Particulate Materials

New Theoretical Approach

B. Geelhoed

faculty Reactor Institute

[illegible]

Sampling of particulate materials

[illegible]

New theoretical approach

[illegible]

[illegible]

Cover design by: A. A. E. Stolte

The research described in this thesis was conducted at the Department of Radiochemistry, Interfaculty Reactor Institute, Delft University of Technology. The project was funded by The Netherlands Forensic Institute (NFI).

Sampling of particulate materials

New theoretical approach

Proefschrift

ter verkrijging van de graad van doctor

aan de Technische Universiteit Delft,

op gezag van de Rector Magnificus prof. dr. ir. J. T. Fokkema,

voorzitter van het College voor Promoties,

in het openbaar te verdedigen op

dinsdag 7 september 2004 om 13:00 uur

door

Bastiaan GEELHOED

doctorandus in de natuurkunde

geboren te Amsterdam

Dit proefschrift is goedgekeurd door de promotoren:

Prof. dr. ir. H.J. Glass
Prof. dr. ir. J.J.M. de Goeij

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. ir. H.J. Glass	Universiteit van Exeter, promotor
Prof. dr. ir. J.J.M. de Goeij	Technische Universiteit Delft, promotor
Prof. dr. J. Pilz	Universiteit van Klagenfurt
Prof. dr. M.H. Ramsey	Universiteit van Sussex
Prof. dr. F.M. Dekking	Technische Universiteit Delft
Prof. dr. P.H. Franses	Erasmus Universiteit Rotterdam
Dr. A. Bolck	Nederlands Forensisch Instituut

Dr. ir. P. Bode heeft in belangrijke mate aan de totstandkoming van het proefschrift bijgedragen.

Published and distributed by:

DUP Science, an imprint of Delft University Press
P.O. Box 98
2600 MG Delft
The Netherlands
Telephone: +31 15 27 85 678
Telefax: +31 15 27 85 706
E-mail: Info@Library.TUDelft.NL

ISBN 90-407-2517-9

Keywords: statistics, granular, estimation, standardization

Printed in The Netherlands

Contents

Chapter 1 Introduction

Standardization of sampling requires that the mass or the volume of the sample is prescribed. In current standards, a prescribed value for the sample mass is derived using empirical relations between assumed properties of the batch and the variance of the sampling error. The potentially inaccurate empirical relations and assumed batch properties may lead to an underestimation or overestimation of the potential magnitude of the sampling error. Therefore, the objective of the research described in this thesis is the development of a new, non-empirical theory for the sampling of particulate materials, to allow for calculation of the minimum sample mass in sampling standards. The positioning of this research is further clarified in several introductory paragraphs.

1.1 Sampling of materials

Chemical, physical or biological properties of gaseous, liquid or solid materials may be important for economic, agricultural, environmental and/or health-related reasons. These properties are determined by corresponding measurements, and several stages can be distinguished relating the material under study ('the batch') to the final analysis result.

Usually a batch contains more material to be analyzed than can be covered by a single sample analysis. Therefore, an analyzable fraction has to be extracted. This analyzable fraction is termed here as the 'laboratory sample'. Depending on the characteristics of the material and of the analysis technique, a sample size exists for the laboratory sample. Unfortunately, it cannot be guaranteed that the properties of a sample of this optimum size drawn directly from the batch are representative for the properties of interest of the batch. Therefore, a larger bulk sample, representing (in properties of interest) the batch, is drawn first: the 'bulk sample'. After homogenization of this bulk sample by *e.g.* milling, blending and/or mixing, a laboratory sample, representing the bulk sample, is drawn. Often, the laboratory sample is not immediately fit for analysis by the required analysis technique and an additional preparation is needed to attain a test portion. Sample dissolution is an example of a sample preparation step of the laboratory sample. Summarizing, the stages leading from batch to the final analysis result are: drawing of the bulk sample, homogenization of the bulk sample, drawing of the laboratory sample, sample preparation towards a test portion and analysis of this test portion (see Figure 1.1).

Each of the above-described stages should be performed in such a way that the difference, here indicated as the total error, between the analysis result and the corresponding true value of the measurand in the batch remains as small as possible, ideally zero. In practice, these differences are not zero, thus cumulating in a finite value for the total error.

In the following, it is assumed that the batch property, which is estimated, is the mass or volume concentration of a component in the batch[1]. The first source of error is the difference between the concentration in the bulk sample and the concentration in the batch. This is a sampling error. When sampling a batch consisting of particles, the sampling error caused by the random sampling of non-identical particles is often denoted as the 'fundamental error' (Gy, 1979).

