Option #1: Final Technical Report

Grant Number: OH03073-02

Grant Title: Statistical Issues in Sampling for Compliance to Exposure Standards

Principal Investigator: Michael J. Symons
Department of Biostatistics

Co-Investigators: Michael R. Flynn
Department of Environmental Sciences

Stephen M. Rapport
Department of Environmental Sciences

Kinh Nhue (Young) Truong
Department of Biostatistics

School of Public Health
University of North Carolina
Chapel Hill, NC 27599-7400

January 1996

50272-101

REPORT DOCUMENTATION PAGE	1. REPORT NO.	2.	PB97-113682

4. Title and Subtitle Statistical Issues in Sampling for Compliance to Exposure Standards	5. Report Date 1996/01/00
	6.
7. Author(s) Symons, M. J., M. R. Flynn, S. M. Rapport, and K. N. Truong	8. Performing Organization Rept. No.
9. Performing Organization Name and Address Department of Biostatistics, School of Public Health, University of North Carolina, Chapel Hill, North Carolina	10. Project/Task/Work Unit No.
	11. Contract (C) or Grant(G) No. (C) (G) R18-OH-03073
12. Sponsoring Organization Name and Address	13. Type of Report & Period Covered
	14.

15. Supplementary Notes

16. Abstract (Limit: 200 words) Statistical approaches for analysis of data from the limited number of samples collected by an industrial hygienist for checking compliance to an occupational standard were considered. Sampling for compliance usually has been guided by judgment selection, rather than true randomness, resulting in the creation of compliance samples which approximate a censored sample from the upper tail of the exposure distribution. A graphical based idea was developed, which appears useful with small numbers of compliance samples which are typically collected, less than ten. Assuming lognormality of the workplace exposures, considered as common with time weighted averages over hours of collection, and assuming the industrial hygienist has the ability to select the greatest and the least exposures, one is able to sketch a lognormal probability paper based inference scheme. Simple linear regression analysis provides an estimation of the median, the variance among jobs, and a subtotal of variance components. These variance components could include worker differences, variability in occasions from day to day sampling, and measurement error. Incorporating historical data will be necessary due to the few samples which are available from a compliance survey.

17. Document Analysis a. Descriptors

b. Identifiers/Open-Ended Terms NIOSH-Publication, NIOSH-Grant, Grant-Number-R18-OH-03073, End-Date-09-14-1995, Grants-other, Workplace-studies, Industrial-hygiene, Sampling-methods, Statistical-analysis

c. COSATI Field/Group

18. Availability Statement	19. Security Class (This Report)	21. No. of Pages 36
	22. Security Class (This Page)	22. Price

(See ANSI-Z39.18) See Instructions on Reverse OPTIONAL FORM 272(4-77) (Formerly NTIS-35) Department of Commerce

Table of Contents

Page

LIST OF FIGURES

LIST OF TABLES

SIGNIFICANT FINDINGS

Related to the distributional shape of occupational exposure measurements, two findings are highlighted:

(1) Although lognormality appears applicable for time weighted averages over hours, mixtures of lognormals suggest themselves for short-term time weighted averages. These seem to correspond to production cycles which last longer than the collection time and the exposure generation level(s) reflect portions of the production cycle.

(2) Enormous sample sizes are required to explore distributional shape via skewness-kurtosis moment diagrams. Since mixtures have natural relevance to levels of exposure generation, even a rich family of unimodal distributions, like the Burr XII distribution, cannot approximate a bimodal mixture.

Related to the analysis of exposures measured for high risk employees, the original proposal of using extreme value theory was extended:

(3) By using a Type II left-censored likelihood model, all of the n compliance samples were incorporated, as well as the N-n unobserved samples. However, this likelihood requires an infallible selection of the n largest from the N possible jobs/sites at a workplace. Statistical inference requires a large number (n greater than 30) of samples.

(4) With fallible selection of the n largest from the N possible samples, a truncated from below likelihood is described, but its sample size requirement is even greater than with the left-censored model.

(5) A probability paper regression model is sketched and may have relevance for the small sample sizes that are usual with compliance sampling. However, lognormality is required and the percentile rank of each sample is needed. Use of the n/2 smallest and the n/2 largest exposure jobs/sites at a workplace provides the regression inputs that minimizes the variance of key parameter estimates.

Related to the serial correlation analysis premise for a sequence of short-term time weighted average samples, alternative modelling is recommended:

(6) Rather than modelling the covariance structure with a single mean as in a time series approach, a fixed effect to the overall mean is proposed, with different levels of exposure generation in the production cycle corresponding to the fixed effects.

USEFULNESS OF FINDINGS

The practice of collecting a few samples from a workplace for the purpose of checking compliance to an occupational exposure standard for a chemical agent yields the most statistical information about the parameters of the distribution of exposures when (1) lognormality of the exposures can be assumed; (2) the number of samples is split equally between the largest <u>and</u> least exposures; and (3) the selection of the greatest and least exposures is nearly infallible.

ABSTRACT

An industrial hygienist samples a workplace for compliance to a chemical standard by selecting those sites, workers, and/or times that are most likely to reveal the highest levels of exposure. Judgment selection, rather than randomness, guides sampling for compliance. Consequently, the n compliance samples approximate a censored sample from the upper tail of the exposure distribution. The fraction of these n samples which are actually among the upper n order statistics reflects a matching of the industrial hygienist's skill and experience with the workplace anatomy and the production process complexity. When the industrial hygienist is infallible at selecting the n largest of the N possible exposures, a Type II-left censored likelihood utilizes all the measured exposures and the size of N, the number of samples that could have been observed at the workplace. When the industrial hygienist is not perfect at selecting the n largest of the N possible exposures, a proposal is made to consider the compliance samples as a random sample from a distribution truncated from below. However, the truncation likelihood requires one additional parameter, the unknown truncation point, which hinders estimation of the exposure distribution parameters. Limiting forms of the truncation from below model reflects (1) random sampling, as the truncation point goes to zero, and (2) Type-II left censoring, as the truncation point approaches the $(N - n + 1)/N$th percentile of the exposure distribution. An assay of an industrial hygienist's proficiency for selecting the n largest exposures of the N possible should aid the statistical procedure by providing information on the unknown truncation point.

With either of these two likelihood based analyses and the compliance sampling of only a few jobs at the highest exposures, statistical inference is weak. Specifically, accuracy and precision are typically poor with small samples. The ability to document the components of variance within jobs, especially those due to occasion, workers, or laboratory error, is not possible; only a sub-total of the relevant components of variance is estimable. In particular, an independent estimate of at least one component of variance, usually variability across occasions/days, is needed with compliance samples.

Lognormality is a common, useful characterization of modeling workplace exposure distributions. Deviations from this strong right skewed shape may follow a mixture of two or

more exposure distributions reflecting shifts in mean exposures. Consequently, modeling mean structure rather than emphasizing covariance structure is recommended, whether samples are for compliance or from random selection. Random sampling of a workplace has its purpose of detailing the components of variance in occupational exposures, such as between occasions for a specific worker and laboratory analytical error. However, the size of these thorough surveys must be multiples of those typical of a compliance sampling. The number of compliance samples is restricted by the number of workplaces to be inspected, by the cost of collecting and processing each sample, and by the nature of the adverse health effect(s) likely from exposure above the chemical standard.

In conclusion, a graphical based idea is introduced that appears useful with the small number (n less than 10) of compliance samples typically collected. Presuming lognormality of the workplace exposures, as is common with time weighted averages over hours of collection, and the ability of the industrial hygienist to select the greatest <u>and</u> the least exposures, a lognormal probability paper based inference scheme is sketched. A simple linear regression analysis provides estimation of the median, variance among jobs, and a subtotal of variance components which could include worker differences, variability in occasions (day-to-day sampling) and measurement error. However, the degrees of freedom for estimating this subtotal of variance components is only $n - 2$. The incorporation of historical data, particularly to estimate variance components reliably, will probably be needed in view of the few samples available from a compliance survey.

The object of statistical inference is the probability of exceeding an exposure standard in a future sample or sampling of the workplace under study.

1. ELEMENTS OF OCCUPATIONAL EXPOSURE MEASUREMENTS

A useful orientation to occupational exposure measurements is a components of variance model, forcing a recognition of several sources of variation. Strictly, as per Samuels, et al (1985), a model allowing for possibly different mean levels of exposure by season within year or by exposure generation state of the production process is a mixed model. Consider the following model for workplace exposure measurements:

$$y_{ijklm} = \mu + \alpha_i + b_{ij} + c_{ijk} + d_{ijkl} + e_{ijklm}, \tag{1.1}$$

where

μ = overall mean exposure level for a year;

α_i = increment/decrement in mean level for the ith season (or for ith level of exposure generation);

b_{ij} = effect of sampling jobs within season i, having mean zero and variance σ_b^2;

c_{ijk} = effect of sampling days/workers within job j of season i, having mean zero and variance σ_c^2;

d_{ijkl} = effect of sampling time periods of day/worker k in job j of season i, having mean zero and variance σ_d^2;

e_{ijklm} = effect of measuring the exposure of the sample collected from time period ℓ day k in worker/job j of season i, having mean zero and variance σ_e^2;

and

y_{ijklm} = exposure measurement m of time period ℓ of day/worker k in job j of season i, having mean $\mu + \alpha_i$ and variance $\sigma_b^2 + \sigma_c^2 + \sigma_d^2 + \sigma_e^2$.

