

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)





大綱

- •1.F分配SPSS
- 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.SPSS變異數分析
- 6.Post Hoc檢定





F分配

- 考量 $U \setminus V$ 分別為具有自由度 u 及 v 之卡方分配的隨機變數且 $U \setminus V$ 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

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

• Gamma function: $\Gamma(n) = (n-1)!$

卡方分配: https://github.com/rwepa/market_survey_research/blob/main/es_04_mean_estimation.pdf P.49



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: $F_{1-\alpha}(v_1,v_2) = \frac{1}{F_{\alpha}(v_2,v_1)}$

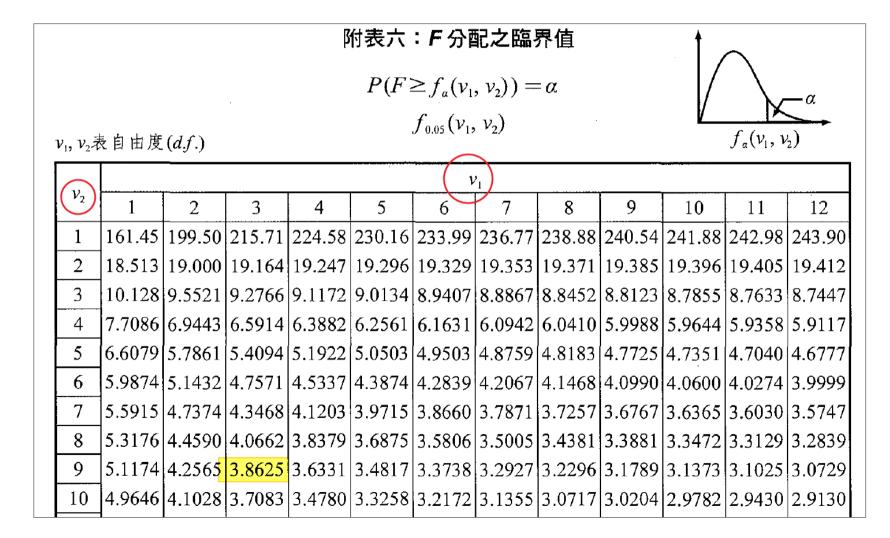


F分配-範例

- 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 · $v_1 = 3 \cdot v_2 = 9$ 所對應之位置為 $f_{0.05}(3,9) = 3.8625$ ·



F分配-查表法





F分配SPSS

• IDF.F(1-alpha,v1,v2) · $F_{1-0.05}(3,9)$ 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]





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
 - <mark>因變數</mark>, 依變數, 應變數, 被解釋變數 dependent variable,
 - 結果變數 outcome variable.





ANOVA資料架構

處理		觀測值		總和	平均值
1	y_{11}	y_{12}	y_{1n}	y_1 .	$ar{y}_{1.}$
2	y_{21}	<i>y</i> ₂₂	y_{2n}	$y_{2.}$	$ar{y}_{2.}$
•••					
a	y_{a1}	y_{a2}	\mathcal{Y}_{an}	y_a .	$ar{\mathcal{Y}}_a$.
				$\mathcal{Y}_{}$	$ar{\mathcal{Y}}_{}$

- 平均值模型(Means model)
 - $y_{ij} = \mu_i + \varepsilon_{ij}$, i = 1, 2, ..., a; j = 1, 2, ..., n °
- 效果模型(Effects model)
 - 考慮 $\mu_i = \mu + \tau_i, i = 1, 2, ..., a$ •
 - $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$, 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 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。

•
$$y_{i.} = \sum_{j=1}^{n} y_{ij}$$
, $i = 1, 2, ..., a$

•
$$\bar{y}_{i.} = \frac{y_{i.}}{n}$$

•
$$y_{..} = \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}$$

•
$$\bar{y}_{..}=\frac{y_{..}}{N}$$
, $N=an$ 總觀測值個數

 處理		觀測信	總和	 平均值	
1	y_{11}	<i>y</i> ₁₂	y_{1n}	$y_{1.}$	$\bar{y}_{1.}$
2	<i>y</i> ₂₁	y ₂₂	 y_{2n}	$y_{2.}$	$ar{y}_{2.}$
•••			•••		
a	y_{a1}	y_{a2}	y_{an}	y_a .	$\bar{y}_{a.}$
				<i>y</i>	<i>y</i>

