STOCKHOLM UNIVERSITY
DEPT OF MATHEMATICS
Div. of Mathematical statistics

MT 5006
EXAMINATION
January 13 2023

Categorical Data Analysis – Examination

January 13, 2023, 9.00-14.00

Examination by: Ola Hössjer, ph. 070 672 12 18, ola@math.su.se
Allowed to use: Miniräknare/pocket calculator and tables included in the appendix of this exam.
Grading: Each correct solution to an exercise yields 10 points.
Limits for grade: A, B, C, D, and E are 45, 40, 35, 30, and 25 points of 60 possible points (including bonus of 0-10 points from computer assignments).

Reasoning and notation should be clear. You might answer in Swedish or English.

Start by reading through the whole exam. Exercises need not to be ordered from simpler to harder.

Problem 1

An investigator visited a school campus. During one hour he asked students who passed by to fill in a questionnaire about their habits regarding which type of movies they watched. Each question had three alternatives as to whether a person never (0), sometimes (1) or frequently (2) watched movies related to a certain theme. A total of 102 students responded, and the result of the first two questions, fantasy and sports, is summarized in the following 3×3 table:

	Sports (j)			
Fantasy (i)	0	1	2	Total
0	3	2	12	17
1	8	14	7	29
2	20	33	3	56
Total	31	49	22	102

a. Formulate an appropriate model for the data, and in particular a null hypothesis H_0 of independence between fantasy and sport watching habits. (Hint: Let μ_{ij} be the expected number of observations in cell (i, j). It is helpful to introduce probabilities μ_{ij}/μ_{++}, where μ_{++} is the sum of all μ_{ij}.) (2p)

b. Compute for $0 \leq i, j \leq 2$ the predicted number $\hat{\mu}_{ij}$ of students that would answer i and j under H_0. (2p)

c. Perform a chisquare test of H_0. Is the null hypothesis rejected at level 5%? (3p)

d. In order to investigate the direction of dependence, compute

$$\hat{\gamma} = \frac{C - D}{C + D}, \tag{1}$$

where C and D are the number of concordant and discortand pairs of students. What is your conclusion? (3p)

Problem 2

Assume that 0 and 1 are merged into one category (=1) in Problem 1, for fantacy as well as sports. This gives a condensed 2×2 table of students that never/sometimes (=1) or frequently (=2) watch fantacy and sports, with cell counts $\tilde{N}_{ij}$ for $1 \leq i, j \leq 2$. (A tilde refers to a quantity of the condensed table.)

a. Based on data from the condensed table, compute an estimate $\hat{\theta}$ of the odds ratio θ between fantacy and sports. (2p)

b. Assume that the number of students $\tilde{N}_{ij} \sim \text{Po}(\tilde{\mu}_{ij})$ are independent and Poisson distributed random variables for $1 \leq i, j \leq 2$. Prove that approximately

$$\text{Var}[\log(\hat{\theta})] = \frac{1}{\tilde{\mu}_{11}} + \frac{1}{\tilde{\mu}_{12}} + \frac{1}{\tilde{\mu}_{21}} + \frac{1}{\tilde{\mu}_{22}}. \tag{2}$$

(Hint: You may use, without proof, the result

$$\log(\hat{\theta}) \approx \log\theta + \frac{\tilde{N}_{11} - \tilde{\mu}_{11}}{\tilde{\mu}_{11}} + \frac{\tilde{N}_{22} - \tilde{\mu}_{22}}{\tilde{\mu}_{22}} - \frac{\tilde{N}_{12} - \tilde{\mu}_{12}}{\tilde{\mu}_{12}} - \frac{\tilde{N}_{21} - \tilde{\mu}_{21}}{\tilde{\mu}_{21}}$$

of a first order Taylor expansion of $\log(\hat{\theta})$ around $\log(\theta)$.) (3p)

c. Use the result in 2b to compute an approximate 95% confidence interval for the odds ratio θ in 2a. (Hint: Start to find a confidence interval of $\log(\theta)$, and estimate all $\tilde{\mu}_{ij}$ in (2) by $\tilde{n}_{ij}$, the observed value of $\tilde{N}_{ij}$.) (4p)

d. Discuss how reliable the interval in 2c is. (1p)

