统计方法与机器学习 理论作业1 参考答案

1

由条件可知 $a=4, n=6, SS_T=10, SS_E=2.5, SS_A=7.5$

故其方差分析表为：

来源	平方和 SS	自由度 df	均方和 MS	F 值
因子 A	7.5	3	2.5	20
误差 E	2.5	20	0.125	
总和	10	23		

2

（1） 令两组数据的总均值

$$\bar{z}=\frac{m\cdot(\bar{x}+\bar{y})}{2m}=\frac{\bar{x}+\bar{y}}{2} \tag{1}$$

取检验统计量 $f=\frac{SS_A}{SS_E/(2m-2)}$，其中

$$\begin{aligned} SS_A&=m(\bar{x}-\bar{z})^2+m(\bar{y}-\bar{z})^2\\ SS_E&=\sum_{i=1}^{m}(x_i-\bar{x})^2+\sum_{i=1}^{m}(y_i-\bar{y})^2 \end{aligned} \tag{2}$$

则在显著性水平 α 下，若 $f\geq F_{1-\alpha}(1,2m-2)$，则拒绝原假设 H_0，接受备择假设 H_1，反之则接受原假设 H_0。

（2）

> **Lemma:** 设 $0<\alpha<1, n\in\mathbb{N}^+$，$t(n), F(m,n)$ 分别为 t 分布和 F 分布，则 $t^2_{\frac{\alpha+1}{2}}(n)=F_\alpha(1,n)$
>
> **Proof:** 设 $t\sim t(n), f\sim F(1,n)$，则有 $f=t^2$
>
> 于是对任意 $x\in\mathbb{R}$，有 $P(-x<t<x)=P(f<x^2)$
>
> 现取 $x=x_0$ 使得 $P(f<x_0^2)=\alpha$
>
> 因此
>
> $$\begin{aligned} \alpha&=P(-x_0<t<x_0)\\ &=1-2(1-P(t<x_0))\\ &=2P(t<x_0)-1 \end{aligned} \tag{3}$$
>
> 于是有 $P(t<x_0)=\frac{\alpha+1}{2}$
>
> 也即 $t^2_{\frac{\alpha+1}{2}}(n)=F_\alpha(1,n)$

对于本题，若使用One-Way ANOVA，则其检验统计量为

$$
\begin{aligned}
f &= \frac{SS_A}{SS_E/(2m-2)} \\
&= \frac{m\left(\bar{x}-\frac{\bar{x}+\bar{y}}{2}\right)^2+m\left(\bar{y}-\frac{\bar{x}+\bar{y}}{2}\right)^2}{\frac{1}{2m-2}\cdot(m-1)(s_x^2+s_y^2)} \\
&= \frac{m(\bar{x}-\bar{y})^2}{s_x^2+s_y^2}
\end{aligned} \tag{4}
$$

而若使用两样本独立 t 检验，则检验统计量为

$$
t = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{1}{2}s_x^2+\frac{1}{2}s_y^2}\cdot\sqrt{\frac{2}{m}}} = \frac{\sqrt{m}(\bar{x}-\bar{y})}{\sqrt{s_x^2+s_y^2}} \tag{5}
$$

易见 $f=t^2$

又由引理可知，对于显著性水平 α，若 $|t|\ge t_{1-\frac{\alpha}{2}}(2m-2)$，则 $f\ge F_{1-\alpha}(1,2m-2)$，也即表明其拒绝域等价。

因此对于该问题，One-Way ANOVA模型与二样本独立 t 检验等价。

3

(1) 设 μ_i 为第 i 个水平下的总体均值，ε_{ij} 为第 i 个水平下第 j 个观测值的随机误差，则其One-Way ANOVA模型为：

$$
y_{ij}=\mu_i+\varepsilon_{ij},\begin{cases} i=1,2,\cdots,a \\ j=1,2,\cdots,m_i \end{cases} \tag{6}
$$

其中 $\varepsilon_i \overset{i.i.d}{\sim} N(0,\sigma^2)$

(2)

$$
\begin{aligned}
H_0 &: \mu_1=\mu_2=\cdots=\mu_a \\
H_1 &: \exists i\ne j\in\{1,2,\cdots,a\},\textbf{s.t}\ \ \mu_i\ne\mu_j
\end{aligned} \tag{7}
$$

