Chapter 7
Some Statistical Issues

7.1 Models and Parameters

All the models presented in the previous chapters are parametric. They belong to different types and serve different purposes.

Probabilistic models are introduced and discussed in Chaps. 2 and 3. Part of their role in the study of infectious diseases is to formulate assumptions regarding disease progression within an infected host. Examples include the parametric life distribution models characterized by shapes of hazard functions in Chap. 2, as well as the process of infectious contacts from the viewpoint of an infected individual through counting processes in Chap. 3. These probabilistic models, especially the distribution models for random counts and counting processes, are also statistical models that take into account the data-generating process that will be further discussed in this chapter.

Phenomenological population models, both in stochastic and in deterministic frameworks, are the main focus in Chaps. 4–6. They are based on conceptual assumptions regarding the population and the interface among the agent, the host, and the environment. Section 4.4 has provided some detailed discussions regarding these assumptions applied to the initial phase of an outbreak. However, phenomenological population models often carry tacit assumptions at the individual level. For example, deterministic transmission models that can be represented by systems of ordinary differential equations implicitly assume that infected individuals pass each stage of the natural history with exponentially distributed durations. The stochastic models with a Markov structure make the same assumptions.

Hidden assumptions at the level of individuals determine certain crucial epidemiological characteristics as well as the effectiveness of certain disease control measures in a phenomenological way. For example, the relationship between the distributions of the infectious periods and the probabilities of invasion and extinction (Sect. 4.2); the relationship between the distributions of the infectious periods (as

P. Yan, G. Chowell, *Quantitative Methods for Investigating Infectious Disease Outbreaks*, Texts in Applied Mathematics 70,
https://doi.org/10.1007/978-3-030-21923-9_7

well as the latent periods) and the initial growth (4.43); the assumption of the exponentially distributed infectious periods in the SIR model as a primary feature with respect to the peak prevalence of infected individuals and some important preserved quantities (Sect. 5.3.3); the prevalence of individuals in each class of the SEIRS model (5.67) with expressions of $[x(\infty), \epsilon(\infty), y(\infty), z(\infty)]$ under endemic equilibrium and their special cases such as (5.71) in Sect. 5.5.2; the expression of the controlled reproduction number R_c (6.4) in Sect. 6.2.1; and the effects of the distributions of the latent and infectious periods on certain control measures (Sect. 6.4).

There is also a different kind of phenomenological population models with relatively simple forms and without assumptions regarding the agent, the host, the environment, and their interactions in the population. They describe data in a phenomenological way and are often useful to answer some key public health questions during an outbreak investigation. These are the growth curve models. Later in Chap. 8, we shall see some applications of these models to real outbreak data.

The most important function of models are to order our thoughts and to sharpen vague intuitive notions. Whatever their types are, they are connected to the formulation of the research questions and objectives of the subject matter. Different questions and objectives require differently formulated models.

Even the same model can be parameterized differently for different research questions. For example, a simple logistic function has many different expressions such as (4.59) and (5.14). The logistic function may be written as

$$F(t) = \frac{K}{1 + e^{-\rho(t-\alpha)}}.$$

This function can be considered as a *descriptive* model to fit disease incidence data, either cumulatively, or incidence numbers of new occurrences (e.g., daily, weekly, etc.) The three parameters (ρ, α, K) are directly and indirectly connected to important public health questions during a disease outbreak, such as: "when do we expect the outbreak to peak?", "how long do we expect the outbreak to last?" and "how big is the outbreak going to be?" This is because the parameter ρ represents the initial growth of a sigmoid growth function; α represents the inflexion point at which $F'(t)$ arrives at the maximum value as well as when the outbreak is at its midpoint $F(\alpha) = K/2$; and K represents the asymptotic limit $K = \lim_{t\to\infty} F(t)$. However, this model provides little understanding of the process such as the disease transmission process.

On the other hand, the logistic function expressed as

$$F(t) = \frac{mi_0\,(R_0 - 1)}{i_0 R_0 + (m\,(R_0 - 1) - i_0 R_0)\,e^{-(R_0-1)\gamma t}}$$

in (5.14) has four parameters: (R_0, γ, m) and the initial condition $i_0 = F(0)$. In fact, it is a *mechanistic* model because all these parameters are associated with scientific hypotheses about the transmission dynamics with the SIS structure, such as the

basic reproduction number R_0, the mean duration of the (exponentially distributed) infectious period γ^{-1}, and the population size m. This expression is used to describe the prevalence of the number of individuals who are "currently" infectious at time t. Statistically speaking, only three out of the four parameters are identifiable from data, because $\rho = (R_0 - 1)\gamma$ can be regarded as a single scale parameter of time t as the initial growth rate and $K = m(1 - 1/R_0)$ is the asymptotic limit. The identifiable parameters are (ρ, i_0, K).

7.1.1 Statistical Models

In their book *Generalized Linear Models,* McCullagh and Nelder (1983) partitioned the model into three components: (1) the random component, (2) the systematic component, and (3) the link function. We adopt the same terminology when combining statistical models with disease transmission models for analyses of outbreak investigation data.

The random component models the data-generating process through probability distributions, denoted here as $f(y; \theta)$. They are the foundation of statistical inference for estimation and testing hypotheses. The discrete distributions and counting processes in Chap. 3 are important statistical models to represent the data-generating process of random counts as realizations of the underlying stochastic processes that generate disease outbreak data that form time-series composed of non-negative integers. The continuous lifetime distributions in Sect. 2.2 are statistical models if the questions under investigation are regarding estimation and testing hypothesis of time-to-event, such as the incubation period defined as the time elapsed from infection to the onset of clinical symptoms, time to recovery, time to death, etc., based on longitudinally observed or retrospectively assessed data. (Under a different context, these continuous lifetime distributions are implicitly built into the phenomenological population models, such as the exponential distribution of the infectious periods.)

The systematic component describes the systematic effects of interest, within the data-generating mechanism. In classic linear regression analysis, this component is formulated through a set of covariates $\underline{x} = (x_1, \ldots, x_p)$ in a linear form $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$. When we say a model is linear, we mean linearity in parameters $\underline{\beta}$, not the covariates $\underline{x}$.

If a random sample $(y_1, \ldots, y_n)$ arises from independent observations $y_i \sim f(y_i; \theta_i)$, $i = 1, \ldots, n$, the reduction of dimension of the parameter space where $p \ll n$ is a mapping

$$(\theta_i : i = 1, \ldots, n) \mapsto \left(\beta_j : j = 1, \ldots, p\right)$$

through the covariates $\underline{x}$ by the linear function through the link function $h(\theta)$ so that

$$h(\theta) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{7.1}$$

Such a model is a generalized linear model (McCullagh and Nelder 1983). Data arising from random counts are often fitted to one of the discrete distributions in Chap. 3 associated with a positive parameter $\theta > 0$ and a logarithm link function $h(\theta) = \log\theta$. The corresponding generalized linear models are the log-linear models. Binary data are often fitted to the binomial, geometric, or negative-binomial distributions. These distributions are often associated with a proportion parameter $0 < \theta < 1$. A logit link function $h(\theta) = \log \frac{\theta}{1-\theta}$ is often chosen, which gives rise to the logistic regression models. Continuous lifetime data are often fitted with the lifetime distributions in Chap. 2 that may be associated with a log-linear model or a proportional hazard model. The latter, in a broader sense, can also be viewed as a generalized linear model.

The systematic components in models for infectious disease outbreak investigations are typically nonlinear functions with respect to their parameters. These are phenomenological population models. Some of them have explicit analytic forms. We shall see many examples in Chap. 8. Others are implicit, including the transmission dynamic models expressed as a system of differential equations discussed in the proceeding chapters. These nonlinear models create additional challenges in computation algorithms, such as the optimization algorithms in the search for the maximum likelihood estimates or the least square estimation. They are highly sensitive to the initial parameter estimates in those algorithms. In the special case of the generalized linear models, initial estimates are not necessary. Therefore, it is important to carefully evaluate the values of the log-likelihood or the sum of square errors (SSE) upon convergence over a wide range of possible initial estimates.

It is important to recognize that disease transmission is only part of the data generating process, and many of the disease transmission models do not directly predict observable events as reflected by data. Other data generating mechanisms, such as case-definition, how data are organized and reported, length-biasedness, retrospective ascertainment of time of events, reporting delays, among many other issues, also need to be described using statistical models. The link function connects these two components and links them to the distribution of the data.

Lindsey (2001) provides comprehensive discussions on nonlinear models in medical statistics.

7.1.2 Fitting Models to Data and Model Criticism

In fitting a statistical model to data, the information in the data is split into two parts, one to assess the unknown parameters, and the other for model criticism. Both assessment of parameters and model criticism are equally important aspects in statistical inference.

Sprott (2000) points out that the sample information is divided into "Likelihood θ" and "Model f" through the factorization of a likelihood function according to the minimal sufficient division or the maximal ancillary division. A classic example

is the factorization of the Poisson likelihood. Consider an i.i.d. random sample $(y_1, \ldots, y_n)$ from the Poisson distribution with mean value μ, then $t = \sum_{i=1}^{n} y_i$ is Poisson distributed with mean $n\mu$. The joint distribution is

$$f(y_1, \ldots, y_n; \mu) = \prod_{i=1}^{n} \frac{\mu^{y_i} e^{-\mu}}{y_i!} = f(t; \mu) f(y_1, \ldots, y_n | t),$$

where the first factor

$$f(t; \mu) = \frac{e^{-n\mu} (n\mu)^t}{t!}$$

is the likelihood function of μ as represented by the minimal sufficient statistics $t = \sum_{i=1}^{n} y_i$, with one degree of freedom, for the assessment of the parameter μ. The second factor

$$f(y_1, \ldots, y_n | t) = \frac{t!}{\prod_{i=1}^{n} y_i!} \prod_{i=1}^{n} \left(\frac{1}{n}\right)^{y_i}$$

is a multinomial distribution for the residual, with $n - 1$ degrees of freedom. If data cast doubt on this multinomial distribution, they equally cast doubt on the assumed Poisson model.

Residuals are often associated with regression models. If a regression model such as (7.1) involves p unknown parameters, fitting such a model to data of sample size $n \gg p$ yields a residual consisting of $n - p$ degrees of freedom. Residual analyses in the form of goodness-of-fit play a crucial role for model criticism on three levels. At the first level, large residual values indicate a lack of fit. This is often used in conjunction with the testing of hypothesis $H_0 : \beta_j = 0$, $j = 1, \ldots, p$. It is the criticism of a sub-model within a larger model to single out important covariates x_j that are statistically significant. The second level is the testing against some fundamental assumptions in these models. For example, the logistic regression models are often associated with the assumption of proportional odds ratios and the proportional hazard model assumes proportional hazard functions. In good statistical practice, one always needs to take due diligence to test against these assumptions whenever these models are applied. Various testing statistics are available in the literature, such as the Z statistics to test against the proportional hazard assumption in almost every survival analysis textbook. The third level is the testing against the probability distributions, for instance, if data used in a logistic regression model arise from a binomial distribution or if data used in a proportional hazard model arise from a Weibull distribution. This may be optional if the primary interest is in the parameters β_j, $j = 1, \ldots, p$ while treating the underlying distribution as a nuisance parameter problem.

7.1.3 Fitting Phenomenological Population Models to Time-Series Data

Fitting phenomenological population models to time-series data collected during an epidemic, often called *curve-fitting*, is commonly practiced for the purposes of parameter estimation and prediction (Smirnova and Chowell 2017). These models can be mechanical disease transmission models with strong assumptions on the transmission process, or other forms of simpler, but nonetheless highly nonlinear descriptive models for the data generating processes. We may regard fitting phenomenological population models to disease outbreak investigation data as nesting a phenomenological population model inside the systematic component of a statistical model in the form of *generalized nonlinear regression*.

In general, we denote the time series of T longitudinal observations by

$$\underline{y} = (y_1, y_2, \ldots, y_T)$$

where $t = 1, 2, \ldots, T$ are discrete or time units, such as daily, weekly, etc., typical in disease outbreak investigations. We regard these data as realizations of random counts $Y(t)$ manifested through a dynamic system, such as in the SEIR system as discussed in Sect. 5.4, appropriately grouped into discrete time units as Y_t.

The systematic component of the model is denoted by $\mu(t; \Theta)$, which is a nonlinear function specified by a set of parameters $\Theta = (\theta_1, \ldots, \theta_m)$. The marginal distribution for Y_t may be only specified to its first moment $E[Y_t] = \mu(t; \Theta)$, or the first two moments, both as functions of Θ, or fully specified such as the Poisson distribution $Poisson(\mu(t; \Theta))$.

Parameter Estimation

We consider the quasi-likelihood estimating equations for the generalized linear models (McCullagh and Nelder 1983)

$$\sum_{t=1}^{T} \frac{\partial \mu(\mu; \Theta)}{\partial \theta_j} \frac{y_t - \mu(t; \Theta)}{V[Y_t; \Theta]} = 0, \ \ j = 1, \ldots, m \tag{7.2}$$

are still valid, where $V[Y_t; \Theta]$ is the variance of Y_t. More generally, the denominator $V[Y_t; \Theta]$ may also involve a correlation matrix, which relaxes the independency assumption among Y_t, $t = 1, 2, \ldots, T$, which are called the generalized estimating equations (Liang and Zeger 1986). The generalized estimating equations can be also applied to zero-mean martingales in Sect. 3.3.2 (Godambe and Heyde 1987).

One of the common choices is assuming $V[Y_t; \Theta] = \alpha \mu(t; \Theta)$, where $\alpha > 0$ is a scalar parameter. This variance form may well approximate variance structures such as $Var[Y_t] = E[Y_t] + E[Y_t]^2/\kappa$ when $E[Y_{t_i}]^2/\kappa$ do not vary greatly with t. The

variance structure of the negative binomial distribution (3.20) follows this form, as well as the mixed Poisson distribution in which the mixing distribution is inverse-Gaussian (see Chap. 3). In this case,

$$\sum_{t=1}^{T} \frac{\partial \mu(\mu;\Theta)}{\partial \theta_j} \frac{y_t - \mu(t;\Theta)}{\alpha \mu(t;\Theta)} = 0, \ \ j = 1, \ldots, m$$

which may be useful to handle data with overdispersion ($\alpha > 1$).

The estimating equations given by (7.2) take the form of the score functions of the likelihood functions of many well-known distributions, such as the Gaussian, Poisson, binomial, among many others, provided that the distributions are correctly specified. Without specifying the distribution, they are unbiased estimating equations that lead to asymptotically unbiased point estimates regardless of any misspecification of the variance–covariance structure. If the variance–covariance structure is correctly specified, they lead to the variance estimation of the parameter estimates. However, the estimated variances of the parameter estimates will be in error with misspecification of the variance–covariance structure.

These estimating equations are usually associated with generalized linear models. In contrast, phenomenological models are nonlinear, and in some cases, are implicitly defined through differential equations without analytic solutions. This poses computational challenges because $\frac{\partial \mu(\mu;\Theta)}{\partial \theta_j}$ is either complicated or prohibitive.

In the following two special cases, optimization algorithms can be employed without the evaluation of $\frac{\partial \mu(\mu;\Theta)}{\partial \theta_j}$.

One is the EE given by

$$\sum_{t=1}^{T} \frac{\partial \mu(\mu;\Theta)}{\partial \theta_j} \frac{y_t - \mu(t;\Theta)}{\mu(t;\Theta)} = 0, \ \ j = 1, \ldots, m. \tag{7.3}$$

It assumes that $Var[Y_t] = E[Y_t] = \mu(t;\Theta)$. As mentioned in Chap. 3, among power series distributions used for random counts, $Var[Y_t] = E[Y_t]$ characterizes the Poisson distribution (Kosambi 1949). In fact, (7.3) is the score function of the likelihood function assuming that $\underline{y} = (y_1, y_2, \ldots, y_T)$ are realizations of independent Poisson random counts. The log-likelihood function is

$$l(\Theta) = \sum_{t=1}^{T} \left[y_t \log \mu(t;\Theta) - \mu(t;\Theta) \right]. \tag{7.4}$$

Therefore, solving (7.3) is equivalent to maximizing (7.4). The maximum likelihood estimate can be expressed as

$$\widehat{\Theta} = \arg\max \sum_{t=1}^{T} \left[y_t \log \mu(t;\Theta) - \mu(t;\Theta) \right]. \tag{7.5}$$

One can use numerical optimization methods in MatLab or R (R Core Team). In R, a general-purpose optimization method based on the downhill simplex method (Nelder-Mead) or the quasi-Newton algorithms are readily available. However, for a nonlinear function $\mu(t;\Theta)$, the optimization algorithms to maximize the log-likelihood are highly sensitive to the initial parameter estimates, which may lead to local maxima. It is important to carefully evaluate the values of the log-likelihood upon convergence over a wide range of possible initial estimates.

An alternative method is the least square estimate, achieved by searching for the set of parameters $\widehat{\Theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_m)$ that minimizes the sum of squared differences between the observed data and the corresponding model solution denoted by$\mu(t;\Theta)$, $t = 1, 2, \ldots, T$. That is, the objective function is given by

$$\widehat{\Theta} = \arg\min \sum_{t=1}^{T} [y_t - \mu(t;\Theta)]^2 . \tag{7.6}$$

This is equivalent to solving the EE

$$\sum_{t=1}^{T} \frac{\partial \mu(\mu;\Theta)}{\partial \theta_j} [y_t - \mu(t;\Theta)] = 0, \;\; j = 1, \ldots, m \tag{7.7}$$

assuming that $Var[Y_t]$ is independent of the mean and does not involve the set of parameters Θ. Although this method does not assume any specific distribution for Y_t except for its first moment $E[Y_t] = \mu(t;\Theta)$, the least square method is equivalent to the maximum likelihood estimation if data Y_t are Gaussian distributed. If the random counts are highly skewed, the least square method may not perform well. In Matlab (The Mathworks, Inc.), two numerical optimization methods are available to solve the nonlinear least squares problem: The trust-region reflective algorithm and the Levenberg-Marquardt algorithm. As with the maximum likelihood estimates, the optimization algorithms to minimize the sum of square errors (SSE) are highly sensitive to the initial parameter estimates, which may lead to local minima. It is important to carefully evaluate the values of the SSE upon convergence over a wide range of possible initial estimates.

