붓스트랩 방법에 의한 95/95 확률 및 신뢰도를 갖는 허용구간의 포함확률 보정

이윤희[1], 김홍기[2], 신희성[3], 김호동[4]

요 약

붓스트랩 기법에 의한 k 인자 허용구간방법을 95/95 확률 및 신뢰도를 갖는 허용구간에 활용하기 위하여 모의실험을 수행하였다. 그 결과 소표본 및 적당한 크기의 표본에서 추정된 신뢰도값은 실제 신뢰도값 95와 약 6~21% 정도의 차이를 나타냈고, 이 차이는 표본크기가 커질수록 점점 줄어들었다. 더불어 기존방법에 보간법 등을 가미한 방법들을 제안하여 이들에 의한 결과를 기존결과와 비교하였다.

주요용어 : 허용구간, 허용한계, 붓스트랩, 포함확률

1. 서 론

표본으로부터 미래의 동일한 조건에서 얻어지는 데이타의 포함구간 예측을 위하여 허용구간방법이 활용되고 있다. 대부분의 분야에서 표본의 모집단이 정규분포하는 것으로 간주하고 k 인자 허용구간을 결정한다. 그러나 표본의 모집단이 정규분포하지 않을 경우, k 인자 허용구간은 잘못된 결과를 유발시킬 수 있다. 이런 점에 착안하여 최근에도 표본의 모집단분포에 종속적인 허용구간 결정방법이 활발히 연구되고 있다. Fernholz등[1]은 전통적인 k 인자 허용구간방법에 붓스트랩기법을 활용함으로써 모든 분포함수에 로버스트(robust)한 보정된 포함확률을 갖는 k 인자 허용구간방법을 개발하였다. Fernholz[2]는 참고문헌[1]에서 제시한 방법이 분포함수에 로버스트할 뿐 아니라 이상치(outlier)에 대해서도 로버스트하다는 것을 보였다.

본 논문은 표본의 모집단이 비정규분포인데 이를 간과하고 전통적인 k 인자 허용구간을 이용했을 때 발생하는 비정규성 정도를 k 인자 허용구간의 k 값에 반영하고자 하는 연구의 한 부분이다. 본 논문에서는 참고문헌[1]에서 제시한 방법을 보수적 수준인 95/95 확률 및 신뢰도의 허용구간에 적용하였고, 보간법 등을 가미하여 기존방법을 부분적으로 개선하고 이것을 여러 분포함수 대상으로 보정된 포함확률과 그에 따른 신뢰도를 구한 후 참값과 비교를 통하여 이들 방법을 검증하였다.

2. 허용구간(tolerance intervals)[3-4]

허용구간은 모집단분포함수의 특정한 확률(비율)부분을 결정하는 경계값을 표본의 평균과

1) 충남대학교 통계학과 박사과정 수료, (305-764) 대전광역시 유성구 궁동 220
2) 충남대학교 통계학과 교수, (305-764) 대전광역시 유성구 궁동 220
3) 한국원자력연구소 사용후핵연료특성계량화기술개발과제 책임연구원, (305-353) 대전광역시 유성구 덕진동 150
4) 한국원자력연구소 사용후핵연료특성계량화기술개발과제 책임연구원, (305-353) 대전광역시 유성구 덕진동 150

표준편차로 추정한 일종의 신뢰구간(confidence interval)이다. 표본통계량 L과 U가 다음 식을 만족할 때, 임의의 구간 $[L,U]$를 모집단의 누적분포함수(cumulative distribution function) F에서 정의한 p/γ 확률 및 신뢰도를 갖는 허용구간이라고 한다.

$$P\{F(U)-F(L)\geq p\}\geq\gamma \quad (1)$$

허용구간의 L과 U를 허용한계(tolerance limits)라고 한다. 모집단의 분포함수 F를 미지의 평균 μ와 표준편차 σ를 갖는 정규분포로 가정하면, p/γ 확률 및 신뢰도를 갖는 허용구간 $[L,U]$는 μ와 σ의 각 표본추정량 $\overline{X}=\sum_{i=1}^{n}X_i/n$과 $S=\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2/(n-1)}$에 의해서 $[\overline{X}-kS, \overline{X}+kS]$ 된다. 이 때 k는 정규확률밀도함수에 의한 허용구간내의 면적이 신뢰도 γ 에서 적어도 p가 되게 하는 허용인자(tolerance factor)다.

