Exercise1

my

2021/12/5

目录

1 研究早教的班级规模对教育绩效和个人发展的影响

1.1 在数据框中创建一个名为 kinder 的新因子变量；将种族变量重新编码成四个等级；将 na.rm = TRUE 添加到函数中，作为丢弃缺失数据

```
setwd("D:/QSS/Chapter2_Causality/Exercise")
STAR <- read.csv("STAR.csv")
```

```
View(STAR)
summary(STAR)
```

```
##       race          classtype       yearssmall          hsgrad
##  Min.   :1.000   Min.   :1.000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:1.000
##  Median :1.000   Median :2.000   Median :0.0000   Median :1.000
##  Mean   :1.341   Mean   :2.052   Mean   :0.9542   Mean   :0.833
##  3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:2.0000   3rd Qu.:1.000
##  Max.   :6.000   Max.   :3.000   Max.   :4.0000   Max.   :1.000
##  NA's   :3                                        NA's   :3278
##      g4math        g4reading
##  Min.   :487.0   Min.   :528.0
##  1st Qu.:688.0   1st Qu.:696.0
##  Median :710.0   Median :723.0
##  Mean   :708.8   Mean   :721.2
##  3rd Qu.:732.5   3rd Qu.:750.0
##  Max.   :821.0   Max.   :836.0
##  NA's   :3930    NA's   :3972
```

```
STAR$kinder <- NA # 创建一个因子变量，初始值均为 NA 缺失值
STAR$kinder[STAR$classtype == "1"] <- " 小班" # 通过是否满足特征来指定不同的类别
STAR$kinder[STAR$classtype == "2"] <- " 普通班"
STAR$kinder[STAR$classtype == "3"] <- " 辅导班"
STAR$kinder <- as.factor(STAR$kinder) # 将字符向量转化成因子变量
levels(STAR$kinder) # 查看各个类别级别
```

```
## [1] "辅导班" "普通班" "小班"
```

```
table(STAR$kinder) # 查看各个级别观察值的数量
```

```
##
## 辅导班 普通班   小班
##   2231   2194   1900
```

```
# View(STAR)

# 数据框中原本有种族变量。y 要求只覆盖而不新建因子变量
STAR$race[STAR$race == "1"] <- " 白人" # 通过是否满足特征来指定不同的类别
STAR$race[STAR$race == "2"] <- " 黑人"
STAR$race[STAR$race == "4"] <- " 西班牙裔"
```

```
STAR$race[STAR$race == "3" | STAR$race == "5" | STAR$race == "6" ] <- " 其他"
STAR$race <- as.factor(STAR$race) # 将字符向量转化成因子变量
levels(STAR$race) # 查看各个类别级别
```

```
## [1] "白人"     "黑人"     "其他"     "西班牙裔"
```

```
table(STAR$race) # 查看各个级别观察值的数量
```

```
##
##     白人     黑人     其他 西班牙裔
##     4234     2058       25        5
```

```
# View(STAR)
```

1.2 小班的阅读和数学成绩与普通班相比如何？使用平均数来进行比较，同时删除缺失值；请比较它们的测试分数的标准差来了解估计效果的大小

```
my.mean <- function(x){ # 定义一个可以删除缺失值的求平均数的函数
  out <- mean(x, na.rm = TRUE)
  return(out)
}
tapply(STAR$g4math, STAR$kinder, my.mean) # 比较数学成绩
```

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

```
tapply(STAR$g4reading, STAR$kinder, my.mean) # 比较阅读成绩
```

```
##   辅导班   普通班     小班
## 720.7155 719.8900 723.3912
```

```
my.sd <- function(x){ # 定义一个可以删除缺失值的求标准差的函数
  out <- sd(x, na.rm = TRUE)
  return(out)
}
tapply(STAR$g4math, STAR$kinder, my.sd)
```

```
##   辅导班   普通班     小班
## 44.74373 41.02063 43.57318
```

```
tapply(STAR$g4reading, STAR$kinder, my.sd)
```

```
##    辅导班    普通班      小班
## 52.44263 53.16788 51.54494
```

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

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

1.2.1 结论

- 小班和普通班的数学成绩基本相差不大，小班的阅读成绩比普通班大很多
- 小班数学成绩的标准差大于普通班，小班数学成绩较分散
- 小班阅读成绩的标准差小于普通班，小班阅读成绩较集中

1.3 将小班的高分（定义为第 66 百分位）和低分（第 33 百分位）与普通班的相应分数进行比较

```
my.quantile <- function(x){ # 定义一个可以删除缺失值的求第 33 和 66 百分位的函数
  out <- quantile(x, probs = seq(from = 0, to = 1, by = 1/3), na.rm = TRUE)
  return(out)
}
tapply(STAR$g4math, STAR$kinder, my.quantile)
```

```
## $辅导班
##        0% 33.33333% 66.66667%      100%
##       487       696       725       821
##
## $普通班
##        0% 33.33333% 66.66667%      100%
##       487       696       725       821
##
## $小班
##        0% 33.33333% 66.66667%      100%
##       487       695       726       821
```

```
tapply(STAR$g4reading, STAR$kinder, my.quantile)
```

```
## $辅导班
```

```
##         0% 33.33333% 66.66667%       100%
##        528        705        738        836
##
## $普通班
##         0% 33.33333% 66.66667%       100%
##        528        705        740        836
##
## $小班
##         0% 33.33333% 66.66667%       100%
##        528        705        741        836
```

