STOCKHOLM UNIVERSITY
DEPT OF MATHEMATICS
Div. of Mathematical statistics

MT 5019
EXAMINATION
January 9 2025

Categorical Data Analysis – Examination

January 9, 2025, 14:00-19:00

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: Announced through course homepage and web based course forum.

Each correct solution to an exercise yields 10 points.
Limits for grade: A, B, C, D, and E are 36, 32, 28, 24, and 20 points of 48 possible points (including bonus of 0-8 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

Two chess players 0 and 1 competed in a championship. There were 15 games (excluding those that resulted in a draw), and each game went on for a day. In the morning before a new game started, a well known psychological expert tried to predict the outcome. He did not know the results of previous games, but made each new prediction based the competitors' action and appearance. For each day without a draw, let X be the number of the player that the expert guessed would win that day, whereas Y is the number of the player that actually did win the same day. The result of the 15 predictions is summarized in the 2×2 contingency table below.

a. Let N_{ij} be the number of observations of cell i, j. Define the joint distribution of $(N_{00}, N_{01}, N_{10}, N_{11})$ under multinomial sampling. (1p)

b. Now condition on row sums and define the odds ratio θ that player 1 wins, where the ratio is taken between the two scenarios that the expert predicts as winner player

	$Y=0$	$Y=1$	Total
$X=0$	4	3	7
$X=1$	2	6	8
Total	6	9	15

1 and 0 respectively. Formulate the null hypothesis H_0 that the outcome of each game is independent of the expert's guess, and the alternative hypothesis H_a that the outcome of the game is such that the expert is more successful than random guessing. (2p)

c. Now condition on the row and column sums, so that the contingency table is solely determined by N_{11}. Write down the distribution of N_{11} under H_0. Then use Fisher's exact test for computing the mid P-value when testing H_0 against H_a. (Hint: You may use that $\binom{7}{3} = 35$, $\binom{8}{6} = 28$ and $\binom{15}{9} = 5005$.) (4p)

d. What is the distribution of N_{11} in c) for a general odds ratio θ? (Hint: Start by conditioning on the two row sums and consider the joint distribution of N_{01} and N_{11}. Then condition on the column sum $N_{01} + N_{11}$ as well.) (3p)

Problem 2

The table below shows the outcome of car accidents for drivers and passengers with or without safety belt. Data was collected in Florida during 2008, and for each person it was registered whether their accident was fatal or not. Denote by $\{n_{ij};\, 1 \le i, j \le 2\}$ the cell counts of the table, where i is the row number and j the column number. Regard this as Poisson sampling, so that n_{ij} are observations of independent Poisson variables $N_{ij} \sim \text{Po}(\mu_{ij})$.

	Injury	
Safety belt use	1: Fatal	2: Nonfatal
1: No	1085	55 623
2: Yes	703	444 239

a. Let π_1 (π_2) be the probability that a person who did not (did) use safety belt had a fatal injury. Express these probabilities in terms of the expected cell counts μ_{ij}. (2p)

b. What is the joint distribution of N_{11} and N_{21} when one conditions on the two row sums $N_{1+} = n_1 = n_{1+}$ and $N_{2+} = n_2 = n_{2+}$? (Hint: Start by defining the marginal distributions of N_{11} and N_{21}, given their respective row sums n_1 and n_2.) (2p)

c. Introduce an appropriate estimator $\hat{r} = \hat{\pi}_1/\hat{\pi}_2$ of the relative risk $r = \pi_1/\pi_2$. Use b) and a first order Taylor expansion of the function $f(\hat{\pi}_1, \hat{\pi}_2) = \log(\hat{r})$ around (π_1, π_2)

to prove that

$$\text{Var}\left[\log(\hat{r})\right] \approx \frac{1-\pi_1}{n_1\pi_1} + \frac{1-\pi_2}{n_2\pi_2}.$$

(3p)

d. Use c) to find an approximate 95% two-sided confidence interval for r. Conclude from this whether or not safety belt use has a significant effect at level 5% on the probability of a fatal injury. (3p)

Problem 3

An investigation of mortality in leukemia was conducted among survivors of the atom bomb 1945 in Hiroshima. Individuals were categorized according to their age group Z, their radiation dose X and whether they died in leukemia or not (Y) within a certain number of years, as summarized in the following threeway contingency table:

	Did not die in leukemia ($j=1$)		Died in leukemia ($j=2$)	
Age	Low dose ($i=1$)	High dose ($i=2$)	Low dose ($i=1$)	High dose ($i=2$)
0-20 years ($k=1$)	39 160	3 882	25	26
20-50 years ($k=2$)	41 664	4 291	39	26
50- years ($k=3$)	15 163	1 337	13	10

M	$G^2(M)$
(XY, XZ, YZ)	1.67
(XY, YZ)	24.44
(XY, XZ)	2.69
(XZ, YZ)	123.28
(XZ, Y)	124.27
(X, YZ)	146.02
(XY, Z)	25.42
(X, Y, Z)	147.00

a. It is assumed that all cell counts are independent Poisson distributed random variables. The second table above gives the deviance $G^2(M)$ for a number of loglinear models M. Compute the number of parameters $p(M)$ of all these models, and give the general principles for how you obtained these numbers. (2p)

b. Select the best model according to Akaike's model selection criterion AIC. (2p)

c. Let n_{ijk} and μ_{ijk} be the observed and expected count of a cell with $X = i$, $Y = j$ and $Z = k$, so that, for instance, $n_{221} = 26$ is the number of individuals of age 0-20 years with a high radiation dose who died of leukemia. Find the fitted expected cell count $\hat{\mu}_{221}$ for model $M_0 = (XY, Z)$. (Hint: For this model M_0, μ_{ijk} is a function of μ_{ij+}, μ_{++k} and μ_{+++}.) (2p)

d. Find $\hat{\mu}_{221}$ for model $M_1 = (XY, XZ)$. (2p)

e. Perform a likelihood ratio test between M_0 and M_1 at level 5%. (2p)

Problem 4

We continue studying the dataset of Problem 3. But now we are primarily interested in the effect that radiation dose has on leukemia mortality. Thus we treat death in leukemia Y as an outcome variable, radiation dose X as a predictor and age Z as a confounder. We restrict ourselves to the loglinear model (XY, XZ).

a. Define the loglinear parameters of (XY, XZ). In particular, specify which of them you put to zero in order to avoid overparametrization. (2p)

b. Show that $P(Y = 2|X = i, Z = k)$, the conditional probability of death in leukemia given radiation dose and age, defines a logistic regression model. Express its parameters as functions of the loglinear parameters from a). (2p)

c. Define the conditional odds ratio $\theta_{XY(k)}$ of death in leukemia, where the ratio is between individuals with a high and low radiation dose. Write $\theta_{XY(k)}$ in terms of the logistic regression parameters from b). Is there homogeneous association between X and Y? (2p)

d. Define the corresponding marginal odds ratio θ_{XY}, and prove that it equals the conditional odds ratio in c). (2p)

e. Compute the maximum likelihood estimator $\hat{\theta}_{XY(k)}$ of the conditional odds ratio. (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