Transmission and ‘likelihood ratio’

Alain Favre & Luc-Olivier Pochon

October 27, 2003

1 Introduction

The purpose of this note is to explain some results given in a paper of Martin Zwick [2]. We want to study the link between the likelihood ratio (always positive according to the Gibbs theorem[1]):

$$T(model) = \sum p \log \frac{p}{q}$$

and the different levels of entropy (transmission).

We shall use: p, the observed frequencies, and q, the frequencies computed to maximise $\sum q \log q$ which fullfills the model constraints. Moreover the entropy of an observation is defined as:

$$U(p) = -\sum p \log p$$

The question is thus to prove that:

$$U(q) - U(p) = \sum p \log \frac{p}{q} \tag{1}$$

which amounts to prove:

$$\sum (p - q) \log q = 0 \tag{2}$$

We shall first deal with a general model showing that if there exists a solution which fullfills the model constraints, then (2) can be shown. Afterwards, for didactic purposes we shall consider 2 particuliar cases. Firstly the case of 2 variables each taking 2 values (This result obviously extend to 2 variables with any finite number of values), then we shall consider the case where the model constraints are expressed as a set of sums of $q_{ijk...}$ with more than 2 variables as proposed in Zwick [2].

The section 6 recall the notion of information in this context.

[1] The Gibbs theorem [3] prove that if $\{p_i\}_{i=1,n}$ et $\{q_i\}_{i=1,n}$ are 2 families of numbers the sum of which are 1, then :

$$-\sum p_i \log p_i \leq -\sum p_i \log q_i$$

This results is based on the convexity of the $x \log x$ function.

2 General model

We shall prove that

$$U(q) - U(p) = \sum p \log \frac{p}{q}$$

Which amount to say: $\sum_I (p - q) \log q = 0$, as soon as $(q_i)_{i=1,n}$ maximises $-\sum q \log q$ with the linear constraints:

$$\begin{array}{l} C_0 : \sum p_i = \sum q_i = 1 \\ C_k : \sum_{I_k} q_i = \sum_{I_k} p_i = \alpha_k \;\; k = 1, m \end{array}$$

With a set of I_k containing the indices 1 to n (without beeing necessarily disjoints).

The equation $\sum (p - q) \log q = 0$ is true as soon as the coefficients $\log q$ can be expressed as a sum the terms of which are constant or at least constants over one I_k.

We note $J_l = \{i | l \in I_i\}$ and write:

$$\varphi_k = \sum_{i \in I_k} q_i - \alpha_k$$

we shall then use the Lagrange function(for constrained maxima):

$$F = - \sum_{i=1,n} q_i \log q_i + \sum \lambda_j \varphi_j$$

By derivation we obtain:

$$\frac{\partial F}{\partial q_l} = -(1 + \log q_l) + \sum_{J_l} \lambda_j$$

and with a derivative equating 0 we have:

$$\log q_l = - \sum_{j \in J_l} \lambda_j - 1$$

$$\begin{array}{l} \sum (p_i - q_i) \log q_i = \\ \sum (p_i - q_i)(- \sum_{J_i} \lambda_j - 1) = \\ - \sum_{J_i} \lambda_k \sum_{I_k} (p_j - q_j) - \sum (p_i - q_i) = 0 \end{array}$$

Using this model depends on the existence of a solution of the Lagrange equations system. Existing solutions can be derived from the iterative proportional fitting algorithm' [1]. What follows is considered in this context.

3 Case of a 4 position table

	B	$\bar{B}$	
A	p_{11}	p_{12}	t_1
$\bar{A}$	p_{21}	p_{22}	t_2
	u_1	u_2	1

in this case the solution which maximises the entropy is given by $q_{ij} = t_i * u_j$ (we shall see this hereafter). Therefore:

$$\sum_{i,j=1}^{2} (p_{ij} - q_{ij}) \log q_{ij} =$$
$$\sum_{i,j=1}^{2} (p_{ij} - q_{ij})(\log t_i + \log u_j) =$$
$$\sum_{i,j=1}^{2} (p_{ij} - q_{ij}) \log t_i + \sum_{i,j=1}^{2} (p_{ij} - q_{ij}) \log u_j = 0$$

Since we have:

$$\sum_{i,j=1}^{2} (p_{ij} - q_{ij}) \log t_i =$$
$$\sum_{i=1}^{2} \log t_i \sum_{j=1}^{2} (p_{ij} - q_{ij}) =$$
$$\sum_{i=1}^{2} \log t_i [(p_{i1} + p_{i2}) - (q_{i1} + q_{i2})] =$$
$$\sum_{i=1}^{2} \log t_i \quad (t_i - t_i) = 0$$

4 Proving that $q_{ij} = t_i * u_j$

The values of q maximising $U(q)$ are the theoretical frequencies defined as the products of the margin distributions: $q_{ij} = t_i * u_j$

How to maximise:

$$-\sum_{i,j=1}^{2} q_{ij} \log q_{ij}$$

With the constraints:

$$\begin{cases} q_{11} + q_{12} = p_{11} + p_{12} = t_1 \\ q_{21} + q_{22} = p_{21} + p_{22} = t_2 \\ q_{11} + q_{21} = p_{11} + p_{21} = u_1 \\ q_{12} + q_{22} = p_{12} + p_{22} = u_2 \end{cases}$$

We write:

$$\begin{cases} \varphi_1 = q_{11} + q_{12} - t_1 \\ \varphi_2 = q_{21} + q_{22} - t_2 \\ \varphi_3 = q_{11} + q_{21} - u_1 \\ \varphi_4 = q_{12} + q_{22} - u_2 \end{cases}$$

The solution is obtained by equating the derivatives to 0 of the function

$$-\sum_{i,j=1}^{2} q_{ij} \log q_{ij} + \sum_{j} \lambda_j \varphi_j$$

derivating with respect to à q_{ij}, we obtain:

$$\begin{cases} -(1 + \log q_{11}) = \lambda_1 + \lambda_3 \\ -(1 + \log q_{12}) = \lambda_1 + \lambda_4 \\ -(1 + \log q_{21}) = \lambda_2 + \lambda_3 \\ -(1 + \log q_{22}) = \lambda_2 + \lambda_4 \end{cases}$$

and there are 4 more constraints: $\varphi_j = 0$.

