박만식교수지도

석사학위청구논문

이산형 자료를 기반으로 한 학습곡선의 모형화

2014

성신여자대학교 대학원

통계학과

최민지

이산형 자료를 기반으로 한 학습곡선의 모형화

박만식교수지도

이 논문을 석사학위 논문으로 제출함

2013년 11월

성신여자대학교 대학원

통계학과

최민지

인 준 서

최민지의 석사학위 논문으로 인준함.

심사위원 _____인

심사위원 _____인

심사위원 _____인

성신여자대학교 대학원

논 문 개 요

어떤 작업을 수행할 때 처음에는 작업자가 능숙하지 못해 그 작업을 처리하는데에 시간이 오래 걸린다. 그러나 작업 수행이 반복되면서 작업자가 작업에 능숙해져 작업을 처리하는 시간이 점차 줄어드는 것을 확인할 수 있다. 이러한 효과를 학습곡선효과(learning-curve effect)라 한다. 최근 기업에서도 이 개념을 확대한 '경험곡선(experience curve)'을 사용하는데, 반복횟수가 늘어날수록 노동의 시간 뿐만 아니라 전체 비용도 줄어든다는 개념이다. 본 연구에서는 작업이 능숙해짐에 따라 성취도가 증가하는 형태를 살펴보고자 한다. 즉, 작업을 수행할 때에 초기에는 작업자의 성취도가 낮지만, 작업을 반복함으로써 성취도가 증가하다가 일정 시간이 지난 후 안정화 되는 형태를 살펴보고자 한다. 자료는 특정 사건의 발생 횟수인 이산형 자료(count data)이고, 누적분포함수를 이용하여 학습곡선을 모형화 하고자 한다. 모수추정을 위한 방법으로 최대우도추정법과 베이지안추정법을 사용하였다. 우도함수를 이용하여 모수를 추정하는 방법 중 수치적 방법인 뉴튼- 랩슨방법과 우도함수를 이용하여 근사적 평균과 분산을 구하는 델타방법을 이용한다. 제안한 모형의 특성을 살펴보고 모형 평가를 위해 모의실험을 시행하였다. 영국 광산 사고 자료, 국내 내륙 지진 자료와 국내 해역 지진 자료를 제안한 방법으로 분석하였다.

주요어 : 포아송분포, 음이항분포, 학습곡선, 로지스틱 분포함수, 최대우도추정법, 베이지안추정법, 영국 광산 사고 자료

목 차

그 림 목 차

표 목 차

제 1 장

서론

‘과제를 수행하는 횟수가 증가할수록 같은 일을 하는데 드는 시간이 줄어든다’라는 ‘학습곡선효과(learning-curve effect)’는 19세기 독일의 심리학자 헤르만 에빙하우스(Hermann Ebbinghaus)가 처음으로 이 개념을 주장하였다. 즉, 작업을 수행할 때에 처음에는 작업자가 능숙하지 못해 작업을 처리하는데에 시간이 오래 걸린다. 그러나 작업 수행이 반복되면서 작업자가 작업에 능숙해져 작업을 처리하는 시간이 점차 줄어드는 것을 확인할 수 있다. 이런 작업 처리 시간은 점차 줄어드는 형태를 나타내다가 일정한 기간이 지나고 나면 작업 처리 시간이 특정 시간으로 안정화 되는 것을 확인할 수 있다. 최근에는 기업에서도 학습곡선의 범위를 확대하여 ‘경험 곡선(experience curve)’이라는 개념으로 많이 사용되고 있다. 경험곡선은 동일한 작업에 대하여 경험, 즉 반복 횟수가 증가할수록 노동의 시간, 즉 작업 처리 시간 뿐만 아니라 전체 비용도 줄어든다는 개념이다. 이런 현상에서 볼 수 있듯이 과거에 비해 학습곡선이 보편적으로 사용됨을 알 수 있다. 일반적인 학습곡선은 초기 시점에 비해 일정 기간이 지나면서 학습 효율이 증가하다 점점 안정화 되는 것을 확인할 수 있다.

학습곡선은 다양한 분야에서 일정 시간에 지남에 따라 학습 효과 전후에 차이가 있는지를 분석하는 방법 중 하나이다. 먼저 산업 공학 분야에서 살펴보면, 백우종 (2008) 은 연료전지 비용 적용을 위해 비용

요소를 고려하기 위해 학습곡선을 사용하여 연료전지 가격은 내연기관에 비교할만한 수준으로까지 감소할 수 있다는 가능성을 시사하였다. 정종교 (2007)는 한계시간평균모형과 누적시간평균모형을 통하여 생산성향상 전후로 공수 절감효과를 나타내었다. 활동 전에는 학습효과가 매우 느리게 나타났지만 활동 후에는 활동 전에 비하여 크게 향상됨을 보였다. 뿐만 아니라 디자인 공학에서 홍자인 (2007)은 한계시간모델을 이용하여 휴대폰의 수행 시간에 대하여 수행예측시간과 수행측정시간에 대해 *t*-검정을 통하여 유의한 차이가 없음을 보였다. 특히 의학 분야에서도 수술 건수 당 수술 시간에 관한 분석 등 학습곡선을 이용한 분석이 많이 활용되고 있다. 정호석 (2009)은 복막외 복강경하 전립선적출술의 경우 초기에는 9시간 이상이 소요되어 개복으로 전환하기도 하였다. 하지만 수술 경험이 쌓이면서 지속적으로 수술 시간이 짧아지고 있는 학습곡선효과를 확인하였다. 이성진과 박수은 (2007)은 유리체절제술에서의 수술의 실패에 관하여 연구하였다. 연구 결과 수술실패율이 3년의 경험 이후로 감소하는 것을 확인하였다. Jeff *et al.* (2013)는 양성 부인과에서의 로봇을 이용한 자궁 절제술의 수술시간에 관하여 연구하였는데, 수술 시간은 20~30건의 수술 경험 이후 유의하게 감소하는 것을 확인하였다. Lee *et al.* (2013)는 고관절 치환술에서의 비구컵 위치에 관한 학습곡선 연구를 하였다. 누적 실패 횟수에 대하여 의사 2명 모두 30건의 수술 이후 실패 횟수가 줄어든 것을 확인할 수 있었다. Kuo *et al.* (2013)는 괄약근간 절제술에 대한 연구에서는 학습효과 전후로 평균 수술시간이 줄어듬을 확인하였다. 또한, 이 수술시간은 의사의 수술 경험 횟수가 18번 이전과 이후에 차이가 있음을 밝혔다. Akin *et al.* (2013)은 비뇨기과의 복강경 수술에서의 합병증율에 대해서 연구하였는데, 합병증율은 초기 시점으로부터 3년이 지난

후에 안정화되는 형태는 나타냈다. Chuan *et al.* (2012)은 살아있는 기증자의 간 이식에 관하여 연구하였다. 수술 시간과 기증자의 수술 중의 혈액 손실이 유의하게 감소하는 것을 확인할 수 있었다. 또한, Forbes *et al.* (2004)는 한 명의 의사의 혈관 내 동맥류 복구 수술의 성공률은 경험이 쌓임에 따라 발전하는 것을 cumulative sum(CUSUM)을 이용하여 밝혔다. Ferguson *et al.* (2005)는 의사 혼자 복강경 근치 전립선 절제술(laparoscopic radical prostatectomy)을 시행 할 경우의 수술시간을 측정하였다. 수술 기간을 12구간으로 나눈 후 비교한 결과, 첫 번째 구간과 네 번째 구간에서 수술시간이 유의하게 줄어들었음을 확인하였다. 이와같이 의학분야에서의 학습곡선이 이용된 논문에는 Ballantyne *et al.* (2005), Biau *et al.* (2008), Cook *et al.* (2004), Filho (2002), Lee *et al.* (2006), Li *et al.* (2012), Lim *et al.* (2011), Schauer *et al.* (2002), Sim *et al.*(2006)와 Tekkis *et al.* (2005) 이 있다. 그 외에도 이슬지 (2011) 는 이항반응 자료(베르누이자료)에 대하여 누적확률분포의 특성을 이용하여 학습곡선을 모형화하였다. 임경민 (2010)은 일정한 제품을 생산하는 과정에서 누적생산량이 2배가 될 때 단위비용의 절감정도를 학습률로 정의하였다. 분석 결과 풍력발전의 누적 설비용량이 증가함에 따라 학습률이 증가했다. Lieberman (1984)은 화학 산업에서의 비용이 시간이 지남에 따라 감소하는 것을 확인하였다. Williams와 Vivarelli (2000) 는 모의실험를 수행하였는데, 그 결과 학습곡선과 신뢰구간이 초기시점과 선형적으로 감소하는 시점에서 특이하게 나타나는 것을 확인하였다. 이 외에 다양한 분야에서 사용되는 논문에는 Adler와 Clark (1991), Kim and Park (2012), Mazzola와 McCardle (1996), Simith *et al.* (2004), Smunt (1999)와 Williams and Vivarelli (2000)이 있다.

본 논문에서는 작업이 능숙해짐에 따라 성취도가 증가하는 형태

그림 1: 1978년부터 2012년까지의 국내 내륙 지진 발생 추이(건수)

를 살펴보고자 한다. 즉, 작업을 수행할 때에 처음에는 작업자의 성취도가 낮게 나타난다. 그러나 반복적으로 작업을 수행함으로써 작업자의 성취도가 증가하는 형태를 나타내다가 일정 기간이 지나고나면 성취도가 안정화 되는 것을 살펴보고자 한다. 이런 형태의 학습곡선은 자연현상이나 사회현상에서도 찾아볼 수 있다. 본 눈문에서는 이슬지(2011)의 확장으로 이항반응자료 뿐만 아니라 이산형 자료에 대해서 학습곡선을 모형화 하고자 한다. 한 예제로 사례연구에 나오는 국내 내륙 지진자료에 대해서 살펴보도록 한다. 그림 1은 1978년부터 2012년까지의 국내 내륙에서의 지진 발생 추이이다. 초기 시점에서는 지진 발생 건수가 낮다가 시간이 지남에 따라 증가하는 패턴을 보인다. 그러다가 일정 시간후에는 안정화 되는 모습을 볼 수 있다. 이런 자연현상과 사회현상 등에서 나타나는 학습곡선의 형태를 모형화 하고자 한다. 물론 지진과 같은 통제 불가능한 자연현상은 학습에 의해 발생하

는 사건이 아니므로 학습곡선을 적용시키는데 어려움이 있다. 본 논문에서는 제공된 자료만을 이용하여 자연현상에 대한 추세를 알아보기 위하여 학습곡선을 적용시킨다. 우선 특정 사건에 관한 초기 시점에서의 발생 횟수와 일정 시간이 지난 후 안정화 되었을 때의 평균 발생 횟수를 추정한다. 또한 초기 시점에서의 발생 횟수와 안정화 되었을 때의 발생 횟수의 평균 값을 지나는 시점인 변곡점과 학습이 일어나는 기간에 대해 추정하고자 한다. 분석을 통해 구한 추정값들을 통하여 학습곡선을 모형화 한다.

본 논문의 순서는 다음과 같다. 제2장에서는 이산형 자료를 이용하여 학습곡선을 모형화 하고, 제3장에서는 제2장에서 제시한 모형에 대해 모의실험을 진행한다. 제4장에서는 영국 광산 사고 자료, 국내 내륙 지진 자료와 국내 해양 지진 자료를 토대로 모형 적합을 한다. 마지막으로 제5장에서는 결론 및 앞으로의 연구방향에 대해 논의한다.

제 2 장

이산형 자료를 이용한 학습곡선의 모형화

작업을 수행할 때에 처음에는 능숙하지 못해 시간이 오래 걸리지만, 작업 수행 횟수가 반복되면서 수행자가 작업에 능숙해져 작업을 처리하는 시간이 점차 줄어들면서 일정한 시간이 지나면 안정화 되는 것을 확인할 수 있다. 이러한 현상을 학습곡선 효과라고 할 수 있다. 일반적인 학습곡선은 능숙도가 증가하면서 처리하는 시간이 줄어드는 현상을 나타냈다. 그와 반대로 작업이 능숙해짐에 따라 성취도가 증가하는 경우도 있다. 즉, 처음 작업을 수행할 때에는 성취도가 낮지만 작업 수행 횟수가 반복됨에 따라 성취도가 증가하여 안정화 되는 모형을 나타낸다. 이런 경우는 자연현상이나 사회현상에서도 찾아볼 수 있다. 예를 들어, 서론에서 살펴본 그림 1이 하나의 현상이다. 본 논문에서는 이런 현상에 대하여 관측값이 점차 증가하다가 시간이 지남에 따라 안정화되는 현상을 모형화 하고자 한다. 본 논문에서 다룰 사건은 이산형 자료이다. 이 자료는 특정 사건의 발생 횟수로, 일정한 시간 간격을 두고 관찰한 자료를 사용한다. 이산형 자료를 갖는 분포에는 여러 종류가 있는데 그 중에서도 포아송분포(Poisson distribution)와 음이항분포(negative binomial distribution)를 이용하고, 누적분포함수(cumulative distribution function; *cdf*)의 특성을 이용하여 학습곡선의 통계적 모형화를 하고자 한다.

2.1 이산형 자료의 확률 분포

이산형 자료를 갖는 분포에는 여러 종류가 있다. 예를 들어, 이항분포(binomial distribution), 기하분포(geometric distribution), 포아송분포, 음이항분포, 초기하분포(hypergeometric distribution)와 다항분포(multinomial distribution) 등이 있다. 이 중에서 본 논문에서 다룰 분포는 특정 사건의 발생 횟수와 관련이 있는 포아송분포와 음이항분포이다. 포아송분포와 음이항분포에 대해 알아보도록 하자.

2.1.1 포아송분포

주어진 시간 동안 일어날 확률이 매우 적은 특정한 사건이 발생하는 횟수는 포아송분포를 따른다고 할 수 있다. 시점을 $t = 1, \cdots, T$ 라 할 때, 특정 시점 t에서의 사건 발생 횟수 Y_t는 평균 발생 횟수가 λ_t인 포아송분포를 따르고 이를 다음과 같이 표현한다.

$$Y_t \sim Poi(\lambda_t).$$

포아송분포의 확률질량함수(probability mass function; *pmf*)는 아래와 같다.

$$g(y_t|\lambda_t) = \frac{e^{-\lambda_t}\lambda_t^{y_t}}{y_t!}, \quad y_t = 0, 1, \cdots. \tag{2.1}$$

포아송분포의 적률생성함수(moment generating function; *mgf*)는 다음과 같이 나타낼 수 있다.

$$M(s) = E(e^{sY_t}) = e^{\lambda_t(e^s-1)}, \quad -\infty < s < \infty.$$

적률생성함수를 이용하여 평균과 분산을 구하면 다음과 같다.

$$E(Y_t) = \frac{d}{ds}M(s)\bigg|_{s=0} = \lambda_t,$$

$$Var(Y_t) = \frac{d}{ds^2}M^2(s)\bigg|_{s=0} - \left[\frac{d}{ds}M(s)\bigg|_{s=0}\right]^2 = \lambda_t.$$

위에서 살펴보았듯이 포아송분포는 평균과 분산이 동일하다는 특징을 갖고 있다.

2.1.2 음이항분포

성공 확률이 p_t인 베르누이 시행(bernoulli trial)에서 첫 번째 성공이 일어날 때까지의 전체 시행 횟수를 X_t라 한다면, 이 X_t는 기하분포를 따르고, 확률질량함수는 다음과 같다.

$$g(x_t|p_t) = p_t(1-p_t)^{x_t-1}, \quad x_t = 1,2,3,\cdots.$$

성공 확률이 p_t인 베르누이 시행에서 r번 성공할 때까지의 전체 시행 횟수를 W_t라 할 때, 이 W_t는 음이항분포을 따르고, 확률질량함수는 다

음과 같다.

$$g(w_t|p_t,r) = \frac{\Gamma(w_t)}{\Gamma(r)\Gamma(w_t-r+1)} p_t^r (1-p_t)^{w_t-r}, \quad w_t = r, r+1, r+2, \cdots.$$

여기서, $\Gamma(a) = (a-1)!$이다. 음이항분포에서 확률변수를 전체 시행 횟수로 할 수도 있지만, 실패 횟수로도 할 수 있다. r번 성공할 때까지의 실패 횟수를 Y_t라 하면 전체 시행 횟수는 $W_t = Y_t + r$이 된다. Y_t를 사용하여 음이항분포를 나타내면 확률질량함수는 다음과 같다.

$$g(y_t|p_t,r) = \frac{\Gamma(y_t+r)}{\Gamma(r)\Gamma(y_t+1)} p_t^r (1-p_t)^{y_t}, \quad y_t = 0, 1, 2, \cdots. \tag{2.2}$$

음이항분포의 적률생성함수는 다음과 같이 나타낼 수 있다.

$$\begin{aligned} M(s) = E(e^{sY_t}) &= \sum_{y_t=0}^{\infty} e^{sy_t} \frac{\Gamma(y_t+r)}{\Gamma(r)\Gamma(y_t+1)} p_t^r (1-p_t)^{y_t} \\ &= \left(\frac{p_t e^s}{1-(1-p_t)e^s} \right)^r, \qquad s < -\ln(1-p_t). \end{aligned}$$

적률생성함수를 이용하여 평균과 분산을 구하면 다음과 같다.

$$E(Y_t) = \frac{d}{ds} M(s) \bigg|_{s=0} = \frac{r(1-p_t)}{p_t}, \tag{2.3}$$

$$Var(Y_t) = \frac{d}{ds^2} M^2(s) \bigg|_{s=0} - \left[\frac{d}{ds} M(s) \bigg|_{s=0} \right]^2 = \frac{r(1-p_t)}{p_t^2}.$$

일반적으로 음이항분포를 표현할 때에는 성공 확률 p_t를 이용하여 나타낸다. 포아송분포의 모수는 평균 발생 횟수이다. 포아송분포에서 사

용되는 모수와 동일한 형태의 모수를 사용하기 위해 음이항분포의 평균인 식(2.3)을 λ_t라고 하여 확률질량함수를 재표현하고자 한다. 먼저 음이항분포의 확률질량함수는 식(2.2)이다. 평균 발생 횟수 $E(Y_t)$를 λ_t라 하면 이 때의 성공 확률은 $p_t = r/(r+\lambda_t)$가 된다. p_t를 이용한 재표현된 음이항분포의 확률질량함수는 식(2.4)이다.

$$g(y_t|\lambda_t, r) = \frac{\Gamma(y_t + r)}{\Gamma(r)\Gamma(y_t + 1)} \left(\frac{r}{r+\lambda_t}\right)^r \left(\frac{\lambda_t}{r+\lambda_t}\right)^{y_t}, \quad y_t = 0, 1, \cdots. \quad (2.4)$$

이를 다음과 같이 표현한다.

