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

MT 5006
SOLUTIONS
January 10 2020

Solutions for Examination
Categorical Data Analysis, January 10, 2020

Problem 1

a. Since a fixed number (=9) of games without a draw are played, 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 π_1 and π_2 for the first and second rows respectively, so the likelihood is

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

The null hypothesis is $H_0 : \pi_1 = \pi_2 = \pi$. This gives a likelihood

$$l(\pi, \pi) = 7056 \cdot \pi^9(1-\pi)^9$$

under H_0.

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

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

In the sequel, for ease of notation we will write $P(i) = P(N_{11} = i|n_{1+}, n_{2+}, n_{+1}, n_{+2})$.

c. A one-sided alternative

$$H_a : \pi_1 > \pi_2$$

corresponds to Mary being a more skilled chess player. Using the probabilities in the table, we find a

$$\begin{aligned} P-\text{value} &= P_{H_0}(N_{11} \geq 6|n_{1+}, n_{2+}, n_{+1}, n_{+2}) \\ &= P(6) + P(7) + P(8) + P(9) \\ &= 0.1451 + 0.0267 + 0.0017 + 0.0000 \\ &= 0.1735, \end{aligned}$$

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

d. A two-sided alternative
$$H_a : \pi_1 \neq \pi_2$$
corresponds to one of the players being more skilled than the other. Because of symmetry in (1) (and from the displayed table) we notice that $P(i) = P(9-i)$. Since this probability is a decreasing function $i = 5, 6, 7, 8, 9$ we find that the two-sided mid P-value is

$$\begin{aligned} P - \text{value} &= 0.5 \textstyle\sum_{i;P(i)=P(n_{11})} P(i) + \sum_{i;P(i)<P(n_{11})} P(i) \\ &= 0.5[P(3)+P(6)] + P(0) + P(1) + P(2) + P(7) + P(8) + P(9) \\ &= P(6) + 2(P(7) + P(8) + P(9)) \\ &= 0.1451 + 2(0.0267 + 0.0017 + 0.0000) \\ &= 0.2019. \end{aligned}$$

Problem 2

a. Let n_{ij} be the number of observations in cell (i, j), which is an observation of the random variable N_{ij}. The joint distribution of all cell counts is multinomial
$$\boldsymbol{N} = (N_{ij})_{i,j=1}^{3} \sim \text{Mult}(90, (\pi_{ij})_{i,j=1}^{3}).$$
Since the cell probabilities sum to 1 ($\sum_{i,j=1}^{3} \pi_{ij} = 1$), there are 8 free parameters, for instance
$$\boldsymbol{\theta} = (\pi_{11}, \pi_{12}, \pi_{13}, \pi_{21}, \pi_{22}, \pi_{23}, \pi_{31}, \pi_{32}).$$
This gives a likelihood
$$\begin{aligned} l(\boldsymbol{\theta}) &= \tfrac{90!}{\prod_{i,j=1}^{3} n_{ij}!} \textstyle\prod_{(i,j)\neq(3,3)} \pi_{ij}^{n_{ij}} \cdot (1 - \sum_{(i,j)\neq(3,3)} \pi_{ij})^{n_{33}} \\ &= \tfrac{500!}{4!8!14!6!12!15!5!10!16!} \pi_{11}^{4}\pi_{12}^{8}\pi_{13}^{14}\pi_{21}^{6}\pi_{22}^{12}\pi_{23}^{15}\pi_{31}^{5}\pi_{32}^{10}(1 - \textstyle\sum_{(i,j)\neq(2,2)} \pi_{ij})^{16}. \end{aligned}$$

b. The expected cell counts under H_0 are
$$\mu_{ij} = E(N_{ij}) = n_{++}\pi_{i+}\pi_{+j} = 90 \cdot \frac{1}{3} \cdot \frac{1}{3} = 10.$$
This gives a X^2-statistic
$$X^2 = \sum_{i,j=1}^{3} \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}} = \frac{1}{10} \sum_{i,j=1}^{3} (n_{ij} - 10)^2 = \frac{1}{10} \left[(4-10)^2 + \ldots (16-10)^2 \right] = 16.2.$$
Since the saturated model has 8 parameters and H_0 no freely variable parameter, the number of degrees of freedom is $8 - 0 = 8$. Therefore, since $X^2 > \chi_8^2(0.05) = 15.5$, we confirm Ben's suspicion that the claimed properties of the lottery are wrong, by rejecting H_0 at level 5%.

