



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

박만식교수지도

석사학위 청구논문

일반화 이산형 분포를 기반으로한  
학습곡선의 모형화

2015

성신여자대학교 대학원

통계학과

강민아

# 일반화 이산형 분포를 기반으로한 학습곡선의 모형화

박만식교수지도

이 논문을 석사학위 논문으로 제출함

2014년 11월

성신여자대학교 대학원

통계학과

강민아

# 인준서

강민아의 석사학위 논문으로 인준함.

심사위원 \_\_\_\_\_인

심사위원 \_\_\_\_\_인

심사위원 \_\_\_\_\_인

성신여자대학교 대학원

# 논문개요

학습현상은 Wright(1936)에 의해 항공 산업에서 항공기의 누적 생산량 증가에 따라 항공기 당 직접노동시간이 규칙적으로 감소한다는 것이 관찰되어 최초로 연구되었다. 일반적으로 어떤 작업을 반복적으로 수행할 때 단위생산시간은 감소하는데 이것은 작업자의 숙련도 향상과 작업 절차 조정 및 사용 기계의 작업능력 향상에 의해서 나타난다. 이와 같은 현상을 공수체감현상 또는 학습현상이라고 하고, 학습현상을 그래프나 수식으로 나타낸 것이 학습곡선(learning curve)이다. 그리고 이 때 향상된 생산성을 학습곡선효과(learning-curve effects)라고 한다. 일반적인 학습현상은 무한정 나타나는 것이 아니라 일정 기간까지는 변화율이 크지만 어느 시점이 되면 완화된 경향이 있다. 본 논문에서는 사건 발생 횟수와 자료에 내재된 변동성이 학습현상과 같은 경향으로 변화하는 자료를 적합하기 위해 일반화 포아송분포, 일반화 음이항분포를 이용하여 학습곡선을 모형화 하고자 한다. 모수는 최대우도추정방법을 사용하여 추정하였고, 제안한 모형의 특징을 살펴보고 모형의 성능을 비교해보기 위하여 모의실험을 실시하였다. 그리고 오렌지주스 캔 자료와 자동차 채광창 자료를 가지고 새롭게 제안한 모형을 적합하였다.

**주요어:** 일반화 포아송분포, 일반화 음이항분포, 학습곡선, 로지스틱 분포함수, 뉴튼-랩슨방법

# 목 차

<b>I.</b>	<b>서론</b> . . . . .	<b>1</b>
<b>II.</b>	<b>일반화 이산형 자료를 이용한 학습곡선의 모형화</b> . . . . .	<b>7</b>
2.1	일반화 라그랑지확률 분포 . . . . .	7
2.1.1	일반화 포아송분포 . . . . .	8
2.1.2	일반화 음이항분포 . . . . .	10
2.1.3	일반화 이산형분포의 특성 . . . . .	12
2.2	학습곡선 모형 구축 . . . . .	14
2.3	모수추정방법 . . . . .	16
2.3.1	일반화 포아송분포 . . . . .	17
2.3.2	일반화 음이항분포 . . . . .	19
<b>III.</b>	<b>모의실험</b> . . . . .	<b>20</b>
3.1	모의실험 방법 . . . . .	21
3.2	모의실험 결과 . . . . .	22
3.2.1	모수추정 결과 . . . . .	22
3.2.2	모형의 성능 비교 . . . . .	34
<b>IV.</b>	<b>실증연구</b> . . . . .	<b>40</b>
4.1	오렌지주스 자료 . . . . .	40
4.2	자동차 채광량 자료 . . . . .	42
<b>V.</b>	<b>결론 및 토의</b> . . . . .	<b>45</b>

참고 문헌 . . . . .	48
<b>Abstract . . . . .</b>	<b>52</b>

# 그림 목 차

그림 1.	오렌지주스 자료(Montgomery, 1991) . . . . .	5
그림 2.	산포모수에 따른 일반화 포아송분포와 포아송분포의 차이 . . . . .	13
그림 3.	산포모수에 따른 일반화 음이항분포와 음이항분포의 차이 . . . . .	14
그림 4.	학습곡선 모형 . . . . .	15
그림 5.	최대우도방법을 이용한 참값과 추정값, 그리고 95% 신뢰구간들( $GP, k = -0.0095$ ) . . . . .	26
그림 6.	최대우도방법을 이용한 참값과 추정값, 그리고 95% 신뢰구간들( $GP, k = -0.005$ ) . . . . .	27
그림 7.	최대우도방법을 이용한 참값과 추정값, 그리고 95% 신뢰구간들( $GNB, \beta = 1$ ) . . . . .	32
그림 8.	최대우도방법을 이용한 참값과 추정값, 그리고 95% 신뢰구간들( $GNB, \beta = 3$ ) . . . . .	33
그림 9.	학습곡선 모형에 적합한 오렌지주스 자료(Montgomery, 1991) . . . . .	40
그림 10.	학습곡선 모형에 적합한 채광창 자료(DeVor <i>et al.</i> , 2007) . . . . .	42

# 표 목 차

표 1.	9개의 모의실험 시나리오 . . . . .	20
표 2.	일반화 포아송자료의 일반화 포아송모형 적용 결과 .	23
표 3.	일반화 포아송자료의 일반화 포아송모형 적용 결과 (계속) . . . . .	24
표 4.	일반화 음이항자료의 일반화 음이항모형 적용 결과 .	28
표 5.	일반화 음이항자료의 일반화 음이항모형 적용 결과 (계속) . . . . .	29
표 6.	일반화 포아송자료에 대한 모형 성능 비교 . . . . .	35
표 7.	일반화 포아송자료에 대한 모형 성능 비교(계속) . . .	36
표 8.	일반화 음이항자료에 대한 모형 성능 비교 . . . . .	37
표 9.	일반화 음이항자료에 대한 모형 성능 비교(계속) . . .	38
표 10.	오렌지주스자료 추정 결과 . . . . .	41
표 11.	채광창자료 추정 결과 . . . . .	43

# 제 1 장

## 서론

품질 관리를 위해 산업분야에서는 관리도를 이용하여 제조 공정의 결함이나 오류를 찾아 수정하여 제품의 품질이 향상될 수 있도록 조절을 한다. 이와 같이 사람은 특정한 일의 경험을 쌓아감에 따라 작업의 수행능력이 향상되어 작업에 드는 수행 시간이 줄어드는 경향이 있다. 이러한 현상을 ‘공수체감현상’ 또는 ‘학습현상’이라 한다. 학습현상은 Wright(1936)에 의해 처음 보고되었는데 2차 세계 대전 이전에 항공 산업에서 각 교대가 작업을 반복할수록 제조되는 양은 배가 됨에 따라 직접노동시간은 일정한 비율로 감소한다는 것을 확인하였다. 이런 현상이 나타나는 이유는 사람이 어떤 특정한 작업에 대하여 경험을 쌓아감에 따라 숙련도가 향상되고, 생산 또는 제조 설계 과정이 개선되기 때문이다. 이러한 학습현상을 그래프나 수식으로 나타낸 것을 학습곡선이라 하고 이 때의 향상된 생산성을 학습곡선효과라고 한다. 본 논문에서 학습곡선은 시간이 흐르면서 작업수행정도가 향상되다가 일정 시점이 지나면 작업수행정도가 안정화되는 형태를 보인다. 통상적으로 학습곡선은 어떤 제품에 대한 생산 계획을 세울 때 기술자가 얼마나 경험을 하느냐에 따라 작업수행능력이 얼마나 되는지, 언제쯤 수행능력이 안정화 되는지 확인이 가능하다. 따라서 학습곡선을 이용하여 사업 계획을 수립한다면 생산 일정이나 인력을 어떻게 조절해야 하는지 파악 가능 하여 경제적으로 효율적인 관리가 가능해질 것이다.

현재 학습곡선은 앞에서 설명한 산업분야 이외의 여러 분야에 적용되고 있다. 먼저 산업분야에서 안명관(2005)은 항공기 시스템 구매자와 공급자 간의 관계를 중심으로 학습곡선을 고려한 수량할인 계약과 도매가격 계약 정책이 기대이익을 최대화시키려는 공급자와 구매자의 문제를 해결 가능한지 분석하고 공급사슬에서의 안정적인 계약 형태를 모색하여 최적 조정 방안을 제시하였다. 정종교(2007)는 항공기 날개구조물 조립공정에 여러 학습곡선모형을 이용하여 생산성향상 활동이 어떻게 변화하는지 확인하였는데 생산성향상활동을 펼치지 않아도 가장 적합한 학습곡선 모형을 적용하여 계속되는 생산량에 대한 투입공수를 추정함으로써 목표 공수 산정이 가능하고 그 목표공수를 관리할 수 있다고 설명하였다. Ngwenyama *et al.*(2007)은 학습곡선을 사용하여 IT생산성을 극대화하기 위한 소프트웨어 업그레이드시기에 대한 의사 결정 분석 모델을 만들어 제시하였다. Schoots *et al.*(2008)은 수소 생산 시설의 건설 및 수소의 생산 공정 향상을 위해 1940년에서 2007년까지의 관찰 데이터를 사용하여 학습곡선을 개발하여 제시하였다.

의학 분야에서 정호석(2009)은 복막외 복강경하 근치적 전립선 적출술을 시행 받은 환자들을 대상으로 수술 전후의 임상경과와 단기수술 결과를 학습곡선을 이용하여 분석한 결과 복막외 복강경하 근치적 전립선 적출술은 장기간의 학습시간과 노력이 필요하나 개복근치적 적출술에 비해 수술 중이나 후에 합병증이 적고 빠른 회복이 가능하다고 시사하였다. 김지선(2014)은 치과위생사들을 대상으로 두가지 3차원 디지털 구강스캐너를 사용한 인상채득 시 반복적으로 촬영을 진행함에 따라 총 촬영시간에 변화가 있는지 학습곡선을 이용하여 비교하였는데 첫번째 구강스캐너는 학습률은 좋으나 총 촬영소요시간이

길고 사용자의 임상경력에 영향을 받았고 두번째 구강스캐너는 학습률은 낮으나 총 촬영소요시간이 짧고 사용자의 임상경력에도 영향을 받지 않는 것을 확인하였고, Tamas *et al.*(2006)은 로봇 담낭 절제술의 이점을 학습곡선을 이용하여 설명하였는데 반복 작업을 함에 따라 수술 시간이 감소한다고 설명하였다. 그리고 Leigh *et al.*(2009)은 학습곡선을 이용하여 자궁 내막 암에 대한 솔반 대동맥 림프절 절제술과 로봇 자궁 절제술이 수술이 진행됨에 따라 효율이 향상됨을 확인하였다. Wolff *et al.*(2014)은 신장 이식 후 수술 합병증이 외과 의사의 경험과 관계가 있는지 개별적인 외과 의사의 합병증 비율이 그들의 학습과정에서 개선이 되는지 확인하기 위해 학습곡선을 이용하여 알아보았는데 경험이 풍부한 외과 의사가 비교적 짧은 학습기간 후에 신장 이식에서 좋은 결과를 얻을 수 있음을 보였다. Ospina *et al.*(2014)은 6가지 마취 절차에 대한 누적합계 학습곡선을 만들어 마취하였을 때 정신 운동 학습성의 변화를 품질 기준과 관련하여 의미 있는 정보를 제공하고 조기에 발견할 수 있는 모형을 형성하였다. Andréasson *et al.*(2014)은 안정화된 점액종 복막을 가진 환자의 수술 결과를 보기 전에 얼마나 많은 과정이 필요한지 누적합 학습곡선을 이용하여 분석한 결과 200개 이상의 절차를 필요로 하는 매우 까다로운 절차라는 것을 확인하였다.

그 외의 분야에서 최준영과 송경빈(2004)은 고효율 기기의 보급을 촉진하기 위해 보조금을 지급하는 제도가 얼마나 효과가 있는지 학습곡선을 이용하여 분석한 결과 고효율기기를 구입할 때 보조금을 지급하는 보급지원정책은 기기의 생산가격에 영향을 주는 것을 확인하였고, 정광태와 홍자인(2008)은 제품디자인에 대한 학습성을 비교 평가하기 위한 기준으로 학습곡선을 적용해 보았는데 디자인된 제품의 학습 정도와 사용편의성을 비교 평가하는데 학습곡선이 효과적으

로 활용될 수 있음을 확인했다. 뿐만 아니라 정소기와 차경천(2014)은 학습곡선을 이용하여 유선통신 서비스 상품별 고객 불만 감소에 대한 분석을 하였는데 시간이 지남에 따라 고객 불만이 감소하였고 초고속 인터넷, 전화, IPTV 등은 인력투입, 네트워크장애, 계절요인으로 인해 고객 불만을 증가시키거나 감소시켰고 서비스 상품별 다양한 변수는 고객의 채감 품질을 높이지만 고객 불만은 감소하지 않는다는 것을 보였다. Wang *et al.*(2001)은 학습곡선을 사용하여 하나의 전체 생산성을 유지하고 구현하는 성능을 점검하기 위한 적합한 시간을 결정하는 방법을 제안하였고, Nemet(2006)는 학습곡선을 이용하여 태양광 발전의 비용에 영향을 미치는 가장 중요한 요소를 식별하는 간단한 모델을 제시하였다. Teng *et al.*(2014)은 무역 신용은 판매뿐만 아니라 기회비용과 기본적인 위험도 증가하고 생산비용은 학습곡선현상에 따라 감소하는 경향을 고려하여 어떤 사람의 최적의 무역 신용 기간과 동시에 생산량의 크기를 결정하기 위해 판매자의 관점으로부터 경제적 생산량 모델을 제안하였다.

학습곡선 모형화에 대한 논문으로는 이슬지(2011)가 이항반응자료를 중심으로 이항분포를 이용하여 학습곡선을 모형화한 이후 Kim *et al.*(2012)은 연속형자료를 중심으로 여러 누적분포함수를 이용하여 학습곡선을 모형화하였다. 그리고 Park *et al.*(2012)은 외과 의사가 복강경절제술의 경험에 따라 로봇절제술에 얼마나 빨리 적응을 하는지에 대한 연구에서 공간통계연구 분야에서 사용되는 비선형회귀모델인 구형 공분산함수를 사용하여 학습곡선을 모형화하였고, 최민지(2014)는 이산형자료에 대하여 포아송분포, 음이항분포를 이용하여 학습곡선을 모형화하였다. 앞에서 언급한 논문들은 특정 사건의 발생 횟수나 시간의 추세가 변화하는 자료에 대하여 모형화하였다. 본 논문은

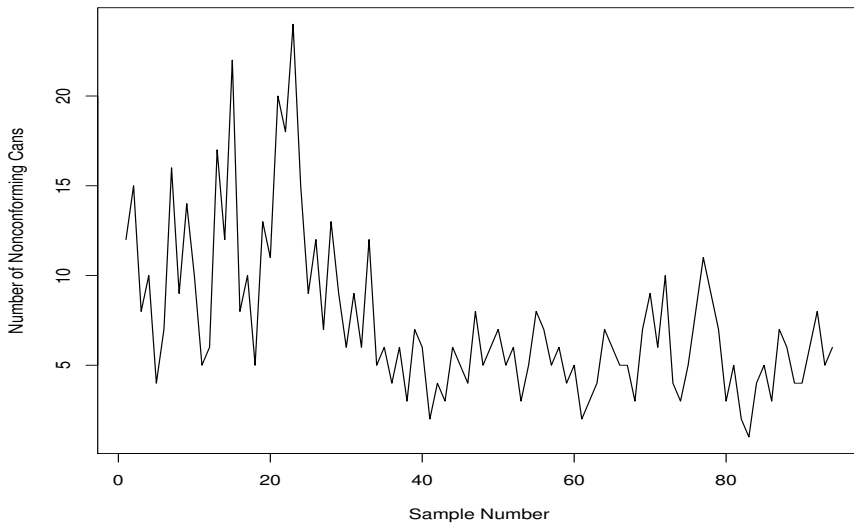


그림 1: 오렌지주스 자료(Montgomery, 1991)

일반화 포아송분포, 일반화 음이항분포를 이용하여 발생 횟수의 추세 뿐만 아니라 자료의 변동성도 변화하는 자료에 대하여 모형화하고자 한다. 일반화 포아송분포와 일반화 음이항분포는 각각 포아송분포와 음이항분포를 확장한 분포로서 기존의 모형보다 변동성이 과도 또한 과소하게 나타나는 경우에도 모형을 적용할 수 있다. 예로 제조공장에서의 제조품의 불량 횟수에 대하여 살펴보고자 한다. 그림 1을 보면 시간이 지날수록 작업수행이 익숙해짐에 따라 제조품의 불량 횟수가 감소하다가 완만해지고 변동성 또한 감소하는 것을 알 수 있다. 이러한 형태의 자료를 적합할 수 있는 학습곡선을 모형화 하고자 한다.

