QSS_Chapter3

my

2021/12/9

目录

```r
setwd("D:/QSS/Chapter3_Measurement")
afghan <- read.csv("afghan.csv")
View(afghan)
summary(afghan)
```

```
##    province           district           village.id         age
##  Length:2754        Length:2754        Min.   :  1.0   Min.   :15.00
##  Class :character   Class :character   1st Qu.: 53.0   1st Qu.:22.00
##  Mode  :character   Mode  :character   Median :104.5   Median :30.00
##                                        Mean   :103.6   Mean   :32.39
##                                        3rd Qu.:153.0   3rd Qu.:40.00
##                                        Max.   :204.0   Max.   :80.00
```

```
## 
##    educ.years        employed          income          violent.exp.ISAF
##  Min.   : 0.000   Min.   :0.0000   Length:2754        Min.   :0.0000  
##  1st Qu.: 0.000   1st Qu.:0.0000   Class :character   1st Qu.:0.0000  
##  Median : 1.000   Median :1.0000   Mode  :character   Median :0.0000  
##  Mean   : 4.002   Mean   :0.5828                      Mean   :0.3749  
##  3rd Qu.: 8.000   3rd Qu.:1.0000                      3rd Qu.:1.0000  
##  Max.   :18.000   Max.   :1.0000                      Max.   :1.0000  
##                                                       NA's   :25      
##  violent.exp.taliban  list.group        list.response  
##  Min.   :0.0000      Length:2754        Min.   :0.000  
##  1st Qu.:0.0000      Class :character   1st Qu.:1.000  
##  Median :0.0000      Mode  :character   Median :2.000  
##  Mean   :0.3289                         Mean   :1.611  
##  3rd Qu.:1.0000                         3rd Qu.:2.000  
##  Max.   :1.0000                         Max.   :4.000  
##  NA's   :54
```

```
prop.table(table(ISAF = afghan$violent.exp.ISAF,
                 Taliban = afghan$violent.exp.taliban))
```

```
##     Taliban
## ISAF         0         1
##    0 0.4953445 0.1318436
##    1 0.1769088 0.1959032
```

```
## 打印出前十位回复者的收入数据
head(afghan$income, n = 10)
```

```
##  [1] "2,001-10,000"  "2,001-10,000"  "2,001-10,000"  "2,001-10,000" 
##  [5] "2,001-10,000"  NA              "10,001-20,000" "2,001-10,000" 
##  [9] "2,001-10,000"  NA
```

```
## 查看他们的收入数据是否缺失
head(is.na(afghan$income), n = 10)
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE
```

```
sum(is.na(afghan$income))  # 缺失值总数
```

```
## [1] 154
```

```
mean(is.na(afghan$income)) # 缺失值比例
```

```
## [1] 0.05591866
```

```
prop.table(table(ISAF = afghan$violent.exp.ISAF,
                 Taliban = afghan$violent.exp.taliban, exclude = NULL))
```

```
##       Taliban
## ISAF             0           1        <NA>
##   0    0.482933914 0.128540305 0.007988381
##   1    0.172476398 0.190994916 0.007988381
##   <NA> 0.002541757 0.002904866 0.003631082
```

```
afghan.sub <- na.omit(afghan)  # 对整个数据使用列表式删除
nrow(afghan.sub)
```

```
## [1] 2554
```

```
length(na.omit(afghan$income))
```

```
## [1] 2600
```

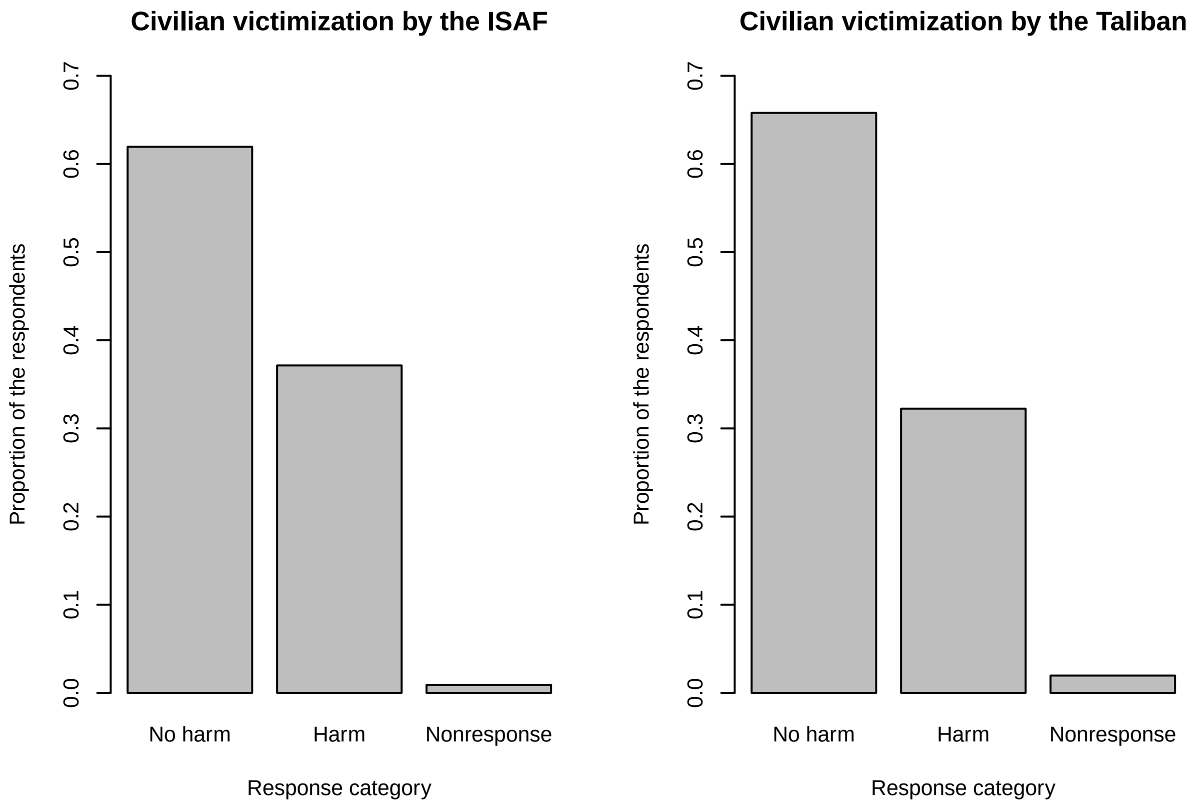
```
# 统计给出不同回复的比例
ISAF.ptable <- prop.table(table(ISAF = afghan$violent.exp.ISAF,
                                exclude = NULL))
ISAF.ptable
```

```
## ISAF
##           0           1        <NA>
## 0.619462600 0.371459695 0.009077705
```

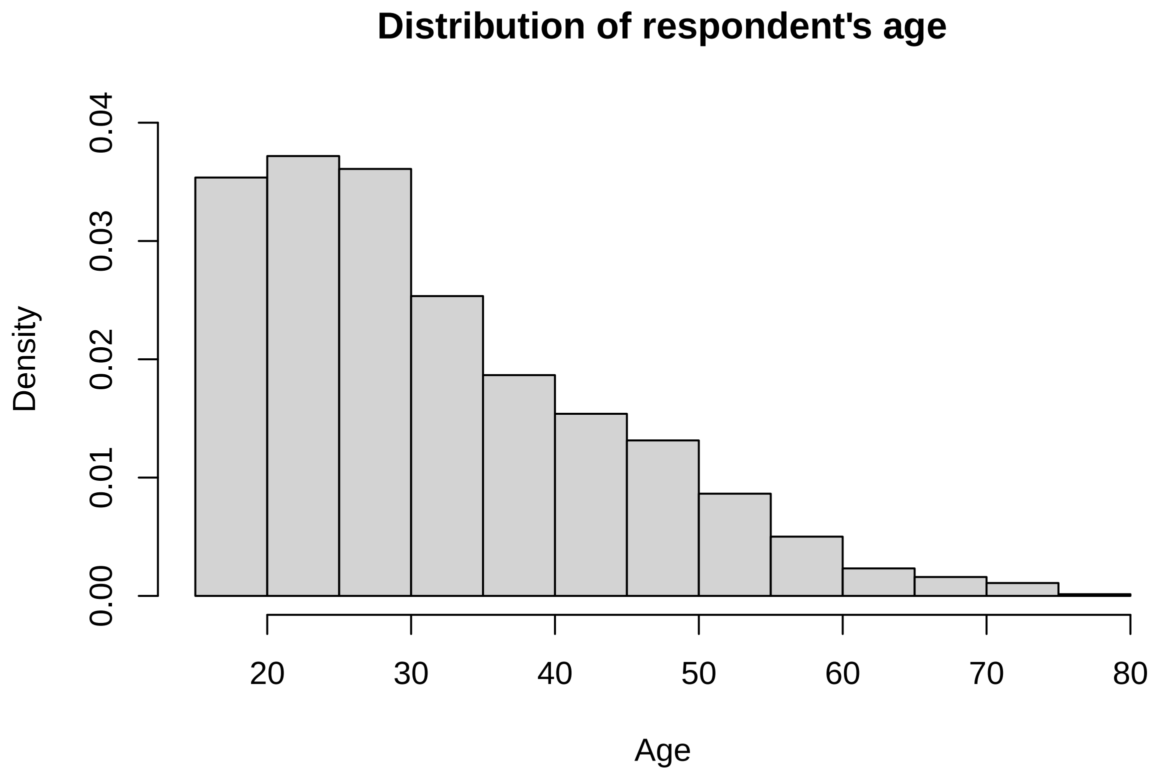
```
Taliban.ptable <- prop.table(table(Taliban = afghan$violent.exp.taliban,
                                exclude = NULL))
Taliban.ptable
```

```
## Taliban
##          0          1       <NA>
## 0.65795207 0.32244009 0.01960784
```

