變異數分析(ANOVA)

大數據分析

- R/Python/Julia/SQL程式設計與應用 (R/Python/Julia/SQL Programming and Application)
- 資料視覺化 (Data Visualization)
- 機器學習 (Machine Learning)
- 統計品管 (Statistical Quality Control)
- 最佳化 (Optimization)

李明昌博士
alan9956@gmail.com
http://rwepa.blogspot.com/

2024.05.22

大綱

- 1.F分配-Rcmdr
- 2.變異數分析簡介
- 3.單因子變異數分析
- 4.二因子變異數分析
- 5.變異數分析-Rcmdr

1.F分配-Rcmdr

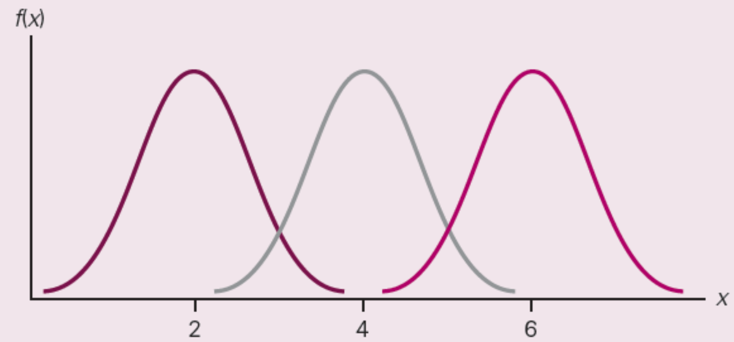
常態分配

- 常態分配兩個參數:
  - 平均數 μ
  - 標準差 σ

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,,\quad -\infty<x<\infty$$

- μ 增加時曲線平移到右邊, 例: μ=2, 4, 6

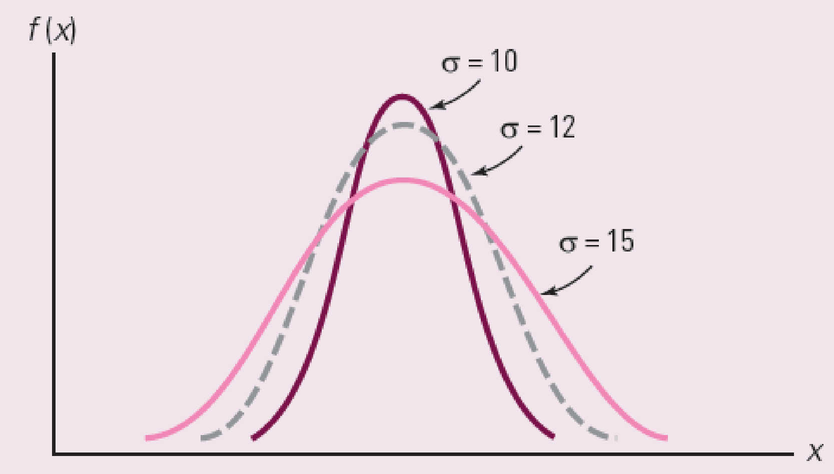
變異數相同但平均數不同的常態分配



常態分配(續)

- 下圖描述 σ 的效果。大的 σ 值曲線會變寬，而小的σ 值曲線會變窄。

平均數相同但變異數不同的常態分配



標準常態分配

- 一個常態分配的平均數為零且其標準差為 1 稱為標準常態分配 (standard normal distribution)。
- 若Z 為一常態隨機變數且 $R.V.Z\sim N(0,1)$ ，則稱 Z 具有標準常態分配，其機率密度函數為

$$f(z)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}z^2} \quad -\infty<z<\infty$$

- 一般統計學教科書附錄為標準常態分配之累積機率值。

標準常態分配-查表法

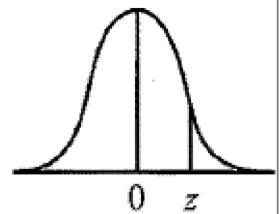
- https://github.com/rwepa/DataDemo/blob/master/prob.pdf

標準常態分配之累積機率值(續)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064		
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443		
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808		
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157		
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486		
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794		
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078		
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817