Uncertainty in Estimated Parameters

The Likelihood Surface and the Likelihood Ratio Statistics The relative likelihood, $R(\Theta; \underline{y})$, is defined by

$$0 < R(\Theta; \underline{y}) = \frac{L(\Theta; \underline{y})}{\sup_{\Theta} L(\Theta; \underline{y})} = \frac{L(\Theta; \underline{y})}{L(\widehat{\Theta}; \underline{y})} \leq 1$$

where $L(\Theta; \underline{y})$ is the likelihood function of Θ given data $\underline{y}$. It ranks the parameters Θ over the scale from 0 to 1, and the maximum likelihood estimate $\widehat{\Theta}$ is the most plausible value of Θ in that it makes the observed data $\underline{y}$ most probable (Sprott 2000; Kalbfleisch 1985). One could also define the likelihood region such that $R(\Theta; \underline{y}) \geq \varsigma$ where $0 < \varsigma < 1$ as plausible parameter values. By varying the threshold ς, one can define such things as "very plausible," "plausible," or "implausible." These concepts give rise to the likelihood ratio statistics that can be used to construct confidence intervals and test hypotheses. More importantly, visualization of the contour of the likelihood surface in the neighborhood reveals the amount of information the data contain with respect to each parameter. This is feasible when the number of parameters in Θ is less than or equal to 2.

It is more convenient to work on the logarithmic scale. The relative log-likelihood is defined by

$$-\infty < r(\Theta) = l(\Theta) - l(\widehat{\Theta}) \leq 0$$

where $l(\Theta) = \log L(\Theta; \underline{y})$.

A likelihood region is defined on the parameter space such that $R(\Theta; \underline{y}) \geq \varsigma$ for a selected value $0 < \varsigma < 1$. Calculating the $100(1-p)\%$ confidence regions based on the likelihood ratio can be done directly by selecting the value for the likelihood region , that is $r(\Theta) = l(\Theta) - l(\widehat{\Theta}) \geq \log \varsigma$, so that the coverage probability

$$CP \approx \Pr(\chi^2_{\text{df}} \leq -2 \log \varsigma) = 1 - p.$$

The 95% confidence interval for a single parameter θ can be derived by choosing $\varsigma = 0.147$ such that

$$CP \approx \Pr(\chi^2_{(1)} \leq -2 \log 0.147) = 0.95. \tag{7.8}$$

The left panel of Fig. 7.1 shows the 95% confidence interval for the mean value μ of the Poisson distribution a small sample of random count data based on (7.8). The most plausible value is $\widehat{\mu} = 2.2$ and the plausible range is $1.4046 \leq \mu \leq 3.2505$ such that $r(\mu) = l(\mu) - l(\widehat{\mu}) \geq \log 0.147 = -1.9173$.

The 95% confidence region for two parameters (α, β) is the contour of $r(\alpha, \beta)$ by choosing $\varsigma = 0.05$ such that

$$CP \approx \Pr(\chi^2_{(2)} \leq -2 \log 0.05) = 0.95. \tag{7.9}$$

The right panel of Fig. 7.1 shows the 95% joint confidence region of two parameters: the median parameter λ^{-1} and the 95th percentile t_{95} of the log-logistic distribution for the incubation period fitted to a small sample of data collected on people with SARS symptoms, using (7.9).

In nonlinear models, parameters are often inter-related in complex ways. In a two-parameter setting, it is more likely to encounter "banana" log-likelihood

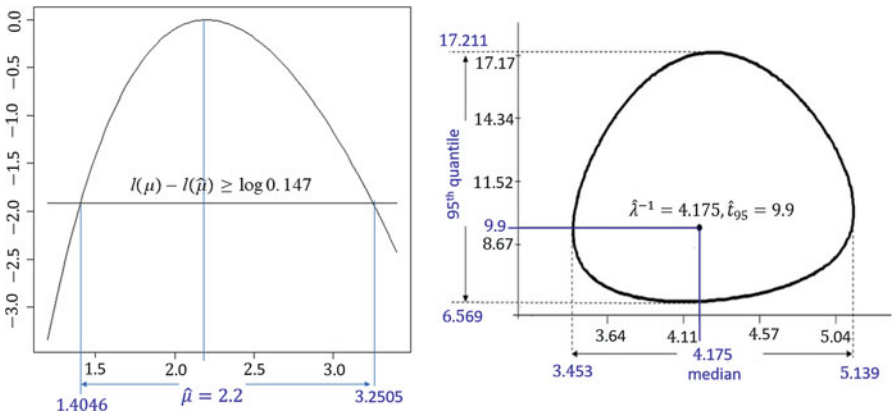


Fig. 7.1 Left: the relative log-likelihood of the Poisson distribution from the sample of random counts: {0, 5, 2, 3, 2, 3, 1, 0, 2, 4} with m.l.e. $\widehat{\mu} = 2.2$ (1.4, 3.25); Right: the joint of likelihood region for the median and the 95th percentile of the log-logistic distribution for the incubation distribution based on a small sample of SARS patients

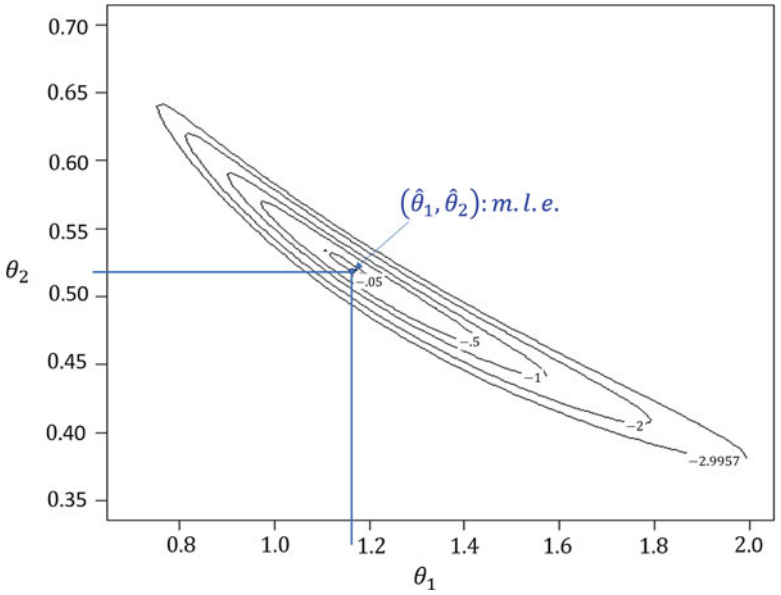


Fig. 7.2 A banana shaped log-likelihood contour showing the correlation of two parameters θ_1 and θ_2 as suggested by data, where $\widehat{\theta_1}$ and $\widehat{\theta_2}$ are the maximum likelihood estimates. The outmost contour line corresponds to the likelihood region with 95% coverage probability

contours as schematically illustrated in Fig. 7.2. The 95% joint likelihood region is the outmost contour, which given the marginal 95% confidence limits for θ_1 between 0.65 and 2.0, and the marginal 95% confidence limits for θ_2 between

0.38 and 0.65. However, it is equally plausible to have these two combinations: ($\theta_1 = 0.7, \theta_2 = 0.65$) or ($\theta_1 = 2.0, \theta_2 = 0.38$). These two pairs may represent very different epidemic scenarios, but they are equally accepted by data. This poses an *identifiability* problem. On the other hand, the likelihood contour also rules out implausible scenarios, such as ($\theta_1 = 1.6, \theta_2 = 0.6$), even though both values are well within their 95% confidence limits.

In the case of more than two parameters, it is still worthwhile to visualize the likelihood surface either as a 3-D function or cross-sectional log-likelihood contours. These will provide more reliable precision intervals than marginal confidence intervals for each parameter, reveal correlation among parameters, and provide better ways to communicate uncertainty. However, these are very time-consuming.

With respect to the testing of the hypothesis $H_0 : \Theta = \Theta_0$, the likelihood ratio statistics is given by

$$D = -2r(\Theta_0) = -2\left[l(\Theta_0) - l(\widehat{\Theta})\right].$$

The significant level is

$$SL = \Pr(D \geq D_{\text{obs}} | H_0 \text{ is true}) \approx \Pr(\chi^2_{\text{df}} \geq D_{\text{obs}}) \tag{7.10}$$

where the degree of freedom, df, is equal to the number of functionally independent parameters in the model. In testing a null hypothesis for a single parameter $H_0 : \theta = \theta_0$, the degree of freedom is df = 1. In testing a null hypothesis for two parameters $H_0 : \alpha = \alpha_0$ and $\beta = \beta_0$, df = 2.

The marginal 95% confidence interval for a single parameter in the presence of many other parameters can be also derived by numerically inverting the testing of null hypothesis $H_0 : \theta = \theta_0$ and calculate the significance level at different θ_0 using the $\chi^2_{(1)}$ approximation in (7.10), until $SL = 0.05$. In the case of m parameters in Θ, it involves two steps:

1. under the null hypothesis, fixing $\theta = \theta_0$, and conduct a maximum likelihood estimation for the remaining $m - 1$ parameters, denoted by Θ^*, and evaluate the value of the log-likelihood $l(\widehat{\Theta^*}|\theta = \theta_0)$;
2. under the alternative hypothesis, conduct a maximum likelihood estimation of all the parameters in Θ.

The likelihood ratio statistics is

$$D = -2\left[l(\widehat{\Theta^*}|\theta = \theta_0) - l(\widehat{\Theta})\right] \tag{7.11}$$

approximated by the $\chi^2_{(1)}$ distribution.

Assessing Uncertainty in the Estimated Parameters Through Bootstrapping The likelihood approach applies only when the joint distribution of $\underline{Y} = (Y_t, Y_{t,} \ldots, Y_T)$ is completely and correctly specified. Other approaches based on the asymptotic properties of the generalized estimating equations are not practical

because of the nonlinear functions employed in $\mu(\mu; \Theta)$ that make the calculations for $\partial\mu(\mu; \Theta)/\partial\theta_j$ prohibitive.

The general bootstrap method (Efron and Tibshirani 1994) based on assumed variance structures to assess uncertainty in the estimated parameters is useful. It is widely applied in quantifying parameter uncertainty and constructing confidence intervals in mathematical modeling studies (see, e.g., Chowell et al. 2006a,b). In this method, multiple observations are repeatedly sampled from the best-fit model by assuming that each point of the time series follows a specific distribution, typically a Poisson or a negative binomial distribution, centered on the estimated mean at that time point. The step-by-step algorithm to quantify parameter uncertainty follows:

1. Derive the parameter estimates $\widehat{\Theta} = (\widehat{\theta_1}, \ldots, \widehat{\theta_m})$ by fitting the model to the time series data $\underline{y} = (y_1, y_2, \ldots, y_T)$ to obtain the best-fitted model $\mu(t; \widehat{\Theta})$, $t = 1, \ldots, T$.
2. Generate replicated simulated datasets through re-sampling. To do so, we first use the best-fit model $\mu(t; \widehat{\Theta})$ to calculate the expected values of the time-series.
3. Each simulated data set is generated by random numbers assuming Poisson or negative binomial error structures based on the expected values. Specifically, for time t, a random number y_t^* with mean value $\mu(t; \widehat{\Theta})$ is drawn from a Poisson or a negative binomial distribution. This forms a simulated time series $\underline{y}^* = (y_1^*, y_2^* \ldots, y_T^*)$. Repeating this simulation s times, we obtain s replicated simulated datasets, denoted by $\underline{y}^*_{(1)}, \underline{y}^*_{(2)}, \ldots, \underline{y}^*_{(s)}$.
4. Re-estimate parameters for each of the s simulated realizations. Estimated parameter sets given by $\widehat{\Theta}_i$, $i = 1, \ldots, s$.
5. Using the set of re-estimated parameters $\widehat{\Theta}_i$, $i = 1, \ldots, s$, it is possible to characterize their empirical distributions, correlations, and construct confidence intervals.

In addition, since infectious disease outbreaks are not repeatable under identical conditions (in the sense of a designed random experiment), the computer-based re-sampling provides a virtual experiment with the resulting uncertainty around the model fit given by $\mu(t; \widehat{\Theta}_1)$, $\mu(t; \widehat{\Theta}_2), \ldots, \mu(t; \widehat{\Theta}_s)$. This is very useful for assessing uncertainty of key disease transmission parameters such as the basic reproduction number (Anderson and May 1991; Diekmann et al. 1990; van den Driessche and Watmough 2002). This parameter is a function of several parameters that characterize the transmission and control process, e.g., transmission rates and infectious periods of the epidemiological classes that contribute to new infections. Uncertainty around this key parameter, estimated through re-sampling, can be viewed through such a perspective.

Residual Analysis

While residuals (differences between model fit and observations), $r_t = y_t - \mu(t; \widehat{\Theta})$, can inform systematic deviations of the model fit to the data, it is also possible to

quantify the error of the model fit to the data using performance metrics (Kuhn and Johnson 2013). These metrics are also useful to quantify the error associated with forecasts. A widely used performance metric is the mean square error (MSE), which is given by

$$MSE = \frac{1}{T}\sum_{t=1}^{T}\left[y_t - \mu(t;\widehat{\Theta})\right]^2. \tag{7.12}$$

Another commonly used residual is the Pearson residual, defined as

$$r_t^{(P)} = \frac{y_t - \mu(t;\widehat{\Theta})}{\widehat{V}_t^{1/2}}$$

where $\mu(t;\widehat{\Theta})$ is the expected value of Y_t and $\widehat{V}_t$ is the estimated variance $Var(Y_t)$. In particular, if the variance structure corresponds to (7.3), then

$$r_t^{(P)} = \frac{y_t - \mu(t;\widehat{\Theta})}{\sqrt{\mu(t;\widehat{\Theta})}}.$$

The performance metric is the weighted mean square error (WMSE)

$$WMSE = \sum_{t=1}^{T}\frac{\left[y_{t_i} - \mu(t;\widehat{\Theta})\right]^2}{\mu(t;\widehat{\Theta})} = \sum_{t=1}^{T}\frac{(O-E)^2}{E} \tag{7.13}$$

where O stands for "Observed" and E stands for "Expected." Although (7.13) assumes the Poisson variance structure, minimizing the weighted sum of squares $\sum_{t=1}^{T}\frac{\left[y_{t_i}-\mu(t;\widehat{\Theta})\right]^2}{\mu(t;\widehat{\Theta})}$ does not correspond to the unbiased estimating equation (7.3) and hence it does not correspond to the maximum likelihood estimation by maximizing (7.4). In fact, the weighted least square estimation by minimizing $\sum_{t=1}^{T}\frac{\left[y_{t_i}-\mu(t;\widehat{\Theta})\right]^2}{\mu(t;\widehat{\Theta})}$ would have yielded the equation

$$\sum_{t=1}^{T}\frac{\partial\mu(\mu;\Theta)}{\partial\theta_j}\frac{y_t-\mu(t;\Theta)}{\mu(t;\Theta)} + \frac{1}{2}\sum_{t=1}^{T}\frac{\partial\mu(\mu;\Theta)}{\partial\theta_j}\left[\frac{y_t-\mu(t;\Theta)}{\mu(t;\Theta)}\right]^2 = 0$$

of which the first term is the left hand of (7.3). The weighted least square estimates are asymptotically biased because (7.3) gives the asymptotically unbiased estimates. In addition to the above arguments, the distributions of the residuals defined by $r_t^{(P)}$ are skewed for non-Gaussian distributions. Consequently, using WMSE as the performance measure for Poisson distributed random counts may not be the best choice.

An improved version of (7.13) is based on the Anscombe residuals (Anscombe 1953). Descriptions for this residual measure can be found in the book by McCullagh and Nelder (1983). For Poisson distributed random counts, the residuals are defined as

$$r_t^{(A)} = \frac{\frac{3}{2}\left[y_t^{2/3} - \mu(t;\Theta)^{2/3}\right]}{\mu(t;\Theta)^{1/6}},$$

and the performance metric is

$$ANSCOMBE = \sum_{t=1}^{T} \left(\frac{\frac{3}{2}\left[y_t^{2/3} - \mu(t;\Theta)^{2/3}\right]}{\mu(t;\Theta)^{1/6}} \right)^2. \tag{7.14}$$

The Anscombe residuals $r_t^{(A)}$ are approximately Gaussian distributed.

Model Criticism

In fitting models to time-series data, one step is to choose an appropriate model for the systematic component. For example, if the time-series exhibits a trend that resembles a sigmoid curve, one may consider a simple growth function, such as a logistic function, and conduct residual analyses and hypothesis tests against some versions of generalized logistic models to examine whether the model captures the general characteristics of the observed time series.

On the other hand, depending on the data fitting methods, there are subtle assumptions on the random component. If the data fitting method is based on a likelihood function, the full specification of the distribution of data must be given. Model criticism will be something like: do observed random counts as a finite time-series arise as an independent sample of, say, a negative-binomial distribution, with its mean values further modelled by a deterministic function of t? Even with empirical curve-fitting, such as the least square method, statistical assumptions such as independency among data points, the relationship between the variance and the mean are still made. These are all subject to criticism in the light of data.

Data usually admit more than one model. Even when a specific model is preferred, for scientific or practical reasons, alternative models also need to be taken into consideration.

In modelling disease outbreak data, it is very common that the available data cannot identify all the parameters involved. What we mean by "not identifiable" is that, in a multiple parameter setting, more than one set of combinations of parameters manifest the same expected value that fits well to data. Since the model is split into random, systematic, and link components, the problem of *identifiability* carries over to the identifiability of these components. This makes model criticism more complex.