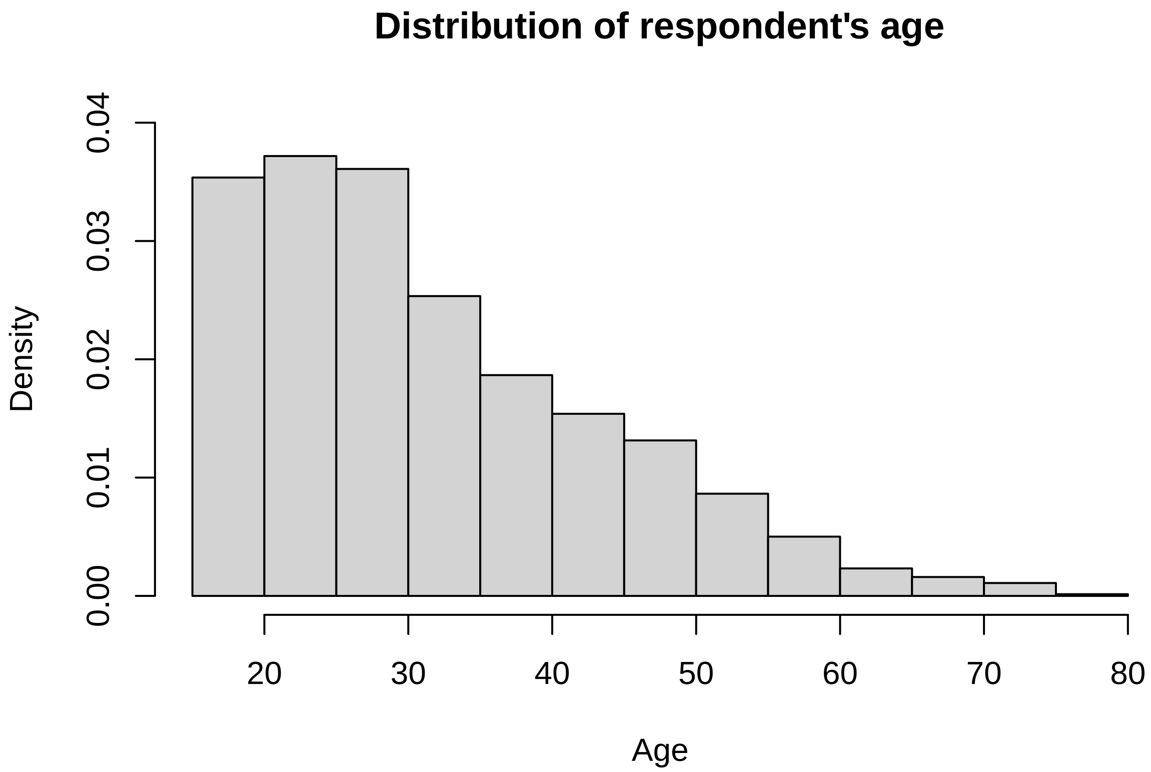
```
# 在一个图形文件中将多个相邻图打印出来
par(mfrow=c(1, 2), cex = 0.7)
# 画出民众受到 ISAF 和塔利班的伤害情况的两个条形图
barplot(ISAF.ptable,
        names.arg = c("No harm", "Harm", "Nonresponse"), # 指定每个小节标签
        main = "Civilian victimization by the ISAF",
        xlab = "Response category",
        ylab = "Proportion of the respondents",
        ylim = c(0, 0.7)
        )
barplot(Taliban.ptable,
        names.arg = c("No harm", "Harm", "Nonresponse"),
        main = "Civilian victimization by the Taliban",
        xlab = "Response category",
        ylab = "Proportion of the respondents",
        ylim = c(0, 0.7)
        )
```



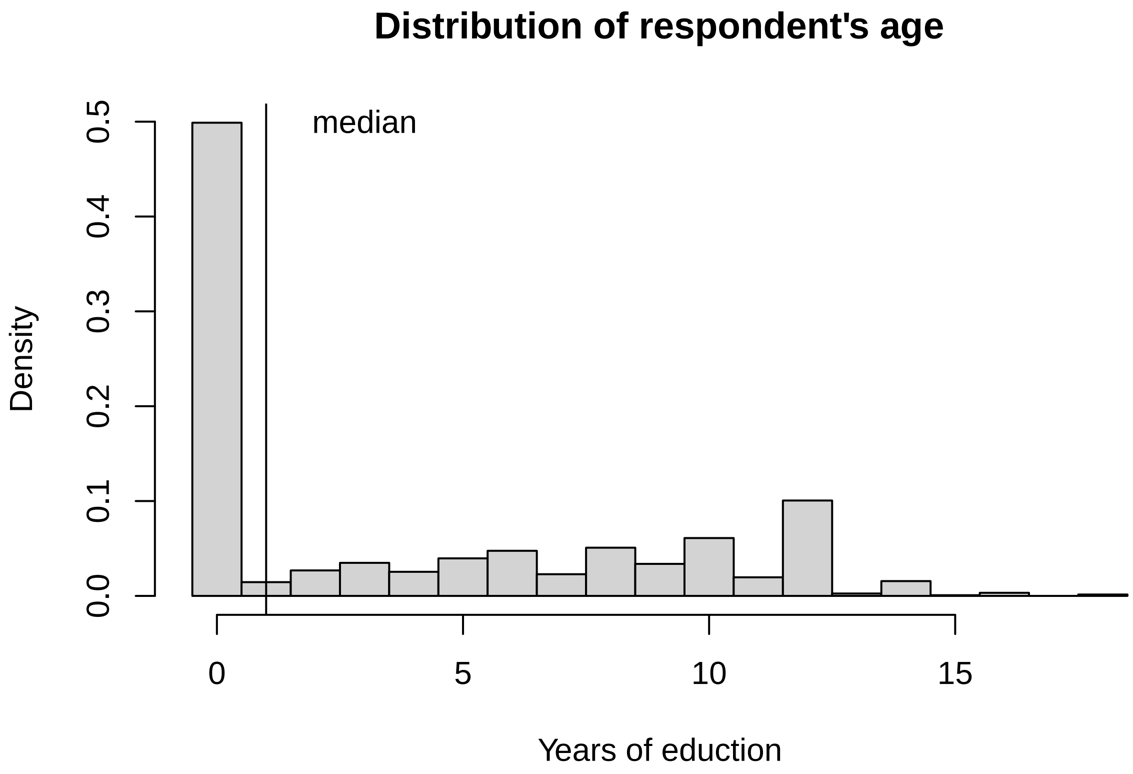
```
hist(afghan$age, freq = FALSE, ylim = c(0, 0.04), xlab = "Age",
     main = "Distribution of respondent's age")
```



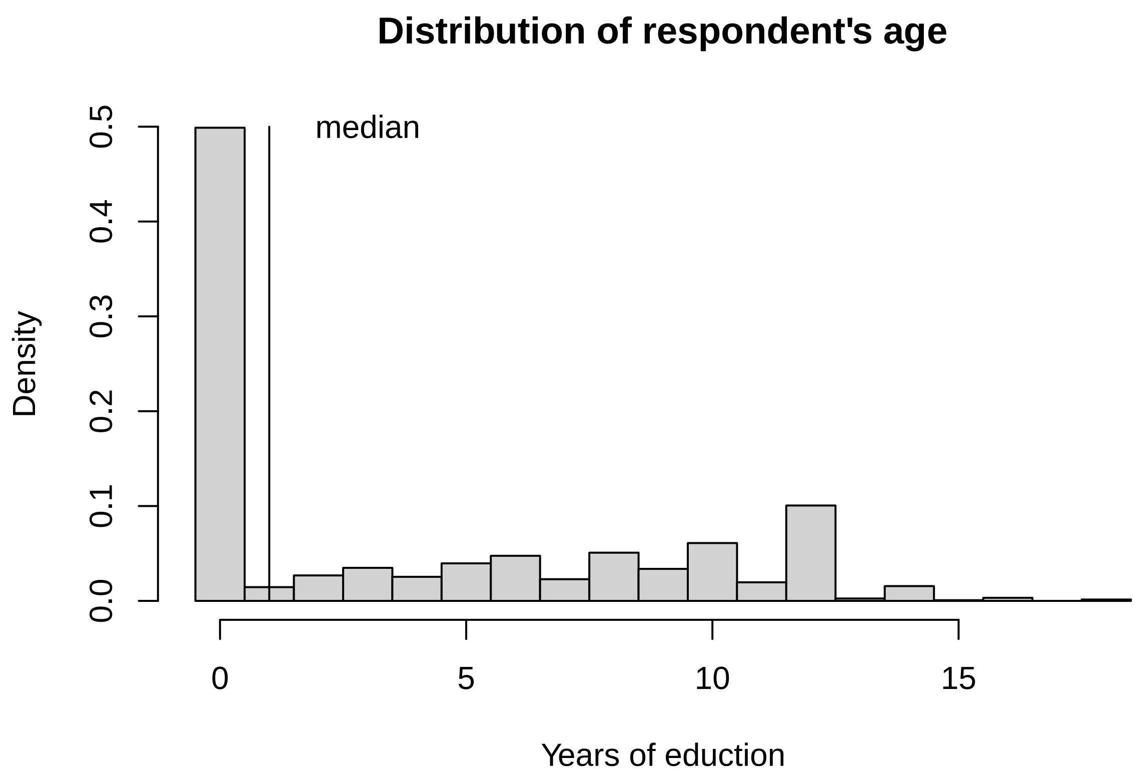
```r
hist(afghan$age, freq = FALSE, ylim = c(0, 0.04), xlab = "Age",
     main = "Distribution of respondent's age")
```



