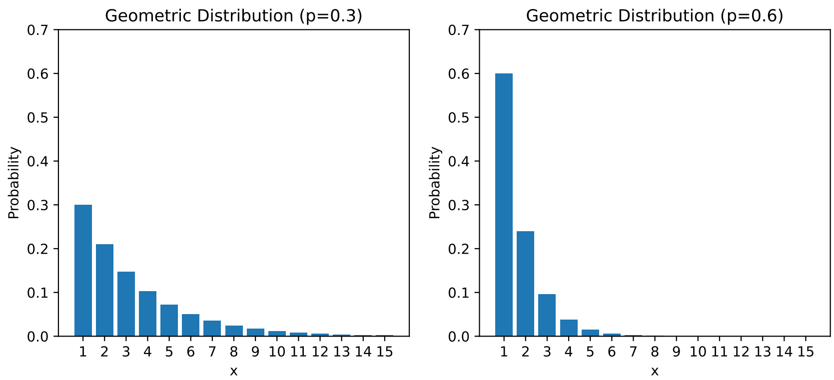

圖 5: 不同參數的幾何分配圖形

從圖 5 可以看出當 p 越大時，分配的圖形越陡，p 越小時，分配的圖形則較平緩。

1.5 負二項分配

負二項分配為幾何分配的相加，參數有 p (成功機率)、r (成功次數)，x 表示的則是成功 r 次所需花費的次數，其機率質量函數與不同參數的圖形如下：

$$P(X = k) = \binom{k+r-1}{k} p^k (1-p)^r$$

在畫圖之前，我們知道負二項分配為幾何分配的相加，因此推論其走向會與幾何分配類似，也就是當 p 越大時，分配的圖形越陡。而 r 改變的會是圖形的位置，因為幾何分配的平均數是 $\frac{1}{p}$，負二項分配則是 $\frac{r}{p}$，因此不難猜測增加 r 會使的圖形向右移。

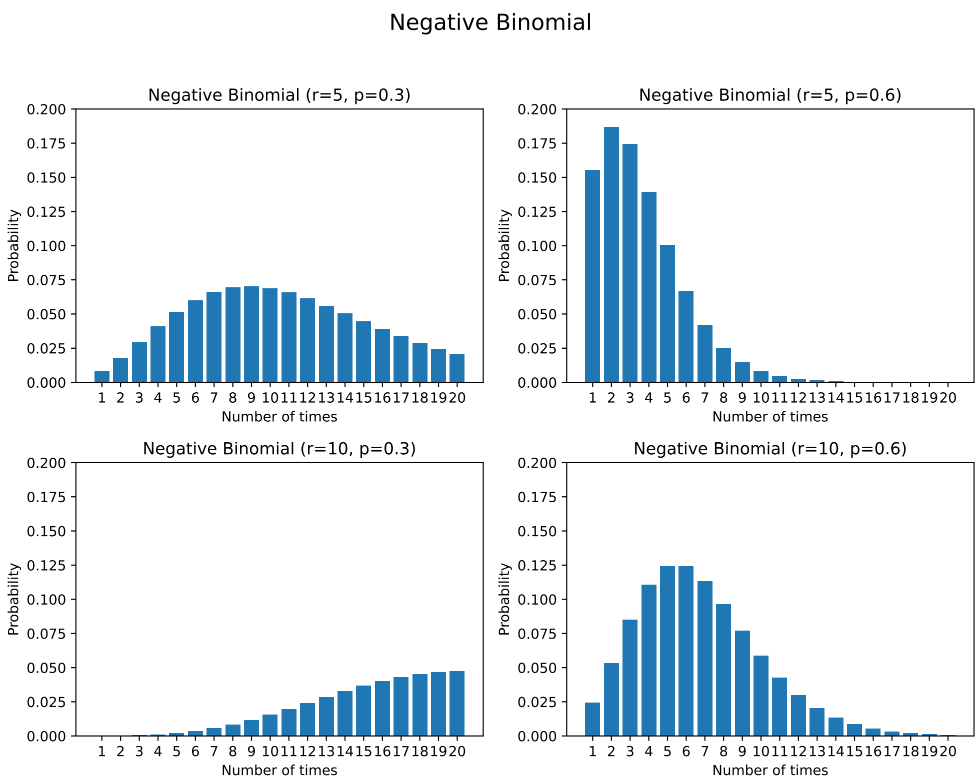

圖 6: 不同參數的負二項分配圖形

從圖 6 的左上圖與右上圖可以看出，與我們推論的一樣，當 p 增加 r 保持不變時，圖形會變得更陡。

從圖 6 的左上圖與左下圖可以看出，其中心位置往右偏移，這也與我們猜測的一樣，當 r 增加時 p 保持不變時，圖形會向右移動。

1.6 離散均勻分配

離散均勻分配也是在離散分配中相對簡單的分配，參數有 a、b，分別為整數並決定上下界，其圖形與機率質量函數如下：

$$P(X = k) = \frac{1}{b - a}$$

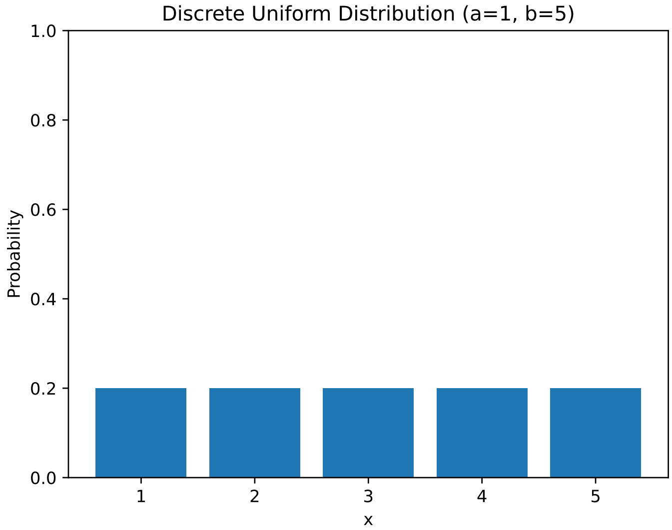

圖 7: 離散均勻分配 (1,5)

1.7 超幾何分配

超幾何分配的參數有三個，N 表總樣本數，K 樣本中符合特定描述的樣本數，n 表抽取數目，其不同參數下的圖形與機率質量函數如下：

$$P(X=k)=\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

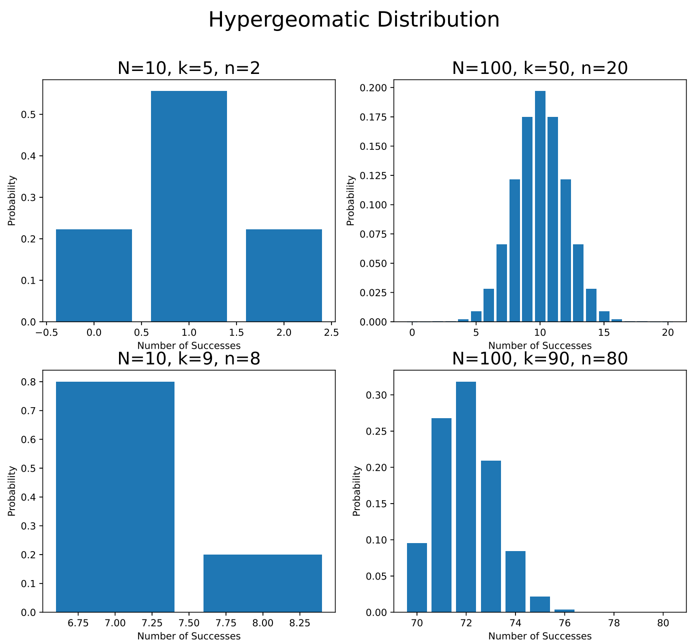

圖 8: 超幾何分配不同參數下的圖形

圖 8 右上圖的參數是左上圖的 10 倍，我們可以發現當把參數都乘以 10，其分配形狀是不會改變的，兩者都是對稱分配。但當我們把 K、n 改變，可以看到分配會從對稱變為右偏圖形 (上右圖與下右圖)。

