STOCKHOLM UNIVERSITY
DEPT OF MATHEMATICS
Div. of Mathematical statistics
Ola Hössjer

MT 5006
SOLUTIONS
January 12 2022

# Solutions for Examination Categorical Data Analysis, January 12, 2022

## Problem 1

a. Since a pre-determined number (=9) of persons were given the medicine, the row sums $n_{1+} = n_{2+} = 9$ are fixed. Therefore the most appropriate sampling scheme is independent binomial rows. We regard $(n_{11}, n_{21})$ as data, since they determine uniquely the number of observations in the other two cells. The success probabilities are $\pi_1$ and $\pi_2$ for the first and second rows respectively, so the likelihood is

$$\begin{aligned} l(\pi_1, \pi_2) &= P(N_{11} = n_{11}, N_{21} = n_{21} | \pi_1, \pi_2) \\ &= \binom{n_{1+}}{n_{11}} \pi_1^{n_{11}} (1-\pi_1)^{n_{1+}-n_{11}} \cdot \binom{n_{2+}}{n_{21}} \pi_2^{n_{21}} (1-\pi_2)^{n_{2+}-n_{21}} \\ &= \binom{9}{6} \pi_1^6 (1-\pi_1)^3 \cdot \binom{9}{3} \pi_2^3 (1-\pi_2)^6 \\ &= 7056 \pi_1^6 (1-\pi_1)^3 \pi_2^3 (1-\pi_2)^6. \end{aligned}$$

b. The null hypothesis is $H_0 : \pi_1 = \pi_2$.

c. Fisher's exact test conditions on fixed row and column sums, with a hypergeometric distribution

$$P_{H_0}(N_{11} = n_{11} | n_{1+}, n_{2+}, n_{+1}, n_{+2}) = \frac{\binom{n_{1+}}{n_{11}} \binom{n_{2+}}{n_{+1} - n_{11}}}{\binom{n}{n_{+1}}} = \frac{\binom{9}{n_{11}} \binom{9}{9 - n_{11}}}{\binom{18}{9}}.$$

d. The odds ratio is

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}. \quad (1)$$

With estimates $\hat{\pi}_i = n_{i1}/n_{i+}$ plugged into (1), for $i = 1, 2$, we find that

$$\hat{\theta} = \frac{\hat{\pi}_1(1-\hat{\pi}_2)}{\hat{\pi}_2(1-\hat{\pi}_1)} = \frac{n_{11} n_{22}}{n_{12} n_{21}} = \frac{6 \cdot 6}{3 \cdot 3} = 4.$$

e. A one-sided alternative

$$H_a : \pi_1 > \pi_2 \Longleftrightarrow H_a : \theta > 1,$$

corresponds to a positive effect of the medicine. Using the probabilities in the table, we find a

$$\begin{aligned} P-\text{value} &= P_{H_0}(N_{11} \geq 6 | n_{1+}, n_{2+}, n_{+1}, n_{+2}) \\ &= 0.1451 + 0.0267 + 0.0017 + 0.0000 \\ &= 0.1735, \end{aligned}$$

and conclude that $H_0$ cannot be rejected at level 5%.

# Problem 2

a. The expected cell counts $\mu_{ij} = n\pi_{ij}$ are fitted as

$$\hat{\mu}_{ij} = \frac{n_{i+}n_{+j}}{n}$$

under $H_0$, as summarized in the following table (see also Table 2 of Appendix B of the problem sheet):

| | $j$ | | |
|---|---|---|---|
| $i$ | 1 | 2 | 3 |
| 1 | 4.80 | 11.20 | 4.00 |
| 2 | 11.76 | 27.44 | 9.80 |
| 3 | 7.44 | 17.36 | 6.20 |

This gives a chisquare statistic

$$X^2 = \sum_{ij} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} = 5.062,$$

which has a $\chi^2_{\text{df}}$-distribution under $H_0$, with df $= (3-1)(3-1) = 4$. Since $\chi^2_4(0.05) = 9.488$ exceeds $X^2$, $H_0$ is not rejected at level 0.05 ($P$-value 0.281).

b. Since the marginal probabilities $\pi_{i+}$ and $\pi_{+j}$ are known under $H_0$, so are the cell probabilities $\pi_{ij} = \pi_{i+}\pi_{+j}$ and expected cell counts $\mu_{ij} = n\pi_{i+}\pi_{+j}$. We have that $\mu_{ij} = 25$ if $i = j = 2$, $\mu_{ij} = 12.5$ if just one of $i$ and $j$ equals 2 and $\mu_{ij} = 6.25$ if neither $i$ nor $j$ equals 2. The redefined chisquare statistic is

$$X^2 = \sum_{ij} \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}} = 9.88.$$

