Author	Brian W Woodget
Owner	Royal Society of Chemistry
Title	Additional notes on Statistical Sampling
Classification	F180, Analytical Chemistry
Keywords	ukoer, sfsoer, oer, open educational resources, metadata, analytical science, cpd training resource, analytical chemistry, measurement science, skills for analytical science, sampling, statistical sampling.
Description	A pdf document explaining in diagrammatic terms and some equations, how statistical sampling can help in the design of sampling strategies. This document will extend the discussion and coverage on 'Sampling' to be found in Chapter 2 of the double modular teaching & learning programme in Analytical Science.
Creative Commons licence	http://creativecommons.org/licenses/by-nc-nd/2.0/uk/
Language	English
File size	190 kBytes
File format	Microsoft Word (1997 – 2003)

A double module teaching & learning programme in Analytical Science, at 2nd year England & Wales Undergraduate level

https://edocs.hull.ac.uk/muradora/objectView.action?parentId=hull%3A2199&type=1&start=10&pid=hull%3A2351

Maximising accuracy and reliability in sampling

1.0 Introduction

As already stated in Chapter 2 of this resource, the sampling of materials, will normally incur a measure of uncertainty.

Definition of 'Measurement uncertainty'
An estimate of the range within which a true value of a measurand (analyte concentration/abundance) will lie

This uncertainty can only be minimised by careful attention to the development of the sampling plan devised for the taking of the samples and the care exercised in the proper storage and labelling of the samples taken. Heterogeneous materials will incur higher sampling errors than homogeneous materials, **with the error increasing with the extent of heterogeneity**.

When attempting to sample any heterogeneous material, the measure of uncertainty, can be statistically reduced, by the taking and analysis of more samples. Consider for instance the sampling of a field using the sample plot [Figure (01)].

The field has been divided into 16 strata, with two samples being taken randomly, from each stratum. These samples are denoted by clear squares.

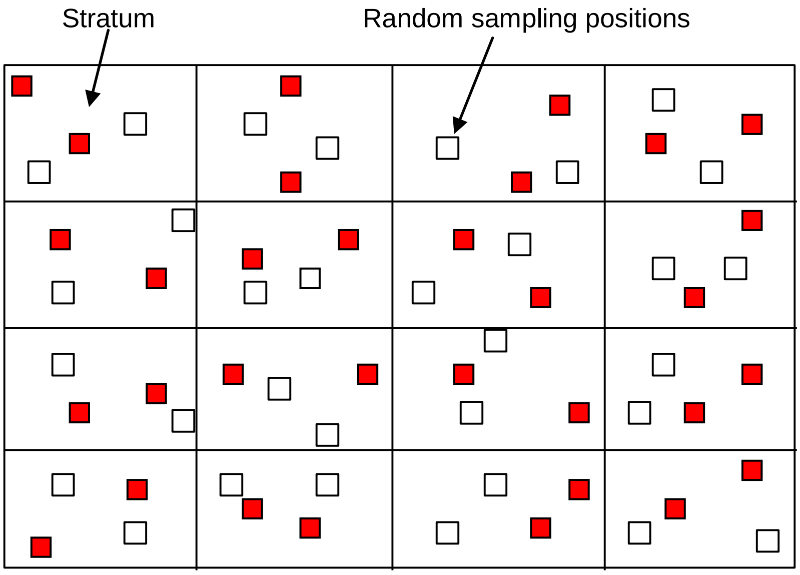

Figure (01) – stratified random sampling design

In order to reduce the likely measure of uncertainty, we can either increase the number of strata, again taking two sample increments form each strata (clear squares) or alternatively increase the number of samples taken from each stratum. Figure (01) illustrates the latter decision being made, the red squares (🟥) showing the additional random samples to be taken from each stratum.

The decision now has to be made, both with the original 32 sample increments, or with the new 64 sample increments, as to how many samples are to be sent for analysis. The decisions we can make are illustrated in Figure (02)

