ESTIMATING PERCENTILES FROM ENVIRONMENTAL SAMPLES WHEN ALL OBSERVATIONS ARE NONDETECTABLE

by

Dennis E. Smith
and
Kevin C. Burns

— STATISTICS —

— OPERATIONS RESEARCH —

— MATHEMATICS —

DESMATICS, INC.

P.O. Box 618
State College, PA 16804

DESMATICS, INC.

P.O. Box 618
State College, PA 16804
Phone: (814)-238-9621
Fax: (814)-238-5731
e-mail: desmat1@aol.com

Applied Research in Statistics and Systems Analysis

ESTIMATING PERCENTILES FROM ENVIRONMENTAL SAMPLES WHEN ALL OBSERVATIONS ARE NONDETECTABLE

by

Dennis E. Smith
and
Kevin C. Burns

Technical Report No. 164-1

June 1996

This research was supported by the Naval Research Laboratory
under Contract No. N00014-96-C-2014

TABLE OF CONTENTS

1. PROBLEM DEFINITION 1
2. BACKGROUND 2
3. PROPOSED PROCEDURE 3
4. MATHEMATICAL DETAILS 4
5. SOME RESULTS 6
6. REFERENCES 8

ABSTRACT

Many environmental sampling problems involve some specified regulatory or contractual limit (RL). Often the interest is in estimating the percentile of the underlying contaminant concentration distribution corresponding to RL. In previous reports, we have discussed the problem of determining a lower $100(1-\alpha)\%$ confidence limit for that percentile when no observations are observable, but are all known to be less than a detection limit DL, where $DL \leq RL$. In this report we extend those results to the situation in which more than a single detection limit is involved.

1. PROBLEM DEFINITION

Many environmental sampling problems involve some specified regulatory or contractual limit (RL). Such problems exist whether sampling air, water, soil, or living organisms. For example, one might be analyzing air samples in buildings for CO, water samples from lakes for pesticides, soil samples from dump sites for arsenic, or leaf samples from trees for lead. Often the interest is in estimating p_{RL}, a specified percentile of the underlying contaminant concentration distribution corresponding to RL.

The problem addressed in this paper is the estimation of the desired percentile p_{RL} based on a sample of n observations, all of which are nondetectable, where an observation x_i is known only to be less than some detection limit $DL_i \leq RL$. That is, we are considering a sample in which all observations are censored. We will assume, of course, that the sample is a representative sample.

Given n observations, each known to be less than RL, a binomial lower limit on p_{RL} is given by:

$$p_{RL} \leq \alpha^{1/n},$$

where $(1-\alpha)$ is the desired confidence level. This lower limit makes no use of the information that the observation x_i is less than DL_i (which may be much less than RL).

2. BACKGROUND

There are a number of procedures that have been proposed for dealing with estimation problems when some observations in a set of data are censored, and reported only as less than a detection limit. These include, for example, simple substitution methods, maximum likelihood estimation, and regression methods.

Haas and Scheff [1] and Helsel and Gilliom [2] have evaluated the performance of a number of suggested approaches. Any of the methods can be used to provide an estimate of p_{RL} when there are a number of uncensored observations in the sample. However, none can be used to deal with the problem defined in the previous paragraphs.

In two previous reports [4,5], we proposed a procedure that is applicable to the situation where all observations are left-censored at the same value $DL \leq RL$. Our first report [4] was based on the assumption of an underlying lognormal distribution. That is the usual assumption for contaminants present in small quantities.

However, there are some cases in which the assumption of a normal distribution may be more reasonable. For example, if the cost of sampling is small relative to the cost of chemical analysis, composite samples may be used. Whatever the underlying distribution of contaminant concentrations, the distribution of concentrations in the composite samples will tend toward normality. Therefore, our second report [5] was based on the assumption of an underlying normal distribution. This report extends the results of that report by considering the situation in which multiple detection limits are involved.

3. PROPOSED PROCEDURE

Given a sample $\underline{x} = \{x_1, x_2, \ldots x_n\}$ from the distribution of the random variable X, we want a lower $100(1-\alpha)\%$ confidence limit for $p_{RL} = Pr\{X < RL\}$. It is assumed that X is normally distributed and that each observation $x_i \le DL_i \le RL$, where DL_i denotes the detection limit for the ith observation and RL denotes the regulatory limit of interest.