We straigtforward obtain: $\log \frac{q_{11}}{q_{12}} = \lambda_4 - \lambda_3 = \log \frac{q_{21}}{q_{22}}$ ce qui implique:

$$\begin{cases} q_{11} = \mu * q_{12} \\ q_{21} = \mu * q_{22} \end{cases}$$

And it comes out that:

$$q_{ij} = t_i * u_j \quad \forall i, j$$

5 Case of three variables

The three variables are defined as A, B et C with the model $AB : BC : AC$. We denote p_{ijk} the observed frequencies.

The equation $\sum (p-q) \log q = 0$ is true as soon as the coefficients $\log q_{ijk}$ can be expressed as a sum: $cst + \eta_{1i} + \eta_{2j} + \eta_{3k}$ and that q_{ijk} satisfy the constraints[2]:

$$\begin{cases} \sum_{j,k=1}^{n,p} q_{ijk} = \sum_{j,k=1}^{n,p} p_{ijk} = t_{1i} \ \forall i = 1, m \\ \sum_{i,k=1}^{m,p} q_{ijk} = \sum_{i,k=1}^{m,p} p_{ijk} = t_{2j} \ \forall j = 1, n \\ \sum_{i,j=1}^{m,n} q_{ijk} = \sum_{i,j=1}^{m,n} p_{ijk} = t_{3k} \ \forall k = 1, p \end{cases}$$

Because:

$$\begin{array}{l} \sum_{i,j,k=1}^{m,n,p} (p_{ijk} - q_{ijk}) \log q_{ijk} = \\ \sum_{i,j,k=1}^{m,n,p} (p_{ijk} - q_{ijk})(cst + \eta_{1i} + \eta_{2j} + \eta_{3k}) = \\ \sum_{i,j,k=1}^{m,n,p} (p_{ijk} - q_{ijk}) cst + \sum_{i,j,k=1}^{m,n,p} (p_{ijk} - q_{ijk}) \eta_{1i} + \\ \sum_{i,j,k=1}^{m,n,p} (p_{ijk} - q_{ijk}) \eta_{2j} + \sum_{i,j,k=1}^{m,n,p} (p_{ijk} - q_{ijk}) \eta_{3k} = 0 \end{array}$$

Since:

$$\begin{array}{l} \sum_{i,j,k=1}^{m,n,p} (p_{ijk} - q_{ijk}) cst = \\ cst \sum_{i,j,k=1}^{m,n,p} (p_{ijk} - q_{ijk}) = cst(1-1) = 0 \end{array}$$

and that:

$$\begin{array}{l} \sum_{i,j,k=1}^{m,n,p} (p_{ijk} - q_{ijk}) \eta_{1i} = \\ \sum_{i=1}^{m} \eta_{1i} \sum_{j,k=1}^{n,p} (p_{ijk} - q_{ijk}) = 0 \end{array}$$

It is sufficient to show that $\log q_{ijk}$ is expressed as such a sum when the q_{ijk} maximise $-\sum q \log q$ with the constraints given hereafter.

We write:

$$\begin{cases} \varphi_{1i} = \sum_{j,k} q_{ijk} - t_{1i} \ \forall i \\ \varphi_{2j} = \sum_{i,k} q_{ijk} - t_{2j} \ \forall j \\ \varphi_{3k} = \sum_{i,j} q_{ijk} - t_{3k} \ \forall k \end{cases}$$

The solution is obtained by computing and equating to 0 the derivatives of

$$F = -\sum_{i,j,k=1}^{m,n,p} q_{ijk} \log q_{ijk} + \sum_{p,q} \lambda_{pq} \varphi_{pq}$$

with respect to à q_{ijk}, we obtain:

$$\frac{\partial F}{\partial q_{ijk}} = -(1 + \log q_{ijk}) + \lambda_{1i} + \lambda_{2j} + \lambda_{3k} = 0 \ \forall i, j, k$$

[2]The constant cst could be added to one of the coefficients η

This immediately implies the wanted result:

$$\log q_{ijk} = -(1 + \lambda_{1i} + \lambda_{2j} + \lambda_{3k}) \ \forall i, j, k$$

To compute the solution, the following constraints[3] are also considered $\varphi_{pq} = 0$. However the solution is computed by the iterative proportional fitting algorithm [1].

6 Last comments

For the case where there are 3 sets of datae: observed (obs), modeled (mod), reference (ref). We consider: $T(obs/mod)$ (the error of the model or the lost constraints) and $T(obs/ref)$ the difference of which is $T(obs/ref) - T(obs/mod)$ represent the constraints captured in the model.

The information (relative) brought by the model is computed as $1 - \frac{T(obs/mod)}{T(obs/ref)}$

References

[1] Haberman, S. J. (1972). Log-linear fit for contingency tables - Algorithm AS51. *Applied Statistics*, 21, 218225.

[2] Zwick, M. (2002). An overview of reconstructability analysis. *Proceedings of* 12^{th} *International World Organisation of Systems and Cybernetics*, Pittsburgh, March 24-26.

[3] Watanabe, S. (1969). *Knowing and Guessing. A quantitative Study of Inference and Information.* New York: John Wiley and Sons.

[3] there are $(n \times m \times p) + (n + m + p)$ of them and the same number of variables.