The Korean Communications
in Statistics Vol. 10 No. 2, 2003
pp. 371-385

Contour Plot to Explore the Structure of Categorical Data

Hyun Chul Kim[1], Moon Yul Huh[2], Hee Suk Chung[3]

Abstract

In this paper, contour plot is considered as a method to explore the structure of categorical data. For this purpose, the paper suggests a method to sort two-way contingency table with respect to the expected marginals. It is found that the suggested plot provides us with valuable information for the underlying data structure. Firstly, we can investigate independency between the categories by examining the differences of expected frequency contours and observed frequency contours. With the plot, we can also visually investigate the existence of outliers inherent in the data. These properties of the suggested contour plot will be demonstrated by several sets of real data.

Keywords : Categorical Data, Contour Plot, Nominal Data, Sorting by Marginals.

제 1 장 서론

등고선 그림은 3차원 곡면을 2차원 평면 상에 나타내는 방법이다. 즉 (x, y, z)로 표시되는 3차원 점들이 있을 때, z가 주어지면 이 값에 대응하는 (x, y)의 점들을 연결하는 곡선을 등고선이라고 한다. 등고선을 그리는 알고리즘으로는 평행한 격자선의 교차로 만들어진 사각형 그물망을 이용하는 방법과 그 사각형의 대각선을 연결하여 만들어진 삼각형의 그물망을 이용하는 방법이 있다 (Scott, 1992). 또 격자점 사이를 직선이 아닌 곡선으로 연결하여 매끄러운 등고선을 그릴 수도 있다 (Bates 등, 1993).

분할표 형태로 주어진 자료에 대해 등고선 그림을 그린다 해도 범주가 연속이면 보간을 해도 큰 문제가 없다. 예를 들어 이변량 정규분포로부터 얻은 자료를 2차원으로 표현할 때 등고선이 유

1) Associate Professor, Department of Informatics and Statistics, Kunsan National University, Kunsan, KOREA
E-mail : kimhc@kunsan.ac.kr
2) Professor, Department of Statistics, SungKyunKwan University, Seoul, KOREA
E-mail : myhuh@skku.ac.kr
3) Automobile Insurance Department, Firstfire & Marine Insurance Co., Ltd, Seoul, KOREA
E-mail : cstone3@hanmail.net

용하게 사용될 수 있다. x와 y가 연속값이기 때문에 격자와 격자 사이의 값이 실제로 존재하여 보간으로 그 값을 구하는 것은 오차가 있을지라도 의미가 있는 것이다. 같은 이유로 자료에 결측값이 발생한다 하더라도 보간에 의해 결측값을 대체할 수 있다.

간단한 모의자료를 통해 등고선 그림의 활용을 살펴보자. 자료를 만든 방법은 x와 y를 각각 -5부터 5까지의 정수로 하고, 각 x, y 값의 조합에 대해 $z = x^2 + y^2$를 계산해서 z를 구하였다. 이 값을 등고선 그림으로 나타내면 <그림 1>과 같다. 이 등고선 그림은 R (The R Development Core Team, 2003)로 그린 것이다.

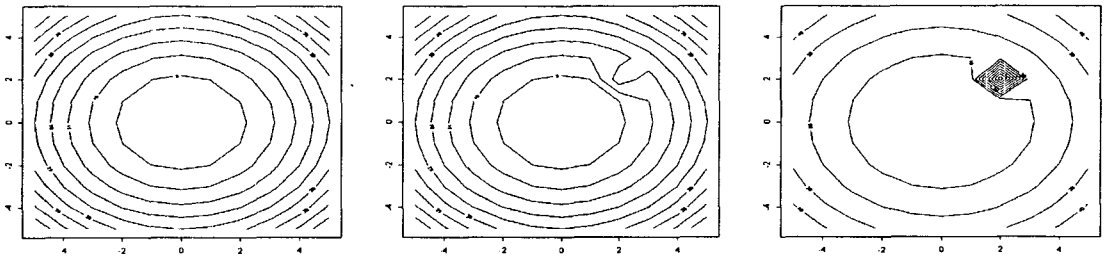
이 그림은 전체의 분포를 이해하는데 도움이 된다. 비록 등고선 그림은 색이나 숫자로 등고선의 의미를 이해해야 하기 때문에 익숙해지지 않으면 분포를 보기 힘들다는 단점을 가지고 있지만 여러 각도에서 바라보아야 하는 3차원 자료를 한 장의 그림으로 나타낼 수 있다는 장점이 있다.

이상값이 있을 때 등고선 그림이 받는 영향을 살펴보기위해 $(x, y) = (2, 2)$에서 $z = 8$ 대신 18로 입력된 경우와 98로 입력된 경우를 살펴본다. 먼저 약한 이상값 (18)이 있는 경우를 살펴보면 <그림 1>의 ②와 같이 튀어나온 부분으로 이상값이 반영된다. 심한 이상값 (98)이 있을 때는 ③과 같이 뾰족한 봉우리가 나타난다. 한 가지 유의할 점은 봉우리가 정확하게 (2, 2)에 만들어지지 않을 수도 있다는 점이다.

등고선 그림을 범주형 자료에 사용할 때는 문제가 된다. 각 범주의 거리를 정확한 수치로 환산할 수 없거나, 실제로는 존재하지 않는 범주와 범주사이의 값이 등고선 그림에 나타나게 되고, 이는 자칫 사용자가 오해하게 만드는 원인이 될 수도 있기 때문이다.

범주형 자료의 분포를 살펴 볼 수 있는 도구로 막대그래프를 사용할 수 있으며, 2차원 분할표의 경우에는 3차원 막대그래프를 사용할 수 있다. 3차원 막대그래프는 입체로 표현되기 때문에 앞쪽에 큰 막대가 배치되면 읽는데 어려움이 발생할 수 있다. 그러나 등고선 그림을 사용하면 이런 어려움을 피할 수 있다. 그런데 각 범주가 명목형 변수일 경우에는 범주의 배열 순서에 의미가 없으므로 각 범주 순서의 조합에 따라 다양한 그림이 나타날 수 있다. 그러므로 등고선 그림을 사용하거나 3차원 막대그래프를 사용하려면 시각적 의미전달 효과가 가장 좋은 범주 조합을 선택할 필요가 있다.

