Seroincidence package methodology

European Centre for Disease Prevention and Control (ECDC)

2015-05-22

Contents

1. Introduction

The simplest possible model for serum antibody decay is an exponential decay function (see Appendix). This is the default longitudinal model used in **seroincidence** package. Various modifications can be made, however, to allow for different interpretations of low antibody concentrations in cross-sectional data. The following three approaches are implemented in the current version of the package:

1. Exponential decay: seroconversion is instantaneous, and is followed by exponential decrease to zero (see Appendix for more information).
2. Exponential decay with left censoring: seroconversion is instantaneous, and it is followed by exponential decay, but antibody concentrations below a given (fixed) level are treated as censored (see Appendix for more information). This implies that observations below censoring level are considered to have been generated by seroconversions from at least a given time

in the past (corresponding to the censoring level). In simple terms: for very low antibody concentrations we can only conclude that seroconversion occurred long ago, but not precisely how long.

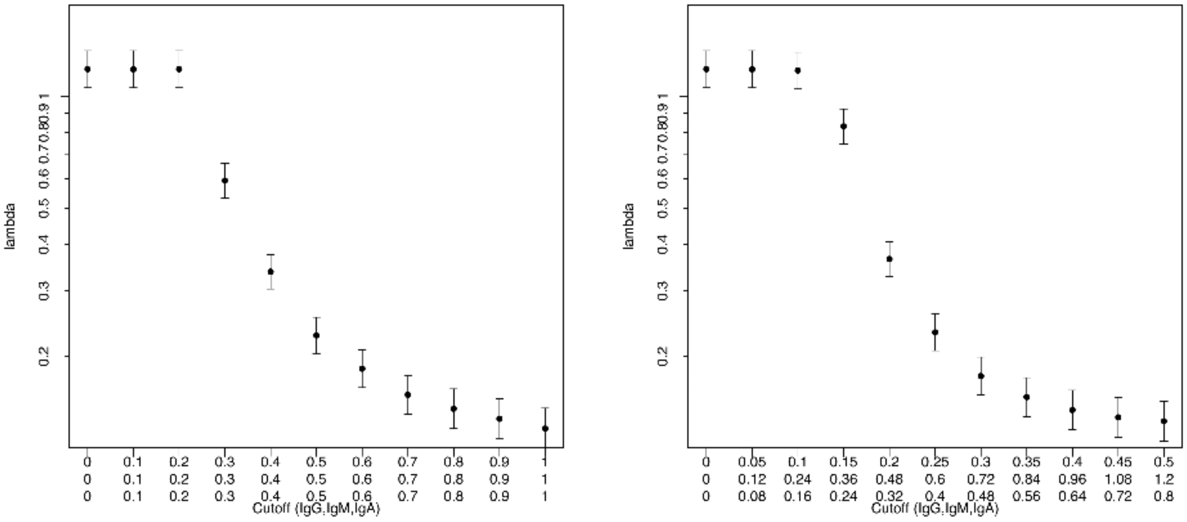
3. Exponential decay with nonzero baseline: seroconversion is again instantaneous, and is again followed by exponential decay, but decay now is not towards zero but instead, approaches some baseline antibody concentration that may be different for any individual (see Appendix for more information). Low antibody concentrations in the cross-sectional sample now may correspond to baseline levels, in which case seroconversion would have occurred an infinite period ago.

Because of their different interpretations of low cross-sectional serum antibody concentrations, both alternatives (left censoring and decay to baseline) lead to decreased estimates of seroincidence, compared to the default model.

Finally, note that the methods described in the Appendix allow for arbitrary antibody decay patterns, so that adaptations can be made to accommodate any longitudinal model for the shape of the decaying arm of the seroresponse.

