一、在一个农业实验中，育种人员测试了3种不同的种子的粮食产量（单位：共计/亩），结果记录在数据yield.txt中。（20’）请问（1）种子的品种是否影响粮食产量；（10’）（2）如果受品种影响，那么哪一个品种和其他品种的产量有差异。（10’）

**参考答案：**

读入数据

> dat <- read.table("yield.txt",

+ header = T,

+ sep = "\t",

+ colClasses = c("numeric", "factor"))

检验各组数据是否满足正态分布。

H0：种子1的产量满足正态分布；H1：种子1的产量不满足正态分布。

> shapiro.test(dat[which(dat$seed==1),1])

Shapiro-Wilk normality test

data: dat[which(dat$seed == 1), 1]

W = 0.93955, p-value = 0.548

因为p值大于0.05，因此接受原假设，即认为种子1的产量满足正态分布。

同样的，种子2和种子3的shapiro检验结果的p值分别为0.47和0.97，均大于0.05，故认为这两组数据也满足正态分布。

检验三组数据是否满足方差齐次性。

H0：三组数据方差没有差别；H1：方差有差别。

> bartlett.test(dat$yield~dat$seed)

Bartlett test of homogeneity of variances

data: dat$yield by dat$seed

Bartlett's K-squared = 4.0299, df = 2, p-value = 0.1333

因为p大于0.05，所以接受原假设，即认为三组数据满足方差齐次性。

（1）用方差分析方法检验种子的品种是否影响粮食产量

H0：种子的品种不影响粮食产量；H1：至少一个种子的品种影响粮食产量。

> summary(aov(data = dat,yield~seed))

Df Sum Sq Mean Sq F value Pr(>F)

seed 2 4364 2182.2 10.04 0.000586 \*\*\*

Residuals 26 5649 217.3

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

检验结果显示p值小于0.05，因此拒绝原假设，即粮食产量受种子品种的影响。

（2）用pairwised t检验分析哪个品种的产量和其他品种不同

> pairwise.t.test(dat$yield,dat$seed,p.adjust.method = "none")

Pairwise comparisons using t tests with pooled SD

data: dat$yield and dat$seed

1 2

2 0.00056 -

3 0.98870 0.00106

P value adjustment method: none

结果显示种子1和种子3的t检验的p值大于0.05，而种子2和其他两者的t检验p值均小于0.05，因此种子2和其他种子相比产量有差异。

二、为研究茶多酚保健饮料对急性缺氧的影响，某研究者将60只小白鼠随机分为低、中、高三个剂量组和一个对照组，每组15只老鼠。对照组给予蒸馏水0.25ml灌胃，低中高分别给予递增剂量的饮料，并将饮料溶于0.2~0.3ml蒸馏水后灌胃。每天一次，40天后，对小鼠进行耐缺氧存活时间试验，结果见数据文件。（30’）

1. 在本次试验中，为研究不同剂量的茶多酚保健饮料对延长小白鼠的平均耐缺氧存货时间有无差异。问因素或者处理是什么？，与之相对的有多少个分组或者水平？（5’）
2. 请将数据读入R中，并用Box plot查看总体数据的情况，最后查看一下数据的最小值、中位数、平均数等信息。（5’）
3. 有研究员对上述资料采用了两样本均数t检验进行了两两比较。问这样处理是否合理，为什么？应采用何种处理方法。（5’）
4. 为了用更好的方法来处理数据，请说明数据应满足的哪三个基本条件？试检验这批数据是否满足这些条件。（提示：一般需要满足三个条件；shapiro.test()函数可检验正态性，bartlett.test()可检验多个正态总体的方差齐次性）（5’）
5. 请根据上述验证情况，对数据进行分析，并给出差异的配对组别（10’）

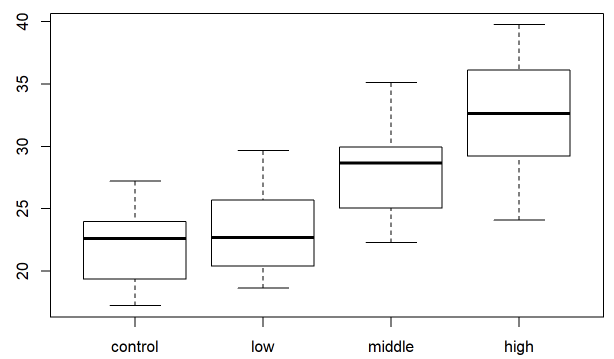
参考答案：

1. 一个因素，保健饮料剂量；四个水平

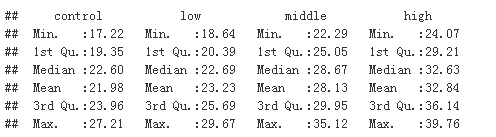
2. 三个命令行各二分

>df <- read.table(file = "data.txt",header = T)

