EXPONENTIAL DECONTAMINATION MODELS FOR COUNT DATA

By

Richard G. Cornell and Ashok K. Bansal

September 1, 1970

Technical Report Number 22

NASA Grant Number NGR-10-004-029

Department of Statistics
Florida State University
Tallahassee, Florida
32306

EXPONENTIAL DECONTAMINATION MODELS FOR COUNT DATA

Richard G. Cornell
Florida State University
Tallahassee, Florida
U.S.A.

and

Ashok K. Bansal
Delhi University
Delhi, India

Abstract: Several models are developed for the estimation of the rate of exponential die-off from decontamination data. Calculations with illustrative data are reported which indicate that the estimation of this rate and its variance are sensitive to changes in modelling assumptions. Since extrapolation using this estimated rate is used in the specification of planetary quarantine standards, special care should be taken in the selection of an appropriate model and corresponding estimation procedure for the analysis of each set of decontamination data to be used for this purpose.

1. INTRODUCTION AND SUMMARY

This report is concerned with the development of decontamination models based upon the assumption that the probability, θ_t, that a single microorganism is alive at time t after initial exposure to a decontamination procedure is exponential, that is, is of the form

$$\theta_t = \mu^t, \ 0 < \mu < 1. \tag{1.1}$$

Observations are assumed to consist of counts x of the number of organisms alive per unit of test material and corresponding sampling or dilution fractions, d. It is common practice to report only the ratios $y = x/d$, which are the estimated total counts that would have resulted if complete counts had been feasible.

The y ratios, or the x and d values if they are recorded separately, are used to estimate μ. The negative inverse of the logarithem to the base ten of this estimate is called the D-value by microbiologists. Its importance, and hence the importance of accurate and precise estimation of μ, in the specification of space quarantine standards is emphasized in Section 2. The remainder of the paper is concerned with the development of models based on (1.1) which enable the estimation of μ under a variety of assumptions. For each model μ is estimated using the data of Section 3 in order to illustrate the importance of choosing a model and a corresponding method of analysis which incorporate the most realistic assumptions.

The first model presented in Section 4 is the one assumed, at least implicitly, when unweighted least squares calculations are applied to the paired values of t and $v \equiv \log y$. An assumption of the model is that the v values have equal variance, or equivalently when the x's are

similar enough to warrant the assumption that they have equal means, that the x values also have equal variance. The estimation of the variance of $\hat{\mu}$ is particularly sensitive to departures from this assumption in the illustrative example.

An alternative model is presented in Section 5 which utilizes both the x and d values instead of their ratios and which is based on independent binomial distributions for the x counts. For our illustrative data this model leads to a maximum likelihood estimate of μ similar to the the least square estimate for the first model but the estimated variance of this second estimate is considerably less. This shows the importance of dealing with the observed x counts instead of the derived counts, if possible, and of avoiding the assumptions of constant variance of the v logarithms when it is unrealistic as it is shown to be for our example.

A Poisson approximation to the binomial model is presented in Section 6. Similar estimates of μ and variances of these estimates are calculate for these two models. The advantages of this Poisson model are that it leads to simpler maximum likelihood estimation calculations when n is not assumed known than does the binomial model and that it is more easily extended to the situation where n, the initial loading, varies for different exposure times. This extension is devloped in Section 7. For our illustration the estimate of μ is quite different when n is assumed to fluctuate widely than when it is relatively stable, again illustrating the importance of making realistic initial assumptions.

For all but the binomial model an estimate of μ is obtained both with n, or its mean when n is variable as in Section 7, assumed known and when it is estimated from all the data. In the former case the common practice is followed of setting n equal to y_o, the y value based on an initial count with t = 0, and the estimate of μ is obtained using the remainder of the observations. In the latter situation all of the observations, including y_0, are used in the same manner to estimate both n and μ. The two estimates of n and the corresponding estimates of μ differ considerably for each model. This illustrates that n should not be experimentally determined and then assumed known unless it is well determined, that is, unless repeated initial counts have been made.

The assumption of exponential die-off as given by (1.1) is basic to all of the models developed herein and is an assumption often made in analyzing decontamination data. The most common departure is for the rate of die-off, that is, the parameter μ, to change with exposure time t. One way to overcome this difficulty is to apply equation (1.1), and the models derived from it, over only part of the range of t. For instance, rapid early die-off is often ignored and models such as those presented here are then used to describe the data when the rate of die-off is more stable.

