

D26 用統計描述資料的樣態

[簡報閱讀](#)[範例與作業](#)[問題討論](#)[學習心得\(完成\)](#)

>

為什麼需要敘述統計？

>

重要知識點

>

母體 (population) 與 樣本 (Sample) 的關係

>

透過統計量，知道目前真相調查的現況

>

統計量的總類

>

集中趨勢 – 平均數

>

集中趨勢 – 中位數



為什麼需要敘述統計？

為了解台灣男生和女生在身高上誰比較高，收集到 20 位男生和 20 位女生的身高，資料如下，肉眼看原始資料，無法很快速的比較出結果。

男生身高

透過莖葉圖，快速紙筆計算
中位數和眾數

集中趨勢 - 確定一組數據的
均衡點

樣本平均數與樣本中位數，
那一個才是一組樣本數據...

176	159	165	165
169	151	156	163
169	144	170	177
165	160	164	171

女生身高

169	170	162	154
183	173	169	167
170	185	162	175
168	151	181	170
182	156	159	160

拿一張白紙，試著回答上述兩個問題，在開始今天的課程。

Q1：你會怎麼角度描述你看到資料後的結果？

Q2：請問男生和女生在身高上誰比較高？

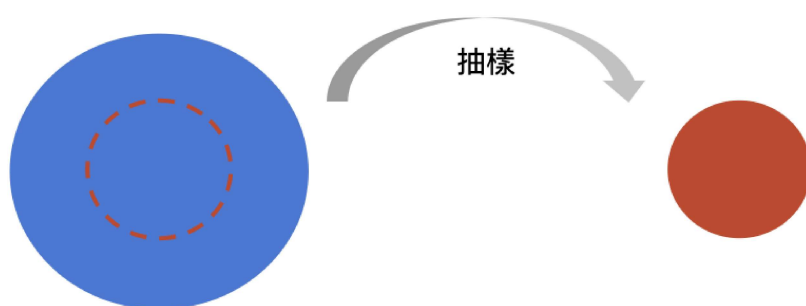
重要知識點

重要知識點

- 能判別你手上拿的資料是母體還是樣本
- 挑選適當統計量，運用python，描述出資料的輪廓進行分析。

母體 (population) 與 樣本 (Sample) 的關係

無法檢測所有的母體的特性，所以用局部的方式來觀察全部的特性，取局部的資料點的科學方法，稱作**抽樣方法**，所以每一次抽都會抽出不一樣的資料點。



母體：

- 台灣的所有男生和女生的身高

樣本：

- 樣本是母體的一部分
- 透過不同的抽樣方法，抽取出的值

所以樣本，是我們觀察到的**現象**，而統計推論的方法，就是一種以小窺大的一門技術，希望得到母體的真實狀況，也就是**真相**，透過**現象(樣本)**企圖了解**真相(母體)**。

透過統計量，知道目前真相調查的現況

- 統計量很陌生？
- 沒錯，熟知的平均數、眾數都是一種統計量。

x_1 ←	164	175	183	173
	176	159	165	165
	169	151	156	163
	169	144	170	177
	165	160	164	171 → x_{20}

統計量：

- 描述一變數或樣本之特徵的數值，樣本數的函數。
- 可以讓人類可以很快了解資料的分布和訊息。

統計量的總類

- 敘述性（描述性）統計 (descriptive statistics)，為了能全面性的瞭解資料的特

- 假設現在有一組樣本， x_1, x_2, \dots, x_n ，已經被確定抽出來的數值為 x_1, x_2, \dots, x_n

大寫是未被抽取出來的樣本代表符號

小寫是透過抽樣方法得到樣本的數值

Python 有三種語法來計算統計量：

- Python內建的 statistics
- Numpy
- Stats (from scipy)