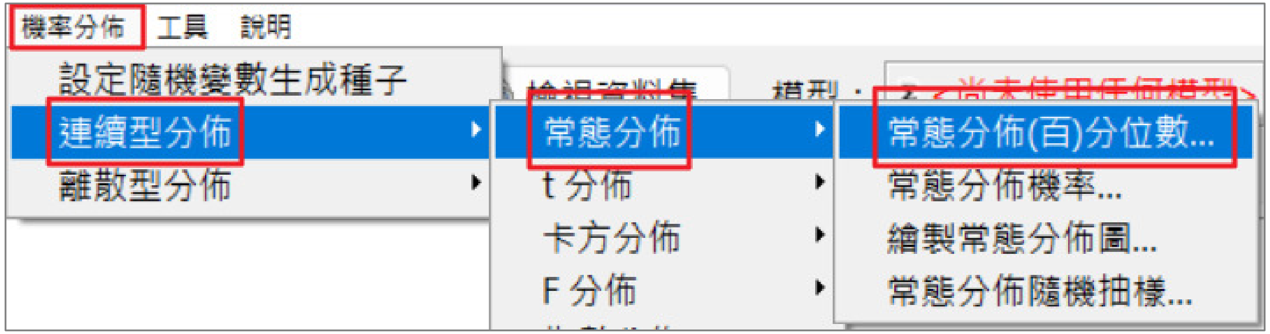
附表三：標準常態分配之累積機率值

$$P(Z \le z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

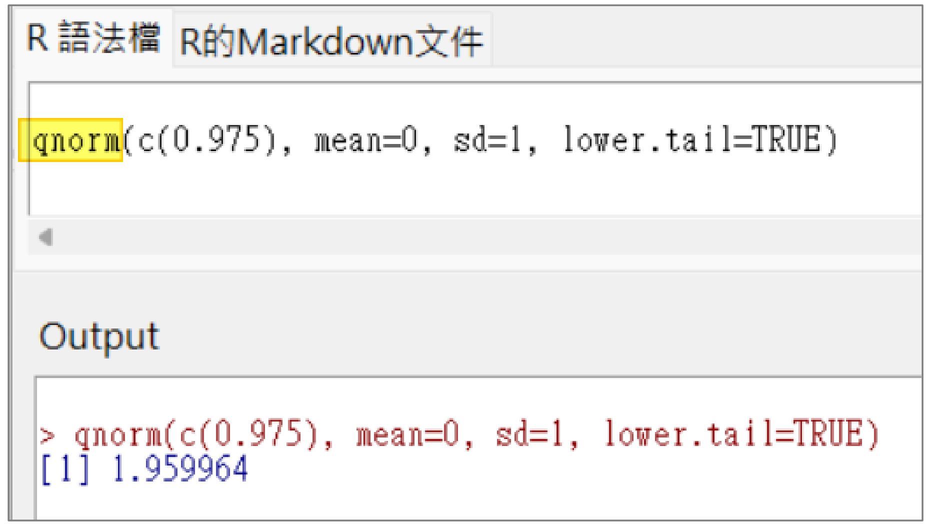


- z=1.96
- $P(Z \leq 1.96) = 0.975$

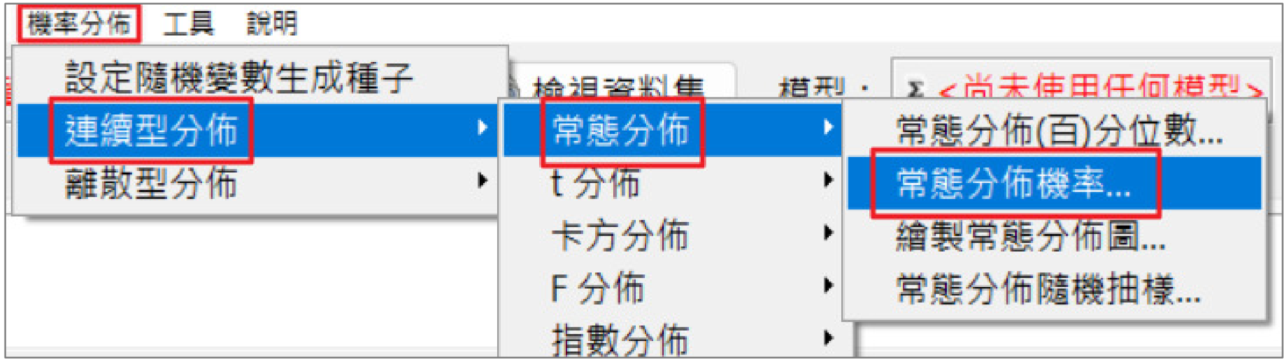
常態分配Rcmdr-qnorm



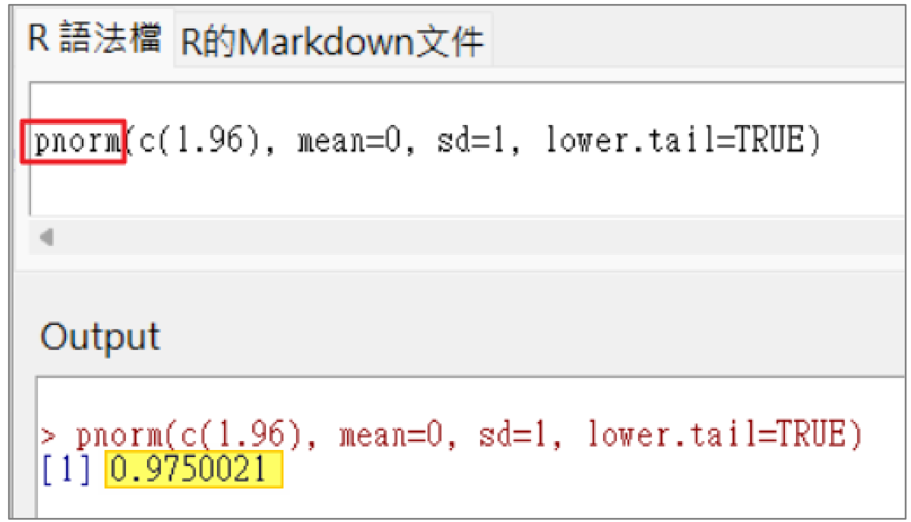
- qnorm:
- 機率值→變數值
- 0.975 → 1.96



常態分配Rcmdr-pnorm



- pnorm:
- 變數值→機率值
- 1.96 → 0.975



卡方分配

- 假設一組隨機樣本 $X_1,\ X_2,\ \cdots, X_n$ 取自一常態分配 $N(\mu, \sigma^2)$ 之母體，則自由度為 $n-1$ 的卡方分配（chi-square distribution）。

$$R.V.\quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

- 若k個隨機變數是相互獨立且符合標準常態分配，期望為0且變異數為1，則隨機變數Z的平方和稱為服從自由度為 k 的卡方分配，$X = \sum_{i=1}^{k} Z_i^2$ ，即 $X \sim \chi^2(k)$或 $X \sim \chi_k^2$，機率密度函數為：

$f_k(x) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{\frac{-x}{2}}, x \geq 0$, 如果 $x \leq 0$，則 $f_k(x) = 0$， Γ為 $Gamma$ 函數，

Gamma function: $\Gamma(n) = (n-1)! = (n-1) \times (n-2) \ldots \times 3 \times 2 \times 1$

t 分配

- 考慮Z為具有標準常態分配之隨機變數且V為具有自由度 (degree of freedom) v 的卡方分配之隨機變數。若 Z、V 為互相獨立，則稱 T為具有自由度 v 之 t 分配或 Student t 分配 ，一般以 $R.V.T \sim t(v)$ 表示。

$$R.V.T = \frac{Z}{\sqrt{\frac{V}{v}}}$$

常態分配 + 卡方分配 → t分配

- t-分配其機率密度函數

$$f(t) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right) \cdot \sqrt{\pi v}} \left(1 + \frac{t^2}{v}\right)^{-\left(\frac{v+1}{2}\right)} \text{ , } -\infty < t < \infty$$

F分配

- 考量 U、V 分別為具有自由度 u 及 v 之卡方分配的隨機變數且U、V 互相獨立，則稱隨機變數 $W = \frac{U/u}{V/v}$ 具有自由度 $(u\,,\,v)$ 之 F分配 。
- 一般以 $RV.W \sim F(u,v)$ 表示，其機率密度函數為

$$f(x) = \frac{\Gamma\left(\frac{u+v}{2}\right)\left(\frac{u}{v}\right)^{\frac{u}{2}} x^{\frac{u}{2}-1}}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)\left[\left(\frac{u}{v}\right)x+1\right]^{\frac{u+v}{2}}} \quad , \quad 0 < x < \infty$$

F-分配（續）

- 若$R.V.\ X \sim F(u,v)$ ，則 $R.V.\ \frac{1}{X} \sim F(v,u)$
- 令 σ_1^2 、 σ_2^2 分別表示兩個常態分配母體之變異數且 S_1^2 、S_2^2 分別為取自於此兩母體之樣本變異數且其樣本個數分別為 n 及 m ，

  則 $W = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$ 稱為具有自由度 $(n-1,\ m-1)$ 之F分配。
- F分配轉換技巧：$F_{1-\alpha}(v_1,v_2) = \frac{1}{F_\alpha(v_2,v_1)}$

F分配-範例

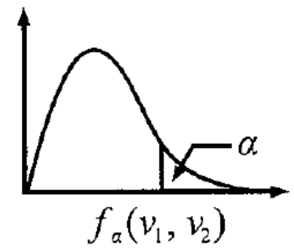
- 考慮 R.V.X~F(3,9) ，求 k 使得 $P(X > k)$=0.05?
- 解：
- 本例為計算 $f_{0.05}(3,9)$值為何?由附表之 F-分配之臨界值表（參見次頁）可知，$v_1 = 3$，$v_2 = 9$ 所對應之位置為 $f_{0.05}(3,9) = 3.8625$，
- 即 $k = f_{0.05}(3,9) = 3.8625$

