Fitting A Second-Order Parametric Distribution Conditioned on an Explanatory Variable Using Maximum Likelihood Estimation

David E. Burmaster
Alceon Corporation
PO Box 382669 Harvard Square Station
Cambridge, MA 02238-2669
tel: 617-864-4300; fax: 617-864-9954
deb@Alceon.com

Kimberly M. Thompson
Harvard School of Public Health
Center for Risk Analysis
718 Huntington Avenue
Boston, MA 02115
tel: 617-432-4285; fax: 617-432-0190
kimt@hsph.harvard.edu

Key Words

variability, uncertainty, second-order random variables, maximum likelihood estimation, Radon 222

Abstract

We show how to use the method of maximum likelihood estimation (MLE) to fit a second-order parametric distribution conditioned on a single explanatory variable to data. To illustrate the method, we demonstrate how a second-order LogNormal distribution, conditioned on the population served, can model the variability and the parametric uncertainty in data collected by the US Environmental Protection Agency for the concentration of Radon 222 in drinking water supplied from ground water, even though 28 percent of the data fall at or below the Minimum Reporting Level.

1.0 Introduction

Using data collected by a regulatory agency, we show how to use the method of maximum likelihood estimation (MLE) to fit a second-order parametric distribution conditioned on an explanatory variable. [EndNote 1] In this paper, we combine and extend three methods: the animation method used to visualize data during exploratory data analysis (Burmaster & Thompson, 1999), the maximum likelihood method used to fit first-order distributions for body weight as a function of age (Burmaster & Crouch, 1997), and the maximum likelihood method for fitting second-order distributions to data (Burmaster & Thompson, 1998).

Linear (or nonlinear) regression is used to study relationships between measured variables, say $\{x_i, y_i\}$ (Weisberg, 1980). With regression, an analyst seeks a linear (or nonlinear) function that predicts the expected value of a dependent (or response) variable, y, as a function of an independent (or explanatory or predictor) variable, x. An equation specifies the simplest regression model:

$$E[\, y \mid x \,] = f[x] \qquad \text{Eqn 1}$$

where $E[\, y \mid x \,]$ denotes the expected value of random variable y conditioned on (as a function of) the single explanatory variable x (see Eqn 26-1 and discussion in Stuart & Ord, 1991). The regression model also specifies an error model (e.g., $\varepsilon \sim$iid Normal[0, σ]) for the distribution of the differences between the individual measurements and the prediction, $\varepsilon_i = y_i - E[\, y \mid x \,]$. For rigorous discussions of linear and nonlinear regression, see Draper and Smith (1981).

Here, we model the full second-order distribution (not just the expected value) of a dependent variable as a function of an independent variable. To illustrate this maximum likelihood technique for modeling the variability and the parametric uncertainty in data as a second-order distribution conditioned on one explanatory variable, we re-analyze data collected by the US Environmental Protection Agency (US EPA) for the concentration of radon in public drinking water supplies drawn from ground water wells. We re-analyze these data to illustrate the general method, not to draw policy conclusions, since we understand that the Agency has already begun to collect new data for its regulatory program (Barry, 1998).

2.0 The NIRS Data

As explained by Longtin (1988; 1990) and Barry and Brattin (1998), the National Inorganic and Radionuclides Survey (NIRS) selected a stratified random sample from the 47,700 community drinking water systems supplied from ground water as inventoried by the US Environmental Protection Agency (US EPA) in 1984. Of the 1,000 water service districts initially included in the NIRS survey, 990 responded (8 of which were later excluded for data quality reasons). Six water samples per system were analyzed to compute the mean concentration of Radon 222 in picocuries per liter (denoted [Rn222] in pCi/L) in the finished water in each distribution system. The resulting 982 data points in the final NIRS data comprise a representative sample of

water supply systems (but not a representative sample of human population) stratified by geographical region.

The NIRS data exhibit a large degree of variability. [EndNote 2] The reported measurements for [Rn222] range from less than 100 pCi/L to greater than 25,000 pCi/L. (Longtin, 1990). The US EPA concluded that a single Minimum Reporting Level (MRL = 100 pCi/L) applies to all the data in the Survey (Barry, 1998). [Helsel (1990) presents some of the basic issues when analyzing censored data.] Since the US Environmental Protection Agency has stated that every value reported below 100 piC/L is unreliable, we censored the data at the single MRL as did Barry and Brattin (1998). Of the 982 data points, 275 (or 28 percent) of the data points fall at or below the MRL and 707 (or 72 percent) of the data points fall above the MRL.

3.0 Barry's and Brattin's Analysis of the NIRS Data

Barry and Brattin (1998) grouped the NIRS data into 5 size categories ranging from "Very Very Small" to "Large." See Table 1, compiled from Table 3 in Barry and Brattin. Barry and Brattin used probability plots, maximum likelihood estimation (MLE), and other statistical techniques to analyze and model various patterns in these left-censored data. Barry and Brattin conclude (1998, page 598) that the distribution of [Rn222] in community ground water systems are "... reasonably well described..." by 5 different LogNormal distributions. For each of the size categories, Barry and Brattin report the best-fit LogNormal parameters, $\hat{\mu}$ and $\hat{\sigma}$; the estimated standard errors for the parameters, $\sigma(\hat{\mu})$ and $\sigma(\hat{\sigma})$; and the estimated correlation between the parameters, $\hat{\rho}$. Barry and Brattin continue,

> "However, despite the good overall [LogNormal] fit, some NIRS data .... exhibit systematic deviations from lognormality, displaying longer (heavier) than lognormal tails. Although these deviations generally occur beyond the 93rd - 95th percentiles, for situations in which high-end estimation is important, a better fitting long-tailed distribution or compound distribution might be better suited to these strata." (emphasis added).

We note that Burmaster and Wilson (1999) recently reanalyzed the same data using a mixture model with two LogNormal distributions as the components.

4.0 A Second-Order Model of the NIRS Data Conditioned on Population

We pursue and quantify Barry and Brattin's suggestion "... that a compound distribution might be better suited [to these data]" by fitting to the NIRS data a LogNormal distribution conditioned on smooth functions of the population served. In this re-analysis, we use the following variables and units:

C = [Rn222] in the water (pCi/L), and
P = population served by the water supply system (number of people).

To simplify the notation, we use lnC and lnP to represent ln[C] and ln[P] as appropriate.

Following Barry and Brattin (1998), we fit a LogNormal distribution to the NIRS data:

$$\ln[X] \sim \text{Normal}[\,\mu, \sigma\,] \qquad \text{Eqn 2}$$

where μ and σ are the common parameters for the Normal or Gaussian distribution (see, e.g., Evans et al, 1993). We convert all values less than or equal to the MRL to the MRL itself (Barry, 1998). Working in logarithmic space for concentration, we fit the PDF and CDF for the Normal distribution to lnC (Aitchison & Brown, 1957; Crow & Shimizu, 1988; Burmaster & Hull, 1997):

$$\text{PDF}[\text{Normal}[\,\ln C \mid \mu, \sigma\,]] = \frac{1}{\sigma \sqrt{2\pi}} \exp\left[-\frac{1}{2} \frac{(\ln C - \mu)^2}{\sigma^2} \right] \qquad \text{Eqn 3}$$

$$\text{CDF}[\text{Normal}[\,\ln C \mid \mu, \sigma\,]] = \frac{1}{2} \left[1 + \text{erf}\left[\frac{\ln C - \mu}{\sigma \sqrt{2}} \right] \right] \qquad \text{Eqn 4}$$

Here, exp[•] denotes the exponential function, ln[•] denotes the natural (or Napierian) logarithm function, and erf[•] denotes the error function (Abramowitz & Stegun, 1964).