2. Influence of censoring on seroincidence

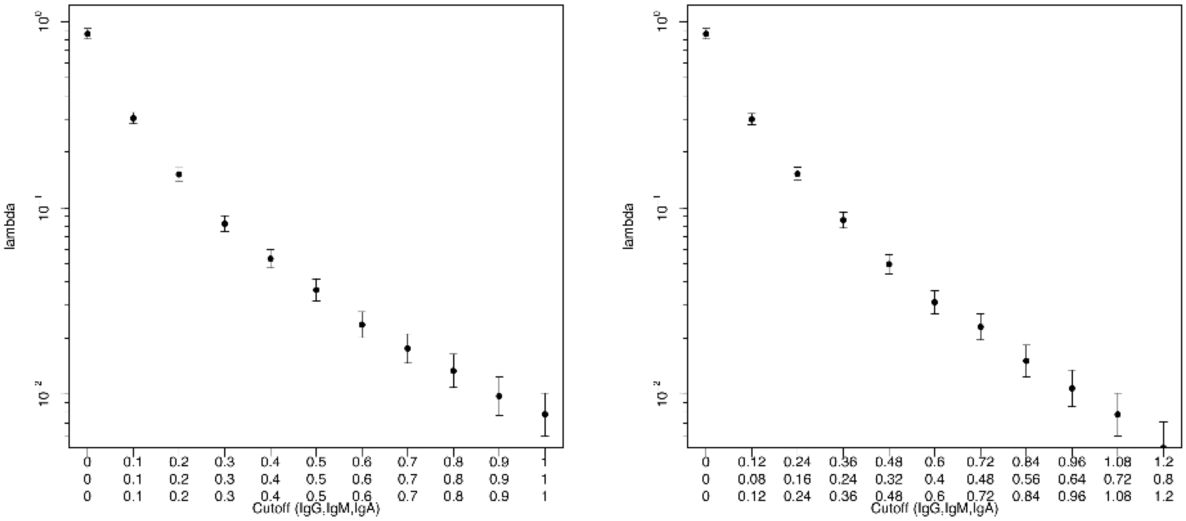
In case serum antibody data are known to be censored, a cutoff level can be defined in the seroincidence calculator script as a list, with separate values for each of the three antibody classes (IgG, IgM, and/or IgA).



(a) Same cutoff

(b) Different cutoffs

Figure 1: Influence of cutoff levels on seroincidence estimates for a cross-sectional sample of *Campylobacter* antibodies. (a) Identical cutoffs for all three antibody types; (b) Different cutoffs for IgG, IgM and IgA.



(a) Same cutoff

(b) Different cutoffs

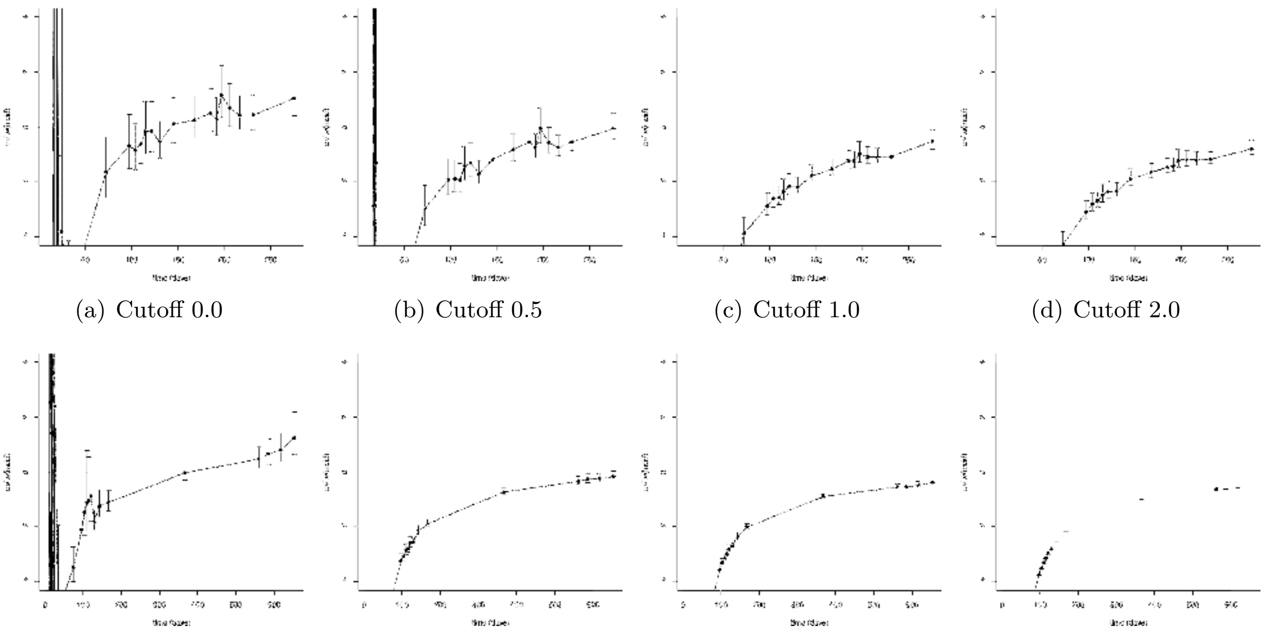
Figure 2: Influence of cutoff levels on seroincidence estimates for a cross-sectional sample of *Salmonella* antibodies (mixed ELISA). (a) Identical cutoffs for all three antibody types; (b) Different cutoffs for IgG, IgM and IgA.