F分配-查表法

附表六：*F* 分配之臨界值

$$P(F \geq f_\alpha(v_1, v_2)) = \alpha$$

$$f_{0.05}(v_1, v_2)$$

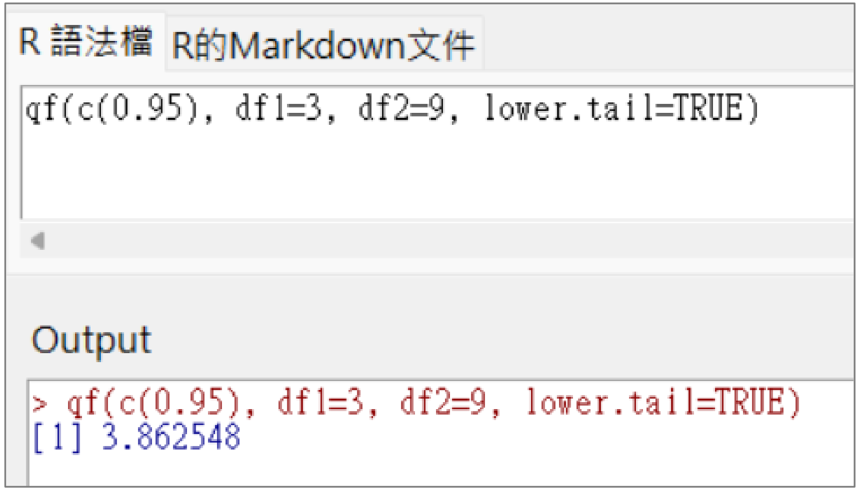


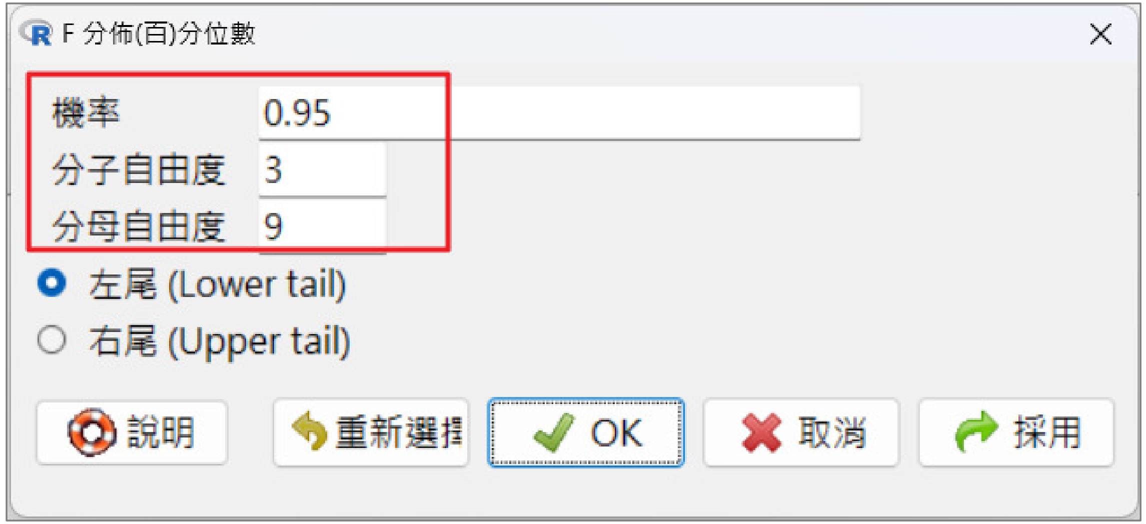
v_1, v_2表自由度$(d.f.)$

v_2 \ v_1	1	2	3	4	5	6	7	8	9	10	11	12
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	242.98	243.90
2	18.513	19.000	19.164	19.247	19.296	19.329	19.353	19.371	19.385	19.396	19.405	19.412
3	10.128	9.5521	9.2766	9.1172	9.0134	8.9407	8.8867	8.8452	8.8123	8.7855	8.7633	8.7447
4	7.7086	6.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	5.9988	5.9644	5.9358	5.9117
5	6.6079	5.7861	5.4094	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725	4.7351	4.7040	4.6777
6	5.9874	5.1432	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990	4.0600	4.0274	3.9999
7	5.5915	4.7374	4.3468	4.1203	3.9715	3.8660	3.7871	3.7257	3.6767	3.6365	3.6030	3.5747
8	5.3176	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881	3.3472	3.3129	3.2839
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789	3.1373	3.1025	3.0729
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782	2.9430	2.9130

F分配Rcmdr

- qf( )，$F_{1-0.05}(3,9)$結果與上一頁查表結果相同。





2.變異數分析簡介

- The Correlation Between Relatives on the Supposition of Mendelian Inheritance. *Ronald A. Fisher*. Philosophical Transactions of the Royal Society of Edinburgh. 1918. (volume 52, pages 399–433)
- https://zh.wikipedia.org/zh-tw/羅納德·愛爾默·費雪 , 英國統計學家

變異數分析簡介

- 變異數分析原文為ANOVA(ANalysis Of VAriance)
- 變異數分析是用來檢定3個(含)以上母體平均數是否相等或是統計上顯著的。
- 變異數分析不是檢定變異數。
- 為何取名變異數分析?
- 範例:
  - 不同的行銷策略是否會影響產品之平均銷售量?
  - 不同的教育程度與不同的性別對工作滿意度是否有影響?
  - 不同排水道的排水化學污染指數(drain chemical pollution index,DCPI)是否有不同影響?
  - 不同土壤性質對農作物生長是否有不同影響?[公務人員考試試題]
    - https://www.public.tw/prog/Howard/LODManage/DataManage/Upload/LOD002/2021010119 0525.pdf

變異數分析簡介 (續)

- 實驗單位(experiment unit)：接受試驗的人或物。
  - 例如：產品、員工、機器為其實驗單位。
- 因子(factor)：研究者所能控制或調整的因素。
  - 例如：加熱溫度、行銷策略、教育程度為因子。
- 處理(treatment)：因子之各種水準或類別，為類別型變數。
  - 例如：加熱溫度 (低溫、中溫、高溫) 。
- 依變數(dependent variable)：實驗單位對不同處理方法的反應變數。
  - 例如：銷售量、工作滿意度、農作物生長為依變數。
  - 一般使用自變數為X，依變數為Y。

自變數 vs.反應變數

- https://github.com/rwepa/DataDemo?tab=readme-ov-file#variables
- X: 自變數
  - 獨立變數 independent variable,
  - 預測變量 predictor variable,
  - 解釋變量 explanatory variable,
  - 共變量 covariate.
- Y: 反應變數 response variable
  - 因變數, 依變數, 應變數, 被解釋變數 dependent variable,
  - 結果變數 outcome variable.

