STOCKHOLM UNIVERSITY
DEPT OF MATHEMATICS
Div. of Mathematical statistics

MT 5006
EXAMINATION
February 20 2019

# Categorical Data Analysis – Examination

February 20, 2019, 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 through the whole exam at first. Exercises need not to be ordered from simpler to harder.

---

# Problem 1

In a small case-control study, 10 patients with lung cancer ($Y = 1$) and 10 healthy controls ($Y = 0$) were asked about whether they smoked regularly ($X = 1$) or not ($X = 0$). It was assumed that the number of smokers among controls and cases had independent binomial distributions $\text{Bin}(10, \pi_0)$ and $\text{Bin}(10, \pi_1)$, with probabilities $\pi_0 = P(X = 1|Y = 0)$ and $\pi_1 = P(X = 1|Y = 1)$ respectively. The result of the study is summarized in the following table:

| | $Y = 0$ | $Y = 1$ | Total |
|---|---|---|---|
| $X = 0$ | 9 | 5 | 14 |
| $X = 1$ | 1 | 5 | 6 |
| Total | 10 | 10 | 20 |

a. The investigator wanted to test the null hypothesis $H_0$ that smoking does not elevate the risk of lung cancer against the one-sided alternative $H_a$ that it does. Express $H_0$ and $H_a$ in terms of an odds ratio $\theta$ that involves the two probabilities $\pi_0$ and $\pi_1$. (2p)

b. The two columns sums $n_{+0} = n_{+1} = 10$ are fixed by the design of the study. Now suppose we condition on row sums as well. Give an expression for the $H_0$-distribution $P(N_{11} = n_{11}|H_0, N_{1+} = 6)$ of the number of patients that smoke, conditional on the total number of smokers. (3p)

c. Compute the $p$-value of Fisher's exact test of $H_0$ against $H_a$. (Hint: Use that $\binom{20}{6} = 38760$.) (2p)

d. Use a) in order to derive the distribution $P(N_{11} = n_{11}|\theta, N_{1+} = 6)$ of the number of patients that smoke, conditional on the total number of smokers, for any value of the odds ratio $\theta$. (3p)

# Problem 2

Suppose the dataset of Problem 1 is a subset of a larger case-control study with 50 controls and 50 cases, so that the number of smokers within each group still have independent binomial distributions with probabilities $\pi_0$ and $\pi_1$, but with 50 instead of 10 trials. The result of the study is obtained by multiplying all four cell counts in the table of Problem 1 by 5. Consequently, in the full dataset there are 25 smokers and 25 non-smokers among the cases, whereas 5 people smoke and 45 do not smoke among the controls.

a. Let $\Delta = \pi_1 - \pi_0$ be the difference in proportion of smokers among cases and controls. Compute a two-sided Wald-based test of $H_0 : \Delta = 0$ against $H_a : \Delta \neq 0$. Is there a significant difference in proportion of smokers between cases and controls at level 5%? Compare the result with that of Problem 1c) and explain the difference. (3p)

b. Let $r = \pi_1/\pi_0$ be the relative risk of smoking between cases and controls. Compute the maximum likelihood estimator $\hat{r}$ of $r$. Then use the multivariate delta method to prove that

$$\text{Var}\left[\log(\hat{r})\right] \approx \frac{1-\pi_0}{n_0\pi_0} + \frac{1-\pi_1}{n_1\pi_1},$$

where $n_j = N_{+j}$. (4p)

c. Use c) in order to find an approximate 95% two-sided confidence interval for $r$. Conclude from this whether smoking increases the risk of lung cancer. (3p)

# Problem 3

A number of patients underwent heart valve replacement surgery at two different clinics ($Z = 1$ and $Z = 2$), where either an aortic ($X = 1$) or mitral ($X = 2$) valve was replaced. A biostatistician investigated whether any complications had occurred ($Y = 2$) or not

($Y = 1$), one year after the surgery. She analyzed all surgeries at the two clinics during one year, as summarized in the following two partial tables:

No complications $Y = 1$:

| Type of valve | Clinic $Z = 1$ | Clinic $Z = 2$ | Sum |
|---|---|---|---|
| $X = 1$ | 210 | 200 | 410 |
| $X = 2$ | 210 | 190 | 400 |
| Sum | 420 | 390 | 810 |

Complications $Y = 2$:

| Type of valve | Clinic $Z = 1$ | Clinic $Z = 2$ | Sum |
|---|---|---|---|
| $X = 1$ | 30 | 100 | 130 |
| $X = 2$ | 10 | 50 | 60 |
| Sum | 40 | 150 | 190 |

