STOCKHOLM UNIVERSITY
DEPT OF MATHEMATICS
Div. of Mathematical statistics

MT 5019
EXAMINATION
January 04 2024

Categorical Data Analysis – Examination

January 4, 2024, 8.00-13.00

Examination by: Ola Hössjer, ph. 070 671 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.

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

Problem 1

A total of 30 randomly chosen car owners were asked whether they regularly use a hands-free mobile ($X = 1$) or not ($X = 0$) while driving, and whether they had an accident situation the last five years ($Y = 1$) or not ($Y = 0$). A total of n_{ij} individuals belonged to category $X = i, Y = j$, according to the following contingency table:

	$Y = 0$	$Y = 1$	Total
$X = 0$	10	3	13
$X = 1$	8	9	17
Total	18	12	30

a. Regard the data n_{ij} of this table as the outcome of a multinomial distribution with $n_{++} = 30$ observations and cell probabilities π_{ij} for $0 \le i, j \le 1$. Formulate the null hypothesis H_0 that mobile usage and accident proneness are independent. (2p)

b. Fisher's exact test of H_0 uses only N_{11}, and it 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	5	6
$P_{H_0}(N_{11} = n_{11}\|\ldots)$	0.0000	0.0000	0.0004	0.0056	0.0354	0.1228	0.2455
n_{11}	7	8	9	10	11	12	
$P_{H_0}(N_{11} = n_{11}\|\ldots)$	0.2894	0.2010	0.0804	0.0175	0.0019	0.0001	

c. Formulate the alternative hypothesis H_a that mobile usage increases accident risk in terms of an odds ratio. (1p)

d. Compute the P-value and mid P-value of a one-sided test where H_0 is tested against H_a. (2p)

e. Fisher's exact test with nominal significance level α rejects H_0 if $P \leq \alpha$ or if mid $P \leq \alpha$, depending on whether the P-value or mid P-value is used. Determine whether these two tests are conservative (actual significance level $\leq \alpha$) or anti conservative (actual significance level $> \alpha$) when $\alpha = 0.05$ and $\alpha = 0.07$ respectively. (2p)

Problem 2

For a US social survey of 2006, individuals of different ages were asked about their job satisfaction Y. Age was categorized into three levels; depending on whether the interviewed person was young (< 30), middle-aged ($30-50$) or old (> 50). The investigator wanted to find out whether increased age was associated with increased job satisfaction, and reported data in the following twoway 3×3 table:

	Job Satisfaction Y		
Age X	1 (=not satisf)	2 (=fairly satisf)	3 (=very satisf)
1 (< 30	34	53	88
2 (30-50)	80	174	304
3 (> 50)	29	75	172

a. Assume that the cell counts N_{ij} for all cells (i, j) are independent and Poisson distributed random variables with expected values μ_{ij}. Define the local odds ratio θ_{ij} for the 2×2 subtable with upper left corner (i, j) for all such subtables. (Hint: For each subtable, θ_{ij} is the ratio of the oddses of lower job satisfaction, between the younger and older age group of that subtable. It can be expressed as a function of the four expected cell counts μ_{ij} of the subtable.) (2p)

b. Compute estimates $\hat{\theta}_{ij}$ of all θ_{ij}. (2p)

From the result in 2b it seems that the higher job-satisfaction of middle-aged compared to young is more due to a higher fraction of young having a job they are not satisfied with, than a higher fraction of middle-aged having a job they are very satisfied with. We may formalize this as the alternative hypothesis H_a of a test where the null hypothesis $H_0 : \theta_{11} = \theta_{12}$ is compared against $H_a : \theta_{11} > \theta_{12}$.

c. Use the multivariate delta method to find an approximation of $\mathrm{Var}\left[\log(\hat{\theta}_{11}/\hat{\theta}_{12})\right]$, and then compute an estimate $\widehat{\mathrm{Var}}\left[\log(\hat{\theta}_{11}/\hat{\theta}_{12})\right]$. (Hint: When $\mathrm{Var}\left[\log(\hat{\theta}_{11}/\hat{\theta}_{12})\right]$ is calculated, the two estimated local odds ratios $\hat{\theta}_{11}$ and $\hat{\theta}_{12}$ involve four cell counts N_{ij} each. But some of these cell counts occur in both of $\hat{\theta}_{11}$ and $\hat{\theta}_{12}$, so you will take the variance of a sum of six terms. In this variance calculation, the multivariate delta method is used, based on an approximation $\log(N_{ij}) \approx \log(\mu_{ij}) + (N_{ij} - \mu_{ij})/\mu_{ij}$ of the logarithm of each cell count that appears among the six terms of the variance expression.) (3p)

d. Use the result in 2c to compute an approximate one-sided 95% confidence interval for θ_{11}/θ_{12}, by first calculating a confidence interval for $\log(\theta_{11}/\theta_{12})$ and then transforming back this interval to the original scale. Conclude whether H_0 should be rejected or not. (Hint: The upper end point of the confidence interval for $\log(\theta_{11}/\theta_{12})$ is ∞, and its lower end point includes the 95% quantile 1.645 of the standard normal distribution.) (3p)

Problem 3

The 3×3 contingency table of Problem 2 can be analyzed in terms of a multinomial distribution with n observations and cell probabilities π_{ij}.