3.單因子變異數分析

ANOVA資料架構

處理	觀測值				總和	平均值
1	y_{11}	y_{12}		y_{1n}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	$\cdots$	y_{2n}	$y_{2.}$	$\bar{y}_{2.}$
$\cdots$				$\cdots$		
a	y_{a1}	y_{a2}		y_{an}	$y_{a.}$	$\bar{y}_{a.}$
					$y_{..}$	$\bar{y}_{..}$

- 平均值模型(Means model)
  - $y_{ij} = \mu_i + \varepsilon_{ij}, i = 1,2,\dots,a;\ j = 1,2,\dots,n$。
- 效果模型(Effects model)
  - 考慮 $\mu_i = \mu + \tau_i, i = 1,2,\dots,a$。
  - $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$, μ 總平均, τ_i表示第i個處理效果。

- 處理有 a個(或因子有 a個水準)
- y_{ij}: 第 i 個處理，第 j 個觀測值
- μ_i: 第i處理的平均數, 即 $\bar{y}_{i.}$
- ε_{ij}: 隨機誤差, $E(\varepsilon_{ij}) = 0$, 即$E(y_{ij}) = \mu_i$

資料模型

- 檢定a個處理的平均值是否相等
  - $H_0: \mu_1 = \mu_2 = \cdots = \mu_a$
  - H_1: 至少有一對(i, j)使得$\mu_i \neq \mu_j$
- 重複量測表示同一組實驗單位測量兩次以上。
- $y_{i.} = \sum_{j=1}^{n} y_{ij}\, , i = 1,2,\ldots,a$
- $\bar{y}_{i.} = \frac{y_{i.}}{n}$
- $y_{..} = \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}$
- $\bar{y}_{..} = \frac{y_{..}}{N}, N = an$總觀測值個數

處理	觀測值				總和	平均值
1	y_{11}	y_{12}		y_{1n}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	$\cdots$	y_{2n}	$y_{2.}$	$\bar{y}_{2.}$
$\cdots$				$\cdots$		
a	y_{a1}	y_{a2}		y_{an}	$y_{a.}$	$\bar{y}_{a.}$
					$y_{..}$	$\bar{y}_{..}$

$$\because \mu = \frac{\sum_{i=1}^{a} \mu_i}{a} = \frac{\sum_{i=1}^{a} (\mu + \tau_i)}{a} = \frac{a\mu + \sum_{i=1}^{a} \tau_i}{a} = \mu + \frac{\sum_{i=1}^{a} \tau_i}{a}$$

$\therefore \sum_{i=1}^{a} \tau_i = 0$, 其中 $\tau_i = \mu_i - \mu$

- $H_0: \tau_1 = \tau_2 = \cdots = \tau_a = 0$
- H_1: 至少有一個i 使得$\tau_i \neq 0$

ANOVA假設檢定之條件

- 依變數(Dependent variable)是連續型變數。
- 依變數的母體符合常態分配。
- 考慮隨機誤差 ε_{ij}獨立且服從常態隨機變數。
- ε_{ij}平均數為0，變異數為σ^2，即變異數具同質性，各組母體變異數假設相等。
- y_{ij}為彼此獨立 $y_{ij} \sim N(\mu + \tau_i, \sigma^2)$。
- ANOVA一般採用完全隨機設計(completely randomized design)，即研究者將不同的處理方法以隨機方式分派給實驗單位。

總平方和分解法

$$(a+b)^2 = a^2 + 2ab + b^2$$

- 總平方和 (Total Sum of Squares)

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{..}\right)^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}\left[\left(\bar{y}_{i.} - \bar{y}_{..}\right) + \left(y_{ij} - \bar{y}_{i.}\right)\right]^2$$

$$= n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 + 2\sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{i.} - \bar{y}_{..})\left(y_{ij} - \bar{y}_{i.}\right) + \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i.}\right)^2$$

$$\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i.}\right) = y_{i.} - n\bar{y}_{i.} = y_{i.} - n\frac{y_{i.}}{n} = 0$$

$$= n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i.}\right)^2$$

= 平均與總平均差的平方和 + 處理內觀測值與處理平均差的平方和

= 處理平均間差異 + 隨機誤差

- $SS_T = SS_{Treatment} + SS_E$ = 處理平方和(處理間) + 誤差平方和(處理內)

變異數分析表

變異來源	平方和	自由度	均方	F_0
處理間	$SS_{Treatment}$ （組間變異）	$a-1$	$MS_{Treatment}=\dfrac{SS_{Treatment}}{a-1}$	$F_0=\dfrac{MS_{Treatment}}{MS_E}$
誤差	SS_E （組內變異）	$N-a$	$MS_E=\dfrac{SS_E}{N-a}$	
總和	SS_T （總變異）	$N-1$		

註:

1. 總樣本數 $N=a\times n$
2. 處理間平方和 $SS_{Treatment}=n\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2$
3. 誤差平方和 $SS_E=SS_T-SS_{Treatment}$
4. 總平方和 $SS_T=\sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_{..}\right)^2$
5. 誤差項自由度 $(N-1)-(a-1)=N-a$
6. 如果 $F_0\geq F_\alpha(a-1,N-a)$, 則表示 p值 $<\alpha$, 即拒絕H_0, 接受不同處理平均數有不同.

ANOVA範例

- 某市場調查公司欲調查市面上四種品牌之相同口味飲料之平均銷售量是否相同，針對每一品牌隨機選定5個地區作調查，各地區單月之銷售量如下表(單位：千箱) 所示。

地區 \ 品牌	A	B	C	D
1	26.5	29.0	26.9	30.5
2	28.7	27.6	28.3	31.2
3	25.2	25.4	27.8	29.9
4	29.3	28.3	26.2	28.1
5	25.3	29.7	25.8	30.3

ANOVA範例（續）

1. 請寫出此問題之假設。
2. 請寫出此問題之變異數分析表。
3. 請根據變異數分析表之結果，以α=0.05 檢定此四種品牌飲料之平均銷售量是否相等。

ANOVA範例（續）

(1)令 μ_i 表第i 種品牌銷售量之平均數，則此問題之假設：

H_0：$\mu_A = \mu_B = \mu_C = \mu_D$，$H_1$：$\mu_A$、$\mu_B$、$\mu_C$、$\mu_D$不全相等

(2)每種品牌之樣本平均數 $\overline{X}_A$、$\overline{X}_B$、$\overline{X}_C$、$\overline{X}_D$ 及總樣本平均數 $\overline{X}$ 如下：

$$\overline{X}_A = \frac{1}{5}(26.5 + 28.7 + 25.2 + 29.3 + 25.3) = 27$$

$$\overline{X}_B = \frac{1}{5}(29.0 + 27.6 + 25.4 + 28.3 + 29.7) = 28$$

$$\overline{X}_C = \frac{1}{5}(26.9 + 28.3 + 27.8 + 26.2 + 25.3) = 27$$

$$\overline{X}_D = \frac{1}{5}(30.5 + 31.2 + 29.9 + 28.1 + 30.3) = 30$$

$$\overline{X} = \frac{1}{20}(5 \times 27 + 5 \times 28 + 5 \times 27 + 5 \times 30) = 28$$