>boxplot(df)



>summary(df)



3. 不合理，给出使用方差分析的充分理由即可。如会增大犯一型错误的概率。

4. 需满足1. 各处理样本是相互独立的随机样本。2.各处理样本相应总体服从正态分布。3.个处理样本总体方差齐次

一般实验设计与实施满足第一个要求，检验要求2

#用for循环也可

>shapiro.test(df[,1])

##

## Shapiro-Wilk normality test

##

## data: df[, 1]

## W = 0.95001, p-value = 0.5245

>shapiro.test(df[,2])

##

## Shapiro-Wilk normality test

##

## data: df[, 2]

## W = 0.93877, p-value = 0.3671

>shapiro.test(df[,3])

##

## Shapiro-Wilk normality test

##

## data: df[, 3]

## W = 0.94679, p-value = 0.4754

设显著水平0.05，P值均大于0.05故满足假设2

检验要求3

#数据格式转换，自己写for循环也可以# aggregate函数都可以

>dff <- reshape2::melt(df)

## No id variables; using all as measure variables

#R中特有的因子型数据类型，有时候不加会有问题。

>dff$variable <- as.factor(dff$variable) #检验

>bartlett.test(value ~ variable,data = dff)

## Bartlett test of homogeneity of variances

##

## data: value by variable

## Bartlett's K-squared = 2.1206, df = 3, p-value = 0.5478

设显著水平0.05，P值均大于0.05故满足假设3

5. 改卷的时候一定要注意假设检验的四个步骤，缺一不可

1.H0 四个总体均值相等 HA不等，alpha 为0.05

2.套代码

3.计算p值

4.做出统计推断

5.给出显著的组别

除了低剂量与对照之外其余都有差异

>res <- aov(value ~ variable,data = dff)

>summary(res)

## Df Sum Sq Mean Sq F value Pr(>F)

## variable 3 1109.1 369.7 24.46 3.03e-10 \*\*\*

## Residuals 56 846.5 15.1

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

>TukeyHSD(res)

## Tukey multiple comparisons of means

## 95% family-wise confidence level

##

## Fit: aov(formula = value ~ variable, data = dff)

##

## $variable

## diff lwr upr p adj

## low-control 1.254667 -2.5044218 5.013755 0.8132480

## middle-control 6.151333 2.3922449 9.910422 0.0003513

## high-control 10.860000 7.1009116 14.619088 0.0000000

## middle-low 4.896667 1.1375782 8.655755 0.0057640

## high-low 9.605333 5.8462449 13.364422 0.0000001

## high-middle 4.708667 0.9495782 8.467755 0.0084746

三、已知有三种药物都能促进小鼠肠道对营养的吸收，现将初始状态相近的两批成年小鼠分别进行给药，经过一段时间后测量其体重，然后得到这段时间内小鼠体重增加的值。问：这三种药导致的平均体重增加值有无统计学差异？（20’）

|  |  |  |
| --- | --- | --- |
| 药物1（增加的体重值g） | 药物2（增加的体重值g） | 药物3（增加的体重值g） |
| 40 | 50 | 60 |
| 10 | 20 | 30 |
| 35 | 45 | 100 |
| 25 | 55 | 85 |
| 20 | 20 | 20 |
| 15 | 15 | 55 |
| 35 | 80 | 45 |
| 15 | -10 | 30 |
| -5 | 105 | 77 |
| 30 | 75 | 105 |
| 25 | 10 |  |
| 70 | 60 |  |
| 65 | 45 |  |
| 45 | 60 |  |
| 50 | 30 |  |

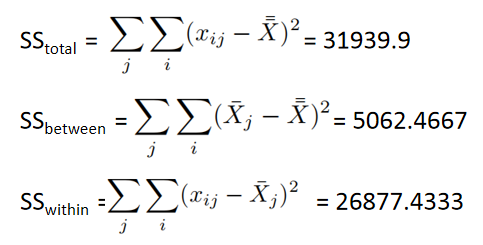
解：

求均值的统计学差异，用ANOVA来解决

H0：三种药物对小鼠体重影响的均值相同；

H1：三种药物对小鼠体重影响的均值不同。

根据表格计算出：总均值=43.55，总样本N=40



|  |  |  |  |
| --- | --- | --- | --- |
| 变异来源 | SS | df | MS |
| 总变异 | 31939.9 | 39 |  |
| 组间变异 | 5062.4667 | 2 | 2531.2333 |
| 误差 | 26877.4333 | 37 | 726.4171 |

所以F=（SSbetween/df1）/（SSwithin/df2）=2531.2333/726.4171=3.4845

查表，当df1=2，df2=37，F0.05=3.25，所以F>F0.05，所以P<0.05.