Another approach for dealing with departues from assumption (1.1) is to assume a different distribution of the underlying tolerances to exposure of the microorganisms being studied. Assumption (1.1) can be thought of as an assumption that the probability a single microorganism dies before time t is $1 - \mu^t$. This is the cumulative tolerance distribution of a single-parameter exponential distribution.

Thus assumption (1.1) is equivalent to an assumption of exponentially distributed tolerances in the population of microorganisms. The well-known probit and logit methods of analysis have been developed and extensively applied for the more flexible two-parameter normal and logistic tolerance distributions, respectively. Another possible tolerance distribution which is a direct extension of (1.1) is the two-parameter gamma distribution. Research is certainly needed to determine appropriate tolerance distributions for a variety of organisms and types of exposure. However, assumption (1.1) is widely used and holds in many experimental situations, so it is appropriate to concentrate in this report on models derived from that assumption.

2. IMPORTANCE TO SPACE QUARANTINE

Decontamination models are of general interest in microbiology. They are of particular interest in the development of procedures to achieve a high probability of sterilization, for instance, in the food and medical supply industries. However, the main concern which motivated this work is the sterilization of interplanetary spacecraft. International planetary quarantine agreements have been reached which imply that the probability of contaminating Mars must be less than 10^{-r} for a single flight, where r is of the order of magnitude of three. The number of viable contaminants on a spacecraft before final decontamination can be made less than 10^{δ}, where δ is no larger than eight. So it is pertinent to set n_0, the number of viable organisms assumed to be present at the beginning of decontaminttion, equal to 10^{δ} and to determine the exposure time, say

t_r, needed to reduce the probability that the corresponding x is zero to $1-10^{-r}$. For the models in this paper, it is appropriate to set $t_r = (r + \delta)\delta$, where

$$\delta = -1/(\log_{10}\mu) > 0 \qquad (2.1)$$

can be defined from (1.1) as the value of t for which $\theta_t = 0.1$. An estimate of δ, usually called a D-value, can be used to form an estimate $\hat{t}_r$ of t_r given by

$$\hat{t}_r = (r + \delta)\, D. \qquad (2.2)$$

The simplicity of (2.2) explains the practice of microbiologists of thinking in terms of the estimator D of δ instead of in terms of an estimator of μ. Moreover, the use of (2.2) can be thought of as plotting the expected value of n evaluated at $\delta = D$ verses t on semi-logarithmic graph paper and then extrapolating linearly to find the value of t for which this line reaches the sterilization probability standard. The extent of the extrapolation required, which is specified by the multiplier $(r + \delta)$ in (2.2) and which we have indicated could be as large as eleven, means that it is imperative for sterilization applications to estimate δ, or alternatively μ, as well as possible. Once an estimate of μ, denoted by $\hat{\mu}$, is calculated, the corresponding D-value can be computed by replacing μ by $\hat{\mu}$ in (2.1).

3. ILLUSTRATIVE DATA

The parameter μ will be estimated using each of the models presented in this paper for the data presented in Table 1. The t and y values listed

there have been taken from a graph presented by Bruch (1965) in a discussion of dry-heat sterilization at a Spacecraft Sterilization Technology Conference. It has been assumed that the y values were determined by diluting the test material for each t value so that a count of the order of magnitude of 300 or less could be obtained, making such a count and expanding that count by division by the dilution fraction. Counts and dilution fractions which could have led to the values graphed by Bruch have been entered in the x and d columns, respectively, for illustrative purposes. The m column lists values of $m = d_n$ for n set equal to 6.3×10^5, the initial expanded count recorded by Bruch before exposure of his test organisms to dry heat. The first column of trial numbers in Table 1 indexes the exposure times used in the experiment.

TABLE 1

NUMBER (x) OF BACILLUS SUBTILIS VAR. NIGER SPORES RECOVERED PER STRIP VERSUS EXPOSURE TIME (t) IN HOURS AT 120° C TEMPERATURE IN AIR AT ATMOSPHERIC PRESSURE

