Journal of the Korean Data & Information Science Society
2017, **28**(5), 959–970

http://dx.doi.org/10.7465/jkdi.2017.28.5.959
한국데이터정보과학회지

이변량 왜도, 첨도 그리고 표면그림

홍종선[1] · 성재현 [2]

[12]성균관대학교 통계학과

접수 2017년 8월 10일, 수정 2017년 9월 15일, 게재확정 2017년 9월 18일

Abstract

본 연구에서는 두 변수의 상관계수를 반영한 이변량 자료의 왜도와 첨도 통계량을 제안하고, 시각적으로 표현할 수 있는 표면그림을 개발한다. 이변량 왜도 통계량은 이변량 확률표본 자료의 치우침 방향과 정도를 표현하는 실수 한 쌍으로 정의한다. 첨도는 양의 값을 가지며 이변량 정규분포를 기준으로 꼬리 부분의 두터운 정도를 파악할 수 있다. 그리고 표면그림은 분위벡터를 바탕으로 평면에 구현한다. 다양한 형태의 이변량 자료를 생성하여 표면그림을 작성하고 왜도와 첨도를 계산하여 탐색해 본 결과, 왜도와 첨도 값들은 표면그림으로 구현한 이변량 자료의 특징을 잘 반영하는 것을 발견하였다. 그러므로 본 논문에서 제안한 왜도, 첨도 그리고 표면그림은 이변량 분포를 분석하는 기술통계학적 방법으로 활용할 수 있다.

주요용어: 마할라노비스 거리, 분위벡터, 상자그림, 표면그림, 혼합분포.

1. 서론

일변량 자료에 대한 기술통계량들 (descriptive statistics)중에서 3차 중심적률 형태로 정의되는 왜도 (skewness)는 대칭분포의 경우에 0이며 분포의 오른쪽 꼬리가 왼쪽보다 길게 뻗어 있는 경우는 양수값을, 반대로 왼쪽 꼬리가 오른쪽보다 길게 뻗어 있는 경우는 음수값을 가진다. 그리고 첨도 (kurtosis) 통계량은 4차 중심적률의 형태로 정의되며 정규분포의 첨도를 0으로 설정하면, 정규분포보다 꼬리가 두터운 분포의 첨도는 양수가 되며 반대로 정규분포보다 짧은 꼬리를 갖는 분포의 첨도는 음수가 된다 (Bickel과 Doksum, 2001; Hogg 등 2013; Lehmann과 Casella, 1998; Lindgren, 1993; Mood 등, 1974; Rohatgi, 1976).

확률벡터 $\underline{X} = (X_1, X_2, \ldots, X_k)^t$는 평균벡터 $\underline{\mu} = (\mu_1, \mu_2, \ldots, \mu_k)^t$와 분산-공분산행렬 Σ을 갖는 분포를 따르며, $\underline{Y}$는 $\underline{X}$와 독립이고 $\underline{X}$와 동일한 분포를 따른다고 할 때 Mardia (1970)는 이변량 이상의 다변량에서 분포의 형태를 측정할 수 있는 다변량 왜도와 첨도를 다음과 같이 각각 제안하였다 (Seber, 1984).

$$\text{왜도} = E[\{(\underline{X}-\underline{\mu})^t\Sigma^{-1}(\underline{Y}-\underline{\mu})\}^3],$$
$$\text{첨도} = E[\{(\underline{X}-\underline{\mu})^t\Sigma^{-1}(\underline{X}-\underline{\mu})\}^2].$$

이변량 분포가 정규분포를 따른다면, 상관계수에 관계없이 Mardia가 제안한 왜도와 첨도는 각각 0과 8이다. 그런데 Mardia의 왜도는 공간상의 분포의 치우침 또는 대칭성을 측정하는 통계량임에도 불구

[1] 교신저자: (03063) 서울 종로구 성균관로 25-2, 성균관대학교 통계학과, 교수. E-mail: cshong@skku.edu
[2] (03063) 서울 종로구 성균관로 25-2, 성균관대학교 통계학과, 대학원생.

하고 항상 0이상의 값만을 가짐으로써 치우침이나 대칭성의 정도만을 측정해줄 뿐 치우침의 방향을 보여주지는 못한다. 또한 이변량 정규분포에서 두 변수의 상관계수가 $+1$ 또는 -1로 접근함에 따라 관측값들이 일변량 정규분포의 형태로 접근함에도 불구하고 Marida 첨도는 항상 동일한 8값을 가진다. 즉 Mardia가 제안한 왜도와 첨도는 좋은 성질을 가짐에도 불구하고 변수의 상관관계에 따른 분포의 변화를 반영하지 못하는 문제점이 존재한다. 그러므로 본 연구에서는 두 변수 사이의 상관관계를 고려한 새로운 왜도와 첨도 통계량을 제안하며 이는 이변량 결합분포에서의 치우침의 방향과 정도에 관한 정보를 제공함으로써 탐색적 자료 분석을 위한 기술통계학적 방법으로 활용될 수 있다.

일변량 자료를 시각적으로 표현하는 여러 방법 중에서 상자그림 (box plot)은 왜도와 첨도에 대응하는 자료의 특성을 잘 구현하는 특징이 있다 (Hong과 Kim, 2016). 그러나 상자그림은 이변량 이상의 자료를 표현할 수 없기 때문에 본 연구에서는 이변량 자료를 시각적으로 구현할 수 있는 표면그림 (surface plot)을 제안한다. 상자그림은 분위수 (quantile)를 이용하지만, 표면그림은 Hong 등 (2016), Hong과 Lee (2016), 그리고 Hong과 Kim (2017)이 제안한 분위벡터 (quantile vector)를 바탕으로 구현한다.

본 연구의 구성은 다음과 같다. 2절에서는 이변량 확률표본에 대한 왜도 통계량을 제안한다. 이 통계량은 두 변수의 치우침 방향을 표현해야 하기 때문에 한 쌍의 실수값으로 표현한다. 다양한 정규혼합분포로부터 확률표본을 생성하고 왜도값을 살펴보면서 자료의 치우침 방향과 정도를 탐색한다. 3절에서는 이변량 첨도 통계량을 제안한다. 두 변수의 상관관계가 없는 이변량 표준 정규분포에서의 첨도는 1의 값을 가지며, 1보다 크거나 작은 첨도값을 갖는 다양한 정규혼합분포로부터 확률표본을 생성하여 첨도값의 변화를 살펴본다. 그리고 4절에서는 이변량 자료를 시각적으로 구현할 수 있는 표면그림을 제안한다. 왜도가 음수 또는 양수인 경우와 첨도가 1보다 크거나 작은 경우에 대응하는 확률표본을 표면그림으로 구현함으로써 본 연구에서 제안한 왜도와 첨도 통계량과 표면그림을 이용한 기술통계 분석에서의 일치성에 대해 토론한다. 마지막으로 5절에서는 결론을 유도하며 향후 연구 과제로 삼변량 이상 자료에 대한 다변량 왜도와 첨도 및 삼변량 표면그림에 대하여 언급한다.

