THE CHINESE UNIVERSITY OF HONG KONG
Department of Statistics

STAT4006: Categorical Data Analysis
Problem Sheet 2 - Solutions

1. Calculate $n_{1+} = 10568$, $n_{2+} = 10392$, $n_{+1} = 229$, $n_{+2} = 20731$, $n = 20960$. They lead to the expected cell values

$$\hat{E}_{11} = \frac{n_{1+}n_{+1}}{n} = 115.4615; \hat{E}_{12} = \frac{n_{1+}n_{+2}}{n} = 10452.54; \hat{E}_{21} = \frac{n_{2+}n_{+1}}{n} = 113.5385; \hat{E}_{22} = \frac{n_{2+}n_{+2}}{n} = 10278.46$$

and the Pearson Chi-squared test statistic $\chi^2 = \sum_{Cells} \frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}} = 31.9589$. The degrees of freedom is $(2-1)(2-1) = 1$. Therefore the p-value is $P(\chi^2 > 33.43) \approx 0 < 0.0001$. There is strong evidence to reject the null hypothesis that aspirin intake an heart attacks are independent.

2. The null hypothesis is H_0: "the frequency of fatal accidents is independent of the size of the automobiles". The alternate hypothesis is H_1: "they are not independent". We choose Fisher's Exact Test because the expected counts in each cell under independence (given by $E_{ij} = n_{i+}n_{+j}/n$) are $(3, 5), (3, 5)$, thus our half of the cells have a count below 5 and our rule of thumb is not satisfied.

   To find the "extreme" tables we solve the inequality $|\frac{x}{8} - \frac{6-x}{8}| \geq |\frac{1}{8} - \frac{5}{8}|$, which has solutions $x \geq 5$ or $x \leq 1$. Thus the p-value is $1 - P(X = 2) - P(X = 3) - P(X = 4) = 0.1189$. Therefore we do not reject the null hypothesis at the 5% level. The fatality of an accident and the size of the car involved are not associated.

3. (a) The difference of proportions is 0.008547, the relative risk is 8.1822, the odds ratio is 8.2528. Both the relative risk and odds ratio show a strong association between fatality and seat belt use, but the difference of proportions shows the magnitude of the difference between the two groups is quite small. When the probabilities of fatality and non-fatality are both close to zero, the odds ratio is approximately equal to relative risk.

   (b) For each measure of association we compute the Wald CI.

   - Difference of proportions:

$$0.008547 \pm 1.96 \times \sqrt{\frac{n_{11}n_{12}}{n_{1+}^3} + \frac{n_{21}n_{22}}{n_{2+}^3}} = (0.008060, 0.009033).$$

     This 95% CI does not contain 0, so we conclude that a seat belt user is more likely to fall in the non-fatal group.

   - Relative Risk: find the CI for $\log(RR)$ first

$$\log(8.1822) \pm 1.96 \times \sqrt{\frac{1}{n_{11}} - \frac{1}{n_{1+}} - \frac{1}{n_{21}} - \frac{1}{n_{2+}}} = (2.001846, 2.202083)$$

     therefore the 95% CI for RR is $(e^{2.001846}, e^{2.202083}) = (7.402705, 9.043828)$. This 95% CI does not contain 1, so we conclude that a seat belt user is more likely to fall in the non-fatal group.

   - Odds ratio: find the CI for $\log(\theta)$ first

$$\log(8.2528) \pm 1.96 \times \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{22}} + \frac{1}{n_{12}} + \frac{1}{n_{21}}} = (2.010114, 2.211001)$$

therefore the 95% CI for θ is $(e^{2.010114}, e^{2.211001}) = (7.464166, 9.12485)$. his 95% CI does not contain 1, so we conclude that a seat belt user is more likely to fall in the non-fatal group.

