2000 Proceedings of the Autumn Conference, Korean Statistical Society

지수분포의 검정을 위한 수정된 W-통계량[1)]

김 남현[2)]

요 약

Shapiro와 Wilk(1972)는 위치모수와 척도모수가 미지인 경우 지수분포의 검정통계량을 제안하였다. 그것은 척도모수의 일반화 최소제곱추정량과 표본분산의 비로 구성되었다. 그러나 이 검정통계량은 일치성을 갖지 않는다. 본 논문에서는 척도모수의 두 개의 점근유효추정량으로 구성된 통계량을 고려하고 이의 극한분포를 구하였다. 또한 두 개의 통계량의 검정력을 비교한 결과 제안된 통계량이 변동계수가 1보다 크거나 같은 분포에서 더 좋은 검정력을 가짐을 볼 수 있었다.

주요용어 : 지수분포, 적합도검정, 점근유효추정량, Brownian bridges

1. 서론

지수분포의 분포함수

$$F(x;\alpha,\beta)=1-\exp\left(-\frac{x-\alpha}{\beta}\right),\ x>\alpha,\ \beta>0,\ -\infty<\alpha<\infty \tag{1.1}$$

을 $\exp(\alpha,\beta)$를 쓰기로 하자. 또한 $\alpha=0$, $\beta=1$인 표준지수분포 $F(x;0,1)$은 $F_0(x)$로 나타내고 $f_0(x)$를 F_0에 해당하는 확률밀도함수, F_0^{-1}은 F_0의 역함수라고 하자.

$X_1,\cdots,X_n$이 연속확률분포함수 $G(x)$에서의 확률표본이고 이 표본의 순서통계량을 $X_{(1)}\le X_{(2)}\le\cdots\le X_{(n)}$이라고 할 때, $X_1,\cdots,X_n$이 지수분포의 모형에 적합한지, 즉

$$H_0:\ G(x)=\exp(\alpha,\beta) \tag{1.2}$$

을 검정하자. 대립가설 H_A는 $H_A:\ G(x)\neq\exp(\alpha,\beta)$이라고 하자. (1.2)의 귀무가설 H_0를 검정하기 위해 제안된 통계량은 무수히 많다. 이 경우 대부분 α는 기지 또는 0으로, β는 미지로 가정하고 있다. D'Agostino와 Stephens (1986, 4, 5, 10장)에서 이에 대한 전반적인 설명과 참고문헌을 제시하고 있다.

본 논문에서는 모수 α와 β가 모두 미지일 때의 검정법을 고려한다. 이 경우의 검정법은 많이 고려되고 있지 않다. 한 가지 이유는 지수분포의 통계적인 성질을 이용하여 α가 미지인 경우의 검정법 대신 $\alpha=0$인 검정법을 이용할 수 있기 때문이다(**보조정리 1** 참조). 그러나 이러한 방법으로 모수 α를 제거하는 것이 가설 (1.2)를 검정하는데 있어서 α를 추정하는 것보다 더 좋은 검정법을 제공할 수 있는지는 확실하지 않다. Spinelli와 Stephens (1987)은 여러 가지 대립가설 하에서 모의실험(simulation)을 실시하여 두 가지 방법을 비교하였다.

α와 β가 모두 미지인 경우 소개된 대표적인 통계량은 Shapiro와 Wilk (1972)의 W_E-통계

1) 이 논문은 2000년 홍익대학교 학술연구 조성비에 의하여 연구되었음.
2) 121-791 서울시 마포구 상수동 72-1 홍익대학교 기초과학과 조교수.

량이다. 이 통계량은 β의 일반화 최소자승 추정량(generalized least squares estimator)과 표본분산 $s^2=S^2/(n-1)=\sum_{j=1}^{n}(X_j-\overline{X})^2/(n-1)$을 비교하는 것으로 β의 두 추정량의 비에 기초한 통계량이다. W_E-통계량의 형태는

$$W_E=\frac{n(\overline{X}-X_{(1)})^2}{(n-1)S^2} \tag{1.3}$$

이며 이는 양쪽 검정통계량이다. Spinelli와 Stephens (1987)에서 지적한대로 W_E-통계량은 대부분의 대립가설에서 좋은 검정력을 갖는다. 그러나 W_E-통계량의 가장 심각한 단점은 이 검정법이 일치성(consistency)을 갖지 않는다는데 있다.

