KJA
Korean Journal of Anesthesiology

Statistical Round
pISSN 2005-6419 · eISSN 2005-7563

Statistical data preparation: management of missing values and outliers

Sang Kyu Kwak[1] and Jong Hae Kim[2]

Departments of [1]Medical Statistics, [2]Anesthesiology and Pain Medicine, School of Medicine, Catholic University of Daegu, Daegu, Korea

Missing values and outliers are frequently encountered while collecting data. The presence of missing values reduces the data available to be analyzed, compromising the statistical power of the study, and eventually the reliability of its results. In addition, it causes a significant bias in the results and degrades the efficiency of the data. Outliers significantly affect the process of estimating statistics (*e.g.*, the average and standard deviation of a sample), resulting in overestimated or underestimated values. Therefore, the results of data analysis are considerably dependent on the ways in which the missing values and outliers are processed. In this regard, this review discusses the types of missing values, ways of identifying outliers, and dealing with the two.

Key Words: Bias, Data collection, Data interpretation, Statistics.

서 론

사회과학 및 자연과학 등의 전반적 분야의 관측 또는 실험되어 얻어지는 데이터를 수집함에 있어서 결측값(missing value)과 이상점(outlier)은 빈번히 나타나는 현상이다.

결측값은 정보의 손실, 연구대상의 탈락, 연구대상의 무응답 등으로 인하여 발생한다. 결측값이 발생하면 연구에서 최초에 의도한 연구대상자의 수를 축소해 연구의 신뢰도를 떨어뜨린다. 또한, 표본을 바탕으로 모집단을 추정함에서도 결과에 편향(bias)을 초래하여 데이터의 효율성을 저하시킨다. 이런 결측값은 분석하기 전 분석의 간편성을 위해서 무시되기도 하고, 또는 통계적 분석 기법을 통하여 추정된 대체 값을 사용하기도 한다. 일반적으로 결측값은 분석할 때는 효율성(efficiency), 결측값 처리 및 분석의 복잡성, 관측값과 결측값 간의 편향을 고려하여야 한다.

이상점이란 각 변수의 분포에서 비정상적으로 분포를 벗어나는 값, 즉 극단적 값을 의미한다. 만약 몸무게에 대한 데이터를 수집하였을 때 한 명의 몸무게가 250 kg이라면 이 값은 몸무게에 대한 정상적 분포를 벗어나는 값이므로 이상점으로 판단할 수 있다. 이상점은 데이터 입력 오류, 연구대상의 응답 오류 등으로 인하여 발생한다. 이상점은 데이터의 분포에서 비정상적으로 분포를 벗어난 값이다. 비정상적으로 극단값을 갖는 경우나 비현실적인 변숫값들이 이에 해당한다. 표본조사에서 이상점이 발생하면 모집단의 평균 등의 통계량을 추정하는 과정에서 이상점이 편향을 야기시켜 추정값이 과대 또는 과소 추정된다. 이런 이상점이 포함된 데이터는 분석하기 전 이상점의 값을 다시 추적하여 확인 및 수정하거나 이상점을 결측값으로 처리하여 대체 값을 사용하기도 한다.

Corresponding author: Jong Hae Kim, M.D.
Department of Anesthesiology and Pain Medicine, School of Medicine, Catholic University of Daegu, 33, Duryugongwon-ro 17-gil, Nam-gu, Daegu 42472, Korea
Tel: 82-53-650-4979, Fax: 82-53-650-4517
Email: usmed@cu.ac.kr
ORCID: http://orcid.org/0000-0003-1222-0054

Received: May 11, 2017.
Revised: June 1, 2017 (1st); June 19, 2017 (2nd).
Accepted: June 20, 2017.

Korean J Anesthesiol 2017 August 70(4): 407-411
https://doi.org/10.4097/kjae.2017.70.4.407

따라서, 수집된 데이터에서 결측값과 이상점을 어떻게 처리하느냐에 따라 데이터 분석 결과가 상이하게 나타날 수 있으므로 결측값과 이상점을 올바르게 처리하여 분석을 진행하는 것이 필요하다. 이 종설에서는 결측값의 종류와 결측값을 처리하는 방법에 대해서 알아보고 이상점을 검정하는 방법과 처리하는 방법에 대한 내용을 다루어 보고자 한다.

결측값의 종류

결측값에 관한 기존 연구들은 결측값의 종류을 결측값이 발생하는 유형에 따라 완전무작위 결측 발생(MCAR: missing completely at random), 임의적 결측 발생(MAR: missing at random), 비임의적 결측 발생(NMAR: not missing at random) 등으로 분류한다(Table 1)[1].

