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

MT 5006
SOLUTIONS
February 20 2019

Solutions for Examination
Categorical Data Analysis, February 20, 2019

Problem 1

a. The null and alternative hypotheses are

$$\begin{array}{ll} H_0: & P(Y=1|X=0) = P(Y=1|X=1), \\ H_a: & P(Y=1|X=0) < P(Y=1|X=1), \end{array} \tag{1}$$

for a one-sided test, since the aim is to test whether smoking does not change (H_0) or increases (H_a) the risk of lung cancer. This is equivalent to testing

$$\begin{array}{ll} H_0: & \theta = 1, \\ H_a: & \theta > 1, \end{array} \tag{2}$$

where

$$\begin{array}{rcl} \theta & = & \frac{P(Y=1|X=1)/[1-P(Y=1|X=1)]}{P(Y=1|X=0)/[1-P(Y=1|X=0)]} \\ & = & \frac{P(Y=1|X=1)/P(Y=0|X=1)}{P(Y=1|X=0)/P(Y=0|X=0)} \end{array} \tag{3}$$

is the odds ratio of lung cancer between those that smoke and those that don't. Bayes' Theorem implies $P(Y=j|X=i) = P(X=i|Y=j)P(Y=j)/P(X=i)$. We insert this formula into to all four terms that appear on the right hand side of (3), and notice after rearrangement that each $P(Y=j)$ and $P(X=i)$ term appears twice, in the numerator and denominator, and hence cancels out. This implies that

$$\theta = \frac{P(X=1|Y=1)/P(X=0|Y=1)}{P(X=1|Y=0)/P(X=0|Y=0)} = \frac{\pi_1/(1-\pi_1)}{\pi_0(1-\pi_0)} \tag{4}$$

can be expressed in terms of π_0 and π_1, two quantities that can be estimated in a case-control study.

b. The sought for null distribution of N_{11} is hyptergeometric, with

$$P(N_{11} = n_{11}|H_0, N_{1+} = 6) = \frac{\binom{10}{n_{11}} \cdot \binom{10}{6-n_{11}}}{\binom{20}{6}}, \tag{5}$$

for $n_{11} = 0, 1, \ldots, 6$.

c. Since $n_{11} = 5$, we use (5) in order to find the one-sided

$$\begin{aligned}
P\text{-value} &= P(N_{11} \geq 5 | H_0, N_{1+} = 6) \\
&= P(N_{11} = 5 | H_0, N_{1+} = 6) + P(N_{11} = 6 | H_0, N_{1+} = 6) \\
&= \frac{\binom{10}{5} \cdot \binom{10}{1}}{\binom{20}{6}} + \frac{\binom{10}{6} \cdot \binom{10}{0}}{\binom{20}{6}} \\
&= \frac{252 \cdot 10 + 210 \cdot 1}{38760} \\
&= \frac{2730}{38760} \\
&= 0.0704.
\end{aligned}$$

d. Since the number of smokers in each column have independent binomial distributions $N_{10} \sim \text{Bin}(10, \pi_0)$ and $N_{11} \sim \text{Bin}(10, \pi_1)$, it follows that

$$\begin{aligned}
P(N_{11} = n_{11} | \theta, N_{1+} = 6) = P(N_{10} &= 6 - n_{11}, N_{11} = n_{11} | N_{1+} = 6) \\
&\propto P(N_{10} = 6 - n_{11}, N_{11} = n_{11}) \\
&= P(N_{10} = 6 - n_{11}) P(N_{11} = n_{11}) \\
&= \binom{10}{6-n_{11}} \pi_0^{6-n_{11}} (1 - \pi_0)^{4+n_{11}} \cdot \binom{10}{n_{11}} \pi_1^{n_{11}} (1 - \pi_1)^{10-n_{11}} \\
&\propto \binom{10}{6-n_{11}} \binom{10}{n_{11}} \theta^{n_{11}},
\end{aligned} \tag{6}$$

where in the first step we dropped θ in the notation for conditional probabilities, and in the second and lasts steps the proportionality constants are independent of n_{11}. In the last step of (6) we also used (4). Since the probabilities in (6) must sum to 1, we find that

$$P(N_{11} = n_{11} | \theta, N_{1+} = 6) = \frac{\binom{10}{6-n_{11}} \binom{10}{n_{11}} \theta^{n_{11}}}{\sum_{k=0}^{6} \binom{10}{6-k} \binom{10}{k} \theta^k},$$

for $n_{11} = 0, \ldots, 6$, which simplifies to (5) when $\theta = 1$.

