응용통계연구 제 11권 2호, 1998년, pp. 287-302

일반적 통계량의 분포함수에 대한 안부점 근사

나종화[1)]

요 약

표본평균(sample mean)의 밀도함수(density function)와 분포함수(distribution function)에 대한 안부점 근사(saddlepoint approximation)는 Daniels(1954, 1987), Lugannani와 Rice(1980)등에 의하여 유도 되었으며, 이 근사식들의 정확도는 대표본(large sample)의 경우는 물론 소표본(small sample)의 경우에도 매우 뛰어난 것으로 알려 져 있다. 최근 Easton과 Ronchetti(1986)는 일반적 통계량(general statistics)의 밀도함수에 대한 안부점 근사법을 제안하였고, 분포함수에 대한 근사로는 밀도함수에 대한 안부점 근사식을 직접 수치적으로 적분하는 방법을 제안하였다. 본 논문에서는 일반적 통계량의 분포함수에 대한 안부점 근사법을 제안하고, 이를 표본분산(sample variance)과 스튜던트화 평균(studentized mean)의 분포함수에 대한 근사에 적용하였다.

1. 서 론

안부점 근사의 기법은 Daniels(1954)가 처음 통계학 분야에 소개한 이래 많은 연구가 진행되어 왔으며, 최근에는 통계학의 여러 분야에서 각종 통계량의 밀도함수나 분포함수의 근사등에 중요한 도구로 사용되고 있다. 그 가운데 Barndorff-Nielsen과 Cox(1979)와 Reid(1988)는 주로 표본 평균과 관련한 여러 가지 형태의 밀도함수와 분포함수에 대한 안부점 근사와 그의 통계적 응용에 대하여 자세히 다루고 있다. 안부점 근사와 관련한 많은 연구들은 주로 근사대상인 통계량의 누율생성함수(cumulant generating function : CGF)에 대한 정보를 정확히 아는 경우에 대하여 진행되어 왔으며 그 정확성에 대해서는 많은 연구에서 검증되어 왔다. 최근에는 좀 더 다양한 형태의 통계량과 통계적 응용분야에 안부점 근사 방법이 사용되고 있으며, 그 가운데 Srivastava와 Yau(1989)는 비선형 통계량(nonlinear statistics)에 대하여 Wood, Booth와 Butler(1993)는 비정규 극한분포(nonnormal limiting distribution)를 가지는 통계량의 분포에 대하여 연구하였다. 특히, Davison과 Hinkley(1988)와 Wang(1988)은 붓스트랩(bootstrap) 분포의 근사에 안부점 근사의 기법을 적용함으로써 이 분야에 대한 가능성을 제시하였다. Field와 Ronchetti(1990)는 M-통계량을 비롯한 로버스트 통계량(robust statistics)들의 분포에 대한 안부점 근사법을 제시 하였으며, 일반화 선형모형(generalized linear model) 분야에서도 추정된 회귀모수의 분포등에 대한 근사에 많은 연구가 현재 진행되고 있다.

Easton과 Ronchetti(1986)는 일반적 통계량의 밀도함수(density)에 대한 안부점 근사식을 유도하였고, 분포함수에 대한 근사는 밀도함수에 대한 근사식으로부터 직접 수치적 적분을 통하여 근사하는 방법을 제안하였다. 본 논문에서는 표본평균의 분포함수의 근사에

1) (361-763) 충북 청주시 흥덕구 개신동 산48, 충북대학교 통계학과, 조교수

대한 Daniels(1987)의 기법을 사용하여 일반적 통계량의 꼬리확률에 대한 근사식을 제안하고, 이에 대한 통계적 응용으로 비선형 통계량인 표본분산(sample variance)과 스튜던트화 평균(studentized mean)의 분포함수에 대한 근사에 적용함으로써 제안된 근사식의 정확성을 확인하였다.

2. 일반적 통계량의 분포함수에 대한 안부점 근사

2.1. EASTON-RONCHETTI의 방법

일반적 통계량의 밀도함수에 대한 Easton과 Ronchetti(1986)의 방법은 다음과 같다. 확률표본 $X_1, \cdots, X_n$에 기초한 임의의 통계량 $V_n(X_1, \cdots, X_n)$의 밀도함수 $f_n(x)$에 대한 안부점 근사는 통계량 V_n의 누율생성함수(CGF)를 $K_n(t)$라 할 때 다음의 식으로 주어진다.

$$f_n(x) = \left[\frac{n}{2\pi R_n''(t_0)}\right]^{1/2} e^{n(R_n(t_0)-t_0x)}\{1+O(n^{-3/2})\} \tag{2.1}$$

위 식에서 $R_n(t) = K_n(nt)/n$이고 t_0는 다음의 안부점 방정식을 만족하는 근이다.