a. Express π_{ij} in terms of the mean parameters μ_{ij}. (2p)

In order to quantify the direction of dependency between age and job satisfaction, we may use

$$\gamma = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d},$$

a number between -1 and 1 that compares the probabilities Π_c and Π_d that a randomly chosen pair (X, Y) and (X', Y') of observations are concordant and discordant respectively. An estimate of this quantity is

$$\hat{\gamma} = \frac{C - D}{C + D} = \frac{\hat{\Pi}_c - \hat{\Pi}_d}{\hat{\Pi}_c + \hat{\Pi}_d}, \qquad (1)$$

where C and D refer to the number of concordant and discordant pairs of interviewed persons from the data set. (Since the total number of cell pairs is $n(n-1)/2$, we have that $C/(n(n-1)/2) = \hat{\Pi}_c$ and $D/(n(n-1)/2) = \hat{\Pi}_d$ are estimators of Π_c and Π_d.)

b. Compute $\hat{\gamma}$, by first computing C and D. What is your conclusion? (3p)

c. Define Π_c and Π_d in terms of the cell probabilities π_{ij}. (Hint: Make use of $P[(X, Y) = (i, j), (X', Y') = (h, k)] = \pi_{ij}\pi_{hk}$ and check which pairs $(i, j), (h, k)$ of cells are concordant/discordant.) (2p)

d. Use the hint in 3c to prove that $\Pi_c = \Pi_d$, and hence $\gamma = 0$, under the null hypothesis that age and job satisfaction are independent. (3p)

Problem 4

An epidemiologist studied possible association between exposure to a certain pollutant (X), lung cancer (Y) and smoking (Z) for a group of workers at a large factory. She modeled all three variables as binary, with levels 0 and 1 corresponding to absence and presence of exposure, cancer or smoking. The number of individuals n_{ijk} with $X = i, Y = j, Z = k$ is summarized in threeway $2 \times 2 \times 2$ contingency table, that consists of the following two partial tables for persons with our without cancer:

Observed values n_{i0k}:

	$k=0$	$k=1$
$i=0$	93	39
$i=1$	101	50

Observed values n_{i1k}:

	$k=0$	$k=1$
$i=0$	12	22
$i=1$	31	72

It is assumed that n_{ijk} are observations of independent and Poisson distributed random variables N_{ijk} with means μ_{ijk}, for all cells (i, j, k).

a. The epidemiologist hypothesized that lung cancer is associated with each one of the two risk factors separately, but not jointly. Therefore, she wanted to test the loglinear model $M_0 = (XY, YZ)$. Specify μ_{ijk} and the parameter vector $\boldsymbol{\beta}$ for this model, when $X = 0$, $Y = 0$ and $Z = 0$ are chosen as baseline levels. (2p)

b. Use the result in 4a to prove that $\mu_{ijk} = \mu_{ij+}\mu_{+jk}/\mu_{+j+}$ for model M_0, where pluses indicate summation over indeces. (Hint: It is possible to use 4a and write the expected cell counts as products $\mu_{ijk} = B_{ij}C_{jk}$ for some B_{ij} and C_{jk}.) (2p)

c. Use 4b and the cell counts of the two partial tables, their row sums n_{ij+}, column sums n_{+jk} and total sums $n_{+0+} = 283$ and $n_{+1+} = 137$, to find the fitted expected cell counts $\hat{\mu}_{ijk} = \hat{\mu}_{ijk}(M_0)$ of model M_0 for all cells. (2p)

d. Perform a likelihood ratio test

$$G^2(M_0) = 2\left[L(M_1) - L(M_0)\right] = 2 \sum_{i,j,k=0}^{1} n_{ijk} \log \frac{n_{ijk}}{\hat{\mu}_{ijk}} \tag{2}$$

between M_0 and the saturated model M_1, and conclude whether M_0 is rejected or not. (2p)

e. Prove that the LR test statistic $G^2(M)$ is given by (2) for any loglinear model M tested against the saturated model M_1, provided the baseline parameter λ is a parameter of M (and with $\hat{\mu}_{ijk} = \hat{\mu}_{ijk}(M)$ in (2)). (2p)

Problem 5

The epidemiologist of Problem 4 is primarily interested in the effect of exposure on lung cancer, since smoking is a previously known risk factor. She defined an ANOVA type multiple logistic regression model from the loglinear model M_0, with lung cancer Y as response variable, exposure X as the predictor of main interest, and smoking Z as a confounder.

a. Prove that

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

and in particular write α, β_i^X and β_k^Z as functions of the loglinear parameters. Specify which parameters you put to zero. (3p)

b. Give an expression for the conditional odds ratio $\theta_{XY(k)}$ between exposure and lung cancer of model M_0, when conditioning on smoking. Is the association between exposure and lung cancer homogeneous? (2p).

c. Prove that the marginal odds ratio θ_{XY} between exposure and lung cancer equals the conditional odds ratio in 5b. (Hint: Make use of the hint of Problem 4b and write the expected cell counts of the marginal table as $\mu_{ij+} = B_{ij}C_{j+}$. The result of 5c indicates that Z could be removed from the model.) (3p)

d. Compute the maximum likelihood estimator $\hat{\beta}_1^X$ of β_1^X. (Hint: Use parts 5b and 5c.) (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