Categorical data analysis: Two way contingency tables

Marianne Huebner, Michigan State University

2015-05-01

Students in grades 4-6 were asked whether good grades, athletic ability, or popularity was most important to them. The responses were divided by school area into urban, suburban, and rural. Is there an association between the type of school area and the students' choice of good grades, athletic ability, or popularity as most important?

*Reference: Chase, M.A and Dummer, G.M. (1992), "The Role of Sports as a Social Determinant for Children," Research Quarterly for Exercise and Sport, 63, 418-424. Dataset available through the Statlib Data and Story Library (DASL).*

grades<-c(38, 58, 69)  
popular<-c(34, 28, 17)  
sports<-c(28, 14, 14)  
kids<-rbind(grades,popular, sports)  
colnames(kids)<-c("rural" , "suburban", "urban")

Adding row and column sums:

addmargins(kids)

## rural suburban urban Sum  
## grades 38 58 69 165  
## popular 34 28 17 79  
## sports 28 14 14 56  
## Sum 100 100 100 300

Adding row and column sums, calculating proportions

addmargins(kids)

## rural suburban urban Sum  
## grades 38 58 69 165  
## popular 34 28 17 79  
## sports 28 14 14 56  
## Sum 100 100 100 300

prop.table(kids) # cell proportion= cell count/total

## rural suburban urban  
## grades 0.12666667 0.19333333 0.23000000  
## popular 0.11333333 0.09333333 0.05666667  
## sports 0.09333333 0.04666667 0.04666667

apply(prop.table(kids),1,sum) # row proportions

## grades popular sports   
## 0.5500000 0.2633333 0.1866667

apply(prop.table(kids),2,sum) # column proportions

## rural suburban urban   
## 0.3333333 0.3333333 0.3333333

Performing a test of independence of the two categories school area and important aspects:

fit<-chisq.test(kids)  
fit

##   
## Pearson's Chi-squared test  
##   
## data: kids  
## X-squared = 21.6274, df = 4, p-value = 0.0002377

Full information for each cell of observed, expected, and residuals

fit$observed

## rural suburban urban  
## grades 38 58 69  
## popular 34 28 17  
## sports 28 14 14

fit$expected

## rural suburban urban  
## grades 55.00000 55.00000 55.00000  
## popular 26.33333 26.33333 26.33333  
## sports 18.66667 18.66667 18.66667

fit$residuals

## rural suburban urban  
## grades -2.292280 0.4045199 1.887760  
## popular 1.494011 0.3247849 -1.818795  
## sports 2.160247 -1.0801234 -1.080123

Contribution of each cell to the test statistic:

(fit$observed-fit$expected)^2/ fit$expected

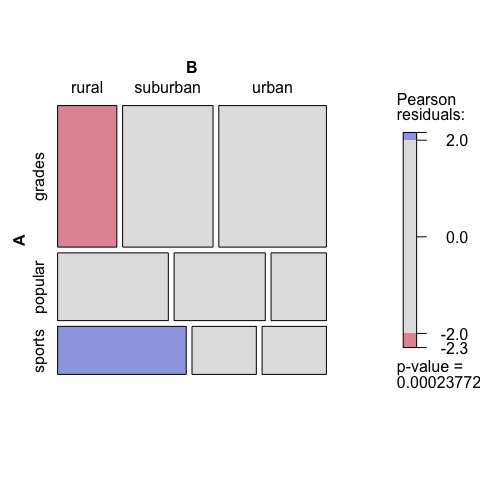
## rural suburban urban  
## grades 5.254545 0.1636364 3.563636  
## popular 2.232068 0.1054852 3.308017  
## sports 4.666667 1.1666667 1.166667

Visualization of a two-way contingency table:

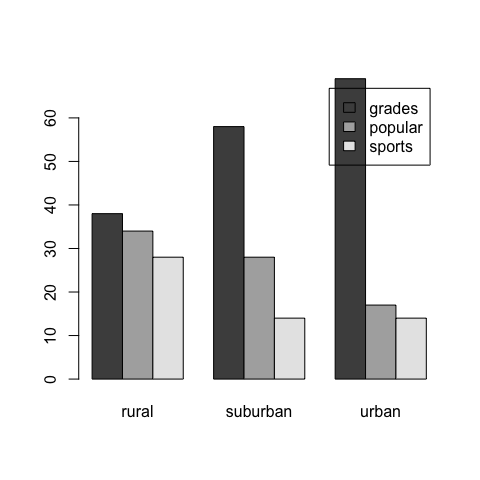
library(vcd)

## Loading required package: grid

mosaic(kids, shade=TRUE, legend=TRUE)



barplot(kids, beside=T, legend=T)



# if column sums do not add up to 100 as in the example, the following will produce such barplots  
#x<-prop.table(kids,2)  
#barplot(x, beside=T, legend=T)

**Results: For 4-6 graders the importance of grades, popularity, athletic ability seems to be associated with the type of school area (urban, suburban, or rural). In particular, grades seem to be less important than expected in a rural school and athletic ability more important than expected in a rural school. The side-by-side barplots seem to indicate that grades become increasingly important for more urban type school areas.**