February 1994
Publication # 9285.7-15-1
EPA 540-R-93-081

APPENDIX A: HOW TO CALCULATE THE GEOMETRIC STANDARD DEVIATION FROM BLOOD LEAD DATA, IF YOU MUST

A.1 A DIRECT METHOD FOR CALCULATING THE GEOMETRIC STANDARD DEVIATION

One of the simplest approaches to calculating a GSD from a sample of blood lead and environmental lead data is based on the idea that children with similar environmental lead exposures will have similar geometric mean blood lead levels. For children of a given age with similar soil lead (denoted PbS), dust lead (denoted PbD), and other lead exposures, we can reasonably characterize the variability in blood lead level (denoted GSD) calculated with respect to the actual geometric mean blood lead level of this group of children (denoted GMB) without modelling blood lead levels. The procedure shown here is the simplest procedure we have found, but even with this procedure, the user must be prepared to do a great deal of statistical calculation. We will illustrate how an empirical GSD may be calculated from data after we describe the procedure:

STEP 1: Divide the data set into subgroups, where each group has children of a given age, with soil lead levels in a given interval, dust lead levels in a given interval, and with distinct levels of other important variables. Each such group corresponds to a "box" or cell of soil and dust lead levels, and levels of other variables if used.

STEP 2: From each individual blood lead (denoted PbB) in each cell, calculate ln(PbB), where ln denotes the natural logarithm.

STEP 3: Within each cell, calculate the mean and the standard deviation of the ln(PbB) values. Then, for that cell,

GMB = exp(mean of ln(PbB) values within the cell)
GSD = exp(standard deviation of ln(PbB) values within the cell)

where exp denotes the process of calculating the exponential function of the indicated quantity. Exponential and natural logarithm functions are available in most statistical packages for microcomputers and on most scientific calculators.

STEP 4: Calculate the inter-individual GSD for this neighborhood by finding the median or middle value of the GSD values in the sample. The median is found by ordering the GSD values from all cells from smallest to largest. If the number of GSD values is odd, the median is the middle value; if sample size is an even number, find the average of the two middle values. Since the number of observations in each box or cell is different, each GSD should be counted a number of times according to the number of degrees of freedom (cell count minus one) for that GSD.

STEP 5: Users with some statistical background may wish to examine the within-cell GSD's for patterns based on the data, such as by plotting GSD against GMB or against the within-cell value of age, PbS, PbD, or other stratifying variables.

STEP 6: Users with more statistical background may wish to use other approaches to calculating a "typical" GSD, such as by calculating a mean or pooled variance of the within-cell variances of ln(PbB). We would caution such users that the data should be carefully evaluated for outliers, either in raw PbB values or in the calculated GSD. One convenient approach for visual detection of outliers is a normal probability of within-cell variances after a variance-stabilizing transformation such as the cube root of the within-cell variance of ln(PbB).

EXAMPLE: In a sample of 166 children from the Midvale, Utah study of 1989 (Bornschein et al., 1990), we found that the estimation of blood lead levels could be considerably improved by determining whether or not the children lived in houses in which paint had recently been removed. There is substantial evidence that inadequately controlled lead paint abatement may increase blood levels in resident young children by 2 to 4 ug/dL on average in the first 6 to 12 months after paint removal (Rabinowitz et al., 1984; Marcus et al., 1991; Menton et al., 1993). Interviews with the family provided such information for 162 of the 166 children, which was used as an additional stratifying variable.

The worksheet for determining subgroups are shown in Table A-1. Each table gives the blood leads of children of a given age, divided by whether or not there was recent paint removal in the residence. Within each table, the children are divided according as the PbS and PbD values at their residence. Each cell in the table corresponds to intervals of 250 ppm of PbS and 250 ppm of PbD. The soil lead levels were averages of non-missing values of perimeter, bare area, play area, and garden soils. It should be noted that data for most of the cells are not available with such detailed sub-division of the data set, and that at higher soil and dust lead concentrations, there is usually only one observation per cell. There was only one case in which two children from the same family had the same age, in years, and analyses without this duplication would produce very similar results. Otherwise, all blood lead levels (denoted PbB) within each cell come from different families. This is believed to

