응용통계연구
제 5 권 제 1 호, 1992년

통계 Package에서의 Box Plot Algorithm†

김승환*, 전홍석**

요 약

여러 통계 Package에서 사용된 Box Plot Algorithm을 통해 그들을 분석하고 비교하여 최적의 Box Plot Algorithm을 산출하고자 한다.

1. 서 론

John W. Tukey[12]에 의하여 제안된 Box-Plot는 자료의 기초 통계량인, 최대값, 최소값, 중앙값 및 사분위수 등을 이용하여 분포의 모형을 쉽게 보여주는 도구로서 EDA의 과정에서 가장 많이 쓰이는 도구중의 하나이다. 또한 다중비교에 있어서도 복잡한 계산 과정을 거치지 않고 시각적으로 처리 효과 차이를 비교할 수 있다. 그러나 몇가지의 상업적인 컴퓨터 패케지를 이용하여보면, 서로 다른 결과를 주기 때문에 통계 비전문인에게는 약간의 혼란을 줄 우려가 있다.

이러한 혼란의 원인은 사분위수를 계산하는 알고리즘이 여러가지에 이르기 때문이다. 대략 8가지 정도의 알고리즘이 제안되고 있는데, 이 논문에서는 그들간의 비교검토에 의하여 좀더 적합한 알고리즘을 알아보고자한다.

2. Box-Plot의 산법

2.1 각 Package의 Box-Plot의 결과 소개 및 문제의 제기

Frigge, M., Hoaglin, D.C. and Iglewicz, B.[2]는 53, 56, 75, 81, 82, 85, 87, 89, 95, 99, 100의 동일한 자료로 여러 통계 Package에서 Box-Plot결과를 산출하여 서로 다르게 나오는 문제를 분석하였다. 본 논문은 이 문제를 국내에서 널리 사용되고 있는 SPSS PC$^+$ V2.0, STATGRAPHICS

† 이 연구는 1991년도 인하대학교 연구비 지원에 의하여 수행되었음.
* (402-751) 인천시 남구 용현동 253 인하대학교 수학과(수리통계학 전공) 박사과정
** (402-751) 인천시 남구 용현동 253 인하대학교 통계학과 부교수

V3.0, MINITAB Relase 5.1.1, MINITAB Relase 7, SAS V6.04의 PC Version에 초점을 맞추어 산법과 출력결과를 비교하고, 중위수의 신뢰구간에 대해 구체적으로 연구하고자 한다.

표 1은 동일한 자료로부터 여러 가지의 통계 Package를 통해 얻은 결과이다. 각 Package의 결과가 중위수의 경우는 85.0으로 모두 같게 나왔지만 중위수의 신뢰구간과 Q_1, Q_3, 인접치, Fence의 값이 약간의 차이를 보임을 알 수 있다. 이결과로 중위수를 계산하는 산법은 모두 같지만 사분위수를 구하는 산법이 서로 같지 않음을 알 수 있다. MINITAB의 경우, DESCRIPTIVE STATISTIC Procedure에서 계산된 사분위수와 Box-Plot Procedure에서 계산된 사분위수가 서로 틀려 혼란을 가중시킬 수 있다.

또한, 중위수의 신뢰구간과 인접치, Fence의 계산은 모두 사분위수의 산법에 영향을 받으므로, 개별적인 산법의 분석이 필요함을 말해주고 있다.

표 1. 각 Package에서의 Box-Plot 결과 비교

구 분	SPSS	STATGRAPHICS	MINITAB 5.1.1	MINITAB 7	SAS
UPPER FENCE	125.0	125.0	113.0	113.0	125.0
위 인접치	100.0	100.0	100.0	100.0	100.0
Q_3	95.0	95.0	92.0	92.0	95.0
중위수의 UL	*	94.5	91.6	95.33	*
중위수	85.0	85.0	85.0	85.0	85.0
중위수의 LL (약 95%)	*	75.5	78.3	73.44	*
Q_1	75.0	75.0	78.0	78.0	75.0
아래 인접치	53.0	53.0	75.0	75.0	53.0
LOWER FENCE	45.0	45.0	57.0	57.0	45.0

**표 2.1에서 "*"는 해당 Package에서 계산하지 못함을 나타낸다.

2.2 일반적인 Box-Plot의 산법 소개

2.2.1 중위수(Median)

자료의 수가 홀수일 때 : (n+1)/2번째 순서치.

