ORIGINAL ARTICLE

투고일 : 2017. 5. 11 심사일 : 2017. 6. 19 게재확정일 : 2017. 6. 30

2

치의학 연구에서 이상치의 처리

연세대학교 치과대학 BK21 플러스 통합구강생명과학 사업단, 치의학 교육연구센터

김 기 열

ABSTRACT

Outlier detection in dental research

BK21 PLUS Project, Dental Education Research Center, Yonsei University College of Dentistry
Ki-Yeol Kim, Ph.D.

In clinical dental research, errors occur in spite of careful study design and conduct. Data cleaning procedures intend to identify and correct these errors or at least to minimize their influence on study. Outlier is the one of these errors. Outlier detection is the first step in data analysis process which has a serious effect in the field of dental research. Hence, this paper aims to introduce the methods to detect the outliers and to examine their influences in statistical data analysis.

Key words : outliers; dental research; interquartile range; Grubb's Test

Corresponding Author
Ki-Yeol Kim, Ph.D.
BK21 PLUS Project, Dental Education Research Center, Yonsei University College of Dentistry
250 Seongsanno, Seodaemun-gu, Seoul 120-752, Korea
Tel : +82-2-2228-3043, Fax : +82-2-392-2959, E-mail : kky1004@yuhs.ac

본 연구는 연세대학교 치과대학 2017년도 교수연구비에 의하여 이루어졌음 (6-2017-0029).

Ⅰ. 서론

데이터 분석 단계 중 시간이 가장 많이 소요되는 단계는 데이터 탐색(EDA, exploratory data analysis) 단계이며, Forbes 에서 인용한 CrowdFlower 의 설문 결과에 따르면 데이터 분석의 80% 정도는 데이터 수집 및 전처리 과정에 사용한다고 한다[1]. EDA 라고 하는 데이터 전처리 단계는 데이터 셋 확인, 결측치(missing values) 처리, 이상치(outliers) 처리 등을 포함하는 과정이다.

결측치란 일부 요인에서 관측값이 얻어지지 않은 것을 말하며, 이를 무시하고 관측된 자료만 분석하게 되면 편향(bias)이 발생할 수 있다. 따라서 결측치가 발생하면 적절한 방법에 의해 대체된 값으로 채워넣고 분석을 하기도 한다[2]. 이상치란 대부분의 데이터 값들과 동떨어진 관측치로, 분석결과를 왜곡할 가능성이 있는 값을 말한다. 이상치는 정도를 벗어난 값으로 잘못된 분석결과의 원인이 될 수 있으므로, 데이터 모델링이나 분석를 실행하기 전에 이상치를 찾는 것은 중요하다[3, 4]. 이상치가 데이터 오류나 노이즈로 간주되기는 하지만, 그것들도 중요한 정보를 포함할 수 있으므로 무조건 분석에서 제외시키는 것보다 결측치 처리 방법과 같이 대체값을 사용할 수 있다.

본 연구에서는 이상치를 찾는 방법을 소개하고, 통계 프로그램을 사용하여 예제 데이터로부터 이상치를 찾아본다. 또한, 이상치의 포함여부에 따른 분석결과를 비교하여 이상치의 효과를 확인해 본다. 통계 프로그램으로는 연구자들이 주로 사용하고 있는 SPSS 와 최근 치의학 분야에서도 사용자가 증가하고 있는 R 을 이용하였다.

Ⅱ. 이상치(outlier)

이상치(outlier) 란 잘못 평가된 값으로 잘못된 분석결과를 초래할 수 있는 값을 의미한다. 즉, 데이터의 전체적인 패턴에서 벗어난 관측값을 말한다.

1. 이상치 종류

이상치의 종류는 Univariate와 Multivariate 로 나눌 수 있다. 하나의 변수 분포에서 나타난 이상치는 univariate outlier라고 하며, multivariate는 n 개 변수에서 나타나는 outlier라고 생각하면 된다.

나이와 체질량지수를 예로 들어보자. 나이와 체질량지수 각 값에 대해서는 이상치가 나타나지 않는다(그림 1). 그런데, 나이와 체질량지수를 함께 산점도로 나타내면 세 개의 이상치를 확인할 수 있다. 이처럼 한

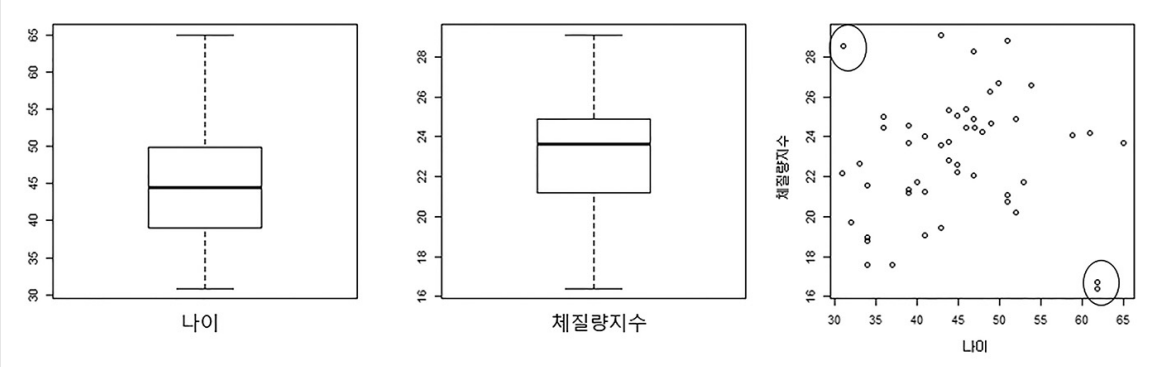

그림 1. 나이와 체질량지수 측정값을 탐색하는 그림. 그림의 왼편과 가운데는 나이와 체질량지수를 각각 요약한 상자그림이며, 오른편 그림은 나이와 체질량지수의 관계를 탐색한 산점도 이다.