TABLE A-1. CELLS OF BLOOD LEAD LEVELS IN 165 MIDVALE CHILDREN, BY PAINT REMOVAL STATUS, AGE, AND INTERVALS OF 250 μg/g IN SOIL AND DUST LEAD[1]

Paint Removal	Age	Soil Pb	Dust Pb	Blood Lead (μg/dL) Smallest →→Largest					
.	0	375	375	5.5	.	.	.	.	.
.	2	.	625	6.	.	.	.	.	.
.	2	375	125	4.	.	.	.	.	.
.	3	625	625	3.	.	.	.	.	.
.	4	125	125	6.5	.	.	.	.	.
0	0	125	.	3.	6.	.	.	.	.
0	0	125	125	1.	.	.	.	.	.
0	0	125	375	0.5	.	.	.	.	.
0	0	375	125	3.	4.5	.	.	.	.
0	0	375	375	5.5	.	.	.	.	.
0	0	375	1125	3.	7.	.	.	.	.
0	0	375	1375	3.5	.	.	.	.	.
0	0	1125	875	13.5	.	.	.	.	.
0	1	.	625	5.5	.	.	.	.	.
0	1	125	375	2.5	3.	.	.	.	.
0	1	375	375	4.	5.5	7.	.	.	.
0	1	375	625	4.5	.	.	.	.	.
0	1	375	875	3.5	.	.	.	.	.
0	1	625	375	3.	7.	8.	10.	.	.
0	1	625	625	3.5	6.	6.	.	.	.
0	1	875	625	3.	.	.	.	.	.
0	1	875	1125	6.	.	.	.	.	.
0	1	1125	875	1.	10.5	.	.	.	.
0	1	1375	1125	6.	.	.	.	.	.
0	1	1625	625	3.	.	.	.	.	.
0	2	.	375	4.	6.	.	.	.	.

TABLE A-1 (cont'd). CELLS OF BLOOD LEAD LEVELS IN 165 MIDVALE CHILDREN, BY PAINT REMOVAL STATUS, AGE, AND INTERVALS OF 250 µg/g IN SOIL AND DUST LEAD[1]

Paint Removal	Age	Soil Pb	Dust Pb	Blood Lead (µg/dL) Smallest →→Largest					
0	2	.	625	7.	.	.	.	.	.
0	2	125	.	5.	.	.	.	.	.
0	2	125	125	2.5	5.5	5.5	8.	12.	.
0	2	125	625	6.	.	.	.	.	.
0	2	375	375	1.5	.	.	.	.	.
0	2	375	1125	4.5	.	.	.	.	.
0	2	625	375	14.5	.	.	.	.	.
0	2	625	625	4.	7.	11.5	.	.	.
0	2	875	1125	5.5	.	.	.	.	.
0	2	1125	.	13.	.	.	.	.	.
0	2	1125	625	9.5	.	.	.	.	.
0	2	1125	875	19.	.	.	.	.	.
0	3	.	625	5.	.	.	.	.	.
0	3	125	125	2.5	.	.	.	.	.
0	3	125	375	2.	7.5	.	.	.	.
0	3	375	125	6.5	.	.	.	.	.
0	3	375	375	3.	4.	.	.	.	.
0	3	375	1375	13.	.	.	.	.	.
0	3	1125	625	16.5	.	.	.	.	.
0	3	1375	1125	5.	.	.	.	.	.
0	4	.	375	2.	.	.	.	.	.
0	4	125	125	4.	7.5	.	.	.	.
0	4	125	375	5.5	6.	.	.	.	.
0	4	375	.	2.	.	.	.	.	.
0	4	375	625	1.5	7.	.	.	.	.

TABLE A-1 (cont'd). CELLS OF BLOOD LEAD LEVELS IN 165 MIDVALE CHILDREN, BY PAINT REMOVAL STATUS, AGE, AND INTERVALS OF 250 μg/g IN SOIL AND DUST LEAD[1]

