Journal of the Korean Data & Information Science Society
2015, **26**(1), 161–168

http://dx.doi.org/10.7465/jkdi.2015.26.1.161
한국데이터정보과학회지

실시간 에볼라 바이러스 전염병 모형의 전염확률분포추정[†]

최일수[1] · 이성석[2]

[1]전남대학교 통계학과 · [2]서원대학교 경영학과

접수 2014년 12월 21일, 수정 2014년 12월 29일, 게재확정 2015년 1월 10일

요 약

전염병은 기초 감염 재생산 수가 시간에 따라 달라져서 상황을 관리하기 어렵기 때문에 확산을 통제하기 어려울 뿐 아니라 정확한 예측은 더욱 어렵다고 알려져 있다. 최근에 많은 모형들이 새롭게 제시되고 있으며 그에 따라 현저하게 다른 결과가 도출되고 있다. 연속된 시간에서 기초 감염 재생산 수는 일반적으로 확률과정이론이 적용되고 있다. 본 논문에서는 에볼라 바이러스 전염병 모형에서 전염확률분포의 추정 방법을 제시하였다. 이 방법은 대규모 전염병 발생에서 실시간 추정을 가능하게 함으로 적절한 질병관리를 용이하게 한다. 기니에서 발생한 에볼라 바이러스 자료를 제시한 방법으로 분석하였다.

주요용어: 기초 감염 재생산 수, 분기과정, 에볼라 바이러스, 전염확률분포, 최우추정법.

1. 머리말

서아프리카에서 2014년에 대규모로 발생한 에볼라 바이러스는 지금까지의 에볼라 발생규모 중에서 가장 큰 것으로 기록되고 있다. 2013년 12월에 기니 (Guinea)에서 최초로 시작된 에볼라 바이러스는 시에라리온 (Sierra Leone), 라이베리아 (Liberia), 나이지리아 (Nigeria)까지 확산되고 있다. 에볼라 바이러스와 같은 돌연변이로 발생한 질병의 신속한 파급이나 생물학 무기의 투하에 따른 대규모 환자의 발생 등과 같은 응급상황에 대비하기 위하여 통계적 분석방법의 끊임없는 개선이 요구되고 있다.

전염병은 기초 감염 재생산 수 (basic reproductive number; R_0)가 시간에 따라 달라지기 때문에 상황을 관리하기가 어려울 뿐만 아니라 확산을 통제하기도 곤란하게 되고 정확한 예측은 더욱 힘든 것으로 알려졌다 (Evans와 Mammadov, 2014). 이런 이유로 최근에 많은 모형들이 새롭게 제시되고 있으며 그에 따라 현저하게 다른 결과가 도출되고 있다 (Althaus, 2014; Chowell 등, 2004; Chowell와 Nishiura, 2014; Nishiura와 Chowell, 2014). 기초 감염 재생산 수를 추정하는 이러한 모형들은 전염병을 시간종속적인 평균의 개념을 갖는 동역학으로 처리하는 SIR (susceptible- infectious-recovery) 모형과 확률적 추계모형과 분기과정을 사용하여 전염병의 기초 감염 재생산 수를 산출해 내는 방법으로 분류될 수 있다.

확률적 분기과정을 사용하면, 우선 분기과정에서의 기초 감염 재생산 수가 상대적으로 간편하게 구해지고 전염병의 최종 감염자 수를 예측하는 데도 효과적일 뿐만 아니라 확률 SIR모형을 추정하는 데도 효과적인 방법인 것으로 알려져 있다 (Becker와 Britton, 1999; Ball과 Donnelly, 1995; Oh, 2014; Hwang과 Oh, 2014). 본 논문에서는 확률적 추계모형을 기본으로 하는 확률과정에서 기초 감염 재생산 수와 전염가능 기간을 동시에 추정하는 White와 Pagano (2008)의 추정 과정에서 사용되는 전염확률분포 (disease transmission distribution)를 효과적으로 추정하는 방법을 제시하였다.

