A Study of the Effects of Histogram Binning on the Accuracy and Precision of Particle Sizing Measurements

Eric Olson

This work and derivation yielded two equations of note for a geometrically spaced histogram, one for the binning constant, c_b, and one for the number of bins, k. With these equations, a conservative histogram may be constructed that has minimal effect on the accuracy and peak width of a number-weighted particle size distribution. Examples including a standard reference material and a real sample are given to show the effects of using these equations. Additionally, a sample of randomly distributed values is also analyzed, both with a traditional algorithm that lead to a discontinuous, discrete distribution, an effect known as over-binning, and the proposed equations, which lead to a continuous distribution.

Most particle sizing instruments that may be considered particle counters measure particles one at a time and populate a frequency histogram with the results. There are two common types of histograms: those that are arithmetically spaced and those that are geometrically spaced. Any histogram can be described using a range, the number of bins, and a binning constant. The binning constant may be additive (e.g., an arithmetically spaced histogram), or it may be multiplicative (e.g., a geometrically spaced histogram).

Arithmetically spaced histograms are the most common and are described in any course on statistics, one example being the normal or Gaussian distribution. Geometrically spaced histograms occur very often in nature, particularly in the field of fine particle characterization. An example of a geometrically spaced histogram is that of a lognormal distribution. The statistical characterization and properties of geometrically spaced histograms are taught far less often in general statistics courses. The theoretical process of generating a statistically robust, geometrically spaced histogram is mentioned in the literature, but the algorithms to do so are absent.

Instruments that measure particle size distribution work on a wide array of technologies and underlying concepts. These underlying concepts often decide the type of weighting by which the data will be reported. As some examples, sieving will produce a mass-weighted particle size distribution (1), laser diffraction will produce a volume-weighted particle size distribution (2), and dynamic light scattering (DLS) will produce an intensity-weighted particle size distribution (3). Another common class of instruments are particle counters, which produce a number-weighted particle size distribution.

Particle counters are somewhat unique in that they measure several particles in serial succession, one at a time. By contrast, all other technologies will measure a given property of an ensemble of particles at the same time then relate that property to particle size. Not only are counters able

Submitted: March 5, 2018
Accepted: April 13, 2018

to give a particle size distribution with the highest level of resolution, but they can sometimes offer information on particle shape as well as concentration. Particle counters, like most other platforms, are comprised of several underlying technologies. The three most common particle counters are those that rely on images (e.g., dynamic or static image analysis) (4, 5); those that produce a signal through the process of blocking light (e.g., light obscuration or single particle optical sensing [SPOS]) (6, 7); and those that make use of the electrical sensing zone method (8), also known as the Coulter principle.

Because each particle is individually measured, the number (or frequency) of observed particles of a given size range can be incrementally measured. This is the basis for the construction of a frequency histogram. The y-axis is typically used for the number of particles measured, a frequency, or probability value. The x-axis is divided into particle size ranges, often referred to as bins or class intervals.

The number of bins and the subsequent spacing between bins used in the construction of the frequency histogram varies depending on the instrument software. For instance, the Malvern Morphologi G3, a static image analysis instrument, will typically use 1000 to 2000 bins as a default. The Particle Sizing Systems Accusizer 770, an SPOS instrument, can use 64, 128, or 256 bins, with 128 bins being the software default. The Particle Sizing Systems Accusizer 780, another SPOS instrument, can use 128, 256, 512, or 1024 bins, with 256 bins being the software default. The Micromeritics Elzone II and the Beckman Coulter Multisizer 4, both of which use the Coulter principle, utilize 300 and 400 bins, respectively.

The use of more bins over a given fixed range will often result in higher resolution, but resolution gained in this manner does have a practical limit. The limit occurs when the difference between two bins becomes less than the resolution of the instrument. This phenomenon is called "over-binning" and is well documented for histograms that are constructed with arithmetic spacing.

One of the results of over-binning is the distribution becomes non-continuous and polymodal. Another way of saying this is the data are no longer continuous, but rather become discrete. In most traditional statistical treatments of continuous data, there is the assumption the data are normally distributed. The normal distribution, also known as a Gaussian distribution, is symmetric about its mean and monomodal (9). Once the data become discrete, more advanced statistical analysis treatments are available (e.g., likelihood testing, logistic modeling, etc.) (10), but because the underlying assumptions of traditional statistics are false for discrete data, the treatment of discrete data by traditional techniques is not applicable.

