Revised Recommendations for Analyzing Crater Size-Frequency Distributions

CATSMEOWg: Crater Analysis, Techniques, Statistics, Methods, Estimation, and Observational Workers group

Stuart J. Robbins*,[1], Jamie Riggs[2], Brian P. Weaver[3], Edward B. Bierhaus[4], Clark R. Chapman[1], Michelle R. Kirchoff[1], Kelsi N. Singer[1], Lisa R. Gaddis[5]

*stuart@swri.boulder.edu / @DrAstroStu, [1]Southwest Research Institute, [2]Northwestern University, [3]Los Alamos National Lab, [4]Lockheed Martin Space Systems Company , [5]United States Geological Survey

Abstract

The modern field of impact crater studies has seen only one main attempt to standardize how crater population data are displayed and presented [1]. Since that 1979 work, the field, computer power, display capabilities, and relevant statistical methods have all progressed, and several assumptions from four decades ago have proven incorrect or incomplete. Our work is one of the first comprehensive attempts to suggest new, revised methods for crater population display and analysis. Revisions are being made for publication in the journal *MAPS*.

Size-Frequency Distributions (SFDs)

Traditional Types of Graphs [1]:

- Differential SFD: Craters are binned; bins are scaled by the bin width.
- Relative SFD ("R-Plot"): Craters are binned; bins are scaled by the bin width; values are divided by power law with exponent –3.
- Cumulative SFD: Craters are binned or unbinned; each smaller bin or diameter value is the number of craters at that diameter and larger.
- Incremental SFD: Craters are binned (*i.e.*, a basic histogram).

Traditional Approach: Data are binned, bin width usually $\cdot 2^{1/2}$.

Issues with Traditional Approach:

- How to deal with 0 or 1 crater per bin?
- Where does the bin's *x* value (diameter) really go?
- How does one incorporate any uncertainty in crater diameter?

New Recommendation: Kernel Density Estimator (KDE) to build an Empirical Density Function (EDF).

1. Represent each crater as a normalized (area under curve = 1) probability function, such as a Gaussian[1] (KDE). Mean is measured diameter, standard deviation is estimated uncertainty on diameter[2].
2. Repeat for every crater.
3. Add each KDE together to get the EDF.
4. The raw product from the EDF is a Differential SFD.
5. Can easily convert to all other traditional display formats.

[1]Any "reasonable" kernel shape could be used and the final EDF is very similar, regardless. Common shapes are a square/boxcar/top-hat, triangle, cosine, Epanechnikov, and Gaussian.

[2]Any measured crater diameter has an uncertainty, be it from individual repeatability or another researcher's reproducibility. Robbins *et al.* (2014) [2] showed ≈10%·*D* is a reasonable estimate of this quantity.

red box on right describes how the new recommendations solve traditional issues

Error Bars / Confidence Interval (CI)

Traditional Approach [1]: Uncertainty is assumed Poisson, $N^{1/2}$.

Issues with Traditional Approach:

- Philosophically, is Poisson appropriate?
- What about any source of uncertainty beyond the counting (Poisson) uncertainty?
- Cumulative: Represents uncertainty of all counts larger than or equal to a certain diameter bin, not just the new information in that bin. Is this appropriate?
- Error "bars" imply an uncertainty in a bin's count; is it more appropriate to think of this as a confidence envelope?

New Recommendation: Modified Bootstrap with Replacement [3]

1. Create the main crater EDF.
2. For each Monte Carlo run m_i, run M times ...
   a. For each n_i, for a total of N times (where N = # craters in population): sample, at random, a crater diameter from the EDF.
   b. Create EDF.
   c. Save EDF at each diameter position (d_i) of interest.
3. For each d_i:
   a. Sort M values from the bootstrapped EDFs.
   b. Find the number in the sorted list where the EDF at that d_i is (call it θ_{EDF}).
   c. For a certain confidence level (*e.g.*, 68%) (call it ψ):
      - Lower bound is value at: $\theta_{PDF}\cdot(1-\psi)$
      - Upper bound is value at: $(M-\theta_{PDF})\cdot\psi+\theta_{PDF}$

red box on right describes how the new recommendations solve traditional issues

Why Use These Techniques?

SFDs

1) Eliminates issue of 0 or 1 crater per bin because there are no longer "bins" in the display.
2) Eliminates issue of where to place a bin's diameter because there are no longer "bins" in the display.
3) Built-in method to account for uncertainty in crater diameter.
4) Easy to construct and convert between different display "types" (R, cumulative, differential, etc.).
5) Trends visible in the traditional methods are still present, so the institutional knowledge of how to interpret slopes, etc. remains.

