第一章　征服一般线性模型
——General Linear Model 菜单详解（上）

在我看来，学习统计软件时不去了解所用统计原理和模型，就像是学英语时只背单词而不记语法一样，是永远不可能说出一口流利的英文的。

——张文彤

请注意，本章的标题用了一些修辞手法，一般线性模型可不是三言两语就能说清楚的，因为它包括的内容实在太多了。那么，究竟什么是一般线性模型呢？用最通俗的话讲：方差分析模型和线性回归模型都是分析一个/多个自变量对一个/多个连续性应变量的影响，并且都假设应变量和自变量是线性数量关系，因此他们都算是一般线性模型的特例，凡是和他们沾边的都可以用 GLM 来做。包括成组设计的方差分析（即单因素方差分析）、配伍设计的方差分析（即两因素方差分析）、多因素方差分析、多元方差分析、重复测量的方差分析、协方差分析、多元线性回归分析等等。因此，能真正掌握 GLM 菜单的用法，会使大家的统计分析能力有极大地提高。

☞ 什么？GLM 还可以完成线性回归分析！是的，它的确可以。只不过他作回归分析的功能没有专用的回归分析模块强，所以该功能较少被用到。

⚠ 有些书中将 General Linear Models 翻译成广义线性模型，实际上 GLM 模块的功能只限于一般线性模型。广义线性模型对应的英文是 Generalized Liner Models，指的是范围更广、功能更强大的另一大类模型，在 SPSS 中可以编程拟合，但并非 GLM 菜单提供的对话框就可以完成的，在本书的最后一章中对其有基本的介绍。

好了，本章要讲述的菜单名为 General Linear Model，既然一般线性模型的能力如此强大，那么下属的四个过程各自的功能是什么呢？请看：

- ✧ Univariate 过程：四个菜单中的大哥大，当应变量只有一个时，我们所进行的分析就要用它来完成。显然，它是用的最多的一个。
- ✧ Multivariate 过程：当结果变量（应变量）不止一个时，当然要用他来分析啦！
- ✧ Repeated Measures 过程：顾名思义，重复测量的数据就要用它来分析，这一点可能要强调一下，用前两个菜单似乎都可以分析出来结果，但在许多情况下该结果是不正确的。在相应章节我会详细讲述。
- ✧ Variance Components 过程：用于对层次数据拟合方差成份模型，它是普通线性模型向随机效应的进一步扩展，是一种可以考察各个层次因素的变异大小，从而为哪些层次上可能存在组内聚集性、如何可能减小数据变异提供信息的统计方法，也就是现在非常热门的多水平模型的最原始形式。

由于 Univariate 过程是最常用的一个过程，本章我们就围绕它来进行讲解，另外三个过程将放在下一章学习。

www.statstar.com

§1.1　方差分析模型简介

现实世界中变量间的联系是错综复杂的，比如要研究性别对身高的影响，显然就要考虑到年龄、遗传、营养状况等因素的作用。这时 t 检验作为单因素分析方法就无能为力了，而方差分析可以在控制其他因素影响的同时研究两者之间的关系，分析的效率更高，适用范围更广。

同时，许多时候各自变量之间还会存在交互作用，如研究催化剂对化学反应的催化能力，如果该催化剂只在某个温度范围内效果最佳，则只单独研究该催化剂的催化作用是没有实际意义的，此时这种交互作用也成为了我们研究的重点，即必须要研究在什么温度条件下该催化剂的催化能力最佳。对交互作用的分析也是方差分析模型的特长。

1.1.1　模型入门

在均数的比较一章中我们实际上已经接触到了方差分析的基本思想，这里首先来复习一下单因素方差分析中变异的分解公式：

总变异=处理因素导致的变异+随机变异

实际上，该公式只是为了便于大家理解，标准的单因素方差分析模型如下：

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

其中 X_{ij} 表示第 i 组的第 j 个观察值；μ 表示总体的平均水平；α_i 表示影响因素在 i 水平下对应变量的附加效应，并假设所有 α_i 之和应当为 0；ε_{ij} 为一个服从正态分布 $N(0, \sigma^2)$ 的随机变量，代表随机误差。一般情况下，我们做假设检验实际上就是检验各个 α_i 是否均为 0，如都为 0，即各组总体均数都相等，则 X_{ij} 就会成为服从正态分布 $N(\mu, \sigma^2)$ 的一个变量。

