응용통계연구 제 **20**권 2호, *2007*년, *pp. 395–407*

일변량 및 이변량 자료에 대하여 특이값의 영향을 평가하기 위한 그래픽 방법

장대흥[1)]

요 약

통계자료분석 시 데이터에 특이값이 존재하면 이 특이값이 자료분석을 위한 여러 가지 측도들을 크게 왜곡시킨다. 우리는 특이값의 영향을 평가하기 위한 도구로서 간단한 그림도구인 민들레꽃씨그림을 제안할 수 있다. 평균-분산 민들레꽃씨그림은 자료 각각에 대하여 가중치를 1에서 0으로 변화시켜가며 구한 가중산술평균과 가중분산을 연결한 그림이고 공분산-상관계수 민들레꽃씨그림은 자료 각각에 대하여 가중치를 1에서 0으로 변화시켜가며 구한 가중공분산과 가중상관계수를 연결한 그림이다. 이러한 그림도구는 학부생들을 위한 기초통계학 교수 시 유용하게 쓰일 수 있다.

주요용어: 평균-분산 민들레꽃씨그림, 공분산-상관계수 민들레꽃씨그림.

1. 서론

통계자료분석 시 데이터에 특이값(이상값)이 존재하면 이 특이값이 자료분석을 위한 여러 가지 측도들을 크게 왜곡시킨다. 대부분의 기초통계학 교재에서는 특이값에 관한 언급이 있기는 하나 자료분석을 위한 여러 가지 측도들이 특이값에 얼마나 영향을 받는 지에 대해서는 설명이 미흡하다. '일변량 자료의 요약' 부분에서 산술평균과 분산을 구할 때 이 측도들이 특이값에 매우 민감하다는 언급과 '이변량 자료의 요약' 부분에서 표본상관계수를 구할 때 이 측도들이 특이값에 매우 민감하다는 언급은 있으나 이러한 측도들이 특이값에 얼마나 영향을 받는 지에 대해서는 설명이 미흡하다. 특이값의 영향을 측정하기 위한 도구로서 우리는 통상 두 가지 방법을 사용한다. 이 두 가지 방법은 Hampel(1974)이 제안한 '영향력함수(influence function)' 방법과 Cook(1986)이 제안한 '국소영향력(local influence)' 방법이다. 그러나 우리는 학부생들을 위한 기초통계학 교수 시 특이값의 영향을 측정하기 위한 도구로서 이 두 가지 방법들을 사용하는 것은 대단히 어려운 일이다. 그래서 우리는 특이값의 영향을 측정하기 위한 도구로서 간단한 그림도구를 제안할 수 있다. 이러한 방법들은 학부생들이 이해하기에 충분히 쉬운 도구들이다.

1) (608-737) 부산광역시 남구 대연3동 599-1 부경대학교 수리과학부 통계학전공, 교수
E-mail: dhjang@pknu.ac.kr

2. 일변량 자료에 대하여 특이값의 영향을 평가하기 위한 그래픽 방법

일변량 자료에 특이값이 존재하게 되면 대표값과 산포도에 관계되는 수치적 측도들이 영향을 받게 된다. 대표값에서는 산술평균이 대표적으로 특이값에 민감한 수치적 측도이고 산포도에서는 분산, 표준편차, 범위 등이 대표적으로 특이값에 민감한 수치적 측도들이다. 이러한 수치적 측도들이 특이값에 얼마나 민감한지를 나타내는 방법으로서 그래픽 방법을 사용한다면 학부생들을 위한 기초통계학 교수 시 많은 도움을 얻을 수 있을 것이다.

일변량 자료에 대하여 특이값의 영향을 평가하기 위하여 우리는 상자그림을 이용할 수 있다. 전체 n개의 자료에 대한 상자그림과 전체 n개의 자료에서 하나씩의 자료를 차례로 뺀 후 남은 $(n-1)$ 개의 자료들에 대한 상자그림들을 병렬상자그림(side-by-side box plot)으로 그려 보면 일변량 자료에 대하여 특이값의 영향을 평가하여 볼 수 있다. 이 때 상자그림들은 산술평균을 같이 나타내는 상자그림이어야 한다.