Problem 3

It is known that one of the risk factors of a certain disease that predominantly occurs among men, is a gene on the Y-cromosome. This gene has one normal ($Z = 0$) and one risk increasing variant ($Z = 1$). In order to test if there are other genetic risk factors, disease status X and Y was investigated for 1000 pairs of fathers and sons, with $X = 1$ (0) for an affected (unaffected) father, and similarly $Y = 1$ (0) for an affected (unaffected) son. A father and son always share the same variant $Z = 0$ or 1, since the son inherits genes on the Y-chromosome from his father. The result of the study was:

Genetic variant $Z = 0$:

Father's aff status	Son's aff status $Y = 0$	$Y = 1$	Sum
$X = 0$	841	27	868
$X = 1$	30	4	34
Sum	871	31	902

Genetic variant $Z = 1$:

Father's aff status	Son's aff status $Y = 0$	$Y = 1$	Sum
$X = 0$	27	22	49
$X = 1$	20	29	49
Sum	47	51	98

a. Estimate the marginal odds ratio θ_{XY} between the father's and son's affection status, as well as the two conditional odds ratios $\theta_{XY(k)}$ for each variant $k = 0, 1$ of the gene. Based on this, discuss if there seems to be other genetic risk factors. (2p)

b. Assume that the number of father-son pairs N_{ijk} with $X = i$, $Y = j$ and $Z = k$ are independent Poisson random variables, with expected values μ_{ijk}. In order to test whether Z is the only genetic risk factor, we will consider the loglinear model (XZ, YZ). Express μ_{ijk} in terms of the loglinear parameters. After setting some loglinear parameters to zero in order to avoid overparametrization, which ones remain? (2p)

c. Use the result in 3b to prove that $\mu_{ijk} = \mu_{i+k}\mu_{+jk}/\mu_{++k}$. (2p)

d. Use 3c to find ML estimates $\hat{\mu}_{ijk}$ of the expected counts for all cells (i, j, k). (Hint: The row sums, column sums and total number of observations of each partial table for $Z = 0$ and $Z = 1$ will be helpful.) (2p)

e. Perform an LR test (at level 5%) in order to investigate whether (XZ, YZ) adequately describes data. (2p)

Problem 4

In order to incorporate the possibility of other shared risk factors in Problem 3, we will look at the loglinear model (XY, XZ, YZ) which has all second order interactions included, and the lowest level of each variable chosen as baseline.

a. Give the loglinear parametrization of (XY, XZ, YZ) by adding one more interaction term compared to Problem 3b. (1p)

b. In order to predict the son's affection staus based on his gene variant Z and the father's affection status X, consider an ANOVA type logistic regression model with Y as outcome. Show that

$$\text{logit}\left[P(Y = 1|X = i, Z = k)\right] = \alpha + \beta_i^X + \beta_k^Z, \tag{3}$$

and in particular write α, β_i^X and β_k^Z as functions of the loglinear parameters. Also show that the parameter vector is $(\alpha, \beta_1^X, \beta_1^Z)$ if $X = 0$ and $Z = 0$ are chosen as baseline levels for the loglinear parameters. (3p)

c. A data analyis gave parameter estimates

$$(\hat{\alpha}, \hat{\beta}_1^X, \hat{\beta}_1^Z) = (-3.3818, 0.8347, 3.0497),$$

and estimated covariances

$$\begin{pmatrix} \widehat{\mathrm{Var}}(\hat{\alpha}) & \widehat{\mathrm{Cov}}(\hat{\alpha}, \hat{\beta}_1^X) & \widehat{\mathrm{Cov}}(\hat{\alpha}, \hat{\beta}_1^Z) \\ \widehat{\mathrm{Cov}}(\hat{\beta}_1^X, \hat{\alpha}) & \widehat{\mathrm{Var}}(\hat{\beta}_1^X) & \widehat{\mathrm{Cov}}(\hat{\beta}_1^X, \hat{\beta}_1^Z) \\ \widehat{\mathrm{Cov}}(\hat{\beta}_1^Z, \hat{\alpha}) & \widehat{\mathrm{Cov}}(\hat{\beta}_1^Z, \hat{\beta}_1^X) & \widehat{\mathrm{Var}}(\hat{\beta}_1^Z) \end{pmatrix} = \begin{pmatrix} 0.0342 & -0.0096 & -0.0295 \\ -0.0096 & 0.1255 & -0.0520 \\ -0.0295 & -0.0520 & 0.0977 \end{pmatrix}$$

