▩ 응용논문

지니(Gini)의 평균차이를 이용한 공정산포 추정
On the Estimation of the Process Deviation Based on the Gini's Mean Difference

남 호 수*
Nam, Ho Soo
이 병 근*
Lee, Byung Gun
정 현 석*
Jung, Hyun Seok

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 σ, the measure of the process deviation through a lots of simulations in various types of distributions.

The Gini's mean difference uses the differences of all possible pairs of data. This point will improve the efficiency of estimation. In various classes of distributions, the Gini's mean difference shows good performance, in sense of bias of estimates or mean squared errors.

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.

* 동서대학교 정보시스템공학부 산업공학전공

본 논문에서는 변동을 추정하는 여러 가지 기법들을 설명하고, 이들 추정법들의 장 · 단점을 논하고자 하며, 특히 여러 추정법들을 시뮬레이션을 통하여 비교함으로서 Gini의 평균차이(mean difference)에 기초한 추정법의 우수성(효율성)을 입증하고, 이를 기존의 표본표준편차 또는 범위에 기초한 산포추정의 대안으로 제시하고자 한다.

2. 산포(변동)의 측도

2.1 모형

산포의 추정에서 고려하고자 하는 모형은 다음과 같다.

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

여기서 μ는 특성치의 평균이며, ε_i는 오차항(error term)을 나타낸다. 이러한 형태의 모형은 계량형 관리도, 공정능력분석 및 측정시스템 분석 등의 문제로 간단하게 확장될 수 있다.

일반적으로 위의 모형에서 이론전개의 편의성을 위하여 오차항에 대하여 다음과 같은 가정을 많이 한다.

$$\varepsilon_i \sim N(0, \sigma^2)$$

즉, 오차항은 서로 독립이고, 동일한 분포(평균이 0이고, 분산이 σ^2인 정규분포; $N(0, \sigma^2)$)를 따른다는 가정을 하게 된다.

이러한 가정이 만족된다면 특성치의 평균값(모평균)은 표본평균에 의하여, 그리고 모분산(σ^2)은 표본분산에 의하여 추정될 수 있으며, 이는 불편성(unbiasedness) 및 최소분산성(minimum variance)의 측면에서 최적의 추정량이 될 수 있다. 그러나 특성의 모집단의 분포에 대한 이러한 가정은 현실적으로 많은 경우에 만족되지 않을 뿐만 아니라, 가정과 상당한 거리가 있는 경우도 허다하다.

그러나 대부분의 통계적 공정관리 기법에서 이를 간과하고 있는 경향이 있으며, 특히 현장에서는 만족되지 않는 가정에 기초한 결과를 맹신하는 것을 종종 볼 수 있다. 물론 이러한 문제에 대한 의논이 비정규(non-normal)분포상에서의 공정능력 평가 등의 문제에서 제기(Rodriguez(1992), 백재욱, 조진남(1999), 등)되기는 하였으나, 정규분포가 아닌 특정 분포상에서의 추론 및 접근법을 제한적으로 설명하고 있다. 이러한 대안적 접근법에서도 여전히 사용되는 산포의 추정방법은 표본표준편차(sample standard deviation) 또는 범위(range)에 기초한 추정인 것이 대부분이다.

특성치가 특정 분포를 따른다는 가정 하에서의 접근법 및 이론전개는 비록 그 분포가 기존의 정규분포가 아닌 다른 형태의 분포라 할지라도 그 의미가 적을 수밖에 없을 것이다. 왜냐하면 현실적으로 그 기법을 적용할 때는 특성치의 분포를 모르는 경우가 대부분이기 때문이다. 또한, 표본의 크기가 충분히 클 때 의미를 가질 수 있는 추정량의 쓰임새도 공정관리기법으로는 떨어질 수밖에 없다.

본 논문에서는 우선 다양한 공정관리 기법에서 공통적으로 필요한 산포의 추정, 그 중에서도 특성치의 모집단분포에 무관한(distribution-free), 그리고 특성치 데이터의 이상점(outliers)에 덜 민감한(robust) 추정법을 논하고자 한다. 여기서 모형에 대하여 오차항의 기대값이 0이라는 것과 분산이 σ^2이라는 것만 가정하기로 한다.

