bulk solids handling Volume 1 Number 3, September 1981

Determination of Sample Size for Estimation of the Mean Grade

Tuncel M. Yegulalp, USA

Die Bestimmung der Probengröße für die Abschätzung der mittleren Güte
Détermination du calibre d'un échantillon pour estimer la qualité moyenne
Determinación del tamaño de muestras para la estimación del grado medio
平均等級評価用サンプルサイズの決定
样品大小的决定作为平均品级的估计
تحديد حجم العينة لتقدير المرتبة الوسط . بقلم ت . م . يجولالب .

Summary

Practical guidelines for the accurate assessment of the requisite sample size for the estimation of the mean grade of a quantity or volume of mineralised bulk material are presented, and an illustrative numerical example is given.

1. Introduction

In mineral processing, sampling of bulk material plays an exceptionally important role. It provides the necessary estimates of the material properties before and after processing. Statistical theory dealing with the properties of the arithmetic mean of a sample as an estimator of the unknown population mean is well developed especially for normally distributed random variables. In the following we shall discuss some elementary statistical relationships and their use in determining the minimum required sample size in sampling of bulk materials with detailed numerical examples.

2. Sampling

The act of sampling aimed to determine one or more properties of a certain material of a certain volume consists of:
a) selection of a relatively small portion of this volume and
b) determination of its properties by means of physical and/or chemical testing. Since the results of these tests are going to be used in estimating the properties of the whole volume, it is desirable to have procedures which will produce estimates with acceptably small errors with acceptable costs.

The selection process produces a sample which may consist of one or more number of specimens that are tested or assayed. In the mineral industries the word *sample* is used to designate one such specimen. In the widely accepted terminology of mathematical statistics the word *sample* means a collection of one or more observations (assays, test results, etc.) of the attribute in question. In the following, to be consistent with the terminology of mathematical statistics we shall refer to a single member of a sample as a specimen, and a collection of specimens as a sample.

Tuncel M. Yegulalp, Associate Professor of Mining Engineering, Henry Krumb School of Mines, Columbia University, New York, NY 10027, USA

3. Sources of Variation in Sampling

If the sampled material is perfectly homogeneous with respect to its properties in question, any fraction of the bulk would be exactly the same as far as those properties are concerned. For example, if a certain bulk material is heterogeneous in its chemical composition but homogeneous in its density it would be considered homogeneous if it was being sampled to determine its density. The test results of different specimens may show some variation even if the material is homogeneous. This type of variation is attributable to the testing or assaying procedures and their repeatability. If the material is heterogeneous, it is expected that the specimens will differ in their measured properties. This is common occurrence in sampling crushed or ground bulk of ore and ore concentrates in mineral processing. The source of this variation, in addition to the variation caused by the assaying process, can be attributed to two main causes [2, 3] when we deal with particulate material:

1. The smaller fragments of the particulate material are themselves heterogeneous. This occurs especially when sampling crushed ores where the particle size is larger than the liberation size of the minerals in the composition of the ore.
2. Segregation of certain types of fragments due to handling of the bulk material. This source of variation can be overcome by blending the particulate material. Blending, however, does not eliminate the heterogeneity of individual particles [2, 3].

The statistical theory of sampling considers a sample of size N (N specimens) selected randomly and independently. A random selection process must be such that every member of the population being sampled would have the same chance to be selected. The independency of specimens implies that the selection of one specimen does not influence the selection of another member in the lot. These requirements are not (in spite of all efforts) usually met ideally in practice either due to the two major sources of variation discussed above or due to the nature of the physical sampling technique or both. For example, a perfectly mixed batch of ore may become segregated during the sampling process due to the nature of mechanical handling system.

The act of sampling can be considered as a process which allows one to observe the realizations of a random variable whose statistical properties are determined by the inherent characteristics of the sampled population and the sampling and measurement process. Naturally, the apparent variability of the assay values can be greatly influenced by the sampling and assaying processes. In statistics a random variable is defined as a function which assigns real numbers to the outcomes of a random experiment. In our case the random experiment consists of: a) selection of a fraction of a given volume of bulk by giving equal chance to all other fractions in the lot, b) further reduction of its volume and selection of a fraction of the specimen volume for testing, and c) testing or assaying the final fraction for its physical and/or chemical properties. The random variable being observed here is defined by the sampling process. A change in the volume of the specimen in the sampling process will correspond to a different experiment, thus will result in the realization of a different random variable. When sampling from a well mixed bulk of particulate material it can be postulated that the random variables observed by choosing different specimen volumes would possess the same distribution function and the same mean but different variances. The theoretical and experimental studies indicate that, other things being constant, reduction of specimen volume increases the variance. A practical relationship has been proposed by Gy [2, 3, 4] which is useful in determining the minimum acceptable specimen weight as a function of the diameter of the coarsest fragment in the bulk. This relationship is expressed as:

$$M_s \geq \frac{Cd^3}{\sigma^2} \qquad (1)$$

Where:

σ^2 variance of the tolerated sampling error
C a constant characterizing the material to be sampled
d diameter of the coarsest fragment
M_s weight of the specimen.

