# One-Sided Approximate Prediction Intervals for at Least $p$ of $m$ Observations From a Gamma Population at Each of $r$ Locations

**Dulal Kumar BHAUMIK and Robert David GIBBONS**

Center for Health Statistics
Departments of Biostatistics and Psychiatry
University of Illinois, Chicago, IL 60612

We develop simultaneous approximate statistical prediction limits for a gamma-distributed random variable. Specifically, we develop an upper prediction limit (UPL) for $p$ of $m$ future samples at each of $r$ locations, based on a previous sample of $n$ measurements. A typical example is the environmental monitoring problem in which the distribution of an analyte of concern is typically non-Gaussian, simultaneous determinations are required for multiple locations (e.g., ground-water monitoring wells), and, in the event of an initial exceedance of the prediction limit, one or more verification samples are obtained to confirm evidence of an impact on the environment. For example, consider a ground-water monitoring program with $r$ wells and the requirement that at least $p = 1$ of the $m = 2$ samples in each of the $r$ wells be below the UPL. We provide derivation of simultaneous approximate gamma UPLs, illustration of the relevance of the gamma distribution to environmental data, a limited simulation study of type I and II error rates achieved using the method and comparison with normal and nonparametric alternatives, tables that aid computation, and an example using ground-water monitoring data.

KEY WORDS: Censored data; Environmental monitoring; Gamma distribution; Prediction limits.

## 1. INTRODUCTION

A common problem in environmental monitoring is the detection of a possible environmental impact on the basis of a small number of samples obtained from a potential area of concern. As an example, in ground-water detection monitoring programs at waste disposal facilities (i.e., landfills), a series of $m$ samples from each of $r$ monitoring wells located hydraulically downgradient of the facility are often compared with statistical prediction limits computed from $n$ measurements obtained from one or more upgradient sampling locations (see Gibbons 1994; Gibbons and Coleman 2001). Because groundwater moves slowly and the cost of the analytical work is high, $m$ is typically quite small (e.g., $m \leq 3$). Furthermore, given the lack of statistical independence of samples obtained either at the same point in time or at points close in time, sequential sampling (at intervals sufficiently spaced to guarantee independence) becomes the rule. For example, Davis and McNichols (1987) developed sequential normal prediction limits that include $p$ of $m$ samples at each of $r$ downgradient monitoring wells. For the common case of $m = 2$ and $p = 1$, in the event of an initial exceedance of the upper prediction limit (UPL), the owner/operator of the facility can obtain a single verification resample from that well for that constituent. Thus statistically significant exceedance is recorded only if both the initial sample and the verification resample both exceed the UPL for that constituent. Alternative verification resampling plans in practical use include pass one of two resamples (i.e., $m = 3$ and $p = 1$) and pass two of two resamples (i.e., $m = 3$ and $p = 2$) in the event of an initial exceedance (see Gibbons 1994).

Unfortunately, many constituents commonly monitored as part of environmental investigations are not normally distributed and may not be amenable to commonly used methods of transformation that help bring about the assumed normality of the statistical methodology. Toward this end, Gibbons (1990, 1991, 1994) and Davis and McNichols (1999) generalized the result of Davis and McNichols (1987) to the nonparametric case. Although these nonparametric prediction limits solve the problem of distributional misspecification, the downside of the nonparametric approach is that statistical power can be quite low for the small background sample sizes (i.e., $n$) that are typical of many environmental monitoring programs (Gibbons 1994).

A special class of constituents that are rarely, if ever, normally distributed is volatile organic compounds (e.g., benzene, toluene, vinyl chloride). These constituents are of particular interest in environmental investigations because they are anthropogenic. Despite this, low levels of these constituents are often found in background monitoring wells and even quality control samples, such as field blanks and trip blanks, which contain only distilled water. The low-level detection of such substances in upgradient background monitoring wells is often due to cross-contamination from air or gas or the analytical process itself (see Gibbons and Coleman 2001 for a review). The empirical distribution of these constituents is far from Gaussian, with the bulk of the distribution at or below the limit of detection (Currie 1968) and an occasional more extreme detected concentration. Beyond the use of nonparametric prediction limits, little other statistical work has been done in this area. (See Gibbons 1987 for an approximate parametric approach based on a Poisson process.) Because many of the volatile organic compounds pose serious health risks, it is of critical importance to human health and the environment that the most statistically powerful approaches be used to determine evidence of their release. To


TECHNOMETRICS, FEBRUARY 2006, VOL. 48, NO. 1
DOI 10.1198/004017005000000355

*Table 1. Vinyl Chloride Data From Clean Upgradient Ground-Water Monitoring Wells in ($\mu g/L$)*

| | | | | |
|---|---|---|---|---|
| 5.1 | 1.2 | 1.3 | .6 | .5 |
| 2.4 | .5 | 1.1 | 8.0 | .8 |
| .4 | .6 | .9 | .4 | 2.0 |
| .5 | 5.3 | 3.2 | 2.7 | 2.9 |
| 2.5 | 2.3 | 1.0 | .2 | .1 |
| .1 | 1.8 | .9 | 2.0 | 4.0 |
| 6.8 | 1.2 | .4 | .2 | |

illustrate the problem, consider the vinyl chloride concentrations obtained from clean upgradient monitoring wells given in Table 1.

Figure 1 presents a gamma probability plot corresponding to the data given in Table 1. The figure reveals an excellent fit of these data to the gamma distribution. Note that because vinyl chloride is not naturally occurring, it should never be detected in clean upgradient monitoring wells, but this clearly is not true in practice.

These findings suggest that the gamma distribution generally characterizes these environmental data well and should provide a useful alternative to the normal, lognormal, and nonparametric approaches that dominate routine practice in this area.

The gamma distribution has been used extensively in the physical sciences. For example, in industrial engineering, Davis (1952) and Barlow and Proschan (1965) pointed out the importance of a gamma distribution for the failure times of complex systems under continuous repair and maintenance. Using a gamma distribution is more appropriate than using a normal distribution when a tail property is revealed in the data and when variability and concentration are related, as they are in the case of many environmental constituents (see Gibbons and Coleman 2001). Moreover, a gamma distribution approaches a normal distribution when its shape parameter becomes large. When the underlying distribution is gamma, using Student's $t$-statistic is less justified, because it is not based on the correct sufficient statistics. For small sample sizes when the distributional properties cannot be easily verified, routine use of the normal distribution is often misleading. Taken as a whole, simultaneous gamma UPLs are potentially quite useful for environmental monitoring applications.

