American Industrial Hygiene Association Journal

ISSN: 0002-8894 (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/aiha20

Sample Size Formulae for Estimating the True Arithmetic or Geometric Mean of Lognormal Exposure Distributions

Paul Hewett

To cite this article: Paul Hewett (1995) Sample Size Formulae for Estimating the True Arithmetic or Geometric Mean of Lognormal Exposure Distributions, American Industrial Hygiene Association Journal, 56:3, 219-225, DOI: 10.1080/15428119591017042

To link to this article: https://doi.org/10.1080/15428119591017042

AIHA JOURNAL Advances in Industrial & Environmental Hygiene

AM. IND. HYG. ASSOC. J. 56:219–225 (1995)

SAMPLE SIZE FORMULAE FOR ESTIMATING THE TRUE ARITHMETIC OR GEOMETRIC MEAN OF LOGNORMAL EXPOSURE DISTRIBUTIONS

Paul Hewett

National Institute for Occupational Safety and Health, 944 Chestnut Ridge Road, Morgantown, WV 26505

Formulae are presented for calculating the approximate sample size needed to estimate the true arithmetic mean or true geometric mean exposure for an exposure group to within a specified accuracy (±x% of the true arithmetic or geometric mean) with a specified level of confidence. These formulae are intended for use in prospective or cross-sectional occupational health studies, or when building an exposure database for use in assessing long-term changes in worker health status. They are applicable where the investigator is satisfied that the distribution of exposures within a group can be approximated by a lognormal distribution. The formulae were validated by computer simulation and show that large sample sizes are required when the existing parameter estimates were derived from a limited number of prior measurements and/or the true exposure distribution has a large geometric standard deviation. When summed across all exposure groups, an unreasonable total sample size may result. The total sample burden can be reduced in several ways: (1) A pilot study should be used to provide reasonably precise initial estimates of the distribution parameters for each exposure group. This may require 20 or more measurements per group. (2) Workers should be grouped into exposure groups where the group geometric standard deviation is two or less. (3) The desired accuracy should be kept at a reasonable level, perhaps between 20 and 30% of the true parameter. Accuracy levels less than 20% can result in large total sample size requirements.

In the practice of occupational epidemiology there are several criteria that should be met in order to conclude that there is a causal relationship between exposure to a toxic agent and a specific disease or disease process.[1] Among these are the criteria that the prevalence or incidence of the disease should be greater in exposed workers than in unexposed workers, and that there should be a dose-response relationship, or biologic gradient. In the occupational setting, exposure usually serves as a surrogate for dose; thus analyses in most occupational health studies involve exposure-response relationships. Here it is critical to obtain sufficient exposure data in order to accurately estimate the exposure-response relationship. One of the characteristics of occupational studies is that the exposures are assumed to be lognormally distributed within each exposure group.[2,3] The purpose of this article is to present simple formulae for calculating the approximate sample size required to estimate either the true arithmetic or true geometric mean of a lognormal distribution to within a specified accuracy (±x% of the true arithmetic or geometric mean) with a specified level of confidence. These formulae should be of interest to the industrial hygienist involved in prospective or cross-sectional studies, or establishing exposure databases to be used to assess long-term changes in worker health status.

BACKGROUND

One problem with collecting exposure data for occupational health studies is that individual exposure measurements are not reliable estimates of a worker's yearly average, or even weekly average, exposure. In contrast, the biological measurements comprising the medical component of the study often are accurate measurements of the response. In the case of cumulative toxicants the body itself has integrated previous exposures such that the parameter measured reflects the dose absorbed over a period of time. Since single exposure measurements are, by themselves, crude estimates of the worker yearly average exposure, multiple measurements are necessary to precisely estimate individual yearly exposures. Since few investigators have the resources to do repeat sampling of each worker's exposure, workers are commonly aggregated, *a priori*, into groups where the exposures, and presumably the risk of disease, are assumed to be similar. The group mean exposure is then assigned to all workers within the group, whether sampled or unsampled.[4] An exposure group can refer to a grouping of similar jobs[5] or to an area within a plant, or across an industry, where tasks, hazardous substances, and the engineering controls are similar.[4] Other names for an exposure group include exposure zone, occupational title, occupational group, homogeneous risk group, homogeneous exposure group, and "uniformly exposed" group.

