# Chapter14

*Dai Shaoqing*

*2017.10.8*

- **Name**: Note of Applied Statistics with R(Chapter 14 demo code) Case and Practice
- **Purpose**: Case and practice
- **Author**: Dai shaoqing
- **Created**: 10/08/2017
- **Copyright**: (c) Dai shaoqing dsq1993qingge@163.com 2017

```
#load library
library(openxlsx)
library(ggplot2)
library(psych)
library(gridExtra)
```

## 1 描述性统计与抽样分布

1.

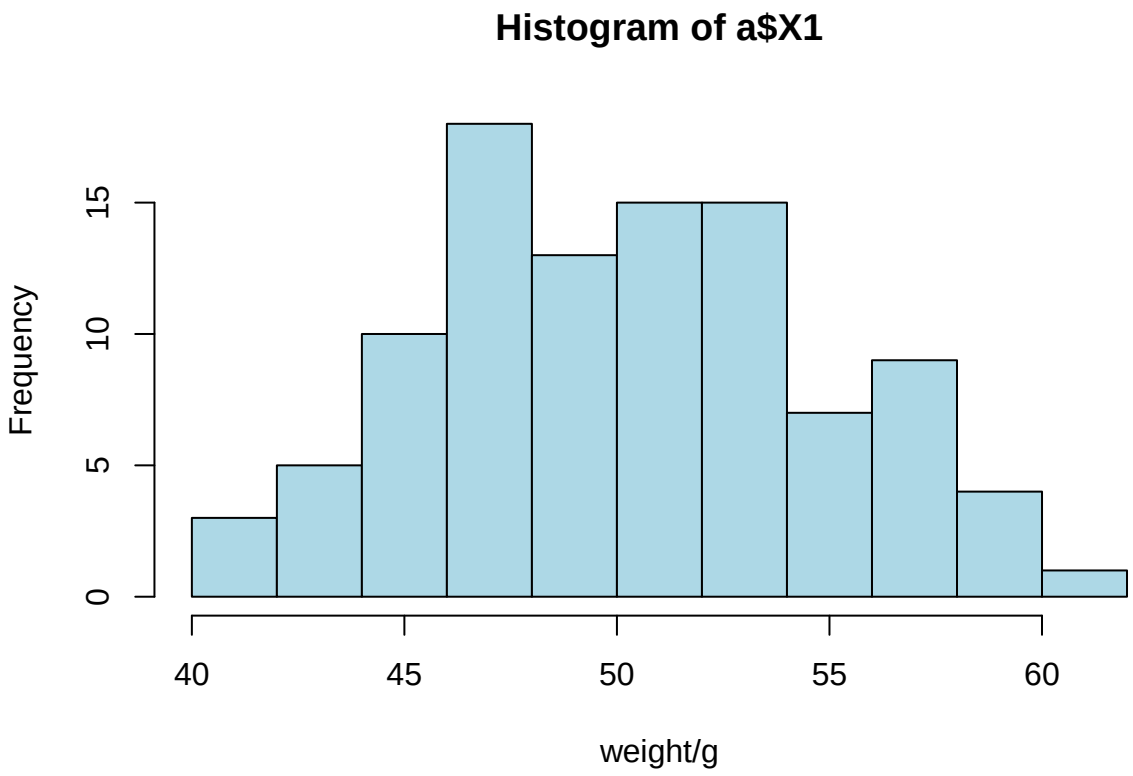
（1）频数分布表

```
#Input data
a<-read.xlsx("F:/R/applicationstatics/data/exercise1.xlsx",sheet = 1,colNames = F)
table(a)
```

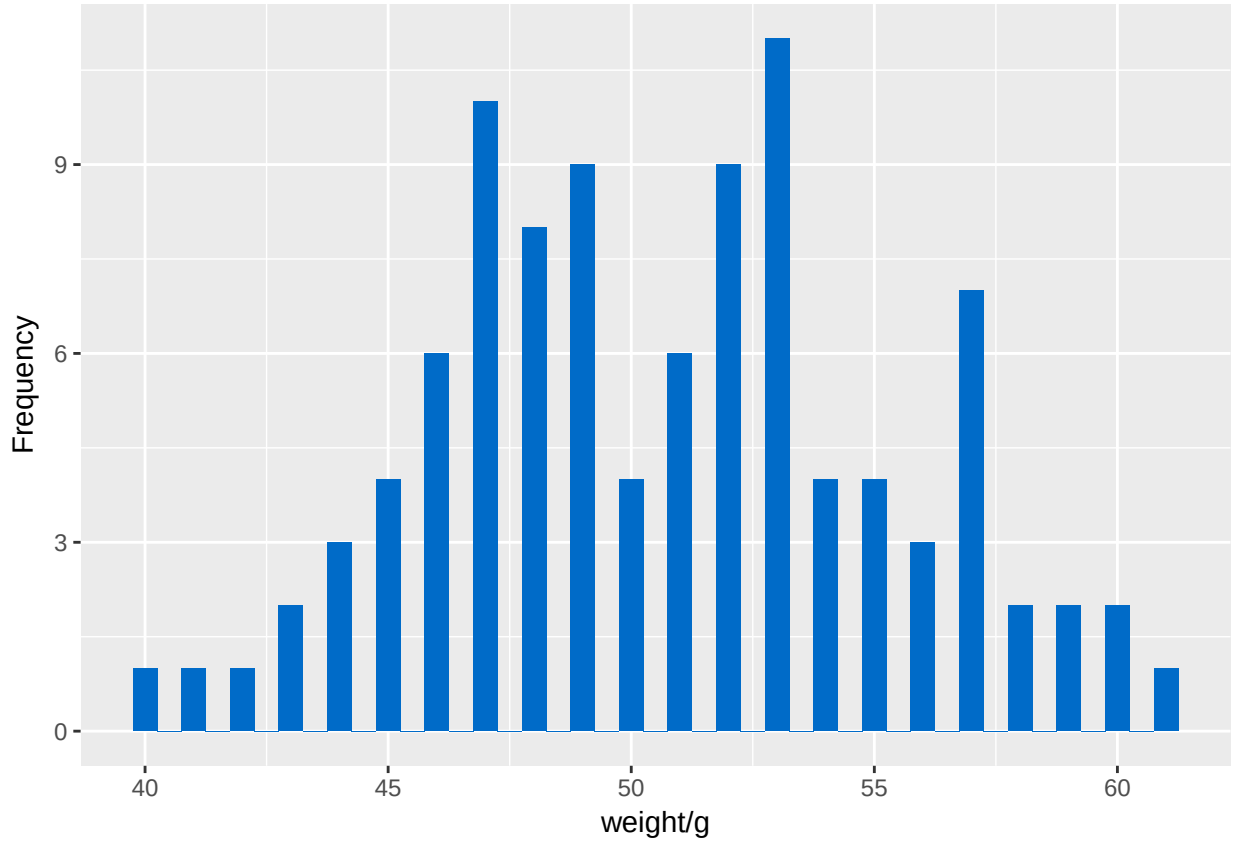
```
## a
## 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61
##  1  1  1  2  3  4  6 10  8  9  4  6  9 11  4  4  3  7  2  2  2  1
```

