The Korean Communications in Statistics
Vol. 13 No. 2, 2006, pp. 297-307

Mutual Information and Redundancy for Categorical Data

Chong Sun Hong[1] and Beom Jun Kim[2]

Abstract

Most methods for describing the relationship among random variables require specific probability distributions and some assumptions of random variables. The mutual information based on the entropy to measure the dependency among random variables does not need any specific assumptions. And the redundancy which is a analogous version of the mutual information was also proposed. In this paper, the redundancy and mutual information are explored to multi-dimensional categorical data. It is found that the redundancy for categorical data could be expressed as the function of the generalized likelihood ratio statistic under several kinds of independent log-linear models, so that the redundancy could also be used to analyze contingency tables. Whereas the generalized likelihood ratio statistic to test the goodness-of-fit of the log-linear models is sensitive to the sample size, the redundancy for categorical data does not depend on sample size but its cell probabilities itself.

Keywords : Entropy; Goodness of fit; Joint Independence; Log-linear Model; Redundancy.

1. 서론

변수들간의 선형관계를 파악하는 피어슨의 선형상관계수나 순위로 주어진 변수들의 관계를 측정하는 순위상관계수와는 달리 상호정보(mutual information)는 연속형이나 이산형, 명목형과 순서형, 정규성 여부, 선형과 비선형 등 여러 가지 가정에 관계없이 변수들 사이의 관계를 설명할 수 있다. Shannon (1948)에 의해 처음 소개된 상호정보는 어떤 확률변수에 포함되어 있는 다른 확률변수의 정보량인 엔트로피(entropy)에 대한 전달과 측정의 이론으로부터 출발하였다. 그 후 Gelfand와 Yaglom (1959), Cover와 Thomas (1991)를 비롯한 많은 학자들에 의해 개념이 확장되어 일반화되었고, 상호정보와 유사한 개념인 리둔던시(redundancy)[3]는 Fraser와 Swinney (1986)에

1) Professor, Department of Statistics, Sungkyunkwan University, Seoul, 110-745, Korea. Correspondence : cshong@skku.ac.kr
2) Lecturer, Department of Statistics, Sungkyunkwan University, Seoul, 110-745, Korea.
3) 통계학용어집에 '과잉' 또는 '중복'으로 해석되고 있지만, 본 논문의 내용과 일치하지 않는 의미이기 때문에 '리둔던시'로 표현함.

의해 제안되었으며, Palus와 Pivka (1995) 등 많은 학자들에 의해 개념이 확장되었다.

본 논문에서는 구체적인 이론적 확률분포나 특정 종속모형의 가정이 필요없이 확률변수들간의 관계를 파악하기 위해 상호정보와 유사한 개념인 리둔던시를 이용하고 범주형 자료의 모형에 이를 적용시켜보고자 한다.

2절에서는 확률변수에 대한 정보를 나타내는 엔트로피와 상호정보 및 리둔던시의 개념을 설명한 후 리둔던시와 엔트로피 및 상호정보와의 관계에 대해 알아보고, 3절에서는 리둔던시를 범주형 자료의 경우에 적용하여 변수들 간의 종속여부와 범주형 자료를 설명하는 특정한 로그선형모형의 적합도 검정통계량과의 관계를 유도하며, 4절에서는 3절에서 논의한 모형들에 대해 자료를 생성하여, 표본크기 변화에 따라 리둔던시와 적합도 검정통계량 값의 결과를 바탕으로 각 성격과 특성을 살펴본다. 이에 대한 결론을 5절에서 논의한다.

2. 엔트로피와 상호정보 그리고 리둔던시

Shannon (1948)은 어떤 시스템의 무질서한 정도를 나타내는 엔트로피(entropy)를 정의하였고, 이에 대한 정보량은 $H(X)$로 표시하였다. X를 이산형 확률변수라고 하고, 확률밀도함수를 $p(x)=\Pr\{X=x\}$라 하면, 이산형 확률변수 X에 대해 엔트로피 $H(X)$는 식 (2.1)과 같이 정의할 수 있다. 이절에서 설명하는 엔트로피의 개념은 Abramson (1963), Gallage (1968) 그리고 Cover와 Thomas (1991)를 참조하였다.

$$H(X)=-\sum_{x\in X} p(x)\log p(x). \tag{2.1}$$

식(2.1)을 일반화하여 결합분포 $p(x_1,x_2,\ldots,x_n)$를 갖는 n개의 확률변수 $X_1,X_2,\ldots,X_n$에 대한 결합 엔트로피(joint entropy)는 식 (2.2)와 같이 정의된다.

$$H(X_1,X_2,\ldots,X_n)=-\sum\sum\cdots\sum p(x_1,x_2,\ldots,x_n)\log p(x_1,x_2,\ldots,x_n). \tag{2.2}$$

그리고 Y가 주어진 경우 X의 조건부 엔트로피(conditional entropy)는 식 (2.3)과 같이 정의된다.