Figures 1 and 2 show the change in seroincidence estimate, for a given set of cross-sectional antibody data (*Campylobacter* and *Salmonella*), with increasing cutoff levels. Cutoffs need not be identical for IgG, IgM and IgA antibodies, as illustrated in these graphs.

3. Checks using the longitudinal data

The longitudinal data may be used to obtain a crude estimate of bias in the estimated incidences (Figure 3a, 3e).

It is seen that for samples taken 100–250 days post symptom onset, the estimated incidences appear unbiased (the 95% interval includes 0). Increasing the cutoff (here applied identical values for all three antibody classes) causes underestimation of the incidence (Figures 3b,c,d and 3f,g,h). The high degree of variation in antibody levels in late samples (more than 100 days post symptom onset) renders estimation problematic.



(a) Cutoff 0.0 (b) Cutoff 0.5 (c) Cutoff 1.0 (d) Cutoff 2.0

(e) Cutoff 0.0 (f) Cutoff 0.5 (g) Cutoff 1.0 (h) Cutoff 2.0

Figure 3: Estimated incidences for longitudinal sera at different times since onset of illness symptoms. delta(incid) = difference between estimated incidence and 365/(time since onset). Top: Campylobacter, Delft ELISA, Bottom: Campylobacter, SSI ELISA.

Appendices

This is a slightly modified description from (Teunis et al. 2012), extended with adaptations for censoring and elevated baseline levels.

A.1. Characteristics of the serum antibody response

The longitudinal model used for describing serum antibody responses has been published (J. Simonsen et al. 2009, Teunis et al. (2012)). As longitudinal data may include early samples taken shortly after infection, possibly during the initial increase in antibody concentrations (during seroconversion) the longitudinal model includes the rising phase of the seroresponse. This avoids misspecification of the peak antibody level, resulting in estimates for the (joint distribution of) peak levels, time to peak, and decay rates of the serum antibody concentration (Teunis et al. 2012, Versteegh et al. (2005)).

However, seroconversion usually is rapid: the increase in antibody levels is orders of magnitude faster than the subsequent decrease. For seroconversion rate calculations we ignore the time to reach peak levels. For simplicity it will be assumed that the antibody response decays exponentially but this is not an essential assumption and it can be relaxed to any monotonically decreasing function of time. So, the antibody response y to incidence at $t = 0$ is an instantaneous increase to peak level (A) followed by an exponential decrease (with rate k):

$$y = A\mathrm{e}^{-kt}$$

A.2. Random sampling of randomly repeating responses

In a cross-sectional sample of responses, each subject will be sampled at a random time within its current inter-event interval Δt (interval, for short). The within–interval sampling distribution is uniform:

$$u_f(\tau|\Delta t) = \frac{[0 \le \tau < \Delta t]}{\Delta t} \tag{1}$$

The subscript f is introduced to show that this distribution is fundamental to sampling from the period between events that are separated by a constant time interval Δt. As the intervals Δt generally will vary, the within–interval sampling distribution and the distribution of intervals combine to a population distribution of backward recurrence times. The backward recurrence time distribution can be translated into a corresponding distribution of response values (antibody concentrations), due to the monotonicity of the response.

