이상치 탐색을 위한 통계적 방법

Statistical methods for outlier detection

김진휘 주임연구원
건강보험심사평가원 심사연구부

Key Points
☑ 이상치 탐색은 중요 정보 도출 및 통계 분석을 위한 사전 작업등 다각도로 활용
☑ 이상치 탐색 목적과 자료 특성에 따라 다양한 방법론들이 개발
☑ 자료의 특성과 활용 목적을 고려한 이상치 탐색 방법론 적용 필요

Key Words
이상치, 극단치, 열외군
outliers, extreme value

1. 들어가며

이상치는 관측치들이 주로 모여 있는 곳에서 멀리 떨어져 있어 특정 그룹으로 분류되지 못하는 값으로, 정상군의 상한과 하한의 범위를 벗어난 자료를 의미한다. 이상치는 자료들의 분포에 따라 대푯값에 영향을 주므로 자료의 신뢰도와 정확도 향상을 위한 보장 측면에서 이상치 탐색과 처리는 중요한 과정이다.

이상치 탐색은 중요한 정보를 도출하거나 통계 분석을 위한 사전 작업으로 보건의료 영역에서 다양한 방법들이 활용되고 있다. 건강보험심사평가원에서도 심사, 진료 경향 모니터링 등 다양한 영역에서 이상 징후 감지, 중요 정보 탐색, 안정적인 결과 도출 목적으로 이상치 탐색 방법들이 적용되고 있다. 그러나 이상치는 정의 방법에 따라 열외군의 비율이 달라질 수 있으며, 동일한 방법에서도 자료의 분포에 따라 차이가 있을 수 있다.

이 글에서는 이상치 탐색 방법론들을 소개하고, 자료의 구조에 따라 방법론을 분류하여 자료 특성과 활용 목적에 적합한 분석 방법론을 검토하였다.[1)]

1) 이 글은 2019년 선정연, 김기영, 김진휘「이상치 탐색을 위한 통계적 방법과 활용 방안」연구보고서 내용의 일부를 재구성하여 작성된 것이다.

2. 이상치 탐색 방법

가. 이상치 탐색의 개념

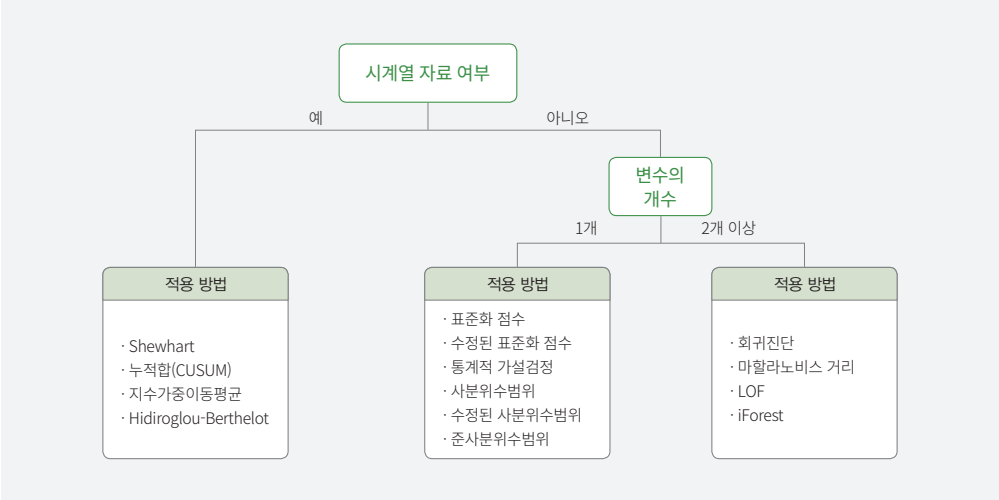
통계학 측면에서 이상치는 관측치들이 주로 모여 있는 곳에서 멀리 떨어져 있는 관측치로 정의된다(Kim, 2006). 이상치는 입력 오류 등 자료 오염으로 인해 발생한 비합리적인 이상치와, 정확하게 측정되었으나 다른 자료들과 전혀 다른 경향이나 특성을 보이는 합리적인 이상치로 구분할 수 있다. 이상치 탐색은 분석 결과의 안정성을 위한 이상치 제거와 자료 대체, 중요한 정보 탐색을 위한 목적으로 활용된다.

