CHAIN BINOMIAL EPIDEMIC MODELS

by

LOUIS HENRY CAIROLI

B.S., John Carroll University, 1971
M.S., John Carroll University, 1974
M.S., Syracuse University, 1976

A MASTER'S REPORT

Submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1988

Approved by:

Major Professor

Table of Contents

List of Tables

Chapter One:

Introduction: Historical Sketch

The accurate mathematical description of the progress of infectious diseases in a population has been a long sought goal of the medical community. Several factors have hampered this formulation until the beginning of this century. Two major factors were a lack of sufficient mathematics to describe the intrinsic properties of the epidemic and secondly, imprecise biological explanations detailing the behavior of how the disease was spread prevented an accurate mathematical formulation of the epidemic. Medical researchers were looking for a useful mathematical model which would describe and allow estimation of such quantities as the course and duration of the epidemic, and the rate of infection given the initial number of infectives and susceptibility rates.

Early biomathematical researchers were concerned principally with fitting curves to epidemic data and using this as a model to deduce the laws that governed epidemics. Brownlee (1906) was the first to publish a paper on the modern mathematical theory of epidemics. He studied the fit of members of Pearson's system of frequency curves to epidemic curves obtained from data of several outbreaks of diseases in the late 19th century.

Concurrently Hamer (1906) presented his classic paper which contained the basic biological elements of an epidemic theory. This has become the foundation for many of the present concepts of the epidemiology of communicable diseases. Hamer assumed that the number of new cases which developed from a given number of infectious cases would be proportional to: (1) the number of existing infectious cases, (2) the number of existing susceptibles, and (3) a constant which depended on factors influencing contact rate between an infectious person and a susceptible. Using these assumptions, Hamer formulated a crude model of the epidemic curve for measles, and the periodic recurance of epidemics.

The work of Ross (1911) and more importantly, Kermack and McKendrick in a series of papers (1927, 1932, 1933, 1937, 1939) developed a deterministic mathematical epidemic theory. Initial development of their equations grew from Ross's work. They introduced a greater degree of generality to their model than previous researchers by describing in a series of differential equations the rates of infectives, susceptibles, and the contact rate between the two groups.

Unfortunately their differential equations did not have simple solutions. Although they were able to achieve approximate integral equations for comparison to survival curves in experimental mouse epidemics, their results failed to describe actual epidemics for human populations. The failure of these deterministic models to agree with published epidemic data led to the consideration of stochastic epidemic models where chance and variation were important considerations.

In a deterministic model, for a given set of initial conditions only one single sequence of events for the epidemic could occur. This is clearly an unrealistic hypothesis since during each disease there are many factors which can exert an influence to change the spread of the disease. If the assumption that the effect of factors influencing the spread of the disease is a random process, then probability concepts may be included in the structure of the model.

In 1926 McKendrick published the first stochastic epidemic model. His deterministic theory considered the number of new cases in a short period of time to be proportional to the number of susceptibles and infectious cases while his new stochastic theory proposed the probability of a new case was proportional to the time interval. McKendrick's efforts attracted little attention when published and he soon continued with Kermack to pursue the deterministic theory. Similar stochastic models were not again considered seriously for almost twenty years.

Although published posthumously (1976), the stochastic model Reed and Frost used in 1928 illustrated the probabilistic nature of epidemics. The model was used in lectures at Johns Hopkins University to simulate epidemics as a teaching tool. The model predicted that under specific conditions the expected number of cases occurring at any stage would have a binomial distribution which depended upon the number of susceptibles and infectives of the previous stage. This led to the formulation of the chain-binomial theory of epidemics.

Independently, Greenwood (1931) also proposed a stochastic model which assumed that the distribution of the number of new cases was independent of the number of present cases. Greenwood assumed a constant infection probability for this model.

Developments in mathematical epidemiology continued through the 1940's and 1950's culminating in Kendall's (1956) paper in which he solved the deterministic differential equations of Kermack and McKendrick, and in Whittle's (1955) paper in which a stochastic threshold theorem was proposed. This result proposed conditions which allowed the calculation of the probability that an epidemic of specific intensity may take place. The theory of chain-binomial models was further enriched by the efforts of Greenwood (1946, 1949), Abbey (1952), and Bailey (1953,1957).

In the 1960's, epidemiologists applied the now established and accepted deterministic and stochastic theories of epidemics to different outbreaks of disease with various degrees of success. Various diseases such as measles, influenza, scarlet fever, and the common cold were modeled using chain binomial methods.

As mathematics and statistics evolved and was applied to the problem of epidemics, more questions were developed and solved. Bailey (1968) and Gani (1969, 1971) developed Markov chain methods in chain binomial models. Ludwig (1975) derived final size distributions for epidemics with arbitrary time dependent infectiousness. Becker, in a series of papers (1977,1980,1981), combined the approaches of Reed-Frost and Greenwood into a general chain binomial model.

The area of stochastic mathematical models for infectious diseases is still very active with numerous open questions. Increased concerns about the statistical aspects of infectious disease modeling continue to attract the attention of researchers. New applications of various other mathematical and statistical techniques to describe the complex behaviors of diseases promise that this will remain a fertile area for researchers to explore for some time.

This report will present several chain binomial model formulations. These models will be fitted to a set of epidemic data and the adequacy of these models to describe this data set will be compared.

Chapter Two:

Theory of Chain Binomial Models

2.1 Epidemiological Ideas

The mathematical formulation of discrete time epidemic models flows from attempts by several investigators to present models which realistically describe the progress of a disease through a population. The usual starting point in model building is the set of assumptions about those factors which control the spread of a disease. These assumptions should create a model which describes actual disease patterns. The epidemic model is then useful as a predictive tool for epidemiologists.

Epidemiologists are concerned with estimating such quantities as the maximum number of cases at the peak of the epidemic, the duration, and total number of cases for the epidemic. The model should be relatively simple mathematically, yet accurate in describing essential features of the epidemic. Chain binomial models satisfy both these criteria. These models have been useful in describing viral diseases such as measles, chicken pox, influenza, and the common cold.

