▨ 연구논문

공정평균을 관리하기 위한 로버스트 관리도 - Robust Control Chart for the Control of the Process Mean -

이 병근*
Lee, Byung Gun
정 현석*
Jung, Hyun Seok
남 호수*
Nam, Ho Soo

Abstract

Control chart is a very extensively used tool in testing whether a process is in a state of statistical control or not. In this paper, a robust control chart for variables is proposed, which is based on the Huber's M-estimator. The Huber's M-estimator is a well-known robust estimator in sense of distributional robustness. In the proposed chart, the estimation of the process deviation is modified to have a stable level and high power. To compare the performances of the proposed control chart with the classical $\bar{x}$-chart, some Monte Carlo simulations are performed. The simulation results show that the robust control chart has good performance.

1. 서론

일반적으로 생산공정에서 제품의 품질을 항상 균일하게 생산해 내는 것은 불가능하다. 공정의 설계가 잘 되어 있고, 그 공정을 잘 운영하더라도 제품의 품질 특성에 영향을 주는 요인은 항상 내재되어 있기 때문이다. 만약 품질변동이 우연원인(chance cause) 즉, 작업자간의 숙련도 차이, 원자재간의 미세한 품질차이, 동일한 생산설비간의 차이, 등에 의해서만 일어난다면 생산공정은 정상상태에 있다고 볼 수 있다. 그러나 작업자의 부주의, 불량자재의 사용, 생산설비상의 이상, 등의 만성적으로 존재하는 것이 아닌 이상원인(assignable cause)이 발생하면 공정의 관리상태에 의문이 생기게 된다. 이러한 공정의 관리상태를 검토해 볼 수 있는 도구로써 흔히 사용되는 것이 관리도(control chart)이다.

* 동서대학교 정보시스템공학부 산업공학전공

본 논문에서는 계량형 데이터의 관리도에서 흔히 사용되는 $\bar{x}$-관리도가 이상점(outliers)에 민감하기 때문에 이에 대한 단점을 보완한 로버스트 관리도(robust control chart)를 제안하고자 한다. 제안하고자 하는 관리도는 공정의 평균을 관리하는 도구로써 공정평균의 로버스트한 추정량인 M-추정량에 기초하고 있다. 또한, 기존의 $\bar{x}$-관리도와 제안된 관리도를 비교하기 위하여 Monte Carlo 모의실험을 실시하였으며, 검정력함수(power function) 또는 검사특성곡선(operating charistic curve)을 통한 두 관리도를 비교분석하여 제안된 관리도의 우수성을 입증하고자 한다.

2. 로버스트 관리도

2.1 M-추정량

Huber(1964, [2])가 M-추정량을 제안한 이래, 많은 연구자들([1], [3], [5])에 의하여 M-추정량의 표본평균에 대한 로버스트성 및 효율성에 관한 연구가 이루어졌다. 일반적으로 다음과 같은 일표본 위치문제(one-sample location) 모형에서,

$$x_i = \mu + \varepsilon_i, \quad i = 1, 2, \cdots, n,$$

위치모수에 대한 Huber의 M-추정량은 목적함수(object function)

$$Q(\mu) = \sum_{i=1}^{n} \rho(x_i - \mu)$$

를 최소화하는 $\hat{\mu}$, 또는 음방정식(implicit equation)

$$\sum_{i=1}^{n} \psi(x_i - \mu) = 0$$

의 해인 $\hat{\mu}$으로 정의된다. 여기서 ρ가 볼록함수이고 미분가능할 때는 $\rho' = \psi$가 된다. 예를 들어 $\rho(x) = x^2/2$이면 $\psi(x) = x$이 되고, 이때 대응되는 μ의 추정량은 $\bar{x}$가 된다. M-추정량은 ρ-함수 또는 대응되는 ψ-함수의 형태에 따라 추정량이 달라지며, Huber가 제안한 ψ-함수의 형태는 다음과 같다.

$$\psi(x) = \begin{cases} x, & |x| \le c \\ c \cdot sign(x), & |x| > c \end{cases}$$