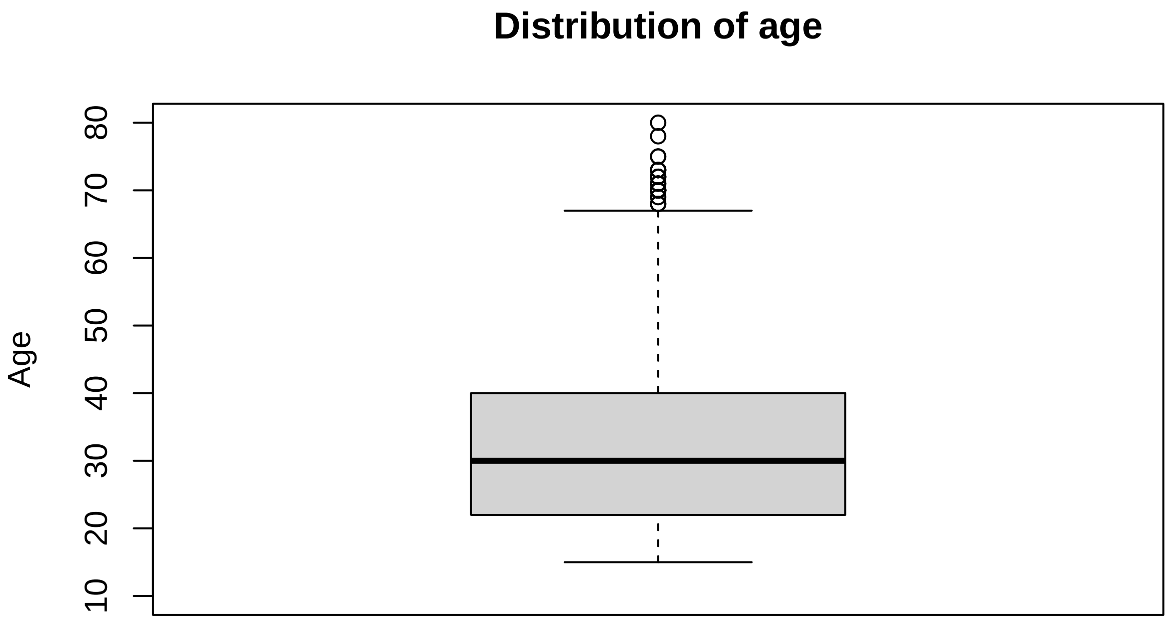
```
hist(afghan$educ.years, freq = FALSE,
     breaks = seq(from = -0.5, to = 18.5, by = 1),
     xlab = "Years of eduction",
     main = "Distribution of respondent's age")
text(x = 3, y = 0.5, "median") # 文本标签 “median” 出现在 (3, 0.5) 的位置
abline(v = median(afghan$educ.years)) # 在中位数处绘制一条垂直线
```



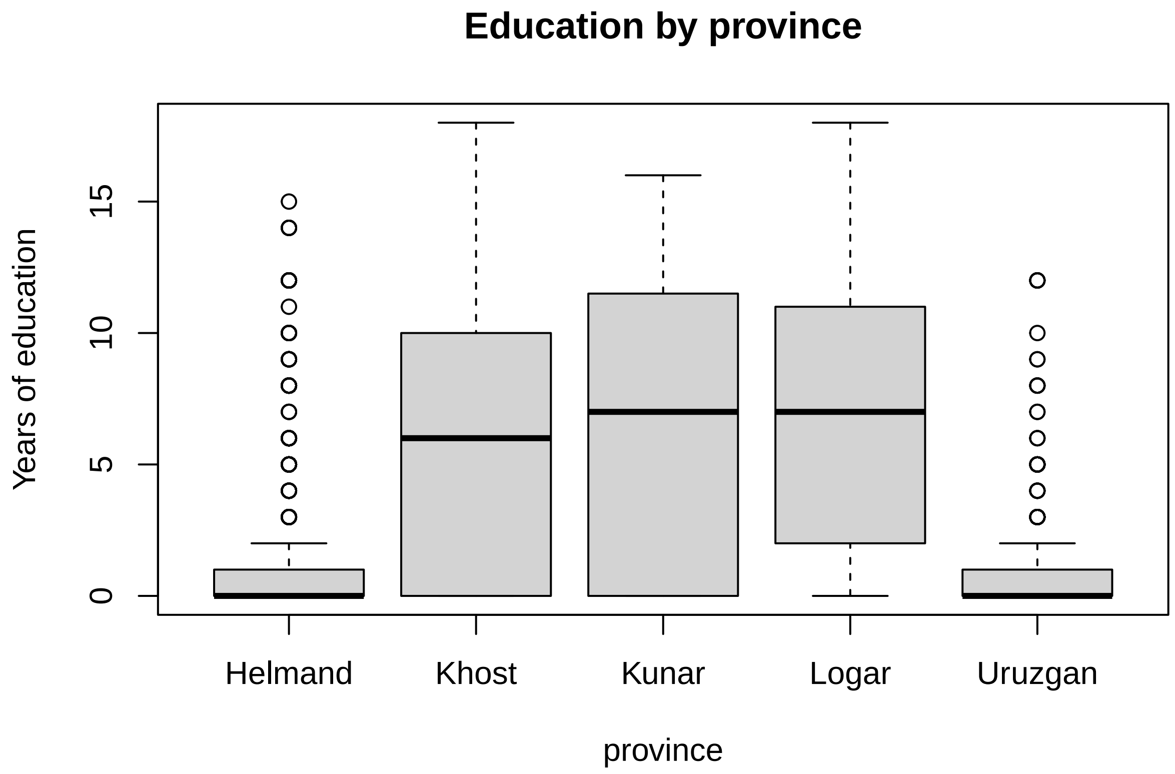
```
hist(afghan$educ.years, freq = FALSE,
     breaks = seq(from = -0.5, to = 18.5, by = 1),
     xlab = "Years of eduction",
     main = "Distribution of respondent's age")
text(x = 3, y = 0.5, "median") # 文本标签 “median” 出现在 (3, 0.5) 的位置
lines(x = rep(median(afghan$educ.years), 2), y = c(0, 0.5)) # 线在直方图底部和顶部之间延申
```



```
# 年龄的分布情况
boxplot(afghan$age, main = "Distribution of age", ylab = "Age",
        ylim = c(10, 80))
```



```
# 各省教育年份的分布情况
boxplot(educ.years ~ province, data = afghan,
        main = "Education by province", ylab = "Years of education")
```



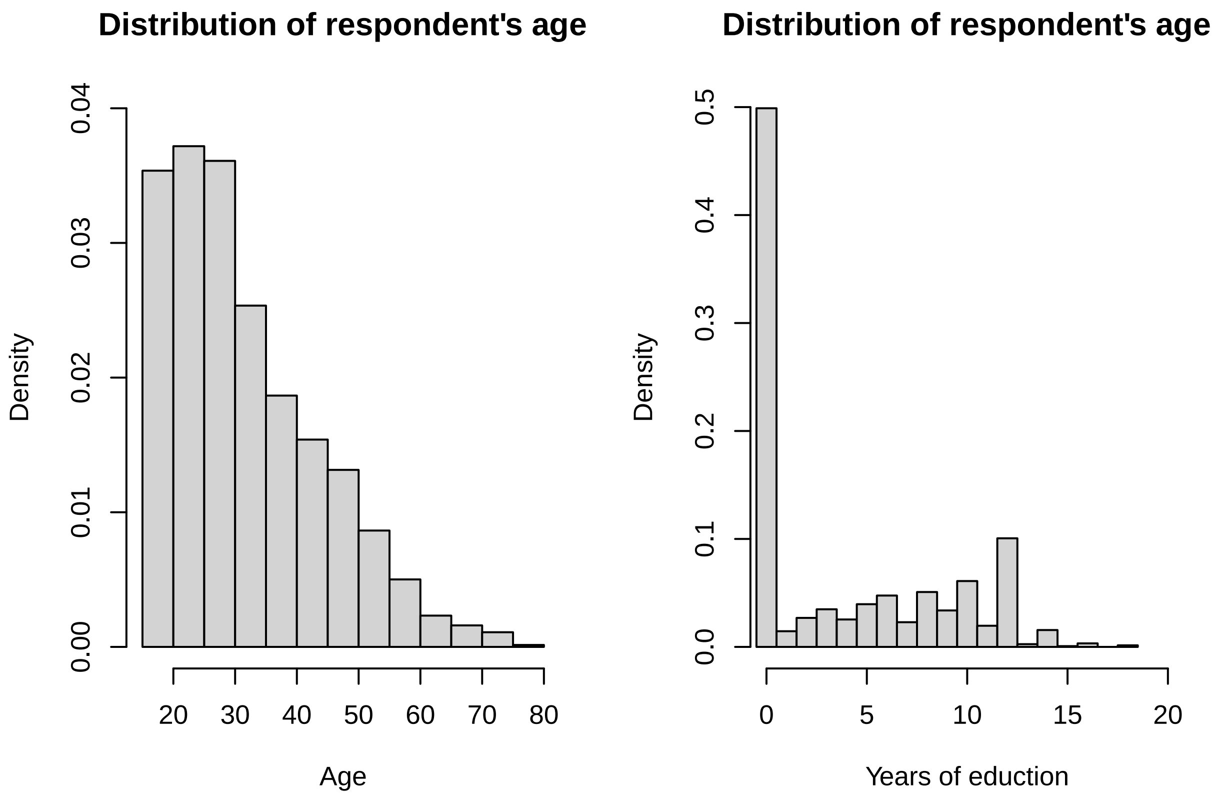
```
# 计算各个省份对相应问题的肯定回答的比例
tapply(afghan$violent.exp.taliban, afghan$province, mean, na.rm = TRUE)
```

```
##    Helmand      Khost      Kunar      Logar    Uruzgan
## 0.50422195 0.23322684 0.30303030 0.08024691 0.45454545
```

```
tapply(afghan$violent.exp.ISAF, afghan$province, mean, na.rm = TRUE)
```

```
##   Helmand     Khost     Kunar     Logar   Uruzgan
## 0.5410226 0.2424242 0.3989899 0.1440329 0.4960422
```

```
# pdf(file = "hist.pdf", height = 4, width = 8)

par(mfrow = c(1, 2), cex = 0.8)

hist(afghan$age, freq = FALSE,
     xlab = "Age", ylim = c(0, 0.04),
     main = "Distribution of respondent's age")

hist(afghan$educ.years, freq = FALSE,
     breaks = seq(from = -0.5, to = 18.5, by = 1),
     xlab = "Years of eduction",
     xlim = c(0, 20),
     main = "Distribution of respondent's age")
```

