

# 统计方法与机器学习

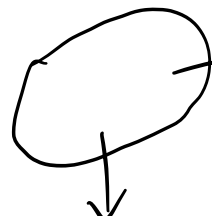
## 第一章：方差分析

倪 蓓

DaSE@ECNU  
(lni@dase.ecnu.edu.cn)

假设检验: 控制第一类错误率

## 目录


 → 一包白色粉末 (混合物)  
 ↓  
 检测 20 种毒物  
 assume only 10

$H_{1,0}$	毒	1	存在	$A_1$
$H_{2,0}$	毒	2	存在	$A_2$
$\vdots$	毒	$\vdots$	存在	
$H_{20,0}$		20		$A_{20}$

每种检验错误率 5%

### 1 多重比较

水平均值差的置信区间  
 多重比较问题  
 Tukey 方法

$$\begin{aligned}
 P(A) &= P\left(\bigcap_{i=1}^{20} A_i\right) = 1 - P(\bar{A}) \\
 &= 1 - P\left(\overline{\bigcap_{i=1}^{20} A_i}\right) \\
 &= 1 - P\left(\bigcup_{i=1}^{20} \bar{A}_i\right) \\
 &= \prod_{i=1}^{20} P(A_i) \\
 &= \prod_{i=1}^{20} (1 - P(\bar{A}_i)) \\
 &= (1 - 0.05)^{20} = 60\%
 \end{aligned}$$

$\bar{A}$ : 无毒

# 水平均值差的置信区间

## 概述

- 在单因子方差分析模型中，经检验，因子  $A$  是显著的。
- 有充分的理由认为因子  $A$  的各个水平中至少存在一对水平的均值是不相等的。
- 但这并不说明，所有的水平均值都不相等的。

$$H_0: \mu_1 = \mu_2 = \dots = \mu_a = 0$$

# 水平均值差的置信区间

## 概述

- 问题：我们想知道哪些水平的均值是不相等的。
- 一个自然的想法：给定一对水平  $(i, i')$ ，构造  $\mu_i - \mu_{i'}$  的区间估计。

# 水平均值差的置信区间

## 回顾：枢轴量法

- 分布为

$$\bar{y}_{i.} \sim N(\mu_i, \sigma^2 m^{-1}) \quad \text{和} \quad \bar{y}_{i' .} \sim N(\mu_{i'}, \sigma^2 m^{-1})$$

- $\bar{y}_{i.}$  和  $\bar{y}_{i' .}$  是独立的。
- 于是,

$$\bar{y}_{i.} - \bar{y}_{i' .} \sim N(\mu_i - \mu_{i'}, 2\sigma^2 m^{-1}).$$

- 但是, 这个分布中  $\sigma^2$  是未知的。
- 我们用  $\hat{\sigma}^2$  代替  $\sigma^2$ 。

# 水平均值差的置信区间

## 回顾：枢轴量法

- 因为

$$SS_E / \sigma^2 \sim \chi^2(n - a)$$

且与  $\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}$  独立。

- 方差的估计为

$$\hat{\sigma}^2 = \frac{SS_E}{n - a}$$

- 因此，枢轴量为

$$\frac{(\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}) - (\mu_i - \mu_{i'})}{\sqrt{\frac{2}{m} \hat{\sigma}}} \sim t(n - a).$$

# 水平均值差的置信区间

## 概述

- 问题：我们想要知道哪些水平的均值是不相等的。
- 一个自然的想法：给定一对水平  $(i, i')$ ，构造  $\mu_i - \mu_{i'}$  的区间估计。
- 置信水平为  $1 - \alpha$  的置信区间为

$$(\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}) \pm \sqrt{\frac{2}{m}} \hat{\sigma} \cdot t_{1-\alpha/2}(n - a)$$

# 水平均值差的置信区间

## 概述

- 置信区间与双侧假设检验是存在对应关系的。
- 置信水平为  $1 - \alpha$  的置信区间为

$$(\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}) \pm \sqrt{\frac{2}{m}} \hat{\sigma} \cdot t_{1-\alpha/2}(n - a)$$

可以转化为两正态总体均值差的检验问题

$$H_0 : \mu_i = \mu_{i'} \quad \text{vs} \quad H_0 : \mu_i \neq \mu_{i'}$$

的接受域。

- 如果置信区间覆盖零，则认为  $\mu_i$  与  $\mu_{i'}$  无明显差异；
- 若置信区间未覆盖零，则认为  $\mu_i$  与  $\mu_{i'}$  之间存在明显的差异。



