Final Progress Report

Project Title: Statistical Problems in Occupational Safety and Health

Grant Number: OH03628

Funding Period: May 1, 2000–April 30, 2004

Sponsor: National Institute of Occupational Safety and Health

PI: Thomas Mathew, Department of Mathematics and Statistics, University of Maryland, 1000 Hilltop Circle, Baltimore, Maryland 21250

Co-investigator: K. Krishnamoorthy, Department of Mathematics, University of Louisiana at Lafayette, Lafayette, Louisiana 70504

Table of Contents

Abstract

The technology for monitoring occupational exposure depends crucially on the statistical analysis of the exposure data, for the purpose of developing appropriate strategies to monitor and improve occupational safety and health: in particular, for assessing health risks, for identifying cost-effective intervention strategies. for developing exposure-response relationships, for evaluating low-cost, easy to use. non-invasive test kits, and a host of other activities that will enhance occupational health. Furthermore, the collection of exposure data involves considerable human and financial resources. Consequently, it is very crucial that the data be used effectively. Needless to say, the use of valid statistical methods for exposure data analysis is critical to this endeavor. The funded project was aimed to develop such statistical methods.

The following have been accomplished based on the funded project. Satisfactory procedures have been obtained for analyzing exposure data using the lognormal distribution; in particular, satisfactory confidence intervals and test procedures have been developed. These findings are important because of the relevance and widespread use of the lognormal distribution for exposure data analysis. Furthermore, the newly developed procedures are quite accurate in meeting the required statistical properties in terms of the coverage probability of the confidence interval, and type I error probability of the test. For the problem of testing the equivalency of a sampling device to an OSHA standard, a statistically rigorous procedure has been developed. This is a problem that is routinely encountered while developing cheaper or more efficient exposure assessment strategies. A thorough investigation of the computation of tolerance limits has been carried out for a random effects model. This problem is of significance since an upper tolerance limit provides a limit below which most of the exposure measurements are expected to lie. Furthermore, a random effect model is relevant for exposure assessment due to the fact that such a model provides a way to account for the dependency among the exposure group. For such models, appropriate statistical procedures have also been developed for checking if the proportion of workers for whom the mean exposure exceeds the occupational exposure limit is above a threshold. We have provided a solution to the hypothesis testing problem involving the proportion of workers whose mean exposure exceeds the occupational exposure limit. Our solutions are again based on a one-way random effects model for the log transformed shift-long personal exposure measurements, where the random effect in the model represents an effect due to the worker. Numerical results show that our procedures exhibit satisfactory performance regardless of the sample size. A similar procedure is then employed for testing hypotheses concerning the overall mean exposure.

Any statistical procedure for analyzing exposure data will be useful only if the procedure is easy to compute and implement. Furthermore, it is also important that the procedure be applicable to small sample sizes, since exposure monitoring could be time consuming and expensive, and large amounts of data are typically unavailable. The funded work was motivated by all of these considerations, and the results we obtained meet these requirements.

Significant Findings

In the context of exposure data analysis, many of the problems that come up are quite different from the problems addressed in the mainstream statistics literature, and available in standard software packages. Routine application of standard statistical techniques to these problems, though easy, can be inappropriate and inefficient; this will result in ineffective use of the data and, more importantly, invalid conclusions. New methodologies and novel approaches were required to handle some of these problems. The project aimed to: (i) develop such methodologies and approaches, (ii) illustrate their relevance and applicability for exposure data analysis, and (iii) address the related computational issues and provide programming codes, so that occupational hygienists and other practitioners can easily apply the proposed procedures.