（2）频数分布图

```
hist(a$X1,col="lightblue",xlab="weight/g")
```




```
ahist<-ggplot(a)+geom_histogram(mapping = aes(a$X1),fill=rgb(red = 0, green = 107, blue = 200, max = 255
ahist
```




（3）数据整体呈一个“双峰”分布。而且刚好 50 g 的食品非常少。大部分集中在 47 和 53 附近。

2.

（1）

```
#Input data and clean data
b<-read.xlsx("F:/R/applicationstatics/data/exercise1.xlsx",sheet = 2)
b<-b[,-c(1:9)]
b<-data.frame(gl=b[,1],p=b[,2],c=b[,3])
```

(2) 甲班中等成绩的人最多，而优良成绩的人比不及格和及格的人少。乙班成绩为良的最多，而且不及格人数与及格人数均比甲班少。仅有中等成绩的人比甲班少，其他均多于甲班。

(3)

甲乙两个班成绩分布差异较大。甲班中等成绩人居多，而且相比较而言，中等成绩人数量十分突出。而乙班则较为均衡，良成绩的人较少些。

3.

```
c<-read.xlsx("F:/R/applicationstatics/data/exercise1.xlsx",sheet = 3)
describe(c)
```

```
##          vars  n mean   sd median trimmed  mad min max range skew kurtosis
## 网民年龄    1 25   24 6.65     23   23.33 5.93  15  41    26 0.95     0.13
##            se
## 网民年龄 1.33
```

```
table(c)
```