$$H(X|Y)=-\sum_{y\in Y}\sum_{x\in X} p(x,y)\log p(x|y). \tag{2.3}$$

위 정의에서 Y를 확장한 결합 조건부 엔트로피(joint conditional entropy)는 다음과 같이 정의된다.

$$H(X_1,\ldots,X_n|Y_1,\ldots,Y_m)=-\sum\cdots\sum p(x_1,\ldots,x_n,y_1,\ldots,y_m)\times\log p(x_1,\ldots,x_n|y_1,\ldots,y_m).$$

상호정보(mutual information)는 식 (2.1)과 같은 Shannon의 엔트로피로를 통해 얻을 수 있고, 식 (2.4)와 같은 관계를 갖는다(DeGroot 1962).

$$\begin{aligned} I(X_1;X_2) &= H(X_1)-H(X_1|X_2) \\ &= H(X_1)+H(X_2)-H(X_1,X_2). \end{aligned} \tag{2.4}$$

변수간의 통계적 종속을 측정하기 위한 결합 확률밀도함수 $p(x_1,x_2)$를 갖는 이변량 확률변수 (X_1,X_2)에 대한 상호정보는 식 (2.5)와 같이 정의된다(Abramson 1963,

Gallager 1968).

$$I(X_1;X_2)=\sum\sum p(x_1,x_2)log(p(x_1,x_2)/p(x_1)p(x_2)). \tag{2.5}$$

이변량 확률변수 X_1과 X_2에 대한 상호정보 $I(X_1;X_2)$는 $X_2=x_2$가 발생하고, $X_1=x_1$의 발생에 관한 정보의 양이다. 그러므로 상호정보는 두 관측값의 일반적인 종속에 관한 측도로 사용된다. 3변량 확률변수에 대한 상호정보는 식 (2.6)과 같이 정의된다.

$$I(X_1;X_2;X_3)=\sum\sum\sum p(x_1,x_2,x_3) \times log\left(\frac{p(x_1,x_2)p(x_1,x_3)p(x_2,x_3)}{p(x_1)p(x_2)p(x_3)p(x_1,x_2,x_3)}\right). \tag{2.6}$$

식 (2.6)을 n개의 확률변수로 확장할 수 있으며, p개의 확률벡터에 대하여도 확장할 수 있다. 확률벡터에 대해 일반화시킨 상호정보는 다음과 같이 엔트로피로 설명된다 (Wienholt와 Sendhoff 1996).

$$\begin{aligned} I(\vec{X}_1;\cdots;\vec{X}_p)&=H(\vec{X}_1)+\cdots+H(\vec{X}_p) \\ &\quad -H(\vec{X}_1,\vec{X}_2)-\cdots-H(\vec{X}_{p-1},\vec{X}_p) \\ &\quad \vdots \\ &\quad (-1)^{p+1}H(\vec{X}_1,\vec{X}_2,\cdots,\vec{X}_p). \end{aligned}$$

상호정보에서도 조건부를 고려할 수 있는데, 예를 들어 Y_1에 대해 조건부 상호정보 (conditional mutual information) $I(X_1;X_2|Y_1)$는 식 (2.7)과 같이 정의한다.

$$\begin{aligned} I(X_1;X_2|Y_1)&=\sum\sum\sum p(x_1,x_2,y_1)log\left(\frac{p(x_1,x_2|y_1)}{p(x_1|y_1)p(x_2|y_1)}\right) \\ &=H(X_1|Y_1)+H(X_2|Y_1)-H(X_1,X_2|Y_1). \end{aligned} \tag{2.7}$$

Fraser와 Swinney (1986)는 상호정보와 다른 형태의 통계적 종속을 추정하기 위한 방법을 제안하였다. n변량 리둔던시(redundancy)는 n개의 확률변수 $X_1,X_2,...,X_n$에 포함되어 있는 공통정보의 양을 의미하고 앞에서 정의한 엔트로피에 대해 표현하면 식 (2.8)과 같다(Prichard와 Theiler 1995, Palus 1993).

$$R(X_1;X_2;...;X_n)=\sum_{i=1}^{n}H(X_i)-H(X_1,X_2,...,X_n). \tag{2.8}$$

한 변수 이상의 상호정보에 대한 정의와 n개의 확률변수에 대해 한 변수와 나머지 $n-1$ 변수간의 상호정보의 정의를 이용하고, 식 (2.6)의 상호정보와 유사한 형태로 나타내면 식 (2.9)와 같다(Wienholt와 Sendhoff 1996).

$$\begin{aligned} R(X_1;X_2;...;X_n)&=I(X_1;X_2,...,X_n)+I(X_2;X_3,...,X_n)+...+I(X_{n-1};X_n) \\ &=\sum\cdots\sum p(x_1,...,x_n)log\left(\frac{p(x_1,...,x_n)}{p(x_1)\cdots p(x_n)}\right). \end{aligned} \tag{2.9}$$

따라서 확률변수 X_1과 X_2에 대한 리둔던시는 식 (2.4)의 상호정보와 일치한다. 즉, $R(X_1;X_2)=I(X_1;X_2)$이다. p개의 확률벡터에 대한 리둔던시는 다음과 같이 정의한다.