본 논문의 순서는 다음과 같다. 제 2장에서는 일반화 포아송분포와 일반화 음이항분포를 이용한 학습곡선 모형에 대하여 설명하고 제 3장에서는 모의실험을 통하여 제안한 모형에 대한 특징을 설명한다.

제 4장에서는 오렌지주스 캔 자료와 자동차 채광창 자료를 가지고 모형  
형을 적합해보고 제 5장에서는 결론과 향후 연구에 대해 논의한다.

## 제 2 장

# 일반화 이산형 자료를 이용한 학습곡선의 모형화

어떤 일을 반복하여 수행할수록 작업이 소요되는 시간은 감소하고 작업의 질 또한 향상되는 것을 학습현상이라 한다. 본 논문에서는 특정 사건의 발생 횟수의 변동성이 변화하는 자료를 가지고 일반화 포아송분포와 일반화 음이항분포를 이용하여 학습곡선을 모형화하고자 한다.

### 2.1 일반화 라그랑지확률 분포

본 논문에서 활용한 분포는 일반화 포아송분포와 일반화 음이항분포로 이 분포들은 일반화 라그랑지분포(general lagrangian distribution)에 속한다.  $g(z)$ 와  $h(z)$ 는  $z$ 의 분포함수라고 할때,  $h(0) \neq 0$ ,  $h(1) = 1$ ,  $g(1) = 1$ 이고, 임의의 자연수  $x \in N$ 에 대하여 아래와 같은 조건을 충족시킨다고 하자.

$$D^{x-1}[(h(z))^x g'(z)] \Big|_{z=0} \geq 0.$$

라그랑지 변환  $u = z/h(z)$ 의 역함수를 고려하여 아래 식 (2.1)을 구현할 수 있다.

$$g(z) = \sum_{k=0}^{\infty} \frac{u^k}{k!} D^{k-1} [(h(z))^k Dg(z)]_{z=0}. \quad (2.1)$$

이 때  $z = \sum_{k=1}^{\infty} (u^k/k!) D^{k-1} [(h(z))^k]_{z=0}$ 이다.  $z = \psi(u)$ 의 일반화 라그랑지 분포의 확률생성함수(probability generate function; *pgf*)는 식 (2.1)에 의해 다음과 같다.

$$g(z) = f(\psi(u)) = E(u^x) = \sum_{k=0}^{\infty} \frac{u^x}{x!} D^{x-1} [(h(z))^x g'(z)] \Big|_{z=0}. \quad (2.2)$$

따라서 일반화 라그랑지 확률분포의 확률질량함수(probability mass function; *pmf*)는 식(2.1)에 의해 다음과 같이 나타낼 수 있다.  $x = 0, 1, 2, \dots$ 에 대하여

$$P(X = x) = \frac{1}{x!} D^{x-1} [(h(z))^x g'(z)] \Big|_{z=0}. \quad (2.3)$$

여기서,  $P(X = 0) = g(0)$ 이다. 이때 확률질량함수의  $g(z)$ 와  $h(z)$ 에 따라 일반화 포아송분포 또는 일반화 음이항분포로 표현된다.

### 2.1.1 일반화 포아송분포

Consul and Jain(1973)이 소개한 일반화 포아송분포(*generalized Poisson distribution*)은 기존의 포아송분포를 확장한 모형이다. 일반화 라그랑지 확률분포의 확률질량함수에서  $g(z)$ 와  $h(z)$ 가 다음과 같을 때 일반

화 포아송분포의 확률질량함수가 된다.

$$g(z) = e^{\theta(z-1)}, \quad h(z) = e^{\lambda(z-1)}.$$

$\theta > 0$ 이고,  $\max(-1, -\theta/m) \leq \lambda \leq 1$ 이고,  $m(\geq 4)$ 이 큰 양의 정수값 일 때 식 (2.3)에 의해 일반화 포아송분포의 확률질량함수는  $x = 0, 1, 2, \dots$ 에 대하여

$$L(f; g; x) = \frac{1}{x!} \theta (\theta + \lambda x)^{x-1} e^{-\theta - \lambda x}.$$

여기서,  $\theta = \mu_t / (1 + k\mu_t)$ ,  $\lambda = k\mu_t / (1 + k\mu_t)$ 로 변환하면 일반화된 포아송분포의 확률질량함수의 정의는 다음과 같다(Famoye, 1993). 임의의  $y_t = 0, 1, 2, \dots$ 에 대하여

$$f(y_t | \mu_t, k) = \left( \frac{\mu_t}{1 + k\mu_t} \right)^{y_t} \frac{(1 + ky_t)^{y_t-1}}{y_t!} \exp \left\{ -\frac{\mu_t(1 + ky_t)}{1 + k\mu_t} \right\}. \quad (2.4)$$

본 논문에서는 식 (2.4)을 일반화 포아송분포라 정의한다. 여기서  $\mu_t$ 는 평균모수이고,  $k$ 는 산포모수이다. 만약  $k = 0$ 이면 평균과 분산이 같은 포아송분포가 된다.

확률변수  $Y_t$ 가 일반화 포아송분포를 따른다고 하면 다음과 같이 표현한다.

$$Y_t \sim GP(\mu_t, k).$$

일반화 라그랑지분포의 계승적률(factorial moment)는  $r = 1, 2, \dots$ 에 대

하여

$$\mu'_r = \left. \frac{d^r f(z)}{du^r} \right|_{z=1} = \left. \frac{d^r f(\Psi(u))}{du^r} \right|_{z=1}.$$

예를 들면 첫번째와 두번째 계승적률을 구하면 아래와 같다.

$$\begin{aligned} \mu'_1 &= \left. \frac{\partial f(\Psi(u))}{\partial u} = f'(z)g(z)(1-ug'(z))^{-1} \right|_{z=1}, \\ \mu'_2 &= \left. \frac{\partial^2 f(\Psi(u))}{\partial u^2} \right. \\ &= \left. \frac{[f''(z)g(z) + g'(z)f'(z)]g(z)}{(1-ug'(z))^2} + \frac{[f'(z)g(z)]}{(1-ug'(z))^2} \left[ g'(z) + ug''(z) \frac{\partial z}{\partial u} \right] \right|_{z=1}. \end{aligned}$$

$z = u = 1$ 이고  $g(z) = e^{\lambda(z-1)}$ 와  $f(z) = e^{\theta(z-1)}$ 를 대입 하면 일반화 포아송분포의 평균과 분산을 다음과 같이 구할 수 있다.

$$\mu = \theta(1-\lambda)^{-1}, \quad \sigma^2 = \mu'_2 + \mu'_1 - (\mu'_1)^2 = \theta(1-\lambda)^{-3}.$$

따라서 본 논문에서 이용하는 일반화 포아송분포 형태의 평균과 분산을 다음과 같다.

$$E(y_t) = \mu_t, \quad \text{Var}(y_t) = \mu_t(1+k\mu_t)^2.$$

## 2.1.2 일반화 음이항분포

일반화 포아송분포와 같이 Consul and Jain(1973)이 소개한 일반화 음이항분포(generalized negative-binomial distribution)는 기존의 음이항분포를 확장한 모형이다. 일반화 라그랑지분포의 확률질량함수인 식 (2.3)에서  $g(z)$ 와  $h(z)$ 가 다음과 같을 때 일반화 음이항분포의 확률질량

함수가 된다.

$$g(z) = (1 - \alpha_t)^n (1 - \alpha_t z)^{-n}, \quad h(z) = (1 - \alpha_t)^{\beta-1} (1 - \alpha_t z)^{-(\beta-1)}.$$

여기서,  $0 < \alpha_t < 1$ 이고  $|\alpha_t \beta| < 1$ 일 때 일반화 음이항분포의 확률질량 함수는 다음과 같이 정의된다. 확률변수  $y_t = 0, 1, 2, \dots$  일 때

$$f(y_t | \alpha_t, \beta, n) = \frac{n}{n + \beta y_t} \binom{n + \beta y_t}{y_t} \alpha_t^{y_t} (1 - \alpha_t)^{n + \beta y_t - y_t}. \quad (2.5)$$

일반화 음이항분포의  $k$ 차 적률(moment)은 다음과 같이 나타낼 수 있다.

$$\mu'_k(n) = n \alpha_t \sum_{j=0}^{k-1} \binom{k-1}{j} \left[ \mu'_j(n + \beta - 1) + \frac{\beta}{n + \beta - 1} \mu'_{j+1}(n + \beta - 1) \right].$$

$1 \leq \beta < \alpha_t^{-1}$ 일 때 적률을 이용하여 평균과 분산을 구하면 다음과 같다.

$$E(y_t) = \frac{n \alpha_t}{1 - \alpha_t \beta}, \quad \text{Var}(y_t) = \frac{n \alpha_t (1 - \alpha_t)}{(1 - \alpha_t \beta)^3}.$$

일반화 음이항분포는  $\beta = 0$ 이고  $n$ 이 양의 정수일 때 이항분포가 되고,  $\beta = 1$ 일 때 음이항분포가 된다. 또한  $\beta = 1/2$ 이면 분포의 평균과 분산이 대략적으로 같아져 포아송분포와 유사하다.

일반화 포아송분포에서 사용되는 모수와 동일한 형태의 모수를 사용하기 위해 일반화 음이항분포의 평균이 일반화 포아송분포의 평균인  $\mu_t$ 라 하면 이때  $\alpha_t = \mu_t / (n + \mu_t \beta)$ 가 된다. 따라서  $\alpha_t$ 를  $\mu_t$ 로 재표현

한 일반화 음이항분포는 아래와 같다. 확률변수  $y_t = 0, 1, 2, \dots$  일 때

$$f(y_t | \mu_t, \beta, n) = \frac{n}{n + \beta y_t} \binom{n + \beta y_t}{y_t} \left( \frac{\mu_t}{n + \mu_t \beta} \right)^{y_t} \left( 1 - \frac{\mu_t}{n + \mu_t \beta} \right)^{n + \beta y_t - y_t}. \quad (2.6)$$

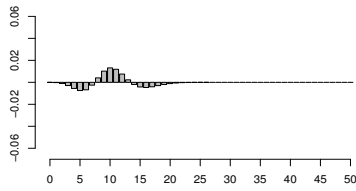
이는 다음과 같이 표현한다.

$$Y_t \sim GNB(\mu_t, \beta, n).$$

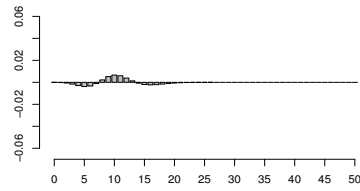
### 2.1.3 일반화 이산형분포의 특성

그림 2는 어떤 특정 사건이 일어난 횟수가 0에서 50이고 평균발생횟수  $\mu = 10$ 인 경우에 산포모수  $k = -0.0095, -0.005, 0.01, 0.1$ 인 일반화 포아송분포의 확률값에서 포아송분포의 확률값을 뺀 값의 막대 그래프이다. 산포모수가 음수일 때 어떤 사건이 일어난 횟수가 0회에서 5회일 때는 일반화 포아송분포의 확률값이 포아송분포의 확률값보다 작지만 6회에서 11회까지는 크고 12회부터 다시 작아지는데 20회 이후 부터 차이가 거의 없다. 산포모수가 양수면 음수일 때와 반대로 포아송분포의 확률값보다 앞부분은 커지고 다시 작아졌다가 커지는 형태를 보이는데 산포모수의 절대값이 증가할수록 그 차이는 크게 난다. 이는 일반화 포아송분포의 산포모수가 음수이면 포아송분포보다 첨도가 크고 변동폭이 작고, 산포모수가 양수면 첨도가 작고 변동폭이 큰 경향을 보이는 분포가 된다는 것을 의미한다.

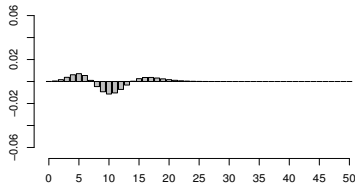
그림 3은 어떤 특정 사건이 일어난 횟수가 0에서 50이고 평균발생횟수  $\mu = 10$ 인 경우에 산포모수  $\beta = 3, 5, 10, 30$ 인 일반화 음이항분포의 확률값에서 음이항분포의 확률값을 뺀 값의 막대 그래프이다. 산포모



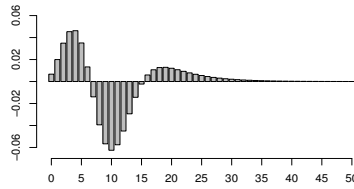
(a)  $k = -0.0095$



(b)  $k = -0.005$



(c)  $k = 0.01$



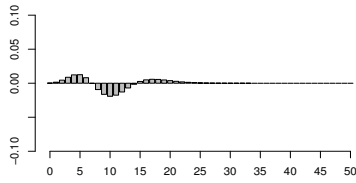
(d)  $k = 0.1$

그림 2: 산포모수에 따른 일반화 포아송분포와 포아송분포의 차이

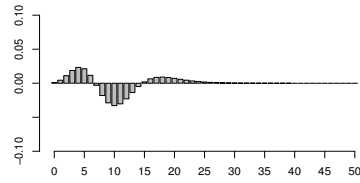
수가 3일 때 어떤 사건이 일어난 횟수가 0회에서 5회일 때는 일반화 음이항분포의 확률값이 음이항분포의 확률값보다 크고 6회에서 13회는 작고 다시 14회부터 크다. 일반화 음이항분포도 일반화 포아송분포와 마찬가지로 산포모수의 값이 증가할수록 음이항분포의 확률값과 차이가 커짐을 알 수 있다. 이는 일반화 음이항분포의 산포모수가 증가할수록 음이항분포보다 자료의 첨도가 감소하고 변동폭이 증가하는 분포 형태를 갖는다는 것을 의미한다.

일반화 이산형분포 간의 관계를 보면  $\mu_t$ 가 10인  $k$ 는 다음과 같이 나타낼 수 있다.

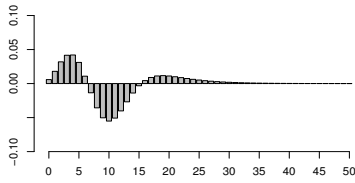
$$k + 0.1 = \left[ n \times \ln \left( \frac{n + 10\beta}{n + 10(\beta - 1)} \right) \right]^{-1}. \quad (2.7)$$



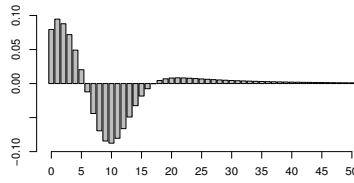
(a)  $\beta = 3$



(b)  $\beta = 5$



(c)  $\beta = 10$



(d)  $\beta = 30$

그림 3: 산포모수에 따른 일반화 음이항분포와 음이항분포의 차이

일반화 포아송분포의 산포모수  $k$ 와 일반화 음이항분포의 산포모수  $\beta$ 의 관계에서  $\beta \geq 1$ 이므로  $k$ 가 음수인 경우는 적용되지 않는다. 따라서 일반화 포아송모형이 일반화 음이항모형보다 더 포괄적으로 여러 형태의 자료를 적합할 수 있다.