To a large degree, the name reflects the amount of information available for categorizing workers.

The Exposure Group Paradigm

The practice of using a group mean exposure as a surrogate for the individual means of group members is based on the paradigm that it is possible to group workers such that the distribution of exposures for each worker is similar to the distribution of exposures for all individuals in the group. Thus the group mean exposure can be expected to be a reasonable estimate of each worker's individual mean exposure. The advantage of this approach is that fewer total measurements are required to estimate the mean exposure of the single group exposure distribution than are necessary to estimate the means of multiple worker exposure distributions.

The process of identifying exposure groups and assigning workers to them has not been critically examined. Of particular importance is the level of acceptable variability in the distribution of individual worker mean exposures within each group. In addition, the robustness of the lognormal assumption has not been evaluated for those situations where the true distribution departs from the lognormal, e.g., when a bimodal mixture of lognormal distributions is actually present due to seasonal variation or due to the combination of two or more differently exposed groups into a single group. Therefore, the sample size formulae to be presented are appropriate in those instances where the investigator is comfortable with the assumptions inherent in the use of the exposure group concept. The reader should consult Rappaport[6] for a review of the exposure group concept.

The Lognormal Distribution

The lognormal distribution has been described as the underlying limiting distribution of workplace exposures.[2] A lognormal distribution (with zero lefthand boundary) is fully described by two parameters: the geometric mean (GM) and the geometric standard deviation (GSD). Estimates of the true GM and GSD of a lognormal distribution of exposures can be calculated using either historical exposure data or data from a pilot study:

$$GM = \exp\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] = \exp(\bar{y}_l) \quad (1)$$

$$GSD = \exp(s_l) = \exp\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}_l)^2} \quad (2)$$

where y_i is the natural log of the i^{th} observation ($y_i = \ln x_i$), n is the number of available measurements, $\bar{y}$ is the mean of the log transformed data, and s_l is the standard deviation of the log transformed data.

There are several ways to estimate the true mean, μ, and variance, σ^2, of a lognormally distributed variable from a sample of size n. The simplest, and most obvious, is to calculate the simple arithmetic mean and variance of the x_i. The arithmetic mean and variance are unbiased but not minimum variance estimators for lognormally distributed data. A minimum variance unbiased estimator of μ can be calculated using a formula proposed by Finney.[7,8]

$$MLE_c = \exp(\bar{y})\psi\left(\frac{s_l^2}{2}\right) \quad (3)$$

where MLE_c is the corrected maximum likelihood estimate of the mean of a lognormal distribution, $\bar{y}$ and s_l^2 are the mean and variance of the log transformed data. The ψ function was defined for any argument g as:

$$\psi(g) = \left[1 + \frac{(n-1)g}{n} + \frac{(n-1)^3 g^2}{n^2(n+1)2!} + \frac{(n-1)^5 g^3}{n^3(n+1)(n+3)3!} + \cdots\right]. \quad (4)$$

Equation 4 is easily calculated using a programmable calculator or personal computer. Calculation to at least five terms is accurate to the third decimal place. A minimum variance unbiased estimate of the variance, σ^2, can be similarly calculated:[7,8]

$$s_c^2 = \exp(2\bar{y})\left[\psi(2s_l^2) - \psi\left(\frac{n-2}{n-1}s_l^2\right)\right] \quad (5)$$

where the s_c^2 is the corrected maximum likelihood estimate of the sample variance.