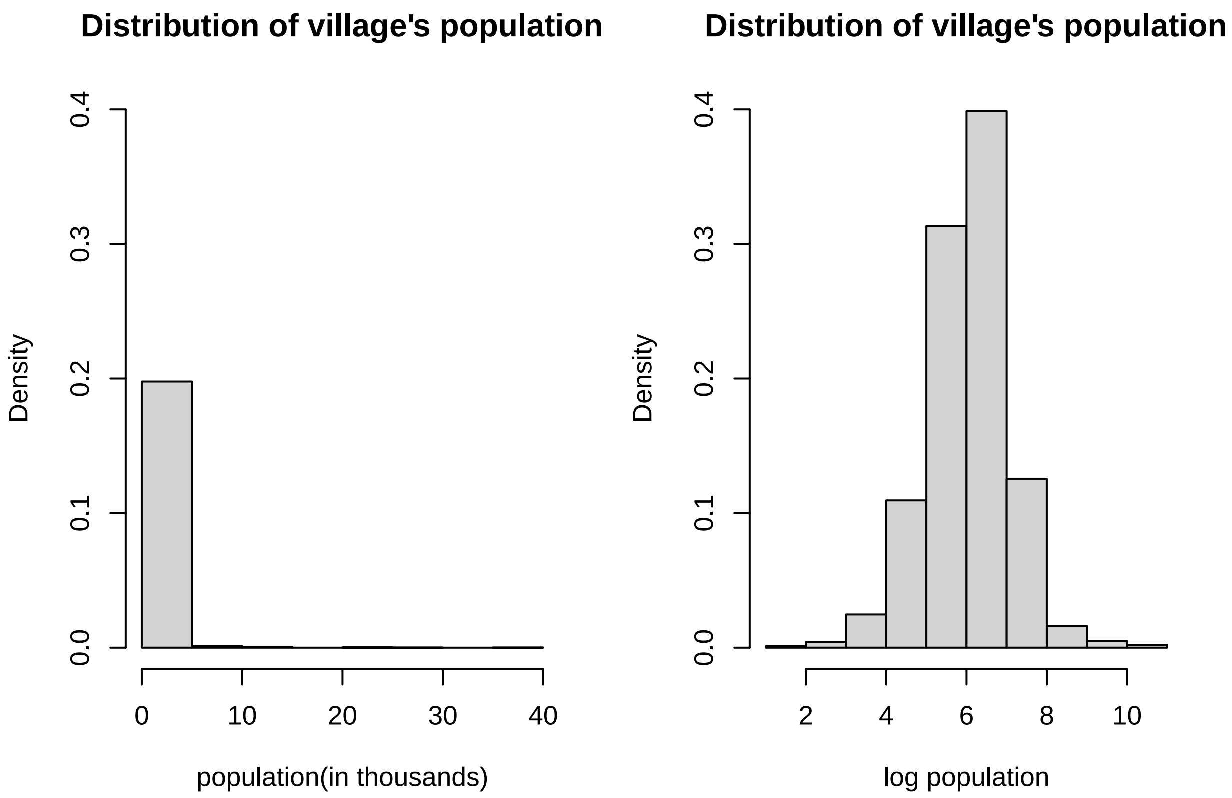


```
# dev.off()
```

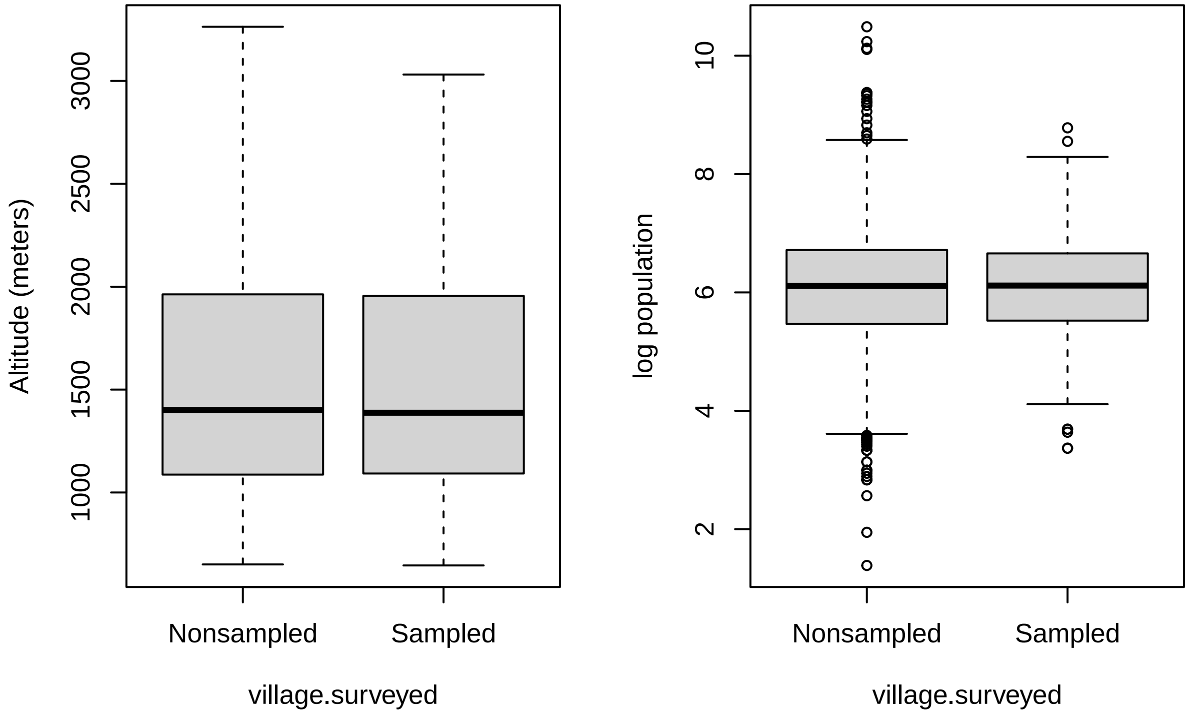
1 调查抽样

1.1 随机化的作用

```
afghan_village <- read.csv("afghan-village.csv")
# View(afghan_village)

# 以原始尺度（以千计和对数尺度）展示阿富汗村庄人口的直方图。没有对数转换，人口分布就会偏离
par(mfrow = c(1, 2), cex = 0.8)
hist((afghan_village$population / 1000), freq = FALSE,
     breaks = seq(from = 0, to = 40, by = 5),
     ylim = c(0, 0.4),
     xlab = "population (in thousands) ",
     main = "Distribution of village's population")

hist(log(afghan_village$population), freq = FALSE,
     ylim = c(0, 0.4),
     xlab = "log population",
     main = "Distribution of village's population")
```



```
boxplot(altitude ~ village.surveyed, data = afghan_village,
        ylab = "Altitude (meters)", names = c("Nonsampled", "Sampled"))

boxplot(log(population) ~ village.surveyed, data = afghan_village,
        ylab = "log population", names = c("Nonsampled", "Sampled"))
```



1.2 拒访和其他偏误来源

```
tapply(is.na(afghan$violent.exp.taliban), afghan$province, mean)
```

```
##     Helmand      Khost      Kunar      Logar    Uruzgan
## 0.030409357 0.006349206 0.000000000 0.000000000 0.062015504
```

```
tapply(is.na(afghan$violent.exp.ISAF), afghan$province, mean)
```

```
##     Helmand      Khost      Kunar      Logar    Uruzgan
## 0.016374269 0.004761905 0.000000000 0.000000000 0.020671835
```

```
mean(afghan$list.response[afghan$list.group == "ISAF"]) -
    mean(afghan$list.response[afghan$list.group == "control"])
```

```
## [1] 0.04901961
```

```
table(response = afghan$list.response, group = afghan$list.group)
```

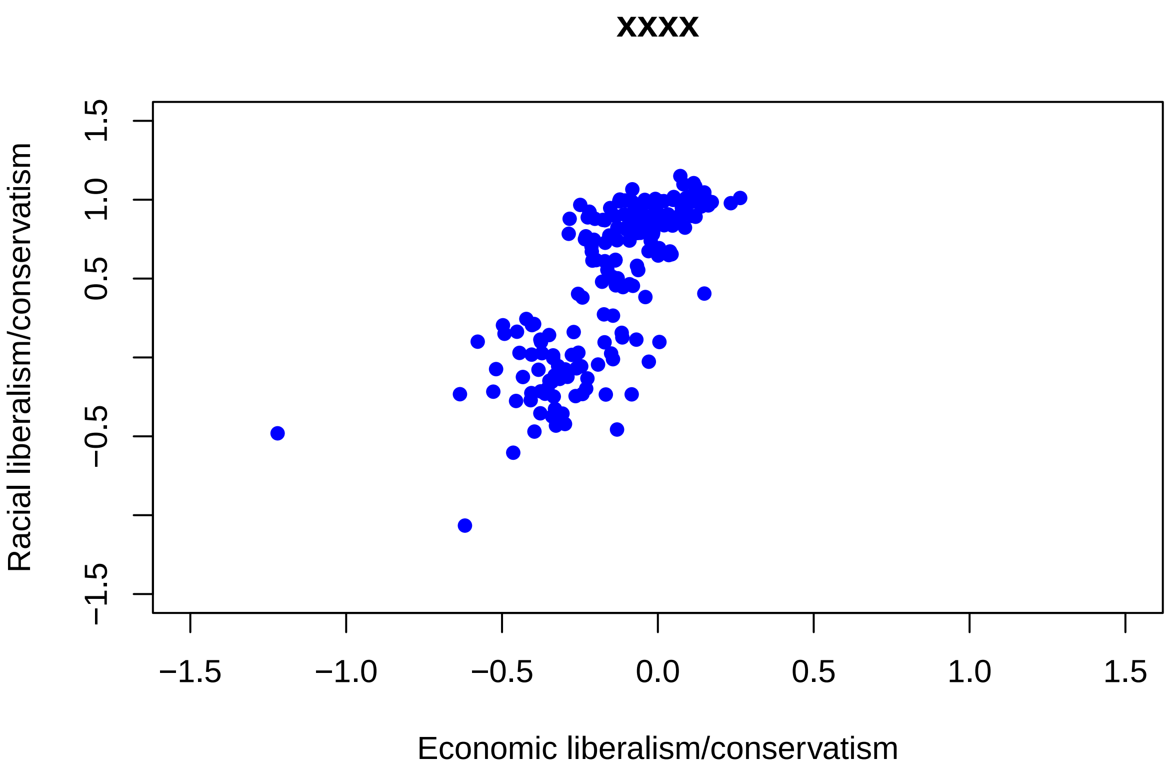
```
##         group
## response control ISAF taliban
##        0     188  174       0
##        1     265  278     433
##        2     265  260     287
##        3     200  182     198
##        4       0   24       0
```