The intervals Δt are supposed to vary randomly with distribution $p(\Delta t)$. This introduces a subtle question about the probability of sampling the response within an infinitesimal range of intervals $[\Delta t, \Delta t + \mathrm{d}\Delta t]$. One would be inclined to assume that the probability is given only by the probability density $p(\Delta t)$, but one should keep in mind that responses with longer intervals are more likely to be sampled than those with shorter intervals. Obviously, the probability of the range of sampling intervals $[\Delta t, \Delta t + \mathrm{d}\Delta t]$ should be proportional to the interval Δt itself (size biased sampling, (Feller 1968, Higgins (2008), Scalia Tomba et al. (2010))). Then, the probability of sampling a response within an infinitesimal range of intervals $[\Delta t, \Delta t + \mathrm{d}\Delta t]$ is

$$\frac{p(\Delta t)\Delta t}{\overline{\Delta t}_p}\mathrm{d}\Delta t$$

Here $\overline{\Delta t}_p$ is the expected interval length from the distribution p. We define the incidence as its reciprocal $1/\overline{\Delta t}_p$.

Combining this probability with the uniform distribution of sampling time within the interval Δt leads to a population distribution of time since last event (backward recurrence time)

$$\begin{aligned} u(\tau) &= \int_{\Delta t=0}^{\infty} u_f(\tau|\Delta t)\frac{p(\Delta t)\Delta t}{\overline{\Delta t}_p}\mathrm{d}\Delta t \\ &= \frac{1}{\overline{\Delta t}_p}\int_{\Delta t=\tau}^{\infty} p(\Delta t)\mathrm{d}\Delta t = \frac{1-P(\tau)}{\overline{\Delta t}_p} \end{aligned}$$

where P is the cumulative distribution of p. In going from the first to the second equality we used the fact that Δt should exceed τ (equation (1)). Note that the distribution $u(\tau)$ is monotonically non-increasing.

A.3. Delta distribution

If $p(\Delta t)$ is the delta-distribution $p(\Delta t) = \delta(\Delta t - \Delta t^*)$ representing identical intervals, then the uniform distribution $u_f(\tau|\Delta t^*)$ is regained: for the delta distribution $p(\Delta t) = \delta(\Delta t - \Delta t^*)$ the expected value is $\overline{\Delta t}_p = \Delta t^*$. It follows that

$$u(\tau|\Delta t^*) = \frac{1}{\Delta t^*}\int_{\tau}^{\infty} \delta(\Delta t - \Delta t^*)\mathrm{d}\Delta t = \frac{[0 \leq \tau < \Delta t^*]}{\Delta t^*}$$

A.4. Distribution of response values given the distribution of backward recurrence times

Given incidence at $\tau = 0$, the response decays as $y(\tau) = A\mathrm{e}^{-k\tau}$, with τ backward recurrence time. Consider

$$\begin{aligned} P(y' \leq y) &= P\left(A\mathrm{e}^{-k\tau'} \leq A\mathrm{e}^{-k\tau}\right)[y \leq A] = P(\tau \leq \tau')[y \leq A] \\ &= (1 - U(\tau))\,[y \leq A] = (1 - U(\log(A/y)/k))\,[y \leq A] \end{aligned} \tag{2}$$

Differentiating this expression leads to

$$\begin{aligned} \rho(y) &= \frac{\mathrm{d}}{\mathrm{d}y}P(y' \leq y) = -u\left(\log(A/y)/k\right)\left(\frac{1}{k}\right)\left(\frac{-1}{y}\right)[y \leq A] \\ &= \frac{[y \leq A]}{ky}u\left(\log(A/y)/k\right) \end{aligned} \tag{3}$$

Given the antibody response y with respect to time of incidence, $\rho(y)$ is the corresponding density for sampled response values.

