生物统计学作业整理

计算所-刘栋梁

2019-06-20

Preface

只对作业三到六进行了整理，作业一、二和七相对不重要，大家直接看答案即可。

肯定有不正确的地方，欢迎大家讨论，预祝大家考试顺利。

homework-3

question 1

一、实验室欲购进一批灯泡，打算在两个供货商之间选择一家购买。选购考虑的主要因素就是灯泡使用寿命的方差大小，为此需要对供货商提供的20个样品进行检测，得到的数据如下表所示。(20’)

供货商1	**供货商2**
6802	5884
5730	5871
5823	5797
5915	5957
5774	5803
5880	5862
5870	5814
5773	5885
5830	5856
5841	5940
5763	5945
5851	5803
5789	5864
5796	5851
5818	5714
5685	5943
5602	5830
5841	5858
5723	5922
5757	5866

1、检验两家供货商的灯泡使用寿命的方差有无显著差异（α=0.05）（10’）

解：检验两供应商的灯泡使用寿命的方差有无显著差异即为两个样本方差的同质性检验，可用F检验。

（1）.假设$H_0:\sigma_1^2=\sigma_2^2$,即两供应商的灯泡使用寿命的方差无显著差异；$H_A:\sigma_1^2\neq\sigma_2^2$。

（2）.确定显著水平α=0.05。

（3）.检验计算：

```r
sup_1<-
c(6802,5730,5823,5915,5774,5880,5870,5773,5830,5841,5763,5851,5789,5796,5818,5685,5602,
5841,5723,5757)
sup_2<-
c(5884,5871,5797,5957,5803,5862,5814,5885,5856,5940,5945,5803,5864,5851,5714,5943,5830,
5858,5922,5866)
n1<-length(sup_1)
n2<-length(sup_2)
var1<-var(sup_1)
var2<-var(sup_2)
F<-var1/var2
F
```

```
## [1] 15.2795
```

```r
p<-2*(1-pf(F,df1=(n1-1),df2=(n2-1)))
p
```

```
## [1] 1.799681e-07
```

答：拒绝原假设，认为两家供货商灯泡使用寿命的方差存在显著差异。

2、选择最合适的检验方法检验两家供应商的灯泡使用寿命有无差别。（10’）

解：检验两家供应商的灯泡使用寿命有无差别即总体方差未知，且两样本都属于小样本，故应该用两组平均数差异显著性的双尾t检验。

先做方差齐性检验，第一问其实已经得出结论：方差有显著差异。当然我们也可以直接通过stats包中的var.test进行检测。

假设$H_0:\sigma_1^2=\sigma_2^2$,即两供应商的灯泡使用寿命的方差无显著差异；$H_A:\sigma_1^2\neq\sigma_2^2$。

```r
var.test(sup_1,sup_2,conf.level = 0.95,alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  sup_1 and sup_2
## F = 15.279, num df = 19, denom df = 19, p-value = 1.8e-07
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   6.047811 38.602899
## sample estimates:
## ratio of variances
##            15.2795
```

因为两个供货商的方差存在显著差异，所以var.equal=false。以下做t检验。

原假设：两供应商的灯泡使用寿命无显著差别。

备择假设：两供应商的灯泡使用寿命有显著差别。

```r
t.test(sup_1,sup_2,var.equal=FALSE,alternative = "two.sided",conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  sup_1 and sup_2
## t = -0.36748, df = 21.476, p-value = 0.7169
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -133.69464   93.49464
## sample estimates:
## mean of x mean of y
##   5843.15   5863.25
```

答：因为p=0.7169>0.05,因此不拒绝原假设，认为两供应商的灯泡使用寿命无显著差别。

R language application (25')

Please use R to resolve the following issues and display your R code and results.

1. For a normal random variable X with mean 4.0, and standard deviation 1.0,

a)find the probability that X is less than 2.0. (4')

b)find the value K so that P(X>K) = 0.05. (4')

```r
# find the probability that X is less than 2.0.
pnorm(2,mean = 4,sd = 1)
```

```
## [1] 0.02275013
```

```r
# find the value K so that P(X>K) = 0.05.
qnorm(0.05,mean = 4,sd =1,lower.tail = FALSE)
```

```
## [1] 5.644854
```

2. When tossing a fair coin 8 times,

a)find the probability of seeing no heads (Hint: this is a binomial distribution.) (3')

b)find the probability of seeing exactly 4 heads. (3')

c)find the probability of seeing more than 5 heads. (3')

It is a fair coin,so the prop = 0.5.

```r
# find the probability of seeing no heads
dbinom(0,8,0.5)
```

```
## [1] 0.00390625
```

```
# find the probability of seeing exactly 4 heads.
dbinom(4,8,0.5)
```

```
## [1] 0.2734375
```

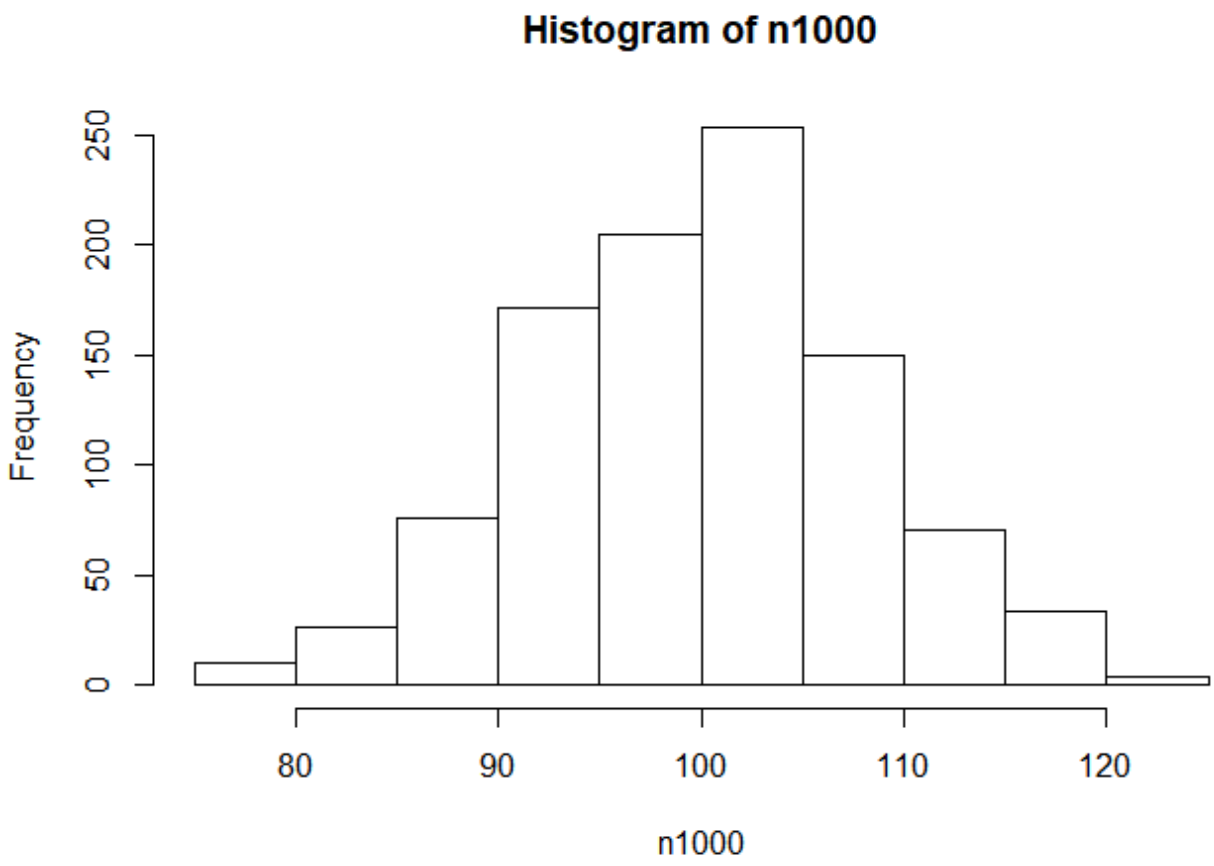
```
# find the probability of seeing more than 5 heads.
pbinom(5,8,0.5,lower.tail = FALSE)
```

```
## [1] 0.1445313
```

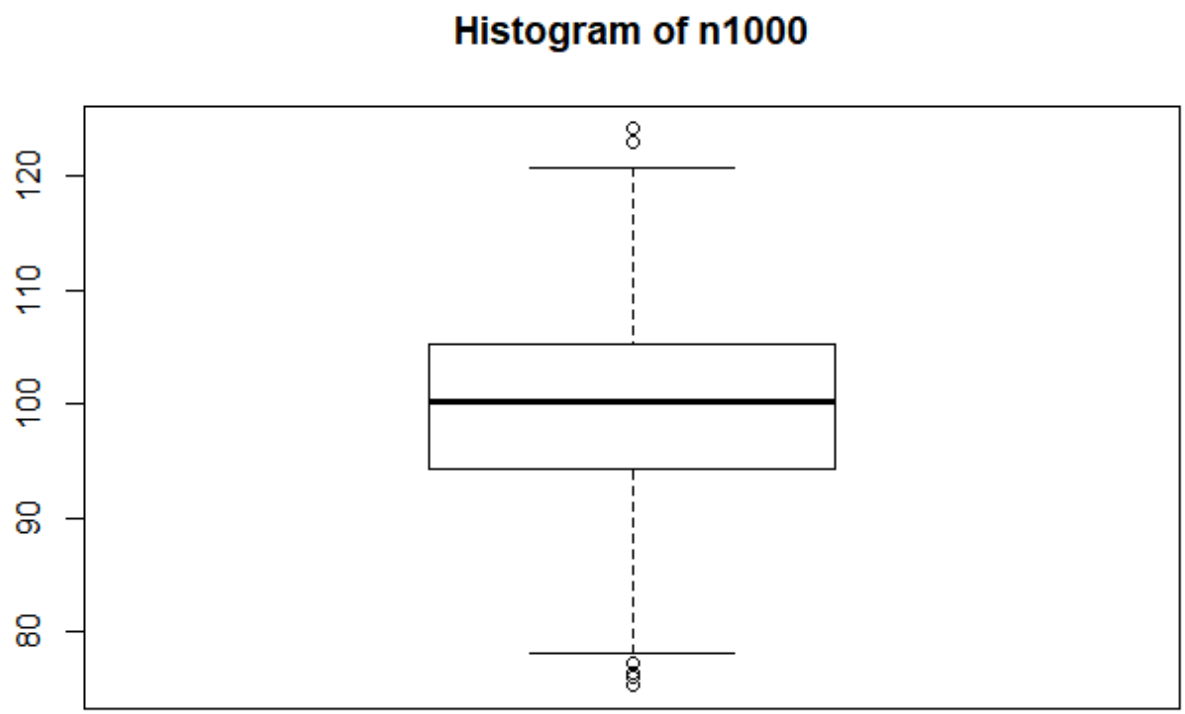
3. Simulate a sample of 1000 random data points from a normal distribution with mean 100 and standard deviation 8, and store the result in a vector.

a)plot a histogram and a boxplot of the vector you just created. (4')

```
set.seed(46)
n1000 <- rnorm(1000,mean = 100,sd = 8)
hist(n1000, main = "Histogram of n1000")
```



```r
boxplot(n1000,main = "Histogram of n1000")
```



b)using the data above, test the hypothesis that the mean equals 100 (using t.test). (4')

(1) .假设H_0：$\mu_1 = \mu_2$,即样本与总体均值之间并无显著差异；H_A：$\mu_1 \neq \mu_2$。

(2) .确定显著水平α=0.05。

(3) .检验计算：

```r
t.test(n1000,mu=100)
```

```
##
##  One Sample t-test
##
## data:  n1000
## t = -0.31056, df = 999, p-value = 0.7562
## alternative hypothesis: true mean is not equal to 100
## 95 percent confidence interval:
##   99.40978 100.42893
## sample estimates:
## mean of x
##  99.91936
```

因为p值大于0.05，所以不拒绝原假设，认为样本与总体之间并无显著差异。

R language application. (25')

(Please use R to read-in and manipulate data, code and results should be displayed.)