Trial Number	t	y	d	x	m
0	0	6.3×10^{5}	10^{-4}	63	6.3×10^{1}
1	0.167	4.6×10^{5}	10^{-4}	46	6.3×10^{1}
2	0.500	3.4×10^{5}	10^{-4}	34	6.3×10^{1}
3	0.750	2.5×10^{5}	10^{-4}	25	6.3×10^{1}
4	1.0	1.6×10^{5}	10^{-3}	161	6.3×10^{2}
5	1.5	3.4×10^{4}	10^{-2}	336	6.3×10^{3}
6	2.0	3.4×10^{3}	10^{-2}	34	6.3×10^{3}
7	3.0	1.4×10^{2}	1	136	6.3×10^{5}
8	4.0	3.4×10^{1}	1	34	6.3×10^{5}
9	5.0	2.0×10^{0}	1	2	6.3×10^{5}

4. A LEAST SQUARES MODEL

A simple model can be obtained by noting from (1.1) that $\log \theta_t = t \log \mu$. Since y/n is an estimate of θ_t, where n is the number of organisms exposed initially, $v = \log y$ satisfies, at lease approximately,

$$v = \log n + t \log \mu. \tag{4.1}$$

The initial loading n is assumed to be the same for each t. Moreover, it is assumed that the variance of v is the same for each t. This latter assumption allows the simple unweighted least squares technique to be applied to (v,t) pairs to estimate the slope $\log \mu$ and, if it is not assumed known, the intercept $\log n$. The former estimate is the negative inverse of the D- value.

The assumption of equal variability for the v's is not entirely unreasonable. Each x is of the same order of magnitude, say γ. Assume in addition that each x is a sample from the same population with mean γ. Also assume that each x value has variance γ. This latter assumption is reasonable since it holds for the Poisson distribution. With these assumptions, the mean and variance of y would be γ/d and γ/d^2 and the variance of $v = \log y$ would be approximately equal to $1/\gamma$, a constant independent of d and t.

The least squares estimate of $\log \mu$ is

$$\widehat{\log \mu} = \sum uv / \sum u^2 \tag{4.2}$$

with $u = (t - \bar{t})$ if n is unknown and with $u = t$ otherwise, where $\bar{t}$ denotes the sample mean of the t's.

In the former case, the least squares estimate of log n is

$$\widehat{\log n} = -(\widehat{\log \mu})\bar{t}. \tag{4.3}$$

The variance of the least squares estimate of log μ is

$$V(\widehat{\log \mu}) = V(v)/\sum u^2, \tag{4.4}$$

where V(v) is the variance of v and is estimated by summing the squared deviations of the v's from their estimated values calculated by substituting least squares estimates into the right side of equation (4.1). This sum of squares is then divided by the number of entries in the sum less the number of least squares estimates calculated. After use of (4.4), the estimated variance of $\hat{\mu} = \exp(\widehat{\log \mu})$ is computed by noting that

$$V(\hat{\mu}) \underset{\sim}{\approx} \hat{\mu}^2 V(\widehat{\log \mu}). \tag{4.5}$$

Using all the data in Table 1 in equations (4.2) and (4.3) leads to the estimates $\hat{\mu} = 0.070$ and $\hat{n} = 7.9 \times 10^5$. Setting $n = y_0 = 6.3 \times 10^5$ and using equation (4.2) excluding y_0 yields $\hat{\mu} = 0.082$. The estimated variances of $\hat{\mu}$ for these two situations, calculated from (4.4) and (4.5), are 0.941×10^{-4} and 0.590×10^{-4}, respectively. These illustrative calculations show that quite different estimates of μ, and the variance of this estimate, can be obtained when n is assumed known and equal to y_0 as opposed to when n is estimated from all the data. This difference is particularly noteworthy in view of the extrapolation based on μ to be made in the space quarantine setting as described in Section 2.

In the calculation of the estimated variances of $\hat{\mu}$, estimated variances of v utilizing deviations of the v_i from their estimates were calculated and equalled 0.507 and 0.505 according to whether or not n was set equal to y_0. If instead the variance of v were estimated, in accord with our justification for assuming that the v's have common variance, by $1/\bar{x}$, where $\bar{x}$ is the sample mean of the x's and an estimate of γ, then we would obtain 0.011 in both cases as our estimate of the variance of v. The disparity between these estimates indicates that the assumption of constant variance for v = log y is unreasonable for our illustrative data and that none of these variance estimates is very meaningful since each is calculated assuming a common variance for all of the v's. However, one of the first, and larger, pair would be used, if any, since that pair of variances is based on observed deviations of the v's from their estimates and not just on modelling assumptions and because the x values are not usually available when least squares calculations are used on the y's. However, this comparison illustrates the fact that the assumption of equal variance for the v's upon which commonly used unweighted least squares calculations are based can not be taken for granted.