For use in Eqns 3 and 4, we searched for two continuous and smooth functions, $\mu[P]$ and $\sigma[P]$, to provide a unified model for the pooled NIRS data as a function of population served.

5.0 Maximum Likelihood Estimation

When using the method of maximum likelihood estimation (MLE), the analyst selects that value of the parameter θ for which the probability of the sample is a maximum (Keeping, 1995). When using MLE, the analyst first computes the loglikelihood function, J, for a particular parametric model as a function of all the parameters in the model, θ, with the given data taken as fixed values. Using nonlinear optimization, the analyst then finds the maximum value of $J[\theta]$ and the corresponding values for all the parameters in the vector $\theta = \hat{\theta}$. For large sample sizes (and under certain regularity conditions met in this analysis), the uncertainty in the parameters θ follows a multivariate Normal distribution (see, e.g., Evans et al, 1993) with (i) the mean vector equal to the optimal values of the parameters, $\hat{\theta}$, and (ii) the variance covariance matrix equal to (Appendix 1, Cox & Snell, 1989):

$$\text{VarCov}[\theta] = \text{Inverse}[- \nabla_\theta[\nabla_\theta[J]]_{\hat{\theta}}] \qquad \text{Eqn 5}$$

Here, ∇_θ is the gradient operator with respect to θ. One may compute the correlation matrix for the parameters (denoted $\text{Corr}[\theta]$) from the $\text{VarCov}[\theta]$ matrix (see, for example, Anderson, 1958). Here, VarCov matrix contains the primary information and the Corr matrix contains redundant information. For complete discussions of MLE, see Edwards (1992) or a graduate-level text on statistics (e.g., Mood et al, 1974).

When the data include values at or below the MRL, the loglikelihood function for the model is a function of its two parameters, which in turn are functions of the explanatory variable:

$$J[\mu[P], \sigma[P]] \qquad \text{Eqn 6}$$

$$= \sum_{C_i \le \text{MRL}} \ln[\text{CDF}[\text{Normal}[\ln\text{MRL} \mid \mu[P], \sigma[P]]]]$$

$$+ \sum_{C_i > \text{MRL}} \ln[\text{PDF}[\text{Normal}[\ln C_i \mid \mu[P], \sigma[P]]]]$$

where $\mu[P]$ and $\sigma[P]$ are smooth, continuous functions of the explanatory variable, P. This loglikelihood function includes what we know about the probabilistic structure of the

sampling and laboratory procedures and excludes what we do not know about the probabilistic structure of the sampling and laboratory procedures (Swan, 1969).

6.0 Results

We used Mathematica® Version 3.0.1 for all computations and graphs in this paper (Wolfram, 1996; Wickham-Jones, 1994).

Figure 1 shows four different scatterplots of the NIRS data. In each of the four panels, vertical lines mark the boundaries between the size categories analyzed by Barry and Brattin (1998). In Panel A, the scatterplot for all $\{P_i, \ln C_i\}$ pairs, most of the points clump together at the left side of the graph. In Panel B, a second scatterplot for $\{P_i, \ln C_i\}$ pairs for smaller P_i, most of the points still clump together, even though the abscissa plots only $25 \leq P \lesssim 12{,}000$. In Panel C, the scatterplot for all $\{\ln P_i, \ln C_i\}$ pairs, the 982 points fall more evenly across the graph. The more even horizontal spacing of the vertical lines shows more clearly how Barry and Brattin grouped the data in different size categories. In Panel D, a second scatterplot for $\{\ln P_i, \ln C_i\}$ pairs for larger $\ln P_i$, only 7 points plot above $\ln P = 10$ (i.e., $P \gtrsim 22{,}000$).

Figure 2 shows the first 8 frames from a longer computer animation (Burmaster & Thompson, 1999) of many LogNormal probability plots made by sliding a vertical window across the scatterplot from left to right (Korn & Graubnard, 1998). More specifically, Frame 1 shows the first probability plot when the window selected the 60 data points with the smallest values of P. Frame 2 shows the second probability plot when the moving window selected the next 60 data points, with an offset of 12 data points to the right. The data in Frame 2 mostly overlap with the data in Frame 1, since Frame 2 is offset from Frame 1 by only a dozen points. The animation continues, with the 60-point window moving 12 points to the right for each successive frame. The frames, of course, have strong serial correlation. During this part of the analysis, we viewed many different animations, each with different parameters for the window and its movement. We experimented with windows ranging in width from 20 to 100 points, and we experimented with offsets ranging from 5 percent to 50 percent of the window's width. We show (excerpts from the) animation labeled W_j[width = 60, offset = 12] because these parameters balance the two competing goals for visualization: detail and smoothness. We also viewed many different animations at different speeds and in the forward and reverse directions. Overall, the animations reveal that the lnC values in

NIRS data have a consistent LogNormal structure for windows of different widths and offsets.

Figure 3 shows different scatterplots for the estimated parameters $\hat{\mu}_j$ (in Panels A and B) and $\hat{\sigma}_j$ (in Panels C and D) as a function of $\ln P_j$ as the window $W_j[60, 12]$ moves across the scatterplot. As expected, the pairs {median $\ln P_j$, $\hat{\mu}_j$} in Panel A show less scatter than do the pairs {median $\ln P_j$, $\hat{\sigma}_j$ } in Panel C, but the overall patterns -- and LOESS regression on the scatterplots (Cleveland, 1993) -- support these functional forms for the explanatory variables in Eqns 3 and 4:

$$\mu[P] = \alpha \exp\left[- \frac{\beta}{100} \ln P \right] = \alpha P^{-\beta/100} \qquad \text{Eqn 7}$$

$$\sigma[P] = \gamma \exp\left[- \frac{\delta}{100} \ln P \right] = \gamma P^{-\delta/100} \qquad \text{Eqn 8}$$

Following the advice in Chapter 7 of Dennis and Schnabel (1983) to increase the performance and sensitivity of the optimizer routines, we scaled the two coefficients in the exponents of Eqns 7 and 8 by factors of 100 so that all four parameters (α, β, γ, and δ) have similar magnitudes (in the decade from 1 to 10). When these equations parameterize the loglikelihood function (Eqn 6), Table 2 shows the optimal values of $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$, and $\hat{\delta}$ as a column vector, $\hat{\theta}$. From the optimization for the 982 data points, $J[\hat{\theta}] = J[\hat{\mu}[P], \hat{\sigma}[P]] = J[\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\delta}] = -1{,}454.54$. [EndNote 3] Table 2 also shows the numerical values in the matrix $\text{VarCov}[\hat{\theta}]$, along with two results computed the matrix, namely, the vector $\sigma[\hat{\theta}]$ and the matrix $\text{Corr}[\hat{\theta}]$. [EndNote 4] With n = 982 data points, the asymptotic multivariate Normal approximation from MLE theory applies (see, for example, Theorem 18 on page 359 in Mood, Graybill, and Boes (1974).