In order to detect air quality, a city's environmental protection department conducted a random test of PM2.5 in the air every few weeks. It is known that the average value of PM2.5 per cubic meter of air in the city is 82ug/m3. In the most recent test, the value of PM2.5 per cubic meter of air(ug/m3) is shown in the homework3_data.

1)Show your work directory (2')

```
getwd()
```

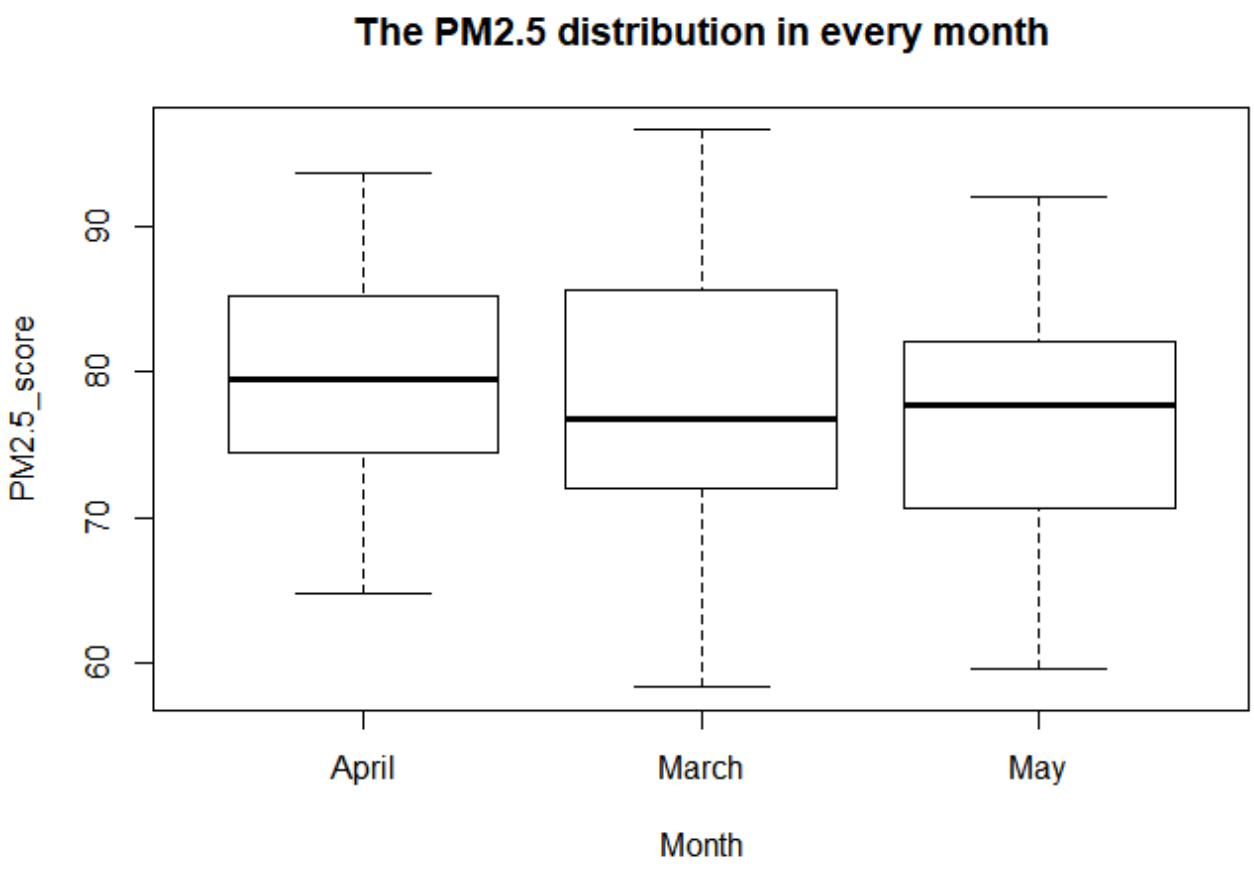
```
## [1] "C:/Users/liu/Documents/biostats_homework_summary"
```

2)use R to read in the data (3')

```
# File is in your work directory, '~' means your  work directory.
pm25 <- read.delim("./data/homework3_data.txt")
```

3)use boxplot to show the PM2.5 distribution in every month (10')

```
boxplot(PM2.5_score~Month,pm25,main = "The PM2.5 distribution in every month")
```



4)get the data of month equal to March and store in data_march (10')

```
data_march <- subset(pm25,Month=="March")
head(data_march)
```

```
##   PM2.5_score Month
## 1        81.6 March
## 2        86.6 March
## 3        80.8 March
## 4        85.8 March
## 5        78.6 March
## 6        58.3 March
```

question 4

Suppose we draw a sample of size 20 of birthweights from a hospital, the details can be found in the homework data. The mean of national-wide birthweights is 118. (30')

```
library(readr)
birthweights <- read_csv("./data/homework4_data.csv")
```

```
## Parsed with column specification:
## cols(
##   Individual_ID = col_double(),
##   Birthweight = col_double()
## )
```

```
head(birthweights)
```

```
## # A tibble: 6 x 2
##   Individual_ID Birthweight
##           <dbl>       <dbl>
## 1             1         123
## 2             2          98
## 3             3         115
## 4             4         120
## 5             5         105
## 6             6         135
```

```
summary(birthweights$Birthweight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    98.0   111.2   119.5   120.0   128.8   141.0
```

1)What is the probability that the mean birthweight of the sample falls between 100.0 and 126.0? Please list the formulas for this and also the R code for it.(5')

because σ^2 is unknown and n is less than 20,so we should use t-distribution to calcuate probability that we need.

$$S_{\overline{X}} = \frac{S}{\sqrt{n}} \tag{1}$$

$$\mathrm{P}(100.0 \leq \mu \leq 126.0) = pt_{n-1}\left(\frac{126-\overline{X}}{S_{\overline{X}}}\right) - pt_{n-1}\left(\frac{100-\overline{X}}{S_{\overline{X}}}\right) \tag{2}$$

```
n <- 20
mu <- 118
df <- n-1
mean <- mean(birthweights$Birthweight)
sd <- sd(birthweights$Birthweight)
Sx <- sd/sqrt(n)
p_126 <- pt((126-mu)/Sx,df)
p_100 <- pt((100-mu)/Sx,df)
p <- p_126-p_100
p
```

```
## [1] 0.9961359
```

So the probability that the mean birthweight of the sample falls between 100.0 and 126.0 is 0.9961359.

2)What is the 95% confidence interval of the sample mean?(5')

$$P\left(\overline{X} - t_{df,1-\alpha/2}\frac{s}{\sqrt{n}} \leq \mu_0 \leq \overline{X} + t_{df,1-\alpha/2}\frac{s}{\sqrt{n}}\right) = 1-\alpha \tag{3}$$

```
a <- 0.05
P <- 1-a/2
L1 <- mean - qt(P,n-1)*Sx
L2 <- mean + qt(P,n-1)*Sx
L1;L2
```

```
## [1] 114.3777
```

```
## [1] 125.6223
```

So the 95% confidence interval of the sample mean is [114.378,125.622].

3)Can we say the underlying mean birthweight from this hospital is higher than the national average?

Please list the formulas for this and also the R code for it.(5')

（1）.假设H_0：$\mu_1 \leqslant \mu_2$,这家医院的新生儿体重小于或等于全国平均水平；H_A：$\mu_1 > \mu_2$。

（2）.确定显著水平α=0.05。

(3) .检验计算:

$$S_{\overline{X}} = \frac{S}{\sqrt{n}} \tag{4}$$

$$\mathrm{t} = \frac{\overline{\mathrm{x}} - \mu_0}{S_{\overline{X}}} \tag{5}$$

$$\mathrm{p} = 1 - \mathrm{pt}_{\mathrm{n}-1}(\mathrm{t}) \tag{6}$$

```
t.test(birthweights$Birthweight,alternative = "greater",mu = 118)
```

```
##
##  One Sample t-test
##
## data:  birthweights$Birthweight
## t = 0.74454, df = 19, p-value = 0.2328
## alternative hypothesis: true mean is greater than 118
## 95 percent confidence interval:
##  115.3552      Inf
## sample estimates:
## mean of x
##       120
```

p>0.05,so maybe we can not say the underlying mean birthweight from this hospital is higher than the national average.

4)Test the hypothesis that the mean birthweight of sample size 20 is different from the national average (Significance level 0.05). Please list the formulas for this and also the R code for it.(5')

(1) .假设H_0：$\mu_1 = \mu_2$,即这家医院新生儿体重和全国新生儿体重平均值之间并无显著差异；H_A：$\mu_1 \neq \mu_2$。

(2) .确定显著水平α=0.05。

(3) .检验计算:

$$S_{\overline{X}} = \frac{S}{\sqrt{n}} \tag{7}$$

$$\mathrm{t} = \frac{\overline{\mathrm{x}} - \mu_0}{S_{\overline{X}}} \tag{8}$$

$$\mathrm{p} = 2 \times \min\left(\mathrm{pt}_{\mathrm{n}-1}\left(\frac{\overline{\mathrm{x}} - \mu_0}{\mathrm{s}/\sqrt{\mathrm{n}}}\right), 1 - \mathrm{pt}_{\mathrm{n}-1}\left(\frac{\overline{\mathrm{x}} - \mu_0}{\mathrm{s}/\sqrt{\mathrm{n}}}\right)\right) \tag{9}$$

```
t.test(birthweights$Birthweight,alternative = "two.sided",mu = 118)
```

```
##
##  One Sample t-test
##
## data:  birthweights$Birthweight
## t = 0.74454, df = 19, p-value = 0.4657
## alternative hypothesis: true mean is not equal to 118
## 95 percent confidence interval:
##  114.3777 125.6223
## sample estimates:
## mean of x
##       120
```

p-value = 0.4657>0.05,不拒绝原假设，认为这家医院新生儿体重和全国新生儿体重平均值之间并无显著差异。

5)Compute the power of the test performed in (4) with significance level 0.05.(5')

$$Power = \mathrm{F}\left[\mathrm{t}_{\alpha/2} + \frac{\left|\overline{\mathrm{X}} - \mu_0\right|}{\mathrm{s}/\sqrt{\mathrm{n}}}\right] \tag{10}$$

```
power <- pt(qt(a/2,df)+abs(mean-mu)/Sx,n-1)
power
```

```
## [1] 0.09667934
```

So,the power of the test performed in (4) with significance level is 0.09667934.

