THE CHINESE UNIVERSITY OF HONG KONG
Department of Statistics

STAT4006: Categorical Data Analysis
Problem Sheet 1 - Solutions

1. **(Exercise 1.1 from Agresti (2007))**

   (a) Response variable: attitude towards gun control. Explanatory variables: gender, mother's education.

   (b) Response variable: heart disease. Explanatory variables: blood pressure, cholesterol level.

   (c) Response variable: vote for president. Explanatory variables: race, religion, annual income.

   (d) Response variable: quality of life. Explanatory variables: marital status.

2. **(Exercise 1.9 from Agresti (2013))** The hypotheses for each test are

$$H_0 : P(\text{Green}) = 0.667, H_1 : P(\text{Green}) \neq 0.667.$$

   - The Wald test statistic is $\frac{\frac{854}{1103}-0.667}{\sqrt{\frac{854}{1103}\frac{249}{1103}/1103}} = 8.546492$. Therefore the p-value is $2[1-\Phi(8.546492)] < 0.0001 < 0.05$. H_0 is rejected at the 5% level.
   - The Score test statistic is $\frac{\frac{854}{1103}-0.667}{\sqrt{\frac{0.667(1-0.667)}{1103}}} = 7.579617$. Therefore the p-value is $2[1-\Phi(7.579617)] < 0.0001 < 0.05$. H_0 is rejected at the 5% level.
   - The LR test statistic is $2\log(\frac{854/1103}{0.667}) \times 854 + 2\log(\frac{249/1103}{0.333}) \times 249 = 61.44677$. Therefore the p-value is $P(\chi^2_1 > 61.44677) < 0.0001 < 0.05$. H_0 is rejected at the 5% level.

   Each test has the same conclusion. The theory behind the 2:1 ratio is not supported by the evidence.

3. **(Exercise 1.26 from Agresti (2013))**

   (a) The likelihood ratio CI for π comes from all π_0 which satisfy $-2\log[(1-\pi_0)^n/(1-\hat{\pi})^n] \leq \chi^2_1(\alpha)$. Note that $\chi^2_1(\alpha) = z^2_{\alpha/2}$ because χ^2_1 is the distribution of the square of a standard normal random variable. With $\hat{\pi} = 0$, this becomes $\log[(1-\pi_0)^n] \geq -z^2_{\alpha/2}/2$, so the CI is $[0, 1-e^{-z^2_{\alpha/2}/2n}]$. When $\alpha = 0.05$, we approximate using $e^x \approx 1+x$ for small x as follows

$$1 - e^{-z^2_{0.05/2}/2n} \approx 1 - (1 - z^2_{0.025}/2n) \approx 1.92/n$$

   yielding a CI for π of $[0, 1.92/n]$.

   (b) The Score CI for π comes from all π_0 which satisfy $\left|\frac{0-\pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}\right| \leq z_{\alpha/2}$. This has solution $\pi_0 \leq \frac{z^2_{\alpha/2}/n}{1+z^2_{\alpha/2}/n} = \pi_0 \leq \frac{z^2_{\alpha/2}}{n+z^2_{\alpha/2}}$. Therefore the CI for π is $[0, \frac{z^2_{\alpha/2}}{n+z^2_{\alpha/2}}]$.

4. The Score CI for π comes from all π_0 satisfying

$$\left|\frac{p-\pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}\right| < z_{\alpha/2} \Rightarrow p^2 - 2p\pi_0 + \pi^2 < z^2_{\alpha/2}[\pi_0(1-\pi_0)/n].$$

This is a quadratic equation with solutions

$$\frac{p + z^2_{\alpha/2}/2n \pm z_{\alpha/2}\sqrt{p(1-p)/n + z^2_{\alpha/2}/4n^2}}{1 + z^2_{\alpha/2}/n}$$

hence the CI is

$$\left(\frac{p + z^2_{\alpha/2}/2n}{1 + z^2_{\alpha/2}/n} - \frac{z_{\alpha/2}\sqrt{p(1-p)/n + z^2_{\alpha/2}/4n^2}}{1 + z^2_{\alpha/2}/n}, \frac{p + z^2_{\alpha/2}/2n}{1 + z^2_{\alpha/2}/n} + \frac{z_{\alpha/2}\sqrt{p(1-p)/n + z^2_{\alpha/2}/4n^2}}{1 + z^2_{\alpha/2}/n}\right).$$