The curves in Panels A and C in Figure 3 show Eqns 7 and 8 fitted to the scatterplots with parameters $\hat{\theta}$, and the multiple curves in Panels B and D in Figure 3 show the same equations fitted to the scatterplots with parameters $\hat{\theta}$ and $\text{VarCov}[\hat{\theta}]$. In Panels A and B, respectively, the single curve and the bundle of curves for Eqn 7 provide a reasonable visual fit to all the points in the scatterplot, and in Panels C and D, respectively, the single curve and the bundle of curves for Eqn 8 provide a reasonable fit to the points in the scatterplot. In the various panels in Figure 3, any single curve represents variability in the data as a function of lnP, and the multiplicity of curves represents the parametric uncertainty in the fit to the data as a function of lnP. In Panels

B and D, as expected, the uncertainty grows larger as lnP diverges from the median $\ln P_i$.

Figure 4 presents four visualizations of the fitted LogNormal distribution as conditioned on the explanatory variable, population served. In all four panels, the dots depict the 982 original data points. In Panel A, the two curved lines represent the 50th percentile (median) and the 95th percentile of the first-order distribution plotted on lnP-lnC axes. Notice that the fitted median crosses below the MRL near lnP = 9. In Panel B, the two bundles of curved lines represent the same percentiles of the second-order distribution plotted on the same axes. In Panel C, the two bundles of curved lines represent the same percentiles of the second-order distribution plotted on P-lnC axes. Panel D plots the same bundles of curves for a shortened abscissa on P-lnC axes. In all four panels, any single curve represents the variability in the data as a function of the explanatory variable, while the multiplicity of curves represents the parametric uncertainty in the fitted distribution as a function of the explanatory variable. Note: Panels B and C present identical information, but they highlight different aspects of the fitted distribution.

Finally, the two panels in Figure 5 present the estimated PDFs for the uncertainty in the median (solid lines) and 95th percentile (dashed lines) of the fitted distribution at lnP = 10 and lnP = 12. These two graphs visualize the cross-sections that one would see at those locations in Panel B of Figure 4. Estimated by passing a kernel density across the bundles of curves at the two cross sections, the graphs quantify the large uncertainty inherent in the fitted distribution for $\ln P \gtrsim 10$.

7.0 Discussion

We demonstrate how to use the method of maximum likelihood estimation (MLE) to fit a second-order parametric distribution conditioned on a single explanatory variable to data. This method has many strengths and a few limitations. Among its strengths, this method:

- lets an analyst fit a second-order distribution to data, thus separating and quantifying the variability and the parametric uncertainty in the distribution as a function of an explanatory variable in the presence of heavy censoring.

- achieves reasonable and parsimonious fits to data. Considering the results here, the results in Table 2 provide a reasonable fit to the NIRS data in

compact and efficient form. The model also provides a way to predict [Rn222] as a function of population outside the observed range with explicit parametric uncertainty.

- works with combinations of distributions from many parametric families.

- works with censored and/or binned data, e.g., measurements reported as "nondetect" with a stated detection limit , even when each datum has a different detection limit.

- produces joint confidence regions with the proper correlations among the fitted parameters. As the number of data points grows large, the uncertainty in the model converges asymptotically to Normal theory, thus producing the joint confidence regions for the parameters as ellipsoids.

- produces results that are easily visualized and used in "two-dimensional" Monte Carlo simulations.

This analysis has one major limitation. It takes no account of spatial or geological information in the NIRS Survey data. For example, latitude, longitude, geology, and/or depth of well may have explanatory power as great as or greater than the variable P. We encourage other people to re-analyze these data using geostatistics in the future.

In this re-analysis, we have followed a maxim from computer science, "visualize, visualize, visualize!" Having reviewed all the visual evidence shown in Figures 1 through 5, we think that a LogNormal distribution conditioned on a single explanatory variable provides a reasonable fit and good insight into the NIRS data. While we cannot think of a way to use a traditional goodness-of-fit (GoF) test (D'Agostino & Stephens, 1986) on the results in this paper, we do generally follow the methods in the first four chapters of a recent book on lack-of-fit tests (Hart, 1997).

This re-analysis reveals that the NIRS data have an underlying statistical structure and simplicity not previously appreciated. A 4-parameter LogNormal distribution captures the first-order structure, and a 4×4 matrix of constants adds the second-order structure.

Further, this re-analysis has revealed that the parametric uncertainty in the median concentration falls below the MRL for $\ln P \gtrsim 9$, a result not apparent in previous analyses

(Barry & Brattin, 1998; Burmaster & Wilson, 1999). In Panel B of Figure 5, the estimated PDFs for the uncertainty in the median and the 95th percentile of variability begin to overlap by lnP = 12. These observations point to the conclusion that the NIRS data -- unless supplemented with additional measurements -- have little usefulness for community water supply systems with lnP $\gtrsim$ 10. This re-analysis re-confirms that, as a matter of [Rn222] alone, small community water supply systems deserve the greatest attention and surveillance.

Even though this re-analysis captures much of the structure of the data by quantifying the contribution of lnP as an explanatory variable, we hasten to add that Eqns 3, 4, 7, and 8 together do not form a perfect model of the data. While we think that Eqns 3, 4, 7, and 8 provide a reasonable and parsimonious model for lnC as a function of lnP, Edmund A.C. Crouch of Cambridge Environmental, Inc. has fit three alternative models to the same data and found even better fits (Crouch, 1998). Crouch's alternative models have 6, 7, and 8 parameters. The addition of each extra parameter provides a statistically significantly better fit as compared to the previous model. From Crouch's results, we learn that community water supplies with small values of lnP are best modeled by a mixture distribution with two LogNormal components, an insight that suggests avenues for future research and regulatory response. Crouch's work frames the central challenge -- to find the best statistical model for any particular number of parameters. For example, if one can find a 6-parameter model that outperforms a 7-parameter model, it is a triumph of ingenuity. [EndNote 5]

Panel B in Figure 4 neatly illustrates a fundamental property of (second-order) distributions -- the smallest uncertainty occurs near the median of the variability. In other words, the uncertainty grows larger towards either tail of the variability. In the panel, note that both the uncertainty bands shown are narrowest near lnP = 6. In addition, note that, for a fixed value of lnP, the uncertainty band for the median is narrower than the uncertainty band for the 95th percentile.

Finally, the method demonstrated with the NIRS data suggests several avenues for future research and application. In particular, we want to apply the method to fit bivariate (or multivariate) second-order distributions to appropriate datasets, and we hope that others will also apply the method to fit second-order distributions conditioned on two or more explanatory variables.