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

$$\therefore \sum_{i=1}^{a} \tau_i = 0, \not \sqsubseteq \tau_i = \mu_i - \mu$$

- H_0 : $\tau_1 = \tau_2 = \cdots = \tau_n = 0$
- 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

• 總平方和 (Total Sum of Squares)

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

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

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

$$= n \sum_{i=1}^{a} (\bar{y}_{i.} - \bar{y}_{..})^{2} + \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^{2} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.}) = y_{i.} - n\bar{y}_{i.} = y_{i.} - n\frac{y_{i.}}{n} = 0$$

- = 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)	<i>a</i> − 1	$MS_{Treatment} = \frac{SS_{Treatment}}{a - 1}$	$F_0 = \frac{MS_{Treatment}}{MS_E}$
誤差	SS _E (組內變異)	N-a	$MS_E = \frac{SS_E}{N - a}$	
總和	SS _T (總變異)	<i>N</i> – 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 $SS_T = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} \bar{y}_{..})^2$
- 5. 誤差項自由度 (N-1)-(a-1)=N-a
- 6. 如果 $F_0 \ge F_\alpha(a-1,N-a)$,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。

品牌 地區	A	В	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



- 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 $\alpha=0.05$ 檢定此四種品牌飲料之平均銷售量是否相等。



(1)令 Hi表第i種品牌銷售量之平均數,則此問題之假設:

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

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

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

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

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

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

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



```
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
```



- 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 · SSE · 35.28

$$MSE = \frac{SSE}{n-k} = \frac{35.28}{16} = 2.205$$

•
$$f_0 = \frac{MSB}{MSE} = \frac{10}{2.205} = 4.535$$



• 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	30	3	10	$f_0 = 4.535$
隨機誤差	35.28	16	2.205	
總和	65.28	19		

- (3) 因為 $\frac{MSB}{MSE}$ ~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。





二因子變異數分析(續)

- 考量 x_{ijk} 為獨立之常態隨機變數 \cdot $i=1,2,\cdots,a$ \cdot $j=1,2,\cdots,b$ \cdot $k=1,2,\cdots,n$ \cdot 且變異數均相等 \cdot
- 母體平均數 $\mu_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$,其中 $\alpha_i \setminus \beta_j \setminus (\alpha\beta)_{ij}$ 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 : $\alpha_1 = \alpha_2 = ... = \alpha_a = 0$ 成立時,即第一因子不影響依變數之條件下, $\frac{MSA}{MSE} \sim F(a-1,ab(n-1))$
- (2)在 H'_0 : $\beta_1 = \beta_2 = ... = \beta_b = 0$ 成立時,即第二因子不影響依變數之條件下, $\frac{MSB}{MSE} \sim F(b-1,ab(n-1))$
- (3)在 H_0'' : $(\alpha\beta)_{11} = (\alpha\beta)_{12} = \dots = (\alpha\beta)_{1b} = \dots = (\alpha\beta)_{a1} = (\alpha\beta)_{a2} = \dots = (\alpha\beta)_{ab}$ 成立時,即兩因子交互作用不影響依變數之條件下, $\frac{MSAB}{MSE} \sim F((a-1)(b-1), ab(n-1))$



二因子變異數分析表

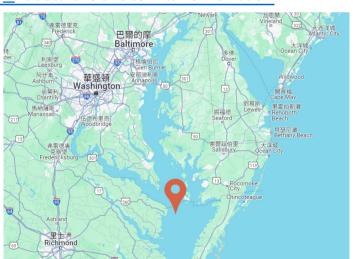
變異來源	平方和	自由度	均方	f值
A因子	SS_A	a-1	MS_A	$f_1 = \frac{MS_A}{MS_E}$
B因子	SS_B	b - 1	MA_B	$f_2 = \frac{MS_B}{MS_E}$
交互作用 (AB)	SS_{AB}	(a-1)(b-1)	MS_{AB}	$f_3 = \frac{MS_{AB}}{MS_E}$
誤差	SS_E	ab(n-1)	MS_E	_
總和	SS_T	abn-1		