2 度量政治极化

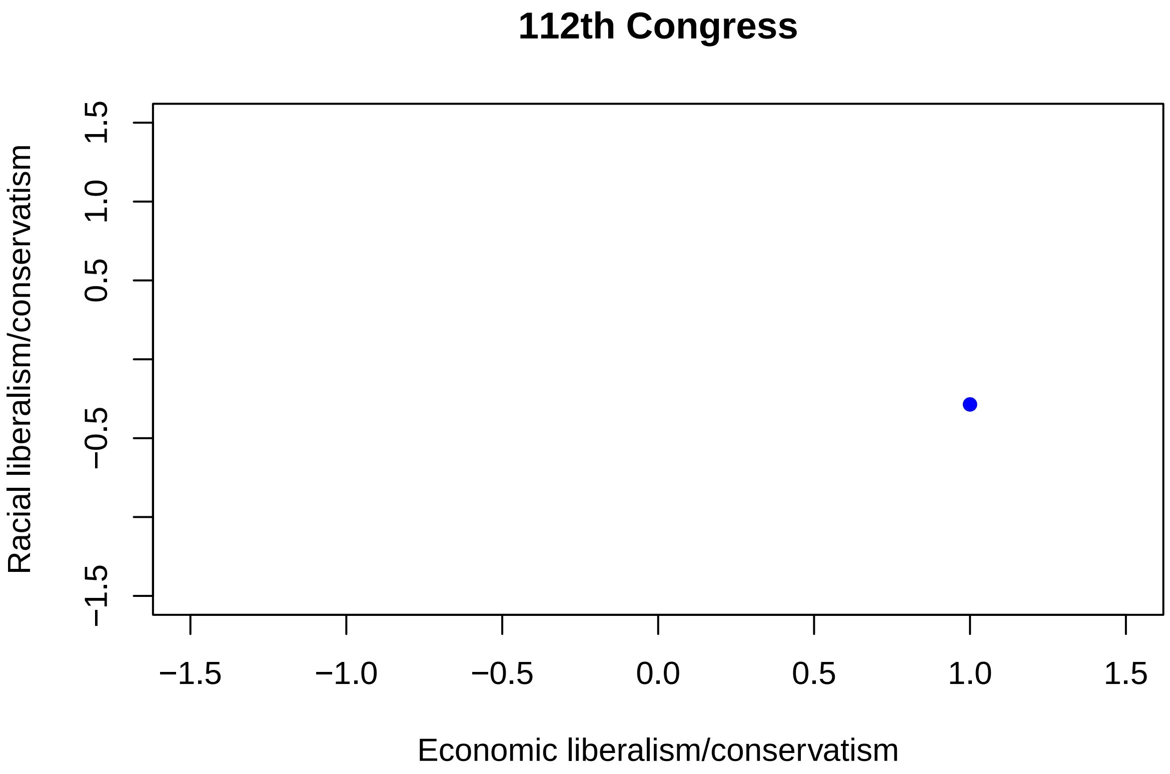
3 概括双变量关系

3.1 散点图

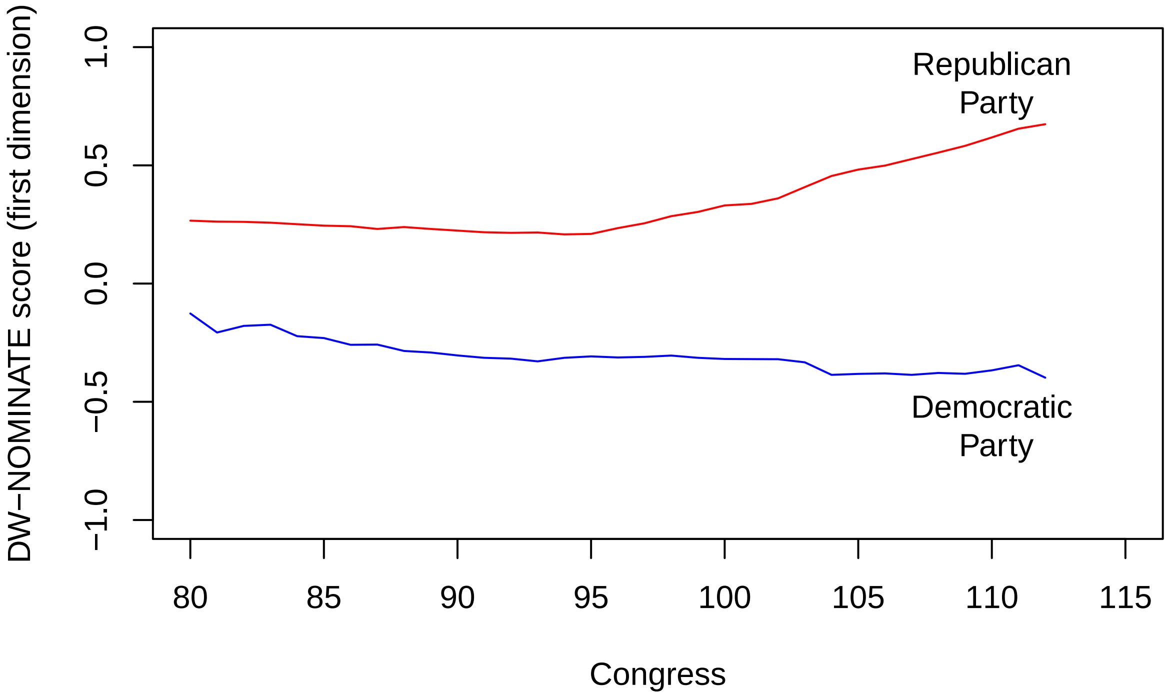
```r
congress <- read.csv("congress.csv")
# View(congress)

rep <- subset(congress, subset = (party == "Republican"))
dem <- congress[congress$party == "Democrat", ] # 另一种取子集的方法

rep80 <- subset(rep, subset = (congress == 80))
dem80 <- subset(dem, subset = (congress == 80))
rep112 <- subset(rep, subset = (congress == 112))
dem112 <- subset(dem, subset = (congress == 112))

xlab <- "Economic liberalism/conservatism"
ylab <- "Racial liberalism/conservatism"
lim <- c(-1.5, 1.5)

# example(points)

# plot(dem80$dwnom1, dem80$dwnom2, pch = 16, col = "blue",
#      xlim = lim, ylim = lim, xlab = xlab, ylab = ylab,
#      main = "80th Congress") # 支持材料里的
# plot(dem80$dwnom1, dem80$dwnow2, pch = 16, col = "blue",
#      xlim = lim, ylim = lim, xlab = xlab, ylab = ylab,
#      main = "80th Congress") # 自己敲的
plot(dem80$dwnom1, dem80$dwnom2, pch = 16, col = "blue",
     xlim = lim, ylim = lim, xlab = xlab, ylab = ylab,
     main = "xxxx")
```



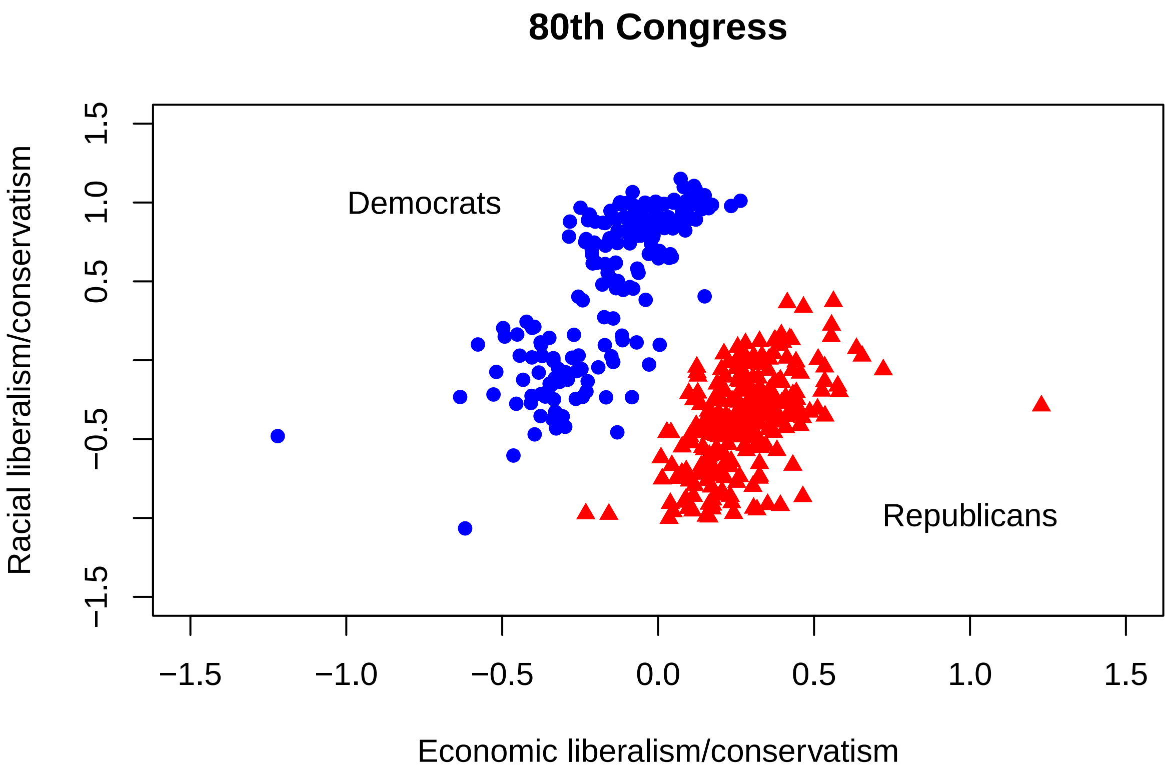
```
# points(rep80$dwnom1, rep80$dwnom2, pch = 17, col = "red")
# text(-0.75, 1, "Demorats")
# text(1, -1, "Republicans")
plot(dem112$dwnom1, dem112$dwnow2, pch = 16, col = "blue",
      xlim = lim, ylim = lim, xlab = xlab, ylab = ylab,
      main = "112th Congress")
```