In addition to the environmental monitoring application just described, there are several other important areas of application for simultaneous gamma UPLs. For industrial quality control problems, in which endpoints often have long-tailed distributions and/or heteroscedasticity, simultaneous gamma prediction limits can be used to determine whether the engineering process is in control. Yet another potential application of this methodology is in the field of molecular genetics, particularly recent advances in microarray technology (Chee et al. 1996). Although there has been considerable interest in identifying differential expression levels of genes in different pathological states (e.g., colon cancer tumor tissue vs. normal colon tissue), little work has been done in comparing expression levels obtained from a single individual or a small number of individuals with a rare condition with normal controls. Gibbons et al. (2005) considered using nonparametric UPLs in this context. Using simultaneous gamma UPLs in this context potentially can provide a more statistically powerful alternative that is consistent with the distributional form of expression-level data (Newton, Kendziorski, Richmond, Blattner, and Tsui 2001).

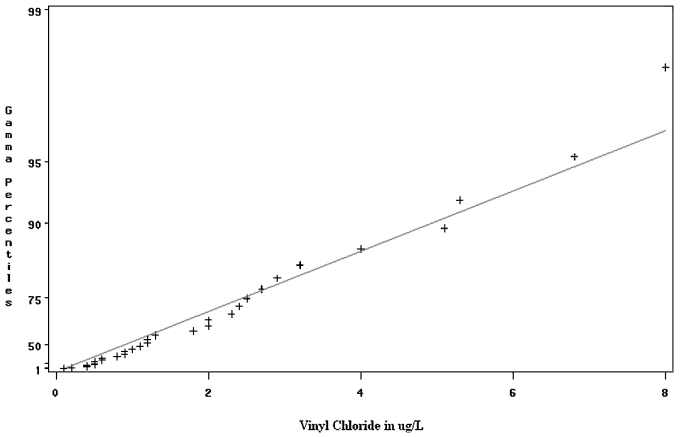


*Figure 1. Gamma Probability Plot for Vinyl Chloride.*

In the following sections we describe the statistical foundation for the gamma distribution, detail the construction of approximate gamma UPLs, and illustrate the use of the methodology in connection with an environmental monitoring application.

## 2. STATISTICAL FOUNDATION

Suppose that $x$ follows a gamma distribution with the shape parameter $\kappa$ and scale parameter $\theta$. Let $x_1, x_2, \ldots, x_n$ be a random sample of size $n$ drawn from this population to estimate the unknown parameters. Denote the arithmetic and geometric means based on this random sample by $\bar{x}$ and $\tilde{x}$. The following basic results of a gamma distribution are derived from work of Grice and Bain (1980). Here $\bar{x}$ and $\tilde{x}$ are jointly sufficient and complete statistics for $\theta$ and $\kappa$, and the distribution of $\bar{x}/\tilde{x}$ does not depend on $\theta$. Hence, using Basu's theorem (Basu 1955), we can claim that $\bar{x}$ and $\bar{x}/\tilde{x}$ are independent, where $\bar{x}/\tilde{x}$ plays the role of an ancillary statistic. In what follows we show that the independence of $\bar{x}$ and $\bar{x}/\tilde{x}$ plays a vital role in constructing the prediction limit.

The maximum likelihood estimators of $\theta$ and $\kappa$, denoted by $\hat{\theta}$ and $\hat{\kappa}$, are solutions to the following equations:

$$\ln(\hat{\kappa}) - \psi(\hat{\kappa}) = \ln(\bar{x}/\tilde{x}) \qquad \text{and} \qquad \hat{\kappa}\hat{\theta} = \bar{x}, \tag{1}$$

where $\psi$ denotes a digamma or Euler's psi function. More results on a gamma distribution have been given by Johnson, Kotz, and Balakrishnan (1994). Note that the mean and variance of $x$ are

$$E(x) = \kappa\theta \qquad \text{and} \qquad V(x) = \kappa\theta^2. \tag{2}$$

Suppose that $X$ is a new observation that follows a gamma distribution with the shape parameter $\kappa$ and scale parameter $\theta$, and $2X/\theta$ has a chi-squared distribution with degrees of freedom $2\kappa$. Denote the $(1-\alpha)100$th percentile point of a gamma distribution with $2\kappa$ degree of freedom by $\chi^2_{1-\alpha}(2\kappa)$.

Note that the exact simultaneous gamma UPL depends on the parameters of the distribution, which are unknown. In practice, we then substitute the estimates of the parameters for the true values and compute the limit based on the exact specification. Note that the UPL is no longer exact when the estimated

parameters are substituted for their true values. Furthermore, through a simulation study, we found that the behavior of these UPLs is quite poor in terms of providing the intended nominal confidence levels. In light of these results, we now propose an alternative approximate simultaneous UPL.

The approximate UPL that we propose is based on a scaled standard deviation from the sample mean. Let $S = \sqrt{\hat{\kappa}}\hat{\theta}$. Using the second expression of (2), we can claim that $S$ is an estimate of the population standard deviation. Thus the right-sided UPL for a single future observation $X$ coming from the aforementioned gamma distribution, when expressed as $\bar{x} + kS$ for a positive constant $k$, can be expressed as $\bar{x}q(k, \hat{\kappa})$, where

$$\bar{x} + kS = \bar{x}\left(1 + \frac{kS}{\bar{x}}\right) \tag{3}$$

$$= \bar{x}\left(1 + k\frac{\sqrt{\hat{\kappa}}\hat{\theta}}{\hat{\kappa}\hat{\theta}}\right) \tag{4}$$

$$= \bar{x}\left(1 + \frac{k}{\sqrt{\hat{\kappa}}}\right) = \bar{x}q(k, \hat{\kappa}) \tag{5}$$

and $(1 + k/\sqrt{\hat{\kappa}}) = q(k, \hat{\kappa})$. Equation (4) is obtained from (3) using the expression of $S$ just defined and the second part of (1). From the likelihood equation given in (1), we see that $\hat{\kappa}$ is a function of $\bar{x}/\tilde{x}$ and the UPL $\bar{x}q(k, \hat{\kappa})$ satisfies a desirable requirement, that it be a function of sufficient statistics $\bar{x}$ and $\bar{x}/\tilde{x}$. Note that when the sample size is small, the bias in the maximum likelihood estimate of $\kappa$ can be substantial; however, our method for computing the value of $k$ used in deriving the UPL is relatively unaffected by this small-sample bias. Let us denote $q(k, \hat{\kappa})$ by $Q(k, \bar{x}/\tilde{x})$, when $\hat{\kappa}$ is replaced by its solution obtained from (1).