본 논문에서는 W_E-통계량의 이러한 단점을 보완하기 위한 수정된 형태의 W_E-통계량을 소개하고 이의 통계적인 성질을 알아본다. 또한 모의실험을 통하여 이들 통계량의 검정력을 비교하여 그 특성을 알아보고자 한다.

2. 수정된 W_E-통계량

1절에서 언급한대로 식(1.3)의 W_E-통계량의 가장 큰 단점은 이 통계량에 기초한 검정법이 일치성을 갖지 않는다는 것이다. 다시 말해서 이 검정법의 검정력이 표본의 수가 증가해도 1로 가까이 가지 않는 분포가 존재한다는 말이다. 지수분포의 경우 모집단의 표준편차와 평균의 비인 변동계수(coefficient of variation)가 $C_V=\sigma/\mu=1$이므로

$$nW_E \xrightarrow{p} 1/C_V^2=1, \quad n\to\infty \text{ 일 때} \tag{2.1}$$

이다. 그러므로 $C_V=1$인 다른 분포에서도 nW_E는 1로 수렴할 것이다. 베타분포 $B(a,b)$에서 $a<1$, $b=a(a+1)/(1-a)$인 경우, 예를 들면 (a,b)가 $(1/4,5/12)$일 때를 고려할 수 있다(Spinelli and Stephens(1987)).

일반적으로 W_E와 같이 β의 두 추정량의 비로 구성된 통계량에 기초한 검정법의 일치성은 두 추정량의 통계적인 성질에 달려있다. 직관적으로 W_E에 기초한 검정법이 일치성을 갖지 않는 이유는 분모의 표본분산이 지수분포의 경우 점근유효추정량(asymptotically efficient estimator)이 아니기 때문이다. 실제로

$$\sqrt{n}(s^2-\beta^2) \xrightarrow{d} N(0,8\beta^4) \tag{2.2}$$

이고 피셔정보(Fisher information)는 $I(\beta^2)=1/4\beta^4$이다(Ferguson(1996, 7절, 19절)). 따라서 분모의 추정량을 β^2의 점근유효추정량으로 대치하면 W_E-통계량의 단점을 보완할 수 있을 것이라 기대된다. β^2의 추정량으로

$$L_n=\frac{1}{n-1}\sum_{j=2}^{n}(X_{(j)}-X_{(1)})^2/v_{jn},\ v_{jn}=F_0^{-1}\left(\frac{j}{n+1}\right)=-\log\left(1-\frac{j}{n+1}\right) \tag{2.3}$$

을 고려하면 L_n은 β^2의 점근유효추정량(정리 2 참조)이므로 수정된 W_E-통계량으로 N_E-통계량,

$$N_E=\frac{n(\overline{X}-X_{(1)})^2}{(n-1)^2L_n}=\frac{n(\overline{X}-X_{(1)})^2}{(n-1)\sum_{j=2}^{n}(X_{(j)}-X_{(1)})^2/v_{jn}} \tag{2.4}$$

을 제안한다. de Wet과 Venter(1973)는 일반적인 척도모수 분포모임(scale parameter family of distributions)에 대해서 식(2.4)와 유사한 형태의 통계량을 제안하고 이의 점근분포(asymptotic distribution)을 i.i.d. 확률변수의 무한급수의 형태로 구하였다. 이에 따르면 일반적인 분포에서 $(n-1)N_E \le 1+o_p(1)$ 이므로 N_E-통계량은 적당한 c에 대해서 $N_E<c$일 때 H_0를 기각하는 것이 합리적이다. N_E-통계량은 de Wet과 Venter(1973)의 통계량을 β뿐만 아니라 α도 미지인 경우의 지수분포에 적용한 것이라고 생각할 수 있다.

