베이지안 음이항 분기과정을 이용한 한국 메르스 발생 연구[†]

박유하¹ · 최일수²

¹²전남대학교 통계학과

접수 2016년 12월 25일, 수정 2017년 1월 16일, 게재확정 2017년 1월 19일

요 약

전염병 확산에 대한 확률과정모형으로 활용되는 분기과정은 실제 데이터를 통해 모수를 추정할 수 있다는 장점이 있다. 음이항 분포를 분기과정의 생산 분포 모형으로 적용할 수 있는데 음이항 분포를 적용하기 위해서는 평균과 산포 모수를 추정하여야한다. 기존의 생물학 연구와 역학 연구 분야에서는 이를 최대우도법을 이용하여 추정하고 있다. 그러나 대부분의 역학 자료의 특성상 분기과정에서 이용되는 음이항 분포는 소표본이어서 최대우도 추정량의 정도를 충족시킬 수 없다. 본 논문에서는 소표본 자료에서 좋은 통계량의 성질을 만족한다고 알려져 있는 베이지안을 이용하여 모수를 추정하는 방법을 제안한다. 2015년 국내 메르스 사례에 베이지안 방법을 적용하여 모수를 추정하고 사후 분포를 적합하였다. 그 결과 어떠한 사전 분포를 가정하더라도 안정적으로 모수를 추정하는 것을 알 수 있었다. 추정된 산포 모수를 이용하여 분기과정에서의 전염병 소멸 확률을 유도하였다.

주요용어: 베이지안 추론, 분기과정, 산포 모수, 수학적 모델링, 음이항 분포, 전염병 확산.

1. 연구의 배경

2015년 한국에서 발생한 메르스 (middle east respiratory syndrome coronavirus; MERS-CoV)는 중동 지역에서 집중적으로 발생한 바이러스로 사스 (severe acute respiratory syndrome coronavirus; SARS-CoV)와 유사하게 주로 고열, 기침, 호흡곤란 등 호흡기 증상을 일으키는 질병이다 (Hwang과 Oh, 2016). 메르스의 전파경로는 아직 정확히 밝혀지지 않았지만, 메르스는 동물에 의한 사람으로의 직·간접적인 전파 또는 사람 간 접촉에 의한 전파로 전염된다. 사우디아라비아에서 발생한 메르스는 낙타가 매개체인 것으로 추정되며 2015년 한국에서 발생한 메르스는 주로 병원 내에서 메르스 감염 환자와의 접촉에 의해 전파되었다. 공기 중으로 감염되지 않고 기침이나 콧물 등 호흡기 분비물에 의해 바이러스가 전파되는 메르스는 특정 감염자로부터 대다수의 2차 감염자 (offspring; secondary cases; 줄여서 2차 감염자)가 발생하는 주요전파사건 (Superspreading Event, SSE)이다. 메르스는 단기간에 급속도로 유행하였고 한국은 전국적으로 메르스의 공포에 휩싸였다. 총 186명이 메르스를 확진 받았으며 이중 다섯 명이 슈퍼전파자(superspreader)이었다 (Korea Centers for Disease Control and Prevention, 2015).

이러한 질병 확산과 관련한 연구는 생물학, 역학, 질병관리 등 다양한 분야에서 관심을 가지고 연구되어지고 있다. Lloyd-Smith 등, (2005)은 역학 모형으로 편미분을 이용한 수학적 모형이 아닌 확률과정을 이용한 통계적 모형을 제시하였다. 주요전파사건으로 분류되는 여러 질병에 대한 통계적 역

[†] 본 연구는 농촌진흥청 공동연구사업 (과제번호: PJ01156302)의 지원에 의해 이루어진 것임.

¹ (61186) 광주광역시 북구 용봉로 77, 전남대학교 통계학과, 석사과정.

² 교신저자: (61186) 광주광역시 북구 용봉로 77, 전남대학교 통계학과, 교수. E-mail: ichoi@jnu.ac.kr

학 모형으로 single-type 분기과정 (branching process)을 적합하였고, 질병의 소멸 확률 (extinction probability)을 계산하였다. 또한 Alexander (2010)는 돌연변이가 고려된 병원균의 확산 모형을 multi-type 분기과정을 이용하여 적합하였고, 이를 사용하여 소멸 확률을 계산하였다. 이 모형에서 분기과정 생산 분포의 모수는 지정된 값을 지닌다.