이상치 탐색 방법론은 다양한 관점에서 분류가 가능하나, 일반적으로 자료의 크기, 차원, 구조와 같은 자료의 특성과 통계적 접근 방법을 기준으로 분류할 수 있다(표 1)[그림 1].

(표 1) 접근 방법에 따른 이상치 탐색 방법의 분류

접근 방법	이상치 탐색 방법 분류
자료의 크기	소표본, 대표본
자료의 차원	일차원, 이차원, 다차원
변수의 개수	일변량, 이변량, 다변량
목표 변수의 유무	지도 방법, 비지도 방법
통계적 방법	모수적 방법, 비모수적 방법, 준모수적 방법

자료: 건강보험심사평가원. 이상치 탐색을 위한 통계적 방법과 활용 방안. 2019.



[그림 1] 자료의 구조에 따른 이상치 탐색 방법의 분류

자료: 건강보험심사평가원. 이상치 탐색을 위한 통계적 방법과 활용 방안. 2019.

나. 이상치 탐색 방법

1) 단변량 자료에서 이상치 탐색

단변량 자료에서 이상치 탐색 방법은 변수가 하나인 자료에서 이상치 영역을 우선적으로 정의하여 이상치를 탐색하는 방법이다. 관찰치가 정의된 이상치 영역 포함 여부에 따라 판단되는 개념이므로 자료의 분포 형태를 확인한 후에 탐색 목적에 적합한 방법을 선택해야 한다.

[그림 2] 단변량 자료의 이상치 탐색 원리

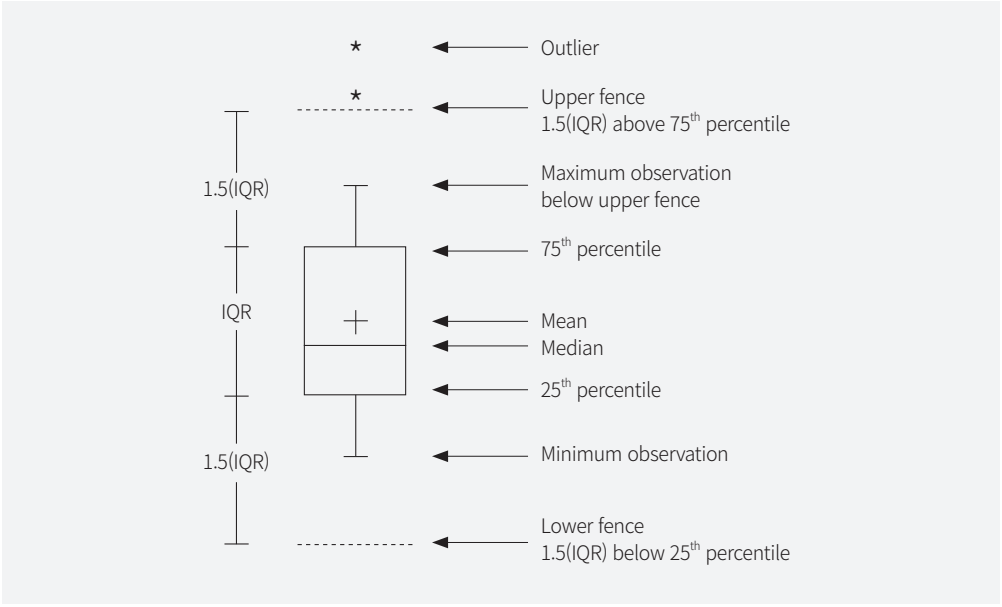
자료: 건강보험심사평가원. 이상치 탐색을 위한 통계적 방법과 활용 방안. 2019.

