Undetected Components in Natural Mixtures: How Many? What Concentrations? Do They Account for Chemical Noise? What Is Needed to Detect Them?

ARTICLE *in* ANALYTICAL CHEMISTRY · MARCH 2011

Impact Factor: 5.64 · DOI: 10.1021/ac102818a · Source: PubMed

2 AUTHORS:

Chris Enke
University of New Mexico
198 PUBLICATIONS **4,582** CITATIONS
SEE PROFILE

Luc Nagels
University of Antwerp
83 PUBLICATIONS **1,056** CITATIONS
SEE PROFILE

analytical chemistry

ARTICLE

pubs.acs.org/ac

Undetected Components in Natural Mixtures: How Many? What Concentrations? Do They Account for Chemical Noise? What Is Needed to Detect Them?

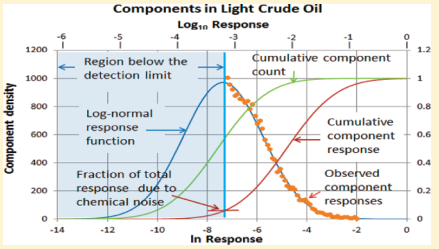
Christie G. Enke*,† and Luc J. Nagels‡

†University of New Mexico, Albuquerque, New Mexico, United States

‡Chemistry Department, University of Antwerpen, Belgium

Supporting Information

ABSTRACT: By definition, information about the set of components in a complex mixture below the detection limit is not directly available. However, if the composition of natural mixtures follows a natural law, the application of this law would enable the prediction of analytically important characteristics of that "hidden" fraction of the mixture. We have found that the analytical responses of compounds in three disparate natural mixtures (extracellular metabolites, light crude oil, and plant extracts) follow a log-normal (LN) distribution to a very high degree of correlation. Through the application of the LN model, the total number of components potentially detectable and the LN parameters of their analytical response distribution have been determined. From this distribution, one can predict the degree of analytical selectivity and dynamic range that would be required to detect any additional fraction of the components present. The data analyses of the studied mixtures reveal that the LN distribution parameters differ from one mixture type to another and that important information regarding the sample and the method employed is obtained. Further, the background level or "chemical noise" in the determinations studied agrees with the predicted cumulative responses of the undetected components. If generally applicable, the LN model will provide characterization parameters for mixture types, a means to assess completeness of analytical methods, and a model for theorists in mixture composition.



An increasingly common analytical goal is the determination of the total composition of complex natural mixtures such as cellular fluids, plant extracts, petroleum, environmental samples, and protein digests. Despite detecting hundreds (or even thousands) of the most abundant compounds, there is the sense that many more remain unseen. This perception is reinforced by the fact that at each lower concentration level the number of components appears to increase exponentially.[1] This paper offers a way to estimate the number and relative responses of the unseen parts of our samples, a quantitative evaluation of chemical noise, and insights into what is needed to detect a larger fraction of the mixture components.

By definition, information about the components in a mixture that are below the detection limit is not experimentally available. This fact is unsettling as we have no way to know what fraction of the total number of compounds present we are missing nor can we know how many more would be revealed by an increase in the resolution of component differentiation or the dynamic range of component detection. Clearly, if we could know how many components were below the detection limit and what their analytical responses would be, we would have a way to quantify analytical completeness and instrumental effectiveness for any given sample type.

This dilemma could be resolved if the concentration distribution of components in natural complex mixtures tended to follow a useful mathematical model. The lack of a verified model for the relative abundances of mixture components is a major impediment to effective developments in the theory[1a−c,2] and methodology of complex mixture analysis. However, it is reasonable to assume that such a model exists. Distribution functions of sizes or other measurable parameters in nature and in society often follow "power laws"[3] which include the Pareto distribution[4] (developed around the distribution of wealth in society) and Zipf's law[5] (analyzing the frequency of word use in literature). For these functions, a log−log plot of the density histogram is curved, i.e., not a simple exponential function which would be linear, and the distribution is said to be heavy-tailed.[6] The Parabolic fractal distribution has been fit to the size of human settlements (cities), the file size distribution of Internet traffic, the volume of oil fields, levels of gene expression,[7] and mRNA levels in cells,[8] among many others. The log-normal (LN) distribution is also heavy-tailed or skewed to the right. Whereas the normal distribution is the one we are all familiar with, LN distributions are increasingly considered nature's primary distribution function: see Limpert[9] who lists more than 60 applications of LN distributions in 10 disciplines of science, Grönholm[10] who relates skewed functions to increased entropy, and Aitchison and Brown[11] who apply the

Received: October 26, 2010
Accepted: February 1, 2011
Published: March 02, 2011

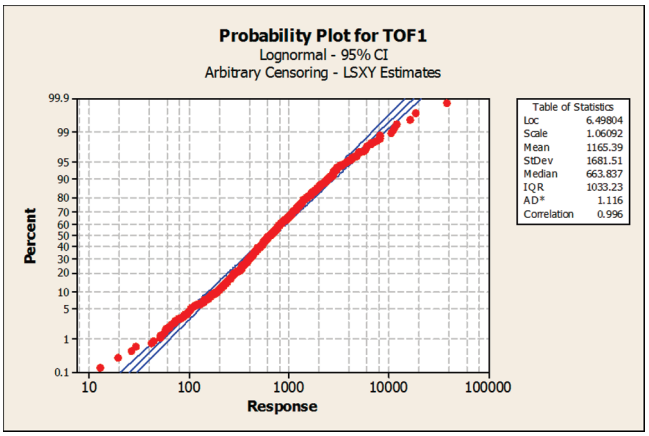


Figure 1. Minitab Q-Q plot of the responses of the Q-TOF instrument to the first metabolic data set. Data quantiles are plotted on the *Y* axis against the theoretical quantiles on the *X* axis (red dots). The high value for the correlation confirms the LN distribution function for the response values. The blue lines are the 95% confidence level boundaries indicating very high correspondence to the LN model except at the extremes. The location (Loc) and Scale parameters are measures of the width and position of the distribution as described in the text. The method for the use of the Minitab program to produce such plots is presented in the Supporting Information.