(3)

$$
f=\frac{SS_A/(a-1)}{SS_E/\left(\sum\limits_{i=1}^{a}m_i-a\right)} \tag{8}
$$

其中

$$
\begin{aligned}
SS_A &= \sum_{i=1}^{a}m_i(\bar{y}_i-\bar{y})^2 \\
SS_E &= \sum_{i=1}^{a}\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_i)^2
\end{aligned} \tag{9}
$$

(4)

来源	平方和 SS	自由度 df	均方和 MS	F 值
因子 A	$\sum_{i=1}^{a}m_i(\bar{y}_i-\bar{y})^2$	$a-1$	$\frac{1}{a-1}\sum_{i=1}^{a}m_i(\bar{y}_i-\bar{y})^2$	$\frac{(\sum_{i=1}^{a}m_i-a)\sum_{i=1}^{a}m_i(\bar{y}_i-\bar{y})^2}{(a-1)\sum_{i=1}^{a}\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_i)^2}$
误差 E	$\sum_{i=1}^{a}\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_i)^2$	$\sum_{i=1}^{a}m_i-a$	$\frac{1}{\sum_{i=1}^{a}m_i-a}\sum_{i=1}^{a}\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_i)^2$	
总和	$\sum_{i=1}^{a}\sum_{j=1}^{m_i}(y_{ij}-\bar{y})^2$	$\sum_{i=1}^{a}m_i-1$		

(5) 接下来我们来推导检验统计量 f 的分布。

首先我们对 SS_E 进行变形以导出其分布：

$$
\begin{aligned}
SS_E &= \sum_{i=1}^{a}\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_i)^2 \\
&= \sum_{i=1}^{a}\sum_{j=1}^{m_i}\left((\mu+\alpha_i+\varepsilon_{ij})-\frac{1}{m_i}\sum_{j=1}^{m_i}(\mu+\alpha_i+\varepsilon_{ij})\right)^2 \\
&= \sum_{i=1}^{a}\sum_{j=1}^{m_i}(\varepsilon_{ij}-\bar{\varepsilon}_i)^2
\end{aligned} \tag{10}
$$

由于 $\varepsilon_{ij} \overset{i.i.d}{\sim} N(0,\sigma^2)$

因此

$$
\frac{1}{\sigma^2}\sum_{j=1}^{m_i}(\varepsilon_{ij}-\bar{\varepsilon}_i)^2 \sim \mathcal{X}^2(m_i-1) \tag{11}
$$

于是由卡方分布的可加性可知

$$
\frac{SS_E}{\sigma^2} = \sum_{i=1}^{a}\left(\frac{1}{\sigma^2}\sum_{j=1}^{m_i}(\varepsilon_{ij}-\bar{\varepsilon}_i)^2\right) \sim \mathcal{X}^2(n-a) \tag{12}
$$

接下来我们来推导关于 SS_A 的分布。

当 H_0 成立时，$\alpha_i = 0$

于是

$$
\begin{aligned}
SS_A &= \sum_{i=1}^{a} m_i(\bar{\varepsilon}_i-\bar{\varepsilon})^2 \\
&= \sum_{i=1}^{a} m_i\bar{\varepsilon}_i^2 - n\bar{\varepsilon}^2
\end{aligned} \tag{13}
$$

令 $x_i = \frac{\sqrt{m_i}\bar{\varepsilon}_i}{\sigma}$，易见 $x_i \overset{i.i.d}{\sim} N(0,1)$

故

$$
\frac{SS_A}{\sigma^2} = \sum_{i=1}^{a} x_i^2 - \left(\sum_{i=1}^{a}\sqrt{\frac{m_i}{n}}x_i\right)^2 \tag{14}
$$

由于每组中的样本个数不同，我们不再能够使用一般的代数变形将其变为与某个分布相关的形式，故而我们需要对其进行整体变换使其容易化简。

由 $x_i \overset{i.i.d}{\sim} N(0,1)$ 可知

$$
p(x_1,\cdots,x_a) = (2\pi)^{-\frac{n}{2}}\exp\{-\frac{1}{2}\sum_{i=1}^{a}x_i^2\} \tag{15}
$$

