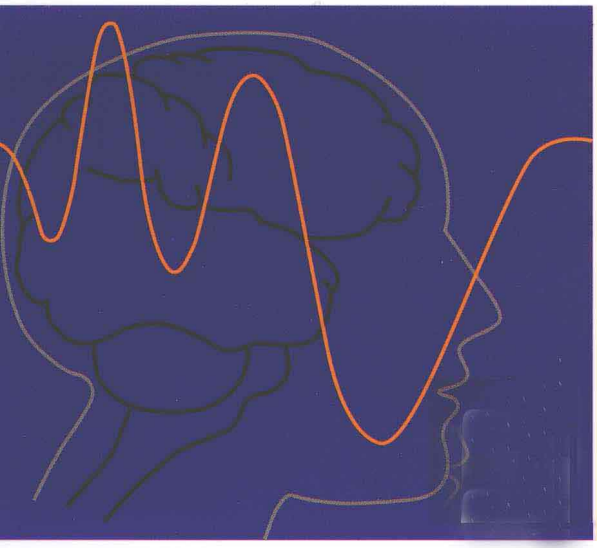
心理与教育研究方法丛书

心理统计学

第三版 下

EXPLAINING PSYCHOLOGICAL STATISTICS (THIRD EDITION)



（美）BARRY H. COHEN 著

高定国 等译 周欣悦 等审校

WILEY

华东师范大学出版社

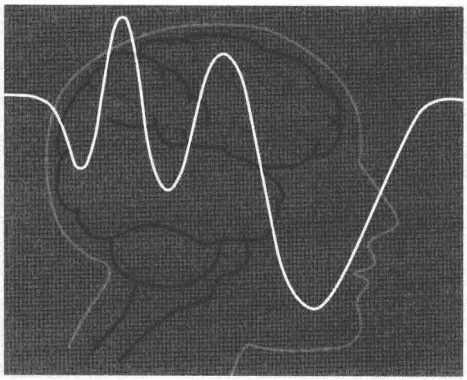
心理与教育研究方法丛书

心理统计学

第三版

（美）Barry H.Cohen 著

高定国　等译　周欣悦　等审校



华东师范大学出版社

图书在版编目(CIP)数据

心理统计学:第3版/(美)科恩著;高定国等译. —上海:华东师范大学出版社,2010.10
(心理与教育研究方法丛书)
ISBN 978-7-5617-8205-7

Ⅰ.①心… Ⅱ.①科…②高… Ⅲ.①心理统计学 Ⅳ.①B841.2

中国版本图书馆CIP数据核字(2010)第208955号

心理与教育研究方法丛书
心理统计学(第三版)

撰　　著　(美)Barry H. Cohen
译　　者　高定国等
审　　校　周欣悦等
策划编辑　彭呈军
审读编辑　李小娜
责任校对　邱红穗
装帧设计　卢晓红

出版发行　华东师范大学出版社
社　　址　上海市中山北路3663号　邮编200062
网　　址　www.ecnupress.com.cn
电　　话　021-60821666　行政传真021-62572105
客服电话　021-62865537　门市(邮购)电话021-62869887
地　　址　上海市中山北路3663号华东师范大学校内先锋路口
网　　店　http://ecnup.taobao.com/

印 刷 者　浙江省临安曙光印务有限公司
开　　本　787×960　16开
印　　张　61
字　　数　1146千字
版　　次　2011年2月第1版
印　　次　2011年2月第1次
印　　数　4100
书　　号　ISBN 978-7-5617-8205-7/B·594
定　　价　98.00元

出 版 人　朱杰人

第13章 多重比较

本章所涉及到的前几章知识：

符号：

$\overline{X}$：样本均数

k：单因素方差分析中的样组数

s_p^2：汇合方差

MS_w：组内均方

（单因素方差分析公式中的分母项）

公式：

公式 7.5B：汇合方差 t 检验

概念：

方差齐性

一类和二类错误

A部分　基本概念

在第 12 章 A 部分，我描述了一个比较三种疗法法（即安慰剂、维生素 C 及多种维生素）对疾病疗效的实验。在这个例子中，我拒绝了三个总体均数相等的零假设，并且我提到，要知道哪些均数对之间有显著差异，还需要进一步的检验。在 ANOVA 中 F 值显著既不能告诉我们是否多种维生素治疗的效果显著不同于只补充维生素 C 的疗效，也不能告诉我们是否多种维生素治疗的效果显著不同于安慰剂的疗效。很明显，我们下一步似乎应该用 t 检验来比较每一对均数，总共进行三对 t 检验。这样的安排并非不合理，但正如我即将介绍的，还存在改善余地。随着实验条件数和样组数增加，实施所有可能的 t 检验就有些问题了。对于一个实验实施多次 t 检验的缺陷，以及专门用来修正这些 t 检验的各种程序，就是本章的主要议题。