Problem 2

a. Since $N_{10} \sim \text{Bin}(50, \pi_0)$ and $N_{11} \sim \text{Bin}(50, \pi_1)$ have independent binomial distributions, the maximum likelihood estimate of Δ is

$$\hat{\Delta} = \hat{\pi}_1 - \hat{\pi}_0 = \frac{n_{11}}{n_1} - \frac{n_{10}}{n_0} = \frac{25}{50} - \frac{5}{50} = 0.4,$$

with variance

$$\text{Var}(\hat{\Delta}) = \text{Var}(\hat{\pi}_0) + \text{Var}(\hat{\pi}_1) = \frac{\pi_0(1 - \pi_0)}{n_0} + \frac{\pi_1(1 - \pi_1)}{n_1}$$

and standard error

$$\begin{aligned}
\text{SE}(\hat{\Delta}) &= \sqrt{\widehat{\text{Var}}(\hat{\Delta})} \\
&= \sqrt{\frac{\hat{\pi}_0(1-\hat{\pi}_0)}{n_0} + \frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1}} \\
&= \sqrt{\frac{0.5(1-0.5)}{50} + \frac{0.1(1-0.1)}{50}} \\
&= 0.0825.
\end{aligned}$$

The Wald-based test statistic is

$$z_W = \frac{\hat{\Delta}}{\mathrm{SE}(\hat{\Delta})} = \frac{0.4}{0.0825} = 4.85.$$

Since $|z_W| > 1.96 = z_{0.025}$, the 0.975-quantile of a standard normal distribution, we reject H_0 at level 5%. This is in contrast to Problem 1c), where the null hypothesis was *not* rejected (P-value larger than 0.05), even though that test was one-sided (which in this case makes it easier to reject the null hypothesis). The main reason why H_0 was not rejected Problem 1c) was the small data set (five times as small compared to Problem 2), which makes it harder to detect correlation between lung cancer and smoking.

b. The maximum likelihood estimator of r is

$$\hat{r} = \frac{\hat{\pi}_1}{\hat{\pi}_0} = \frac{0.5}{0.1} = 5.$$

By means of a first order Taylor expansion, we have that

$$\begin{aligned}\log(\hat{r}) &= \log(\hat{\pi}_1) - \log(\hat{\pi}_0)\\ &\approx \log(\pi_1) + \tfrac{\hat{\pi}_1-\pi_1}{\pi_1} - \log(\pi_0) - \tfrac{\hat{\pi}_0-\pi_0}{\pi_0}.\end{aligned}$$

Consequently, the variance of $\log(\hat{r})$ satisfies

$$\begin{aligned}\mathrm{Var}[\log(\hat{r})] &\approx \mathrm{Var}\left(\tfrac{\hat{\pi}_0-\pi_0}{\pi_0}\right) + \mathrm{Var}\left(\tfrac{\hat{\pi}_1-\pi_1}{\pi_1}\right)\\ &= \tfrac{\mathrm{Var}(\hat{\pi}_0)}{\pi_0^2} + \tfrac{\mathrm{Var}(\hat{\pi}_1)}{\pi_1^2}\\ &= \tfrac{\pi_0(1-\pi_0)}{n_0\pi_0^2} + \tfrac{\pi_1(1-\pi_1)}{n_1\pi_1^2}\\ &= \tfrac{1-\pi_0}{n_0\pi_0} + \tfrac{1-\pi_1}{n_1\pi_1}.\end{aligned} \tag{7}$$

c. The standard error of $\log(\hat{r})$ is obtained by plugging in estimates of π_0 and π_1 into the variance formula (7) and then taking the square root. That is,

$$\begin{aligned}\mathrm{SE}[\log(\hat{r})] &= \sqrt{\widehat{\mathrm{Var}}[\log(\hat{r})]}\\ &= \sqrt{\tfrac{1-\hat{\pi}_0}{n_0\hat{\pi}_0} + \tfrac{1-\hat{\pi}_1}{n_1\hat{\pi}_1}}\\ &= \sqrt{\tfrac{n_{00}}{n_0 n_{10}} + \tfrac{n_{01}}{n_1 n_{11}}}\\ &= \sqrt{\tfrac{45}{50\cdot 5} + \tfrac{25}{50\cdot 25}}\\ &= \sqrt{0.2}\\ &= 0.4472.\end{aligned}$$

This gives a Wald-based confidence interval

$$\begin{aligned}&(\log(\hat{r}) - 1.96\cdot\mathrm{SE}[\log(\hat{r})], \log(\hat{r}) + 1.96\cdot\mathrm{SE}[\log(\hat{r})])\\ &= (\log(5) - 1.96\cdot 0.4472, \log(5) + 1.96\cdot 0.4472)\\ &= (0.733, 2.486)\end{aligned}$$

for $\log(r)$ with approximate coverage probability 95%. Applying the exponential transformation to both sides of this interval, we finally obtain a confidence interval

$$(\exp(0.733), \exp(2.486)) = (2.081, 12.013)$$

for r with approximate coverage probability 95%. Since 1 does not belong to this interval, the null hypothesis $r = 1$ (or equivalently $\Delta = 0$) is rejected, as in Problem 2a).

