Homework 3

Shun-Lung Chang, Dilip Hiremath

```r
# library packages
library(knitr)
library(vcd)
library(DescTools)
library(psych)
```

```r
# import dataset
load("data/creditcard.Rdata")
```

1. First of all, tabulate the CreditCard variable.

```r
table(creditcard$CreditCard)
```

```
  0   1
383 307
```

(a) (half a point) How many applicants was the credit card approved?

The credit card was approved for 307 applicants.

(b) (1 point) What are the odds of getting a credit card application denied?

```r
383 / 307
```

```
[1] 1.247557
```

The odds of credit card application being denied is 1.247557.

(c) (1 point) What is the "risk" (i.e. probability) of getting a credit card application approved?

The risk (probability) of getting a credit card application approved is 0.4449275.

```r
307 / (383 + 307)
```

```
[1] 0.4449275
```

2. Cross-tabulate the variables Children and CreditCard.

```r
table(creditcard$CreditCard, creditcard$Children)
```

```
      0   1
  0 213 170
  1 161 146
```

(a) (1 point) How many applicants with children in their household was the credit card denied?

The credit card denied for 170 applicants with children in their household.

```r
table(creditcard$Children, creditcard$CreditCard)
```

```
      0   1
  0 213 161
  1 170 146
```

(b) (1 point) How many applicants without children in their household got the credit card application approved?

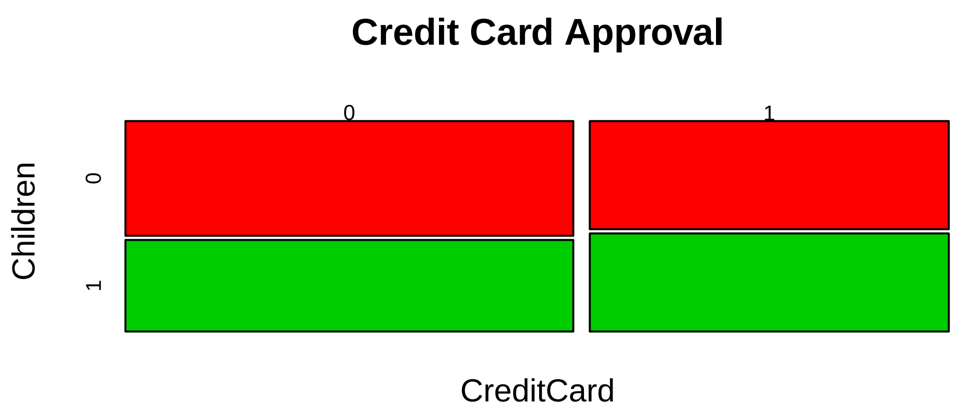
```r
xtabs(~ creditcard$Children + creditcard$CreditCard)
```

```
                   creditcard$CreditCard
creditcard$Children   0   1
                  0 213 161
                  1 170 146
```

The credit card application approved for 161 applicants without children in their household.

(c) (half a point) Draw a mosaicplot visualising the contingency table of credit card approval and whether children live in the applicant's household.

```r
mosaicplot(CreditCard ~ Children, data = creditcard, color = 2:3, main = "Credit Card Approval")
```



3. You continue with your analysis of the relationship between Children and CreditCard.

(a) (1 point) Are applicants with children in their household less likely (as measured in odds) to get credit card applications approved than others? Calculate the odds ratio for getting a credit card application approved comparing applicants with children in their household with those without.

```r
(146 / 170) / (161 / 213)
```

```
[1] 1.136208
```

```r
# or
oddsratio(xtabs(~ creditcard$Children + creditcard$CreditCard), log = FALSE)
```

```
 odds ratios for creditcard$Children and creditcard$CreditCard

[1] 1.136208
```

The odds of getting a credit card application approved comparing applicants with children in their household with those without is 1.136208.

(b) (1 point) Are applicants with children less likely (as measured in risk) to get credit card applications denied than applicants without children? Calculate the relative risk for getting a credit card application denied comparing applicants with children in their household to those without.

```r
(213 / (213 + 161)) / (170 / (170 + 146))
```

```
[1] 1.058635
```

The risk for getting a credit card application denied comparing applicants with children in their household to those without is 1.058635.

(c) (half a point) Looking at the mosiac plot created in Question 2c, how strong is the relationship between the two variables Children and CreditCard.

Based on the mosaicpot the created in question 2c, the relationship between the two variables Children and CreditCard is not that strong.

