Journal of the Korean Data & Information Science Society
2012, **23**(4), 657–663

http://dx.doi.org/10.7465/jkdi.2012.23.4.657
한국데이터정보과학회지

그럽 및 코크란 검정을 이용한 임상자료의 이상치 판단[†]

손기철[1] · 신임희[2]

[12]대구가톨릭대학교 의과대학

접수 2012년 5월 10일, 수정 2012년 5월 29일, 게재확정 2012년 6월 7일

요약

많은 분야에서 수집된 자료 중 데이터의 범위에서 많이 벗어난 아주 작은 값이나 아주 큰 값인 아상치가 종종 존재한다. 이런 이상치의 발생원인은 자료의 입력과정에서의 오류 또는 응답 과정에서 응답자의 특이한 답변 때문이다. 만약 자료에서 이상치가 존재할 경우 자료의 요약값인 평균과 분산에 많은 영향을 미쳐서 잘못된 정보가 산출된다는 문제점이 있다. 따라서 연구자는 자료에서 이상치가 존재하는지를 주의깊게 살펴보아야 한다. 특히 사람을 대상으로 실시한 임상자료의 경우 자료의 비용측면에서나 결과의 일관성 측면에서 이상치의 판단은 더욱 중요한 문제이다. 따라서 본 논문에서는 이상치를 판단하는 방법인 Grubb 검정과 Cochran 검정을 이용하여 임상자료에서의 이상치를 판단하는 방법을 소개하고자 한다.

주요용어: 그럽 검정, 이상치, 임상자료, 코크란 검정.

1. 서론

설문조사나 면접조사 등으로 수집된 자료에서 빈번하게 일어나는 문제는 자료의 값 중 아주 작은 값이거나 아주 큰 값인 이상치에 대한 것이다. 이상치는 자료를 입력하는 과정에 발생할 수 있으며 응답자가 응답을 하는 과정에서도 발생할 수 있다. 예를 들어 어떤 설문조사의 문항에 대한 응답이 1에서 5까지의 러커트 척도라고 가정하자. 그렇다면 자료의 값이 1에서 5사이의 값이어야 하는데도 불구하고 23, 43 등의 값이 존재한다면 이것은 자료의 입력에서 문제가 있었음 알 수 있다. 또한 성인의 몸무게를 조사한다고 하였을 때, 조사되는 값은 적게는 30kg에서 많게는 150kg까지의 값을 예상할 수 있다. 그러나 어떤 응답자가 10kg 또는 300kg이라고 응답을 하였다면 이는 응답자가 잘못 응답한 것임을 알 수 있다. 이러한 이상치는 자료를 요약하고 분석하는 데 커다란 영향을 미치게 된다. 따라서 수집된 자료에서 이상치의 존재를 판단하는 것은 중요한 문제가 아닐 수 없다.

이상치를 판단하는 방법은 매우 다양하게 존재한다. 우선 잘 알려진 방법들로는 다음과 같은 것들이 있다, 기본적으로 상자그림으로 이상치를 판단할 수 있다. 상자그림에서 윗 울타리나 아랫 울타리를 벗어나는 값을 이상치로 판단한다. 또한 정규분포 하에서 평균 μ에서 ± 3배의 표준편차보다 큰 값을 이상치로 판단한다. 여기서 σ는 표준편차이다. Gentleman과 Wilk (1975)는 해당 자료를 제외하고 분석한 잔차와 전체 자료에서 분석한 잔차를 이용하여 이상치를 판단하는 방법을 연구하였다. 또한

[†] 본 연구는 보건복지부 보건의료연구개발사업 (과제고유번호: A084177)의 지원과 보건복지부 권역별전문질환설치 지원–류마티스 퇴행성관절염센터 (과제고유번호: 090-091-2700-2744-300)의 지원에 의하여 이루어진 것임.