결측값 종류의 설명을 위해서 전체 연구대상자가 I명이고 총 측정시점이 J번 일 때 다음과 같이 기호를 정의한다.

Y_{ij}: i번째 환자의 j번째 측정값, i = i, ⋯ , I , j = 1, ⋯ , J

$\mathbf{Y}_{i(observation)}$: i번째 환자의 관측값으로 구성된 백터

$\mathbf{Y}_{i(missing)}$: i번째 환자의 결측값으로 구성된 백터

$\mathbf{R}_i = (R_{i1}, R_{i2}, \cdots R_{ij} \cdots R_{ij})$: i번째 환자의 j번째 측정값이 결측값인지를 나타내는 지시 함수 백터

R_{ij} = 1만약 Y_{ij} 가 결측값일 경우

R_{ij} = 0만약 Y_{ij} 가 관측값일 경우

완전무작위 결측 발생(MCAR: Missing completely at random)

만약 $\mathbf{Y}_{i(observation)}$, $\mathbf{Y}_{i(missing)}$ 그리고 $\mathbf{R}_i$가 모두 독립인 경우 $\mathbf{Y}_{i(missing)}$를 완전무작위 결측값이라고 한다. 즉, 결측값이 발생한 경우가 다른 값에 영향을 받지 않고 완전히 랜덤하게 발생하였다는 것이다. 완전무작위 결측값은 갑자기 연구대상자가 측정시점에 나타나지 않아 총 측정시점 중간에 발생할 수도 있고, 연구에서 중도탈락을 하여 특정 측정시점부터 연구종료시까지 발생할 수도 있다.

예를 들어, A라는 환자가 j번째 시점에서 개인적인 일로 방문을 못하는 경우와 동의철회, 주요검사 누락, 사망, 추적실패, 중대한 이상반응 등의 이유로 연구참여에서 중도탈락하였다면 이 경우 발생한 결측값은 완전무작위 결측값이다.

임의적 결측 발생(MAR: Missing at random)

만약 $\mathbf{R}_i$가 $\mathbf{Y}_{i(observation)}$과는 종속이지만 $\mathbf{R}_i$가 $\mathbf{Y}_{i(missing)}$과는 독립인 경우 $\mathbf{Y}_{i(missing)}$를 임의적 결측값이라고 한다. 즉, 결측값이 발생하였는가의 여부가 $\mathbf{Y}_{i(observation)}$와 관련이 있고 $\mathbf{Y}_{i(missing)}$과는 관련이 없다는 것이다. 임의적 결측값은 특정 시점에서 연구대상자가 참여한 연구 성과에 만족하지 못할 시 발생할 수 있다.

예를 들어, A라는 환자가 j번째 시점에서 이전의 측정결과를 확인하고 연구에 만족하지 못하는 경우 j번째 측정을 의도적으로 하지 않을 수 있다. 이 경우 발생된 결측값은 임의적 결측값이다. 또한, 임의적 결측값은 완전무작위 결측값보다 임상연구에서 훨씬 보편적으로 발생하는 결측 유형이라고 할 수 있다.

비임의적 결측 발생(NMAR: Not missing at random)

만약 $\mathbf{R}_i$가 $\mathbf{Y}_{i(observation)}$과 종속이고 $\mathbf{R}_i$가 $\mathbf{Y}_{i(missing)}$과도 종속인 경우 $\mathbf{Y}_{i(missing)}$를 비임의적 결측값이라고 한다. 즉, 결측값이 발생하였는가의 여부가 $\mathbf{Y}_{i(observation)}$과 $\mathbf{Y}_{i(missing)}$에 모두 관련이 있다는 것이다. 비임의적 결측값은 특정 시점에서 연구대상자가 참여한 연구 성과에 만족하지 못하는 것과 동시에 방문 전 개인적으로 측정한 경우 발생할 수 있다.

예를 들어, A라는 환자가 j번째 시점에서 이전의 측정결과를 확인하고 연구에 만족하지 못하고 있으면서, 다음 측정을 위해 방문을 하기 전 개인적으로 측정하여 이전의 측정결과와 비슷한 결과를 확인한다면 j번째 시점에서 측정을 의도적으로 하지 않을 수 있다. 이 경우 발생한 결측값은 비임의적 결측값이다.