PROPOSED SAMPLE SIZE FORMULAE

For those substances that produce health effects following long-term exposures, the usual goal of the epidemiologist/industrial hygienist is to precisely determine the true mean exposure for all workers in an exposure group over some observation interval, usually each year of the study. The approximate sample size for estimating the true arithmetic mean exposure (μ) can be calculated using the following formula:

$$n_\mu \cong \frac{t^2_{\alpha/2,n-1}s_c^2}{(f \cdot MLE_c)^2} \quad (6)$$

where n_μ is the approximate number of exposure measurements, t is the t-value for a $1-\alpha$ confidence level and $n-1$ degrees of freedom, and f is a fraction of the estimated true mean (Equation 3) and represents the desired accuracy in the estimate. Both the MLE_c and s_c are calculated from the best available data for each exposure group: compliance data, company records, and/or a pilot study. Note that the n used to calculate degrees of freedom for the appropriate t-value (and hereafter called n_{pilot}) comes from the prior exposure data and should not be confused with n_μ calculated by Equation 6.

This formula is a slightly modified version of the standard sample size formula for estimating the true mean of a distribution (see Appendix, Equation A1).[9-11] Equation 6 can be used to estimate the number of measurements required for a certain level

of confidence that a future sample mean, in this case the mean yearly exposure, will be within plus or minus a specified percentage of the true mean. For example, if it is desired to estimate μ and be 95% confident that the estimate will be within ±20% of μ, f will be 0.20, and the t-value will be $t_{0.05/2}$ with n−1 degrees of freedom. The corrected maximum likelihood estimate of the mean (MLE_c) and the corrected maximum likelihood estimate of the variance (s_c^2) are the preferred estimates of the true mean and variance when n_{pilot} is small (20 or less).[12] However, the simple arithmetic mean and variance can be substituted, particularly when it is suspected that the distribution is not lognormal.

The t-value in Equation 6 is most appropriate for normally distributed variables. The variable of interest here is the arithmetic mean of a lognormal distribution. The Central Limit Theorem holds that, with few exceptions, the distribution of sample means of size n is asymptotically normal as n increases, without regard to the shape of the underlying distribution.[13] As will be shown later, it requires GSDs greater than 3 and small n_{pilot} before the sample size calculated by Equation 6 fails to provide the desired accuracy in the estimate.

There are several instances where the parameter to be estimated is the geometric mean. Seixas et al.[14] pointed out that there are types of statistical analyses where the geometric mean may be preferred over the arithmetic mean. Rappaport et al.[15] recommended that the parameters of individual exposure distributions (GM and GSD) be estimated in order to assess the influence of individual work practices on exposures. In the case where the GM is of interest, the following equation can be used to estimate an approximate sample size:

$$n_{GM} \cong \frac{t^2_{\alpha/2,n-1}(\ln GSD)^2}{\left[\frac{1}{2}\ln\left(\frac{1+f}{1-f}\right)\right]^2}. \quad (7)$$

Equation 7 is also a modified version of the standard sample size formula when used to calculate the sample size for estimating the log transformed GM (see Appendix for the derivation and a simpler version useful when f < 0.3). Since we do not normally think in terms of log transformed variables, the author modified the equation so that the variables entered are readily interpreted. The GSD is estimated from pilot study data using Equation 2 and the accuracy (f) is specified by the investigator. Equation 7 can be used to approximate the number of measurements required for a certain level of confidence that a future sample GM will be within plus or minus a specified percentage of the true GM.

The reader should note that an efficient two-stage sampling scheme can be considered in those instances where it can be assumed that conditions have not changed since the initial sample was collected.[9-11] The investigator collects an initial sample of size n_{pilot}, calculates n_μ or n_{GM} (Equations 6 and 7), but collects only $n_\mu - n_{pilot}$ or $n_{GM} - n_{pilot}$ additional exposure measurements, which are then combined with the initial sample.

METHODOLOGY

A computer simulation was used to test the two sample-size formulae. "Computer simulation" refers to the generation of artificial data using a specified set of distribution parameters and manipulating and analyzing the data to test some hypotheses that cannot be addressed by direct means. The goals for this simulation were to (1) show that Equation 6 is adequate for estimating the arithmetic mean for a variety of lognormal distributions, and (2) verify that Equation 7, despite its unfamiliar appearance, is appropriate for estimating the geometric mean.