所有可能 t 检验的次数

我们通过一个例子来理解实施多个 t 检验的主要缺点。这个例子是一

个多样组实验，并且其零假设很可能为真。假设一位奇怪的研究者相信一个成年人的 IQ 分数在某种程度上取决于这个人是在一周中的哪一天出生的。为了检验这个想法，研究者测量了 7 个不同组各自的平均 IQ：一个组的人全是星期日出生的，另一个组的人全是星期一出生的，依此类推。正如上一章提到的，对于 7 组的研究，可能的 t 检验是 21 次。怎么能容易地算出这个数字呢？当我们从这 7 组当中选出任意一组作为第一组时，有 7 种可能。接下来，我们从其余 6 组里面挑选一组与其比较时，我们又有 6 种选择。所以总共有 $7 \times 6 = 42$ 对。然而这其中有一半的对子和另一半对子是相同的，只不过顺序相反。例如，对于 t 检验来说，先选出星期一再选出星期四所得的结果，和先选出星期四再选出星期一得出的结果是一样的。因此，实际上一共有 $42/2 = 21$ 对互不相同的 t 检验。求可能 t 检验次数的一般性公式如下：

$$\frac{k(k-1)}{2} \qquad \text{公式 13.1}$$

其中 k 是样组数。

以实验为单位的 α

假设上面那位研究者不知道 ANOVA，并进行了所有可能的 21 对 t 检验，每一次检验都设.05 的显著性水平。假设这 7 个组的总体均数实际上是相等的。也就是说对于这 21 对 t 检验而言，零假设都为真。因此，这些 t 检验中有任何显著性结果出现（如星期一同星期四的数据有显著差异），研究者就犯了一类错误。而对那位研究者来说，他并不知道零假设为真，于是他有可能会试着发表结果。例如，出生于星期一的人比出生于星期四的人聪明，而这样的结果显然是误导性的“虚惊”。即使这 21 对 t 检验中只有一对犯了一类错误，我们也可以说“整个实验”犯了一类错误，而这是研究者希望避免的。对于某个实验来说，犯任何一类错误的概率，就叫以实验为单位的 α（experimentwise alpha）。需要注意的是，以后我们会更多用到以族系为单位的 α（familywise alpha）这个概念。这是因为一组检验可以被更精确界定。然而，对于单因素方差检验，这种区分并不重要，所以我在本章会继续使用“以实验为单位”这个术语。当在一个多样组实验中毫无限制地进行 t 检验时，以实验为单位的 α（α_{EW}）会比每一对 t 检验所用的 α 更大（称为逐对比较的 α；testwise α）。另外，随着样组数增加，由于整个实验犯一类错误的概率会增大，因此 α_{EW} 也会随之增大。

我们可以通过一个简单的例子来获知 α_{EW} 能变到多大。假设一个研究者对一个毫无效果的两样组实验（即 $\mu_1 = \mu_2$）重复了 21 遍。那么，他的研究出现一个或多个显著性结果的概率

(即研究者犯至少一次一类错误的概率)是多少呢？如果直接计算就显得很重复和乏味——我们需要求出犯一次一类错误的概率，然后再求出恰好犯两次一类错误的概率，最后直至犯全部 21 次一类错误的概率。更简单的方法是找出不犯一类错误的概率，然后用 1 减去这个概率，就得到了至少犯一次一类错误的概率。我们从求在一次比较中不犯一类错误的概率开始。如果零假设为真，且 $\alpha = .05$，则不犯一类错误的概率是 $1 - .05 = .95$。现在我们需要接着找出在 21 次比较中都不犯一类错误的概率。如果这 21 对比较互相之间都是独立的(即每一次重复实验都使用新的随机样本)，则依照第 4 章 C 部分介绍的乘法律，总概率是各概率之积。因此，在 21 个独立检验中都不犯一类错误的概率是 $.95 \times .95 \times .95 \cdots$ 共乘 20 次，或 $.95^{21}$，结果等于 .34。因此，在 21 次检验中至少犯一次一类错误的概率是 $1 - .34 = .66$，或接近三分之二。关于此计算的一般公式是：

$$\alpha_{EW} = 1 - (1 - \alpha)^j \qquad \text{公式 13.2}$$

其中 j 是独立检验的总次数。公式 13.2 并不能很好运用于同一个多样组实验的多个 t 检验中。这是因为 t 检验并不完全是相互独立的(参见 C 部分)，但是这确实能告诉我们当实施多个 t 检验时，α_{EW} 能变得多大。

复杂比较和事前比较

在计算了 α_{EW} 后，我们明白了为什么我们需要用其他程序来控制 α_{EW}，而不是单单进行多次 t 检验。对于上面例子中那位奇怪的研究者，即使假定零假设为真，他在比较不同日出生者的 IQ 时会有多于 .05 的机会发现至少一个显著性结果。在描述能将 α_{EW} 控制在合理水平之下的各种方法之前，我需要介绍一些新术语。例如，在 ANOVA 之后进行一个两样本均数间的 t 检验就是一个比较(comparison)①。当比较仅仅涉及两样组时，我们也称其为成对比较(pairwise comparison)。这是另一种称呼两样组 t 检验的方法。复杂比较(complex comparison)涉及多于两组的样本。请看下面的例子。一名奇怪的研究者假设周末出生的人会比在工作日出生的人更聪明。当我们把周六和周日出生者的智商均数和其余五天出生者的智商均数进行比较时，这就是一个复杂比较。由于成对比较更常用，也更容易介绍，因此我在这一部分介绍这种比较。我在 B 部分会介绍复杂比较。