[1] (705-718) 대구광역시 남구 대명4동 3056-6, 대구가톨릭대학교 의과대학, 전임강사.
[2] 교신저자: (705-718) 대구광역시 남구 대명4동 3056-6, 대구가톨릭대학교 의과대학 의학통계학 교실, 교수. E-mail: ihshin@cu.ac.kr

Barret과 Lewis (1994), Walfish (2006)는 이상치를 판단하는 여러 방법들을 정리하여 놓았다. 또한 Seo와 Yoon (2011)은 서포트 백터 회귀를 이용한 이상치 진단의 실질적인 방법을 제안하였으며, Ahn 과 Seo (2011)는 동적 잔차도를 활용하여 가면화 효과 등으로 이상치로부터 영향을 받아 정확하게 이상치를 발견하지 못하는 경우를 개선하였다. Song 등 (2011)은 수질 자료를 이용하여 투기시점으로 여겨지는 이상점을 탐지하는 알고리즘을 R언어로 구현하였다.

만약 $K(\geq 2)$개의 그룹에서 동일한 자료를 수집하거나 또는 한 대상에 대해서 자료를 반복하여 수집하는 경우 어느 그룹 또는 어느 대상의 응답이 이상치인지 판단할 필요가 있다. 예를 들어, 10개의 학교에서 각 50명씩의 몸무게를 조사하였다면 각 학교 몸무게의 평균과 분산은 크게 다르지 않아야 할 것이다. 그러나 특정 학교의 평균 또는 분산이 특별히 작거나 크다면 그 학교의 응답값이 이상치를 의심하여야 한다. 이에 따라 본 논문에서는 다양한 이상치 판단 방법들 중 그룹별 평균과 분산에 기초하여 이상치를 판단하는 방법인 Grubb 방법과 Cochran 방법 (Burke, 2001)을 소개하고 또한 임상자료에 적용하여 보았다. 따라서 본 논문은 다음과 같이 구성하였다. 2절에서 평균을 기초로 한 이상치 검정법인 Grubb 방법과 분산을 기초로 한 이상치 검정법인 Cochran 방법을 소개한다. 그리고 3절에서는 임상자료를 2절에서 소개한 방법을 적용하여 두 가지 방법에서 판단된 이상치에 대해 비교하고 4절에서는 토의사항을 기술한다.

2. 이상치 판단 방법

Table 2.1과 같이 자료가 수집되었다고 가정하자. 그룹은 K개 이고 각 그룹에서 조사된 응답자수는 n이다. 그리고 x_{ki}는 k번째 그룹에서 i번째에 대응하는 자료이며 $k = 1, \cdots, K$, $i = 1, \cdots, n$이다. 또한 $\bar{x}_k$와 s_k^2는 각 k번째 그룹에 해당하는 평균과 표본분산이다.

Table 2.1 Data structure

Group	Number of response	Response				Mean	Sample variance
1	n	x_{11}	x_{12}	$\cdots$	x_{1n}	$\bar{x}_1$	s_1^2
2	n	x_{21}	x_{22}	$\cdots$	x_{2n}	$\bar{x}_2$	s_2^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
K	n	x_{K1}	x_{K2}	$\cdots$	x_{Kn}	$\bar{x}_K$	s_K^2

다음은 K개의 그룹 중 어느 그룹의 자료가 이상치 인지를 판단하기 위해 사용되는 Grubb 방법과 Cochran 방법을 살펴보도록 한다.

2.1. Grubb 방법

평균을 기준으로 이상치를 판단하기 위한 Grubb 방법은 다음과 같이 3가지의 식이 존재한다.

$$G_1 = \frac{|\bar{x} - \bar{x_k}|}{s}, G_2 = \frac{\bar{x_{(n)}} - \bar{x_{(1)}}}{s}, G_3 = 1 - \left(\frac{(n-3) \times s_{n-2}^2}{(n-1) \times s^2} \right). \tag{2.1}$$

여기서, n의 각 그룹의 자료의 수이고, $\bar{x}$와 s는 각 전체 자료에 대한 평균과 표준편차이다. 또한 $\bar{x}_k$는 k번째 그룹의 평균값에 해당하고, $\bar{x}_{(n)}$과 $\bar{x}_{(1)}$은 각 K개의 평균 중 가장 큰 값과 가장 작은 값에 해당한다. 그리고 s_{n-2}^2은 이상치로 여겨지는 두 개의 그룹에 대한 값을 제외한 자료에 대한 표본분산이다. Grubb 방법의 귀무가설은 '모든 그룹의 평균이 동일하다'이고, 이에 대한 검정통계량으로 G_1, G_2, G_3을 사용한다. 만약 각 신뢰수준 (95% 또는 99%)에서 모든 G_1, G_2, G_3값 중 하나의 값이라도 각각의