$$R_n'(t_0) = x \tag{2.2}$$

또한 $R_n'(\cdot)$, $R_n''(\cdot)$ 및 $R_n^{(r)}(\cdot)$은 각각 $R_n(\cdot)$의 1, 2차 및 r차 미분을 의미한다. 위의 안부점 근사식 (2.1)의 정확도에 대해서는 여러 연구에서 검증되어 있으나 통계량 V_n의 누율생성함수(CGF)를 구할 수 있는 경우는 극히 제한적이기 때문에 다양한 형태의 통계량에 대해서 이 식을 적용하기 어렵다. Easton과 Ronchetti(1986)은 통계량 V_n의 처음 4차 누율들(cumulants)을 이용하여 $K_n(t)$를 다음의 식

$$\tilde{K}_n(t) = \kappa_{1n}t + \frac{\kappa_{2n}}{2!}t^2 + \frac{\kappa_{3n}}{3!}t^3 + \frac{\kappa_{4n}}{4!}t^4 \tag{2.3}$$

으로 근사하였다. 여기서 $\kappa_{1n} = \mu_n = E(V_n)$, $\kappa_{2n} = \sigma_n^2 = Var(V_n)$이고 κ_{3n}, κ_{4n}은 각각 통계량 V_n의 3차와 4차 누율을 의미한다. 즉, 식 (2.1)과 (2.2)에서 $K_n(t)$ 대신 $\tilde{K}_n(t)$를 사용하여 일반적 통계량 V_n의 밀도함수에 대한 근사를 다음의 식으로 구하였다.

$$\tilde{f}_n(x) \approx \left[\frac{n}{2\pi\tilde{R}''(t_0)}\right]^{1/2} e^{n\{\tilde{R}_n(t_0)-t_0x\}} \tag{2.4}$$

여기서 $\tilde{R}_n(t) = \tilde{K}_n(nt)/n$이고 t_0는 다음의 안부점 방정식을 만족하는 근이다.

$$\tilde{R}_n'(t_0) = x \tag{2.5}$$

위 식에서 사용된 $\tilde{R}_n'(\cdot)$과 $\tilde{R}_n''(\cdot)$은 각각 t에 대한 1차와 2차 미분을 의미한다. Easton과 Ronchetti(1986)는 임의의 통계량 V_n의 분포함수에 대한 근사를 식 (2.4)의 밀도함수에 대한 안부점 근사식으로 부터 다음의 적분

$$\int_{-\infty}^{x} \tilde{f}_n(t)dt \tag{2.6}$$

을 직접 수치 적분하여 근사하였다. 이는 밀도함수에 대한 안부점 근사식 $\tilde{f}_n(x)$가 일반적으로 복잡한 형태를 취할 뿐 아니라 밀도함수가 가져야 할 다음의 성질, 즉 $\int_{-\infty}^{\infty} \tilde{f}_n(t)dt = 1$을 만족시키지 않기 때문에 근사된 결과를 다시 조정하는(renormalizing) 과정을 거쳐야 하며 이 역시 많은 계산이 요구되어 실질적인 응용차원에서 어려움이 있다. 이러한 문제에 대한 해결책으로 본 논문에서는 Easton과 Ronchetti(1986)가 제시한 방법을 분포함수의 근사에 직접 적용하는 방법을 제안하였다.

2.2. 제안된 방법

$X_1, \cdots, X_n$을 밀도함수가 f인 분포로부터의 확률표본이라 하고, 임의의 통계량 $V_n = V_n(X_1, \cdots, X_n)$의 밀도함수를 f_n이라 하자. V_n의 적률생성함수(moment generating function)와 누율생성함수(cumulant generating function)를 각각 $M_n(u) = \int e^{ux} f_n(x)dx$와 $K_n(u) = \log M_n(u)$이라 하고, 특성함수(characteristic function)를 $\psi(u) = M_n(iu)$ 이라 하자. $M_n(u)$가 원점을 포함하는 임의의 구간에 포함되는 실수값 u에 대하여 존재한다고 가정하자. 퓨리에 역변환공식(Fourier inversion formula)에 의하여 V_n의 꼬리확률(tail probability)은 다음과 같이 주어진다.

$$\begin{aligned} P(V_n \geq v) &= \int_v^{\infty} \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-its} M_n(it)dtds \\ &= \frac{1}{2\pi i} \int_{\tau-i\infty}^{\tau+i\infty} e^{n[R_n(t)-tv]} \frac{1}{t} dt \end{aligned} \tag{2.7}$$

여기서 τ는 임의의 실수이고, $R_n(t)$는 다음과 같이 주어진다.

$$R_n(t) = K_n(nt)/n \tag{2.8}$$

식 (2.7)의 계산을 위해 다음과 같은 변환을 생각하자.

$$R_n(t) - tv = \frac{\eta^2}{2} - \eta\hat{\eta} \tag{2.9}$$

또한 $R_n'(t) = v$의 근을 t_0라 하고 식 (2.9)의 좌우변의 최소값(minimum value)이 일치하도록 $\hat{\eta}$와 η를 찾으면 다음과 같다.

$$\hat{\eta} = sgn(t_0)\{2[t_0 v - R_n(t_0)]\}^{1/2} \tag{2.10}$$

$$\eta = \hat{\eta} + \{2(R_n(t) - tv + t_0 v - R_n(t_0))/(t - t_0)\}^{1/2} \cdot (t - t_0) \tag{2.11}$$

따라서 식 (2.7)은 다음의 식으로 주어진다.