$$Y_t \sim NB(\lambda_t).$$

2.2 학습곡선의 모형화를 위한 모형 구축

본 논문에서 다룰 자료는 일정한 시간 간격을 두고 특정 사건을 반복적으로 관찰한 이산형 자료이다. 추정해야할 모수는 평균 발생 횟수인 λ_t로서 4개의 모수를 이용하여 그림 2를 따르는 학습곡선 모형을 구축하고자 한다. 그림 2에서의 θ_1은 특정 사건에 대해 초기 시점에서의 발생 횟수를 나타낸다. θ_2는 일정 시간이 지난 후 안정화 되었을 때의 평균 발생 횟수를 뜻한다. 또한, 위치모수(location parameter)인 θ_3와 척도모수(scale parameter)인 θ_4를 고려하여, $\Theta = (\theta_1, \theta_2, \theta_3, \theta_4)'$ 추정하고자 한다. 즉, λ_t가 $\lambda_t(\Theta)$가 된다. 위치모수 θ_3는 초기 시점에서의 발생 횟수와 안정화 되었을 때의 발생 횟수의 평균값인 $(\theta_1 + \theta_2)/2$를 지날 때의 시점이다. 즉, 특정 사건이 계속 증가하다가 안정기로 접어드는 시점이다. 또한 θ_1과 θ_2는 0 이상의 값을 가지며, $0 \le \theta_1 < \theta_2$인 조건을

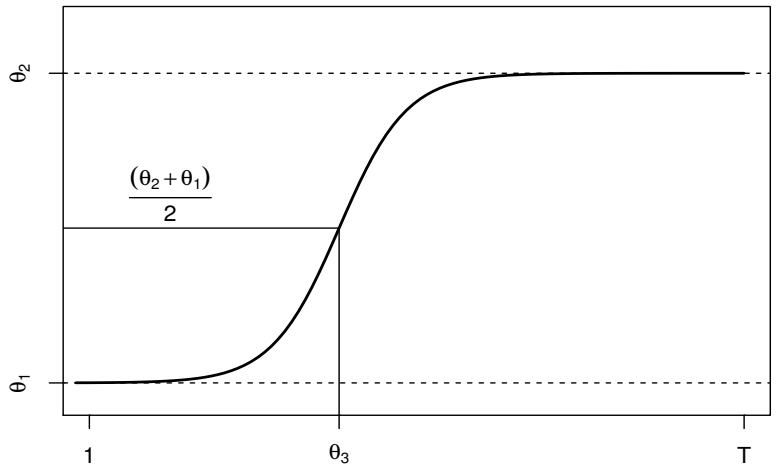

그림 2: 누적분포함수를 이용한 학습곡선의 모형화

따라야 한다. 평균 발생 횟수인 $\lambda_t(\Theta)$를 정의하면 식(2.5)와 같다.

$$\lambda_t(\Theta) = \theta_1 + (\theta_2 - \theta_1) \cdot F(t|\theta_3, \theta_4). \tag{2.5}$$

여기서, $F(\cdot)$은 임의의 시점 t의 누적분포함수로서 다음과 같이 표현된다.

$$F(t|\theta_3, \theta_4) = \int_{-\infty}^{t} f(x|\theta_3, \theta_4)dx, \quad -\infty < t < \infty.$$

본 논문에서는 t의 함수로 로지스틱분포(logistic distribution)를 이용하고자 한다. 로지스틱분포를 따르는 어떤 확률변수 X가 위치모수 θ_3와 척도모수 θ_4를 가진다고 하자. 이 로지스틱 분포는 $x = \theta_3$를 기준으로 대칭으로 나타난다. 즉, 종 모양을 갖는 정규분포(normal distribution)와 유사한 모양을 가지며, 정규분포에 비해 두꺼운 꼬리를 갖는다. 로지

스틱분포의 확률밀도함수와 누적분포함수는 각각 다음과 같다.

$$f(x|\theta_3,\theta_4) = \frac{\exp(-(x-\theta_3)/\theta_4)}{\theta_4\left(1+\exp(-(x-\theta_3)/\theta_4)\right)^2},$$

$$F(x|\theta_3,\theta_4) = \left[1+\exp\left(-\frac{x-\theta_3}{\theta_4}\right)\right]^{-1}, \quad -\infty < x < \infty.$$

여기서, $-\infty < \theta_3 < \infty, 0 < \theta_4 < \infty$이다. 시점 t에서 관측하는 발생 횟수 Y_t가 식(2.5)의 $\lambda_t(\Theta)$를 가지는 포아송분포 혹은 음이항분포를 따른다고 하자. 식(2.5)의 $F(\cdot)$가 로지스틱분포의 누적분포함수라 하면 $\lambda_t(\Theta)$는 아래와 같이 나타낼 수 있다.

$$\begin{aligned}\lambda_t(\Theta) &= \theta_1 + (\theta_2-\theta_1)\cdot F(t|\theta_3,\theta_4) \\ &= \theta_1 + (\theta_2-\theta_1)\cdot\left[1+\exp\left(-\frac{t-\theta_3}{\theta_4}\right)\right]^{-1}.\end{aligned} \tag{2.6}$$

2.3 모수 추정

2.1장에서 정의한 포아송분포와 음이항분포, 그리고 2.2장에서 정의한 $\lambda_t(\Theta)$ 를 이용하여 모수를 추정하고자 한다. 본 논문에서 추정해야 할 모수는 $\lambda_t(\Theta)$로 4개의 모수들로 이루어져 있다. 따라서 $\lambda_t(\Theta)$를 추정하기 위해서는 각각의 모수들을 추정해야 한다. 본 논문에서는 모수를 추정하기 위해서 최대우도추정법(maximum likelihood estimation method)과 베이지안추정법(Bayesian estimation method)을 사용하고자 한다. 우도함수(likelihood function) 를 이용하여 모수를 추정하는 방법 중 수치적 방법인 뉴튼-랩슨방법(Newton-Raphson method)과 우도함수를 이용하여 근사적 평균과 분산을 구하는 델타방법(Delta method)을

이용한다. 또한, 사전분포(prior distribution)와 우도함수를 이용하여 사후분포(posterior distribution)를 추정하는 방법인 베이지안추정법에 대해서 알아보도록 하자.

2.3.1 최대우도추정법

모수를 추정할 때 사용되는 방법들 중 하나는 최대우도추정법이다. 최대우도추정법은 우도함수 또는 로그우도함수를 이용하여 모수를 추정하는 방법이다. 로그우도함수는 확률질량(밀도)함수의 결합으로 이루어진 우도함수에 로그를 취한 것이다. 우도함수를 이용하여 모수를 추정하는 방법들 중 수치적으로 방정식을 해결하는 방법에는 뉴튼-랩슨방법이 있다. 이 뉴튼-랩슨방법은 먼저 초기 추정값을 지정한 후, 식(2.7)인 점화식(recursive formula)을 이용하여 다음 단계의 추정값, 즉 m 단계의 $\widehat{\Theta}^{(m)}$과 이전 단계인 $(m-1)$단계의 $\widehat{\Theta}^{(m-1)}$ 의 차이가 거의 없어질 때까지 동일한 과정을 반복적으로 시행한다.

$$\widehat{\Theta}^{(m)} = \widehat{\Theta}^{(m-1)} + \left[\mathbf{I}^{(m-1)}\right]^{-1} \mathbf{u}^{(m-1)}. \tag{2.7}$$

여기서, $\mathbf{u}$는 스코어 통계량으로서 로그우도함수를 모수로 편미분한 결과로 4×1인 벡터값을 갖는다. $\mathbf{I}$는 피셔의 정보량(Fisher's Information)으로 로그우도함수를 2번 편미분한 결과로 4×4인 행렬이다. $\mathbf{y} = (y_1, y_2, \cdots, y_T)'$라 할 때, 로그우도함수는 다음과 같다.

$$l(\Theta|\mathbf{y}) \equiv \ln L(\Theta|\mathbf{y}) = \ln \prod_{t=1}^{T} g(\mathbf{y}|\lambda_t(\Theta)) = \sum_{t=1}^{T} \ln g\left(\mathbf{y}|\lambda_t(\Theta)\right).$$

또한, 우도함수를 이용하여 모수의 추정값에 대한 근사적 평균과

분산을 구하는 방법으로 델타방법이 있다. 어떤 확률변수의 확률분포를 알고 있다 하더라도 변수의 분포식 혹은 우도함수가 복잡한 함수의 형태라면 기댓값과 분산을 구하는 것은 쉽지 않다. 이런 경우에는 테일러 급수(Tayler series)에 기반을 두어 계산 과정을 간단하게 하여 평균과 분산을 쉽게 구할 수 있는 델타방법을 사용한다. 임의의 $t = 1, \cdots T$에 대한 λ_t에 대한 $100 \times (1-\alpha)\%$ 근사 신뢰구간(asymptotic confidence interval)은 다음과 같이 나타낼 수 있다.

$$\lambda_t(\widehat{\Theta}) \pm Z_{\alpha/2}\sqrt{\widehat{Var}(\lambda_t(\widehat{\Theta}))}.$$

이 때, 델타방법을 이용하여 구한 $\lambda_t(\hat{\Theta})$의 기댓값은 다음과 같다.

$$E[\lambda_t(\hat{\Theta})] \approx \lambda_t(\Theta) + \sum_{i=1}^{4} \frac{\partial \lambda_t}{\partial \theta_i} E(\widehat{\theta}_i - \theta_i) = \lambda_t(\Theta).$$

델타방법을 이용하여 구한 $\lambda_t(\hat{\Theta})$의 분산은 다음과 같다.

$$\begin{aligned} Var(\lambda_t(\hat{\Theta})) &\approx E\left[\left(\lambda_t(\widehat{\Theta}) - \lambda_t(\Theta)\right)^2\right] \\ &= \sum_{i=1}^{4}\left(\frac{\partial \lambda_t}{\partial \theta_i}\right)^2 Var(\hat{\theta}_i) + 2\sum_{i>j}^{4}\left(\frac{\partial \lambda_t}{\partial \theta_i}\right)\left(\frac{\partial \lambda_t}{\partial \theta_j}\right) Cov(\hat{\theta}_i, \hat{\theta}_j). \end{aligned}$$

2.3.1.1 포아송분포

확률변수 Y_t가 $Poi(\lambda_t(\Theta))$를 따른다고 할 때, 포아송분포의 확률질량함수는 식(2.1)이다. 편의상 $\lambda_t(\Theta)$는 λ_t로 표기하기로 한다. 로그우도

함수 $l(\cdot)$는 다음과 같다.

$$l(\lambda_t|\mathbf{y}) = -\sum_{t=1}^{T}\lambda_t + \sum_{t=1}^{T} y_t \ln\lambda_t - \sum_{t=1}^{T}\ln y_t!.$$

로그우도함수를 λ_t로 1차 편미분하면 다음과 같다.

$$\frac{\partial}{\partial\lambda_t} l(\lambda_t|\mathbf{y}) = \sum_{t=1}^{T}\left(\frac{y_t}{\lambda_t} - 1\right).$$

스코어 통계량 $\mathbf{u}$는 로그우도함수를 각각의 모수에 대하여 편미분을 하여 구하는데, 그 식은 다음과 같이 나타낼 수 있다.

$$\mathbf{u}(\Theta|\mathbf{y}) = \sum_{t=1}^{T}\frac{\partial l(\lambda_t|\mathbf{y})}{\partial\Theta} = \{U_i(\Theta|\lambda_t)\}_{i=1,2,3,4}. \tag{2.8}$$

위의 식에서의 $U_i(\cdot)$는 $i=1,2,3,4$에 대해서 다음과 같다.

$$U_i(\Theta|\mathbf{y}) = \sum_{t=1}^{T}\frac{\partial l(\lambda_t|\mathbf{y})}{\partial\theta_i} = \sum_{t=1}^{T}\frac{\partial l(\lambda_t|\mathbf{y})}{\partial\lambda_t}\frac{\partial\lambda_t}{\partial\theta_i}.$$

λ_t를 θ_i로 편미분하면 다음과 같다.

$$\frac{\partial\lambda_t}{\partial\theta_i} = \begin{cases} 1 - F(t|\theta_3,\theta_4), & i=1 \\ F(t|\theta_3,\theta_4), & i=2 \\ -\dfrac{\theta_2-\theta_1}{\theta_4}F(t|\theta_3,\theta_4)[1-F(t|\theta_3,\theta_4)], & i=3 \\ -\dfrac{\theta_2-\theta_1}{\theta_4^2}(t-\theta_3)F(t|\theta_3,\theta_4)[1-F(t|\theta_3,\theta_4)], & i=4. \end{cases}$$

정보행렬(information matrix) **I**는 식(2.8)의 스코어 통계량 **u**를 모수 Θ에 대하여 한 번 더 편미분한 값이다.

$$\mathbf{I} = \left\{ I_{ij} \right\}_{i,j=1,2,3,4} = \left\{ \frac{\partial}{\partial \theta_j} U_i(\Theta|\mathbf{y}) \right\}_{i,j=1,2,3,4}. \tag{2.9}$$

위 식에서의 $i, j = 1,2,3,4$에 대한 I_{ij}는 다음과 같다.

$$I_{ij} = \frac{\partial U_i(\Theta|\mathbf{y})}{\partial \theta_j} = \frac{\partial U_i(\Theta|\mathbf{y})}{\partial \lambda_t} \frac{\partial \lambda_t}{\partial \theta_j} = \left[\frac{\partial}{\partial \lambda_t} \frac{\partial l(\lambda_t|\mathbf{y})}{\partial \lambda_t} \frac{\partial \lambda_t}{\partial \theta_i} \right] \frac{\partial \lambda_t}{\partial \theta_j}.$$

2.3.1.2 음이항분포

확률변수 Y_t가 음이항분포 $NB(\lambda_t(\Theta))$를 따른다고 할 때, 음이항분포의 확률질량함수는 식(2.4)이고, 로그우도함수는 다음과 같다.

$$l(\lambda_t|\mathbf{y}) = \sum_{t=1}^{T} \left(\frac{\Gamma(y_t + r)}{\Gamma(r)\Gamma(y_t + 1)} \right) + \sum_{t=1}^{T} y_t \ln \left(\frac{\lambda_t}{r + \lambda_t} \right) + r \sum_{t=1}^{T} \ln \left(\frac{r}{r + \lambda_t} \right).$$

로그우도함수에 대한 1차 편미분은 다음과 같다.

$$\frac{\partial}{\partial \lambda_t} l(\lambda_t|\mathbf{y}) = \sum_{t=1}^{T} \frac{y_t}{\lambda_t} - \sum_{t=1}^{T} \frac{y_t}{r + \lambda_t} - r \sum_{t=1}^{T} \frac{1}{r + \lambda_t}.$$

이하 스코어 통계량과 정보행렬은 각각 식(2.8)과 식(2.9)로 포아송분포와 동일하다.

2.3.2 베이지안추정법

베이지안추정법은 모수에 관한 과거의 경험이나 지식 등의 주관적인 견해를 바탕으로 사전분포를 정의하고 관측된 자료와 사전 정

보를 이용하여 사후분포를 추론한다. 사전분포를 $\pi(\Theta)$라고 설정하고, 자료 $\mathbf{y}$에 대한 모수 Θ의 우도함수 $L(\mathbf{y}|\Theta)$를 정의한다. 베이즈 정리(Bayes' Theorem)에 의하여 주어진 자료 $\mathbf{y}$에 대한 모수 $\lambda_t(\Theta)$의 조건부분포(conditional distribution)는 다음과 같이 나타난다.

$$g(\Theta|\mathbf{y}) = \frac{L(\Theta|\mathbf{y})\pi(\Theta)}{g(\mathbf{y})} = \frac{L(\Theta|\mathbf{y})\pi(\Theta)}{\int_{\Theta} L(\Theta|\mathbf{y})\pi(\Theta)d\Theta}. \tag{2.10}$$

식(2.10)에서의 $g(\Theta|\mathbf{y})$를 사후분포라 하며, 이 사후분포는 사전분포와 우도함수의 곱에 비례한다. 각 모수에 대한 사전분포 정의는 아래와 같다.

$$\theta_1 \sim Unif(0,a), \quad \theta_2 \sim Unif(\theta_1,a),$$
$$\theta_3 \sim Unif(0,T), \quad \theta_4 \sim IGamma(b,c).$$

여기서, $Unif(\cdot)$은 균일분포(uniform distribution)이고, $IGamma(\cdot)$는 역감마분포(inverse gamma distribution)를 의미한다. θ_2는 $0 \leq \theta_1 < \theta_2$인 조건을 만족해야 하므로 $Unif(\theta_1,a)$을 θ_2의 사전분포로 정의한다. 따라서 θ_2의 사전분포함수(prior distribution function)는 아래와 같다.

$$f(\theta_2|\theta_1,a) = \frac{1}{a-\theta_1}, \quad \theta_1 < \theta_2 < a,$$

θ_4의 사전분포함수는 아래와 같다.

$$f(\theta_4|b,c) = \frac{b^a}{\Gamma(a)}\theta_4^{(-a-1)}\exp(-b/\theta_4), \quad 0 < \theta_4 < \infty.$$

제 3 장

모의실험

제2장에서 모수 추정법으로 최대우도추정법과 베이지안추정법을 살펴보았다. 학습곡선 모형이 자료에 얼마나 적합한지를 평가하기 위해 모의실험을 수행하였다. 모의실험의 기본절차는 다음과 같다. 먼저 모의실험에 사용할 자료를 생성하기 위하여 분포와 모수를 설정한다. 설정된 분포와 모수를 기반으로 분석에 쓰일 자료를 생성한다. 생성된 자료를 이용해서 모수 추정을 한다. 본 논문에서 자료 생성을 위한 분포로는 포아송분포와 음이항분포를 사용한다. 모수 Θ_0는 $\Theta_0 = (\theta_{10}, \theta_{20}, \theta_{30}, \theta_{40})'$라고 정의한다. 먼저, 특정 사건에 대해 초기 시점에서의 발생 횟수인 θ_{10}은 1, 2, 5로, 일정 시간이 지난 후 안정화 되었을때의 발생 횟수인 θ_{20}는 20으로 지정하였다. 위치모수인 θ_{30}은 40, 척도모수인 θ_{40}는 2, 5, 10으로 정하여 9개의 모의실험을 수행하였다. 모의실험에서의 시점은 $t = 1, \cdots, T$로 하였고, 표본 크기가 T인 자료를 B 개 생성하였다. 각 시나리오별 모수의 참값을 정리하면 표 1과 그림 3으로 나타낼 수 있다. 모수의 참값은 포아송모형과 음이항모형 모두 동일하다. 여기서 포아송모형은 포아송분포를 따르는 학습곡선의 모형이고, 음이항모형은 음이항분포를 따르는 학습곡선의 모형이다. 또한 $Poi(\lambda_t(\Theta_0))$로부터 생성된 자료는 포아송자료, $NB(\lambda_t(\Theta_0))$로부터 생성된 자료는 음이항자료라 한다.