여기서 c는 조율상수(tuning constant)로서 $c = 1.5$는 많은 모의실험을 통하여 합리적인 값으로 사용되는 상수 중의 하나이다. c의 값이 작아지면 로버스트 추정량(robust estimator)을 구할 수 있으나 효율이 떨어지게 되고, 반면 큰 c의 값에 대하여 효율은 높아지지만 이상점의 영향을 제어하기 힘들어진다. Huber가 제안한 M-추정량의 동기는 일반적으로 오염된 자료(contaminated data), 또는 길고 두터운 꼬리를 가지는 분포(long and heavy-tailed distribution)에서의 표본에서 표본평균 $\bar{x}$가 갖는 비로버스트성(non-robustness)에 있다. 즉, 표본에 이상점(outliers)이 있을 경우 표본평균은 이상점에 심각하게 영향을 받는다. 그러나 ψ-함수의 구조에서 볼 수 있듯이 Huber의 M-추정량에서는 이상점의 영향이 감소되므로 이상점에 대하여 둔감한 로버스트 추정량을 구할 수 있다.

2.2 M-추정량의 수치해법

표본평균 $\bar{x}$와는 달리 M-추정량은 위치-척도-등가변환(location-scale-equivariant)의 조건을 만족하지 못한다[5]. 즉, 일반적으로 M-추정량에서는 다음의 조건이 성립하지 않는다.

$$\hat{\mu}(bx_1+a, \cdots, bx_n+a) = b\hat{\mu}(x_1, \cdots, x_n) + a$$

따라서 척도에 무관하게 정의되어 사용될 수 있는 척도불변(scale invariant) M-추정량은 다음과 같은 음방정식의 해 $\hat{\mu}$로 정의될 수 있다.

$$\sum_{i=1}^{n} \psi\left(\frac{x_i-\mu}{s}\right) = 0 \tag{2.1}$$

여기서 s는 척도(σ)의 로버스트한 추정량으로서, 절대편차중앙값(median absolute deviation; MAD)에 기초한 추정량이 흔히 사용된다. 즉,

$$s = 1.483 \cdot median_i\{|x_i - median_j(x_j)|\}$$

여기서 1.483은 정규분포에서 s가 σ의 일치추정량이 되도록 하는 상수이다.

식 (2.1)의 해는 μ의 초기추정치(initial estimate) $\hat{\mu}_0$에 대하여 일차 테일러 급수전개(first-order Taylor series expansion)를 한 다음 반복법에 의하여 얻을 수 있으며, 초기추정치로는 중앙값이 흔히 사용된다. 즉, M-추정량을 얻기 위한 알고리즘은 다음과 같다.

$$\hat{\mu}_m = \hat{\mu}_{m-1} + s \cdot \frac{\sum\psi[(x_i - \hat{\mu}_{m-1})/s]}{\sum\psi'[(x_i - \hat{\mu}_{m-1})/s]}, \quad m=1,2, \cdots$$

한편, 적절한 조건하에서 M-추정량($\hat{\mu}$)은 정규분포로 근사될 수 있는데, 표본크기 n이 충분히 클 경우 다음과 같은 점근 정규성을 갖는다.

$$\sqrt{n}(\hat{\mu}-\mu) \xrightarrow{d} N(0, \sigma^2)$$

여기서 σ^2은 $\sqrt{n}\hat{\mu}$의 점근분산으로, 일단계 M-추정량(one-step M-estimator)을 사용할 경우, 다음과 같이 추정될 수 있다.

$$\hat{\sigma}^2 = s^2 \cdot \frac{\frac{1}{n}\sum\psi^2\left(\frac{x_i-\hat{\mu}_0}{s}\right)}{\left[\frac{1}{n}\sum\psi'\left(\frac{x_i-\hat{\mu}_0}{s}\right)\right]^2} \tag{2.2}$$

2.3 M-추정량에 기초한 로버스트 관리도

일반적으로 관리도는 한 개의 중심선(center line; CL)과 한 쌍의 관리한계선(upper and lower control limit; UCL and LCL)으로 구성된다. 앞 절에서 설명된 M-추정량의 관리도에 대한 적용을 위하여 우선 다음의 모형을 고려해 보자.