As previously mentioned, over-binning is well documented for histograms that are constructed with arithmetic spacing; however, there is virtually no solution proposed for histograms that are constructed with geometric spacing such as particle size distribution data. Most particle size distribution data are lognormal. Thus, they are typically presented using a semi-log plot with the x-axis being divided in a geometric manner.

Presented here is a mathematically derived algorithm that may be used to decide the proper number of bins to use based on the number of particles measured, the range of the data, the standard deviation, and the desired confidence level. This algorithm should give the best resolution possible without forcing the data to become discrete. To test the validity of the algorithm, two objectives were proposed:

- Study the effects of histogram binning on the accuracy and precision of particle sizing measurements.
- Test the hypothesis whether the proposed technique would decrease the probability of over-binning.

Review of literature

The classical frequency histogram is formed by constructing a set of non-overlapping intervals, called bins, and counting the number of points in each bin. Again, by classical theory, the bins all should be evenly spaced (i.e., arithmetically spaced). The first work in the literature to discuss the construction of histograms was by Sturges (11). Sturges assumed a binomial distribution could be used as a model of an optimally constructed histogram with appropriately scaled data. This led to what is known as Sturges' rule where k is the number of bins and n is the total sample size (see **Equation 1**).

$$k = 1 + \log_2 n \qquad \text{[Eq. 1]}$$

Sturges' rule has since become the widely recommended rule in most statistics texts and often is used in statistical packages as a default. One of the assumptions of Sturges' rule is the data follow a perfectly shaped Gaussian distribution (i.e., the distribution has no skewness or kurtosis). To accommodate skewness, Doane (12) proposed increasing the number of bins as a function of the standardized skewness coefficient, γ, as shown in **Equation 2**.

$$k = \left(1 + \log_2 n\right) + \left(\log_2\left(1 + \gamma\sqrt{\frac{n}{6}}\right)\right) \qquad \text{[Eq. 2]}$$

Both Sturges' rule and Doane's rule are known to lead to oversmoothed histograms, especially for large sample sizes (13, 14). To minimize the mean square error across the distribution function, Scott (15) derived an equation (see **Equation 3**) to find the best width of each bin, h, rather than the number of bins, k, which is a simple function of the standard deviation of the distribution, σ.

$$h = \frac{3.5\sigma}{\sqrt[3]{n}} \quad \text{[Eq. 3]}$$

This was then improved upon by Freedman and Diaconis (16) who derived a more robust equation (see **Equation 4**) that replaced σ with a multiple of the interquartile range (*IQ*). Note the interquartile range is the difference, X_{75}-X_{25}.

$$h = \frac{2(IQ)}{\sqrt[3]{n}} \quad \text{[Eq. 4]}$$

Terrell (17) derived a similar relationship (see **Equation 5**).

$$h = \frac{2.603(IQ)}{\sqrt[3]{n}} \quad \text{[Eq. 5]}$$

Sturges' rule, Scott's rule, and the Freedman-Diaconis rule constitute nearly all references in the literature and textbooks. In that order, Sturges' rule is the least stringent and the Freedman-Diaconis rule is the most stringent. Sturges' rule requires approximately 64% of the number of bins as Scott's Rule, which requires approximately 74% of the number of bins as the Freedman-Diaconis rule.

Keeping in mind one of the first assumptions of all three rules was the bins should all be evenly spaced (i.e., arithmetically spaced), Scott (15) proposed using the t-distribution as a reference density to change the normal rule when the data are skewed or heavily tailed. Scott's work derived both a skewness factor and a kurtosis factor by which either Scott's rule or the Freedman-Diaconis rule should be multiplied.