2 連續型機率分配的圖形

以下將會繪製數個連續型分配的機率質量函數圖形。

2.1 指數分配

指數分配的參數有 λ，分配多用來描述事件間的時間間隔，其機率密度函數與不同參數的圖形如下：

$$f(x;\lambda) = \lambda e^{-\lambda x}$$

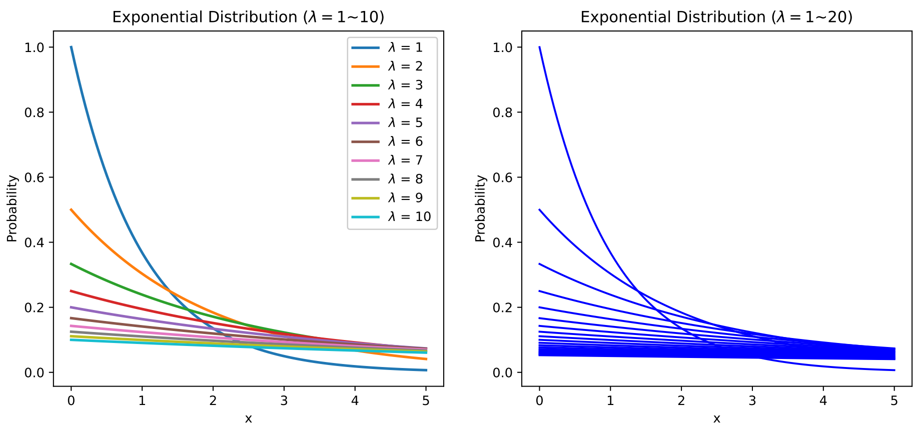

圖 9: 指數分配不同參數的圖形

從圖 9 可以看出來，當 λ 增加時，圖形會逐漸變得平緩，但不管 λ 值如何，其機率密度函數都會隨著 x 增加而逐漸趨近於 0。

2.2 迦馬分配

迦馬分配有兩個參數，α 為形狀參數，θ 為尺度參數，其中 θ 與指數分配中的 λ 相同，兩者都是屬於 Poisson family，其機率密度函數與不同參數的圖形如下：

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

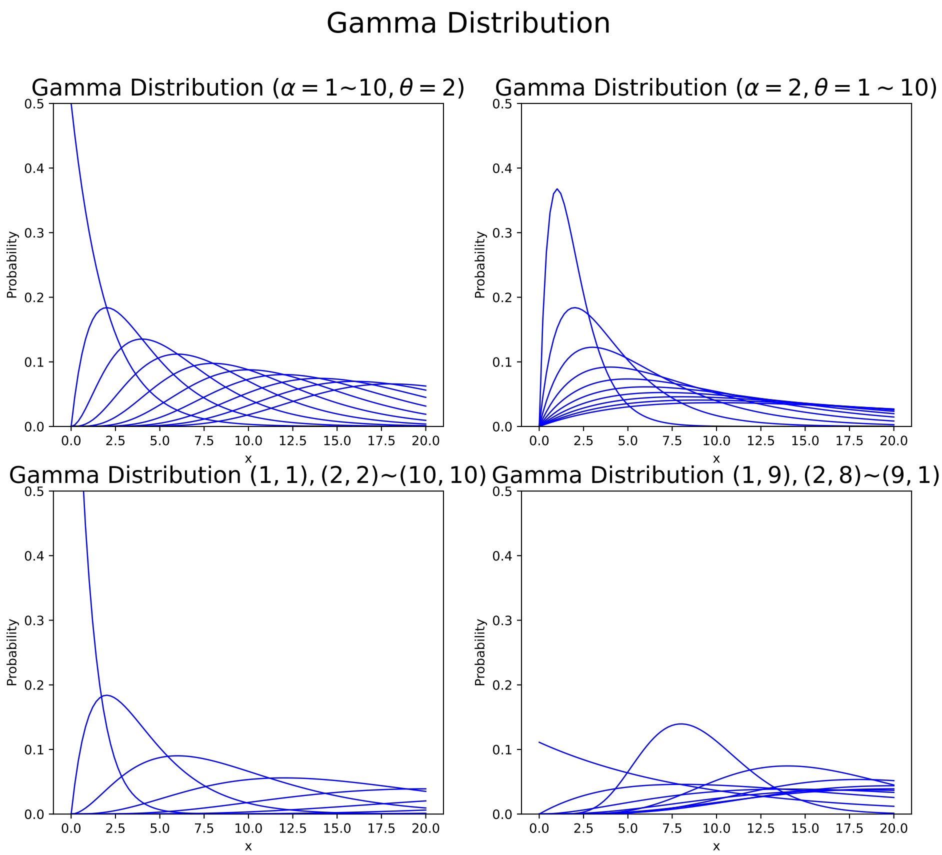

圖 10: 迦馬分配不同參數的圖形

圖 10 左上圖可以看出固定 θ 時，改變 α，會讓分配形狀從右偏到對稱再到左偏，符合其形狀參數的名字。右上圖則是固定 α 改變 θ，其變化的趨勢是隨著 θ 增加，圖形會逐漸平緩，與指數分配的變化趨勢雷同。