기각역보다 크게 되면 귀무가설을 기각하고 K개의 그룹 중 k번째 그룹을 이상치의 후보로 판단할 수 있다. Grubb 검정에 대한 기각역 (G_1^*, G_2^*, G_3^*)은 Table 2.2에 제시되어 있다. 예를 들어, 각 그룹의 자료의 수 (n)가 10이고 95%신뢰수준에서 Grubb 검정을 한다면 G_1, G_2, G_3에 대한 각각의 기각역은 2.18, 3.68, 0.77이다. 그러나 원하는 n과 신뢰수준에 해당하는 값이 없을 경우 연구자가 직접 Little 등 (1987) 이 제시한 식으로 기각역을 계산 하여야 한다.

Table 2.2 Critical region of Grubb test

n	95% Confidence level			99% Confidence level			n	95% Confidence level			99% Confidence level		
	G_1^*	G_2^*	G_3^*	G_1^*	G_2^*	G_3^*		G_1^*	G_2^*	G_3^*	G_1^*	G_2^*	G_3^*
3	1.15	2.00	-	1.16	2.00	-	30	2.75	4.89	0.4	3.1	5.19	0.47
4	1.46	2.43	1	1.49	2.44	1.00	35	2.81	5.03	0.36	3.18	5.33	0.43
5	1.67	2.75	0.98	1.75	2.80	1.00	40	2.87	5.15	0.33	3.24	5.45	0.39
6	1.82	3.01	0.94	1.94	3.10	0.98	50	2.96	5.35	0.28	3.34	5.65	0.33
7	1.94	3.22	0.90	2.10	3.34	0.96	60	3.03	5.50	0.25	3.41	5.80	0.29
8	2.03	3.40	0.85	2.22	3.54	0.93	70	3.08	5.64	0.22	3.47	5.94	0.26
9	2.11	3.55	0.81	2.32	3.72	0.89	80	3.13	5.73	0.20	3.52	6.03	0.24
10	2.18	3.68	0.77	2.41	3.88	0.86	90	3.17	5.82	0.18	3.56	6.12	0.21
12	2.29	3.91	0.7	2.55	4.13	0.80	100	3.21	5.90	0.17	3.6	6.20	0.20
13	2.33	4.00	0.67	2.61	4.24	0.77	110	3.24	5.97	0.16	3.63	6.27	0.18
15	2.41	4.17	0.62	2.71	4.43	0.71	120	3.27	6.03	0.15	3.66	6.33	0.17
20	2.56	4.49	0.52	2.88	4.79	0.61	130	3.29	6.09	0.14	3.69	6.39	0.16
25	2.66	4.73	0.45	3.01	5.03	0.53	140	3.32	6.14	0.13	3.71	6.44	0.15
30	2.75	4.89	0.40	3.10	5.19	0.47							

2.2. Cochran 방법

분산을 기준으로 이상치를 판단하기 위한 Cochran 검정의 통계량은 다음과 같다.

$$C_k = \frac{s_k^2}{\sum_{k=1}^{K} s_k^2}, k = 1, \cdots, K. \tag{2.2}$$

여기서, C_k는 k번 째 그룹에 해당하는 Cochran의 C통계량이고 s_k^2는 k번째 그룹의 표본분산이다. Cochran 검정의 귀무가설은 '모든 그룹의 분산이 동일하다' 이고, 이를 검정하기 위한 기각역은 다음과 같이 계산된다.

$$C^*(\alpha, n, K) = \left[1 + \frac{K-1}{F(\alpha/K, (n-1), (K-1)(n-1))}\right]^{-1}. \tag{2.3}$$

여기서, α는 유의수준, n은 각 그룹에서 자료의 수, K는 그룹의 수이고 F는 유의수준 α와 자유도 $(n-1)$, $(K-1)(n-1)$을 가지는 F분포로부터 계산된 값이다. 만약 C_k값이 기각역 C^*보다 크게 되면 귀무가설을 기각하고 적어도 하나의 그룹에 해당하는 분산이 다른 분산보다 크다는 대립가설을 채택하게 된다. 즉 k번째 그룹의 자료가 이상치의 후보로 판단된다.