respectively. Use this to compute a 95% confidence interval for the conditional odds ratio θ_{XY} between the father and son's affection status. What conclusion can be drawn from this regarding additional common risk factors for the father and son? (Remark: For the chosen model, the conditional odds ratio $\theta_{XY} = \theta_{XY(k)}$ does not depend on the level k of the confounder Z, but it is still not the same as the marginal odds ratio between X and Y.) (3p)

d. Compute a 95% confidence interval for the affection probability $P(Y = 1|X = 1, Z = 0)$ of a son who carries the normal variant 0 and whose father is affected. (3p)

Problem 5

Based on data from 2021, an insurance company wanted to assess accident risks for a certain type of car, in order to define premiums for 2022 in a correct way. They categorized car owners as old and young ($X = 0, 1$) and whether they lived in rural and urban areas ($Z = 0, 1$). Cars with owners of age $X = i$ that live in a $Z = k$ region, experienced a total number

$$N_{ik} \sim \mathrm{Po}(\mu_{ik})$$

of accidents in 2021, where N_{ik} are independent for different (i, k).

a. Write down the log likelihood of data in terms of all μ_{ik}. (2p)

A loglinear model

$$\mu_{ik} = t_{ik} \exp(\lambda + \lambda_i^X + \lambda_k^Z) \tag{4}$$

was used, with offset variables t_{ik}, the total time of exposure (in years) of all cars of type (i, k) during 2021.

b. Which are the baseline levels for age and region if the parameter vector is $\boldsymbol{\beta} = (\lambda, \lambda_1^X, \lambda_1^Z)$? Which parameters in (4) are put to zero? (1p)

c. Derive the likelihood equations for the ML estimates $(\hat{\lambda}, \hat{\lambda}_1^X, \hat{\lambda}_1^Z)$. (Hint: View $(\lambda, \lambda_1^X, \lambda_1^Z)$ as the parameter vector of the log likelihood in 5a. The derivative can still be expressed in terms of all μ_{ik}, using the fact that $\partial \mu_{ik}/\partial \lambda = \mu_{ik}$ and $\partial \log(\mu_{ik})/\partial \lambda = 1$. The partial derivatives with respect to the other parameters are similar but slightly different.) (3p)

d. It is known that an accident with this type of car on average costs 100 kSEK (100 000 Swedish crowns), and it is assumed that accident rates for 2021 and 2022 are the same. Based on this, the insurance company decided to set a premium

$$\hat{P}_{ik} = 110 \cdot \frac{\hat{\mu}_{ik}}{t_{ik}} = 110 \cdot \exp\left(\hat{\lambda} + \hat{\lambda}_i^X + \hat{\lambda}_k^Z\right)$$

for a one year contract of type (i, k) in units of kSEK, in order to cover other expenses and make a profit. Based on the parameter estimates

$$\hat{\boldsymbol{\beta}} = (\hat{\lambda}, \hat{\lambda}_1^X, \hat{\lambda}_1^Z) = (-3.10, 0.25, 0.48),$$

compute the premium $\hat{P}_{10}$ for young drivers that live in a rural area. (2p)

e. Return to the likelihood analysis in 5a-5c, and let

$$\boldsymbol{J}(\boldsymbol{\beta}) = \begin{pmatrix} J_{11}(\boldsymbol{\beta}) & J_{12}(\boldsymbol{\beta}) & J_{13}(\boldsymbol{\beta}) \\ J_{21}(\boldsymbol{\beta}) & J_{22}(\boldsymbol{\beta}) & J_{23}(\boldsymbol{\beta}) \\ J_{31}(\boldsymbol{\beta}) & J_{32}(\boldsymbol{\beta}) & J_{33}(\boldsymbol{\beta}) \end{pmatrix}$$

be the Fisher information matrix, assuming that parameters are numbered as $\lambda = \beta_1$, $\lambda_1^X = \beta_2$ and $\lambda_1^Z = \beta_3$. Describe how $J(\boldsymbol{\beta})$ is obtained from the log likelihood, and compute $J_{33}(\boldsymbol{\beta})$. (2p)

Good luck!

