華中師範大學

《量化社会科学导论》
第三章：度量

张梦毅　2021年12月15日

華中師範大學
CENTRAL CHINA NORMAL UNIVERSITY

CONTENTS

01 章节引入

02 抽样调查

03 潜变量度量

04 课后习题

華中師範大學
CENTRAL CHINA NORMAL UNIVERSITY

01 章节引入

1.1 快速回顾

数据来源：

- 政府机构公开的数据集（比如人口普查、选举结果、经济指标等）
- 手工编码的数据（第二章习题2.8.3刺杀领导人的数据）
- 自行设计并进行**调查**（最低工资对就业的影响；劳动力市场中的种族歧视）

数据的偏差：

- 2.4.2中给选民发放不同的鼓励信息的明信片，**样本选择**是否受伦理和操作等影响？
- 2.1中给雇主发放求职简历，未回复的和回复的雇主若有系统性的不同对**结论**有影响吗？

数据的可观测性：

- 有些概念和变量几乎不可观测，如何**度量**到这些**潜变量**？
- 有时很事先并不明确数据的具体模式规律，如何做**探索性研究**？

1.2 章节结构

知识点清单：

- **抽样调查**
  - 可视化单变量分布
  - 抽样调查的实施
    - 代表性
    - 偏误
    - 衡量敏感问题
- **潜变量度量**
  - 借用理论模型
  - 概括双变量关系
  - 使用聚类方法（k-均值算法）

借助的案例及处理方法：

- **案例：**战争时期平民受伤情况的度量
- **数据集：**“afghan.csv”，”afghan-village.csv”
- **R中的相关处理和工具**
  - 处理缺失数据
  - 条形图
  - 直方图
  - 箱型图
  - 打印和保存图表

- **案例：**政治极化的度量
- **数据集：**“congress.csv”
- **R中的相关处理和工具**
  - 散点图
  - 相关性
  - 分位数-分位数图（Q-Q图）
  - R中的列表和矩阵
  - kmeans()函数

華中師範大學
CENTRAL CHINA NORMAL UNIVERSITY

02 抽样调查

2.1 案例概况

调查方法是最常见的数据收集模式

案例：战争时期平民受伤情况的度量

- 调查于2011年1-2月在阿富汗南部（叛乱的中心地区）开展，对2754名受访者进行调查（最初联系了3097名，参与率89%）
- 由于当地文化禁止访问者和女性公民交谈，受访者全为男性
- 询问“在过去一年中，由于国际安全援助部队ISAF/塔利班政府的行动，你和家人是否受到伤害”（伤害泛指人身伤害和财产损失）

表1 阿富汗调查数据

变量名	描述
province	受访者居住的省份
age	受访者的年龄
educ.year	受访者的教育年数
income	受访者月收入（5个等级）
violent.exp. ISAF	受访者是否有受到ISAF的威胁
violent.exp. taliban	受访者是否有受到塔利班的威胁
list.group	随机分配列表实验到不同组（control, ISAF, taliban）
list.response	列表实验的回答（0-4）

2.2 使用描述性统计数据总结分布

- **计算受到两种势力伤害的比例**
  - 37.3%(=17.7%+19.6%)受到ISAF伤害
  - 32.8%(=13.2%+19.6%)受到塔利班伤害
  - 阿富汗平民受到两方势力同等程度的伤害
- **处理R中缺失的数据**
  - R中缺失数据被编码为NA
  - 可以使用**is.na()函数**判断缺失值，若其参数为NA，则函数返回TRUE的逻辑值，否则返回FALSE
  - 配合sum()和mean()函数计算缺失数据的总数和比例（**P36，逻辑值强行转成二元变量**）

```
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))  # 缺失值综述
```

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

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

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

- **处理R中缺失的数据**
  - 一些包含max(), min()和median()在内的很多统计计算的函数**自带参数na.rm**，可将其设置为TRUE，可在应用函数之前删除所有丢失的数据
  - 不止如此，apply组函数、rowSums()等函数都可以添加na.rm参数
  - **table()函数**会自动忽略缺失数据，也可以通过将附加**参数exclude**设置为NULL包括所有数据，也包括缺失数据

```
tapply(STAR$g4math, STAR$kinder, mean, na.rm = TRUE)
```

```
##    辅导班    普通班      小班
## 707.6335 709.5214 709.1851
```

```
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
```

- **处理R中缺失的数据**
  - **na.omit()函数**可用于直接删除数据框中的某个观察值。若该观察值中**至少有一个**缺失数据，返回没有这些观察值的另一个数据集
  - 若将na.omit()应用于整个数据集，将返回一个比较小的数据子集，比只对收入变量应用整个函数所返回的子集小很多

```
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
```

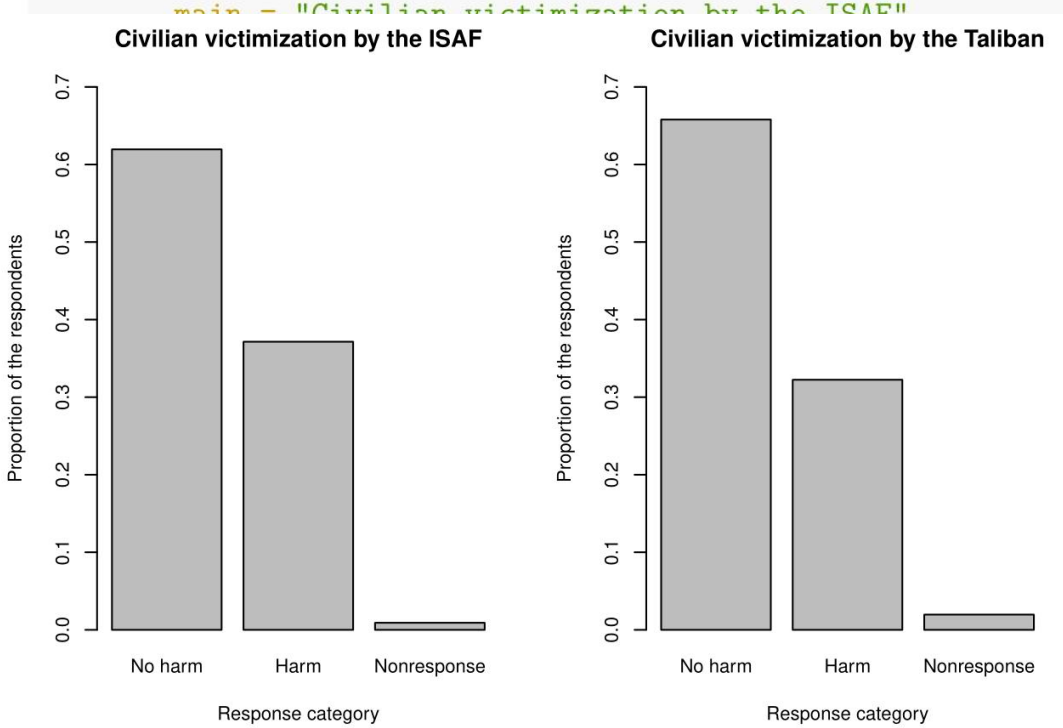
- **列表式删除**(listwise deletion/casewise deletion)
- **成对删除**(pairwise deletion)
  - 如果某条记录在其中一个配对变量中的数据缺失，则在进行这对配对变量的统计量计算时把含有缺失值的数据删除，在计算其他变量的统计量时不受影响。

2.3 可视化单变量分布

2.3.1 条形图（可视化分类变量）

- barplot()函数
  - main：绘图主标题的字符串
  - ylab, xlab：分别用于标记纵轴和横轴的字符串
  - ylim, xlim：长度为2的数字向量指定y轴和x轴间隔
  - names.arg：特有的可选参数，使用指定每个小书标签的字符常量
- par()函数
  - par(mfrow=c(X, Y))表示创建一个X行Y列的子图，图片按行一个个填充进格子中
  - cex参数更改字符或符号的大小，默认等于1

```
# 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"),
```



2.3 可视化单变量分布

- 当图象中含有中文且需要导出pdf时
  - 导入三个包

```
install.packages("showtext")
install.packages("sysfonts")
install.packages("showtextdb")
```

