Izvestiya NAN Armenii. Matematika, vol. 44, no. 2, 2009, pp. 39-50.

EPIDEMICS: THE FITTING OF THE FIRST DYNAMIC MODELS TO DATA

K. DIETZ

University of Tübingen, Tübingen, Germany
E-mail: *klaus.dietz@uni-tuebingen.de*

Dedicated with appreciation to Klaus Krickeberg on his 80th birthday.

"There is probably no more legitimate use of the instrument of statistics than its application to the study of epidemic diseases."

Arthur Ransome (1868)

1. INTRODUCTION

Mathematical infectious disease epidemiology uses dynamic models with interpretable parameters like contact rates and recovery rates to describe individual epidemics and their periodicities. There is a vast literature, scattered in mathematical and recently also in physical journals, which is concerned with the stability of equilibrium points and the identification of threshold parameters for bifurcations. Most of these papers contain no empirical data at all. But for an applied science it is important to make predictions, which can be tested against observations. In view of my historic interests I shall concentrate on the first models, which have the advantage of being still simple while incorporating all the essential elements. Serfling (1952) concludes his historical review of epidemic theory as follows: "However, advance in epidemic theory depends also upon tests of hypotheses and a crucial test must be based on concrete and accurate data. In the past, these have been inadequate".

Since the emergence of new infections like AIDS, SARS, vCJD and others, the field of mathematical infectious disease epidemiology has undergone a tremendous development. A search in PubMed in December 2008 for "mathem* model* infect* disease" yielded 1082 articles, among them 93 reviews. There is a remarkable trend towards the application of disease-specific models to local outbreaks.

The credit for the first epidemic model is often given to Daniel Bernoulli who already in 1760 studied the potential effect of eradicating smallpox by inoculation on

the life expectancy (Dietz and Heesterbeek, 2002). However Bernoulli considered an endemic situation with a constant force of infection, i.e. a fixed value for the yearly incidence of a susceptible individual. He constructed the first so-called catalytic model in the sense of Muench (1934, 1959). For these models it is not necessary to interpret the force of infection as a function of the prevalence of the infection, i.e. the proportion of infective individuals in the total population.

Catalytic models could also be applied to non-infectious diseases because they describe the age-specific proportion of individuals who have experienced a disease. Therefore they are excluded in the following. Also the attempts of Farr (1840) and Brownlee (e.g. 1909) to describe epidemics using implicitly the normal distribution or Pearson curves are not considered, because the parameters are not interpretable.

Among the models with discrete time I shall concentrate on the chain-binomial models of En'ko (1889, 1989), of Reed and Frost (Frost, 1976) and of Greenwood (1931). The most important deterministic model with continuous time is proposed by Kermack and McKendrick (1927). Finally we shall consider the stochastic general epidemic model McKendrick (1926). After a brief introduction of each of these models I shall provide examples of fitting them to observed epidemics and shall discuss the problems of interpreting the estimated parameters, especially if several models are fitted to the same data.

2. CHAIN BINOMIAL MODELS

The first chain-binomial model is due to P. D. En'ko (1844-1916). A brief account of his life can be found in Dietz (1988) and Heyde and Seneta (2001). From 1874 he was the senior doctor at the St. Petersburg Alexander Institute where he gathered data on epidemics of measles and scarlet fever and in 1889 he published his remarkable paper on the course of epidemic diseases. His model can be expressed by the following equations:

$$\begin{aligned} C_{t+1} &= S_t \left\{ 1 - \left(1 - \frac{C_t}{N_t - 1} \right)^{kN_t} \right\}, \\ S_{t+1} &= S_t \left(1 - \frac{C_t}{N_t - 1} \right)^{kN_t}, \\ N_{t+1} &= N_t - C_t. \end{aligned}$$

Here C_t denotes the number of infectious individuals at time t, S_t is the number of susceptible individuals at time t, and N_t is the total size of the population at time t.

The parameter k determines the number of contacts of a susceptible individual:

$$A_t = kN_t$$

The probability

$$\frac{C_t}{N_t - 1}$$

of making a contact with an infective in a homogeneously mixing population of size N_t is used to calculate the probability of making at least one contact with an infective if A_t contacts are made by one susceptible. The equation for the susceptible individuals gives the remaining numbers of susceptible individuals who have avoided contacts with cases. The equation for the total population takes into account that the size of the population is diminished during the course of the epidemic because cases are isolated. This aspect is quite remarkable in En'ko's model and has been ignored in later epidemic models.

