The Statistics Of Super-Emitters: Modeling Heavy-Tailed Datasets As Power Laws

Marc Mansfield
Utah State University, Vernal, Utah 84078
marc.mansfield@usu.edu

ABSTRACT

Many observational datasets of emissions, including emissions from the oil and gas sector, follow heavy-tailed distributions, for which a small fraction of the measured emitters are much larger than more typical emitters, and account for a large fraction of the total measured emission. Such distributions are problematic, because they are expected to exhibit large, negatively biased sampling errors. As a result, there are very real concerns that current bottom-up emissions inventories may underestimate the true emission. Based on the Generalized Central Limit Theorem, there are good reasons for expecting such datasets to obey power-law distribution functions, and I have been able to fit a number of datasets of methane in the environment with such distribution functions. Analyses of three datasets, comprising methane emissions from abandoned oil and gas wells in Pennsylvania, methane in soil gas near coal bed methane wells in Utah, and methane dissolved in ground water in West Virginia, respectively, are presented here. I have also developed an error-analysis algorithm for such distributions. In calculations on artificial datasets, I have verified that the error-analysis algorithm works well. Unfortunately, it may require information about the underlying distribution that may not be available in real-world applications. Overcoming this drawback is a current focus of my research.

INTRODUCTION: HEAVY-TAILED DATASETS PROVIDE SPECIAL CHALLENGES FOR ESTIMATING EMISSIONS FROM THE OIL AND GAS SECTOR

Figures 1 – 3 represent three datasets of methane measurements in the environment. Figure 1 shows emissions measured from abandoned oil and gas wells in Pennsylvania [Kang, et al., 2014]. Each vertical bar represents an individual measurement, of which there are a total of 38, arranged in order from smallest to largest. The measurements extend over several orders of magnitude, so the y-axis is logarithmic. Some wells, represented at the left end of the chart, showed very little methane leakage, considerably less than 1 mg/hr. However, other wells were detected to have leakage rates approaching 10^5 mg/hr. The four highest wells, or about 10% of the total, are responsible for about 95% of the total leakage. The median emission is 29 mg/hr, while the mean is over 6000 mg/hr. The huge difference between mean and median occurs because a few wells at the high end of the dataset dominate the total.

Figure 2 shows methane concentrations in the soil gas near coal-bed methane wells in Utah [Stolp, et al., 2006]. These measurements were taken by inserting probes into the soil within 1 to 2 m of the wellhead, and to a depth of about 1 to 2 m. Again, each vertical bar represents an individual measurement. The bars in pink represent measurements assigned to the "tail" of the

dataset and have been isolated for further analysis. The bars in light blue represent measurements not included in the tail. Measurements in the tail extend from 6 ppm to over 60,000 ppm. (The broad swath of measurements at 5 ppm were actually reported as "< 10 ppm," the detection limit of many of the earlier measurements. Since they have been excluded from the tail they have no impact on the subsequent analysis.) Again, the mean is much larger than the median.

Figure 3 shows concentrations of methane in the ground water in West Virginia, as measured in existing water wells [White and Mathes, 2006]. Again, the pink bars represent the "tail" of the measurement set. The "non-tail" measurements are again represented in light blue, and for this dataset, were all reported as 0. Since 0 cannot be represented on a log-plot, those particular data points have been assigned to their own axis, outside of the bar chart.