圖 10 的左下圖與右下圖呈現的是兩個參數同時改變的狀況，兩個參數同時增加時，位置與平緩度都會同時改變；一個增加一個減少時則看不太出變化趨勢。

2.3 貝塔分配

貝塔分配有兩個參數，α、β，兩者都是形狀參數，而 x 的值域界於 $0 \sim 1$，其機率密度函數與不同參數的圖形如下：

$$f(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

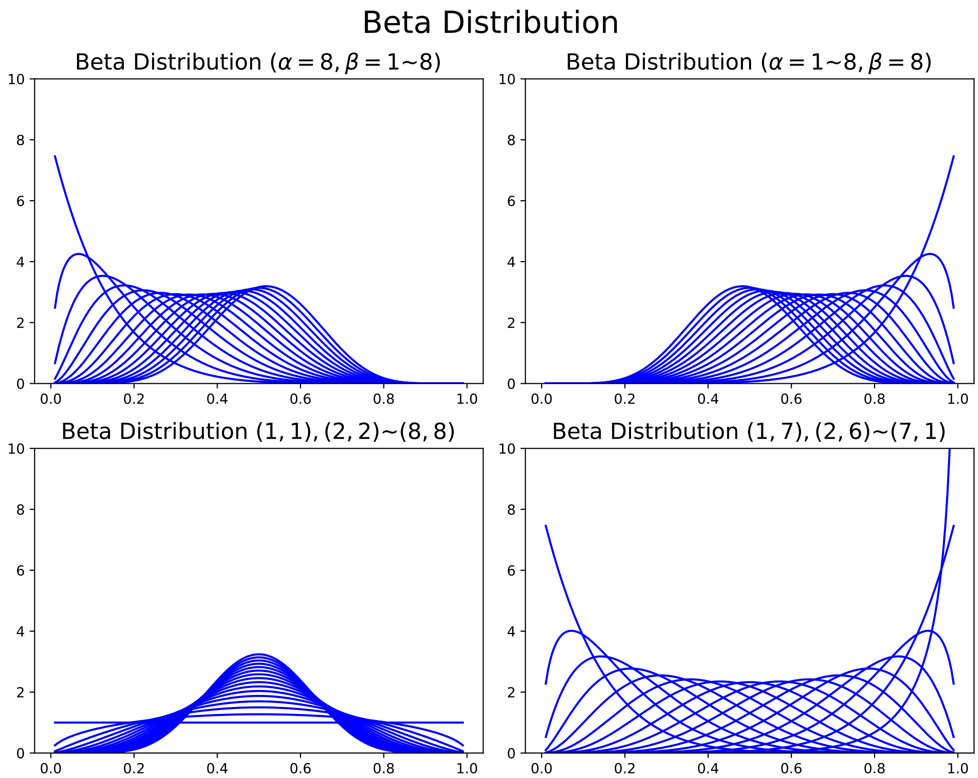

圖 11: 貝塔分配不同參數的圖形

由圖 11 上左圖與上右圖可以看出當 $\alpha > \beta$ 時，分配會右偏；$\alpha < \beta$ 時，分配會左偏；當 $\alpha = \beta$ 時會是對稱分配。

當 $\alpha = \beta$ 時，隨著兩個參數一起增加，圖形的峰態會增加，也就是中間的區域佔的機率會較大。值得注意的是，從圖 11 左下圖可以看到有一條水平線，是 $\beta(1,1)$ 的圖形，我們過去學過 $\beta(1,1) = U(0,1)$，這個圖形也確實驗證了這樣的結果。

2.4　t 分配

t 分配只有一個參數 $df = v$，其分配形狀與常態分配類似，我們知道當 df 足夠大的時候，t 分配會漸進 $N(0,1)$，除此之外，當 $v = 1$ 時，也剛好會是 $Cahchy(0,1)$，下面是驗證這兩個性質的結果與 t 分配的機率密度函數：

$$f(t;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

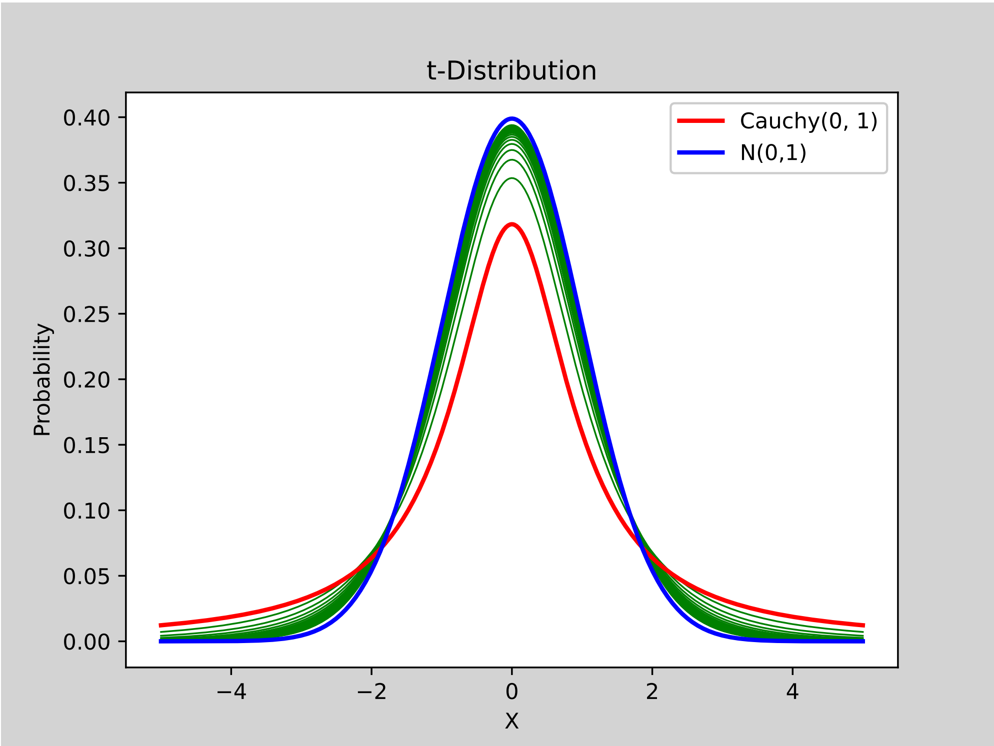

圖 12: t 分配不同參數的圖形

由圖 12 可以看出，t 分配的變動範圍會落在 $Cahchy(0,1)$ 與 $N(0,1)$ 之間，隨著 v 變大，t 分配的雙尾會逐漸縮小，機率分配集中在 0 周圍，而且看得出 $v = 0 \sim 4$ 的時候變化幅度最大，後面逐漸縮小變化幅度。

2.5 卡方分配

卡方分配只有一個參數 $df = k$，我們也知道它即是 $Gamma(\frac{k}{2}, 2)$，因此不難猜測會與迦馬分配，當 α 固定不動改變 θ 的變動一樣，由右偏往左偏變化。下面是機率分配函數與不同參數的卡方分配圖形：

$$f(x;k) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2}})x^{\frac{k}{2}-1}e^{\frac{-x}{2}}$$

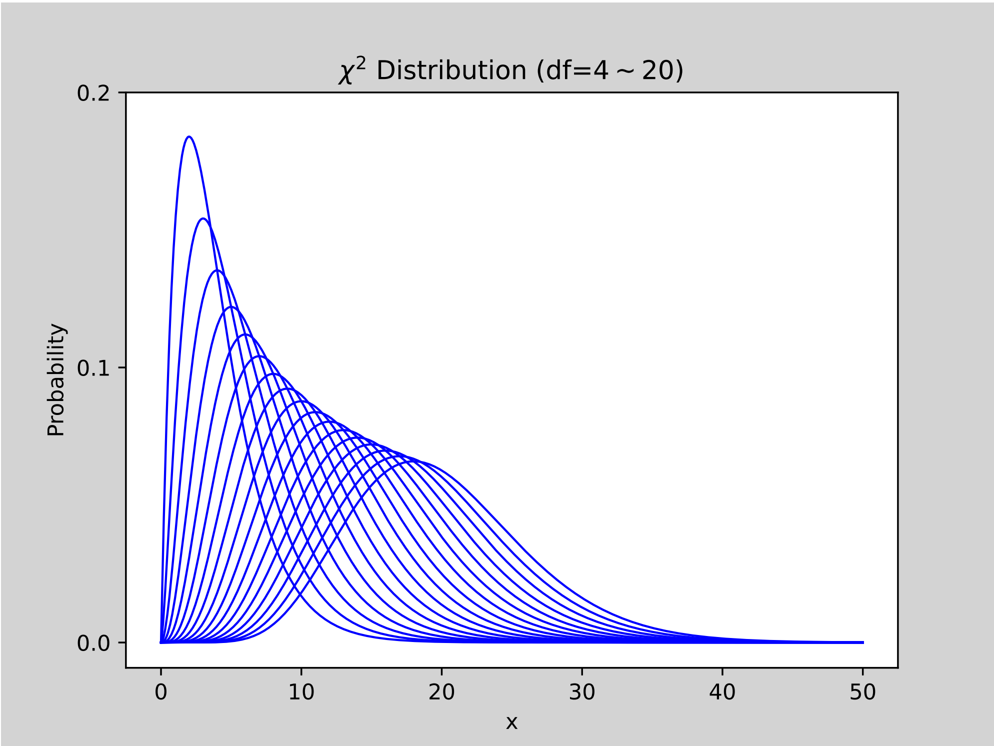

圖 13: 卡方分配不同參數的圖形

從圖 13 我們可以知道，為何卡方自由度增加時，臨界值也會跟著增加，因為分配漸漸往右移動，使得臨界值也隨著往右移動。另外，卡方分配不像 t 分配一樣變動會遞減，可以看出隨著 df 增加，分配圖形幾乎是等距往右移，不會漸漸減少變動幅度。

