02441 Applied Statistics and Statistical Software

Exercise 4C - Popular

Subjects were students in grades 4-6 from three school districts in Ingham and Clinton Counties, Michigan. Chase and Dummer stratified their sample, selecting students from urban, suburban, and rural school districts with approximately 1/3 of their sample coming from each district. Students indicated whether good grades, athletic ability, or popularity was most important to them. They also ranked four factors: grades, sports, looks, and money, in order of their importance for popularity. The questionnaire also asked for gender, grade level, and other demographic information.

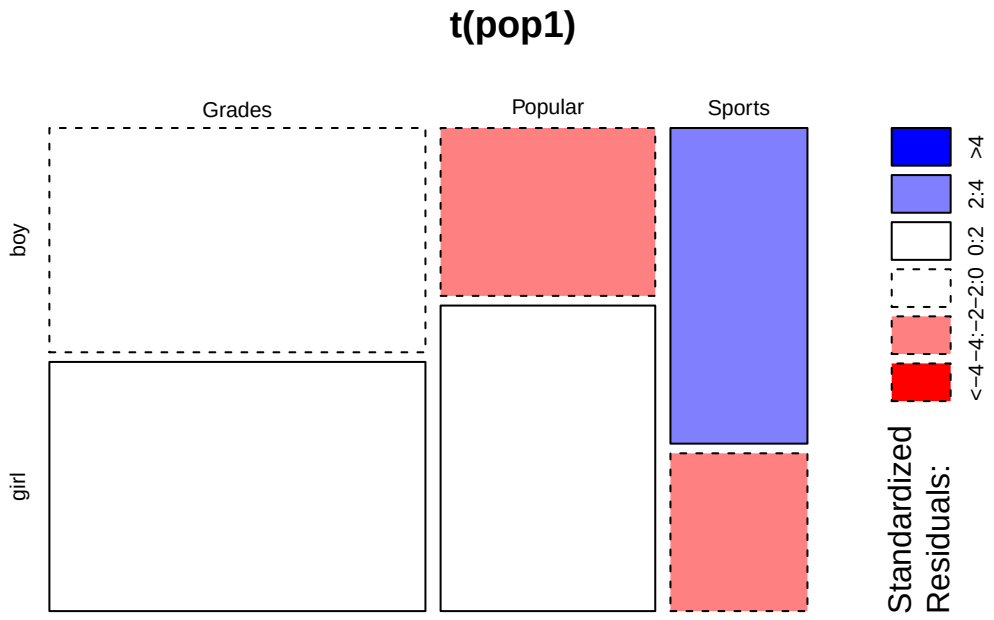
Variable name	Description
Gender	Boy or girl
Grade	4, 5, or 6
Age	Age in years
Race	White, Other
Urban/Rural	Rural, Suburban, or Urban school district
School	Brentwood Elementary, Brentwood Middle, Ridge, Sand, Eureka, Brown, Main Portage, Westdale Middle
Goals	Student's choice in the personal goals question where options were 1 = Make Good Grades 2 = Be Popular, 3 = Be Good in Sports
Grades	Rank of "make good grades"(1=most important for popularity, 4=least important)
Sports	Rank of "being good at sports"(1=most important for popularity, 4=least important)
Looks	Rank of "being handsome or pretty"(1=most important for popularity, 4=least important)
Money	Rank of "having lots of money"(1=most important for popularity, 4=least important)

1. Analyze the relationship between gender and goals.

```
pop <- read.table("popular.txt", sep = '\t', header = TRUE)
pop1 <- table(pop$Gender, pop$Goals)
chisq.test(pop1)
```

```
##
##  Pearson's Chi-squared test
##
## data:  pop1
## X-squared = 21.455, df = 2, p-value = 2.193e-05
```

```
mosaicplot(t(pop1), shade = TRUE)
```



There is a significant interaction between Gender and Goals. It turns out that sports are more important for boys than for girls. On the other hand boys don't find it as important to be popular as girls do.

2. Analyze the relationship between age and goals.

```
pop2 <- table(pop$Age, pop$Goals)
chi2 <- chisq.test(pop2)
```

```
## Warning in chisq.test(pop2): Chi-squared approximation may be incorrect
```

```
chi2$expected
```

```
##
##          Grades    Popular     Sports
##   7   0.5167364  0.2949791  0.1882845
##   9  51.6736402 29.4979079 18.8284519
##   10 68.2092050 38.9372385 24.8535565
##   11 97.6631799 55.7510460 35.5857741
##   12 26.8702929 15.3389121  9.7907950
##   13  2.0669456  1.1799163  0.7531381
```

```
pop2
```

```
##
##      Grades Popular Sports
##   7       1       0      0
##   9      50      28     22
##   10     72      39     21
##   11     94      63     32
##   12     29      11     12
##   13      1       0      3
```

Some expected values are below 5. The Chi-Square test results are therefore unreliable. Fisher's exact test

can't be used in this case either because the table would require too many probability calculations to be carried out, i.e. of any possible table which is more extreme than the observed. Most benchtop computers do not have sufficient RAM for this purpose.