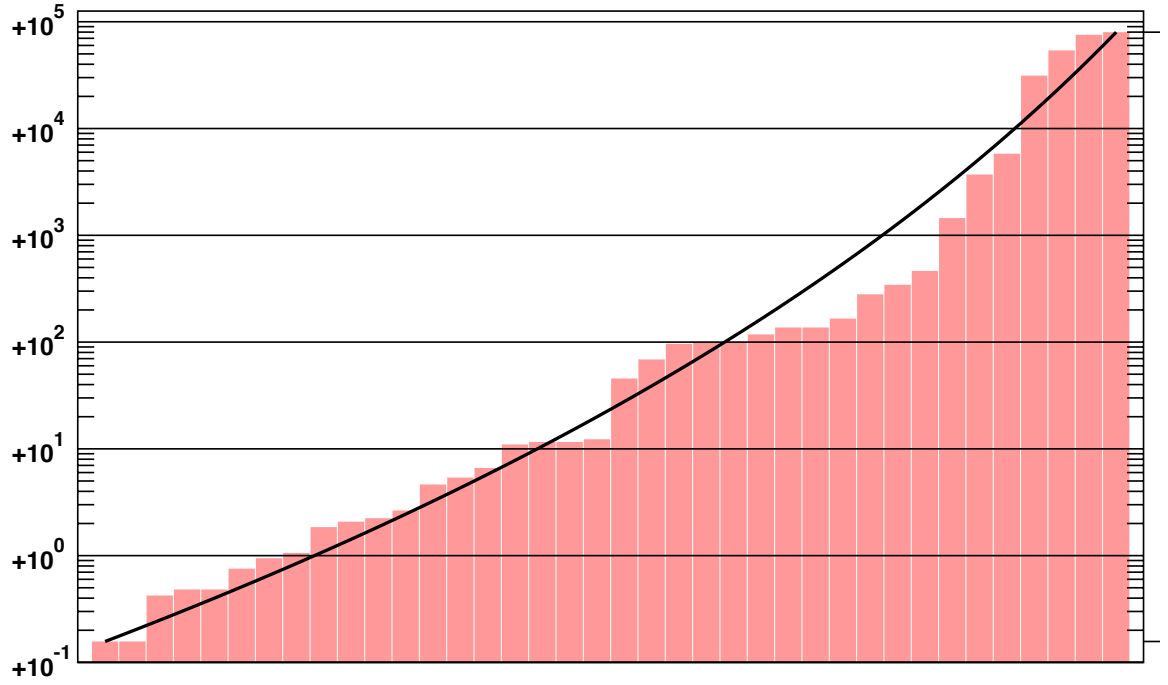

Figure 1. Methane emissions from abandoned oil and gas wells, Pennsylvania, mg/hr/well, Kang, et al., 2014.

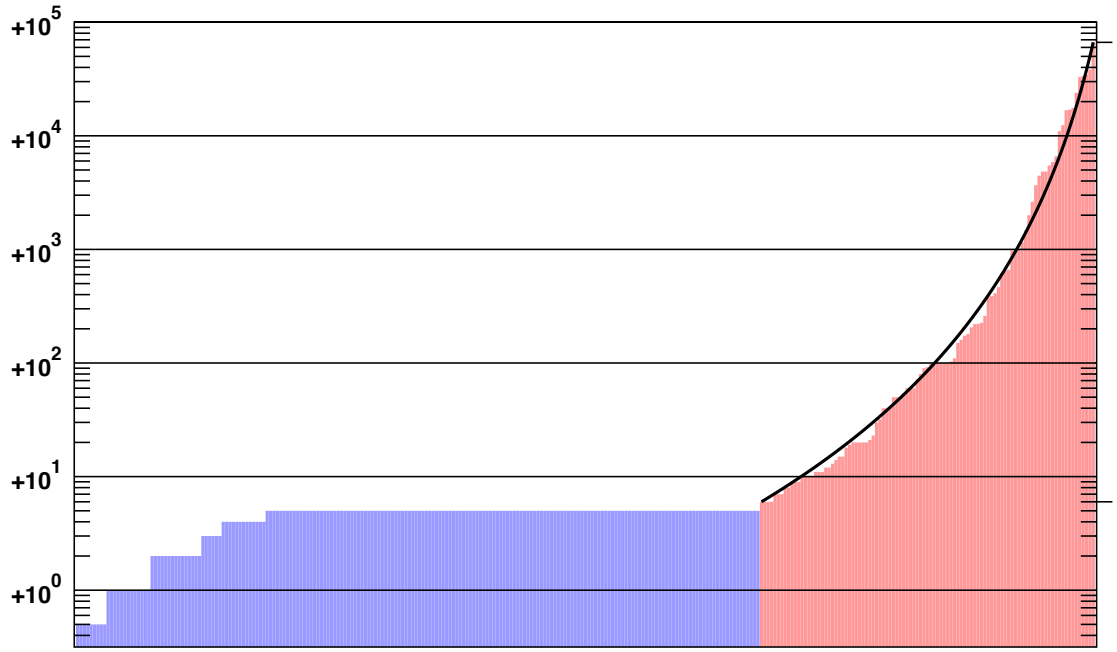

Figure 2. Methane concentrations in soil gas adjacent to coal-bed methane wells, Utah, ppm, Stolp, et al., 2006.