[†] 이 논문은 2012년도 전남대학교 학술연구비 지원에 의하여 연구되었음.
[1] (550-757) 광주광역시 북구 용봉로 77, 전남대학교 통계학과, 교수.
[2] 교신저자: (362-742) 충북 청주시 서원구 무심서로 377-3, 서원대학교 경영학과, 교수.
E-mail: ssrhee@seowon.ac.kr

2. 통계적 방법

2.1. 전염병 확률모형

시간 t에서 새롭게 발생한 특정한 전염 질병 건수를 N_t라 할 때, 기간 T 동안의 질병 발생자료는 $N_t,\ t = 1, \cdots, T$이다. 환자가 전염병에 감염됐을 때 처음에는 외부적으로는 증상을 보이지 않고 감염 증상을 발전시키는 단계인 잠복기를 거치게 되고 잠복기가 지나고 나면 외부적으로 증상이 나타나는 전염병 발현 단계에 이르게 된다 (Na 등, 2010). 일반적으로 환자의 발생자료는 전염병 발현 단계에서 조사되어진다. 질병 발생단계 확률모형은 여러 가지 형태로 정의되고 있지만 본 논문에서는 White와 Pagano (2008)가 제시한 모형을 사용하였다. 전염병 발생 환자는 기댓값이 기초 감염 재생산 수 R_0를 갖는 포아송분포에 따르고 감염가능간격에 따른 전염 확률분포는 전염가능 최대기간 k를 갖는 다항분포에 따른다. 초기 감염자 수를 N_0라고 할 때 초기 감염자에 의해 전염된 환자수를 $X_{0\cdot}$라고 하면 $X_{0\cdot}$은 평균이 $R_0 N_0$인 포아송분포에 따른다. 또한 전염가능기간이 k이면 $X_{0\cdot}$은 이어지는 k일의 전염 환자수이고 이의 분포는 다항분포를 갖게 된다. 임의의 i번째 일의 환자수가 N_i일 때 X_{ij}를 N_i 환자에 의해 감염된 j번째 날의 환자 발생수라고 정의하면, $X_{i\cdot}$은 N_i환자에 의해 감염된 환자의 총수를 나타낸다. 만일 X_{ij}를 알고 있다면, 우도함수를 다음과 같이 구성할 수 있다.

$$\begin{aligned} L(R_0, \boldsymbol{P}|\boldsymbol{N}, \boldsymbol{X}) = & \left[\frac{e^{-N_0 R_0}(N_0 R_0)^{X_0}}{X_{0\cdot}!}\right]\left[\binom{X_{0\cdot}}{X_{01}\cdots X_{0,k}} p_1^{X_{01}}\cdots p_k^{X_{0k}}\right] \\ & \times \left[\frac{e^{-N_1 R_0}(N_1 R_0)^{X_1}}{X_{1\cdot}!}\right]\left[\binom{X_{1\cdot}}{X_{12}\cdots X_{1,1+k}} p_1^{X_{12}}\cdots p_k^{X_{1,1+k}}\right]\times\cdots \\ & \times \left[\frac{e^{-N_T R_0}(N_T R_0)^{X_T}}{X_{T\cdot}!}\right]\left[\binom{X_{T\cdot}}{X_{T,T+1}\cdots X_{T,T+k}} p_1^{X_{T,T+1}}\cdots p_k^{X_{T,T+k}}\right] \end{aligned}$$

그러나, 일반적으로 환자 발생수 N_i는 가용한 자료이지만 X_{ij}는 추출해 낼 수가 없다. 따라서 가용한 N_i를 사용하여 우도함수를 재구성하면 다음과 같은 포아송 우도함수를 얻는다.

$$L(R_0, \boldsymbol{p}) = \prod_{t=1}^{T} \frac{e^{-\mu_t}\mu_t^{N_t}}{N_t!} \tag{2.1}$$

여기서 $\mu_t = R_0 \sum_{j=1}^{\min(k,t)} N_{t-j} p_j$를 나타낸다.

2.2. 기초 감염 재생산 수

식 (2.1)의 값을 이용하여 $p_j,\ j = 1, \cdots, k$가 주어질 때, 기초 감염 재생산 수 R_0의 최우추정값 $\widehat{R}_0$은 로그우도함수로부터,

$$l = logL = -\sum_{t=1}^{T} R_0 \sum_{j=1}^{\min(k,j)} N_{t-j} p_j + (logR_0)\sum_{t=1}^{T} N_t + \log\left(\sum_{j=1}^{\min(k,t)} N_{t-j} p_j\right)\sum_{t=1}^{T} N_t - \log(N_t!)$$

$$\frac{\partial l}{\partial R_0} = -\sum_{t=1}^{T}\sum_{j=1}^{\min(k,t)} N_{t-j} p_j + \sum_{t=1}^{T} N_t / R_0$$

$\frac{\partial l}{\partial R_0} = 0$에서, 다음과 같다.

