STOCKHOLM UNIVERSITY
DEPT OF MATHEMATICS
Div. of Mathematical statistics

MT 5006
EXAMINATION
January 10 2020

Categorical Data Analysis – Examination

January 10, 2020, 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.
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 at first. Exercises need not to be ordered from simpler to harder.

Problem 1

Two grand masters of chess, Mary and Ann, contested against a computer program. Games with a draw were not reported, and each player continued until 9 games had resulted in a win, either for the chess player or the computer. The result of these 18 games is summarized in the following table:

	Chess player wins?		
Player	Yes	No	Total
Mary	6	3	9
Ann	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 that Mary and Ann wins a game respectively, that is not a draw. Which constraint on π_1 and π_2 corresponds to the null hypothesis H_0 that both players are equally skilled. Write down the likelihood under this constraint. (3p)

b. 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). (3p)

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

c. Write down the alternative hypothesis H_a that Mary is more skilled than Ann, and compute the corresponding one-sided P-value for the given data set, using Fisher's exact test. (2p)

d. Write down the alternative hypothesis H_a' that the two players are not equally skilled and compute the corresponding two-sided mid P-value for the given data set, using Fisher's exact test. (2p)

Problem 2

An amusement park has a lottery that consists of two different wheels, each one with three possible outcomes 1, 2, and 3. When these two wheels are rolled, with outcomes X and Y, a profit is returned when $X = Y$, whereas otherwise a loss occurs. The manager of the amusement part claims that X and Y are independent and uniformly distributed, i.e. $P(X = i) = 1/3$ and $P(Y = j) = 1/3$ for $i, j = 1, 2, 3$. This corresponds to a null hypothesis

$$H_0 : \pi_{ij} = P(X = i, Y = j) = \pi_{i+}\pi_{+j} = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}$$

for the joint distribution of the outcomes of the two wheels. A visitor of the amusement park, Ben, suspects that H_0 is incorrect, and that wins occur less often than predicted by H_0. In order to test H_0, Ben collected data from 90 outcomes of the lottery, with the following result:

	Y			
X	1	2	3	Sum
1	4	8	14	26
2	6	12	15	33
3	5	10	16	31
Sum	15	30	45	90

a. Assume multinomial sampling for the joint distribution of the cell counts $\{N_{ij}, 1 \leq i, j \leq 3\}$. Then formulate the likelihood for the saturated model in terms of 8 parameters. (Hint: The cell probabilities π_{ij} sum to 1.) (2p)

b. Compute a X^2-statistic in order to test H_0 against H_a that the saturated model holds but not H_0. Is H_0 rejected at level 5%? (4p)

c. After further investigation, Ben suspects that X and Y are indeed independent, the distribution of X is indeed uniform as claimed, but Y is biased towards higher values. Therefore, Ben formulates a second null hypothesis

$$H_0' : \pi_{ij} = \pi_{i+}\pi_{+j} = \frac{1}{3} \cdot \pi_{+j},$$

where all π_{+j} are left unspecified. Compute a X^2-statistic in order to test H_0' against H_a', that the saturated model holds but not H_0'. Is H_0' rejected at level 5%? (4p)

Problem 3

A research study compared the ninth classes of two schools (S) in a city. It was registered whether the average grade of each student exceeded a certain threshold or not (G), 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. Regard n_{egs} as observations of independent and Poisson distributed variables $N_{egs} \sim \text{Po}(\mu_{egs})$. Specify the parameters of two loglinear models M_1 = (EG,ES) (for which school and grade are conditionally independent given economy class) and M_0 = (EG,S) (for which school is jointly independent of grade and economic level). Note in particular which of these parameters you put to 0 in order to avoid over-parametrization (using the highest level of each variable as baseline). (3p)

b. Give formulas for the fitted values $\hat{\mu}_{egs}^{(0)}$ and $\hat{\mu}_{egs}^{(1)}$ for M_0 and M_1 respectively in terms of the appropriate marginal sums of n_{egs}. (4p)

c. The fitted cell counts in b) are given in the two tables of Appendix A. Use this to compute the likelihood ratio test statistic for choosing between M_0 and M_1 at level 0.05. Is the null hypothesis of no interaction between school and economy level rejected? (3p)