结论：拒绝H0，接受H1，三种药物对小鼠体重影响的均值不同。

四、在一生物实验中，为了研究不同饲养条件对大鼠体重的影响，现将60只8周龄体重相等的大鼠随机分为六组，分别放入以下饲养条件中培养两周：

A：饲养温度（。C） 4（A1），25（A2），30（A3）

B：饲料：普通饲料（B1），高脂饲料（B2）

最后测得体重为下表：

|  |  |  |  |
| --- | --- | --- | --- |
|  | A1 | A2 | A3 |
| B1 | 282.1 | 296.7 | 300.1 |
| 264.2 | 318 | 307.5 |
| 274.2 | 295.3 | 294.2 |
| 276.4 | 292.8 | 312 |
| 283.7 | 304.5 | 300.2 |
| 288 | 305.9 | 292.6 |
| 274.3 | 312.3 | 302 |
| 278.4 | 311.4 | 306.9 |
| 293.5 | 307.6 | 313.3 |
| 271.5 | 292.7 | 312.4 |
| B2 | 284.5 | 296.6 | 304 |
| 263.7 | 323.9 | 312.4 |
| 292.3 | 296.5 | 297.8 |
| 270.3 | 298.1 | 318.4 |
| 281.3 | 310 | 302.4 |
| 286.9 | 312.5 | 295.3 |
| 271.2 | 317.6 | 305.5 |
| 275.6 | 305.5 | 309.9 |
| 289.4 | 305.8 | 319.2 |
| 289.8 | 295.4 | 320.2 |

1. 检验体重数据对于因素A和因素B是否是正太的？是否满足方差齐性的要求？**（10分）**
2. 试分析因素A、因素B以及两因素的相互作用对大鼠体重有无显著影响？**（10分）**
3. 对AiBj条件下平均产量作多重比较，并指出与A2B1组有显著差异的组。**（10分）**

参考答案：

**（1）**

> data<-read.csv("体重.csv",header = T,sep = ",")

> shapiro.test(data$Weight[data$A=="A1"])**（1分）**

Shapiro-Wilk normality test

data: data$Weight[data$A == "A1"]

W = 0.96118, p-value = 0.5676

> shapiro.test(data$Weight[data$A=="A2"])**（1分）**

Shapiro-Wilk normality test

data: data$Weight[data$A == "A2"]

W = 0.93573, p-value = 0.1989

> shapiro.test(data$Weight[data$A=="A3"])**（1分）**

Shapiro-Wilk normality test

data: data$Weight[data$A == "A3"]

W = 0.96408, p-value = 0.6281

> shapiro.test(data$Weight[data$B=="B1"])**（1分）**

Shapiro-Wilk normality test

data: data$Weight[data$B == "B1"]

W = 0.95026, p-value = 0.1718

> shapiro.test(data$Weight[data$B=="B2"])**（1分）**

Shapiro-Wilk normality test

data: data$Weight[data$B == "B2"]

W = 0.96797, p-value = 0.4853

> bartlett.test(Weight~A,data=data) **（2分）**

Bartlett test of homogeneity of variances

data: Weight by A

Bartlett's K-squared = 0.19888, df = 2, p-value = 0.9053

> bartlett.test(Weight~B,data=data) **（2分）**

Bartlett test of homogeneity of variances

data: Weight by B

Bartlett's K-squared = 0.20953, df = 1, p-value = 0.6471

是正太的并且满足方差齐性 **（1分）**

**（2）**

> fit<-aov(Weight~A+B+A:B,data = data)**（7分）**

> summary(fit)

Df Sum Sq Mean Sq F value Pr(>F)

A 2 9080 4540 56.809 5.22e-14 \*\*\*

B 1 127 127 1.589 0.213

A:B 2 17 9 0.108 0.897

Residuals 54 4316 80

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

因素A对大鼠体重有显著影响**（3分）**

**（3）**

> TukeyHSD(fit) **（7分）**

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = Weight ~ A + B + A:B, data = data)