6)To see the significance difference between the sample mean and the national mean and ensure the type II error to be β =0.05, what is the appropriate sample size with significance level is 0.01?(5')

$$\mathrm{n} = \frac{\left(\mathrm{z}_{1-\beta} + \mathrm{z}_{1-\alpha/2}\right)^2 \mathrm{S}^2}{\left(\overline{\mathrm{x}} - \mu_0\right)^2} \tag{11}$$

```
# 使用pwr包进行估计
# install.packages("pwr")
library("pwr")
d = abs(mean-mu)/sd
pwr.t.test(d=d,power=0.95,sig.level=0.01,type="one.sample",alternative="two.sided")
```

```
##
##      One-sample t test power calculation
##
##               n = 646.0381
##               d = 0.1664842
##       sig.level = 0.01
##           power = 0.95
##     alternative = two.sided
```

So,the appropriate sample size with significance level=0.01 is 647.

homework-4

question 1

一、为研究某种新药对抗凝血酶活力的影响，随机安排新药组病人12例，对照组病人10例，分别测定器抗凝血酶活力（单位：），其结果如下：

```
  新药组：126 125 138 128 123 138 142 116 110 108 113 140

  对照组：160 175 177 170 175 153 168 159 160 162
```

试分析新药组和对照组病人的抗凝血酶活力有无差别（α = 0.05）

解：

（1）检验两组样本方差是否相同 (15')

假设检验：

H_0:两组样本方差不存在差异。

H_1:两组样本方差不存在差异。

先做正态检验，再做方差检验。

```
new_drug <- c(126,125,138,128,123,138,142,116,110,108,113,140)
control <- c(160,175,177,170,175,153,168,159,160,162)
# 原假设：样本符合正态分布；备择假设：样本不符合正态分布。
shapiro.test(new_drug)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  new_drug
## W = 0.92304, p-value = 0.3121
```

```
shapiro.test(control)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  control
## W = 0.92374, p-value = 0.3892
```

```
var.test(new_drug,control)
```

```
##
##  F test to compare two variances
##
## data:  new_drug and control
## F = 2.1512, num df = 11, denom df = 9, p-value = 0.26
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5498769 7.7181417
## sample estimates:
## ratio of variances
##           2.151159
```

结果显示两样本符合正态分布，p=0.26>0.05,可认为两组样本方差一致（α = 0.05）

（2）选择最合适的检验方法检验新药组和对照组病人的抗凝血酶活力有无差别。(15')

假设检验：

H_0:两组样本抗凝血酶活力无差别。

H_1:两组样本抗凝血酶活力有差别。

```
t.test(new_drug,control,var.equal = TRUE,alternative = "two.sided",conf.level = 0.95)
```

```
##
##  Two Sample t-test
##
## data:  new_drug and control
## t = -8.9578, df = 20, p-value = 1.947e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -49.70504 -30.92829
## sample estimates:
## mean of x mean of y
##  125.5833  165.9000
```

p < 0.05,可认为新药组和对照组病人的抗凝血酶活力有显著差别。（α = 0.05）

question 2

二、对7位健康成年人的血液测量其中的尿酸浓度，分别用手工（X）和仪器（Y）两种方法测量，结果如下表所示，请用wilcoxon signed-rank test来检测两种测量方法的精度是否存在差异？（α = 0.05）(20')

手工(X)	**4.5**	**6.5**	**6**	**9.2**	**10**	**12**	**8.3**
仪器(Y)	4	7.2	8	14	8.8	10	11.5

解：假设检验：

H_0:两种测量方法的精度不存在差异。

H_1:两种测量方法的精度存在差异。

```r
handmade <- c(4.5,6.5,6,9.2,10,12,8.3)
device <- c(4,7.2,8,14,8.8,10,11.5)
wilcox.test(handmade,device,alternative = "two.sided",paired = TRUE,exact=FALSE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  handmade and device
## V = 8.5, p-value = 0.3972
## alternative hypothesis: true location shift is not equal to 0
```

因为p-value = 0.3972>0.05，所以不拒绝原假设，认为两种测量方法的精度不存在差异。

question 3

三、在某保险种类中，一次关于2018年的索赔数额（单位：元）的随机抽样为（按升幂排列）：

```
   4152, 4579, 5053, 5112, 5745, 6250, 7081, 9048,

   12095, 14430, 17220, 20610, 22836, 48950, 67200
```

已知2017年的索赔数额的中位数为7520元。问2018年索赔的中位数与前一年是否有所变化？ （α = 0.05）(15')

Hint: You can use wilcox.test

解：假设检验问题： H_0:2018年索赔的中位数与前一年无变化。

H_0:2018年索赔的中位数与前一年有变化。

```r
insurance <-
c(4152,4579,5053,5112,5745,6250,7081,9048,12095,14430,17220,20610,22836,48950,67200)

wilcox.test(insurance,mu=7520)
```

```
##
##  Wilcoxon signed rank test
##
## data:  insurance
## V = 87, p-value = 0.1354
## alternative hypothesis: true location is not equal to 7520
```

p-value = 0.1354>0.05,所以不拒绝原假设，认为2018年索赔的中位数与前一年无变化。 （α = 0.05）

question 4

Type 1 diabetes is a multigenic disease caused by T-cell mediated destruction of the insulin producing β-cells. Although conventional (targeted) approaches of identifying causative genes have advanced our knowledge of this disease, many questions remain unanswered.

Here we have a gene data from NOD mouse after(case) and before(control) treatment. The data can be found in "Data.txt".Use the information mentioned above to answer the following questions:

a)use paired t-test to find genes which have significant expression (p<0.05) between case and control sample. Give the number of differential expressed genes and give the names of top 10 significantly differential expression genes. hint: "apply(data,1,function(x){...})" can apply function to every row in data more quickly than "for{}", "names()" or "rownames()" can be used to extract names of differentially expressed genes.(20')

```
# 导入数据
expression_data <- read.table('./data/Data.txt',header = T,stringsAsFactors =F )
# View(expression_data) ##观察数据可发现前十列为control组，后十列为case组


# 定义函数根据方差检验结果做成对t.test
t.test.p.value <- function(x){
  p_value <- t.test(x[1:10],x[11:20],paired = TRUE)$p.value
  p_value
}

p.t.test <- apply(expression_data,1,t.test.p.value)

#  Give the number of differential expressed genes
sum(p.t.test<0.05)
```

```
## [1] 2296
```

```
#  give the names of top 10 significantly differential expression genes
names(p.t.test[order(p.t.test,decreasing=F)[1:10]])
```

```
##  [1] "21744" "4008"  "4817"  "28474" "6816"  "2593"  "17191" "6786"
##  [9] "17677" "8678"
```

b)Adjust the p-values in question a) with bonferroni and FDR method to find differentially expressed genes in stringent way(list the differentially expressed gene names and the adjusted p-value). (15')

Hint: you can do the adjustment according to the fomular, or use "p.adjust()" instead.

```
p.bonf <- p.adjust(p.t.test,'bonferroni')
p.fdr <- p.adjust(p.t.test,'fdr')
p.bonf[p.bonf<0.05]
```

```
## named numeric(0)
```

```
p.fdr[p.fdr<0.05]
```

```
## named numeric(0)
```

经过校正后发现,并无显著差异表达基因。

homework-5

question-1

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

解：

（1）种子的品种是否影响粮食产量；（10’）

用方差分析方法检验种子的品种是否影响粮食产量

H_0：种子的品种不影响粮食产量；

H_1：至少一个种子的品种影响粮食产量。

```
yield <- read.delim("./data/yield.txt")

# H0:数据符合正态分布 H1：数据不符合正态分布。
shapiro.test(yield$yield[yield$seed==1])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  yield$yield[yield$seed == 1]
## W = 0.93955, p-value = 0.548
```

```
shapiro.test(yield$yield[yield$seed==2])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  yield$yield[yield$seed == 2]
## W = 0.93528, p-value = 0.4667
```

```
shapiro.test(yield$yield[yield$seed==3])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  yield$yield[yield$seed == 3]
## W = 0.97869, p-value = 0.9561
```

```
# 满足正态分布。
# H0：数据方差齐性 H1：数据方差非齐性
bartlett.test(yield~seed,data = yield)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  yield by seed
## Bartlett's K-squared = 4.0299, df = 2, p-value = 0.1333
```

```
# 满足方差齐次性。

seed_aov <- aov(yield~factor(seed),data = yield)
summary(seed_aov)
```

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

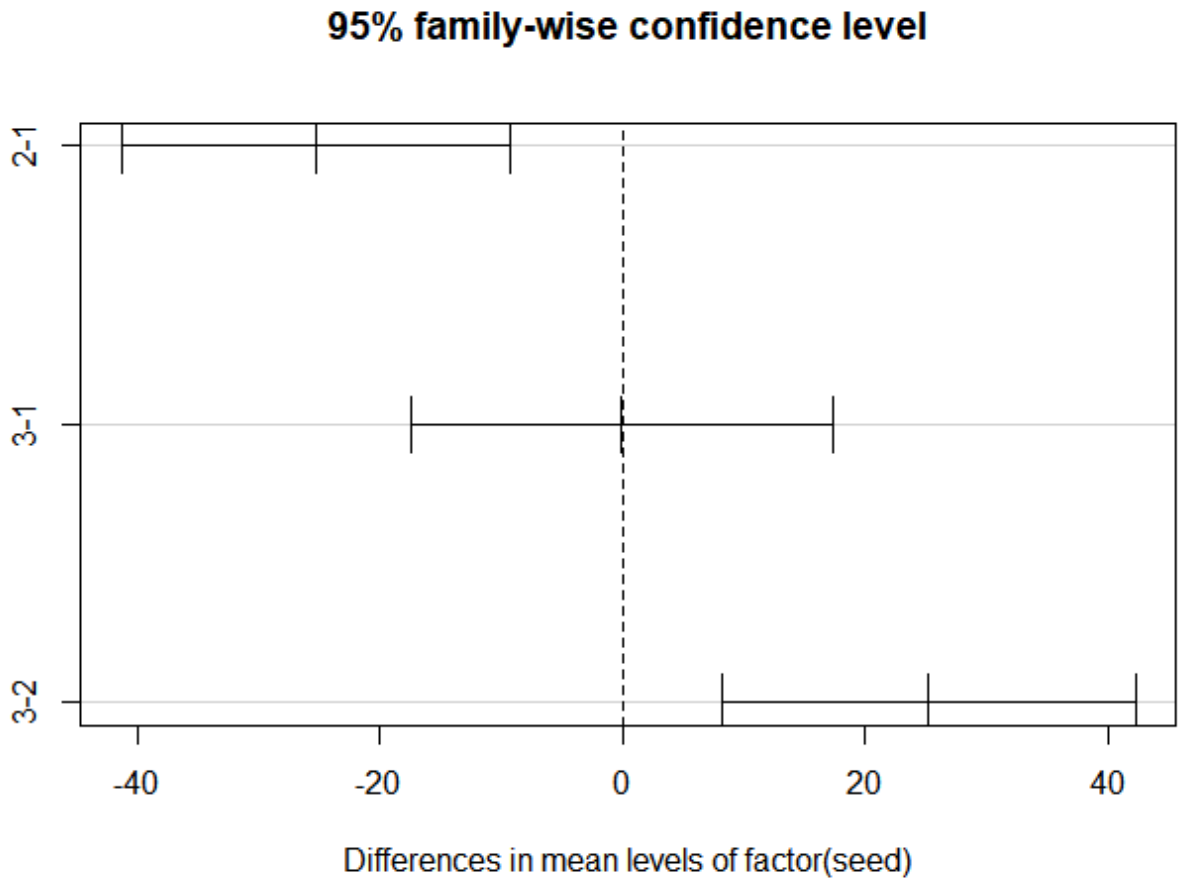
答：检验结果显示p值小于0.05，因此拒绝原假设，种子的品种会影响粮食产量。

（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’）

```
tukey_results <- TukeyHSD(seed_aov);tukey_results
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = yield ~ factor(seed), data = yield)
##
## $`factor(seed)`
##          diff        lwr       upr     p adj
## 2-1 -25.32727 -41.330375 -9.324171 0.0015630
## 3-1  -0.10000 -17.473293 17.273293 0.9998872
## 3-2  25.22727   8.208575 42.245971 0.0029482
```

```
plot(tukey_results)
```




```
# 置信区间包含0说明差异并不显著。
```