  - 添加头文件

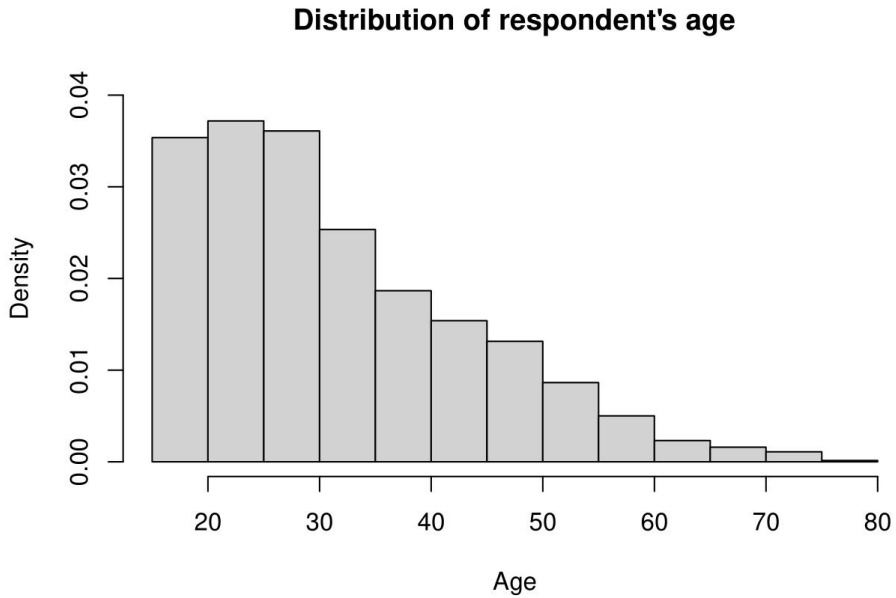
```
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE,
                      fig.showtext = TRUE #图表中可以输出中文
                      )
```
```

2.3 可视化单变量分布

- ggplot2包
  - 由Hadley Wickham创建的一个十分强大的可视化R绘图包
  - 绘图理念：Plot(图)= data(数据集)+ Aesthetics(美学映射)+ Geometry(几何对象)
  - 八大基本要素：
    - 数据：作图用的原始数据
    - 几何图形 geom_:表示数据的几何形状
    - 美学 aes(): 几何或者统计对象的美学，比如位置，颜色，大小，形状等
    - 刻度 scale_(): 数据与美学维度之间的映射，比如图形宽度的数据范围，
    - 统计转换 stat_: 数据的统计，比如百分位，拟合曲线或者和
    - 坐标系统 coord_: 数据的转换
    - 面 facet_: 数据图表的排列
    - 主题 theme(): 图形的整体视觉默认值，如背景、网格、轴、默认字体、大小和颜色
  - 可以随时更换不同要素的参数调整图形，更具灵活性

2.3 可视化单变量分布

2.3.2 直方图（可视化数值变量分布）



- **直方的密度（高度）**
  - 公式：密度 = $\frac{\text{直方中观测数所占比例}}{\text{直方的宽度}}$
  - 直方图中每个直方形的面积等于直方形中观察到的比例，所有直方形的面积总和为1
  - 对比较两种分布很有用，因为即使观测数量不同，密度尺度在各分布之间是可比较的

- **hist()函数**
  - freq：默认为TRUE，绘制频率（即计数）而非密度
  - breaks：为一个向量设置直方之间的断点，若不指定默认为[0,1), [1,2), ……也可接受一个整数来指定直方形的数目

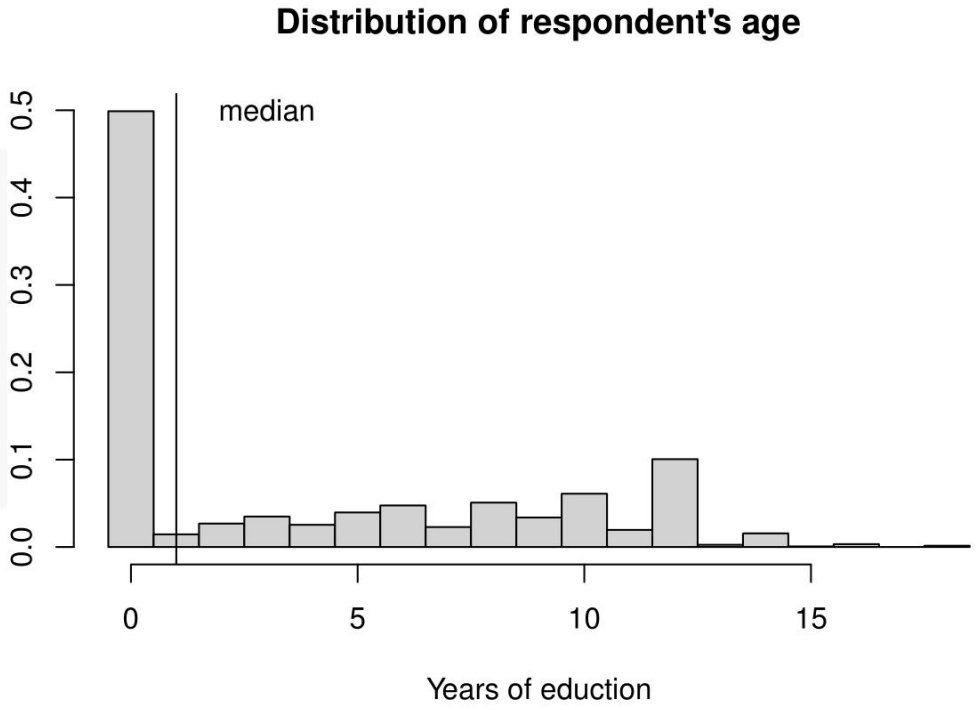
```
hist(afghan$age, freq = FALSE, ylim = c(0, 0.04), xlab = "Age",
     main = "Distribution of respondent's age")
```

2.3 可视化单变量分布

- **text()函数**
  - text(*x, y, z*)函数添加以坐标向量(*x, y*)指定的点为中心的字符文本*z*
- **abline()函数**
  - abline(*h = x*)：在点x处放一条水平线
  - abline(*v = x*)：在点x处放一条垂直线
  - abline(*a = y, b = s*)：用截距*y*和斜率*s*做出的线
- **lines()函数**
  - 有两个参数x和y，两个参数必须是分别具有相同数量x坐标和y坐标的向量
  - x中的第一个坐标和y中第一个坐标形成的点，连接到每个参数的第二个坐标表示的点，以此类推
- **rep()函数**
  - 第一个参数取想要重复的值，第二个参数取需要重复的次数，即所得到的向量的长度

2.3 可视化单变量分布

```r
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
```

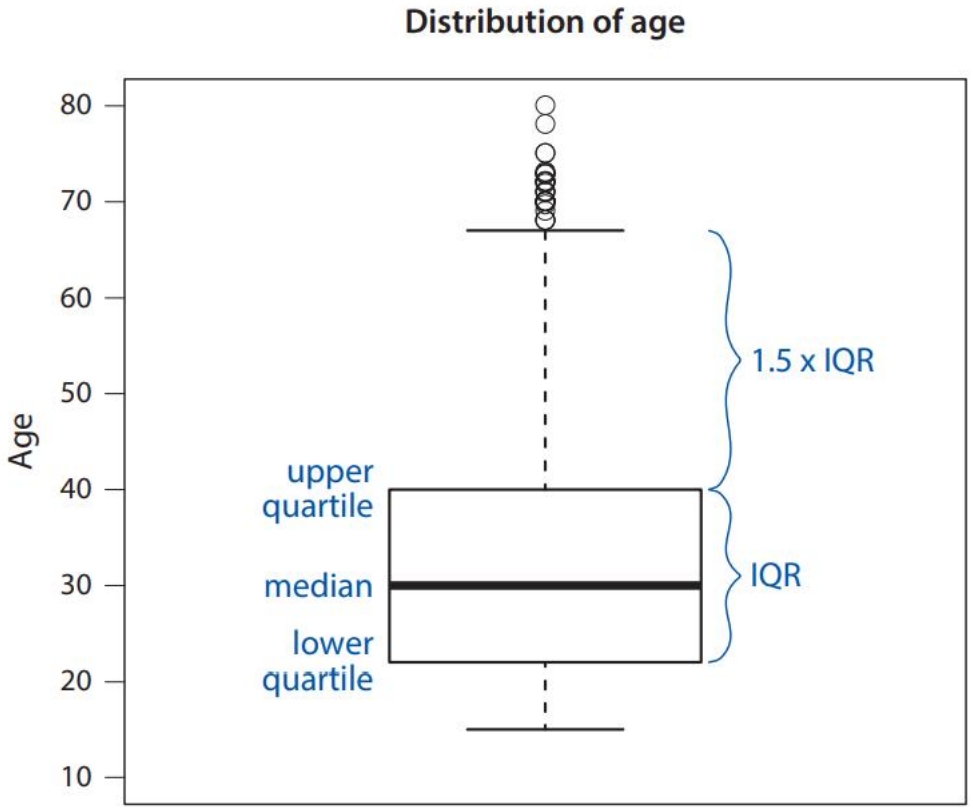