$$R(\vec{X}_1;\vec{X}_2;\cdots;\vec{X}_p)=\sum_{i=1}^{p}H(\vec{X}_i)-H(\vec{X}_1,\vec{X}_2,...,\vec{X}_p).$$

$\vec{Y}_1$이 주어진 경우 $\vec{X}_1$과 $\vec{X}_2$의 조건부 리둔던시는

$$R(\vec{X}_1;\vec{X}_2|\vec{Y}_1)=H(\vec{X}_1|\vec{Y}_1)+H(\vec{X}_2|\vec{Y}_1)-H(\vec{X}_1,\vec{X}_2|\vec{Y}_1)$$

으로 정의하며(Palus와 Pivka 1995), $R(\vec{X}_1;\vec{X}_2|\vec{Y}_1)=I(\vec{X}_1;\vec{X}_2|\vec{Y}_1)$임을 확인할 수 있다. p개의 확률벡터에 대한 조건부 리둔던시를 엔트로피와 확률로 나타내면 다음과 같다.

$$\begin{aligned}R(\vec{X}_1;\ldots;\vec{X}_p|\vec{Y}_1)&=\sum_{i=1}^{p}H(\vec{X}_i|\vec{Y}_1)-H(\vec{X}_1,\ldots,\vec{X}_p|\vec{Y}_1)\\&=\sum\cdots\sum p(\vec{x}_1,\ldots,\vec{x}_p,\vec{y}_1)\,log\left(\frac{p(\vec{x}_1,\ldots,\vec{x}_p|\vec{y}_1)}{p(\vec{x}_1|\vec{y}_1)\cdots p(\vec{x}_p|\vec{y}_1)}\right).\end{aligned}\tag{2.10}$$

3. 범주형 자료에서 리둔던시

이변량 확률변수 X_1과 X_2에 대해 X_1은 $i=1,\ldots,I$의 값을 갖고, X_2는 $j=1,\ldots,J$의 값을 갖는다고 하면 (i,j)칸의 확률은 p_{ij}이고, p_{i+}와 p_{+j}는 각 범주의 주변 확률이다. X_1과 X_2의 리둔던시는 식 (2.9)를 이용하여 식 (3.1)로 정의하고, 이는 식 (2.5)와 비교하여 상호정보와 같음을 다시 확인한다.

$$R(X_1;X_2)=I(X_1;X_2)=\sum_i\sum_j p_{ij}\,log\left(\frac{p_{ij}}{p_{i+}p_{+j}}\right).\tag{3.1}$$

식 (3.1)에서 정의한 리둔던시의 오른쪽 항에 있는 로그 함수에서 분모부분은 표본크기 N의 범주형 자료에서 완전독립성 모형 하에서 기대 칸확률(expected cell ptobability)과 동일하다: 즉, $\hat{p}_{ij}=p_{i+}\,p_{+j}$. 그러므로 식 (3.1)으로부터 2차원 범주형 자료에서의 리둔던시는 완전독립성 모형에 대한 적합도 검정통계량인 일반화 가능도비 검정통계량(generalized likelihood ratio test statistic) G^2에 대해 $1/2N$배의 값을 갖는 것을 알 수 있다. 이 관계는 Brillinger (2004), Brillinger와 Guha (2006) 등 여러 문헌에서 언급되었는데 이러한 관계를 확장하여 n차원 범주형 자료를 설명하는 여러 종류의 로그선형모형들에 대한 리둔던시와 일반화 가능도비 검정통계량과의 관계를 다음과 같이 정리할 수 있다

정리 1.

n차원 범주형 자료에서 리둔던시 $R(X_1;\ldots;X_n)$는 식 (2.9)로부터 다음과 같이 정리된다.

$$\begin{aligned}R(X_1;\ldots;X_n)&=\sum\cdots\sum p(x_{1c_1},\ldots,x_{nc_n})log\left(\frac{p(x_{1c_1},\ldots,x_{nc_n})}{p(x_{1c_1})\cdots p(x_{nc_n})}\right)\\&=\frac{1}{2N}G^2,\end{aligned}\tag{3.2}$$

여기서 G^2는 n차원 범주형 자료에 대한 로그선형모형 중에서 완전독립성 모형

(complete independence model)인 $[X_1][X_2]\cdots[X_n]$ 모형 하에서 일반화 가능도비 통계량이다.

증명

식 (3.2)의 오른쪽 항에 있는 로그 함수에서의 분모부분에 해당하는 $p(x_{1c_1})\cdots p(x_{nc_n})$은 n차원 범주형 자료에 대한 완전독립성 모형인 $[X_1][X_2]\cdots[X_n]$ 모형 하에서 기대 칸확률이다. 그러므로 범주형 자료에서의 리둔던시 $R(X_1;\ldots;X_n)$는 완전독립성 모형인 $[X_1][X_2]\cdots[X_n]$ 모형 하에서 일반화 가능도비 검정통계량 G^2의 함수로 나타난다. ∎

따름정리 1.