그림 3에 대해서 살펴보자. 그림 3에 나타난 7개의 점선은 평균

표 1: 9개의 모의실험 시나리오

	Scenario								
Θ_0	1.	2.	3.	4.	5.	6.	7.	8.	9.
θ_{10}	1	1	1	2	2	2	5	5	5
θ_{20}	20	20	20	20	20	20	20	20	20
θ_{30}	40	40	40	40	40	40	40	40	40
θ_{40}	2	5	10	2	5	10	2	5	10

발생 횟수 증가율의 시점이다. 즉, $(\theta_{20}-\theta_{10})\times a$로 구할 수 있다. 발생 횟수의 5% 증가율 시점은 $(\theta_{20}-\theta_{10})\times 0.05$로 구할 수 있다. 시나리오 1, 시나리오 2와 시나리오 3은 $\theta_{10}=1$, $\theta_{20}=20$, $\theta_{30}=40$으로 동일하고, θ_{40}은 각각 2, 5, 10이다. θ_{40}가 증가함에 따라 학습곡선이 완만하게 증가하는 것을 확인할 수 있다. 즉, 안정화 단계로 접어드는 기간이 증가함을 뜻한다. 다음으로 시나리오 1, 시나리오 4와 시나리오 7은 $\theta_{20}=20$, $\theta_{30}=40$, $\theta_{40}=1$로 동일하고, θ_{10}은 각각 1, 2, 5이다. θ_{10} 이 증가함에 따라 변곡점에서 평균 발생 횟수가 증가하였다. 증가율에 대해서 살펴보도록 하자. 증가율은 θ_{40}가 증가함에 따라 증가율의 시점 간격이 점점 커짐을 확인할 수 있다. 시나리오 1, 시나리오 2와 시나리오 3 을 살펴보면 θ_{40}가 증가할 수록 증가율 시점 간격이 점점 넓어지는 것을 확인할 수 있다. 시나리오 4, 시나리오 5와 시나리오 6을 살펴보면 θ_{40}가 증가할 수록 증가율 시점 간격이 점점 넓어지는 것을 확인할 수 있다. 시나리오 7, 시나리오 8과 시나리오 9를 살펴보면 역시 θ_{40}가 증가할 수록 증가율 시점 간격이 점점 넓어지는 것을 확인할 수 있다.

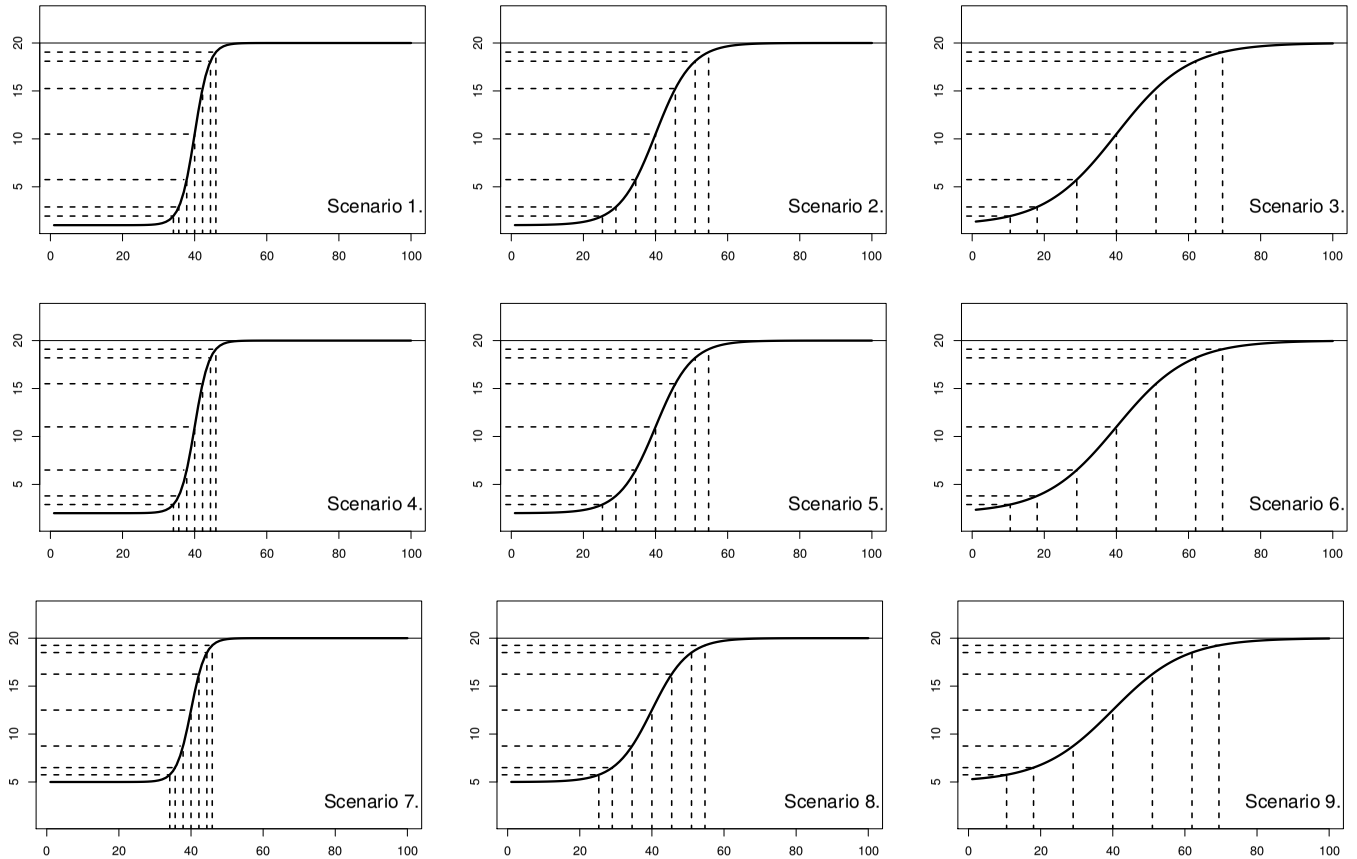

그림 3: 9개의 모의실험 시나리오

(왼쪽 점선부터 증가율이 5%, 10%, 25%, 50%, 75%, 90%, 95% 인 시점)

3.1 모의실험 방법

모수 추정을 위한 모의실험 방법은 아래와 같다. 편의상 시나리오 1을 기준(표 1 참조)으로 모수의 참값은 $\theta_{10}=1$, $\theta_{20}=20$, $\theta_{30}=40$, $\theta_{40}=2$로 나타내겠다.

1) 발생 시점 $t=1,\cdots,T$에 대하여 로지스틱분포의 누적분포함수를 이용하여 $\lambda_t(\Theta_0)$를 구한다. $\lambda_t(\Theta_0)$는 식(2.6)을 참고한다. 예를 들어 $t=5$인 경우에 다음과 같이 계산된다.

$$\begin{aligned}\lambda_5(\Theta_0) &= \theta_{10}+(\theta_{20}-\theta_{10})\times\left[1+\exp\left(-\frac{5-\theta_{30}}{\theta_{40}}\right)\right]^{-1}\\ &= 1+(20-1)\times\left[1+\exp\left(-\frac{5-20}{2}\right)\right]^{-1}=1.0105.\end{aligned}$$

2) 1)단계에서 구한 모든 시점에서의 $\lambda_t(\Theta_0)$를 이용하여, 특정한 분포(포아송분포 또는 음이항분포)로부터 표본 크기가 T인 B 개의 자료를 만든다. 즉, $b=1,\cdots,B$ 에 대하여 다음과 같이 나타낸다.

$$\mathbf{y}^{(b)}=\left\{y_1^{(b)},\cdots,y_t^{(b)},\cdots,y_T^{(b)}\right\}'.$$

여기서, $y_t^{(b)}\sim Poi(\lambda_t(\Theta_0))$ 또는 $y_t^{(b)}\sim NB(\lambda_t(\Theta_0))$이다.

3) 2)단계에서 만든 B개의 자료에 대하여 모수를 추정한다. 모수 추정방법으로는 최대우도추정법과 베이지안추정법을 사용한다. 각 추정결과에 대해서 추정값으로는 평균, 중위수, 분산(variance)과 평균제곱오차(mean squared error; MSE)를 사용한다. b번째 자료를 특정한 모형에 적합한 후 얻게 되는 i번째 모수의 추정값을 $\widehat{\theta}_i^{(b)}$라 하자. 그러면 $i=1,2,3,4$에 대한 분산과 평균제곱오차는

다음과 같이 정의한다.

$$\widehat{\mathrm{Var}}(\widehat{\theta}_i) = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\theta}_i^{(b)} - \bar{\widehat{\theta}}_i\right)^2,$$
$$\widehat{\mathrm{MSE}}(\widehat{\theta}_i) = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\theta}_i^{(b)} - \theta_{i0}\right)^2.$$

여기서, $\bar{\widehat{\theta}}_i = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}_i^{(b)}$ 이다.

4) 추정결과에 따른 추정값을 이용하여 λ_t의 $100\times(1-\alpha)\%$ 경험적 신뢰구간(empirical confidence interval)을 구한다. 경험적 신뢰구간을 2가지 방법으로 구하였다. 첫 번째 방법은 $\widehat{\lambda}_t$를 만든 후 정렬하여 사용하는 방법이다. 두 번째 방법은 $\widehat{\theta}_i$의 순서통계량을 이용하는 방법이다. 이 방법은 B개의 $\widehat{\theta}_i$를 정렬한 후, 새로운 추정값인 $\widehat{\lambda}_t^*$를 생성하여 사용하는 방법이다. 각 방법의 내용은 다음과 같다.

(1) $\widehat{\lambda}_t$의 순서통계량을 이용한 $100\times(1-\alpha)\%$ 경험적 신뢰구간

$$\left(\{\widehat{\lambda}_t\}_{\left(\frac{(B+1)\cdot\alpha}{2}\right)}, \{\widehat{\lambda}_t\}_{\left((B+1)\left(1-\frac{\alpha}{2}\right)\right)}\right). \tag{3.1}$$

$\widehat{\lambda}_{t(j)}$는 정렬된 $\widehat{\lambda}_t$ 중 j번째 순서통계량이다.

(2) $\widehat{\theta}_i$의 순서통계량을 이용한 $100\times(1-\alpha)\%$ 경험적 신뢰구간

$$\left(\{\widehat{\lambda}_t^*\}_{\left(\frac{(B+1)\cdot\alpha}{2}\right)}, \{\widehat{\lambda}_t^*\}_{\left((B+1)\left(1-\frac{\alpha}{2}\right)\right)}\right). \tag{3.2}$$

여기서, $\widehat{\lambda}_t^*$는 아래와 같다.

$$\widehat{\lambda}^*_{t(j)} = \hat{\theta}_{1(j)} + (\hat{\theta}_{2(j)} - \hat{\theta}_{1(j)}) \left[1 + \exp\left(-\frac{t - \hat{\theta}_{3(B+1-j)}}{\hat{\theta}_{4(j)}} \right) \right]^{-1}. \tag{3.3}$$

여기서, $\hat{\theta}_{i(j)}$는 정렬된 B개의 $\hat{\theta}_i$ 중 j번째 순서통계량이고, $\widehat{\lambda}^*_{t(j)}$는 각각의 $\hat{\theta}_{i(j)}$로 만들어진 값이다. 위 식에서 $\widehat{\theta}_3$는 j번째 순서통계량 대신 $(B+1-j)$번째 순서통계량을 사용하였다. θ_3는 위치모수로 안정기로 접어드는 시점을 뜻한다. 하한 신뢰구간은 참값보다 안정화가 늦게 일어나고, 상한 신뢰구간은 참값보다 안정화가 먼저 일어난다. 따라서 내림차순 정렬 순서를 사용하기 위해 $(B+1-j)$번째를 사용한다. 베이지안추정법을 이용한 $100\times(1-\alpha)\%$ 경험적 신뢰구간도 식(3.1)과 식(3.2)로 동일하게 구한다.

5) 근사 신뢰구간은 2.3장에서 정의한 델타방법을 이용하여 구한다.

$$\left(\lambda_t(\hat{\Theta}) - Z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\lambda_t(\hat{\Theta}))}, \lambda_t(\hat{\Theta}) + Z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\lambda_t(\hat{\Theta}))} \right). \tag{3.4}$$

$\lambda_t(\widehat{\Theta})$는 B개의 $\widehat{\theta}_i$의 평균을 이용하여 구한다. $\widehat{\mathrm{Var}}(\lambda_t(\widehat{\Theta}))$ 를 구할 때 사용되는 분산과 공분산(covariance) 은 최대우도추정법을 이용하여 구한 분산과 공분산을 사용한다. B개의 $\mathrm{Var}(\lambda_t(\widehat{\Theta}))$ 의 평균을 $\widehat{\mathrm{Var}}(\lambda_t(\widehat{\Theta}))$로 이용한다.

3.2 모의실험결과

모의실험에서는 최대우도추정법과 베이지안추정법을 이용하여 모수를 추정하고, 최대우도추정법에서의 추정결과와 베이지안추정법에서의 추정결과를 이용하여 신뢰구간을 구한다. 또한 각각의 포아송자료를 음이항모형으로, 음이항자료를 포아송모형에 적합시켜 각 모형의 성능을 평가하였다. 9개의 시나리오에 대해서 $T = 100, B = 199$로, 음이항분포에서의 성공 횟수 r은 100으로 하여 분석하였다. 베이지안추정법에서 B개의 추정값은 각각 11000번을 반복하여 얻은 사후 추정값 결과 중 초기 1000개의 추정값을 제외한 나머지 10000개의 사후 평균으로 구해진 값을 사용한다.

3.2.1 모수추정 결과

모수 추정 결과는 포아송자료를 포아송모형에 적용한 결과와 음이항자료를 음이항모형에 적용한 결과를 살펴보도록 한다. 추정 결과와 함께 추정 결과를 이용한 경험적 신뢰구간과 근사적 신뢰구간을 구해보도록 한다.

표 2는 포아송자료의 9개의 시나리오에 대한 모수들의 참값과 추정값이다. 추정값으로는 최대우도 추정값과 베이지안 추정값들의 평균, 중위수, 분산과 평균제곱오차를 살펴보도록 한다. 먼저 추정값들 중 추정 방법에 따른 평균과 중위수를 살펴보았다. 추정값과 모수의 참값 간에 크게 차이가 나타나지 않았다. 분산과 평균제곱오차에 대해서 살펴보면 최대우도추정법에서의 시나리오 9와 베이지안추정법에서의 시나리오 8과 시나리오 9를 제외하고 모든 시나리오에서 θ_{30}의 분산과 평균제곱오차가 가장 크게 나타났다. 또한, θ_{10}과 θ_{40}의 값이

표 2: 포아송자료를 포아송모형에 적용한 추정 결과

			Maximum Likelihood				Bayesian			
Scenario	Param	Truth	Mean	Median	Var	MSE	Mean	Median	Var	MSE
1	θ_{10}	1	0.994	0.995	0.028	0.028	1.029	1.030	0.029	0.030
	θ_{20}	20	20.107	20.104	0.308	0.320	20.114	20.112	0.309	0.322
	θ_{30}	40	40.136	40.065	0.449	0.467	40.121	40.039	0.456	0.470
	θ_{40}	2	2.011	1.944	0.235	0.235	2.004	1.965	0.244	0.244
2	θ_{10}	1	0.969	0.970	0.066	0.067	1.010	1.009	0.064	0.064
	θ_{20}	20	20.021	19.962	0.440	0.441	20.045	19.979	0.442	0.444
	θ_{30}	40	40.017	40.058	1.109	1.109	40.056	40.105	1.118	1.121
	θ_{40}	5	5.034	4.998	0.566	0.567	5.052	5.003	0.570	0.573
3	θ_{10}	1	1.070	1.054	0.272	0.276	1.173	1.113	0.157	0.187
	θ_{20}	20	20.020	19.898	0.940	0.940	20.085	19.959	0.933	0.940
	θ_{30}	40	40.140	39.949	3.861	3.881	40.392	40.266	3.993	4.146
	θ_{40}	10	9.771	9.595	2.586	2.638	9.769	9.701	2.007	2.060
4	θ_{10}	2	1.994	1.980	0.060	0.060	2.033	2.023	0.060	0.061
	θ_{20}	20	20.015	19.964	0.369	0.369	20.018	19.958	0.372	0.373
	θ_{30}	40	40.026	40.052	0.479	0.479	40.004	40.043	0.489	0.489
	θ_{40}	2	1.980	1.949	0.332	0.332	1.962	1.956	0.343	0.344
5	θ_{10}	2	1.958	1.992	0.108	0.110	1.988	2.023	0.109	0.109
	θ_{20}	20	19.984	19.976	0.483	0.483	20.016	20.014	0.490	0.490
	θ_{30}	40	39.955	39.989	1.383	1.385	39.993	40.063	1.395	1.395
	θ_{40}	5	4.918	4.882	0.920	0.926	4.972	4.949	0.968	0.969
6	θ_{10}	2	1.864	1.884	0.465	0.484	1.866	1.794	0.302	0.320
	θ_{20}	20	20.221	20.224	0.907	0.955	20.366	20.395	0.930	1.064
	θ_{30}	40	40.158	40.365	5.511	5.535	40.414	40.551	5.667	5.839
	θ_{40}	10	10.180	9.949	3.850	3.883	10.468	10.410	3.173	3.392
7	θ_{10}	5	4.922	4.934	0.181	0.187	4.968	4.981	0.181	0.182
	θ_{20}	20	20.009	19.985	0.307	0.307	20.005	19.980	0.308	0.308
	θ_{30}	40	40.050	39.918	0.902	0.905	40.020	39.856	0.910	0.911
	θ_{40}	2	2.019	2.023	0.662	0.662	1.982	1.946	0.638	0.639
8	θ_{10}	5	4.903	4.941	0.336	0.345	4.887	4.949	0.386	0.398
	θ_{20}	20	20.101	20.069	0.451	0.461	20.156	20.107	0.489	0.513
	θ_{30}	40	39.955	39.895	2.708	2.710	39.956	39.844	2.759	2.760
	θ_{40}	5	5.073	4.776	2.329	2.335	5.260	4.918	3.069	3.137
9	θ_{10}	5	4.714	4.886	1.376	1.458	4.289	4.281	1.258	1.764
	θ_{20}	20	20.277	20.122	0.889	0.966	20.749	20.589	1.394	1.955
	θ_{30}	40	40.210	40.003	8.598	8.641	40.363	39.989	10.230	10.362
	θ_{40}	10	10.537	10.203	8.951	9.240	12.122	11.902	12.456	16.959

Param,Parameter; Var,Variance; MSE,mean squared error

커짐에 따라 대부분 추정값의 분산과 평균제곱오차가 커짐을 확인할 수 있다. 최대우도추정법의 추정값들을 살펴보도록 하자. 먼저, θ_{10}을

고정시킨 경우를 살펴보도록 하자. 시나리오 4, 시나리오 5와 시나리오 6은 $\theta_{10} = 2$일 때, θ_{40}는 각각 2, 5, 10이다. θ_{40}가 커짐에 따라 모든 모수의 분산과 평균제곱오차가 커짐을 확인할 수 있다. 특히, θ_{40}가 5에서 10으로 증가할 때, θ_{30}의 분산은 1.383에서 5.511로, θ_{40}의 분산은 0.920에서 3.850으로 크게 커짐을 확인할 수 있다.

