STAT 659 Spring 2016
Homework 4 Solution

2.29

Let θ be the odd ratio between the experimental group and the control group. Then the null hypothesis is $\theta = 1$ and the alternative is $\theta > 1$. The potential value for n_{11} is 7, so the p-value is $P(n_{11} = 7) = 0.00316 < 0.05$. Thus we reject H_0 and we are 95 percent sure that the results were significantly better for treatment than control.

2.30

According to the table 2.17, we know $n_{1+} = 23, n_{2+} = 18, N = n_{1+} + n_{2+} = 41, n_{+1} = 36$. The potential values for n_{11} are $21, 22, 23$. So the corresponding p-value for fisher's exact test is $\sum_{i=21}^{23} P(n_{11} = i) = 0.3808 \approx 0.381$. So we fail to reject null hypothesis at confidence level 95 percent.

2.31

(a) The two sided exact p-value is $P(n_{11} = 18) + P(n_{11} = 19) + P(n_{11} = 21) + P(n_{11} = 22) + P(n_{11} = 23) = 0.6384$, which indicates that the $\theta = 1$.

(b) The mid-P-value is $P(n_{11} = 21)/2 + P(n_{11} = 22) + P(n_{11} = 23) = 0.2431$, which is much smaller. Since the distribution of n_{11} is hypergeometric distribution which is discrete, the ordinary p-value is too conservative. The mid-p-value can reduce the conservativeness which equals to half the probability of observed result plus the probabilities of more extreme results.

2.33

(a) The data table is as follows:

victims	defenders	Yes for death penalty	No for death penalty	row sum
white	white	19	132	151
	black	11	52	63
black	white	0	9	9
	black	6	97	103
col sum		36	290	

(b) For white victims,

defenders	Yes for death penalty	No for death penalty
white	19	132
black	11	52

Conditional on the white victims, the odd ratio for death penalty between white defenders and black defenders is $19.5/132.5 \cdot 52.5/11.5 = 0.6719$, which means conditional on white victims, the odds of death penalty for white defenders is 0.6719 times as that for black defenders. For black victims,

defenders	Yes for death penalty	No for death penalty
white	0	9
black	6	97

Conditional on the black victims, the odd ratio for death penalty between white defenders and black defenders is $0.5/9.5 \cdot 97.5/6.5 = 0.7895$. It means conditional on black victims, the odds of death penalty for white defenders is 0.7895 of that for black defenders.

(c) The marginal odds ratio between defendants' race and death penalty verdict is $19/141 \cdot 149/17 = 1.181$, which is greater than conditional odds ratios. So marginal association has different directions from the conditional associations, which is the Simpson's paradox.

2.35

This is also a case of Simpson's paradox. Although at each age level, the death rate of South Carolina is higher than that of Maine, we do not know the death number of the two states at each age level. So it is possible that the age distribution of people in Maine is higher than in South Carolina so that the number of deaths at each age level in Maine is higher than in South Carolina. When you calculate the marginal death rate, you sum all the deaths at each age level, which can cause the marginal death rate of Maine is higher.

2.37

(a) The data table is as follows:

	race	murder victims proportion
male	nonwhite	0.0263
	white	0.0049
female	nonwhite	0.0072
	white	0.0023

So conditional on gender male, the odds ratios between race and whether a murder victim is $\frac{0.0049}{1-0.0049} \cdot \frac{1-0.0263}{0.0263} = 0.1823$, which means conditional on males, the odds of murder victim for white newborn males is 0.1823 times as that for nonwhite newborn males. Conditional on gender female, the odds ration is $\frac{0.0023}{1-0.0023} \cdot \frac{1-0.0072}{0.0072} = 0.3179$, which means the odds of murder victim for white newborn females is 0.3179 times as that for nonwhite newborn females.

(b) The marginal odd ratio is $\frac{1-(0.0263+0.0072)/2}{(0.0263+0.0072)/2} \cdot \frac{(0.0049+0.0023)/2}{1-(0.0049+0.0023)/2} = 0.2120$.

3.2

(a) Since the intercept is -0.0003 which is much smaller than the slope, then P equals to 3.04 which is the proportion of vote for Buchanan in 2000 of that for Perot in 1996.

(b) $\hat{\pi}_i = -0.0003 + 0.0304 * 0.0774 = 0.002053$. So $\pi - \hat{\pi}_i = 0.005847$ and $\pi/\hat{\pi}_i = 3.848$. Then the observation is an outlier since it is much larger than the estimated result.

3.3

(a) The prediction equation is $y = 0.00255 + 0.00109x$. The intercept is the estimated probability for a baby with sex organ malformation when her mother's alcohol consumption is zero; the slope is the increase of probability of the baby with sex organ malformation when her mother's alcohol consumption increases to the next level.

(b) $Y(0) = 0.00255, Y(7) = 0.00255+7*0.00109 = 0.01018$. The relative risk is $0.01018/0.00255 = 3.9922 \approx 4.00$.

3.4

(a) The re-fitted model is $y = 0.0026 + 0.0007x$, so $Y(0) = 0.0026, Y(7) = 0.0075$. The relative risk is $0.0075/0.0026 = 2.885$, which is smaller than the result in 3.3. So the fitting result is sensitive to this single malformation observation.