Modeling the spread of these diseases among individuals in a population is a complex task. It is necessary to make several mathematical and biological assumptions about the factors which control the disease process. Mathematically, the population under consideration is assumed to be closed and homogeneously mixed.

The population consists of two classes of individuals. Those who already have the disease are called infectives, and those capable of receiving the disease are called susceptibles. The models assume that all individuals have equal susceptibility, capability to transmit the disease, and the ability to be removed from observation when the transmitting period is over.

After one or more members of the population is infected, the period of time during which the development of the disease is purely internal to the infected person is called the latent or incubation time period. In chain binomial models this latent period is assumed to be constant. This discrete time unit is used in these models to chart the progress of the disease.

The infectious period is the time during which the disease may be transmitted to other members of the population. This time period is contracted to a single point. The infected person may spread the disease upon "adequate contact" to susceptibles in the population. This adequate contact is the probability of contact at any time between an infective and a susceptible sufficient to transmit the infection. Denote this parameter p where $0 < p = 1-q < 1$, and q is the probability of no contact with the infection.

After the disease shows its symptoms, the infected members are removed (isolated) from the rest of the population until recovery. At each time step, a new generation or set of cases following a binomial distribution depending on the parameter p is presented.

The epidemic continues until at some stage there are no new cases generated. An epidemic is defined as the transient outbreak of a disease which is terminated when there are no new infectives.

2.2 Reed-Frost and Greenwood Models

Let N be the initial size of the population. The disease process starts with I_0 individuals ($1 \leq I_0 < N$) becoming infected at time $t = 0$. The remaining $N - I_0$ members of the population are susceptibles. Let S_t denote the number of susceptibles, and I_t the number of infectives just prior to time t. The recursive relationships $S_t = I_{t+1} + S_{t+1}$ for $t = 0,1,\ldots,$ and $N = I_0 + S_0$ hold.

The probability of new infectives during time (t, t+1) may be viewed under two mutually exclusive assumptions:

(I) The probability of new infectives is dependent on the number of infectives in the population at time t.

(II) The probability of new infectives is independent of the number of infectives in the population at time t.

These assumptions parallel the development of chain binomial models. In 1928, Lowell J. Reed and Wade H. Frost in lectures at Johns Hopkins University developed the first chain binomial epidemic model using assumption (I). Although Reed and Frost never published their results, they used these models to help explain disease progress to their students. Independently, in 1931, Major Greenwood published his chain binomial model based on assumption (II). These two models have been accepted by epidemiologists and other health researchers as useful tools in describing the progress of viral diseases.

In the Reed-Frost model the probability of infection during time (t, t+1) depends on the number of infectives present in the population. The corresponding conditional probability of having I_{t+1} individuals infected prior to time $t+1$ is expressed

$$P[I_{t+1} = i_{t+1} \mid I_t = i_t,\ S_t = s_t] = \binom{s_t}{i_{t+1}} (1-q^{i_t})^{i_{t+1}}\, q^{i_t s_{t+1}}$$

$$= \frac{s_t!}{s_{t+1}!\, i_{t+1}!} (1-q^{i_t})^{i_{t+1}}\, q^{i_t s_{t+1}} \qquad \text{for } i_t \geq 0. \quad (2.1)$$

The quantity $1-q^{i_t}$ is the probability of adequate contact with at least one of the s_t susceptibles.

In the Greenwood model, under the assumption of independence, the conditional probability is

$$P[I_{t+1} = i_{t+1} \mid I_t = i_t,\ S_t = s_t] = \binom{s_t}{i_{t+1}} (p^{i_{t+1}})\, q^{s_{t+1}}$$

$$= \frac{s_t!}{s_{t+1}!\, i_{t+1}!} (p)^{i_{t+1}} (q)^{s_{t+1}} \qquad \text{for } i_t \geq 1. \quad (2.2)$$

The relationship $p + q = 1$ holds, where p is the probability of adequate contact. Notice that the Greenwood model is slightly simpler mathematically as a result of the assumption of independence.

2.3 Generalized Chain Binomial Model

Neils Becker (1981) combined the formulations of Reed-Frost and Greenwood into a general chain binomial epidemic model. Using the same notation as above with initial values $I_0 = a$ and $S_0 = k$, then

$$P[S_{t+1} = x,\ I_{t+1} = s - x \mid S_t = s,\ I_t = i]$$

$$= \binom{s}{x} (q_i)^x (1-q_i)^{s-x} \quad \text{for } x = 0,1,\ldots,s. \tag{2.3}$$

The advantage of these models is that it is possible to follow the course of the disease through a population by following the epidemic chain. Consider the chain of infectives specified by counting the number of cases in each generation. The notation 1-3-2-0 is used to denote the epidemic chain in a household consisting of a single introductory case, three first generation cases, two second generation cases, and no new cases in later generations.

In the general chain binomial model, it is possible to write the probability of any chain $a\text{-}i_1\text{-}i_2\text{-}\ldots\text{-}i_r$ of infectives for times $t = 0,1,\ldots$ as

$$P[a\text{-}i_1\text{-}\ldots\text{-}i_r] = \frac{s_0!}{i_1!i_2!\ldots i_r!s_{r+1}!} \prod_{t=0}^{r} q_{i_t}^{s_{t+1}} p_{i_t}^{i_{t+1}} \tag{2.4}$$

where $i_0 = a$, $s_t = k\text{-}i_1\text{-}i_2\text{-}\ldots\text{-}i_t$ for $t = 1,2,\ldots$

Notice that if q_i represents the probability that a susceptible escapes infection when there are i infectives, the Reed-Frost model

may be obtained from the general model by setting $q_i = q^i$, and $q_i = q$ for the Greenwood model.