The t-distribution was used in Equations 6 and 7, since it is anticipated that limited prior exposure data will be available, likely resulting in inaccurate estimates of the true mean and variance or the true GM and GSD. To test and verify Equations 6 and 7, the following scenario was postulated. A limited number of exposure data are available for a single exposure group. These data will be used to provide initial estimates of the true exposure parameters for this group (Equations 2, 3, and 5), which in turn will be used to calculate the approximate sample size for this group during the first year of a prospective study. The goal of the industrial hygienist is to estimate the true exposure parameters for this group to within some reasonable degree of accuracy, which will be ±20% (f = 0.2) at a 95% level of confidence ($(1-\alpha)100\%$) in the simulation to follow.

To simulate this scenario, artificial datasets were created using a personal computer. Pilot study datasets of sizes n_{pilot} = 5, 10, 20, 50 were drawn from lognormal distributions having a true GM of 10 and true GSDs of 1.5, 2, 3, and 4. Equations 2, 3, and 5 were used to calculate initial estimates of the true parameters of the lognormal distribution from the pilot study datasets. These estimates were then used to calculate sample sizes for the main study: n_μ and n_{GM} (Equations 6 and 7). Artificial datasets of size n_μ and n_{GM} were then simulated, and the MLE_c and GM were calculated, respectively. This process of generating an artificial pilot study, calculating n_μ and n_{GM}, generating artificial datasets of size n_μ and n_{GM}, and calculating MLE_c and GM, respectively, was repeated 5000 times. The statistic of interest was the observed confidence level, or the fraction of the 5000 artificial datasets that had MLE_cs or GMs within ±20% of the true mean or GM, respectively. Equation 4 was evaluated to eight terms in all calculations.

RESULTS AND DISCUSSION

The results of the simulations are presented in Tables I and II. Table I indicates that when f = 0.20, the estimated confidence levels for Equation 6 approach the target level of 0.95 for most combinations of GSD and n_{pilot}. The exceptions were for large GSDs (≥3) and small pilot study samples sizes (<20), indicating that the distributions of MLE_c values departed significantly from the normal distribution. This simulation was repeated for other accuracy values (f = 0.05 and f = 0.50) with similar results (Table I).

Exact, but asymmetrical, confidence intervals can be calculated for means of samples of specific size drawn from a lognormal distribution,[16,17] but such calculations are cumbersome and do not provide an easy route to estimating the "correct" sample size. In reality, the process of sample size estimation is seldom exact. In most cases the number of prior measurements is limited, and the goal is simply to calculate an approximate sample size. In this context, Equation 6 can be considered appropriate for the range of GSDs and sample sizes evaluated in

TABLE I. Estimates of True Confidence Levels When using Equation 6 for Various Combinations of f, n_{pilot}, and GSD[A]

		Pilot Study Sample Size			
f	*GSD*	*5*	*10*	*20*	*50*
0.05	1.5	0.9408	0.9416	0.9506	0.9498
	2	0.9314	0.9348	0.9466	0.9524
	3	0.8992	0.9156	0.9320	0.9540
	4	0.8780	0.8942	0.9272	0.9612
0.20	1.5	0.9386	0.9450	0.9468	0.9424
	2	0.9422	0.9434	0.9450	0.9486
	3	0.9012	0.9180	0.9290	0.9514
	4	0.8658	0.8936	0.9288	0.9570
0.50	1.5	0.9574	0.9488	0.9496	0.9554
	2	0.9464	0.9452	0.9466	0.9566
	3	0.9264	0.9220	0.9420	0.9520
	4	0.9038	0.9162	0.9354	0.9564

[A] The target confidence level was set at 0.95 (α = 0.05, target confidence level = 1 − α). Each number represents the fraction of 5000 simulated "main study" means that fell within ±f(100%) of the true mean of a lognormal distribution having the indicated GSD.

Table I. The loss of confidence was not excessive even for large GSDs and small n_{pilot}.