Derivation

In an arithmetic progression, the number of bins, *k*, should be related to the range of the data (*b*-*a*) and the minimum bin width, *h* (see **Equation 6**).

$$k = \frac{b-a}{h} \quad \text{[Eq. 6]}$$

Furthermore, the area, *A*, can be calculated as a Riemann sum over the range, *b*-*a*, assuming the frequency at each point to be y_i as shown in **Equation 7**.

$$A = \sum_{i=a}^{b} y_i h \quad \text{[Eq. 7]}$$

In a geometric progression, which may be used to describe a lognormal distribution, the area also can be calculated as a Riemann sum if c_b is the binning constant used to construct the geometric progression of the x-axis (see **Equation 8**).

$$A = \sum_{i=a}^{b} y_i \ln c_b \quad \text{[Eq. 8]}$$

Thus, the width of the first bin, which is also the width of the smallest bin, *h*, is a function of the binning constant and the starting value, *a* (see **Equation 9**).

$$h = a(c_b - 1) \quad \text{[Eq. 9]}$$

In statistics, a rearrangement of the t-test yields **Equation 10** in which the minimum number of samples, *n*, is a function of the statistical confidence level, $z_{\alpha/2}$, the standard deviation, σ, and the margin of error, ω. The margin of error also can be thought of as the smallest difference in measurements that is considered real or the maximum difference between the observed sample mean and the value of the population mean. At the 95% confidence level, if one assumes *n* to be sufficiently large, then $z_{\alpha/2} = 1.96$.

$$n \geq \left(\frac{z_{\alpha/2}\sigma}{\omega}\right)^2 \quad \text{[Eq. 10]}$$

Substitute *h* for ω as in **Equation 11**:

$$n \geq \left(\frac{z_{\alpha/2}\sigma}{a(c_b - 1)}\right)^2 \quad \text{[Eq. 11]}$$

Solve for c_b (see **Equation 12**):

$$c_b \leq \frac{1.96\sigma}{a\sqrt{n}} + 1 \quad \text{[Eq. 12]}$$

To solve for the number of bins, *k*, the range of values, *b*-*a*, can be expressed as a sum of all the individual bin widths, as shown in **Equation 13**.

$$\sum_{i=1}^{k}\left(ac_b^i - ac_b^{i-1}\right) = (b-a) \quad \text{[Eq. 13]}$$

Thus, in solving for *k* (**Equations 14** and **15**),

$$c_b^k = 1 + \frac{(b-a)}{a} \quad \text{[Eq. 14]}$$

$$k = \frac{\ln\left(\frac{b}{a}\right)}{\ln c_b} \qquad \text{[Eq. 15]}$$

Results and discussion

The feasibility of this process, which uses **Equations 12** and **15**, was initially applied to data from a static image analysis system(Malvern Morphologi G3). If successful, future investigations may find the applicability to other technologies and instrumentation. Two sets of data, one representing the analysis of a reference standard and the other representing an example sample, were chosen. The example sample or "real-world" sample was chosen because the plot of the data on a linear-linear scale was obviously lognormal in shape. The reference standard data were exported to Microsoft Excel, and the descriptive statistics were completed as given in **Table I**.

Using **Equation 15**, assuming the 95% confidence level, the number of bins, k = 108. The range was increased 10% above and below the maximum and minimum, respectively, so the data would not exceed the minimum and maximum bin sizes. Thus, a set of geometrically spaced bins was created with a minimum value of 5.679 µm, using 108 bins, to a maximum bin size of 160.089 µm. This was done using a geometric multiplier, c_b, of 1.031399 as calculated using **Equation 12**.

By comparison, a density histogram was also created using the data and the default bin choices: 1000 bins covering a range from 0.5 µ to 1000 µm. This was done using a geometric multiplier of 1.007638. The corresponding plots were overlaid in **Figure 1**. As displayed in **Figure 1**, the overall shape of the distribution using the default or original bins was a distribution with a mode around 40 µm and a shoulder with a mode around 70 µm. The distribution, on a semi-log scale, did appear to have a fair degree of fronting or positive skewness. The distribution also may be described as jagged or non-smooth. The distribution using the proposed re-binning parameters was nearly identical to the original plot but may be described as less jagged or smoother.