다음의 표 2.1은 渡部洋 등(1988)에 나타나는 일본 내의 어느 지진 관측소에서 측정한 연간유감지진 발생횟수(1961년에서 1981년까지)를 나타낸 표이다.

표 2.1: 일본 내의 어느 지진 관측소에서 측정한 연간유감지진 발생횟수

년	순서	지진횟수
'61	1	8
'62	2	11
'63	3	5
'64	4	7
'65	5	2
'66	6	3
'67	7	8
'68	8	8
'69	9	4
'70	10	3
'71	11	8
'72	12	272
'73	13	103
'74	14	48
'75	15	26
'76	16	8
'77	17	10
'78	18	16
'79	19	7
'80	20	34
'81	21	6

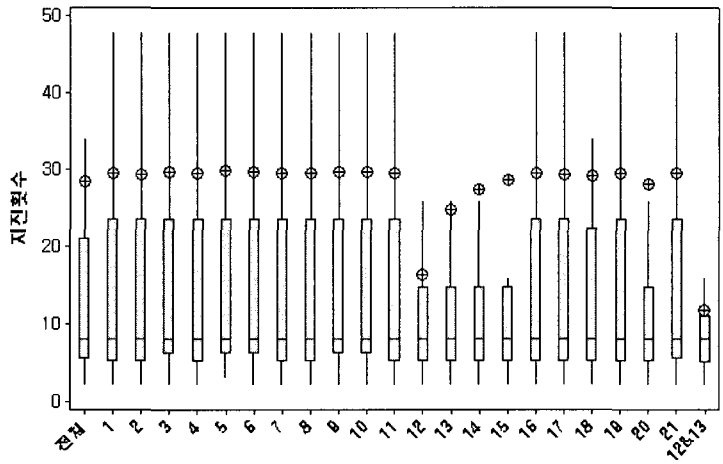

그림 2.1: 병렬상자그림 1

이 자료에 대하여 특이값의 영향을 평가하기 위하여 전체 21개의 자료에 대한 상자그림과 전체 21개의 자료에서 하나씩의 자료를 차례로 뺀 후 남은 20개의 자료들에 대한 상자그림들을 병렬상자그림(상자와 수염만 그린)으로 그려 보면 다음 그림 2.1과 같다. 그림에서 십자표시가 산술평균을 나타내고 가장 오른쪽 상자그림은 12번째(272, 제일 큰 특이값)와 13번째(103, 두 번째 큰 특이값) 두 개의 자료를 동시에 뺀 후 남은 19개 자료에 대한 상자그림이다. 12번째 자료를 뺀 경우의 상자그림에서 산술평균(십자표시)의 급격한 변화(산술평균이 28.40에서 16.25로 변함)를 볼 수 있다. 이를 통하여 12번째 자료가 가장 큰 영향을 주는 특이값임을 알 수 있다.

전체 21개의 자료에 대한 상자그림, 전체 21개의 자료에서 12번째의 자료를 뺀 후 남은 20개 자료에 대한 상자그림, 그리고 12번째와 13번째 두 개의 자료를 동시에 뺀 후 남은 19개 자료에 대한 상자그림을 병렬상자그림으로 그려 보면 다음 그림 2.2와 같다. 중앙값은 변화가 없으나(8로 동일함.) 산술평균에는 많은 변화(28.43→ 16.25→11.68)가 있음을 알 수 있다. 이를 통하여서도 12번째 자료가 가장 큰 영향을 주는 특이값임을 알 수 있다.

자료를 $x_1, x_2, \ldots, x_n$이라 할 때 가중산술평균과 가중분산을 우리는 다음과 같이 정의할 수 있다.

$$\bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}, s_{x_w}^2 = \frac{\sum_{i=1}^n w_i (x_i - \bar{x}_w)^2}{\sum_{i=1}^n w_i}. \tag{2.1}$$

제 7차 수학과교육과정에 보면 도수분포표에서의 평균의 정의 및 계산은 7학년(중학교 1학년)에서 가르치도록 되어 있고 도수분포표에서의 분산의 정의 및 계산은 10학년(고등학교 1학년)에서 가르치도록 되어 있다. 이러한 도수분포표에서의 평균과 분산의 개념은 위에서 정의한 가중평균과 가중분산과 같은 개념이다. 이러한 가중산술평균과 가중분산의 개

