Fitting Lognormal Distributions to Data from a Size-Sorting Particle Device

Stanley Shulman, Paul Baron, Dawn Ramsey
National Institute for Occupational Safety and Health
Cincinnati, OH 45226

Keywords: nonlinear optimization

1) Abstract

The cascade impactor is a device for sampling airborne particles to determine the particle mass-weighted aerodynamic–diameter distribution. The particles enter at the top of the device and deposit by decreasing particle diameter on one of the multiple stages or on the final filter. Past experimental data has established the fraction of each particle size expected to deposit on each stage. From the measured masses, estimates of the parameters of the lognormal distribution (either a single distribution or a mixture of two) must be obtained. Procedures based on weighted least squares are evaluated for fitting one or two lognormal distributions to both actual and simulated data. Several different scenarios are considered: 1)geometric means and standard deviations producing one or two modes; 2) distributions with substantial mass either in small, medium, or large particle size stages. For mixtures of two lognormal distributions the weighted least squares solutions can produce imprecise estimates of the lognormal parameters.

2) Introduction

An important characteristic of air sampled particles is their size distribution, which can have important effects on their health consequences. The cascade impactor is a device whereby sampled particles can be separated into multiple samples (nine, in the example discussed here) by particle diameter. The mass of each stage can be individually determined. Particles in an air stream enter at the top of the device, and pass through a series of nozzles, each of which causes the air passing through it to impinge on a flat collection plate. The air stream from each nozzle makes a right angle bend at the surface of each corresponding collection plate. Particles in the air stream are deflected from the air stream onto the collection plate when their inertia is too large to follow the right angle bend. The stages are organized so that the largest nozzle is in the first stage and decreasing sized nozzles are in the subsequent stages. Thus, the largest particles are deposited on the first stage and particles with decreasing sizes are deposited on subsequent stages. After the final stage, the remaining particles passing through all the stages are collected on a filter. The mass of all the particles on a particular stage is weighed and can be related to the size of particles expected to deposit on that stage. Historically, aerodynamic diameter has been assumed to have a lognormal distribution. Thus, the problem is how to obtain estimates of the geometric mean and standard deviation of the lognormal distribution from the nine masses that deposit on the impactor stages. Also, since the distribution of these nine masses may be a mixture of two lognormal distributions, it is important to be able to estimate a mixture of two lognormal distributions—thus, five parameters [1].

The quality of the estimates produced by weighted least squares will be investigated in this paper, both from simulated data and from sample data collected in a laboratory.

The data types to be studied are the following:
a) Sampled data: 2 modes, 2 lognormal components, a mixture of components with geometric means in small and large particle sizes
b) Simulated data: 1 mode, 1 lognormal component, geometric means in medium or large particle sizes
c) Simulated data: 1 mode, 2 lognormal components, a mixture of components with geometric means in small and large particle sizes

3) Statistical Models

The data were obtained from a set of stages in a Model 290 personal cascade impactor (Thermo Electron Corp.), for which there are nine stages. It is thought that the deposition of particles in a cascade impactor is Poisson deposition, each particle acting independently. However, with the huge number of particles collected, this is a small component of the total variability. Most of the variability will be due to sampling or weighing. By theory there are available estimates of the average deposition on each stage of the device, as a function of particle size. However, the actual collection efficiency of each stage is typically obtained from published data that were fitted to a hyperbolic function. The actual deposition will vary with each run of the device. Also, the plates on which the collected particles deposit are weighed, and are subjected to any errors associated with that process.

In a generation chamber, mass is generated at some rate, say r(t)dt (mass/time), in the time interval (t, t+ dt). The mass is thought to be distributed across size distributions (diameter) according to a lognormal distribution (f(t)). Thus, for a period of time T, the total generated mass is :

$M`(T)=\int_{[0,T]} r(t)\, dt,$

where the limits of integration are 0 and T, and the distribution in any infinitesimal size range, (d, d+dd) is:

$dM`(t,d)= r(t)f(d)\, dd,$

where the f(d) is the probability distribution of mass as a function of diameter, assumed here to be lognormal. If the rate is constant, then the expectation of the total mass generated may be expressed as the sum of the mass (M`(T)) generated in the diameter size intervals [d(i-1),d(i)], i=2,…I, :