ANOVA範例（續）

$$SST = (26.5-28)^2 + (28.7-28)^2 + (25.2-28)^2 + (29.3-28)^2$$
$$+ (25.3-28)^2 + (29.0-28)^2 + (27.6-28)^2 + (25.4-28)^2$$
$$+ (28.3-28)^2 + (29.7-28)^2 + (26.9-28)^2 + (28.3-28)^2$$
$$+ (27.8-28)^2 + (26.2-28)^2 + (25.8-28)^2 + (30.5-28)^2$$
$$+ (31.2-28)^2 + (29.9-28)^2 + (28.1-28)^2 + (30.3-28)^2$$
$$= 65.28$$

$$SSB = 5 \times (27-28)^2 + 5 \times (28-28)^2 + 5 \times (27-28)^2 + 5 \times (30-28)^2 = 30$$

$$SSE = SST - SSB = 65.28 - 30 = 35.28$$

ANOVA範例（續）

- SSB之自由度為 $4-1=3$，$MSB=\frac{SSB}{k-1}=\frac{30}{3}=10$
- SSE之自由度為 $n-k=20-4\ =\ 16$，
$$MSE=\frac{SSE}{n-k}=\frac{35.28}{16}=2.205$$
- $f_0=\frac{MSB}{MSE}=\frac{10}{2.205}=4.535$

ANOVA範例（續）

- 變異數分析表如下：

變異來源	平方和	自由度	均方	值
處理方法 隨機誤差	30 35.28	3 16	10 2.205	$f_0 = 4.535$
總和	65.28	19		

- (3) 因為$\frac{MSB}{MSE} \sim F(3,16)$，因此其拒絕域$\{f_0 \geq f_{0.05}(3,16)\} = \{f_0 \geq 3.239\}$。
- 檢定值 $f_0 = 4.535 > 3.239$ 落在拒絕域中，因此拒絕 H_0 ，即四種不同品牌飲料之平均銷售量有顯著地差異。

4.二因子變異數分析

二因子變異數分析（續）

- 考量 x_{ijk} 為獨立之常態隨機變數，$i = 1,2,\cdots,a$ ， $j = 1,2,\cdots,b$，$k = 1,2,\cdots,n$，且變異數均相等。
- 母體平均數 $\mu_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$，其中$\alpha_i$ 、β_j、$(\alpha\beta)_{ij}$ 分別表示第一因子、第二因子及兩因子交互作用造成之偏差，考慮以下三種情形：

二因子變異數分析-假設

- (1) 在 $H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_a = 0$ 成立時，即第一因子不影響依變數之條件下，$\frac{MSA}{MSE} \sim F\big(a-1, ab(n-1)\big)$
- (2)在 $H_0': \beta_1 = \beta_2 = \ldots = \beta_b = 0$ 成立時，即第二因子不影響依變數之條件下，$\frac{MSB}{MSE} \sim F\big(b-1, ab(n-1)\big)$
- (3)在 $H_0'': (\alpha\beta)_{11} = (\alpha\beta)_{12} = \ldots = (\alpha\beta)_{1b} = \ldots = (\alpha\beta)_{a1} = (\alpha\beta)_{a2} = \ldots = (\alpha\beta)_{ab}$ 成立時，即兩因子交互作用不影響依變數之條件下，$\frac{MSAB}{MSE} \sim F\big((a-1)(b-1), ab(n-1)\big)$

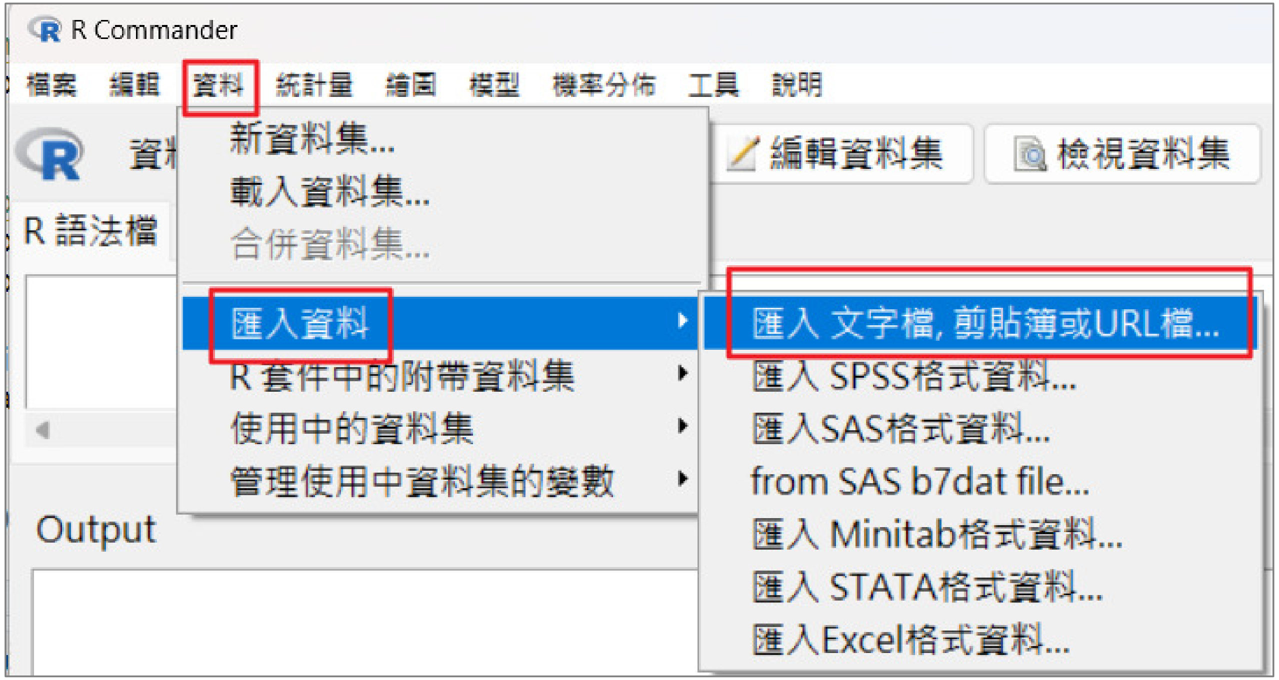
二因子變異數分析表

變異來源	平方和	自由度	均方	f值
A因子	SS_A	$a-1$	MS_A	$f_1=\dfrac{MS_A}{MS_E}$
B因子	SS_B	$b-1$	MA_B	$f_2=\dfrac{MS_B}{MS_E}$
交互作用 (AB)	SS_{AB}	$(a-1)(b-1)$	MS_{AB}	$f_3=\dfrac{MS_{AB}}{MS_E}$
誤差	SS_E	$ab(n-1)$	MS_E	
總和	SS_T	$abn-1$		