Let's simplify the data instead. There haven't been many 7th and 13th graders. We therefore merge them with their neighboring classes.

```
pop2b <- pop2[2:5,]
pop2b[1,] <- pop2b[1,] + pop2[1,]
pop2b[4,] <- pop2b[4,] + pop2[6,]
chisq.test(pop2b)
```

```
##
##  Pearson's Chi-squared test
##
## data:  pop2b
## X-squared = 6.6186, df = 6, p-value = 0.3576
```

There is no significant dependency between Age and Goals.

3. Could you suggest other analysis that may be interesting?

Chi-Square tests and Mosaic Plots can be applied to any contingency table. Contingency tables are used to analyze the relationship between two categorical variables. Hence, the individual observations in the two categorical variables mjust be countable.

Use the `str` command to see which possible contingency tables can be established.

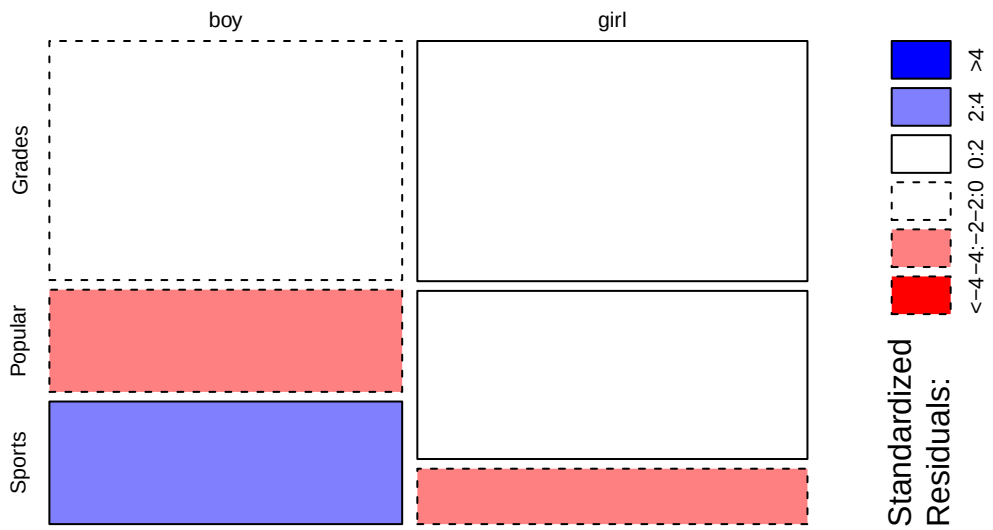
```
str(pop)
```

```
## 'data.frame':    478 obs. of  11 variables:
##  $ Gender     : Factor w/ 2 levels "boy","girl": 1 1 2 2 2 2 2 2 2 2 ...
##  $ Grade      : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ Age        : int  11 10 11 11 10 11 10 10 10 10 ...
##  $ Race       : Factor w/ 2 levels "Other","White": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Urban.Rural: Factor w/ 3 levels "Rural","Suburban",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ School     : Factor w/ 9 levels "Brentwood Elementary",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Goals      : Factor w/ 3 levels "Grades","Popular",..: 3 2 2 2 2 2 2 1 3 3 ...
##  $ Grades     : int  1 2 4 2 4 4 3 3 3 4 ...
##  $ Sports     : int  2 1 3 3 2 2 4 4 2 3 ...
##  $ Looks      : int  4 4 1 4 1 1 1 2 1 2 ...
##  $ Money      : int  3 3 2 1 3 3 2 1 4 1 ...
```

```
pop3 <- table(pop$Gender, pop$Goals)
chisq.test(pop3)
```

```
##
##  Pearson's Chi-squared test
##
## data:  pop3
## X-squared = 21.455, df = 2, p-value = 2.193e-05
```

