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

MT 5019
SOLUTIONS
January 4 2024

Solutions for Examination Categorical Data Analysis, January 4, 2024

Problem 1

a. The null hypothesis H_0 that hands-free mobile usage and accident proneness are independent, corresponds to $\pi_{ij} = \pi_{i+}\pi_{+j}$, where $\pi_{i+} = \pi_{i0}+\pi_{i1}$, and $\pi_{+j} = \pi_{0j}+\pi_{1j}$.

b. Fisher's exact test uses a hypergeometric distribution

$$P_{H_0}(N_{11} = n_{11}|n_{0+}, n_{1+}, n_{+0}, n_{+1}) = \frac{\binom{n_{0+}}{n_{+1}-n_{11}}\binom{n_{1+}}{n_{11}}}{\binom{n_{++}}{n_{+1}}} = \frac{\binom{13}{12-n_{11}}\binom{17}{n_{11}}}{\binom{30}{12}},$$

for the number N_{11} of mobile users with accidents, and $n_{11} = 0, 1, \ldots, 12$. This distribution is based on drawing n_{+1} persons with accidents from a sample of size n_{++} that consists of n_{0+} persons who do not use the mobile while driving, and n_{1+} who do. It is also possible to reverse the role of columns and rows, and draw n_{1+} mobile users from a sample that consists of n_{+0} persons without accidents and n_{+1} with accidents. This gives

$$P_{H_0}(N_{11} = n_{11}|n_{0+}, n_{1+}, n_{+0}, n_{+1}) = \frac{\binom{n_{+0}}{n_{1+}-n_{11}}\binom{n_{+1}}{n_{11}}}{\binom{n_{++}}{n_{1+}}} = \frac{\binom{18}{17-n_{11}}\binom{12}{n_{11}}}{\binom{30}{17}},$$

for $n_{11} = 0, 1, \ldots, 12$.

c. A one sided alternative where mobile usage increases accident risk corresponds to an alternative hypothesis $H_a : \theta > 1$, where $\theta = (\pi_{11}\pi_{00})/(\pi_{01}\pi_{10})$ is the odds ratio.

d. Since H_0 is rejected for large values of N_{11} for the one-sided alternative hypothesis in c, and $n_{11} = 9$, we get

$$\begin{aligned} P\text{-value} &= P_{H_0}(N_{11} \geq 9) \\ &= 0.0804 + 0.0175 + 0.0019 + 0.0001 = 0.100, \\ \text{mid } P\text{-value} &= 0.5P_{H_0}(N_{11} = 9) + P_{H_0}(N_{11} \geq 10) \\ &= 0.5 \cdot 0.0804 + 0.0175 + 0.0019 + 0.0001 = 0.060. \end{aligned} \quad (1)$$

e. It follows from 1d and

$$P_{H_0}(N_{11} \geq 10) = 0.0175 + 0.0019 + 0.0001 < 0.05$$

that both the P-value and the mid P-value are smaller than 0.05 when $n_{11} \geq 10$. Since both the P-value and the mid P-value are larger than 0.05 when $n_{11} = 9$ (cf. (1)) we deduce that the actual significance level of a test with nominal significance level $\alpha = 0.05$, is

$$\begin{array}{rcl} P_{H_0}(P\text{-value} \leq 0.05) & = & P_{H_0}(N_{11} \geq 10) < 0.05, \\ P_{H_0}(\text{mid-}P\text{-value} \leq 0.05) & = & P_{H_0}(N_{11} \geq 10) < 0.05. \end{array}$$

This implies that both tests are conservative (that is, the actual significance levels are smaller than the nominal significance level 0.05). On the other hand, when $n_{11} = 9$, it follows from (1) that the P-value is larger than 0.07, whereas the mid P-value is smaller than 0.07. Consequently, when the nominal significance level is $\alpha = 0.07$, then the actual significance levels are

$$\begin{array}{rcl} P_{H_0}(P\text{-value} \leq 0.07) & = & P_{H_0}(N_{11} \geq 10) < 0.07, \\ P_{H_0}(\text{mid-}P\text{-value} \leq 0.07) & = & P_{H_0}(N_{11} \geq 9) > 0.07. \end{array}$$

This implies that the P-value based test is conservative when $\alpha = 0.07$, whereas the other test based on the mid P-value is anti conservative when $\alpha = 0.07$.