It has a $\chi^2_{\text{df}}$-distribution under $H_0$, with df $= 3 \times 3 - 1 = 8$. Since $\chi^2_8(0.05) = 15.51$ exceeds $X^2$, $H_0$ is not rejected at level 0.05 ($P$-value 0.274).

c. Merging rows 1 and 2, and also columns 1 and 2, we get a reduced $2 \times 2$ contingency table

| | $j$ | |
|---|---|---|
| $i$ | 1 | 2 |
| 1 | 52 | 17 |
| 2 | 28 | 3 |

of cell counts $\tilde{n}_{ij}$. Its odds ratio is estimated as

$$\hat{\theta} = \frac{\tilde{n}_{11}\tilde{n}_{22}}{\tilde{n}_{12}\tilde{n}_{21}} = \frac{52 \cdot 3}{17 \cdot 28} = 0.328.$$

d. The log odds is estimated as $\log(\hat{\theta}) = -1.1156$, with a standard error

$$\mathrm{SE} = \sqrt{\frac{1}{\tilde{n}_{11}} + \frac{1}{\tilde{n}_{12}} + \frac{1}{\tilde{n}_{21}} + \frac{1}{\tilde{n}_{22}}} = \sqrt{\frac{1}{52} + \frac{1}{17} + \frac{1}{28} + \frac{1}{3}} = 0.6687.$$

This gives a 95% confidence interval

$$(-1.1156 \pm 1.96 \cdot 0.6687) = (-2.426, 0.195)$$

for $\log(\theta)$, and

$$(e^{-2.426}, e^{0.195}) = (0.0884, 1.2153)$$

for $\theta$. The interval is wide because of the small number of observations in cell (2,2), and $H_0$ (that rows and columns of the $2 \times 2$ table are independent, i.e. $\theta = 1$) cannot be rejected. For the same reason, the interval is also quite inaccurate, since it relies on $\log(\hat{\theta})$ having a distribution close to normal, which is an asymptotic result for large cell counts.

# Problem 3

a. Model $M_1 = (\mathrm{EGS})$ has Poisson distributed cell counts

$$N_{egs} \sim \mathrm{Po}\left(\exp(\lambda + \lambda_e^E + \lambda_g^G + \lambda_s^S + \lambda_{eg}^{EG} + \lambda_{es}^{ES} + \lambda_{gs}^{GS} + \lambda_{egs}^{EGS})\right),$$

with $e \in \{1,2,3\}$ and $g, s \in \{1,2\}$. If the highest level of each variable is used as baseline, any parameter with at least one of its indeces $e$, $g$ or $s$ equal to the highest level is put to zero. This gives 12 parameters, included in the vector

$$(\lambda, \lambda_1^E, \lambda_2^E, \lambda_1^S, \lambda_1^G, \lambda_{11}^{EG}, \lambda_{21}^{EG}, \lambda_{11}^{ES}, \lambda_{21}^{ES}, \lambda_{11}^{GS}, \lambda_{111}^{EGS}, \lambda_{211}^{EGS}). \tag{2}$$

b. Model $M_0$ is obtained from (2) by removing all interaction parameters that involve $S$. The remaining 7 parameters are included in the vector

$$(\lambda, \lambda_1^E, \lambda_2^E, \lambda_1^S, \lambda_1^G, \lambda_{11}^{EG}, \lambda_{21}^{EG}). \tag{3}$$

c. Since $E$ and $G$ are jointly independent of $S$ under $M_0$, it follows that the expected cell counts are

$$\mu_{egs} = \mu_{+++}\pi_{egs} = \mu_{+++}\pi_{eg+}\pi_{++s} = \mu_{+++}\frac{\mu_{eg+}}{\mu_{+++}}\frac{\mu_{++s}}{\mu_{+++}} = \frac{\mu_{eg+}\mu_{++s}}{\mu_{+++}}. \tag{4}$$

d. The fitted cell counts for model $M_0$ are obtained by plugging $\hat{\mu}_{eg+} = n_{eg+}$, $\hat{\mu}_{++s} = n_{++s}$ and $\hat{\mu}_{+++} = n_{+++} = n$ into (4). This yields

$$\hat{\mu}_{egs} = \frac{n_{eg+} n_{++s}}{n}, \tag{5}$$

where $n_{++1} = 128$, $n_{++2} = 93$ and $n = n_{++1} + n_{++2} = 221$. By adding the tables for the two schools we find that $n_{11+} = 25$, $n_{12+} = 9$, $n_{21+} = 57$, $n_{22+} = 59$, $n_{31+} = 27$ and $n_{32+} = 44$. Insertion into (5) gives (see also Table 3 of Appendix B of the problem sheet)