The main difficulty in using the lognormal distribution is that the parameters of interest (for example, the lognormal mean) depend on both the mean and the variance of the normal distribution of the logged data. This makes the analysis quite difficult. The available procedure is computationally intensive, and researchers in occupational hygiene have noted this difficulty. We used the novel concepts of generalized p-values and generalized confidence intervals, recently introduced in the statistics literature, and developed simple alternative methods for exposure data analysis using the lognormal distribution. The procedure based on the generalized p-value and the generalized confidence interval are quite simple to compute, and they are also highly accurate. Furthermore, they are applicable to small samples. We have developed the theoretical framework, and have also numerically investigated the accuracy of this novel approach. Our work clearly shows that for analyzing exposure data based on the lognormal distribution, generalized p-values and generalized confidence intervals are preferable to ad hoc or approximate methods, in terms of simplicity and accuracy.

The OSHA regulations allow the use of an alternate sampling device for exposure monitoring provided the device has been been demonstrated to be equivalent to the standard device. OSHA also defines the accuracy of the monitoring device for measuring airborne chemicals such as benzene and sulfur dioxide. Typically, the OSHA criterion is that 90% of the readings of the sampling device should be within plus or minus 25% of the readings obtained by the standard device or within plus or minus 25% of the actual airborne chemical concentration. Earlier work dealing with testing this criterion appears to be unsatisfactory since the statistical problem was not rigorously defined or solved. In our work, we developed two statistical tests for establishing that an alternative measuring device of airborne chemicals or dust is equivalent to the OSHA standard. Furthermore, we showed numerically that the tests are very accurate. The statistical procedures were also illustrated using a real example.

Random effects models are becoming increasingly relevant in analyzing occupational exposure data. The problems and the difficulties that one encounters in this context should be clear from the articles by Lyles, Kupper, and Rappaport (*Journal of Agricultural, Biological, and Environmental Statistics* (1997, pp. 64-86), and *The Annals of*

Occupational Hygiene (1997, pp. 63-76)). The quantity of interest in these articles is the proportion of workers for whom the mean exposure exceeds the occupational exposure limit, and the above authors have developed appropriate statistical procedures for checking if this proportion exceeds a specified level. They used large sample procedures, whose performance is not always satisfactory. Furthermore, large sample procedures are very often inapplicable, since typical samples will seldom be large. We have succeeded in developing procedures that are applicable to small samples. The procedures are once again based on the concepts of generalized p-values and generalized confidence intervals. Our procedures are quite easy to compute and implement, and their performance is quite satisfactory. We have also illustrated our procedures for analyzing actual exposure data.

The use of statistical tolerance intervals for exposure assessment has been noted by several researchers in occupational hygiene. Roughly speaking, an upper tolerance limit provides a limit below which most (say, 95%) of the exposure measurements are expected to lie, with a high level of confidence. (Similarly, one can also define a lower limit, or both lower and upper limits). For a random effects model, the derivation of a tolerance limit presents some difficulties, because approximations have to be used due to the fact that the tolerance interval condition results in a limit that depends on unknown parameters. We again used the idea of a generalized confidence interval, and derived appropriate tolerance limits. We provided computational algorithms and the sources of free softwares that can be used to compute relevant table values.

In the context of exposure samples with some values below the detection limit (DL), we have investigated the strategy to be adopted when these values have to be replaced by artificial data. Given the rather widespread practice of replacing values below the DL with DL/2, DL/$\sqrt{2}$, zero, and so on, without giving any careful thought to its implications, our proposal is to come up with a procedure that provides satisfactory performance, and yet is consistent with the standard practice of replacing the below DL values by artificial data. We are advocating an *imputation approach* to the problem. That is, replace the below DL sample data with observations estimated based on the data that is above the DL, instead of replacing them with arbitrary values such as DL/2. Numerical results show that our imputation approach is extremely accurate. This work is under progress.

Translation of Findings

The funded project addressed statistical models and problems that often arise in industrial hygiene, and appeared in the occupational hygiene literature. We proposed to develop new theory, when necessary, and to apply already developed theory, for the purpose of analyzing exposure data.