범주형 자료의 분석에서 중요한 내용은 변수 사이의 연관성 (혹은 반대로 독립성)과 이상값의 탐색이다. 본 연구에서는 2차원 범주형 자료를 등고선 그림으로 표현함으로써 두 변수간의 연관성 파악과 이상값 탐색이 효율적으로 이루어지는 것을 보이고자 한다.



① 원자료 ② 약한 이상값 (18)의 경우 ③ 심한 이상값 (98)의 경우

<그림 1> 모의자료에 대한 등고선 그림

제 2 장 범주형 자료의 등고선 그림

1차원 범주형 자료를 표현하는 가장 대표적인 방법은 막대그래프이다. 범주형 자료를 그래프로 표현할 때 각 범주의 순서는 의미가 없다. 따라서 일정한 기준에 따라 순서를 재배열해서 효율적인 그래프를 작성할 필요가 있다. 여기서 생각할 수 있는 배열은 막대의 크기 (z값) 순으로 재배열하는 것이다. 이렇게 정렬하면 큰 막대부터 작은 막대 순으로 나타나기 때문에 전체 분포의 모양을 한 눈에 알아보는 데 유리하다.

2차원 이상의 범주형 자료를 정렬하는 문제는 간단하지 않다. 행 (또는 열)단위로 정렬을 해야 하기 때문이다. 그렇다고 가능한 모든 범주 값의 조합에 대해 그림을 그리고 가장 나은 것을 찾는 것 역시 매우 어렵다. 4x4 분할표의 경우 전체 배열 방법은 576가지가 된다.

본 논문에서 분할표를 재배열하는 기준으로 주변합의 크기를 사용하고자 한다. 주변합의 크기로 분할표를 정렬하는 것은 결과적으로 기대도수를 크기 순으로 정렬하는 것과 같다.

총 n개의 관측값을 갖는 $I \times J$ 분할표를 생각해 보자. 이 분할표에서 i 행과 j 열의 주변합을 각각 $n_{i\cdot}$, $n_{\cdot j}$ 라 하자. 이 때 (i, j) 번째 기대도수, E_{ij}는 행과 열의 주변확률을 각각 $p_{i\cdot}$, $p_{\cdot j}$라고 했을 때 다음 식과 같이 얻을 수 있다.

$$E_{i,j} = n \times p_{i\cdot} \times p_{\cdot j} = \frac{n_{i\cdot} \times n_{\cdot j}}{n}, \quad i=1,\cdots,I, \quad j=1,\cdots,J$$

주어진 i 행에 속한 모든 칸의 기대도수 E_{ij}는 모두 같은 주변합 $n_{i\cdot}$을 갖는다. 따라서 E_{ij}는 $n_{\cdot j}$에만 의존하게 되어 i 행을 주변합의 크기 $n_{\cdot j}$의 크기로 정렬하면 i 행의 기대도수는 크기 순으로 정렬된다. 같은 이유로 j 열에 대해서도 같은 정렬이 이루어진다. 따라서 2차원 분할표를 주변합의 크기에 의해 정렬하면 각 칸은 기대도수의 크기 순으로 정렬되는 결과를 얻게 된다.

<표 1>의 자료는 Snee (1974)가 595명의 학생을 대상으로 모발 색과 눈동자 색 사이의 관계를 알아보기 위해 조사한 자료이다. 이 자료를 주변합 크기의 내림차순으로 정렬하면 <표 2>와 같다. 여기서 괄호 안의 값은 기대도수이다. 기대도수의 크기가 모든 행과 열에서 크기 순으로 정렬된 결과를 볼 수 있다. 이 기대도수에 대한 막대그래프를 그리면 <그림 2>의 ②와 같다. 이제 각 범주가 비록 명목척도의 값이지만 이를 연속으로 간주하고 등고선 그림을 그려 보자. <그림 2>의 ④는 기대도수에 대한 등고선 그림이다.

한편 <그림 2>의 ①과 ③은 기대도수와 같은 순서로 관측도수를 정렬한 뒤에 그린 막대그래프와 등고선 그림이다. 이미 크기 순으로 잘 정렬된 관측도수에 대해서 그린 등고선은, 각 범주값들 사이에 대한 보간 결과를 실제 있는 값으로 오해하지만 않는다면, 막대그래프에 비해 전체 분포 모양을 이해하는데 훨씬 도움이 된다. 막대그래프는 다른 것들에 비해 큰 막대들이 전면에 배치되면 뒤쪽에 숨은 막대들을 파악하기 힘들게 되는 단점이 있기 때문이다.

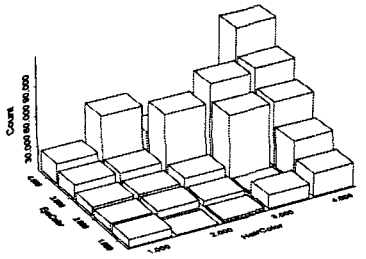
<표 1> 모발 색과 눈동자 색 자료

눈동자색 \ 모발색	검정색	고동색	붉은색	금발	합
초록색	5	29	14	16	64
엷은 갈색	15	54	14	10	93
푸른색	20	84	17	94	215
고동색	68	119	26	7	220
합	108	286	71	127	592

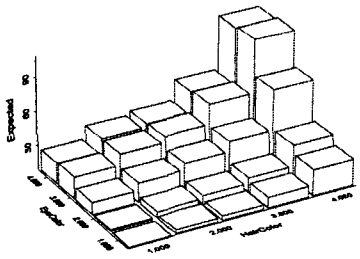
<표 2> 주변합을 이용한 분할표의 정렬 결과

눈동자색 \ 모발색	고동색(4)	금발(3)	검정색(2)	붉은색(1)	합
고동색(4)	119(106)	7(47)	68(40)	26(26)	220
푸른색(3)	84(104)	94(46)	20(39)	17(26)	215
엷은 갈색(2)	54(45)	10(20)	15(17)	14(11)	93
초록색(1)	29(31)	16(14)	5(12)	14(8)	64
합	286	127	108	71	592

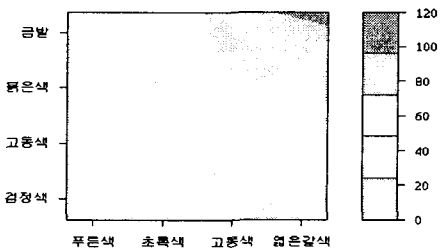
주: 1. 범주의 이름 뒤에 있는 괄호 안은 막대그래프에 표시된 이름
2. 관측도수의 괄호 안에 있는 값은 기대도수



