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.

5th exercise set (9.10.)

Background theory

Independent multinomial sampling (independent binomial sampling being a special case) and multinomial sampling have been discussed during the lectures. Sample sizes for the independent samples are fixed in the former; the total sample size is fixed in the latter. A third possibility is that the frequencies (N_i) in the cells of a contingency table with c cells each follow a Poisson distribution ($N_i \sim \mathsf{Poi}(\mu_i)$):

$$\mathsf{P}(N_i = n_i) = \; e^{-\mu_i} \frac{\mu_i^{n_i}}{n_i!}, \quad n_i = 0, 1, 2, \ldots, \quad i = 1, \ldots, c.$$

The total sample size $N = N_1 + \cdots + N_c$ is then random. The mean and variance of N_i equal both $\mu_i > 0$. Assuming independence of the cell frequencies, the joint probability mass function for N_is is the product of the probability mass functions for each frequency:

$$\mathsf{P}(N_1 = n_1, N_2 = n_2, \ldots, N_c = n_c) = \prod_{i=1}^{c} e^{-\mu_i} \frac{\mu_i^{n_i}}{n_i!}.$$

By the properties of the Poisson distribution, the total count $N = N_1 + \cdots + N_c$ follows $\mathsf{Poi}(\sum_{i=1}^{c} \mu_i)$.

An example might be asylym applicants at Finland 1.1.–31.8.2015 (the most recent published figures) by nationality[1]:

nationality	count
Iraq	3228
Somalia	1282
Albania	583
Afghanistan	411
Syyria	192
Russia	140
other	1179
total	7015

Each count would be treated as a realisation of an independent Poisson variate with a (potentially different) mean $\mu_i, i = 1, \ldots, 7$. (The counts need not be in descending order. The ordering is for illumination of the most important source countries of asylum seekers.)

[1] Statistics on asylum and refugees. The Finnish immigration service. (`Http://www.migri.fi/about_us/statistics/statistics_on_asylum_and_refugees` (read 5.10.2015)

1. Let the frequencies (N_i) of the table cells (c altogether) follow independent $\mathsf{Poi}(\mu_i)$ distributions, and let the total observed sample size be $n = n_1 + \cdots + n_c$. Prove that the joint probability mass function of the N_is, conditional on the observed total sample size n, is multinomial:

$$\frac{n!}{n_1! n_2! \ldots n_c!} \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_c^{n_c}.$$

Here $\pi_i = \mu_i / (\sum_{j=1}^{c} \mu_j)$. (Hint: Agresti 2013, 8.)

2. The contingency table examined is

		Y		
		y_1	y_2	Σ
X	x_1	n_{11}	n_{12}	n_{1+}
	x_2	n_{21}	n_{22}	n_{2+}
	Σ	n_{+1}	n_{+2}	n

(in obvious notation). Fisher's exact test is based on the hypergeometric distribution

$$\mathsf{P}(N_{11} = n_{11}) = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{21}}}{\binom{n}{n_{+1}}} = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{1+} - n_{11}}}{\binom{n}{n_{+1}}}.$$

Above N_{11} is the random frequency of the 1,1-cell and lower case letters denote observed values. The distribution was derived in the lecture by arguments explicitly employing combinatorics. The distribution can be obtained by the reasoning below as well.

Suppose that the table is a result of independent binomial sampling or that the two rows are independent binomial samples with fixed sample sizes n_{1+} and n_{2+}. Let the null hypothesis be equality of proportions (probability of y_1 equals π) in the two corresponding populations. Prove that the hypergeometric distribution arises as the conditional distribution given the sample sizes n_{1+} and n_{2+}. (Hint: Formulate the joint probability mass function of the two samples assuming independent binomial sampling. Note that n_{+1} is likewise a binomial random variate. Devide the joint probability mass function by the conditioning probability mass function.)[2]

3. The (multivariate) Multinomial distribution

$$\mathsf{P}(N_1 = n_1, N_2 = n_2, \ldots, N_c = n_c) = \frac{n!}{n_1! n_2! \ldots n_c!} \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_c^{n_c}$$

is focused at. Above $\sum_{i=1}^{c} N_i = n$ and $\sum_{i=1}^{c} \pi_i = 1$.

a) Prove that the covariance between the frequencies in categories j and k is

$$\mathsf{cov}(N_j, N_k) = -n\pi_j\pi_k.$$

[2]This derivation is given e.g. in Y.M. Bishop, S.E. Fienberg and P.W. Holland (1975): *Discrete Multivariate Analysis*. MIT Press. Cambridge, MA. P. 364.