5. A BINOMIAL MODEL

The reasonableness of the model given by (4.1) depends on the assumption that each x observation has the same mean. This assumption is often regarded as inadequate as it has been shown to be for our illustration even though the x's are usually within a power of ten of each other. A more reasonable assumption, but one which leads to unequal variances for the v's and to a slightly more complicated model and estimation

equations, is that x follows a binomial distribution with each of dn organisms observed at time t independently having probability μ^t of survival. This leads to a model specified by the binomial probability function f(x) for x, where

$$f(x) = \binom{dn}{x}(\mu^t)^x(1-\mu^t)^{dn-x}, \quad x = 0,1,\ldots,dn, \qquad (5.1)$$

and where $\binom{dn}{x}$ is the binomial coefficient representing the number of ways of selecting x organisms from a total available of dn.

The model given by equation (5.1) with both n and each of the sampling fractions assumed known is often encountered in bioassay and epidemiology as documented by Cornell and Speckman (1967) who reference, illustrate and compare several methods for estimating μ. Peto (1953) works out the maximum likelihood procedure which when applied to the data in Table 1 with $n = y_0$ yields $\hat{\mu} = 0.078$, which is very close to the comparable least squares estimate of 0.082 given in Section 4. This method entails maximizing the joint likelihood function which equals the products of the right-sides of equation (5.1) formed using all of the combinations of d, t and x values observed in the experiment. The approximate (asymptotic) variance formula for $\hat{\mu}$, which is also given by Peto, leads to an estimated variance of $\hat{\mu}$ of 0.167×10^{-5}. This is considerably smaller than the corresponding variance estimate cited in Section 4 and illustrates, at least when n is assumed known, that taking into account that the x variables do not all have the same mean and using the actual x observations and corresponding dilution factors d instead of just their y ratios makes more efficient estimation of μ possible. The main drawback to the use of this binomial model instead

of the least squares model of Section 4 is the difficult computations required to estimate both n and μ using all the data. This difficulty is overcome to some extent in the next section.

6. A POISSON MODEL

Since θ_t is usually small relative to nd for $t > 0$ because of the exponential die-off assumed in equation (1.1), the binomial probability model given by (5.1) can alternatively be approximated using the Poisson probability function

$$f(x) = e^{-\lambda}\lambda^x/x!, x = 0, 1, \ldots, \qquad (6.1)$$

where $\lambda = nd\theta_t = nd\mu^t$ is the mean of x given t for the distribution specified by (6.1) as well as by (5.1). This model leads to easier estimation calculations than the binomial model in Section 5, particularly when n as well as μ is to be estimated.

The model given by (6.1) has been introduced by Williams (1961) in the context of a genetic study for equally spaced t values. He develops tables to assist in the maximum likelihood estimation of μ and studies the significance of departures from the model. To apply his work directly to the model depicted by (6.1), the dilution factors d would have to be the same for each t.

A formula which can be solved iteratively for a maximum likelihood estimate of μ in model (6.1) for any set of exposure times, thus avoiding the restrictions in Williams' model, is

$$n \sum d_i t_i \hat{\mu}^{t_i} - T = 0 \qquad (6.2)$$

where $T = \sum t_i x_i$ and n is assumed to be known. When n is not known and $X \equiv \sum x_i$, joint maximum likelihood estimates of μ and n are given by

$$(\sum t_i d_i \hat{\mu}^{t_i} / \sum d_i \hat{\mu}^{t_i}) - T/X = 0; \qquad (6.3)$$

$$\hat{n} = X / \sum d_i \hat{\mu}^{t_i}. \qquad (6.4)$$

A method of solving such equations iteratively which also leads to variance estimates is discussed, for instance, by Rao (1952, pages 165-172).

For the data in Table 1 with $n = y_0$, equation (6.2) yields $\hat{\mu} = 0.077$. Its estimated variance is $\hat{\mu}^2 / m_i t_i^2 \hat{\mu}^{t_i} = 0.163 \times 10^5$. These results are very close to those obtained with the binomial model in Section 5. When n is not assumed to equal y_0, equations (6.3) and (6.4) lead to $\hat{\mu} = 0.056$ and $\hat{n} = 1.435 \times 10^6$. These results indicate the sensitivity of the estimate $\hat{\mu}$ to the value taken for n and also the descrepancies which can arise between an estimate of n based on a single initial count and an estimate based upon all the data. This emphasizes again the need for several observations at $t = 0$ if n is to be experimentally determined and then assumed known in further calculations used to estimate μ.