대개 Cochran 검정에 대한 기각역은 Table 2.3과 같이 테이블로 주어지나 자유도 $n-1$이나 그룹의 수 K에 해당하는 값이 없을 경우 연구자가 직접 기각역을 계산하여야 한다. 무료 프로그램인 R 프로그램에서 "outliers" 패키지를 설치하여 qcochran$(1-\alpha, n, K)$ 함수를 이용하여 Cochran 검정의 기각역을 계산할 수 있다. 예를 들어, 유의수준 (α)이 0.05이고 자료의 수 (n)가 10이고 그룹의 수 (K)가 5라면 R 프로그램 창에서 qcochran(0.99,10,5)라고 입력하면 0.4854로 계산되어 진다.

Table 2.3 Critical region of Cochran test with significance level 0.01

K	$n-1$													
	1	2	3	4	5	6	7	8	9	10	16	36	144	∞
2	0.9999	0.995	0.9794	0.9586	0.9373	0.9172	0.8988	0.8823	0.8674	0.8539	0.7949	0.7067	0.6062	0.5000
3	0.9933	0.9423	0.8831	0.8335	0.7933	0.7606	0.7335	0.7107	0.6912	0.6743	0.6059	0.5153	0.4230	0.3333
4	0.9676	0.8643	0.7814	0.7212	0.6761	0.6410	0.6129	0.5897	0.5702	0.5536	0.4884	0.4057	0.3251	0.2500
5	0.9279	0.7885	0.6957	0.6329	0.5875	0.5531	0.5259	0.5037	0.4854	0.4697	0.4094	0.3351	0.2644	0.2000
6	0.8828	0.7218	0.6258	0.5635	0.5195	0.4866	0.4608	0.4401	0.4229	0.4084	0.3529	0.2858	0.2229	0.1667
7	0.8376	0.6644	0.5685	0.5080	0.4659	0.4347	0.4105	0.3911	0.3751	0.3616	0.3105	0.2494	0.1929	0.1429
8	0.7945	0.6152	0.5209	0.4627	0.4226	0.3932	0.3704	0.3522	0.3373	0.3248	0.2779	0.2214	0.1700	0.1250
9	0.7544	0.5727	0.4810	0.4251	0.3870	0.3592	0.3378	0.3207	0.3067	0.2950	0.2514	0.1992	0.1521	0.1111
10	0.7175	0.5358	0.4469	0.3934	0.3572	0.3308	0.3106	0.2945	0.2813	0.2704	0.2297	0.1811	0.1376	0.1000
12	0.6528	0.4751	0.3919	0.3428	0.3099	0.2861	0.2680	0.2535	0.2419	0.2320	0.1961	0.1535	0.1157	0.0833
15	0.5747	0.4069	0.3317	0.2882	0.2593	0.2386	0.2228	0.2104	0.2002	0.1918	0.1612	0.1251	0.0934	0.0667
20	0.4799	0.3297	0.2654	0.2288	0.2048	0.1877	0.1748	0.1646	0.1567	0.1501	0.1248	0.0960	0.0709	0.0500
24	0.4247	0.2871	0.2295	0.1970	0.1759	0.1608	0.1495	0.1406	0.1338	0.1283	0.1060	0.0810	0.0595	0.0417
30	0.3632	0.2412	0.1913	0.1635	0.1454	0.1327	0.1232	0.1157	0.1100	0.1054	0.0867	0.0658	0.0480	0.0333
40	0.2940	0.1915	0.1508	0.1281	0.1135	0.1033	0.0957	0.0898	0.0853	0.0816	0.0668	0.0503	0.0363	0.0250
60	0.2151	0.1371	0.1069	0.0902	0.0796	0.0722	0.0668	0.0625	0.0594	0.0567	0.0461	0.0344	0.0245	0.0167
120	0.1225	0.0759	0.0585	0.0489	0.0429	0.0387	0.0357	0.0334	0.0316	0.0302	0.0242	0.0178	0.0125	0.0083
∞	0	0	0	0	0	0	0	0	0	0	0	0	0	0

3. 예제

사용한 자료는 노르웨이의 오슬로와 네덜란드의 위트르헤트에서 30가구를 대상으로 개의 혈장을 10개 실험실에서 수집한 동물임상자료 (Ulleberg 등, 2011)이다. 또한 각 연구소별로 세 번씩 반복 측정을 하였고 자료의 일부는 Table 3.1에 제시하였다. 개의 혈장 자료에서 각 가구에서 10개의 실험실의 자료가 동일한지 아니면 어느 실험실의 자료가 이상치의 후보인지를 확인하기 위해 2절에서 제시한 Grubb 검정과 Cochran 검정을 적용하여 보았다.