The model written as such is a heuristic approximation to a stochastic model because the first and the second equation are based on the binomial distribution, which is only valid if the exponent is an integer. This model therefore implicitly assumes that every susceptible makes exactly the same number of contacts with other individuals in the community. It is remarkable that En'ko compares the predictions of his model with the observations, which he collected over many years. In Table 1 the two epidemics are shown with the best fit together with the parameter values chosen by him. The goodness of fit is strikingly good. The appropriate test by Pearson was only published 11 years later (Pearson, 1900). En'ko was much ahead of his time. Since his work was published in Russian in a journal, which was not widely read, his work was nearly ignored until it was rediscovered some 100 years after its publication.

Year of epidemic:	1874		1879	
Parameters and case numbers	Observed	Fitted	Observed	Fitted
S_0		133		50
k		0.9		0.5
N_0		400		400
C_0	1	1	2	2
C_1	70	79	28	30
C_2	45	53	14	18
C_3	2	0	1	0

Observed and fitted number of cases of measles epidemics at institutions in St. Petersburg

The next chain binomial model is due to the collaboration of Reed and Frost at the Johns Hopkins University. They used the model in jointly teaching epidemiology starting in 1925. It was only published much later because Frost did not consider it such an important contribution. The equation

$$C_{t+1} = S_t(1 - q^{C_t})$$

describes the successive generations of an infectious disease in a closed community. The Reed-Frost model assumes, that the infectious period is rather short compared to the latent period so that subsequent generations of the epidemic can be identified. If one assumes that the number of contacts which one individual makes follows a Poison distribution, then the Reed-Frost model can be considered as a generalization of the En'ko model (Dietz, 1988).

It took many years before the model was fitted to data by Helen Abbey (1952). She states: "Although there is considerable discussion in the literature of epidemic models, very little testing of the models on actual observations has been done, partly because the necessary data are difficult or impossible to obtain." Reed and Frost modified the Soper model (1929) to make allowance for the fact that only one new case would be produced if a given susceptible would have contact with two or more cases. One can write the Reed-Frost model in a deterministic way but also in a stochastic way by replacing the expectations with the full binomial distribution. Helen Abbey states: "Although there is a great deal of published data on the reported cases of infectious diseases, most of this is not useful in testing in adequacy of the model because of variable amounts of underreporting and of lack of information about the number of susceptibles." Because the model assumes uniform mixing among the members of a closed population, it is most applicable in institutions or within families.

It is also important that there are no carriers, which means that it is assumed that every infection brings out clinical symptoms, which can be easily recognized and diagnosed. Therefore the model would be particularly appropriate for measles, rubella and chicken pocks. Helen Abbey uses epidemics, which were reported from the Medical Research Council Special Report "Epidemics in Schools". These are reports of the incidence of epidemic diseases in Naval and boarding schools in England during the years 1932 - 1939. She collects data from some 20 epidemics. In order to divide the daily observations of onsets of cases into generations of cases, one has to

make a certain assumption about the length of the incubation period, i.e. the time from the infection until the onset of symptoms. This already introduces a certain uncertainty about the number of cases in the different generations. Helen Abbey then uses a maximum likelihood estimate in order to estimate the contact parameter of the Reed-Frost model.

Finally she compares the observed and the expected number of cases with the chi-square test. She claims that the chi-square test is applicable because the expected numbers in each time period are calculated from the observed numbers in the previous period and are therefore independent of the previous expected numbers. In some cases the problem of a poor fit arises from the fact that the initial number of susceptible individuals, which enters into the model does not correspond to the actual epidemic. This is due to underreporting of infections in the past and she shows that the fit of the model can be extremely improved if one also estimates the initial number of susceptible individuals. For the reported number of susceptible individuals practically all chi-square values are so high that the corresponding P-Value is less than 0.00001. With an estimated number of susceptible individuals, however, the number of epidemics, which yield an acceptable chi-square value, is considerably increased.

However there remain epidemics, which do not fit according to this criterion. Helen Abbey then goes into an investigation about the possible reasons for the lack of fit. At first she makes a sensitivity analysis with respect to the choice of the incubation interval. She finds that the chi-square values are very sensitive to this choice but concludes that the lack of agreement of the theory from the observations is not due to a particular choice among the intervals. She then examines the possibility that the total population consists of two subpopulations with different contact probabilities. By introducing this extra variation she can improve considerably the fit to the data. In summary, the Reed-Frost theory fails to fit the observations if one takes into account the reported number of susceptible individuals. If one estimates also the initial number of susceptible individuals, then the fit can considerably be improved. In most cases the total number of cases observed equals the estimated number of susceptible individuals.