---

① 译者注：在本书中，作者对“comparison”(比较)和“contrast”(对照)是混用的，并没有对二者区分。

在 ANOVA 后每一对比较所采用的 α 值叫做每次比较的 α 水平(alpha per comparison),或称 α_{pc}。由于这个术语比以检验为单位的 α 水平(testwise alpha)更普遍也更易记住,因此我将在本书中使用它。正如你将看到的,校正 α_{pc} 也是控制 α_{EW} 的一种方法。

我们还需要对以下两种比较进行区分。一种是在进行一个多样组实验之前就已经计划好的检验,而另一种是在看过数据后再选择进行的检验。我们把提前计划好的检验称为事前检验(priori comparisons),而把研究者在检视了各个样本均数后决定进行的检验称为事后检验(posteriori comparisons),也更常被称为 post hoc(即在事实之后)comparison。事前检验不像事后检验那样有增大 α_{EW} 的风险。由于事前检验是相当成熟的,因此我将会在 B 部分的后半部分讨论这个问题。在本部分及 B 部分的前半部分,我讨论的都是事后检验。

Fisher 氏被保护 t 检验

在之前"一周中各日"的例子中,我描述了那个研究者在不知道 ANOVA 的情况下(没有获得一个显著性 ANOVA 结果)就不受限制地运用 t 检验,从而导致了不同的后果。在实际研究中,一个研究者应该知道运用一些措施来防止 α_{EW} 过高。最简单的程序是:除非单因素 ANOVA 中的 F 值显著,否则不允许进行多个 t 检验。如果采取这个程序,那么"一周中各日"的实验就只会有.05 的概率(设 $\alpha=.05$)获得显著性的 F 值;也就是说,在 100 个完全无效的实验当中,仅仅有 5 个会被选出来并进行之后的多重 t 检验。在这种情况下,如果一个研究者"幸运"地在一个完全无效的实验中获得了一个显著的 F 值,那么在多重 t 检验中至少获得一个显著性结果的概率也就很大了,但是要求 F 值显著,意味着对 95%($1-\alpha$)完全无效的实验将不会进行其后的 t 检验。

只仅仅在一个显著 ANOVA 后进行 t 检验的程序是由 Fisher(1951)发明的,因此也叫做 Fisher 氏被保护 t 检验(Fisher's protected t tests)。t 检验"被保护"是因为研究者必须先得到一个显著的 F 值,所以当零假设实际上为真时,t 检验经常不允许被实施。另外,正如你即将看到的,这里 t 检验的计算方式和一般 t 检验稍有不同,但却更具检验力。为了解释 Fisher 氏被保护 t 检验计算公式,我将从公式 7.5B 入手(由于这里假设两个总体均数之差为零,因此分子中 $\mu_1-\mu_2$ 项去掉了):

$$t=\frac{(\bar{X}_1-\bar{X}_2)}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \qquad \text{公式 7.5B}$$

由于我们假设方差齐性，因此运用汇合方差 s_p^2 是合理的。如果方差齐性假设在整个多样组实验中都成立（即所有总体的方差都相等），那么 MS_W（又称 MS_{error}）就是对公共方差的最好估计，因此可用它来替代公式 7.5B 中的 s_p^2。更具体来说，我会假设在“一周中各日”的实验中，在进行 ANOVA 后，相比仅仅汇合某个 t 检验中涉及的那两个样本方差，汇合所有 7 组样本的方差（即 MS_W）能对 σ^2 做更好的估计。把 s_p^2 换成 MS_W，就得到公式 13.3：

$$t = \frac{(\bar{X}_i - \bar{X}_j)}{\sqrt{MS_W\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \qquad \text{公式 13.3}$$

下标 i 和 j 表明任何两个样本均数都可以进行比较。

如果总体方差齐性假设不成立，则运用 MS_W 便不合适了。如果方差齐性假设不适合于某一对条件并且样本大小也各不相同，那么我们应该针对这一对均数进行某种形式的单独方差 t 检验（参加第 7 章 C 部分）。（由于这种情况太复杂，因此这里我仅仅处理所有样组方差齐性假设成立的情况。）如果所有参与 ANOVA 的样本大小相等，则公式 13.3 中 n_i 和 n_j 的下标就都可以去掉，得到公式 13.4：

$$t = \frac{(\bar{X}_i - \bar{X}_j)}{\sqrt{\frac{2MS_W}{n}}} \qquad \text{公式 13.4}$$

请注意，不管是哪两组在进行比较，公式 13.4 的分母都是一样的。样本量相等情况下的分母不变使得我们能够进一步简化这个程序，称为 Fisher 氏最小显著差检验（Fisher's least significant difference，简称 LSD 检验）。我将在 B 部分更详细地介绍这种方法。