令 $\mathbf{X} = (x_1,\cdots,x_a)^T$

可以证明存在正交矩阵 $\mathbf{A}$，使得 $A_{1i} = \sqrt{\frac{m_i}{n}}$

令 $\mathbf{Y} = \mathbf{AX}$

于是

$$
\sum_{i=1}^{n} y_i^2 = Y^TY = X^TA^TAX = X^TX = \sum_{i=1}^{n} x_i^2 \tag{16}
$$

且

$$p(y_1,\cdots,y_a)=(2\pi)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{a}y_i^2\right)\right\} \tag{17}$$

因此有 $y_i \overset{i.i.d}{\sim} N(0,1)$

于是

$$\begin{aligned}\frac{SS_A}{\sigma^2}&=\sum_{i=1}^{a}x_i^2-\left(\sum_{i=1}^{a}\sqrt{\frac{m_i}{n}}x_i\right)^2\\&=\sum_{i=1}^{a}y_i^2-y_1^2\\&=\sum_{i=2}^{a}y_i^2\sim\mathcal{X}^2(a-1)\end{aligned} \tag{18}$$

且由于 $\sum_{j=1}^{m_i}(\varepsilon_{ij}-\bar{\varepsilon}_i)^2$ 与 $\bar{\varepsilon}_i$ 独立

故 SS_A 与 SS_E 独立

由此可知检验统计量

$$f\sim F(a-1,n-a) \tag{19}$$

4

由于

$$\begin{aligned}SS_T&=\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{m}(y_{ijk}-\bar{y}_{...})^2\\&=\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{m}\left[(\bar{y}_{i..}-\bar{y}_{...})+(\bar{y}_{.j.}-\bar{y}_{...})+(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...})+(\bar{y}_{ijk}-\bar{y}_{ij.})\right]^2\end{aligned} \tag{20}$$

故要证 $SS_T=SS_A+SS_B+SS_{AB}+SS_E$，即要证平方和展开式中六个交叉项为 0。

下面依次证明之：

$$\begin{aligned}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{m}\left(\bar{y}_{i..}-\bar{y}_{...}\right)\left(\bar{y}_{.j.}-\bar{y}_{...}\right)&=m\sum_{i=1}^{a}(\bar{y}_{i..}-\bar{y}_{...})\sum_{j=1}^{b}(\bar{y}_{.j.}-\bar{y}_{...})\\&=m\sum_{i=1}^{a}(\bar{y}_{i..}-\bar{y}_{...})(b\bar{y}_{...}-b\bar{y}_{...})\\&=0\end{aligned} \tag{21}$$

$$\begin{aligned}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{m}\left(\bar{y}_{i..}-\bar{y}_{...}\right)\left(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...}\right)&=m\sum_{i=1}^{a}(\bar{y}_{i..}-\bar{y}_{...})\sum_{j=1}^{b}\left(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...}\right)\\&=m\sum_{i=1}^{a}(\bar{y}_{i..}-\bar{y}_{...})(b\bar{y}_{i..}-b\bar{y}_{i..}-b\bar{y}_{...}+b\bar{y}_{...})\\&=0\end{aligned} \tag{22}$$

$$
\begin{aligned}
\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{m}\left(\bar{y}_{i..}-\bar{y}_{...}\right)\left(\bar{y}_{ijk}-\bar{y}_{ij.}\right) &= \sum_{i=1}^{a}(\bar{y}_{i..}-\bar{y}_{...})\sum_{j=1}^{b}\sum_{k=1}^{m}\left(\bar{y}_{ijk}-\bar{y}_{ij.}\right) \\
&= \sum_{i=1}^{a}(\bar{y}_{i..}-\bar{y}_{...})\sum_{j=1}^{b}\left(m\bar{y}_{ij.}-m\bar{y}_{ij.}\right) \\
&= 0
\end{aligned} \tag{23}
$$

$$
\begin{aligned}
\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{m}\left(\bar{y}_{.j.}-\bar{y}_{...}\right)\left(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...}\right) &= m\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{.j.}-\bar{y}_{...})\left(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...}\right) \\
&= m\sum_{j=1}^{b}(\bar{y}_{.j.}-\bar{y}_{...})\sum_{i=1}^{a}\left(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...}\right) \\
&= m\sum_{j=1}^{b}(\bar{y}_{.j.}-\bar{y}_{...})\left(a\bar{y}_{.j.}-a\bar{y}_{...}-a\bar{y}_{.j.}+a\bar{y}_{...}\right) \\
&= 0
\end{aligned} \tag{24}
$$