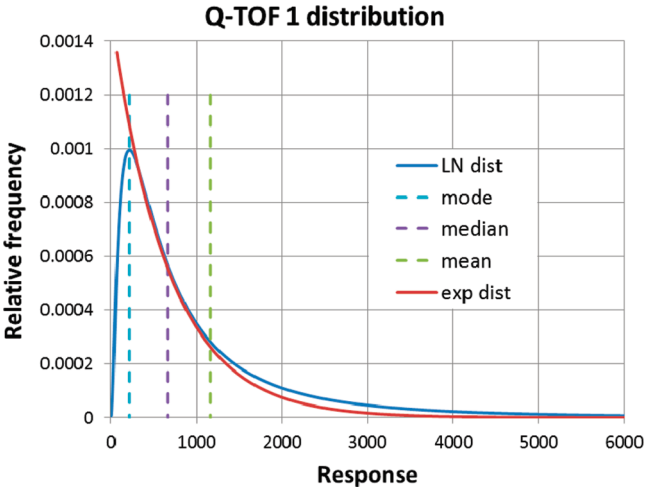


Figure 2. Blue line is the LN distribution corresponding to the parameters of the distribution fit in Figure 1. The LN distribution is skewed to the larger values when plotted on linear axes. The region of responses between 6000 and 20 000 is not shown, as the lines are indistinguishable from the *X* axis. Because of the skew, the mean, median, and mode (peak) values are all different as shown. The median and mean values were given in the legend of the Figure 1 plot. The mode is $\exp(\mu-\sigma 2)$ where μ is the Loc value and σ is the Scale value in Figure 1. An exponential distribution is also shown to illustrate the significant difference in shape between the exponential and log-normal distributions.

model to distribution of incomes and the analysis of consumer behavior.

When a concentration distribution histogram for detected components in a complex mixture is plotted, it has an exponential appearance, but the plot necessarily truncates at the detection limit and does not provide a satisfying fit to an exponential function.[12] The number of species in the decades below the detection limit cannot continue to increase indefinitely, so the plot of component density vs concentration must go through a maximum in the low concentration region and return to zero. These qualities are shared by the heavy-tailed distributions mentioned above. If a widely applicable concentration distribution model exists, how could it be applied and what could we learn from it? These critical questions are addressed in this study.

RESULTS AND DISCUSSION

In this work, we demonstrate that component responses from liquid chromatography/mass spectrometry metabolic profiling analyses,[13] from a mass spectrometric petroleum analysis[14] and from a chromatographic mixture analysis,[12] conform to a LN distribution. The very high correlation with which the analytical responses of these disparate samples match the LN distribution suggests that this distribution will be widely found in nature and that the parameters of the distribution function will vary among sample types and the compound types to which the analytical method responds. The determination of the LN distribution parameters enables the estimation of the total number of potentially detectable sample components and their analytical response distribution even for that fraction of the sample below the detection limit. The methods used to fit the data and the significance of the result is discussed for each of the three sample types individually.

Log-Normal Distribution of Metabolic Profiling Analyses. Researchers in the Theodoridis lab recently compared metabolic profiling data simultaneously generated by two different mass spectrometric detection systems following UPLC separation.[13] In this work, they generated response data for the detected components in their samples of rat urine. Results were made available to us for the current study. (Please see Acknowledgments.) For each of 10 samples, the quadrupole-time-of-flight (Q-TOF) instrument detected 689 metabolites while the hybrid triple quadrupole linear ion trap (Q-TRAP) system detected 412 metabolites.

Each of the 20 data sets resulting from the experiment described above was tested for its distribution function using the Minitab 16 statistical analysis program. In its analysis, Minitab produces a Q-Q plot (where Q stands for quantile) according to the distribution function selected. Q-Q plots are used to compare data sets with a theoretical model and to compare two theoretical distributions with each other. As such, it is an appropriate tool for testing a data set against any of the selectable distribution functions. The Q-Q plot for the metabolite responses of the Q-TOF instrument to the first sample is shown in Figure 1. The LN distribution function was selected. The distribution parameters are optimized by a least-squares minimization algorithm. The LN distribution resulted in an excellent fit with a correlation of 0.996; other tested functions (Weibull, exponential, normal, and logistic) produced Q-Q plots with very poor correlations.