Confidence Intervals

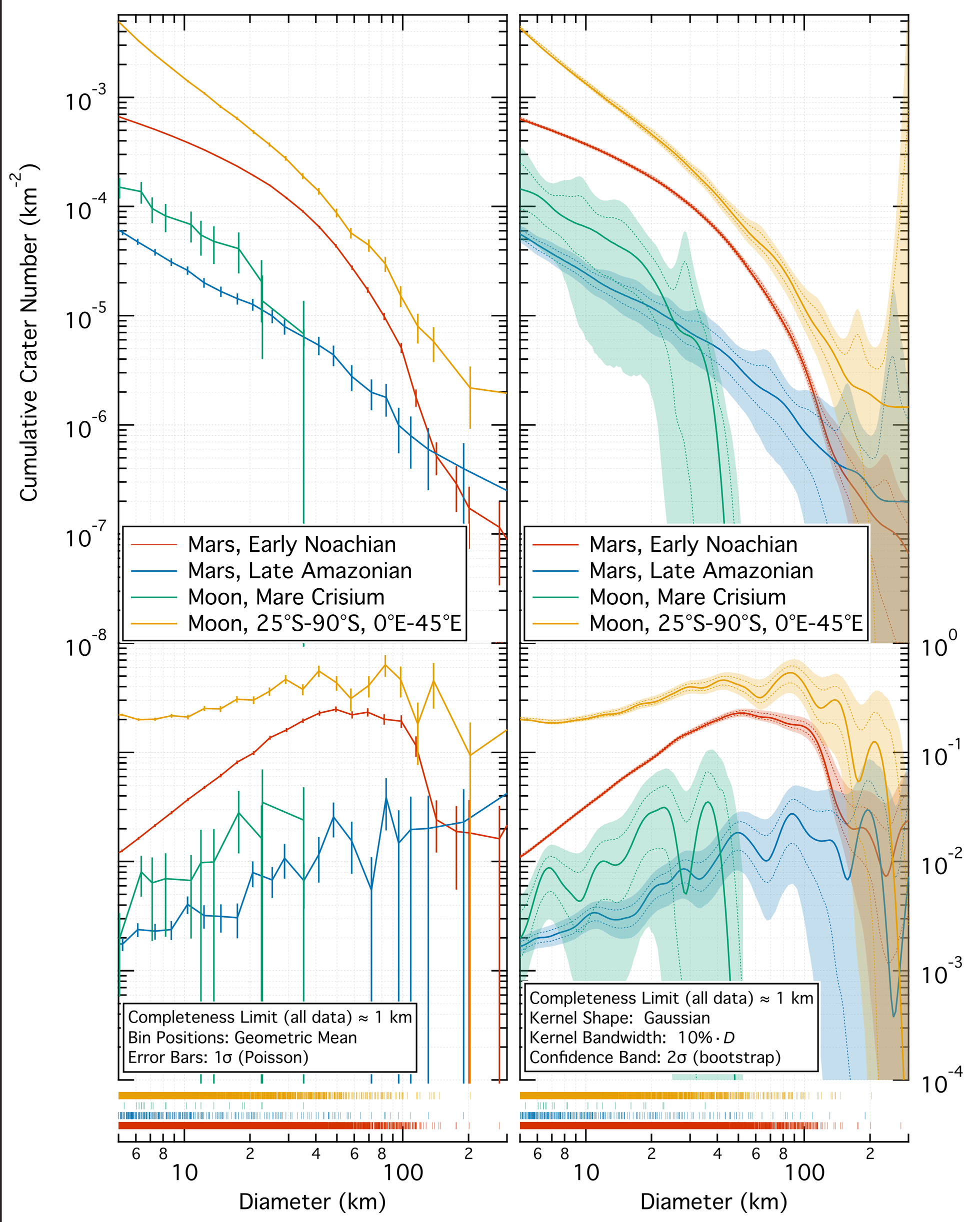
1) Eliminates assumption and structure imposed by Poisson: (a) mean equals square-root of standard deviation, (b) error bars are symmetric.
2) More statistically robust.
3) Still easy to calculate, though may take several minutes on larger datasets.
4) Increased uncertainty where large differences are between adjacent crater diameters in a sorted list.

Power-Law Fitting via ML

1) Computationally simple and easier than least-squares.
2) Much more statistically robust than least-squares.
3) Much less biased than least-squares.
4) Operates on original data rather than dependent on how data are displayed.

An Example SFD with CI

Classic Binned plots. New EDF plots.



- Bottom: "Rug plot" has tick marks for each crater diameter.
- New plots mirror basic trends of traditional plots
- New plots introduce different qualities at large diameters due to "gaps" in the data. These gaps manifest as large dips *BUT* with increased uncertainty.
- New plots allow visually simple "at a glance" analysis of overlaps, differences, and uncertainties.
- Rug plot allows simple "at a glance" of crater diameters in sample.
- Shaded confidence interval shows ±2σ while dashed envelope is ±1σ.

