多因素方差分析模型入门

@文彤老师

模型表达式入门

- 假设现在希望描述某个人群的月收入状况，那么根据统计学知识，均数能够表示集中趋势，标准差能够表示离散趋势，则任何一位受访者i的月收入X_i该如何表达？

$$X_i = \mu + \varepsilon_i$$

  - 显然，这里的的ε_i应当服从正态分布，其均数为0，标准差为相应总体标准差
  - 在只有样本信息时，样本均数和标准差就是上述参数的最佳估计值。

模型表达式入门

- 现在希望比较三种职业的月收入有无差异，这三类职业分别是医生、律师和软件工程师。
  - 如果我们仍然希望能够对每一个个体的数据加以表达，应当如何做？

$$X_{1j} = \mu_1 + \varepsilon_{1j}$$

$$X_{2j} = \mu_2 + \varepsilon_{2j}$$

$$X_{3j} = \mu_3 + \varepsilon_{3j}$$

模型表达式入门：单因素方差分析

观测变量

控制因素

每个人具体的月薪	职　业
Xxx,xxx,xxx,xxx Xxx,xxx,xxx,xxx	医　生
Xxx,xxx,xxx Xxx,xxx,xxx,xxx	律　师
Xxx,xxx,xxx,xxx Xxx,xxx	软件工程师

三个水平

模型表达式入门

- 将上面三个式子可以合并如下：

$$X_{ij} = \mu_i + \varepsilon_{ij}$$

- 为了进一步分析的方便，一般都会寻找一个均数的参照水平，将其余组的平均水平与之相比

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

- 显然，这样的组合会有许多种，因此模型在实际分析的时候往往会加上一些限制条件，比如假设参照水平是最后一个组的均数，这被称为拟合的约束条件

模型表达式入门

- 由于在常见的研究中，我们更关心各组均数的差别，对于标准差的差别则比较忽视，因此在最初的方差分析模型中，往往将不同组的ε_{ij}假设为服从相同的正态分布（就是说相同）

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \varepsilon_{ij} \sim (0, \delta^2)$$

  - 注意：在后来发展的混合效应模型和多水平模型中，各组间离散程度的差异也进入了研究视野，此时模型不一定会加入此限制

模型表达式入门

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

- 如果职业1和职业2的平均收入不相等，则应当有α1≠α2
  - H0: α1=α2
- 如果三种职业的平均收入无差异，则应当有α1=α2=α3=0，此时如果采用适当的参照水平，就有
  - H0：αi＝0，H1：至少有一个αi≠0

案例：胶合板磨损深度的比较

- 现希望比较四种胶合板的耐磨性，分别从这四个品牌的胶合板中抽取了5个样品，在相同的转速下磨损相同时间，测量其被磨损的深度（mm），现希望对此进行分析，数据见veneer.sav
  - 方差齐性检验
  - 模型参数估计值与设计矩阵