答：品种二和其他品种的产量有差异。

或者用pairwised-t检验分析哪个品种的产量和其他品种不同。

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

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  yield$yield and yield$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。

question-2

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’）

解：

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’）

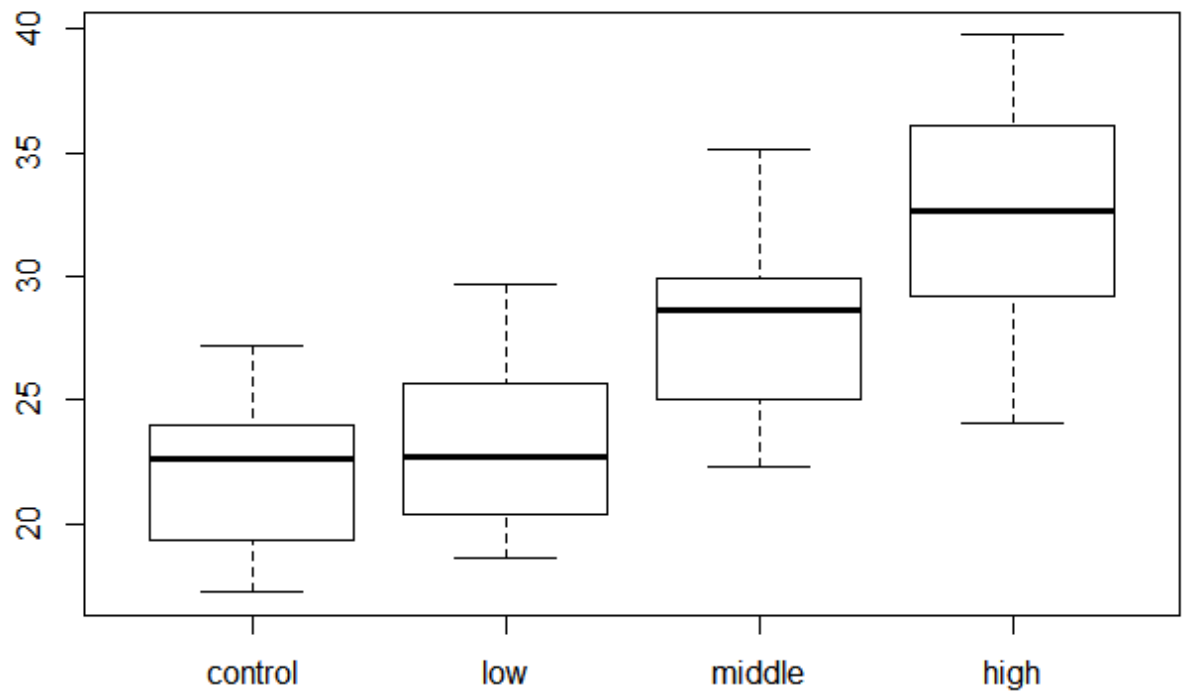
答:

因素：不同剂量的茶多酚保健饮料

四个处理：对照、低、中、高组

2.请将数据读入R中，并用Boxplot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’）

```
data <- read.delim("./data/data2.txt")
boxplot(data)
```




```
summary(data)
```

```
##     control           low            middle           high      
##  Min.   :17.22   Min.   :18.64   Min.   :22.29   Min.   :24.07  
##  1st Qu.:19.35   1st Qu.:20.39   1st Qu.:25.05   1st Qu.:29.21  
##  Median :22.60   Median :22.69   Median :28.67   Median :32.63  
##  Mean   :21.98   Mean   :23.23   Mean   :28.13   Mean   :32.84  
##  3rd Qu.:23.96   3rd Qu.:25.69   3rd Qu.:29.95   3rd Qu.:36.14  
##  Max.   :27.21   Max.   :29.67   Max.   :35.12   Max.   :39.76
```

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’）

答：不合理，t检验适合两组数据的检验，用于多组数据会增大犯一型错误的概率。对于多组数据的差异研究，应该采用方差分析。

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’）

答：

方差分析的基本假设：

（1）各总体的方差必须相等。

```
# H0：数据方差齐性 H1：数据方差非齐性
bartlett.test(data)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  data
## Bartlett's K-squared = 2.1206, df = 3, p-value = 0.5478
```

满足方差齐性。

（2）各总体必须服从正态分布。

```
# H0:数据符合正态分布 H1：数据不符合正态分布。
shapiro.test(data$control)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$control
## W = 0.95001, p-value = 0.5245
```

```
shapiro.test(data$low)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$low
## W = 0.93877, p-value = 0.3671
```

```
shapiro.test(data$middle)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$middle
## W = 0.94679, p-value = 0.4754
```

```
shapiro.test(data$high)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$high
## W = 0.95715, p-value = 0.6431
```

经检验，各总体服从正态分布。

（3）各观测值相互独立

可以通过控制抽样过程来控制独立性，无具体的检验方法。

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’）

H_0四个总体均值相等。 H_A至少有一个不等，α =0.05

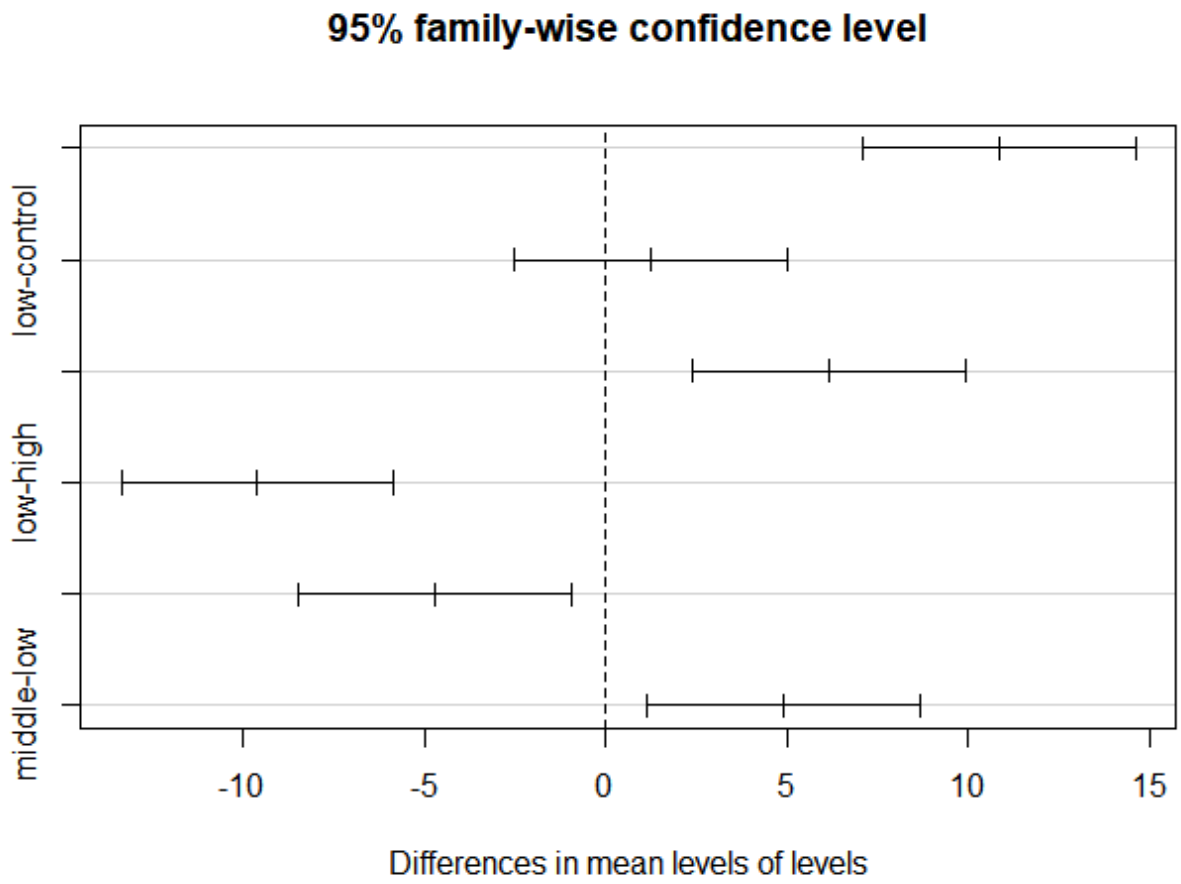
```
time <- c(data$control,data$low,data$middle,data$high)

levels <- factor(c(rep("control",15),rep("low",15),rep("middle",15),rep("high",15)))
time_aov <- aov(lm(time~levels))
summary(time_aov)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## levels       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(time_aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = lm(time ~ levels))
##
## $levels
##                     diff        lwr        upr     p adj
## high-control   10.860000   7.100912 14.6190884 0.0000000
## low-control     1.254667  -2.504422  5.0137551 0.8132480
## middle-control  6.151333   2.392245  9.9104218 0.0003513
## low-high       -9.605333 -13.364422 -5.8462449 0.0000001
## middle-high    -4.708667  -8.467755 -0.9495782 0.0084746
## middle-low      4.896667   1.137578  8.6557551 0.0057640
```

```
plot(TukeyHSD(time_aov))
```



答:除low组和control组经检验不具有显著性差异，其他组都具有显著差异。

question-3

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均值的影响都相同；

H1：三种药物对小鼠体重均值的影响至少有一个不同。

```
# 输入数据
weight1 <- c(40,10,35,25,20,15,35,15,-5,30,25,70,65,45,50)
weight2 <- c(50,20,45,55,20,15,80,-10,105,75,10,60,45,60,30)
weight3 <- c(60,30,100,85,20,55,45,30,77,105)
# 正态性检验
shapiro.test(weight1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  weight1
## W = 0.9778, p-value = 0.9523
```

```
shapiro.test(weight2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  weight2
## W = 0.98349, p-value = 0.9878
```

```
shapiro.test(weight3)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  weight3
## W = 0.94005, p-value = 0.5536
```

```
weights <- c(weight1,weight2,weight3)
drugs <- factor(c(rep("drug1",15),rep("drug2",15),rep("drug3",10)))

# 方差齐性检验
bartlett.test(weights,g = drugs)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  weights and drugs
## Bartlett's K-squared = 2.4482, df = 2, p-value = 0.294
```

