

**REVISED RECOMMENDED METHODS FOR ANALYZING CRATER STATISTICS.** S.J. Robbins<sup>\*,1</sup>, J. D. Riggs<sup>2</sup>, B.P. Weaver<sup>3</sup>, E.B. Bierhaus<sup>4</sup>, C.R. Chapman<sup>1</sup>, M.R. Kirchoff<sup>1</sup>, K.N. Singer<sup>1</sup>. \*stuart@boulder.swri.edu. <sup>1</sup>Southwest Research Institute, 1050 Walnut Street, Suite 300, Boulder, CO 80302. <sup>2</sup>Northwestern University. <sup>3</sup>Statistical Sciences, CCS-6, Los Alamos National Laboratory. <sup>4</sup>Lockheed Martin Space Systems Company.

**Introduction:** 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. The field has generally refrained from using many of the statistical tools available to them, and instead, as a starting point, rely on now-traditional treatments that made many simplifying assumptions. Therefore, different researchers often deal with crater data differently, showing what appear to be standardized graphs but in fact contain differences that can change how the crater data are interpreted.

This was one of the driving motivations behind the 2015 "Workshop on Issues in Crater Studies and the Dating of Planetary Surfaces." During the workshop, the issue of standardization was raised, as was proper treatment of impact crater statistics. Since that time, several of us have worked together to investigate the problem and develop techniques and recommendations. We hope to spark discussion and this abstract and our PCC presentation will summarize a paper that is currently under review with *MAPS*.

**Philosophical Approach:** Traditionally, craters are treated as discrete objects that either exist or do not, and they are measured in an objective manner to yield their properties. As an alternative, we propose craters be thought of as a sample from a probability distribution: Impact craters are the outcome of an unknown probability distribution, and they are affected by deterministic and stochastic processes; moreover, "crater counts" (a term that also carries baggage) are the analyst's sampling of a population which is already a sample of the underlying processes. From the "counts," we want to work backwards to understand the probability distribution from which the craters were formed and processes that have affected the distribution (understanding there are sampling issues in our measurement, as well).

**Sources of Uncertainty:** Many sources of uncertainty are often known within the crater community but are rarely discussed. While researcher variability was known at least as early as 1970 [2], they are not factored into error bars nor standard crater analyses. Similarly, while crater diameter measurement has uncertainty (from technique, repeatability by one analyst, or reproducibility by others), this is also not factored in.

**Display Change:** Thinking of impact crater observations from a probability distribution, with numerous sources of bias and uncertainty, which themselves sample another probability distribution, suggest the data be displayed as a "Probability Density Function" (PDF) which is built by the "Kernel Density Estima-

tor" (KDE) technique [3,4]. This creates the same distribution as the traditional "differential" plot from [1].

The KDE models each impact crater  $i$  as a normalized probability distribution (*e.g.*, a Gaussian) with  $\mu_i$  at the observed crater diameter  $D_i$  and  $\sigma_i$  as an uncertainty in that crater diameter. Research has shown that crater diameter repeatability is  $\approx 10\%$  and *does* approximately follow a Gaussian distribution [2,5,6], so this can be used as  $\sigma$  ( $\sigma_i = 0.1D_i$ ). To construct the KDE, one determines where to sample diameters ( $d_s$ ) to create a visually continuous KDE (since KDEs are continuous, unbounded distributions) and then calculates the sum of all craters' probability distributions at each  $d_s$ .

Once the PDF is constructed, it can be converted to the incremental, summed from large to small diameters to create the cumulative plot (CSFD), or divided by  $D^{-3}$  to create the traditional R plot (Fig. 1). Except for very sparse crater datasets (*e.g.*, at large diameters relative to the rest of the data), the PDF will reproduce traditionally binned data almost exactly. Differences in sparse data are due to binning acting as a more significant smoother than the KDE and individual's variations in where to place the diameter bin. The former can be demonstrated, in part, visually by adding what is known as a "rug plot" to the horizontal axis which shows a small symbol for every  $D_i$ .

**Confidence Envelope:** Traditionally, crater analysts use 1-sigma error bars based on Poisson counting uncertainty ( $\sigma = N^{1/2}$ , where  $N$  is the number of craters sampled). This  $\sigma$  is based solely upon counting statistics; therefore, it does not include any other uncertainty sources described in this abstract. Also, depending on the purpose of a CSFD, uncertainties are incorrectly represented as  $N^{1/2}$  for they do not include the uncertainty of just new counts in the bins.

Statistical literature provides a Monte Carlo technique, known as the bootstrap, for calculating confidence intervals (CIs) [7]. The bootstrap has the benefit of working in situations where one is not able or not willing to assume an underlying structure (*e.g.*, Gaussian) for the probability distribution. The idea behind the bootstrap is to generate  $M$  bootstrap datasets where  $M$  is large (*e.g.*,  $\geq 1000$ ; the recommendation has increased over the years as computing power has increased). Mechanically, for each bootstrap dataset  $m_i$ ,  $N$  (the size of the original crater dataset) crater diameters are drawn with replacement from the PDF at random; because the method is "with replacement," one may sample the same crater more than once (some variations exist). A PDF is created for each  $m_i$ . For each  $d_s$ , the new PDF's value is stored. After  $M$  runs, at each  $d_s$ , sort the  $M$  values of the Monte Carlo PDFs.

From the PDF of the observed population, the value corresponding to the desired confidence interval above and below it would be recorded for each  $d_s$ . *E.g.*, if CI = 95%,  $M = 100$ , and at a given  $d_s$  the observed PDF is at the 60<sup>th</sup> position in the sorted list, then the CI for that  $d_s$  corresponds to the 3<sup>rd</sup> value and 98<sup>th</sup> value ( $60 \times (1 - 0.95)$  and  $60 + (100 - 60) \times 0.95$ ).