2.6 F 分配

F 分配有兩個參數，$d1, d2$，兩者都是自由度，其不同參數下的圖形與機率密度函數如下：

$$f(x; d1, d2) = \frac{\Gamma\left(\frac{d1+d2}{2}\right)}{\Gamma\left(\frac{d1}{2}\right)\Gamma\left(\frac{d2}{2}\right)} \left(\frac{d1/d1x}{d1/d1x + d2}\right)^{\frac{d1}{2}} \left(\frac{d2}{d1x + d2}\right)^{\frac{d2}{2}}$$

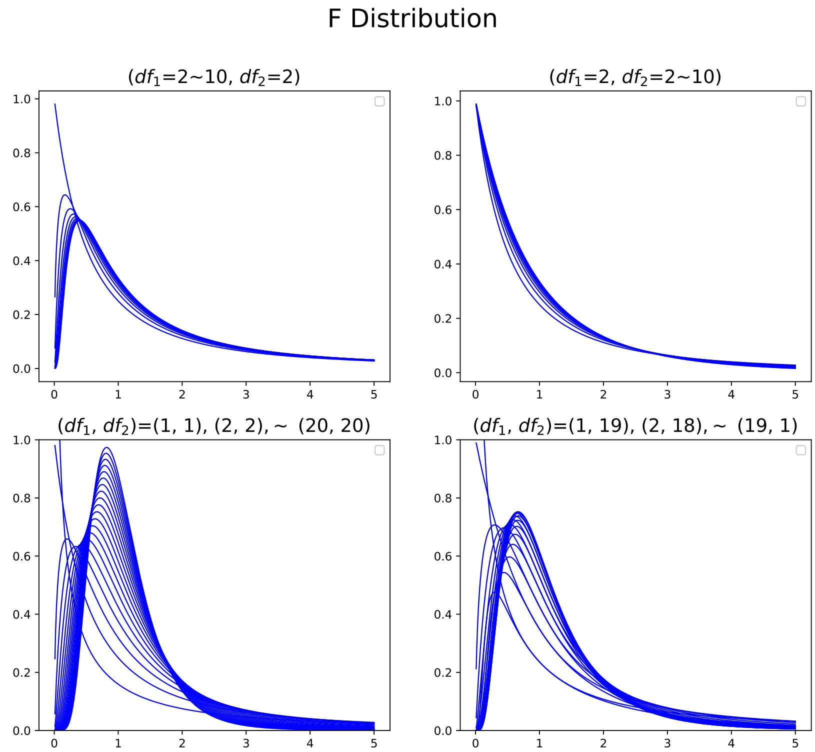

圖 14: F 分配不同參數的圖形

從 14 左上、右上圖可知，固定 $d2$ 改變 $d1$ 會使得圖形些微的往右移，變動幅度不大；若是固定 $d1$ 改變 $d2$ 圖形會由平緩漸漸變陡，值得注意的是，當 $d1 < d2$ 時，分配形狀會與 $d1 > d2$ 不一樣，是屬於單調遞減函數。再來，從圖 14 左

下、右下圖可知，同時改變兩個參數也會改變分配的位置，當兩個參數相等時，一起上升會使的圖形右移、變陡。

2.7 柯西分配

科西分配有兩個參數，與常態分配類似，具有位置參數 m 與尺度參數 γ，但不同的是柯西分配屬於厚尾分配，雖然類似鐘形分配，但與常態分配還是有一定差距，以下是其機率密度函數，與更改不同尺度參數的柯西分配圖形：

$$f(x;m,\gamma) = \frac{1}{\pi\gamma\left(1+\left(\frac{x-m}{\gamma}\right)^2\right)}$$

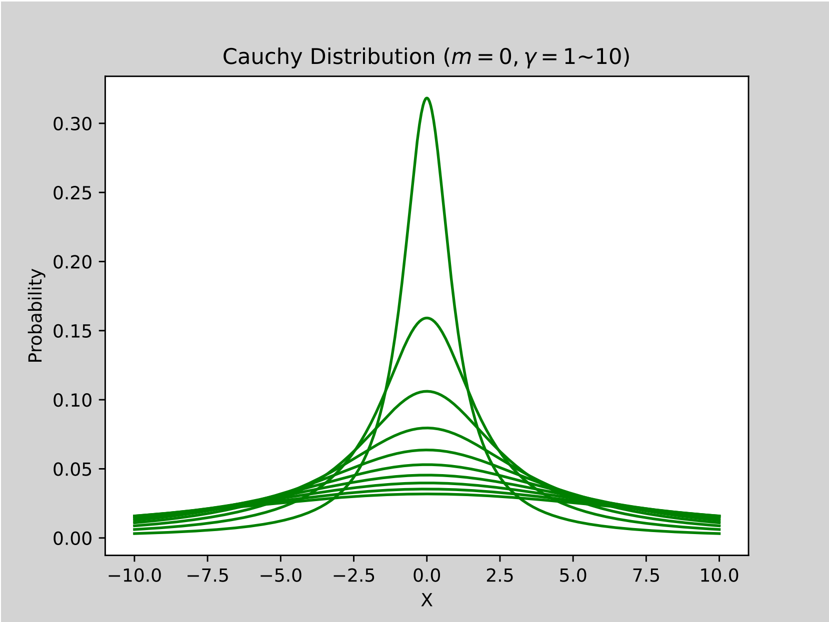

圖 15: 柯西分配不同尺度參數的圖形

從圖 15 可以看出，增加尺度參數會使得圖形趨向均勻分配，兩尾的面積會逐漸增加，中間的面積會逐漸減少，若尺度參數增加至更大，其分配函數會往一條水平線趨近。

3 亂數產生

此小節將會利用 python 的亂數產生器，產生三個分配的樣本，分別是：

- 卡方分配 (χ^2 distrbution)
- 貝塔分配 (Beta distrbution)
- 迦馬分配 (Gamma distrbution)

