韓國經營科學會誌
第31권 第1號
2006年 3月

시뮬레이션 입력 모형화 : 확률분포 모수 추정을 위한 표본크기 결정

박 성 민*

Simulation Input Modeling : Sample Size Determination for Parameter Estimation of Probability Distributions

Sungmin Park*

■ Abstract ■

In simulation input modeling, it is important to identify a probability distribution to represent the input process of interest. In this paper, an appropriate sample size is determined for parameter estimation associated with some typical probability distributions frequently encountered in simulation input modeling. For this purpose, a statistical measure is proposed to evaluate the effect of sample size on the precision as well as the accuracy related to the parameter estimation, square rooted mean square error to parameter ratio. Based on this evaluation measure, this sample size effect can be not only analyzed dimensionlessly against parameter's unit but also scaled regardless of parameter's magnitude. In the Monte Carlo simulation experiments, three continuous and one discrete probability distributions are investigated such as : 1) exponential : 2) gamma : 3) normal : and 4) poisson. The parameter's magnitudes tested are designed in order to represent distinct skewness respectively. Results show that : 1) the evaluation measure drastically improves until the sample size approaches around 200 : 2) up to the sample size about 400, the improvement continues but becomes ineffective : and 3) plots of the evaluation measure have a similar plateau pattern beyond the sample size of 400. A case study with real datasets presents for verifying the experimental results.

Keyword : Simulation Input Modeling, Sample Size, Square Rooted Mean Square Error to Parameter Ratio, Parameter Estimation, Probability Distribution

논문접수일 : 2005년 7월 4일 논문게재확정일 : 2006년 1월 3일
* 천안대학교 경상학부

1. 연구주제

시뮬레이션 입력 모형화(simulation input modeling)는 ; 1) 수집된 데이터 수를 고려하여 히스토그램(histogram) 혹은 quantile-quantile(Q-Q) plot 등을 이용한 그래프 분석 ; 2) 선택된 확률분포 집단의 모수(parameter) 추정 ; 그리고 3) 확률분포에 대한 통계적 적합도검정(goodness-of-fit test) 등을 수행해, 시뮬레이션 입력모형으로 활용될 확률분포를 결정하는 과정으로 요약될 수 있다[3, 10]. 한편, 데이터 수집시, 상당한 시간·자원이 소요될 수 있어, 비용 측면을 고려해 수집 데이터 수가 축소될 가능성도 배제할 수 없다[8]. 하지만, 데이터 수 감소는 ; 1) 그래프 분석의 선명도를 낮춰, 입력모형으로 고려될 수 있는 잠재적 확률분포 집단 선택을 어렵게 할 뿐만 아니라 ; 2) 선택된 확률분포 모수 추정치 정밀도(precision) 및 정확도(accuracy) 저하 등을 초래할 수 있다. 본 논문에서는, '시뮬레이션 입력 모형화를 위해 수집된 데이터 수'를 '표본크기'라 지칭한다.

한편, 표본크기는, 적합도검정과도 밀접한 관련이 있는데 ; 1) 작은 표본크기로는, 적합분포(fitted distribution)와 표본사이의 불일치에 둔감한(insensitive) 검정이 유도될 가능성이 큰 반면 ; 2) 표본크기가 커질수록 검정력(power of test) 또한 증가되어, 통계적 검정으로는 적합한 입력모형 도출이 어려울 수 있다[4, 9]. 특히, 실무자는 그래프 분석을 통해 시각적으로는 적합분포라 판단할지라도, 일단 고려대상 분포가 적합도검정에서 기각되면, 입력모형으로 채택하기를 거부하려는 경향이 있을 수 있다[15].

분포 식별을 위한 히스토그램 작성시 ; 1) 표본크기는 75~100개 이상 ; 2) 계급구간(class interval) 수는 5~20개 ; 그리고 3) 계급구간 수는 표본크기 제곱근에 근사하게 설정될 수 있다[14]. Gross[6]는, 대기행렬 시뮬레이션 입력모형을 갖는 출력 민감도 분석시, 입력모형 형상(shape) 파악을 위한 최소 표본크기로 대략 500개를 제안하지만, 구체적 근거를 제시한 바 없다. 반도체 '수율'(yield) 분석시, 비대칭도 (skewness)가 심한 포아송 분포의 'sigma 수준'별 임계값(critical value) 결정에 필요한 표본크기로서 대략 400개 이상이면 안정적인 것으로 보고된 바 있다[1].

본 연구에서는, 시뮬레이션 입력 모형화에서 자주 고려되는 연속·이산 확률분포를 대상으로, 모수 추정을 위한 적정 표본크기를 결정하고자 한다. 즉, 표본크기 증가에 따른 ; 1) 모수 추정오차를 정량적으로 측정 ; 이를 바탕으로 2) 추정오차가 효과적으로 감소되어, 모수 추정치 정밀도·정확도를 적정하게 보장하는 표본크기를 검출하고자 한다. 2장는 실험인자 설명, 3장은 실험대상 확률분포 비대칭도 검토, 4장는 추정오차에 대한 표본크기 효과를 정량화하는 평가측도(evaluation measure) 제안, 5장는 실험결과 그래프 요약·분석, 그리고 6장은 사례분석을 제시한다.