Another possibility of improving the fit is to assume that the contact rates decline with time during the epidemic. She comes to the remarkable conclusion that "the evidence in this paper does not suggest that any of these factors are likely to be important sources of the discrepancies between the theory and the observations".

The main reason for the lack of fit was probably the fact that she had considered epidemics in too large communities for which the assumption of uniform mixing was violated.

The approach of using the chi-square test by Abbey has been examined in detail by Almond (1954). She comes to the conclusion that for sufficiently large numbers of susceptible individuals the approximation is acceptable provided that one reduces the degrees of freedom if the contact parameter is estimated from the data.

Finally the chain binomial model of Greenwood (1931) is considered. This model is similar to the Reed-Frost model but it is assumed that the actual number of infective individuals does not matter for determining the infection probability as long as there is at least one case in the household. The Greenwood assumption would be appropriate when the household is saturated with infectious material even if only one case is present.

Wilson et al. (1939) criticise Greenwood because he lumps chains of infections together. Wilson et al. have independent data from Providence, which they analyze according to the Greenwood model and find significant discrepancies (see Table 2). The example given is concerned with households of size three with one initial case and two initial susceptible individuals. In the first column, the number of cases in each generation specifies the epidemic chains. For initially two susceptible individuals and one initial case the Greenwood model and the Reed-Frost model are not distinguishable. If one lumps the last two chains together, as was done by Greenwood, then one obtains a perfect fit. One reason for the lack of fit of the two models could be the assumption of homogeneous mixing with the same infection parameter for all 416 households. In later attempts to describe such observed chain frequencies, one has assumed that the contact parameter follows a beta distribution and thereby has improved the fit considerably (Bailey, 1975; Becker, 1989).

The data set, which has been analyzed in greatest detail, has been published by Heasman and Reid (1961). The observations are based on a two-year surveillance of some sixty households with respect to the common cold. All households consist of three children together with their parents. For single introductory cases 16 different chains are possible. They are listed with their observed frequencies in Table 3. The corresponding probabilities for the two classic models Reed-Frost and Greenwood are given in the subsequent columns. The next column gives the observed frequencies. The last two columns contain the expected frequencies. The problem with applying

the chi-square goodness-of-fit test is the large number of cells with small expected values. Schenzle (1982) calculates the P-values by combining the expected values according to three rules: P_1 is calculated by pooling the cells according to the total number of cases, P_2 according to the length of the chain and finally P_3 according to the number of cases in the first generation. The table contains the corresponding P-values for these three types of grouping. In contrast to the analysis by Heasman and Read he concludes that even the Reed-Frost model does not provide an adequate fit. The Greenwood model yields smaller P-values for all three modes of grouping. Schenzle produces 6 other models some of which give a much better fit than the classical models by Reed-Frost and Greenwood.

Fitting the Greenwood model to observations of measles in households with one primary case and two susceptible individuals

Type of chain	Probabilities	Observations	Obs.lumped	Expectations	Exp.lumped
$\{1\}$	q^2	51	51	51.2	51.2
$\{1 \to 1\}$	$2pq^2$	67	67	66.5	66.5
$\{1 \to 1 \to 1\}$	$2p^2q$	36	298	123.0	298.3
$\{1 \to 2\}$	p^2	262		175.3	
	1	416	416	416.0	416.0

Fitting the Reed-Frost model and the Greenwood model to 664 individual chains of common cold in households of five individuals

Schenzle comes to the conclusion that the data do not allow discriminating between the models, probably because the data are lumped together. Despite the fact that this data set has been analysed and reanalysed several times, it does not provide much epidemiological insight. For drawing meaningful conclusions one would need to know more about the sources of heterogeneity among the various household epidemics, i.e. reference to season and to the identity of the individual household.

From these examples one can conclude that the classical chain binomial models are too simple to describe the data in a realistic way. Later analysis have shown that one has to add variability among parameters and/or time dependence in the parameters. In spite of these shortcomings the classical models provide important building blocks for more realistic models if the data are available in sufficient detail.

