第八章 方差分析与回归分析

Instructor: 郝壮

haozhuang@buaa.edu.cn
School of Economics and Management
Beihang University

April 13, 2022

第八章方差分析与回归分析

- §8.1 方差分析
- §8.2 多重比较 (拓展阅读)
- §8.3 方差齐性分析 (方差分析基本假定的检验, 略)
- §8.4 一元线性回归 (下半学期讲)
- §8.5 一元非线性回归 (下半学期讲)

§8.1 方差分析(ANOVA, Analysis of Variance)

前几章讨论的是一个总体或两个总体的统计分析问题.

实际工作中我们经常碰到多个正态总体均值的比较问题, **检验多个总体均值是否相等的统计方法, 称为方差分析**(analysis of variance, 缩写ANOVA). 如:

- 行业对被投诉次数是否有显著影响?
- 家电的品牌对它们的销售量是否有影响?
- 教学中采用不同的教法对教学效果(学生成绩)是否存在影响?
- 多个处理组的随机控制实验中, 各处理是否有不同的效果?

上述问题都是考察如何衡量分类变量对数值变量的影响. 都可以用方差分析的方法解决.

8.1.1 方差分析问题的提出

例8.1.1 在饲料养鸡增肥的研究中, 某研究所提出三种饲料配方: A_1是以鱼粉为主的饲料, A_2是以槐树粉为主的饲料, A_3是以苜蓿粉为主的饲料. 为比较三种饲料的效果, 特选24只相似的雏鸡随机均分为三组, 每组各喂一种饲料, 60天后观察它们的重量. 试验结果如下表所示:

表8.1.1 鸡饲料试验数据

饲料A	鸡重(克)							
A_1	1073	1009	1060	1001	1002	1012	1009	1028
A_2	1107	1092	990	1109	1090	1074	1122	1001
A_3	1093	1029	1080	1021	1022	1032	1029	1048

本例中, 我们要比较的是三种饲料对鸡的增肥作用是否相同.

为此, 把饲料称为因子, 记为A, 三种不同的配方称为因子A的三个水平,记为A_1, A_2, A_3, 使用配方A_i下第 j 只鸡60天后的重量用y_{ij}表示, $i = 1, 2, 3, j = 1, 2, ..., 8$.

我们的目的是比较三种饲料配方下鸡的平均重量是否相等. 为此, 需要做一些基本假定, 把所研究的问题归结为一个统计问题, 然后用方差分析的方法进行解决.

8.1.2 单因子方差分析(One-way ANOVA)的统计模型

在**例8.1.1**中我们只考察了一个因子, 称其为单因子试验.

通常, 在单因子试验中, 记因子为 A, 设其有r个水平(levels of factor), 记为$A_1, A_2, \ldots, A_r$, 在每一水平下考察的指标可以看成一个总体, 现有 r 个水平, 故有 r 个总体. 假定:

- **每一总体均为正态总体**, 记为 $N(\mu_i, \sigma_i^2), i = 1, 2, \ldots, r$
- **各总体的方差相同**: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_r^2 = \sigma^2$
- **从每一总体中抽取的样本是相互独立的**, 即所有的试验结果 y_{ij} 都相互独立.

我们要比较各水平下的均值是否相同,即要对如下的一个假设进行检验:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_r$$

备择假设为$H_1 : \mu_1, \mu_2, \ldots, \mu_r$不全相等.

如果H_0成立, 因子A的r个水平均值相同, 称因子A的r个**水平间没有显著差异**, 简称**因子A不显著**;

反之, 当H_0不成立时, 因子A的r个水平均值不全相同, 这时称因子A的**不同水平间有显著差异**, 简称**因子A显著**.

思考：多总体均值比较能否使用t检验？

1. 需要进行多次比较： t检验一次研究两个样本; 例8.1.1中有3个总体, 意味着要进行3次两样本的t检验

2. 增加犯第一类错误的概率：假设每次t检验犯第一类错误的概率是0.05, 那么整体犯第一类错误的概率将达到

$$1-(1-0.05)^3=0.143$$

为对假设进行检验, 需要从每一水平下的总体抽取样本, 设从第i个水平下的总体获得m个试验结果, 记 y_{ij}表示第i个总体的第j次重复试验结果. 共得如下$n = r \times m$个试验结果:

$$y_{ij}, \quad i = 1, 2, \ldots, r, \quad j = 1, 2, \ldots, m$$

其中r为水平数, m为重复数, i为水平编号, j 为重复序号(每组观察编号).

在水平A_i下的试验结果y_{ij}与该水平下的指标均值μ_i (sample mean) 一般总是有差距的, 记 $\varepsilon_{ij} = y_{ij} - \mu_i$称为随机误差. 于是有

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

称为试验结果 y_{ij} 的**数据结构式**(data structure formula).