## 2.2 학습곡선 모형 구축

본 논문에서 다룰 자료는 일정한 시간에 특정 사건이 일어난 횟수와 자료의 변동성이 변화하는 자료이고 이에 대하여 4개의 모수를 이용하여 그림 4와 같은 학습곡선 모형을 구축하고자 한다. 이 때, 그림 4에서  $\theta_2$ 는 초기 시점에서 특정 사건이 발생한 횟수를 나타내고  $\theta_1$ 은 일정 시간이 지난 후 안정화 되었을 때 특정 사건이 발생한 횟수를

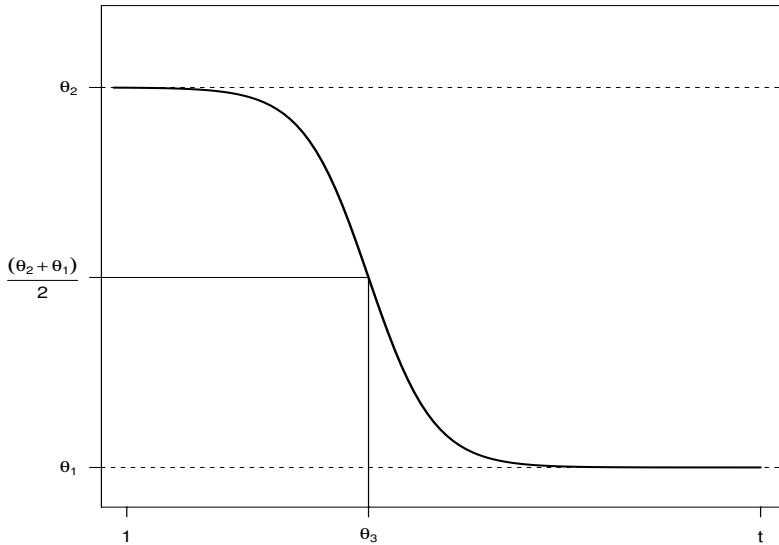


그림 4: 학습곡선 모형

나타낸다. 즉, 더 이상 줄일 수 없는 특정 사건의 발생 횟수를 뜻한다.  $\theta_3$ 과  $\theta_4$ 는 각각 학습곡선의 위치모수(location parameter)와 척도모수(scale parameter)이고 이 4개의 모수를 고려하여 학습곡선 모형을 구축한다. 이때  $\theta_1$ 과  $\theta_2$ 는 0 이상 값을 가지며  $0 \leq \theta_1 < \theta_2$ 를 만족해야 한다. 여기서  $\Theta = (\theta_1, \theta_2, \theta_3, \theta_4)^T$ 이고  $t = 1, \dots, T$ 에 대하여 평균발생횟수인  $\mu_t(\Theta)$ 는 다음과 같다.

$$\mu_t(\Theta) = \theta_1 + (\theta_2 - \theta_1)[1 - F(t | \theta_3, \theta_4)]. \quad (2.8)$$

본 논문은  $t$ 의 함수로 로지스틱분포(logistic distribution)를 이용하였다.

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

위치모수가  $\theta_3$ 이고 척도모수가  $\theta_4$ 인 로지스틱 누적분포함수는 다음과 같다.

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

따라서 식 (2.8)은 다음과 같이 나타낼 수 있다.

$$\mu_t(\Theta) = \theta_1 + (\theta_2 - \theta_1) \left[ 1 - \left\{ 1 + \exp\left(-\frac{t - \theta_3}{\theta_4}\right) \right\}^{-1} \right].$$

## 2.3 모수추정방법

학습곡선 모형의 모수  $\Theta$ 를 추정하기 위해서 최대우도추정방법(maximum likelihood estimation method)을 이용하였다. 최대우도추정 방법은 우도함수 또는 로그우도함수를 이용하여 모수를 추정하는 방법이다. 이런 우도식이나 로그우도식의 수치적 해는 뉴턴-랩슨방법(Newton-Raphson method)에 의해서 설명될 수 있다. 뉴턴-랩슨방법은 초기에 추측한 값을 가지고 다음에 추측한 추정값의 차이가 없어질 때까지 반복하여 모수를 추정하는 방법이다. 뉴턴-랩슨방법을 이용할 때 활용하는 반복 식이 식 (2.9)이다. 여기서, 추정값  $\hat{\Theta} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4)^T$ 이다.

$$\hat{\Theta}^{(m)} = \hat{\Theta}^{(m-1)} + \left[ \mathbf{I}^{(m-1)} \right]^{-1} \mathbf{U}^{(m-1)}. \quad (2.9)$$

로그우도함수의 특정 모수에 대한 일차 편미분을 그 모수의 스코어함수(score function)라고 하고 스코어함수의 제곱의 기댓값을 피셔 정보(Fisher information)이라고 하며 식 (2.9)에서  $\mathbf{I}^{(m-1)}$ 를 말한다.

그리고 델타 방법(delta method)을 이용하여 신뢰구간을 추정하는데 이 방법은 테일러 전개에 기반을 두어 함수의 평균과 분산을 근사적으로 계산하여 구할 수 있다.  $\mu_t$ 에 대한  $100 \times (1 - \alpha)\%$  근사신뢰구간(approximate confidence interval)은 다음과 같다(델타 방법에 대한 자세한 내용은 최민지(2013)논문을 참고한다).

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

### 2.3.1 일반화 포아송분포

$\mathbf{y} = (y_1, y_2, \dots, y_T)'$ 이고 확률변수  $Y_t$ 가  $GP(\mu_t(\Theta))$ 를 따를 때, 일반화 포아송분포의 확률질량함수는 식 (2.4)이고, 로그우도함수는 다음과 같다.

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

로그우도함수를  $\mu_t$ 로 1차 편미분하면 다음과 같다.

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

식 (2.9)에서  $\mathbf{U}^{(m-1)}$ 는 스코어 통계량(score statistic)이고 각각의 모수에 대한 편미분을 하여 구한다.

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

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

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

$\mu_t$ 를  $\theta_i$ 로 편미분하면 다음과 같다.

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

정보행렬(information matrix)  $\mathbf{I}$ 는 식 (2.9)의 스코어 통계량  $\mathbf{U}$ 를 모수  $\Theta$ 에 대하여 편미분한 값이다.

$$\mathbf{I} = \{I_{ij}\}_{i,j=1,2,3,4} = \left\{ \frac{\partial}{\partial \theta_j} U_i(\Theta|\mathbf{y}) \right\}_{i,j=1,2,3,4}. \quad (2.10)$$

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

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

## 2.3.2 일반화 음이항분포

$\mathbf{y} = (y_1, y_2, \dots, y_T)'$ 이고 확률변수  $Y_t$ 가  $GNB(\mu_t(\Theta))$ 를 따른다고 할 때, 일반화 음이항분포의 확률질량함수는 식 (2.6)이고, 로그우도함수는 다음과 같다.

$$\begin{aligned} l(\mu_t | \mathbf{y}) &= \sum_{t=1}^T \ln(\Gamma(y_t + n\beta)) - \sum_{t=1}^T \ln(\Gamma(y_t + 1)) \\ &\quad - \sum_{t=1}^T \ln(\Gamma(n + y_t\beta - y_t + 1)) + \sum_{t=1}^T y_t \ln\left(\frac{\mu_t}{n + \mu_t\beta}\right) \\ &\quad + \sum_{t=1}^T (n + y_t\beta - y_t) \ln\left(1 - \frac{\mu_t}{n + \mu_t\beta}\right). \end{aligned}$$

로그우도함수를  $\mu_t$ 로 1차 편미분하면 다음과 같다.

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

스코어 통계량과 피셔 정보는 일반화 포아송분포에서 구한 것과 동일하다.

## 제 3 장

### 모의실험

제 2장에서 구축한 통계적인 모형의 특징을 살펴보고 모형의 성능을 비교하기 위해 최대우도추정방법을 이용하여 모의실험을 수행하였다. 모의실험은 일반화 포아송분포와 일반화 음이항분포에서 임의로 자료를 생성한다. 이때 각 분포의 모수는 임의로 설정한 후 자료를 생성하여 통계적인 모형의 모수를 추정한다. 통계적인 모형은 포아송분포, 음이항분포, 일반화 포아송분포와 일반화 음이항분포를 따르는 모형을 말한다.

표 1: 9개의 모의실험 시나리오

$\Theta_0$	Scenario								
	1.	2.	3.	4.	5.	6.	7.	8.	9.
$\theta_{10}$	1	1	1	1	1	1	1	1	1
$\theta_{20}$	10	10	10	10	10	10	10	10	10
$\theta_{30}$	100	200	300	100	200	300	100	200	300
$\theta_{40}$	5	5	5	25	25	25	45	45	45

통계적인 모형 식 (2.8)은 어떤 현상이 초기에 발생하는 횟수인  $\theta_1$ , 어떤 현상이 안정화 되었을 때의 발생하는 횟수인  $\theta_2$ , 학습곡선의 변곡점에 해당하는 시점인  $\theta_3$ 와 학습곡선의 변동의 정도를 나타내는  $\theta_4$ 와 같은 4개의 모수로 구성되어 있다. 표 1을 보면 모수의 값의 변화에 따른 모형의 특성을 알아보기 위해 총 9가지 시나리오로 모의실험을 하였다. 여기서 참값  $\Theta_0 = (\theta_{10}, \theta_{20}, \theta_{30}, \theta_{40})^T$ 이다. 모의실험에서 시점

$t = 1, \dots, T$ 이고, 이러한 표본을  $B$ 개 생성하여 유의성을 보았다.

### 3.1 모의실험 방법

모의실험 절차는 다음과 같다.

- 1) 로지스틱 누적분포함수로 만든 학습곡선 모형을 이용하여 식 (2.8)에 나타난 평균발생횟수  $\mu_t(\Theta_0)$ 를 구한다.
- 2) 1)에서 구한  $\mu_t(\Theta_0)$ 를 이용하여 이산형 확률분포를 따르는 자료  $B$ 개를 각각 생성한다. 이때 일반화 포아송분포의 산포모수  $k$ 가  $\{-0.0095, -0.005, 0, 0.01, 0.1, 0.5\}$ 인 경우, 일반화 음이항분포의 산포모수  $\beta$ 가  $\{1, 3, 5, 10, 20, 30\}$ 인 경우를 고려하였다.
- 3) 2)에서 생성된 자료를 가지고 각 분포에 적합하여 모수를 추정한다. 이 때 모수추정방법은 최대우도추정방법을 이용한다.
- 4)  $B$ 개의 자료에 대한 평균, 분산, 평균제곱오차(MSE)를 계산하여 모수 추정결과를 확인한다.
- 5) 델타방법을 이용하여 근사신뢰구간을 구한다.  $\mu_t$ 에 대한  $100 \times (1 - \alpha)\%$  근사신뢰구간은 다음과 같다.

$$\left( \mu_t(\hat{\Theta}) - Z_{\alpha/2} \sqrt{\widehat{Var}(\mu_t(\hat{\Theta}))}, \mu_t(\hat{\Theta}) + Z_{\alpha/2} \sqrt{\widehat{Var}(\mu_t(\hat{\Theta}))} \right). \quad (3.1)$$

$\mu_t(\hat{\Theta})$ 는  $B$ 개의  $\hat{\theta}_i$ 의 평균을 이용하여 구하고  $\widehat{Var}(\mu_t(\hat{\Theta}))$ 는  $B$ 개의  $Var(\mu_t(\hat{\Theta}))$ 을 평균내어 구한다.

- 6) 뉴턴-랩슨 방법을 이용하여 구한 추정값의 경험적 신뢰구간(empirical confidence interval)을 구한다.  $\mu_t$ 에 대한  $100 \times (1 - \alpha)\%$  경

협적 신뢰구간은  $\mu_t$ 를 생성한 후 정렬하여 구한다.

$$\left( \{\hat{\mu}_t\}_{\left(\frac{(B+1)\alpha}{2}\right)}, \{\hat{\mu}_t\}_{\left((B+1)(1-\frac{\alpha}{2})\right)} \right). \quad (3.2)$$

## 3.2 모의실험 결과

모의실험에서는 최대우도추정방법을 이용하여 모수를 추정하고 결과를 이용하여 신뢰구간을 구하였다. 일반화 이산형분포는 산포모수 별 자료를 생성하여 각각의 모형에 적합해본 후 모형의 성능에 대해 평가하기 위해 각 모형의 AIC(Akaike, 1974), BIC(Schwarz, 1978), AICc(Burnham *et al.*, 2004) 값을 구하여 비교하였다. 9개의 시나리오에 대해서  $T = 500, B = 199$ 로 설정하였고 일반화 음이항분포의 모수  $n$ 은 100으로 하여 분석하였다.

### 3.2.1 모수추정결과

표 2와 표 3은 9개의 시나리오와 산포모수에 따라 일반화 포와송 자료를 일반화 포아송모형에 적합했을 때 추정결과를 정리한 표이다. 결과를 보면 모든 추정값의 평균이 모수의 참값과 유사하게 나타났다.  $\theta_{30}$ 의 변화에 따른 분산과 평균제곱오차를 살펴보기 위해 시나리오 1, 2, 3을 보면,  $\theta_{30}$ 의 값이 증가할수록  $\hat{\theta}_2$ 와  $\hat{\theta}_3$ 의 분산과 평균제곱오차는 감소한다. 이러한  $\hat{\theta}_2$ 와  $\hat{\theta}_3$ 의 분산과 평균제곱오차가 감소하는 형상은 시나리오 1, 2, 3보다 시나리오 7, 8, 9에서 확연히 드러난다. 그와 반대로  $\hat{\theta}_1$ 과  $\hat{\theta}_4$ 의 분산과 평균제곱오차는 증가한다. 그리고 추정값의 분산과 평균제곱오차를 비교해보면  $\hat{\theta}_1$ 의 분산과 평균제곱오차의 차이가 없고 비교적  $\hat{\theta}_3$ 와  $\hat{\theta}_4$ 의 분산과 평균제곱오차의 차이가 있다. 그 차이는

표 2: 일반화 포아송자료의 일반화 포아송모형 적용 결과

Scenario	Param	Truth	$k = -0.0095$			$k = -0.005$			$k = 0$		
			Mean	Var	MSE	Mean	Var	MSE	Mean	Var	MSE
1	$\theta_{10}$	1	0.995	0.002	0.002	1.002	0.003	0.003	0.998	0.003	0.003
	$\theta_{20}$	10	9.872	0.033	0.049	9.876	0.037	0.052	9.859	0.046	0.066
	$\theta_{30}$	100	100.2	1.807	1.845	100.0	2.283	2.283	100.5	2.001	2.217
	$\theta_{40}$	5	4.899	0.974	0.984	4.959	1.103	1.105	4.730	1.282	1.355
2	$\theta_{10}$	1	1.003	0.004	0.004	1.003	0.004	0.004	1.000	0.004	0.004
	$\theta_{20}$	10	9.925	0.014	0.019	9.914	0.016	0.023	9.892	0.022	0.034
	$\theta_{30}$	200	200.1	2.055	2.060	200.1	1.674	1.682	200.1	2.058	2.079
	$\theta_{40}$	5	4.898	1.284	1.295	4.860	1.117	1.136	4.814	1.210	1.245
3	$\theta_{10}$	1	1.003	0.005	0.006	1.000	0.005	0.005	0.998	0.006	0.006
	$\theta_{20}$	10	9.924	0.012	0.017	9.927	0.011	0.016	9.925	0.012	0.017
	$\theta_{30}$	300	300.1	2.025	2.036	300.1	1.911	1.930	300.1	1.912	1.915
	$\theta_{40}$	5	4.933	1.354	1.358	4.809	1.211	1.248	4.932	1.351	1.355
4	$\theta_{10}$	1	1.005	0.003	0.003	1.005	0.003	0.003	1.004	0.003	0.003
	$\theta_{20}$	10	9.791	0.067	0.111	9.773	0.093	0.144	9.789	0.097	0.142
	$\theta_{30}$	100	101.8	13.03	16.26	101.6	14.41	16.89	101.8	15.93	19.19
	$\theta_{40}$	25	23.81	6.977	8.383	23.84	6.404	7.741	23.95	6.430	7.529
5	$\theta_{10}$	1	0.999	0.005	0.005	0.996	0.005	0.005	0.999	0.005	0.005
	$\theta_{20}$	10	9.915	0.018	0.025	9.904	0.019	0.028	9.897	0.022	0.033
	$\theta_{30}$	200	200.7	9.495	10.05	200.5	12.18	12.45	201.1	11.95	13.26
	$\theta_{40}$	25	24.65	5.888	6.010	24.69	6.352	6.446	24.56	5.486	5.682
6	$\theta_{10}$	1	0.992	0.010	0.010	0.991	0.010	0.011	0.990	0.013	0.013
	$\theta_{20}$	10	9.938	0.009	0.013	9.930	0.009	0.014	9.916	0.015	0.022
	$\theta_{30}$	300	300.6	10.64	10.95	300.8	9.580	10.26	300.5	10.55	10.85
	$\theta_{40}$	25	24.98	6.478	6.478	24.85	6.273	6.295	24.80	9.568	9.607
7	$\theta_{10}$	1	1.014	0.005	0.005	1.008	0.005	0.005	1.010	0.005	0.005
	$\theta_{20}$	10	9.661	0.220	0.335	9.708	0.183	0.268	9.636	0.222	0.355
	$\theta_{30}$	100	104.3	55.57	74.35	103.9	48.06	63.33	104.3	54.06	72.15
	$\theta_{40}$	45	43.23	22.35	25.48	43.27	18.75	21.74	42.79	21.14	26.05
8	$\theta_{10}$	1	1.002	0.008	0.008	1.009	0.008	0.008	1.009	0.011	0.011
	$\theta_{20}$	10	9.881	0.034	0.048	9.841	0.046	0.072	9.865	0.041	0.060
	$\theta_{30}$	200	201.9	24.76	28.20	202.5	23.56	29.80	202.0	23.88	28.07
	$\theta_{40}$	45	44.10	18.04	18.85	43.40	14.81	17.39	43.69	17.38	19.11
9	$\theta_{10}$	1	1.003	0.030	0.030	1.021	0.036	0.036	1.024	0.032	0.033
	$\theta_{20}$	10	9.910	0.018	0.026	9.895	0.022	0.033	9.900	0.023	0.033
	$\theta_{30}$	300	301.1	27.32	28.60	301.0	26.05	27.11	301.6	26.76	29.26
	$\theta_{40}$	45	43.84	18.10	19.44	44.24	21.05	21.62	43.90	21.91	23.12

