第十章 方差分析

10.1 方差分析引论

10.1.1 方差分析的概念

假设检验一次只能研究两个样本。随着个体显著检验的次数增加，偶然因素导致差别的可能性增加。方差分析同时考虑所有样本，因此排除错误累计的概率，避免拒绝真实的原假设。

方差分析是通过检验各总体的均值是否相等来判断分类型自变量对数值型自变量是否有显著影响。方差分析的几个概念：

- 因素（因子）：指方差分析要检验的对象（通常来说是变量X等，例如比较不同行业的服务质量，那么行业就是这个因子）；
- 水平（处理）：因素的不同表现，例如分析不同行业的服务质量，可以是“餐饮业”“零售业”等；
- 观测值：在每个水平下得到的样本数据。

10.1.2 方差分析的基本思想和原理

图像描述： 用散点图可以简单判断多个水平是否有显著差异。

误差分解： 通过对数据误差来源的分析判断总体均值是否相等，进而分析各个自变量对因变量是否有显著影响。

- 组内误差（SSE）：来自水平内部的数据误差，反映了一个样本内部数据的离散程度；
- 组间误差（SSA）：不同水平之间的数据误差，反应了不同样本之间数据的离散程度；
- 总误差（SST）：全部数据误差大小的平方和。

误差分析：

- 如果没有显著影响：组间误差只包含随机误差，组间误差误差与组内误差经过平均后的数值应该很接近；
- 如果有影响：组间误差还包括系统误差，比值会大于1，越大，就说明影响越显著。

10.1.3 方差分析的几种假定

（1）每个总体都应服从正态分布；

（2）各个总体的方差 σ^2 必须相同；

（3）观测值是独立的。

10.1.4 问题的一般想法

$$\begin{array}{ll} H_0: \mu_1 = \mu_2 = \cdots = \mu_k & \text{自变量对应变量没有显著影响} \\ H_1: \mu_1, \mu_2, \ldots, \mu_k \text{不全相等} & \text{自变量对应变量有显著影响} \end{array} \tag{1}$$

10.2 单因素方差分析

用 A 表示因素，k 个水平分别用 $A_1, A_2, \ldots, A_k$ 表示，观测值为 $x_{ij},\ (i = 1, 2, \ldots, k;\ j = 1, 2, \ldots n)$，表示第 i 个水平中的第 j 个观测值。

10.2.2 分析步骤

提出假设

$$\begin{array}{ll} H_0: \mu_1 = \mu_2 = \cdots = \mu_k & \text{自变量对应变量没有显著影响} \\ H_1: \mu_1, \mu_2, \ldots, \mu_k \text{不全相等} & \text{自变量对应变量有显著影响} \end{array} \tag{2}$$

构造检验的统计量

1. 计算各水平的均值 $\overline{x}_i = \sum_{i=1}^{n_i} x_{ij}/n_i\ ,\ i = 1, 2, \ldots, k$；
2. 计算全部观测值的总均值 $\overline{\overline{x}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}$；
3. 计算各误差平方和
   - 总平方和 SST

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{\overline{x}})^2 \tag{3}$$

   - 组间平方和

$$SSA = \sum_{i=1}^{k} n_i (\overline{x}_i - \overline{\overline{x}}) \tag{4}$$

   - 组内平方和

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)^2 \tag{5}$$

4. 计算各统计量
   - SSA 的自由度为 $k-1$，SST 的自由度为 $n-1$，SSE 的自由度为 $n-k$

$$MSA = \frac{SSA}{k-1}, \quad MSE = \frac{SSE}{n-k} \tag{6}$$

- 比较 MSA 和 MSE，得到需要的检验统计量 F

$$F = MSA/MSE \sim F(k-1, n-k) \tag{7}$$

做出统计决策：使用单侧检验，如果 $F > F_\alpha$ 则拒绝原假设，否则接受原假设。

方差分析表： 如果需要可以看 P215

10.2.3 关系强度的测量

用组间平方和（SSA）占总平方和（SST）的比例反应关系强度。得到关系强度，$1 - R^2$ 表示残差影响（其他因素的影响）。

$$R^2 = \frac{SSA}{SST} \tag{8}$$

10.2.4 方差分析中的多重比较

方差分析的多重比较用于确定不确定性出现在哪些水平（处理）之间。课本使用最小显著差异方法（LSD），具体步骤为：

1. 提出 C_k^2 个假设，$H_0 : \mu_i = \mu_j;\ H_1 : \mu_i \neq \mu_j$
2. 计算检验统计量 $|\bar{x}_i - \bar{x}_j|$
3. 计算 $LSD = t_{\alpha/2}\sqrt{MSE(\frac{1}{n_i}) + \frac{1}{n_j}}$
4. 根据显著性水平 α 做出决策，检验统计量大于 LSD 则拒绝原假设，否则接受原假设。

10.3 双因素方差分析

单因素方差分析研究的是自变量（因素）对因变量有没有影响，双因素方差分析研究的是两个因素对因变量的影响是否有相互作用。如果两个因素各自影响因变量，则称为无重复双因素分析；如果两种因素搭配产生一种新的影响则称为可重复双因素分析。

在双因素方差分析的数据结构中，行和列分别安排一个因素。k 个列因素，r 个行因素，一共 kr 个观察数据。（标号行在前列在后）

10.3.2 无交互作用的双因素分析

预备值：分别逐行、逐列计算观测值的均值 $\bar{x}_{i\cdot}$ 和 $\bar{x}_{\cdot j}$，以及全部 kr 个样本数据的总平均值 $\bar{\bar{x}}$；

分析步骤：

1. 分别对两个因素提出假设（原假设为均值全相等，备择假设是不全相等）。
2. 计算 $SST = SSC + SSR + SSE$，自由度分别为 $kr-1, k-1, r-1, (k-1)(r-1)$

$$\begin{aligned} SST &= \sum_{i=1}^{k}\sum_{j=1}^{r}(x_{ij} - \bar{\bar{x}})^2 \\ &= \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{r}(\bar{x}_{i\cdot} - \bar{\bar{x}})^2}_{\text{行因素产生的误差平方和}SSR} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{r}(\bar{x}_{\cdot j} - \bar{\bar{x}})^2}_{\text{列因素产生的误差平方和}SSC} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{r}(x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}})^2}_{\text{除行、列因素其他因素产生的误差平方和}SSE} \end{aligned} \tag{9}$$

3. 为构造检验统计量需要计算 $MSR, MSC，MSE$

$$MSR = \frac{SSR}{k-1} \quad , MSC = \frac{SSC}{r-1} \quad , MSE = \frac{SSE}{(k-1)(r-1)} \tag{10}$$

4. 计算 F 值，并与 F_α 比较（单侧），做出决策。

$$\begin{aligned} F_R &= \frac{MSR}{MSE} \sim F(k-1,(k-1)(r-1)) \\ F_C &= \frac{MSC}{MSE} \sim F(r-1,(k-1)(r-1)) \end{aligned} \tag{11}$$

5. 方差分析表（见 P223）

10.3.3 有交互作用的双因素方差分析

太多了，不整了。基本上就是除了行和列分别有 r 和 k 个值，每个值对应有 m 个样本，然后代入公式。