▨연구논문

분계점 붓스트랩 방법을 이용한 자기상관을 갖는 공정의 $\overline{X}$ 관리도

김윤배
성균관대학교 시스템경영공학부
박대수
한국통신 경영연구소

$\overline{X}$ control charts of autocorrelated process using threshold bootstrap method

Yun Bae Kim
School of Systems Management Engineering, Sungkyunkwan University
Daesu Park
Management Research Lab. Korea Telecom

Abstract

$\overline{X}$ control chart has proven to be an effective tool to improve the product quality. Shewhart charts assume that the observations are independent and normally distributed. Under the presence of positive autocorrelation and severe skewness, the control limits are not accurate because assumptions are violated. Autocorrelation in process measurements results in frequent false alarms when standard control charts are applied in process monitoring. In this paper, Threshold Bootstrap and Moving Block Bootstrap are used for constructing a confidence interval of correlated observations. Monte Carlo simulation studies are conducted to compare the performance of the bootstrap methods and that of standard method for constructing control charts under several conditions.

1. 서론

Shewhart[8]에 의해 처음 소개된 관리도는 유의한 공정평균의 변동을 검출하는데 사용되는 온라인 공정관리기법으로서 요구되는 가정은 관측치 X_i가 따르는 분포 F의 정규성과 관측치 들 사이의 독립성이다. 전자는 F가 어느 정도 대칭함수이거나 부분군의 크기 n이 충분히 크다면 완화될 수 있지만 후자는 완화시키기 어려운 문제이다.

관측치들이 독립성 가정을 벗어나게 되면 관측치들 사이에 자기상관이 존재한다. 일반적인 예로써, 주식시장의 주식가격이나 자동으로 조정되는 생산 설비로부터 얻어진 측정치들이 있다. 관측치에 자기상관이 존재할 때, X_i들 사이의 자기상관을 포함하게 되는 $\overline{X}$의 분산은 표준방법의 관리도에 의해서는 잘 추정되지 못하므로 표준방법에 의한 $\overline{X}$ 관리도는 유효하지 않게 된다. 본 연구에서는 관측치들의 분포 F에 대한 모수적인 가정없이 양의 자기상관이 존재하는 관측치들을 위한 관리도를 개발하기 위하여 붓스트랩 방법을 적용하기로 한다. 붓스트랩 방법은 분포에 대한 가정을 요구하지 않는 재표집 과정이며, 분포가 알려져 있지 않은 통계량의 분포를 추정하기 위하여 개발되었다[1]. 본 연구에서는 관측치들이 정상(stationary) 상태의 자기상관을 가지는 경우에 대하여 개발된 수정된 붓스트랩 방법인 이동블록 붓스트랩(Moving Block Bootstrap)[5][6]과 분계점 붓스트랩(Threshold Bootstrap)[4] 구조를 이용하여 관측치들간의 자기상관을 보존하면서 관리도의 관리한계선을 구축하기 위한 통계량들을 재표집한다. 재표집된 통계량에 대한 분포를 추정하기 위하여 관심있는 통계량에 대한 분포가 알려져 있지 않은 경우에 신뢰구간을 구하는 방법인 붓스트랩 백분위수 방법[2]을 적용하여 관리한계선을 구한다.

관리도를 구축하기 위하여 붓스트랩 방법을 사용하는 것에는 여러 가지 이점이 있다. 우선, 관측치들의 분포와 자기상관의 구조에 대한 사전지식이 필요하지 않게 되며, 관리도를 구축하기 위하여 많은 관측치들의 수를 요구하지 않는다. 붓스트랩 방법을 사용하기 위하여 요구되는 것은 상대적으로 많은 계산량을 처리할 수 있는 계산능력의 확보이나, 근래에는 컴퓨터 기술의 발달로 충분히 계산을 수행하는 것이 가능하다.

본 논문은 전체 6절으로 구성되어 있으며, 제 2 절에서는 표준 $\overline{X}$ 관리도에 대해 살펴본 후에 관측치들간의 자기상관이 존재할 경우의 표준 $\overline{X}$ 관리도에 나타난 문제점을 살펴본다. 제 3 절에서는 관리도의 작성에 이용되는 붓스트랩 방법들에 대하여 살펴본다. 상호 독립적이고 동일한 분포를 따르는 관측치들에 대한 고전적 붓스트랩 방법과 자기상관이 존재하는 관측치들을 위하여 개량된 분계점 붓스트랩 방법의 구조와 통계적인 특성을 살펴본다. 제 4 절에서는 관리도를 구축하기 위한 부분군의 크기를 결정하는 방법과, 관리도 작성에 필요한 통계량의 분포에 대한 붓스트랩 방법에 의한 유사 통계량의 분포의 근사정도에 대하여 살펴보고 관리도에 적용시키는 방법을

제시하며, 제 5 절에서는 여러 가지 형태의 관측치에 대한 시뮬레이션 결과를 표준 $\overline{X}$ 관리도를 사용하였을 때의 결과와 비교한다. 마지막으로 제 6 절에서는 실험의 결과에 대하여 기술한다.

2. 표준관리도 구조

Shewhart에 의해 개발된 관리도는 통계적 공정관리에서 유용하게 사용되는 도구이다. 예를 들어, $\overline{X}$ 관리도는 공정평균 μ를 모니터하기 위해 사용되는 관리도이다. 이 관리도는 N개의 관측치 $\{X_1, X_2, \cdots, X_N\}$ 중에서 크기 n인 부분군의 표본평균 $\overline{X}_n$들을 타점하며, 실험 통계량으로서 $\overline{X}_n$를 사용하여 어떠한 목적값 μ_0에 대하여 가설 $H_0: \mu = \mu_0$에 대한 $H_1: \mu \neq \mu_0$를 검증하게 된다.

여기서, $\{X_1, X_2, \cdots, X_N\}$는 평균 μ, 표준편차 σ인 미지의 분포 F를 따르는 특정 품질특성치의 표본측정치라고 가정한다. 공정의 관리상태를 판정하기 위하여 관리한계선을 사용하며, 각 시점에 대해 타점되는 $\overline{X}_n$값들이 관리한계선 내에 존재하면 공정이 관리상태에 있다고 판정된다. H_0가 사실일 때 $\overline{X}_n - \mu$의 표본분포를 찾아야 관리한계선을 확정할 수 있다. 더욱 정확하게 말하면, 주어진 위험수준 α를 가지고 H_0 하에서

$$\Pr[L < \overline{X}_n - \mu < U] = 1 - \alpha$$

가 되는 U와 L을 위치시킬 필요가 있게된다. 여기에서 확률은 관측치의 분포 F에 관하여 계산된다. 그 때의 관리하한선과 관리상한선은 $(\mu_0 + L)$과 $(\mu_0 + U)$로 정해지며, α의 수준은 공정이 관리상태에 있음에도 불구하고 타점되는 $\overline{X}_n$의 값이 관리한계선을 벗어나는 제 1종오류확률(false alarm rate)로서 해석되어진다.

