4. 指數分佈具有無記憶的特性（Memoryless Property），在工程界和金融界有廣泛的應用。假設 $x_1, x_2, \ldots\ldots x_n$ 是指數分佈 $exp(\lambda)$ 的一組隨機抽樣樣本，其概率密度分佈函數是 $f(x)=\lambda e^{-\lambda x}$（$x>0$），求該指數分佈的一階和二階動差估計量。

5. 讀取數據集 Bwages.csv，求出在信心水準為 0.95 下，變數 wage 的信賴區間。

6. 使用樣本均值和變異數繪製變數 wage 擬合的常態分佈的概率密度圖，與 wage 的頻率直方圖作比較，看看我們的假設是否合理。

7. 下表分別給出了兩位文學家馬克 吐溫 (Mark Twain) 的 8 篇小品文以及斯諾特格拉斯（Snodgrass）的 8 篇小品文中由 3 個字母組成的單字的比例：

Mark Twain	0.225	0.262	0.217	0.240	0.230	0.229	0.235	0.217
Snodgrass	0.209	0.205	0.196	0.210	0.202	0.207	0.224	0.223

假設這兩組數據分別來自常態母體，且母體變異數相等，但參數皆未知。兩樣本相互獨立。問兩位作家所寫的小品文中包含由 3 個字母組成的單字的比例是否有顯著性的差異，取 $\alpha=0.05$。

8. 隨機選取兩個地區各 10 個人，測量他們的身高（cm），得到以下的數據：

編號	1	2	3	4	5	6	7	8	9	10
地區1	168	180	181	172	165	160	166	165	177	174
地區2	169	170	176	173	166	167	166	173	171	170

設各對數據的差（即同一編號下的一對數據的差）是來自互相獨立的常態母體之樣本，均值和變異數均未知，取顯著性水平 $\alpha=0.05$，運用 Python 判斷這兩個地區的人的身高是否有顯著性差異。

9. 讀取數據集 Bwages.csv，對變數 wage 進行 t 檢定（虛無假設 wage 的均值為 11，信心水準為 0.95），根據輸出結果得出你的結論。

10. 讀取數據集 history.csv，假設所有風格的對沖基金的收益率獨立，檢定新興市場（Emerging Markets）風格對沖基金和全球宏觀（Global Macro）風格對沖基金收益率是否有顯著差異，根據輸出結果得出你的結論。

11. 對上題的兩種風格的對沖基金收益率進行配對樣本 t 檢定，根據輸出結果得出你的結論。

— 第 17 章 —

變異數分析

前面的章節主要是針對單個變數進行統計分析，如描述性統計、參數估計等。然而，僅對單個變數進行分析所能獲得的訊息是有限的。以某公司股票的股價為例，對其進行單變數分析，我們最多只能獲得其概率分佈，然後據此來對股價進行預測。透過這樣的方式，盡管也能進行預測，但預測的可信度是非常有限的。然而，如果我們知道明天該公司會宣告發放股利，那麼由此可判斷明天股價上漲的概率極高。所以，在分析某個變數的時候，往往還會需要其他變數的訊息。這樣，統計分析工作就涉及到了多個變數。不同於單變數分析，對於多個變數的統計分析，我們更多地關注多個變數之間存在的聯繫，以幫助提高預測的準確性。下面先來介紹一種對變數間關係的定性分析方法——變異數分析（Analysis of Variance，ANOVA）。

17.1 | 變異數分析之思想

在正式介紹變異數分析之前，我們先來看一個問題：「不同行業股票的收益率是否相同？」。如果答案是「不同行業的股票有著明顯不同的收益率」，並在不同行業的股票收益率相互獨立的前提下，我們可以進一步提出疑問：食品行業的收益率會比金融行業的收益率更高嗎？再考慮總體經濟學中的失業問題中，失業率會因為地區的不同而不同。如果地區確實是影響失業率的一個重要的因素，那麼台北的失業率會比高雄的失業率更高還是更低？