기초감염재생산수 (basic reproductive number) R_0 는 질병에 걸릴 가능성이 있는 집단에서 감염자에 의한 2차 감염자 수의 평균을 의미하는데 이를 보고 질병이 확산되는 정도에 대하여 대략적으로 알 수 있다. 따라서 R_0 를 통해서 주요전파사건 여부를 판단할 수 있다. 기초감염재생산수를 추정하기 위한 모형은 전염병을 시간종속적인 평균의 개념을 갖는 동역학으로 처리한 SIR (susceptible-infectious-recovery) 모형과 확률적 추계모형과 분기과정을 이용한 모형으로 분류될 수 있다 (Choi와 Rhee, 2015; Ryu와 Choi, 2015).

질병 발생이 끝날 확률을 말하는 소멸 확률은 R_0 와 k 로부터 계산된다. Lloyd-Smith 등, (2005)은 분기과정을 사용하여 2차 감염자의 수 Z 가 k 명일 때의 확률 분포를 의미하는 생산 분포 $P(Z = k)$ 에 대하여 포아송 분포, 기하 분포, 음이항 분포를 가정하고, 주요전파사건에 대하여 각 생산 분포에 따른 소멸 확률을 구하였다. 소멸 확률은 질병에 대한 위험 (risk)을 말해주는 지표로 소멸 확률이 높을수록 질병이 불규칙적으로 발생하고 일시적인 유행을 갖는다. 반면 낮은 소멸 확률은 질병이 만성적으로 생존함을 의미한다. 여기서 환자의 감염력 (individual's infectious history)을 설명하는 분포로 음이항 분포는 기하 분포보다 적용 범위가 넓다고 알려져 있다 (Burnham과 Anderson, 2002, Lloyd-Smith 등, 2005). 그러므로 음이항 분포의 모수에 대한 추정량의 정도 (precision)를 충족시켜 역학 자료의 모형 적합성을 높이는 것이 중요하다고 할 수 있다.

음이항 분포의 모수는 평균 R_0 와 산포 모수 k 이다. Lloyd-Smith 등 (2005)은 음이항 분포의 산포 모수 k 를 추정하기 위해서 최대우도법 (maximum-likelihood method)을 이용하였다. 최대우도 추정치는 소표본 오차 (small-sample bias)가 적고 효율성의 성질이 더 좋다고 알려졌다 (Anscombe, 1950; Pieters 등, 1977). 산포 모수 k 를 추정하는 연구는 붓스트랩 (bootstrap) 등의 방법으로 활발히 진행되고 있지만 수치해석적으로 구하기 때문에 소표본 추정의 편의는 모수의 참값을 과대 추정하는 단점이 있다 (Saha과 Paul, 2005; Pieters, 1977; Piegorsch, 1990).

베이지안 방법 (Bayesian method)은 현재 주어진 관측치에 의존하여 모수를 추정하므로 기존 방법에 비해 통계적으로 더 좋은 특성을 갖는 모수를 추정할 수 있다. 또한 베이지안 방법은 데이터가 적을 때 특히 그 유용성이 좋다고 알려져 있다 (Hoff, 2009). 본 논문에서 베이지안 방법을 사용하여 음이항 분포를 생산 분포로 하는 분기과정의 모수를 추정하는 방법을 제시한다. 또한 2015년 한국에서 발생한 메르스 자료에 이를 적합 시켜 소멸 확률을 구하였다. 2015년 한국의 메르스를 비롯하여 대부분의 역학 자료는 소표본이므로 분기과정을 이용한 모형을 적합할 시에 기존의 방법으로 적합 시킨 모형보다 베이지안 방법으로 적합 시킨 모형이 더 타당하다고 할 수 있다.

2. 연구 자료

본 연구에서는 2015년 한국 메르스의 분기과정을 파악하기 위해서 보건복지부 질병관리본부와 WHO에서 조사되어 공개된 자료를 활용하였다. 2015년 5월 20일 확진된 한국의 첫 메르스 환자는 증상 발현 후 약 10일 동안 여러 병·의원을 다니며 가족과 다른 환자, 의료진과 접촉하여 다수의 2차 환자를 유발하였다. 한 달이 채 되지 않는 기간 동안 한국에서 메르스 확진을 받은 전체 환자의 수는 186명이다.