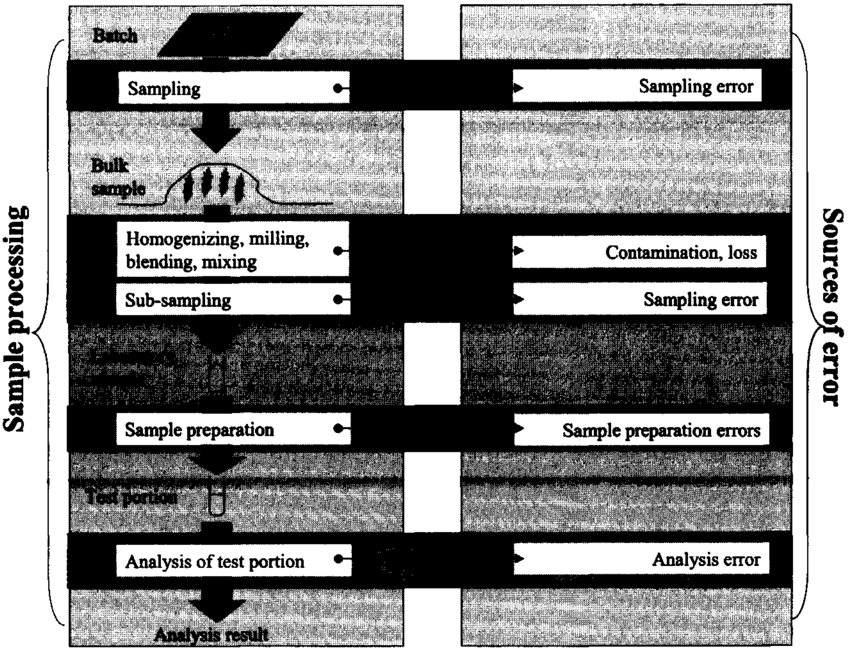


Figure 1.1. Sources of error during all the stages of the process going from batch to final analysis result.

Secondly, during the homogenization, mixing, milling and blending, loss of or contamination with the compound of interest may occur in the bulk sample. There may also be a sampling error during subsampling, defined analogously to the sampling error during the drawing of the bulk sample: the difference between the concentration in the subsample or laboratory sample and the concentration in the bulk sample.

Finally, sources of error can be sample preparation steps necessary for the analysis of the laboratory sample. An example is incomplete digestion of the laboratory

[1] The mass or volume concentration of a component in a batch, sample or particle is defined as the mass or volume of the component in the batch, sample or particle respectively, divided by the total mass or volume of the batch, sample or particle respectively.

sample if the test portion should be in the liquid state. Also the analysis technique itself includes sources of error; the analysis error is defined as the difference between the analysis result and the true value of the analyte concentration in the test portion (ISO, 1993).

The analysis error and errors due to sample preparation are not the main subject of this thesis, because these errors are generally well-characterized and well-understood as part of the method validation, and assessed using quality control materials. Also the sources of error due to loss or contamination during mixing, milling and blending of the bulk sample can be minimized by a good choice of sample preparation equipment, and are therefore not the main topic of this thesis.

Sampling errors, on the other hand, are more difficult to control. In the next paragraph, it is discussed how this problem is approached in current sampling standards. The Dutch standard NEN 5742 is taken as an example of a typical standard. This standard relies on a sampling theory based on empirical relations between assumed batch properties. Since these relations and assumed batch properties are potentially inaccurate, the actual sampling errors may be larger than expected.

1.2 Sampling standards and theories

The difference between the estimated value derived from measurement of a specific sample and the corresponding true value of the measurand[2] in a batch (the value of the total error) is unknown since it would require the true batch value. To reduce the occurrence of large positive or negative values of the sampling errors, but also to standardize sampling, sampling standards have been devised. Examples of sampling standards are the NEN 5742 (see NEN, 2001), ASTM C1075-93 (see ASTM, 1997), ASTM D1900-94 (see ASTM, 2002), ISO 11648-2 (see ISO, 2001) and ISO 11648-1 (see ISO, 2003) publications.

In the Dutch standard NEN 5742, of which a more detailed description is presented in the Appendix, a prescribed value of the mass that has to be sampled is given. This value is chosen in such a way, that the standard “guarantees” that the relative standard deviation of the sampling error does not exceed 10%. However, it will be seen that the reliability of this assertion is questionable.