總結起來，上述兩個問題的前一半關注的都是，一個因子（Factor）變數（如行業和地區）是否會影響某一個變數（如收益率和失業率）的數值。因子變數的取值可以是不同的狀態，我們稱這些狀態為水平。例如，行業因子變數，其取值不

是 1.414、3.1415926 諸如此類的數值，而是「食品行業」「金融行業」這樣的水平。欲研究的被影響變數被稱為反應變數（Response Variable）。這兩個問題的後一半歸結起來關注的是：兩個不同的水平下（食品行業 VS 金融行業，台北 VS 高雄）反應變數（如股票收益率，失業率）是如何取值的？哪種情況下反應變數的取值更高？

若要探究一個因子變數對反應變數的影響，變異數分析是一個較為適合的工具。變異數分析從反應變數（如上述的股票收益率和失業率）的變異數入手，研究諸多因子（如行業、地區等因素）中哪些因子對觀測變數有顯著影響。變異數分析的重點不在於預測（它無法預測出明天金融行業股票的走勢如何），而在於分析和比較各組之間的差異。例如，分析食品行業和金融行業股票收益率的差異。如果我們發現，這兩個行業的股票收益率是有顯著差異的，則可以得到下述結論：行業是影響股票收益率的一個重要的因素。

準確地說，變異數分析的研究對像是各個組別反應變數均值之間可能存在的差異，其中組別的劃分是以因子變數為依據的。由於需要藉助變異數來觀察均值是否相同，所以被叫做變異數分析。透過變異數分析，可以檢定分組所依據的因子變數對反應變數是否具有重要的影響。如果反應變數在不同組別中的均值是相同的，則可以認為分組所依據的因子變數對反應變數沒有影響（如果所有地區的平均失業率都是一樣的，地區對失業率就沒有重要的影響）。反之，可以推斷分組所依據的因子變數是影響該反應變數的重要因素。請注意，不一定要求所有水平下反應變數的均值都不同，才能說明該因子變數是有重要影響的。只要存在至少兩個組別的均值顯著不同，就可以認為該因子變數對反應變數是有影響的。例如，哪怕只有台北和高雄的失業率不同，其它地區的失業率都是一樣的，也可以說明地區是對失業率有影響的。

根據所研究的因素的數量，可以將變異數分為單因素變異數分析、多因素變異數分析和析因變異數分析。單因素變異數分析即是只研究一個因子的變異數分析，如前面談到的失業率、股票收益率等例子都屬於單因素變異數分析。多因素變異數分析則是研究多個因子的變異數分析，最常見的多因素變異數分析為二因素變異數分析，即研究兩個因子的變異數分析，比如探討施肥量和灌溉量對於糧食產量的影響

即是一個二因素變異數分析。多因素變異數分析研究的是每個因子是否對因變數有著重要的影響，而不是這些因子整體對因變數是否有著重要影響。析因變異數分析則是在多因素變異數分析的基礎上加入了因子之間的乘項，其原因是一個因子對反應變數的影響大小可能受到另一個因子的水平的影響。舉個簡單的例子。假設有兩個因子——是否酗酒與年齡段。是否酗酒有兩個水平，即「是」和「否」；年齡段也有兩個水平，即「青年」和「老年」。我們都知道酗酒對身體有負面影響，同時老年人酗酒對身體的傷害比年輕人酗酒對身體的傷害更大。也就是說，是否酗酒對身體的影響在不同的年齡段水平是不一樣的。為了體現出這種影響，可以加入是否酗酒與年齡段的乘項，進行析因變異數分析。