만약, F가 정규분포이고 부분군 크기 n이 크다면, 표준 $\overline{X}$ 관리도는

$$\text{관리하한선} = \mu_0 - \frac{3\sigma}{\sqrt{n}}$$

$$\text{관리상한선} = \mu_0 + \frac{3\sigma}{\sqrt{n}}$$

을 가지는 n개의 연속적인 X_i들로 구성된 각 부분군의 표본평균을 타점한다. 이 경우에, α는 0.27%이며, 중심선은 μ_0가 된다.

실제로, μ_0와 σ는 알려져 있지 않기 때문에 공정이 관리상태에 있을 때 취해진 과거의 자료들을 사용하여 추정한다. 과거의 자료들이 k개의 부분군들로 구성되어져 있고, 각 부분군이 크기 n으로 구성되어 있다고 가정하자. i번째 부분군에 대한 표본평균과 표본분산을 각각 $\overline{X}_i$와 S_i^2이라고 정의하면, μ_0와 σ^2는

$$\overline{X}_N = \frac{\sum_{i=1}^{k} \overline{X}_i}{k} \quad , \quad S^2 = \frac{\sum_{i=1}^{k} S_i^2}{k}$$

을 사용하여 추정한다.

F의 정규성의 가정 하에서, $\frac{k(n-1)S^2}{\sigma^2}$ 은 $k(n-1)$의 자유도를 가지는 χ^2분포를 따르며, H_0하에서 $\frac{\sqrt{n}(\overline{X}_n - \mu_0)}{S}$는 $k(n-1)$의 자유도를 가지는 t분포를 따른다. 실제로 $k(n-1)$은 일반적으로 큰 수이기 때문에, t분포는 표준정규분포에 충분히 근접하게 되며 $\overline{X}$ 관리도에 대한 관리한계선은

$$\text{관리하한선} = \overline{X}_N + z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

$$\text{관리상한선} = \overline{X}_N + z_{(1-\frac{\alpha}{2})} \frac{S}{\sqrt{n}}$$

이 된다. 여기서, z_c는 표준정규분포의 c 백분위수를 의미하며, 부분군의 크기 n은 4에서 6사이의 값을 취하는 것이 일반적이다.

3. 고전적인 붓스트랩 방법과 분계점 붓스트랩

붓스트랩 방법은 주어진 관측치의 여러 가지 통계량들에 대한 표준오차, 신뢰구간 등을 추정하는 데 사용할 수 있는 비모수적인 통계적 추론방법이다. 이 방법의 기본 원리는 관측치 $X=\{X_1, X_2, \cdots, X_N\}$가 주어졌을 때, 재표집 과정을 통하여 어떠한 통계량 $\hat{\theta}=\hat{\theta}(X)$에 대한 여러 값들을 만들어낸 후에 이 값들을 이용하여 $\hat{\theta}$에 대한 표본분포를 만들어내는 것이다. 붓스트랩 방법은 적은 수의 관측치와 최소한의 가정을 가지고도 수행될 수 있는 통계적 추론과정이며, 이 방법을 사용하기 위한 부가적인 계산능력이 필요하지만 컴퓨터를 이용한 계산능력은 이를 가능하게 하였다.

미지의 확률분포 함수 F로부터 서로 독립인 데이터를 추출한다고 하면, 모수 $\theta=\theta(F)$에 대한 추정값은 일반적으로 아래와 같이 정의된다.

$$\hat{\theta}=\theta(\hat{F})$$

여기서 $\hat{F}$는 경험적 확률분포함수를 나타내고, $\hat{F}$는 독립성과 동일성이 보장되는 N개의 모든 관측치 X_i에 대해 추출될 확률을 $\frac{1}{N}$로 보장한다.

$\sigma(F)$를 미지의 표본분포 F와 함수관계에 있는 표준오차라고 하면, 붓스트랩으로 추정한 표준오차는 $\sigma(\cdot)$를 F대신 $\hat{F}$으로 대체했을 때의 값이다. 즉

$$se_{boot}(\hat{\theta})=\sigma(\hat{F}) \tag{1}$$

만약 표본평균이 추정하고자 하는 통계량이면

$$\sigma(F)=\left\{\frac{\mu_2(F)}{N}\right\}^{\frac{1}{2}} \tag{2}$$

이 된다. 여기서 $\mu_2(F)$는 $\int(x-\mu)^2dF$이다. 따라서 표본평균에 대한 붓스트랩 표준오차는 아래와 같이 계산할 수 있다.

$$se_{boot}(\overline{X}_N)=\sigma(\hat{F})=\left\{\frac{\mu_2(\hat{F})}{N}\right\}^{\frac{1}{2}} \tag{3}$$

여기서 $\mu_2(\hat{F})$는 $\dfrac{\sum_{i=1}^{N}(X_i-\overline{X}_N)^2}{N}$ 이고, $\hat{F}$의 2차 중심적률이다.

$\overline{X}_N$이외의 통계량들에 대한 $\sigma(F)$의 공식은 일반적으로 알려져 있지 않고 $se_{boot}(\hat{\theta})$은 Monte Carlo 시뮬레이션을 통하여 추정할 수 있으며 그 알고리듬은 아래와 같다.

1) 관측치 $\{X_1, X_2, \cdots, X_N\}$의 집단에서 크기가 N인 붓스트랩 표본을 복원추출한다.

$$\hat{F} \rightarrow (X_1^*, X_2^*, \cdots, X_N^*)$$

2) 붓스트랩 재표집으로 구한 표본을 이용한 통계량을 B회 계산한다.

$$\hat{\theta}_i^*=\hat{\theta}(X_1^*, \cdots, X_N^*), \quad i=1,\ldots,B$$

3) 통계량 $\hat{\theta}$의 표준오차의 붓스트랩 추정치를 아래와 같이 계산한다.

$$se^*(\hat{\theta}^*)=\left\{\frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i^*-\hat{\theta}_\cdot^*)^2\right\}^{\frac{1}{2}} \quad (4)$$

여기서, $\hat{\theta}_\cdot^*=\frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^*$ 이며, B가 충분히 크면 $se^*(\hat{\theta}^*)$는 $se_{boot}(\hat{\theta})$로 수렴한다. 대부분의 붓스트랩 실험에서 B의 크기는 50에서 200정도가 $se_{boot}(\hat{\theta})$에 대하여 적합하다고 Efron과 Tibshirani[3]는 권장하고 있다.

