用機率分佈描述亂中有序的世界 - 連續型分配

簡報閱讀 範例與作業 問題討論 學習心得(完成)

生活中的機率

你知道嗎，生活中蘊藏許多的機率分佈

1000 個商品不良品的個數	大樂透，會中頭獎的機率	台灣男生的身高

上面這三種和機率分佈有關，在上完兩天後，你可以學到怎樣分辨出生活中常見問題的機率，輔助生活中大小事情的決策判斷。

重要知識點

- 生活中常見的連續分配的意義與性質
- 用模擬學習中央極限定理
- 運用python語法了解各分配的特性與應用

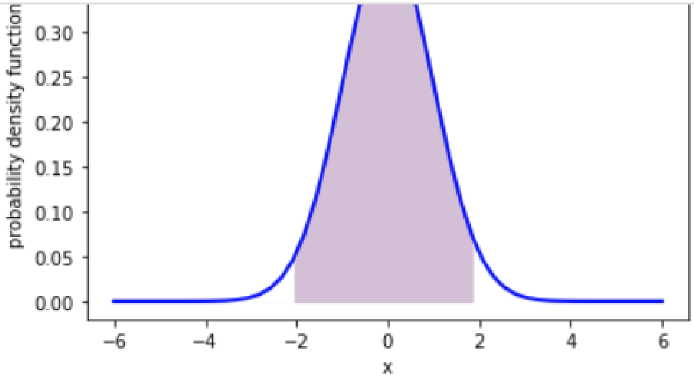
今天的課程，介紹連續型分配，包含：

- 連續型均勻分配
- 常態分配

連續型分配的定義

- 連續型機率分配之基本性質：無法列出所有實驗的可能值（不可數）
- $f(x)$ 為 X 的機率分配函數（機率密度函數），則 $f(x)$ 符合以下條件：
  1. $f(x) \geqq 0$
  2. $f(x)$ 圖形下總面積為 1
- $f(x)$ 並非機率
- 即 $f(a) \neq P(X = a)$，a 為任意數
- $f(x)$ 圖形下面積才為機率
- $P(-2 \leqq X \leqq 2)$：X 介於 [-2,2] 中間的機率為

$$\int_{-2}^{2} f(x)dx$$



連續型分配 - 平均數和變異數的定義

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

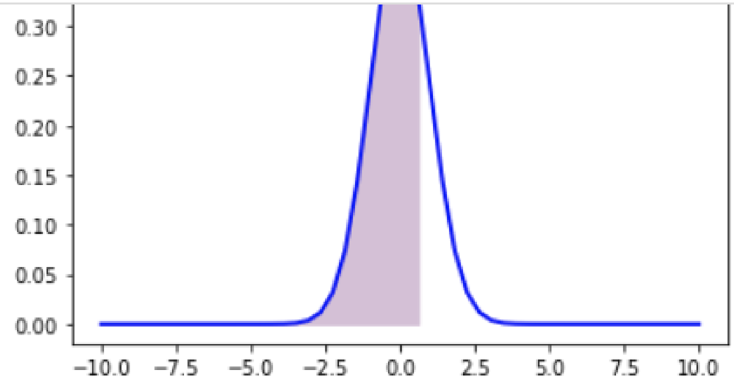
$$\sigma^2 = V(X) = E(X-\mu)^2$$
$$= E(X^2) - \mathrm{E(X)}^2$$
$$= \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2$$

累積分佈函數 (cdf)

