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

MT 5019
SOLUTIONS
January 9 2025

# Solutions for Examination
# Categorical Data Analysis, January 9, 2025

## Problem 1

a. The joint distribution of the cell counts is multinomial

$$(N_{00}, N_{01}, N_{10}, N_{11}) \sim \text{Mult}(15; \pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}),$$

where $\pi_{ij} = P(X = i, Y = j)$ for each new game without a draw.

b. Introduce the conditional probabilities

$$\pi_{j|i} = P(Y = j|X = i) = \frac{\pi_{ij}}{\pi_{i+}}$$

of the outcome of a game, given that the expert has guessed player $i$ as a winner, for $i = 0, 1$. The null hypothesis can be formulated as

$$H_0 : \pi_{j|0} = \pi_{j|1} \text{ for } j = 0, 1 \Longleftrightarrow \theta = 1 \Longleftrightarrow \pi_{ij} = \pi_{i+}\pi_{+j} \text{ for } 0 \le i, j \le 1,$$

where

$$\theta = \frac{\pi_{1|1}/(1 - \pi_{1|1})}{\pi_{1|0}/(1 - \pi_{1|0})} = \frac{\pi_{0|0}\pi_{1|1}}{\pi_{0|1}\pi_{1|0}} = \frac{\pi_{00}\pi_{11}}{\pi_{01}\pi_{10}} \tag{1}$$

is the odds ratio. The alternative hypothesis

$$H_a : \pi_{1|1} > \pi_{1|0} \Longleftrightarrow \frac{\pi_{1|1}}{1 - \pi_{1|1}} > \frac{\pi_{1|0}}{1 - \pi_{1|0}} \Longleftrightarrow \theta > 1$$

is one-sided.

c. Let $n_{ij}$ be the observed cell counts. If we condition on the two row sums $n_{i+}$ and the two column sums $n_{+j}$, then $N_{11}$ has a hypergeometric distribution under the null hypothesis, i.e.

$$\begin{aligned} &P(N_{11} = k|H_0, N_{0+} = 7, N_{1+} = 8, N_{+0} = 6, N_{+1} = 9) \\ &= P(N_{11} = k|H_0, N_{1+} = 8, N_{+1} = 9) \\ &= \tbinom{7}{9-k}\tbinom{8}{k}/\tbinom{15}{9} \end{aligned}$$

for $2 \le k \le 8$. Notice that we only need to include one row sum (say $n_{1+}$) and one column sum (say $n_{+1}$) in the conditioning, since $n_{++} = 15$ is fixed.

The null hypothesis is rejected for large values of $N_{11}$, since these are more likely to occur when the alternative hypothesis is true. Since $n_{11} = 6$, this gives a

$$\begin{aligned}
\text{mid } P\text{-value} &= 0.5P(N_{11} = 6|H_0, N_{1+} = 8, N_{+1} = 9) \\
&+ P(N_{11} = 7|H_0, N_{1+} = 8, N_{+1} = 9) \\
&+ P(N_{11} = 8|H_0, N_{1+} = 8, N_{+1} = 9) \\
&= 0.5 \cdot \tbinom{7}{3}\tbinom{8}{6}/\tbinom{15}{9} + \tbinom{7}{2}\tbinom{8}{7}/\tbinom{15}{9} + \tbinom{7}{1}\tbinom{8}{8}/\tbinom{15}{9} \\
&= (0.5 \cdot 35 \cdot 28 + 21 \cdot 8 + 7 \cdot 1)/\tbinom{15}{9} \\
&= 665/5005 \\
&= 0.1329.
\end{aligned}$$

Hence we cannot reject the null hypothesis at nominal level 5% if we use mid $P$-value $\leq$ 0.05 as criterion for rejecting $H_0$.

d. When one only conditions on row sums, we have that

$$\begin{aligned}
N_{01} &\sim \text{Bin}(7, \pi_{1|0}), \\
N_{11} &\sim \text{Bin}(8, \pi_{1|1})
\end{aligned}$$

are independent and binomially distributed. Therefore the joint distribution of $N_{01}$ and $N_{11}$ is

$$\begin{aligned}
&P(N_{01} = n_{01}, N_{11} = n_{11}|N_{0+} = 7, N_{1+} = 8) \\
&\quad = P(N_{01} = n_{01}|N_{0+} = 7) \cdot P(N_{11} = n_{11}|N_{1+} = 8) \\
&\quad = \tbinom{7}{n_{01}}\pi_{1|0}^{n_{01}}(1 - \pi_{1|0})^{7-n_{01}} \cdot \tbinom{8}{n_{11}}\pi_{1|1}^{n_{11}}(1 - \pi_{1|1})^{8-n_{11}}.
\end{aligned} \tag{2}$$