$Z_{(1)},\ldots,Z_{(n)}$을 $\exp(0,1)$에서의 순서통계량이라고 할 때, 식(2.3)의 L_n에서 v_{jn}은 $\widetilde{v_{jn}}=E(Z_{(j)})$의 근삿값이므로 v_{jn}을 $\widetilde{v_{jn}}=\sum_{i=1}^{j}\frac{1}{n-i+1}$으로 대치하면 β^2의 추정량으로

$$\widetilde{L_n}=\frac{1}{n-1}\sum_{j=2}^{n}(X_{(j)}-X_{(1)})^2/\widetilde{v_{jn}} \tag{2.5}$$

을 고려할 수 있고, 수정된 W_E-통계량으로 $\widetilde{N_E}$-통계량,

$$\widetilde{N_E}=\frac{n(\overline{X}-X_{(1)})^2}{(n-1)^2\widetilde{L_n}} \tag{2.6}$$

을 제안할 수 있다.

보조정리 1. $X_{(1)},\ldots,X_{(n)}$이 $\exp(\alpha,\beta)$로 부터의 크기 n인 순서통계량이면, $Y_{(i)}=X_{(i+1)}-X_{(1)}$, $i=1,\ldots,n-1$은 $\exp(0,\beta)$로 부터의 크기 $n-1$인 순서통계량이다.

정리 1. 식(2.4)의 N_E와 식(2.6)의 $\widetilde{N_E}$는 위치, 척도 불변인 통계량이다.

증명. $X_i=\alpha+\beta Z_i$, $i=1,\ldots,n$을 대입해 보면 자명하다. □

정리 2. 식(2.4)의 N_E와 식(2.6)의 $\widetilde{N_E}$의 분모, 분자인 $(\overline{X}-X_{(1)})^2$, L_n, $\widetilde{L_n}$는 $X_1,\ldots,X_n$이 $\exp(\alpha,\beta)$에서의 표본일 때 β^2의 점근유효추정량이다. 즉

$$\sqrt{n}((\overline{X}-X_{(1)})^2-\beta^2)\xrightarrow{d}N(0,4\beta^2) \tag{2.7}$$

$$\sqrt{n}(L_n-\beta^2)\xrightarrow{d}N(0,4\beta^4) \tag{2.8}$$

이고, $\widetilde{L_n}$에 대해서도 식(2.8)이 성립한다.

증명. Chernoff, Gastwirth, and Johns(1967)의 정리를 적용한다. □

G_n을 순서통계량 $X_{(1)},\ldots,X_{(n)}$의 경험적 분포함수(empirical distribution fucntion), G_n^{-1}를 그것의 역함수, $v(t)=F_0^{-1}(t)=-\log(1-t)$라고 하자. 또한 $\rho_n(t)$를

$$\rho_n(t):=\sqrt{n}(1-t)(G_n^{-1}(t)-v(t)) \tag{2.9}$$

로 정의된 quantile process라고 하자.

정리 3. (Csörg˝o 과 Horváth(1993)의 정리 6.2.1) 적절한 확률공간에서

$$n^{(1/2)-\nu} \sup_{1/(n+1)\le t\le n/(n+1)} \frac{|\rho_n(t)-B_n(t)|}{(t(1-t))^{\nu}} = \begin{cases} O_p(\log n), & \nu=0 \\ O_p(1), & 0<\nu\le 1/2 \end{cases}$$

을 만족하는 Brownian bridges $\{B_n(t), 0\le t\le 1\}_n$가 존재한다.

정리 4. 식(2.4)의 N_E는 극한분포

$$n\left(\frac{1}{N_E}-1-a_n\right) \xrightarrow{d} \int_0^1 \frac{B^2(t)-t(1-t)}{(1-t)^2 v(t)}\,dt-\left(\int_0^1 \frac{B(t)}{1-t}\,dt\right)^2 \tag{2.10}$$

을 갖는다. 여기서

$$a_n=\frac{1}{n}\int_{1/(n+1)}^{n/(n+1)} \frac{t(1-t)}{(1-t)^2 v(t)}\,dt$$

이다.

증명. del Barrio, Cuesta, Matrán와 Rodríguez(1999)의 방법과 유사하다.