2. 실험인자

- 확률분포(d) : 시뮬레이션 입력 모형화에서, 사용빈도가 큰 세 개 연속 및 한 개 이산 확률분포를 아래처럼 실험대상 확률분포로 선정한다 ; 1) 지수(exponential) ; 2) 감마(gamma) ; 3) 정규(normal) ; 및 4) 포아송(poisson). 식 (1)~(4)는, 순서대로 위 네 개 확률분포를 따르는 확률변수 x의 확률 밀도·질량함수를 나타낸다[9].

exponential(β) :

$$f(x)=\begin{cases}\beta^{-1}e^{-x/\beta} & \text{if } x\geq 0\\ 0 & \text{otherwise}\end{cases} \tag{1}$$

gamma (α,β) :

$$f(x)=\begin{cases}(\beta^{-\alpha}x^{\alpha-1}e^{-x/\beta})/\Gamma(\alpha) & \text{if } x\geq 0\\ 0 & \text{otherwise}\end{cases} \tag{2}$$

normal(μ,σ^2) :

$$f(x)=(\sqrt{2\pi}\sigma)^{-1}e^{-(x-\mu)^2/(2\sigma^2)} \quad \text{for all real numbers } x \tag{3}$$

poisson(λ) :

$$p(x)=\begin{cases}(e^{-\lambda}\lambda^{x})/(x!) & \text{if } x\in\{0,1,\ldots\}\\ 0 & \text{otherwise}\end{cases} \tag{4}$$

- 모수(θ) : 확률분포별 모수는 다음과 같이 선정된다 ; 1) exponential(β), $\beta=0.5, 5.0, 10.0$; 2) gamma (α, $\beta=1$), $\alpha=3.0, 5.0, 10.0$; 3) normal($\mu=0,\sigma^2$), $\sigma^2=1.0, 4.0, 9.0$; 4) poisson(λ), $\lambda=0.5, 5.0, 10.0$. 지수확률분포 exponential(β)는 감마확률분포의 형상모수(shape parameter) α를 '1'로 고정한 gamma ($\alpha=1,\beta$)로서, 척도모수(scale parameter) β만을 변화시켜 비대칭도가 일정한 확률분포를 표현한다. 반면, 감마확률분포에서는 scale parameter β를 '1'로 고정하고, shape parameter α를 변화시켜 비대칭도에 차이를 갖도록 설계한다. 정규확률분포에서는 위치모수(location parameter) μ를 '0'으로 고정하고, scale parameter σ^2만 변화되어 완전한 정규성(normality) 대칭도를 갖는다. 한편, 포아송확률분포에서는, 선정된 세 개 모수로써 비대칭도에 차이를 부여한다.
- 표본크기(n) : 확률분포 모수 추정을 위한 표본크기 n은, $[n_{min}, n_{max}]=[20,1000]$ 범위를 고려하되 ; 1) $n=400$ 이하에서는 $n_{new}\leftarrow n_{old}+20$으로 증가되는 표본크기를 대상으로 좁게 ; 2) $n=400$ 초과시는 $n_{new}\leftarrow n_{old}+100$으로 증가되는 표본크기를 대상으로 넓게 실험을 설계한다.

3. 비대칭도

각 확률분포의 선정 모수별, 비대칭도(ν) 검토를 위해, 식 (2)~(4)의 적률생성함수(moment generating function) $M(t)$의 3차 미분계수(derivative)를 도출하면, 순서대로 식 (5)~(7)과 같이 정리된다[7].

gamma (α,β) :

$$M'''(t)=(\alpha)(\alpha+1)(\alpha+2)(1-\beta t)^{(-\alpha-3)}\beta^3 \tag{5}$$

normal(μ,σ^2) :

$$M'''(t)=e^{(\mu t+\sigma^2t^2/2)}(\mu+\sigma^2t)(3\sigma^2+(\mu+\sigma^2t)^2) \tag{6}$$

poisson(λ) :

$$M'''(t)=e^{\lambda(e^t-1)}(\lambda e^t)(1+3(\lambda e^t)+(\lambda e^t)^2) \tag{7}$$

식 (5)~(7)로 정리된 $M'''(t)$와 확률변수 x의 평균 μ_x, 분산 σ_x^2을 비대칭도 정의식 식 (8)에 대입하면 <표 1>과 같이 정리된다.

$$\begin{aligned}\nu &= \mathrm{E}[(x-\mu_x)^3]/\sigma_x^3 \\ &= (\mathrm{E}(x^3)-3\mu_x\mathrm{E}(x^2)+3\mu_x^2\mathrm{E}(x)-\mu_x^3)/\sigma_x^3\end{aligned} \tag{8}$$

〈Table 1〉 Skewness of the probability distributions with selected parameters

Probability distributions	Moments			Parameters	Skewness
	1st	2nd	3rd		
d	$M'(0)\equiv E(x)$	$M''(0)\equiv E(x^2)$	$M'''(0)\equiv E(x^3)$	θ	ν
exponential(β)	β	$2\beta^2$	$6\beta^3$	$\beta=0.5$ $\beta=5.0$ $\beta=10.0$	2.00000
gamma(α,β)	$\alpha\beta$	$\alpha(\alpha+1)\beta^2$	$\alpha(\alpha+1)(\alpha+2)(\beta)^3$	$\alpha=3.0,\ (\beta=1.0)$ $\alpha=5.0,\ (\beta=1.0)$ $\alpha=10.0,\ (\beta=1.0)$	1.15470 0.89443 0.63246
normal(μ,σ^2)	μ	$\sigma^2+\mu^2$	$\mu(\mu^2+3\sigma^2)$	$(\mu=0.0),\ \sigma^2=1.0$ $(\mu=0.0),\ \sigma^2=4.0$ $(\mu=0.0),\ \sigma^2=9.0$	0.00000
poisson(λ)	λ	$\lambda(1+\lambda)$	$\lambda(1+3\lambda+\lambda^2)$	$\lambda=0.5$ $\lambda=5.0$ $\lambda=10.0$	1.41421 0.44721 0.31623