Then we condition on the columns sums as well, although we only write out $N_{+1}$ in the conditioning. In order to treat the two rows symmetrically, it is convenient to write out both row sums in the conditioning. This gives

$$\begin{aligned}
&P(N_{11} = k|N_{0+} = 7, N_{1+} = 8, N_{+1} = 9) \\
&\quad = P(N_{01} = 9 - k, N_{11} = k|N_{0+} = 7, N_{1+} = 8, N_{+1} = 9) \\
&\quad = P(N_{01} = 9 - k, N_{11} = k|N_{0+} = 7, N_{1+} = 8)/P(N_{+1} = 9|N_{0+} = 7, N_{1+} = 8), \\
&\quad \propto P(N_{01} = 9 - k, N_{11} = k|N_{0+} = 7, N_{1+} = 8) \\
&\quad = P(N_{01} = 9 - k|N_{0+} = 7) \cdot P(N_{11} = k|, N_{1+} = 8) \\
&\quad = \tbinom{7}{9-k}\pi_{1|0}^{9-k}(1 - \pi_{1|0})^{7-(9-k)} \cdot \tbinom{8}{k}\pi_{1|1}^{k}(1 - \pi_{1|1})^{8-k} \\
&\quad \propto \tbinom{7}{9-k}\tbinom{8}{k}\left[\pi_{0|0}\pi_{1|1}/(\pi_{0|1}\pi_{1|0})\right]^k \\
&\quad = \tbinom{7}{9-k}\tbinom{8}{k}\theta^k,
\end{aligned}$$

where in the fourth and fifth steps we used (2) and in last step we inserted the definition (1) of the odds ratio. The two expressions to the right and left of a proportionality sign $\propto$ differ by a multiplicative constant, not depending on $k$. The proportionality constant of the last step is chosen so that all probabilities sum to one. This gives a non-central hypergeometric distribution

$$P(N_{11} = k|N_{0+}, N_{1+} = 8, N_{+1} = 9) = \frac{\binom{7}{9-k}\binom{8}{k}\theta^k}{\sum_{j=2}^{8}\binom{7}{9-j}\binom{8}{j}\theta^j},$$

for $2 \leq k \leq 8$. The special case $\theta = 1$ was used in c) to find the mid $P$-value.

# Problem 2

a. Let $X$ refer to seat belt use, $Y$ to the fatality of the accident and let $\pi_{ij} = \mu_{ij}/\mu_{++}$ be the probability that each new observed accident belongs to cell $i, j$. The probability of a fatal accident for persons without seat belt is

$$\pi_1 = P(Y = 1|X = 1) = \frac{\pi_{11}}{\pi_{11} + \pi_{12}} = \frac{\mu_{11}/\mu_{++}}{\mu_{11}/\mu_{++} + \mu_{12}/\mu_{++}} = \frac{\mu_{11}}{\mu_{11} + \mu_{12}}.$$

The corresponding probability for those that use seat belt, is

$$\pi_2 = P(Y = 1|X = 2) = \frac{\mu_{21}}{\mu_{21} + \mu_{22}}.$$

b. When one conditions on row sums, the two cell counts $N_{11}$ and $N_{12}$ are independent random variables with binomial distributions $N_{11} \sim \text{Bin}(n_{1+}, \pi_1)$ and $N_{21} \sim \text{Bin}(n_{2+}, \pi_2)$ respectively. It follows that their joint distribution is

$$P(N_{11} = n_{11}, N_{21} = n_{21}) = \binom{n_{1+}}{n_{11}} \pi_1^{n_{11}} (1 - \pi_1)^{n_{12}} \cdot \binom{n_{2+}}{n_{21}} \pi_2^{n_{21}} (1 - \pi_2)^{n_{22}}. \quad (3)$$

c. As an estimator of the relative risk $r = \pi_1/\pi_2$ we use

$$\hat{r} = \frac{\hat{\pi}_1}{\hat{\pi}_2} = \frac{N_{11}/n_{1+}}{N_{21}/n_{2+}}. \quad (4)$$