$$x_{ij} = \mu_0 + \varepsilon_{ij}, \quad i=1,2,\cdots,k; \quad j=1,2,\cdots,n \tag{2.3}$$

여기서 k는 부분군(subgroup)의 개수이고 n은 부분군의 크기(subgroup size)이며 μ_0는 공정의 목표치 또는 공정평균을 의미한다. x_{ij}는 i번째 부분군에서 얻어진 j번째 데이터이며, ε_{ij}는 오차항으로 서로 독립이고 기대값은 0, 분산은 σ^2인 분포를 따른다. 만약 ε_{ij}가 $N(0, \sigma^2)$을 따른다면 표본평균 $\bar{x}$는 가장 적합한 공정평균의 추정량이 될 것이다. 그러나 일반적으로 오차항이 정규분포를 따른다는 가정이 무리일 경우가 빈번하며, 이 경우 표본평균은 데이터에 포함되어 있을 수 있는 이상점에 영향을 받아 왜곡된 값을 공정의 평균으로 추측하는 경우가 생길 수 있다. 즉, 표본평균에 기초한 $\bar{x}$-관리도는 공정의 산포가 클 경우, 또는 대형오차(gross-errors) 등으로 이상점이 발생할 때 공정평균의 변화를 제대로 탐지해내지 못하게되는 단점이 있다.

본 논문에서는 오차항에 대하여 특정한 형태의 분포를 가정하지 않고, 다만 오차항의 기대값과 분산이 각각 0, σ^2이라는 가정 하에서 M-추정량에 기초한 관리도를 제안하고자 한다. 우선

각 부분군에 속한 n개의 데이터에 기초하여 M-추정량을 구하면 k개의 M-추정량들이 얻어지게 된다. 이 추정량들을 관리도에 타점하게 되는데, 중심선은 이들의 평균이 된다. 즉,

$$CL = \frac{1}{k}\sum_{i=1}^{k}\hat{\mu}^{(i)}$$

여기서, $\hat{\mu}^{(i)}$는 i번째 부분군에서 얻어진 M-추정량이다.

한편, 관리한계선을 정하는 방법은 두 가지로 생각할 수 있는데, 첫 번째 방법은 각 부분군에서 M-추정량을 구하고 식(2.2)에 기초하여 M-추정량의 점근분산을 추정하여 관리한계선으로 사용하는 것인데 이 방법은 문제점을 갖고 있다. 식(2.2)의 분산추정식은 M-추정량의 점근분산을 추정한 것인데 이는 표본크기(부분군의 크기)가 충분히 클 때 점근분산의 일치추정량으로서 의미를 갖는다. 그러나 일반적으로 관리도에서 부분군의 크기는 3~7 정도로 아주 작은 편이다. 설상가상으로 식(2.2)의 추정량에 사용되는 MAD는 공정산포의 로버스트한 추정량이기는 하나 정규분포에서 효율이 떨어지며 과소추정(under-estimate)되는 경향이 있다. 각 부분군에서 구한 $\hat{\sigma}$들의 평균에 기초하여 관리한계선을 설정할 경우 소표본(small sample)에서는 지나치게 좁은 관리한계선을 갖게되어 한계를 벗어나는 점들이 빈번해진다. 즉, 검정의 측면에서 볼 때 제1종 오류의 제어가 안되는 것이다. 이러한 단점을 보완하기 위하여 모형 (2.3)에서 기술한 바와 같이 부분군들의 공정산포가 동일하다는 가정에 근거하여 전체의 자료, 즉 $k\times n(=N)$개의 데이터에 기초한 공정산포를 추정하여 관리한계선을 설정하고자 한다. 이 경우 전체 데이터의 개수 (N)가 충분히 커지기 때문에 공정의 산포가 안정적으로 추정될 수 있다.