並將產生的亂數與理論的 pdf、cdf 比較，檢視產生的亂數是否真的服從該分配。

3.1　卡方分配

選擇 $n = 1000$ 產生卡方亂數，分別繪製直方圖、箱型圖、常態分位數圖，最後再將樣本與 Empirical CDF 比較：

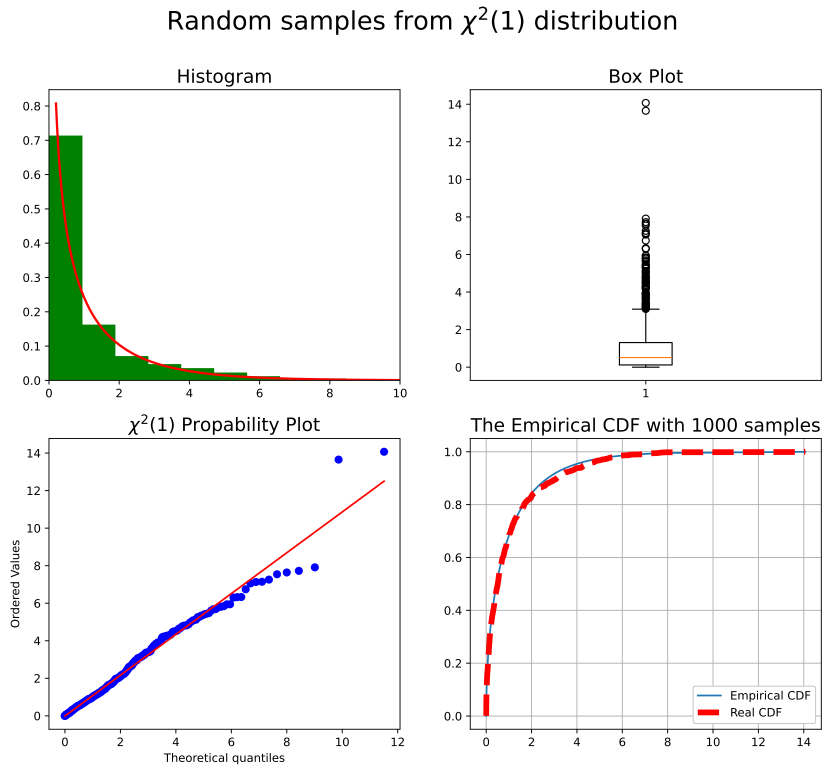

圖 16: 卡方 1 的亂數

從圖 16 的直方圖、分位數圖 ECDF 可以看出，亂數生成的樣本確實與實際理論的 $chi^2(1)$ 分配一樣，資料點幾乎貼著理論值，幾乎沒有誤差。

3.2　迦馬分配

這次選擇 $n = 10000$ 產生迦馬分配亂數，分別繪製直方圖、箱型圖、常態分位數圖，最後再將樣本與 Empirical CDF 比較：

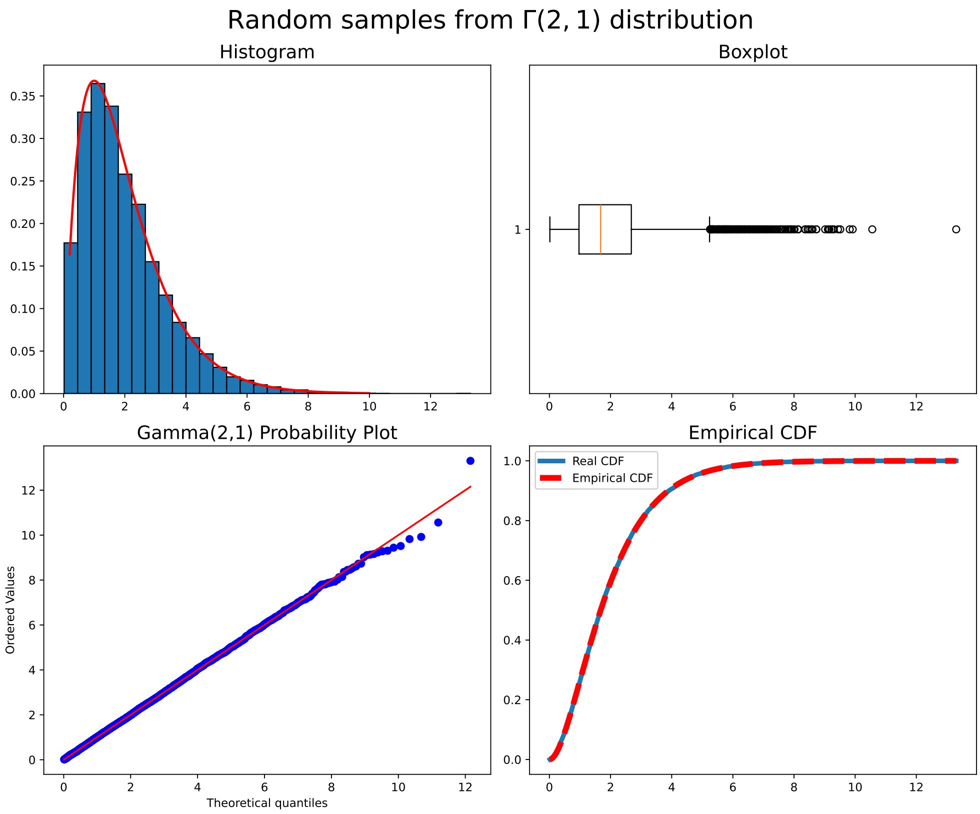

圖 17: $\Gamma(2,1)$ 的亂數

從圖 16 的直方圖與分位數圖可以看出，亂數生成的樣本確實與實際理論的 $\Gamma(2,1)$ 分配一樣，除了幾個極端值，資料點幾乎貼著理論值，幾乎沒有誤差。

3.3 貝塔分配

這次選擇 $n = 100$ 產生貝塔分配亂數，分別繪製直方圖、箱型圖、常態分位數圖，最後再將樣本與 Empirical CDF 比較：

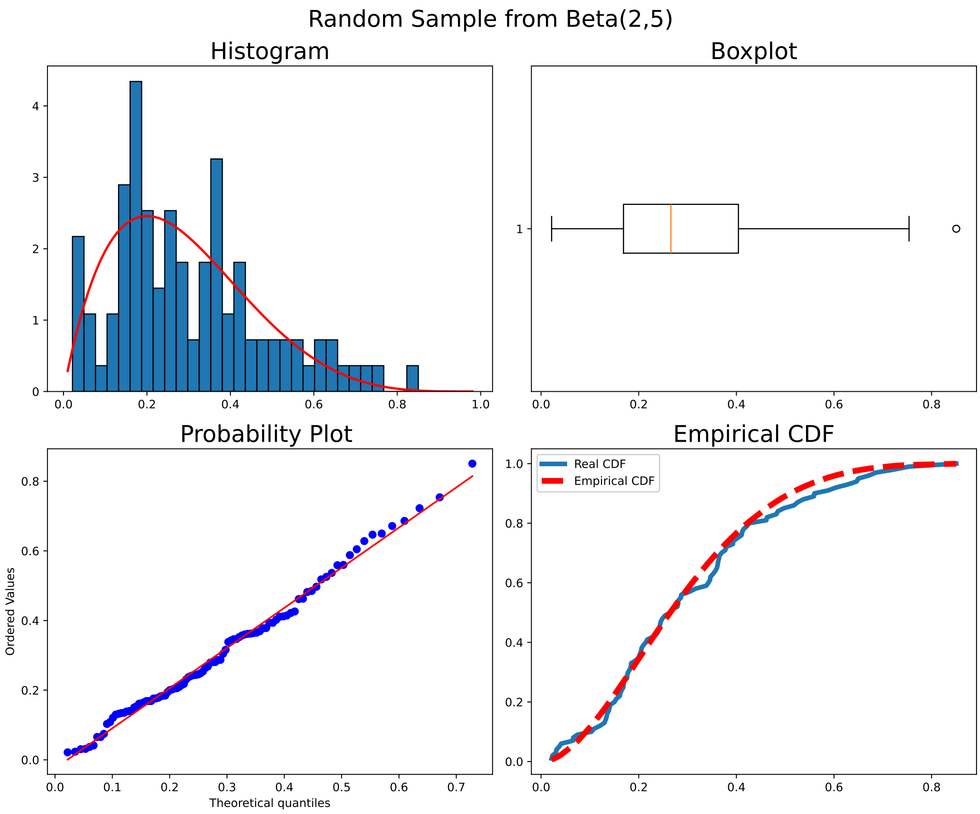

圖 18: $Beta(2,5)$ 的亂數

當 $n = 100$ 時，我們可以發現不管在直方圖、分位數圖還是 ECDF 圖，都比起選擇 $n = 1000, 10000$ 存在著較大的誤差，雖然存在誤差，但亂數還是大致照著理論分配走。