## 3. CONSTRUCTION OF THE PREDICTION LIMIT

Constructing of simultaneous gamma UPLs involves determining the value of $k$ for a given problem based on $n$, $m$, $r$, $p$, and $1 - \alpha$. Assume as an example that we have $m$ independent samples from each of $r$ locations. The $j$th observation, $j = 1, 2, \ldots, m$, from the $i$th location, $i = 1, 2, \ldots, r$, is denoted by $y_{ij}$, and the $p$th smallest observation is denoted by $y_{i(p)}$. Let $y^* = \max_i y_{i(p)}$. Thus if $y^* \le Q(k, \bar{x}/\tilde{x})\bar{x}$, then at least $p$ observations from each of the $r$ locations will be less than $Q(k, \bar{x}/\tilde{x})\bar{x}$. As noted previously, the distribution of $2n\bar{x}/\theta$ is chi-squared with $2n\kappa$ degrees of freedom. Hence, from $y^* \le Q(k, \bar{x}/\tilde{x})\bar{x}$, we obtain

$$\frac{y^*/\theta}{2n\bar{x}/\theta} \le \frac{Q(k, \bar{x}/\tilde{x})}{2n} \quad \Rightarrow \quad \frac{2nz^*}{Q(k, \bar{x}/\tilde{x})} \le U, \tag{6}$$

where $z^* = y^*/\theta$, and $U = 2n\bar{x}/\theta$. Note that the distributions of $\bar{x}/\tilde{x}$, $z^*$ and that $U$ are independent and that $U$ follows a central chi-squared distribution with degrees of freedom (df) $2n\kappa$ (see Grice and Bain 1980). We now discuss how to determine $k$, the constant involved in the UPL.

### 3.1 Determination of $k$

$z_{ij} = y_{ij}/\theta$ follows a gamma distribution with scale parameter 1 and shape parameter $\kappa$. The density function and the cumulative density function of $z_{ij}$ are denoted by $\phi$ and $\Phi$. The variables $z_{i(p)} = y_{i(p)}/\theta$ are independently and identically distributed random variables for $i = 1, 2, \ldots, r$. The density function of $z_{i(p)}$ is denoted by $f_{z_{i(p)}}(z)$, where

$$f_{z_{i(p)}}(z) = mC_{(p-1)}^{(m-1)}\Phi^{(p-1)}(z)\phi(z)[1 - \Phi(z)]^{m-p} = g(z), \text{ say,} \tag{7}$$

where $C_i^j$ denotes the number of combinations of selecting $i$ elements out of a total of $j$ elements. Let $G(z) = \int_0^z g(t)\,dt$, the cumulative density function of $z_{i(p)}$, and $z^* = \max_i z_{i(p)}$. Denote the density function of $z^*$ by $h(z^*)$. Then

$$h(z^*) = rG^{r-1}(z^*)g(z^*) \tag{8}$$

and

$$\begin{aligned}\Pr &= P\{\text{at least } p \text{ of } m \text{ future observations are below } Q(k, \bar{x}/\tilde{x}) \cdot \bar{x} \text{ at each of } r \text{ locations}\} \\ &= P\left(y^* \le Q(k, \bar{x}/\tilde{x})\bar{x}\right) \\ &= P\left(U \ge \frac{2nz^*}{Q(k, \bar{x}/\tilde{x})}\right) \\ &= E_{\hat{\kappa}}\left(E_{z^*}\left(P_U\left(U \ge \frac{2nz^*}{Q(k, \bar{x}/\tilde{x})}\right) \Big| z^*, \hat{\kappa}\right) \Big| \hat{\kappa}\right) \\ &= E_{\hat{\kappa}}\left(E_{z^*}\left(\int_{l(k,\hat{\kappa})z^*}^{\infty} f(\chi^2)\,d\chi^2 \Big| \hat{\kappa}\right)\right),\end{aligned} \tag{9}$$

where $f(\chi^2)$ is the density function of a central chi-squared distribution with $2n\kappa$ df and $l(k, \hat{\kappa}) = 2n/Q(k, \bar{x}/\tilde{x}) = 2n/q(k, \hat{\kappa})$. Inequality (6) and the fact that $\bar{x}$ and $\bar{x}/\tilde{x}$ are independent are used to obtain the second line of (9). Using the expression of the density function of $z^*$ given in (9),

$$\begin{aligned}\Pr = E_{\hat{\kappa}}\Bigg(&\int_0^{\infty}\left(\int_{l(k,\hat{\kappa})z^*}^{\infty} f(\chi^2)\,d\chi^2\right)r \\ &\times\left[\int_0^{z^*} mC_{(p-1)}^{(m-1)}\Phi^{(p-1)}(t)\phi(t)[1 - \Phi(t)]^{m-p}\,dt\right]^{r-1} \\ &\times mC_{(p-1)}^{(m-1)}\Phi^{(p-1)}(z^*)\phi(z^*)[1 - \Phi(z^*)]^{m-p}\,dz^* \Big| \hat{\kappa}\Bigg).\end{aligned} \tag{10}$$

We further simplify the expression given in the right side of (10). Let $v = \Phi(z^*)$, $B(a, b)$ be the usual beta function with parameters $a$ and $b$, and let $I(x, a, b)$ be the cumulative density function of a beta random variable with parameters $a$ and $b$. Using these transformations, we obtain

$$\begin{aligned}\Pr = E_{\hat{\kappa}}\Bigg(&\int_0^1\left(\int_{l(k,\hat{\kappa})\Phi^{-1}(v)}^{\infty} f(\chi^2)\,d\chi^2\right) \\ &\times rI^{r-1}(v, p, m + 1 - p)\frac{v^{p-1}(1 - v)^{m-p}}{B(p, m + 1 - p)}\,dv \Big| \hat{\kappa}\Bigg).\end{aligned} \tag{11}$$

Let $A = (\int_0^1(\int_{l(k,\hat{\kappa})\Phi^{-1}(v)}^{\infty} f(\chi^2)\,d\chi^2)rI^{r-1}(v, p, m + 1 - p) \times \frac{v^{p-1}(1-v)^{m-p}}{B(p,m+1-p)}\,dv)$. Note that $A$ depends on $\hat{\kappa}$, and hence it is subject to the sampling variation. We compute the right side of (11) numerically.