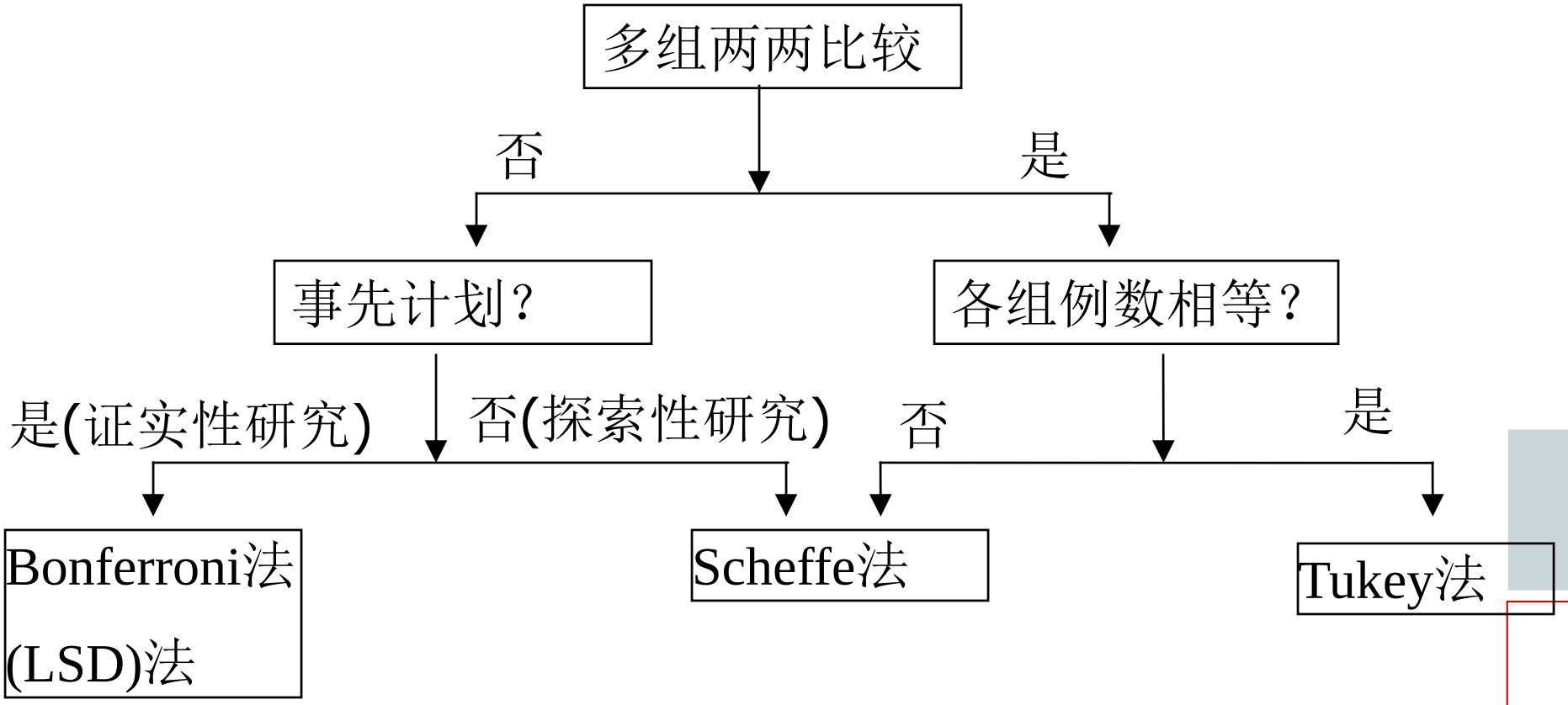
两两比较方法

- LSD法：实际上就是t检验的变形，只是在变异和自由度的计算上利用了整个样本信息，因此仍然存在放大一类错误的问题
- Scheffe法：当各组人数不相等，或者想进行复杂的比较时，用此法较为稳妥。但它相对比较保守

两两比较方法

- S-N-K法：是运用最广泛的一种两两比较方法。它采用Student Range 分布进行所有各组均值间的配对比较。该方法保证在H0真正成立时总的α水准等于实际设定值，即控制了一类错误。
- 方差不齐时的两两比较方法：一般认为是Games-Howell法稍好一些，但最好直接使用非参数检验方法

两两比较方法的选择


多组两两比较
否
是
事先计划？
各组例数相等？
是(证实性研究)
否(探索性研究)
否
是
Bonferroni法
(LSD)法
Scheffe法
Tukey法

多因素方差分析模型

- 目的：在同时考虑若干个控制因素的情况下，分别分析它们的改变是否造成观察变量的显著变动

	医生	律师	软件工程师
男性	XX,XX, XX,XX	XX,XX, XX,XX	XX,XX
女性	XX,XX, XX,XX	XX,XX, XX,XX	XX,XX, XX,XX

模型表达式入门

- 如果只研究职业的影响

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

- 如果只研究性别的影响

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

- 同时考虑职业和性别对收入的影响

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

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

方差分析模型常用术语

- 因素（Factor）
  - 因素是可能对因变量有影响的变量，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因变量的影响是否相同。
- 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。