$$\widehat{R}_0 = \frac{\sum_{t=1}^{T} N_t}{\sum_{t=1}^{T} \sum_{j=1}^{\min(k,t)} N_{t-j} p_j} \tag{2.2}$$

또한 $\widehat{R}_0$의 분산은 $\frac{\partial^2 l}{\partial R_0}^2 = -\sum_{t=1}^{T} N_t / R_0^2$으로부터 다음과 같다.

$$Var(\widehat{R}_0) = \frac{\widehat{R}_0^2}{\sum_{t=1}^{T} N_t}$$

충분히 큰 전염가능기간 k를 고려하는 경우가 일반적이라면, 즉 이전에 발생한 모든 환자가 어느 시점에서 전염가능성을 갖고 있다고 가정하면, $\widehat{R}_0$은 다음과 같이 계산된다.

$$\widehat{R}_0 = \frac{\sum_{t=1}^{T} N_t}{\sum_{t=1}^{T} \sum_{j=1}^{t-1} N_{t-j} p_j} \tag{2.3}$$

3. 전염확률분포 추정

식 (2.1)의 우도함수를 최대화하는 R_0와 $p_j,\ j = 1, \cdots, k$를 동시에 구해야 한다. 이때 $p_j,\ j = 1, \cdots, k$를 전염확률분포라고 한다. 전염확률분포의 모수의 개수가 k이므로 k가 크면 추정해야 할 모수가 지나치게 많아지는 문제가 발생한다. 따라서 추정해야할 모수의 수가 많은 것을 피하기 위해 전염확률밀도함수가 $f(x|\boldsymbol{\theta})$를 따른다고 가정하면 추정해야 할 모수의 수가 $\boldsymbol{\theta}$의 개수로 줄어들게 되고, 이를 이용하여 다음과 같이 $p_j,\ j = 1, \cdots, k$를 구할 수 있다.

$$p_j \propto \int_{j-1}^{j} f(x|\boldsymbol{\theta}) dx$$

본 논문에서는 전염가능 기간 k를 충분히 큰 값 (즉, $k = t - 1$)으로 가정하고, Wallinga와 Teunis (2004)가 제안한 것으로 R_0를 계산하는 방법은 사례 m가 사례 l을 감염시킬 상대우도 p_{lm}을 이용한다. t_l를 사례 l가 발생한 시간이라 하면 $t_l - t_m\ (l > m)$은 사례 m이 사례 l을 감염시킨 간격이 되고, 주어진 시점 t에 대해 $t_l - t_m$의 모든 경우의 수를 구할 수 있다. $t_l - t_m$의 모든 경우의 수를 이용하여 식 (2.3)의 $\widehat{R}_0$를 계산하기 위하여 본 논문에서는 다음 두 가지 경우의 전염확률분포를 정의한다.

1) **경험적 전염확률분포 (empirical transmission distribution)**

$$p_j \propto \frac{C_j}{Call}$$

여기서 C_j는 $t_l - t_m = j$의 모든 경우의 수, $Call$은 $t_l - t_m$의 모든 경우의 수를 나타낸다.

2) 감마 전염확률분포 (gamma transmission distribution)

$t_l - t_m$의 밀도함수를 모수가 $\alpha,\ \beta$인 감마분포에 따른다고 가정하고, p_j를 다음과 같이 정의한다.

$$p_j \propto \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \int_{j-1}^{j} x^{\alpha-1} e^{-x/\beta} dx$$

실용적으로는 $t_l - t_m$에서 α, β의 최우추정값을 Minka (2002), Rahman과 Muraduzzaman (2010)의 방법을 사용하여 구하고, 이를 이용해서 p_j를 계산한다.

4. 적용 및 모형비교

세계보건기구 (WHO)의 보고에 의하면 2014년 3월 26일부터 10월 25일까지 기니에서 1553건의 에볼라 바이러스 환자가 발생했다. Figure 4.1은 기니에서 발생한 누적 감염환자수를 나타내며, Figure 4.2는 동일기간의 일일 발생환자수를 보여준다. 본 논문에서는 기니에서 발생한 에볼라 바이러스 감염환자의 자료에 대하여 앞에서 제시한 두 가지 방법으로 기초 감염 재생산 수와 전염확률분포를 추정하였다.

