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

MT 5022
SOLUTIONS
January 7 2026

Solutions for Examination Classification and Analysis of Categorical Data, January 7, 2026

Problem 1

a. The scheme is independent binomial sampling for the two rows of the 2 × 2 table of cell counts N_{ij}. That is, the number of correct guesses $N_{i1} \sim \text{Bin}(9, \pi_{1|i})$ made by person i, is binomially distributed. Since N_{01} and N_{11} are independent, we get a likelihood function

$$l(\pi_{1|0}, \pi_{1|1}) = \binom{9}{5}(1-\pi_{1|0})^4\pi_{1|0}^5 \cdot \binom{9}{7}(1-\pi_{1|1})^2\pi_{1|1}^7 = 4536(1-\pi_{1|0})^4\pi_{1|0}^5(1-\pi_{1|1})^2\pi_{1|1}^7.$$

b. The null hypothesis that the expert and amateur are equally skilled in selecting the correct pieces of art, corresponds to

$$H_0 : \pi_{1|0} = \pi_{1|1} \Longleftrightarrow r = 1,$$

where

$$r = \frac{\pi_{1|1}/(\pi_{0|1} + \pi_{1|1})}{\pi_{1|0}/(\pi_{0|0} + \pi_{1|0})} = \frac{\pi_{1|1}}{\pi_{1|0}}$$

is the risk ratio between the expert's and amateur's probabilities of selecting a correct painting.

c. When we condition of column sums as well, we get a hypergeometric distribution under H_0 for N_{11}, the number of correct guesses made by the expert. More specifically,

$$\begin{aligned} P_{H_0}(N_{11} = n_{11}|N_{+1} = 12) &= P(N_{01} = 12 - n_{11}, N_{11} = n_{11}|N_{+1} = 12) \\ &= \tbinom{9}{12-n_{11}}\tbinom{9}{n_{11}}/\tbinom{18}{12}, \end{aligned} \tag{1}$$

for $n_{11} = 3, 4, \ldots, 9$. Notice that we only conditioned on one column sum $N_{+1} = 12$ in (1). Since the total cell count $n_{++} = 18$ is known, this is equivalent to conditioning on $N_{+0} = 18 - 12 = 6$ as well.

d. We will test H_0 against the one-sided alternative hypothesis

$$H_a : \pi_{1|1} > \pi_{1|0}$$

that the expert does better than the amateur. This corresponds to rejecting H_0 for large values of N_{11}. It follows from (1) and the given hint, that

$$P_{H_0}(N_{11} = 7|N_{+1} = 12) = \frac{\binom{9}{5}\binom{9}{7}}{\binom{18}{12}} = \frac{126 \cdot 36}{18564} = 0.2443,$$

$$P_{H_0}(N_{11} = 8|N_{+1} = 12) = \frac{\binom{9}{4}\binom{9}{8}}{\binom{18}{12}} = \frac{126 \cdot 9}{18564} = 0.0611,$$

and

$$P_{H_0}(N_{11} = 9|N_{+1} = 12) = \frac{\binom{9}{3}\binom{9}{9}}{\binom{18}{12}} = \frac{84 \cdot 1}{18564} = 0.0045.$$

This gives a

$$\begin{aligned} P - \text{value} &= 0.2443 + 0.0611 + 0.0045 = 0.3099, \\ \text{mid } P\text{-value} &= 0.5 \cdot 0.2443 + 0.0611 + 0.0045 = 0.1878. \end{aligned}$$

Suppose we reject H_0 if either the P-value or the mid P-value is at most 0.05. Then we cannot reject the null hypothesis that the expert and amateur are equally good in terms of distinguishing original art, by any of these two tests. (Neither of these two tests will have 0.05 as significance level though. For the P-value based test, we reject H_0 when $N_{11} \geq 9$, with a significance level of $P_{H_0}(N_{11} = 9|N_{+1} = 12) = 0.0045 < 0.05$. For the mid P-value based test, we reject H_0 when $N_{11} \geq 8$, with a significance level of $P_{H_0}(N_{11} = 8|N_{+1} = 12) + P_{H_0}(N_{11} = 9|N_{+1} = 12) = 0.0611 + 0.0045 = 0.0656 > 0.05$. We conclude that the P-value based test is conservative, whereas the mid P-value based test is anti-conservative.)