在現實世界中，影響一個反應變數的因素往往有很多種，多因素變異數分析即體現了這一點。但是，盡管有著很多影響因素，有時我們只想研究其中的一兩種，而不是全部。值得注意的是，在變異數分析中，如果發現一個因素對反應變數有著重要的影響，這並不能保證該因素真的對反應變數有影響。之所以得到這樣的結果的原因可能是，有另外一個與該因素相關的因素對反應變數產生了影響，我們把這種因素叫做干擾因素（Confounding Factor）。為了避免干擾因素的影響，需要加入其他變數以控制干擾因素。如果加入的是因子變數，我們採取的就是隨機區組設計（Randomized Block Design）。如果加入的是連續變數，那麼該變數就是共變數，我們所進行的就是共變異數分析（Analysis of Covariance，ANCOVA）。

17.2 | 變異數分析之原理

變異數分析的目的在於分析因子對反應變數有無顯著影響；亦即，在因子的不同水平下，反應變數的均值是否有顯著差異。一般來說，影響反應變數的因素有兩大類：

1. 不可控的隨機因素

即使兩塊一模一樣的土地、施加完全一樣的肥料、灌溉一樣數量的水、給予完全一樣的光照，得到的糧食產量也不見得會完全一樣。有太多無法控制的隨機因素會影響產量，例如這塊土地種植的大豆的基因或許比另一塊土地上大豆的基因好。即使

是同一品種、同一棵植物上獲取的大豆也不見得相同。若要研究行業對股票的收益率的影響，除了行業間收益率可能存在差異以外，還存在其他不可控的隨機因素會影響股票的收益率。

2. 研究中施加對結果形成影響的可控因素（因子）

若要研究施肥量對於糧食產量的影響，施肥量則是對結果會產生影響的可控因素。

這些因素都會使我們收集到的反應變數數據產生波動。變異數分析透過分析不同來源的波動（不可控隨機 VS 可控因素）對總波動（反應變數的總體變化）的貢獻大小，進而確定可控因素（因子）對反應變數影響力的大小。如果反應變數的波動主要由可控因素引起，可控因素對於總波動的貢獻較大，則說明可控因素對於反應變數有顯著的影響。例如，如果不同施肥量條件下的糧食的產量大小相似、不同組別之間產量無變化，僅有的變化是由種子質量等不可控的隨機因素引起，我們無法得出「施肥量是影響糧食產量的一個重要因素」的結論。如果產量的變化很大程度上是由「施肥量」這個因素引起，即使這個產量整體變化差異很小，也可以說「施肥量」是產量的一個影響因素，只不過這個因素的影響作用有限。

17.2.1 離差平方和

現在以單因素變異數分析為例，說明變異數分析的假設檢定過程。假設現在因子變數共有 M 個水平，每個水平下試驗或觀測對象有 N_j 個（ $j=1,2,\ldots,M$ ）。令 Y_{ij} 表示第 j 個水平組別下第 i 個反應變數，其中 $i=1,2,\ldots,N_j$。令 μ_j 代表第 j 個水平組別下反應變數的均值，μ_0 代表所有反應變數的均值。若因子水平對反應變數無影響，則不同因子水平下反應變數的均值是相同的，這就是變異數分析之虛無假設：

$$H_0:\ \mu_1=\mu_2=\ldots=\mu_M=\mu_0$$

現在我們觀測到不同因子水平下之樣本數據 y_{ij}（ $j=1,2,\ldots,M$ ，$i=1,2,\ldots,N_j$ ）這樣第 j 組之樣本均值為：

$$\overline{y}_j=\frac{y_{1j}+y_{2j}+\ldots+y_{N_jj}}{N_j}=\frac{1}{N_j}\sum_{i=1}^{N_j}y_{ij},\quad j=1,2,\ldots,M.$$

而全樣本之平均值為：

$$\overline{y}=\frac{1}{N}\sum_{j=1}^{M}\sum_{i=1}^{N_j}y_{ij}=\frac{1}{N}\sum_{j=1}^{M}N_j\left(\frac{1}{N_j}\sum_{i=1}^{N_j}y_{ij}\right)=\frac{1}{N}\sum_{j=1}^{M}N_j\overline{y}_j$$

其中 $N=\sum_{j=1}^{M}N_j$ 為全樣本之數量。變異數分析之實質是檢定 $\overline{y}$ 是否與 $\overline{y}_j$ 相異。

