

07 Probability and Statistics 統計學

期中測驗

- 採網路線上測驗
- 時間: 29 日 下午 2:00 - 11:59
- 範圍: 以習題為範圍

統計分析是以數據為基礎，對數據進行科學處理、分析。統計分析包括兩大部分

1. 描述統計(Descriptive Statistics)
2. 推論統計(Inferential Statistics)

數據類型

- 定性數據(Qualitative Data)：對事物性質進行描述數據。ex 股票所屬行業數據即為定性數據。ex 國巨屬於半導體業。
- 定量數據(Quantitative Data)：事物數量特徵的數據，由數字所組成。ex 每個股票的開盤價。

圖表

當拿到數據後拿來分析，最常使用圖表的方式呈現數據。有以下圖表可以呈現數據。

- [直方圖](#)：二維統計圖表，橫軸是類別或是數值間的範圍，縱軸是事件發生的頻率。
- [圓餅圖](#)：圓形統計圖表，每個扇區的弧長表示的數量的比例。
- [折線圖](#)：折線圖是用直線段將各數據點連接起來而組成的圖形，以折線方式顯示數據的變化趨勢。
- [散佈圖](#)：兩個變數之間關係的圖，又稱相關圖
- [頻率分布表](#)：統計學中表示樣本數據頻率分布規律的表格。

數據的位置

分析數據時，想要了解數據分布的位置，在統計分析中有專門的指標用於描述數據的位置。常用的指標有以下四種方式：

- 樣本平均數(Sample Mean)
- 中位數(Median)
- 眾數(Mode)
- 百分位數(Percentile)

樣本平均數(Sample Mean)

假設現在有 n 個樣本觀測值(X_1, X_2, \dots, X_n)，在統計學中樣本平均數有兩種計算方式：

- 算術平均數(Arithmetic Mean)

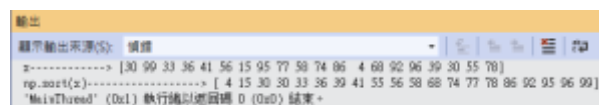
$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

- 幾何平均數(Geometric Mean)

$$G_n = \sqrt[n]{\prod_{i=1}^n x_i} = \sqrt[n]{x_1 x_2 x_3 \cdots x_n}$$

使用亂數產生一個一維陣列

```
import numpy as np  
  
x = np.random.randint(0, 100, size = 20)  
  
print('x----->', x)  
  
print('np.sort(x)----->', np.sort(x))
```



```
輸出  
顯示輸出來源(S): 選擇  
x-----> [30 99 33 36 41 56 15 95 77 58 74 86  4 68 92 96 39 30 55 78]  
np.sort(x)-----> [ 4 15 30 30 33 36 39 41 55 56 58 68 74 77 78 86 92 95 96 99]  
'MainThread' (0x1) 執行緒以退還碼 0 (0x0) 結束。
```

平均值

```
print('x.mean()----->', x.mean()) # 平均值
```

中位數(Median)

將樣本從小到大做排序，如果樣本是奇數個中位數即為最中間的值，如果是偶數個，中位數是中間兩個數值的平均。

```
print('np.median(x)---->', np.median(x)) # 中位數(Median)
```

眾數(Mode)

樣本中出現次數最多的數值。

```
from scipy import stats  
print('stats.mode(x)----->', stats.mode(x)[0]) # 眾數(Mode)
```

數據的離散度

也稱為數據的變異性，常用的離散度指標有：

- 全距(Range): 全距 = 最大值 - 最小值
- 平均絕對偏差(Mean Absolute Deviation): 樣本與平均值(Mean)的差值來計算，當差值越大則表示數據值偏離均值越遠
- 變異數(Variance): 變異數也是用來描述數據的離散程度，標準差就是變異數得平方根
- 標準差(Standard Deviation)

```
import pandas as pd  
ironman = pd.Series([60, 70, 80, 80, 90, 100])  
print(ironman.var())  
print(ironman.std())
```

隨機變數(Random Variable)

統計的母體是某個要預測事物的所有可能發生結果的集合，隨機變數則是一個不確定性事件結果的數值函數(Function)，也就是說把不確定性事件的結果用數值來表示，即為隨機變數。根據隨機變數可能取值的結果，可以分成兩種：

1. 離散型隨機變數(Discrete Random Variable)：隨機變數取值的範圍有限個(ex 擲硬幣只有兩種結果，所以取擲的範圍有限制)
2. 連續型隨機變數(Continuous Random Variable)：隨機變數可以在一個區間上任意取值(ex 股票中的收益率就是連續型隨機變數)

離散型隨機變數(Discrete Random Variable)

設 x 是一個隨機變數，如果它全部可能的取值只有有限個，則稱 x 為一個離散型隨機變數。

假設現在有 n 個數據，每個數據發生的機率即為概率質量函數(Probability Mass Function)：

$$P(X = x_i) = P_i, i = 1, 2, \dots$$

可以使用 Python 中的 `choice()` 來產生特定機率隨機數：

```
import numpy as np
import pandas as pd

ironman = np.random.choice([1,2,3,4,5], size=1000, p=[0.25,0.1,0.35,0.2,0.1])
valueCount = pd.Series(ironman).value_counts()
print('valueCount/1000---->', valueCount/1000)

valueCount/1000----> 3    0.354
1    0.243
4    0.184
5    0.118
2    0.101
dtype: float64
```

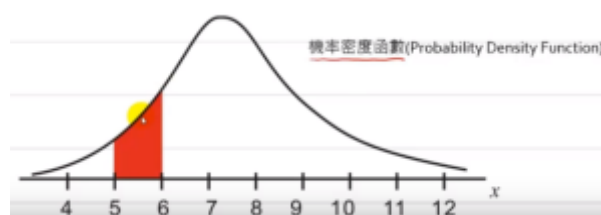
連續型隨機變數(Continuous Random Variable)

即在一定區間內變數取值有無限個，或數值無法一一列舉出來。用來表示某個區間內每個數值所發生的機率為

機率密度函數(Probability Density Function)

$$\int_{-\infty}^{\infty} f(y) dy = \int_a^b f(y) dy = 1$$

以下圖為例，在 5~6 發生的機率即為紅色的面積



範例: 鐵達尼號

<https://www.kaggle.com/c/titanic/data>

匯入資料與轉換成為 DataFrame

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import scipy.stats as stats
PD_DATA = pd.read_csv("train.csv", usecols=['Age'])
print(PD_DATA)
```

資料原始分佈情況

- 使用 鐵達尼號資料 (Titaniccsv), 繪製直方圖
- 資料是否為常態分配

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import pylab
import scipy.stats as stats
PD_DATA = pd.read_csv("train.csv", usecols=['Age'])print(PD_DATA)
def diagnostic_plots(df, variable):
    plt.figure(figsize=(15,6))
    plt.subplot(1,2,1)
    df[variable].hist()
    plt.subplot(1,2,2)
    stats.probplot(df[variable], dist="norm", plot = pylab)
    plt.show()
diagnostic_plots(PD_DATA, 'Age')
```