$$
\begin{aligned}
\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{m}\left(\bar{y}_{.j.}-\bar{y}_{...}\right)\left(\bar{y}_{ijk}-\bar{y}_{ij.}\right) &= \sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{.j.}-\bar{y}_{...})\sum_{k=1}^{m}\left(\bar{y}_{ijk}-\bar{y}_{ij.}\right) \\
&= \sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{.j.}-\bar{y}_{...})\left(m\bar{y}_{ij.}-m\bar{y}_{ij.}\right) \\
&= 0
\end{aligned} \tag{25}
$$

$$
\begin{aligned}
\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{m}\left(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...}\right)\left(\bar{y}_{ijk}-\bar{y}_{ij.}\right) &= \sum_{i=1}^{a}\sum_{j=1}^{b}\left(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...}\right)\sum_{k=1}^{m}\left(\bar{y}_{ijk}-\bar{y}_{ij.}\right) \\
&= \sum_{i=1}^{a}\sum_{j=1}^{b}\left(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...}\right)\left(m\bar{y}_{ij.}-m\bar{y}_{ij.}\right) \\
&= 0
\end{aligned} \tag{26}
$$

由此可得 $SS_T = SS_A + SS_B + SS_{AB} + SS_E$

5

> **Lemma:** 设 $\{X_i\}(i=1,2,\cdots,n)$ 为一独立随机变量序列，且 $E(X_i)=\mu_i, Var(X_i)=\sigma^2$，则
>
> $$E\left[(X_i-\bar{X})^2\right] = (\mu_i-\bar{\mu})^2 + \frac{n-1}{n}\sigma^2 \tag{27}$$
>
> 其中
>
> $$\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mu_i \tag{28}$$
>
> **Proof:** 由 $\{X_i\}$ 相互独立可知，当 $i \neq j$ 时，$Cov(X_i, X_j) = 0$
>
> 故
>
> $$Cov(X_i,\bar{X}) = \frac{1}{n}Cov(X_i,X_i) = \frac{1}{n}Var(X_i) = \frac{1}{n}\sigma^2 \tag{29}$$
>
> 于是有

> $$
> \begin{aligned}
> Var(X_i-\bar{X}) &= Var(X_i)+Var(\bar{X})-2Cov(X_i,\bar{X}) \\
> &= \sigma^2+\frac{1}{n}\sigma^2-2\frac{1}{n}\sigma^2 \\
> &= \frac{n-1}{n}\sigma^2
> \end{aligned} \tag{30}
> $$
>
> 因此
>
> $$
> \begin{aligned}
> E\left[(X_i-\bar{X})^2\right] &= \Big(E(X_i-\bar{X})\Big)^2+Var(X_i-\bar{X}) \\
> &= (\mu_i-\bar{\mu})^2+\frac{n-1}{n}\sigma^2
> \end{aligned} \tag{31}
> $$

设 $y_{ijk}=\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}+\varepsilon_{ijk}$，其中 $\varepsilon_{ijk}\overset{i.i.d}{\sim}N(0,\sigma^2)$

则

$$
\begin{aligned}
E(y_{ijk}) &= \mu+\alpha_i+\beta_j+(\alpha\beta)_{ij} \\
Var(y_{ijk}) &= \sigma^2
\end{aligned} \tag{32}
$$

于是

$$
\begin{aligned}
E(\bar{y}_{i..}) &= \frac{1}{bm}\sum_{j=1}^{b}\sum_{k=1}^{m}E(y_{ijk}) \\
&= \frac{1}{b}\sum_{j=1}^{b}(\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}) \\
&= \frac{1}{b}(b\mu+b\alpha_i+0+0) \\
&= \mu+\alpha_i
\end{aligned} \tag{33}
$$

$$
\begin{aligned}
Var(\bar{y}_{i..}) &= \frac{1}{b^2m^2}\sum_{j=1}^{b}\sum_{k=1}^{m}Var(y_{ijk}) \\
&= \frac{1}{bm}\sigma^2
\end{aligned} \tag{34}
$$