7.2 Data

7.2.1 Some Features of Infectious Disease Outbreak Data

A striking feature of data collected during an infectious disease outbreak is that they do not arise from designed experiments, which are either impossible or unethical in the context of epidemics among humans.

Data are not repeatable. Outbreaks of the same disease do not start with identical conditions. Moreover, environmental and behavioral changes occur, and pathogens mutate. Even with data collected as a long sequence of time-series, or data collected from multiple data sources, or even Big Data, one may view them as high-dimensional data based on a single realization of a random event.

Furthermore, outcomes are not independent over disjoint time intervals and between individuals. One example is the phenomenon of *herd immunity* where individuals that are vaccinated indirectly protect those who are not vaccinated.

These data features determine the statistical models and methods that are different from those based on designed experiments with i.i.d. samples.

7.2.2 What Do We Mean by "Large Number"?

In the classic statistics textbooks, "large number" is associated with the *law of large numbers*, the central limit theory, the asymptotic confidence intervals, and asymptotically unbiased estimates, such as the estimates based on unbiased estimating equations. In such context, it is called the sample size, which is understood as the number of repeated independent random experiments under identical conditions. An infectious disease outbreak dataset, no matter how many observations, is considered a small sample.

In a different context, when we say that deterministic models are approximations of the mean field of the corresponding stochastic processes, we do not mean large populations, but large repeated realizations of the same outbreak under identical conditions. To a certain extent, when the population size m in disease transmission models becomes large, the stochastic effects of correlations among the numbers of individuals in different compartments are reduced, even negligible, such as $\frac{\beta}{m} cov\{S(t), I(t)\}$ in (5.23). This may lead to a smooth realized epidemic curve that resembles that predicted by a deterministic model. However, it is not the average of all possible epidemic curves in large numbers of repetitions of the epidemic under identical conditions. This distinction was illustrated in Figs. 5.1 and 5.2 in Chap. 5. The deterministic models can be viewed as approximations of the mean field of the corresponding stochastic processes in the context of a large number of repetitions of the epidemic under identical conditions. The population size is not equivalent to the sample size.

In fitting models to time-series data, increasing the number of observations means more accumulation of data over time to achieve longer time-series. A single time-series is still regarded as sample size = 1. Longer time series also improves the precision of the estimates, but only to a certain extent, and is not equivalent to having a large sample. Increasing observations over time often forces us to change to more complex models whereas increasing the sample size does not.

7.2.3 Lack of Information or Not Identifiable?

In a single parameter setting, the lack of information from data simply means very imprecise estimation. The confidence interval is extremely wide, or one-sided, or unable to yield the point estimate (e.g., the likelihood function is maximized at the boundary of the parameter space). It is often characterized by a very flat likelihood function over the parameter space (Raue et al. 2009; Roosa and Chowell 2019).

In multiple parameter settings data cannot provide accurate estimates for some of the parameters or cannot test against certain hypotheses when fitting the model to a single data source (i.e., single time-series) especially in models that involve sub-models for the random, systematic, and link components. People often say, ambiguously, that data do not have enough information or are not able to identify certain parameters. We would like to point out some subtle differences.

Shared Information in a Multiple Parameter Setting

We start with the classic statistical problem of the i.i.d. sample $\left(x_1, \cdots, x_n\right)$ arising from the Gaussian distribution $N(\mu, \sigma^2)$ with the parameter of interest being the variance σ^2. It is well known that the maximum likelihood estimate of σ^2 is

$$\widehat{\sigma^2} = \begin{cases} \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\,, & \text{if } \mu \text{ is known;} \\ \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\,, & \text{if } \mu \text{ is unknown and } \widehat{\mu}=\overline{x} \end{cases}.$$

It is also well known that when μ is known, $\widehat{\sigma^2}$ is an unbiased estimator, whereas if μ is unknown $\widehat{\sigma^2}$ is only asymptotically unbiased for σ^2, but biased in any finite population. It is unbiased for $\frac{n-1}{n}\sigma^2$. The unbiased estimator for σ^2 is $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$. This is because the information for σ^2 in the data is shared with the parameter μ and the minimum sufficient statistics for μ is $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Data can be re-arranged through one-to-one mapping:

$$\left(x_1, \cdots, x_n\right) \mapsto (x_1-\overline{x}, \cdots, x_{n-1}-\overline{x}, \overline{x})$$

in which $\overline{x}$ contains all the information in the data for μ and $(x_1-\overline{x}, \cdots, x_{n-1}-\overline{x})$ are the residuals with $n-1$ degrees of freedom left for the estimation of σ^2.

This argument is formalized in the Fisher-Neyman factorization of the likelihood function

$$f\left(x_{1,}\cdots,x_n\right)=\frac{1}{\left(\sqrt{2\pi}\sigma\right)^n}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}=\frac{1}{\left(\sqrt{2\pi}\sigma\right)^n}e^{-\frac{n}{2\sigma^2}(\overline{x}-\mu)^2}e^{-\frac{n-1}{2\sigma^2}s^2},$$

where $s^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2$. The pair $(\overline{x},s^2)$ is the minimal sufficient statistics for (μ,σ^2). In other words, data are reduced to the pair $(\overline{x},s^2)$ and the information for variance is summarized as the mean of the square errors $(x_i-\overline{x})^2$ from $n-1$ out of the data of sample size n.

A more revealing example is the Neyman-Scott paradox (Neyman and Scott 1948). Consider $2n$ independent measures $\left(x_{1,}\cdots,x_n\right)$ and $\left(y_{1,}\cdots,y_n\right)$, where $X_i\sim N(\mu_i,\sigma^2)$, $Y_i\sim N(\mu_i,\sigma^2)$, $i=1,\ldots,n$. There are $n+1$ unknown parameters $(\mu_{1,}\cdots,\mu_n,\sigma^2)$. If we use the likelihood function given by the joint distribution

$$f\left(x_{1,}\cdots,x_n,y_{1,}\cdots,y_n\right)=\frac{1}{\left(\sqrt{2\pi}\sigma\right)^{2n}}e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i-\mu_i)^2+\sum_{i=1}^{n}(y_i-\mu_i)^2\right)},$$

the maximum likelihood estimates are

$$\widehat{\mu_i}=\frac{1}{2}\left(x_i+y_i\right),\ i=1,\ldots,n$$
$$\widehat{\sigma^2}=\frac{1}{4n}\left(x_i-y_i\right)^2.$$

In this case, it can be shown that $\widehat{\sigma^2}$ is not only biased, but also asymptotically biased:

$$E\left[\widehat{\sigma^2}\right]\to\frac{\sigma^2}{2},\ \text{as}\ n\to\infty.$$

This analysis under-estimates σ^2 by 50% because half of the information in data about σ^2 is lost in estimating $(\mu_{1,}\cdots,\mu_n)$.

Later in Chap. 8, we shall see an example with discussions, where the model evolves from relatively simple to more complex by adding a shape parameter, adapted to an increasing number of observations of a time-series during a Zika outbreak investigation. In the simpler form, all three parameters directly address three public health questions of interest. Data collected in the early period have little information on these parameters in terms of very wide confidence limits. As data accumulate, the estimation for these parameters becomes more and more precise. Meanwhile, data start to force us to add a shape parameter to the model, which does not directly address any of the questions of public health interest. It does,

however, improve the model's goodness of fit and also correct potential biases in the estimated parameters of interest, which is increasingly apparent as suggested by data. This added parameter needs to be estimated using the information from data at a cost of the precision in the estimation of parameters of interest. Therefore, at some midpoint in the outbreak, discussions are needed to address the pros and cons of expanding the model at that moment when the time-series might not be long enough to accommodate another parameter. It is only after another month of data accumulates that it becomes obvious that the more complex model is necessary and meanwhile the estimation of parameters of interest becomes more precise.

In a different example, Lagakos et al. (1988) presented data based on 258 adults with transfusion-associated AIDS. Data were fitted to a Weibull distribution (2.12) with scale parameter λ and shape parameter $\varsigma > 0$. The purpose is to estimate the incubation period from the time of transfusion and the onset of AIDS illnesses. Data are not i.i.d. in the sense of a random sample from an experiment, but from a different type of observational scheme (to be discussed later). Figure 7.3 illustrates the contours of the surface of the log-likelihood. It shows that data do not have enough information for the scale parameter λ except for $\lambda \leq 0.128$, and the m.l.e. does not exist. Together with estimated $\widehat{\varsigma} \approx 2.1$, at best the data tell us that the incubation period is very long, with the lower bound of the estimated median incubation period ≥ 6.6 years. On the other hand, data are informative about ς, with $1.85 \leq \varsigma \leq 2.38$ based on the approximate 95% confidence limits based on the likelihood ratio statistics. It implies an approximately linear increasing hazard function.

The take home message from the above discussions includes:

1. when there are multiple unknown parameters in the model, information in the data are shared among the parameters;

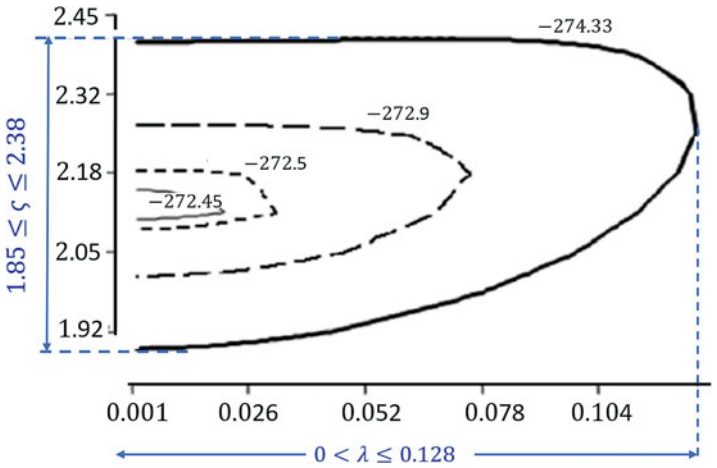


Fig. 7.3 Contours of the surface of the log-likelihood $l(\lambda, \varsigma)$ of the Weibull distribution along with the 95% confidence region using the likelihood ratio statistics for the two parameters based on data from Lagakos et al. (1988)

2. in the presence of *nuisance parameters,* without additional statistical modelling to handle the nuisance parameters, data may not have enough information to estimate the parameter of interest, in the sense of precision as well as potential biases;
3. the amount of information in data with respect to the parameters of interest is also affected by how the data are collected.

Identifiability Among Parameters and Components of Models

For two parameters (α, β), if different combinations (α_1, β_1), (α_2, β_2), ... produce the same expected value of the model that gives equally good fit to data, we say that data are not able to identify them. This problem can arise for different reasons.

One of the sources that can cause the identifiability problem is due to high correlation among parameters in a nonlinear function. The banana shaped likelihood contour in Fig. 7.2 illustrates such a correlation. In Chap. 8, we shall see many banana shaped likelihood contours. In some applications, the paired combination (α, β) corresponds to a specific scenario of scientific or epidemiologic interest. For example, one parameter may represent the number of infected individuals at the beginning of the epidemic and another parameter may represent the rate of growth of the epidemic. Data may not be able to identify whether the observed phenomenon is due to a small number of initially infected individuals combined with a fast growth rate, or a large number of initially infected individuals combined with a slow growth rate. On the one hand, lack of parameter identifiability is not the same concept as the lack of information because, through the contours of the likelihood function, data do provide a great deal of information on how the parameters are correlated and are able to estimate these parameters. On the other hand, as the number of observations increases, more precise estimates can be obtained, which in many cases solves the identifiability problem.

Another source of identifiability issues is the redundancy of parameters. For example, the logistic growth function (5.14), parameterized as

$$\frac{mi_0(\beta - \gamma)}{\beta i_0 + (m\,(\beta - \gamma) - \beta i_0)\, e^{-(\beta-\gamma)t}}$$

has a shape of a sigmoid curve. Data may allow rather precise estimation of the initial growth $\rho = \beta - \gamma$, but cannot identify β or γ separately unless one of them is known. All these parameters have specific scientific interpretations. In the context of the SIS compartment model, the hypothesis $H_0 : \gamma = 0$ distinguishes whether the model is SI or SIS. In this sense, the parameters are not redundant. However, from the data point of view, only three of the four parameters can be estimated. When data admit a logistic functional form, they cannot identify the SIS model. In this case, increasing the number of observations will not solve the identifiability problem of parameters. Re-parameterization of the logistic function will, though it cannot solve the problem of identifying underlying model structures.

The identifiability problem often arises in models that involve sub-models. For instance, it is well known that univariate frailty models (Sect. 2.6) through mixture of distributions are not identifiable from the survival information alone. Similarly, in a model that has random, systematic, and link components, some parameters are specific to the disease transmission process and other parameters are specific to another aspect of the data-generating process beyond disease transmission. A single time-series data usually cannot identify some of the embedded processes or components, especially if two components are linked through convolution.

Now we move on to the *link* component of the model (which has been less discussed so far).

A typical example is back-calculation, in which the disease transmission process is modelled through an intensity function $i(t;\underline{\theta})$, which describes the incidence of new infections over time, where new infections are not directly observable. Data $\underline{y}$ are generated based on the occurrence of the subsequent observable events, such as the onset of clinical symptoms as a consequence of infection. A statistical model may assume the observable events arise from a counting process with the intensity function $\mu(t;\underline{\theta},\underline{\psi})$, which is further modelled as a convolution

$$\mu(t;\underline{\theta},\underline{\psi}) = \int_0^t i(u;\underline{\theta}) f(t-u|u;\underline{\psi}) du \tag{7.15}$$

or $\mu(t;\underline{\theta},\underline{\psi}) = \sum_{u=0}^{t} i(u;\underline{\theta}) f(t-u|u;\underline{\psi})$, depending on whether one takes a continuous time or a discrete time framework. The quasi-likelihood generalized estimating equation (7.2) becomes

$$\sum_{t=1}^{T} \frac{\partial \mu(t;\underline{\theta},\underline{\psi})}{\partial \theta_j} \frac{y_t - \mu(t;\underline{\theta},\underline{\psi})}{V[Y_t;\underline{\theta},\underline{\psi}]} = 0,\ j = 1,\ldots,m. \tag{7.16}$$

In this model, the convolution (7.15) serves the same role as (7.1) in the generalized linear model.

The systematic component is $i(t;\underline{\theta})$ which captures the data-generating process due to disease transmission specified by a vector of parameters $\underline{\theta}$. It may be stochastic or deterministic transmission models that are explicitly linked to the underlying scientific hypotheses regarding the agent–host–environment interface. The number of unknown parameters in $\underline{\theta}$ depends on the complexity of the model. Donnelly and Ferguson (1999) formulated dynamic models for the population biology of the bovine spongiform encephalopathy (BSE) and then embedded this model into a back-calculation framework along with the maximum likelihood estimation. This approach gave estimated annual incidence of animals infected with BSE in Great Britain. In most other cases, it is convenient to adopt some flexible empirical parametric functions for $i(t;\underline{\theta})$ with relatively few parameters such as a generalized logistic function. There are also the flexible step-function models such as

$$i(t;\underline{\theta}) = \theta_j,\ j = 1,\cdots,q \tag{7.17}$$

involving q steps. Each step θ_j is a parameter. The longer the steps, the fewer the number of parameters.

The component $f(t-u|u;\underline{\psi})$ is a model that captures the details of all other data-generating processes since infection, as the conditional probability of being captured in the data at time t after an amount of time $x = t-u$, given infection at time $u < t$, specified by a vector of parameters $\underline{\psi}$. It may include aspects such as the factors that determine diagnoses of infections like the onset of clinical symptoms or external influences such as public health campaigns and screening. It may also include the process of reporting diagnosed infections to a central registry, delays in reporting, whether data are collected prospectively or retrospectively, and so on.

The estimating equation (7.16) incorporates the random components by specifying that $E[Y_t] = \mu(t;\underline{\theta},\underline{\psi})$ and appropriate variance structure $V[Y_t;\underline{\theta},\underline{\psi}]$.

The parameter of interest is the vector $\underline{\theta}$ because the objective is to estimate the incidence of new infections over time. However, data can only identify the convolution $\mu(t;\underline{\theta},\underline{\psi})$ as a whole, but not the systematic component $i(t;\underline{\theta})$ and the link component $f(t-u|u;\underline{\psi})$ separately.

7.2.4 *Observable Data and Unobservable Events*

At the Population Level

The ideal sequence of a scientific investigation is: formulation of research questions—obtaining appropriate data—analysis of data—interpretation of results. However, the investigation based on observational data collected during an infectious disease outbreak is an extreme departure from the ideal sequence, and most of the data are collected for other purposes unrelated to the question under investigation.

For example, a research question may be addressed using an SIR model in Sect. 5.3 with two parameters (β,γ). Using martingales (briefly introduced in Sect. 3.3.2), Becker (1989), Rida (1991), Becker and Hasofer (1997), Becker and Britton (2001), Hohle and Jørgensen (2003), and others have developed estimating equations that yield asymptotically unbiased estimates

$$\widehat{\beta} = \frac{C(t_N)}{\int_0^{t_N} \frac{S(x)I(x)}{m}dx},\ \widehat{\gamma} = \frac{C(t_N)}{\int_0^{t_N} I(x)dx},$$

with standard error estimates

$$s.e.\left(\widehat{\beta}\right) = \frac{\widehat{\beta}}{\sqrt{C(t_N)}},\ s.e.\left(\widehat{\gamma}^{-1}\right) = \frac{\widehat{\gamma}^{-1}}{\sqrt{C(t_N)}},$$

where t_N is the time of the observed end of the epidemic, and $C(t) = m - S(t)$ is the cumulatively infected individuals at time t, and m is the population size. Data

are assumed to arise as continuous and complete observation for $\{[S(t), I(t)] : 0 \leq t \leq t_N\}$. However, such data do not exist in reality, simply because a transmission from an infected individual to a susceptible individual is an unobservable event.

