Tutorial 4

ST2137-2420

Material

This tutorial covers the topics and concepts from chapter 4 of the course textbook. The questions offer practice on understanding and visualising categorical data. The final section introduces yet another measure that we can use to assess association between two categorical random variables.

As always, work out each question with both R and Python unless otherwise stated.

Dataset: Student Performance

1. The columns `address` and `paid` are both binary variables. Conduct a χ^2-test at 5% significance level to assess if there is any significant association between the two variables.
2. Create a barchart to visualise these two columns.
3. Create a table to present row-wise proportions instead of raw counts. The `address` variable should appear in the rows.
4. *Relative risk* is another measure of association between two categorical variables. With the above two variables, define the following:
   - Let $\hat{p}_1$ be the proportion of rural residents who paid for extra classes.
   - Let $\hat{p}_2$ be the proportion of urban residents who paid for extra classes. Then

   $$\text{relative risk} = \hat{p}_1 / \hat{p}_2$$

   Compute the relative risk for the above variables.
5. What are the range of values for RR? How is it similar/different to Odds Ratio?
6. Read up on the `cut` function in R and `pd.cut` in `pandas`. Use it to divide the `G3` column into letter grades:

G3 score	Letter grade
[0, 10]	F
(10, 12]	D
(12, 15]	C
(15, 18]	B
(18, 20]	A

7. There are 6 Likert-scale survey questions (`famrel` to `health`). Which one of them is most strongly associated with `letter_grade` (from question 6)? Which measure of association did you use?
8. Recreate the following mosaic plot for `Dalc` and `Walc`. Interpret what the plot shows. In addition, identify the cell counts corresponding to the labelled rectangles (“Cell Count?”).

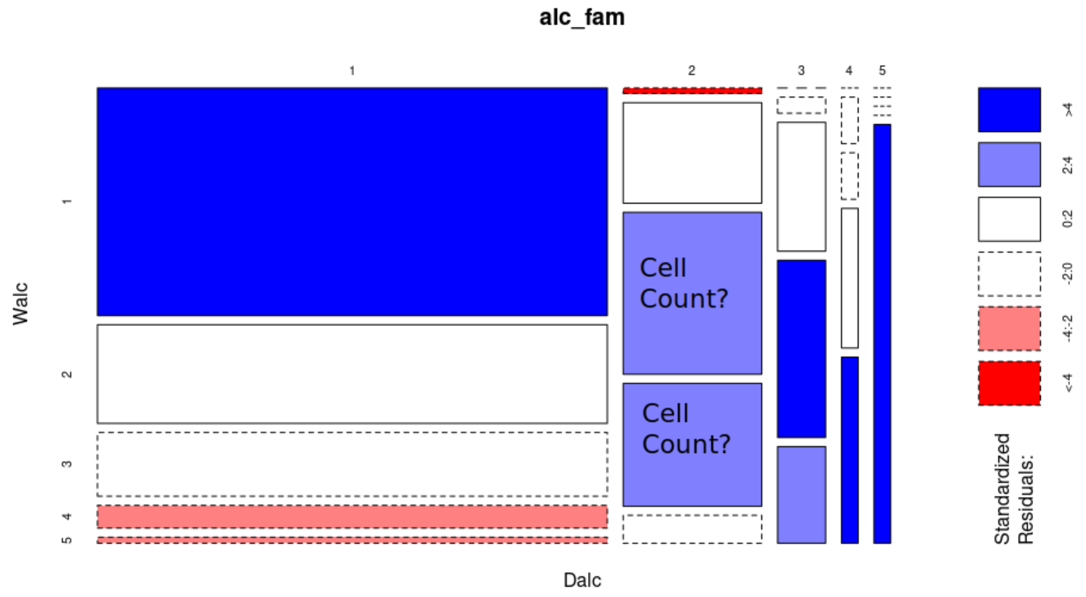

Figure 1: Mosaic plot

Mutual Information

If X and Y are two discrete random variables, then the mutual information between them is given by

$$I(X,Y) = \sum_{i,j} P(X=i, Y=j) \times \ln\left(\frac{P(X=i,Y=j)}{P(X=i)P(Y=j)}\right)$$

To clarify, you have to consider all possible values that X takes on, and all possible values that Y takes on in the summation.

Here is some intuition behind MI: $I(X,Y)$ measures how different the joint pmf is from the product of the marginals. If X and Y are independent, then $I(X,Y)$ will be 0. In this case, X and Y hold no information about each other - X would probably not be useful in predicting Y. In general, the more positive this value $I(X,Y)$ is, the more associated these two variables are, and the more useful X could be for predicting Y.

In order to estimate probabilities, a typical approach is to compute the sample proportions.

9. Find the mistakes in the following functions that have been written to compute MI.

```
## Solution 1:
mi1 <- function(x, y) {
  total <- length(x)
  output = 0

  pX = table(x)/length(x)
  pY = table(y)/length(y)
  X <- unique(pX) probablity should not be unique
  Y <- unique(pY)

  pXY <- as.vector(table(x,y)/length(x))
  XY <- unique(pXY)
  for(i in X) {
    for (j in Y){
      for (k in XY){
        output <- output + k * log(k/(i+j)) log(k/(i*j))
      }
    }
  }
  return(output)
}

## Solution 2:
mi2 <- function(X, Y) {
  total <- 0
  for (i in 1:length(X)) {
    for (j in 1:length(Y)) {
      ProbX <- length(which(X == X[[i]])) / length(X) X==X[i]
      ProbY <- length(which(Y == Y[[j]])) / length(Y) Y==Y[i]
      ProbXY <- ProbX * ProbY  Assume that variable X and Y are independent
      prob <- ln((ProbX + ProbY) / ProbXY) There is no ln() method in R, ProbX*ProbY
      total1 <- ProbXY * prob OK
    }
    total <- total + total1
  }
  total
}
```

10. Implement your own function to compute Mutual Information in R. Use it to compute the MI between `address` and `paid`.