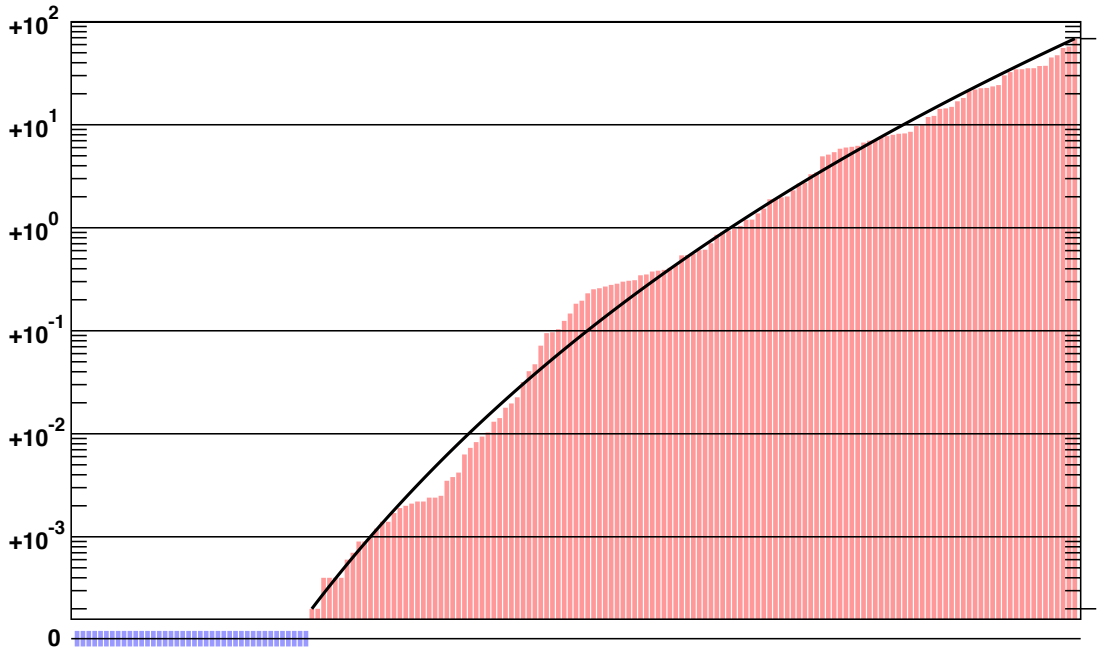

Figure 3. Methane concentrations in ground water, West Virginia, mg/L, White and Mathes, 2006.

In all three cases, measurements extend over four to six orders of magnitude, the means are much larger than the medians, and the standard deviations are larger than the means. In all three cases, a few large measurements completely dominate the measurement set. These few, large measurements have earned the moniker "super-emitter" in the literature. (To be precise, only the Pennsylvania well data are measurements of emissions to the atmosphere, so I sometimes also use the term "hot spots.") Such datasets are also said to follow distributions with "heavy" or "fat" tails, which is just a way of recognizing that data in the upper-end tail dominate the total.

There are many other examples in the literature of measurement datasets with this same general structure. They present a significant problem in trying to estimate emissions from the oil and gas sector, and perhaps from other classes of emissions as well. For example, one wonders if we have been able to adequately sample the super-emitters. When distributions are so highly skewed, we expect large sampling errors with negative bias. There are growing suspicions that bottom-up inventories of the oil and gas sector are too low, and no doubt, the super-emitter issue is part of the problem. Furthermore, the standard statistical tools and algorithms used to estimate sampling error cannot be applied.