$M`(T)=Tr=Tr\int f(d)dd=Tr\Sigma_i\int_{[d(i-1),d(i)]}f(d)dd=\Sigma_i M`_i(T)$

(For the nine stage impactor studied here, i=1,2,..,10, and d_1=0 and $d_{10}=\infty$.) The amount collected on the impactor stages is affected by loss and deposition functions, combined into {K(i,d)}, functions of the I stages of the impactor. The mass from the sampled atmosphere during the time interval [0,T] that deposits on the ith stage is given by:

M_i= Tr $\int$K(i,d) f(d) dd, where the integration is over all diameters, since any of the $M`_i(T)$s may contribute to $M_i(t)$.

The above are just average values, and ignore error. {K(i,d)} are determined experimentally, but the parameterization of the f(d) must be determined by statistical fitting. Let m_i denote the measured mass on stage i. A model for m_i is:

$m_i=Tr\int K(i,d)\ f(d)\, dd + g_i + e_i,$ (1)

where it is assumed that the e_is are (probably) normally distributed weighing error, the g_is combine the sampling error associated with both K(i,d) and f(d) and possible lack of fit of the lognormal distribution. The g_is may combine both normal and lognormal components, as would the total error terms, {$g_i + e_i$}. In this work the {$g_i + e_i$} are treated as statistically independent. This assumption makes sense for the weighing errors, but not necessarily for the sampling errors {g_i}. Also, population parameters will be denoted by names with upper case letters, and estimates of these parameters will be denoted by changing the upper case letters to lower case- for example, M_i and m_i defined above.

Least squares minimization of Σ_i {m_i -Tr $\int$ K(i,d) f(d) dd }^{2} is a means of estimation of the parameters of the lognormal distribution f(d). However, since there is loss of material during sampling (included in the K(i,d) terms) and since the estimation of the total mass Tr is of little interest, the criterion is altered to select parameters that minimize

$\Sigma_i\{\ m_i / \Sigma_j m_j\ -\ Tr\int K(i,d)\ f(d)\ dd/\ [Tr\ \Sigma_j \int K(j,d)\ f(d)\ dd]\ \}^2$

$=\Sigma_i\{fr_i-\int K(i,d)f(d)(dd)/[\Sigma_j\int K(j,d)f(d)dd]\}^2$ (2)

where $m_i / \Sigma_j m_j = fr_i$, the fraction of total mass on the ith stage. Equation (2) is the minimization of the difference between the sample frequency distribution and the theoretical frequency distribution for the I stages.

Since the fr_i do not have equal variances, weighted least squares is appropriate:

The minimization (2) becomes the minimization:

$\Sigma_i\{fr_i-\int K(i,d)f(d)(dd)/[\Sigma_j\int K(j,d)f(d)dd]\}^2/Var(fr_i)$ (3)

A justification for the use of weighted least squares in equations (2) and (3) is as follows. Suppose that the the total mass is treated as given. Then all the variability in the fr_i is in the numerator, m_i. If the errors {$g_i + e_i$ } are normally distributed, then the {fr_i} are normally distributed and weighted least squares is appropriate. Even when the total mass is treated as random, since it is the same for each fr_i, and since it is so large compared to each m_i value, the {fr_i} will be approximately normal if the m_i values are. Analysis of an example data set suggests that the m_i are approximately normally distributed. (See section (7).) If so, then the weighted least squares approach of (1) and (2) seems appropriate.

The solution may be obtained iteratively, by approximating the integral by a sum, and substituting different values for the unknown parameters, continuing until convergence is obtained. The actual problem that is solved is:

$$\Sigma_i \{fr_i - \Sigma_j K(i,d_j) \int_{[d(j-1),d(j)]} f(d)dd / [\Sigma_k \Sigma_j K(k,d_j) \int_{[d(j-1),d(j)]} f(d)dd]\}^2 / Var(fr_i) \quad (4)$$

where the index j is from 1 to N, for an N value large enough to provide a good approximation to the integral. Other approximations could have been used, but the intervals are very small, and the actual approximation should make little significant difference. To verify this, both the approximation of the integral used in reference [1] and the midpoint rule for approximating an integral [2] also were used, and results differed little from those presented.