Several estimators of gamma distribution parameters, along with the estimators' asymptotic distributions, were given by Johnson et al. (1994). To estimate the parameters of this gamma distribution, in addition to the maximum likelihood procedure, we have also attempted two other procedures given in 17.57a, 17.57b, and 17.58 of Johnson et al. (1994, p. 364). For various values of $\bar{x}/\tilde{x}$, we observe that all three procedures produce quite similar values of $\hat{\kappa}$. To compute the right side of (11), we use the asymptotic normal distribution of $\hat{\kappa}$. We estimate the asymptotic mean and variance of $\hat{\kappa}$ following the expressions given by Bowman and Shenton (1988). Computation of the right side of (11) involves the central chi-squared distribution, inverse-gamma cdf, incomplete gamma function, and beta function. We use subroutines from Press, Teukolsky, Vetterling, and Flannery (1997) for these computations. To determine $k$, we numerically solve the equation $\Pr = 1 - \alpha$, where Pr is defined in (11) and $1 - \alpha$ is the confidence level for the prediction limit.

## 4. ILLUSTRATION

### 4.1 Analysis of the Vinyl Chloride Data

Returning to the vinyl chloride data given in Table 1 and Figure 1, we now illustrate computation of simultaneous gamma UPLs for various values of $r$, $p$, and $m$. To begin, we have $n = 34$ vinyl chloride measurements with mean $\bar{x} = 1.879$ μg/L and standard deviation $(S) = 1.823$ μg/L. Our estimates of $\kappa$ and $\theta$ are $\hat{\kappa} = 1.063$ and $\hat{\theta} = 1.769$. Now suppose that we have a site with a single future monitoring location, $r = 1$, and that in the event of an initial exceedance of the UPLs a single verification resample is permitted, failure being indicated only if both the initial and resample both exceed the UPL. This is the same as requiring that at least $p = 1$ out of $m = 2$ samples are in bounds. In this case the 95% confidence UPL is obtained as $\bar{x} + .576S = 2.931$ μg/L. Now consider the same sampling scheme but with the number of monitoring locations increased to $r = 10$. In this case the 95% confidence UPL is obtained as $\bar{x} + 1.835S = 5.224$ μg/L. Shifting the resampling program to including one of two resamples in bounds in the event of an initial exceedance (i.e., $p = 1$ and $m = 3$) decreases the UPL to $\bar{x} + .901S = 3.521$ μg/L. Finally, increasing the number of resamples in bounds from one to two (i.e., $p = 2$ and $m = 3$) increases the UPL to $\bar{x} + 2.441S = 6.330$ μg/L.

To determine the accuracy of the computed UPLs, we computed actual confidence levels via simulation. Using the estimated parameters listed before, we simulated 34 new background measurements from a gamma distribution with parameters $\kappa = 1.063$ and $\theta = 1.769$. For each of the foregoing scenarios, (a) $r = 1$, $p = 1$, and $m = 2$; (b) $r = 10$, $p = 1$, and $m = 2$; (c) $r = 10$, $p = 1$, and $m = 3$; and (d) $r = 10$, $p = 2$, and $m = 3$; UPLs were computed for the simulated background data ($n = 34$). Based on the specific conditions listed in (a)–(d), $m$ new downgradient monitoring measurements were generated at each of $r$ locations. If fewer than $p$ of $m$ of those generated measurements were in bounds at any of the $r$ locations, then a failure was recorded. This process was then repeated 10,000 times, and the simulated confidence levels were computed. Simulated confidence levels were .9511, .9408, .9506, and .9417 for conditions (a)–(d). These results indicate that the UPLs achieve their intended confidence levels.

### 4.2 Simulation Study

Although these results are encouraging, they are based on a single condition for a large background sample. To examine the robustness of these results to other values of $\kappa$ and smaller background sample sizes, we performed a limited simulation study. To determine whether our gamma UPLs offer any advantages over normal UPLs, we also included the normal simultaneous UPLs described by Davis and McNichols (1987). In our study we set $\theta = 1.0$ and considered values of $\kappa$ equal to .25, .5, 1.0, and 1.5. Background sample sizes of $n = 8$, 20, 50, and 100 were examined. Finally, we examined three sampling plans: (a) $p = 1$ and $m = 2$, (b) $p = 1$ and $m = 3$, and (c) $p = 2$ and $m = 3$. In all cases, we considered $r = 10$ locations. Results of the simulation study are presented in Table 2.

Inspection of Table 2 reveals that the gamma UPLs have type I error rates that are close to the nominal level of 5% in all cases for $n \geq 20$. For $n = 8$, the type I error rates generally are remarkably close to the intended nominal rate of 5%, even when $\kappa$ is as low as .25. We note that the worst behavior for $n = 8$ is when $p = 2$ and $m = 3$, a condition rarely encountered in practice. Nonetheless, the simulated type I error rate is no more than 10%. In contrast, the normal UPLs performed quite poorly throughout. For $p = 2$ and $m = 3$, the normal UPLs had simulated type I error rates as high as 25%. Conversely, for $p = 1$ and $m = 3$, the normal UPLs were in many cases overly conservative, with simulated type I error rates as low as 1%.

### 4.3 Statistical Power

To investigate the statistical power of the procedure and to compare it against normal and nonparametric alternatives, a second limited simulation study was performed. The background distribution was gamma with $\kappa = 1.0$ and $\theta = .2$; $n = 20$ background measurements and $m = 2$ future measurements were generated for each of $r = 13$ locations; and $r = 13$ was selected because the 95% confidence nonparametric prediction limit for pass one of two samples in each of 13 locations is defined as the maximum of $n = 20$ background measurements. The alternative hypothesis was generated by sampling the future samples out a distribution with $\kappa = 1.0$ and $\theta = .4$, .6, .8, and 1.0. The resulting power curve, displayed in Figure 2, reveals that both the gamma and nonparametric UPLs achieve their nominal type I error rate of 5% (at $\theta = .2$), but the normal limit has a type I error rate of approximately 10%. In terms of power, the normal UPL is most powerful, but at the expense of an inflated false-positive rate. The gamma UPL limit achieves its intended 5% false-positive rate and has increased power relative to the nonparametric alternative throughout the range of effect sizes.