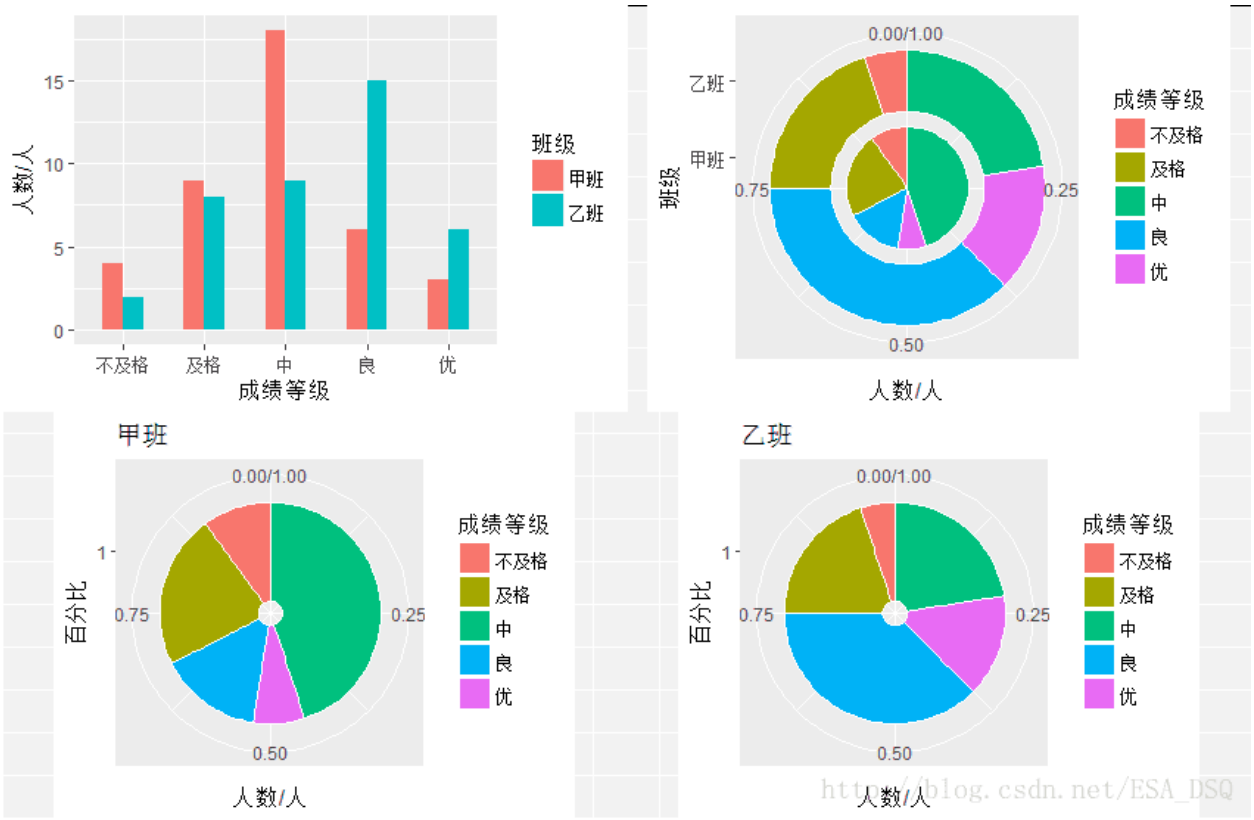

人数/人
15
10
5
0
不及格
及格
中
良
优
成绩等级
班级
甲班
乙班
0.00/1.00
0.25
0.50
0.75
1
百分比
成绩等级
不及格
及格
中
良
优
甲班
乙班
人数/人
https://blog.csdn.net/ESA_DSQ

Figure 1:

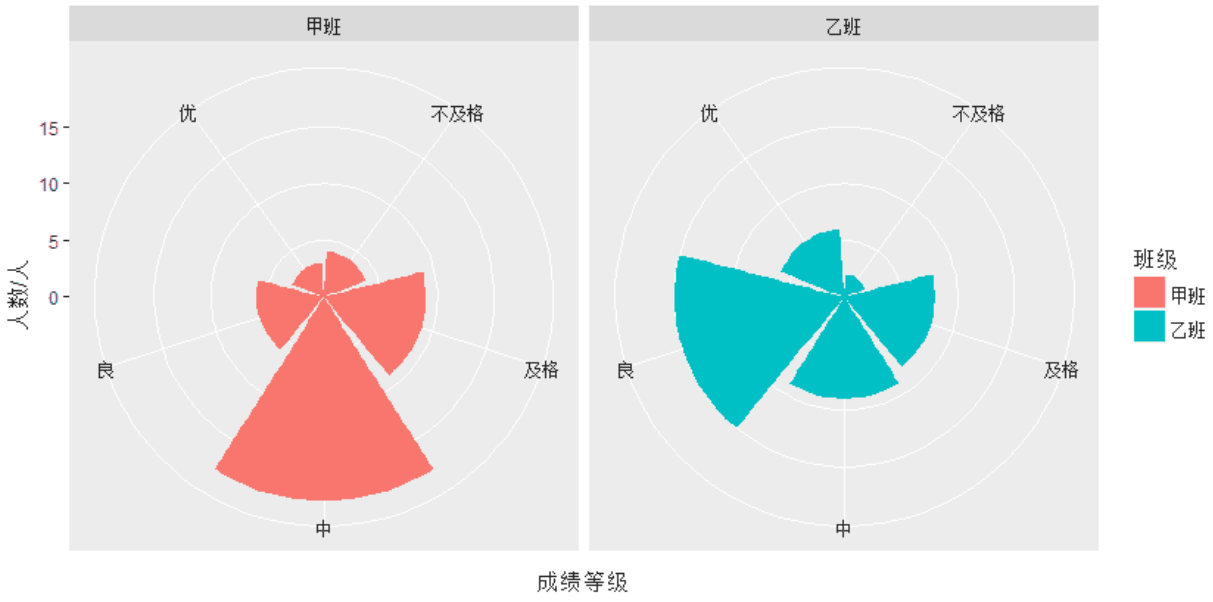





Figure 2:

```
## c
## 15 16 17 18 19 20 21 22 23 24 25 27 29 30 31 34 38 41
##  1  1  1  1  3  2  1  2  3  2  1  1  1  1  1  1  1  1
```

（1）众数：19 和 23、中位数：23

（2）四分位数：19（上四分位数）、27（下四分位数）

（3）平均数：24、标准差：6.65

（4）偏态系数：0.95、峰态系数：0.13

（5）网民整体分布呈现一个右偏的尖峰分布，但是平均数与中位数较为接近。整体分布还是较为平稳。

4.

```
d<-read.xlsx("F:/R/applicationstatics/data/exercise1.xlsx",sheet = 4)
stem(d[,1])
```

```
##
##   The decimal point is at the |
##
##   5 | 5
##   6 |
##   6 | 678
##   7 | 134
##   7 | 88
```

```
summary(d)
```

```
##     排队时间
##  Min.   :5.5
##  1st Qu.:6.7
##  Median :7.1
##  Mean   :7.0
##  3rd Qu.:7.4
##  Max.   :7.8
```

```
describe(d)
```

```
##          vars n mean   sd median trimmed  mad min max range  skew kurtosis
## 排队时间    1 9    7 0.71    7.1       7 0.59 5.5 7.8   2.3 -0.72    -0.46
##            se
## 排队时间 0.24
```

（1）茎叶图 5 | 5 6 | 6 | 678 7 | 134 7 | 88

（2）平均数：7.0，标准差 0.71。

（3）第一种方式标准差要远大于第二种方式，所以第一种方式离散程度较大。

（4）我会选择第二种，首先，第二种平均等待时间小于第一种，同时标准差则远小于第一种。也就是说明平均的等待时间小于第一种，同时等待时间也不会偏离 7 分钟太多。

5.