Table 1. Types of Missing Values

결측값의 종류	내용	예시
완전 무작위 결측 발생	결측값이 발생한 경우가 다른 값에 영향을 받지 않고 완전히 랜덤하게 발생	동의철회, 주요검사 누락, 사망, 추적실패, 중대한 이상반응 등
임의적 결측 발생	특정 시점에서 연구대상자가 참여한 연구 성과에 만족하지 못할 시 발생	연구 참여 불만족으로 인하여 측정 거부
비임의적 결측 발생	특정 시점에서 연구대상자가 참여한 연구 성과에 만족하지 못하는 것과 동시에 방문 전 개인적으로 측정을 한 경우 발생	연구 참여 불만족으로 인하여 개인적 측정 결과가 여전히 불만족으로 인한 측정 거부

결측값 처리 방법

결측값을 가지는 데이터를 분석하기 위한 연구가 많이 진행되었다[2,3]. 본 절에서는 결측값을 어떻게 처리하여 분석하는지를 살펴보도록 한다.

완전히 관측된 데이터 분석(Complete case analysis)

결측값이 있는 데이터는 모두 무시하고 모든 변수 및 시점에서 관측된 데이터만 이용하여 분석하는 방법이다. 분석이 간편하다는 장점이 있지만, 많은 표본수가 줄어들고 검정력이 낮아지므로 통계적 추론에도 문제가 발생하는 단점이 있다. 대부분의 데이터 분석 소프트웨어 프로그램인 SPSS, SAS 등에서 가장 보편적으로 사용되는 결측값 처리 방법이다.

이용가능한 데이터 분석(Available case analysis)

각각의 분석에서 사용 가능한 데이터를 이용하여 분석하는 방법이다. 완전히 관측된 데이터 분석하는 경우보다는 표본수가 많으나 표본의 수가 분석에서 사용하는 변수마다 달라지는 단점이 있다.

대체 분석(Imputation analysis)

결측값을 통계적 방법을 통하여 대체하여 결측값이 없는 완전한 데이터를 생성하여 분석하는 방법이다. 대체는 명시적 모형(explicit modeling)에 의한 대체와 내재적 모형(implicit modeling)에 의한 대체로 두 가지 종류가 존재한다. 우선 명시적 모형에 의한 대체는 각 변수들이 특정한 확률분포를 따른다는 가정하에 각 분포의 모수(parameter)를 추정하여 대체하는 방법으로 평균값 대체법, 중앙값 대체법, 확률 대체법, 비율 대체법, 회귀 대체법, 확률적 회귀 대체법, 분포를 가정한 대체법 등이 존재한다. 또한, 내재적 모형에 의한 대체는 가능한 한 정확한 값을 통하여 대체하기 위한 알고리즘 계산에 중점을 둔 방법으로 핫덱(hotdeck) 대체법, 콜드덱(colddeck) 대체법, 대입법(substitution) 등이 존재한다. 그리고 명시적 모형과 내재적 모형이 혼합된 방식의 대체도 존재한다. 본 종설에는 각각의 대체법을 자세히 설명하지는 않도록 하겠다.

이상점 검정 방법

이상점은 특정 분포를 따르지 않고 비정상적으로 분포를 벗어나는 값이므로 분포 형태가 없는 비모수적 분포에서는 이상점을 판단할 필요가 없다. 그러나 많은 데이터가 정규분포를 포함해서 분포를 가정하고 있기 때문에 데이터를 분석하기에 앞서 이상점을 판단하여야 한다. 이상점을 판단하는 방법은 매우 다양하게 존재한다. 우선 일반적으로 이상점은 데이터의 중심으로부터 상대적으로 벗어난 거리를 사용해서 검정한다. 정규분포 하에서 평균에서 3배의 표준편차보다 크거나 작은 값을 이상점으로 판단하는 방법이 이 경우에 해당한다. 그러나 평균과 표준편차의 값은 이상점이 존재할 경우 매우 민감하게 반응하는 통계량이므로 적합한 방법이라고 볼 수 없다. 따라서 평균과 표준편차에 민감하지 않은 통계량인 중위수와 사분위수를 이용하여 이상점을 검정한다. 상자그림으로 이상점을 판단하는 경우에 해당한다(Fig. 1). 상자그림에서 윗 울타리나 아랫 울타리를 벗어나는 값을 이상점으로 판단한다.