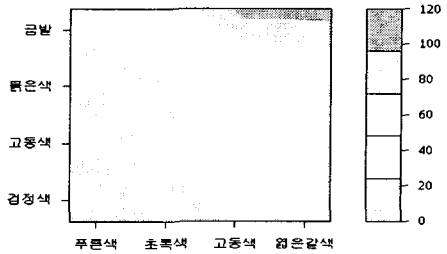
① 관측도수의 막대그래프



② 기대도수의 막대그래프



③ 관측도수 등고선



④ 기대도수 등고선

<그림 2> 기대도수 (주변합) 크기 순으로 정렬한 그림

또 관측도수와 기대도수를 기대도수 크기 순으로 정렬하여 그린 등고선 그림을 서로 나란히 그려서 비교하면 범주형 자료의 독립성에 대한 시각적인 정보를 얻을 수 있다. 기대도수에 의한 그림과 관측도수에 의한 그림을 서로 비교하여 서로 비슷한 그림이 만들어지면 이는 두 범주형 변수가 서로 독립임을 의미하는 것으로 해석할 수 있으며, 서로 눈에 띠게 다르면 서로 독립이 아님을 의미하는 것으로 해석할 수 있다. 자세한 내용은 3장에서 다룬다.

제 3 장 등고선 그림의 응용

3.1 독립성 탐색

$I \times J$의 2차원 분할표에서 카이제곱 통계량은 다음과 같이 정의한다.

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

여기서 O_{ij} 와 E_{ij} 는 각각 관측도수와 기대도수를 나타낸다. 카이제곱 통계량에서 기대도수는 독립을 가정하고 계산되므로, 이 기대도수와 관측도수 사이에 차이가 크게 나면 독립이 아닌 것으로 검정되고, 차이가 적으면 독립이라는 가정이 타당한 것으로 보아 독립인 것으로 검정된다. 따라서 기대도수에 의한 등고선 그림과 관측도수에 의한 등고선 그림을 서로 비교하여 비슷한 그림이 만들어지면 이는 두 범주형 변수가 서로 독립임을 의미하는 것으로 해석할 수 있으며, 서로 눈에 띠게 다르면 서로 독립이 아님을 의미하는 것으로 해석할 수 있다.

한편 독립성을 검정할 때 사용하는 카이제곱 통계량은 범주의 수가 많아지거나 관측값의 수가 적을 때는 임계값이 커져 귀무가설을 기각하지 못하는 문제가 발생한다. 그러나 등고선 그림으로 이 값을 나타내면 분포를 비교적 쉽게 알 수 있으며, 한 그림에 나타낼 수 있는 장점으로 인해 연관성에 대해 효율적으로 탐색할 수 있다. 이것을 모의자료를 통해 살펴보도록 하자.

3.1.1 상관관계가 있는 모의자료의 생성방법

상관관계가 있는 이변량 정규난수를 생성한 후에 이 자료로 분할표를 만들어 모의자료로 사용했다. 먼저 상관관계가 있는 이변량 정규난수를 다음과 같이 생성하였다.

1단계: 표준 정규분포를 갖는 두 개의 난수 (X_1, X_2)를 생성

2단계: $Y_1 = \rho X_1 + (1-\rho)^{1/2} X_2$를 계산하면 이렇게 만들어진 (X_1, Y_1)는 상관계수가 ρ인 이변량 정규난수임.

분포의 왜곡을 방지하기 위해 각 칸에 들어갈 확률이 동일해지도록 구간을 나누었다. z의 값에 따라 각각 다음 <표 3>과 같이 범주를 나누었다. 또 상관계수는 1.0, 0.8, 0.3, 0.0의 4 가지 경우에 대해 각각 크기 100의 모의자료를 만든 후 등고선 그림을 통해 독립성을 탐색을 해 보았다.

<표 3> 모의자료의 범주 구분

z의 범위	$z \leq -1.2$	$-1.2 < z \leq -0.4$	$-0.4 < z \leq 0.4$	$0.4 < z \leq 1.2$	$z > 1.2$
x 범주 이름	A	B	C	D	E
y 범주 이름	a	b	c	d	e

4 개 모의자료에 대한 피어슨 카이제곱 통계량, p-값, 피어슨의 연관성 측도가 <표 4>에 나타나있다. 여기서 모집단의 상관계수가 0.3과 0인 경우, 피어슨 연관성 측도가 0.4236 및 0.3550과 같이 매우 크게 나타났다. 이는 범주형 난수를 만드는 과정에서 연속형 변수를 무리하게 몇 개의 구간으로 나누기 때문에 나타나는 문제다 (Lee와 Huh, 2003).

3.1.2 상관정도에 따른 독립성 탐색

먼저 <그림 3>에 있는 상관계수가 1인 모집단에서 얻은 모의자료의 등고선을 보면, 기대도수의 등고선은 우측 상단이 가장 높은 값을 갖고 왼쪽 대각선으로 내려오면서 빈도가 작아지는 것을 발견할 수 있다. 우리가 기대도수를 정렬하는 방법으로 행과 열을 내림차순으로 정렬하였다. 따라서 분할표 상에서 왼쪽 윗부분이 큰 값을 갖게 정렬된다. R에서 등고선 그림을 그리면 좌우가 바뀌어 나타난다. 그런데 <그림 3>의 ②를 보면 관측도수로 그린 등고선들은 거의 대각선에 가깝게 나타난다. 따라서 ①과 ②의 두 등고선의 구조가 전혀 달라 우리는 이 분할표의 자료가 서로 독립이 아님을 알 수 있다. 한편 이 둘의 차이를 보여주는 잔차의 등고선 그림을 그려 보았다. 잔차의 등고선 그림인 ③에는 뚜렷한 패턴이 존재함으로써 잔차가 랜덤하지 않음을 알 수 있다. 따라서 이 그림 역시 두 범주 집단간에 서로 독립이지 않음을 보여주는 증거이다. 여기서 잔차 (e_{ij})는 다음과 같다.

$$e_{ij} = O_{ij} - E_{ij}$$

<그림 3>부터 <그림 6>을 참고하면 다음과 같은 몇 가지 특징을 발견할 수 있다. 첫째, 당연한 결과이지만 기대도수 등고선 그림은 상관계수의 크기에 상관없이 항상 일정한 방향으로 패턴을 갖는 모양으로 나타난다. 둘째, 관측도수 등고선 그림은 상관계수의 크기가 작아짐에 따라 점점 기대도수 등고선 그림에 가깝게 변해 간다. 셋째, 잔차의 등고선 그림은 상관계수가 클 때는 뚜렷한 패턴을 보이다가 상관계수가 점점 작아짐에 따라 랜덤한 모양으로 변해 간다.

