生物統計學講義

第　四　回

704621-4

社團法人 考友社 出版發行

生物統計學講義　第四回

目錄

第五講　變異數分析、相關及迴歸

命題大綱

一、變異數分析
　㈠變異數分析概說
　㈡單因子變異數分析
　㈢二因子變異數分析
二、相關係數
　㈠定義
　㈡公式和定理
　㈢資料散佈圖
　㈣斯皮爾曼等級相關係數
　㈤範例
三、迴歸分析
　㈠簡單迴歸模式
　㈡簡單迴歸模式的推論統計
　㈢新觀察值的預測
　㈣殘差分析

重點整理

一、變異數分析

㈠變異數分析概說：

1. 定義：

⑴變異數分析（Analysis of Variance，ANOVA）是用來檢定各組資料的平均數（mean）是否有差異。檢定兩組資料間的平均數有無差異，可用前面所說的平均數差的檢定方法，也就是所謂的 t 檢定（或此講的變異數分析），而若是對於三組資料或三組以上資料的平均數作檢定，就必須要用變異數分析。

⑵變異數分析的假設條件爲各組樣本所來自的母體變異數相等，其目的是在探究各組的反應，是否因處理的不同而有所差異，作爲日後擬定決策時的參考。它亦可用來檢定兩個以上的母體的平均數是否相等。

⑶當顯著水準 $\alpha=0.05$ 時，想知道 5 組獨立樣本中，哪兩組的平均數有顯著性差異存在時，需作十種檢定：

第一種檢定：P（接受 $H_{0,1}|H_{0,1}$ 是眞）=0.95

第二種檢定：P（接受 $H_{0,2}|H_{0,2}$ 是眞）=0.95

⋮

第十種檢定：P（接受 $H_{0,10}|H_{0,10}$ 是眞）=0.95

因此，P（接受所有 $H_{0,.}|H_{0,.}$ 全是眞）=$(0.95)^{10}$=0.5987

P（拒絕至少一 $H_{0,.}|H_{0,.}$ 全是眞）=1–$(0.95)^{10}$=0.4013

雖然在每一種檢定中，犯第一類型錯誤的機率只有 0.05，但在此十種檢定中彼此是獨立的情況下，犯第一類型錯誤的機率卻高達 40%，更何況作此十種檢定時，所用的資料有些是相同，彼此之間並不是完全獨立（independent）的，若是不完全獨立，其犯第一類型錯誤的機率將更高，所以採用兩兩比較的 t 檢定法，並不是很適宜的。而變異數分析則是一種最常用且可彌補此項缺點的檢定方法，它探討各項變易來源（組間、組內之變異），利用變異數的大小比值，來比較兩組以上的母體平均數是否有顯著性的差異。

2. 名詞解釋：

⑴反應（response）：將個體的特性狀況，以數量的形式來表達，作爲分析的依據，亦稱爲依變數（dependent variable）。

⑵因子（factor）：影響個體特性狀況的因素，亦稱爲自變數（independent variable）。

⑶實驗單位（experimental unit）：接受實驗，並產生特性資料，以供分析的個體或群體。

⑷因子水準（factor level）：每因子的分類狀況，皆代表一個水準。

⑸實驗因子（experimental factor）：若因子的不同水準是因隨機方法外加於各個實驗單位，則此因子稱爲實驗因子，如探究不同的溫度對植物生長的情形，由於在實驗的過程中，溫度可以由實驗者來控制，故溫度爲一實驗因子。

⑹類別因子（classification factor）：若因子的不同水準，並非實驗者所能控制，而係實驗單位本身所具有，則此因子稱爲類別因子。如「性別」即爲一種類別因子。

⑺單因子分析（single-factor analysis）：僅探究某一個因子對反應的影響，則稱爲單因子分析。

⑻多因子分析（multi-factor analysis）：若同時探究兩個或兩個以上的因子對反應的影響，則稱爲多因子分析。

⑼處理（treatment）：即實驗單位所能承受的不同實驗狀況。在單因子分析中，每一個因子水準，即爲一種處理。而在多因子分析中，因子水準的每種組合，都是一種處理。

3.模式假定：

模式能將眞實世界予以簡化，是由數學方程式組成。方程式中的參數能反應眞實世界的本質。所有的模式背後都有一組假定，如果假定不成立，模式將不適用，變異數分析亦不例外。說明如下：

