cda 第二次作业

林振炜

2019 年 10 月 19 日

0.1 1 analysis the GDS5037 data, Suppose the samples are randomly chosen

- in sa group, the proportion of male and female, respectively

```
library(dplyr)
library(cdabookfunc)
library(cdabookdb)
data('aspirin')
genderdata <- read.csv('GDS5037/genderdata.csv')
s <- genderdata[genderdata$ID == 'SA',] %>%group_by(gender)%>%summarise(sa=n());s
```

```
## # A tibble: 2 x 2
##   gender    sa
##   <fct>  <int>
## 1 female    28
## 2 male      10
```

```
pfemale = 28/38;pfemale
```

```
## [1] 0.7368421
```

```
pmale = 10/38;pmale
```

```
## [1] 0.2631579
```

- construct table for gender and status, calculate the sample odds ratio.

```
control <- genderdata[genderdata$ID == 'control',] %>%group_by(gender)%>%summarise(contro
```

```
## # A tibble: 2 x 2
##   gender control
##   <fct>    <int>
## 1 female      11
## 2 male         9
```

```
gen_sta <- merge(s,control,by = 'gender');gen_sta
```

```
##   gender sa control
## 1 female 28      11
## 2   male 10       9
```

```
rownames(gen_sta) <- gen_sta[,1]
gen_sta <- gen_sta[,-1]
gen_sta <- as.matrix(gen_sta);
addmargins(gen_sta)
```

```
##        sa control Sum
## female 28      11  39
## male   10       9  19
## Sum    38      20  58
```

```
oddsratio(gen_sta)
```

```
##   oddsratio
## 1  2.290909
```

- in (2) test the independence between gender and status

```
# x2 test
x2_result <- chisq.test(gen_sta);x2_result
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  gen_sta
## X-squared = 1.3151, df = 1, p-value = 0.2515
```

```
#G2_test
gen_sta_excepted <- x2_result$expected
gen_sta_excepted
```

```
##               sa   control
## female 25.55172 13.448276
## male   12.44828  6.551724
```

```
gsq <- 2*sum(gen_sta*log(gen_sta/gen_sta_excepted))
pvalue <-  1 - pchisq(gsq,2)
gsq;pvalue
```

```
## [1] 2.037953
```

```
## [1] 0.3609641
```

- test the independence between gender and status

```
control <- genderdata[genderdata$ID == 'control',] %>%group_by(gender)%>%summarise(control
gen_sta <- merge(s,control,by = 'gender');
MMA <- genderdata[genderdata$ID == 'MMA',] %>%group_by(gender)%>%summarise(MMA=n());
gender_sta <- merge(gen_sta,MMA,by = 'gender');gender_sta;
```

```
##   gender sa control MMA
## 1 female 28      11  35
## 2   male 10       9  15
```

```
rownames(gender_sta) <- gender_sta[,1]
gender_sta <- gender_sta[,-1]
gender_sta <- as.matrix(gender_sta);
addmargins(gender_sta)#the count of every cell is more than 5
```

```
##        sa control MMA Sum
## female 28      11  35  74
## male   10       9  15  34
## Sum    38      20  50 108
```

```r
G2 <- function(data){
  x2_result <- chisq.test(data)
  expected <- x2_result$expected
  Gsq <- 2 * sum(data * log(data / expected))
  pvalue <- 1 - pchisq(Gsq, 2)
  return(pvalue)
}

#test detached group
oddsratio(gender_sta[,c(1,2)])
```

```
##   oddsratio
## 1  2.290909
```

```r
p12 <- G2(gender_sta[,c(1,2)]);p12
```

```
## [1] 0.3609641
```

```r
oddsratio(gender_sta[,c(3,2)])
```

```
##   oddsratio
## 1  1.909091
```

```r
p32 <- G2(gender_sta[,c(3,2)]);p32
```

```
## [1] 0.4976395
```

```r
oddsratio(gender_sta[,c(1,3)])
```

```
##   oddsratio
## 1       1.2
```

```r
p13 <- G2(gender_sta[,c(1,3)]);p13
```

```
## [1] 0.9302105
```

```
trans12 <- matrix(data=c(1,1,0,0,0,1), nrow = 3, ncol = 2, byrow = F, dimnames = NULL)
trans13 <- matrix(data=c(1,0,1,0,1,0), nrow = 3, ncol = 2, byrow = F, dimnames = NULL)
trans23 <- matrix(data=c(0,1,1,1,0,0), nrow = 3, ncol = 2, byrow = F, dimnames = NULL)
p_combine <- function(trans){
  df <- gender_sta%*%trans
  oddsratio(df)
  p <- G2(df);
return(p)}

p12_3 <- p_combine(trans12);p12_3
```

```
## [1] 0.9536745
```

```
p13_2 <- p_combine(trans13);p13_2
```

```
## [1] 0.3700692
```

```
p23_1 <- p_combine(trans23);p23_1
```

```
## [1] 0.6917503
```