Most data at the population level are random counts based on observable events aggregated into time intervals. They may arise from multiple sources involving multiple agencies regarding the same outbreak. For example, a central public health agency of a country may compile a registry of reported "cases" of a certain reportable disease, which are forwarded from similar disease registry systems in state, provincial, territorial, or local authorities. Meanwhile, a different agency, or institute, or a collaborative sentinel hospital may have a database with variable population coverage regarding hospitalizations, discharges, and other events on the severe end of the same disease. In recent years, syndromic surveillance based on early warning indicators, such as emergency department attendances or emergency telephone calls, have gained much attention for their potential use to detect outbreaks at early stages. In the era of Big Data, the computer algorithm Google Trends aggregates Google Search queries by monitoring millions of users' health tracking behavior online.

The increasing number of data sources comes with pros and cons. On the pros side, multiple data sources, at least conceptually, help to identify high dimensional parameters in a complex model. On the cons side, it becomes increasingly difficult to sort out the data-generating process for each data source and increasingly difficult to develop the corresponding statistical models, not to mention model criticism.

What Is a "Case"?

In workshops on mathematical epidemiology, we have encountered questions from mathematicians such as:

> How do we reconcile differences between the incidence of new infections predicted by our models and the incidence of new cases in surveillance data we try to fit?

In the preceding paragraphs, we have illustrated the gap between the unobservable events predicted by mathematical models and data that are collected at the population level. Here we would like to highlight part of the data collection process, which is the preciseness in definitions and terminologies.

A "case" is one of the most commonly used terms in epidemiology and public health surveillance. A system based on reporting diagnosed diseases to a central registry is often called "case-reporting-surveillance." For each surveillance system, there is a case-definition (which may evolve and change over time).

We would say that a case is a file associated with an individual diagnosed with some "case-defining" illnesses, and this case must exist in some central registry in the system. Inside the "case" (more precisely, in the file), there are multiple observable events associated with different time points. Some events are relevant to the underlying epidemiology, such as the onset of clinical symptoms (but only among those according to the "case-definition"), the diagnosis of such symptoms

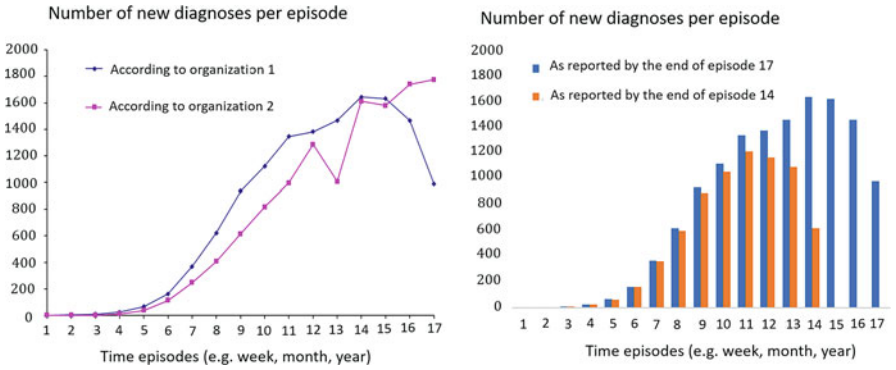


Fig. 7.4 Left: Illustration of two presentations of the "number of cases" per episode for Disease X from the same dataset; Right: Illustration of "number of new diagnoses" per episode being modified as new reports come in, due to reporting delays

(sometime later by a medical doctor among individuals who seek medical attention), and some subsequent clinical events during the follow-up such as morbidity, mortality or recovery. Other events are not directly relevant to epidemiology, but are equally important, not only for bookkeeping of the surveillance system but also serve as a bridge between the unobservable events and the observable data. These events may include the creation of this case (file) with the date, the arrival of the file to a local or a central registry with dates, the entry of the case into the database with date. However, the most important event, the infection and the associated time at infection, is not observable and is not documented in the case.

It is not uncommon in published reports and public health literature that disease trends are presented as "number of cases" over time when the meaning of a "case" is ambiguous. In Fig. 7.4 (for illustration purposes only), the actual disease, the time episodes, the population and the organizations involved are anonymized. The left panel shows two completely different trends about the same disease from the same data disseminated by two organizations and both use the word "case" ambiguously. Upon careful inspection, it turns out that each case-file has two dates associated with two different events: one is the date of a new diagnosis signed by a doctor and the other is the date on the stamp (which could be a computer digital signature) by the system when the file is entered into the registry. Organization 1 presents the trend of "new diagnoses" over time according to the most recent data; whereas, Organization 2 presents the trend of "new reports received by the system" over time. Both organizations call them "numbers of new cases."

The gap between the time of diagnosis and the time when this individual case is reported and entered into the database of a public health registry is called *reporting delay*. If this gap is ignorable (close to zero), then the two trend curves in the left panel of Fig. 7.4 should be nearly identical. If the gap is large but there is little variation from individual to individual, then the two curves should look alike, with a shift of fixed number of periods. If the gap has random variations, then the two curves will not look alike.

When there are substantial reporting delays, Organization 1 would argue that it is the new diagnosis, not the new entry to the system, that is relevant to the trend of the epidemic. However, the numbers of new diagnoses per episode will be modified as newly reported cases come in, especially for the recent episodes, as shown in the right panel of Fig. 7.4. Furthermore, there is usually a declining trend near the end, as cases with most recent time of diagnoses are still not yet reported. Therefore, the trend based on time of diagnoses must be statistically adjusted, especially for the recent past.

Organization 2 would argue that the number of new "cases" defined as new entries of the disease per episode is a static number. By the end of every episode, a new number is added to the database regardless of time at diagnoses and the reporting delay is irrelevant. After all, it presents useful trend information on the case-load seen by the registry. In our opinion, this could be misleading, especially when reporting delays are long and variable (Tariq et al. 2019).

While trends of specific events over time are meaningful, presenting trends of "cases" may not be. A good book-keeping practice in the registration system is to line-list all the events longitudinally for each reported individual whether the event is clinically relevant or not. All the documented events may serve some purposes. The researcher will decide which key event is the most relevant event to the question, but will also use some events and corresponding time lines to adjust biases such as reporting delay. The latter is part of statistical modelling. Naturally this will demand more resources and due diligence both on the system and the researcher.

In analyses involving diverse data sources, many agencies may contribute data originally collected for other purposes with more ambiguity in terminology. It will be more challenging for a researcher to get into the depth of each data source and statistically model the data-generating process. In the era of Big Data, will artificial intelligence be able to model all the data-generating processes and conduct model criticism? Quoting from Cox and Donnelly (2011):

> A large amount of data is in no way synonymous with a large amount of information. In some settings at least, if a modest amount of poor quality data is likely to be modestly misleading, an extremely large amount of poor quality data may be extremely misleading.

At Individual Levels

Phenomenological population models, especially those that mechanically model disease transmission dynamics, carry tacit assumptions at the level of individuals, such as the distributions of the latent periods and the infectious periods, as well as the infectious contact process. All events associated with these assumptions are not observable. One cannot pinpoint the time that an infection, which is the transfer of the infectious agent from an infected individual to a susceptible individual, occurs; nor can one ascertain the time when an infected individual is no longer latent and starts to be infectious. Therefore, there is no ideal data that can be directly used to validate these models.

In many diseases, the onset of clinical symptoms can be ascertained either precisely or within a narrow time interval. If the time of infection can also be ascertained within an acceptable range, for instance, through contact-tracing, then the *incubation period* is defined as the duration from the infection to the onset of symptoms and can be measured with some acceptable uncertainty. In some diseases, symptom onset may be used as a proxy for the beginning of the infectiousness and the *incubation period* may be a proxy for the *latent period*. However, there are diseases where a proportion of infected individuals may remain asymptomatic and are still able to transmit the infection.

The diagnosis of an infection, either due to onset of clinical symptoms or other screening/testing mechanisms, is always observable and ascertained to a specific point in time. This is also the event that generates most of the data. However, this event involves two mechanisms. One is driven by the progression of the disease natural history. The other is influenced by external factors. With respect to Fig. 7.4, Organization 1 is only partially right by saying that the new diagnoses are relevant to the trend of the epidemic. Before the time-series, represented as numbers of new diagnosis over time, become fully informative about the disease spread, statistical models are required to capture the entire data-generating process, incorporating the disease progression, the external factors such as how long since symptom onset and reasons for seeking diagnosis, as well as duration from the initial diagnosis to the entry of the case to the data registry.

For diseases with disease induced mortality, death caused by the disease is an observable event.

For models involving intervention, such as vaccination, treatment, isolation, etc., all of these are also observable.

Serial Interval, Generation Interval, Generation Time, and So On

In recent literature, these "intervals" have been widely cited, measured, and applied to outbreak investigations, especially during the early transmission phase. However, there is a lot of ambiguity. For instance, there are occasions that the same terminology is associated with two different definitions, whereas there are other occasions that the same definition is assigned to different terminologies by different researchers.

To our knowledge, the earliest definition of *serial interval* dates back to Hope Simpson (1948):

> The period from the observation of symptoms in one case to the observation of symptoms in a second case directly infected from the first is the (clinical) serial interval. It is an observable epidemiological unit.

Bailey (1975) wrote that:

> The period from the observation of symptoms in one case to the observation of symptoms in a second case directly infected from the first is the *serial interval*. Thus the serial interval

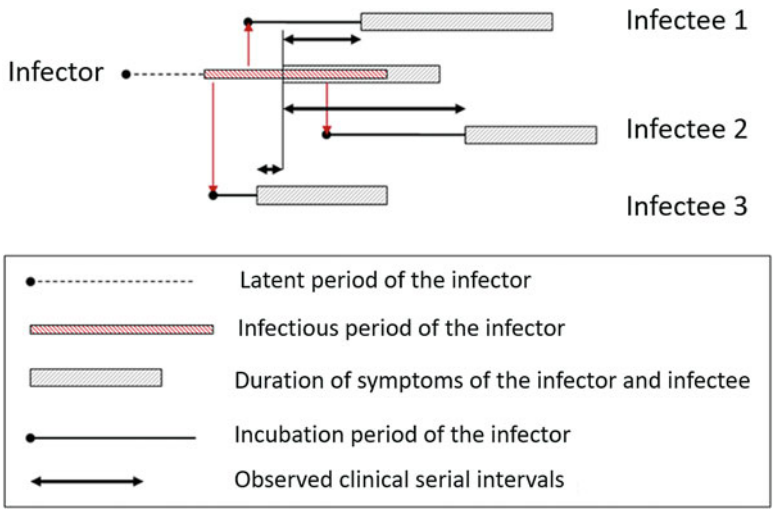


Fig. 7.5 A schematic presentation of three serial intervals produced by the same infector with 3 infectees

is the *observable epidemiological unit*, and it reflects to some extent the life cycle of the infectious organism. Nevertheless, it can not be readily related to the mechanism of transfer.

Decades later, Lipsitch et al. (2003) defined serial interval in the same way as that in Hope Simpson (1948) and Bailey (1975) but also specified that it is defined as the average between the two observed onset of symptoms. Lipsitch et al. (2003) applied such a measure during the outbreak investigation of the transmission of the severe acute respiratory syndrome (SARS) to estimate the basic reproduction R_0.

The features in the above definition of the serial interval include:

1. involving a pair of infected individuals, an infector, and an infectee;
2. observable;
3. book keeping device to track generations, White and Pagano (2008);
4. depending on the latent period and the infectious period (of the infector) as well as the incubation period (of the infectee).

For example, Infectees 1 and 3 in Fig. 7.5 are both infected before the symptom onset of the infector, with Infectee 1 following the natural sequence that its own clinical onset takes place after its infector's onset; whereas Infectee 3 has the reversed sequence, with its own clinical onset taking place before its infector's onset. This can happen in theory, if both the infectious periods and the incubation periods are highly variable.

There has been some confusion in the literature, as various terms have been used to refer to the same concepts. Before Hope Simpson, Pickles (1939) used the term *transmission interval* for what was later defined as serial interval with reference to empirical observations of a hepatitis epidemic in the United Kingdom (Nishiura 2010).

A different measure is the interval between the time of infection and time of transmission by linking two individuals, the infector and the infectee. This is formally named as the *transmission interval* in Fine (2003). Fine (2003) made it clear that (1) the transmission interval (i.e., the interval between successive infections) and (2) the clinical onset serial interval (i.e., the interval between successive clinical onsets) are different both conceptually and quantitatively. This distinction was also made clear in Svensson (2007).

A different name was given to the transmission interval as the *generation interval* in Wallinga and Lipsitch (2007), Roberts and Heesterbeek (2007), described as the duration between "the time of infection of an individual to the time of infection of a secondary case by that individual." Svensson (2007), Nishiura (2010), Kenah et al. (2008), and many others called it the *generation time.* Minor differences in the definitions among these authors are whether this interval is defined as a random variable, or is defined according to its mean value. Svensson (2007), from a sampling point of view, further points out the difference in distribution between the *primary generation time* as measured prospectively from the time of the infection of the infector to the transmission to the infectee, and the *secondary generation time* as measured retrospectively from the time of the transmission to the infectee to the infection of the infector.

The features in the above definition are that

1. involving a pair of infected individuals, an infector and an infectee;
2. both the infection of the infector and the passing of infection to an infectee are unobservable;
3. depending on the latent period and the infectious period.

The generation intervals (or generation times, transmission intervals) are distinguished from the serial intervals by definitions, but sometimes the two are used interchangeably in the literature. Figure 7.6 compares different time periods between the first patient and the second patient, adopted from the Field Epidemiology Manual published by the European Centre for Disease Prevention and Control (ECDC). In this diagram, generation time is synonymous to serial interval, both refer to the interval between successive clinical symptoms.

Anderson and May (1991) defined the *generation time* as the sum of the latent period and the infectious period: T_E+T_I. Daley and Gani (1999) defined the *average generation time* as $\mu_E + \mu_I/2$ where $\mu_E = E[T_E]$ and $\mu_I = E[T_I]$. According to strict arguments by Fine (2003), the sum of the latent and (part of) the infectious periods is concerned with the course of a single infection and is different from the interval between successive infections.

In a recent article, Champredon and Dushoff (2015) defined the *intrinsic generation interval* with its distribution defined by the p.d.f.

$$g(x) = \frac{\beta(x)A(x)}{R_0}, \ x > 0 \tag{7.18}$$

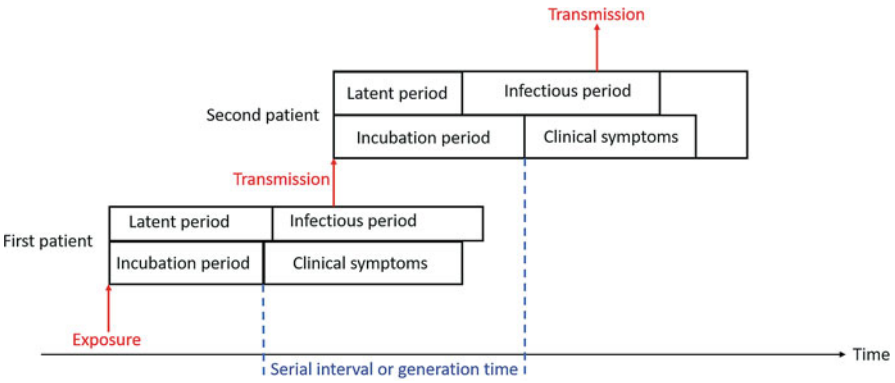


Fig. 7.6 Relationships between time periods

where $R_0 = \int_0^\infty \beta(x)A(x)dx$ is the basic reproduction number as formulated in Chap. 4 following the renewal-type equation (4.35). This concept is behind the developments of the theories in Wallinga and Lipsitch (2007) and Roberts and Heesterbeek (2007) that link the distribution of the generation intervals given by (7.18) to the estimation of R_0 with implicit assumptions that the generation intervals, according to their respective definitions, follow such a distribution. The intrinsic generation interval is defined along the course of a single infection. In particular, if $\beta(x) = \beta$, then $g(x) = A(x)/\mu_I$, $x > 0$. In structured models that involve latent periods T_E followed by infectious periods T_I, assuming independency, $A(x) = \int_0^x \overline{F}_I(x-u)f_E(u)du$, that gives

$$g(x) = \int_0^x f_E(u)\frac{\overline{F}_I(x-u)}{\mu_I}du. \tag{7.19}$$

The mean value is

$$\mu_G = \mu_E + \frac{1}{2}(1+\phi^2)\mu_I = \mu_E + \frac{\mu_I}{2} + \frac{var[T_I]}{2\mu}$$

as previously given as (4.48), where $\mu_E = E(T_E)$ and ϕ is the coefficient of variation of the infectious period T_I defined as the ratio of standard deviation to the mean. If the infectious periods are exponentially distributed with mean μ_I, then $\mu_G = \mu_E + \mu_I$ which is the mean generation time according to Anderson and May (1991). If the infectious periods are the same constant μ_I, then $\mu_G = \mu_E + \mu_I/2$ which is the mean generation time according to Daley and Gani (1999). If the infectious periods can be expressed as the sum of n independently and identically distributed exponential distributions with mean μ_I/n, then $\phi^2 = 1/n$ and $\mu_G = \mu_E + \frac{n+1}{2n}\mu_I$. This expression can be found in Roberts and Heesterbeek (2007).

Champredon and Dushoff (2015) further discussed forward generation interval from the perspective of an infector and backward generation interval from the perspective of infectee. We have discussed in Chap. 4 that the distribution given by (7.19) also has a sampling perspective. It coincides with that based on the convolution of the latent period T_E and the equilibrium distribution given by p.d.f. $f_W(x) = \overline{F}_I(x)/\mu_I$, under suitable assumptions concerning equilibrium conditions of the epidemic at the population level. In this case, it is appropriate to define the (intrinsic) generation time as

$$T_G = T_E + W.$$

It has the sampling property that an arbitrary observer makes a snapshot sample at an arbitrary time. All individuals who are "currently infectious" form a prevalence cohort. The observer looks backward to the time of infection and measures the time from infection to the observation time. The distribution has the p.d.f. given by (7.19).