$A

diff lwr upr p adj

A2-A1 25.39 18.576905 32.203095 0.0000000

A3-A1 26.75 19.936905 33.563095 0.0000000

A3-A2 1.36 -5.453095 8.173095 0.8805321

$B

diff lwr upr p adj

B2-B1 2.91 -1.717782 7.537782 0.2128406

$`A:B`

diff lwr upr p adj

A2:B1-A1:B1 25.09 13.277922 36.90208 0.0000009

A3:B1-A1:B1 25.49 13.677922 37.30208 0.0000006

A1:B2-A1:B1 1.87 -9.942078 13.68208 0.9970585

A2:B2-A1:B1 27.56 15.747922 39.37208 0.0000001

A3:B2-A1:B1 29.88 18.067922 41.69208 0.0000000

A3:B1-A2:B1 0.40 -11.412078 12.21208 0.9999985

A1:B2-A2:B1 -23.22 -35.032078 -11.40792 0.0000050

A2:B2-A2:B1 2.47 -9.342078 14.28208 0.9892562

A3:B2-A2:B1 4.79 -7.022078 16.60208 0.8358408

A1:B2-A3:B1 -23.62 -35.432078 -11.80792 0.0000035

A2:B2-A3:B1 2.07 -9.742078 13.88208 0.9952518

A3:B2-A3:B1 4.39 -7.422078 16.20208 0.8800277

A2:B2-A1:B2 25.69 13.877922 37.50208 0.0000005

A3:B2-A1:B2 28.01 16.197922 39.82208 0.0000001

A3:B2-A2:B2 2.32 -9.492078 14.13208 0.9919360

A1B1，A1B2组和A2B1组有显著差异**（3分）**

4.文件protein.csv是15只小鼠随机分为3组并且分别经历A（安慰剂）、B药物、C药物处理后（见下表）某个组织的蛋白质组数据，请做一下分析。

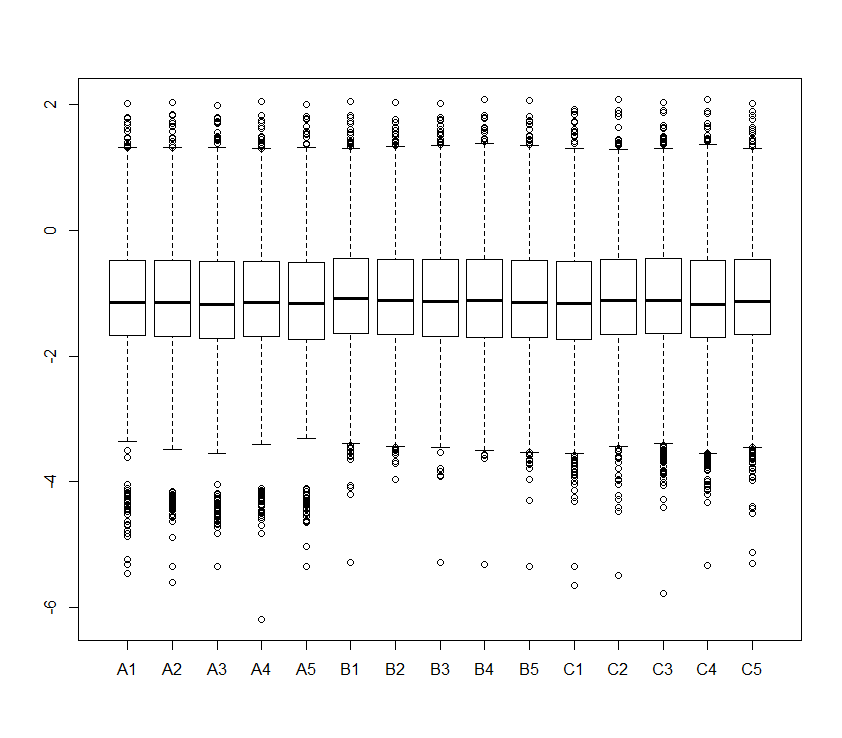
（1）对数据进行标准化（每列除以各自均值），每个数取其底为10的对数后绘制箱线图。**（5分）**

（2）首先对原始数据进行标准化（每列除以各自均值），再用方差分析(不做正太性以及方差齐性检验)筛选差异表达基因（p-value <0.05）并显示其数量。**（10分）**

（3）用p.adjust(data,method= "bonferroni")函数对（2）题p值进行校正，然后再筛选异表达基因（p-value <0.05）并显示其数量。**（5分）**

* **（1）**
* data2<-read.csv("protein.csv",sep=",",head=T,row.names = 1)
* data2<-apply(data2,2,function(x){x/mean(x)}) **（4分）**

> boxplot(log10(data2)) **（1分）**



**（2）**

data2<-apply(data2,2,function(x){x/mean(x)}) **（2分）**

> f.raw<-factor(rep(c("A","B","C"),each=5))

> anova.p<-apply(data2,1,function(x){x<-as.numeric(x);anova(lm(x~f.raw))$`Pr(>F)`[1]}) **（5分）**

> data2<-data.frame(data2,anova.p)

> data.screen<-data2[which(anova.p<0.05),] **（2分）**

> dim(data.screen) **（1分）**

[1] 1048 16

**（3）**

> anova.p.adj=p.adjust(anova.p,method= "bonferroni")**（2分）**

> data2<-data.frame(data2,anova.p.adj)

> data.screen.adj<-data2[which(anova.p.adj<0.05),] **（2分）**

> dim(data.screen.adj) **（1分）**

[1] 203 17