The centroid method was originally derived as a mathematical approximation to the more difficult principal-axes procedure when computers were not generally available. Even though the centroid solution yields the same complexity of variables and factors, and also has the same variance contributions of the factors as the principal-axes procedure, it does not share the other important mathematical properties of the principal-axes solution, which include uniqueness and orthogonality. A treatment of the centroid method in factor analysis is given in Cureton & D'Agostino [2, Chapter 2].

References

[1] Burt, C. (1917). *The Distribution and Relations of Educational Abilities*. King & Son, London.

[2] Cureton, E.E. & D'Agostino, R.B. (1983). *Factor Analysis: An Applied Approach*. Lawrence Erlbaum, Hillsdale.

[3] Thurstone, L.L. (1931). Multiple factor analysis, *Psychological Review* **38**, 406–427.

(*See also* **Factor Analysis, Overview**)

RALPH B. D'AGOSTINO, SR &
HEIDY K. RUSSELL

Cephalic Index *see* Anthropometry

Chain Binomial Model

Let time be discrete and indexed $t = 0, 1, \ldots$. Let S_t be the number of individuals at risk for the event of interest (e.g. infection or death) at the beginning of time interval t, and let I_t be the number that experienced the event of interest at the beginning of time interval t. The event has a duration of at least one time interval. We let $p_t = 1 - q_t = f(t, \theta, I_t)$ be the probability that an at-risk individual has a new event at the beginning of time interval time $t+1$, with parameter $\boldsymbol{\theta}$. As shown, this probability can be a function, $f(\cdot)$, of t and I_t. We usually start with a closed population of $n = S_0 + I_0$ individuals. Then I_{t+1} is a **binomial** random variable that follows the conditional probability mass function

$$\Pr(I_{t+1} = i_{t+1} | S_t = s_t, p_t) = \binom{s_t}{i_{t+1}} p_t^{i_{t+1}} q_t^{s_t - i_{t+1}}, \quad s_t \geq i_{t+1}. \quad (1)$$

In many cases, S_t is updated via the relationship

$$S_{t+1} = S_t - I_{t+1}, \quad (2)$$

although other relationships are possible (see below). The conditional expectation and variance of I_{t+1}, respectively, are

$$\mathrm{E}(I_{t+1} | s_t, p_t) = s_t p_t, \quad (3)$$

$$\mathrm{var}(I_{t+1} | s_t, p_t) = s_t p_t q_t. \quad (4)$$

Eqs (1) and (2) form the classical chain binomial model. Formal mathematical treatment of the model involves formulation of the discrete, two-dimensional **Markov chain** $\{S_t, I_t\}_{t=0,1,\ldots}$. I_t is the (binomial) random variable of interest, and S_t is updated using (2). The probability of a particular chain, $\{i_0, i_1, i_2, \ldots, i_r\}$, is given by the product of conditional binomial probabilities from (1) as

$$\Pr(I_1 = i_1 | S_0 = s_0, p_0) \Pr(I_2 = i_2 | S_1 = s_1, p_1) \times \cdots \Pr(I_r = i_r | S_{r-1} = s_{r-1}, p_{r-1}) \quad (5)$$
$$= \prod_{t=0}^{r-1} \binom{s_t}{i_{t+1}} p_t^{i_{t+1}} q_t^{s_t - i_{t+1}}.$$

The conditional expected value of I_{t+1} from (3) suggests the deterministic system of first-order difference equations

$$i_{t+1} = s_t p_t, \qquad s_{t+1} = s_t - i_{t+1}, \quad (6)$$

which can be analyzed as an approximation to the mean of the sample paths of the **stochastic process** $\{S_t, I_t\}_{t=0,1,\ldots}$. This system reduces to

$$s_t = s_{t-1} q_{t-1} = s_0 \prod_{l=0}^{t-1} q_l, \quad (7)$$

which is analyzed using methods from discrete mathematics (see, for example, [7] and [11]).