c. The estimated expected cell counts under H_0' equal
$$\hat{\mu}_{ij} = n_{++}\pi_{i+}\hat{\pi}_{+j} = n_{++} \cdot \frac{1}{3} \cdot \frac{n_{+j}}{n_{++}} = \frac{n_{+j}}{3} = \begin{cases} 5, & j = 1, \\ 10, & j = 2, \\ 15, & j = 3. \end{cases}$$

This gives a X^2-statistic

$$X^2 = \sum_{i,j=1}^{3} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} = \frac{(4-5)^2}{5} + \ldots \frac{(16-15)^2}{15} = 1.33.$$

Since H_0' has 2 freely variable parameters (π_{+1} and π_{+2} for instance, since $\pi_{+3} = 1 - \pi_{+1} - \pi_{+2}$), the test has 8-2=6 degrees of freedom. Since $X^2 < \chi_6^2(0.05) = 12.59$, we don't reject Ben's suggested model H_0' for the lottery.

Problem 3

a. Model $M_1 =$ (EG,ES) has Poisson distributed cell counts

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

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 indices e, g or s equal to the highest level is put to zero. This gives 9 parameters, included in the vector

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

Model $M_0 =$ (EG,S) is obtained from (2) by removing the two interaction parameters between E and S. The remaining 7 parameters are included in the vector

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

b. Write the expected cell counts as $\mu_{egs} = \mu_{+++}\pi_{egs}$. Since E, G and S are jointly independent under M_0, we have that $\pi_{egs} = \pi_{eg+}\pi_{++s}$. The fitted values of μ_{egs} for model $M_0 =$ (EG,S) are therefore

$$\hat{\mu}_{egs}^{(0)} = \hat{\mu}_{+++}\hat{\pi}_{eg+}\hat{\pi}_{++s} = n \cdot \frac{n_{eg+}}{n} \cdot \frac{n_{++s}}{n} = \frac{n_{eg+}n_{++s}}{n}, \tag{4}$$

where n_{++s} are total number of students in each school ($n_{++1} = 128$, $n_{++2} = 93$) and $n = n_{++1} + n_{++2} = 221$ the total number of students in both schools. By adding the tables for the two schools we obtain all n_{eg+} ($n_{11+} = 25$, $n_{12+} = 9$, $n_{21+} = 57$, $n_{22+} = 59$, $n_{31+} = 27$ and $n_{32+} = 44$). Insertion into (4) gives the values of $\hat{\mu}_{egs}^{(0)}$ in the upper table of Appendix A, for instance

$$\hat{\mu}_{111}^{(0)} = \frac{n_{11+}n_{++1}}{n} = \frac{25 \cdot 128}{221} = 14.48.$$

For model $M_1 =$ (EG,ES) we have that $\pi_{egs} = \pi_{e++}\pi_{g|e}\pi_{s|e}$. Since $\pi_{g|e} = \pi_{eg+}/\pi_{e++}$ and $\pi_{s|e} = \pi_{e+s}/\pi_{e++}$, we find that $\pi_{egs} = \pi_{eg+}\pi_{e+s}/\pi_{e++}$. Consequently,

$$\hat{\mu}_{egs}^{(1)} = n \cdot \frac{\hat{\pi}_{eg+}\hat{\pi}_{e+s}}{\hat{\pi}_{e++}} = n \cdot \frac{(n_{eg+}/n) \cdot (n_{e+s}/n)}{n_{e++}/n} = \frac{n_{eg+}n_{e+s}}{n_{e++}}, \tag{5}$$

with n_{eg+} as in (4), whereas the values of n_{e+s} ($n_{1+1} = 21$, $n_{2+1} = 68$, $n_{3+1} = 39$, $n_{1+2} = 13$, $n_{2+2} = 48$, $n_{3+2} = 32$) are obtained from the row sums of the two schools. By adding the two row sums from the two schools, for each economy level e, we end up with all n_{e++} ($n_{1++} = 34$, $n_{2++} = 116$, $n_{3++} = 71$). Insertion of these numbers into (5) gives the values of the lower table of Appendix A, for instance