4차원 범주형 자료에서 리둔던시 $R(X_1;X_2;X_3;X_4)$는 다음과 같이 표현된다.

$$R(X_1;X_2;X_3;X_4)=\sum\sum\sum\sum p(x_{1c_1},x_{2c_2},x_{3c_3},x_{4c_4}) \times log\left(\frac{p(x_{1c_1},x_{2c_2},x_{3c_3},x_{4c_4})}{p(x_{1c_1})p(x_{2c_2})p(x_{3c_3})p(x_{4c_4})}\right),$$

여기서 $p(x_{1c_1})p(x_{2c_2})p(x_{3c_3})p(x_{4c_4})$는 4차원 범주형 자료에서 완전독립성 모형인 $[X_1][X_2][X_3][X_4]$ 모형의 기대 칸확률이고, 이값은 $p_{c_1+++}p_{+c_2++}p_{++c_3+}p_{+++c_4}$이다. 따라서 4차원 범주형 자료에서 리둔던시 $R(X_1;X_2;X_3;X_4)$는 귀무가설 H_0가 4차원 범주형 자료에서 완전독립성 모형인 $[X_1][X_2][X_3][X_4]$ 모형 하에서 일반화 가능도비 검정통계량 G^2의 $1/2N$배이다. ∎

정리 2.

$n+m$차원 범주형 자료에서 n차원 확률벡터 $\vec{X}$와 m차원 확률벡터 $\vec{Y}$가 결합된 형태의 리둔던시 $R(\vec{X};\vec{Y})$는 식 (2.9)로부터 다음과 같이 정리된다.

$$\begin{aligned} R(\vec{X};\vec{Y}) &= \sum\cdots\sum p(\vec{x}_{c_n},\vec{y}_{d_m})log\left(\frac{p(\vec{x}_{c_n},\vec{y}_{d_m})}{p(\vec{x}_{c_n})p(\vec{y}_{d_m})}\right) \\ &= \frac{1}{2N}G^2, \end{aligned} \tag{3.3}$$

여기서 G^2는 $n+m$차원 범주형 자료에 대한 로그선형모형 중에서 n차원 확률벡터 $\vec{X}$와 m차원 확률벡터 $\vec{Y}$가 서로 독립적임을 나타내는 결합독립성 모형(joint independence model)인 $[\vec{X}][\vec{Y}]$ 모형 하에서 일반화 가능도비 통계량이다.

증명

식 (3.3)에 있는 로그 함수에서의 분모부분에 해당하는 $p(\vec{x}_{c_n})p(\vec{y}_{d_m})$는 n차원 범주형 확률벡터 $\vec{X}$와 m차원 범주형 확률벡터 $\vec{Y}$가 결합적으로 독립인 $[\vec{X}][\vec{Y}]$ 모형 하에서의 기대 칸확률이다. 그러므로 범주형 자료에서 n차원의 확률벡터와 m차원의 확

률벡터가 결합된 형태의 리둔던시 $R(\vec{X};\vec{Y})$는 결합독립성 모형인 귀무가설 H_0: $[\vec{X}][\vec{Y}]$ 하에서 일반화 가능도비 통계량 G^2의 함수로 정의된다. ■

따름정리 2.

5차원 범주형 자료에서 리둔던시 $R(X_1,X_2,X_3;Y_1,Y_2)$는 다음과 같이 표현된다.

$$R(X_1,X_2,X_3;Y_1,Y_2)=\sum\cdots\sum p(x_{1c_1},x_{2c_2},x_{3c_3},y_{1d_1},y_{2d_2}) \times log\left(\frac{p(x_{1c_1},x_{2c_2},x_{3c_3},y_{1d_1},y_{2d_2})}{p(x_{1c_1},x_{2c_2},x_{3c_3})p(y_{1d_1},y_{2d_2})}\right),$$

여기서 $p(x_{1c_1},x_{2c_2},x_{3c_3})p(y_{1d_1},y_{2d_2})$는 5차원 범주형 자료에서 확률벡터 $\vec{X}=(X_1,X_2,X_3)'$와 $\vec{Y}=(Y_1,Y_2)'$가 독립인 $[X_1X_2X_3][Y_1Y_2]$ 모형의 기대 칸확률이며, $p_{c_1c_2c_3++}$ $p_{+++d_1d_2}$으로 표현된다. 따라서 5차원 범주형 자료에서 리둔던시 $R(X_1,X_2,X_3;Y_1,Y_2)$는 귀무가설 H_0가 확률벡터 $\vec{X}$와 $\vec{Y}$의 결합독립 모형인 $[X_1X_2X_3][Y_1Y_2]$모형 하에서 일반화 가능도비 통계량 G^2의 $1/2N$배이다. ■

정리 2에서 논의한 확률벡터들이 결합된 형태의 리둔던시 $R(\vec{X};\vec{Y})$는 p개의 확률벡터인 $\vec{X}_1,\vec{X}_2,\ldots,\vec{X}_p$로 확장할 수 있으며, p개의 확률벡터들이 결합된 형태의 리둔던시 $R(\vec{X}_1;\vec{X}_2;\ldots;\vec{X}_p)$는 결합독립성 모형인 $[\vec{X}_1][\vec{X}_2]\cdots[\vec{X}_p]$ 모형 하에서 일반화 가능도비 검정통계량 G^2의 $1/2N$배로 표현하며 확장할 수 있다.