$Z_1,\ldots,Z_n$이 $\exp(0,1)$에서의 표본일 때

$$R_n:=\left(\frac{1}{n}\sum_{i=1}^n Z_i^2/v_{jn}\right)/\,\overline{Z}^2-1,\quad R_n^*:=\overline{Z}^2 R_n=\frac{1}{n}\sum_{i=1}^n Z_i^2/v_{jn}-\overline{Z}^2$$

이라고 하면, **보조정리 1**과 **정리 1**로부터 nR_n-a_n에 대해서 위의 결과가 성립함을 보이면 충분하다. 또한 $n(R_n^*-a_n)=O_p(1)$ 이면

$$n(R_n-a_n)-n(R_n^*-a_n)=(1-\overline{Z}^2)/\,\overline{Z}^2 nR_n^*=O_p(1)\sqrt{n}(R_n^*-a_n+a_n)\xrightarrow{p}0$$

이므로 $n(R_n^*-a_n)$이 식(2.10)의 우변의 극한분포를 갖음을 보이면 된다.

$$\begin{aligned} nR_n^* &= \int_{1/(n+1)}^{n/(n+1)} \frac{\left(\sqrt{n}(G_n^{-1}(t)-v(t))\right)^2}{v(t)}\,dt-\left(\int_{1/(n+1)}^{n/(n+1)}\sqrt{n}(G_n^{-1}(t)-v(t))dt\right)^2+O\left(\frac{\log n}{n}\right) \\ &= \int_{1/(n+1)}^{n/(n+1)} \frac{\rho_n^2(t)}{(1-t)^2 v(t)}\,dt-\left(\int_{1/(n+1)}^{n/(n+1)}\frac{\rho_n(t)}{(1-t)}\,dt\right)^2+O\left(\frac{\log n}{n}\right) \end{aligned}$$

이므로 **정리 3**을 이용하면 결과를 얻을 수 있다. 또한

$$A_n:=\int_{1/(n+1)}^{n/(n+1)} \frac{B^2(t)-t(1-t)}{(1-t)^2 v(t)}\,dt$$

라고 하면 $EA_n^2<\infty$이므로 식(2.10)의 우변의 첫째항은 A_n의 L_2-극한으로 정의된다. □

3. 모의실험결과 및 결론

2절에서 소개한 수정된 W_E-통계량인 식(2.4), (2.6)의 N_E, $\widehat{N_E}$의 효율성을 소표본일 때 조사하기 위해서, 모의실험(simulation)을 행하였다. 이 때 표본의 수는 $N=5000$을 사용하였다. 대립가설의 분포로는 웨이블, 감마분포 등을 포함하여 여러 가지 분포가 고려되었다. $g(x)$가 확률밀도함수, $G(x)$가 확률분포함수일 때 고장률 또는 위험률(failure rate, hazard rate)을 $h(x)=g(x)/(1-G(x))$, $x>0$이라고 하면, 이 중 $\chi^2(4)$, $U(0,1)$, Weib(1.5), $(1/2)N$은 증가고장률(increasing failure rate : IFR) 분포이고, $\chi^2(1)$, Weib(0.8),

$\log\mathrm{norm}(1)$, $(1/2)C$는 감소고장율(decreasing failure rate : DFR) 분포이다. IFR 분포는 변동계수 C_V가 $C_V<1$이고 DFR 분포는 $C_V>1$이다.

<표 1>에서는 세 가지 통계량의 검정력을 유의수준 $100\alpha=10\%$에서 비교하고 있다. 이를 보면, 우선 특이할만한 사실은 베타분포 $B(1/4,5/12)$에서 Shapiro와 Wilk(1972)의 W_E-통계량은 표본크기가 커질수록 검정력이 점점 감소하나 제안된 N_E, $\widehat{N_E}$-통계량은 이 경우 검정력이 매우 우수함을 볼 수 있다.

또한 대립가설이 DFR 분포일 경우에는 N_E, $\widehat{N_E}$-통계량이 W_E-통계량보다 대체적으로 ($\mathrm{Weib}(0.8)$ 제외) 우수한 검정력을 보여주고 있고 N_E와 $\widehat{N_E}$는 거의 비슷한 양상을 나타낸다. 그러나 대립가설이 IFR 분포일 경우에는 W_E-통계량이 특히 표본크기가 작을 때 ($n=10,20$) N_E, $\widehat{N_E}$ 보다 훨씬 우수한 검정력을 보여주고 $\widehat{N_E}$가 N_E보다 고려된 모든 경우에 좋은 검정력을 보여준다.