If f(d) is unimodal then

$$f(d)=f_1(d)= 1/[(2\pi)^{0.5}\ln(GS_1)(d)] \exp\{-0.5[(\ln(d)-\ln(GM_1)]^2/[\ln(GS_1)]^2\}, \quad (5)$$

where GM_1 and GS_1 are, respectively, the geometric mean and standard deviation of the lognormal distribution. For a mixture of two lognormal distributions,

$$f(d)=Kf_1(d)+(1-K)f_2(d), \quad (6)$$

for K between 0 and 1, and $f_2(d)$ parameterized like $f_1(d)$, but with GM_2 and GS_2 as geometric mean and standard deviation.

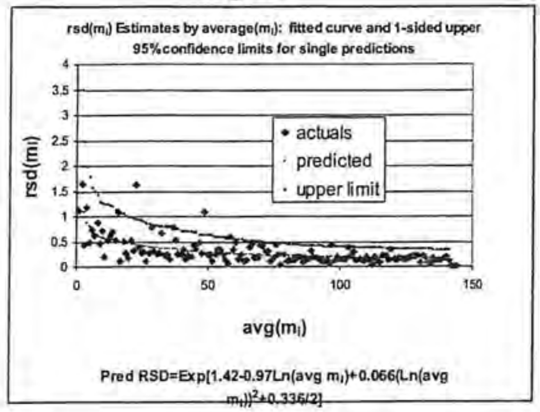
4) Form of Variance

Models considered for the variance of fr_i are:

1) Var(a): $Var(fr_i)=1$

2) Var(b): $Ln[rsd(m_i)]=\alpha + \beta\, Ln[avg(m_i)] + \gamma\{Ln[avg(m_i)]\}^2$. RSDs for m_i are converted to Var(fr_i) by Taylor series linearization. Rsds (rsd=std(m_i)/avg(m_i), for ratio of standard deviation to average of m_is) were modeled from historical data (16 data sets, with between 4 and 9 impingers; each of the 16 data sets yielded nine (mean,rsd) pairs or 144=16*9 data pairs). The fitted curve is shown in Figure 1. (The Taylor series linearization produces the result $RSD^2(m_i/\Sigma_j m_j) \approx RSD^2(m_i)+RSD^2(\Sigma_j m_j) -2Covariance(m_i,\Sigma_j m_j)/[E(m_i)E(\Sigma_j m_j)]$, where E denotes the statistical expectation.) Whereas the average(m_i) in Figure 1 are based on averages of between 4 and 9 impingers, the variances required in eq. (4) will use the fitted curve predictions, as if they were for m_i values for single samplers.

Figure 1



5) Determining Minimizing Parameter Values of Eq. (4)

The functions that define {K(i,d)} are complicated [3], and it is difficult to establish uniqueness of solutions of eq. (4).

$$K(i,d)=[1-B_1(d)] *B_2(d)\Pi_{[j=1 \text{ to } (i-1)]} \{[100-C_j(d)]/100\} [C_i(d)]/100],$$

where $B_1(d)=1/[1+(d/31.13)^{(-1.076-3.147d/31.13)}]$

$$B_2(d)= 1/[1+(d/28.92)^{(3.52-5.96(d/28.92)+9.78(d/28.92)(d/28.92)}]$$

$$C_i(d)=100[\tanh(d/a_i)^{b_i}]$$

for a_i and b_i dependent on the stage i=1,2…9. These factors have been experimentally determined.

For the results presented , estimated parameters are obtained for two different starting values for the optimization algorithm[4]. When the minima do not agree, estimates associated with the smaller minimum of eq(4) are used.

A problem that occurs is the presence of negative mass determinations. For the work done here, these determinations were reassigned the value 0.00001. There is also the problem to provide a variance for negative mass determinations to use in eq(4) for the weighted least squares estimation, for the Var(b) approach. A related problem was for the simulations, whether an upper limit should be placed on the RSD values used for the random error of eq(1). The RSDs were based on the predicted values shown in Figure 1. The decision was made, for masses less than 8 micrograms (a little less than FR_i=0.01 for total mass sampled of 1000 micrograms) to assign the RSD corresponding to 8 micrograms to that sample, an RSD of 0.87, based on the curve plotted in Figure 1. This same approach was used in the actual weighting in eq(4). Furthermore, for the model fitting, if the variance of fr_i under the Var(b) scheme was such that the largest variance was greater than 100 times the smallest variances, then the smallest variances were replaced by 0.01 times the largest variance.