2. 이변량 왜도

이변량 자료에서 치우침의 정도와 방향을 측정하기 위하여 실수쌍으로 표현되는 새로운 왜도를 제안한다. 각각의 부호와 절대값을 통해서 자료가 치우친 방향과 비대칭성의 정도를 동시에 파악할 수 있는 대안적인 왜도를 다음과 같이 3차 중심 적률 형태로 제안한다.

정의 2.1 이변량 확률변수 X와 Y의 표준화 변수 Z_X와 Z_Y을 다음과 같이 3차 중심적률의 형태로 표현한 통계량을 이변량 왜도로 정의한다.

$$\text{왜도} = \left(\frac{E[Z_X Z_Y^2]}{|\Sigma|^{3/2}}, \frac{E[Z_X^2 Z_Y]}{|\Sigma|^{3/2}} \right),$$

여기서 $Z_X = (X - E[X])/\sqrt{Var(X)}$, $Z_Y = (Y - E[Y])/\sqrt{Var(Y)}$이며 Σ는 Z_X와 Z_Y의 분산-공분산행렬.

일변량에서 왜도는 3차 중심적률 형태이므로 이변량에서도 3차 중심적률을 고려하되, 비대칭성의 방향을 반영하기 위해 $Z_X Z_Y^2$와 $Z_X^2 Z_Y$의 형태를 사용한다. 정의 2.1에서의 왜도의 첫번째 성분은 Z_X에 의해서 그리고 두 번째 성분은 Z_Y에 의하여 부호가 결정된다. 치우침 또는 비대칭성의 정도는 $Z_X Z_Y^2$와 $Z_X^2 Z_Y$의 절대값의 크기에 각각 의존한다. 그리고 일반적으로 많이 사용하는 Z_X, Z_Y의 마할라노비스 거리 (Mahalanobis distance)를 새로 정의된 이변량 왜도에 반영한다. 표준화 확률벡터

$\underline{Z} = (Z_X, Z_Y)^t$의 마할라노비스 거리 $\sqrt{\underline{Z}^t \Sigma^{-1} \underline{Z}}$에서 역행렬 Σ^{-1}를 고려하듯이 분산-공분산행렬의 행렬식 $|\Sigma|$의 3/2제곱을 $Z_X Z_Y^2$와 $Z_X^2 Z_Y$의 기대값에 나누어줌으로써 정의 2.1과 같이 제안한다.

또한 n개의 이변량 확률표본 $\{(X_i, Y_i); i = 1, \ldots, n\}$에 대한 왜도는 다음과 같이 계산할 수 있다.

$$\left(\frac{\sum_{i=i}^{n} Z_{X_i} Z_{Y_i}^2}{n(1-\hat{\rho}^2)^{3/2}}, \frac{\sum_{i=i}^{n} Z_{X_i}^2 Z_{Y_i}}{n(1-\hat{\rho}^2)^{3/2}} \right),$$

여기서 $Z_{X_i} = (X_i - \bar{X})/sd(X)$, $Z_{Y_i} = (Y_i - \bar{Y})/sd(Y)$, $\hat{\rho} = \widehat{corr(X,Y)}$이다. 표본왜도에서 $\hat{\rho}$의 절대값이 0에서 1로 갈수록 분포는 원형에서 가느다란 타원 형태로 변하기 때문에 치우침의 정도가 강해진다고 볼 수 있으므로 이변량 왜도값이 증가하는 것은 타당하다.

이변량 정규분포에서의 이변량 왜도는 상관계수에 의존하지 않으며 (0, 0)값을 가진다. 즉 일변량 정규분포의 왜도와 동일하게 이변량 왜도의 성분이 0이므로 비대칭성은 존재하지 않는다고 판단할 수 있다. 비대칭성이 존재할 때 이변량 왜도값의 변화를 탐색하기 위해 다음과 같은 이변량 확률변수 $\underline{X}=(X,Y)^t$의 정규혼합분포인 Distribution 2.1을 고려한다.

$$\text{Distribution 2.1} : \underline{X} \sim \lambda\, N(\underline{\mu_1}, \Sigma_1) + (1-\lambda)\, N(-\underline{\mu_1}, \Sigma_1),$$

여기서 $\underline{\mu_1} = \begin{pmatrix} \mu_{1x} \\ \mu_{1y} \end{pmatrix}$, $\Sigma_1 = \begin{bmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{bmatrix}$, $\lambda \in (0,1)$.

일변량 혼합분포의 평균과 분산을 구한 Hogg 등 (2013)의 방법을 확장하여 이변량 정규혼합분포의 평균벡터 $\underline{\mu}$와 분산-공분산행렬 Σ 그리고 상관계수 ρ는 각각 다음과 같이 구한다.

$$\underline{\mu} = (2\lambda - 1)\underline{\mu_1},$$

$$\Sigma = \begin{bmatrix} 1 + (\mu_{1x} + \mu_x)^2 - 4\lambda\mu_{1x}\mu_x & \rho_1 + (\mu_{1x} + \mu_x)(\mu_{1y} + \mu_y) - 2\lambda(\mu_{1x}\mu_y + \mu_{1y}\mu_x) \\ \rho_1 + (\mu_{1x} + \mu_x)(\mu_{1y} + \mu_y) - 2\lambda(\mu_{1x}\mu_y + \mu_{1y}\mu_x) & 1 + (\mu_{1y} + \mu_y)^2 - 4\lambda\mu_{1y}\mu_y \end{bmatrix},$$

$$\rho = \frac{\rho_1 + (\mu_{1x} + \mu_x)(\mu_{1y} + \mu_y) - 2\lambda(\mu_{1x}\mu_y + \mu_{1y}\mu_x)}{\sqrt{1 + (\mu_{1x} + \mu_x)^2 - 4\lambda\mu_{1x}\mu_x}\sqrt{1 + (\mu_{1y} + \mu_y)^2 - 4\lambda\mu_{1y}\mu_y}}.$$

λ에 따른 혼합분포의 치움침을 조절하기 위하여 λ의 극단적인 값인 0.1과 0.4의 경우를 제외하여 λ를 0.3으로 설정한다. 이변량 정규혼합분포에서 $\lambda = 0.3$일 때 중심이 (0,0)이며 반지름이 $\sqrt{2}$인 원의 궤적에 있는 점의 좌표에서 $\underline{\mu_1}$를 설정하여 Distribution 2.1로부터 12종류의 Model을 설정하고 $\rho_1 = -0.6$, 0, 0.6일때 각 Model로부터 1,000개의 확률표본을 추출하여 왜도를 계산하는 모의실험을 500회 반복하였다. 이렇게 구한 이변량 왜도를 Table 2.1의 'Skewness'열에 나타내었고 Figure 2.1에 구현하였다. 또한 Mardia 왜도를 Table 2.1의 'Mardia'열에 나타내었다.