⑴常態性（Normality）：

常態性假定要求誤差項必須遵循常態分配。此一假定將用於假說檢定中，尤其是小樣本且常態性的假定又無法成立時，因 ANOVA 模式不適合作爲分析的工具。變異數分析之目的爲檢定組間均值是否有所差異，若常態性不成立，檢定將會失效，變異數分析也就沒有意義。

⑵恆常性（Constancy）：

恆常性假定要求誤差項之變異數爲一固定常數，不會隨著解釋變數值成遞減或遞增的型態。換句話說，每個組別必須有相同的變異數，各組別之變異數都須相同，才可進行組間均值之比較，否則各組別變異數不相同，是無法比較組間均值是否有所差異的。

⑶獨立性（Independency）：

獨立性假定要求誤差項彼此之間獨立。一般而言，只要樣本來自於橫斷面的抽樣，獨立性絕大多數都可成立；來自於縱斷面的時間序列，獨立性則大多不能成立。例如，探討經濟成長率在兩個不同政黨的執政期間是否有所差異時，前後兩期的觀察值彼此是非獨立的。亦即前一期偏高，本期亦可能偏高。

常態性、恆常性、獨立性三假定是否成立，可藉由殘差分析來確定。運用單因子變異數分析模式時，可先進行分析，取得殘差值後，再進行殘差分析。

㈡單因子變異數分析：

1. 定義：

⑴單因子變異數分析（One-way Analysis of Variance）是指以一個自變數（因子）來解釋反應變數來源的分析方法。一個自變數，稱爲一個因子（factor）。由於只有一個自變數，所以稱爲單因子。實驗時就此因子來分類，分爲 k 個因子水準（也就是 k 組），每個因子水準的樣本個數爲 n_j，$j=1, 2, \ldots, k$，且各個因子水準（組）的處理平均效果爲 u_j，$j=1, 2, \ldots, k$。

⑵由實驗過程中得知每一來源的處理效果，探討各個因子水準的平均處理效果是否有差異存在，確定變異的來源（source of variation）。完全隨機實驗設計（completely randomized experimental design）是以隨機選取獨立樣本的方式，分別給予不同的處理，比較兩個或多個群體在不同處理方式之後的效果差異。就單一因子所做的完全隨機實驗設計，即爲一單因子的實驗模式。

⑶單因子實驗資料如下表㈠：

表㈠　單因子實驗資料表

	處理（treatment）						
	1	2	3	…	…	…	k
	x_{11}	x_{12}	x_{13}				x_{1k}
	x_{21}	x_{22}	x_{23}				x_{2k}
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	$x_{n_1 1}$	$x_{n_2 2}$	$x_{n_3 3}$				$x_{n_k k}$
合計	$T_{\cdot 1}$	$T_{\cdot 2}$	$T_{\cdot 3}$				$T_{\cdot k}$
平均	$\bar{x}_{\cdot 1}$	$\bar{x}_{\cdot 2}$	$\bar{x}_{\cdot 3}$				$\bar{x}_{\cdot k}$

其中 $T_{\cdot k}=\sum_{i=1}^{n_k} x_{ik}$ ，$\bar{x}_{\cdot k}=T_{\cdot k}/n_k$ 。假設每一行的資料所來自的母體爲一平均數 μ_j ，標準差 σ_j 的常態分配， j=1, 2, …， k ，則：