EndNotes

1. For the definition and development of second-order random variables, see Burmaster and Wilson (1996) and the references therein. For a full discussion on how to fit first- and second-order LogNormal distributions (with nonconditioned parameters) to data using maximum likelihood estimation, see Burmaster and Thompson (1998) and references therein. Basically, a second-order random variable distinguishes and separates variability and (parametric) uncertainty:

   • Variability represents true heterogeneity in the (statistical) sample or population which cannot be reduced through further measurement or study (although such heterogeneity may be disaggregated into different components associated with different subgroups in the population). Variability is a fundamental property of Nature and of the population. Variability in a population is best analyzed and modeled in terms of a probability distribution, often a first-order parametric distribution with constant parameters.

   • Uncertainty represents ignorance -- or lack of perfect knowledge -- about a phenomenon for a population as a whole which may sometimes be reduced through further measurement or study. Uncertainty is a property of the analyst performing the assessment. Uncertainty about the variability in a population can be analyzed and modeled in terms of a probability distribution, usually a second-order parametric distribution with nonconstant (distributed) parameters.

2. One may download a table containing the 982 data points in the NIRS Survey from this URL: `http://www.Alceon.com`

3. We used Mathematica's FindMinimum function to find the minimum of -J[μ[P], σ[P]] incorporating Eqns 6, 7, and 8 with these options: AccuracyGoal and PrecisionGoal each set to 11 digits and Method set to Automatic. After some initial runs to debug the code and to narrow the search, the final optimization took a few minutes on desktop computer running at 120 MHz.

4. Throughout this paper, we report the results by showing digits beyond the ones usually considered the "significant digits" for two reasons. First, for

values that have explicit standard deviations or standard errors associated with them, we present the extra digits because the standard deviations or standard errors directly show the tolerances properly associated with the results. Second, for the values in the VarCov[•] and Corr[•] matrices in a later section, we present the extra digits to prevent round-off errors in the matrix algebra from making the matrices inconsistent.

5. Statisticians often use the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC) to decide whether the increase in fit is worth the extra (computational) cost of including an additional explanatory variable or parameter in an analysis. See, for example, the discussion in Gershenfeld (1999) or Loader (1999).

Dedication

We dedicate this paper to Professor J.C.R. Licklider.

Acknowledgments

We thank Edmund A.C. Crouch of Cambridge Environmental, Inc., for first demonstrating this technique to us and for many helpful comments during the preparation of this paper. We thank Timothy M. Barry of the US Environmental Protection Agency for providing the data from the NIRS Survey and Ronald J. Bosch of the Harvard School of Public Health for giving us help and encouragement throughout this research. We also thank two anonymous reviewers for suggesting many improvements to the paper.

Alceon Corporation funded this research.

Trademarks

Mathematica® is a registered trademark of Wolfram Research, Inc., `http://www.Wolfram.com`

Alceon® is a registered trademark of Alceon Corporation, `http://www.Alceon.com`

References

Aitchison & Brown, 1957
Aitchison, J. and J.A.C. Brown, 1957, The Lognormal Distribution, Cambridge University Press, Cambridge, UK

Abramowitz & Stegun, 1964
Abramowitz, M. and I.A. Stegun, Eds, 1964, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, National Bureau of Standards, Applied Mathematics Series Number 55, Issued June 1964, Tenth Printing with corrections in December 1972, US Government Printing Office, Washington, DC

Anderson, 1958
Anderson, T.W., 1958, An Introduction to Multivariate Statistical Analysis, Wiley, New York, NY

Barry, 1998
Barry, T.M., 1998, Email message dated 28 January 1998 sent to D.E. Burmaster

Barry & Brattin, 1998
Barry, T.M. and W.J. Brattin, 1998, Distribution of Radon-222 in Community Groundwater Systems: Analysis of Type I Left-Censored Data with Single Censoring Point, Human and Ecological Risk Assessment, Volume 4, Number 2, pp 579 - 603, April 1998

Burmaster & Crouch, 1997
Burmaster, D.E. and E.A.C. Crouch, 1997, Lognormal Distributions for Body Weight as a Function of Age for Males and Females in the United States, 1976 - 1980, Risk Analysis, Volume 17, Number 4, pp 499 - 506

Burmaster & Hull, 1997
Burmaster, D.E. and D.A. Hull, 1997, Using LogNormal Distributions and LogNormal Probability Plots in Probabilistic Risk Assessments, Human and Ecological Risk Assessment, Volume 3, Number 2, pp 235 - 255

Burmaster & Thompson, 1999
Burmaster, D.E. and K.M. Thompson, 1999, Using Animated Probability Plots to Explore the Suitability of Mixture Models with Two Component Distributions, Risk Analysis, revised 22 February 1999, available as a PDF file from `http://www.Alceon.com`

Burmaster & Thompson, 1998
Burmaster, D.E. and K.M. Thompson, 1998, Fitting Second-Order Parametric Distributions to Data Using Maximum Likelihood Estimation, Human and Ecological Risk Assessment, Volume 4, Number 2, pp 319 - 339, April 1998, available as a PDF file from `http://www.Alceon.com`

Burmaster & Wilson, 1999
Burmaster, D.E. and A.M. Wilson, 1999, Fitting Second-Order Mixture Models to Data with Many Censored Values Using Maximum Likelihood Estimation, Risk Analysis, in press, available as a PDF file from `http://www.Alceon.com`

Burmaster & Wilson, 1996
Burmaster, D.E. and A.M. Wilson, 1996, An Introduction to Second-Order Random Variables in Human Health Risk Assessment, Human and Ecological Risk Assessment, Volume 2, Number 4, pp 892 - 919, available as a PDF file from `http://www.Alceon.com`

Cleveland, 1993
Cleveland, W.S., 1993, Visualizing Data, AT&T Bell Laboratories, Hobart Press, Summit, NJ

Cox & Snell, 1989
Cox, D.R. and E.J. Snell, 1989, Analysis of Binary Data, Second Edition, Chapman & Hall, London, UK

Crouch, 1998
Crouch, E.A.C., 1998, Email message dated 14 July 1998 sent to D.E. Burmaster

Crow & Shimizu, 1988
Crow, E.L. and K. Shimizu, Eds., 1988, Lognormal Distributions, Theory and Applications, Marcel Dekker, New York, NY

D'Agostino & Stephens, 1986
D'Agostino, R.B. and M.A. Stephens, 1986, Goodness-of-Fit Techniques, Marcel Dekker, New York, NY

Dennis & Schnabel, 1983
Dennis, Jr, J.E., and R.B. Schnabel, 1983, Numerical Methods for Unconstrained Optimization and Nonlinear Equations, Prentice-Hall, Englewood Cliffs, NJ

Draper & Smith, 1981
Draper, N.R. and H. Smith, 1981, Applied Regression Analysis, Second Edition, John Wiley & Sons, New York, NY

Edwards, 1992
Edwards, A.W.F., 1992, Likelihood, John Hopkins University Press, Baltimore, MD

Evans et al, 1993
Evans, M., N. Hastings, and B. Peacock, 1993, Statistical Distributions, Second Edition, John Wiley & Sons, New York, NY

Gershenfeld, 1999
Gershenfeld, N., 1999, The Nature of Mathematical Modeling, Cambridge University Press, Cambridge, UK

Hart, 1997
Hart, J.D., 1997, Nonparametric Smoothing and Lack-of-Fit Tests, Springer Verlag, New York, NY