본 연구에서는 2015년 5월 초부터 6월 말까지의 두 달 동안 감염된 총 186명의 환자를 대상으로 메르스 감염 경로를 추적하였다. 감염 경로가 확실하지 않은 17명의 환자를 제외한 169명의 환자에 대한 자

료로부터 2015년 5월 20일 발생한 한국의 첫 환자로부터의 감염 경로를 파악하여 분기과정을 구성하면 Figure 2.1과 같다.

Figure 2.1 Branching process of MERS in Korea (2015)

3. 베이지안 음이항 분기과정 모형

3.1. 베이지안 음이항 분기과정 모형

분기과정은 Galton-Watson 분기과정이라고도 하며 1873년에 Francis Galton에 의해서 처음으로 제안된 것이다 (Kendall, 1966). 본 논문에서는 single-type discrete time 분기과정을 2015년 한국 메르스 자료에 적용하였다. single-type discrete time 분기과정은 같은 세대 (generation)에서 질병을 전파할 가능성이 있는 모든 환자는 다음 세대에서 동일한 성질의 질병을 갖는 2차 감염자를 생산하고 이때, 2차 감염자의 수 Z ($Z = 0, 1, 2, \dots, n$)는 서로 독립이고 동일한 생산 분포를 따른다고 가정한다. 생산 분포는 2차 감염자의 수 Z 가 k 명일 때의 확률 분포를 의미하며, $P(Z = k) = p_k$ 로 주어진다. i 번째 개체의 2차 감염자의 수를 Z_i 이라고 하고, n 번째 세대에서 감염자 수의 총합을 X_n 라고 하면, X_n 은 $1, 2, \dots, Z_{n-1}$ 감염자에게서 전염된 2차 감염자 수의 총합이다 (Kimmel, 2002; Haccou, 2005; Pinsky, 2011). 즉,

$$X_n = \sum_{i=1}^{X_{n-1}} Z_i. \quad (3.1)$$

감염재생산수 ν 는 특정한 환자에 의해서 감염된 2차 감염자의 감염 기간을 의미하며 환자의 감염력에서의 모든 변동을 나타낸다 (Lloyd-Smith 등, 2005). 또한 감염재생산수 ν 의 값은 모평균이 기초감염재생산수 R_0 인 연속확률분포로부터 추정할 수 있다. 주요전파사건 (superspreading events)은 감염재생산수 ν 의 분포가 오른쪽 꼬리 분포의 형태로 나타난다는 특징을 갖는다 (Dye 등, 2003; Anderson, 2004; Lloyd-Smith 등, 2005).

본 논문에서는 생산 분포를 음이항 분포로 가정한다. 우선, 전염병의 확률적 효과가 포아송 과정을 따른다고 가정하면 2차 감염자의 수 Z 는 평균이 감염재생산수 ν 인 포아송 혼합분포 (poisson mixture distribution) $Z \sim \text{Poisson}(\nu)$ 를 따른다고 할 수 있다 (Lloyd-Smith 등, 2005).

여기서, 감염재생산수 ν 의 분포를 감염재생산수 ν 가 평균이 R_0 이고 산포 모수 (dispersion parameter)가 k 인 감마 분포를 따를 때 즉, $\nu \sim \text{Gamma}(k, R_0/k)$ 라고 한다면 생산 분포는 결국 음이항 분포 ($Z \sim \text{NB}((1 + R_0/k)^{-1}, k)$)로 주어진다 (Patil, 1970; Pielou, 1977; Pinsky과 Karlin, 2011).

분기과정에서 생산 분포는 확률생성함수 (probability generating function, pgf)로 나타낸다. 그러므로 베이저안 음이항 분기과정모형은 다음과 같은 음이항 분포를 생산분포로 갖는 분기과정으로 표현된다. 즉,

$$g(s) = E(s^{Z_n}) = \left(\frac{q}{1 - ps} \right)^n, |s| \leq 1. \quad (3.2)$$

이때, $p = (1 + R_0/k)^{-1}$ 이며, $g_0(s) = s$, $g_1(s) = g(s)$, $\dots$, $g_{n+1}(s) = g(g_n(s))$, $n = 1, 2, 3, \dots$ 을 만족한다.

