暨南大學
JINAN UNIVERSITY

多元统计分析及R语言建模

第2章 多元数据的数学表达及R使用

王斌会 教授

多元统计分析及R语言建模 → 2 多元数据的数学表达及R使用

内容与要求

内容：

多元数据的基本格式，如何收集和整理多元统计分析资料、数据的数学表达、数据矩阵及R表示、数据的R语言表示、R调用多元的数据和多元的数据的简单R语言分析。

要求：

要求学生熟练如何收集和整理多元统计分析资料、数据的数学表达、掌握多元数据的数字特征的解析表达式、数字特征的基本性质。熟悉有关统计软件。利用统计软件来练习矩阵的有关计算。练习在已给数据下，求样本均值、样本离差阵、样本协差阵等。

多元分析资料的一般格式

	变量 X_1	变量 X_2 …	变量 X_p
记录 1	x_{11}	x_{12} …	x_{1p}
记录 2	x_{21}	x_{22} …	x_{2p}
	……		
记录 n	x_{n1}	x_{n2} …	x_{np}

矩阵化表示

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$= (x_1, x_2, \ldots, x_p) = (x_{ij})_{n\times p}$$

举例

【例2.1】为了了解股民的投资状况，研究股民的股票投资特征，我们在2002年组织统计系本科生进行小范围的“股民投资状况抽样调查”。本次调查的抽样框主要涉及广东省的6个城市（广州、深圳、珠海、中山、佛山和东莞，其中，广州、深圳各100份，其他城市各80份），共发放问卷520份，回收有效问卷514份。问卷中设计了18个问题。为了简化分析，本例只考虑：年龄、性别、风险意识、是否专兼职、职业状况、教育程度和投资结果共7个变量进行分析。

#本例性别、风险、专兼职、职业、教育和结果为定性变量，年龄是定量变量，有时为了分析问题方便，也可将其定量化，例如

- 年龄（age）：19岁以下（1）；20至29岁（2）；30 至39岁（3）；40至49岁（4）；50至59岁（5）；60岁 及以上（6）；缺失（*）。
- 性别（sex）：男（1），女（2）。
- 风险（risk）：有（1）；无（2）。
- 专兼职（post）：专职（1）；业余（2）。
- 职业（career）：干部（1）；管理（2）；3科教（3）；金融（4）；工人（5）；农民（6）；个体（7）；无业（8）。
- 教育（edu）：文盲（1）；小学（2）；中学（3）；高中（4）；中专（5）；

大专（6）；大学（7）；研究生（8）。

- 投资结果（result）：赚钱（1）；不赔不赚（2）；赔钱（3）。

数据的表达

一元数据

期望

$$\mu = E(X) = \sum_i x_i p_i$$

方差

$$\sigma^2 = Var(X) = \sum_i (x_i - \mu)^2 p_i$$

样本均值和方差

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s^2 = \frac{l_{xx}}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

多元数据

期望

$$E(X) = (E(x_1), E(x_2), \cdots, E(x_p))$$

协方差

$$\sum = Var(X) = \begin{bmatrix} cov(x_1,x_1) & cov(x_1,x_2) & \cdots & cov(x_1,x_p) \\ cov(x_2,x_1) & cov(x_2,x_2) & \cdots & cov(x_2,x_p) \\ \vdots & \vdots & \ddots & \vdots \\ cov(x_p,x_1) & cov(x_p,x_2) & \cdots & cov(x_p,x_p) \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

【例 2.2】测得 12 名学生的生长发育指标：身高(x1)、体重(x2)的数据，试用 R 语言表述该数据。

x1	171，	175，	159，	155，	152，	158，	154，	164，	168，	166，	159，	164
x2	57，	64，	41，	38，	35，	44，	41，	51，	57，	49，	47，	46

在R中可以用函数c()来创建向量：

```
x1=c(171,175,159,155,152,158,154,164,168,166,159,164)
x2=c(57,64,41,38,35,44,41,51,57,49,47,46)
```

在R中结果输出如下：

```
> x1
 [1] 171 175 159 155 152 158 154 164 168 166 159 164
> x2
 [1] 57 64 41 38 35 44 41 51 57 49 47 46
```

