Journal of Korean
Data & Information Science Society
2003, Vol. 14, No.2 pp. 303~311

A Nonparametric Goodness-of-Fit Test for Sparse Multinomial Data[1)]

Jangsun Baek[2)]

Abstract

We consider the problem of testing cell probabilities in sparse multinomial data. Aerts, et al.(2000) presented $T_1 = \sum_{i=1}^{k}(\hat{p}_i - p_i)^2$ as a test statistic with the local polynomial estimator $\hat{p}_i$, and showed its asymptotic distribution. When there are cell probabilities with relatively much different sizes, the same contribution of the difference between the estimator and the hypothetical probability at each cell in their test statistic would not be proper to measure the total goodness-of-fit. We consider a Pearson type of goodness-of-fit test statistic, $T = \sum_{i=1}^{k}(\hat{p}_i - p_i)^2 / p_i$ instead, and show it follows an asymptotic normal distribution.

Keywords : 국소다항추정량, 적합도 검정, 희박다항 자료

1. 서론

다항분포자료를 관측할 때 우리는 종종 여러개의 칸(cell)에서 하나도 없거나 1개 혹은 2개와 같은 아주 적은 갯수의 관측값들을 발견하게 된다. 이와 같은 자료를 희박다항자료(sparse multinomial data)라 한다. 좀 더 정확하게 말하여 i번째 칸의 확률이 p_i인 총 k개의 칸을 가진 다항분포로부터 $(N_1, N_2, \cdots, N_k)^T$의 칸도수(cell frequency)가 관측되었다고 하자. 이때 $\sum_{j=1}^{k} N_j = n$은 총도수를 나타낸다. 따라서 희박다항분포자료는 총도수 n이 칸의 총갯수 k에 비하여 상대적으로 매우 작을 때, 즉 n/k이 작을 때 발생한다.

1) 이 논문은 2000년도 전남대학교 학술연구비 지원에 의하여 연구되었음.
2) 광주광역시 북구 용봉동 300 전남대학교 수학통계학부 부교수
E-mail: jbaek@chonnam.ac.kr

만약 다항분포자료의 칸들이 순서대로 배열되어 있다면 (예를 들어 연속형 자료가 그 크기 순서대로 몇 개의 구간으로 그룹화되어 각 그룹내의 도수로 관측된 경우) 칸 도수들은 인접칸들에 걸쳐 평활함으로써 칸 확률을 추정할 수 있다. 일반적으로 다항분포에 대한 커널을 이용한 이와 같은 비모수적 평활 방법들이 Burman(1987a), Aerts, et al.(1997a), Baek(1998)등에 의해 개발 되었고, 그 추정량에 대한 이론적인 희박점근 일치성이 증명되었다.

주어진 다항자료의 분포 $\boldsymbol{p}=(p_1,\cdots,p_k)$가 특정분포 $\boldsymbol{p_0}=(p_{10},\cdots,p_{k0})$를 따르는지 검정하고자 할 때 우리는 귀무가설 $H_0:\ \boldsymbol{p}=\boldsymbol{p_0}$ 에 대하여 일반적으로 피어슨의 $X^2=\sum_{i=1}^{k}(N_i-np_{i0})^2/np_{i0}$ 검정통계량이나 우도비 검정통계량 $G^2=2\sum_{i=1}^{k}N_i\log(N_i/np_{i0})$ 등을 적용하게 된다. 이들 통계량들은 귀무가설하에서 $n\to\infty$이고 $inf_i\ np_i\to\infty$이면 $X^2(k-1)$ 분포를 따르게 된다. 그러나 희박다항분포에서는 $n\to\infty$이고 $k\to\infty$이므로 p_i가 아주 작아진다. 따라서 $X^2(k-1)$의 점근분포로 수렴하기 위한 $inf_i\ np_i\to\infty$조건을 충족시키지 못한다. Morris(1975)는 X^2와 G^2가 $n\to\infty$, $k\to\infty$ 일 때 점근적으로 정규분포에 수렴함을 보였다. 따라서 희박다항자료의 적합도검정에서 X^2와 G^2를 그것의 점근적 정규분포와 함께 사용할 수는 있으나 그 수렴속도가 너무 느리다. Lawal and Upton(1980)은 희박자료에 대해 X^2와 G^2를 로그정규근사하는 방법을 제안하였다.