For the nth moment of the backward recurrence time one derives

$$\begin{aligned} E(\tau^n) &= \int_{\tau=0}^{\infty} \tau^n u(\tau)\mathrm{d}\tau = \int_{\tau=0}^{\infty}\left(\tau^n \frac{1}{\overline{\Delta t}_p}\int_{\Delta t=\tau}^{\infty} p(\Delta t)\mathrm{d}\Delta t\right)\mathrm{d}\tau \\ &= \frac{1}{\overline{\Delta t}_p}\int_{\Delta t=0}^{\infty} p(\Delta t)\left(\int_{\tau=0}^{\Delta t}\tau^n \mathrm{d}\tau\right)\mathrm{d}\Delta t \\ &= \frac{1}{\overline{\Delta t}_p}\int_{\Delta t=0}^{\infty} p(\Delta t)\frac{(\Delta t)^{n+1}}{n+1}\mathrm{d}\Delta t = \frac{\overline{(\Delta t)^{n+1}}_p}{(n+1)\overline{\Delta t}_p} \end{aligned}$$

A.5. Family of Gamma distributions

The distribution of intervals $p(\Delta t)$ may have any form. Let the family of $\Gamma(\lambda, m)$ distributions be given:

$$p_m(\Delta t|\lambda) = \frac{\lambda^{m+1}\Delta t^m}{m!}\mathrm{e}^{-\lambda\Delta t}$$

From

$$\int_{x=0}^{\infty} x^{m+1}\mathrm{e}^{-x}\mathrm{d}x = (m+1)!$$

one easily derives that $\overline{\Delta t}_p = (m+1)/\lambda$.

We shall need the cumulative distribution

$$\begin{aligned} P_m(\tau|\lambda) &= \int_{\Delta t=0}^{\tau} p_m(\Delta t|\lambda)\mathrm{d}\Delta t = 1 - \int_{\Delta t=\tau}^{\infty} p_m(\Delta t|\lambda) d\Delta t \\ &= 1 - \frac{\Gamma(m+1,\lambda\tau)}{m!} \end{aligned} \tag{4}$$

where $\Gamma(m+1,\lambda\tau)$ denotes the (upper) incomplete gamma function (not to be confused with the gamma distribution).

Note: working out the integral

$$P_m(\tau|\lambda) = \int_{\Delta t=0}^{\tau} p_m(\Delta t|\lambda)\mathrm{d}\Delta t,$$

it is straightforward to show that

$$\Gamma(m+1,\lambda\tau) = \frac{m!}{\lambda}\sum_{j=0}^{m} p_j(\tau|\lambda) \tag{5}$$

For the distribution of backward recurrence times one finds

$$u_m(\tau|\lambda) = \frac{\lambda}{m+1}\int_{\Delta t=\tau}^{\infty} p_m(\Delta t|\lambda)\mathrm{d}\Delta t = \frac{\lambda}{(m+1)!}\Gamma(m+1,\lambda\tau) \tag{6}$$

when $m = 0$

$$u_0(\tau|\lambda) = \lambda\Gamma(1,\lambda\tau) = p_0(\tau|\lambda),$$

the famous property of the Poisson process.

A.6. Heterogeneity in serum antibody responses

Incorporation of responses that also vary in amplitude A and decay rate k can be easily described with a joint density $g(A,k)$. Of course, such heterogeneity changes the population based distribution of y to

$$\rho(y) = \frac{1}{y}\int_{k=0}^{\infty}\int_{A=y}^{\infty}\frac{1}{k}u\left(\frac{\log(A/y)}{k}\right)g(A,k)\mathrm{d}A\mathrm{d}k$$

instead of corresponding equation (3). Note that the factor $[y \leq A]$ in equation (3) forces the lower limit y in the integration over A. In case the peak antibody levels A and antibody decay rates k are available as a Monte Carlo sample $\{A_n, k_n\}$ of their joint distribution the integration can be approximated by