综上, **单因子方差分析的统计模型可表示为**

$$\begin{cases} y_{ij} = \mu_i + \varepsilon_{ij}, i = 1, 2, \ldots, r, j = 1, 2, \ldots, m \\ \varepsilon_{ij}\text{相互独立, 且都服从}\varepsilon_{ij} \sim N(0, \sigma^2) \end{cases}$$

总均值与效应:

- 称诸 μ_i 的平均$\mu = \frac{1}{r}(\mu_1 + \ldots + \mu_r) = \frac{1}{r}\sum_{i=1}^{r} \mu_i$**为总均值/一般平均 (Grand Mean)**.
- 称第 i 水平下的均值 μ_i与总均值μ 的差: $a_i = \mu_i - \mu$为因子A的**第i水平的主效应**, 简称A_i的**水平效应/或处理效应 (treatment effect)**.

该模型可以改写为

$$\begin{cases} y_{ij} = \mu + a_i + \varepsilon_{ij}, i = 1, 2, \ldots, r, j = 1, 2, \ldots, m \\ \sum_{i=1}^{r} a_i = 0 \\ \varepsilon_{ij}\text{相互独立, 且都服从}\varepsilon_{ij} \sim N(0, \sigma^2) \end{cases}$$

假设检验可改写为

$$H_0 : a_1 = a_2 = \ldots = a_r = 0$$
$$H_1 : a_1, a_2, \ldots, a_r\text{不全为0}$$

为了检验该假设, 我们要构造一个检验统计量(F 检验). 为此, 需引入**平方和分解(Decomposition of sum of squares)**的概念(方差分析的基本原理).

8.1.3 平方和分解(误差分解)(Decomposition of sum of squares)

一.试验数据

表8.1.2 单因子方差分析试验数据

因子水平	试验数据				和	平均
A_1	y_{11}	y_{12}	$\cdots$	y_{1m}	T_1	$\bar{y}_{1\cdot}$
A_2	y_{21}	y_{22}	$\cdots$	y_{2m}	T_2	$\bar{y}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	y_{r1}	y_{r2}	$\cdots$	y_{rm}	T_r	$\bar{y}_{r\cdot}$
					T	$\bar{y}$

$$T_i = \sum_{j=1}^{m} y_{ij} \quad \bar{y}_{i\cdot} = \frac{T_i}{m} \quad i = 1, 2, \cdots, r$$

$$T = \sum_{i=1}^{r} T_i \quad \bar{y} = \frac{T}{r \cdot m} = \frac{T}{n}$$

$n = r \cdot m =$ 总试验次数

二.组内偏差(within group deviation)与组间偏差(between-group deviation)

数据y_{ij}与总平均$\bar{y}$间的偏差可用$y_{ij}-\bar{y}$表示, 它可分解为二个偏差之和$y_{ij}-\bar{y}=(y_{ij}-\bar{y}_{i\cdot})+(\bar{y}_{i\cdot}-\bar{y})$. 记

$$\bar{\varepsilon}_{i\cdot}=\frac{1}{m}\sum_{j=1}^{m}\varepsilon_{ij},\quad \bar{\varepsilon}=\frac{1}{r}\sum_{i=1}^{r}\bar{\varepsilon}_{i\cdot}=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{m}\varepsilon_{ij}$$

由于

$$y_{ij}-\bar{y}_{i\cdot}=(\mu_i+\varepsilon_{ij})-(\mu_i+\bar{\varepsilon}_{i\cdot})=\varepsilon_{ij}-\bar{\varepsilon}_{i\cdot}$$

所以$y_{ij}-\bar{y}_{i\cdot}$**仅反映组内数据与组内平均的随机误差**, 称为**组内偏差**; 而

$$\bar{y}_{i\cdot}-\bar{y}=(\mu_i+\bar{\varepsilon}_{i\cdot})-(\mu+\bar{\varepsilon})=a_i+\bar{\varepsilon}_{i\cdot}-\bar{\varepsilon}$$

除了反映随机误差外, 还反映了第i个水平的效应, 称为**组间偏差**.

三.偏差平方和(sum of squares of deviations)及其自由度(degree of freedom)

定义： 把k个数据$y_1, y_2, \ldots, y_k$分别对其均值$\bar{y} = (y_1 + \ldots + y_k)/k$ 的偏差平方和

$$Q = (y_1 - \bar{y})^2 + \cdots + (y_k - \bar{y})^2 = \sum_{i=1}^{k} (y_i - \bar{y})^2$$

称为k个数据的**偏差平方和**, 有时简称平方和(sum of squares). 它常用来**度量若干个数据分散的程度**.

三.偏差平方和(sum of squares of deviations)及其自由度(degree of freedom)

$$Q = (y_1 - \bar{y})^2 + \cdots + (y_k - \bar{y})^2 = \sum_{i=1}^{k} (y_i - \bar{y})^2$$

- 在构成偏差平方和Q的k个偏差$y_1 - \bar{y}, \ldots, y_k - \bar{y}$间有一个恒等式 $\sum_{i=1}^{k} (y_i - \bar{y}) = 0$, 这说明在$Q$中独立的偏差只有$k-1$个.
- 把**平方和中独立偏差个数**称为该平方和的**自由度**, 常记为f. 如Q的自由度为$f_Q = k - 1$. **自由度是偏差平方和的一个重要参数**.