For example, to calculate the probability of the chain 1-2-1-0 using the general chain binomial model for a household of five:

$$P[1\text{-}2\text{-}1] = P[S_1 = 2 | S_0 = 4, I_0 = 1] P[S_2 = 1 | S_1 = 2, I_1 = 2] P[S_3 = 0 | S_2 = 1, I_2 = 1]$$

$$= \binom{4}{2} q_1^2 p_1^2 \binom{2}{1} q_2^1 p_2^1 \binom{1}{0} q_3^0 p_3^0$$

$$= 12 q_1^3 p_1^2 q_2 p_2 \text{ where } p_i = 1 - q_i , \quad i = 1,2,\ldots$$

This expression may be converted to either the Reed-Frost or Greenwood model formulation by the above transformations. For the above example, $P[1\text{-}2\text{-}1] = 12p^3q^5(1+q)$ for the Reed-Frost model, and $12p^3q^4$ for the Greenwood model.

Direct calculation of the probabilities for either model is possible by using (2.6) for the Reed-Frost model, and equation (2.7) for the Greenwood model.

$$P[i_0\text{-}\cdots\text{-}i_k] = \frac{s_0!}{i_1!i_2!\cdots i_k!s_{k+1}!} \prod_{j=0}^{k-1} (1-q^{i_j})^{i_{j+1}} q^{\sum_{j=0}^{k} i_j s_{j+1}} \qquad (2.6)$$

$$P[i_0\text{-}\cdots\text{-}i_k] = \frac{s_0!}{i_1!i_2!\cdots i_k!s_{k+1}!} p^{\sum_{j=1}^{k} i_j} q^{\sum_{j=1}^{k+1} s_j} \qquad (2.7)$$

It is possible to enumerate and calculate all possible chains and their associated probabilities for small values for N. Tables 1 to 3 provide these summaries.

Table 1. Individual chains for households of three.

Introduction	Chain Type	Frequency	
		Reed-Frost	Greenwood
Single	1	q^2	
	1-1	$2pq^2$	
	1-1-1	$2p^2q$	
	1-2	p^2	
Double	2	q^2	q
	2-1	$1-q^2$	p

Table 2. Individual chains for households of four.

Introduction	Chain Type	Frequency	
		Reed-Frost	Greenwood
Single	1	q^3	q^3
	1-1	$3pq^4$	$3pq^4$
	1-1-1	$6p^2q^4$	$6p^2q^4$
	1-2	$3p^2q^3$	$3p^2q^2$
	1-1-1-1	$6p^3q^3$	$6p^3q^3$
	1-1-2	$3p^3q^2$	$3p^3q^2$
	1-2-1	$3p^3q(1+q)$	$3p^3q$
	1-3	p^3	p^3

Table 2. Individual chains for households of four.

Introduction	Chain Type	Frequency	
		Reed-Frost	Greenwood
Double	2	q^4	q^2
	2-1	$2pq^3(1+q)$	$2pq^2$
	2-1-2	$2p^2q^2(1+q)$	$2p^2q$
	2-2	$p^2(1+q)^2$	p^2
Triple	3	q^3	q
	3-1	$1-q^3$	p

In addition to finding the probability of an individual chain, it is also possible to determine the distribution of the total number of cases in an epidemic for these models. The distribution is obtained by adding together all the probabilities for the relevant chains. Tables 4 to 6 provide the relevant summaries.

Table 3. Individual chains for households of five.

Introduction	Chain Type	Frequency	
		Reed-Frost	Greenwood
Single	1	q^4	q^4
	1-1	$4pq^6$	$4pq^6$
	1-1-1	$12p^2q^7$	$12p^2q^7$
	1-2	$6p^2q^6$	$6p^2q^4$
	1-1-1-1	$24p^3q^7$	$24p^3q^7$
	1-1-2	$12p^3q^6$	$12p^3q^5$
	1-2-1	$12p^3q^5(1+q)$	$12p^3q^4$
	1-3	$4p^3q^4$	$4p^3q^2$
	1-1-1-1-1	$24p^4q^6$	$24p^4q^6$
	1-1-1-2	$12p^4q^5$	$12p^4q^5$
	1-1-2-1	$12p^4q^4(1+q)$	$12p^4q^4$
	1-2-1-1	$12p^4q^4(1+q)$	$12p^4q^3$
	1-2-2	$6p^4q^2(1+q)^2$	$6p^4q^2$
	1-1-3	$4p^4q^3$	$4p^4q^3$
	1-3-1	$4p^4q(1+q+q^2)$	$4p^4q$
	1-4	p^4	p^4

Table 3. Individual chains for households of five.

Introduction	Chain Type	Frequency	
		Reed-Frost	Greenwood
Double	2	q^6	q^3
	2-1	$3pq^6(1+q)$	$3pq^4$
	2-1-1	$6p^2q^6(1+q)$	$6p^2q^4$
	2-2	$3p^2q^4(1+q)^2$	$3p^2q^2$
	2-1-1-1	$6p^3q^5(1+q)$	$6p^3q^3$
	2-1-2	$3p^3q^4(1+q)$	$3p^3q^2$
	2-2-1	$3p^3q^2(1+q)^3$	$3p^3q$
	2-3	$p^3(1+q)^3$	p^3
Triple	3	q^6	q^2
	3-1	$2pq^4(1+q+q^2)$	$2pq^2$
	3-1-1	$2p^2q^3(1+q+q^2)$	$2p^2q$
	3-2	$p^2(1+q+q^2)$	p^2
Quadruple	4	q^4	q
	4-1	$1-q^4$	p

Table 4. Total size of epidemic for households of three.

Introduction	Total Cases	Frequency	
		Reed-Frost	Greenwood
Single	1	q^2	
	2	$2pq^2$	
	3	$p^2(1+2q)$	
Double	2	q^2	q
	3	$1-q^2$	p

Table 5. Total size of epidemic for households of four.

Single Introductory Case Frequencies

Cases	Reed-Frost	Greenwood
1	q^3	q^3
2	$3pq^4$	$3pq^4$
3	$3p^2q^3(1+2q)$	$3p^2q^2(1+2q^2)$
4	$p^3(1+3q+6q^2+6q^3)$	$p^3(1+3q+3q^2+6q^3)$

Table 5. Total size of epidemic for households of four.

Double Introductory Case Frequencies

Cases	Reed-Frost	Greenwood
2	q^4	q^2
3	$2pq^3(1+q)$	$2pq^2$
4	$p^2(1+q)(1+q+2q^2)$	$p^2(1+2q)$

Triple Introductory Case Frequencies

Cases	Reed-Frost	Greenwood
3	q^3	q
4	$1-q^3$	p

Table 6. Total size of epidemic for households of five.