Table 3.1 Data for plasma of dogs

Sample number	Laboratory												
	1			2			3			···	10		
	Replication			Replication			Replication				Replication		
	1	2	3	1	2	3	1	2	3		1	2	3
1	151	147	151	153	144	152	149	174	215		151	146	146
2	95	93	96	95	98	91	84	98	124		100	89	91
3	96	93	99	107	95	100	85	103	150	···	100	90	96
4	111	107	110	121	113	109	103	119	163		125	112	114
5	95	93	98	101	89	92	79	98	139		88	83	84
						⋮							
26	314	309	314	334	334	313	355	421	757		328	317	331
27	619	616	619	664	664	620	650	763	1420		680	636	657
28	418	415	426	448	448	449	457	531	1010	···	434	416	425
29	542	539	544	567	567	551	550	703	289		558	540	565
30	715	709	715	749	749	718	709	915	1774		754	727	750

Table 3.2에서 각 가구별 실험실에 대한 평균과 분산을 일부 제시하였다. 30가구 전체에서 세 번째 실험실의 평균 및 분산은 각 다른 실험실의 값들과 크게 차이가 있음을 보이고 있다. 예를 들어, 가구1에 대한 값을 보면, 세 번째 실험실의 평균과 분산은 각 179.33과 1110.00이다. 그러나 첫 번째 실험실의 평균과 분산은 각 149.67과 5.33이고 두 번째 실험실의 평균과 분산은 각 149.67과 24.33이다. 따라서 세 번째 실험실의 자료가 이상치인지를 Grubb 검정과 Cochran 검정을 사용하여 판단하여 보았다.

Table 3.2 Mean and variance of laboratory by sample

Sample	Laboratory1		Laboratory2		Laboratory3		...	Laboratory10		Total	
	Mean	Variance	Mean	Variance	Mean	Variance		Mean	Variance	Mean	Variance
1	149.67	5.33	149.67	24.33	179.33	1110.33		147.67	8.33	147.73	292.00
2	94.67	2.33	94.67	12.33	102.00	412.00		93.33	34.33	91.97	146.10
3	96.00	9.00	100.67	36.33	112.67	1126.33		95.33	25.33	96.90	240.51
4	109.33	4.33	114.33	37.33	128.33	965.33		117.00	49.00	110.93	222.75
5	95.33	6.33	94.00	39.00	105.33	940.33		85.00	7.00	93.60	153.70
6	102.00	7.00	105.33	2.33	124.67	1172.33		101.33	12.33	105.07	227.03
7	130.00	9.00	129.00	48.00	154.00	1164.00		121.33	16.33	126.97	258.03
8	109.00	9.00	108.00	9.00	129.33	1136.33		104.67	10.33	108.23	210.87
9	104.00	4.00	110.67	105.33	127.67	1605.33		102.67	80.33	104.23	260.05
10	87.67	6.33	89.33	4.33	97.67	710.33		82.67	16.33	83.60	168.04
11	158.67	6.33	164.33	42.33	209.33	4384.33		163.67	37.33	162.17	660.07
12	170.33	6.33	171.00	16.00	216.67	3892.33		167.00	49.00	170.37	645.07
13	142.00	1.00	157.00	7.00	188.33	2904.33		165.67	14.33	147.47	644.74
14	142.67	10.33	155.67	22.33	188.67	2454.33		146.33	25.33	150.93	428.48
15	206.67	20.33	220.67	2.33	296.33	12530.33		224.67	6.33	219.67	1800.30
16	191.33	1541.33	227.67	81.33	289.33	9996.33	...	237.33	209.33	226.93	1737.58
17	90.33	142.33	96.67	345.33	126.67	3237.33		90.67	520.33	96.20	507.89
18	85.67	2.33	86.00	39.00	112.00	1561.00		83.67	12.33	84.63	269.14
19	246.00	7.00	258.00	93.00	351.00	16393.00		266.33	44.33	259.63	2296.03
20	317.00	49.00	324.33	281.33	464.33	32817.33		321.00	37.00	330.13	4721.64
21	760.67	37.33	796.33	302.33	1269.00	311653.00		794.33	184.33	801.23	65846.46
22	245.67	5.33	250.67	170.33	373.33	24862.33		255.33	46.33	253.20	4213.34
23	689.67	4.33	712.33	277.33	1040.00	196123.00		722.00	93.00	726.90	26977.68
24	452.00	4.00	478.00	100.00	740.33	101349.33		469.33	74.33	472.67	20616.02
25	398.67	5.33	403.33	22.33	601.00	68383.00		416.67	20.33	422.37	9857.96
26	312.33	8.33	322.00	117.00	511.00	46476.00		325.33	54.33	355.43	7037.43
27	618.00	3.00	634.00	676.00	944.33	172886.33		657.67	484.33	659.33	22927.54
28	419.67	32.33	442.33	114.33	666.00	90121.00		425.00	81.00	450.27	12572.82
29	541.67	6.33	565.00	172.00	847.33	152154.33		554.33	166.33	556.80	26650.37
30	713.00	12.00	734.33	162.33	1132.67	319090.33		743.67	212.33	738.07	53847.37