$$\begin{aligned} P(V_n \geq v) &= \frac{1}{2\pi i} \int_{\tau-i\infty}^{\tau+i\infty} e^{\frac{n}{2}\eta^2 - n\eta\hat{\eta}} \frac{1}{\eta} d\eta \\ &+ \frac{1}{2\pi i} \int_{\tau-i\infty}^{\tau+i\infty} e^{\frac{n}{2}(\eta-\hat{\eta})^2} \left(\frac{1}{t}\frac{dt}{d\eta} - \frac{1}{\eta}\right) d\eta \cdot e^{n[R_n(t_0)-t_0 v]} \end{aligned} \tag{2.12}$$

이제 표본 평균의 분포함수에 대한 안부점 근사에서 사용된 Daniels(1987)의 기법을 적용하자. 식 (2.12)에서 $g(\eta) = \frac{1}{t}\frac{dt}{d\eta} - \frac{1}{\eta}$이라 하고, 이를 $\hat{\eta}$에서 테일러(Taylor) 전개후 각 항을 정리하면 통계량 V_n의 분포함수에 대한 다음의 근사식을 얻을 수 있다. (부록 참고)

$$Pr\{V_n \le v\} = \begin{cases} \Phi(w) + \phi(w)\left\{\frac{1}{w} - \frac{1}{\zeta} + O(n^{-3/2})\right\}, & v \neq E(V_n) \\ \frac{1}{2} - \frac{R_n^{(3)}(0)}{6\sqrt{2\pi(R_n''(0))^3}}\{1 + O(n^{-3/2})\}, & v = E(V_n) \end{cases} \tag{2.13}$$

위 식에서 $\phi(\cdot)$와 $\Phi(\cdot)$는 각각 표준정규분포의 밀도함수와 분포함수를 나타내고 w와 ζ는 다음과 같이 정의되는 통계량이다.

$$w = \sqrt{n}\hat{\eta} = [2n\{t_0 v - R_n(t_0)\}]^{1/2} sgn(t_0) \tag{2.14}$$

$$\zeta = t_0\{nR_n''(t_0)\}^{1/2} \tag{2.15}$$

여기서 R_n과 t_0는 식 (2.1)에서의 정의와 동일하고 $sgn(t_0)$는 안부점 근사식의 해 t_0가 양, 음의 또는 0의 값을 가질 때 $+1, -1, 0$ 의 값을 취하는 부호함수(sign function)이다.

본 논문에서는 일반적 통계량 V_n의 누율생성함수를 구하기 어려운 경우, 이에 대한 근사로서 통계량의 처음 4차까지의 누율(cumulants)에 기초한 Easton과 Ronchetti(1986)의 근사식 (2.3)을 사용하고자 한다. 이 방법은 정확도가 뛰어나고, 통계적 추론에 자주 사용되는 꼬리부분(tail part)의 영역에서는 2개항을 사용한 ($O(n^{-1})$의 오차를 가지는) Edgeworth 근사식 보다 뛰어남을 다음절에서 다루게 될 표본분산 및 스튜던트화 평균에 대한 근사를 통해 확인하였다. 특히 스튜던트화 평균의 예제에서는 통계량의 처음 4차 까지의 누율 계산이 불가능하여 요구되는 차수까지의 근사식을 이용하였다.

3. 통계적 응용

3.1. 표본분산의 분포함수

$X_1, \cdots, X_n$을 평균이 μ, 분산이 σ^2인 분포함수 $F(x)$로부터의 확률표본이라 하자. 모집단의 r차 적률(moment), 중심적률(central moment), 그리고 누율(cumulant)을 각각 μ_r, μ_r, κ_r이라 표기하자. 표본분산(sample variance) $S_n^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$의 분포함수에 대한 근사를 생각하자. 표본의 크기가 충분히 큰 경우 중심극한정리(central limit theorem)로부터 다음의 사실이 성립한다.

$$\sqrt{n}(S_n^2 - \sigma^2) \to N(0, \mu_4 - \sigma^4) \tag{3.1}$$

이제 표본분산의 분포함수에 대한 새로운 근사법으로, 2.2절에서 제시된 안부점 근사법을 적용하기 위해 통계량 S_n^2의 처음 4차 까지의 누율을 계산하면 다음과 같이 주어짐을 보일

수 있다.

$$\begin{aligned}
\kappa_1(S_n^2) &= E(S_n^2) = \sigma^2 = \kappa_2 \\
\kappa_2(S_n^2) &= Var(S_n^2) = \frac{\kappa_4}{n} + \frac{2\kappa_2^2}{n-1} \\
\kappa_3(S_n^2) &= \frac{\kappa_6}{n^2} + \frac{12\kappa_4\kappa_2}{n(n-1)} + \frac{4(n-2)}{n(n-1)^2}\kappa_3^2 + \frac{8}{(n-1)^2}\kappa_2^3 \\
\kappa_4(S_n^2) &= \frac{\kappa_8}{n^3} + \frac{24}{n^2(n-1)}\kappa_6\kappa_2 + \frac{32(n-2)}{n^2(n-1)^2}\kappa_5\kappa_3 + \frac{8(4n^2-9n+6)}{n^2(n-1)^3}\kappa_4^2 \\
&\quad + \frac{144}{n(n-1)^2}\kappa_4\kappa_2^2 + \frac{96(n-2)}{n(n-1)^3}\kappa_3^2\kappa_2 + \frac{48}{(n-1)^3}\kappa_2^4
\end{aligned} \tag{3.2}$$

위의 식 (3.2)의 결과는 모집단의 처음 8차 까지의 누율들에 대한 정보 만으로 구해지는 값이며, 이는 누율생성함수(cumulant generating function)를 알고 있는 어떠한 형태의 모수적 모형에 대해서도 적용할 수 있다.

