第1章 多重对比

1.1 引言

我们在前一章以范文一为例，讨论了一个自变量、一个应变量、而且自变量只有两个水平时的最简单实验设计。研究者感兴趣的是对产品质量进行降与无序对消费者的质量质感度所产生的影响。现在，我们把另外一种排序即升序排列也加入进来。我们的原理还是一样。根据具体性原则，当产品按质量指标降序排列时，质量指标因为其具体性与直接可比性会更吸引网上消费者的注意力。根据损失厌恶原则，消费者对失去的价值会比对得到的同等价值更敏感。根据这两个原因，当产品按质量做降序排列时，因为此时质量的具体性并给人以失去感，消费者会比在无序排列时有更高的质量敏感度。同样，因为具体性，当产品按质量做升序排列时，消费者会比在无序排列时有更高的质量敏感度；而此时的质量“得到感”又不至于降低消费者的质量敏感度，所以总体来讲，当产品按质量做升序排列时，消费者会比在无序排列时有更高的质量敏感度。这个理论推理可以用以下表格来总结。

	损失厌恶	具体性
随机	标杆水平=0	标杆水平=0
降序	质量失去感+	具体性+
升序	质量得到感+	具体性+

有了这样的理论推理，我们的假设是：

H1：产品的质量排序方法影响消费者的质量敏感度。

这是一个综合性假设，我们也可以有更具体的假设：

H1a：当产品按质量做降序排列时，消费者会比在无序排列时有更高的质量敏感度。

H1b：当产品按质量做升序排列时，消费者会比在无序排列时有更高的质量敏感度。

我们可不可以提出另外一个假设来比较降序与升序呢？因为质量损失感比质量得到感更强烈，而升序与降序的具体性是相同的，所以，我们有理由提出：

H1c：当产品按质量做降序排列时，消费者会比在升序排列时有更高的质量敏感度。

我们设计了这三种网站。试验对象被随机地分配到某一个设计中。我们得到了三个样本。现在我们要如何做假设检验呢？按照前一章所述的方法，读者会说：我们做三次 t 检验不就得了！

1.2 错误率类型

真的希望统计分析就是如此简单！但事与愿违；从两个组到三个组是设计上小小的一步，却是统计分析复杂性上大大的一步。在统计界，在处理多组对比时，提出了很多方法，各个方法各有长短。

为什么情况会变得复杂呢？这都起源于假设错误率在多组对比中的不同。

先回顾我们所熟悉的假设检验中的错误率。在两个组的对比中，我们设错误率为α。也就是说，对于原假设 H0：μ1=μ2，如果我们发现μ1 与μ2 相差在两个标准差之外（对于 T 检验而言），那么，我们得出结论说：数据不支持原假设。但我们的结论有 5%的概率是错的，因为即使μ1 真的等于μ2，由于抽

样本身的随机性，有 5%的机会这两个组样本的均值会相距太远，导致我们得到一个错误的结论。这个错误率在这里叫作**两组对比错误率**(error rate per comparison, ERPC)。

如果我们有多个组要比较，比如在这个例子中有三个组，那么直观而又简单地讲，每一次对比都会有 5%的错误率。我们在三次对比中至少犯一个错误的概率是多大呢？这并不难，至少犯一个错误的概率是 1 减去三个都不犯错误的概率，所以是 $1-(1-0.05)^3=0.1426$。

我们至少犯一个错误的概率是概率居然不低。这也难怪，人要是不想犯错误，最好什么都不做。做了，就会有犯错误的概率；你做的事情越多，出错的机会就越多，要想一件事都不出错的机会就越小。所以难怪当我们做三次对比时出错率就会升高。这种出错率，即在多次对比时至少犯一个错误的概率叫作**家族出错率** (error rate familywise) 。

可见，想做很多事，又想少犯错是一件不容易的事。可恶的是，我们有一个**总体假设**（overall hypothesis）：产品的质量排序方法影响消费者的质量敏感度。也就是说，至少要有一个组它所产生的质量敏感度与其它组不同。假定三个组的均值相同，那么如果只看其中两组，我们观察到不同均值的概率只有 5%，但是因为我们比了三次，这个“一不小心被样本欺骗”的风险就大大增加。所以，就算我们现在观察到的确有一个组，它与其它组的均值有显著不同，我们也不敢说：我们结论的可靠性是 95%。