따라서 제안하고자 하는 관리도에서 공정산포는 다음과 같은 방법에 의하여 추정될 수 있다.

$$\hat{\sigma}^2 = s^2 \cdot \frac{\frac{1}{N}\sum_{l=1}^{N}\psi^2\left(\frac{x_l - median\{x_l\}}{s}\right)}{\left[\frac{1}{N}\sum_{l=1}^{N}\psi'\left(\frac{x_l - median\{x_l\}}{s}\right)\right]^2} \tag{2.4}$$

여기서 $s = 1.483 \cdot median_l\{|x_l - median_l(x_l)|\}$, $l=1,2,\cdots,N$이다. 식(2.4)에 근거하여 3σ-관리한계선을 설정하면 관리상한선과 하한선은 다음과 같이 주어진다.

$$UCL = CL + 3\frac{\hat{\sigma}}{\sqrt{n}} \; ; \quad LCL = CL - 3\frac{\hat{\sigma}}{\sqrt{n}}$$

3. 몬테칼로 모의실험을 통한 비교

2장에서 제안된 로버스트 관리도를 기존의 $\bar{x}$-관리도와 비교해보기 위하여 다양한 상황에서 몬테칼로(Monte Carlo) 모의실험을 실시하였으며, 모의실험에서 고려한 모형은 다음과 같다.

$$x_{ij} = 10.0 + \varepsilon_{ij} + \delta \; ; \quad j=1,2,\cdots,5, \quad i=1,2,\cdots,2000$$

여기서 δ는 공정평균의 변화를 의미하는 값으로서 다음과 같이 변화시키면서 관리도의 탐지능력을 측정하였다.

$$\delta = a\times\frac{\sigma}{2} \; ; \quad a=0,1,2,\cdots,6$$

모의실험에서 고려한 오차항의 분포는 오염된 정규분포(contaminated normal distribution; $CN(\alpha, \sigma)$)로서 분포함수 $F(x)$는 다음과 같다. 즉, α를 표준정규분포에서의 오염비율(contaminated rate)이라 하고, $\Phi(x)$를 표준정규분포의 분포함수라 하면

$$F(x) = (1-\alpha)\Phi(x) + \alpha\Phi(x/\sigma)$$

으로 나타낼 수 있다. 또한, 모의실험에서 사용된 각 부분군에서 공정평균의 추정량은 일단계

M-추정량(one-step M-estimator)이며 초기추정량으로 중앙값(median)을, 조율상수 (c)로서 1.5를 사용하였다. 한편, 모의실험의 모든 계산은 S-PLUS(Statistical Sciences[4])를 이용하여 이루어 졌다.

[표 3.1]은 모의실험(2000번 반복)에서 얻어진 표본평균($\bar{x}$) 및 M-추정량(M)들의 평균(Mean)과 평균제곱오차(MSE)를 계산한 결과이다. [표 3.1]에서 목표값은 10.0이며, 편의(bias)의 측면에서 보면 표본평균과 M-추정량은 비슷한 정도의 편의를 가지고 있음을 알 수 있다. 그러나 평균제곱오차의 측면에서 보면 $N(0, 1)$경우를 제외한 모든 경우에서 M-추정량의 MSE가 표본평균의 그것보다 훨씬 작은 값을 가짐을 알 수 있으며, 이로부터 M-추정량의 로버스트한 성질을 알 수 있다.