When analyzing data, the biostatistician assumed that the number of patients $N_{ijk} \sim \text{Po}(\mu_{ijk})$ with $X = i$, $Y = j$ and $Z = k$ were independent and Poisson distributed for different combinations $i, j, k \in \{1, 2\}$ of type of valve, absence/presence of complications, and clinic. Since it was known that clinic 1 had a higher fraction of successful surgeries, and that complications occurred more often for aortic than for mitral valve replacements, she wanted to find the simplest possible model with these features. Therefore she hypothesized a loglinear model $H_0 : M = (XY, YZ)$ for the expected cell counts $\mu_{ijk}$.

a. Express $\mu_{ijk}$ in terms of the loglinear parameters for model $M = (XY, YZ)$. Put some of these parameters to zero in order to avoid overparametrization. How many parameters remain? (2p)

b. Prove that $\mu_{ijk} = \mu_{ij+}\mu_{+jk}/\mu_{+j+}$ for model $M = (XY, YZ)$. (Hint: You may either use a) or look at $\pi_{ijk} = \mu_{ijk}/\mu_{+++}$.) (2p)

c. Use b) in order to find ML estimates $\hat{\mu}_{ijk}$ of the expected cell counts for model $M$. The row sums, columns sums, and total number of observations of each partial table will be helpful. (3p)

d. Perform a likelihood ratio test between $M$ and the saturated model $(XYZ)$ in order to check (at significance level 5%) whether $M$ fits data well. (3p)

# Problem 4

Consider the loglinear model $M = (XY, YZ)$ of Problem 3, regarding $Y$ as an outcome and $X, Z$ as predictor variables.

a. Show that the conditional distribution of $Y$ given $X$ and $Z$ defines an ANOVA type logistic regression model

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

and write $\alpha$, $\beta_i^X$, and $\beta_k^Z$ as functions of the loglinear parameters of Problem 3. Then show that $\alpha$, $\beta_2^X$ and $\beta_2^Z$ are the only nonzero parameters of the model if $X = 1$, $Y = 1$, and $Z = 1$ are chosen as baseline levels for the loglinear model. (3p)

b. Define $\theta_{XY(k)}$, the conditional odds ratio of having a complication one year after surgery, between patients who had their mitral and aortic valves replaced, conditional on clinic $Z = k$. Then express $\theta_{XY(k)}$ in terms of the logistic parameters in (1). Is there homogeneous association between $X$ and $Y$? (4p)

c. Express the average causal effect (ACE) that type of valve has on the probability of complication, in terms of the three nonzero logistic parameters in a). (Hint: The expression will involve $N_{++1}$ and $N_{++2}$, where $N_{ijk}$ are the cell counts of Problem 3.) (3p)

# Problem 5

Suppose we have data from two binary variables $X, Y$, summarized in a $2 \times 2$ table, whose entries $N_{ij}$ are the number of observations with $X = i$ and $Y = j$ for $i, j \in \{0, 1\}$. It is assumed that the two rows of the table have independent binomial distributions, i.e. $N_{i1} \sim \text{Bin}(n_i, \pi_i)$ are independent for $i = 0, 1$, with $n_i = N_{i0} + N_{i1}$, and

$$\begin{aligned} \pi_0 &= \exp(\alpha)/[1 + \exp(\alpha)], \\ \pi_1 &= \exp(\alpha + \beta)/[1 + \exp(\alpha + \beta)], \end{aligned}$$

for some parameters $\alpha$ and $\beta$.

a. Determine the log likelihood $L(\alpha, \beta)$. (Any term of $L(\alpha, \beta)$ that neither depends on $\alpha$ nor $\beta$ can be denoted as $C$, without further specification.) (2p)

b. Use a) to compute the second derivate matrix $d^2L(\alpha, \beta)/d(\alpha, \beta)^2$. (3p)

c. Let $\hat{\beta}$ be the maximum likelihood estimate of $\beta$. Based on b), find an approximation of $\text{Var}(\hat{\beta})$. (Hint:. The formula

$$\begin{pmatrix} a & c \\ c & b \end{pmatrix}^{-1} = \frac{1}{ab - c^2} \begin{pmatrix} b & -c \\ -c & a \end{pmatrix}$$

might be useful.) (3p)

d. Use c) to show that the standard error of $\hat{\beta}$ is

$$\text{SE}(\hat{\beta}) = \sqrt{\frac{1}{n_{00}} + \frac{1}{n_{01}} + \frac{1}{n_{10}} + \frac{1}{n_{11}}}.$$

(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 |