Paint Removal	Age	Soil Pb	Dust Pb	Blood Lead (μg/dL) Smallest →→Largest					
0	4	625	375	5.	.	.	.	.	.
0	4	875	625	7.5	.	.	.	.	.
0	4	1125	2375	5.	.	.	.	.	.
0	4	2125	875	8.	.	.	.	.	.
0	5	125	125	2.	4.5	8.5	.	.	.
0	5	125	375	2.5	3.5	10.	.	.	.
0	5	375	375	4.	6.	.	.	.	.
0	5	625	375	5.	.	.	.	.	.
0	5	625	625	4.	.	.	.	.	.
0	5	625	1375	4.5	.	.	.	.	.
0	5	1125	875	13.5	.	.	.	.	.
1	0	125	125	3.	16.5	.	.	.	.
1	0	125	375	0.5	1.5	3.5	5..	6.	.
1	0	375	375	8.5	.	.	.	.	.
1	0	375	875	5.	.	.	.	.	.
1	1	.	375	8.	22.5	.	.	.	.
1	1	125	125	5.5	.	.	.	.	.
1	1	125	375	5.5	.	.	.	.	.
1	1	375	375	3.5	.	.	.	.	.
1	1	375	625	5.5	.	.	.	.	.
1	1	375	875	16.5	.	.	.	.	.
1	1	625	125	2.5	.	.	.	.	.
1	1	625	375	9.	9.	.	.	.	.
1	1	875	625	6.	.	.	.	.	.
1	1	1125	1375	5.5	.	.	.	.	.

TABLE A-1 (cont'd). CELLS OF BLOOD LEAD LEVELS IN 165 MIDVALE CHILDREN, BY PAINT REMOVAL STATUS, AGE, AND INTERVALS OF 250 µg/g IN SOIL AND DUST LEAD[1]

Paint Removal	Age	Soil Pb	Dust Pb	Blood Lead (µg/dL) Smallest →→Largest					
1	2	.	.	8.5	.	.	.	.	
1	2	125	125	3.	4.	4.5	5.5	.	.
1	2	125	375	2.5	3.5	5.	5.	5.5	19.5
1	2	375	.	6.	.	.	.	.	.
1	2	375	375	3.	10.	.	.	.	.
1	2	375	625	8.5	.	.	.	.	.
1	2	625	625	6.5	.	.	.	.	.
1	2	1875	625	10.5	.	.	.	.	.
1	3	.	.	8.5	.	.	.	.	.
1	3	.	625	4.	.	.	.	.	.
1	3	125	375	4.5	.	.	.	.	.
1	3	375	625	5.5	5.5	8.	.	.	.
1	3	375	875	4.	.	.	.	.	.
1	3	625	875	2.	.	.	.	.	.
1	3	875	3625	2.	.	.	.	.	.
1	3	1625	1375	15.5	.	.	.	.	.
1	3	1875	625	7.5	.	.	.	.	.
1	4	.	375	3.5	18.	.	.	.	.
1	4	.	625	3.5	.	.	.	.	.
1	4	125	125	4.5	5.	5.	5.	5.	.
1	4	125	375	2.	3.5	4.	8.	.	.
1	4	375	.	7.	.	.	.	.	.
1	4	875	625	9.	.	.	.	.	.
1	4	875	1125	7.5	.	.	.	.	.
1	4	1625	1375	9.5	.	.	.	.	.

TABLE A-1 (cont'd). CELLS OF BLOOD LEAD LEVELS IN 165 MIDVALE CHILDREN, BY PAINT REMOVAL STATUS, AGE, AND INTERVALS OF 250 µg/g IN SOIL AND DUST LEAD[1]

Paint Removal	Age	Soil Pb	Dust Pb	Blood Lead (µg/dL) Smallest →→Largest					
1	4	3125	1625	13.	.	.	.	.	.
1	5	.	.	4.5	.	.	.	.	.
1	5	125	125	4.	5.	.	.	.	.
1	5	125	375	4.	7.5	.	.	.	.
1	5	375	375	1.5	.	.	.	.	.
1	5	625	375	6.5	.	.	.	.	.
1	5	1875	625	5.5	.	.	.	.	.

[1]An isolated decimal point denotes a missing value.

give a much more valid estimate of variability than within-family GSD's for children of different ages, but similar genetic and non-lead environmental factors and similar family behavior patterns.

The statistics for GMB and GSD were calculated as described in Step 3, for each cell where enough data were available (at least 2 PbB values in order to calculate GSD). The results are shown in Table A-2. The PbS and PbD values are the cell midpoints, and provide convenient plot points. Some of the GSD values are very high, as for the cell whose two values are PbB = 1.5 and 10 μg/dL.