In summary,

1. Serial intervals are between observable events, subject to observation errors and time-length bias (to be discussed next), but cannot be readily related to the mechanism of transmission. Their distributions, even correctly estimated by data, may not be used to approximate the distributions of the transmission intervals or the intrinsic generation times.
2. Transmission intervals (also known as generation intervals, generation times) are not directly observable. They are further distinguished by forward measuring from the time of an infector to the time of transmission to an infectee, versus backward measuring from the time of infection of an infectee to the time of infection of its infector. These two measurements follow different distributions (Svensson 2007).
3. The intrinsic generation interval includes the definitions given by Anderson and May (1991) and Daley and Gani (1999) as special cases. By definition, it is related to the basic reproduction number R_0 that carries information about disease transmission. This relation is established under equilibrium conditions (implicitly assumed) as R_0 itself is also defined under such conditions. This is more apparent in a structured model with latent and infectious periods, where the intrinsic generation time can be written explicitly as $T_G = T_E + W$, where W corresponds to the equilibrium distribution of the infectious periods. The intrinsic generation intervals do not involve pairs of infectors and infectees and are not observable.

7.3 Time-Length Bias

The time-length bias discussed here is in the same nature of the famous "survivorship bias." During World War II, researchers from the Centre for Naval Analyses conducted a study of the damage made to planes that had returned from missions.

The statistician A. Wald noticed that the study only considered the planes that had survived their missions. Those that had been shot down were not present for the damage assessment. The holes in the returning aircraft represents areas where the aircraft could take damage and return home safely. Wald (1943) proposed that the Navy reinforce areas where the returning planes were unscratched, since those areas, if hit, would cause the plane to be lost. The same type of bias also applies in observational data in the study of disease outbreaks.

Observational data during an infectious disease outbreak are often length-biased with respect to key epidemiological durations, such as the latent periods, infectious periods, incubation periods, generation times, etc. In some cases, individuals associated with longer durations are more likely to be included in the data. In other cases, individuals associated with shorter durations are more likely to be included in the data. At the population level, length biases at individual levels further lead to mis-interpretation of the disease trends. Figure 7.4 has an illustration of disease trends by date of onset, affected by reporting delays.

For a comprehensive review on length-biased sampling and length-biased distribution, we recommend Chapter 1 of Qin (2017).

Disease progression within an infected host involves sequences of events. Each pair of successive events is composed of an *initiating event* that leads to a *subsequent event* over a random duration $X \geq 0$. An individual is denoted by the index i. We used $T_i^{(1)}$ for the time of onset of the initiating event and $T_i^{(2)}$ for the time of onset of the subsequent event. The duration of interest is $X_i = T_i^{(2)} - T_i^{(1)}$.

7.3.1 Prevalence Cohorts and Left-Truncation

The prevalence cohort as illustrated in Fig. 4.8 leads to length-bias that systematically includes individuals with longer durations. Assuming that X_i among individuals are (in theory) i.i.d. with p.d.f. $f_X(x)$, the distribution of X_i as observed in the prevalence cohort is length biased because the arbitrary sampling time t must satisfy $T_i^{(1)} < t \leq T_i^{(2)}$. The length biased duration is denoted by $X_i^{(B)}$. We also write $X_i^{(B)} = W_i + V_i$, where $W_i = t - T_i^{(1)}$ and $V_i = T_i^{(2)} - t$.

We further assume that

1. The occurrence of the initiating events, which is a stochastic process, follows constant incidence rate (i.e., the equilibrium condition).
2. X_i is independent of the occurrence of the initiating event.

Under these conditions, $X_i^{(B)}$, W_i, and V_i have the following equilibrium distributions (Wang 2005) with p.d.f.s

$$X_i^{(B)} \sim \frac{x f_X(x)}{\mu}, \text{ and } W_i \sim V_i \sim \frac{\overline{F}_X(x)}{\mu} \tag{7.20}$$

where $\mu = E[X] = \int_0^\infty x f_X(x)dx = \int_0^\infty \overline{F}_X(x)dx$.

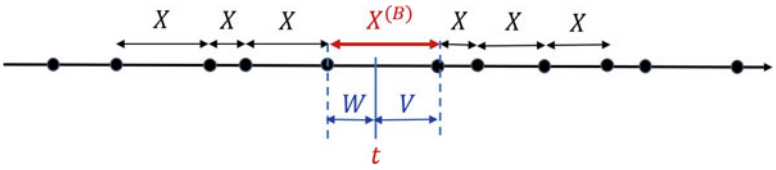


Fig. 7.7 Illustration of a repeated testing scheme on an infectious disease

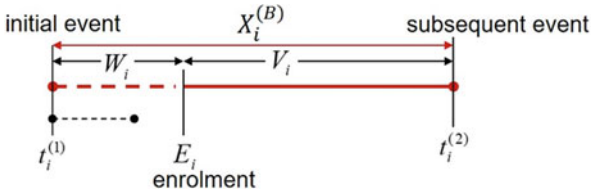


Fig. 7.8 Illustration of the observation scheme with left-truncation. The individual with short duration (dotted line) is not included at the time of enrollment

The same problem is formulated differently as illustrated in Fig. 7.7. Consider a repeat testing scheme for an infectious disease such as HIV. Individuals get tested repeatedly with i.i.d. inter-testing intervals X with p.d.f. $f_X(x)$. Denote $X^{(B)}$ the special interval between the last negative and the first positive tests. Assuming that the epidemic in the population is at equilibrium so that the occurrence time t of the "sero-conversion event" is distributed uniformly between the two tests. This interval is length-biased with equilibrium distributions given by (7.20). These distributions are applied for designing better repeated testing algorithms to reduce the prevalence of not yet diagnosed HIV infections (Yan and Zhang 2018).

A more general setting is the left-truncation in survival analysis. The initial events occur over time t following a random process with intensity $\lambda(t)$. For each individual i, the subsequent event occurs following the initial event after a random duration X_i. Assuming $X_i's$ are i.i.d. as the random variable X, and the objective is to estimate this distribution. However, data are collected by enrollment. The time at enrollment E_i for individual i must satisfy that the initial event has taken place while the subsequent event has not. Thus individuals with longer X_i have more chance to be enrolled. Observed data are length-biased following the distribution $X^{(B)}$. Each individual $X_i^{(B)}$ arises from the conditional distribution of X given $X \geq W_i$. An illustration is given in Fig. 7.8.

The observed part from enrollment until the end point is V_i which has conditional the survival function and the conditional p.d.f.

$$\overline{F}(v_i|w_i) = \frac{\overline{F}(w_i + v_i)}{\overline{F}(w_i)} = \frac{\overline{F}(t_i^{(2)} - t_i^{(1)})}{\overline{F}(E_i - t_i^{(1)})},$$

$$f(v_i|w_i) = \frac{f(w_i + v_i)}{\overline{F}(w_i)} = \frac{f(t_i^{(2)} - t_i^{(1)})}{\overline{F}(E_i - t_i^{(1)})}.$$

This is the residual life distribution discussed in Sect. 2.3 in Chap. 2. If the time of initial event $t_i^{(1)}$ cannot be ascertained but follows a random process with intensity $\lambda(t)$ until enrollment, then the p.d.f. of V_i is

$$f(v_i|w_i) = \frac{\int_{-\infty}^{E_i} \lambda(t) f(t_i^{(2)} - t)dt}{\int_{-\infty}^{E_i} \lambda(t)\overline{F}(E_i - t)dt}.$$

If $\lambda(t) = \lambda$ is constant, then the above becomes

$$\begin{aligned} f(v_i|w_i) &= \frac{\int_{-\infty}^{E_i} f(t_i^{(2)} - t)dt}{\int_{-\infty}^{E_i} \overline{F}(E_i - t)dt} \\ &= \frac{\int_{v_i}^{\infty} f(x)dx}{\int_0^{\infty} \overline{F}(x)dx} = \frac{\overline{F}(v_i)}{\mu} = f_V(v_i), \end{aligned}$$

where $v_i = t_i^{(2)} - E_i$ and $\mu = \int_0^{\infty} \overline{F}(x)dx$. In this case, we have recovered the equilibrium distribution. Data arising from left-truncated data under equilibrium can be used to estimate the distribution $\overline{F}_X(x)$ because the distribution of the observed part $v_i = t_i^{(2)} - E_i$ contains all the information, independent of the truncation time $w_i = E_i - t_i^{(1)}$.

If $\lambda(t)$ is constant but the time $t_i^{(1)}$ can be all ascertained (retrospectively) upon enrollment, then $w_i = E_i - t_i^{(1)}$ is observable and data consist of a pair (x_i, w_i) for each individual where $x_i = t_i^{(2)} - t_i^{(1)}$. Do data have enough information to estimate the distribution of X without modelling $\lambda(t)$?

The good news is that the hazard function under left-truncation is invariant for $x > w$. For each individual, conditioning on $X > w$, the hazard function calculated from the conditional distribution is

$$h_{\text{left-truncation}}(x|w) = \frac{f(x)/\overline{F}(w)}{\overline{F}(x)/\overline{F}(w)} = \frac{f(x)}{\overline{F}(x)} = h(x), \ x > w.$$

All the survival analysis models and methods focusing on hazard rate estimation and comparison (e.g., proportional hazard regression model, other hazard based models, etc.) apply to data with left-truncation. Existing software may be directly applied with minimum modification (e.g., SAS, S-plus, R, etc.). However, the identifiable part of the survival function is the residual survival function

$$\overline{F}(x|w) = \exp\left(-\int_w^x h(u)du\right), \ x > w, \tag{7.21}$$

not the entire distribution $\overline{F}(x) = \exp\left(-\int_0^x h(u)du\right)$.

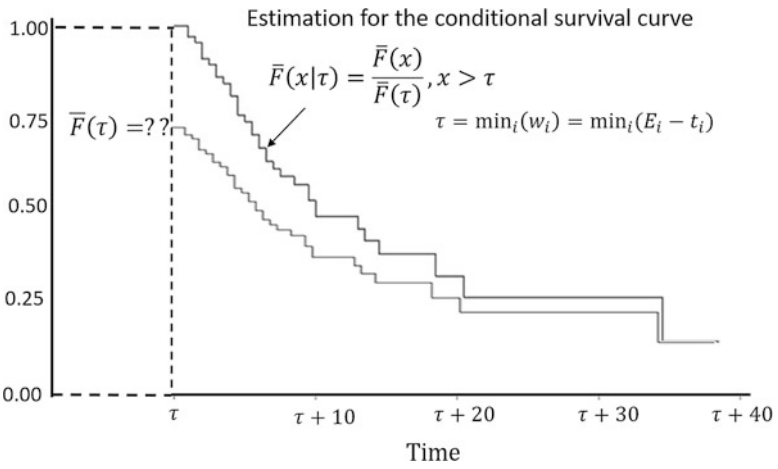


Fig. 7.9 Non-parametrically, left-truncated data are unable to identify the entire survival function. It is only possible to estimate the residual survival function conditioning on $X > \tau$

The above argument can be demonstrated by non-parametric methods. The Kaplan-Meier estimation in survival analysis is based on the empirical estimation of the hazard function at discrete time points. If there is no left-truncation in data, the discrete hazard function yields the empirical survival function as a decreasing step function starting $\overline{F}^{\mathrm{KM}}(0) = 1$. For left-truncated data in pairs (x_i, w_i) and let $\tau = \min(w_i)$, because the hazard function is invariant under left-truncation, the Kaplan-Meier estimation can be still applied (as built into various statistical software packages). The decreasing step function is now understood as the non-parametric estimation for the residual survival function

$$\overline{F}(x|\tau) = \frac{\overline{F}(x)}{\overline{F}(\tau)}, \ x \geq \tau = \min(w_i),$$

starting at $\overline{F}^{\mathrm{KM}}(\tau|\tau) = 1$. The non-identifiable part is $\overline{F}(\tau) \leq 1$. This is illustrated in Fig. 7.9.

There is a rich literature concerning lifetime and life history data with left-truncation. We recommend the books: Lawless (2003) and Cook and Lawless (2007, 2018).

7.3.2 Retrospective Ascertainment and Right-Truncation

As illustrated in Fig. 7.8, in left-truncated data, the inclusion criteria is that, at time of enrollment, the initiating event has occurred but the subsequent event has not. This observation scheme produces time-length bias in favor of longer duration.

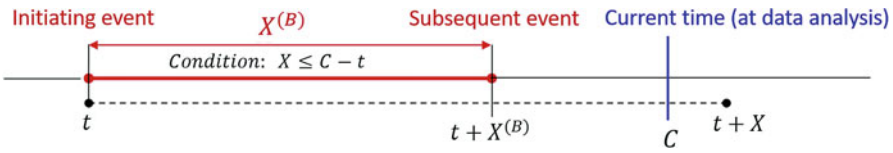


Fig. 7.10 Illustration of the observation scheme with right-truncation. The individual with long duration (dotted line) is not included by time C

The opposite observation scheme is illustrated in Fig. 7.10. The inclusion criteria is based on the occurrence of the subsequent event. Individuals whose initiating events have occurred, but the subsequent events have not, are not included by the time of data analysis. This observation scheme produces time-length bias in favor of shorter duration.

Data arise from the conditional distribution

$$F^*(x|\tau) = \frac{F_X(x)}{F_X(\tau)},\ 0 \le x \le \tau = \max_i \left\{ C - t_i^{(1)} \right\}. \tag{7.22}$$

The probability of inclusion is the cumulative probability $F_X(\tau) = \Pr\{X \le \tau\}$, which itself is the object of the estimation. If τ is sufficiently large, F^* will be a good approximation to F_X. However, sufficiently large τ implies that C is large and one needs to wait for much longer time before starting the analysis. This is not desirable for an emerging infectious disease where one needs information quickly.

Assessment of the Incubation Period Distribution

The incubation period is defined as the duration from the time at infection to the time of onset of clinical symptoms. For acute infectious diseases such as the severe acute respiratory syndrome (SARS) in 2003, knowledge of the incubation period distribution must be generated quickly at the very early stage of the epidemic for guidance to determine the length of quarantine of individuals exposed to infection sources. If a potentially exposed individual has not shown symptoms of the disease after x days of quarantine, the risk of releasing this individual into the susceptible population who subsequently becomes symptomatic (and infectious) is the survivor function of the incubation period. For other infectious diseases that are also chronic, such as HIV/AIDS, the incubation period distribution is the crucial link between an observable event based on clinical presentation such as the diagnosis of AIDS and the unobservable event based on disease transmission such as the infection of HIV.

The incubation period distribution can be estimated using standard survival analysis techniques by following selected cohorts of infected individuals whose dates of exposure to the infection sources are known. However, in emerging new infectious diseases, such as AIDS in the 1980s, SARS in 2003, the pandemic H1N1 (pH1N1) influenza in 2009, knowledge of this distribution must be generated quickly before any formal cohort follow-up studies become feasible.

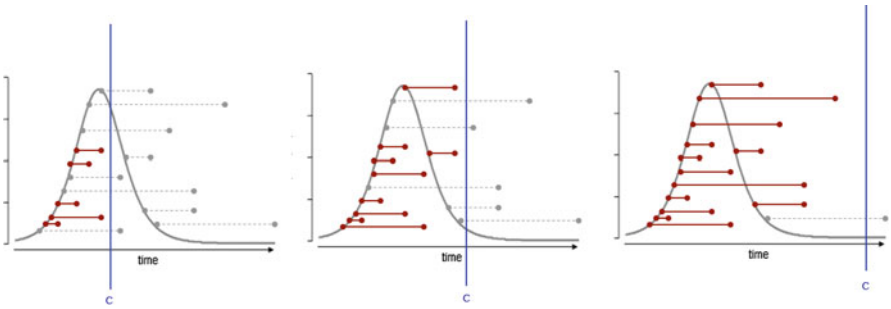


Fig. 7.11 Schematic illustration of retrospectively ascertained incubation periods analyzed during, immediately after and long after a disease outbreak

Early data often arise from selected individuals diagnosed with symptoms with retrospectively ascertained dates of exposure to the infection sources. These early assessments tend to be shorter than assessments made at some later time during the epidemic, which may still be shorter than the actual distribution. This is schematically illustrated in Fig. 7.11. This phenomenon is often misinterpreted by the media and the public to make hypotheses about the mutation of the pathogen, rather than time-length bias in data.

Example 29 This example is well cited in the literature, such as Examples 3.5.3 and 4.3.3 in Lawless (2003). For the incubation period from HIV infection to the development of AIDS illnesses, early studies were based on retrospectively ascertained data for blood transfusion-associated cases, assembled by the U.S. Centers for Disease Control, with transfusion as the only known risk factor. The data were studied by Lui et al. (1986), Medley et al. (1987), Lagakos et al. (1988), Kalbfleisch and Lawless (1989), among many others. A comprehensive survey of various statistical methods of these studies was included in Chap. 4 of Brookmeyer and Gail (1994). Lui et al. (1986) published the earliest results based on data available as of April 1985. The authors acknowledged the bias due to right-truncation and illustrated that the sample average was only 2.6 years based on the naïve approach, whereas based on the conditional distribution (7.22) along with a Weibull distribution model, the estimated mean incubation period was 4.5 years. Kalbfleisch and Lawless (1989) analyzed the data as reported by July 1986 with median estimation approximately 8.5 years.

For right-truncated data regarding the incubation period, uncertainty with respect to the time of infection, $T_i^{(1)}$, is a common problem. Tuite et al. (2010) analyzed 3152 laboratory confirmed pH1N1 cases in Ontario with symptom onset between April 13 and June 20, 2009. A subset of 316 cases containing sufficient information on exposure date and disease onset were used to estimate the incubation period distribution. The dates of exposure were imputed as the midpoint between the earliest and the most recent dates of exposure. Farewell et al. (2005) studied a subset of 128 cases out of a total of 1755 reported cases in a Hong Kong Hospital

Authority database during the 2003 SARS outbreak. The data consist of the date of the appearance of the symptoms of SARS, but the dates of exposure can be only ascertained to an earliest and latest possible date of exposure. The authors explored statistical methodology for retrospective data with the timing of the initiating event being uncertain, except for lying in a given time interval, and what might reasonably be inferred about such a maximum incubation time based on the moderately sized samples that would typically be available in the early course of an epidemic.