Notice that the last component of variance, σ_e^2, is of laboratory measurement and calibration error; the other components are due to sampling: periods within day/worker and days/workers within jobs. These latter components are usually substantial. Also for 8-hour samples the variance component within days is zero. More relevant to compliance sampling is the model:

$$y_{ijkl} = \mu + \alpha'_i + b_{ij} + c_{ijk} + e'_{ijkl} \tag{1.2}$$

where μ, b_{ij} and c_{ijk} are as above but

α'_i = increment/decrement in mean level for exposure generation state i, including any seasonal shifts;

e'_{ijkl} = combined effects of d_{ijkl} and e_{ijklm} in the above model, hence having mean zero and variance $\sigma_{e'}^2 = \sigma_d^2 + \sigma_e^2$;

and

y_{ijkl} = exposure measurements like those of compliance sampling for short term, e.g., 15 minute, exposure limits.

Consider the additional notes:

(a) With compliance sampling, strategically selected workers and times during the production process are chosen to observe the greatest exposures, so that the measurements have mean $\mu + \alpha'_{i\cdot}$ and variance $\sigma^2 = \sigma_b^2 + \sigma_c^2 + \sigma_d^2 + \sigma_e^2$. The increment $\alpha'_{i\cdot}$ is for an exposure generation peak and at that time (season?) when ventilation is poorest.

(b) With samples from one or two days rather than over weeks or a month, there is little ability to estimate adequately the variance component due to occasions, σ_c^2 , from the sample.

(c) The logarithm of the exposure measurements are approximately normal in their distribution. Any monotone transformation preserves the exceedance probability since the probability of an untransformed exposure exceeding the standard equals the probability of a logged exposure exceeding the logarithm of the standard.

The proposed object of inference is the probability of exceeding the standard in future exposure measurements.. With distribution function F(.) of the y_{ijkl} and random selection of m occasions in a year for a highest exposure job(s), the exceedance probability is the complement of the compliance probability, $P(E) = P(\overline{C}) = 1 - P(C)$

and

$$P(C) \geq [F(\{y_{ijkl} - \mu - \alpha'_{i\cdot}\}/\sigma)]^m . \tag{1.3}$$

Several remarks of clarification follow:

(i) This exceedance probability is for a particular high exposure job. The exceedance probability for any worker at the workplace must be larger, but how much larger depends upon the mean levels and their variances for the other generation states.

(ii) Inference from a compliance sampling for the model in equation (2) may yield estimates of $\mu + \alpha'_i$ and the subtotal of variance components $\sigma^2 = \sigma_b^2 + \sigma_d^2 + \sigma_e^2$. The true variance, $\sigma_b^2 + \sigma_c^2 + \sigma_d^2 + \sigma_e^2$, will be estimable only by an independent estimate of the day to day component, σ_c^2, being added to that estimate of subtotal variance from the compliance samples.

(iii) Regarding the distributional shape of exposure measurements, lognormality seems very applicable. However, a mixture of lognormal distributions appears a relevant alternative, especially within a workday which have components corresponding to exposure generation states. Unfortunately very large sample sizes (n at least 100) are needed either to assess the distributional shape or to provide an alternative nonparametric analysis.

2. INFERENCE FROM COMPLIANCE SAMPLES

In compliance sampling of job sites, workers, and/or times that are most likely to reveal the greatest exposure levels, an industrial hygienist uses judgment and workplace knowledge. In contrast, a randomly selected sample reflects the full range of workplace exposure levels, as shown in the top panel of Figure 1. When the industrial hygienist is infallible at selecting the n largest of the N available, the compliance sample exhibits Type II-left censoring, as portrayed in the middle panel of Figure 1. If the industrial hygienist is not perfect at selecting the n largest of N available, a truncated from below model is proposed and is shown in the bottom panel of Figure 1. This model requires one additional parameter, the lower limit above which all the observations lie and these observations are taken as a random sample from the truncation model.

2.a. Type II-Left Censored Likelihood

Inference with Type II-left censoring requires two features of the compliance sampling of a workplace. First, the industrial hygienist must be able to pick the n largest exposure sites. A combination of simplicity of the workplace process and an experienced industrial hygienist contributes to the n largest exposures being collected. Of course, the complexity of the workplace or an inexperienced industrial hygienist makes the truncation from below model more appropriate. Second, an accounting of N, the number of samples that are available to the industrial hygienist is needed and from the first feature, confidence that the N-n uncollected samples are below the smallest of those which were collected is utilized.

To illustrate the estimation details with the Type II-left censoring and truncation from below models, a Burr XII density was chosen. Specifically, the density and CDF are as follows:

$$f(z; c,k) = ckz^{c-1} \quad (1+z^c)^{-k-1}, \tag{2.1}$$

and

$$F(z; c,k) = 1 - (1+z^c)^{-k}, \tag{2.2}$$

for the positive random variable Z with mean ν_1 and standard deviation ν_2 determined by the positive shape parameters c and k. Specifically,

$$\nu_1 \;=\; k\Gamma(1 + c^{-1})\Gamma(k - c^{-1})/\Gamma(k + 1). \tag{2.3}$$

and

$$\nu_2^2 \;=\; k\Gamma(1 + 2c^{-1})\Gamma(k-2c^{-1})/\Gamma(k + 1) - \nu_1^2 \tag{2.4}$$

Choosing c = 5 and k = 5.75, an approximate normal shape is available for Z; see Burr 1942) and Johnson and Kotz (1970).

When the logged exposure Y is also approximately normal with mean μ and variance σ^2, the density and CDF for Y are derived using the linear relationship

$$Z = \frac{\nu_2}{\sigma}(Y - \mu) + \nu_1 \,. \tag{2.5}$$

This equation follows directly from equating the standardized Z and Y variates, namely $(Z - \nu_1)/\nu_2 = (Y - \mu)/\sigma$. The CDF for Y is the CDF for Z in (2.2) with the linear relationship (2.5) substituted for Z. Specifically,

$$F(y; \mu, \sigma, c, k) = 1 - [1 + (\frac{\nu_2}{\sigma}(y - \mu) + \nu_1)^c]^{-k} \tag{2.6}$$

and differentiating with respect to Y yields its density

$$f(y; \mu, \sigma, c, k) = \frac{ck\nu_2}{\sigma} (\frac{\nu_2}{\sigma}(y - \mu) + \nu_1)^{c-1} [1 + \frac{\nu_2}{\sigma} (y-\mu) + \nu_1)^c]^{-k-1}. \tag{2.7}$$

The natural logarithm of the Type II-left censored likelihood is

$$\sum_{i=1}^{n} \ln\{f(y_i; \mu, \sigma, c, k)\} + (N-n) \ln\{F(y_{(1)}; \mu, \sigma, c, k)\}, \tag{2.8}$$

where $y_{(1)}$ is the natural logarithm of the smallest of the n observed exposures. Maximum likelihood estimates for μ and σ^2 were obtained by an iterative routine using the initial estimates:

$$\mu_0 = [y_{(n/2)} + y_{(1)} - \sigma_o(z_{1-n/2N} + z_{1-n/N})] \tag{2.9}$$

and

$$\sigma_0 = [y_{(n/2)} - y_{(1)}] / [z_{1-n/2N} - z_{1-n/N}], \tag{2.10}$$

or the sample mean and sample standard deviation of the logged exposures. Each pair of initial estimates was used in attempting to achieve convergence. Appendix A contains a copy of this estimation routine.

The recognition of a compliance sample from an industrial hygienist's infallible selection of the n largest from among N possible exposures as a Type II - left censored sample allows the analysis to proceed utilizing readily available software. Following Ware and Demets (1976), note that the negative of the logarithm of the exposures converts this sample to a Type II - right censored sample, i.e., the usual survival analysis scenario. Lawless (1982, pp. 220-237) sketches the estimation theory for the lognormal-normal model and SAS's PROC LIFEREG provides a reliable computer algorithm for fitting a normal model to the logarithm of right censored data.

2.b. Truncation from Below Likelihood

The natural logarithm of the truncated from below likelihood is

$$\sum_{i=1}^{n} \ln\{f(y_i; \mu, \sigma, c, k)\} - n \ln\{1 - F(\gamma; \mu, \sigma, c, k)\}. \tag{2.11}$$

By inspection, $\hat{\gamma} = y_{(1)}$ maximizes the right-most term for all values of μ and σ^2. Given $\hat{\gamma} = y_{(1)}$, maximum likelihood estimates for μ and σ^2 were obtained by an iterative routine using the initial estimates from (2.9) and (2.10) or sample mean and sample standard deviation. Appendix B contains a copy of the estimation routine.

Estimation for a truncated likelihood is difficult, as is documented by Schneider (1986, Chapter 3), especially when the truncation point is unknown. He provides a sketch of maximum likelihood estimation. Practical solutions require the unknown mean to be within the range of the observations, and when the truncation point is unknown, very large ($n > 100$) samples are required. When the percentage of observations truncated is known, as may be relevant for compliance sampling, the mean and standard deviation are linearly related. This observation forms the basis of the probability paper regression strategy, sketched in section 2.d. Nevertheless, with less than 10 compliance samples available, reliability of the estimates is limited. The EM algorithm, restricted maximum likelihood estimation and moment estimators are also presented, but maximum likelihood estimation is stressed.