Problem 2

a. Since θ_{ij} is the odds ratio of a table with rows i and $i+1$, and columns j and $j+1$, it follows that

$$\theta_{ij} = \frac{\mu_{ij}\mu_{i+1,j+1}}{\mu_{i,j+1}\mu_{i+1,j}} \tag{2}$$

for four different combinations of i and j ($1 \leq i, j \leq 2$).

b. Estimates $\hat{\theta}_{ij}$ are obtained by replacing all μ_{ij} in (2) with n_{ij}, so that

$$\begin{array}{rcl} \hat{\theta}_{11} & = & (n_{11}n_{22})/(n_{12}n_{21}) = (34 \cdot 174)/(80 \cdot 53) = 1.395, \\ \hat{\theta}_{12} & = & (n_{12}n_{23})/(n_{22}n_{13}) = (53 \cdot 304)/(174 \cdot 88) = 1.052, \\ \hat{\theta}_{21} & = & (n_{21}n_{32})/(n_{31}n_{22}) = (80 \cdot 175)/(29 \cdot 174) = 1.189, \\ \hat{\theta}_{22} & = & (n_{22}n_{33})/(n_{23}n_{32}) = (174 \cdot 172)/(75 \cdot 304) = 1.313, \end{array}$$

c. We have that

$$\begin{array}{rcl} \text{Var}\left[\log(\hat{\theta}_{11}/\hat{\theta}_{12})\right] & = & \text{Var}\left[\log\left\{(\frac{N_{11}N_{22}}{N_{12}N_{21}})/(\frac{N_{12}N_{23}}{N_{13}N_{22}})\right\}\right] \\ & = & \text{Var}\,[\log(N_{11}) - \log(N_{21}) - 2\log(N_{12}). \\ & & \quad +2\log(N_{22}) + \log(N_{13}) - \log(N_{23})] \\ & \approx & \frac{1}{\mu_{11}} + \frac{1}{\mu_{21}} + \frac{4}{\mu_{12}} + \frac{4}{\mu_{22}} + \frac{1}{\mu_{13}} + \frac{1}{\mu_{23}}, \end{array}$$

where in the third step we used independence of the six terms (since N_{ij} are independent), and computed the variance of a Taylor expansion for each one of them, according to

$$\text{Var}\left[\log(N_{ij})\right] \approx \text{Var}\left[\log(\mu_{ij}) + \frac{N_{ij}-\mu_{ij}}{\mu_{ij}}\right] = \frac{\text{Var}(N_{ij})}{\mu_{ij}^2} = \frac{\mu_{ij}}{\mu_{ij}^2} = \frac{1}{\mu_{ij}}.$$

The corresponding estimate of the variance is obtained by replacing all μ_{ij} with n_{ij}, i.e.

$$\begin{aligned}\widehat{\text{Var}}\left[\log(\hat\theta_{11}/\hat\theta_{12})\right] &= \tfrac{1}{n_{11}} + \tfrac{1}{n_{21}} + \tfrac{4}{n_{12}} + \tfrac{4}{n_{22}} + \tfrac{1}{n_{13}} + \tfrac{1}{n_{23}} \\ &= \tfrac{1}{34} + \tfrac{1}{80} + \tfrac{4}{53} + \tfrac{4}{174} + \tfrac{1}{88} + \tfrac{1}{304} \\ &= 0.1550.\end{aligned}$$

d. We first compute a one-sided confidence interval for $\log(\theta_{11}/\theta_{12})$ as

$$\begin{aligned}\left(\log(\tfrac{\hat\theta_{11}}{\hat\theta_{12}}) - 1.645\sqrt{0.1550}, \infty\right) &= \left(\log(\tfrac{1.395}{1.052}) - 1.645\sqrt{0.1550}, \infty\right) \\ &= (-0.3654, \infty).\end{aligned}$$

The corresponding one-sided confidence interval for θ_{11}/θ_{12} is

$$(\exp(-0.3654), \infty) = (0.694, \infty).$$

Since 1 is included in the interval, we cannot reject the null hypothesis at level 5%.