Helsel, 1990
Helsel, D.R., 1990, Less Than Obvious, Environmental Science and Technology, Volume 24, Number 12, pp 1766 - 1774

Keeping, 1995
Keeping, E.S., 1995, Introduction to Statistical Inference, Dover, New York, NY

Korn & Graubnard, 1998
Korn, E.L. and B.I Graubard, 1998, Scatterplots with Survey Data, The American Statistician, Volume 52, Number 1, pp 58 - 69, February 1998

Loader, 1999
Loader, C., 1999, Local Regression and Likelihood, Spring-Verlag, New York, NY

Longtin, 1990
Longtin, J.P., 1990, Occurrence of Radionuclides in Drinking Water, In Radon, Radium, and Uranium in Drinking Water edited by C.R. Cothern and P.A. Rebers, Lewis Pubishers, Ann Arbor, MI

Longtin, 1988
Longtin, J.P., 1988, Occurrence of Radon, Radium, and Uranium in Groundwater, Journal of the American Water Works Association, pp 84 - 93, July 1988

Mood et al, 1974
Mood, A.M., F.A. Graybill, and D.C. Boes, 1974, Introduction to the Theory of Statistics, Third Edition, McGraw Hill, New York, NY

Swan, 1969
Swan, A.V., 1969, Computing Maximum Likelihood Estimates for Parameters of the Normal Distribution from Grouped and Censored Data, Applied Statistics, Volume 18, pp 65 - 69

Stuart & Ord, 1991
Stuart, A. and J.K. Ord, 1991, Kendall's Advanced Theory of Statistics, Fifth Edition, Volume 2, Oxford University Press, New York, NY

Weisberg, 1980
Weisberg, S., 1980, Applied Linear Regression, John Wiley & Sons, New York, NY

Wickham-Jones, 1994
Wickham-Jones, T., 1994, Mathematica Graphics, Techniques & Applications, Springer-Verlag, Telos, Santa Clara, CA

Wolfram, 1996
Wolfram, S., 1996, The Mathematica® Book, Third Edition, Wolfram Media, Champaign, IL

Type of System	Population Range	Number of Systems	$\hat{\mu}$	$\sigma(\hat{\mu})$	$\hat{\sigma}$	$\sigma(\hat{\sigma})$	$\hat{\rho}$
Very Very Small (VVS)	[25, 100)	335	5.625	0.082	1.448	0.067	-0.142
Very Small (VS)	[100, 500)	334	5.457	0.080	1.405	0.065	-0.148
Small (S)	[500, 3300)	232	4.862	0.092	1.236	0.082	-0.331
Medium (M)	[3300, 10000)	53	4.912	0.130	0.860	0.113	-0.274
Large (L)	[10000, 126000]	28	4.902	0.183	0.862	0.161	-0.312
Total		982					

Table 1
Five Size Categories for the NIRS Data Analyzed by Barry and Brattin

$$\hat{\theta} = \begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \\ \hat{\gamma} \\ \hat{\delta} \end{bmatrix} = \begin{bmatrix} 6.51126 \\ 3.61110 \\ 1.95792 \\ 6.88712 \end{bmatrix}$$

$$\sigma[\hat{\theta}] = \begin{bmatrix} 0.202713 \\ 0.528859 \\ 0.215377 \\ 1.968740 \end{bmatrix}$$

$$\text{VarCov}[\hat{\theta}] = \begin{bmatrix} 0.041092 & 0.103122 & -0.011406 & -0.107335 \\ 0.103122 & 0.279691 & -0.032201 & -0.318413 \\ -0.011406 & -0.032201 & 0.046387 & 0.409875 \\ -0.107335 & -0.318413 & 0.409875 & 3.875950 \end{bmatrix}$$

$$\text{Corr}[\hat{\theta}] = \begin{bmatrix} 1. & 0.961903 & -0.261252 & -0.268950 \\ 0.961903 & 1. & -0.282705 & -0.305818 \\ -0.261252 & -0.282705 & 1. & 0.966634 \\ -0.268950 & -0.305818 & 0.966634 & 1. \end{bmatrix}$$

Table 2: Parameters Fit by MLE

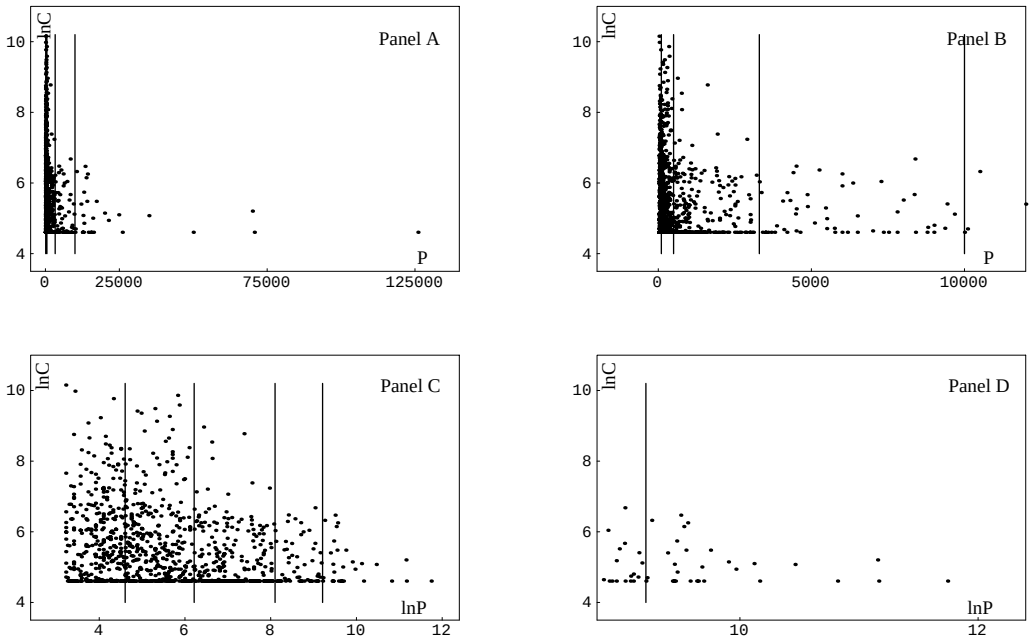

Figure 1: Panel A: Scatterplot of the $\{P_i, \ln C_i\}$ pairs with vertical lines showing the size categories analyzed by Barry and Brattin (1998) over the full range of P_i. Panel B: Same as Panel A, but for small values of P_i. Panel C: Scatterplot of the $\{\ln P_i, \ln C_i\}$ pairs with vertical lines showing the size categories analyzed by Barry and Brattin (1998) over the full range of $\ln P_i$. Panel D: Same as Panel C, but for large values of $\ln P_i$.

