STOCKHOLM UNIVERSITY
DEPT OF MATHEMATICS
Div. of Mathematical statistics

MT 5006
EXAMINATION
February 3, 2021

Categorical Data Analysis – Examination

Febaruary 3, 2021, 9.00-16.00

Examination by: Ola Hössjer, ph. 070 672 12 18, `ola@math.su.se`
Allowed to use: Miniräknare/pocket calculator (including the use of R as pocket calculator), table in the appendix of this exam, course literature and other course material. You are not allowed to ask anyone for help.
Inlämning/Handing in: The solutions should by sent by email as a pdf file to the eaminator by 16.00. Either scanned hand-written notes, or a pdf file generated from a word processor. *Återlämning/Return of exam:* Will be communicated by email.

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 0

Verify that you have solved all exerices without help from anyone. This is required in order for the solutions to be corrected. (0p)

Problem 1

An old British lady claimed an ability to distinguish whether milk or tea was put into a cup at first. In order to find out whether her claim was true or not, a statistician designed a clinical trial with fifteen cups of tea. Seven of these cups were first prepared with milk ($X = 0$), whereas the remaining eight cups were first filled with tea ($X = 1$). The cups were given to the lady in a random order, and for each cup she guessed whether it was filled with milk ($Y = 0$) or tea ($Y = 1$) at first. The result of the study is summarized in the following 2×2 contingency table:

	$Y=0$	$Y=1$	Total
$X=0$	5	2	7
$X=1$	2	6	8
Total	7	8	15

a. Let N_{ij} be the number of observations of cell i,j. Define the likelihood, i.e. the joint distribution of N_{00} and N_{10} under independent binomial rows sampling, in terms of the success probabilities $\pi_i = P(Y=0|X=i)$, $i=0,1$, of the two binomial distributions. (2p)

b. Formulate the null hypothesis H_0 that the lady guesses at random, against the one-sided alternative that she performs better than random guessing. Express these hypotheses in terms of the probabilities in 1a) as well as in terms of an odds ratio θ. (2p)

c. Suppose instead the design of the study was as follows: The lady was told that seven cups of tea were first prepared with milk and the remaining eight cups had tea poured into them at first. Then she was asked to pick out the seven cups that had milk at first. This corresponds to conditioning not only on the row sums but also on the column sums of the table, so that all its entries are determined by N_{00}. Write down the distribution of N_{00} under H_0. Then use Fisher's exact test for computing the P-value when testing H_0 against H_a. (Hint: You may use that $\binom{7}{5}=21$, $\binom{8}{2}=28$ and $\binom{15}{7}=6435$.) (3p)

d. What is the distribution of N_{00} in 1c) for a general odds ratio θ? (Hint: Start with the joint distribution of N_{00} and N_{10} found in 1a) for indepdendent binomial rows sampling, and then condition on column sums. This corresponds to a likelihood $P(N_{00}=k|N_{+0}=7) \propto P(N_{00}=k)P(N_{10}=7-k)$ for $k=0,\ldots,7$, where the proportionality constant does not depend on k.) (3p)

Problem 2

It is generally believed that SARS-Cov-2 infected individuals who wear face mask lower the risk of spreading the virus to others (so called source control). It is more debated however to what extent uninfected individuals are protected by wearing face mask. In order to answer this question, in a Danish study, 6024 unaffected inviduals were asked to either wear face mask ($X=0$) or not ($X=1$). For the $n=4862$ individuals that completed the study it was checked, about a month later, whether they had been infected ($Y=1$) or not ($Y=0$) during this period. The following table summarizes the number n_{ij} of individuals with $X=i$ and $Y=j$, for $0 \le i,j \le 1$:

	Infected after a month?			
Face mask?	No ($Y=0$)	Yes ($Y=1$)	Total	n_{i1}/n_{i+}
Yes ($X=0$)	2350	42	2392	0.01756
No ($X=1$)	2417	53	2470	0.02146
Sum	4767	95	4862	

a. The study can be regarded as independent binomial rows sampling, with affection probabilities $\pi = P(Y = 1|X = 1)$ among inviduals that do not wear face mask, whereas $\pi + \Delta = P(Y = 1|X = 0)$ is the affection probability among individuals that wear face mask. Define the log likelihood function $L(\pi, \Delta)$ of the data set $\{n_{ij};\, 0 \le i, j \le 1\}$. (2p)

b. Regarding π as a nuisance parameter, we want to test

$$\begin{aligned} H_0 : &\quad \Delta = 0, \\ H_a : &\quad \Delta < 0. \end{aligned}$$

