threeway

Heejung Shim

4991 Wisconsin male high school seniors according to socio-economic status (low, lower middle, upper middle, and high), the degree of parental encouragement they receive (low and high) and whether or not they have plans to attend college (no, yes). Fienberg (1977, p. 101)

```
encouraged <- gl(2, 1, 16, labels=c("low", "high"))
soc_stratum <- gl(4, 2, 16, labels=c("lower", "lower middle", "upper middle", "higher"))
plans <- gl(2, 8, 16, labels=c("no", "yes"))
counts <- c(749, 233, 627, 330, 420, 374, 153, 266, 35, 133, 38, 303, 37, 467, 26, 800)
(wisconsin <- data.frame(counts, encouraged, soc_stratum, plans))
```

```
##    counts encouraged  soc_stratum plans
## 1     749        low        lower    no
## 2     233       high        lower    no
## 3     627        low lower middle    no
## 4     330       high lower middle    no
## 5     420        low upper middle    no
## 6     374       high upper middle    no
## 7     153        low       higher    no
## 8     266       high       higher    no
## 9      35        low        lower   yes
## 10    133       high        lower   yes
## 11     38        low lower middle   yes
## 12    303       high lower middle   yes
## 13     37        low upper middle   yes
## 14    467       high upper middle   yes
## 15     26        low       higher   yes
## 16    800       high       higher   yes
```

```
wt <- xtabs(counts ~ soc_stratum + encouraged + plans, wisconsin)
ftable(wt)
```

```
##                            plans  no yes
## soc_stratum  encouraged
## lower        low                 749  35
##              high                233 133
## lower middle low                 627  38
##              high                330 303
## upper middle low                 420  37
##              high                374 467
## higher       low                 153  26
##              high                266 800
```

a 20-year follow-up study on the effects of smoking

Appleton, French and Vanderpump (1996)

In the period 1972 - 74, a larger study categorized women into smokers and nonsmokers and according to their age group. In the follow-up, the researchers recorded whether the subjects were dead or still alive. Only smokers or women who had never smoked are presented here. Relatively few smokers quit and these women have been excluded from the data. The cause of death is not reported here. Here is the data.

```
library(faraway)
data(femsmoke)
str(femsmoke)
```

```
## 'data.frame':    28 obs. of  4 variables:
##  $ y     : num  2 1 3 5 14 7 27 12 51 40 ...
##  $ smoker: Factor w/ 2 levels "yes","no": 1 2 1 2 1 2 1 2 1 2 ...
##  $ dead  : Factor w/ 2 levels "yes","no": 1 1 1 1 1 1 1 1 1 1 ...
##  $ age   : Factor w/ 7 levels "18-24","25-34",..: 1 1 2 2 3 3 4 4 5 5 ...
```

```
ct <- xtabs(y ~ smoker + dead + age, femsmoke)
ftable(ct)
```

```
##             age 18-24 25-34 35-44 45-54 55-64 65-74 75+
## smoker dead
## yes    yes          2     3    14    27    51    29  13
##        no          53   121    95   103    64     7   0
## no     yes          1     5     7    12    40   101  64
##        no          61   152   114    66    81    28   0
```

Question of interest is the relationship between smoking and risk of death. We can combine the data over age groups and produce:

```
(cta <- xtabs(y ~ smoker+dead, femsmoke))
```

```
##       dead
## smoker yes  no
##    yes 139 443
##    no  230 502
```

```
prop.table(cta, 1)
```

```
##       dead
## smoker       yes        no
##    yes 0.2388316 0.7611684
##    no  0.3142077 0.6857923
```

```r
## pearson's chisquared test
summary(cta)
```

```
## Call: xtabs(formula = y ~ smoker + dead, data = femsmoke)
## Number of cases in table: 1314
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 9.121, df = 1, p-value = 0.002527
```

76% of smokers have survived for 20 years while only 69% of nonsmokers have survived. Thus, smoking appears to have beneficial effect on longevity. This dependency between smoking and risk of death is significant.