분계점 붓스트랩은 시뮬레이션의 단일실행의 결과를 해석하기 위하여 고전적인 붓스트랩의 재표집 단위인 단독 관측치를 관측치내에 존재하는 종속관계를 유지하도록 'chunk'로 바꾸었다. 재표집 단위를 cycle로 바꾼 상태에서 재표집된 붓스트랩 모의자료 군이 원시자료에 존재하는 종속관계를 유지하는 것을 보였다. 표본평균이나 표본의 중앙값 등과 같이 관측치를 관통하는 수준으로 분계점을 정하면 관측치는 분계점보다 높은 자료의 연속인 상위런과 낮은 자료의 연속인 하위런으로 구분된다. 재표집의 기본단위를 중복되지 않고 연속되는 두 개의 런(상위런과 하위런의 결합)으로 정하고, 이를 'cycle'이라고 하였다. Park[7]은 분계점 붓스트랩의 수행도를 높이기 위하여 하나이상의 cycle의 집합인 chunk를 재표집 단위로 사용하였다. 본 연구에서는 chunk가 하나의 cycle로 이루어진다고 보았으며, 이 때에 분계점을 정하면 관측치 자체가 chunk를 결정하게 되므로 재표집의 단위가 자동으로 정해지는 것이 분계점 붓스트랩의 장점일 수 있다.

확률구조 P에 의해 생성된 정상상태의 자기상관이 존재하는 관측치 $\{X_1, X_2, \cdots, X_N\}$이 있고, 임의의 분계점이 R개의 chunk를 만든다고 가정하자. $C_i=\{X_{i,1}, X_{i,2}, \cdots, X_{i,n_i}\}$가 크기를 n_i로 하는 i번째 chunk라면, $i=1,2,\cdots,R$이고 $\sum_{i=1}^{R}n_i=N$이 된다. chunk의 크기 n_i와 chunk의 수 R은 둘 다 확률변수이다. 이 두 개의 확률변수들은 관측자료군의 자기상관과 분계점에 매우 강한 종속관계를 갖는다. 분계점 붓스트랩은 집합 $\{C_1, C_2, \cdots, C_R\}$에서 랜덤하게 chunk를 추출하여 붓스트랩 표본을 만든다. 미지의 P는 재표집 구조에 근거하여 P^*로 추정할 수 있다. 어떠한 확률과정의 평균을 μ라고 하고, 관측치 $\{X_1, X_2, \cdots, X_N\}$의 표본평균 $\overline{X}_N$의 표준오차를 추정하려 한다고 가정하자. $\{X_1^*, X_2^*, \cdots, X_N^*\}$가 P^*에서 얻은 붓스트랩 샘플이라면, $\overline{X}_N$ 표준오차에 대한 붓스트랩 추정치는 재표집 확률구조 P^*에 의거한 $\overline{X}_N^*$의 표준오차와 일치하게 된다.

Kim 등[4]이 제안한 분계점 붓스트랩(threshold bootstrap: TB) 알고리듬은 다음과 같다.

step 0 : N개로 이루어진 자기상관이 존재하는 자료군을 획득한다.
step 1 : 표본평균이나 표본의 중앙값 등을 사용하여 분계점을 설정한다
step 2 : 자료군은 분계점보다 높은 자료의 연속인 "high-run"과 낮은 자료의 연속인 "low-run"으로 자동으로 구분되어진다.
step 3 : 중복되지 않고 연속되는 두 개의 런(high런과 low런의 결합)을 chunk라 정의한다. 복원추출된 chunk들을 연결하여 붓스트랩 샘플을 원시자료의 크기 (N) 만큼 생성한다.

* R개의 chunk가 추출될 확률은 $\frac{1}{R}$이며, 균등분포(0,1)를 따르는 난수를 생성시켜 재추출할 chunk를 선택한다. 만약 붓스트랩 샘플이 원시자료군의 크기인 N을 넘어가면, 마지막으로 선택된 chunk내의 관측치를 N까지만 취한다.

step 4 : 재표집된 붓스트랩 샘플로 관심 있는 통계량을 계산한다.
step 5 : step 3과 step 4를 총 B번 반복한다.
step 6 : B번의 반복으로 계산된 통계량의 표본분포를 구한 후 관심 있는 통계량의 표준편차나 신뢰구간 등을 추정한다.

4. $\overline{X}$ 관리도에 기초한 붓스트랩 방법의 개발

만약 정상상태의 자료열 $\{X_1, X_2, \cdots, X_k\}$과 $\{X_{k+n}, X_{k+n+1}, \cdots, X_{k+n+l}\}$이 $n > m$일 때 상호독립이면 자료열은 정상상태의 m-dependent하다. 부분군의 크기가 n인 $\{X_1, X_2, \cdots, X_n\}$이 정상상태의 m-dependent 표본 관측치이고 $m \le n$일 때, $\overline{X}_n$의 분산은 $Var[\overline{X}_n] = \frac{Var[X]}{n}\left\{1+2\sum_{i=1}^{m}(1-\frac{i}{n})\rho_i\right\}$가 된다. $Var[\overline{X}_n]$의 정교한 추정치는 $\overline{X}$ 관리도의 정확한 관리한계선을 제공하기 때문에, 부분군의 크기를 결정하는 것은 관리한계선을 구하는데 중요한 역할을 하게 된다. 부분군의 크기를 m이하로 결정하게 되면, 관측치들간에 존재하는 자기상관을 잘 표현하지 못하게 되어 우수한 $Var[\overline{X}_n]$의 추정치를 얻을 수 없게 된다. 본 연구에서는 분계점 붓스트랩 방법에서의 재표집의 단위인 chunk의 크기를 이용하여 부분군의 크기를 정하도록 한다.

분계점 붓스트랩의 평균 chunk크기를 이용하여 부분군의 크기를 정하는 것이 적절함을 보이기 위하여 N=500,000인 표준정규분포와 지수분포, 이웃하는 자료간의 자기상관(ρ_1) 정도가 0.5인 감마분포(2,1), 여러 가지 ϕ값을 가지는 AR(1) 자료들을 이용하여 평균 chunk크기에 대한 실험을 한 결과가 <표 1>에 나타나 있다.

<표 1>에서 ρ_1는 임의의 chunk의 첫 번째 자료와 연속한 이웃 chunk의 첫 번째 자료와의 자기상관을 나타낸다.