集中趨勢 – 平均數

樣本平均數（簡稱平均數）：

$$\bar{x} = \frac{x_1 + \dots + x_n}{n}$$

164	175	183	173
176	159	165	165
169	151	156	163
169	144	170	177
165	160	164	171

$$\frac{\sum_{i=1}^{20}(x_1 + \dots + x_{20})}{20} = 165.95$$

集中趨勢 - 中位數

定義：數據中有一半小於中位數，一半大於中位數。

Step 1：首先必須將標誌值按大小排序

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \dots \leq x_{(n)}$$

Step 2：

$$M = \begin{cases} \frac{x_{(n+1)}}{2}, & \text{if } n = \text{奇數} \\ \frac{x_{\frac{n}{2}} + x_{(\frac{n}{2}+1)}}{2}, & \text{if } n = \text{偶數} \end{cases}$$

集中趨勢 - 眾數

定義：指一組數據中出現次數最多的數據值，眾數也是最容易取樣到的數據。

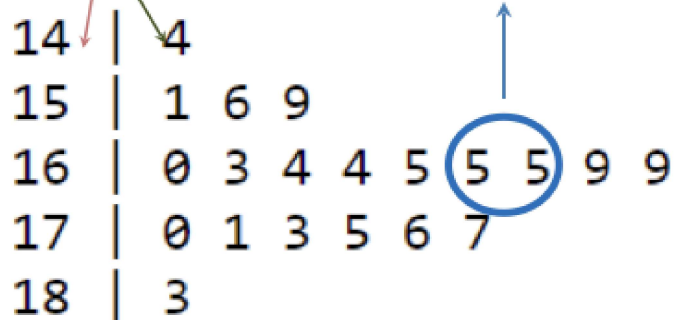
例子：{2,3,3,3} 中，出現最多的是3，因此眾數是3，眾數可能是一個數，但也可能是多個數。

164	175	183	173
176	159	165	165
169	151	156	163
169	144	170	177
165	160	164	171

中位數 = 165

- 20為偶數， $20/2=10$ ，所以按照大小排後，第10個位置和11個位置的平均即為中位數。

此為莖葉圖



眾數 = 165

```

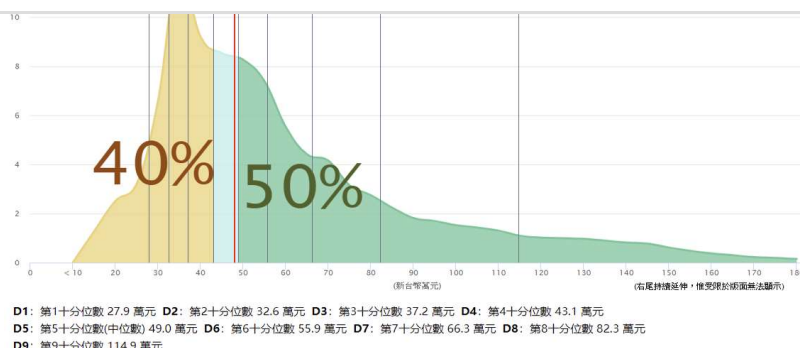
14 | 4
15 | 1 6 9
16 | 0 3 4 4 5 5 5 9 9
17 | 0 1 3 5 6 7
18 | 3
    
```

集中趨勢 - 確定一組數據的均衡點

統計量	應用(優點)	缺點	Python 語法
平均數 (Mean)	<ul style="list-style-type: none"> 平均數適合用於數值型資料。 進行不同組資料的比較，以看出組與組之間的差別。 	<ul style="list-style-type: none"> 不能用於分類資料和順序資料。 樣本平均數不是一個強健的 (robust) 統計量，容易受離群值影響而劇烈變 	<ul style="list-style-type: none"> <code>np.mean(boys)</code> <code>statistics.mean(boys)</code>
中位數 (Medium)	<ul style="list-style-type: none"> 較不受極端值影響，變化較大的資料，用中位數呈現較佳。 	<ul style="list-style-type: none"> 只利用了部分數據，可靠性比較差 	<ul style="list-style-type: none"> <code>np.median(boys)</code> <code>statistics.median(boys)</code>
眾數 (Mode)	<ul style="list-style-type: none"> 資料有很大的變動,且某個數據出現的次數最多,適用眾數來描述資料 	<ul style="list-style-type: none"> 只利用了部分數據，可靠性比較差 可能不存在或存在多個 	<ul style="list-style-type: none"> <code>stats.mode(boys)</code> <code>statistics.mode(boys)</code> 統計量的眾數，如果有多個眾數，取最小的值當眾數