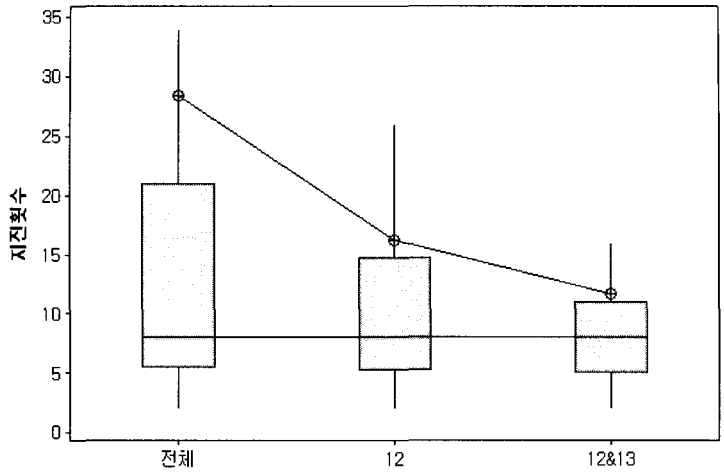

그림 2.2: 병렬상자그림 2

념은 중고등학교 때 배운 개념이므로 대학생들에게 생소한 정의가 아니다. n개의 자료 각각에 대하여 가중치 $w_i (i = 1, 2, \ldots, n)$를 1에서 0으로 변화시켜가며 가중산술평균과 가중분산을 구한 후 이렇게 구한 가중산술평균과 가중분산들을 연결하면 하나의 그림이 완성되는 데 이러한 그림을 평균-분산 민들레꽃씨그림(dandelion seed plot)이라 명하자. 모양이 민들레꽃씨처럼 생겨서 붙인 이름이다. 그림 2.3은 표 2.1의 자료에 대하여 그린 평균-분산 민들레꽃씨그림이다. 21개의 자료 각각에 대하여 가중치 $w_i (i = 1, 2, \ldots, n)$를 1에서 0으로 0.01씩 감소시키며 가중산술평균과 가중분산을 구한 후 이렇게 구한 가중산술평균과 가중분산들을 연결하였다. 평균-분산 민들레꽃씨그림에서 각각의 자료를 나타낼 때 자료의 인덱스를 표기할 수도 있고 자료의 값을 표기할 수도 있다. 그림 2.3은 자료의 인덱스를 표기한 평균-분산 민들레꽃씨그림이다.

그림 2.3을 보면 12번째 자료에 대응되는 연결곡선이 다른 연결곡선보다 가장 길게 뻗어 있음을 알 수 있다. 12번째 자료에서 가중산술평균과 가중분산이 극적인 변화(가중산술평균은 28.43에서 16.25로, 가중분산은 3464.34에서 522.89로 변함.)를 겪고 있음을 알 수 있다. 그러므로 12번째 자료가 가장 큰 영향을 주는 특이값임을 알 수 있다. 그림 2.1에서 12번째 자료를 제거하였을 때 산술평균의 급격한 변화를 보았다. 병렬상자그림에서는 12번째 자료를 포함할 때와 포함하지 않을 때 두 가지 경우를 비교할 수 있는 반면 평균-분산 민들레꽃씨그림에서는 12번째 자료를 서서히 빼내면서, 즉 가중치를 1에서 0으로 서서히 변화시켜가며 가중산술평균과 가중분산의 변화를 볼 수 있다.

평균-분산 민들레꽃씨그림에서의 가중산술평균과 가중분산의 변화량은 산술평균과 분산에 대한 표준화민감도곡선(standardized sensitivity curve, Maronna 등(2006) 참조.)값에 각각 비례한다. n개의 표본 $x_1, x_2, \ldots, x_n$이 주어져 있는 데 새로운 관측값 x_0가 이 표본에 합류하는 경우 표준화민감도곡선은 다음과 같이 정의되어진다.

