平均數事前比較的重要性及其統計方法

范德鑫

國立臺灣師範大學教育心理與輔導系

摘要

變異數分析(analysis of variance , ANOVA) 統計方法，目前已被廣泛的應用於各種學術領域。過去國內一些研究者在使用這種統計方法時，習慣採用綜合的F考驗，如果發現F考驗達統計的顯著(Statistical significance)) 則繼續進行事後比較 (Post hoc Comparison) ，若無顯著，則告一段落。不可否認的，這種做法對於探究性的研究極為有用，但有些研究應該使用計劃性的比較 (Planned comparison) 或稱事前的比較(Priori comparison)較具科學價值。

本文旨在敘述事前比較的意義、重要性、統計方法以及與統計方法有密切關係的錯誤率的問題，其中包括下列三種統計方法：(1)多重t考驗(Multiple t tests)，(2)正交比較(Orthogonal contrasts)，(3)杜納考驗(Dunn's tests)

壹.事前比較的意義及其重要性

近年來單變量變異數分析被廣泛地應用於心理學、社會學、經濟學、政治學、農業學、生物學、教育和其他領域。此種統計方法所以具有如此廣大的應用性，主要歸因於下列兩點：(1) 它不但具有考驗處理樣本平均數差異的功能，而且不受平均數個數之限制；(2) 它能同時處理兩個以上的自變項，除可分別呈現每個自變項的效果外，尚可說明兩個或更多自變項的交互的效果(Howell, 1985)

過去國內很多研究者曾使用過簡單的變異數分析作為進行實徵性研究的統計方法。他們大都按照下面的程序進行：首先進行綜合的F考驗 (overall F-test) ，若F考驗統計上未達到顯著，則保留虛無假設(H。) ，視樣本平均數之差異為抽樣誤差所造成。反之，F考驗達統計上的顯著，則繼續使用事後考驗(Post hoc tests) ，並找到那幾個平均數之間有差異。

這種等到搜集、檢視資料和發現綜合的 F考驗顯著之後才進行的多重比較稱為事後比較。它是一般變異數分析的重要補充統計方法，這種比較在初步分析發現資料有真正的處理效果後，對資料做進一步的探究極為有用 (Hays, 1981)，同時引導將來研究方面亦極有貢獻。然而，從科學的重要性言，事後比較卻不如事前比較(Kerlinger, 1986)。事前比較又稱為計劃性的比較。使用事前比較者常常在資料搜集到之前，根據有關的理論、前人研究的結果或自己

嚴謹的邏輯概念確定某些特定的待答問題。研究者不必進行綜合的F考驗，而是完全針對這些特定問題進行考驗。若特定的問題愈少，使用事前比較之優點更為明顯。為使讀者明瞭特定問題的形式，茲舉下面的例子說明之。

假設有位研究者想知道說服的方法對改變人們對少數民族態度之影響，於是可設計一個說服實驗計劃如下：

實驗組			控制組		
Ⅰ	Ⅱ	Ⅲ	Ⅰ	Ⅱ	Ⅲ
電影	演講	電影與演講	無任何處理	無關的電影	無關的演講

實驗組的第Ⅰ，Ⅱ個處理分別為欣賞對少數民族有利的電影，聽相同主題且對少數民族有利的演講；第Ⅲ個處理為同時接受上述兩項處理。因考慮有些受試者態度的改變，可能是由於看任何電影或聽任何演講的結果，故於控制組中分成三種處理，除了第一個為無任何處理外，還增加兩個處理，一是看與少數民族毫無關係之電影，另一則是聽與少數民族無關之演講，且演講者與實驗組中之演講者相同。又假設隨機抽取３００名受試者，並隨機分派到上述六個情境，即每個情境中 50 名受試，每一名受試者於實驗前後各做態度量表一次，並把態度分數的改變視為依變項。

在此研究中，實驗者於實驗之始，對下面幾個問題感到興趣：

1. 就整個實驗組而言，處理效果是否不同於控制組？
2. 電影與演講合併組之效果是否不同於電影組或演講組？
3. 控制組中，無關的電影與無關的演講處理組較無任何處理組是否有效果？

換言之，實驗者在資料搜集之前，僅對上述特定問題感興趣，希望分析資料以回答這些問題。雖然綜合的變異數分析與F考驗能指出任一系統效果，然而，實驗者感興趣的問題並不在此，卻在所提的特定問題的答案，特定母群平均數之差異（Hays，1981）。在這個例子中，問題或假設都是無方向性的(nondirectional)，然而，若研究者有其根據，亦可提出方向性的問題。

事前比較受研究者或統計學者的重視，約有下列兩個理由：

（1）事後比較的研究者，對於其實驗處理的效果並無明確的概念，到底那一對處理之效果有異，完全要等到搜集資料之後才能知曉。然而事前比較研究者在研究之初，就已經提出較具邏輯的科學性的問題或假設，然後搜集資料以驗證其提出的特定的問題，故事前比較較具科學的價值(Kerlinger,1986)。Glass和Hopkins(1984)亦指出，只有當研究者在選擇要考驗的假設時，無機會受到資料的影響，多重比較(multiple comparison)方法之分配理論和機率

敘述才可能正確，而事前比較正具有這個特性。

（2）事後比較僅適用於雙側考驗，而事前比較不僅可用於雙側考驗，亦可用於單側考驗。其實，事前比較的理論根據與單側考驗極其相似，欲正確的話，研究者必須在研究之初就需做決定（Glass & Hopkins ,1984）。故一般言，事前比較較受研究者喜愛，因為它具有較大的統計效力（Howell , 1982 ; Kirk, 1982)

茲舉一例說明在任何顯著水準下，事前比較具有較大統計效力之理由。假設有三個處理實驗（如啓發式教學法，編序教學法和演講法）之虛無假設(H_0 : $u_1 = u_2 = u_3$) 為真，研究者在研究之始就已經決定採用.05的顯著水準，考驗某兩個平均數的差異。若拒絕H_0，則犯第一類型錯誤之機率為.05；反之，研究者同樣採.05 的顯著水準，並一直等到考驗最大差異之後再決定結果，則犯第一類型的真正機率接近.15而非.05。理由是:只要研究者見到資料，雖然僅考驗產生最大差異的一對平均數，實際上，暗地裡，卻是考驗了三個差異($\overline{X}_1$和$\overline{X}_2$，$\overline{X}_1$和$\overline{X}_3$，$\overline{X}_2$和$\overline{X}_3$)，說明白些，虛無假設為真，然而只是其中一個差異造成了拒絕虛無假設(即犯了第一類型錯誤)，事後比較者只要考驗最大差異的一對平均數，必定就能找到。反之，事前比較者，由於未見資料之前已經指定好要考驗那一對平均數，故只有三分之一的機會能決定那對產生差異的比較。因為在其他條件相等的情況下，事後考驗將產生較大的第一類型錯誤率。雖有些研究者在使用事後比較時採用較低的顯著水準，但這種做法卻減低了統計的效力。為此，研究者感興趣的比較數若是少於所有可能的比較時，事前比較較受人喜歡，若希望作所有可能的比較，則使用事前或事後比較就不是那麼重要了。