```r
hist(afghan$educ.years, freq = FALSE,
     breaks = seq(from = -0.5, to = 18.5, by = 1),
     xlab = "Years of eduction",
     main = "Distribution of respondent's age")
text(x = 3, y = 0.5, "median") # 文本标签 “median” 出现在 (3, 0.5) 的位置
lines(x = rep(median(afghan$educ.years), 2), y = c(0, 0.5)) #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

2.3 可视化单变量分布

2.3.3 箱形图（可视化数值变量分布）

- 适合将**几个变量**的分布并排放置
- 将**中位数、四分位数和IQR**作为单个对象一起可视化
- 框中包含50%的数据（从下四分位数到上四分位数）
- IQR表示两个四分位数之间的间距
- 两条虚线分别表示低于四分位数和高于四分位数的1.5IQR内的数据
- 虚线之外的结果用空心圆表示
- 可为不同的观察组创建箱型图，其中的组由一个**因子变量**定义



2.3 可视化单变量分布

- boxplot(y ~ x, data = d)
  - “~”是一种运算符，“y ~ x”表示一个公式formula
  - 公式fomula是一个把响应变量（在~左侧）和解释变量（在~右侧）联系起来的对象

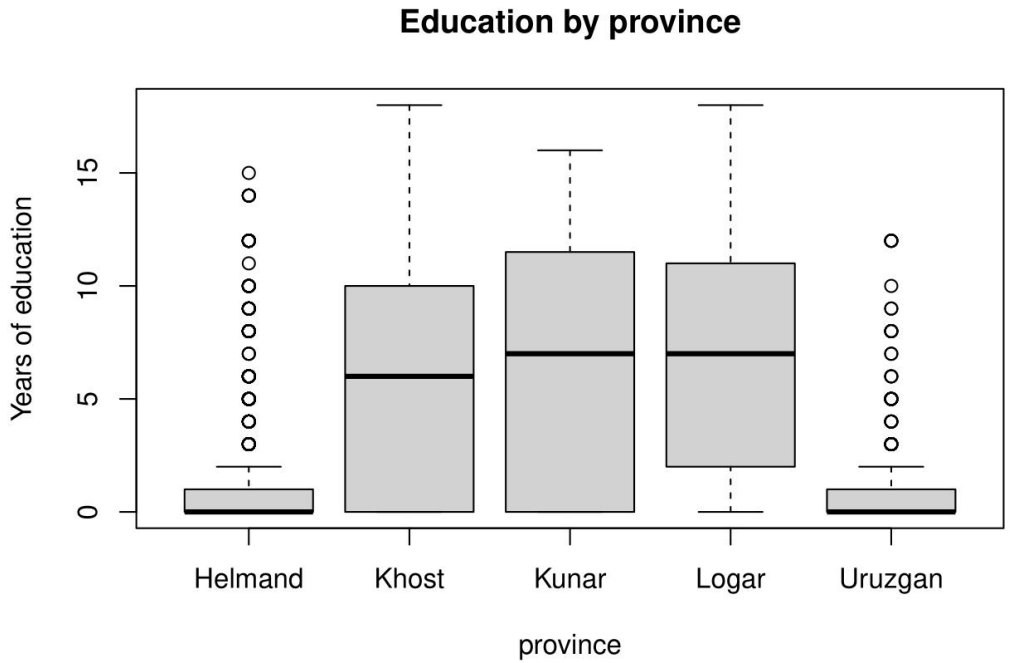
```
# 各省教育年份的分布情况
boxplot(educ.years ~ province, data = afghan,
        main = "Education by province", ylab = "Years of education")
```

```
# 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)
```

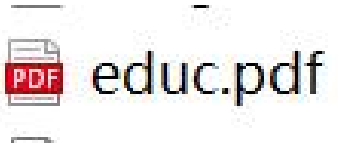
```
##    Helmand      Khost      Kunar      Logar    Uruzgan
## 0.5410226 0.2424242 0.3989899 0.1440329 0.4960422
```



2.3.4 打印及保存图表

- 指“**将图像单独保存为一个文件**”
- 使用pdf()函数在绘图命令之前打开PDF设备
- 可以以英寸为单位指定图形区域的高度和宽度
- 之后使用dev.off()函数关闭设备来使用命令保存或打印图形

```
pdf(file = "educ.pdf", height = 5, width = 5)
boxplot(educ.years ~ province, data = afghan,
        main = "Education by Province", ylab = "Years of education")
dev.off()
```



2.4 实施抽样调查

抽样调查是研究人员从总体中选择一个子集的过程，该子集称为样本，可以此了解目标人群的特征，实际上是**通过调查一小部分人来了解大部分人**

2.4.1 样本的代表性

- **概率抽样**
  - 最基本的概率抽样过程称为**简单随机抽样（SRS）**
  - 从目标人群中不做替换地随机选择确定数量的个体，每个个体都具有相同的备选概率
  - 这时候的代表性指如果多次重复同样的步骤，抽样结果的特征不一定会和人口的特征完全相同，但是平均而言是相同的
  - 保证样本特征（**无论是观察到的还是未观察到的**）能与总体相应特征大致相同
  - 但是现实中获取一个代表总体的抽样框很难。若用电话号码、住宅地址和电子邮箱地址去划定总体样本框，依旧可能存在**样本选择偏误**的问题（有些人没有电话号码或有多个电话号码）

2.4 实施抽样调查

- **配额抽样**
  - 诸如年龄、性别、教育和种族等基本人口统计数据被用来构建不同类别的定额
  - 但是，即使用了人口可被观察到的特征选择了定额的代表，也可能有着**观察不到的特征**使样本与总体存在差异
  - 类似于个人在观察性研究中自我选择接受干预一样，研究人员可能会无意中选择那些**与未接受调查**的人有本质性不同的人
- **多级整群抽样**
  - 首先对较大单位进行抽样，再在每个被选取的较大单位中随机选择较小单位，以此类推
  - 在阿富汗调查的案例中，先在5个受关注的省份的每一个省份中抽样了各地区，然后在每个选定地区的村庄进行调查；在每个抽样的村庄中，根据村庄的位置以大致随机的方式选择一个家庭；最后对16岁及16岁以上的男性受访者进行调查

