STOCKHOLM UNIVERSITY
DEPT OF MATHEMATICS
Div. of Mathematical statistics

MT 5022
EXAMINATION
January 7 2026

Classification and Analysis of Categorical Data – Examination

January 7, 2026, 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: 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 36, 32, 28, 24, and 20 points of 44 possible points (including bonus of 0-4 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 professional art expert (1) and a self-trained amateur (0) competed in a competition. Both of them were shown the same 9 pairs of paintings. Among each such pair, one painting was made by a well known artist, whereas the other one was a forgery. The expert and amateur were asked to decide for each pair of paintings, which of them was an original and which was not. Let N_{i1} refer to the number of times person i made a correct decision, whereas $N_{i0} = 9 - N_{i1}$ is the number of wrong guesses. It is assumed that the two persons answered independently of each other, and moreover, that all guesses of person i were independent with the same probability $\pi_{j|i}$ of having outcome j, where $j = 1$ corresponds to a correct guess and $j = 0$ to a wrong one. The outcome of the competition is shown in the following table, where X refers to individual and Y to whether a guess is correct or not.

	$Y = 0$	$Y = 1$	Total
$X = 0$	4	5	9
$X = 1$	2	7	9
Total	6	12	18

a. Determine the sampling scheme, and write down the likelihood function in terms of the two parameters $\pi_{1|0}$ and $\pi_{1|1}$. (2p)

b. Formulate the null hypothesis H_0 that the two persons are equally skilled in distinguishing original pieces of art from forgeries, both in terms of probabilities $\pi_{j|i}$ and in terms of the risk ratio between the expert's and amateur's success probabilities. (2p)

c. Now condition on the column sums as well, so that the contingency table is solely determined by N_{11}. Write down (without proof) the distribution of N_{11} under H_0. (3p)

d. Use Fisher's exact test for computing the P-value and mid P-value when testing H_0 against the alternative H_a that the expert does better than the amateur. Can we say that the expert is better than the amateur in terms of distinguishing original art? (Hint: You may use that $\binom{9}{3} = 84$, $\binom{9}{4} = 126$ and $\binom{18}{12} = 18564$.) (3p)

Problem 2

A large company is located in two different regions. The leaders of the company decided to investigate whether employees in the two regions had the same degree of job satisfaction or not. They collected data for a sample of $n = n_{++}$ employees in terms of a 2×2 table, with n_{ij} the number of employees in region $i = 1, 2$ who where either satisfied ($j = 1$) or not ($j = 2$) with their job, with the following result:

	$j = 1$	$j = 2$	Total
$i = 1$	133	52	185
$i = 2$	120	36	156
Total	253	88	341

a. Each cell count n_{ij} is an observation of N_{ij}, where N_{ij} is a random variable with expected value $n\pi_{ij}$, and π_{ij} is the probability that a sampled individual belongs to cell (i, j). Write down the likelihood function l of data, if multinomial sampling is assumed. (2p)

b. Define the odds ratio OR between job satisfaction in region 1 and 2 in terms of the model parameters. Then compute an approximate 95% confidence interval for OR. You may use without proof the fact that

$$\text{Var}\left[\log(\widehat{\text{OR}})\right] \approx \frac{1}{n\pi_{11}} + \frac{1}{n\pi_{12}} + \frac{1}{n\pi_{21}} + \frac{1}{n\pi_{22}},$$

where $\widehat{\text{OR}}$ is a certain estimate of OR. Is there any significant difference between the two regions in terms of job satisfaction? (4p)

c. Alternatively, one may use

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

in order to quantify association between region X and job satisfaction Y, where Π_c (Π_d) is the probability that a pair of individuals is concordant (discordant). Compute a 95% confidence interval for γ. (Hint: Start by writing Π_c and Π_d in terms of the cell probabilities $\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}$. Then express γ as a function of OR in b).) (4p)

Problem 3

In a cross-sectional study; gender, seat belt use (S), location (L), and injury (I) were reported for a number of passengers involved in automobile or light truck accidents during one year in the state of Maine, US. The table below contains data for all males, with $S = 0$ corresponding to no seat belt use, $S = 1$ to seat belt use, $L = 0$ to urban, and $L = 1$ to a rural area:

No injury $I = 0$:

Seat belt use	Location $L=0$	$L=1$	Sum
$S=0$	10381	6123	16504
$S=1$	10969	6693	17662
Sum	21350	12816	34166

Injury $I = 1$:

Seat belt use	Location $L=0$	$L=1$	Sum
$S=0$	812	1084	1896
$S=1$	380	513	893
Sum	1192	1597	2789

a. Assume that the number of individuals N_{sli} with $S = s$, $L = l$, and $I = i$ are independent Poisson random variables, with expected values μ_{sli}. In order to test whether location and seat belt use are conditionally independent given injury, we will consider the loglinear model $M = (SI, LI)$. Express μ_{sli} in terms of the loglinear parameters. After setting some loglinear parameters to zero in order to avoid overparametrization, which ones and how many remain? (2p)

b. Prove that $\mu_{sli} = \mu_{s+i}\mu_{+li}/\mu_{++i}$ for model M. (Hint: You may either use the representation in a), or look at $\pi_{sli} = \mu_{sli}/\mu_{+++}$.) (2p)

c. Use b) in order to find ML estimates $\hat{\mu}_{sli}$ of the expected counts for all cells (s, l, i). (Hint: The row sums, column sums, and total number of observations of each partial table for $I = 0$ and $I = 1$ will be helpful.) (3p)

d. Perform a likelihood ratio test between (SI, LI) and the saturated model (SLI) in order to check (at level 5%) whether M adequately describes data. (3p)

Problem 4

Consider the loglinear model $M = (SI, LI)$ of Problem 3. We will now regard injury I as an outcome variable, whereas seat belt use S and location L are predictors.

a. Show that the conditional distribution of I given L and S defines the ANOVA type logistic regression model

$$\text{logit}\,[P(I = 1|S = s, L = l)] = \alpha + \beta_s^S + \beta_l^L, \tag{1}$$

and write α, β_s^S, and β_l^L as functions of the loglinear parameters. Then show that α, β_1^S, and β_1^L are the only nonzero parameters of the logistic model, if $I = 0$, $S = 0$, and $L = 0$ are chosen as baseline levels for the loglinear model. (3p)

b. Let $\theta_{SI(l)}$ be the conditional odds ratio of having an injury between people who use seat belt and not, given that their location variable is l. Express $\theta_{SI(l)}$ in terms of the logistic parameters in (1). Is their homogeneous association between seat belt use and injury? (Hint: It might be convenient to first look at the log conditional odds ratio $\log(\theta_{SI(l)})$.) (3p)

c. Prove that the conditional odds ratio in b) equals the marginal odds ratio θ_{SI} of having an injury between people who use seat belt and not. (Hint: Use Bayes' Theorem in order to work with $P(S = s|I = i, L = l)$ instead of $P(I = i|S = s, L = l)$.) (2p)

d. Use the data set of Problem 3 in order to estimate θ_{SI}. (2p)

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