Problem 3

a. The loglinear parametrization of (XY, YZ) is

$$\mu_{ijk} = \exp(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}) \tag{8}$$

for $1 \le i, j, k \le 2$. Assume that $X = 1$, $Y = 1$, and $Z = 1$ are chosen as baseline levels. Then those loglinear parameters are put to zero for which at least one index i, j or k equals 1. The remaining 6 parameters are

$$\boldsymbol{\beta} = (\lambda, \lambda_2^X, \lambda_2^Y, \lambda_2^Z, \lambda_{22}^{XY}, \lambda_{22}^{YZ}). \tag{9}$$

b. One possible solution is to look at the cell probabilities $\pi_{ijk} = \mu_{ijk}/\mu_{+++}$. Since X and Z are conditionally independent given Y for model (XY, YZ), it follows that

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

and hence

$$\mu_{ijk} = \mu_{+++}\pi_{ijk} = \mu_{+++} \cdot \frac{\frac{\mu_{ij+}}{\mu_{+++}} \cdot \frac{\mu_{+jk}}{\mu_{+++}}}{\frac{\mu_{+j+}}{\mu_{+++}}} = \frac{\mu_{ij+}\mu_{+jk}}{\mu_{+j+}}. \tag{10}$$

As an alternative proof, one may start writing (8) as a product $\mu_{ijk} = A_{ij}B_{jk}$ of two terms $A_{ij} = \exp(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY})$ and $B_{jk} = \exp(\lambda_k^Z + \lambda_{jk}^{YZ})$. Then

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

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+} in (10) by estimates n_{ij+}, n_{+jk} and n_{+j+}. From the given marginals of the partial table with $Y = j$ we can read off all n_{ij+}, n_{+jk} and n_{+j+}, for instance

$$\hat{\mu}_{111} = \frac{n_{11+}n_{+11}}{n_{+1+}} = \frac{410 \cdot 420}{810} = 212.59.$$

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

No complications $Y = 1$:

Type of valve	Clinic		Sum
	$Z = 1$	$Z = 2$	
$X = 1$	212.59	197.41	410
$X = 2$	207.41	192.59	400
Sum	420	390	810

Complications $Y = 2$:

Type of valve	Clinic		Sum
	$Z = 1$	$Z = 2$	
$X = 1$	27.37	102.63	130
$X = 2$	12.63	47.37	60
Sum	40	150	190

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

$$\begin{array}{rcl} G^2 & = & 2\sum_{ijk} n_{ijk} \log \frac{n_{ijk}}{\hat{\mu}_{ijk}} \\ & = & 2\left(210 \cdot \log \frac{210}{212.59} + \ldots + 50 \cdot \log \frac{50}{47.37}\right) \\ & = & 1.18 \\ & < & \chi^2_2(0.05) = 5.99, \end{array}$$

where in the last step we used that df $= 8 - 6 = 2$, since the saturated model (XYZ) has $2 \times 2 \times 2 = 8$ parameters, whereas the conditional independence model (XY, YZ) has 6 parameters according to (9). Thus we cannot reject conditional independence between X and Z given Y at level 5%.

Problem 4

a. As in Problem 3, we let $\pi_{ijk} = \mu_{ijk}/\mu_{+++} = P(X = i, Y = j, Z = k)$ refer to cell probabilities, i.e. the joint distribution of all three variables, and $\pi_{i+k} = P(X = i, Z = k)$ to the joint distribution of valve type and clinic. From equation (8) we find that

$$\begin{array}{rcl} \text{logit}[P(Y = 2|X = i, Z = k)] & = & \log[P(Y = 2|X = i, Z = k)/P(Y = 1|X = i, Z = k)] \\ & = & \log[(\pi_{i2k}/\pi_{i+k})/(\pi_{i1k}/\pi_{i+k})] \\ & = & \log(\pi_{i2k}/\pi_{i1k}) \\ & = & \log(\mu_{i2k}/\mu_{i1k}) \\ & = & (\lambda + \lambda^X_i + \lambda^Y_2 + \lambda^Z_k + \lambda^{XY}_{i2} + \lambda^{YZ}_{2k}) \\ & - & (\lambda + \lambda^X_i + \lambda^Y_1 + \lambda^Z_k + \lambda^{XY}_{i1} + \lambda^{YZ}_{1k}) \\ & = & \alpha + \beta^X_i + \beta^Z_k, \end{array} \tag{11}$$

where in the last step we used that

$$\begin{array}{rcl} \alpha & = & \lambda^Y_2 - \lambda^Y_1, \\ \beta^X_i & = & \lambda^{XY}_{i2} - \lambda^{XY}_{i1}, \\ \beta^Z_k & = & \lambda^{YZ}_{2k} - \lambda^{YZ}_{1k}. \end{array} \tag{12}$$