四.总平方和分解公式

各y_{ij}间总的差异大小可用**总偏差平方和(total sum of squares, SST)**

$$S_T = \sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij} - \bar{y})^2$$

表示, 其自由度为$f_T = n - 1$.

四.总平方和分解公式

仅由随机误差引起的数据间的差异可以用组内偏差平方和

$$S_e = \sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij} - \bar{y}_{i.})^2$$

表示, 也称为**误差偏差平方和(SSE)**, 其自由度为 $f_e = n - r$

由于组间差异除了随机误差外, 还反映了效应间的差异, 故由**效应不同引起的数据差异可用组间偏差平方和**

$$S_A = m\sum_{i=1}^{r}(\bar{y}_{i.} - \bar{y})^2$$

表示. 也称为**因子A的偏差平方和(SSA)**,其自由度为 $f_A = r - 1$.

四.总平方和分解公式

定理8.1.1 在上述符号下, 总平方和S_T可以分解为因子平方和S_A与误差平方和S_e之和, 其自由度也有相应分解公式, 具体为:

$$S_T = S_A + S_e, \quad f_T = f_A + f_e$$

上两式通常称为**总平方和分解式**.

更常见的表达形式为

$$SST = SSA + SSE$$

总偏差平方和=组间偏差平方和/因子A偏差平方和/因子平方和+组内偏差平方和/误差偏差平方和/误差平方和

The total sum of squares(SST) = sum of squares between treatments/factors (SSA) + sum of squares of the residual error (SSE)

四.总平方和分解公式

证明： 注意到

$$\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y}_i)(\bar{y}_{i\cdot}-\bar{y})=\sum_{i=1}^{r}\left[(\bar{y}_{i\cdot}-\bar{y})\sum_{j=1}^{m}(y_{ij}-\bar{y}_{i\cdot})\right]=0,$$

故有

$$\begin{aligned}S_T&=\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y})^2=\sum_{i=1}^{r}\sum_{j=1}^{m}\left[(y_{ij}-\bar{y}_{i\cdot})+(\bar{y}_{i\cdot}-\bar{y})\right]^2\\&=S_e+S_A+2\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y}_{i\cdot})(\bar{y}_{i\cdot}-\bar{y})=S_e+S_A,\end{aligned}$$

四.总平方和分解公式

基于上述定理, 我们可以引出**构造检验统计量的直觉:**

- 数学上, 方差分析就是要分析数据的总误差中有没有处理误差(因子偏差平方和) (即S_T主要是由S_A解释还是主要由S_e解释).
- 固定S_T, S_A越大, S_e越小, 即S_A/S_e越大, 越有理由认为存在处理误差, 即拒绝H_0, 认为因素对观测指标有影响.
- 需要考虑比值S_A/S_e的统计分布.

8.1.4 检验方法

偏差平方和Q的大小与自由度有关, 为了比较偏差平方和, 引入了**均方和**(mean squares)的概念. 它定义为$MS = Q/f_Q$, 其意为平均每个自由度上有多大的平方和, 它比较好地度量了一组数据的离散程度.

如今要对因子平方和 S_A与误差平方和 S_e之间进行比较, 用其均方和 $MS_A = S_A/f_A$, $MS_e = S_e/f_e$ 进行比较更为合理, 故可用

$$F = \frac{MS_A}{MS_e} = \frac{S_A/f_A}{S_e/f_e}$$

作为检验H_0的统计量.

定理8.1.2 在单因子方差分析模型

$$\begin{cases} y_{ij} = \mu + a_i + \varepsilon_{ij}, i = 1, 2, \ldots, r, j = 1, 2, \ldots, m \\ \sum_{i=1}^{r} a_i = 0 \\ \varepsilon_{ij}\text{相互独立, 且都服从}\varepsilon_{ij} \sim N(0, \sigma^2) \end{cases}$$

及前述符号下, 有

- (1) $S_e/\sigma^2 \sim \chi^2(n-r)$, 从而$E(S_e) = (n-r)\sigma^2$
- (2) $E(S_A) = (r-1)\sigma^2 + m\sum_{i=1}^{r} a_i^2$, 进一步, 若$H_0$成立, 则$S_A/\sigma^2 \sim \chi^2(r-1)$
- (3) S_A与S_e独立

证明应用随机变量基本性质, 见教材.