Assessment of the Reporting Delay and Estimation of the Number of Occurred But Not Yet Reported Events

In most public health disease surveillance systems, data are compiled upon the *reporting* of the diagnosis of the disease. Official reports often present aggregated counts based on the number of new diagnoses or the disease onset per unit of time. It is thought that these "epicurves" represent, to some degree, the epidemiology of the disease transmission. However, there is a time-length bias in under-reporting: the more recent the diagnosis (or onset), the more severe is the under-reporting. This is reflected by an artificial decline of trend near the end of the time-series. This time-length bias is more profound when data are compiled and analyzed when the outbreak is still ongoing, but it is also more important to present real time trends during the outbreak investigation.

Figure 7.12 illustrates the epicurves of the severe acute respiratory syndrome (SARS) during the spring of 2003 in Canada, Singapore, and Hong Kong, compiled from publicly available information from respective government websites by dates of onset. We immediately see the same phenomenon as illustrated in the right panel of Fig. 7.4. Figure 7.13 illustrates the phenomenon again using epicurves presented by the Ministry of Health of Mexico, by compiling the numbers of onset of symptoms of the H1N1 influenza outbreak in Mexico from April to December of 2009, as officially released by different dates.

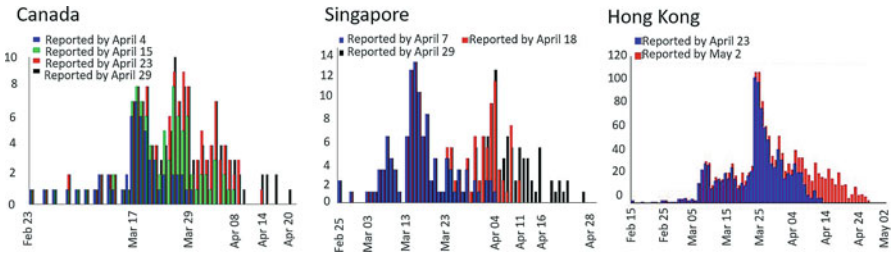


Fig. 7.12 Illustration of SARS Epicurves by dates of onset as reported on different dates during the 2003 SARS outbreak. (Sources: Health Canada; Singapore Ministry of Health; Department of Health, Hong Kong, China)

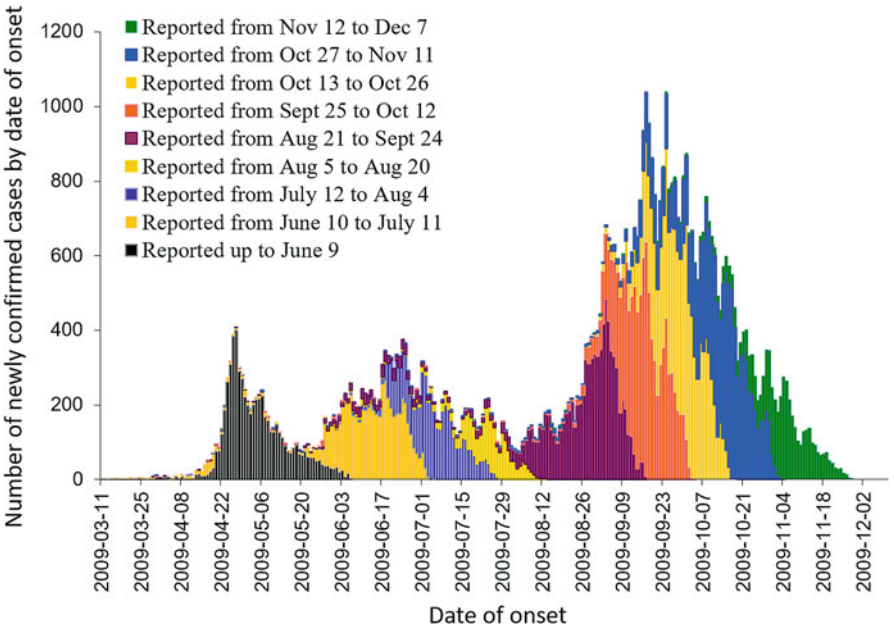


Fig. 7.13 Illustration of the H1N1 influenza outbreak by dates of onset in Mexico as reported on different dates during the 2009 outbreak. (Source: Ministry of Health of Mexico, http://portal.salud.gob.mx/contenidos/noticias/influenza/estatisticas.html)

Such a phenomenon is caused by the reporting delay. Reporting delays are measured at individual levels as the gap between the time of disease onset (or diagnosis) and the time when this individual case is reported and entered into the database of a public health registry.

Let C be the most current time when data are analyzed. We use a generic term "event" for the initiating event, such as disease onset (or diagnosis). The aggregated counts are denoted by

$$N(t; C) = \#\{\text{events occurred at time } t \text{ as reported by time } C\},$$
$$N(t) = N(t; \infty) = \#\{\text{events occurred at time } t\}.$$

$N(t; C)$ is always a proportion of $N(t)$ and the adjustment for reporting delay reduces to the problem of estimating this proportion.

For simplicity, let us assume (for the time being) that the reporting delay can be represented by a random variable X, which is i.i.d. among all individuals. The cumulative distribution is $F(x) = \Pr(X \le x)$. Then

$$F(C - t) = \Pr(X \le C - t)$$

has the same meaning as the probability of events that happened at time $t \leq C$ have been reported by C. Therefore,

$$N(t) = \frac{N(t; C)}{F(C - t)}$$

and the reporting delay adjustment becomes the problem of estimating $F(x)$.

Data on reporting delay is always right-truncated, because only upon reporting can one retrospectively measure the delay (see Fig. 7.10).

There are two levels of time-length biasness involved. Reporting delays produce the time-length bias in aggregated counts over time at the population level with an artificial declining trend near the end of the time-series. The observed reporting delays are also length-biased due to right-truncation. Individuals associated with shorter delays are over-represented in data.

The right-truncation bias in reporting delays is much less recognized than the reporting delay phenomenon at the population level, because frontline workers who make diagnoses, reports, and analyses do not see long delays, even after using naive statistical analyses such as summary statistics directly on observations. Results from formal statistical analysis taking into account right-truncation are often counter-intuitive.

When the first reporting delay adjusted trend of the diagnoses of AIDS in Canada was published in 1993 (Fig. 7.14), it was met with much criticism because it implied a much longer delay most public health workers in the field than felt. It was also dramatically different from a naive analysis based on summary statistics directly calculated from measured delays: median around 1.6 months and only 3% of all individuals were reported after 14 months since diagnoses. The reporting delay

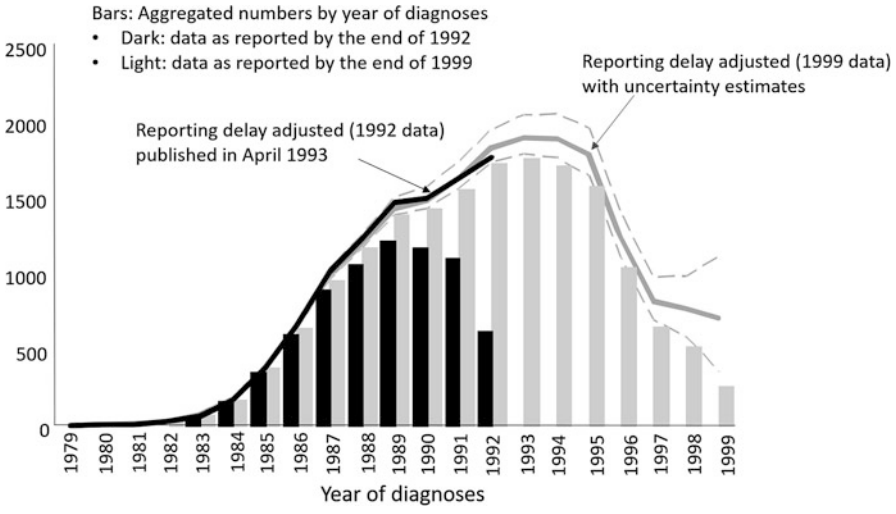


Fig. 7.14 Reporting delay adjusted trends of number of AIDS diagnoses by year from *AIDS in Canada, AIDS Surveillance Report (Health Canada)* in selected years

adjusted trend in Fig. 7.14 based on diagnoses to the end of 1992 suggested a delay with a median of at least 9 months. Despite the doubts and criticisms, history (e.g., reported AIDS incidence by year of diagnosis by the end of 1999) showed that the counter-intuitive delay-adjusted trend was more realistic.

A Simple Method to Estimate Reporting Delay Adjusted Incidence Trends We take a discrete time framework. Let C denote the "current time" which is the time when data are used for analysis. Let $t = 0, 1, 2, \cdots, C$ denote the time of the occurrence of the events under the study where $t = 0$ is the earliest possible time when the events could happen in the population. Let $x = 0, 1, 2, \cdots, C$ denote the report delay and $x = 0$ means that the event is reported within the same time unit. In this setting, data can be grouped into counts

$$n_{tx} = \#\{\text{events occurred at time } t \text{ and reported at time } t + x\},\ x \le C - t.$$

These counts are then arranged into a 2-way contingency table of which the lower triangle remains empty due to right-truncation, as represented in Table 7.1. The row totals are

$$N(t; C) = \sum_{x=0}^{C-t} n_{tx},\ t = 0, 1, \cdots, C.$$

Clearly, as t is getting closer to the current time C, the more likely that $N(t; C)$ under counts the true number of events. The column totals are

$$n_{+x} = \sum_{t=0}^{C-x} n_{tx} = \sum_{i} I(x_i = x).$$

Table 7.1 The upper-triangle table for $\{n_{tx}\}$ with column totals represent the number of events with $X = x$ and row totals represent the number of events over time as reported by C

	Reporting delay x							Row totals
	0	1	$\cdots$	x	$\cdots$	$C-1$	C	
0	n_{00}	n_{01}	$\cdots$	n_{0x}	$\cdots$	$n_{0,C-1}$	n_{0C}	$N(0; C)$
1	n_{10}	n_{11}	$\cdots$	n_{1x}	$\cdots$	$n_{1,C-1}$		$N(1; C)$
$\vdots$	$\vdots$	$\vdots$		$\vdots$				
t	n_{t0}	n_{t1}	$\cdots$	n_{tx}				$N(t; C)$
$\vdots$	$\vdots$	$\vdots$		$\vdots$				
$C-x$	$n_{C-x,0}$	$n_{C-x,1}$	$\cdots$	$n_{C-x,x}$				
$\vdots$	$\vdots$	$\vdots$						
$C-1$	$n_{C-1,0}$	$n_{C-1,1}$						
C	n_{C0}							$N(C; C)$
Col. totals		n_{+1}		n_{+x}			n_{+C}	

representing the total number of events with delay $X = x$ as observed in the data. Let's also denote N_{+x} as the total number of cases with delay $X \leq x$, among events during times $t = 0, 1, \cdots, C - x$, which is the sum of numbers in the 2-way contingency table inside the rectangle area defined by $t = 0, \cdots, C - x$ and $X = 0, \cdots x$:

$$N_{+x} = \sum_{t=0}^{C-x} \sum_{j=0}^{x} n_{tj} = \sum_{i} I(x_i \leq x \leq \tau_i).$$

The ratio

$$g(x) = \frac{n_{+x}}{N_{+x}}, \quad x = 1, \cdots, C$$

gives an estimate for the proportion of events with delay $X = x$ out of those with delay $X \leq x$. Therefore, $1 - g(x)$ gives the estimate for

$$\frac{\#\{X \leq x\} - \#\{X = x\}}{\#\{X \leq x\}} = \frac{\#\{X \leq x - 1\}}{\#\{X \leq x\}}$$

which is the proportion of events with delay $X \leq x - 1$ out of those with delay $X \leq x$. Rewriting $x = C - t + 1$, $1 - g(C - t + 1)$ gives an estimate for

$$\frac{\#\{X \leq C - t\}}{\#\{X \leq [C + 1] - t\}}$$

which is the proportion of events at time t and reported by time C (current), out of those at time t and reported by $C + 1$. This is the estimate for the conditional probability

$$\Pr\{X \leq C - t | X \leq C + 1 - t\} = \frac{F(C - t)}{F([C + 1] - t)}, \quad 1 \leq t \leq C.$$

and hence

$$N(t; C) = N(t; C + 1) \times \frac{F(C - t)}{F([C + 1] - t)} = N(t; C + 1) \times [1 - g(C - t + 1)].$$

Therefore, a one-step prediction for the number of events that occurred at time t as seen in by time $C + 1$, based on current observation $N(t; C)$ is established as

$$\begin{aligned} \widehat{N}(t; C + 1) &= \frac{N(t; C)}{1 - g(C - t + 1)} \\ &= N(t; C) \frac{N_{+,C-t+1}}{N_{+,C-t+1} - n_{+,C-t+1}}, \quad 1 \leq t \leq C. \end{aligned}$$

This also predicts the off-diagonal elements of $\{n_{tx}\}$ in Table 7.1 as

$$\widehat{n}_{t,C-t+1} = N(t;C)\frac{n_{+,C-t+1}}{N_{+,C-t+1} - n_{+,C-t+1}}, \quad 1 \le t \le C.$$

In particular, $\widehat{n}_{1C} = N(1;C)\frac{n_{+C}}{N_{+C}-n_{+C}}$ and $\widehat{n}_{C1} = N(C;C)\frac{n_{+1}}{N_{+1}-n_{+1}}$.

Similarly, it can be shown that

$$\widehat{N}(t;C+2) = \frac{N(t;C)}{[1-g(C-t+1)][1-g(C-t+2)]}, \quad 2 \le t \le C.$$

gives a 2-step prediction for the number of events that occurred at time t as seen by time $C+2$, where the denominator is the estimate for $\Pr\{X \le C-t | X \le C+2-t\} = \frac{F(C-t)}{F([C+2]-t)}$. It further predicts the elements in the lower triangle of Table 7.1 as

$$\widehat{n}_{t,C-t+2} = \widehat{N}(t;C+1)\frac{n_{+,C-t+2}}{N_{+,C-t+2} - n_{+,C-t+2}}, \quad 2 \le t \le C.$$

In particular, $\widehat{n}_{2C} = \widehat{N}(2;C+1)\frac{n_{+C}}{N_{+C}-n_{+C}}$ and $\widehat{n}_{C2} = \widehat{N}(C;C+1)\frac{n_{+2}}{N_{+2}-n_{+2}}$.

Continuing, the maximum is to predict C steps for $t = C$; $C-1$ steps for $t = C-1$; and so on, until all the empty elements in the lower triangle of Table 7.1 are filled by predicted values according to the iterative formulae

$$\widehat{n}_{t,C-t+k} = \widehat{N}(t;C+k-1)\frac{n_{+,C-t+k}}{N_{+,C-t+k} - n_{+,C-t+k}}, \quad k \le t \le C$$

and $k = 1, \ldots, C$. Therefore,

$$\frac{N(t;C)}{[1-g(C-t+1)][1-g(C-t+2)]\cdots[1-g(C)]}, \quad 1 \le t \le C$$

gives the farthest prediction for the number of events that occurred at time $t \le C$ based on current data as seen in the future as data allow, because the longest observable reporting delay is C. The denominator is the estimate for $\Pr\{X \le C-t | X \le C\} = \frac{F(C-t)}{F(C)}$. If it is appropriate to assume $F(C) \approx 1$, the reporting delay adjustment can be written as

$$\widehat{N}(t) = \frac{N(t;C)}{\prod_{x=C-t+1}^{C}[1-g(x)]} \approx \frac{N(t;C)}{\widehat{F}(C-t)}. \tag{7.23}$$

Brookmeyer and Gail (1994) contain a detailed chapter on reporting delays in AIDS surveillance systems. We adopt their example below for illustration of the algorithm.

Table 7.2 The upper-triangle table Table 7.1 filled by numbers from Brookmeyer and Gail (1994) with $C = 4$

Time of diagnosis	Reporting delay x						# diagnoses as reported $N(t;C)$
	0	1	2	3	4		
0	50	20	10	6	2		88
1	100	55	20	12		(88)	187
2	171	115	45		(273)		331
3	207	118		(586)			325
4	220		(836)				220
Col.totals n_{+x} : $x = 1, ..., 4$		308	75	18	2		

Example 30 In Table 7.2, the column totals give $n_{+x} = \sum_{t=0}^{C-x} n_{tx}$:

$$n_{+1} = 308,\ n_{+2} = 75,\ n_{+3} = 18,\ n_{+4} = 2.$$

Numbers in brackets are $N_{+x} = \sum_{t=0}^{C-x} \sum_{j=0}^{x} n_{tj}$:

$$N_{+1} = 836,\ N_{+2} = 586,\ N_{+3} = 273,\ N_{+4} = 88.$$

These quantities yield:

$$g(x) = \left(\frac{308}{836}, \frac{75}{586}, \frac{18}{273}, \frac{2}{88}\right).$$

The reported number diagnoses at the most recent time $t = 4 = C$ is $N(4;4) = 220$. The probability of being reported by time $C = 4$ is

$$[1 - g(1)][1 - g(2)][1 - g(3)][1 - g(4)] = 0.503.$$

According to (7.23), we adjust this number as $\widehat{N}(4) = 220/0.503 = 437$. Similarly, we get

$$\widehat{N}(1) = \frac{N(1;4)}{1 - g(4)} = \frac{187}{0.977} = 191$$

$$\widehat{N}(2) = \frac{N(2;4)}{[1 - g(3)][1 - g(4)]} = \frac{331}{0.913} = 363$$

$$\widehat{N}(3) = \frac{N(3;4)}{[1 - g(2)][1 - g(3)][1 - g(4)]} = \frac{325}{0.796} = 408$$

$$\widehat{N}(4) = \frac{N(4;4)}{[1 - g(1)][1 - g(2)][1 - g(3)][1 - g(4)]} = \frac{220}{0.503} = 437.$$

Caveats This simple method has a few caveats.

