统 计 学 Statistics

陈 灯 塔

Econ, XJTUCC

CDT.WISE@G

2021 年 5 月 4 日

内容梗概

大数据

统计学会有什么变革

- 还需要抽样吗？
- 假设检验还有用处吗？
- 所有犯罪分子都无利可逃了吗？
- AI 能探索出新理论吗？

方差分析

方差分析 (Analysis of variance, ANOVA), 是假设检验问题，即检验多个 (三组及三组以上) 子总体 (群体) 的均值是否显著差异: 通过检验各子总体的均值是否相等，来判断分类变量对数值变量是否有显著影响

检验均值，为什么叫方差分析？

对于相同的样本均值差异，如果样本内部差异越小，表明各群组的均值越不相等。样本均值间差异，以及样本内部差异都由方差来度量，即计算组间方差和组内方差，所以叫方差分析。

传统方法

平方和

不同的分组也称为水平或处理 (treatment). 假定从第 i 个组中抽取一个容量为 n_i 的简单随机样本, x_{ij} 为第 i 组的第 j 个观测值. 总均值

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i} x_{ij} = \frac{1}{n}\sum_{i=1}^{g} n_i\bar{x}_i$$

其中 $n = \sum_{i=1}^{g} n_i$, 组 i 的均值 $\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$

$$\mathrm{TSS} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2$$

$$\mathrm{BSS} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\bar{x}_i - \bar{x})^2 = \sum_{i=1}^{g} n_i(\bar{x}_i - \bar{x})^2$$

$$\mathrm{WSS} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$$

平方和

total sum of squares = within-group sum of squares + between-group sum of squares

$$\mathrm{TSS} = \mathrm{WSS} + \mathrm{BSS}$$

- 组内平方和 (within-group sum of squares) $\sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2$ SSE
- 组间平方和 (between-group sum of squares) $\sum_{i=1}^{g} n_i(\bar{x}_i-\bar{x})^2$ SSA

原假设下 (各组均值无差异), 通常 $\bar{x}_i \approx \bar{x}$, 此时 $\mathrm{BSS} \approx 0$, $\mathrm{TSS} \approx \mathrm{WSS}$, 从而 $\frac{\mathrm{BSS}}{\mathrm{WSS}} \to 0$. 对于相同的样本均值差异，如果样本内部差异越小，表明各群组的均值越不相等 ($\frac{\mathrm{BSS}}{\mathrm{WSS}} \to \infty$)

基本假定

方差分析的基本假定

- 每个子总体都应服从正态分布
- 各个子总体的方差必须相同
- 观察值是独立的

原假设

$$\mathbb{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_g$$

$$\mathbb{H}_1 : \mu_1, \mu_2, \cdots, \mu_g \text{ 非同一值}$$

检验统计量 $F = \frac{\mathrm{BSS}/(g-1)}{\mathrm{WSS}/(n-g)} \sim F(g-1, n-g)$

注意：拒绝原假设，只表明至少有两个群体的均值不相等，并不意味着所有的均值都不相等

两两比较

为什么不做两两比较？

- 要检验四个总体的均值是否相等，每次检验两个的作法共需要进行 6 次不同的检验，每次检验犯第一类错误的概率为 α，连续作 6 次检验犯第　类错误的概率增加到 $1-(1-\alpha)6=0.265$, 大于 0.05。相应的置信水平会降低到 0.956=0.735
- 方差分析方法则是同时考虑所有的样本，因此排除了错误累积的概率，从而避免拒绝一个真实的原假设

可能错误

又是以讹传讹？(类似的说法在课本 p213L-11)　　若原假设成立，组间均方 (平方和除以相应的自由度) 与组内均方的数值就应该很接近，它们的比值就会接近 1

应该是接近于零，因为组间均方 $\to 0$, by $\mathrm{BSS} = \sum_{i=1}^{g} n_i(\bar{x}_i - \bar{x})^2 \to 0$

哑变量法

哑变量

哑变量 (dummy variable, 虚拟变量, 0-1 变量) 回归估计的方法

$$Y = a + \sum_{i=2}^{g} b_i X_i + e$$

其中 $X_i = 1_{X=i}$. 该设定下，以第一组为基准组, 显然

$$\mu_1 = \mathrm{E}(Y|X=1) = a$$

$$\mu_i = \mathrm{E}(Y|X=i) = a + b_i \qquad i = 2, 3, \cdots, g$$

表明

$$b_i = \mu_i - \mu_1 \qquad i = 2, 3, \cdots, g$$

即第 i 组与第一组在均值上的差别。检验单个 $b_i = 0$ 是 t 检验, 即第 i 组与第一组在均值上是否显著差别。方差检验就是回归的 F 检验

$$b_2 = b_3 = \cdots = b_g = 0 \iff \mu_1 = \mu_2 = \cdots = \mu_g$$

总结

1. 方差分析 (了解)
2. 哑变量法也许更灵活 (假设条件可以放宽)，更方便 (与基准的差异可以看到) (了解)

Keyword: 方差分析, 哑变量

Homework: ($*$ by group)

- Time loss of smart phones, pads and computers in classroom
- 基本概念要牢记，牢记，牢记!
- Read
  1. Textbook: full book for final exam
  2. Reference book: corresponding chapters