

Model-Based Imputation Approach for Data Analysis in the Presence of Non-detects

K. KRISHNAMOORTHY^{1*}, AVISHEK MALLICK¹ and THOMAS MATHEW²

¹*Department of Mathematics, University of Louisiana at Lafayette, Lafayette, LA 70504, USA;*

²*Department of Mathematics and Statistics, University of Maryland, Baltimore, MD 21250, USA*

Received 19 June 2008; in final form 24 November 2008

A model-based multiple imputation approach for analyzing sample data with non-detects is proposed. The imputation approach involves randomly generating observations below the detection limit using the detected sample values and then analyzing the data using complete sample techniques, along with suitable adjustments to account for the imputation. The method is described for the normal case and is illustrated for making inferences for constructing prediction limits, tolerance limits, for setting an upper bound for an exceedance probability and for interval estimation of a log-normal mean. Two imputation approaches are investigated in the paper: one uses approximate maximum likelihood estimates (MLEs) of the parameters and a second approach uses simple *ad hoc* estimates that were developed for the specific purpose of imputations. The accuracy of the approaches is verified using Monte Carlo simulation. Simulation studies show that both approaches are very satisfactory for small to moderately large sample sizes, but only the MLE-based approach is satisfactory for large sample sizes. The MLE-based approach can be calibrated to perform very well for large samples. Applicability of the method to the log-normal distribution and the gamma distribution (via a cube root transformation) is outlined. Simulation studies also show that the imputation approach works well for constructing tolerance limits and prediction limits for a gamma distribution. The approach is illustrated using a few practical examples.

Keywords: confidence interval; exceedance probability; left-censored data; prediction limits; quantiles; tolerance limits; Wilson–Hilferty approximation

INTRODUCTION

While analyzing exposure data or environmental data, a problem frequently encountered by practitioners is that the samples contain non-detectable amounts of the contaminant under study. The definition of the detection limit (DL) of an instrument or a method varies across different industries. For example, the definition commonly used in radiological analysis laboratories is ‘the minimum concentration of a substance that can be measured and reported with 99% confidence that the analyte concentration is greater than zero’. In occupational hygiene $3 \times$ standard deviation of the sampling or analytic method at low levels of the true concentration is defined as a limit of detection (LOD), used as a method-estimate limit at which the false-positive rate for

assertions of a substance’s presence when actually absent is under control. A $DL = 2 \times LOD$ is sometimes taken as the true concentration at which false-negative rate is well controlled. Although the definition of the DL varies among different areas of application, all of them essentially amount to saying that the DL is ‘the lowest concentration of a chemical that can reliably be distinguished from a zero concentration’, as stated by the US Environmental Protection Agency (USEPA, 1984). The DL of an instrument or a method is usually estimated by a sample of readings obtained using the instrument or method; however, for simplicity and convenience, the DL is commonly assumed to be a known fixed quantity while analyzing the samples containing non-detects. For further details, see the books by Gibbons and Coleman (2001, Chapters 5–7) and Helsel (2005, Chapter 3).

Statisticians recognize the problem of analyzing data with values below a DL as one dealing with censored data or, more specifically, as one involving ‘Type I singly left censored samples’. That is,

*Author to whom correspondence should be addressed.
Tel: +337-482-5283; fax: +337-482-5346;
e-mail: krishna@louisiana.edu

a censoring value x_0 is specified, and the measurements that fall below x_0 are not included in the sample. Assuming normality, Cohen (1959) derived the maximum likelihood estimates (MLEs) for normal parameters, based on censored data. As the MLEs are solutions of some complex equations, they are not easy to compute. Cohen (1961) and Schmee *et al.* (1985) provided some table values that facilitate the computation of MLEs. From a practical point of view, the MLEs by themselves are of limited use; confidence intervals (CIs) and hypothesis tests are certainly required.

Faced with a sample that contains values below the DL, a common strategy among practitioners is to use a substitution method, i.e. replace each non-detectable value by $DL/2$, $DL/\sqrt{2}$, $2DL/3$ or even zero, and then use the standard procedures available for complete sample analysis. The Code of Federal Regulations (CFR) reporting requirements for lead and copper levels, in the context of tap water monitoring, states that ‘All levels below the lead and copper MDLs must be reported as zero’ [here MDL stands for the *Method Detection Limit*; see 40 CFR, Chapter 1, p. 478 (7-1-02 edn)]. Such an approach maybe good enough for practical accuracy if the percentage of measurements less than the DL is very small. However, there are many practical situations where the percentage of measurements less than the DL is substantial. It appears that there is no rationale that justifies the common practice of replacing each non-detectable value by a fraction of the DL; more importantly, the resulting inference procedures can be very inaccurate. Simulation results reported in Lubin *et al.* (2004, Table 3) show that a CI for the normal mean, obtained after replacing each value below the DL by $DL/2$, can have coverage probabilities well below the nominal level. Recently, Hewett and Ganser (2007) compared several methods for estimating the 95th percentile of right-skewed exposure distributions. These authors, based on extensive simulation studies, concluded that the ML method is one of the best while the substitution method (replacing the non-detects by a fraction of the DL) is the worst with respect to mean squared error criterion.

We carried out some preliminary work to assess the performance of the substitution method of replacing the values below the DL by $DL/2$. For this, we assumed a $N(0, \sigma^2)$ distribution, and assumed that $DL = \mu - \sigma = -\sigma$ (for simulation purposes). Note that we have chosen DL such that $\sim 16\%$ of the population data are less than the DL. Monte Carlo estimates (based on 100 000 simulation runs) of the coverage probabilities of the CIs for μ and of one-sided confidence limits for the percentiles are given in Table 1. It should be noted that the one-sided $100(1 - \alpha)\%$ upper confidence limit for the $100p$ th percentile is referred to as the p -content $-1 - \alpha$ coverage upper tolerance limit, and similarly,

Table 1. Coverage probabilities of normal mean and percentiles when non-detects are replaced by $DL/2$; (a) 95% confidence interval for the mean, (b) 95% lower confidence limit for the 5th percentile and (c) 95% upper confidence limit for the 95th percentile; percentage of non-detects is 16

n	10	20	30	40	50	60	70
σ	1	2	2	1	2	2	1
a	0.94	0.89	0.84	0.79	0.73	0.67	0.61
b	0.81	0.55	0.32	0.17	0.09	0.03	0.02
c	0.84	0.80	0.76	0.72	0.69	0.66	0.63

$100(1 - \alpha)\%$ lower confidence limit for the $100(1 - p)$ th percentile is referred to as the p -content $-1 - \alpha$ coverage lower tolerance limit. These one-sided tolerance limits for a normal distribution are given in equation (10) and the material following equation (10).

We observe from Table 1 that the coverage probabilities of the CI for a normal mean based on the substitution method could go as low as 61% when the nominal level is 95%. Furthermore, we see that the coverage probabilities of confidence limits for the 5th percentiles are in general much lower than the nominal level, they could go as low as 2% when the nominal level is 95%. The performance of the substitution method for finding an upper confidence limit for the 95th percentile is more satisfactory compared to that for the lower percentiles; still the coverage probabilities could go as low as 63%.

We also evaluated similar coverage probabilities when $DL = -\sigma/2$ so that the percentage of data in the population below the DL is 31. In this case, the coverage probabilities of CIs for a normal mean could be very small. For example, when $n = 50$, $DL = -\sigma/2$ and the coverage probability of the usual 95% CI based on the substitution method (substituting $DL/2$ for non-detects) is $\sim 15\%$. Thus, the widespread substitution method could produce inaccurate results, and the conclusions drawn from such results can be seriously flawed.