4. Determination of Minimum Sample Size

The arithmetic average of random independent observations of a Normally distributed random variable is known to be the best estimator of the unknown population mean. The Central Limit Theorem [1] asserts that the arithmetic average of random independent observations of a random variable will be distributed Normally when the sample size is infinitely large. It has been shown in the literature [6, 7] that even for moderate sample sizes the statistical behaviour of the sample average is acceptably close to that of the Normal distribution. It is also known that by increasing the number of specimens (sample size) one obtains sample averages closer to the unknown value of the population mean. The Law of Large Numbers [5] asserts that when the sample size is infinitely large, the sample mean becomes equal to the population mean. In practice, however, we deal with small sample sizes for obvious economic reasons, and we do not expect our sample average to become equal to the unknown population mean. What is customary then is to establish an interval around the sample average which will contain the unknown mean with a certain predetermined probability. Such an interval is called "confidence interval" in statistics. Given the probability (1-α), the length of the confidence interval becomes a function of the population variance and the sample size. This relationship is derived from the following probability statement:

$$\text{Prob.}\left[\bar{X} - Z_{1-\alpha/2}\,\sigma/\sqrt{N} \leq \mu \leq \bar{X} + Z_{1-\alpha/2}\,\sigma/\sqrt{N}\right] = 1-\alpha \qquad (2)$$

Where:

$\bar{X}$ sample arithmetic average
σ population standard deviation
μ population mean
N sample size.

$Z_{1-\alpha/2}$: $1-\alpha/2$ percentile of standard Normal variable, i.e.,

$$\int_{-\infty}^{Z_{1-\alpha/2}} f(x)\,dx = 1-\alpha/2$$

The length L of the confidence interval is:

$$L = 2\,Z_{1-\alpha/2}\,\sigma/\sqrt{N} \qquad (3)$$

It is clear from this expression that the length of the confidence interval is linearly proportional to the population standard deviation, and inversely related to the square root of the sample size. If σ was known, this expression could be used to determine the minimum sample size needed to obtain a confidence interval which will contain the unknown mean μ with $1-\alpha$ probability. The expression for the minimum sample size is:

$$N \geq 4\,Z^2_{1-\alpha/2}\,\sigma^2/L^2 \qquad (4)$$

For example, if we wish to estimate the average grade of a batch of ore concentrate such that the true average grade is contained within an interval of 1% metal, with 90% probability ($1-\alpha = 0.90$), and if the population standard deviation was known to be also 1% metal, we then proceed by finding $1-\alpha/2 = 0.95$ percentile of standard normal variable from the tables of standard normal integral ($Z_{0.95} = 1.96$) and by the use of above relationship find that the smallest sample size is 16. If the population standard deviation was 2% metal, the minimum sample size would be 62. If we wished to estimate the mean with more precision, say $L = 0.5$% metal, we would need at least $N = 62$ observations for $\sigma = 1$% and $N = 256$ observations for $\sigma = 2$%.

Since σ is usually not known by the sampler, and the sampling process is in fact also directed towards estimating the population standard deviation as well as the mean, the above discussion has a very limited practical consequence. It shows, however, an important relationship between the sample size, the inherent variability of the sampled population and the precision at which the mean μ is to be estimated. This precision is increased by reducing L and/or increasing $1-\alpha$, both resulting in larger sample sizes. Similarly, the variability represented by σ will increase the sample size quadratically. Given the sampling method, the specimen volume, and the bulk to be sampled, σ is an unknown constant. Reduction of σ is only possible by employing a different sampling method, further crushing or grinding the bulk to be sampled, or increasing the specimen volume. All of these changes will introduce a new random variable with the same mean but smaller variances.

As we noted above, the population standard deviation σ is usually not known by the sampler. The variance is estimated from the sample by the following well known formula:

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{X})^2 \qquad (5)$$

Where:

S^2 estimator for σ^2
$\bar{X}$ sample arithmetic average
x_i assay values
N sample size.