The derivation of Equation 7 (see Appendix) involved the conversion of a symmetrical confidence interval about the true GM from the linear scale to the log scale. The simulation results in Table II indicate that the true confidence levels for Equation 7 also approach the target value of 0.95 for f = 0.20 and the various combinations of GSD and pilot study sample size. The simulation was repeated for other f values (Table II). The confidence levels for f = 0.05 were close to the target value of 0.95 for all combinations of GSD and pilot study sample size. The confidence levels for f = 0.50 were lower than 0.95 in all cases, but not excessively lower. Interestingly, the confidence levels decreased with increasing pilot study sample size. (This counterintuitive result was traced to a restriction in the simulation program specifing that the main study sample size be set to one in the event that the calculated sample size was less than one.)

In conclusion, if it is desired to estimate the true arithmetic yearly exposure for an exposure group, Equation 6 can be used, with some caution when n_{pilot} is small and the estimated GSD is large, to estimate an appropriate sample size. If the true geometric mean is of interest, Equation 7 can be used, with some caution when the desired accuracy is low (f is large). (Note that either equation could also be used to calculate appropriate sample sizes for estimating the mean or geometric mean exposure of an individual worker.)

For example, consider a prospective exposure-response study of workers exposed to welding fumes (the data are real, the scenario is fictitious). Using 17 measurements from a pilot study, the MLE_c, s_c, and GSD for one exposure group were estimated to be 10.2 mg/m^3, 4.6 mg/m^3, and 1.55 (see Equations 3, 5, and 2, respectively). Equation 6 can be used to calculate the number of measurements needed to estimate the mean exposure experienced by this exposure group during the first observation interval, e.g., first six months of the study, to within plus or minus 25% at a 95% confidence level. Use of the above parameter estimates, the appropriate t-value ($t_{0.05/2,17-1} = 2.120$), and Equation 6 suggests that 15 measurements are required in order to be 95% confident that the estimated mean exposure for the group will be within 25% of the true mean. Another exposure group at the same plant appeared to experience lower but more variable exposures. The MLE_c, s_c, and GSD for this group were estimated, using 18 measurements collected during the pilot study, to be 2.6 mg/m^3, 2.2 mg/m^3, and 2.16, respectively. Use of these parameter estimates, the appropriate t-value, ($t_{0.05/2,18-1} = 2.110$), and Equation 6 suggests that 51 measurements are required to estimate the mean exposure for the second exposure group to the same degree of accuracy and confidence level.

Equation 7 should be used if the GM is the parameter of interest. Only the estimated GSD and the appropriate t-value are required for this calculation. Using the appropriate GSD and t-values in Equation 7 results in approximate sample sizes of 13 and 40 for the first and second exposure groups, respectively.

The actual level of accuracy achieved will be determined not only by sample size, but also by how well the distribution of exposures fits the lognormal model, and how the measurements are collected. Random sampling of workers is advisable in order to assess the homogeneity of the group.[6,18] Random sampling by days is also desirable[18] to reduce the effect of autocorrelation of exposures, but may not be practical given limited resources. Replicate sampling of workers may be necessary to compare group exposure variability with individual exposure variability.[6] Departures from the exposure group paradigm may not be known until additional data is collected, at which time more appropriate worker/job combinations can be devised.

Up to this point the focus has been on the validity of Equations 6 and 7. The author has purposely avoided discussing the actual sample sizes predicted by these formulae. Table III lists sample sizes for different pilot study sizes, GSDs, and several reasonable values of f. For the sake of illustration the author

TABLE II. Estimates of True Confidence Levels When using Equation 7 for Various Combinations of f, n_{pilot}, and GSD[A]

		Pilot Study Sample Size			
f	*GSD*	*5*	*10*	*20*	*50*
0.05	1.5	0.9476	0.9494	0.9488	0.9478
	2	0.9462	0.9500	0.9516	0.9506
	3	0.9520	0.9494	0.9432	0.9510
	4	0.9466	0.9444	0.9550	0.9490
0.20	1.5	0.9512	0.9398	0.9492	0.9474
	2	0.9494	0.9450	0.9462	0.9444
	3	0.9444	0.9440	0.9474	0.9392
	4	0.9452	0.9490	0.9464	0.9458
0.50	1.5	0.9380	0.9262	0.9256	0.9182
	2	0.9336	0.9294	0.9190	0.9230
	3	0.9412	0.9276	0.9156	0.9148
	4	0.9282	0.9240	0.9212	0.9234

[A] The target confidence level was set at 0.95 (α = 0.05, target confidence level = 1 − α). Each number represents the fraction of 5000 simulated "main study" geometric means that fell within ±f(100%) of the true geometric mean of a lognormal distribution having the indicated GSD.