2.2 산포의 추정량

우선 이 절에서는 앞절에서 소개된 모형 하에서, 단 오차항의 분포에 대하여 특정형태의 분포를 가정하지 않고 고려될 수 있는 산포의 추정량들 가운데, 지니(Gini)의 평균차이(Gini's mean difference, Gini(1912))를 비롯한 흔히 사용되는 몇 가지를 소개하고자 한다.

2.2.1 표본표준편차(sample standard deviation)에 의한 추정

표본표준편차에 기초한 σ의 추정은 다음과 같이 할 수 있으며,

$$\widehat{\sigma_S} = \frac{s}{c_2(n)} = \frac{1}{c_2(n)}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

여기서 s는 흔히 쓰이는 표본표준편차이고, $c_2(n)$은 데이터가 정규분포에서의 표본일 때 $E(\widehat{\sigma_s}) = \sigma$ 즉 불편성(unbiasedness)을 갖게 해 주는 상수로서 n의 함수로 표현되며, $c_2(n)$은 다음과 같다.

$$c_2(n) = \sqrt{\frac{2}{n-1}}\frac{\Gamma(n/2)}{\Gamma[(n-1)/2]}$$

$c_2(n)$의 값은 1보다 작으나 표본의 크기 n이 커짐에 따라 1에 가까워지며, 대부분의 경우 표본표준편차 s를 간편하게 산포 σ의 추정치로 쓴다. 그러나 표본의 크기가 작을 경우에는 $E(s)<\sigma$ 으로 s는 σ를 과소추정(under estimate)하게되는 경향이 있다.

2.2.2 범위를 이용한 추정

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

$$\widehat{\sigma_R} = \frac{R}{d_2(n)}$$

여기서 $R = \max\{x_i\} - \min\{x_i\}$이고, $d_2(n)$는 데이터가 정규분포에서의 표본일 때 $E(\widehat{\sigma_R}) = \sigma$ 즉 범위에 기초한 추정량이 불편성을 갖게 해 주는 상수로서 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.

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.

$$\widehat{\sigma_G} = \frac{k}{{}_nC_2}\sum_{j=2}^{n}\sum_{i=1}^{j-1}|x_i-x_j|$$

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.

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

즉, 지니(Gini)의 평균차이는 이에 기초한 산포의 추정량으로서, 관리도 및 공정능력 평가를 위한 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 산포에 대한 로버스트 추정법으로 고려될 수 있다.

2.2.4 절대편차중앙값(median absolute deviation; MAD)에 기초한 추정

$$\widehat{\sigma_M} = c \cdot median\{|x_i - \hat{\mu}|\}$$

여기서 $\hat{\mu} = \text{median}\{x_i\}$이고, c는 정규분포에서 일치추정량이 되게 하는 조건상수로서 $c=1.4826$이 많이 쓰인다. 이 추정량은 고위붕괴점(high breakdown point; 50%의 붕괴점)을 갖는 로버스트한 추정량으로 알려져있다(송문섭(1996)). 그러나 MAD에 기초한 추정량은 고위붕괴점의 장점에도 불구하고 편의가 크고, 과소추정(under estimate)되는 경향이 있어서 극단적인 이상점이 존재하는 모형 또는 두터운 꼬리를 갖는 분포(heavy tailed distribution)를 제외한 대부분의 분포에서는 효율이 떨어지는 것으로 알려져 있다.

3. 몬테칼로 시뮬레이션

본 절에서는 앞에서 소개된 추정량들의 편의 및 효율성을 비교하기 위하여 다양한 분포군에서 모의실험을 시행해 보고자 한다. 우선 이를 위하여 다음과 같은 모형을 고려해 보자.

$$x_i = 10.0 + \varepsilon_i \; ; \quad i=1,2,\cdots,5$$

여기서 ε_i는 오차항으로서 $E(\varepsilon_i)=0$, $Var(\varepsilon_i)=\sigma^2$이다.

모의실험에서 고려한 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. 한편, 모의실험의 모든 계산은 S-PLUS(Statistical Sciences[4])를 이용하여 이루어 졌다.

[표 3.1]은 모의실험(5000번 반복)에서 얻어진 산포(σ)의 추정량들에 대한 평균(MEAN)과 평균제곱오차(mean squared error; MSE)를 계산한 결과이다. [표 3.1]에서 파라메타의 목표값(σ)은 분포에 따라 다른데 산포를 표에 계산해 두었다. 여기서 평균제곱오차는

$$\text{MSE} = \frac{1}{5000} \sum (\text{산포의 추정치} - \text{산포}(\sigma))^2$$

으로 계산된다.