河口灣鐵含量範例

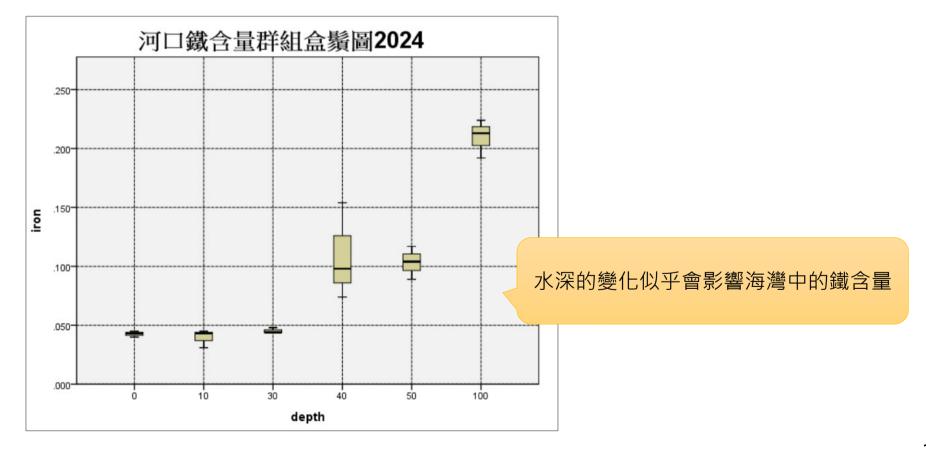
- 考量美國東岸面積最大的河口灣乞沙比克灣 (Chesapeake Bay · https://en.wikipedia.org/wiki/Chesapeake_Bay) ,研究人員想知道水深的變化是否會影響水中的鐵含量。
- 實驗人員在六個水深處進行了三次測量: $0 \times 10 \times 30 \times 40 \times 50$ 和 100 英尺。反應變數是鐵含量,以 mg/L 為單位測量。資料包含在stats4nr包中的iron:
- 參考: https://stats4nr.com/analysis-of-variance
- 資料下載: https://github.com/rwepa/market_survey_research/blob/main/iron.csv





使用「群組」技巧

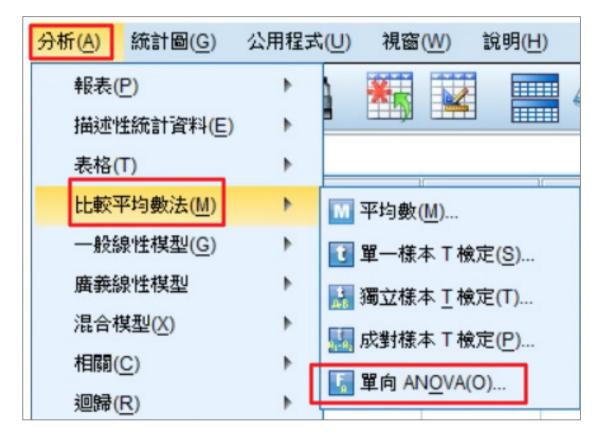
• 統計圖 \ 圖表建置器 \ 簡易盒鬚圖





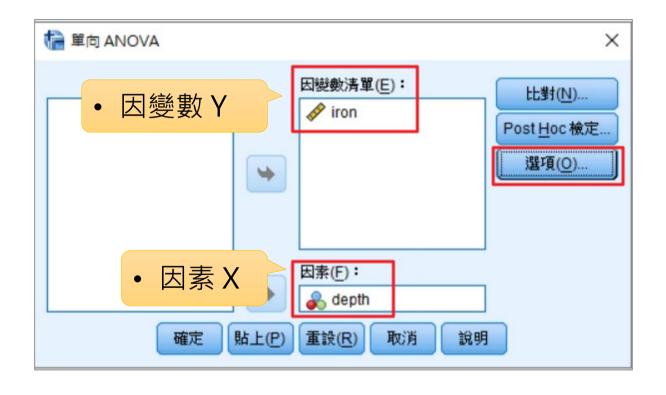
單向ANOVA

• 分析 \ 比較平均數法 \ 單向ANOVA





單向ANOVA (續)







單向ANOVA (續)