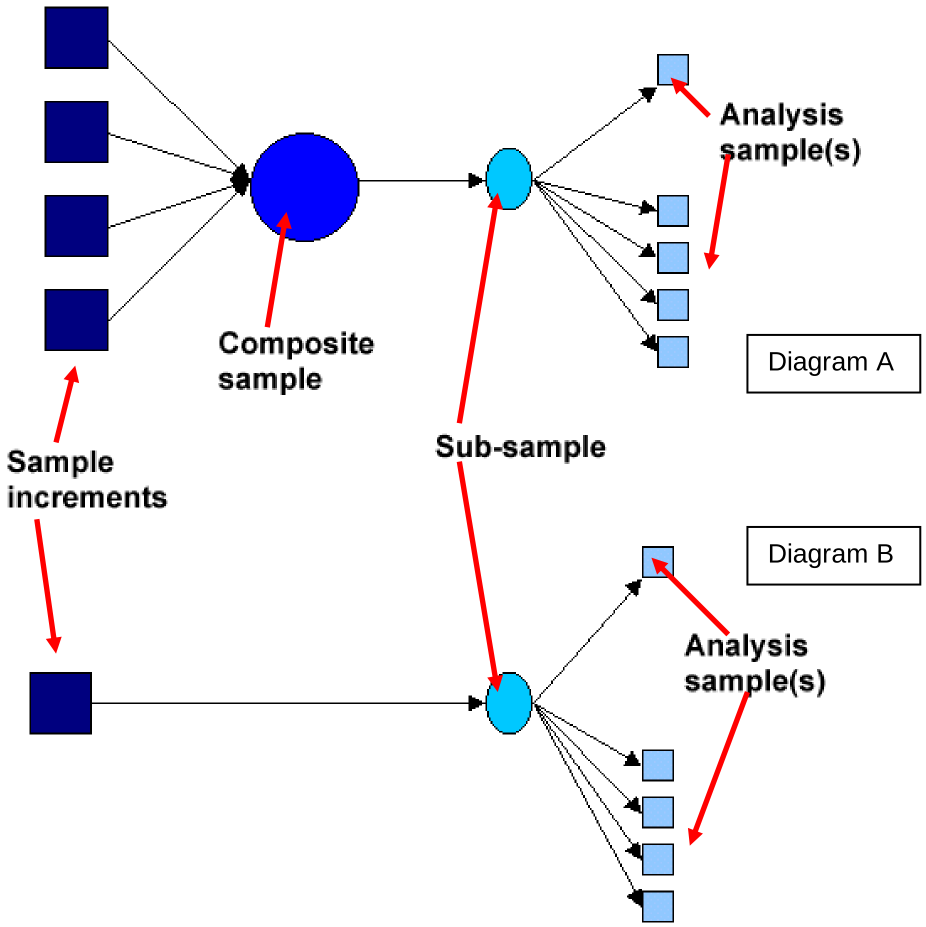

Figure (02) – selection of samples for analysis

'Diagram A' in Figure (02) shows the sample increments being composited, sub-sampled and either a single sample then supplied for analysis or replicate samples supplied for analysis. [**Maximum of 4 analyses – assuming 4 replicate analyses on each sub-sample**]
'Diagram B' in Figure (02) shows each increment being sub-sampled and either a single or replicate analysis samples being derived from this. [**Maximum of 256 analyses when related to Figure (02) – that is 64 sample increments with 4 replicate analyses on each**]

'A' represents the cheapest option with the highest measurement uncertainty whilst 'B' represents the most expensive option with the lowest measurement uncertainty, especially if replicate samples are chosen for analysis. A compromise alternative could be to composite together samples taken from each sampling stratum, sub-sample and then to select for analyse, either a single or replicate portions. [**Maximum of 16 or 64 analyses when related to Figure (02)**] The real choice will depend upon the reason for the analysis and the level of information required.

- If for instance we wish to know the average levels of soil nutrients in the field and are prepared to accept a high level of measurement uncertainty, then the scenario represented by 'Diagram A' may be satisfactory with replicate analyses of the single sample being carried out.

- If on the other hand we suspect that parts of the field may be suffering from chemical contamination, then samples from all strata will need to be analysed, probably by compositing sample increments from that stratum and performing a single analysis on the composite sample. If contamination is found to be present in one or more of the strata, then these can be re-examined at a later time.

2.0 Statistical sampling

2.1 Introduction

Note: this section on 'Statistical Sampling' assumes some knowledge of statistical procedure as used within Analytical Science. It highlights just those elements that have an influence on sampling procedures. For a fuller description of Statistics, learners are referred to Chapter 5 of the Teaching and Learning package in Analytical Science.

Definition of Statistical sampling

Statistical sampling is based on the premise that all particles or portions of the material (population) to be sampled have an equal probability of being chosen as the sample.