4. Now, you assess the relationship between Children and CreditCard using the χ^2-statistic.

(a) (1 point) Calculate the χ^2-statistic to assess the relationship between Children and CreditCard.

```r
chisq.test(creditcard$Children,creditcard$CreditCard)$statistic
```

```
X-squared
0.5682733
```

The χ^2-statistic is 0.56827, which indicates a weak relationship between children and credit card approval.

(b) (1.5 points) Calculate the expected frequencies under the assumption that Children has no effect on credit card approval. For which cells are expected frequencies higher than the observed ones?

```r
# observation
xtabs(~ creditcard$Children + creditcard$CreditCard)
```

```
                   creditcard$CreditCard
creditcard$Children   0   1
                  0 213 161
                  1 170 146
```

```r
# expectation
chisq.test(creditcard$Children, creditcard$CreditCard)$expected
```

```
                   creditcard$CreditCard
creditcard$Children        0        1
                  0 207.5971 166.4029
                  1 175.4029 140.5971
```

For cell representing, expected frequencies are higher than the observed ones for people who got credit card approved without children and for people who did not get credit card approved with children.

```r
chisq.test(creditcard$Children, creditcard$CreditCard)$expected >
    xtabs(~ creditcard$Children + creditcard$CreditCard)
```

```
                   creditcard$CreditCard
creditcard$Children     0     1
                  0 FALSE  TRUE
                  1  TRUE FALSE
```

5. In the following, perform the analyses separately for female and male applicants.

The dataset is partitioned into *creditcard_f* and *creditcard_m* by gender.

```r
creditcard_f <- creditcard[creditcard$Gender == 0, ]
creditcard_m <- creditcard[creditcard$Gender == 1, ]
```

(a) (half a point) Calculate the χ^2-statistic to assess the relationship between Children and CreditCard.

```r
chisq.test(creditcard_f$Children, creditcard_f$CreditCard)$statistic
```

```
X-squared
 1.083713
```

Chi-squared statistic for the relationship between males with Children and CreditCard is 1.0837.

```r
chisq.test(creditcard_m$Children, creditcard_m$CreditCard)$statistic
```

```
X-squared
 3.033659
```

Chi-squared statistic for the relationship between females with Children and CreditCard is 3.033659.

(b) (1.5 point) Calculate the expected frequencies under the assumption that Children has no effect on credit card approval. For which cells are expected frequencies higher than the observed ones?

```r
chisq.test(creditcard_f$Children, creditcard_f$CreditCard)$expected
```

```
                     creditcard_f$CreditCard
creditcard_f$Children        0        1
                    0 70.31081 58.68919
                    1 50.68919 42.31081
```

```r
chisq.test(creditcard_f$Children, creditcard_f$CreditCard)$expected >
    xtabs(~creditcard_f$Children + creditcard_f$CreditCard)
```

```
                     creditcard_f$CreditCard
creditcard_f$Children     0     1
                    0  TRUE FALSE
                    1 FALSE  TRUE
```

```r
xtabs(~creditcard_f$Children + creditcard_f$CreditCard)
```

```
                     creditcard_f$CreditCard
creditcard_f$Children  0  1
                    0 66 63
                    1 55 38
```

The expected frequencies under the assumption that females with Children have no effect on credit card approval is shown in the first table.

For cells representing females without children who get credit card denied and females get credit card approved with children have expected frequencies higher than the observed ones.

```r
chisq.test(creditcard_m$Children, creditcard_m$CreditCard)$expected
```

```
                     creditcard_m$CreditCard
creditcard_m$Children        0         1
                    0 137.1581 107.84188
                    1 124.8419  98.15812
```

```r
chisq.test(creditcard_m$Children, creditcard_m$CreditCard)$expected >
    xtabs(~ creditcard_m$Children + creditcard_m$CreditCard)
```

```
                     creditcard_m$CreditCard
creditcard_m$Children     0     1
                    0 FALSE  TRUE
                    1  TRUE FALSE
```

```r
xtabs(~ creditcard_m$Children + creditcard_m$CreditCard)
```

```
                     creditcard_m$CreditCard
creditcard_m$Children   0   1
                    0 147  98
                    1 115 108
```

The expected frequencies under the assumption that males with Children have no effect on credit card approval is shown in the first table.