While the Q-Q plots are extremely useful for testing a data set against various theoretical distributions, they do not provide a sense of what the distribution plot looks like. For that, we can use the fit parameters and the distribution equation to produce a probability density plot. Figure 2 is such a plot. The peak is skewed heavily giving rise to the long tail toward higher response values and the exponential-appearing rise as one goes from high to lower responses. Because of the asymmetry of the curve, the mean, median, and mode have different values. The legend of the

dx.doi.org/10.1021/ac102818a |*Anal. Chem.* 2011, 83, 2539–2546

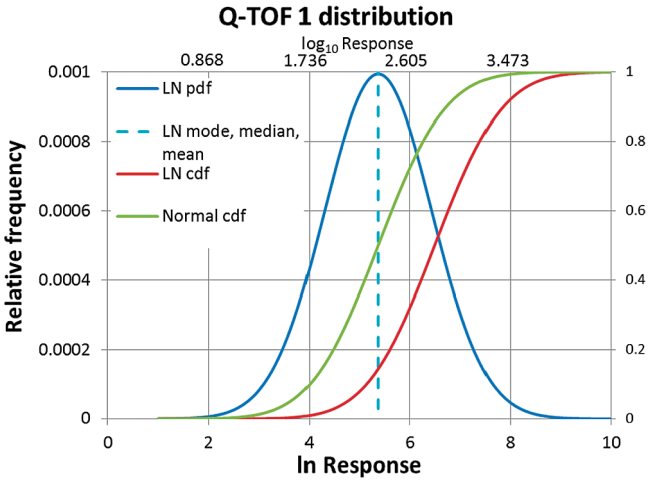


Figure 3. Changing the X axis of Figure 2 to the natural log of the response values results in a "normal" or Gaussian distribution curve (blue). In this symmetrical curve, the mode, median, and mean all have the same value (5.372) which is the mode value in Figure 2. The LN cdf curve (cumulative distribution function, red) is the cumulative fraction of responses from small to large, and the Normal cdf curve (green) is the cumulative fraction of the number of components. The equivalent $\log_{10}$ response is shown on the upper X axis. The total dynamic range of component responses is thus about 3 orders of magnitude.

Minitab data fit of Figure 1 gives the values for the mean and the median. The mode is calculated from the equation $e^{\mu-\sigma^2}$ where μ and σ are the Loc and Scale values given in Figure 1. While the scale value is related to the width of the LN function, as is σ in the normal distribution, its major effect in the LN distribution is the degree of skew, with the skew and width approaching zero as Scale approaches zero.

An exponential curve is also plotted in Figure 2. The equation used is $\lambda e^{-\lambda x}$ where λ is the single fit parameter (a value of 0.0015 was used for this plot) and x is the response. Changing the value of λ for the exponential curve results in shifting it to approximate different portions of the LN curve, but a broader match cannot be obtained. Therefore, the LN and exponential curves are significantly different, even in just the rising portion of the LN curve.

Another way to view the LN distribution is with a logarithmic axis for the response. This sort of plot is shown in Figure 3 for the same data used in Figures 1 and 2. With the natural log of the response values, the distribution is the normal distribution exactly with its usual meanings for μ and σ. This enables the application of all the familiar methods of statistical analysis to be used for the characterization of this data set and its comparison with others. The green curve is the cumulative fraction of the total number of components at each increasing natural log (ln) Response. There were 690 total components detected in this analysis, so for example, we could see that 10% or 69 of them had response values less than e^4. The red curve is the cumulative response curve which shows that the components with responses lower than the peak value account for ~15% of the total response.

There were 20 results in this study, 10 from each instrument for each of the 10 samples. All 20 results were analyzed using Minitab in the same way as the sample Q-TOF 1. The results of these analyses are given in Table 1. In this table, we observe that the fit parameters are very similar for all 10 samples. We also see that, for each sample, there is considerable difference between the values obtained from the Q-TOF instrument and the Q-Trap instrument. The differences in the Loc values are simply due to the response factor of the instruments. This is confirmed by the fact that the ratio of the Loc values, given in the last column, is essentially constant. The unexpected aspect of this analysis is the difference in the Scale values between the two instruments. As described above, the Scale value in the LN distribution is equal to the σ value in the normal distribution in the semilog plot. In other words, there is a difference in the dynamic range of the component responses between the two instruments. Since they were both responding to the same samples at the same time, this difference must be ascribed to a systematic difference in the component response factors between these two systems. A simple multiplicative factor would not explain these results, so it would seem that a nonlinearity in response of one or both of these systems may be involved. Further investigation would be required to determine the cause of the nonlinear response, but the fact that it exists is clearly revealed by this analysis. This conclusion adds another aspect of comparison to the differences in metabolite identification presented in the Theodoridis group paper.[13]

Log-Normal Concentration Distribution of Components in Light Crude Oil. A data set of compounds detected by very high resolution mass spectrometry in a light crude oil sample was provided to us by researchers in Alan Marshall's lab at Florida State University. (Please see Acknowledgments.) Altogether, compounds with 18 364 different chemical formulas were detected and arranged in order of response intensity. See Table 2. From this table, one can see that the number of components detected continues to increase as the response value decreases. This means that only the high side of the distribution curve has been sampled; the falloff in component density at the low response end is missing. Minitab is capable of handling missing or "censored" data at either end of the distribution, but the number of components for which the response is missing must be provided. This number is not experimentally available, so it will have to be estimated by other means. Fortunately, we have devised a method for doing this.

One of the qualities of the LN distribution is that, when histogram data are graphed as a log−log plot, the result is a parabolic curve. We call this the LLP plot for log−log parabolic. From the parameters of the quadratic equation that fit this LLP curve, the parameters of the LN distribution can be obtained. The algebraic derivations are given in the Supporting Information. From the LN distribution, the number of missing components can be estimated and a Minitab analysis can be employed. This process is illustrated below for the data in Table 2.