这个例子说明：要做一件“好事”容易，要做“一辈子好事”难！而当我们要对一个总体假设保持一样的 5%错误率，就好象要求人多做事，但又不能多犯错误一样，难！

那么为什么我们要把家族出错率定得这么低呢，叫人为难呢？首先，这是因为研究者想测试一个总体假设。既然你对总体假设感兴趣，那么没办法，作为一个假设，按照老规矩，我们的要求是错误率为 5%。你也许会说：这样的话，我不要总体假设了！这算是一种知难而退的选择：达不到标准或者怕达不到标准，那就不“吹牛”，不说：我的组中至少有一组是不同的（你的研究还有价值么？）。第二，就算研究者不明明地提出一个总体假设，这个问题还是无法回避。比方说，我们设计了 3 种新药，以一种原有的旧药作为控制组，我们有 4 个组。结果发现，有一个新药 X 比旧药好。就算没有总体假设，我们敢不敢就说：在 5%的错误率下，X 比旧药好？我们不敢这样说。为什么呢？因为就算所有的新药都实际上与旧药一样好，我们有 14%的机会（见以上的计算）在 3 种所谓的新药中发现它比旧药好。实际上，就算你直接在旧药中抽样 3 次，与一个给定的旧药样本比，就有 14%的机会可以找到显著不同。这个例子说明一个问题：就算没有总体假设，只要我们做多组对比，我们就需要控制家族出错率。否则我们的结果就不可靠，我们发现至少一组有显著不同（也即实际上犯第一类错误）的概率就会虚高。第三，有时候，研究者是先观察了各组的均值，然后猜测某些组可能有显著不同，所以想测试一下。这种情况叫做后验检验。后难检验因为是看了数据才做出的，所以被这个样本的特殊性欺骗的可能性就增加了。可能两组间实际上没有不同，但在这个样本中恰巧就有不同了。在这种情况下，我们希望对后验检验采用更严格的标准，即家族出错率。换句话说，如果一个事后诸葛亮要证明他的判断是正确的，人们对他的要求就比事前诸葛亮要更严。

你也许会不服气，会说：这是因为以上的对比都是针对某一个控制组。在实际中，研究者可能只想比较其中的一些，比如，A 也 B 比，C 与 D 比。如果我们把它们当作两个实验来对比，我们可以在 95%的置信度说 A 与 B 或 C 与 D 有没有不同；因为这是两个不同的实验，每个实验又是最简单的，我们不用考虑什么家族出错率。如果我们把它们放在一个实验中，我们反倒要多了一个家族出错率的负担！A 也 B 比，C 与 D 比，风牛马不相及，为什么要求家族出错率？

这个反驳没错。在这种情况下，我们称这些对比是相互独立的。我们会在后面给出相互独立的对比的具体定义。如果一组先验假设（事先所做的假设）是相互独立的，的确没有必要要求家族出错率。但是如果它们不独立，就像那个药物比较的例子，那就免不了家族出错率的负担了。你还可以讨价还价：那如果我的对比有些是想到独立的，有些是不独立的，比如，A、B 与 C 比，E 与 F 比，那么家族出错率是要求在 A、B、C 上面呢，还是包括它们所有？这个问题留给读者。

我们先做一个总结：在只有一个因子，但有两个以上水平的情况下，如果各组之间的对比是独立的，我们可以用 T 检验。如果各组之间的对比不独立，我们必需顾及家族出错率。

1.3 对比的正交性

我们首先必须知道各组之间的对比是否是独立的。任何一个比较，从数学上看，不过是一个**对比**（contrast）。比如，μ1=μ2 可以转化为(1)μ1+(-1)μ2，我们想知道它是不是显著不为 0。这种只涉及到两个组的比较叫作**两组直接对比**(pairwise comparison)。更一般地，任何一个对比都是各组均值的一个线性组合。如果我们有三个组，我们的一部分组合可以是：

$d_1 = \bar{Y}_{\cdot 1} - \bar{Y}_{\cdot 2}$

$d_2 = \bar{Y}_{\cdot 2} - \bar{Y}_{\cdot 3}$

$d_3 = \bar{Y}_{\cdot 1} - \bar{Y}_{\cdot 3}$

$d_4 = \bar{Y}_{\cdot 1} - \frac{1}{2}\bar{Y}_{\cdot 2} - \frac{1}{2}\bar{Y}_{\cdot 3}$

