Buffo, A.; Alopaeus, V.

Experimental determination of size distributions

Published in: Measurement Science and Technology

DOI: 10.1088/0957-0233/27/4/045301

Published: 14/03/2016

Please cite the original version:
Experimental determination of size distributions: analyzing proper sample sizes

This content has been downloaded from IOPscience. Please scroll down to see the full text.

2016 Meas. Sci. Technol. 27 045301


View the table of contents for this issue, or go to the journal homepage for more

Download details:

IP Address: 130.233.162.40
This content was downloaded on 30/05/2016 at 11:32

Please note that terms and conditions apply.
Experimental determination of size distributions: analyzing proper sample sizes

A Buffo and V Alopaeus

Department of Biotechnology and Chemical Technology, School of Chemical Technology, Aalto University, Kemistintie 1, 02150 Espoo, Finland

E-mail: antonio.buffo@aalto.fi

Received 26 November 2015, revised 3 February 2016
Accepted for publication 12 February 2016
Published 14 March 2016

Abstract

The measurement of various particle size distributions is a crucial aspect for many applications in the process industry. Size distribution is often related to the final product quality, as in crystallization or polymerization. In other cases it is related to the correct evaluation of heat and mass transfer, as well as reaction rates, depending on the interfacial area between the different phases or to the assessment of yield stresses of polycrystalline metals/ alloys samples. The experimental determination of such distributions often involves laborious sampling procedures and the statistical significance of the outcome is rarely investigated. In this work, we propose a novel rigorous tool, based on inferential statistics, to determine the number of samples needed to obtain reliable measurements of size distribution, according to specific requirements defined a priori. Such methodology can be adopted regardless of the measurement technique used.

Keywords: size distribution, sample size, statistical significance

1. Introduction

Measurements of the size distribution of drops, bubbles, and particles are very common in the chemical engineering and process industry due to the importance that these distributions have on the evaluation of mass, heat, and momentum exchange between the different phases of a multiphase system (Clift et al 1978), or on the quality of the final product in crystallization and polymerization (Myerson 2001). Moreover, the important role played by experimentally measured size distribution in the validation of predictive models, mainly based on the Computational Fluid Dynamics and Population Balance approach, is worth mentioning (Laakkonen et al 2007, Marchisio and Fox 2013, Buffo and Marchisio 2014). Another example that deserves a mention is the case of grain size distribution affecting the yield stresses of polycrystalline metals/ alloys (Berbenni et al 2007).

Nowadays, many experimental techniques are available for characterizing the size distribution of a dispersed phase: light scattering for solid particles (for a review of the methodology, see Xu 2015), capillary suction probes (Greaves and Kobbacy 1984, Barigou and Greaves 1991, 1992a, Barigou and Greaves 1992b, Alves et al 2002), phase Doppler anemometry (Mudde et al 1997, Kulkarni et al 2001) and digital imaging (Honkanen et al 2010, Maaß et al 2011, Lau et al 2013, Panckow et al 2015) for fluid particles. Besides the practical advantages (or limitations) that every technique possesses, an essential aspect of these measurements is represented by the number of samples needed to capture the population distribution reliably. Unfortunately, due to the complexity and time-consuming nature of the experimental procedures, this aspect is often overlooked, potentially leading to measurements that are not a reliable representation of the underlying size distribution from a statistical point of view. The standard procedure usually adopted is empirical and consists of evaluating the variation of a mean size value and its standard deviation with the number of sampled particles, as exemplified in Maaß et al (2011): when these values no longer vary with the sample size, the sampling procedure is stopped and the measurement...
is considered reliable. However, it is important to point out that this procedure might be not sufficient for characterizing the distribution without a proper definition of confidence intervals. This is especially true when considering that the characterization of a size distribution also involves high-order moment statistics, such as area or volume distribution, where the correct evaluation is very sensitive to the distribution tails. For this reason, we propose a new theoretical approach for the determination of the optimal sample size, after a small preliminary sampling. This procedure is applied here to a perfect precision measurement, but, as we will show later in this contribution, this aspect can be addressed by establishing a connection between the uncertainty of the measurement and the desired confidence interval.