2.4 实施抽样调查

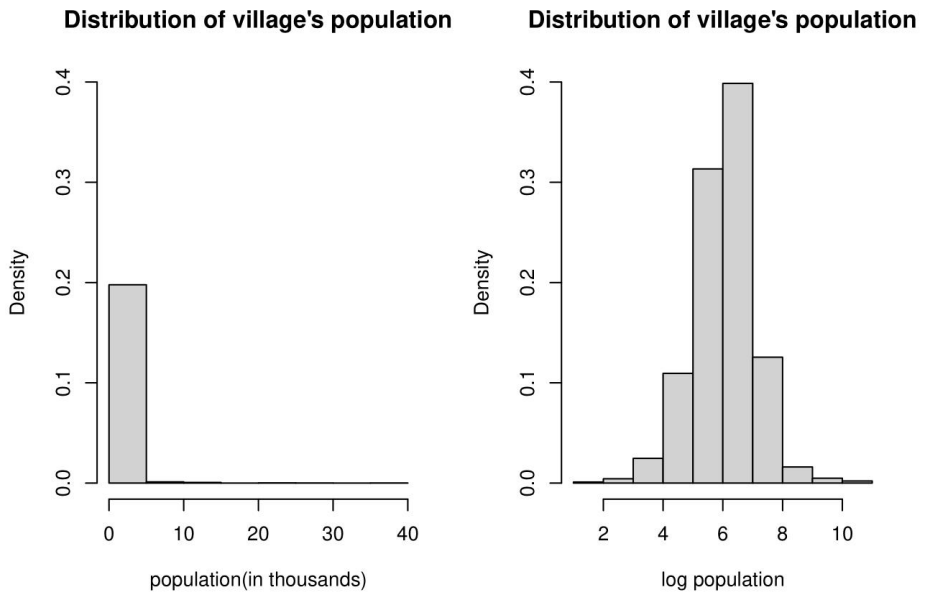
- **检验阿富汗的数据中随机抽样村庄代表性(afghan-village.csv)**
  - **自然对数转换**
    - 可用来纠正收入和人口等变量的偏度
    - 使结果分布看起来不那么极端
    - 若不进行对数转换，人口分布极不平衡，因为存在大量小村庄和少数大村庄
    - **使用log()函数**，默认情况下使用e作为基，也可以指定不同的基

表2 阿富汗村庄数据

变量名	描述
Village.surveyed	该村庄是否有被问卷调查
altitude	村庄的地理位置高度
educ.year	村庄的人数

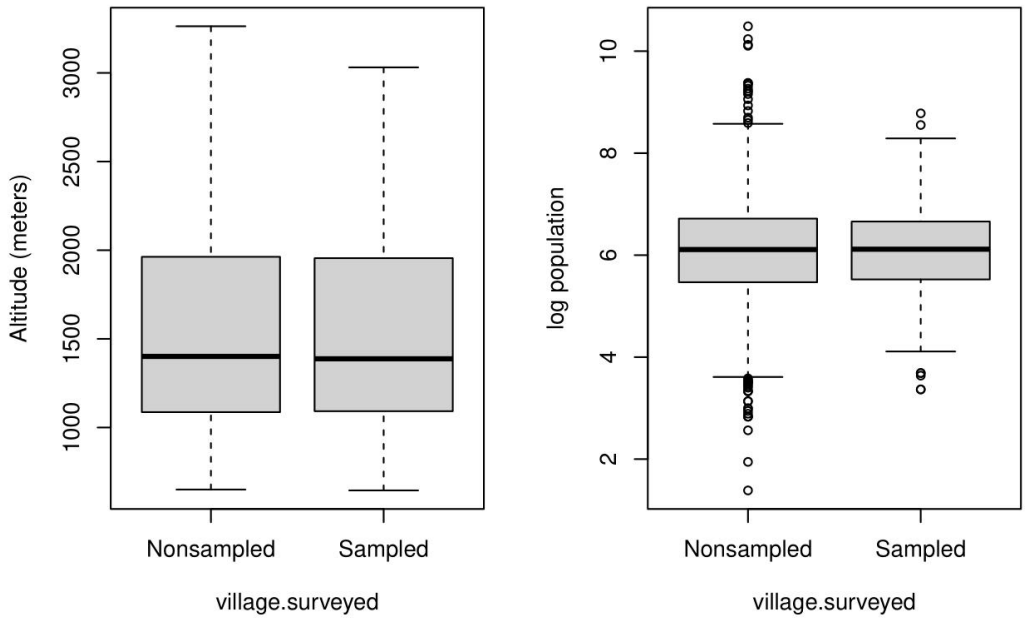
2.4 实施抽样调查

```
# 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")
```



- 使用箱型图比较抽样和非抽样村庄中各变量的分布情况

```
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"))
```



2.4.2 各类偏误来源

```
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
```

- **系统性偏误**
  - 大多数情况下代表总体的抽样框很难获得
- **拒访**
  - 即使有代表性的抽样框可供选择，对随机选择的个人进行访谈也可能很难
  - **对象拒访**：潜在的受访者拒绝参与调查（阿富汗调查的案例中，拒绝率为11%）
  - **题目拒访**：受访者参加了调查却拒绝回答了其中某个问题（收入变量缺失率为5%）
  - 当**回答了问题的人和不回答问题的人有根本性差异**时，两种拒访都会造成有偏误的推断
  - 例如，塔利班和ISAF关于平民受害问题的拒访率在各省之间有所不同
- **错误报告**
  - 受访者可能不希望采访者知道他们真实想法而撒谎（比如受到社会期许偏误的影响）

2.4 实施抽样调查

2.4.3 衡量敏感问题

类似于腐败、非法行为、种族偏见和性行为的敏感问题，极易受到**社会期许偏误**影响，若访谈是公开进行，一些涉及个人的伦理道德事项和潜在风险的调查会受到**个人或组织的抵制**

- **项目计数技术/列表实验**
  - 首先将样本随机分为两个可比较的组，让控制组和实验组在数值上的**差异由实验干预**所造成
  - 受访者报告的平均对象数量差异就是支持ISAF比例的估计值
  - 缺点：实验组中选择最大值或最小值可以显示一个人的真实答案（**地板效应**and**天花板效应**）