Appendix A - Table for chi-square distribution

Table 1: Quantiles of the chi-square distribution with df $= 1, 2, \ldots, 12$ degrees of freedom

	degrees of freedom											
prob	1	2	3	4	5	6	7	8	9	10	11	12
0.8000	1.64	3.22	4.64	5.99	7.29	8.56	9.80	11.03	12.24	13.44	14.63	15.81
0.9000	2.71	4.61	6.25	7.78	9.24	10.64	12.02	13.36	14.68	15.99	17.28	18.55
0.9500	3.84	5.99	7.81	9.49	11.07	12.59	14.07	15.51	16.92	18.31	19.68	21.03
0.9750	5.02	7.38	9.35	11.14	12.83	14.45	16.01	17.53	19.02	20.48	21.92	23.34
0.9800	5.41	7.82	9.84	11.67	13.39	15.03	16.62	18.17	19.68	21.16	22.62	24.05
0.9850	5.92	8.40	10.47	12.34	14.10	15.78	17.40	18.97	20.51	22.02	23.50	24.96
0.9900	6.63	9.21	11.34	13.28	15.09	16.81	18.48	20.09	21.67	23.21	24.72	26.22
0.9910	6.82	9.42	11.57	13.52	15.34	17.08	18.75	20.38	21.96	23.51	25.04	26.54
0.9920	7.03	9.66	11.83	13.79	15.63	17.37	19.06	20.70	22.29	23.85	25.39	26.90
0.9930	7.27	9.92	12.11	14.09	15.95	17.71	19.41	21.06	22.66	24.24	25.78	27.30
0.9940	7.55	10.23	12.45	14.45	16.31	18.09	19.81	21.47	23.09	24.67	26.23	27.76
0.9950	7.88	10.60	12.84	14.86	16.75	18.55	20.28	21.95	23.59	25.19	26.76	28.30
0.9960	8.28	11.04	13.32	15.37	17.28	19.10	20.85	22.55	24.20	25.81	27.40	28.96
0.9970	8.81	11.62	13.93	16.01	17.96	19.80	21.58	23.30	24.97	26.61	28.22	29.79
0.9980	9.55	12.43	14.80	16.92	18.91	20.79	22.60	24.35	26.06	27.72	29.35	30.96
0.9990	10.83	13.82	16.27	18.47	20.52	22.46	24.32	26.12	27.88	29.59	31.26	32.91
0.9991	11.02	14.03	16.49	18.70	20.76	22.71	24.58	26.39	28.15	29.87	31.55	33.20
0.9992	11.24	14.26	16.74	18.96	21.03	22.99	24.87	26.69	28.46	30.18	31.87	33.53
0.9993	11.49	14.53	17.02	19.26	21.34	23.31	25.20	27.02	28.80	30.53	32.23	33.90
0.9994	11.78	14.84	17.35	19.60	21.69	23.67	25.57	27.41	29.20	30.94	32.65	34.32
0.9995	12.12	15.20	17.73	20.00	22.11	24.10	26.02	27.87	29.67	31.42	33.14	34.82
0.9996	12.53	15.65	18.20	20.49	22.61	24.63	26.56	28.42	30.24	32.00	33.73	35.43
0.9997	13.07	16.22	18.80	21.12	23.27	25.30	27.25	29.14	30.97	32.75	34.50	36.21
0.9998	13.83	17.03	19.66	22.00	24.19	26.25	28.23	30.14	31.99	33.80	35.56	37.30
0.9999	15.14	18.42	21.11	23.51	25.74	27.86	29.88	31.83	33.72	35.56	37.37	39.13