To produce the LLP curve, the values in column 2 were plotted (Y axis) against the values in the fourth column (X axis) using Excel as shown in Figure 4. The R^2 value for the fit to a parabolic function ($ax^2 + bx + c = 0$) is a remarkable 0.9945. As proven in the derivation in the Supporting Information, the relationships between a, b, and c in the LLP equation and the LN distribution parameters are

$$\sigma = \left(-\frac{1}{2a}\right)^{1/2} \qquad \mu = -\frac{b+1}{2a}$$

$$A = \sqrt{\frac{\pi}{-a}} \exp\left(\frac{4ca - b^2 - 2b - 1}{4a}\right) \qquad (1)$$

dx.doi.org/10.1021/ac102818a |Anal. Chem. 2011, 83, 2539–2546

Table 1. Results of Minitab Analysis of the Metabolomics Response Data Fits

#	Scale, TOF	Scale, Trap	Loc, TOF	Loc, Trap	Corr, TOF	Corr, Trap	Loc Trap/Loc TOF
1	1.061 ± 0.054	0.831 ± 0.054	6.498 ± 0.079	17.013 ± 0.080	0.996	0.988	2.618
2	1.086 ± 0.055	0.866 ± 0.057	6.738 ± 0.081	17.154 ± 0.084	0.993	0.993	2.546
3	1.173 ± 0.060	0.918 ± 0.060	6.655 ± 0.088	17.036 ± 0.089	0.994	0.992	2.560
4	1.101 ± 0.056	0.879 ± 0.058	6.846 ± 0.082	17.210 ± 0.085	0.997	0.995	2.514
5	1.122 ± 0.057	0.898 ± 0.059	6.688 ± 0.084	17.108 ± 0.087	0.994	0.993	2.558
6	1.162 ± 0.060	0.880 ± 0.058	6.759 ± 0.087	17.153 ± 0.085	0.997	0.994	2.538
7	1.156 ± 0.059	0.872 ± 0.057	6.624 ± 0.086	17.073 ± 0.084	0.992	0.994	2.577
8	1.166 ± 0.059	0.874 ± 0.057	6.598 ± 0.087	17.051 ± 0.084	0.993	0.994	2.584
9	1.115 ± 0.057	0.836 ± 0.054	6.640 ± 0.083	17.121 ± 0.081	0.994	0.990	2.579
10	1.116 ± 0.057	0.883 ± 0.058	6.828 ± 0.083	17.234 ± 0.085	0.994	0.992	2.524
Avg.	1.155 ± 0.149	0.874 ± 0.057	6.687 ± 0.084	17.115 ± 0.084	0.994	0.993	2.560

Table 2. Light Crude Data from FSU[a]

1	2	3	4	1	2	3	4
ln hi	ln lo	#/0.1 ln	ln (#/0.1 ln)	ln hi	ln lo	#/0.1 ln	ln (#/0.1 ln)
−1.9	−2	3	1.09861229	−5	−5.1	358	5.88053299
−2	−2.1	14	2.63905733	−5.1	−5.2	335	5.81413053
−2.1	−2.2	9	2.19722458	−5.2	−5.3	424	6.04973346
−2.2	−2.3	10	2.30258509	−5.3	−5.4	434	6.07304453
−2.3	−2.4	15	2.7080502	−5.4	−5.5	497	6.20859003
−2.4	−2.5	15	2.7080502	−5.5	−5.6	529	6.27098843
−2.5	−2.6	12	2.48490665	−5.6	−5.7	565	6.33682573
−2.6	−2.7	11	2.39789527	−5.7	−5.8	644	6.46769873
−2.7	−2.8	22	3.09104245	−5.8	−5.9	700	6.55108034
−2.8	−2.9	15	2.7080502	−5.9	−6	679	6.52062113
−2.9	−3	21	3.04452244	−6	−6.1	736	6.60123012
−3	−3.1	33	3.49650756	−6.1	−6.2	818	6.70686234
−3.1	−3.2	30	3.40119738	−6.2	−6.3	784	6.66440902
−3.2	−3.3	36	3.58351894	−6.3	−6.4	839	6.73221071
−3.3	−3.4	45	3.80666249	−6.4	−6.5	830	6.7214257
−3.4	−3.5	49	3.8918203	−6.5	−6.6	851	6.74641213
−4.5	−4.6	227	5.42495002	−6.6	−6.7	854	6.74993119
−4.6	−4.7	218	5.38449506	−6.7	−6.8	890	6.79122146
−4.7	−4.8	238	5.47227067	−6.8	−6.9	879	6.7787849
−4.8	−4.9	277	5.62401751	−6.9	−7	924	6.82871207
−4.9	−5	307	5.72684775	−7	−7.1	958	6.86484778

[a] The natural logs of the magnitudes were calculated, and the numbers of components in each 0.1 range (columns 1 and 2) of ln magnitude were counted (column 3).

From the *a*, *b*, and *c* values of the parabola (−0.1971, −2.9034, and −3.8134), we used eq 1 to estimate the parameters (μ = −4.828, σ = 1.5927, and *A* = 8.726) for the LN function

$$p(\ln x) = \frac{A}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) \qquad (2)$$

where *x* is the response and *A* is a factor to adjust for the total area under the curve. The resulting log-normal distribution is plotted on a semilog graph along with the response data and is shown in Figure 5.