累積分佈函數，又叫分佈函數，是機率密度函數的積分，能完整描述一個實隨機變數 X 的機率分佈。

對於所有實數 x，累積分佈函數定義如下：

- $F(x) = \mathrm{P}(X \leqq x)$
- $F(1) = \mathrm{P}(X \leqq 1)$

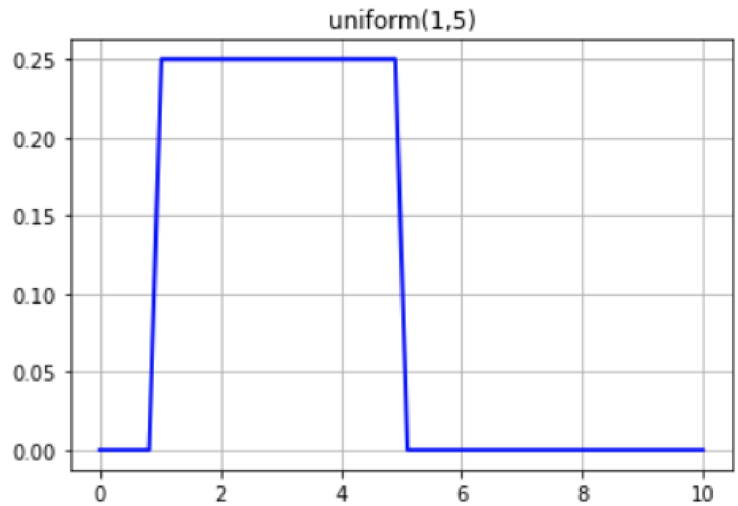


連續型均勻分配

一個均勻分佈在區間 [a,b] 上的連續型隨機變量 X，記做 X ~ U [a,b]

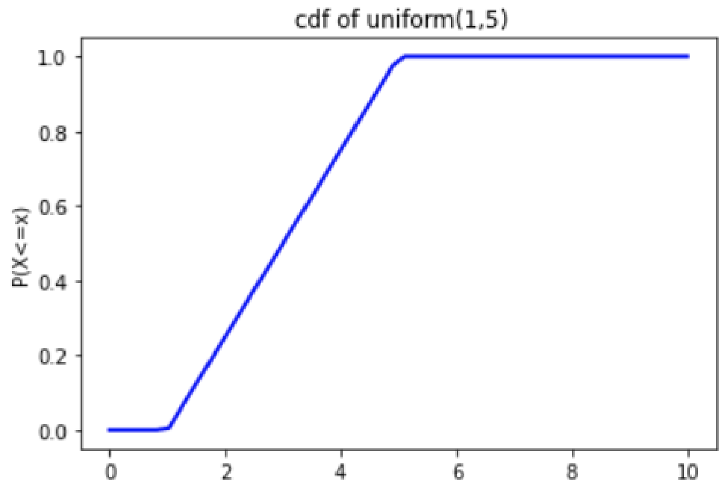
機率密度函數(pdf)：

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & \text{for a} \le x \le b \\ 0, elsewhere \end{cases}$$



累積分佈函數(cdf)：

$$F(x) = \begin{cases} \dfrac{x-a}{b-a}, for\ a \le x < b \\ 1, for\ x \ge b \end{cases}$$



連續型均勻分配的平均數和變異數為：

$$E(X) = \frac{a+b}{2}$$

$$V(X) = \frac{(b-a)^2}{12}$$

U[1,5] 的平均數和變異數

$$\mu = 3$$

$$\sigma^2 = \frac{16}{12} = \frac{4}{3}$$

這邊我們運用 python 裡面的， scipy 套件裡的 stats 相關函數進行運算。

連續型均勻分配	Python 語法 (scipy.stats)
機率質量函數	uniform.pdf
累積機率函數	uniform.cdf
樣本點	uniform.rvs
統計量計算	uniform.stats

日常生活中的常態分配

- 你知道智商的平均是甚麼分配？沒錯就是常態分配。
- 常態分配(Normal Distribution) 是統計學中最重要的一個分配，亦為連續分配中最重要的一個分配。
- 常態分佈(normal distribution)，又稱高斯分佈(Gaussian distribution)

德國鈔票高斯人像：

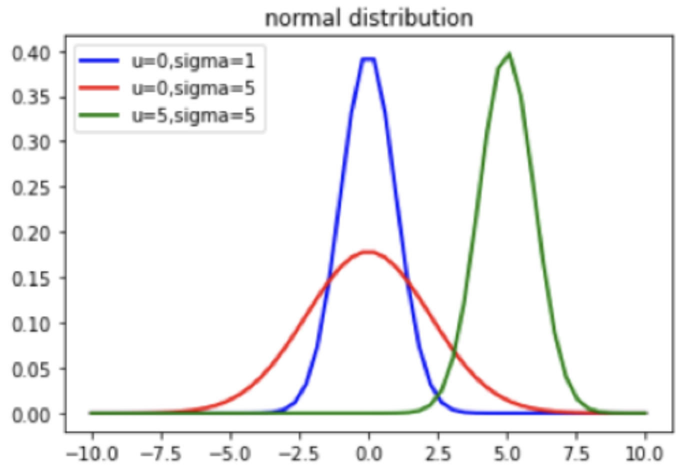
圖片資料來源：case.ntu.edu.tw

常態分配

- 機率密度函數(pdf)為

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- 平均數為 μ 決定 pdf 位置
- 標準差為 σ 決定 pdf 寬度