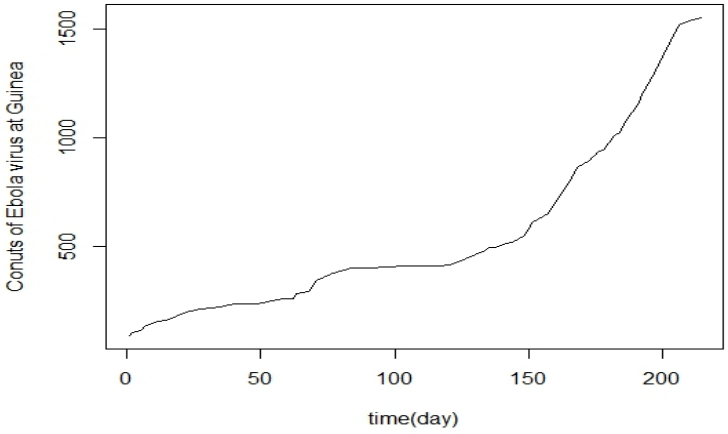

Figure 4.1 Data of the cumulative numbers of infected cases in Guinea

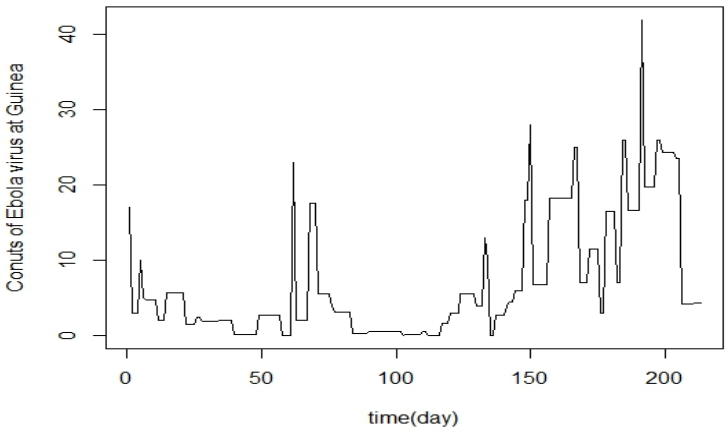

Figure 4.2 Data of the numbers of daily infected cases in Guinea

Table 4.1은 경험적 전염확률분포와 감마 전염확률분포를 사용하여 기초 감염 재생산 수 R_0의 추정값과 이의 표준오차를 보여준다. $t = 50$일 때의 기초감염 재생산 수의 추정값은 각각 0.58054, 0.60113,

$t = 100$일 때의 기초감염 재생산 수의 추정값은 각각 0.742558, 0.789158, $t = 150$일 때의 기초감염 재생산 수의 추정값은 각각 0.898116, 0.959672, $t = 200$일 때의 기초감염 재생산 수의 추정값은 각각 1.598011, 1.856321로, 경험적 전염확률분포를 사용한 추정값이 감마 전염확률분포를 사용한 추정값보다 작게 나타났다. Figure 4.3에서는 초기에는 경험적 전염확률분포를 사용한 추정값이 감마 전염확률분포를 사용한 추정값보다 크지만, $t = 29$부터는 감마 전염확률분포를 사용한 추정값이 더 크게 나타났다. 또한 기초 감염 재생산 수의 추정값 $\widehat{R}_0$은 경험적 전염확률분포를 사용한 경우는 $t = 159$ (8월 30일)를 지나면서, 감마 전염확률분포를 사용한 경우는 $t = 156$ (8월 27일)을 지나면서 1보다 커져 대규모 발생의 단계로 들어가고 있음을 알 수 있다. Figure 4.4는 $t = 200$에서 경험적 전염확률분포, Figure 4.5는 감마 전염확률분포를 나타낸다. 경험적 확률분포함수는 $t = 200$인 경우 199개의 모수를 추정해야하는 단점을 갖고 있으나, 감마 확률분포함수는 두 개의 모수 α, $1/\beta$로 전염확률분포를 간략하게 추정하는 장점을 갖게 된다. Table 4.1은 $t = 50$, 100, 150, 200에서 α, $1/\beta$의 추정값을 보여주고 있다. 이를 살펴보면 α, $1/\beta$의 추정값에서 t가 커짐에 따라 감마분포의 기댓값 $\alpha\beta$가 증가됨을 알 수 있다.