An obvious methodology for data analysis in the presence of non-detects is to analyze the sample data above the DL using the available methodology for censored data analysis. In particular, tobit analysis can be used for this purpose; see Amemiya (1984). However, for obtaining CIs or test procedures, one may have to rely on large sample theory, and the methodology may be unsuitable when the sample sizes are small. Wild *et al.* (1996) considered non-parametric bootstrap, parametric bootstrap, hybrid bootstrap and Gibbs sampling methods for finding interval estimates for an exceedance probability based on a left censored sample from a log-normal distribution. Their limited numerical studies indicate that the Gibbs sampling is the only satisfactory method. However, the problem addressed by Wild *et al.*

(1996) is only for the estimation of an exceedance probability.

In the present article, our goal is to present a unified approach for addressing a variety of data analysis issues in the presence of non-detects, for normal and related distributions. We have investigated an imputation approach that is particularly suitable for sample sizes that are small or only moderately large. The basic idea is quite simple, and has been used earlier; see, for example, Lubin *et al.* (2004), where imputation is in fact recommended to deal with samples that contain below DL values (see also Baccarelli *et al.*, 2005). The approach is one of the substitution; i.e. replace the values below the DL by imputed values. The imputation is done subject to the condition that the values to be imputed are below the DL, and the imputation also uses parameter estimates based on the data that are above the DL. The data analysis is then carried out using the methods available for complete samples. This substitution method is actually repeated several times to have several imputed samples, and the results based on these imputed samples are combined or averaged. This will guarantee that the final results do not depend on a specific imputed value.

The methodology developed in this article is specifically for the normal, log-normal and gamma distributions. The problems addressed in this article include the computation of CIs, tolerance intervals, prediction intervals and CIs for exceedance probabilities. The degrees of freedom that is used while implementing our procedure (for example, for obtaining a CI) is adjusted to reflect the fact that we do not have a complete sample. We have first developed our procedures for the normal distribution. Application to the log-normal distribution is quite straightforward. For a gamma distribution also, we have used the normal-based procedures in view of the observation that the cube root of a gamma random variable is well approximated by a normal random variable; this approximation is due to Wilson and Hilferty (1931).

In this article, two imputation approaches are investigated: one is based on the MLEs of the parameters and the second approach uses some *ad hoc* estimates. Monte Carlo simulation is used to investigate the performance of our procedures. For a given percentage of values below the DL, our proposed imputation approach exhibits excellent performance when the sample size is small to moderately large. However, as the sample size gets large, the *ad hoc* procedure performs poorly; but the MLE-based procedure continues to perform reasonably well unless the sample size gets very large. This may appear contradictory; however, it appears that the performance of the imputation approach depends on the 'actual number' of sample values below the DL and not on the percentage of such values. In other words, if we fix the percentage of values below the DL, the num-

ber of values below the DL goes up with the sample size, and this affects the performance of the imputation approach. However, the confidence levels can be calibrated so that the MLE-based imputation approach continues to provide coverage probabilities close to the nominal level. In the paper, we have also included tables of such calibrated confidence levels; these depend only on the sample size.

Throughout the paper, we have reported a number of examples in order to illustrate the simplicity and applicability of the proposed imputation approach. The examples and the simulation results point out an interesting fact: in situations where the imputation approach is satisfactory, the expected length of the CIs as well as the expected value of the tolerance limits are very close to what one would obtain based on a complete sample even in cases where percentage of values below the DL is as big as 50%. The numerical results and the examples suggest that the imputation approach is a simple and accurate methodology to deal with the problem of non-detects in samples from normal, log-normal and gamma populations.

Even though the imputation approach is discussed in Lubin *et al.* (2004), we would like to point out some important differences between their work and the present work. First of all, Lubin *et al.* (2004) use the tobit regression approach to estimate the unknown parameters, whereas we use the likelihood of the censored observations. Estimates based on tobit regression has to be numerically obtained; however, a simple approximation is available for computing the censored data MLEs. Along with the imputation, Lubin *et al.* (2004) have also used the bootstrap to implement their methodology. In this article, we present methods which simply use the standard formulas for CIs, prediction intervals, tolerance intervals etc., once the imputation has been done. However, we do adjust the degrees of freedom in order to reflect the fact that we do not have a complete sample. In their recent article, Baccarelli *et al.* (2005, p. 905) point out that 'An additional advantage of imputation methods is that, once distribution-based values are imputed for the non-detectable values, any statistical method appropriate for complete data can be used.' Our results confirm this, except that the degrees of freedom has to be adjusted in order to improve performance. The procedures developed in this article are quite accurate regardless of the sample size, except that the confidence level needs to be calibrated for large samples.

THE IMPUTATION APPROACH

Consider a sample of size n from a continuous distribution with cumulative distribution function (cdf) $F(x; \theta)$, where θ is a vector parameter. The general idea behind the model-based imputation approach is as follows. Suppose $n - k$ of the sample values

are above the DL and k of them are below the DL. To make inferences about the distribution, we can impute the values below the DL to form a complete sample and then use an available approach applicable to the complete sample case. Toward this, we note that the distribution of the measurements below the DL is given by

$$P(X \leq x | X < \text{DL}) = \frac{F(x; \theta)}{F(\text{DL}; \theta)}.$$

Let $\hat{\theta}$ be an estimate of θ based on the $n - k$ measurements above the DL, along with the DL. Then, $\hat{P}_{\text{DL}} = F(\text{DL}; \hat{\theta})$ is an estimate of $P_{\text{DL}} = F(\text{DL}; \theta)$. We can impute the values below the DL by random numbers generated from the estimated distribution function. Let

$$x_i^* = F^{-1}(u_i \hat{P}_{\text{DL}}; \hat{\theta}), \quad i = 1, \dots, k,$$

where u_i s are independent uniform (0, 1) random numbers. Notice that x_i^* are random numbers from the estimated distribution function $\frac{F(x; \hat{\theta})}{F(\text{DL}; \hat{\theta})}$ and are all less than or equal to DL because $x_i^* = F^{-1}(u_i \hat{P}_{\text{DL}}; \hat{\theta}) \leq F^{-1}(\hat{P}_{\text{DL}}; \hat{\theta}) = \text{DL}$. Inference about θ , or the population, can be made by considering the measured values above the DL, along with the imputed values, as a complete sample and then applying available procedures for complete samples. To avoid the variability between the results based on different imputations, one should combine results from multiple imputations.

In the following sections, we describe the imputation procedures for the normal and related distributions.

THE NORMAL CASE

MLEs for normal parameters based on left censored data

When the sample includes values below the DL, we have Type I singly left censored samples. Based on such a sample from a normal distribution, Cohen (1959, 1961) derived the MLEs for the mean μ and variance σ^2 , which can be computed numerically as solutions of some non-linear equations. To present the equations, let k denote the number of observations below the DL in a sample of size n and let x_i ($i = 1, 2, \dots, n - k$) denote the observations above the DL. Define

$$\xi = \frac{\text{DL} - \mu}{\sigma}, \quad Z(\xi) = \frac{\Phi(\xi)}{1 - \Phi(\xi)} \quad \text{and} \quad (1)$$

$$Y(h, \xi) = \frac{hZ(-\xi)}{1 - h},$$

where ϕ and Φ denote, respectively, the density function and distribution function of the standard normal random variable and $h = k/n$. Let

$$\bar{x}_l = \frac{1}{n-k} \sum_{i=1}^{n-k} x_i \quad \text{and} \quad s_l^2 = \frac{1}{n-k} \sum_{i=1}^{n-k} (x_i - \bar{x}_l)^2,$$

where x_i s are the measured values above the DL. The MLEs of μ , σ^2 and ξ are the solutions of

$$\begin{aligned} \mu &= \bar{x}_l - \lambda(h, \xi)(\bar{x}_l - \text{DL}), \\ \sigma^2 &= s_l^2 + \lambda(h, \xi) (\bar{x}_l - \text{DL})^2 \\ \frac{1 - Y(h, \xi)(Y(h, \xi) - \xi)}{(Y(h, \xi) - \xi)^2} &= \frac{s_l^2}{(\bar{x}_l - \text{DL})^2}, \quad (2) \end{aligned}$$

where $\lambda(h, \xi) = Y(h, \xi)/[Y(h, \xi) - \xi]$. Let $\hat{\xi}$ be the solution of the third equation in equation (2). Then, the MLEs of μ and σ^2 can be computed by substituting $\lambda(h, \hat{\xi})$ for $\lambda(h, \xi)$ in the first two equations of equation (2).