In fitting the one-component and two-component models described in sections 6)-8), geometric standard deviation estimates were required to exceed 1.05. For the two-component models, the mixing parameter estimate k was forced to be between 0.01 and 0.99, and the smaller geometric mean was forced to be smaller than 7.

6) Simulation: One Lognormal Component with GM in Medium or Large Particle Sizes

The simulation studies were done for two different scenarios- either geometric mean of 8 or 12 μm. Results are shown in Table I. In these simulations the error term (g_i + e_i) of eq(1) was given a normal distribution with mean 0 and variance determined from the fitted RSD curve of Figure 1. The expected per cent distribution of mass is shown (=100 FR_i), and estimates are shown for the two weighting schemes shown in Section 4. Sixty simulations were done for each of the four (Scenario, Var) combinations in Table I, and all simulations converged to solutions.

Apparently outlying estimates were removed if the estimated geometric mean exceeded 40 (this occurred in one simulation), or if $\Sigma_j \int_{[d(j-1),d(j)]} f(d)dd] < 0.99$, since this sum should be close to 1. Twelve of the 13 outiers of the "<0.99" form occurred in (Scenario II, Var(a)). The effect of retaining more of these is to increase the estimated geometric mean.

Table I: One Lognormal Component Models
Normally Distributed Error (Using Fitted RSDs of Figure 1) Added to True Mass

	Scenario I. GM=8, GS=2
Stage, 100FR_i, without error	1 2 3 4 5 6 7 8 9 8.8, 12, 18.9,26.1, 22.1, 11.2, 0.8, 0.08, 0.01
	Var(a) Var(b)
N +avg gm(std,rsd) +avg gs(std, rsd) Correlation (gm,gs)	60 60 8.08(0.77,0.095) 8 .06(0.67,0.083) 2.03(0.153,0.075) 2.00(0.10,0.05) 0.53 0.58

	Scenario II GM=12, GS=2.5
Stage, 100FR_i, without error	1 2 3 4 5 6 7 8 9 14.3, 16.2,19.7, 21.5, 16.7, 9.9, 1.3, 0.3, 0.05
	Var(a) Var(b)
N* +avg gm(std,rsd) +avg gs(std, rsd) Correlation (gm,gs)	47 59 11.36(1.66,0.15) 11.49(1.69;0.15) 2.38(0.26, 0.11) 2.37(0.20,0.084) 0.79 0.76

*Numbers vary because of convergence problems under Scenario II.

+ avg gm and avg gs denote the arithmetic means of gm or gs estimates; std is standard deviation of estimates; rsd is relative standard deviation

In Scenario I averages for both gm & gs are close to the true values when either Var(a) and Var(b) is used; standard deviations are smaller when Var(b) is used. In Scenario II, both Var(a) and Var(b) result in underestimates of the true parameters, though, as was remarked above, that result for Var(a) weighting is not true if more simulations were retained.

7) Bimodal Sampled Data: Mixture of Two Lognormals

Figure 2

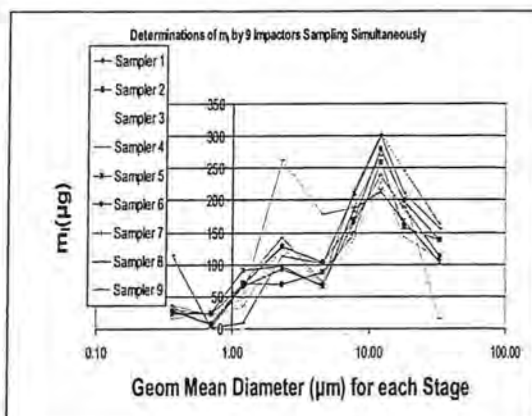


Table II

Results for the nine samplers, from Var(b) weighting:

	gm_1	gs_1	gm_2	gs_2	k
avg	12.25	1.46	2.59	1.82	0.74
std	0.85	0.22	0.97	0.62	0.15
rsd	0.070	0.15	0.37	0.34	0.20

Figure 2 and Table II present the results of a lab experiment carried out with nine impactors sampling simultaneously in a chamber. The experiment was performed to generate size distribution data to allow comparison with two simulated lognormal modes. A chamber was used in which dust was generated as airborne particles so that they could be collected simultaneously and equally by the nine impactors. The dust was prepared by mixing two commercial aluminum oxide powders (Micro-Abrasives Corp., Westfield MA) so that a clearly biomodal aerosol would be produced.