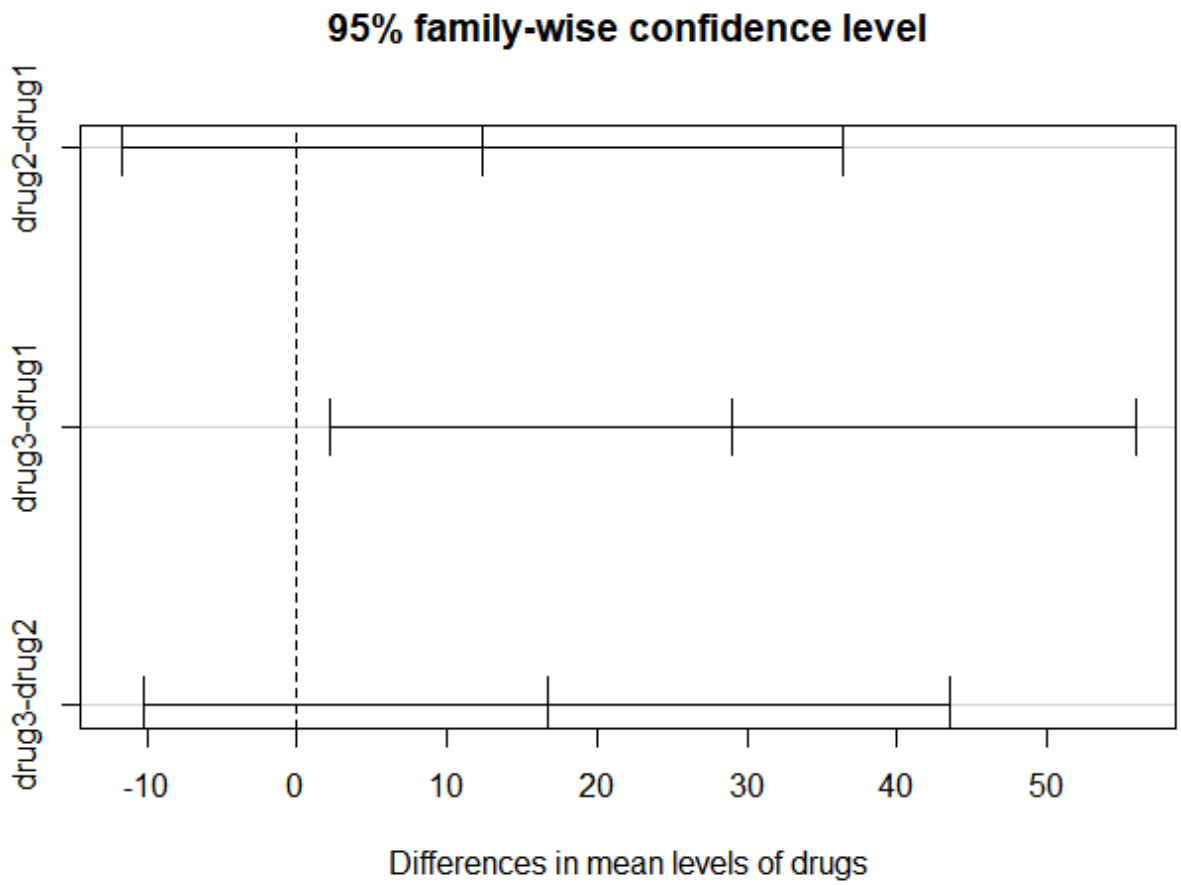
```
#方差分析
weight_aov <- aov(lm(weights~drugs))
summary(weight_aov)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## drugs        2   5062  2531.2   3.485 0.0411 *
## Residuals   37  26877   726.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(weight_aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = lm(weights ~ drugs))
##
## $drugs
##                 diff        lwr      upr     p adj
## drug2-drug1 12.33333 -11.694604 36.36127 0.4302024
## drug3-drug1 29.03333   2.169283 55.89738 0.0317315
## drug3-drug2 16.70000 -10.164050 43.56405 0.2944100
```

```
plot(TukeyHSD(weight_aov))
```



答:经检验，药物3与药物1导致的平均体重增加值之间存在显著差异，其他药物导致的平均体重增加值之间不存在显著差异。

question-4

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（℃） 4（A1），25（A2），30（A3）

B：饲料：普通饲料（B1），高脂饲料（B2）

	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

解：

（1）检验体重数据对于因素A和因素B是否是正态的？是否满足方差齐性的要求？（10分）

将数据整理成csv并导入。

```
library(readr)
rat_weight <- read_csv("./data/biostats_homework_weight.csv")
```

```
## Parsed with column specification:
## cols(
##   weight = col_double(),
##   A = col_character(),
##   B = col_character()
## )
```

```
shapiro.test(rat_weight$weight[rat_weight$A=="A1"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rat_weight$weight[rat_weight$A == "A1"]
## W = 0.96118, p-value = 0.5676
```

```
shapiro.test(rat_weight$weight[rat_weight$A=="A2"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rat_weight$weight[rat_weight$A == "A2"]
## W = 0.93573, p-value = 0.1989
```

```
shapiro.test(rat_weight$weight[rat_weight$A=="A3"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rat_weight$weight[rat_weight$A == "A3"]
## W = 0.96408, p-value = 0.6281
```

```
shapiro.test(rat_weight$weight[rat_weight$B=="B1"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rat_weight$weight[rat_weight$B == "B1"]
## W = 0.95026, p-value = 0.1718
```

```
shapiro.test(rat_weight$weight[rat_weight$B=="B2"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rat_weight$weight[rat_weight$B == "B2"]
## W = 0.96797, p-value = 0.4853
```

```
bartlett.test(weight~A,data = rat_weight)
```

```
##
##  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 = rat_weight)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  weight by B
## Bartlett's K-squared = 0.20953, df = 1, p-value = 0.6471
```

答：经检验，体重数据对于因素A和因素B是正态的，并满足方差齐性的要求。

（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）

```
rat_fit<-aov(weight~A+B+A:B,data = rat_weight)
summary(rat_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
```

答:因素A对大鼠体重有显著影响

（3）对A_iB_j条件下平均产量作多重比较，并指出与A_2B_1组有显著差异的组。（10分）

```
TukeyHSD(rat_fit)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = weight ~ A + B + A:B, data = rat_weight)
##
```

```
## $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
```

答:A_1B_1，A_1B_2组和A_2B_1组有显著差异（3分）

homework-6

question-1

一、某地29名13岁儿童身高（cm），体重（kg）和肺活量（L）数据见data，求： （1）由身高，体重推算肺活量的回归方程； （2）求出的方程是否有意义； （3）剩余标准差

```
library(readxl)
homework_6_1_data <- read_excel("./data/homework-6.1-data.xlsx")
y <- homework_6_1_data$y
x1 <- homework_6_1_data$x1
x2 <- homework_6_1_data$x2
ff <- lm(y~x1+x2)
ff
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Coefficients:
## (Intercept)           x1           x2
##   -0.565664     0.005017     0.054061
```

```
summary(ff)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.54117 -0.25524 -0.00266  0.22039  0.55425
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.565664   1.240127  -0.456  0.65208
## x1           0.005017   0.010575   0.474  0.63920
## x2           0.054061   0.015984   3.382  0.00228 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3137 on 26 degrees of freedom
## Multiple R-squared:  0.546,  Adjusted R-squared:  0.511
## F-statistic: 15.63 on 2 and 26 DF,  p-value: 3.485e-05
```

答:

(1)回归方程:$y = -0.565664 + 0.005017x1 + 0.054061x2$。

(2)因为F=15.63，p=3.485e-05<0.01，所以方程有意义。(这个具体为什么用F检验，看书吧)

(3)剩余标准差(Residual standard error): 0.3137 on 26 degrees of freedom

question-2

二、某农场通过试验取得早稻收获量与春季降雨量和春季温度的数据如下:

收获量y(kg/mm2)	降雨量x1(mm)	温度x2(℃)
2250	25	6
3450	33	8
4500	45	10
6750	105	13
7200	110	14
7500	115	16
8250	120	17

建立早稻收获量对春季降雨量和春季温度的二元线性回归方程，计算各回归系数的置信区间，并对回归模型的线性关系和回归系数进行检验（α =0.05）。

```
library(readr)
homework_6_2_data <- read_delim("./data/homework-6.2-data.txt",
    "\t", escape_double = FALSE, trim_ws = TRUE)
```

```
## Parsed with column specification:
## cols(
##   y = col_double(),
##   x1 = col_double(),
##   x2 = col_double()
## )
```

```
y <- homework_6_2_data$y
x1 <- homework_6_2_data$x1
x2 <- homework_6_2_data$x2
ff <- lm(y~x1+x2)
ff
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Coefficients:
## (Intercept)           x1           x2
##      -0.591       22.386      327.672
```

```
# 置信区间
confint(ff)
```

```
##                    2.5 %     97.5 %
## (Intercept) -1402.707516 1401.52552
## x1             -4.268921   49.04184
## x2             53.364699  601.97873
```

降雨量x1的置信区间为（-4.268921，49.04184），含义是在温度不变的条件下，降雨量每变动1mm，收获量的平均变动在-4.268921到49.04184 kg/mm2之间。

温度x2的置信区间为（53.364699，601.97873），含义是在降雨量不变的条件下，温度每变动1℃，收获量的平均变动在53.364699到601.97873 kg/mm2之间。

```
summary(ff)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##        1        2        3        4        5        6        7
## -275.101   90.464  216.483  140.280  150.676 -316.599   -6.203
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.591    505.004  -0.001   0.9991
## x1            22.387      9.601   2.332   0.0801 .
## x2           327.672     98.798   3.317   0.0295 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 261.4 on 4 degrees of freedom
## Multiple R-squared:  0.9913, Adjusted R-squared:  0.987
## F-statistic: 228.4 on 2 and 4 DF,  p-value: 7.532e-05
```

线性关系检验是检验因变量y与k个自变量之间的关系是否显著，也称总体显著性检验。具体步骤如下：

第1步 提出假设：

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \tag{12}$$

$$H_1 : \beta_1, \beta_2, \ldots, \beta_k \text{至少有一个不等于} 0 \tag{13}$$

第2步 计算检验统计量F：

$$F = \frac{SSR/k}{SSE/(n-k-1)} \sim F(k, n-k-1) \tag{14}$$

第3步 做出决策。给定显著水平α，根据分子的自由度=k，分母的自由度=$n-k-1$计算出统计量的P值。若P< α，拒绝原假设，表明y与k个自变量之间的线性关系显著。 根据以上R输出结果，检验统计量F= 228.4，显著水平P= 7.532e-05< 0.05，拒绝H_0，即收获量y与降雨量x1和温度x2之间的线性关系显著。

4、要判断每个自变量对因变量的影响是否都显著，需要对各回归系数β_i分别进行t检验，具体步骤如下： 第1步 提出假设。对于任意参数$\beta_i (i = 1, 2, \ldots, k)$，有

$$H_0 : \beta = 0, H_1 : \beta \neq 0 \tag{15}$$

第2步 计算检验的统计量t:

$$t_i = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}} \sim t(n-k-1) \tag{16}$$

其中：是回归系数的抽样分布的标准差。 第3步 做出决策。给定显著性水平α，根据自由度= n-k-1计算出统计量的P值。若P< α，则拒绝原假设，表明回归系数β_i显著。 根据R输出结果，降雨量x1和温度x2的回归系数相应的显著水平分别为0.0801和0.0295，只有温度对应的显著性水平小于0.05通过检验，这表明影响收获量的自变量中，只有温度对收获量的影响显著，而降雨量对收获量的影响不显著。

question-3

三、某葡萄酒爱好者想探索葡萄酒的品质与哪些因素相关。他有一个数据集包含了（1 -固定酸度，2 -挥发性酸度，3 -柠檬酸，4 -残余糖，5 -氯化物，6 -自由二氧化硫量，7 -二氧化硫总量，8 -密度，9 - pH值，10 -硫酸盐，11 -酒精浓度，和12 -品质(0 - 10分)。

1.查看数据集的前五行和数据集的总结。

```
library(readr)
winequality <- read_csv("./data/homework-6.3-winequality-red.csv")
```

```
## Parsed with column specification:
## cols(
##   `fixed acidity` = col_double(),
##   `volatile acidity` = col_double(),
##   `citric acid` = col_double(),
##   `residual sugar` = col_double(),
##   chlorides = col_double(),
##   `free sulfur dioxide` = col_double(),
##   `total sulfur dioxide` = col_double(),
##   density = col_double(),
##   pH = col_double(),
##   sulphates = col_double(),
##   alcohol = col_double(),
##   quality = col_double()
## )
```

```
head(winequality,n=5)
```