貳. 錯誤率概念單位

在進行處理平均數多重比較之前，我們需先指定錯誤率。對兩個處理水準的實驗言，「顯著水準」一詞的意義很明確；但一個實驗包含三個或更多處理水準時，顯著水準的意義若不特別交待，則易令人混淆不清，其原因在於顯著水準或錯誤率可以用各種不同概念單位(conceptual units)來界定。其界定的方式通常有四種，即每一比較的錯誤率(error rate per comparison，α_{PC})，每個實驗的錯誤(error rate per experiment，α_{PE})，整個實驗的錯誤(error rate per experimentwise，α_{EW})，和族屬的錯誤率(error rates on family of comparison)，其中以前三種使用較為普遍(Howell, 1982)。在只有一個實驗處理的實驗裡，只有一個族屬，故每一族屬的錯誤率便等於一個實驗的錯誤率，本文偏於討論一個實驗處理的實驗，故在此僅就三種錯誤率討論之。

一．每一比較的錯誤率

一．每一比較的錯誤率

每一比較的錯誤率就是指任何一比較所產生的第一類型錯誤的機率，或說錯誤宣稱某一比較顯著之機率，具體言之，假定考驗很多很多的比較假設，再算算錯誤假設的數目，則可得：

$$\alpha_{pc}=\frac{\text{錯誤宣稱比較顯著之個數}}{\text{比較之數目}}。$$

使用t統計數，在α顯著水準下，考驗事前正交比較， 錯誤率的概念的單位就是這個個別的『比較』。雖然能夠控制每一比較的錯誤率，但整組中幾個比較犯一個或更多第一類型錯誤率將隨著考驗數目之增加而增大(Kirk ,1982) 。

二．每一實驗的錯誤率

每一實驗的錯誤率是指進行任一實驗，錯誤陳述的期望次數。假設我們要一群男女學生對50 個字作一分鐘的聯想。在α＝.05下,就每一個字,看看是否男女之聯想字數有異。假設虛無假設為真，由於我們需進行大約50 個獨立的t考驗，若顯著水準定為.05，則可預料將有2.5個(即 50 X 0.05)考驗，其顯著是由於機遇所造成。在這個例子中，每一實驗的錯誤率為 2.5，然而每一比較的錯誤率卻為 .05，故每一實驗的錯誤率與其他兩種錯誤率不同，它是一個實驗錯誤的期望數，而非機率。

三．整個實驗的錯誤率

整個實驗的錯誤率係指實驗結果包含至少一個或更多錯誤陳述的機率。換言之，把一個實驗中所得到的全部結論看成一個單位，而在某處犯第一類型錯誤之機率就是整個實驗的錯誤率。

茲舉Ryan(1959) 的例子,說明此三種錯誤率。假設我們對相同變項進行1000 個實驗，每一實驗包含 10 個平均數之比較，又假設每一比較的α為.01，在 10000 個顯著陳述中有 90 個犯第一類型錯誤,而這 90 個錯誤中，發生在70 個實驗中，則

$$\text{每一比較的錯誤率}=\frac{\text{錯誤宣稱比較顯著的個數}}{\text{比較數}}=\frac{90}{10000}=.009\fallingdotseq .01=\alpha$$

$$\text{每一實驗的錯誤率}=\frac{\text{錯誤宣稱比較顯著的個數}}{\text{實驗數}}=\frac{90}{1000}=0.09$$

$$\text{整個實驗的錯誤率} = \frac{\text{一個或更多的錯誤實驗數}}{\text{實驗總數}} = \frac{70}{1000} = 0.07$$

雖然根據每一比較，犯第一類型錯誤之機率約為.01（實際是.009），而有7 %的實驗包含至少有一個犯第一類型錯誤。這個意思是，如果有一本論文刊物都是包含十個比較的實驗，在α=. 01 下，若所有之比較，虛無假設皆為真，則實驗報告中至少有一個的結論其錯誤的機率是7%。整個實驗錯誤率之使用是依據『一個實驗中犯一個錯誤的陳述與犯多個是一樣嚴重』這個前提的。在這個例子中，每一實驗的錯誤陳述的平均數為.09。

在只有一個比較的實驗中，上述三種錯誤率是完全相同的，但是隨著比較數的增加，三種錯誤率之差距愈大。若設α表示真正錯誤率，α_{pc}表示任一比較之錯誤率，且c表示比較的數目，則：

每一比較的錯誤率(α_{pc}) $=\alpha'$

每一實驗的錯誤率(α_{PE}) $=c\alpha'$

整個實驗的錯誤率(比較相互獨立) (α_{EW})$=1-(1-\alpha')^c$

整個實驗的錯誤率(比較間非相互獨立) (α_{EW}) $\leqq 1-(1-\alpha')^c$

通常整個實驗的錯誤率不大於每個實驗的錯誤率(即$\alpha_{EW} \leqq \alpha_{PE}$) ，然而不小於每個比較的錯誤率。換言之，$\alpha_{EW}$的界限為$\alpha_{pc} \leqq \alpha_{EW} \leqq \alpha_{PE}$。

當使用F統計數在考驗綜合的虛無假設($H_0: u_1 = u_2 = \cdots = u_p$) 時，在$\alpha$顯著水準下，單一處理變異數分析中，錯誤的的概念單位是這個『實驗』。假設拒絕綜合的虛無假設，則興趣將轉移到確定那一對平均數的比較(若有的話) 達到顯著，通常學者建議，事後正交考驗之總錯誤率應等於綜合的F考驗的錯誤率，故在事後考驗中，原則上係以『整個實驗』為錯誤率的概念單位，而事前比較則以『比較』為錯誤率的概念單位(Kirk，1982)