We approximate $\log(\hat{r}) = f(\hat{\pi}_1, \hat{\pi}_2)$ by the first order Taylor expansion

$$\log(\hat{r}) \approx \log(r) + \frac{\partial \log(r)}{\partial \pi_1}(\hat{\pi}_1 - \pi_1) + \frac{\partial \log(r)}{\partial \pi_2}(\hat{\pi}_2 - \pi_2) = \log(r) + \frac{\hat{\pi}_1 - \pi_1}{\pi_1} - \frac{\hat{\pi}_2 - \pi_2}{\pi_2}$$

of the logarithm of the relative risk. It follows from (3) that $N_{11}$ and $N_{21}$ are independent binomial random variables. Therefore, $\hat{\pi}_1$ and $\hat{\pi}_2$ are independent binomial proportions with $\text{Var}(\hat{\pi}_i) = \pi_i(1 - \pi_i)/n_{i+}$ and

$$\begin{aligned} \text{Var}\left[\log(\hat{r})\right] &\approx \text{Var}\left[\tfrac{\hat{\pi}_1 - \pi_1}{\pi_1} - \tfrac{\hat{\pi}_2 - \pi_2}{\pi_2}\right] \\ &= \tfrac{\text{Var}(\hat{\pi}_1 - \pi_1)}{\pi_1^2} + \tfrac{\text{Var}(\hat{\pi}_2 - \pi_2)}{\pi_2^2} \\ &= \tfrac{\pi_1(1-\pi_1)/n_{1+}}{\pi_1^2} + \tfrac{\pi_2(1-\pi_2)/n_{2+}}{\pi_2^2} \\ &= \tfrac{1-\pi_1}{n_{1+}\pi_1} + \tfrac{1-\pi_2}{n_{2+}\pi_2}. \end{aligned}$$

d. The point estimates of the two fatal accident probabilities are

$$\begin{aligned} \hat{\pi}_1 &= 1085/(1085 + 55623) = 0.0191, \\ \hat{\pi}_2 &= 703/(703 + 444239) = 0.00158. \end{aligned}$$

This gives a point estimate

$$\hat{r} = \frac{0.0191}{0.00158} = 12.11$$

of the relative risk (4). The standard error of the estimator $\log(\hat{r})$ of $\log(r)$ is

$$\begin{aligned} \text{SE} &= \sqrt{\widehat{\text{Var}}[\log(\hat{r})]} \\ &= \sqrt{\frac{1-\hat{\pi}_1}{n_{1+}\hat{\pi}_1} + \frac{1-\hat{\pi}_2}{n_{2+}\hat{\pi}_2}} \\ &= \sqrt{\frac{1-\hat{\pi}_1}{n_{11}} + \frac{1-\hat{\pi}_2}{n_{21}}} \\ &= \sqrt{\frac{1-0.0191}{1085} + \frac{1-0.00158}{703}} \\ &= 0.0482, \end{aligned}$$

and the associated approximate 95% confidence interval for $\log(r)$ is

$$(\log(12.11) - 1.96 \cdot \text{SE}, \log(12.11) + 1.96 \cdot \text{SE}) = (2.400, 2.588).$$

If we transform this interval back to the original relative risk scale, we finally get an approximate 95% confidence interval

$$I = (e^{2.400}, e^{2.588}) = (11.02, 13.31)$$

for $r$. Since $1 \notin I$, we conclude that seat belt use has a significant effect on the fatality of an accident at level 5%.