개의 값에서 나타나지 않는 이상치가 n개의 값의 관계를 함께 고려했을 때 이상치로 나타날 수 있다.

2. 이상치의 원인

이상치가 발생하는 원인은 다음과 같다

* 데이터 입력 오류 : 데이터를 수집하는 과정에서 발생할 수 있는 에러를 말한다. 예를들어 100을 입력해야 하는데, 1000 을 입력하면 10배의 값으로 입력이 된다. 이렇게 입력된 값은 전체 데이터의 분포를 보면 쉽게 발견을 할 수 있다.
* 측정 오류 : 데이터를 측정하는 과정에서 발생하는 에러를 말한다. 예를 들어 몸무게를 측정하는데, 9개의 체중기는 정상 작동, 1개는 비정상 작동을 한다고 가정 할 때, 한 사용자가 비정상적으로 작동하는 체중계를 이용할 경우에 에러가 발생하게 된다.
* 실험 오류 : 실험을 할 때 생기는 에러를 말한다. 100미터 달리기를 하는데, 한 선수가 '출발' 신호를 못듣고 늦게 출발했다고 가정하자. 이때 그 선수의 기록은 다른 선수들보다 늦을 것이고, 그의 경기시간은 이상치가 될 수 있다. 즉, 실험조건이 동일하지 않은 경우에 발생할 수 있다.
* 고의적인 이상치 : self-reported measures 에서 나타나는 에러를 말한다. 예를들어 음주량을 묻는 조사가 있다고 가정하자. 대부분의 10대 들은 자신들의 음주량을 적게 기입할 것이고, 오직 일부만 정확한 값을 적을 것이다. 이런 경우, 정확하게 기입한 값이 이상치로 보일 수도 있다.
* 표본추출 에러 : 데이터를 샘플링하는 과정에서 나타나는 에러를 말한다. 대학 신입생들의 키를 조사하기 위해 샘플링을 하는데, 농구선수가 포함되었다면 농구선수의 키는 이상치가 될 수 있다. 이것은 샘플링을 잘못한 경우이다.

3. 이상치 검출

이상치는 그림을 이용한 탐색을 통해서 발견할 수 있다. 주로 상자그림(Box-plot), 히스토그램, 산점도(Scatter Plot)를 사용하며, 수치적으로는 다음의 기준에 의해 이상치를 찾을 수 있다.

이상치를 찾는 몇 가지 방법을 소개하고, R과 SPSS 를 사용하여 실행해본다.

3-1. IQR rule for outliers (Interquartile Range)[5]

상자그림은 데이터의 분포를 탐색하는 편리한 그래픽 방법이다. 상자그림은 중앙값과 하사분위, 상사분위(25^{th}, 75^{th} percentiles(Q1, Q3)로 정의된다)를 사용하여 작성되며, 상사분위 값과 하사분위 값의 차이를 사분위간 범위 (Interquartile range, IQR)라고 한다.

데이터 값들의 분포를 상자그림으로 탐색할 경우, 다음 그림과 같이 이상치를 발견할 수 있다(그림 2).

[예 1 : R] 순서대로 나열된 90개의 관측값들을 사용하여 이상치를 탐색하는 예이다.

30, 171, 184, 201, 212, 250, 265, 270, 272, 289, 305, 306, 322, 322, 336, 346, 351, 370, 390, 404, 409, 411, 436, 437, 439, 441, 444, 448, 451, 453, 470, 480, 482, 487, 494, 495, 499, 503, 514, 521, 522, 527, 548, 550, 559, 560, 570, 572, 574, 578, 585, 592, 592, 607, 616, 618, 621, 629, 637, 638, 640, 656, 668, 707, 709, 719, 737, 739, 752, 758, 766, 792, 792, 794, 802, 818, 830, 832, 843, 858, 860, 869, 918, 925, 953, 991, 1000, 1005, 1068, 1441