樣本平均數與樣本中位數，那一個才是一組樣本數據的理想量測集中程度？

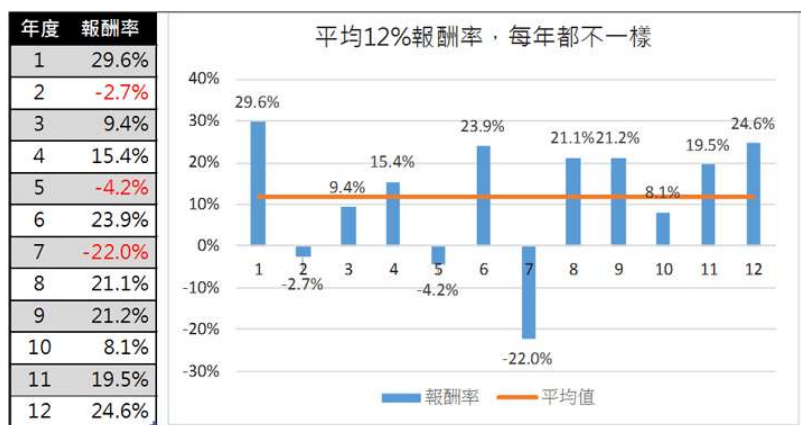
- 隨著世界的變化愈來愈大，可預見的未來，中位數在各領域將會變得愈來愈重要。
- 107年國人薪資分布，不是均勻分配，使用中位數人民比較有感。



資料來源：[薪資平台](#)

理專告訴你平均12%的報酬，就是穩定獲利？

只透過資料的集中程度來做決策，是很危險的事情，還需要哪些指標輔助？



圖表來源：[理財專家沒告訴你的真相](#)

離散程度 – 全距

定義：最大值與最小值

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \dots \leq x_{(n)}$$

Step 2 :

$$\text{全距 (Range)} = x_{(n)} - x_{(1)}$$

離散程度 – 變異數

定義：變異數即在量測所有資料到平均數的平均距離

假設假設現在有一組樣本， x_1, x_2, \dots, x_n ， \bar{x} 為此組樣本資料之平均數，則樣本變異數(s^2)為

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(n - 1)}$$

$$\text{標準差} = \sqrt{\text{變異數}}$$

離散程度 – 百分位數 (Percentile)

定義：如果將一組數據從小到大排序，併計算相應的累計百分位，則某一百分位所對應數據的值就稱為這一百分位的百分位數，以 P_k 表示第k百分位數。

99人。

求男生身高的20百分位數(P20)？

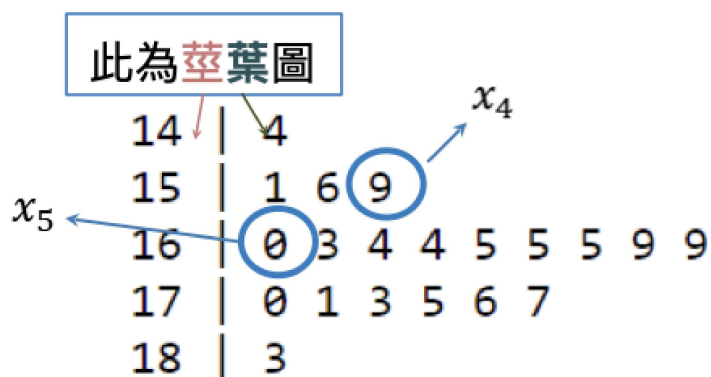
164	175	183	173
176	159	165	165
169	151	156	163
169	144	170	177
165	160	164	171