控制组：不包含敏感对象

Karzai Government; National Solidarity Program; Local Farmers

实验组：增加一个敏感对象

Karzai Government; National Solidarity Program; Local Farmers; Foreign Forces

两组人都只回答支持列表中的几个人，不回答具体支持谁

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

## [1] 0.04901961
```

该结果说明约有5%的阿富汗公民支持ISAF

2.4 实施抽样调查

- **随机回答技术RRT**
  - 被调查者以一个预定的基础概率P从两个或两个以上的问题中选择一个问题进行回答
  - 研究者并不知道被调查者回答的具体是哪个问题，但是可以根据概率论知识计算出来
  - 缺点：只能获得与总体水平有关的结论，无法进行单位水平参数的研究，即无法对一些产生敏感特性的原因进行剖析
  - 常用两个模型：沃纳模型，西蒙斯模型
  - 以西蒙斯模型为例：一般建议两个问题，一个是敏感性问题，一个是无关紧要的问题

事件	符号	概率
回答问题1	A	p
回答问题2	$\bar{A}$	$1-p$
回答结果为“是”	B	λ

假设受到塔利班伤害的人的比例为π
根据全概率公式可得:

$$P(B)=P(A\cap B)+P(\bar{A}\cap B)$$
$$=P(A)P(B|A)+P(\bar{A})P(B|\bar{A})$$

$$\lambda=p\pi+(1-p)\pi_2$$

假设调查了n人，有m人回答“是”
使用m / n近似估计λ，则

$$\hat{\pi}=\frac{1}{2p-1}(p-1+\frac{m}{n}),\ p\neq\frac{1}{2}$$

華中師範大學
CENTRAL CHINA NORMAL UNIVERSITY

03 潜变量度量

3.1 案例概况

度量问题是人类行为研究中理论和实证分析的交集，对于不可观测的概念度量往往需要结合理论模型

案例：政治极化度量

- 意识形态（自由主义or保守主义）可以有效地描述人的政治取向
- 美国国会每年都要对数百个法案进行投票，利用这种公开可用的投票记录，可以推测出他们的意识形态
- 使用空间度量模型将立法者的意识形态与他们的投票联系起来

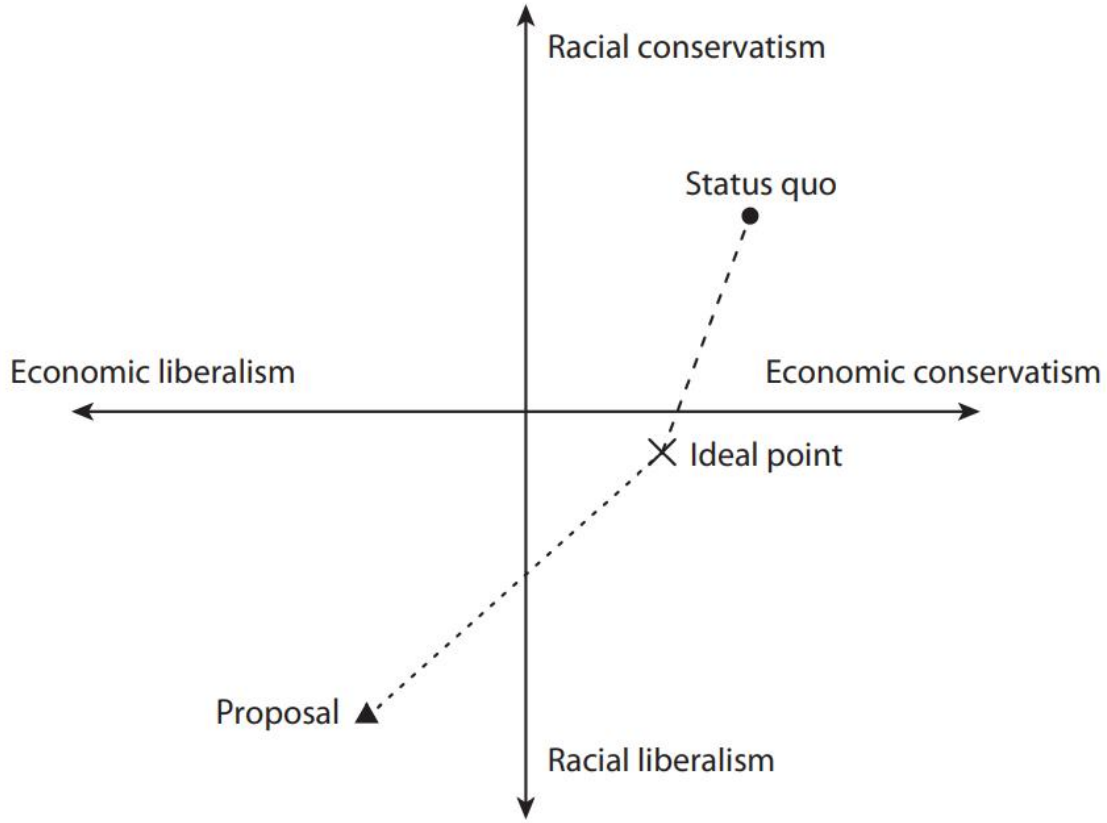


图 立法者意识形态投票的空间模型

3.1 案例概况

- 数据集”congress.csv”
- 将立法者理想点的估计s值，称为DW-NOMINATE分数，正（负）的分数越大（小）表示越偏向自由（保守）主义
- 数据包含所有在第80届（1947年-1948年）至第112届（2011年-2012年）众议院议员们的理想点

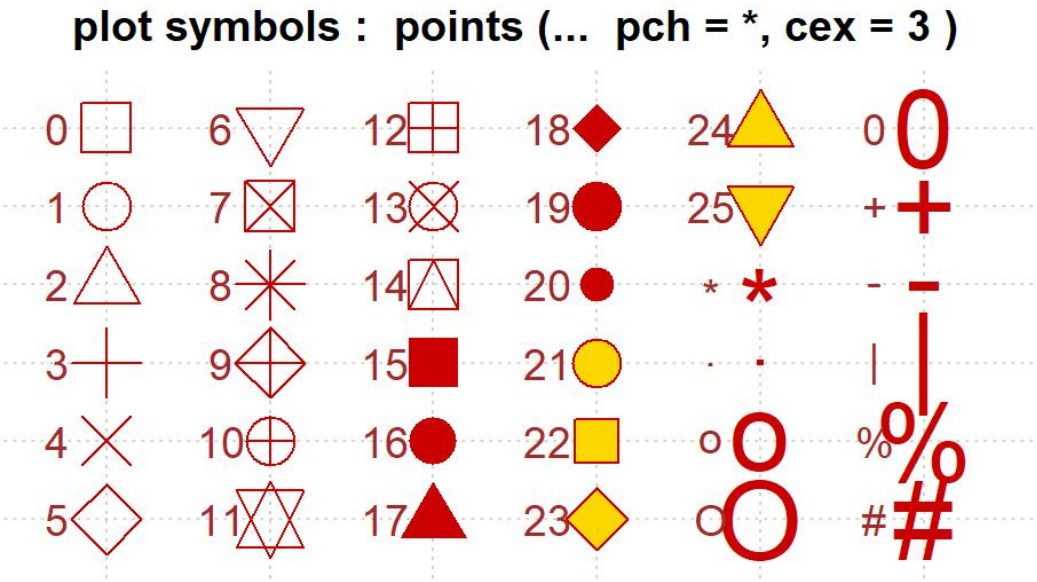
表3 立法理想点数据

变量名	描述
name	议员的名字
state	议员所在的州
distinc	议员所在的区
party	议员的党派
Congress	国会会议代码
dwnom1	DW Nominate分数（第一维度）
dwnom2	DW Nominate分数（第二维度）

3.2 概括双变量关系

3.2.1 散点图

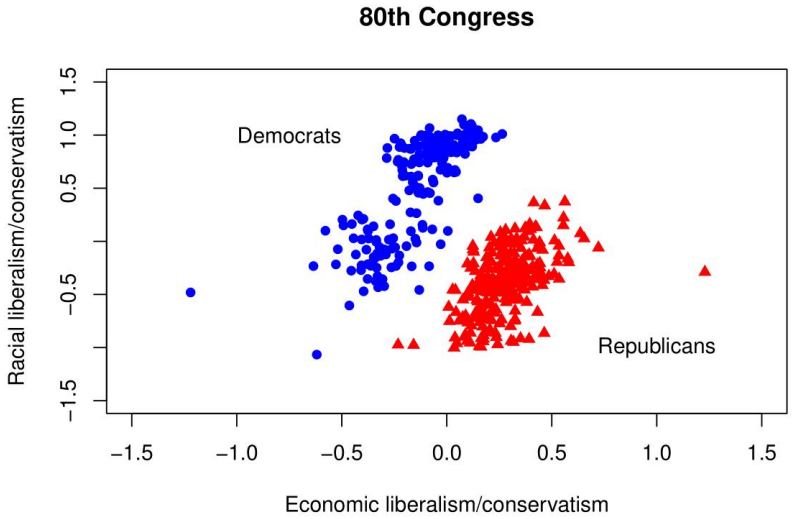
- plot()函数和point()函数
  - plot(x, y)和point(x, y)中的x和y分别是横坐标和纵坐标的向量
  - pch参数可以指定不同的绘图符号
  - col参数可以指定要使用的颜色，比如”blue”
  - lty参数指定要绘制的线条类型，使用字符或数值，包括”solid”或1（默认）为实线，”dashed”或2为线条状虚线……
  - lwd参数指定行的粗细，默认值为1

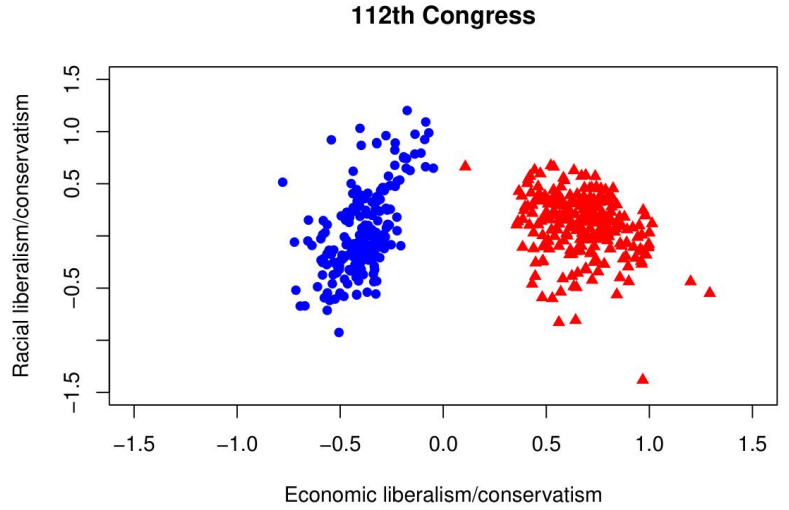


3.2 概括双变量关系