Problem 2

a. If multinomial sampling is used, then

$$\boldsymbol{N} = (N_{11}, N_{12}, N_{21}, N_{22}) \sim \text{Mult}(n; \pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$$

has a multinomial distribution. The probability of observing $N_{ij} = n_{ij}$ for all cells $1 \leq i, j \leq 2$, is

$$P((N_{11}, N_{12}, N_{21}, N_{22}) = (n_{11}, n_{12}, n_{21}, n_{22})) = \frac{n!}{n_{11}!n_{12}!n_{21}!n_{22}!}\pi_{11}^{n_{11}}\pi_{12}^{n_{12}}\pi_{21}^{n_{21}}\pi_{22}^{n_{22}}.$$

Since the four cell probabilities sum to 1 ($\sum_{ij} \pi_{ij} = 1$) there are only three free parameters in the model. Putting $\pi_{22} = 1 - \pi_{11} - \pi_{12} - \pi_{21}$, we get a likelihood function

$$l(\pi_{11}, \pi_{12}, \pi_{21}) = \frac{n!}{n_{11}!n_{12}!n_{21}!n_{22}!}\pi_{11}^{n_{11}}\pi_{12}^{n_{12}}\pi_{21}^{n_{21}}(1 - \pi_{11} - \pi_{12} - \pi_{21})^{n_{22}}.$$

b. The odds that a person from region i is satisfied with the job, is

$$\frac{\pi_{i1}/\pi_{i+}}{\pi_{i2}/\pi_{i+}} = \frac{\pi_{i1}}{\pi_{i2}}.$$

This gives an odds ratio

$$\mathrm{OR} = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \tag{2}$$

of job satisfaction between regions 1 and 2. Plugging in estimates $\hat{\pi}_{ij} = n_{ij}/n$ into (2) we get an estimate

$$\widehat{\mathrm{OR}} = \frac{\hat{\pi}_{11}\hat{\pi}_{22}}{\hat{\pi}_{12}\hat{\pi}_{21}} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{133 \cdot 36}{52 \cdot 120} = 0.7673$$

of OR. The standard error of $\log(\widehat{\mathrm{OR}})$ is an estimate of its approximate standard deviation, i.e.

$$\begin{aligned}\mathrm{SE} &= \sqrt{1/(n\hat{\pi}_{11}) + 1/(n\hat{\pi}_{12}) + 1/(n\hat{\pi}_{21}) + 1/(n\hat{\pi}_{22})}\\ &= \sqrt{1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22}}\\ &= \sqrt{1/133 + 1/52 + 1/36 + 1/120}\\ &= \sqrt{0.0629}\\ &= 0.2507.\end{aligned}$$

This gives a confidence interval

$$\begin{aligned} & (\log(0.7673) - 1.96 \cdot \mathrm{SE}, \log(0.7673) + 1.96 \cdot \mathrm{SE})\\ = & (-0.2649 - 1.96 \cdot \mathrm{SE}, -0.2649 + 1.96 \cdot \mathrm{SE})\\ = & (-0.7563, 0.2265)\end{aligned}$$

for log(OR) with approximate confidence level 95%, and a corresponding confidence interval

$$I = (\exp(-0.7563), \exp(0.2265)) = (0.4694, 1.2543) \tag{3}$$

for OR. Since $1 \in I$, there is no significant job satisfaction difference between the two regions, although the point estimate indicated that the employees in region 2 were a bit more satisfied.

c. The probabilities of concordant or discordant pairs, $\{(1,1),(2,2)\}$ and $\{(1,2),(2,1)\}$ respectively, can be written in terms of the cell probabilities π_{ij}, according to

$$\begin{aligned}\Pi_c &= 2\pi_{11}\pi_{22},\\ \Pi_d &= 2\pi_{12}\pi_{21}.\end{aligned}$$

This follows if we think of the company as large. Then choosing two individuals is effectively like picking two balls with replacement from an urn that has four types of balls, with probabilities $\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}$ of drawing different kinds of balls. From this and (2) we find that

$$\gamma = \frac{2\pi_{11}\pi_{22} - 2\pi_{12}\pi_{21}}{2\pi_{11}\pi_{22} + 2\pi_{12}\pi_{21}} = \frac{\mathrm{OR} - 1}{\mathrm{OR} + 1}, \tag{4}$$

Since γ is a monotone increasing function of OR, we obtain a confidence interval

$$\left(\frac{0.4694-1}{0.4694+1}, \frac{1.2543-1}{1.2543+1}\right) = (-0.3611, 0.1128)$$

for γ by transforming the end points of (3) with the transformation we found in (4). The two intervals will therefore have the same coverage probability of the true values of OR and γ respectively.