#将向量按列和并

rbind(x1,x2)

```
> cbind(x1,x2
      x1 x2
 [1,] 171 57
 [2,] 175 64
 [3,] 159 41
 [4,] 155 38
 [5,] 152 35
 [6,] 158 44
 [7,] 154 41
 [8,] 164 51
 [9,] 168 57
[10,] 166 49
[11,] 159 47
[12,] 164 46
```

#利用x1数据创建矩阵

matrix(x1,nrow=3,ncol=4)

```
> matrix(x1,nrow=3,ncol=4)
     [,1] [,2] [,3] [,4]
[1,]  171  155  154  166
[2,]  175  152  164  159
[3,]  159  158  168  164
```

#创建按照行排列的矩阵

matrix(x1,nrow=3,ncol=4，byrow=T)

```
> matrix(x1,nrow=3,ncol=4,byrow=T)
     [,1] [,2] [,3] [,4]
[1,]  171  175  159  155
[2,]  152  158  154  164
[3,]  168  166  159  164
```

2 多元数据的数学表达及R使用 ⟹ 2.3 数据矩阵

#创建两个相同的矩阵
A=B=matrix(1:12,nrow=3,ncol=4)

```
> A;B
     [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
     [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
```

#矩阵转置
t(A)

```
> t(A)
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
[3,]    7    8    9
[4,]   10   11   12
```

#矩阵加法

A+B

```
> A+B
     [,1] [,2] [,3] [,4]
[1,]    2    8   14   20
[2,]    4   10   16   22
[3,]    6   12   18   24
```

#矩阵加法

A+B

```
> A-B
     [,1] [,2] [,3] [,4]
[1,]    0    0    0    0
[2,]    0    0    0    0
[3,]    0    0    0    0
```

```
#矩阵相乘
A=matrix(1:12,nrow=3,ncol=4)
B=matrix(1:12,nrow=4,ncol=3)
A%in%B
```

```
     [,1] [,2] [,3]
[1,]   70  158  246
[2,]   80  184  288
[3,]   90  210  330
```

```
#获取对角线元素
A=matrix(1:16,nrow=4,ncol=4)
diag(A)
```

```
> diag(A)
[1]  1  6 11 16
```

```
#利用对角线元素创建对角矩阵
diag(diag(A))
```

```
> diag(diag(A))
     [,1] [,2] [,3] [,4]
[1,]    1    0    0    0
[2,]    0    6    0    0
[3,]    0    0   11    0
[4,]    0    0    0   16
```

```
#创建3阶单位矩阵
diag(3)
```

```
> diag(3)
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
```

#求逆矩阵

A=matrix(rnorm(16),4,4)

solve(A)

```
> solve(A)
           [,1]        [,2]       [,3]        [,4]
[1,]  0.2485820 -0.35092701  0.19955797 -0.03945507
[2,] -1.0308041 -0.18913835 -0.03197376  0.90517759
[3,] -0.9322897  0.22579897  0.25031260  0.46842094
[4,]  2.3471280  0.08939981 -0.36256217 -0.26584474
```

#求矩阵特征根与特征向量

A=diag(4)+1

A.e=eigen(A,symmetric=T)

```
$values
[1] 5 1 1 1

$vectors
     [,1]          [,2]       [,3]       [,4]
[1,] -0.5  0.000000e+00  0.0000000  0.8660254
[2,] -0.5 -6.408849e-17  0.8164966 -0.2886751
[3,] -0.5 -7.071068e-01 -0.4082483 -0.2886751
[4,] -0.5  7.071068e-01 -0.4082483 -0.2886751
```