Considering the fact that the variance is not known but estimated by S^2, the following probability statement can be made:

Prob.

$$\left[\bar{X} - t_{1-\alpha/2,\,N-1} S/\sqrt{N} \leq \mu \leq \bar{X} + t_{1-\alpha/2,\,N-1} S/\sqrt{N} \right] = 1-\alpha \qquad (6)$$

Where:

$\bar{X}$ sample arithmetic average
S sample standard deviation
N sample size
μ population mean
$t_{1-\alpha/2,\,N-1}$ $1-\alpha/2$ percentile of the Student's t distribution with $N-1$ degrees of freedom.

The length of the confidence interval can be expressed by:

$$L = 2t_{1-\alpha/2,\,N-1} S/\sqrt{N} \qquad (7)$$

From equation (7) we obtain the relationship for minimum sample size as:

$$N \geq 4t^2_{1-\alpha/2,\,N-1} S^2/L^2 \qquad (8)$$

A careful examination of the equations (5) and (8) will show that the values of S and t are not independent of N. Furthermore, one needs a sample of a certain size, say N', to obtain S initially. This value will change as new specimens are sampled, assayed and their values used in recalculating S. This difficulty suggests that when the population variance is unknown it is not possible to determine the minimum sample size by exact means. A practical solution to this problem can be obtained by sequentially sampling and calculating successive approximations of N. To accomplish this, one starts with a small sample size, and estimates the standard deviation by means of equation (5). Using the appropriate values of $t_{1-\alpha/2,\,N-1}$, S and L one obtains a new value for N. If this value is larger than the present sample size, additional sampling is done and the procedure is repeated. As the sample size increases the value of S will approach the unknown population standard deviation with decreasing error. However, since S^2 is a function of random independent observations x_i ($i = 1, \ldots, N$), it is also a random variable whose mean is equal to the unknown population variance; and thus the successive values of S^2, as sample size increases, do not approach the population variance σ^2 from one direction monotonically. They may fluctuate around σ^2 with decreasing error. The consequence of these random fluctuations is that the calculated values of N will also fluctuate. It is, therefore, not advisable to increase the sample size to that of the newly calculated value of N, especially when the number of observations is small and the cost of sampling is to be minimized. The ideal procedure is to increase the sample size by smallest possible increments and continually calculate N until the value of N and the actual sample size are equal or sufficiently close.

The relationship given in (8) suggests certain practical means of designing a sample plan by considering the ratio L^2/S^2. For example, if one wishes to estimate the population mean such that the 95% confidence interval length is equal to a certain multiple of S, then an initial determination of the sample size is possible. By rearranging the terms of the relationship (8) one obtains the relationship:

$$U \geq \frac{4t^2_{1-\alpha/2,\,N-1}}{N} \qquad (9)$$

Where: $U = L^2/S^2$

The right hand side of this inequality is determined by certain values of α and N. The value of the percentiles of the t distribution are available in most of the standard textbooks on statistics. Fig. 1 shows the relationship between U and N for two most commonly used confidence levels. Table 1 lists U values for sample sizes from 2 to 500 for confidence levels 0.90, 0.95, and 0.99. These can be used directly to determine the required minimum sample size for a predetermined value fo U. For example, if one wishes to estimate the population mean using a 95% confidence interval whose length is equal to S, one would look for the intersection of the horizontal line crossing the U axis at $U = 1$ with the curve corresponding to $1-\alpha = 0.95$. In this particular case, the minimum sample size would be $N = 18$. Similarly, from Table 1, one can obtain U values corresponding to given sample sizes directly.