由**定理8.1.2**, 若H_0成立, 则检验统计量$F = \frac{MS_A}{MS_e} = \frac{S_A/f_A}{S_e/f_e}$服从自由度为$f_A$和$f_e$的$F$分布, 因此拒绝域为

$$W = \{F \geq F_{1-\alpha}(f_A, f_e)\}$$

将上述计算过程列成一张表格, 称为**方差分析表**:

表8.1.3 单因子方差分析表

来源	平方和	自由度	均方和	F比	p值
因子	S_A	$f_A = r-1$	$MS_A = S_A/f_A$	$F = MS_A/MS_e$	p
误差	S_e	$f_e = n-r$	$MS_e = S_e/f_e$		
总和	S_T	$f_T = n-1$			

对给定的α, 可作如下判断:

- 若$F \geqslant F_{1-\alpha}(f_A, f_e)$, 则认为因子$A$显著;
- 若$F < F_{1-\alpha}(f_A, f_e)$, 则说明因子$A$不显著.

- 该检验的 p 值也可利用统计软件求出, 若以Y记服从$F(f_A, f_e)$的随机变量, 则检验的 p 值为 $p = P(Y \geqslant F), Y \sim F(f_A, f_e)$.

常用的各偏差平方和的计算公式如下：

$$S_T = \sum_{i=1}^{r}\sum_{j=1}^{m} y_{ij}^2 - \frac{T^2}{n}$$

$$S_A = \frac{1}{m}\sum_{i=1}^{r} T_i^2 - \frac{T^2}{n}$$

$$S_e = S_T - S_A$$

一般可将计算过程列表进行.

例8.1.2 采用例8.1.1的数据, 将原始数据减去1000, 列表给出计算过程：

表8.1.4 例8.1.2的计算表

水平	数据(原始数据-1000)								T_i	T_i^2	$\sum_{j=1}^{m} y_{ij}^2$
A_1	73	9	60	1	2	12	9	28	194	37636	10024
A_2	107	92	-10	109	90	74	122	1	585	342225	60355
A_3	93	29	80	21	22	32	29	48	354	125316	20984
和									1133	505177	91363

可算得各偏差平方和为:

$$S_T = 91363 - \frac{1133^2}{24} = 37876, \qquad f_T = 24 - 1 = 23$$
$$S_A = \frac{50517}{8} - \frac{1133^2}{24} = 9660, \qquad f_A = 3 - 1 = 2$$
$$S_e = S_T - S_A = 37876 - 9660 = 28216, \qquad f_e = 3(8-1) = 21$$

把上述诸平方和及其自由度填入方差分析表

表8.1.5 例8.1.2的方差分析表

来源	平方和	自由度	均方和	F比	p值
因子	9660	2	4830	3.59	0.0456
误差	28216	21	1344		
总和	37876	23			

- 若取$\alpha = 0.05$, 则$F_{0.95}(2, 21) = 3.47$, 由于$F = 3.59 > 3.47$, 故认为因子A(饲料)是显著的, 即三种饲料对鸡的增肥作用有明显的差别.

8.1.5 参数估计(拓展内容)

在检验结果为显著时, 我们可进一步求出总均值μ, 各主效应a_i和误差方差σ^2的估计.

一. 点估计

由诸y_{ij}相互独立, 且$y_{ij} \sim N(\mu + a_i, \sigma^2)$, 因此可用极大似然方法求出总均值$\mu$, 各主效应$a_i$和误差方差$\sigma^2$的估计.

1. 似然函数为(利用y_{ij}的密度函数, 求各y_{ij}同时发生的概率)

$$L\left(\mu, a_1, \cdots, a_r, \sigma^2\right) = \prod_{i=1}^{r}\prod_{j=1}^{m}\left\{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{\left(y_{ij}-\mu-a_i\right)^2}{2\sigma^2}\right\}\right\}$$

2. 对数似然函数

$$\ln L\left(\mu, a_1, \cdots, a_r, \sigma^2\right) = -\frac{n}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{r}\sum_{j=1}^{m}\left(y_{ij}-\mu-a_i\right)^2$$

3. 一阶条件

$$\begin{cases} \frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\mu-a_i) = 0 \\ \frac{\partial \ln L}{\partial a_i} = \frac{1}{\sigma^2}\sum_{j=1}^{m}(y_{ij}-\mu-a_i) = 0, \quad i = 1,\cdots,r \\ \frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\mu-a_i)^2 = 0 \end{cases}$$

+约束条件$\sum_{i=1}^{r} a_i = 0$

4. 最大似然估计

$$\widehat{\mu} = \bar{y}$$

$$\widehat{a}_i = \bar{y}_{i\cdot} - \bar{y}, \quad i = 1,\cdots,r$$

$$\widehat{\sigma}_M^2 = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y}_{i\cdot})^2 = \frac{S_e}{n}\text{(误差平方和/n)}$$

由MLE的不变性, 各水平均值$\mu_i = \mu + a_i$的MLE为$\hat{\mu}_i = \bar{y}_i$

$\hat{\sigma}_M^2$ 不是σ^2的无偏估计, 可修偏$\hat{\sigma}^2 = MS_e = S_e/f_e$ (其中$f_e = n - r$).