#矩阵的Choleskey分解

A.c=chol(A)

```
         [,1]      [,2]      [,3]      [,4]
[1,] 1.414214 0.7071068 0.7071068 0.7071068
[2,] 0.000000 1.2247449 0.4082483 0.4082483
[3,] 0.000000 0.0000000 1.1547005 0.2886751
[4,] 0.000000 0.0000000 0.0000000 1.1180340
```

```
#矩阵奇异值分解
A=matrix(1:18,3,6)
A.s=svd(A)
```

```
$d
[1] 4.589453e+01 1.640705e+00 3.627301e-16

$u
           [,1]       [,2]       [,3]
[1,] -0.5290354  0.74394551  0.4082483
[2,] -0.5760715  0.03840487 -0.8164966
[3,] -0.6231077 -0.66713577  0.4082483

$v
            [,1]        [,2]        [,3]
[1,] -0.07736219 -0.71960032 -0.18918124
[2,] -0.19033085 -0.50893247  0.42405898
[3,] -0.30329950 -0.29826463 -0.45330031
[4,] -0.41626816 -0.08759679 -0.01637004
[5,] -0.52923682  0.12307105  0.64231130
[6,] -0.64220548  0.33373889 -0.40751869
```

```
#矩阵的维数
A=matrix(1:12,3,4)
dim(A)
```

```
[1] 3 4
```

```
#矩阵的行数
nrow(A)
```

```
[1] 3
```

```
#矩阵的行数
ncol(A)
```

```
[1] 4
```

```
#矩阵按行求和
rowSums(A)
```

```
[1] 22 26 30
```

```
#矩阵按行求均值
rowMeans(A)
```

```
[1] 5.5 6.5 7.5
```

```
#矩阵按列求和
colSums(A)
```

```
[1]  6 15 24 33
```

```
#矩阵按列求均值
colMeans(A)
```

```
[1]  2  5  8 11
```

apply()函数

apply(X, MARGIN, FUN, ...)

```
#矩阵按行求和
apply(A,1,sum)
```

```
[1] 22 26 30
```

```
#矩阵按行求均值
apply(A,1,mean)
```

```
[1] 5.5 6.5 7.5
```

#矩阵按列求和

apply(A,2,sum)

```
[1]  6 15 24 33
```

#矩阵按列求均值

apply(A,2,mean)

```
[1]  2  5  8 11
```

#矩阵按列求方差

A=matrix(rnorm(100),20,5)

apply(A,2,var)

```
[1] 1.2748524 1.8964186 1.2920973 0.6991467 0.5818300
```

#矩阵按列求函数结果

B=matrix(1:12,3,4)

apply(B,2,function(x,a) x*a, a=2)

```
     [,1] [,2] [,3] [,4]
[1,]    2    8   14   20
[2,]    4   10   16   22
[3,]    6   12   18   24
```

数据框

数据框（data frame）是一种矩阵形式的数据，但数据框中各列可以是不同类型的数据。

数据框录入限制条件

1. 分量必须是向量（数值，字符，逻辑）、因子、数值矩阵、列表或者其他数据框 。
2. 矩阵、列表和数据框为新的数据框提供了尽可能多的变量，因为它们各自拥有列、元素或者变量。
3. 数值向量、逻辑值、因子保持原有格式，而字符向量会被强制转换成因子并且它的水平就是向量中出现的独立值。
4. 在数据框中以变量形式出现的向量长度必须一致，矩阵结构必须有一样的行数。

```
#由x1和x2构建数据框
X=data.frame(x1,x2)
```

```
    x1 x2
1  171 57
2  175 64
3  159 41
4  155 38
5  152 35
6  158 44
7  154 41
8  164 51
9  168 57
10 166 49
11 159 47
12 164 46
```

```
#赋予数据框新的列标签
X=data.frame('身高'=x1,'体重'=x2)
```

```
    身高 体重
1   171   57
2   175   64
3   159   41
4   155   38
5   152   35
6   158   44
7   154   41
8   164   51
9   168   57
10  166   49
11  159   47
12  164   46
```

从剪切板读取


选择需要进行计算的数据块（比如上例中名为UG的数据），拷贝之。
01
在R中使用dat <-read.table("clipboard",header=T)
02

从文本文件读取

```
#读取名为textdata的txt格式文档
X=read.table("textdata.txt")
```

```
    V1 V2
1   x1 x2
2  171 57
3  175 64
4  159 41
5  155 38
6  152 35
7  158 44
8  154 41
9  164 51
10 168 57
11 166 49
12 159 47
13 164 46
```

第一行作为标题时

读取csv格式和excel格式

1.下载读取excel文件的包“readxl”

2. 调用包：library(readxl)