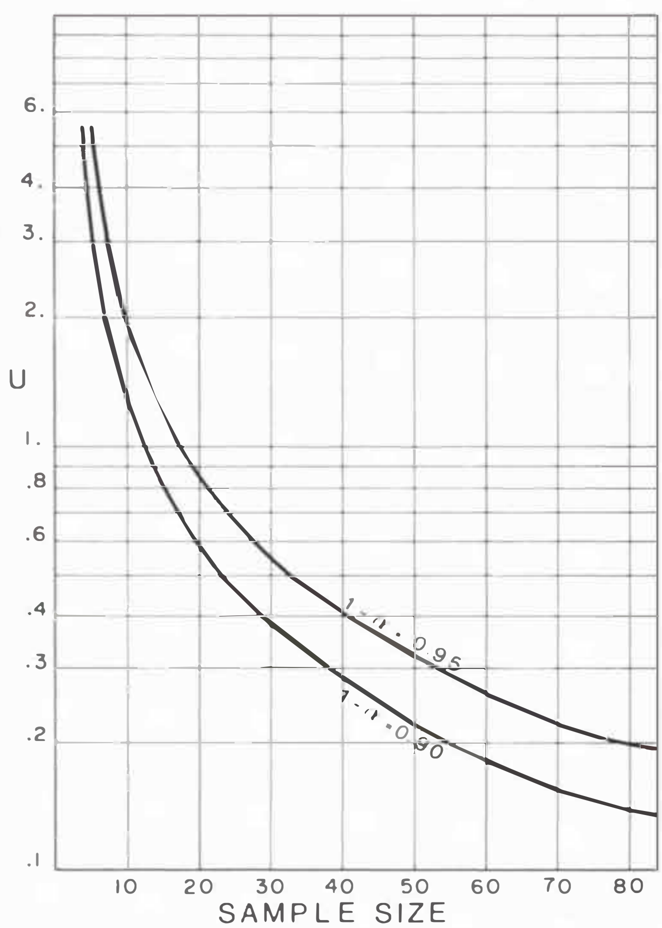

Fig. 1: Relationship between U and N for 2 levels of confidence

Table 1

Sample Size N	$U = L^2/S^2$		
	$1-\alpha = 0.90$	$1-\alpha = 0.95$	$1-\alpha = 0.99$
2	79.63	323.09	8105.19
3	11.37	24.65	131.21
4	5.52	10.11	34.11
5	3.63	6.18	16.93
6	2.69	4.40	10.83
7	2.15	3.43	7.87
8	1.81	2.78	6.13
9	1.54	2.37	5.02
10	1.34	2.04	4.23
11	1.19	1.81	3.65
12	1.08	1.61	3.22
13	0.97	1.46	2.88
14	0.90	1.33	2.59
15	0.83	1.22	2.37
16	0.77	1.13	2.18
17	0.72	1.06	1.89
18	0.67	0.99	1.77
19	0.63	0.93	1.66
20	0.60	0.87	1.64
21	0.56	0.83	1.54
22	0.54	0.79	1.46
23	0.51	0.75	1.38
24	0.49	0.71	1.32
25	0.47	0.68	1.25
26	0.45	0.65	1.20
27	0.43	0.63	1.14
28	0.41	0.60	1.10
29	0.40	0.58	1.05
30	0.39	0.55	1.02
40	0.28	0.41	0.73
50	0.22	0.32	0.57
60	0.19	0.27	0.47
80	0.14	0.20	0.35
100	0.11	0.16	0.28
200	0.05	0.08	0.14
500	0.02	0.03	0.05

5. A Numerical Example

To illustrate the use of the sequential sampling scheme described above we shall consider the following numerical example.

A batch of copper concentrate is to be sampled with the purpose of determining its average copper content, such that the true average grade μ would not deviate from the arithmetic average of sample assays more than $\pm$ 1 % Cu with 95 % confidence, i.e.,

$$\text{Prob.}\left[\,|\bar{X}-\mu| \leq 1\,\%\ \text{Cu}\right] = 0.95.$$

Furthermore, the sampler would like to make this assertion with as little as possible sampling.

Given the specimen volume, the assay value corresponding to each specimen is an observation of a random variable whose mean is the same as the average grade of the whole bulk, and the variance is a function of the specimen volume. Suppose that in this example the average grade of the con-liberation size of Cu minerals in the concentrate, the sampler decides to take 1 lb specimens by a random selection process. The random variable being observed by this selection process would have the same mean as the average grade of the bulk concentrate. Let the variance associated with the selection process be 4 (σ = 2 % Cu). The sampler without the knowledge of the value of the variance decides to sample few specimens and determine their copper content, and decide whether to sample more or not on the basis of his analysis of the presently available assay data. Table 2 shows the successive outcomes of such an experiment. The first column indicates the order that specimens are sampled by means of random selection. The second column lists the assay values corresponding to these specimens. The third column lists, cumulatively, the number of specimens sampled up to that point (sample size). The fourth column is the arithmetic average of assay values of the specimens up to and including the corresponding row. In the fifth column the sample standard deviation is listed. The sixth column shows the required minimum sample size calculated by using (8) for $1-\alpha$ = 0.95, and the last column shows the calculated minimum sample size for $1-\alpha$ = 0.90. A close look at Table 2 shows that the calculated sample size fluctuates widely for small sample sizes but later converges towards the actual sample size. For $1-\alpha$ = 0.95 the required sample size becomes stable between 18 and 19 after the actual sample size reaches 22. If the sampler was using the sequential scheme, he would keep sampling until the sample size reaches 21, where the calculated minimum sample size becomes less than the actual sample size. The same situation occurs for $1-\alpha$ = 0.90 at N = 16. Note that if the population standard deviation was known (2 % Cu) the required minimum sample size could have been calculated by using (4). In this particular example, N would be 16 for $1-\alpha$ = 0.95, and 11 for $1-\alpha$ = 0.90. The need for larger sample sizes when the variance is also estimated by sampling is the penalty for not knowing the true value of the variance.