3. 붓스트랩기법에 의한 보정된 포함확률을 갖는 k 인자 허용구간

3.1 D_n 통계량

정규분포 가정하에서 구한 허용구간이 실제 모집단 분포함수의 p 확률을 만족하는지를 알기 위해서는 표본의 경험적 분포함수(empirical distribution function)를 이용해 추정할 수 있다. 참고문헌[1]에서는 모집단의 임의의 확률분포함수 F_0와 표본의 경험적 분포함수 F_n에서 허용구간에 의해 결정되는 확률의 차이를 다음과 같은 확률변수로 정의하였다.

$$D_n=\sqrt{n}\left(F_n(\overline{X}+kS)-F_n(\overline{X}-kS)-\left(F_0(\overline{X}+kS)-F_0(\overline{X}-kS)\right)\right) \quad (2)$$

식(2)에서 F_0가 유한한 평균 및 표준편차를 갖고 연속인 분포함수라면, 확률변수 D_n은 점근적으로(asymptotically) 정규분포를 한다는 것이 알려져 있다[1]. (0,1) 사이의 고정된 γ 에 대해서, $P\{D_n\leq d_\gamma\}=\gamma$을 만족하는 d_γ는 확률변수 D_n의 γ 분위수(quantile)이고, 이로부터 다음 식이 정의된다.

$$P\{F_0(\overline{X}+kS)-F_0(\overline{X}-kS)\geq F_n(\overline{X}+kS)-F_n(\overline{X}-kS)-d_\gamma/\sqrt{n}\}=\gamma \quad (3)$$

식(3)의 부등호 우측부분을 $p_n=F_n(\overline{X}+kS)-F_n(\overline{X}-kS)-d_\gamma/\sqrt{n}$ 하면, 식(3)은 다음 식과 같이 표현된다.

$$P\{F_0(\overline{X}+kS)-F_0(\overline{X}-kS)\geq p_n\}=\gamma \quad (4)$$

만약 모집단의 분포함수 F_0가 기지이면 p_n은 허용한계 $\overline{X}\pm kS$에서의 모집단분포함수의 면적값의 하한(lower bound)이고 이 값은 표본 및 모집단분포함수에 의하여 결정된다. 일반적으로 F_0는 미지이므로 D_n의 분포함수는 미지이고, 따라서 d_γ와 p_n도 미지이다. 여기서 표본에 의해 종속적으로 결정되는 포함확률 p_n을 추정하기 위해 붓스트랩 방법이 활용된다.

3.2 붓스트랩 통계량 D_n^*

미지인 모집단의 분포함수에 기인한 표본통계량의 분포는 붓스트랩 표본통계량의 분포에 의해 근사시킬 수 있다[5-6]. 이것에 의해 식(2)에서 D_n의 분포는 D_n의 붓스트랩 표본통계량에 의해 추정될 수 있고, 이 붓스트랩 통계량은 다음 식과 같이 표현된다.

$$D_n^*=\sqrt{n}\left(F_n^*(\overline{X}^*+kS^*)-F_n^*(\overline{X}^*-kS^*)-\left(F_n(\overline{X}^*+kS^*)-F_n(\overline{X}^*-kS^*)\right)\right) \quad (5)$$

식(5)에서 F_n^*는 붓스트랩 표본에서 정의되는 경험적 분포함수이고 $\overline{X}^*$와 S^*는 붓스트랩 표본의 평균과 표준편차이다. p_n을 추정하기 위해서는 식(3)의 d_γ를 추정해야 하고 이를 위해 주어진 표본에서 B 회의 붓스트랩 표본을 반복 추출하여, 각 붓스트랩 표본에서 식(5)에 의한 D_n^*

를 계산한다. B 개의 D_n^* 값들 중에서 γ 분위수에 해당되는 d_γ^*는 d_γ의 추정량이라고 할 수 있고 이러한 논리에 근거하여 식(3)의 d_γ 대신에 d_γ^*를 대치하면 다음과 같은 식이 성립된다.