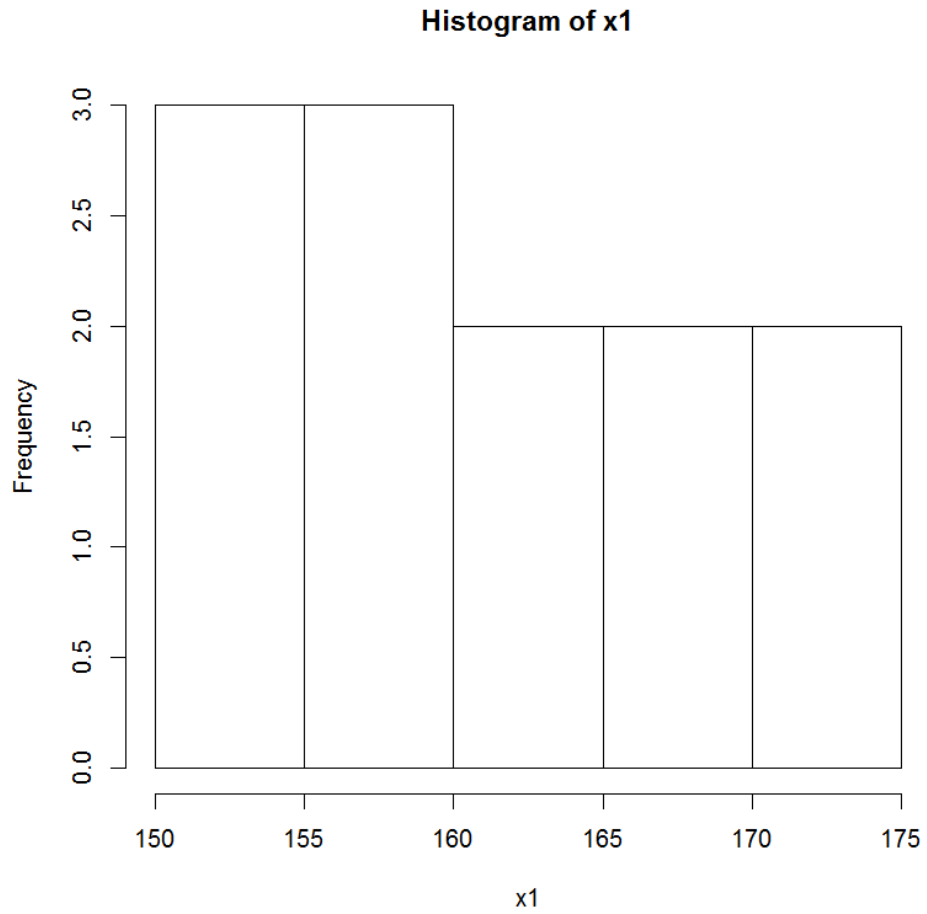
3. 读取文件：X=read_excel(“data.xls”)

X=read.csv("textdata.csv")