다음으로는 θ_{40}을 고정시킨 경우를 살펴보도록 하자. 시나리오 2, 시나리오 5와 시나리오 8은 $\theta_{40} = 2$일 때, θ_{10} 는 각각 1, 2, 5이다. θ_{10}이 커짐에 따라 대부분 모수의 분산과 평균제곱오차가 커짐을 확인할 수 있다. θ_{10}이 1에서 2, 2에서 5로 증가할 때, θ_{10}의 분산은 0.066에서 0.108로, 0.108에서 0.336으로 증가하는 것을 확인할 수 있다. 그러나 θ_{20}의 분산은 0.440에서 0.483으로, 0.483 에서 0.451로 증가하다가 감소하는 것을 확인할 수 있다. θ_{30}와 θ_{40}는 θ_{10}이 커짐에 따라 분산과 평균제곱오차의 값이 커진다. 9개의 시나리오들 중 9번째 시나리오, 즉, $\theta_{10} = 5$, $\theta_{20} = 20$, $\theta_{30} = 40$, $\theta_{40} = 10$인 경우의 θ_{30}와 θ_{40} 의 분산과 평균제곱오차가 크게 나타났다.

베이지안추정법의 추정값들을 살펴보도록 하자. 먼저, θ_{10}을 고정시킨 경우를 살펴보도록 하자. 시나리오 4, 시나리오 5와 시나리오 6은 $\theta_{10} = 2$일 때, θ_{40}는 각각 2, 5, 10이다. θ_{40}가 커짐에 따라 모든 모수의 분산과 평균제곱오차가 커짐을 확인할 수 있다. 특히, θ_{40}가 2에서 5로, 5 에서 10으로 증가할 때, θ_{30}의 분산은 0.489에서 1.395로, 1.395에서 5.667로 커졌고, θ_{40}의 분산은 0.343 에서 0.968로, 0.968에서 3.173으로 크게 커짐을 확인할 수 있다.

다음으로는 θ_{40}을 고정 시킨 경우를 살펴보도록 하자. 시나리오 2, 시나리오 5와 시나리오 8은 $\theta_{40} = 2$일 때, θ_{10}는 각각 1, 2, 5이다. θ_{10}이 커짐에 따라 대부분 모수의 분산과 평균제곱오차가 커짐을 확인할 수

있다. θ_{10}이 1에서 2, 2에서 5로 증가할 때, θ_{10}의 분산은 0.064에서 0.109로, 0.109에서 0.386으로 증가하는 것을 확인할 수 있다. 그러나 θ_{20}의 분산은 0.442에서 0.490으로, 0.490 에서 0.489로 증가하다가 감소하는 것을 확인할 수 있다. θ_{30}와 θ_{40}는 θ_{10}이 커짐에 따라 분산과 평균제곱오차의 값이 커진다. 9개의 시나리오들 중 9번째 시나리오, 즉, $\theta_{10} = 5$, $\theta_{20} = 20$, $\theta_{30} = 40$, $\theta_{40} = 10$인 경우의 θ_{30}와 θ_{40} 의 분산과 평균제곱오차가 크게 나타났다.

다음은 각 시점마다 199번 반복하여 얻은 포아송자료의 신뢰구간들을 구하였다. 신뢰구간은 참값을 이용한 발생 횟수와 최대우도 추정값를 이용한 평균 적합선, 2종류의 95% 경험적 신뢰구간과 근사 신뢰구간을 구하였다. 평균 적합선은 최대우도법을 이용하여 구한 각 시나리오별 평균 추정값을 이용하여 구하였다. 2종류의 95% 경험적 신뢰구간은 식(3.1)과 식(3.2)를 이용하였다. 근사 신뢰구간은 식(3.4)를 이용하였다. 그림 4를 살펴보자. 전반적으로 평균 적합선이 참값과 매우 유사하게 나타났다. 첫 번째 경험적 신뢰구간과 근사 신뢰구간은 매우 유사하게 나타났다. 두 번째 경험적 신뢰구간은 평균 발생 횟수가 증가하기 시작하는 시점에서는 신뢰구간이 넓게, 변곡점을 지나 안정화되기 직전의 시점에서는 신뢰구간이 비교적 좁게 나타났다. $\theta_{10} = 2$, $\theta_{20} = 20$, $\theta_{30} = 40$인 시나리오 4, 시나리오 5와 시나리오 6을 살펴보자. 이 시나리오들의 θ_{40}는 각각 2, 5, 10으로 θ_{40}가 커짐에 따라 변곡점 이전 시점에서 두 번째 경험적 신뢰구간이 넓게 나타났다. $\theta_{20} = 20$, $\theta_{30} = 40$, $\theta_{40} = 5$인 시나리오 2, 시나리오 5와 시나리오 8을 살펴보자. 이 시나리오들의 θ_{10}은 각가 1, 2, 5 로 θ_{10}이 커짐에 따라 발생 횟수가 증가하는 구간에서 두 번째 신뢰구간이 넓게 나타났다.

그림 5는 각 시점마다 199번 반복하여 얻은 포아송자료의 베이지

안 추정값 신뢰구간이다. 베이지안 추정값의 평균을 이용한 평균 적합선, 최대우도 추정법과 동일한 2 종류의 95% 경험적 신뢰구간을 구하였다. 그림 5를 살펴보자. 전반적으로 평균 적합선이 참값과 매우 유사하게 나타났다. 두 번째 경험적 신뢰구간은 평균 발생 횟수가 증가하기 시작하는 시점에서는 신뢰구간이 넓게, 변곡점을 지나 안정화되기 직전의 시점에서는 신뢰구간이 비교적 좁게 나타났다. $\theta_{10} = 2$, $\theta_{20} = 20$, $\theta_{30} = 40$인 시나리오 4, 시나리오 5와 시나리오 6을 살펴보자. 이 시나리오들의 θ_{40}는 각각 2, 5, 10으로 θ_{40}가 커짐에 따라 변곡점 이전 시점에서 두 번째 경험적 신뢰구간이 넓게 나타났다. $\theta_{20} = 20$, $\theta_{30} = 40$, $\theta_{40} = 5$인 시나리오 2, 시나리오 5와 시나리오 8을 살펴보자. 이 시나리오들의 θ_{10}은 각가 1, 2, 5로 θ_{10}이 커짐에 따라 발생 횟수가 증가하는 구간에서 두 번째 경험적 신뢰구간이 넓게 나타났다.

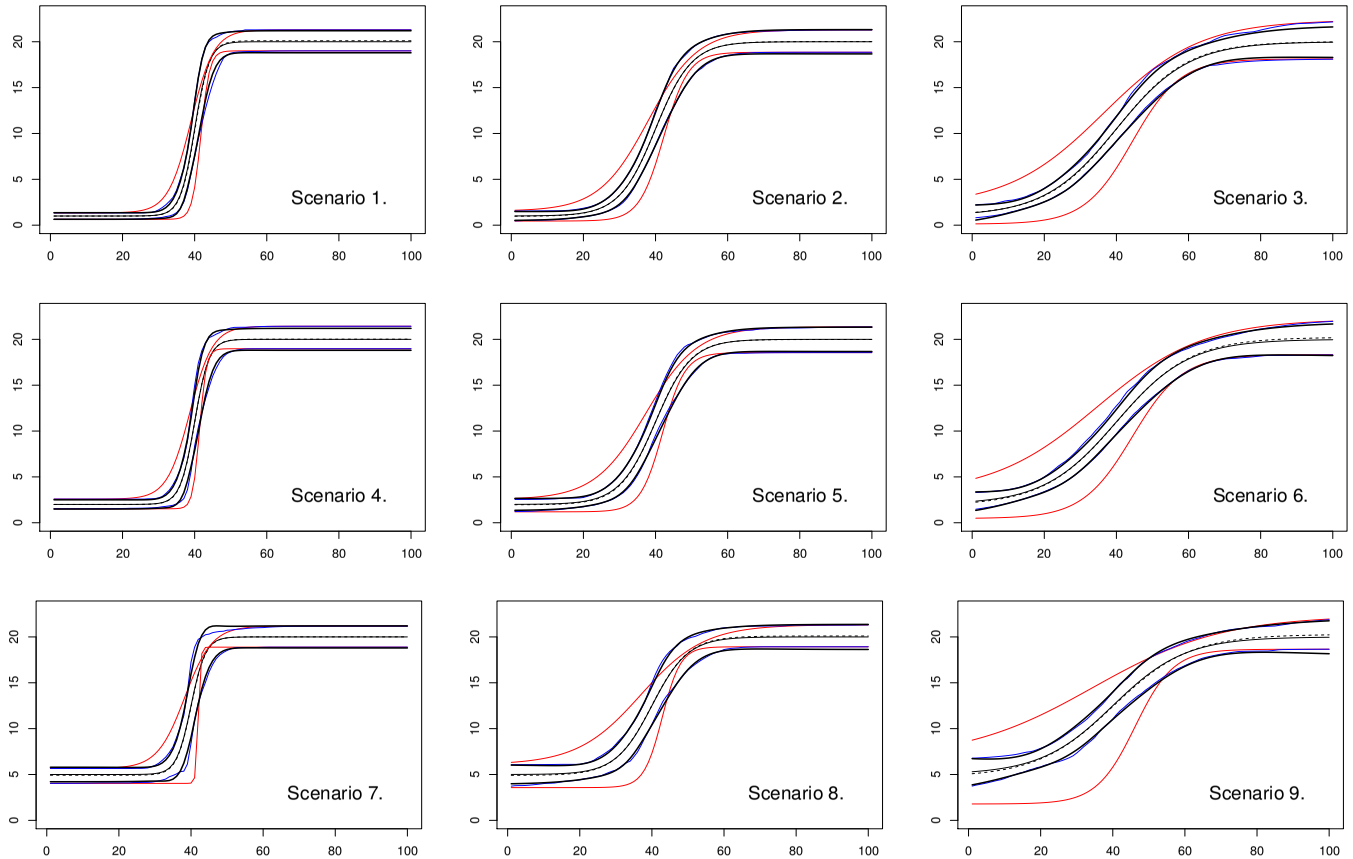

그림 4: 최대우도추정법을 이용한 참값과 평균 추정값, 그리고 95% 신뢰구간들(포아송분포)

(실선:참값; 점선:평균; 파란 실선:첫 번째 경험적 신뢰구간
빨간 실선:두 번째 경험적 신뢰구간, 굵은 실선:근사적 신뢰구간)

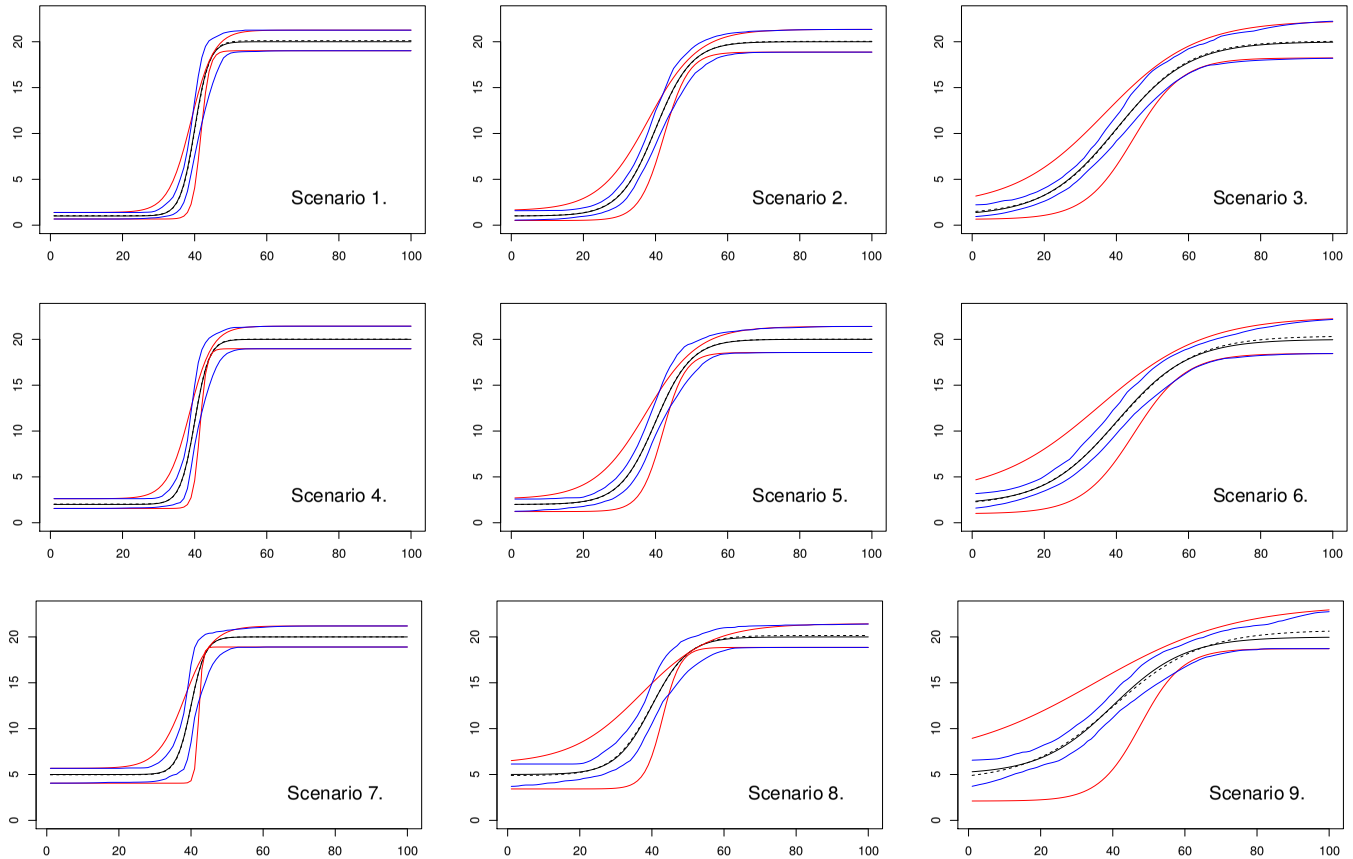

그림 5: 베이지안추정법을 이용한 참값과 평균 추정값, 그리고 95% 신뢰구간들(포아송분포)
(실선:참값; 점선:평균; 파란 실선:첫 번째 경험적 신뢰구간; 빨간 실선:두 번째 경험적 신뢰구간)

표 3은 음이항자료의 9개의 시나리오에 대한 모수들의 참값과 추정값이다. 추정방법으로는 최대우도 추정값과 베이지안 추정값들의 평균, 중위수, 분산과 평균제곱오차를 살펴보도록한다. 먼저 추정값들 중 추정 방법에 따른 평균과 중위수를 살펴보았다. 대부분의 추정값과 모수의 참값 간에 크게 차이가 나타나지 않았다. 분산과 평균제곱오차에 대해서 살펴보면 베이지안추정법에서의 시나리오 8과 시나리오 9를 제외한 모든 시나리오에서 θ_{30}의 분산과 평균제곱오차가 가장 크게 나타났다. 또한, θ_{10}과 θ_{40}의 값이 커짐에 따라 대부분 추정값의 분산과 평균제곱오차가 커짐을 확인할 수 있다. 최대우도추정법의 추정값들을 살펴보도록 하자. 먼저, θ_{10}을 고정시킨 경우를 살펴보도록 하자. 시나리오 7, 시나리오 8과 시나리오 9는 $\theta_{10} = 5$일 때, θ_{40}는 각각 2, 5, 10이다. θ_{40}가 커짐에 따라 모수의 분산과 평균제곱오차가 커짐을 확인할 수 있다. 특히, θ_{40}가 2에서 5로, 5에서 10으로 증가할 때, θ_{30}의 분산은 1.009에서 2.982로, 2.982에서 17.476으로 커졌고, θ_{40}의 분산은 0.803에서 1.851로, 1.851에서 13.990으로 크게 커짐을 확인할 수 있다.

다음으로는 θ_{40}을 고정시킨 경우를 살펴보도록 하자. 로 시나리오 3, 시나리오 6과 시나리오 9는 $\theta_{40} = 10$일 때, θ_{10}는 각각 1, 2, 5이다. θ_{10}이 커짐에 따라 대부분 모수의 분산과 평균제곱오차가 커짐을 확인할 수 있다. θ_{10} 이 1에서 2, 2에서 5로 증가할 때, θ_{10}의 분산은 0.289에서 0.520으로, 0.520에서 1.814로 증가하는 것을 확인할 수 있다. 그러나 θ_{20}의 분산은 0.997에서 0.672로, 0.672에서 1.500으로 감소하다가 증가하는 것을 확인할 수 있다. 나머지 θ_{30}와 θ_{40}의 분산은 θ_{10}이 커짐에 따라 분산과 평균제곱오차의 값이 커진다.

