研究論文

지니(Gini)의 평균차이에 기초한 $\overline{X}$-관리도

남호수* · 강중철**
* 동서대학교 정보시스템공학부, ** 동서대학교 e-비즈니스학부

An $\overline{X}$-Control Chart Based on the Gini's Mean Difference

Ho Soo Nam* · Jung Cheol Kang**
* Division of Information System Engineering
** Division of e-Business, Dongseo University

Keywords: $\overline{X}$-Control Chart, Gini's Mean Difference, Process Deviation

Abstract

Estimation of the process deviation is an important problem in statistical process control, especially in the control chart, process capability analysis or measurement system analysis. In this paper we suggest the use of the Gini's mean difference for the estimation of the process deviation when we design the control limits in construction of the control charts.

The efficiency of the Gini's mean difference was well explained in Nam, Lee and Jung(2000). In this paper we propose an $\overline{X}$ control chart which use the control limits based on the Gini's mean difference. In various classes of distributions, the proposed control chart shows good performance.

1. 서론

일반적으로 일정한 조건으로 작업을 하더라도 얻어지는 제품(또는 반제품)의 품질 특성치는 어떤 값을 중심으로 하여 산포하고 있다. 공정의 설계가 잘 되어 있고, 그 공정을 잘 운영하더라도 제품의 품질 특성에 영향을 주는 요인은 항상 내재되어 있기 때문에 생산공정에서 제품의 품질을 항상 균일하게 생산해 내는 것은 불가능하다. 이러한 요인에 의하여 제품의 특성은 변동을 하게 마련이다. 품질특성의 변동을 측정하는 일은 여러 가지 측면에서 중요한 작업중의 하나이다. 일반적으로 통계적 공정관리 기법에서 사용되는 도구들, 예를 들면 가설검정, 관리도법, 공정능력분석, 측정시스템분석 등의 여러 분야에서 산포 또는 변동의 제대로 된 추정치는 공정진단 및 분석의 기초가 된다고 할 수 있겠다.

일반적으로 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.

본 논문에서는 남호수, 이병근, 정현석(2000)의 논문에서 제안한 Gini의 평균차이(mean difference)에 기초한 산포추정법의 우수성(효율성)에 기초하여, 이를 이용하여 관리한계선을 설정하는 형태의 $\overline{X}$-관리도를 제안하고자 한다.

제안하고자 하는 관리도는 공정의 평균을 관리하는 도구로써 가장 널리 사용되는 $\overline{X}$-관리도 형태이며, 기존의 관리도에서 공정산포의 추정법으로 흔히 사용되는 표준편차 또는 범위(range) 대신에 Gini의 평균차이(mean difference)에 기초한 추정량이 사용된다.

제안된 관리도는 Monte Carlo 모의실험을 통하여 기존의 $\overline{X}$-관리도와 비교해 보고자 하며, 검정력함수(power function) 또는 검사특성곡선(operating characteristic curve)의 측면에서 제안된 관리도의 우수성을 입증하고자 한다.

2. 지니의 평균차이를 이용한 $\overline{X}$-관리도

2.1 $\overline{X}$-관리도

일반적으로 관리도는 한 개의 중심선(center line; CL)과 한 쌍의 관리한계선(upper and lower control limit; UCL and LCL)으로 구성된다. 제안하고자 하는 관리도를 구체적으로 설명하기 위하여 우선 다음과 같은 모형을 고려해 보자.

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

여기서 k는 부분군(subgroup)의 개수이고 n은 부분군의 크기(subgroup size)이다. x_{ij}는 i번째 부분군에서 얻어진 j번째 데이터이며, ε_{ij}는 오차항으로 서로 독립이고 기대값은 0, 분산은 σ^2인 분포를 따른다. 또한 μ_0는 공정의 목표치 또는 공정평균을, σ는 공정의 산포를 의미한다. 우선 각 부분군에 속한 n개의 데이터에 기초하여 $\overline{X}_i$를 구하고 이를 타점통계량으로 한다. 여기서 공정산포의 추정은 범위나 표준편차를 이용하는 기존의 방법과는 달리 본 논문에서는 지니의 평균차이를 이용하여 각 부분군에서의 산포를 추정하고자 하며, 이에 기초하여 관리한계선을 설정하는 방법을 제안하고자 한다. 즉, 중심선은 기존의 $\overline{X}$ 관리도와 같이 다음의 형태로 표현될 수 있다.

$$CL = \frac{1}{k}\sum_{i=1}^{k}\overline{X}_i = \overline{\overline{X}}$$

한편, $\overline{X}$ 관리도에서 관리한계선(control

limits; UCL 또는 LCL)은

중심선 ± 3×표준오차(타점통계량)

$$= \overline{\overline{X}} \pm 3\times\hat{\sigma}/\sqrt{n}$$

으로 주어지게 되는데, 여기서 공정산포 σ는 각 부분군에서 얻어진 산포의 추정량들의 평균으로 추정되어질 수 있다.