< 표 1 > 여러 가지 형태의 관측치에 대한 평균 chunk크기

		평균Chunk크기*	ρ_n
iid	N(0,1)	4	0
	EXP(1)	5	0
ρ_1=0.5	Gamma(2,1)	7	0
AR(1)	ϕ=0.1	5	0.000001
	ϕ=0.2	5	0.000064
	ϕ=0.3	5	0.000729
	ϕ=0.4	6	0.001638
	ϕ=0.5	6	0.007812
	ϕ=0.6	7	0.016796
	ϕ=0.7	8	0.040353
	ϕ=0.8	10	0.085899
	ϕ=0.9	14	0.205891

* Chunk의 평균크기는 가장 근접한 정수 값을 취함.

표준 $\overline{X}$ 관리도를 구축할 때에는 관리한계선을 설정하기 위한 통계량으로서 $\frac{\sqrt{n}(\overline{X}-\mu)}{S}$를 사용한다. 관측치 X_i의 분포 F가 서로 독립이고 동일한 정규분포를 따른다고 가정하면 이 통계량은 표준정규분포를 따르게 되며 이를 이용하여 관리한계선을 구축한다. 하지만, 이러한 가정들이 의심스러울 경우에는 $\frac{\sqrt{n}(\overline{X}_n-\mu)}{S}$을 사용할 수가 없다. 본 연구에서는 그러한 경우에 분계점 붓스트랩과 이동블록 붓스트랩을 이용하여 관리한계선을 구축하는 방법을 제시하고 있으며 통계량으로서 $\sqrt{n}(\overline{X}_n-\mu)$를 사용한다. 관측치가 주어져 있을 때의 $\sqrt{N}(\overline{X}_N^*-\overline{X}_N)$의 표본분포는 $\sqrt{N}(\overline{X}_N-\mu)$의 표본분포의 붓스트랩 근사이다. 따라서, $\sqrt{N}(\overline{X}_N-\mu)$의 표본분포는 붓스트랩 표본분포에 의해서 근사되어질 수 있다. 즉,

$$\Pr[\sqrt{N}(\overline{X}_N^*-\overline{X}_N)\le x \mid X_1, X_2, \cdots, X_N] \approx \Pr[\sqrt{N}(\overline{X}_N-\mu)\le x] \qquad (6)$$

이다.

$\overline{X}$ 관리도를 구축하는데 있어서 N은 공정으로부터 이용 가능한 관측치들의 총 수가 된다. $\overline{X}$ 관리도는 크기가 n인 부분군의 표본평균을 타점하기 때문에 관리한계선은 $\sqrt{n}(\overline{X}_n-\mu)$의 표본분포로부터 얻어져야만 한다. 하지만, $\sqrt{n}(\overline{X}_n-\mu)$의 분포는 관측치간의 자기상관과 관측치의 분포 F에 의해서 많은 영향을 받게 된다. 관측치 N이 크다면, $\sqrt{n}(\overline{X}_n-\mu)$의 표본분포를 이용할 수 있지만, 관측치 N이 작을 경우에는 분포를 추정하기가 어렵게 된다. 이러한 경우에 $\sqrt{n}(\overline{X}_n-\mu)$의 분포는

$$\Pr[\sqrt{n}(\overline{X}_n^*-\overline{X}_N)\le x \mid X_1, X_2, \cdots, X_N]\approx \Pr[\sqrt{n}(\overline{X}_n-\mu)\le x] \qquad (7)$$

를 따르는 붓스트랩에 의해 근사되어질 수 있어야 한다. n과 N이 ∞로 접근함에 따라, $\sqrt{n}(\overline{X}_n-\mu)$과 $\sqrt{N}(\overline{X}_N-\mu)$는 $N(0, Var[X]\left\{1+2\sum_{i=1}^{m}\rho_i\right\})$으로 수렴하게 되며, $\sqrt{n}(\overline{X}_n^*-\overline{X}_N)$과 $\sqrt{N}(\overline{X}_N^*-\overline{X}_N)$ 대해서도 마찬가지로 수렴 해야된다. 그러나, 부분군의 크기 n이 실제로 크지 않기 때문에 $\sqrt{n}(\overline{X}_n^*-\overline{X}_N)$과 $\sqrt{n}(\overline{X}_n-\mu)$에 대한 정규분포의 가정을 할 수 없다.

본 연구에서는 서로 독립이거나 자기 자기상관이 존재하는 N=100,000인 여러 가지 형태의 자료를 대상으로 $\sqrt{n}(\overline{X}_n^*-\overline{X}_N)$의 표본분포가 $\sqrt{n}(\overline{X}_n-\mu)$의 표본분포를 얼마나 근접하게 근사하는가를 보이기 위하여 실험을 수행하였다. 붓스트랩 표본평균 $\overline{X}_n^*$을 발생시키는 방법으로 이동블록 붓스트랩과 분계점 붓스트랩 방법을 사용하였다. 두 종류 분포의 근사성을 보기 위해 히스토그램과 분산을 서로 비교하였다.

분계점 붓스트랩을 이용한 방법 :

1) 분계점 붓스트랩의 평균 chunk크기를 이용하여 부분군의 크기 n을 결정한다. 이 때, 관측치들은 $\frac{N}{n}$개의 부분군으로 나누어지며, $\frac{N}{n}$개의 부분군에 대한 $\sqrt{n}(\overline{X}_n-\mu)$값들을 히스토그램으로 작성한다.

2) 분계점 붓스트랩을 적용하여 N개의 붓스트랩 모의자료군을 형성한 후 $\frac{N}{n}$개의 $\overline{X}_n^*$을 계산한다.

3) $\frac{N}{n}$개의 $\overline{X}_n^*$을 이용하여 얻은 $\sqrt{n}(\overline{X}_n^*-\overline{X}_N)$값들을 히스토그램으로 작성한다.

이동블록 붓스트랩을 이용한 방법 :

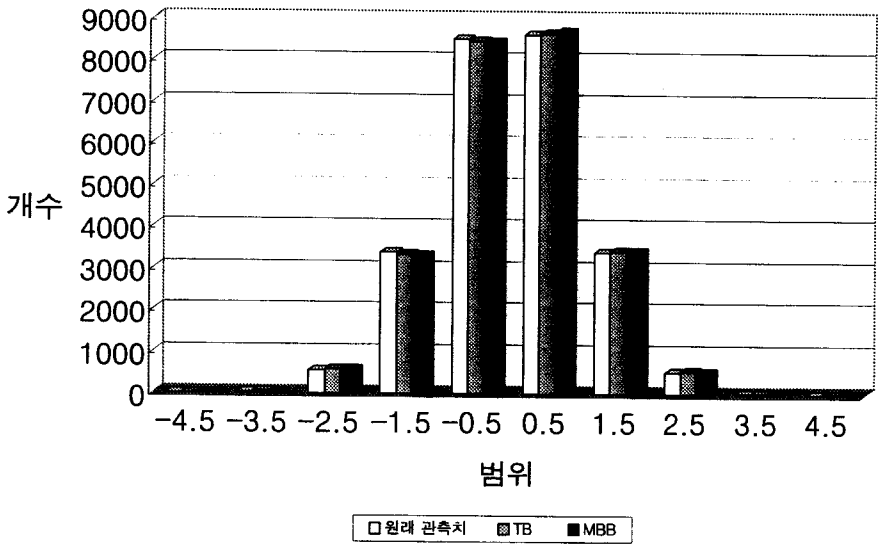
1) 분계점 붓스트랩의 평균 chunk크기를 이용하여 부분군의 크기 n을 결정한다. 이 때, 관측치들은 $\frac{N}{n}$개의 부분군으로 나누어지며, $\frac{N}{n}$개의 부분군에 대한 $\sqrt{n}(\overline{X}_n-\mu)$값들을 히스토그램으로 작성한다.