$x_{ij}=\mu_j=e_{ij}$

e_{ij} 爲誤差項（error term）。

令 μ 爲整個母體的平均，則 $\mu=\sum_{j=1}^{k}\mu_j/k$ ，且：

$\mu_j=\mu+\tau_j$

τ_j 爲第 j 個處理的效果（effect），爲一未知參數。

因此，在處理效果固定的模式（fixed effect model）下：

$x_{ij}=\mu+\tau_j+e_{ij}$

i=1, 2, …， n_j

j=1, 2, …， k

且其基本假設爲：

① k 組觀察所得的資料爲 k 個獨立的隨機樣本。

② k 個母體的變異數相等，即 $\sigma_1^2=\sigma_2^2=\cdots=\sigma_k^2=\sigma^2$ 。

③ $\sum_{j=1}^{k}\tau_j=0$

④隨機誤差項爲獨立且相同之常態分配， $e_{ij}\sim N(0,\ \sigma^2)$。

2. 假設檢定的步驟：

⑴建立假設：檢定因子處理效果是否相同。

虛無假設 H_0 ： $\mu_1=\mu_2=\cdots=\mu_k=\mu$（表示各組母體平均數相等）

對立假設 H_1 ：並非所有的 μ_i 都相等（表示各組母體平均數至少有一不相等）

或

虛無假設 H_0 ： $\tau_1=\tau_2=\cdots=\tau_k=0$（表示處理的效果相同）

對立假設 H_1 ：並非所有的 τ_j 都相等（表示處理的效果至少有一不同）

⑵計算各種變異數：此時變異數的來源是來自於因子及隨機誤差項。

總變異 = 因子變異 + 隨機誤差變異

SS_t = SS_b + SS_w

① SS_t（sum of square of total）：

$$SS_t=\sum_{j=1}^{k}\sum_{i=1}^{n_j}x_{ij}^2-T_{\cdot\cdot}^2/N$$

② SS_b（sum of square due to factor, among groups）：

$$SS_b = \sum_{j=1}^{k} T_{\cdot j}^{\,2} / n_j - T_{\cdot\cdot}^{\,2} / N$$

③ SS_w（sum of square of error, within group）：

$$SS_w = \sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}^2 - \sum_{j=1}^{k} T_{\cdot j}^{\,2} / n_j = SS_t - SS_b$$

⑶計算各種自由度：

① SS_t 的自由度爲 $N-1$

② SS_b 的自由度爲 $k-1$

③ SS_w 的自由度爲 $N-k$

⑷計算各種均方（mean of square）：

① $MS_b = SS_b/(k-1)$

② $MS_w = SS_w/(N-k)$

⑸計算 F 值：$F = MS_b / MS_w \sim F_{(k-1,\ N-k)}$

⑹查表找 F 臨界值：若顯著水準爲 α，則由 F 分配表，找出臨界值 $F_{(\alpha,\ k-1,\ N-k)}$，當 $F > F_{(\alpha,\ k-1,\ N-k)}$ 時，拒絕 H_0，表示各組處理的平均效果有顯著性差異存在。反之則接受 H_0。

表(二)　單因子變異數分析表

變異來源	變異數	自由度	均方	F 值
因子（組間）	SS_b	$k-1$	MS_b	$F = MS_b / MS_w$
誤差（組內）	SS_w	$N-k$	MS_w	
總　和	SS_t	$N-1$		

3.事後比較：

⑴延續 2.之步驟，若拒絕 H_0，我們想要知道哪兩組處理的平均效果有顯著性差異存在，此時可以兩兩作比較，此即爲事後的多重比較（multiple comparison）。若有 k 組，則有 C_2^k 種檢定法，檢定的次數愈多，個別比較的顯著水準就愈小。故有時會對顯著水準作調整，若顯著水準爲 α，則每一個別比較的顯著水準爲 $\alpha^* = \alpha / C_2^k$，此爲 Bonferroni adjustment。若有 5 組，則有 C_2^5=10 種檢定。在顯著水準 α=0.05 的情況下，每一個別比較的顯著水準爲 α^*=0.05/ C_2^5=0.005。多重比較的虛無假設爲 H_0：$\mu_i = \mu_j$，$i \neq j$，也就是兩兩成對組合之間的處理效果相同，兩組的平均數相等。

⑵多重比較的方法：

①最小顯著差異法（The Least Significant Difference Method）：

檢定統計量 $t_{ij}=\frac{\bar{x}_i-\bar{x}_j}{\sqrt{MS_w(1/n_i+1/n_j)}}$ 爲一自由度 $N-k$ 之雙尾之 t 檢定。而 $LSD=t_{a/2,\ N-k}\sqrt{MS_w(1/n_i+1/n_j)}$ 稱爲最小顯著差異（the least significant difference）。

②Scheffe's Significant Difference Test：

$$SSD=\sqrt{(k-1)F_{\alpha,k-1,N-k}}\ \sqrt{MS_w(1/n_i+1/n_j)}$$

③ Tukey's Honestly Significant Difference Test：

$$HSD=q_{\alpha,\ k,\ N-k}\sqrt{\frac{MS_w}{n}}$$

其中 k 爲實驗之組數，N 爲實驗之總個數，n 爲各組內的實驗個數，$q_{\alpha,\ k,\ N-k}$ 爲一常數。若各組內的實驗個數不相等時，則：

$$HSD=q_{\alpha,\ k,\ N-k}\sqrt{MS_w(1/n_i+1/n_j)/2}$$

④ Duncan's Multiple Range Test：

$$R_p=r_\alpha(p,f)\sqrt{\frac{MS_w}{n}}$$

其中 p=2,…, k ，f 爲組內變異（誤差項）的自由度，n 爲各組內的實驗個數，r_α(p, f)爲一常數。若各組內的實驗個數不相等時，則上式中之 n，以$\{n_j\}$的調和平均數 $n_h=k/(\sum_{j=1}^{k}n_j)$ 代之。首先，各種處理的平均數須先依遞增順序排序，而後依序計算成對兩組間的平均數差，p 爲成對兩組間的間隔數加 1，例如，在 k =5 的情況下，若要將排序後最大值之組與最小值之組作比較時，p=5，又若是要將最大值之組與第二小值之組作比較時，p=4，以此類推。