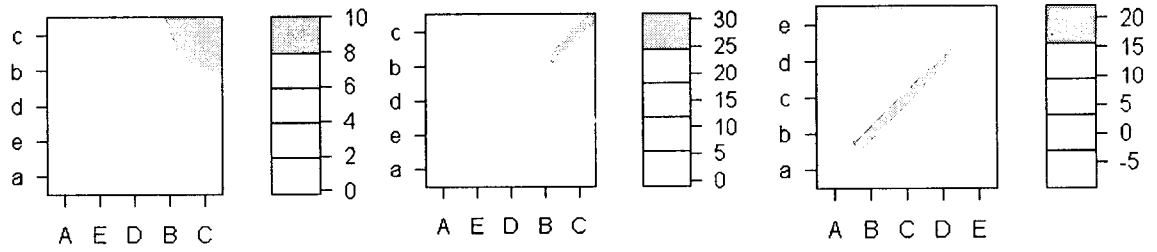
따라서 이상을 종합해 볼 때 기대도수의 등고선 그림과 관측도수의 등고선 그림이 비슷하면 범주 변수들 간에 서로 독립임을 보여주며, 반대로 두 그림이 서로 달라지면 달라질수록 서로 독립이 아님을 보여주는 것으로 해석할 수 있다. 한편 이 두 그림의 비교에 잔차의 등고선 그림은 보조 자료의 역할을 할 수 있다. 잔차의 등고선 그림은 랜덤한 정도로 범주 변수가 독립인 정도를 판단할 수 있다.

<표 4>에서 3번째 자료의 경우, 상관계수가 0.3인 모집단으로부터 추출 된 것으로서 두 변수 사이에는 연관성이 있는 모집단으로부터 생성되었다. 그러나 χ^2-검정의 p-값이 0.147로 나타나

두 변수 간에 독립적이라는 가설을 기각할 수 없다. 즉, 두 변수가 서로 독립적이라는 결론을 내리게 된다. 이 문제를 등고선 그림을 사용하여 분석해보자. <그림 5>에 보면 관측도수의 등고선 그림도 기대도수의 등고선 그림과는 사뭇 다른 모양을 보여주고 있으며, 잔차의 등고선 그림은 일정한 패턴을 보여주고 있다. 따라서 등고선 그림에 의하면 두 범주 사이에 서로 독립이 아니라고 결론을 내릴 수 있다.

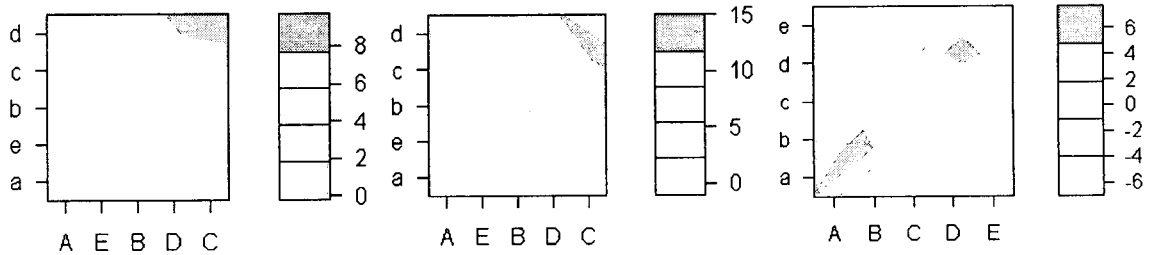
<표 4> 모의자료의 각종 통계량 비교

모집단의 상관계수(ρ)	피어슨의 χ^2-통계량	χ^2-검정의 p-값	피어슨의 연관성측도
1.0	400.00	0.000	0.8944
0.8	99.11	0.000	0.7055
0.3	21.87	0.147	0.4236
0.0	14.42	0.567	0.3550



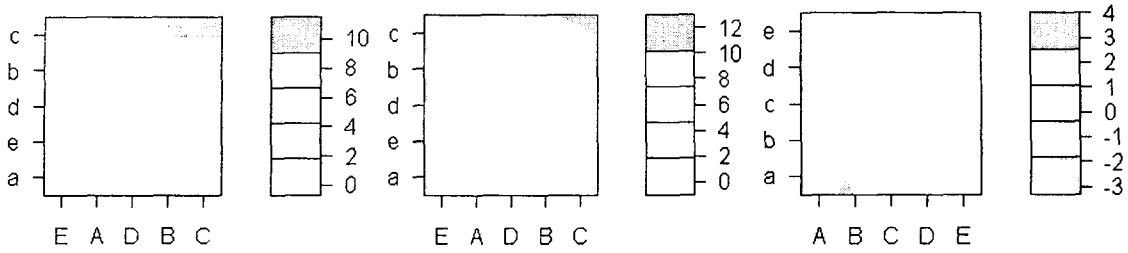
① 기대도수 등고선 ② 관측도수 등고선 ③ 잔차 등고선

<그림 3> 주변합으로 정렬한 등고선 그림 (ρ=1.0)



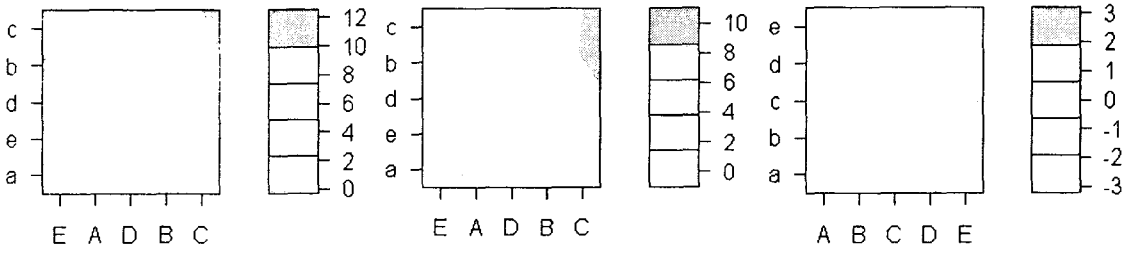
① 기대도수 등고선 ② 관측도수 등고선 ③ 잔차 등고선

<그림 4> 주변합으로 정렬한 등고선 그림 (ρ=0.8)