常態分配的平均數和變異數為：

$$\mathrm{E}(X)=\mu$$

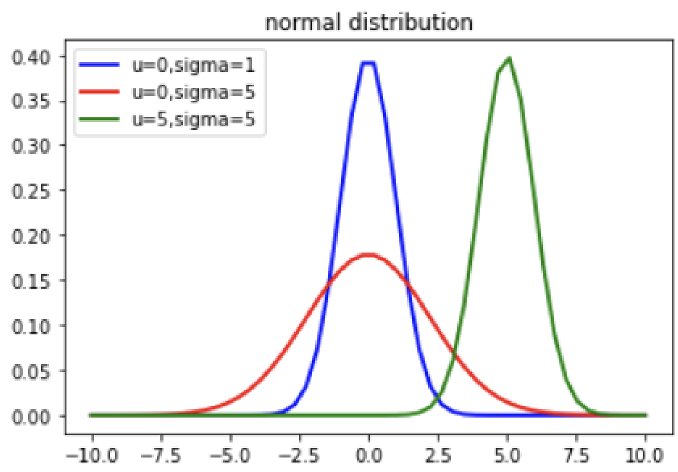
$$V(X)=\sigma^2$$

常態分配	Python 語法 (scipy.stats)
機率質量函數	norm.pmf
累積機率函數	norm.cdf
樣本點	norm.rvs
統計量計算	norm.stats

常態分配的特性

型曲線分佈。

- 觀察值之範圍為負無限大至正無限大之間。
- 變項之平均數、中位數和眾數為同一數值。
- $\mu = 0$，$\sigma^2 = 1$，稱為標準常態分配（藍色）
- 標準常態分配，記做 Z。



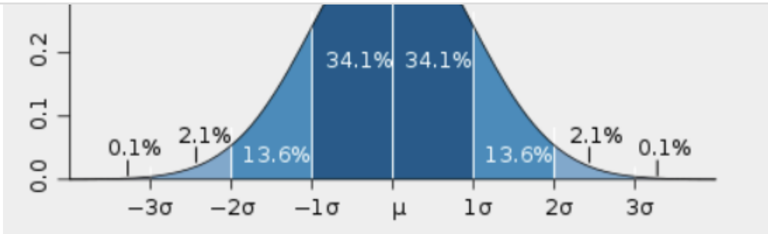
一般常態轉成標準常態分配

$X \sim N(\mu,\sigma^2)$ 透過標準化，將一般常態轉成標準常態分配，標準常態分配的符號為 Z。

$$Z = \frac{X-u}{\sigma}$$

標準常態的特性：資料點都集中在中心點的兩旁

- 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。



中央極限定理

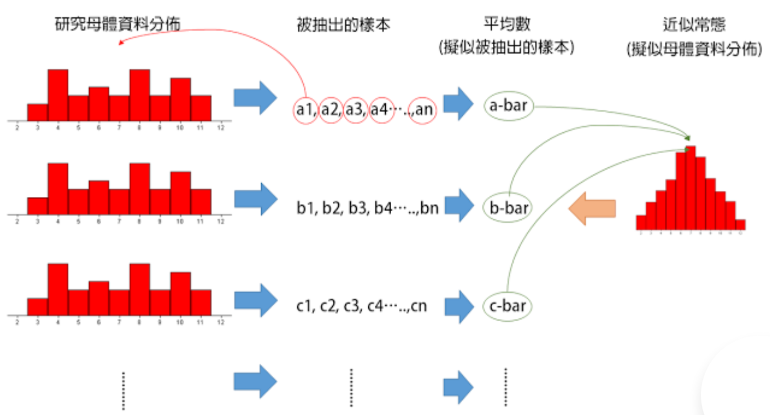
- 從平均數 μ，標準差 σ 的母體中抽樣大小為 n 的簡單隨機樣本，當樣本數 n 夠大時，樣本平均數($\bar{X}_n$) 的抽樣分配會近似於常態分配。
- 一般統計食物，大部分的應用假設 n 大於 30 時，樣本平均數($\bar{X}_n$) 的抽樣分配接近常態分配。
- 當母體為常態分配，不管 n，樣本平均數($\bar{X}_n$) 的抽樣分配皆為常態分配。

