# Categorical Data Analysis
# Homework II

05/05/2021

1. A study on educational aspirations of high school students (S. Crysdale, *Int. J. Compar. Sociol.* **16**: 19-36, 1975) measured aspirations with the scale (some high school, high school graduate, some college, college graduate). The student counts in these categories were $(9, 44, 13, 10)$ when family income was low, $(11, 52, 23, 22)$ when family income was middle, and $(9, 41, 12, 27)$ when family income was high.

   (a) Test independence of educational aspirations and family income using $X^2$ or $G^2$. Explain the deficiency of this test for these data.

   (b) Find the standardized residuals. Do they suggest any association pattern?

   (c) Conduct an alternative test that may be more powerful. Interpret.

2. The table below shows the results of a retrospective study comparing radiation therapy with surgery in treating cancer of the larynx. The response indicates whether the cancer was controlled for at least two years following treatment. Some software output has been listed.

   (a) Report and interpret the $p$-value for Fisher's exact test with (i) $H_a : \theta > 1$ and (ii) $H_a : \theta \neq 1$. Explain how the $p$-values are calculated.

| | Cancer Controlled | Cancer Not Controlled |
|---|---|---|
| Surgery | 21 | 2 |
| Radiation therapy | 15 | 3 |

   (b) Interpret the confidence intervals for $\theta$. Explain the difference between them and how they were calculated.

   (c) Find and interpret the one-sided mid $p$-value. Give advantages and disadvantages of this type of $p$-value.

```
            Fisher's Exact Test
----------------------------------
Cell (1,1) Frequency (F)        21
Left-sided Pr <= F          0.8947
Right-sided Pr >= F         0.3808
Table Probability (P)       0.2755
Two-sided Pr<= P            0.6384
----------------------------------

Odds Ratio                  2.1000

Asymptotic Conf Limits:     95% Lower Conf Limit   0.3116
                            95% Upper Conf Limit  14.1523
Exact Conf Limits:          95% Lower Conf Limit   0.2089
                            95% Upper Conf Limit  27.5522
```

3. For independent uniform prior distributions for two binomial parameters $p_1$ and $p_2$, derive the form of prior density for $r = p_1/p_2$.

4. An experiment analyzes imperfection rates for two processes used to fabricate silicon wafers for computer chips. For treatment $A$ applied to 10 wafers, the numbers of imperfections are 8, 7, 6, 6, 3, 4, 7, 2, 3, 4. Treatment $B$ applied to 10 other wafers has 9, 9, 8, 14, 8, 13, 11, 5, 7, 6 imperfections. Treat the counts as independent Poisson variates with means $\mu_A$ and $\mu_B$.

   (a) Fit the model $\log \mu = \alpha + \beta x$, where $x = 1$ for treatment $B$ and $x = 0$ for treatment $A$. State the relationship between $\beta$ and $\mu_A$ and $\mu_B$, and interpret its estimate.

   (b) Test $H_0 : \mu_A = \mu_B$ using the Wald or likelihood-ratio test for $H_0 : \beta = 0$, Interpret.

   (c) Construct a 95% confidence interval for $\mu_A/\mu_B$.

   (d) Test $H_0 : \mu_A = \mu_B$ based on this result: If $Y_1$ and $Y_2$ are independent Poisson with means $\mu_1$ and $\mu_2$, then given $n = Y_1 + Y_2$, $Y_1$ is $\mathrm{Binom}(n, p_1)$ with $p_1 = \mu_1/(\mu_1 + \mu_2)$.

   (e) Is there evidence of overdispersion in the Poisson model? [Hint: Fit the model allowing overdispersion also.]

   (f) For the overall sample of 20 observations, the sample mean and variance are 7.0 and 10.2. Fit the loglinear model having only an intercept term under Poisson and negative binomial assumptions. Compare the results and confidence intervals for the overall mean response. Why do they differ?

5. The following table is based on a study with British doctors.

   (a) For each age, find the sample coronary death rate per 1,000 person-years for non-smokers and smokers. To compare them, take their ratio and describe its dependence on age.

(b) Fit a main-effect model for the log rates using age and smokers as factors. Discuss the goodness of fit.

(c) From (a), discuss why it is sensible to add a quantitative interaction of age and smoking. For the interaction model, how does the log ratio of coronary death rates change with age? Assign scores to age, fit the model and interpret.

| | Person-Years | | Coronary Deaths | |
|---|---|---|---|---|
| Age | Nonsmokers | Smokers | Nonsmokers | Smokers |
| 35–44 | 18,793 | 52,407 | 2 | 32 |
| 45–54 | 10,673 | 43,248 | 12 | 104 |
| 55–64 | 5710 | 28,612 | 28 | 206 |
| 65–74 | 2585 | 12,663 | 28 | 186 |
| 75–84 | 1462 | 5317 | 31 | 102 |

6. (a) For $n$ independent observations from a Poisson distribution, show that Fisher scoring gives $\mu^{(t+1)} = \bar{y}$ for all $t > 0$. By contrast, what happens with Newton-Raphson?

   (b) Write out the form of likelihood equations for Poisson loglinear models, i.e.,

$$\log \mu(\boldsymbol{x}_i) = \sum_{j=0}^{p} \beta_j x_{ij}, \text{ for } i = 1, \ldots, n.$$

.

7. For the horseshoe crab data attached to Homework II, fit a logistic regression model for the probability of a satellite, using colour alone as the predictor.

   (a) Treat colour as nominal. Explain why the model is saturated. Express its parameter estimates in terms of the sample logits for each colour.

   (b) Conduct a likelihood-ratio test that if colour has effect.

   (c) Fit a model that treats colour as quantitative. Interpret the fit and test that if colour has effect.

   (d) Test the goodness of the model in (c). Interpret.

8. Refer to the death penalty example. We fitted a logistic model by treating death penalty as the response (1=yes), defendant's race (1=white) and victims' race (1=white) as indicator predictors. The results are shown in the second table.

| Victims' Race | Defendant's Race | Death Penalty Yes | Death Penalty No | Percent Yes |
|---|---|---|---|---|
| White | White | 53 | 414 | 11.3 |
| | Black | 11 | 37 | 22.9 |
| Black | White | 0 | 16 | 0.0 |
| | Black | 4 | 139 | 2.8 |
| Total | White | 53 | 430 | 11.0 |
| | Black | 15 | 176 | 7.9 |

*Source:* M. L. Radelet and G. L. Pierce, *Florida Law Rev.* **43**: 1–34 (1991). Reprinted with permission from the *Florida Law Review.*

(a) Interpret parameter estimates. Which group is most likely to have the yes response? Find the estimated probability in that case.

(b) Interpret 95% confidence intervals for conditional odds ratios.

(c) Test the effect of defendant's race, controlling for victims' race, using a (i)Wald test and (ii) likelihood-ratio test. Interpret.

(d) Test the goodness of fit. Interpret.

```
      Criteria For Assessing Goodness Of Fit
   Criterion                DF        Value
   Deviance                  1          0.3798
   Pearson Chi-Square        1          0.1978
   Log Likelihood                    -209.4783
                        Standard    Likelihood Ratio     Chi-
Parameter   Estimate      Error     95% Conf Limits     Square
Intercept    -3.5961     0.5069    -4.7754   -2.7349    50.33
def          -0.8678     0.3671    -1.5633   -0.1140     5.59
vic           2.4044     0.6006     1.3068    3.7175    16.03

                     LR Statistics
        Source      DF     Chi-Square       Pr > ChiSq
        def          1           5.01         0.0251
        vic          1          20.35         <.0001
```