$$P\{F_0(\overline{X}+kS)-F_0(\overline{X}-kS)\geq F_n(\overline{X}+kS)-F_n(\overline{X}-kS)-d_\gamma^*/\sqrt{n}\}\simeq\gamma \qquad (6)$$

식(6)에서 보는 바와 같이 허용한계 $\overline{X}\pm kS$에 의한 모집단 분포함수에서의 포함확률이 적어도 붓스트랩 추정량을 대치한 하한이 될 확률은 근사적으로 γ가 된다. 이 결과는 붓스트랩에 의한 추정량을 이용하여 식(3)의 하한을 추정하는 것을 의미하고, 따라서 식(6)에서의 하한을 $p_n^*=F_n(\overline{X}+kS)-F_n(\overline{X}-kS)-d_\gamma^*/\sqrt{n}$ 하면 식(6)은 다음 식과 같이 표현된다.

$$P\{F_0(\overline{X}+kS)-F_0(\overline{X}-kS)\geq p_n^*\}\simeq\gamma \qquad (7)$$

식(7)의 부등호 우측부분인 p_n^*를 표본에 의해 종속적으로 결정되는 보정된 포함확률(corrected-content)이라 하고, 이 확률값은 정규분포를 가정하여 구한 허용구간의 초기 포함확률 p를 표본의 모집단의 특성을 붓스트랩으로 반영하여 보정한 값이다. 이 방법은 모든 분포함수에서 보정된 p_n^*가 γ 확률을 만족하기 때문에, 기존의 정규분포 가정에서 정의되었던 k 인자 허용구간에 비해 그 활용범위가 매우 넓다. 확률 γ가 모든 분포함수에서 유지되는 이러한 특징을 분포함수에 대하여 로버스트하다고 표현한다. 식(2)를 기점으로 만들어진 보정확률 및 신뢰도를 갖는 k 인자 허용구간은 모든 분포함수에 로버스트할 뿐 아니라, 식(2)에서 정의된 통계량의 영향함수(influence function)와 분기점(breakdown point)도 이상치에 대하여 로버스트하다고 알려져 있다[2].

4. 모의실험

4.1 벤치마크

참고문헌[1]의 결과와 비교하기 위하여 모의실험의 모든 조건을 동일하게 하였다. Fig. 1에서 제시한 바와 같이 정규분포, 라플라스(Laplace) 분포 및 Student-t 분포를 고려하였다. 각 분포로부터 확률 및 신뢰도가 75/90과 90/95인 경우에는 표본크기 20, 80 및 320을, 95/95인 경우에는 표본크기 500을 추가하였고, 각 표본크기에 대하여 10000 번씩 반복 추출하였다. 보정확률을 구하기 위해 붓스트랩 재표본 횟수를 2000 번으로 하였다. 각 분포함수에서 반복 추출한 전체 표본 중 식(7)의 확률부분을 만족하는 표본의 비율을 계산하여 B-Conf.(Bootstrap Confidence)를 구하였고, 마찬가지로 식(1)의 확률부분을 만족하는 표본의 비율을 계산하여 S-Conf.(Standard Confidence)를 구하였다. 또한 평균보정확률 $\overline{p}_n^*$와 보정확률의 표준편차 sd^*를 구하였다.

4.2 경험적 분포함수 해석의 개선

참고문헌[1]에서 제안된 방법은 허용한계 값이 표본의 i 번째 데이터 $x_{(i)}$와 $i+1$ 번째 데이터 $x_{(i+1)}$ 사이에 포함되어 있을 때, 이 값에서의 경험적 분포함수 값으로 $x_{(i)}$의 확률값 $F_n(x_{(i)})$를 갖게 한다. 그러나 허용한계 값이 $x_{(i+1)}$에 더 가까운 값일 경우에 이 방법에 의한 확률결정은 불합리해 보인다. 이를 개선하는 방법으로 허용한계 값이 두 데이터 중에서 더 가까운 쪽의 확률값을 갖도록 하는 방법을 제안하였고, 또한 두 점의 직선의 방정식을 이용하여 허용한계 값의 확률값을 결정하는 보간법에 의한 방법을 제안하였다. 본 논문에서는 기존방법을 EM(Existing Method), 가까운 쪽의 확률할당 방법을 NPM(Near Point Method) 그리고 보간법에 의한 방법을 IM(Interpolation Method)이라고 명명한다.

