第三次数据分析作业

张申铎 2176112379

目录

1 Problem 1

1. (P125 3.1 题)对于**两因素等重复试验下的方差分析模型**

$$\begin{cases} y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \\ \quad i = 1,2,...,a, \quad j = 1,2,...,b, \quad k = 1,2,...,c \\ \varepsilon_{ijk} \sim N(0,\sigma^2), \text{ 且诸}\varepsilon_{ijk}\text{相互独立} \\ \sum_{i=1}^{a} \alpha_i = 0, \sum_{j=1}^{b} \beta_j = 0, \sum_{i=1}^{a} \gamma_{ij} = 0, \sum_{j=1}^{b} \gamma_{ij} = 0, \end{cases}$$

请证明最后一行的四个等式成立。

$$\sum_{i=1}^{a}\alpha_i=\sum_{i=1}^{a}(\mu_i-\mu)=\frac{1}{b}\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij}-a\frac{1}{ab}\sum_{i=1}^{a}\sum_{i=1}^{b}\mu_{ij}=0$$
$$\sum_{j=1}^{b}\beta_j=\sum_{j=1}^{b}(\mu_j-\mu)=\frac{1}{a}\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij}-b\frac{1}{ab}\sum_{j=1}^{b}\sum_{i=1}^{a}\mu_{ij}=0$$
$$\sum_{i=1}^{a}\gamma_{ij}=\sum_{i=1}^{a}((\mu_{ij}-\mu)-(\alpha_i+\beta_j))=\sum_{i=1}^{a}(\mu_{ij}-\mu)-a\beta_j=a\beta_j-a\beta_j=0$$
$$\sum_{j=1}^{b}\gamma_{ij}=\sum_{j=1}^{b}((\mu_{ij}-\mu)-(\alpha_i+\beta_j))=\sum_{j=1}^{b}(\mu_{ij}-\mu)-b\alpha_i=b\alpha_i-b\alpha_i=0$$

2 Problem 2

2. (P125 3.2 题)对上述两因素等重复试验下的方差分析模型，试推导

$$\begin{cases}E\left(SS_A\right)=(a-1)\sigma^2+bc\sum_{i=1}^{a}\alpha_i^2\\ E\left(SS_B\right)=(b-1)\sigma^2+ac\sum_{j=1}^{b}\beta_j^2\\ E\left(SS_{AB}\right)=(a-1)(b-1)\sigma^2+c\sum_{i=1}^{a}\sum_{j=1}^{b}\gamma_{ij}^2\\ E\left(SS_E\right)=ab(c-1)\sigma^2\end{cases}.$$

我们有

$$SS_A=bc\sum_{i=1}^{a}(\bar{y}_{i..}-\bar{y})^2$$
$$SS_B=ac\sum_{j=1}^{b}(\bar{y}_{.j.}-\bar{y})^2$$
$$SS_{AB}=c\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y})^2$$
$$SS_E=\sum_{i,j,k}(y_{ijk}-\bar{y}_{ij})^2$$