現在樣本觀測值 y_{ij} 與全樣本均值 $\overline{y}$ 之差異可分為兩個部分：

$$y_{ij}-\overline{y}=y_{ij}-\overline{y}_j+\overline{y}_j-\overline{y}\qquad(21.1)$$

其中 $y_{ij}-\overline{y}_j$ 被稱為組內偏差，$\overline{y}_j-\overline{y}$ 被稱為組間偏差。接下來，將式 (21.1) 左右兩邊平方並加總，可得反映樣本數據波動情況的指標——總離差平方和（Total Sum of Squares，TSS），又稱為總變異：

$$\sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(y_{ij}-\overline{y}\right)^2=\sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(y_{ij}-\overline{y}_j\right)^2+\sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(\overline{y}_j-\overline{y}\right)^2+2\sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(y_{ij}-\overline{y}_j\right)\left(y_{ij}-\overline{y}_j\right)\quad(21.2)$$

由於 $\sum_{i=1}^{N_j}\left(y_{ij}-\overline{y}_j\right)=0$，故式 (21.2) 中最後一項為 0，可簡化為：

$$\begin{aligned}\sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(y_{ij}-\overline{y}\right)^2&=\sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(y_{ij}-\overline{y}_j\right)^2+\sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(\overline{y}_j-\overline{y}\right)^2\\&=\sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(y_{ij}-\overline{y}_j\right)^2+\sum_{j=1}^{M}N_j\left(\overline{y}_j-\overline{y}\right)^2\end{aligned}$$

其中，等號右邊第一項為組內偏差平方和，又被稱為誤差平方和（Error Sum of Squares，ESS）、隨機變異、或組內變異；第二項為組間偏差平方和，又被稱為因子平方和（Factor Sum of Squares，FSS）、因子變異、或組間變異。

17.2.2 自由度

從 TSS、FSS、ESS 之數學表達式可以看出，反應變數的個數對離差平方和大小可能有影響，變數個數越多，離差平方和就有可能越大；變數個數越少，離差平方和就有可能越小。為了消除變數個數對離差平方和大小的影響，我們用離差平方和進行平均，得到平均變異（Mean Square），來衡量不同來源的波動。在引入各平均變異的定義之前，我們先來瞭解一下自由度的概念。

自由度是指當以樣本的統計量來估計母體的參數時，樣本中能夠獨立或自由變動的數據的個數。例如在前面章節中提到，樣本變異數的計算式為：

$$S = \frac{1}{n-1}\sum_{i=1}^{n} N_j \left(x_i - \bar{x}\right)^2 \quad (21.3)$$

其中分母 $n-1$ 就是自由度。為什麼式 (21.3) 中可以自由變動的樣本個數為 $n-1$ 而不是 n？因為該式隱藏了 $\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$ 的這一均值約束，為了滿足這個均值約束，n 個樣本不能都自由變動。考慮三個樣本數據：a,b,c，若這三個數的平均數是 3，即 $\frac{1}{3}\left(a+b+c\right)=3$，那麼可自由變動的數就只有 2 個，例如一旦 a,b 確定，那麼 c 只能為 $3\times 3-\left(a+b\right)$ 而不能自由變動。按照這樣的思路，可以確定出 TSS、FSS 和 ESS 的自由度。

總結起來，我們進行變異數分析的對象共有 N 個樣本觀測值，分佈在 M 個組中，第 j 個組的樣本量為 N_j。

- TSS 是衡量的是 N 個樣本之總波動水平，這裡所有的 N 個樣本並不獨立，它們滿足一個約束條件（均值為 $\bar{y}$）：

$$\sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(y_{ij} - \bar{y}\right) = 0$$