Let $g = s_l^2 / (\bar{x}_l - \text{DL})^2$. As $\hat{\xi}$ is implicitly a function of g , we can write the first two equations of equation (2) as

$$\begin{aligned} \hat{\mu} &= \bar{x}_l - (\bar{x}_l - \text{DL})\lambda(g, h) \quad \text{and} \\ \hat{\sigma}^2 &= s_l^2 + (\bar{x}_l - \text{DL})^2 \lambda(g, h). \end{aligned} \quad (3)$$

Schmee *et al.* (1985) have provided table values of $\lambda(g, h)$ for values of g up to 10, and $h = .1(.1).9$. Hass and Scheff (1990) have developed an approximation to λ that fits the table values of Schmee *et al.* (1985) within 6% relative error; the approximation is given by

$$\begin{aligned} \ln \lambda(g, h) &\simeq \frac{0.182344 - 0.3756}{g + 1} + 0.10017g + 0.78079y \\ &\quad - 0.00581g^2 - 0.06642y^2 \\ &\quad - 0.0234gy + 0.000174g^3 \\ &\quad + 0.001663g^2y - 0.00086gy^2 \\ &\quad - 0.00653y^3, \end{aligned} \quad (4)$$

where $y = \ln \frac{h}{1-h}$.

Imputation method based on the MLEs: inference for the normal mean and variance

We shall now describe the imputation method for a normal distribution with mean μ and variance σ^2 . In this case, an estimate of P_{DL} is given by $\hat{P}_{\text{DL}} = \Phi((\text{DL} - \hat{\mu})/\hat{\sigma})$, where the MLEs $\hat{\mu}$ and $\hat{\sigma}$ are given in equation (3). Using this estimate of P_{DL} , we can impute the values that are below the DL by

$$x_i^* = \hat{\mu} + z_i \hat{\sigma}, \quad i = 1, \dots, k, \quad (5)$$

where $z_i = \Phi^{-1}(u_i \hat{P}_{\text{DL}})$, $i = 1, \dots, k$ and u_i s are uniform (0, 1) random numbers. By treating the $n - k$ measured values and the k imputed values as a random sample of size n from the $N(\mu, \sigma^2)$ distribution and applying standard procedures with some adjustments for imputation, one can make inferences about the mean, variance or quantiles. Details follow.

Let \bar{x}^* and s^* , respectively, denote the mean and standard deviation of the pooled data containing $n - k$ measurements that are above the DL and the k imputed values in equation (5). Then, a $(1 - \alpha)$ CI for μ is given by

$$\bar{x}^* \pm t_{n-k-1;1-\alpha/2} \frac{s^*}{\sqrt{n}},$$

where $t_{m,p}$ denotes the 100 p th percentile of a t distribution with degrees of freedom (df) m . Note that we have used the df $n - k - 1$ instead of $n - 1$ because we imputed k values. The above imputations should be repeated several times, and the results should be averaged to get a single interval. For example, the average of the end points of the intervals based on several imputations can be used to obtain a single CI for the mean. This is equivalent to averaging the means and standard deviations from several imputations. That is, let

$$\bar{x}^* = \frac{1}{m} \sum_{j=1}^m \bar{x}_j^* \quad \text{and} \quad s^* = \frac{1}{m} \sum_{j=1}^m s_j^*, \quad (6)$$

where \bar{x}_j^* and s_j^* denote, respectively, the mean and standard deviation based on the j th imputation, $j=1, \dots, m$. Then our proposed CI for the mean is

$$\bar{x}^* \pm t_{n-k-1;1-\alpha/2} \frac{s^*}{\sqrt{n}}. \quad (7)$$

If one is interested in testing, for example, $H_0 : \mu \leq \mu_0$ versus $H_a : \mu > \mu_0$, then the p-value based on a set of m imputed samples can be computed as

$$P\left(t_{n-k-1} \geq \frac{\bar{x}^* - \mu_0}{s^*/\sqrt{n}}\right). \quad (8)$$

Remark 1.

Little and Rubin (2002, p. 86) provided a formula for variance estimate which is a combination of the two variance components, namely, within imputation variance and between imputation variance. They also provided an expression involving the two variance components for the df to compute CIs using Student's t -distribution as the reference distribution. Our preliminary simulation studies showed that the CI for a normal mean on the basis of the suggested variance estimates exhibited similar performance as that of the CI in equation (7). Furthermore, no such variance formulas are available for estimating normal quantiles (which is a function of μ and σ) or computing tolerance intervals and so we shall use the variance formula equation (6) in the sequel.

To find a $1 - \alpha$ CI for σ^2 , let \bar{s}^{*2} be the average of the variances based on a set of imputed samples. Then,

$$\left(\frac{(n-k-1)\bar{s}^{*2}}{\chi_{n-k-1;1-\alpha/2}^2}, \frac{(n-k-1)\bar{s}^{*2}}{\chi_{n-k-1;\alpha/2}^2} \right), \quad (9)$$

where $\chi_{m,p}^2$ denotes the p th quantile of a chi-square distribution with df = m , is a $1 - \alpha$ CI for σ^2 .

The following algorithm can be used to construct a CI or to compute the P -value for testing the mean.

Algorithm 1

Based on a given value of the DL and the data above the DL:

Compute the MLEs $\hat{\mu}$ and $\hat{\sigma}^2$ using equations (3) and (4).

Compute $\hat{P}_{DL} = \Phi\left(\frac{DL - \hat{\mu}}{\hat{\sigma}}\right)$

For $j=1, m$

Generate u_1, \dots, u_k from uniform $(0, 1)$

Set $z_i = \Phi^{-1}(u_i \hat{P}_{DL})$ and $x_i^* = \hat{\mu} + z_i \hat{\sigma}$, $i=1, \dots, k$

Compute the mean \bar{x}_j^* and variance s_j^{*2} using x_1^*, \dots, x_k^* and the data above the DL.

(end j loop)

compute $\bar{x}^* = \frac{1}{m} \sum_{j=1}^m \bar{x}_j^*$ and $s^* = \frac{1}{m} \sum_{j=1}^m s_j^*$

Then $\bar{x}^* \pm t_{n-k-1;\alpha/2} \frac{s^*}{\sqrt{n}}$ is a $1 - \alpha$ CI for μ

$P\left(t_{n-k-1} \geq \frac{\bar{x}^* - \mu_0}{s^*/\sqrt{n}}\right)$ is the P -value for testing

$H_0 : \mu \leq \mu_0$ versus $H_a : \mu > \mu_0$.

Tolerance intervals for a normal distribution

A $(p, 1 - \alpha)$ tolerance interval is such that it contains at least a proportion p of the normal population with confidence $1 - \alpha$. Similarly, a $(p, 1 - \alpha)$ upper tolerance limit U is such that at least a proportion p of the normal population is less than U with confidence $1 - \alpha$. A lower tolerance limit is similarly defined. For details, we refer to Guttman (1970) or Krishnamoorthy and Mathew (2009).