The usual confidence limit for a percentile, which is also known as a tolerance limit, is of the form $\bar{x} + k \cdot s$, where $\bar{x}$ and s are the sample mean and standard deviation, respectively. A tolerance limit p* for p_{RL} can be expressed as:

$$Pr\{Pr(X \le \bar{x} + k \cdot s) \ge p^*\} = (1-\alpha).$$

Of course, in the situation we are considering, none of the x_i values are known.

However, since larger values of k correspond to larger values of p*, a conservative lower bound for p_{RL} can be found by minimizing k subject to the restriction that $\bar{x} + k \cdot s = RL$. That is, we want to minimize $k = (RL - \bar{x})/s$ subject to the constraints $0 \le x_i \le DL_i$ for all i. This procedure finds the worst-case sample, subject to the constraints. It is shown in the next section that each of the n observations in this worst-case sample is either equal to the corresponding detection limit DL_i or is equal to zero.

Given k, a lower bound for p_{RL} can be found from a table of normal tolerance limits, using the desired confidence level. If the required software is available, exact values can be obtained using the noncentral t distribution function, as described in the next section. An extensive discussion of the noncentral t distribution and its use in computing tolerance limits can be found in [3].

4. MATHEMATICAL DETAILS

Our objective is to minimize, subject to the constraints that $0 \le x_i \le DL_i \le RL$, the function $(RL - \bar{x})/s$. Since this function is positive in this interval, this is equivalent to maximizing its reciprocal. For analytical convenience, we work with the squared reciprocal:

$$f(\underline{x}) = \frac{s^2}{(RL - \bar{x})^2} = \frac{\sum(x_i - \bar{x})^2}{(n-1)(RL - \bar{x})^2}$$

Consider the partial derivative of this function with respect to an individual observation:

$$\frac{\partial f(\underline{x})}{\partial x_j} = \frac{(RL - \bar{x})(x_j - \bar{x}) + (1/n)\sum(x_i - \bar{x})^2}{.5(n-1)(RL - \bar{x})^3}$$

Let $g(x_j)$ denote the numerator of this function.

It can be verified that:

$$g'(x_j) = (1/n)\sum_{i \neq j}(RL - x_i) > 0,$$

so $g(x_j)$ is increasing. Since the numerator of $f'(x_j)$ is increasing and the denominator is decreasing, $f'(x_j)$ must be increasing. Therefore, $f(x_j)$ is maximized either at zero or at DL_j.

Now consider $f(\underline{x})$ as a function of x_i and x_j, with detection limits DL_i and DL_j, respectively. Suppose that $f(\underline{x})$ is maximized when $x_i = 0$ and $x_j = DL_j$. Note that $f(0, DL_j) = f(DL_j, 0)$. Now, if $DL_i > DL_j$, then either:

(1) $f(DL_i, 0) > f(DL_j, 0)$

or (2) $f(0, 0) > f(DL_j, 0)$,

which violates the assumption that $f(0, DL_j)$ is a maximum. Therefore, $DL_i \le DL_j$.

Thus, the procedure to be followed to maximize $f(\underline{x})$ is to sort the detection limits in ascending order, $DL_{(1)} \le DL_{(2)} \le \ldots \le DL_{(n)}$, and then compute:

$f(DL_{(1)}, DL_{(2)}, \ldots, DL_{(n)})$,

$f(0, DL_{(2)}, \ldots, DL_{(n)})$,

:

:

:

and $f(0, 0,\ldots,DL_{(n)})$.

One of these n calculations will result in a maximum value of $f(\underline{x})$.

5. SOME RESULTS

Following [3], the tolerance limit equality in Section 3 can be reexpressed as:

$$\Pr\{T_{n-1} \le k \bullet n^{1/2} \mid \delta\} = (1-\alpha),$$

where T_{n-1} has a noncentral t distribution with (n-1) degrees of freedom and noncentrality parameter δ. The noncentrality parameter is given by:

$$\delta = n^{1/2}\Phi^{-1}(p^*), \text{ so } p^* = \Phi(\delta \bullet n^{-1/2}),$$

where Φ denotes the standard normal distribution function. Therefore, given k, n and the desired confidence level $(1-\alpha)$, one can search for δ and solve for p*, the lower bound on p_{RL}.