Mardia 왜도는 스칼라 형태이기 때문에 분포의 치우침 방향을 확인하기 어렵지만 이변량 왜도는 벡터 형태여서 치우침의 방향을 확인할 수 있다. 예를 들어 Table 2.1의 Model 2인 경우 ($\underline{\mu_1} = (1,1)^t$) 혼합분포의 평균좌표는 $(-0.4, -0.4)$이다. 평균좌표를 기준으로 확률표본을 분류해보면 X와 Y는 모두 평균을 기준으로 우측으로 치우쳐있다 (skewed to the right). 따라서 이변량 왜도의 두 성분 모두 양의 값을 가지며 $\rho_1 = -0.6, 0, 0.6$일 때 왜도값은 각각 (0.28, 0.28), (0.38, 0.38), (1.12, 1.12)이다. 반면 Mardia 왜도는 $\rho_1 = -0.6, 0, 0.6$에서 0.43, 0.23, 0.13으로 치우침의 방향을 확인할 수 없다. 다른 예로 Model 5 ($\underline{\mu_1} = (-1,1)^t$)의 경우 혼합분포의 평균좌표는 $(0.4, -0.4)$이다. 평균좌표를 기준으로 확률표본을 분류해보면 X는 평균을 기준으로 좌측으로 치우쳐있으며 (skewed to the left), Y는 평균을 기준으로 우측으로 치우쳐있다 (skewed to the right). 따라서 왜도의 첫 번째 성분은 음수, 두 번째 성분은 양수이며 $\rho_1 = -0.6, 0, 0.6$일 때 왜도값은 (-1.12, 1.12), (-0.38, 0.38), (-0.28, 0.28)이다. 반면

Table 2.1 Bivariate skewness of various model

$\rho_1 = -0.6$							
Model	$\underline{\mu_1}$	Skewness	Mardia	Model	$\underline{\mu_1}$	Skewness	Mardia
1	(1.3,0.55)	(0.14, 0.23)	0.42	7	(-1.3, -0.55)	(-0.14,-0.23)	0.42
2	(1,1)	(0.28, 0.28)	0.43	8	(-1, -1)	(-0.28,-0.28)	0.43
3	(0.6,1.28)	(0.24, 0.15)	0.43	9	(-0.6, -1.28)	(-0.24,-0.15)	0.41
4	(-0.6,1.28)	(-0.7, 0.44)	0.20	10	(0.6, -1.28)	(0.7,-0.44)	0.20
5	(-1,1)	(-1.12, 1.12)	0.15	11	(1, -1)	(1.12,-1.12)	0.15
6	(-1.3,0.55)	(-0.36, 0.61)	0.19	12	(1.3, -0.55)	(0.36,-0.61)	0.20
$\rho_1 = 0$							
Model	$\underline{\mu_1}$	Skewness	Mardia	Model	$\underline{\mu_1}$	Skewness	Mardia
1	(1.3,0.55)	(0.16, 0.28)	0.24	7	(-1.3, -0.55)	(-0.16,-0.28)	0.22
2	(1,1)	(0.38, 0.38)	0.23	8	(-1, -1)	(-0.38,-0.38)	0.21
3	(0.6,1.28)	(0.3, 0.19)	0.23	9	(-0.6, -1.28)	(-0.3,-0.19)	0.22
4	(-0.6,1.28)	(-0.3, 0.19)	0.24	10	(0.6, -1.28)	(0.3,-0.19)	0.20
5	(-1,1)	(-0.38, 0.38)	0.22	11	(1, -1)	(0.38,-0.38)	0.22
6	(-1.3,0.55)	(-0.16, 0.28)	0.22	12	(1.3, -0.55)	(0.16,-0.28)	0.23
$\rho_1 = 0.6$							
Model	$\underline{\mu_1}$	Skewness	Mardia	Model	$\underline{\mu_1}$	Skewness	Mardia
1	(1.3,0.55)	(0.36, 0.61)	0.19	7	(-1.3, -0.55)	(-0.36,-0.61)	0.20
2	(1,1)	(1.12, 1.12)	0.13	8	(-1, -1)	(-1.12,-1.12)	0.13
3	(0.6,1.28)	(0.7, 0.44)	0.19	9	(-0.6, -1.28)	(-0.7,-0.44)	0.20
4	(-0.6,1.28)	(-0.24, 0.15)	0.42	10	(0.6, -1.28)	(0.24,-0.15)	0.42
5	(-1,1)	(-0.28, 0.28)	0.40	11	(1, -1)	(0.28,-0.28)	0.43
6	(-1.3,0.55)	(-0.14, 0.23)	0.38	12	(1.3, -0.55)	(0.14,-0.23)	0.42

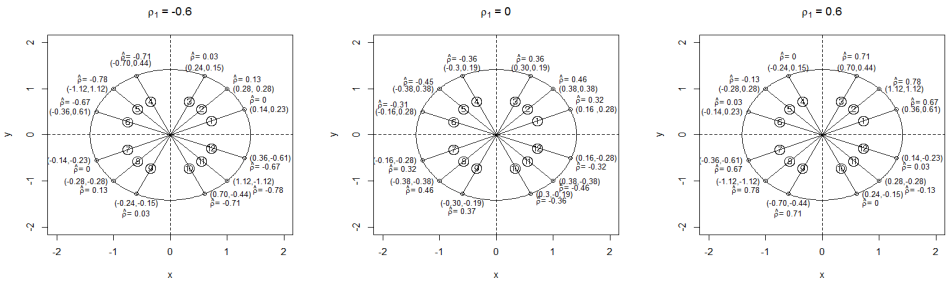


Figure 2.1 Bivariate skewness with various mean vectors and correlation coeffcicents

Mardia 왜도는 $\rho_1 = -0.6, 0, 0.6$에서 0.15, 0.22, 0.40으로 분포의 치우침 방향을 알 수 없다. 이변량 왜도의 부호는 자료가 치우쳐 있는 방향을 나타내며, 치우쳐 있는 방향과 상관계수의 크기에 따라 왜도의 절대값의 차이가 발생함을 탐색할 수 있다. 12종류의 Model의 왜도에 따라 치우쳐있는 방향을 요약하면 Table 2.2와 같다.