X ~ 任一分配符合

1. 平均數存在
2. 0 < 變異數 < ∞ 存在

$\bar{X}_n$ —> 常態分配

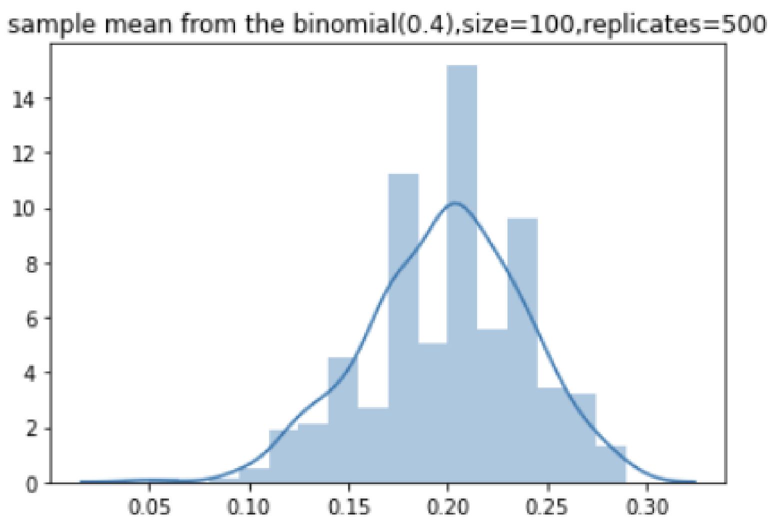
用圖說解釋中央極限定理



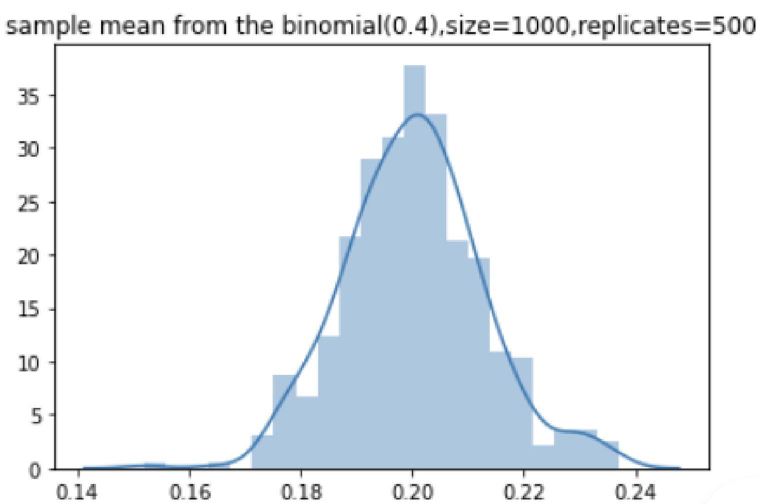
用模擬理解中央極限定理，X~bernoulli(p)

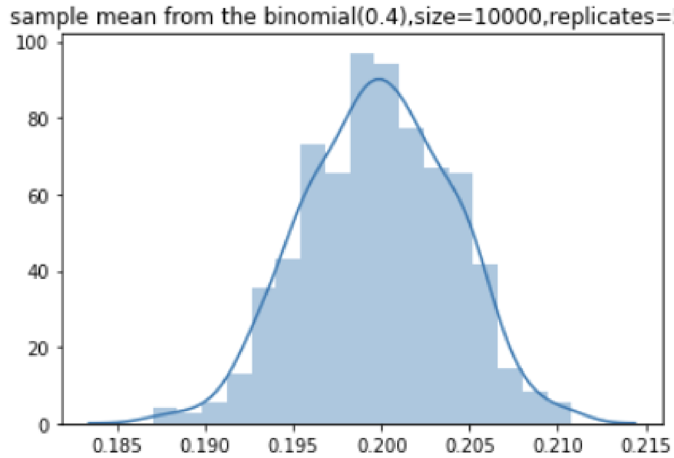
從伯努利分配中，抽取 n 個樣本，$X_1, \ldots, X_n$，重複抽了 500 次，每一次都計算 $\bar{X}_n$，長條圖呈現 500 次 $\bar{X}_n$ 結果。