It seems that smoking reduce risk of death (significantly),

but not for nearly all individual age groups (albeit insignificant).

```r
cta <- xtabs(y ~ smoker+dead, femsmoke)
prop.table(cta, 1)
```

```
##       dead
## smoker       yes        no
##    yes 0.2388316 0.7611684
##    no  0.3142077 0.6857923
```

```r
summary(cta)
```

```
## Call: xtabs(formula = y ~ smoker + dead, data = femsmoke)
## Number of cases in table: 1314
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 9.121, df = 1, p-value = 0.002527
```

```r
ct1 <- xtabs(y ~ smoker+dead, femsmoke, subset=(age=="18-24"))
prop.table(ct1, 1)
```

```
##       dead
## smoker        yes         no
##    yes 0.03636364 0.96363636
##    no  0.01612903 0.98387097
```

```r
summary(ct1)
```

```
## Call: xtabs(formula = y ~ smoker + dead, data = femsmoke, subset = (age ==
##     "18-24"))
## Number of cases in table: 117
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 0.4776, df = 1, p-value = 0.4895
##  Chi-squared approximation may be incorrect
```

```r
ct2 <- xtabs(y ~ smoker+dead, femsmoke, subset=(age=="25-34"))
prop.table(ct2, 1)
```

```
##       dead
## smoker        yes         no
##    yes 0.02419355 0.97580645
##    no  0.03184713 0.96815287
```

```r
summary(ct2)
```

```
## Call: xtabs(formula = y ~ smoker + dead, data = femsmoke, subset = (age ==
##     "25-34"))
## Number of cases in table: 281
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 0.14673, df = 1, p-value = 0.7017
##  Chi-squared approximation may be incorrect
```

```r
ct3 <- xtabs(y ~ smoker+dead, femsmoke, subset=(age=="35-44"))
prop.table(ct3, 1)
```

```
##       dead
## smoker        yes         no
##    yes 0.12844037 0.87155963
##    no  0.05785124 0.94214876
```

```r
summary(ct3)
```

```
## Call: xtabs(formula = y ~ smoker + dead, data = femsmoke, subset = (age ==
##     "35-44"))
## Number of cases in table: 230
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 3.444, df = 1, p-value = 0.06349
```

```r
ct4 <- xtabs(y ~ smoker+dead, femsmoke, subset=(age=="45-54"))
prop.table(ct4, 1)
```

```
##       dead
## smoker       yes        no
##    yes 0.2076923 0.7923077
##    no  0.1538462 0.8461538
```

```r
summary(ct4)
```

```
## Call: xtabs(formula = y ~ smoker + dead, data = femsmoke, subset = (age ==
##     "45-54"))
## Number of cases in table: 208
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 0.9278, df = 1, p-value = 0.3354
```

```r
ct5 <- xtabs(y ~ smoker+dead, femsmoke, subset=(age=="55-64"))
prop.table(ct5, 1)
```

```
##       dead
## smoker       yes        no
##    yes 0.4434783 0.5565217
##    no  0.3305785 0.6694215
```

```r
summary(ct5)
```

```
## Call: xtabs(formula = y ~ smoker + dead, data = femsmoke, subset = (age ==
##     "55-64"))
## Number of cases in table: 236
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 3.172, df = 1, p-value = 0.0749
```

```r
ct6 <- xtabs(y ~ smoker+dead, femsmoke, subset=(age=="65-74"))
prop.table(ct6, 1)
```

```
##       dead
## smoker       yes        no
##    yes 0.8055556 0.1944444
##    no  0.7829457 0.2170543
```

```r
summary(ct6)
```

```
## Call: xtabs(formula = y ~ smoker + dead, data = femsmoke, subset = (age ==
##     "65-74"))
## Number of cases in table: 165
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 0.08609, df = 1, p-value = 0.7692
```

```r
ct7 <- xtabs(y ~ smoker+dead, femsmoke, subset=(age=="75+"))
prop.table(ct7, 1)
```

```
##       dead
## smoker yes no
##    yes   1  0
##    no    1  0
```

```r
summary(ct7)
```

```
## Call: xtabs(formula = y ~ smoker + dead, data = femsmoke, subset = (age ==
##     "75+"))
## Number of cases in table: 77
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = NaN, df = 1, p-value = NA
##  Chi-squared approximation may be incorrect
```

The marginal association where we add over the age groups is different from the conditional association observed within age groups.

This is an example of Simpson's paradox.

Let's see why it happens.

```
prop.table(xtabs(y ~ smoker+age, femsmoke), 2)
```

```
##       age
## smoker     18-24     25-34     35-44     45-54     55-64     65-74
##    yes 0.4700855 0.4412811 0.4739130 0.6250000 0.4872881 0.2181818
##    no  0.5299145 0.5587189 0.5260870 0.3750000 0.5127119 0.7818182
##       age
## smoker       75+
##    yes 0.1688312
##    no  0.8311688
```

We can see smokers are more concentrated in the younger age group and younger people are more likely to live for another 20 years. This explains why the marginal table gave an apparent advantage to smokers which is, in fact, not real because once we control for age, we see that smoking has a negative effect on longevity.

test if all three factors independent, clearly not!

```
modi <- glm(y ~ smoker + dead + age, femsmoke, family=poisson)
deviance(modi)
```

```
## [1] 735.0028
```

```
df.residual(modi)
```

```
## [1] 19
```

```
pchisq(735.0028, df=19, lower.tail = FALSE)
```

```
## [1] 1.362676e-143
```

test if age independent of smoking and death, clearly not!

```
modj <- glm(y ~ smoker*dead + age, femsmoke, family=poisson)
deviance(modj)
```

```
## [1] 725.8025
```

```r
df.residual(modj)
```

```
## [1] 18
```

```r
pchisq(725.8025, df=18, lower.tail = FALSE)
```

```
## [1] 1.889863e-142
```

are smoking and death conditionally independent given age? could be

```r
modc <- glm(y ~ smoker*age + age*dead, femsmoke, family=poisson)
deviance(modc)
```

```
## [1] 8.326939
```

```r
df.residual(modc)
```

```
## [1] 7
```

```r
pchisq(deviance(modc), df.residual(modc), lower.tail=FALSE)
```

```
## [1] 0.3046493
```

table has some zeros which mean the chisq assumption may not hold

model comparison test is more reliable

```r
modd <- glm(y ~ (smoker + dead + age)^2, femsmoke, family=poisson)
# modd <- glm(y ~ smoker*age + age*dead + smoker*dead, femsmoke, family=poisson)  # equivalent
anova(modc, modd, test="Chi")
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ smoker * age + age * dead
## Model 2: y ~ (smoker + dead + age)^2
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         7     8.3269
## 2         6     2.3809  1    5.946  0.01475 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# looks like smoking and death are related after all
```