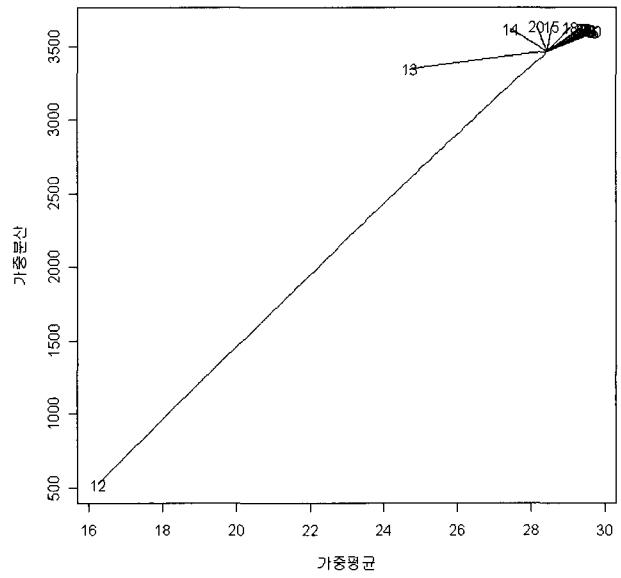

그림 2.3: 평균-분산 민들레꽃씨그림

$$SC_{mean}(x_0) = (n+1)(\bar{x}(x_1, x_2, \ldots, x_n, x_0) - \bar{x}(x_1, x_2, \ldots, x_n)),$$
$$SC_{variance}(x_0) = (n+1)(s^2(x_1, x_2, \ldots, x_n, x_0) - s^2(x_1, x_2, \ldots, x_n)).$$

여기서, $\bar{x}(x_1, x_2, \ldots, x_n, x_0)$는 $x_0, x_1, x_2, \ldots, x_n$을 이용하여 구한 산술평균이고 $\bar{x}(x_1, x_2, \ldots, x_n)$은 $x_1, x_2, \ldots, x_n$을 이용하여 구한 산술평균이다. 한편 $s^2(x_1, x_2, \ldots, x_n, x_0)$는 $x_0, x_1, x_2, \ldots, x_n$을 이용하여 구한 분산이고 $s^2(x_1, x_2, \ldots, x_n)$은 $x_1, x_2, \ldots, x_n$을 이용하여 구한 분산이다. 그러므로 가중산술평균과 가중분산의 변화량은 산술평균과 분산에 대한 표준화민감도곡선값에 각각 비례하게 되는 것이다. 영향력함수는 표준화민감도곡선의 극한개념이 된다.

다음 표 2.2는 渡部洋 등(1988)에 나타나는 일본 내의 21개 지진 관측소에서 측정한 연간유감지진 발생횟수(1961년에서 1981년까지)에 대한 자료 중 3개 지진 관측소에서 측정한 연간유감지진 발생횟수(1961년에서 1981년까지)에 대한 자료를 나타내는 표이다. 이 자료에 대한 민들레꽃씨그림은 그림 2.4와 같다. 그림 2.4에서는 3개 지진관측소를 서로 비교하는 것이 중요하므로 자료의 값(최대값)을 표기한 평균-분산 민들레꽃씨그림을 그렸다.

A와 B 두 지진관측소들이 영향이 큰 특이값들(A는 112, B는 130)을 갖고 있으나 C 지진과측소는 영향이 큰 특이값이 없음을 알 수 있다.

표 2.2: 일본 내의 3개 지진관측소에서 측정한 연간유감지진 발생횟수

년	A	B	C
'61	20	72	70
'62	17	52	63
'63	10	45	41
'64	10	40	69
'65	2	40	79
'66	37	33	69
'67	10	50	64
'68	75	93	59
'69	82	86	78
'70	112	32	53
'71	65	38	57
'72	72	44	60
'73	48	130	61
'74	61	49	68
'75	18	49	56
'76	18	36	56
'77	37	38	49
'78	21	49	72
'79	13	50	54
'80	25	43	73
'81	13	35	47

3. 이변량 자료에 대하여 특이값의 영향을 평가하기 위한 그래픽 방법

이변량 자료를 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$이라 할 때 가중산술평균, 가중분산, 가중공분산, 가중상관계수들을 우리는 다음과 같이 정의할 수 있다.

