DRUG DEVELOPMENT AND INDUSTRIAL PHARMACY, 19(14), 1747-1753 (1993)

MAXIMUM LIKELIHOOD ESTIMATION FOR THE ANDERSEN SAMPLER FOR LOG-NORMALLY DISTRIBUTED PARTICLES

Dr. Walter Kremers
Medical Department
Ciba-Geigy AG
CH-4002 Basel

ABSTRACT

Given a sample of particles from cascade impactors the geometric mean and variance are commonly estimated by plotting cumulative particle size as a function of particle size on log probability paper, assuming a log normal distribution and drawing a line. Here, theoretical estimates are described base upon the likelihood of the sample, a simple numerical method is described to obtain these estimates.

INTRODUCTION

A Common method for determination of airborne particles or droplets involves the use of cascade impactors, or an Andersen Sampler (Vaughan, 1989). This cascade consists of a series of ever more efficient impactors with respect to aerodynamic diameter. Ideally each impactor should remove all particles larger than a given diameter. In this way the impactors "group" the sampled particles according to size. Modeling particle size as log-normally distributed, the cumulative fraction removed is plotted as a function of the log of impactor stage cut-off diameter for successively larger groups. Then a line is fit to the plotted points, often by eye. Alternatively one may obtain the least squares fit which can be calculated by considering the line which minimizes the sum of the squared vertical deviations between the plotted points and the line. Then the GM (Geometric Mean), e^{μ}, and GV (Geometric Variance), $e^{\sigma 2}$, are calculated from this line or read from the plot.

When responses are i) independent, ii) have mean linear in the predictor, iii) have constant variance and iv) are normally distributed then this regression line gotten

by minimizing the sum of the squared vertical deviations gives the best line according to various criteria. For example this fitted line is a) the Maximum Likelihood Estimate (MLE) (Arnold, 1981) and b) the Best Linear Unbiased Estimate (BLUE) (Mood et alii, 1974) of the mean of the response given the predictor. If i), ii) and iii) hold and iv) is relaxed then this line is still BLUE.

MLEs essentially do what the name implies. They take those parameter values which most support the observed data, and are based upon a well understood theoretical foundation. Best linear unbiased estimates first restrict attention to those estimates which are linear in the data, and on the average are correct, and then select that estimate which has the smallest mean square deviation between it and the actual term being estimated. The MLEs come from a wider class than the BLUEs and are generally preferred when the model upon which it is based is applicable.

For the ideal cascade impactor none of conditions i) - iv) hold. Thus there is not a theoretical justification for fitting the minimum squared deviations line. Furthermore for the case when particle size follows a log-normal distribution the estimates based upon the likelihood approach are easily obtained numerically using standard statistical packages, or alternately with standard programing languages when accompanied by a numerical library for statistical functions. One easily implemented method is the EM algorithm (Dempster et alii, 1977; Boyles, 1983; Wu, 1983) and is presented here.

METHODS

The EM Algorithm

For the algorithm on must first define an unobserved but estimable random variable. This random variable should be such that the MLEs could be easily obtained if it were in fact observed. Furthermore, one must have the facility to calculate the conditional expected value of this random variable given the actual data and current values for the unknown parameters.

Once the unobserved random variable with the above properties has been identified the EM algorithm proceeds as follows:

0-Step) Obtain an initial estimate of the parameter to be estimated and take these as the current parameter values. These estimates may be taken from naive estimates or might be based upon an inspection of the data.

E-Step) Taking the current parameter values calculate the conditional expectation of the "unobserved random variable." Proceed to the M-Step.

M-Step) Calculate the parameter MLEs based upon the conditional expectation. Call these the current parameter values. If the convergence criterion has

not met return to the E-Step. Otherwise stop and take the current values as the numerical MLEs.

Likelihood based upon group frequencies from a Log-Normal distribution

For a description of the actual likelihood let the I observed groups be bounded by $0 = l_0 < l_1 < l_2 < \ldots l_I = \text{infinity}$, and let $f_1, \ldots f_I$ denote the observed fractions sampled in the I groups. Then the probability of a log normally distributed particle being observed in group i is

$$P(\, l_{i-1} < \text{particle} < l_i \,) = P(\, \log(l_{i-1}) < \log(\text{particle}) < \log(l_i) \,)$$
$$= \Phi((\log(l_i) - \mu)/\sigma) - \Phi((\log(l_{i-1}) - \mu)/\sigma)$$

where $\Phi(x)$ is the cumulative distribution function for the standard normal distribution, specifically the integral of $\phi(x) = (2\pi)^{-1} \exp(-x^2/2)$, the probability distribution function for the standard normal distribution. Then the likelihood based upon the observed data apart from a constant and a power is

$$\Pi\, P(\log(\text{particle}) \text{ given } \log(l_i < \text{particle} < l_{i-1})^{f_i}$$
$$= \Pi\, (\, \Phi((\log(l_i) - \mu)/\sigma) - \Phi((\log(l_{i-1}) - \mu)/\sigma) \,)^{f_i}$$

Those values for μ and σ^2 (or μ and σ^2) which maximize this expression are then the MLEs.

