ON THE BINNING AND ASSOCIATED UNCERTAINTY OF CRATER DIAMETER SIZE-FREQUENCY DISTRIBUTIONS. B.P. Weaver[1], S.J. Robbins[2], C.S. Plesko[3], J.D. Riggs[4]. [1]Statistical Sciences, CCS-6, Los Alamos National Laboratory; [2]Southwest Research Institute, 1050 Walnut Street, Suite 300, Boulder, CO 80302; [3]Applied Physics, Theoretical Design, XTD-NTA, Los Alamos National Laboratory; [4]Northwestern University. theguz@lanl.gov

Introduction: In 1978, in a locked conference room, twelve planetary scientists worked tirelessly to examine the plethora of crater population data in the literature with the goal of deriving a "best practices" set of recommendations. While the setting is perhaps dramatized, the result was the 1979 paper, "Standard Techniques for Presentation and Analysis of Crater Size-Frequency Data" [1]. It described a set of recommendations for how crater population data – specifically cataloged craters' diameters – should be tabulated and it describes two methods for display: Cumulative Size-Frequency Distribution (CSFD) and Relative Size-Frequency Distribution (R-plot). It also described the Differential Size-Frequency Distribution (DSFD or ISFD for "incremental" as it is sometimes referred to today). These all display crater diameter on the abscissa and some form of crater density or crater number on the ordinate axis. In the nearly four decades since that effort, several issues have arisen related to these types of display – specifically how diameter bins are assigned and uncertainties calculated – and a reexamination of these and whether they are appropriate is the subject of this abstract.

Binning Scheme: All of the recommended display types share a binning scheme in common. Crater diameters are binned on the *x*-axis, usually such that they appear evenly spaced when that axis is in $\log_{10}$ or, as advocated often by W.K. Hartmann, the axis tick marks are in 2^N where N is in the set $\mathbb{Z}$ (counting numbers and their additive inverses). Bins are usually set in multiplicative intervals of $2^{1/2}\cdot D$, such that if the first bin begins at D = 1 km, the second is 2 km, third is 4 km, and so on in a geometric progression. Then, the craters with diameters within each bin are summed or scaled in a method specific with the graph type.

The issue with this binning scheme is: where does the crater datum reside for that bin on the abscissa? If the data were approximately normally distributed, then the bin location would simply be the mean of the bin boundaries. In the above example, this would be 1.5 km, 3 km, 6 km, etc. However, crater population data typically follow a power law with an exponent of anywhere from –2 to –8, while model crater production functions (*e.g.*, [2]) have exponents of –2 to about –3.5. This means that, not only are there many more small craters within any given bin, but that one cannot assume an *a priori* distribution from which to easily calculate where the bins should be.

The variability in where bins are placed will have the effect of changing the amplitude of any of the SFDs. For example, if bin locations are placed at too large diameters, then the spatial density of craters will be displayed and interpreted as larger than it truly is, which would result in, *e.g.*, older model ages for a given surface. The opposite is also true.

With this described, we now describe the three main plotting techniques and how uncertainties are ascribed.

CSFD: A cumulative size-frequency distribution is created such that the number of craters within the largest diameter bin is assigned the value on the ordinate axis. The next-largest bin is the sum of its craters and the largest bin. And so on. This has the effect that the number of craters in any given bin is the total number of all craters larger than the minimum diameter for that bin.

The CSFD can then be normalized to the surface area on which craters were identified. This is often written as $N(D)$ notation, where D is the minimum bin diameter. For example, the lunar chronology [*e.g.*, 2, 3] is defined for $N(1)$ only – the spatial density of craters on a given terrain larger than or equal to 1 km in diameter.

Per the 1979 paper [1], uncertainties are ascribed as $N^{1/2}$ where N is the number of craters in any given bin of the CSFD. This is based on the assumption that identifying impact craters is inherently a Poisson-like counting endeavor.

DSFD or ISFD: The differential or incremental SFD is the same as the CSFD, minus one step: Each bin is simply the number of craters in that bin (and then potentially normalized to the surface area). Uncertainties are again $N^{1/2}$, but they will necessarily be larger per bin than the corresponding CSFD (except for the largest diameter bin).