故真正獨立的變數只有 $N-1$ 個，TSS 的自由度為 $N-1$。

- FSS 衡量的是由於因子水平變化導致的反應變數取值之波動。但是，M 個因子組別之均值並不獨立，$\bar{y}_j, j=1,\ldots,M$ 滿足一個約束條件：

$$\sum_{j=1}^{M} N_j\left(\bar{y}_j - \bar{y}\right) = 0$$

因此也丟失一個自由度，FSS 的自由度是 $M-1$，其平均組間變異（因子變異數）為：

$$MSF = \frac{FSS}{M-1} = \frac{1}{M-1}\sum_{j=1}^{M} N_j\left(\bar{y}_j - \bar{y}\right)^2$$

- ESS 反應的是由於樣本與其所處因子水平的組別均值之差異而產生的波動，需滿足 M 個約束條件，

$$\sum_{i=1}^{N_j}\left(y_{ij} - \bar{y}_j\right) = 0, \quad j=1,\ldots,N$$

進而失去了 M 個自由度，所以 ESS 的自由度是 $N-M$，其平均組內變異（隨機變異數）為：

$$MSE = \frac{1}{N-M}\sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(y_{ij} - \bar{y}_j\right)^2$$

TSS、FSS 和 ESS 的自由度滿足以下關係：$N-1=\left(M-1\right)+\left(N-M\right)$。

17.2.3 顯著性檢定

假設反應變數 Y_{ij} 滿足條件：根據因子水平劃分的任一 j 組，$Y_{ij}\left(i=1,2,\ldots,N_j\right)$ 為一組獨立同分佈變數，且服從常態分佈，即 $Y_{ij}\sim N\left(\mu_j,\sigma_0^2\right)$。基於這個假設，可證明出平均組間變異和平均組內變異之期望值滿足下列式子：

$$\mathrm{E}(MSF)=\sigma_0^2+\frac{1}{M-1}\sum_{j=1}^{M}N_j\left(\mu_j-\mu_0\right)^2$$

$$\mathrm{E}(MSE)=\sigma_0^2$$

在虛無假設 $H_0:\mu_1=\mu_2=\ldots=\mu_M=\mu_0$ 下，$\mathrm{E}(MSF)=\mathrm{E}(MSE)=\sigma_0^2$，而且變異數分析之統計量：

$$\varphi=\frac{MSF}{MSE}=\frac{FSS/(M-1)}{ESS/(N-M)}$$

服從 $F(M-1,N-M)$ 分佈。φ 統計量越大，說明平均組間變異 MSF 與平均組內變異 MSE 差異很大，且 $MSF>MSE$，故 MSF 成為樣本總波動之主要貢獻，因子影響十分顯著；φ 統計量很小時，說明平均組內變異 MSE 是主要的變異來源，因子影響不顯著。我們可以查閱 F 分佈的臨界值表，或者計算 p 值來判斷該統計量是否顯著。

現在，我們來總結一下變異數分析的一般步驟：

1. 根據感興趣的因素（因子）之不同取值（水平）將反應變數分成 M 個組；
2. 提出虛無假設 H_0：因子對於反應變數沒有影響；對立假設 H_1：因子對於反應變數有影響；
3. 求出樣本數據中每組的樣本平均值及全樣本均值，算出平均組內變異 MSF 和平均組間變異 MSE；
4. 建構 φ 統計量並計算 φ 值：

$$\varphi=\frac{MSF}{MSE}\sim F(M-1,N-M)$$

5. 由顯著性水平 α（通常取 0.01,0.05,0.1 等）查 F 分佈表的臨界值 $F_\alpha(M-1,N-M)$（或計算 φ 值對應的 p 值）來判斷是接受虛無假設還是拒絕虛無假設：

- 如果 $\varphi>\mathrm{F}_\alpha(\mathrm{M\text{-}1,N\text{-}M})$（或 $p<\alpha$）則拒絕虛無假設，可以推斷認為因子變數對反應變數有影響；
- 如果 $\varphi<\mathrm{F}_\alpha(\mathrm{M\text{-}1,N\text{-}M})$（或 $p>\alpha$）則接受虛無假設，可以推斷認為因子變數對反應變數沒有影響。