<표 1>에 제시된 바와 같이, 세 개 연속확률분포의 경우, 완전한 대칭도 '0'을 갖는 정규확률분포로부터 비대칭도가 '2'인 비정규(nonnormal) 지수확률분포까지를 고려한 실험설계이다. 반면, 포아송 이산확률분포의 경우, 대략 [0.3, 1.4] 범위내 세 개 비대칭도가 실험대상으로 설계된다.

4. 평가측도

모수 추정오차의 정량적 평가측도로서, 식 (9) 평균제곱오차(mean square error, mse)를 고려할 수 있다. 식 (9)는, 모수 θ를 위한 추정치 $\hat{\theta}$에 대한 ; 1) 샘플링 오차와 관련된 정밀도를 측정하는 분산(variance) ; 및 2) 모형화 오차와 관련된 정확도를 측정하는 편의(bias) 제곱을 동시에 평가하는 통계량이다[1, 10, 12, 13].

$$\begin{aligned} \text{mse} &= \mathrm{E}[(\hat{\theta}-\theta)^2] \\ &= \mathrm{E}[\hat{\theta}^2] - \mathrm{E}[2\hat{\theta}\theta] + \mathrm{E}[\theta^2] \\ &= \mathrm{E}[\hat{\theta}^2] - 2\theta\mathrm{E}[\hat{\theta}] + \theta^2 \\ &= \mathrm{E}[\hat{\theta}^2] - \mathrm{E}[\hat{\theta}]^2 + \mathrm{E}[\hat{\theta}]^2 - 2\theta\mathrm{E}[\hat{\theta}] + \theta^2 \\ &= \mathrm{var}[\hat{\theta}] + (\mathrm{E}[\hat{\theta}]-\theta)^2 \\ &= \text{variance} + \text{bias}^2 \end{aligned} \tag{9}$$

하지만, mse 사용상 한계점으로 다음 세 가지가 지적될 수 있다 ; 1) 이론적으로 θ를 알고 있을 때 계산 가능한 점 ; 2) θ 및 $\hat{\theta}$의 이차항(quadratic term)으로 계산되므로, θ 단위(unit)와는 불일치되어 비교·해석시 직관성이 저하된다는 점 ; 그리고 3) θ 크기(magnitude) 변화에 따라 mse 절대적 값 자체도 비례하여 변동될 수 있다는 점이다.

그럼으로, '0'이 아닌 θ가 주어지면 ; 1) θ 단위에 무차원적(dimensionless)이면서 ; 동시에 2) θ 크기에 상관없이 척도화된 평가측도로 θ 추정오차를 정량화할 필요가 있다. 상기 두 가지 요구를 보완해, 식 (10)을 제안하고, '제곱근 평균제곱오차 대비 모수 비율(square rooted Mean square error to parameter ratio, sMp)'이라 명칭한다. 단, $\theta=0$인 경우, 제곱근 평균제곱오차(즉, $\sqrt{\text{MSE}}=\sqrt{\mathrm{E}[\hat{\theta}^2]}$)로써 추정오차 검토를 대신하고자 한다.

$$\text{sMp} = \sqrt{\text{MSE}}/\theta \quad \text{if } \theta \neq 0 \tag{10}$$

5. 실험 및 분석

- 시나리오 : 2장에서 설명된 실험인자 ; 1) 확률분포(d) ; 2) 모수(θ) ; 및 3) 표본크기(n)를 달리 조합하여, Monte Carlo 시뮬레이션 실험 시나리오를 설계한다. 즉, (d,θ,n) 조합별 확률표집(random sampling) dataset을 대상으로 $\hat{\theta}$을 계산하고, 이 과정을 r=100회 반복(replication)한다. 각 조합마다, 100개 $\hat{\theta}$의 sMp를 도출한다. 한편, (d,θ) 조합별로는, $n_{max}=1000$인 dataset이 총 100개 생성된다. Minitab^{R}[11]이 제공하는 확률표집 기능을 이용해 dataset을 준비한다.
- 표본 비대칭도 : 시나리오 (d,θ) 조합별 $N=n_{max}\times r=10^5$개 총 표본의 비대칭도 ($\hat{\nu}(N)$)는 식 (11)을 이용해, <표 2>와 같이 정리된다. 단, 식 (11)에서 x_i, $\bar{x}_{(N)}$, $s_{(N)}$은 순서대로 총 표본 데이터, 평균, 표준편차를 지칭한다. <표 1>과 <표 2>를 비교하면, 실험설계시 의도된 비대칭도를 갖는 실험대상 dataset 생성이 확인된다.