자료의 수가 짝수일 때 : n/2번째와 (n+1)/2번째 순서치의 산술평균.

2.2.2 중위수(Median)의 신뢰구간

MINITAB V.5.1.1과 STATGRAPHICS에서는 아래의 식으로 모집단 중위수 (M)의 신뢰구간을 $\hat{M} \pm C \times S$로 계산하고 있다.

여기서, $S = \frac{1.25 \times IQR}{1.35\sqrt{n}}$, $IQR = Q_3 - Q_1$이다.

C는 상수로 약 1.3~2.0 사이의 값이며 모집단 중위수에 대한 95% 신뢰구간일 경우 C=1.96이지만, 경험적으로 C=1.7이 많이 사용된다.

또한, MINITAB Release 7.에서는 Sign Confidence interval을 사용하여 계산한다. η를 미지의 모집단 중위수라고 하자. n개의 표본을 얻을 때 X를 η보다 작은 표본의 수라 하면 X는 시행횟수 n, 성공확률 1/2을 갖는 이항분포를 따르게 된다. 그러므로 구간(d번째 작은 표본, d번째로 큰 표본)을 포함하는 확률은 $1 - 2Pr(X<d)$로 계산할 수 있다. 이 확률이 0.95인 구간이 η에 대한 95% 신뢰구간이 된다. 하지만, 실제 표본의 대부분은 사용자가 원하는 확률을 정확하게 만족시키지 못하기 때문에 비선형 보간법(Nonlinear Interpolation)을 이용하여 원하는 확률의 신뢰구간을 추정하고 있다[8].

2.2.3 사분위수(Q_1, Q_3)

Frigge, M., Hoaglin, D.C. and Iglewicz, B.[2]는 일반적으로 사용되고 있는 사분위수의 산법을 아래와 같이 소개 하였다.

1) Weighted Average at $X_{(n/4)}$([10])

$Q_1 = (1-g)\,X_{(j)} + gX_{(j+1)}$, $n/4 = j + g$; j:정수, $g:0 \le g < 1$

Q_3는 Q_1의 식에 n/4 대신 3n/4를 대입하여 계산한다.

2) Observation Numbered Closest to n/4.([10])

$Q_1 = X_{(j)}$, where j는 $n/4 + 0.5$의 정수 부분

$Q_3 = X_{(j)}$, where j는 $3n/4 + 0.5$의 정수 부분

3) Empirical Distribution Function([10])

$Q_1 = X_{(j)}$, if $g = 0$
$Q_1 = X_{(j+1)}$, if $g > 0$] $n/4 = j + g$; j:정수, $g:0 \le g < 1$

Q_3는 n/4 대신 3n/4로 계산한다.

4) Weighted Average Aimed at $X_{((n+1)/4)}$([10])

$Q_1 = (1-g)\,X_{(j)} + gX_{(j+1)}$, $(n+1)/4 = j + g$

$Q_3 = (1-g)\,X_{(j)} + gX_{(j+1)}$, $3(n+1)/4 = j + g$

5) Empirical Distribution Function with Averaging([10])

$Q_1 = (X_{(j)} + X_{(j+1)})/2$, if $g = 0$
$Q_1 = X_{(j+1)}$, if $g > 0$] $n/4 = j + g$; j:정수, $g:0 \le g < 1$

Q_3는 n/4 대신 3n/4를 대입한다.

6) Standard Fourths of Hinges([12])

$Q_1 = (1-g)X_{(j)} + gX_{(j+1)}$

$Q_3 = (1-g)X_{(n+1-j)} + gX_{(n+1-j-1)}$

$[(n+3)/2]/2 = j+g$ ($g=0$ or $g=0.5$)

단, [X]는 X를 초과하지 않는 X의 최대값이다.