3.2. 스튜던트화(STUDENTIZED) 평균의 분포함수

$X_1, \cdots, X_n$을 평균이 μ이고 분산이 $\sigma^2(>0)$인 분포로 부터의 확률표본이라 하자. $\bar{X} = \sum_{i=1}^{n} X_i/n$이고 $\hat{\sigma^2} = \sum_{i=1}^{n}(X_i - \bar{X})^2/n$일 때, 스튜던트화 평균 $T = \sqrt{n}(\bar{X} - \mu)/\hat{\sigma}$의 분포는 모집단이 정규분포를 따르는 경우를 제외하고 일반적으로 알려져 있지 않다. 따라서 다양한 모집단의 형태에 대한 통계량 T의 분포에 대한 정확한 근사법이 필요하다. 본 연구에서는 2.2절에서 제시한 안부점 근사에 기초한 T의 분포함수에 대한 근사법을 제시하고 자 한다. 이를 위해서는 통계량 T의 처음 4차 까지의 누율에 대한 정보가 필요하다. 그러나 통계량 T의 누율들에 대한 구체적인 형태로의 표현이 앞에서 다룬 표본분산의 경우와는 달리 대단히 어려운 점이 있기 때문에 본 눈문에서는 이에 대한 한가지 해결책으로 처음 4차 까지의 누율들에 대한 근사식을 사용하고자 한다. 통계량 T의 누율에 대한 근사과정은 다음과 같다.

먼저, $W_i = (X_i - \mu)/\sigma$이라 하고 Z_1과 Z_2를 다음과 같이 정의하자.

$$Z_1 = \sqrt{n}\bar{W} = \sqrt{n} \cdot \sum_{i=1}^{n} W_i/n, \qquad Z_2 = \sqrt{n} \cdot \sum_{i=1}^{n}(W_i^2 - 1)/n \tag{3.3}$$

통계량 T는 Z_1과 Z_2의 함수로 다음과 같이 표현됨을 보일 수 있다.

$$\begin{aligned}
T &= Z_1\left\{1 + \frac{1}{\sqrt{n}}Z_2 - \frac{1}{n}Z_1^2\right\}^{-1/2} \\
&= Z_1\left\{1 - \frac{1}{2\sqrt{n}}Z_2 + \frac{1}{2n}Z_1^2 + \frac{3}{8n}Z_2^2 \right. \\
&\quad \left. - \frac{6}{8n\sqrt{n}}Z_1^2Z_2 - \frac{5}{16n\sqrt{n}}Z_2^3 + O_p(n^{-2})\right\}
\end{aligned} \tag{3.4}$$

식 (3.4)로 부터 통계량 T의 처음 4차 까지의 적률은 다음과 같이 주어짐을 보일 수 있다.

$$\begin{aligned}
\mu_1'(T) &= E(T) = -\frac{\rho_3}{2\sqrt{n}} + O(n^{-3/2}) \\
\mu_2'(T) &= E(T^2) = 1 + \frac{(2\rho_3^2+3)}{n} + O(n^{-2}) \\
\mu_3'(T) &= E(T^3) = -\frac{7\rho_3}{2\sqrt{n}} + O(n^{-3/2}) \\
\mu_4'(T) &= E(T^4) = 3 + \frac{(28\rho_3^2 - 12\rho_4 + 24)}{n} + O(n^{-2})
\end{aligned} \tag{3.5}$$

위 식에서 ρ_3와 ρ_4는 다음의 식으로 정의되는 양이다.

$$\rho_3 = E(W^3), \qquad \rho_4 = E(W^4) - 3 \tag{3.6}$$

따라서, 안부점 근사식에 사용될 처음 4차 까지의 누율들은 적률과 누율의 관계식 (Kendll과 Stuart(1977) 또는 McCullagh(1987))으로 부터 다음과 같이 주어진다.

$$\begin{aligned}
\kappa_1(T) &= -\frac{\rho_3}{2\sqrt{n}} + O(n^{-3/2}) \\
\kappa_2(T) &= 1 + \frac{(7\rho_3^2+12)}{n} + O(n^{-2}) \\
\kappa_3(T) &= -\frac{2\rho_3}{n} + O(n^{-3/2}) \\
\kappa_4(T) &= \frac{(12\rho_3^2 - 2\rho_4 + 6)}{n} + O(n^{-2})
\end{aligned} \tag{3.7}$$