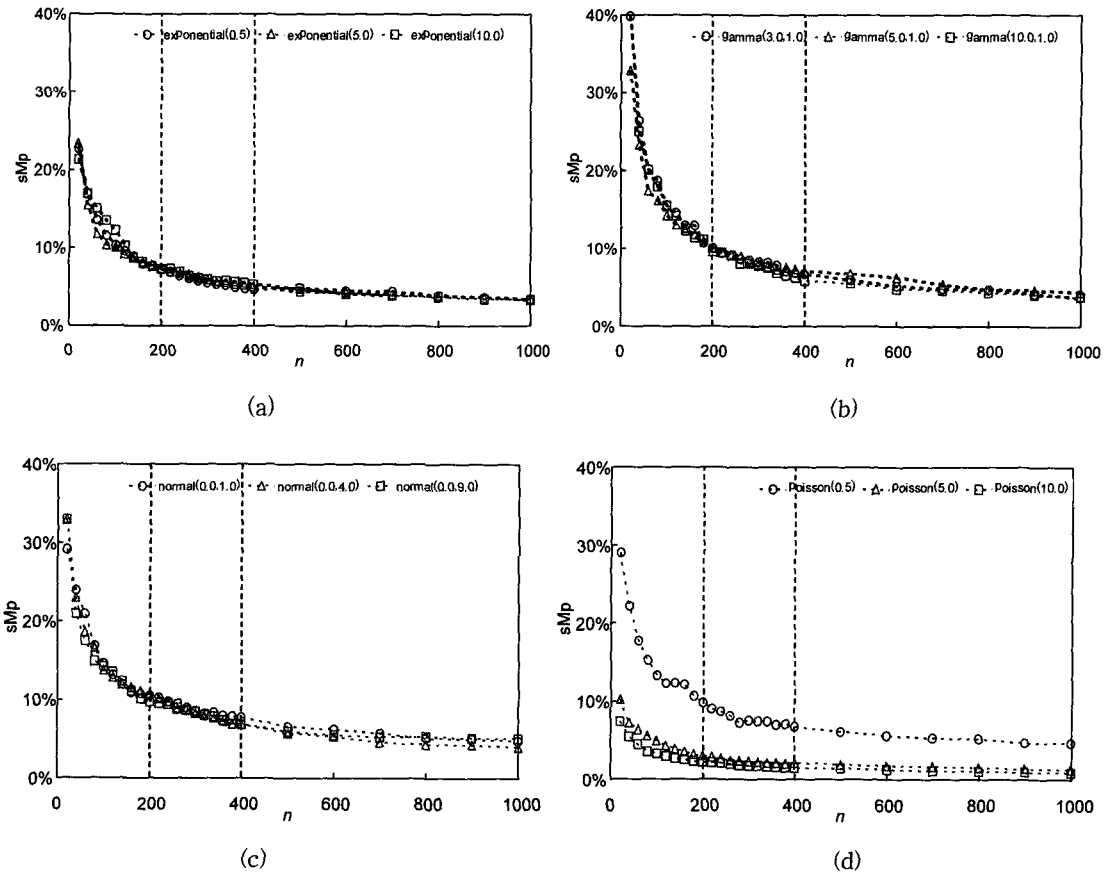
$$\hat{\nu}(N) = \frac{\sum_{i=1}^{N}[x_i - \bar{x}_{(N)}]^3/N}{s^3_{(N)}} \tag{11}$$

- sMp plots : 실험결과 sMp 통계치를 요약하면 [그림 1]과 같다. [그림 1] 네 개 panel은 순서대로 ; 1) exponential($\theta=\beta$) ; 2) gamma ($\theta=\alpha,\beta=1$) ; 3) normal($\mu=0,\theta=\sigma^2$) ; 및 4) poisson($\theta=\lambda$)에 대한, (d,θ,n) 조합별 100개 $\hat{\theta}$의 sMp plots를 제시한다. [그림 1(b)] sMp plots 작성을 위한 감마확률분포 모수 추정치 $\hat{\alpha}$은, Choi and Wette[5]의 최대가능도추정량(maximum likelihood estimator, mle)을 이용해 계산된다.

[그림 1] 네 개 panel 모두에서, 유사한 sMp plot 패턴이 시각적으로 확인된다 ; 1) $[n_{min},n_{max}]=[20,$

〈Table 2〉 Sample skewness with experimental dataset

Probability distributions	Parameters	Total sample size	Sample skewness
d	θ	N	$\hat{\nu}(N)$
exponential(β)	$\beta=0.5$	10^5	1.98204
	$\beta=5.0$		2.01906
	$\beta=10.0$		1.94522
gamma (α,β)	$\alpha=3.0,\ (\beta=1.0)$	10^5	1.16819
	$\alpha=5.0,\ (\beta=1.0)$		0.89700
	$\alpha=10.0,\ (\beta=1.0)$		0.62861
normal(μ,σ^2)	$(\mu=0.0),\ \sigma^2=1.0$	10^5	0.00046
	$(\mu=0.0),\ \sigma^2=4.0$		0.01514
	$(\mu=0.0),\ \sigma^2=9.0$		-0.00498
poisson(λ)	$\lambda=0.5$	10^5	1.40259
	$\lambda=5.0$		0.43399
	$\lambda=10.0$		0.31728



[Figure 1] sMp plots for probability distributions; (a) exponential; (b) gamma regarding α ; (c) normal regarding σ^2 ; and (d) poisson

1000]에서, θ 단위에 무차원적인 sMp (%)로써 추정오차에 대한 표본크기 효과가 정량화됨 ; 2) θ 크기에 상관없이 척도화되어, 각 panel별 세 개 sMp plots 모두 sMp 범위 [0.0, 0.5]에서 함께 비교됨 ; 3) n범위 [20, 200]에서, n증가할수록 sMp가 급감하여, sMp가 대부분 10%이하로 축소됨과 동시에, sMp plot 변곡점(inflection point) 존재 ; 4) n 범위 [200, 400]에서, n 증가에 따른 sMp 감소가 지속은 되나 비효과적 ; 그리고 5) n 범위 [400, 1000]에서는, 추가 표본에 따른 sMp 감소가 미미한 정체기 패턴을 확인할 수 있다.