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

시나리오 1, 2, 3을 보면  $\theta_{30}$ 가 작을수록 증가한다.  $\theta_{40}$ 의 변화에 따른 추정값의 분산과 평균제곱오차를 살펴보기 위해 시나리오 1, 4, 7을 보

표 3: 일반화 포아송자료의 일반화 포아송모형 적용 결과(계속)

Scenario	Param	Truth	k = 0.01			k = 0.1			k = 0.5		
			Mean	Var	MSE	Mean	Var	MSE	Mean	Var	MSE
1	$\theta_{10}$	1	1.001	0.003	0.003	0.999	0.003	0.003	1.002	0.008	0.008
	$\theta_{20}$	10	9.862	0.045	0.064	9.751	0.135	0.197	9.198	1.066	1.709
	$\theta_{30}$	100	100.2	2.273	2.312	100.3	6.665	6.778	100.9	47.50	48.34
	$\theta_{40}$	5	4.830	1.468	1.497	4.621	3.395	3.538	3.806	12.89	14.31
2	$\theta_{10}$	1	0.997	0.003	0.003	1.004	0.005	0.005	1.019	0.010	0.010
	$\theta_{20}$	10	9.887	0.023	0.036	9.817	0.067	0.101	9.288	0.650	1.156
	$\theta_{30}$	200	200.0	2.266	2.269	200.0	5.999	5.999	200.3	49.39	49.47
	$\theta_{40}$	5	5.034	1.457	1.458	4.788	3.234	3.279	4.074	16.63	17.49
3	$\theta_{10}$	1	1.005	0.006	0.006	0.992	0.008	0.008	1.001	0.016	0.016
	$\theta_{20}$	10	9.938	0.009	0.013	9.848	0.051	0.074	9.482	0.345	0.614
	$\theta_{30}$	300	300.0	2.440	2.441	300.1	6.496	6.511	300.0	62.92	62.92
	$\theta_{40}$	5	4.893	1.496	1.508	4.782	3.144	3.191	4.330	21.87	22.32
4	$\theta_{10}$	1	1.001	0.004	0.004	1.011	0.004	0.004	0.993	0.009	0.009
	$\theta_{20}$	10	9.775	0.110	0.161	9.610	0.314	0.466	9.120	1.318	2.092
	$\theta_{30}$	100	101.9	16.32	20.00	103.0	41.09	50.03	106.6	219.1	262.0
	$\theta_{40}$	25	24.09	7.829	8.666	23.54	14.93	17.06	21.97	80.24	89.43
5	$\theta_{10}$	1	1.003	0.005	0.005	1.010	0.006	0.006	1.012	0.025	0.025
	$\theta_{20}$	10	9.907	0.025	0.034	9.805	0.081	0.119	9.303	0.706	1.193
	$\theta_{30}$	200	200.8	12.71	13.36	201.6	28.15	30.67	204.3	210.9	229.6
	$\theta_{40}$	25	24.49	7.114	7.371	24.45	14.02	14.32	23.14	103.0	106.5
6	$\theta_{10}$	1	0.999	0.010	0.010	1.007	0.017	0.017	0.998	0.050	0.050
	$\theta_{20}$	10	9.915	0.016	0.023	9.841	0.045	0.071	9.376	0.451	0.840
	$\theta_{30}$	300	300.3	13.21	13.28	300.4	30.86	31.04	303.1	216.8	226.4
	$\theta_{40}$	25	24.65	7.976	8.096	24.74	16.71	16.78	23.75	129.0	130.51
7	$\theta_{10}$	1	1.005	0.004	0.004	1.008	0.006	0.006	1.000	0.016	0.016
	$\theta_{20}$	10	9.580	0.297	0.473	9.501	0.447	0.696	8.803	2.284	3.717
	$\theta_{30}$	100	105.1	77.85	104.0	106.7	114.1	159.1	114.3	643.7	847.0
	$\theta_{40}$	45	42.93	27.58	31.88	42.43	36.14	42.77	39.14	196.5	230.9
8	$\theta_{10}$	1	1.000	0.010	0.010	1.006	0.013	0.013	0.998	0.035	0.035
	$\theta_{20}$	10	9.839	0.053	0.079	9.712	0.177	0.260	9.220	0.944	1.553
	$\theta_{30}$	200	202.0	28.89	32.86	203.9	71.68	87.23	209.6	458.4	550.8
	$\theta_{40}$	45	44.11	19.63	20.42	43.05	34.96	38.78	41.66	208.8	219.9
9	$\theta_{10}$	1	1.034	0.037	0.039	1.006	0.058	0.058	0.928	0.159	0.208
	$\theta_{20}$	10	9.886	0.024	0.037	9.826	0.068	0.098	9.309	0.632	1.130
	$\theta_{30}$	300	301.3	31.16	33.02	301.8	59.61	62.75	308.3	338.1	425.7
	$\theta_{40}$	45	43.50	20.29	22.53	43.72	56.53	58.18	43.63	259.2	306.7

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

면  $\theta_{40}$ 의 값이 증가할수록 모든  $\hat{\Theta}$ 의 분산과 평균제곱오차는 증가하고 모든  $\hat{\Theta}$ 의 분산과 평균제곱오차의 차이 또한 증가한다. 그리고 일반화

포아송분포의 산포모수  $k$ 가 증가할수록 모든 추정값의 분산과 평균제곱오차가 증가하는데 특히 산포모수가 0.1인 경우와 0.5인 경우 매우 큰 것을 확인하였다.

그림 5는 산포모수가  $-0.0095$ 인 일반화 포아송자료 199개를 일반화 포아송모형에 적용하여 학습곡선의 평균적합모형, 95% 근사신뢰구간과 95% 경험적 신뢰구간을 구하여 그린 그림이다. 95% 근사신뢰구간은 식 (3.1)을 이용하여 구하였고, 95% 경험적 신뢰구간은 식 (3.2)를 이용하여 구하였다. 그림 5의 평균적합모형을 보면 참 모형과 매우 유사하게 나타났다. 두가지 신뢰구간은 초기 시점에서 근사신뢰구간보다 경험적 신뢰구간이 작은 값에서 추정되는 것으로 나타났지만 이후 서로 매우 유사한 형태를 가지고 있고 신뢰구간의 폭은 초기 시점에서  $\theta_{30}$ 가 감소할수록 넓어지고 학습곡선의 변곡점 이후  $\theta_{30}$ 가 증가할수록 넓어진다. 그리고 전 시점에서  $\theta_{40}$ 가 증가할수록 신뢰구간의 폭이 넓어졌다.

그림 6은 산포모수가  $-0.005$ 인 일반화 포아송자료 199개를 일반화 포아송모형에 적용하여 학습곡선의 평균적합모형, 95% 근사신뢰구간과 95% 경험적 신뢰구간을 구하여 그린 그림이다. 그림 6의 평균적합모형을 보면 그림 5와 같이 참 모형과 매우 유사하게 나타났다. 신뢰구간들도 초기 시점에서 차이가 있지만 시점이 지날수록 매우 유사하고, 신뢰구간의 폭은 초기 시점에서  $\theta_{30}$ 가 감소할수록 넓어지고 학습곡선의 변곡점 이후  $\theta_{30}$ 가 증가할수록 넓어진다. 그리고 전 시점에서  $\theta_{40}$ 가 증가할수록 신뢰구간의 폭이 넓어졌다. 그림 5와 비교해보았을 때 모든 시나리오에서 각 시점의 모든 신뢰구간이 더 넓어지는 경향이 있다. 즉, 산포모수의 값이 증가할수록 신뢰구간이 더 넓어졌다.

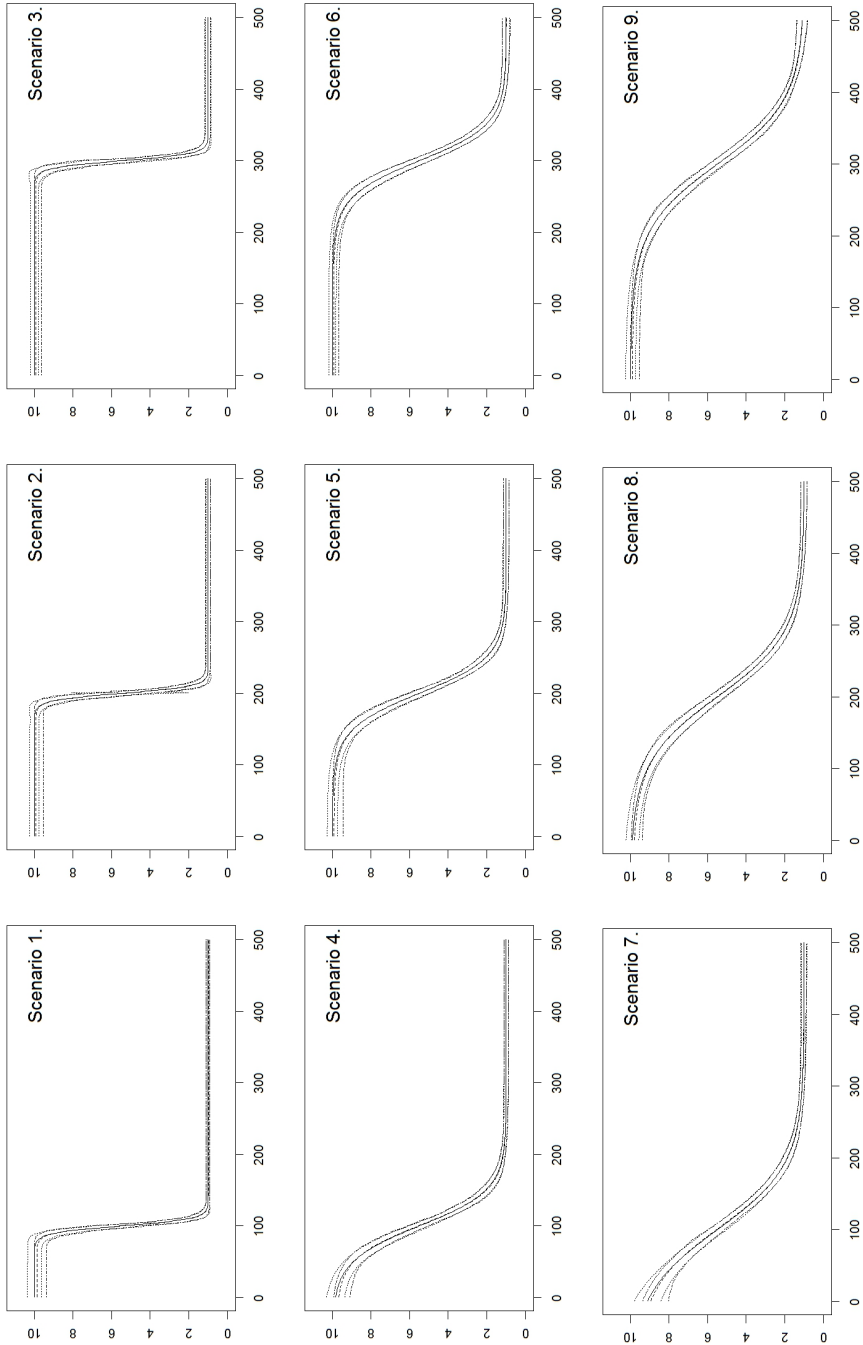


그림 5: 최대우도방법을 이용한 참값과 추정값, 그리고 95% 신뢰구간들(GP,  $k = -0.0095$ )  
(실선: 참값, 꺾은 선: 추정값, 점선: 95% 근사신뢰구간, 점-꺾은 선: 95% 경험적 신뢰구간)

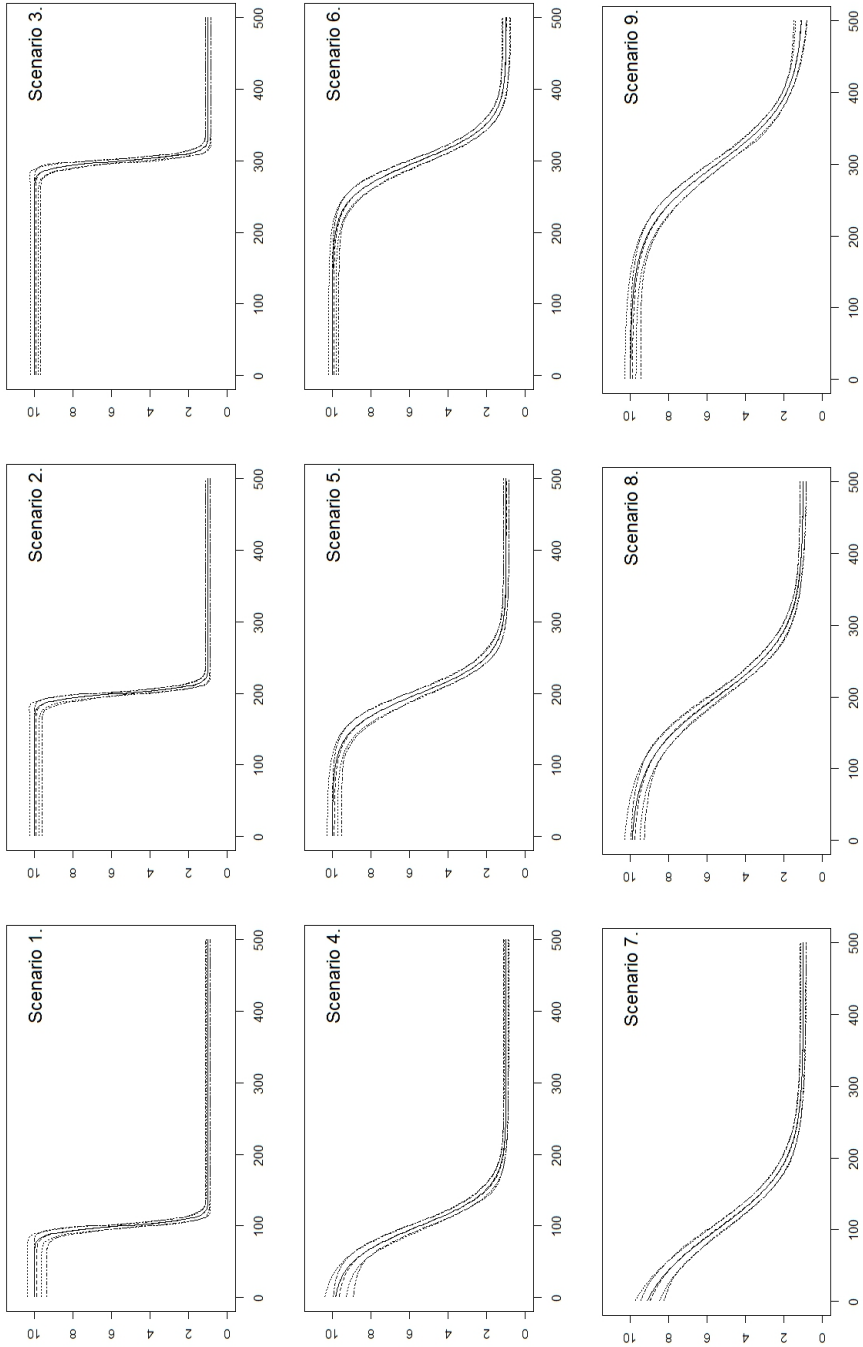


그림 6: 최대우도방법을 이용한 참값과 추정값, 그리고 95% 신뢰구간들 ( $GP, k = -0.005$ )  
 (실선: 참값, 꺾은 선: 추정값, 점선: 95% 근사신뢰구간, 점-꺾은 선: 95% 경험적 신뢰구간)