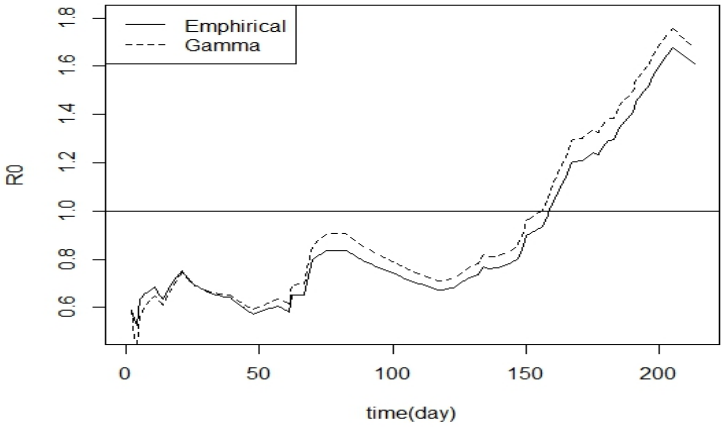

Figure 4.3 The estimates of the basic reproductive numbers R_0 for Ebola virus counts in Guinea using our two methods

Table 4.1 The parameter estimates and their standard error in parentheses for Ebola virus counts in Guinea using our two methods

t	Empirical transmission	Gamma transmission		
	$\widehat{R}_0$	$\widehat{R}_0$	$\widehat{\alpha}$ (shape)	$\widehat{1/\beta}$ (rate)
50	0.58054 (0.04743)	0.60113 (0.04911)	2.01408 (0.00251)	0.07887 (0.00011)
100	0.742558 (0.04149)	0.789158 (0.044102)	1.914493 (0.001686)	0.037854 (0.000037)
150	0.898116 (0.040449)	0.959672 (0.043221)	1.876710 (0.001343)	0.024833 (0.000020)
200	1.598011 (0.044519)	1.683745 (0.046908)	1.856321 (0.001148)	0.001846 (0.000013)

5. 결론 및 토의

전염병과 같이 대규모 확산이 발생하는 상황에서 이의 확산을 방지하고 질병관리 정책을 세우기 위해서 기초 감염 재생산 수 R_0를 추정하는 것이 매우 중요하다. White와 Pagano (2008)가 제시한 기초 감

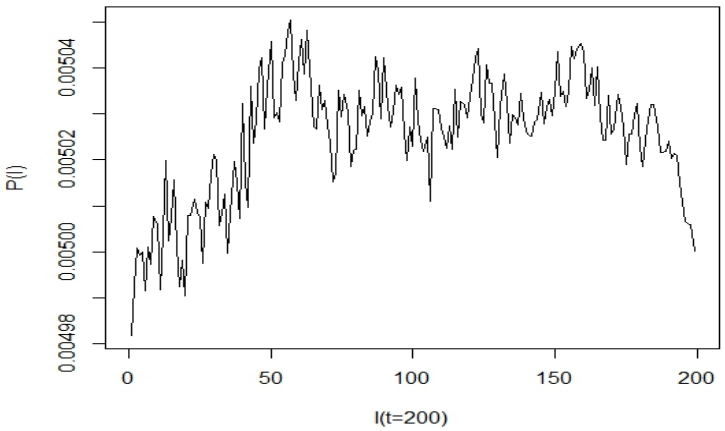

Figure 4.4 The estimated empirical transmission distribution (t=200)

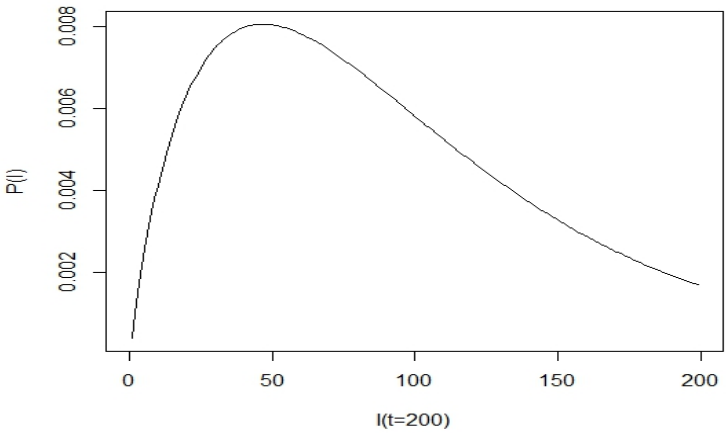

Figure 4.5 The estimated gamma transmission distribution (t=200)

염 재생산 수 R_0를 계산하는 방법에서 전염확률분포를 임의적으로 정해놓고 있으나 본 논문에서는 경험적 전염확률분포와 감마 전염확률분포를 제안하였고 또한 이를 이용하여 실제로 기니에서 발생한 에볼라 바이러스의 사례에 적용하여 보았다. 특히 감마 전염확률분포를 이용하는 경우 모수의 개수를 최소화시켜서 모수추정을 용이하게 구할 수 있다는 장점이 있으며 실시간 추정이 가능하여 실제 상황에서 적용하는데 매우 적절한 방법이었다. 단지 분기과정에서 확률함수를 유도할 때 사망자수를 동시에 고려하지 않았다는 한계를 갖고 있으나 이는 차후에 개선되어야 할 것이다.