정리 3.

$n+m$차원 범주형 자료에서 m차원 확률벡터 $\vec{Y}$가 조건으로 주어졌을 때, 확률변수 $X_1,\ldots,X_n$에 대한 조건부 리둔던시 $R(X_1;\ldots;X_n|\vec{Y})$는 식 (2.10)으로부터 다음과 같이 정리한다.

$$\begin{aligned}R(X_1;\ldots;X_n|\vec{Y}) &= \sum\cdots\sum p(x_{1c_1},\ldots,x_{nc_n},\vec{y}_{d_m}) \\ &\quad\times log\left(\frac{p(x_{1c_1},\ldots,x_{nc_n}|\vec{y}_{d_m})}{p(x_{1c_1}|\vec{y}_{d_m})\cdots p(x_{nc_n}|\vec{y}_{d_m})}\right) \\ &= \frac{1}{2N}G^2,\end{aligned} \tag{3.4}$$

여기서 G^2는 범주형 자료에 대한 로그선형모형 중에서 $n+m$차원 범주형 자료에서 m차원 확률벡터 $\vec{Y}$가 조건으로 주어졌을 때, 확률변수 $X_1,\ldots,X_n$이 상호 독립적임을 나타내는 조건부 독립성 모형(conditional independence model)인 $[X_1\vec{Y}][X_2\vec{Y}]\cdots[X_n\vec{Y}]$ 모형 하에서 일반화 가능도비 통계량이다.

증명

식 (3.4)에서의 $p(x_{1c_1}|\vec{y}_{d_m})\cdots p(x_{nc_n}|\vec{y}_{d_m})$는 확률벡터 $\vec{Y}$가 조건으로 주어졌을 때, 범

주형 확률변수 $X_1,...,X_n$이 독립인 $[X_1\overrightarrow{Y}][X_2\overrightarrow{Y}]\cdots[X_n\overrightarrow{Y}]$ 모형 하에서의 기대 칸확률이다. 그러므로 $n+m$차원 범주형 자료에서 m차원 확률벡터 $\overrightarrow{Y}$가 조건으로 주어졌을 때, 확률변수 $X_1,...,X_n$에 대한 조건부 리둔던시 $R(X_1;...;X_n|\overrightarrow{Y})$는 조건부 독립성 모형의 귀무가설 H_0: $[X_1\overrightarrow{Y}][X_2\overrightarrow{Y}]\cdots[X_n\overrightarrow{Y}]$ 모형 하에서 일반화 가능도비 통계량 G^2의 함수로 정의된다. ∎

따름정리 3.

5차원 범주형 자료에서 조건부 리둔던시 $R(X_1;X_2;X_3|Y_1,Y_2)$는 다음과 같이 정리된다.

$$R(X_1;X_2;X_3|Y_1,Y_2)=\sum\sum\sum\sum\sum p(x_{1c_1},x_{2c_2},x_{3c_3},y_{1d_1},y_{2d_2})\times log\left(\frac{p(x_{1c_1},x_{2c_2},x_{3c_3}|y_{1d_1},y_{2d_2})}{p(x_{1c_1}|y_{1d_1},y_{2d_2})p(x_{2c_2}|y_{1d_1},y_{2d_2})p(x_{3c_3}|y_{1d_1},y_{2d_2})}\right),$$

여기서 $p(x_{1c_1}|y_{1d_1},y_{2d_2})p(x_{2c_2}|y_{1d_1},y_{2d_2})p(x_{3c_3}|y_{1d_1},y_{2d_2})$는 5차원 범주형 자료에서 확률벡터 $\overrightarrow{Y}=(Y_1,Y_2)'$가 주어졌을 때, 확률변수 X_1,X_2,X_3가 조건부 독립인 $[X_1\overrightarrow{Y}][X_2\overrightarrow{Y}][X_3\overrightarrow{Y}]$ 모형의 기대 칸확률이고, $p_{c_1++|d_1d_2}p_{+c_2+|d_1d_2}p_{++c_3|d_1d_2}$으로 표현된다. 그러므로 조건부 리둔던시 $R(X_1;X_2;X_3|Y_1,Y_2)$는 귀무가설 H_0가 조건부 독립성 모형인 $[X_1Y_1Y_2][X_2Y_1Y_2][X_3Y_1Y_2]$ 모형 하에서 일반화 가능도비 검정통계량 G^2의 $1/2N$배이다. ∎