二. μ_i置信区间

由**定理8.1.2**, $\bar{y}_{i\cdot} \sim N\left(\mu_i, \sigma^2/m\right)$, $S_e/\sigma^2 \sim \chi^2\left(f_e\right)$, 且两者独立, 故

$$\frac{\sqrt{m}\left(\bar{y}_{i\cdot} - \mu_i\right)}{\sqrt{S_e/f_e}} \sim t\left(f_e\right)$$

由此可给出A_i的水平均值μ_i的$1-\alpha$的置信区间为

$$\left[\bar{y}_{i\cdot} - \hat{\sigma} \cdot t_{1-\alpha/2}\left(f_e\right)/\sqrt{m}, \quad \bar{y}_{i\cdot} + \hat{\sigma} \cdot t_{1-\alpha/2}\left(f_e\right)/\sqrt{m}\right]$$

其中$\hat{\sigma}^2 = MS_e$.

例8.1.3 继续例8.1.2, 此处我们给出诸水平均值的估计. 因子A的三个水平均值的估计分别为

$$\begin{aligned}\widehat{\mu}_1 &= 1000 + \frac{194}{8} = 1024.25\\ \widehat{\mu}_2 &= 1000 + \frac{585}{8} = 1073.13\\ \widehat{\mu}_3 &= 1000 + \frac{354}{8} = 1044.25\end{aligned}$$

从点估计来看, 水平2(以槐树粉为主的饲料)是最优的.

误差方差的无偏(和一致)估计为

$$\hat{\sigma}^2 = MS_e = 1343.6171$$

可得出标准差的一致估计(但有偏) $\hat{\sigma} = \sqrt{\hat{\sigma}^2} = 36.66$

由此可以给出诸水平均值的置信区间.

若取$\alpha = 0.05$, 则$t_{1-\alpha/2}(f_e) = t_{0.95}(21) = 2.0796$,
$\hat{\sigma} t_{0.975}(21)/\sqrt{8} = 26.95$, 则三个水平均值的0.95置信区间分别为

$$\mu_1 : 1024.25 \pm 26.95 = [997.30, 1051.20]$$
$$\mu_2 : 1073.125 \pm 26.95 = [1046.18, 1100.08]$$
$$\mu_3 : 1044.25 \pm 26.95 = [1017.30, 1071.20]$$

在单因子试验的数据分析中可得到如下三个结果：

- 因子是否显著;
- 试验的误差方差σ^2的估计;
- 诸水平均值μ_i的点估计与区间估计.

在因子A显著时, 通常只需对较优的水平均值作参数估计,
在因子A不显著场合, 参数估计无需进行.

8.1.6 重复数不等情形 (拓展内容)

在重复数不等场合的方差分析与重复数相等情况极为相似, 只在几处略有差别. 在此仅介绍差别部分:

一. 数据:

假设从第i个水平下的总体获得m_i个试验结果, 记为$y_{i1}, y_{i2}, \ldots, y_{im_i}$, $i = 1, 2, \ldots, r$, 统计模型为:

$$\begin{cases} y_{ij} = \mu_i + \varepsilon_{ij}, i = 1, 2, \ldots, r, j = 1, 2, \ldots, m_i \\ \varepsilon_{ij}\text{相互独立, 且都服从}\varepsilon_{ij} \sim N(0, \sigma^2) \end{cases}$$

二. 总均值：

诸μ_i的加权平均(所有试验结果的均值的平均)

$$\mu = \frac{1}{n}\left(m_1\mu_1 + \ldots + m_r\mu_r\right) = \frac{1}{n}\sum_{i=1}^{r} m_i\mu_i$$

称为**总均值或一般平均**.

三. 效应约束条件：

$$\sum_{i=1}^{r} m_i a_i = 0$$

四. 各平方和的计算： S_A的计算公式略有不同

$$S_A = \sum_{i=1}^{r} m_i\left(\bar{y}_i - \bar{y}\right)^2 = \sum_{i=1}^{r}\frac{T_i^2}{m_i} - \frac{T^2}{n}$$

例8.1.4 某食品公司对一种食品设计了四种新包装. 为考察哪种包装最受顾客欢迎, 选了10个地段繁华程度相似. 规模相近的商店做试验, 其中二种包装各指定两个商店销售, 另二个包装各指定三个商店销售. 在试验期内各店货架排放的位置. 空间都相同, 营业员的促销方法也基本相同, 经过一段时间, 记录其销售量数据, 列于表8.1.6左半边, 其相应的计算结果列于右侧.

表8.1.6 销售量数据及计算表

包装类型	销售量			m_i	T_i	T_i^2/m_i	$\sum_{j=1}^{m_i} y_{ij}^2$
A_1	12	18		2	30	450	468
A_2	14	12	13	3	39	507	509
A_3	19	17	21	3	57	1083	1091
A_4	24	30		2	54	1458	1476
				$n = 10$	$T = 180$	$\sum_{i=1}^{r} \frac{T_i^2}{m_i} = 3498$	$\sum_{i=1}^{r}\sum_{j=1}^{m_i} y_{ij}^2 = 3544$

由此可求得各类偏差平方和如下

$$S_T = 3544 - 3240 = 304, \quad f_T = 10 - 1 = 9$$
$$S_A = 3498 - 3240 = 258, \quad f_A = 4 - 1 = 3$$
$$S_e = 304 - 258 = 46, \quad f_e = 10 - 4 = 6$$

方差分析表如下表所示.