또다른 특징은 분포형태가 직선에 가깝게 될수록 왜도의 절대값이 커진다는 점이다. 이는 분포가 직선 형태가 되어 길어질수록 분포의 치우침이 커지기 때문이다. 예를 들어 $\underline{\mu_1}$이 2, 4사분면에 위치한 Model 5와 11인 경우에 $\rho_1 = 0.6$일 때는 $\rho_1 = 0$일 때와 비교하여 왜도의 크기는 유의미한 차이가 없으나 $\rho_1 = -0.6$인 경우는 $\rho_1 = 0$일 때보다 큰 값을 갖는다. 반면 Mardia 왜도는 Model 5의 $\rho_1 = 0.6$, $\rho_1 = -0.6$에서 각각 0.40, 0.15 값을 가지며 치우침이 클 때 오히려 왜도값이 작아짐을 알 수 있다. 마찬가지로 $\underline{\mu_1}$이 1,3사분면에 위치한 Model 2와 8에서 $\rho_1 = -0.6$와 0 두 경우의 이변량 왜도값은 유의

Table 2.2 Explanation of skewness's sign

Sign of skewness	Explanation
$(+,+)$	The bivariate data is skewed from 3rd to 1st quadrant. Both X and Y are skewed to positive direction.
$(-,+)$	The bivariate data is skewed from 4th to 2nd quadrant. X is skewed to negative but Y is skewed to positive direction.
$(-,-)$	The bivariate data is skewed from 1st to 3rd quadrant. Both X and Y are skewed to negative direction.
$(+,-)$	The bivariate data is skewed from 2nd to 4th quadrant. X is skewed to positive but Y is skewed to negative direction.

미한 차이가 없으나 $\rho_1 = 0.6$인 경우에는 $\rho_1 = 0$일 때와 비교하여 큰 왜도값을 가짐을 탐색할 수 있다. 이 경우에서도 Mardia 왜도는 분포의 치우침이 커질 때 오히려 작아짐을 알 수 있다. 따라서 분포의 치우침의 방향과 정도를 확인할 수 있는 이변량 왜도가 Mardia 왜도보다 분포 형태를 탐색하는데 있어서 더 효과적이다.

Model과 상관계수에 따른 왜도 각각의 값을 Figure 2.2에 구현하였다. Figure 2.2를 바탕으로 왜도의 부호를 통해서 자료가 치우친 방향을 확인할 수 있으며, 왜도의 절대값이 크면 클수록 치우침의 정도가 크다고 말할 수 있다. 또한 평균벡터 $\underline{\mu_1}$이 수직축과 수평축에 가까울수록 왜도 성분 각각은 0에 근접한 값을 가지는 것을 파악할 수 있다.

3. 이변량 첨도

일변량 첨도는 4차 중심적률의 함수로 표현하며 Mardia 첨도 역시 4차 중심적률을 기반으로 하고있다. 그러나 Mardia 첨도의 문제점을 해결하기 위하여 본 연구에서는 X와 Y 각각의 2차 중심적률의 곱인 $Z_X^2 Z_Y^2$를 고려한다. $\underline{Z} = (Z_X, Z_Y)^t$라 할 때, $\underline{Z}$의 분산-공분산행렬의 행렬식 $|\Sigma|$의 제곱을 $Z_X^2 Z_Y^2$의 기대값에 나누어줌으로써 새로운 형태의 이변량 첨도를 정의 3.1과 같이 제안한다.

정의 3.1 표준화 이변량 확률변수 Z_X와 Z_Y에 대한 다음의 통계량을 이변량 첨도로 정의한다.

$$\text{첨도} = \frac{E[Z_X^2 Z_Y^2]}{|\Sigma|^2}. \qquad \square$$

또한 n개의 이변량 확률표본 $\{(X_i, Y_i); i = 1, \ldots, n\}$에 대한 이변량 첨도는

$$\textstyle\sum_{i=1}^{n} Z_{X_i}^2 Z_{Y_i}^2 / n(1-\hat{\rho}^2)^2$$

를 이용해서 계산할 수 있다. Mardia 첨도는 이변량 정규분포이면 두 변수의 상관계수에 관계없이 항상 8을 갖는다. 반면 본 연구에서 제안한 이변량 첨도는 상관계수가 0인 이변량 표준 정규분포에서 1의 값을 가지지만 상관계수의 절대값이 커질수록 증가한다. 우선 1보다 큰 첨도를 가지는 분포를 생성하기 위하여 다음과 같이 꼬리부분이 두터운 정규혼합분포 Distribution 3.1을 고려한다.

$$\text{Distribution 3.1} : \underline{X} \sim 0.5\ N(\underline{0}, I_2) + 0.5\ N(\underline{0}, \Sigma_2),$$

여기서 $I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} \sigma^2 & \rho_2\sigma^2 \\ \rho_2\sigma^2 & \sigma^2 \end{bmatrix}$.