Many of the problems connected with super-emitters center around the problem of knowing their true spectrum. For example, just by looking at the data in Figure 1, can anyone tell me what is the largest emission coming out of any one abandoned well in Pennsylvania? Suppose that Kang et al. could have afforded measurements at 400 or 500 wells, rather than just 38. How many more super-emitters would they have found, and how strong would they be? If, after studying 500 wells, they had found that the data still conform to a 95%-10% leakage law, then our current best estimate of the leakage per well would be much larger than 6000 mg/hr.

HEAVY-TAILED DATASETS CAN BE FIT TO POWER LAWS

The first step in a strategy for analyzing heavy-tailed datasets is to fit the dataset to some mathematical form. I have been able to fit a number of methane emissions datasets, including the three represented in Figures 1 – 3, with power-law distributions. These are distribution functions of the form:

$$P(x) = \begin{cases} 0, & if\ x \le a \\ \dfrac{\beta}{x^{\lambda}}, & if\ a < x < b \\ 0, & if\ b \le x \end{cases}$$

where λ is a positive constant exponent and β is a normalization constant. The distribution function is formally defined between a lower and an upper cutoff, *a* and *b*. The lower cutoff is employed because power-law behavior often is seen only in the high-end tail of the dataset (the pink bars in Figures 1 – 3). The upper and lower cutoffs serve another function. Depending on the value of λ, it may be permissible to dispense with one or the other of the cutoffs, and let $a \to 0$ or $b \to \infty$. On the other hand, we may need to maintain either *a* or *b* as finite numbers in order to avoid infinities. The earth and its inhabitants cannot produce an infinite amount of

methane, and our mathematical models must be adapted accordingly. In this case, we can think of b as the largest possible emission.

The index λ controls how rapidly the super-emitters thin out at the high end of the distribution. For example, when $\lambda = 0$, the distribution is actually uniform between a and b, there is no thinning out. With λ between 0 and 2, the super-emitters thin out so slowly that they have a strong influence on the mean of the distribution. When $\lambda > 2$, the super-emitters thin out rapidly enough that the mean of the distribution becomes insensitive to b, at which point it might be permissible to let $b \rightarrow \infty$, or at least assume that b is large and unspecified. Above a λ of about 2 or 3, the distributions begin to lose their heavy-tailed character.

I use a maximum likelihood estimation to fit a power law to any given dataset. The details will be published elsewhere [Mansfield, to be submitted]. The three datasets in Figures 1 – 3 can all be fit to power laws. Table 1 summarizes the results. The solid curves in Figures 1 – 3 are the expectations provided by the power-law fits. (The concavity or convexity of the curves depend on whether λ is less or greater than 1.) r* is an objective measure of the quality of fit. Essentially, it means that on a particular statistical test for power law behavior, the Pennsylvania Wells dataset outperforms 68% of all datasets known to be drawn from a $\lambda = 1.08$ power law. I assume any r* above about 0.1 to 0.2 to indicate a successful fit. The other remarkable thing in Table 1 is that we are observing power-law behavior over 4 to 5 orders of magnitude.

Table 1. Power Law Fits

	λ	r*	Range (max/min)
Pennsylvania Wells	1.08	0.68	500,000
Utah Soil Gas	1.21	0.77	11,000
West Virginia Ground Water	0.92	0.64	340,000

I have also fit seven other datasets of methane in the environment to power laws. They were all selected without cherry picking: I studied a total of ten environmental pollutant datasets that appeared heavy-tailed on paper. All ten could be fit to a power law using the maximum likelihood analysis [Mansfield, to be submitted].

GENERALIZATIONS OF THE CENTRAL LIMIT THEOREM PROVIDE JUSTIFICATION FOR THE USE OF POWER LAWS TO REPRESENT EMPIRICAL DATA

There are good reasons, based on the Central Limit Theorem and its generalizations, for using power laws to model datasets. According to the theorem, Gaussian distributions and power laws are "stable." For example, if x and y are variables that each obey its own Gaussian distribution, then it can be shown that their sum, $x + y$, also obeys a Gaussian distribution. The theorem implies that if the outcome of a process is the result of the sum of a large number of random variables, then that outcome will obey a Gaussian distribution. This is the reason that Gaussian distributions appear so frequently in nature [Adams, 2009; Zolotarev, 1986].