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 (still using the highest levels of each variable 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 β_0 and two effect parameters β_1 and β_2. Show in particular how β_0, 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. Let $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ be the parameter vector of the logistic regression model in a). The maximum likelihood estimates are

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = (0.4884, -1.5100, -0.4539). \tag{1}$$

Compute and interpret $\hat{\theta}_1 = \exp(\hat{\beta}_1)$, $\hat{\theta}_2 = \exp(\hat{\beta}_2)$, and $\hat{\theta}_3 = \exp(\hat{\beta}_1 - \hat{\beta}_2)$. (2p)

c. Argue that the dataset can be reduced to the marginal table with cell counts n_{eg+}. Then derive the log likelihood function $L(\boldsymbol{\beta})$ and the likelihood score vector $u(\boldsymbol{\beta}) = dL(\boldsymbol{\beta})/d\boldsymbol{\beta} = (u_0(\boldsymbol{\beta}), u_1(\boldsymbol{\beta}), u_2(\boldsymbol{\beta}))$ for the logistic regression model in a) in terms of all n_{eg+}, and with $u_j(\boldsymbol{\beta}) = \partial L(\boldsymbol{\beta})/\partial \beta_j$. The likelihood equations are $u_j(\hat{\boldsymbol{\beta}}) = 0$ for $j = 0, 1, 2$. Use (1) to verify this for at least one j. (5p)

Problem 5

Continuing Problem 2, we will look at those rolls of the two wheels for which both outcomes i and j of X and Y are either 1 or 2, and discard all other combinations of X and Y. This corresponds to the following data set $\{n_{ij}; 1 \le i, j \le 2\}$:

	Y		
X	1	2	Sum
1	4	8	12
2	6	12	18
Sum	10	20	30

Assume that the cell counts n_{ij} are observations of $\{N_{ij},\ 1 \le i, j \le 2\}$, whose joint distribution is multinomial with a total number $n = 30$ of rolls, and cell probabilities

$$p_{ij} = \frac{\pi_{ij}}{\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22}}, \quad 1 \le i, j \le 2,$$

with π_{ij} as in Problem 2.

a. Let θ be the ratio of the oddses of the second wheel having a high outcome ($Y = 2$), when the first wheel has a high ($X = 2$) and low ($X = 1$) outcome respectively. Express θ in terms of all p_{ij}. (2p)

b. Compute an estimator $\hat{\theta}$ of θ by replacing all p_{ij} with $\hat{p}_{ij} = n_{ij}/n$. Then use the multivariate delta method to prove the approximation

$$\mathrm{Var}(\log(\hat{\theta})) \approx \frac{1}{np_{11}} + \frac{1}{np_{12}} + \frac{1}{np_{21}} + \frac{1}{np_{22}}$$

for the variance of $\log(\hat{\theta})$. (Hint: Start with a) and a Taylor expansion of $\log(\hat{\theta}) - \log(\theta)$, viewed as a function of all $\hat{p}_{ij}$. You will need $\mathrm{Cov}(\hat{p}_{ij}, \hat{p}_{kl})$ for all $1 \leq i, j, k, l \leq 2$.) (4p)

c. Use b) in order to compute a Wald type confidence interval of θ with approximate coverage probability 95%. Can independence between X and Y be rejected at level 5%? (Hint: Start by computing a confidence interval of $\log(\theta)$.) (4p)

Good luck!

Appendix A - Fitted cell counts from Problem 3

Fitted cell counts of model M_0:

$\hat{\mu}^{(0)}_{eg1}$:

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

$\hat{\mu}^{(0)}_{eg2}$:

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

Fitted cell counts of model M_1:

$\hat{\mu}^{(1)}_{eg1}$:

	g	
e	1	2
1	15.44	5.56
2	33.41	34.59
3	14.83	24.17

$\hat{\mu}^{(1)}_{eg2}$:

	g	
e	1	2
1	9.56	3.44
2	23.59	24.41
3	12.17	19.83

Appendix B - Table for chi-square distribution

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