① 기대도수 등고선 ② 관측도수 등고선 ③ 잔차 등고선

<그림 5> 주변합으로 정렬한 등고선 그림 (ρ=0.3)



① 기대도수 등고선 ② 관측도수 등고선 ③ 잔차 등고선

<그림 6> 주변합으로 정렬한 등고선 그림 (ρ=0.0)

3.2 이상값 탐색

이제 등고선 그림을 이용하여 이상값을 탐색하는 방법을 살펴보자. Hong과 Lee (2001)는 이상값이 존재하는 분할표를 생성하기 위해 Simonoff (1988)가 제안한 칸 확률에 의한 다항확률 표본추출을 이용하였다. 여기서는 이 자료에 대해서 등고선 그림만을 가지고 이상값을 탐색해 보려고 한다. 원자료는 <표 5>와 <표 6>에 있으며, 자세한 자료는 Hong과 Lee를 참고하기 바란다.

우리는 앞에서 재배열된 자료에 대한 등고선 그림의 패턴이 기대도수 등고선 그림의 패턴과 많은 차이가 나는 경우 서로 독립이 아니라고 판단했었다. 이런 차이가 그림 전체에서 나타나는 경우와 일부에서 나타나는 경우로 나누어 생각해 볼 수 있다. 전반적인 흐름에서 차이가 난다면 독립성을 의심해 볼 수 있고, 부분적으로 나타나면 이상값을 의심해 볼 수 있다.

한편 이런 이상값 탐색에 잔차의 등고선 그림을 함께 그려보면 더욱 좋다. 어느 칸이 이상값이면 그 이상값이 포함된 행과 열의 주변합에 영향을 주게 되고 따라서 같은 행과 열의 기대도수는 동시에 영향을 받게 된다. 이런 사실은 특정한 행과 열에서 잔차의 절대값 크기가 다른 행과 열에 비해 두드러지면 그 행과 열이 만나는 칸이 이상값일 수 있다는 의미가 된다. 특히 좀 더 로버스트한 잔차 (표준화된 잔차, 수정된 잔차, 삭제된 잔차 등)를 사용하면 이상값을 좀 더 잘 판별할 수도 있을 것이다. 여기서는 삭제된 잔차 (deleted residual), r_{ij} 를 사용하였다 (Simonoff, 1988).

$$r_{ij} = (O_{ij} - E^*_{ij})/\sqrt{E^*_{ij}}$$

여기서 E^*_{ij}는 (i,j) 칸을 제외한 준독립모형 (quasi-independence model) 하에서 (i,j) 칸의 기대도수로 다음과 같이 정의된다.

$$E^*_{ij} = (n_{i.} - O_{ij})(n_{.j} - O_{ij}) / (n - n_{i.} - n_{.j} + O_{ij})$$

3.2.1 하나의 이상값이 존재하는 경우

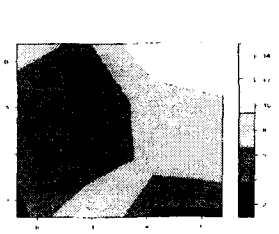
<표 5>에 주어진 4x4 2차원 분할표는 등고선 그림을 통해 하나의 이상값을 탐색하는 것을 보여 주기 위한 모의자료이다. 이 표에서 (d, D) 칸에 1이 아닌 11이 입력되었다고 하자. 큰 이상값이 사용된 것이다. 이렇게 바뀌면 기대도수는 4 ($= 24 \times 19 / 110$)가 된다. 이 때의 관측도수에 의한 등고선 그림인 <그림 7>의 ①을 보자. (d, D) 부근에서 등고선들의 배열을 흐트러뜨리는 뾰족한 봉우리를 볼 수 있어 칸 (d, D)가 이상값으로 의심된다. 삭제된 잔차 등고선을 그리면 (d, D)에 급격한 봉우리가 나타남으로써 (d, D)가 기대값보다 큰 이상값임을 뒷받침한다.

두 번째로 <표 5>의 (a, B) 칸에 12 대신 2를 사용했을 경우를 살펴보자. 이렇게 되면 기대도수는 7 ($= 21 \times 30 / 90$)이 된다. 이 경우의 등고선 그림은 <그림 8>에 있다. 여기에서도 우리는 ① 번 그림에서 (a, B)가 이상값일 것이라고 예상할 수 있다. 삭제된 잔차의 등고선 그림인 ②를 보면 (a, B)에서 등고선의 모양이 급격한 계곡을 이루고 있어 이상값임을 알 수 있다.

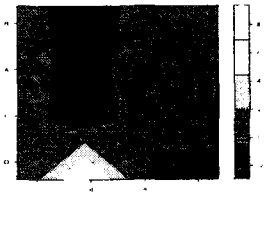
또 삭제된 잔차 그림에서 한 가지 공통적인 특징이 발견된다. 하나의 이상값이 있을 때 그 이상값을 포함하는 행과 열의 등고선이 주변과 유난히 다른 색깔로 나타난다는 것이다.