```
## 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")
```



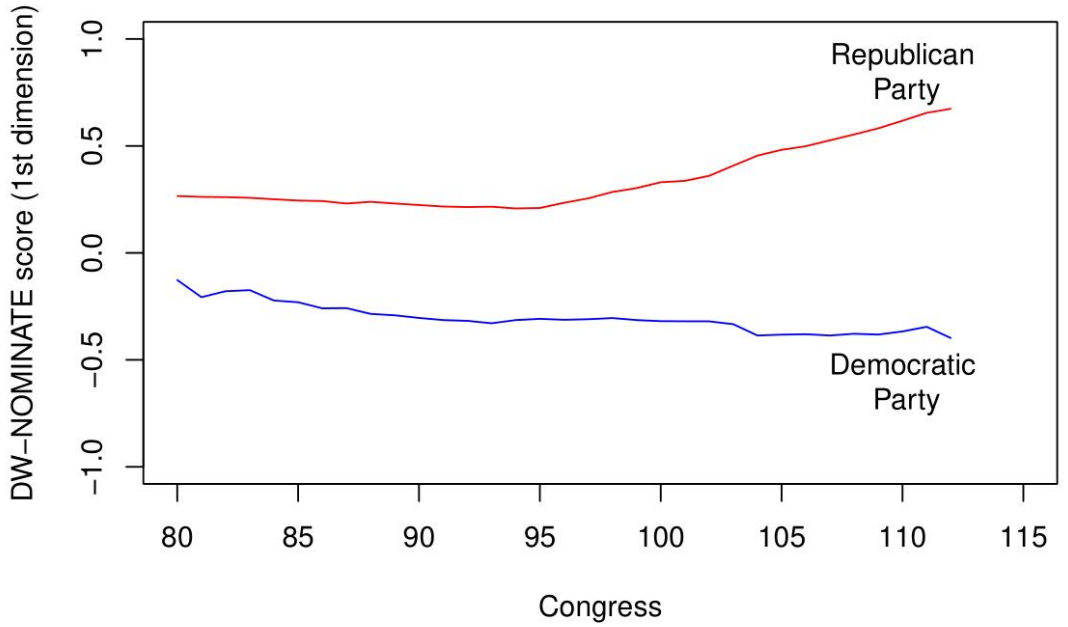


结论：在112届国会中，两党在种族维度上的意识形态差异已经不大，经济维度可能成为党派差异主要原因

3.2 概括双变量关系

- **党派的中位理想点可以代表经济自由主义/保守主义维度中每个党派的中心，观察该中心如何随着时间变化**
- plot()函数中**type参数**设置为“l”，可以将点连接成实线
  - "p" 绘散点图
  - "b" 所有点被实线连接
  - "o" 实线通过的所有点
  - "h" 绘出点到x轴的竖线
  - "s" 绘出阶梯形曲线
- text()函数中使用”**\n**”可以更改为新行
- **结论：**双方的意识形态中心随着时间的推移而产生分歧，民主党变得更加自由主义，共和党变得更加保守主义，很多学者将这种现象称为**政治两极分化**

```
## 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")
```

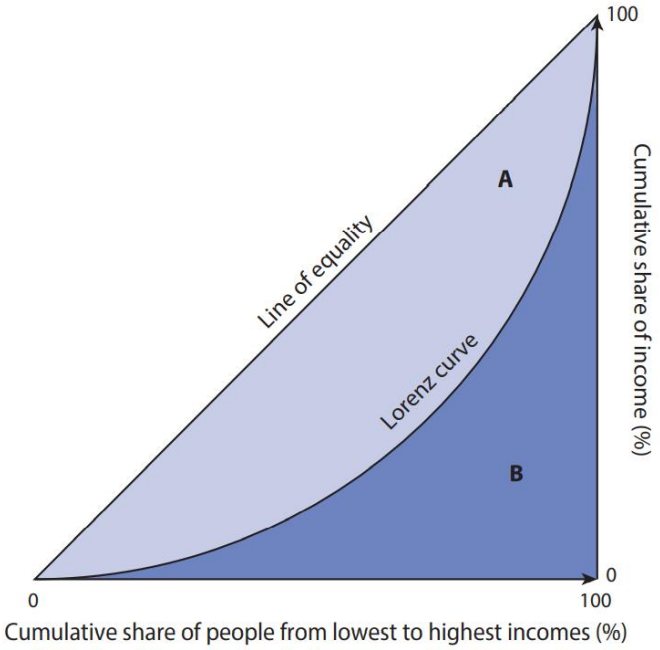


3.2 概括双变量关系

3.2.2 相关性

- 猜想：党派差异的扩大可能是因为收入不平等的加剧
- **基尼系数**
  - 衡量某个社会收入平等和不平等的程度，从0（每个人拥有相同数量的财富）到1（一个人拥有所有的财富）
  - 基尼系数可以被定义为**平等线与洛伦兹曲线之间的区域除以平等线之下的区域**
  - 基尼系数 $= \frac{\text{平等线和洛伦兹曲线之间的面积}}{\text{平等线下面的面积}}$

    $= \frac{\text{区域}A\text{的面积}}{\text{区域}A\text{的面积}+\text{区域}B\text{的面积}}$
  - A的面积越大，基尼系数越大，不平等程度越高



- **洛伦兹曲线**
  - 横轴代表最低收入到最高收入的人群的累计份额
  - 纵轴代表收入等于或小于某个特定收入百分比的人的收入累计份额
  - 连接了这两个统计数据的线被称为**洛伦兹曲线**
  - 若所有人的收入完全相同，那么洛伦兹曲线将与45度线相同。即x%的人口都将刚好占国民收入的x%，这样的线称为**平等线**

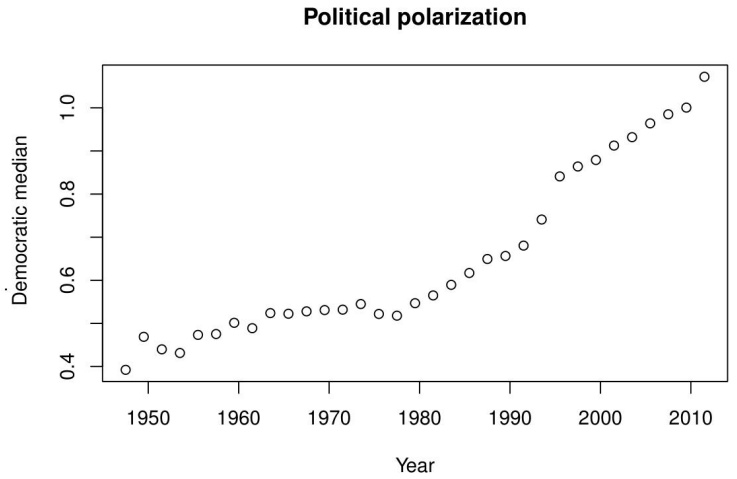
3.2 概括双变量关系

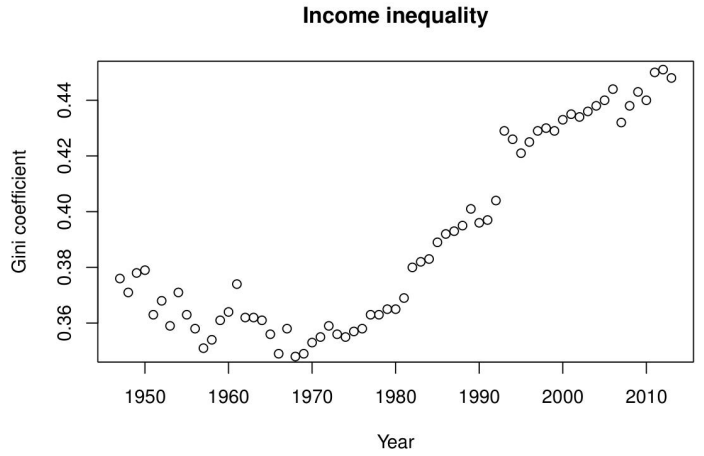
- 数据集”USGini.csv”
- 数据集包含1947年至2013年的基尼系数
- 检验政治两极化与收入不平等之间的关系，并排创建两个时间序列图
- 第一张图显示随时间推移的党派差距，第二张显示同一时期的基尼系数
- 可以观察到，**美国政治两极分化和收入不平等现象一直在稳步增加**