Problem 3

a. The loglinear parametrization of (SI, LI) is

$$\mu_{sil} = \exp(\lambda + \lambda_s^S + \lambda_l^L + \lambda_i^I + \lambda_{si}^{SI} + \lambda_{li}^{LI}) \tag{5}$$

for $0 \le s, l, i \le 1$. Assume that $S = 0$, $L = 0$, and $I = 0$ are chosen as baseline levels. Then those loglinear parameters are put to zero for which at least one index s, l or i equals 0. The remaining 6 parameters are

$$\boldsymbol{\beta} = (\lambda, \lambda_1^S, \lambda_1^L, \lambda_1^I, \lambda_{11}^{SI}, \lambda_{11}^{LI}). \tag{6}$$

b. One possible solution is to look at the cell probabilities $\pi_{sli} = \mu_{sli}/\mu_{+++}$. Since S and L are conditionally independent given I for model (SI, LI), it follows that

$$\pi_{sli} = \pi_{++i}\pi_{sl|i} = \pi_{++i}\pi_{s+|i}\pi_{l+|i} = \pi_{++i} \cdot \frac{\pi_{s+i}}{\pi_{++i}} \cdot \frac{\pi_{+li}}{\pi_{++i}} = \frac{\pi_{s+i}\pi_{+li}}{\pi_{++i}},$$

and hence

$$\mu_{sli} = \mu_{+++}\pi_{sli} = \mu_{+++} \cdot \frac{\frac{\mu_{s+i}}{\mu_{+++}} \cdot \frac{\mu_{+li}}{\mu_{+++}}}{\frac{\mu_{++i}}{\mu_{+++}}} = \frac{\mu_{s+i}\mu_{+li}}{\mu_{++i}}. \tag{7}$$

c. The maximum likelihood estimates

$$\hat{\mu}_{sli} = \frac{n_{s+i}n_{+li}}{n_{++i}}$$

of the expected cell counts are obtained by replacing μ_{s+i}, μ_{+li} and μ_{++i} in (7) by estimates n_{si+}, n_{+li} and n_{++i}. From the given marginals of the partial table with $I = i$ we can read off all n_{s+i}, n_{+li} and n_{++i}, for instance

$$\hat{\mu}_{000} = \frac{n_{0+0}n_{+00}}{n_{++0}} = \frac{16504 \cdot 21350}{34166} = 10313.$$

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

No injury $I = 0$:

Seat	Location		
belt use	$L = 0$	$L = 1$	Sum
$S = 0$	10313	6191	16504
$S = 1$	11037	6625	17662
Sum	21350	12816	34166

Injury $I = 1$:

Seat	Location		
belt use	$L = 0$	$L = 1$	Sum
$S = 0$	810.3	1085.7	1896
$S = 1$	381.7	511.3	893
Sum	1192	1597	2789

d. The log likelihood ratio statistic for testing (SI, LI) against the saturated model (SIL), is

$$\begin{aligned} G^2 &= 2\textstyle\sum_{sli} n_{sli} \log \frac{n_{sli}}{\hat{\mu}_{sli}} \\ &= 2\left(10381 \cdot \log \frac{10381}{10313} + \ldots + 513 \cdot \log \frac{513}{511.3}\right) \\ &= 2.318 \\ &< \chi_2^2(0.05) = 5.99, \end{aligned}$$

where in the last step we used that df $= 8 - 6 = 2$, since the saturated model (SLI) has $2 \times 2 \times 2 = 8$ parameters, whereas the conditional independence model (SI, LI) has 6 parameters according to (6). Thus we cannot reject conditional independence between S and L given I at level 5%.

Problem 4

a. As in Problem 3, we let $\pi_{sli} = \mu_{sli}/\mu_{+++} = P(S = s, L = l, I = i)$ refer to cell probabilities, i.e. the joint distribution of all three variables, and $\pi_{sl+} = P(S = s, L = l)$ to the joint distribution of seat belt use and location. From equation (5) we find that

$$\begin{aligned} \text{logit}[P(I = 1|S = i, L = l)] &= \log[P(I = 1|S = s, L = l)/P(I = 0|S = s, L = l)] \\ &= \log[(\pi_{sl1}/\pi_{sl+})/(\pi_{sl0}/\pi_{sl+})] \\ &= \log(\pi_{sl1}/\pi_{sl0}) \\ &= \log(\mu_{sl1}/\mu_{sl0}) \\ &= (\lambda + \lambda_s^S + \lambda_l^L + \lambda_1^I + \lambda_{s1}^{SI} + \lambda_{l1}^{LI}) \\ &- (\lambda + \lambda_s^S + \lambda_l^L + \lambda_0^I + \lambda_{s0}^{SI} + \lambda_{l0}^{LI}) \\ &= \alpha + \beta_s^S + \beta_l^L, \end{aligned} \tag{8}$$

where in the last step we used that

$$\begin{aligned} \alpha &= \lambda_1^I - \lambda_0^I, \\ \beta_s^S &= \lambda_{s1}^{SI} - \lambda_{s0}^{SI}, \\ \beta_l^L &= \lambda_{l1}^{LI} - \lambda_{l0}^{LI}. \end{aligned} \tag{9}$$