定量变量分析

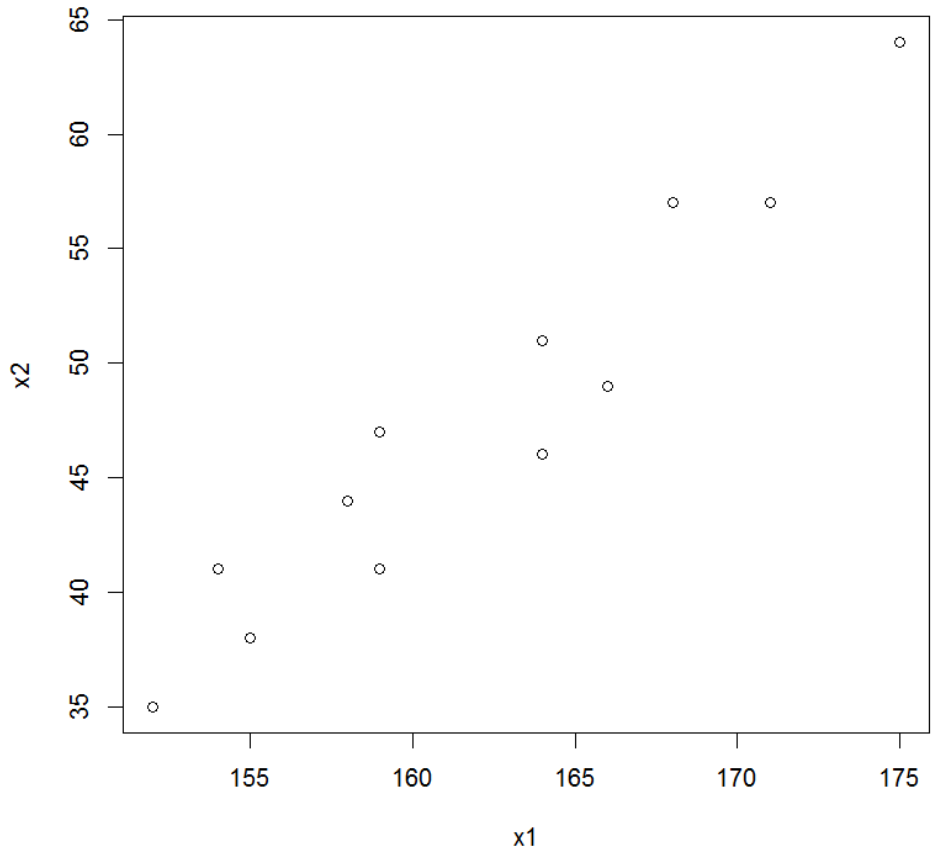
#身高的直方图

hist(x1)



#身高与体重散点图

plot(x1,x2)



定性变量分析

```
#将剪切板数据读入数据框d2.1中
d2.1=read.table("clipboard",header=T)
#显示数据前6行
head(d2.1)
```

	年龄	性别	风险	专兼职	职业	教育	结果
1	20-29	男	有	兼职	金融	高中	赚钱
2	50-59	女	有	兼职	科教	中学	持平
3	40-49	女	无	专职	科教	中学	赔钱
4	30-39	男	有	兼职	工人	中专	赚钱
5	50-59	女	有	专职	农民	大专	赚钱
6	40-49	女	有	兼职	管理	小学	赚钱

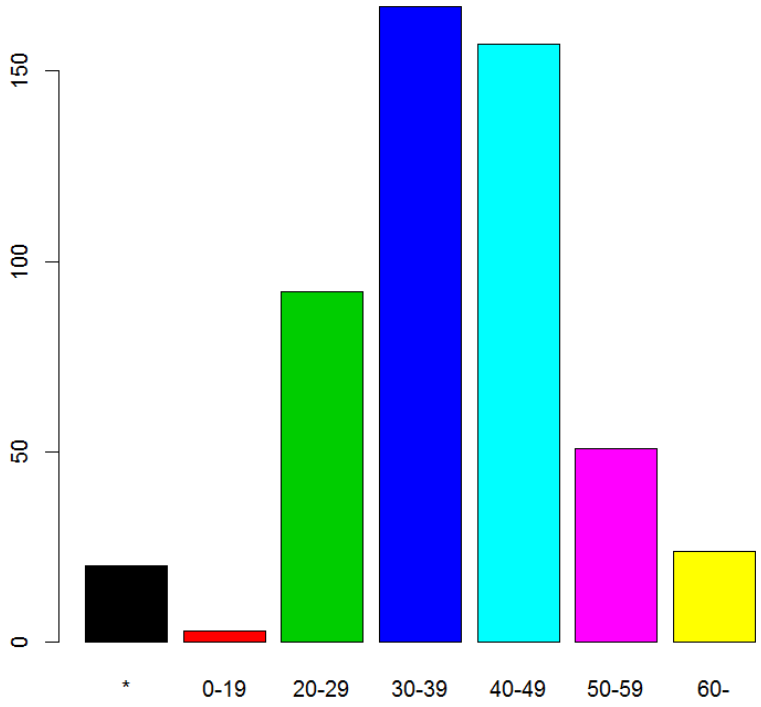
```
#绑定数据
attach(d2.1)
#一维列联表
table(年龄)
```

```
年龄
   *  0-19 20-29 30-39 40-49 50-59   60-
  20     3    92   167   157    51    24
```