- p-value 在格子拆分和合并的情况下，其 p 值都大于 0.05，所以在 95% 的置信水平下，可以认为 gender and status are independent.

0.2 2.6

- (1) 在超过 35 岁的女性人群中，有 0.001304 吸烟的同时死肺癌，有 0.000121 不吸烟的同时死于肺癌，造成这接近 10 倍的差异的原因应该是 “是否吸烟”

```
smo_lung<- matrix(data=c(0.001304,1-0.001304,0.000121,1-0.000121), nrow = 2, ncol = 2, byr

oddsratio(smo_lung)
```

```
##   oddsratio
## 1  10.78963
```

```
p <- G2(smo_lung);p
```

```
## [1] 0.9994262
```

- (2) the estimated odds of smoking equal 1.79 times the estimated odds of nonsmoking. the reason is the value is too small compared with 1.

0.3 2.9

- 对于 $(0 < s < 20)$ 有 $\theta_1 = \frac{\pi_1}{\pi_0} \times \frac{1-\pi_0}{1-\pi_1} = 11.7$
- 对于 $(s >= 20)$ 有 $\theta_2 = \frac{\pi_2}{\pi_0} \times \frac{1-\pi_0}{1-\pi_2} = 26.1$
- 因此对于 $(0 < s < 20)$ 和 $(s >= 20)$ 有 $\theta_3 = \frac{\pi_2}{\pi_1} \times \frac{1-\pi_1}{1-\pi_2} = \frac{\pi_2}{\pi_0} \times \frac{1-\pi_0}{1-\pi_2} / \frac{\pi_1}{\pi_0} \times \frac{1-\pi_0}{1-\pi_1} = \theta_2/\theta_1 = 26.1/11.7 = 2.2$

0.4 2.12

```
column <- c('yes','no')
row <- c('aspirin','placebo')
aspirin_heart<- matrix(data=c(198,193,19934-198,19942-193), nrow = 2, ncol = 2, byrow = F,
```

```
##          yes    no
## aspirin 198 19736
## placebo 193 19749
```

```
oddsratio(aspirin_heart)
```

```
##   oddsratio
## 1  1.026582
```

```
lodd <- log(oddsratio(aspirin_heart))
se <- sqrt(1/198+1/19736+1/193+1/19749)
ci <- c(exp(lodd+1.96*se),exp(lodd-1.96*se));ci
```

```
## $oddsratio
## [1] 1.252916
##
## $oddsratio
## [1] 0.8411352
```

- a 数据如上
- 从 oddsratio 可以看出 aspirin 导致心脏疾病的 odds(导致与不导致的比率) 是 placebo 的 1.0265 倍
- 95% 的置信区间为 [0.841,1.253]