$$\rho(y) = \frac{1}{y}\frac{1}{N}\sum_{n=1}^{N}\frac{1}{k_n}u\left(\frac{\log(A_n/y)}{k_n}\right)[y \leq A_n] \tag{7}$$

where N is the size of the Monte-Carlo sample, maintaining any correlation between peak levels A_n and decay rates k_n. Assuming e.g. that infections occur with Gamma distributed intervals with rate λ and shape factor m:

$$\begin{aligned}\rho(y|\lambda, m) &= \frac{1}{(m+1)N}\sum_{n=1}^{N}\left(\frac{\lambda}{k_n A_n}\left(\frac{y}{A_n}\right)^{-1+\lambda/k_n}\right.\\ &\times \left.\sum_{j=0}^{m}\frac{(\lambda/k_n\log(A_n/y))^j}{j!}[y \leq A_n]\right)\end{aligned}$$

For a cross-sectional sample $\{Y_1, Y_2, \ldots, Y_{N_c}\}$ a likelihood

$$\ell(\lambda, m) = \prod_{i=1}^{N_c}\rho(Y_i|\lambda, m)$$

can be calculated to allow estimation of the parameters (λ, m) of the process generating infections. The corresponding incidence is the average rate of infections $\lambda/(m+1)$.

A.7. Censored observations

In case observations are censored at y_c such that an observed $Y = \max(Y, y_c)$, then for $y_c < Y$ the density $\rho(y)$ as in eq. (3) holds, but the likelihood of any $Y \leq y_c$ can be calculated

$$\ell(\lambda, m|y \leq y_c) = \int_{z=0}^{y_c}\rho(z|\lambda, m)\mathrm{d}z = R(y_c|\lambda, m)$$

and the likelihood function for a cross-sectional sample $\{Y_1, Y_2, \ldots, Y_{N_c}\}$ becomes

$$\ell(\lambda, m) = \prod_{i=1}^{N_c}\left(\rho(Y_i|\lambda, m)[y_c < Y_i] \times R(y_c|\lambda, m)[Y_i \leq y_c]\right)$$

We need the cumulative distribution $R(y)$ as given by eq. (2)

$$R(y) = 1 - U\left(\frac{\log\left(\frac{A}{\min(y,A)}\right)}{k}\right) = 1 - U\left(\frac{\mathrm{lm}(y, A)}{k}\right) \tag{8}$$

with obvious definition of $\mathrm{lm}(y, A)$.

For the family of gamma distributions for the intervals Δt, eq. (8) leads to

$$R_m(y) = 1 - U_m\left(\frac{\mathrm{lm}(y,A)}{k}|\lambda\right)$$

Substituting eq. (5) into eq. (6), and using the cumulative distribution P_j

$$R_m(y) = 1 - \frac{1}{m+1}\sum_{j=0}^{m} P_j\left(\frac{\mathrm{lm}(y,A)}{k}|\lambda\right)$$

which can also be expressed in terms of incomplete Gamma functions, using eq. (4)

$$R_m(y) = 1 - \frac{1}{m+1}\sum_{j=0}^{m} 1 - \frac{\Gamma\left(j+1, \lambda\frac{\mathrm{lm}(y,A)}{k}\right)}{j!}$$

or, expanding lm

$$R(y) = \frac{1}{m+1}\sum_{j=0}^{m}\frac{1}{j!}\Gamma\left(j+1, \frac{\lambda}{k}\log\left(\frac{A}{\min(y,A)}\right)\right)$$

The latter form is useful for numeric calculations as there is a reliable and efficient implementation available in R.

A.8. Nonzero baseline

The procedure outlined in (Teunis et al. 2012) is valid for any monotonically non-increasing serum antibody response. If for instance, after seroconversion, antibody levels decrease to some baseline level greater than zero (however long ago seroconversion has occurred), the antibody response is

$$y = y_0 + A\mathrm{e}^{-kt}$$

Such responses have been assumed for back-calculations (J. Simonsen et al. 2009).