The Reed–Frost Model

History

The probabilistic form of the Reed–Frost epidemic model was introduced by the biostatistician Lowell J. Reed and the epidemiologist Wade Hampton Frost around 1930, as a teaching tool at Johns Hopkins University. It was developed as a mechanical model consisting of colored balls and wooden shoots. Although Reed and Frost never published their results, the work is described in articles and books by others (see [1, Chapters 14 and 18] and [2, Chapters 2 and 3]). An excellent description of the early Reed–Frost model is given by Fine [6]. The deterministic version of the Reed–Frost model has been traced back to the Russian epidemiologist P.D. En'ko, who used the model to analyze epidemic data in the 1880s (see [5]). The Reed–Frost version of the chain binomial and its extensions is used to study the dynamics of epidemics in small populations, such as families or day care centers, and to estimate transmission probabilities from epidemic data.

Formulation

In this case, S_t is the number of susceptible persons at the beginning of time interval t, and I_t is the number of persons who were newly infected at the beginning of time interval t. An infected person is infectious for exactly one time interval and then is removed; that is, becomes immune. Thus, a person infected at the beginning of time interval t will be infectious to others until the beginning of time interval $t + 1$. We let R_t be the number of removed persons at the beginning of time interval t, and then, by definition,

$$R_{t+1} = R_t + I_t = R_0 + \sum_{r=0}^{t} I_r. \qquad (8)$$

Since the population is closed, we have $S_t + I_t + R_t = n$ for all t. We let $p = 1 - q$ be the probability that any two specified people make sufficient contact in order to transmit the infection, if one is susceptible and the other infected, during one time interval. We note that p is a form of the **secondary attack rate**. We assume **random mixing**. Then, if during time interval t there are I_t infectives, the probability that a susceptible will escape being infected over the time interval is q^{I_t}, and the probability that they will become a new case at the beginning of time interval $t + 1$ is $1 - q^{I_t}$. Thus $q_t = q^{I_t}$, and substituting into (1) yields

$$\Pr(I_{t+1} = i_{t+1} | \dot{S}_t = s_t, I_t = i_t) \\ = \binom{s_t}{i_{t+1}} (1 - q^{i_t})^{i_{t+1}} q^{i_t(s_t - i_{t+1})}, \quad s_t \geq i_{t+1}. \qquad (9)$$

The epidemic process starts with $I_0 > 0$, and terminates at stopping time T, where

$$T = \inf_{t \geq 0} \{t : S_t I_t = 0\}. \qquad (10)$$

The possible chains for a population of size 4 with one initial infective - that is, $S_0 = 3$, $I_0 = 1$ - are shown in Table 1.

The probability of no epidemic is defined as the probability that there will be no further cases beyond the initial cases. This probability is

$$\Pr(I_1 = 0 | S_0 = s_0, p_0) = q^{i_0 s_0}. \qquad (11)$$

For example, if $S_0 = 10$, $I_0 = 1$, and $p = 0.05$, then the probability of no further cases beyond the initial case is 0.599. From (3), the conditional expected number of new cases in time interval t is $\mathrm{E}(I_{t+1} | s_t, p_t) = s_t(1 - q^{i_t})$. On the average, the epidemic process will not progress very far if the expected number of cases in the first generation is less than or equal to one; that is, $\mathrm{E}(I_1 | s_0, p_0) = s_0(1 - q^{i_0}) \leq 1$. In many cases, $i_0 = 1$, so that there will be few secondary cases if $s_0 p \leq 1$. Then, for example, if $S_0 = 10$, $I_0 = 1$, there will be few secondary cases if $p \leq 0.1$.

From (7), the deterministic counterpart of the Reed–Frost model is

$$s_t = s_0 q^{\sum_{l=0}^{t-1} i_l}, \qquad (12)$$

Table 1 Possible individual chains when $S_0 = 3$, $I_0 = 1$

Chain	Probability	Final size
$\{i_0, i_1, i_2, \ldots, i_T\}$		R_T
$\{1\}$	q^3	1
$\{1, 1\}$	$3pq^4$	2
$\{1, 1, 1\}$	$6p^2q^4$	3
$\{1, 2\}$	$3p^2q^3$	3
$\{1, 1, 1, 1\}$	$6p^3q^3$	4
$\{1, 1, 2\}$	$3p^3q^2$	4
$\{1, 2, 1\}$	$3p^3q(1 + q)$	4
$\{1, 3\}$	p^3	4

which has been thoroughly analyzed (see, for example [7] and [11]).