One possible measurement of smoothness is the second derivative of the data. The magnitude of the second derivative is proportional to the smoothness of the data. As shown in **Table II**, the mean absolute value of the second derivative of the originally binned data was 30.95. By comparison, the mean absolute value of the second derivative of the proposed binned data was 2.96, which was a ratio of 10.45. There appeared to be a small bias of the mean in both the originally binned data and the proposed data of 1.41% and 2.72%, respectively. This was to be expected and was in line with theoretical errors associated with most approximation methods (e.g., Simpson's method, Riemann sums, Trapezoidal method, etc.) (18). For both sets of binned data, the standard deviation was biased low (underestimated) by approximately 19.2% and 17.6%, respectively. This was a considerable magnitude but was most likely explained by outliers in the raw data that may be lost in the averaging and smoothing processes associated with binning.

The example sample data were exported to Excel and the descriptive statistics were completed as given in **Table III**. Again, using **Equation 15**, assuming the 95% confidence level, the number of bins, k = 32. The range was increased 10% above and below the maximum and minimum, respec-

Table I: Descriptive statistics for the reference standard.

Descriptor	Value
Mean	41.29
Standard deviation	15.76
Range	139.20
Maximum	145.51
Minimum	6.31
Count	30000

Figure 1: Binning effects on a reference standard.

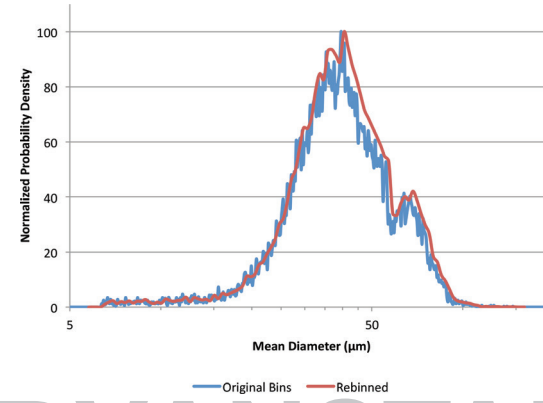


Table II: Calculated statistics for the reference standard.

	Raw data	Histogram data using original bins	Histogram data using proposed bins
Mean	41.29	41.87	42.41
Standard deviation	15.76	12.74	12.98
Mean d^2y/dx^2	n/a	30.95	2.96

Table III: Descriptive statistics for the example sample.

Descriptor	Value
Mean	4.55
Standard deviation	3.77
Range	32.56
Maximum	34.10
Minimum	1.54
Count	2373

ALL FIGURES COURTESY OF THE AUTHOR.

Figure 2: Binning effects on the example sample.

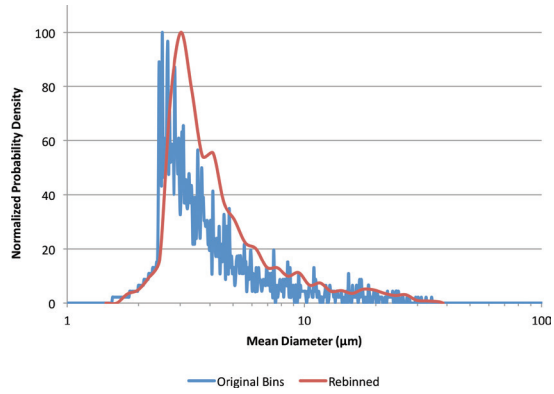
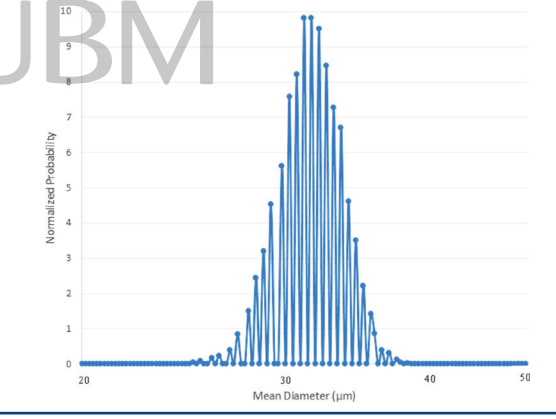


Table IV: Calculated statistics for the example sample.

	Raw data	Histogram data using original bins	Histogram data using proposed bins
Mean	4.55	3.07	3.25
Standard deviation	3.77	0.66	0.73
Mean d^2y/dx^2	n/a	4,058.29	73.57

Figure 3: Random data binned with the traditional spacing.



tively, so the data would not exceed the minimum and maximum bin sizes.