1. It only provides partial reporting delay adjustment if $F(C) < 1$. The assumption $F(C) \approx 1$ may be suitable for sufficiently large C so that the system can capture very long delays.
2. It assumes that the reporting delays X_i are i.i.d. among all individuals. This is debatable. It also assumes that the distribution is $F(x)$ is stationary, that is, it does not depend on the time t when the events occur. The latter is mostly untrue in practice because the system can improve, deteriorate, or fluctuate over time. The distribution is most likely to be non-stationary, as $F(x|t)$. There is a rich literature on reporting delay adjustments applied to different disease reporting systems, with statistical models designed to handle non-stationary reporting delays distributions. For example, Kalbfleisch and Lawless (1991) and Lawless (1994).

Likelihood Based Approaches for Analyzing Right-Truncated Data

Here we present formal likelihood based approaches for statistical inferences of the distribution $F(x) = \Pr(X \le x)$ when X is right-truncated.

The reverse hazard function is defined by

$$h_X^r(x) = \frac{f_X(x)}{F_X(x)}, \tag{7.24}$$

where $f_X(x) = \Pr\{X = x\}$ when X is discrete and $f_X(x)$ is the p.d.f. of X when X is continuous. The cumulative distribution function $F(x)$ is uniquely determined by $h_X^r(x)$ through

$$F_X(x) = \begin{cases} \prod_{l=x+1}^{\infty}\{1 - h_X^r(l)\}, & X \text{ discrete} \\ \exp\left\{-\int_x^{\infty} h_X^r(u)du\right\}, & X \text{ continuous} \end{cases}, \quad x > 0.$$

These are analogous to the relationships between the survival function and the hazard function, such as $\overline{F}_n = \prod_{j=0}^{n-1}(1 - h_j)$ (3.1) in discrete case, and $\overline{F}_X(x) = \exp\left(-\int_0^x h_X(u)du\right)$ (2.5) in continuous case.

Lagakos et al. (1988) proposed to use (7.24) as the key quantity for statistical inference for right-truncated data. If $h_X^r(x)$ is identifiable for $0 \le x \le \tau$, then the conditional distribution (7.22) is identifiable through

$$\frac{F_X(x)}{F_X(\tau)} = \begin{cases} \prod_{l=x+1}^{\tau}\{1 - h_X^r(l)\}, & X \text{ discrete} \\ \exp\left\{-\int_x^{\tau} h_X^r(u)du\right\}, & X \text{ continuous} \end{cases}, \quad 0 \le x \le \tau.$$

Suppose that a sample of right-truncated data is represented by (x_i, τ_i), $i = 1, \ldots, n$ subject to the condition $x_i \leq \tau_i$ is observed, in which τ_i is the right-truncation time for individual i. This corresponds to Fig. 7.10, $\tau_i = C - t_i$.

If X is continuous, assuming X follows the distribution $F(x; \underline{\theta})$ which is fully specified by a vector of parameters $\underline{\theta}$, one may consider to model the reverse hazard function parametrically as $h^r_X(x; \underline{\theta})$. One may consider maximizing the (conditional) likelihood given by

$$L(\underline{\theta}) \propto \prod_{i=1}^{n} \frac{f(x_i; \underline{\theta})}{F(\tau_i; \underline{\theta})} = \prod_{i=1}^{n} h^r_X(x_i; \underline{\theta}) \exp\left\{ - \int_{x_i}^{\tau_i} h^r_X(u; \underline{\theta}) du \right\}. \tag{7.25}$$

In the discrete time framework, one may consider the likelihood function

$$L \propto \prod_{i=1}^{n} \frac{f(x_i)}{F(\tau_i)} = \prod_{i=1}^{n} h^r_X(x_i) \prod_{l=x_i+1}^{\tau_i} \{1 - h^r_X(l)\} \tag{7.26}$$

and treat each value $h^r_X(x)$ for $x = 0, \ldots, \tau = \max(\tau_i)$ as a "parameter." In this case, nonparametric method can be only used to estimate $F(x|\tau) = F_X(x)/F_X(\tau) = \prod_{l=x+1}^{\tau} \{1 - h^r_X(l)\}$, $0 \leq x \leq \tau$, because $h^r_X(x)$ is only defined up to τ.

The Non-parametric Maximum Likelihood Estimation Lawless (1994) re-wrote (7.26) as

$$\prod_{i=1}^{n} h^r_X(x_i) \prod_{l=x_i+1}^{\tau_i} \{1 - h^r_X(l)\} = \prod_{x=1}^{\tau} \left[h^r_X(x)^{n_{+x}} \{1 - h^r_X(x)\}^{N_{+x} - n_{+x}} \right]. \tag{7.27}$$

The maximum likelihood estimate is

$$\widehat{h^r}_X(x) = \frac{n_{+x}}{N_{+x}} = \frac{\sum_i I(x_i = x)}{\sum_i I(x_i \leq x \leq \tau_i)}, \; x = 0, 1, \cdots, \tau = \max(\tau_i). \tag{7.28}$$

which has been written as $g(x) = \frac{n_{+x}}{N_{+x}}$ previously in the simple reporting delay adjustment algorithm. Standard multinomial large sample theory provides an estimate of the asymptotic covariance matrix

$$diag\left(\frac{\widehat{h^r}_X(x)\{1 - \widehat{h^r}_X(x)\}}{N_{+x}} \right).$$

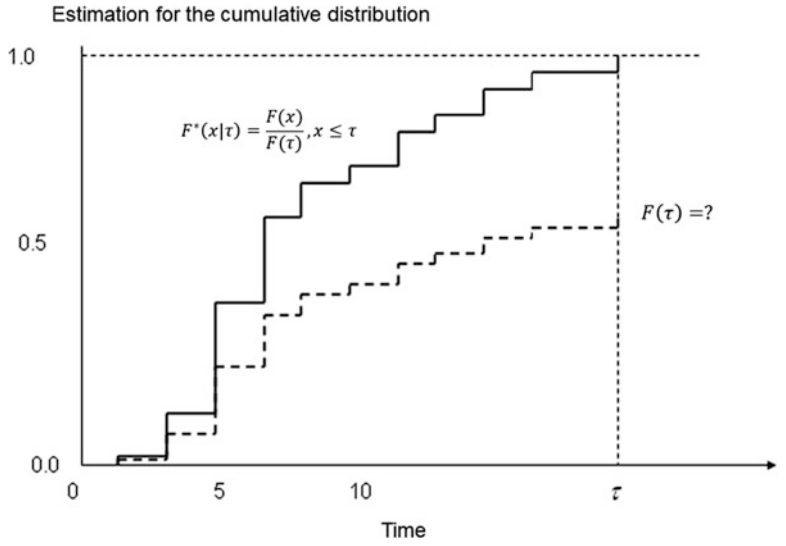


Fig. 7.15 Non-parametrically, right-truncated data are unable to identify the entire cumulative distribution. It is only possible to estimate the conditional distribution conditioning on $X \le \tau$

These yield the estimation for the conditional probability:

$$\widehat{F}^*(x|\tau) = \widehat{\frac{F_X(x)}{F_X(\tau)}} = \prod_{l=x+1}^{\tau}\{1 - \widehat{h^r}_X(l)\} = \prod_{l=x+1}^{\tau}\left(1 - \frac{n_{+l}}{N_{+l}}\right), \quad 1 \le x \le \tau = \max(\tau_i). \tag{7.29}$$

The asymptotic variance estimate (Lawless 1994, 2003) is

$$\widehat{var}\left\{\widehat{F}^*(x|\tau)\right\} = \left\{\widehat{F}^*(x|\tau)\right\}^2 \sum_{x=1}^{\tau} \frac{\widehat{h^r}_X(x)}{N_{+x}\{1 - \widehat{h^r}_X(x)\}}. \tag{7.30}$$

For nonparametric estimation, the identifiable part of $F_X(x)$ is $\{h_X^r(x) : 1 \le x \le \tau\}$. Data do not have information for $\{h_X^r(x) : x = \tau + 1, \cdots, \infty\}$. Therefore, nonparametrically, one cannot fully identify the distribution $F_X(x)$. The best one can achieve is to estimate $F^*(x|\tau)$, $x \le \tau$. Analogous to Fig. 7.9, the non-parametric estimation for $\widehat{F}^*(x|\tau)$ is plotted in Fig. 7.15.

Will a Fully Parametric Model $f(x;\underline{\theta})$ Be Able to Identify the Distribution from Right-Truncated Data? To study this question, we consider a family of the scale-shape distributions where the c.d.f. is defined by $F(x;\lambda,\varsigma) = F_0((\lambda x)^{\varsigma})$, where $F_0(x)$ is a standard distribution not involving unknown parameters with $F_0(0) = 0$ and $F_0(\infty) = 1$, subject to the condition $\lim_{\theta\to 0}\frac{F_0(\theta x^{\varsigma})}{F_0(\theta)} = x^{\varsigma}$ (where $\theta = \lambda^{\varsigma}$). Given the shape parameter ς, if the scale parameter λ is very small, it

approaches a simple power function. In other words, for long underlying durations X, the beginning part of the c.d.f. when $x \in (0, 1]$ behaves like the power function x^{ς}. This family includes the Weibull and the log-logistic distributions.

Now consider a pair of observations (x, τ) subject to $x \leq \tau$. The condition $\lim_{\theta \to 0} \frac{F_0(\theta x^{\varsigma})}{F_0(\theta)} = x^{\varsigma}$ becomes

$$\frac{F(x; \lambda, \varsigma)}{F(\tau; \lambda, \varsigma)} = \frac{F_0\left(\lambda^{\varsigma} x^{\varsigma}\right)}{F_0\left((\lambda\tau)^{\varsigma}\right)} = \frac{F_0\left(\theta\,(x/\tau)^{\varsigma}\right)}{F_0\left(\theta\right)} \to (x/\tau)^{\varsigma}\,, \text{ as } \theta = (\lambda\tau)^{\varsigma} \to 0.$$

Weibull distribution: $F_0(x) = 1 - e^{-x}$.

$$\frac{F_0\left((\lambda x)^{\varsigma}\right)}{F_0\left((\lambda\tau)^{\varsigma}\right)} = \frac{1 - e^{-\theta(x/\tau)^{\varsigma}}}{1 - e^{-\theta}} = \left(\frac{x}{\tau}\right)^{\varsigma} \times$$
$$\left[1 + \frac{\theta}{2}\left(1 - \left(\frac{x}{\tau}\right)^{\varsigma}\right) + \frac{\theta^2}{12}\left(1 - \left(\frac{x}{\tau}\right)^{\varsigma}\right)\left(1 - 2\left(\frac{x}{\tau}\right)^{\varsigma}\right) + O\left(\theta^3\right)\right]$$

Log-logistic distribution: $F_0(x) = \frac{x}{1+x}$

$$\frac{F_0\left((\lambda x)^{\varsigma}\right)}{F_0\left((\lambda\tau)^{\varsigma}\right)} = \left(\frac{x}{\tau}\right)^{\varsigma} \frac{1+\theta}{1+\theta\,(x/\tau)^{\varsigma}}$$
$$= \left(\frac{x}{\tau}\right)^{\varsigma}\left[1 + \theta\left(1 - \left(\frac{x}{\tau}\right)^{\varsigma}\right) + \theta^2\left(\frac{x}{\tau}\right)^{\varsigma}\left(1 - \left(\frac{x}{\tau}\right)^{\varsigma}\right) + O\left(\theta^3\right)\right]$$

In these cases, if $\varsigma > 1$, $\theta = (\lambda\tau)^{\varsigma} \to 0$ implies that for any $\varepsilon > 0$, $\lambda\tau$ must be sufficiently small such that $\lambda\tau < \varepsilon^{1/\varsigma}$. Translating to plain language, it implies that if the truncation time τ is short whereas the underlying distribution of X is long (i.e., small value of the scale parameter λ), the conditional distribution $\frac{F_0((\lambda x)^{\varsigma})}{F_0((\lambda\tau)^{\varsigma})} \approx \left(\frac{x}{\tau}\right)^{\varsigma}$ does not contain λ.

Now we consider a sample of right-truncated data that is represented by (x_i, τ_i), $i = 1, \ldots, n$ subject to the condition $x_i \leq \tau_i$ and let $\tau = \max(\tau_i)$. The above discussion intuitively leads to

1. If the maximum observation window $\tau = \max(\tau_i)$ is relatively short and the underlying distribution for X is long, such that $\lambda\tau \ll 1$, for the above distributions with $\varsigma > 1$, data do not contain enough information about the scale parameter λ.
2. The second and higher order terms of the series expansion of the Weibull and the log-logistic distributions contain the factor $1 - \left(\frac{x}{\tau}\right)^{\varsigma}$. This implies that only a subset of the data such that x_i are neither too close to zero nor too close to $\max(\tau_i)$, but close to $2^{-1/\varsigma} \max(\tau_i)$, may contain some information for λ. This also requires that $\max(\tau_i)$ to be sufficiently large, as well as the specific shape of the underlying distribution.

We demonstrate this through two examples, both related to the estimation of the incubation period based on right truncated data.

Example 31 (Example 29 Continued) Using the subset of 258 adult transfusion associated AIDS by C =June 30, 1986 (Lagakos et al. 1988, Table 1) with $\tau = \max(\tau_i) = 8$ years for the incubation period from HIV infection to the onset of AIDS, data are fitted to a Weibull distribution based on the conditional likelihood (7.25). The estimated shape parameter is $\widehat{\varsigma} \approx 2.1$ with the 95% confidence limits $1.85 \leq \varsigma \leq 2.38$. It implies that the incubation distribution during the first 8 years since infection increases approximately linearly. With respect to λ, data could only provide a one-sided 95% confidence limit $\lambda \leq 0.128$. Together with estimated $\widehat{\varsigma} \approx 2.1$, this gives estimated median incubation ≥ 6.6 years. The contour of the likelihood surface is very flat, which has been shown in earlier discussions (see Fig. 7.3). The upper limit $\lambda^{up} = 0.128$ gives $\lambda^{up}\tau \leq 0.128 * 8 \approx 1$, and hence $\lambda\tau < 1$.

Example 32 This example shows a case $\lambda\tau$ is rather large and both λ and ς are precisely estimated. Dr. Ian Johnson at University of Toronto (personal communication) kindly provided 42 probable SARS cases on April 11, 2003 to assess the incubation distribution. They had been retrospectively ascertained to single exposure dates, ranging from March 6 to March 29, 2003. The longest observable window was $\tau = \max(\tau_i) = 36$ days, from the earliest exposure date March 6 to the time when data are compiled April 11. The maximum observed incubation period in the data was $\max_i\{x_i\} = 10$ days. A log-logistic distribution was used in the conditional likelihood (7.25) which gives the maximum likelihood estimate $\widehat{\lambda} = 0.2395$ and $\widehat{\varsigma} = 3.413$. In this case, $\widehat{\lambda\tau} = 8.622$. We reparameterize the log-logistic distribution in terms of the median λ^{-1} and the 95th quantile. The median was estimated as $\widehat{\lambda}^{-1} = 4.175$ (days) with 95% confidence limits 3.453–5.139 days. The 95th quantile is defined as t_{95} such that $\Pr\{X \leq t_{95}\} = 0.95$. It was estimated as $\widehat{t_{95}} = 9.9$ (days) with 95% confidence limits 6.569–17.211 days. For the goodness-of-fit of the log-logistic model, we compare the cumulative distributions over the parameters ranges of the log-logistic distribution (smooth lines) as well as the non-parametric estimate based on (7.29) as what data suggest. Figure 7.16 illustrates these estimates.

7.4 Some More Discussions About Back-Calculation

Back-calculation has been briefly mentioned twice in this chapter. Once was for the demonstration of the convolution of the systematic component and the link component in a generalized nonlinear model. Another time was for the discussion of the non-identifiability problem.

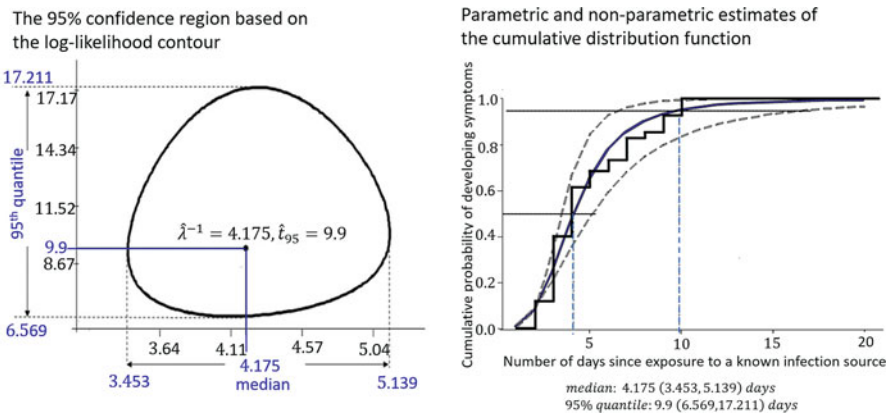


Fig. 7.16 Illustration of fitting the log-logistic distribution to right-truncated incubation times based on a small sample during the 2003 SARS outbreak and comparison with the non-parametric estimate

The convolution, given by

$$\mu(t;\underline{\theta},\underline{\psi}) = \begin{cases} \int_0^t i(u;\underline{\theta}) f(t-u|u;\underline{\psi})du, & \text{continuous time model} \\ \sum_{u=0}^{t} i(u;\underline{\theta}) f(t-u|u;\underline{\psi}), & \text{discrete time model} \end{cases}.$$

is to link the data based on the occurrence of the subsequent observable events from a counting process with mean function $\mu(t;\underline{\theta},\underline{\psi})$ in order to estimate the parameters in the systematic part of the model $i(u;\underline{\theta})$, which is the incidence intensity of the non-observable initial events. The observable events could be onset of clinical symptoms, and the initiating events could be the infection of an agent which then leads to clinical symptoms modelled according to an incubation distribution.