第四种组合表示的是**非两组直接对比**(nonpairwise comparison)。

以我们的例子为例，前三个对比对应我三个提出来的假设。第四个可以理解为：随机排序与有规律排序（升序和降序平均起来）有没有不同？这当然也可以作为一个假设提出来。

这些对比是否独立呢？按数学定义，只有当各对比的系数所代表的向量的积为 0 时，也即当它们正交时，才算是独立。按此定义，以上的对比都是不独立的。

如果我们有四个组：

	1	2	3	4
对比 1	1	-1	0	0
对比 2	1	0	-1	0
对比 3	0	0	1	-1
对比 4	1/2	1/2	-1/2	-1/2

对比 1 与 3 ， 1 、 3 与对比 4 是正交的，但其它对比之间都不正交。对于 C 个组而言，最多正交的对比数不会超过 C-1。

这些对比是谁定的呢？在一个实验中，是否要检验这些对比完全是研究者的决定。这样的对比可以多一点，甚至多于 C，也可以少于 C-1，可以是正交的，也可以不是正交的。但是，一旦做了决定，这个决定就会影响到什么样的多组对比方法是可以合理应用的。

1.4 先验与后验假设

那么在一个因子、多个对比的情况下，研究者往往做哪些假设呢？假设可以按先验(a priori)与后验(a posteriori)分。先验假设是研究者在看到数据之前就计划好的。后验假设则是研究者在观察到数据之后，觉得某两组之间的均值可能有显著不同，对它们所做的检验。在行为研究中，大部分时候，研究者都会事先计划好所有的假设。做后验检验的时候并不多。后验检验是探索性的，它为下一次的研究提供一个可能的方向。后验检验也是数据驱动的，所以往往会被样本的特殊性所“欺骗”，误以为有一些关系是存在的。

那么先验假设有哪些呢？最常见的有几种：第一是对所有组进行两组对比。第二是所有组相对于一个控制组做对比。第三是选择性地对一些组做对比。显然，前两种情况的各对比之间是不正交的。第三种情况需视具体假设而定。这三种情况当然也可以出现在后验检验中。

假设的数量可多可少，这也取决于研究者所想问的问题的多少。

关于假设的这些林林总总的不同将影响多组对比方法的选用。

1.5 正交的先验假设

看上去很多因素会影响家族错误率，从而多组对比方法的选用。这些因素包括对比的数量、对比的正交性、先验与后验检验之分等。我们从研究最常见的需求入手，逐一介绍各种方法。

最简单的情况是当假设是正交的、又是先验的情况。如前所述，如果有四个组，各组的抽样又是相互独立的，研究者的先验假设是比较 A 组与 B 组、C 组与 D 组，或者类似地，比较 A 与 C、B 与 D，或者(A+B)/2 与(C+D)/2、A 与 B 等等。因为这些对比之间的正交性，常规的 t 检验就可行了。在这里，常规的 t 检验指的是当我们假定**观察点是相互独立的、各组观察点是服从同一个正态分布**的情况。请记住这些条件，因为一旦这些条件不满足，使用常规的 t 检验就会不合理，你的论文可能就被拒绝了。如果先定义一些符号：

d：两组的均值的差

α：要求的错误率，在正交对比的情况下，它是**两组对比错误率**，一般定在 0.05

C：组的个数

df：组内自由度，即总体样本数减去 C，或者说是(R-1)C

对于任何正交的、先验的对比中的任何一个，我们定义一个服从 t 分布的 t 值，t_d；对于原假设 H0: d=0，如果：

$$|t_d| \geq t(\alpha, df)$$

那么拒绝 H0。这里，$t(\alpha, df)$ 是在给定错误率α下，自由度为 df 的 t 分布的取值。这里：

$$t_d = \frac{\bar{Y}_{\cdot j} - \bar{Y}_{\cdot k}}{\sqrt{2\frac{MSW}{R}}}$$

TODO：范文中的数据

要注意，当我们采用常规的 t 检验时，家族错误率就大于α。研究者假定这除了这些正交的对比外，论文的读者不会对其它对比感兴趣。这种方法最大的优点是直接、容易测试、容易理解。不幸的是，既然这些组都是一个自变量的不同水平，读者往往对每个组之间的不同都感兴趣。一个常见的问题是：既然这个实验因子可取这些水平，那么哪个最好？正交的对比往往无法回答这样的问题。