사분위간 (Q1, Q2, Q3) 구하기

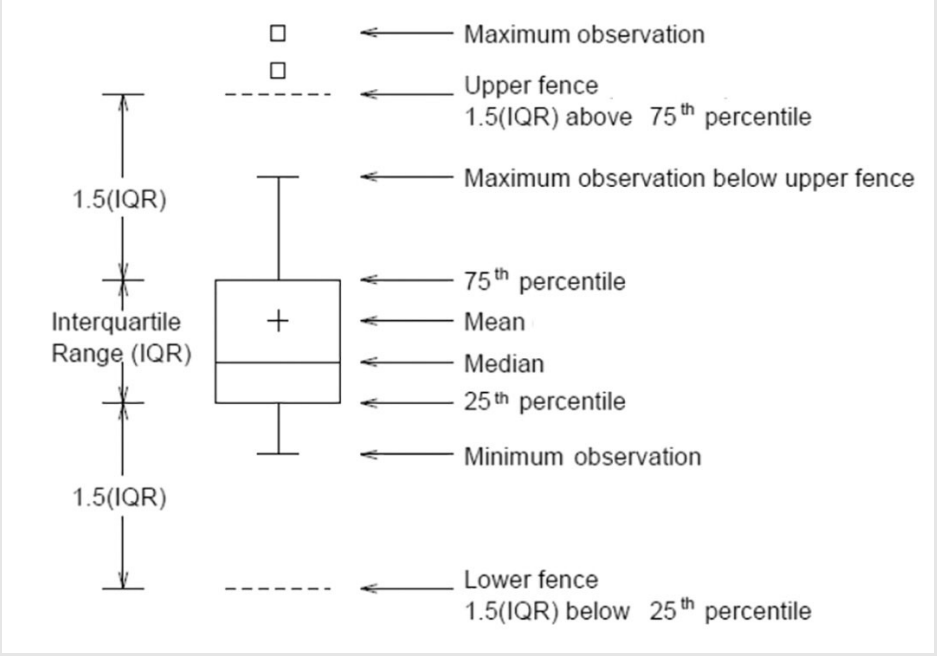

그림 2. 상자그림을 이용한 데이터의 분포를 탐색하는 그림. Median (Q2) 은 중앙값, 25th percentile (Q1) 과 75th percentile (Q3) 은 전체 데이터 값을 크기 순서대로 나열했을 때 25%, 75% 순위에 해당되는 값이며, IQR 은 interquartile range (Q3-Q1)로 이 두 값의 차이를 말한다. Q3+IQR x 1.5는 상한 극한값(upper fence), Q1-IQR x 1.5는 하한 극한값(lower fence)으로 이 범위를 벗어나는 값을 이상치라고 한다. 데이터의 최대값, 최소값이 상한 극한값과 하한 극한값 내에 위치하면 수염의 길이가 짧게 그려진다.

- R 에서 데이터 입력 후, summary 함수로 quantile 정보를 볼 수 있다.

```
〉 data=c(30, 171, 184, 201, 212, 250, 265, 270, 272, 289, 305, 306, 322, 322, 336, 346, 351, 370, 390, 404, 409, 411, 436, 437, 439, 441, 444, 448, 451, 453, 470, 480, 482, 487, 494, 495, 499, 503, 514, 521, 522, 527, 548, 550, 559, 560, 570, 572, 574, 578, 585, 592, 592, 607, 616, 618, 621, 629, 637, 638, 640, 656, 668, 707, 709, 719, 737, 739, 752, 758, 766, 792, 792, 794, 802, 818, 830, 832, 843, 858, 860, 869, 918, 925, 953, 991, 1000, 1005, 1068, 1441)
```

```
〉 summary(data)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 30.0   436.2   559.5   576.1   738.5  1441.0
```

데이터의 범위는 30.0 - 1441.0 이며, Q1= 436.2, Q2=559.5, 평균=576.1, Q3=738.5 이다.

- quantile 함수로도 볼 수 있다.

```
〉 quantile(data)
    0%    25%    50%    75%   100%
 30.00  436.25  559.50  738.50 1441.00
```

사분위간 범위(IQR, Interquartile range) 구하기

- IQR은 Q3 - Q1 값을 뜻한다.

> Q3 = quantile (data)[4]
> Q1 = quantile (data)[2]
> IQR = Q3 − Q1
> IQR
[1] 303

IQR Rule에서 이상치는 다음과 같이 정의된다.
이상치 > Q3 + IQR x 1.5
이상치 < Q1 − IQR x 1.5

R에서 IQR Rule을 이용한 이상치 검출방법

> upperOutlier = data[which(data > Q3+IQR x 1.5)]
> lowerOutlier = data[which(data < Q1−IQR x 1.5)]

> upperOutlier
[1] 1441
> lowerOutlier
numeric(0)

위의 결과로부터 사분위간 범위를 사용한 경우 최대값인 1441 이 이상치로 나타났다(그림 3).

데이터의 최대값은 1441이고 이 값이 이상치로 나타났으며, 이 값을 제외한 데이터의 범위는 30−1068 이었다. 따라서 상자그림의 수염은 최소값인 33, 최대값인 1068 로 그려졌다.

[예 1 : SPSS] SPSS 에서는 '데이터 탐색' 메뉴를 사용하여 이상치를 탐색할 수 있다.

분석−기술통계량−데이터 탐색

데이터의 90번째 값이 이상치임을 알 수 있다. 즉, 마지막 관측값인 1441 이 이상치이다(그림 4).

3−2. Grubb's Test[6]

이 방법은 측정된 데이터의 분산에 따라서 결정되는 방법이며, 'Maximum normed residual test' 라고도 불린다. Grubbs' test 는 데이터가 정규분포를 보일 때 사용된다.

1) Grubb's test의 가설

귀무가설 : 데이터에는 이상치가 없다.