베이지안추정법의 추정값들을 살펴보도록 하자. 먼저, θ_{10}을 고정시킨 경우를 살펴보도록 하자. 시나리오 7, 시나리오 8과 시나리오 9는

표 3: 음이항자료를 음이항모형에 적용한 추정 결과

			Maximum Likelihood				Bayesian			
Scenario	Param	Truth	Mean	Median	Var	MSE	Mean	Median	Var	MSE
1	θ_{10}	1	0.998	0.996	0.032	0.032	1.034	1.027	0.032	0.033
	θ_{20}	20	20.015	20.060	0.412	0.412	20.028	20.072	0.413	0.414
	θ_{30}	40	40.014	39.972	0.548	0.548	39.998	39.967	0.548	0.548
	θ_{40}	2	1.939	1.946	0.267	0.270	1.935	1.942	0.262	0.266
2	θ_{10}	1	0.949	0.937	0.075	0.078	0.993	0.980	0.071	0.071
	θ_{20}	20	20.051	20.054	0.593	0.596	20.085	20.091	0.598	0.605
	θ_{30}	40	40.079	40.056	1.436	1.443	40.124	40.108	1.467	1.482
	θ_{40}	5	5.079	5.060	0.819	0.825	5.097	5.085	0.811	0.821
3	θ_{10}	1	1.056	1.082	0.289	0.292	1.173	1.114	0.163	0.192
	θ_{20}	20	20.013	19.940	0.997	0.997	20.088	20.027	1.009	1.017
	θ_{30}	40	39.972	40.012	3.840	3.841	40.254	40.240	4.013	4.077
	θ_{40}	10	9.713	9.562	2.818	2.900	9.691	9.673	2.127	2.222
4	θ_{10}	2	2.001	2.013	0.064	0.064	2.042	2.048	0.064	0.066
	θ_{20}	20	20.006	20.057	0.463	0.464	20.016	20.049	0.461	0.462
	θ_{30}	40	39.966	39.964	0.573	0.574	39.945	39.949	0.559	0.562
	θ_{40}	2	1.949	1.902	0.340	0.343	1.933	1.881	0.346	0.350
5	θ_{10}	2	1.982	1.982	0.119	0.119	2.012	2.019	0.120	0.120
	θ_{20}	20	20.049	20.021	0.583	0.585	20.091	20.071	0.587	0.595
	θ_{30}	40	40.045	39.969	1.673	1.675	40.086	39.994	1.697	1.704
	θ_{40}	5	4.892	4.865	0.864	0.876	4.955	4.917	0.940	0.942
6	θ_{10}	2	1.971	1.970	0.520	0.521	1.953	1.891	0.366	0.368
	θ_{20}	20	19.973	19.986	0.672	0.673	20.154	20.127	0.765	0.788
	θ_{30}	40	39.96	39.714	4.811	4.812	40.243	39.992	4.971	5.030
	θ_{40}	10	9.852	9.973	3.555	3.577	10.215	10.335	3.279	3.325
7	θ_{10}	5	4.953	4.946	0.171	0.173	5.004	5.002	0.172	0.172
	θ_{20}	20	20.091	20.069	0.393	0.401	20.089	20.080	0.397	0.405
	θ_{30}	40	39.915	39.845	1.009	1.016	39.862	39.810	1.018	1.037
	θ_{40}	2	2.017	1.893	0.803	0.803	1.956	1.832	0.836	0.838
8	θ_{10}	5	4.915	4.997	0.282	0.289	4.886	4.949	0.384	0.398
	θ_{20}	20	20.093	20.085	0.582	0.590	20.154	20.104	0.489	0.512
	θ_{30}	40	40.030	40.012	2.982	2.983	39.946	39.833	2.769	2.772
	θ_{40}	5	5.172	5.192	1.851	1.880	5.258	4.918	3.074	3.141
9	θ_{10}	5	4.398	4.600	1.814	2.177	4.031	3.991	1.250	2.189
	θ_{20}	20	20.342	20.223	1.500	1.618	20.899	20.802	2.036	2.845
	θ_{30}	40	39.233	39.370	17.476	18.065	39.806	39.472	13.517	13.555
	θ_{40}	10	10.803	10.187	13.990	14.635	12.437	12.290	15.378	21.319

Param,Parameter; Var,Variance; MSE,mean squared error

$\theta_{10} = 5$일 때, θ_{40}는 각각 2, 5, 10 이다. θ_{40}가 커짐에 따라 모수의 분산과 평균제곱오차가 커짐을 확인할 수 있다. 특히, θ_{40}가 2에서 5로, 5에서

10으로 증가할 때, θ_{30}의 분산은 1.018에서 2.769로, 2.769에서 13.517로 커졌고, θ_{40}의 분산은 0.836 에서 3.074로, 3.074에서 15.378로 크게 커짐을 확인할 수 있다. 다음으로는 θ_{40}을 고정시킨 경우를 살펴보도록 하자. 시나리오 3, 시나리오 6과 시나리오 9는 $\theta_{40} = 10$일 때, θ_{10}는 각각 1, 2, 5이다. θ_{10}이 커짐에 따라 대부분 모수의 분산과 평균제곱오차가 커짐을 확인할 수 있다. θ_{10}이 1에서 2, 2에서 5로 증가할 때, θ_{10}의 분산은 0.163에서 0.366으로, 0.366에서 1.250으로 증가하는 것을 확인할 수 있다. 그러나 θ_{20} 의 분산은 1.009에서 0.765로, 0.765에서 2.036으로 감소하다가 증가하는 것을 확인할 수 있다. 나머지 θ_{30}와 θ_{40}는 θ_{10}이 커짐에 따라 분산과 평균제곱오차의 값이 커진다.

다음은 각 시점마다 199번 반복하여 얻은 음이항자료의 신뢰구간들을 구하였다. 신뢰구간은 참값을 이용한 발생 횟수와 최대우도 추정값를 이용한 평균 적합선, 2종류의 95% 경험적 신뢰구간과 근사 신뢰구간을 구하였다. 평균 적합선은 베이지안추정법을 이용하여 구한 각 시나리오별 평균 추정값을 이용하여 구하였다. 2종류의 95% 경험적 신뢰구간은 식(3.1)과 식(3.2)를 이용하였다. 근사 신뢰구간은 식(3.4)를 이용하였다. 그림 6을 살펴보자. 전반적으로 평균 적합선이 참값과 매우 유사하게 나타났다. 첫 번째 경험적 신뢰구간과 근사 신뢰구간은 매우 유사하게 나타났다. 두 번째 경험적 신뢰구간은 평균 발생 횟수가 증가하기 시작하는 시점에서는 신뢰구간이 넓게, 안정화 되기 직전의 시점에서는 신뢰구간이 좁게 나타났다. $\theta_{10} = 5$, $\theta_{20} = 20$, $\theta_{30} = 40$ 인 시나리오 7, 시나리오 8과 시나리오 9를 살펴보자. 이 시나리오들의 θ_{40}는 각각 2, 5, 10으로 역시 θ_{40}가 커짐에 따라 변곡점 이전의 시점에서 두 번째 경험적 신뢰구간이 넓게 나타났다. $\theta_{20} = 20$, $\theta_{30} = 40$, $\theta_{40} = 10$ 인 시나리오 3, 시나리오 6 과 시나리오 9를 살펴보자. 이 시나리오들

의 θ_{10} 은 각가 1, 2, 5로 역시 θ_{10}이 커짐에 따라 발생 횟수가 증가하는 구간에서 두 번째 신뢰구간이 넓게 나타났다.

그림 7은 각 시점마다 199번 반복하여 얻은 음이항자료의 베이지안 추정값 신뢰구간이다. 베이지안 추정값의 평균을 이용한 평균 적합선, 최대우도추정법과 동일한 2종류의 95% 신뢰구간을 구하였다. 전반적으로 평균 적합선이 참값과 매우 유사하게 나타났다. 두 번째 경험적 신뢰구간은 평균 발생 횟수가 증가하기 시작하는 시점에서는 신뢰구간이 넓게, 변곡점을 지나 안정화 되기 직전의 시점에서는 신뢰구간이 비교적 좁게 나타났다. $\theta_{10} = 5, \theta_{20} = 20, \theta_{30} = 40$인 시나리오 7, 시나리오 8과 시나리오 9를 살펴보자. 이 시나리오들의 θ_{40}는 각각 2, 5, 10 으로 역시 θ_{40}가 커짐에 따라 변곡점 이전의 시점에서 두 번째 경험적 신뢰구간이 넓게 나타났다. $\theta_{20} = 20, \theta_{30} = 40, \theta_{40} = 10$인 시나리오 3, 시나리오 6과 시나리오 9를 살펴보자. 이 시나리오들의 θ_{10}은 각가 1, 2, 5 로 역시 θ_{10}이 커짐에 따라 발생 횟수가 증가하는 구간에서 두 번째 신뢰구간이 넓게 나타났다.

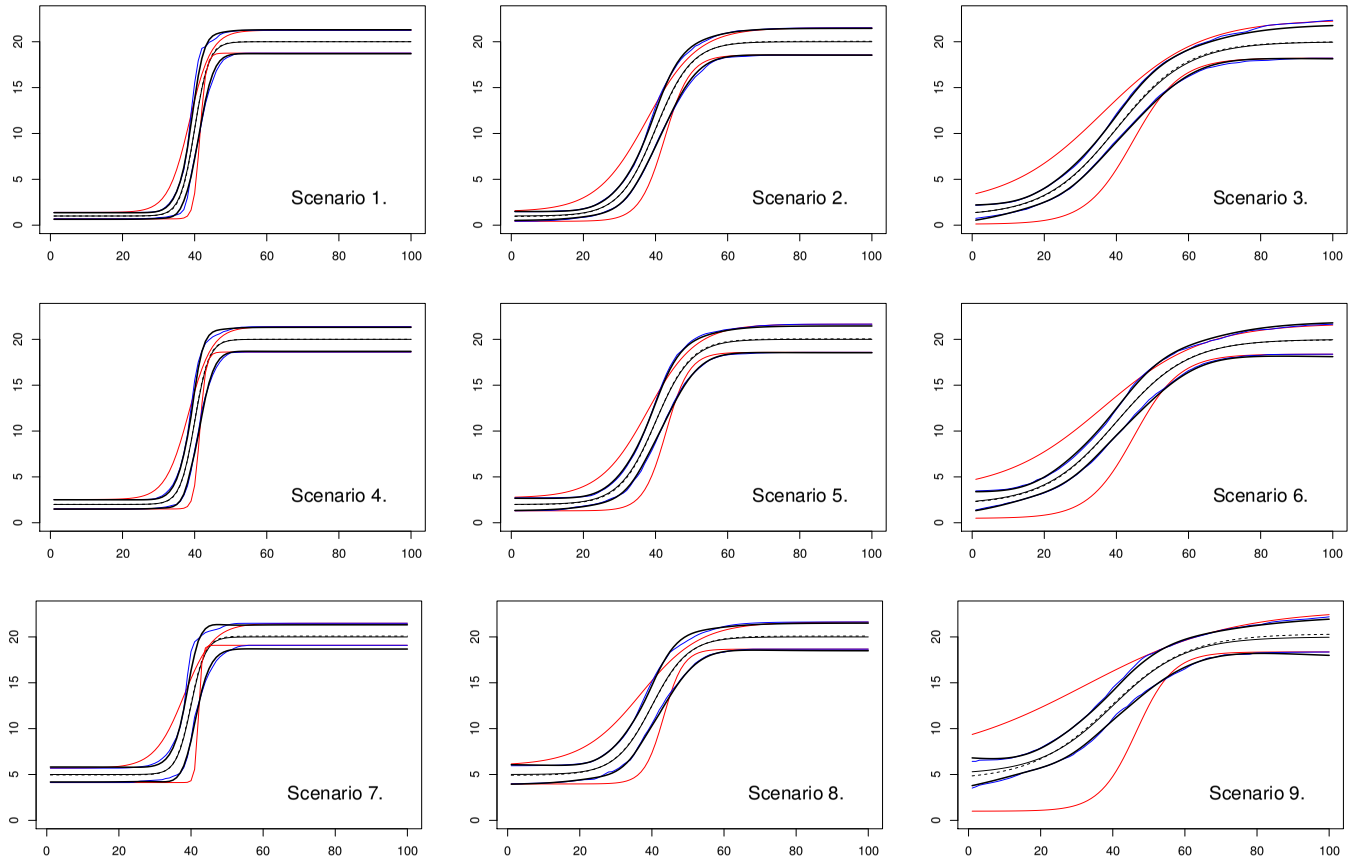

그림 6: 최대우도법을 이용한 참값과 평균 추정값, 그리고 95% 신뢰구간들(음이항분포)

(실선:참값; 점선:평균; 파란 실선:첫 번째 경험적 신뢰구간
빨간 실선:두 번째 경험적 신뢰구간, 굵은 실선:근사적 신뢰구간)

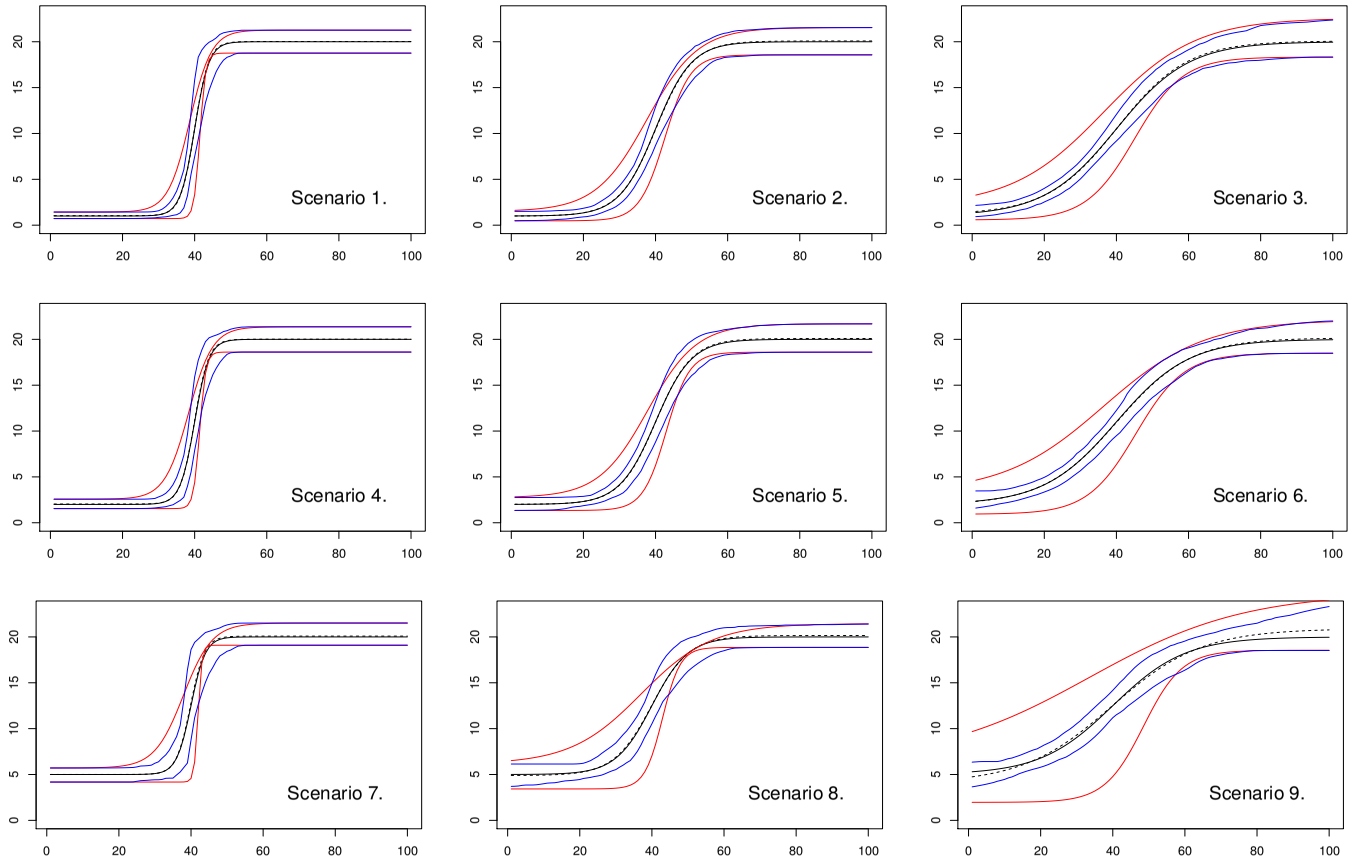

그림 7: 베이지안추정법을 이용한 참값과 평균 추정값, 그리고 95% 신뢰구간들(음이항분포)
(실선:참값; 점선:평균; 파란 실선:첫 번째 경험적 신뢰구간; 빨간 실선:두 번째 경험적 신뢰구간)

지금까지의 모의실험 결과를 정리하면 다음과 같다. 먼저 포아송 자료를 포아송모형에 적용시킨 추정결과를 살펴보자. 대부분의 시나리오에서 θ_{30}의 분산과 평균제곱오차가 가장 크게 나타났다. 또한 θ_{10}과 θ_{40}이 증가함에 따라 대부분의 모수의 분산과 평균제곱오차가 커짐을 확인할 수 있었다. 다음으로 최대우도추정값을 이용한 신뢰구간과 베이지안추정값을 이용한 신뢰구간을 확인하였다. 평균 적합값은 참값과 매우 유사하게 나타났고, 첫 번째 경험적 신뢰구간과 근사적 신뢰구간은 거의 유사하게 나타났다. 두 번째 경험적 신뢰구간은 발생 횟수가 증가하는 시점에서는 신뢰구간이 넓게, 안정화 되기 직전의 시점에서는 신뢰구간이 좁게 나타났다. 다음은 음이항자료를 음이항모형에 적용시킨 추정결과를 살펴보자. 대부분의 시나리오에서 θ_{30}의 분산과 평균제곱오차가 가장 크게 나타났다. 또한 θ_{10}과 θ_{40}이 증가함에 따라 대부분의 모수의 분산과 평균제곱오차가 커짐을 확인할 수 있었다. 다음으로 최대우도추정결과를 이용한 신뢰구간과 베이지안 추정결과를 이용한 신뢰구간을 확인하였다. 평균 적합값은 참값과 매우 유사하게 나타났고, 첫 번째 경험적 신뢰구간과 근사적 신뢰구간은 거의 유사하게 나타났다. 두 번째 경험적 신뢰구간은 발생 횟수가 증가하는 시점에서는 신뢰구간이 넓게, 안정화 되기 직전의 시점에서는 신뢰구간이 좁게 나타났다.

3.2.2 분포와 추정법 간의 비교

다음으로는 모의실험에 사용한 분포와 추정법에 대해서 모형간 비교를 해보도록 한다. 먼저 추정법을 비교하기 위한 평가방법으로 4가지 방법을 이용하였다. 4 가지 방법으로는 중위절대편차(Median

Absolute Deviation; MAD), 상대표준편차(Relative Standard Deviation; RSD), 2종류의 편향(Bias)을 이용하였다. 다음으로는 포아송자료와 음이항자료를 각각 음이항모형과 포아송모형에 적합시켜보도록한다. 포아송자료를 음이항모형에 적합시켰을 때, 최대우도추정법에서의 AIC(Akaike information criterion) 와 BIC(Bayesian information criterion)를, 베이지안 추정법에서의 DIC(Deviance information criterion)를 이용하여 정분류율 비교해보도록 한다.

