

機率論與統計學

沈威宇

2025 年 2 月 4 日

目錄

第一章 機率論 (Probability Theory) 與統計學 (Statistics)	1
第一節 機率論	1
一、 試驗、事件與機率	1
二、 分割/劃分 (Partitions)	2
三、 Probability space (機率空間)	2
四、 Random Variable (隨機變數)	2
五、 Probability Mass Function (機率質量函數) or Probability Function (機率函數)	2
六、 期望值 (Expected Value)	2
七、 獨立事件 (Independent Events)	2
八、 互斥事件 (Mutually Exclusive Events)	2
九、 貝葉斯/貝氏定理 (Bayes' Theorem)	3
第二節 一維數據分析	3
一、 眾數 (Mode, Mo)	3
二、 中位數 (Median, Me)	3
三、 算術平均數 (Arithmetic mean)	3
四、 加權平均數	3
五、 幾何平均數	3
六、 百分位數 (Percentile, Percentile score)	3
七、 四分位數 (Quantile)	5
八、 百分位排名/等級 (Percentile rank)	6
九、 全距	6
十、 四分位距	6
十一、 母體變異數 (Population variance) 和母體標準差 (Population standard deviation)	6
十二、 樣本變異數 (Sample variance) 和樣本標準差 (Sample standard deviation)	6

十三、	線性變換	7
十四、	標準化	7
第三節	二維數據分析	7
一、	散布圖	7
二、	皮爾森積動差相關係數 (Pearson product-moment correlation coefficient, PPMCC, PCCs) /相關係數	8
三、	判定係數 (Coefficient of determination)	8
四、	(線性) 迴歸直線/最適直線	8
第四節	多維資料分析	9
一、	迴歸直線	9
第五節	參考文獻	11

第一章 機率論 (Probability Theory) 與統計學 (Statistics)

第一節 機率論

一、試驗、事件與機率

- 試驗 (Experiment)：指一個可以重複進行並且每次結果可能不同的過程。具有可重複性，即試驗可以在相同條件下重複進行，與隨機性，即每次試驗的結果可能不同，具有隨機性和不確定性。
- 樣本空間 (Sample Space)：一試驗所有可能結果的集合。例如，擲一枚硬幣的樣本空間是 {正面, 反面}。
- 事件 (Event)：樣本空間的子集。例如，擲一枚骰子得到一個偶數的事件是 {2, 4, 6}。
- 機率 (Probability)：事件發生的可能性，為 0 到 1 之間的數字。機率越接近 1，事件發生的可能性就越大。
- 空事件：機率為零的事件。
- 全事件：機率為一的事件。
- 和事件：事件 A 和事件 B 的和事件為 $A \cup B$ 。
- 積事件：事件 A 和事件 B 的積事件為 $A \cap B$ 。
- 餘事件：樣本空間 S 中，事件 A 的餘事件 $A' = S \setminus A$ 。
- 古典機率 (Classical Probability)：如果一個事件的所有可能結果數目是有限，且樣本空間中每個結果發生的機會相等，則事件發生的機率可以通過以下公式計算：

$$P(A) = \frac{\text{發生事件 } A \text{ 的結果數}}{\text{所有可能結果數}}$$

- 條件機率 (Conditional Probability)：在已知某事件發生的情況下，另一事件發生的機率。通常表示為 $P(A|B)$ ，即在事件 B 發生的情況下事件 A 發生的機率。

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- 客觀機率/頻率機率：根據過往的經驗或統計數據而得到的客觀數值，通常以過往事件發生的頻率或多次重複試驗來得到該事件發生的機率。
- 主觀機率：沒有統計數據支持的機率數值。

二、 分割/劃分 (Partitions)

若 $\{A_i\}_{i \in I}$ 是樣本空間 Ω 的一組分割，那麼滿足以下條件：

$$\begin{aligned} \forall i \in I : A_i &\subseteq \Omega, \\ \forall i, j \in I \wedge i \neq j : A_i \cap A_j &= \emptyset, \\ \bigcup_{i \in I} A_i &= \Omega. \end{aligned}$$

三、 Probability space (機率空間)

A probability space (Ω, Σ, μ) is a measure space where $\mu : \Sigma \rightarrow [0, 1]$ and $\mu(\Omega) = 1$, in which μ is called the probability measure.