<표 5> 이상값 탐색을 위한 4x4 2차원 분할표 (원자료)

	A	B	C	D	주변합
a	8	12	8	3	31
b	5	8	5	2	20
c	9	14	9	3	35
d	4	6	3	1	14
주변합	26	40	25	9	100

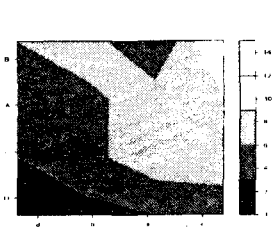


① 관측도수 등고선

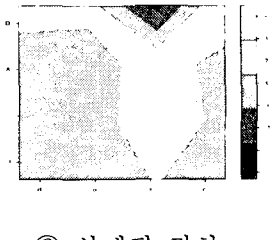


② 삭제된 잔차 등고선

<그림 7> (d, D)가 큰 이상값인 경우



① 관측도수 등고선



② 삭제된 잔차 등고선

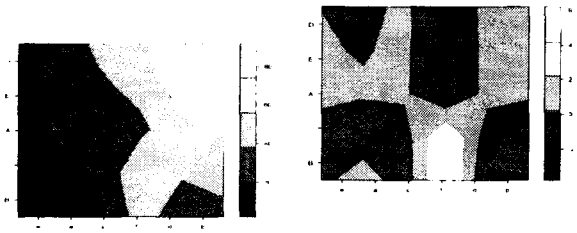
<그림 8> (a, B)가 작은 이상값인 경우

<표 6> 이상값 탐색을 위한 6x5 2차원 분할표 (원자료)

	A	B	C	D	E	합
a	22	19	14	29	26	110
b	58	31	46	91	74	300
c	28	14	21	42	35	140
d	52	25	40	78	65	260
e	7	3	4	9	8	31
f	34	17	25	51	43	170
합	201	109	150	300	251	1011

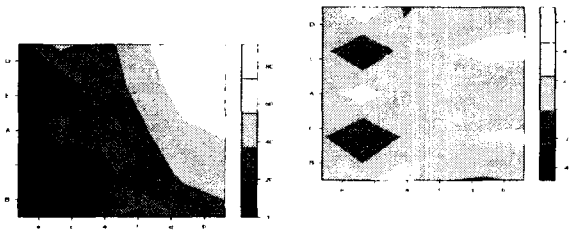
3.2.2 동일한 행 혹은 열에 두 개의 이상값이 존재하는 경우

이제 <표 6>의 모의자료를 통해 6x5 2차원 분할표에서 등고선 그림으로 두 개의 이상값을 탐색해 보자. <그림 9>는 <표 6>의 f 행의 (f, B), (f, C) 칸에 17과 25 대신 47, 55를 사용한 경우이다. 둘 다 큰 이상값이 사용된 것이다. 이렇게 바꾸면 기대도수는 각각 30과 39가 된다. 먼저 그림 ①을 보자. f열의 B, C가 포함된 등고선이 전체 패턴을 파괴하고 있다. 이 등고선이 부드럽게 우측 아래 방향으로 이어지는 것이 훨씬 자연스러운 등고선이 될 것이기 때문이다. 한편 관측도수 등고선의 일부분에서 패턴이 파괴되어 나타나므로 이상값이 있는 경우로 판단된다. 그림 ②의 삭제된 잔차 등고선을 보면, (f, B), (f, C) 주변에서 뾰족한 봉우리가 나타나 이들이 이상값임을 확인할 수 있다.



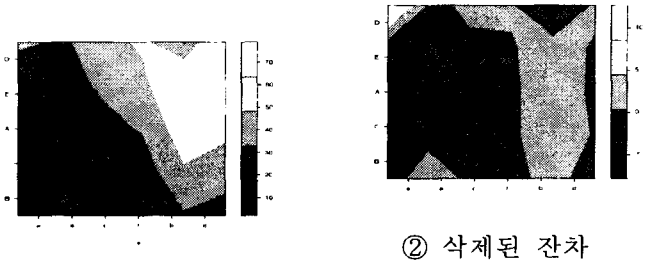
① 관측도수 등고선 ② 삭제된 잔차 등고선

<그림 9> (f, B), (f, C)가 큰 이상값



① 관측도수 등고선 ② 삭제된 잔차 등고선

<그림 10> (c, C), (c, E)가 작은 이상값



① 관측도수 등고선 ② 삭제된 잔차 등고선

<그림 11> (b, D)는 기대도수보다 작고, (e, D)는 큰 이상값인 경우

이번에는 같은 행에 기대도수보다 작은 이상값이 두 개 나타나는 경우를 살펴보자. (c, C), (c, E) 칸의 관측도수는 원래 21과 35인데 이곳에 각각 1과 5가 나타났다고 가정하자. 이렇게 이상값이 나타나면 각각의 기대도수는 12와 21이 된다. <그림 10>의 ①은 관측도수의 등고선 그림이다. 이 그림을 보면 c 열이나 a 열이 톱니 모양으로 나타나 이상값이 의심되는 곳이다. 이제 그림 ②의 삭제된 잔차 등고선을 보면 c 열과 C, E 행에 유난히 크거나 작은 잔차들이 모여 있는 것을 볼 수 있다. 또 (c, C), (c, E) 주변에서 깊은 계곡이 나타나 이상값임을 확인할 수 있다.

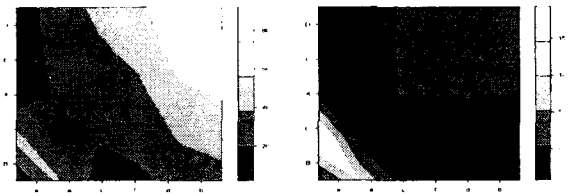
이제 두 칸이 이상값을 갖되, 하나는 기대값보다 큰 이상값이고, 다른 하나는 작은 이상값인 경우를 살펴보자. <표 6>의 (b, D), (e, D) 칸의 원자료는 91과 9인데, 실제로는 31과 39로 이상값이 나타난 경우이다. 이렇게 바꾸면 기대도수는 각각 66, 17이 된다. <그림 11>의 ①에서 관측도수 등고선을 보면 (b, D)는 기대도수보다 작은 이상값으로, (e, D)는 기대도수보다 큰 이상값으로 의심된다. 그림 ②의 삭제된 잔차 등고선을 보면 여기에서도 (b, D)는 뾰족한 봉우리로 나타나고 (e, D)는 깊은 계곡으로 나타난다. 따라서 (b, D) 칸은 기대도수보다 큰 이상값으로, (e, D) 칸은 기대도수보다 작은 이상값으로 판단할 수 있다.

3.2.3 서로 다른 행과 열에 두 개의 이상칸이 존재하는 경우

이번에는 서로 다른 행과 열에 각각 하나씩의 이상값이 있는 경우를 살펴보자. 역시 <표 6>의 원자료를 이용한다.