한편, [표 3.1]에서 MSE의 측면을 고려할 때, 비교의 용이성을 위하여 효율(efficiency)을 구하여 괄호()속에 표시하였는데, 각 분포 및 표본의 크기에서 효율은 다음과 같이 구하여 진다.

$$\text{효율} = \frac{\text{최소 MSE}}{\text{MSE}} \times 100\,(\%)$$

[표 3.1]에서 편의의 측면에서 보면 범위를 이용한 추정법이 상대적으로 작은 편의를 보이고 있으나 그 차이가 큰 것은 아니다. MAD를 이용한 산포의 추정은 소표본(small sample)에서 편의 및 MSE의 측면에서 좋지 못한 성능을 가지고 있다.

한편, MSE의 측면에서는 $N(0, 1)$경우를 제외한 모든 고려된 분포에서 Gini 추정법에 의한 산포추정량이 가장 작은 값을 가지고 있어서 산포를 안정적으로 추정하고 있음을 보여주고 있다.

[표 3.1] 산포추정량의 성능비교

분 포			$N(0, 1)$	$CN(0.1, 3)$	$CN(0.2, 3)$	$CN(0.1, 5)$	$CN(0.2, 5)$
산포 (σ)			**1**	**1.3416**	**1.6125**	**1.8439**	**2.4083**
MEAN	$n=5$	$\widehat{\sigma_S}$	1.0028	1.2610	1.5088	1.5650	2.0927
		$\widehat{\sigma_R}$	1.0026	1.2669	1.5232	1.5701	2.1091
		$\widehat{\sigma_G}$	1.0029	1.2422	1.4765	1.5018	1.9944
		$\widehat{\sigma_M}$	0.8251	0.9293	1.0566	0.9727	1.2080
	$n=10$	$\widehat{\sigma_S}$	1.0028	1.2832	1.5408	1.6590	2.2314
		$\widehat{\sigma_R}$	1.0021	1.3386	1.6302	1.7785	2.4368
		$\widehat{\sigma_G}$	1.0025	1.2386	1.4764	1.5132	2.0160
		$\widehat{\sigma_M}$	0.9141	0.9959	1.1270	1.0367	1.2127
	$n=25$	$\widehat{\sigma_S}$	1.0027	1.3160	1.5660	1.7403	2.3006
		$\widehat{\sigma_R}$	1.0059	1.5092	1.8484	2.1788	2.9498
		$\widehat{\sigma_G}$	1.0015	1.2436	1.4642	1.5104	1.9800
		$\widehat{\sigma_M}$	0.9650	1.0565	1.1515	1.0819	1.2074
MSE	$n=5$	$\widehat{\sigma_S}$	0.1321(100)	0.4481(89.3)	0.6825(89.5)	1.4130(82.1)	2.2463(86.8)
		$\widehat{\sigma_R}$	0.1397(94.6)	0.4742(84.4)	0.7208(84.7)	1.4270(81.3)	2.3208(84.0)
		$\widehat{\sigma_G}$	0.1347(98.1)	0.4000(100)	0.6107(100)	1.1598(100)	1.9505(100)
		$\widehat{\sigma_M}$	0.2604(50.7)	0.5103(78.4)	0.7911(77.2)	1.1696(99.1)	2.3349(83.5)
	$n=10$	$\widehat{\sigma_S}$	0.0568(100)	0.2482(78.4)	0.3663(83.3)	0.9291(70.4)	1.3053(79.6)
		$\widehat{\sigma_R}$	0.0657(86.5)	0.3431(56.7)	0.5059(60.3)	1.2345(53.0)	1.7608(59.0)
		$\widehat{\sigma_G}$	0.0579(98.1)	0.1946(100)	0.3051(100)	0.6545(100)	1.0395(100)
		$\widehat{\sigma_M}$	0.1238(45.9)	0.2635(73.9)	0.4349(70.2)	0.8182(80.0)	1.7138(60.7)
	$n=25$	$\widehat{\sigma_S}$	0.0209(100)	0.1156(70.2)	0.1543(83.0)	0.4514(71.7)	0.5707(91.7)
		$\widehat{\sigma_R}$	0.0324(64.5)	0.3309(24.5)	0.4244(30.2)	1.2240(26.4)	1.5731(33.3)
		$\widehat{\sigma_G}$	0.0212(98.6)	0.0812(100)	0.1281(100)	0.3235(100)	0.5234(100)
		$\widehat{\sigma_M}$	0.0553(37.8)	0.1496(54.3)	0.2953(43.4)	0.6525(49.6)	1.5371(34.1)