有的时候以上模型也被写为 $X_{ij} = \mu_i + \varepsilon_{ij}$，此时检验的含义就变成了各组均数 μ_i 是否相同。显然，他和上面的公式是等价的，但这种写法应当更加容易理解。比如在上一章中我们比较三组石棉矿工的用力肺活量有无差别，那么石棉肺患者组中某一位观察对象的用力肺活量数值就等于石棉肺患者组的平均水平再加上一个随机误差项。而如果三组总体均数相同，则它就应当等于总体均数（平均水平）再加上一个随机误差项，实际上就变成了同一个变量分布中的某一点。

多因素方差分析模型只是对以上公式的进一步扩展，以两因素方差分析模型为例，公式如下：

$$X_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

其中 α_i、β_j 分别表示 A 因素 i 水平和 B 因素 j 水平的附加效应，ε_{ijk} 仍为随机误差变量。此时如果要说明因素 A 有无影响，就是检验如下假设：

H_0：$\alpha_i=0$，H_1：至少有一个 $\alpha_i \neq 0$

如果要说明因素 B 有无影响，就是检验如下假设：

H_0：$\beta_i=0$，H_1：至少有一个 $\beta_i \neq 0$

www.StatStar.com

而所说的模型无显著性就是指上面两个 H_0 同时成立（均不能被拒绝）。

更为复杂的是如考虑交互作用的情形，模型如下：

$$X_{ijk} = \mu + \alpha_i + \beta_j + \alpha_i \beta_j + \varepsilon_{ijk}$$

其中 α_i、β_j 分别表示 A 因素 i 水平和 B 因素 j 水平的附加效应。$\alpha_i \beta_j$ 则为两者的交互效应。

可能有的朋友已经看得头痛了，但无论怎样，只要记住方差分析的原理即可：根据资料类型以及研究目的，可将总变异分解为两个或多个部分，除一部分代表随机误差的作用外，每个部分的变异可由某因素的作用（或交互作用）来解释，通过比较可能由某因素所致的变异与随机误差的大小，借助 F 分布做出推断，即可了解该因素对结果变量的影响是否存在。

1.1.2　常用术语

方差分析中的常用术语有：

1．因素（Factor）：因素是可能对应变量有影响的变量，一般来说，因素会有不止一个水平，而分析的目的就是考察或比较各个水平对应变量的影响是否相同。例如影响农作物产量的因素有气温、降雨量、日照时间等。在方差分析中，因素的取值范围不能无限，只能有若干个水平，即应当为分类变量。

2．水平（Level）：因素的不同取值等级称作水平，例如性别有男、女两个水平。需要注意的是有时候水平是人为划分出来的，比如身高被分为高、中、低三个水平。

3．单元（Cell）：指各因素水平之间的组合，我们所说的方差齐就是指的各个单元间的方差齐。注意有的试验设计并不会给出所有可能的水平组合，如拉丁方设计。

4．元素（Element）：指用于测量应变量值的最小单位，比如研究石棉矿工用力肺活量，则肺活量是从每一位矿工身上测得，矿工就是试验的元素。一个单元格内可以有多个元素，也可以只有一个，甚至于没有元素。

5．均衡（Balance）：如果一个实验设计中任一因素各水平在所有单元格中出现的次数相同，且每个单元格内的元素数相同，则该试验是均衡的，否则，就被称为不均衡。不均衡的实验设计在分析时较为复杂。