$$\bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}, s_{x_w}^2 = \frac{\sum_{i=1}^n w_i (x_i - \bar{x}_w)^2}{\sum_{i=1}^n w_i},$$

$$\bar{y}_w = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i}, s_{y_w}^2 = \frac{\sum_{i=1}^n w_i (y_i - \bar{y}_w)^2}{\sum_{i=1}^n w_i}, \tag{3.1}$$

$$cov_w(x,y) = \frac{\sum_{i=1}^n w_i (x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\sum_{i=1}^n w_i}, corr_w(x,y) = \frac{cov_w(x,y)}{\sqrt{s_{x_w}^2}\sqrt{s_{y_w}^2}}.$$

n개의 이변량 자료 각각에 대하여 가중치 $w_i (i = 1, 2, \ldots, n)$를 1에서 0으로 변화시켜가며 가중산술평균, 가중분산, 가중공분산, 가중상관계수들을 구한 후 이렇게 구한 가중산술

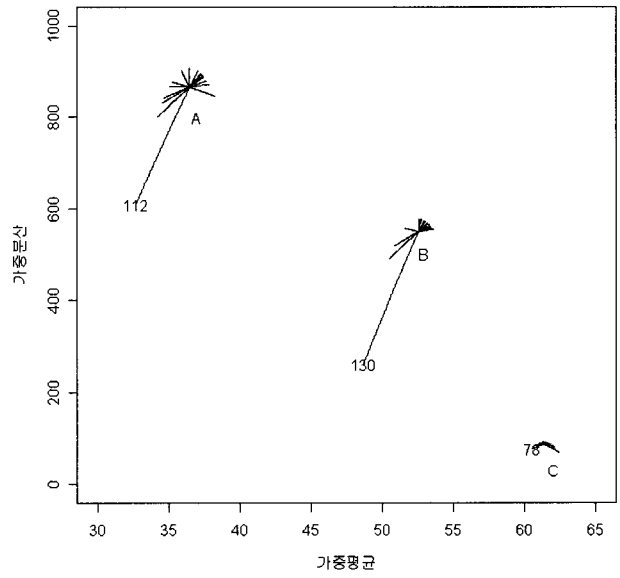

그림 2.4: 3개 지진관측소 자료에 대한 평균-분산 민들레꽃씨그림

평균과 가중분산들을 연결하여 각각 변수 X와 Y에 대한 평균-분산 민들레꽃씨그림을 작성하고 가중공분산과 가중상관계수들을 연결하여 공분산-상관계수 민들레꽃씨그림을 그릴 수 있다.

다음 표 3.1은 Maronna 등(2006)에 나타나는 생화학 자료이다.

이 자료에 대하여 평균-분산 민들레꽃씨그림을 그리면 다음 그림 3.1과 그림 3.2와 같고 공분산-상관계수 민들레꽃씨그림을 그리면 그림 3.3과 같다. 변수 X(인산염) 12개의 자료 각각에 대하여 가중치 $w_i (i = 1, 2, \ldots, n)$를 1에서 0으로 0.01씩 감소시키며 가중산술평균과 가중분산을 구한 후 이렇게 구한 가중산술평균과 가중분산들을 연결하여 그림 3.1을 완성하였고, 변수 Y(염화물) 12개의 자료 각각에 대하여 가중치 $w_i (i = 1, 2, \ldots, n)$를 1에서 0으로 0.01씩 감소시키며 가중산술평균과 가중분산을 구한 후 이렇게 구한 가중산술평균과 가중분산들을 연결하여 그림 3.2를 완성하였다. 12개의 자료 각각에 대하여 가중치 $w_i (i = 1, 2, \ldots, n)$를 1에서 0으로 0.01씩 감소시키며 가중공분산과 가중상관계수를 구한 후 이렇게 구한 가중공분산과 가중상관계수들을 연결하여 그림 3.3을 완성하였다.

