STAT 659 Spring 2016
Homework 3 Solution

2.16

(a) The response variable is the smoking status and the explanatory variable is the lung cancer.

(b) It is the case-control study.

(c) No, because the marginal distribution of lung cancer is fixed. But we can infer the proportion of smokers for people with lung cancer and compare it with the proportion of smokers for people without lung cancer.

(d) The proportion of smokers for people with lung cancer is $p_1 = 0.9704$ while the proportion of smokers for people without lung cancer is $p_2 = 0.9168$. So the relative risk is $p_1/p_2 = 1.058$ which means the people with lung cancer is 1.058 times more likely to be smokers than the people without lung cancer. The odd ratio is 2.97, which means the relative risk of smokers for people with lung cancer is 2.97 times that for people without lung cancer.

2.18

(a) The sample size $n = \sum_{i=1}^{3}\sum_{j=1}^{3} n_{i,j} = 1362$. Under the $H_0 : p_{i,j} = p_{i+} \cdot p_{+j}$, the expected count for the first cell is $n \cdot p_{1+} \cdot p_{+1} = 1362 \cdot 0.123 \cdot 0.213 = 35.8$.

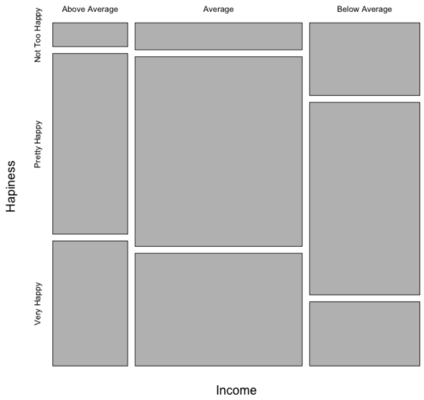
(b) Under the null hypothesis, the test statistics has a chi-square distribution with degree of freedom $(I-1)(J-1) = 4$. Then, the p-value is $P(\chi_4^2 > 73.4) = 4.33 \cdot 10^{-15} \approx 0$ We reject the null hypothesis and conclude that there exists association between happiness and income.

(c) For the corner cells having counts 21 and 83, the standardized residuals are negative. It means there are fewer people who are not too happy with above average income and who are very happy with below average income than expected that under the independence assumption.

(d) For the corner cells having counts 110 and 94, the standardized residuals are positive. It means that there are more people who are very happy with above average income and who are not too happy with below average income than expected under the independence assumption.

(e) From the previous questions, we can see a linear trend is possible for the level of income and the feeling of happiness. The variable income is ordinal, we can assign scores to rows and columns and perform a M^2 ordinal test. Let $u_i = 3, 2, 1$ be scores for the rows and $v_j = 1, 2, 3$ be scores for the columns, where $1 \leq i, j \leq 3$. Then $r = \frac{\sum\limits_{i,j}(u_i - \bar{u})(v_j - \bar{v})p_{i,j}}{\sqrt{\sum\limits_{i}(u_i - \bar{u})^2 p_{i+} \times \sum\limits_{j}(v_j - \bar{v})^2 p_{+j}}} = 0.2027$ and $M^2 = (n-1) * r^2 = 55.9$. P-value ≤ 0.0001. So the independence assumption should be rejected.

(f) The mosaic plot for the data is shown below:



The mosaic plot suggest that happiness increases when people have more income. Thus, a linear trend between income and hapiness exists.

2.19

(a) From the table, we can easily obtain that $n = 2613, p_{1+} = 0.83735, p_{2+} = 0.16265, p_{+1} = 0.44891, p_{+2} = 0.20054, p_{+3} = 0.35055$. Then the Pearson $X^2 = \sum\limits_{i=1}^{2}\sum\limits_{j=1}^{3} \frac{(n_{i,j} - np_{i+}p_{+j})^2}{np_{i+}p_{+j}} = 167.85$, while the likelihood ratio test statistic is $G^2 = 2\sum\limits_{i=1}^{2}\sum\limits_{j=1}^{3} n_{i,j}\log(\frac{n_{i,j}}{np_{i+}p_{+j}}) = 187.58$. Both two test statistics have a chi-square distribution with $df = (I-1)(J-1) = 2$ and the p-values for them are zero. Therefore, we reject the null hypothesis that the party identification and race are independent.

(b) The standardized residuals for each cell are shown below:

-11.853	0.692	11.775
11.853	-0.692	-11.775

We can see there are fewer democrat identifications for white race/republican identifications for black race than expected counts under the independence assumption; there are more democrat identifications for black people/republican identifications for white people than expected counts under independence assumption.