In using this definition, it is also assumed that the analyte to be measured, has a normal distribution within the population. By this we mean that if it were possible to take an infinite number of samples from the population and to measure the analyte concentration in each, then the frequency distribution of the analyte should show a normal distribution, as illustrated in Figure (03)

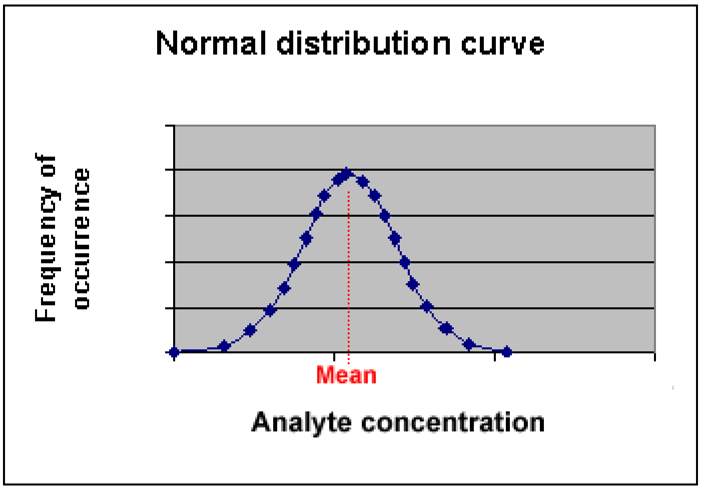

Figure (03) – normal distribution curve

The normal distribution curve is symmetrical about the mean value of the measurements. The spread of the curve (ie: the range of concentration found for the analyte) is a function of the standard deviation (S) for the analysis. The larger the standard deviation, the greater is the spread.

The total variance (S^2_{total}) shown by an analyte following analysis, is an additive combination of sampling ($S^2_{sampling}$) and analysis ($S^2_{analysis}$) variances [equation 01].

$S^2_{total} = S^2_{sampling} + S^2_{analysis}$ [Equation 01]

Where we are sampling a material, which can be considered as essentially homogeneous ($1dm^3$ of a single phase liquid for example), then the majority of this variance may well be due to the analysis stages - those stages in the analytical process following that of sampling. [see slide 3 in Chapter 2 of this resource]. However where there is a heterogeneous distribution of analyte within the material matrix (many solid samples for example), then the majority of the variation will arise as a result of real differences in analyte distribution. Under these circumstances, the analysis stage will make only a minor, and sometimes almost negligible, contribution to the total variance**. It is normally assumed that when one variance exceeds the other by greater than 10 times, then the smaller of these can be ignored.**

Let us consider some of these points in terms of statistical and algebraic equations. The algebraic terms used are given in table (01), shown at the end of this section.

Equation (01) has shown that the total variance in any analysis is a combination of variances from all stages in the analysis. Where the analysis is complex, the $S^2_{analysis}$ term may need to be subdivided to take into account variances introduced during the:

- Sample preparation stage;
- Separation and or concentration stage;
- Measurement stage.

These stages, in some situations, can result in large variances being introduced into the total variance [S^2_{total}] and analytical development work may need to be carried out to reduce their overall significance. For the purpose of this discussion however, all of these potential variances will be considered under the one single 'analysis' variance.

From equation (01), the sampling variance may be represented as shown in [equation 02]

$$S^2_{sampling} = S^2_{total} - S^2_{analysis} \quad \text{[Equation 02]}$$

The analysis variance may be calculated by carrying out the whole analysis on samples that are known to be homogeneous. A minimum of seven replicate samples will need to be analysed and the variance calculated by squaring the resultant standard deviation.

The sampling variance is also an additive combination of two other individual variances:

- That due to the real variation of analyte distribution within the material to be sampled, termed the population variance [$S^2_{population}$];
- That due to the actual sampling variance [$S^2_{practical}$] – how good we are at taking the sample

$$S^2_{sampling} = S^2_{population} + S^2_{practical} \quad \text{[Equation 03]}$$

By careful attention to the sampling regime, it should be possible to reduce the value of [$S^2_{practical}$], however the sampler has no control over the population distribution [$S^2_{population}$]. The size of these two components will influence the number of samples that need to be taken, in order to achieve an acceptable measure of measurement uncertainty.

Some important sampling and measurement scenarios may now be considered.

Parameter	**Symbol**
True mean	μ
Estimated mean	$\bar{x}$
Standard deviation	S
Variance	S^2
Statistical term from 't' test table	t
Number of samples taken	n_s
Number of replicate analyses	n_a

Table (01) – algebraic symbols used in this statistical discussion

2.2 Measurement situations

2.2.1 Where the sampling variance is significant and the measurement variance is insignificant.

The first step is to decide upon the acceptable level of measurement uncertainty for the target analyte – that is the difference between the true mean concentration/abundance [μ] and the mean of the results obtained [$\bar{x}$]. The relationship between these two mean values is given by equation (04)

$$\mu = \bar{x} \pm tS/\sqrt{n_S} \quad \text{[Equation 04]}$$

The next step is to decide upon the level of confidence that is acceptable for the measurement. It is usual to choose a 95% confidence level, which means that in 5% of cases, μ would be outside the calculated uncertainty limit. For a normal distribution situation and where the value of $n > 30$:

The value of 't' approximates to 2.