[표 3.1] $\bar{x}$와 M-추정량의 비교

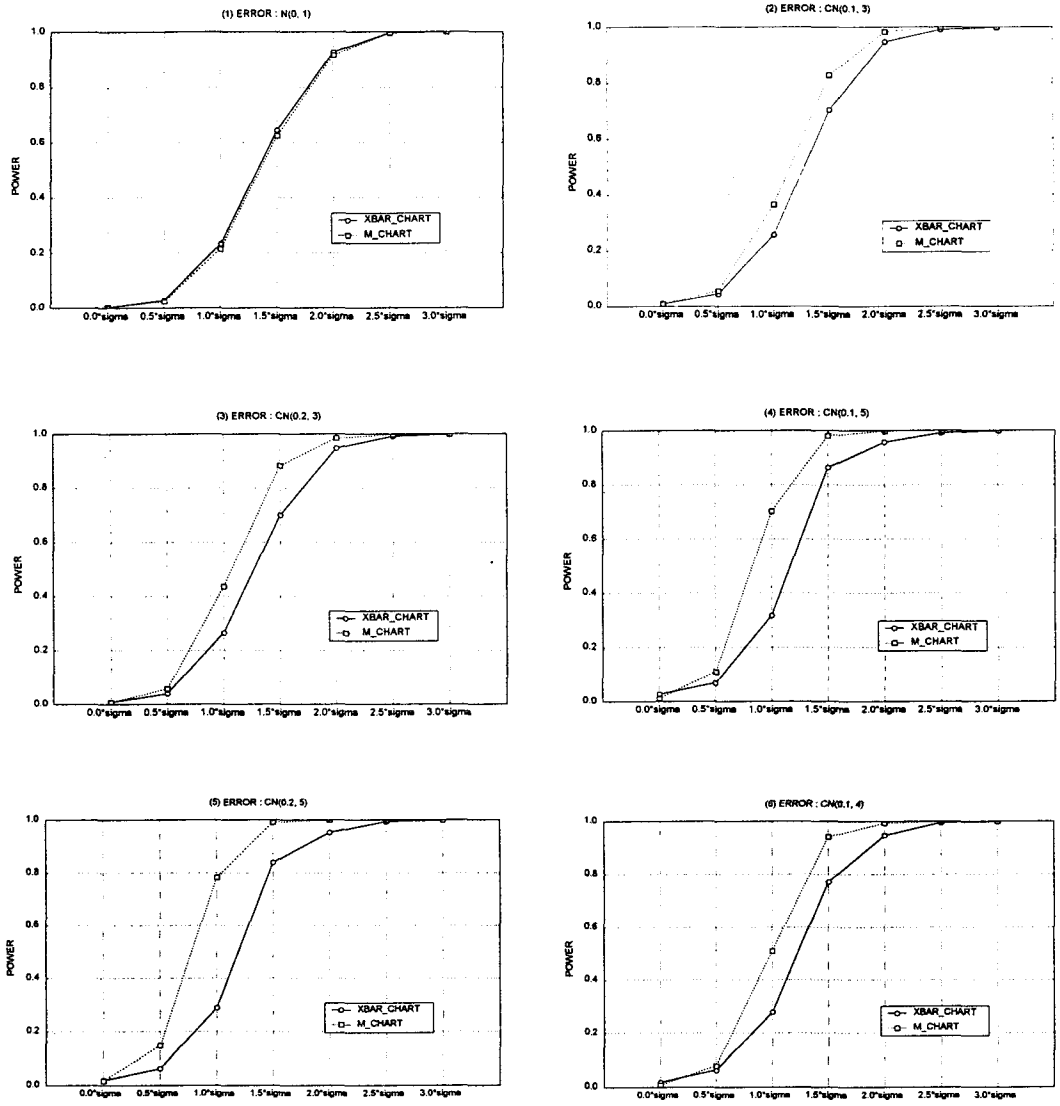
오차분포		$N(0, 1)$	$CN(0.1, 3)$	$CN(0.2, 3)$	$CN(0.1, 5)$	$CN(0.2, 5)$	$CN(0.1, 4)$
Mean	$\bar{x}$	9.9891	9.9914	10.0089	9.9831	9.9991	10.0213
	M	9.9912	9.9899	10.0065	9.9851	10.0026	10.0058
MSE	$\bar{x}$	0.1997	0.3433	0.5166	0.6961	1.2098	0.5186
	M	0.2035	0.2756	0.3586	0.3714	0.4824	0.3260

한편, [표 3.2]는 다양한 오차항의 분포에서 주어진 δ에 대하여 관리도가 공정의 변화를 탐지해낼 가능성을 관리한계선을 이탈하는 비율로 나타낸 값이다. 여기서 탐지력은 경험적 검정력(empirical power)으로 2000개의 점들 중에서 관리한계선을 벗어나는 점의 개수를 비율로 나타낸 것이다. [표 3.2] 및 [그림 3.1]에서 볼 수 있듯이 공정평균이 목표값 μ_0(=10)에서 δ만큼 벗어날 때 $\bar{x}$-관리도(XBAR_CHART)는 오차항의 분포가 $N(0, 1)$일 경우를 제외한 모든 고려된 분포에서 로버스트 관리도(M_CHART)보다 검정력의 의미를 갖는, 즉, 실제의 공정평균이 목표값에서 벗어났을 때 그것을 제대로 탐지해내는 능력이 떨어짐을 알 수 있다.