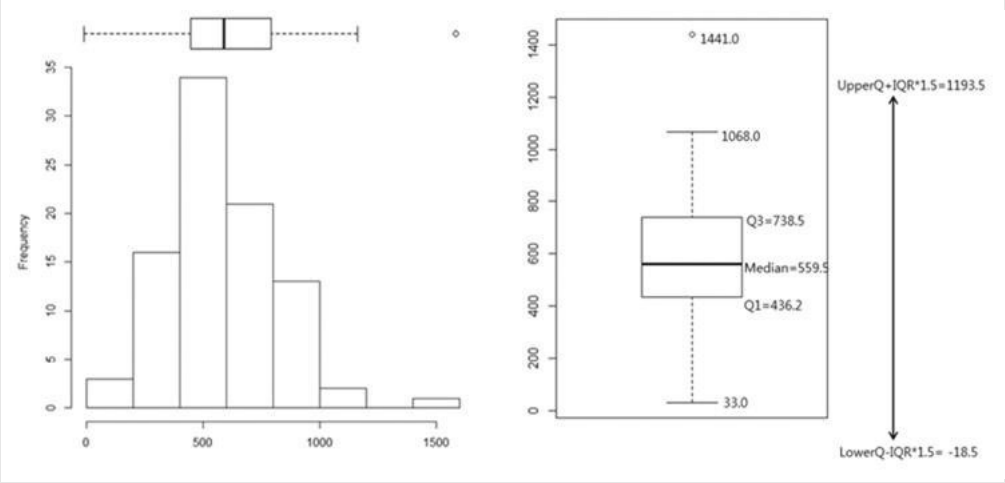

그림 3. 히스토그램과 상자그림을 함께 사용하여 데이터의 분포를 확인한 그림

백분위수

	백분위수						
	5	10	25	50	75	90	95
가중평균(정의 1) x	207.05	273.70	429.75	559.50	742.25	868.10	995.05
Tukey의 Hinges x			436.00	559.50	739.00		

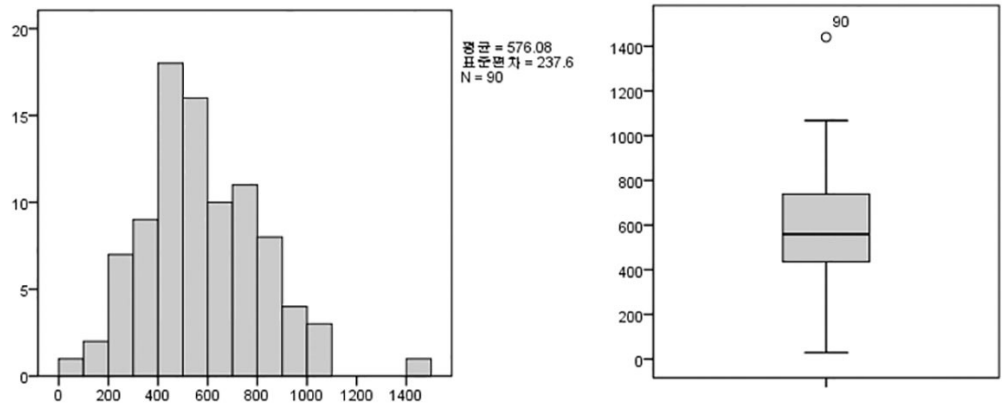

그림 4. SPSS 의 데이터 탐색기능을 사용하여 이상치를 찾는 그림. 오른편 상자그림에서 90번째 관측치는 Q3 + IQR x 1.5 (1193) 보다 큰 값을 가지므로 이상치 임을 알 수 있다.

대립가설 : 데이터에는 최소한 한 개의 이상치가 있다.

2) Grubb's test의 원리

데이터의 각 값들로부터 평균과의 차이가 가장 큰 값을 찾고, 그 값이 특정한 값보다 더 큰 지를 확인하는 방법이다. 여기서 평균과의 차이는 절대값을 이용하고, T 로 표기한다. T 는 다음과 같이 정의할 수 있다.

$T = \frac{\max|\bar{X} - X_i|}{s}$, $\bar{X}$ 는 데이터의 평균, X_i는 i번째 관측값, s는 표준편차이다.

유의수준에 따라 T 와 비교할 특정 값이 다음과 같이 계산되며, T 가 이 값보다 큰지를 검정하게 된다.

$T > \frac{N-1}{\sqrt{N}}\sqrt{\frac{t^2_{\alpha/(2N),N-2}}{N-2+t^2_{\alpha/(2N),N-2}}}$, N은 데이터의 개수, α는 유의수준을 나타낸다.

3) R에서 Grubb's Test를 이용한 이상치 검출방법

실행을 위하여 라이브러리 'outliers' 를 설치한다.

```
install.packages("outliers")
library(outliers)
```

- grubbs.test 는 최소한 한 개의 이상치가 있는지 확인한다.

```
> grubbs.test(data)

        Grubbs test for one outlier

data:  data
G = 3.64020, U = 0.84944, p-value = 0.007089
alternative hypothesis: highest value 1441 is an outlier
```

p=0.007089 로 귀무가설은 기각되고, 데이터는 최소한 한 개의 이상치를 포함하고 있음을 알 수 있다.

데이터 셋에서 이상치 목록을 확인하기 위하여 다음의 추가적인 함수의 정의가 필요하다. grubbs.flag 함수를 정의하고, 실행시키는 모든 이상치를 선별해 낼 수 있다(그림 5).

이 외에 백분위 수에서 5^{th} ~ 95^{th} 범위에서 벗어나는 값을 이상치로 간주하기도 하며. 평균으로 부터 3배의 표준편차를 벗어난 값을 이상치로 간주하기도 한다. 이 경우, 평균과 표준편차는 이상치에 매우 민감한 통계량이므로 이를 이용하는 것은 매우 비효율적일 수 있다. 따라서, 평균과 표준편차보다 의미있는 통계량으로써 중위수(median)와 MAD(median absolute deviation) 을 사용하기도 한다.