7) Ideal or Machine Fourths.([3])

$Q_1 = (1-g)X_{(j)} + gX_{(j+1)}$, $n/4+(5/12) = j+g$

$Q_3 = (1-g)X_{(j)} + gX_{(j+1)}$, $3n/4+(7/12) = j+g$

8) Weighted Average Aimed at $X_{(n/4+0.5)}$.([1])

$Q_1 = (1-g)X_{(j)} + gX_{(j+1)}$, $n/4+0.5 = j+g$

$Q_3 = (1-g)X_{(j)} + gX_{(j+1)}$, $3n/4+0.5 = j+g$

이상 8가지의 산법은 각 Package에서 표 2와 같이 사용되고 있다.

표 2. 통계 Package에서 사용하고 있는 산법

산법	사용되고 있는 Package
1)	SAS(option)
2)	SAS(option)
3)	SAS(option)
4)	SAS Univariate procedure, MINITAB DESCRIPTIVE procedure
5)	SPSS MANOVA procedure, STATGRAPHICS BOX PLOT procedure
6)	MINITAB BOX PLOT procedure
7)	·
8)	·

4), 6)을 보면 같은 MINITAB에서도 서로 다른 산법을 사용하고 있음을 알 수 있다.

2.2.4 상,하 인접치(Lower & Upper Adjacent Value)

인접치(Adjacent Value)를 계산하는 방법은 아래와 같다.

1) 하인접치(Lower Adjacent Value) : $MIN\{X;X \geq Q_1 - k \times IQR\}$

2) 상인접치(Upper Adjacent Value) : $MAX\{X;X \leq Q_3 + k \times IQR\}$

단, IQR은 Interquartile Range의 약자로 $Q_1 - Q_3$ 이다.

2.2.5 상, 하 Fence(Lower & Upper Fence)

Fence를 계산하는 방법은 아래와 같다.

1) Lower Fence : $Q_1 - k \times IQR$

2) Upper Fence : $Q_3 + k \times IQR$

Fence를 벗어나는 자료는 이상치(Outlier)로 취급한다.

2.3 Box-Plot 산법에 대한 비교 및 분석

2.3.1 사분위수 산법

Frigge, et. al.[2]는 사분위수의 산법을 아래와 같은 방법으로 비교하였다.

Q_1은 자료의 수가 증가함에 따라 깊이(Depth)가 선형으로 증가하는 특성을 가지고 있어야 하고 정의대로 근사적, 혹은 정확하게 25% 가 되는 값, 혹은 그 값에 가까운 값이 되어야 한다. 이를 알아보기 위해서는 자료의 수에 따른 깊이(Depth)의 변화를 알아보는 방법이 효과적일 것이다. 그림 1으로 1), 4), 7), 8)식이 선형으로 증가함을 알 수 있다. 이중 어느 식이 실제 자료의 25% 값에 가깝게 가는 가를 알아 보기 위해 대표본의 정규모의난수를 발생시켜 자료의 25% 점을 계산하여 비교해 본 결과 4), 7), 8)식에서 계산된 Q_1은 거의 정확한 값임을 알 수 있다.