In developing a sampling plan, equation (04) can be used to calculate how many increments are needed to be taken in order to achieve our accepted level of measurement uncertainty. Firstly however, we need to obtain a value for 'S' (standard deviation) for this particular sampling and analysis. This can be achieved by taking a large number of sample increments, analysing the samples and calculating the standard deviation and/or the variance.

If 'E' represents the level of total uncertainty we have accepted for this analysis then:

$$E = tS/\sqrt{n_S} \quad \text{[Equation 05]}$$

Squaring each side gives:

$$E^2 = t^2S^2/n_S \quad \text{[Equation 06]}$$

Or by rearranging gives:

$$\mathbf{n_S = t^2S^2/ E^2} \quad \text{[Equation 07]}$$

It is therefore possible to calculate the number of sample increments that need to be taken by calculating 'S', accepting 'E' and interpolating 't' from Student's 't' test table.

Example (i)

We shall assume that for a single consignment of material from which 30 replicate sample increments were removed and analyses carried out, the value of the standard deviation for these measurements was 0.187. If we set the total uncertainty we are prepared to accept between the estimated mean and the actual value to be 0.15, we can now calculate the number of increments we shall need to take from any future consignments of this type.

Equation (07) gives us the formula we can use for this calculation in order to calculate n_s and assuming that we are prepared to have a 95% confidence level then:

$$n_s = [2^2 \text{ X } (0.187)^2]/ (0.15)^2$$

$$= [4 \text{ X } 0.035]/0.0225$$

$$\mathbf{= 6.2}$$

This calculation shows us that we need to take 7 samples in order to comply with our accepted level of measurement uncertainty. However, we assumed that the value of 't' was 2, indicating a minimum of 15 sample replicates.

So we now need to carry out some trial and error calculations to hone in on a more accurate number. From table (02) shown at the end of this section, we can see that value of 't' that corresponds to the number of degrees of freedom (n – 1).

For n = 7, t = 2.45 :

This calculates to a possible error of 0.173, above that deemed to be acceptable.

For n = 8, t = 2.36 :

This calculates to a possible error of 0.156, again just above the target of 0.15.

For n = 9 , t = 2.31 :

This calculates to a possible error of 0.144, just below the target value.

So the correct answer is 8 or 9.

2.2.2 Where the sampling variance is insignificant and the measurement variance is significant.

As with the example above, the value of 'S' needs to be calculated and 'E' needs to be agreed.

To calculate 'S', a single representative sample is taken and analysed 'n' times. Equation (08) may then be applied to calculate the number of replicate analyses that need to be carried out from the taking of a single sample.

$$\mathbf{n_a = t^2S^2/ E^2} \quad \text{[Equation 08]}$$

Note: as with the example shown above, it will be necessary to carry out a trail and error calculation to establish an accurate number once the approximate value has been calculated.

2.2.3 Where the sampling and measurement variances are both significant

This is a more difficult situation, as we now need to know individual standard deviations for both the sampling and analysis stages. Equation (09) could be used if sufficient data were available or could be calculated:

$$E_{total} = t \sqrt{[(S_s^2/n_s) + (S_a^2/n_s n_a)]} \qquad \text{[Equation 09]}$$

Equation (10) may be rewritten as:

$$(E_{total})^2 = t^2 [(S_s^2/n_s) + (S_a^2/n_s n_a)] \qquad \text{[Equation 10]}$$

Where it is suspected that there may be a small but real variation of a component within a consignment, then by using the statistical method of ANOVA (analysis of variance), it is possible to discriminate between the random error obtained from the analysis and the real variation in quality. For a description of how this may be achieved, learners are directed to: http://en.wikipedia.org/wiki/ANOVA

Degrees of Freedom	Value of ‘t’ at a confidence level of 95%
1	12.7
2	4.30
3	3.18
4	2.78
5	2.57
6	2.45
7	2.36
8	2.31
9	2.26
10	2.23
11	2.20
12	2.18
13	2.16
14	2.14
>30	1.96 (2)

Table (02) – values of ‘t’ at the 95% confidence level

3.0 Sampling strategies

In Chapter 2 of this resource, the competing decisions between cost and reliability when devising a sampling plan were introduced and expanded upon earlier in this document. Figure (02) was used to show this decision making process diagrammatically. Figure (02) is reproduced below:

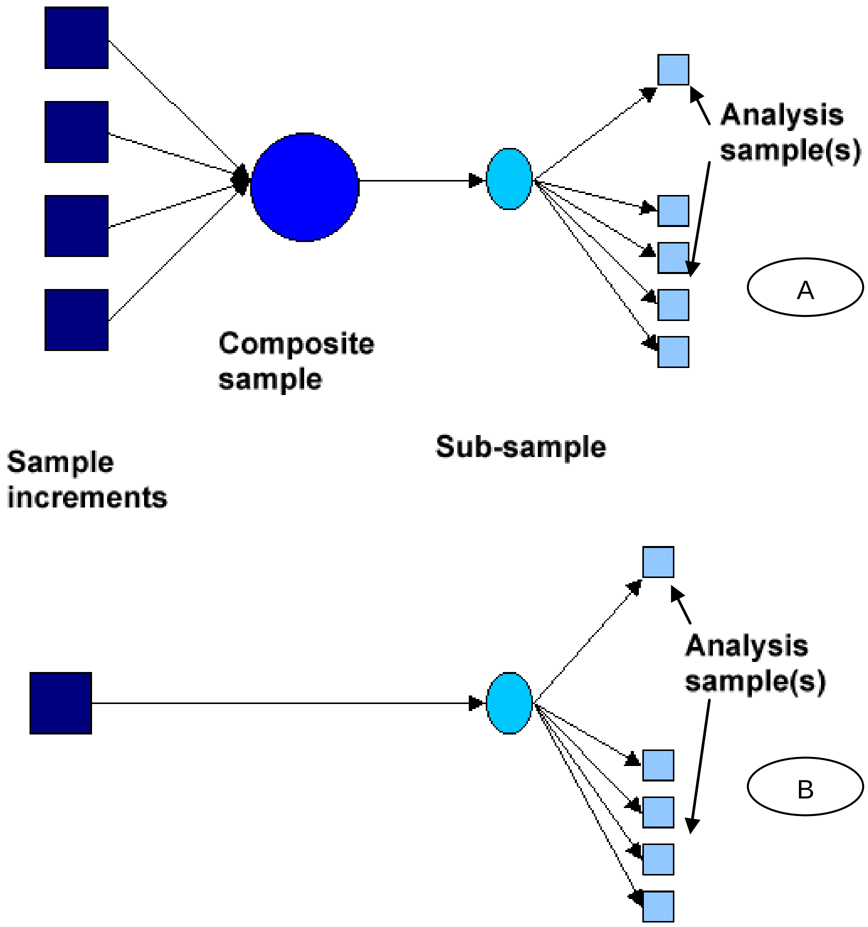

Figure (02 repeated) – selection of samples for analysis

Given the statistical argument that was introduced in section (2) we can now use these equations to justify the sampling and analysis suggestions made in the earlier section.

If one analysis is made on n_s sample increments then the relationship between true mean and the estimated mean is given by equation (04):

$$\mu = \bar{x} \pm tS/\sqrt{n_s} \qquad \text{[Equation 04]}$$

Where ‘S’ relates to the to the overall standard deviation for the sampling and analysis and ‘S^2’ is the total variance. The measurement uncertainty at a chosen confidence level is given by:

$$t\,[S/\sqrt{n_s}] \qquad \text{[Equation 11]}$$

If each single sample increment is analysed ‘N’ times, then the measurement uncertainty now becomes:

$$t\,\sqrt{[(S_a^2/N + S_s^2)/\,n_s\,]} \qquad \text{[Equation 12]}$$

Equation (12) may be rewritten as:

$$t\,\sqrt{[(S_a^2/n_sN) + (S_s^2/\,n_s)]} \qquad \text{[Equation 13]}$$

For maximum analytical reliability we need to minimise the value of equation (13). The term 'S_a^2', may be reduced by either choosing a more precise analytical method or by increasing the number of samples taken 'n_s'. However if the sampling variance 'S_s' is much greater than the analysis variance 'S_a', then little is to be gained by using different analytical methods. Rather it is preferable to take a larger number of sample increments, since the value of 't' is dependent upon the value of 'n_s', and gradually decreases for increasing sample increment numbers, up to 15.

Formulae shown in (Equations 11 to 13) may be modified again to take into account the sampling and analysis strategy illustrated in Figure (02) whereby the sample increments are composited and following sub-sampling, undergo replicate analysis.

$$t \sqrt{[(S_a^2/N) + (S_s^2/ n_s)]} \quad \text{[Equation 14]}$$

In this instance, as only a single composite sample has been submitted for analysis, the value of 'n_s' becomes 1. Uncertainty measurements are therefore much larger, than when single increments are all analysed, however the costs are considerably reduced.

RSC | Advancing the Chemical Sciences