Let us now consider the alternate method of determining the sample size by means of the ratio $U = L^2/S^2$. For a selection of U = 1, the required minimum sample size is 18 for $1-\alpha$ = 0.95 and 13 for $1-\alpha$ = 0.90. In the first case (N = 18) the length of the confidence interval is 2.14 % Cu (S = 2.14 % Cu), and in the second case it is 2.31 % Cu (S = 2.31 % Cu). If the sampler was not satisfied with this result, he would select a smaller value for U and continue to sample. For example, if he decides to select U = 0.8, he then will have to sample additional 4 and 3 specimens for 95 % and 90 % confidence intervals, respectively. In this case he will find that the 95 % confidence interval length is 1.71 % Cu (0.8 $S = L$) and the 90 % confidence interval length is 1.81 % Cu.

References

[1] Cramér, H., "The Elements of Probability Theory and Some Applications", John Wiley & Sons, N.Y. 1955

[2] Gy, P.M., "Sampling of Particulate Materials — Theory and Practice", Elsevier, Amsterdam, 1979

[3] Gy, P.M., "The Sampling of Particulate Materials — A General Theory", Inter. Jour. of Mineral Processing, V.3, pp. 289—312, 1976

[4] Gy, P.M., "Calculateur d'échantillonage — Sampling Slide Rule", Soc. Ind. Min. Saint-Etienne, France, 1965

[5] Hogg, R.V., and Craig, A.T., "Introduction to Mathematical Statistics", The Macmillan Co., N.Y., 1958

[6] Shewhart, W.A., "Economic Control of Quality of Manufactured Product", Princeton, N.J., Van Nostrand Co. Inc., 1931

[7] Shewhart, W.A., "Statistical Method from the Viewpoint of Quality Control", Washington D.C., The Graduate School, Dept. of Agriculture, 1939

Table 2

Specimen No.	Assay Value X (% Cu)	Sample Size N	Sample Mean $\bar{X}$ (% Cu)	Sample S. Deviation S (% Cu)	Calculated Sample Size for $L = 2\%$ Cu	
					$1-\alpha = 0.95$	$1-\alpha = 0.90$
1	22.75	1	22.75	—	—	—
2	21.55	2	22.15	0.85	117	29
3	25.82	3	23.37	2.20	90	42
4	26.66	4	24.20	2.44	61	33
5	26.47	5	24.65	2.34	43	25
6	25.08	6	24.72	2.10	30	18
7	23.93	7	24.61	1.94	23	15
8	22.86	8	24.39	1.90	21	14
9	24.32	9	24.38	1.78	17	11
10	25.99	10	24.54	1.75	16	11
11	30.60	11	25.09	2.47	31	20
12	25.07	12	25.09	2.36	27	19
13	26.86	13	25.23	2.31	26	17
14	27.28	14	25.38	2.28	25	17
15	22.78	15	25.20	2.30	25	16
16	23.52	16	25.10	2.26	24	15
17	24.27	17	25.05	2.20	22	15
18	25.44	18	25.07	2.14	21	14
19	22.58	19	24.94	2.15	21	14
20	22.84	20	24.83	2.15	21	14
21	24.17	21	24.80	2.10	20	14
22	22.59	22	24.70	2.10	19	13
23	26.31	23	24.77	2.08	19	13
24	27.00	24	24.86	2.08	18	13
25	26.99	25	24.95	2.08	18	13
26	22.43	26	24.85	2.10	19	13
27	25.86	27	24.89	2.07	18	13
28	27.43	28	24.98	2.09	19	13
29	26.99	29	25.05	2.08	19	13
30	25.88	30	25.08	2.05	18	13