STOCKHOLM UNIVERSITY
DEPT OF MATHEMATICS
Div. of Mathematical statistics

MT 5006
EXAMINATION
January 12 2022

Categorical Data Analysis – Examination

January 12, 2022, 8.00-13.00

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.
Återlämning/Return of exam: Will be communicated on the course homepage and by email upon request.

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.

Read first through the whole exam. Exercises need not to be ordered from simpler to harder.

Problem 1

A new medicine for a certain disease was tested in a clinical trial, which consisted of 18 patients with the disease. A randomly chosen subset of nine patients were given medicine, while the others received a placebo treatment. It was checked one month later whether the treatment improved health status or not, with the following result:

	Improved health?		
Treatment	Yes	No	Total
Medicine	6	3	9
Placebo	3	6	9
Total	9	9	18

a. Define the most appropriate sampling distribution for data and write down the likelihood $l(\pi_1, \pi_2)$ in terms of the probabilities π_1 and π_2 of improved health among those that received medicine and placebo. (3p)

b. Write down the null hypothesis H_0 that the medicine has no effect. (1p)

c. Let N_{ij} refer to the number of observations in row i and column j, $1 \leq i, j \leq 2$. Fisher's exact test uses only N_{11} and is based on a certain conditional distribution $P_{H_0}(N_{11} = n_{11}|\ldots)$, displayed below. Determine the condition (the dots) and write down the formula for this conditional distribution (you don't have to prove it). (2p)

n_{11}	0	1	2	3	4
$P_{H_0}(N_{11} = n_{11}\|\ldots)$	0.0000	0.0017	0.0267	0.1451	0.3265
n_{11}	5	6	7	8	9
$P_{H_0}(N_{11} = n_{11}\|\ldots)$	0.3265	0.1451	0.0267	0.0017	0.0000

d. Define the odds ratio θ of improved health between patients that take the medicine or placebo. Then compute an estimate $\hat{\theta}$ of θ. (2p)

e. Write down the alternative hypothesis H_a that the medicine improves health and compute the corresponding one-sided P-value for the given data set, using Fisher's exact test. (2p)

Problem 2

In a genetic study the objective is to determine whether two genes of a certain plant are located on the same chromosome or not. Gene 1 codes for colour, with variants A and a. Since each plant has two copies of Gene 1 (inherited from each of its two parents) there are three possible combinations - aa=white, Aa=pink and AA=red. In the same way Gene 2 codes for size, with variants b and B, giving rise to bb=small, Bb=medium size and BB=large.

A total of $n = 100$ pairs of plants were crossed and the variants of Gene 1 and 2 for their offspring reported:

	Gene 2			
Gene 1	bb	Bb	BB	Total
aa	3	12	5	20
Aa	14	23	12	49
AA	7	21	3	31
Total	24	56	20	100

Regard an offspring's pair of variants at Gene 1 and 2 as two categorical variables, with levels numbered as $i = 1, 2, 3$ (rows) and $j = 1, 2, 3$ (columns), and denote the observed cell counts by n_{ij}.

a. The experiment can be regarded as multinomial sampling with cell probabilities π_{ij}. The null hypothesis

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$$

that rows and columns are independent corresponds to the two genes belonging to different chromosomes. Compute the chisquare test statistic X^2 for the above data set. What is the approximate distribution of X^2 under H_0? Can H_0 be rejected at level 0.05? (Hint: Table 2 of Appendix B might be of help.) (3p)

b. Due to the nature of the experiment it is known that $\pi_{1+} = \pi_{3+} = \pi_{+1} = \pi_{+3} = 0.25$, and $\pi_{2+} = \pi_{+2} = 0.5$. Use this extra information in order to redefine the chisquare statistic. Is H_0 rejected at level 0.05? (2p)

c. Suppose that colour and size are recessive traits, meaning that neither aa and Aa nor bb and bB can be distinguished. Then two rows and two columns have to merged. Estimate the odds ratio θ of the 2×2 table with cell counts $\tilde{n}_{ij}$ so obtained. (2p)

d. Formulate H_0 for the reduced table, and give a 95% confidence interval for θ, making use of

$$\widehat{\mathrm{Var}}(\log(\hat{\theta})) = \frac{1}{\tilde{n}_{11}} + \frac{1}{\tilde{n}_{12}} + \frac{1}{\tilde{n}_{21}} + \frac{1}{\tilde{n}_{22}}.$$

Comment also on the reliability of this interval. (3p)

Problem 3

A socio-economic study compared the ninth classes of two schools (S) in a city. For each student it was registered whether his or her average grade (G) exceeded a certain threshold or not, as well as the total salary of the parents, dichotomized into an economy variable (E) with three levels. The two tables below summarize data in terms of observed counts n_{egs} for all $e \in \{1, 2, 3\}$ and $g, s \in \{1, 2\}$, numbering the categories of the ordinal variables E and G from lower to higher.