2.c. Comparison of Left Censored and Truncation Below Estimation

The estimation routines for the Type II-left censoring and single truncation from below models were examined using simulated data. For $N=200$ available logged exposures from a normal distribution with mean 3.1 and variance $(0.25)^2$, five repetitions using each of the above described routines were performed at six conditions. These conditions correspond to two choices for sample size: $n = 25$ and 50, and three levels of industrial hygiene success at selecting the n largest of N available exposures: (i) the largest n were selected; (ii) n randomly selected from the largest $[3n/2] + 1$; and (iii) n randomly selected from the largest $2n$. The latter two selections reflect increasing fallibility in the n samples being <u>the</u> n largest.

The results from these simulated data suggest a strong dependence on utilizing the correct likelihood and having a large sample (n much larger than 30 or 50) to obtain accurate and reliable estimation. Three points relevant to inference from compliance samples appear as follows:

(1) Large samples ($n > 30$) are needed, even with infallible industrial hygienist's ability to select the n largest of N exposures. See Table 1 and the row section for "Left Censored (50)" to see that model parameter estimates are beginning to reflect the population values.

(2) Accurate assaying of the industrial hygienist's ability to select the n largest of N exposures is needed. See Tables 2 and 3 for two levels of "contaminated" Left Censoring data generation and the degradation of the left censored likelihood estimation. An underestimation of the mean and an overestimation of the variance is visible from these simulations.

(3) Very large samples ($n > 50$) are needed with fallible industrial hygienist's ability to select the n largest of N exposures. See Table 2 and 3 for the weak performance of the truncated below likelihood estimation even with $n = 50$ at either of the two levels of "contaminated" left censoring data. (Based upon the needs of this truncation below likelihood in Schneider (1986) and a hint of better performance with greater "contamination" in Table 3 than Table 2, large samples in the upper half of the exposure distribution may be a reasonable strategy with fallible industrial hygienist's selection of the n largest of the N exposures.)

2.d. Probability Paper Regression

Since the sample size requirement of likelihood estimation appropriate for compliance sampling greatly exceeds the number of compliance samples usually available (n less than 10), a "probability paper" regression strategy suggested itself out of the "linear dependence" of the mean and standard deviation with the truncated likelihood based estimation for the case of known percentile of the unknown truncation limit. Two crucial inputs are required:

(1) an assumption of distributional form for the observations. Lognormality is commonly presumed and has relevance for workplace exposure measurements.

(2) an indication of the percentile-rank of each observation. The industrial hygienist's ability to select the n largest among N possible samples characterizes this needed input.

The data available for analysis are exposure measurements at a few (n less than 10) known percentiles of the exposure distribution, that is,

$$(P_i, X_i) \equiv (Z_i, Y_i), \tag{2.12}$$

for i = 1, 2, ..., n selected jobs, and where $Y_i = \ln(X_i)$ denotes the presumed normally distributed transformed exposures and $Z_i = \Phi^{-1}(P_i)$. Note that with normal probability paper P_i is plotted at the corresponding percentile (Z_i) of the standard normal to estimate the straight line relationship with the transformed measurement (Y_i). See Figure 2 for a schematic of the probability paper regression layout and Kotz, et al (1982). In this application the $P_i \equiv \Phi(Z_i)$ is specified by the industrial hygienist's knowledge of the n selected exposures among the distribution of N possible samples (jobs) rather than as determined when all N samples are available in usual probability paper plotting of the empirical distribution.

The regression model is

$$Y_i = \mu + \sigma Z_i + \varepsilon_i \tag{2.13}$$

for i = 1, 2, ..., n selected jobs, where μ is the mean (median) at Z = 0 (0.50 = $\Phi(0)$) and σ is the standard deviation for the variation between jobs. Since the expected value of the ε_i being zero depends upon accurate specification of the P_i and appropriateness of the normality so that $Z_i = \Phi^{-1}(P_i)$, the variance of the errors, ε_i, being σ_e^2 reflects the subtotal of the remaining components of variance: between workers within jobs, between days, and measurement error for collecting and converting the workplace physical samples into a chemical exposure concentration. If the sampling of jobs does not reflect different workers or different days, then σ_e^2 reflects a subtotal of relevant variance components less those components not reflected in the workplace sampling. A schematic of requisite data for the envisioned probability paper regression is provided in Table 4.

Least squares estimates for the simple linear regression model in (2.13) are the standard ones:

$$\hat{\mu}(=\hat{\beta}_0) = \overline{Y} - \hat{\sigma}\,\overline{Z} \tag{2.14}$$

and

$$\hat{\sigma}(=\hat{\beta}_1) = \sum_{i=1}^{n}(Y_i - \overline{Y})(Z_i - \overline{Z}) \Big/ \sum_{i=1}^{n}(Z_i - \overline{Z})^2, \tag{2.15}$$

having variances

$$\upsilon(\hat{\mu}) = \hat{\sigma}_e^{\,2}\left(\frac{1}{n} + \frac{(\overline{Z})^2}{\Sigma(Z_i - \overline{Z})^2}\right) \tag{2.16}$$

and

$$\upsilon(\hat{\sigma}) = \hat{\sigma}_e^{\,2} / \sum_{i=1}^{n}(Z_i - \overline{Z})^2, \tag{2.17}$$

where

$$\hat{\sigma}_e^{\,2} = \sum_{1}^{n}\left[Y_i - (\hat{\mu} + \hat{\sigma}Z_i)\right]^2 / n - 2; \tag{2.18}$$

see for example Draper and Smith (1966). Notice that only n – 2 degrees of freedom are available to estimate the subtotal of variance components $\sigma_e^{\,2}$ by (2.18) and that a constant variance at each Z_i requires the same components of variance be reflected in the Y_i observed at each Z_i. With different variance components present for some Y_i, a weighted least square estimation should be considered.

Upon inspection of the variance for the standard deviation between jobs in (2.17), a large spread in the selected percentiles of the Y_i is desirable to reduce $v(\hat{\sigma})$. With lognormality assuring the straight line in the probability plot of Figure 2, (2.17) is minimized by taking the n/2 largest exposures <u>and</u> the n/2 smallest exposures among the N possible in the workplace. This is a modification from taking all n of the largest exposures and should greatly improve the information about the parameters of the distribution of exposures for the workplace. This symmetric selection of the Z_i about zero also minimizes the $v(\hat{\mu})$ in (2.16). Randomly selected jobs would yield measurements at equally spaced percentiles on average and would be less informative than the n/2 greatest and n/2 least exposures.

Further discussion of the probability paper regression is provided in the conclusion and future research sections.

3. ILLUSTRATIVE EXCEEDANCE PROBABILITY CALCULATIONS

Exceedance probability calculations are presented in this section. Typically these cannot be completed from the information available from a single compliance sample. Consequently, the required model parameters must be provided by analysis of the compliance sample, including descriptive features of the workplace, and from other sources, such as the literature or previous sampling(s) of the workplace. Several workplace exceedance probability calculations are illustrated given the exceedance probabilities for an exposure generation site or sites. Lognormality is used for illustration. In the absence of a parametric model, nonparametric estimates could be used, but they require much larger sample sizes. The exceedance probabilities for an exposure generation site or sites require parameter estimation from a compliance sample as sketched in the previous section.

Scenario: Suppose n = 6 of the N=100 workplace employees have a log-exposure distribution 80% time at N $(3.1, (0.25)^2)$ and 20% time at N $(2.4, (0.52)^2)$. The remainder of the 94 workers are exposed 100% time at N $(1.2, (0.52)^2)$. With a short-term exposure limit (STEL) for 15 minutes samples of 4.0, several exceedance probability calculations are possible.

<u>Example 3.1</u>: k independent samples at one site

The exceedance probability is the complement of the compliance probability:

$$P\,(\text{exceedance} \mid \mu_i, \sigma_i, k) = 1 - \Phi^k\left(\frac{4.0-\mu_i}{\sigma_i}\right),$$

where $\Phi(.)$ is the CDF for the standard normal. For example, k = 3 independent samples from the high risk sites ($\mu_1 = 3.1$, $\sigma_1 = 0.25$) yields

$$1-\Phi^3\left(\frac{4.0-3.1}{0.25}\right) = 1-(0.9998)^3 \quad = \quad 0.005999.$$

<u>Example 3.2:</u> one random sample from the workplace

The exceedance probability for the workplace is the weighted sum of those exceedance probabilities at each site (by the addition law of probability and multiplication law at each site):

$$= \left(\frac{6}{100}\right)(0.8)\left[1-\Phi\left(\frac{4.0-3.1}{0.15}\right)\right] + \left(\frac{6}{100}\right)(0.2)\left(1-\Phi\left(\frac{4.0-2.4}{0.52}\right)\right)$$

$$+ \frac{94}{100}(1.00)\left(1-\Phi\left(\frac{4.0-1.2}{0.52}\right)\right)$$

= 0.048 (1-0.9998) + 0.012 (1-0.9990) + 0.94 (1-0.9999+)

= 0.0000096 + 0.000012 + 0.0000 = 0.00002

Example 3.3: One sampling of the workplace - 3 independent samples at level 1, 2 at level 2, and 1 at level 3:

$= (1-(0.9998)^3) + (1-0.9990)^2) + (1-0.9999+)$

= 0.0005999 + 0.001999 + 0.00000 = 0.008

Example 3.4: Two independent samplings of the workplace

Again, the exceedance probability is the complement of compliance:

For workplace sampling 3.2,

P (exceedance) = $1 - (0.99998)^2$ = 0.00004.

