CATEGORICAL DATA ANALYSIS, 5 credits (intermediate studies), 3.9.–22.10.2015. Literature: Alan Agresti. An Introduction to Categorical Data Analysis, 2. edition. Lecturer: Pekka Pere.

Examination 23.10.2015

Lay out your reasoning and calculations in detail. The questions are each worth six points. Answer all questions. You may answer in English or Finnish. Please give feedback about the course at

`https://elomake.helsinki.fi/lomakkeet/64563/lomake.html`

You can give feedback until the 31st of October (Sunday). You can take this exam sheet with you so that you have the web address readily available.

1.

a) A 2×2 contingency table is studied (the variables are X and Y). Define the odds ratio and interpret it carefully.

b) A $2 \times 2 \times K$ contingency table is studied (the variables are X, Y, and Z). When does a homogeneous XY association apply? Interpret this situation verbally. When does conditional independence of X and Y apply? Interpret this situation verbally.

c) A logistic regression model is fitted to describe the data of point b). How do homogenous association and conditional independence relate to the parameters of the model? Explain carefully.

2. Kotimaa enterprises, in co-operation with TNS-Gallup and the Research Centre of the Evangelical Lutheran Church of Finland, has conducted a survey ($n = 4930$) about the attitudes of Finns of issues relevant to the Church (table).[1]

Are you	yes	cannot tell	no	Σ
religious	1516	725	2689	4930
spiritual	1586	884	2460	4930
nonreligious	1049	656	3255	4930

a) A 99% confidence interval for the proportion of religious Finns is needed. Explain carefully two likelihood based methods which could be used.

b) Which one of the methods is better in general (for calculating a confidence interval for a proportion)? Explain carefully.

c) Calculate a 99% confidence interval for the proportion of religious Finns (proportion of those who answer yes to the question in the table). An appropriately calculated confidence interval calculated by either method can compose a correct answer.

d) A factory makes 300 parachutes with a new technique. Each of these parachutes is tested (used), and they all open as they should. Calculate a 95% confidence interval for the proportion of parachutes (made with the new technique) which would not open when used.

[1] The composer of the exercise thanks Jaakko Tapaninen for supplying the data 1.10.2013.

3. Ahmed and Hammarstedt (2009)[2] wondered if homosexuals are discriminated against in the market for rental apartments. They sent two almost identical queries for an apartment to 408 landlords (altogether 816 queries) offering an apartment for rental. The first query was formulated so that it was apparent that a heterosexual couple was looking for an apartment. The second query made correspondingly clear that the couple was homosexual. The landlords responded (or did not) as described in the table.

		answered the heterosexual couple		
		yes	no	Σ
answered the homosexual couple	yes	177	4	181
	no	50	177	227
	Σ	227	181	408

a) What would be the null hypothesis of marginal homogeneity be in this context and what would it mean?

b) Test the null hypothesis of marginal homogeneity with this data at the 5% risk level (two-sided test).

c) Explain carefully the intuition of the test in general.

4.

a) Define a logistic regression model with a single predictor (x). Explain it in words.

b) Derive the probability function associated with the model ($\pi(x)$ in the notation of the book).

c) At what value of x does the equality $\pi(x) = 1/2$ hold? Prove it.

d) What is the effect on odds of a unit change in x?

[2] A.M. Ahmed and M. Hammarstedt (2009): Detecting Discrimination against Homosexuals: Evidence from a Field Experiment on the Internet. *Economica*, 76, 588–597.