(b) If using the new score, the fitted model is $y = 0.0026 + 0.0005x$. $Y(0) = 0.0026, Y(4) = 0.0046$ and relative risk is 1.769. So the result is also sensitive to the choice of score.

(c) When fitting the logistic model, the fitted model is $\log(\text{odds}) = -5.96+0.32x$. We can see the slope is positive, which means the odds will increase with the alcohol consumption.

3.5

For linear probability model, if the score is $(0, 2, 4, 6)$, the fitted linear model is $y = 0.0176 + 0.0181x$; if the score is $0, 1, 2, 3$, the fitted model is $y = 0.0176 + 0.0362x$; if the score is $1, 2, 3, 4$, the fitted model is $y = -0.0186 + 0.0362x$. We can see the slope depends on the spacing of the scores, for the first score with spacing 2, it has half the slope of the second and the third score. But the fitted values for all above models are the same, which are

$0.0176, 0.0538, 0.09, 0.1262$ respectively.

Additional Problem A

(a) The odds ratios for each department and corresponding 95 percent confidence intervals are as follows:

department	odds ratio	confidence interval
1	0.3492	$(0.2087, 0.5844)$
2	0.8025	$(0.3404, 1.8920)$
3	1.1331	$(0.8545, 1.5024)$
4	0.9213	$(0.6863, 1.2367)$
5	1.2216	$(0.8251, 1.8088)$
6	0.8279	$(0.4552, 1.5057)$

(b) The marginal odds ratio is 1.841, which shows different directions from conditional odds ratios. Hence, Simpson's paradox seems to be present.

(c) The null hypothesis for Cochran-Mantel-Haenszel test is the gender of applicants and the admission decisions are conditional independent, that is, $\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(6)} = 1$. The alternative is that at least one conditional odds ratio is not one. When using R to perform the test, the test statistic is 1.5246, which has a chi-squared one distribution. The p-value is 0.2169, so the result indicates the null hypothesis is not violated. The null hypothesis of Breslow-Day test is the conditional association between the gender and the admission decisions are same, that is, $\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(6)}$. The alternative is that at least one conditional odds ratio is different. The test statistic is 18.8255, which has a chi-squared distribution with $df = K - 1 = 5$. The corresponding P-value is 0.0021, so by this result, we should reject H_0.

(d) When deleting the data for department one, the test statistic for Breslow-Day test is 2.53 with P-value 0.6399. So the conditional odds ratio is homogeneous and the common odds ratio will be reasonable. To construct CI for common odds ratio, we can first use logit estimator to estimate it. It is $\hat{\theta}_L = \exp\left\{\frac{\sum_{k=1}^{K} w_k \log \hat{\theta}_k}{\sum_{k=1}^{K} w_k}\right\}$ where $\hat{\theta}_k$ are estimated conditional odds ratios and the weights $w_k = \left(\frac{1}{n_{11k}} + \frac{1}{n_{12k}} + \frac{1}{n_{21k}} + \frac{1}{n_{22k}}\right)^{-1}$. Then the 90 percent confidence interval is $\exp\left\{\log \hat{\theta}_L \pm 1.645 \times \left(\sum_{k=1}^{K} w_k\right)^{-1/2}\right\}$, which is $(0.895, 1.1898)$.

Additional Problem B

(a) The conditional odds ratios and the confidence interval are as follows:

District	odds ratio	confidence interval
NC	0.6809	$(0.2519, 1.8404)$
NE	0.5926	$(0.1331, 2.6384)$
NW	0.1974	$(0.0444, 0.8766)$
SE	0.4233	$(0.1387, 1.2914)$
SW	0.3559	$(0.0899, 1.4097)$

(b) The marginal odds ratio is 0.4470,which shows the same direction with the conditional odds ratios. Hence, it seems that Simpson's paradox is not present.

(c) The null hypothesis is that conditional on the district, the race and the Merit pay have no association, $\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(5)} = 1$. The alternative is the at least one conditional odds ratio is not one. The test statistic for CMH test is 7.815, which has a chi-squared one distribution. The corresponding P-value is 0.00518, so we should reject the null hypothesis. The test statistic for Breslow-day test is 2.151, which has a chi-squared distribution with $df = 4$. Then the p-value is 0.708 which indicates that the conditional odds ratios are all equal to one.

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

Additional Problem C

For a 2x2x2 table,

		Z=1	Z=2
X=1	Y=1	n_{111}	n_{112}
	Y=2	n_{121}	n_{122}
X=2	Y=1	n_{211}	n_{212}
	Y=2	n_{221}	n_{222}

Fix Z, the homogeneous equal XY conditional odds ratios is $\frac{n_{111}n_{221}}{n_{211}n_{121}} = \frac{n_{112}n_{222}}{n_{212}n_{122}}$, or $n_{111}n_{221}n_{212}n_{122} = n_{211}n_{121}n_{112}n_{222}$. Fix X, the homogeneous equal XY conditional odds ratios is $\frac{n_{111}n_{122}}{n_{112}n_{121}} = \frac{n_{211}n_{222}}{n_{212}n_{221}}$, or $n_{111}n_{122}n_{212}n_{221} = n_{112}n_{121}n_{211}n_{222}$. These are equivalent. Thus, homogeneous association is a symmetric property.