다항분포확률에 대한 국소다항커널추정량을 $\hat{p}_i$ 라 할 때 Aerts, et al.(2000) 에서는 $T_1=\sum_{i=1}^{k}(\hat{p}_i-p_i)^2$ 을 적합 검정통계량으로 하여 그것의 귀무가설하에서의 정규성을 증명하였다. 또한 Liero(2001)은 국소대립가설 하에서 T_1의 검정력의 변화를 연구하였다. Simonoff(1985)는 최대벌점우도(maximum penalized likelihood) 추정량을 사용하여 그것을 표준화하여 구성한 검정통계량을 제시하였다. 그러나 그의 검정통계량은 귀무가설하의 분포를 이론적으로 구하지 못하여 모의 실험을 통한 임계치를 구한 후 검정을 실시할 수 밖에 없다.

Aerts, et al.(2000)의 검정통계량 형태는 각 칸의 추정확률과 귀무가설하의 확률간의 차이를 똑같은 가중치를 가지고 단순히 합하여 구한 것이므로 만약 각 칸 확률들 간에 차이가 크면 전체적인 적합도를 측정하는데 어려움이 발생한다. 반면 Simonoff(1985)의 검정통계량은 각 칸에서의 추정확률과 귀무가설하에서의 확률간의 차이를 각 칸 확률로 나누어 전체를 합한 형태이므로 더욱 바람직한 형태를 가진다고 할 수 있다. 본 연구에서는 이러한 두가지 검정통계량의 단점을 극복할 수 있는 검정통계량으로서 피어슨 통계량 형태의 $T=\sum_{i=1}^{k}(\hat{p}_i-p_i)^2/p_i$ 를 제시하고 그것의 점근분포를 유도하였다.

2. 희박다항자료에 대한 비모수 적합도 검정통계량

순서화된 범주를 갖는 범주형 자료를 분석하고자 할 때, 이러한 다항자료의 첫번째 관심은 칸 확률 $\boldsymbol{p}=\{p_j\}$를 추정하는 것이다. 우리가 추정하려는 칸 확률 p_i는 [0, 1]에서 연속인 잠재밀도함수 $f(\cdot)$에 의해 결정된다고 가정하자. 즉, $p_i = \int_{(i-1)/k}^{i/k} f(u)\ du$, $i=1,\cdots,k$이며 $f(\cdot)$는 $(p+1)$차 미분가능하다고 가정한다. 칸 확률은 일반적으로 도수 추정량 $\overline{p_i}=N_i/n$ 을 사용하여 왔다. 그러나 이 추정량은 각 칸에서의 관측도수가 많을 때 유용하며 희박한 다항자료에서는 희박일치성을 만족하지 못한다.

순서화된 다항분포자료에 대한 이산형 확률의 비모수적 추정은 도수추정량 $\overline{p_i}=N_i/n$들을 평활함으로서 시행된다. $x_l=(l-1/2)/k$, $l=1,\cdots,k$을 등간격을 갖는 각 구간의 중심점이라하면, Aerts, et al.(1997)과 Baek(1998)에서 사용한 평활 추정량은 $(x_l, \overline{p_l})$, $l=1,\cdots,k$ 를 평활하여 구성된 것이다. 즉 칸 확률 p_i에 대한 국소다항 추정량은 평활모수 h 를 갖는

$$\widehat{p}_i = e_1^T (X_i^T W_i\ X_i)^{-1}\ X_i^T\ W_i\ \overline{P}$$

의 형태를 갖는다. 이때 $\overline{P}=(\overline{p_i},\cdots,\overline{p_k})^T$이고 e_1^T는 $(p+1)$차원의 $(1,0,\cdots,0)$벡터이며 W_i와 X_i는 아래와 같다.

$$W_i = diag\left(\frac{1}{h}K\left(\frac{x_1-x_i}{h}\right),\cdots,\frac{1}{h}K\left(\frac{x_k-x_i}{h}\right)\right),$$

$$X_i=\begin{pmatrix} 1 & x_1-x_i & \cdots & (x_1-x_i)^p \\ \vdots & & & \vdots \\ 1 & x_k-x_i & \cdots & (x_k-x_i)^p \end{pmatrix}.$$

추정량 $\widehat{p}_i$는 $n\to\infty$, $k\to\infty$일 때 진실된 확률 p_i에 수렴하는 희박점근 일치성을 갖는다.