$\hat{\mu}_{eg1}$:

| | $g$ | |
|---|---|---|
| $e$ | 1 | 2 |
| 1 | 14.48 | 5.21 |
| 2 | 33.01 | 34.17 |
| 3 | 15.64 | 25.48 |

$\hat{\mu}_{eg2}$:

| | $g$ | |
|---|---|---|
| $e$ | 1 | 2 |
| 1 | 10.52 | 3.79 |
| 2 | 23.99 | 24.83 |
| 3 | 11.36 | 18.52 |

In order to test

$$\begin{aligned} H_0 &: \ M_0 \text{ holds}, \\ H_a &: \ M_1 \text{ holds but not } M_0, \end{aligned}$$

we use the likelihood ratio statistic

$$\begin{aligned} G^2(M_0|M_1) &= 2\sum_{e,g,s} n_{egi} \log(n_{egi}/\hat{\mu}_{egi}) \\ &= 2\,(15 \cdot \log(15/14.48) + \ldots + 19 \cdot \log(19/18.52)) \\ &= 1.6125 \\ &< \chi^2_{12-7}(0.05) = 11.07. \end{aligned}$$

Since $H_0$ is not rejected, there is no significant difference between the two schools at level 0.05.

# Problem 4

a. The parameters of $M_0$ are listed in (3), and therefore the logistic regression model satisfies

$$\begin{aligned} &\text{logit}\,(P(G = 2|E = e, S = s)) \\ &\quad = \log\,(P(G = 2|E = e, S = s)) - \log\,(P(G = 1|E = e, S = s)) \\ &\quad = \log\,(P(E = e, G = 2, S = s)) - \log\,(P(E = e, G = 1, S = s)) \\ &\quad = \log(\pi_{e2s}) - \log(\pi_{e1s}) \\ &\quad = \log(\mu_{e2s}) - \log(\mu_{e1s}) \\ &\quad = \left(\lambda + \lambda^E_e + \lambda^G_2 + \lambda^S_s + \lambda^{EG}_{e2}\right) - \left(\lambda + \lambda^E_e + \lambda^G_1 + \lambda^S_s + \lambda^{EG}_{e1}\right) \\ &\quad = (\lambda^G_2 - \lambda^G_1) + (\lambda^{EG}_{e2} - \lambda^{EG}_{e1}) \\ &\quad = \alpha + \beta_e, \end{aligned} \tag{6}$$

where in the fourth step we made use of $\pi_{egs} = \mu_{egs}/\mu_{+++}$ and in the last step we introduced $\alpha = \lambda^G_2 - \lambda^G_1$ and $\beta_e = \beta^E_e = \lambda^{EG}_{e2} - \lambda^{EG}_{e1}$ for $e = 1, 2, 3$. Since $\lambda^{EG}_{31} = \lambda^{EG}_{32} = 0$, it follows that $\beta_3 = 0$, so there are only three parameters $(\alpha, \beta_1, \beta_2)$.

Model (6) is an ANOVA type logistic regression model for an outcome variable $G$ and two categorical predictor variables $E$ and $S$, of which the second has no effect.

b. It follows from (6) that

$$\theta = e^{\beta_1} = \frac{e^{\alpha+\beta_1}}{e^{\alpha}} = \frac{P(G=2|E=1,S=s)/P(G=1|E=1,S=s)}{P(G=2|E=3,S=s)/P(G=1|E=3,S=s)}$$

is the odds ratio for a student from a low income family to have high grades relative to one from a high income family (regardless of school). The estimated odds ratio is $\hat{\theta} = \exp(-1.51) = 0.221$. As a remark we notice that this conforms with Problem 3, since

$$\hat{\theta} = \frac{n_{12+}n_{31+}}{n_{11+}n_{32+}} = \frac{9 \cdot 27}{25 \cdot 44} = 0.221.$$

c. A 95% Wald type confidence interval is

$$\left(\hat{\beta}_1 \pm 1.96\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_1)}\right) = \left(-1.51 \pm 1.96 \cdot \sqrt{0.2109}\right) = (-2.4101, -0.6099)$$

for $\beta_1$ and

$$(\exp(-2.4101), \exp(-0.6099)) = (0.0898, 0.5434)$$

for $\theta$.