Single Introductory Case Frequencies

Cases	Reed-Frost	Greenwood
1	q^4	q^4
2	$4pq^6$	$4pq^6$
3	$6p^2q^6(1+2q)$	$6p^2q^4(1+2q^3)$
4	$4p^3q^4(1+3q+6q^2+6q^3)$	$4p^3q^2(1+3q^2+3q^3+6q^5)$
5	$p^4(1+4q+10q^2+20q^3+30q^4+36q^5+24q^6)$	$p^4(1+4q+6q^2+16q^3+12q^4+12q^5+24q^6)$

Double Introductory Case Frequencies

Cases	Reed-Frost	Greenwood
2	q^6	q^3
3	$3pq^6(1+q)$	$3pq^4$
4	$3p^2q^4(1+q)(1+q+2q^2)$	$3p^2q^2(1+2q^2)$
5	$p^3(1+q)(1+2q+4q^2+6q^3+6q^4+6q^5)$	$p^3(1+3q+3q^2+6q^3)$

Triple Introductory Case Frequencies

Cases	Reed-Frost	Greenwood
3	q^6	q^2
4	$2pq^4(1+q+q^2)$	$2pq^2$
5	$p^2(1+q+q^2)(1+q+q^2+2q^3)$	$p^2(1+2q)$

These calculations are rather awkward and time consuming even for small size populations. Recurrance relations have been determined for both models.

Let ${}_aP_{N,j}$ denote the probability in a population of size N with a introductory infections and j total number of cases ($a \leq j \leq N$). Each stage follows a binomial distribution with k infectives and N-a-k susceptibles at time t = 1 with probability for the Greenwood model given by:

$$\binom{N-a}{k} p^k q^{N-a-k} \tag{2.8}$$

The probability of j-a new cases, including the k infected, to have a total of j cases is then just ${}_kP_{N-a,j-a}$. The recurrance relationship can be expressed

$${}_aP_{N,j} = \sum_{k=1}^{j-a} \binom{N-a}{k} p^k q^{N-a-k} {}_kP_{N-a,j-a} \quad \text{where } {}_aP_{Na} = q^{N-a}. \tag{2.9}$$

In a similar manner for the Reed-Frost model

$${}_aP_{N,j} = \sum_{k=1}^{j-a} \binom{N-a}{k} (1-q^a)^k q^{a(N-a-k)} {}_kP_{N-a,j-a} \text{ where } {}_aP_{Na} = q^{a(N-a)} \tag{2.10}$$

The Reed-Frost and Greenwood models require estimates of the basic parameter q of the model. This is accomplished by using maximum likelihood methods. It is also possible to test the fit of these models to data by using the chi square goodness of fit test. Examples of these methods will be presented in Chapter 3.

2.4 Markov Formulation

A relatively new and promising analytical tool for chain binomial models was developed by Gani and Jerwood (1971). They reformulated these models in terms of Markov chains. The Greenwood model may be written as

$$P[S_{t+1} = s_{t+1} | S_t = s_t] = \frac{s_t!}{s_{t+1}!(s_t - s_{t+1})!} p^{s_t - s_{t+1}} q^{s_{t+1}}. \quad (2.11)$$

The model in (2.11) satisfies a univariate Markov chain for S_{t+1} with $t = 0,1,\ldots$

The Reed-Frost model may be rewritten as

$$P[S_{t+1} = s_t - i_{t+1},\ I_{t+1} = i_{t+1} |\ S_t = s_t,\ I_t = i_t]$$

$$= \frac{s_t!}{i_{t+1}!(s_t - i_{t+1})!} (1-q^{i_t})^{i_{t+1}} q^{i_t(s_t - i_{t+1})}. \quad (2.12)$$

The Reed-Frost model may then be described in terms of a bivariate Markov chain for the pair of variables S_t and I_t.

To see the advantages of this formulation, consider the mathematical results of this setting for chains terminating for the first time at $T = t$ when $X_t = X_{t-1}$. The interpretation of X_t is the variable number of susceptibles in the epidemic. Let $\{X_t\}$ be a Markov chain with finite state space and transition matrix

$M = \{m_{ij}\}$ where $\quad 0 \le m_{ij} \le 1 \quad$ for $i \neq j$

$\quad\quad\quad\quad\quad\quad\quad\quad 0 < m_{jj} \le 1 \quad$ for $i, j = 0, 1, \ldots k.$

Define the elements of the transition matrix **M** for the Greenwood model to be:

$$m_{ij} = \begin{cases} \binom{j}{j} p^{i-j} q^{j} & \text{for} \quad 0 \le j \le i \\ 0 & \text{for} \quad i < j \le k \end{cases} \tag{2.16}$$

The (k+1) x (k+1) transition matrix **M** has the lower triangular form:

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ p & q & 0 & \cdots & 0 \\ p^2 & 2pq & q^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ p^k & kp^{k-1}q & \binom{k}{2} p^{k-2} q^2 & \cdots & q^k \end{bmatrix} \tag{2.17}$$

Let **A** be the vector of initial probabilities where the i th row $\mathbf{A}_i = [0, 0, \ldots, 0, 1]$. $\mathbf{A}_i$ is the 1 x (k+1) row vector with 1 in the (i+1)st position. Define $\mathbf{R} = [1, q, q^2, \ldots, q^k]'$ to be the column vector of diagonal elements from **M**, the transition probability matrix. Define $\mathbf{B} = \mathbf{M} - \mathbf{R}$. The probability of the epidemic stopping in state j at time T = t given the epidemic started with $X_0 = i$ at time zero may be expressed as

$$P[T = t, X_t = X_{t-1} = j | X_0 = i] = (\mathbf{B}^{t-1})_{ij} m_{jj}. \tag{2.18}$$

Summing equation (2.18) over $0 \le j \le k$ yields

$$\begin{aligned} P[T = t | X_0 = i] &= \sum_{j=0}^{k} P[T = t, X_t = X_{t-1} = j | X_0 = i] \\ &= \sum_{j=0}^{k} (\mathbf{B}^{t-1})_{ij} m_{jj} \\ &= \mathbf{A}_i' \mathbf{B}^{t-1} \mathbf{R} \qquad \text{for} \quad 1 \le t < \infty. \end{aligned} \tag{2.19}$$

It is then possible to construct the probability generating function (p.g.f.) of the epidemic termination time T. The p.g.f. is defined to be $\sum_{t=0}^{\infty} \theta^t P[T = t | X_0 = i]$.