본 논문에서는 2.2절에서 제시한 안부점 근사식을 사용하여 스튜던트화 평균의 분포함수에 대한 근사를 실시하고자 한다. 특히, 스튜던트화 평균의 경우처럼 평균벡터의 함수로 표현되는, 즉 $f(\bar{X}_1, \cdots \bar{X}_p)$의 형태를 취하는, 통계량들에 대해서는 본 논문에서와 같이 누율들에 대한 근사가 이론적으로 가능하며 이들 통계량에 대한 Edgeworth 근사 또한 타당성(validity)을 가지는 것이 밝혀져 있다. (Hall(1992)참고)

한편, 통계량 T의 분포함수에 대한 정규근사는

$$Pr\{T \le x\} \approx \Phi(x) \tag{3.8}$$

이고, Edgeworth 근사식은 다음과 같이 주어짐을 보일 수 있다.

$$Pr\{T \le x\} \approx \Phi(x) + \frac{1}{\sqrt{n}} p_1(x)\phi(x) + \frac{1}{n} p_2(x)\phi(x) \tag{3.9}$$

여기서, $\Phi(x)$와 $\phi(x)$는 각각 표준정규분포의 분포함수와 밀도함수이고, $p_1(x)$와 $p_2(x)$는 다음의 식으로 주어진다.

$$\begin{aligned}
p_1(x) &= \frac{\rho_3}{6}(2x^2+1) \\
p_2(x) &= x\left\{\frac{\rho_4}{12}(x^3-3) - \frac{\rho_3^2}{18}(x^4+2x^2-3) - \frac{1}{4}(x^2+3)\right\}
\end{aligned}$$

4. 모의실험 및 결과

4.1. 표본분산의 경우

모집단의 분포가 표준지수분포(standard exponential distribution), 오염정규분포(contaminated normal distribution)를 따르는 경우에 대하여 살펴보기로 하자. 확률밀도함수가 $f(x) = e^{-x}$ $(x > 0)$으로 주어지는 표준지수분포의 경우 r차 누율은 다음과 같다.

$$\kappa_r = (r-1)! \qquad (r = 1, 2, \cdots) \tag{4.1}$$

확률밀도함수가

$$w_1\phi(x; \mu_1, \sigma_1) + w_2\phi(x; \mu_2, \sigma_2) \quad (w_1 + w_2 = 1)$$

으로 정의되는 오염정규분포의 경우, r차 적률(moments)은 다음과 같다.

$$\mu_r' = \sum_{j=0}^{[r/2]} \frac{r!}{(2j)!(r-2j)!} E(U^{2j})\{w_1\mu_1^{r-2j}\sigma_1^{2j} + (1-w_1)\mu_2^{r-2j}\sigma_2^{2j}\} \quad (r = 1, 2, \cdots) \tag{4.2}$$

위 식에서 기호 $[a]$은 a값을 넘지않는 최대의 정수를 나타낸다.

이상에서 정의된 각 분포의 누율(cumulant) 또는 적률(moment)의 정보로 부터 식 (3.2)의 표본분산 S_n^2에 대한 처음 4차 까지의 누율을 쉽게 계산할 수 있다. 아래의 표 4.1과 표 4.2는 S_n^2의 분포함수에 대한 정규근사와 안부점 근사의 결과를 비교한 것이다. 표 4.1은 모집단이 표준지수분포(standard exponential distribution)인 경우이고, 표 4.2는 오염정규분포 $0.2\phi(x, 0, 1) + 0.8\phi(x, 0, 1/2)$에 대한 결과이다. 각 모집단으로 부터 표본의 크기는 $n = 10, 30$인 경우에 대해서 결과를 제시 하였으며, 모든 결과에서 제시된 정확한값(EXA)은 각 분포로부터 500만번 시뮬레이션(simulation)을 통해 구한 것이고, NOR과 SAD는 각각 정규근사와 안부점 근사를 사용한 결과이다. 이 경우 안부점 근사의 정확도 (precision)는 전반적으로 정규근사에 비해 뛰어나며, 특히 우측 꼬리부분(right tail)의 확률 영역에서 뛰어난 정확도를 유지하므로 정밀한 추론이 요구되는 많은 통계적 문제에 효율적으로 사용될 수 있다.

4.2. 스튜던트화 평균(STUDENTIZED MEAN)의 경우

모집단의 분포가 오염정규분포(contaminated normal distribution), 균일분포(uniform distribution) 인 경우에 대하여 알아보자. 먼저 모집단이 오염정규분포 $0.2N(0,1)+0.8N(0,1/2)$를 따르는 경우, 모집단의 3, 4차 표준화된 누율(standardized cumulants)은