什麼是正確的錯誤率概念單位呢？這個問題的答案應決定於比較的性質(Kirk , 1982) 。一個實驗若只有一個比較，顯著水準的解釋非常明確。反之，一個實驗包括多個比較時，情況就複雜了。事前就計劃的正交比較，目前習慣上贊成以『比較』作為錯誤率的概念單位。由於非正交比較涉及重覆的資料，某一考驗的結果與其他考驗結果有關，因此習慣上贊成採用較大的錯誤率的概念單位，即在單一處理的實驗中以『實驗』為錯誤率的概念單位。而在多個實驗處理的情況下，則以『族屬』(family) 為錯誤率的概念單位。數理統計學者已經發展出多種的考驗統計數，以控制各種不同考驗總錯誤率不大於α，例如 :(1) 控制組與p -1 個實驗組比較，(2) 任何一組c個平均數的比較，(3) 所有$\frac{p(p-1)}{2}$個兩個平均數的比較，和(4) 所有可能的平均數比較。雖然有例外情形，但前兩類考驗統計數極適合評鑑事前的非正交比較；而後兩類則較常用於事後的非正交比較。由於各種不同的考驗統計數之效

力有顯著之差異，故錯誤率概念單位並非一成不變的。實驗者所面臨的問題是如何選擇能提供保護和最大統計效力的考驗統計數。一般而言，用來考驗選擇過的，有限數目比較的考驗統計數，較考驗所有配對比較或所有可能比較的考驗統計數有效力。因此如有可能，事先指明一選擇性的，有限的比較個數對實驗者較為有利(Kirk , 1982) 。

參. 事前比較的統計方法

至此，我們已經了解事前比較是在資料搜集之前有計劃的比較。目前有許多種可作為事前比較之統計方法，茲敘述如下：

一. 多重t考驗

使用各個的 t統計數以考驗各對處理組，這是常被研究者作為事前考驗平均數差異的方法。但是這種方法，除了在特殊情形使用外，學者們並不推荐使用，主要的原因在於每個比較的誤差項不同，必須分別計算，十分費時，且由於誤差項自由度減小，降低了各種考驗的統計效力。菲色最小顯著差異法(Fisher's Least Significant Difference (L . S . D) Procedure)由於使用綜合的變異數分析的誤差項，合併了組內的變異數和自由度，故可解決上述的問題 ；因此，最小顯著差異考驗又被稱為保護t(Protected t) 。然而上述兩種方法仍共同存在一個缺點，即無差異(indiscriminate) 的使用經常造成巨大的每個實驗和整個實驗的錯誤率。

最小顯著差異考驗是一種t考驗的稍微改變，只是在樣本t公式中，共同變異數估計值(Sp^2)以綜合的變異數分析中之誤差項的變異數估計值 (MSerror) 來替代而已。這種作法不足為奇，因為誤差項均方(MSerror)是每一組內變異數的平均數。倘若實驗中只有兩組，變異數分析中之MSerror與兩個樣本之t考驗中之Sp^2是完全相同的，若是進行幾個組間的比較，我們仍以MSerror替代Sp^2 ，原因是MSerror是依據所有各組之變異數求得，而非僅擬進行比較的那兩組。使用誤差項的好處，是其自由度係使用誤差項的自由度而非$n_1 + n_2 - 2$ ，因此，若用MSerror替代Sp^2，則實得的 t值(obtained value of t，t_{obt})公式為

$$t_{obt} = \frac{\bar{X}_i - \bar{X}_j}{MSerror \sqrt{(1/n_i + 1/n_j)}}$$

茲利用下面實驗設計之資料，說明此種方法之計算過程。

實驗處理（教學方法）				
	Ⅰ	Ⅱ	Ⅲ	Ⅳ
平均數	4.75	8.60	3.33	8.25
人數	4	5	3	4

MSerror = 6.11　　d f error = 12

假設綜合的F值達到顯著，則可繼續進行各個平均數間之比較。例如，想知道第Ⅰ種教學方法與第Ⅱ種教學方法是否有別，其考驗方法如下：

$$t'_{obt}=\frac{4.75-8.60}{\sqrt{6.11(1/4+1/5)}}=-2.32$$，而臨界t值（t_{cri}）

為$t_{.025,12}=-2.18$，因為t_{obt}小於t_{cri}，所以拒絕H_0，得到的結論是：第一組與第二組之平均數有差異

二．直線比較

使用直線比較，進行兩平均數或兩組平均數的比較，平均數直線組合之形式如下：

$$L=a_1\bar{x}_1+a_2\bar{x}_2+a_3\bar{x}_3+\cdots\cdots+a_k\bar{x}_k=\Sigma a_i\bar{x}_i$$

從上面公式可知，直線組合就是實驗處理平均數之加權總和。當設定$\Sigma a_i=0$時，直線組合就變成直線比較，若能對a_i作適當的選擇，直線比較極為有用。例如，有三個處理平均數分別為$\bar{x}_1$，$\bar{x}_2$，$\bar{x}_3$．若假設$a_1=1$，$a_2=-1$，$\Sigma a_i=0$且$L=(1)\bar{X}_1+(-1)\bar{X}_2+(0)\bar{X}_3=\bar{X}_1-\bar{X}_2$。在這個例子中，L就是第一組與第二組平均數間的差異。反之，我們選定$a_1=1/2$，$a_2=1/2$，且$a_3=-1$，

$$L=(1/2)\bar{X}_1+(1/2)\bar{X}_2+(-1)(\bar{X}_3)=\frac{\bar{X}_1+\bar{X}_2}{2}-\bar{X}_3$$

在此例中，L表示前二組平均數和第三組處理的差異。

比較的平方和（S S比較）

直線比較的其中一個優點是：它們很容易被轉換成平方和，同時它們表示各組處理平均數間差異平方之總和。如果按照變異數分析之標準應用，使用各組的總和（T_i）而不用平均數，又限制$\Sigma a_i=0$，則可令$L=a_1T_1+a_2T_2+a_3T_3+\cdots+a_kT_k=\Sigma a_iT_i$，並得比較的平方和（S S比較）為$L^2/n\Sigma a_i^2$，公式中n表示每一處理分數的個數。又S S比較即為自由度是1的處理平方和（S S處理）的一部分。茲利用一簡單例子加以說明。假設有一研究，包含兩個處理且每一

處理皆為 10 個觀測值，其處理之總和(T_i) 分別為25和30。根據傳統變異數分析方法，其處理之平方和(SS處理)$=\Sigma T_i^2/n-(\Sigma x)^2/N=(25^2+30^2)/10-(55)^2/20=1.25$，如欲計算兩處理總和間直線比較，只要令$a_1=1$，$a_2=-1$即可得：