(Hint: Formulate a Multinomially distributed vector $\sum_{i=1}^{n} \mathbf{Y}_i$, where $\mathbf{Y}_i = [Y_{i1} \dots Y_{ic}]'$. Derive the covariances of the components of $\mathbf{Y}_i$ and, on the grounds of independence of the observations, the covariances of the components of $\sum_{i=1}^{n} \mathbf{Y}_i$.

b) Prove that the correlation between the frequencies in categories j and k is

$$\mathsf{cor}(N_j, N_k) = \frac{-\pi_j \pi_k}{\sqrt{\pi_j(1-\pi_j)\pi_k(1-\pi_k)}}.$$

c) Let the number of categories be $c = 2$. Prove that the correlation between the frequencies in categories 1 and 2 is

$$\mathsf{cor}(N_1, N_2) = -1.$$

Provide intuition for the result.[3]

4. It will be shown that the statistics z_s^2 and X^2

$$z_s^2 \equiv \frac{(\hat{\pi}_1 - \hat{\pi}_2)^2}{\hat{\pi}(1-\hat{\pi})/n_{1+} + \hat{\pi}(1-\hat{\pi})/n_{2+}} = \sum_{i=1}^{2}\sum_{i=j}^{2} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \equiv X^2$$

explored in exercises 3.6 and 4.5 are score statistics. The first two rows of the data

n_{11}	n_{12}	n_{1+}
n_{21}	n_{22}	n_{2+}
n_{+1}	n_{+2}	n

are assumed to compose of two independent binomial samples with (fixed) sample sizes n_{1+} and n_{2+}. The probabilities for the cells are as follows:

π_1	$1-\pi_1$	1
π_2	$1-\pi_2$	1

The null hypothesis is that $\pi_1 = \pi_2$.

In the case of a multivariate parameter $\boldsymbol{\pi} = [\pi_1 \dots \pi_k]'$ the score statistic is

$$\nabla l(\hat{\boldsymbol{\pi}}_0)' \boldsymbol{I}(\hat{\boldsymbol{\pi}}_0)^{-1} \nabla l(\hat{\boldsymbol{\pi}}_0),$$

if the null hypothesis binds all parameters. Above $l(\boldsymbol{\pi})$ is the logarithm of the likelihood function for the parameters,$\nabla l(\hat{\boldsymbol{\pi}}_0)$ is the gradient of the logarithm of the likelihood function

$$\left[\frac{\partial}{\partial \pi_1} l(\boldsymbol{\pi}) \dots \frac{\partial}{\partial \pi_k} l(\boldsymbol{\pi})\right]'$$

evaluated at the restricted (as expressed by the null hypothesis) MLE $\hat{\boldsymbol{\pi}}_0$ and $\boldsymbol{I}(\hat{\boldsymbol{\pi}}_0)^{-1}$ is the inverse of the Fisher information matrix ($k \times k$) evaluated at the restricted MLE. In the present circumstance $k = 2$ and $\boldsymbol{\pi} = [\pi_1 \;\; \pi_2]'$.