즉, 각 부분군에서 얻어진 σ의 추정량을 $\hat{\sigma_i}$ 이라고 하면 σ는 다음과 같이 추정되어질 수 있다.

$$\hat{\sigma} = \sum_{i=1}^{k} \hat{\sigma_i}/k$$

2.2 공정산포의 추정

우선 이 절에서는 앞절에서 소개된 모형하에서, 단 오차항의 분포에 대하여 특정형태의 분포를 가정하지 않고 고려될 수 있는 각 부분군에서의 산포 추정량들 가운데, 널리 사용되는 범위(range) 및 표본표준편차에 기초한 공정산포(σ)의 추정량과 지니(Gini)의 평균차이(Gini's mean difference, Gini(1912))에 기초한 공정산포의 추정법을 간단하게 소개하고자 한다.

2.2.1 표본표준편차(sample standard deviation)에 의한 추정)

표본표준편차에 기초한 σ의 추정은 다음과 같이 할 수 있으며,

$$\hat{\sigma_{si}} = \frac{s_i}{c_4(n)} = \frac{1}{c_4(n)}\sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_{ij}-\overline{x_i})^2}$$

여기서 s_i는 흔히 쓰이는 표본표준편차이고, $c_4(n)$은 데이터가 정규분포에서의 표본일 때 $E(\hat{\sigma_{si}})=\sigma$, 즉 불편성(unbiasedness)을 갖게 해 주는 상수로서 n의 함수로 표현되며, $c_4(n)$은 다음과 같다(Devore(1995)).

$$c_4(n) = \sqrt{\frac{2}{n-1}}\frac{\Gamma(n/2)}{\Gamma[(n-1)/2]}$$

여기서 $\Gamma(\cdot)$은 감마함수이며, $c_4(n)$의 값은 1보다 작으나 표본의 크기 n이 커짐에 따라 1에 가까워지며, 대부분의 경우 표본표준편차 s_i를 간편하게 산포 σ의 추정치로 쓴다. 그러나 표본의 크기가 작을 경우에는 $E(s_i) < \sigma$으로 s_i는 σ를 과소추정(under estimate)하게되는 경향이 있다. 따라서 관리도에서 한계선을 설정할 때 사용되는 산포의 추정값으로 표본표준편차 s_i보다는 $\hat{\sigma_{si}}$를 쓰는 것이 합리적이라고 할 수 있겠다.

2.2.2 범위(range)를 이용한 추정

범위(range)를 이용한 산포의 추정은 다음과 같이 할 수 있으며,

$$\hat{\sigma_{Ri}} = \frac{R_i}{d_2(n)}$$

여기서 i번째 부분군의 범위는

$$R_i = \max_j\{x_{ij}\} - \min_j\{x_{ij}\}$$

이고, $d_2(n)$은 데이터가 정규분포에서의 표

본일 때 $E(\widehat{\sigma_{Ri}}) = \sigma$, 즉 범위에 기초한 추정량이 불편성(unbiasedness)을 갖게 해 주는 상수로서 표본의 크기(부분군의 크기) n의 함수로 표현된다.

범위에 기초한 추정량은 관리도 또는 공정능력 평가에 가장 널리 사용되고 있는 산포의 추정량이다. 이는 Montgomery(1991)에 의하면 소표본일 경우 표본표준편차에 근접하는 상대효율(relative efficiency)을 보여주고 있다. 그러나 표본크기가 커지면 범위에 기초한 추정량의 효율은 상당히 떨어지는 것으로 알려져 있다.

실제로 본 연구에서 모의실험을 통하여 살펴본 바에 의하면 표본의 크기가 $n=5$ 정도 일 때는 범위를 이용한 추정과 표준편차에 기초한 추정은 그 효율성이 비슷함을 알 수 있었다.