Numerical MLEs based upon group frequencies from a Log-Normal distribution

Observe, for the log-normal distribution MLEs based upon the actual values (rather than group frequencies) are the sample and variance of the logs of the observations. That is, for $X_1, \ldots X_n$ indepent, identically distributed log normal random variables, the MLEs for μ and σ^2 are

$$\Sigma_j \log(X_j) / n, \text{ and}$$
$$\Sigma_j (\log(X_j) - (\Sigma_j \log(X_j) / n))^2 / n,$$

(Arnold, 1981; Mood et alii, 1974). Thus the EM algorithm can be used to obtain the MLEs if one can calculate the conditional expectation $\log(X_j)$ given $l_{i-1} < X_j < l_i$ or equivalently given $\log(l_{i-1}) < \log(X_j) < \log(l_i)$. Using the fact that the $\log(X_j)$ are normally distributed, it can be shown that

$$E_j = E(\log(X_j) \text{ given } \log(l_i) < \log(X_j) < \log(l_{i-1}))$$
$$= (\, \mu - \sigma^2 * (\phi((\log(l_i) - \mu)/\sigma) - \phi((\log(l_{i-1}) - \mu)/\sigma)$$
$$/\, (\Phi((\log(l_i) - \mu)/\sigma) - \Phi((\log(l_{i-1}) - \mu)/\sigma)\,)$$

The formula for ϕ is described above. Although Φ is an integral without an expression in simple form many mathematical libraries contain routines for its calculation. Other wise a sufficiently accurate approximation is given by Hill (1973) and can be adapted to the problem here (see Appendix A). Therefore, for the case of

grouped random variables from a log normal distribution a use of the EM algorithm is the following.

0-Step) Define $\log^*(l_0) = \log(l_1)$, and $\log^*(l_i) = \log(l_i)$ for $i>0$. Then take $\Sigma_i f_i \log^*(l_i)$, and $\Sigma_i f_i(\log^*(l_i) - \Sigma_i f_i \log^*(l_i))^2$ as initial current estimates for μ and σ^2.

E-Step) Calculate the E_j using the current values for μ and σ^2. Proceed to the M-step.

M-Step) Update the current values for μ and σ^2 by $\Sigma_i f_i \log(E_i)$, and $\Sigma_i f_i(\log(E_i) - \Sigma_i f_i \log(E_i))^2$. If there is no change to five significant digits in both μ and σ^2, then stop and take the current values as the MLEs. Otherwise return to the E-step.

The initial values described in the 0-step are easily calculated. Alternatively initial values could be taken from the least squares fit of the normal scores on log(size) as described in the introduction. This however is somewhat more involved. and in either case the EM-algorithm should lead to the same values, the MLEs.

If one wants a higher degree of accuracy than that expected by requiring that the iterations agree to f significant digits one may take a different criterion, for example agreement to eight significant digits.

Numerical Example

Consider the data of Table 1 taken from Wiggens (1991). For these the values for μ and σ^2 agree to 5 significant digits after 6 iterations of the EM-algorithm yielding the MLEs 1.87 for the geometric mean and 2.04 for the geometric standard deviation. The least squares approach regressing probit on size yields the estimates 1.82 and 2.11 for GM and GSD. Thus we see that the model approach gives slightly different values for GM and GSD than the naive fitting of a line. These results were obtained using SAS program Version 6.06 (SAS Institute Inc., 1990). The SAS program is available upon request from the author.

Regression of Scores on Log(size) and Log(size) on Scores

Whereas Wiggens (1991) apparently regresses score on log(size), one could regress log(size) on score which gives the estimates 1.81 and 2.14. Thus, even if one adopts the least square criteria to define the "best line" this line is not unique without specifying which variable is going to be regressed on the other. Further this choice is ambiguous in the literature. While Wiggens (1991) and Andersen Samplers (1985) plot log(size) on the vertical (ordinate), Sucker et alii (1978) and Pharmacopeial Forum (1991) plot this variable on the horizontal (abscissa). All state that the best

TABLE 1
Example Data from Andersen Impactor

Lower Particle Size	Upper Particle Size	Observed Cumulative Percent
0.00	0.52	5.0
0.52	0.70	2.0
0.70	1.09	20.0
1.09	2.02	50.0
2.02	2.99	75.0
2.99	4.37	92.0
4.37	6.20	95.0
6.20	9.0	98.0
9.00		100.0

TABLE 2
Parameter Values as Function of Iteration of the EM-Algorthm
Iteration 0 denotes the starting values.