The distribution of GSD values for all cells is shown in Table A-3, in the form of a "stem-and-leaf" plot (Tukey, 1977). No weighting scheme has been applied. Many users would prefer a weighted GSD where the number of observations in each cell is taken into account. This can be done by counting each GSD estimate as representing the number of degrees of freedom (denoted DF) in the GSD estimate. In this application, DF = N − 1, where N is the number of PbB values in the cell. A DF-weighted stem-and-leaf plot is shown in Table A-4. In the unweighted case, the median GSD = 1.694 may be taken as a representative value for this community. In the weighted DF case, a somewhat larger median GSD = 1.768 may be used.

TABLE A-2. GEOMETRIC MEAN AND GEOMETRIC STANDARD DEVIATION OF BLOOD LEADS IN CELLS OR GROUPS, BY PAINT REMOVAL STATUS, AGE, AND INTERVALS OF 250 µg/g IN SOIL AND DUST LEAD[1]

Paint Removal	Age (Years)	Soil Lead (µg/g)	Dust Lead (µg/g)	N	Geometric Mean Blood Lead (µg/dL)	GSD
.	0	375	375	1	5.5	.
.	2	.	625	1	6.	.
.	2	375	125	1	4.	.
.	3	625	625	1	3.	.
.	4	125	125	1	6.5	.
0	0	125	.	2	4.243	1.633
0	0	125	125	1	1.	.
0	0	125	375	1	0.5	.
0	0	375	125	2	3.674	1.332
0	0	375	375	1	5.5	.
0	0	375	1125	2	4.583	1.821
0	0	375	1375	1	3.5	.
0	0	1125	875	1	13.5	.
0	1	.	625	1	5.5	.
0	1	125	375	2	2.739	1.138
0	1	375	375	3	5.360	1.324
0	1	375	625	1	4.5	.
0	1	375	875	1	3.5	.
0	1	625	375	4	6.402	1.693
0	1	625	625	3	5.013	1.365
0	1	875	625	1	3.	.
0	1	875	1125	1	6.	.
0	1	1125	875	2	3.240	5.273
0	1	1375	1125	1	6.	.

TABLE A-2 (cont'd). GEOMETRIC MEAN AND GEOMETRIC STANDARD DEVIATION OF BLOOD LEADS IN CELLS OR GROUPS, BY PAINT REMOVAL STATUS, AGE, AND INTERVALS OF 250 µg/g IN SOIL AND DUST LEAD[1]

Paint Removal	Age (Years)	Soil Lead (µg/g)	Dust Lead (µg/g)	N	Geometric Mean Blood Lead (µg/dL)	GSD
0	1	1625	625	1	3.	.
0	2	.	375	2	4.899	1.332
0	2	.	625	1	7.	.
0	2	125	.	1	5.	.
0	2	125	125	5	5.055	2.013
0	2	125	625	1	6.	.
0	2	375	375	1	1.5	.
0	2	375	1125	1	4.5	.
0	2	625	375	1	14.5	.
0	2	625	625	3	6.854	1.696
0	2	875	1125	1	5.5	.
0	2	1125	.	1	13.	.
0	2	1125	625	1	9.5	.
0	2	1125	875	1	19.	.
0	3	.	625	1	5.	.
0	3	125	125	1	2.5	.
0	3	125	375	2	3.873	2.546
0	3	375	125	1	6.5	.
0	3	375	375	3	4.762	1.768
0	3	375	1375	1	13.	.
0	3	1125	625	1	16.5	.
0	3	1375	1125	1	5.	.
0	4	.	375	1	2.	.

TABLE A-2 (cont'd). GEOMETRIC MEAN AND GEOMETRIC STANDARD DEVIATION OF BLOOD LEADS IN CELLS OR GROUPS, BY PAINT REMOVAL STATUS, AGE, AND INTERVALS OF 250 µg/g IN SOIL AND DUST LEAD[1]