运用公式 13.3（或公式 13.4），而不用公式 7.5B 来进行接下来的 t 检验有其优点：在公式 13.3 中，t 的临界值基于 df_W，而 df_W 比两组均数 t 检验时采用的自由度更大，因此 t 的临界值就更小。然而，Fisher 氏被保护 t 检验程序有一个严重的局限性。如果你还想了解各种替代方法，则你必须明白这个局限性是什么。

完全零假设和部分零假设

Fisher 氏被保护 t 检验的一个问题是，F 值显著之后起到的“保护作用”只有在完全无效的实验中才能被充分体现出来，就像“一周中各日”的实验那样。“完全无效的实验”是指 ANOVA 的零假设（即所有总体的均数都相等）实际上为真。这种涉及到所有总体均数的零

假设被称为完全零假设(complete null hypothesis)。Fisher 氏被保护 t 检验只能在完全零假设为真时才能把以实验为单位的 α 控制在.05(或其他设定值)之下。其保护作用在零假设只有部分成立时便不能成功发挥功效,而且在这种情况下,α_{EW}很容易就变得不合理的大。我举一个极端的例子来解释 Fisher 方法的局限性。

部分零假设(partial null hypothesis)

假设一个心理学家相信,如果不考虑恐惧症的类型,则所有恐惧症都可通过一些生理学指标鉴定出来,但是一些恐惧症可能会比另一些恐惧症在识别指标上更强。他测量了六种恐惧症(即社交恐惧症、动物恐惧症、广场恐惧症、幽闭恐惧症、恐高症和恐刀症)的相关生理指标,并且同时也测量了一个没有恐惧症的控制组。在这个研究中,我们可以综合这所有六种恐惧症与控制组进行比较(即一个复杂比较),也可以比较不同类型恐惧症之间的差异。依据所选择的变量,这七组可以有许多比较方式,不过我会选择一种简单(且极端)的比较形式来说明我的观点。假设对于所选择的生理指标,恐惧症组的总体均数和控制组均数不同,但是这六类恐惧症的总体均数之间却是完全相同的(即 $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 \neq \mu_7$)。在这种情况下,完全零假设不成立,但部分零假设却是成立的。接下来,假设此研究者很负责任地先进行了单因素方差分析。如果控制组均数只是略微不同于恐惧症组均数,那么研究者获得一个显著 ANOVA 结果的概率也只会略微高于 α。但是,即使各恐惧症组之间分数相等,控制组分数也很有可能和恐惧症组的会相差很多,那么 ANOVA 的结果就很容易显著了。

如果该心理学家发现 F 值在.05 的水平上达到了统计显著性,并且他坚持运用 Fisher 氏被保护 t 检验的策略,那么他就很可能不受约束地对每一对在.05 水平上进行所有成对比较。这一策略包括在六组恐惧组间进行配对检验(运用公式 13.1)。这样的比较共有 $6 \times (6-1)/2 = 30/2 = 15$ 组,而在这些组中零假设实际上为真。如果这 15 组检验都是相互独立的,那么 α_{EW}(运用公式 13.2)则变为 $1-(1-.05)^{15} = 1-.95^{15} = 1-.46 = .54$。虽然这些 t 检验并非完全独立,但我们应该明白的是,一旦决定进行所有 t 检验,犯一类错误的概率就会相当高。(剩下的 6 对检验是把控制组与每一个恐惧组进行比较,因此在这个例子中不会犯任何一类错误。)请注意,如果没有控制组,则完全零假设成立,这时获得显著性 ANOVA 结果的概率就仅仅为 α 了。在这里,运用 Fisher 氏被保护 t 检验是没有缺陷的。不幸的是,当我们加上一个控制组时,获得显著性 ANOVA 结果变得相对容易。这就失去了 Fisher 方法中的保护功能,而使得 α_{EW}升高,变得高于事先设定的用于整个 ANOVA 的 α 值。

三样组的情况

当实验中仅有三组时，即使完全零假设不为真，Fisher 方法依然能提供足够的保护。在三样组中只可能发生一种部分零假设为真的情况，即有两组均数相等，但它们和第三组均数不等。在 ANOVA 结果显著之后继续进行的 t 检验中，最多只会犯一次一类错误（即在对两个均数相等组检验时得到显著结果），因此，也就不会导致 α_{EW} 升高。然而，Fisher 方法在组数大于 3 且完全零假设不为真时依然会导致 α_{EW} 升高，而且组数越多，升高越多。正是由于这个原因，Fisher 氏被保护 t 检验口碑不好，使得研究者即使在只有三组时（这时用 Fisher 氏被保护 t 检验是恰当的）也不愿意运用此程序。这种做法令人遗憾，因为实际上 Fisher 方法在三样组情况下在所有事后检验中拥有最大检验力。当我们分析了其他的方法后，这一点会更加清晰。