The problems that arise in exposure data analysis are far from routine, and do require the development of new statistical methods. The solutions we developed have strong components of statistical theory; something that is not of direct interest to an industrial hygienist. Therefore, we adopted the following strategy to disseminate our work. We published our theoretical work in main stream statistical journals. We also prepared

a version emphasizing only the application of the theoretical results for the purpose of exposure data analysis, along with the computational details and practical examples, for publication in an industrial hygiene journal, so that our work will reach practitioners in occupational hygiene. For example, our work on the lognormal distribution is theoretical in nature and is published in a statistics journal (*Journal of Statistical Planning and Inference*). However, the article by Krishnamoorthy, Mathew and Ramachandran (2004, submitted to the *Journal of Occupational and Environmental Hygiene*) describes our methodology for the industrial hygiene audience. In addition, to aid the users of statistical procedures, we posted relevant computational algorithms, Fortran codes and "help files" at co-investigator's homepage http://www.ucs.louisiana.edu/~kxk4695.

We plan to adopt the same strategy on our currently ongoing work dealing with the detection limit problem.

Scientific Report

The strategies for monitoring occupational exposure depend crucially on the statistical analysis of the exposure data. For analyzing such data, it is quite easy to use a routine software package, ignoring the appropriateness of the underlying statistical model, or the validity of the relevant assumptions. However, this could very well result in incorrect conclusions, and from the point of view of exposure monitoring, the well being of the workers could be compromised or the employers could be unduly burdened. Hence it is of utmost importance that appropriate statistical models and procedures be used. It is also highly desirable that such procedures are easy to compute and implement. Furthermore, it is also important that the procedures be applicable to small samples, because exposure monitoring could be time consuming and expensive, and large amounts of data are typically unavailable. The funded project was aimed to achieve all of the above. We believe that the work that has been accomplished is both promising and satisfactory in terms of meeting the above requirements.

In the context of exposure data analysis using the lognormal distribution, the proposal emphasized the development and applications of the recent concepts of *generalized p-values* and *generalized confidence intervals*. The concept of the generalized p-value was introduced by Tsui and Weerahandi (1989, *Journal of the American Statistical Association*), and that of the generalized confidence interval by Weerahandi (1993, *Journal of the American Statistical Association*). It turns out that these concepts provide a simple procedure for carrying out a variety of analysis based on the lognormal distribution, and based on random effects models. Also, they are accurate even for small samples. Here we shall briefly explain the procedure for obtaining confidence interval for the lognormal mean.

Let X_1, X_2, ..., X_n be a lognormal random sample representing exposure measurements on n workers, so that $Y_i = \ln(X_i)$, $i = 1, 2, ..., n$, is a random sample from a normal distribution with mean and variance, say μ and σ^2, respectively. The arithmetic mean (AM) of the lognormal distribution is then given by $\exp(\mu + \frac{\sigma^2}{2})$. Thus confidence

intervals and tests concerning the AM is equivalent to confidence intervals and tests concerning $\mu + \frac{\sigma^2}{2}$. Let $\bar{Y}$ and S^2 denote the sample mean and variance of the Y_i's. We shall also denote by $\bar{y}$ and s^2, the observed values of $\bar{Y}$ and S^2, respectively. In order to compute a generalized confidence interval for $\mu + \frac{\sigma^2}{2}$, we shall first define a *generalized pivot statistic* T, that is a function of the random variables $\bar{Y}$ and S^2, and their observed values $\bar{y}$ and s^2, where T could also depend on the unknown parameters. However, T must satisfy the following conditions.

(a) The observed value of T is $\mu + \frac{\sigma^2}{2}$

(b) The distribution of T is free of any unknown parameters.