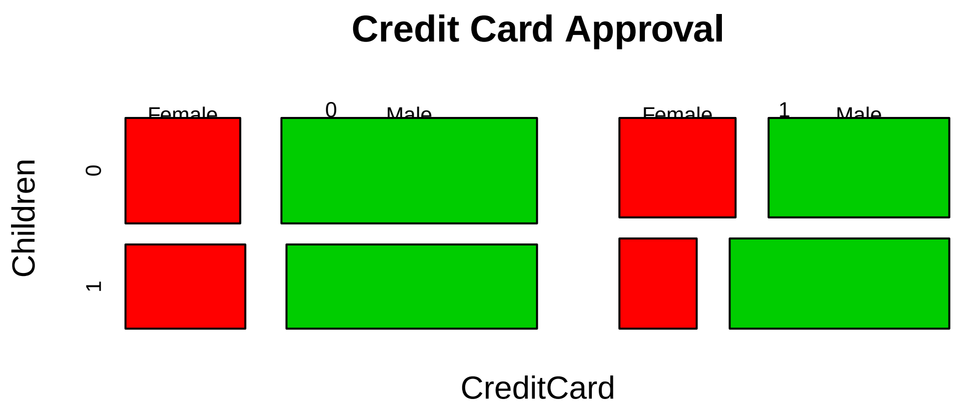
For cells representing males without children who get credit card approved and males who get credit card denied and with children have expected frequencies higher than the observed ones.

(c) (half a point) Do the results differ for the two sexes?

Yes, the results differ for the two sexes as shown in the above tables.

6. (2.5 points) Visualise the relationships using mosaicplots. Do the differences between females and males in relation to credit card approval and children in the household become visible in the plots? Provide reason for your answer!

```r
# better labels
creditcard$gender <- ifelse(creditcard$Gender == 1, "Male", "Female")

mosaicplot(CreditCard ~ Children + gender,
           data = creditcard, color = 2:3, main = "Credit Card Approval")
```



Yes, the differences between females and males in relation to credit card approval and children becomes visible in the plots. We see that for male applicants who have children are less likely to be denied credit card than females do.

7. (2.5points) Compute the odds ratios to assess the relationship between Children and CreditCard for males and females separately. Are the results more in line with the χ^2-statistic or more in line with the mosaicplot?

```r
oddsratio(xtabs(~ creditcard_f$Children + creditcard_f$CreditCard), log = FALSE)
```

```
 odds ratios for creditcard_f$Children and creditcard_f$CreditCard

[1] 0.7238095
```

```r
oddsratio(xtabs(~ creditcard_m$Children + creditcard_m$CreditCard), log = FALSE)
```

```
 odds ratios for creditcard_m$Children and creditcard_m$CreditCard

[1] 1.408696
```

The results are more in line with the mosaicplot, as the odds for males with children are more likely to get their credit card approved.

8. Now, you assess the relationships between SavingsType and CreditCard.

(a) (half a point) Calculate the χ^2-statistic to assess the relationship between SavingsType and CreditCard.

```r
chisq.test(creditcard$SavingsType, creditcard$CreditCard)$statistic
```

```
X-squared
  9.19157
```

The χ^2-statistic for relationship between SavingsType and Creditcard is 9.19157.

(b) (1 point) Calculate the Phi-coefficient, the Contingency Coefficient and Cramer's V to assess the relationship between SavingsType and CreditCard.

```r
# cannot be calculated, since the size of the contigency table is not 2 x 2
# phi(table(creditcard$SavingsType, creditcard$CreditCard))

relation <- list("Contingency Coefficient" =
ContCoef(x = creditcard$SavingsType, y = creditcard$CreditCard),
"Cramer's V" = CramerV(x = creditcard$SavingsType, y = creditcard$CreditCard))

relation
```

```
$`Contingency Coefficient`
[1] 0.114656

$`Cramer's V`
[1] 0.1154171
```

The calculated values are as below:

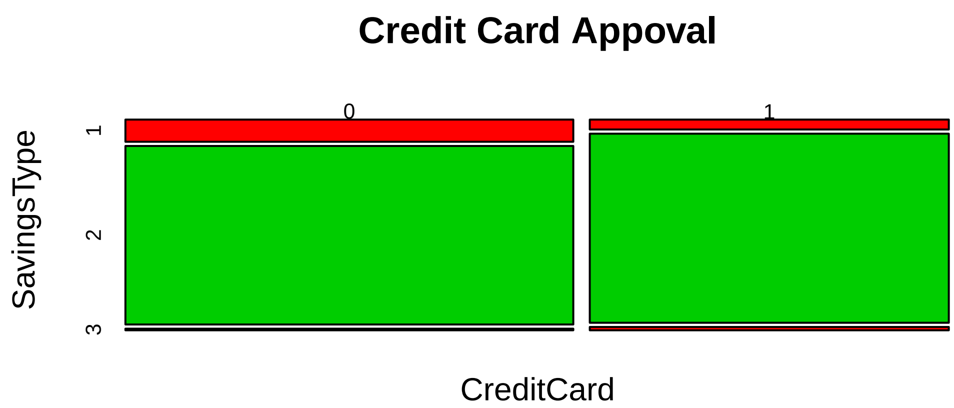
- Contingency Coefficient: 0.114656,
- Cramer's V : 0.1154171.