Fitting a Power Law

Traditional Methods: Least-Squares

Issues with Least-Squares:

- Assumes Gaussian nature of data, which is not valid for crater data.
- Assumes independent and identically distributed uncertainty on each data point, which is not valid for cumulative SFDs.
- Very difficult to factor in diameter uncertainty.

Recommendation: Maximum Likelihood Estimator (MLE)

- Statistically robust for power-law-based data. (see [4])
- Does not matter how data are binned/displayed, it only operates on the "raw" crater diameters.
- Always returns a fit value if $N \geq 2$.
- Is easier to use than least-squares:

a Pareto distribution: $f(D)=\alpha\dfrac{D_{\min}^{\alpha}}{D_i^{\alpha+1}}$

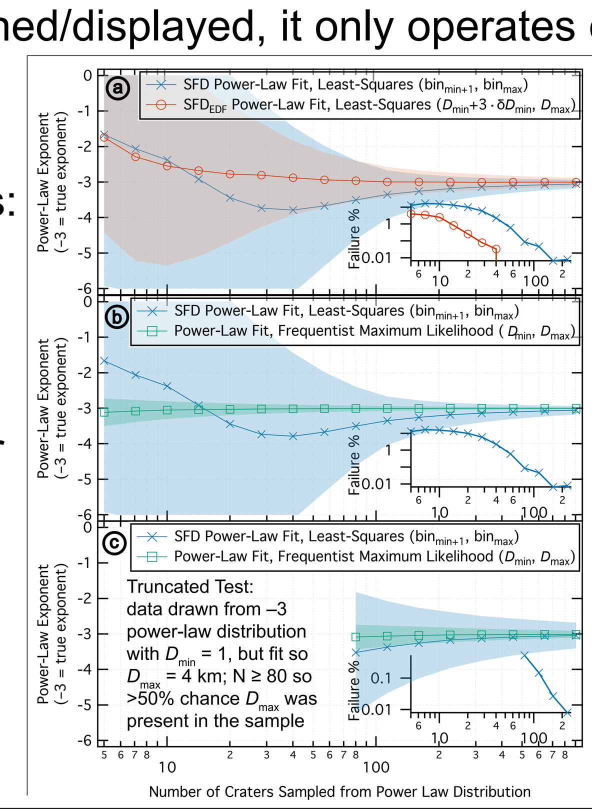
$D_{\min}$ = minimum diameter crater

$$\alpha=\frac{N}{\sum_{i=1}^{N}\ln(D_i/D_{\min})}$$

differential SFD slope: $-\alpha-1$

cumulative SFD slope: $-\alpha$

relative SFD / R-plot slope: $-\alpha+2$



Monte Carlo results of fitting simulated data via ⓐ least-squares for traditional bins or from sampling the EDF, ⓑ least- squares for traditional bins or from MLE to the crater diameters, and ⓒ truncating the data to larger diameters and using using the truncated form of MLE.

They show the MLE fits have less bias (it is closer to the true value) and less variance (it is closer more often) than traditional least-squares results.

Other Recommendations*

Basic Data Display:

1) Include detailed information in any plot, including a legend of all regions graphed, number of craters per region, and surface area. If this will not fit, include in caption.
2) Include both a cumulative and relative SFD because each show different aspects of the population.
3) Include a rug plot showing where the input data lie.

Interpretation:

1) Extreme caution must be taken near the completeness limit of one's data, and the completeness limit may be larger than estimated (see also [2]).

Data Archiving (may be US-Specific):

1) Crater data should be archived as supplemental material or in PDS or the USGS PDS Annex if at all feasible.

*This is a small subset, focused on general display and reporting recommendations that should be universal. Our manuscript goes into significantly more detail.

References & Acknowledgments

[1] Crater Analysis Techniques Working Group (1979). Standard techniques for presentation and analysis of crater size-frequency data. *Icarus* **37**:467-474. doi: 10.1016/0019-1035(79)90009-5

[2] Robbins S. J. et al. (2014). The variability of crater identification among expert and community crater analysts. *Icarus* **234**: 109–131. doi: 10.1016/j.icarus.2014.02.022

[3] Efron B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* **7**: 1–26. doi: 10.1214/aos/1176344552

[4] Clauset A., Shalizi C. R., and Newman M. E. J. (2009). Power law Distributions in Empirical Data. *SIAM Review 51*: 661–703. doi: 10.1137/070710111

This work was supported in part by NASA's TWSC/MDAP Award NNX15AC58G, NASA SSERVI NNH13ZDA006C, Southwest Research Institute, and Los Alamos National Lab.