$$\rho_3 = 0, \qquad \rho_4 = 1/3$$

이고, 균일분포 $U(0,1)$의 경우는

$$\rho_3 = 0, \qquad \rho_4 = -6/5$$

표 4.1: 표준지수분포하의 표본분산의 분포함수

N=10				N=30			
X	EXA	NOR	SAD	X	EXA	NOR	SAD
.2	.063	.185	.114	.2	.001	.060	.046
.4	.229	.251	.141	.4	.052	.122	.080
.6	.398	.327	.181	.6	.209	.219	.141
.8	.537	.411	.341	.8	.413	.349	.301
1.0	.644	.500	.500	1.0	.595	.500	.500
1.2	.726	.588	.750	1.2	.732	.650	.731
1.4	.787	.672	.776	1.4	.825	.780	.807
1.6	.833	.748	.803	1.6	.887	.877	.864
1.8	.868	.814	.829	1.8	.927	.939	.907
2.0	.895	.868	.854	2.0	.952	.973	.938
2.2	.916	.910	.876	2.2	.969	.989	.959
2.4	.932	.941	.895	2.4	.979	.996	.973
2.6	.945	.963	.912	2.6	.986	.999	.983
2.8	.955	.977	.927	2.8	.990	.999	.989
3.0	.963	.987	.939	3.0	.993	.999	.993
3.2	.969	.993	.950	3.2	.995	1.000	.996
3.4	.975	.996	.959	3.4	.996	1.000	.997
3.6	.979	.998	.966	3.6	.997	1.000	.998
3.8	.982	.999	.972	3.8	.998	1.000	.999
4.0	.985	.999	.977	4.0	.998	1.000	.999

표 4.2: 오염정규분포하의 표본분산의 분포함수

N=10				N=30			
X	EXA	NOR	SAD	X	EXA	NOR	SAD
2.0	.062	.096	.134	3.4	.075	.090	.104
2.8	.157	.158	.222	3.8	.137	.141	.168
3.6	.280	.243	.341	4.2	.220	.208	.250
4.4	.412	.346	.463	4.6	.317	.290	.343
5.2	.536	.463	.546	5.0	.422	.386	.437
6.0	.645	.583	.622	5.4	.527	.489	.529
6.8	.734	.696	.700	5.8	.626	.593	.617
7.6	.805	.793	.771	6.2	.713	.690	.699
8.4	.860	.868	.830	6.6	.786	.776	.770
9.2	.901	.922	.877	7.0	.845	.846	.830
10.0	.931	.957	.913	7.4	.890	.900	.877
10.8	.952	.978	.940	7.8	.924	.939	.914
11.6	.968	.990	.959	8.2	.949	.964	.941
12.4	.978	.995	.973	8.6	.966	.980	.961
13.2	.985	.998	.982	9.0	.978	.990	.974
14.0	.990	.999	.988	9.4	.986	.995	.983
14.8	.994	.999	.992	9.8	.991	.997	.989
15.6	.996	.999	.995	10.2	.994	.999	.993

으로 주어진다.

표 4.3과 표 4.4는 각 모집단으로 부터 표본의 크기가 $n = 5, 30$인 경우에 대하여 스튜던트화 평균 T 의 분포함수의 값을 구한 것이다. 표 4.3은 모집단이 오염정규분포의 경우이며, 표 4.4는 균일분포를 따르는 경우에 대한 결과이다. 각 표에서 제시된 정확한 값(EXA)들은 모집단의 분포로 부터 5백만번 시뮬레이션(simulation)한 결과이며, 정규근사(NOR)와 Edgeworth근사(EDGE)는 각각 식 (3.8)과 (3.9)로 부터 구해진 값이며, 안부점 근사(SAD)는 2.2절에서 제시한 방법을 사용하였다. 각 표의 정확한 값(EXA)은 시뮬레이션을 통해 구해진 정확한 확률값이며, 나머지는 각각의 근사식으로 부터 다음의 식을 통해 얻어진 값이다.

$$(\text{근사값} - \text{정확한 값}) \times 10000$$

이상의 결과로 부터 다음과 같은 사실을 확인할 수 있다. 모집단의 분포에 관계없이 안부점 근사의 결과는 정규근사 보다 뛰어나며, 특히 표본의 크기가 작은 경우 정규근사는 상당히 부정확한 반면 안부점 근사는 높은 정확도를 유지한다. 또한 2개항을 사용한 Edgeworth 근사와는 서로 비슷한 정확도(precision)를 유지하나, 통계적 추론에 자주 사용되는 꼬리부분(tail part)의 영역에서는 안부점 근사의 정확도가 뛰어남을 알 수 있다.