주어진 희박다항분포자료가 특정의 분포 $\boldsymbol{p_0}=(p_{10},\cdots,p_{k0})$를 따르는지 검정하고자 할 때 $H_0: \boldsymbol{p}=\boldsymbol{p_0}$에 대하여 Aerts et al.(2000)은 $T_1=\sum_{i=1}^{k}(\widehat{p}_i-p_i)^2$을 검정통계량으로 제안하였다. 그런데 이것은 단순히 추정확률과 귀무가설간의 차이의 제곱을 합한 것이므로 만약 귀무가설분포의 각 칸 확률들이 크기에 있어 심하게 차이가 나는 경우 높은 확률에 있어서의 차이나 낮은 확률에 있어서의 차이나 전체적인 적합도에 대한 공헌은 동일하게 된다. 그러나 전체적인 적합도를 구성하는 요소로서 실제로 낮은 확률에서의 차이는 훨씬 민감하며 또한 높은 확률에서의 차이는 덜 민감하다. 따라서 일반적인 피어슨 X^2 적합도 검정통계량처럼 각 칸에서 추정확률과 귀무확률간의 차이를 각 귀무확률로 나누어 표준화하는 것이 합리적이다.

Simonoff(1985)는 그의 최대벌점우도추정량 $\widetilde{p}_i$에 대하여 $z_i=(\widetilde{p}_i-p_i)/p_i$로서 표

준화한 다음 적합도 검정 통계량으로서 $T_2 = \sum_{i=1}^{k}\left(\frac{z_i - \mu_0(z_i)}{\sigma_0(z_i)}\right)^2$ 을 제안하였다. 이 때 $\mu_0(z_i)$는 z_i의 귀무가설하에서의 평균이며, $\sigma_0(z_i)$는 z_i의 귀무가설하에서의 표준편차이다. T_2를 살펴보면 귀무가설하에서 T_2는 서로 종속적인 k개의 점근적 $X^2(1)$ 확률변수들의 합으로 된 분포를 갖게 된다. 그러나 Simonoff(1985)는 T_2의 점근분포를 유도하지 못하고 대신 모의 실험에 의해 T_2의 분포를 추정하여 그로부터 검정에 필요한 임계치를 구한다.

본 연구에서는 T_1과 T_2의 단점을 극복할 수 있도록 국소다항추정량 $\hat{p}_i$를 이용하여 각 칸에서의 표준화된 차이를 합하여 전체적인 적합도를 측정할 수 있는 검정통계량을 제시하고자 한다. 즉 본 연구에서 제안하는 검정통계량은 $T = \sum_{i=1}^{k}(\hat{p}_i - p_i)^2 / p_i$ 이다. 실제로 검정을 수행하기 위해서는 T의 귀무가설하에서의 점근분포를 유도해야 한다. T의 점근분포를 유도하기 위해서 본 연구에서는 Burman(1987b)의 $\overline{p_i}$ 를 이용한 피어슨 적합통계량의 점근분포 결과를 이용하였다.

T 의 점근분포를 유도하기 위한 **[정리]**의 증명에 필요한 사항인 Aerts et al.(1977b) 의 Example 2와 Burman (1987b) 의 Corollary 3.3(a)를 **[보조정리 1]**과 **[보조정리 2]**로서 각각 아래와 같이 정리한다.

[보조정리 1] 칸 확률을 결정하는 잠재밀도함수(latent density function) $f(u)$ 가 $0 < inf\{f(u): 0 \le u \le 1\} \le sup\{f(u): 0 \le u \le 1\} < \infty$를만족하고, $\frac{\ln k}{\ln \ln n} \to \infty$, $\frac{k \ln k}{n} \to 0$ 이면, $\sup_{1 \le i \le k}\left|\frac{\overline{p_i}}{p_i} - 1\right| \overset{a.s.}{=} O\left(\sqrt{\frac{k \ln k}{n}}\right)$, $n \to \infty$ 이다.

[보조정리 2] $\beta_n = 2k + (1/n)\sum_{i=1}^{k}(1/p_i)$ 이라 하자. 만약 $\min_i\{np_i\} = \varepsilon_n > 0$ 이며 $k\varepsilon_n^3 \to \infty$ 이면, 칸도수 추정량 $\overline{\boldsymbol{p}}$ 에 대하여, $n \to \infty$, $k \to \infty$ 에 따라 $\frac{n}{\sqrt{\beta_n}}\left\{\sum_{i=1}^{k}\frac{(\overline{p_i} - p_i)^2}{p_i} - \left(\frac{k-1}{n}\right)\right\} \overset{d}{\to} N(0, 1)$ 이다.