Because of the mathematical equivalence of the LLP and LN functions, the fit of the data to the LN curve is the same as it was

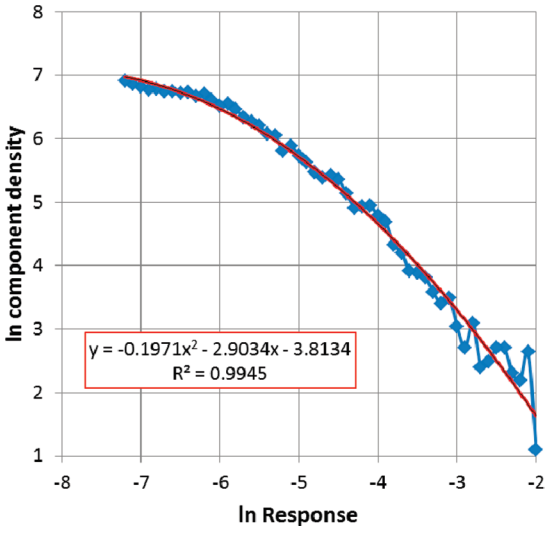


Figure 4. Ln-ln plot of the petroleum data in Table 2. Because of the high value of fit and the equivalence of the LLP and the LN distributions, an LN distribution for this data set is indicated.

in Figure 3. The importance of this plot is that it tells us what fraction of the total components present were in the original data set. The data end at the peak value of the LN pdf curve. As the normal cdf curve tells us, this is at the 50% point in the cumulative number of components. There were 18 364 different chemical formulas detected, so the number of missing results can be estimated at 18 364. This is the number needed for the Minitab analysis of the data.

The resulting Minitab plot is shown in Figure 6. The correlation value for the fit is an excellent 0.998. This is a strong indication that the data in the region plotted has a log-normal distribution. In this plot as well, the values for the lower responses are vacant. It is, therefore, an extrapolation to assume that they will follow the log-normal distribution as well. The excellent fit and the fact that the values are not deviating from the fit boundaries at the low response end of the curve provide some level of confidence in this assumption. The Loc value, which is the mode for the LN distribution is −7.19838. This differs slightly from that derived by the LLP plot approach of $\mu - \sigma^2 = -7.365$. The Scale value of 1.5601 also differs from the value of 1.5927 calculated from the LLP plot parameters. These differences come from the fact that the LLP curve is minimizing the differences in

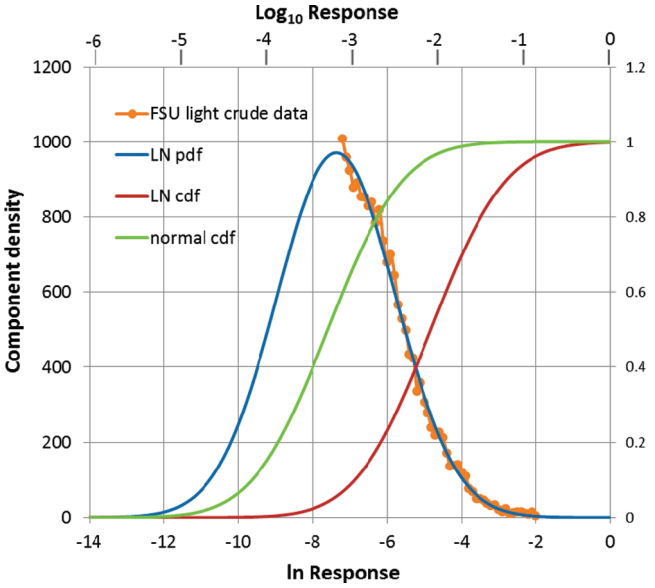


Figure 5. Petroleum density data (orange ●) are plotted on a log-normal scale and the LN function (blue) using the values $\mu = -4.828$, $\sigma = 1.5927$, and $A = 8.726$ as derived from the LLP plot and eqs 2. This plot also includes the PF plot (green) using the values in this figure. Even though optimized differently, the LN and PF plots are very nearly the same. The deviations at the low concentration end are likely due to a decrease in abundance precision in this region. The normalized cumulative (response) curve (red, right axis) shows that the observed compounds account for all but 5% of the total signal while the normalized count integral (green, right axis) reveals that 50% of the compounds present have been detected.

the logs of the density while the Minitab fit is minimizing the differences in the linear density values. The fit obtained by Minitab is, therefore, more true to the LN distribution function. Furthermore, the Minitab analysis provides values for the correlation and confidence intervals for the Loc and Scale values. The value of the LLP analysis was to provide an estimate of the number of missing data points. Its disadvantage is that it requires enough data points from which to produce data in the histogram form.

There is much to learn from this analysis of the crude oil data. We can see that the dynamic range of the technique is just over 2 orders of magnitude, roughly half that of the range of component concentrations. We can predict that the total number of compounds having different chemical formulas is about twice the 18 364 different formulas detected. From the LN cdf curve of Figure 5, we can see that the compounds undetected will produce a cumulative signal which is only 5% of the total response. Because of their large number and small response, they are likely to be distributed widely over the data space and produce a background continuum. The experimental value for this background signal, as reported to us, is estimated at 5%. A reasonable hypothesis is that the unseen components have the predicted LN distribution and that their cumulative signal comprises the background level observed. If this is true, this is the first quantitation of chemical noise and its correlation with the background signal in survey analyses.

Another piece of information we can glean from this analysis is that the measurement quality that prevents the detection of a greater fraction of the compounds is dynamic range, not resolution. To detect virtually all the compounds with different formulas, the dynamic range would need to be increased 100-fold.