결론적으로 소표본이고 대립가설이 $C_V\geq 1$인 분포일 경우에는 W_E-통계량보다는 제안된 N_E, $\widehat{N_E}$-통계량이 더 효율적이며 소표본이고 $C_V<1$인 대립가설에서는 W_E가 $\widehat{N_E}$, N_E 보다 더 효율적이라고 할 수 있다.

N_E와 $\widehat{N_E}$를 비교하면, 각각이 반 정도의 대립가설에서 다른 통계량보다 우수하지만, N_E가 우수할 경우에는 검정력의 차이가 매우 작고, 반면 $\widehat{N_E}$가 우수할 경우에는 그 차이가 대체적으로 크다고 할 수 있다. 따라서 전체적으로 $\widehat{N_E}$가 N_E보다 우수하다는 결론을 내릴 수 있다.

참고문헌

[1] Chernoff, H., Gastwirth, J. L. and Johns, M. V. (1967). "Asymptotic distribution of linear combinations of functions of order statistics with applications to estimation," *Annals of Mathematical Statistics*, 38, 52-73.

[2] Csörg˝o, M. and Horváth, L. (1993). *Weighted Approximations in Probability and Statistics*. Wiley, New York.

[3] D'Agostino, R. B. and Stephens, M. A. (1986). *Goodness-of-fit Techniques*, Marcel Dekker, New York.

[4] de Wet, T. and Venter, J. H. (1973). "A goodness of fit test for a scale parameter family of distributions," *South African Statistical Journal*, 7, 35-46.

[5] del Barrio, E., Cuesta, J. A., Matrán, C. and Rodríguez, J. M. (1999). "Tests of Goodness of fit based on the L_2-Wasserstein distance," *The Annals of Statistics*, 27, 1230-1239.

[6] Ferguson, T. S. (1996). *A course in large sample theory*, Chapman & Hall.

[7] Shapiro, S. S. and Wilk, M. B. (1972). "An analysis of variance test for the exponential distribution (complete samples)," *Technometrics*, 14, 355-370.

[8] Spinelli, J. J. and Stephens, M. A. (1987). "Tests for exponentiality when origin and scale parameters are unknown," *Technometrics*, 29, 471-476.

대립가설	표본크기(n)	W_E	N_E	$\widehat{N_E}$
$B(1/4,5/12)$	10	14.92	33.46	47.94
	20	10.00	69.94	83.64
	50	6.82	99.82	100.00
	100	5.34	100.00	100.00
$\chi^2(4)$	10	18.92	2.24	5.02
	20	32.36	4.84	11.36
	50	68.52	36.58	53.00
	100	93.36	85.60	93.00
$U(0,1)$	10	48.66	4.54	17.24
	20	84.60	34.00	59.90
	50	99.94	97.90	99.92
	100	100.00	100.00	100.00
Weib(1.5)	10	22.34	1.76	4.64
	20	42.18	6.24	16.36
	50	85.12	47.54	68.36
	100	99.28	93.44	98.10
$(1/2)N$	10	16.50	1.62	4.68
	20	32.12	3.86	10.12
	50	68.36	20.60	41.60
	100	94.20	62.22	80.96
$\chi^2(1)$	10	25.64	38.26	39.98
	20	40.82	63.02	60.34
	50	75.00	92.10	92.50
	100	94.66	99.60	99.74
Weib(0.8)	10	25.42	25.54	25.26
	20	39.14	38.76	34.64
	50	74.28	61.22	57.96
	100	94.78	80.86	80.26
lognorm(1)	10	19.98	23.88	24.06
	20	29.14	34.34	32.98
	50	44.76	55.62	49.48
	100	60.80	72.92	69.58
$(1/2)C$	10	47.46	56.96	55.94
	20	73.26	80.88	78.48
	50	96.30	97.86	97.06
	100	99.84	99.94	99.92

<표 1> W_E, N_E, $\widehat{N_E}$-통계량의 검정력 비교

Weib(m) : 확률밀도함수 $g(x)=mx^{m-1}e^{-x^m}$, $x>0$

lognorm(m) : $g(x)=C\exp[-(\log x)^2/(2m^2)]$, $x>0$

$(1/2)N$: $Y\sim N(0,1)$일 때 $X=|Y|$의 분포

$(1/2)C$는 Y가 중앙값이 0인 코쉬분포를 따를 때 $X=|Y|$의 분포