For workplace sampling 3.3,

P (exceedance) = $1 - (0.992)^2$ = 0.016.

Several comments are noted:

(1) Compliance sampling provides information on the site with log exposures $N(3.1, (0.25)^2)$, if the industrial hygienist samples only at those sites.

(2) The total variance may be underestimated if selections over days are not in the compliance sampling. Use of historic estimates for one or two components of variance should be considered rather than assuming zero values; see Buringh and Lanting (1991).

(3) The exceedance probability for a sample at site 2: $N(2.4, (0.52)^2)$ is greater than at site 1: $N(3.1, (0.25)^2)$ due to the larger variance. Careful estimation of variances is important.

(4) When variances are estimated and not known, the exceedance probabilities should be calculated from Student's t distribution rather than the standard normal; see Symons, Chen and Flynn (1993).

4. DISTRIBUTIONAL SHAPE AND CORRELATION OF MEASUREMENTS

Full-shift samples taken across days show weak correlation, as shown by Frances, et al (1989) and Kumagai, et al (1993). Departures from lognormality should be less due to the longer averaging times. Consequently, attention is directed at consecutive 15-minute time weighted measurements within a day.

Figure 3 from George (1993) contains illustrative data. These are 15-minute time weighted averages at an autobody wipedown using an alcohol and water mixture as preparation for spray painting. A clear cycle is visible in Figure 3, Panel A, and attributable to higher exposures corresponding to actual "wiping down" and lower exposures at in-between times.

4a. Modelling Mean or Covariance Structure and Mixture Alternative

Two modelling strategies are conceptually compared. Frances, et al (1989), Roach (1977), Spear, et al (1986) and Rappaport and Spear (1988) documented the tendency of such serial air concentrations to be autocorrelated using analysis based upon a stationary time series model. Since the mean is taken to be constant over the time period, only the covariances of the observations are available to explain the obvious variability in the sequence. Alternatively, variation in the means is proposed, one level for the exposure generation phase of the process and another (lower) level for the non-generation phase. Within these periods, independence of the observations can be considered; i.e., are they random deviations about their respective mean levels? Visual support for the two levels of exposure is readily apparent in Figure 3, Panel A, readily accommodating the exposure generation process.

The focus on mean structure rather than covariance modelling can be presented in two ways. If the component source of each observation, namely, exposure generation and non-generation, is not available, then a mixture model is proposed; see Figure 3: Panel B. Specifically,

$$f(x; \phi) = \pi\, N(x; \mu_1, \sigma_1^2) + (1 - \pi)\, N(x; \mu_2, \sigma_2^2), \qquad (4.1)$$

where the fraction π of the exposures comes from the component with the larger mean $\mu_1 > \mu_2$, say), the remainder are from the non-generation phase, and perhaps $\sigma_1^2 = \sigma_2^2$. The interpretation of π is the portion of the time that exposure generation is operative.

But by observing the process, it is clear whether or not exposure generation is the operative phase of the industrial process. With supplemental notation of which phase the process is in, then the distribution for exposure generation observations is

$$N(\mu_1, \sigma_1^2) \tag{4.2}$$

and otherwise it is normal with mean μ_2 and variance σ_2^2, and parameter estimation is straight forward. The fraction of exposure generation π can be estimated by observing the process.

Nicas and Spear (1993) describe a task-based model, of which the above is a special case. Their use was for the analysis of full or partial shift time weighed averages. They also permitted a serial correlation of observations within task, but in application they considered statistical independence as useful. The wipe-down data come from different cars and with different wipe-down rags that are necessary in preparing a surface for painting. The emphasis here is for different mean levels; statistical independence is thereby better approximated than with a single mean.

4b. Sampling Properties of Moment-Ratio Diagram

Simulations of lognormally distributed samples were used to study the skewness-kurtosis plots on a moment ratio diagram; see Kotz, et al (1982). A range of geometric standard deviations were utilized to cover the environmental variability for workplace exposures. The sampling variability of the skewness-kurtosis pair was generally not reflective of the population values, especially as skewness increases. The results were improved somewhat by considering transformations of skewness and kurtosis; the best results were obtained for the natural logarithm of the positive square root of skewness with the negative of natural logarithm of kurtosis. See Figure 4 and note the underestimation with the large geometric standard deviation.

A very large sample size (n > 1000) is needed before results as per Figure 4 at the smaller geometric standard deviations are reasonable. In view of these enormous sample size requirements and that deviation from lognormality is more likely a mixture of lognormals, and therefore easily outside the shape range of the unimodal Burr XII distribution, no further work was expended in the fitting of Burr XII distributions. The reasons for the very large sample requirements of the skewness-kurtosis plots are likely due to two aspects:

- First, the variance of the r^{th} power of a positive random variable involves the expectation of the $2r^{th}$ power of that random variable. Empirically, sample size requirements seemed to increase approximately by a factor of 10 with each increment in power: where 10's suffice for inference on means; 100's are needed for variances, 1000's for skewness and 10,000's for kurtosis, at least roughly.
- Second, skewness (kurtosis) is a ratio of the 3^{rd} (4^{th}) central moment over the three halves (four halves) power of the standard deviation. Taking logarithms changes the ratio to a difference of logarithms, a generally helpful manipulation.

5. DISCUSSION AND CONCLUSIONS

In compliance sampling an industrial hygienist visits a workplace to assess the chances of chemical exposures exceeding an occupational standard. Study of the production process provides a basis for enumerating the N possible sites/jobs at the workplace that could be sampled. Using knowledge of the flow of materials and products in the steps of manufacturing and occasionally sampling results from an earlier visit to this workplace or surveys at similar manufacturing plants, a small number of samples n are collected at the sites/jobs expecting to yield the greatest exposures.

Without any distributional assumptions, such as lognormality, of the exposure measurements (X), a $1 - \alpha$ confidence limit can be calculated for the binomial proportion p, the probability that a randomly selected job/site exposure measurement exceeds the occupational standard, S. Specifically,

$$p = P[X > S] \equiv P[Y = \ln X > \ln S]. \tag{5.1}$$

When none of the measured exposures exceed the standard and the industrial hygienist is infallible at selecting the greatest exposures, a $1 - \alpha$ lower confidence limit for p is based upon the likelihood of the data as follows:

$$(1 - p_l)^N \leq \alpha \tag{5.2}$$

or

$$p_l \geq 1 - \exp(\ln \alpha / N). \tag{5.3}$$

Notice that even $n = 1$, a single observed exposure measurement below the standard, but assumed to be the maximum among the N possible exposures that could have been selected, is sufficient for calculating the lower limit p_l. For levels of significance $\alpha = 0.01$ and 0.05 and workplace sizes $N = 50$, 100 and 250, lower confidence limits on the probability of exceeding the occupational standard in a single future random sample from the workplace are:

$\alpha =$		0.05			0.01	
$N =$	50	100	250	50	100	250
$p_l =$	0.0582	0.0295	0.0119	0.0880	0.0450	0.0183

Even with large workplaces, the exceedance probabilities without any statistical distributional assumptions are above 1%. Depending upon the magnitudes of the exposure measurements relative to the standard and utilizing lognormality of these measurements, much smaller estimates are possible as illustrated in section 3.

Although the lognormality can be supported for many situations as discussed in section 4, the ability of the industrial hygienist to infallibly select the greatest exposures is required. Further research is needed to explore empirically this premise. Cato (1995) examined the correlation of industrial hygienists' prior ranking of selected samples with their actual measured values. Analysis of fifty-one surveys from 19 industrial hygienists from the North Carolina OSHA bureaus of compliance and consultative services and the Department of Environment, Health, and Natural Resources industrial hygiene consultant program found 49 percent of the correlation coefficients were 0.80 or larger. The ability to recognize differences in exposures seemed well supported, but directly unexamined was the related question of the ability of industrial hygienists to select the greatest exposures among all of these at a workplace. Several factors appear important to each of these abilities, namely, the magnitude of the difference in measured concentrations, hygienists' certification status, hygienists' overall experience, and assessment complexity.

The practice of collecting a few samples (n less than 10) for purposes of checking compliance to an occupational standard dramatically limits the statistical approaches for analysis of such data. This report has highlighted the upper tail sampling nature of these data and detailed two likelihoods for their analysis: Type II-left censored and truncation from below. Unfortunately, performance of these procedures requires large samples (n greater than 30) before reliable estimation can be expected. And, the ability of the industrial hygienist to select from among the greatest, if not the greatest, exposures is required for these statistical analysis strategies.