2) 부분군의 크기 n을 재표집을 위한 블록크기로 정한다. 이 때, 재표집의 기본단위인 B_i는 $B_i=\{X_i, X_{i+1}, \cdots, X_{i+n-1}\}$, $i=1,2,\cdots,N-n+1$가 된다.

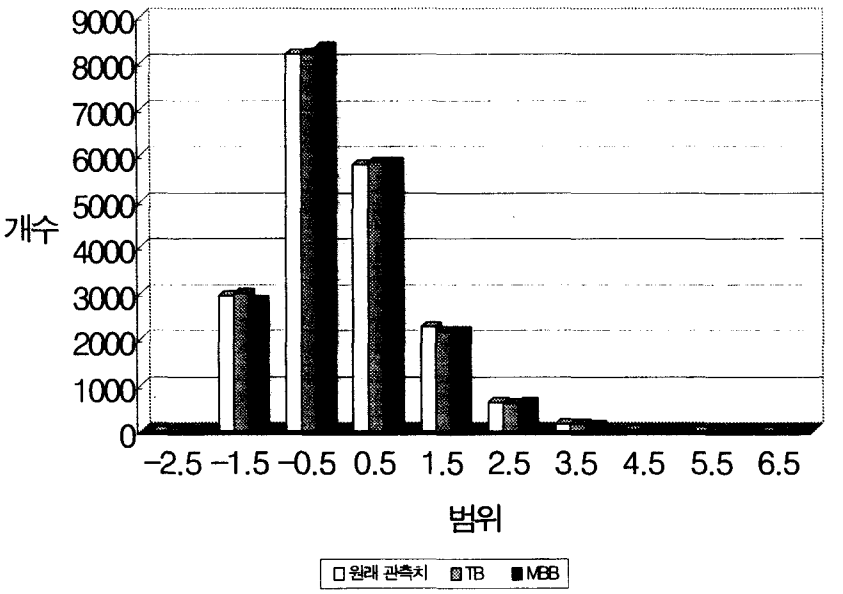
3) B_i에 대한 $\frac{N}{n}$번의 재표집을 통하여 얻은 $\sqrt{n}(\overline{X}_n^*-\overline{X}_N)$값들을 히스토그램으로 작성한다.

이러한 실험 결과가 다음의 <그림 1>, <그림 2>, <그림 3>, <그림 4>에 나타나 있다. 그림에서 TB는 분계점 붓스트랩 방법을 사용한 결과이며, MBB는 이동블록 붓스트랩을 사용한 결과이다.

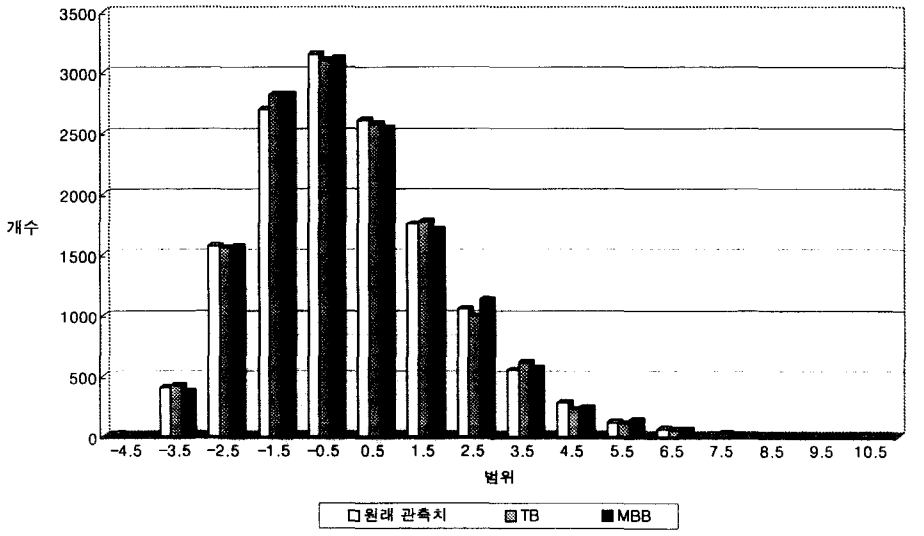
<그림 1>부터 <그림 4>까지를 살펴보면, 관측치 X_i가 정규분포를 따르지 않거나 관측치간에 자기상관이 존재하는 경우에도, 재표집에 의한 $\sqrt{n}(\overline{X}_n^*-\overline{X}_N)$의 표본분포가 $\sqrt{n}(\overline{X}_n-\mu)$의 표본분포를 잘 묘사하는 것을 알 수 있다.



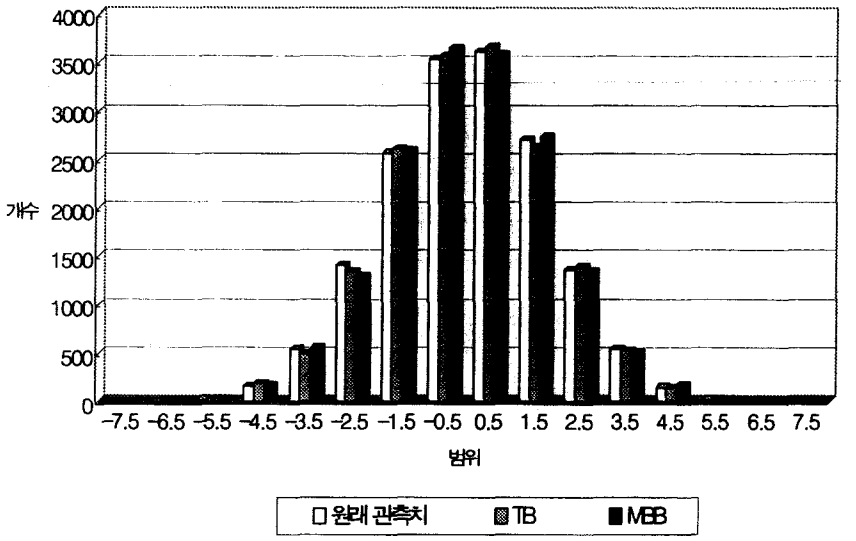
< 그림 1 > 표준정규분포에 대한 $\sqrt{n}(\overline{X}_n-\mu)$와 $\sqrt{n}(\overline{X}_n^*-\overline{X}_N)$의 비교