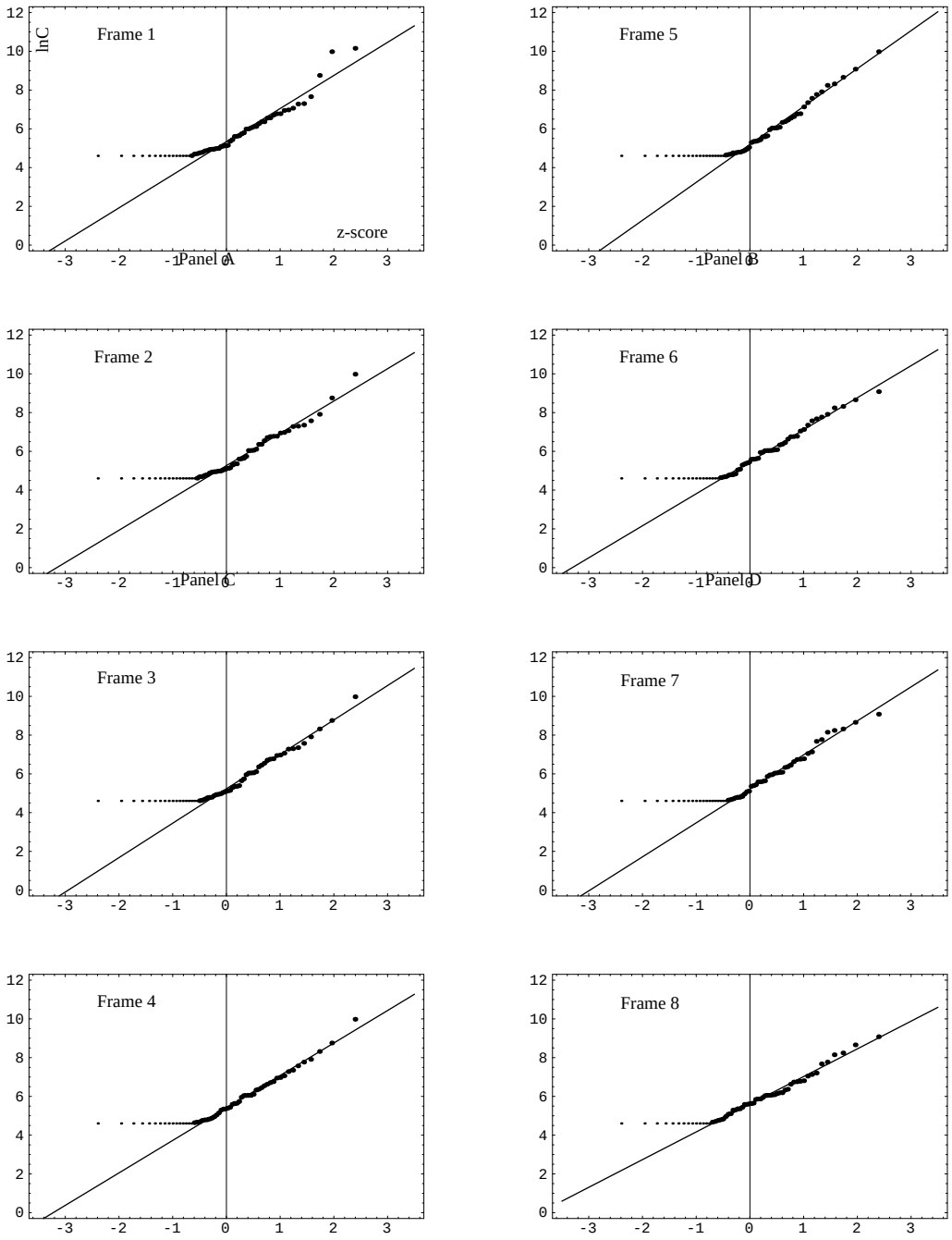

Figure 2: The first eight frames from a computer animation. Each LogNormal probability plot contains 60 data points, with each frame offset 12 data points from the previous one, starting with the lowest $\ln P_i$. The larger points show the values reported above the MRL, and the smaller points show the values reported at or below the MRL.

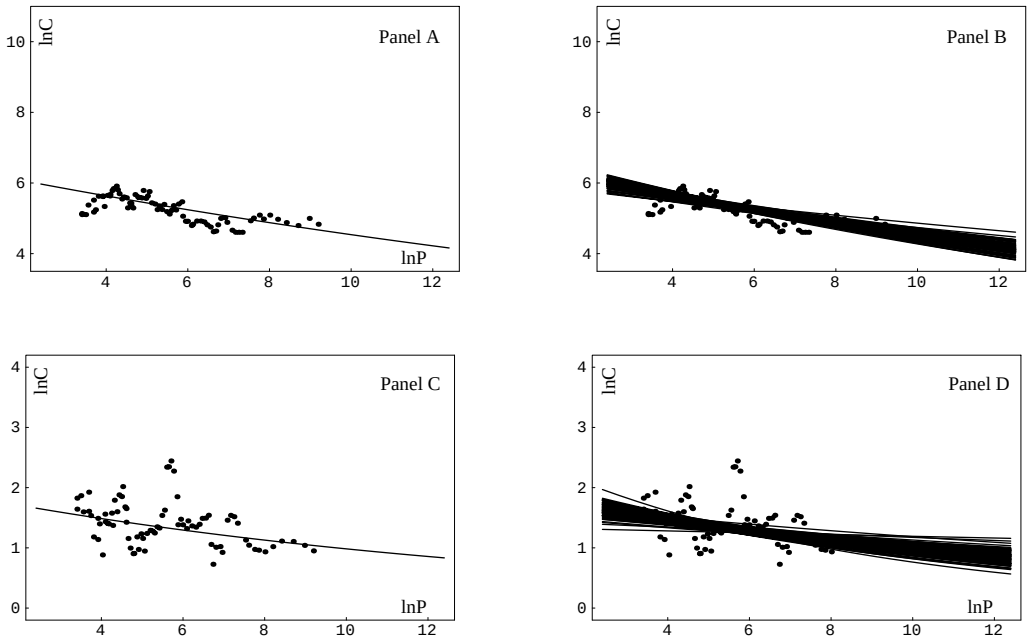

Figure 3: Panel A: Scatterplot of the {median $\ln P_j$, μ_j} pairs for window W_j[60, 12] with a line showing the median of the model fit by MLE. Panel B: Same as Panel A, but with multiple lines showing the uncertainty in the median of the model fit by MLE. Panel C: Scatterplot of the {median $\ln P_j$, σ_j} pairs for window W_j[60, 12] with a line showing the standard deviation of the model (σ) fit by MLE. Panel D: Same as Panel C, but with multiple lines showing the uncertainty in the standard deviation of the model (σ) fit by MLE.