If $X = 1$, $Y = 1$, and $Z = 1$ are chosen as baseline levels for the loglinear model, then any loglinear parameter with $i = 1$, $j = 1$ or $k = 1$ among its indexes is zero. In view of (12), this implies $\beta^X_1 = \beta^Z_1 = 0$. The only remaining parameters are $(\alpha, \beta^X_2, \beta^Z_2) = (\lambda^Y_2, \lambda^{XY}_{22}, \lambda^{YZ}_{22})$.

b. By the definition of the conditional odds ratio and (11), we have that

$$\begin{array}{rcl} \theta_{XY(k)} & = & [P(Y = 2|X = 2, Z = k)/P(Y = 1|X = 2, Z = k)] \\ & & /\,[P(Y = 2|X = 1, Z = k)/P(Y = 1|X = 1, Z = k)] \\ & = & \exp(\alpha + \beta^X_2 + \beta^Z_k)/\exp(\alpha + \beta^X_1 + \beta^Z_k) \\ & = & \exp(\beta^X_2 - \beta^X_1) \\ & = & \exp(\beta^X_2), \end{array} \tag{13}$$

since $\beta^X_1 = 0$. Alternatively, we use that

$$\begin{array}{rcl} \log(\theta_{XY(k)}) & = & \text{logit}P(Y = 2|X = 2, Z = k) - \text{logit}P(Y = 2|X = 1, Z = k) \\ & = & (\alpha + \beta^X_2 + \beta^Z_k) - (\alpha + \beta^X_1 + \beta^Z_k) \\ & = & \beta^X_2. \end{array}$$

There is homogeneous association, since $\theta_{XY(k)}$ does not depend on the value k of the clinic Z. This also follows from the fact that there is no third order association XYZ between all three variables in the loglinear model.

c. Write $\pi(i,k) = P(Y = 2|X = i, Z = k)$ for the probability that a patient who has a valve of type i replaced at clinic k has complications within one year after the surgery. The average causal effect (ACE) of type of valve is

$$\begin{aligned}\text{ACE} &= \tfrac{N_{++1}}{N_{+++}}[\pi(2,1) - \pi(1,1)] + \tfrac{N_{++2}}{N_{+++}}[\pi(2,2) - \pi(1,2)] \\ &= \tfrac{460}{1000}\left[\tfrac{\exp(\alpha+\beta_2^X)}{1+\exp(\alpha+\beta_2^X)} - \tfrac{\exp(\alpha)}{1+\exp(\alpha)}\right] \\ &+ \tfrac{540}{1000}\left[\tfrac{\exp(\alpha+\beta_2^X+\beta_2^Z)}{1+\exp(\alpha+\beta_2^X+\beta_2^Z)} - \tfrac{\exp(\alpha+\beta_2^Z)}{1+\exp(\alpha+\beta_2^Z)}\right].\end{aligned}$$

As opposed to $\theta_{XY(k)}$, it depends not only on β_2^X, but also on α and β_2^Z.

Problem 5

a. Since the two rows of the contingency table have independent binomial distributions, it follows that the likelihood of data is

$$\begin{aligned} l(\alpha,\beta) &= \tbinom{n_0}{n_{01}}\pi_0^{n_{01}}(1-\pi_0)^{n_0-n_{01}} \cdot \tbinom{n_1}{n_{11}}\pi_1^{n_{11}}(1-\pi_1)^{n_1-n_{11}} \\ &= \tbinom{n_0}{n_{01}}\left(\tfrac{\pi_0}{1-\pi_0}\right)^{n_{01}}(1-\pi_0)^{n_0} \cdot \tbinom{n_1}{n_{11}}\left(\tfrac{\pi_1}{1-\pi_1}\right)^{n_{11}}(1-\pi_1)^{n_1} \\ &= c \cdot \exp(n_{01}\alpha)(1-\pi_0)^{n_0} \cdot \exp(n_{11}(\alpha+\beta))(1-\pi_1)^{n_1}, \end{aligned}$$

where $c = \binom{n_0}{n_{01}} \cdot \binom{n_1}{n_{11}}$ is a constant, not depending on α or β. Taking the logarithm we obtain a log likelihood

$$\begin{aligned} L(\alpha,\beta) &= C + n_{01}\alpha + n_0\log(1-\pi_0) + n_{11}(\alpha+\beta) + n_1\log(1-\pi_1) \\ &= C + n_{01}\alpha - n_0\log[1+\exp(\alpha)] + n_{11}(\alpha+\beta) - n_1\log[1+\exp(\alpha+\beta)], \end{aligned} \tag{14}$$

where $C = \log(c)$.