The Dutch standard NEN 5742 defines scope (sediments and soils), measurands (metals, inorganic compounds, semi-volatile organic compounds and physico-chemical soil properties) and sampling devices to be used. The standard prescribes the way of sampling, which includes a prescribed value for the mass of the sample. Finally, packaging, conservation and transport of the drawn samples and the essential elements

[2] From hereon, the following shortcuts will be used:

- Value of the measurand in a batch /true value of the concentration of an analyte in a batch: batch value / batch concentration
- Value of the measurand in a sample /true value of the concentration of an analyte in a sample: sample value / sample concentration
- Value of the measurand in a particle /true value of the concentration of an analyte in a particle: particle value / particle concentration

of reporting are described.

The prescribed value for the sample mass is calculated in the NEN 5742 using Gy's theory of particulate materials (Gy, 1979) and assumed properties of the sampled batch. It is assumed that the maximum particle size is 10 mm, density of the particles is 2600 kg·m^{-3} and the fraction of particles containing the property of interest is 0.1. Using these assumptions a prescribed sample mass is obtained for which Gy's theory predicts that its relative standard deviation is 10% or less.

An obvious drawback is that these assumptions limit the general applicability of this Dutch standard. Even if the assumptions are correct, the relative standard deviation may still be larger than 10% due to possible flaws in Gy's theory. Moreover, if an alternative prescribed sample mass were to be calculated on the basis of different estimates for the maximum particle size, density and fraction of particles containing the property of interest still using Gy's theory, the relative standard deviation could be larger than 10% due to errors in the assumed batch properties.

1.3 Scope of the thesis

The work presented in this thesis aims to improve the scientific basis for sampling standards by developing a new sampling theory for the sampling of randomly mixed batches of particulate material. New fundamental theoretical work has been done with associated computational and experimental verification. The development of the new sampling theory is outlined in the next paragraphs.

1.4 Characterization of materials

All matter may be regarded as a collection of indivisible units. Applied to sampling, 'indivisible' means that the unit does not split or break up into several pieces during sampling. This is the starting point for the development of a new sampling theory. For the sampling of particulate materials, the units are represented by solid particles.

The shape of the particles can take any form, from simple spheres or cubic particles to particles that are shaped like dendrites. Also the dimensions can vary greatly, from spherical colloids with diameters ranging from several nm to 1 μm to pieces of wood of several decimetres length. Soil is an example of a particulate material that has a broad range of particle sizes, with typical dimensions varying from several micrometers to several centimetres. Crushed rocks have generally also much variation in size; often the particle mass distribution is lognormally distributed. Biological materials like soybeans or coffee beans are more uniformly distributed and have less variation in size and shape.

Variation in composition of particles can lead to differences in estimates derived from distinct samples. This implies that variation in particle composition is a source of sampling error when sampling particulate materials. Samples containing only a few particles are prone to large sampling errors, because the average composition of the particles in the sample can be very different from the average particle composition in the batch. Because homogenization reduces the average particle size, the laboratory sample contains generally more particles than the bulk sample. This is one of the reasons that sampling errors are generally much larger during the drawing of the bulk sample than during the drawing of the laboratory sample from the homogenized bulk sample. Application of the new sampling theory is therefore expected to be more relevant to the drawing of bulk samples than to the subsampling, although, in principle, the new theory should also be applicable for subsampling.

Finally, it is noted that in this thesis the term particulate material is also used for materials with interstitial fluid or gas, provided that the property of interest in the sample and batch are exclusively determined by the solid fraction.

1.5 Development of a new theory

As discussed in Paragraph 1.2, there are several potential drawbacks with the calculation of the required sample mass in the NEN 5742, which uses Gy's theory. For the development of the new theory to be used for calculation of the minimum sample size in standards, the following eight criteria were chosen:

Criterion 1

- The theory must provide an equation for the variance of the sample concentration, containing the mass or volume sampled and an arbitrary number of additional parameters.

This equation can subsequently be used to derive an equation for the minimum mass or volume to be sampled when it is demanded that the relative standard deviation is equal to or smaller than a preselected value (often set to 0.1). A theoretical basis for the derivation of the equation for the variance is obtained when the second criterion is met:

Criterion 2

- The theory must be based on a model of the drawing of a sample on the level of (groups of) particles.

Because real batches may contain a wide range of different types of particles, the sampling theory should be applicable for batches containing multiple distinct types of particles, with arbitrary particle masses and concentrations. This yields a third criterion:

Criterion 3

- The theory must be applicable to batches containing any number of distinct types of particles, with arbitrary particle masses and concentrations.