기초감염재생산수 R_0 는 Z 의 평균이며 $g'(1)$ 로 나타낼 수 있기 때문에 모수 추정에 있어 기초감염재생산수 R_0 는 표본의 평균으로 간단히 추정할 수 있다. 그러나 산포 모수 k 의 경우 간단히 추정하기 어려우므로 본 연구에서는 베이저안 추론을 이용하였다.

베이즈 정리 (Bayes' rule)에 의해서 실제 관측된 관측치 $\mathbf{Z} = (z_1, z_2, \dots, z_n)'$ 는 평균이 ν 인 포아송 분포를 따르고 ν 는 평균이 R_0 이고 산포 모수가 k 인 감마 분포를 따른다고 하자.

$$Z|\nu \sim \text{Poisson}(\nu), \quad (3.3)$$

$$\nu \sim \text{Gamma}(k, R_0/k). \quad (3.4)$$

이때 사후분포는

$$f(k|Z) \propto f(Z|\nu) f(\nu|k, R_0/k) f(k, R_0/k) \quad (3.5)$$

$$\propto \prod_{i=1}^n \frac{\nu^{z_i} e^{-\nu}}{z_i!} \frac{R_0/k}{\Gamma(k)} (R_0\nu/k)^{k-1} e^{-R_0\nu/k} \cdot f(k, R_0/k). \quad (3.6)$$

그러므로 $p = (1 + R_0/k)^{-1}$ 일 때,

$$f(Z|\nu) f(\nu|k, R_0/k) \propto \prod_{i=1}^n \binom{k + z_i - 1}{k} (1 - p)^k (p)^{z_i} \quad (3.7)$$

로 나타낼 수 있고 최종적으로 식 (3.5)는

$$f(k|Z) \propto \prod_{i=1}^n \binom{k + z_i - 1}{k} (1 - p)^k f(k, R_0/k) \quad (3.8)$$

가 된다.

위에서 알 수 있는 것처럼 모수에 대한 베이저안 추론은 모수의 사후분포에 근거한다. 식 (3.5)에서 모수의 사후분포는 자료가 주어졌을 때 모수의 사전 분포에 의존한다는 것을 알 수 있다. 베이저안 방법에서 사전 분포는 자료를 얻기 전에 모수에 대한 정보를 나타내기 때문에 적절한 사전분포의 선택이 중요하다. 사전 분포는 두 가지의 경우 즉, 과거의 경험이나 이해에 따른 주관적 사전 정보를 요약한 주

관적 사전 분포 (subjective prior)와 사전 정보가 확실하지 않거나 없을 때 사용하는 객관적 사전 분포 (objective prior) 또는 무정보적 사전 분포 (non-informative prior)로 구분된다. 각각의 경우에서 연구자의 선택에 따라 매우 다양한 분포를 가정할 수 있다 (Hogg, 2009).

본 연구에서는 실제 자료에 근거하여, 모수 k 에 대해 최소한의 사전 정보만을 이용한 무정보적 사전 분포로 일양분포 (uniform prior)를 고려하였다. 일양 사전분포의 모수는 k 의 범위에 기초하여 (0, 20)으로 가정하였다. 또한 Lloyd-Smith 등 (2005)의 연구 결과에 기초하여, 주관적 사전 분포로 지수 분포 (exponential prior)와 로그정규분포 (log-normal prior)도 고려하였다. 2015년 한국 메르스 자료의 기본감염재생성수 R_0 는 1.98이므로 지수 사전분포의 모수를 2로 가정하였으며 로그-척도모수는 0이고 형상모수는 1인 로그정규 사전분포를 가정하였다.

3.2. 시뮬레이션과 모형 진단

사후밀도함수를 계산하는 과정에서 필요한 복잡한 형태의 함수 적분을 하기 위해서 베이지안 추론에서 대표적으로 사용하는 시뮬레이션 방법인 마코브 체인 몬테칼로 (Markov Chain Monte Carlo; MCMC) 알고리즘을 이용할 수 있다. 본 연구에서는 MCMC 기법의 일종인깁스 표집기 (Gibbs sampler)를 사용하였다. 깁스 표집기를 위한 v 의 사후분포는 앞 절에서 설명한 것처럼 식 (3.8)과 같고 이에 대한 모수 k 의 분포는 일양 사전분포, 지수 사전분포, 로그정규 사전분포를 가정하였다. 각 사전분포 별로 깁스 표집기에서 10,000의 반복 샘플링을 하였으며 500개를 번인 (burn-in)하였다.