If \bar{x} and s denote the mean and standard deviation based on a complete sample of size n from a normal distribution, a $(p, 1 - \alpha)$ upper tolerance limit for the normal distribution is given by (see Guttman, 1970)

$$\bar{x} + K_1 s, \quad \text{with} \quad K_1 = \frac{1}{\sqrt{n}} t_{n-1;1-\alpha}(z_p \sqrt{n}), \quad (10)$$

where z_p is the p th quantile of the standard normal distribution and $t_{m;\alpha}(\delta)$ denotes the α th quantile of a non-central t distribution with df = m and non-centrality parameter δ . The quantity K_1 is referred to as a tolerance factor. A lower tolerance limit can be obtained by replacing the plus sign in equation (10) by the minus sign. Note that a $(p, 1 - \alpha)$ upper tolerance limit also provides a $1 - \alpha$ upper confidence limit for the p th quantile. Similarly, a $(p, 1 - \alpha)$ lower tolerance limit [which can be obtained by replacing the K_1 in equation (10) by $-K_1$] also provides a $1 - \alpha$ lower confidence limit for the $(1 - p)$ th quantile.

If the sample contains k values below the DL, then the upper tolerance limit based on a set of m imputed samples is given by

$$\bar{x}^* + K_{1k}\bar{s}^*, \quad \text{with} \quad K_{1k} = \frac{1}{\sqrt{n}}t_{n-k-1;1-\alpha}(z_p\sqrt{n}), \quad (11)$$

where \bar{x}^* and \bar{s}^* are as defined in equation (6).

An exact two-sided tolerance interval for the normal distribution is given by $\bar{x} \pm Ks$, where K is the tolerance factor. Odeh (1978) computed the exact tolerance factor K for $n = 2(1)98, 100$, $p = 0.75, 0.90, 0.95, 0.975, 0.99$ and 0.995 and $1 - \alpha = 0.5, 0.75, 0.90, 0.95, 0.975, 0.99$ and 0.995 . Eberhardt *et al.* (1989) provided a Fortran program to compute the values of K . The PC calculator that accompanies the book by Krishnamoorthy (2006) computes the one-sided tolerance limits and exact tolerance intervals for a normal distribution. This calculator is free and can be downloaded from <http://www.ucs.louisiana.edu/~kxk4695>.

An accurate approximation for the factor K which is due to Wald and Wolfowitz (1946) is given by

$$K = \left(\frac{(n-1)\chi_{1;p}^2(1/n)}{\chi_{n-1;\alpha}^2} \right)^{1/2}, \quad (12)$$

where $\chi_{1;p}^2(1/n)$ denotes the p th quantile of a non-central chi-square distribution with $df = 1$ and non-centrality parameter $1/n$ and $\chi_{m;\alpha}^2$ denotes the α th quantile of a central chi-square distribution with $df = m$. This approximation is extremely satisfactory even for small sample sizes (as small as 3) if p and $1 - \alpha$ are ≥ 0.9 .

To compute the tolerance factor c for our imputation approach, the df associated with the denominator chi-square random variable in equation (12) should be $n - k - 1$. That is, if $c_{2k} = \left(\frac{(n-1)\chi_{1;p}^2(1/n)}{\chi_{n-k-1;\alpha}^2} \right)^{1/2}$, a $(p, 1 - \alpha)$ tolerance interval based on the imputation approach is given by

$$\bar{x}^* \pm c_{2k}\bar{s}^*, \quad (13)$$

where \bar{x}^* and \bar{s}^* are as defined in equation (6).

Estimation of an exceedance probability

Suppose we want to assess the exceedance probability based on the sample data x_1, \dots, x_n from a normal distribution. Since the exceedance probability at the threshold t is given by $S_t = P(X > t)$, one-sided lower tolerance limits discussed earlier can be used to find a lower confidence limit for S_t . For example, if a $(p, 1 - \alpha)$ lower tolerance limit for a normal distribution is greater than t , then we can conclude that S_t is at least p with confidence $1 - \alpha$. As a consequence, it is not difficult to see that an exact one-sided $1 - \alpha$ lower confidence limit for S_t can be obtained by solving the equation

$$t_{n-1;1-\alpha}(z_p\sqrt{n}) = \frac{\bar{x} - t}{s/\sqrt{n}} \quad (14)$$

for p . Once $n, 1 - \alpha$ and the quantity on the right hand side of equation (14) are given, the above equation can be solved using the PC calculators mentioned in the preceding subsection.

If the sample contains k values below the DL, then a $1 - \alpha$ level lower limit for S_t is the value of p that satisfies

$$t_{n-k-1;1-\alpha}(z_p\sqrt{n}) = \frac{\bar{x}^* - t}{\bar{s}^*/\sqrt{n}}, \quad (15)$$

where \bar{x}^* and \bar{s}^* are as defined in equation (6).

Imputation based on some ad hoc estimators

We already noted that exact MLEs are not simple to calculate as they require table values of λ [see equation (3)] which are not easy to compute. An approximation to the MLEs can be obtained using the approximation for λ in equation (4). An alternative simple approach of imputing is as follows. Let \bar{x}_l and s_l^2 be the mean and variance based on measured values above the DL. We can estimate P_{DL} by

$$\hat{P}_{DL} = \Phi\left(\frac{DL - \bar{x}_l}{s_l}\right). \quad (16)$$

Notice that the location parameter μ is overestimated and as a result \hat{P}_{DL} defined above underestimates P_{DL} . To compensate for this underestimation, we impute the values below the DL by

$$x_i^* = \hat{\mu}_l + z_i s_l, \quad i = 1, \dots, k,$$

where $\hat{\mu}_l = ((n-k)\bar{x}_l + k \times DL)/n$ and $z_i = \Phi^{-1}(u_i \hat{P}_{DL})$, $i = 1, \dots, k$, the u_i s being uniform (0, 1) random numbers. Using the facts that Φ^{-1} is an increasing function and $\bar{x}_l \geq DL$, it can be seen that $x_i^* s_l$ are less than or equal to DL. Notice that $\hat{\mu}_l$ also slightly overestimates the true mean μ because the values below the DL are all replaced by the DL to compute $\hat{\mu}_l$. Nevertheless, as will be seen later in our simulation studies, the imputation based on the above *ad hoc* estimators produces very satisfactory results in many cases.

Accuracy studies of the imputation approach

We shall now study the properties of the proposed approaches in the preceding sections using simulation. Specifically, we like to study the coverage properties of the CIs and tolerance intervals. Algorithm 2 given below describes a Monte Carlo method of evaluating the coverage probabilities of CIs for a normal mean, based on the imputation method that uses the approximate MLE. The coverage probabilities of tolerance intervals can be

Table 2. Coverage probabilities of the 95% CIs for a normal mean based on (a) complete data, (b) imputation using the approximate MLE and (c) imputation using *ad hoc* estimators