It is noted that the equation for the variance, mentioned in the first criterion, may contain parameters other than the mass or volume sampled. Examples are the mass or volume of the batch, the particle mass or volume and the concentration in a particle. For calculation of a numerical size-variance equation (*e.g.* variance = 12.34 divided by the sample mass expressed in kg), a numerical evaluation of the other variables is required. This is the underlying reason for the formulation of criterion 4 to 6.

Criterion 4

- The theory must allow determination of the parameters of the size-variance equation, using the measured sample concentrations of one or more samples of a given size.

The equation relating the variance to the mass or volume sampled may depend on the properties of the particles in the batch. Examples are the distribution of particle masses, particle volumes or concentrations in the particles of the batch. In these cases, some prior knowledge of the properties of the particles in the batch is required. When this knowledge is available, a good sampling theory should provide the opportunity to use this prior knowledge for calculation of the variance. This may lead to a better estimate of the true variance. These considerations lead to the following criterion:

Criterion 5

- The theory must allow determination of the parameters of the size-variance equation, using prior knowledge of the properties of the particles in the batch.

Analogous to the above situation, knowledge of the properties of the particles in the sample may be available. Examples of properties of particles in the sample are the distribution of particle masses, particle volumes or concentrations of the particles in the sample. Therefore,

Criterion 6

- The theory must allow determination of the parameters of the size-variance equation, using posterior knowledge of the properties of the particles in the sample.

Both mass and volume concentrations are important in practice. Therefore, the seventh criterion is:

Criterion 7

- The sampling theory must be applicable to mass and volume concentrations.

In Chapter 2, it will be demonstrated that in current models for the drawing of a sample, a constant number of particles is assumed or it is assumed that the number of particles in the sample is distributed according to a binomial distribution. These artificial conditions are too stringent because in the practice the total number of particles sampled will be variable even under similar sampling operating conditions. Therefore, it is difficult to sample either a constant number of particles or a binomially distributed number of particles. Instead, it is much easier to sample a constant mass or volume. In addition, sampling standards give a prescribed sample mass or volume. This suggests the need for modelling the sample drawing as a process leading to a constant sample mass or volume. Therefore, from a pragmatic point of view the following criterion is important:

Criterion 8

- The theory must be able to describe the sample drawing as a process leading to an approximately constant sample mass or volume.

For theories that are based on a mathematical model for the drawing of a sample on the level of (groups of) particles (*i.e.* theories that meet the second criterion), the modelling of sampling as a process leading to a fixed sample mass or volume has the additional advantage that the equation for the variance of the sample concentration will depend on the mass or volume sampled. Hence, for these theories, the first criterion is always met.

The above eight criteria are used in this work to develop a new sampling theory.

1.6 Outline of this thesis

Because the aim of sampling is to extract only a small part of the batch, it would not be useful to develop a potentially complicated sampling theory, which is valid for every sample size. Similarly, it is not very useful to develop a potentially complicated sampling theory, which is valid for samples containing only a few particles. Therefore, in this thesis, the batch-to-sample size ratio and the sample-to-particle size are defined, respectively. Numerous results throughout this thesis are valid in the limit of an infinite batch-to-sample size ratio and in the limit of an infinite sample-to-particle size ratio, and good approximations for large, but finite, values for the batch-to-sample size ratio and for large, but finite, values for the sample-to-particle size ratio.

In Chapter 2, existing sampling theories are reviewed by considering the eight criteria outlined in Paragraph 1.5. It will be demonstrated that none of the sampling theories meet all eight criteria. This all justifies the need for development of a new sampling theory, which meets all eight criteria.

In Chapter 3, a mathematical algorithm is presented to serve as a model for ideal sampling from a random arrangement of particles. The concepts of ideal sampling and random arrangement are clarified and the details of the algorithm are discussed. The validity of the proposed algorithm to describe real sampling processes is demonstrated on basis of computer simulations.

In Chapter 4, a mathematical calculation will be performed using the algorithm proposed in Chapter 3. As a final result, equations for the expected value and variance of the sample concentration are derived in the limit of an infinite batch-to-sample size ratio and an infinite sample-to-particle size ratio.

In Chapter 5, the estimation of the variance will be investigated using sample information only. It will be shown that the Horvitz-Thompson estimator can provide a general and unbiased estimate for the variance of the π-expanded estimator. Based on the results of Chapter 4, an expression for the Horvitz-Thompson estimator in the limit of an infinite batch-to-sample size ratio and an infinite sample-to-particle size ratio is derived.

For finite ratios, the sample concentration and the obtained variance estimator may be slightly biased. Chapter 6 describes how an indicative contour-plot can be obtained via simulations on a wide range of distinct batch compositions. The maximum absolute value of the relative bias can be obtained as a function of both the batch-to-sample size ratio and the sample-to particle size ratio.