1.3.1 结论

- 小班和普通班的数学成绩的高低分差不多
- 小班和普通班的阅读成绩的高低分差不多

1.4 有些学生在 STAR 课程的四年中都在小班中上课，其他人被分到小班一年，之后去了其他班级。那么数据集中每种类型的学生数量为多少？使用 kinder 和 yearssmall 创建比例列表。参加更多年的小班对考试成绩有更大的影响吗？比较那些在小班不同年数的学生的阅读和数学考试分数的平均数和中位数

```
sSTAR <- subset(STAR, subset = (kinder == " 小班")) # 对数据集进行分集，得到只包含小班的数据集
prop.table(table(sSTAR$yearssmall)) # 创建比例列表
```

```
##
##         1         2         3         4
## 0.3031579 0.1431579 0.1026316 0.4510526
```

```
tapply(sSTAR$g4math, sSTAR$yearssmall, my.mean) # 比较在小班不同年数的学生的数学成绩的平均数
```

```
##        1        2        3        4
## 703.0000 699.1429 704.5000 710.0519
```

```
tapply(sSTAR$g4reading, sSTAR$yearssmall, my.mean) # 比较在小班不同年数的学生的阅读成绩的平均数
```

```
##        1        2        3        4
## 724.6667 700.5714 709.1481 724.6651
```

```
my.median <- function(x){ # 定义一个可以删除缺失值的求中位数的函数
  out <- median(x, na.rm = TRUE)
```

```
  return(out)
}
tapply(sSTAR$g4math, sSTAR$yearssmall, my.median) # 比较在小班不同年数的学生的数学成绩的中位数
```

```
##   1   2   3   4
## 706 712 707 711
```

```
tapply(sSTAR$g4reading, sSTAR$yearssmall, my.median) # 比较在小班不同年数的学生的阅读成绩的中位数
```

```
##     1     2     3     4
## 724.5 711.0 707.0 726.0
```

1.5 STAR 计划是否缩小了不同种族群体之间的成绩差距？找出没有接受额外辅导的、被分配到普通班的学生中白人和少数族裔（黑人或西班牙裔）学生的平均阅读和数学成绩，与被分配到小班的学生进行比较

```
rSTAR <- subset(STAR, subset = (kinder == " 普通班")) # 对数据集进行分集，得到只包含普通班的数据集
# 前面已经分出了只包含小班的数据集

# 普通班白人和少数族裔的平均阅读分数差异
DifReadRegular <- my.mean(rSTAR$g4reading[rSTAR$race == " 白人"]) - my.mean(rSTAR$g4reading[rSTAR$r
# 普通班白人和少数族裔的平均数学分数差异
DifMathRegular <- my.mean(rSTAR$g4math[rSTAR$race == " 白人"]) - my.mean(rSTAR$g4math[rSTAR$race !=
DifReadRegular
```

```
## [1] 34.13144
```

```
DifMathRegular
```

```
## [1] 12.42599
```

```
# 小班白人和少数族裔的平均阅读分数差异
DifReadSmall <- my.mean(sSTAR$g4reading[sSTAR$race == " 白人"]) - my.mean(sSTAR$g4reading[sSTAR$rac
# 小班白人和少数族裔的平均数学分数差异
DifMathSmall <- my.mean(sSTAR$g4math[sSTAR$race == " 白人"]) - my.mean(sSTAR$g4math[sSTAR$race != '
DifReadSmall
```

```
## [1] 27.36424
```

```
DifMathSmall
```

```
## [1] 12.46733
```

1.6 幼儿园班级规模对人的长期影响。比较分配给不同班级类型的学生的高中毕业率；根据小班的学习年数，检查毕业率是否有所不同。调查 STAR 计划是都减少了白人和少数族裔学生毕业率之间的种族差距

```
hsgradrate <- function(x){ # 定义一个求毕业率的函数
  newx <- na.omit(x)
  out <- sum(newx) / length(newx)
  return(out)
}
tapply(STAR$hsgrad, STAR$kinder, hsgradrate) # 比较不同班级类型的学生的高中毕业率
```

```
##     辅导班     普通班       小班
## 0.8392857 0.8251619 0.8359202
```

```
tapply(sSTAR$hsgrad, sSTAR$yearssmall, hsgradrate) # 比较小班不同年数的学生的高中毕业率
```

```
##         1         2         3         4
## 0.7852761 0.7589286 0.7727273 0.8775510
```

```
# 普通班白人和少数族裔的高中毕业率差异
DifGraRegular <- hsgradrate(rSTAR$hsgrad[rSTAR$race == " 白人"]) - hsgradrate(rSTAR$hsgrad[rSTAR$ra
DifGraRegular
```

```
## [1] 0.1181304
```

```
# 小班白人和少数族裔的高中毕业率差异
DifGraSmall <- hsgradrate(sSTAR$hsgrad[sSTAR$race == " 白人"]) - hsgradrate(sSTAR$hsgrad[sSTAR$race
DifGraSmall
```

```
## [1] 0.1195707
```