4 抽樣分配

在這小節將會介紹三種抽樣分配，以程式繪圖去驗證我們當初在數理統計上的結果，檢驗理論是否是真的正確，並且每個分配都會分別試驗抽取不同的樣本數，來檢視樣本大小對抽樣分配理論的重要性，以下是三個抽樣分配與結果：

- $x_i \overset{\text{i.i.d.}}{\sim} Poisson(1)$ then $\sum_{i=1}^{n} x_i \sim Poisson(n)$
- $x_i \overset{\text{i.i.d.}}{\sim} exp(\lambda)$ then $2\lambda \sum_{i=1}^{n} x_i \sim \chi^2(2n)$
- 中央極限定理

分別為卜瓦松加成性、指數分配加成性與中央極限定理的驗證。

4.1 卜瓦松分配加成性

在機率概論曾學過，n 個相同 λ 的獨立卜瓦松分配，相加還會是一個卜瓦松分配，而且 λ 值會變成每個卜瓦松分配的 λ 相加，也就是 $n\lambda$。為了驗證這項結果，我將產生 20 個獨立且相同的 $Poisson(1)$ 分配，三次分別抽取 100,1000,10000 個樣本，並驗證這 20 個獨立分配相加後會服從 $Poisson(20)$。

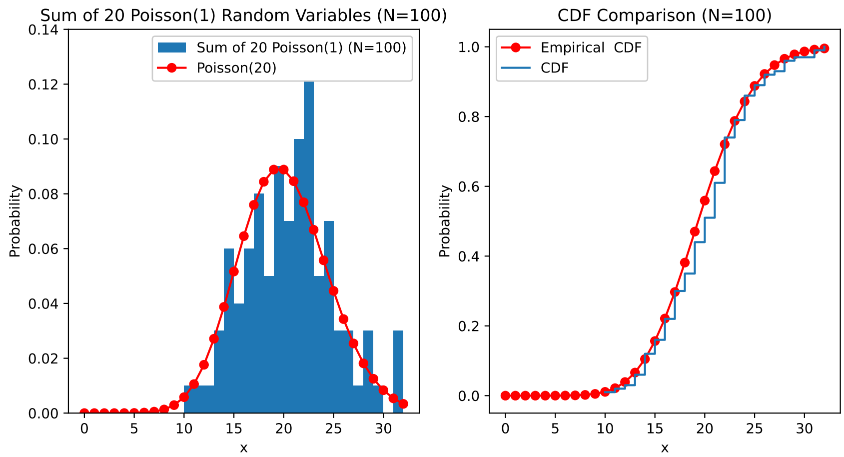

圖 19: 20 個樣本數 100 的 $poisson(1)$ 分配相加

從圖 19 可以看出在 $n = 100$ 時直方圖與實際上 $poisson(20)$ 的分配還是存在著明顯差距，但轉換成 cdf 後，與理論分配的差距並不算大，我們接著繼續看 $n = 1000, 10000$ 的情況：

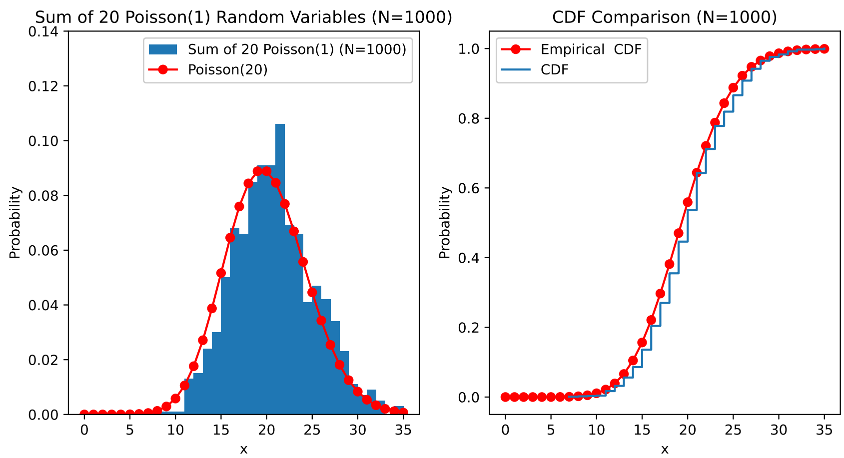

圖 20: 20 個樣本數 1000 的 $poisson(1)$ 分配相加

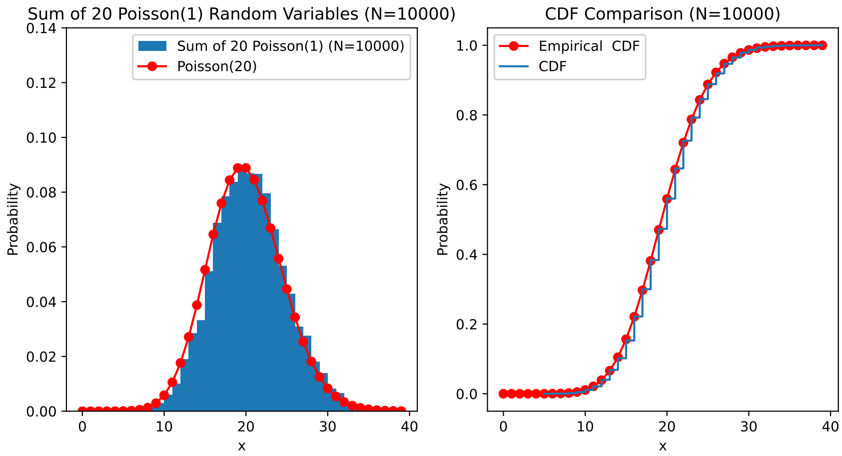

圖 21: 20 個樣本數 10000 的 $poisson(1)$ 分配相加

從圖 20、圖 21 可以看到當 n 足夠大的時候，抽樣分配不管是在直方圖還是 CDF 都已經和理論值呈現同樣的分布趨勢。

但若將 3 種情況的平均數計算出來，分別為：

$$\mu_{100} = 19.95\,, \mu_{1000} = 20.12\,, \mu_{10000} = 19.9$$

可以發現三者與理論平均值的差距不會隨著 n 增加而縮小，可見平均數不需要樣本夠大即可趨近於理論值。

4.2 指數分配加成性

在數理統計的假設檢定中，我們常把指數分配轉換為卡方分配進行檢定統計量的計算，因為透過卡方分配我們才能做查表，以做出拒絕或不拒絕的結論。在這個例子中，我將分別產生 30,100,1000 個獨立且相同的 $exp(1)$ 樣本，每個樣本包含 100 個觀察值，驗證把樣本相加再乘以 2λ 後，會服從 $\chi^2(2n)$。