⑶多重比較法的檢定步驟：

①選定顯著水準（significance level）。

②計算所有可能兩兩組合之間樣本平均數的差異（the difference of two sample mean）。

③計算 LSD（或 SSD, HSD, R_p）。

④若任兩組之間的差異大於 LSD（或 SSD, HSD, R_p），則表示兩組之間有顯著差異；若任兩組之間的差異小於 LSD（或 SSD, HSD, R_p），則表示兩組之間沒有顯著性差異存在。

4.範例：

精選試題

一、某藥商想知道三種藥物 P、Q、R 在降血壓上的效果。今有 15 名患者，隨機分成 A、B、C 三組，每組 5 人，分別給予三種藥物 P、Q、R 治療，在服用一個月後，測其血壓降低值，結果如下：（mmHg）

	A	B	C
1	11	10	10
2	10	14	11
3	8	13	8
4	9	12	9
5	12	11	12

試以 α=0.05 之顯著水準，檢定不同的藥物治療是否有顯著性的差異？

答：㈠虛無假設 H_0：不同的藥物治療沒有差異

對立假設 H_1：不同的藥物治療有差異

㈡變異數分析表：

變異來源	SS	df	MS	F
組間	13.33	2	6.67	2.67
組內	30	12	2.5	

㈢ α=0.05， $F_{(0.05,2,12)}$=3.89

㈣因爲 2.67<3.89，所以不能拒絕 H_0

即不同的藥物治療沒有顯著性的差異存在。

二、從某醫院 A、B、C 三種疾病之衆多病患中，隨機抽取若干名，調查其初診年齡，結果如下，問該醫院 A、B、C 三種病人之初診年齡是否有所不同？(α=0.05)

	A	B	C
1	29	32	24
2	28	35	25
3	27	30	26
4	28	34	24
5	29	29	25
6	27		27
7			24

答：㈠虛無假設 H_0 ：三種病人的初診年齡沒有不同

對立假設 H_1 ：三種病人的初診年齡有不同

㈡臨界值： α=0.05， $F_{(0.05,2,15)}$=3.68

㈢拒絕域： F>3.68

㈣變異數分析表：

變異來源	SS	df	MS	F
組間	142.94	2	71.47	28.2
組內	38	15	2.53	

㈤因爲 28.2>3.68，所以拒絕 H_0

即 A、B、C 三種病人之初診年齡有顯著的不同。

三、某復健師想瞭解不同性別患者，在不同教導方式下，學會使用某種復健器有所需的時間（單位：分）是否有所差異，記錄三日來所需的時間，得到如下的結果。

教導方式 / 性別	I	II	III
男性	45, 22, 50	45, 26, 19	34, 52, 76
女性	56, 47, 68	33, 47, 52	42, 55, 62

試以 α=0.05 之顯著水準，對這組資料作一些檢定與結論。是否患者性別不同及教導方式不同對所需的時間有差異？

答：㈠建立假設：

1. 虛無假設 H_0 ： $\tau_1=\tau_2$ （表示性別不同對其所需的學習時間沒有差異）

   對立假設 H_1 ： $\tau_1 \neq \tau_2$ （表示性別不同對其所需的學習時間有差異）

2. 虛無假設 H_0 ： $r_{\mathrm{I}}=r_{\mathrm{II}}=r_{\mathrm{III}}$ （表示教導方式不同對其所需的學習時間沒有差異）

   對立假設 H_1 ：並非所有的 r_j 都相等（表示教導方式不同對其所需的學習時間有差異）

3. 虛無假設 H_0 ：性別與教導方式沒有交互作用存在

   對立假設 H_1 ：性別與教導方式有交互作用存在

㈡二因子變異數分析表：

變異來源	變異數	自由度	均方	F 值
性別因子	480.5	1	480.5	2.487
教導方式因子	847	2	423.5	2.192
交互作用	301	2	150.5	0.799