본 논문에서는 난수를 생성하여 모수를 추정한 후, 수렴이 잘 되었는지를 판단하기 위하여 Gelman-Rubin 통계량을 사용하였다. 체인 (Chain)은 10개로 설정하였고 각각의 체인에 대한 샘플링을 10,000번 하였다. Gelman-Rubin 통계량은 1과 가까운 값을 가질수록 잘 수렴한다 (Gelman 과 Rubin, 1992; Monahan, 2001; Jo와 Lee, 2016).

3.3. 소멸 확률

본 연구에서는 2015년 한국 메르스에 대해서 질병의 위험 정도를 분석하기 위해 소멸 확률을 구하였다. X_n 의 확률생성함수가 $g_n(s)$ 라면, n 번째 세대에서 질병의 소멸 확률 (extinction probability) $P(X_n = 0)$ 은 $g_n(0)$ 으로 나타낼 수 있다. 여기서 이론적으로 $g'(1) < 1$ 이면, 질병은 소멸된다. 만약 $g'(1) = 1$ 이면, 질병은 소멸되지만 소멸 확률이 정해지지 않게 된다. 또한 만약 $g'(1) > 1$ 이면, 질병은 확산된다. 따라서 $g'(1) = R_0 \leq 1$ 일 때, 생산 분포의 확률생성함수를 $g(s)$ 라고 하면 소멸 확률을 $g(s) = s$ 로 정의한다. 이러한 소멸 확률은 수학적으로 구하기 어렵기 때문에 일반적으로 수치해석적인 방법을 사용하여 유도한다.

4. 자료 분석 결과

2015년 한국 메르스 자료에 대하여 산포 모수 k 를 베이지안 추론을 이용하여 추정된 결과는 Table 4.1이다. Figure 4.1은 시뮬레이션을 통해 생성된 표본들의 수렴성을 나타내는 표본의 경로그림이다. 고려된 사전분포들이 모두 수렴함을 알 수 있다. 그러나 로그 정규분포일 때의 진폭이 다른 두 사전분포에 비해 넓어 수렴에는 불리한 결과를 갖고 있음을 알 수 있다. $k \sim \exp(2)$ 일 경우, k 의 중앙값을 1.409, k 의 표준편차를 0.093으로 추정하여 가장 나은 성능을 보였다. 반면 $k \sim \lognormal(0, 1)$ 일 경우, k 의 중앙값을 1.414로 k 의 표준편차를 0.094로 추정하였다. 그러나 세 가지 분포를 가정한 경우 모두 추정된 k 의 중앙값은 비슷하였고 k 의 표준편차의 추정값 또한 별다른 차이가 없었다.

Table 4.1 Bayesian estimates based on prior distributions

prior	mean	sd	MC error	2.5%	median	97.5%
unif(0, 20)	1.430	0.093	0.0018	1.252	1.428	1.616
exp(2)	1.411	0.093	0.0021	1.237	1.409	1.596
lognormal(0, 1)	1.417	0.094	0.0018	1.242	1.414	1.608

Figure 4.1 MCMC trace plot according to uniform, exponential and log-normal prior

Figure 4.2는 체인 (chain)을 10개로 설정하여 구한 Gelman-Rubin 통계량을 나타낸 그림과 통계량이다. 푸른색은 체인 내 변이 (within-chain variance)의 추정치를, 초록색은 체인 간 변이 (between-chain variance)의 추정치를 나타내며 붉은색은 체인 내 변이의 추정치와 체인 간 변이의 추정치의 비 (ratio)를 나타낸다. 이러한 비를 Gelman-Rubin 통계량이라고 하며 자료 분석 결과 1로 수렴하는 것을 알 수 있다. 따라서 어떠한 사전 분포를 가정하더라도 안정적으로 모수가 수렴하는 것을 알 수 있다.