In Chapter 7, for four samples, the estimators developed in this study and associated analytical uncertainties will be evaluated. Using the experimental results, the new theory is verified by comparing the level of contradiction of the theory with the levels of contradiction of the theories of Wilson and Gy. Also the normality of the distribution of the estimated sample concentrations will be investigated. It will be shown that the new theory is more internally consistent than the theories of Gy and Wilson and that it also provides a more normal estimator for the batch concentration.

In Chapter 8, an expression for the minimum sample mass, based on the properties of the particles in the batch, and an estimator for the minimum sample mass, based on the properties of the particles in the sample, will be derived. The applicability of the obtained estimator for the minimum sample mass will be investigated using simulations. It will be demonstrated that the estimator is applicable for sample-to-particle size ratio larger than 10. Knowledge of the particle masses in the sample can be used for estimation of the minimum sample mass, when knowledge of the distribution of particle concentrations is not available. Hence, it will be demonstrated that no *a priori* knowledge of the distribution of particle concentrations in the sample is required.

1.7 References

ASTM (1997) C1075-93 Standard Practices for Sampling Uranium-Ore Concentrate.

ASTM (2002) D1900-94 Standard Practice for Carbon Black—Sampling Bulk Shipments.

P.M. Gy (1979) *Sampling of Particulate Materials, Theory and Practice,* 1st edition, Elsevier, Amsterdam, 431 pp.

ISO (1993) International Vocabulary of Basic and Standard Terms in Metrology, Geneva, Switzerland (ISBN 92-67-10175-1).

ISO (2001) 11648-2 Statistical aspects of sampling from bulk materials -- Part 2: Sampling of particulate materials.

ISO (2003) 11648-1 Statistical aspects of sampling from bulk materials -- Part 1: General principles.

NEN 5742 (September 2001) Soil - Sampling of soil and sediments for the determination of metals, inorganic compounds, semi-volatile organic compounds and physico-chemical soil characteristics.

Chapter 2 Review of current sampling theories[3,4]

Current empirical and non-empirical sampling theories are reviewed. None of the empirical and non-empirical theories meet all eight criteria identified in Chapter 1. This justifies the development of a new sampling theory to meet all criteria.

2.1 Introduction

The mathematical model of the second criterion described in Paragraph 1.5 ("The theory must be based on a model of the drawing of a sample on the level of (groups of) particles") provides a scientific basis for the equation for the variance. The equation for the variance provided by a theory that is not based on an underlying mathematical model can only have an empirical basis. Therefore, theories that do not meet the second criterion are termed empirical theories, while theories that meet the second criterion are termed non-empirical theories. Current empirical and non-empirical sampling theories are separately reviewed in Paragraph 2.2 and 2.3 respectively using the eight criteria that were presented in Paragraph 1.5.

It is noted that sampling theories were previously reviewed by Smith and James (1981). In this review, the sampling theories are classified according to their "level of sophistication". Three situations are distinguished: (i) sampling theories applicable to binary mixtures, (ii) sampling theories applicable to real mixtures, and (iii) sampling theories applicable to mixtures that are potentially segregated. However, most of the characteristics with respect to the eight criteria, outlined in Paragraph 1.5, are ignored. Therefore, a new comparison of sampling theories, using the eight new criteria, is presented in this chapter.

[3] The main aspects of this chapter have been published as: B. Geelhoed and H.J. Glass (2004) Comparison of theories for the variance caused by the sampling of random mixtures of non-identical particles, *Geostandards and Geoanalytical Researchs* (IN PRESS).

[4] Because a wide range of sampling theories is reviewed, it was not feasible to adopt a consistent set of symbols with corresponding definitions. The set adopted in this chapter is not consistent, because: (i) different symbols are adopted to denote the same entity in different sampling theories, and (ii) for each symbol, the definition adopted may vary between different sampling theories. When required, the appropriate definition is given. The symbols and corresponding definitions, which can be extracted from the list of symbols given at the end of this thesis, do not necessarily apply for Chapter 2.

As discussed in Chapter 1, the measured value can contain an analysis error. This error is independent of the sampling error and will not be discussed further here.