# Problem 3

a. This data set is a threeway contingency table, with $I = 2$ levels for $X$, $J = 2$ levels for $Y$ and $K = 3$ levels for $Z$. The saturated model $(XYZ)$ has $IJK = 12$ parameters. All the eight submodels $M$ of the table below share one intercept parameter $\lambda$, and $(I-1)+(J-1)+(K-1) = 4$ marginal parameters. The number of parameters for the three types of second order interaction is

$$\begin{aligned} XY: &\quad (I-1)(J-1) = 1, \\ XZ: &\quad (I-1)(K-1) = 2, \\ YZ: &\quad (J-1)(K-1) = 2. \end{aligned}$$

By adding the relevant number of parameters for the different models we fill in the second column of the following table:

| $M$ | $p(M)$ | $G^2(M) + 2p(M)$ |
|---|---|---|
| $(XY, XZ, YZ)$ | 1+4+1+2+2=10 | 21.67 |
| $(XY, YZ)$ | 1+4+1+2=8 | 40.44 |
| $(XY, XZ)$ | 1+4+1+2=8 | 18.69 |
| $(XZ, YZ)$ | 1+4+2+2=9 | 141.28 |
| $(XZ, Y)$ | 1+4+2=7 | 138.27 |
| $(X, YZ)$ | 1+4+2=7 | 160.02 |
| $(XY, Z)$ | 1+4+1=6 | 37.42 |
| $(X, Y, Z)$ | 1+4=5 | 157.00 |

b. Akaike's Information Criterion for submodel $M$ is

$$\begin{aligned} \text{AIC}(M) &= -2L(M) + 2p(M) \\ &= -2L(XYZ) + 2\left[L(XYZ) - L(M)\right] + 2p(M) \\ &= -2L(XYZ) + G^2(M) + 2p(M). \end{aligned} \tag{5}$$

Since the first term $-2L(XYZ)$ on the right hand side of (5) only involves the saturated model $XYZ$ it does not depend on $M$. Therefore, minimizing AIC($M$) is equivalent to minimizing $G^2(M) + 2p(M)$. By adding twice the values $p(M)$ of the middle column in the above table to the known deviance values, we obtain the values of the right column. We find that $G^2(M) + 2p(M)$ is minimized by $M = (XY, XZ)$, which is the best model according to the AIC criterion.

c. For model $M_0 = (XY, Z)$ we have that

$$\mu_{ijk} = \frac{\mu_{ij+}\mu_{++k}}{\mu_{+++}} \Longrightarrow \hat{\mu}_{ijk} = \frac{n_{ij+}n_{++k}}{n_{+++}}.$$

In particular,

$$\hat{\mu}_{221} = \frac{n_{22+}n_{++1}}{n_{+++}} = \frac{(26+26+10)(39160+3882+25+26)}{39160+3882+\ldots+13+10} = \frac{62 \cdot 43093}{105636} = 25.29.$$

d. For model $M_1 = (XY, XZ)$ we have that

$$\mu_{ijk} = \frac{\mu_{ij+}\mu_{i+k}}{\mu_{i++}} \Longrightarrow \hat{\mu}_{ijk} = \frac{n_{ij+}n_{i+k}}{n_{i++}}.$$

In particular,

$$\hat{\mu}_{221} = \frac{n_{22+}n_{2+1}}{n_{2++}} = \frac{(26+26+10)(3882+26)}{3882+4291+1337+26+26+10} = \frac{62 \cdot 3908}{9572} = 25.31.$$

e. The log likelihood ratio statistic between $H_0 : M_0$ and $H_a : M_1 \setminus M_0$ is

$$\begin{aligned} G^2(M_0|M_1) &= 2\left[L(M_1) - L(M_0)\right] \\ &= G^2(M_0) - G^2(M_1) \\ &= 25.42 - 2.69 \\ &= 22.73 \\ &> \chi^2_2(0.05) = 5.99, \end{aligned}$$

where in the last step we used that there are $8 - 6 = 2$ degrees of freedom, the number of additional parameters in $M_1$ compared to $M_0$. Hence we reject the null hypothesis $H_0$ at level 5%.

# Problem 4

a. The expected cell counts $\mu_{ijk}$ of the loglinear model $(XY, XZ)$ satisfy

$$\log(\mu_{ijk}) = \lambda + \lambda^X_i + \lambda^Y_j + \lambda^Z_k + \lambda^{XY}_{ij} + \lambda^{XZ}_{ik}, \quad 1 \le i, j \le 2, 1 \le k \le 3.$$