0.5 2.13

```
data("afterlife");afterlife1
```

```
##          Belief
## Gender    Yes No or Undecided
##   Females 509             116
##   Males   398             104
```

```
addmargins(afterlife1)
```

```
##          Belief
## Gender     Yes No or Undecided  Sum
##   Females  509             116  625
##   Males    398             104  502
##   Sum      907             220 1127
```

```
se <- sqrt(((509/625)*(1-509/625)/625)+((398/502)*(1-398/502)/502))
di <- (509/625)-(398/502);di
```

```
## [1] 0.02157131
```

```
ci <- c(di-qnorm(0.95)*se,di+qnorm(0.95)*se);ci
```

```
## [1] -0.01766588  0.06080851
```

```
oddsratio(afterlife1)
```

```
##   oddsratio
## 1  1.146595
```

```
lodd <- log(oddsratio(afterlife1))
se <- sqrt(1/509+1/116+1/398+1/104)
ci <- c(exp(lodd+qnorm(0.95)*se),exp(lodd-qnorm(0.95)*se));ci
```

```
## $oddsratio
## [1] 1.469161
##
## $oddsratio
## [1] 0.8948511
```

```
p <- G2(afterlife1);p
```

```
## [1] 0.6628491
```

- 90% confidence interval is [-0.226,0.269], 表明相信有来世的人群中，男性的比率比女性大的 90% 的置信区间是 [-0.226,0.269]
- 90% confidence interval is [0.894,1.469], 表明在男性中，相信有来世与无来世的比率比女性中的之比的 90% 的置信区间为 [0.894,1.469]
- 根据 G2 检验的结果可以知道其 p-value 为 0.6628，在 95% 的置信水平下可以认为其是独立的，即是否相信来世与性别无关

0.6 2.16

- response variable is lung cancer, explanatory variable is have smoked
- case control study, a study that investigated the relationship between smoking and lung cancer
- we can use it to compare smokers with nonsmokers, because it uses cross-sectional design.

```
column <- c('cases','control')
row <- c('yes','no')
data216<- matrix(data=c(688,21,650,59), nrow = 2, ncol = 2, byrow = F,dimnames = list(row,
```

```
##     cases control
## yes   688     650
## no     21      59
```

```r
oddsratio(data216)
```

```
##   oddsratio
## 1  2.973773
```

```r
p <- G2(data216);p
```

```
## [1] 4.825515e-05
```

- p 值足够小，可以在 95% 的置信水平下认为 smoke 和 lung cancer 有相关关系

0.7 2.27

```r
column <- c('high','h_gra','college','c_gra')
row <- c('low','middle','high')
data227<- matrix(data=c(9,44,13,10,11,52,23,22,9,41,12,27), nrow = 3, ncol = 4, byrow = T,
```

```
##        high h_gra college c_gra
## low       9    44      13    10
## middle   11    52      23    22
## high      9    41      12    27
```

```r
independent_test_of_table(data227, "X2")
```

```
## $method
## [1] "X2"
##
## $statistic
## [1] 8.870942
##
## $df
## [1] 6
##
## $p.value
## [1] 0.1809674
```

```r
independent_test_of_table(data227, "G2")
```

```
## $method
## [1] "G2"
##
## $statistic
## [1] 8.916528
##
## $df
## [1] 6
##
## $p.value
## [1] 0.1783272
```

```r
x2_result <- chisq.test(data227)
x2_result$stdres
```

```
##              high      h_gra    college      c_gra
## low     0.4061328  1.5828205 -0.1286367 -2.1078423
## middle -0.1898118 -0.5440627  1.3041565 -0.4031584
## high   -0.1903291 -0.9459053 -1.2374420  2.4360173
```

```r
independent_test_of_table(data227, "M2",u=c(5,15,25),v = c(1,2,3,4))
```

```
## $method
## [1] "M2"
##
## $statistic
## [1] 4.748927
##
## $df
## [1] 1
##
## $p.value
## [1] 0.02931658
```

- 根据 x2 和 G2 的结果，其 p 值都大于 0.05，则在 95% 的显著水平我们可以认为 aspiration 与 family income 是相互独立的，缺陷在于，这里面没有考虑 aspiration 的序数关系，他们是有序的