Ⅲ. 이상치가 분석결과에 미치는 영향

데이터 샘플에서 관찰된 한 값이 그 외 다른 관측값들과 거리가 있을 때 이상치라고 한다. 이것은 측정된 데이터들의 가변성, 변동성(variability) 때문일 수 있고, 실제로 잘못된 실험에 의한 에러일 수도 있다. 후자의 경우에는 분명히 데이터 분석 이전에 이상치를 제거 해야 한다

이상치는 데이터 분석이나, 통계 모델링의 결과에 심각하게 변화를 줄 수 있다. 이상치가 주는 영향에 대해서 정리하면 다음과 같다.

- 분산이 증가하고 통계분석시 검정력이 감소한다.
- 이상치가 랜덤하게 분포하지 않으면(non-randomly), 데이터의 정규성(normality)이 감소한다.
- 회귀분석, 분산분석등 통계적 분석시 통계적 모형에 대한 가정에 영향을 줄 수 있다.

예를 들어 데이터가 다음과 같을 때,

```
grubbs.flag <- function(x) {
  outliers <- NULL
  test <- x
  grubbs.result <- grubbs.test(test)
  pv <- grubbs.result$p.value
  while(pv < 0.05) {
    outliers <- c(outliers,as.numeric(strsplit(grubbs.result$alternative," ")[[1]][3]))
    test <- x[!x %in% outliers]
    grubbs.result <- grubbs.test(test)
    pv <- grubbs.result$p.value
  }
  return(data.frame(X=x,Outlier=(x %in% outliers)))
}
```

```
## 실행
> grubbs.flag(data)


## 결과
X Outlier
1     30   FALSE
2    171   FALSE
3    184   FALSE
4    201   FALSE
...
88 1005   FALSE
89 1068   FALSE
90 1441    TRUE
```

그림 5. 구체적으로 몇 번째 데이터가 이상치 인지 확인하는 단계를 보여주는 그림. 실행결과 90번째 자료인 1441 이 이상치임을 확인할 수 있다.

이상치가 없는 경우	이상치를 포함한 경우
4,4,5,5,5,5,6,6,6,7,7	4,4,5,5,5,5,6,6,6,7,7,300
Mean=5.45	Mean=30.00
Median=5.00	Median=5.50
Mode=5.00	Mode=5.00
Standard deviation=1.04	Standard deviation=85.03

이상치의 포함 여부에 따라 평균, 표준편차가 상당히 다른 것을 볼 수 있다. 이상치가 없는 경우 평균 5.45, 표준편차 1.04인 반면, 이상치가 포함되는 경우에는 평균 30, 표준편차 85.03 이 된다. 이상치에 의해서 추정한 값이 완전히 바뀐다는 것을 알 수 있다. 이러한 값들은 통계적인 분석을 할 때 영향을 주게 된다.

[예 2 :R] 나이와 BMI 의 관계를 알아보는 데이터에서 두 요인간의 관계를 탐색한 자료이다(그림 6).

다음은 나이와 BMI 의 관계를 R을 사용하여 회귀분석을 실행한 결과이다(표 1).

이상치를 포함 한 경우, Age 는 BMI 에 대한 유의한 요인이 아니었으나(p=5907), 이상치를 제거하고 분석한 결과, Age 는 BMI 에 대하여 강한 영향력을 갖는 요인이었다(p=0.004)

[예 3 : SPSS] 교정치료 예정인 환자 40명을 대상

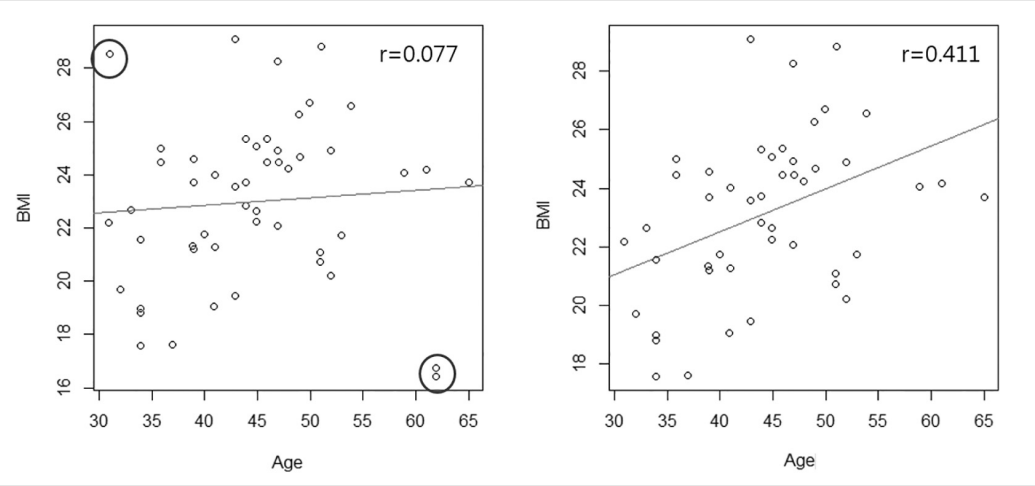

그림 6. 나이와 BMI 의 관계를 알아보는 그림. r은 상관계수를 나타낸다. 왼편은 이상치가 3 개 포함된 경우이고 오른편 그림은 이상치를 제거한 경우이다. 이상치를 제거함으로써 두 요인간의 상관관계가 증가하였다(0.077 에서 0.411). 직선은 두 변수간의 관계를 보여준다. 왼편 그림은 두 변수간의 관계가 거의 없어서 수평의 형태를 보이며, 오른편 그림은 양의 관계를 보여준다.