TABLE III. Comparison of Sample Sizes Necessary for Estimating Arithmetic and Geometric Mean of a Lognormally Distributed Variable[A]

			$f = 0.20$		$f = 0.30$		$f = 0.50$	
n_{pilot}	True GSD	$t_{0.05/2,n-1}$	n_μ	n_{GM}	n_μ	n_{GM}	n_μ	n_{GM}
5	1.5	2.776	34	31	15	13	6	4
	2		119	90	53	39	19	12
	3		452	226	201	97	72	31
	4		1124	360	500	154	180	49
10	1.5	2.262	23	20	10	9	4	3
	2		79	60	35	26	13	8
	3		300	150	133	64	48	20
	4		746	239	332	102	119	33
20	1.5	2.093	20	18	9	8	3	2
	2		68	51	30	22	11	7
	3		257	129	114	55	41	18
	4		639	205	284	88	102	28
50	1.5	2.010	18	16	8	7	3	2
	2		62	47	28	20	10	6
	3		234	119	105	51	38	16
	4		589	189	261	81	94	26
⩾50	1.5	1.96	17	15	8	7	3	2
	2		59	45	26	19	9	6
	3		225	113	100	48	36	15
	4		560	180	249	77	90	24

[A] Sample sizes calculated for various combinations of pilot study sample size (n_{pilot}), estimated group GSD, and desired accuracy (f)

assumed, for all calculations of n_μ, that the pilot study estimates of the mean and variance exactly equaled the true mean and variance,* and that for all calculations of n_{GM} that the GSD estimated from the pilot study exactly equaled the true GSD. In reality, these estimates will vary around the true mean and variance and the true GSD. As expected, as n_{pilot} increased, n_μ and n_{GM} decreased (Table III). This is in response to the decrease in $t_{\alpha/2,n-1}$. As the pilot study GSD increased, the sample size increased. The troubling aspect of these numbers is their magnitude, particularly for small n_{pilot} and large GSD. The total sample size, summed across all exposure groups that comprise the study, could easily become unrealistic. On a positive note, progressively fewer measurements will be required each year of a prospective study as the exposure database becomes larger, resulting in more stable estimates of the true, underlying distributions for each group.** Even so, when n_{pilot} is large, much greater than 50, and the t-value approaches the z-value for $\alpha/2$ (bottom of Table III), the sample sizes are still quite large for GSDs greater than 3.

Consideration of the numbers in Table III led to the following observations:

(1) Accurate estimation of the true arithmetic mean or geometric mean of a single exposure group can require a sizable commitment in sampling resources when n_{pilot} is small and/or the estimated GSD is large. This underscores the value of acquiring reasonably accurate exposure data during the pilot study phase of a prospective study. Pilot study sample sizes beyond 20 do not greatly reduce n_μ or n_{GM}, which suggests that as a general rule n_{pilot} should approximately equal 20. In addition, attention should be given to devising exposure groups so that the GSD for each group is as low as possible.

(2) The sample sizes are greatly reduced when the required accuracy is reduced (f is increased). The total sample burden can be reduced by relaxing expectations. However, the resulting inaccurate estimates of the group arithmetic mean or geometric mean exposures may reduce the precision of any subsequent exposure-response analyses.

(3) The sample sizes required to obtain an accuracy on the order of many medical measurements ($f < 0.20$) can be quite large. This leads to the general statement that the accuracy of the estimation process for long-term exposures for chronic disease agents, as applied to each individual, is unlikely to ever approach the accuracy of most medical measurements. Limited resources place limits on the accuracy of our estimates of individual worker mean exposures. The sample sizes predicted by Equations 6 and 7 suggest that the accuracy in any study that collects a small number of exposure measurements per exposure group is most likely poor. The only consolation, albeit hollow, is that the medical measurements are not necessarily specific to the disease process and may also miss the target.