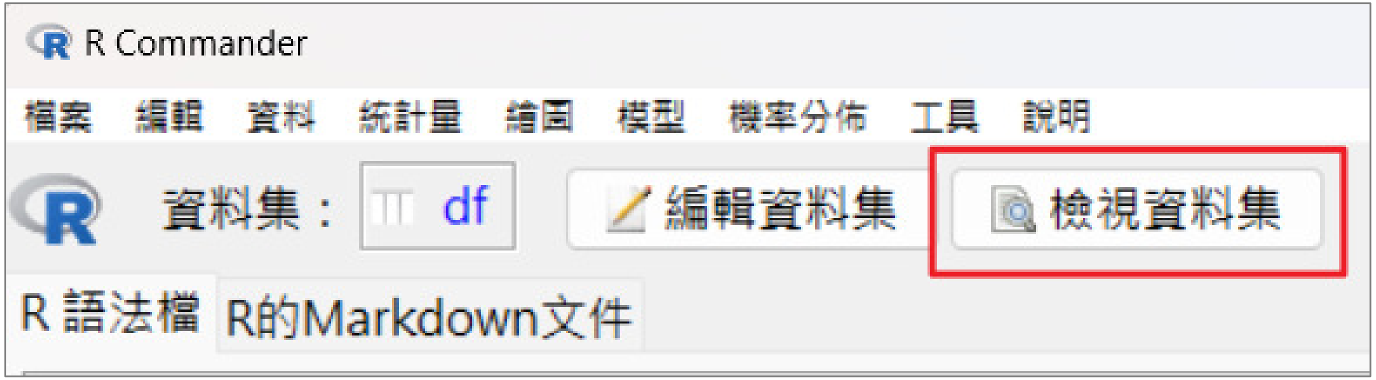
5.變異數分析-Rcmdr

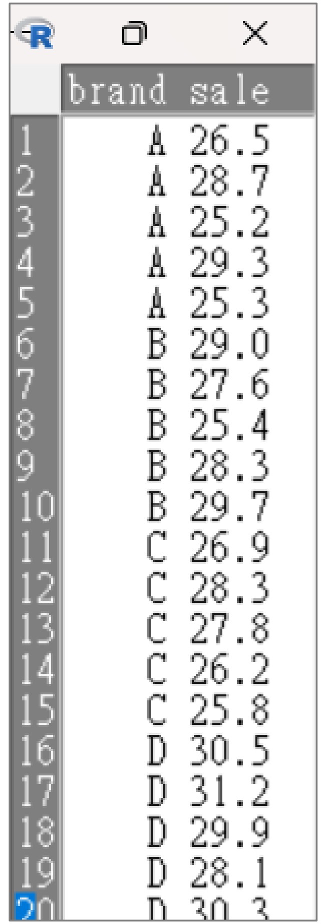
beverage.csv

- URL: https://raw.githubusercontent.com/rwepa/DataDemo/master/beverage.csv



檢視資料集





	brand	sale
1	A	26.5
2	A	28.7
3	A	25.2
4	A	29.3
5	A	25.3
6	B	29.0
7	B	27.6
8	B	25.4
9	B	28.3
10	B	29.7
11	C	26.9
12	C	28.3
13	C	27.8
14	C	26.2
15	C	25.8
16	D	30.5
17	D	31.2
18	D	29.9
19	D	28.1
20	D	30.3

資料結構 str

- str(df)

```
str(df)
```

執行語法

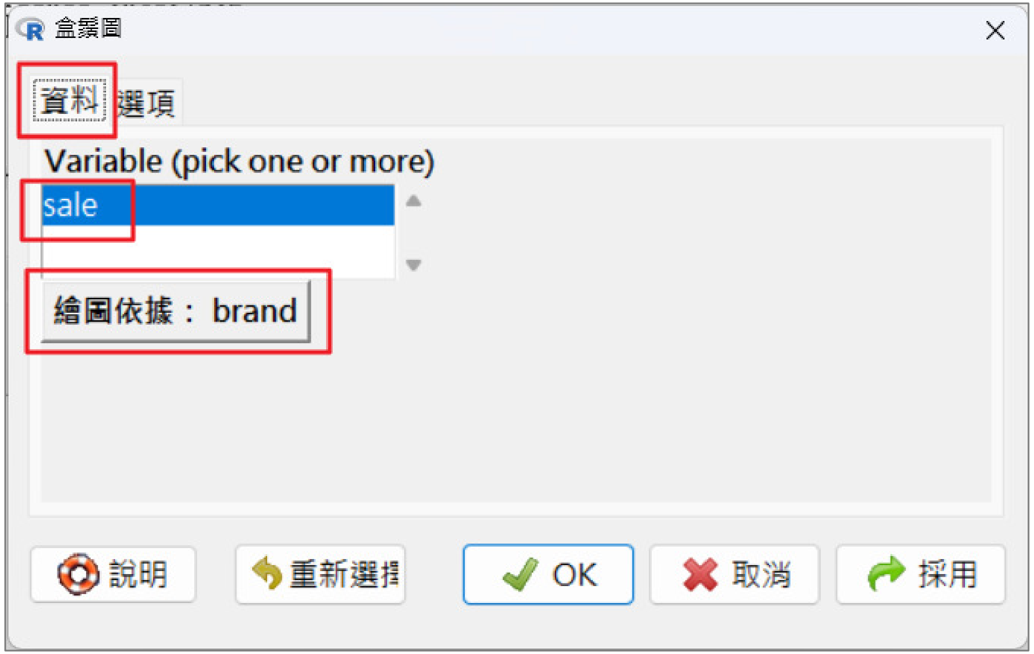
Output

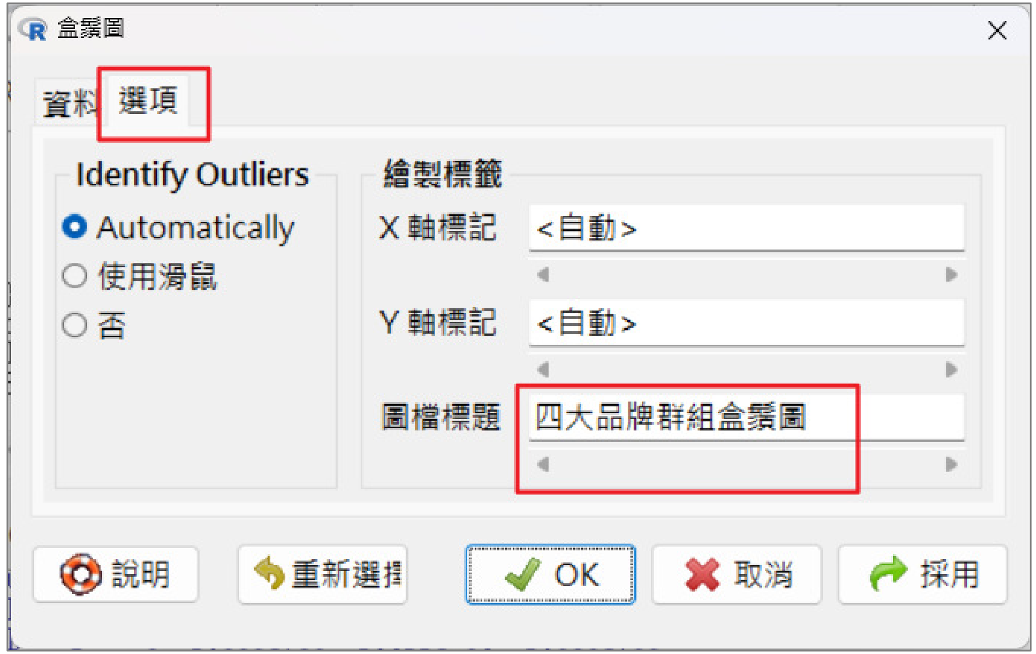
```
> df <- read.table("https://raw.githubusercontent.com/rwepa/DataDemo/master/beverage.csv", header=TRUE,
+   stringsAsFactors=TRUE, sep=",", na.strings="NA", dec=".", strip.white=TRUE)

> str(df)
'data.frame':	20 obs. of  2 variables:
 $ brand: Factor w/ 4 levels "A","B","C","D": 1 1 1 1 1 2 2 2 2 2 ...
 $ sale : num  26.5 28.7 25.2 29.3 25.3 29 27.6 25.4 28.3 29.7 ...
```