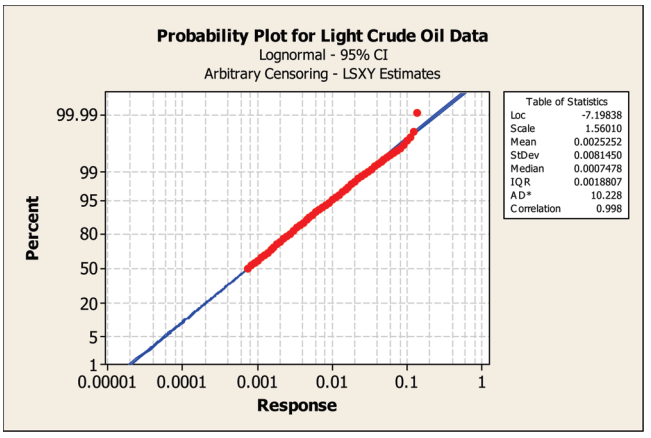


Figure 6. Minitab plot of the light crude oil data. The statistical analysis of the fit parameters at the 95% confidence level are Loc = −7.19838 ± 0.02 and Scale = 1.56010 ± 0.018. The Minitab fit can be done by least-squares (this plot) or by maximum likelihood. The Loc and Scale values from the maximum likelihood fit are −7.2027 and 1.5535 which are within the confidence limits of the least-squares fit.

No limitation from the number of independent resolution channels is expected or observed. The number of different compounds in these samples is larger even than the 36 000 different formulas predicted. This number would be multiplied by the average number of isomers of each formula and increased by the number of compounds that are not ionized or that are outside the mass range studied.

Fitting the chromatographic data. In an earlier article,[12] Nagels et al. determined the histogram of relative peak areas (as a % of the total area of the chromatogram) from a data set comprising HPLC chromatograms of 62 plant extracts. In each chromatogram, the number of components greatly exceeded the chromatographic peak capacity, resulting in substantial peak overlap. It was clear that most of the observed peaks were composed of several (equal, but more often smaller) underlying "component" peaks. The statistics of peak overlap have been well developed as to position[15] but not intensity.[1a,2] On the basis of these statistics, computer simulations were performed, where the cumulative distribution function of underlying "component" peaks was iteratively constructed. However, the computer simulation could not estimate the total number of compounds or the shape of the response distribution in the lower concentration regions.

In order to use the data in Table 1 of the Nagels' paper with the statistical analysis methods used in the previous two examples, we first calculated the relative density of calculated peaks for each of the peak area boundaries. The original data and the details of the calculations involved are given in the Supporting Information. Then, the ln of the peak density was plotted against the ln of the average peak area in each peak area range. The results are shown in Figure 7. The curve fit demonstrates that the calculated peak densities cover only the higher portion of the distribution. We estimated that ∼2/3 of the detectable components were below the lowest peak area data point. This is an even larger fraction of undetected components than for the light crude response distribution shown in Figure 5. Thus, in order to use Minitab to test for the distribution type and obtain its parameters, we had to get an estimate for the fraction of the components that are not represented in the data set.

First, we use the *a*, *b*, and *c* parameters in the LLP fit to calculate estimated LN parameters ($\mu = -1.61296$, $\sigma = 1.500$, and

dx.doi.org/10.1021/ac102818a |*Anal. Chem.* 2011, 83, 2539–2546

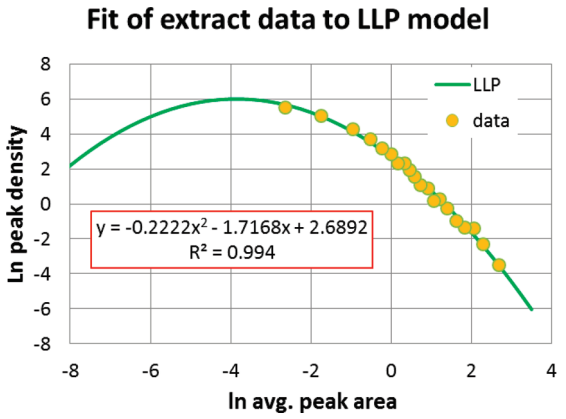


Figure 7. Plant extract data were fit to a ln-ln parabolic curve using Excel. The fit equation and R^2 value are shown, indicating a very good fit. These values were used to produce the LLP curve shown by the green line. Data from the five largest peak area boundaries were not included in this plot. The number of large peaks is few, giving a poor statistical density, but on the log scale, they heavily influence the result.

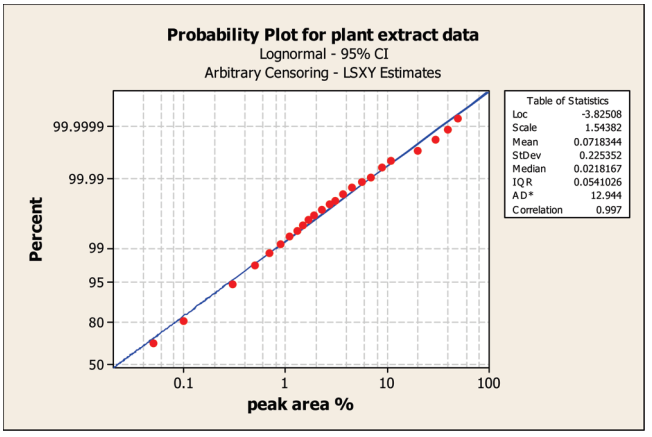


Figure 8. Minitab plot for the extract data. All data points were used in this analysis. This plot provides a more reliable estimate of the LN parameters than the LLP plot in Figure 7. The Loc is 3.82508 ± 0.004, and the Scale is 1.54382 ± 0.003. The high correlation of 0.997 is highly supportive of the LN distribution hypothesis over the data range plotted.

