STATISTICS OF SUPER-EMITTERS: Modeling heavy-tailed datasets with power-law distributions

Marc Mansfield
Bingham Research Center
Utah State University
Vernal, Utah

ACKNOWLEDGMENTS
Utah Science Technology and Research (USTAR) Initiative
Bureau of Land Management

April 16, 2015
EPI Emissions Inventory Conference
San Diego
Methane emissions from abandoned oil and gas wells, McKean and Potter Counties, PA (mg/hour/well)

Median = 29
Mean = 6676
St. Dev. = 19416
Max = 79900

Methane in soil gas near CBM wells, Carbon and Emery Counties, UT (ppm)

Median = 5
Mean = 1192
Std. Dev. = 5990
Max = 66500

Methane in ground water, WV (mg/L)

- Median = 0.18
- Mean = 5.6
- Std. Dev. = 12
- Max = 68.5

Super-emitters:
High-end members of the dataset, “hot spots.”
Responsible for most of the emission.
(70%-30%, 80%-20% rules, etc.)

Distributions have “heavy” or “fat” tails:
Much of the weight of the distribution is in the tail.
Mean >> median

Have we adequately sampled the super-emitters?

Perhaps this explains growing suspicions than bottom-up inventories are too low.

The things you learned in Statistics 101 are of no help here.
Strategy to Analyze Heavy-Tailed Datasets

Step 1: Fit to a distribution

Fit dataset to a distribution, e.g., power-law.

\[ P(x) = \frac{\beta}{x^\lambda} \]

Usually between upper and lower cutoffs: \( a < x < b \)

“Maximum Likelihood Estimation”

Upper cutoff is necessary whenever \( \lambda < 2 \).

(Earth can only produce a finite amount of methane.)

\( \lambda \) controls how rapidly the super-emitters thin out.
Why power laws?

Generalized Central Limit Theorem:

Gaussian distributions and power laws are “stable distributions.”
Sums of large number of random variables: Gaussian
Sums of large number of heavy-tailed random variables: Power law
Products of large numbers of random variables: Log-normal

Long story short: Power laws are to heavy-tailed datasets what the Gaussian distribution is to run-of-the-mill datasets.

“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 & Porter, Science, 335, 666 (2012).

Like the Gaussian distribution, power-law distributions pop up everywhere:

- Personal wealth or income
- Stellar masses
- Species among genera
- City sizes
- Lunar craters
- Files in internet traffic
- Citations of scientific papers
- Occurrence frequency of words
# Power Law Fits

(See also solid curves on bar charts.)

<table>
<thead>
<tr>
<th>Location</th>
<th>r*</th>
<th>Range (max/min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pennsylvania Wells</td>
<td>1.08</td>
<td>0.68</td>
</tr>
<tr>
<td>Utah Soil Gas</td>
<td>1.21</td>
<td>0.77</td>
</tr>
<tr>
<td>West Virginia Ground Water</td>
<td>0.92</td>
<td>0.64</td>
</tr>
</tbody>
</table>

Indicates quality of fit
Strategy to Analyze Heavy-Tailed Datasets

Step 2: 95% confidence limits

“Based on the dataset in hand, we can state, with 95% confidence, that the true mean lies somewhere between A and B.”

Fitted distribution ≠ “true” distribution
Many others are also good fits

Determine 95% confidence limits by averaging over all possible distributions.

This average is inherent in the formula they teach in Stat 101. Not guaranteed to work for heavy-tailed sets.
95%-confidence algorithm for power law distributions works very well

**IF**

I know the upper cutoff, \( b \).

(Related to the infinities inherent in the power law.)

Sometimes we might have independent information:

* e.g., methane in soil gas \(< 1,000,000\ ppm *

There may be other clues.

(I’m omitting the details.)

Without \( b \), the 95%-confidence interval becomes blurred and fuzzy.

Large \( N \) helps.

\( \lambda \ < \ 1 \) or \( \lambda \ > \ 3 \) helps.
95% confidence limits (using best available procedure) become spread out and fuzzy.
I do not expect a similar problem for log-normal laws

BUT

which law is appropriate?

(It might be possible for the dataset itself to answer this question.)
Conclusions

• Many heavy-tailed datasets of environmental pollutants can be fit to power laws.

• 95%-confidence limit calculation often becomes “fuzzy.” We can determine a confidence interval, but cannot always give it a definite percentage score. This is related to the inherent unpredictability of $b$. 