In some cases, the distribution of the total number of cases, R_T, is the random variable of interest. We let J be the random variable for the total number of cases in addition to the initial cases, so that $R_T = J + I_0$. If we let $S_0 = k$ and $I_0 = i$, then the probability of interest is

$$\Pr(J = j | S_0 = k, I_0 = i) = m_{ijk}, \qquad (13)$$

where $\sum_{j=0}^{k} m_{ijk} = 1$. Then, based on probability arguments (see, for example, [1]), we have the recursive expression

$$m_{ijk} = \binom{k}{j} m_{ijj} q^{(i+j)(k-j)}, \quad j < k, \quad (14)$$

and

$$m_{ikk} = 1 - \sum_{j=0}^{k-1} m_{ijk}. \qquad (15)$$

The Reed–Frost model has several extensions and special cases. If it is hypothesized that the probability that a susceptible becomes infected does not depend on the number of infectives that he or she is exposed to, then

$$p_t = \begin{cases} p, & \text{if } I_t > 0, \\ 0, & \text{if } I_t = 0. \end{cases} \qquad (16)$$

This model is known as the Greenwood model [8].

Longini & Koopman [12] modified the Reed–Frost model for the common case in which there is a constant source of infection from outside the population that does not depend on the number of infected persons in the population. We let $a_t = 1 - b_t$ be the probability that a susceptible person is infected during interval t due to contacts with infected persons outside the population, where

$$a_t > 0, \quad \text{if } t \le T,$$
$$a_t = 0, \quad \text{if } t > T,$$

and T is a stopping time. Then $p_t = 1 - b_t q^{I_t}$. If we let $B = \prod_{t=0}^{T} b_t$, then B is the probability that a person escapes infection from sources outside of the population over the entire period $[0, T]$. We then define CPI $= 1 - B$ as the community probability of infection. Longini & Koopman derive the probability mass function

$$m_{ijk} = \binom{k}{j} m_{ijj} B^{(k-j)} q^{(i+j)(k-j)}, \quad j < k. \quad (17)$$

Usually, $i = 0$ for this model. This model reduces to (14) when $B = 1$.

Another extension of the Reed–Frost model is for infectious diseases that do not confer immunity following infection. In this case, there is no removed state, so that $S_t + I_t = n$. Then, since $S_{t+1} = n - I_{t+1}$, the model is a discrete, one-dimensional Markov chain $\{I_t\}_{t=0,1,\ldots}$. The transition probabilities for this process are

$$\begin{aligned} &\Pr(I_{t+1} = i_{t+1} | I_t = i_t) \\ &\quad = \binom{n - i_t}{i_{t+1}} (1 - q^{i_t})^{i_{t+1}} q^{i_t(n - i_t - i_{t+1})}, \\ &\quad i_t + i_{t+1} \le n. \end{aligned} \qquad (18)$$

In this case, the disease in question can become "endemic". An interesting analytic question involves the study of the mean stopping time for the endemic process. From (6), the deterministic counterpart of this model is

$$i_{t+1} = (n - i_t)(1 - q^{i_t}), \qquad (19)$$

which is a form of the discrete logistic function. The stochastic behavior of (18) has been analyzed by Longini [10], and the dynamics of (19) have been analyzed by Cooke et al. [4].

There are many other extensions of the Reed–Frost model depending on the particular infectious disease being analyzed, but a further key extension is to allow the infectious period to extend over several time intervals. In this case $p_t = f(t, \theta, I_0, I_1, \ldots, I_t)$, and $\{S_t, I_t\}_{t=0,1,\ldots}$ is not a Markov chain. Special methods are used to analyze this model [14].

Inference

Data are usually in the form of observed chains, $\{i_0, i_1, \ldots, i_r\}$, for one or more populations, or final sizes, R_T, for more than one population. With respect to the former data form, suppose that we have N populations and let $\{i_{k0}, i_{k1}, \ldots, i_{kr}\}$ be the observed chain for the kth population. Then, from (5), the

likelihood function for estimating $p = 1 - q$ is

$$L(p) = \prod_{k=1}^{N} \prod_{t=0}^{r-1} \binom{s_{kt}}{i_{kt+1}} (1 - q^{i_{kt}})^{i_{kt+1}} q^{i_{kt}(s_{kt} - i_{kt+1})}. \tag{20}$$