故由引理可知

$$
\begin{aligned}
E(MS_A) &= E\left(\frac{SS_A}{a-1}\right) \\
&= \frac{1}{a-1}E\left(bm\sum_{i=1}^{a}(\bar{y}_{i..}-\bar{y}_{...})^2\right) \\
&= \frac{bm}{a-1}\sum_{i=1}^{a}E\left[(\bar{y}_{i..}-\bar{y}_{...})^2\right] \\
&= \frac{bm}{a-1}\sum_{i=1}^{a}\left[(\mu+\alpha_i-\mu)^2+\frac{a-1}{a}\cdot\frac{1}{bm}\sigma^2\right] \\
&= \frac{bm}{a-1}\sum_{i=1}^{a}\alpha_i^2+\sigma^2
\end{aligned} \tag{35}
$$

同理，由

$$
\begin{aligned}
E(\bar{y}_{.j.}) &= \frac{1}{am}\sum_{i=1}^{a}\sum_{k=1}^{m}E(y_{ijk}) \\
&= \mu + \beta_j \\
Var(\bar{y}_{.j.}) &= \frac{1}{a^2m^2}\sum_{i=1}^{a}\sum_{k=1}^{m}Var(y_{ijk}) \\
&= \frac{1}{am}\sigma^2
\end{aligned} \tag{36}
$$

可知

$$
\begin{aligned}
E(MS_B) &= E\left(\frac{SS_B}{b-1}\right) \\
&= \frac{1}{b-1}E\left(am\sum_{j=1}^{b}(\bar{y}_{.j.}-\bar{y}_{...})^2\right) \\
&= \frac{am}{b-1}\sum_{j=1}^{b}E\left[(\bar{y}_{i..}-\bar{y}_{...})^2\right] \\
&= \frac{am}{b-1}\sum_{j=1}^{b}\left[(\mu+\beta_j-\mu)^2+\frac{b-1}{b}\cdot\frac{1}{am}\sigma^2\right] \\
&= \frac{am}{b-1}\sum_{j=1}^{b}\beta_j^2+\sigma^2
\end{aligned} \tag{37}
$$

$$
\begin{aligned}
E(MS_E) &= E\left(\frac{SS_E}{ab(m-1)}\right) \\
&= \frac{1}{ab(m-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{m}E\left[(y_{ijk}-\bar{y}_{ij.})^2\right] \\
&= \frac{1}{ab(m-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{m}\left[(\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}-\mu-\alpha_i-\beta_j-(\alpha\beta)_{ij})^2+\frac{m-1}{m}\sigma^2\right] \\
&= \sigma^2
\end{aligned} \tag{38}
$$

又

$$
\begin{aligned}
E(SS_T) &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{m}E\left[(y_{ijk}-\bar{y}_{...})^2\right] \\
&= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{m}\left[(\alpha_i+\beta_j+(\alpha\beta)_{ij})^2+\frac{abm-1}{abm}\sigma^2\right] \\
&= bm\sum_{i=1}^{a}\alpha_i^2+am\sum_{j=1}^{b}\beta_i^2+m\sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha\beta)_{ij}^2+(abm-1)\sigma^2
\end{aligned} \tag{39}
$$

故

$$
\begin{aligned}
E(SS_{AB}) &= E(SS_T-SS_A-SS_B-SS_E) \\
&= bm\sum_{i=1}^{a}\alpha_i^2+am\sum_{j=1}^{b}\beta_i^2+m\sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha\beta)_{ij}^2+(abm-1)\sigma^2 \\
&\quad -bm\sum_{i=1}^{a}\alpha_i^2-(a-1)\sigma^2-am\sum_{j=1}^{b}\beta_j^2-(b-1)\sigma^2-ab(m-1)\sigma^2 \\
&= m\sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha\beta)_{ij}^2+(a-1)(b-1)\sigma^2
\end{aligned} \tag{40}
$$

因此

$$
\begin{aligned}
E(MS_{AB}) &= E\left(\frac{SS_{AB}}{(a-1)(b-1)}\right) \\
&= \frac{1}{(a-1)(b-1)}\left[m\sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha\beta)_{ij}^{2}+(a-1)(b-1)\sigma^{2}\right] \\
&= \frac{m}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha\beta)_{ij}^{2}+\sigma^{2}
\end{aligned}
\tag{41}
$$