Figure 2 indicates two modes for most of the samplers. Table II indicates that , for the component with lesser weighting (smaller geometric mean), rsds are much larger. Results on rsds are similar for Var(a) weighting.

These nine sets of masses were also tested for normality, by means of a mixed model in which the stages were treated as fixed and the samplers as random. The residuals from the model appear to be approximately normally distributed. This indicates that for the simulations discussed in sections 6 and 8 it is not unreasonable to add normally distributed error to the expected values of the masses.

8) Simulations: Mixture of Two Lognormals Components with GMs in Small and Large Particle Sizes

For the estimates shown in Table III, gm_1 is the larger of the 2 geometric mean estimates. Initial values were obtained by adding normally distributed error corresponding to Figure 1 to the randomly generated starting values of 5 parameter estimates (GM_1, GS_1, GM_2,GS_2, K). . The expected per cent distribution of mass is shown (=100 FR_i), and estimates are shown for the two weighting schemes shown in Section 4 (Var(a) and Var(b)).

For each of the four (Scenario, Var) combinations, 120 simulations were produced. Some did not converge- no more than three for each of the four combinations. Some apparently outlying estimates were removed, for instance, if the larger geometric mean estimate exceeded 40 , or if either the geometric standard deviation estimate exceeded 100, or if $\Sigma_j\int_{[d(j-1),d(j)]}f(d)dd]<0.99$, since this sum should be close to 1. (Inclusion of simulations that do not exclude results that fail the "<0.99" criterion produces larger gms, gss, and rsds for the larger gm component under the Var(b) weighting.)

Table III

Two-Component Lognormal: $Kf_1 + (1\text{-}K) f_2$

	Scenario I GM_1=12,GS_1=2.5,GM_2=2.5,GS_2=2,K =0.5	
Stage, 100FR_i, without error	1 2 3 4 5 6 7 8 9 6.4, 7.4, 9.8, 14, 19.5, 29.1, 9.9, 3.2, 0.8	
	Var(a)	Var(b)
N*	98	103
+avg gm_1(std,rsd)	11.57(5.89,0.51)	11.27(5.55,0.49)
+avg gs_1(std,rsd)	1.94(0.82,0.42)	1.89(0.72,0.38)
+gm_2(std,rsd)	2.61(0.91,0.35)	2.48(0.92,0.37)
+gs_2(std, rsd)	1.76(0.48,0.27)	1.75(0.46, 0.26)
+k(std, rsd)	0.49(0.29, 0.59)	0.50(0.29,0.58)
	Scenario II GM_1=12,GS_1=2.5,GM_2=2.5,GS_2=2,K=0.75	
Stage, 100FR_i, without error	1 2 3 4 5 6 7 8 9 10, 11.4, 14.3, 17.4, 18.2 ,20.4, 6.0, 1.9, 0.5	
	Var(a)	Var(b)
N*	101	103
+avg gm_1(std,rsd)	12.58(4.32,0.34)	11.59(3.88,0.33)
+avg gs_1(std,rsd)	1.85(0.73,0.39)	2.10(0.71,0.34)
+gm_2(std,rsd)	3.25(1.69, 0.52)	2.73(1.68, 0.62)
+gs_2(std, rsd)	1.79(0.60,0.34)	1.72(0.58, 0.34)
+k(std, rsd)	0.57(0.31, 0.54)	0.67(0.29,0.43)

* Numbers vary because of convergence problems .