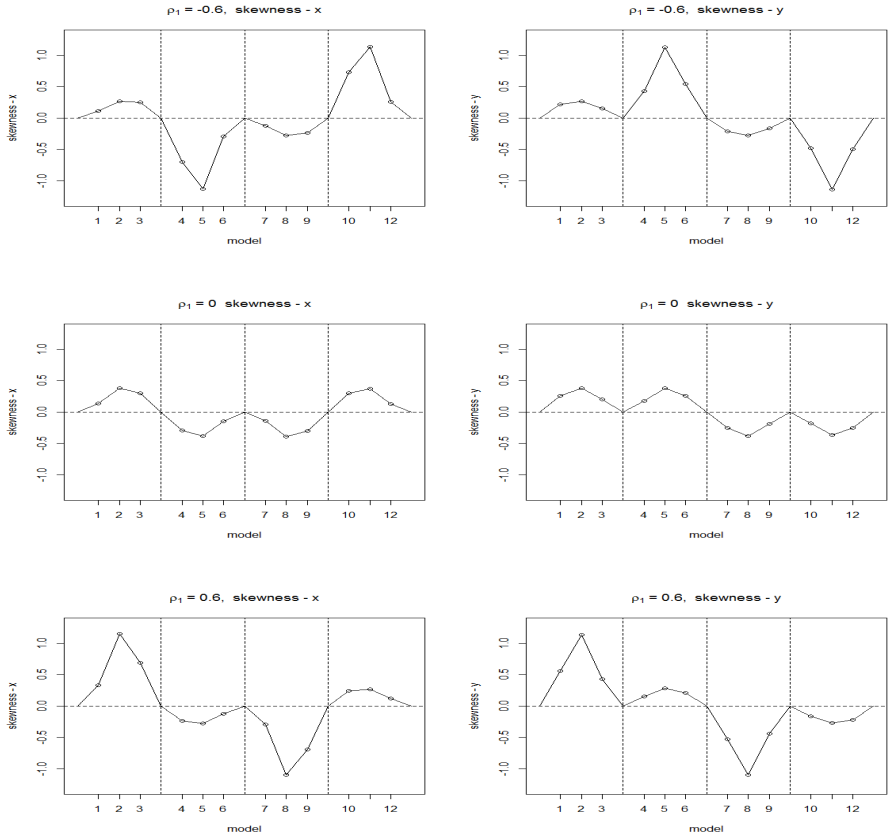


Figure 2.2 Each skewness values with various models and correlation coefficients

모의실험 결과를 살펴보기 전에 Distribution 3.1에서 $\lambda = 0.5$, $\underline{\mu} = \underline{0}$ 이므로 분산-공분산행렬, 상관계수는 다음과 같이 계산할 수 있다.

$$\Sigma = \begin{bmatrix} \frac{1+\sigma^2}{2} & \frac{\rho_2\sigma^2}{2} \\ \frac{\rho_2\sigma^2}{2} & \frac{1+\sigma^2}{2} \end{bmatrix}, \quad \rho = \frac{\rho_2\sigma^2}{(1+\sigma^2)}.$$

Distribution 3.1에서 σ=1, 1.5, 2, 2.5. ρ_2= 0, ±0.3, ±0.6으로 변화시키면서 1,000개의 확률표본을 추출하여 이변량 첨도를 계산하는 모의실험을 500회 반복하여 얻은 첨도값을 분산과 상관계수의 크기에 따라 Table 3.1의 'Kurtosis' 열에 정리하였으며, Figure 3.1에 시각적으로 구현하였다. 또한 Tble 3.1의 'Mardia' 열에 Mardia 첨도를 표기하였다.

Table 3.1과 Figure 3.1을 통하여 Distribution 3.1에서 분산 σ^2와 상관계수 ρ_2의 절대값이 증가할수록 이변량 첨도는 증가함을 알 수 있다. 결론적으로 분산과 상관계수의 절대값이 증가하면 분포의 꼬리부분이 길어지므로 새로 제안된 이변량 첨도는 큰 값을 갖는다. 반면 Mardia 첨도는 $\sigma^2 = 1.5^2$, $\sigma^2 = 2^2$, $\sigma^2 = 2.5^2$일 때 ρ_2의 절대값이 커져서 분포의 꼬리가 두꺼워져도 첨도값은 작아진다. 따라서 Marida 첨도보다 새로 제안된 이변량 첨도가 분포의 꼬리가 두꺼운 정도와 얇은 정도를 적절하게 측정

Table 3.1 Bivariate kurtosis with various variances and correlation coefficients

	$\sigma^2 = 1$		$\sigma^2 = 1.5^2$		$\sigma^2 = 2^2$		$\sigma^2 = 2.5^2$	
	kurtosis	Mardia	kurtosis	Mardia	kurtosis	Mardia	kurtosis	Mardia
$\rho_2 = 0.6$	1.64	8.47	2.67	8.86	3.86	10.04	4.85	11.17
$\rho_2 = 0.3$	1.14	8.11	1.45	9.05	1.80	10.64	2.07	11.9
$\rho_2 = 0$	1.00	7.87	1.14	9.1	1.36	10.81	1.53	12.08
$\rho_2 = -0.3$	1.14	8.09	1.45	9.05	1.80	10.67	2.07	11.93
$\rho_2 = -0.6$	1.64	8.49	2.67	8.86	3.86	10.02	4.85	11.15

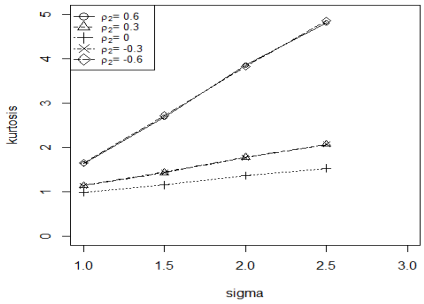


Figure 3.1 Bivariate kurtosis with various variances and correlation coefficients

하는 통계량으로 판단된다.

다음으로 1보다 작은 첨도값을 갖는 분포를 생성하기 위해 2절에서와 동일하게 Distribution 2.1에서 설정한 12개의 Model에서 가중값 λ을 0.5로 하는 정규혼합모형 Distribution 3.2를 고려한다.

$$Distribution\ 3.2 : 0.5\ N(\underline{\mu_1}, \Sigma_1) + 0.5\ N(-\underline{\mu_1}, \Sigma_1).$$

동일한 가중값을 부여한 12종류의 Model 중에서 처음 6개는 나머지 6개와 동일하다. 따라서 ρ_1의 값이 -0.6, 0, 0.6인 경우에 처음 6개의 Model을 고려하고, 각 Model에서 1,000개의 확률표본을 추출하여 이변량 첨도를 계산하는 모의실험을 500회 반복하면서 첨도값을 구하여 Table 3.2에 정리하였다.