Chain	Reed-Frost	Greenwood	Observ.	Pred (R-F)	Pred (G)
$\{1\}$	q^4	q^4	423	405.2	400.0
$\{1 \to 1\}$	$4pq^6$	$4pq^6$	131	147.1	147.8
$\{1 \to 1 \to 1\}$	$12p^2q^7$	$12p^2q^7$	36	45.3	46.5
$\{1 \to 1 \to 1 \to 1\}$	$24p^3q^7$	$24p^3q^7$	14	10.5	11.1
$\{1 \to 1 \to 2\}$	$12p^3q^6$	$12p^3q^5$	8	6.0	7.1
$\{1 \to 2\}$	$6p^2q^6$	$6p^2q^4$	24	25.6	34.0
$\{1 \to 2 \to 1\}$	$12p^3q^5(1+q)$	$12p^3q^4$	11	12.7	8.1
$\{1 \to 1 \to 1 \to 1 \to 1\}$	$24p^4q^6$	$24p^4q^6$	4	1.4	1.5
$\{1 \to 1 \to 1 \to 2\}$	$12p^4q^5$	$12p^4q^5$	2	0.8	0.8
$\{1 \to 1 \to 2 \to 1\}$	$12p^4q^4(1+q)$	$12p^4q^4$	2	1.7	1.0
$\{1 \to 1 \to 3\}$	$4p^4q^3$	$4p^4q^3$	2	0.3	0.4
$\{1 \to 2 \to 1 \to 1\}$	$12p^4q^4(1+q)$	$12p^4q^3$	3	1.7	1.1
$\{1 \to 2 \to 2\}$	$6p^4q^2(1+q)^2$	$6p^4q^2$	1	2.0	0.6
$\{1 \to 3\}$	$4p^3q^4$	$4p^3q^2$	3	2.5	3.5
$\{1 \to 3 \to 1\}$	$4p^4q(1+q+q^2)$	$4p^4q$	0	1.1	0.5
$\{1 \to 4\}$	p^4	p^4	0	0.1	0.1
			P_1	0.059	0.001
			P_2	0.012	0.003
			P_3	0.351	0.176

3. THE KERMACK-McKENDRICK MODEL

Kermack and McKendrick (1927) proposed the following epidemic model for the spread of an infection in a homogeneously mixing population:

$$\begin{aligned} \frac{dx}{dt} &= -\kappa xy, \\ \frac{dy}{dt} &= \kappa xy - \gamma y, \\ \frac{dz}{dt} &= \gamma y. \end{aligned}$$

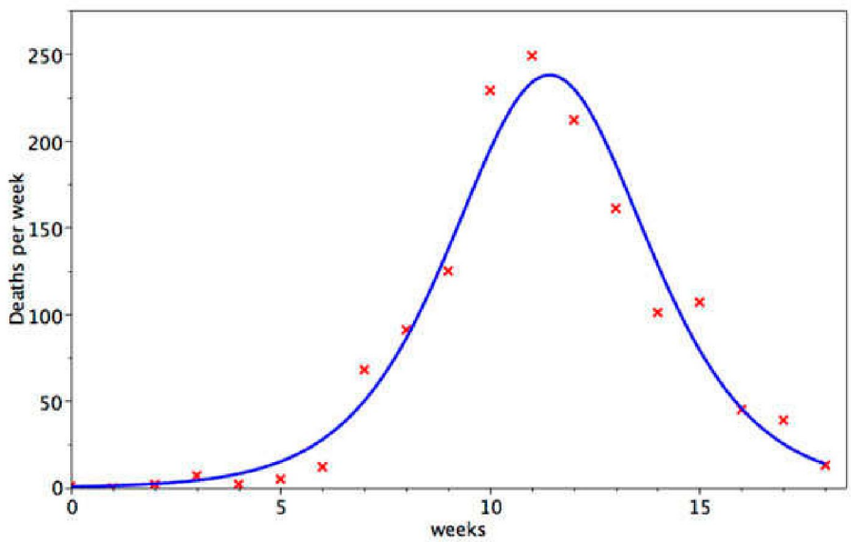
Here x, y and z denote the number of susceptible, infective and immune individuals. The parameter κ is the contact rate and the parameter γ is the recovery rate, i.e. the rate of transfer from the infective into the immune state. Implicitly the model assumes an exponentially distributed time in the infective state. Kermack and McKendrick originally described a more complicated model for which the distribution of the time in the infective state is arbitrary, but in simplifying the assumptions, they come up with the present equations, which often are simply referred to as the Kermack-McKendrick model ignoring that this is just a special case of a more general approach. Kermack and McKendrick noticed that there is no explicit time-dependent solution for this

non-linear system of differential equations, but they derive an approximate solution by using a Taylor expansion up to the second degree of the exponential function, which occurs in the solution for the third equation.