d. The saturated logistic regression model satisfies

$$\mathrm{logit}\,(P(G=2|E=e,S=s)) = \alpha + \beta_e^E + \beta_s^S + \beta_{es}^{ES}$$

for $e \in \{1,2,3\}$ and $s \in \{1,2\}$. All effect parameters are put zero when at least one index satisfies $e = 3$ or $s = 2$. This gives a parameter vector $(\alpha, \beta_1^E, \beta_2^E, \beta_1^S, \beta_{11}^{ES}, \beta_{21}^{ES})$ with $3 \times 2 = 6$ parameters, equal to the number of possible combinations $(e,s)$ of the two predictor variables. Therefore the ANOVA type model (6) has

$$\mathrm{df} = 6 - 3 = 3.$$

# Problem 5

a. We find that

$$\log f(y;\theta,\phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi) \Longrightarrow u(y) = \frac{y - b'(\theta)}{a(\phi)},$$

so that

$$0 = E(u(Y)) = \frac{E(Y) - b'(\theta)}{a(\phi)} \Longrightarrow \mu = E(Y) = b'(\theta),$$

and hence

$$u(y) = \frac{y - \mu}{a(\phi)}.$$

b. Suppose $y \in \{0, \phi, 2\phi, \ldots\}$. The probability function of the ODP distribution is

$$\begin{aligned} f(Y) &= P(Y=y) = P(Y/\phi = y/\phi) = e^{-\mu/\phi} \frac{(\mu/\phi)^{y/\phi}}{(y/\phi)!} \\ &= \exp\left( \frac{y\log(\mu)-\mu}{\phi} - (y/\phi)\log(\phi) - \log((y/\phi)!) \right). \end{aligned}$$

It belongs to the exponential dispersion family, with natural parameter $\theta = \log(\mu)$, $a(\phi) = \phi$, $b(\theta) = \mu = e^{\theta}$ and $c(y, \phi) = -(y/\phi)\log(\phi) - \log((y/\phi)!)$.

c. Since $\theta_i = \log(\mu_i)$ is the natural parameter, $\log(\mu_i)$ is the canonical link function and

$$\log(\mu_i) = \sum_{j=1}^{p} x_{ij}\beta_j = \boldsymbol{x}_i\boldsymbol{\beta},$$

so that

$$Y_i/\phi \sim \text{Po}(\mu_i/\phi) = \text{Po}(\exp(\boldsymbol{x}_i\boldsymbol{\beta})/\phi).$$

d. Let

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log f(y_i; x_i, \boldsymbol{\beta})$$

be the log likelihood. It then follows that

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \log f(y_i)}{\partial \beta_j} \overset{\text{hint}}{=} \sum_{i=1}^{n} x_{ij} \frac{\partial \log f(y_i)}{\partial \theta_i} \overset{a)}{=} \sum_{i=1}^{n} \frac{x_{ij}(y_i - \mu_i)}{a(\phi)} \overset{b)}{=} \sum_{i=1}^{n} \frac{x_{ij}(y_i-\mu_i)}{\phi}. \tag{7}$$

Since

$$\text{Var}(Y_i) = \phi^2 \text{Var}(Y_i/\phi) = \phi^2 \cdot \mu_i/\phi = \phi\mu_i,$$

and all $Y_i$ are independent, this gives

$$\begin{aligned} J_{jk} &= \text{Cov}\left( \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j}, \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_k} \right) \\ &= \text{Cov}\left( \textstyle\sum_{i=1}^{n} \frac{x_{ij}(Y_i-\mu_i)}{\phi}, \sum_{i=1}^{n} \frac{x_{ik}(Y_i-\mu_i)}{\phi} \right) \\ &= \textstyle\sum_{i=1}^{n} \text{Cov}\left( \frac{x_{ij}(Y_i-\mu_i)}{\phi}, \frac{x_{ik}(Y_i-\mu_i)}{\phi} \right) \\ &= \textstyle\sum_{i=1}^{n} x_{ij}x_{ik}\text{Var}(Y_i)/\phi^2 \\ &= \textstyle\sum_{i=1}^{n} x_{ij}x_{ik}\mu_i/\phi. \end{aligned} \tag{8}$$

Alternatively one may differentiate (7) with respect to $\beta_k$ and use that

$$J_{jk} = -E\left( \frac{\partial L^2(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} \right) = -\frac{\partial L^2(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} = -\sum_{i=1}^{n} \frac{\partial^2 \log f(y_i)}{\partial \beta_j \partial \beta_k} = \sum_{i=1}^{n} \frac{x_{ij}\partial \mu_i/\partial \beta_k}{\phi},$$

where the second equality holds for a generalized linear model with a canonical link function. Together with $\partial \mu_i/\partial \beta_k = x_{ik}\mu_i$, this implies (8).