Sketched, but not illustrated, is a proposal to capitalize on the general relevance of exposures being lognormally distributed and to utilize the ability of an industrial hygienist to select the largest exposures at a workplace. A regression of logged exposures on their specified percentile rank in the distribution of all exposures at a workplace may provide considerable information about the parameters of the exposure distribution. The concept is expressed in a probability paper plotting framework and inspires a modification of the compliance sampling strategy of selecting the n largest exposures. By selecting only the n/2 largest exposures and applying the reverse of this skill to select the n/2 smallest exposures, maximum information on the variance between jobs is obtained from n samples. Crucial but untested is the empirical ability of industrial hygienists to reliably sample the extremes of the distribution of exposures at a workplace.

Two issues of correlation of occupational exposures can be identified. First is the serial correlation of exposures collected at adjacent points in time, especially within a day/shift. Correlation between days are of negligible concern by comparison. Second is the correlation induced by the nested collection of samples in a random effects modelling of occupational exposures. Independence then of observations can be enhanced by sampling at spaced out times within a day. Sampling across days which are spaced out requires repeat visits to a workplace and pose expensive logistical problems at least for compliance sampling. Further, denoting different jobs as levels of the first random component, and sampling across jobs, independence of the exposure measurements is reasonable. However, the components due to days, workers in jobs, and the measurement error from collecting and processing a physical sample become a subtotal of variance components which are estimable only as a subtotal. The choice of jobs as the primary component of variance is somewhat arbitrary, and with only a one day visit to the plant, variability due to occasions/days will not be reflected in the residual subtotal of variance components. Historical data could be employed to adjust upward the estimated subtotal of variance components and reflect a nonzero component due to occasions/days.

6. FUTURE RESEARCH

Some research ideas that were formulated but not fully examined under the funding for this grant are itemized in this section.

6.1 Probability Paper Regression

The proposal is responsive to the small number of samples typically collected for compliance purposes. However, lognormality of the exposures, independence of the observations, and the required percentile rank of each observation are required for this simple linear regression analysis. Trial application of the procedure is the next step in examination of its utility.

6.2 Comparative Modelling: Mean versus Covariance Structure

In view of Figure 3, Panel A, comparatively modelling these cyclical data with a traditional time series approach and a two component mean model would be informative. The difficulty with the time series approach is in the lack of regularity in the cycles related to production realities. The exposure-generation and nonexposure-generation phases of the cycle seem appropriate, but there may still be a need for correlation between adjacent measurements within the same phase of the cycle.

6.3 Likelihood Augmentation for Single Truncation from Below

Corresponding to the log-likelihood (2.11) for the truncated from below sampling model, consider supplemental information on the percentile rank of the truncation point γ. It is a characteristic of the industrial hygienist's skill and experience applied to the workplace complexity. As a finite population of N possible exposures X, let π_γ be the fraction of exposures actually exceeding γ , when the industrial hygienist is attempting to select the largest n of N at this workplace. Then $\pi_\gamma = 1 - F(\gamma) = S(\gamma)$, where F(.) is the CDF of exposures X, and $\pi_\gamma \leq n/N$. The closer π_γ is to n/N, perfect selection of the n largest of N possible samples, the more skilled the industrial hygienist and/or the simpler the workplace.

Suppose a binomial survey of M workplace exposures yields m which exceed γ, the characterization of the industrial hygienist's compliance sampling at the workplace being examined. The likelihood of these supplemental data are proportional to

$$p^m(1-p)^{M-m}, \qquad (6.3.1)$$

where $p = S(\gamma; \mu, \sigma^2)$. With independence of these data from the n compliance samples, the supplemented truncation from below likelihood is proportional to

$$\left\{\left[\prod_{i=1}^{n} f(x_i; \mu, \sigma^2)\right] p^{-n}\right\}\left\{p^m (1-p)^{M-m}\right\}. \tag{6.3.2}$$

Two limits of this likelihood support the exploration of the truncation from below model. First, if $m = n = M$, likelihood (6.3.2) reduces to that for random sampling of the workplace. Indeed if all of the exposures exceed γ, then $p = S(\gamma; \mu, \sigma^2)$ being certain is suggested. Second, if $m = N = M$, likelihood (6.3.2) reduces to the Type II-left censored likelihood. Such a survey of all possible exposures suggests strong evidence of the industrial hygienist's ability to always select exposures among the n largest of N available. Although supplemental information summarized by the binomial likelihood (6.3.1) could "steer" the estimation toward the industrial hygienist's γ by knowing better $\pi_\gamma = S((\gamma; \mu, \sigma^2)$, the availability of such information must be explored.

SELECTED REFERENCES

Buringh, E. and R. Lanting (1991). Exposure Variability in the Workplace: Its Implications for Assessment of Compliance. American Industrial Hygiene Association Journal 52:6-13.

Burr, I.W. (1942). Cumulative Frequency Functions. Annals of Mathematical Statistics, 13: 215-232.

Cato, S. (1995). An examination of the ability of industrial hygienists to distinguish differences in worker exposures. MS Technical Report. Department of Environmental Sciences and Engineering, University of North Carolina at Chapel Hill.

Draper, N.R. and H. Smith (1966). Applied Regression Analysis. Wiley, New York.

Francis, M., S. Selvin, R. Spear, and S. Rappaport (1989). The effect of autocorrelation on the estimation of workers' daily exposures. American Industrial Hygiene Association Journal, 50 (1):37-43.

George, D.K. (1993). An Examination of Applied Issues in Occupational Exposure Assessment. Ph.D. Dissertation. Department of Environmental Sciences and Engineering, University of North Carolina at Chapel Hill.

Johnson, N.L. and S. Kotz (1970). Continuous Univariate Distributions - 1. Houghton Mifflin, Boston, pp 30-33.

Kotz,, S., N.L. Johnson and C.B. Read (1982). Encyclopedia of Statistical Sciences. Wiley-Intersciences, New York (Volume 1: Burr Distributions, p. 335-340;) Volume 2: Probabilty Plotting, p. 232-237).

Kumagai, S., I. Matsunaga and Y. Kusaka (1993. Autocorrelation of short-term and daily average exposure levels in workplaces. American Industrial Hygiene Association Journal, 54:341-350.

Lawless, J.F. (1982). Statistical Models and Methods for Lifetime Data. Wiley, New York.

Nicas, M. and R.C. Spear (1993). A task-based statistical model of a worker's exposure distribution: Part I: Description of the model. American Industrial Hygiene Association Journal, 54:211-220.

Rappaport, S.M. and R.C. Spear (1988). Physiological damping of exposure variability during brief periods. Annals of Occupational Hygiene, 32:21-33.

Roach, S.A. (1977). A most rational basis for air sampling programs. Annals of Occupational Hygiene, 20:65-84.

Samuels, S.J., G.K. LeMasters and A. Carson (1985). Statistical Methods for Describing Occupational Exposure Measurements. American Industrial Hygiene Association Journal, 46: 427-433.

SAS (1988). SAS/STAT USER'S Guide. Release 6.03 Edition, Chapter 21 (LIFEREG).

Schneider, H. (1986). Truncated and Censored Samples from Normal Populations. Marcel Dekker, New York.

Spear, R.C., S. Selvin and M. Francis (1986). The influence of averaging times on the distribution of exposures. American Industrial Hygiene Association Journal, 47:365-368.

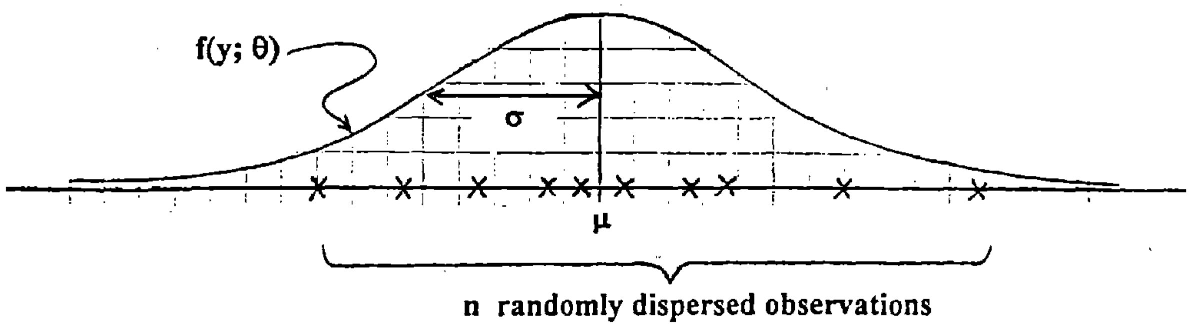
Symons, M.J., C.C. Chen and M.R. Flynn (1993). Bayesian Nonparametrics for Compliance to Exposure Standards. JASA, 88:1237-1241.

Ware, J.H. and D.L. Demets (1976). Reanalysis of some baboon descent data. Biometrics, 32:459-463.