This quadratic function can be solved explicitly. If one assumes that initially there is one infective individual, no immune individuals and x_0 susceptible individuals, then one can express the solution in terms of the basic reproduction number R_0 which is the number of secondary cases which one infective could infect in a completely susceptible population. This number is the key threshold parameter in epidemic theory. Heesterbeek (2002) describes the intricate path until this insight became common knowledge among infectious disease epidemiologists. The incidence of new infections is given by the following equation:

$$\begin{aligned} \frac{dz}{dt} &= \frac{\gamma x_0}{2R_0^2}\, c_1 \operatorname{sech}^2(c_1\gamma t - c_2), \\ c_1 &= \sqrt{(R_0-1)^2 + \frac{2R_0^2}{x_0}}, \\ c_2 &= \tanh^{-1}\left(\frac{R_0-1}{c_1}\right). \end{aligned}$$

Observed and fitted weekly incidence of pneumonic plague in Harbin 1910/1911. The fit is based on the classical Kermack-McKendrick model.



Here the function sech is the inverse of cosh. The advantage of this explicit solution over the fitting of the epidemic curve by the normal distribution or some other Pearson curve is the interpretability of the parameters. Kermack and McKendrick fit their model to the number of deaths from plague "in the island of Bombay over the period 17 December 1905 to 21 July 1906". The observed numbers on the ordinate represent the number of deaths per week and the abscissa denotes the time in weeks. They give the numerical values for the three parameters in this model from which one could deduce that $x_0 = 7722$, $R_0 = 1.32$ and the average duration of the infectious period is 5.6 days. Here they refer to bubonic plague, i.e. they are aware that this is not a direct transmission from man to man.

Therefore they are very cautious and say: "A close fit is not to be expected and deductions as to the actual values of the various constants should not be drawn." Figure 1 shows the fitting of this model to an epidemic of pneumonic plague in Harbin (International Plague Conference, 1912). Here transmission is from man to man and the parameters can be interpreted: they are $x_0 = 2985$, $R_0 = 2.00$ and a mean infectious period of 11 days. The last estimate is biologically realistic if one takes into account that a more detailed model would break this interval down into a latent and an infectious period. This approach is obviously superior over the fitting by curves with non-interpretable parameters.

3.1. **The general stochastic epidemic in a finite population.** In a seminal paper McKendrick (1926, 1997; Dietz 1997) introduced a stochastic model with continuous time for the spread of an epidemic in a finite population. It took more than twenty years until this model was analysed mathematically in more detail by Bartlett, Bailey and others and another twenty-five years until it was first fitted to a smallpox epidemic in Abakaliki, Nigeria, which took place in 1967 in a religious group that refused vaccination. The first attempts to fit the model were based on the information that 30 cases occurred in a homogeneously mixing group of 120 susceptible individuals. The following formula allows to estimate the basic reproduction number R_0 if one knows the initial and the final proportion of susceptible individuals s_0 and s_1, respectively:

$$R_0 = \frac{\ln s_0 - \ln s_1}{s_0 - s_1}$$

Assuming that initially all members of the community were susceptible, one arrives at a surprisingly low basic reproduction number for smallpox of 1.15. If, however,

a certain proportion of the individuals were already immune at the beginning, then the basic reproduction number would be higher. The low estimate of the basic reproduction number has been found in many papers because the detailed report about this epidemic had been ignored. In the meantime this report is available from the homepage of WHO (Thompson and Foege, 1968). Eichner and Dietz (2003) have reanalysed this data set taking into account the following details: the vaccination history of members of this community; the distribution of the time between infection and the onset of fever and the distribution of the prodromal period. One important result is that the infectivity is estimated to be much higher during the period of rash compared to the prodromal period. Altogether the basic reproduction number for this community is estimated to be 6.87 (95 % CI: 4.52, 10.1), i.e. much higher than the previous value of 1.15.

4. CONCLUDING REMARKS

The last example shows clearly that parameter estimates are highly dependent on the underlying model assumptions. Therefore one has to be cautious in interpreting numerical estimates. Ideally two aspects must be fulfilled: a reliable documentation of the events and a realistic model for the description of the observed process. The sophistication of statistical methods has increased considerably. Lately Markov Chain Monte Carlo methods, martingales and Bayesian analyses have been applied to the problem of estimating key epidemiological parameters. Much remains to be done.

References

[1] H. Abbey, "An examination of the Reed-Frost theory of epidemics," Human Biology, **24**, 201-233 (1952).