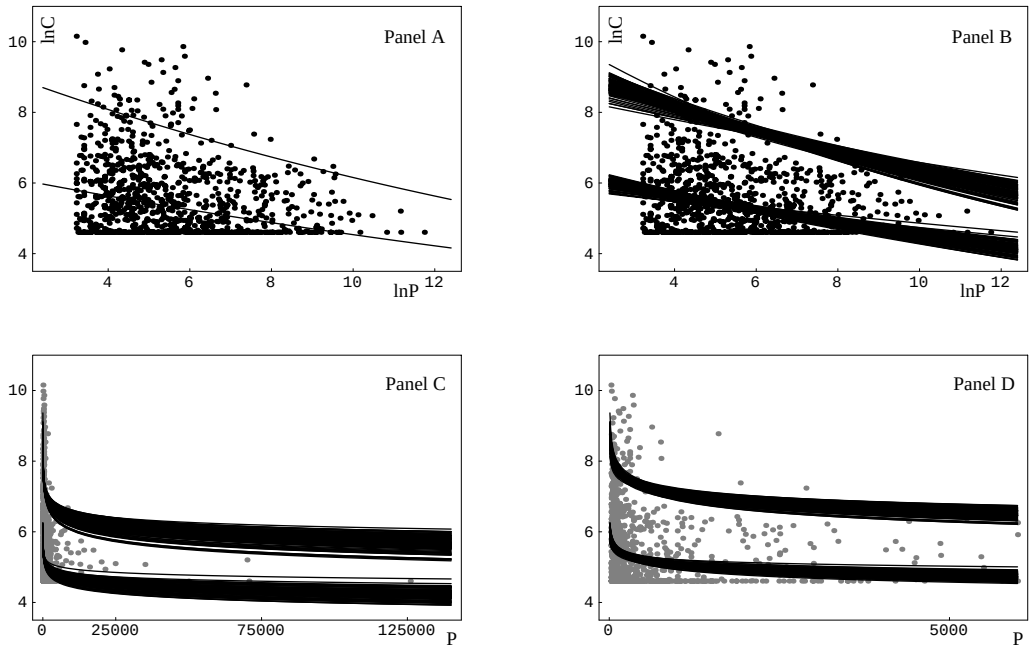

Figure 4: Panel A: Scatterplot of $\{\ln P_i, \ln C_i\}$ pairs showing the lines for the median and the 95th percentile for the model fit by MLE. Panel B: Same as Panel A, with multiple lines for the median and for the 95th percentile showing the uncertainty. Panel C: Scatterplot of $\{P_i, \ln C_i\}$ pairs over the full range of P_i showing the uncertainty in the median and the 95th percentile of the model. Panel D: Same as Panel C, but for small P_i.

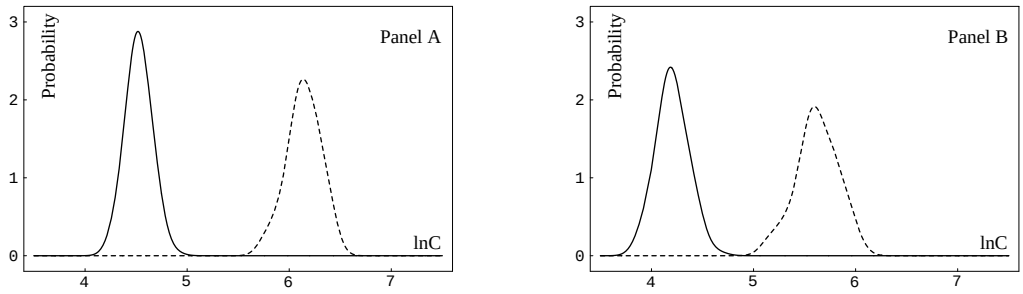

Figure 5: Panel A: Estimated distribution for the uncertainty in the median (solid line) and the 95th percentile (dashed line) of the variability at lnP = 10. Panel B: Same, but for the variability at lnP = 12.

NIRS Radon 222 Data

Lrg Sys Pop	Lrg Sys pCi/L	Med Sys Pop	Med Sys pCi/L	Sml Sys Pop	Sml Sys pCi/L	VSml Sys Pop	VSml Sys pCi/L	VVSml Sys Pop	VVSml Sys pCi/L
15320	24	6177	24	842	18	455	1	41	1
12810	30	3400	28	750	21	150	1	48	1
70661	32	7800	56	752	28	464	21	42	1
26000	53	9000	100	600	28	225	24	70	26
12500	53	3583	100	1114	33	400	25	53	31
50000	100	4200	100	1290	35	448	32	92	36
126000	100	3500	100	680	35	150	37	50	37
15528	100	5764	100	857	37	400	40	43	38
16250	100	8400	100	743	38	451	43	43	39
14500	100	8778	100	2073	38	120	55	95	40
12643	100	4629	100	2700	48	380	59	40	43
10099	110	6500	100	638	49	156	59	29	48
13000	129	3500	100	1231	50	125	59	45	51
21300	140	3800	100	793	53	270	73	65	54
16000	149	7520	100	700	60	163	100	75	56
35000	160	3688	100	1400	68	450	100	100	60
12700	161	6000	100	1300	73	210	100	60	63
24800	164	10000	100	1250	73	350	100	75	69
20000	172	7350	100	592	74	256	100	100	72
70000	182	4500	104	1862	81	380	100	78	77
12000	222	7000	104	800	95	110	100	88	85
14000	240	4100	108	1480	97	440	100	100	100
17204	240	5500	111	900	100	150	100	32	100
12953	311	9357	112	1200	100	118	100	65	100
13750	468	5750	112	1180	100	203	100	26	100
14205	520	8800	115	521	100	270	100	28	100
10500	558	3862	120	975	100	400	100	30	100
13400	646	9005	122	760	100	429	100	80	100
		5100	130	2983	100	183	100	48	100
		5500	147	2831	100	455	100	55	100
		6500	159	1683	100	240	100	50	100
		9670	167	2150	100	350	100	95	100
		4481	168	3000	100	112	100	28	100
		7800	178	1350	100	220	100	35	100
		4506	194	750	100	375	100	73	100
		5461	199	3106	100	416	100	59	100
		4870	207	800	100	280	100	31	100
		9430	223	1000	100	120	100	99	100
		4060	242	720	100	382	100	100	100
		4300	246	803	100	120	100	30	100
		8000	249	760	100	164	100	30	100
		4860	290	700	100	369	100	56	100
		8356	291	2560	100	138	100	84	100
		4203	307	860	100	107	100	100	100
		3360	308	1545	100	350	100	40	100
		6000	372	2870	100	400	100	31	100
		6350	403	700	100	200	100	60	100
		7272	421	2300	100	272	100	42	100
		6000	522	1950	100	440	100	80	100
		4400	541	550	100	350	100	80	100
		5250	583	1440	100	130	100	40	100
		4500	650	1377	100	300	100	50	100
		8383	796	700	100	312	100	28	100
				600	100	500	100	36	100
				1050	100	293	100	33	100
				1868	100	440	100	50	100
				1578	100	143	100	51	100
				2643	100	171	100	70	100
				1800	100	180	100	44	100
				1227	100	256	100	100	100
				965	100	500	100	27	100
				699	100	294	100	90	100
				860	100	182	100	100	100