Problem 3

a. The cell probabilities π_{ij} are proportional to μ_{ij} with sum 1, i.e. $\pi_{ij} = \mu_{ij}/\mu_{++} = \mu_{ij}/n$. In the last step we used that $n = \mu_{++}$ for multinomial sampling, since $\mu_{ij} = n\pi_{ij}$.

b. The number of concordant and discordant pairs are

$$\begin{aligned} C &= 34(174+304+75+172) + 53(304+172) + 80(75+172) + 174\cdot 172 = 99566, \\ D &= 53(80+29) + 88(80+174+29+75) + 174\cdot 29 + 304(29+75) = 73943 \end{aligned}$$

respectively. Therefore, an estimator of γ is

$$\hat\gamma = \frac{99566-73943}{99566+73943} = 0.148.$$

This indicates a positive association between age and job satisfaction.

c. A pair $(X,Y) = (i,j)$ and $(X',Y') = (h,k)$ of cells is concordant if $i < h, j < k$ or $i > h, j > k$. For a large population (not the sample with n individuals!), we may regard (X,Y) and (X',Y') as drawn independently with replacement, so that

$$\begin{aligned} P\left[(X,Y) = (i,j), (X',Y') = (h,k)\right] &= P\left[(X,Y) = (i,j)\right] P\left[(X',Y') = (h,k)\right] \\ &= \pi_{ij}\pi_{hk}.\end{aligned}$$

Therefore, the probability of a concordant pair is

$$\begin{array}{rcl}\Pi_c & = & \sum_{i,j} \pi_{ij} \sum_{h,k;h>i,k>j} \pi_{hk} + \sum_{h,k} \pi_{hk} \sum_{i,j;i>h,j>k} \pi_{ij} \\ & = & 2 \sum_{i,j} \pi_{ij} \sum_{h,k;h>i,k>j} \pi_{hk}.\end{array}$$

Since a cell pair is discordant if $i < h, j > k$ or $i > h, j < k$, an analogous calculation gives

$$\Pi_d = 2 \sum_{i,j} \pi_{ij} \sum_{h,k;h>i,k<j} \pi_{hk}$$

for the probability of such a pair.

d. Since $\pi_{ij} = \pi_{i+}\pi_{+j}$ under the null hypothesis that age and job satisfaction are independent, it follows that

$$\begin{array}{rcl}\Pi_c & = & 2 \sum_{i,j} \pi_{i+}\pi_{+j} \sum_{h,k;h>i,k>j} \pi_{h+}\pi_{+k} \\ & = & 2 \left(\sum_i \pi_{i+} \sum_{h;h>i} \pi_{h+} \right) \left(\sum_j \pi_{+j} \sum_{k;k>j} \pi_{+k} \right).\end{array}$$

A similar calculation gives

$$\begin{array}{rcl}\Pi_d & = & 2 \left(\sum_i \pi_{i+} \sum_{h;h>i} \pi_{h+} \right) \left(\sum_j \pi_{+j} \sum_{k;k<j} \pi_{+k} \right) \\ & = & 2 \left(\sum_i \pi_{i+} \sum_{h;h>i} \pi_{h+} \right) \left(\sum_k \pi_{+k} \sum_{j;j>k} \pi_{+j} \right).\end{array}$$

By interchanging the role of indeces j and k in the last sum, we conclude that $\Pi_c = \Pi_d$, and hence $\gamma = 0$.