四、 Random Variable (隨機變數)

The random variables are using real numbers to indicate events in the sameple space, either discrete or continuous, called discrete random variables and continuous random variables.

五、 Probability Mass Function (機率質量函數) or Probability Function (機率函數)

The probability mass function is a measurable function of which the domain is random variables in the sample space and the codomain is a probability space.

六、 期望值 (Expected Value)

隨機變數的長期平均值。對於可能值之集合為 Y 的離散隨機變數 X ，其期望值計算公式為：

$$E(X) = \sum_{x_i \in Y} (x_i \cdot P(x_i))$$

七、 獨立事件 (Independent Events)

指在一次試驗中，兩個或多個事件彼此之間沒有影響。即：

$$\left(\forall J \neq \emptyset \subseteq \{a \mid a \in \mathbb{N} \wedge 1 \leq a \leq n\} : P\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} P(A_j) \right) \iff ((A_1, A_2, \dots, A_n) \text{ 為獨立事件})$$

八、 互斥事件 (Mutually Exclusive Events)

指在一次試驗中，兩個或多個事件不可能同時發生，即一些互斥事件中的任兩個的和事件的機率為零。

九、 貝葉斯/貝氏定理 (Bayes' Theorem)

若 $\{A_i\}_{i \in I}$ 是樣本空間 Ω 的一個分割，則：

$$\forall 1 \leq j \leq |I| : P(A_j|B) = \frac{P(A_j) \times P(B|A_j)}{\sum_{k=1}^{|I|} P(A_k) \times P(B|A_k)}$$

第二節 一維數據分析

今有一由小到大排列的實數序列 $\mathbf{X} = \{x_1, x_2, \dots, x_n\}$ 。

一、 眾數 (Mode, Mo)

出現次數最多者。

二、 中位數 (Median, Me)

$$\begin{cases} x_{\frac{n+1}{2}}, & n \text{ is odd.} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & n \text{ is even.} \end{cases}$$

三、 算術平均數 (Arithmetic mean)

$$\mu = \frac{\sum_{i=1}^n x_i}{n}$$

四、 加權平均數

令 (x_1, x_2, \dots, x_n) 對應的權數為 (w_1, w_2, \dots, w_n) 。加權平均數為：

$$\frac{\sum_{i=1}^n x_i w_i}{\sum_{i=1}^n w_i}$$

五、 幾何平均數

$$\sqrt[n]{\prod_{i=1}^n x_i}$$

六、 百分位數 (Percentile, Percentile score)

令：第 k 百分位數為 P_k ， $m = n \frac{k}{100}$ ， $i = \lfloor m \rfloor$ ， $j = i + 1$ ， $g = m - i$ ， $h = \frac{k}{100}(n - \alpha - \beta + 1) + \alpha$ ， $r = (n - 1) \frac{k}{100}$ ， $s = \lfloor r \rfloor + 1$ ， $t = s + 1$ 。

令：

$$\text{RoundHalfToEven}(x) = \begin{cases} \lfloor x \rfloor, & \text{if } x - \lfloor x \rfloor < 0.5 \\ \lceil x \rceil, & \text{if } x - \lfloor x \rfloor > 0.5 \\ 2 \left\lfloor \frac{x}{2} \right\rfloor, & \text{if } x - \lfloor x \rfloor = 0.5 \text{ and } \lfloor x \rfloor \text{ is even} \\ 2 \left\lfloor \frac{x}{2} \right\rfloor + 1, & \text{if } x - \lfloor x \rfloor = 0.5 \text{ and } \lfloor x \rfloor \text{ is odd} \end{cases}$$

$$x_h = x_{\lfloor h \rfloor} + (h - \lfloor h \rfloor)(x_{\lceil h \rceil} - x_{\lfloor h \rfloor}).$$

各種百分位數定義主要分為兩類：

- 包含性定義（Inclusive definition）：較常用。至少有 $k\%$ 的項 $\leq P_k$ ，且至少有 $(100 - k)\%$ 的項 $\geq P_k$ 。
- 排除性定義（Exclusive definition）：較少用。至少有 $k\%$ 的項 $< P_k$ ，且至少有 $(100 - k)\%$ 的項 $> P_k$ 。