본 논문의 **[정리]**에 필요한 조건들을 정리하면 다음과 같다:

(C1) $K(\cdot)$는 유한 경계 $[-L, L]$을 갖는 평균이 0 이고 Lipschitz 연속인 밀도함수이다.

(C2) $f^{(p+1)}(\cdot)$이 [0, 1]에서 연속이다.

(C3) $h \to 0$, $hk \to \infty$ 이다.

[정리] $f(u)$가 $0 < inf\{f(u) : 0 \le u \le 1\} \le sup\{f(u) : 0 \le u \le 1\} < \infty$ 를

만족하고, $n \to \infty$ 에 따라 조건들 (C1), (C2), (C3) 와 $\frac{\ln k}{\ln \ln n} \to \infty$, $\frac{k \ln k}{n} \to 0$, 평활모수 h 와 $\beta_n = 2k + (1/n) \sum_{i=1}^{k} (1/p_i)$ 에 대하여 $\frac{h^{p+1} k^2 \ln k}{\sqrt{\beta_n}} \to 0$, $\frac{n}{\sqrt{\beta_n}} h^{2(p+1)} \to 0$ 을 만족하면,

$$\frac{n}{\sqrt{\beta_n}} \left\{ \sum_{i=1}^{k} \frac{(\hat{p}_i - p_i)^2}{p_i} - \left(\frac{k-1}{n} \right) \right\} \xrightarrow{d} N(0, 1) \text{ 이다.}$$

[증명] 우선 검정통계량을 다음과 같이 분해하자.

$$\begin{aligned} \sum \frac{(\hat{p}_i - p_i)^2}{p_i} &= \sum \frac{(\hat{p}_i - \overline{p}_i + \overline{p}_i - p_i)^2}{p_i} \\ &= \sum \frac{1}{p_i} \{ (\hat{p}_i - \overline{p}_i)^2 + 2(\hat{p}_i - \overline{p}_i)(\overline{p}_i - p_i) + (\overline{p}_i - p_i)^2 \} \\ &= \sum \frac{(\overline{p}_i - p_i)^2}{p_i} + \sum \frac{(\hat{p}_i - \overline{p}_i)^2}{p_i} + 2 \sum \frac{(\hat{p}_i - \overline{p}_i)(\overline{p}_i - p_i)}{p_i} \\ &\equiv A + B + C. \end{aligned}$$

증명에 필요한 표기 및 함수들은 Aerts et al.(1977a)를 따른다.

$s_{k,l}(x) = \frac{1}{kh} \sum_{j=1}^{k} \left(\frac{x_j - x}{h} \right)^l K\left(\frac{x_j - x}{h} \right)$ 라 하면, $N_{i,p}$ 는 (r, s) 번째 요소로서 $s_{k, r+s-2}(x_i)$ 를 갖는 $(p+1) \times (p+1)$ 행렬이다. 그리고 $M_{i,p}(u)$ 는 $N_{i,p}$ 에서 첫 번째 열만 $(1, u, \ldots, u^p)^T$ 로 대체된 행렬이다. $L_{i,p}(u) = \{ |M_{i,p}(u)| / |N_{i,p}| \} K(u)$ 이라 하면, $\overline{p}_i = \frac{N_i}{n}$ 이며, 국소다항 추정량은 $\hat{p}_i = \frac{1}{kh} \sum_{j=1}^{k} L_{i,p}\left(\frac{x_j - x_i}{h} \right) \overline{p}_j$ 이다. 그러면 $\hat{p}_i - \overline{p}_i$ 의 평균 및 분산은 다음과 같이 전개할 수 있다.

$$\begin{aligned} E(\hat{p}_i - \overline{p}_i) &= E(\hat{p}_i) - p_i \\ &= \frac{f^{(p+1)}(x_i)}{(p+1)!} \frac{1}{kh} \sum_{j=1}^{k} \left(\frac{x_j - x_i}{h} \right)^{p+1} L_{i,p}\left(\frac{x_j - x_i}{h} \right) \frac{h^{p+1}}{k} + o\left(\frac{h^{p+1}}{k} \right) \\ &\sim \frac{f^{(p+1)}(x_i)}{(p+1)!} \left\{ \int_{-L}^{L} v^{p+1} L_p(v)\, dv \right\} \frac{h^{p+1}}{k} + o\left(\frac{h^{p+1}}{k} \right). \end{aligned}$$