2. Theoretical analysis

As previously mentioned, the experimental determination of a size distribution is usually carried out through a sampling procedure, with the aim of measuring a relative amount of particles, typically by number or volume, according to their size. Let us assume that the result of our measurement is an array of $N$ values of particle size, $L = \{L_1, L_2, \ldots, L_N\}$. Clearly this is just a sample of the unknown size distribution that we wish to measure. In order to characterize the population distribution from the sample, we need to rely on statistical estimators. For example, the mean value of this sample is equal to

$$\bar{L} = \frac{1}{N} \sum_{i=1}^{N} L_i, \quad (1)$$

while an unbiased estimator for the variance of the sample is the following:

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (L_i - \bar{L})^2. \quad (2)$$

We can also define the generic raw sample moment $M_k$ as

$$M_k = \sum_{i=1}^{N} L_i^k, \quad (3)$$

with $k = 0, 1, 2, \ldots, N_k$. It is also possible to show that mean and variance may be written as a function of the sample raw moments:

$$\bar{L} = \frac{M_1}{M_0}, \quad (4)$$

$$S^2 = \frac{M_0 M_2 - M_1^2}{M_0 (M_0 - 1)}, \quad (5)$$

where $M_0$ is equal to $N$, the total number of sampled particles. If the latter value is sufficiently high, we can be reasonably sure that the random error associated with the measurement is smoothed out and that our statistics are adequate for characterizing the distribution. According to the central limit theorem, we know that the mean values $\bar{L}$ of repeated measurement arrays are distributed according to a normal distribution with standard deviation equal to

$$\sigma_{SE} = \frac{\sigma}{\sqrt{N}}. \quad (6)$$

This is usually referred to as the ‘standard error of the mean’. Since the standard deviation of the distribution is unknown (i.e. can only be approximated with a very large sample size), we can use its estimation:

$$\sigma_{SE} \approx \frac{S}{\sqrt{N}}. \quad (7)$$

This relation tells us that the ‘standard error of the mean’ goes to zero when the square root of the number of samples $N$ goes to infinity. However, the choice of $N \to \infty$ is not feasible from the experimental point of view, therefore we need to make use of a few more concepts of inferential statistics.

Since the variance of the measured distribution is unknown, we need to use the $t$-test to assess the confidence interval of our measured mean $\bar{L}$; however, the preliminary number of samples $N$ is usually large enough ($N > 100$) that, for large degrees of freedom, the Student’s $t$-distribution approaches the normal distribution. For this reason, we can write the following relation for the confidence interval $\Delta \bar{L}$:

$$\Delta \bar{L} = 2 t_{N-\alpha/2} \frac{S}{\sqrt{N}}. \quad (8)$$

where $t_{\infty, \alpha/2}$ is the so-called $t$ value for infinite degrees of freedom of a two-tail Student’s $t$–distribution and $\alpha$ defines the confidence interval; in fact we can state that with a confidence level of $100(1 - \alpha)\%$ the mean of the population lies in the confidence interval $\Delta \bar{L}$. From the point of view of a statistical test, once $\alpha$ is chosen and the values of $S$ and $N$ are calculated from the sample, the confidence interval $\Delta \bar{L}$ is determined. In addition, the previous relation is particularly useful, since it is possible to reverse this concept and thus estimate the number of samples $N'$ needed to have a certain target confidence interval $\Delta \bar{L}$ and a certain confidence level of $100(1 - \alpha)\%$, decided a priori:

$$N' = \left( \frac{2 t_{N-\alpha/2} S}{\Delta \bar{L}} \right)^2. \quad (9)$$

The specification of the target confidence interval $\Delta \bar{L}$ and of the level of statistical significance $\alpha$ is an important choice for the experimental practitioner, since these two related aspects may depend on the experimental uncertainty of the particular measurement technique and on the desired accuracy of the experiment. Moreover, by substituting equation (5) into equation (9), we can write the following expression as a function of sampling moments:

$$N' = \left( \frac{2 t_{N-\alpha/2} S}{\Delta \bar{L}} \right)^2 \frac{M_0 M_2 - M_1^2}{M_0 (M_0 - 1)}. \quad (10)$$