Iteration	μ	σ^2	σ	GM	GV	GSD
0	0.423743	0.229173	0.478720	1.52767	1.257560	1.614007
1	0.619400	0.463800	0.681029	1.85781	1.590105	1.975909
2	0.629200	0.501000	0.707814	1.87611	1.650371	2.029549
3	0.629000	0.505700	0.711126	1.87573	1.658146	2.036283
4	0.628900	0.506300	0.711548	1.87555	1.659141	2.037142
5	0.628900	0.506400	0.711618	1.87555	1.659307	2.037285
6	0.628900	0.506400	0.711618	1.87555	1.659307	2.037285

line be fit but none state which variable is to be fitted on the other, although this is commonly taken to mean the least squares of the vertical is to be regressed on the horizontal. Because the least squares criteria gives different results for the two different cases the case must be specified for the estimated GM and GSD to be unique when calculated using the least squares approach.

CONCLUSIONS

Standard regression types of estimates can often be motivated on the grounds that they are optimal according to various criteria when deviations of the observations

from the line are independent and identically distributed normal random variables. For the case of cumulative group frequencies from the log-normal distribution, the deviations from the normal score - log(size) plot do not satisfy these prerequisites. Nonetheless, MLEs can be obtained numerically from such data, for example with the EM algorithm presented here. Because of its theoretical basis, yielding the parameters which give the most plausibility to the data, the MLEs should be preferred over the regression method for the estimation of either μ and σ^2 or GM and GSD.

REFERENCES

Anderson Samplers (1985) Operating Manual for Andersen 1 ACFM Non.Viable Ambient Particle Sizing Samplers. Andersen Samplers, Atlanta.

Arnold, S.F. (1983) The Theory of Linear Models and Multivariate Analysis. Wiley, New York.

Hill, I.D. (1973) "Algorithm AS 66, The Normal Integral." *Applied Statistics*, 22:424-427.

Mood, A.M., Graybill, F.A., Boes, D.C. (1974) Introduction to the Theory of Statistics, Third Edition. McGraw-Hill, New York.

Pharmacopeial Forum (1991) "Stimuli to the Revision Process," 18:1705-1708.

SAS Institute INC. (1990) SAS(R) Language: Reference, Version 6, First Edition. SAS Institute Inc, Cary, NC.

Sucker, H. Fuchs, P., Speiser R. (1978) Pharmzeutische Technologie. Georg Theime Verlag, Stuttgart.

Vaughan, N.P. (1989) The Andersen impactor:Calibration, wall losses and numerical simulation. J. Aersol Sc. 20: 67-90.

Wiggins, N.A. (1991) "The Development of a Mathematical Approximation technique to determine the mass median aerodynamic diameter (MMAD) and geometric standard deviation of drug particles in an inhalation aerosol spray." *Drug Development and Industrial Pharmacy*, 17: 1971-1986.

APPENDIX

```
***** Numerical approximation of normal probability,  *****;
***** P(Z < z) as part of a SAS data step             *****;
upper = 0 ;
if (z < 0) then do ;
  z = - z ;
  upper = 1 - upper ;
  end ;
if (z > 8) then do ;
  normprob = 1 ;
```

```
  end ;
else do ;
  y = 0.5 * z**2 ;
  if (z < 1.28) then do ;
    normprob = 0.5 - z * (0.398942280444 - 0.399903438504 * y /
       (y + 5.75885480458 - 29.8213557808 /
       (y + 2.62433121679 + 48.6959930692 /
       (y + 5.92885724438)))) ;
    end ;
  else do ;
    normprob = 0.398942280385 * exp(-y) /
       (z - 3.8052E-8 + 1.00000615302 /
       (z + 3.98064794E-4 + 1.98615381364 /
       (z - 0.151679116635 + 5.29330324926 /
       (z + 4.8385912808 - 15.1508972451 /
       (z + 0.742380924027 + 30.789933034 / (z + 3.99019417011))))));
    end ;
  end ;
if ^upper then normprob = 1 - normprob ;
```