Paint Removal	Age (Years)	Soil Lead (µg/g)	Dust Lead (µg/g)	N	Geometric Mean Blood Lead (µg/dL)	GSD
0	4	125	125	2	5.477	1.560
0	4	125	375	2	5.745	1.063
0	4	375	.	1	2.	.
0	4	375	625	2	3.240	2.972
0	4	625	375	1	5.	.
0	4	875	625	1	7.5	.
0	4	1125	2375	1	5.	.
0	4	2125	875	1	8.	.
0	5	125	125	3	4.245	2.065
0	5	125	375	3	4.440	2.061
0	5	375	375	2	4.899	1.332
0	5	625	375	1	5.	.
0	5	625	625	1	4.	.
0	5	625	1375	1	4.5	.
0	5	1125	875	1	13.5	.
1	0	125	125	2	3.	16.5
1	0	125	375	5	0.5	1.5
1	0	375	375	1	8.5	.
1	0	375	875	1	5.	.
1	1	.	375	2	8.	22.5
1	1	125	125	1	5.5	.
1	1	125	375	1	5.5	.
1	1	375	375	1	3.5	.

TABLE A-2 (cont'd). GEOMETRIC MEAN AND GEOMETRIC STANDARD DEVIATION OF BLOOD LEADS IN CELLS OR GROUPS, BY PAINT REMOVAL STATUS, AGE, AND INTERVALS OF 250 µg/g IN SOIL AND DUST LEAD[1]

Paint Removal	Age (Years)	Soil Lead (µg/g)	Dust Lead (µg/g)	N	Geometric Mean Blood Lead (µg/dL)	GSD
1	1	375	625	1	5.5	.
1	1	375	875	1	16.5	.
1	1	625	125	1	2.5	.
1	1	625	375	1	9.	9.
1	1	875	625	2	6.	.
1	1	1125	1375	1	5.5	.
1	2	.	.	1	8.5	.
1	2	125	125	4	3.	4.
1	2	125	375	6	2.5	3.5
1	2	375	.	1	6.	.
1	2	375	375	2	3.	10.
1	2	375	625	1	8.5	.
1	2	625	625	1	6.5	.
1	2	1875	625	1	10.5	.
1	3	.	.	1	8.5	.
1	3	.	625	1	4.	.
1	3	125	375	1	4.5	.
1	3	375	625	3	5.5	5.5
1	3	375	875	1	4.	.
1	3	625	875	1	2.	.
1	3	875	3625	1	2.	.
1	3	1625	1375	1	15.5	.
1	3	1875	625	1	7.5	.

TABLE A-2 (cont'd). GEOMETRIC MEAN AND GEOMETRIC STANDARD DEVIATION OF BLOOD LEADS IN CELLS OR GROUPS, BY PAINT REMOVAL STATUS, AGE, AND INTERVALS OF 250 µg/g IN SOIL AND DUST LEAD[1]

Paint Removal	Age (Years)	Soil Lead (µg/g)	Dust Lead (µg/g)	N	Geometric Mean Blood Lead (µg/dL)	GSD
1	4	.	375	2	3.5	18.
1	4	.	625	1	3.5	.
1	4	125	125	5	4.5	5.
1	4	125	375	4	2.	3.5
1	4	375	.	1	7.	.
1	4	875	625	1	9.	.
1	4	875	1125	1	7.5	.
1	4	1625	1375	1	9.5	.
1	4	3125	1625	1	13.	.
1	5	.	.	1	4.5	.
1	5	125	125	2	4.	5.
1	5	125	375	2	4.	7.5
1	5	375	375	1	1.5	.
1	5	625	375	1	6.5	.
1	5	1875	625	1	5.5	.

A.2 A MORE SOPHISTICATED STATISTICAL METHOD FOR ESTIMATING THE GEOMETRIC STANDARD DEVIATION

The GSD actually represents the residual variability in the logarithm of the predicted blood lead level. A direct regression method that is an overly simplified approximation to the IEUBK model at steady state exposure may be useful in deriving a residual GSD from a blood lead and environmental lead study. The method is based on the concepts that: (1) the IEUBK model at low to moderate steady-state exposure yields predicted blood leads that are

TABLE A-3. STEM AND LEAF PLOT OF GEOMETRIC STANDARD DEVIATION FOR MIDVALE CHILDREN[1,2]

MINIMUM:	1.048	
LOWER QUARTILE:	1.332	
MEDIAN:	1.694	
UPPER QUARTILE:	2.071	
MAXIMUM:	5.273	
1		0011
1	H	22333333
1		55
1	M	6667
1		88
2	H	00000
2		3
2		5
2		7
2		9
OUTSIDE VALUES		
3		13
5		2