From (2.19), the p.g.f. is $\sum_{t=1}^{\infty} \theta^t A_i' B^{t-1} R$ and rewritten as

$$= \sum_{t=1}^{\infty} A_i' \theta^{t-1} B^{t-1} \theta R$$

$$= A_i'(I - \theta B)^{-1} \theta R \qquad \text{for } 0 \le \theta \le 1 \qquad (2.20)$$

where I is the (k+1) x (k+1) identity matrix. The inverse in (2.20) always exists since $|\theta B| < 1$.

Gani and Jerwood refer to the distribution in (2.19) as a Markov geometric distribution because its structure resembles the ordinary geometric distribution.

The expected length of the epidemic and its variance is found by computing the derivatives of (2.20) and evaluating the resulting expressions at $\theta = 1$.

The expected length of epidemic, denoted E[T], is found to be

$$E[T] = A_i' (I - B)^{-2} R. \qquad (2.21)$$

The variance, V[T], is obtained from the well known formula

$$V[X] = E[X(X-1)] + E[X] - \{E[X]\}^2. \qquad (2.22)$$

Using (2.22), the expression for the variance is

$$V[T] = 2A_i'B(I-B)^{-3}R + A_i'(I-B)^{-2}R - \{A_i'(I-B)^{-2}R\}^2. \qquad (2.23)$$

Chapter Three:

Application of Chain Binomial Models to Epidemic Data

3.1 Epidemic Data

A classic data set in the area of mathematical epidemiology was that presented by Heasman and Reid (1961). They presented data on 45-50 London households consisting of five people, two parents and three children. 664 family epidemics of acute coryza (common cold) were investigated over a two year period. The date of onset of illness and the number of upper respiratory infections experienced by family members were recorded on time charts for each family in days. By examining these charts the progress of the disease could be described in terms of the chains of infections as noted in Chapter Two. Table 7 presents data for single primary cases (one case to begin the home epidemic).

Household data as presented in Table 7 provide the ideal population for testing the adequacy of chain binomial models. It is only for such small groups as households that the different possible chains can be readily classified. Furthermore, many of the simplifying underlying assumptions of the models presented are likely to be satisfied within this type of population than in a general community setting.

The usual chi square goodness of fit test may be used to test model fit to the data. A comparison of actual versus expected

frequencies under the hypothesized model is used to check model adequacy.

Table 7. Heasman and Reid epidemic data.

Chain	Observed Frequency
1-0	423
1-1-0	131
1-2-0	24
1-1-1-0	36
1-3-0	3
1-1-2-0	8
1-2-1-0	11
1-1-1-1-0	14
1-4	0
1-3-1	0
1-1-3	2
1-2-2	1
1-2-1-1	3
1-1-2-1	2
1-1-1-2	2
1-1-1-1-1	4
	664

3.2 The General Chain Binomial Model

The formulation of a general chain binomial model is very attractive in the sense that this model combines the Reed-Frost and Greenwood formulations. Additionally, maximum likelihood (ML) estimates may be obtained for the q_i, $i = 1,2,\ldots,k-1$ where q_i denotes the probability of a given susceptible escaping infection when exposed to the i th infective of any generation.

The log likelihood function is expressed as

$$\ell n\, L(q_1,\ldots,q_{k-1}) = \sum_{j=1}^{k-1} \{x_j \ell n\, q_j + (m_j - x_j)\, \ell n\, (1-q_j)\} + K, \qquad (3.1)$$

where m_j denotes the total number of exposures to the j th infective, x_j denotes the total number of those escaping infection, and K is a constant.

Computing partial derivatives of (3.1) with respect to the q_i, setting the derivatives equal to zero, and solving, yield the ML estimates

$$\hat{q}_j = \frac{x_j}{m_j} \qquad \text{for } j = 1,2,\ldots,k-1. \qquad (3.2)$$

For the Heasman-Reid data, the likelihood function becomes:

$$L(q_1, q_2, q_3) = (nq_1^4)^{423}(4nq_1^6p_1)^{131}\ldots(np_1^4)^4 \qquad (3.3)$$

$$L(q_1, q_2, q_3) = Kq_1^{3000}(1-q_1)^{397}q_2^{70}(1-q_2)^{18}q_3^3 \qquad (3.4)$$

Values of m_j and x_j can easily be read from (3.4) as follows:

$$x_1 = 3000 \qquad x_2 = 70 \qquad x_3 = 3 \qquad (3.5)$$
$$m_1 = 3397 \qquad m_2 = 88 \qquad m_3 = 3.$$

The ML estimates from (3.5) and (3.2) are $\hat{q}_1 = .8831$ (variance $= .000155$), $\hat{q}_2 = .7954$ (variance $= .0002451$), and $\hat{q}_3 = 1.0$. Using these estimates with $n = 664$ in the general model allows the calculation of expected frequencies for each of the chains (Table 8).

Since $\hat{q}_3$ is based on only three Bernoulli trials and several chains occur with frequencies fewer than five, these chains are pooled together to form one class. In this pooled case only q_1 and q_2 are estimated.

The chi square goodness of fit test yields a value of $\chi^2 = 9.573$ with 5 df. This result is not significant at the 5% level. The general chain binomial model does adequately describe the data.

Table 8. Fitted General Model.