< 그림 2 > 지수분포(1)에 대한 $\sqrt{n}(\overline{X}_n - \mu)$와 $\sqrt{n}(\overline{X}_n^* - \overline{X}_N)$의 비교



< 그림 3 > ρ_1=0.5 감마분포(2,1)의 $\sqrt{n}(\overline{X}_n - \mu)$와 $\sqrt{n}(\overline{X}_n^* - \overline{X}_N)$의 비교



< 그림 4 > $\phi=0.5$인 AR(1)에 대한 $\sqrt{n}(\overline{X}_n-\mu)$와 $\sqrt{n}(\overline{X}_n^*-\overline{X}_N)$의 비교

또한, $Var\,[\sqrt{n}(\overline{X}_n-\mu)]$와 $Var\,[\sqrt{n}(\overline{X}_n^*-\overline{X}_N)]$를 비교한 결과가 <표 2>에 나타나 있다. AR(1) 모형의 경우 $Var\,[X_i]$와 $Var\,[\sqrt{n}(\overline{X}_n-\mu)]$는 다음과 같은 식을 사용하여 계산하였다.

$$Var\,[X_i]=\frac{Var\,[\varepsilon_i]}{1-\phi^2},\qquad \varepsilon_i \sim N(0,1) \tag{8}$$

$$Var\,[\sqrt{n}(\overline{X}_n-\mu)] = Var[X]\left\{1+2\sum_{i=1}^{n}(1-\frac{i}{n})\rho_i\right\} \tag{9}$$

< 표 2 > $Var[\sqrt{n}(\overline{X}_n-\mu)]$와 $Var[\sqrt{n}(\overline{X}_n^*-\overline{X}_N)]$의 비교

		n	$Var[X_i]$	$Var[\sqrt{n}(\overline{X}_n-\mu)]$	MBB를 이용한 $Var[\sqrt{n}(\overline{X}_n^*-\overline{X}_N)]$	TB를 이용한 $Var[\sqrt{n}(\overline{X}_n^*-\overline{X}_N)]$
iid	$N(0,1)$	4	1	1	0.993	1.002
	$EXP(1)$	5	1	1	1.003	1.002
AR(1)	$\phi=0.5$	6	1.333	3.125	3.067	3.113

붓스트랩 추정치 $Var[\sqrt{n}(\bar{X}_n^* - \bar{X}_N)]$는 (N/n)개의 붓스트랩 평균의 표본분산이다. <표 2>에 의하면 TB와 MBB 분산 추정치 모두 참값에 근접하고 있는 것을 볼 수 있다.

이제 붓스트랩 방법을 이용하여 자기상관이 있는 관측치의 $\bar{X}$ 관리도를 구축하는 방법을 설명하겠다. 관측치 $\{X_1, \cdots, X_N\}$이 주어져 있고 false alarm 비율이 α일 때, $\bar{X}$ 관리도 작성을 위한 통계량으로서 표준 $\bar{X}$ 관리도에서 사용되는 $\frac{\sqrt{n}(\bar{X}_n - \mu)}{S}$ 대신에 여기서는 $\sqrt{n}(\bar{X}_n - \mu)$을 사용한다. 우선, 분계점 붓스트랩의 재표집 단위인 chunk의 평균크기를 이용하여 부분군의 크기 n을 결정한 후에, $\sqrt{n}(\bar{X}_n - \mu)$의 표본분포를 추정하기 위하여 K개의 $\sqrt{n}(\bar{X}_n^* - \bar{X}_N)$ 값들을 발생시킨다. 여기서 K는 임의의 숫자이며 붓스트랩 재표집 횟수와 관련 되어있다. 즉 관측치의 크기가 N이고 부분군의 크기가 n이며 재표집의 횟수가 B라 하면 K는 N/n에 가장 가까운 정수 값에 B를 곱한 값이 된다. K개의 $\sqrt{n}(\bar{X}_n^* - \bar{X}_N)$ 값들에 대하여 $\frac{\alpha}{2}$ 백분위수, $(1-\frac{\alpha}{2})$백분위수에 해당하는 값을 취할 수 있으며 이 값들을 $\hat{\tau}_{\frac{\alpha}{2}}$, $\hat{\tau}_{1-\frac{\alpha}{2}}$라고 하면, 이 두 값이

$$\Pr[\tau_{\frac{\alpha}{2}} \le \sqrt{n}(\bar{X}_n - \mu) \le \tau_{1-\frac{\alpha}{2}}] = 1-\alpha \tag{10}$$

를 만족시키는 $\sqrt{n}(\bar{X}_n - \mu)$의 분포의 $\tau_{\frac{\alpha}{2}}$, $\tau_{1-\frac{\alpha}{2}}$에 해당하는 값의 추정치로서 사용된다. 이렇게 될 때의 관리한계선은 결국 다음과 같이 된다.

$$\text{관리하한선} = \bar{X}_N - \frac{\hat{\tau}_{\frac{\alpha}{2}}}{\sqrt{n}} \tag{11}$$

$$\text{관리상한선} = \bar{X}_N + \frac{\hat{\tau}_{1-\frac{\alpha}{2}}}{\sqrt{n}} \tag{12}$$

지금까지 살펴본 결과에 따르면, X_i의 분포인 F에 대한 가정이 없음을 알 수 있다. 그러므로, 자기상관이 존재하는 관측치뿐만 아니라 관측치 X_i의 분포가 정규분포가 아니라는 가정 하에서도 붓스트랩을 이용한 관리도가 잘 적용될 것이다.