만약, sMp plot 변곡점이 발생하는 최소 n을 '필요' 표본크기라고 정의하고, sMp plot 직선화가 발생하는 최소 n을 '충분' 표본크기라고 정의하면, Monte Carlo 시뮬레이션 결과, 대략 '필요' 표본크기 $n=200$ 그리고 '충분' 표본크기 $n=400$임을 확인할 수 있다.

<표 3>은 [그림 1]에 제시한 sMp 통계치를 요약한다. <표 3>의 마지막 세 행은, 순서대로 식 (12)~(14)에 정의된 것처럼, 최소 표본크기 $n=20$에서의 sMp 대비, 표본크기 증가에 따른 상대적 sMp 감소량을 제시한다. <표 3>에서 보이는 것처럼 ; 1) 초기 표본크기 증가분 $n=200\leftarrow20$에서 평균 69.38% ;

〈Table 3〉 Summary of sMp statistics from Monte Carlo simulation experiments

(unit : %)

	sMp											
	exponential($\theta=\beta$)			gamma($\theta=\alpha,\beta=1$)			normal($\mu=0,\theta=\sigma^2$)			poisson($\theta=\lambda$)		
n	$\beta=0.5$	$\beta=5.0$	$\beta=10.0$	$\alpha=3.0$	$\alpha=5.0$	$\alpha=10.0$	$\sigma^2=1.0$	$\sigma^2=4.0$	$\sigma^2=9.0$	$\lambda=0.5$	$\lambda=5.0$	$\lambda=10.0$
20	22.70	23.40	21.33	39.82	32.95	40.02	29.12	33.01	32.91	29.02	10.26	7.43
40	16.87	15.43	16.98	26.42	23.27	25.02	23.92	22.93	20.89	22.20	7.26	5.50
60	13.49	11.72	15.01	20.18	17.36	20.07	20.89	18.61	17.47	17.77	6.43	4.51
80	11.50	10.35	13.45	18.71	16.14	17.95	16.82	16.55	14.85	15.24	5.58	3.59
100	10.27	10.10	12.19	15.54	14.25	15.47	14.56	13.70	14.14	13.33	4.98	3.34
120	9.55	9.18	10.21	14.51	13.00	14.17	13.11	12.80	13.48	12.33	4.37	3.02
140	8.88	8.68	8.71	12.93	12.28	12.58	11.95	11.99	12.34	12.32	3.92	2.71
160	7.93	8.02	8.18	12.94	11.82	11.43	10.87	11.54	11.09	12.15	3.59	2.58
180	7.60	7.78	7.61	10.65	10.98	11.10	10.75	11.10	10.12	10.72	3.26	2.40
200	7.20	7.22	7.24	10.04	10.01	9.57	10.54	10.65	9.74	9.81	3.04	2.22
220	6.88	6.96	7.27	9.37	9.48	9.46	10.26	10.11	9.52	9.04	2.91	2.20
240	6.45	6.73	6.89	9.11	9.23	9.07	9.75	9.70	9.44	8.72	2.74	2.14
260	6.06	6.58	6.36	8.67	9.04	8.09	9.50	8.82	8.95	8.17	2.45	2.02
280	5.73	6.24	6.07	8.52	8.34	8.12	8.94	8.75	8.62	7.31	2.42	1.87
300	5.50	5.96	5.94	8.28	8.05	7.87	8.57	8.27	8.45	7.60	2.37	1.84
320	5.24	5.57	5.67	8.18	7.72	7.57	8.28	8.00	8.05	7.48	2.26	1.78
340	5.18	5.51	5.78	7.82	7.33	6.85	8.32	7.66	7.81	7.47	2.17	1.70
360	4.93	5.32	5.65	7.11	7.57	6.53	7.92	7.24	7.36	7.04	2.16	1.66
380	4.74	5.15	5.46	6.85	7.31	6.27	7.90	6.92	6.90	7.13	2.08	1.62
400	4.65	5.00	5.28	6.72	7.09	5.85	7.70	6.77	6.85	6.80	2.15	1.54
500	4.83	4.37	4.72	5.90	6.75	5.55	6.54	5.74	5.93	6.16	1.95	1.48
600	4.44	4.27	4.08	5.18	6.28	4.83	6.21	5.33	5.49	5.66	1.82	1.33
700	4.33	3.99	3.85	4.82	5.40	4.59	5.71	4.53	5.25	5.36	1.73	1.20
800	3.88	3.80	3.67	4.69	4.72	4.34	5.11	4.30	5.32	5.26	1.59	1.11
900	3.61	3.63	3.43	4.18	4.62	4.05	4.96	4.25	5.09	4.74	1.44	1.03
1000	3.37	3.43	3.35	3.70	4.40	3.78	4.77	3.98	5.08	4.75	1.38	0.99
Eq.(12)	68.26	69.13	66.04	74.78	69.61	76.09	63.80	67.74	70.41	66.19	70.41	70.14
Eq.(13)	9.84	8.36	9.35	6.66	7.25	9.03	8.78	10.10	8.13	7.73	7.43	8.87
Eq.(14)	6.47	4.03	6.44	5.53	7.13	4.43	6.07	5.34	2.58	4.88	5.58	6.69