표 1. 나이와 BMI 의 회귀분석 결과. 이상치 포함여부에 따른 결과를 비교하였다.

이상치가 포함된 경우	이상치를 제거한 경우
〉 anova(lm(BMI~Age))	〉 anova(lm(BMI[-c(1,6,49)]~Age[-c(1,6,49)]))
Analysis of Variance Table	Analysis of Variance Table
Response: BMI	Response: BMI[-c(1, 6, 49)]
Df Sum Sq Mean Sq F value Pr(〉F)	Df Sum Sq Mean Sq F value Pr(〉F)
Age 1 2.78 2.7804 0.2932 0.5907	Age[-c(1, 6, 49)] 1 58.224 58.224 9.149 0.004103 **
Residuals 48 455.17 9.4827	Residuals 45 286.378 6.364

으로 구강위생교육을 실시한 다음, 두 군으로 나누어 한 군은 전동칫솔을, 다른 한 군은 수동칫솔을 사용하게 하였다. 6개월 후에 치태지수를 측정하여 차이가 생기는지 알아보는 연구이다. 데이터는 두 군의 구강위생교육을 시작하기 전의 치태지수 이며, 두 군간의 차이가 없어야 한다. 실제로 두 집단의 치태지수에 차이가 없는지 검정해 보았다(표 2)[9].

분석하기 전, 이상치가 있는지 탐색해 보았다.

표 2. 진동칫솔과 수동칫솔을 사용한 후 치태지수 데이터. n은 환자수를 나타낸다.

전동칫솔, n = 19	0.46, 1.33, 0.74, 0.26, 0.55, 0.46, 0.33, 0.49, 0.63, 0.35, 0.24, 0.31, 0.38, 0.46, 0.43, 0.55, 0.14, 0.44, 0.16
수동칫솔, n = 21	0.66, 0.69, 0.9, 0.7, 1.14, 0.51, 0.83, 0.45, 0.74, 0.55, 0.34, 0.14, 0.34, 0.57, 0.51, 0.63, 0.28, 0.46, 0.2, 0.23, 0.44

기술통계

칫솔종류			통계량	표준오차
치태지수	전동칫솔	평균	.4584	.05976
		평균의 95% 신뢰구간 하한	.3329	
		상한	.5840	
		5% 절사평균	.4277	
		중위수	.4400	
		분산	.068	
		표준편차	.26048	
		최소값	.14	
		최대값	1.33	
		범위	1.19	
		사분위수 범위	.24	
		왜도	2.132	.524
		첨도	6.667	1.014
	수동칫솔	평균	.5386	.05414
		평균의 95% 신뢰구간 하한	.4256	
		상한	.6515	
		5% 절사평균	.5278	
		중위수	.5100	
		분산	.062	
		표준편차	.24808	
		최소값	.14	
		최대값	1.14	
		범위	1.00	
		사분위수 범위	.35	
		왜도	.511	.501
		첨도	.298	.972

백분위수

			백분위수						
		칫솔종류	5	10	25	50	75	90	95
가중평균(정의 1)	치태지수	전동칫솔	.1400	.1600	.3100	.4400	.5500	.7400	.
		수동칫솔	.1460	.2060	.3400	.5100	.6950	.8860	1.1160
Tukey의 Hinges	치태지수	전동칫솔			.3200	.4400	.5200		
		수동칫솔			.3400	.5100	.6900		

그림 7. 전동칫솔과 수동칫솔을 사용한 경우 치태지수를 요약한 그림

SPSS 를 사용하여 분석한 결과이다(그림 7).

두 군의 치태지수의 범위는 (0.14, 1.33), (0.14, 1.14) 이며, 사분위간 범위는 (0.31, 0.55), (0.34, 0.70) 이다. IQR rule 에 의하여, 전동칫솔은(-0.05, 0.91), 수동칫솔은 (-0.20, 1.24) 외의 데이터가 이상치이며, 전동칫솔 자료의 2번째 자료가 이상치에 해당함을 보여준다(그림 7, 8).

이상치를 포함한 상태로 전동칫솔, 수동칫솔을 사용한 환자군 간의 치태지수를 t-test 를 사용하여 비교한 결과이다(그림 9).

분석결과, p 값은 0.325(등분산성이 만족되는 경우)이며, 두 군간에 차이가 없음을 나타낸다.

이상치를 제거하고 분석한 결과는 다음과 같았다(그림 10).

p값은 0.066 으로, 유의성의 크기가 많이 변함을 알 수 있다. 이는 이상치 제거에 의해 전동칫솔을 사용한 환자집단의 치태지수 값의 표준편차가 감소하고, 이에 따라 검정통계량의 절대값이 증가하였기 때문이다(-0.997 에서 -1.895).