이때 마지막 줄은 $\mu_l = \int_{-L}^{L} v^l K(v) dv$ 에 대하여 N_p 가 (r, s) 번째 요소로서 μ_{r+s-2} 를 갖는 $(p+1) \times (p+1)$ 행렬이며, $M_p(u)$ 는 N_p 에서 첫 번째 열만 $(1, u, \ldots, u^p)^T$ 로 대체된 행렬이라 할 때, $L_{i,p}(u) \to L_p(u) = \{ |M_p(u)| / |N_p| \} K(u)$, $n \to \infty$ 이기 때문이다.

$\widehat{p}_i - \overline{p}_i$ 의 분산은 $Var(\widehat{p}_i - \overline{p}_i) = Var(\widehat{p}_i) + Var(\overline{p}_i) - 2Cov(\widehat{p}_i, \overline{p}_i)$ 이며, 우선 $Var(\widehat{p}_i)$ 를 구하면 아래와 같다.

$$\begin{aligned} Var(\widehat{p}_i) &= \frac{f(x_i)}{nk^2h}\frac{1}{kh}\sum_{j=1}^{k} L_{i,p}^2\left(\frac{x_j - x_i}{h}\right) + o\left(\frac{1}{nk^2h}\right) \\ &\sim \frac{f(x_i)}{nk^2h}\int_{-L}^{L} L_p^2(v)\,dv + o\left(\frac{1}{nk^2h}\right). \end{aligned}$$

$p_i = \frac{1}{k}f(x_i) + O\left(\frac{1}{k^3}\right)$ 이므로 $Var(\overline{p}_i) = \frac{p_i(1-p_i)}{n} = \frac{f(x_i)}{nk} + O\left(\frac{1}{nk^2}\right)$ 이며,

$$\begin{aligned} E(\widehat{p}_i\overline{p}_i) &= E\left[\left\{\frac{1}{kh}\sum_{j=1}^{k} L_{i,p}\left(\frac{x_j - x_i}{h}\right)\overline{p}_j\right\}\overline{p}_i\right] \\ &= \frac{1}{kh}\sum_{j=1}^{k} L_{i,p}\left(\frac{x_j - x_i}{h}\right)E(\overline{p}_j\overline{p}_i) \\ &= \frac{1}{kh}\sum_{j=1}^{k} L_{i,p}\left(\frac{x_j - x_i}{h}\right)E\left(\frac{N_jN_i}{n^2}\right) \\ &= \frac{1}{kh}\sum_{j=1}^{k} L_{i,p}\left(\frac{x_j - x_i}{h}\right) \\ &\quad \times\left[\left\{\frac{np_i(1-p_i) + p_i^2}{n^2}\right\}I(i=j) + \left\{\frac{p_ip_j(1-n)}{n^2}\right\}I(i\neq j)\right] \end{aligned}$$

이다. $L_{i,p}\left(\frac{x_j - x_i}{h}\right) = O(1)$ 이기 때문에 위 식을 다시 정리하면

$$\begin{aligned} E(\widehat{p}_i\overline{p}_i) &= \frac{p_i(1-n)}{n^2}\frac{1}{kh}\sum_{j=1}^{k} L_{i,p}\left(\frac{x_j - x_i}{h}\right)p_j + \frac{p_i}{n}\frac{1}{kh}\sum_{j=1}^{k} L_{i,p}\left(\frac{x_i - x_i}{h}\right) \\ &\sim f(x_i)O\left(\frac{1}{nk^2h}\right) + O\left(\frac{1}{n^2k^2h}\right) \end{aligned}$$

이다. 또한 $E(\widehat{p}_i) = p_i + O(h^{p+1}/k)$, $E(\overline{p}_i) = p_i$ 이므로

$$\begin{aligned} E(\widehat{p}_i)E(\overline{p}_i) &= p_i^2 + p_i\,O\left(\frac{h^{p+1}}{k}\right) \\ &\sim \frac{f^2(x_i)}{k^2} + O\left(\frac{h^{p+1}}{k^2}\right) \end{aligned}$$