(c) Take the likelihood ratio test statistic for an example, we can partition the table into two parts: (i)Democrat and Independent;(ii)Democrat + independent and Republican. For the first component, the $G_1^2 = 24.046$ and for the second component, the $G_2^2 = 163.53$. The test statistics for both components suggest the independence assumption does not hold for each sub-table.

2.20

The data are shown below:

Breastcancer	alone	withspouse	withothers
Yes	$144 \times 0.41 = 59.04$	$209 \times 0.522 = 109.098$	$89 \times 0.596 = 53.044$
No	$144 \times 0.59 = 84.96$	$209 \times 0.478 = 99.902$	$89 \times 0.404 = 35.956$

The Pearson test statistic $X^2 = \sum_i^2 \sum_j^3 \frac{(n_{i,j} - np_{i+}p_{+j})^2}{np_{i+}p_{+j}} = 8.351$ and the likelihood ratio test statistic $G^2 = 2\sum_i^2 \sum_j^3 n_{i,j} log(\frac{n_{i,j}}{np_{i+}p_{+j}}) = 8.3969$. Both test statistics have a chi-square distribution with $df = 2$. The p-value for Pearson test statistic is $P(\chi_2^2 > 8.351) = 0.0154 \approx 0.02$; the likelihood ratio test statistic has p-value $0.015 \approx 0.02$. It shows that the race and the party identification are not independent.

2.21

(a) No, since each object can belong to multiple columns, the three factors are dependent.

(b) To test whether the factor A is responsible for increase in teenage crime, the contingency table can be constructed as below:

Gender	A	NotA
Men	60	40
Women	75	25

2.22

(a) Here we only consider using the Pearson statistic to do the test and the test by likelihood ratio statistic is similar. $X^2 = \sum_{i=1}^2 \sum_{j=1}^3 \frac{(n_{i,j} - np_{i+}p_{+j})^2}{np_{i+}p_{+j}} = 84.188$ and X^2 has a chi-square distribution with $df = 4$ under the independence assumption. The p-value is $P(\chi_4^2 > 84.188) = 0$. So the independence assumption should be rejected.

(b) The standardized Person residuals for each cell are shown below:

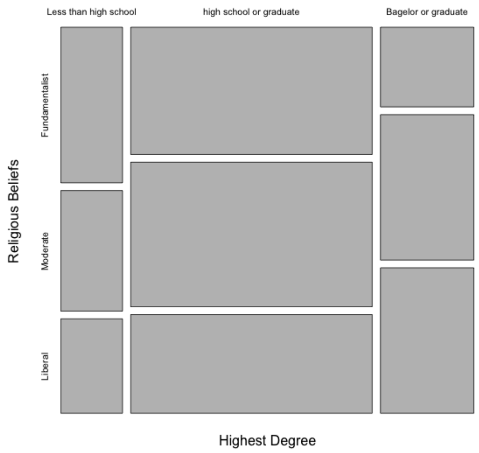
Diagnosis	Drugs	NoDrugs
Schizophrenia	7.8745	-7.8745
Affectivedisorder	1.6023	-1.6023
Neurosis	-2.3853	2.3853
Personalitydisorder	-4.8417	4.8417
Specialsymptoms	-5.1395	5.1395

The absolute values of standardized residuals for Schizophrenia, Neurosis, Personality disorder and Special symptoms are greater than 2. There are more Schizophrenia patients with drugs than expected under independence assumption; there are fewer Neurosis, Personality disorder and Special symptoms patients with drugs than expected under independence assumption.

(c) The Pearson statistic for component (i) is 0.8917, which has a chi-square distribution with $df = 1$ and p-value 0.3450; the Pearson statistic for component (ii) is 0.01488, which has a chi-square distribution with $df = 1$ and p-value 0.9029; the Pearson statistic for component (iii) is 83.8839, which has a chi-square distribution with $df = 2$ and p-value close to 0. So reject the null hypothesis of independence. There is significant difference on comparing these three types of diagnosis and the usage of drugs.

2.23

The mosaic plot for the data is shown below:



If the two variables are independent, then the cells in a given row of the mosaic plots should have the same heights. From the plots, we can see the counts of liberal level with highest degree Bachelor or graduate are more than expected under independence assumption. So the independence assumption is violated by the data.