对其分别取期望, 有

$$\begin{aligned}
\mathbb{E}SS_A &= bc\sum_{i=1}^{a}\mathbb{E}(\alpha_i+\bar{\epsilon}_{i..}-\bar{\epsilon}_{ij})^2 = bc\sum_{i=1}^{a}\mathbb{E}(\alpha_i^2+\bar{\epsilon}_{i..}^2+\bar{\epsilon}_{ij}^2+2\alpha_i\bar{\epsilon}_{i..}-2\alpha_i\bar{\epsilon}_{ij}-2\bar{\epsilon}_{i..}\bar{\epsilon}) \\
&= bc\sum_{i=1}^{a}(\alpha_i^2+\frac{\sigma^2}{bc}+\frac{\sigma^2}{abc}-2\mathbb{E}\bar{\epsilon}_{i..}\frac{1}{a}\sum_{i=1}^{a}\bar{\epsilon}_{i..}) = bc\sum_{i=1}^{a}(\alpha_i^2+\frac{\sigma^2}{bc}+\frac{\sigma^2}{abc}-2\frac{\sigma^2}{abc}) \\
&= bc\sum_{i=1}^{a}\alpha_i^2+(a-1)\sigma^2 \\
\mathbb{E}SS_B &= ac\sum_{j=1}^{b}\mathbb{E}(\beta_j+\bar{\epsilon}_{.j.}-\bar{\epsilon}_{ij})^2 = ac\sum_{j=1}^{b}\mathbb{E}(\beta_j^2+\bar{\epsilon}_{.j.}^2+\bar{\epsilon}_{ij}^2+2\beta_j\bar{\epsilon}_{.j.}-2\beta_j\bar{\epsilon}_{ij}-2\bar{\epsilon}_{.j.}\bar{\epsilon}) \\
&= ac\sum_{j=1}^{b}(\beta_j^2+\frac{\sigma^2}{ac}+\frac{\sigma^2}{abc}-2\mathbb{E}\bar{\epsilon}_{.j.}\frac{1}{b}\sum_{j=1}^{b}\bar{\epsilon}_{.j.}) = ac\sum_{j=1}^{b}(\beta_j^2+\frac{\sigma^2}{ac}+\frac{\sigma^2}{abc}-2\frac{\sigma^2}{abc}) \\
&= ac\sum_{j=1}^{b}\beta_j^2+(b-1)\sigma^2 \\
\mathbb{E}SS_{AB} &= \mathbb{E}c\sum_{i=i}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y})^2 = \mathbb{E}c\sum_{i=i}^{a}\sum_{j=1}^{b}(\bar{\epsilon}_{ij}-\bar{\epsilon}_{i..}-\bar{\epsilon}.j.+\bar{\epsilon}+\gamma_{ij})^2
\end{aligned}$$

交叉项消失因为残差的均值为 0 并且 γ_{ij} 对任何一个角标加和一次为 0

$$\begin{aligned}
&= c\sum_{i=i}^{a}\sum_{j=1}^{b}(\mathbb{E}(\bar{\epsilon}_{ij}-\bar{\epsilon}_{i..}-\bar{\epsilon}.j.+\bar{\epsilon})^2+c\mathbb{E}\gamma_{ij}^2) \\
&= c\sum_{i=i}^{a}\sum_{j=1}^{b}(\frac{\sigma^2}{c}-\frac{\sigma^2}{bc}-\frac{\sigma^2}{ac}+\frac{\sigma^2}{abc})+c\gamma_{ij}^2) \\
&= (a-1)(b-1)\sigma^2+c\sum_{i=i}^{a}\sum_{j=1}^{b}\gamma_{ij}^2 \\
\mathbb{E}SS_E &= \sum_{i=1}^{a}\sum_{j=1}^{b}\mathbb{E}\sum_{k=1}^{c}(y_{ijk}-\bar{y}_{ij.})^2 \\
&= ab(c-1)\sigma^2
\end{aligned}$$

3 Problem 3

3. (P126 3.5 题)为了了解生产某种电子设备的公司在过去三年中的科研经费投入(分为低、中、高三档)对当年生产能力提高的影响，调查了共计 27 家生产该设备的公司，对当年生产能力较之三年前的提高量作评估，得数据如下表所示：

表 1. 不同科研经费投入下生产能力的提高量.

投入	生产能力提高量											
低	7.6	8.2	6.8	5.8	6.9	6.6	6.3	7.7	6.0			
中	6.7	8.1	9.4	8.6	7.8	7.7	8.9	7.9	8.3	8.7	7.1	8.4
高	8.5	9.7	10.1	7.8	9.6	9.5						

假设生产能力提高量服从方差分析模型.

(1) 建立方差分析表，在显著性水平 $\alpha = 0.05$ 下检验过去三年科研经费投入的不同是否对当年生产力的提高有显著影响；

(2) 分别以 μ_L, μ_M, μ_H 记在过去三年科研经费投入为低、中、高情况下当年生产能力提高量的均值，分别给出 μ_L, μ_M, μ_H 的置信度为 95%的置信区间以及差值

的置信度不小于 95%的 Bonferroni 同时置信区间。

是否过去三年科研经费投入越高，当年的生产能力的改善越明显？

3.1 (1)