In principle, the target confidence interval $\Delta \bar{L}$ should be at least equal to or larger than the resolution of the specific measurement technique adopted. This means, for example, that if our technique can only measure particle size up to 0.1 mm, we cannot choose a value lower than 0.1 mm as a confidence interval for the mean size. It should be noted that this choice is crucial for the determination of $N'$: the smaller the confidence...
interval, the more precise our statistics level but the larger the value of the sample needed. In the absence of this information about the measurement technique, we can continue our discussion by hypothetically assuming that the amplitude of the target confidence interval is a (small) fraction \( c \) of the mean sampling value \( \bar{L} \):

\[
\Delta \bar{L} = c \bar{L}
\]

(11)

where \( c \) is a constant less than one. It should be noted that this choice is arbitrary: in fact equation (9) can be used to evaluate \( N^* \) notwithstanding this definition of the target confidence interval. Nevertheless, imposing the target \( \Delta \bar{L} \) as a fraction of the preliminary sampled mean value \( \bar{L} \) represents a reasonable approach, since it gave us a physically relevant value to use in our examples. By substituting equation (11) into equation (10), we can write:

\[
N^* = \left( \frac{2t_{\alpha/2}}{c \bar{L}} \right)^2 \frac{M_\alpha M_2 - M_1^2}{M_0(M_0 - 1)},
\]

(12)

By collecting all the constants of the expression in one symbol \( C \), and expressing the mean value \( \bar{L} \) as a function of sample moments, we can write:

\[
N^* = \frac{C^2 M_0(M_0 M_2 - M_1^2)}{M_1^2 M_0(M_0 - 1)},
\]

(13)

where the constant \( C = 2t_{\alpha/2}/c \).

Equation (13) tells us which number \( N^* \) of particle is needed to obtain a statistically significant measurement of the mean value \( \bar{L} \) of a distribution. If we were only interested in assessing the mean size of the measured distribution, equation (13) would give us the desired answer. However, in the case of size distribution measurements, other properties related to the high-order moments of the distribution are usually of interest, such as the interfacial area or the total volume of the particles. In fact, the mean size is related to the first-order moment, as shown in equation (4), while the mean area or the mean volume is proportional to the following quantities:

\[
\bar{A} = \frac{1}{N} \sum_{i=1}^{N} A_i \propto \frac{1}{N} \sum_{i=1}^{N} L_i^1 = \frac{M_0}{M_0},
\]

(14)

\[
\bar{V} = \frac{1}{N} \sum_{i=1}^{N} V_i \propto \frac{1}{N} \sum_{i=1}^{N} L_i^3 = \frac{M_3}{M_0}.
\]

(15)

As stated by the previous expressions, the total interfacial area is proportional to the moment of order two with respect to particle size, while the total volume is proportional to the moment of order three. Here we are assuming that the particles all have the same shape, represented by the areic \( k_A \) and volumetric \( k_V \) shape factors, a common assumption made when such measurements are carried out. We can also define the variances for the mean \( \bar{A} \) and \( \bar{V} \) in the following way:

\[
S_A^2 = \frac{1}{N - 1} \sum_{i=1}^{N} (A_i - \bar{A})^2 = \frac{M_0 M_4 - M_2^2}{M_0(M_0 - 1)},
\]

(16)

\[
S_V^2 = \frac{1}{N - 1} \sum_{i=1}^{N} (V_i - \bar{V})^2 = \frac{M_0 M_6 - M_3^2}{M_0(M_0 - 1)}.
\]