```
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")
```

表4 美国基尼系数数据

变量名	描述
year	年份
gini	美国基尼系数





3.2 概括双变量关系

- **z分数**
  - 表示观测值高于或低于平均值的标准差的数量
  - 变量x的第i次观察的z分数等于 $\frac{x_i - \bar{x}}{S_x}$
  - z分数将变量标准化，因此其度量单位并不重要，即ax_i+b的z分数与x_i的z分数相同

$$z\text{-score of } (ax_i + b) = \frac{(ax_i + b) - \text{mean of } (ax + b)}{\text{standard deviation of } (ax + b)}$$

$$= \frac{a \times (x_i - \text{mean of } x)}{a \times \text{standard deviation of } x}$$

$$= z\text{-score of } x_i,$$

- **相关系数**
  - 平均而言，两个变量如何相对各自的平均值一起移动
  - 两变量x和y的相关性可定义为两个变量的z分数的平均乘积：

$$\text{correlation}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(z\text{-score of } x_i \times z\text{-score of } y_i).$$

  - 相关性的分母实际上是n-1而不是n，但是只要样本量足够大，这种差异就不会影响到其结论
  - 相关性基于z分数，因此即使使用不用的**单位**进行度量，相关性也不会变化
  - 可以使用cor()**函数**计算相关性

3.2 概括双变量关系

- **计算基尼系数和政治两极分化之间的相关性**
  - 基尼系数每年都有一个值，而每届美国国会为期两年
  - 选择在每届国会的第二年提取基尼系数
  - **结果：**
    - 相关性为正，且相当高，说明政治上的两极化和收入不平等向相同的方向发展
    - 注意：**相关性不一定意味着因果性**

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

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

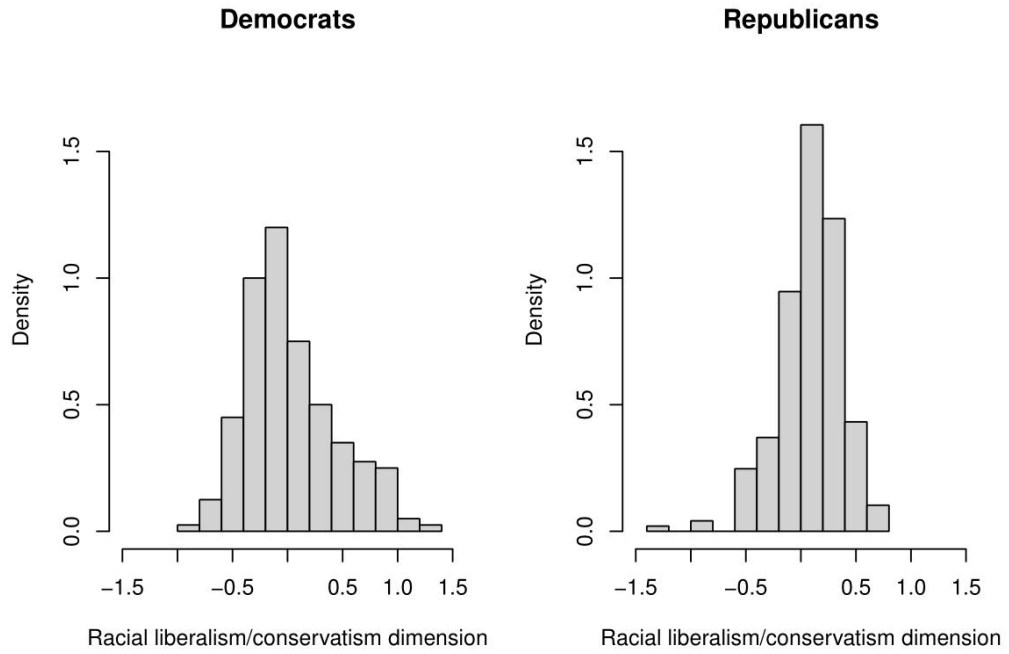
3.2 概括双变量关系

3.2.3 分位数-分位数 / Q-Q 图

- 比较两个变量的整体分布（不使用平均数or中位数）
- 比较两个分布可以使用直方图
- **分位数-分位数 / Q-Q 图**
  - **分位数的散点图**，描绘一个变量的每个分位数的数值与另一个变量的相应分位数的数值之间的关系
  - 如果两个分布相同，那么所有分位数都具有相同的值，此时Q-Q图将产生**45度线**
  - 如果Q-Q图中的点形成比45度线更平坦的线，则表明横轴上的分布比纵轴上的分布更分散
  - 使用**qqplot()函数**指定参数x和y来生成Q-Q图

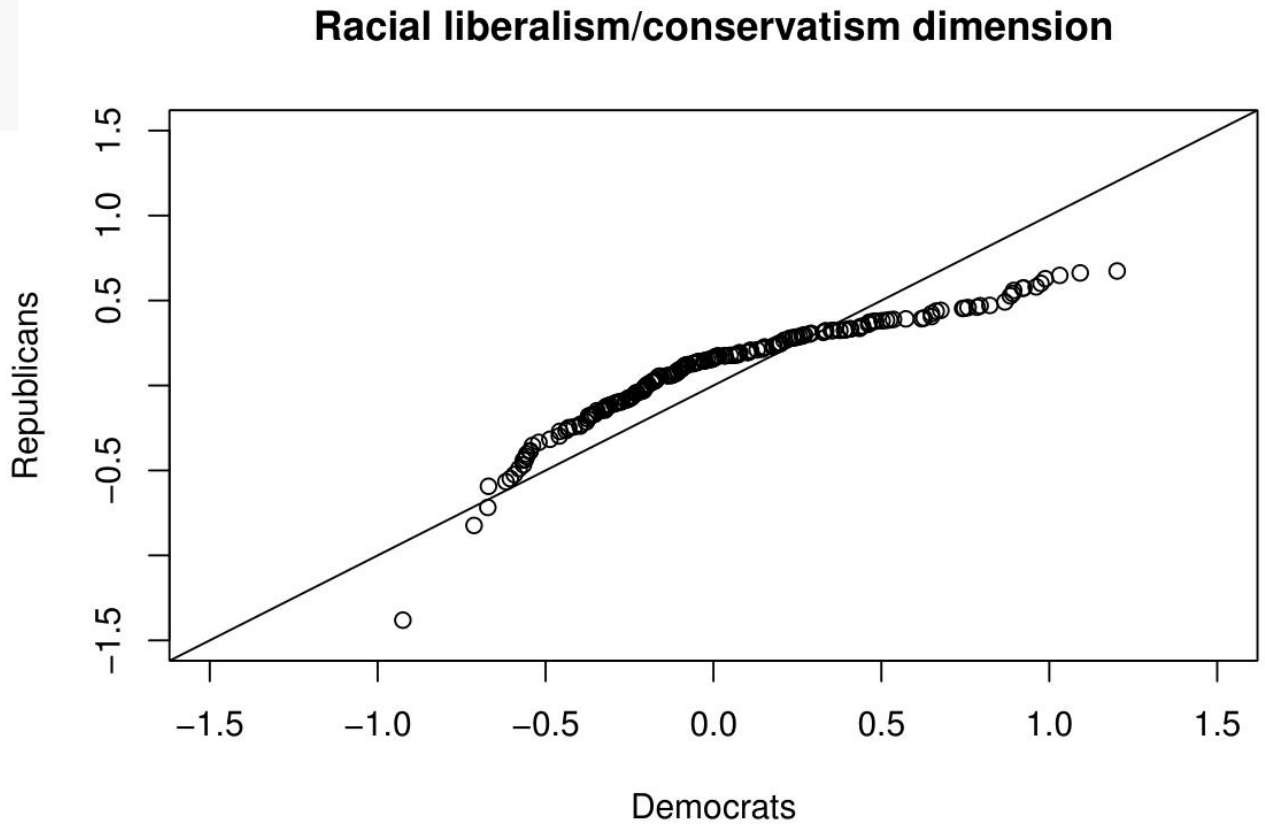
比较第112届国会中种族维度理想点的分布