17.3 | 變異數分析之 Python 實作

statsmodel 中的 anova 模組可以幫助我們進行變異數分析。利用 anova 模組進行變異數分析的一般途徑是：先建立一個線性迴歸模型[3]，然後再將該模型作為參數傳入專門的函數之中，下面，我們呼叫該模組分別進行單因素變異數分析與多因素變異數分析。

17.3.1 單因素變異數分析

現在來考慮一個簡單的問題，即行業對股票收益率的影響。直覺上來說，行業對股票收益率有重要的影響，不同行業的股票的平均收益率是不一樣的。以美國為例，網際網路等高科技行業的平均股票收益率就比其他行業要高。接下來，針對中國大陸股票市場，我們用變異數分析來驗證一下這個根據經驗形成的直覺。不過值得注意的是，變異數分析的假設條件是各組別各反應變數之間要求獨立同分佈，此處例子可能不恰當，僅以說明 Python 實踐變異數分析之過程。

選擇測試數據為 2014 年的中國大陸股票市場各產業指數的年度收益率，表 17.1 為該數據的前幾項：

▷ 表 17.1: 股票收益率表

Code	Year	Return	Industry
000001	2014	0.57298	貨幣金融服務
000002	2014	0.827567	房地產業
000003	2014	0.336481	醫藥製造業
000004	2014	0.64	房地產業
000005	2014	0.477997	房地產業

3 關於線性迴歸模型，請參見下一章。

接下來，用 Python 撰寫程式碼進行變異數分析。

```
#讀取數據
>>> import pandas as pd
>>> import statsmodels.stats.anova as anova
>>> from statsmodels.formula.api import ols

>>> year_return=pd.read_csv('TRD_Year.csv',\
...                         encoding='gbk')
>>> year_return.head()
   Code  Year    Return Industry
0     1  2014  0.572980   貨幣金融服務
1     2  2014  0.827567     房地產業
2     4  2014  0.336481    醫藥製造業
3     5  2014  0.640000     房地產業
4     6  2014  0.477997     房地產業

#進行變異數分析：
>>> model=ols('Return ~ C(Industry)',\
...           data=year_return.dropna()).fit()
>>> table1 = anova.anova_lm(model)
>>> print(table1)
                 df      sum_sq   mean_sq         F        PR(>F)
C(Industry)    74.0   60.517228  0.817800  4.177614  4.382045e-28
Residual     2302.0  450.634318  0.195758       NaN           NaN
```

上述結果表明，$p = 4.38e-028$，在 0.05 的顯著性水平下，p 值遠遠小於 0.05，故我們應該拒絕虛無假設，認為不同行業股票 2014 年的年收益率不一樣。因此，我們的直覺得到了驗證，即行業是影響股票收益率的一個重要因素的結論。

17.3.2 多因素變異數分析

假設我們想要探討婚姻狀況和受教育水平這兩個因素對個人收入的影響，我們運用多因素變異數分析技術。PSID.csv 包含了 1993 年美國個人收入的相關數據，我們可以利用該數據來進行多因素變異數分析。在 Python 實作上，類比於單因素分析，在線性迴歸模型裡加入多個要研究的因素即可。具體程式碼如下：


```
#讀取數據
>>> PSID=pd.read_csv('PSID.csv')
>>> PSID.head(3)
   Unnamed: 0  intnum  persnum  age  educatn  earnings  hours
kids   married
0           1       4        4   39     12.0     77250   2940
2  married
1           2       4        6   35     12.0     12000   2040
2  divorced
2           3       4        7   33     12.0      8000    693
1   married

#多因素變異數分析
>>> model=ols('earnings ~C(married)+C(educatn)',\
...           data=PSID.dropna()).fit()
>>> table2 = anova.anova_lm(model1)
>>> print(table2)
                df      sum_sq       mean_sq          F        PR(>F)
C(married)     6.0  1.956487e+10  3.260811e+09  15.551238  9.355695e-18
C(educatn)    19.0  2.082990e+11  1.096311e+10  52.284500  9.947527e-180
Residual    4829.0  1.012553e+12  2.096818e+08        NaN           NaN
```