$$\hat{\mu}^{(1)}_{111} = \frac{n_{11+}n_{1+1}}{n_{1++}} = \frac{25 \cdot 21}{34} = 15.44.$$

c. 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

$$\begin{aligned} G^2(M_0|M_1) &= G^2(M_0) - G^2(M_1) \\ &= 2\textstyle\sum_{e,g,s} n_{egs} \log(n_{egs}/\hat{\mu}^{(0)}_{egs}) \\ &- 2\textstyle\sum_{e,g,s} n_{egs} \log(n_{egs}/\hat{\mu}^{(1)}_{egs}) \\ &= 2\textstyle\sum_{e,g,s} n_{egs} \log(\hat{\mu}^{(1)}_{egs}/\hat{\mu}^{(0)}_{egs}) \\ &= 2\,(15 \cdot \log(15.44/14.48) + \ldots + 19 \cdot \log(19.83/18.52)) \\ &= 0.4906 \\ &< \chi^2_{9-7}(0.05) = 5.99. \end{aligned}$$

Since H_0 is not rejected, there is no significant difference between the economy levels of 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

$$\begin{aligned} &\operatorname{logit}(P(G=2|E=e,S=s)) \\ &\quad = \log(P(G=2|E=e,S=s)) - \log(P(G=1|E=e,S=s)) \\ &\quad = \log(P(E=e,G=2,S=s)) - \log(P(E=e,G=1,S=s)) \\ &\quad = \log \pi_{e2s} - \log \pi_{e1s} \\ &\quad = \log \mu_{e2s} - \log \mu_{e1s} \\ &\quad = \left(\lambda + \lambda^E_e + \lambda^G_2 + \lambda^S_s + \lambda^{EG}_{e2}\right) - \left(\lambda + \lambda^E_e + \lambda^G_1 + \lambda^S_s + \lambda^{EG}_{e1}\right) \\ &\quad = (\lambda^G_2 - \lambda^G_1) + (\lambda^{EG}_{e2} - \lambda^{EG}_{e1}) \\ &\quad = \beta_0 + \beta_e, \end{aligned} \tag{6}$$

where $\beta_0 = \lambda^G_2 - \lambda^G_1 = -\lambda^G_1$ and $\beta_e = \beta^E_e = \lambda^{EG}_{e2} - \lambda^{EG}_{e1} = -\lambda^{EG}_{e1}$ for $e = 1, 2, 3$. Since $\lambda^{EG}_{31} = 0$ it follows that $\beta_3 = 0$, so there are only three parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$.

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

$$\theta_1 = e^{\beta_1} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = \frac{P(G=2|E=1,S=s)/P(G=1|E=1,S=s)}{P(G=2|E=3,S=s)/P(G=1|E=3,S=s)}$$

is a conditional odds ratio, i.e. the odds of a student from a low income family to have high grades relative to the corresponding odds of a student from a high income family (regardless of school, i.e. homogeneous association). The corresponding estimated conditional odds ratio is $\hat{\theta}_1 = e^{-1.51} = 0.221$. Similarly, one finds that $\hat{\theta}_2 = e^{-0.4539} = 0.635$ is the estimated odds for a student from a middle income family to have high grades relative to one from a high income family (regardless of school, i.e. homogeneous association), whereas $\hat{\theta}_3 = e^{-1.51-(-0.4539)} = 0.348$ is the estimated odds for a student from a low income family to have high grades relative to one from a middle income family (regardless of school, i.e. homogeneous association). As a remark we notice that all these three estimated conditional odds ratios agree with the corresponding estimated marginal odds ratios, for instance

$$\hat{\theta}_1 = \frac{n_{12+}n_{31+}}{n_{11+}n_{32+}} = \frac{9 \cdot 27}{25 \cdot 44} = 0.221.$$

c. Since the probability $P(G = g|E = e, S = s) = P(G = g|E = e)$ of grade g does not depend on school s, the likelihood of the logistic regression model is