Chain	Expected	Observed	Fitted
1-0	nq_1^4	423	403.83
1-1-0	$4nq_1^6p_1$	131	147.26
1-1-1-0	$12nq_1^7p_1^2$	36	45.61
1-1-1-1-0	$24nq_1^7p_1^3$	14	10.66
1-1-1-1-1	$24nq_1^6p_1^4$	4	1.41
1-1-1-2	$12nq_1^5p_1^4$	2	0.79
1-1-2	$12nq_1^4p_1^3q_2$	8	6.15
1-1-2-1	$12nq_1^4p_1^3p_2$	2	1.58
1-1-3	$4nq_1^3p_1^4$	2	0.34
1-2	$6nq_1^2p_1^2q_2^2$	24	26.86
1-2-1	$12nq_1^3p_1^2q_2p_2$	11	12.22
1-2-1-1	$12nq_1^2p_1^3q_2p_2$	3	1.62
1-2-2	$6nq_1^2p_1^2p_2^2$	1	1.77
1-3	$4nq_1p_1^3q_3$	3	3.74
1-3-1	$4nq_1p_1^3p_3$	0	0.00
1-4	np_1^4	0	0.12
		664	663.96

3.3 Greenwood Model

The likelihood function provided by the general chain binomial model (3.4) may be transformed into the likelihood function for the Greenwood model by substituting $q_i = q$. The likelihood function then becomes

$$L(q) = K\, q^{3000}\, (1-q)^{397}\, q^{70}\, (1-q)^{18}\, q^{3} \tag{3.6}$$

$$L(q) = K\, q^{3073}\, (1-q)^{415}. \tag{3.7}$$

ML estimates for q are found in the usual manner from the log likelihood function. $\hat{q} = .88102$ with variance .0001579 was found.

Expected frequencies may then be found for each chain. Table 9 presents the summary for this model. Using the same number of classes as in the generalized model case, the chi square statistic tests the adequacy of the model fit to the data.

The chi square goodness of fit test yields a value of $\chi^2 = 16.147$ with 6 df. This value is significant at the 5% level. Hence, the Greenwood model does not adequately describe the data. The rejection of this model in favor of other models calls into question the assumption of a constant infection rate for this disease. By using this technique of fitting different models with various assumptions for the disease, researchers are able to test their assumptions about disease factors. Various aspects of diseases which may not totally be known can be discovered in this manner.

Table 9. Fitted Greenwood Model.

Chain	Expected	Observed	Fitted
1-0	nq^4	423	400.01
1-1-0	$4nq^6p$	131	147.78
1-1-1-0	$12nq^7p^2$	36	46.48
1-1-1-1-0	$24nq^7p^3$	14	11.06
1-1-1-1-1	$24nq^6p^4$	4	1.49
1-1-1-2	$12nq^5p^4$	2	0.85
1-1-2	$12nq^5p^3$	8	7.12
1-1-2-1	$12nq^4p^4$	2	0.96
1-1-3	$4nq^3p^4$	2	0.36
1-2	$6np^2q^4$	24	33.98
1-2-1	$12nq^4p^3$	11	8.08
1-2-1-1	$12nq^3p^4$	3	1.24
1-2-2	$6nq^2p^4$	1	0.62
1-3	$4nq^2p^3$	3	3.47
1-3-1	$4nq^4p^4$	0	0.47
1-4	np^4	0	0.13
		664	663.75

3.4 Reed-Frost Model

The transformation $q_i = q^i$ for i = 1,2,3 in the general chain binomial model converts the general chain binomial model's likelihood function (3.4) into the Reed-Frost likelihood function

$$L(q) = K\, q^{3000}\, (1-q)^{397}\, (q^2)^{70}\, (1-q^2)^{18}\, (q^3)^3 \qquad (3.8)$$

$$L(q) = K\, q^{3149}\, (1-q)^{415}\, (1+q)^{18}. \qquad (3.9)$$

The ML estimate of q is obtained in the usual manner from the log likelihood function. Computing the derivative of the log of the likelihood function (3.9) with respect to q yields a quadratic expression in q. Solving this resulting quadratic expression in q yields the ML estimate $\hat{q} = .8838$ with variance .0001547. Using this estimate with n = 664 provides the expected frequencies of Table 10. Chains with fewer than five observed frequencies were pooled.

The chi square goodness of fit test value of $\chi^2 = 9.127$ with 6 df was not significant at the 5% level. One may then conclude that the Reed-Frost chain binomial model does adequately describe the data.

Table 10. Fitted Reed-Frost Model.

Chain	Expected	Observed	Fitted
1-0	nq^4	423	405.12
1-1-0	$4nq^6p$	131	147.08
1-1-1-0	$12nq^7p^2$	36	45.31
1-1-1-1-0	$24nq^7p^3$	14	10.53
1-1-1-1-1	$24nq^6p^4$	4	1.38
1-1-1-2	$12nq^5p^4$	2	0.78
1-1-2	$12nq^5p^3$	8	5.96
1-1-2-1	$12nq^4p^4(1+q)$	2	1.67
1-1-3	$4nq^3p^4$	2	0.33
1-2	$6np^2q^4$	24	25.63
1-2-1	$12nq^4p^3(1+q)$	11	12.70
1-2-1-1	$12nq^3p^4(1+q)$	3	1.67
1-2-2	$6nq^2p^4(1+q)^2$	1	2.01
1-3	$4nq^4p^3$	3	2.54
1-3-1	$4nq\ p^4(1+q+q^2)$	0	1.14
1-4	np^4	0	0.12
		664	663.97

3.5 Modified Reed-Frost Model

One of the basic assumptions of the Reed-Frost model is that the probability q of any given susceptible escaping infection by any infected person is constant. Epidemiologically, this assumption may not always be true. There are various social and genetic factors which may have a substantial effect on this probability causing it to vary from household to household. These factors, including age, nutrition, sex, and hereditary immune system components, give support to a non-constant probability q of escaping infection.

The easiest method of considering the variability of q is to allow q to vary according to some known distribution. A reasonable choice for this distribution is the beta distribution. The beta distribution provides q with values between 0 and 1 in addition to allowing a great deal of flexibility in the shape the distribution may assume. This accounts for some of the various factors which may effect the values of q in households.