Tukey 氏 HSD 检验

为了不论组数多少，也不论零假设是完全还是部分为真，都可以使 α_{EW} 维持在预定水平，即得到完全保护，Tukey 设计了一个在多样组实验中检验每一对可能均数差的替代性程序。与 Fisher 氏最小显著差检验形成对照，这种方法称为 Tukey 氏真实显著差程序（Tukey's honestly significant difference procedure）。（我在下一节将讨论 HSD 和 LSD 中“差异”的含义。）这暗指 Fisher 方法中包含着某种“欺骗”，因为它只有在完全零假设为真的实验条件下才能提供保护。那么，当对所有可能的均数都进行比较时，我们需要怎样的保护呢？假设我们进行了一个多样组实验，并且希望找出均数不同的样本。如果你需要找出至少一对均数有显著差异，那么最好的做法是先把最大的样本均数和最小的样本均数进行比较。我们可以这样理解：当完全零假设成立时，按照至少犯一次一类错误的概率来算，把最大的均数和最小的均数比较就相当于比较了所有的两两配对情况。换句话说，当差异最大的两组均数在统计上都不显著时，那么其他均数之间的差异也不会显著了。如果有一个程序可以在比较最小和最大均数时提供保护，使我们不犯一类错误，那么这个程序也可以在比较所有两两配对时防止我们犯一类错误。这就是 Tukey 氏程序的内在逻辑。Tukey 氏检验所基于的统计分布被称为 student 化全距统计（studentized range statistic）。[①]

---

① 译者注：也可以翻译为“学生化全距统计”。

Student 化全距统计

每当我们从两个均数相同的总体中抽取两个样本，并且求这两个样本间的均数差时（之后还会从样本数据估计 σ^2），就会涉及到 t 分布。当你在相同条件下抽取 3 个或更多样本并比较最大均数和最小均数之间的差异时，此差异就往往会大于在抽取 2 个样本时所得到的差异。实际上，每次抽取的样本越多，最大均数和最小均数之间的差异往往就越大。由于所有的总体均数都相等，因此这些差异仅仅是基于偶然性的。为了保护我们不受这些差异所愚弄，我们需要一个能够解释因样本越多导致越大均数差异的临界值。幸运的是，Stduent 化全距统计的分布就能让我们求出所需的临界值，并且可以根据一个多样组实验中样本的个数来进行校正。像一般 t 值一样，此统计量也是“student 化”的（有时也称 Student 氏 t）。因此，依赖于分母中的样本方差去估计通常未知的总体方差。然而，这些临界值所要满足的前提假设是所有的样本大小相等。因此，Tukey 氏程序的公式（公式 13.5）和当样本量相等时的 Fisher 氏公式（公式 13.4）非常相似，如下：

$$q = \frac{(\bar{X}_i - \bar{X}_j)}{\sqrt{\frac{MS_W}{n}}} \qquad \text{公式 13.5}$$

其中 q 代表 student 化全距统计量的临界值。附录 A 的表 A.11 中列出了临界值。我将在 B 部分介绍此表的使用方法。可能你已注意到，公式 13.4 分母中的 2 在这个公式中没有出现。这并不代表这两个检验在结构上存在真实差异。为了简化计算，Tukey 决定把因子 2（因为 2 出现在根号中，所以实际上是$\sqrt{2}$）包括在表 A.11 中；因此他在原本的 student 化全距统计量临界值的基础上都乘以$\sqrt{2}$，这就构成了表 A.11。

Tukey 氏检验的优点和缺点

在 Tukey 氏 HSD 检验中，用于决定临界值 q 的 α 即为以实验为单位的 α。不论一共进行了多少次检验，也不论部分零假设是否为真，α_{EW} 依然能被控制在之前设置的水平上。如果你要维持 α_{EW} 在 .05 的水平上，那么正如我们一般所做的那样，α_{pc}（以检验为单位的 α 水平）就必须小于 .05 的水平，这样才可以使得所有检验对的 α_{EW} 不会超过 .05。组数越多，就有越多可

能进行比较的均数对，α_{pc}就必须减小。然而，使用 Tukey 氏 HSD 检验并不需要决定适当的 α_{pc}值，因为其临界值可以直接在表 A.11 中查到（当然，一个更大的临界值对应的 α_{pc}值越小）。你可能已经看出，表 A.11 中的临界值随着研究中组数增加而增加。我在 B 部分介绍 Tukey 氏检验的使用时，会演示如何使用表 A.11。

在三样组研究中，Tukey 氏 HSD 检验的一个缺点是，和 LSD 检验相比，它会导致检验力减少，因而引起二类错误增加。在处理多于三组的情形时，LSD 检验对于给定 α 值而言依然比 HSD 拥有更大的检验力。但是这时的比较并不公平，因为 LSD 检验此时（$k > 3$ 时）所拥有的额外检验力是通过允许 α_{EW}超过最初设定的值而得到的。到目前为止，你应该很熟悉这样一个事实：即总是能通过增加 α 值来增加检验力。但是，这样的方法是允许增加犯一类错误的概率，从而减少犯二类错误的概率。对于大部分研究者来说，通过 α_{EW}的增加来提高多组实验的检验力是不可接受的。因此，当实验组数多于三组时，研究者一般偏向于使用 Tukey 氏 HSD 检验，而不是 LSD 检验。