$L=\Sigma a_i T_i=(1)25+(-1)(30)=-5$。比較的平方和

$$(SS比較)=\frac{L^2}{n\Sigma a_i^2}=\frac{(-5)^2}{10[(1)^2+(-1)^2]}=1.25。$$

此值與前面之處理 平方和完全相等，由於這個例子是只有兩個平均數的比較，故比較的平方和與處理平方和完全一樣，若是三個平均數的比較，則比較平方和為處理平方和的一部分。

再舉三個處理的例子說明：

設 $n=10$，$T_1=15$，$T_2=20$，$T_3=30$ 。在綜合的變異數分析中處理的平方和為：

$$(SS處理)=\frac{15^2+20^2+30^2}{10}-\frac{65^2}{30}=11.67$$ 令$a_1=1, a_2=1, a_3=-2$

則$L=\Sigma a_i T_i=(1)(15)+(1)(20)+(-2)(30)=-25$

$$所以SS比較=\frac{L^2}{n\Sigma a_i^2}=\frac{(-25)^2}{10(6)}=10.42,$$

此比較的平方和是自由度為1 的綜合的SS處理之一部分。自由度是1,因為吾人實際上是比較兩個平均——前兩個處理之平均與第三組之平均。現在假設吾人獲得另一個直線比較，設 $a_1=1$ ，$a_2=-1$ ，$a_3=0$ ，則 $L=\Sigma a_i T_i=(1)(15)+(-1)(20)+(0)(30)=-5$，

$$SS比較=\frac{L^2}{n\Sigma a_i^2}=\frac{(-5)^2}{10(2)}=1.25。$$

這個比較的平方和也是自由度為1 的處理平方和之一部分。若將兩個比較的平方和相加正等於處理之平方和。即11.67=10.42+1.25 。由此可知,兩個比較能解釋所有的處理的平方和。下面大部分討論的的方法皆根據上面所述的直線比較。具體言之，比較的平方和常常定義成：

$$SS比較=\frac{(\Sigma a_i T_i)^2}{n\Sigma a_i^2}=\frac{L^2}{n\Sigma a_i^2}$$

三．正交比較

正交比較是常被用來獲得一組處理間比較的方法。這種方法與其他方法之最大不同在於一組正交比較中各個比較間是相互獨立的。換言之，兩處理組間比較達到顯著差異，與另一個比較是否顯著是沒有關係的，它們之間是相互獨立的。由於是正交，故一組比較的平方和之總和等於處理平方和，這些比較因此能解釋所有處理平均數間的變異情形。從計算的觀點言，正交比較異於其他類型的比較係在於比較係數選擇上的差別。

（一）正交係數

假設各處理樣本大小相同，如果要使比較的平方和等於處理平方和，則比較係數需符合下列三個標準：

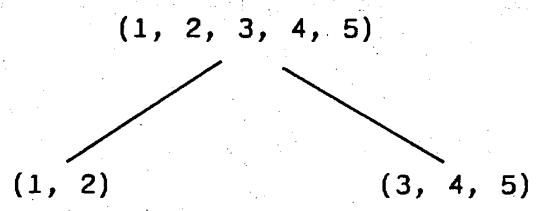
1．$\sum a_i = 0$

2．比較數等於處理之自由度個數。

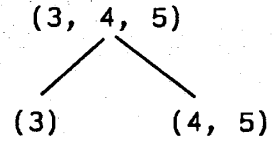
3．$\sum a_i b_i = 0$（a_i，b_i表示不同比較的相對應係數）

第一個限制是為了使比較變成平方和，第二個限制純粹是為了能找出所有處理平方和的各組成部分，最後一個限制的目的在於保障比較之間相互獨立，因而能夠將不相重疊的部分相加在一起。

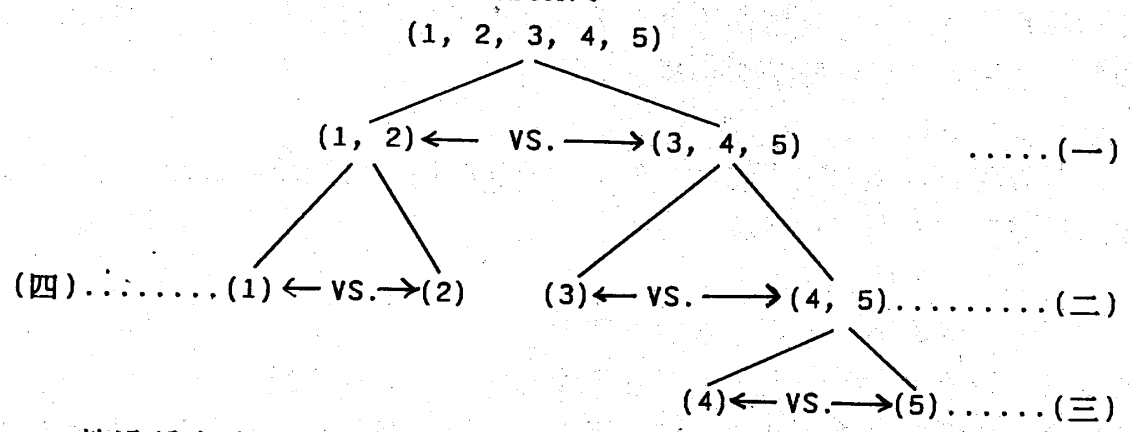
要找出符合$\sum a_i b_i = 0$要求的比較係數，乍看之下，似乎是非常困難的過程，事實上並非如此，在此介紹一種簡單的方法，雖然此法則未必能找到所有可能的組，但卻能指出大部分的比較組。若能把處理平方和分解成樹枝狀則不難找到。我們知道五個處理的綜合的F考驗是同時處理所有5個處理平均數。如果要將處理1，2的組合與處理3，4，5的組合作比較，那麼則可將此處理1，2看成是樹的一分枝，而處理3，4，5看成是另一分枝，其樹與樹枝的關係圖形如下：



選擇正交比較係數時，圖左邊處理平均數之比較係數值（a_i）等於右邊處理的個數，反之亦然，但其中一組要加上負號形成之（通常是右邊分支的加負號）。因此，五個處理的比較係數分別為3，3，-2，-2，-2。既然將一主枝分成兩個分枝，我們就不能拿一分枝中之處理與另一分枝的處理相比較，但是同一分枝中之處理可繼續相比較。因此又可形成：