The Reed-Frost model may be expressed as

$$P[S_{t+1} = u,\ I_{t+1} = s-u \mid S_t = s,\ I_t = i] = \binom{s}{u} q^u (1-q)^{s-u} \qquad (3.10)$$

and assumes q is the same for all members of a household but varies between households. This is accomplished by the mixing distribution

$$dF(q) = \frac{1}{B(x,y)} q^{x-1} (1-q)^{y-1} dq \quad \text{for} \quad 0 \le q < 1,\ x,y > 0 \qquad (3.11)$$

where

$$B(x,y) = \frac{\Gamma(x)\,\Gamma(y)}{\Gamma(x+y)} \quad \text{and} \quad \Gamma(x) = (x-1)!$$

The required expectations can then be obtained by averaging the frequencies for each chain type over the mixing distribution (3.11). Since each of the chains of the Reed-Frost model involves linear expressions of the form $q^h p^k$ where $h + k = s$, the expectations may be found by

$$E[q^h p^k] = \int_0^1 \frac{1}{B(x,y)} q^{h+x-1} (1-q)^{k+y-1} \, dq \tag{3.12}$$

$$= \frac{B(x+h,\ y+k)}{B(x,y)}$$

$$= \frac{x(x+1)\ldots(x+h-1)y(y+1)\ldots(y+k-1)}{(x+y)(x+y+1)\ldots(x+y+h+k-1)} \, . \tag{3.13}$$

Bailey (1953), Griffiths (1973), and others have suggested that the reparametrization

$$q = \frac{x}{x+y} \qquad \text{and} \qquad z = \frac{1}{x+y} \, , \tag{3.14}$$

is useful in allowing ML estimates to be found.

Further notation to simplify the expectation expressions for the chains is defined by

$$z(n) = \prod_{i=0}^{n} (1 + iz), \tag{3.15}$$

$$z_q(n) = \prod_{i=0}^{n} (q + iz), \tag{3.16}$$

$$z_p(n) = \prod_{i=0}^{n} (p + iz). \tag{3.17}$$

Application of (3.15)-(3.17) allows the expected probabilites for households of size five to be expressed (Table 11) in terms of the two parameters q and z.

Table 11. Expected Probabilities for Modified Reed-Frost Model.

Chain	Probabilities	Expected Value of Probability
1-0	nq^4	$z_q(3)/z(3)$
1-1-0	$4nq^6p$	$4z_q(5)\ z_p(0)/z(6)$
1-1-1-0	$12nq^7p^2$	$12z_q(6)z_p(1)/z(8)$
1-1-1-1-0	$24nq^7p^3$	$24z_q(6)z_p(2)/z(9)$
1-1-1-1-1	$24nq^6p^4$	$24z_q(5)z_p(3)/z(9)$
1-1-1-2	$12nq^5p^4$	$12z_q(4)z_p(3)/z(8)$
1-1-2	$12nq^5p^3$	$12z_q(5)z_p(2)/z(8)$
1-1-2-1	$12nq^4p^4(1+q)$	$12z_q(3)z_p(3)(1+q+12z)/z(8)$
1-1-3	$4nq^3p^4$	$4z_q(2)z_p(3)/z(6)$
1-2	$6np^2q^4$	$6z_q(5)z_p(1)/z(7)$
1-2-1	$12nq^4p^3(1+q)$	$12z_q(4)z_p(2)(1+q+13z)/z(8)$
1-2-1-1	$12nq^3p^4(1+q)$	$12z_q(3)z_p93)(1+q+12z)/z(8)$
1-2-2	$6nq^2p^4(1+q)^2$	$6z_q(1)z_p(3)[76z^2+(17+19q)z+(1+q)^2]/z(7)$
1-3	$4nq^4p^3$	$4z_q(3)z_p(2)/z(6)$
1-3-1	$4nq\ p^4(1+q+q^2)$	$4z_q(0)z_p(3)[38z^2+(12+9q)z+(1+q)^2]/z(7)$
1-4	np^4	$z_p(3)/z(3)$

The likelihood function may now be formed by taking the product of the expected value of each of the chains raised to the power of its observed frequency. Estimates for q and z may be found by iteratively solving the derivatives of the likelihood function. The IMSL subroutine ZXMWD was used with the log likelihood function

$$\begin{aligned}
\ell n\, L = C &+ 664\, \ell n\, q + 664\, \ell n(q+z) + 663\, \ell n(q+2z) \\
&+ 661\, \ell n(q+3z) + 230\, \ell n(q+4z) + 217\, \ell n(q+5z) \\
&+ 50\, \ell n(q+6z) + 241\, \ell n\, p + 110\, \ell n(p+z) \\
&+ 50\, \ell n(p+2z) + \ell n[76z^2 + (17+19q)z + (1+q^2)] \\
&+ 14\, \ell n(p+3z) + 5\, \ell n(1+q+12z) + 11\, \ell n(1+q+13z) \\
&- 664\, \ell n(1+z) - 664\, \ell n(1+2z) - 664\, \ell n(1+3z) \\
&- 241\, \ell n(1+4z) - 241\, \ell n(1+5z) - 241\, \ell n(1+6z) \\
&- 105\, \ell n(1+7z) - 80\, \ell n(1+8z) - 18\, \ell n(1+9z),
\end{aligned} \tag{3.18}$$

where C is a constant.

The estimates of q and z were found to be:

$$\hat{q} = .8887 \qquad \hat{z} = .0222$$

$$\mathrm{Var}(\hat{q}) = .3433 \times 10^{-4} \qquad \mathrm{Var}(\hat{z}) = .1237 \times 10^{-4}$$

$$\mathrm{Cov}(\hat{q},\hat{z}) = .1900 \times 10^{-4}$$

Using these estimates of q and z, the expected frequencies for the chains were calculated (Table 12).

The goodness of fit test for this modified Reed-Frost model with beta distribution yielded $\chi^2 = 2.94$ with 5 df. This value, when compared to the tabled chi square value of 11.070 for 5% level of significance, revealed an excellent fit of this model to the data. The excellent fit of this modified model to the data implies the presence of household to household variability for this disease. This aspect of the disease may not have been determined from merely observing the data.