各種百分位數定義：

- inverted_cdf (method 1 of H & F):

$$P_k = \begin{cases} x_j, & g > 0 \\ x_i, & g = 0 \end{cases}$$

- averaged_inverted_cdf (method 2 of H & F): 離散定義中最常用。

$$P_k = \begin{cases} x_j, & g > 0 \\ \frac{x_i + x_j}{2}, & g = 0 \end{cases}$$

- closest_observation (method 3 of H & F):

$$P_k = x_{\text{RoundHalfToEven}(m)}$$

- interpolated_inverted_cdf (method 4 of H & F):

$$\alpha = 0$$

$$\beta = 1$$

$$P_k = x_h$$

- hazen (method 5 of H & F):

$$\alpha = \frac{1}{2}$$

$$\beta = \frac{1}{2}$$

$$P_k = x_h$$

- weibull (method 6 of H & F): Excel PERCENTILE.EXC 使用其乘以% 為值。

$$\alpha = 0$$

$$\beta = 0$$

$$P_k = x_h$$

- linear (method 7 of H & F): Excel PERCENTILE.INC 使用其乘以% 為值。連續定義中最常用。

$$\alpha = 1$$

$$\beta = 1$$

$$P_k = x_h$$

- median_unbiased (method 8 of H & F):

$$\alpha = \frac{1}{3}$$

$$\beta = \frac{1}{3}$$

$$P_k = x_h$$

- normal_unbiased (method 9 of H & F):

$$\alpha = \frac{3}{8}$$

$$\beta = \frac{3}{8}$$

$$P_k = x_h$$

- lower (NumPy old method):

$$P_k = x_s$$

- higher (NumPy old method):

$$P_k = x_t$$

- nearest (NumPy old method):

$$P_k = x_{\text{RoundHalfToEven}(r)+1}$$

- midpoint (NumPy old method):

$$P_k = \frac{x_s + x_t}{2}$$

七、 四分位數 (Quantile)

與百分位數同有該等各種定義，僅將其中之 100 均改為 4、第 k 四分位數稱 Q_k 、PERCENTILE.INC 改為 QUANTILE.INC、PERCENTILE.EXC 改為 QUANTILE.EXC，並另有下列其他定義方法。

- 定義一：先取中位數為 Q_2 ，將序列以中位數為界分為兩半，若 n 為奇數則中位數不包含在兩半，分別取兩半之中位數為 Q_1 、 Q_3 。
- 定義二：先取中位數為 Q_2 ，將序列以中位數為界分為兩半，若 n 為奇數則中位數包含在兩半，分別取兩半之中位數為 Q_1 、 Q_3 。
- 定義三：先取中位數為 Q_2 ，若 n 為偶數則將序列以中位數為界分為兩半，分別取兩半之中位數為 Q_1 、 Q_3 ；若 n 除以 4 的商為 q 且餘數為 1，則 $Q_1 = 0.25x_q + 0.75x_{q+1}$ ； $Q_3 = 0.75x_{3q+1} + 0.25x_{3q+2}$ ；若 n 除以 4 的商為 q 且餘數為 3，則 $Q_1 = 0.75x_{q+1} + 0.25x_{q+2}$ ； $Q_3 = 0.25x_{3q+2} + 0.75x_{3q+3}$ 。

八、 百分位排名/等級 (Percentile rank)

令百分位等級 PR，累積次數 CF 為小於等於感興趣值的項數，次數 F 為於等於感興趣值的項數，CF' 為小於感興趣值的項數。

- 定義一：

$$PR = 100 \frac{CF - 0.5F}{n} = 100 \frac{CF' + 0.5F}{n}$$

- 定義二 (Excel PERCENTRANK.INC 定義)，最常用：

$$PR = \frac{CF'}{n - 1} 100\%$$

- 定義三 (Excel PERCENTRANK.EXC 定義)：

$$PR = \frac{CF' + 1}{n + 1} 100\%$$

九、 全距

$$R = \max(\mathbf{X}) - \min(\mathbf{X})$$

十、 四分位距

$$Q_3 - Q_1$$

十一、 母體變異數 (Population variance) 和母體標準差 (Population standard deviation)