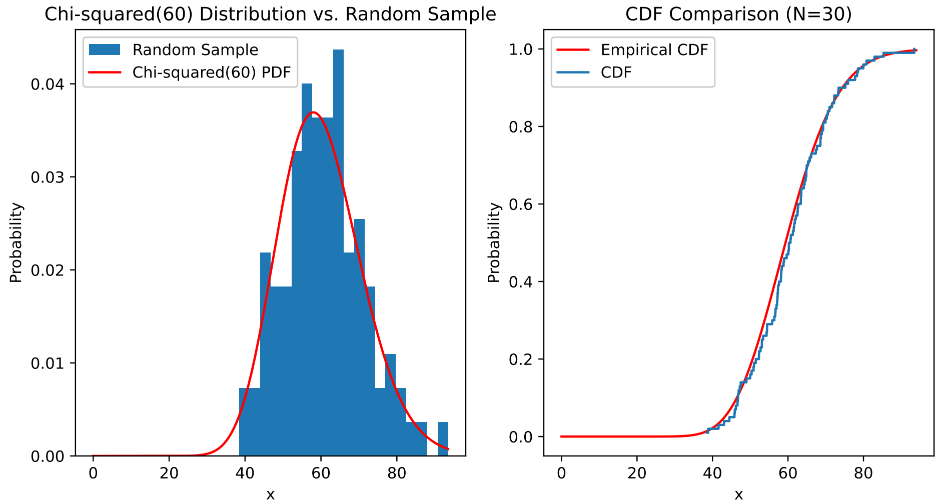

圖 22: 30 個 $exp(1)$ 樣本進行轉換

$N = 30$ 時，與理論誤差不算太大，直方圖與 CDF 圖幾乎都依靠著理論分配，故猜測隨著 N 再繼續增加的話，效果不會太明顯。

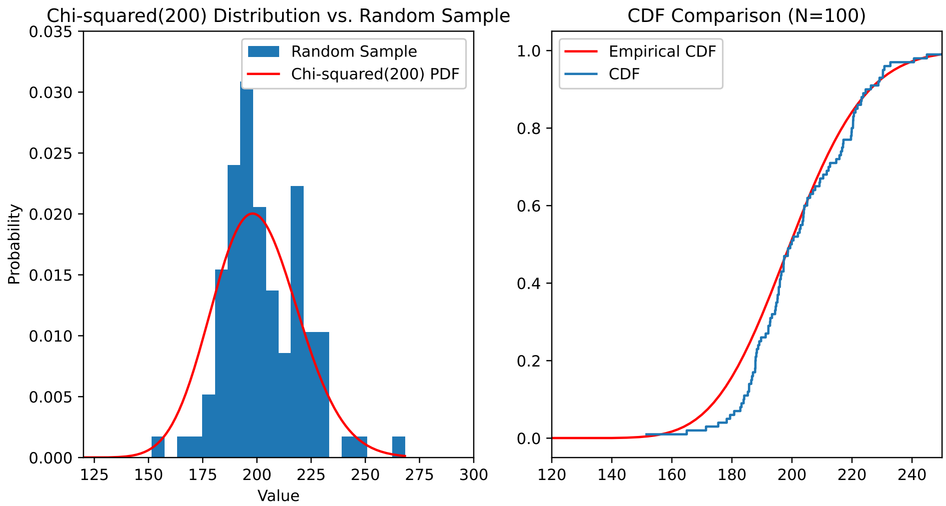

圖 23: 100 個 $exp(1)$ 樣本進行轉換

$N = 100$ 時，誤差竟然是比 $N = 30$ 時更大，本文反覆抽樣了數次，發現誤差仍然都與 $N = 30$ 時大一些，為了更了解樣本數會不會影響此抽樣分配的精準度，故將 N 增加至 1000：

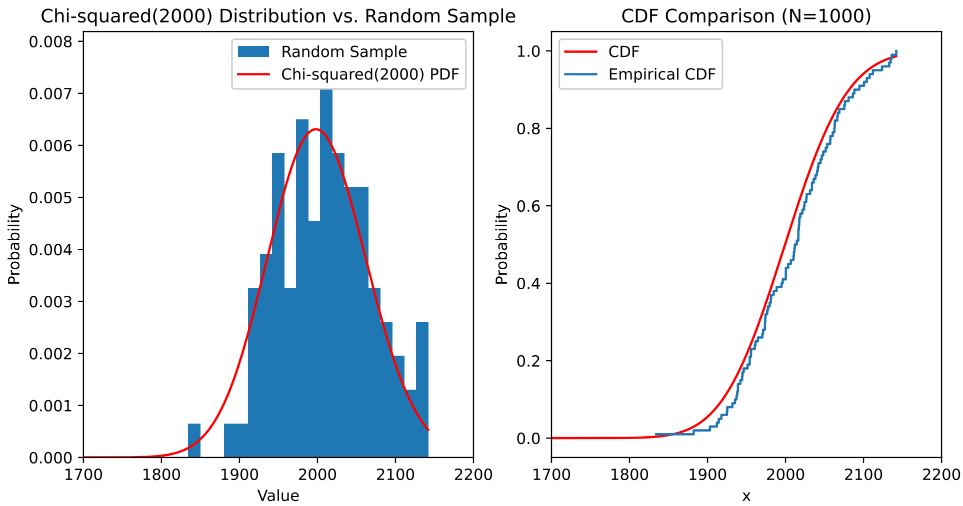

圖 24: 1000 個 $exp(1)$ 樣本進行轉換

當 $N = 1000$ 時，不像前一小節在 $N = 1000$ 時已經相當貼近理論分配，可以從 24 的直方圖與 CDF 圖看出雖然趨勢與理論分配相同，但都還是存在不小的誤差，故推論此抽樣分配在樣本數的要求上並沒有麼嚴格，其不會隨著樣本數增加，而增加精準度。

若將 3 種情況的平均數計算出來，分別為：

$$\mu_{30} = 60.84\,, \mu_{100} = 203.01\,, \mu_{1000} = 2009.34$$

可以發現三者與理論平均值的差距不會隨著 N 增加而縮小，可見平均數不需要樣本夠大即可趨近於理論值。

4.3 中央極限定理

中央極限定理為相當廣為人知的理論，其說明不論從何種分配抽取樣本，只要抽取樣本數足夠大，並抽自同一分配，則這些樣本的樣本平均數會服從常態分配，其中平均數恰為母體平均數，變異數則是母體變異數除以 n (抽取的樣本數)。以下我將產生 20,100,1000,100000 個 $exp(\lambda = 2)$ 樣本，以驗證中央極限定理的結論，確認樣本平均數是否會服從 $N(\frac{1}{2}, \frac{1}{4n})$