Thus, a set of geometrically spaced bins was created with a minimum value of 1.38 µm, using 32 bins, with a maximum bin size of 38.52 µm. This was accomplished using a geometric multiplier, cb, of 1.109637 as calculated using **Equation 12**.

By comparison, a density histogram also was created using the data and the default bin choices: 1000 bins covering a range from 0.5 µm to 1000 µm. This was accomplished using a geometric multiplier of 1.007638. The corresponding plots were overlaid in **Figure 2**.

As displayed in **Figure 2**, the overall shape of the distribution using the default or original bins was a monomodal distribution with a mode around 2.5 µm. The distribution, on a semi-log scale, did appear to have a significant degree of tailing or negative skewness. This negative skewness was obvious even in the semi-log plot presented as **Figure 2**. The distribution also may be described as jagged or non-smooth. The distribution using the proposed re-binning parameters had the same general shape, but the mode was around 3.0 µm, and it was much smoother.

As shown in **Table IV**, the mean absolute value of the second derivative of the originally binned data was 4058.29. By comparison, the mean absolute value of the second derivative of the proposed binned data was 73.57, which was a ratio of 55.2. There appeared to be a significant bias of the mean in both the originally binned data and the proposed data of 32.5% and 28.5%, respectively. For both sets of binned data, the standard deviation was biased rather low (underestimated) by approximately 81.6%. This was a considerable magnitude, but was likely attributable to the extreme jaggedness of the originally binned data and the degree of skewness of the distribution.

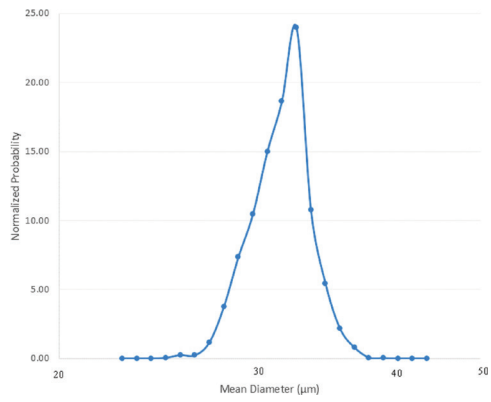
To test the hypothesis whether the new binning technique would decrease the probability of overbinning in which the bins are separated by more resolution than the instrument is capable, a random set of data (n =10,000) was generated such that all data points were a factor of 0.5 µm or n*0.5 µm apart where n is an integer. The randomly generated distribution had a mean of 32 µm and a standard deviation of 2 µm. The concept of the smallest change that is considered real, ω, was introduced with **Equation 10**. In the case of the Malvern Morphologi G3, ω is approximately 0.5 µm due to limitations imposed by optical microscopy (19).

The original number of 1000 bins was used with a minimum size of 0.5 µm and a maximum size of 1000 µm with a c_b = 1.007638. The difference between the bins closest to 32 µm was found to be 32.103 µm - 31.860 µm = 0.243 µm, which was less than ω for the Malvern Morphologi G3.

By comparison, using **Equation 15**, assuming the 95% confidence level, the number of bins, k = 22. Thus, a set of geometrically spaced bins was created with a minimum value of 22.95 µm, using 22 bins. This was done using a geometric multiplier, c_b, of 1.03209 as calculated using **Equation 12**. Using this modified binning scheme, the difference between the bins closest to 32 µm was found to be 32.48433 µm - 31.47433 µm = 1.009997 µm, which was greater than ω for the Malvern Morphologi G3.

The distribution of the randomly generated numbers was then plotted in Excel. When using the traditional binning parameters, the 10,000 data points fell into several discontinuous bins as shown in **Figure 3**. Again, this is to be expected because the bin width was less than the resolution of the data, 0.5 µm. By comparison, the 10,000 data points fell within several continuous bins when the bin

Figure 4: Random data binned with the proposed algorithm.



width was set using the proposed algorithms as shown in **Figure 4**, which is to be expected since the bin width was greater than the resolution of the data.

Conclusion

This original work and derivation yielded two equations of note, one for the binning constant, c_b, (**Equation 12**) and one for the number of bins, k (**Equation 15**). With these equations, a histogram may be constructed that has minimal effect on the accuracy and peak width compared to the default process.