표 4: 일반화 음이항자료의 일반화 음이항모형 적용 결과

Scenario	Param	Truth	$\beta = 1$			$\beta = 3$			$\beta = 5$		
			Mean	Var	MSE	Mean	Var	MSE	Mean	Var	MSE
1	$\theta_{10}$	1	1.000	0.002	0.002	1.005	0.003	0.003	0.992	0.003	0.003
	$\theta_{20}$	10	9.838	0.056	0.083	9.826	0.061	0.091	9.823	0.067	0.099
	$\theta_{30}$	100	100.2	2.434	2.482	100.3	3.032	3.114	100.3	3.492	3.606
	$\theta_{40}$	5	4.810	1.384	1.420	4.779	1.243	1.292	4.680	1.838	1.940
2	$\theta_{10}$	1	0.999	0.004	0.004	1.005	0.005	0.005	1.008	0.003	0.003
	$\theta_{20}$	10	9.885	0.022	0.036	9.893	0.023	0.035	9.875	0.037	0.053
	$\theta_{30}$	200	200.2	2.212	2.251	200.1	2.795	2.812	200.4	3.674	3.837
	$\theta_{40}$	5	4.974	1.425	1.426	4.973	2.049	2.050	4.790	1.758	1.802
3	$\theta_{10}$	1	1.003	0.007	0.007	0.994	0.006	0.006	1.005	0.007	0.007
	$\theta_{20}$	10	9.938	0.011	0.015	9.902	0.021	0.031	9.905	0.019	0.028
	$\theta_{30}$	300	300.2	1.876	1.911	300.2	2.781	2.820	300.1	3.835	3.846
	$\theta_{40}$	5	4.933	1.479	1.484	4.845	1.635	1.659	4.824	2.490	2.521
4	$\theta_{10}$	1	1.004	0.004	0.004	1.002	0.004	0.004	1.004	0.003	0.003
	$\theta_{20}$	10	9.753	0.113	0.173	9.770	0.121	0.174	9.661	0.187	0.301
	$\theta_{30}$	100	101.7	18.27	21.26	102.0	22.38	26.20	102.6	24.56	31.40
	$\theta_{40}$	25	24.16	8.553	9.254	24.12	9.076	9.858	23.89	10.31	11.55
5	$\theta_{10}$	1	0.999	0.005	0.005	1.001	0.005	1.004	1.011	0.006	0.006
	$\theta_{20}$	10	9.886	0.027	0.041	9.863	0.038	0.057	9.843	0.055	0.079
	$\theta_{30}$	200	201.2	14.05	15.47	201.2	16.45	17.79	201.6	16.29	18.93
	$\theta_{40}$	25	24.58	6.222	6.400	24.34	8.268	8.703	23.67	10.07	11.83
6	$\theta_{10}$	1	1.002	0.012	0.012	1.004	0.012	0.012	1.008	0.011	0.011
	$\theta_{20}$	10	9.907	0.017	0.026	9.900	0.027	0.037	9.907	0.028	0.036
	$\theta_{30}$	300	300.8	11.42	12.13	300.7	15.48	15.96	300.6	18.07	18.39
	$\theta_{40}$	25	24.39	7.773	8.144	24.56	8.256	8.454	24.55	11.17	11.38
7	$\theta_{10}$	1	1.006	0.005	0.005	1.008	0.005	0.005	1.012	0.006	0.006
	$\theta_{20}$	10	9.637	0.247	0.378	9.534	0.364	0.581	9.498	0.434	0.686
	$\theta_{30}$	100	104.7	63.45	85.99	105.8	83.21	117.2	105.7	103.9	135.9
	$\theta_{40}$	45	42.82	21.27	26.01	42.15	31.23	39.36	42.14	34.96	43.16
8	$\theta_{10}$	1	0.992	0.011	0.011	1.014	0.010	0.011	1.018	0.011	0.011
	$\theta_{20}$	10	9.848	0.048	0.071	9.800	0.077	0.117	9.794	0.092	0.134
	$\theta_{30}$	200	202.5	28.26	34.52	202.66	39.17	46.22	202.8	43.65	51.47
	$\theta_{40}$	45	44.06	14.43	15.32	43.43	19.95	22.43	43.37	23.27	25.93
9	$\theta_{10}$	1	1.016	0.031	0.031	1.016	0.046	0.046	1.005	0.045	0.045
	$\theta_{20}$	10	9.892	0.024	0.035	9.874	0.031	0.047	9.877	0.040	0.055
	$\theta_{30}$	300	301.4	28.64	30.60	301.22	35.64	37.12	301.3	41.19	43.01
	$\theta_{40}$	45	43.71	19.65	21.32	44.10	29.35	30.17	44.05	37.29	38.20

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

표 4와 표 5는 9개의 시나리오에 따라 일반화 음이항자료를 일반화 음이항모형에 적합했을 때 모수의 참값과 추정값의 평균, 분산, 평

표 5: 일반화 음이항자료의 일반화 음이항모형 적용 결과(계속)

Scenario	Param	Truth	$\beta = 10$			$\beta = 20$			$\beta = 30$		
			Mean	Var	MSE	Mean	Var	MSE	Mean	Var	MSE
1	$\theta_{10}$	1	1.006	0.003	0.003	0.999	0.005	0.005	0.998	0.003	0.003
	$\theta_{20}$	10	9.733	0.138	0.209	9.487	0.544	0.807	9.617	0.308	0.455
	$\theta_{30}$	100	100.4	6.076	6.231	101.2	23.58	25.09	100.2	13.09	13.14
	$\theta_{40}$	5	4.797	2.823	2.865	3.915	8.833	10.01	4.562	5.481	5.673
2	$\theta_{10}$	1	1.000	0.004	0.004	1.001	0.008	0.008	1.005	0.006	0.006
	$\theta_{20}$	10	9.825	0.060	0.091	9.692	0.221	0.316	9.672	0.176	0.283
	$\theta_{30}$	200	200.4	5.408	5.538	200.6	35.42	35.79	200.1	15.64	15.66
	$\theta_{40}$	5	4.514	3.016	3.251	4.038	11.496	12.42	4.578	6.528	6.706
3	$\theta_{10}$	1	1.010	0.006	0.006	0.999	0.011	0.011	1.006	0.010	0.010
	$\theta_{20}$	10	9.866	0.046	0.063	9.679	0.194	0.297	9.767	0.104	0.158
	$\theta_{30}$	300	300.1	6.539	6.549	300.6	35.90	36.24	300.0	20.63	20.63
	$\theta_{40}$	5	4.859	3.978	3.997	4.467	10.45	10.73	4.776	7.396	7.446
4	$\theta_{10}$	1	1.001	0.004	0.004	0.997	0.007	0.007	1.003	0.005	0.005
	$\theta_{20}$	10	9.623	0.305	0.447	9.338	0.868	1.306	9.446	0.529	0.836
	$\theta_{30}$	100	101.9	52.73	56.29	105.4	106.8	135.9	104.9	72.382	95.905
	$\theta_{40}$	25	24.36	14.66	15.07	22.65	37.92	43.43	22.61	26.076	31.799
5	$\theta_{10}$	1	0.998	0.007	0.007	0.999	0.011	0.011	0.996	0.008	0.008
	$\theta_{20}$	10	9.784	0.103	0.150	9.695	0.261	0.354	9.717	0.173	0.253
	$\theta_{30}$	200	201.3	29.83	31.55	202.8	98.27	105.7	201.3	56.63	58.44
	$\theta_{40}$	25	24.60	14.07	14.23	23.97	41.70	42.77	24.18	24.43	25.11
6	$\theta_{10}$	1	1.000	0.015	0.015	1.007	0.026	0.026	1.012	0.019	0.019
	$\theta_{20}$	10	9.857	0.055	0.075	9.680	0.234	0.336	9.738	0.136	0.205
	$\theta_{30}$	300	300.5	28.80	29.02	302.1	91.04	95.56	301.3	49.14	50.94
	$\theta_{40}$	25	24.60	17.65	17.80	23.34	52.79	55.56	23.70	25.39	27.07
7	$\theta_{10}$	1	1.009	0.007	0.007	1.002	0.010	0.010	1.009	0.008	0.008
	$\theta_{20}$	10	9.366	0.618	1.021	9.075	1.291	2.145	9.288	0.853	1.360
	$\theta_{30}$	100	108.6	133.2	207.0	112.7	332.0	492.9	108.4	211.6	283.0
	$\theta_{40}$	45	41.89	42.54	52.24	41.05	97.85	113.5	42.04	63.36	72.15
8	$\theta_{10}$	1	0.993	0.013	0.013	0.993	0.013	0.013	0.994	0.016	0.016
	$\theta_{20}$	10	9.702	0.165	0.254	9.702	0.165	0.254	9.626	0.314	0.454
	$\theta_{30}$	200	204.1	72.59	89.23	204.1	72.59	89.23	204.4	146.1	165.7
	$\theta_{40}$	45	43.54	36.53	38.66	43.54	36.53	38.66	43.72	70.85	72.50
9	$\theta_{10}$	1	1.023	0.058	0.058	1.012	0.103	0.109	1.012	0.073	0.073
	$\theta_{20}$	10	9.811	0.078	0.114	9.615	0.247	0.396	9.745	0.142	0.207
	$\theta_{30}$	300	302.4	66.77	72.57	306.7	170.1	216.1	302.5	117.8	124.1
	$\theta_{40}$	45	43.34	54.42	57.17	41.79	142.5	158.4	43.58	88.04	90.06

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

균제곱오차를 정리한 표이다. 결과를 보면 일반화 포아송자료를 일반화 포아송모형에 적합했을 때와 같이 모든 추정값의 평균이 모수의

참값과 유사하게 나타났다.  $\theta_{30}$ 의 변화에 따른 분산과 평균제곱오차를 살펴보기 위해 시나리오 1, 2, 3을 보면  $\theta_{30}$ 가 증가할수록  $\hat{\theta}_2$ 와  $\hat{\theta}_3$ 의 분산과 평균제곱오차는 감소한다. 이러한  $\hat{\theta}_2$ 와  $\hat{\theta}_3$ 의 분산과 평균제곱오차가 감소하는 형상은 시나리오 1, 2, 3보다 시나리오 7, 8, 9에서 확연히 드러난다. 그와 반대로  $\hat{\theta}_1$ 과  $\hat{\theta}_4$ 의 분산과 평균제곱오차는 증가한다. 그리고 추정값의 분산과 평균제곱오차를 비교해보면  $\hat{\theta}_1$ 의 분산과 평균제곱오차의 차이가 없고 비교적  $\hat{\theta}_3$ 와  $\hat{\theta}_4$ 의 분산과 평균제곱오차의 차이가 있다. 그 차이는 시나리오 1, 2, 3을 보면  $\theta_{30}$ 가 작을수록 증가한다.  $\theta_{40}$ 의 변화에 따른 추정값의 분산과 평균제곱오차를 살펴보기 위해 시나리오 1, 4, 7을 보면  $\theta_{40}$ 의 값이 증가할수록 모든  $\hat{\theta}$ 의 분산과 평균제곱오차는 증가하고 모든  $\hat{\theta}$ 의 분산과 평균제곱오차의 차이도 증가한다. 그리고 일반화 음이항분포의 산포모수  $\beta$ 의 값이 증가할수록 분산과 평균제곱오차가 증가하는 것을 확인하였다.

그림 7는 산포모수가 1인 일반화 음이항자료 199개를 일반화 음이항모형에 적용했을 때 얻은 학습곡선의 평균적합모형, 95% 근사신뢰구간과 95% 경험적 신뢰구간을 구하여 그린 그림이다. 95% 근사신뢰구간은 식 (3.1)을 이용하여 구하였고, 95% 경험적 신뢰구간은 식 (3.2)을 이용하여 구하였다. 그림 7을 보면 모든 시나리오에서 평균적합모형이 참 모형과 매우 유사하게 나타났다. 두가지 신뢰구간은 초기 시점에서 근사신뢰구간보다 경험적 신뢰구간이 작은 값에서 추정되는 것으로 나타났지만 시점이 지날수록 서로 매우 유사한 형태를 가지고 있다.

그림 8은 산포모수가 3인 일반화 음이항자료 199개를 일반화 음이항모형에 적용했을 때 얻은 학습곡선의 평균적합모형, 95% 근사신뢰구간과 95% 경험적 신뢰구간을 구하여 그린 그림이다. 그림 8을 보

면 그림 7과 같이 모든 시나리오에서 평균적합모형이 참 모형과 매우 유사하게 나타났고 두가지 신뢰구간은 초기 시점에서 차이가 있지만 시점이 지날수록 매우 유사하다. 그림 7와 비교해보았을 때 모든 시나리오에서 각 시점의 모든 신뢰구간이 더 넓어지는 경향이 있다. 즉, 일반화 음이항모형에서도 산포모수의 값이 증가할수록 신뢰구간이 더 넓어졌다.

모의실험 결과, 통계적인 모형의 특성은  $\hat{\Theta}$ 의 평균이 참값과 매우 유사하고  $\theta_{30}$ 가 감소할수록,  $\theta_{40}$ 가 증가할수록, 산포모수가 증가할수록 추정값의 분산과 평균제곱오차가 증가하고 분산과 표준제곱오차의 차이 또한 증가한다. 그리고 그림 5, 6, 7, 8을 봤을 때 평균적합모형이 참 모형과 매우 유사하게 나타났고 95% 근사신뢰구간과 95% 경험적 신뢰구간은 초기시점에서 차이가 있지만 시점이 지날수록 매우 유사했다. 두가지 신뢰구간은 초기시점에는  $\theta_{30}$ 가 감소할수록 넓고 학습곡선의 변곡점 이후  $\theta_{30}$ 가 증가할수록 넓어지는 특징을 보였고, 전 시점을 봤을 때  $\theta_{40}$ 가 증가할수록 넓어지고 산포모수의 값이 증가할수록 넓어졌다.