$n = 15$												
$P_{DL} = 0.10$				$P_{DL} = 0.20$			$P_{DL} = 0.30$			$P_{DL} = 0.40$		
σ	a	b	c	a	b	c	a	b	c	a	b	c
1.0	0.95 (1.09)	0.95 (1.11)	0.95 (1.11)	0.95 (1.09)	0.95 (1.14)	0.95 (1.12)	0.95 (1.09)	0.95 (1.16)	0.95 (1.12)	0.95 (1.09)	0.96 (1.21)	0.96 (1.11)
3.0	0.95 (3.27)	0.95 (3.36)	0.95 (3.34)	0.95 (3.27)	0.95 (3.42)	0.95 (3.37)	0.95 (3.26)	0.95 (3.48)	0.95 (3.34)	0.95 (3.27)	0.95 (3.63)	0.95 (3.32)
$n = 20$												
σ	a	b	c	a	b	c	a	b	c	a	b	c
1.0	0.95 (0.93)	0.95 (0.94)	0.95 (0.94)	0.95 (0.92)	0.95 (0.95)	0.95 (0.94)	0.95 (0.92)	0.94 (0.96)	0.94 (0.92)	0.95 (0.92)	0.94 (0.97)	0.94 (0.90)
3.0	0.95 (2.77)	0.95 (2.82)	0.95 (2.82)	0.95 (2.78)	0.95 (2.86)	0.95 (2.82)	0.95 (2.77)	0.95 (2.89)	0.95 (2.79)	0.95 (2.77)	0.95 (2.95)	0.95 (2.73)
$n = 30$												
$P_{DL} = 0.20$			$P_{DL} = 0.30$			$P_{DL} = 0.40$			$P_{DL} = 0.50$			
σ	a	b	c	a	b	c	a	b	c	a	b	c
1.0	0.95 (0.74)	0.95 (0.75)	0.95 (0.74)	0.95 (0.74)	0.95 (0.76)	0.95 (0.73)	0.95 (0.74)	0.93 (0.76)	0.94 (0.71)	0.95 (0.74)	0.95 (0.78)	0.95 (0.68)
3.0	0.95 (2.22)	0.95 (2.26)	0.95 (2.23)	0.95 (2.22)	0.95 (2.26)	0.95 (2.19)	0.95 (2.21)	0.95 (2.29)	0.95 (2.13)	0.95 (2.22)	0.95 (2.32)	0.95 (2.03)
$n = 50$												
σ	a	b	c	a	b	c	a	b	c	a	b	c
1.0	0.95 (0.57)	0.95 (0.57)	0.95 (0.57)	0.95 (0.57)	0.94 (0.57)	0.94 (0.55)	0.95 (0.57)	0.90 (0.57)	0.91 (0.53)	0.95 (0.57)	0.90 (0.57)	0.91 (0.50)
3.0	0.95 (1.70)	0.95 (1.71)	0.95 (1.70)	0.95 (1.70)	0.94 (1.71)	0.94 (1.66)	0.94 (1.70)	0.95 (1.72)	0.90 (1.60)	0.95 (1.70)	0.90 (1.72)	0.90 (1.51)

Expected lengths are given in parentheses.

Table 4. Coverage probabilities of the one-sided tolerance limits based on (a) complete data, (b) imputation using approximate MLE and (c) imputation using the *ad hoc* estimators

Upper limits												
$n = 15$												
$P_{DL} = 0.10$			$P_{DL} = 0.20$			$P_{DL} = 0.30$			$P_{DL} = 0.40$			
σ	a	b	c	a	b	c	a	b	c	a	b	c
1.0	0.95 (2.51)	0.95 (2.62)	0.95 (2.61)	0.95 (2.53)	0.96 (2.71)	0.96 (2.67)	0.95 (2.51)	0.96 (2.82)	0.95 (2.71)	0.95 (2.51)	0.96 (3.02)	0.95 (2.78)
3.0	0.95 (7.56)	0.95 (7.83)	0.96 (7.81)	0.95 (7.56)	0.96 (8.12)	0.96 (8.01)	0.95 (7.55)	0.96 (8.46)	0.95 (8.14)	0.95 (7.55)	0.97 (9.55)	0.95 (8.34)
$n = 20$												
σ	a	b	c	a	b	c	a	b	c	a	b	c
1.0	0.95 (2.36)	0.95 (2.43)	0.95 (2.42)	0.95 (2.37)	0.96 (2.49)	0.95 (2.46)	0.95 (2.36)	0.96 (2.56)	0.95 (2.48)	0.95 (2.37)	0.97 (2.68)	0.94 (2.50)
3.0	0.95 (7.09)	0.96 (7.29)	0.96 (7.28)	0.95 (7.11)	0.96 (7.47)	0.96 (7.38)	0.95 (7.10)	0.96 (7.70)	0.95 (7.44)	0.95 (7.09)	0.97 (7.99)	0.95 (7.46)
$n = 30$												
$P_{DL} = 0.20$			$P_{DL} = 0.30$			$P_{DL} = 0.40$			$P_{DL} = 0.50$			
σ	a	b	c	a	b	c	a	b	c	a	b	c
1.0	0.95 (2.20)	0.95 (2.27)	0.95 (2.25)	0.95 (2.21)	0.96 (2.33)	0.94 (2.26)	0.95 (2.20)	0.97 (2.38)	0.94 (2.24)	0.95 (2.20)	0.97 (2.47)	0.91 (2.22)
3.0	0.95 (6.61)	0.96 (6.83)	0.95 (6.76)	0.95 (6.60)	0.96 (6.96)	0.95 (6.74)	0.95 (6.61)	0.96 (7.14)	0.93 (6.70)	0.95 (6.61)	0.97 (7.43)	0.92 (6.66)
$n = 50$												
σ	a	b	c	a	b	c	a	b	c	a	b	c
1.0	0.95 (2.05)	0.95 (2.10)	0.95 (2.08)	0.95 (2.05)	0.95 (2.12)	0.93 (2.06)	0.95 (2.06)	0.96 (2.13)	0.94 (2.06)	0.95 (2.05)	0.96 (2.15)	0.91 (2.03)
3.0	0.95 (6.16)	0.95 (6.29)	0.94 (6.23)	0.95 (6.17)	0.96 (6.37)	0.94 (6.19)	0.95 (6.17)	0.96 (6.37)	0.94 (6.19)	0.95 (6.16)	0.96 (6.47)	0.92 (6.10)
Lower limits												
$n = 15$												
$P_{DL} = 0.10$			$P_{DL} = 0.20$			$P_{DL} = 0.30$			$P_{DL} = 0.40$			
σ	a	b	c	a	b	c	a	b	c	a	b	c
1.0	0.95 (-2.53)	0.95 (-2.64)	0.95 (-2.64)	0.95 (-2.51)	0.95 (-2.72)	0.95 (-2.68)	0.95 (-2.52)	0.94 (-2.85)	0.94 (-2.73)	0.95 (-2.52)	0.93 (-3.08)	0.92 (-2.77)
3.0	0.95 (-7.56)	0.95 (-7.90)	0.95 (-7.88)	0.95 (-7.56)	0.94 (-8.21)	0.94 (-8.09)	0.95 (-7.55)	0.93 (-8.55)	0.93 (-8.71)	0.95 (-7.56)	0.92 (-9.20)	0.91 (-8.28)
$n = 20$												
σ	a	b	c	a	b	c	a	b	c	a	b	c
1.0	0.95 (-2.36)	0.95 (-2.44)	0.95 (-2.44)	0.95 (-2.37)	0.94 (-2.51)	0.94 (-2.48)	0.95 (-2.37)	0.93 (-2.59)	0.93 (-2.49)	0.95 (-2.35)	0.92 (-2.69)	0.91 (-2.45)
3.0	0.95 (-7.10)	0.95 (-7.33)	0.95 (-7.33)	0.95 (-7.10)	0.95 (-7.53)	0.95 (-7.44)	0.95 (-7.09)	0.93 (-7.73)	0.93 (-7.43)	0.95 (-7.10)	0.92 (-8.10)	0.91 (-7.39)
$n = 30$												
$P_{DL} = 0.20$			$P_{DL} = 0.30$			$P_{DL} = 0.40$			$P_{DL} = 0.50$			
σ	a	b	c	a	b	c	a	b	c	a	b	c
1.0	0.95 (-2.20)	0.94 (-2.28)	0.94 (-2.26)	0.95 (-2.20)	0.93 (-2.33)	0.93 (-2.25)	0.95 (-2.20)	0.91 (-2.39)	0.89 (-2.20)	0.95 (-2.20)	0.90 (-2.48)	0.82 (-2.11)
3.0	0.95 (-6.60)	0.94 (-6.86)	0.94 (-6.79)	0.95 (-6.58)	0.93 (-6.95)	0.93 (-6.72)	0.95 (-6.61)	0.92 (-7.17)	0.90 (-6.60)	0.95 (-6.63)	0.90 (-7.43)	0.81 (-6.32)
$n = 50$												
σ	a	b	c	a	b	c	a	b	c	a	b	c
1.0	0.95 (-2.06)	0.94 (-2.11)	0.94 (-2.09)	0.95 (-2.06)	0.93 (-2.13)	0.92 (-2.06)	0.95 (-2.06)	0.90 (-2.15)	0.86 (-1.99)	0.95 (-2.05)	0.88 (-2.18)	0.73 (-1.87)
3.0	0.95 (-6.17)	0.94 (-6.31)	0.94 (-6.26)	0.95 (-6.17)	0.93 (-6.37)	0.92 (-6.18)	0.95 (-6.17)	0.90 (-6.46)	0.86 (-5.98)	0.95 (-6.16)	0.88 (-6.58)	0.74 (-5.63)

Expected value of the limits are given in parentheses.