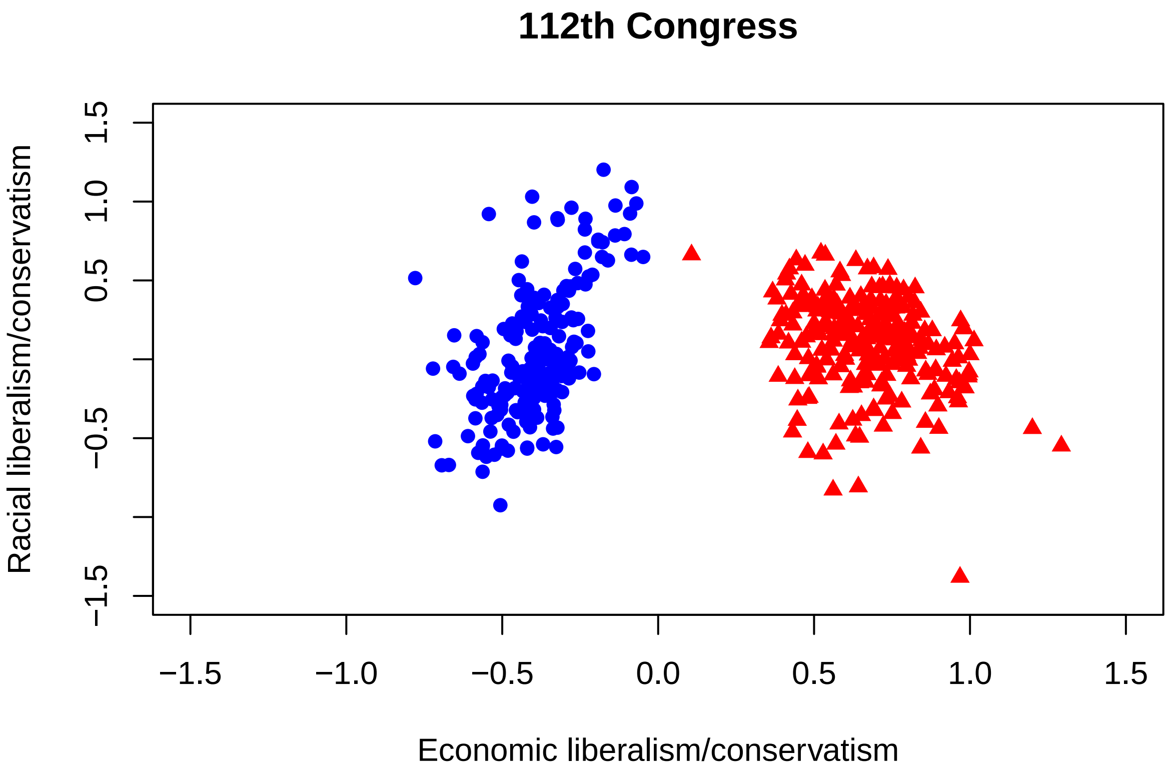
```
# plot(dem112$dwnom1, dem112$dwnom2, pch = 16, col = "blue",
#      xlim = lim, ylim = lim, xlab = xlab, ylab = ylab,
#      main = "112th Congress")
# points(rep112$dwnom1, rep112$dwnom2, pch = 17, col = "red")

dem.median <- tapply(dem$dwnom1, dem$congress, median)
rep.median <- tapply(rep$dwnom1, rep$congress, median)

plot(names(dem.median), dem.median, col = "blue", type = "l",
     xlim = c(80, 115), ylim = c(-1, 1), xlab = "Congress",
     ylab = "DW-NOMINATE score (first dimension)")
lines(names(rep.median), rep.median, col = "red")
text(110, -0.6, "Democratic\n Party")
text(110, 0.85, "Republican\n Party")
```



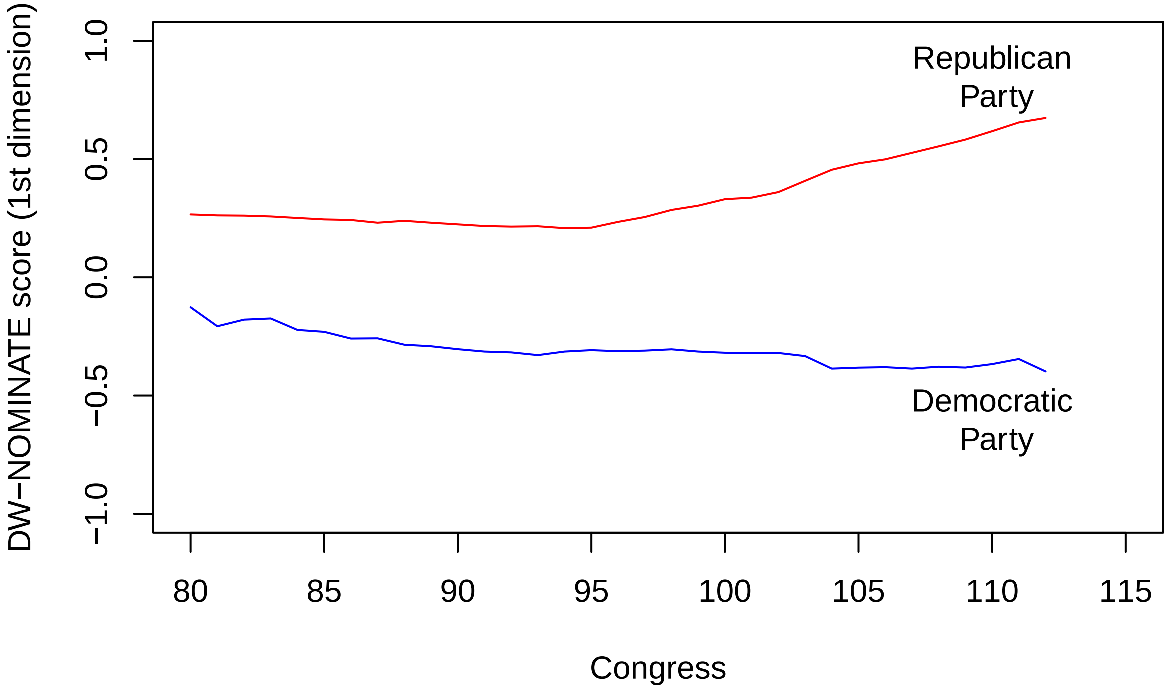
```
congress <- read.csv("congress.csv")

## 按党派取出子集
rep <- subset(congress, subset = (party == "Republican"))
dem <- congress[congress$party == "Democrat", ] # 另一种取子集的方式
## 取出第 80 届和第 112 届两个党派的子集
rep80 <- subset(rep, subset = (congress == 80))
dem80 <- subset(dem, subset = (congress == 80))
rep112 <- subset(rep, subset = (congress == 112))
dem112 <- subset(dem, subset = (congress == 112))

## 使用同一组坐标轴标签和数值范围创建多个散点图
xlab <- "Economic liberalism/conservatism"
ylab <- "Racial liberalism/conservatism"
lim <- c(-1.5, 1.5)
## 绘制第 80 届国会的散点图
plot(dem80$dwnom1, dem80$dwnom2, pch = 16, col = "blue",
     xlim = lim, ylim = lim, xlab = xlab, ylab = ylab,
     main = "80th Congress") # 民主党
points(rep80$dwnom1, rep80$dwnom2, pch = 17, col = "red") # 共和党
text(-0.75, 1, "Democrats")
text(1, -1, "Republicans")
```



```
## 绘制第 112 届国会的散点图
plot(dem112$dwnom1, dem112$dwnom2, pch = 16, col = "blue",
     xlim = lim, ylim = lim, xlab = xlab, ylab = ylab,
     main = "112th Congress")
points(rep112$dwnom1, rep112$dwnom2, pch = 17, col = "red")
```



```r
## 得到每届国会的民主党和共和党的中位立法者 (现在只看第一维度经济维度
dem.median <- tapply(dem$dwnom1, dem$congress, median)
rep.median <- tapply(rep$dwnom1, rep$congress, median)

## 创建一个折线图，观察两党的中位议员如何随时间变化
plot(names(dem.median), dem.median, col = "blue", type = "l",
     xlim = c(80, 115), ylim = c(-1, 1), xlab = "Congress",
     ylab = "DW-NOMINATE score (1st dimension)") # 民主党
lines(names(rep.median), rep.median, col = "red") # 共和党
text(110, -0.6, "Democratic\n Party")
text(110, 0.85, "Republican\n Party")
```