5. (a) We test the hypothesis $H_0 : \pi_1 = 0.1, \pi_2 = 0.1, \pi_3 = 0.25, \pi_4 = 0.2$ at $\alpha = 0.05$. Perform Pearson's χ^2 test by calculating Table 1 and the test statistic $\chi^2 = \sum_{j=1}^5 \frac{(O_j - E_j)^2}{E_j} = 5.7753 < 9.488 = \chi^2_{4,0.05}$ and so

Cell	1	2	3	4	5
O_j	10	13	21	23	29
E_j	9.6	9.6	19.2	33.6	24.0
$(O_j - E_j)^2/E_j$	0.0167	1.2042	0.1688	3.3440	1.0417

Table 1: Calculating O_js and E_js

H_0 is not rejected. If we use the likelihood ratio test, the test statistic is

$$G^2 = 2\sum_{j=1}^5 O_j \log \frac{O_j}{E_j} = 6.0036 < 9.488$$

and again we do not reject H_0.

(b) The null hypothesis is $H_0 : \pi_1 = \pi_2, \pi_3 = \pi_4, \pi_5 = 1 - 2\pi_1 - 2\pi_3$. The loglikelihood under H_0 is

$$L(\pi) = (n_1 + n_2)\log \pi_1 + (n_3 + n_4)\log \pi_3 + n_5 \log \pi_5.$$

To maximize this, differentiate and solve:

$$\frac{\partial L}{\partial \pi_1} = 0 \Rightarrow \frac{n_1 + n_2}{\pi_1} = \frac{2n_5}{1 - 2\pi_1 - 2\pi_3}$$
$$\frac{\partial L}{\partial \pi_3} = 0 \Rightarrow \frac{n_3 + n_4}{\pi_3} = \frac{2n_5}{1 - 2\pi_1 - 2\pi_3}.$$

These equations imply $\hat{\pi}_1 = \hat{\pi}_2 = (n_1 + n_2)/2n$ and $\hat{\pi}_3 = \hat{\pi}_4 = (n_3 + n_4)/2n$ and $\hat{\pi}_5 = n_5/n$.

(c) Using the MLE's under H_0 (as found above), we have $\hat{\pi}_1 = \hat{\pi}_2 = 23/192$ and $\hat{\pi}_3 = \hat{\pi}_4 = 11/48$ and $\hat{\pi}_5 = 29/96$. We use these to calculate the E_js. The Pearson χ^2 test statistic is $0.482213 < 5.991 = \chi^2_{0.05,2}$. The likelihood ratio test statistic is $G^2 = 0.483362 < 5.991$. For both tests, H_0 is not rejected.

6. We test whether the data can come from a $Po(\mu)$ distribution. Our hypotheses are H_0 : "the data follows a Poisson distribution" against H_1 : "the data does not follow a Poisson distribution". The MLE for μ and $\hat{\mu} = \bar{x} = 4.1$. Use the Poisson disribution p.m.f. $P(X = x) = e^{-\mu}\mu^x/x!$ to find the expected values in each cell under H_0. Calculate Table 2: Note we turned the last column from "9" to "≥ 9", so when we fit the

Values of X	0	1	2	3	4	5	6	7	8	≥ 9
O_j	5	11	18	29	26	25	15	10	7	4
E_j	2.486	10.192	20.894	28.555	29.269	24.001	16.400	9.906	4.923	3.674

Table 2: Calculating O_js and E_js

cell probabilities $\hat{\pi}_j = e^{-\hat{\mu}}\hat{\mu}^j/j!$ for cells $j = 0, 1, \ldots, 8$, the last cell has fitted probability $\hat{\pi}_9 = 1 - \sum_{j=1}^{8} \hat{\pi}_j$. Three of the ten cells have an expected count less than 5, so we do not quite satisfy our rule of thumb for invoking normality. We therefore combine cells "8" and "≥ 9", and the rule of thumb is satisfied.

We calculate $X^2 = 4.229$ and $G^2 = 3.626$. Both are well below the critical value of $\chi^2_7(0.05) = 14.067$. Therefore the null hypothesis of H_0 : "the data follows a Poisson distribution" cannot be rejected. The Poisson distribution is a good fit for the data.

THE END