$n = 100$

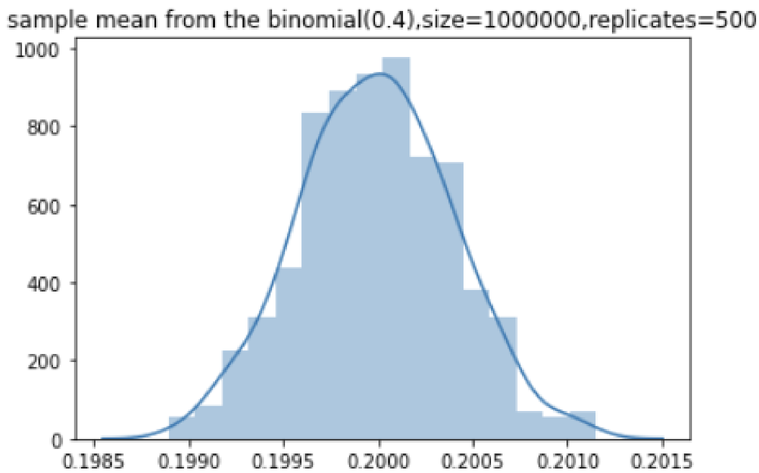


$n = 1000$





$n = 10000$



當樣本數 n 夠大時，樣本平均數($\bar{X}_n$) 的抽樣分配會近似於常態分配

今天課程學到了 - 常見連續型分配

分配	符號	分配特性
連續型均勻分配 uniform	X~U[a,b]	大自然很少連續型均勻分配
常態分配	X~N(μ, σ^2)	• 以平均數為中線，構成左右對稱之單峰、鐘型曲線分布。 • 觀察值之範圍為負無限大至正無限大之間。 • 變項之平均數、中位數和眾數為同一數值。 • μ=0, σ^2=1，稱為標準常態分配 • 具有平均數 μ，標準差 σ 的母體中抽樣樣本大小為 n 的簡單隨機樣本，當樣本數 n 夠大時，樣本平均數($\bar{X}_n$)的抽樣分配會近似於常態分配

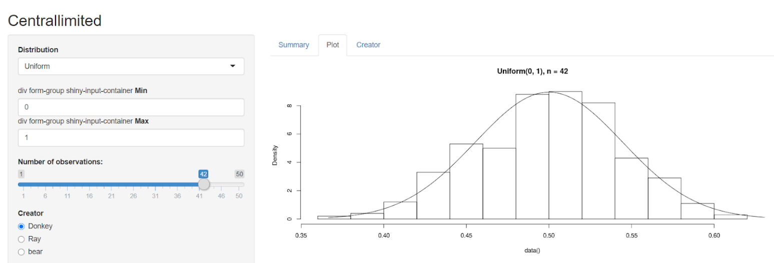
- 生活中常見的連續分配的意義與性質，包含連續均勻分配和自然界常見的常態分配。
- 用模擬學習中央極限定理，了解許多分配當收集的樣本數量夠多時，樣本平均($\bar{X}_n$) 的抽樣分配會趨近常態分配。
- 運用 python 語法了解各連續分配的特性與應用 pdf、cdf、rvs 和 stats

延伸閱讀

中央極限定理模擬軟體

網站：[Centrallimited]

動手玩玩看，原始分配不同，當抽樣的數量變多時，樣本平均數會趨近於常態分配。



scipy.stats 離散分配官方教學

- 均勻分配：[scipy.stats.randint]

and completes them with details specific for this particular distribution.

Notes

The probability mass function for randint is:

$$f(k) = \frac{1}{high - low}$$

for `k = low, ..., high - 1`.

randint takes `low` and `high` as shape parameters.

The probability mass function above is defined in the "standardized" form. To shift distribution use the `loc` parameter. Specifically, `randint.pmf(k, low, high, loc)` is identically equivalent to `randint.pmf(k - loc, low, high)`.

- 常態分配：scipy.stats.bernoulli

scipy.stats.norm¶

scipy.stats.norm(*args, **kwds) = <scipy.stats._continuous_distns.norm_gen object> [source]

A normal continuous random variable.

The location (`loc`) keyword specifies the mean. The scale (`scale`) keyword specifies the standard deviation.

As an instance of the rv_continuous class, norm object inherits from it a collection of generic methods (see below for the full list), and completes them with details specific for this particular distribution.

Notes

The probability density function for norm is:

$$f(x) = \frac{\exp(-x^2/2)}{\sqrt{2\pi}}$$

for a real number x.

The probability density above is defined in the "standardized" form. To shift and/or scale the distribution use the `loc` and `scale` parameters. Specifically, `norm.pdf(x, loc, scale)` is identically equivalent to `norm.pdf(y) / scale` with `y = (x - loc) / scale`. Note that shifting the location of a distribution does not make it a "noncentral" distribution; noncentral generalizations of some distributions are available in separate classes.

下一步：閱讀範例與完成作業