This method removes assumptions about the distribution (*e.g.*, Poisson), assumptions about symmetry of the confidence bands, and emphasizes that the crater population has a confidence band rather than error bars on specific bins.

**Fitting a Power-Law:** By default, most crater analysts will use whatever their analysis software of choice has to fit a power law to their (binned) data. This is most often a form of linear or non-linear least-squares (LS) (note— linear LS does not mean the fit takes the form of a line). If one were to use this with the PDF, we recommend re-sampling the PDF at each original crater diameter as the  $x$  value and taking the PDF  $\pm$  CI at that point as the  $y \pm \sigma_y$  value.

However, it has long been known in the statistics literature and the crater literature [8] that LS produces biased results, especially in data that are non-uniform or non-Gaussian because a core assumption of LS is  $\delta y_i$  are independent and identically distributed. A significantly less biased method is the Maximum Likelihood Estimator (MLE). MLE uses the original crater diameters and does not rely on any binning or smoothing to produce the fit parameter. There is a simple analytic form of the MLE for a Pareto distribution (power law truncated at a minimum diameter), and a version for a truncated Pareto distribution (bounded at large and small diameters) exists but must be solved numerically. MLE has the added benefit that it does not rely on the data display or format, so error bars on the MLE are the same regardless of whether the data are represented as a differential, CSFD or other form.

We used both a frequentist and Bayesian likelihood estimate, as well as LS to the resampled PDF and traditional bins, to fit power laws to data randomly sampled from a power law distribution with differential slope  $-3$  for  $N$  craters  $M$  times, where  $N = 5-905$  and  $M = 40,000-2500$  (depending on  $N$ ; Bayesian prior set to  $-3$  exponent with small level of Gaussian noise added). The results (Fig. 2) demonstrate the likelihood estimates produce the true slope with less bias and less variance than LS fitting. When using LS fitting, the fit to the PDF reproduced the true power law with less bias and variance than using the traditional bins. Finally, in cases of very small  $N$  where binning is practically meaningless, the MLE *always* returned a fit exponent, while the binned data often returned an inconclusive fit (exponent  $\notin (-13, +7)$ ).

**Summary:** We have presented several new suggestions for the treatment of crater population data that highlights some of our recent work. We think the methods described herein are a quantitative treatment

of important characteristics of crater population data, factoring in many known but previously ignored uncertainties, and are a better statistical treatment for the field.

**References:** [1] Crater Analysis Techniques Working Group (1979) doi:10.1016/0019-1035(79)90009-5. [2] Greeley & Gault (1970) doi:10.1007/BF00561875. [3] Rosenblatt (1956) doi:10.1214/aoms/1177728190. [4] Parzen (1962) doi:10.1214/aoms/1177704472. [5] Robbins *et al.* (2014) doi:10.1016/j.icarus.2014.02.022. [6] Robbins *et al.* (in rev.) "Craters of the Pluto-Charon System." [7] Efron (1979) doi:10.1214/aos/1176344552. [8] Chapman & Haefner (1967) doi:10.1029/JZ072i002p00549.

**Funding:** This work was funded by NASA TWSC/MDAP NNX15AC58G, NASA SSERVI NNH13ZDA006C, Southwest Research Institute, and Los Alamos National Lab.

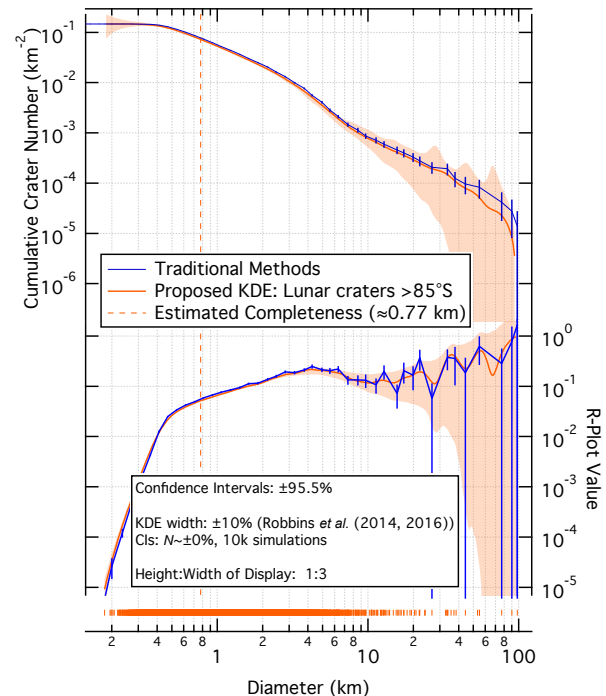


Figure 1. Example CSFD and R-plot with the traditional bins and new PDF. This illustrates several of our recommendations, both stated in this text (*e.g.*, rug plot) and not (*e.g.*, dotted vertical line estimating the completeness limit).

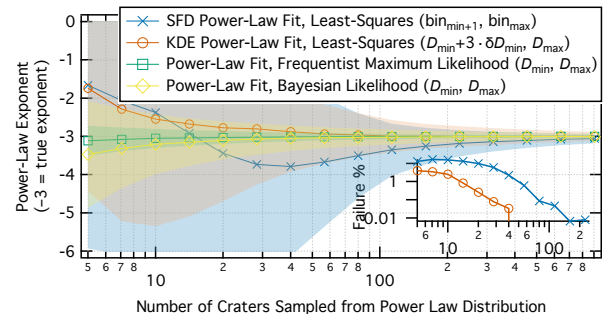


Figure 2: Monte Carlo test of fits with traditional LS and recommended LE; failure  $\equiv$  exponent  $\notin (-13, +7)$ .