Table 3.2 Bivariate kurtosis values with various mean vectors and correlation coefficients

	$\rho_1 = -0.6$		$\rho_1 = 0$		$\rho_1 = 0.6$	
	Kurtosis	Mardia	Kurtosis	Mardia	Kurtosis	Mardia
Model 1, 7	0.72	6.63	1.37	7.11	6.60	7.15
Model 2, 8	0.63	6.58	1.78	7.10	13.86	7.38
Model 3, 9	0.63	6.64	1.42	7.08	7.53	7.19
Model 4, 10	7.47	7.20	1.42	7.12	0.69	6.61
Model 5, 11	14.01	7.35	1.78	7.08	0.63	6.59
Model 6, 12	6.51	7.18	1.37	7.06	0.72	6.63

Table 3.2를 통하여 $\rho_1 = 0$일 경우에 평균벡터 $\underline{\mu_1}$의 위치에 상관없이 유사한 첨도값을 갖는다고 파악할 수 있다. 반면 Model 1, 2, 3 (또는 Model 7, 8, 9)에서 즉, $\underline{\mu_1}$이 1 (또는 3)사분면에 위치하는 모형에서 $\rho_1 = 0.6$이면 혼합분포의 상관계수 ρ가 큰 값을 가지며 정의 3.1의 첨도 분모가 작아져서 첨도값은 $\rho_1 = 0$인 경우에 비교하여 커지며, 반대로 $\rho_1 = -0.6$이면 ρ가 작은 값을 가지며 첨도의 분모가 커지기

때문에 첨도값은 작아진다. 그리고 Model 4, 5, 6 (또는 Model 10, 11, 12)에서 즉, $\underline{\mu_1}$이 2 (또는 4)사분면에 위치하는 Model에서 $\rho_1 = -0.6$인 경우에는 ρ값이 커지므로 $\rho_1 = 0$일 경우와 비교하여 첨도가 커지며, $\rho_1 = 0.6$인 경우에는 ρ가 작은 값을 가지므로 첨도값은 작아진다. 반면 Mardia 첨도값은 분포의 꼬리가 두꺼워질수록 값이 증가하는 것은 맞으나 그 증가 정도가 미미해서 증가 정도를 통해 분포의 두꺼움 정도를 파악하기 어렵다.

결론적으로 Distribution 3.1 모의실험에서와 같이 Mardia 첨도는 분포의 꼬리가 두꺼워질수록 적은 값은 반환하는 경우도 존재하며 Distribution 3.2 모의실험에서와 같이 분포의 꼬리가 두꺼워질 때 첨도가 커지는 경우에도 그 증가 정도가 미미하여 실제 어느 정도 두꺼운 꼬리를 갖는지 확인하기 쉽지 않다. 따라서 분포 형태 탐색에 있어서 이변량 첨도가 Mardia 첨도보다 효율적이다.

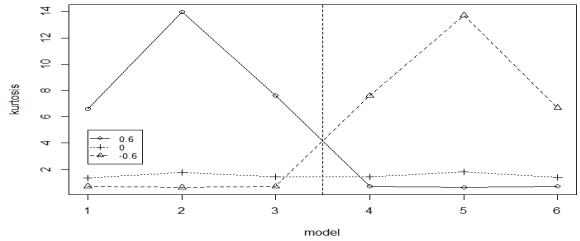


Figure 3.2 Bivariate kurtosis with various mean vectors and correlation coefficients

Distribution 3.2 모의실험의 이변량 첨도의 평균벡터의 위치와 상관계수의 변화에 따른 첨도값의 결과를 Figure 3.2에 표현하였다. Figure 3.2을 통해서 제안된 이변량 첨도는 이변량 분포에서 꼬리의 두꺼운 정도를 측정하는 통계량이며, 두 변수의 상관관계가 높아져서 일변량 정규분포에 근사하는 형태가 될수록 꼬리가 길어지므로 이변량 첨도값이 증가하는 것을 탐색할 수 있다. 그러므로 본 연구에서 제안한 이변량 첨도는 꼬리의 두꺼운 정도를 측정할 수 있는 적절한 척도이다.

4. 표면그림

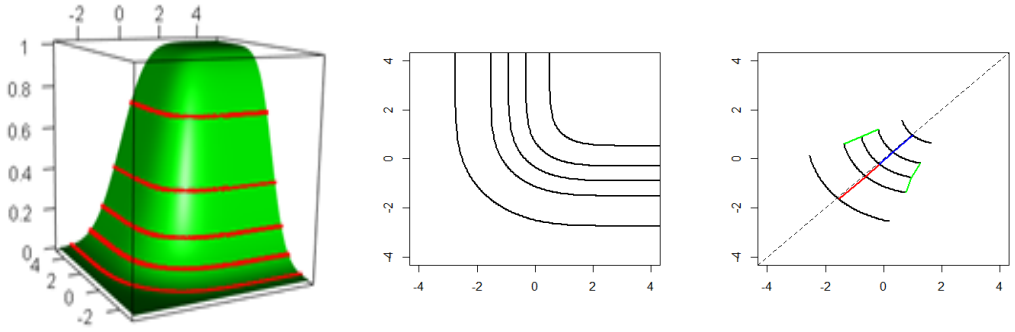
일변량 상자그림 (box-plot)은 자료의 크기 순서를 분위수 기반으로 구현할 수 있으나, 이변량 이상의 자료에서는 각 관찰값의 크기 순서를 정할 수 없기 때문에 상자그림을 이차원으로 확장할 수 없다. 본 연구에서는 이변량 자료를 시각적으로 구현하는 방법을 제안한다. Hong 등 (2016), Hong과 Lee (2016), 그리고 Hong과 Kim (2017)이 제안한 이변량 확률표본을 순서화 할 수 있는 분위벡터 (quantile vector)를 이용하여 자료의 5, 25, 50, 75, 95 백분위벡터를 찾은 후 이를 각각 최소 (minimum) 벡터, 하사분위 (lower quartile) 벡터, 중위 (median) 벡터, 상사분위 (upper quartile) 벡터 그리고 최대 (maximum) 벡터로 설정하며 이를 시각적으로 구현한 형태를 표면그림 (surface plot)이라 제안한다. 표면그림을 작성하는 방법은 다음과 같다.

표면그림 (surface plot) 작성하는 세 단계

Step 1: 이변량 확률표본 자료의 5, 25, 50, 75, 95 백분위벡터를 얻는다. 예를들어 상관계수가 0인 이변량 표준정규분포의 CDF를 3차원에서 표현한 Figure 4.1의 (a)에서 5, 25, 50, 75, 95 분위벡터를 진한 선으로 표현하였다.

Step 2: 다섯 종류의 분위벡터를 이차원 평면에 구현한다. Step 1에서 구한 5개의 분위벡터를 이차원 평면에 표현하면 Figure 4.1 (b)와 같다.

Step 3: 분포의 밀도가 존재하지 않는 부분을 분위벡터에서 삭제한다. 즉, 이차원 평면에서 위에서 아래 방향과 왼쪽에서 오른쪽 방향으로의 직선으로 표현되는 분위벡터는 삭제한다. 그리고 사분위 벡터 (quartile vector)인 하사분위벡터, 중위벡터, 상사분위벡터들의 끝부분을 연결하여 곡면 (curved surface)을 만들고, 최소벡터와 최대벡터의 평균점을 선으로 연결한다. (Figure 4.3 (c) 참조).