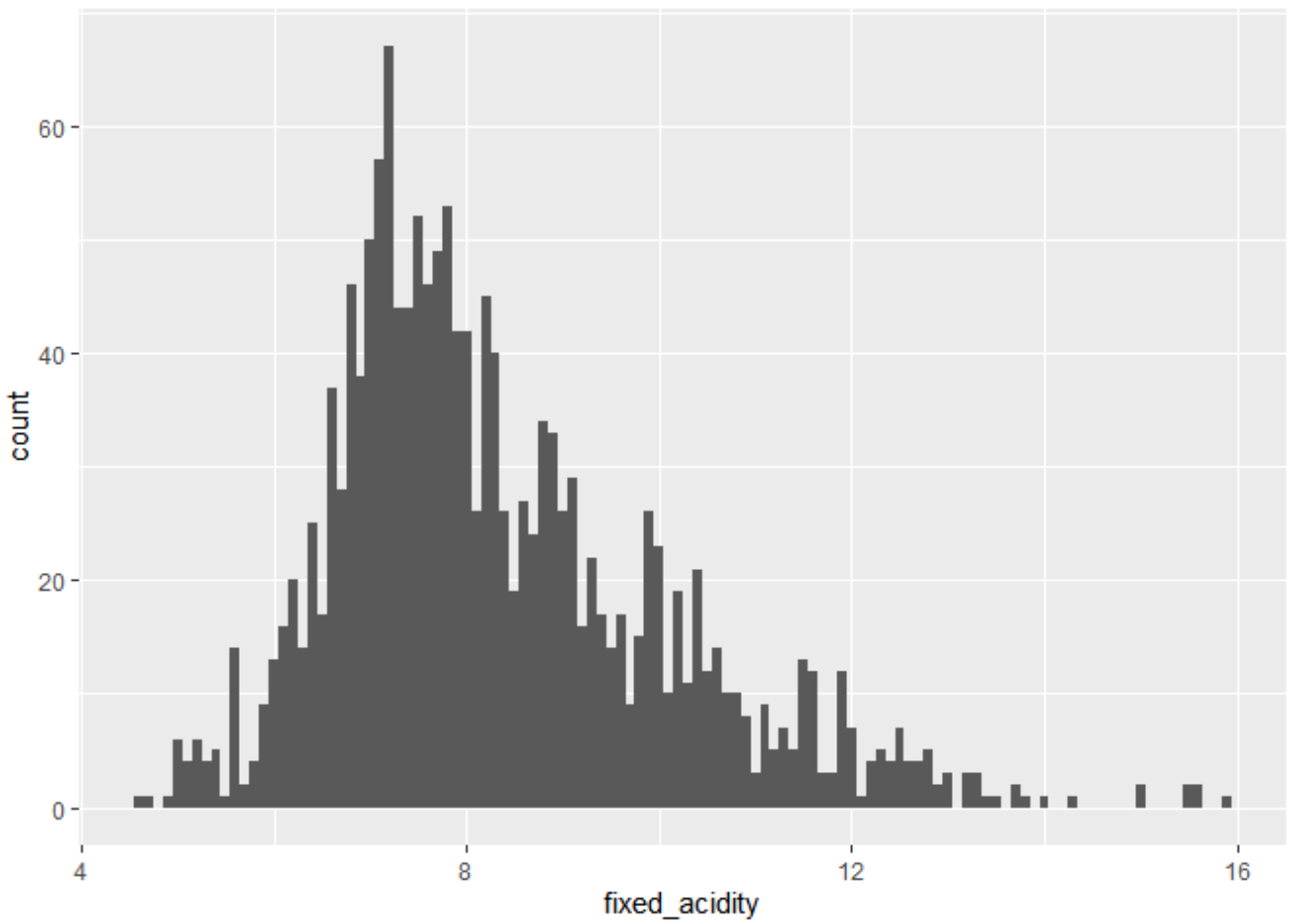
```
## # A tibble: 5 x 12
##   `fixed acidity` `volatile acidi~ `citric acid` `residual sugar` chlorides
##             <dbl>            <dbl>         <dbl>            <dbl>     <dbl>
## 1             7.4             0.7           0                 1.9     0.076
## 2             7.8             0.88          0                 2.6     0.098
## 3             7.8             0.76          0.04              2.3     0.092
## 4            11.2             0.28          0.56              1.9     0.075
## 5             7.4             0.7           0                 1.9     0.076
## # ... with 7 more variables: `free sulfur dioxide` <dbl>, `total sulfur
## #   dioxide` <dbl>, density <dbl>, pH <dbl>, sulphates <dbl>,
## #   alcohol <dbl>, quality <dbl>
```

```
summary(winequality)
```

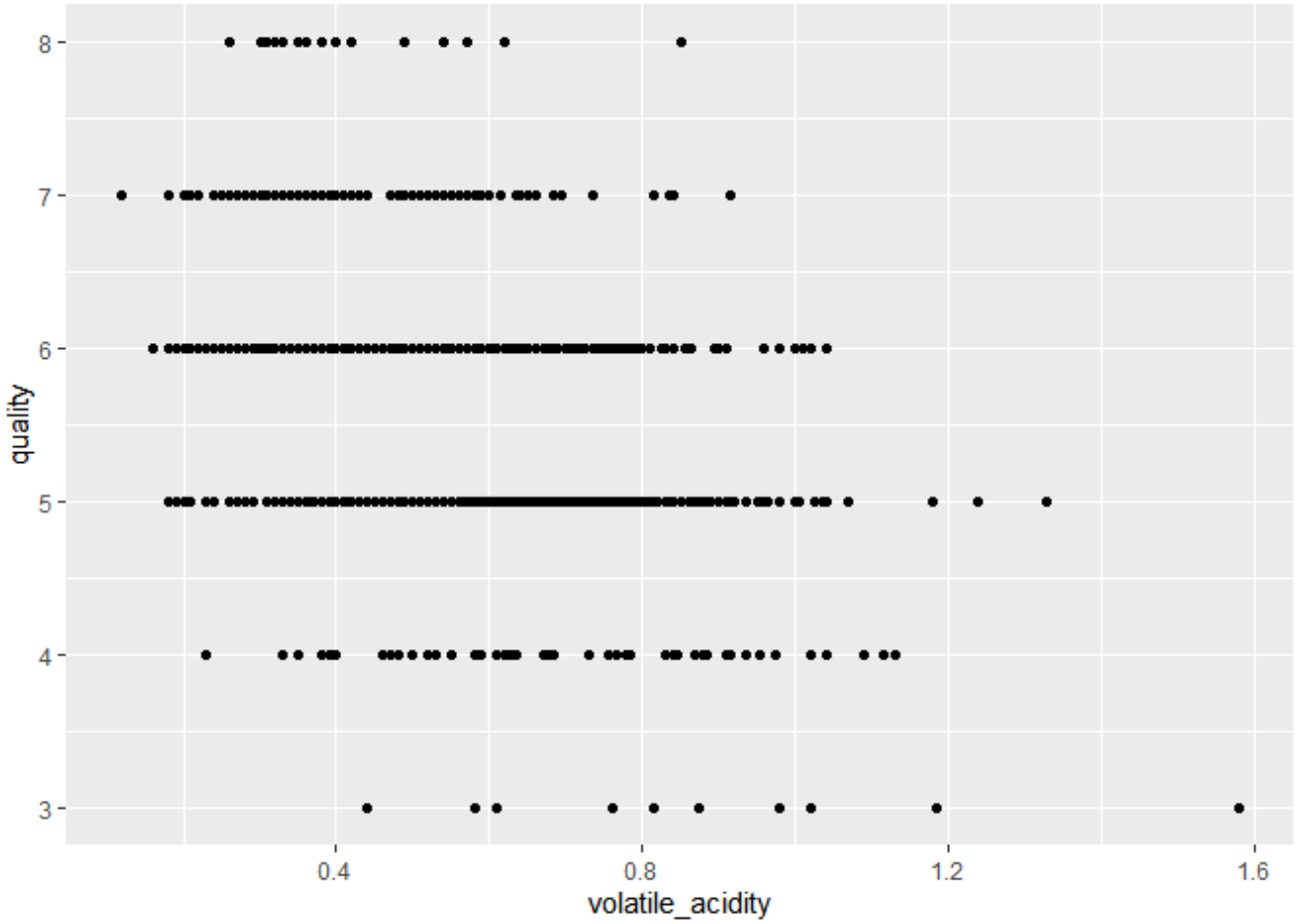
```
##  fixed acidity   volatile acidity  citric acid    residual sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chlorides       free sulfur dioxide total sulfur dioxide
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00
##  Median :0.07900   Median :14.00       Median : 38.00
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00
##     density             pH          sulphates         alcohol
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50
##  Median :0.9968   Median :3.310   Median :0.6200   Median :10.20
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.636
##  3rd Qu.:6.000
##  Max.   :8.000
```

2.通过直方图展示固定酸度的分布和展示挥发性酸度与品质的散点图。

```
library(ggplot2)
fixed_acidity <- winequality$`fixed acidity`
hist=ggplot(winequality,aes(x=fixed_acidity))+geom_histogram(stat="bin",binwidth = 0.1)
hist
```




```r
volatile_acidity <- winequality$`volatile acidity`
quality <- winequality$quality
point <- ggplot(winequality,mapping = aes(x=volatile_acidity,y=quality))+geom_point()
point
```



3.计算这些变量与品质的相关性。

```
quality_cor <- cor(winequality[1:11],winequality$quality)
quality_cor
```

```
##                              [,1]
## fixed acidity         0.12405165
## volatile acidity     -0.39055778
## citric acid           0.22637251
## residual sugar        0.01373164
## chlorides            -0.12890656
## free sulfur dioxide  -0.05065606
## total sulfur dioxide -0.18510029
## density              -0.17491923
## pH                   -0.05773139
## sulphates             0.25139708
## alcohol               0.47616632
```

4.通过方差分析不同品质的葡萄酒的酒精浓度是否有差异。

```
alcohol <- winequality$alcohol
alcohol_anova <- data.frame(alcohol,quality)
# 方差齐性检验
bartlett.test(alcohol_anova)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  alcohol_anova
## Bartlett's K-squared = 121.32, df = 1, p-value < 2.2e-16
```

```
# 正态检验
for (i in c(3:8)) {
  print(shapiro.test(alcohol_anova$alcohol[alcohol_anova$quality==i]))
}
```

```
##
##  Shapiro-Wilk normality test
##
## data:  alcohol_anova$alcohol[alcohol_anova$quality == i]
## W = 0.9423, p-value = 0.5788
##
##
##  Shapiro-Wilk normality test
##
## data:  alcohol_anova$alcohol[alcohol_anova$quality == i]
## W = 0.93444, p-value = 0.00607
##
##
##  Shapiro-Wilk normality test
##
## data:  alcohol_anova$alcohol[alcohol_anova$quality == i]
## W = 0.84302, p-value < 2.2e-16
##
##
##  Shapiro-Wilk normality test
##
## data:  alcohol_anova$alcohol[alcohol_anova$quality == i]
## W = 0.96945, p-value = 2.885e-10
##
##
##  Shapiro-Wilk normality test
##
## data:  alcohol_anova$alcohol[alcohol_anova$quality == i]
## W = 0.99166, p-value = 0.3108
##
##
##  Shapiro-Wilk normality test
##
## data:  alcohol_anova$alcohol[alcohol_anova$quality == i]
## W = 0.96336, p-value = 0.6676
```

```r
#这部分有的没有通过正态性检验，答案中没有考虑这一点，大家见仁见智吧。
#H0:不同品质的葡萄酒的酒精浓度没有差异；HA:不同品质的葡萄酒的酒精浓度有显著差异
fit <- aov(winequality$alcohol~winequality$quality)
summary(fit)
```

```
##                       Df Sum Sq Mean Sq F value Pr(>F)
## winequality$quality    1  411.5   411.5   468.3 <2e-16 ***
## Residuals           1597 1403.3     0.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

答：不同品质的葡萄酒的酒精浓度有显著差异

5.通过多元线性回归建立一个品质预测模型，并说明哪些变量与品质显著相关。