4.3 결과분석

세 가지 분포함수를 기준으로 확률 및 신뢰도 75/90과 90/95 각 경우의 결과인 B-Conf., S-Conf. 및 평균보정확률 등을 참고문헌[1]의 결과와 함께 Table 1과 2에 제시하였다. 본 연구에서 기존방법을 재현했던 EM에 의한 결과가, 두 경우에 모든 분포함수에서 기존결과와 거의 일치하였다. 이로서 EM이 기존방법을 잘 재현한다고 볼 수 있다.

Table 1과 2에서 보는 바와 같이 정규분포에서는 S-Conf.가 신뢰도 γ에 근사하지만 비정규분포에서는 큰 차이를 보이고 있다. 표본의 모집단이 비정규분포인데도 전통적인 k 인자 허용구간방법을 사용한다면 요구하는 신뢰도 γ와 포함확률 p가 모두 만족되지 않는다는 것을 알 수 있다. 반면에 B-Conf.는 75/90에서는 모든 분포함수 및 모든 표본크기에서 대체로 신뢰도 γ를 만족하였으나 90/95에서는 크기가 20인 소표본에서 실제값 95와 약 5 % 정도의 차이를 보이는 것으로 나타나 포함확률이 커지면 특히 소표본에서는 이 방법에 의한 B-Conf.가 오차를 갖는다는 것을 알 수 있다. 한편 주어진 포함확률은 붓스트랩 기법에 의해 실제 모집단분포의 포함확률로 보정되었다.

참고문헌[1]에서는 95/95 확률 및 신뢰도를 갖는 허용구간에 대한 모의실험결과가 없는 관계로 기존방법을 활용하기 위해 사전모의실험결과의 필요성을 인식하여 모의실험 수행 후 그 결과를 Table 3에 제시하였다. 95/95에서는 소표본 및 중표본에서 B-Conf.가 실제값과 약 6~21 %의 큰 차이를 나타냈고 320의 대표본에서도 약 2 % 정도의 차이를 보이고 있다. 이 오차는 표본크기가 500 이상으로 커질 때 비로소 신뢰도를 만족하여 줄어드는 것을 볼 수 있다. 이 결과로부터 95/95인 경우에는 소표본을 비롯해 꽤 큰 표본에서도 B-Conf.를 만족하지 못하는 비실용적인 면을 발견하였다.

본 연구에서 제안한 NPM과 IM에 의한 결과는 75/90에서는 대부분의 경우 기존결과와 거의 일치하였고, 90/95인 경우는 크기가 80 이상에서는 거의 일치하였으나 크기가 20인 소표본에서는 B-Conf.가 기존방법에 비해 참값에 근접한 값을 보였다. 95/95의 결과에서는 80이상의 대표본에서는 앞에서 기술했듯이 기존결과와 동일하게 약 2~6 % 정도 참값과의 차이를 나타냈으고, 표본크기 20에서는 기존방법과 달리 참값에 근접한 값을 보였으나 평균보정확률이 기존방법에 비해 참값과 차이를 보이는 것으로 나타났다. 본 연구에서 제안한 NPM과 IM이 95/95인 경우에 기존방법이 갖고 있는 문제점을 해결하지는 못하였다.

5. 결론 및 향후계획

본 논문에서는 기존방법을 보수적 수준인 확률 및 신뢰도 95/95를 갖는 허용구간에 적용한 결과, 정규분포의 소표본 및 중표본에서 B-Conf.가 참값과 약 6~21 % 정도의 차이를 보임으로써 신뢰도가 유지되지 않는 것을 발견하였다. 한편, 본 연구에서 제안한 NPM과 IM의 결과로는 대표본에서는 기존방법과 거의 일치하였고, 소표본에서는 기존방법이 큰 오차를 보인 반면, 이들 두 방법은 참값에 거의 근접한 값을 보였다. 그러나 평균보정확률이 기존방법과는 달리 참값과 차이를 보임으로서 기존방법에 비해 크게 좋다고 말할 수는 없었다. 따라서 붓스트랩에 의한 95/95 확률 및 신뢰도를 갖는 허용구간방법을 소표본 및 중표본에 적용할 때는 신뢰도가 유지되지 않는다는 점을 고려하여 세심한 주의를 해야 한다.