[2] J. Almond, "A note on the χ^2 test applied to epidemic chains," Biometrics, **10**, 459-477 (1954).

[3] N.T.J. Bailey, *The Mathematical Theory of Infectious Diseases and its Application* (Griffin, London, 1975).

[4] N.G. Becker, *Analysis of Infectious Disease Data* (Chapman and Hall, London, 1989).

[5] J. Brownlee, "Certain considerations of the causation and course of epidemics," Proceedings of the Royal Society Medicine (Epidemiological Section), **2**, 243-258 (1909).

[6] K. Dietz, "The first epidemic model: a historical note on P.D. En'ko," Australian Journal of Statistics, **30A**, 56-65 (1988).

[7] K. Dietz, "Introduction to McKendrick (1926) Applications of mathematics to medical problems," in *Breakthroughs in Statistics Volume III*, S. Kotz and N.L. Johnson (eds), Springer, 17-26 (1997).

[8] K. Dietz and J.A.P. Heesterbeek, "Daniel Bernoulli's epidemiological model revisited," Mathematical Biosciences, **180**, 1-21 (2002).

[9] M. Eichner and K. Dietz, "Transmission potential of smallpox: estimates based on detailed data from an outbreak," American Journal of Epidemiology, **158**, 110-117 (2003).

[10] P.D. En'ko, "On the course of epidemics of some infectious diseases," (In Russian) Vrach, **10**, 1008-1010, 1039-1042, 1061-1063 (1889).

[11] P.D. En'ko, "On the course of epidemics of some infectious diseases" (Translation from Russian by K. Dietz), International Journal of Epidemiology, **18**, 749-755 (1989).
[12] W. Farr, "Progress of epidemics," Second Report of the Registrar General, 91-98 (1840).
[13] W.H. Frost, "Some conceptions of epidemics in general," American Journal of Epidemiology, **103**, 141-151 (1976).
[14] M. Greenwood, "On the statistical measure of infectiousness," Journal of Hygiene Cambridge, **31**, 336-351 (1931).
[15] M.A. Heasman and D.D. Reid, "Theory and observations in family epidemics of the common cold," British Journal of Preventive and Social Medicine, **15**, 12-16 (1961).
[16] J.A.P. Heesterbeek, "A brief history of R_0 and a recipe for its calculation," Acta Biotheoretica, **50**, 189-204 (2002).
[17] C.C. Heyde and E. Seneta, *Statisticians of the Centuries* (Springer, New York, 2001).
[18] *Report of the International Plague Conference: held at Mukden, April, 1911* (Bureau of Printing, Manila, 1912).
[19] W.O. Kermack and A.G. McKendrick, "A contribution to the mathematical theory of epidemics," Proceedings of the Royal Society, **115A**, 700-721 (1927).
[20] A.G. McKendrick, "Applications of mathematics to medical problems" Proceedings of the Edinburgh Mathematical Society, **14**, 98-130 (1926).
[21] A.G. McKendrick, "Applications of mathematics to medical problems," in *Breakthroughs in Statistics Volume III*, S. Kotz and N.L. Johnson (eds), Springer, 27-57 (1997).
[22] H. Muench, "Derivation of rates from summation data by the catalytic curve," Journal of the American Statistical Association, **29**, 25-38 (1934).
[23] H. Muench, *Catalytic Models in Epidemiology* (Cambridge, MA, Harvard U.P, 1959).
[24] K. Pearson, "On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," Philosophical Magazine, **50**, 157-175 (1900).
[25] A. Ransome, "On epidemics, studied by means of statistics of disease," British Medical Journal, **2**, 386-388 (1868).
[26] D. Schenzle, "Problems in drawing epidemiological inferences by fitting epidemic chain models to lumped data," Biometrics, **38**, 843-847 (1982).
[27] R.E. Serfling, "Historical review of epidemic theory," Human Biology, **24**, 145-166 (1952).
[28] H.E. Soper, "The interpretation of periodicity in disease prevalence," Journal of the Royal Statistical Society, **92**, 34-73 (1929).
[29] D. Thompson and W. Foege, *Faith Tabernacle smallpox epidemic. Abakaliki, Nigeria* WHO/SE/68.3 (http://whqlibdoc.who.int/ smallpox/WHO_SE_68.3.pdf) (1968).
[30] E.B. Wilson, C. Bennett, M. Allen and J. Worcester, "Measles and scarlet fever in Providence, R.I., 1929-34 with respect to age and size of family," Proceedings of the American Philosophical Society, **80**, 357-476 (1939).

December 12, 2008