R-plot: The relative plot does not have the same option as the C or D/I SFDs, for it must be normalized to the surface area. The ordinate axis value is:

$$R = \bar{D}^3 n / A\left(D_b - D_a\right)$$

where $\bar{D}$ is the geometric mean of the diameters in the bin:

$$\bar{D} = \left(\prod_{j=1}^{n} d_j \right)^{1/n},$$

where n is the number of craters per bin, d_j is the diameter of the j^{th} crater in the bin, D_a and D_b are the diameter limits of the bin (such that $D_b > D_a$), and A is the surface area of the terrain on which craters were identified.

Uncertainty is calculated as $R/N^{1/2}$, again assuming Poisson uncertainties.

Assigning Uncertainties: The root theory behind assigning these uncertainties is that crater populations

should follow Poisson statistics since researchers are counting discrete events. In a Poisson distribution with a mean N, one-sigma uncertainties are $N^{1/2}$. These are where 68% of the data should fall. Practically speaking, this can be interpreted as, (a) if history were run again, 68% of the time, craters in that $N \pm N^{1/2}$ range at any given diameter would be found; and/or (b) if another researcher (or the same researcher at a different time) were to identify craters in the region, craters within that $N \pm N^{1/2}$ range at any given diameter will be found 68% of the time.

Recent questions have arisen about this, however, specifically with respect to assigning CSFD uncertainties and uncertainties of very small numbers of craters – *e.g.*, 1 crater.

For the latter, consider the following thought experiment: You observe 1 crater in Arizona, the famous "Meteor Crater." Does that mean there are 0–2 craters in Arizona (since $1^{1/2} = 1$)? Whether strict Poisson statistics are appropriate in this case is questionable as related to a practical interpretation of the data.

For the former, the issue can be thought of in this manner: For each successively small diameter bin on a CSFD, the uncertainty grows because there are more craters. However, the relative uncertainty – the uncertainty divided by its value – will shrink because

$$\lim_{N \to \infty} N^{1/2} / N = 0 \, .$$

However, because there will be a limiting point where craters are too small to be detectable, there will be successively fewer and fewer new craters that are a part of the smallest bins. Since there is less new information that is incorporated into each diameter bin, should the assigned uncertainty reflect that?

Potential Solutions: At the May workshop, one potential better practice we will argue for is that a binning scheme is not necessary. Instead, we propose what is known in probability and statistics as the "empirical survivor function" (*i.e.*, this is defined as one minus the empirical cumulative distribution function), or "survivor function" for short. The survivor function is defined as the probability that a crater is larger than or equal to some value. Note that this is essentially the CSFD from [1] where each diameter is its own "bin."

Plotting the survivor function on a log-log plot (*i.e.*, plot of the log probability against the $\log_{10}(D)$) can be used to identify a suitable power law model. For example, if the (transformed) survivor function appears as a straight line on the log-log plot, then we know a power law is a suitable model. Similarly, plotting the survivor function on other scales may help identify other suitable models for the crater data. For example, plotting the log of the diameters against the quantile function of a standard normal distribution would inform that a lognormal distribution could be a suitable model.

Once a suitable model for the data has been identified, Bayesian or maximum likelihood estimation is presented for model fitting. Using such methods allows us to assign defendable uncertainties to the model fit and for any function that is dependent on the fit. Furthermore, any uncertainty that is calculated will borrow information from the entire fit of the data, potentially providing more precision in the estimates. This is quite different than the current approach of assuming an independent Poisson count for a bin which only uses the information for that bin (or for all bins larger than it).

We recognize that any significant suggested change from current standard practices can result in resistance from the community who would prefer a simple, easy-to-understand implementation even if it may contain small errors. These issues will also be addressed in our talk at the May workshop where we will present the current technique and its associated issues; what we think is an ideal technique from a statistical standpoint and how to implement it; and we will discuss potential middle-ground methods that are easier to implement and understand, are better than current practice, but they are not as rigorous as our ideal technique.

References: [1] Crater Analysis Techniques Working Group (1979). *Icarus*, doi: 10.1016/0019-1035(79)90009-5. [2] Neukum, G. *et al*. (2001). *Space Sci. Rev.* [3] Robbins, S.J. (2014). *EPSL*, doi: 10.1016/j.epsl.2014.06.038.