TODO：解释 (A+B)/2 与(C+D)/2、A 与 B 这种对比会出现在哪些场合，为什么 A、B 出现在两个不同的对比中，它们还可以算是正交？

1.6 所有的组间对比- Tukey 检验

一个最常见的问题可能是对所有的组进行两两对比，就如本章引言所讲的例子。这时，正交性就会不满足，家族错误率的阴影就倏然出现。那怎样才能保证家族错误率不超过所定的标准（一般是 0.05）呢？思路很简单，多做事，还要少犯错或者只犯与少做事的一样多的错，在每一件事上的错误率就要比常人更低。Tukey 检验(Tukey's test)就是这样一种方法。这种方法是由统计学家 John Tuckey 在 50 年代提出的，又叫 Tukey's HSD (Honestly Significant Difference)。在常规 t 分布，我们知道 5%的错误率对应的是均值左右两个标准差的区间，为了提高每次两组对比的准确度，Tukey 检验对每一个对比采用比两个标准差更大的范围，这就意味着每个对比犯错的概率更小。Tukey 会保证家族错误率在指定的范围内。具体来讲是这样的：对于两个组，定义：

$$q = \frac{\bar{Y}_{\cdot j} - \bar{Y}_{\cdot k}}{\sqrt{\frac{MSW}{R}}} = t_d\sqrt{2}$$

Tukey 证明 q 服从一种叫作**学生氏化的范围分布**(studentized range distribution)。显然，对于一样的置信度，它所要求的置信区间更大。按所要求的准确度（95%），它的关键值可以按组数 C 与自由度（(R-1)C）查 q 的分布表得到。比如当 C=4，R=6，df=4×6-4=20，关键值 q=3.96，转成等价的 t 关键值是 2.8。

Tukey 检验是这样做的，对于原假设 H0: d=0，如果：

$$|t_d| \geq q(1-\alpha, C, df)/\sqrt{2}$$

那么拒绝 H0。读者也许会问，为什么这个检验不按 q 的定义算出 q 的值，再与 q 的关键值比较，就好像做 t 检验一样？这里的做法反而是这样的：不管实际中两个组之间的具体 q 值，先按置信度、组数与自由度找到关键值，把它转化为相应的 t 关键值，然后把实际中两个组的 t 与这个转化而来的 t 关键值比较。这样做的原因为二：一是大家为计算 t 值比较熟悉；二是我们可以很明显比较常规 t 检验所需要的 t 关键值与这个转化而来的 t 关键值的不同。比如当 C=4，R=6，df=20，常规 t 的关键值是 2.0860。这就说明如果使用常规 t 的关键值，我们比使用 Tukey 检验更有可能找到显著不同，所以 Tukey 检验的统计效能更高，但代价是犯第一类错误的概率更大。

Tukey 检验总是假定你要做所有的组间对比，也即一共 C(C-1)/2 个对比。就算你不做所有的组间对比，关键 q 值与所对应的 t 关键值也不会随着变化。显然，如果我们不做所有的对比，我们犯错误的概率就小了，理论上讲关键 q 值与所对应的 t 关键值应该变小才对。所以，如果你不做所有的组间对比，使用 Tukey 检验就“亏了”，你会把一些本来显著的不同当作不显著，或者说，统计效能降低了。亏得多少取决于不做的组间对比的个数。

Tukey 检验把家族错误率控制在α。Tukey 检验可以用在先验假设的检验上。但是 Tukey 检验只适用于两组直接对比，不适用于由几组加权求和所得的非两组直接对比。关于观察点是相互独立的、各组观察点是服从同一个正态分布，而且各组样本量相同的假定条件同样适用在 Tukey 检验上。

TODO：范文中的数据

1.7 部分的组间对比- DUNN 检验

那么如果研究者不想做所有的两两对比，该怎么办呢？在先验假设的前提下(这时，假设的个数当然是知道的)，可以使用 Dunn 检验。Dunn 检验假定假设的个数是 c，这个 c 往往小于等于 C(C-1)/2。与 Tukey 检验的思路一样，对于每一个对比，Dunn 检验也选用比常规 t 分布的关键值更大的关键值，Dunn 检验要求每一个的对比的置信度要达到α'=α/c。对于原假设 H0: d=0，如果：