Table 5. Calibrated confidence levels for the MLE-based (0.95, 0.95) one-sided tolerance limits

P_{DL}	Lower limits					Upper limits				
	n					n				
	30	40	50	100	200	30	40	50	100	200
0.3	0.960	0.965	0.970	0.970	0.980	0.940	0.940	0.940	0.940	0.940
0.4	0.970	0.975	0.975	0.980	0.985	0.940	0.940	0.940	0.940	0.940
0.5	0.980	0.980	0.985	0.985	0.990	0.940	0.930	0.930	0.930	0.930

An example for normal distribution

Example 1. To illustrate the methods of finding CIs for a normal mean and variance, let us consider Example 4.7 in Martz and Waller (1982). The data for this example represent failure time in years for 22 bushing failures of 115 kV power generator. The data are given in Table 8 below. As shown in Martz and Waller (1982), the normal distribution provides a good fit for the data. We shall first give CIs for the mean and variance based on the complete data. The summary statistics are $n = 22$, $\bar{x} = 14.01$, $s = 3.362$ and $t_{21; .975} = 2.0796$. Thus, the 95% CI based on the complete data is (12.52, 15.50).

To illustrate the methods of the preceding sections, let us choose a right-censoring time of 15.5 years so that $k = 7$ failure times were not observed. In this case, the MLEs based on the approximate method is $\hat{\mu} = 13.781$ and $\hat{\sigma} = 3.145$ and the ones based on the exact method are $\hat{\mu} = 13.856$ and $\hat{\sigma} = 3.137$. Using the exact MLEs and 10 imputations, we found $\bar{x}^* = 13.810$ and $\bar{s}^* = 3.172$. Noting that $t_{13; .975} = 2.1604$, we get the 95% CI for the mean as $\bar{x}^* \pm t_{n-k-1; 1-\alpha/2} \frac{\bar{s}^*}{\sqrt{n}} = (12.35, 15.27)$. Using the *ad hoc* procedure with 10 imputations, we found $\bar{x}^* = 13.820$ and $\bar{s}^* = 3.069$ and the 95% CI is given by (12.41, 15.23). Notice that both CIs are in agreement with (12.52, 15.50) based on the entire data.

The usual $1 - \alpha$ CI for the variance is given by $(\frac{(n-1)s^2}{\chi_{n-1; 1-\alpha/2}^2}, \frac{(n-1)s^2}{\chi_{n-1; \alpha/2}^2})$. This yields a 95% CI for the variance based on the complete data as (6.69, 23.09). The MLE-based imputation yielded $\bar{s}^{*2} = 10.06$, which is the average of the variances of 10 imputed samples. Using $\bar{s}^{*2} = 10.06$, $\chi_{n-k-1; \alpha/2}^2 = \chi_{13; .025}^2 = 5.009$ and $\chi_{13; .975}^2 = 24.736$ in equation (9), we get the 95% CI as (5.29, 26.11). The value of \bar{s}^{*2} based on the *ad hoc* method with 10 imputations is 9.42 and the 95% CI is (4.95, 24.45). Here, the *ad hoc* method produced the CI that is close to the one based on the complete data.

LOG-NORMAL DISTRIBUTION

The procedures for computing tolerance limits or bounds for exceedance probabilities can be extended to any other distribution that has one–one relation with the normal distribution. For instance, if y_1, \dots, y_n is a sample from a log-normal distribution with parameters μ and σ^2 , then $x_1 = \ln(y_1), \dots, x_n = \ln(y_n)$ can be

regarded as a sample from a $N(\mu, \sigma^2)$ distribution. Therefore, normal-based methods can be applied to log-transformed data to compute tolerance limits and then by taking antilog we get tolerance limits for the log-normal distribution.

We shall use an example to illustrate the results of the preceding sections for constructing one-sided upper tolerance limits, an upper limit for an exceedance probability and a lower prediction limit for a log-normal distribution.

Example 2. The uncensored sample for this example is taken from Wild *et al.* (1996), and the data represent oil mist measurements obtained from a machining workshop in France. The 14 uncensored measurements are 1.7, 1.8, 2.1, 2.3, 2.3, 2.5, 2.8, 2.9, 2.9, 3.0, 3.0, 3.8, 3.8 and 5.3.

Normal probability plot fits the log-transformed data very well and so we can assume that the sample is from a log-normal distribution. In the following, we shall illustrate the methods described in the preceding sections.

We first note that $n = 14$ and the mean and standard deviation of the log-transformed data are $\bar{x} = 1.0097$ and $s = 0.3060$.

Upper tolerance limit: The tolerance factor for constructing the normal-based (0.90, 0.95) one-sided tolerance limits is $\frac{1}{\sqrt{14}} t_{13; .95} (1.2816\sqrt{14}) = 2.1088$. Thus, based on the complete log-transformed data, we get

$$\begin{aligned} &\bar{x} + 2.1088 \times s \\ &= 1.0097 + 2.1088 \times 0.3060 \\ &= 1.655, \end{aligned}$$

and the required upper tolerance limit is $\exp(1.665) = 5.233$.

To illustrate the procedures for a left-censored sample, let us assume that the DL is 2.4. In this case, the number of values below the DL is $k = 5$. The exact likelihood method yielded $\hat{\mu} = 1.0024$ and $\hat{\sigma} = 0.3253$. Using these exact MLEs and 10 imputations, we computed $\bar{x}^* = 0.9988$ and $\bar{s}^* = 0.3204$ for the log-transformed data. The required tolerance factor for applying multiple imputation is $\frac{1}{\sqrt{14}} t_{8; 0.95} (z_{0.90}\sqrt{14}) = 2.3706$. Finally, we computed the (0.90, 0.95) upper tolerance limit for the log-transformed data as 1.758 and so the desired upper tolerance limit for original measurements is $\exp(1.758) = 5.801$. Using the *ad hoc* method with 10 imputations, we computed $\bar{x}^* = 0.9975$ and

Table 9. Coverage probabilities of the (0.95, 0.95) upper tolerance limits for a gamma (a,1) distribution based on a (a) complete data, (b) imputation using approximate MLE and (c) imputation using the *ad hoc* estimators