### 4.4 Outliers

From time to time, it is not at all uncommon to have an occasional elevated value or outlier of the distribution of measure-

*Table 2. Simulation Study of Gamma and Normal Simultaneous UPLs, $\theta = 1.0$ and $r = 10$ Locations*

| Sampling plan p of m | n | $\kappa$ | Type I error rate | |
|---|---|---|---|---|
| | | | Gamma | Normal |
| 1 of 2 | 8 | .25 | .065 | .158 |
| | 20 | .25 | .059 | .100 |
| | 50 | .25 | .050 | .066 |
| | 100 | .25 | .050 | .057 |
| | 8 | .50 | .074 | .136 |
| | 20 | .50 | .057 | .097 |
| | 50 | .50 | .061 | .099 |
| | 100 | .50 | .051 | .071 |
| | 8 | 1.00 | .087 | .135 |
| | 20 | 1.00 | .062 | .096 |
| | 50 | 1.00 | .052 | .085 |
| | 100 | 1.00 | .059 | .080 |
| | 8 | 1.50 | .078 | .103 |
| | 20 | 1.50 | .059 | .098 |
| | 50 | 1.50 | .053 | .082 |
| | 100 | 1.50 | .058 | .082 |
| 1 of 3 | 8 | .25 | .056 | .072 |
| | 20 | .25 | .052 | .034 |
| | 50 | .25 | .050 | .019 |
| | 100 | .25 | .050 | .014 |
| | 8 | .50 | .050 | .065 |
| | 20 | .50 | .043 | .036 |
| | 50 | .50 | .044 | .030 |
| | 100 | .50 | .044 | .019 |
| | 8 | 1.00 | .059 | .058 |
| | 20 | 1.00 | .051 | .049 |
| | 50 | 1.00 | .052 | .036 |
| | 100 | 1.00 | .041 | .025 |
| | 8 | 1.50 | .079 | .081 |
| | 20 | 1.50 | .056 | .052 |
| | 50 | 1.50 | .056 | .044 |
| | 100 | 1.50 | .048 | .035 |
| 2 of 3 | 8 | .25 | .073 | .248 |
| | 20 | .25 | .062 | .183 |
| | 50 | .25 | .053 | .131 |
| | 100 | .25 | .050 | .111 |
| | 8 | .50 | .088 | .221 |
| | 20 | .50 | .051 | .166 |
| | 50 | .50 | .071 | .166 |
| | 100 | .50 | .045 | .110 |
| | 8 | 1.00 | .102 | .183 |
| | 20 | 1.00 | .066 | .149 |
| | 50 | 1.00 | .056 | .154 |
| | 100 | 1.00 | .048 | .116 |
| | 8 | 1.50 | .105 | .148 |
| | 20 | 1.50 | .076 | .142 |
| | 50 | 1.50 | .063 | .144 |
| | 100 | 1.50 | .063 | .135 |

ments. The question is how robust the gamma UPLs are to the presence of outliers. To shed light on this question, we simulated data with a single outlier (set at the 95th percentile of the background distribution) for $n = 20$ (i.e., 5% outliers), and two outliers for $n = 50$ (i.e., 4% outliers). We based the simulation on $p = 1$, $m = 2$, $r = 10$, $\kappa = .5$, and $\theta = 1.0$. Results of the simulation with outliers revealed small decreases in the type I error rate from 5% to 3.0% for $n = 50$ and to 3.2% for $n = 20$. Note that an outlier study for $n = 8$ is not meaningful, because even a single outlier results in 12% outliers. These results suggest that the gamma UPLs are fairly robust to the presence of a small number of extreme outlying values; however, every attempt should be made to screen the background data for anomalous values before the analysis.

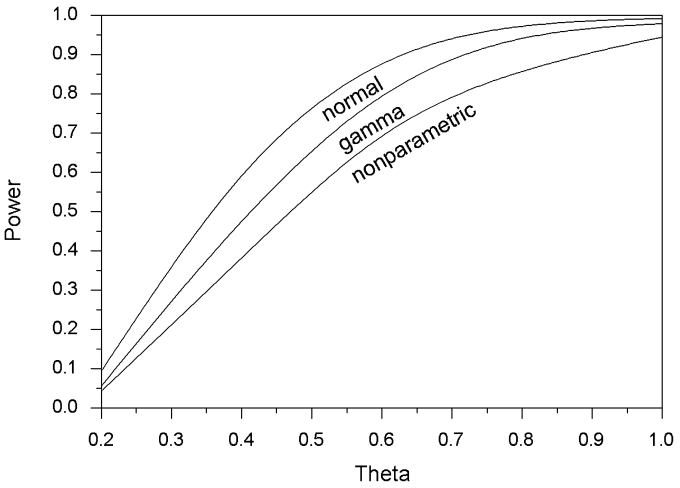


*Figure 2. Power Curves for Gamma, Normal, and Nonparametric UPLs. Background $\theta = .2$, $\kappa = 1.0$, $p = 1$, $m = 2$, and $r = 13$.*

## 5. THE PROBLEM OF NONDETECTS

A potential complication of the methodology described thus far is that it is common for a proportion of the analytical measurements to be reported as less than a limit of detection (Currie 1968). Even if every effort is made to limit censoring of the data by reporting all measured concentrations, there will still be instances where the analyte is simply not detected in the sample and no concentration estimate is available. In such cases the U.S. Environmental Protection Agency (USEPA) (1992) has advised that either one-half of the detection limit be imputed for the nondetected measurements, or that, under the assumption of normality or lognormality, either Aitchison's (1955) adjustment or Cohen's (1959, 1961) estimator be used to provide estimates of the mean and variance of the censored distribution. The adjusted mean and variance are then used in computing normal or lognormal UPLs with only limited success (see Gibbons 1994, pp. 214–215). Alternatively, when censoring is 50% or more, the USEPA has advocated using nonparametric UPLs as described by (Gibbons 1990, 1991; Davis and McNichols 1999).