(c) (half a point) Why is there no result for the Phi-Coefficient?

Since the size of the contigency table is not 2 x 2, there is no result for Phi-coefficient.

(d) (half a point) Visualize the relationship between credit card approval and savings type. Which of the statistics used for this data comes closer to the visual representation of the relationship's strength?

```r
mosaicplot(CreditCard ~ SavingsType, data = creditcard, color = 2:3,
           main = "Credit Card Appoval")
```



Spearmen correlation coefficient comes closer to the visual representation of the relationship's strength, since it measures the relationship between two discrete variables.

9. (2.5 points) Since there are too few cases for savings type 3 (CDs), you exclude applicants having this savings type and re-run the analysis of Question 8 again. Summarize the results and comment on the differences.

```
# remove savings type 3
creditcard_remove_3 <- creditcard[!creditcard$SavingsType == 3, ]
```

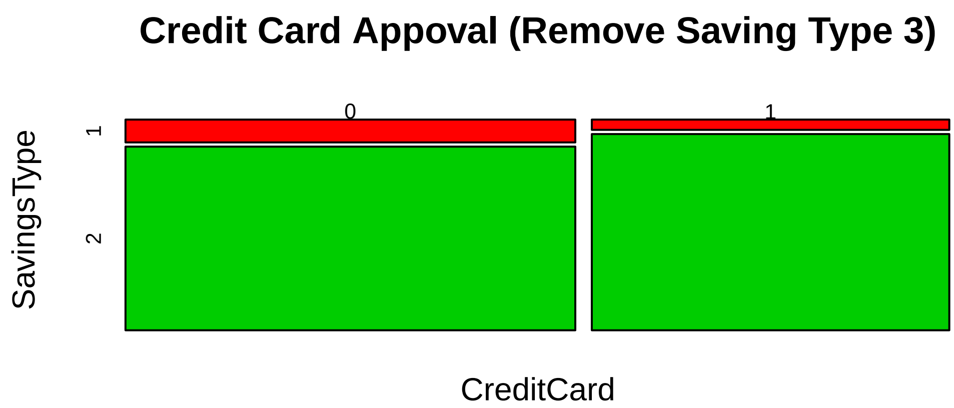
```
chisq.test(creditcard_remove_3$SavingsType, creditcard_remove_3$CreditCard)
```

```

	Pearson's Chi-squared test with Yates' continuity correction

data:  creditcard_remove_3$SavingsType and creditcard_remove_3$CreditCard
X-squared = 7.3615, df = 1, p-value = 0.006663
```

```
relation <- list("Phi-coefficient" =
                     phi(table(creditcard_remove_3$SavingsType, creditcard_remove_3$CreditCard)),
                 "Contingency Coefficient" =
ContCoef(x = creditcard_remove_3$SavingsType, y = creditcard_remove_3$CreditCard),
"Cramer's V" = CramerV(x = creditcard_remove_3$SavingsType,
                       y = creditcard_remove_3$CreditCard))

relation
```

```
$`Phi-coefficient`
[1] 0.11

$`Contingency Coefficient`
[1] 0.1085816

$`Cramer's V`
[1] 0.1092274
```

```
mosaicplot(CreditCard ~ SavingsType, data = creditcard_remove_3, color = 2:3,
           main = "Credit Card Appoval (Remove Saving Type 3)")
```



After removing savings type 3, the following values are obtained:

- Chi-squared statistic is 7.3615,
- Phi-coefficient is 0.11,
- Contingency Coefficient is 0.1085816,
- Cramer's V is 0.1092274.

We see that the chi-squared value has reduced from 9.19 to 7.36 Contigency coefficient has not changed from 0.114656 to 0.1085816 and Cramer's V has changed from 0.1154171 to 0.1092274.

The decrease indicates that the null hypothesis would be less likely to be rejected after type 3 is removed.