이다. 그러므로

$$\begin{aligned} Cov(\widehat{p}_i, \overline{p}_i) &= E\{(\widehat{p}_i - E(\widehat{p}_i))(\overline{p}_i - E(\overline{p}_i))\} \\ &= E(\widehat{p}_i\overline{p}_i) - E(\widehat{p}_i)E(\overline{p}_i) \\ &\sim f(x_i)O\left(\frac{1}{nk^2h}\right) + O\left(\frac{1}{n^2k^2h}\right) - \frac{f^2(x_i)}{k^2} + O\left(\frac{h^{p+1}}{k^2}\right) \end{aligned}$$

이다. 따라서 $n \to \infty$, $k \to \infty$, $nk^2h \to \infty$이면 $Var(\hat{p}_i)$, $Var(\overline{p}_i)$, $Cov(\hat{p}_i, \overline{p}_i)$ 모두 0 으로 수렴하므로 $Var(\hat{p}_i - \overline{p}_i) \to 0$ 이다. 그러므로 $\hat{p}_i - \overline{p}_i \xrightarrow{p} E(\hat{p}_i - \overline{p}_i)$ 이며, 즉 $\hat{p}_i - \overline{p}_i = \frac{f^{(p+1)}(x_i)}{(p+1)!}\left\{\int_{-L}^{L} v^{p+1} L_p(v)\, dv\right\}\frac{h^{p+1}}{k} + o_p\left(\frac{h^{p+1}}{k}\right)$ 이다.

$\int_{-L}^{L} v^{p+1} L_p(v)\, dv$를 L^* 라 하자. 그리고 다시 $p_i = \frac{1}{k} f(x_i) + O\left(\frac{1}{k^3}\right)$ 임을 이용하여 B를 구하면 다음과 같다.

$$
\begin{aligned}
B &= \sum_{i=1}^{k} \frac{\left\{\frac{f^{(p+1)}(x_i)}{(p+1)!} L^* \frac{h^{p+1}}{k} + o_p\left(\frac{h^{p+1}}{k}\right)\right\}^2}{p_i} \\
&\sim \sum_{i=1}^{k} \frac{\left(\frac{f^{(p+1)}(x_i)}{(p+1)!} L^*\right)^2 \frac{h^{2(p+1)}}{k^2} + o_p\left(\frac{h^{2(p+1)}}{k^2}\right)}{\frac{1}{k} f(x_i)} \\
&= h^{2(p+1)} \int_0^1 \frac{(f^{(p+1)}(u))^2}{f(u)((p+1)!)^2}\, du L^* + o_p(h^{2(p+1)}) \\
&= h^{2(p+1)} \left(\int_0^1 \frac{(f^{(p+1)}(u))^2}{f(u)((p+1)!)^2}\, du L^* + o_p(1) \right).
\end{aligned}
$$

[보조정리 1]에 의하여 $\overline{p}_i - p_i = O_p\left(\sqrt{\frac{k \ln k}{n}}\right)$ 이므로 C를 다시 정리하면 다음과 같다.

$$
\begin{aligned}
C &= 2 \sum_{i=1}^{k} \frac{(\hat{p}_i - \overline{p}_i)(\overline{p}_i - p_i)}{p_i} \\
&= 2 \sum_{i=1}^{k} \frac{\left\{\frac{f^{(p+1)}(x_i)}{(p+1)!} L^* \frac{h^{p+1}}{k} + o_p\left(\frac{h^{p+1}}{k}\right)\right\} O_p\left(\frac{k \ln k}{n}\right)}{p_i} \\
&\sim O_p\left(\frac{k^2 \ln k}{n}\right) \sum_{i=1}^{k} \frac{\left\{\frac{f^{(p+1)}(x_i)}{(p+1)!} L^* \frac{h^{p+1}}{k^2} + o_p\left(\frac{h^{p+1}}{k^2}\right)\right\}}{\frac{1}{k} f(x_i)} \\
&\sim O_p\left(\frac{k^2 \ln k}{n}\right)\left[\left\{\int_0^1 \frac{f^{(p+1)}(u)}{f(u)(p+1)!}\, du\right\} L^* h^{p+1} + o_p(h^{p+1})\right] \\
&= O_p\left(\frac{h^{p+1} k^2 \ln k}{n}\right)\left[\left\{\int_0^1 \frac{f^{(p+1)}(u)}{f(u)(p+1)!}\, du\right\} L^* + o_p(1)\right].
\end{aligned}
$$

이제 표준화된 통계량을 위의 A, B, C 를 이용하여 다시 표현하면 아래와 같다.