像上面這種比較是可以的，在這個例子中，處理1 ，2 並未參與比較，其係數皆為 0，所以五個處理比較係數分別為0 ，0 ，2 ，-1 ，-1 。其中處理3之係數為2 ，原因是它和兩個處理比較。而處理4 ，5與另一個處理相比較故係數同為-1 。像這種程序可以一直進行，直到所有的比較出現，同時這個時後之比較數也一定等於處理的自由度數了。



按這種方法我們即可達到如上圖的比較，這些比較的比較係數如下表：

	處理				
	1	2	3	4	5
(一) a_i	3	3	-2	-2	-2
(二) b_i	0	0	2	-1	-1
(三) c_i	0	0	0	1	-1
(四) d_i	1	-1	0	0	0

欲證明這些係數是正交，只要證明所有相對應的比較係數相乘積之和等於 0 即可。例如：

$$\Sigma a_i b_i = (3)(0)+(3)(0)+(-2)(2)+(-2)(-1)+(-2)(-1) = 0$$

又，$\Sigma a_i c_i = (3)(0)+(3)(0)+(-2)(0)+(-2)(1)+(-2)(-1) = 0$

因此，我們知道(一) 和(二) 的比較，以及(一) 和(三) 的比較相互之間是獨立的。同樣的，計算方法亦能顯示所有的其他比較之間也是相互獨立的。

讀者須注意：在分枝時，如果開始是比較處理1 與處理2 ，3 ，4 ，5之組合，結果的正交比較將完全不同，因此，實驗者必須先決定那些比較較為重要，再計劃如何分枝。

茲舉一個食物、水剝奪與動物學習行為影響的實驗例子，說明如何進行正交比較。假設實驗分成控制組與實驗組，控制組又含有兩個處理，其一是可以自由吃食或喝水，另一處理是動物一天吃喝兩次，分別稱為處理1與處理2。實驗組則包括食物剝奪，水剝奪，食物和水同時剝奪三種處理，分別稱處理3，處理4，處理5。

假設在進行實驗之前，我們決定(1)比較合併的控制組(處理1，2)和合併的實驗組。(2)比較控制組的兩個處理，(3)比較合併的水剝奪，食物剝奪與兩者皆剝奪(4)水剝奪處理與食物剝奪處理。根據上述可寫出下面幾個比較：

1.(1，2)對(3，4，5)
2.(1)對(2)
3.(3，4)對(5)
4.(3)對(4)

我們決定每一比較之顯著水準訂為.05進行比較。假設實驗結果如下：

	處理			
控制組		實驗組		
1.自由吃喝	2.每天吃喝兩次	3.剝奪食物	4.剝奪水	5.剝奪食物和水
18	20	6	15	12
20	25	9	10	11
21	23	8	9	8
16	27	6	12	13
15	25	11	14	11
T_i 90	120	40	60	55

GT=365

其變異數分析摘要表如下：

變異來源	自由度	平方和	均方	F
處理	4	816	204.00	36.43
誤差	20	112	5.60	
全體	24	928		

合併實驗組與合併控制組之差異考驗如下表：

T_i	90	120	40	60	55
a_i	3	3	−2	−2	−2
a_iT_i	270	360	−80	−120	−110

$$L=\sum a_iT_i=320$$

$$SS_{1,2\ vs.3,4,5}=\frac{L^2}{n\sum a_i^2}=\frac{320^2}{5(30)}=682.67$$

考驗兩控制組差異，結果如下：

T_i	90	120	40	60	55
b_i	1	-1	0	0	0
b_iT_i	90	-120	0	0	0

$$L=\sum b_iT_i=-30$$

$$SS_{1\ VS.\ 2}=\frac{L^2}{n\sum b_i^2}=\frac{(-30)^2}{5(2)}=90.00$$

比較處理5與處理3,4合併平均數

T_i	90	120	40	60	55
C_i	0	0	1	1	-2
C_iT_i	0	0	40	60	-110

$$L=\sum C_iT_i=-10$$

$$SS_{3,4\ VS.\ 5}=\frac{L^2}{n\sum c_i^2}=\frac{(-10)^2}{5(6)}=3.33$$

最後，比較處理3與處理4

T_i	90	120	40	60	55
d_i	0	0	1	−1	0
d_iT_i	0	0	40	−60	0

$$L = \Sigma d_I T_i = -20$$

$$SS_{3\ VS.\ 4} = \frac{L^2}{n\Sigma d_i^2} = \frac{(-20)^2}{5(2)} = 40.00$$

若將上述四個比較的平方和相加，我們會發現 682.67 + 90.00 + 3.33 + 40.00＝816.00 ，它們之和與處理的平方和完全相等。換言之，我們可以把處理的平方和分成四個獨立的部分。

算出各個比較的平方和之後，則需進一步加以考驗。考驗各個比較的方法並不難，既然各個比較的平方和等於處理之平方和，那麼每一比較的考驗可以視為另一種形式的處理平方和之比較，由於每一比較的自由度是 1，故均方項與平方和項相同。再將各個比較的均方除以誤差項的均方，則可得自由度1 和 k(n-1) 的F值。此實驗的比較摘要表如下：

來源	自由度	平方和	均方	F
處理	4	816.00	204.00	36.43*
1.2 VS.3,4,5	1	682.67	682.67	121.91**
1 VS. 2	1	90.00	90.00	16.07**
3,4 VS. 5	1	3.33	3.33	< 1
3 VS. 4	1	40.00	40.00	7.14*
誤差	20	112.00	5.60	

*P<.05;**P<.01　注意文中之說明

從上表可知，除了兩單剝奪情境平均數與雙剝奪平均數間無差異外，其他三個比較皆達到統計上的顯著差異。

(二) 顯著水準

表中每一個F值皆在.05顯著水準下考驗的，因此我們有每一個比較.0 5的錯誤率。雖然這是最常用的方法，但是，仍有些學者認為採取較嚴謹的顯著水準會更好些。如果在實驗過程中，我們在$\alpha = .05$的水準下，進行五個獨立比較，對整個虛無假設而言，整個實驗的錯誤率為.23(即$1 - .95^5$)，因此當虛無

假設事實上為真時，即時每一考驗是在$\alpha = .05$ 下進行，四個中仍將有一個會達到顯著。這種錯誤水準高得令人難以接受。因此，有些學者認為可以減低每一比較的錯誤率來降低總的錯誤率。如果我們採顯著水準為.01,則整個實驗的錯誤率接近.05(即$1-.99^5$)，這種水準就較易為人接受了。