If $I = 0$, $S = 0$, and $L = 0$ are chosen as baseline levels for the loglinear model, then any loglinear parameter with $i = 0$, $s = 0$ or $l = 0$ among its indexes is zero. In view of (9), this implies $\beta_0^S = \beta_0^L = 0$. The only remaining parameters are $(\alpha, \beta_1^S, \beta_1^L) = (\lambda_1^I, \lambda_{11}^{SI}, \lambda_{11}^{LI})$.

b. By the definition of the conditional odds ratio and (8), we have that

$$\begin{aligned} \theta_{SI(l)} &= [P(I = 1|S = 1, L = l)/P(I = 0|S = 1, L = l)] \\ &\quad / [P(I = 1|S = 0, L = l)/P(I = 0|S = 0, L = l)] \\ &= \exp(\alpha + \beta_1^S + \beta_l^L)/\exp(\alpha + \beta_0^S + \beta_l^L) \\ &= \exp(\beta_1^S - \beta_0^S) \\ &= \exp(\beta_1^S), \end{aligned} \tag{10}$$

since $\beta_0^S = 0$. Alternatively, we use that

$$\begin{aligned} \log(\theta_{SI(l)}) &= \text{logit}P(I = 1|S = 1, L = l) - \text{logit}P(I = 1|S = 0, L = l) \\ &= (\alpha + \beta_1^S + \beta_l^L) - (\alpha + \beta_0^S + \beta_l^L) \\ &= \beta_1^S. \end{aligned}$$

There is homogeneous association, since $\theta_{SI(l)}$ does not depend on the value l of the location variable L. This also follows from the fact that there is no third order association SLI between all three variables in the loglinear model.

c. From the Bayes' Theorem we find that

$$\begin{aligned} P(I = i|S = s, L = l) &= \tfrac{P(I=i|L=l)P(S=s|I=i,L=l)}{P(S=s|L=l)} \\ &= \tfrac{P(I=i|L=l)P(S=s|I=i)}{P(S=s|L=l)}, \end{aligned} \tag{11}$$

where in the last step we used the conditional independence of S and L given I. Inserting (11) into the upper two lines of (10), we notice that all terms $P(I = i|L = l)$ and $P(S = s|L = l)$ cancel out. Consequently,

$$\theta_{SI(l)} = \frac{P(S = 1|I = 1)/P(S = 1|I = 0)}{P(S = 0|I = 1)/P(S = 0|I = 1)}. \tag{12}$$

The right hand side of (12) is the marginal odds ratio θ_{SI} between S and I. In order to reformulate θ_{IS}, as stated in the problem, we apply Bayes' Theorem again; $P(S = s|I = i) = P(I = i|S = s)P(S = s)/P(I = i)$. If we insert this expression into (12), all terms $P(S = s)$ and $P(I = i)$ will cancel out. This gives

$$\theta_{SI(l)} = \frac{P(I = 1|S = 1)/P(I = 0|S = 1)}{P(I = 1|S = 0)/P(I = 1|S = 0)},$$

which is the marginal odds ratio θ_{SI} of having an injury between people who use seat belt and not.

d. Since $P(I = i|S = s) = P(I = i, S = s)/P(S = s) = \pi_{s+i}/\pi_{s++}$, the marginal odds ratio in (12) can be rewritten as

$$\theta_{SI} = \frac{\pi_{1+1}\pi_{0+0}}{\pi_{0+1}\pi_{1+0}}.$$

The cell probabilities π_{sli} can be estimated by $\hat{\pi}_{sli} = n_{sli}/n$, and the marginal cell probabilities for seat belt use and injury, by $\hat{\pi}_{s+i} = n_{s+i}/n$, where $n = n_{+++}$ is the total cell count. If we plug the marginal cell probability estimates into (12), we obtain an estimate

$$\begin{aligned} \hat{\theta}_{SI} &= n_{0+0}n_{1+1}/(n_{0+1}n_{1+0}) \\ &= 16504 \cdot 893/(1896 \cdot 17662) \\ &= 0.44 \end{aligned}$$

of the marginal odds ratio. That is, according to this estimate the odds of injury among passengers is lowered to 44% when they start using seat belt.