A generalized pivot statistic T that satisfies these conditions is given by

$$T = \bar{y} - \frac{\bar{Y} - \mu}{S/\sqrt{n}} \frac{s}{\sqrt{n}} + \frac{1}{2} \frac{\sigma^2}{S^2} s^2 = \bar{y} - \frac{Z}{U/\sqrt{n-1}} \frac{s}{\sqrt{n}} + \frac{1}{2} \frac{s^2}{U^2/(n-1)},$$

where $Z = \sqrt{n}(\bar{Y} - \mu)/\sigma \sim N(0,1)$ independently of $U^2 = (n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. Here χ^2_r denotes the central chisquare distribution with df $= r$. From the second expression for T given above, it is clear that the distribution of T is free of any unknown parameters. The observed value of T is obtained by replacing $\bar{Y}$ and S^2 by $\bar{y}$ and s^2, respectively, in the first expression for T given above, and this observed value is easily seen to be $\mu + \frac{\sigma^2}{2}$. Thus T satisfies the conditions (a) and (b) given above. If $T(1-\alpha)$ denotes the $100(1-\alpha)$th percentile of T, then $T(1-\alpha)$ is the $100(1-\alpha)\%$ generalized upper confidence limit for η. A lower confidence limit, or two-sided confidence limits can be similarly defined. For testing hypotheses concerning $\mu + \frac{\sigma^2}{2}$, a p-value (the *generalized p-value*) can also be defined.

We found that the generalized p-value and the generalized confidence interval are quite easy to calculate, by simulation. Our extensive numerical results show that the generalized confidence interval practically coincides with the interval obtained by a procedure due to Land. The importance of this approach becomes even more evident once it is noted that: (i) the procedures can be easily extended for comparing two lognormal means, both for testing hypotheses and obtaining confidence intervals for the ratio of two lognormal means, and (ii) the procedures are applicable to small samples. As far as we are aware, no other small sample procedures exist, for comparing two lognormal means. These aspects are all pointed out in our article, Krishnamoorthy and Mathew (2003). Using an example, we also noted that comparing the means of the logged data in two samples (using the usual t-test) can produce a different conclusion, as opposed to comparing the lognormal means. In other words, the conclusions regarding exposure assessments at two different sites can be quite different, depending on whether we compare the actual lognormal means of exposure data, or the means of the logged exposure data!

Procedures based on the generalized p-value and generalized confidence interval are versatile enough to handle any parameter associated with a single lognormal distribution,

or two lognormal distributions. For example, the variance of the lognormal distribution is given by $\exp(2\mu+\sigma^2)[\exp(\sigma^2)-1]$. It is clearly of interest to characterize the variability among the exposure measurements; this leads us to the computation of confidence intervals and tests for the variance $\exp(2\mu+\sigma^2)[\exp(\sigma^2)-1]$. The variance could be a major component in rigorous exposure monitoring programs. In Krishnamoorthy, Mathew and Ramachandran (2004, submitted to the *Journal of Occupational and Environmental Hygiene*), the ideas of generalized p-values and generalized confidence intervals are developed for obtaining tests and confidence intervals for the lognormal variance.

In an article that appeared in the *American Industrial Hygiene Association Journal*, Krishnamoorthy and Mathew (2002a) have developed statistical methods for establishing the equivalency of a sampling device to the OSHA standard. Typically, the OSHA criterion is that 90% of the readings of the sampling device should be within plus or minus 25% of the readings obtained by the standard device or within plus or minus 25% of the actual airborne chemical concentration. Earlier work dealing with testing this criterion appears to be unsatisfactory since the statistical problem was not rigorously defined or solved. The OSHA requirements for equivalency can be regarded as a hypothesis testing problem about θ = proportion of the alternative device readings that are within $\pm 25\%$ of the standard readings. The problem then is to test the hypotheses

$$H_0 : \theta \leq 0.90 \quad \text{vs.} \quad H_a : \theta > 0.90.$$

In our work, we proposed two statistical tests, depending on whether the data follow the lognormal distribution or not. The tests are illustrated by applying them to establish the equivalence of an alternate cotton dust sampler to the Lumsden-Lynch vertical elutriator.