P 值 9.355695e-18 和 9.947527e-180 均遠小於 0.05，即婚姻狀況和受教育水平的係數都顯著，說明這兩個因素對收入水平有著重要的影響。

17.3.3 析因變異數分析

析因變異數分析與多元素變異數分析差不多，僅是多了一個因子的乘項。例如說，在上面的例子中，添加 married 與 educatn 的乘項，以檢定這兩者對收入的影響的大小是否與另一個因子的水平有關。

```
>>> model=ols('earnings ~ C(married)*C(educatn)', data=PSID.
dropna()).fit()
>>> table3 = anova.anova_lm(model)
>>> print(table3)
                          df      sum_sq       mean_sq          F  \
C(married)               6.0  1.956487e+10  3.260811e+09  15.477314
C(educatn)              19.0  2.082990e+11  1.096311e+10  52.035962
C(married):C(educatn)  114.0  2.322040e+10  2.036878e+08   0.966796
Residual              4745.0  9.996921e+11  2.106833e+08        NaN
```

```
9.
10.                                PR(>F)
11.  C(married)                1.160545e-17
12.  C(educatn)               1.334463e-178
13.  C(married):C(educatn)     5.823260e-01
14.  Residual                           NaN
```

第三個係數的 P 值 0.5823>0.05，即結果並不顯著。所以，婚姻狀況和受教育水平對收入的影響大小並不依賴於另一者的水平。

習題

1. 某證券公司對五個地區的分公司的單日開戶數量（單位：個）進行分析，每個地區獲取 10 個營業部的數據，得到資料如下表：

地點	單日開戶數量									
D_1	16	14	10	9	15	14	8	6	13	9
D_2	20	15	16	12	11	10	6	7	9	11
D_3	10	11	13	11	9	8	13	11	6	7
D_4	8	11	6	7	7	9	12	15	10	13
D_5	7	6	8	8	13	6	10	7	5	9

判斷 5 個地區的單日開戶數量是否有顯著性差異，取顯著性水平 $\alpha = 0.05$。

2. 假設有下面一個情景，為考察中央銀行調息對此後一周的加權指數的影響，收集數據如下：

調息方案	指數變化				
方案A	3.12%	2.45%	-1.56%	1.13%	0.55%
方案B	-2.24%	0.56%	1.33%	-1.05%	1.41%
方案C	1.53%	2.41%	-1.88%	-0.31%	-0.65%
方案D	1.05%	1.86%	-0.81%	2.34%	0.33%

取顯著性水平 $\alpha = 0.05$，判斷 4 個調息方案對上證綜指是否有顯著性影響。

3. 同一型號的電池委託 A、B 和 C 三家工廠生產，為比較其質量，從各廠中隨機抽取 5 節電池，試驗後得出其壽命（單位：小時）如下：

工廠	電池壽命				
A	48	38	45	47	40
B	26	35	30	29	31
C	38	40	49	51	51

在顯著性水平 0.05 下檢定電池的平均壽命有無顯著性的差異。

4. 讀取文件 managers.csv，提取 HAM1,HAM3,HAM4 組成一個新的數據集 MANA。

   (a) 求 HAM1, HAM3, HAM4 三個變數的組內變異（SSE）；

   (b) 計算 MANA 三個變數的組間變異（SSA）；

   (c) 計算 MANA 的離差平方和（SST）；

   (d) 分別求出 SST，SSA 和 SSE 的自由度，並說明他們滿足的關係；

   (e) 建構顯著性檢定的 F 統計量，求出統計量的值，比較和相應的 F 分佈的分位值的大小關係，判斷三者的收益率是否存在顯著性的差異。

5. 使用 Python 對數據集 MANA 進行變異數分析，並和題 4 的結果進行比較。