The strategy for health researchers interested in using these models is clear. They should try several of these chain binomial models to see which of them "best" fits the epidemic data. Once the "best" model fit has been determined, the researchers should examine the model to determine what the epidemiological implications of the model are. In this manner additional research information can be obtained for the disease.

Table 12. Fitted Modified Reed-Frost Model.

Chain	Observed	Fitted
1-0	423	420.81
1-1-0	131	133.55
1-1-1-0	36	40.01
1-1-1-1-0	14	10.42
1-1-1-1-1	4	1.81
1-1-1-2	2	1.09
1-1-2	8	6.11
1-1-2-1	2	2.41
1-1-3	2	0.53
1-2	24	23.15
1-2-1	11	13.34
1-2-1-1	3	2.41
1-2-2	1	3.21
1-3	3	2.84
1-3-1	0	2.18
1-4	0	0.24
	664	664.11

3.6 Conclusions

A comparison of the fit of the four chain binomial models presented in this report reveals several observations. The overall usefulness of this class of models can be seen in the similarity of the estimates of the escape rate of infection q provided by all the models.

The best fit was achieved by the modified Reed-Frost model. This model provided additional insight into the nature of the disease by showing variabilty from household to household of the escape rate from infection.

The worst fit was the Greenwood model. This lack of fit was not surprising in light of the additional information about the disease provided by the modified model. The Greenwood assumption of a constant infection rate from household to household is clearly not true for this disease.

This class of models provides insight into the difficult job of modeling the progress of disease in households. Health researchers require models which accurately describe the progress of diseases and which provide reliable estimates of infection rate so that public health policies may be determined to benefit the general public. Chain binomial models satisfy these criteria.

Chapter Four:

Summary

Chain binomial models provide an extremely useful mathematical description of epidemic processes in small, household size groups. These relatively simple yet versitile models allow health researchers great flexibility in accurately modeling viral diseases. The models can provide estimates for infection rates by application of well known statistical procedures from the theory of maximum likelihood estimation.

The adequacy of these models to describe epidemic data may be tested by using chi square goodness of fit tests. By combining the fit of the models with the estimates obtained from these models, additional information concerning the underlying biological assumptions of the disease may be discovered. Such additional insights are extremely useful in man's war against disease.

Bibliography

Abbey, H. (1952) "Examination of Reed-Frost epidemic equations. "Human Biology, 24: 201-233.

Bailey, N. T. J. (1953) "The use of chain-binomials with a variable chance of infection for the analysis of intra-household epidemics." Biometrika, 40: 177-185.

Bailey, N. T. J. (1957) The Mathematical Theory of Epidemics. Griffen, London.

Bailey, N. T. J. (1968) "A perturbation approximation to the simple stochastic epidemic in a large population." Biometrika, 55: 199-210.

Becker, N. (1977) "Estimation for discrete time branching processes with applications to epidemics." Biometrics, 33:515-522.

Becker, N. (1980) "An epidemic chain model." Biometrics, 36:249-254.

Becker, N. (1981) "A general chain binomial model for infectious diseases." Biometrics, 37: 251-258.

Brownlee, J. (1906) "Statistical Studies in Immunity: The Theory of an Epidemic." Proceedings Royal Society Edinburgh, 26: 484-521.

Frost, W. H. (1976) "Some conceptions of epidemics in general." American Journal of Epidemiology, 103: 141-151.

Gani, J. (1969) "A chain binomial study of inoculation in epidemics." Bulletin of the International Statistical Institute, 43(2):203-204.

Gani, J. and D. Jerwood. (1971) "Markov chain methods in chain binomial epidemic models." Biometrics, 27: 591-603.

Greenwood, M. (1931) "On the statistical measure of infectiousness." Journal of Hygiene Cambridge, 31: 336-351.

Greenwood, M. (1946) "The statistical study of infectious diseases." Journal Royal Statistical Society, Part II, 109: 85-110

Greenwood, M. (1949) "The infectiousness of measles." Biometrika, 36: 1-8.

Griffiths, D. A. (1973) "Maximum likelihood estimation for the beta-binomial distribution and an application to the household diistribution of the total number of cases of a disease." Biometrics, 29: 637-48.

Hamer, W. H. (1906) "Epidemic disease in England." Lancet, 1: 733-739.

Heasman, M. A. and D. D. Reid. (1961) "Theory and observation in family epidemics of the common cold." British Journal of Preventive Medicine, 15: 12-16.

Jacquez, J. (1987) "A note on chain-binomial models of epidemic spread." Mathematical Biosciences, 87:73-82.

Kendall, D. G. (1956) "Deterministic and stochastic epidemics in closed populations." Proceedings Third Berkeley Symposium on Mathematical Statistics and Probability, 4:149-165.

Kermack, W. O. and A. G. McKendrick. (1927-1939) "Contributions to the mathematical theory of epidemics." (Parts I-V)
Proc. Royal Society, Part A, 115: 700-721 (1927).
Proc. Royal Society, Part A, 138: 55-83 (1932).
Proc. Royal Society, Part A, 141: 94-122 (1933).
Journal of Hygiene Cambridge, 37: 172-187 (1937).
Journal of Hygiene Cambridge, 39: 271-288 (1939).

Ludwig, D. (1975) "Final size distribution for epidemics." Mathematical Biosciences, 23: 33-46.

Ross, R. (1911) The Prevention of Malaria (2nd Edition) Murray, London.

Whittle, P. (1955) "The outcome of a stochastic epidemic - A note on Bailey's paper." Biometrika, 42: 116-122.

CHAIN BINOMIAL EPIDEMIC MODELS

by

LOUIS HENRY CAIROLI

B.S., John Carroll University, 1971
M.S., John Carroll University, 1974
M.S., Syracuse University, 1976

AN ABSTRACT OF A MASTER'S REPORT

Submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1988

ABSTRACT

Four chain binomial models were discussed. The Reed-Frost, modified Reed-Frost, Greenwood, and Becker generalized chain binomial models were presented. Comparisons of these models to epidemiological data were made. Parameter estimates for infection escape rate were obtained by maximum likelihood methods. The adequacy of model fit to data was judged by chi square goodness of fit tests.