3.2.2.1 예측 평가 기준을 통한 추정법 비교

모형의 추정법을 평가하는 방법에는 여러가지가 있다. 그 중에서도 중위절대편차, 상대표준편차와 2종류의 편향의 절대값을 이용하여 동일한 분포에서의 추정법을 평가하였다. 첫 번째 방법은 중위절대편차로 다음과 같이 나타낼 수 있다.

$$\text{MAD} = median\left|\lambda_{t0} - \bar{\hat{\lambda}}_t\right|.$$

두 번째 방법은 상대표준편차로 변동계수(Coefficient of Variability; CV)라고도 불리우며, 다음과 같이 나타낸다.

$$\text{RSD} = \frac{1}{\bar{\hat{\lambda}}_t}\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\lambda}_t^{(b)} - \bar{\hat{\lambda}}_t\right)^2}.$$

세 번째 방법은 편향의 절대값으로 추정값의 평균과 모수의 참값의 차이를 이용하였다.

$$\text{Bias 1} = \left|\bar{\hat{\lambda}}_t - \lambda_{t0}\right|. \tag{3.5}$$

표 4: 예측 평가 기준을 통한 추정법 비교 결과

		Poisson data				Negative-binomial data			
		MAD	RSD	Bias 1	Bias 2	MAD	RSD	Bias 1	Bias 2
1	MLE	0.099	0.093	0.069	0.069	0.015	0.099	0.027	0.037
	BE	0.108	0.091	0.093	0.079	0.028	0.096	0.043	0.052
2	MLE	0.017	0.095	0.020	0.032	0.035	0.100	0.037	0.048
	BE	0.041	0.092	0.047	0.039	0.046	0.097	0.047	0.051
3	MLE	0.026	0.078	0.036	0.049	0.035	0.079	0.049	0.068
	BE	0.133	0.073	0.110	0.055	0.071	0.074	0.078	0.078
4	MLE	0.015	0.076	0.029	0.038	0.006	0.080	0.026	0.039
	BE	0.056	0.076	0.058	0.051	0.020	0.079	0.046	0.057
5	MLE	0.031	0.076	0.041	0.059	0.036	0.080	0.039	0.052
	BE	0.026	0.075	0.031	0.039	0.043	0.079	0.050	0.057
6	MLE	0.036	0.069	0.058	0.090	0.053	0.072	0.047	0.088
	BE	0.047	0.066	0.096	0.095	0.036	0.069	0.043	0.044
7	MLE	0.030	0.059	0.050	0.072	0.088	0.062	0.083	0.102
	BE	0.050	0.059	0.053	0.056	0.087	0.062	0.085	0.101
8	MLE	0.045	0.062	0.054	0.093	0.060	0.064	0.057	0.082
	BE	0.079	0.063	0.075	0.107	0.080	0.063	0.082	0.108
9	MLE	0.050	0.059	0.069	0.122	0.074	0.075	0.089	0.206
	BE	0.165	0.063	0.231	0.207	0.178	0.063	0.226	0.281

MLE,maximum likelihood; BE, Bayesian estimation, Bias 1, Equation(3.5); Bias 2, Equation(3.6).

네 번째 방법은 경험적인 방법이다. 2종류의 경험적 신뢰구간 중 두 번째 신뢰구간은 각각의 $\widehat{\theta}_i$의 순서통계량을 이용하였다. 이 경험적 방법도 $\widehat{\theta}_i$의 순서통계량을 이용하여 계산하였다. 경험적 방법으로 구한 편향의 절대값 식은 다음과 같다.

$$\text{Bias } 2 = \left| \overline{\widehat{\lambda^*}_t} - \lambda_{t0} \right|. \tag{3.6}$$

여기서, $\overline{\widehat{\lambda^*}_t}$는 식(3.3)의 평균이다. 표 4는 추정법 비교를 위한 4가지의 예측 평가 기준값들이다. 상대표준편차, Bias 1과 Bias 2는 각 값들의 평균을 이용하였다. 모든 시나리오의 평가 기준값들 전부 0.3를 넘지 않는 값을 갖는다. 자료의 분포에 따른 시나리오를 살펴보자. 시나리오

1을 살펴보면 포아송자료와 음이항자료 모두 상대표준편차를 제외한 나머지 3개의 기준값이 최대우도추정법에서 더 작게 나타났다. 시나리오 9에서 포아송자료는 모든 기준값에 최대우도추정법에서 더 작게 나타났으며, 음이항자료는 상대표준편차를 제외한 기준값에서 더 작게 나타났다.

그림 8부터 그림 11은 모의실험의 추정결과를 이용하여 Bias 1과 Bias 2를 그래프로 나타낸 것이다. $y = 0$을 축으로 하여 위 쪽은 Bias 1이고, 아래 쪽은 Bias 2이다. 그림 8과 그림 9는 포아송자료를 포아송모형에 적용한 최대우도추정법과 베이지안추정법의 결과이고, 그림 10과 그림 11은 음이항자료를 음이항모형에 적용한 최대우도추정법과 베이지안추정법의 결과이다. 먼저 그림 8과 그림 9를 살펴보자. 두 그림에서 모두 θ_4가 증가함에 따라 각가의 증가율 시점의 간격이 커지는 것을 확인할 수 있었다. 즉 시나리오 1에서 시나리오 2, 시나리오 3으로 θ_4가 커질 때, 증가율의 시점 간격이 점점 넓어졌다. θ_3를 기준으로 그래프들을 살펴보면 대부분의 시나리오에서 θ_3에서의 두 절대 편향의 변동폭이 증가하는 것을 확인할 수 있다. 최대우도추정법에서 시나리오 7을 살펴보면 시점이 40, 즉 발생 횟수의 증가율이 50%인 시점에서의 Bias 2의 변동폭이 모든 시나리오의 모든 시점을 통틀어서 가장 큰 변동폭을 나타내고 있다. 다음으로 Bias 1과 Bias 2를 비교해보자. 최대우도추정법에서는 Bias 1과 Bias 2가 모든 시점에서 0.6보다 작게 나타났으며, 베이지안추정법에서는 모든 시점에서 1.0보다 작게 나타났다. 또한, 베이지안의추정법에서 Bias 2가 Bias 1에 비해서 변동폭이 큰 경우가 있다. 그림 9의 시나리오 3을 살펴보면 시점이 40인 이후에 Bias 2의 편향이 더 작게 나타나는 것을 확인할 수 있다.

다음으로 음이항자료를 음이항모형에 적용시킨 결과에 대해 살펴

보도록 하자. 그림 10과 그림 11에서 모두 θ_4가 증가함에 따라 각가의 증가율 시점의 간격이 커지는 것을 확인할 수 있었다. 즉 시나리오 1에서 시나리오 2, 시나리오 3으로 θ_4가 커질 때, 증가율의 시점 간격이 점점 넓어졌다. 다음으로 θ_3를 기준으로 그래프들을 살펴보면 대부분의 시나리오에서 θ_3에서의 두 절대 편향의 변동폭이 증가하는 것을 확인할 수 있다. 최대우도추정법에서 시나리오 7을 살펴보면 시점이 40, 즉 발생 횟수의 증가율이 50%인 시점에서의 Bias 2의 변동폭이 모든 시나리오의 모든 시점을 통틀어서 가장 큰 변동폭을 나타내고 있다. 또한, 시나리오 9의 Bias 2 역시 시점이 40일 때 변동이 크게 나타나는 모습을 보였다. 다음으로 Bias 1과 Bias 2를 비교해보자. 최대우도추정법에서는 Bias 1과 Bias 2가 모든 시점에서 0.6보다 작게 나타났으며, 베이지안추정법에서는 모든 시점에서 1.0보다 작게 나타났다.

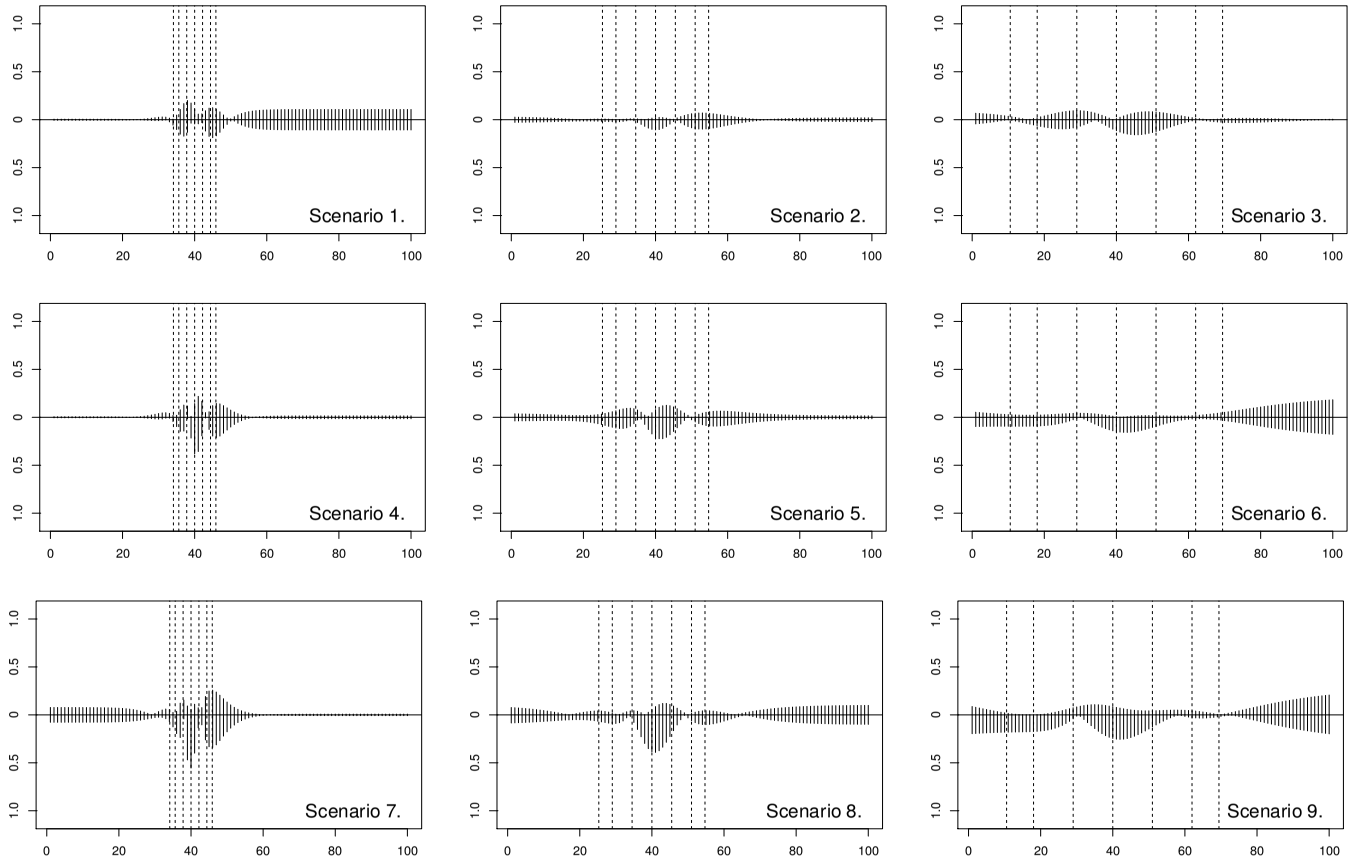

그림 8: 포아송자료를 포아송모형에 적용한 최대우도추정값의 예측 평가 기준 추정값
(왼쪽 점선부터 증가율이 5%, 10%, 25%, 50%, 75%, 90%, 95%인 시점)

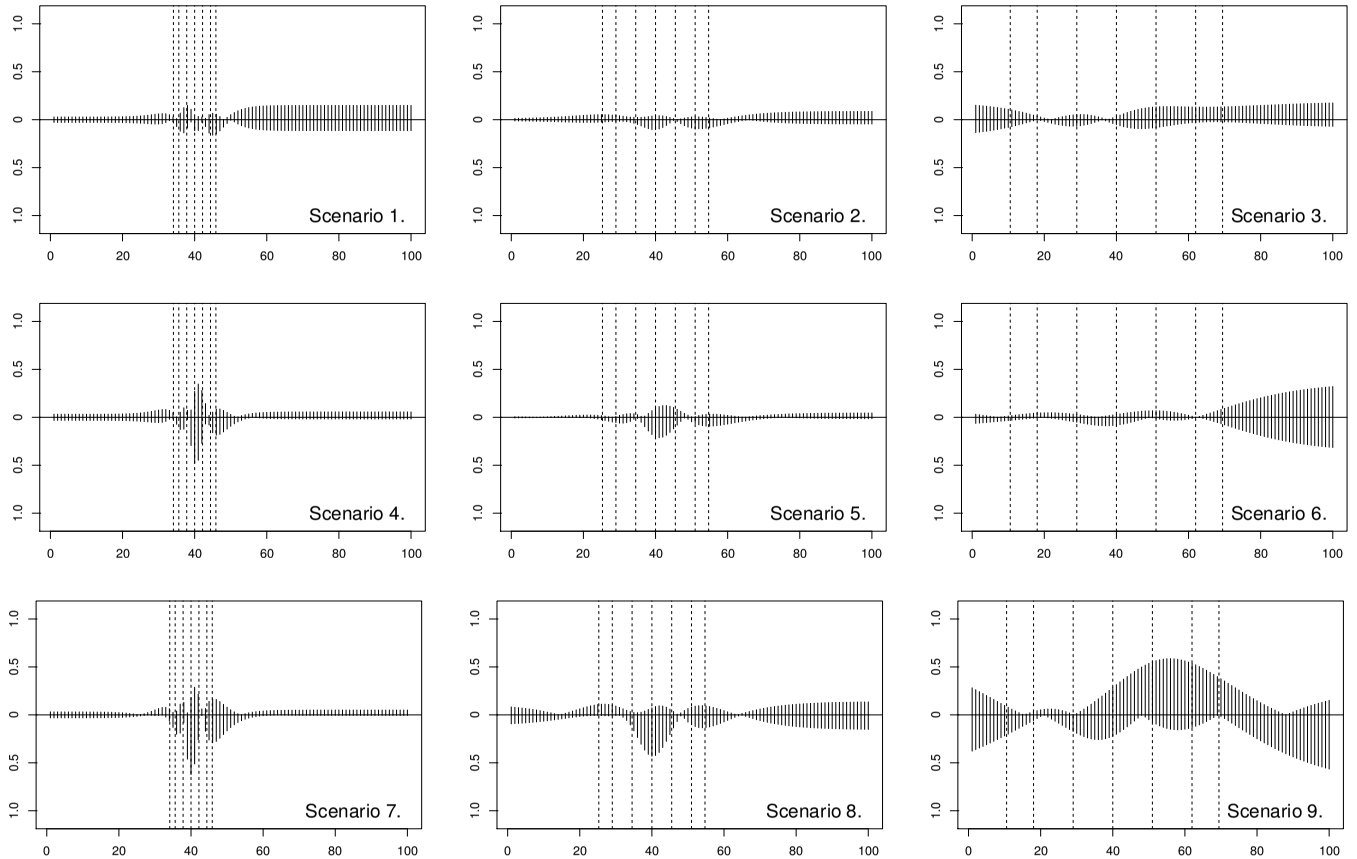

그림 9: 포아송자료를 포아송모형에 적용한 베이지안 추정값의 예측 평가 기준 추정값
(왼쪽 점선부터 증가율이 5%, 10%, 25%, 50%, 75%, 90%, 95%인 시점)

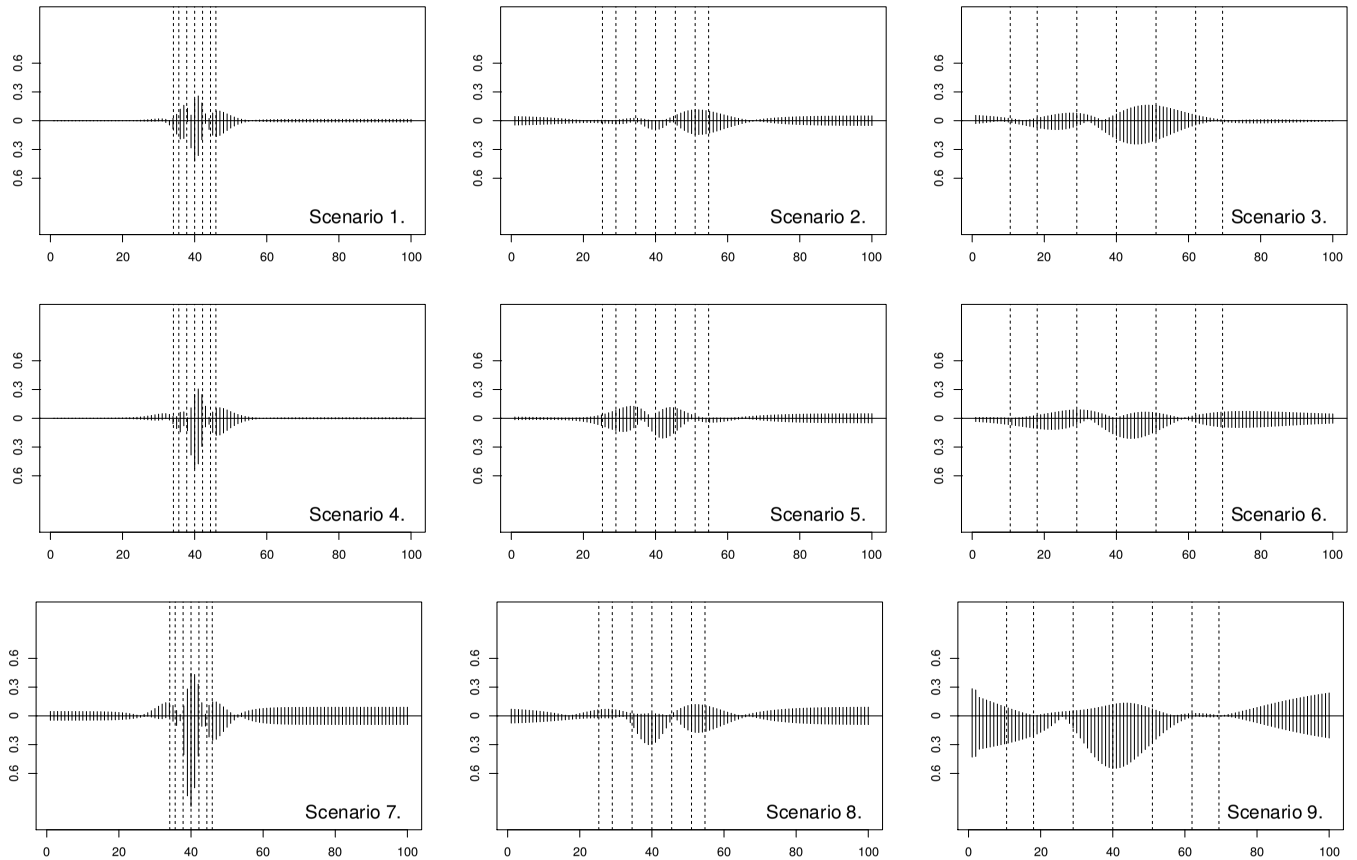

그림 10: 음이항자료를 음이항모형에 적용한 최대우도추정값의 예측 평가 기준 추정값
(왼쪽 점선부터 증가율이 5%, 10%, 25%, 50%, 75%, 90%, 95%인 시점)