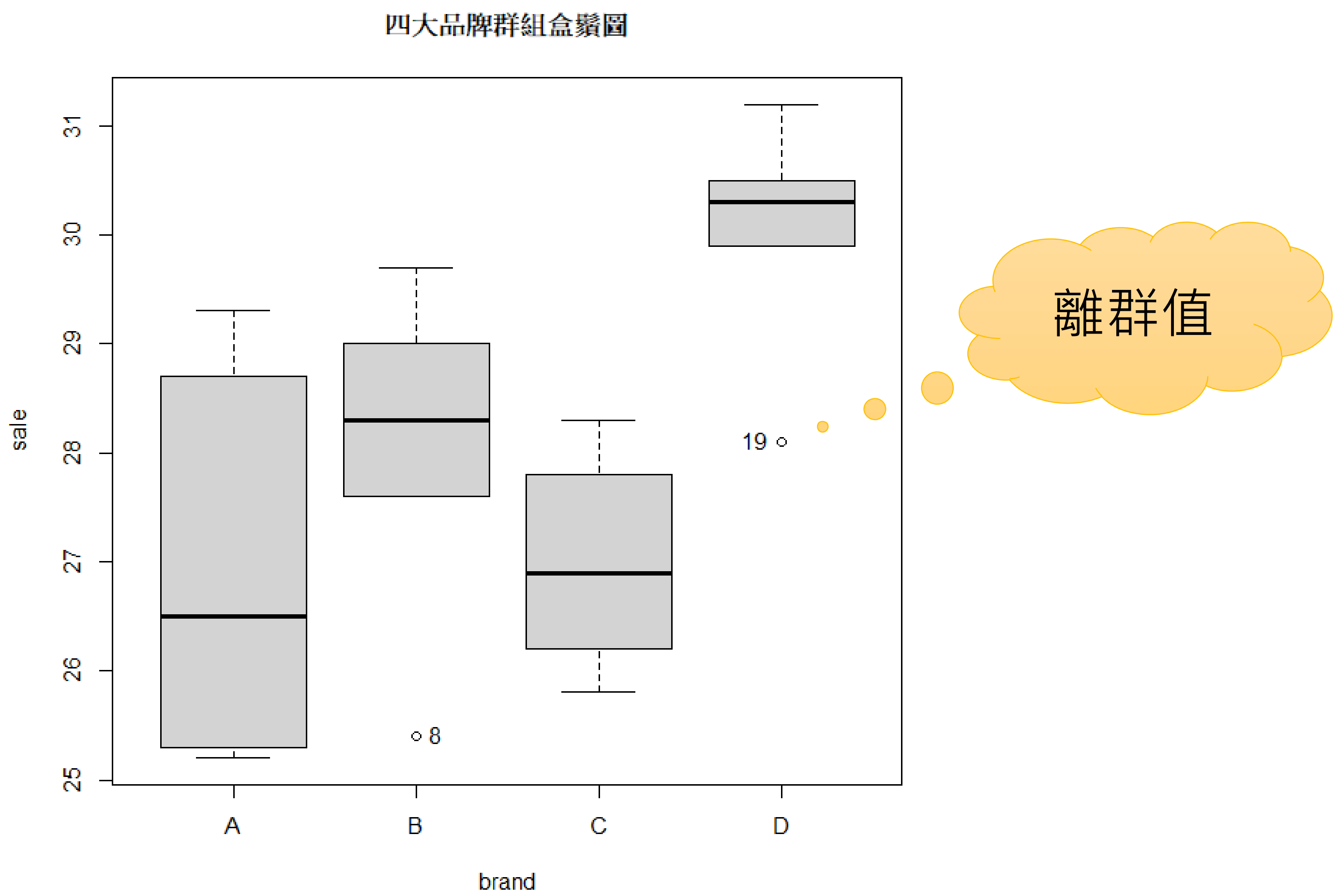
- data.frame ：資料框
- Factor w/ 4 ：因子 四種水準, 顯示為 A, B, C, D, R內部儲存為整數 1, 2, 3, 4
- num ：數值 numeric 的縮寫