5. 모의실험

본 실험에서는 양의 자기상관이 존재하는 감마분포와 AR(1)모형에 대해 총 100번의 독립반복실험을 하여 표준 $\overline{X}$ 관리도와 붓스트랩을 이용한 관리도의 성능을 비교했다. 관측치의 수는 N=200이며 이동블록 붓스트랩과 분계점 붓스트랩을 각각 사용하여 각각 K=1,000개의 $\sqrt{n}(\overline{X}_n^* - \overline{X}_N)$를 생성하였다. 표준 $\overline{X}$ 관리도 작성에 필요한 σ를 추정할 때, 부분군의 표본분산을 이용한 $Sp=\sqrt{\sum_{i=1}^{k} S_i^2/k}$ (k는 부분군의 개수)와 총 관측치 N을 이용한 표본분산 $S=\sqrt{\sum_{i=1}^{k}(\overline{X}_i - \overline{X}_N)^2/(k-1)}$ 두 가지 경우로 나누어 수행했다. <표 3>의 근사적 참값은 N=500,000인 관측치들을 대상으로 얻은 경험적 분포로부터 취해진 값이며, 괄호 안의 숫자는 $\tau_{1-\frac{\alpha}{2}}$와 $\tau_{\frac{\alpha}{2}}$값에 대한 붓스트랩 추정치의 표준편차이다. 관측치 개수를 500,000으로 한 이유는 참값을 알기 위해 의도적으로 큰 값을 취했다.

< 표 3 > 자기상관이 존재하는 관측치에 대한 표준 관리한계선과 붓스트랩 관리한계선

				관리 하한선	관리 상한선	$\tau_{\frac{\alpha}{2}}$	$\tau_{1-\frac{\alpha}{2}}$
AR(1)	ϕ=0.1 (n=5)	근사적 참값		-0.948	0.948	-2.120	2.119
		부트스트랩	MBB	-0.906	0.947	-2.059(0.296)	2.084(0.305)
			TB	-0.917	0.955	-2.084(0.275)	2.103(0.280)
		표준방법	S(pooled)	-0.842	0.871	-1.916	1.916
			전체 S	-0.860	0.889	-1.956	1.956
	ϕ=0.5 (n=6)	근사적 참값		-1.440	1.440	-3.527	3.527
		부트스트랩	MBB	-1.383	1.354	-3.348(0.641)	3.355(0.616)
			TB	-1.370	1.339	-3.316(0.616)	3.319(0.567)
		표준방법	S(pooled)	-0.805	0.773	-1.932	1.932
			전체 S	-0.929	0.897	-2.236	2.236
	ϕ=0.9 (n=14)	근사적 참값		-3.671	3.644	-13.734	13.635
		부트스트랩	MBB	-3.130	3.087	-11.817(3.947)	11.446(3.430)
			TB	-3.141	3.084	-11.859(3.918)	11.436(3.363)
		표준방법	S(pooled)	-0.697	0.753	-2.714	2.714
			전체 S	-1.093	1.149	-4.196	4.196

<표 3>에서 붓스트랩을 적용하여 작성한 관리한계선이 표준방법을 적용한 관리도보다 참값에 훨씬 근접함을 알 수 있다. $\overline{X}_n^*$을 발생시키는데 있어서 이동블록 붓스

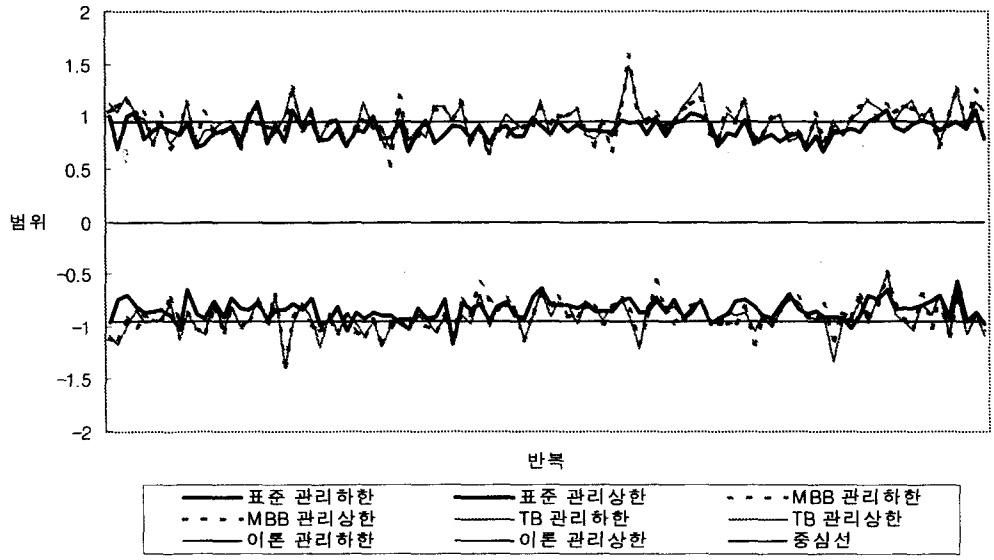
트랩과 분계점 붓스트랩을 사용한 결과는 거의 유사하나 독립반복시행에 따른 $\tau_{1-\frac{\alpha}{2}}$와 $\tau_{\frac{\alpha}{2}}$값의 변동에서 분계점 붓스트랩이 약간 우수함을 알 수 있다. 또한, 관측치들 간의 자기상관이 증가할수록 표준방법을 적용한 관리한계선은 참값을 심하게 하향 추정하게 되는 것을 볼 수 있다. 그 이유는 다음과 같이 설명된다.

양의 자기상관이 존재하는 관측치들에 대한 $\overline{X}_n$의 분산은

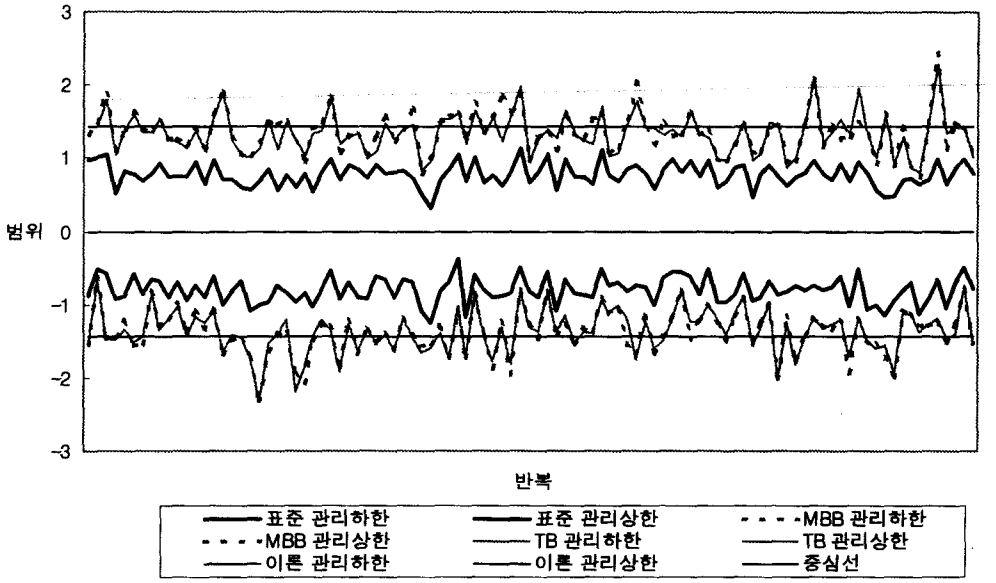
$$VIF = \frac{Var[X]}{n} \cdot 2\sum_{i=1}^{\infty}(1-\frac{i}{n})\rho_i$$

여기서 VIF는 Variance Inflation Factor로서 iid 때보다 양의 자기상관정도로 인하여 분산이 팽창되는 정도를 나타냄.