Step1：先求排序的位置：

- 位置 = $1 + (n-1) \times p$
- $1 + 19 \times 0.2 = 4.8$

Step2：取 x_4 和 x_5 數值，內插法換算出數值：

$$x_4 + (x_5 - x_4) \times (4.8 - 4) = 159.8$$



離散趨勢 - 確定一組數據分散程度

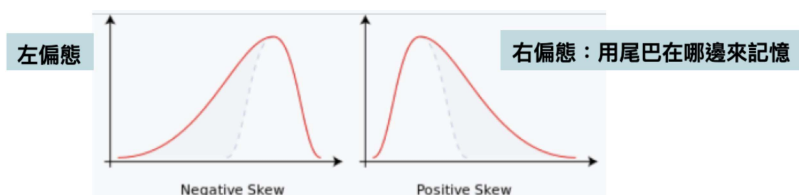
變異數 (Variance)	反映數據集的偏離程度。當標準差較大時，表示大部分數值與其平均值之間差異較大，反之代表這些數值較接近平均值。		<code>print(rangeV(boys))</code> <code>np.var(boys, ddof=1)</code> <code>statistics.variance(boys)</code> <small>*ddof=1，代表計算樣本</small>
百分位數 (Percentile)		反映較多數據的離散程度。不過其在使用中需要樣本量大才會穩定	<code>np.percentile(boys, 20)</code> <code>stats.scoreatpercentile(boys, 20)</code>

分布型態 - 偏度 (Skewness)

衡量資料分佈的不對稱性與偏斜方向程度，偏度分為兩種：

- **負偏態或左偏態**：左側的尾部更長，分布的主體集中在右側，左側有較多極端值。
- **正偏態或右偏態**：右側的尾部更長，分布的主體集中在左側，右側有較多極端值，日常生活數據常見右偏態分布。

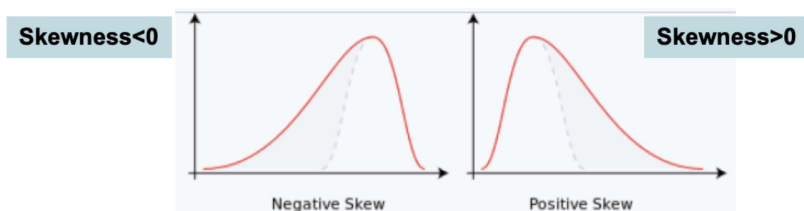
如果分布對稱，那麼平均值=中位數，偏度為零（此外，如果分布為單峰分布，那麼平均值=中位數=眾數）。



資料來源：[Skewness](#)

樣本偏度的公式如下：

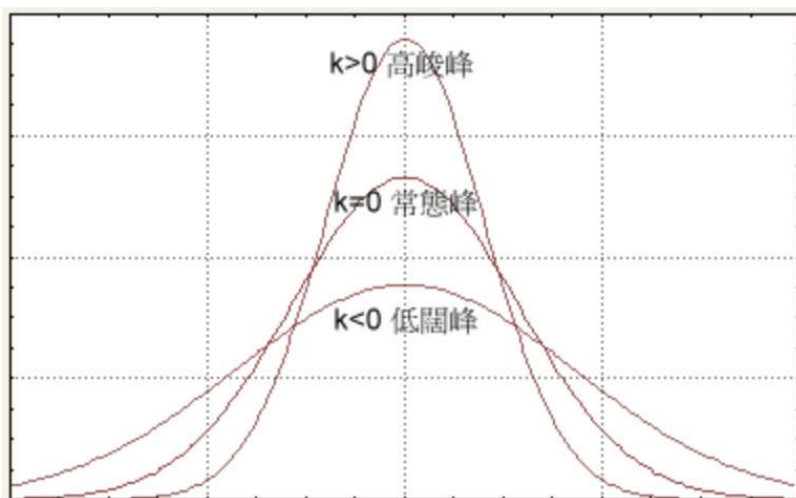
$$skewness = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{3/2}}, \bar{x} \text{ 是樣本平均數}$$



資料來源：[Skewness](#)