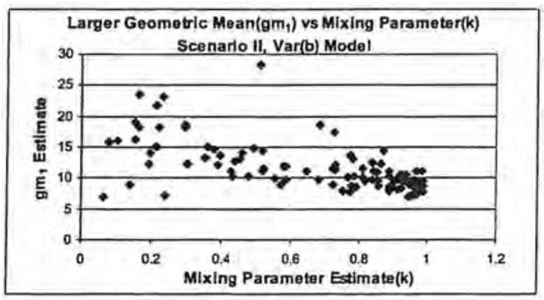
\+ avg gm or avg gs denotes arithmetic mean of gm or gs estimates; "std" is estimated standard deviation of estimates; "rsd" is estimated relative standard deviation

For both scenarios I and II, for both Var(a) and Var(b), estimated correlations of k with each of 4 other parameters exceed 0.6 in absolute value, as shown in Table IV. When there is high correlation between estimators, the model parameters can still fit the data well, but there is uncertainty in the estimator values. High correlation means that various combinations of parameters may similarly describe the same set of data. Figure 3 indicates that, as the mixing parameter decreases, the larger geometric mean increases. A decrease in the value of the mixing parameter means an increase in (1-k), the weight given to the second lognormal component in eq(6). As the weight given to that component increases, the value of the geometric mean and standard devation for that component also increases, in order to cover the part of the distribution that is no longer well accounted for by the component with larger geometric mean. Alternatively, it can be seen in Table III that the rsds of the estimators for both scenarios and weighting schemes all exceed 0.3. For the mixing parameter, the estimated rsds exceed 0.4.

Table IV
Correlation of Estimates for Scenario II (n=103), Var(b) Weighting

	gm_1	gs_1	gm_2	gs_2	K
gm_1	1				
gs_1	-0.48	1			
gm_2	0.57	-0.74	1		
gs_2	0.51	-0.84	0.88	1	
K	-0.61	0.91	-0.92	-0.92	1

Figure 3



9) Conclusions

Comparison of the one lognormal and two lognormal component simulation results indicates that the one lognormal component results have smaller rsds.

The problem, especially for the two-component models, is that each evaluation by the instrument produces just nine data values. The estimators of the five parameters in the two-component model are highly correlated, and have large relative standard deviations, and great care must be taken if the parameter estimates that correspond to the minimum of the weighted least squares objective function are to be attained.

This suggests that, where possible, the use of single impactor samples should be avoided, so that the precision of estimates can be improved by averaging the determinations of several instruments sampling simultaneously, or by using repeated measurements with a single impactor for processes with a temporally stable output. However, limited sampler replication will not remove the problem of large rsd of most parameter estimates in the 2-component model.

10) Acknowledgments

The authors appreciate helpful comments from NIOSH associates Mark Hoover, Steven Page, and Martin Petersen.

11) References

1) Ramachandran, G.R. , Johnson, E.W., and Vincent, J.H. (1996) Inversion Techniques for Personal Cascade Impactor Data. Journal of Aerosol Science. Vol 27, pp.1083-1097.

2)Epperson, J.F.(2002) An Introduction to Numerical Methods and Analysis. Wiley, pp. 265-7.

3) O'Shaughnessy, P. T., and Raabe, O. G. (2003) A Comparison of Cascade Impactor Data Reduction Methods. Aerosol Science and Technology. Vol. 37, pp.187-200.

4) Hartmann, W. M. and So, Y. Nonlinear Least Squares and Maximum Likelihood Estimation Using Proc NLP and SAS/IML. Computer Technology Workshop, Joint Statistical Meeting, Orlando, Florida, August 1995.

ASA

AMERICAN STATISTICAL
ASSOCIATION

Papers presented at the
Annual Meeting of the American Statistical Association
Joint Statistical Meetings
Toronto, Ontario, Canada, August 8-12, 2004
and other ASA-sponsored conferences

Search Menu

Full Text Search

Program Book

Table of Contents

Help

Exit

American Statistical Association
1429 Duke Street, Alexandria, VA 22314-3415
(703) 684-1221 Fax: (703) 684-2037
Email: *asainfo@amstat.org*, Web site: *www.amstat.org*

MIRA DIGITAL PUBLISHING

Manufactured by Mira Digital Publishing.
For support e-mail cdsupport@mirasmart.com
or call 314.333.5165.

To return to this screen at any time, click the "HOME" icon
on the toolbar at the top of your screen.

JSM 2004
STATISTICS AS A UNIFIED DISCIPLINE
CD Table of Contents
(Click on a title to view article)