2.2.3 Gini의 평균차이(mean difference)에 기초한 추정

지니(Gini)의 평균차이(Gini's mean difference)는 Gini(1912))에 의하여 제안된 산포의 추정방법으로 다음과 같이 나타낼 수 있다.

$$\widehat{\sigma_{Gi}} = \frac{k}{{}_nC_2} \sum_{l=2}^{n} \sum_{j=1}^{l-1} | x_j - x_l |$$

여기서 $k = \sqrt{\pi}/2$를 주로 쓰며, 이는 데이터가 정규분포에서의 표본일 경우 추정량의 불편성을 담보한다. 일반적으로 분포함수 $F(x)$를 갖는 연속확률변수 X에 대한 평균차이 Δ_R은 다음과 같이 표현될 수 있으며,

$$\Delta_R = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |x - y| \, dF(x) \, dF(y)$$

Δ_R의 적률방식에 의한 추정량(method of moment estimator)은 다음과 같이 나타낼 수 있다.

$$\widehat{\Delta_R} = \frac{1}{{}_nC_2} \sum_{l=2}^{n} \sum_{j=1}^{l-1} | x_j - x_l |$$

즉, 지니(Gini)의 평균차이는 이에 기초한 산포의 추정량으로서, 관리도 및 공정능력 평가를 위하여 표본의 크기가 작을 때 데이터의 재사용(reuse) 또는 반복사용을 통한 정보활용도를 높여 산포를 추정하기 때문에 추정의 정도(precision) 또는 효율성을 높일 수 있으며, 이상점(outliers)에 덜 민감하여 산포 σ에 대한 로버스트 추정법으로 고려될 수 있다.

지니의 평균차이에 기초한 산포의 추정법은 그 동안 계산의 번거로움 때문에 기피되어온 면이 없지 않다. 그러나 오늘날과 같이 컴퓨터의 사용 및 사용자의 인터페이스가 강화된 프로그램 환경에서 계산알고리즘의 복잡성 또는 번거로움은 더 이상 설득력을 얻기 힘들 것으로 생각된다.

3. 모의실험을 통한 비교

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

$$x_{ij} = 10.0 + \varepsilon_{ij} + \omega$$

$$; \ j = 1, 2, \cdots, 5, \quad i = 1, 2, \cdots, 10000$$

여기서 각 부분군의 크기(n)는 5개이며, ω는 공정평균의 변화를 의미하는 값으로서 ω의 값이 0일 경우에 관리한계선을 벗어나는 점들의 비율은 가설검정에서 유의수준(significance level)과 같은 의미로 해석될 수 있겠다.

또한, ω의 값이 증가할수록 공정평균이 목표값(μ_0)에서 멀어짐을 뜻하며, 이 때는 관리도의 탐지력은 다음과 같이 표현될 수 있다.

$$\text{탐지력} = \frac{\text{관리한계선을 이탈하는 점들의 수}}{\text{모든 점들의 수}}$$

여기서 오차항의 분포에 따라 관리도의 탐지력을 측정하기 위하여 이동시키는 ω의 값이 다른데, 이를 5개의 수준으로 표시하고자 한다. 즉, 공정평균의 변화수준을 δ라고 표현하면 δ의 수준이 0일 때는 ω의 값이 0이 되고 δ의 수준이 증가할수록 ω의 값은 증가하게 된다.

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

$$F(x) = (1-\kappa)\Phi(x) + \kappa\Phi(x/\gamma)$$

으로 나타낼 수 있다. 또한, 모의실험의 모든 계산은 S-PLUS(Statistical Science(1994))를 이용하여 이루어 졌다.

[표 3.1]은 다양한 오차항의 분포에서 주어진 수준 δ에 대하여 관리도가 공정의 변화를 탐지해낼 가능성을 관리한계선을 이탈하는 점들의 비율(탐지력)로 나타낸 값이다. [표 3.1]에서 볼 수 있듯이 공정평균이 목표값 μ_0(=10)에서 δ수준만큼(수준이 높을수록 목표값에서 많이 벗어남) 벗어날 때, 오차항의 분포가 표준정규분포일 경우에는 Gini의 평균차이를 이용한 관리도(G)는 표본표준편차에 기초한 산포추정량을 이용한 관리도(S) 또는 범위에 기초한 산포추정량을 이용한 관리도(R)에 비하여 관리도의 성능에 큰 차이가 없어 보인다.