$A = 4.41837$) and plot the LN curve. A plot of the LN pdf and cdf for these parameters provided an estimate of the fraction of components missing of 67%. This estimate provided the number to use with the Minitab analysis for the missing components. The resulting Minitab plot is shown in Figure 8. Again, the fit parameters strongly support the LN distribution over the range of the data. The confidence intervals for the Loc and Scale are very small, giving a high degree of confidence in these results.

Finally, the Loc and Scale values were used to produce a more exact LN plot than the tentative one (not shown) used to obtain an approximation of the fraction of missing components. This more exact plot is shown in Figure 9. The data are shown as *x*'s. As one can see, the fit of the data to the LN pdf curve is excellent, strongly supporting the LN distribution over the range of the data points.

What we can learn from this plot, assuming the LN model applies over the entire response range, is that the peak areas cover a dynamic range on the order of 10^4 and that roughly 70% of the components were below the range over which the number and

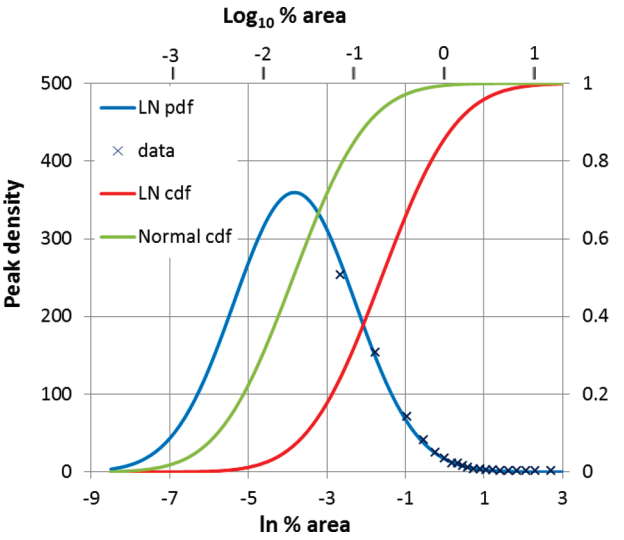


Figure 9. LN plot of the plant extract data using the LN distribution parameters obtained from the Minitab fit of Figure 7. It confirms the excellent fit of the data to the LN curve (blue). The plot indicates that the response distribution has a range just exceeding 4 orders of magnitude. Just as in Figure 5, the red LN cdf curve is the cumulative curve of peak areas while the green Normal cdf curve is the cumulative curve of number of peaks.

areas of the peaks were calculated. (The 70% value is obtained from the point where the Normal cdf line crosses the ln% area of −3, the value of the ln of the smallest peaks in the data set.) This corresponds very well with the 67% value that was assumed from Figure 6 and which gave rise to the number of missing components for the Minitab analysis.

The calculation of the number of components (detectable by the methodology employed) under the peak requires an additional bit of data which is the actual count of detected components that comprise some portion of the high concentration end of the curve. In this case, it was determined that there were 23 components with a response greater than 1%.[12] We then determined what fraction of the distribution curve area lies to the right of the 1% mark on the concentration or response axis of Figure 9. In this case, that fraction is 0.025. If there are 23 components in this fraction of the histogram, the total number of potentially detectable components in each chromatogram is approximately 920.

As the component peaks get smaller and more numerous, their responses would spread randomly across the retention time axis and the sum of their unresolved responses would produce a background signal. In the case of the plant extract data, the background level comprised 45% of the total response. From Figure 9, the ln % area at the 0.45 point on the LN cdf curve occurs at −1.75. This corresponds to an area of 0.174% of the total response. From Table 1 in the Nagels paper, it is in this region where the direct observation of individual peaks was no longer possible. Here again, the chemical noise or background signal is accounted for by the sum of the responses of the undetected components, and again, the agreement between the calculated sum of responses and the level of background supports the LN distribution hypothesis over the region of undetected components.

CONCLUSIONS

From the fact that the distribution of analytical response values appears to follow the LN distribution for three quite

disparate natural samples (all that we tested), it is reasonable to conclude that the LN distribution may well be generally applicable. We invite and encourage others with response data sets for complex mixtures to expand the number and types of mixtures that have been tested for this property. On the basis of the differences in the results obtained for the mixtures studied, we expect that various mixtures will have different statistical parameters and that these parameters could become a useful characterization of mixture complexity and dynamic range.

We have also seen that the production of the distribution plots such as Figures 3, 5, and 9 enables an assessment of the completeness of the survey analysis and indicates whether an increase in resolution, dynamic range, or both is needed to further extend the analysis. Further, one could predict the degree of improvement that would be achieved by an increase in resolution or dynamic range. On the basis of the differences in the two instruments responding to the same metabolic samples, it also appears that these statistical analyses will be useful in characterizing the analytical method employed as well.

The implementation of the LN statistical analysis is quite accessible as outlined in the Supporting Information. If the total number of responses is in the low thousands, the response data can be entered into Minitab directly. Only when a significant fraction of the components is missing does one need to have a way to estimate this number. In this case, the LLP plot method is also easy to implement and should not be a deterrent to obtaining the useful information it can provide.

The chromatographic peaks in the Nagels data were very heavily overlapped. Besides the problem of lesser peaks being hidden by more prominent peaks, the responses of the observed peaks included the areas of the hidden peaks they overlapped. In Nagels work, these errors in count and amplitude were teased out of the chromatograms using extensive computer simulation and peak overlap theory. This was necessary because the peak capacity was insufficient to resolve the components which were above the detection limit and the overlap distorted the observed peak areas. Potential component overlap needs to be taken into account before fitting new data to the LN model. As shown in the cases of the petroleum and metabolite samples using techniques with much greater resolving power, the implementation of the log-normal curve fit can be very straightforward.