2.27

(a) We can construct contingency table as follows:

family income	high school	high school graduate	some college	college graduate
Low	9	44	13	10
Middle	11	52	23	22
High	9	41	12	27

The sample size $n = 273$. The Pearson test statistic $X^2 = \sum_{i=1}^{3}\sum_{j=1}^{4} \frac{(n_{i,j} - np_{i+}p_{+j})^2}{np_{i+}p_{+j}} = 8.8709$ and $G^2 = 8.92$, both have chi-square distribution with degree of freedom $(I-1)(J-1) = 6$. The corresponding p-value are $P(\chi_6^2 > 8.8709) = 0.1810$ and $P(\chi_6^2 > 8.92) = 0.1783$, which indicates that the independence assumption is not violated. But the family income and the level of education are ordinal variables(have ranks), so we perform a nominal test for ordinal data which is not appropriate.

(b) The standardized Pearson residuals are as follows:

family income	high school	high school graduate	some college	college graduate
Low	0.4061	1.5828	-0.1286	-2.1078
Middle	-0.1898	-0.54406	1.30416	-0.40316
High	-0.1903	-0.9459	-1.2374	2.4360

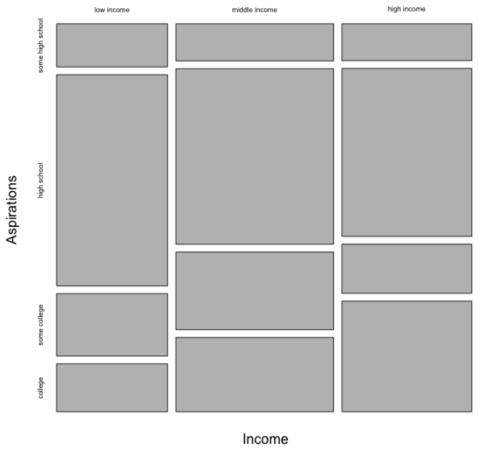
We can see the counts of some high school level aspiration for low income family is more than expected counts under independence assumption while the counts of highest college graduate level aspiration for high income family is more than expected counts under the independence. Thus a linear trend between aspirations and family income exists.

(c) We can assign scores to row levels and column levels. Let $u_1 = 1, u_2 = 2, u_3 = 3$ be scores of low family income, middle family income and high family income separately and let $v_1 = 1, v_2 = 2, v_3 = 3, v_4 = 4$ be scores of some high school, high school graduate, some college and college graduate separately. Then

$$r = \frac{\sum_{i,j}(u_i - \bar{u})(v_j - \bar{v})p_{i,j}}{\sqrt{\sum_i (u_i - \bar{u})^2 p_{i+} \times \sum_j (v_j - \bar{v})^2 p_{+j}}} = 0.097/\sqrt{0.604 \cdot 0.895} = 0.13193$$

The test statistic $M^2 = (n-1)r^2 = 4.74$ which has a chi-square distribution with $df = 1$. The p-value is 0.03, so we reject the independence assumption for family income and the aspirations.

(d) The mosaic plot suggests larger percentage of students with middle income family has some college aspiration and higher percentage of students from higher income level has aspiration for college graduate. Thus a linear trend between aspirations and family income exists.



The remaining problems are only for students who have taken STAT 414, 610 or STAT 630.

2.25

(a) $\sum_j \hat{\mu}_{i,j} = \sum_j n_{i+}n_{+j}/n = n_{i+} \cdot \sum_j n_{+j}/n = n_{i+} \cdot n/n = n_{i+}$. Similarly, $\sum_i \hat{\mu}_{i,j} = \sum_i n_{i+}n_{+j}/n = n_{+j} \cdot \sum_i n_{i+}/n = n_{+j} \cdot n/n = n_{+j}$.

(b) $\hat{\mu}_{1,1} \cdot \hat{\mu}_{2,2}/\hat{\mu}_{1,2} \cdot \hat{\mu}_{2,1} = \frac{n_{1+}n_{+1}n_{2+}n_{+2}/n^2}{n_{1+}n_{+2}n_{2+}n_{+1}/n^2} = 1$

2.26

(a) Due to the definition of chi-square distribution, Z has a chi-square distribution with $df = 1$.

(b) Suppose $Y_1 = \sum_{i=1}^{df_1} Z_{1,i}^2$ and $Y_2 = \sum_{j=1}^{df_2} Z_{2,j}^2$, where $Z_{1,i}, Z_{2,j}$ are standard normal variables,$i = 1, 2, \cdots, df_1, j = 1, 2, \cdots, df_2$. Since Y_1 and Y_2 are independent chi-squared variates, then $Z_{1,i}$ and $Z_{2,j}$ are mutually independent. So Y is still sum of $df_1 + df_2$ squared standard normal variables. By definition, Y has a chi-squared distribution with degree of freedom $df_1 + df_2$.