$$|t_d| \geq t(\alpha/c, df)$$

那么拒绝 H0。要注意，虽然这个对比看似一个关键值不同的 t 检验，但是如果要得到对应于一个α、c、与 df 的关键值，我们不能从 t 的分布表中找，而要用 Dunn 检验自己的分布表。比如当 C=4，R=6，df=20，如果我们做四个两两对比，c=6，我们所得的关键值为 2.93。这个值大于常规 t 的关键值 2.09，也大于 Tukey 检验的 2.8。这个例子似乎说明在一样保证家族错误率的情况下，Dunn 检验能只发现更少的显著不同。事实是，如果 c 比较大（接近于或等于 C(C-1)/2）时，Dunn 检验的关键值往往大于 Tukey 检验的关键值；如果 c 比较小，那么 Dunn 检验的统计效能往往更高。这是因为 Dunn 检验的关键值是 c 的函数，会随着 c 减小而减小，所以当 c 减小时这个关键值“不浪费”。而 Tukey 检验的关键值不会随对比的数量而变化，所以在对比少时就显得“浪费”了。那么你也许会问，到底多大最大，多小算小，好叫我可以明确地决定选用 Dunn 检验或 Tukey 检验？答案却没这么简单。答案是，根据α、C、与 df，你先查到 Tukey 检验的关键值；然后，按α、c、与 df，你再查到 Dunn 检验的关键值；最后，谁小就选谁。

Dunn 检验把家族错误率控制在α以内。Dunn 检验可以用在先验假设的检验上。但是 Dunn 检验适用于两组直接对比，不适用于非两组直接对比的情形，这与 Tukey 检验一样。关于观察点是相互独立的、各组观察点是服从同一个正态分布，而且各组样本量相同的假定条件同样适用在 Dunn 检验上。如果先验假设是所有的两两对比，选用 Tukey 检验，否则研究者要根据估计的样本量，在看到数据之前选定 Tukey 检验或 Dunn 检验。

TODO：范文中的数据

1.8 与控制组对比-DUNNETT 检验

一种常见的对比是几个组与一个控制组比较。比如，我们把 A、B、与 C 各组分别与 D 组比。显然，用 Tukey 检验太“浪费”了，统计效能不高。这时，Dunn 检验可以考虑。但是，在这个特殊情况下，Dunnett 检验更有效。与 Tukey 检验的思路一样，对于每一个对比，Dunnett 检验也选用比常规 t 分布的关键值更大的关键值。对于原假设 H0: d=0，如果：

$$|t_d| \geq D(\alpha, C, df)$$

那么拒绝 H0。

这里的 $D(\alpha, C, df)$ 表示 Dunnett 检验所用的统计分布表，可以根据α、c、与 df 的值查到关键值。比如当 C=4，R=6，df=20，对于三个对比(这时 Dunnett 检验只能有三个对比)，关键值是 2.54，而在相同的对比数时，Dunn 检验的关键值是 2.61。这时，选用 Dunnett 检验就更合理。

Dunnett 检验把家族错误率控制在α。Dunnett 检验可以用在先验假设的检验上。但是 Dunnett 检验只适用于几个组与一个控制组比较的情形，不适用于其它情形。关于观察点是相互独立的、各组观察点是服从同一个正态分布，而且各组样本量相同的假定条件同样适用在 Dunnett 检验上。

TODO：范文中的数据

1.9 任何对比-SCHEFFÉ 检验

以上的检验方法都是用在两组直接对比的情形。Scheffé 检验允许对任何数目、任何组合的对比进行检验，并不局限于两组直接对比的情形，也不要求对比个数小于 C(C-1)/2。可以说，它是一种适用更广、更灵活的方法。但这并不等于说它是最有效的方法。这个灵活性的代价是有时它的效能比较低。与其它检验的思路一样，对于每一个对比，Scheffé 检验也选用比常规 t 分布的关键值更大的关键值。对于原假设 H0: d=0，如果：

$$|t_d| \geq \sqrt{(C-1)F(\alpha, C-1, df)}$$

那么拒绝 H0。

这里的 F 是我们所熟悉的 F 分布，自由度为组的个数 C-1 与 df。这个关键值与对比的个数无关，如同 Tukey 检验时的做法。当 C=4，R=6，df=20，不管多少个对比，关键值是 3.05。在四个对比的情况下，这个值比 Tukey 检验与 Dunn 检验更保守。当然，如果对比的数目增加，Scheffé 检验的优势就会显现出来。

Scheffé 检验把家族错误率控制在α。Scheffé 检验可以用在先验假设的检验上。但是 Scheffé 检验不只适用于两组直接对比，也适用于其它对比的情形，比如(A+B)/2 与(C+D)/2 和 A 与 B。Scheffé 检验在对比的数目上没有限制。关于观察点是相互独立的、各组观察点服从同一个正态分布，而且各组样本量相同的假定条件同样适用在 Scheffé 检验上。