정리 3에서 논의한 조건부 리둔던시 $R(X_1;...;X_n|\overrightarrow{Y})$에서 확률변수 $X_1,...,X_n$를 p개의 확률벡터 $\overrightarrow{X}_1,...,\overrightarrow{X}_p$로 대체하여 $R(\overrightarrow{X}_1;\overrightarrow{X}_2;...;\overrightarrow{X}_p|\overrightarrow{Y})$에 대하여 확장할 수 있다. 예를 들어 $\overrightarrow{X}_1=(X_1,X_2)'$, $\overrightarrow{X}_2=(X_3)$, 그리고 $\overrightarrow{Y}=(Y_1,Y_2)'$인 경우, $R(\overrightarrow{X}_1;\overrightarrow{X}_2|\overrightarrow{Y})$는 다음과 같이 정리된다.

$$R(\overrightarrow{X}_1;\overrightarrow{X}_2|\overrightarrow{Y})=\sum_{c_1}\sum_{c_2}\sum_{c_3}\sum_{d_1}\sum_{d_2}p(x_{1c_1},x_{2c_2},x_{3c_3},y_{1d_1},y_{2d_2})\times log\left(\frac{p(x_{1c_1},x_{2c_2},x_{3c_3}|y_{1d_1},y_{2d_2})}{p(x_{1c_1}x_{2c_2}|y_{1d_1},y_{2d_2})p(x_{3c_3}|y_{1d_1},y_{2d_2})}\right).$$

이 경우에 대응하는 로그선형모형은 $[\overrightarrow{X_1Y}][\overrightarrow{X_2Y}]=[X_1X_2|Y_1Y_2][X_3|Y_1Y_2]$이 된다. 따라서 조건부 리둔던시 $R(X_1;...;X_n|\overrightarrow{Y})$는 조건부 독립성 모형 $[X_1\overrightarrow{Y}]\cdots[X_n\overrightarrow{Y}]$ 모형 하에서 일반화 가능도비 통계량 G^2의 함수로 표현되듯이, 확률벡터의 조건부 리둔던시 $R(\overrightarrow{X}_1;\overrightarrow{X}_2;...;\overrightarrow{X}_p|\overrightarrow{Y})$는 $[\overrightarrow{X}_1\overrightarrow{Y}][\overrightarrow{X}_2\overrightarrow{Y}]\cdots[\overrightarrow{X}_p\overrightarrow{Y}]$ 모형 하에서 일반화 가능도비 통계량 G^2의 함수로 정의된다.

정리 1부터 정리 3까지 얻은 결과를 바탕으로, 다차원 범주형 자료의 적합성을 분석하는 통계적 방법으로 사용하는 여러 로그선형모형 중에서 완전독립성 모형, 결합독립성 모형, 그리고 조건부 독립성 모형을 검정하는 일반화 가능도비 통계량은 리둔

던시와의 함수 관계를 갖고 있음을 발견하였으며, 여러 종류의 독립성 모형에 대한 리둔던시는 대응하는 독립성 모형의 일반화 가능도비 통계량을 두배의 표본 크기로 나눈 값으로 정의되고 있음을 유도하였다.

4. 예제

로그선형모형 중에는 세 종류의 독립성 모형(완전독립성 모형, 결합독립성 모형, 그리고 조건부 독립성 모형)이 존재한다. 이와 같은 세 종류의 독립성 모형의 적합도 검정통계량인 일반화 가능도비 통계량 G^2와 범주형 자료에 대한 리둔던시는 함수 관계를 갖고 있다는 것을 3절에서 논의했기 때문에 본절에서는 이러한 모형에 적합한 자료를 생성하여 표본의 크기에 따라 리둔던시와 일반화 가능도비 통계량의 값의 변화를 살펴보고자 한다.

완전독립성 모형인 경우는 이미 2차원 범주형 자료에서 많은 사람들이 논의했기 때문에 생략하고, 결합독립성 모형인 [12][3] 모형에 적합한 삼차원인 $2\times2\times3$ 범주형 자료를 고려한다. 임의의 $\{p_{ij+}\}$와 $\{p_{++k}\}$를 생성하고, $\{p_{ijk}=p_{ij+}\times p_{++k}\}$를 작성한 칸 확률표는 <표 4.1>과 같다.

<표 4.1> [12][3] 모형에 적합한 확률

[12][3]모형	$k=1$		$k=2$		$k=3$	
	$j=1$	$j=2$	$j=1$	$j=2$	$j=1$	$j=2$
$i=1$	0.03	0.10	0.08	0.15	0.04	0.24
$i=2$	0.04	0.01	0.10	0.04	0.10	0.07

<표 4.2> [12][3] 모형에서 리둔던시와 일반화 가능도비 통계량

표본크기 N	리둔던시 $R(X_1,X_2;X_3)$	일반화 가능도비 통계량 $G^2\sim$[12][3]모형	p-값
100	0.0233	4.668606	0.5870
200		9.337212	0.1555
500		23.343031**	0.0007
1,000		46.686062**	0.0001

** : 대응하는 p-값이 유의수준 0.01보다 작아 유의한 경우.

표본크기가 N인 경우에 각 칸에 해당하는 빈도수는 $N\times p_{ijk}$로 생성하고, 다양한 표본크기에 따라 표본을 생성하여, 각각의 경우에 일반화 가능도비 통계량 G^2값과 리둔던시 $R(X_1,X_2;X_3)$값을 구한 결과는 <표 4.2>와 같다, 이것은 정리 2의 $R(\vec{X};\vec{Y})$에서 $\vec{X}=(X_1,X_2)'$이고 $\vec{Y}=(X_3)$인 경우이다.