The use of cheap and easy to use alternative sampling devices for exposure measurement is clearly of considerable interest, and related research appears quite often in the industrial hygiene literature. The procedures we have developed are quite simple and very easy to use; the required tables are provided in our paper.

We have been surprised to see the applicability of the concepts of the generalized p-values and the generalized confidence intervals to a variety of problems relevant to exposure assessment. In Krishnamoorthy and Mathew (2002b), we have considered a one-way random effects model for the log transformed shift-long personal exposure measurements, where the random effect in the model represents an effect due to the worker. We have addressed a hypothesis testing problem involving the proportion of workers for whom the mean exposure exceeds the occupational exposure limit, based on balanced data. Numerical results show that our procedures exhibit satisfactory performance regardless of the sample size; in particular, for small samples. A similar procedure is then employed for testing hypotheses concerning the overall mean exposure. Earlier work on this problem use large sample approximations. The above work deals with a balanced data situation (i.e., the same number of exposure measurements are available on each worker). In Krishnamoorthy and Guo (2005), similar problems are addressed for the case of unbalanced data.

In Krishnamoorthy and Mathew (2004), we have again considered a one-way random

effects model for the log transformed exposure measurements, and have made a rather exhaustive study of the problem of constructing tolerance intervals. The relevance of tolerance intervals for exposure assessment has been pointed out by Tuggle (1982, *American Industrial Association Journal*) and Lyles and Kupper (1996, *American Industrial Association Journal*). Specifically, if the upper tolerance limit based on a sample of exposure measurements is less than the occupational exposure limit (OEL), then it indicates that majority of the exposure data are within the OEL, and hence exposure monitoring might be reduced or terminated until a process change occurs. In our article, we have derived upper tolerance limits using the generalized confidence interval idea. Both balanced and unbalanced data situations are considered.

We shall now describe our work in progress dealing with the important problem of exposure samples containing values below the detection limit (DL). Suppose we have a sample of n exposure measurements of which k are below the DL. Let us assume that the sample is from a continuous distribution with cumulative distribution function (cdf) $F(x;\theta)$, where θ is a vector parameter. An estimate of the proportion of the data below the DL is given by

$$P_{DL} = F(DL;\hat{\theta}),$$

where $\hat{\theta}$ is an estimate of θ based on the $n-k$ measurements above the DL. Using this probability estimate, we can impute (i.e., estimate) the values that are below the DL. For instance, we can generate $x_1, ..., x_k$ from the estimated cdf $F(x;\hat{\theta})$ so that $F(x_i;\hat{\theta}) \leq P_{DL}$, and impute the values below the DL by $x_1, ..., x_k$. The x_i's can then be generated as

$$x_i = F^{-1}(u_i;\hat{\theta}),\ i = 1, ..., k,$$

where $u_1, ..., u_k$ are uniform random numbers from $(0, P_{DL})$. Inference about θ or a function of θ can be made using the k imputed values and the $n-k$ measurements above the DL.

We shall now describe the progress we have made in the case of the normal distribution. Note that our results also apply for exposure data following the lognormal distribution. For the normal distribution, an estimate of the proportion of exposure data below the DL is given by

$$P_{DL} = \Phi((\ln(\mathrm{DL}) - \bar{y}_l)/s_l),$$

where $\bar{y}_l$ and s_l denote respectively the mean and standard deviation based on the $n-k$ data above the DL, and Φ denotes the cdf of the standard normal distribution. Using this probability estimate, we can impute (i.e., "estimate") the values that are below the DL by

$$x_i = \frac{(n-k)\bar{x}_l + k(DL)}{n} + z_i s_l,\quad i = 1, ..., k,$$

where $z_1, ..., z_k$ are the simulated random numbers from the standard normal distribution. These random numbers should be generated so that $\Phi(z_i) \leq P_{DL}$ for $i = 1, ..., k$. This can be done by generating uniform random numbers $u_1, ..., u_k$ from $(0,\ P_{DL})$, and