方差分析模型常用术语

- 单元（Cell）
  - 单元亦称试验单位（Experimental Unit），指各因素的水平之间的每种组合。指各因素各个水平的组合，例如在研究性别（二水平）、血型（四水平）对成年人身高的影响时，该设计最多可以有2*4＝8个单元。注意在一些特殊的试验设计中，可能有的单元在样本中并不会出现，如拉丁方设计。

方差分析模型常用术语

- 元素（Element）
  - 指用于测量因变量值的观察单位，比如研究职业与收入间的关系，月收入是从每一位受访者处得到，则每位受访者就是试验的元素
  - 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。
    - 这主要在一些特殊的设计方案中出现，如正交设计

方差分析模型常用术语

- 均衡（Balance）
  - 如果在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均相同，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，需要对方差分析模型作特别设置才能得到正确的分析结果。

方差分析模型常用术语

- 固定因素（Fixed Factor）
  - 指的是该因素在样本中所有可能的水平都出现了。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。
  - 绝大多数情况下，研究者所真正关心的因素都是固定因素。
    - 性别：只有两种
    - 疗法：只有三种

方差分析模型常用术语

- 随机因素（Random Factor）
  - 该因素所有可能的取值在样本中没有都出现，目前在样本中的这些水平是从总体中随机抽样而来，如果我们重复本研究，则可能得到的因素水平会和现在完全不同！
  - 这时，研究者显然希望得到的是一个能够“泛化”，即对所有可能出现的水平均适用的结果。这不可避免的存在误差，需要估计误差的大小，因此被称为随机因素。

方差分析模型常用术语

- 协变量（Covariates）
  - 指对因变量可能有影响，需要在分析时对其作用加以控制的连续性变量
  - 实际上，可以简单的把因素和协变量分别理解为分类自变量和连续性自变量
  - 当模型中存在协变量时，一般是通过找出它与因变量的回归关系来控制其影响

方差分析模型常用术语

- 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。

效应的检验方法

- 无论模型结构多复杂，假设检验都是基于变异分解的原理进行的，都是F检验。
- 根据变异分解式，可以将总的样本离均差平方和分解成各个部分，随后各个离均差平方和除以自由度可得到均方，进而将各效应的均方和误差均方相比较，就得到了F统计量

方差分析模型的检验层次

- 对总模型进行检验
- 对模型中各交互效应、主效应进行检验
  - 交互项有统计学意义：分解为各种水平的组合情况进行检验
  - 交互项无统计学意义：进行主效应各水平的两两比较

方差分析模型的适用条件

- 从模型表达式出发得到的提示
  - 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：即个单元格内的所有观察值系从正态总体中抽样得出
  - 方差齐：各个单元格中的数据离散程度均相同，即各单元格方差齐

方差分析模型的适用条件

- 实际运用
  - 在多因素方差分析中，由于个因素水平组合下来每个单元格内的样本量可能非常少，这样直接进行正态性、方差齐检验的话检验效能很低，实际上没什么用
  - 因此真正常见的做法是进行建模后的残差分析

案例：超市规模、货架位置与销量的关系

- 现希望现希望考察对超市中销售的某种商品而言，是否其销售额会受到货架上摆放位置的影响，除此以外，超市的规模是否也会有所作用？甚或两者间还会存在交互作用？
- Berenson和Levine（1992）着手研究了此问题，他们按照超市的大小（三水平）、摆放位置（四水平）各随机选取了两个点，记录其同一周内该货物的销量。

案例：超市规模、货架位置与销量的关系

		货物摆放位置			
		A	B	C	D
超市规模	小型	45、50	56、63	65、71	48、53
	中型	57、65	69、78	73、80	60、57
	大型	70、78	75、82	82、89	71、75

- 方差齐性检验问题
- 边际均数和轮廓图
- 残差分布图

案例：广告宣传效果的比较

- 现希望研究四种广告的宣传效果有无差异，具体的广告类型为：店内展示、发放传单、推销员展示、广播广告。在本地区共有几百个销售网点可供选择，出于经费方面的考虑，在其中随机选择了18个网点进入研究，各网点均在规定长度的时间段内使用某种广告宣传方式，并记录该时间段内的具体销售额。为减小误差，每种广告方式在每个网点均重复测量两次。数据见ranavona.sav。