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.

1st exercise set (11.9.)

1. Let there be $n > 0$ elements in set A.

a) Explain why

$$\binom{n}{k}$$

($0 \leq k \leq n$) is the number of different arrangements of the objects when A is composed of k objects of one kind and $n - k$ of another kind ("a" and "b"). (Hint: Mark by N the number of different arrangements. Reason the number of arrangements if the a-objects could be differentiated. Reason next the number of arrangements if also the b-objects could be differentiated. Set the number of arrangements you have reasoned equal to $n!$ (explain this as well) and solve N.)

b) Explain why

$$\frac{n!}{n_{1!}n_{2!}\ldots n_k!}$$

($i = 1, \ldots, k$; $n_1 + \cdots + n_k = n$) is the number of different arrangements of the objects when A is composed of k subsets each with n_i similar objects in the subset but different from the other objects.

2.

a) Explain carefully the symbols in the binomial formula

$$\mathsf{P}(Y = y) = \binom{n}{k}\pi^y(1 - \pi)^{n-y}$$

and the justification of it.

b) Explain carefully the symbols in the multinomial formula

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

and the justification of it.

3. Let Y_i:t ($i = 1, \ldots, n$) be independently distributed Bernoulli random variates. Y_i equals 1 with probability π and 0 with probability $1 - \pi$ ($\pi \in (0, 1)$).

a) Derive the mean and variance of Y_i.

b) Derive the mean and variance of $P = \sum_{i=1}^n Y_i/n$. Is P an unbiased estimator for π?

c) Explain carefully why P follows (approximatively) the Normal distibution when n is large. What are the mean and variance of this Normal distribution? (Hint: Central limit theorem.)

4. Let the log-likelihood function depend on a single parameter θ:

$$l(\theta; \mathbf{y}) \equiv l(\theta).$$

Here $\mathbf{y}$ is the vector of observations. Explain the geometric intuition of the likelihood ratio

$$2[l(\hat{\theta}) - l(\theta_0)],$$

Wald

$$\left.\sqrt{i(\theta)}\right|_{\theta=\hat{\theta}} (\hat{\theta} - \theta_0) = \sqrt{i(\hat{\theta})}(\hat{\theta} - \theta_0),$$

and Rao's score

$$\left.\frac{l'(\theta)}{\sqrt{i(\theta)}}\right|_{\theta=\theta_0} = \frac{l'(\theta_0)}{\sqrt{i(\theta_0)}}$$

test statistics. Above θ_0 is the value of θ under the null hypothesis, $\hat{\theta}$ is the maximum likelihood estimator (MLE) of θ, $l'(\theta)$ is the derivative of the log-likelihood function with respect to θ and

$$i(\theta) = \mathsf{E}\left[-\frac{\partial^2 l(\theta)}{\partial \theta^2}\right]$$

is the Fisher or expected information for θ. (Hint: Figures 1–3 in A. Buse (1982): The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note. *American Statistician*, 36, 153–157.)

5.

Let us see how the previous results relate to likelihood inference in the present context. Let y_i (1 or 0) be observed values of a Bernoulli distributed random variate Y_i ($i = 1, \dots, n$) and $y = \sum_{i=1}^{n} y_i$.

a) Derive the log of the likelihood function for π:

$$l(\pi) = y \log(\pi) + (n - y) \log(1 - \pi).$$

b) Derive the first derivative of it:

$$l'(\pi) = \frac{y - n\pi}{\pi(1 - \pi)}.$$

c) Derive the MLE for π:

$$\hat{\pi} = p = \frac{y}{n} = n^{-1} \sum_{i=1}^{n} y_i.$$

d) In the single parameter case the asymptotic variance of the MLE is (under standard conditions) the inverse of the Fisher information for the parameter:

$$i(\pi) = \mathsf{E}\left[-\frac{\partial^2 l(\pi)}{\partial \pi^2}\right].$$

Calculate $[i(\pi)]^{-1}$ and compare the result with the variance of P derived above.