b. By differentiating (14) with respect to α and β, we find that the likelihood score vector has components

$$\begin{aligned} \partial L(\alpha,\beta)/\partial\alpha &= n_{01} - n_0\pi_0 + n_{11} - n_1\pi_1, \\ \partial L(\alpha,\beta)/\partial\beta &= n_{11} - n_1\pi_1, \end{aligned} \tag{15}$$

using the fact that $d\log[1+\exp(x)]/dx = \exp(x)/[1+\exp(x)]$. Differentiating (15) with respect to α and β we find that

$$\begin{aligned} \partial^2 L(\alpha,\beta)/\partial^2\alpha &= -n_0\pi_0(1-\pi_0) - n_1\pi_1(1-\pi_1) =: -a, \\ \partial^2 L(\alpha,\beta)/\partial\alpha\partial\beta &= -n_1\pi_1(1-\pi_1) =: -b, \\ \partial^2 L(\alpha,\beta)/\partial^2\beta &= -n_1\pi_1(1-\pi_1) =: -b, \end{aligned} \tag{16}$$

where $d[\exp(x)/(1+\exp(x))]/dx = \exp(x)/[1+\exp(x)]^2$ was used. In conclusion, the second derivative matrix of the log likelihood is

$$\frac{d^2L(\alpha,\beta)}{d(\alpha,\beta)^2} = -\begin{pmatrix} a & b \\ b & b \end{pmatrix}, \tag{17}$$

with a and b as in (16).

c. Since the Hessian matrix in (17) does not depend on data, the Fisher information matrix equals

$$\boldsymbol{J}(\alpha,\beta) = -E\left[\frac{d^2L(\alpha,\beta)}{d(\alpha,\beta)^2}\right] = -\frac{d^2L(\alpha,\beta)}{d(\alpha,\beta)^2} = \begin{pmatrix} a & b \\ b & b \end{pmatrix}.$$

Inverting this matrix and using the hint, we find an approximation

$$\mathrm{Cov}(\hat{\alpha},\hat{\beta}) \approx \boldsymbol{J}(\alpha,\beta)^{-1} = \begin{pmatrix} a & b \\ b & b \end{pmatrix}^{-1} = \frac{1}{(a-b)b}\begin{pmatrix} b & -b \\ -b & a \end{pmatrix}$$

of the covariance matrix of $(\hat{\alpha},\hat{\beta})$. From the second diagonal element of this matrix we obtain an approximation

$$\begin{aligned} \mathrm{Var}(\hat{\beta}) &\approx \tfrac{a}{(a-b)b} = \tfrac{n_0\pi_0(1-\pi_0)+n_1\pi_1(1-\pi_1)}{n_0\pi_0(1-\pi_0)\cdot n_1\pi_1(1-\pi_1)} \\ &= \tfrac{1}{n_0\pi_0(1-\pi_0)} + \tfrac{1}{n_1\pi_1(1-\pi_1)} \\ &= \tfrac{1}{n_0(1-\pi_0)} + \tfrac{1}{n_0\pi_0} + \tfrac{1}{n_1(1-\pi_1)} + \tfrac{1}{n_1\pi_1} \end{aligned} \tag{18}$$

of the variance of $\hat{\beta}$ (the estimated log odds ratio).

d. The standard error of $\hat{\beta}$ is obtained by first replacing π_i by estimates $\hat{\pi}_i = n_{i1}/n_i$ for $i = 0, 1$, in the variance formula (18), and then taking the square root. This gives

$$\begin{aligned} \mathrm{SE}(\hat{\beta}) &= \sqrt{\tfrac{1}{n_0(1-\hat{\pi}_0)} + \tfrac{1}{n_0\hat{\pi}_0} + \tfrac{1}{n_1(1-\hat{\pi}_1)} + \tfrac{1}{n_1\hat{\pi}_1}} \\ &= \sqrt{\tfrac{1}{n_{00}} + \tfrac{1}{n_{01}} + \tfrac{1}{n_{10}} + \tfrac{1}{n_{11}}}. \end{aligned}$$