2) 차기 표본크기 증가분 $n=400 \leftarrow 220$에서 평균 8.46% ; 그리고 3) 표본크기 증가분 $n=1000 \leftarrow 500$에서 평균 5.43% sMp 감소가 있음을 확인할 수 있다.

$$\text{Sample size effect}_{(n=200 \leftarrow 20)} = \frac{sMp_{(n=20)} - sMp_{(n=200)}}{sMp_{(n=20)}} \times 100\ (\%) \quad (12)$$

$$\text{Sample size effect}_{(n=400 \leftarrow 220)} = \frac{sMp_{(n=220)} - sMp_{(n=400)}}{sMp_{(n=20)}} \times 100\ (\%) \quad (13)$$

$$\text{Sample size effect}_{(n=1000 \leftarrow 500)} = \frac{sMp_{(n=500)} - sMp_{(n=1000)}}{sMp_{(n=20)}} \times 100\ (\%) \quad (14)$$

한편, [그림 1(b)] 감마확률분포 $\hat{\alpha}$의 sMp plots가 [그림 1] 나머지 세 개 panels와 비교해 상대적으로 높은 값을 갖는 현상은, mle 사용에 기인하는 것으로 추정된다. 또한, [그림 1(d)] 포아송확률분포 $\hat{\lambda}$의 sMp plots에서, 비대칭도가 상대적으로 큰 $\lambda=0.5$인 sMp plot이, 나머지 $\lambda=5.0, 10.0$인 두 개 sMp plots 대비, 높은 경향을 갖는 것이 확인된다.

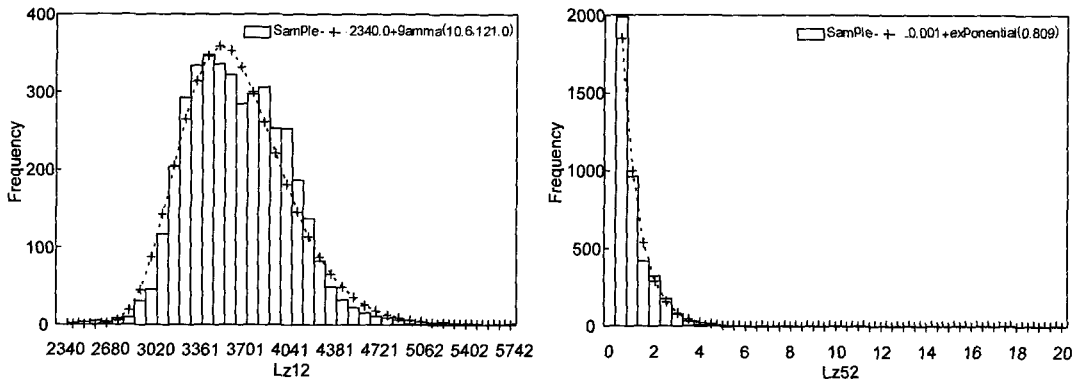
6. 사례분석

인터넷통신 네트워크 성능분석 관련, 하향속도(download speed, 'Lz12') 및 하향단절률(download disconnection rate, 'Lz52') 두 가지 네트워크 품질특성으로써, 나머지 품질특성을 대표하여, 시뮬레이션 입력 모형화를 위한 적정 표본크기를 결정하고자 한다. Lz12 및 Lz52 각각 총 표본 N=4015인 dataset을 준비하고, 각 품질특성 모집단을 대표하기에 충분하다고 가정한다. 단, 기밀성(confidentiality)과 관련하여, 사례분석 데이터는 임의의 단위를 갖는다고 가정한다.

Arena^R[2]를 이용해, Lz12 및 Lz52 dataset의 histogram 및 적합분포 함수식을 도출하면 [그림 2]와 같다. [그림 2]의 적합분포 함수식은, 식 (15) Arena^{R}의 제곱오차 판정시준(square error criterion)에 근거해 선택된다. 식 (15)에서, k는 히스토그램 계급구간 수, f_i 및 $\hat{f}_i$은 i^{th} 계급구간 표본 및 적합분포 함수식의 상대빈도를 순서대로 나타낸다.

$$\sum_{i=1}^{k}(f_i - \hat{f}_i)^2 \quad (15)$$

<표 4> panel (a), (b)는 Lz12 및 Lz52 dataset 대상, Arena^{R}를 이용해 도출된 적합분포 함수식과 제곱오차를 정리한 것이다[8, 9]. 단, <표 5>에서 확인되는 것처럼, <표 4>에서 가장 좋은 적합분포 함수식이라도 ; 1) chi-square test p-value < 0.005 ; 2) Kolmogorov-Smirnov test p-value < 0.010으로, 제1종오류(type I error) = 0.10 수준에서 Lz12, Lz52 모두 통계적 적합도검정에서는 기각



(a) (b)

[Figure 2] Histogram and fitted probability distribution using Arena^R for ; (a) Lz12 ; and (b) Lz52

〈Table 4〉 Fitted probability distribution functions and the square errors for ; (a) Lz12 ; and (b) Lz52