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

（a）首先认为 n=100 的情况下属于大样本，可以认为近似正态分布，所以重复抽样的样本均值的抽样分布也遵循正态分布，所以样本均值抽样分布的期望值为 200，方差为 25

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\frac{N-n}{N-1} n \geq 5\%N$$

（b）不重复抽样的样本均值的抽样分布同样遵循近似正态分布，总体样本为 10000 和 1000 时，简单随机样本的样本量 n=100，5%N=500 和 50，所以当总体样本为 10000 时样本均值抽样分布的期望为为 200，方差为 24.75。而当总体样本仅为 1000 时，不满足 n 5%N 的条件，可以按重复抽样计算样本均值的抽样分布：也就是期望值为 200，方差为 25。

**2** 参数估计与假设检验

1. 样本数 n=36>30，可以认为大样本数据非正态分布，且总体的均值未知，因此，采用 z 分布计算置信区间，样本均值为 3.317，标准差为 1.609，置信区间计算公式为：

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

分别带入计算可得。90% 置信概率的置信区间为 [2.863,3.770]，95% 置信概率的置信区间为 [2.772,3.861]，99% 置信概率的置信区间为 [2.586,4.047]。

2. 总体均值之差估计（且 n1=n2，总体标准差已知）所需样本容量的公式为：

$$n = \frac{(z_{\alpha/2})^2 \cdot (\sigma_1^2 + \sigma_2^2)}{E^2} E = z_{\alpha/2}\sqrt{\frac{(\sigma_1^2 + \sigma_2^2)}{n}}$$

误差范围不超过 5，即将 E=5 带入，即可得到 n 的最小值，即 n=56.700，即 n=57。

3. 假设 $H_0$： =0.618，备择假设 $H_1$： 0.618。该问题为总体方差未知的正态小样本均值检验。故选用 t 分布检验统计量。

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t(n-1)$$

可以得到 t=1.932318，而显著性水平 =0.05 的 t 分布临界值为 2.093024。因为 t $<$ 2.093024，所以拒绝 H0，无法认为该工厂生产的工艺品框架宽与长度的平均比例为 0.618。

4.

（1）第一类错误是弃真错误，也就是原假设为真，却拒绝了原假设。

（2）第二类错误是取伪错误，也就是原假设为假，但未拒绝原假设。

（3）连锁店的顾客们会将取伪错误看得较为严重，因为顾客肯定希望能获得更多的利益，也就是说希望土豆片比 60 克多，如果商家检验结果是取伪错误——就是事实上，土豆片不到 60 克，但是检验结果却是大于 60 克。而供应商则会将弃真错误看得较为严重，因为对供应商来说，土豆片少一点，相当于材料费少了些，对于他们收益是好的，所以他们希望的是土豆片比 60 克少或者刚好 60 克，如果商家检验结果是弃真错误——就是事实上，土豆片是大于 60 克的，但是检验结果却是小于 60 克。

相关代码及自编假设检验函数。

```
a<-read.xlsx("F:/R/applicationstatics/data/exercise2.xlsx",sheet=1)
mean(a[,1])
```

```
## [1] 3.316667
```

```
sd(a[,1])
```

```
## [1] 1.609348
```

```
# 方差已知的区间估计
conf.int=function(x,sigma,alpha) {
  mean=mean(x)
```

```
  n=length(x)
  z=qnorm(1-alpha/2,mean=0,sd=1,lower.tail = T)
  c(mean-sigma*z/sqrt(n),mean+sigma*z/sqrt(n))
}
```

```
# 方差未知的区间估计
t.test(a,alternative = "two.sided",conf.level = 0.9)
```

```
##
##  One Sample t-test
##
## data:  a
## t = 12.365, df = 35, p-value = 2.491e-14
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  2.863482 3.769852
## sample estimates:
## mean of x
##  3.316667
```

```
t.test(a,alternative = "two.sided",conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  a
## t = 12.365, df = 35, p-value = 2.491e-14
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  2.772142 3.861192
## sample estimates:
## mean of x
##  3.316667
```

```
t.test(a,alternative = "two.sided",conf.level = 0.99)
```

```
##
##  One Sample t-test
##
## data:  a
## t = 12.365, df = 35, p-value = 2.491e-14
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  2.586075 4.047258
## sample estimates:
## mean of x
##  3.316667
```

```
# 样本容量
#sample number function
samplemin.int=function(sigma1,sigma2,error,alpha) {
  z=qnorm(1-alpha/2,mean=0,sd=1,lower.tail = T)
  n=z^2*(sigma1^2+sigma2^2)/error^2
  cat("The number of Sample is more than", n)
}
```

```r
#function calculated
samplemin.int(12,15,5,0.05)
```

```
## The number of Sample is more than 56.69993
```

```r
b<-read.xlsx("F:/R/applicationstatics/data/exercise2.xlsx",sheet=2)

#function meantest
meantest.int=function(x,meanpop,sigmapop,alpha,pop=TRUE) {
  mean=mean(x)
  sd=sd(x)
  n=length(x)
  t0=qt(1-alpha/2,df=n-1,lower.tail = T)
  z0=qnorm(1-alpha/2,mean=0,sd=1,lower.tail = T)
  if (pop) {
    p=(mean-meanpop)/(sigmapop/sqrt(n))
    status=p-z0
  } else {
    sigmapop=sd
    p=(mean-meanpop)/(sigmapop/sqrt(n))
    status=p-t0
  }
  cat("Hypothesis Test:",status>0)
}

#function calculated
meantest.int(b[,1],0.618,1,0.05,pop = F)
```

```
## Hypothesis Test: FALSE
```