繪圖 \ 盒鬚圖


盒鬚圖
資料
選項
Variable (pick one or more)
sale
繪圖依據： brand
說明
重新選擇
OK
取消
採用

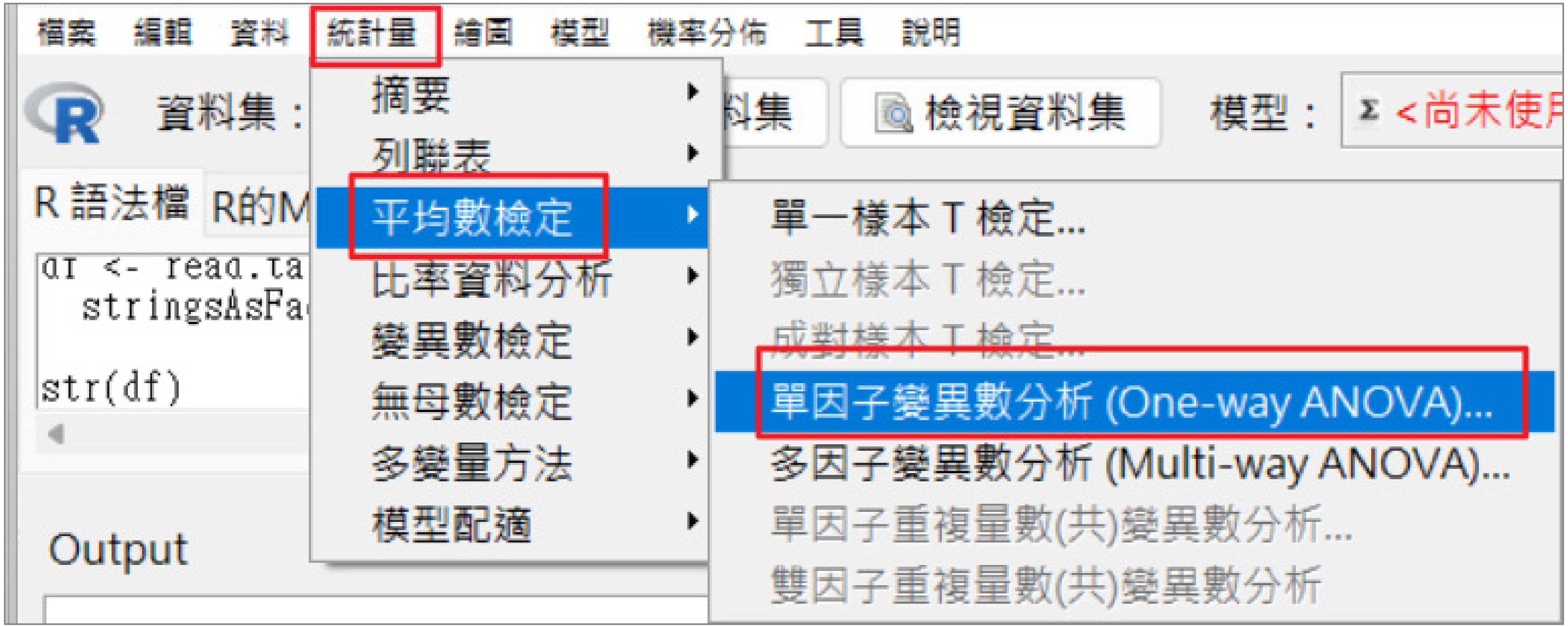

盒鬚圖
資料
選項
Identify Outliers
Automatically
使用滑鼠
否
繪製標籤
X 軸標記
<自動>
Y 軸標記
<自動>
圖檔標題
四大品牌群組盒鬚圖
說明
重新選擇
OK
取消
採用

四大品牌群組盒鬚圖

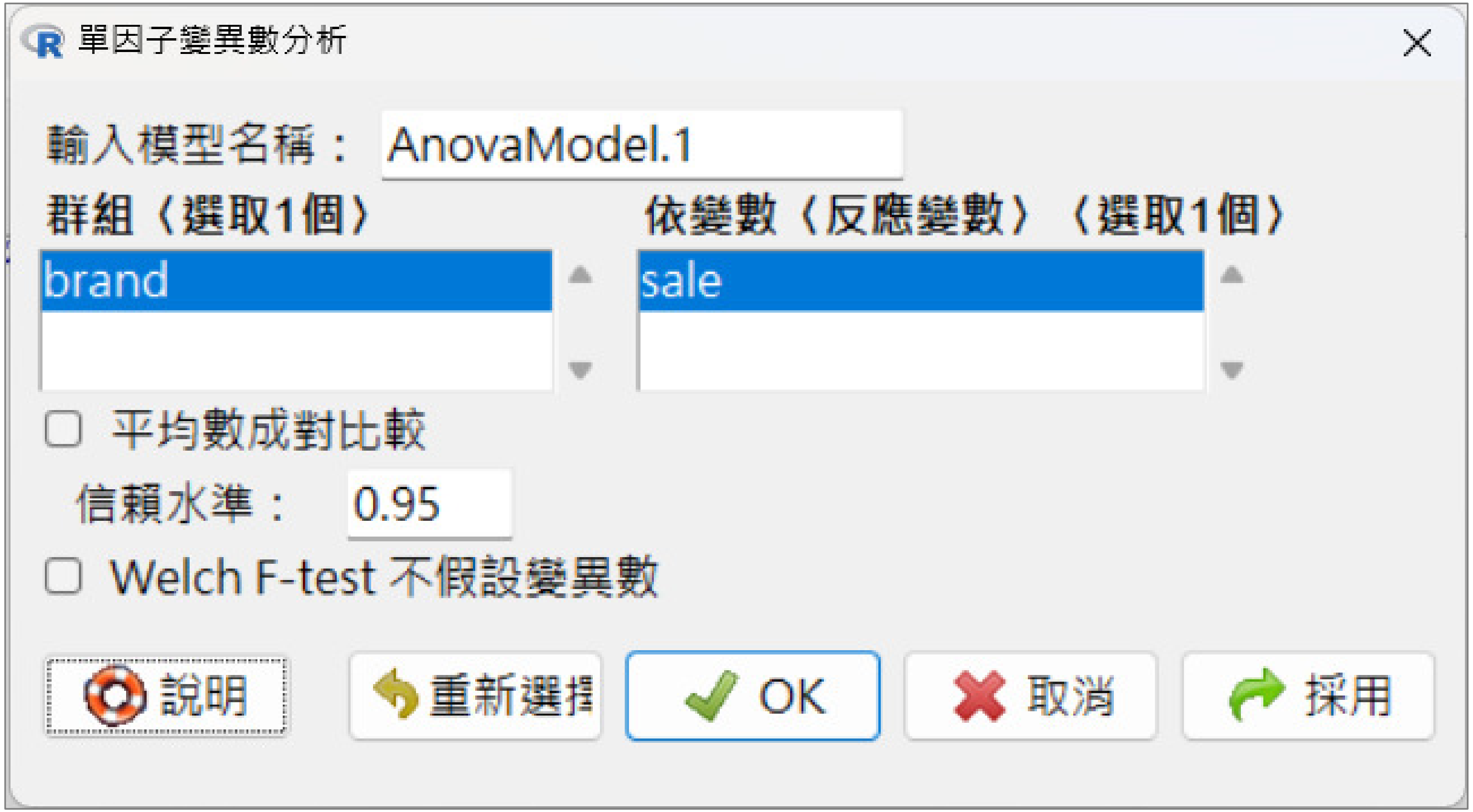


單因子變異數分析

- 統計量 \ 平均數檢定 \ 單因子變異數分析



單因子變異數分析 (續)


單因子變異數分析
輸入模型名稱：
AnovaModel.1
群組（選取1個）
brand
依變數（反應變數）（選取1個）
sale
平均數成對比較
信賴水準：
0.95
Welch F-test 不假設變異數
說明
重新選擇
OK
取消
採用

aov 函數

- $H_0: \mu_A = \mu_B = \mu_C = \mu_D$
- H_1: 四個品牌飲料平均銷售量不全相等

結果與人工
計算相同

```
> AnovaModel.1 <- aov(sale ~ brand, data = df)

> summary(AnovaModel.1)
            Df Sum Sq Mean Sq F value Pr(>F)
brand        3  30.00  10.000   4.535 0.0175 *
Residuals   16  35.28   2.205
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> with(df, numSummary(sale, groups = brand, statistics=c('mean', 'sd')))
  mean       sd data:n
A   27 1.907878      5
B   28 1.650757      5
C   27 1.051190      5
D   30 1.161895      5
```

p值 < 0.05, 拒絕H_0

謝謝您的聆聽

Q & A

李明昌

alan9956@gmail.com

http://rwepa.blogspot.tw/