$$\begin{aligned} l(\boldsymbol{\beta}) &= \textstyle\prod_{e,s}[P(G=2|E=e)^{n_{e2s}} P(G=1|E=e)^{n_{e1s}}] \\ &= \textstyle\prod_{e=1}^{3}[P(G=2|E=e)^{n_{e2+}} P(G=1|E=e)^{n_{e1+}}] \\ &= \textstyle\prod_{e=1}^{3}[(e^{\beta_0+\beta_e}/(1+e^{\beta_0+\beta_e}))^{n_{e2+}} \cdot (1/(1+e^{\beta_0+\beta_e}))^{n_{e1+}}] \\ &= \textstyle\prod_{e=1}^{3}[e^{n_{e2+}(\beta_0+\beta_e)}/(1+e^{\beta_0+\beta_e})^{n_{e++}}], \end{aligned}$$

where $\beta_3 = 0$ according to a). This gives a log likelihood function

$$\begin{aligned} L(\boldsymbol{\beta}) &= \log l(\boldsymbol{\beta}) \\ &= \textstyle\sum_{e=1}^{3}[n_{e2+}(\beta_0+\beta_e) - n_{e++}\log(1+e^{\beta_0+\beta_e})] \\ &= n_{12+}(\beta_0+\beta_1) - n_{1++}\log(1+e^{\beta_0+\beta_1}) \\ &+ n_{22+}(\beta_0+\beta_2) - n_{2++}\log(1+e^{\beta_0+\beta_2}) \\ &+ n_{32+}\beta_0 - n_{3++}\log(1+e^{\beta_0}). \end{aligned}$$

Write the score vector as $\boldsymbol{u}(\boldsymbol{\beta}) = (u_0(\boldsymbol{\beta}), u_1(\boldsymbol{\beta}), u_2(\boldsymbol{\beta}))$, where $u_j(\boldsymbol{\beta}) = \partial L(\boldsymbol{\beta})/\partial \beta_j$. The first component equals

$$\begin{aligned} u_0(\boldsymbol{\beta}) &= n_{12+} - n_{1++}e^{\beta_0+\beta_1}/(1+e^{\beta_0+\beta_1}) \\ &+ n_{22+} - n_{2++}e^{\beta_0+\beta_2}/(1+e^{\beta_0+\beta_2}) \\ &+ n_{32+} - n_{3++}e^{\beta_0}/(1+e^{\beta_0}), \end{aligned}$$

whereas the other two components are

$$\begin{aligned} u_1(\boldsymbol{\beta}) &= n_{12+} - n_{1++}e^{\beta_0+\beta_1}/(1+e^{\beta_0+\beta_1}), \\ u_2(\boldsymbol{\beta}) &= n_{22+} - n_{2++}e^{\beta_0+\beta_2}/(1+e^{\beta_0+\beta_2}). \end{aligned}$$

We find that

$$\begin{aligned} u_1(\hat{\boldsymbol{\beta}}) &= n_{12+} - n_{1++}e^{\hat{\beta}_0+\hat{\beta}_1}/(1+e^{\hat{\beta}_0+\hat{\beta}_1}) \\ &= 9 - 34 \cdot e^{0.4884-1.5100}/(1+e^{0.4884-1.5100}) \\ &= 0, \end{aligned}$$

and similarly $u_0(\hat{\boldsymbol{\beta}}) = u_2(\hat{\boldsymbol{\beta}}) = 0$.

Problem 5

a. The ratio of the oddses of the second wheel having a high outcome ($Y = 2$), when the first wheel has a high ($X = 2$) and low ($X = 1$) outcome respectively, is

$$\begin{aligned} \theta &= \tfrac{P(Y=2|X=2)/P(Y=1|X=2)}{P(Y=2|X=1)/P(Y=1|X=1)} \\ &= \tfrac{(p_{22}/p_{2+})/(p_{21}/p_{2+})}{(p_{12}/p_{1+})/(p_{11}/p_{1+})} \\ &= (p_{11}p_{22})/(p_{12}p_{21}). \end{aligned} \tag{7}$$