우선 Grubb 검정을 위하여 검정통계량인 G_1과 G_2를 계산하여 Table 3.3에 제시하였다. 사용한 예제에서는 각 실험실별 자료의 수가 3이므로 G_3는 계산하지 않았다. 그리고 자료의 수가 3이므로 Grubb 검정의 기각역은 95% 신뢰수준에서 G_1^*와 G_2^*는 각 1.15, 2.00이고 99% 신뢰수준에서는 각 1.16이고 2.00이다. 모든 가구에서 검정통계량 (G_1, G_2)이 기각역 (G_1^*, G_2^*)보다 크다. 따라서 평균을 근거로 한 이상치 검정인 Grubb 검정으로는 세 번째 실험실 자료는 이상치라고 할 수 있다.

Table 3.3 Test statistic of Grubb for third laboratory data

Sample	G_1	G_2	Sample	G_1	G_2	Sample	G_1	G_2
1	1.849	2.965	11	1.836	2.66	21	1.823	2.611
2	0.83	2.647	12	1.823	2.822	22	1.851	2.496
3	1.017	2.257	13	1.609	2.573	23	1.906	2.835
4	1.166	2.389	14	1.823	2.834	24	1.864	2.695
5	0.946	2.151	15	1.807	2.773	25	1.799	2.649
6	1.301	2.168	16	1.497	2.407	26	1.854	2.368
7	1.683	2.47	17	1.352	2.248	27	1.882	2.71
8	1.453	2.364	18	1.668	2.621	28	1.924	2.774
9	1.454	2.378	19	1.907	2.671	29	1.78	2.593
10	1.085	2.314	20	1.953	2.877	30	1.7	2.558

다음으로 분산을 근거로 세 번째 실험실 한 이상치 검정을 위해 Cochran 검정을 하였다. 검정통계량인 C_k값은 Table 3.4에 제시하였다. 사용한 예제에서 자료의 수와 그룹의 수에 해당하는 Cochran 검정의 기각역 C^*은 0.5358이다. 2번째 가구를 제외하고 모든 가구에서 검정통계량이 기각역 보다 크다. 따라서 분산을 근거로 한 이상치 검정인 Cochran 검정으로 세 번째 실험실 자료에서 가구2를 제외한 모든 자료가 이상치라고 할 수 있다.

Table 3.4 Test statistic of Cochran for third laboratory data

Sample	C_k	Sample	C_k	Sample	C_k	Sample	C_k	Sample	C_k	Sample	C_k
1	0.848	6	0.615	11	0.939	16	0.760	21	0.604	26	0.799
2	0.480	7	0.760	12	0.881	17	0.651	22	0.750	27	0.981
3	0.571	8	0.856	13	0.709	18	0.872	23	0.984	28	0.986
4	0.580	9	0.813	14	0.872	19	0.942	24	0.661	29	0.735
5	0.729	10	0.819	15	0.913	20	0.972	25	0.888	30	0.706

앞에서 살펴본 바와 같이 세 번째 실험실자료는 Grubb 검정과 Cochran 검정으로 이상치로 의심이 가는 것으로 판단되었다. 따라서 세 번째 실험실에서 반복 측정한 자료의 값 중 어느 반복에서 잘못 측정되었는지 또는 세 번째 실험실의 측정도구의 오류가 없는지를 살펴보아야 할 것이다.

