D05 NumPy 統計函式 Universal Functions (ufunc)



**重要知識點**



陣列的統計功能分為四大類，今天的內容會依照這四大類介紹相關的函式及使用。

* 順序統計量 (Order Statistics)
* 平均數與變異數
* 相關性
* 直方圖 (Histogram)

**NumPy 陣列統計函式 - 順序統計量**

**numpy.maximum(), numpy.minimum()**

* 以 element-wise 比較 2 個陣列並回傳各元素的最大值或最小值。如果比較的元素中的 nan 的話，則會回傳 nan。
* maximum() 與 minimum() 在進行比較時，若有需要會利用到廣播 (bradcasting)。

**numpy.fmax(), numpy.fmin()**

* 以 element-wise 比較 2 個陣列並回傳各元素的最大值或最小值。與 maximum() / minimum() 不同的是，如果比較的元素中只有一個是 nan 的話，回傳非 nan 的值，如果兩個元素都是 nan 則回傳 nan。
* 同樣在進行比較時，若有需要會利用到廣播 (bradcasting)。

**numpy.nanmax(), numpy.nanmin()**

* 回傳陣列中有非 nan 元素值的最大值或最小值。
* 可以指定要比較的軸，以及回傳值是否要保留維度。

**百分位數：percentile(), nanpercentile()**

* 計算百分位數，percentile() 與 nanpercentile() 不同的地方在於後者會忽略 nan。
* 欲取得的百分位數引數，可以傳入純量或是陣列的值 (介於0 - 100 之間)，也可以指定要比較的軸，以及回傳值是否要保留維度。

**分位數：quantile(), nanquantile()**

* 計算分位數，quantile() 與 nan quantile() 不同的地方在於後者會忽略 nan。如果元素中包含 nan 的話，則 quantile() 會回傳 nan。
* 欲取得的分位數引數，可以傳入純量或是陣列的值 (介於0 - 1 之間)，也可以指定要比較的軸，以及回傳值是否要保留維度。

**NumPy 陣列統計函式 - 平均數與變異數**

**平均值：mean(), nanmean()**

* mean() 和 nanmean() 不同的地方在於後者會忽略 nan。如果元素中包含 nan 的話，則 mean() 會回傳 nan。下面的例子使用 np.isnan() 判斷陣列中是否包含 nan，如果無 nan 的話就呼叫 mean() 計算平均值，反之則呼叫 nanmean() 進行計算。
* 可以指定要計算平均數的軸，以及回傳值是否要保留維度。dtype 引數是計算使用的型別，若輸入陣列是整數的話，則會用 float64 型別計算，若輸入的是浮點數的話，則是依輸入陣列的型別做為 dtype。

**平均值：average()**

* 使用 average() 計算平均值的話，可以輸入權重值做為引數。
* 須注意權重的總和不能為 0，否則會產生錯誤。

**計算中位數：median()，nanmedian()**

* median() 和 nanmedian() 不同的地方在於後者會忽略 nan。如果元素中包含 nan 的話，則 median() 會回傳 nan。
* 可以指定要計算中位數的軸，以及回傳值是否要保留維度。要留意的是，如果軸或是陣列總數不是單數的話，中位數的值會是中間 2 個元素值相加除以 2。

**計算標準差：std()，nanstd()**

* std() 和 nanstd() 不同的地方在於後者會忽略 nan。如果元素中包含 nan 的話，則 std() 會回傳 nan。
* 可以指定要計算標準差的軸，以及回傳值是否要保留維度。若是對於精度可能造成的誤差影響，可以改變 dtype 提高精度。
* 如果要計算樣本標準差的話，可將 ddof (自由度) 引數傳入 1，在計算平均方差 (mean squared deviation) 時分母就會以 N - ddof 做計算。

**計算變異數：var()，nanvar()**

* var() 和 nanvar() 不同的地方在於後者會忽略 nan。如果元素中包含 nan 的話，則 var() 會回傳 nan。
* 可以指定要計算變異數的軸，以及回傳值是否要保留維度。若是對於精度可能造成的誤差影響，可以改變 dtype 提高精度。
* 如果要計算樣本變異數的話，可將 ddof (自由度) 引數傳入 1，在計算平均方差 (mean squared deviation) 時分母就會以 N - ddof 做計算。