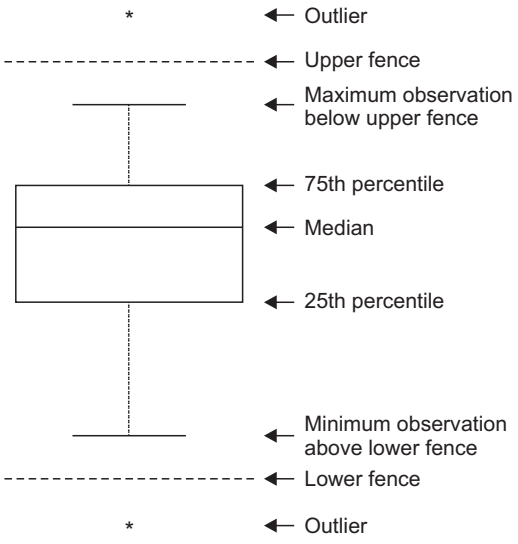

Fig. 1. Boxplot with outliers. The upper and lower fences represent values more and less than 75th and 25th percentiles (3rd and 1st quartiles), respectively, by 1.5 times the difference between the 3rd and 1st quartiles. An outlier is defined as the value above or below the upper or lower fences.

이 외에도 이상점을 검정하기 위한 연구는 활발히 진행되었다. 우선 회귀분석시, 해당 자료를 제외하고 분석한 잔차와 전체 자료에서 분석한 잔차를 이용하여 이상점을 판단하는 방법이 있다[4]. 또한 서포트 백터 회귀를 이용한 이상점 진단의 방법도 있다[5]. 만약 K개의 그룹에서 동일한 자료를 수집하거나 또는 한 대상에 대해서 자료를 반복하여 수집하는 경우 어느 그룹 또는 어느 대상의 응답이 이상점인지 판단할 필요가 있다. 각 그룹별 평균과 분산에 기초하여 이상점을 판단하는 방법도 연구되었다[6]. 따라서 하나의 변수만 사용하는 경우에는 상자그림으로 이상점 여부를 확인하고, 관계를 알아보기 위해 여러 가지 변수를 사용하는 경우에는 이상점을 판단하는 검정을 사용하여 이상점 여부를 확인하

여야 한다. 본 종설에는 각각의 이상점 검정 방법에 대해서 자세히 설명하지는 않도록 하겠다.

이상점 처리 방법

데이터에서 검정된 이상점을 처리하는 방법은 크게 세 가지로 구분할 수 있다. 먼저, 이상점을 제외하는 방법(trimming)과 이상점의 값을 다른 값으로 교체하거나 분석할때 이상점에 대한 가중치를 정상값과는 다르게 조정하여 이상점의 영향력을 감소시키는 방법, 마지막으로 로버스트(robust) 기법을 적용하여 이상점의 값을 추정하는 방법 등이 있다.

이상점을 제외하는 방법(Trimming)

이상점으로 검정된 데이터를 제외하고 데이터를 분석하는 방법으로, 평균과 같이 분석의 결과로 산출된 값들의 분산은 작아지지만 실제 값보다 과소 또는 과대 추정되어 편의가 발생한다. 따라서 이상점 데이터도 실제 관측된 값이므로 이상점을 제외하고 분석하는 것은 관측된 값을 모두 반영하는 방법이 아니므로 적합한 이상점 차리 방법은 아니다.

윈저화 방법(Winsorization)

윈저화 방법은 분석시 이상점에 대한 가중치를 조정하거나 이상점으로 검정된 값을 다른 값으로 대체하는 방식으로 이상점을 처리한다. 가중치 조정방법(weight modification method)은 이상점 값을 다른 값으로 바꾸거나 제외하지 않고 가중치를 조정함으로써 이상점의 영향을 감소시키는 방법이다. 관측값 변경(value modification)은 이상점을 제외한 나머지 데이터 중에서 최댓값 또는 최솟값에 가까운 값으로 이상점을 대체하는 방법이다.

로버스트 추정방법

로버스트 추정방법은 이상점에 해당하는 값을 추정하는 방법으로 모집단 분포가 정해져 있을 경우 적합한 방법이며, 추정된 값은 일관성을 지니고 있다. 최근 많은 연구가 진행되고 있으며, 여러 가지 통계모형이 제안되었지만, 방법적 구조가 너무 복잡하여 거의 사용되지 않고 있다.