定性变量分析（单因素）

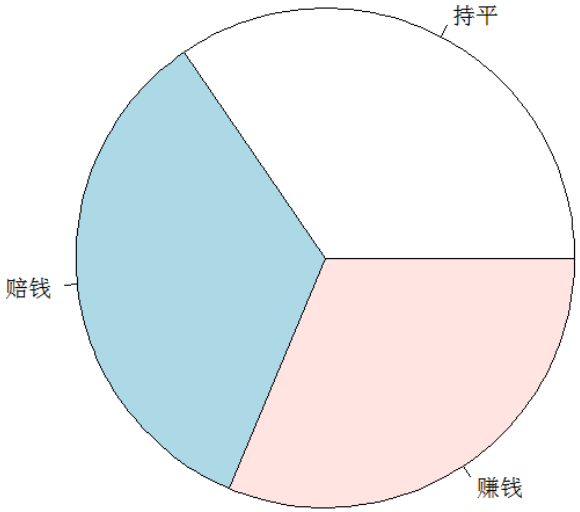
#条形图

```
barplot(table(年龄),col=1:7)
```



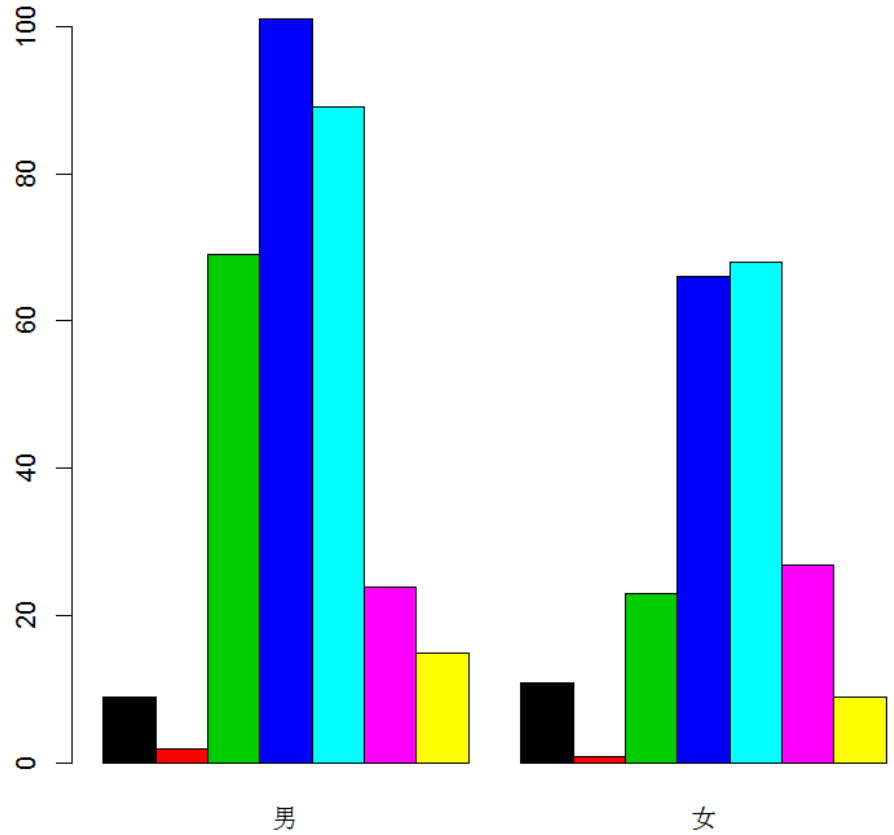
#饼图

```
pie(table(结果))
```

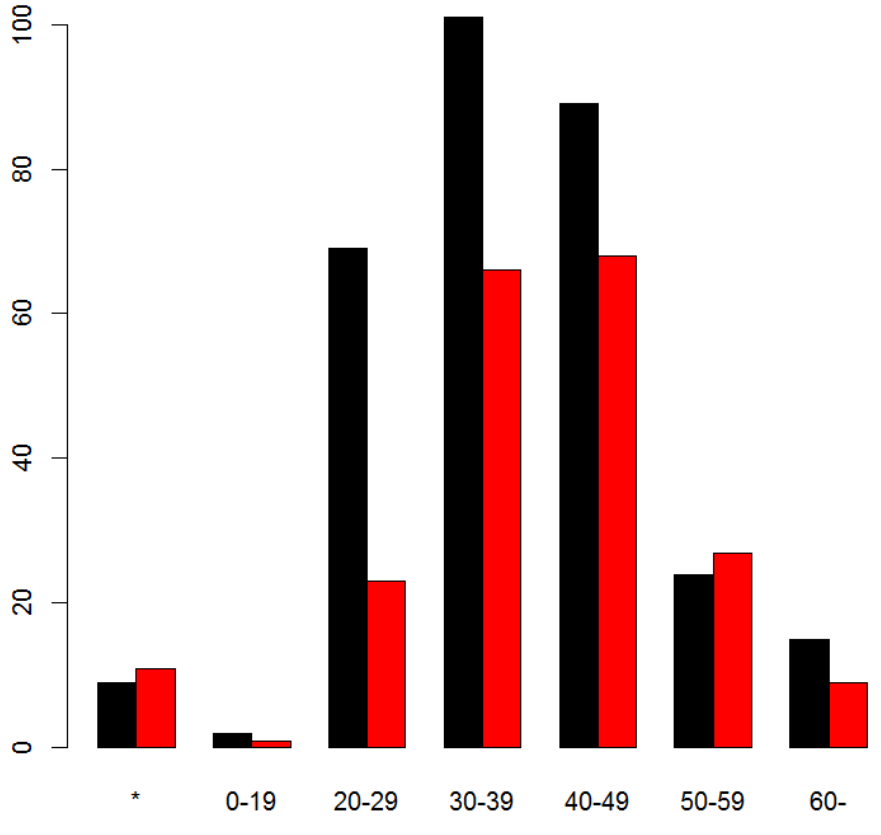


定性变量分析（双因素）

```
#以性别分组的年龄条图
barplot(table(年龄,性别),
        beside =T, col = 1:7)
```




```
#以年龄分组的性别条图
barplot(table(性别,年龄),
        beside=T,col =1:2)
```



定性变量分析（三因素）

#以年龄、性别排列的结果频数三维列联表

ftable(年龄,性别,结果)

年龄	性别	结果 持平	赔钱	赚钱
*	男	4	3	2
	女	3	7	1
0-19	男	0	0	2
	女	1	0	0
20-29	男	21	17	31
	女	10	7	6
30-39	男	31	30	40
	女	30	20	16
40-49	男	31	30	28
	女	25	30	13
50-59	男	5	11	8
	女	8	10	9
60-	男	7	5	3
	女	2	5	2

#以性别、年龄排列的结果频数三维列联表

ftable(性别,年龄,结果)

性别	年龄	结果 持平	赔钱	赚钱
男	*	4	3	2
	0-19	0	0	2
	20-29	21	17	31
	30-39	31	30	40
	40-49	31	30	28
	50-59	5	11	8
	60-	7	5	3
女	*	3	7	1
	0-19	1	0	0
	20-29	10	7	6
	30-39	30	20	16
	40-49	25	30	13
	50-59	8	10	9
	60-	2	5	2

定性变量分析（三因素）

#ft=ftable(性别,结果,年龄)

```
                年龄  * 0-19 20-29 30-39 40-49 50-59 60-
性别 结果
男   持平          4    0    21    31    31     5    7
     赔钱          3    0    17    30    30    11    5
     赚钱          2    2    31    40    28     8    3
女   持平          3    1    10    30    25     8    2
     赔钱          7    0     7    20    30    10    5
     赚钱          1    0     6    16    13     9    2
```

#求ft的行和

rowSums(ft)

```
[1]  99  96 114  79  79  47
```

#求ft的列和

colSums(ft)

```
[1]  20   3  92 167 157  51  24
```

#整理得

性别	结果	年龄							
		*	0-19	20-29	30-39	40-49	50-59	60-	合计
男	持平	4	0	21	31	31	5	7	99
	赔钱	3	0	17	30	30	11	5	96
	赚钱	2	2	31	40	28	8	3	114
女	持平	3	1	10	30	25	8	2	79
	赔钱	7	0	7	20	30	10	5	79
	赚钱	1	0	6	16	13	9	2	47
合计		20	3	92	167	157	51	24	514

注意

detach(d2.1)

当数据框不使用时，解除绑定！！