表8.1.7 例8.1.4的方差分析表

来源	平方和	自由度	均方和	F比	p值
因子A	258	3	86	11.22	0.0071
误差e	46	6	7.67		
总和T	304	9			

若取$\alpha = 0.01$, 由于p值为0.0071, 故拒绝原假设, 认为各水平间有显著差异.

由于因子显著, 我们还可以给出诸水平均值的估计. 因子A的四个水平均值的估计分别为

$$\widehat{\mu}_1 = 30/2 = 15, \quad \widehat{\mu}_2 = 39/3 = 13$$
$$\widehat{\mu}_3 = 57/3 = 19, \quad \widehat{\mu}_4 = 54/2 = 27$$

由此可见, 第四种包装方式效果最好. 误差方差的无偏估计为

$$\hat{\sigma}^2 = MS_e = 7.67$$

进一步, 可以给出诸水平均值的置信区间. 在这里要用不同的m_i代替那里相同的m. 此处, $\hat{\sigma} = \sqrt{7.67} = 2.7695$, 若取$\alpha = 0.05$,
则$t_{1-\alpha/2}(f_e) = t_{0.975}(6) = 2.4469$, $\hat{\sigma} t_{0.975}(6) = 6.7767$, 于是效果较好的第四个水平均值的0.95置信区间为

$$\mu_4 : 27 \pm 6.7767/\sqrt{2} = [22.21, \quad 31.79]$$

利用STATA做 Anova分析

分析不同种肥料对苹果重量影响是否不同. weight: 苹果的克重
treatment: 肥料种类

```
help anova
use https://www.stata-press.com/data/r16/apple
```

	treatment	weight
1	1	117.5
2	1	113.8
3	1	104.4
4	2	48.9
5	2	50.4
6	2	58.9
7	3	70.4
8	3	86.9
9	4	87.7
10	4	67.3

利用STATA做 Anova分析

```
anova weight treatment

                   Number of obs =         10     R-squared     =  0.9147
                   Root MSE      =    9.07002     Adj R-squared =  0.8721

    Source | Partial SS          df          MS         F     Prob>F
-----------+----------------------------------------------------------
     Model |   5295.5443           3   1765.1814      21.46  0.0013
           |
 treatment |   5295.5443           3   1765.1814      21.46  0.0013
           |
  Residual |   493.59167           6   82.265278
-----------+----------------------------------------------------------
     Total |    5789.136           9   643.23733
```

Post-Anova analyses 估计各个水平均值

```
. margins treatment

Adjusted predictions                            Number of obs     =         10

Expression   : Linear prediction, predict()

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   treatment |
           1 |      111.9   5.236579    21.37   0.000     99.08655    124.7134
           2 |   52.73333   5.236579    10.07   0.000     39.91989    65.54678
           3 |      78.65   6.413473    12.26   0.000      62.9568     94.3432
           4 |       77.5   6.413473    12.08   0.000      61.8068     93.1932
------------------------------------------------------------------------------
```

Post-Anova analyses 估计各个水平均值: 选取水平1作为基准水平

```
. regress, baselevels

      Source |       SS           df       MS      Number of obs   =        10
-------------+----------------------------------   F(3, 6)         =     21.46
       Model |  5295.54433         3  1765.18144   Prob > F        =    0.0013
    Residual |  493.591667         6  82.2652778   R-squared       =    0.9147
-------------+----------------------------------   Adj R-squared   =    0.8721
       Total |    5789.136         9  643.237333   Root MSE        =      9.07

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   treatment |
           1 |          0  (base)
           2 |  -59.16667   7.405641    -7.99   0.000    -77.28762   -41.04572
           3 |     -33.25   8.279758    -4.02   0.007    -53.50984   -12.99016
           4 |      -34.4   8.279758    -4.15   0.006    -54.65984   -14.14016
             |
       _cons |      111.9   5.236579    21.37   0.000     99.08655    124.7134
------------------------------------------------------------------------------
```

回归分析与Anova关系: 回归结果

```
//try linear regression with a vector of dummy variables
//same results as in  "regress, baselevels"

regress weight i.treatment
```

```
      Source |       SS           df       MS      Number of obs   =        10
-------------+----------------------------------   F(3, 6)         =     21.46
       Model |  5295.54433         3  1765.18144   Prob > F        =    0.0013
    Residual |  493.591667         6  82.2652778   R-squared       =    0.9147
-------------+----------------------------------   Adj R-squared   =    0.8721
       Total |    5789.136         9  643.237333   Root MSE        =      9.07

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   treatment |
           2 |  -59.16667   7.405641    -7.99   0.000    -77.28762   -41.04572
           3 |     -33.25   8.279758    -4.02   0.007    -53.50984   -12.99016
           4 |      -34.4   8.279758    -4.15   0.006    -54.65984   -14.14016
             |
       _cons |      111.9   5.236579    21.37   0.000     99.08655    124.7134
------------------------------------------------------------------------------
```