- 从 standardized residuals 可以看出，其中有两个绝对值超过 2，故可以认为应该是不独立的，需要进行相关分析
- 使用 M2 方法来进行假设检验，并对序数进行设置，工资中认为 low-income 为 5，middle 为 15，high 为 25，学历方面 high school 为 1，high school graduate 是 2，some college 是 3，college graduate 是 4，检验结果中 p 值为 0.029, 在 95% 的置信水平下可以拒绝原假设，即可以认为 aspiration 和 family income 存在相关关系。

0.8 2.29

```
column <- c('yes','no')
row <- c('prednisolone','control')
d29<- matrix(data=c(7,0,8,15), nrow = 2, ncol = 2, byrow = F,dimnames = list(row,column));
```

```
##              yes no
## prednisolone   7  8
## control        0 15
```

```
fisher.test(d29, alternative = "g")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  d29
## p-value = 0.003161
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  2.645931      Inf
## sample estimates:
## odds ratio
##        Inf
```

- 使用 Fisher's exact test,p 值为 0.0031，所以在 95% 的置信水平下，可以认为 oddsratio 大于 1，即尼龙松治疗是有效果的

0.9 2.33

```r
data233=c(19,11,0,6,151-19,63-11,9,103-6)
dim(data233) <- c(2,2,2)
firstdim <- c('white','black')
secdim <- c('white','black')
thidim <- c('yes','no')
dimnames(data233) <- list(firstdim,secdim,thidim)
#first question
ftable(data233)
```

```
##                yes  no
##
## white white   19 132
##       black    0   9
## black white   11  52
##       black    6  97
```

```r
#add 0.5
data <- data233 + 0.5
ftable(data)
```

```
##                 yes    no
##
## white white   19.5 132.5
##       black    0.5   9.5
## black white   11.5  52.5
##       black    6.5  97.5
```

```r
# victim is white
data[,1,]
```

```
##         yes    no
## white 19.5 132.5
## black 11.5  52.5
```

```r
independent_test_of_table(data[,1,],'X2')
```

```
## $method
```

```
## [1] "X2"
##
## $statistic
## [1] 0.5949342
##
## $df
## [1] 1
##
## $p.value
## [1] 0.4405174
```

```
independent_test_of_table(data[,1,],'G2')
```

```
## $method
## [1] "G2"
##
## $statistic
## [1] 0.934627
##
## $df
## [1] 1
##
## $p.value
## [1] 0.3336635
```

```
# victim is black
data[,2,]
```

```
##       yes   no
## white 0.5  9.5
## black 6.5 97.5
```

```
independent_test_of_table(data[,2,],'X2')
```

```
## $method
## [1] "X2"
##
```

```
## $statistic
## [1] 3.099661e-31
##
## $df
## [1] 1
##
## $p.value
## [1] 1
```

```
independent_test_of_table(data[,2,],'G2')
```

```
## $method
## [1] "G2"
##
## $statistic
## [1] 0.02616369
##
## $df
## [1] 1
##
## $p.value
## [1] 0.8715012
```

```
#defendant race and penalty
margin.table(data, c(2, 3))
```

```
##       yes  no
## white  31 185
## black   7 107
```

```
independent_test_of_table(margin.table(data, c(2, 3)),'X2')
```

```
## $method
## [1] "X2"
##
## $statistic
## [1] 4.164978
```

```
##
## $df
## [1] 1
##
## $p.value
## [1] 0.04126795
```

```
independent_test_of_table(margin.table(data, c(2, 3)),'G2')
```

```
## $method
## [1] "G2"
##
## $statistic
## [1] 5.41384
##
## $df
## [1] 1
##
## $p.value
## [1] 0.01997773
```

- 结果如上
- 从 partial data 来看，在 95% 的置信水平上, 可以认为，当受害者为白人时，是否判死刑与被告人的人种相互独立；当受害者为黑人时，是否判死刑与被告人的人种相互独立；从 margin data 来看，在 95% 的置信水平上, 可以认为，被告人的种族与被告人是否判死刑有关的
- 从 2 的结果来看，Simpson's paradox 成立,partial data 的结果与 marginal data 的结果不一致