2.2 Empirical sampling theories

Ingamells and Switzer (Ingamells and Switzer, 1973) proposed a single constant K_s that relates the mass of the sample to the relative standard deviation of the mass concentration in the sample a_{sample}. The original notation is:

$$R = \sqrt{K_s / w} \tag{2.1}$$

in which R is the relative standard deviation in percent and w is the sample mass. In this thesis, the symbol $M_{sample}(S)$ will be used instead of w (the symbol used in the original publication) to denote the mass of the sample S. Using partly the notation adopted in this thesis (see List of Symbols for a summary) and rewriting Equation 2.1, an expression is obtained for the relative variance $V_{rel}(a_{sample})$ and the variance $V(a_{sample})$:

$$V_{rel}(a_{sample}) \equiv \frac{V(a_{sample})}{a_{batch}^2} = 10^{-4} \frac{K_S}{M_{sample}(S)} \tag{2.2}$$

in which a_{batch} is the batch concentration and the symbol '≡' stands for 'is by definition equal to'. The factor 10^{-4} arises so as to express R as a percentage.

For a material for which no sampling constant is specified, Ingamells and Switzer proposed to estimate K_s by analyzing multiple samples, using the following equation:

$$K_s = \frac{10^4 w \sum_{i=1}^{N_{det}} (c_i - \bar{c})^2}{(N_{det} - 1)\bar{c}^2} \tag{2.3}$$

in which N_{det} is the total number of samples drawn to estimate K_s, c_i is the concentration as measured in the ith sample and $\bar{c}$ is the arithmetic mean of the N_{det} determinations. A requirement for application of the above equation is that the analytical error is negligible compared to the variations induced in the variables c_i by the sampling error. A second requirement is that a minimum number of samples is analyzed. Ingamells and Switzer recommend that at least 10 samples should be analyzed for a reliable estimate of K_s. However, no theoretical basis for this number is given.

The theory of Visman (Visman, 1969) is applicable to incremental sampling. In incremental sampling, N_{inc} increments of mass w are drawn to constitute a composite sample. Visman proposed the following equation for the variance of the mass concentration in the composite sample:

$$V(a_{sample}) = \frac{A}{N_{inc}w} + \frac{B}{N_{inc}} \tag{2.4}$$

where A is the "homogeneity" constant and B is the "segregation" constant. As the segregation constant B increases, the number N_{inc} of samples assumes greater importance. If there is no segregation, B is zero and it makes no difference into how many increments the total mass of samples is divided. Note that since the value w is given in the standard unit of mass (kg) A must also be given the in standard unit of mass (kg); B is dimensionless. When specified values for A and B are used, the variance of the sample concentration is calculated using Equation 2.4. When a material is sampled for which no values for A and B are specified, A and B must be estimated by analyzing two series of samples, one series of "small" samples and one series of "large" samples. Ingamells (Ingamells, 1974) proposes the following equations:

$$A = w_1 w_2 \left(s_1^2 - s_2^2\right) / \left(w_2 - w_1\right) \tag{2.5}$$

$$B = s_2^2 - A/w_2 \tag{2.6}$$

where s_1^2 is the observed small-sample variance, s_2^2 is the observed large-sample variance, w_1 is the small-sample mass and w_2 the large-sample mass (i.e. $w_1 < w_2$).

The theory of Rasemann and Herbst (Rasemann and Herbst, 2000) is also not based on a mathematical algorithm used as a model for the drawing of a sample on the level of (groups of) particles. In this theory, the parameters that describe the statistical distribution of the number of particles in the sample, N_{sample}, cannot be derived mathematically. Therefore, it is assumed that the expected value of N_{sample}, $E(N_{sample})$, and variance of N_{sample}, $V(N_{sample})$, are determined empirically. For the derivation of an equation that relates the identity of the particles in the batch to the variance of the sample concentration, it is assumed that the identity of each particle in the sample is chosen at random from the entire collection of particles in the batch. Without reference to a derivation, Rasemann and Herbst gave the following equation for the sample-to-sample variance:

$$V(a_{sample}) = V\left(\frac{a_i m_i}{M_{sample}}\right) E(N_{sample}) + V(N_{sample}) E^2\left(\frac{a_i m_i}{M_{sample}}\right) \tag{2.7}$$

in which $V(a_i m_i/M_{sample})$ and $E(a_i m_i/M_{sample})$ are respectively the variance and expected value of the product of the concentration in the i^{th} particle in the sample and the particle mass divided by the sample mass. Because the identities of the particles in the sample were chosen independently, these quantities are equal for all i between 1 and N_{sample}. When the sample mass is constant, the following identities hold:

$$V\left(\frac{a_i m_i}{M_{sample}}\right)=\frac{V(a_i m_i)}{M^2_{sample}(S)} \tag{2.8}$$

$$E^2\left(\frac{a_i m_i}{M_{sample}}\right)=\frac{E^2(a_i m_i)}{M^2_{sample}(S)} \tag{2.9}$$

where $V(a_i m_i)$ and $E(a_i m_i)$ are respectively the variance and expected value of the product of the concentration in a particle and the particle mass. The equation for the variance becomes:

$$V(a_{sample})=\frac{1}{M^2_{sample}(S)}\left(V(a_i m_i)E(N_{sample})+V(N_{sample})E^2(a_i m_i)\right) \tag{2.10}$$