School 1 (128 students):

Economy level	Grade	
	Low	High
Low	15	6
Medium	31	37
High	14	25

School 2 (93 students):

Economy level	Grade	
	Low	High
Low	10	3
Medium	26	22
High	13	19

a. Specify the parameters of the saturated loglinear model $M_1 =$ (EGS), and in particular which of them you put to 0 in order to avoid over-parametrization. (2p)

b. Which of the parameters in a) are included in a loglinear model $M_0 =$ (EG,S) for which school is jointly independent of grade and economic level? (2p)

c. Prove that model M_0 has expected cell counts $\mu_{egs} = \mu_{eg+}\mu_{++s}/\mu_{+++}$. (Hint: You may look at $\pi_{egs} = \mu_{egs}/\mu_{+++}$.) (2p)

d. Formulate and compute the likelihood ratio test statistic for choosing between M_0 and M_1 at level 0.05. Is the null hypothesis of no difference between schools rejected? (Hint: Numbers from Table 3 of Appendix B might be of help.) (4p)

Problem 4

Consider the three-way $3 \times 2 \times 2$ contingency table of Problem 3. Assume that school S and economy level E are predictor variables (with their highest levels used as baseline), and that grade G is a binary outcome variable. We will study the logistic regression model derived from the loglinear model M_0.

a. Write down the probability $P(G = 2|S = s, E = e)$ of a student having high average grade as a certain function of an intercept parameter α and two effect parameters β_1 and β_2. Show in particular how α, β_1 and β_2 are functions of the loglinear parameters of M_0, and discuss which constraints you impose in order to avoid overparametrization. (3p)

b. The maximum likelihood estimates are

$$(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2) = (0.4884, -1.5100, -0.4539),$$

Compute and interpret $\hat{\theta} = \exp(\hat{\beta}_1)$. (2p)

c. The estimated covariance matrix for the three parameters is

$$\begin{pmatrix} \widehat{\text{Var}}(\hat{\alpha}) & \widehat{\text{Cov}}(\hat{\alpha}, \hat{\beta}_1) & \widehat{\text{Cov}}(\hat{\alpha}, \hat{\beta}_2) \\ \widehat{\text{Cov}}(\hat{\alpha}, \hat{\beta}_1) & \widehat{\text{Var}}(\hat{\beta}_1) & \widehat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_2) \\ \widehat{\text{Cov}}(\hat{\alpha}, \hat{\beta}_2) & \widehat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_2) & \widehat{\text{Var}}(\hat{\beta}_2) \end{pmatrix} = \begin{pmatrix} 0.0598 & -0.0598 & -0.0598 \\ -0.0598 & 0.2109 & 0.0598 \\ -0.0598 & 0.0598 & 0.0943 \end{pmatrix}.$$

Use this information in order to obtain a 95% confidence interval for θ. (3p)

d. What is the number of extra degrees of freedom for the saturated logistic regression model for $P(G = 2|S = s, E = e)$, compared to the logistic regression model in Problem 4a)-4c)? (2p)

Problem 5

A random variable Y belongs to the exponential dispersion family if its density or probability function has the form

$$f(y; \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right), \quad (1)$$

with natural parameter θ, scale parameter ϕ and functions a, b and c that characterize the form of the distribution.

a. Express the score function $u(y) = \partial \log(f(y))/\partial\theta$ in terms of y, ϕ and the mean $\mu = E(Y)$. (Hint: Use that $E(u(Y)) = 0$ in order to find a relation between μ and $b'(\theta)$.) (2p)

b. The overdispersed Poisson (ODP) distribution is defined in terms of

$$Y/\phi \sim \mathrm{Po}(\mu/\phi).$$

Find an expression for

$$f(y) = P(Y = y) = P(Y/\phi = y/\phi),$$

$y \in \{0, \phi, 2\phi, \ldots\}$. Verify that ODPs belong to the exponential dispersion family (1), by finding the natural parameter θ and the three functions $a(\phi)$, $b(\theta)$ and $c(y, \phi)$. (3p)

c. Suppose we have data $(\boldsymbol{x}_i, Y_i)$ for $i = 1, \ldots, n$, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ contains p predictors and Y_i has an ODP distribution with dispersion parameter ϕ and mean parameter μ_i. Write down the distribution of Y_i for a generalized linear model with canonical link function and regression parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$. (2p)

d. Find the elements J_{jk} of the Fisher information matrix $\boldsymbol{J} = (J_{jk})_{j,k=1}^{p}$ for $\boldsymbol{\beta}$. (Hint: Use part a),b) and $\partial \log(f(y_i))/\partial\beta_j = x_{ij}\partial \log(f(y_i))/\partial\theta_i$ for $i = 1, \ldots, n$.) (3p)

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

Appendix B - Details from Problems 2 and 3

Table 2: Numbers $n_{i+}n_{+j}/n$ from Problem 2, with $n = n_{++} = \sum_{i,j=1}^{3} n_{ij}$.

	j		
i	1	2	3
1	4.80	11.20	4.00
2	11.76	27.44	9.80
3	7.44	17.36	6.20

Table 3: Numbers $n_{eg+}n_{++s}/n$ from Problem 3, with $n = n_{+++}$.

$s = 1$:

	g	
e	1	2
1	14.48	5.21
2	33.01	34.17
3	15.64	25.48

$s = 2$:

	g	
e	1	2
1	10.52	3.79
2	23.99	24.83
3	11.36	18.52