Because the volatile organic compounds used to illustrate the use of simultaneous gamma UPLs in this article are plagued with problems of nondetects, we studied the behavior of these UPLs for various degrees of censoring in a limited simulation study based on the vinyl chloride data in the previous example for $r = 1$, $p = 1$, and $m = 2$. We studied the effects of 10%, 25%, and 50% censoring on the simulated confidence levels obtained over 10,000 simulations. As in the previous illustration, the complete data ($n = 34$ measurements) were generated from a gamma distribution with parameters $\kappa = 1.063$ and $\theta = 1.769$. Given these population parameters, to achieve 10%, 25%, and 50% censoring, the data were censored at the following cutpoints: .2209, .5745, and 1.3341. A separate simulation study was performed for each cutpoint (i.e., level of censoring). Any simulated measurement below one of these censoring points was replaced by one-half of the value of the censoring point. This is consistent with the USEPAs use of imputation of

*Table 3. Cutpoints and the Corresponding Simulated Confidence Levels for Various Combinations of Gamma Parameters at Different Censoring Levels*

| | Censoring levels | | |
|---|---|---|---|
| *Parameters* | *10%* | *25%* | *50%* |
| $\kappa = 1.063$ | .22090[a] | .57450 | 1.33410 |
| $\theta = 1.769$ | (.9506)[b] | (.9491) | (.9468) |
| $\kappa = .5$ | .00789 | .05076 | .22740 |
| $\theta = 1.0$ | (.9523) | (.9529) | (.9552) |
| $\kappa = .25$ | .0000675 | .00264 | .04376 |
| $\theta = 1.0$ | (.9525) | (.9553) | (.9590) |

[a]Cutpoints.
[b]Simulated confidence levels.

one-half of the detection limit for nondetects. Results of this study revealed that the simultaneous gamma prediction limits are in fact remarkably robust to both modest and more extreme levels of censoring. Simulated confidence levels were .9506 for 10% censoring, .9491 for 25% censoring, and .9468 for 50% censoring.

To ensure the method's robustness to censored data, generalized to other values of the population parameters, we repeated the simulation study for gamma distributions with parameters $\kappa = .5$ and $\theta = 1.0$ and $\kappa = .25$ and $\theta = 1.0$. Table 3 reveals that the robustness observed for the original example data generalizes to other values of $\kappa$ and $\theta$ as well.

The net result is that the routine use of gamma UPLs for data that contain nondetected concentrations will yield unbiased UPL estimates with the intended nominal type I error rate. For routine practice, this is a major advantage of this methodology.

## 6. TABLED VALUES OF GAMMA UPPER PREDICTION LIMIT FACTORS

To aid applications, we have constructed Table 4, which contains gamma UPL factors $k$ for various combinations of $n$, $r$, $p$, $m$, and $\hat{\kappa}$ commonly encountered in practice. Using the values of $k$ given in Table 4, the UPL $\bar{x} + kS$ can be computed directly. Furthermore, the values of $\hat{\kappa}$ (5.16, 1.14, .62, .43, and .34) correspond to values of $\bar{x}/\tilde{x}$ equal to .10, .50, 1.00, 1.50, and 2.00. As such, only the statistics $\bar{x}$, $\tilde{x}$, and $S$ are required. In terms of the sampling plan, Table 4 provides values of $k$ for $p = 1$ and $m = 2$, $p = 1$ and $m = 3$, and $p = 2$ and $m = 3$, the three primary sampling plans used in most environmental monitoring programs. The table includes background sample sizes of $n = 8$, 16, 32, and 48 and $r = 1$, 2, 4, 8, 16, 32, and 64 locations. These values of $n$, $r$, $p$, $m$, and $\hat{\kappa}$ should cover a great many routine applications and can be easily interpolated to include others.

Inspection of Table 4 reveals several interesting features. First, for $p = 1$ and $m = 3$, some of the values of $k$ are negative. This indicates that when we require only one of three measurements in bounds, the gamma UPL can be less than the arithmetic mean for certain combinations of $n$, $r$, and $\hat{\kappa}$. Second, the largest values of $k$ are found for the smallest values of $n$ and $\hat{\kappa}$. Third, as expected, increasing the number of future comparisons $r$ increases $k$. Finally, of the three sampling plans, one of three in bounds produces the smallest UPLs, followed by one of two in bounds, followed by two of three in bounds.

*Table 4. Gamma UPL Factors k as a Function of n, r, p, m, and $\hat{\kappa}$, $UPL = \bar{x} + kS$*