Because the identity of the i^{th} particle in the sample is chosen at random from the entire collection of particles in the batch, $V(a_i m_i)$ and $E(a_i m_i)$ can be related to the identities of the particles in the batch:

$$E(a_i m_i)=\frac{1}{N_{batch}}\sum_{j=1}^{N_{batch}} a_j m_j \tag{2.11}$$

$$V(a_i m_i)=\frac{1}{N_{batch}}\sum_{j=1}^{N_{batch}} \left(a_j m_j - E(a_i m_i)\right)^2 \tag{2.12}$$

where N_{batch}, a_j and m_j are the total number of particles in the batch, the mass concentration in and mass of the j^{th} particle of the batch respectively. The index j ranges from 1 to N_{batch}. Hence, the sample-to-sample variance is modelled on the level of the individual particles of the batch.

The characteristics of the theories of Ingamell/Switzer, Visman, and Rasemann/Herbst can now be projected against the eight criteria:

Criterion 1

- Met, the theories relate the sample mass $M_{sample}(S)$ (denoted as w in the theory of Ingamells/Switzer and $N_{inc}w$ in the theory of Visman) to the variance of the sample concentration, see Equations 2.2, 2.4, and 2.10.

Criterion 2

- Not met, the theories are not based on a model for the drawing of a sample.

Criterion 3

- Met, there are no restrictions to the theories that prohibit the application to batches that contain multiple distinct types of particles.

Criterion 4

- Met in the theories of Ingamells/Switzer and Visman, because the parameters K_s, A and B can be estimated using multiple samples (see Equations 2.3, 2.5 and 2.6).
- Not met in the theory of Rasemann/Herbst. This theory provides no method to estimate the parameters $E(a_i m_i)$ and $V(a_i m_i)$ using the measured sample concentrations of one or more samples. This applies even if $E(N_{sample})$ and $V(N_{sample})$ have been determined empirically.

Criterion 5

- Not met in the theories of Ingamells/Switzer, and Visman. These theories do not provide a method to estimate K_s, A, and B using knowledge of the properties of the particles in the batch.
- Met in the theory of Rasemann/Herbst, where values of the parameters $E(a_i m_i)$ and $V(a_i m_i)$ can be calculated using the properties of the particles in the batch (see Equation 2.11 and 2.12). If it is assumed that $E(N_{sample})$ and $V(N_{sample})$ were determined empirically, the value of the variance can be calculated for an arbitrary value of the sample mass.

Criterion 6

- Not met, no method is specified to estimate the parameters K_s, A, B, $E(a_i m_i)$, $V(a_i m_i)$, $E(N_{sample})$ and $V(N_{sample})$ using knowledge of the properties of the particles in the sample.

Criterion 7

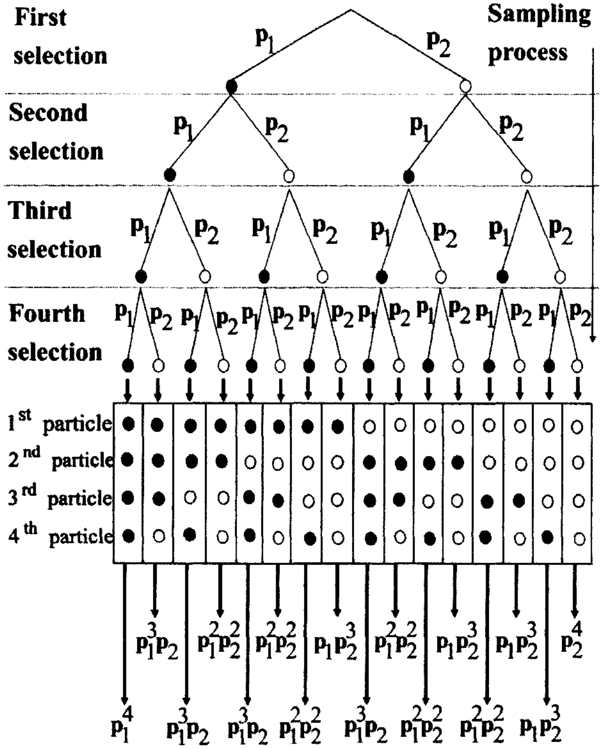
- Not met, the theories apply only for mass concentrations.