3번째 자료에 있어서 변수 X에 대한 가중산술평균과 가중분산 그리고 가중공분산과 가중상관계수에서 상대적으로 변화(변수 X에 대한 가중산술평균은 1.788에서 1.868로, 변수 X에 대한 가중분산은 0.235에서 0.178로, 가중공분산은 -0.439에서 -0.625로, 가중상관계수는 -0.494에서 -0.805로 변함.)가 있음을 알 수 있다. 특히 가중상관계수에서의 변화가 심함을 알 수 있다. 반면 변수 Y에 대한 가중산술평균과 가중분산에서는 상대적으로 큰 변

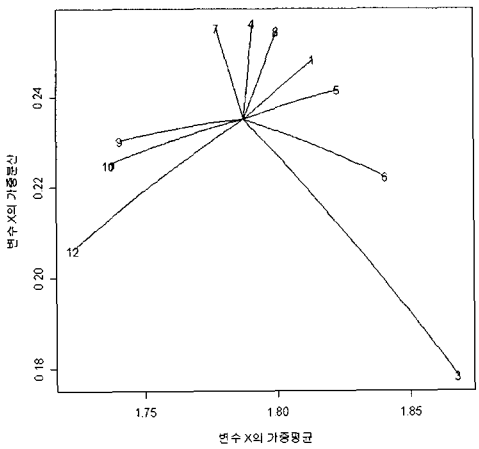

그림 3.1: 변수 X에 대한 평균-분산 민들레꽃씨그림

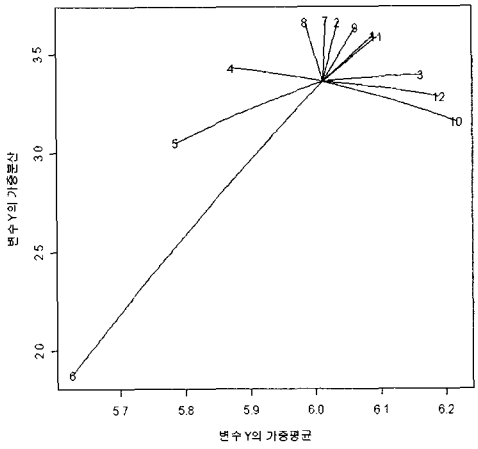

그림 3.2: 변수 Y에 대한 평균-분산 민들레꽃씨그림

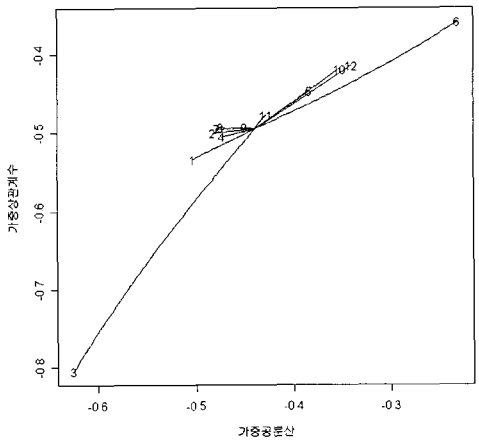

그림 3.3: 공분산-상관계수 민들레꽃씨그림

표 3.1: 생화학 자료

관측치	인산염	염화물
1	1.50	5.15
2	1.65	5.75
3	0.90	4.35
4	1.75	7.55
5	1.40	8.50
6	1.20	10.25
7	1.90	5.95
8	1.65	6.30
9	2.30	5.45
10	2.35	3.75
11	2.35	5.10
12	2.50	4.05

화가 없음(변수 Y에 대한 가중산술평균은 6.013에서 6.164로, 변수 Y에 대한 가중분산은 3.358에서 3.389로 변함.)을 알 수 있다. 그림 3.1, 3.2, 3.3에서 3번째 자료에 대응되는 연결곡선을 보면 이러한 사실들을 확인할 수 있다. 그림 3.1, 3.2, 3.3을 종합하면 3번째 자료가 가장 큰 영향을 주는 특이값임을 알 수 있다. 특이값이라 할 수 있는 3번째 자료에 비하여 6번째 자료는 변수 Y에 대한 가중산술평균과 가중분산 그리고 가중공분산에서는 상대적으로 변화(변수 Y에 대한 가중산술평균은 6.013에서 5.627로, 변수 Y에 대한 가중분산은 3.358에서 1.882로, 가중공분산은 -0.439에서 -0.232로 변함.)가 있으나 변수 X에 대한 가중산술평균과 가중분산 그리고 가중상관계수에서 상대적으로 큰 변화가 없음(변수 X에 대한 가중산술평균은 1.788에서 1.841로, 변수 X에 대한 가중분산은 0.235에서 0.222로, 가중상관계수는 -0.494에서 -0.359로 변함.)을 알 수 있다. 그림 3.1, 3.2, 3.3에서 6번째 자료에 대응되는 연결곡선을 보면 이러한 사실들을 확인할 수 있다. 그림 3.1, 3.2, 3.3을 종합하면 6번째 자료가 특이값이라 보기가 어렵다.