6．固定因素（Fixed Factor）与随机因素（Random Factor）：两者都是因素的不同种类，固定因素指的是该因素在样本中所有可能的水平都出现了。换言之，该因素的所有可能水平仅此几种，针对该因素而言，从样本的分析结果中就可以得知所有水平的状况，无需进行外推。比如要研究糖尿病、IGT（糖耐量异常）和正常人的血糖有无差别，则按照糖尿病有无可将所有人分为糖尿病、IGT 和正常人三种，此时该因素就被认为是固定因素。另外，有些人为设定的因素，比如下面引例中的 group，他被分为工前、工中、工后这三个值，如果我们所做的检验就是想弄清楚这三个水平对应变量有无影响，不需要外推到其他水平（如工后半小时），则它也可以被认为是固定因素。

和固定因素相对应的是随机因素，他指的是该因素所有可能的取值在样本中没有都出现，或不可能都出现。如下面引例中的 worker，它表示的是工人这个配伍因素，实际上总体中当然不可能只有这 10 个工人，他们只是所有工人的代表而已。因此要用样本中

www.statstar.com

10 个工人对应变量的影响情况来推论总体中全体工人对应变量的影响情况，包括未出现的那些工人，这不可避免的存在误差（即随机效应），需要估计该误差的大小，因此被称为随机因素。又如我们要研究什么温度下催化剂的效果最好，样本中取了 30、40、50℃三个水平，如果我们希望在分析结果中能同时外推 35℃、45℃这些水平的情况，此时温度也是随机因素。

一般来说固定因素和随机因素在分析时应分别指定，如果将随机因素按固定因素来分析，则可能得出错误的分析结果。但是如果所有单元格内都至多只有一个元素，则随机效应无法被估计出来，此时两种做法的统计分析结论完全相同。

7．交互作用（Interaction）：如果一个因素的效应大小在另一个因素不同水平下明显不同，则称为两因素间存在交互作用。当存在交互作用时，单纯研究某个因素的作用是没有意义的，必须分另一个因素的不同水平研究该因素的作用大小。

如果所有单元格内都至多只有一个元素，则交互作用无法测量，只能不予考虑，最典型的例子就是配伍设计的方差分析。

1.1.3　方差分析模型的适用条件

方差分析并不是万金油，他也有自己的适用条件（以 H_0 假设成立为前提），具体说有以下几点：

- ✧ 各样本的独立性：只有各样本为相互独立的随机样本，才能保证变异的可加性（可分解性）。
- ✧ 正态性：即所有观察值系从正态总体中抽样得出。
- ✧ 方差齐：这里所说的方差齐是指假设总的模型无意义时方差齐，亦即每一个单元格中的方差齐。

道理大家其实都懂，但做起来就不一定都明白了，为什么多因素方差分析一般都不提这些条件呢？首先在以上条件中，对独立性的要求是最严的，但他一般都可以保证。其次为正态性和方差齐性，具体来说就是：

- ✧ 单因素方差分析：适用条件是必须要考虑的问题，尤其是正态性和方差齐性一般都需要进行考察。
- ✧ 配伍设计的方差分析：不考虑这两个问题，这是由于正态性和方差齐性的考察是以单元格为基本单位的，此时每个格子中只有一个元素，当然没法分析了！不止是配伍设计的方差分析，只要是无重复数据的方差分析，如交叉设计、正交设计等，都不考虑这两个问题。
- ✧ 有重复数据的多因素方差分析：由于方差齐性的考察是以单元格为基本单位，此时单元格数目众多，真正分到每个格子中的样本例数一般都只有 3~5 例，此时很难检验出差别；或者出现另一种极端情况，因为极个别格子方差不齐而导致检验不能通过，这种情况实际上对分析结果影响并不太严重。因此在多因素方差分析中，方差齐性往往只限于理论探讨的程度。真正在应用时，只要数据分布不是明显偏态，不存在极端值即可，这两种情况对结果的影响要比方差不齐严重得多。

www.StatStar.com

根据 Box 的研究结果，如果各组的例数相同（即均衡），或总体呈正态分布，则方差分析模型对方差略微不齐有一定的耐受性，只要最大／最小方差之比小于 3，分析结果都是稳定的。

在实际操作中，对数据正态性的考察有一个办法，就是拟合完毕后作出残差分布图，如果残差呈随机分布，则可知（单元格内）原始数据满足正态条件。

www.StatStar.com