2.3.2 Fence의 계산을 위한 k 값

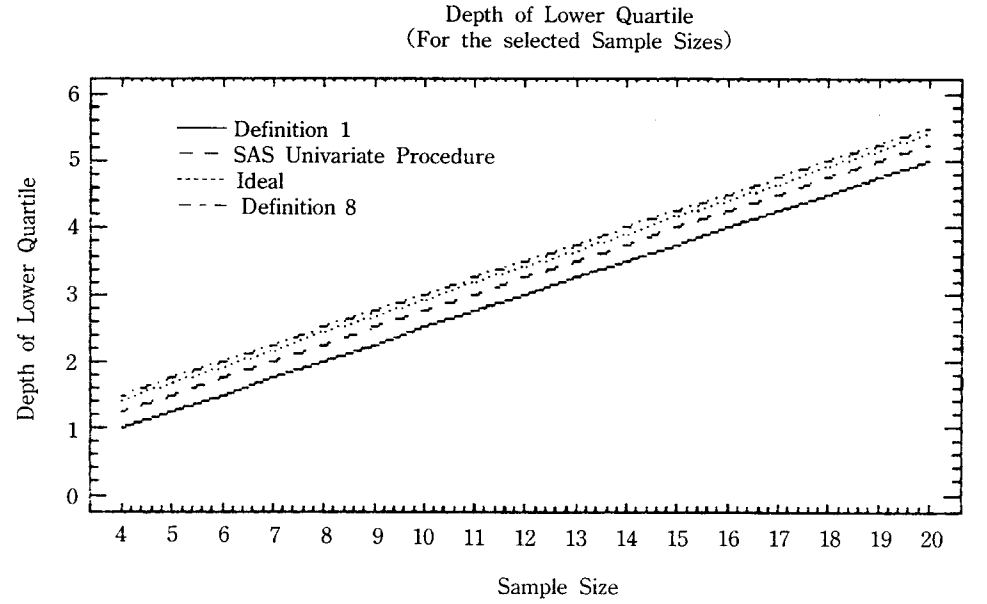
k값은 자료의 이상치를 판정하는 데에 사용되는 값으로 Hoaglin, Iglewicz, Tukey(1986)[4]는 n개의 임의추출표본 중에서 하나 또는 그 이상이 이상치로 판정될 확률에 대해 연구하여 그 확률을 1-B(k, n)으로 표시하였고 그 값은 표 3과 같다.

표 3. n과 k에 따른 이상치 판정확률(1-B(k, n))

n	k			
	1.0	1.5	2.0	3.0
10	0.424	0.198	0.094	0.026
20	0.577	0.232	0.082	0.011
30	0.705	0.284	0.094	0.008
50	0.837	0.365	0.094	0.004
100	0.967	0.523	0.115	0.003

주 : Q_1 , Q_3 를 계산하는 식은 6)식을 사용하였다.

k=1.5의 값을 사용하면 정상적인 100개의 정규임의 추출표본 중에서 1개 혹은 그 이상의 자료를 이상치로 판단할 확률이 0.523임을 표 3에서 알 수 있다. 현재는 대개 k를 1.5 정도로 가장 적당한 값으로 보는 견해가 지배적이고, 이 논문에서 취급한 모든 통계 Package에서 k=1.5의 값을 사용하고 있다.



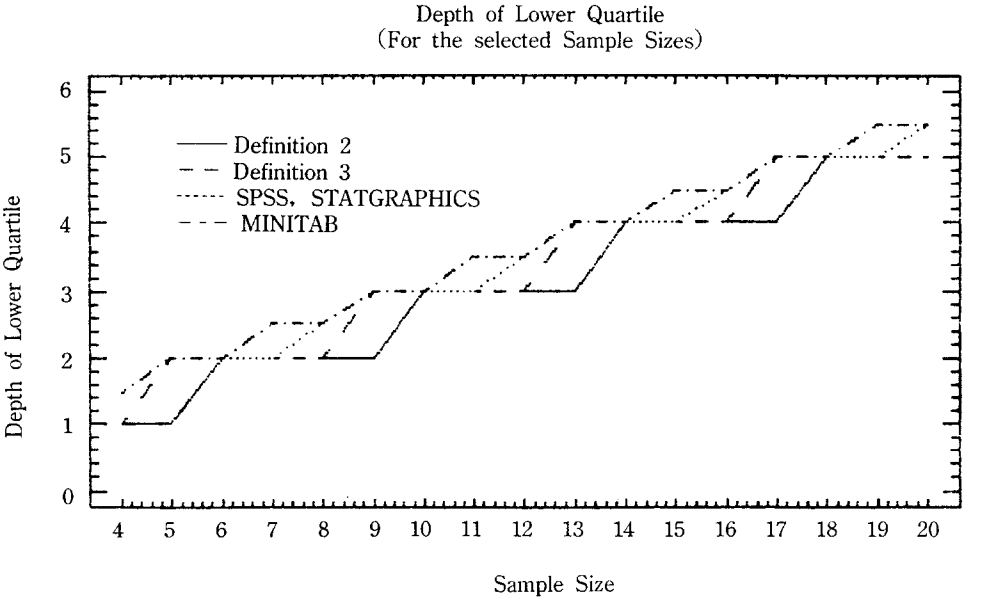


그림 1. 표본크기에 따른 Q_1의 깊이(Depth)

2.3.3 중위수에 대한 신뢰구간 산법

위에서 두가지 방법을 소개하였다.