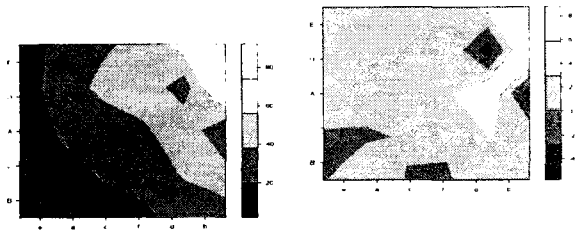
<그림 12>는 (a, B), (e, C)에 각각 기대도수보다 큰 이상값이 나타나는 경우를 보여 준다. 원자료는 (a, B)와 (e, C) 칸의 관측도수가 각각 19와 4였는데, 41과 44가 나타난 경우이다. 이렇게 이상값이 나타나면 기대도수는 각각 16과 13이 된다. ① 번 관측도수 등고선 그림을 보면 왼쪽 아래 부분에 비스듬한 밝은 색 띠가 나타난 부분과 바로 우측으로 비스듬히 윗부분에 나타난 단절된 마디를 함께 볼 때 2 개의 이상값이 있을 것으로 짐작할 수 있다. ② 번의 삭제된 잔차 등고선 그림은 바로 같은 위치에서 급경사의 능선이 나타남으로써 (a, B), (e, C)가 이상값임을 보여준다.

<그림 13>은 (b, A), (d, D)에 각각 기대도수보다 작은 이상값이 나타나는 경우를 보여 준다. 원자료는 (b, A)와 (d, D) 칸의 관측도수가 각각 58과 78이었는데, 18과 28이 나타난 경우이다. 이렇게 이상값이 나타나면 기대도수는 각각 45와 57이 된다. ① 번 관측도수에 의한 등고선 그림을 보면 오른쪽 위 부근에 두 개의 웅덩이가 만들어져 여기가 이상값일 것으로 예상된다. ② 번의 삭제된 잔차 등고선 그림은 같은 장소에 급격한 웅덩이 둘이 만들어져 이런 판단을 뒷받침한다. 따라서 (b, A)와 (d, D)가 기대도수보다 작은 이상값임을 확인할 수 있다.



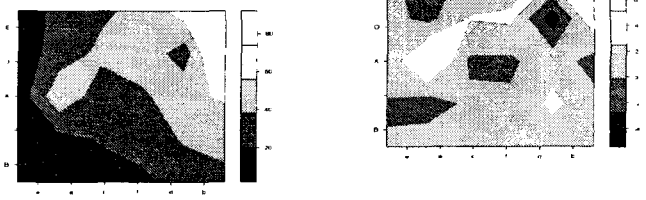
① 관측도수 등고선　② 잔차 등고선

<그림 12> (a, B), (e, C)가 큰 이상값



① 관측도수 등고선　② 삭제된 잔차 등고선

<그림 13> (b,A), (d,D)가 작은 이상값



① 관측도수 등고선　　② 삭제된 잔차 등고선

<그림 14> (a, A)는 기대도수보다 크고, (d, D)는 작은 이상값인 경우

마지막으로 <그림 14>를 보자. (a, A), (d, D)의 원자료는 각각 22와 78이었는데 52와 28로 서로 다른 종류의 이상값이 발생한 경우이다. 이렇게 되면 기대도수는 각각 33, 53이 되어 하나는 기대도수보다 크고, 다른 하나는 기대도수보다 작은 값이 입력된 경우가 된다. 역시 관측도수에 의한 등고선 그림 ①은 (d, D)에는 웅덩이가 (a, A)에는 주변보다 높은 능선이 길게 이어져 이상값일 가능성이 있는 것을 보여준다. 삭제된 잔차의 등고선 그림 ②를 보면 (a, A)에는 급경사의 봉우리가 만들어져 기대도수보다 큰 이상값임을 보여준다. 또 (d, D)는 주변보다 급경사의 웅덩이가 만들어져 기대도수보다 작은 이상값임을 알 수 있다.

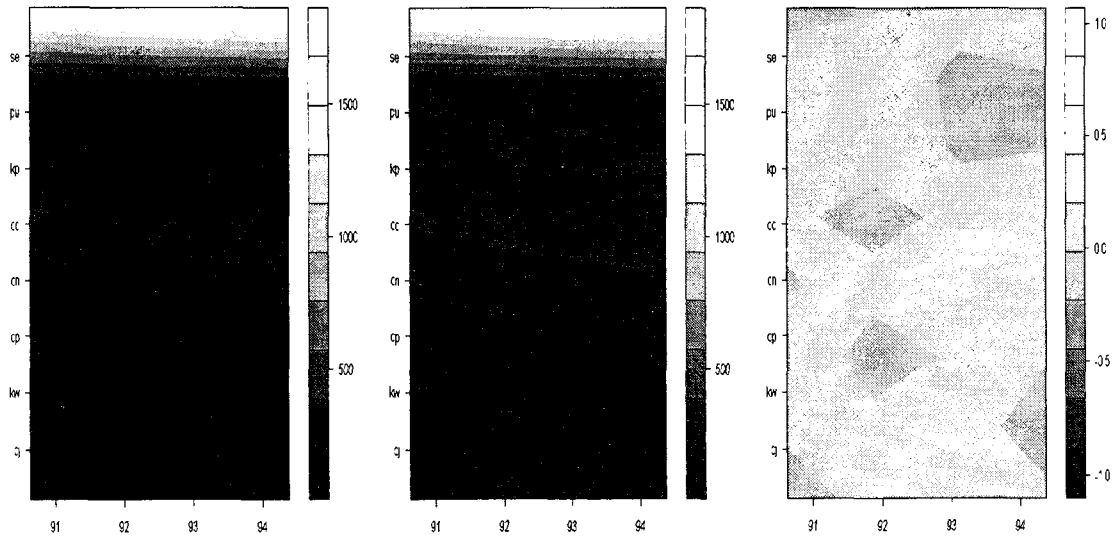
제 4 장 사례분석

4.1 서로 독립이 아닌 자료

이미 앞에서 <그림 2>에 기대도수와 관측도수에 대한 막대그래프와 등고선 그림을 그려서 보인 적이 있다. 두 범주형 변수의 독립성을 탐색하기 위해 두 등고선 그림을 비교해 보면 현격한 차이를 나타내고 있다. 따라서 두 변수가 독립이라는 귀무가설을 기각할 수 있을 것으로 보인다. 실제로 χ^2-검정을 하면, χ^2-통계량은 138.3, 자유도는 9로 p-값은 0.0001 이므로 독립이라는 귀무가설은 유의수준 1%에서도 기각된다.