7. A MODEL WITH VARIABLE LOADINGS

The Poisson model in Section 6 forms a base for easily allowing for the variation in the initial loading n over the tubes prepared for different exposure times. Since n is large in spacecraft decontamination applications, it is reasonable to approximate its distribution by that of a

continuous random variable and since n can not be negative, the gamma distribution is the natural one to use. Its probability density function, with the mean and variance of n equal to α/β and α/β^2, respectively, is

$$f(n) = \beta^{\alpha} n^{\alpha-1} e^{-\beta n}/\Gamma(\alpha), \alpha, \beta, \; n > 0.$$

The gamma distribution is flexible, assuming an exponential form when $\alpha = 1$ and moving through stages of decreasing skewness towards normality as α increases.

Multipling the right sides of equations (6.1) and (7.1) leads to

$$f(x,n) = \beta^{\alpha} (d\mu^t)^x e^{-n(\beta+d\mu^t)} n^{x+\alpha-1}/[\Gamma(\alpha)\, x!].$$

Integrating out n yields a negative binomial probability function for x given by

$$f(x) = \frac{\Gamma(x+\alpha)}{\Gamma(\alpha)x!}\left(\frac{\beta}{\beta+d\mu^t}\right)^{\alpha}\left(\frac{d\mu^t}{\beta+d\mu^t}\right)^{x}, \; x = 0, 1, \ldots; \alpha, \beta > 0. \qquad (7.2)$$

The mean of x is equal to $\alpha d\mu^t/\beta$, which is just the mean λ for the binomial and Poisson models of Sections 5 and 6 with n replaced by its mean α/β.

The parameter μ has been estimated using (7.2) with the mean α/β of n set equal to y_0, the assumed known and fixed value of n used in calculations for previous models. The parameter β has been set successively equal to 1, 1/2, and 0.4×10^{-5}. When $\beta = 1$ the variance and mean of n are equal, as when n has a Poisson distribution. Taking $\beta = 1/2$ makes the variance of n twice its mean, representing slightly more variability than when n has a Poisson distribution. When $\beta = 0.4 \times 10^{-5}$ the variance of n is much larger than its mean and this illustrates the use of this model when there is extremely diffuse information about n. For these three situations the maximum likelihood estimates of μ are 0.078 for $\beta = 1$

and 1/2, as compared to 0.077 for the corresponding Poisson model in Section 6, and 0.089 for $\beta = 0.4 \times 10^{-5}$. This illustrates that allowing for variability in n does not affect the estimate of μ when this variability is relatively small but does when it is large, and hence should be taken into account in this instance through the use of model (7.2) as opposed to those given by (5.1) and (6.1).

The parameters α and β as well as μ in (7.2) can be regarded as unknown and estimated from the data by the maximum likelihood procedure, although the calculations are formidable. For our illustration, the resultant estimates of μ, α and β are 0.064, 4.10 and 0.273×10^{-5}, respectively. The latter two estimates lead to 15.02×10^{5} and 55.01×10^{25} as the estimated mean and variance of n. The point values of these estimates should not be emphasized too much because the likelihood contour is relatively flat around its maximum point, that is, the estimates of μ, α and β could be varied quit a bit without reducing the likelihood of the sample markedly. However, these estimates do indicate that for this example it is not very reasonable to assume that n is the same for each exposure time in the Poisson model in Section 6 or in the binomial model in Section 5 which it approximates. This calculation, like the first one in this section, also illustrates the sensitivity of the estimate of μ to assumptions made about n in developing the model upon which the estimation calculations are based.

ACKNOWLEDGMENTS

Support for this work was received from grant number NGR-10-004-029 from the National Aeronautics and Space Administration. Partial support

for the Florida State University Computing Center, whose facilities were used extensively, was received from the National Science Foundation.

REFERENCES

Bruch, C. W. : 1965, Spacecraft Sterilization Technology, National Aeronautics and Space Administration Special Publication SP-108, Washington, D. C., 207-229.

Cornell, R. G. and Speckman, J. A. : 1967, Biometrics 23, 717-737.

Peto, S. : 1953, Biometrics 9, 320-335.

Rao, C. R. : 1952, Advanced Statistical Methods in Biometric Research, John Wiley and Sons, Inc., New York.

Williams, E. J. : 1961, Biometrics 17, 584-606.