```r
ff<-lm(quality~.,data = winequality)
summary(ff)
```

```
##
## Call:
## lm(formula = quality ~ ., data = winequality)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.68911 -0.36652 -0.04699  0.45202  2.02498
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.197e+01  2.119e+01   1.036   0.3002
## `fixed acidity`        2.499e-02  2.595e-02   0.963   0.3357
## `volatile acidity`    -1.084e+00  1.211e-01  -8.948  < 2e-16 ***
## `citric acid`         -1.826e-01  1.472e-01  -1.240   0.2150
## `residual sugar`       1.633e-02  1.500e-02   1.089   0.2765
## chlorides             -1.874e+00  4.193e-01  -4.470 8.37e-06 ***
## `free sulfur dioxide`  4.361e-03  2.171e-03   2.009   0.0447 *
## `total sulfur dioxide` -3.265e-03  7.287e-04  -4.480 8.00e-06 ***
## density               -1.788e+01  2.163e+01  -0.827   0.4086
## pH                    -4.137e-01  1.916e-01  -2.159   0.0310 *
## sulphates              9.163e-01  1.143e-01   8.014 2.13e-15 ***
## alcohol                2.762e-01  2.648e-02  10.429  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.648 on 1587 degrees of freedom
## Multiple R-squared:  0.3606, Adjusted R-squared:  0.3561
## F-statistic: 81.35 on 11 and 1587 DF,  p-value: < 2.2e-16
```

挥发性酸度，氯化物，自由二氧化硫量，二氧化硫总量，pH值，硫酸盐和酒精浓度与品质显著相关。

question 4

四、A study was performed to investigate new automobile purchases. A sample of 20 families was selected. Each family was surveyed to determine the age of their oldest vehicle and their total family income. A follow-up survey was conducted six months later to determine if they had actually purchased a new vehicle during that time period (y = 1 indicates yes and y = 0 indicates no). The data from this study are shown in the Table1.

(a) Fit a logistic regression model to the data.(40')

(b) Interpret the model coefficients β1 and β2 and write the logistic regression model formula.(20')

(c) What is the estimated probability that a family with an income of $45,000 and a car that is five years old will purchase a new vehicle in the next six months?(40')

```
income_age <- read.csv("./data/homework-6.4-data.txt")
fit<-glm(y~.,data = income_age,family = binomial())
summary(fit)
```

```
##
## Call:
## glm(formula = y ~ ., family = binomial(), data = income_age)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5635  -0.8045  -0.1397   0.9535   1.7915
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.047e+00  4.674e+00  -1.508    0.132
## Income       7.382e-05  6.371e-05   1.159    0.247
## Age          9.879e-01  5.274e-01   1.873    0.061 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27.726  on 19  degrees of freedom
## Residual deviance: 21.082  on 17  degrees of freedom
## AIC: 27.082
##
## Number of Fisher Scoring iterations: 5
```

```
coef(fit)
```

```
##   (Intercept)        Income           Age
## -7.047061e+00  7.381679e-05  9.878861e-01
```

```
test_data<-data.frame(Income=45000,Age=5)
test_data$probe<-predict(fit,newdata = test_data,type = "response")
test_data
```

```
##   Income Age     probe
## 1  45000   5 0.7710279
```

答：

公式：$odds = exp(-7.047061 + 7.381679e-05x1 + 9.878861e-01x2)$

question 5

五、数据文件“Drivers.csv”为对45名司机的调查结果，其中四个变量的含义为：

1）x1：表示视力状况，它是一个分类变量，1表示好，0表示有问题；

2）x2：年龄，数值型；

3）x3：驾车教育，它也是一个分类变量，1表示参加过驾车教育，0表示没有；

4）y：一个分类型输出变量，表示去年是否出过事故，1表示出过事故，0表示没有；

问题：

（1）请在R语言中调用logistic回归函数，计算视力状况、年龄、驾车教育与是否发生事故的logistic回归模型，并以“odds=……”的形式写出回归公式。（10分）

（2）指出（1）得到的模型中哪些因素对是否发生事故有显著性影响。如果存在对是否发生事故没有显著性影响的因素，请去除这些因素后重新计算logistic回归模型，并以“p=……”的形式写出回归公式。（20分）

（3）A是一名参加过驾车教育，但视力有问题的50岁老司机；B是一名没有参加过驾车教育，但视力良好的20岁新手。现在A、B都想在某保险公司投保，但按公司规定，被保险人必须满足“明年出事故的概率不高于40%”的条件才能予以承保。请预测A、B两者明年出事故的概率，并告诉保险公司谁可以投保。（20分）

```
library(readr)
drivers <- read_csv("./data/homework-6.5-Drivers.csv")
```

```
## Parsed with column specification:
## cols(
##   x1 = col_double(),
##   x2 = col_double(),
##   x3 = col_double(),
##   y = col_double()
## )
```

```
fit.full<-glm(y~., data=drivers, family=binomial)
```

回归公式为：$odds = exp(0.597610 - 1.496084x1 - 0.001595x2 + 0.315865x3)$

```
fit_x1<-glm(y~x1, data=drivers, family=binomial)
summary(fit_x1)
```

```
##
## Call:
```

```
## glm(formula = y ~ x1, family = binomial, data = drivers)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4490  -0.8782  -0.8782   0.9282   1.5096
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.6190     0.4688   1.320   0.1867
## x1           -1.3728     0.6353  -2.161   0.0307 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 62.183  on 44  degrees of freedom
## Residual deviance: 57.241  on 43  degrees of freedom
## AIC: 61.241
##
## Number of Fisher Scoring iterations: 4
```

回归公式为$p = exp(0.6190 - 1.3728x1)/(1 + exp(0.6190 - 1.3728x1))$

```
test_data<- data.frame(x1=c(0,1))
test_data$probe<-predict(fit_x1,test_data ,type='response')
test_data
```

```
##   x1 probe
## 1  0  0.65
## 2  1  0.32
```

答：所以A、B两者明年出事故的概率分别为0.65和0.32,因只有B明年出事故的概率不高于40%，故只有B可以投保。

question 6

六、Many digitized image of a fine needle aspirate (FNA) of a breast mass are collected and computed to predict the diagnosis of breast cancer(data.csv).

Attribute information

1) ID number

2) Diagnosis (M = malignant, B = benign)

Ten real-valued features are computed for each cell nucleus:

```
a) radius (mean of distances from center to points on the perimeter)

b) texture (standard deviation of gray-scale values)

c) perimeter

```

```
d) area

e) smoothness (local variation in radius lengths)

f) compactness (perimeter^2 / area - 1.0)

g) concavity (severity of concave portions of the contour)

h) concave points (number of concave portions of the contour)

i) symmetry

j) fractal dimension ("coastline approximation" - 1)
```

The mean, standard error, and "worst" or largest (mean of the threelargest values) of these features were computed for each image,resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

All feature values are recorded with four significant digits. In total, there are 357 benign and 212 malignant samples.

You may need to use proper regression algorithm to train your data, and make predictions.

Instructions:

1)Use all mean features(such as: radius_mean,texture_mean...) to construct a logistic regression model

```
library(readr)
breast_cancer <- read_csv("./data/homework-6.6-data.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   diagnosis = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
labels <- breast_cancer$diagnosis
labels[labels=="M"] <- 1
labels[labels=="B"] <- 0
labels <- as.integer(labels)
breast_cancer$lables <- labels
fit.full <-
glm(lables~radius_mean+texture_mean+perimeter_mean+area_mean+smoothness_mean+compactnes
s_mean+concavity_mean+`concave points_mean`+symmetry_mean+fractal_dimension_mean,data =
breast_cancer,family = binomial(),control=list(maxit=100))
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(fit.full)
```

```
##
## Call:
## glm(formula = lables ~ radius_mean + texture_mean + perimeter_mean +
##     area_mean + smoothness_mean + compactness_mean + concavity_mean +
##     `concave points_mean` + symmetry_mean + fractal_dimension_mean,
##     family = binomial(), data = breast_cancer, control = list(maxit = 100))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.95590  -0.14839  -0.03943   0.00429   2.91690
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -7.35952   12.85259  -0.573   0.5669
## radius_mean              -2.04930    3.71588  -0.551   0.5813
## texture_mean              0.38473    0.06454   5.961  2.5e-09 ***
## perimeter_mean           -0.07151    0.50516  -0.142   0.8874
## area_mean                 0.03980    0.01674   2.377   0.0174 *
## smoothness_mean          76.43227   31.95492   2.392   0.0168 *
## compactness_mean         -1.46242   20.34249  -0.072   0.9427
## concavity_mean            8.46870    8.12003   1.043   0.2970
## `concave points_mean`    66.82176   28.52910   2.342   0.0192 *
## symmetry_mean            16.27824   10.63059   1.531   0.1257
## fractal_dimension_mean -68.33703   85.55666  -0.799   0.4244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 751.44  on 568  degrees of freedom
## Residual deviance: 146.13  on 558  degrees of freedom
## AIC: 168.13
##
## Number of Fisher Scoring iterations: 9
```

> Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred:这个问题实际是因为数据集问题，这个数据集本身接近线性可分了，所以导致模型过拟合。具体原因大家自己检索一下吧。可以看看这篇博客：https://www.cnblogs.com/runner-ljt/p/4574275.html

2)Then try to reduce the number of features from your last model, construct another regression model, and you will need to write down the equation of your logistic regression model(Tips: $LogitP = \alpha + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_p X_p$)

除去不显著相关的变量。

```
fit.reduced<-glm(lables~texture_mean+area_mean+smoothness_mean+`concave
points_mean`,data = breast_cancer,family = binomial(),control=list(maxit=100))
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(fit.reduced)
```

```
##
## Call:
## glm(formula = lables ~ texture_mean + area_mean + smoothness_mean +
##     `concave points_mean`, family = binomial(), data = breast_cancer,
##     control = list(maxit = 100))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.31798  -0.15623  -0.04212   0.01662   2.84201
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -23.677816   3.882774  -6.098 1.07e-09 ***
## texture_mean            0.362687   0.060544   5.990 2.09e-09 ***
## area_mean               0.010342   0.002002   5.165 2.40e-07 ***
## smoothness_mean        59.471304  25.965153   2.290    0.022 *
## `concave points_mean`  76.571210  16.427864   4.661 3.15e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 751.44  on 568  degrees of freedom
## Residual deviance: 156.44  on 564  degrees of freedom
## AIC: 166.44
##
## Number of Fisher Scoring iterations: 8
```

公式：$LogitP = -23.677816 + 0.362687x2 + 0.010342x4 + 59.47130x5 + 76.571210x8$ 3)Use proper test to test the difference between two models

对两个模型进行ANOVA分析。

```
#H0:两模型无显著差别；HA:两模型有显著差别
anova(fit.full,fit.reduced,test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: lables ~ radius_mean + texture_mean + perimeter_mean + area_mean +
##     smoothness_mean + compactness_mean + concavity_mean + `concave points_mean` +
##     symmetry_mean + fractal_dimension_mean
## Model 2: lables ~ texture_mean + area_mean + smoothness_mean + `concave points_mean`
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       558     146.13
## 2       564     156.44 -6   -10.31   0.1122
```

p=0.1122>0.05，两模型无显著差别。

4)You may split the data properly, use part of them to train your regression model and use another part to make predictions. Lastly, you may try to calculate the accuracy of your model.(Tips: To split the data, you can use the first 398 rows as training data, use the last 171 rows as prediction data.The predict function return a value between 0 and 1, 0.~0.5 belong to the first class, and 0.5~1 belong to second class in binary classification problems)