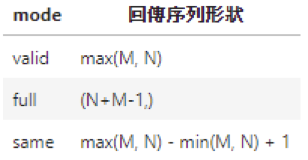
**NumPy 陣列統計函式 - 相關性**

**相關係數：corrcoef()**

* corrcoef() 計算 Pearson 積差相關係數。引數 rowvar 預設值為 True，代表將每一個 row 當做是一筆變數。

**互相關 (Cross-correlation)：correlate()**

* 計算 2 個一維序列的互相關。mode 引數及回傳序列形狀如下表：



* *N*為第*1*個序列的元素數，*M*為第*2*個序列的元素數。

**共變異數：cov()**

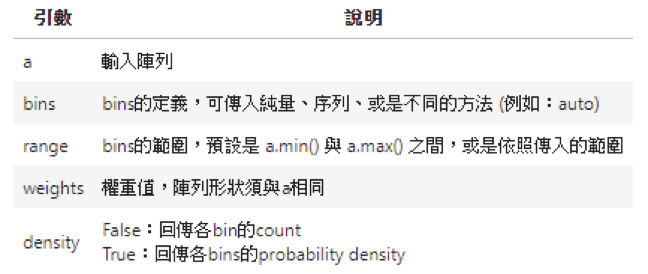
* 函式引數說明如下：



**NumPy 陣列統計函式 – 直方圖 (Histogram)**

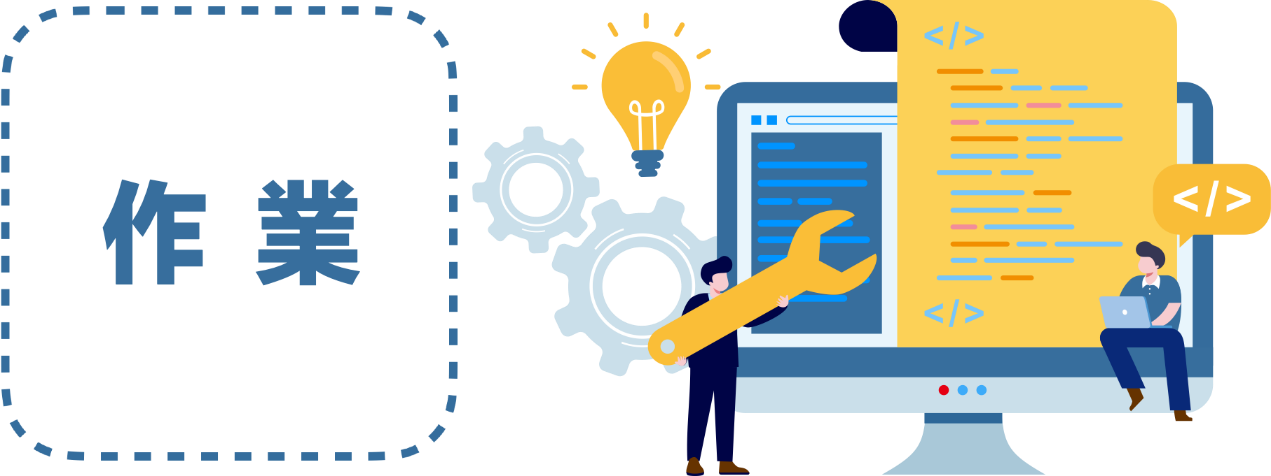
**NumPy 提供 np.histogram() 函式來計算 histogram，基本語法及引數說明如下：**

* numpy.histogram(a, bins=10, range=None, normed=None, weights=None, density=None)



**知識點回顧**

數學及統計運算是 NumPy 最主要的功能，今天介紹 NumPy 統計四大分類及說明各個函式的使用，請照範例程式碼提供的函式運用示範。



題目：

english\_score = np.array([55,89,76,65,48,70])

math\_score = np.array([60,85,60,68,np.nan,60])

chinese\_score = np.array([65,90,82,72,66,77])

上 3 列共六位同學的英文、數學、國文成績，第一個元素代表第一位同學，舉例第一位同學英文 55 分、數學 60 分、國文 65 分，今天第五位同學因某原因沒來考試，導致數學成績缺值，運用上列數據回答下列問題。

1. 請計算各科成績平均、最大值、最小值、標準差，其中數學缺一筆資料可忽略？
2. 第五位同學補考數學後成績為 55，請計算補考後數學成績平均、最大值、最小值、標準差？
3. 用補考後資料找出與國文成績相關係數最高的學科？