**3** 方差分析与回归分析

```r
#load library
library(MASS)
library(openxlsx)
library(psych)
library(corrplot)
```

1.

```r
#question1
a<-read.xlsx("F:/R/applicationstatics/Data/exercise3.xlsx",sheet=4)

#variance analysis
a.aov<-aov(battery~company,data=a)
summary(a.aov)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## company      2  615.6  307.80   17.07 0.00031 ***
## Residuals   12  216.4   18.03
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

根据方差分析结果，三个企业生产的电池的平均寿命之间有显著差异。根据 LSD 方法进行检验。LSD 检验统计

量公式如下：

$$LSD = t_{\alpha/2}\sqrt{MSE(\frac{1}{n_i}+\frac{1}{n_j})}$$

带入计算可得，LSD=5.760。然后可以计算可得：| A- B|=14.4>5.760，| A- C|=1.8<5.760，| B- C|=12.6>5.760。所以 A 企业和 B 企业，B 企业和 C 企业之间是有差异的。

2. 本问题为双因素的问题，所以采用双因子方差分析结果（分别选用的无交互作用和有交互作用的）如下：

```
#Input data
b<-read.xlsx("F:/R/applicationstatics/Data/exercise3.xlsx",sheet=6)

#no interaction
b.aov<-aov(value~location+competition,data=b)
summary(b.aov)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## location     2   1736   868.1  23.448 7.38e-07 ***
## competition  3   1078   359.4   9.709 0.000123 ***
## Residuals   30   1111    37.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#interaction
bi.aov<-aov(value~location*competition,data=b)
summary(bi.aov)
```

```
##                      Df Sum Sq Mean Sq F value   Pr(>F)
## location              2 1736.2   868.1  34.305 9.18e-08 ***
## competition           3 1078.3   359.4  14.204 1.57e-05 ***
## location:competition  6  503.3    83.9   3.315   0.0161 *
## Residuals            24  607.3    25.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

（1）从双因素方差分析的结果来看，F 统计值通过了 0.001 大于设定的显著性水平的显著性检验，可以认为竞争者的数量对销售额有显著影响。

（2）从双因素方差分析的结果来看，F 统计值通过了 0.001 大于设定的显著性水平的显著性检验，可以认为超市的位置对销售额有显著影响。

（3）从双因素方差分析的结果来看，F 统计值通过了 0.05 小于设定的显著性水平的显著性检验，可以认为竞争者的数量和超市的位置对销售额无交互影响。

3.

（1）$r_{y,x1}$=0.309，$r_{y,x2}$=0.01。并绘制了散点图，从相关系数来看，y 与 $x_1$ 有线性关系，y 与 $x_2$ 无线性关系。几何散点图来看，二者的线性关系也不是非常显著。

```
c<-read.xlsx("F:/R/applicationstatics/Data/exercise3.xlsx",sheet=3)
layout((matrix(c(1,2),nrow = 1,byrow = T)))
```

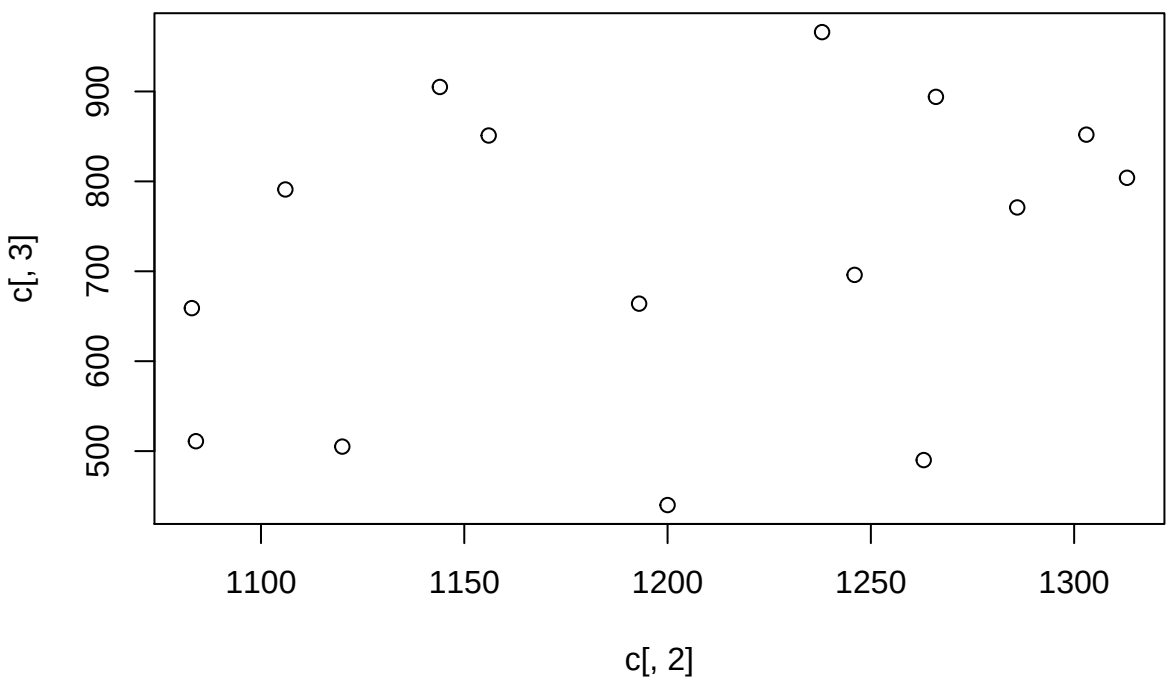
```
cor(c[,2],c[,3])
```

```
## [1] 0.3089521
```