(a) (b) (c)

Figure 4.1 Quantile vector and surface plot

Figure 4.2은 이변량 표준정규분포에서 상관계수 ρ를 변화시켜가면서 구현한 표면그림이다. 상관계수가 음수인 $\rho = -0.45$의 경우는 가운데 50% 확률표본을 나타내는 곡면이 2와 4사분면 방향으로 길게 표현되고 폭은 짧으며 최대벡터와 최소벡터의 간격은 길다. 반면에 상관계수가 양수인 $\rho = 0.45$의 경우에는 가운데 50% 확률표본을 나타내는 곡면이 짧으며 폭은 증가하여 상관계수가 음수인 경우보다 넓다. 그리고 최대벡터와 최소벡터의 간격은 상대적으로 짧다. 참고로 Figure 4.2의 세 종류의 표면그림에 대응하는 분포의 이변량 왜도는 모두 (0,0)이며, 이변량 첨도는 2.22 ($\rho = -0.45$), 1 ($\rho = 0$), 2.22 ($\rho = 0.45$)이다.

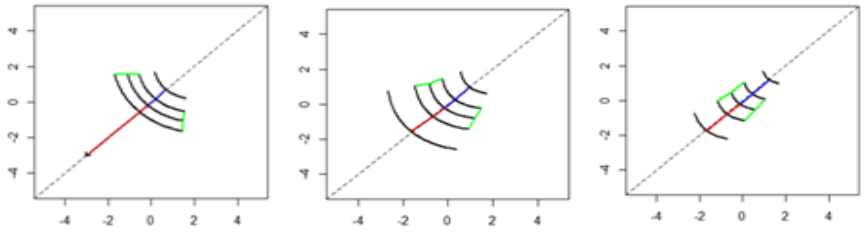
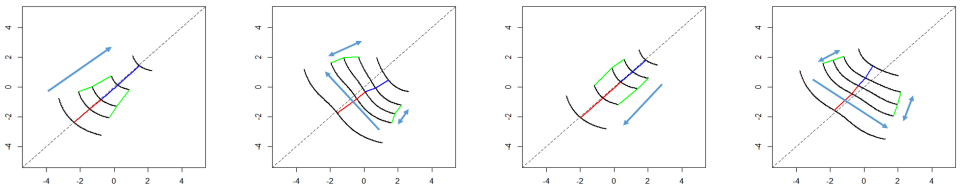


Figure 4.2 Surface plots based on Standard normal distributions: $\rho = -0.45, 0, 0.45$

일변량 상자그림과 유사하게 표면그림에서도 그림의 형태를 통해 이변량 왜도 및 첨도에 관한 정보를 얻을 수 있는지를 살펴보기 위하여 2, 3절에서 논의한 Distribution에 대한 표면그림을 Figure 4.3과 Figure 4.4에 구현하였다.

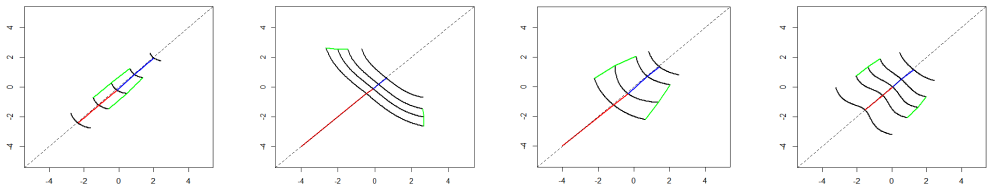


(a) (b) (c) (d)

Figure 4.3 Surface plots for Distribution 2.1

우선 Distribution 2.1의 $\rho_1 = 0$일 때 Model 2, 5, 8, 11에 대응하는 표면그림은 Figure 4.3과 같다. Model 2 (Figure 4.3 a)에 대응하는 표면그림은 이변량 왜도가 모두 양수인 경우인데 1사분면 방향으로 분위벡터 간격이 증가한다. Model 8 (Figure 4.3 c)에 대응하는 표면그림은 왜도가 모두 음수인 반대의 경우로 3사분면 방향으로 분포가 더 멀리 형성되므로 표면그림 역시 3사분면 방향으로 분위벡터의 간격이 넓어진다. Model 5 (Figure 4.3 b)에 대응하는 표면그림은 왜도가 음수와 양수이므로 2사분면 방향으로 분포가 더 멀리 형성되기 때문에 표면그림 자체가 4사분면에서 2사분면 방향으로 올라가며 가운데 50% 확률표본을 나타내는 곡면의 좌측 선분이 우측 선분보다 더 길다. Model 11 (Figure 4.3 d)에 대응하는 표면그림은 왜도가 양수와 음수이므로 Model 5와 비교하여 반대로 설명할 수 있다.

이변량 첨도와 표면그림의 관계를 살펴보기 위하여 첨도값이 1보다 큰 Distribuion 3.2의 Model 2 ($\rho_1 = 0.6$)와 Model 5 ($\rho_1 = -0.6$)을 선택하고, 첨도값이 1보다 작은 Distribution 3.2의 Model 2 ($\rho_1 = -0.6$)와 Model 5 ($\rho_1 = 0.6$)을 선택한다. $\underline{\mu_1}$의 좌표를 시점으로, $-\underline{\mu_1}$의 좌표를 종점으로 하는 벡터를 따라 혼합분포가 분포하여 꼬리가 길며 첨도가 큰 경우인 Figure 4.4의 (a), (b)는 50% 확률표본을 나타내는 곡면이 두 평균 좌표가 만드는 벡터를 따라 길고 얇게 형성되며 이때의 첨도값은 13.86으로 1보다 매우 큰 값을 갖는다. 반면 첨도가 낮은 경우인 Figure 4.4의 (c), (d)는 곡면이 두 평균이 만드는 벡터 방향을 따라서 두껍고 짧게 형성되며 첨도값은 1보다 작은 0.81을 갖는다.