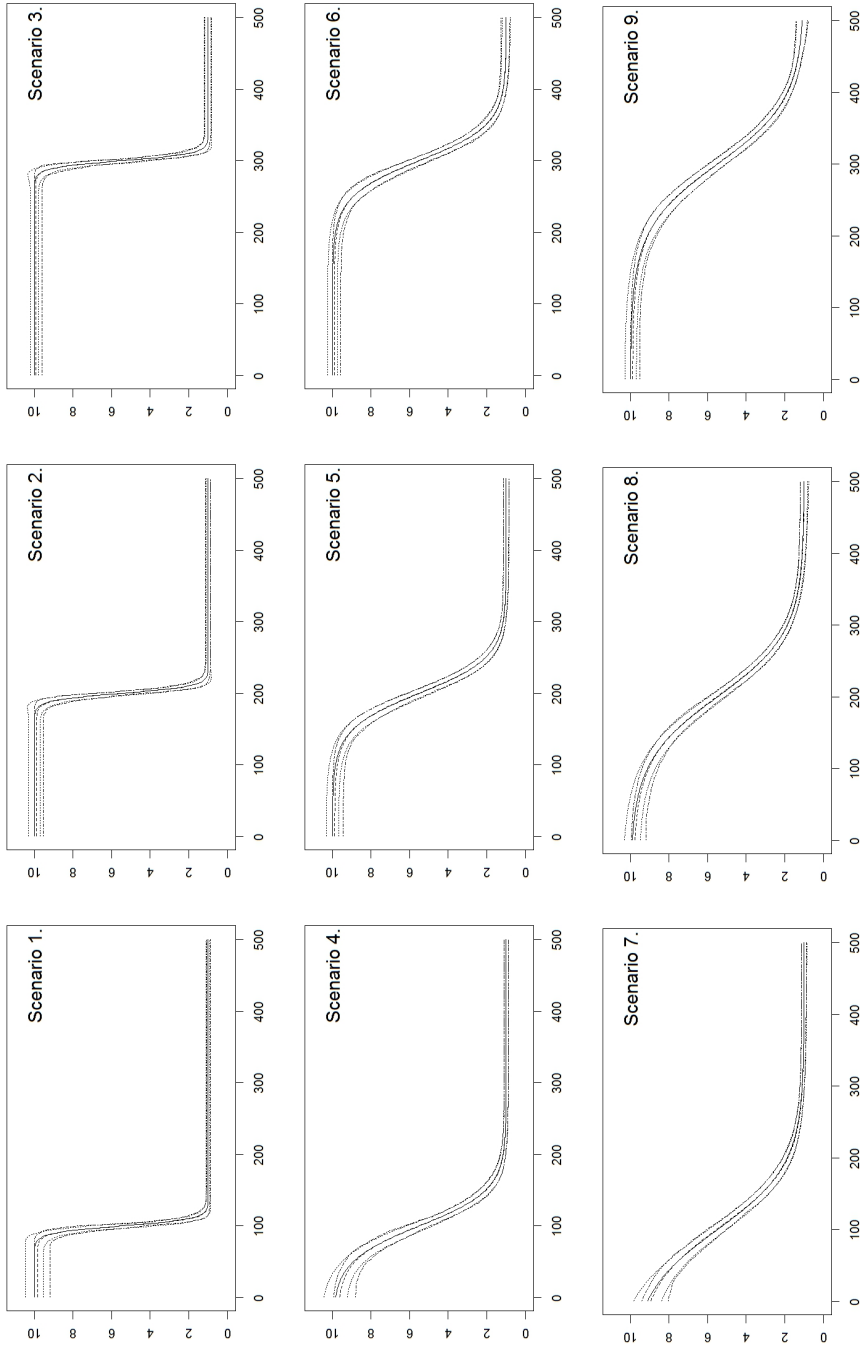


그림 7: 최 대우도방법을 이용한 참값과 추정값, 그리고 95% 신뢰구간들( $GNB, \beta = 1$ )  
(실선: 참값, 꺾은 선: 추정값, 점선: 95% 근사신뢰구간, 점-꺾은 선: 95% 경험적 신뢰구간)

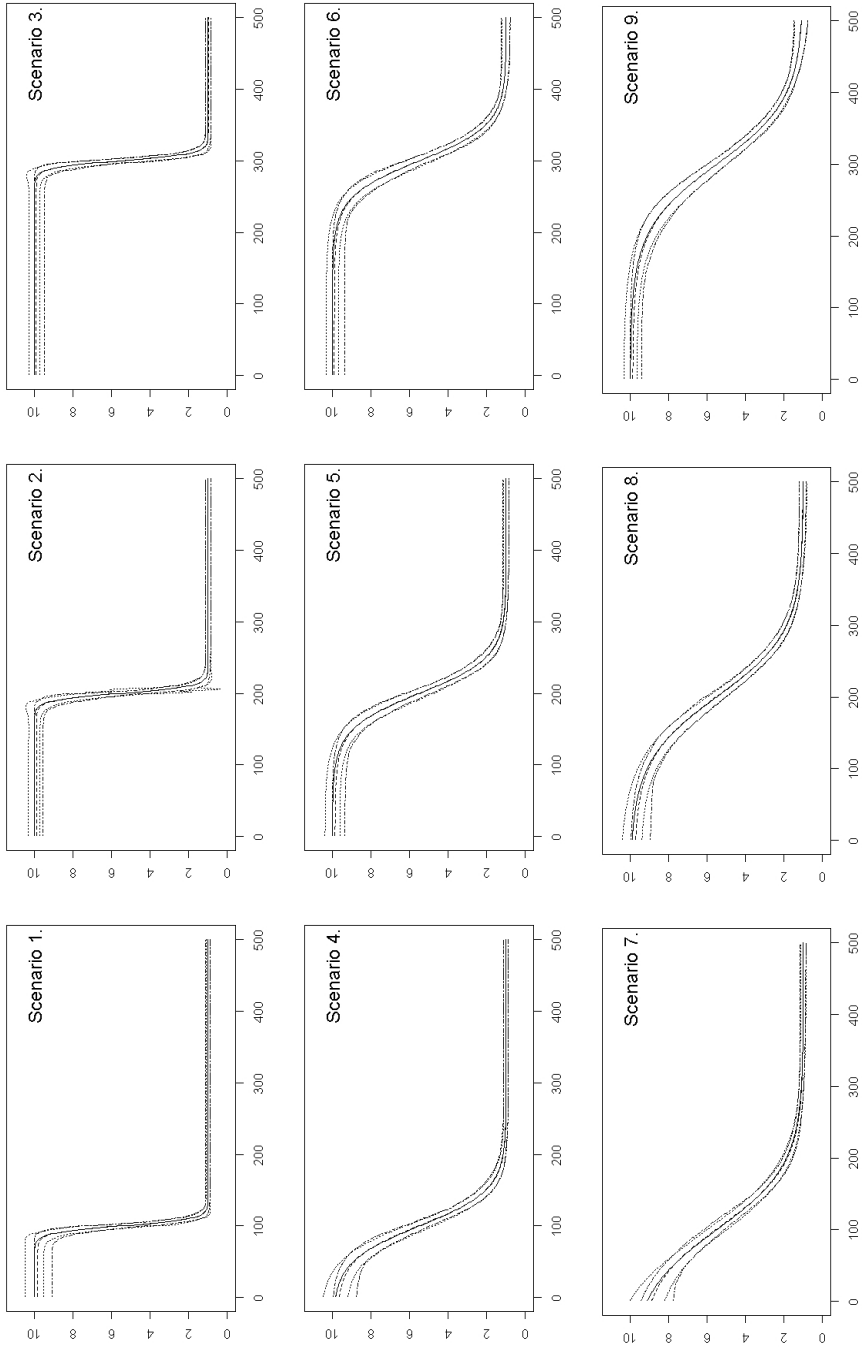


그림 8: 최 대우도방법을 이용한 참값과 추정값, 그리고 95% 신뢰구간들( $GNB, \beta = 3$ )  
 (실선: 참값, 꺾은 선: 추정값, 점선: 95% 근사신뢰구간, 점-꺾은 선: 95% 경험적 신뢰구간)

### 3.2.2 모형의 성능 비교

앞장에서는 일반화 포아송자료와 일반화 음이항자료를 만들어 일반화 포아송모형과 일반화 음이항모형에 적합을 시켜 분석하였다. 모형의 성능 비교에서는 일반화 포아송자료와 일반화 음이항자료를 포아송모형, 음이항모형, 일반화 포아송모형과 일반화 음이항모형에 적합 시켜 모형선택기준(AIC, BIC, AICc)으로 비교하였다.

표 6과 표 7은 일반화 포아송자료를 4개의 모형 중 모형선택기준이 가장 작은 횃수를 세어 비율로 나타낸 것이다. 예를 들어 일반화 포아송자료에서 시나리오 1이고, 산포모수가  $-0.0095$ 이고, 모형선택기준이 AIC일 때 일반화 포아송모형을 선택한 비율은 0.61이다. 표를 살펴보면 산포모수가 음수일 때  $\theta_{30}$ 이 증가할수록  $\theta_{40}$ 가 감소할수록 포아송모형과 음이항모형이 선택되는 빈도는 감소하고 일반화 포아송모형이 선택되는 빈도가 증가한다. 그리고 모든 시나리오에서 일반화 음이항모형은 선택되지 않았다. 산포모수가 0.01인 경우에는  $\theta_{30}$ 이 증가할수록 포아송모형과 음이항모형이 선택되는 빈도가 감소하고 일반화 포아송모형과 일반화 음이항모형이 선택되는 빈도는 증가했고,  $\theta_{40}$ 의 변화에는 큰 영향을 받지 않았다. 산포모수가 0.1이상 일 때는  $\theta_{30}$ 나  $\theta_{40}$ 의 변화에 영향을 받지 않았다. 전체적으로 보면  $\theta_{30}$ 가 증가할수록  $\theta_{40}$ 가 감소할수록 일반화 포아송모형이 선택되는 빈도가 증가한다. 그리고 산포모수가  $-0.005$ 일 때는 포아송모형이 선택되지만  $-0.0095$ 일 때는 일반화 포아송모형이 선택되었고 산포모수가 양수일 때 0.01 경우 일반화 포아송모형보다 음이항모형이 선택되지만 산포모수 0.1 부터는 포아송모형이나 음이항모형은 선택이 안되고 일반화 포아송모형과 일반화 음이항모형이 약 50%씩 선택되는 경향이 있다.

표 6: 일반화 포아송자료에 대한 모형 성능 비교

Scenario	Model	$k = -0.0095$			$k = -0.005$			$k = 0$		
		AIC	BIC	AICc	AIC	BIC	AICc	AIC	BIC	AICc
1	<i>P</i>	0.38	0.92	0.38	0.60	0.87	0.61	0.60	0.69	0.60
	<i>NB</i>	0.01	0.01	0.01	0.10	0.10	0.10	0.29	0.31	0.29
	<i>GP</i>	0.61	0.07	0.61	0.30	0.03	0.29	0.10	0.01	0.10
	<i>GNB</i>	–	–	–	–	–	–	0.01	–	0.01
2	<i>P</i>	0.21	0.69	0.21	0.60	0.89	0.61	0.60	0.71	0.60
	<i>NB</i>	0.01	0.01	0.01	0.04	0.04	0.04	0.27	0.28	0.27
	<i>GP</i>	0.78	0.30	0.78	0.36	0.08	0.36	0.12	0.01	0.12
	<i>GNB</i>	–	–	–	–	–	–	0.01	0.01	0.01
3	<i>P</i>	0.17	0.55	0.17	0.48	0.82	0.48	0.66	0.75	0.67
	<i>NB</i>	–	–	–	0.02	0.02	0.02	0.24	0.24	0.24
	<i>GP</i>	0.83	0.45	0.83	0.50	0.16	0.50	0.10	0.01	0.10
	<i>GNB</i>	–	–	–	–	–	–	–	–	–
4	<i>P</i>	0.46	0.95	0.47	0.53	0.82	0.55	0.59	0.72	0.59
	<i>NB</i>	0.03	0.03	0.03	0.18	0.18	0.18	0.28	0.28	0.28
	<i>GP</i>	0.51	0.02	0.50	0.29	–	0.27	0.14	–	0.13
	<i>GNB</i>	–	–	–	–	–	–	–	–	–
5	<i>P</i>	0.27	0.72	0.28	0.55	0.87	0.55	0.60	0.70	0.60
	<i>NB</i>	0.01	0.01	0.01	0.07	0.07	0.07	0.30	0.30	0.30
	<i>GP</i>	0.72	0.27	0.72	0.38	0.06	0.38	0.11	–	0.11
	<i>GNB</i>	–	–	–	–	–	–	–	–	–
6	<i>P</i>	0.13	0.55	0.14	0.54	0.85	0.55	0.64	0.74	0.64
	<i>NB</i>	–	–	–	0.03	0.03	0.03	0.25	0.25	0.25
	<i>GP</i>	0.87	0.45	0.86	0.43	0.12	0.42	0.11	0.01	0.11
	<i>GNB</i>	–	–	–	–	–	–	–	–	–
7	<i>P</i>	0.53	0.92	0.53	0.56	0.80	0.57	0.62	0.73	0.63
	<i>NB</i>	0.06	0.06	0.06	0.19	0.19	0.19	0.26	0.27	0.26
	<i>GP</i>	0.41	0.02	0.41	0.25	0.01	0.24	0.11	0.01	0.10
	<i>GNB</i>	–	–	–	0.01	–	0.01	0.01	–	0.01
8	<i>P</i>	0.29	0.80	0.29	0.63	0.89	0.65	0.62	0.70	0.63
	<i>NB</i>	0.01	0.01	0.01	0.06	0.06	0.06	0.28	0.29	0.28
	<i>GP</i>	0.70	0.19	0.70	0.31	0.05	0.30	0.09	0.01	0.08
	<i>GNB</i>	–	–	–	–	–	–	0.02	–	0.02
9	<i>P</i>	0.13	0.55	0.13	0.52	0.88	0.52	0.62	0.70	0.63
	<i>NB</i>	0.01	0.01	0.01	0.03	0.03	0.03	0.28	0.28	0.28
	<i>GP</i>	0.87	0.44	0.86	0.46	0.09	0.46	0.10	0.02	0.09
	<i>GNB</i>	–	–	–	–	–	–	–	–	–

P, Poisson; NB, negative binomial; GP, generalized Poisson; GNB, generalized negative binomial

표 8과 표 9를 보면 일반화 포아송자료와 같이 일반화 음이항자료를 4개의 모형 중 모형선택기준이 가장 작은 횃수를 세어 비율로 나

표 7: 일반화 포아송자료에 대한 모형 성능 비교(계속)

Scenario	Model	$k = 0.01$			$k = 0.1$			$k = 0.5$		
		AIC	BIC	AICc	AIC	BIC	AICc	AIC	BIC	AICc
1	<i>P</i>	0.19	0.20	0.19	-	-	-	-	-	-
	<i>NB</i>	0.63	0.79	0.64	-	-	-	0.01	0.01	0.01
	<i>GP</i>	0.10	0.02	0.09	0.50	0.50	0.50	0.50	0.50	0.50
	<i>GNB</i>	0.08	-	0.08	0.50	0.50	0.50	0.50	0.50	0.50
2	<i>P</i>	0.11	0.12	0.11	-	-	-	-	-	-
	<i>NB</i>	0.57	0.84	0.57	-	-	-	-	-	-
	<i>GP</i>	0.19	0.03	0.19	0.57	0.57	0.57	0.59	0.59	0.59
	<i>GNB</i>	0.13	0.02	0.13	0.43	0.43	0.43	0.41	0.41	0.41
3	<i>P</i>	0.02	0.02	0.02	-	-	-	-	-	-
	<i>NB</i>	0.61	0.86	0.61	-	-	-	-	-	-
	<i>GP</i>	0.16	0.06	0.15	0.51	0.51	0.51	0.57	0.57	0.57
	<i>GNB</i>	0.22	0.07	0.22	0.49	0.49	0.49	0.43	0.43	0.43
4	<i>P</i>	0.22	0.22	0.22	-	-	-	-	-	-
	<i>NB</i>	0.61	0.76	0.62	-	-	-	0.01	0.01	0.01
	<i>GP</i>	0.11	0.02	0.10	0.49	0.49	0.49	0.47	0.47	0.47
	<i>GNB</i>	0.06	0.01	0.06	0.51	0.51	0.51	0.52	0.52	0.52
5	<i>P</i>	0.14	0.14	0.14	-	-	-	-	-	-
	<i>NB</i>	0.63	0.81	0.64	-	-	-	-	-	-
	<i>GP</i>	0.12	0.04	0.12	0.55	0.55	0.55	0.52	0.52	0.52
	<i>GNB</i>	0.11	0.02	0.11	0.45	0.45	0.45	0.48	0.48	0.48
6	<i>P</i>	0.04	0.04	0.04	-	-	-	-	-	-
	<i>NB</i>	0.61	0.88	0.61	-	-	-	-	-	-
	<i>GP</i>	0.18	0.03	0.18	0.53	0.53	0.53	0.48	0.48	0.48
	<i>GNB</i>	0.17	0.06	0.17	0.47	0.47	0.47	0.52	0.52	0.52
7	<i>P</i>	0.24	0.24	0.24	-	-	-	-	-	-
	<i>NB</i>	0.57	0.74	0.59	-	-	-	-	-	-
	<i>GP</i>	0.09	0.02	0.09	0.48	0.48	0.48	0.51	0.51	0.51
	<i>GNB</i>	0.11	0.01	0.09	0.52	0.52	0.52	0.49	0.49	0.49
8	<i>P</i>	0.10	0.11	0.10	-	-	-	-	-	-
	<i>NB</i>	0.64	0.86	0.65	-	-	-	-	-	-
	<i>GP</i>	0.10	0.02	0.10	0.46	0.46	0.46	0.49	0.49	0.49
	<i>GNB</i>	0.16	0.02	0.16	0.54	0.54	0.54	0.51	0.51	0.51
9	<i>P</i>	0.04	0.04	0.04	-	-	-	-	-	-
	<i>NB</i>	0.59	0.88	0.60	-	-	-	-	-	-
	<i>GP</i>	0.14	0.03	0.14	0.53	0.53	0.53	0.43	0.43	0.43
	<i>GNB</i>	0.23	0.05	0.22	0.47	0.47	0.47	0.57	0.57	0.57