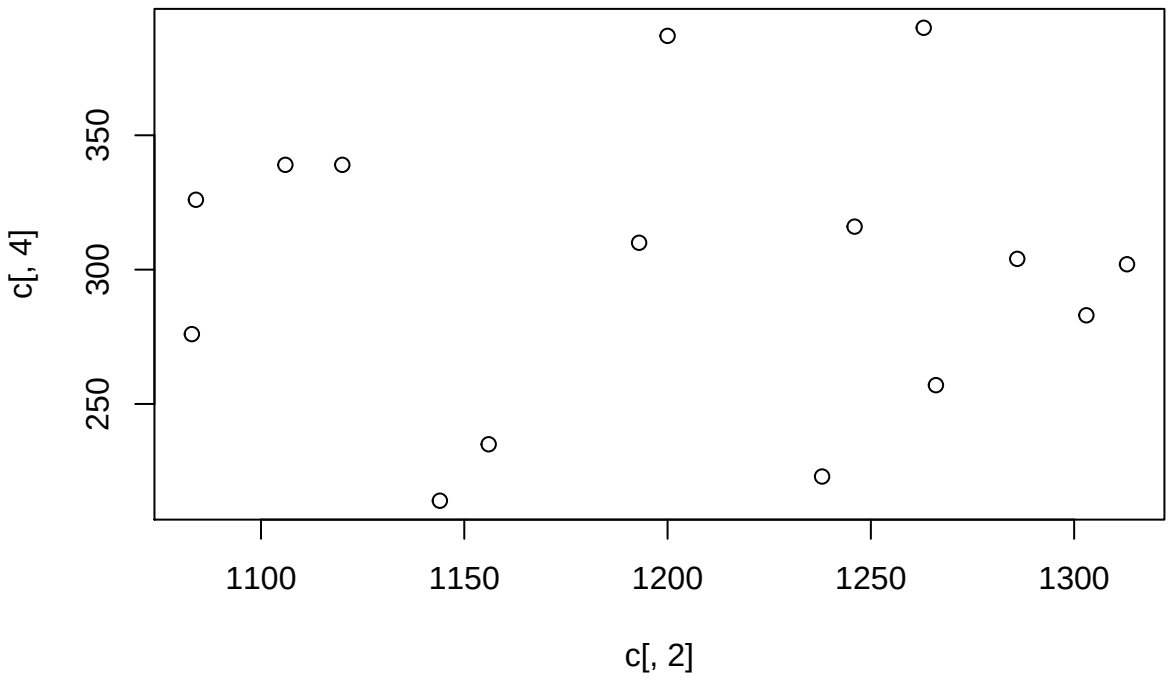
```
cor(c[,2],c[,4])
```

```
## [1] 0.001214062
```

```
plot(c[,2],c[,3])
```




```
plot(c[,2],c[,4])
```




（2）用购进价格来预测销售价格可能更有用，销售费用对销售价格影响较小。

```
c.lm<-lm(formula = c$销售价格 y~c$购进价格 x1+c$销售费用 x2)
summary(c.lm)
```

```
##
## Call:
## lm(formula = c$销售价格y ~ c$购进价格x1 + c$销售费用x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -189.02  -25.69   17.89   44.16   64.90
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  375.6018   339.4106   1.107   0.2901
## c$购进价格x1   0.5378     0.2104   2.556   0.0252 *
## c$销售费用x2   1.4572     0.6677   2.182   0.0497 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.75 on 12 degrees of freedom
## Multiple R-squared:  0.3525, Adjusted R-squared:  0.2445
## F-statistic: 3.266 on 2 and 12 DF,  p-value: 0.07372
```