4. 결론

본 논문에서는 특이값의 영향을 측정하기 위한 도구로서 간단한 그림도구인 평균-분산 민들레꽃씨그림과 공분산-상관계수 민들레꽃씨그림을 제안하였다. 이러한 방법은 학부생들을 위한 기초통계학 교수 시 유용하게 쓰일 수 있다.

부록 A: 평균-분산 민들레꽃씨그림과 공분산-상관계수 민들레꽃씨그림

본문에 있는 그림들은 R 패키지로 작성한 것들이다. 다음 두 개의 프로그램은 평균-분

```
# 데이터 입력
list1=c(8,11,5,7,2,3,8,8,4,3,8,272,103,48,26,8,10,16,7,34,6)
n=length(list1)
m=101
mat1=c(rep(0,m*n))
wmean=matrix(mat1,nrow=n)
wvariance=matrix(mat1,nrow=n)
# 가중평균과 가중분산의 계산
k=1
for(i in 1:n)
{
for(j in seq(1,0,by=-0.01))
{
w=c(rep(1,n))
w[i]=j
wmean[i,k]=sum(w*list1)/sum(w)
wvariance[i,k]=sum(w*(list1-wmean[i,k])^2)/sum(w)
k=k+1
}
k=1
}
x.min=min(wmean)
x.max=max(wmean)
y.min=min(wvariance)
y.max=min(wvariance)
# 평균-분산 민들레꽃씨 그림(자료의 인덱스를 표시함.)
for(i in 1:n)
{
plot(wmean[i,],wvariance[i,],type="l",xlim=c(x.min,x.max),ylim=c(y.min,y.max),
xlab="가중평균",ylab="가중분산",main="민들레꽃씨 그림(자료의 인덱스를 표시함.)")
text(wmean[i,m], wvariance[i,m],i)
par(new=F)
}
par(new=F)
#평균-분산 민들레꽃씨 그림(자료의 값을 표시함.)
for(i in 1:n)
{
plot(wmean[i,],wvariance[i,],type="l",xlim=c(x.min,x.max),ylim=c(y.min,y.max),
xlab="가중평균",ylab="가중분산",main="민들레꽃씨 그림(자료의 값을 표시함.)")
text(wmean[i,m], wvariance[i,m],list1[i])
par(new=F)
}
par(new=F)
```