Problem 4

a. The expected cell counts of the $M_0 = (XY, ZY)$ loglinear model are

$$\mu_{ij} = \mu_{ij}(M_0, \boldsymbol{\beta}) = \exp\left(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} \right), \quad 0 \le i, j, k \le 1. \quad (3)$$

If $i = j = k = 0$ are baseline levels, then all parameters with at least one 0 index are put to zero. This gives a parameter vector

$$\boldsymbol{\beta} = (\lambda, \lambda_1^X, \lambda_1^Y, \lambda_1^Z, \lambda_{11}^{XY}, \lambda_{11}^{YZ}), \quad (4)$$

with $p = 6$ components.

b. It follows from (3) that

$$\mu_{ijk} = B_{ij} C_{jk}, \quad (5)$$

where, for instance, $B_{ij} = \exp(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY})$ and $C_{jk} = \exp(\lambda_k^Z + \lambda_{jk}^{YZ})$. Then, summing over one of i or k, or over both indeces simultaneously in (5), we find that

$$\begin{array}{rcl}\mu_{ij+} & = & B_{ij} C_{j+}, \\ \mu_{+jk} & = & B_{+j} C_{jk}, \\ \mu_{+j+} & = & B_{+j} C_{j+}.\end{array}$$

Consequently,

$$\frac{\mu_{ij+}\mu_{+jk}}{\mu_{+j+}} = \frac{B_{ij}C_{j+} \cdot B_{+j}C_{jk}}{B_{+j}C_{j+}} = B_{ij}C_{jk} = \mu_{ijk}.$$

Alternatively, we may work directly with the cell probabilities $\pi_{ijk} = \mu_{ijk}/\mu_{+++}$. Since X and Z are conditionally independent given Y for model (XY, YZ), it follows that

$$\pi_{ijk} = \pi_{+j+}\pi_{ik|j} = \pi_{+j+}\pi_{i+|j}\pi_{+k|j} = \pi_{+j+} \cdot \frac{\pi_{ij+}}{\pi_{+j+}} \cdot \frac{\pi_{+jk}}{\pi_{+j+}} = \frac{\pi_{ij+}\pi_{+jk}}{\pi_{+j+}},$$

and hence

$$\mu_{ijk} = \mu_{+++}\pi_{ijk} = \mu_{+++} \cdot \frac{\frac{\mu_{ij+}}{\mu_{+++}} \cdot \frac{\mu_{+jk}}{\mu_{+++}}}{\frac{\mu_{+j+}}{\mu_{+++}}} = \frac{\mu_{ij+}\mu_{+jk}}{\mu_{+j+}}.$$

c. The maximum likelihood estimates

$$\hat{\mu}_{ijk} = \frac{n_{ij+}n_{+jk}}{n_{+j+}}$$

of the expected cell counts are obtained by replacing μ_{ij+}, μ_{+jk} and μ_{+j+} by estimates n_{ij+}, n_{+jk} and n_{+j+}. From the two partial tables we can compute row sums n_{ij+}, columns sums n_{+jk}, and total number of observations $n_{+0+} = 283$ and $n_{+1+} = 137$. This gives

$$\hat{\mu}_{000} = \frac{n_{00+}n_{+00}}{n_{+0+}} = \frac{(93 + 39) \cdot (93 + 101)}{283} = 90.49.$$

Continuing in this way for the other cells (i, j, k), we get the following predicted expected cell counts $\hat{\mu}_{ijk}$:

No cancer $Y = 0$:

Exposure	Smoking $Z = k$		
$X = i$	$k = 0$	$k = 1$	Sum
$i = 0$	90.49	41.51	132
$i = 1$	103.51	47.49	151
Sum	194	89	283

Cancer $Y = 1$:

Exposure	Smoking $Z = k$		
$X = i$	$k = 0$	$k = 1$	Sum
$i = 0$	10.67	23.33	34
$i = 1$	32.33	70.67	103
Sum	43	94	137

d. The log likelihood ratio statistic for testing (XY, YZ) against the saturated model (XYZ), is

$$\begin{aligned} G^2 &= 2\textstyle\sum_{ijk} n_{ijk} \log \frac{n_{ijk}}{\hat{\mu}_{ijk}} \\ &= 2\left(93 \cdot \log \tfrac{93}{90.49} + \ldots + 72 \cdot \log \tfrac{72}{70.67}\right) \\ &= 0.733 \\ &< \chi^2_2(0.05) = 5.99, \end{aligned}$$

where in the last step we used that df $= 8 - 6 = 2$, since the saturated model has $2 \times 2 \times 2 = 8$ parameters, and the conditional independence model (XY, YZ) has 6 parameters according to (4). We thus cannot reject conditional independence between X and Z given Y at level 5%, indicating that smoking and exposure don't have a joint effect on lung cancer.

e. For any model M, the maximum likelihood is

$$l(M) = \max_{\boldsymbol{\beta}} \prod_{ijk} e^{-\mu_{ijk}} \frac{\mu_{ijk}^{n_{ijk}}}{n_{ijk}!} = \prod_{ijk} e^{-\hat{\mu}_{ijk}} \frac{\hat{\mu}_{ijk}^{n_{ijk}}}{n_{ijk}!},$$

where $\hat{\mu}_{ijk} = \hat{\mu}_{ijk}(M, \hat{\boldsymbol{\beta}}(M))$ are the fitted cell counts for model M, based on plugging the ML estimate $\hat{\boldsymbol{\beta}}(M)$ of that model into (3). This gives a log likelihood

$$L(M) = \log(l(M)) = \text{constant} + \sum_{ijk} \left[n_{ijk} \log(\hat{\mu}_{ijk}) - \hat{\mu}_{ijk} \right], \tag{6}$$

with a constant ($= -\sum_{ijk} n_{ijk}!$) that is the same for any M. Since the saturated model has the same fitted and observed counts, $\hat{\mu}_{ijk}(M_1, \hat{\boldsymbol{\beta}}(M_1)) = n_{ijk}$, it follows that the deviance equals

$$\begin{aligned} G^2(M) &= 2\left[L(M_1) - L(M)\right] \\ &= 2\textstyle\sum_{ijk} \left[n_{ijk} \log(\frac{n_{ijk}}{\hat{\mu}_{ijk}}) - (n_{ijk} - \hat{\mu}_{ijk}) \right], \\ &= 2\textstyle\sum_{ijk} n_{ijk} \log(\frac{n_{ijk}}{\hat{\mu}_{ijk}}). \end{aligned}$$

In the last step we used that the baseline parameter λ is part of M. Indeed, differentiating (6) with respect to λ we find that

$$\left.\frac{\partial L(M)}{\partial \lambda}\right|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}} = \sum_{ijk} (n_{ijk} - \hat{\mu}_{ijk}) = 0,$$

since, by equation (3), $d\log(\mu_{ijk})/d\lambda = 1$ and $d\mu_{ijk}/d\lambda = \mu_{ijk}$.

Problem 5

a. Let $\pi_{ijk} = \mu_{ijk}/\mu_{+++}$ refer to the multinomial cell probabilities of the contingency table. We have that

$$\begin{aligned} \text{logit}P(Y=1|X=i, Z=k) &= \log\left(\tfrac{\pi_{i1k}/\pi_{i+k}}{\pi_{i0k}/\pi_{i+k}}\right) = \log\left(\tfrac{\pi_{i1k}}{\pi_{i0k}}\right) \\ &= \log\left(\tfrac{\mu_{i1k}/\mu_{+++}}{\mu_{i0k}/\mu_{+++}}\right) = \log\left(\tfrac{\mu_{i1k}}{\mu_{i0k}}\right) \\ &= \alpha + \beta_i^X + \beta_k^Z, \end{aligned} \tag{7}$$

where in the last step we used (3), with $\alpha = \lambda_1^Y - \lambda_0^Y$, $\beta_i^X = \lambda_{i1}^{XY} - \lambda_{i0}^{XY}$ and $\beta_k^Z = \lambda_{1k}^{YZ} - \lambda_{0k}^{YZ}$. Since $i = j = k = 0$ are baseline levels for the loglinear model, it follows that the three nonzero parameters of the logistic regression model are