반면, S와 R은 관리도의 성능이 모든 고려된 분포에서 비슷하게 나타남을 알 수 있다. 그러나 고려된 대부분의 분포에서 G관리도의 성능은 기존의 S 또는 R 관리도에 비하여 탐지력이 높음을 알 수 있다.

한편, 각 부분군의 크기(n)가 5인 경우의 결과만 기술하였으나 부분군의 크기가 5 근방에서 변화하더라도 결과에 큰 차이는 없는 것으로 나타났다.

다만 n이 클 경우 전술한 바와 같이 범위를 이용한 산포의 추정(R)이 비효율적이어서 탐지력이 떨어지는 면이 있는 것으로 나타났다.

이러한 결과에 근거해 볼 때, $\overline{X}$-관리도에서 관리한계선(관리상한선 및 관리하한선)의 설정, 즉, 산포의 추정은 Gini의 평균차이를 이용한 방법이 기존의 표본범위 또는 표본표준편차를 이용한 방법보다 우수하다고 할 수 있겠다.

4. 결론

본 논문에서는 공정평균을 관리하기 위한 관리도로서 기존의 범위(range) 또는 표준편

[표 3.1] 제안된 관리도와 기존의 관리도의 공정변화 탐지력 비교

오차항의 분포	Method	δ(level)				
		0	1	2	3	4
$N(0, 1)$	S	0.0036	0.2177	0.4669	0.6308	0.9335
	R	0.0037	0.2173	0.4672	0.6290	0.9334
	G	0.0035	<u>0.2178</u>	<u>0.4675</u>	<u>0.6312</u>	<u>0.9341</u>
$CN(0.1, 3)$	S	0.0128	0.2233	0.4367	0.7115	0.9245
	R	0.0124	0.2188	0.4301	0.7050	0.9223
	G	0.0131	<u>0.2387</u>	<u>0.4551</u>	<u>0.7323</u>	<u>0.9312</u>
$CN(0.2, 3)$	S	0.0102	0.0695	0.4950	0.7855	0.9216
	R	0.0100	0.0667	0.4819	0.7759	0.9182
	G	0.0104	<u>0.0785</u>	<u>0.5251</u>	<u>0.8043</u>	<u>0.9304</u>
$CN(0.1, 5)$	S	0.0274	0.2944	0.6179	0.8521	0.9533
	R	0.0267	0.2915	0.6122	0.8500	0.9527
	G	0.0287	<u>0.3435</u>	<u>0.6697</u>	<u>0.8737</u>	<u>0.9611</u>
$CN(0.2, 5)$	S	0.0209	0.3184	0.7911	0.8885	0.9397
	R	0.0203	0.3107	0.7849	0.8848	0.9380
	G	0.0251	<u>0.3760</u>	<u>0.8259</u>	<u>0.9045</u>	<u>0.9513</u>

S: 표본표준편차에 기초한 $\overline{X}$-관리도

R: 범위에 기초한 $\overline{X}$-관리도

G: 지니의 평균차이에 기초한 $\overline{X}$-관리도

*참고: 밑줄친 데이터는 각 δ의 수준에서 가장 큰 탐지력을 나타낸 값이다.

차에 기초한 $\overline{X}$-관리도와는 달리 지니의 평균차이에 기초하여 공정산포(σ)를 추정하고, 이에 근거하여 관리한계선을 설정하는 형태의 관리도를 제안하였다.

제안된 관리도와 기존의 $\overline{X}$-관리도의 공정변화 탐지능력을 비교하기 위하여 다양한 상황에서 모의실험을 해본 결과 제안된 관리도의 우수함이 밝혀졌다.

참고문헌

[1] Devore, Jay L.(1995). *Probability and S tatistics for Engineering and the Scie nces,* Duxbury.

[2] Gini, C.(1912). variabilita e mutabilita, c ontributo allo studio delle distribuzioni e delle relazioni statistiche. *Studio Eco nomico-Giuridici della R. Universita d*

i Cagliari. 3, part 2, 3-159.
[3] Montgomery, D. C.(1991). *Introduction to Statistical Quality Control*, Wiley.
[4] Statistical Sciences(1994). *S_PLUS for Windows User's Manual*, Siattle: Statistical Sciences.
[5] 남 호수, 이 병근, 정 현석(2000). "지니(Gini)의 평균차이를 이용한 공정산포의 추정", 산업경영시스템학회지, 23권, 제58집, pp. 113-118.