（3）从 F 检验统计值来看，P 值通过了 0.1 的显著性水平检验，与题目所要求的 0.05 不符合。所以模型的线性关系不显著。

（4）判定系数 $R^2$ 为 0.352，说明销售价格变动的 35% 是由购进价格和销售费用决定的。线性关系较弱。

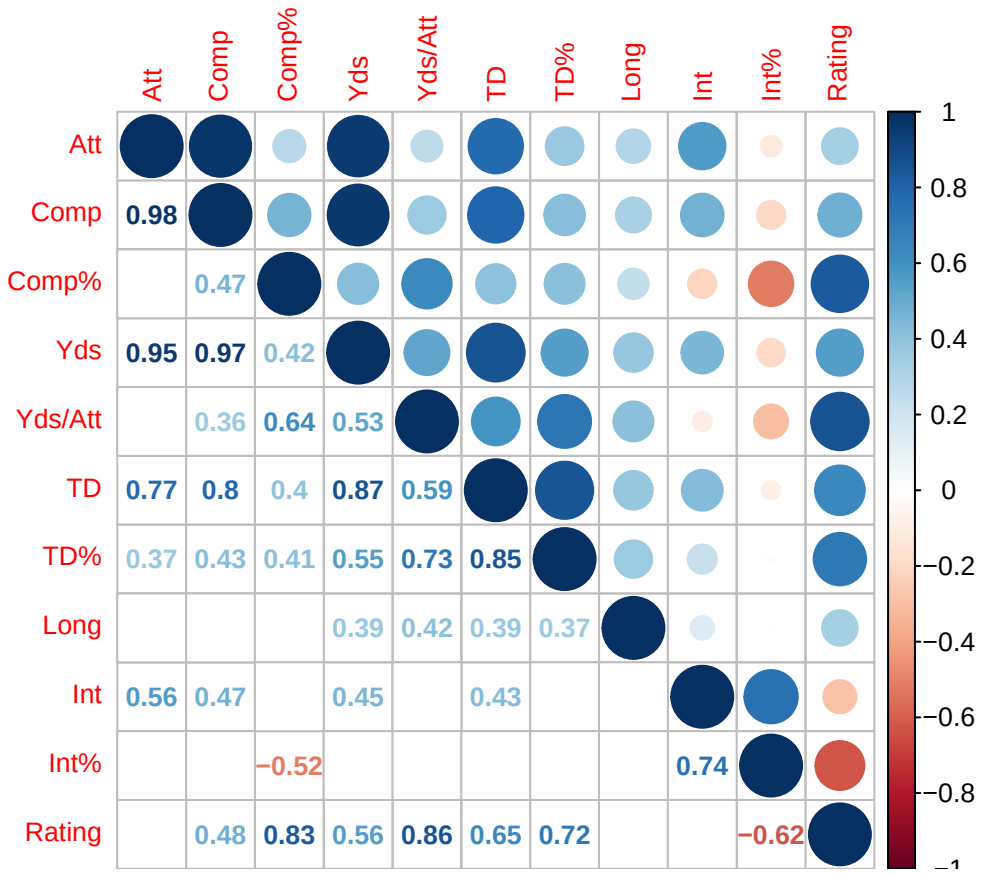
```
cor(c[,3],c[,4])
```

```
## [1] -0.8528576
```

（5）$r_{x1,x2}$=-0.853，说明购进价格与销售费用呈现负相关的关系。（6）模型存在多重共线性，建议使用逐步回归方法去除变量进行回归分析。

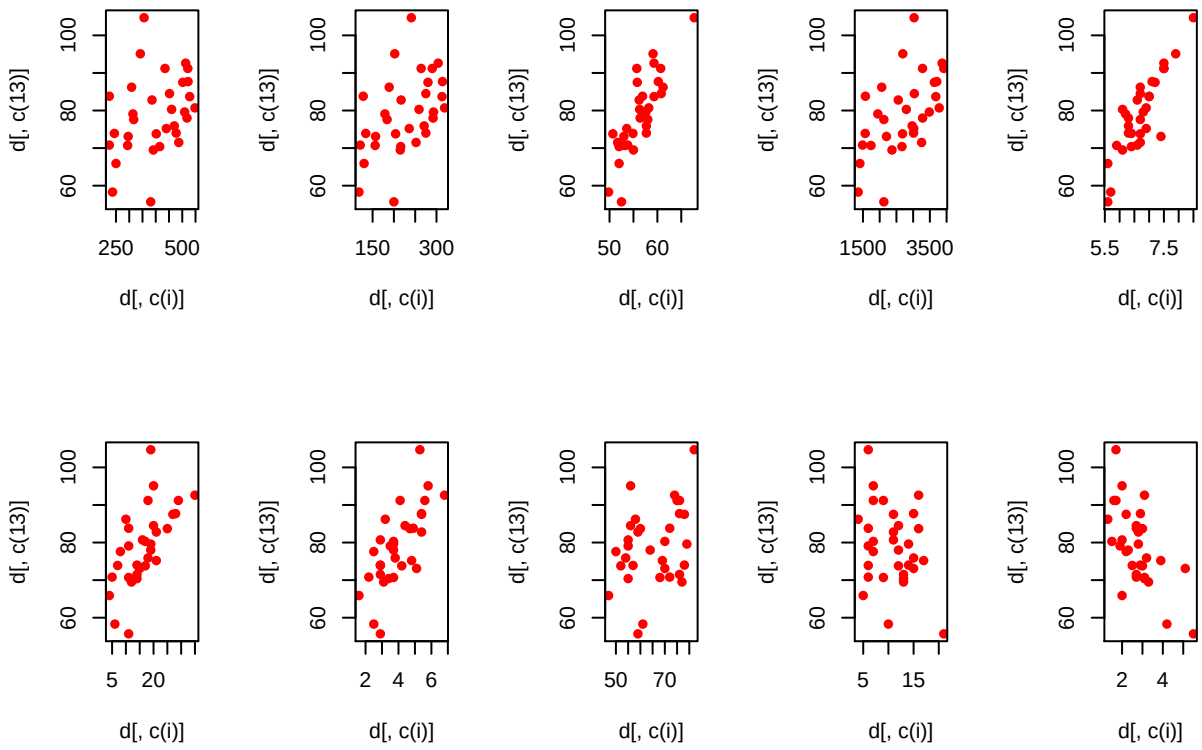
4. 首先对变量进行相关分析。

```
d<-read.xlsx("F:/R/applicationstatics/Data/exercise3.xlsx",sheet=5)
dcor<-corr.test(d[,c(3:13)])
dcorp<-dcor$p
dcorp[upper.tri(dcorp)]=0
corrplot.mixed(dcor$r,lower = "number",upper = "circle",diag = "u",
               tl.pos = "lt",tl.cex=0.8,number.cex=0.8,p.mat=dcorp,sig.level=0.05,insig=c("blank"))
```