Our previous paper [5] provided estimates of p_{RL}, given by $p_{RL} = \Phi(k)$, and lower 95% ($\alpha = .05$) bounds for p_{RL} for various sample sizes and values of r = RL/DL. Tables 1 and 2 extend these results by considering multiple detection limits. Specifically, the tables presents the estimates for the cases where 20%, 50%, and 80% of the detection limits are DL and the remaining ones are .5DL. Also included are the binomial lower 95% limits on p_{RL} and the case where 100% of the detection limits are DL, which were presented in the previous paper.

Note that the procedure addressed in this paper provides point estimates of p_{RL} in each case, which the binomial approach does not (except for the uninformative 1.0). Likewise, because the 95% confidence bounds do use the information given by the detection limits, the procedure performs better than the binomial method, except for situations in which r is close to 1.0.

It appears that the procedure discussed in this paper should prove useful in many cases where a sample is encountered in which all observations are less than detection limits, This is particularly true for larger values of r and smaller values of F.

Sample Size	F	r = 1.0	r = 1.5	r = 2.0	r = 2.5	r = 3.0
10	.20	.932	.997	>.9999	>.9999	>.9999
	.50	.802	.971	.998	.9999	>.9999
	.80	.802	.951	.996	.9999	>.9999
	1.00	.624	.951	.996	.9999	>.9999
20	.20	.937	.998	>.9999	>.9999	>.9999
	.50	.809	.974	.998	>.9999	>.9999
	.80	.675	.954	.997	.9999	>.9999
	1.00	.588	.954	.997	.9999	>.9999
30	.20	.939	.998	>.9999	>.9999	>.9999
	.50	.810	.975	.998	>.9999	>.9999
	.80	.676	.956	.997	.9999	>.9999
	1.00	.572	.956	.997	.9999	>.9999

Table 1: Estimated Values of p_{RL} (r = RL/DL) when a Fraction, F, of the Observations in the Sample Have Detection Limit DL and the Remainder Have Detection Limit .5DL

Sample Size	F	r = 1.0	r = 1.5	r = 2.0	r = 2.5	r = 3.0	Binomial
10	.20	.756	.941	.990	.999	>.9999	.741
	.50	.586	.835	.946	.986	.975	.741
	.80	.455	.791	.934	.984	.997	.741
	1.00	.411	.791	.934	.984	.997	.741
20	.20	.834	.977	.998	>.9999	>.9999	.861
	.50	.666	.903	.980	.997	.9998	.861
	.80	.525	.863	.973	.996	.9997	.861
	1.00	.440	.863	.973	.996	.9997	.861
30	.20	.861	.986	.999	>.9999	>.9999	.905
	.50	.698	.925	.987	.9992	>.9999	.905
	.80	.555	.889	.982	.998	.9999	.905
	1.00	.451	.889	.982	.998	.9999	.905

Table 2: Conservative Lower 95% Bounds for p_{RL} (r = RL/DL)when a Fraction, F, of the Observations in the Sample Have Detection Limit DL and the Remainder Have Detection Limit .5DL

6. REFERENCES

[1] Haas, C.N. and Scheff, P.A., "Estimation of Averages in Truncated Samples," Environmental Science and Technology, 24, 912-919, 1990.

[2] Helsel, D.R. and Gilliom, R.J., "Estimation of Distributional Parameters for Censored Trace Level Water Quality Data: 2. Verification and Applications," Water Resources Research, 22, 147-155, 1986.

[3] Owen, D. B., "A Survey of Properties and Applications of the Noncentral t-Distribution," Technometrics, Vol. 10, pp. 445- 478, 1968.

[4] Smith, D.E. and Burns, K.C., Estimating a Percentile of a Contaminant Concentration Distribution When All Observations Are Less Than a Detection Limit, Desmatics, Inc. Technical Report No. 157-1, July 1994.

[5] Smith, D.E. and Burns, K.C., Drawing Inferences from Environmental Samples When All Observations Are Less Than a Detection Limit, Desmatics, Inc. Technical Report No. 157-3, March 1995.