STAT 477/577 - Technology Guide

Module 2 - Section 4

Test of Independence

Below is an explanation of the R commands and functions needed to conduct a test of independence for two categorical variables.

For the example from the lecture notes, we first need to read in the data from the file `smoking.csv`.

```
smoking.data<- read.csv(file.choose(), header = T)
```

Then we need to set the category order for both variables.

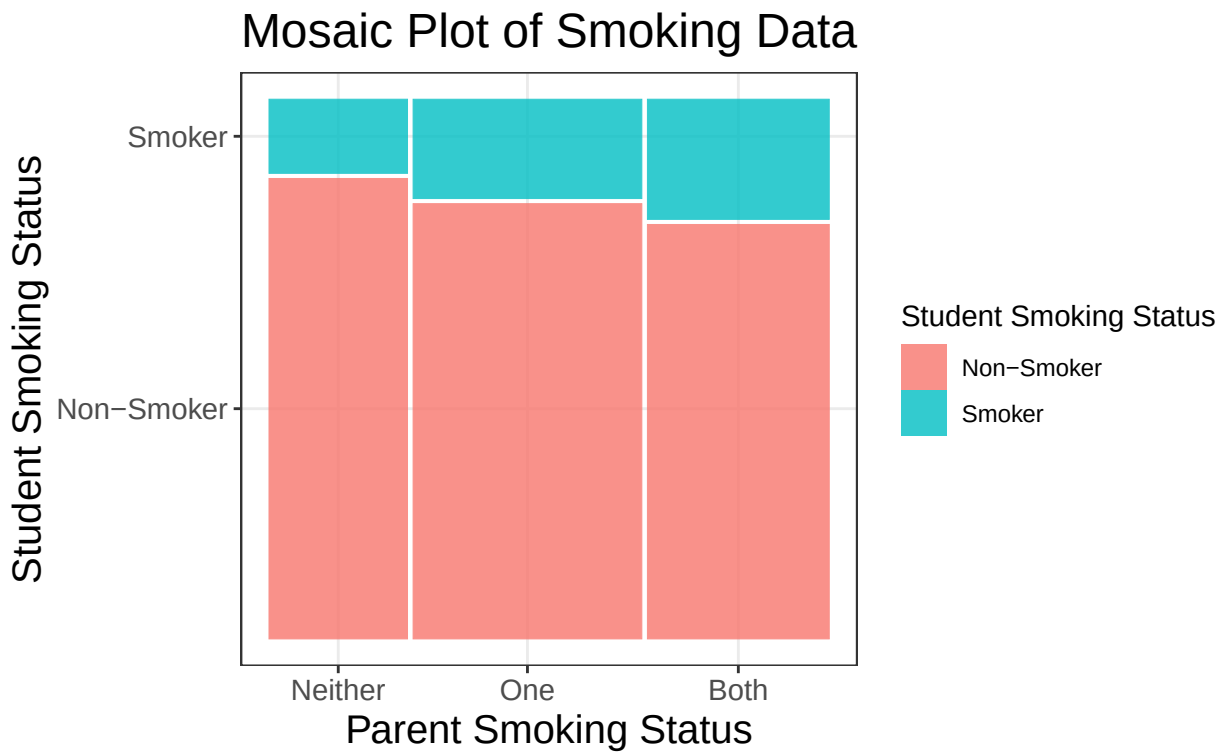
```
smoking.data$Student<- factor(smoking.data$Student,
                         levels = c("Non-Smoker", "Smoker"))
smoking.data$Parent<- factor(smoking.data$Parent,
                         levels = c("Neither", "One", "Both"))
```

We will then use R to calculate the contingency table

```
smoking.table<- table(smoking.data$Parent, smoking.data$Student)
```

and to graph the mosaic plot for the two variables.

```
ggplot(data = smoking.data)+
  geom_mosaic(aes(x = product(Student, Parent), fill = Student),
                na.rm = TRUE, divider =  mosaic("h"))+
  theme_bw()+
  theme(axis.title.y = element_text(size = rel(1.4)),
        axis.title.x = element_text(size = rel(1.4)),
        axis.text.x = element_text(size = rel(1.2)),
        axis.text.y = element_text(size = rel(1.2)),
        strip.text.y = element_text(size = rel(1.4)),
        plot.title = element_text(hjust=0.5, size = rel(1.6)))+
  labs(x = "Parent Smoking Status",
       y = "Student Smoking Status",
       fill = "Student Smoking Status",
       title = "Mosaic Plot of Smoking Data")
```



To perform the test of independence, we will use the R function `chisq.test`. Using the smoking data from lecture, the chi-square test of independence can be calculated using the command

```
smoking.test<- chisq.test(smoking.table)
```

The output of the test is contained in the variable `smoking.test`. Here is the output:

```
smoking.test

##
##  Pearson's Chi-squared test
##
## data:  smoking.table
## X-squared = 37.566, df = 2, p-value = 6.959e-09
```

The expected values in the contingency table and the contribution of each cell in the table are saved as the variables

```
smoking.test$expected

##
##           Non-Smoker   Smoker
##   Neither   1102.712 253.2882
##   One       1820.776 418.2244
##   Both      1447.513 332.4874
```

and

```
(smoking.test$residuals)^2
```

```
##
##            Non-Smoker      Smoker
##   Neither  3.86551335 16.82884348
##   One      0.00271743  0.01183057
##   Both     3.14881241 13.70862455
```