[3]Points a) and b) are exercise 1.19 in Alan Agresti (2013): *Categorical Data Analysis, 3. laitos*. Wiley. Hoboken, New Jersey.

a) Prove that

$$l(\boldsymbol{\pi}) = n_{11}\log\pi_1 + n_{12}\log(1-\pi_1) + n_{21}\log\pi_2 + n_{22}\log(1-\pi_2)$$

b) Prove that

$$\nabla l(\boldsymbol{\pi}) = \left[\frac{n_{11}}{\pi_1} - \frac{n_{12}}{1-\pi_1} \qquad \frac{n_{21}}{\pi_2} - \frac{n_{22}}{1-\pi_2}\right]'.$$

c) Prove that

$$\nabla l(\hat{\boldsymbol{\pi}}_0) = \left[n\frac{n_{11}n_{22} - n_{12}n_{21}}{n_{+1}n_{+2}} \qquad n\frac{n_{12}n_{21} - n_{11}n_{22}}{n_{+1}n_{+2}}\right]'.$$

(Hint: $\hat{\pi}_{1,0} = \hat{\pi}_{2,0} = n_{+1}/n$, where $\hat{\pi}_{1,0}$ and $\hat{\pi}_{2,0}$ are the restricted MLEs of π_1 and π_2.)

d) Prove that

$$-\begin{bmatrix} \frac{\partial^2}{\partial\pi_1^2}l(\boldsymbol{\pi}) & \frac{\partial^2}{\partial\pi_1\partial\pi_2}l(\boldsymbol{\pi}) \\ \frac{\partial^2}{\partial\pi_2\partial\pi_1}l(\boldsymbol{\pi}) & \frac{\partial^2}{\partial\pi_2^2}l(\boldsymbol{\pi}) \end{bmatrix} = \begin{bmatrix} \frac{n_{11}}{\pi_1^2} + \frac{n_{12}}{(1-\pi_1)^2} & 0 \\ 0 & \frac{n_{21}}{\pi_2^2} + \frac{n_{22}}{(1-\pi_2)^2} \end{bmatrix}.$$

e) Prove that the Fisher information matrix and the inverse of it evaluated at the restricted MLE $\hat{\boldsymbol{\pi}}_0$ are

$$\boldsymbol{I}(\boldsymbol{\pi}) = \begin{bmatrix} \frac{n_{1+}}{\pi_1(1-\pi)} & 0 \\ 0 & \frac{n_{2+}}{\pi_2(1-\pi_2)} \end{bmatrix}$$

and

$$\boldsymbol{I}(\hat{\boldsymbol{\pi}}_0)^{-1} = \begin{bmatrix} \frac{n_{+1}n_{+2}}{n^2 n_{1+}} & 0 \\ 0 & \frac{n_{+1}n_{+2}}{n^2 n_{2+}} \end{bmatrix}.$$

f) Finally prove that

$$\nabla l(\hat{\boldsymbol{\pi}}_0)'\boldsymbol{I}(\hat{\boldsymbol{\pi}}_0)^{-1}\nabla l(\hat{\boldsymbol{\pi}}_0) = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1+}n_{2+}n_{+1}n_{+2}}$$

or that the score statistic matches with the statistics z_s^2 and X^2 (by exercises 3.6 and 4.5).

5. The tables below relate to matched pairs data (notation obvious):

		Y		
		y_1	y_2	Σ
X	x_1	π_{11}	π_{12}	π_{1+}
	x_2	π_{21}	π_{22}	π_{2+}
	Σ	π_{+1}	π_{+2}	1

		Y		
		y_1	y_2	Σ
X	x_1	n_{11}	n_{12}	n_{1+}
	x_2	n_{21}	n_{22}	n_{2+}
	Σ	n_{+1}	n_{+2}	n

a) The interesting question for matched pairs data is, does marginal homogeneity $\pi_{1+} = \pi_{+1}$ apply. Prove that $\pi_{12} = \pi_{21}$ under marginal homogeneity.

b) Does marginal homogeneity imply that $\pi_{2+} = \pi_{+2}$? State your reasons for your answer.

c) Does marginal homogeneity imply that $\pi_{11} = \pi_{22}$ as well? State your reasons for your answer.

6. McNemar's test statistic for marginal homogeneity can be written in alternative ways:

$$\begin{aligned} z &= \frac{n_{12} - 0,5 \times n^*}{\sqrt{n^* \times 0,5 \times 0,5}} \\ &= \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} \\ &= \frac{-(n_{21} - 0,5 \times n^*)}{\sqrt{n^* \times 0,5 \times 0,5}} \end{aligned}$$

(cf. formula (8.1) in the book). Derive the second and third equality above.