Selectivity is part of any analytical technique in that there will always be classes of compounds which are excluded by sample preparation or instrument response. In this case, when the LN analysis is employed to determine the total number of components present in a sample, it will necessarily be just the total number of compounds to which the technique employed could have responded.

We believe, on the basis of these three sample types and the widespread occurrence of the LN distribution in nature, a wide range of complex natural mixtures will prove to have log-normal response distributions. If that proves to be the case, mixture types could be characterized by the analytical challenges they present, analytical methods could be quantitatively evaluated by the degree of completeness they are able to achieve for a given application, scientists studying component overlap and interference will have a solid model for mixture composition with which to work, and it will support inquiry into the processes by which such mixtures are generated in nature.

These extensions of the model will raise the question of the extent to which the response distribution and the concentration distribution are related. In most analytical applications, the instrumental response value or peak area calculation is proportional to the concentration of the analyte. Of course, the response factor may be significantly different for each analyte. Assuming the response factors are not correlated to the concentration (each working curve is linear), the shape of the response distribution and the concentration distribution should be essentially the same. Therefore, it is reasonable to assume that the response and concentration distributions for any given complex sample will be nearly equivalent.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information. Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Address: 33 Vista de Oro, Placitas, NM 87043. E-mail: enke@unm.edu.

■ ACKNOWLEDGMENT

The crude oil FTICR MS data were given to us by Amy M. McKenna, Ryan P. Rodgers, and Alan G. Marshall of Florida State University, supported by NSF Division of Materials Research through DMR-06-54118, NSF CHE-10-49753, and the State of Florida. The metabolite data were supplied by Georgios Theodoridis, Aristotle University, Thessaloniki, Greece, and Ian D Wilson, AstraZeneca, Alderley Park, Cheshire, U.K. We also thank Joe Davis (Southern Illinois Univ.), Peter Wentzell (Dalhousie Univ.), and John Engen (Northeastern Univ.) for comments on an early draft manuscript.

■ REFERENCES

(1) (a) Fellinger, A.; Pietrogrande, M. C. *Anal. Chem.* **2001**, *73* (21), 618A–626A. (b) Dondi, F.; Bassi, A.; Cavazzini, A.; Pietrogrande, M. C. *Anal. Chem.* **1998**, *70* (4), 766–773. (c) El Fallah, M. Z.; Martin, M. *Chromatographia* **1987**, *24*, 115–22. (d) Dondi, F.; Pietrogrande, M. C.; Felinger, A. *Chromatographia* **1997**, *45*, 435–440. (e) Davis, J. M.; Arriaga, E. A. *Anal. Chem. (Washington, DC, U. S.)* **2010**, *82* (1), 307–315. (f) Dondi, F.; Kahie, Y. D.; Lodi, G.; Remelli, M.; Reschiglian, P.; Bighi, C. *Anal. Chim. Acta* **1986**, *191*, 261–73. (g) Herman, D. P.; Gonnord, M. F.; Guiochon, G. *Anal. Chem.* **1984**, *56* (6), 995–1003.

(2) Wentzell, P. D.; Vega Montoto, L. *Chemom. Intell. Lab. Syst.* **2003**, *65* (2), 257–279.

(3) Searls, D. B. *Nature* **2002**, *420* (14 November 2002), 211–217.

(4) Pareto, V. *Cours d'Economie Politique*. Droz: Geneva, Switzerland, 1896.

(5) Zipf, G. K. *The psycho-biology of Language: an Introduction to Dynamic Philology*; Rutledge: London, 1936.

(6) Mitzenmacher, M. *Internet Math.* **2004**, *1* (2), 226–251.

(7) Hoyle, D. C.; Rattray, M.; Jupp, R.; Brass, A. *Bioinformatics* **2002**, *18* (4), 576–584.

(8) Bengtsson, M.; Stahlberg, A.; Rorsman, P.; Kubista, M. *Genome Res.* **2005**, *15*, 1388–1392.

(9) Limpert, E.; Stahel, W. A.; Abbt, M. *Bioscience* **2001**, *51* (5), 341–352.

(10) Grönholm, T.; Annila, A. *Math. Biosci.* **2007**, *210*, 659–667.

(11) Aitchison, J.; Brown, J. A. C. *The Lognormal Distribution with special reference to its uses in ecomomi*; Cambridge University Press: Cambridge, U.K., 1957; p 176.

(12) Nagels, L. J.; Creten, W. L.; Vanpeperstraete, P. M. *Anal. Chem.* **1983**, *55* (2), 216–20.

(13) Gika, H. G.; Theodoridis, G. A.; Earil, M.; Snyder, R. W.; Sumner, S. J.; Wilson, I. D. *Anal. Chem. (Washington, DC, U. S.)* **2010**, *82* (19), 8226–8234, please see Acknowledgements.

(14) McKenna, A. M.; Rodgers, R. P.; Marshall, A. G. *FTMS responses at detected m/z values for a sample of light crude petroleum*, private communication; please see Acknowledgment.

(15) (a) Davis, J. M. *J. Chromatogr., A* **1999**, *831* (1), 37–49. (b) Davis, J. M.; Giddings, J. C. *Anal. Chem.* **1983**, *55* (3), 418–24. (c) Martin, M.; Guiochon, G. *Anal. Chem.* **1985**, *57* (1), 289–95.