780	100	500	100	33	100
1200	100	210	100	40	100
1122	100	200	100	59	100
1480	100	184	100	40	100
750	100	227	100	63	100
3063	100	116	100	60	100
1050	100	350	100	36	100
1405	100	285	100	50	100
980	100	146	100	40	100
550	100	454	100	74	100
1100	100	500	100	70	100
921	100	300	100	30	101
2250	100	399	100	100	102
1500	100	189	101	81	102
1360	100	150	101	65	102
1500	100	160	102	40	104
1400	100	220	104	58	104
2000	100	320	104	71	106
1500	100	250	105	54	106
650	100	240	106	40	107
786	100	269	106	52	107
2493	100	483	106	36	109
1512	100	273	106	26	110
1500	100	258	107	50	111
640	100	400	108	25	112
1320	100	106	109	42	117
1033	100	140	109	100	117
625	100	106	111	35	117
1000	100	350	111	100	118
1279	100	285	113	35	120
800	100	355	114	45	121
871	100	120	114	38	121
2100	101	360	115	100	122
838	101	258	115	40	126
750	102	264	116	25	129
875	105	201	117	56	129
1644	106	135	117	35	132
570	107	380	121	68	138
1382	112	450	121	30	139
1000	114	180	121	87	140
690	116	130	121	33	141
600	116	250	122	28	142
552	118	218	123	82	143
550	118	270	124	50	143
2500	118	105	125	54	144
1293	119	250	125	100	144
693	120	385	127	84	144
628	127	275	131	30	146
2449	127	132	131	27	147
564	130	400	133	60	148
639	132	170	134	99	149
1018	132	180	135	36	155
900	133	295	136	55	156
3000	133	500	137	100	157
800	135	164	138	96	157
675	136	343	140	57	157
776	136	144	140	45	161
900	137	362	140	30	162
508	138	140	141	89	164
1610	138	316	141	84	165
1761	139	200	143	50	165
2000	141	200	143	25	166
3015	142	240	144	65	166
1596	144	283	144	30	168
950	147	250	146	30	172
2000	149	283	147	72	173

868	149	141	147	96	175
900	149	409	148	75	177
2600	150	486	150	74	180
600	150	225	151	60	180
994	153	225	151	80	189
650	159	110	151	60	192
2330	160	407	152	68	193
966	165	208	152	60	194
3000	165	400	153	100	194
2200	165	500	153	55	195
782	168	212	159	96	195
1000	172	320	162	96	198
1700	173	104	162	50	199
825	174	425	169	36	199
1500	178	160	170	85	201
2150	181	305	173	50	202
905	186	209	175	90	205
2350	190	350	176	70	206
1200	198	119	176	65	208
850	200	180	176	35	210
580	200	450	176	60	212
1500	201	162	177	33	212
563	202	124	180	50	213
974	203	132	182	73	213
2500	211	260	184	49	213
600	214	130	185	60	220
1000	215	135	185	38	220
839	216	185	186	89	222
700	216	200	186	65	226
1000	217	240	189	90	229
2505	222	240	192	88	229
1200	223	200	194	40	230
890	227	102	195	25	230
1150	237	120	203	100	232
2580	239	295	203	75	232
835	243	115	205	54	244
590	244	190	206	60	244
1050	250	266	207	90	253
813	257	140	209	70	257
3000	258	160	210	100	258
2025	260	207	216	80	260
695	278	346	217	42	267
1000	279	330	218	40	268
1250	279	280	218	77	271
2325	280	500	223	76	272
2950	289	220	224	27	272
800	300	110	226	46	274
1960	312	300	227	26	275
2450	316	159	229	55	276
1300	317	108	230	40	281
1080	336	130	230	57	283
750	346	150	231	30	283
2111	355	259	235	50	285
2000	376	150	236	88	287
2506	381	270	237	60	301
2274	382	200	239	60	311
611	412	250	243	30	311
1750	413	133	245	52	311
1000	414	377	247	60	322
3300	417	200	249	71	324
828	420	133	249	90	328
2400	426	200	251	25	328
1360	429	144	253	70	330
1971	431	350	254	90	334
1200	459	235	256	80	336
2500	459	350	260	66	340

600	464	118	260	50	346
1072	467	180	266	70	348
555	479	130	269	100	349
705	480	115	269	70	351
2700	481	250	270	90	352
2300	483	110	271	48	353
1900	497	132	280	50	353
3200	503	300	282	45	381
1065	514	324	282	99	386
840	514	470	284	36	387
583	521	224	287	75	388
635	528	110	290	71	398
790	531	105	291	25	399
1400	533	252	293	25	402
520	540	375	294	73	413
530	550	393	300	45	419
1900	564	105	301	35	421
1495	575	166	306	35	422
1100	586	340	309	40	424
1000	602	120	326	30	426
1200	604	148	327	50	426
550	614	220	332	58	428
1875	620	145	336	80	432
555	648	110	341	45	435
648	702	150	352	80	435
1053	710	263	354	42	440
700	774	180	355	30	450
800	832	357	357	60	451
1095	1170	133	361	25	455
525	1250	220	367	100	464
680	1350	183	371	50	467
2891	1390	152	373	50	486
1925	1610	174	376	50	487
760	3220	216	378	100	497
753	5120	120	382	98	505
1600	6480	140	386	60	525
623	7830	177	386	25	525
		165	389	56	532
		130	397	41	561
		300	401	56	567
		200	402	78	571
		305	412	65	577
		450	424	28	579
		397	432	29	581
		500	435	40	584
		148	436	75	587
		114	438	60	588
		146	445	93	599
		350	452	80	624
		450	455	90	627
		450	465	60	631
		214	475	37	633
		120	478	100	636
		330	479	68	669
		105	484	60	670
		216	486	51	674
		120	495	54	688
		135	496	70	691
		128	504	32	700
		122	519	75	705
		120	537	25	710
		356	553	75	751
		115	580	40	758
		355	586	95	767
		387	598	85	785
		115	604	52	786

440	608	30	821
190	615	70	826
140	627	100	829
230	631	48	846
325	632	45	860
115	635	35	865
132	651	57	874
170	655	27	878
290	656	28	879
115	668	42	880
180	680	50	907
172	683	100	962
300	684	65	999
147	689	30	1040
340	704	64	1060
205	719	30	1070
160	723	65	1070
200	736	60	1080
500	766	96	1130
275	779	45	1150
145	783	30	1170
103	800	96	1240
150	802	60	1240
175	817	42	1250
233	842	65	1260
280	853	90	1320
157	854	75	1320
120	871	60	1340
145	892	50	1350
133	896	59	1370
150	903	68	1400
363	929	60	1410
102	997	30	1450
240	1010	28	1480
125	1200	98	1520
186	1220	35	1560
140	1280	50	1630
224	1320	99	1710
126	1330	75	1830
156	1550	88	1840
328	1570	53	1840
130	1610	35	1950
390	1770	50	2060
400	1810	25	2120
160	2010	45	2160
250	2040	92	2250
252	2140	80	2320
152	2210	40	2370
300	2230	75	2560
130	2600	68	2660
328	2690	84	2680
220	2700	35	2730
264	2930	100	2750
111	3110	84	2750
300	3240	60	2860
208	3260	96	3160
425	3330	66	3200
300	3600	48	3460
190	3730	72	3740
300	3890	42	3800
110	4230	36	4100
444	4370	90	4190
255	5230	90	4260
275	5730	72	4360
156	6990	69	4690
300	7300	63	4850

208	9230	43	5760
280	10600	63	6020
145	11600	30	6350
132	12300	42	8780
200	13200	56	10200
353	14600	76	17500
340	19200	31	21600
		25	25700