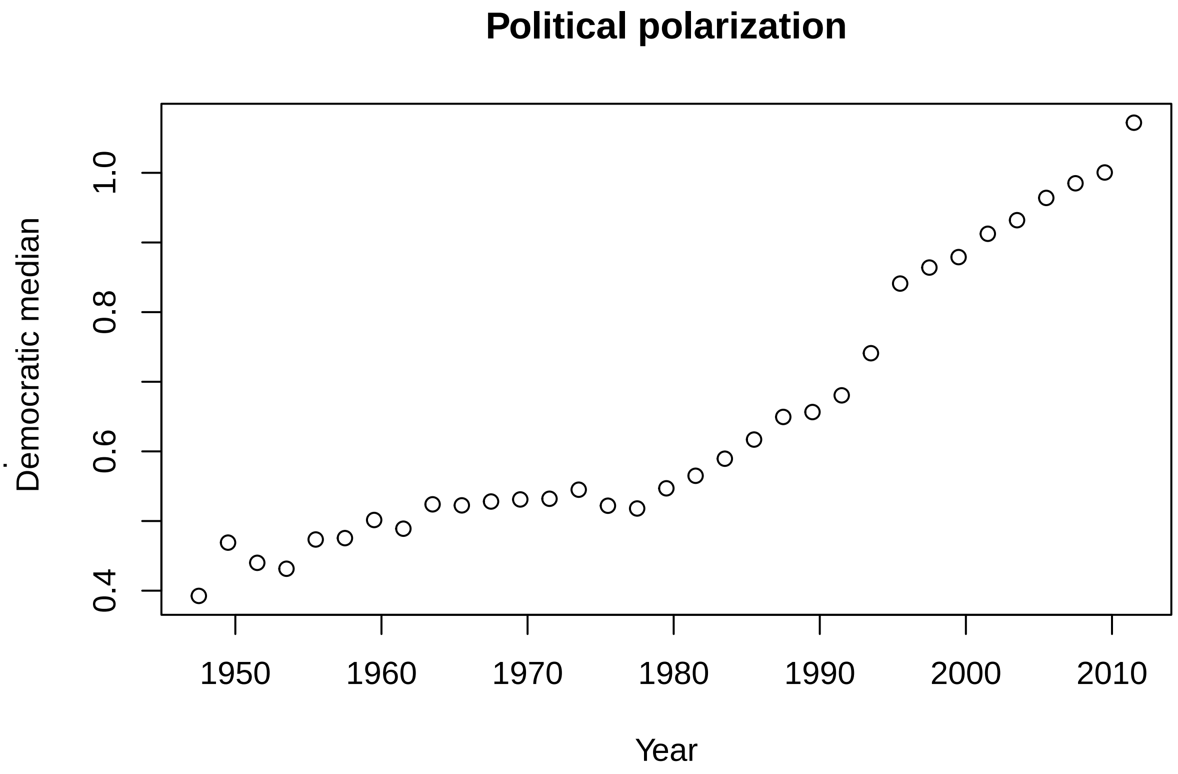


3.2 相关性

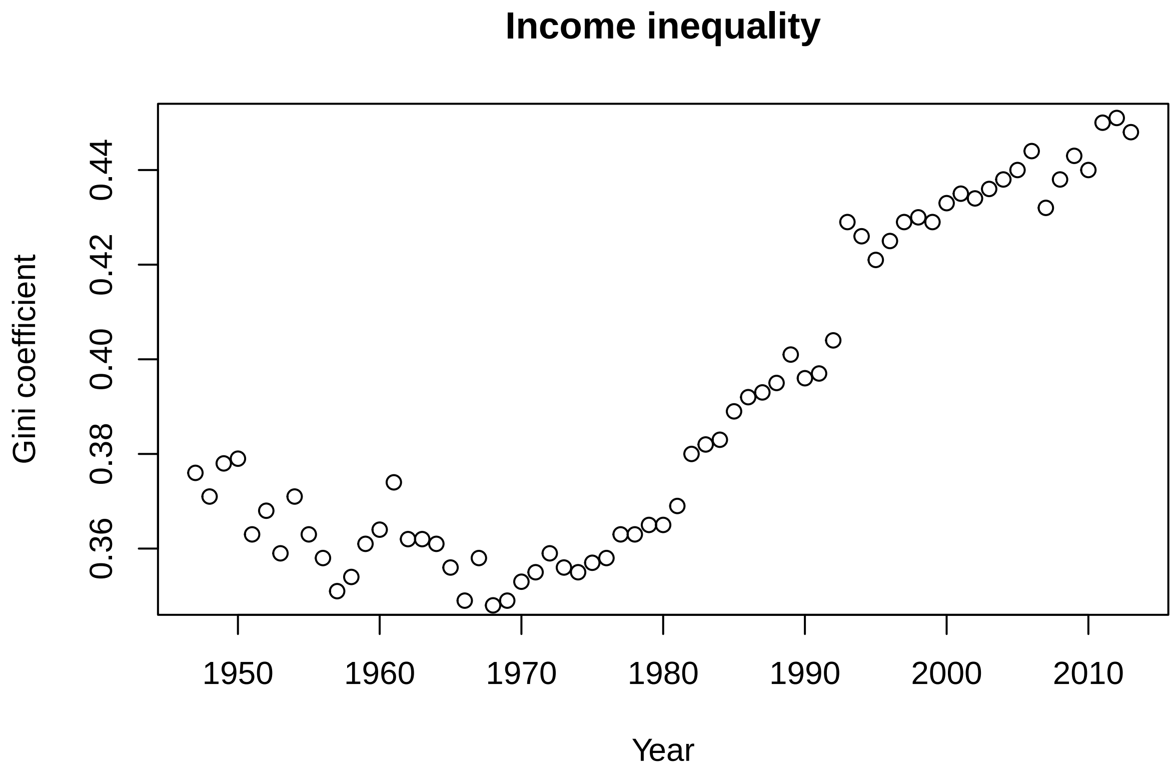
```
gini <- read.csv("USGini.csv")
range(gini$year) # 1947 年到 2013 年
```

```
## [1] 1947 2013
```

```
plot(seq(from = 1947.5, to = 2011.5, by = 2), rep.median - dem.median,
     xlab = "Year", ylab = "Republican median -\n Democratic median",
     main = "Political polarization")
```



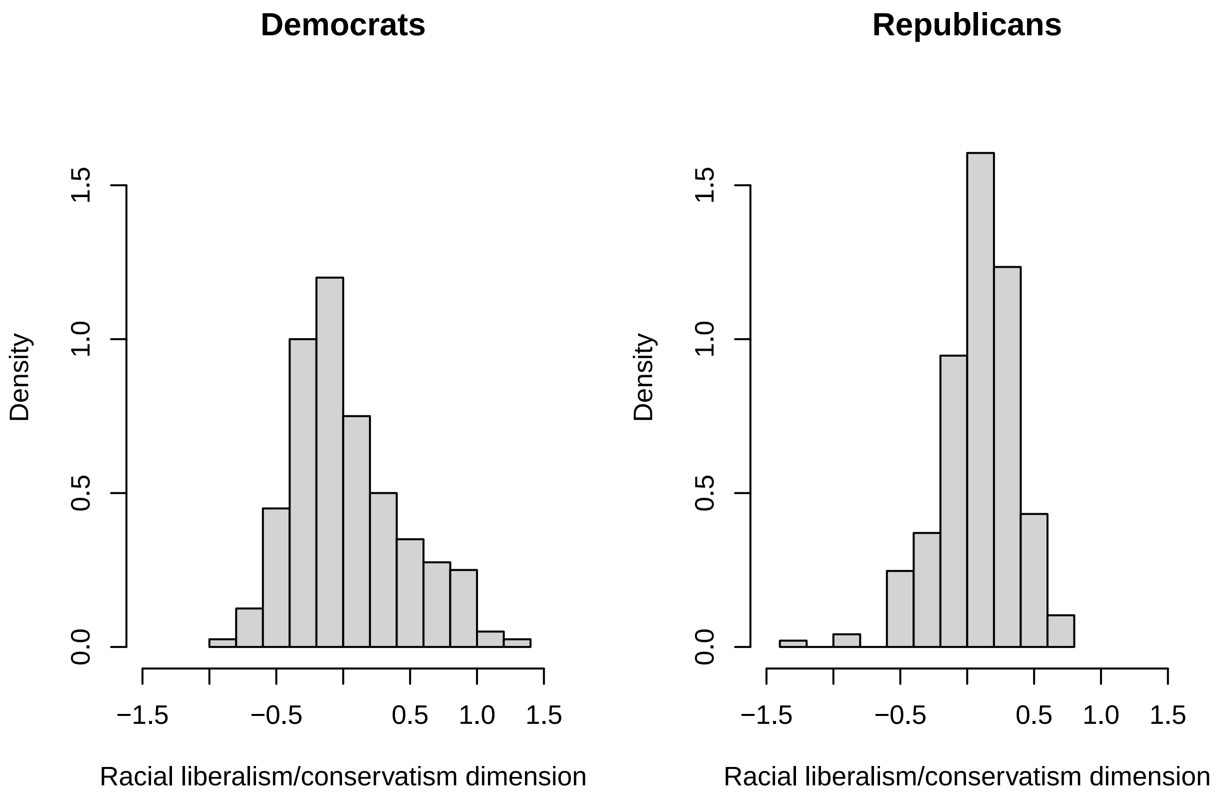
```
plot(gini$year, gini$gini,
     ylim = c(0.35, 0.45), xlab = "Year",
     ylab = "Gini coefficient", main = "Income inequality")
```



```
cor(gini$gini[seq(from = 2, to = nrow(gini), by = 2)],
    rep.median - dem.median)
```