回归分析与Anova关系：回归结果, 不同的基准期

```
\\set category 2 as base level using b2 to obtain the point estimate and confidence
\\ intervals outputting in _cons
. regress weight b2.treatment

      Source |       SS           df       MS      Number of obs   =        10
-------------+----------------------------------   F(3, 6)         =     21.46
       Model |  5295.54433         3  1765.18144   Prob > F        =    0.0013
    Residual |  493.591667         6  82.2652778   R-squared       =    0.9147
-------------+----------------------------------   Adj R-squared   =    0.8721
       Total |    5789.136         9  643.237333   Root MSE        =      9.07

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   treatment |
           1 |   59.16667   7.405641     7.99   0.000     41.04572    77.28762
           3 |   25.91667   8.279758     3.13   0.020     5.656828    46.17651
           4 |   24.76667   8.279758     2.99   0.024     4.506828    45.02651
             |
       _cons |   52.73333   5.236579    10.07   0.000     39.91989    65.54678
------------------------------------------------------------------------------

.
```

ANOVA和线性回归的关联

Theoretically, ANOVA focuses on test for differences in means. Mathematically, ANOVA is simply a special case of regression analysis where all of the predictor variables are categorical.

References:

https://stats.stackexchange.com/questions/175246/why-is-anova-equivalent-to-linear-regression

https://www.researchgate.net/post/Can_anyone_help_me_to_get_the_core_differences_between_regression_model_and_ANOVA_model2

小结

1. 了解什么是方差分析及本质(多总体均值检验)

2. 理解并熟悉掌握方差分析的基本思想及一般原理(误差分解原理)

3. 掌握单因素方差分析统计模型, 假设检验, 一般步骤及其统计软件实现

4. 在未来的学习中, 思考方差分析和哑变量向量的线性回归(一种对分类变量做回归的实现方法)间的异同

作业8 暨上机实验5

8.1课后习题： 1,5

上机实验5要求同学们对8.1课后习题5的分析提供统计软件分析代码, 并对方差分析结果进行解释.

8.2.1 效应差的置信区间

如果方差分析的结果因子A显著, 则等于说有充分理由认为因子A各水平的效应不全相同, 但这并不是说它们中一定没有相同的. 就指定的一对水平A_i与A_j, 我们可通过求$\mu_i - \mu_j$的区间估计来进行比较.

由

$$\begin{cases} y_{ij} = \mu_i + \varepsilon_{ij}, i = 1, 2, \ldots, r, j = 1, 2, \ldots, m_i \\ \varepsilon_{ij}\text{相互独立, 且都服从}\varepsilon_{ij} \sim N(0, \sigma^2) \end{cases}$$

可以推出

$$\bar{y}_{i\cdot} - \bar{y}_{j\cdot} \sim N\left(\mu_i - \mu_j, \left(\frac{1}{m_i} + \frac{1}{m_i}\right)\sigma^2\right)$$

又由**定理8.1.2**, $S_e/\sigma^2 \sim \chi^2(f_e)$, 且与$\bar{y}_{i\cdot} - \bar{y}_{j\cdot}$相互独立, 故

$$\frac{(\bar{y}_{i\cdot} - \bar{y}_{j\cdot}) - (\mu_i - \mu_j)}{\sqrt{\left(\frac{1}{m_i} + \frac{1}{m_j}\right)\frac{S_e}{f_e}}} \sim t(f_e)$$

由此给出$\mu_i - \mu_j$的置信水平为$1-\alpha$的置信区间为

$$\bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm \sqrt{\left(\frac{1}{m_i} + \frac{1}{m_j}\right)}\hat{\sigma} \cdot t_{1-\frac{a}{2}}(f_e)$$

其中 $\hat{\sigma}^2 = S_e/f_e$ 是σ^2的无偏估计.

这里的置信区间与第六章中的两样本的t区间基本一致, 区别在于这里σ^2的估计使用了全部样本而不仅仅是两个水平A_i, A_j下的观测值.

例8.2.1 继续**例8.1.2**, $\hat{\sigma} = \sqrt{1343.61} = 36.66, \quad f_e = 21,$
取$\alpha = 0.05$, 则$t_{1-\alpha/2}(f_e) = t_{0.975}(21) = 2.0796,$
$\sqrt{\frac{1}{8} + \frac{1}{8}}\hat{\sigma} t_{0.975}(21) = 38.12$, 于是可算出各个置信区间为

$$\begin{aligned}
\mu_1 - \mu_2: &\quad -48.88 \pm 38.12 = [-87, -10.76] \\
\mu_1 - \mu_3: &\quad -20 \pm 38.12 = [-58.12, 18.12] \\
\mu_2 - \mu_3: &\quad 28.8750 \pm 38.12 = [-9.24, 67]
\end{aligned}$$

可见第一个区间在0的左边, 所以我们可以概率95%判断认为$\mu_1 < \mu_2$

其它二个区间包含0点, 虽然从点估计角度看水平均值估计有差别, 但这种差异在0.05水平上是不显著的.