统计学家常常认为，由于 Tukey 氏 HSD 检验能更好地控制犯一类错误的概率，因此它比 Fisher 氏 LSD 检验更为保守。如果其他条件不变，则越保守的统计程序检验力就越小。在减小犯一类错误的同时，增大了犯二类错误的概率（β）。这就反过来减小了检验力（$1-\beta$）。对于那些因允许 α_{EW}增大而获得更大检验力的方法，我们称其为自由（liberal）方法。Fisher 氏被保护 t 检验是在所有进行事后比较的方法中（除了不被保护 t 检验外）最自由的，因此也有更大检验力。Tukey 氏 HSD 检验是事后比较方法中最为保守的方法之一。

Tukey 方法不要求总体 ANOVA 结果显著。这点和 Fisher 氏 LSD 检验不同。在总体 ANOVA 不显著时运用 Tukey 氏 HSD 检验仍然有可能（虽然可能性不大）找到一对显著不同的均数。要求在运用 Tukey 氏 HSD 检验前获得一个显著的 ANOVA，会导致检验力略微降低。由于 Tukey 氏程序已经被认为是足够保守的了，因此这种检验就没有必要了。另一方面，在获得一个显著的 ANOVA 后，运用 Tukey 氏 HSD 检验（或者甚至 Fisher 氏 LSD 检验）仍然可能（虽然可能性不大）找不到显著的均数差异。一个显著的 ANOVA 唯一能保证的是，在均数之间，确实存在显著差异，但是这些显著的结果可能是需要复杂比较才能发现。我会在 B 部分介绍复杂比较。

最后，Tukey 氏 HSD 检验的一个小缺陷是其准确性取决于所有的样本数是否相等。在真实研究中，样本数有些许偏差的情况经常发生。这时我们只需要通过计算所有样本数的调和均数就可以解决（见公式 12.24）。然而，如果样本数之间有较大差异，则应该采用其他的事后比较方法。

其他事后成对比较方法

大部分其他用于成对比较的事后比较方法都处于在从 Tukey 氏 HSD 检验到 Fisher 氏被保护 t 检验这个由保守到自由的连续体之间。这些程序的差异主要是在一类错误和检验力权衡时采取了不同尺度。不幸的是，一些更具检验力的检验通过放松对一类错误的控制而获得的额外检验力。我下面列出的一些其他成对比较方法只是所有方法的一小部分，但是这些都是最为人知的一些方法。

Newman-Keuls 检验

近年来，在单因素方差分析后进行成对比较时，Tukey 氏 HSD 检验的主要竞争对手是 Newman-Keuls 检验(N－K 检验)。它也因其临界值基于 student 化全距统计量而被称为 Student-Newman-Keuls 检验(SNK)。此方法最大的优势是它通常比 Tukey 氏 HSD 检验有更大的检验力，而又比 Fisher 氏被保护 t 检验更加保守。因此，N－K 检验被广泛认为是 LSD 和 HSD 之间一个不错的折中办法，此法能对犯一类错误的概率进行足够的控制。以前，N－K 检验的一个主要缺陷是其应用过程比较复杂。N－K 检验并不对每一个成对比较运用同一个临界值。取而代之的是，它需要对所有均数进行排序，使用均数之间的距离大小(range)(排序后，两个相邻均数的距离大小是 2；如果两个均数之间还有另一个均数，则距离大小是 3，依此类推)而不是用总体 ANOVA 中的组数去表 A.11 查找临界值。现在大部分的电脑程序包都提供了支持 N－K 检验的选项，所以计算的复杂性就不那么要紧了。然而，N－K 检验有个更严重的缺陷。同 Tukey 氏 HSD 不同，N－K 检验不能将 α_{EW} 控制在用于决定 student 化全距统计量临界值的那个水平上。正因为如此，统计学家并不推荐此方法。现在，大家普遍认同，N－K 检验之所以能比 Tukey 氏 HSD 拥有更大的检验力，主要是因为前者允许 α_{EW} 不合理地扩大(当组数扩大时，情况会更糟)，所以 N－K 检验在心理学研究中的运用呈下降趋势。

Dunnett 氏检验

我们来回忆一下在我描述部分零假设时所用到的涉及 6 个不同恐惧症组和 1 个非恐惧症控制组的那个例子。如果研究者不想比较 6 个恐惧症组之间的差异，而仅仅希望把每一恐惧症组和控制组进行比较，那么最好的方法是用 Dunnett(1964)设计的成对比较方法。然而，

Dunnett 氏检验要求运用特殊的临界值表，并且其适用条件也相当特殊，在这里我就不介绍了。尽管如此，我们需要知道，Dunnett 氏检验在一些统计软件包中都有提供。并且，如果满足其适用条件，它是能控制 α_{EW} 不增大的检验中具有最大检验力的方法。

REGWQ 检验

REGWQ 检验似乎能达到一些研究者认为 N－K 检验能实现（但实际上 N－K 检验却没能实现）的目标——即修正 Tukey 氏检验以使其控制 α_{EW} 不超过已设定 α 值（通常.05）的条件下获得更大的检验力。同 N－K 检验类似，REGWQ 检验也是基于 q 分布（即 student 化全距统计量）。同时，REGWQ 检验根据对每一对均数排序后之间分隔多少级距离而采取特定的临界值。REGWQ 检验通过校正每一次比较中的 α 来决定 q 的临界值。然而，这种方法所造成的不同 α 值使之无法在传统的临界值表上表示出来。没有临界值表也是这种方法直到最近才开始使用的原因。但是，现在大部分统计软件包（如 SPSS 和 SAS）都提供 REGWQ 检验作为可选项目。因此，此方法在以后可能被更多地使用。（REGWQ 检验得名于发展此检验的 4 位研究者：Ryan、Einot、Garbriel 和 Welsch；Q 意味着此检验基于 student 化全距统计量。）