Now the distribution of response values given the distribution of the backward recurrence times becomes

$$\begin{aligned} P(y' \leq y) &= P\left(y_0 + A\mathrm{e}^{-k\tau'} \leq y_0 + A\mathrm{e}^{-k\tau}\right)[y_0 < y \leq A] \\ &= P(\tau \leq \tau')[y_0 < y \leq A] = (1 - U(\tau))\,[y_0 < y \leq A] \\ &= \left(1 - U\left(\frac{1}{k}\log\left(\frac{A}{y-y_0}\right)\right)\right)[y_0 < y \leq A] \end{aligned}$$

Differentiation produces the density

$$\rho(y) = \frac{\mathrm{d}}{\mathrm{d}y}P(y' \leq y) = \frac{1}{k(y-y_0)}u\left(\frac{1}{k}\log\left(\frac{A}{y-y_0}\right)\right)$$

Including heterogeneity in A, k and y_0 via their joint distribution

$$\rho(y) = \int_{k=0}^{\infty} \int_{A=y}^{\infty} \int_{y_0=0}^{y} \frac{1}{k(y-y_0)} u\left(\frac{1}{k}\log\left(\frac{A}{y-y_0}\right)\right) g(A,K,y_0)\mathrm{d}y_0\mathrm{d}A\mathrm{d}k$$

or, when a Monte-Carlo sample of $\{A_n, k_n, y_{0,n}\}$ triplets is available

$$\rho(y) = \frac{1}{N}\sum_{n=1}^{N} \frac{u\left(\frac{1}{k_n}\log\left(\frac{A_n}{y-y_{0,n}}\right)\right)}{k_n(y-y_{0,n})}[y_{0,n} < y < A_n]$$

For the gamma family of distributions for Δt,

$$\begin{aligned}\rho(y) &= \frac{1}{N(m+1)}\sum_{n=1}^{N}\left(\frac{\lambda}{k_n A_n}\left(\frac{y-y_{0,n}}{A_n}\right)^{-1+\lambda/k_n}\right. \\ &\times \left.\sum_{j=0}^{m}\frac{1}{j!}\left(\frac{\lambda}{k_n}\log\left(\frac{A_n}{y-y_{0,n}}\right)\right)^{j}[y_{0,n} < y \leq A_n]\right)\end{aligned}$$

For the Poisson process ($m = 0$) this reduces to

$$\rho(y) = \frac{1}{N}\sum_{n=1}^{N}\frac{\lambda}{k_n A_n}\left(\frac{y-y_{0,n}}{A_n}\right)^{-1+\lambda/k_n}[y_{0,n} < y \leq A_n]$$

References

Feller, W. 1968. *An Introduction to Probability Theory and Its Applications. John Wiley & Sons.* Vol. 2.

Higgins, P. M. 2008. *Number Story. From Counting to Cryptography. London: Springer Verlag.*

Scalia Tomba, G. P., Å. Svenson, T. Asikainen, and J. Giesecke. 2010. "Some Model Based Considerations on Observing Generation Times for Communicable Diseases." *Mathematical Biosciences* 223 (1): 24–31.

Simonsen, J., K. Mølbak, G. Falkenhorst, K. A. Krogfelt, A. Linneberg, and P. F. Teunis. 2009. "Estimation of Incidences of Infectious Diseases Based on Antibody Measurements." *Statistics in Medicine* 28 (14): 1882–95. doi:10.1002/sim.3592.

Teunis, P. F., J. C. van Eijkeren, C. W. Ang, Y. T. van Duynhoven, J. B. Simonsen, M. A. Strid, and W. van Pelt. 2012. "Biomarker Dynamics: Estimating Infection Rates from Serological Data." *Statistics in Medicine* 31 (20): 2240–48. doi:10.1002/sim.5322.

Versteegh, F. G., P. L. Mertens, H. E. de Melker, J. J. Roord, J. F. Schellekens, and P. F. Teunis. 2005. "Age-Specific Long-Term Course of IgG Antibodies to Pertussis Toxin After Symptomatic Infection with Bordetella Pertussis." *Epidemiology and Infection* 133 (4): 737–48.