In the situation where, due to limited resources, n is fixed, Equations 6 and 7 can be rearranged to estimate the accuracy for any given n:

* The following formulae were used to calculate the mean and variance of a lognormally distributed variable:[3,6]

$$\mu = GM \cdot \exp(0.5 \ln^2 GSD);$$

$$\sigma^2 = GM^2 \cdot [\exp(\ln^2 GSD)(\exp(\ln^2 GSD) - 1)]$$

** The reader may wonder why, if stable (large sample) estimates of the parameters are available, sampling is continued. The reason is that the mean exposure for each group may change from year to year in response to changes in production, ageing or replacement of equipment, changes in environmental controls, and so on.

$$f_\mu = \frac{t_{\alpha/2,n-1} s_c}{\sqrt{n}\ MLE_c}$$

$$f_{GM} = \frac{k-1}{k+1}, \quad \text{where } k = \exp\left[\frac{2t_{\alpha/2,n-1}\ \ln(GSD)}{\sqrt{n}}\right].$$

REFERENCES

1. **Hill, A.B.:** The environment and disease: association or causation. *Proceedings Royal Soc. Med. 58*:295–300 (1956).
2. **Esmen, N.A. and Y.Y. Hammad:** Log-normality of environmental sampling data. *J. Environ. Sci. Health A12(1&2)*:29–41 (1977).
3. **Leidel, N.A., K.A. Busch, and J.R. Lynch:** *Occupational Exposure Sampling Strategy Manual.* Cincinnati, OH: National Institute for Occupational Safety and Health, 1977.
4. **Corn, M. and N.A. Esmen:** Workplace exposure zones for classification of employee expousres to physical and chemical agents. *Am. Ind. Hyg. Assoc. J. 40*:47–57 (1979).
5. **Gamble, J. and R. Spirtas:** Job classification and utilization of complete work histories in occupational epistomology. *J. Occup. Med. 18*:399–404 (1976).
6. **Rappaport, S.M.:** Assessment of long-term exposures to toxic substances in air. *Ann. Occup. Hyg. 35*:61–121 (1991).
7. **Finney, D.J.:** On the distribution of a variate whose logarithm is normally distributed. *J. Royal Stat. Soc. 7(suppl)*:155–161 (1941).
8. **Aitchison, J. and J.A.C. Brown:** *The Lognormal Distribution.* Cambridge: Cambridge University Press, 1957.
9. **Stein, C.:**A two-sample test for a linear hypothesis whose power is independent of the variance. *Ann. Math. Stat. 16*:243–258 (1945).
10. **Seelbinder, B.M.:** On Stein's two-stage sampling scheme. *Ann Math. Stat. 24*:640–649 (1953).
11. **Steel, R.G.D. and J.H. Torrie:** *Principles and Procedures in Statistics,* 2nd ed. New York: McGraw-Hill Book Company, 1980. pp. 119–120.
12. **Attfield, M.D. and P. Hewett:** Exact expressions for the bias and variance of estimators of the mean of a lognormal distribution. *Am. Ind. Hyg. Assoc. J. 53*:432–435 (1992).
13. **Dowdy, S. and S. Wearden:** *Statistics for Research.* New York: John Wiley & Sons, 1983.
14. **Seixas, N.S., T.G. Robins, and L.H. Moulton:** The use of geometric and arithmetic mean exposures in occupational epidemiology. *Am. J. Ind. Med. 14*:465–477 (1988).
15. **Rappaport, S.M., R.C. Spear, and S. Selvin:** The influence of exposure variability on dose-response relationships. *Ann. Occup. Hyg. 32*:529–537 (1988). Suppl. 1, *Inhaled Particles VI.*
16. **Armstrong, B.:** Letter to the editor. *Am. Ind. Hyg. Assoc. J. 52*:A468 (1991).
17. **Land, C.E.:** Tables of confidence limits for linear functions of the normal mean and variance. In *Selected Tables in Mathematical Statistics,* vol. III. H.L. Harter and D.B. Owen, eds. Providence, RI: American Mathematical Society, 1975. pp. 385–419.
18. **Leidel, N.A. and K.A. Busch:** Statistical design and data analysis requirements. In *Patty's Industrial Hygiene and Toxicology: Theory and Rationale of Industrial Hygiene Practice,* vol. IIIA, L.J. Cralley and L.V. Cralley, eds. New York: Wiley Interscience, 1985. p. 432.