修正 LSD(Fisher-Hayter)检验

Tukey 氏 HSD 很容易使用和理解，但是它太过于保守。在比较各种多重比较的方法后，Seaman、Levin 和 Serlin(1991)通过计算机模拟发现在某些特殊的数据分析条件下，Tukey 方法将以实验为单位的 α 保持在.02 和.03 之间，而你本来想的是将总体 α 设定在.05 的水平上。于是，为了既保持 HSD 的检验力又不让 α_{EW} 上升到.05 以上，Hayter(1986)设计了一个 LSD 和 HSD 的混合检验。Hayter 的新检验包括一个两步程序，而这些程序均源于 Fisher，因此该检验被称为 Fisher-Hayter(F－H)检验或修正 LSD 检验。修正 LSD(modLSD)检验的第一步就是评估单因素 ANOVA 的显著性。只有当 ANOVA 显著时，你才能进入到下一步。下一步是 HSD 的计算，但有一个很重要修正：在表 A.11 里通过设定组数为 $k-1$ 而不是 k 来查找 HSD(即 q)的估计值，从而导致了一个更小的 q 值。因此，一个必须要超越此值才能达到显著性的均数差异也更小了，当然检验力也更大了。

Seaman、Levin 和 Serlin(1991)发现，modLSD(或 F－H)检验跟 REGWQ 检验的检验力差不多，而且总能保持 α_{EW} 在可接受的保守水平。另外，modLSD 有易于解释的优势，且无需精湛的统计软件，只要一个普通的计算器和一个 q 值表就够了。请注意，modLSD 检验在只

有三组时就变为一般的 LSD 检验了。但是，即使组数少到只有四组时，它还是要比 HSD 有更大的检验力。我将在 B 部分演示 modLSD 检验。

事前比较的优势

在数据收集之前就计划好特定的比较，相比事后比较所有可能的均数对或在看到数据后有选择地进行比较(这在本质上和比较所有可能的均数对是一样的)具有优势。这个优势有点类似于有计划地进行单侧检验而不是双侧检验，即你可以运用一个更小的临界值。正如单侧检验那样，事前比较更适用于验证性研究而不是探索性研究。另外，事前比较的效度取决于研究者的"承诺"，即研究者确实是事前已计划好了所要进行的那些比较，而不是其他比较。研究者都知道，在看过数据之后再假装是按照事前计划而进行比较也是很容易的。因此，事前比较总是要有充分的理论证明或先前研究的支持，并且应该被认为是有根据的假设，发表在实证性期刊论文的结果之前。事前比较一般使用 Bonferroni 检验。我接下来就介绍这种检验。

Bonferroni *t* 或 Dunn 氏检验

公式 13.2 表明了当进行多次独立成对比较时 α_{EW}会变得多大。我说过，如果这些检验不是完全相互独立的话(如当我们检验所有可能的成对比较时)，那么用这个公式计算的 α_{EW}是不准确的。然而，α_{EW}有一个上限，它从来不会被超过。根据意大利数学家 Carlo Emilio Bonferroni 的研究，我们可以说，对于给定次数的比较(比较次数用 j 来表示)，α_{EW}决不会超过每次比较的 α 值的 j 倍。下面是 Bonferroni 不等式的一种形式，表示为：

$$\alpha_{EW} \leqslant j\alpha_{pc}$$

Bonferroni 不等式提供了一种非常简单的方法来校正每次比较的 α 值。其中的逻辑是，如果每次比较的显著水平是 α/j，那么 α_{EW}就不会超过 $j(\alpha/j) = \alpha$。因此，α 可以根据 α_{EW}的要求来设定(通常是.05)，然后除以 j 得出每次比较的显著性水平。如公式 13.6 所示：

$$\alpha_{pc} = \frac{\alpha_{EW}}{j} \qquad \text{公式 13.6}$$

例如，如果要求 $\alpha_{EW} = .05$，而且打算进行 5 次比较，那么 α_{pc}就等于 $.05/5 = .01$。如果进

行 5 次比较，每次显著性水平都为.01 的话，那么这 5 次比较至少犯一次一类错误的概率不超过 $5\times.01=.05$。成对比较时，我们可以用一般 t 检验（即被保护 t 检验的公式），而临界 t 值则由公式 13.6 所算出的 α_{pc} 决定。因此，这个检验通常称为 Bonferroni 氏 t 检验。过去，使用 Bonferroni 检验最大的困难是，每次比较所用到的 α_{pc} 通常在 t 检验表中不容易查找到。例如，如果 $\alpha_{EW}=.05$，并且进行 4 次比较，那么每次检验的 α 为 $.05/4=.0125$。怎样在表中查找对应 $\alpha=.0125$ 的临界 t 值呢？确实有公式可以求出近似的临界 t 值，但 Dunn(1961)给出了相关表格以方便查找。这就是为什么这个检验又叫 Dunn 检验，或者 Bonferroni-Dunn 检验。当然，计算机使得检验过程更简单。大多数统计软件在计算单因素方差分析时都会给用户选择使用 Bonferroni 检验。