If we choose the lowest level ($i = j = k = 1$) of each variable as baseline, any parameter with at least one index at its lowest level is put to zero in order to avoid overparametrization. The remaining eight free parameters are

$$(\lambda, \lambda_2^X, \lambda_2^Y, \lambda_2^Z, \lambda_3^Z, \lambda_{22}^{XY}, \lambda_{22}^{XZ}, \lambda_{23}^{XZ}).$$

b. Write $\pi_{ijk} = \mu_{ijk}/\mu_{+++}$ for the cell probabilities under multinomial sampling. Then

$$\begin{aligned}
&\text{logit}P(Y = 2|X = i, Z = k) \\
&= \log P(Y = 2|X = i, Z = k) - \log P(Y = 1|X = i, Z = k) \\
&= \log(\pi_{i2k}/\pi_{i+k}) - \log(\pi_{i1k}/\pi_{i+k}) \\
&= \log(\pi_{i2k}) - \log(\pi_{i1k}) \\
&= \log(\mu_{i2k}/\mu_{+++}) - \log(\mu_{i1k}/\mu_{+++}) \\
&= \log(\mu_{i2k}) - \log(\mu_{i1k}) \qquad (6) \\
&= (\lambda + \lambda_i^X + \lambda_2^Y + \lambda_k^Z + \lambda_{i2}^{XY} + \lambda_{ik}^{XZ}) \\
&-(\lambda + \lambda_i^X + \lambda_1^Y + \lambda_k^Z + \lambda_{i1}^{XY} + \lambda_{ik}^{XZ}) \\
&= (\lambda_2^Y - \lambda_1^Y) + (\lambda_{i2}^{XY} - \lambda_{i1}^{XY}) \\
&= \lambda_2^Y + \lambda_{i2}^{XY} \\
&=: \alpha + \beta_i^X,
\end{aligned}$$

if we use the parameter constraints of the loglinear model from a). It follows from (6) that $Y|X, Z$ is a logistic regression model with two nonzero parameters $\alpha = \lambda_2^Y$ and $\beta_2^X = \lambda_{22}^{XY}$, since $\beta_1^X = \lambda_{12}^{XY} = 0$.

c. We have that

$$\theta_{XY(k)} = \frac{P(Y = 2|X = 2, Z = k)/P(Y = 1|X = 2, Z = k)}{P(Y = 2|X = 1, Z = k)/P(Y = 1|X = 1, Z = k)}. \qquad (7)$$

Taking the logarithm and using (6), it follows that

$$\begin{aligned}
\log \theta_{XY(k)} &= \text{logit}P(Y = 2|X = 2, Z = k) - \text{logit}P(Y = 2|X = 1, Z = k) \\
&= (\alpha + \beta_2^X) - \alpha \\
&= \beta_2^X,
\end{aligned}$$

so that

$$\theta_{XY(k)} = \exp(\beta_2^X). \qquad (8)$$

Since $\theta_{XY(k)}$ does not depend on the level $k$ of $Z$, there is homogeneous association between $X$ and $Y$. This also follows from the fact that there is no third order association between $X$, $Y$ and $Z$ in model $(XY, XZ)$.

d. The marginal odds ratio between $X$ and $Y$ is given by

$$\theta_{XY} = \frac{P(Y = 2|X = 2)/P(Y = 1|X = 2)}{P(Y = 2|X = 1)/P(Y = 1|X = 1)} = \frac{\mu_{22+}\mu_{11+}}{\mu_{12+}\mu_{21+}}, \qquad (9)$$

where the last step follows after some computations, similarly as in the first steps of (6). But since $Y$ and $Z$ are conditionally independent given $X$ for model $(XY, XZ)$, it follows that

$$P(Y = j|X = i, Z = k) = P(Y = j|X = i). \qquad (10)$$

Comparing (7) and (9), we find because of (10) that

$$\theta_{XY(k)} = \theta_{XY}. \qquad (11)$$

e. Equation (10) implies that the likelihoods of $Y|X,Z$ and $Y|X$ are the same. The maximum likelihood estimator $\hat{\theta}_{XY(k)}$ of the conditional odds ratio $\theta_{XY(k)}$ is therefore a function of the twoway marginal table of $X$ and $Y$. In particular, we find from (8) and (11) that

$$\hat{\theta}_{XY(k)} = \exp(\hat{\beta}_2^X) = \hat{\theta}_{XY} = \frac{n_{22+}n_{11+}}{n_{12+}n_{21+}} = \frac{62 \cdot 95987}{77 \cdot 9510} = 8.13$$

for $k = 1, 2, 3$.