첫번째 방법은 중위수의 근사 표본분포를 이용한 방법이다.

중위수의 근사표본분포는 아래와 같다.[13]

$$\hat{M} \sim AN(\mu, 1/4nf^2(\mu))$$

여기서, M은 모집단의 중위수이고 f는 확률밀도함수이다.

위의 식에서 f에 정규분포의 PDF를 대입하여 중위수의 표준편차를 구하면 아래와 같이 된다.

$$\sigma(\hat{M}) = \frac{1.2533\sigma}{\sqrt{n}} \tag{2.1}$$

여기서 σ에 대한 대안으로 Semi-Interquartile Range을 생각할 수 있다.

$$\hat{R} = \frac{1}{2}(\hat{Q}_3 - \hat{Q}_1), \quad R : \text{Semi-Interquartile Range}$$

$\hat{R} \sim AN(0.6745\sigma, (0.7867)^2\sigma^2/n)$[13]이 성립한다. 그러므로 표준편차와 IQR의 관계식은 아래와 같이 쓸 수 있다.

$$\sigma = IQR/1.349 \tag{2.2}$$

(2.2)식을 (2.1)에 대입하면 아래의 식을 얻을 수 있다.

$$\sigma(\hat{M}) = \frac{1.2533\ IQR}{1.349\sqrt{n}} \tag{2.3}$$

지금까지 살펴본 바와 같이 MINITAB V. 5.1.1과 STATGRAPHICS에서 사용한 방법은 모집단이 정규분포라는 가정하에 나온 식이므로 모집단의 분포가 정규분포를 따를 때에는 (2.3)식을 무리없이 사용할 수 있다. 또한 이 때의 상수 "C" 값도 표준정규분포표를 이용하여 결정할 수 있다. 그러나 분포가 미지이거나 정규분포와 많은 차를 가지는 분포라면 위의 방법을 아무 수정없이 사용할 수 없게 된다.

두번째 방법은 비모수적인 방법이므로 모집단의 분포에 무관하고 사용자가 원하는 신뢰한계에 대한 신뢰구간을 비선형 보간법을 이용하여 계산해 줄 수 있다는 점이 통계 Package에서 사용하는 데에 있어 첫번째 방법에 비해 장점으로 생각할 수 있다. 이 방법에서 사용한 비선형보간법은 Hettmansperger과 Sheather[5]에 의해 제안된 방법으로 자료의 분포가 대칭분포(Normal, Cauchy, Uniform)와 비대칭분포인 경우에도 비교적 정확한 신뢰구간을 제공하는 것으로 알려져 있다. 표 4는 비선형 보간법이 어느 정도 정확히 원하는 신뢰한계를 만족시켜줄 수 있는 가를 모의실험을 통해

알아 본 결과이다. 모의실험은 표본의 크기와 모집단분포를 변화시키면서 추출한 임의표본에서 계산된 중위수의 95% 신뢰구간에 모집단 중위수가 포함될 확률을 10,000번씩 계산하여 평균한 값을 "%"로 나타내었다.

표 4.

	n = 10	n = 30	n = 50	n = 100
Normal	94.42	94.15	94.69	94.86
Univorm	93.95	94.27	94.50	95.21
Cauchy	94.14	94.27	94.82	94.98
F	93.66	94.31	94.67	94.91

표 4에서 알 수 있듯이 두번째 방법은 모집단의 분포와는 무관하고 대체적으로 95%에 가까운 결과를 주고 있음을 알 수 있다.