Bonferroni 检验的严重缺点是，它太保守了，通常使得 α_{EW} 低于原来设定的水平。回想一下我们就会发现，这个检验是基于不等式的，预先设定的 α_{EW} 只是一个上限；尤其是要进行多次检验时，即使最差的情况，也很难达到上限。这就是 Bonferroni 检验过于保守的原因，而且在进行所有成对比较时不推荐使用。例如，一个 5 样组实验，你事先计划进行所有 10 次成对比较。Bonferroni 检验就会把 α_{pc} 设为 $.05/10=.005$，但 Tukey 检验会把 α_{pc} 设为大约.0063（一个更大的 α_{pc} 意味着更大的检验力）。然而，当你考虑去掉至少 3 次比较，Bonferroni 氏检验就会比 Tukey 氏检验更有检验力（$.05/7=.007\,14$）。然而，由于要排除某些成对比较不作考虑，只有在看到数据之前就已经计划好才是合理的，因此某些事前计划是必要的。鉴于此，Bonferroni 检验最适合用于事前比较，而只有在没有其他更有效检验方法时，我们才用它进行事后比较。这样一来，在对事前比较进行详细论述之前，我暂时把 Bonferroni 检验放一下。

小　结

1. 如果实验中共有 k 组，则所有可能进行的 t 检验共有 $k(k-1)/2$ 对。

2. 如果进行所有可能的 t 检验，则至少犯一次一类错误的概率（即以实验为单位的 α，α_{EW}）将会大于每一对 t 检验所用的 α 值（即每次比较的 α 水平，α_{pc}）。α_{EW} 的大小取决于 α_{pc}、检验次数以及各检验之间相互独立的程度。

3. 如果比较所有可能的均数对，或者如果 t 检验的均数对是在看过实验数据后确定的，那么 α_{EW} 就很容易变高到不可接受的程度。我们需要运用事后比较程序对 α_{EW} 进行控制。如果你能在看到数据之前计划比较特定的均数对，则你可以运用事前比较程序。

4. 最简单的事后比较程序是仅仅在单因素方差分析结果显著的情况下，再进行 t 检验。

这种方法一般被称为 Fisher 氏被保护 t 检验。在其中的 t 检验公式中，用 MS_w 来替代样本的标准误。当所有样组的样本量相等时，我们能通过计算 Fisher 氏最小显著差检验(LSD)来简化程序。

5. 如果在一个多样组实验中所有总体均数实际上是相等的，则完全零假设为真。Fisher 方法仅在这种情况下能提供足够的保护，以控制一类错误。当部分零假设(即一些而非全部的总体均数相等)为真时，保护便不周全了。在这种情况下且样组数大于 3 时，Fisher 方法容易让 α_{EW} 变高到不可接受的程度。

6. Tukey 设计了真实显著差(HSD)检验。此检验允许在进行任何 t 检验之前设置 α_{EW} 值。无论在实验中有多少个组，也无论部分零假设是否为真，Tukey 氏真实显著差检验都能维持 α_{EW} 值在事先设定的水平上。在进行 HSD 检验之前不需要获得显著的 ANOVA 结果。此方法基于 student 化全距统计量(即 q)。

7. 由于 HSD 检验能够更好地将犯一类错误的概率控制在可接受的低水平上，因此它比 LSD 检验更为保守。一个保守的检验比一个自由的检验具有较小的检验力。一个自由的检验犯二类错误的概率更小，但通常是以增加犯一类错误的概率为代价的。

8. 由于 Newman-Keuls 检验通过对所有均数进行排序，然后依据所比较的均数在序列中的距离大小来校正临界 q 值，因此它比 Tukey 氏 HSD 检验具有更大检验力。然而，N－K 检验似乎是通过允许 α_{EW} 大于现有水平来获得额外的检验力的，所以这种检验并不常用。

9. 当想要把某一特定组(如一个控制组)的均数同研究中的其他各组均数进行比较时，Dunnett 检验是一种很好的方法。REGWQ 检验是对 Tukey 氏检验的修订。它能在不放松控制一类错误的条件下拥有更大的检验力。修正 LSD 检验是 REGWQ 检验的一个更简单的替代方法，它在控制好 α_{EW} 的同时也有足够的检验力。

10. 若其他条件相同，则事前比较比事后比较拥有更大检验力。Bonferroni 检验要将所期待的 α_{EW} 除以计划要比较的次数。这种检验太过于保守因此很少用于事后比较，但是当所计划的比较次数相对少时，该检验有很强的检验力。

练习题

*1. 以下实验中各有多少对互不相同的成对比较：

(1) 实验中有 5 个组？

(2) 实验中有 8 个组？

(3) 实验中有 10 个组？

2. 如果多次重复一个两组实验，每次重复