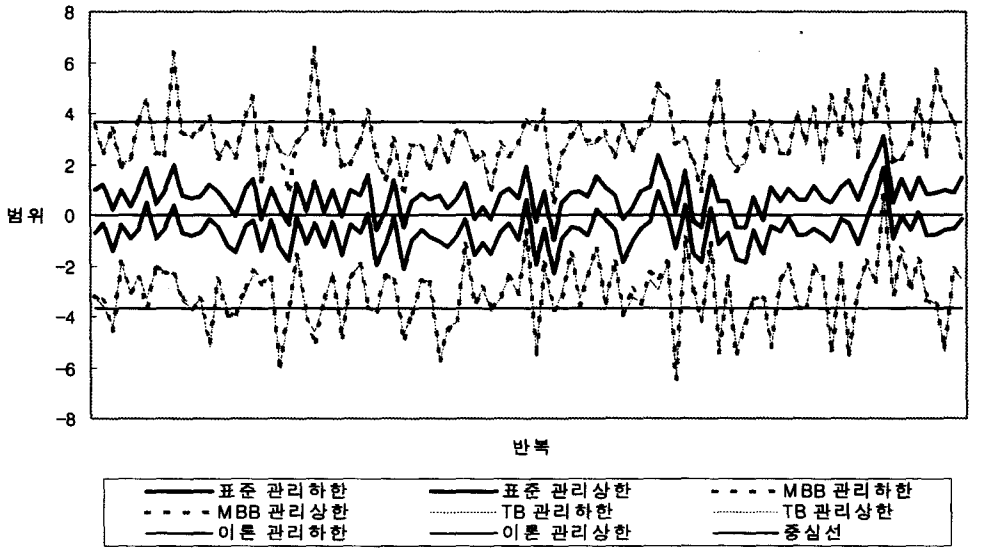
의 영향을 받아, 적절한 부분군의 크기 n에 대하여 자기상관의 정도가 증가할수록 증가하게 된다. 그 결과로, 자기상관의 정도가 심해질수록 관리한계선의 폭도 따라서 증가해야만 한다. 하지만, 표준방법에 의해 설정된 관리한계선은 관측치 X_i의 표본분산 S에 기초하기 때문에, $\overline{X}_n$의 분산의 증가를 반영하지 못한다. 여러 가지 ϕ값을 가지는 AR(1)모형을 따르는 관측치들을 대상으로 이것을 보인 결과를 다음 그림들을 통해 알 수 있다. <그림 5>부터 <그림 7>까지의 관리한계선중 TB와 MBB의 구분이 어려운 것은 두 가지 방법이 거의 근사한 성능을 보이고 있기 때문이다.



< 그림 5 > ϕ=0.1인 AR(1)모형을 따르는 관측치에 대한 관리한계선



< 그림 6 > ϕ=0.5인 AR(1)모형을 따르는 관측치에 대한 관리한계선



< 그림 7 > ϕ=0.9인 AR(1)모형을 따르는 관측치에 대한 관리한계선

6. 결론

통계적 공정관리에서 널리 사용되는 $\overline{X}$ 관리도는 임의의 공정에서 관측된 관측치들이 서로 독립적이고 동일한 정규분포를 따른다는 가정 하에서 $\overline{X}$의 분포를 근거로 관리한계선을 확정하였다. 그러나, 실제 공정에서는 관측치들이 정규분포를 따르지 않거나 관측치들간에 자기상관이 존재하는 경우가 많이 발생하며, 이러한 경우에 표준 $\overline{X}$ 관리도를 적용하면 관측치들이 따르는 분포의 편향성과 관측치들간의 자기상관을 반영하지 못하는 $\overline{X}$의 분포를 추정하게 되어 결과적으로 관리도의 성능이 저하된다. 특히, 양의 자기상관이 존재할 때에 $\overline{X}$의 분산은 자기상관의 정도에 영향을 받아 증가하게 되어 관리도 구축 시에 설정한 false alarm 비율보다 높은 false alarm을 발생하게 된다.

본 연구에서는 자기상관을 갖는 공정에 붓스트랩 방법을 적용하여 관리도를 작성하는 방법을 제안하였다. 제안한 방법은 관측치들간에 존재하는 자기상관을 잘 반영하도록 하기 위하여 적절한 부분군의 크기를 결정하였고 관측치들의 재표집을 통하여 관측치들간의 자기상관을 잘 반영하는 $\overline{X}_n$의 표본분포를 찾아내어 관리한계선을 확정하였다. 표준 관리도와 붓스트랩을 이용한 관리도에 대하여 실험을 수행한 결과 붓스트랩을 이용한 관리도의 성능이 더욱 우수한 것으로 나타났다.

참고문헌

[1] Efron, B.(1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, Vol. 7, pp. 1-26.

[2] Efron, B.(1982), "The Jackknife, the Bootstrap and other Resampling Plans," *Society for Industrial and Applied Mathematics*, Philadelphia.

[3] Efron, B. and Tibshirani, R.(1986), "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy," Statistical Science, Vol. 1, No.1, pp. 54-77.

[4] Kim Y.(1993), "The Threshold Bootstrap: A New Approach to Simulation Output Analysis," *Winter Simulation Conference 93*, pp. 498-502.

[5] Künsch, H.(1989), "The Jackknife and the Bootstrap for General Stationary Observations," *The annals of Statistics*, Vol. 17, pp. 1217-1241.

[6] Liu, R. and Singh, K(1992), "Moving Blocks Jackknife and Bootstrap Capture Weak Dependence," in *Exploring the Limits of Bootstrap*, eds. R. LePage and L. Billard, John Wiley, New York, pp. 225-248.

[7] Park, Daesu and Willemain, T.(1999), "The Threshold Bootstrap and

Threshold Jackknife," *Computational Statistics and Data Analysis*, Vol. 31, pp. 187-202.

[8] Shewhart, W. A.(1931), *Economic Control of Quality Manufactured Product*, D. Van Nostrand Co.. New York.