For discussion purposes, let us suppose an ideal situation where for every individual in the data, the time of the initiating event can be back-dated. In this case, the upper triangle matrix as shown in Table 7.1 can be established, in which

$$n_{tx} = \#\{\text{initiating event at time } t, \text{subsequent event at time } t+x\}.$$

where $t = 0, 1, 2, \cdots, C$ denote the time of the initiating event and $x = 0, 1, \cdots, C-t$. In this way, the back-calculation problem is essentially the same as the right-truncation problem as applied in the reporting delay analysis.

The expected values are

$$E[n_{tx}] = \mu(t,x) = i(t) f(x).$$

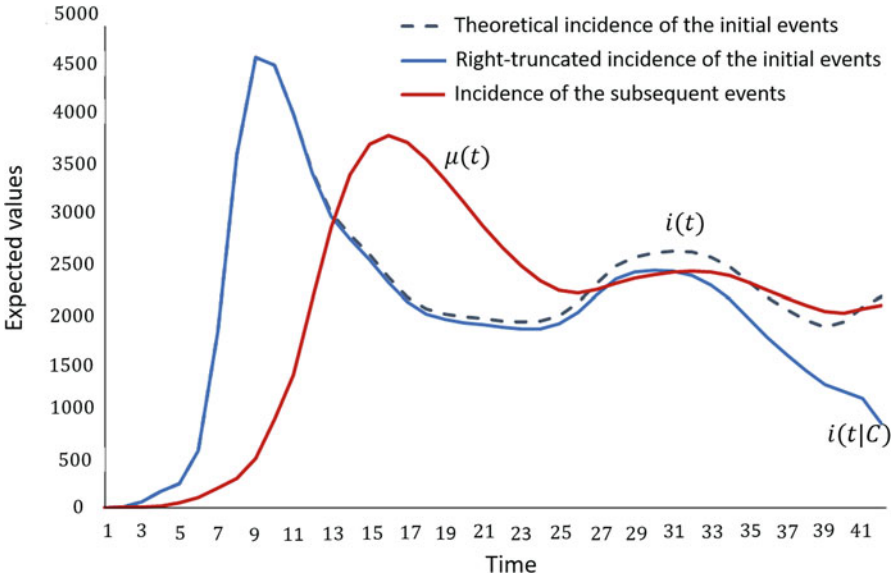


Fig. 7.17 Schematic illustration of the expected counts of the marginal totals $i_1(t|C)$, $i_2(t)$ and the theoretical incidence of the initial events $i_1(t)$ at $C = 41$.

The expected values of the row totals are $E[N(t|C)] \triangleq i(t|C)$, which might be called the right-truncated incidence of the initial events,

$$i(t|C) = i(t) \sum_{x=0}^{C-t} f(x) = i(t)F(C-t), \ t = 0, 1, \cdots, C. \tag{7.31}$$

The expected values of the column totals are

$$\mu(t) = \sum_{s=0}^{t} E[n_{s,t-s}] = \sum_{s=0}^{t} i(s) f(t-s). \tag{7.32}$$

We notice that (7.31) is the model in reporting delay analysis, provided that $F(C-t)$ can be estimated; (7.32) is the model in back-calculation provided that $f(t-s)$ is fully specified. Figure 7.17 conceptually illustrates (7.31), (7.32) and $i(t)$ on the same graph.

Even though one cannot completely identify $i(t)$ from $f(x)$, retrospectively ascertained data still contain some information that one may assess either the trend of the initiating event or the duration distribution with caution.

Since $E[n_{tx}] = \mu(t,x) = [i(t)\,F(C-t)] \times \left[\frac{f(x)}{F(C-t)}\right]$, data (t_i, x_i) are sufficient for the first factor $i(t)\,F(C-t)$. It can be shown that the minimal sufficient statistics for $i(t)F(C-t)$ are the row totals $N(t|C) = \sum_{x=0}^{C-t} n_{tx}$. Conditioning on the row totals, the likelihood function becomes

$$L \propto \prod_{i=1}^{n} \frac{f(x_i)}{F(C-t_i)} = \prod_{i=1}^{n} \frac{f(x_i)}{F(\tau_i)}$$

Therefore the likelihood function (7.26) for analyzing right-truncated data is the conditional likelihood by treating $i(t)$ as the nuisance parameter. As discussed previously, data may not be able to fully identify the distribution $F(x)$ but only up to $\tau = \max(\tau_i)$ as a conditional distribution.

On the other hand, if $F(x)$ is fully specified, then $F(C-t)$ is precisely known. $i(t)$ is estimated through the marginal totals $N(t|C)$ as

$$\widehat{i}(t) = \frac{N(t|C)}{F(C-t)} \tag{7.33}$$

in the same logic as the reporting delay adjustment.

In general, back-calculation methods are developed assuming that each individual in the data only has information on the onset of the subsequent event. Therefore $\{n_{tx}\}$ are not observable and $N(t|C)$ are not observable. The only observable data are the column totals with mean value (7.32).

There is plenty of literature on different back-calculation methods and algorithms, based on continuous time or discrete time models, applied to the studies of HIV/AIDS, viral hepatitis, and many other infectious diseases. We do not intend to write this section about these methods and algorithms, except for the following brief mentioning.

Since the distribution $f(x)$ is fully specified, had the incidence function $i(t)$ been fully specified with all the parameters known, it would have been possible to compute the expected values $E[n_{tx}]$ in each cell of the upper triangle matrix in Table 7.1 based on the observed column totals using a multinomial distribution (Becker et al. 1991). On the other hand, had $\{n_{tx}\}$ been observed, then the back-calculation would have been reduced to a simple algorithm based on (7.33). This is the core of the Expectation-Maximization-Smoothing (EMS) algorithm (Becker et al. 1991) widely used in many back-calculation applications. A generalization of the EMS algorithm based on more than one source of data is given by Yan et al. (2011).

7.5 Problems and Supplements

7.1 Incidence data of the confirmed and probable cases of the Ebola outbreak in the Democratic Republic of Congo (DRC, August 2018–January 2019) are publicly available in the World Health Organization website. Data are manually extracted using WebPlotDigitizer (Rohatgi 2018). In this exercise, data are aggregated as weekly counts and we consider a subset of data with dates of symptom onset starting from August 20, 2018 onwards. We define Week Zero as the week August 20–26, 2018. By the end of Week 9 (i.e., October 28, 2018), there were a total 144 reported cases starting with week of onset at Week Zero. They are cross-tabulated by week of onset and week of report.

Week of onset	Week of report									
	0	1	2	3	4	5	6	7	8	9
0	1	3	2	0	0	0	2	0	0	0
1		3	9	0	0	0	0	0	0	0
2			0	6	0	0	0	0	0	0
3				2	4	1	0	1	0	0
4					0	9	2	0	0	0
5						3	12	2	3	0
6							7	6	13	0
7								0	12	4
8									7	26
9										4

(a) Plot the row totals and the column totals on the same graph and comment on the meaning of these marginal totals.

(b) Reporting delays $x = 0, 1, \ldots$ are calculated by weeks. Diseases that are reported during the same week of the symptoms onset are assigned with $x = 0$. Calculate the frequency of cases with $x = 0, 1, 2, \ldots, 9$ (defined as the numbers of cases with $x = 0, 1, 2, \ldots, 9$ divided by 144) and calculate the cumulative frequency by Week 3. Do you think the reporting delay is that short?

(c) Moving forward, by the end of Week 14 (ending on December 2, 2018) there were a total of 330 reported cases starting with week of onset at Week Zero. The cross-tabulation table is updated. Plot the row totals of the updated table and the row totals of the table ending on Week 9 on the same graph.

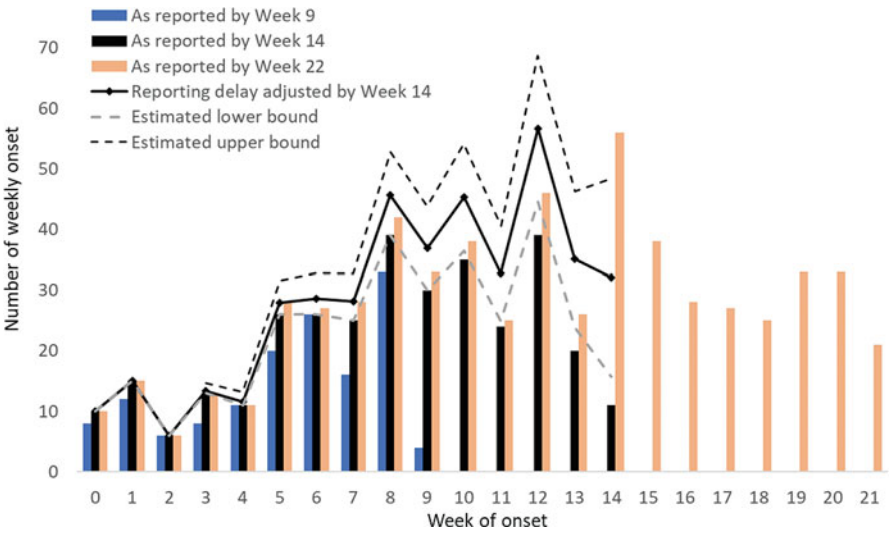
(d) Calculate the frequency cases with $x = 0, 1, 2, \ldots, 14$ (defined as the numbers of cases with $x = 0, 1, 2, \ldots, 14$ divided by 330) and calculate the cumulative frequency by Week 7. Do you think it is because the reporting delay is getting longer or something else?

Wk.of onset	Week of report														
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	1	3	2	0	0	0	2	0	0	0	0	1	0	1	0
1		3	9	0	0	0	0	0	0	0	0	1	2	0	0
2			0	6	0	0	0	0	0	0	0	0	0	0	0
3				2	4	1	0	1	0	0	0	5	0	0	0
4					0	9	2	0	0	0	0	0	0	0	0
5						3	12	2	3	0	0	6	0	0	0
6							7	6	13	0	0	0	0	0	0
7								0	12	4	2	7	0	0	0
8									7	26	1	2	1	2	0
9										4	12	9	3	2	0
10											7	21	3	4	0
11												9	8	7	0
12													22	17	0
13														12	8
14															11

(e) The following table converts the table above to represent n_{tx} as defined in Table 7.1. The column totals represent $n_{+x} = \sum_{t=0}^{C-x} n_{tx} = \sum_i I(x_i = x)$. Use the method in Brookmeyer and Gail (1994) as illustrated in Fig. 7.14 to estimate the weekly number by symptom onset up to the end of Week 9, including cases that were not yet reported.

Week of onset	Reporting delay x														
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	1	3	2	0	0	0	2	0	0	0	0	1	0	1	0
1	3	9	0	0	0	0	0	0	0	1	2	0	0	0	
2	0	6	0	0	0	0	0	0	0	0	0	0	0		
3	2	4	1	0	1	0	0	0	0	0	0	0			
4	0	9	2	0	0	0	0	0	0	0	0				
5	3	12	2	3	0	0	6	0	0	0					
6	7	6	13	0	0	0	0	0	0						
7	0	12	4	2	7	0	0	0							
8	7	26	1	2	1	2	0								
9	4	12	9	3	2	0									
10	7	21	3	4	0										
11	9	8	7	0											
12	22	17	0												
13	12	8													
14	11														
n_{+x}	88	153	44	14	11	2	8	0	5	1	2	1	0	1	0
$N_{+x} = \sum_{t=0}^{C-x} \sum_{j=0}^{x} n_{tj}$															
N_{+x}	88	230	254	229	216	183	161	122	102	77	53	43	30	25	10

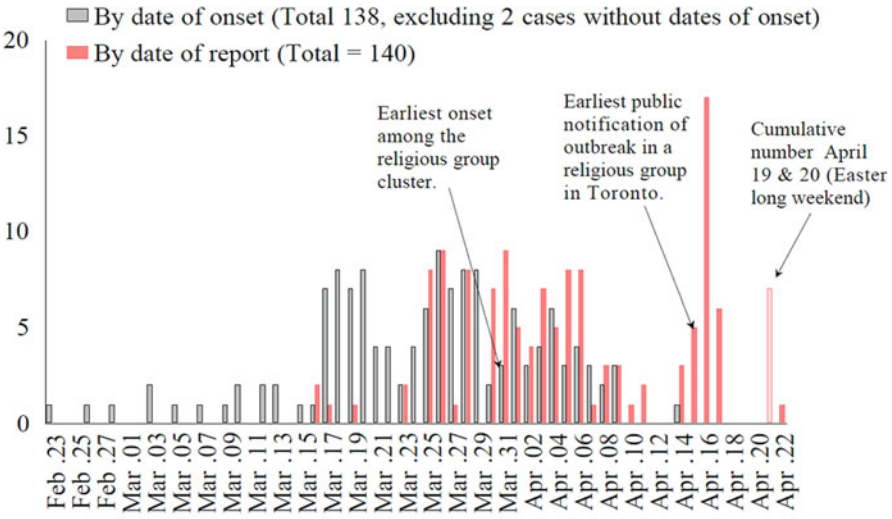
(f) The following figure displays Ebola cases in DRC by week of onset. The bars represent data as reported by Week 9, Week 14, and Week 22. Reporting delays were adjusted using data by the end of Week 14, with point estimates as well as lower and upper 95% confidence limits represented by lines, using the method in Lawless (1994). Consult the original paper and examine the assumptions in the algorithm. Comment on the performance of this method as applied to the Ebola data and discuss potential violations of the assumptions.



7.2 The following figure compares the trends of the 2003 SARS outbreak in Canada based on probable cases reported to Health Canada on April 22, 2003. The dark bars represent the numbers by date of symptom onset whereas the pink bars represent the numbers by date of report. There are several important dates to remember:

- March 13: WHO started worldwide surveillance on atypical pneumonia (later renamed as SARS);
- March 25: SARS became reportable in Canada and surveillance was intensified;
- April 15: health officials made a news release about a new cluster of SARS cases in Toronto related to a religious group;
- April 19–20: Easter long weekend.

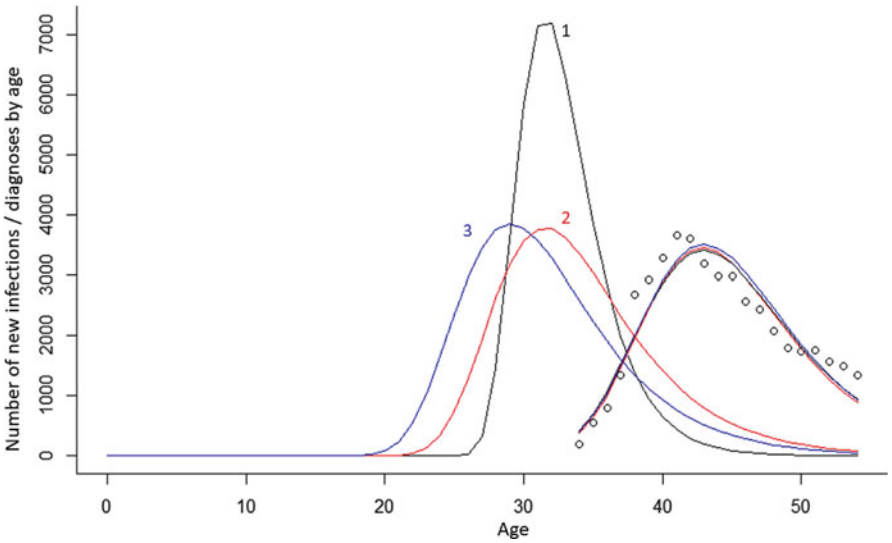
All these events affected the trend based on date of report. Comment on the differences of the trend based on date of onset and date of report. Do you think the trend by date of onset after April 12 was declining? Do you think that the epidemic peaked around April 16? Do you think the simple reporting delay method in Problem 7.5 is suitable?



By date of onset, http://www.hc-sc.gc.ca/pphb-dgspsp/sars-sras/eu-ae/sars20030422_e.html
By date of report, http://www.hc-sc.gc.ca/english/protection/warnings/sars/sars_updates.html

7.3 A case reporting surveillance system is able to document the year of birth as well as the number of new diagnoses of the disease by year with respect to a chronic viral infectious disease. Because the disease natural history is very long (years or decades), the trend of new diagnoses of the disease does not reflect the trend of new infections. There is not much information with respect to the distribution from the time at infection to the time at diagnosis. Empirical evidence has suggested that a log-logistic distribution given by (2.24) is suitable to capture the general shape of such a distribution, with median λ^{-1} and shape parameter ς.

(a) The following figure summarizes some results for a specific birth-cohort. For simplicity, we take year zero to correspond the year of birth and the x-axes in the figure is labelled as age. Surveillance data are shown as circles, starting from age 34 years with peak age in the early 40s. However, auxiliary epidemiologic evidence has shown that the peak of new infections is likely to be in the range between 25 and 34 years of age. The figure shows differently assigned values for λ^{-1} and ς yield differently estimated number of new infections by age/year, but they all provide equally good fit to data. Comment on: in spite of the very different incidence curves (i.e. estimated number of new infections by age/year), is there any feature that is relatively robust with respect to the values of λ^{-1} and ς?



(b) Which of the three incidence curves (labelled as 1, 2 and 3) correspond to which of the following assumptions?

1. a log-logistic distribution with median $\lambda^{-1} = 12$ years and shape parameter $\varsigma = 3.8$;
2. a log-logistic distribution with median $\lambda^{-1} = 12$ years and shape parameter $\varsigma = 10.0$;
3. a log-logistic distribution with median $\lambda^{-1} = 14$ years and shape parameter $\varsigma = 12.0$.