To this end, we use the score statistic

$$z_S = \frac{\hat{\Delta}}{\sqrt{(\frac{1}{n_{0+}} + \frac{1}{n_{1+}})\hat{\pi}(0)(1 - \hat{\pi}(0))}}, \tag{1}$$

where $\hat{\Delta} = n_{01}/n_{0+} - n_{11}/n_{1+}$ is an estimate of Δ, whereas $\hat{\pi}(\Delta)$ is the value of π that maximizes the log likelihood when Δ is fixed. In particular, $\hat{\pi}(0) = n_{+1}/n$ is the maximum likelihood estimate of π under the null hypothesis. Perform the score test, at significance level 5%, in order to find out whether face mask wearing significantly lowers the risk of being infected. (Hint: You may use the value $z_{0.05} = 1.645$ of a standard normal quantile.) (2p)

c. Use 2a) in order to derive expressions for the score function components $u_\Delta(\pi, \Delta) = \partial L(\pi, \Delta)/\partial \Delta$ and $u_\pi(\pi, \Delta) = \partial L(\pi, \Delta)/\partial \pi$. (2p)

d. Use 2c) in order to find expressions for the two diagonal elements $J_{\pi\pi}(\pi, \Delta)$ and $J_{\Delta\Delta}(\pi, \Delta)$ of the Fisher information matrix as well as the value $J_{\pi\Delta}(\pi, \Delta)$ of the two non-diagonal elements of this matrix.(Hint: One possibility is to first derive the elements of the Hessian matrix $\boldsymbol{H}(\pi, \Delta)$. For instance, its first diagonal element is $H_{\pi\pi}(\pi, \Delta) = \partial u(\pi, \Delta)/\partial \pi$.) (2p)

e. Use 2c-2d) to verify that (1) is indeed a score statistic for the Δ parameter. That is, show that

$$z_S = \frac{u_\Delta(\hat{\pi}(0), 0)}{\sqrt{\text{Var}(u_\Delta(\hat{\pi}(0), 0))}},$$

where $u_\Delta(\hat{\pi}(0), 0)$ is the score function at $\Delta = 0$ when the nuisance parameter π is estimated by the quantity $\hat{\pi}(0)$, defined in 2b). (Hint: You may without proof use that $\text{Var}(u_\Delta(\hat{\pi}(0), 0)) = J_{\Delta\Delta}(\hat{\pi}(0), 0) - J_{\pi\Delta}(\hat{\pi}(0), 0)^2/J_{\pi\pi}(\hat{\pi}(0), 0)$.) (2p)

Problem 3

A genetic disorder has one of its risk factors at mitochondrial DNA. This mitichodrial gene has one normal ($Z = 0$) and one risk increasing variant ($Z = 1$). Since mitochondrial DNA is inherited from the mother, in order to test if there are other genetic risk factors, the disease status X and Y was investigated for 1000 pairs of mothers and children, with

$X = 1$ (0) for an affected (unaffected) mother, and similarly $Y = 1$ (0) for an affected (unaffected) child. The mother and the child always share the same variant $Z = 0$ or 1. The result of the study was:

Genetic variant $Z = 0$:

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

Genetic variant $Z = 1$:

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

a. Assume that the number of mother-child 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 $M = (XZ, YZ)$. 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)

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

c. Appendix B contains a table with ML estimates $\hat{\mu}_{ijk}$ of the expected counts for all cells (i, j, k). Use 3b) to veryify that the numeric value of $\hat{\mu}_{111}$ is correct. (Hint: The row sums, column sums and total number of observations of each partial table for $Z = 0$ and $Z = 1$ will be helpful.) (1p)

d. Perform an X^2-test in order to check (at level 5%) whether (XZ, YZ) adequately describes data. (Hint: The table of Appendix B will be helpful.) (3p)

e. Use the two partial tables above to estimate the two conditional odds ratios $\theta_{XY(k)}$ for each variant $k = 0, 1$ of the gene. Based on this (and as a complement to the goodness-of-fit test of M in 3d)) discuss whether M seems to be a good model. (2p)

Problem 4

In order to incorporate the possibility of other shared risk factors in Problem 3, we will look at the loglinear model $M_1 = (XY, XZ, YZ)$ which has all second order interactions included, and the lowest level of each category chosen as baseline.