Howell(1982) 認為考驗綜合的F時，採用.05的α水準較為適當，但考驗個別比較的，則用α'(每個比較之錯誤率) 使得整個實驗的錯誤率接近.05 。就大部分的實際情況中，α'接近α/c，此處之c為比較個數。就前面這個實驗言，考驗處理的均方可使用 .05 的顯著水準，而個別比較則使用 .01 的水準($\alpha' = \alpha/c = .05/4 = .0125 \fallingdotseq .01$)為宜。這個例子中，改變顯著水準的作法是要減少實驗處理內的差異，而保留所有其他的差異，如上表 ** 符號所示。為了減少整個實驗或每個實驗錯誤率，不感興趣的考驗應該省去，如果只對五個比較中的四個感興趣，則只需考驗四個。就整個的錯誤率嚴重性而言，比較數目一般較比較是否正交更為重要(Howell，1 9 8 2) 。

(三) 非正交比較

能完全分割處理平方和(S S 處理) ，解釋所有處理間的變異，這是正交比較的優點。然而許多例子顯示，某些研究者感興趣的比較，並非彼此間成正交。有許多國內外的教育統計學課本常造成讀者錯誤印象，認為比較必須或至少須為正交的。事實上，正交性不若比較的全部數目來得重要，通常進行兩個非正交的比較，較之三個正交的比較為佳。除此之外，雖然整個實驗的錯誤率會受到正交性的影響，但是每一比較的錯誤率和每一實驗錯誤率都完全不受正交性的影響的。舉例言，若每一比較採.05 顯著水準，五個非獨立的比較，整個實驗的錯誤率最大值為.25，與獨立比較的整個實驗錯誤率. 23 相較，差異並不太大。

使用非正交比較唯一產生的實際問題是，比較平方和之總和不等於處理的平方和，但這個問題並不嚴重,正交比較在下面杜納(Dunn) 考驗中將再提及。

(四) 樣本大小不等

每一處理樣本大小相等時，我們對比較係數作兩種限制，為了使直線比較成為處理平方和之一部分a_i之總和必須為0 ，又為使兩比較成為正交，相對應係數相乘積之總和亦須限制為0 。當樣本大小不等時，上述限制須分別改成：$\Sigma n_i a_i = 0$. $\Sigma n_i a_i b_i = 0$ ，這些係數和前述一樣可直接應用於處理之全部，其

$$F = \frac{L^2}{\Sigma n_i a_i^2 (MSerror)}$$

這些限制還是適用於樹狀圖法則的，只不過要將 a_i 設定為另一邊比較的觀測值數目，而不是處理的個數。下面就是四個處理之觀測值數目分別為 9，10，8，10，的一組正交比較。

			處理			
			Ⅰ	Ⅱ	Ⅲ	Ⅳ
比較		係數 n_i	9	10	8	10
Ⅰ和Ⅱ	VS. Ⅲ和Ⅳ	a_i	18	18	-19	-19
Ⅰ	vs. Ⅴ	b_i	10	-9	0	0
Ⅲ	vs. Ⅳ	c_i	0	0	10	-8

四、杜納考驗(Dunn's test)

杜納考驗又稱作彭法羅尼t(Bonferroni t) 。若有幾個平均數，則直線組合的公式為 $L=\Sigma a_i T_i$ ，由於無限多的可能 a_i 值而有無限多個的直線組合，就是我們對直線比較加上 $\Sigma a_i=0$ 之限制，仍然有無限多個存在，就是只看有意義的比較，其可能之比較數仍然相當多。如果我們的興趣僅置於既使可能的比較數目，杜納考驗是極為適宜的，雖然研究者早已了解此種，Dunn (1961) 卻是第一個使其成為可供正式使用者。

(一) 彭法羅尼不等式(Bonferrori Inquality)

一組c個事前比較，第一類型錯誤機率以彭法羅尼不等式處理是比較正式的方式(Marascuilo & Levin，1983) ，杜納考驗正是根據彭法羅尼不等式而設計的。就本文所討論的實驗，它只是說明整個實驗誤率小於或等於 $C\alpha_{pc}$ (此處之 C 表示比較的個數) ，α_{pc} 則為進行任一比較所採用的顯著水準。因此彭法羅尼不等式純粹是說明整個實驗的錯誤率(α_{EW}) 小於或等於每一實驗的錯誤率(α_{PE})

如果每一比較的錯誤率(α') 定為 $\alpha/_c$ (此處之α是整個實驗錯誤率的最大值)則 $\alpha_{EW}\leqq\alpha$ ，即 $C\alpha'=(\alpha/c)=\alpha$。若每一比較的錯誤率不一樣，分別為 $\alpha'_1,\alpha'_2,\alpha'_3,\cdots\cdots,\alpha'_c$ 時 ，則 $\alpha'_1+\alpha'_2+\cdots\cdots+\alpha'_c=\alpha$。同時每一實驗的錯誤率定會等於α。杜納就是利用這種不等式設計考驗使每一比較在 α'下進行，使整個實驗錯誤率不大於α。其實杜納考驗是將標準的 t 考驗稍加改變而成的。假設我們首先考慮進行一對平均數的比較，令 t'_{obt} 表示應用杜納考驗的實得結果，t'_{cri} 表示 t 在 α/C 的臨界值。換言之，若整個實驗的最大錯誤率為.05,則每一比較之錯誤率(α')為 $\alpha/C=0.05/5=.01$。而 $'$ 就是在.01水準下 t '臨界值。如果下面公式算得之值大於 t'_{cri}，則顯示平均數間有差異。

$$t'_{obt}=\frac{\bar{X}_1-\bar{X}_2}{\sqrt{S^2/_n+S^2/_n}}=\frac{\bar{X}_1-\bar{X}_2}{\sqrt{2S^2/_n}}=\frac{\bar{X}_1-\bar{X}_2}{\sqrt{2MSerror/_n}}$$

上式t考驗中，由於變異數分析中之誤差均方與合併的變異數相等，故可使用誤差均方計算。如果使用總和，則上面公式需改成：

$$t'=\frac{T_i-T_j}{\sqrt{2nMSerror}}$$

此式若應用於一對總數間差異的比較，則分母為$\sqrt{n\Sigma a_i{}^2MSerror}$。若寫成通式，考驗任何平均數比較，可使用下列兩公式：

$$L=\Sigma a_i\bar{x}_i \quad ; \quad t'_{obt}=\frac{L}{\sqrt{1/_n(\Sigma a_i{}^2MSerror)}}$$