추후에는 앞에서 제시한 문제점들을 해결할 수 있는 방안에 대해 연구할 계획이며, 이를 바탕으로 본 연구의 궁극적 목적인 보정된 포함확률에 따라 k 값을 보정해주는 방법을 연구할 계획이다. 또한 본 연구에서 사용한 허용한계는 전통적인 추정량으로서 이상치에 대하여 로버스트하지 않다고 알려져 있는데 이를 극복할 수 있는 방법을 연구할 계획이다.

본 연구는 과학기술부의 원자력 연구개발사업의 일환으로 SF특성계량화기술개발과제에서 수행한 것임.

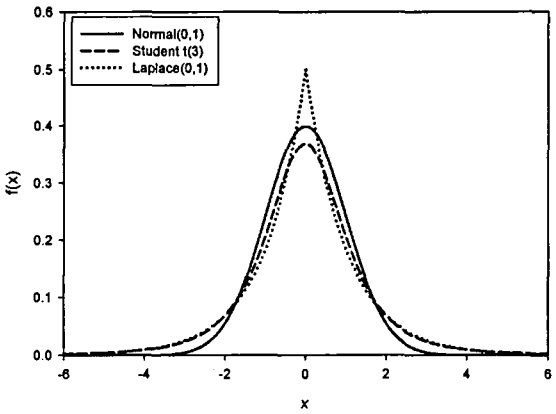

Fig. 1. Three Probability Density Function Curves.

Table 1. Simulation Results for 75/90 Probability and Confidence for Three Populations

Popul-ation	n	k^*	B-Conf.				S-Conf.		$\bar{p}^*_n \pm sd^*$			
			Ref.(1)	This Study			Ref.(1)	This Study	Ref.(1)	This Study		
				EM	IM	NPM				EM	IM	NPM
Normal (0,1)	20	1.505	.8944	.8925	.8789	.8917	.8963	.8953	.7357±.0465	.7352±.0470	.7499±.0324	.7398±.0414
	80	1.292	.8941	.8944	.8948	.8967	.8998	.9005	.7405±.0282	.7401±.0280	.7407±.0268	.7403±.0268
	320	1.214	.8934	.8952	.8946	.8959	.8981	.8969	.7443±.0148	.7438±.0149	.7439±.0148	.7438±.0148
Laplace (0,1)	20	1.505	.9012	.9038	.9035	.9106	.9167	.9183	.7489±.0517	.7458±.0526	.7562±.0348	.7473±.0416
	80	1.292	.8966	.8990	.8954	.8968	.9702	.9796	.7763±.0331	.7710±.0330	.7778±.0319	.7768±.0324
	320	1.214	.8955	.8924	.8925	.8911	.9996	.9996	.7809±.0176	.7904±.0173	.7906±.0172	.7905±.0172
Student t(3)	20	1.505	.9103	.9044	.9012	.9100	.9292	.9209	.7621±.0545	.7608±.0552	.7673±.0365	.7577±.0417
	80	1.292	.9049	.9020	.9030	.9044	.9871	.9854	.8095±.0474	.8091±.0472	.8104±.0458	.8095±.0460
	320	1.214	.9015	.9001	.9006	.8998	1.0000	1.000	.8344±.0341	.8341±.0332	.8343±.0332	.8341±.0332

Table 2. Simulation Results for 90/95 Probability and Confidence for Three Populations