可以发现 Rating 跟 Comp、Comp%，Yds，Yds/Att，TD，TD% 和 Int% 有显著的相关关系，且相关系数均在 0.48 以上。接下来绘制 Rating 跟其余 10 个指标的散点图。

```
layout((matrix(c(1,2,3,4,5,6,7,8,9,10),nrow=2,byrow=T)))
for (i in 3:12) {
  plot(d[,c(i)],d[,c(13)],col="red",pch=16)
}
```




可以看到与其他 10 个指标的散点图，线性关系也较为显著。根据相关系数矩阵结果和散点图，选定 7 个自变量进行逐步回归。结果如下：

```
d.lm<-lm(formula=d$Rating~d$Comp+d$`Comp%`+d$Yds+d$`Yds/Att`+d$TD+d$`TD%`+d$`Int%`)
summary(d.lm)
```

```
##
## Call:
## lm(formula = d$Rating ~ d$Comp + d$`Comp%` + d$Yds + d$`Yds/Att` +
##     d$TD + d$`TD%` + d$`Int%`)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.44418 -0.10043 -0.01134  0.07179  0.45795
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.8195100  0.9227453   0.888    0.383
## d$Comp      -0.0096770  0.0127657  -0.758    0.456
## d$`Comp%`    0.8824402  0.0580504  15.201 8.12e-14 ***
## d$Yds        0.0007261  0.0011967   0.607    0.550
## d$`Yds/Att`  3.9934242  0.4803831   8.313 1.59e-08 ***
## d$TD         0.0053251  0.0412568   0.129    0.898
## d$`TD%`      3.2613349  0.1736048  18.786 7.38e-16 ***
## d$`Int%`    -4.1156277  0.0552894 -74.438  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.2107 on 24 degrees of freedom
## Multiple R-squared:  0.9997, Adjusted R-squared:  0.9996
## F-statistic: 1.073e+04 on 7 and 24 DF,  p-value: < 2.2e-16
```

```
d.lms<-step(d.lm)
```

```
## Start:  AIC=-92.86
## d$Rating ~ d$Comp + d$`Comp%` + d$Yds + d$`Yds/Att` + d$TD +
##     d$`TD%` + d$`Int%`
##
##               Df Sum of Sq     RSS     AIC
## - d$TD         1     0.001   1.067 -94.841
## - d$Yds        1     0.016   1.082 -94.376
## - d$Comp       1     0.026   1.091 -94.106
## <none>                       1.066 -92.863
## - d$`Yds/Att`  1     3.069   4.135 -51.481
## - d$`Comp%`    1    10.262  11.328 -19.230
## - d$`TD%`      1    15.673  16.739  -6.736
## - d$`Int%`     1   246.077 247.142  79.415
##
## Step:  AIC=-94.84
## d$Rating ~ d$Comp + d$`Comp%` + d$Yds + d$`Yds/Att` + d$`TD%` +
##     d$`Int%`
##
##               Df Sum of Sq     RSS     AIC
## - d$Yds        1     0.035   1.102 -95.800
## - d$Comp       1     0.040   1.106 -95.669
## <none>                       1.067 -94.841
## - d$`Yds/Att`  1     5.126   6.193 -40.555
## - d$`Comp%`    1    13.339  14.405 -13.541
## - d$`TD%`      1   185.958 187.024  68.496
## - d$`Int%`     1   248.858 249.925  77.774
##
## Step:  AIC=-95.8
## d$Rating ~ d$Comp + d$`Comp%` + d$`Yds/Att` + d$`TD%` + d$`Int%`
##
##               Df Sum of Sq    RSS     AIC
## - d$Comp       1      0.03   1.14 -96.806
## <none>                       1.10 -95.800
## - d$`Yds/Att`  1     77.60  78.70  38.797
## - d$`Comp%`    1    132.45 133.55  55.719
## - d$`TD%`      1    191.93 193.03  67.508
## - d$`Int%`     1    318.97 320.08  83.690
##
## Step:  AIC=-96.81
## d$Rating ~ d$`Comp%` + d$`Yds/Att` + d$`TD%` + d$`Int%`
##
##               Df Sum of Sq    RSS     AIC
## <none>                       1.14 -96.806
## - d$`Yds/Att`  1     79.79  80.92  37.688
## - d$`Comp%`    1    145.56 146.69  56.724
## - d$`TD%`      1    210.59 211.73  68.467
## - d$`Int%`     1    319.64 320.77  81.760
```

```r
summary(d.lms)
```

```
##
## Call:
## lm(formula = d$Rating ~ d$`Comp%` + d$`Yds/Att` + d$`TD%` + d$`Int%`)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46142 -0.10795 -0.01766  0.10111  0.42857
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.22654    0.79856   1.536    0.136
## d$`Comp%`    0.83826    0.01426  58.802   <2e-16 ***
## d$`Yds/Att`  4.28174    0.09835  43.535   <2e-16 ***
## d$`TD%`      3.27642    0.04632  70.729   <2e-16 ***
## d$`Int%`    -4.13490    0.04745 -87.137   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2052 on 27 degrees of freedom
## Multiple R-squared:  0.9997, Adjusted R-squared:  0.9996
## F-statistic: 1.98e+04 on 4 and 27 DF,  p-value: < 2.2e-16
```

可以看到逐步回归结果只保留了 4 个自变量（Comp%，Yds/Att，TD%，Int%），模型的 $R^2$ 达到了 1.000。说明美式足球员的 Ranting 变化的 100% 能够被如上的四个变量进行解释。标准残差为 0.205。说明预测精度非常高，残差较小，而 F 统计值通过了 0.01 的显著性检验。说明该预测方程可信度较高。