分布型態 - 峰度 (Kurtosis)

峰度 (Kurtosis) 資料分佈的峰態，峰度反映了峰部的尖度，也代表變異數的來源來自於不常出現的尾巴兩端的數值。



資料來源：[Kurtosis](#)，[峰態係數 \(Kurtosis\)](#)

樣本峰度的公式：

$$\text{Kurtosis} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^2} - 3$$

離散趨勢 - 確定一組數據分布的均勻程度

統計量	應用(優點)	Python 語法
偏度 (Skewness)	衡量資料分佈的不對稱性與偏斜方向程度，偏態分佈，眾數的代表性比均值好	<code>stats.skew(data)</code>
峰度 (Kurtosis)	一個分配如果「峰度」越大，表示資料中越可能會出現數值很極端的離群值	<code>stats.kurtosis(data)</code>

知識點回顧

- 分配圖
- 拿到資料後，要先判斷：
 - 母體 vs 樣本
- 挑選適當統計量，描述出資料的輪廓
- 敘述統計值，可以快速知道資料的特徵，進行簡單的比較，但只是現象，不是真相。

參考資料

為什麼標準差，要除以 $(n-1)$ 不是 n

網站：[從標準差除以 \$n\$ 或除以 \$n - 1\$ 談起\(丁村成\)](#)

丁村成

1. 前言

根據民國八十四年教育部頒佈的高級中學數學課程標準，所編寫出的教科書自八十八年九月開始使用。當初大家對統計教材中「標準差是除以 n 或 $n-1$ 」的疑問，在國立編譯館的主導之下，現行版本一律選取了除以 $n-1$ 的情形。如今，雖然教師與學生都已經默默的接受，但是否代表在教與學已經沒有任何爭議了呢？值得我們進一步反思。筆者也藉此機會，探討這一批新教材存活下來的六種教科書，為什麼會找不到一本獨具創意的版本？其問題的癥結也將在文章最後做扼要說明。在新課程標準修訂已接近完成之際，即將有新教材要在九十五年開始實施，筆者願以參與教學的實際經驗，提出最誠摯具體的建議，給下一波要編寫高中數學教科書的專家學者們參考。

2. 從高觀點看標準差之定義

統計學是關於數據資料之收集、整理、分析和推論的一門學科，其內容可區分為敘述統計學 (descriptive statistics) 和推論統計學 (inferential statistics) 兩大部分。敘述統計學在探討數據的收集、資料的整理與描述等。如果研究中可以得到整個母體 (population) 資料 X_1, X_2, \dots, X_N ，那麼其分佈狀況即已完全獲得掌握。我們特別有興趣的母體平均數

$$\mu = \frac{1}{N} \sum_{i=1}^N X_i = \bar{X},$$

母體變異數

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2,$$

母體標準差

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2}.$$

網站：[淺談自由度 \(樣本標準差公式中的分母為什麼要採用 \$n-1\$ \)](#)

淺談自由度 (樣本標準差公式中的分母為什麼要採用 $n-1$)

江振東 教授 / 政治大學統計系

我們都知道，在母體平均數已知的情形下我們可以利用來估計母體變異數。但是在母體平均數未知的情形下，我們則改用來作估計。由於未知，因此直觀上我們可以來作取代，但是分母為什麼也要調整為 $n-1$ 呢？由於，因此。如此一來，我們就可以發現如果是母體變異數的一個好的估計量的話，那麼顯然就會有低估的可能。如果改用來作為估計量，低估的現象應該可以獲得改善。但是為什麼是 $n-1$ ，而不是 $n-2$ ，甚至 $n-3$ 呢？這就牽涉到自由度的問題。

[詳全文\(225KB\)](#)

出處：教育部高中數學學科中心電子報線上系統

[轉寄文章](#)

[列印本文](#)

[關閉視窗](#)

網站：[中位數的時代來了](#)