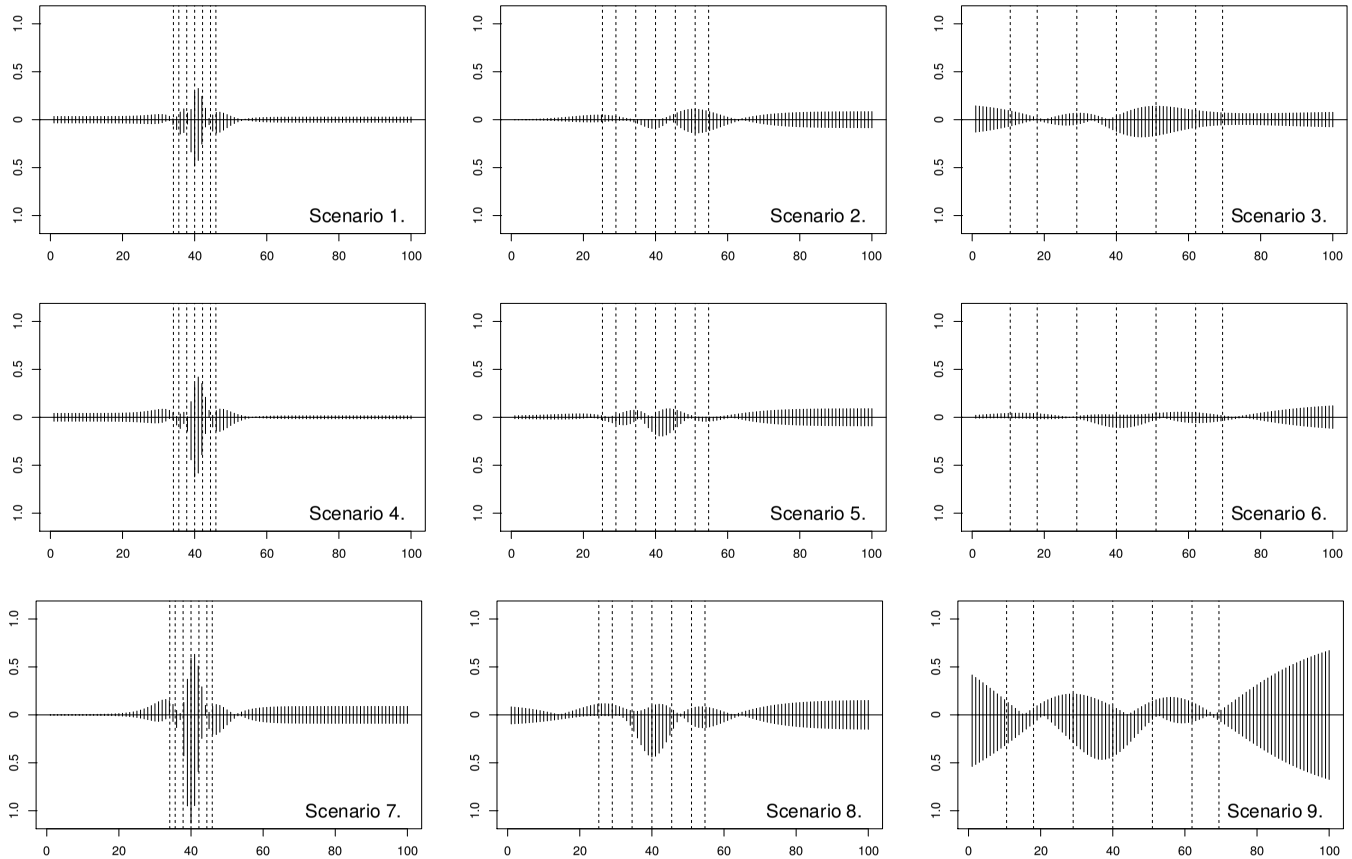

그림 11: 음이항자료를 음이항모형에 적용한 베이지안 추정값의 예측 평가 기준 추정값
(왼쪽 점선부터 증가율이 5%, 10%, 25%, 50%, 75%, 90%, 95%인 시점)

표 5: 오분류 평가에 의한 정분류율

	Poisson → Poisson			NB → NB		
Sce.	AIC	BIC	DIC	AIC	BIC	DIC
1	0.779	0.779	0.719	0.598	0.593	0.628
2	0.789	0.789	0.794	0.643	0.638	0.658
3	0.729	0.724	0.729	0.508	0.513	0.528
4	0.814	0.814	0.683	0.611	0.616	0.623
5	0.789	0.789	0.774	0.528	0.523	0.523
6	0.819	0.814	0.774	0.603	0.603	0.618
7	0.814	0.824	0.683	0.593	0.593	0.583
8	0.754	0.754	0.648	0.613	0.613	0.628
9	0.849	0.849	0.566	0.628	0.633	0.648

Sce, Scenario; NB, Negative binomial.

3.2.2.2 오분류 평가

모의실험에서 포아송자료를 만들어 포아송모형에 적합을 시켜 분석하였다. 오분류 평가에서는 포아송자료를 음이항모형에, 음이항자료를 포아송모형에 적합시켜 보도록 한다. 사전에 실시한 모의실험과 동일하게 최대우도추정법과 베이지안추정법을 이용하였고, 평가 기준은 최대우도추정법은 AIC와 BIC를, 베이지안추정법은 DIC를 기준으로 하였다.

표 5를 살펴보면 제시된 값은 전체 199개 중 오분류 평가를 이용한 정분류의 비율을 나타낸 값이다. 한 예제로 시나리오 1의 포아송 AIC의 정분류율은 0.779 이다. 이 값은 포아송자료를 포아송모형에 적용시켰을 때의 199개의 AIC와 음이항모형에 적용시켰을 때의 199개의 AIC를 비교한 후, 포아송모형의 AIC가 더 작은 경우의 비율로 구할 수 있다. 포아송자료 199개 중 155개의 AIC가 포아송모형에서 더 작게 나타났다. 즉, 약 77.8%가 자료의 분포와 동일한 모형에 잘 적합하였다는 결과를 나타냈다. 나머지 값들도 살펴보도록 하자. 먼저 포아송자료의 결과이다. 시나리오 9를 제외한 나머지 시나리오들이 AIC, BIC와 DIC

의 정분류율이 비슷하게 나타났다. 먼저 AIC의 경우 모든 정분류율이 0.7 이상으로 나타났다. 즉, 199개의 자료 중 70% 이상의 자료가 포아송자료를 포아송모형에 적합하다는 결과를 나타냈다. 다음으로 BIC의 경우도 AIC 와 동일하다. 70% 이상의 자료가 포아송자료를 포아송모형에 적합한 결과가 더 좋음을 나타냈다. 다음으로 DIC의 결과이다. DIC의 경우는 AIC와 BIC에 비해서 작은 값을 나타냈다. 시나리오 9를 제외한 나머지 시나리오에서는 60% 이상의 자료가 모형을 잘 적합하였음을 나타냈다. 다음으로 음이항자료의 결과이다. 모든 시나리오에서 AIC, BIC와 DIC의 정분류율이 비슷하게 나타났다. AIC, BIC와 DIC의 경우 모든 값들이 0.5 이상으로 나타났다. 즉, 199개의 자료 중 50%이상의 자료가 음이항자료를 음이항모형에 적합하다는 결과를 나타냈다. 정리하면 최대우도추정법에서 199개의 자료 중 포아송자료를 포아송모형에 적용하는 것이 적합하다는 결과는 70% 이상이지만, 음이항자료를 음이항모형에 적용하는 것이 적합하다는 결과는 50% 이상으로 비교적 작게 나타났다.

3.2.2장에서는 방법론 평가와 오분류 평가를 하였다. 방법론 평가에서는 중위절대편차, 상대표준편차, Bias 1(식(3.5) 참조)와 Bias 2(식(3.6) 참조)를 이용하여 방법론을 비교하였다. 그 결과, 포아송자료와 음이항자료 둘 다 베이지안추정법보다 최대우도추정법에서의 추정값들이 작은 값을 가지는 경우가 더 많은 것을 확인할 수 있었다. 추정값들 중 Bias 1과 Bias 2의 결과를 시점별로 비교해 보았다. 자료와 추정법에 상관없이 대부분의 시나리오에서 θ_3에서 변동이 커지는 것을 확인할 수 있었다. 또한, θ_4가 커짐에 따라 증가율의 시점 간격이 넓어지는 것도 확인할 수 있었다. 다음으로 오분류 평가에 의한 정분류율을 살펴보았다. 최대우도추정법에서 약 70% 이상 포아송자료를 포

아송모형에 적용하는 것이 적합하다는 결과가 나왔다. 음이항자료를 음이항모형에 적용시킨 결과를 살펴보면 최대우도추정법에서 약 50% 이상 적합하다는 결과를 나타냈다. 이는 포아송자료를 포아송모형의 결과와 비교하였을 때보다 비교적 작은 것을 알 수 있었다.

3.3 결론

3.2.1장과 3.2.2장의 결과를 정리하면 다음과 같다. 3.2.1장에서의 모의실험 결과에서 포아송자료를 포아송모형에 적용시킨 결과와 음이항자료를 음이항모형에 적용시킨 결과 모두에서 평균 추정값이 모수와 유사하게 나타났다. 또한 θ_1과 θ_4를 증가시켰을 때, 대부분의 분산과 평균제곱오차의 값이 커지는 현상을 확인할 수 있었다. 모의실험 추정결과값들을 이용하여 참값, 평균 적합선과 95% 신뢰구간을 구하였다. 신뢰구간을 구할 때에는 2종류의 경험적 신뢰구간와 근사적 신뢰구간을 이용하였다. 최대우도추정법에서의 그래프를 살펴보면 평균 적합선은 참값과 매우 유사하게 나타났다. 첫 번째 경험적 신뢰구간와 근사적 신뢰구간은 비슷하게 나타났다. 두 번째 경험적 신뢰구간은 평균 발생 횟수가 증가하는 시점에서 신뢰구간이 넓게 나타났으며, 변곡점을 지나 안정화 구간에 되기 전에는 신뢰구간이 매우 좁은 형태로 나타났다. 베이지안추정법에서의 그래프도 역시 평균 적합선은 참값과 매우 유사하게 나타났다. 두 번째 경험적 신뢰구간은 신뢰구간은 평균 발생 횟수가 증가하는 시점에서 신뢰구간이 넓게 나타났으며, 변곡점을 지나 안정화 구간에 되기 전에는 신뢰구간이 매우 좁은 형태로 나타났다. 다음으로 모형 평가를 하였다. 방법론 평가에서는 중위절대편차, 상대표준편차, Bias 1과 Bias 2를 이용하여 비교하였다. 결과로는 포아

송자료를 포아송모형에 적용시킨 결과와 음이항자료를 음이항모형에 적용시킨 결과 모두 베이지안추정법에 비해 최대우도추정법에서 더 적합하다는 결과가 나타났다. 추정값 중 Bias 1과 Bias 2를 비교하여 보았다. 그 결과 추정방법과 자료의 분포에 상관없이 대부분의 시나리오에서 θ_3가 시점 40, 즉 변곡점에서 변동이 크게 나타나는 것을 확인할 수 있었다. 다음으로 오분류 평가를 이용하여 정분류율을 살펴보았다. 최대우도추정법에서 포아송자료를 포아송모형에 적용시키는 것이 적합다는 결과은 약 70% 이상으로 나타났다. 그에 비해 음이항자료를 음이항모형에 적용시키는 것이 적합하다는 결과는 약 50%로 포아송자료를 포아송모형에 적용시킨 결과보다 낮은 값을 나타냈다.

제 4 장

실증연구

제3장에서 모의실험을 통하여 학습곡선을 모형화 하였다. 제4장에서는 실제자료에 모수 추정을 위한 분석방법을 직접 적용하여 학습곡선을 모형화 하고자 한다. 첫 번째 자료는 영국 광산 사고 자료로, 시간이 지남에 따라 사고 건수가 점차 줄어들다가 안정화 되는 현상을 나타낸다. 두 번째와 세 번째 자료는 국내 내륙과 해역 지진 발생 건수 자료이다(http://www.kma.go.kr/weather/earthquake/domesticlist.jsp). 지진 발생 건수와 같은 통제 불가능한 자연현상에 대해서는 실질적으로 학습곡선을 적용시키는데 어려움이 있다. 본 논문에서의 지진 자료를 이용한 실증연구는 1978년부터 2012년까지의 지진의 추이만을 고려하여 학습곡선을 적용시킨다. 따라서 추후의 지진 발생 건수가 분석 결과에 의한 안정화 된 이후의 평균 발생 건수와 매우 크게 차이가 날 수 있음을 참고하길 바란다.

4.1 영국 광산 사고 자료

영국 광산 사고 자료(그림 12)는 1851년부터 1962년까지의 122년간의 영국 광산 사고 건수이다(Bradley *et al.*, 1992). 최소 발생 건수는 0, 최대 발생건수는 6이다. 그림 12를 살펴보면, 영국 광산 사고 건수 자료는 모의실험과는 다르게 사고 건수가 점점 감소하는 형태를 띄고 있다. 따라서 이 자료에 대해서는 평균 발생 횟수인 $\lambda_t(\Theta)$에 대해서 누적분

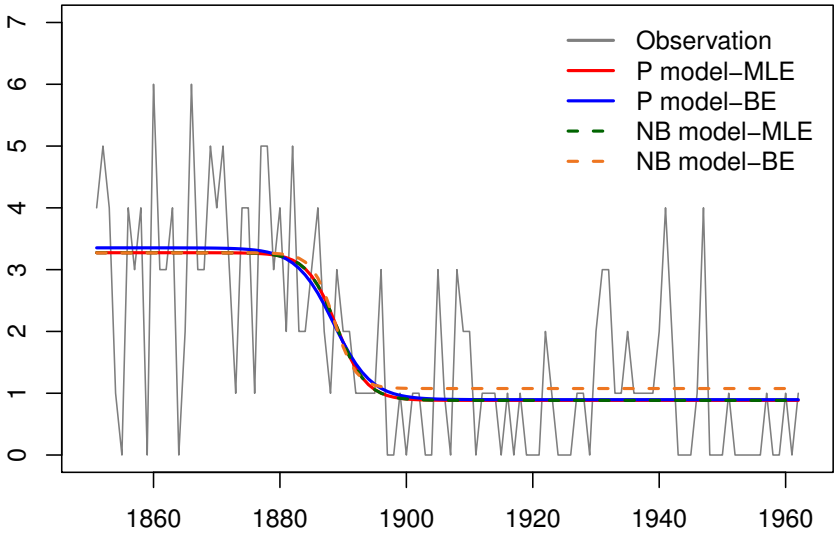

그림 12: 1851 - 1962년의 영국 광산 사고 추이 및 추정 결과

포함수 $F(\cdot)$ 대신에 $1-F(\cdot)$을 사용하기로 한다(보다 자세한 내용은 이슬지 (2011)를 참조하길 바란다).

표 6은 최대우도추정법과 베이지안추정법을 포아송분포와 음이항분포를 적용하여 분석한 결과이다. 포아송분포를 적용하였을 때, 초기 시점의 광산 사고 발생 횟수를 살펴보면 최대우도추정값과 베이지안 추정값은 각각 3.274와 3.353으로 베이지안추정법이 약간 더 크게 나타났다. 음이항분포를 적용하였을 때, 시간이 지난 후 안정화 되었을 때의 평균 사고 발생 횟수를 살펴보면 최대우도추정값과 베이지안 추정값은 각각 0.885와 1.076으로 베이지안추정법이 더 크게 나타났다. 음이항분포의 θ_1과 θ_4의 베이지안 추정값을 제외하고 나머지 추정값들은 서로 다 비슷한 값을 나타냈다. 다음으로 어떤 모형이 적합한지를 살펴보도록 하자. 먼저 포아송분포와 음이항분포의 최대우도추

표 6: 영국 광산 사고 자료 추정 결과

		Maximum likelihood		Bayesian	
Model	Param	Estimate	95% C.I.	Estimate	95% C.I.
	θ_1	0.884	0.656 - 1.113	0.896	0.616 - 1.155
	θ_2	3.274	2.635 - 3.914	3.353	2.669 - 4.619
	θ_3	38.971	33.985 - 43.956	38.523	26.899 - 44.760
Poisson	θ_4	2.368	-0.974 - 5.711	2.965	0.044 - 17.280
	AIC	343.8			
	BIC	354.7			
	DIC			338.981	
	θ_1	0.885	0.655 - 1.114	1.076	1.002 - 1.238
	θ_2	3.274	2.624 - 3.924	3.272	2.662 - 4.029
	θ_3	38.972	33.930 - 44.015	38.580	32.150 - 44.000
NB	θ_4	2.367	-1.011 - 5.745	1.749	0.039 - 7.146
	AIC	344.1			
	BIC	354.9			
	DIC			340.866	

Param,Parameter; NB,Negative binomial; C.I.,Confidence Interval

정량을 비교하기 위해서 AIC와 BIC를 사용하였다. 비교 결과, 포아송 분포의 AIC와 BIC가 더 작게 나타났다. 두 번째로 베이지안추정법을 비교하기 위하여 DIC를 사용하였다. 그 결과 포아송분포의 DIC값이 더 작게 나타났다. 즉, 영국 광산 자료에 대해서는 포아송분포가 더 적합하다고 할 수 있다.

위의 결과를 다른 논문과 비교하고자 한다. Kim and Cheon (2013)은 본 자료에 대하여 구조 변화 시점(structural change-point)이 몇 번째 시점에서 나타나는지를 추정하였다. 추정 결과, $t = 41$에서 구조 변화가 일어났을 것이라 예측하였다. 또한, Bradley *et al.* (1992) 역시 동일한 자료에 대한 구조 변화 시점의 추정 결과, 39, 40, 41로 나타났으며, 이는 1889년 말부터 1892년 초 사이에 변화가 일어났을 것이라 추정하였다. 두 논문의 결과와 본 논문에서 적용한 최대우도 추정값과 베이지안 추

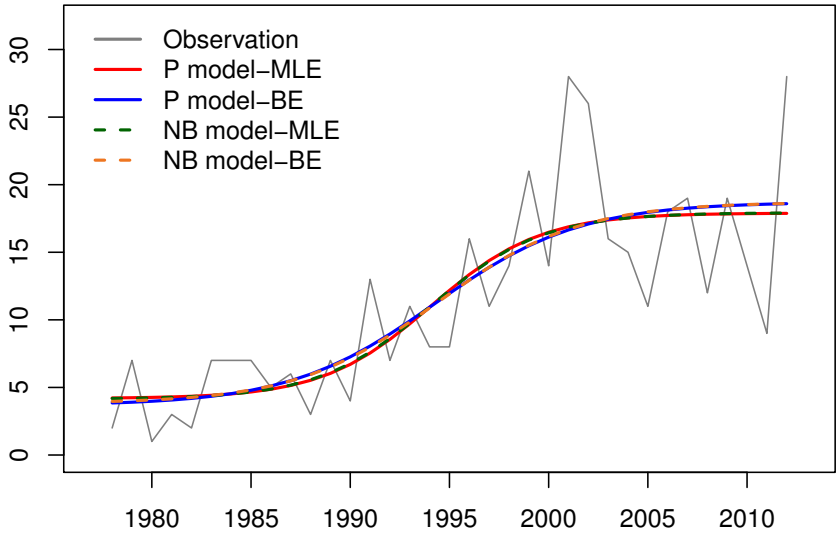

그림 13: 1978 - 2012년의 국내 내륙 지진 발생 추이 및 추정 결과

정값을 비교하였을 때 추정값이 근접하게 나타남을 확인할 수 있었다.