(17)

In order to obtain the sampling size required to have a certain degree of significance for both the mean area and the mean volume, we should therefore use a similar form of equation (9):

\[
N_A^* = \left( \frac{2t_{\alpha/2}}{c \bar{A}} \right)^2,
\]

(18)

\[
N_V^* = \left( \frac{2t_{\alpha/2}}{c \bar{V}} \right)^2,
\]

(19)

where \( \Delta \bar{A} \) and \( \Delta \bar{V} \) are the confidence intervals for the estimation of the second- and third-order moment of the distribution, respectively. Also in this case, by making the same assumptions made for \( \Delta \bar{L} \), namely \( \Delta \bar{A} = c s \bar{A} \) and \( \Delta \bar{V} = c s \bar{V} \), we can write the following equations as functions of the high-order sampling moments:

\[
N_A^* = \frac{C^2 M_0^2(M_0 M_4 - M_2^2)}{M_2^2 M_0(M_0 - 1)},
\]

(20)

\[
N_V^* = \frac{C^2 M_0^2(M_0 M_6 - M_3^2)}{M_3^2 M_0(M_0 - 1)}.
\]

(21)

It should be noted that the constants \( C_A = 2t_{\alpha/2}/c_a \) and \( C_V = 2t_{\alpha/2}/c_v \) are arbitrary, and in principle they could have different values, since they depend on the level of significance required by the measurements with respect to the mean area and the mean volume of the investigated distribution.

Finally, it is possible to formulate a general rule for the desired level of significance of the measurements of a certain moment \( M_k \) of the distribution:

\[
N_k^* = \left( \frac{2t_{\alpha/2}}{c \bar{\delta}^2} \right)^2,
\]

(22)

where \( \Delta \delta \) is the target confidence interval for the moment \( M_k \) and the standard deviation \( s_k \) can be written as:

\[
s_k = \sqrt{\frac{M_0 M_{2k} - M_k^2}{M_0(M_0 - 1)}}.
\]

(23)

Thus, the number of samples needed is a function of the moment of order \( 2k \), \( M_{2k} \), meaning that high-order moments of the distribution significantly influence the number of samples needed for the proper characterization of the distribution.

3. Practical examples

In order to give a practical example of the proposed procedure, let us assume that we are measuring the bubble size distribution present in a small fraction of volume in a bubble column or in a stirred tank reactor. We first measure \( N = 1000 \) samples of bubbles and we wish to determine whether the number is sufficient for reliable statistics or, conversely, to get an idea of the right number of samples to obtain the desired level of statistical significance. Since we wish to keep the present analysis general and keep the uncertainties of each experimental technique separate from this analysis, a random number generator (based on the Marsenne-Twister algorithm) is used to generate an array composed of \( N = 1000 \) values of
bubble size. Let us assume that these are log-normally distributed, with the mean equal to 0.003 m and the standard deviation equal to 20% of the mean value, thus representing a possible outcome of an experiment (Case 1). The shape of the distribution is usually unknown, but in this case we assume a priori a realistic shape, according to the bubble size measurement experiments of (Laakkonen et al, 2006). Figure 1 shows the analysis usually carried out in these cases (as for example in Maß et al, 2011), i.e. the evolution of the developing mean value of interest (for this case $d_{32}$ the mean Sauter diameter, namely the ratio between the moment of order three and the moment of order two with respect to bubble size) as a function of the number of sampled bubbles. Based on visual analysis, it would appear that the developing mean and its cumulative standard deviation (expressed in percentage of maximum diameter) stabilize after approximately the first 500 samples. Therefore, according to the usual procedure and the results obtained, $N=1000$ would be too many for our purposes and we could stop after 500 sampled bubbles.