표 4.3: 오염정규분포하의 스튜던트화 평균의 분포함수

	N=5					N=30			
x	EXA	NOR	EDGE	SAD	x	EXA	NOR	EDGE	SAD
-6.0	.0028	-28	-28	-28	-6.0	.0000	0	0	0
-5.5	.0039	-39	-39	-39	-5.5	.0000	0	0	0
-5.0	.0055	-55	-55	-54	-5.0	.0000	0	0	0
-4.5	.0079	-79	-79	-74	-4.5	.0000	0	0	0
-4.0	.0115	-115	-110	-99	-4.0	.0002	-2	-1	-1
-3.5	.0175	-173	-151	-130	-3.5	.0008	-6	-3	-2
-3.0	.0274	-261	-186	-161	-3.0	.0029	-16	-3	-3
-2.5	.0445	-383	-189	-181	-2.5	.0097	-35	-3	-2
-2.0	.0741	-514	-142	-169	-2.0	.0286	-59	3	3
-1.5	.1255	-587	-69	-117	-1.5	.0738	-70	16	17
-1.0	.2110	-524	-13	-43	-1.0	.1636	-50	35	38
-.5	.3393	-308	5	1	-.5	.3087	-2	50	52
.0	.5001	-1	-1	-1	.0	.4949	51	51	51
.5	.6609	305	-8	-4	.5	.6817	97	45	43
1.0	.7889	524	13	43	1.0	.8297	116	31	28
1.5	.8744	587	69	117	1.5	.9226	105	19	18
2.0	.9257	515	143	170	2.0	.9698	74	12	12
2.5	.9553	384	190	182	2.5	.9897	40	8	7
3.0	.9723	263	188	163	3.0	.9968	18	5	5
3.5	.9823	174	152	133	3.5	.9990	7	4	3
4.0	.9882	117	112	101	4.0	.9997	2	1	1
4.5	.9920	79	79	74	4.5	.9999	0	0	0
5.0	.9944	56	55	54	5.0	.9999	1	0	0
5.5	.9960	40	39	39	5.5	.9999	1	1	0
6.0	.9970	30	30	29	6.0	.9999	1	1	1

표 4.4: 균일분포하의 스튜던트화 평균의 분포함수

N=5					N=30				
x	EXA	NOR	EDGE	SAD	x	EXA	NOR	EDGE	SAD
-6.0	.0066	-66	-66	-66	-6.0	.0000	0	0	0
-5.5	.0083	-83	-83	-82	-5.5	.0000	0	0	0
-5.0	.0106	-106	-106	-104	-5.0	.0000	0	0	0
-4.5	.0138	-138	-137	-130	-4.5	.0001	-1	-1	-1
-4.0	.0185	-185	-179	-164	-4.0	.0003	-3	-2	-2
-3.5	.0251	-249	-220	-198	-3.5	.0011	-9	-4	-4
-3.0	.0351	-338	-242	-226	-3.0	.0035	-22	-6	-6
-2.5	.0507	-445	-214	-230	-2.5	.0105	-43	-5	-5
-2.0	.0763	-536	-136	-184	-2.0	.0296	-69	-2	-2
-1.5	.1209	-541	-61	-82	-1.5	.0748	-80	0	5
-1.0	.2004	-418	-31	28	-1.0	.1652	-66	-1	8
-.5	.3291	-206	-17	66	-.5	.3119	-34	-3	5
.0	.5002	-2	-2	-2	.0	.5005	-5	-5	-5
.5	.6712	202	13	-70	.5	.6885	29	-2	-10
1.0	.7997	416	29	-30	1.0	.8349	64	-1	-10
1.5	.8791	540	60	81	1.5	.9251	80	0	-5
2.0	.9235	537	137	185	2.0	.9702	70	3	3
2.5	.9489	448	217	233	2.5	.9892	45	7	7
3.0	.9645	341	245	229	3.0	.9964	22	6	6
3.5	.9746	251	222	200	3.5	.9988	9	4	4
4.0	.9813	186	180	165	4.0	.9996	3	2	2
4.5	.9859	140	139	132	4.5	.9998	1	1	1
5.0	.9892	108	107	105	5.0	.9999	1	0	0
5.5	.9915	85	84	83	5.5	.9999	1	1	0
6.0	.9933	67	67	66	6.0	.9999	1	1	1

5. 결론

본 논문에서는 표본평균의 분포함수에 대한 안부점 근사의 확장으로 일반적 통계량의 분포함수에 대한 안부점 근사법을 제안하였다. 밀도함수의 근사에 사용된 Easton과 Ronchetti(1986)의 방법을 분포함수의 근사에 직접 적용함으로써 복잡한 수치적 적분이 요구되는 그들의 방법에 대한 대안을 제시하였다. 이에 대한 통계적 응용으로 표본분산 및 스튜던트화 평균의 근사에 적용한 결과, 제안된 방법의 정확도가 상당히 뛰어남을 확인하였다. 대표본의 경우는 물론 소표본의 경우에도 높은 정확도를 유지하며 분포의 꼬리부분에서도 근사의 정도가 뛰어난 이 방법은 사용하기에도 편리하여 통계학의 전반에 걸쳐 제시되는 각종 통계량의 분포함수에 대한 근사에 응용 될 수 있다.

부록

식 (1.12)에서 식 (1.13)이 유도되는 과정을 정리하면 다음과 같다.