For final value data, let a_{ijk} be the observed frequencies of the m_{ijk}, from (17); $i = 1, \dots, I$, $k = 1, \dots, K$, and $j = 1, \dots, k$. Then the likelihood function for estimating p and B is

$$L(p, B) = \prod_{i=1}^{I} \prod_{k=1}^{K} \prod_{j=0}^{k} m_{ijk}^{a_{ijk}}. \tag{21}$$

The logarithms of (20) and (21) are maximized using standard scoring routines (see, for example, Bailey [1], Becker [2], and Longini et al. [12, 13]) (*see* **Optimization and Nonlinear Equations**) or the corresponding **generalized linear model** (see Becker [2] and Haber et al. [9]). Extensions involve making both p and the CPI functions of covariates, such as age, level of susceptibility, or vaccination status (*see* **Vaccine Studies**). Bailey [1, Section 14.3] gives an example in which (20) is used to estimate $\hat{p} = 0.789 \pm 0.015$ (estimate ±1 standard error) for the household spread of measles among children. In the case of the household spread of influenza, Longini et al. [13] use (21) to estimate $\hat{p} = 0.260 \pm 0.030$ for persons with no prior immunity and $\hat{p} = 0.021 \pm 0.026$ for persons with some prior immunity. In addition, they estimate $\widehat{CPI} = 0.164 \pm 0.015$ and $\widehat{CPI} = 0.092 \pm 0.013$ for persons with no and some prior immunity, respectively.

Life Tables

The chain binomial model forms the statistical underpinnings of the **life table** (see Chiang [3, Chapter 10]). In this case, p_t simply depends on the time interval. Then S_t is the random variable of interest, which is formulated in terms of the interval survival probabilities $q_t = 1 - p_t$. Many important life table indices are functions of q_t. For example, the probability that an individual who starts in the cohort at time zero, is still alive at the end of time interval r, denoted q_{0r}, is $q_{0r} = \prod_{t=0}^{r} q_t$. The expected number alive at the beginning of time interval $r + 1$ is $\mathrm{E}(S_{r+1}) = s_0 q_{0r}$. This model is a discrete, one-dimensional Markov chain $\{S_t\}_{t=0,1,\dots}$. From (1) we see that the chain binomial model for S_t is simply

$$\begin{aligned}\Pr(S_{t+1} &= s_{t+1} | S_t = s_t) \\ &= \binom{s_t}{s_{t+1}} q_t^{s_{t+1}} p_t^{s_t - s_{t+1}}, \quad s_t \geq s_{t+1}. \end{aligned} \tag{22}$$

From (5), the probability of a particular chain $\{s_0, s_1, s_2, \dots, s_r\}$ is

$$\begin{aligned}\Pr(S_1 &= s_1 | S_0 = s_0) \Pr(S_2 = s_2 | S_1 = s_1) \\ &\times \cdots \Pr(S_r = s_r | S_{r-1} = s_{r-1}) \\ &= \prod_{t=0}^{r-1} \binom{s_t}{s_{t+1}} q_t^{s_{t+1}} p_t^{s_t - s_{t+1}}. \end{aligned} \tag{23}$$

For an observed chain $\{s_0, s_1, s_2, \dots, s_r\}$, (23) is the likelihood function for estimating $\{q_0, q_1, \dots, q_r\}$. The maximum likelihood estimators are

$$\hat{q}_t = s_{t+1}/s_t, \tag{24}$$

while the approximate variances, for large S_0, are

$$\mathrm{var}(\hat{q}_t) \approx p_t q_t / \mathrm{E}(S_t). \tag{25}$$

In addition, the $\hat{q}_t$ are unique, unbiased estimates of the q_t, and $\mathrm{cov}(\hat{q}_t, \hat{q}_l) = 0, t \neq l$. Estimators of most of the life table functions are based on the estimators $\hat{q}_t$.

References

[1] Bailey, N. (1975). *The Mathematical Theory of Infectious Diseases*, 2nd Ed. Griffin, London.

[2] Becker, N. (1989). *Analysis of Infectious Disease Data*. Chapman & Hall, New York.

[3] Chiang, C. (1984). *The Life Table and its Applications*. Krieger, Malabar.

[4] Cooke, K., Calef, D. & Level, E. (1977). Stability or chaos in discrete epidemic models, in *Nonlinear Systems and Applications - An International Conference*. Academic Press, New York.

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

[6] Fine, P. (1977). A commentary on the mechanical analogue to the Reed-Frost epidemic model, *American Journal of Epidemiology* **106**, 87-100.