4. 결론

수집된 자료에서 이상치를 판단하는 것은 여러 분야에서 중요한 문제이다. 특히 인간을 대상으로 하는 임상시험에서는 더욱 그러하다. 본 논문에서는 평균을 기초로 이상치를 판단하는 Grubb 방법과 분산을 기초로 이상치를 판단하는 Cochran 방법을 소개하고 동물 임상자료인 개의 혈장자료에 적용시켜 보았다. 두 방법을 적용시켜 본 결과 서로 이상치의 후보로 판단되는 자료가 동일하게 이상치로 판정되었다. 즉 자료를 수집하는 반복과정에서 오류가 발생하였는지 또는 자료를 기록하는 과정에서 오류가 발생하였는지에 대한 조사가 필요할 것이다. 더 나아가서 이상치를 줄이기 위한 노력으로 자료의 입력시 잘못 입력되는 오류를 최소화하여야 한다. 이는 어느 방법보다 입력자의 주의가 가장 필요한 부분이다. 또한 응답자가 잘못 응답하는 경우를 해결하기 위한 교육과 환경의 개선이 동시에 진행되어야 할 것이다.

참고문헌

Ahn, B. J. and Seo, H. S. (2011). Outlier detection using dynamic plots. *The Korean Journal of Applied Statistics*, **24**, 979-986.

Barret, V. and Lewis, T. (1994). *Outliers in statistical data*, 3rd Edition, John Wiley, England.

Burke, S. (2001). Missing values, outliers, robust statistics and non-parametric methods. *Statistics and Data Analysis*, LC·GC Europe online Supplement, 19-24.

Gentleman, J. F. and Wilk, M. B. (1975). Detecting outliers II: Supplementing the direct analysis of residuals. *Biometrics*, **31**, 387-410.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical analysis with missing data*, John Wiley & Sons, United States of America.

Seo, H. S. and Yoon, M. (2011). Outlier detection using support vertor machines. *Communications of the Korean Statistical Society*, **18**, 171-177.

Song, G. M., Moon, J. E. and Park, C. (2011). Realization of an outlier detection algorithm using R. *Journal of the Korean Data & Information Science Society*, **22**, 449-458.

Ulleberg, T., Robben, J., Nordahl, K. M., Ulleberg, T. and Heiene, R. (2011). Plasma creatinine in dogs: Intra and inter laboratory variation in 10 European veterinary laboratories. *Acta Velerinaria Scandinavica*, **53**, 1-13.

Walfish, S. (2006). A review of statistical outlier methods. *Pharmaceutical Technology*, **2**, 1-5.

Journal of the Korean Data & Information Science Society
2012, **23**(4), 657–663

http://dx.doi.org/10.7465/jkdi.2012.23.4.657
한국데이터정보과학회지

Outlier detection using Grubb test and Cochran test in clinical data[†]

Ki Cheul Sohn[1] · Im Hee Shin[2]

[12]School of Medicine, Catholic University of Daegu

Received 10 May 2012, revised 29 May 2012, accepted 7 June 2012

Abstract

There are very small values and/or very big values which get out of the normal range for survey data in various fields. The reasons of occurrence for outlier are two. One of them is the error in process of data input and the other is the strange response of the respondent. If the data has outliers, then the summary statistics such as the mean and the variance produce misleading information. Therefore, researcher should be careful in detecting the outlier in data. In particular, it is very important problem for clinical fields because the cost of experiment is very high. This article introduce the Grubb test and Cochran test to detect outliers in the data and we apply this method for clinical data.

Keywords: Clinical data, Cochran-test, Grubb-test, outlier.

[†] This study was supported by the grant of Korea Ministry of Health & Welfare, Republic of Korea (Project No: A084177 and Project No: 090-091-2700-2744-300).

[1] Full time instructor, School of Medicine, Catholic University of Daegu 705-718, Republic of Korea.
[2] Corresponding author: Professor, Department of Medical Statistics, School of Medicine, Catholic University of Daegu 705-718, Republic of Korea. E-mail: ihshin@cu.ac.kr.