8.2.2 多重比较问题

多重比较方法(Multiple comparison procedures):通过对总体均值之间的配对比较进一步检验到底哪些均值之间存在差异

- 在方差分析中, 如果经过F检验拒绝原假设, 表明因子A是显著的, 即r个水平对应的水平均值不全相等, 此时, 我们还需要进一步确认哪些水平均值间是确有差异的, 哪些水平均值间无显著差异.

在此仅给出重复数不等场合的S法(Scheffe Test)的结论(重复数相等也成立).

8.2.4 重复数不等场合的S法(Scheffe Test)

1. 提出假设

$$\boldsymbol{H_0}:\boldsymbol{\mu_i}=\boldsymbol{\mu_j}\leftrightarrow\boldsymbol{H_1}:\boldsymbol{\mu_i}\neq\boldsymbol{\mu_j}$$

2. 计算统计量

$$|\bar{y}_i-\bar{y}_j|$$

3. 拒绝域

$$W=\{|\bar{y}_{i\cdot}-\bar{y}_{j\cdot}|\geq c_{ij}\}$$

其中

$$c_{ij}=\sqrt{(r-1)F_{1-\alpha}\left(r-1,f_e\right)\left(\frac{1}{m_i}+\frac{1}{m_j}\right)\hat{\sigma}^2}$$

例8.2.3 在例8.1.4中, 我们指出包装方式对食品销量有明显的影响, 此处$r=4, f_e=6, \hat{\sigma}^2=7.67$, 若取$\alpha=0.05$, 则$F_{0.95}(3,6)=4.76$. 注意到$m_1=m_4=2, m_2=m_3=3$, 故

$$c_{12}=c_{13}=c_{24}=c_{34}=\sqrt{3\times 4.76\times(1/2+1/3)\times 7.67}=9.6$$
$$c_{14}=\sqrt{3\times 4.76\times(1/2+1/2)\times 7.67}=10.5$$
$$c_{23}=\sqrt{3\times 4.76\times(1/3+1/3)\times 7.67}=8.5$$

由于

$$|\bar{y}_{1\cdot} - \bar{y}_{2\cdot}| = 2 < c_{12}, |\bar{y}_{1\cdot} - \bar{y}_{3\cdot}| = 4 < c_{13}, |\bar{y}_{1\cdot} - \bar{y}_{4\cdot}| = 12 > c_{14}$$
$$|\bar{y}_{2\cdot} - \bar{y}_{3\cdot}| = 6 < c_{23}, |\bar{y}_{2\cdot} - \bar{y}_{4\cdot}| = 14 > c_{24}, |\bar{y}_{3\cdot} - \bar{y}_{4\cdot}| = 8 < c_{34}$$

这说明A_1, A_2, A_3间无显著差异, A_1, A_2与A_4有显著差异, 但A_4与A_3 的差异却尚未达到显著水平. 综合上述, 包装A_4销售量最佳.

STATA multiple comparison

```
help oneway
use https://www.stata-press.com/data/r16/apple
anova weight treatment
oneway weight treatment, noanova scheffe
```

```
            Comparison of Average weight in grams by Fertilizer
                                    (Scheffe)
Row Mean-|
Col Mean |          1          2          3
---------+------------------------------------
       2 |   -59.1667
         |      0.001
         |
       3 |     -33.25    25.9167
         |      0.039      0.101
         |
       4 |      -34.4    24.7667      -1.15
         |      0.034      0.118      0.999
```

```
// first row is the difference of means
// second row is significance level of the difference
```

§8.3 方差齐性检验 (Homogeneity of variance test) (略)

在进行方差分析时要求r个方差相等, 这称为**方差齐性**. 理论研究表明, 当正态性假定不满足时对F检验影响较小, 即F检验对正态性的偏离具有一定的稳健性, 而F检验对方差齐性的偏离较为敏感. 所以r个方差的齐性检验就显得十分必要.

所谓方差齐性检验是对如下一对假设作出检验:

$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots \sigma_r^2 \quad H_1 : \sigma_i^2$不都相等

最常用的检验: **Bartlett检验**, 可用于样本量相等或不等的场合, 但是每个样本量不得低于5.

STATA homogeneity of variance 方差齐次性检验

```
help oneway
use https://www.stata-press.com/data/r16/apple
anova weight treatment

oneway weight treatment

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      5295.54433      3   1765.18144     21.46     0.0013
 Within groups      493.591667      6   82.2652778
------------------------------------------------------------------------
    Total             5789.136      9   643.237333

Bartlett's test for equal variances:  chi2(3) =   1.3900  Prob>chi2 = 0.708

\\ can not reject null hypothesis that variances are equal 不能拒绝方差齐次的原假设
```