이러한 결과에 근거해 볼 때 로버스트한 추정량인 M-추정량에 기초한 관리도의 공정변화 탐지능력이 기존의 $\bar{x}$-관리도에 비하여 우수하다고 할 수 있겠다.

[표 3.2] $\bar{x}$-관리도와 로버스트 관리도의 공정변화 탐지력의 비교

오차분포	Method	δ						
		0.0	$0.5\times\sigma$	$1.0\times\sigma$	$1.5\times\sigma$	$2.0\times\sigma$	$2.5\times\sigma$	$3.0\times\sigma$
$N(0, 1)$	$\bar{x}$-관리도	0.0022	0.0285	0.2325	0.6445	0.9270	0.9965	1.0000
	Robust	0.0021	0.0240	0.2155	0.6255	0.9165	0.9970	1.0000
$CN(0.1, 3)$	$\bar{x}$-관리도	0.0090	0.0440	0.2560	0.7030	0.9470	0.9900	0.9985
	Robust	0.0090	0.0550	0.3655	0.8275	0.9830	0.9995	0.9995
$CN(0.2, 3)$	$\bar{x}$-관리도	0.0085	0.0405	0.2660	0.7000	0.9480	0.9915	0.9995
	Robust	0.0083	0.0585	0.4370	0.8835	0.9855	0.9985	1.0000
$CN(0.1, 5)$	$\bar{x}$-관리도	0.0275	0.0685	0.3185	0.8650	0.9585	0.9925	0.9995
	Robust	0.0120	0.1095	0.7025	0.9820	0.9985	1.0000	1.0000
$CN(0.2, 5)$	$\bar{x}$-관리도	0.0200	0.0620	0.2885	0.8395	0.9535	0.9935	0.9990
	Robust	0.0180	0.1475	0.7820	0.9915	0.9995	0.9990	1.0000
$CN(0.1, 4)$	$\bar{x}$-관리도	0.0185	0.0620	0.2785	0.7725	0.9480	0.9950	0.9990
	Robust	0.0085	0.0785	0.5100	0.9430	0.9935	0.9995	1.0000



[그림 3.1] 오차항의 분포에 따른 두 관리도의 공정변화 탐지력 비교

4. 결론

본 논문에서는 기존의 $\bar{x}$-관리도의 단점을 보완한 로버스트 관리도를 제안하였다. 제안된 관리도는 공정평균의 로버스트 추정량인 M-추정량에 기초하며, 관리한계선이 공정변화의 탐지를 유효하게 할 수 있도록 공정산포의 추정에 관한 새로운 방법에 근거하고 있다.

제안된 관리도와 기존의 $\bar{x}$-관리도의 공정변화 탐지능력을 비교하기 위하여 모의실험을 해본 결과 제안된 로버스트 관리도의 우수함이 밝혀졌다.

참고문헌

[1] Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. and Tukey, J.W. (1972). *Robust Estimates of Location: Survey and Advances*, Princenton University Press.

[2] Huber, P.J. (1964). Robust Estimation of a Location Parameter, *The Annals of Mathmatical Statistics*, Vol. 35, 73-101.

[3] Huber, P.J. (1972). Robust Statistics: A Review, *The Annals of Mathmatical Statistics*, Vol. 43, 1041-1067.

[4] Statistical Sciences(1994). *S_PLUS for Windows User's Manual*, Siattle: Statistical Sciences.

[5] 송 문섭(1996). 로버스트 통계, 자유아카데미.