Ⅳ. Outlier 처리 방법[7)]

데이터에서 이상치를 포함한 채로 분석을 하게 되

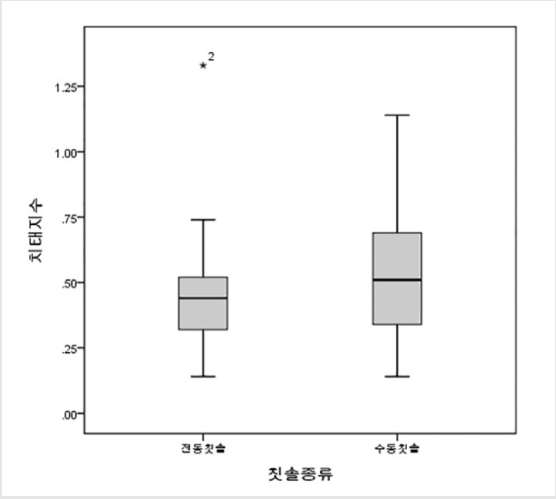

그림 8. 전동칫솔, 수동칫솔을 사용한 환자들의 치태지수를 비교한 그림. 전동칫솔을 사용한 환자 중, 2 번째 값이 이상치임을 알 수 있다. 이 값은 1.33 이다.

집단통계량

	칫솔종류	N	평균	표준편차	평균의 표준오차
치태지수	전동칫솔	19	.4584	.26048	.05976
	수동칫솔	21	.5386	.24808	.05414

독립표본 검정

		Levene의 등분산 검정		평균의 동일성에 대한 T 검정						
		F	유의확률	t	자유도	유의확률 (양측)	평균차이	차이의 표준오차	차이의 95% 신뢰구간 하한	차이의 95% 신뢰구간 상한
치태지수	등분산을 가정함	.296	.589	-.997	38	.325	-.08015	.08043	-.24297	.08267
	등분산을 가정하지 않음			-.994	37.149	.327	-.08015	.08063	-.24350	.08320

그림 9. 이상치를 포함한 상태로 데이터를 분석한 결과.

집단통계량

	치솔종류	N	평균	표준편차	평균의 표준오차
치태지수	전동칫솔	18	.4100	.15707	.03702
	수동칫솔	21	.5386	.24808	.05414

독립표본 검정

		Levene의 등분산 검정		평균의 등일성에 대한 T 검정						
		F	유의확률	t	자유도	유의확률 (양측)	평균차이	차이의 표준오차	차이의 95% 신뢰구간 하한	차이의 95% 신뢰구간 상한
치태지수	등분산을 가정함	2.903	.097	-1.895	37	.066	-.12857	.06784	-.26602	.00888
	등분산을 가정하지 않음			-1.960	34.265	.058	-.12857	.06558	-.26182	.00467

그림 10. 이상치를 제거하고 데이터를 분석한 결과

면, 분석결과의 정확성이 떨어지게 되므로 적절한 방법에 의한 이상치 처리가 필요하다.

- 삭제(Deleting Observations) : 이상치로 판단되는 관측값을 제외하고 분석하는 방법으로, 추정치의 분산은 작아지지만 실제보다 과소(또는 과대) 추정되어 편의가 발생할 수 있다. 이상치를 제외시키기 위해 양 극단의 값을 trimming 하기도 한다. 이상치 자료도 실제 조사된 수치이므로 이상치를 제외하는 것은 현실을 제대로 반영하는 방법으로 적절하지 않을 수도 있다.
- 대체법(Imputation) : 이상치 값을 평균이나 중앙값 등으로 대체하는 방법이다. 대체법은 '데이터의 결측치(missing value) 처리' 에 관한 내용으로 추후 정리하기로 한다.
- 변환(transformation) : 데이터의 변환은 극단적인 값으로 인해 이상치가 발생했다면 자연로그를 취해서 값을 감소시키는 방법으로 실제 값을 변형하는 것을 말한다(그림 11).
- 분류하여 처리 : 만약 이상치가 많을 경우에 서로 다른 그룹으로 통계적인 분석을 실행한다(그림 12, 13). 각각의 그룹에 대해서 통계적인 모형을 생성하고, 결과를 결합(combine)하는 방법을 사용한다[8].

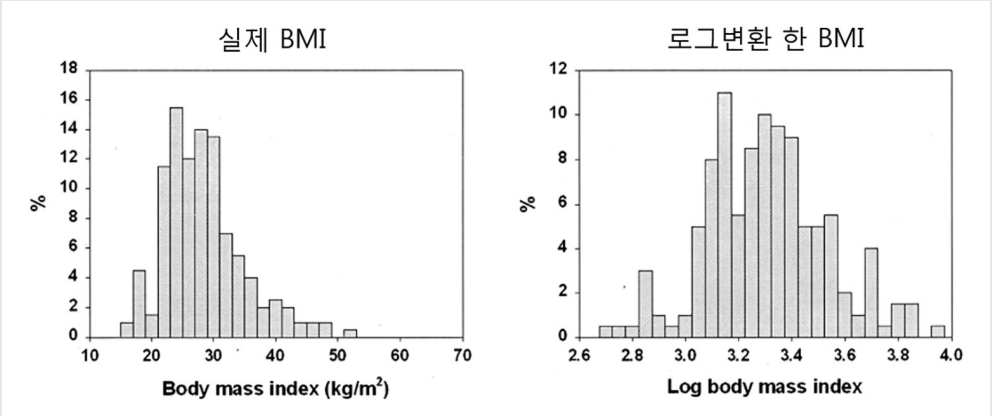

그림 11. 데이터를 로그변환하고 결과를 비교한 그림. 왼편은 극단적인 BMI 값에 의해 전체 데이터의 분포가 오른쪽으로 길게 기울어져 보이며, 로그변환에 의해 극단적으로 큰 값이 나타나지 않고 평균을 중심으로 대칭의 형태로 변환된 것을 볼 수 있다.