我们建立方差分析表如下.

```
Data<-data.frame(X=problem3,a=factor(rep(1:3,c(9,12,6))))
prob3 <- aov(X~a,data = Data)
summary(prob3)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## a            2  20.12   10.06   15.72 4.33e-05 ***
## Residuals   24  15.36    0.64
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

可以看到在显著性水平 0.05 下, 我们可以认为有显著的影响.

3.2 (2)

```
alpha <- 0.05
mu <- c(mean(Data$X[1:9]),mean(Data$X[10:21]),mean(Data$X[22:27]))
Deviation<-qt(1-alpha/2,prob3$df.residual)*sqrt(
  (sum(prob3$residuals^2)/prob3$df.residual)/c(9,12,6))
muInterval <- list(left=(mu-Deviation),right=mu+Deviation)
muInterval
```

```
## $left
## [1] 6.327365 7.656662 8.525885
##
## $right
## [1] 7.428191 8.610005 9.874115
```

故 μ_l, μ_m, μ_h 三个 0.95 置信区间分别为 $(6.327, 7.428), (7.656, 8.610), (8.526, 9.874)$.

下面计算其 Bonferroni 同时置信区间.

```
muDiff <-c(mu[1]-mu[2],mu[1]-mu[3],mu[2]-mu[3])
diffDeviation<-qt(1-alpha/(2*3),prob3$df.residual)*sqrt(
  2*(sum(prob3$residuals^2)/prob3$df.residual)/c(9,12,6))
muDiffInterval <- list(
  left=(muDiff-diffDeviation),right=(muDiff+diffDeviation))
muDiffInterval
```

```
## $left
## [1] -2.226207 -3.162831 -2.255467
##
## $right
## [1] -0.2849045 -1.4816138  0.1221332
```

其差值 $\mu_L - \mu_M, \mu_L - \mu_H, \mu_M - \mu_h$ 的 95%Bonferroni 同时置信区间为

$$(-2.226, -0.285), (-3.163, -1.482), (-2.255, 0.122)$$

.

从此可以看出答案是肯定的. 过去三年科研经费投入越高, 当年生产能力改善越明显.

4 Problem 4

4. (P127 3.7 题)为研制一种治疗枯草热病的药物，将两种成分(A 和 B)各按三种不同剂量(低、中、高)混合，将 36 位自愿受试患者随机分成 9 组，每组 4 人服用不同各种剂量混合下的药物，记录其病情缓解的时间(单位: h)，如下表所示：

表 2. 不同剂量组合下病情缓解的时间.

(1)计算每个水平组合 上的均值 的估计值 (i,j=1,2,3), 并作图(参考书中图 3.2)判断 A 与 B 的交互效应是否显著;

(2)假设所给数据服从方差分析模型，建立方差分析表，A 与 B 的交互效应在显著性水平 下是否显著?

(3)若 A 与 B 的交互效应显著，分别就 A 的各水平 (i=1,2,3)，给出 B 的各水平 上的均值 的置信度不小于 95%的置信区间及两两之差的置信度不少于 95%的 Bonferroni 同时置信区间。固定 B 的各水平 ，关于因素 A 作类似分析，你能否选出最佳的水平组合?

4.1 (1)(2)

```
Data2 <- data.frame(x=problem4,a=gl(3,12,36),b=gl(3,4,36))
prob4 <- aov(x~a+b+a:b,data = Data2)
```

```r
summary(prob4)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## a            2 220.02  110.01  1827.9 <2e-16 ***
## b            2 123.66   61.83  1027.3 <2e-16 ***
## a:b          4  29.43    7.36   122.2 <2e-16 ***
## Residuals   27   1.62    0.06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

可以看到在显著性水平 95% 下,AB 的交互作用效应显著.