단변량 자료에서 이상치 탐색 방법으로는 표준화 점수(Z-score), 통계적 가설 검정, 사분위수 범위 등이 있다. 표준화 점수를 활용한 이상치 탐색 방법에서, 표준화 점수는 평균이 μ이고, 표준편차가 σ인 정규분포를 따르는 관측치들이 자료의 중심(평균)에서 얼마나 떨어져 있는지를 반영한다[그림 2]. 일반적으로 표준화 점수의 절댓값이 3보다 큰 경우에 이상치로 정의되지만, 절대적인 기준은 없으므로 경험에 근거하여 이상치 판단 기준을 결정하는 것이 합리적인 대안이다. 그러나 표준화 점수는 평균과 표준편차에 의존하므로, 산출 과정에 이상치의 영향을 받는다는 제한점이 있다. 표준화 점수의 문제점을 보완하기 위해 평균 대신 중앙값과 중앙값 절대편차(median absolute deviation)를 이용하는 수정된 표준화 점수(modified Z-score)가 있다. Iglewicz와 Hoaglin(1993)은 수정된 표준화 점수의 절댓값이 3.5보다 큰 경우에 이상치로 판단하는 것을 제안하였다.

통계적 가설검정을 활용한 이상치 탐색은 최솟값 혹은 최댓값의 이상치 여부에 대한 검정으로, 이상치로 판단된 관측치를 제외해 나가면서 더 이상 이상치가 존재하지 않을 때까지 반복적으로 검정을 수행하여 이상치를 정의하는 방법이다. 통계적 가설검정을 활용한

이상치 탐색 방법은 딕슨 Q검정(Dixon Q-test), Grubbs test, Generalized ESD(extreme studentized deviate) 검정, 카이제곱 검정(Chi-square test) 등이 있다.

사분위수 범위(interquartile range, IQR)를 활용한 이상치 탐색은 상자그림에서 사분위수 범위의 1.5배를 초과하는 관측치는 이상치, 3배를 초과하는 관측치는 극단적 이상치로 정의하는 방법이다. 상자그림은 최솟값, 최댓값, 제1사분위수(Q_1), 제2사분위수(Q_2), 제3사분위수(Q_3)를 활용하여 자료를 시각적으로 요약한 그래프이다. 상자그림에서 표현되는 최솟값과 최댓값은 이상치를 제외한 자료의 최솟값과 최댓값을 의미하고, 사분위수 범위는 제3사분위수와 제1사분위수의 차이를 말한다[그림 3]. 기존 사분위수 범위를 일반화한 수정된 사분위수 범위와 일부 사분위수 범위를 변형한 준사분위수 범위도 이상치 탐색 방법으로 사용된다.



[그림 3] 상자그림의 사분위수 범위를 활용한 이상치 정의

자료: 건강보험심사평가원. 이상치 탐색을 위한 통계적 방법과 활용 방안. 2019.