$$\begin{aligned} \alpha &= \lambda_1^Y, \\ \beta_1^X &= \lambda_{11}^{XY}, \\ \beta_1^Z &= \lambda_{11}^{YZ}. \end{aligned}$$

b. The log conditional odds ratio between X and Y is

$$\begin{aligned} \log(\theta_{(k)}^{XY}) &= \log \tfrac{P(Y=1|X=1,Z=k)/P(Y=0|X=1,Z=k)}{P(Y=1|X=0,Z=k)/P(Y=0|X=0,Z=k)} \\ &= \text{logit}P(Y=1|X=1, Z=k) - \text{logit}P(Y=1|X=0, Z=k) \\ &= (\alpha + \beta_1^X + \beta_k^Z) - (\alpha + \beta_0^X + \beta_k^Z) \\ &= \beta_1^X, \end{aligned} \tag{8}$$

where in the third step we used (7). Hence

$$\theta_{(k)}^{XY} = \exp(\beta_1^X).$$

The association between exposure and lung cancer is homogeneous, since $\theta_{(k)}^{XY}$ does not depend on the level k of the confounding variable Z.

c. The marginal odds ratio

$$\theta^{XY} = \frac{\mu_{00+}\mu_{11+}}{\mu_{01+}\mu_{10+}}$$

between X and Y can expressed in terms of the expected cell counts μ_{ij+} of the XY marginal table. It follows from (3) that $\mu_{ij+} = B_{ij}C_{j+}$. After some simplifications, this gives

$$\theta^{XY} = \frac{B_{00}B_{11}}{B_{01}B_{10}} = \exp(\lambda_{11}^{XY}) = \exp(\beta_1^X) = \theta_{(k)}^{XY},$$

where in the second step we used that $B_{ij} = \exp(\lambda + \lambda_j^Y + \lambda_i^X + \lambda_{ij}^{XY})$, and that $X = 0$ and $Y = 0$ are baseline levels. This proves that the marginal and conditional odds ratios are the same.

Alternatively, we express conditional and marginal odds ratios in terms of probabilities. We can use Bayes' Theorem to rewrite the conditional odds ratio in (8) as

$$\theta_{(k)}^{XY} = \frac{P(X=1|Y=1,Z=k)/P(X=0|Y=1,Z=k)}{P(X=1|Y=0,Z=k)/P(X=0|Y=0,Z=k)}. \tag{9}$$

The marginal odds ratio between X and Y can similarly be written as

$$\theta^{XY} = \frac{P(X=1|Y=1)/P(X=0|Y=1)}{P(X=1|Y=0)/P(X=0|Y=0)}. \tag{10}$$

But since X and Z are conditionally independent given Y for loglinear model M_0, it follows that

$$P(X=i|Y=j,Z=k) = P(X=i|Y=j)$$

for all i, j, k. Comparing (9) and (10), we conclude that the conditional and marginal odds ratios are the same.

d. We can use parts 5b and 5c to deduce that $\beta_1^X = \log(\theta^{XY})$. The ML estimator of the marginal odds ratio is obtained from the marginal table n_{ij+} of cell counts, as

$$\hat{\theta}^{XY} = \frac{n_{00+}n_{11+}}{n_{01+}n_{10+}} = \frac{132 \cdot 103}{34 \cdot 151} = 2.648.$$

Since the ML estimator of a function of θ^{XY} is the same function of $\hat{\theta}^{XY}$, it follows that the ML estimator of β_1^X is

$$\hat{\beta}_1^X = \log(\hat{\theta}^{XY}) = \log(2.648) = 0.974.$$

We may also take the logarithm of any of the two estimated conditional odds ratios

$$\begin{aligned}\hat{\theta}_{(0)}^{XY} &= (93 \cdot 31)/(101 \cdot 12) = 2.379,\\ \hat{\theta}_{(1)}^{XY} &= (39 \cdot 72)/(50 \cdot 22) = 2.553,\end{aligned}$$

in order to estimate β_1^X, but they are both different from the ML estimator.