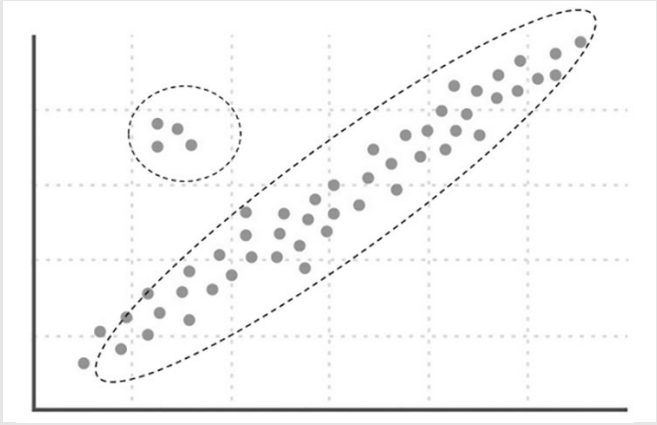

그림 12. 여러 개의 이상치가 하나의 그룹을 형성하는 데이터의 형태를 보여주는 그림

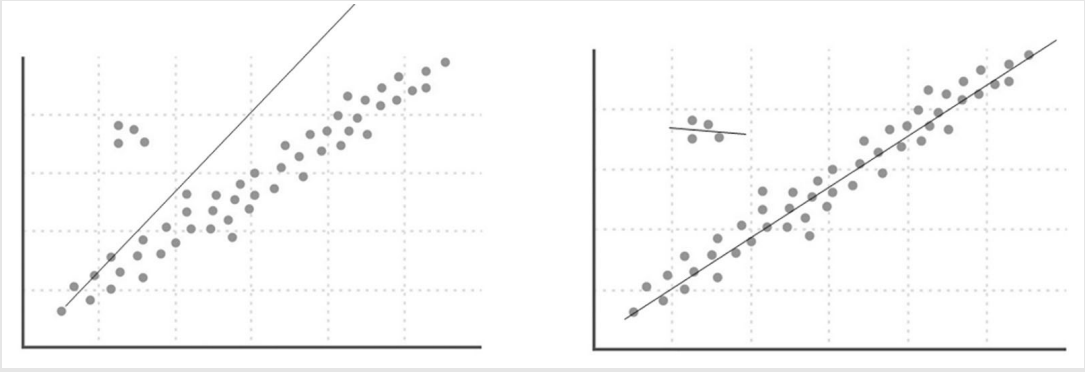

그림 13. 이상치가 많은 경우 두 변수간의 관계를 직선으로 표현한 그림. 하나의 데이터 셋으로 분석한 경우와 서로 다른 그룹으로 하여 각각 분석한 경우를 보여준다.

Ⅴ. 결론

이상치는 데이터의 총합이나 평균, 표준편차등을 추정할 때 큰 영향을 미치게 되고, 이 값들은 통계분석 결과를 왜곡시키므로 이상치의 영향을 감소하기 위한 여러 가지 방법들이 연구되어 왔다. 가장 간단한 방법으로는 이상치를 삭제하고 분석하는 것이다. 또는, 데이터 값을 로그변환 함으로써 극단적인 값들의 효과를 감소시키는 방법도 있다. 변환한 데이터를 분석하는 경우에는 해석에 주의를 기울여야 한다. 또한, 이상치를 다른 값으로 대체하는 방법이 있는데, 가장 간단한 방법은 평균이나 중앙값으로 대체 하는 것이다.

이상치가 측정 또는 입력 오류에 의해서 발생한 경우에는 해당 관측치를 제거하면 되지만, 데이터가 절대적으로 적은 경우에는 제거하는 방법으로 이상치를 처리하면 관측치가 적어지는 문제가 발생한다. 또한, 이상치도 일단 측정된 자료값이므로 제거하는 것은 바람직하지 않다고 생각된다. 그러나 이상치가 추정값에 미치는 영향이 작지 않으므로 이상치의 영향을 감소시키기 위한 적절한 처리방법이 필요하다.

참 고 문 헌

1. CrowdFlower, Data Science. 2016.
2. Yun, S.-C., Imputation of Missing Values. J Prev Med Public Health, 2004. 37(3): p. 209-211.
3. Graham Williams, R.B., Hongxing He, Simon Hawkins and Lifang Gu, A Comparative Study of RNN for Outlier Detection in Data Mining, in CSIRO Technical Report. 2002, CSIRO.
4. Hancong Liu, S.S., Wei Jiang, On-line outlier detection and data cleaning. Computers & Chemical Engineering, 2004. 28(9): p. 1635-1647.
5. Tukey, J.W., Exploratory data analysis. 1977.
6. Grubbs, F.E., Procedures for detecting outlying observations in samples. Technometrics 1969. 11(1): p. 1-21.
7. Ray, S. A Comprehensive Guide to Data Exploration. Available from: https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#two.
8. V. Deneshkumar, K.S., M. Manikandan, Identification of Outliers in Medical Diagnostic System Using Data Mining Techniques. International Journal of Statistics and Applications, 2014. 4(6): p. 241-248.
9. 임회정, SPSS 를 이용한 치의학 통계 입문 및 자료분석, 2008. 나래출판사.