References

Althaus, C. L. (2014). Estimating the reproduction number of Ebola virus (EBOV) during the 2014 outbreak in west Africa. *Plos Currents Outbreaks*, **10**, 1-9.

Ball, F. and Donnelly, P. (1995). Strong approximations for epidemic models. *Stochastic Processes and their Applications*, **55**, 1-21.

Becker, N. and Britton, T. (1999). Statistical studies of infectious disease incidence. *Journal of the Royal Statistical Society B*, **61**, 287-307.

Chowell, G., Hengartner, N. W., Castillo-Chavez, C., Fenimore, P. W. and Hyman, J. M. (2004). The basic reproductive number of Ebola and the effects of public health measures: The cases of Congo and Uganda. *Journal of Theoretical Biology*, **229**, 119-126.

Chowell, G. and Nishiura, H. (2014). Transmission dynamics and control of Ebola virus disease (EVD): A review. *BMC Medicine*, **12**, 196.

Evans, R. J. and Mammadov, M. (2014). Dynamics of Ebola epidemics in west Africa 2014. arXiv preprint arXiv:1412.1579, 1-9.

Hwang, J. and Oh, C. (2014). A study on the spread of the foot-and-mouth disease in Korea in 2010/2011. *Journal of the Korean Data & Information Science Society*, **25**, 271-280.

Minka, T. P. (2002). *Estimating a gamma distribution*, Technical Report, Microsoft Research, Cambridge, UK.

Na, K., Choi, I. and Kim, Y. (2010). Estimation of the incubation period of P. vivax malaria in Korea from 2006 to 2008. *Journal of the Korean Data & Information Science Society*, **21**, 1237-1242.

Nishiura, H. and Chowell, G. (2014). Early transmission dynamics of Ebola virus disease (EVD), west Africa, March to August 2014. *Euro Surveill*, **19**, 5-10.

Oh, C. (2014). Derivation of the likelihood function for the counting process. *Journal of the Korean Data & Information Science Society*, **25**, 169-176.

Rahman, M. and Muraduzzaman, S. M. (2010). Likelihood ratio in estimating gamma distribution parameters, *Journal of the Korean Data & Information Science Society*, **21**, 345-354.

Wallinga, J. and Teunis, P. (2004). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology*, **160**, 509-516.

White, L. F. and Pagano, M. (2008). A likelihood-based method for real-time estimation of the serial interval and reproductive number of an epidemic. *Statistics in Medicine*, **27**, 2999-3016.

Journal of the Korean Data & Information Science Society
2015, **26**(1), 161–168

http://dx.doi.org/10.7465/jkdi.2015.26.1.161
한국데이터정보과학회지

A transmission distribution estimation for real time Ebola virus disease epidemic model[†]

Ilsu Choi[1] · Sung-Suk Rhee[2]

[1]Department of Statistics, Chonnam National University
[2]Department of Business Administration, Seowon University

Received 21 December 2014, revised 29 December 2014, accepted 10 January 2015

Abstract

The epidemic is seemed to be extremely difficult for accurate predictions. The new models have been suggested that show quite different results. The basic reproductive number of epidemic for consequent time intervals are estimated based on stochastic processes. In this paper, we proposed a transmission distribution estimation for Ebola virus disease epidemic model. This estimation can be easier to obtain in real time which is useful for informing an appropriate public health response to the outbreak. Finally, we implement our proposed method with data from Guinea Ebola disease outbreak.

Keywords: Basic reproductive number, branching process, Ebola virus, maximum likelihood estimation, transmission distribution.

[†] This paper was supported in part by Chonnam National University Research Grant for 2012.
[1] Professor, Department of Statistics, Chonnam National University, Gwangju 500-757, Korea.
[2] Corresponding author: Professor, Department of Business Administration, Seowon University, Cheongju 361-742, Korea. E-mail: ssrhee@seowon.ac.kr