<i>n</i> = 15												
$P_{DL} = 0.10$			$P_{DL} = 0.20$			$P_{DL} = 0.30$			$P_{DL} = 0.40$			
<i>a</i>	a	b	c	a	b	c	a	b	c	a	b	c
1.0	0.94 (5.18)	0.95 (5.65)	0.95 (5.62)	0.94 (5.23)	0.96 (6.12)	0.95 (5.97)	0.94 (5.19)	0.96 (6.60)	0.96 (6.15)	0.94 (5.20)	0.97 (8.12)	0.95 (6.15)
3.0	0.95 (9.22)	0.96 (9.64)	0.96 (9.61)	0.95 (9.24)	0.96 (10.08)	0.96 (9.90)	0.95 (9.29)	0.97 (10.66)	0.96 (10.16)	0.95 (9.22)	0.97 (12.19)	0.95 (10.47)
7.0	0.95 (15.55)	0.95 (16.01)	0.95 (15.98)	0.95 (15.62)	0.96 (16.54)	0.96 (16.34)	0.95 (15.53)	0.96 (17.06)	0.95 (16.51)	0.95 (15.59)	0.97 (18.19)	0.95 (16.92)
<i>n</i> = 20												
<i>a</i>	a	b	c	a	b	c	a	b	c	a	b	c
1.0	0.94 (4.70)	0.95 (5.03)	0.95 (5.01)	0.93 (4.70)	0.95 (5.29)	0.95 (5.18)	0.93 (4.69)	0.96 (5.54)	0.95 (5.26)	0.94 (4.70)	0.97 (5.94)	0.95 (5.32)
3.0	0.95 (8.63)	0.96 (8.91)	0.96 (8.89)	0.95 (8.63)	0.96 (9.17)	0.96 (9.04)	0.95 (8.65)	0.96 (9.49)	0.95 (9.14)	0.95 (8.63)	0.96 (9.89)	0.94 (9.17)
7.0	0.95 (14.84)	0.96 (15.16)	0.96 (15.14)	0.95 (14.79)	0.96 (15.39)	0.95 (15.23)	0.95 (14.81)	0.96 (15.76)	0.95 (15.35)	0.95 (14.89)	0.97 (16.36)	0.95 (15.51)
<i>n</i> = 30												
$P_{DL} = 0.20$			$P_{DL} = 0.30$			$P_{DL} = 0.40$			$P_{DL} = 0.50$			
<i>a</i>	a	b	c	a	b	c	a	b	c	a	b	c
1.0	0.93 (4.21)	0.95 (4.58)	0.95 (4.51)	0.93 (4.21)	0.96 (4.72)	0.95 (4.52)	0.93 (4.22)	0.96 (4.92)	0.94 (4.49)	0.94 (4.23)	0.97 (5.27)	0.93 (4.48)
3.0	0.95 (8.04)	0.96 (8.35)	0.95 (8.26)	0.95 (8.03)	0.96 (8.50)	0.94 (8.25)	0.95 (8.02)	0.96 (8.72)	0.93 (8.20)	0.95 (8.02)	0.97 (9.05)	0.92 (8.12)
7.0	0.95 (14.12)	0.96 (14.46)	0.95 (14.35)	0.95 (14.08)	0.96 (14.62)	0.94 (14.31)	0.95 (14.10)	0.96 (14.93)	0.93 (14.28)	0.95 (14.09)	0.97 (15.33)	0.91 (14.19)
<i>n</i> = 50												
<i>a</i>	a	b	c	a	b	c	a	b	c	a	b	c
1.0	0.93 (3.83)	0.95 (4.08)	0.95 (4.02)	0.92 (3.83)	0.95 (4.15)	0.94 (4.00)	0.92 (3.82)	0.96 (4.24)	0.91 (3.92)	0.92 (3.82)	0.97 (4.39)	0.88 (3.83)
3.0	0.95 (7.53)	0.95 (7.71)	0.95 (7.64)	0.95 (7.54)	0.96 (7.81)	0.94 (7.61)	0.95 (7.54)	0.97 (7.91)	0.92 (7.51)	0.95 (7.53)	0.97 (8.09)	0.89 (7.38)
7.0	0.95 (13.46)	0.96 (13.67)	0.95 (13.58)	0.95 (13.46)	0.96 (13.75)	0.94 (13.50)	0.95 (13.46)	0.96 (13.90)	0.91 (13.39)	0.95 (13.84)	0.97 (14.14)	0.89 (13.25)

Expected value of the limits are given in parentheses.

exponentiation, we get the prediction limit 1.566. The prediction limit based on the imputations using the MLEs is given by

$$\begin{aligned} \bar{x}^* - t_{n-k-1;1-\alpha} s^* \sqrt{1 + \frac{1}{n}} \\ = 0.9988 - 1.8596 \times 0.3204 \sqrt{1 + \frac{1}{14}} \\ = 0.3821 \end{aligned}$$

Thus, the 95% lower prediction limit is $\exp(0.3821) = 1.465$. If we use the *ad hoc* method, then we have

$$\begin{aligned} \bar{x}^* - t_{n-k-1;1-\alpha} s^* \sqrt{1 + \frac{1}{n}} \\ = 0.9975 - 1.85955 \times 0.3135 \sqrt{1 + \frac{1}{14}} \\ = 0.3941, \end{aligned}$$

and $\exp(0.3941) = 1.4830$.

GAMMA DISTRIBUTION

Recently, Krishnamoorthy *et al.* (2008) constructed tolerance limits and prediction limits for a gamma distribution using the Wilson and Hilferty (1931) normal approximation. The approximation states that, if Y follows a gamma distribution, the distribution of $Y^{1/3}$ is well approximated by a normal distribution. Krishnamoorthy *et al.* (2008) used the approximation in the following way: if y_1, \dots, y_n is a sample from a gamma distribution, then the transformed sample $x_1 = y_1, \dots, x_n = y_n$ can be regarded as a sample from

a normal distribution with an arbitrary mean μ and arbitrary variance σ^2 . Thus normal-based procedures can be derived based on x_1, \dots, x_n , and the results can then be transformed for the gamma distribution. The resulting procedures are not only simple but also the simulation studies by Krishnamoorthy *et al.* (2008) showed that the results are satisfactory and are comparable to those based on complex methods.

If a sample from a gamma distribution contains some non-detects, then after taking cube root of the measurements above the DL, we simply apply normal-based imputation methods to make inference. As this approach involves an approximation and imputations, we studied its properties in the context of constructing upper tolerance limits. Specifically, we computed the coverage probabilities of upper tolerance limits for a gamma distribution for a few values of the shape parameter a and the results are given in Table 9. The estimated coverage probabilities show similar patterns as those for the complete sample case reported in Krishnamoorthy *et al.* (2008). In particular, the procedure is liberal when a is small; otherwise, it performs satisfactorily for moderate sample sizes and $P_{DL} \leq 0.50$. For large samples and $P_{DL} \geq 0.40$, the *ad hoc* method is liberal while the MLE-based imputation seems to be conservative. The calibrated confidence levels obtained for the normal distribution can be applied to get more accurate results for the gamma distribution when the sample size is large.

We also studied the properties of lower tolerance limits and tolerance intervals for a gamma distribution based on the imputation procedures proposed in the preceding sections (not reported here). In general, we found that the imputation procedures are very satisfactory as long as $P_{DL} \leq 0.40$.

Example 3. In this example, we shall use the data reported in Gibbons (1994, p. 261), which were also used for illustrative purpose by Aryal *et al.* (2007) and Krishnamoorthy *et al.* (2008). The measurements represent alkalinity concentrations in ground water obtained from a ‘greenfield’ site (the site of a waste disposal landfill prior to disposal of waste) and are reproduced here in Table 10.

Let $x_i = y_i^{1/3}$, $i = 1, \dots, 27$. The mean $\bar{x} = 3.8274$ and the standard deviation $s_x = 0.4298$. Krishnamoorthy *et al.* (2008) constructed $(p, 0.95)$ one-sided upper tolerance limits and two-sided tolerance intervals for $p = 0.90, 0.95, 0.99$. The tolerance limits

Table 10. Alkalinity concentrations in ground water

y_i :	28	32	39	40	40	42	42	42	49	51
	51	52	54	54	55	58	59	59	60	63
	66	70	79	82	89	96	118			

Table 11. Tolerance limits based on complete data; $n = 27$, $DL = 50$ and $k = 9$

$(p, 1-\alpha)$	Factor for one-sided	Upper limit	Factor for tolerance interval	Tolerance interval
(0.9, 0.95)	1.8114	97.71	2.1814	(24.1,108.3)
(0.95, 0.95)	2.2601	110.51	2.6011	(19.9,121.0)
(0.95, 0.95)	3.1165	137.94	3.4146	(13.1,148.5)

Table 12. Tolerance limits based on 10 imputations; $DL = 50$, $n = 27$ and $k = 9$

$(p, 1-\alpha)$	Factor for one sided	Upper tolerance limit		Factor for tolerance interval	Tolerance interval	
		MLE	<i>ad hoc</i>		MLE	<i>ad hoc</i>
(0.9, 0.95)	1.9212	103.1	98.8	2.3501	(21.8,116.2)	(22.3,110.9)
(0.95, 0.95)	2.4073	118.0	112.6	2.7989	(17.3,131.1)	(18.1,124.6)
(0.95, 0.95)	3.3345	150.4	142.3	3.6747	(10.6,163.6)	(11.5,154.4)

along with corresponding tolerance factors are reproduced here in Table 11.