稱 $x_i - \mu$ 為離均差， $i = 1, 2, \dots, n$ 。母體變異數 σ^2 或 $\text{Var}(X)$ ，母體標準差 σ ：

$$\begin{aligned}\sigma^2 &= \frac{\sum_{i=1}^n (x_i - \mu)^2}{n} \\ &= \frac{\sum_{i=1}^n x_i^2}{n} - \mu^2 \\ \sigma &= \sqrt{\sigma^2}\end{aligned}$$

十二、 樣本變異數 (Sample variance) 和樣本標準差 (Sample standard deviation)

稱 $x_i - \mu$ 為離均差， $i = 1, 2, \dots, n$ 。樣本變異數 s^2 ，樣本標準差 s ：

$$\begin{aligned}s^2 &= \frac{\sum_{i=1}^n (x_i - \mu)^2}{n - 1} \\ &= \frac{\sum_{i=1}^n x_i^2 - n\mu^2}{n - 1} \\ s &= \sqrt{s^2}\end{aligned}$$

十三、 線性變換

\mathbf{X} 的線性變換 $\mathbf{Y} = \{y_i \mid y_i = ax_i + b, i = 1, 2, \dots, n\}$ ，記作 $\mathbf{Y} = a\mathbf{X} + b$ 。

性質：

$$\mu_{\mathbf{Y}} = a\mu_{\mathbf{X}} + b$$

$$\sigma_{\mathbf{Y}} = |a|\sigma_{\mathbf{X}}$$

$$s_{\mathbf{Y}} = |a|s_{\mathbf{X}}$$

十四、 標準化

標準分數/Z 分數 \mathbf{Z} ：即標準化後的數據

$$\mathbf{Z} = \left\{ z_i \mid z_i = \frac{x_i - \mu}{\sigma}, i = 1, 2, \dots, n \right\}.$$

性質：

$$\mu_{\mathbf{Z}} = 0, \quad \sigma_{\mathbf{Z}} = 1$$

$$\sum_{i=1}^n z_i^2 = n$$

第三節 二維數據分析

今有由小到大排列的實數序列 $\mathbf{X} = \{x_1, x_2, \dots, x_n\}$ 與 $\mathbf{Y} = \{y_1, y_2, \dots, y_n\}$ ，標準差分別為 $\sigma_{\mathbf{X}}, \sigma_{\mathbf{Y}}$ ，算術平均數分別為 $\mu_{\mathbf{X}}, \mu_{\mathbf{Y}}$ ，標準化後的數據 $\mathbf{X}' = \{x'_i \mid x'_i = \frac{x_i - \mu_{\mathbf{X}}}{\sigma_{\mathbf{X}}}, i = 1, 2, \dots, n\}$ 與 $\mathbf{Y}' = \{y'_i \mid y'_i = \frac{y_i - \mu_{\mathbf{Y}}}{\sigma_{\mathbf{Y}}}, i = 1, 2, \dots, n\}$ 。

一、 散布圖

將數據點每個 $(x_i, y_i), i = 1, 2, \dots, n$ 描繪在 xy 平面直角座標平面。

二、 皮爾森積動差相關係數 (Pearson product-moment correlation coefficient, PPMCC, PCCs) / 相關係數

X 與 Y 的相關係數記作 r_{XY} 。定義：

$$\begin{aligned}r_{XY} &= \frac{\sum_{i=1}^n x'_i y'_i}{n} \quad (\text{標準化積和除以項數}) \\S_{XX} &= \sum_{i=1}^n (x_i - \mu_X)^2 = \sum_{i=1}^n x_i^2 - n\mu_X^2 = n\sigma_X^2 \\S_{YY} &= \sum_{i=1}^n (y_i - \mu_Y)^2 = \sum_{i=1}^n y_i^2 - n\mu_Y^2 = n\sigma_Y^2 \\S_{XY} &= \sum_{i=1}^n (x_i - \mu_X)(y_i - \mu_Y) = \sum_{i=1}^n x_i y_i - n\mu_X \mu_Y \\r_{XY} &= \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}} \quad (\text{離均差積和除以根號離均差平方和積})\end{aligned}$$