由圖 25、圖 26 可以看出當 n 較小時，雖然分配圖形與實際誤差較大，但還是能看出其為鐘形分布，且中心點位在 0.5 的位置。

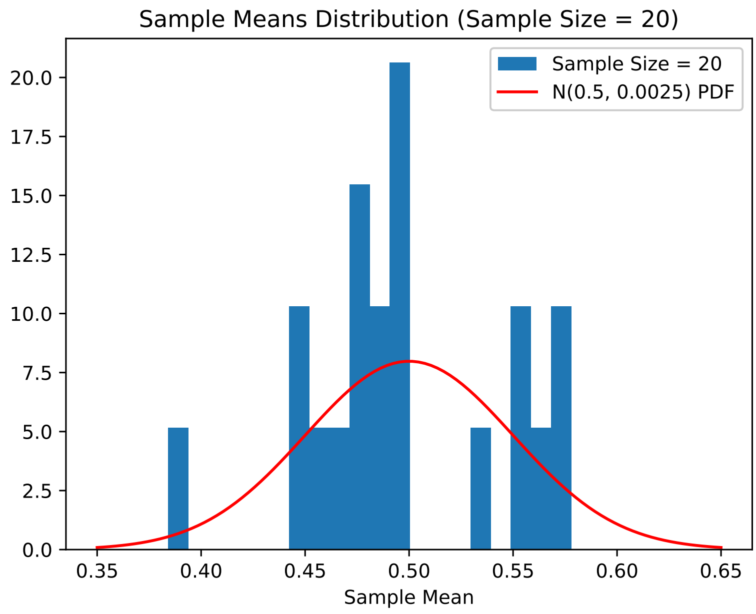

圖 25: $n = 20$

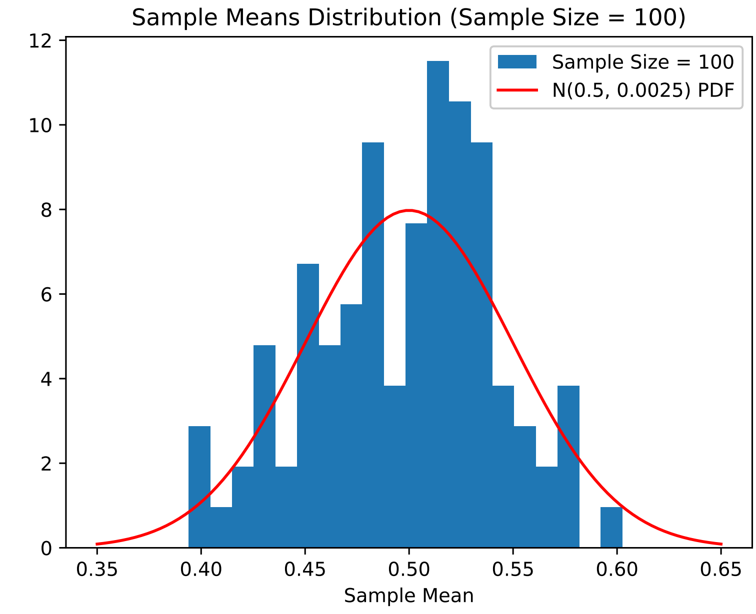

圖 26: $n = 100$

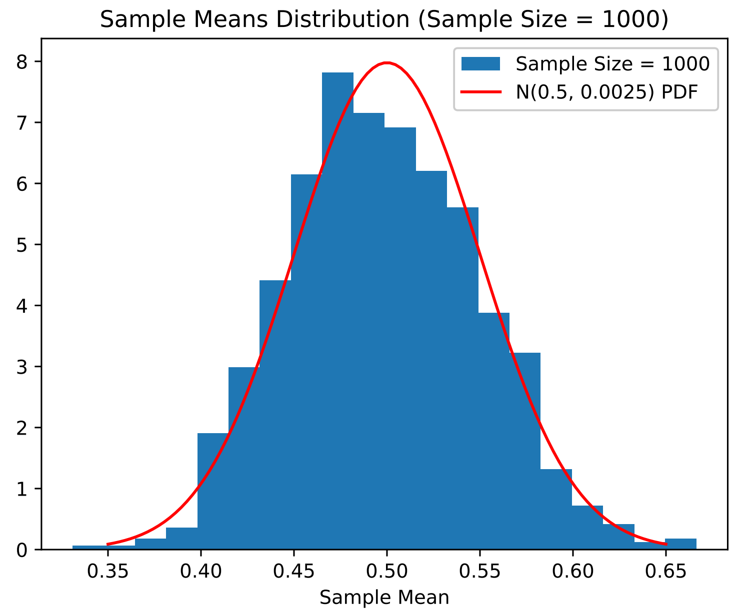

圖 27: $n = 1000$

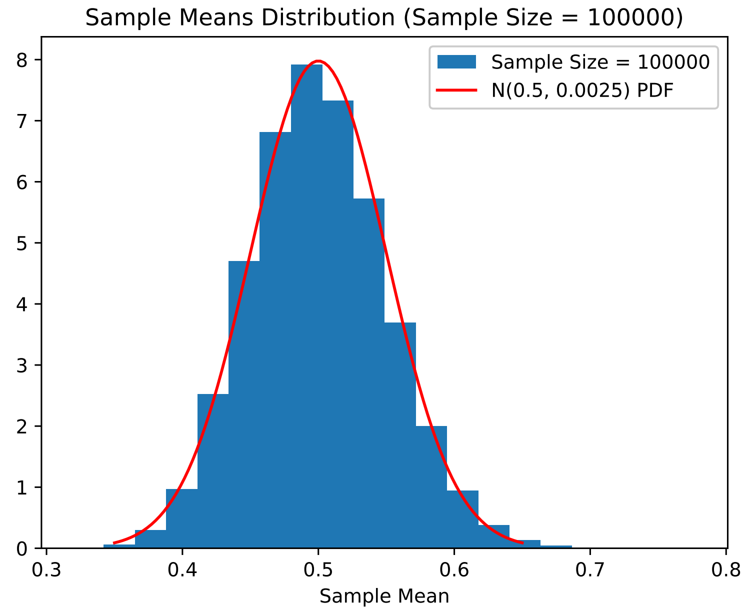

圖 28: $n = 100000$

由圖 27、圖 28 可以看出當 n 非常大時，幾乎與常態分配的 pdf 重疊，可知樣本數對中央極限定理影響相當之大，樣本數越大則理論越可靠，透過這幾張圖我們也驗證了中央分配理論是不可爭的事實。

5 數理統計題目驗證

該小節我們將探討數理統計課本上的一題習題，使用 python 驗證其結果，原題目如下與結果如下：

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。假設隨機變數 Y 代表這四個數字的平均數。請繪製隨機變數 Y 的 PMF。

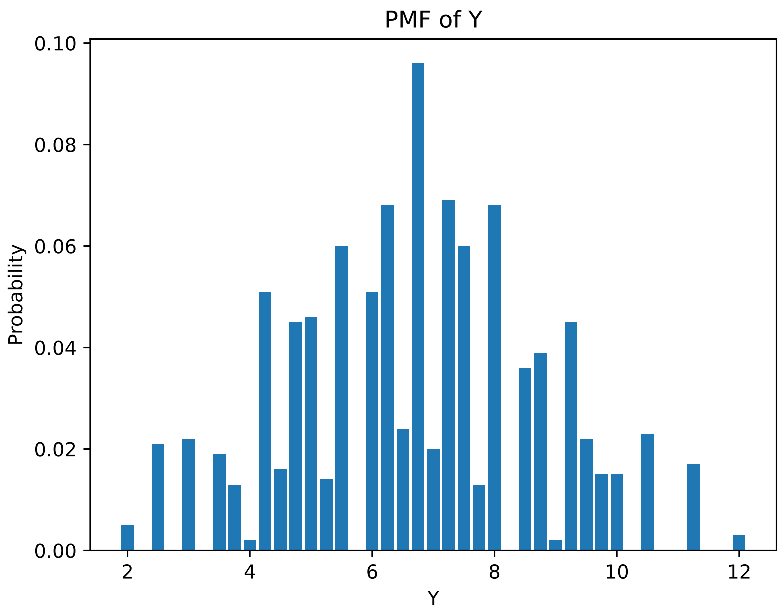

圖 29: 四個數字隨機抽樣的平均數

該題用迴圈與 unique 的語法就能產生平均數，再使用長條圖將這些平均數繪製上去，以下是該題程式碼：

```
Ys = []
n = 1000
for _ in range(n):
    sample = np.random.choice([2, 4, 9, 12], size=4,
       replace=True)
    Y = np.mean(sample)
    Ys.append(Y)

unique_Y, counts = np.unique(Ys, return_counts=True)
density = counts / n
plt.bar(unique_Y, density,width=0.2)
```

6 結論

本文使用 python 繪製了許多常見分配，可以了解改變參數對於分配的實際影響，不管是在連續型分配還是離散型分配。也透過亂數產生，繪製了理論值與亂數值之間的差異圖，從中了解亂數產生的語法以及繪製各種圖形的方式。最後我們透過實際作圖驗證了數理統計中的三個定理，經過實作確實能更清楚瞭解定理為何成立，與成立的條件，把這些結果驗證完畢之後，甚至更改了本文作者過去對數理統計的模糊的概念，收穫良多！