Popul-ation	n	k^*	B-Conf.				S-Conf.		$\bar{p}^*_n \pm sd^*$			
			Ref.(1)	This Study			Ref.(1)	This Study	Ref.(1)	This Study		
				EM	IM	NPM				EM	IM	NPM
Normal (0,1)	20	2.310	.9040	.9035	.9455	.9414	.9506	.9464	.8985±.0393	.8981±.0393	.8910±.0262	.8894±.0304
	80	1.907	.9334	.9349	.9331	.9369	.9520	.9505	.8932±.0206	.8931±.0205	.8954±.0187	.8936±.0192
	320	1.763	.9421	.9423	.9427	.9430	.9483	.9482	.8952±.0097	.8952±.0103	.8955±.0100	.8953±.0100
Laplace (0,1)	20	2.310	.9027	.9007	.9201	.9163	.8754	.8711	.8634±.0387	.8630±.0388	.8663±.0254	.8653±.0310
	80	1.907	.9379	.9427	.9436	.9456	.8610	.8696	.8765±.0203	.8762±.0204	.8784±.0181	.8770±.0186
	320	1.763	.9445	.9416	.9412	.9426	.8834	.8759	.8895±.0113	.8897±.0111	.8900±.0109	.8898±.0110
Student t(3)	20	2.310	.8907	.8890	.9143	.9105	.9024	.8824	.8655±.0385	.8643±.0383	.8662±.0267	.8658±.0332
	80	1.907	.9510	.9462	.9460	.9479	.9278	.9223	.8912±.0225	.8912±.0226	.8927±.0203	.8910±.0207
	320	1.763	.9502	.9496	.9484	.9499	.9810	.9840	.9130±.0179	.9127±.0178	.9130±.0176	.9128±.0176

Table 3. Simulation Results for 95/95 Probability and Confidence for Three Populations

Population	n	k^*	B-Conf.			S-Conf.	$\bar{p}^*_n \pm sd^*$		
			EM	IM	NPM		EM	IM	NPM
Normal (0,1)	20	2.752	.7382	.9427	.9669	.9493	.9485±.0364	.9277±.0216	.9286±.0267
	80	2.272	.8947	.8979	.8984	.9480	.9454±.0182	.9480±.0157	.9463±.0170
	320	2.100	.9285	.9285	.9307	.9482	.9459±.0088	.9463±.0086	.9460±.0086
	500	2.070	.9393	.9365	.9380	.9504	.9464±.0069	.9466±.0068	.9466±.0068
	1000	2.036	.946	.945	.948	.960	.9478±.0047	.9478±.0047	.9477±.0047
Laplace (0,1)	20	2.752	.8459	.9185	.9079	.7792	.9102±.0404	.9033±.0255	.9030±.0298
	80	2.272	.9306	.9273	.9337	.6663	.9158±.0170	.9185±.0146	.9165±.0154
	320	2.100	.9364	.9364	.9383	.4058	.9264±.0089	.9268±.0087	.9266±.0087
	500	2.070	.9366	.9358	.9355	.2935	.9287±.0071	.9289±.0070	.9288±.0071
Student t(3)	20	2.752	.8268	.9050	.8895	.7654	.9081±.0431	.9026±.0271	.9037±.0304
	80	2.272	.9317	.9266	.9360	.7280	.9214±.0174	.9232±.0144	.9211±.0150
	320	2.100	.9470	.9467	.9465	.7502	.9388±.0124	.9392±.0121	.9389±.0121
	500	2.070	.9467	.9451	.9464	.7714	.9430±.0112	.9432±.0111	.9430±.0111

- k^* is calculated by Howe's method[7].

참 고 문 헌

[1] Luisa T. Fernholz and John A. Gillespie(2001), Content-Corrected Tolerance Limits Based on the Bootstrap, Technometrics, Vol. 43, No. 2, 147-155.

[2] Luisa T. Fernholz(2002), Robustness Issues regarding Content-corrected Tolerance Limits , Matrika, Vol. 55, 53-66.

[3] Wald, A and J. Wolfwiz(1947), Tolerance Limits for a Normal Distribution, The Analysis of Mathematical Statistics, Vol. 17, 208-215.

[4] Odeh, R. B.(1980), *Tables for Normal Tolerance Limits, Sampling Plans, and Screening*, New York: Marcel Dekker.

[5] B. Efron(1979), Bootstrap Methods : Another Look at the Jackknife, The Annals of Statistics, Vol. 7, No. 1, 1-26.

[6] B. Efron. and R. J. Tibshirani(1993), *An Introduction to the Bootstrap*, Chapman & Hall.

[7] Guenther, W. C.(1972), Tolerance Intervals for Univariate Distribution, Naval Research Logistics Quarterly, Vol. 19, 309-333.