Figure 1: Schematics for Three Samplings of an Exposure Distribution

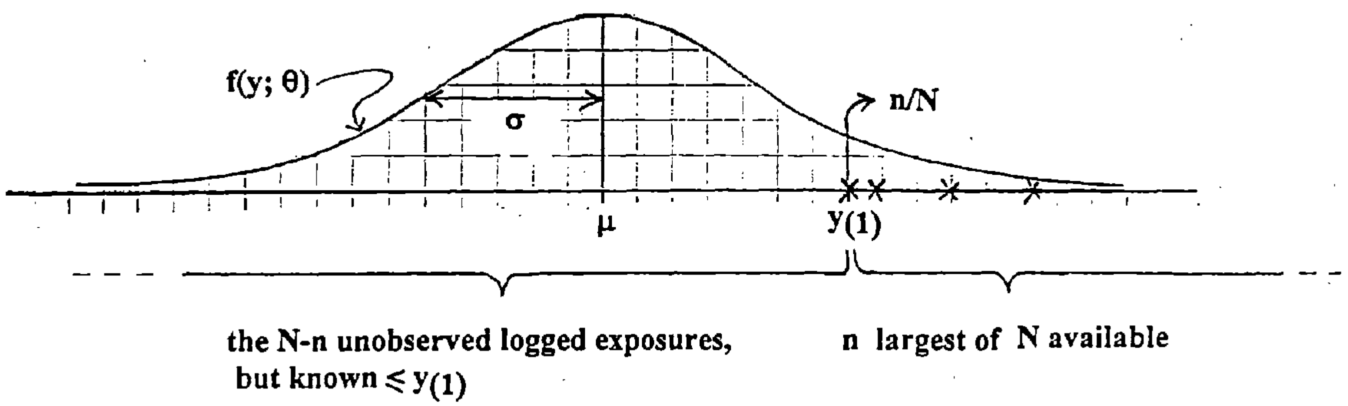
Random Sampling:

$$L = \prod_{i=1}^{n} f(y_i;\theta)$$



Type II Left Censoring:

$$L = \prod_{i=1}^{n} f(y_i;\theta) \prod_{i=n+1}^{N} F(y_{(1)};\theta)$$



Single Truncation from Below:

$$L = \prod_{i=1}^{n} \frac{f(y_i;\theta)}{1 - F(\gamma;\theta)}$$

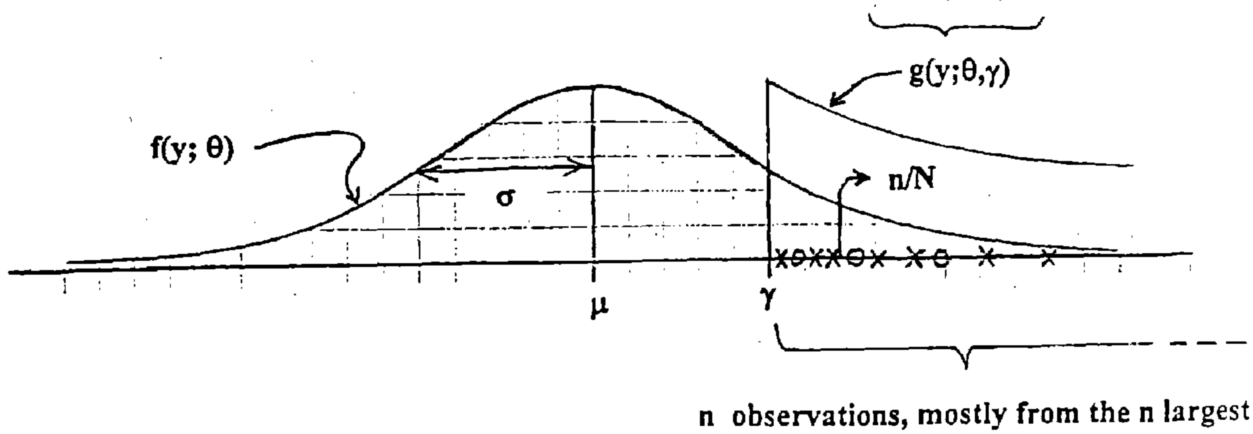
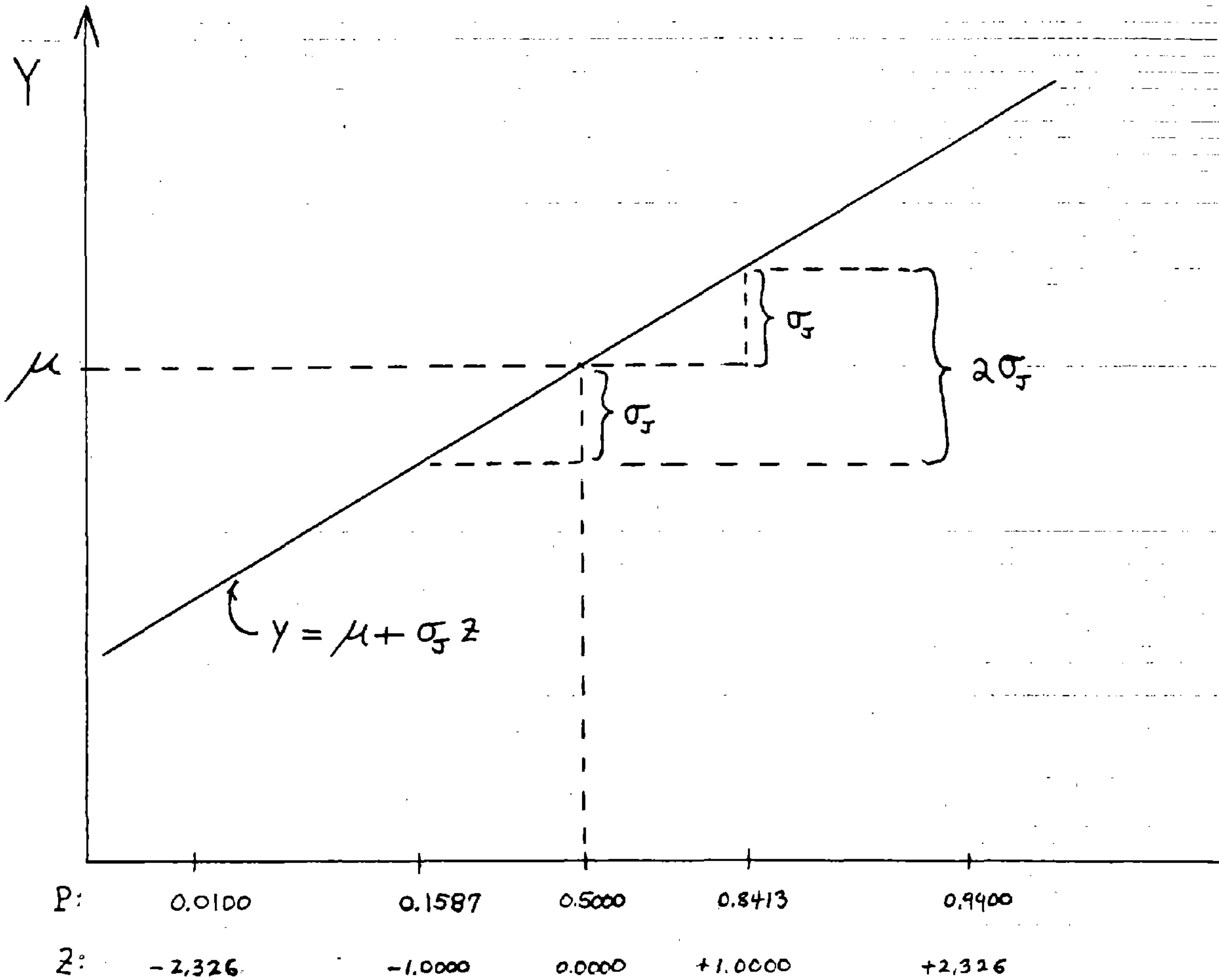


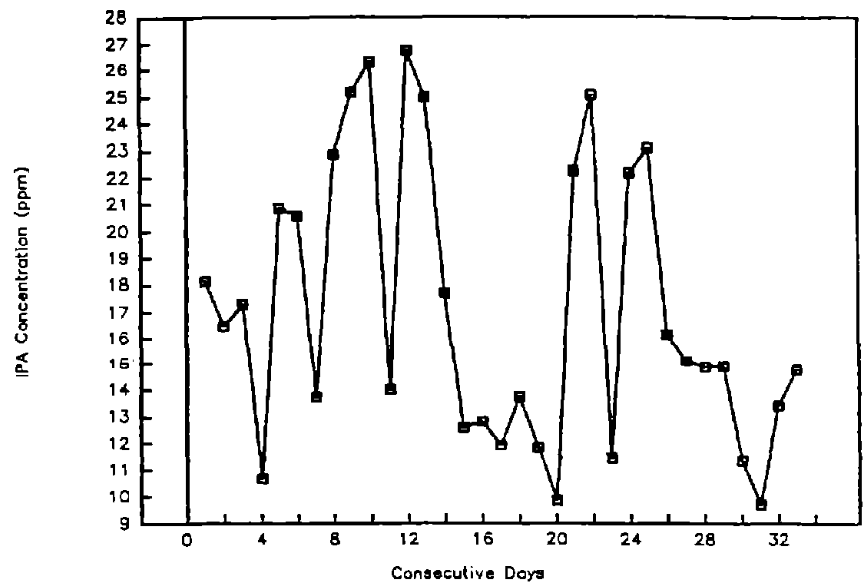
Figure 2. Schematic of Probability Paper Regression*



* The response (Y) is the natural logarithm of measured exposure plotted at the percentile rank (P) of its job title. Assuming lognormality of the distribution of exposures across job titles, the scale of percentile rank (P) is standard normal, viz. $Z = \Phi^{-1}(P)$.

Figure 3. Illustration of Data and Its Distribution from Exposure - Generation and Nongeneration Periods in Production Cycle.