b. Let $\hat{p}_{ij} = n_{ij}/n$, and define

$$\begin{aligned} \hat{\theta} &= (\hat{p}_{11}\hat{p}_{22})/(\hat{p}_{12}\hat{p}_{21}) \\ &= (n_{11}n_{22})/(n_{12}n_{21}) \\ &= (4 \cdot 12)/(8 \cdot 6) \\ &= 1 \end{aligned} \tag{8}$$

be our estimator of θ, obtained by replacing all p_{ij} with $\hat{p}_{ij}$ in (7). Then, by a first order Taylor expansion

$$\begin{aligned} \log(\hat{\theta}) - \log(\theta) &= (\log(\hat{p}_{11}) - \log(p_{11})) + (\log(\hat{p}_{22}) - \log(p_{22})) \\ &- (\log(\hat{p}_{12}) - \log(p_{12})) - (\log(\hat{p}_{21}) - \log(p_{21})) \\ &\approx \hat{p}_{11}/p_{11} + \hat{p}_{22}/p_{22} - \hat{p}_{12}/p_{12} - \hat{p}_{21}/p_{21} \\ &= \textstyle\sum_{i,j}(-1)^{i+j}\hat{p}_{ij}/p_{ij}. \end{aligned} \tag{9}$$

The cell counts n_{ij} are observations of

$$(N_{11}, N_{12}, N_{21}, N_{22}) \sim \text{Mult}(n; p_{11}, p_{12}, p_{21}, p_{22}).$$

From this it follows that

$$\text{Cov}(\hat{p}_{ij}, \hat{p}_{kl}) = \frac{\text{Cov}(N_{ij}, N_{kl})}{n^2} = \begin{cases} p_{ij}(1 - p_{ij})/n, & (i,j) = (k,l), \\ -p_{ij}p_{kl}/n, & (i,j) \neq (k,l). \end{cases} \tag{10}$$

Combining (9) and (10), we find that

$$\begin{aligned} \text{Var}(\log(\hat{\theta})) &\approx \text{Var}(\textstyle\sum_{i,j}(-1)^{i+j}\hat{p}_{ij}/p_{ij}) \\ &= \textstyle\sum_{i,j,k,l}(-1)^{i+j+k+l}\text{Cov}(\hat{p}_{ij}, \hat{p}_{kl})/(p_{ij}p_{kl}) \\ &= \textstyle\sum_{i,j}(-1)^{2(i+j)}p_{ij}/(np_{ij}^2) \\ &- \textstyle\sum_{i,j,k,l}(-1)^{i+j+k+l}p_{ij}p_{kl}/(np_{ij}p_{kl}) \\ &= \textstyle\sum_{i,j}/(np_{ij}) - \left(\textstyle\sum_{i,j}(-1)^{i+j}\right)^2/n \\ &= 1/(np_{11}) + 1/(np_{12}) + 1/(np_{21}) + 1/(np_{22}). \end{aligned} \tag{11}$$

c. We start by estimating the variance of $\log(\hat{\theta})$, replacing p_{ij} by estimates $\hat{p}_{ij} = n_{ij}/n$ in (11). This gives

$$\begin{aligned} \widehat{\text{Var}}(\log(\hat{\theta})) &= 1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22} \\ &= 1/4 + 1/8 + 1/6 + 1/12 \\ &= 0.6250. \end{aligned}$$

Together with the point estimate of θ in (8), we obtain an approximate 95% Wald type confidence interval

$$(\log(\hat{\theta}) - 1.96\sqrt{0.6250}, \log(\hat{\theta}) + 1.96\sqrt{0.6250}) = (-1.5495, 1.5495) =: (a, b)$$

for $\log(\theta)$, with $1.96 = \sqrt{\chi^2_{0.05}(1)} = \sqrt{3.84}$ the 0.975-quantile of a standard normal distribution. Since the logarithmic function is monotone increasing, we take the inverse of this function in order to find a confidence interval

$$(\exp(a), \exp(b)) = (\exp(-1.5495), \exp(1.5945)) = (0.212, 4.709)$$

for θ, with approximate coverage probability 95%. Since 1 is included in this interval we cannot reject independence between X and Y at level 5%.