then setting $z_i = \Phi^{-1}(u_i)$, $i = 1, ..., k$. Suppose one is interested in constructing confidence interval for the mean based on an imputed sample. Let $\bar{x}^*$ and s^* denote respectively the mean and standard deviation of the pooled data containing $n-k$ measurements that are above the DL, and k simulated data using the above equation. Then, a $1-\alpha$ confidence interval for μ based on the imputed sample is given by

$$\bar{x}^* \pm t_{n-k-1}\frac{s^*}{\sqrt{n}}.$$

Notice that the df is adjusted, accounting for the k imputed values. We carried out extensive numerical studies to assess the performance of such an imputation method, and compared it with the usual practice of replacing the values below the DL by DL/2. For this, we assumed a $N(0, \sigma^2)$ distribution for the logged data, and assumed that DL= $\mu - \sigma = -\sigma$ (for simulation purposes). Note that we have chosen DL such that about 16% of the population data are less than the DL. Here are the coverage probabilities of the confidence intervals for μ, and for the 95th percentile, based on the two approaches:

Table 1. Coverage Probabilities of the 95% Confidence Intervals for μ when (1) below DL values are replaced by imputed values, and (2) below DL values are replaced by DL/2

n	20	30	40	40	30	60
σ	2	2	1	2	3	2
(1)	0.94	0.95	0.95	0.94	0.94	0.94
(2)	0.89	0.84	0.79	0.79	0.87	0.66

Table 2. Coverage Probabilities of the 95% Confidence Intervals for the 95th percentile when (1) below DL values are replaced by imputed values, and (2) below DL values are replaced by DL/2

n	20	20	30	30	30	60
σ	2	5	1	2	3	2
(1)	0.95	0.95	0.95	0.95	0.95	0.94
(2)	0.80	0.80	0.75	0.76	0.75	0.65

It is clear that the common practice of replacing the below DL values by DL/2 can really bring down the coverage probability, and as a result, it may lead to incorrect conclusions. In other words, what is routinely done can lead to very inaccurate results.

The work described above for the normal distribution is actually complete. We are now investigating our imputation approach for testing hypothesis, and for computing a confidence interval for the lognormal mean, based on the concepts of generalized p-values and generalized confidence intervals. Preliminary numerical findings show that the imputation approach once again give very accurate results. In other words, once this

work is complete, the imputation idea along with the concepts of generalized p-values and generalized confidence intervals will provide an accurate method for analyzing log-normally distributed exposure data, when the data include below DL values.

Journal Articles

Krishnamoorthy K, Mathew T: Statistical methods for establishing equivalency of a sampling device to the OSHA standard. American Industrial Hygiene Association Journal 63: 567-571, 2002a.

Krishnamoorthy K, Mathew T: Assessing occupational exposure via the one-way random model with balanced data. Journal of Agricultural, Biological and Environmental Statistics 7: 440-457, 2002b.

Krishnamoorthy K, Mathew T: Inferences on the means of lognormal distributions using generalized p–values and generalized confidence intervals. Journal of Statistical Planning and Inference 115: 103-121, 2003.

Krishnamoorthy K, Mathew T: Generalized confidence limits and one-sided tolerance limits in balanced and unbalanced one-way random models. Technometrics 46: 44-52, 2004.

Krishnamoorthy K, Guo H: Assessing occupational exposure via the one-way random model with unbalanced data. Journal of Statistical Planning and Inference 128: 219-229, 2005.

Inclusion of gender and minority study subjects. Not applicable.

Inclusion of Children. Not applicable.

Materials available for other investigators

Necessary computational algorithms with details and Fortran codes to compute confidence intervals for a single lognormal mean, for the ratio of two lognormal means, and associated power computations, are all posted at the co-investigator's website http://www.ucs.louisiana.edu/~kxk4695. All the published articles are also posted at this site. These materials can be accessed and downloaded freely.