若考驗任何總數的差異時則用：

$$L=\Sigma a_i T_i \quad ; \quad t'_{obt}=\frac{L}{\sqrt{n\Sigma a_i{}^2MSerror}}$$

如果L是任何直線組合（即不必現定$\Sigma a_i=0$之直線組合），則C個比較的整個實驗錯誤率最大值是α。簡單言，杜納採用一般的t考驗，卻以改變之t'的臨界值考驗其結果，目的在於限制整個實驗的錯誤率。茲用前面水與食物剝奪對動物學習行為影響研究之資料，說明杜納考驗之進行方法。假設我們事前就決定要考驗下列的假設：

A. $\mu_1-\mu_2=0$
B. $\mu_3-\mu_4=0$
C. $\mu_5-\mu_3=0$
D. $\mu_5-\mu_4=0$
E. $\frac{\mu_3+\mu_4}{2}-\mu_5=0$（或 $\mu_3+\mu_4-2\mu_5=0$）

五個比較中前四個皆為平均數之比數，而最後一個則為平均數之平均與第五個平均數之比較。A，B，E三個比較與我們進行的正交比較完全一樣，它們

之間也是相互獨立的，而比較C 與D 不但與其他三個比較既非相互獨立，彼此間亦非正交。

已知比較數(C) 為5，自由度K(N -1) =20 ，又α=.05 ，則杜納多重比較之臨界值(t'_{cri})為2.85 (此值即$\alpha'=\alpha/C=.05/5=.01$，df=20 之 t 值)

因為在兩平均數比較之$\Sigma a_i^2=2$ ，因此

$$t'_{obt}=\frac{T_i-T_j}{\sqrt{2\,n\,MSerror}}$$

值大於2 . 8 5或T_i-T_j大於$t'_{cri}\sqrt{2nMSerror}$，則拒絕虛無假設。T_i-T_j的臨界差值$=2.85\sqrt{2(5)(5.6)}=2.88(7.483)=21.33$ ，因此任何兩平均數差絕對值大於 21.33 ，則判定為顯著。四個事前比較皆係兩平均數的比較。因此

$\mu_1-\mu_2=0$ 以 $T_1-T_2=90-120=-30$ 考驗之

$\mu_3-\mu_4=0$ 以 $T_3-T_4=40-60=-20$ 考驗之

$\mu_5-\mu_4=0$ 以 $T_5-T_3=55-40=15$ 考驗之

$\mu_5-\mu_4=0$ 以 $T_5-T_4=55-60=-5$ 考驗之

由於我們是進行雙側考驗，故差異為正為負無關，總和相差之絕對值大於21.33 的只有T_1-T_2，所以我們的結論是：$\mu_1\neq\mu_2$ 而保留其他三個假設。

考驗 $\frac{\mu_3-\mu_4}{2}-\mu_5=0$ 與考驗 $\mu_3-\mu_4-2\mu_5=0$ 是完全一樣。在

在此比較中，$L=(1)T_3+(1)T_4+(-2)T_5=60+40-2(55)=-10$ 此例中$\Sigma a^2=10^2+0^2+1^2+1^2+(-2)^2=6$，又臨界差異$=t'\sqrt{n\Sigma a_i^2 MSerror}=2.85\sqrt{5(6)(5.6)}=36.94$ 由於實得的差異(-10)小於臨界差異(36,94)故無法拒絕Ho。

(二) 杜納考驗與正交比較

表面上，杜納考驗似乎許多地方與正交法不相同。然而要深入了解多重比較法之性質必須對這些表面差異詳加研究。第一個差異是杜納考驗使用 t 統計數，而正交比較使用F 統計數。這種差異其實是微不足道的，因為當自由度為1時 t 值與$\sqrt{F}$值完全相等。換言之，如果我們取兩個獨立組並計算 t 值，同時也取相同資料計算F 值，則$t_{obt}=\sqrt{F_{obt}}$ ，此外學者也很容易從一般統計書

的附錄表看出自由度為γ，$\alpha=.05$之t'_{cri}值與自由度為1，γ，$\alpha=.05$之$\sqrt{F_{cri}}$相等。

第二個差異是在正交比較時，我們要解出F值，而在杜納考驗時則計算臨界值T_i-T_j（或L）。這個差異其實也是不值得注意的。在杜納考驗中，為了省掉每次計算t'_{obt}的時間與精力，使用另一種方式找出T_i-T_j的臨界值以作為說明L是否大於或小於此臨界值，而進一步決定拒絕或接受虛無假設。事實上，L小於臨界值T_i-T_j時，算出來的t'_{obt}也是小於t'_{cri}的。

下面值得考慮的差異是：杜納考驗不但可用於正交比較，就是非正交比較一樣適用。若為正交比較時，則整個實驗的錯誤率（α_{EW}）$=1-(1-\alpha')^c$，因此這種情形可以正確指明整個實驗的錯誤率。若是非正交比較，則無法明確指定整個實驗的錯誤率，這時後必須使用不等式$\alpha_{EW}\leq c\alpha'$。雖然這的確是杜納考驗與正交比較的差異，但是仍然不是很大差別。假設五個非正交比較在$\alpha'=.01$時，其整個實驗錯誤率為.049（即$1-.99^5$），五個非正交比較，整個實驗錯誤率的最大值為.05（即5(.01)=.05）。因此使用杜納考驗損失並不大。除此之外，前面已述及實驗者若有足夠理由進行有意義但非正交的比較，那麼他就應該如此做，這種說法正支持了杜納的方法。

杜納考驗與正交比較最後的一個差異是有關於錯誤率的問題。傳統上，非正交比較是以每一比較α的錯誤率下考驗的，而杜納考驗則設定每一實驗錯誤率為α或（$\alpha_{FW}\leq\alpha$）。因此在既定的α水準下，杜納考驗較為保守（不易拒絕H。）。然而這種差異也不是想像中的那麼重要。正交比較不用每一比較的錯誤率而用每一實驗的錯誤率，杜納考驗不用每一實驗的錯誤率而用每一比較的錯誤率都不能說是錯的。事實上，杜納考驗採用的錯誤率α'（即α/C）是每一比較的錯誤率。因此兩種考驗錯誤率的差異只是哲學觀念的差異而已(Howell，1982)。只有強調正交比較者，真正會主張要維持固定的每一比較之錯誤率，而杜納考驗的支持者才真正會主張實驗是錯誤率之基本單位。