| | | | *k* | | |
|---|---|---|---|---|---|
| *n* | *r* | $\hat{\kappa}$ | *p = 1, m = 2* | *p = 1, m = 3* | *p = 2, m = 3* |
| 8 | 1 | 5.16 | .913 | .398 | 1.395 |
| | 2 | 5.16 | 1.252 | .680 | 1.761 |
| | 4 | 5.16 | 1.589 | .941 | 2.100 |
| | 8 | 5.16 | 1.948 | 1.219 | 2.440 |
| | 16 | 5.16 | 2.264 | 1.470 | 2.767 |
| | 32 | 5.16 | 2.566 | 1.716 | 3.063 |
| | 64 | 5.16 | 2.911 | 1.948 | 3.395 |
| | 1 | 1.14 | .852 | .257 | 1.522 |
| | 2 | 1.14 | 1.306 | .560 | 2.098 |
| | 4 | 1.14 | 1.829 | .919 | 2.648 |
| | 8 | 1.14 | 2.351 | 1.273 | 3.206 |
| | 16 | 1.14 | 2.908 | 1.645 | 3.817 |
| | 32 | 1.14 | 3.553 | 2.040 | 4.448 |
| | 64 | 1.14 | 4.114 | 2.494 | 5.040 |
| | 1 | .62 | .790 | .133 | 1.596 |
| | 2 | .62 | 1.353 | .462 | 2.296 |
| | 4 | .62 | 2.021 | .852 | 3.047 |
| | 8 | .62 | 2.726 | 1.283 | 3.903 |
| | 16 | .62 | 3.578 | 1.791 | 4.724 |
| | 32 | .62 | 4.493 | 2.296 | 5.893 |
| | 64 | .62 | 5.252 | 2.827 | 6.685 |
| | 1 | .43 | .711 | .040 | 1.636 |
| | 2 | .43 | 1.332 | .365 | 2.450 |
| | 4 | .43 | 2.121 | .786 | 3.414 |
| | 8 | .43 | 2.990 | 1.295 | 4.648 |
| | 16 | .43 | 4.161 | 1.855 | 5.993 |
| | 32 | .43 | 5.245 | 2.545 | 6.958 |
| | 64 | .43 | 6.443 | 3.622 | 8.250 |
| | 1 | .34 | .642 | −.025 | 1.631 |
| | 2 | .34 | 1.316 | .296 | 2.619 |
| | 4 | .34 | 2.232 | .728 | 3.831 |
| | 8 | .34 | 3.282 | 1.278 | 5.198 |
| | 16 | .34 | 4.561 | 1.929 | 7.103 |
| | 32 | .34 | 6.016 | 2.733 | 8.620 |
| | 64 | .34 | 7.793 | 3.631 | 10.882 |
| 16 | 1 | 5.16 | .792 | .304 | 1.252 |
| | 2 | 5.16 | 1.101 | .543 | 1.569 |
| | 4 | 5.16 | 1.413 | .792 | 1.876 |
| | 8 | 5.16 | 1.716 | 1.028 | 2.180 |
| | 16 | 5.16 | 2.023 | 1.252 | 2.471 |
| | 32 | 5.16 | 2.321 | 1.470 | 2.732 |
| | 64 | 5.16 | 2.598 | 1.695 | 3.024 |
| | 1 | 1.14 | .685 | .136 | 1.258 |
| | 2 | 1.14 | 1.069 | .399 | 1.689 |
| | 4 | 1.14 | 1.502 | .685 | 2.157 |
| | 8 | 1.14 | 1.931 | .979 | 2.608 |
| | 16 | 1.14 | 2.385 | 1.273 | 3.052 |
| | 32 | 1.14 | 2.817 | 1.582 | 3.553 |
| | 64 | 1.14 | 3.316 | 1.905 | 4.036 |
| | 1 | .62 | .577 | .016 | 1.217 |
| | 2 | .62 | 1.016 | .277 | 1.739 |
| | 4 | .62 | 1.509 | .577 | 2.296 |
| | 8 | .62 | 2.021 | .895 | 2.879 |
| | 16 | .62 | 2.585 | 1.249 | 3.504 |
| | 32 | .62 | 3.167 | 1.619 | 4.181 |
| | 64 | .62 | 3.817 | 2.021 | 4.847 |
| | 1 | .43 | .481 | −.070 | 1.176 |
| | 2 | .43 | .939 | .178 | 1.768 |
| | 4 | .43 | 1.473 | .481 | 2.405 |
| | 8 | .43 | 2.085 | .827 | 3.121 |
| | 16 | .43 | 2.753 | 1.208 | 3.851 |
| | 32 | .43 | 3.494 | 1.636 | 4.648 |
| | 64 | .43 | 4.274 | 2.085 | 5.597 |
| | 1 | .34 | .408 | −.128 | 1.124 |
| | 2 | .34 | .883 | .107 | 1.741 |
| | 4 | .34 | 1.461 | .403 | 2.462 |
| | 8 | .34 | 2.110 | .757 | 3.282 |
| | 16 | .34 | 2.855 | 1.156 | 4.168 |
| | 32 | .34 | 3.729 | 1.631 | 5.198 |
| | 64 | .34 | 4.706 | 2.110 | 6.258 |

*Table 4 (continued).*

| | | | $k$ | | |
|---|---|---|---|---|---|
| $n$ | $r$ | $\hat{\kappa}$ | $p=1, m=2$ | $p=1, m=3$ | $p=2, m=3$ |
| 32 | 1 | 5.16 | .735 | .254 | 1.177 |
| | 2 | 5.16 | 1.036 | .484 | 1.470 |
| | 4 | 5.16 | 1.331 | .710 | 1.761 |
| | 8 | 5.16 | 1.620 | .927 | 2.035 |
| | 16 | 5.16 | 1.887 | 1.137 | 2.307 |
| | 32 | 5.16 | 2.167 | 1.340 | 2.566 |
| | 64 | 5.16 | 2.425 | 1.539 | 2.820 |
| | 1 | 1.14 | .604 | .081 | 1.130 |
| | 2 | 1.14 | .961 | .322 | 1.522 |
| | 4 | 1.14 | 1.339 | .575 | 1.918 |
| | 8 | 1.14 | 1.723 | .836 | 2.317 |
| | 16 | 1.14 | 2.112 | 1.095 | 2.731 |
| | 32 | 1.14 | 2.512 | 1.365 | 3.127 |
| | 64 | 1.14 | 2.908 | 1.634 | 3.553 |
| | 1 | .62 | .487 | −.036 | 1.062 |
| | 2 | .62 | .873 | .196 | 1.509 |
| | 4 | .62 | 1.300 | .459 | 1.975 |
| | 8 | .62 | 1.739 | .733 | 2.476 |
| | 16 | .62 | 2.204 | 1.029 | 2.961 |
| | 32 | .62 | 2.701 | 1.335 | 3.504 |
| | 64 | .62 | 3.198 | 1.654 | 4.038 |
| | 1 | .43 | .388 | −.117 | .989 |
| | 2 | .43 | .786 | .100 | 1.473 |
| | 4 | .43 | 1.242 | .358 | 1.998 |
| | 8 | .43 | 1.727 | .643 | 2.544 |
| | 16 | .43 | 2.256 | .951 | 3.155 |
| | 32 | .43 | 2.838 | 1.286 | 3.756 |
| | 64 | .43 | 3.414 | 1.636 | 4.393 |
| | 1 | .34 | .315 | −.169 | .924 |
| | 2 | .34 | .715 | .033 | 1.439 |
| | 4 | .34 | 1.189 | .284 | 1.998 |
| | 8 | .34 | 1.713 | .567 | 2.619 |
| | 16 | .34 | 2.297 | .883 | 3.282 |
| | 32 | .34 | 2.919 | 1.232 | 3.397 |
| | 64 | .34 | 3.584 | 1.605 | 4.706 |
| 48 | 1 | 5.16 | .721 | .238 | 1.147 |
| | 2 | 5.16 | 1.014 | .462 | 1.444 |
| | 4 | 5.16 | 1.304 | .684 | 1.724 |
| | 8 | 5.16 | 1.583 | .899 | 1.989 |
| | 16 | 5.16 | 1.852 | 1.105 | 2.254 |
| | 32 | 5.16 | 2.109 | 1.299 | 2.513 |
| | 64 | 5.16 | 2.370 | 1.496 | 2.755 |
| | 1 | 1.14 | .578 | .064 | 1.095 |
| | 2 | 1.14 | .927 | .296 | 1.470 |
| | 4 | 1.14 | 1.290 | .542 | 1.846 |
| | 8 | 1.14 | 1.660 | .790 | 2.230 |
| | 16 | 1.14 | 2.031 | 1.042 | 2.608 |
| | 32 | 1.14 | 2.409 | 1.295 | 2.986 |
| | 64 | 1.14 | 2.788 | 1.548 | 3.373 |
| | 1 | .62 | .460 | −.051 | 1.012 |
| | 2 | .62 | .831 | .172 | 1.435 |
| | 4 | .62 | 1.233 | .420 | 1.882 |
| | 8 | .62 | 1.658 | .683 | 2.334 |
| | 16 | .62 | 2.096 | .958 | 2.809 |
| | 32 | .62 | 2.555 | 1.244 | 3.273 |
| | 64 | .62 | 3.008 | 1.537 | 3.762 |
| | 1 | .43 | .360 | −.131 | .931 |
| | 2 | .43 | .738 | .077 | 1.390 |
| | 4 | .43 | 1.165 | .321 | 1.876 |
| | 8 | .43 | 1.628 | .587 | 2.390 |
| | 16 | .43 | 2.121 | .873 | 2.927 |
| | 32 | .43 | 2.628 | 1.181 | 3.494 |
| | 64 | .43 | 3.167 | 1.502 | 4.053 |
| | 1 | .34 | .286 | −.181 | .864 |
| | 2 | .34 | .667 | .012 | 1.342 |
| | 4 | .34 | 1.109 | .246 | 1.863 |
| | 8 | .34 | 1.596 | .513 | 2.413 |
| | 16 | .34 | 2.123 | .803 | 3.009 |
| | 32 | .34 | 2.675 | 1.119 | 3.631 |
| | 64 | .34 | 3.282 | 1.461 | 4.250 |