The histogram generated with these equations will be continuous, thus, can be treated using normal statistical models, software, and processes.

A histogram generated using these equations will also have the statistically maximum resolution allowed given the instrument, software, and underlying scientific principles. Thus, the technology will be optimized to yield the most resolute data, which is a primary importance for a particle counter.

It is possible given certain applications that this increase in resolution may allow the instrument to be used to resolve monomers, dimers, trimers, etc., intact particles from fractured particles, partially milled lots from fully milled lots, etc.

In the example given, these algorithms led to a continuous distribution (**Figure 4**) rather than the original discontinuous distribution (**Figure 3**).

Once proven, typical statistical treatments could be used to calculate values (e.g., mean, median, mode, standard deviation, variance, standard error, skewness, kurtosis, etc.) as well as derived values from the integral or cumulative distribution plot. Furthermore, the resolution was optimized in such a way the minimum bin size was a scientifically supportable 0.975 µm rather than 0.243 µm.

If this had been a real-world sample, there could exist the situation where a population with a mean diameter near 1 µm should be contained within two or three bins, depending on the standard deviation of the distribution. Given the optimal bin spacing of 0.975 µm, this would most likely be the case.

If the traditional spacing were used, it is quite possible the same population would have been "resolved" into several discrete peaks, suggesting the sample were a mixture of several very narrowly distributed populations. The problem with this is the peaks would be considered independent populations even though the instrument is incapable of that level of resolution.

A more statistical way to state this is the null hypothesis that the distribution came from the same population would be incorrectly determined to be false. This is known as a false negative result, also known as type II or β-error. Through use of the derived equations, the probability of a type II error is minimized while maximizing the system resolution.

References

1. ISO 2591-1, Test Sieving—Part 1: Methods Using Test Sieves of Woven Wire Cloth and Perforated Metal Plate, 1988.
2. ISO 13320, Particle Size Analysis—Laser Diffraction Methods, 2009.
3. ISO 22412, Particle Size Analysis—Dynamic Light Scattering (DLS), 2017.
4. ISO 13322-1, Particle Size Analysis—Image Analysis Methods, Part 1: Static Image Analysis Methods, 2004.
5. ISO 13322-2, Particle Size Analysis—Image Analysis Methods, Part 2: Dynamic Image Analysis Methods, 2004.
6. ISO 21501-2, Determination of Particle Size Distribution—Single Particle Light Interaction Methods, Part 2: Light Scattering Liquid-borne Particle Counter, 2007.
7. ISO 21501-3, Determination of Particle Size Distribution—Single Particle Light Interaction Methods, Part 3: Light Extinction Liquid-borne Particle Counter, 2007.
8. ISO 13319, Determination of Particle Size Distributions—Electrical Sensing Zone Method, 2007.
9. S. Bolton and C. Bon, *Pharmaceutical Statistics—Practical and Clinical Applications* (Informa Healthcare, New York, NY, 5th ed., 2010).
10. T. Santnerand and D. Duffy, *The Statistical Analysis of Discrete Data* (Springer-Verlag, Berlin, Germany, 1989).
11. H. Sturges, "The Choice of a Class-Interval," *J. Amer. Statist. Assoc.* 21, 65–66 (1926).
12. D.P. Doane, "Aesthetic Frequency Classification," *American Statistician* 30, 181–183 (1976).
13. D.W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization* (John Wiley & Sons, New York, NY, 1992).
14. G. Terrell and D. Scott, *J. Amer. Statist. Assoc.* 80, 209-214 (1985).
15. D.W. Scott, *Biometrika* 66, 605–610 (1979).
16. D. Freedman and P. Diaconis, *Zeit. Wahr. ver. Geb.* 57, 453–476 (1981).
17. G. Terrell, *J. Amer. Statist. Assoc.* 85, 470-477 (1990).
18. J. Stewart, *Calculus, 7th ed.* (Brooks/Cole, Pacific Grove, CA, 2012).
19. E. Abbe, *Über Einen Neuen Beleuchtungsapparat am Mikroskop* (About a New Lighting Device on the Microscope) (1873). **PT**

Eric Olson is a senior research scientist—Physical and Chemical Characterization at PPD Laboratories.