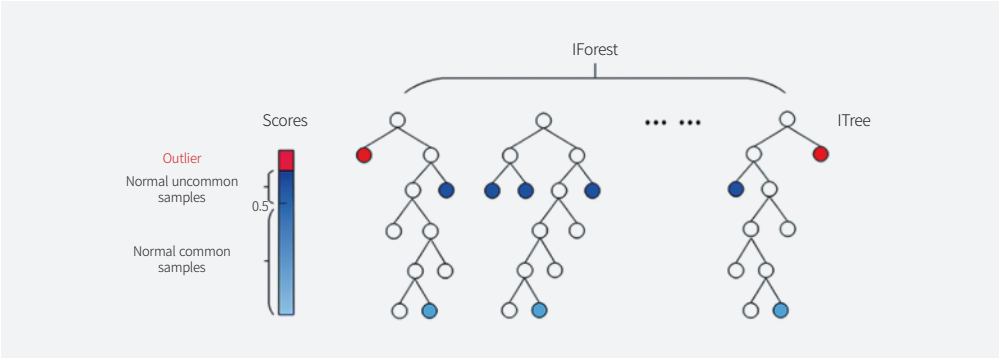
2) 다변량 자료에서 이상치 탐색

다변량 자료에서 이상치 탐색 방법은 연관성이 존재하는 2개 이상의 변수 정보를 활용하여 관측치 사이의 거리, 밀도 등을 기반으로 이상치를 탐색하는 방법이다. 다변량 자료에서 이상치 탐색 방법에는 회귀진단(regression diagnostics), 마할라노비스 거리(Mahalanobis Distance), LOF(Local Outlier Factor), iForest(Isolation Forest) 등이 있다. 회귀진단은 추정된 회귀식에 대한 전반적인 검토를 의미하며, 회귀식 추정에 영향을 미치는 극

단치를 탐색하는 것을 포함한다. 회귀진단을 통한 이상치 탐색 방법의 통계량으로는 레버리지(leverage), 표준화 잔차(standardized residual), 스튜던트 잔차(studentized residual), 외적 스튜던트 잔차(studentized deleted residual), 쿡의 거리(Cook's distance), DFFITS(difference of fits), DFBETAS(difference of betas) 등이 있다.

레버리지는 독립변수의 각 관측치가 독립변수들의 평균에서 떨어진 정도를 반영하는 통계량으로, 0과 1 사이 값을 가지며, 일반적으로 레버리지 평균의 2~4배를 초과하는 관측치를 이상치로 정의한다. 표준화 잔차는 추정된 회귀모형에 의해 산출된 예측치와 실제로 측정된 관측치의 차이를 의미하는 잔차를 표준화한 통계량으로, 일반적으로 표준화 잔차의 절대값이 2 또는 3을 초과하는 관측치를 이상치로 정의한다. 스튜던트 잔차는 잔차를 잔차의 표준오차로 나눈 통계량으로 t-분포를 기반으로 이상치를 탐색하며, 절대값이 3 또는 4를 초과하면 이상치로 의심한다. 스튜던트 제외 잔차는 해당 관측치를 제외하여 추정된 회귀모형으로부터 산출한 스튜던트 잔차를 의미하며, t-분포의 값을 기준으로 해당 관측치를 이상치로 결정한다. 쿡의 거리는 추정된 회귀모형에 대한 각 관측치들의 전반적인 영향력을 측정하기 위해 잔차와 레버리지를 동시에 고려한 척도로, 쿡의 거리가 1보다 큰 경우 이상치로 판단한다. DFFITS 통계량은 모든 관측치를 활용하여 추정된 회귀모형 예측치와 해당 관측치를 제외한 후 추정된 회귀모형의 예측치 변화 정도를 측정하는 방법으로 DFFITS 값이 클수록 이상치일 가능성이 높다. DFBETAS 통계량은 모든 관측치를 활용하여 추정된 회귀모형의 회귀계수와 해당 관측치를 제외한 후 추정된 회귀모형의 회귀계수 변화 정도를 측정하는 방법으로 자료의 수가 적은 경우 DFBETAS 절댓값이 1, 자료의 수가 많은 경우 절댓값이 $2/\sqrt{n}$ 보다 크면 이상치로 판단한다.

마할라노비스 거리는 자료의 분포를 고려한 거리 측도이며, 관측치가 평균으로부터 벗어난 정도를 측정하는 통계량이다. 변수의 개수를 기준으로 카이제곱 분포의 임계값을 초과하는 경우 이상치로 정의된다. LOF는 관측치 주변의 밀도와 근접한 관측치 주변 밀도의 상대적인 비교를 통해 이상치를 탐색하는 방법으로 값이 1에 가까울수록 주변의 관측치와 유사한 밀도임을 의미한다. 1보다 커질수록 밀도가 낮음을 의미하므로 이상치로 의심한다. iForest기법은 관측치 사이의 거리 또는 밀도에 의존하지 않고, 데이터마이닝 기법인 의사결정나무(decision tree)를 이용하여 이상치를 탐지하는 기법이다. 분류모형을 생성하여 모든 관측치를 고립시켜 나가면서 분할 횟수로 이상치를 탐색하며, 모형에서 적은 횟수로 Leaf 노드에 도달하는 관측치일수록 이상치일 가능성이 크다고 판단한다[그림 4].



