derivation of mutual information

Canonical name	DerivationOfMutualInformation
Date of creation	2013-03-22 15:13:38
Last modified on	2013-03-22 15:13:38
Owner	tdunning (9331)
Last modified by	tdunning (9331)
Numerical id	5
Author	tdunning (9331)
Entry type	Derivation
Classification	msc 94A17

The maximum likelihood estimater for mutual information is identical (except for a scale factor) to the generalized log-likelihood ratio for multinomials and closely related to Pearson's χ^2 test. This implies that the distribution of observed values of mutual information computed using maximum likelihood estimates for probabilities is χ^2 distributed except for that scaling factor.

In particular if we sample each of X and Y and combine the samples to form N tuples sampled from $X \times Y$. Now define $T(x, y)$ to be the total number of times the tuple (x, y) was observed. Further define $T(x, *)$ to be the number of times that a tuple starting with x was observed and $T(*, y)$ to be the number of times that a tuple ending with y was observed. Clearly, $T(*, *)$ is just N, the number of tuples in the sample. From the definition, the generalized log-likelihood ratio test of independence for X and Y (based on the sample of tuples) is

$$-2log\lambda = 2 \sum_{xy} T(x, y) \log \frac{\pi_{x|y}}{\mu_x}$$

where

$$\pi_{x|y} = T(x, y) / \sum_{x} T(x, y)$$

and

$$\mu_x = T(x, *)/T(*, *)$$

This allows the log-likelihood ratio to be expressed in terms of row and column sums,

$$-2log\lambda = 2 \sum_{xy} T(x, y) \log \frac{T(x, y)T(*, *)}{T(x, *)T(*, y)}$$

This reduces to the following expression in terms of maximum likelihood estimates of cell, row and column probabilities,

$$-2log\lambda = 2 \sum_{xy} T(x, y) \log \frac{\pi_{xy}}{\mu_{*y}\mu_{x*}}$$

This can be rearranged into

$$-2log\lambda = 2N \left[\sum_{xy} \pi_{xy} \log \pi_{xy} \sum_{x} \mu_{x*} \log \mu_{x*} \sum_{y} \mu_{*y} \log \mu_{*y} \right] = 2N\hat{I}(X;Y)$$

where the hat indicates a maximum likelihood estimation of $I(X;Y)$.

This also gives the asymptotic distribution of $\hat{I}(X;Y)$ as $2N$ times a χ^2 deviate.