3. 출력결과의 문제점

Box-Plot은 우리가 알고자 원하는 자료의 정보를 한눈에 보여 주는 방법이라는 측면으로 볼 때 무엇보다도 한눈에 자료가 가지고 있는 정보를 쉽게 파악할 수 있어야 할 것이다. 이러한 관점에서 Box-Plot는 고해상도 그래픽(High Resolution Graphic)처리를 해야 한다. 현재 국내에서 보편적으로 사용되고 있는 Package 중에 고해상도 그래픽을 제공하는 Package는 SAS for PC, SPSS PC$^+$, STATGRAPHICS, CSS 등이 있지만 SAS for PC, SPSS PC$^+$는 통계적인 문제의 그래픽 처리는 거의 못하는 실정이다. 실제로, 통계적 목적의 그래픽을 제공할 수 있는 Package는 MINITAB Release 7.과 STATGRAPHICS가 있다. MINITAB은 사용자를 만족시킬 만한 수준의 그래픽은 보여주지 못하고 있으며 STATGRAPHICS는 그래픽 면에서는 만족할 만한 수준이다.

4. Box-Plot의 최적 산법과 결론

지금까지의 산법과 출력결과에 대한 분석으로 아래와 같은 결론을 내릴수 있다.

사분위수의 최적산법 : SAS의 Univariate Procedure에서 사용하고 있는 4)의 산법이나 7), 8)의 산법.
Fence를 정하기 위한 최적 k값 : 1.5
중위수의 근사신뢰구간 : Hettmansperger과 Sheather[5]에 의해 제안된 방법.
출력결과 : 고해상도 그래픽을 실현할 수 있어야 함.

참 고 문 헌

[1] Cleveland, W.S.(1985), *The Elements of Graphing Data*, Monterey, CA : Wadwrorth.

[2] Frigge, M., Hoaglin, D.C. and Iglewicz, B., "Some Implementation of the Box-Plot", *The American Statistician*, February 1989, Vol.43, No.1.

[3] Hoaglin, D.C., Iglewicz, B.(1987), "Fine-Tuning Some Resistant Rules for Outlier Labeling.", *Journal of the American Statistical Association*, 82, 1147-1149.

[4] Hoaglin, D.C., Iglewicz, B., and Tukey, J.W.(1986), "Performance of some Resistant Rules for Outlier Labeling.", *Journal of the American Statistical Association*, 81, 991-999.

[5] T.P. Hettmansperger and S.J. Sheather(1986), "Confidence Intervals Based on Interpolated Order Statistics", *Statistical and Probability Letters*, Volume 4, No.2, pp.75-79.

[6] Norusis, M.J., SPSS Inc., *SPSS PC$^+$ for the IBM PC/XT/AT.*

[7] Ryan, B.F., Joiner, B.L., Ryan, T.A. Jr, *Minitab Hand Book & Reference Manual 2nd ed.*, Duxbury Press, Boston.

[8] MINITAB Inc., *MINITAB Reference Manual Release 7*, April 1989.

[9] SAS Institute Inc., *SAS Procedure Guide for Personal Computer Version 6 edition*, Cary, NC Author.

[10] SAS Institute Inc.(1985), *SAS User's Guide : Basics(Version 5 ed.)*.Cary, NC Author.

[11] STSC Inc, *STATGRAPHICS User's Guide*, 1988.

[12] Tukey, J.W.(1970,1977), *Exploratary Data Analysis*. Reading, MA : Addison - Wesley Publishing Co.

[13] Serfling, R.J.(1980), *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, Inc.

Box-Plot Algorithm used in Packages†

Seung Whan Kim* · Hongsuk Jorn**

ABSTRACT

We want to derive the optimal Box-Plot Algorithm by comparing and analyzing Box-Plots used in several packages.

"Box-Plot" is one of the most frequently used EDA procedures. Many computer packages give Box-Plot, but their results are different each other even for the same data set. Of course, the differences are caused by their algorithms. In this note, we compared their algorithms and gave some suggestions.

† This research is partially supported by Inha University research fund, 1991.

* Department of Mathematics, Inha University, 253, Nam-Gu, Yong Hyun-Dong, Incheon, 402-751, Korea

** Department of Statistics, Inha University, 253, Nam-Gu, Yong Hyun-Dong, Incheon, 402-751, Korea
E-mail BG02@KRINHA