Figure 4.2 Gelman-Rubin plot

Figure 4.3은 일양분포, 지수분포와 로그 정규분포 사전확률분포에 따른 베이지안 전염병 소멸확률이 다. 확률생성함수 $y = g(s)$ 와 직선 $y = s$ 가 교차되는 것을 구하는 것으로 사전확률분포에 따른 베이지안 전염병 소멸 확률을 유도하면 Table 4.2와 같다. 추정된 산포 모수 k 의 중앙값을 가진 세 가지의 사전확률분포에서 뉴턴-랩슨 알고리즘을 사용하여 각각 5,000개의 표본을 추출하여 소멸 확률을 구하였

다. 소멸 확률은 일양 사전분포를 가정하는 경우 0.445로 가장 크게 나타났고 지수 사전분포를 가정하는 경우 0.434로 가장 낮게 나타났다.

Table 4.2 Bayesian extinction probability based on prior distributions

prior	estimate of k	extinction probability
unif(0, 20)	1.428	0.445
exp(2)	1.409	0.434
lognormal(0, 1)	1.414	0.439

Figure 4.3 Bayesian extinction probability according to uniform, exponential and log-normal prior

5. 결론

본 연구에서는 분기과정에서 생산 분포가 음이항 분포를 따를 때, 모수의 재구성 (reparameterization)을 통하여 베이지안 추론을 유도하였다. 산포 모수 k 의 사후분포를 유도하기 위해서 첫째, 일양 사전 분포를 따르는 경우, 둘째, 지수 사전 분포를 따르는 경우, 셋째, 로그정규 사전 분포를 따르는 경우를 가정하였다.

본 논문에서 사용된 베이지안 음이항 분기과정은 세 가지의 장점을 가진다. 첫째, 2015년 한국 메르스의 실제 자료를 이용하여 모형에 적용한 결과, 베이지안 방법을 이용하면 산포 모수 k 의 추정값을 안정적으로 할 수 있음을 알 수 있었다. 세 가지의 사전 분포에서 산포 모수 k 의 베이지안 추정값의 차이가 많지 않았으며, 추정값의 표준편차에서도 많은 차이가 나지 않았다. 또한 Gelman-Rubin 통계량을 사용하여 모형 진단을 한 결과 통계량값이 1에 가까워 모수의 수렴성이 확인되었다. 둘째, 베이지안 추론을 이용하면 기존 추정치의 문제점으로 지적되는 모수의 참값을 과대 추정하는 단점을 해결하고 수치해석적으로 구하는 최대우도법의 이론적 오류를 해결할 수 있을 것으로 판단된다. 셋째, 분기과정을 사용하면 질병의 소멸 확률을 계산할 수 있다는 장점이 있다. 본 연구에서는 베이지안 음이항 분기과정을 이용해서 2015년 한국 메르스 자료의 위험 정도를 판단하기 위하여 소멸 확률을 계산하였다. 세 가지의 사전 분포 가정에서 구한 소멸 확률은 43%에서 44%로, 2015년의 한국 메르스는 낮은 소멸 확률을 갖는 특이한 현상을 나타내었다. 따라서 한국 메르스는 숙주인 낙타가 없음에도 불구하고 전염력이 항상 존재하는 질병으로 발전할 수 있는 가능성을 가지고 있음을 알 수 있다.

그러나 본 논문에서는 기존의 최대우도법 추정치와 비교를 실시하지 않았다. 이후의 연구에서는 이와 관련된 연구를 진행할 예정이며 주요전과사건과 역학적 관련성이 있는 질병에 대한 다양한 자료를 수집하여 본 연구에서 사용한 베이지안 음이항 분기과정을 역학 연구에 일반적으로 활용할 수 있음을 연구할 예정이다.