<표 7> 체신청별 주변합으로 정렬한 연도별 우편물 접수량 자료

체신청	91년	92년	93년	94년	합계
서울청	1537	1673	1764	1865	6839
부산청	303	330	347	367	1347
경북청	213	232	244	258	947
충청청	210	228	241	255	934
전남청	154	168	177	187	686
전북청	78	84	89	94	345
강원청	68	74	78	82	301
제주청	17	19	20	21	76
합계	2579	2807	2960	3129	11475



① 기대도수의 등고선 ② 관측도수 등고선 ③ 삭제된 잔차 등고선

<그림 15> 우편물 접수 자료의 등고선 그림

4.2 서로 독립인 자료

<표 7>은 김성주 (1999)에서 재인용한 우리나라 각 체신청별 연도별 우편물 접수량 자료를 다시 정렬한 자료이다. 이 자료는 순서척도자료 (연도)와 명목척도자료 (체신청)가 섞여 있으므로 적절한 정렬을 위해 명목척도인 체신청별 주변합으로 정렬하였다. 이 자료에 대해 등고선 그림을 통해 독립성을 탐색 해보자.

이 자료를 등고선으로 그리면 <그림 15>와 같다. 그림 ①과 ②의 기대도수와 관측도수의 등고선 그림을 보면 전체적으로 거의 흡사한 모양이다. 따라서 이 자료는 독립의 가정을 만족하는 것으로 판단할 수 있다. 또 그림 ③의 삭제된 잔차에 대한 등고선 그림에도 특별한 패턴이 없는 랜덤한 성질을 보여주고 있어 독립임을 보여준다. 한편 이상값으로 의심할 만한 값도 없는 것으로 식별된다. 실제로 통계량을 살펴보면 χ^2-통계량은 10.58982, χ^2-검정의 p-값은 0.8340685, 그리고 피어슨의 연관성 척도는 0.0303646으로 나타났다.

제 5 장 결 론

본 논문에서는 등고선 그림을 사용해 2차원 분할표를 나타내는 방법에 관해 살펴보았다. 무엇보다 일반적으로 좌표에 해당하는 자료가 연속형 자료일 때 사용되던 등고선 그림을 일정한 전제하에서 범주형 자료에서도 사용 할 수 있는 방법을 제시하였다.

등고선 그림은 그림을 그리기 위해서는 격자점에 해당하는 값만 있으면 된다. 다른 좌표들은 보

간에 의해서 결정되기 때문이다. 따라서 이산형 자료나 범주형 자료를 등고선 그림으로 나타낼 때는 보간되는 값들이 현실적인 의미가 없다는 점을 염두에 두고 사용해야 한다. 그런데 보간에 의해 표현한다는 사실은 동시에 자료의 형태에 구애를 받지 않고 그림으로 나타낼 수 있다는 장점을 갖기도 한다는 점에 착안하여 순서척도를 갖는 이산형 자료는 보간만으로, 명목척도를 갖는 범주형 자료의 경우는 간격을 동일하게 부여하여 등고선을 그림으로써 전체의 분포를 파악하는데 도움을 줄 수 있었다.

그러나 범주형 변수가 순서에 의미가 없는 명목척도를 가질 때는 가능한 범주 조합의 수는 범주의 수에 비례해 기하급수적으로 늘어나므로 전체의 분포를 파악하기 가장 좋은 범주 순서를 찾는 방법이 필요하다. 범주형 자료는 열과 행 단위로 자리를 교환해야 하므로 모든 값을 동시에 크기 순으로 정렬할 수 있는 해를 찾기가 쉽지 않다. 따라서 우리는 주변합으로 정렬함으로써 기대도수 크기 순으로 정렬하고, 이 순서로 관측도수를 정렬하는 방법을 제안하였다.

이렇게 범주형 자료를 등고선 그림으로 나타내면 다음과 같은 장점을 얻을 수 있었다. 첫째, 기대도수 등고선 그림과 관측도수 등고선 그림을 서로 비교함으로써 두 범주형 변수가 서로 독립인지에 대한 정보를 얻을 수 있었다. 이는 근본적으로 분할표의 독립성 검정에 사용하는 카이제곱 통계량의 정의와 같다. 둘째, 잔차의 등고선 그림을 추가함으로써 이상값을 탐색하는 데 사용할 수 있었다. 물론 이상값을 탐색하기 위해서는 삭제된 잔차를 이용하여 등고선을 그리는 것이 더욱 유리하다.

본 논문에서 나타나는 그림은 모두 R을 사용하여 작성하였다. 이 그림은 원래 모두 천연색으로 그려졌다. 이 그림의 원본은 다음 사이트에서 볼 수 있다.

http://stat.skku.ac.kr/~myhuh/research/contour.html

참고문헌

[1] 김성주 (1999). 국내 우편통계의 현황과 배달 우편물량에 대한 추정, 응용통계연구 제12권 2호, 315-323.

[2] Bates, D., Reams, F., Wahba, C. (1993). Getting better contour plots with S and GCVPACK , *Computational Statistics & Data Analysis* 15, 329-342.

[3] Hong, Chong Sun and Lee, Jong Cheol (2001). An Identification of Outlying Cells in Contingency Tables via Correspondence Analysis Map, *The Korean Communications in Statistics* 8, 39-49.

[4] Lee, Seung C. and Huh. Moon Y. (2003). A Measure of Association for Complex Data, *Computational Statistics and Data Analysis,* To be published

[5] Scott W. David (1992). *Multivariate Density Estimation : Theory, Practice and Visualization*, New York : Wiley, c1992.

[6] Simonoff, S. Jeffrey (1988). Detection Outlying Cells in Two-Way Contingency Tables Via Backwards-Stepping, *Technometrics* 30, 339-345.

[7] Snee, R. D. (1974). Graphical Display of Two-way Contingency Tables, *The American Statistician* 28, 9-12.

[8] The R Development Core Team (2003). *The R Environment for Statistical Computing and*

Graphics Refernce Index, Ver. 1.6.2.

[2003년 3월 접수, 2003년 6월 채택]