Panel A. Figure 3.8 from George (1993, p. 112): Exposure Over Longest Consecutive Period - Worker 2 (Weekends and Down Time Omitted).



Panel B. Annotated Figure 4.18 from George (1993, p. 173): Log Probability Plot of 15 Minute Average Concentrations Overlayed with Calculated and Predicted Distribution.

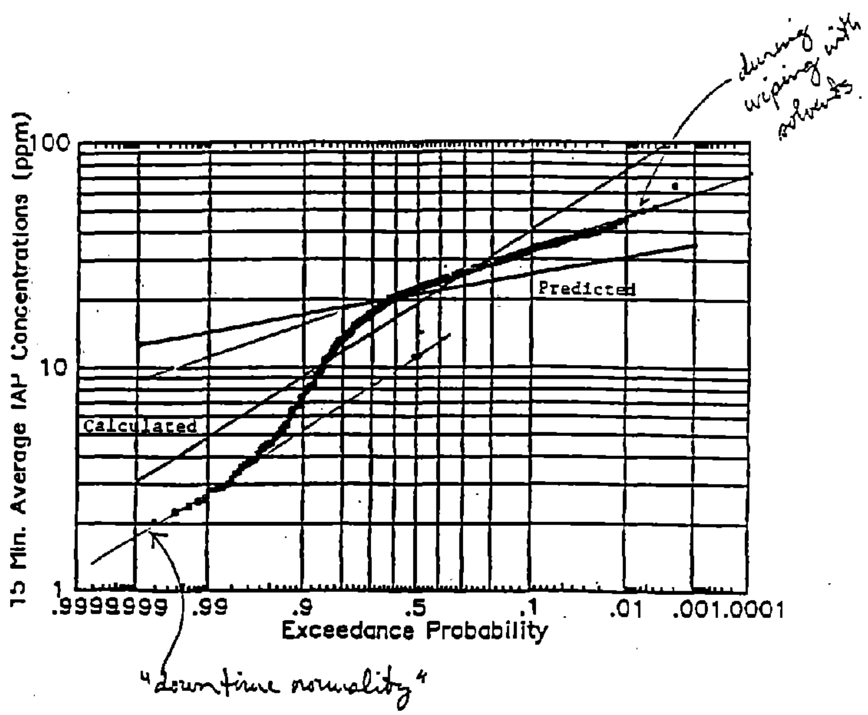
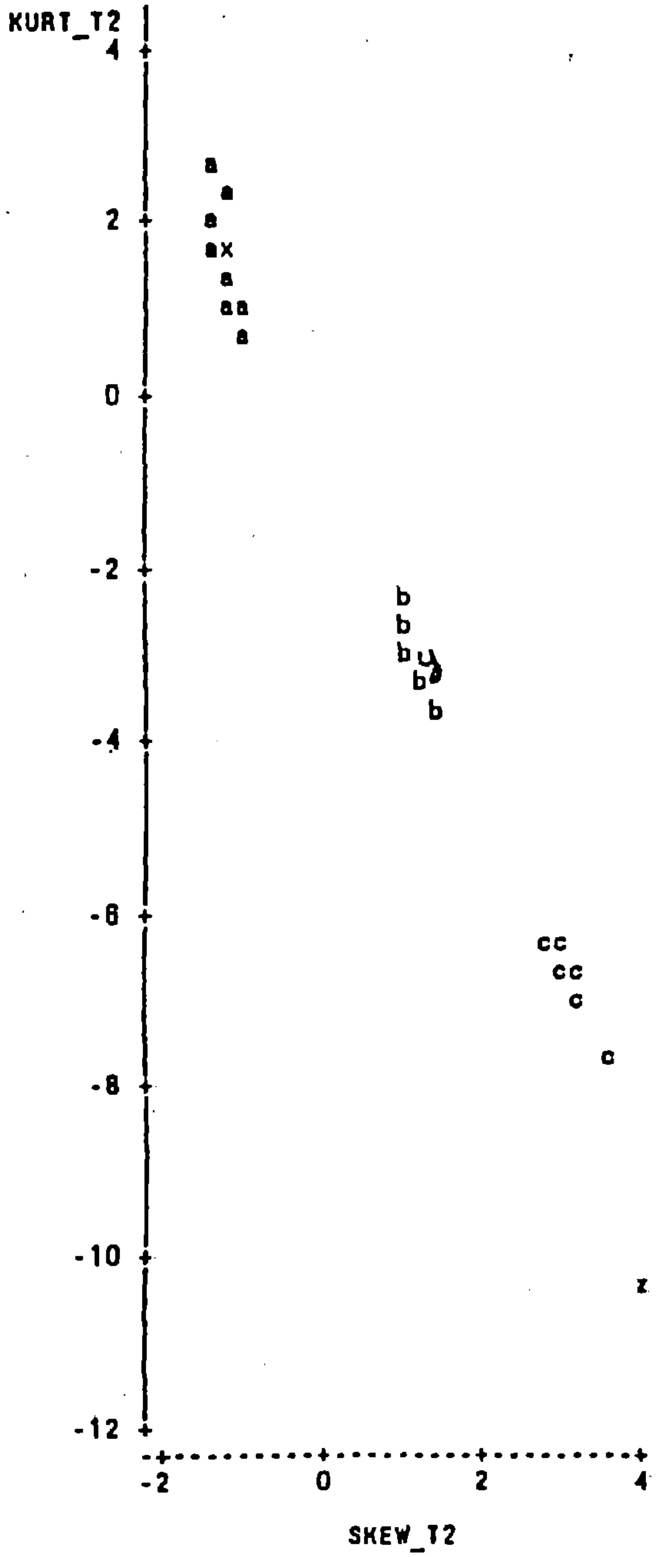


Figure 4. Sampling Results on Transformed Axes of Moment-Ratio Diagram: Three Lognormal Skewness Values and N = 10,000.



$SKEW\_T2 = \ln |b_1|$

$KURT\_T2 = -\ln b_2$

a = sample values with population sigma at 0.1 plotted at x.

b = sample values with population sigma at 0.69 plotted at y.

c = sample values with population sigma at 1.60 plotted at z.

NOTE: 12 obs hidden.

Table 1. Estimation Results by Left Censored and Truncated Below Likelihood Methods from Five simulations of n(=25 and 50) Largest from 200 Lognormal Observations: N(3.1, 0.0625)

		Repetition				
Likelihood (n)	Parameter	1	2	3	4	5
Left Censored (25)	$\hat{\mu}$	3.21	3.01	3.15	3.18	3.17
	$\hat{\sigma}^2$	0.0404	0.1077	0.0477	0.0257	0.0471
	$v(\hat{\mu})$	0.0025	0.0068	0.0030	0.0016	0.0030
	$C(\hat{\mu}, \hat{\sigma}^2)$	−0.0007	−0.0028	−0.0008	−0.0003	−0.0008
	$v(\hat{\sigma}^2)$	0.0002	0.0015	0.0003	0.0001	0.0003
Left Censored (50)	$\hat{\mu}$	3.05	3.04	3.09	3.12	3.09
	$\hat{\sigma}^2$	0.0881	0.0951	0.0667	0.0380	0.0723
	$v(\hat{\mu})$	0.0018	0.0019	0.0013	0.0008	0.0014
	$C(\hat{\mu}, \hat{\sigma}^2)$	−0.0007	−0.0008	−0.0005	−0.0002	−0.0005
	$v(\hat{\sigma}^2)$	0.0005	0.0005	0.0003	0.0001	0.0003
Truncated Below (25)	$\hat{\mu}$	3.49	3.16	3.21	3.28	3.41
	$\hat{\sigma}^2$	0.0105	0.0845	0.0423	0.0180	0.0213
	$v(\hat{\mu})$	0.0028	0.2561	0.1788	0.0410	0.0178
	$C(\hat{\mu}, \hat{\sigma}^2)$	−0.0003	−0.0419	−0.0187	−0.0034	−0.0021
	$v(\hat{\sigma}^2)$	0.0004	0.0074	0.0021	0.0003	0.0003
Truncated Below (50)	$\hat{\mu}$	3.35	2.82	3.17	3.24	3.20
	$\hat{\sigma}^2$	0.0320	0.1320	0.0539	0.0231	0.0367
	$v(\hat{\mu})$	0.0043	0.3878	0.0399	0.0091	0.0098
	$C(\hat{\mu}, \hat{\sigma}^2)$	−0.0008	−0.0632	−0.0064	−0.0011	−0.0017
	$v(\hat{\sigma}^2)$	0.0002	0.0109	0.0011	0.0002	0.0003

Table 2. Estimation Results by Left Censored and Truncated Below Likelihood Methods from Five simulations of n (=25 and 50) Sampled from [3n/2] + 1 Largest of 200 Lognomal Observations: N(3.1, 0.0625).