References

- Alexander, H. K. (2010). *Modelling pathogen evolution with branching processes*, Master Thesis, Queen's University, Ontario.
- Anderson, R. M., Fraser, C., Ghani, A. C., Donnelly, C. A., Riley, S., Ferguson, N. M., Leung, G. M., Lam, T. H. and Hedley, A. J. (2004). Epidemiology, transmission dynamics and control of SARS: The 2002-2003 epidemic. *Philosophical Transactions of the Royal Society B*, **359**, 1091-1105.
- Anscombe, F. J. (1950). Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika*, **37**, 358-382.
- Choi, I. and Rhee, S-S. (2015). A transmission distribution estimation for real time Ebola virus disease epidemic model. *Journal of the Korean Data & Information Science Society*, **26**, 161-168.
- Dye, C. and Gay, N. (2003). Modeling the SARS epidemic. *Science*, **300**, 1884-1885.
- Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, **7**, 457-511.
- Haccou, P., Jagers, P. and Vatutin, V. A. (2005). *Branching processes: Variation, growth, and extinction of populations*, Cambridge University Press, New York.
- Hoff, P. D. (2009). *A first course in Bayesian statistical methods*, Springer Dordrecht Heidelberg London, New York.
- Hwang, S. and Oh, C. (2016). Estimation of the case fatality ratio of MERS epidemics using information on patients' severity condition. *Journal of the Korean Data & Information Science Society*, **27**, 599-607.
- Jo, S., and Lee, J. (2016). Density estimation of summer extreme temperature over South Korea using mixtures of conditional autoregressive species sampling model. *Journal of the Korean Data & Information Science Society*, **27**, 1155-1168.
- Kendall, D. G. (1966). Branching processes since 1873. *Journal of London Mathematics Society*, **41**, 386.
- Kimmel, M. and Axelrod, D. E. (2002). *Branching processes in biology*, Springer-Verlag Inc., New York.
- Korea Centers for Disease Control and Prevention. (2015). Middle east respiratory syndrome coronavirus outbreak in the Republic of Korea. *Osong Public Health and Research Perspectives*, **6**, 269-278.
- Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E., Getz, W. M. (2005). Superspreading and the impact of individual variation on disease emergence. *NATURE*, **438**, 355-359.
- Monahan, J. F. (2001). *Numerical methods of statistics*, Cambridge University Press, New York.
- Patil, G. P. (1970). *Random counts in models and structures*, Pennsylvania State University Press, University Park PA.
- Piegorsch, W. W. (1990). Maximum-likelihood estimation for the negative binomial dispersion parameter. *Biometrics*, **46**, 863-867.
- Pielou, E. C. (1977). *Mathematical Ecology*, Wiley, New York.
- Pieters, E. P., Gates, C. E., Matis, J. H. and Sterling, W. L. (1977). Small sample comparison of different estimators of negative binomial parameters. *Biometrics*, **33**, 718-723.
- Pinsky, M. A. and Karlin, S. (2011). *An Introduction to stochastic modeling*, 4th Ed., Academic Press, San Diego.
- Ryu, S. and Choi, B. (2015). Development of epidemic model using the stochastic method. *Journal of the Korean Data & Information Science Society*, **26**, 301-312.
- Saha, K. and Paul, S. (2005). Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics*, **61**, 179-185.

A study on MERS-CoV outbreak in Korea using Bayesian negative binomial branching processes[†]

Yuha Park¹ · Ilsu Choi²

¹²Department of Statistics, Chonnam University

Received 25 December 2016, revised 16 January 2017, accepted 19 January 2017

Abstract

Branching processes which is used for epidemic dispersion as stochastic process model have advantages to estimate parameters by real data. We have to estimate both mean and dispersion parameter in order to use the negative binomial distribution as an offspring distribution on branching processes. In existing studies on biology and epidemiology, it is estimated using maximum-likelihood methods. However, for most of epidemic data, it is hard to get the best precision of maximum-likelihood estimator. We suggest a Bayesian inference that have good properties of statistics for small-sample. After estimating dispersion parameter we modelled the posterior distribution for 2015 Korea MERS cases. As the result, we found that the estimated dispersion parameter is relatively stable no matter how we assume prior distribution. We also computed extinction probabilities on branching processes using estimated dispersion parameters.

Keywords: Bayesian inference, branching processes, dispersion parameter, epidemic emergence, mathematical modeling, negative binomial distribution.

[†] This study was carried out with the support of the Research Program of Rural Development Administration, Republic of Korea (Project No. PJ01156302).

¹ Master degree, Department of Statistics, Chonnam National University, Gwangju 61186, Korea.

² Corresponding author: Professor, Department of Statistics, Chonnam National University, Gwangju 61186, Korea. E-mail: ichoi@jnu.ac.kr