性質：

$$-1 \leq r \leq 1, \quad 0 \leq r^2 \leq 1$$

$$r_{XY} = r_{YX}$$

相關程度：

- $r = 1$ 稱完全正相關； $r = -1$ 稱完全負相關。
- $r > 0$ 稱正相關； $r < 0$ 稱負相關； $r = 0$ 稱無相關。

線性變換：

令 $X' = aX + b$ 、 $Y' = aY + b$ 。

$$r_{X'Y'} = \frac{ac}{|ac|} r_{XY}$$

三、 判定係數 (Coefficient of determination)

指皮爾森積動差相關係數的平方。

四、 (線性) 迴歸直線/最適直線

令平方和：

$$D = \sum_{i=1}^n (y_i - (mx_i + k))^2$$

解出使 D 最小 (即 D 為最小平方和) 的 m, k 即得 L (即最小平方方法)。

X' 與 Y' 的最適直線 L ： $mx + k$ 為：

$$y' = r_{X'Y'} x'$$

X 與 Y 的最適直線為：

$$y - \mu_Y = m(x - \mu_X)$$

其中：

$$m = r_{XY} \cdot \frac{\sigma_Y}{\sigma_X} = \frac{S_{XY}}{S_{XX}}$$

第四節 多維資料分析

一、迴歸直線

設有 n 個樣本，每個樣本有 m 個特徵。

令矩陣 \mathbf{X} 是 $n \times (m+1)$ 的矩陣，第一 column 是全為 1 的 column（對應截距項），其餘 column 是特徵 x_1, x_2, \dots, x_m 。

令 \mathbf{y} 是 $n \times 1$ 的 column 向量，表示目標變數。

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}$$
$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

迴歸係數 \mathbf{a} 可以用以下公式計算：

$$\mathbf{a} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

得到迴歸方程式：

$$y = (1, x_1, x_2, \dots, x_m) \mathbf{a}$$

Proof.

最小平方法的目標是找到一組係數 \mathbf{a} ，使得實際值 \mathbf{y} 與預測值 \mathbf{Xa} 之間的平方差和最小，即最小化以下目標函數：

$$J(\mathbf{a}) = \sum_{i=1}^n (y_i - \mathbf{X}_i \mathbf{a})^2$$

其中， \mathbf{X}_i 是 \mathbf{X} 的第 i row。寫成矩陣形式：

$$J(\mathbf{a}) = (\mathbf{y} - \mathbf{Xa})^T (\mathbf{y} - \mathbf{Xa})$$

展開：

$$J(\mathbf{a}) = \mathbf{y}^T \mathbf{y} - \mathbf{y}^T \mathbf{Xa} - \mathbf{a}^T \mathbf{X}^T \mathbf{y} + \mathbf{a}^T \mathbf{X}^T \mathbf{Xa}$$

因純量的轉置為其自身，所以：

$$\mathbf{y}^T \mathbf{Xa} = \mathbf{a}^T \mathbf{X}^T \mathbf{y}$$

即：

$$J(\mathbf{a}) = \mathbf{y}^T \mathbf{y} - 2\mathbf{y}^T \mathbf{Xa} + \mathbf{a}^T \mathbf{X}^T \mathbf{Xa}$$

要最小化 $J(\mathbf{a})$ ，我們對 \mathbf{a} 求導數並令其為零：

$$\frac{\partial J(\mathbf{a})}{\partial \mathbf{a}} = -2\mathbf{y}^T \mathbf{X} + \mathbf{X}^T \mathbf{Xa} = 0$$

整理後得到：

$$\mathbf{X}^T \mathbf{X} \mathbf{a} = \mathbf{X}^T \mathbf{y}$$

假設 $\mathbf{X}^T \mathbf{X}$ 是可逆的，我們可以兩邊同時乘以 $(\mathbf{X}^T \mathbf{X})^{-1}$ ：

$$\mathbf{a} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

□

第五節 參考文獻

- R. J. Hyndman and Y. Fan, “Sample quantiles in statistical packages,” The American Statistician, 50(4), pp. 361-365, 1996.
- Numpy. numpy.percentile. <https://numpy.org/doc/stable/reference/generated/numpy.percentile.html>.