Likelihood (n)	Parameter	Repetition 1	2	3	4	5
Left Censored (25)	$\hat{\mu}$	2.97	2.91	2.94	3.06	2.97
	$\hat{\sigma}^2$	0.0955	0.1240	0.1010	0.0439	0.0972
	$v(\hat{\mu})$	0.0060	0.0078	0.0063	0.0027	0.0061
	$C(\hat{\mu}, \hat{\sigma}^2)$	−0.0024	−0.0035	−0.0026	−0.0007	−0.0024
	$v(\hat{\sigma}^2)$	0.0012	0.0020	0.0013	0.0002	0.0012
Left Censored (50)	$\hat{\mu}$	did	2.76	2.82	did	2.86
	$\hat{\sigma}^2$	not	0.1286	0.0988	not	0.0905
	$v(\hat{\mu})$	converge	0.0082	0.0062	converge	0.0057
	$C(\hat{\mu}, \hat{\sigma}^2)$		−0.0038	−0.0025		−0.0022
	$v(\hat{\sigma}^2)$		0.0021	0.0013		0.0010
Truncated Below (25)	$\hat{\mu}$	3.38	2.71	3.34	3.23	3.42
	$\hat{\sigma}^2$	0.0305	0.1527	0.0352	0.0258	0.0232
	$v(\hat{\mu})$	0.0129	1.1449	0.0172	0.0399	0.0053
	$C(\hat{\mu}, \hat{\sigma}^2)$	−0.0021	−0.1615	−0.0029	−0.0044	−0.0008
	$v(\hat{\sigma}^2)$	0.0004	0.0244	0.0006	0.0005	0.0002
Truncated Below (50)	$\hat{\mu}$	3.04	2.42	2.90	3.14	2.60
	$\hat{\sigma}^2$	0.0872	0.1913	0.0962	0.0362	0.1328
	$v(\hat{\mu})$	0.0695	1.2348	0.1786	0.0172	0.8711
	$C(\hat{\mu}, \hat{\sigma}^2)$	−0.0138	−0.1757	−0.0289	−0.0025	−0.1053
	$v(\hat{\sigma}^2)$	0.0030	0.0263	0.0050	0.0004	0.0134

Table 3. Estimation Results by Left Censored and Truncated Below Likelihood Methods from Five Simulations of n (=25 and 50) Sampled from 2n Largest of 200 Lognormal Observations: N(3.1, 0.0625).

		Repetition				
Likelihood (n)	Parameter	1	2	3	4	5
Left Censored (25)	$\hat{\mu}$	2.87	2.88	2.95	3.01	2.94
	$\hat{\sigma}^2$	0.1131	0.1016	0.0759	0.0430	0.0861
	$v(\hat{\mu})$	0.0071	0.0065	0.0048	0.0027	0.0054
	$C(\hat{\mu}, \hat{\sigma}^2)$	−0.0031	−0.0027	−0.0017	−0.0007	−0.0020
	$v(\hat{\sigma}^2)$	0.0016	0.0013	0.0007	0.0002	0.0009
Left Censored (50)	$\hat{\mu}$	did	2.71	2.76	2.96	2.95
	$\hat{\sigma}^2$	not	0.1264	0.1033	0.0592	0.0834
	$v(\hat{\mu})$	converge	0.0081	0.0065	0.0012	0.0017
	$C(\hat{\mu}, \hat{\sigma}^2)$		−0.0037	−0.0027	−0.0004	−0.0006
	$v(\hat{\sigma}^2)$		0.0021	0.0014	0.0002	0.0004
Truncated Below (25)	$\hat{\mu}$	3.37	2.62	3.22	3.26	3.31
	$\hat{\sigma}^2$	0.0228	0.1356	0.0378	0.0164	0.0293
	$v(\hat{\mu})$	0.0040	1.6750	0.0406	0.0096	0.0140
	$C(\hat{\mu}, \hat{\sigma}^2)$	−0.0006	−0.2053	−0.0060	−0.0011	−0.0022
	$v(\hat{\sigma}^2)$	0.0002	0.0265	0.0010	0.0001	0.0004
Truncated Below (50)	$\hat{\mu}$	3.14	2.78	3.03	3.15	3.08
	$\hat{\sigma}^2$	0.0609	0.1130	0.0633	0.0292	0.0600
	$v(\hat{\mu})$	0.0130	0.2477	0.0473	0.0073	0.0343
	$C(\hat{\mu}, \hat{\sigma}^2)$	−0.0029	−0.0417	−0.0082	−0.0011	−0.0062
	$v(\hat{\sigma}^2)$	0.0008	0.0075	0.0016	0.0002	0.0013

Table 4. Schematic of Probability Paper Regression Data Inputs

Job Title (P_i)*	Occasion ° Worker †	D_j ; $j = 1, \ldots, d$ $W_{k(ij)}$; $k = 1, \ldots, m_{ij}$
$J_N \left(\frac{N - \frac{1}{2}}{N} \right)$		
$J_{N-1} \left(\frac{N-1}{N} \right)$		
$J_{N-2} \left(\frac{N-2}{N} \right)$		
.		X_{ijk} (measured exposure)
.		
.		
$J_3 \left(3/N \right)$		
$J_2 \left(2/N \right)$		
$J_1 \left(\frac{1}{N} \right)$		

* Job Titles, ranked $i = 1$ to N, from the lowest to the highest expected exposures. Only the extremely smallest and largest need to be identified (from production/plant knowledge). Each job's percentile rank, P_i, is determined by the job title ranking and its exposure (X_{ijk}) obtained by measuring the kth worker on the jth day.

° Occasions or days on which the plant is visited and selected jobs are sampled.

† The number of workers at the ith job on day j may vary with job titles or days selected for sampling.

APPENDIX A

LEFT CENSORING LIKELIHOOD ALGORITHM

```
(*  ****     Left Censoring likelihood, n=25, N=200     **** *)
Print["    ****    type A data, first replication    ****   "]

(* read data : a25_1.dat  --- type A data with n=25, N=200 *)
y=ReadList["a25_1.dat"]
v1=0.66; v2=0.163; c=5; k=5.75;m=25; n=200;

(* compute the initial values *)
<<Statistics`DescriptiveStatistics`
a1=Median[y]; a2=Min[y];
b1=1.55; b2=1.15; (*b1=P{Z < 1-n/2N}, b2=P{Z < 1-n/N} *);
sigma0=(a1-a2)/(b1-b2);
v0=sigma0^2;
u0=(a1+a2-b1*sigma0-b2*sigma0)/2;

(* specify the left censored likelihood --- logL   *)
z=v2*(y-u)/Sqrt[v]+v1
y1=Min[y]
z1=v2*(y1-u)/Sqrt[v]+v1
logL=Sum[(c-1)*Log[z[[i]]]-(k+1)*Log[1+z[[i]]^c],{i,1,m}]+m*Log[v2*c*k]-
(m/2)*Log[v]+(n-m)*Log[1-(1+z1^c)^(-k)]

(* solve the mean and variance from logL *)
du=D[logL,u]
dv=D[logL,v]
root=FindRoot[{du==0,dv==0},{u,u0},{v,v0}]

(* estimate the covariance matrix --- cov  *)
du2=D[logL,{u,2}]
dv2=D[logL,{v,2}]
duv=D[logL,u,v]
dvu=D[logL,v,u]
(*  *** Covariance matrix of (uhat,vhat) ***   *)
m={{-du2,-duv},{-dvu,-dv2}}/.{root[[1]],root[[2]]}
cov=Inverse[m]
```

Remark: Program is in Mathematica and the type A data set "a25_1.dat" denotes n = 25 largest observations from 200 lognormals: N(3.1, 0.0625). The type B (C) data set was a random selection of n observations from [3n/2] + 1 (2n) largest observations from 200 lognormals: N(3.1, 0.0625).

APPENDIX B

TRUNCATION BELOW LIKELIHOOD ALGORITHM

```
(* ****        Trunctaion likelihood , n=25 , N=200  **** *)
Print["    ****    type A data, first replication     **** "]

(* read data : a25_1.dat --- type A data with n=25, N=200 *)
y=ReadList["a25_1.dat"]

(* specify the parameters for mean, variance, c and k in Burr dist. *)
v1=0.66; v2=0.163; c=5; k=5.75;
n=200; n1=25;

(* compute the initial values for mean and variance *)
<<Statistics`DescriptiveStatistics`
a1=Median[y]; a2=Min[y];
b1=1.55; b2=1.15; (*b1=P{Z < 1-n/2N}, b2=P{Z < 1-n/N} *);
sigma0=(a1-a2)/(b1-b2);
v0=sigma0^2;
u0=(a1+a2-b1*sigma0-b2*sigma0)/2;
sigma0=Sqrt[v0]

(* specify the log likelihood --- logL *)
r=Min[y]
sigma=Sqrt[v]
z=v2*(y-u)/sigma + v1
L=(c-1)*Log[z]-(k+1)*Log[1+z^c]
logL=Sum[L[[i]],{i,1,n1}]-(n1/2)*Log[v]+k*n1*Log[1+(v2*(r-u)/sigma+v1)^c]

(* define the first derivatives w.r.t mean (u) and variance (v)
    &  solve mean and variance from the truncated likelihood *)
du=D[logL,u]
dv=D[logL,v]
root=FindRoot[{du==0,dv==0},{u,u0},{v,v0}]

(* estimate the covariance matrix --- cov *)
du2=D[logL,{u,2}]
dv2=D[logL,{v,2}]
duv=D[logL,u,v]
(*  *** Covarrance matrix of (uhat,vhat,rhat) ***  *)
m={{-du2,-duv},{-duv,-dv2}}/.{root[[1]],root[[2]]}
cov=Inverse[m]
```

Remark: See Remark for Appendix A.