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

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

$$\text{logit}\left[P(Y = 1|X = i, Z = k)\right] = \alpha + \beta_i^X + \beta_k^Z, \tag{2}$$

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 and the delta method to compute a 95% confidence interval for the conditional odds ratio $\theta_{XY(k)}$ between the mother's and the child's affection status, given $Z = k$. What conclusion can be drawn from this regarding additional common risk factors for the mother and child? (Hint: Because of homogeneous association of model M_1, $\theta_{XY(k)} = \theta_{XY}$ does not depend on k. Start looking at $\log(\theta_{XY})$.) (3p)

d. Use (2) and the delta method to compute a 95% confidence interval for the affection probability $P(Y = 1|X = 0, Z = 1)$ of a child who carries the risk variant 1 and whose mother is not affected. (3p)

Problem 5

Suppose we have a data set $(\boldsymbol{x}_i, y_i)$ for $i = 1, \ldots, n$, with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ the values of p predictors and y_i the outcome variable of observation i. It is assumed that the distribution of $Y_i|\boldsymbol{x}_i$ belongs to an exponential dispersion family (EDF), with a density or probability function of the form

$$f_{Y_i|\boldsymbol{x}_i}(y) = f(y; \theta_i, \omega_i, \phi) = \exp\left(\frac{y\theta_i - b(\theta_i)}{\phi/\omega_i} + c(y, \phi)\right), \tag{3}$$

where $\theta_i = \theta(\boldsymbol{x}_i)$ is the natural parameter, ω_i the exposure weight of observation i, whereas ϕ is a scale parameter. The functions b and c characterize the form of the distribution.

a. We will first regard $\theta_1, \ldots, \theta_n$ as parameters. Use (3) to find expressions for the score function

$$u_i(y) = \frac{\partial \log(f(y))}{\partial \theta_i}$$

and Hessian

$$H_i(y) = \frac{\partial^2 \log(f(y))}{\partial \theta_i^2}$$

of observation i. (2p)

b. The binomial proportions model (BPM) is defined in terms of

$$n_iY_i \sim \mathrm{Bin}(n_i, \pi_i), \tag{4}$$

where $\pi_i = \pi(\boldsymbol{x}_i)$. Verify that (4) belongs to the exponential dispersion family (3), by finding θ_i, ω_i, and ϕ, as well as the two functions $b(\theta_i)$ and $c(y, \phi)$. Then use 5a) to find the score function $u_i(y)$ and Hessian $H_i(y)$ of observation i. (Hint: Since n_iY_i has a bionmial distribuiton according to (4), start using the fact that $P(Y_i = y) = P(n_iY_i = n_iy)$.) (4p)

c. Consider a Generalized Linear Model, where each outome variable $Y_i|\boldsymbol{x}_i$ has a BPM distribution (4). The expected success proportion $\pi_i = \pi(\boldsymbol{x}_i; \boldsymbol{\beta})$ of each observation corresponds to using a canonical link function and a regression parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$. In contrast to 5a) we will now regard $\boldsymbol{\beta}$ as the parameter vector. Let

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log[f(y_i; \theta_i, \omega_i, \phi)]$$

be the log likelihood, viewed as a function of $\boldsymbol{\beta}$. Use 5a)-5b) in order to find the compunents of score function $\boldsymbol{u}(\boldsymbol{\beta}) = (u_1(\boldsymbol{\beta}), \ldots, u_p(\boldsymbol{\beta}))^T = dL(\boldsymbol{\beta})/d\boldsymbol{\beta}$ and the elements of the Fisher information matrix $\boldsymbol{J}(\boldsymbol{\beta}) = (J_{jk}(\boldsymbol{\beta}))_{j,k=1}^{p}$. (Hint: View each θ_i as a function of $\boldsymbol{\beta}$ and use the chain rule.) (4p)

Good luck!

Appendix A - 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

Appendix B - Fitted cell counts $\hat{\mu}_{ijk}$ from Problem 3

Genetic variant $Z = 0$:

Mother's aff status	Child's aff status		Sum
	$Y = 0$	$Y = 1$	
$X = 0$	838.2	29.8	868
$X = 1$	32.8	1.17	34
Sum	871	31	902

Genetic variant $Z = 1$:

Mother's aff status	Child's aff status		Sum
	$Y = 0$	$Y = 1$	
$X = 0$	23.5	25.5	49
$X = 1$	23.5	25.5	49
Sum	47	51	98