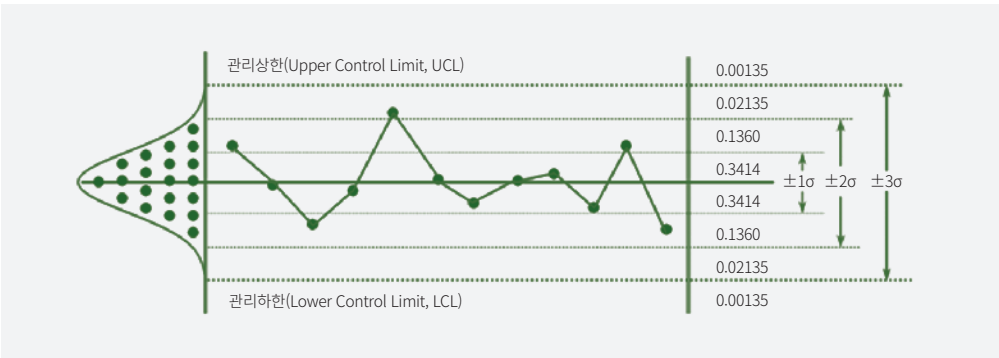
[그림 4] 이상치 탐색을 위한 iForest 방법

자료: Chen et al. Representative subset selection and outlier detection via isolation forest. 2016.

3) 시계열 자료에서 이상치 탐색

시계열 자료에서 이상치 탐색은 대부분 모형 적합을 통해 관측치 사이의 연관성을 제거한 잔차를 산출한 후, 잔차에 대해 방법을 적용한다. 감시(surveillance) 목적의 통계적 공정관리(Statistical Process Control) 기법은 시계열 자료에서 이상치 탐색에 활용할 수 있다. 대표적인 방법으로는 슈하르트(Shewhart) 관리도, 누적합(cumulative Sum) 관리도, 지수가중이동평균(exponentially weighted moving average) 방법, Hidiroglou-Berthelot 방법 등이 있다.

슈하르트 관리도는 관리하고자 하는 값을 중심선(central line)으로 하여 관리 하한(lower control limit)과 관리 상한(upper control limit)을 설정하고, 시간의 경과에 따라 관측값을 표시하는 통계적 과정이다. 관리 상한과 관리 하한을 벗어나면 이상치로 판단하며, 관리 모수와 관리 통계량에 따라 다양한 관리도가 존재한다[그림 5].



[그림 5] 슈하르트 관리도 개요

자료: 건강보험심사평가원. 이상치 탐색을 위한 통계적 방법과 활용 방안. 2019.

누적합 관리도는 과거부터 최근까지 통계량의 누적합을 사용하는 방법으로, 작은 변화가 발생하더라도 그 효과가 누적되어 관리 통계량에 반영된다. 이 방법은 작은 추세 변화를 감지하기 위하여 변이가 큰 자료보다 안정적인 자료에서 유용하게 사용되며, 의사결정 구간을 벗어나는 시점을 이상치로 정의한다. 지수가중이동평균 방법은 최근 관측값에 큰 가중치를 주어 최근 변화를 반영하여 이상치를 탐지하는 방법이다. 관리 상한과 하한을 설정하여 관리 한계를 벗어나는 시점의 관측치를 이상치로 판단한다. Hidroglou-Berthelot 방법은 이전 시점과 현재 시점의 비로 이상치를 탐지하는 방법으로 단위의 크기(size of unit)를 고려하여 이상치에 대한 허용 범위를 정의하는 방법이다.