(a) Model 2 ($\rho_1 = 0.6$) (b) Model 5 ($\rho_1 = -0.6$) (c) Model 2 ($\rho_1 = -0.6$) (d) Model 5 ($\rho_1 = 0.6$)

Figure 4.4 Surface plots for Distribution 3.2

5. 결론

일변량 왜도와 첨도는 분포의 모양이 정규분포를 기준으로 어느 방향으로 어느 정도 치우쳐있는지 그리고 어느 정도 뾰족하고 꼬리가 두꺼운지 등을 판단할 수 있는 통계량이다. Mardia가 제안한 다변량 왜도와 첨도 중에서 특히 이변량 왜도와 첨도는 변수들의 상관관계를 고려하지 않음으로써 치우침의 방

향에 대한 정보를 제공하지 못하므로 분포형태를 측정하는 척도로서 문제점을 갖고 있다. 본 논문에서 제안한 왜도와 첨도는 두 변수들의 상관관계를 반영하고 분포의 중심을 기준으로 자료의 치우침 방향과 정도를 파악할 수 있으며, 이변량 정규분포를 기준으로 꼬리 부분이 두터운 정도를 파악할 수 있으므로 Mardia 왜도와 첨도의 문제점을 개선하였다.이변량 왜도는 실수쌍으로 표현되며 수평축과 수직축을 바탕으로 이변량 자료의 치우침 방향과 정도를 상관계수의 함수로 표현한다. 또한 이변량 첨도는 단일 값으로 표현하며 이변량 분포의 중심에 몰려있는 정도와 꼬리 부분의 두터운 정도를 이변량 표준정규분포를 기준으로 판단할 수 있다. 결론적으로 두 변수의 상관관계에 따라 왜도와 첨도값이 변하기 때문에 정확하게 분포의 성질을 탐색할 수 있는 장점이 있다.

일변량 자료를 시각적으로 표현하는 상자그림은 자료의 특성을 잘 구현하는 특징이 있으며, 분위수 중에서 사분위수와 최소값, 최대값을 바탕으로 표현하기 때문에 왜도와 첨도를 잘 반영하는 시각적인 방법이다. 이변량 이상의 자료를 시각적으로 표현하기 위하여 본 연구에서는 분위벡터를 바탕으로 사분위벡터를 이용하고 최소값과 최대값보다는 5와 95 백분위벡터를 이용하여 시각적으로 구현하는 표면그림을 제안하였다. 다양한 형태의 이변량 분포를 생성하여 표면그림을 작성하고, 본 연구에서 제안한 이변량 왜도와 첨도를 계산하여 탐색해 본 결과 왜도와 첨도값의 부호와 크기가 표면그림으로 표현한 이변량 분포의 특징을 잘 반영하는 것을 탐색하였다. 그러므로 본 논문에서 제안한 왜도와 첨도는 이변량 분포의 형태를 측정하는 기술통계량이며 표면그림은 이변량 분포를 시각적으로 표현하는 방법으로 유용하게 활용할 수 있다.

본 논문에서 제안한 표면그림에서 활용하는 분위벡터는 삼변량 이상의 자료에 대하여도 확장할 수 있지만, 삼변량 자료에 대해서만 시각적인 방법을 구현할 수 있겠다. 그리고 본 논문에서 제안한 왜도와 첨도의 정의는 삼변량 이상 다변량 자료에 대해서도 확장 가능하다. k변량 자료의 왜도는 3차 중심 적률, 그리고 첨도는 4차 중심 적률의 함수로 정의할 수 있으며 왜도는 k개 성분을 갖는 벡터로 그리고 첨도는 단일값으로 표현할 수 있는데 이에 대한 연구는 향후 과제로 남겨두기로 한다.

References

Bickel, P. J. and Doksum, K. A. (2001). *Mathematical statistics: Basic ideas and selected topics*, 2nd Ed., Prentice-Hall, New Jersey.

Hogg, R. V., Mckean, J. W. and Craig, A. T. (2013). *Introduction to mathematical statistics*, 7th Ed., Prentice-Hall, New Jersey.

Hong, C. S. and Kim, H. J. (2016). *Data analysis for regression model using SAS*, Tamjin, Seoul.

Hong, C. S., Han, S. J. and Lee, G. P. (2016). Vector at risk and alternative value at risk. *The Korean Journal of Applied Statistics*, **29**, 689-697.

Hong, C. S. and Kim, J. Y. (2017). Multivariate cte for copula distributions. *Journal of the Korean Data & Information Science Society*, **28**, 421-433.

Hong, C. S. and Lee, G. P. (2016). Properties of alternative var for multivariate normal distributions. *Journal of the Korean Data & Information Science Society*, **27**, 1453-1463.

Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation*, 2nd Ed., Springer, New York.

Lindgren, B. W. (1993). *Statistical theory*, 4th Ed., Chapman & Hall, London.

Mardia, K. V. (1970). Measure of multivariate skewness and kurtosis with applications. *Biometrica*, **57**, 519-530.

Mardia, K. V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *The Indian Journal of Statistics, Series B*, 115-128.

Mood, A. M., Graybill, F. A. and Boes, D. C. (1974). *Introduction to the theory of statistics*, 3rd ed., McGraw-Hill, New York.

Rohatgi, V. K. (1976). *An introduction to probability theory and mathematical statistics*, John Wiley & Sons, New York.

Seber, G. A. F. (1984). *Multivariate observations*, John Wiley & Sons, New York.

Journal of the Korean Data & Information Science Society
2017, **28**(5), 959–970

http://dx.doi.org/10.7465/jkdi.2017.28.5.959
한국데이터정보과학회지

Bivariate skewness, kurtosis and surface plot

Chong Sun Hong [1] · Jae Hyun Sung [2]

[12]Department of Statistics, Sungkyunkwan University

Received 10 August 2017, revised 15 September 2017, accepted 18 September 2017

Abstract

In this study, we propose bivariate skewness and kurtosis statistics and suggest a surface plot that can visually implement bivariate data containing the correlation coefficient. The skewness statistic is expressed in the form of a paired real values because this represents the skewed directions and degrees of the bivariate random sample. The kurtosis has a positive value which can determine how thick the tail part of the data is compared to the bivariate normal distribution. Moreover, the surface plot implements bivariate data based on the quantile vectors. Skewness and kurtosis are obtained and surface plots are explored for various types of bivariate data. With these results, it has been found that the values of the skewness and kurtosis reflect the characteristics of the bivariate data implemented by the surface plots. Therefore, the skewness, kurtosis and surface plot proposed in this paper could be used as one of valuable descriptive statistical methods for analyzing bivariate distributions.

Keywords: Box plot, mahalanobis distance, mixture, quantile vector, surface plot.

[1] Corresponding author: Professor, Department of Statistics, Sungkyunkwan University, Seoul 03063, Korea. E-mail: cshong@skku.edu
[2] Graduate student, Department of Statistics, Sungkyunkwan University, Seoul 03063, Korea.