(a) Lz12

No.	Fitted distribution	Definition	Fitted function	Square error
1	gamma	offset+gamma(α,β)	2340.0+gamma(10.6,121.0)	0.00111
2	erlang	offset+erlang(m,β)	2340.0+erlang(11.0,117.0)	0.00111
3	beta	offset+r×beta(α_1,α_2)	2340.0+3400.0×beta(6.9,11.2)	0.00116
4	lognormal	offset+lognormal(μ,σ^2)	2340.0+lognormal(1310.0,460.0^2)	0.00193
5	normal	normal(μ,σ^2)	normal(3630.0,378.0^2)	0.00195
6	weibull	offset+weibull(α,β)	2340.0+weibull(4.2,1380.0)	0.00255
7	triangular	triangular(min,max,mode)	triangular(2340.0,5740.0,3400.0)	0.01600
8	uniform	uniform(min,max)	uniform(2340.0,5740.0)	0.04010
9	exponential	offset+exponential(β)	2340.0+exponential(1290.0)	0.04730

(b) Lz52

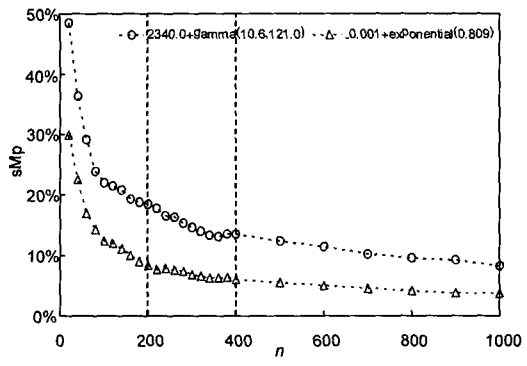
No.	Fitted distribution	Definition	Fitted function	Square error
1	exponential	offset+exponential(β)	-0.001+exponential(0.809)	0.00249
2	erlang	offset+erlang(m,β)	-0.001+erlang(1.000,0.809)	0.00249
3	gamma	offset+gamma(α,β)	-0.001+gamma(0.768,1.050)	0.00610
4	weibull	offset+weibull(α,β)	-0.001+weibull(0.760,0.793)	0.01280
5	beta	offset+r×beta(α,β)	-0.001+20.000×beta(0.922,15.300)	0.01300
6	lognormal	offset+lognormal(α,β)	-0.001+lognormal(2.090,11.500^2)	0.03370
7	normal	normal(μ,σ^2)	normal(0.808,0.992^2)	0.08950
8	triangular	triangular(min,max,mode)	triangular(-0.001,20.000,0.249)	0.24400
9	uniform	uniform(min,max)	uniform(-0.001,20.000)	0.27100

〈Table 5〉 Skewness and goodness-of-fit tests for the best fitted probability distribution function for ; a) Lz12 ; and b) Lz52

			Goodness-of-fit tests	
	Best fitted function	Skewness	Chi-square test	Kolmogorov-Smirnov test
(a) Lz12	2340.0+gamma(10.6,121.0)	$\nu = 0.61430$ $\hat{\nu}(N) = 0.45098$	No. of class intervals = 26 Degrees of freedom = 23 Test statistic = 124.000 p-value < 0.005	Test statistic = 0.032 p-value < 0.010
(b) Lz52	-0.001+exponential(0.809)	$\nu = 2.00000$ $\hat{\nu}(N) = 6.23496$	No. of class intervals = 10 Degrees of freedom = 8 Test statistic = 82.700 p-value < 0.005	Test statistic = 0.098 p-value < 0.010

된다. 또한, Lz52는, 가장 좋은 적합분포 함수식과 총 표본간 비대칭도가, Lz12와 비교해 상대적으로 더 큰 차이를 보인다.

<표 4>에서 검토된 아홉 가지 모형중 ; 1) Lz12, 2340.0+gamma (10.6, 121.0) ; 2) Lz52, -0.001+exponential(0.809)를 모집단 확률분포로 가정한다. 다음, Lz12, Lz52 각각 총 표본 N=4015로부터 확률표집된 표본을 5장 실험체계하에서 분석하면, [그림 3]과 같은 사례분석 dataset의 sMp plots를 작성할 수 있다.



[Figure 3] sMp plots for ; 1) Lz12, 2340.0+gamma (10.6,121.0); and 2) Lz52, −0.001+ exponential(0.809)

[그림 3]과 같이, 잡음(noise)이 상대적으로 큰 실제 데이터를 대상으로 한 사례분석 실험에서도, 5장 Monte Carlo 시뮬레이션 실험과 유사한 표본크기 증가 대비 sMp 감소 패턴을 확인할 수 있다. <표 6>에 제시된 것처럼, 초기 표본크기 증가분 $n=200\leftarrow20$에서 식 (12) 계산결과, Lz12 62.04%, Lz52 72.17%로써 sMp가 큰 폭 감소한다. 한편, 비대칭도가 비슷한 [그림 1(b)] gamma (10.0,1.0) 및 [그림 1(a)] exponential(0.5) sMp plot과 비교하면, [그림 3] 두 개 sMp plots는 상대적으로 높은 위치에 형성되지만, 변곡점 위치는 대략 n=200이하에 존재하는 것으로 판단된다.