3. 나가며

이 연구는 '이상치를 어떻게 탐색할 것인가'라는 문제를 중심으로 체계적인 문헌 검토 과정을 통해 자료의 특성과 목적에 따른 이상치 탐색에 초점을 두고 방법을 검토 · 정리하였다. 이상치는 탐색 목적에 따라 조금은 다르게 정의되지만 일반적으로는 다른 관측치들과 일관성이 없는 것으로 나타나는 관측치나 관측치의 집합으로 정의된다(Barnett and Lewis, 1994).

이상치는 특정 지정된 그룹에 분류되지 못하는 값으로, 정상군의 상한과 하한의 범위를 벗어난 자료를 의미하며, 이상치 탐색은 질 높은 통계 분석 결과 도출에 있어 중요한 역할을 담당한다. 특히 이상치 정의 방법에 따라 열외군의 비율이 달라질 수 있으며, 동일한 방법을 적용하더라도 자료의 분포에 따라 차이가 있을수 있어, 적절한 탐색 방법 선정이 중요하다(Palmer and Reid, 2001).

이상치 탐색은 분석 결과의 안정성을 위한 이상치 제거와 자료 대체, 중요한 정보 탐색을 위한 목적으로도 활용이 가능하다. 이상치가 포함된 자료 분석으로 인해 모형의 오류, 편향된 결과가 도출 될 수 있으므로 일관성 있는 분석 결과를 산출하기 위해서는 이상치 탐색이 우선적으로 수행되어야 한다. 또한 이상치 탐색과정에 일부 극단치에 의해 이상치로 분류되어야 하는 값들이 정상 범주값으로 정의되는 가면효과(masking effect)와 정상 범주값이 이상치와 근접하여 동일하게 이상치로 판별되는 수렁효과(swamping effect)를 주의해야 한다.

이상치 탐색 방법은 보건의료 영역에서 다양하게 활용된다. 특히 진료 경향을 모니터링하거나 동일한 질병군으로 분류하는 과정에 다빈도로 활용된다. 그리고 평균과 같은 대푯값을 산출 시 열외군을 제외하는 과정에 이상치 탐색 방법을 적용하여 자료의 특성을 반영하는 대푯값을 산출할 수 있다.

그러나 이상치를 탐색하는 방법들이 많이 개발 · 활용됨에도 불구하고 대부분의 경우 일반적으로 많이 사용되는 전통적인 이상치 처리 방법 몇 가지를 단순하게 사용하고 있으며,

전문적인 특정 분야에서는 기존 과거에 사용된 방법과 절차를 반복하여 형식적으로 사용하고 있다. 급속히 발전하는 과학 기술과 다양하고 방대해지는 자료의 변화에서 과거의 자료에 기반한 이상치 처리 방법을 동일하게 적용한다는 점은 개선이 필요한 부분이다. 모든 가정과 상황을 맞출 수 없다는 문제점을 가지고 있고, 목적과 자료 특성에 기반하여 이상치 처리 방법을 선택해야 하기 때문에 최상의 이상치 처리 방법은 없다. 따라서 자료 분석의 중요성을 인식하고 이상치 탐색 및 처리 방법 선정 시 자료의 특성을 고려하고, 활용 목적에 맞게 선택하여 적용함으로써 과거보다 더 좋고 효율적인 방법을 찾기 위한 노력이 필요하다.

참고문헌

선정연, 김기영, 김진휘. 이상치 탐색을 위한 통계적 방법과 활용 방안. 건강보험심사평가원. 2019.

Barnett V, Lewis T. Outliers in Statistical Data. 3rd edition. New York: John Wiley & Sons. 1994.

Chen W, Yun Y, Wen M, Lu H, Zhang Z, Liang Y. Representative subset selection and outlier detection via isolation forest. Analytical Methods. 2016;8(39):7225-7231.

Iglewicz B, Hoaglin D. How to detect and handle outliers. Milwaukee: ASQC Quality Press. 1993.

Kim J. Weight Reduction Method for Outlier in Survey Sampling. The Korean Communications in Statistics. 2006;13(1):19-27.

Palmer G, Reid B. Evaluation of the performance of diagnosis-related groups ans similar casemix systems: methodological issues. Health Serv Manage Res. 2001;14(2):71-81.