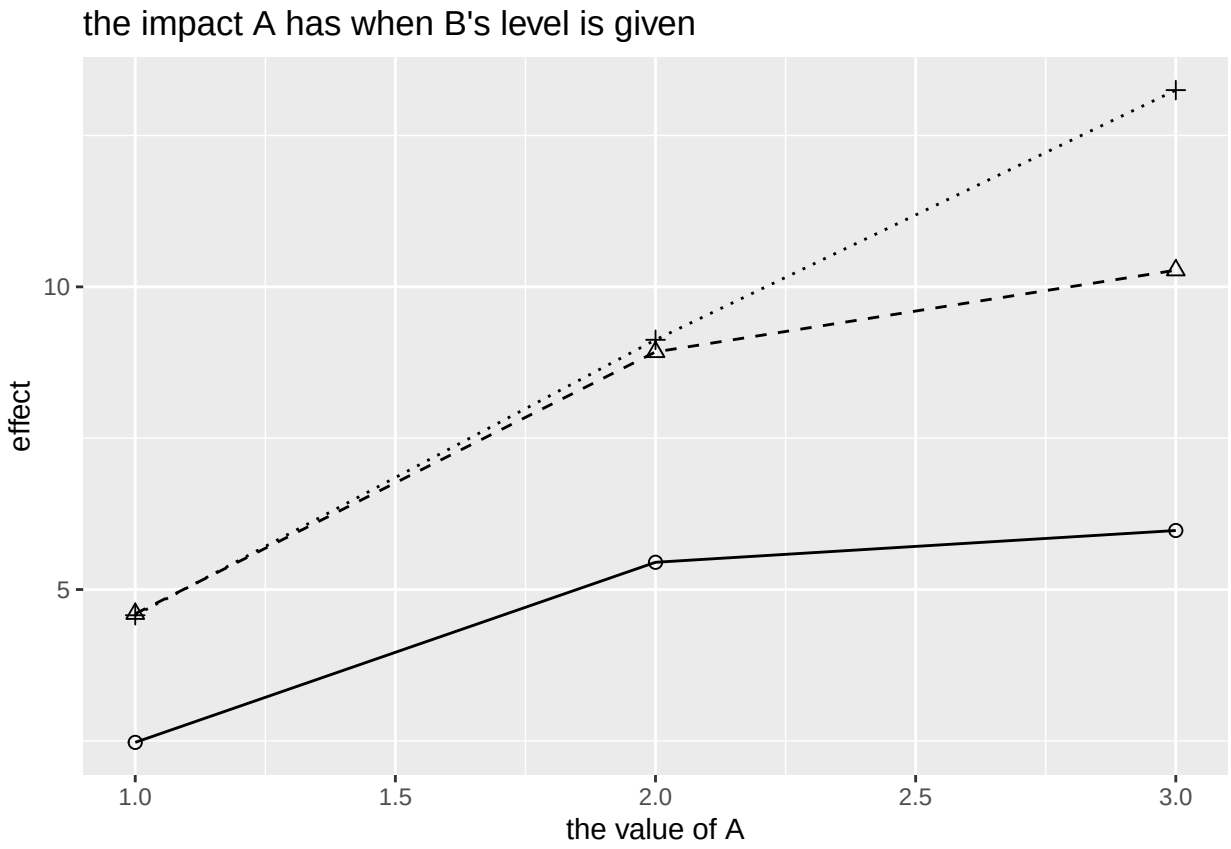
```r
mu <- matrix( apply(matrix(problem4 ,nrow = 4,ncol = 9),2,mean)
              ,nrow = 3,ncol = 3,byrow = TRUE)
mu
```

```
##       [,1]   [,2]   [,3]
## [1,] 2.475  4.600  4.575
## [2,] 5.450  8.925  9.125
## [3,] 5.975 10.275 13.250
```