To ascertain whether it is reasonable to stop the measurement after 500 particles, we need to use some concepts defined in the previous section, such as the confidence bounds for the mean size, mean area, and mean volume. By assuming a confidence level of 99%, table 1 reports these values for our experiment after 500 and after 1000 samples respectively, showing how the amplitude of the confidence bounds decreases as the number of samples increases. The last rows of table 1 show how large the confidence bound is compared to the corresponding mean value: it can be seen that the ratio is higher for the high-order moments of the distribution and decreases with the number of samples. Of course, the smaller the ratio, the higher the level of statistical significance of our sampling procedure. Moreover, this analysis clearly shows the improvement in the level of statistical significance by increasing the sample size, which is not obviously identifiable from the analysis of the developing Sauter mean diameter and its cumulative standard deviation performed in figure 1.

### Table 1. Case 1. Possible outcome of a size distribution experiment.

<table>
<thead>
<tr>
<th></th>
<th>$\bar{L}$ (m)</th>
<th>$\Delta \bar{A}$ ($m^2$)</th>
<th>$\Delta \bar{V}$ ($m^3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analytical</td>
<td>$3.000 \cdot 10^{-3}$</td>
<td>$9.360 \cdot 10^{-6}$</td>
<td>$3.037 \cdot 10^{-8}$</td>
</tr>
<tr>
<td>After $N=500$</td>
<td>$2.987 \cdot 10^{-3}$</td>
<td>$9.292 \cdot 10^{-6}$</td>
<td>$3.013 \cdot 10^{-8}$</td>
</tr>
<tr>
<td>After $N=1000$</td>
<td>$2.981 \cdot 10^{-3}$</td>
<td>$9.250 \cdot 10^{-6}$</td>
<td>$2.992 \cdot 10^{-8}$</td>
</tr>
</tbody>
</table>

Note: Mean value, standard deviation, confidence bound amplitude for a confidence level of 99% and ratio between confidence bound and mean value for the mean size, mean interfacial area and mean volume of the particle size distribution.

### Table 2. Case 1. Confidence intervals and number of samples needed to obtain statistically significant measurements for the different moments of the distribution.

<table>
<thead>
<tr>
<th></th>
<th>$\Delta \bar{L}$ (m)</th>
<th>$\Delta \bar{A}$ ($m^2$)</th>
<th>$\Delta \bar{V}$ ($m^3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>After $N=500$</td>
<td>$1.396 \cdot 10^{-4}$</td>
<td>$9.144 \cdot 10^{-7}$</td>
<td>$4.856 \cdot 10^{-9}$</td>
</tr>
<tr>
<td>After $N=1000$</td>
<td>$9.830 \cdot 10^{-5}$</td>
<td>$6.429 \cdot 10^{-7}$</td>
<td>$3.398 \cdot 10^{-9}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\Delta L/\bar{L}$ (%)</th>
<th>$\Delta \bar{A}/\bar{A}$ (%)</th>
<th>$\Delta \bar{V}/\bar{V}$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>After $N=500$</td>
<td>4.7</td>
<td>9.8</td>
<td>16.1</td>
</tr>
<tr>
<td>After $N=1000$</td>
<td>3.3</td>
<td>7.0</td>
<td>11.3</td>
</tr>
</tbody>
</table>