Power-law distributions are also stable, but they apply when the underlying processes are themselves heavy-tailed. To make a long story short, power laws are to heavy-tailed datasets as Gaussian distributions are to run-of-the-mill datasets. Stumpf and Porter put it this way: "One thus expects power laws to emerge naturally for rather unspecific reasons, simply as a by-product of mixing multiple (potentially rather disparate) heavy-tailed distributions." [Stumpf and Porter, 2012]. Like the Gaussian distribution, and for essentially the same reasons, power-law distributions are frequently encountered in nature. Among other things, they have been invoked as models for such diverse phenomena as personal wealth, personal income, the distribution of species among genera, the sizes of lunar craters, citation frequencies of scientific papers, the distribution of initial stellar masses, the sizes of cities, the sizes of files in internet traffic, and the occurrence frequency of words in English prose. (For a list of references, see Mansfield, to be submitted.)

Another possible mathematical law, known as the log-normal distribution, becomes appropriate under conditions for which the outcome of a process is the *product* of many random variables. Log-normal distributions also have heavy tails, and a parallel investigation examining how well they do in modeling environmental pollutants is called for. However, my research up to this point has focused only on power laws.

ESTIMATING THE SAMPLING ERROR OF HEAVY-TAILED DATASETS PRESENTS SPECIAL CHALLENGES

The next step in analyzing heavy-tailed datasets is to estimate their sampling error. Often in statistics, this takes the form of calculating 95% confidence limits (or some other percentage). Essentially, this means we perform an analysis that allows us to make an assertion of the form: "Based on the dataset in hand, we can state, with 95% confidence, that the true mean lies somewhere between A and B." The values A and B are the confidence limits, and they are functions of the dataset itself. A major goal of my research has been to develop techniques for obtaining confidence limits when the dataset can be represented by a power law.

One obvious but naïve statistical interpretation is to assume that the best fit, or the maximum likelihood, distribution is a close enough approximation to the true distribution that we can use it directly to represent our measurements. The problem with this interpretation is that many other distribution functions are also good fits to the data. The essence of determining sampling error is to include contributions from all distributions that have some likelihood of being the true one. In other words, we determine 95% confidence limits by averaging over all distribution functions that have some likelihood of having produced the given dataset.

The following formula gives the standard result, as taught in Statistics 101, for determining 95% confidence limits:

$$A, B = \mu_M \pm \frac{(1.96)\sigma_M}{\sqrt{n-1}}$$

Here, μ_M and σ_M are the mean and standard deviation, respectively, of the measurement set, and n is the total number of data points in the dataset. This formula also assumes that we have taken a randomized sample. An average over all possible Gaussian distribution functions is implicit in this formula, and for reasons beyond the scope of this talk, it can still be expected to work, under certain conditions, even when the underlying distribution is non-Gaussian. This formula has had numerous applications in statistics for many decades. However, in calculations on artificial datasets, I have demonstrated that this formula generally does not work well for heavy-tailed datasets. (It should work in the limit of *extremely* large n, but, for example, none of the three datasets introduced above have n values anywhere near large enough to justify its use.)

I have developed a procedure for determining 95% confidence limits when it can be assumed that the true distribution is a power law. It cannot be represented with a simple formula, rather it takes the form of a computer algorithm. (I will make the code available upon request, once it has been adequately developed.) In calculations with artificial datasets, I have been able to show that it works very well, with one nagging caveat: To work well consistently the procedure requires prior knowledge of the parameter b, *i.e.*, the value of the largest possible measurement, although when λ is larger than about 2.5, the algorithm becomes relatively insensitive to the value of b.