# 多重比较问题

## 概述

- 由于因子  $A$  总共有  $a$  个不同的水平，总共有

$$\binom{a}{2} = \frac{a(a-1)}{2}.$$

- 对不同的水平组合。对于每一对水平  $(i, i')$ ,

$$(\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}) \pm \sqrt{\frac{2}{m}} \hat{\sigma} \cdot t_{1-\alpha/2}(n-a)$$

是  $\mu_i - \mu_{i'}$  的置信区间，置信水平为  $1 - \alpha$ 。

- 然而，总共有  $a(a-1)/2$  个区间，要求其同时成立，其联合置信水平就无法达到  $1 - \alpha$ 。

# 多重比较问题

## 概述

- 若  $A_1, A_2, \dots, A_k$  表示  $k$  个随机事件，且每个事件发生的概率均为  $1 - \alpha$ ，即  $P(A_i) = 1 - \alpha, i = 1, 2, \dots, k$ ，则其共同发生的概率为

$$P\left(\bigcap_{i=1}^k A_i\right) \leq P(A_1) = 1 - \alpha$$

$$P\left(\bigcap_{i=1}^k A_i\right) = 1 - P\left(\bigcup_{i=1}^k \bar{A}_i\right)$$

$$\begin{aligned} &\geq 1 - \sum_{i=1}^k P(\bar{A}_i) = 1 - k(1 - (1 - \alpha)) \\ &= 1 - k\alpha. \end{aligned}$$

- 这表明了它们同时发生的概率实际上应介于  $1 - k\alpha$  和  $1 - \alpha$  之间，可能比  $1 - \alpha$  小得多。

# 多重比较问题

置信水平  $1-\alpha$

## 概述

- 为了使得它们同时发生的概率不低于  $1 - \alpha$ ，一个很自然的方法是把每一个事件发生的概率提高。
- 具体来说，将  $t_{1-\alpha/2}(n-a)$  调整为  $t_{1-\alpha/(a(a-1))}(n-a)$ ；
- 这样使得每个置信区间的置信水平提高到  $1 - \alpha/(a(a-1)/2)$ ；
- 于是，

$$P\left(\bigcap_{i=1}^{a(a-1)/2} A_i\right) \geq \frac{(1-\alpha)_{a(a-1)/2}}{1 - a(a-1)/2 \cdot \frac{\alpha}{a(a-1)/2}} = 1 - \alpha.$$

- 称该方法为 Bonferroni 方法。
- 虽然简单，但是会导致所得到的置信区间过于保守，精度很差。

# 多重比较问题

## 概述

- 在方差分析中，经  $F$  检验拒绝原假设，表明因子  $A$  是显著的，即  $a$  个水平的均值不全相等。
- 进一步，我们需要确定哪些水平之间是存在差异的，哪些水平之间是没有差异的。
- 在  $a(a > 2)$  个水平均值中同时比较任意两个水平均值间有无明显差异的问题称为**多重比较**。
- 也就是说，在显著性水平为  $\alpha$  同时检验  $a(a - 1)/2$  个假设

$$H_0^{ii'} : \mu_i = \mu_{i'}, \quad 1 \leq i < i' \leq a.$$

- 当  $H_0^{ii'}$  成立时， $|\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}|$  不应过大，过大就应拒绝  $H_0^{ii'}$ 。

在原假设成立的假设下去考察拒绝域

## 多重比较问题

### 概述

- 于是，在同时考察  $a(a-1)/2$  个假设  $H_0^{ii'}$  时，这些  $H_0^{ii'}$  中至少有一个不成立就构成了多重比较检验问题的拒绝域，即拒绝域的形式为

$$W = \bigcup_{1 \leq i < i' \leq a} \{|\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}| \geq c_{ii'}\},$$

其中  $c_{ii'}$  是临界值，由原假设  $H_0^{ii'}$  成立时  $P(W) = \alpha$  而确定。

# Tukey 方法

## 概述

- 我们需要求  $a(a-1)/2$  个临界值  $\{c_{ii'} : 1 \leq i < i' \leq a\}$ ;
- 为了简化这个问题，我们可以对所求的临界值提出一些合理的假设；
- 由于各个水平下重复次数均相等，基于对称性一个很自然的要求是  $c_{ii'}$  是相等的，我们记为  $c$ 。

# Tukey 方法

## 概述

- 考虑多重比较的检验问题

$$H_0^{ii'} : \mu_i = \mu_{i'}, \quad 1 \leq i < i' \leq a$$

- 在原假设成立时,  $\mu_1 = \mu_2 = \cdots = \mu_a = \mu$ .