다음으로는 조건부 독립성 모형인 [13][23] 모형에 적합한 3차원인 $2\times2\times3$ 범주형 자료를 고려해보자. $\{p_{i+k}\}$와 $\{p_{+jk}\}$를 생성하고, $\{p_{ijk}=p_{i+k}\times p_{+jk}/p_{++k}\}$를 작성한

칸 확률표는 <표 4.3>과 같으며, 여러 표본크기의 경우에 일반화 가능도비 통계량과 조건부 리둔던시의 값은 <표 4.4>와 같다.

<표 4.3> [13][23] 모형에 적합한 확률

[13][23]모형	$k=1$		$k=2$		$k=3$	
	$j=1$	$j=2$	$j=1$	$j=2$	$j=1$	$j=2$
$i=1$	0.22	0.07	0.1	0.13	0.02	0.1
$i=2$	0.15	0.02	0.02	0.09	0.03	0.05

<표 4.4> [13][23] 모형에서 조건부 리둔던시와 일반화 가능도비 통계량

표본크기 N	조건부 리둔던시 $R(X_1;X_2\|X_3)$	일반화 가능도비 통계량 $G^2\sim$[13][23]모형	p-값
100	0.0221	4.427733	0.2188
200		8.855466*	0.0313
500		22.138665**	0.0001
1000		44.277329**	0.0001

* : 대응하는 p-값이 유의수준 0.05보다 작아 유의한 경우,
** : 대응하는 p-값이 유의수준 0.01보다 작아 유의한 경우.

완전독립성 모형뿐만 아니라, 결합독립성 모형 그리고 조건부 독립성 모형을 따르는 분할표 자료에서 리둔던시와 일반화 가능도비 통계량의 값을 구한 <표 4.2>와 <표 4.4>를 살펴보면, 리둔던시는 표본크기가 변함에 따라 영향이 전혀 없는 값을 유지하지만, 일반화 가능도비 통계량 값은 표본크기가 증가함에 따라 비례적으로 증가하고 대응하는 p-값은 작아진다. 따라서 적합도 검정통계량인 일반화 가능도비 통계량은 표본크기에 매우 민감하게 반응하지만, 리둔던시는 표본크기에 영향을 받지 않는다는 것을 확인할 수 있었다. 그러므로 리둔던시는 표본크기에 상관없이 로그선형 모형을 따르는 확률에 의존한다. 즉 리둔던시는 표본크기에 영향을 받지 않고 자료에 적합한 모형으로 설명되는 확률 자체의 성격과 특성만을 고려하여, 범주형 변수들 사이의 독립적인 관계만을 설명하는 통계량으로 설명할 수 있다.

5. 결론

범주형 변수들의 관계를 설명하기 위해서 본 논문에서는 리둔던시를 범주형 자료에 적용하여 살펴보았다. 확률변수에 대한 특별한 가정이 필요하지 않는 n개의 범주형 확률변수 $X_1,\ldots,X_n$에 대한 리둔던시 $R(X_1;\ldots;X_n)$와 p개의 확률벡터가 결합된 형태의 리둔던시 $R(\vec{X}_1;\ldots;\vec{X}_p)$, 그리고 확률벡터 $\vec{Y}$가 조건으로 주어졌을 때, 조건부 리둔던시 $R(X_1;\ldots;X_n|\vec{Y})$ (또는 $R(\vec{X}_1;\ldots;\vec{X}_p|\vec{Y})$)는 각각 범주형 자료를 설명하는 로그선형모형들 중에서 완전독립성 모형과 결합독립성 모형, 그리고 조건부 독립성 모형 하에서

의 일반화 가능도비 검정통계량 G^2의 함수로 나타낼 수 있음을 발견하였다.

- n개의 범주형 변수에 대한 리둔던시 $R(X_1;\ldots;X_n)$는 로그선형모형 중에서 완전 독립성 모형(complete independence model)인 $[X_1]\cdots[X_n]$ 모형 하에서 일반화 가능도비 검정통계량 G^2의 $1/2N$배로 정의된다.

- 다차원 범주형 자료에서 p개의 확률벡터들이 결합된 형태의 리둔던시 $R(\overrightarrow{X_1};\ldots;\overrightarrow{X_p})$는 결합독립성 모형(joint independence model)인 $[\overrightarrow{X_1}]\cdots[\overrightarrow{X_p}]$ 모형 하에서 일반화 가능도비 검정통계량 G^2의 $1/2N$배로 정의된다.