[7] Frauenthal, J. (1980). *Mathematical Models in Epidemiology*. Springer-Verlag, Berlin.

[8] Greenwood, M. (1931). The statistical measure of infectiousness, *Journal of Hygiene* **31**, 336-351.

[9] Haber, M., Longini, I. & Cotsonis, G. (1988). Models for the statistical analysis of infectious disease data. *Biometrics* **44**, 163–173.

[10] Longini, I. (1980). A chain binomial model of endemicity, *Mathematical Biosciences* **50**, 85–93.

[11] Longini, I. (1986). The generalized discrete-time epidemic model with immunity: a synthesis, *Mathematical Biosciences* **82**, 19–41.

[12] Longini, I. & Koopman, J. (1982). Household and community transmission parameters from final distributions of infections in households, *Biometrics* **38**, 115–126.

[13] Longini, I., Koopman, J., Haber, M. & Cotsonis, G. (1988). Statistical inference for infectious diseases: risk-specified household and community transmission parameters, *American Journal of Epidemiology* **128**, 845–859.

[14] Saunders, I. (1980). An approximate maximum likelihood estimator for chain binomial models, *Australian Journal of Statistics* **22**, 307–316.

(*See also* **Epidemic Models, Deterministic; Epidemic Models, Stochastic; Infectious Disease Models**)

IRA M. LONGINI, JR

Chalmers, Thomas Clark

Born: December 17, 1917, in Forest Hills, New York.

Died: December 27, 1995, in Hanover, New Hampshire.

Thomas C. Chalmers, M.D., a leader in the design, conduct, and evaluation of **clinical trials**, was born in Forest Hills, New York, where his father was a physician in private practice. Following a tradition set by his father and grandfather, he graduated in 1943 from Columbia University College of Physicians and Surgeons. After additional training in medical research in New York and at the Thorndike Memorial Laboratories of Boston City Hospital, he entered private practice in Cambridge, Massachusetts, in 1947. He soon became concerned over the lack of knowledge on the efficacy of accepted medical therapies. Having learned about **randomization** from **Sir Austin Bradford Hill**, he applied this principle to a study of the treatment of infectious hepatitis among American soldiers in Japan during the Korean War. This study, a 2 × 2 randomized factorial study of diet and bed rest (*see* **Factorial Designs in Clinical Trials**), designed in 1951, included estimates of the required number of patients and an evaluation of ineligible patients, withdrawals, and **compliance**.

Deciding to devote his career to research and education, he was Chief of Medical Services at the Lemuel Shattuck Hospital in Boston (1955–1968), Assistant Director for Research and Education for the Veterans' Administration in Washington (1968–1970), Director of the Clinical Center at the National Institutes of Health in Bethesda (1970–1973), and President and Dean of the Mount Sinai Medical Center and School of Medicine in New York City (1973–1983).

Dr Chalmers returned to Boston in 1983. Over the next 10 years he was on the faculty of the Harvard School of Public Health and Tufts University School of Medicine, was appointed a Distinguished Professor at the Boston Veterans' Administration Medical Center, and was a member and Chairman of the Board of Trustees of the Dartmouth Hitchcock Medical Center. In 1992, at age 75, he co-founded Meta-Works, Inc., a **meta-analysis** consulting company and moved to Lebanon, New Hampshire. He continued to teach in both Boston and New York and was actively involved in numerous meta-analytic studies.

Throughout his career he was a fervent advocate of randomized controlled trials in all areas of medicine and the education of students and physicians in the skills needed to evaluate these trials. His belief in the ethical need for randomization [6] (*see* **Ethics of Randomized Trials**) led to his recommendation to "begin randomization with the first patient" [1]. A corollary was the belief that developing trends should not be known by investigators during the conduct of the trial, but should be monitored by an independent policy advisory committee (*see* **Data and Safety Monitoring Boards**). In subsequent years he was a member (and frequently chairman) of Policy or Data Safety and Monitoring Boards for numerous multicenter clinical trials.

Dr Chalmers moved to Mount Sinai in 1973 because he wanted to influence the education of medical students, and to make both students and faculty aware of the need for properly conducted clinical trials. He became concerned that clinical trials were being conducted with insufficient sample sizes and in a review published in the *New England Journal of*