We shall now assume that the DL is 50 and use the imputation methods to construct the various limits. The MLEs based on the approximate method are $\hat{\mu} = 3.842$ and $\hat{\sigma} = 0.4355$, and those based on the exact method are $\hat{\mu} = 3.824$ and $\hat{\sigma} = 0.4353$. Using these exact MLEs and 10 imputations, we computed $\bar{x}^* = 3.834$ and $\bar{s}^* = 0.4430$, and using the *ad hoc* method with 10 imputations, we obtained $\bar{x}^* = 3.810$ and $\bar{s}^* = 0.4231$. The results based on the approximate MLEs and the *ad hoc* procedures are given in Table 12. We observe from Table 12 that the tolerance intervals obtained using the *ad hoc* method are shorter than those based on the MLEs. Furthermore, the upper tolerance limits based on the *ad hoc* method are closer to the corresponding ones based on the complete sample that are reported in Table 11.

Prediction limits: Recall that a $1 - \alpha$ upper prediction limit for a future observation from a normal distribution is given by $\bar{x} + t_{n-1;1-\alpha} s \sqrt{1 + 1/n}$. Based on complete data, we computed the 95% upper prediction limit as $3.8274 + 1.7056 \times 0.4298 \sqrt{1 + 1/23} = 4.576$. Thus, a 95% upper prediction limit for the gamma distribution is $(4.576)^3 = 95.64$. Using the MLE-based imputation, we get $\bar{x}^* + t_{n-k-1;1-\alpha} \bar{s}^* \sqrt{1 + 1/n} = 3.834 + 1.740 \times 0.4450 \sqrt{1 + 1/27} = 4.623$. Taking third power, we get 98.8. Using *ad hoc* procedure, we have $3.816 + 1.740 \times 0.4231 \sqrt{1 + 1/27} = 4.566$ and $(4.566)^3 = 95.2$. We again see that the *ad hoc* method produced a prediction limit that is closer to the one based on the complete data than the prediction limit that uses the MLE-based imputation.

CONCLUDING REMARKS

Samples that include measurements below the DL are very common in many areas of application, most notably in environmental sampling and industrial hygiene. A unified practical approach that also performs accurately has been lacking for performing data analysis under this scenario. This article is an attempt to fill this void by advocating an imputation approach that replaces the values below the DL with an imputed value. For the normal and related distributions, CIs, tolerance limits, prediction limits etc. are then developed using the available standard procedures, and their performance have been numerically investigated. The confidence levels have been calibrated to improve the accuracy, and recommendations are made regarding the methodology to be adopted in practice. Our overall conclusion is that, the imputation approach is quite satisfactory when samples contain values below the DL. Furthermore, imputation is now a very standard procedure that is widely used (see Rubin, 1987 and Little

and Rubin, 2002), and our proposed methodology is applicable to small samples. It should be noted that the Bayesian methodology is yet another approach that can be used to analyze samples that include non-detects. Programs such as BUGS have been developed that allow the easy implementation of the Bayesian approach. However, this is not investigated in this article since our motivation has been to come up with procedures that perform well in the frequentist sense.

The framework developed in this article should be applicable to other setups and problems where below DL values are encountered, for example, in the context of other distributions, in the context of regression models, etc. Finally, we note that data that exhibit between and within worker variations are of considerable interest, but methodology that can handle such a situation is lacking when the sample includes non-detects. We are currently investigating problems in this direction.

FUNDING

National Institute of Occupational Safety and Health (R01-OH03628-01A1).

Acknowledgements—The authors are grateful to the editor David Bartley and two reviewers for their valuable comments and suggestions.

REFERENCES

- Amemiya T. (1984) Tobit models: A survey. *J Econom*; 24: 3–61.
- Aryal S, Bhaumik DK, Mathew T *et al.* (2007) Approximate tolerance limits and prediction limits for the gamma distribution. *J Appl Stat Sci*; 16: 103–111.
- Baccarelli A, Pfeiffer R, Consonni D *et al.* (2005) Handling of dioxin measurement data in the presence of non-detectable values: Overview of available methods and their application in the Seveso Chloracne study. *Chemosphere*; 60: 898–906.
- Cohen AC. (1959) Simplified estimators for the normal distribution when samples are singly censored or truncated. *Technometrics*; 1: 217–237.
- Cohen AC. (1961) Tables for maximum likelihood estimates: singly truncated and singly censored samples. *Technometrics*; 3: 535–541.
- Eberhardt KR, Mee RW, Reeve CP. (1989) Computing factors for exact two-sided tolerance limits for a normal distribution. *Communications in Statistics—Simulation and Computation*; 18: 397–413.
- Gibbons RD. (1994) *Statistical methods for groundwater monitoring*. New York, NY: Wiley.
- Gibbons RD, Coleman DD. (2001) *Statistical methods for detection and quantification of environmental contamination*. New York, NY: Wiley.
- Guttman I. (1970) *Statistical tolerance regions: classical and Bayesian*. London: Griffin.
- Hass CN, Scheff PA. (1990) Estimation of averages in truncated samples. *Environ Sci Technol*; 24: 912–9.
- Helsel DR. (2005) *Nondetects and data analysis*. New York, NY: Wiley.
- Hewett P, Ganser GH. (2007) A comparison of several methods for analyzing censored data. *Ann Occup Hyg*; 51: 611–32.
- Krishnamoorthy K. (2006) *Handbook of statistical distributions with applications*. New York, NY: Chapman Hall/CRC.
- Krishnamoorthy K, Mathew T. (2009) *Statistical tolerance regions: theory, applications and computation*. New York, NY: Wiley.

- Krishnamoorthy K, Mathew T, Mukherjee S. (2008) Normal based methods for a gamma distribution: prediction and tolerance intervals and stress-strength reliability. *Technometrics*; 50: 69–78.
- Little RJA, Rubin DB. (2002) *Statistical analysis with missing data*. 2nd edn edn. New York, NY: Wiley.
- Lubin JH, Colt JS, Camann D *et al.* (2004) Epidemiologic evaluation of measurement data in the presence of detection limits. *Environ Health Perspect*; 112: 1691–6.
- Martz HF, Waller RA. (1982) *Bayesian reliability analysis*. New York, NY: John Wiley.
- Odeh RE. (1978) Tables of two-sided tolerance factors for a normal distribution. *Communications in Statistics–Simulation and Computation*; 7: 183–201.
- Rubin DB. (1987) *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.
- Schmee J, Gladstein D, Nelson W. (1985) Confidence limits of a normal distribution from singly censored samples using maximum likelihood. *Technometrics*; 27: 119–28.
- US-EPA. (1984) Definition and procedure for determination of the method detection limit: revision 1.11. *Fed Reg*; 49: 43430–1.
- Wald A, Wolfowitz J. (1946) Tolerance limits for a normal distribution. *Ann Math Stat*; 17: 208–15.
- Wild P, Hordan R, Leplay A. (1996) Confidence intervals for probabilities of exceeding threshold limits with censored log-normal data. *Environmetrics*; 17: 247–59.
- Wilson EB, Hilferty MM. (1931) The distribution of chi-squares. *Proc Natl Acad Sci USA*; 17: 684–8.