[1]N = 32 groups, unweighted.
[2]76 groups with missing values excluded from plot.

approximately linear functions of PbS and PbD, with age-dependent regression coefficients; (2) the linear model should be fitted in a logarithmic form so as to estimate relative variability. In order to use the model, it is necessary to create indicator variables for the age of the child in the study. These are:

TABLE A-4. STEM AND LEAF PLOT OF GEOMETRIC STANDARD DEVIATION FOR MIDVALE CHILDREN[1] (Weighted by Degrees of Freedom)

MINIMUM:	1.048	
LOWER QUARTILE:	1.332	
MEDIAN:	1.768	
UPPER QUARTILE:	2.061	
MAXIMUM:	5.273	
1		0000011
1	H	2222233333333
1		55
1	M	66666677
1		8888
2	H	00000000000000
2		3
2		5
2		7777
2		9
OUTSIDE VALUES		
3		13
5		2

[1]N = 58 groups, weighted by degrees of freedom.

AGE0 = 1 if the child is age 0 to 11 months; AGE0 = 0 if not;
AGE1 = 1 if the child is age 12 to 23 months; AGE1 = 0 if not;
AGE2 = 1 if the child is age 24 to 35 months; AGE2 = 0 if not;
AGE3 = 1 if the child is age 36 to 47 months; AGE3 = 0 if not;
AGE4 = 1 if the child is age 48 to 59 months; AGE4 = 0 if not;
AGE5 = 1 if the child is age 60 to 71 months; AGE5 = 0 if not;

AGE6 = 1 if the child is age 72 to 83 months; AGE6 = 0 if not;

and so on. Then the model that may be fitted, using all of the children in the data set for which observed or imputed PbS and PbD values are available, using a nonlinear regression program for parameter estimation, is given by

ln(PbB) = ln(A0*AGE0 + A1*AGE1 + A2*AGE2 + A3*AGE3 + …
+ PbS * (B0*AGE0 + B1*AGE1 + B2*AGE2 + B3*AGE3 + …) +
+ PbD * (C0*AGE0 + C1*AGE1 + C2*AGE2 + C3*AGE3 + …) +
+ X * (D0*AGE0 + D1*AGE1 + D2*AGE2 + D3*AGE3 + …))

Here, X represents other predictive covariates for blood lead. In the Midvale example, X = RMVPAINT = 1 if paint has recently been removed from the premises, and X = 0 if not. In other applications, it may be useful to use water lead or air lead levels as an additional predictor. X may be omitted if necessary. In many applications, the regression parameters may be set equal for some ages. For example, if blood leads stabilize for ages 3 to 5 years, we may set A3 = A4 = A5, B3 = B4 = B5, C3 = C4 = C5, etc. Since the age-dependence of soil and dust lead exposure may differ somewhat from one site to another, depending on climate or other factors, no general prescription for how to carry out such analyses may be given.

When a non-linear regression model is fitted to the data by use of a program that estimates non-linear parameters, it is then possible to calculate the residual standard deviation S for the model, so that

S = standard deviation of ln(observed PbB / predicted PbB)
GSD = exp(S).

For the Midvale example described in this Appendix, we find that for N = 143 children with no missing data for PbD, PbS, or RMVPAINT, S = 0.5701 on the natural log scale, thus GSD = 1.768. The approximate relative standard error of S^2 is $(2 / (N - p))^{0.5}$, where p is the number of nonlinear parameters estimated from the data. With p = 12 parameters (ages 2 to 5 years were grouped), we have $(2 / (143 - 12))^{0.5} = 0.1236$ relative standard deviation for the variance. An approximate 95% confidence interval for the true value of S^2 has a lower bound $(1 - 2 * (2 / (N - p))^{0.5})$, and an upper bound $(1 + 2 * (2 / (N - p))^{0.5})$, times S^2. For Midvale, the limits are

$$(1 - 2 * 0.1236) * (0.5701)^2 = 0.2447$$

$$(1 + 2 * 0.1236) * (0.5701)^2 = 0.4053$$

Thus the lower and upper bounds for S are 0.4947 to 0.6366, and for GSD = exp(S) the confidence limits are 1.640 to 1.890.