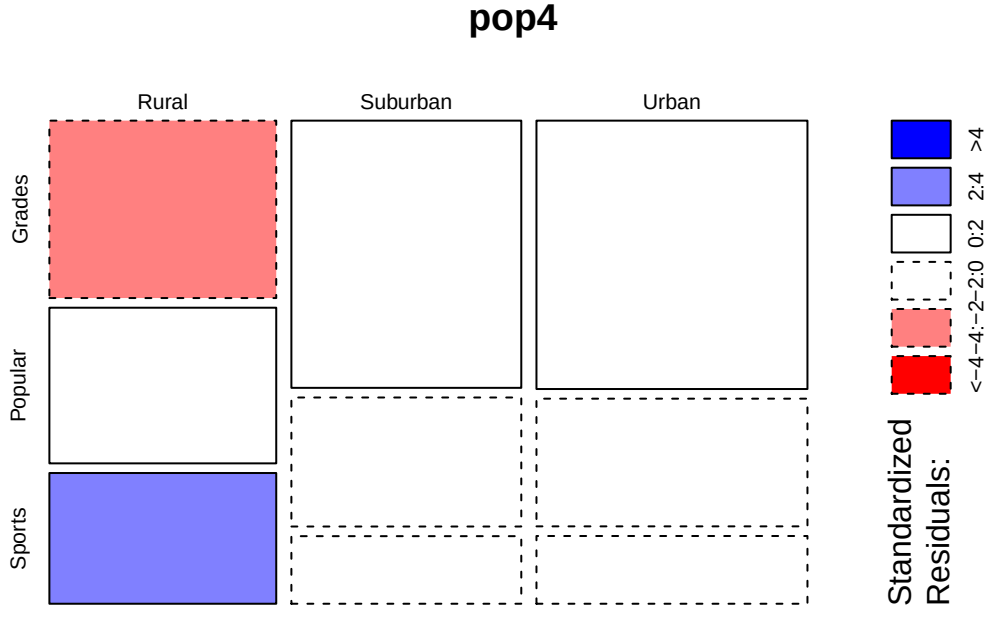
```
mosaicplot(pop3, shade = TRUE)
```




```
pop4 <- table(pop$Urban.Rural, pop$Goals)
chisq.test(pop4)
```

```
##
##  Pearson's Chi-squared test
##
## data:  pop4
## X-squared = 18.828, df = 4, p-value = 0.0008497
```

```
mosaicplot(pop4, shade = TRUE)
```



Three way contincency tables

One can also add a third categorical variable to the `table` command. This way the relationship between two categorical variables can be analyzed while conditioning for the third.

```
pop5 <- table(pop$Gender, pop$Urban.Rural, pop$Goals)
pop5
```

```
## , ,  = Grades
##
##
##        Rural Suburban Urban
##   boy     21       51    45
##   girl    36       36    58
##
## , ,  = Popular
##
##
##        Rural Suburban Urban
##   boy     19       20    11
##   girl    31       22    38
##
## , ,  = Sports
##
##
##        Rural Suburban Urban
##   boy     26       18    16
##   girl    16        4    10
```

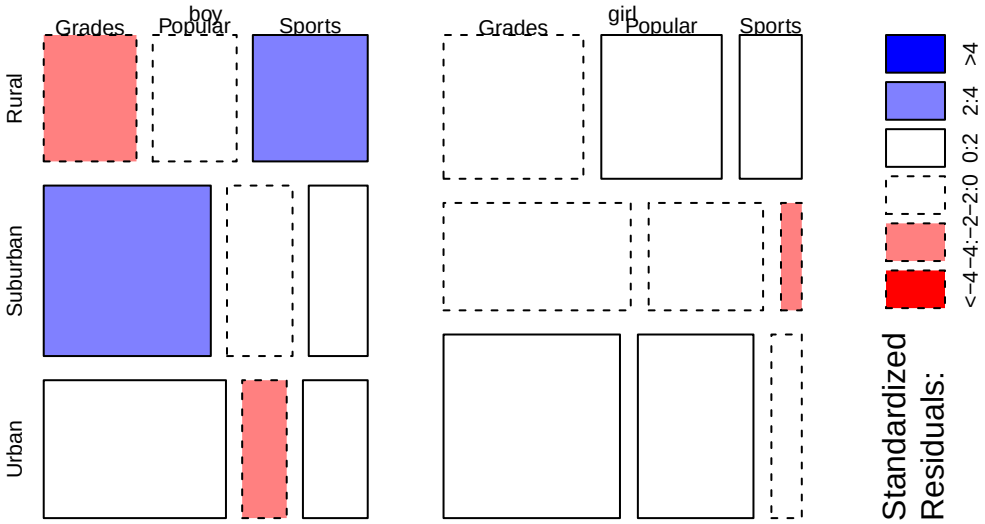
```
chisq.test(pop5[1,,])
```

```
##
##  Pearson's Chi-squared test
##
## data:  pop5[1, , ]
## X-squared = 16.074, df = 4, p-value = 0.002921
```

```
chisq.test(pop5[2,,])
```

```
##
##  Pearson's Chi-squared test
##
## data:  pop5[2, , ]
## X-squared = 7.6821, df = 4, p-value = 0.1039
```

```
mosaicplot(pop5, shade=TRUE)
```


pop5
boy
girl
Grades
Popular
Sports
Rural
Suburban
Urban
Standardized Residuals:
<−4
−4:−2
−2:0
0:2
2:4
>4