```
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")
```



3.2 概括双变量关系

```
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)
```



结果与结论：

- 代表较低分位数的点出现在45度线以上：自由派共和党人比自由派民主党人更保守
- 代表上四分位数的点位于45度以下：保守的民主党人比保守的共和党人更保守
- 连接点的线比45度线更平坦：对于民主党人来说，意识形态的分布比共和党人更分散

3.3 聚类

第112届国会中民主党和共和党意识形态截然不同，但是每个**党派内**是否存在意识形态相似的立法者呢？

3.3.1 R中的矩阵

- **矩阵vs数据框**
  - 数据框使用不同类型的变量（如数字、因子、字符）
  - 矩阵原则上只使用数值（在某些情况下接受逻辑和其他特殊值）
- **matrix()函数**
  - 通过nrow和ncol参数指定矩阵大小
  - byrow=TRUE/FALSE指示按行/列填入数据
  - rownames() 和colnames() 可向行和列添加标签

```
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
```

3.3 聚类

- as.matrix()函数
  - 将数据框对象强制转换为矩阵
  - 但数据框对象的某些功能（如变量类型）将会丢失，会将不同类型的变量转换成单一类型的字符
- colSums()、colMeans（）、rowSum()、rowMean()
  - 可分别计算列（行）的综合和平均值
  - 可使用na.rm参数

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

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

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

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

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

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

```
colSums(x)
```

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

```
rowMeans(x)
```

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

3.3 聚类

- apply()函数
  - 第一个或X参数为数组or矩阵or数据框（**至少是二维**）
  - 第二个或MARGIN参数指定我们希望应用函数的维度（1表示行，2表示列）；若想**在所有元素上应用**函数，可写成 apply(x, 1：2, sum)
  - 第三个或FUN参数指定执行的函数
  - apply组函数也可以处理**具有多个参数**的函数，例如 apply(x1, 1, fn, x2 = b, x3 = c)。将x1作为data传入，将函数的x2和x3作为其他参数

表4 apply组函数总结

函数名称	使用对象	返回结果
apply()	矩阵、数组或数据框	向量、数组或列表
lappy()	列表、数据框或向量	列表
sapply()	列表、数据框或向量	向量、数组或列表
tapply()	不规则阵列	阵列
mapply()	多个列表或向量参数	列表

3.3.2 R中的列表

- **列表vs矩阵**
  - 数据框只接受具有相同长度的向量
  - 列表可以存储**不同类型**的对象作为其元素（比如采用**不同长度**的数字和字符向量）
  - 甚至可以包含多个不同大小数据框作为元素
- **列表提取元素的方式**
  - 使用”$”运算符进行提取
  - 使用双方括号”[[ ]]”进行提取，可用**整数**指示，也可用**元素名称**指示
- **names()和length()等函数也可以作用于列表**

```
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
```

3.3.3 K-means算法

- **步骤**
  - ① 选择k个聚类的初始质心
  - ② 给定质心，将每个观测值分配给最近的质心（使用欧几里得距离）
  - ③ 选择坐标等于相应变量的簇平均值的每个簇的新质心
  - ④ 重复步骤2、3，直到聚类分配不再改变
- **输入数据标准化**
  - 可将所有变量置于同一尺度，聚类结果不取决于每个变量的单位
  - 可以借助z分数完成转换
  - 使用scale()函数标准化一个或一组变量，采用单个变量的向量或多个变量的矩阵
- **函数kmeans()在R中实现了该算法**

3.3 聚类

将k-means算法分别应用于第80届和第112届国会的DW-NOMINATE得分

- cbind()函数可按列组合两个变量以创建矩阵（补充：rbind()）
- DW-NOMINATE得分本身已经按照实际含义进行过缩放，因此不再标准化
- kmeans()函数
  - centers：聚类数
  - iter.max：最大迭代次数，默认为10
  - nstart：随机选择的初始质心数，默认为1，相当于运行次数
  - 输出对象为一个**列表**
    - cluster，各个聚类的编号向量
    - centers，各个聚类的质心矩阵
    - totss，所有聚类变量的离差平方和之和，测度类内部数据点离散程度
    - betweenss，各类别间的聚类变量离差平方和之和
    - size，每个聚类中数据点的数量
    - iter，直到收敛时的迭代次数

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

3.3 聚类

聚成两个类

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

```
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
```

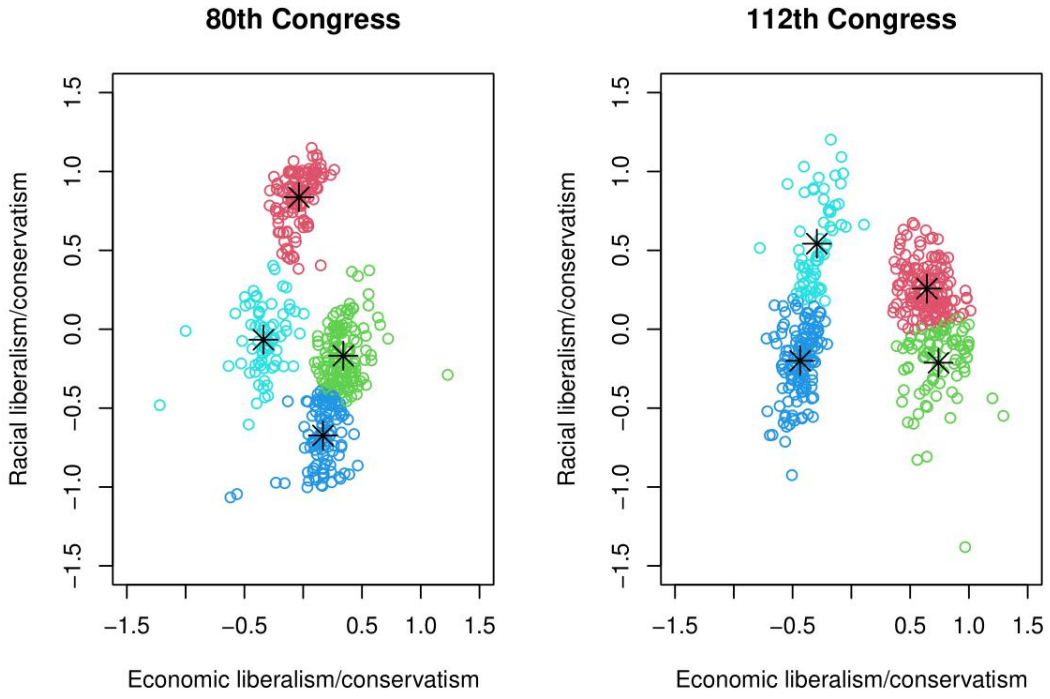
结果与结论：

- 对于112届国会而言，一个聚类只包含共和党人，另一个聚类也几乎只包含民主党人，说明该聚类结果符合党派属性；而对于80届国会而言，有一个聚类同时包含大量共和党和民主党人
- 政治两极分化随着时间的推移而恶化

3.3 聚类

聚成四个类

```
## 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)
```



- 为col参数指定一个整数值的向量，而不是具体的颜色名称，以便每个整数值用于相应的聚类
- 加1是为了避免和质心的绘制颜色（黑色）冲突

结果和结论：

- 将民主党分为2个聚类，将共和党分为2个聚类
- 两党中的党内分歧都体现在种族层面；两党之间差异体现在经济层面

華中師範大學
CENTRAL CHINA NORMAL UNIVERSITY

04 课后习题

4 课后习题

- 改变对待同性恋婚姻的看法
- 中国和墨西哥的政治效力
- 联合国大会投票表决

華中師範大學
CENTRAL CHINA NORMAL UNIVERSITY

谢谢大家！

张梦毅　2021年12月15日