（三）L的信賴界限(Confidence limit on L)

雖然大部分的心理學家與社會學者有從顯著考驗的角度去思考之傾向，但是大部分的統計學者喜歡從信賴界限去思考。Howell（1982）指出將來信賴界限法可能獲得更多人的支持。因此，在此特別提及如何使用信賴界限的觀念於前面所提的差異考驗，讀者若有信賴區間的觀念則對L的信賴界限極易了解。

使用信賴區間時，使用平均數比使用總和較有意義。因此在此採用

$t'_{obt}=\frac{Lm}{\sqrt{\frac{\sum a_i^2 MSerror}{n}}}$ 的公式（公式中之Lm表示平均數之直線組合），

那麼平均數間差異之信賴界限為

$CI=Lm\pm t'_{cri}\sqrt{\frac{\sum a_i^2 MSerror}{n}}$。若為兩平均數之比較則$\sum a_i^2=2$。就前述之資料言，其$CI=Lm\pm 2.85\sqrt{\frac{2(5.6)}{n}}=Lm\pm 2.85\sqrt{2.24}=Lm\pm 4.265$。而$Lm=\bar{x}_1-\bar{x}_2=18-24=-6$。所以$^5\mu_1-\mu_2$之信賴界限（$CI\mu_{1-}\mu_2$）等於$-6\pm 4.265$。因為我們採$\alpha$為.05，且將作五個平均數之比較，其$\alpha'=\alpha/c=.01$，故有百分之九十九（即$1-\alpha'=.99$）之機率，$\mu_1-\mu_2$將在-1.735與-10.2565之間，即$-10.256\leqq(\mu_1-\mu_2)\leqq-1.735$，同樣的，其他兩平均數比較之結果如下：

$-8.265\leqq(\mu_5-\mu_4)\leqq+0.265$

$-1.265\leqq(\mu_5-\mu_3)\leqq+7.265$

$-5.265\leqq(\mu_5-\mu_4)\leqq+3.265$

最後一個比較$(\frac{\mu_3+\mu_4}{2}-\mu_5)$的信賴界限之計算方法與前四個完全相同

$CI=Lm\pm t'_{cri}\sqrt{\frac{\sum a_i^2 MSerror}{n}}$

$Lm=(1/2)(\bar{X}_3)+(1/2)(\bar{X}_4)+(-1)(\bar{X}_5)=-1$

$CI=-1\pm 2.85\sqrt{\frac{1.5(5.6)}{5}}=-1\pm 3.69$

即，$-4.69\leqq\frac{\mu_3+\mu_4}{2}-\mu_5\leqq 2.69$

上下界限若同號則表示拒絕虛無假設。若異號，則需接受虛無假設。前面五個假設中按此法則，只有第一個假設考驗達統計上的顯著外，其他四個皆未達顯著，因此種方法與使用考驗統計數方法，其結果是完全相同的，因為採用的顯著水準（α'）為.01，故各個信賴區間有.99機率包括有關的參數。就整組而言，五個陳述有.95的機率（即1-(0.05)）同時包括有關的參數。

前面討論的幾種事前比較的方法，茲接照它們的錯誤率，比較，考驗統計數等表列於下：

考　　驗	錯　誤　率	比　　較	考驗統計數	雙側或單側
個別的 t 考驗	α_{PC}	兩平均數	t	雙、單側
正交　比較	α_{PC}	正交比較	F	雙　側
杜納　考驗	α_{PE}或α_{EW}	任何比較	變更 t	雙、單側

上述每個比較或考驗在統計分析上各有適用的地位。然而如果是複雜的事前比較，使用杜納考驗可能最為適當(Howell,1982)。傳統的杜納考驗係將每一比較的錯誤率視為一樣，同為α/c，而且所有比較也都採用雙側考驗。然而按照彭法羅尼不等式之定義，各個比較之錯誤率不一定需要相等，可視各個比較的重要性而訂定不同的錯誤率。為此，Dayton和Schafar(1973) 特別設計了極其有用的表格，以為研究者查找彭法羅尼 t 之臨界值。它不但可被用於對α作不等的分配情境，亦可適用於單側(有方向性的) 和雙側驗(無方向性)的統計考驗，讀者可以善加利用。

參考書目

林清山（民７５）心理與教育統計學（修正版）。台北：東華書局。

Dayton, C. M., & Schafer, W.D. (1973). Extended tables of t and Chi Square for Bonferroyi tests with unequal error allocation. Journal of the American Statistical Association, 68,78 83.

Glass, G.V. and Hopkins, K.D. (1984). Statistical methods in education and psychology. Englewood, N.J.: Prentice-Hall.

Hays, W. (1981). Statistics (3rd ed.), New York: Holt, Rinehart, and Winston.

Howell, D. C. (1982). Statistical methods for psychology. Boston: Duxbury Press.

Howell, D. C. (1985). Fundamental Statistics for the behavioral sciences. Boston: Duxbury Press.

Kerlinger, F. N. (1986). Fundations of behavior research (3rd ed.). New York: Holt, Rinehart, and Winston.

Kirk, R. E. (1982). Experimental design (2nd ed.). Moterey, CA: Brooks/Cole.

Marascuilo, L. A., & Levin, J. R. (1983). Multivariate statistics in the social sciences :A researcher's guide. Monterey, CA: Brooks/Cole.

Ryan, T.A. (1959). Multiple comparisons in Puschological research. psychological Bulltin, 56, 26-47.

The Importance and Statistical Procedures of Priori Comparisons among Means

ABSTRACT

Der-shin Fan

The analysis of variance (ANOVA) currently enjoys the status of being probably the most used (some would say abused) statistical technique in every field of research. In the past, when this procedure was used, from the very begining researchers usually ran an overall F-test. If a significant F value was obtained, what they would have shown was simply that the overall null hypothesis was false and continue to run post hoc comparisons to investigate hypothesis involving means of individual groups or sets of groups. On the other hand, If the overall F value was not significant, then the procedure was brought to an end. While post hoc comparisons are important in actual research, especially for exploring one's data and for getting leads for future research, the method of priori comparisons is perhaps more powerful and of greater scientific value.

The purposes of this article are : (1) to describe the meanning and importance of priori comparisons and related error rate and (2) to introduce the priori comparison procedures of multiple t test, orthognoal contrasts, and Dunn's test.