## 7. DISCUSSION

Simultaneous gamma UPLs should provide a useful addition to the arsenal of statistical methods useful in quality control applications in general and environmental statistics in particular. The methodology presented here extends earlier work for the normal distribution (Davis and McNichols 1987) and nonparametric alternatives (Gibbons 1990, 1991; Davis and McNichols 1999) to the case of a gamma-distributed random variable. Use of the gamma distribution permits association between the mean and variance of the distribution, a phenomenon commonly observed in practice. Furthermore, the gamma distribution permits analysis of skewed distributions, only some of which were previously amenable to computation based on a lognormal assumption. Results of a limited simulation study revealed that in contrast with simultaneous gamma UPLs, simultaneous normal UPLs do not achieve their intended nominal type I error rate when applied to data generated from a gamma distribution. Simultaneous nonparametric UPLs do achieve their intended nominal type I error rate, but have reduced statistical power relative to simultaneous gamma UPLs. Finally, we have shown that the gamma UPLs are remarkably robust to censoring of the data based on limits of detection, a condition that typifies environmental data in most, if not all, areas of application. Of course, using the gamma distribution and corresponding UPLs goes well beyond the environmental sciences. Recent statistical work in molecular genetics has revealed that gene expression levels measured using microarray technology are accurately modeled using a gamma distribution (Newton et al. 2001).

*[Received October 2002. Revised February 2005.]*

## REFERENCES


Aitchison, J. (1955), "On the Distribution of a Positive Random Variable Having a Discrete Probability Mass at the Origin," *Journal of the American Statistical Association*, 50, 901–908.

Barlow, R. E., and Proschan, F. (1965), *Mathematical Theory of Reliability*, New York: Wiley.

Barnett, V., and Lewis, T. (1984), *Outliers in Statistical Data* (2nd ed.), New York: Wiley.

Basu, D. (1955), "On Statistics Independent of a Complete Sufficient Statistic," *Sankhyā*, 15, 377–386.

Bowman, K. O., and Shenton, L. R. (1988), *Properties of Estimators for the Gamma Distribution*, New York: Marcel Dekker.

Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S., and Fodor, S. P. (1996), "Accessing Genetic Information With High-Density DNA Microarrays," *Science*, 274, 610–614.

Cohen, A. C. (1959), "Simplified Estimators for the Normal Distribution When Samples Are Singly Censored or Truncated," *Technometrics*, 1, 217–237.

——— (1961), "Tables for Maximum Likelihood Estimates: Singly Truncated and Singly Censored Samples," *Technometrics*, 3, 535–541.

Currie, L. A. (1968), "Limits for Qualitative Detection and Quantitative Determination," *Analytical Chemistry*, 40, 586–593.

Davis, C. B., and McNichols, R. J. (1987), "One-Sided Intervals for at Least $p$ of $m$ Observations From a Normal Population on Each of $r$ Future Occasions," *Technometrics*, 29, 359–370.

——— (1999), "Simultaneous Nonparametric Prediction Limits," *Technometrics*, 41, 89–101.

Davis, D. J. (1952), "An Analysis of Some Failure Data," *Journal of the American Statistical Association*, 47, 113–150.

Gibbons, R. D. (1987), "Statistical Prediction Intervals for the Evaluation of Ground-Water Quality," *Ground Water*, 25, 455–265.

——— (1990), "A General Statistical Procedure for Ground-Water Detection Monitoring at Waste Disposal Facilities," *Ground Water*, 28, 235–243.

——— (1991), "Some Additional Nonparametric Prediction Limits for Ground-Water Detection Monitoring at Waste Disposal Facilities," *Ground Water*, 29, 729–736.

Gibbons, R. D. (1994), *Statistical Methods for Groundwater Monitoring*, New York: Wiley.

Gibbons, R. D., Bhaumik, D. K., Cox, D. R., Grayson, D. R., Davis, J. M., and Sharma, R. P. (2005), "Sequential Prediction Bounds for Identifying Differentially Expressed Genes in Replicated Microarray Experiments," *Journal of Statistical Planning and Inference*, 129, 19–37.

Gibbons, R. D., and Coleman, D. E. (2001), *Statistical Methods for Detection and Quantification of Environmental Contamination*, New York: Wiley.

Grice, J. V., and Bain, L. J. (1980), "Inferences Concerning the Mean of the Gamma Distribution," *Journal of the American Statistical Association*, 75, 929–933.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994), *Continuous Univariate Distributions—1* (2nd ed.), New York: Wiley.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. (2001), "On Differential Variability of Expression Ratios: Improving Statistical Inference About Gene Expression Changes From Microarray Data," *Journal of Computational Biology*, 8, 37–52.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1997), *Numerical Recipes in Fortran 77*, Vol. 1 (2nd ed.), Cambridge, U.K.: Cambridge University Press.

U.S. Environmental Protection Agency (1992), *Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities*, Office of Solid Waste.