As mentioned above, the analysis involves an average over all distribution functions that have some likelihood, or some *a priori* probability, of producing the dataset in question. My current thinking is that the issue revolves around the meaning of the phrase "*a priori* probability." Averages that include all possible power laws with all possible values of the cutoffs do not converge: Distributions that have negligible *a priori* probability of producing the dataset are able to swamp out the contributions from those that do. (For example, any distribution that calls for more methane than the phenomenon could possibly produce of course has zero *a priori* probability, but the whole point of doing the measurements is to learn how much methane the phenomenon is producing.) In regards to the Pennsylvania Wells dataset, I pointed out the difficulty of determining the largest possible measurement from the existing dataset. The conundrum that we face is that the mean of a power law may depend strongly on the unknown value of b. The algorithm works well, but may require more information than may be available in real-world applications. Therefore, in cases where b cannot be constrained, it is important to examine sensitivity of the result to the value of b.

I am currently examining the impacts of three different approaches for constraining the value of b. First, depending on n and λ, it is improbable that we should see a very large gap between the largest measurement in the dataset and the value of b; this approach should be especially useful when $\lambda < 1$ and when n is large. Second, depending on the class of measurement, our knowledge of the physics may allow us to estimate b. For example, in the context of the Utah Soil Gas measurements, methane concentrations can never be greater than complete saturation, or 10^6 ppm. In the context of the West Virginia Ground Water dataset, we have ways of knowing the concentration of a saturated solution. Third, the State of Utah knows how many coal-bed methane wells were in existence at the time that Stolp, et al., performed their measurements. If we assume that all the wells follow the best-fit power law, we can make an order-of-magnitude estimate of b.

This procedure is still an active area of research, and I hope eventually to have more reliable algorithms. Nevertheless, there probably will never be a completely fail-safe algorithm that works when n is small and when λ is between about 1 and 2.5.

Above, I mentioned the possibility of modeling heavy-tailed datasets using log-normal distributions. I do not expect log-normal distributions to present the same challenges that are presented by the power laws. However, that is little help if it turns out that power laws are more appropriate from a physical standpoint. Fortunately, it should be fairly straightforward to judge, based on an analysis of the dataset, if log-normal distributions really are more appropriate.

CONCLUSIONS

The Generalized Central Limit Theorem provides theoretical justification for the use of power laws as mathematical models of heavy-tailed distributions. I have developed a maximum-likelihood estimation technique for fitting power laws to empirical pollution data, and have been successful in representing measurement datasets of methane in the environment.

Estimation of sampling errors has proven to be a significant challenge. The problem is related to an imprecise knowledge of the full spectrum of super-emitters, especially the value of the largest possible emission. I have developed an algorithm for estimating 95% confidence intervals, which is successful if one has prior knowledge of the largest possible emission. In the absence of such prior knowledge, it might still be possible to estimate confidence intervals if λ, the index of the power law, is either less than about 1 or greater than about 2.5, and if the dataset is not too small.

ACKNOWLEDGEMENTS

Partial funding provided by the Utah Science Technology and Research (USTAR) Initiative and by the Bureau of Land Management.

REFERENCES

Adams, W.J. *The Life and Times of the Central Limit Theorem;* 2nd ed., American Mathematical Society, 2009.

Kang, M., et al., "Direct measurements of methane emissions from abandoned oil and gas wells in Pennsylvania," *Proceedings of the National Academy of Sciences*, 2014, 111, 18173-18177.

Stolp, B.J.; Burr, A.L.; and Johnson, K.K. *Methane Gas Concentration in Soils and Ground Water, Carbon and Emery Counties, Utah, 1995-2003*, US Geological Survey, 2006; Scientific Investigations Report 2006-5227.

White, J.S. and Mathes, M.V. *Dissolved-gas concentrations in ground water in West Virginia, 1997-2005* U.S. Geological Survey, 2006; Data Series 156.

Zolotarev, V.M. *One-Dimensional Stable Distributions*, Translations of Mathematical Monographs, Vol. 65, American Mathematical Society, Providence, Rhode Island, 1986.

KEYWORDS

Methane emissions, heavy-tailed datasets, power-law distributions