$$\frac{n}{\sqrt{\beta_n}}\left\{\sum_{i=1}^{k}\frac{(\hat{p}_i-p_i)^2}{p_i}-\left(\frac{k-1}{n}\right)\right\}$$

$$=\frac{n}{\sqrt{\beta_n}}\left[\left\{\sum_{i=1}^{k}\frac{(\bar{p}_i-p_i)^2}{p_i}-\left(\frac{k-1}{n}\right)\right\}+\sum_{i=1}^{k}\frac{(\hat{p}_i-\bar{p}_i)^2}{p_i}+2\sum_{i=1}^{k}\frac{(\hat{p}_i-\bar{p}_i)(\bar{p}_i-p_i)}{p_i}\right]$$

$$=\frac{n}{\sqrt{\beta_n}}\left\{A-\left(\frac{k-1}{n}\right)\right\}+\frac{n}{\sqrt{\beta_n}}B+\frac{n}{\sqrt{\beta_n}}C.$$

첫째 항은 **[보조정리2]**에 의해서 $\frac{n}{\sqrt{\beta_n}}\left\{A-\left(\frac{k-1}{n}\right)\right\}\xrightarrow{d}N(0,1), n\to\infty$ 이다.

둘째항은 원래 $\frac{n}{\sqrt{\beta_n}}B\sim\frac{nh^{2(p+1)}}{\sqrt{\beta_n}}\left[\int_0^1\frac{(f^{(p+1)}(u))^2}{f(u)((p+1)!)^2}du\,L^*+o_p(1)\right]$

이다. 그런데 $\frac{nh^{2(p+1)}}{\sqrt{\beta_n}}\to 0$ 이므로 $\frac{n}{\sqrt{\beta_n}}B=o_p(1)$ 이 된다. 또한

$\frac{n}{\sqrt{\beta_n}}C\sim O_p\left(\frac{h^{p+1}k^2\ln k}{\sqrt{\beta_n}}\right)\left[\int_0^1\frac{f^{(p+1)}(u)}{f(u)(p+1)!}duL^*+o_p(1)\right]$ 이며,

$\frac{h^{p+1}k^2\ln k}{\sqrt{\beta_n}}\to 0$ 이므로 $\frac{n}{\sqrt{\beta_n}}C=o_p(1)$ 이다. 그러므로 Slutsky 정리에 의하여

$\frac{n}{\sqrt{\beta_n}}\left\{\sum_{i=1}^{k}\frac{(\hat{p}_i-p_i)^2}{p_i}-\left(\frac{k-1}{n}\right)\right\}\xrightarrow{d}N(0,1)$ 이다.

참고문헌

1. Aerts, M. A. , Augustyns, I. and Janssen P.(1997a). Smoothing Sparse Multinomial Data using Local Polynomial Fitting, *Nonparameteric Statistics*, 8. 127-147.
2. Aerts, M. A. , Augustyns, I. and Janssen P.(1997b), Sparse consistency and smoothing for multinomial data, *Statistics & Probability Letters*, 33, 41-48.
3. Aerts, M. A. , Augustyns, I. and Janssen P.(2000), Central limit theorem for the total squared error of local polynomial estimators of cell probabilities, *Journal of Statistical Planning and Inference*, 91, 181-193.
4. Baek, J. (1998), A Local Linear Kernel Estimator for Sparse Multinomial Data, *Journal of the Korean Statistical Society*, 27, 515-529.
5. Burman, P. (1987a), Smoothing Sparse Contingency Tables, *Sankhyaser. A.* , 49, 24-36.
6. Burman, P. (1987b), Central Limit Theorem for Quadratic Forms for

Sparse Tables, *Jonrnal of Multivariate Analysis*, 22, 258-277.
7. Lawal, H. B. , and Upton , G. J. G. (1980), An Approximation to the Distribution of the X^2 Goodness-of-Fit Statistic for use with Small Expectations, *Biometrika*, 67, 447-453.
8. Liero, H. (2001), L_2-tests for sparse multinomials, *Statistics & Probability Letters*, 55, 147-158.
9. Morris, C. (1975), Central Limit Theorems for Multinomial Sums, *Annals of Statistics*, 3, 165-188.
10. Simonoff, J. S. (1985), An Improved Goodness-of-Fit Statistic for Sparse Multinomials, *Journal of the American Statistical Society*, 80, 671-677.

[2003년 3월 접수 , 2003년 5월 채택]