所得收入者的可支配所得						
	全體			大學學歷者		
	平均數	中位數	中位數 / 平均數	平均數	中位數	中位數 / 平均數
1993年	41.5	35.6	0.86	66.3	61.2	0.92
2003年	50.3	41.2	0.82	69.9	60.2	0.86
2017年	54.5	44.0	0.80	60.5	48.6	0.80

註：以中位數除以平均數，可測量不均度。大學學歷者不包括研究所學歷者
資料來源：主計總處「家庭收支調查」 單位：萬元 製表：于國欽

所得收入者的可支配所得

一群數據用以描述其中心位置的統計，有平均數（mean）、中位數（median）及眾數（mode），中位數是把一群數據依高低排序，位於中央者，即是中位數。

網站：[scipy.stats](#)


SciPy.org

[SciPy.org](#)
[Docs](#)
[SciPy v1.5.3 Reference Guide](#)

Statistical functions (scipy.stats)

This module contains a large number of probability distributions as well as a growing library of statistical functions.

Each univariate distribution is an instance of a subclass of `rv_continuous` (`rv_discrete` for discrete distributions):

`rv_continuous([momtype, a, b, xtol, ...])` A generic continuous random variable class meant for subclassing.
`rv_discrete([a, b, name, badvalue, ...])` A generic discrete random variable class meant for subclassing.
`rv_histogram(histogram, *args, **kwargs)` Generates a distribution given by a histogram.

Continuous distributions

<code>alpha(*args, **kwargs)</code>	An alpha continuous random variable.
<code>anglit(*args, **kwargs)</code>	An anglit continuous random variable.
<code>arcsine(*args, **kwargs)</code>	An arcsine continuous random variable.
<code>argus(*args, **kwargs)</code>	Argus distribution
<code>beta(*args, **kwargs)</code>	A beta continuous random variable.
<code>betaprime(*args, **kwargs)</code>	A beta prime continuous random variable.
<code>bradford(*args, **kwargs)</code>	A Bradford continuous random variable.
<code>burr(*args, **kwargs)</code>	A Burr (Type III) continuous random variable.
<code>burr12(*args, **kwargs)</code>	A Burr (Type XII) continuous random variable.
<code>cauchy(*args, **kwargs)</code>	A Cauchy continuous random variable.
<code>chi(*args, **kwargs)</code>	A chi continuous random variable.
<code>chi2(*args, **kwargs)</code>	A chi-squared continuous random variable.
<code>cosine(*args, **kwargs)</code>	A cosine continuous random variable.
<code>crystalball(*args, **kwargs)</code>	Crystalball distribution
<code>dgamma(*args, **kwargs)</code>	A double gamma continuous random variable.
<code>dweibull(*args, **kwargs)</code>	A double Weibull continuous random variable.
<code>erlang(*args, **kwargs)</code>	An Erlang continuous random variable.
<code>expon(*args, **kwargs)</code>	An exponential continuous random variable.
<code>exponnorm(*args, **kwargs)</code>	An exponentially modified Normal continuous random variable.
<code>exponweib(*args, **kwargs)</code>	An exponentiated Weibull continuous random variable.
<code>exponpow(*args, **kwargs)</code>	An exponential power continuous random variable.
<code>f(*args, **kwargs)</code>	An F continuous random variable.
<code>fatiguelife(*args, **kwargs)</code>	A fatigue-life (Birnbaum-Saunders) continuous random variable.
<code>fisk(*args, **kwargs)</code>	A Fisk continuous random variable.
<code>foldcauchy(*args, **kwargs)</code>	A folded Cauchy continuous random variable.
<code>foldnorm(*args, **kwargs)</code>	A folded normal continuous random variable.
<code>frechet_r(*args, **kwargs)</code>	A Frechet right (or Weibull minimum) continuous random variable.
<code>frechet_l(*args, **kwargs)</code>	A Frechet left (or Weibull maximum) continuous random variable.
<code>genlogistic(*args, **kwargs)</code>	A generalized logistic continuous random variable.
<code>gennorm(*args, **kwargs)</code>	A generalized normal continuous random variable.
<code>genpareto(*args, **kwargs)</code>	A generalized Pareto continuous random variable.
<code>genexpon(*args, **kwargs)</code>	A generalized exponential continuous random variable.