The procedure that we propose in this contribution is based on the idea of fixing a target confidence interval and the desired level of statistical significance, and then finding the number of samples needed to meet these requirements. In practice, the choice of these constraints is left to the experimental practitioner. In the present analysis, the target confidence intervals are expressed as a fraction of the corresponding preliminary measured moment value (see equation (11)). If we assume that this fraction is equal to 1% and a confidence level of 99% ($\alpha = 0.01$ and thus $t_{\alpha,0.05} = 2.576$) we obtain the values shown in table 2. As can be seen, not only are the 1000 particles insufficient, but we need to sample more than 10 times this number to obtain a statistically reliable value for the mean size, more than 48 times the number for the interfacial area, and more than 128 times the number for the total volume. However, it is important to remember that we are assuming null measurement uncertainty (since our numbers are artificially generated); in real life, every technique has a specific measurement uncertainty intrinsic to the experimental procedure, and this must be taken into account when selecting the desired amplitude of the confidence interval. For example, it is not correct to assume a target confidence interval for the volume $\Delta \bar{V} = 0.299 \, m^3$ unless our experimental technique can detect a similar variation of bubble volume. Figure 2 shows the concept of confidence bounds for the distribution in Case 1: if we measure just 500 particles to build our statistics (first row of figure 2), the statistical error associated with...
the determination of the distribution may be very large, both for the size distribution and the volume distribution function. The gray area shown in figure 2 is a visual representation of the statistical error associated with the sampling; this area is drawn by considering a high number of different possible realizations of the same sampling procedure (thus explaining the presence of bumps in the figure). This proves that the number of particles suggested by the analysis of the developing mean and the cumulative variance (i.e. approx. 500 particles) may not be enough to capture the size distribution with an acceptable level of statistical significance. The only way to reduce this statistical error is to measure more particles, as shown in the other rows of figure 2, where the statistical error is represented for \( N = 1000 \), \( N = 10876 \), and \( N = 128942 \). As is clear from the figure, the gray area associated with the statistical error decreases with the increase in number of sampled particles, until it almost disappears in the last row of figure 2, when the number of particles suggested by our methodology
Table 3. Case 2. Possible outcome of a size distribution experiment.

<table>
<thead>
<tr>
<th>Num.</th>
<th>( \bar{L} ) (m)</th>
<th>( \bar{A} ) (m²)</th>
<th>( \bar{V} ) (m³)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ana.</td>
<td>6.451 \times 10^{-3}</td>
<td>5.580 \times 10^{-5}</td>
<td>5.653 \times 10^{-7}</td>
</tr>
<tr>
<td></td>
<td>6.5 \times 10^{-3}</td>
<td>5.668 \times 10^{-5}</td>
<td>5.776 \times 10^{-7}</td>
</tr>
</tbody>
</table>

\( \Delta L \) (m) | \( \Delta A \) (m²) | \( \Delta V \) (m³) 
6.451 \times 10^{-5} | 5.580 \times 10^{-7} | 5.653 \times 10^{-7} 

\( N_L \) | \( N_A \) | \( N_V \) 
90 560 | 266 331 | 480 937

Table 4. Confidence intervals and number of samples needed to obtain statistically significant measurements for the different moments of the distribution.

4. Conclusion

This communication shows a novel rigorous methodology for the determination of the sample size required to obtain reliable measurements of size distributions. The present analysis demonstrates that the number of samples needed depends on the shape of the unknown distribution and on the requirements chosen for the confidence interval of the desired statistics. For the high-order moments often required in the analysis of chemical processes, the number of samples needed could be higher than generally expected.

Acknowledgment

The authors wish to thank the Academy of Finland for their financial support (Consortium ExtraFin - PORLIS project).

References

Bergomi S, Favier V and Berveiller M 2007 Impact of the grain size distribution on the yield stress of heterogeneous materials Int. J. Plast. 23 114–42
Myerson A 2001 Handbook of Industrial Crystallization (Boston: Butterworth Heinemann)
Xu R 2015 Light scattering: a review of particle characterization applications Particuology 18 11–21

Note: The underlying distribution is the sum of two lognormal distributions, one with a mean equal to 0.003 m and standard deviation equal to 6.0 \times 10^{-4} m, and the other with a mean of 0.01 m and standard deviation of 0.002 m.

\( \Delta L \) (m) | \( \Delta A \) (m²) | \( \Delta V \) (m³) 
6.451 \times 10^{-5} | 5.580 \times 10^{-7} | 5.653 \times 10^{-7} 

\( N_L \) | \( N_A \) | \( N_V \) 
90 560 | 266 331 | 480 937