```
train_data <- breast_cancer[1:398,]
test_data <- breast_cancer[399:569,]
train_fit<-glm(lables~texture_mean+area_mean+smoothness_mean+`concave points_mean`,data = train_data,family = binomial())
summary(train_fit)
```

```
##
## Call:
## glm(formula = lables ~ texture_mean + area_mean + smoothness_mean +
##     `concave points_mean`, family = binomial(), data = train_data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.39278  -0.14454  -0.02447   0.03635   2.60665
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -27.47397    4.74798  -5.786 7.19e-09 ***
## texture_mean            0.46244    0.08434   5.483 4.19e-08 ***
## area_mean               0.01082    0.00235   4.606 4.11e-06 ***
## smoothness_mean        90.11221   30.96961   2.910 0.003618 **
## `concave points_mean`  59.01212   17.51779   3.369 0.000755 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 544.93  on 397  degrees of freedom
## Residual deviance: 108.30  on 393  degrees of freedom
## AIC: 118.3
##
## Number of Fisher Scoring iterations: 8
```

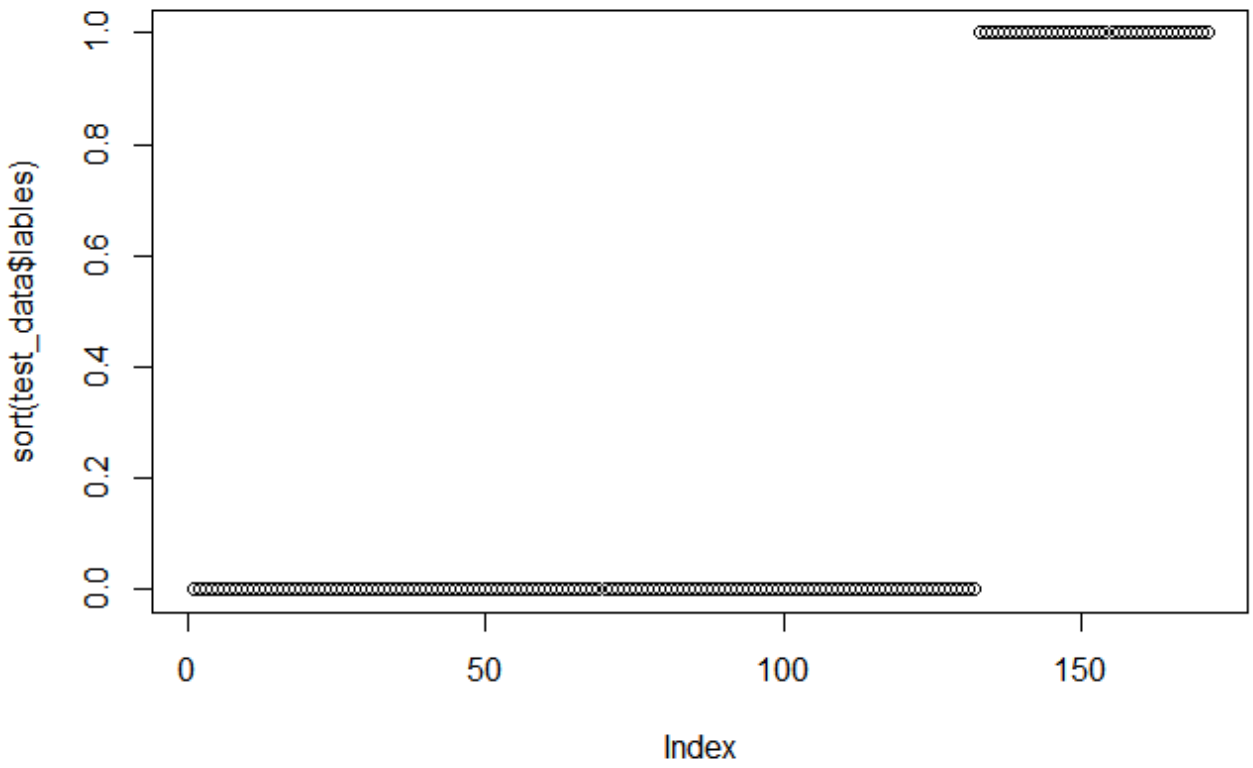
```
test_data$probe<-predict(train_fit,newdata = test_data,type = "response")
lables <- test_data$lables
pred_lables <- ifelse(test_data$probe > 0.5, 1, 0)
mean(pred_lables==lables)
```

```
## [1] 0.9064327
```

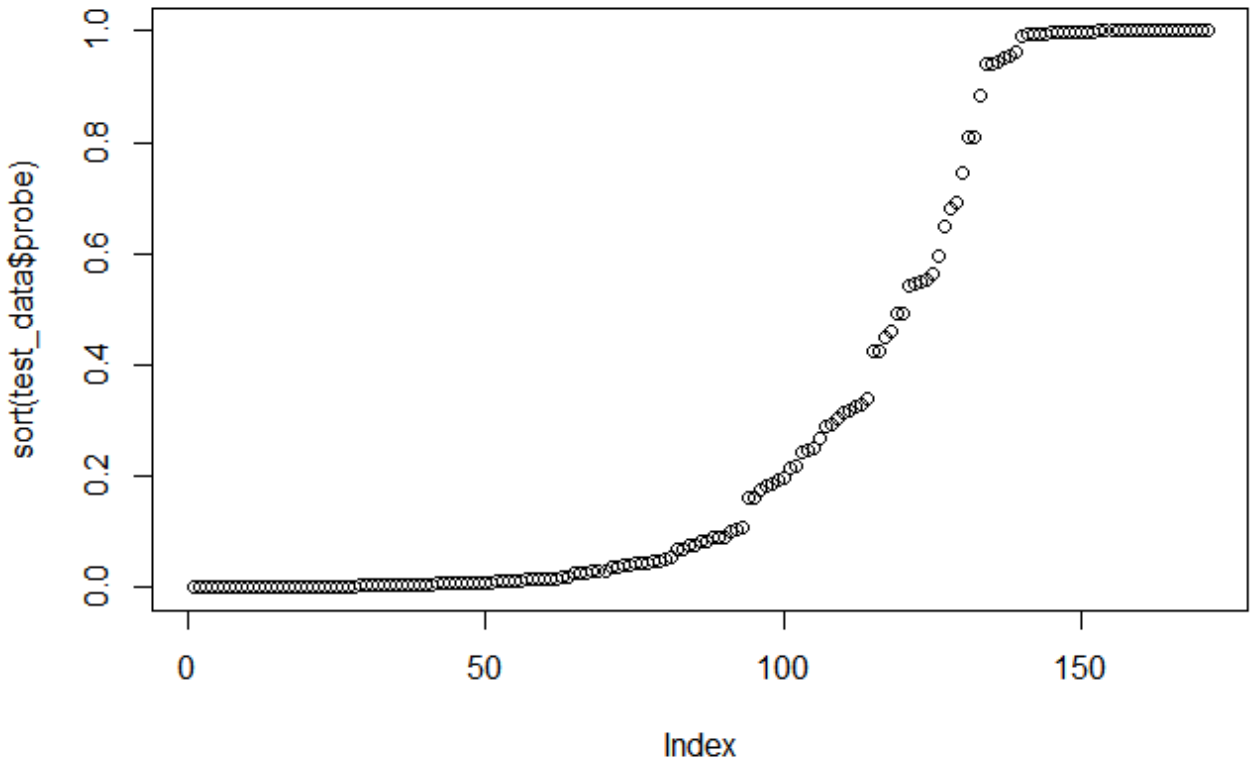
The accuracy is 0.9064

进一步地，可以看看它们的预测情况和ROC曲线以及AUC值。

```
plot(sort(test_data$lables))
```




```
plot(sort(test_data$probe))
```




```r
# install.packages("pROC")
library('pROC')
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

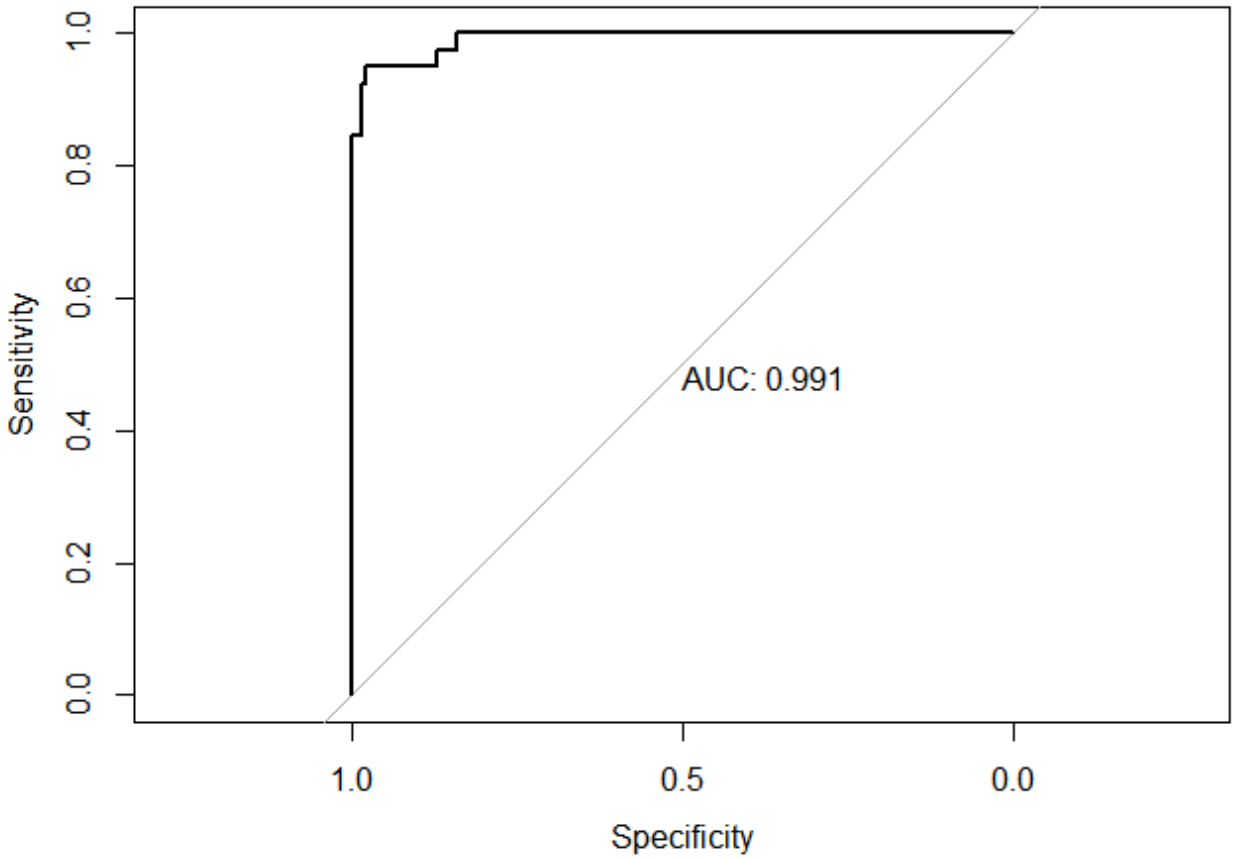
```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```r
model_roc <- roc(test_data$lables,test_data$probe)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```r
plot(model_roc,print.auc = T)
```



AUC = 0.991，可以说明分类效果非常好。