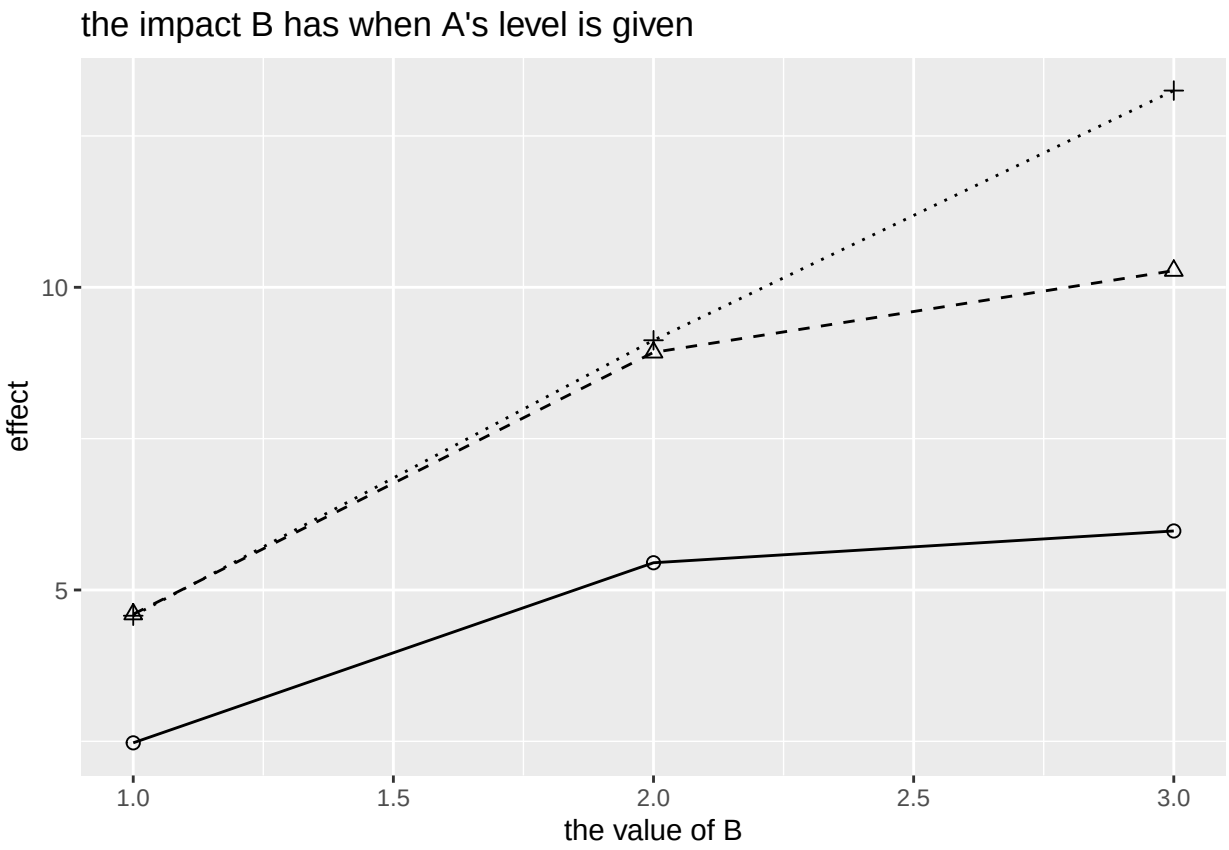
上面矩阵的 ij 个元素代表 A_i, B_j 的组合下的均值.

```r
library(ggplot2)
plot1<-ggplot(data = data.frame(mu))
plot1<-plot1+geom_point(
  aes(x=c(1,2,3),y=mu[,1]),shape=1,size=2)+geom_line(aes(x=c(1,2,3),y=mu[,1]),linetype = 1)
plot1<-plot1+geom_point(
  aes(x=c(1,2,3),y=mu[,2]),shape=2,size=2)+geom_line(aes(x=c(1,2,3),y=mu[,2]),linetype = 2)
plot1<-plot1+geom_point(
  aes(x=c(1,2,3),y=mu[,3]),shape=3,size=2)+geom_line(aes(x=c(1,2,3),y=mu[,3]),linetype = 3)
plot1<-plot1+labs(x="the value of A",y="effect",
                  title = "the impact A has when B's level is given")
plot2<-ggplot(data = data.frame(mu))
plot2<-plot1+geom_point(
  aes(x=c(1,2,3),y=mu[1,]),shape=4,size=2)+geom_line(aes(x=c(1,2,3),y=mu[1,]),linetype = 1)
plot2<-plot1+geom_point(
  aes(x=c(1,2,3),y=mu[2,]),shape=5,size=2)+geom_line(aes(x=c(1,2,3),y=mu[2,]),linetype = 2)
```

```
plot2<-plot1+geom_point(
  aes(x=c(1,2,3),y=mu[3,]),shape=6,size=2)+geom_line(aes(x=c(1,2,3),y=mu[3,]),linetype = 3)
plot2<-plot1+labs(x="the value of B",y="effect",
                  title = "the impact B has when A's level is given")
plot1
```




```
plot2
```



从图上也可以看出其交互作用显著. 因为其在一个水平给定了以后, 另外一个水平改变的均值的量与先前给定的这个水平有关.