〈Table 6〉 Summary of sMp statistics from experiments associated with ; (a) Lz12 ; and (b) Lz52

(unit : %)

(a) Lz12						(b) Lz52					
2340.0+gamma(10.6,121.0)						-0.001+exponential(0.809)					
n	sMp	n	sMp	n	sMp	n	sMp	n	sMp	n	sMp
20	48.50	220	17.73	500	12.32	20	29.89	220	7.67	500	5.49
40	36.31	240	16.54	600	11.48	40	22.50	240	7.88	600	5.04
60	29.12	260	16.26	700	10.23	60	16.93	260	7.49	700	4.59
80	23.85	280	15.27	800	9.59	80	14.26	280	7.35	800	4.18
100	21.98	300	14.62	900	9.24	100	12.35	300	6.79	900	3.81
120	21.49	320	13.92	1000	8.26	120	12.00	320	6.51	1000	3.76
140	20.77	340	13.30			140	11.05	340	6.24		
160	19.27	360	13.12	Eq.(12)	62.04	160	9.94	360	6.27	Eq.(12)	72.17
180	18.78	380	13.47	Eq.(13)	8.72	180	9.01	380	6.30	Eq.(13)	5.47
200	18.41	400	13.50	Eq.(14)	8.37	200	8.32	400	6.03	Eq.(14)	5.81

7. 결 론

시뮬레이션 입력 모형화에서 자주 고려되는 몇 가지 전형적인 연속 및 이산 확률분포를 대상으로, 확률분포 모수 추정을 위한 적정 표본크기 결정에 대해 논의하였다. Monte Carlo 시뮬레이션 및 사례분석 실험을 통해 ; 1) sMp로 정량화된 모수 추정 오차를 효과적으로 감소시키는 '필요' 표본크기 n=200 ; 그리고 2) 추가 표본크기 증가분이 갖는 sMp 감소효과가 미미한 경계값으로서 '충분' 표본크기 n=400이 제안된다. 특히, 기존 mse 통계량 대신 sMp를 제안·사용함으로써, 추정대상 θ 단위 및 크기에 무관하게 표본크기 증가 대비 θ 추정오차 감소효과(즉, 개선효과)를 정량적으로 평가할 수 있었다. 즉, [그림 1] 및 [그림 3]에서 확인할 수 있는 바와 같이, 추정대상 확률분포 모수 단위 및 크기에 상관없이 동일한 panel로 표본크기 증가에 따른 추정오차 감소 패턴을 비교할 수 있었다. 아울러, 6장 사례분석 인터넷통신 네트워크 품질특성 dataset을 대상으로, 5장 Monte Carlo 시뮬레이션 실험으로 도출된 sMp plots 패턴을 검증하였다. 향후, 본 연구에서 고려한 일반적인 확률분포에 적합되기 어려운 특히, 절단표본(truncated sample)을 위한 시뮬레이션 입력 모형화 및 적정 표본크기에 대한 연구를

연계하고자 있다.

참 고 문 헌

[1] 박성민, 김영식, "포아송 분포를 가정한 Wafer 수준 Statistical Bin Limits 결정방법과 표본크기 효과에 대한 평가", 「*IE Interfaces*」, 제17권, 제1호(2004), pp.1-12.

[2] Arena^{R} Version 3.01, Systems Modeling Corp., 1998.

[3] Banks, J., J.S. II Carson, and B.L. Nelson, *Discrete-Event System Simulation*, 2nd Edition, Prentice-Hall, New Jersey, 1996.

[4] Biller, B. and B.L. Nelson, "Answers to the Top Ten Input Modeling Questions," *2002 Winter Simulation Conference Proceedings*, San Diego, CA, December(2002), pp. 35-40.

[5] Choi, S.C. and R. Wette, "Maximum Likelihood Estimation of the Parameters of the Gamma Distribution and Their Bias," *Technometrics*, Vol.11, No.4(1969), pp.683-690.

[6] Gross, D., "Sensitivity of Output Performance Measures to Input Distribution Shape in Modeling Queues-3 : Real Data Scenario," *1999 Winter Simulation Conference Proceedings*, Phoenix, AZ, December(1999), pp.452-457.

[7] Hogg, R.V. and A.T. Craig, *Introduction to Mathematical Statistics*, 4th Edition, Collier Macmillan, New York, 1978.

[8] Kelton, W.D., R.P. Sadowski, and D.A. Sadowski, *Simulation With Arena*, McGraw-Hill, New York, 1998.

[9] Law, A.M. and W.D. Kelton, *Simulation Modeling and Analysis*, 3rd Edition, McGraw-Hill, New York, 2000.

[10] Leemis, L., "Input Modeling," *2003 Winter Simulation Conference Proceedings*, New Orleans, LA, December(2003), pp.14-24.

[11] Minitab^{R} Release 14.1, Minitab Inc., 2003.

[12] Montgomery, D.C., *Introduction to Statistical Quality Control*, 4th Edition, John Wiley & Sons, New York, 2001.

[13] Montgomery, D.C., E.A. Peck, and G.G. Vining, *Introduction to Linear Regression Analysis*, 3rd Edition, John Wiley & Sons, New York, 2001.

[14] Montgomery, D.C. and G.C. Runger, *Applied Statistics and Probability for Engineers*, 2nd Edition, John Wiley & Sons, New York, 1999.

[15] Schmeiser, B., "Advanced Input Modeling for Simulation Experimentation," *1999 Winter Simulation Conference Proceedings*, Phoenix, AZ, December(1999), pp.110-115.