P, Poisson; NB, negative binomial; GP, generalized Poisson; GNB, generalized negative binomial

타낸 것이다. 표를 보면 산포모수가 1인 경우에는  $\theta_{30}$ 가 증가할수록  $\theta_{40}$ 가 감소할수록 포아송모형이 선택되는 빈도는 감소하고 일반화 포

표 8: 일반화 음이항자료에 대한 모형 성능 비교

Scenario		$\beta = 1$			$\beta = 3$			$\beta = 5$		
		AIC	BIC	AICc	AIC	BIC	AICc	AIC	BIC	AICc
1	<i>P</i>	0.38	0.43	0.38	–	–	–	–	–	–
	<i>NB</i>	0.55	0.56	0.55	0.21	0.57	0.22	0.01	0.03	0.01
	<i>GP</i>	0.06	–	0.06	0.38	0.22	0.38	0.47	0.47	0.47
	<i>GNB</i>	0.01	0.01	0.01	0.41	0.21	0.41	0.52	0.50	0.52
2	<i>P</i>	0.32	0.33	0.32	–	–	–	–	–	–
	<i>NB</i>	0.61	0.67	0.61	0.02	0.17	0.02	–	–	–
	<i>GP</i>	0.05	–	0.05	0.51	0.43	0.51	0.50	0.50	0.50
	<i>GNB</i>	0.03	–	0.03	0.48	0.40	0.48	0.50	0.50	0.50
3	<i>P</i>	0.25	0.27	0.25	–	–	–	–	–	–
	<i>NB</i>	0.66	0.72	0.67	0.01	0.05	0.01	–	–	–
	<i>GP</i>	0.07	0.01	0.07	0.53	0.51	0.53	0.52	0.52	0.52
	<i>GNB</i>	0.02	0.01	0.02	0.47	0.45	0.47	0.48	0.48	0.48
4	<i>P</i>	0.44	0.48	0.44	0.01	0.01	0.01	–	–	–
	<i>NB</i>	0.49	0.52	0.49	0.25	0.55	0.25	–	0.07	–
	<i>GP</i>	0.06	–	0.06	0.38	0.23	0.38	0.48	0.45	0.48
	<i>GNB</i>	0.02	–	0.02	0.37	0.22	0.37	0.52	0.49	0.52
5	<i>P</i>	0.34	0.36	0.34	–	–	–	–	–	–
	<i>NB</i>	0.58	0.63	0.58	0.02	0.22	0.02	–	–	–
	<i>GP</i>	0.06	0.01	0.06	0.50	0.43	0.50	0.48	0.48	0.48
	<i>GNB</i>	0.03	–	0.03	0.49	0.35	0.48	0.52	0.52	0.52
6	<i>P</i>	0.32	0.33	0.32	–	–	–	–	–	–
	<i>NB</i>	0.63	0.67	0.63	–	0.04	–	–	–	–
	<i>GP</i>	0.03	–	0.03	0.47	0.45	0.47	0.51	0.51	0.51
	<i>GNB</i>	0.03	–	0.03	0.53	0.51	0.53	0.49	0.49	0.49
7	<i>P</i>	0.36	0.42	0.36	0.02	0.02	0.02	–	–	–
	<i>NB</i>	0.53	0.57	0.53	0.18	0.54	0.18	–	0.05	–
	<i>GP</i>	0.09	0.01	0.09	0.43	0.25	0.43	0.47	0.46	0.47
	<i>GNB</i>	0.02	0.01	0.02	0.38	0.19	0.37	0.53	0.49	0.53
8	<i>P</i>	0.38	0.39	0.38	–	–	–	–	–	–
	<i>NB</i>	0.56	0.60	0.56	0.06	0.23	0.06	–	–	–
	<i>GP</i>	0.05	0.01	0.05	0.44	0.37	0.44	0.47	0.47	0.47
	<i>GNB</i>	0.02	–	0.02	0.50	0.40	0.50	0.53	0.53	0.53
9	<i>P</i>	0.27	0.29	0.27	–	–	–	–	–	–
	<i>NB</i>	0.67	0.71	0.67	0.01	0.10	0.01	–	–	–
	<i>GP</i>	0.04	–	0.04	0.51	0.45	0.51	0.50	0.50	0.50
	<i>GNB</i>	0.02	0.01	0.02	0.48	0.45	0.48	0.50	0.50	0.50

P, Poisson; NB, negative binomial; GP, generalized Poisson; GNB, generalized negative binomial

아송과 일반화 음이항모형은 선택이 안되고 음이항모형이 선택되는 빈도가 증가한다. 그리고 산포모수가 3인 경우  $\theta_{30}$ 가 증가할수록  $\theta_{40}$ 가

표 9: 일반화 음이항자료에 대한 모형 성능 비교(계속)

Scenario		$\beta = 10$			$\beta = 20$			$\beta = 30$		
		AIC	BIC	AICc	AIC	BIC	AICc	AIC	BIC	AICc
1	<i>P</i>	-	-	-	-	-	-	-	-	-
	<i>NB</i>	-	-	-	-	-	-	-	-	-
	<i>GP</i>	0.56	0.56	0.56	0.50	0.50	0.50	0.56	0.56	0.56
	<i>GNB</i>	0.44	0.44	0.44	0.50	0.50	0.50	0.44	0.44	0.44
2	<i>P</i>	-	-	-	-	-	-	-	-	-
	<i>NB</i>	-	-	-	-	-	-	-	-	-
	<i>GP</i>	0.48	0.48	0.48	0.57	0.57	0.57	0.42	0.42	0.42
	<i>GNB</i>	0.52	0.52	0.52	0.43	0.43	0.43	0.58	0.58	0.58
3	<i>P</i>	-	-	-	-	-	-	-	-	-
	<i>NB</i>	-	-	-	-	-	-	-	-	-
	<i>GP</i>	0.56	0.56	0.56	0.52	0.52	0.52	0.62	0.62	0.62
	<i>GNB</i>	0.44	0.44	0.44	0.48	0.48	0.48	0.38	0.38	0.38
4	<i>P</i>	-	-	-	-	-	-	-	-	-
	<i>NB</i>	-	-	-	-	-	-	-	-	-
	<i>GP</i>	0.49	0.49	0.49	0.47	0.47	0.47	0.54	0.54	0.54
	<i>GNB</i>	0.51	0.51	0.51	0.53	0.53	0.53	0.46	0.46	0.46
5	<i>P</i>	-	-	-	-	-	-	-	-	-
	<i>NB</i>	-	-	-	-	-	-	-	-	-
	<i>GP</i>	0.51	0.51	0.51	0.55	0.55	0.55	0.43	0.43	0.43
	<i>GNB</i>	0.49	0.49	0.49	0.45	0.45	0.45	0.57	0.57	0.57
6	<i>P</i>	-	-	-	-	-	-	-	-	-
	<i>NB</i>	-	-	-	-	-	-	-	-	-
	<i>GP</i>	0.45	0.45	0.45	0.45	0.45	0.45	0.53	0.53	0.53
	<i>GNB</i>	0.55	0.55	0.55	0.55	0.55	0.55	0.47	0.47	0.47
7	<i>P</i>	-	-	-	-	-	-	-	-	-
	<i>NB</i>	-	-	-	-	-	-	-	-	-
	<i>GP</i>	0.50	0.50	0.50	0.47	0.47	0.47	0.49	0.49	0.49
	<i>GNB</i>	0.50	0.50	0.50	0.53	0.53	0.53	0.51	0.51	0.51
8	<i>P</i>	-	-	-	-	-	-	-	-	-
	<i>NB</i>	-	-	-	-	-	-	-	-	-
	<i>GP</i>	0.51	0.51	0.51	0.55	0.55	0.55	0.51	0.51	0.51
	<i>GNB</i>	0.49	0.49	0.49	0.45	0.45	0.45	0.49	0.49	0.49
9	<i>P</i>	-	-	-	-	-	-	-	-	-
	<i>NB</i>	-	-	-	-	-	-	-	-	-
	<i>GP</i>	0.56	0.56	0.56	0.51	0.51	0.51	0.45	0.45	0.45
	<i>GNB</i>	0.44	0.44	0.44	0.49	0.49	0.49	0.55	0.55	0.55

P, Poisson; NB, negative binomial; GP, generalized Poisson; GNB, generalized negative binomial

감소할수록 포아송모형은 선택이 안되고 음이항모형은 선택되는 빈도가 감소하고 일반화 포아송모형과 일반화 음이항모형이 선택되는

빈도가 증가한다. 하지만 산포모수가 5 이상이 되면  $\theta_{30}$ 와  $\theta_{40}$ 에 상관 없이 포아송모형과 음이항모형은 선택되지 않고 일반화 포아송모형과 일반화 음이항모형이 약 50%씩 선택되는 경향이 있다.

모형 성능 비교 결과, 두 자료의 공통점은 산포모수가 증가하면 포아송모형과 음이항모형은 선택되지 않고 일반화 포아송모형과 일반화 음이항모형이 선택된다는 것이다. 이 때, 일반화 포아송모형과 일반화 음이항모형 한쪽에 치우치지 않고 약 50%씩 선택된다. 이는 포아송모형과 음이항모형이 잘 설명하지 못하는 부분을 일반화 포아송모형과 일반화 음이항모형이 설명을 가능하게 한다는 것을 의미한다. 이와 같은 결과가 나타난 이유는 식 (2.7)과 같이 일반화 포아송모형의 산포모수를 일반화 음이항모형의 산포모수로 표현할 수 있기 때문이다. 그리고 일반화 포아송자료에 대한 모형 성능 비교에서 산포모수가 음수인 경우 일반화 음이항모형이 선택될 비율이 0이고 일반화 포아송모형을 선택하는 빈도가 큰 이유는 식 (2.7)에서 일반화 음이항분포의 산포모수  $\beta$ 의 범위는 1보다 크거나 같아 일반화 포아송분포의 산포모수  $k$ 의 음수 범위를 포함하지 못하여 산포모수가 음수인 일반화 포아송자료를 일반화 음이항모형은 설명하지 못하고 일반화 포아송모형만이 설명 가능하기 때문이다.

## 제 4 장

### 실증연구

본 논문에서 제안한 모형을 오렌지주스자료와 자동차 채광창 자료에 적합시켜보았다. 품질관리 분야에서는 제조 과정의 오류를 찾아 이를 수정하고 제조품의 품질이 향상 되도록 하는 과정을 거친다. 이러한 제조 과정을 학습곡선에 적합하여 설명하고자 한다.

#### 4.1 오렌지주스 자료

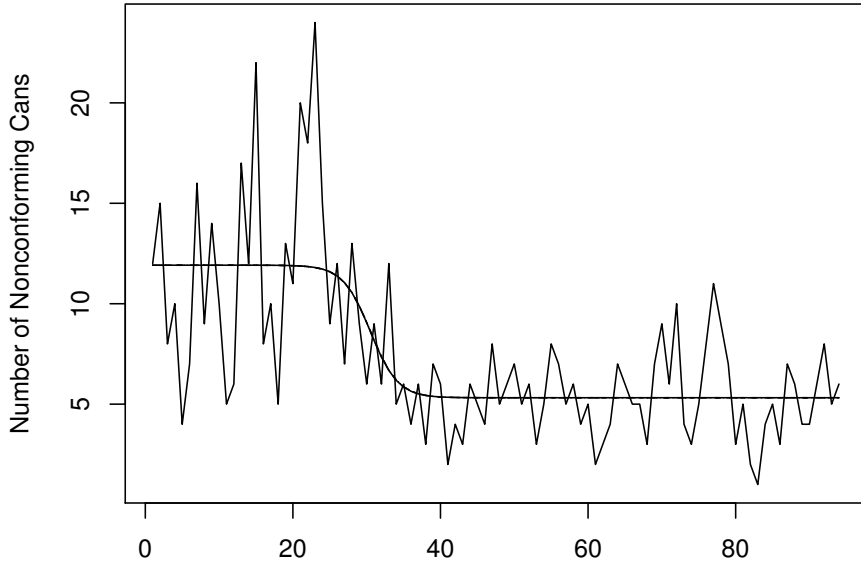


그림 9: 학습곡선 모형에 적합한 오렌지주스 자료(Montgomery, 1991)

표 10: 오렌지주스자료 추정 결과

Model	Parameter	Estimate	95% C.I	AIC	BIC	AICc
<i>P</i>	$\theta_1$	5.320	(4.714, 5.926)	470.6	480.8	471.1
	$\theta_2$	11.92	(10.54, 13.30)			
	$\theta_3$	30.49	(27.03, 33.96)			
	$\theta_4$	1.854	(0.329, 3.380)			
<i>NB</i>	$\theta_1$	5.320	(4.697, 5.942)	468.2	478.3	468.6
	$\theta_2$	11.92	(10.46, 13.37)			
	$\theta_3$	30.50	(26.92, 34.08)			
	$\theta_4$	1.860	(0.255, 3.464)			
<i>GP</i>	$\theta_1$	5.318	(4.651, 5.985)	467.6	480.3	468.3
	$\theta_2$	11.92	(10.23, 13.60)			
	$\theta_3$	30.53	(26.54, 34.51)			
	$\theta_4$	1.871	(0.016, 3.726)			
	$k$	0.019	(-0.001, 0.039)			
<i>GNB</i>	$\theta_1$	5.318	(4.650, 5.986)	467.6	480.3	468.3
	$\theta_2$	11.92	(10.23, 13.60)			
	$\theta_3$	30.53	(26.57, 34.48)			
	$\theta_4$	1.871	(0.030, 3.712)			
	$\beta$	2.417	(0.430, 4.404)			

C.I, confidence interval; P, Poisson; NB, negative binomial; GP, generalized Poisson; GNB, generalized negative binomial

그림 9는 오렌지주스 자료로 냉동 오렌지주스 농축액을 포장하고 있는 6온스 캔의 제조 과정에서 발생한 부적합 판정을 받은 캔의 수를 센 이산형 자료이다(Montgomery, 1991). 이 캔은 기계에 의해 생성되는데 캔의 적합과 부적합 판정 검사는 오렌지주스를 가득 채웠을 때 측면 또는 하단 주변으로 누수되는지 확인하여 새는 곳이 있으면 부적합으로 판단한다. 기계는 연속 운전 되고 3교대에 걸쳐 30분 간격으로 50개의 캔을 추출하는 과정을 30번 반복하여 표본을 얻었다. 30개의 표본을 얻은 후 제조 과정의 오류를 찾아 기계를 조절하여 향상시키고 앞에서 언급한 방식으로 64번 추가적으로 표본을 얻었다. 이 자료를 앞에서 구축한 학습곡선 모형에 적합시켰다.

추정결과 표 10을 보면 4가지 모형에서 모든 추정값  $\hat{\theta}$ 이 유사하게 나타났다. 모형선택기준을 살펴본 결과 4가지 모형 중 일반화 포아송 모형과 일반화 음이항모형의 모형선택기준이 다른 모형보다 좋았다. 이 때 추정된 산포모수는 일반화 포아송모형의 경우  $\hat{k} = 0.019$ 이고, 일반화 음이항모형의 경우  $\hat{\beta} = 2.417$ 이다. 여기서 일반화 포아송모형의 산포모수의 추정값  $\hat{k}$ 이 유의하지 않았는데 이는 근사적 신뢰구간을 이용하였기 때문이다. 모든 모형의  $\hat{\theta}_3$ 을 보면 제조 과정의 오류를 찾아 기계를 조절하여 향상시킨 시점인 30과 유사함을 확인할 수 있다.