$$
\begin{aligned}
P(V_n \geq v) &= \frac{1}{2\pi i}\int_{\tau-i\infty}^{\tau+i\infty} e^{\frac{1}{2}s^2-\sqrt{n}s\hat{\eta}}\frac{1}{s}ds \\
&\quad + \frac{1}{2\pi i}\int_{\hat{\eta}-i\infty}^{\hat{\eta}+i\infty} e^{\frac{n}{2}(\eta-\hat{\eta})^2} g(\eta)d\eta \cdot e^{n[R_n(t_0)-t_0 v]} \\
&\quad (\text{여기서}, g(\eta) = \frac{1}{t}\frac{dt}{d\eta} - \frac{1}{\eta}\text{이다.}) \\
&= \bar{\Phi}(\sqrt{n}\hat{\eta}) + \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{\frac{n}{2}y^2}\left\{ g(\hat{\eta}) + g'(\hat{\eta})iy + \frac{g''(\hat{\eta})}{2!}i^2y^2 + \cdots \right\} dy \cdot \\
&\quad e^{n[R_n(t_0)-t_0 v]} \qquad (*) \\
&= \bar{\Phi}(\sqrt{n}\hat{\eta}) + \phi(\sqrt{n}\hat{\eta})\frac{1}{\sqrt{n}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}\left\{ g(\hat{\eta}) - \frac{1}{n}g''(\hat{\eta})\frac{y^2}{2} + \cdots \right\} dy (**) \\
&= \bar{\Phi}(\sqrt{n}\hat{\eta}) + \phi(\sqrt{n}\hat{\eta})\left\{ \frac{g(\hat{\eta})}{\sqrt{n}} - \frac{1}{n\sqrt{n}}\frac{g''(\hat{\eta})}{2} + \cdots \right\} dy
\end{aligned}
$$

여기서, $g(\hat{\eta}) = \frac{1}{t_0}(\frac{dt}{d\eta})_{\eta=\hat{\eta}} - \frac{1}{\hat{\eta}} = \frac{1}{t_0\sqrt{R''(t_0)}} - \frac{1}{\hat{\eta}}$이고 $g'(\cdot)$, $g''(\cdot)$ 및 $g^{(r)}(\cdot)$은 각각 1, 2, r차 미분을 의미한다. 또한, $\bar{\Phi}(\cdot) = 1 - \Phi(\cdot)$이며 $\Phi(\cdot)$와 $\phi(\cdot)$은 본문에서의 정의와 동일하다. 따라서 다음의 관계가 성립한다.

$$
\begin{aligned}
Pr\{V_n \leq v\} &= 1 - Pr\{V_n \geq v\} \\
&= \Phi(w) + \phi(w)\left\{ \frac{1}{w} - \frac{1}{\zeta} + O(n^{-3/2}) \right\}
\end{aligned}
$$

여기서 w와 ζ는 식 (2.14)및 (2.15)와 동일하다. 위의 식 (*)에서 식 (**)의 과정은 본문에서 식 (2.9)의 좌우변의 최소값이 일치하는 조건을 찾으면 $R_n(t_0) - t_0 v = \hat{\eta}^2/2$이 되기 때문이다.

참고문헌

[1] Barndorff-Nielsen, O. E. and Cox, D. R. (1979). Edgeworth and saddlepoint approximations with statistical applications. *Journal of Royal Statistical Society.* Series B, 41(3),279-312.

[2] Daniels, H. E. (1954). Saddlepoint approximations in statistics. *Annals of Mathematical Statistics.* 25, 631-650.

[3] Daniels, H. E. (1987). Tail probability approximation. *International Statistical Review.* 55(1), 37-48.

[4] Davison, A. C. and Hinkley, D. V. (1988). Saddlepoint approximations in resampling methods. *Biometrika.* 75(3), 417-431.

[5] Easton, G. S. and Ronchetti, E. (1986). General saddlepoint approximations with applications to L statistics. *Journal of the American Statistical Association.* 81(394),420-430.

[6] Field , C. A. and Ronchetti, E. (1990). *Small Sample Asymptotics.* Institute of Mathematical Statistics, Hayward, California.

[7] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York.

[8] Kendall, M. G. and Stuart, A. (1997). *The Advanced Theory of Statistics.* Vol 1, London, Griffin.

[9] Lugannani, R. and Rice, S. (1980). Saddlepoint approximation for the distribution of the sum of independent random variables. *Advances in Applied Probability.* 12, 475-490.

[10] McCullagh, P. (1987). *Tensor Methods in Statistics.* Chapman and Hall, London, New York.

[11] Reid, N. (1998). Saddlepoint methods and statistical inference. *Statistical Science.* 3(2), 213-238.

[12] Srivastava, M. S. and Yau, W. K. (1989). Tail probability approximations of general statistics. *Technical Report.* No. 88-38, University of Pittsburgh, Pittsburgh.

[13] Wang, S. (1990). Saddlepoint approximations in resampling analysis. *Annals of the Institute of Statistical Mathematics.* 42(1), 115-131.

[14] Wood, A. T. A., Booth, J. G., and Butler, R. W. (1993). Saddlepoint approximation to the CDF of some statistics with nonnormal limiting distribution. *Journal of the American Statistical Association.* 88, 680-686.

[1997년 8월 접수, 1998년 3월 최종수정]

Saddlepoint Approximation to the Distribution of General Statistic

Jonghwa Na [1)]

ABSTRACT

Saddlepoint approximation to the distribution function of sample mean(Daniels, 1987) is extended to the case of general statistic in this paper. The suggested approximation methods are applied to derive the approximations to the distributions of some statistics, including sample variance and studentized mean. Some comparisons with other methods show that the suggested approximations are very accurate for moderate or small sample sizes. Even in extreme tail the accuracies are also maintained.

1) Assistant Profrssor, Department of Statistics, Chungbuk National University, Cheongju 361-763, Korea.