4. (a) Calculate

$$C = 178 \times (648 + 442 + 252 + 252) + 183 \times (442 + 252) + 570 \times (252 + 252) + 648 \times 252 = 861310$$
$$D = 108 \times (570 + 648 + 138 + 252) + 183 \times (570 + 138) + 442 \times (138 + 252) + 348 \times 138 = 565032$$

hence $\hat{\gamma} = \frac{C-D}{C+D} = 0.2077$. We conclude there is a weak tendency for the level of education to increase with the liberalness of religious beliefs.

(b) Both variables are given scores $\mathbf{u} = \mathbf{v} = \{1, 2, 3\}$. The null hypothesis H_0 is "religious beliefs and highest degree are independent", the alternate hypothesis is "the variable are not independent". We calculate test statistic

$$r = \frac{\sum_{i=1}^{3}\sum_{j=1}^{3}(u_i - \bar{u})(v_j - \bar{v})p_{ij}}{\sqrt{[\sum_{i=1}^{3}(u_i - \bar{u})^2 p_{i+}][\sum_{j=1}^{3}(v_j - \bar{v})^2 p_{+j}]}}, M^2 = (n-1)r^2.$$

We calculate

$$p_{ij} = \frac{n_{ij}}{n}, n = 2771, p_{1+} = 0.1693, p_{2+} = 0.5991, p_{3+} = 0.2317, p_{+1} = 0.3197, p_{+2} = 0.3908, p_{+3} = 0.2894$$

$$\bar{u} = \sum_i u_i p_{i+} = 2.0624, \bar{v} = \sum_j v_j p_{+j} = 1.9697$$

leading to $r = 0.0597/0.4914 = 0.1215$. Hence $M^2 = 40.8703 > 3.841 = \chi^2_{1,0.05}$. Consequently, we reject H_0 with a two-sided test with $\alpha = 0.05$.

(c) The method we adopted in (b) considers the ordered structure of the data, which makes it more powerful than the χ^2 and G^2 tests for independence/association applied to ordinal data.

5. For example, let X = smoking (yes, no); Y = colour blindness (yes, no); Z = sex (male, female). There is no reason for people with colour blindness to be more likely to smoke, but there will be an apparent marginal association but Z is a confounding variable - males are more likely to be smokers and are much more likely to be colour blind.

6. (a) We arrange the data as six partial tables with row variable "Sex" (Mae, Female) and column variable "Admitted" (Yes, No). The sample conditional odds ratios are 0.4550, 0.8024, 1.0877, 0.9486, 1.1525, 0.8278 for Departments 1 to 6 in order. These suggests that for some departments, females are more likely to be admitted (Departments 1, 2, 4, 6); for others males are more likely to be admitted (3 and 5); for some departments the association appears strong (1, 2, 5 and 6); for others it appears weak (3 and 4).

Table 1 shows the marginal table: The marginal odds ratio is 1.8640, a clear indication that males are

	Yes	No
Male	1165	1469
Female	545	1281

Table 1: Berkeley Data (Marginal)

more likely to be accepted than females. The contradiction the mixed message of the partial associations, which if anything leaned towards the conclusion that conditional on the department, females are more likely to be admitted than males. The marginal and partial associations contradict each other - the data is a (famous) example of Simpson's Paradox.

How can we explain this? So far we have only looked at the association between A and G, conditional on G. Table 2 shows the marginal tables between D and G and D and A: We see that some departments

	Male	Female	Yes	No
Dept1	780	103	558	325
Dept2	564	23	381	206
Dept3	320	589	321	588
Dept4	409	377	260	526
Dept5	188	393	144	437
Dept6	373	341	46	668

Table 2: Berkeley Data (Marginal)

receive many more applications from males than females (1, 2); some many more females than males (3, 5); some roughly equal (4, 6). Some departments admit far more applicants than they reject (1,2); some reject far more applicants than they accept (3, 4, 5, 6). Now we can see where the marginal AG association has come from: males are more likely to apply to departments with higher acceptance rates; females are more likely to apply to departments with much lower acceptance rates. Therefore, over all the departments, males have a higher chance of being accepted than females.

THE END