그림 A.1: 평균-분산 민들레꽃씨그림

산 민들레꽃씨그림과 공분산-상관계수 민들레꽃씨그림을 본문에서 사용한 자료들(표 2.1과 표3.1)을 사용하여 R로 짠 프로그램이다. A.1에서는 자료의 인덱스를 표기한 평균-분산 민들레꽃씨그림과 자료의 값을 표기한 평균-분산 민들레꽃씨그림 두 종류의 평균-분산 민들레꽃씨그림을 제시하였다. A.2에서는 자료의 인덱스를 표기한 공분산-상관계수 민들레꽃씨그림을 제시하였다.

```
# 데이터 입력
list1=c(1.50, 1.65, .90, 1.75, 1.40, 1.20, 1.90, 1.65, 2.30, 2.35, 2.35, 2.50)
list2=c(5.15, 5.75, 4.35, 7.55, 8.50, 10.25, 5.95, 6.30, 5.45, 3.75, 5.10, 4.05)
n=length(list1);m=101
mat1=c(rep(0,n*m))
wmean1=matrix(mat1,nrow=n);wvariance1=matrix(mat1, nrow=n)
wmean2=matrix(mat1,nrow=n);wvariance2=matrix(mat1, nrow=n)
wcovariance=matrix(mat1,nrow=n);wcorr.coeff=matrix(mat1,nrow=n)
# 가중평균, 가중분산, 가중공분산, 가중상관계수의 계산
k=1
for(i in 1:n)
{
for(j in seq(1.0,by=-0.01))
{
w=c(rep(1,n))
w[i]=j
wmean1[i,k]=sum(w*list1)/sum(w)
wvariance1[i,k]=sum(w*(list1-wmean1[i,k])^2)/sum(w)
wmean2[i,k]=sum(w*list2)/sum(w)
wvariance2[i,k]=sum(w*(list2-wmean2[i,k])^2)/sum(w)
wcovariance[i,k]=sum(w*(list1-wmean1[i,k])*(list2-wmean2[i,k]))/sum(w)
wcorr.coeff[i,k]=sum(w*(list1-wmean1[i,k])*(list2-wmean2[i,k]))/
(sqrt(sum(w*(list1-wmean1[i,k])^2))*sqrt(sum(w*(list2-wmean2[i,k])^2)))
k=k+1
}
k=1
}
x1.min=min(wmean1);x1.max=max(wmean1);y1.min=min(wvariance1);y1.max=max(wvariance1)
x2.min=min(wmean2);x2.max=max(wmean2);y2.min=min(wvariance2);y2.max=max(wvariance2)
x12.min=min(wcovariance);x12.max=max(wcovariance);y12.min=min(wcovariance);y12.max=max(wcovariance)
# 변수 X에 대한 평균-분산 민들레꽃씨 그림(자료의 인덱스를 표시함.)
for(i in 1:n)
{
plot(wmean1[i,],wvariance1[i,],type="l",xlim=c(x1.min,x1.max),ylim=c(y1.min,y1.max),
xlab="가중평균",ylab="가중분산",main="변수 X에 대한 평균-분산 민들레꽃씨 그림(자료의 인덱스를 표시함.)")
text(wmean1[i,m],wvariance1[i,m],i)
par(new=T)
}
par(new=F)
# 변수 Y에 대한 평균-분산 민들레꽃씨 그림(자료의 인덱스를 표시함.)
for(i in 1:n)
{
plot(wmean2[i,],wvariance2[i,],type="l",xlim=c(x2.min,x2.max),ylim=c(y2.min,y2.max),
xlab="가중평균",ylab="가중분산",main="변수 X에 대한 평균-분산 민들레꽃씨 그림(자료의 인덱스를 표시함.)")
text(wmean2[i,m],wvariance2[i,m],i)
par(new=T)
}
par(new=F)
# 공분산-상관계수 민들레꽃씨 그림(자료의 인덱스를 표시함.)
for(i in 1:n)
{
plot(wcovariance[i,],wcorr.coeff[i,],type="l",xlim=c(x12.min,x12.max),ylim=c(y12.min,y12.max),
xlab="가중공분산",ylab="가중상관계수",main="공분산-상관계수 민들레꽃씨 그림(자료의 인덱스를 표시함.)")
text(wcovariance[i,m],wcorr.coeff[i,m],i)
par(new=T)
}
par(new=F)
```

그림 A.2: 공분산-상관계수 민들레꽃씨그림

참고문헌

渡部洋, 鈴木規夫, 山田文康, 大塚雄作(1988). *探索的テ-タ解析入門*, 朝倉書店.

Cook, R. D.(1986). Assessment of local influence, *Journal of Royal Statistical Society*, Ser. B, **48**, 133–169.

Hampel, F. R.(1974). The influence curve and its role in robustness, *The Annal of Statistics*, **45**, 383–393.

Maronna, R. A., Martin, D. and Yohai, V. J. (2006). *Robust Statistics*, John Wiley & Sons, New York.

[2007년 1월 접수, 2007년 2월 채택]

A Graphical Method for Evaluating the Effect of Outliers in One- and Two-Variate Data

Dae-Heung Jang[1)]

ABSTRACT

Outliers distort many measures for data analysis. We can propose dandelion seed plot as a graphical tool for evaluating the effect of outliers in one- and two-variate data. We can draw mean-variance dandelion seed plots using linked curves which are made by changing weights from 1 to 0 for each datum. Similarly we can also draw covariance-correlation-coefficient dandelion seed plots. This graphical method can be a useful tool for elementary statistics education in college.

Keywords: Mean-variance dandelion seed plot, covariance-correlation-coefficient dandelion seed plot.

1) Professor, Division of Mathematical Sciences, Pukyong National University, 599-1, Daeyeon-dong, Nam-gu, Busan 608-737, Korea
E-mail: dhjang@pknu.ac.kr