- $n+m$차원 범주형 자료에서 m차원 확률벡터 $\overrightarrow{Y}$가 주어졌을 때, n개의 확률변수 $X_1,\ldots,X_n$에 대한 조건부 리둔던시 $R(X_1;\ldots;X_n|\overrightarrow{Y})$ (또는 p개의 확률벡터들에 대한 $R(\overrightarrow{X}_1;\ldots;\overrightarrow{X}_p|\overrightarrow{Y})$)는 조건부 독립성 모형(conditional independence model)인 $[X_1\overrightarrow{Y}]\cdots[X_n\overrightarrow{Y}]$ 모형 (또는 조건부 결합 독립성 모형 $[\overrightarrow{X}_1\overrightarrow{Y}]\cdots[\overrightarrow{X}_p\overrightarrow{Y}]$ 모형)하에서 일반화 가능도비 검정 통계량 G^2의 $1/2N$배로 정의된다.

위에서 언급한 세 가지 요약에서, 다차원 범주형 자료를 설명하는 로그선형모형들 중에서 완전독립성 모형과 결합독립성 모형 그리고 조건부 독립성 모형의 적합성은 리둔던시를 통해 설명되고 있음을 발견하였다. 이러한 로그선형모형들은 범주형 변수들에 대한 세 종류의 독립적인 관계를 설명하는 세 종류의 모형들이며, 이에 대응하는 로그선형모형들의 기대칸값(expected cell frequency)은 직접해(direct solution)가 존재하는 모형들이다 (직접해에 관한 자세한 설명은 Bishop, Fienberg와 Holland (1975)를 참조).

범주형 자료의 적합도 검정통계량인 일반화 가능도비 통계량은 표본크기에 매우 민감하게 반응하지만 리둔던시는 표본크기에 영향을 받지 않는다. 그러므로 범주형 자료에서의 리둔던시는 자료의 확률 자체의 성격과 특성만을 고려하여 범주형 자료에 적합한 로그선형모형 중 여러 종류의 독립성 모형의 적합성을 설명하는 통계량이라고 판단할 수 있겠다.

이변량 확률변수에 대하여는 이미 식 (3.1)에서 설명하였듯이 상호정보와 리둔던시는 동일한 개념이다. 고차원 확률변수의 리둔던시를 상호정보로 표현한 식 (2.9)를 이용하면, $I(X_1;\cdots;X_n)$, $I(\overrightarrow{X}_1;\cdots;\overrightarrow{X}_p)$, 그리고 $I(X_1;\cdots;X_n|Y)$, $I(\overrightarrow{X}_1;\cdots;\overrightarrow{X}_p|\overrightarrow{Y})$를 리둔던시로 표현할 수 있다. 그리고 3절에서 유도한 정리를 이용하여 상호정보도 범주형 자료에서의 일반화가능도비 검정통계량의 함수로 나타낼 수 있다.

참고문헌

[1] Abramson, N. (1963). *Information Theory and Coding*, Mcgraw Hill, New York.

[2] Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis*, Cambridge, Massachusetts: MIT Press.

[3] Brillinger, R. (2004). Some data analyses using mutual information, *Brazilian Journal of Probability and Statistics*, Vol. 18, 163-183.

[4] Brillinger, R and Guha A. (2006). Mutual Information in the Frequency Domain, *Journal of Statistical Planning and Inference,* To appear.

[5] Cover, T. and Thomas, J. (1991). *Elements of Information Theory*, John Wiley and Sons, New York.

[6] DeGroot, M.H. (1962). Uncertainty, information and sequential experiments. *Annals of Mathematical Statistics*, Vol. 33, 404-419.

[7] Fraser, A. and Swinney, H. (1986). Independent coordinates for strange attractors from mutual information. *Physical Review,* Vol. 33(2), 1134-1140.

[8] Gallager, R.G. (1968). *Information Theory and Reliable Communication*, John Wiley, New York.

[9] Gelfand, I.M. and Yaglom, A.M. (1959). Calculation of the amount of information about a random function contained in another such function. *American Mathematical Society*, Translations, Ser.

[10] Palus, M. (1993). Identifying and quantifying chaos by using information theoretic functions in time series prediction: Forecasting the Future and Understanding the Past. *SantaFe Institute Studies in the Sciences of Complexity*, Vol. 15, 387-413.

[11] Palus, M. and Pivka, D. (1995). Estimating predictability: Redundancy and surrogate data method. *Neural Network World*, Vol. 4, 537-552.

[12] Prichard, D. and Theiler, J. (1995). Generating surrogate data for time series with several simultaneously measured variables. *Physical Review*. Vol. 73, 951-954.

[13] Shannon, C.E, (1948). A mathematical theory of communication. *The Bell System Technical Journal*, Vol. 27, 379-423.

[14] Wienholt, W. and Sendhoff, B. (1996). How to determine the redundancy of noisy chaotic time series. *International Journal of Bifurcation and Chaos*, Vol. 6, 101-117.

[Received March 2006, Accepted May 2006]