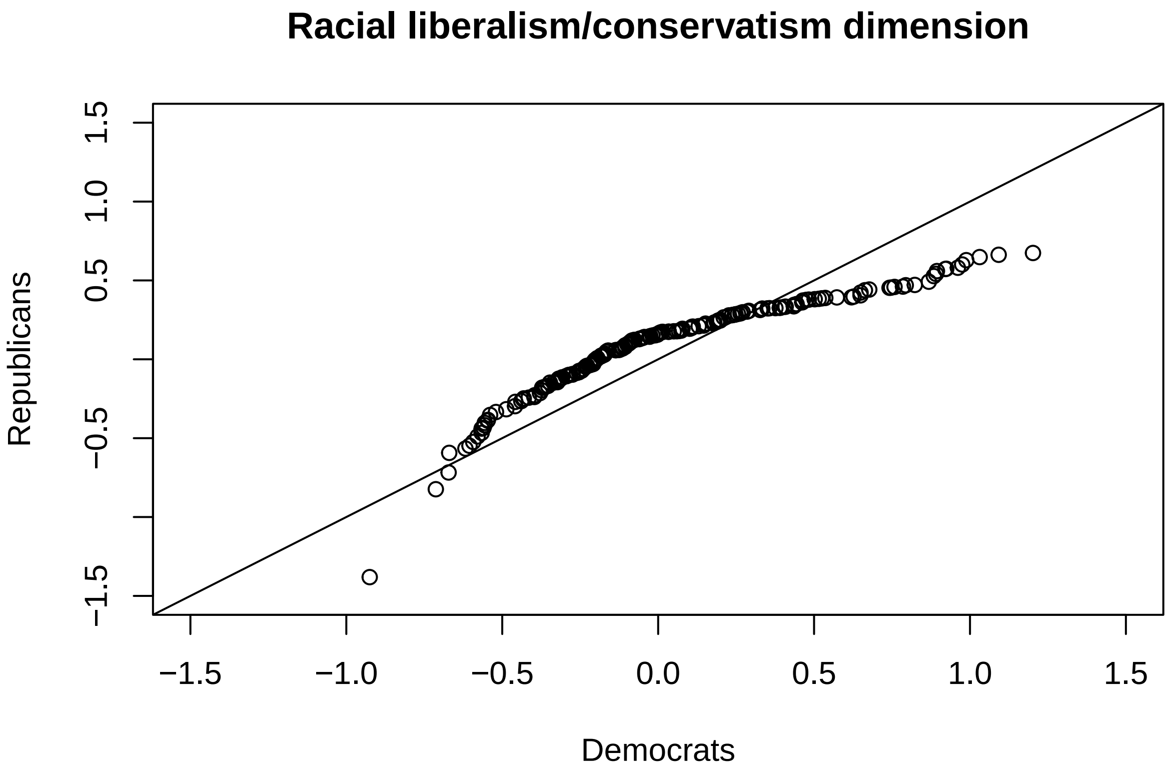
```
## [1] 0.9418128
```

```
par(mfrow = c(1, 2), cex = 0.8)
hist(dem112$dwnom2, freq = FALSE, main = "Democrats",
     xlim = c(-1.5, 1.5), ylim = c(0, 1.75),
     xlab = "Racial liberalism/conservatism dimension")
hist(rep112$dwnom2, freq = FALSE, main = "Republicans",
     xlim = c(-1.5, 1.5), ylim = c(0, 1.75),
     xlab = "Racial liberalism/conservatism dimension")
```



3.3 分位数-分位数图（Q-Q 图）

```
qqplot(dem112$dwnom2, rep112$dwnom2, xlab = "Democrats",
       ylab = "Republicans", xlim = c(-1.5, 1.5), ylim = c(-1.5, 1.5),
       main = "Racial liberalism/conservatism dimension")
abline(0, 1)
```



4 聚类

4.1 R 中的矩阵

```
x <- matrix(1:12, nrow = 3, ncol = 4, byrow = TRUE)
rownames(x) <- c("a", "b", "c")
colnames(x) <- c("d", "e", "f", "g")
dim(x)
```

```
## [1] 3 4
```

```
x
```

```
##   d  e  f  g
## a 1  2  3  4
## b 5  6  7  8
## c 9 10 11 12
```

```r
y <- data.frame(y1 = as.factor(c("a", "b", "c")), y2 = c(0.1, 0.2, 0.3))
class(y$y1)
```

```
## [1] "factor"
```

```r
class(y$y2)
```

```
## [1] "numeric"
```

```r
z <- as.matrix(y)
z
```

```
##      y1  y2   
## [1,] "a" "0.1"
## [2,] "b" "0.2"
## [3,] "c" "0.3"
```

```r
# colSum(), colMeans(), rowSum(), rowMean() 函数
colSums(x)
```

```
##  d  e  f  g 
## 15 18 21 24
```

```r
rowMeans(x)
```

```
##    a    b    c 
##  2.5  6.5 10.5
```

```r
apply(x, 2, sum)
```

```
##  d  e  f  g 
## 15 18 21 24
```

```r
apply(x, 1, mean)
```

```
##    a    b    c 
##  2.5  6.5 10.5
```

```r
apply(x, 1, sd)
```

```
##        a        b        c 
## 1.290994 1.290994 1.290994
```

4.2 R 中的列表

```
x <- list(y1 = 1:10, y2 = c("hi", "hello", "hey"),
          y3 = data.frame(z1 = 1:3, z2 = c("good", "bad", "ugly")))
# 三种从列表中提取元素的方法
x$y1
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

```
x[[2]]
```

```
## [1] "hi"    "hello" "hey"
```

```
x[["y3"]]
```

```
##   z1   z2
## 1  1 good
## 2  2  bad
## 3  3 ugly
```

```
names(x)
```

```
## [1] "y1" "y2" "y3"
```

```
length(x)
```

```
## [1] 3
```

4.3 k 均值算法

```
dwnom80 <- cbind(congress$dwnom1[congress$congress == 80],
                 congress$dwnom2[congress$congress == 80])
dwnom112 <- cbind(congress$dwnom1[congress$congress == 112],
                  congress$dwnom2[congress$congress == 112])

## 聚成两个类
k80two.out <- kmeans(dwnom80, centers = 2, nstart = 5)
k112two.out <- kmeans(dwnom112, centers = 2, nstart = 5)

names(k80two.out)
```

```
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
k80two.out$centers
```

```
##          [,1]       [,2]
## 1 -0.04843704  0.7827259
## 2  0.14681029 -0.3389293
```

```
k112two.out$centers
```

```
##         [,1]       [,2]
## 1 -0.3912687 0.03260696
## 2  0.6776736 0.09061157
```

```
## 创建党派和聚类标签变量的交叉列表来计算属于每个聚类的民主党和共和党议员的数量
table(party = congress$party[congress$congress == 80],
      cluster = k80two.out$cluster)
```

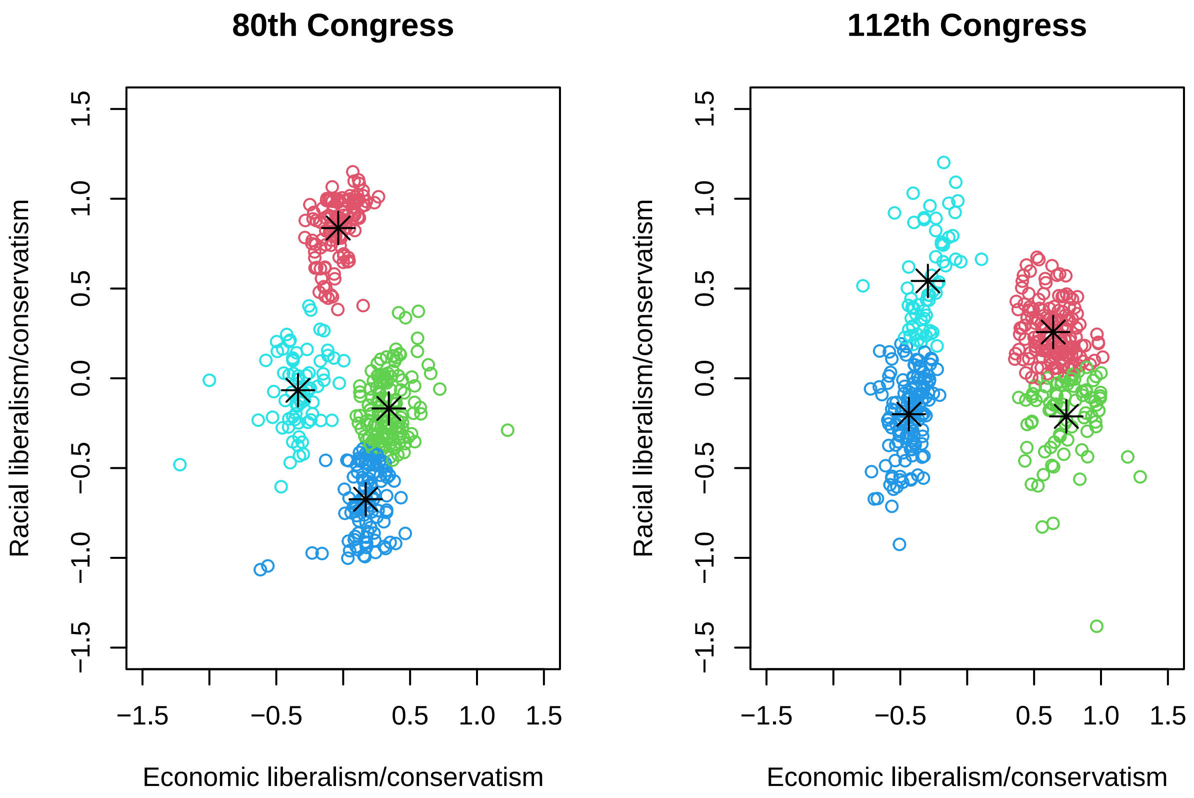
```
##              cluster
## party          1   2
##   Democrat   132  62
##   Other        0   2
##   Republican   3 247
```

```
table(party = congress$party[congress$congress == 112],
      cluster = k112two.out$cluster)
```

```
##              cluster
## party          1   2
##   Democrat   200   0
##   Republican   1 242
```

```
# xlab <- "Economic liberalism/conservatism"
# ylab <- "Racial liberalism/conservatism"
# lim <- c(-1.5, 1.5)



## 聚成四个类
k80four.out <- kmeans(dwnom80, centers = 4, nstart = 5)
k112four.out <- kmeans(dwnom112, centers = 4, nstart = 5)

par(mfrow = c(1, 2), cex = 0.8)
## 绘制第 80 届国会的四个聚类的散点图
```

```
plot(dwnom80, col = k80four.out$cluster + 1, xlab = xlab, ylab = ylab,
                  xlim = lim, ylim = lim, main = "80th Congress")
## 绘制质心
points(k80four.out$centers, pch = 8, cex = 2)

## 绘制第 112 届国会的四个聚类的散点图
plot(dwnom112, col = k112four.out$cluster + 1, xlab = xlab, ylab = ylab,
                  xlim = lim, ylim = lim, main = "112th Congress")
points(k112four.out$centers, pch = 8, cex = 2)
```



```
# palette()
```