4.2 (3)

为了计算均值 μ_{ij} 的 95% 置信区间, 先计算

```
alpha <- 0.05
Deviation<-qt(1-alpha/2,prob4$df.residual)*sqrt(
  (sum(prob3$residuals^2)/prob4$df.residual)/4)
```

则 μ_{ij} 的置信区间的左端点为

```
 mu - Deviation
```

```
##          [,1]    [,2]     [,3]
## [1,] 1.70115 3.82615  3.80115
```

```
## [2,] 4.67615 8.15115  8.35115
## [3,] 5.20115 9.50115 12.47615
```

右端点为

```
mu + Deviation
```

```
##          [,1]     [,2]     [,3]
## [1,] 3.24885  5.37385  5.34885
## [2,] 6.22385  9.69885  9.89885
## [3,] 6.74885 11.04885 14.02385
```

给定水平 A_i, 我们的不同 B_j 水平上均值 μ_{ij} 两两之差为

```
GivenA <- matrix(c(mu[,2]-mu[,1],mu[,3]-mu[,2],mu[,3]-mu[,1]),nrow = 3,ncol = 3 )
GivenA
```

```
##       [,1]   [,2]  [,3]
## [1,] 2.125 -0.025 2.100
## [2,] 3.475  0.200 3.675
## [3,] 4.300  2.975 7.275
```

其第 i 行代表在水平 A_i 给定下 $\mu_{i2}-\mu_{i1}$, $\mu_{i3}-\mu_{i2}$ 与 $\mu_{i3}-\mu_{i1}$ 的值.

为了计算其置信区间首先计算

```
diffDeviation<-qt(1-alpha/(2*3),prob4$df.residual)*sqrt(
  2*(sum(prob4$residuals^2)/prob4$df.residual)/4)
```

则其同时置信区间的左端点为

```
GivenA -diffDeviation
```

```
##          [,1]       [,2]     [,3]
## [1,] 1.682219 -0.4677806 1.657219
## [2,] 3.032219 -0.2427806 3.232219
## [3,] 3.857219  2.5322194 6.832219
```

其同时置信区间的右端点为

```
GivenA + diffDeviation
```

```
##          [,1]      [,2]     [,3]
## [1,] 2.567781 0.4177806 2.542781
```

```
## [2,] 3.917781 0.6427806 4.117781
## [3,] 4.742781 3.4177806 7.717781
```

而给定水平 B_j, 我们的不同 A_i 水平上均值 μ_{ij} 两两之差为

```
GivenB <- matrix(c(mu[2,]-mu[1,],mu[3,]-mu[2,],mu[3,]-mu[1,]),nrow = 3,ncol = 3 )
GivenB
```

```
##       [,1]  [,2]  [,3]
## [1,] 2.975 0.525 3.500
## [2,] 4.325 1.350 5.675
## [3,] 4.550 4.125 8.675
```

其第 i 行代表在水平 B_i 给定下 $\mu_{2i}-\mu_{1i}$, $\mu_{3i}-\mu_{2i}$ 与 $\mu_{3i}-\mu_{1i}$ 的值.

其同时置信区间的左端点为

```
GivenB -diffDeviation
```

```
##          [,1]       [,2]     [,3]
## [1,] 2.532219 0.08221944 3.057219
## [2,] 3.882219 0.90721944 5.232219
## [3,] 4.107219 3.68221944 8.232219
```

其同时置信区间的右端点为

```
GivenB + diffDeviation
```

```
##          [,1]      [,2]     [,3]
## [1,] 3.417781 0.9677806 3.942781
## [2,] 4.767781 1.7927806 6.117781
## [3,] 4.992781 4.5677806 9.117781
```

我们应该选择低水平的 A 与低水平的 B. 因为不管先固定 A 还是先固定 B, 我们的同时置信区间的正负都告诉了我们应该选择低水平的 AB 来达到均值的最小.