예 제

0에서 10까지의 값을 가지는 VAS 변수에 대해 5명으로 구성된 데이터를 생각해보자(Table 2). 우선 5번째 대상자의 값이 결측값이라고 가정하자. 앞에서 언급한 결측값 처리 방법 중 완전히 관측된 데이터 분석 방법을 사용한다면 평균(표준편차)이 2.50 (1.29)이다. 또한, 대체분석방법 중 평균값 대체법을 사용한다면 평균(표준편차)이 2.50 (1.12) 이다. 다음으로 5번째 대상자의 값이 99라고 가정하자. 99라는 값은 VAS값의 범위인 0에서 10을 벗어난 값으로 측정된 값이 아니라 데이터 입력의 오류로 발생한 값이며 이상점은 아니다. 이럴 경우에는 데이터 클리닝(data cleaning)을 통하여 원 자료를 다시 확인하여 오류의 값을 수정하여야 한다. 마지막으로 측정된 VAS값이 평균이 3이고 표준편차가 1인 정규분포를 따르고 5 번째 대상자의 값이 9라고 가정하자. 9는 평균 (3) + 3 × 표준편차 (1) 값인 6보다 큰 값이므로 VAS의 정상 분포를 벗어나는 극단적 값이므로 이상점이다. 또한, 연구 대상자의 응답 오류로 발생한 이상점이라고 볼 수 있다. 5번째 대상자의 값을 이상점으로 처리하지 않고 분석한다면 평균(표준편차)이 4.20 (2.77)로 나타나 평균(표준편차)이 과대 추정되는 문제점이 발생

Table 2. Examples of Missing Value and Outlier

No.	Data with a missing value			Date with an outlier		
	Raw data	Complete case	Imputation with the mean value	Raw data	Complete case	Winsorization with the maximum value
1	1	1	1	1	1	1
2	2	2	2	2	2	2
3	3	3	3	3	3	3
4	4	4	4	4	4	4
5	.*	-†	2.5‡	9§	-†	4‖
			Summary			
N	NA	4	5	5	4	5
Mean	NA	2.50	2.50	4.20	3.00	3.2
SD	NA	1.29	1.12	2.77	0.82	0.84

N: the number of a sample, NA: not applicable. *Missing value, †Discarded value, ‡Imputed mean value, §Outlier, ‖Winsorized maximum value.

한다. 앞에서 언급한 이상점 처리 방법 중 이상점을 제외하는 방법을 사용한다면 평균(표준편차)이 3.00 (0.82)이다. 또한, 윈저화 방법 중 이상점을 제외한 나머지 데이터 중에서 최댓값 또는 최솟값에 가까운 값인 최댓값인 4로 이상점을 대체하면 평균(표준편차)이 3.2 (0.84) 이다.

이와 같이, 결측값과 이상점을 처리한다면 연구대상자의 수를 축소시키지 않을 뿐만 아니라 결과의 편향을 초래하지도 않으며, 통계량을 과대 또는 과소 추정하지 않는다.

결 론

결측값을 가진 데이터를 분석할 때에는 결측값에 대한 발생원인을 고려하여 적절한 방법으로 결측값을 처리하여야 할 것으로 사료된다. 더욱이 결측값의 비율이 높은 데이터일수록 더욱더 신중한 분석이 필요할 것이다. 결측값을 처리하는 방법으로 딱 한가지 방법만 사용하는 대신 여러 가지 방법을 사용하여 보고 비교하는 것도 하나의 좋은 방법으로 생각된다. 수집된 자료에서 이상점을 검정하고 처리하는 과정은 중요하다. 더 나아가서 이상점을 가지는 데이터를 줄이기 위한 노력으로 데이터 입력시 잘못 입력되는 오류를 최소화하여야 한다. 이는 입력자의 주의가 가장 필요한 부분이다. 데이터를 분석하기에 앞서 데이터 전처리 과정으로 결측값과 이상점을 적절하게 처리하여 데이터 분석 결과에서 발생할 수 있는 편향, 과대 또는 과소 추정의 문제를 최소화하는 노력이 필요할 것으로 생각된다.

ORCID

Sang Gyu Kwak, http://orcid.org/0000-0003-0398-5514
Jong Hae Kim, http://orcid.org/0000-0003-1222-0054

References

1. Rubin DB. Inference and missing data. Biometrika 1976; 63: 581-92.
2. Rubin DB. Multiple imputation after 18+ years. J Am Stat Assoc 1996; 91: 473-89.
3. Schafer JL. Multiple imputation: a primer. Stat Methods Med Res 1999; 8: 3-15.
4. Gentleman J, Wilk M. Detecting outliers II: supplementing the direct analysis of residuals. Biometrics 1975; 31: 387-410.
5. Seo HS, Yoon M. Outlier detection using support vector machines. Commun Stat Appl Methods 2011; 18: 171-7.
6. Burke S. Missing values, outliers, robust statistics & non-parametric methods. LC-GC Eur Online Suppl Stat Data Anal 2001; 2: 19-24.