## 4.2 자동차 채광창 자료

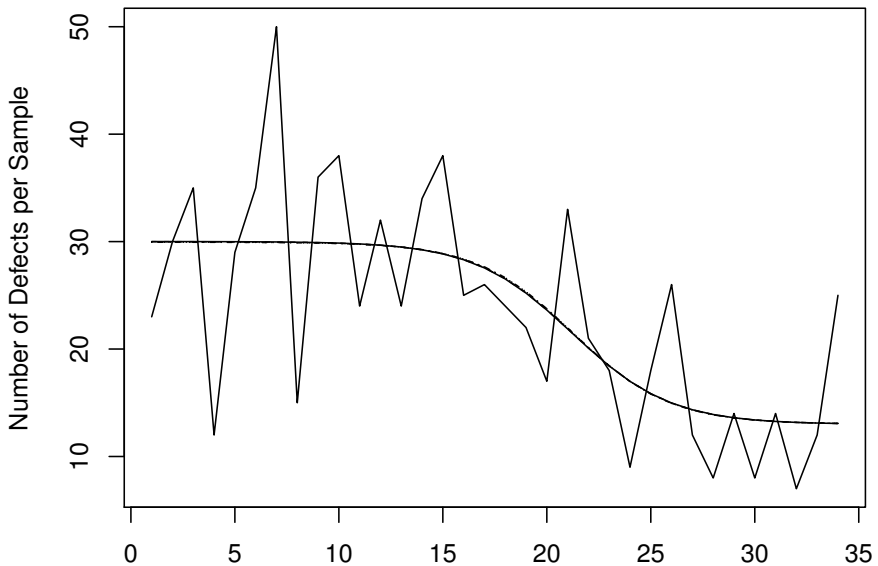


그림 10: 학습곡선 모형에 적합한 채광창 자료(DeVor *et al.*, 2007)

그림 10은 자동차 채광창 자료로 자동차에 채광창을 설치하는 과

표 11: 채광창자료 추정 결과

Model	Parameter	Estimate	95% C.I	AIC	BIC	AICc
<i>P</i>	$\theta_1$	13.00	(9.496, 16.50)	258.5	264.6	259.9
	$\theta_2$	30.01	(26.91, 33.10)			
	$\theta_3$	21.21	(18.36, 24.05)			
	$\theta_4$	2.365	(0.154, 4.576)			
<i>NB</i>	$\theta_1$	13.00	(9.757, 16.24)	249.9	256.0	251.3
	$\theta_2$	29.99	(26.26, 33.73)			
	$\theta_3$	21.22	(17.95, 24.50)			
	$\theta_4$	2.352	(0.160, 4.545)			
<i>GP</i>	$\theta_1$	13.02	(8.850, 17.18)	241.9	249.5	244.0
	$\theta_2$	29.94	(24.71, 35.18)			
	$\theta_3$	21.27	(17.14, 25.40)			
	$\theta_4$	2.303	(-0.303, 4.909)			
	$k$	0.025	(0.007, 0.043)			
<i>GNB</i>	$\theta_1$	13.01	(8.735, 17.29)	241.9	249.5	249.5
	$\theta_2$	29.94	(24.59, 35.30)			
	$\theta_3$	21.27	(16.91, 25.62)			
	$\theta_4$	2.304	(-0.581, 5.188)			
	$\beta$	3.006	(1.241, 4.771)			

C.I, confidence interval; P, Poisson; NB, negative binomial; GP, generalized Poisson; GNB, generalized negative binomial

정에서 생기는 결함이 몇 부분이 생겼는지에 대한 자료이다(DeVor *et al.*, 2007). 많은 고객들의 불만 때문에 조립 공장에서 관련 공정에서 이러한 문제가 발생하는 몇 가지 데이터를 수집하였다. 임의로 지정된 교대가 확인한 데이터가 한번 관찰된 표본이 된다. 9일 동안 27회의 관찰을 하여 자료를 얻었다. 이때 어느 부분에서 결함이 생기는지 관찰 후 결함이 생기는 이유를 찾아 채광창 설치 과정의 설계를 다시 조정하였다. 그 후에 위와 같은 조건으로 7회 표본을 얻었다. 이 자료를 구축한 학습곡선 모형에 적합시켰다.

표 11 채광창 자료의 추정 결과를 살펴보면 모든 추정값  $\hat{\Theta}$ 이 유사하게 나타남을 알 수 있다. 모형선택기준을 보면 4가지 모형 중 일반

화 포아송모형과 일반화 음이항모형이 다른 모형 보다 좋았다. 이 때 추정된 산포모수는 일반화 포아송모형의 경우  $\hat{k} = 0.025$ 이고, 일반화 음이항모형의 경우  $\hat{\beta} = 3.006$ 이다. 표 11에서 일반화 포아송모형과 일반화 음이항모형의  $\hat{\theta}_4$ 가 유의하지 않았는데 이는 근사적 신뢰구간을 이용하였기 때문이다. 그리고 선택된 모형의  $\hat{\theta}_3$ 을 보면 제조 과정의 오류를 찾아 기계를 조절하여 향상 시킨 시점인 27보다 작게 추정된 것을 확인할 수 있다.

## 제 5 장

### 결론 및 토의

본 논문에서는 포아송분포와 음이항분포를 사용하여 학습곡선을 모형화 한 최민지(2014)를 확장하여 일반화 포아송분포와 일반화 음이항분포를 사용하여 학습곡선을 모형화하였다. 특정한 사건의 발생 횟수와 내재된 변동성이 변화하는 자료를 4개의 모수를 이용하여 학습곡선 모형을 구축하였다. 학습곡선 모형의 모수를 추정하기 위해 최대우도추정방법을 이용하였다.

모의실험은 학습곡선의 위치모수와 산포모수를 변화시킨 다양한 시나리오를 가지고 실행하였다. 시나리오와 분포의 산포모수를 변화시켜 각각의 일반화 포아송자료와 일반화 음이항자료를 생성하였고, 이와 같이 생성된 자료된 자료를 일반화 포아송모형과 일반화 음이항모형에 적용하였을 때 각 특성을 알아보기 위해 추정값의 평균과 분산, 평균제곱오차와 각 모형의 신뢰구간을 구하여 확인하였다. 그 결과 모든 추정값의 평균이 참값과 매우 유사하게 나타났고 학습곡선의 위치모수가 감소할수록, 학습곡선의 산포모수가 증가할수록, 자료의 산포모수가 증가할수록 분산과 평균제곱오차가 증가하고 분산과 표준제곱오차의 차이 또한 증가하였다. 그리고 추정값을 이용하여 평균적합모형과 신뢰구간을 구하였을 때 평균적합모형은 참 모형과 매우 유사하게 나타났고 근사신뢰구간과 경험적 신뢰구간은 초기시점에서 차이가 있지만 시점이 지날수록 매우 유사했다. 두가지 신뢰구간은 초

기시점에서 학습곡선의 위치모수가 감소할수록 넓고 완만해지는 시점 이후 학습곡선의 위치모수가 증가할수록 넓어지는 특징을 보였다. 전 시점에서 학습곡선의 산포모수가 증가할수록 넓어지고 자료의 산포모수 값이 증가할수록 넓어졌다.

4가지 모형 성능을 비교 결과 자료의 산포모수가 커질수록 포아송 모형이나 음이항모형은 선택되지 않고 일반화 포아송모형과 일반화 음이항모형이 선택되었다. 이때 산포모수가 양의 값을 갖으면 일반화 포아송모형과 일반화 음이항모형이 약 50%씩 선택되었고 산포모수가 음의 값을 갖으면 일반화 음이항모형은 선택되지 않고 일반화 포아송모형이 선택되었다. 이와 같은 결과는 산포모수가 커지면 포아송모형과 음이항모형은 그 자료를 설명하지 못하고 특히 자료의 산포모수가 음수인 경우는 일반화 포아송모형이 자료를 설명 가능하다는 것을 의미한다. 이는 제 2.1.3장에서 언급한 바와 같이 일반화 포아송분포의 산포모수가 일반화 음이항분포의 산포모수에 대한 식으로 표현이 되므로 자료의 산포모수가 양수일 때는 일반화 음이항모형과 일반화 포아송모형에 약 50%씩 선택이 되고, 이 때 식은 일반화 음이항분포의 산포모수의 범위가 1보다 크거나 같아 일반화 포아송분포의 음수 범위는 포함하지 못하여 자료의 산포모수가 음수일 때는 일반화 음이항모형이 자료를 설명하지 못하고 일반화 포아송모형이 자료를 설명할 수 있는 것이다.

본 논문에서 제안한 모형을 오렌지주스 자료와 채광창 자료에 적용해 보았다. 두 자료 모두 제조 공정에서 일어나는 불량 횟수를 예측하기 위해 모형을 적용한 것으로 두 자료 모두 4가지 모형이 매우 유사하게 추정되었고 오렌지주스 자료는 학습곡선모형의 변곡점을 추정한 값이 제조 과정의 오류를 찾아 기계를 조절하여 향상 시킨 시점인 30

과 유사함을 확인할 수 있었고, 채광창자료는 학습곡선모형의 변곡점의 추정값이 제조 과정의 오류를 찾아 기계를 향상시킨 시점보다 미리 나타난 것을 확인할 수 있었다.

향후 학습곡선모형의 모수를 추정하기 위해 관측된 자료와 사전 분포를 이용하여 사후분포를 추정하는 베이지안추정방법을 이용하여 최대우도추정방법으로 추정한 값과 비교해 보고자한다. 그리고 학습곡선을 모형화 할때 본 논문은 대칭 형태인 로지스틱 분포의 누적분포 함수를 사용하였는데 이와 같은 대칭적인 분포를 사용하지 않고 카이제곱분포와 같은 시점이 0부터 시작하고 비대칭적인 분포의 누적분포 함수를 이용하여 학습곡선을 모형화하고자 한다.

## 참고 문헌

- [1] 김인구, 김형준, 문덕중 (2003). 생산운영관리, 문영사.
- [2] 김지선 (2014). 디지털 구강스캐너 두 종의 학습곡선 비교, 이화여자대학교 임상치의학대학원, 석사학위논문.
- [3] 백우중 (2009). 학습곡선에 의한 연료전지의 비용효과 분석, 동신대학교 대학원, 석사학위논문.
- [4] 안명관 (2005). 항공기 시스템시장에서 학습곡선에 의한 공급사슬 균형의 모색, 경상대학교 대학원, 박사학위논문.
- [5] 이슬지 (2011). 학습곡선의 모형화 : 이항반응자료를 중심으로, 성신여자대학교 일반대학원, 석사학위논문.
- [6] 이슬지, 박만식 (2012). 이항 반응 자료에 대한 학습곡선의 모형화, 한국통계학회, 제19권, 433-450.
- [7] 정광태, 홍자인 (2008). 제품의 학습성을 평가하기 위한 학습곡선 모델의 적용, 대한인간공학회, 제27권, 59-65.
- [8] 정소기, 차경천 (2014). 통신 상품별 VOC 영향 요인과 학습곡선에 관한 연구, 한국통신학회, 제39권, 518-527.
- [9] 정종교 (2007). 항공기 날개구조물 조립공정의 생산성향상활동과 학습곡선에 관한 연구, 창원대학교 산업·정보대학원, 석사학위논문.
- [10] 정호석 (2009). 복막외 복강경하 근치적 전립선적출술 : 103예의 임상경험 및 학습곡선, 전남대학교 대학원, 석사학위논문.
- [11] 최민지 (2014). 이산형 자료를 기반으로 한 학습곡선의 모형화, 성신여자대학교 일반대학원, 석사학위논문.
- [12] 최민지, 박만식 (2014). 개수형 자료에 대한 학습곡선효과의 모형화, 한국통계학회, 제27권, 445-459.

- [13] 최준영, 송경빈 (2005). 학습곡선을 이용한 수요관리의 효과 추정, *대한전기학회*, 제53권, 209-213.
- [14] Akaike, H. (1974), A New Look at the Statistical Model Identification, *IEEE Transactions on Automatic Control* Vol. 19(6), 716-723.
- [15] Andréasson, H., Lorant, T., Pählman, L., Graf, W., and Mahteme, H. (2014). Cytoreductive Surgery plus Perioperative Intraperitoneal Chemotherapy in Pseudomyxoma Peritonei: Aspects of the Learning Curve, *European Journal of Surgical Oncology(EJSO)*, Vol. 40, 930-936.
- [16] Burnham, K. P., Anderson, D. R. (2004), "Multimodel Inference: Understanding AIC and BIC in Model Selection, *Sociological Methods and Research* Vol. 33, 261-304.
- [17] Consul, P. C. and Famoye, F. (2006). *Lagrangian Probability Distributions*, Birkhauser.
- [18] DeVor, R. E., Chang, T., and Sutherland, J. W. (1992). *Statistical Quality Design and Control: Contemporary Concepts and Methods*, Prentice Hall.
- [19] Famoye, F. (1993). Restricted Generalized Poisson Regression Model, *Communications in Statistics-Theory and Methods*, Vol. 22, 1335-1354.
- [20] 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.
- [21] Leigh G. S., Jeffrey M. F., Debra L. R., Matthew J. C., Valmadre, S., Gary S. P., and David E. C. (2009). A Detailed Analysis of the Learning Curve: Robotic Hysterectomy and Pelvic-aortic Lymphadenectomy for Endometrial Cancer, *Gynecologic Oncology*, Vol. 144, 162-167.

- [22] Louis E. Y. (1979). The Learning Curve: Historical Review and Comprehensive Survey, *University of Lowell*, Vol. 10, 302-328.
- [23] Montgomery, D. C. (1991). *Introduction to Statistical Quality Control*, New York, John Wiley&Sons.
- [24] Nemet, G. F. (2006). Beyond the Learning Curve: Factors Influencing Cost Reductions in Photovoltaics, *Energy Policy*, Vol. 34, 3218-3232.
- [25] Ngwenyama, O., Guergachi, A., and McLaren, T. (2007). Using the Learning Curve to Maximize IT Productivity: A Decision Analysis Model for Timing Software Upgrades, *Production Economics*, Vol. 105, 524-535.
- [26] Ospina, O. D. A., Medina, Á. M. R., Marulanda, M. C., and Buitrago, L. M. G. (2014). Cumulative Sum Learning Curves(CUSUM) in Basic Anaesthesia Procedures, *Colombian Journal of Anesthesiology*, Vol. 42, 142-153.
- [27] Park, S. S., Kim, M. C., Park, M. S., Hyung W. J. (2012). Rapid Adaptation of Robotic Gastrectomy for Gastric Cancer by Experienced Laparoscopic Surgeons, *Surgical Endoscopy*, Vol. 36, 60-67.
- [28] Schoots, K., Ferioli, F., and Kramer, G. J. (2008). Learning Curves for Hydrogen Production Technology: An Assessment of Observed Cost Reductions, *International Journal of Hydrogen Energy*, Vol. 33, 2630-2645.
- [29] Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics* Vol.6(2), 461-464.
- [30] Tamas J. V., Smith, W., Ghosh, J., and Mohamed R. A. (2006). Robotic Cholecystectomy: Learning Curve, Advantages, and Limitations, *Journal of Surgical Research*, Vol. 136, 172-178.
- [31] Teng, J. -T., Lou, K. -R., and Wang, L. (2014). Optimal Trade Credit and Lot Size Policies in Economic Production Quantity Models with

Learning Curve Production Costs, *International Journal of Production Economics*, Vol. 155, 318-323.

- [32] Wolff, T., Schumacher M., Dell-Kuster, S., Rosenthal R., Dickemann, M., Steiger, J., Bachmann, A., and Bachmann, L. (2014). Surgical Complications in Kidney Transplantation: No Evidence for a Learning Curve, *Journal of Surgical Education*, Vol. 71, 748-755.
- [33] Wang, F.-K. and Lee, W. (2001). Learning Curve Analysis in Total Productive Maintenance, *Omega*, Vol. 29, 491-499.
- [34] Wright, T. P. (1936). Factors Affecting the Cost of Airplanes, *Journal of Aeronautical Sciences*, Vol. 3, 122-128.

# Abstract

## Modelling of The Learning Curves on the Generalized Discrete Distribution

Mina Kang

Department of Statistics

The Graduate School

Sungshin Women's University

Learning curve was studied at first by Wright(1936), which says that, as a cumulative production of aircraft increases, direct labor hours per aircraft decrease regularly in aviation industry. Unit production time generally decreases when performing a certain operation repeatedly. It results from the improvement of the workers' competence, coordination of work procedures and work capacities of the machine. This is called a learning phenomenon. Rate of change of the learning phenomenon is great for a period of time, but it tends to mitigate at a certain point. This paper models the learning-curve effects using the generalized Poisson and the generalized negative-binomial distribution to fit the count data. These models take into account both time-varying mean of the learning-curve and its excessive variability. This paper carries out simulations to examine the performances of the proposed mod-

els and compares with common models. The real application consists of the orangejuice data and the moonroof data.

**Keywords :** Generalized Poisson distribution, Generalized negative-binomial distribution, Learning curve, Logistic distribution function, Newton-Raphson method