# Tukey 方法

## 概述

- 我们有

$$\begin{aligned} P(W) &= P\left(\bigcup_{1 \leq i < i' \leq a} \{|\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}| \geq c\}\right) \\ &= 1 - P\left(\bigcap_{1 \leq i < i' \leq a} \{|\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}| < c\}\right) \\ &= 1 - P\left(\max_{1 \leq i < i' \leq a} |\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}| < c\right) \\ &= P\left(\max_{1 \leq i < i' \leq a} |\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}| \geq c\right) \\ &= P\left(\max_{1 \leq i < i' \leq a} \left| \frac{(\bar{y}_{i\cdot} - \mu) - (\bar{y}_{i'\cdot} - \mu)}{\hat{\sigma}/\sqrt{m}} \right| \geq \frac{c}{\hat{\sigma}/\sqrt{m}}\right) \\ &= P\left(\max_i \frac{\bar{y}_{i\cdot} - \mu}{\hat{\sigma}/\sqrt{m}} - \min_i \frac{\bar{y}_{i\cdot} - \mu}{\hat{\sigma}/\sqrt{m}} \geq \frac{c}{\hat{\sigma}/\sqrt{m}}\right). \end{aligned}$$

$P(\bigcup_{i=1}^n A_i) = 1 - P(\bigcap_{i=1}^n \bar{A}_i)$   
对立事件

极差  
peers里差异最大的 - 差异最小的



## 概述

不是枢轴量 Tukey 方法  
 $\frac{\bar{y}_{i.} - \mu}{\sqrt{\frac{\sigma^2}{m}}}$  未知, 用估计  $\hat{\sigma}$  替代

- 令

$$q(a, df) = \max_i \frac{\bar{y}_{i.} - \mu}{\hat{\sigma} / \sqrt{m}} - \min_i \frac{\bar{y}_{i.} - \mu}{\hat{\sigma} / \sqrt{m}}.$$

- 因为

$$\frac{\bar{y}_{i.} - \mu}{\hat{\sigma} / \sqrt{m}} \sim t(n - a),$$

- $q(a, df)$  可以看作  $a$  个独立同分布的自由度为  $df$  的  $t$  分布的随机变量的极差;
- 所以, 一般称  $q$  为  $t$  化极差统计量。
- 这个分布并不是我们常见的分布之一, 这个分布与水平数目  $a$  和  $t$  分布的自由度  $df = n - a$  有关, 但与  $\mu, \sigma^2, m$  都无关。

# Tukey 方法

## 概述

- 如何获得  $t$  化极差统计量的分布？
- 该分布可以通过蒙特卡洛的方法获得。
- 具体算法如下。

---

## 算法 $t$ 化极差统计量的蒙特卡洛分布

---

Require: 水平数目  $a$ ,  $t$  分布的自由度  $df$ , 重复次数  $N$ ;

Ensure:  $t$  化极差统计量的  $N$  个观测值

- 1: for  $n = 1, 2, \dots, N$  do
  - 2:     从标准正态分布  $N(0, 1)$  产生  $a$  个随机数:  $x_1, x_2, \dots, x_a$ ;
  - 3:     将  $a$  个数据进行排序, 令  $x_{\max}$  为最大值,  $x_{\min}$  为最小值;
  - 4:     从自由度为  $df$  的  $\chi^2$  分布产生一个随机数  $y$ ;
  - 5:     计算  $q_n = (x_{\max} - x_{\min}) / \sqrt{y/df}$ ;
-

# Tukey 方法

## 概述

- 于是，由

$$P(W) = P(q(a, df) \geq \sqrt{m}c/\hat{\sigma}) = \alpha$$

可推出

$$c = q_{1-\alpha}(a, df)\hat{\sigma}/\sqrt{m}$$

其中， $q_{\alpha}(a, df)$  表示  $q(a, df)$  的  $\alpha$  分位数。

# Tukey 方法

## 步骤

- 在给定的显著性水平  $\alpha$  下，确定  $t$  化极差统计量的分位数  $q_{1-\alpha}(a, df)$ ，并计算  $c = q_{1-\alpha}(a, df)\hat{\sigma}/\sqrt{m}$ ；
- 比较每一组样本均值的差  $|\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}|$  临界值  $c$  的大小；
- 如果

$$|\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}| \geq c$$

- 那么认为水平  $i$  与水平  $i'$  之间有显著差异；反之，则认为这两个水平无差异。