網站：[statistics — Mathematical statistics functions](#)

Python » English » 3.9.0 » Documentation » The Python Standard Library » Numeric and Mathematical Modules »

Table of Contents

statistics — Mathematical statistics functions

- Averages and measures of central location
- Measures of spread
- Function details
- Exceptions
- NormalDist objects
 - NormalDist
 - Examples and Recipes

Previous topic

random — Generate pseudo-random numbers

Next topic

Functional Programming Modules

This Page

[Report a Bug](#)
[Show Source](#)

statistics — Mathematical statistics functions

New in version 3.4.

Source code: [Lib/statistics.py](#)

This module provides functions for calculating mathematical statistics of numeric (real-valued) data.

The module is not intended to be a competitor to third-party libraries such as NumPy, SciPy, or proprietary full-featured statistics packages aimed at professional statisticians such as Minitab, SAS and Matlab. It is aimed at the level of graphing and scientific calculators.

Unless explicitly noted, these functions support int, float, Decimal and Fraction. Behaviour with other types (whether in the numeric tower or not) is currently unsupported. Collections with a mix of types are also undefined and implementation-dependent. If your input data consists of mixed types, you may be able to use map() to ensure a consistent result, for example: map(float, input_data).

Averages and measures of central location

These functions calculate an average or typical value from a population or sample.

mean()	Arithmetic mean ("average") of data.
fmean()	Fast, floating point arithmetic mean.
geometric_mean()	Geometric mean of data.
harmonic_mean()	Harmonic mean of data.
median()	Median (middle value) of data.
median_low()	Low median of data.
median_high()	High median of data.
median_grouped()	Median, or 50th percentile, of grouped data.
mode()	Single mode (most common value) of discrete or nominal data.
multimode()	List of modes (most common values) of discrete or nominal data.
quantiles()	Divide data into intervals with equal probability.

Measures of spread

These functions calculate a measure of how much the population or sample tends to deviate from the

網站：[pandas](#)

<div>10 minutes to pandas</div> <div>Intro to data structures</div> <div>Essential basic functionality</div> <div>IO tools (text, CSV, HDF5, ...)</div> <div>Indexing and selecting data</div> <div>Multindex / advanced indexing</div> <div>Merge, join, concatenate and compare</div> <div>Reshaping and pivot tables</div> <div>Working with text data</div> <div>Working with missing data</div> <div>Categorical data</div> <div>Nullable integer data type</div> <div>Nullable Boolean data type</div> <div>Visualization</div> <div>Computational tools</div> <div>Group by: split-apply-combine</div> <div>Time series / date functionality</div> <div>Time deltas</div> <div>Styling</div> <div>Options and settings</div> <div>Enhancing performance</div> <div>Scaling to large datasets</div> <div>Sparse data structures</div> <div>Frequently Asked Questions (FAQ)</div> <div>Cookbook</div>	<div>The User Guide covers all of pandas by topic area. Each of the subsections introduces a topic (such as "working with missing data"), and discusses how pandas approaches the problem, with many examples throughout.</div> <div>Users brand-new to pandas should start with 10 minutes to pandas.</div> <div>For a high level summary of the pandas fundamentals, see Intro to data structures and Essential basic functionality.</div> <div>Further information on any specific method can be obtained in the API reference.</div> <div><ul style="list-style-type: none">10 minutes to pandas<ul style="list-style-type: none">Object creationViewing dataSelectionMissing dataOperationsMergeGroupingReshapingTime seriesCategoricalsPlottingGetting data in/outGotchasIntro to data structures<ul style="list-style-type: none">SeriesDataFrameEssential basic functionality<ul style="list-style-type: none">Head and tailAttributes and underlying dataAccelerated operationsFlexible binary operationsDescriptive statisticsFunction applicationReindexing and altering labelsIteration.dt accessorVectorized string methods</div>
---	--

下一步：閱讀範例與完成作業