1.10其它方法

在假定满足的情况下，以上的方法似乎可以应付大部分的对比了。除了以上的方法，还有另外一些对比检验方法。

有一类方法叫做分步测试法(stepwise methods)。这些办法往往是先对各组的均值按从小到大的次序进行排列。然后，对最大的不同进行测试，即把第一组与第 C 组对比。如果没有不同，那么其它组之间就不用比了，一定没有不同。如果第一组与第 C 组的对比有不同，那么接下去测第一组与第 C-1 组的对比，一直到没有不同为止。这类方法中有 Newman-Keuls 检验，Ryan (REGWQ)检验等。其中 Newman-Keuls 检验是不推荐的。如果要用，就用 Ryan 检验。

要注意，这类方法只有当我们观察到数据之后才能开始，所以只能用在后验假设的检验上。这类方法的一个主要应用是对因子的各个水平进行排序，并根据各组的差距做选择。在行为研究中，这类应用并不常见，所以我们不对此多做介绍。

另一类方法叫作**有掩护的检验**(protected test)。它的基本做法是先对所有组作为一个总体做 F 检验，如果其中至少有一个均值是不同，然后再对各组进行对比。否则对比停止。在这一类方法中最有名的是 Fish's least significant difference (LSD)，又叫有掩护的 t 检验。它的做法是它的基本做法是先对所有组作为一个总体做 F 检验，如果其中至少有一个均值是不同，然后对各组有常规 t 检验进行测试。但这种方法不能保证家族错误率，所以不被推荐。还有的方法包括 Shaffer-Ryan 检验、Fisher-Hayter 检验。虽然 Shaffer-Ryan 检验可能比前面提到的方法更好，但因为一般的软件并不支持这些方法，它们很少出现在实际应用中。我们对这些方法不做展开。

1.11条件不满足情形下的检验

以上的检验方法都假定理想的条件，即观察点是相互独立的、各组观察点服从同一个正态分布，而且各组样本量相同。如果这些条件不满足，以上的方法就可能不可靠。我们来看一看每一个条件不满足情况下的对策。

先看正态性。如果各组不服从正态分布，以上的方法还可靠吗？好消息是我们所重点讨论的检验方法对这正态性不敏感。在不正态的数据中，它们仍然可靠。

再看样本量的不同。如果样本量不同，我们讨论的方法大都需要修改。

TODO: 多大的不同可以算不同。

在一个简单的修改办法是对 t_d 的计算方法进行修改。记得在对两个方差相同，但样本量不同的组计算总体方差时，我们所用的公式是：

$$s_{jk}^2 = s_{pooled}^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right) = MSW\left(\frac{1}{n_1}+\frac{1}{n_2}\right),$$

根据这个方差，我们可以得到两个组的均值 t 检验是：

$$t_d^* = \frac{\overline{Y}_j - \overline{Y}_k}{\sqrt{MSW\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}},$$

有了这个新的 t_d，这个值就可以与 $q(1-\alpha, C, df)/\sqrt{2}$ 比较，这样就可以检验任何两个组的不同了。这个改进的过程叫做 Tukey-Kramer 检验。但是，这个方法只适用于两组直接对比。

再来看各组方差不同时的情形。先看多大的方差不同可以“忽略”。前人的研究表明，如果样本量相同，如果最大与最小方差的比例不超过 4（标准差相差不超过一倍），在对所有的组间进行比较时，以上的检验方法仍然可靠。其中相对保守的方法比如 Dunn 与 Scheffé 检验受的影响就更小了。 就

算是方差比例非常不同，第一类错误虽会增加，却往往不会大于 0.075。所以如果研究者研究接受这个错误率，以上的方法还可以接受。

最后，我们来看一个更现实的问题。现实中，各组的样本量与方差会同时不同。在这同情况下，一个最常用的检验是 GH 检验。对于任何两组之间的直接对比，GH 检验采用 t_d^*，与 Tukey-Kramer 检验相同。但是，它的关键值不只 $q(1-\alpha, C, df)/\sqrt{2}$，而是 $q(1-\alpha, C, df_{jk})/\sqrt{2}$。我们在前一章已经定义：

$$df_{jk} = \frac{\left(\frac{s_{\cdot j}^2}{n_j} + \frac{s_{\cdot k}^2}{n_k}\right)^2}{\frac{\left(\frac{s_{\cdot j}^2}{n_j}\right)^2}{n_j - 1} + \frac{\left(\frac{s_{\cdot k}^2}{n_k}\right)^2}{n_k - 1}}。$$

GH 并不能使家族错误率维持在 0.05，有时会略高。

1.12 总结