- f值顯著性0.000 < 0.05
- 因此拒絕虛無假設,即至少一個平均值與其他平均值不同的結論, 所有不同水深的鐵含量不相同。

變異數同質性測試

Levene統計資料: p值 0.014 < 0.05

H₀:6個不同深度母體變異數相等

H₁:至少有1個母體變異數不相等

結論是6組不同水深母體變異數不全相等。

	平方和	df	平均值平方	F	顯著性		
群組之間	.065	5	.013	35.107	.000		
在群組內	.004	12	.000				
總計	.069	17					

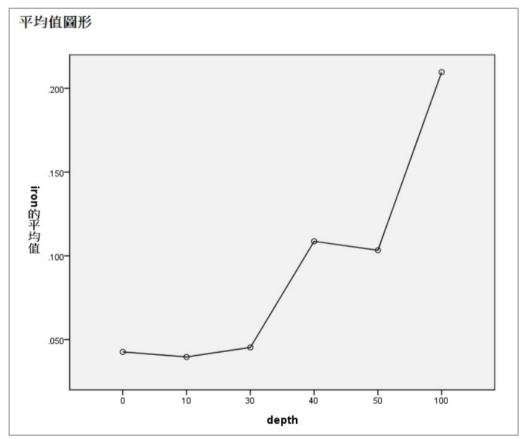
F檢定: p值 < 0.05

結論是6組不同水深母體平均數不全相等。



平均值圖形

• 選項 \ 平均數圖 打勾







Post Hoc檢定

- Post Hoc檢定:可檢定是哪幾組平均值不相等,是屬於事後檢定的方式。
- 假設相同變異數,一般 Scheffe法比較嚴格,LSD比較不嚴格,可考慮採用 Bonferroni法、Tukey法,屬於適中的嚴格程度。
- https://www.ibm.com/docs/zh-tw/spss-statistics/beta?topic=anova-one-way-post-hoc-tests



本例選取

37



因變數: iron Tamhane

lamnane						95% 信	主賴區間
(I) depth	(J) depth	平均差異 (]	-J)	標準錯誤	顯著性	下限	上限
0	10	.00300	00	.004607	1.000	05055	.05655
	30	00266	57	.001972	.986	01500	.00967
	40	06600	00	.023746	.819	46479	.33279
	50	06066	57	.008219	.203	18361	.06227
	100	16700	0*	.009499	.038	31373	02027
10	0	00300	00	.004607	1.000	05655	.05055
	30	00566	57	.004570	.997	06146	.05013
	40	06900	00	.024102	.780	42710	.28910
	50	06366	57	.009195	.082	13939	.01205
	100	17000	0*	.010355	.010	26540	07460
30	0	.00266	57	.001972	.986	00967	.01500
	10	.00566	57	.004570	.997	05013	.06146
	40	06333	33	.023739	.842	46303	.33637
	50	05800	00	.008199	.225	18307	.06707
	100	16433	3*	.009481	.041	31301	01565
40	0	.06600	00	.023746	.819	33279	.46479
	10	.06900	00	.024102	.780	28910	.42710
	30	.06333	33	.023739	.842	33637	.46303
	50	.00533	33	.025044	1.000	28110	.29177
	100	10100	00	.025493	.431	36537	.16337
50	0	.06066	57	.008219	.203	06227	.18361
	10	.06366	57	.009195	.082	01205	.13939
	30	.05800	00	.008199	.225	06707	.18307
	40	00533	33	.025044	1.000	29177	.28110
	100	10633	3*	.012392	.017	18486	02780
100	0	.16700	0*	.009499	.038	.02027	.31373
	10	.17000	0*	.010355	.010	.07460	.26540
	30	. 16433	3*	.009481	.041	.01565	.31301
	40	.10100	00	.025493	.431	16337	.36537
	50	. 10633	3*	.012392	.017	.02780	.18486

- 第3行 平均差異 (I-J) 表示I與J之比較結果。
- 有加註 * 表示達到顯著差異。
- 水深100與大部分不同水深之含鐵量有顯著 差異。
- 100 vs. 40 差異較不顯著, 二者含鐵量相等。
- 0 vs. 100 結果為 -0.167 · 表示 0 之含鐵量較小。

@RWEPA

*. 平均值差異在 0.05 層級顯著。



謝謝您的聆聽 Q&A



李明昌

alan9956@gmail.com

http://rwepa.blogspot.tw/