※ 밑줄: 최소의 편의 및 MSE를 갖는 결과

또한, Gini의 평균차이에 기초한 추정법은 정규모집단에서의 표본일 경우 최적의 추정량으로 알려진 표본표준편차 또는 표본표준편차에 기초한 추정량에 대하여 98%이상의 높은 효율을 유지하고 있어서 로버스트 추정량으로서의 상당한 장점을 갖고 있다고 볼 수 있다.

한편, 계량형 관리도 또는 측정시스템 분석에서 흔히 사용되는 범위에 기초한 산포추정량의

성능은 상당한 문제를 안고 있다. 이 추정량은 표본크기가 10이하의 비교적 작은 경우에 편의가 작은 장점이 있지만, 표본의 크기가 커지면 다소 과대추정(over estimate)되어 편의가 커지는 경향이 있으며, 효율의 측면에서도 표본표준편차보다 나은 면이 없는 것으로 나타났다. 단지 범위를 이용한 추정법이 계산의 편의성만을 가지고 그 쓰임새를 논한다면 의미가 적어질 수밖에 없을 것으로 생각된다.

반면 시뮬레이션 결과에 의하면 Gini의 평균차이에 기초한 추정량은 편의의 측면에서도 큰 차이를 보이지 않고 안정적이며, 정규분포를 제외한 모든 분포에서 비교된 추정량들 중 가장 뛰어난 효율을 갖고 있는 것으로 나타났다.

4. 결 론

본 논문에서는 일반적으로 관리도법 또는 공정능력분석 및 측정시스템 분석에 필요한 산포의 로버스트 추정법으로서 Gini의 평균차이에 기초한 추정법을 논하고, 이의 유용성을 모의실험을 통하여 입증하였다. 관리도 및 공정능력분석 등의 기법에서 산포추정에 관련된 중요한 문제는 데이터의 개수(표본의 크기)가 아주 작다는 것이다. 이러한 경우에는 일반적인 산포의 추정량(예를 들면 표본표준편차)으로는 편의(bias)와 정도(precision)의 측면에서 효율이 떨어질 수 있으며, 흔히 논의되는 추정량 계산의 간편성은 요즘과 같이 계산지원 소프트웨어가 발달한 상황에서는 설득력이 떨어질 수밖에 없을 것으로 생각된다.

Gini의 평균차이를 이용한 산포의 추정은 이러한 측면에서, 특히 소표본에서 데이터가 갖는 산포에 관한 정보를 최대한 활용한다고 볼 수 있다. 기존의 표준편차는 n개의 편차(차이)를 이용하여 산포를 추정하고, 범위를 이용한 추정은 1개의 차이(범위)만 이용한다. 그러나 Gini의 평균차이에서는 $n(n-1)/2$개의 차이들을 이용하게 되므로 작은 표본에서 특히 데이터가 갖는 정보의 재사용(reuse)의 측면에서 유용한 산포의 측도로 고려될 수 있다. 시뮬레이션 결과는 Gini 추정법의 이러한 효율성을 입증해 주고 있다.

참고문헌

[1] Gini, C.(1912). variabilita e mutabilita, contributo allo studio delle distribuzioni e delle relazioni statistiche. *Studio Economico-Giuridici della R. Universita di Cagliari.* 3, part 2, 3-159.

[2] Montgomery, D. C.(1991). *Introduction to Statistical Quality Control*, Wieley.

[3] Rodriguez, R. N.,(1992), "Recent Development in Process Capability Analysis," Journal of Quality Technology, Vol. 24, No. 4, pp. 176-187.

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

[5] 백재욱, 조진남(1999). "공정능력지수에 대한 비평과 올바른 공정능력분석절차", 품질경영학회지, 27권 2호(1999년 6월)

[6] 송 문섭(1996). *로버스트 통계*, 자유아카데미.