4.2 국내 내륙 지진 자료

국내 내륙 지진 자료(그림 13)는 1978년부터 2012년까지 35년 간 국내에서 발생한 지진 목록 중 진앙지가 내륙(남한)인 자료(건수)로, 기상청 홈페이지에서 확인할 수 있다.

표 7은 최대우도추정법과 베이지안추정법을 포아송모형과 음이항모형을 적용하여 분석한 결과와 예측력 비교를 위한 추정값이다. 포아송분포를 적용하였을 때, 초기 시점의 지진 발생 횟수를 살펴보면 최대우도추정값과 베이지안 추정값은 각각 4.183과 3.676으로 최대우도추정법이 더 크게 나타났다. 음이항분포를 적용하였을 때, 시간이

표 7: 국내 내륙 지진 자료 추정 결과

		Maximum likelihood		Bayesian	
Model	Param	Estimate	95% C.I.	Estimate	95% C.I.
	θ_1	4.183	2.212 - 6.155	3.676	0.478 - 6.007
	θ_2	17.898	15.338 - 20.457	18.712	15.720 - 24.520
	θ_3	17.080	14.463 - 19.697	17.255	13.770 - 21.260
Poisson	θ_4	2.739	0.561 - 4.917	3.664	0.965 - 8.705
	AIC	201.1			
	BIC	207.5			
	DIC			181.627	
	θ_1	4.139	2.057 - 6.222	3.816	1.285 - 6.087
	θ_2	17.926	15.116 - 20.736	18.712	15.530 - 24.760
	θ_3	17.064	14.279 - 19.849	17.374	13.820 - 21.820
NB	θ_4	2.808	0.431 - 5.184	3.561	0.859 - 8.127
	AIC	191.3			
	BIC	205.5			
	DIC			180.749	

Param,Parameter; NB,Negative binomial; C.I.,Confidence Interval

지난 후 안정화 되었을 때의 평균 지진 발생 횟수를 살펴보면 최대우도추정값과 베이지안 추정값은 각각 17.926과 18.712 로 베이지안추정법이 크게 나타났다. 다음으로 어떤 모형이 적합한지를 살펴보도록 하자. 먼저 포아송분포와 음이항분포의 최대우도추정량을 비교하기 위해서 AIC와 BIC를 사용하였다. 비교 결과, 음이항분포의 AIC와 BIC가 더 작게 나타났다. 두 번째로 베이지안추정법을 비교하기 위하여 DIC를 사용하였다. 그 결과 음이항분포의 DIC값이 더 작게 나타났다. 즉, 국내 내륙 지진 자료에 대해서는 음이항분포가 더 적합하다고 할 수 있다.

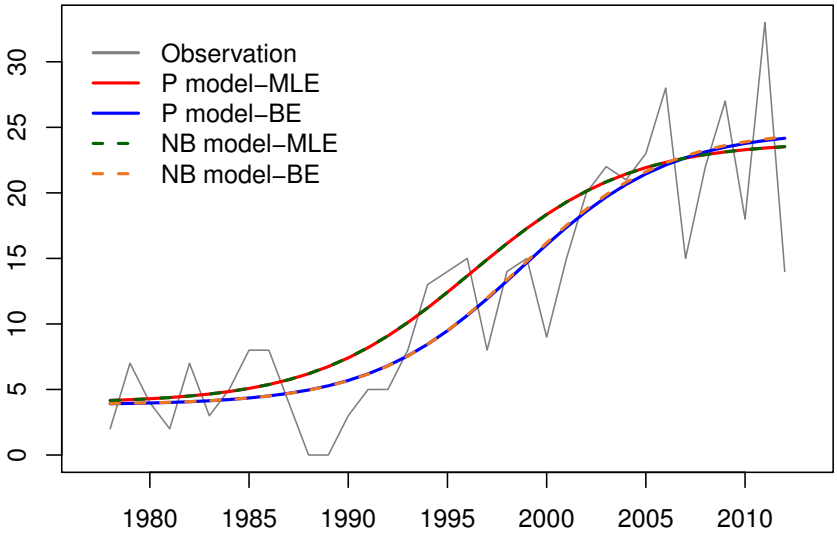

그림 14: 1978 - 2012년의 국내 해역 지진 발생 추이 및 추정 결과

4.3 국내 해역 지진 자료

국내 해역 지진 자료(그림 14)는 1978년부터 2012년까지 35년 간 국내에서 발생한 지진 목록 중 진앙지가 해역인 자료(건수)로, 기상청 홈페이지에서 확인할 수 있다.

표 8은 최대우도추정법과 베이지안추정법을 포아송분포와 음이항분포를 적용하여 분석한 결과와 예측력 비교를 위한 추정값이다. 포아송분포를 적용하였을 때, 초기 시점의 지진 발생 횟수를 살펴보면 최대우도추정값과 베이지안 추정값은 각각 3.964와 3.828로 최대우도추정법이 약간 더 크게 나타났다. 음이항분포를 적용하였을 때, 시간이 지난 후 안정화 되었을 때의 평균 지진 발생 횟수를 살펴보면 최대우도추정값과 베이지안 추정값은 각각 23.896과 24.823로 베이지안추정

표 8: 국내 해역 지진 자료 추정 결과

		Maximum likelihood		Bayesian	
Model	Param	Estimate	95% C.I.	Estimate	95% C.I.
	θ_1	3.964	1.787 - 6.140	3.828	1.638 - 5.343
	θ_2	23.890	18.214 - 29.565	24.759	19.200 - 39.551
	θ_3	19.208	15.750 - 22.667	21.728	17.900 - 29.320
Poisson	θ_4	3.968	1.260 - 6.677	3.750	1.542 - 8.385
	AIC	204.2			
	BIC	210.5			
	DIC			190.870	
	θ_1	3.957	1.697 - 6.217	3.851	1.879 - 5.376
	θ_2	23.896	17.606 - 30.187	24.823	18.630 - 41.010
	θ_3	19.207	15.424 - 22.990	21.684	17.250 - 29.830
NB	θ_4	3.977	1.056 - 6.897	3.688	1.146 - 8.128
	AIC	203.1			
	BIC	209.3			
	DIC			165.815	

Param,Parameter; NB,Negative binomial; C.I.,Confidence Interval

법이 더 크게 나타났다. 다음으로 어떤 모형이 적합한지를 살펴보도록 하자. 먼저 포아송분포와 음이항분포의 최대우도추정량을 비교하기 위해서 AIC와 BIC를 사용하였다. 비교 결과, 음이항분포의 AIC와 BIC가 더 작게 나타났다. 두 번째로 베이지안추정법을 비교하기 위하여 DIC를 사용하였다. 그 결과 음이항분포의 DIC값이 더 작게 나타났다.

제 5 장

결론

본 논문에서는 이산형자료를 이용하여 학습곡선을 모형화하였다. 학습곡선의 형태는 초기 시점에서의 발생 횟수가 시간이 지남에 따라 증가하다가 일정 시간이 지난 후 안정화 되는 형태를 살펴보았다. 이산형자료로는 포아송분포와 음이항분포를 사용하였고, 누적분포함수로는 로지스틱분포함수를 이용하였다. 모수를 추정하기 위한 방법으로는 최대우도추정법과 베이지안추정법을 이용하였다. θ_1과 θ_4를 변화시킨 9개의 모의실험에 대한 모의실험을 진행하였다.

모의실험 결과 추정값들이 모수에 아주 근접하는 결과를 확인하였으며, 초기 시점에서의 발생 횟수인 θ_1과 척도모수인 θ_4가 증가함에 따라 분산과 평균제곱오차의 값이 커지는 것을 확인할 수 있었다. 추정값들을 이용하여 평균 적합선과 95% 신뢰구간을 구하였다. 최대우도추정법에서의 평균 적합선은 참값과 매우 유사하게 나타났으며, 첫 번째 경험적 신뢰구간과 근사적 신뢰구간이 유사하게 나타났다. 두 번째 경험적 신뢰구간은 평균 발생 횟수가 증가하는 시점에서 신뢰구간이 넓게 나타났으며, 안정화 단계에 들어서는 시점에서는 신뢰구간이 좁게 나타났다. 베이지안추정법에서 역시 평균 적합선은 참값과 매우 유사하게 나타났으며, 두 번째 경험적 신뢰구간은 평균 발생 횟수가 증가하는 시점에서 신뢰구간이 넓게 나타났으며, 안정화 단계에 들어서는 시점에서는 신뢰구간이 좁게 나타났다.

예측 평가 기준을 통한 추정법 비교에서는 4개의 기준을 이용하여 평가하였는데, 포아송자료와 음이항자료에 대해서 최대우도추정법이 베이지안추정법에 비해 더 좋게 나타났다. 오분류 평가를 이용한 정분류율은 최대우도추정법에서는 포아송자료를 포아송모형에 적용시키는 것이 적합하다는 결과가 약 70%로 음이항자료를 음이항모형에 적용시키는 것이 적합하다는 결과에 비해 높게 나타났다. 실증연구에서는 영국 광산 자료, 국내 내륙 지진 자료와 국내 해역 지진 자료를 이용하여 분석하였다. 영국 광산 사고 자료는 포아송분포가 적합하다는 결과가, 국내 지진 자료에 대해서는 음이항분포가 적합하다는 결과가 나왔다.

본 논문에서 사용된 자료 중 포아송분포는 평균과 분산이 동일하다는 특징을 갖는다. 따라서 평균이 커질 때 분산도 함께 커지는 과대산포 문제를 갖게 된다. 향후 이런 문제를 해결하기 위해 일반화포아송분포 혹은 일반화음이항분포와 같은 다양한 확률분포를 이용한 학습곡선을 모형화하고자 한다.

참고 문헌

[1] 기상청 홈페이지. http://www.kma.go.kr/weather/earthquake/domesticlist.jsp

[2] 백우종 (2008). 학습곡선에 의한 연료전지의 비용효과 분석, 동신대학교 대학원, 석사학위논문.

[3] 이성진, 박수은 (2007). 유리체절제술의 학습곡선, 대한안과학회, 제48권, 925-934.

[4] 이슬지 (2011). 학습곡선의 모형화 : 이항반응자료를 중심으로, 성신여자대학교 일반대학원, 석사학위논문.

[5] 이슬지, 박만식 (2012). 이항 반응 자료에 대한 학습곡선의 모형화, 한국통계학회, 제19권, 433-450.

[6] 임경민 (2010). 한국 풍력발전의 학습률 추정에 관한 연구, 서울과학기술대학교 대학원, 석사학위논문.

[7] 정종교 (2007). 항공기 날개구조물 조립공정의 생산성향상활동과 학습곡선에 관한 연구, 창원대학교 산업·정보대학원, 석사학위논문.

[8] 정호석 (2009). 복막외 복강경하 근치적 전립선적출술 : 103예의 임상경험 및 학습곡선, 전남대학교 대학원, 석사학위논문.

[9] 홍자인 (2007). 학습곡선이론을 적용한 사용자의 학습성 평가 방법에 관한 연구, 한국기술교육대학교 대학원, 석사학위논문.

[10] Adler, P. S., and Clark, K.B. (1991). Behind The Learning Curve: A Sketch Of The Learning Process, *Management Science*, Vol. 37, 267-281.

[11] Akin, Y., Ates, M., Celik, O., Ucar, M., Yucel, S., and Erdogru T. (2013). Complications of urologic laparoscopic surgery: A center surgeon's experience involving 601 procedures including the learning curve, *The Kaohsiung Journal of Medical Sciences*, Vol. 29, 75-279.

[12] Ballantyne, G. H., Ewing, D., Capella, R. F., Capella, J. F., Davis, D., Schmidt, H. J., Wasielewski, A., and Davies, R.J. (2005). The Learning Curve Measured by Operating Times for Laparoscopic and Open Gastric Bypass: Roles of Surgeon's Experience, Institutional Experience, Body Mass Index and Fellowship Training, *Obesity Surgery*, Vol. 15, 172-182.

[13] Biau, D. J., Williams, S. M., Schlup, M. M., Nizard, R. S., and Porcher, R. (2008). Quantitative and individulaized assessment of the learning curve using LC-CUSUM, *British Journal of Surgery*, Vol. 95, 925-929.

[14] Bradley P. C., Alan E. G., and Adrian F. M. (1992). Hierarchical Bayesian Analysis of Changepoint Problems, *Journal of Applied Statistics*, Vol. 41, 389-405.

[15] Cook, J. A., Ramsay, C. R., and Fayers, P. (2004). Statistical evaluation of learning curve effects in surgical trials, *Clinical Trials*, Vol. 1, 421-427.

[16] Ferguson, G. G., Ames, C. D., Weld, K. J., Yan, Y., Venkatesh, R., and Landman, J. (2005). Prospective evaluation of learning curve for laparoscopic radical prostatectomy: Identification of factors improving operative times, *Adult urology*, Vol. 66, 840-844.

[17] Filho, G. R. O.(2002). The Construction of Learning Curve for Basic Skills in Anesthetic Procedures: An Application for the Cumulative Sum Method, *Economics, Education, And Health Systems Research*, Vol. 92, 411-416.

[18] Forbes, T. L., DeRose, G., Kribs, S. W., and Harris, K. A. (2004). Cumulative sum failure analysis of the learning curve with endovascular

adbominal aortic aneurysm repair, *Journal of Vascular Surgery*, Vol. 39, 102-108.

[19] Jeff F. L., Melissa F., and Huang J. Q. (2013). Learning curve analysis of the first 100 robotic-assisted laparoscopic hysterectomies performed by a single surgeon, *International Journal of Gynecology and Obstetrics*, In Press.

[20] Kim, J. H. and Cheon, S. Y. (2013). Bayesian Multiple Change-point Estimation and Segmentation, *Journal of the Korean Data Analysis Society*, 143.

[21] Kim, S. Y. and Park, M. S. (2012). Parametric Models of the Learning Curve Effects with Continuous Responses, *Journal of the Korean Data Analysis Society*, Vol. 14, 567-576.

[22] Kuo, L. J., Hung, C. S., Wang, W., Tam, K. W., Lee, H. C., Liang, H. H., Chang, Y. J., Huang, M. T., and Wei, P. L. (2013). Intersphincteric resection for very low rectal cancer: clinical outcomes of open versus laparoscopic approach and multidimensional analysis of the learning curve for laparoscopic surgery, *Journal of Surgical Research*, Vol. 183, 524-530.

[23] Lee, J. H., Ryu, K. W., Lee, J. H., Park, S. R., Kim, C. G., Kook, M. C., Nam, B. H., Kim, Y. W., and Bae, J. M. (2006). Learning Curve for Total Gastrectomy with D2 Lymph Node Dissection: Cumulative Sum Analysis for Qualified Surgery, *Annals of Surgical Oncology*, Vol. 13, 1175-1181.

[24] Lee, Y. K., David, J. B., Yoon, B. H., Kim, T. Y., Ha, Y. C., and Koo, K. H. (2013). Learning Curve of Acetabular Cup Positioning in Total Hip Arthroplasty Using a Cumulative Summation Test for Learning Curve (LC-CUSUM), *The Journal of Arthroplasty*.

[25] Li, C., Mi, K., Wen, T. f., Yan, L. n., Li, B., Yang, J. y., Xu, M. q., Wang, W. T., and Wei, Y. g. (2012). A learning curve for living donor liver transplantation, *Digestive and Liver Disease*, Vol. 44, 597-602.

[26] Lieberman, M. B. (1984). The learning curve and pricing in the chemical processing industrie, *Rand Journal of Economics*, Vol. 15, 213-228.

[27] Lim, P. C., Kang, E., and Park, D. H. (2011). A comparative detail analysis of the learning curve and surgical outcome for robotic hysteretomy with lymphadenectomy versus laparoscopic hysterectomy with lymphadenectomy in treatment of endometrial cancer: A case-matched controlled study of the first one hundred twenty two patients, *Gynecologic Oncology*, Vol. 120, 413-418.

[28] Mazzola, J. B. and McCardle, K. F. (1996). A Bayesian Approach to Managing Learning-Curve Uncertainty, *Management Science*, Vol. 42, 680-692.

[29] Schauer, P., Ikramuddin, S., Hamad, G., and Gourash, W. (2002). The Learing curve for laparoscopic Roux-en-Y gastric bypass is 100 cases, *Surgical Endoscopy*, Vol. 17, 212-215.

[30] Sim, H. G., Yip, S. K. H., Lau, W. K. O., Tan, Y. H., Wong, M. Y. C., and Cheng, C. W. S. (2006). Team-based approach reduces learning curve in robot-assisted laparoscopic radical prostatectomy, *International Journal of Urology*, Vol. 13, 560-564.

[31] Smith, A. C., Frank, L. M., Wirth, S., Yanike, M., Hu, D., Kubota, Y., Graybiel, A. M., Suzuki, W. A., and Brown, E. N. (2004). Dynamic Analysis of Learning in Behavioral Experiments, *The Journal of Neuroscience*, Vol. 24, 447-461.

[32] Smunt, T. L. (1999). Log-linear and non-log-linear learning curve models for production research and cost estimation, *International Journal of Production Research*, Vol. 37, 3901-3911.

[33] Tekkis, P. P., Senagore, A. J., Delaney, C. P., and Fazio, V. W. (2005). Evaluation of the Learning Curve in Laparoscopic Colorectal Surgery, *Annals of Surgery*, Vol. 242, 83-91.

[34] Williams, C. K. I. and Vivarelli, F. (2000). Upper and Lower Bounds on the Learning Curve for Gaussian Processes, *Machine Learning*, Vol. 40, 77-102.

Abstract

Modelling of The Learning Curves on the Count Responses

Minji Choi
Department of Statistics
The Graduate School
Sungshin Women's University

As a worker performs a certain job over and over again, he tends to spend less time to complete the job with lower risk of failure. This is generally called the learning-curve effect. In this paper, we focus on parametric models to capture the learning-curve effect inherent in count data by using some discrete-type probability distributions and logistic cumulative distribution function. This work extends Lee and Park (2012), which takes binary data into account for modeling of the learning-curve effect. We use Maximum likelihood method and the Bayesian method to estimate parameters, which determine the characteristics of the learning curves. We consider various simulation scenarios in order to clarify the behaviors of the models proposed in this study and examine their performance. We also compare the two discrete-type models: Poisson model versus Negative-binomial one by means of some information criteria. The real application consists of the anal-

yses of British coal-mine accident data and the earthquake count data in South Korea.

Keywords : Poisson distribution, Negative-binomial distribution, Learning-curve effect, Logistic distribution function, Maximum likelihood estimation method, Bayesian method, British coal-mine data.