APPENDIX

Derivation of the Sample Size Formula for Calculating n_{GM}

The standard sample size formula for estimating the mean of a normally distributed variable is:

$$n = \frac{Z^2_{\alpha/2}\ \sigma^2}{\delta^2} \qquad \textbf{(A1)}$$

where σ is a large sample estimate of the true arithmetic variance (most statistics textbooks state that $n > 30$ is sufficient), and δ is the tolerated error. The goal, when using Equation A1, is to be $(1 - \alpha)$ 100% confident that the future sample mean will fall within $\pm\delta$ of the true mean. The tolerated error, δ, can be considered to be a fraction of the true mean, μ, such that $\delta^2 = (\bar{x} - \mu)^2 = (\mu - \bar{x})^2 = (\pm f\mu)^2$, where f is a fraction of μ and $\bar{x}$ is the future mean of the yet uncollected study data. Since μ and σ^2 are rarely known, they are estimated by $\bar{x}$ and s^2 derived from existing data, and the Z-value in Equation A1 is replaced by a t-value with $n-1$ degrees of freedom.[9–11]

It is assumed here that the goal, when estimating the true GM, is to be able to observe a sample GM within some symmetrical interval about the true GM, for example ±20%. However, this symmetrical interval on the linear scale is not symmetrical on the log scale. For example, consider the problem where the estimate should fall between GM_1 and GM_2, where $GM_1 = \mu_g(1 - f)$ and $GM_2 = \mu_g(1 + f)$, where μ_g is the true GM. The quantity δ on the log scale will not be equal for the two alternatives:

$$\delta_1 = \ln \mu_g - \ln GM_1 = \ln \frac{\mu_g}{GM_1}$$

$$= \ln\left(\frac{\mu_g}{\mu_g(1-f)}\right) = -\ln(1-f)$$

$$\delta_2 = \ln GM_2 - \ln \mu_g = \ln\left(\frac{GM_2}{\mu_g}\right)$$

$$= \ln\left(\frac{\mu_g(1+f)}{\mu_g}\right) = \ln(1+f).$$

An average difference (δ_{ave}) was defined to convert a symmetrical interval about the true GM from the linear scale to the log scale:

$$\delta_{ave} = \frac{1}{2}(\delta_1 + \delta_2) = \frac{1}{2}\left[\ln\left(\frac{1+f}{1-f}\right)\right].$$

By definition the log transformed data from a lognormal distribution are normally distributed. Thus, it was assumed that the sample size formula for estimating the log transformed geometric mean is similar to Equation A1, with $t_{\alpha/2,n-1}$ replacing $Z_{\alpha/2}$ and $(\ln GSD)^2$ as the estimated variance, where the GSD is estimated from prior data:

$$n_{GM} \cong \frac{t^2_{\alpha/2,n-1}(\ln GSD)^2}{\delta^2}.$$

Substituting δ_{ave} for δ results in the final formula for n_{GM}:

$$n_{GM} \cong \frac{t^2_{\alpha/2,n-1}(\ln GSD)^2}{\left[\frac{1}{2}\ln\left(\frac{1+f}{1-f}\right)\right]^2}.$$

This formula is expected to yield an approximate sample size. Note that the following simpler formula yields nearly identical sample sizes when $f < 0.3$:

$$n_{GM} \cong \frac{t^2_{\alpha/2,n-1}(\ln GSD)^2}{f^2}.$$