Criterion 8

- Met in the theories of Ingamells/Switzer and Visman. The theories describe sampling as a process leading to a constant sample mass, denoted as w in the theory of Ingamells/Switzer and $N_{inc}w$ in the theory of Visman.
- Met in the theory of Rasemann/Herbst. The theory is able to describe sampling as a process leading to a constant sample mass.

2.3 Non-empirical sampling theories

Hassialis (Taggart, 1945) proposed a binomial sampling scheme. With this model, it is assumed that the particles belong to only two classes: particles with and particles without the property of interest. It is further assumed that a sample contains a fixed number of N particles that were selected during N selections. With each selection, the probabilities of selecting a particle belonging to either the first or the second class do not depend on any of the previously selected particles. A graphical illustration of the model of the sample drawing process is given in Figure 2.1.


First selection
Sampling process
Second selection
Third selection
Fourth selection
p_1
p_2
$p_1^3p_2$ $p_1^2p_2^2$ $p_1^2p_2^2$ $p_1p_2^3$ $p_1^2p_2^2$ $p_1p_2^3$ $p_1p_2^3$ p_2^4
p_1^4 $p_1^3p_2$ $p_1^3p_2$ $p_1^2p_2^2$ $p_1^3p_2$ $p_1^2p_2^2$ $p_1^2p_2^2$ $p_1p_2^3$

1st particle	●	●	●	●	●	●	●	●	○	○	○	○	○	○	○	○
2nd particle	●	●	●	●	○	○	○	○	●	●	●	●	○	○	○	○
3rd particle	●	●	○	○	●	●	○	○	●	●	○	○	●	●	○	○
4th particle	●	○	●	○	●	○	●	○	●	○	●	○	●	○	●	○

Figure 2.1. Sampling process according to the model of Hassialis. In the model of Hassialis, only two classes of particles are distinguished. The probabilities of drawing a particle of type 1 and 2 are denoted as p_1 *and* p_2 *respectively. In this example, the number of particles sampled is 4. At the end of each branch, the sample composition and probability are given.*

Using the binomial distribution, Hassialis derived the following formula for the variance, σ_p^2, of the numerical fraction of particles with the property of interest[5]:

[5] The numerical fraction of particles with the property of interest in the sample is defined as the ratio of the number of particles in the sample with the property of interest to the total number of particles in the sample. Other fractions are the mass fraction or voluminous fraction of particles with the property of interest in the sample. These are defined as mass or volume of the particles with the property in the sample respectively, divided by the total mass or volume of the sample respectively.

$$\sigma_p^2 = \frac{p_1 p_2}{N} \tag{2.13}$$

where p_1 and p_2 are the probability of selecting a particle with and without the property of interest respectively. Hassialis assumed that p_1 and p_2 are equal to the numerical fraction of particles with and without the property of interest respectively, i.e. $p_1 = p_1' \equiv N_{1,\text{batch}} / N_{\text{batch}}$ and $p_2 = p_2' \equiv N_{2,\text{batch}} / N_{\text{batch}}$ where:

p_1' = the numerical fraction of particles in the batch with the property of interest,

p_2' = the numerical fraction of particles in the batch without the property of interest,

$N_{1,\text{batch}}$= the number of particles in the batch with the property of interest and

$N_{2,\text{batch}}$= the number of particles in the batch without the property of interest.

Wilson (Wilson, 1964) generalized the model of Hassialis by using a multinomial distribution.

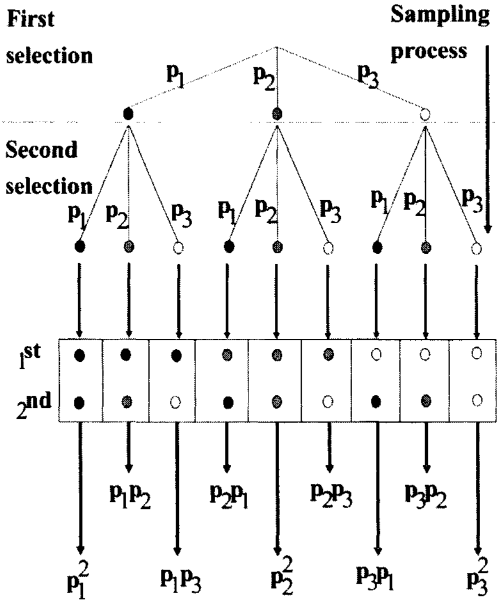


Figure 2.2. Sampling process according to the model of Wilson. In this example, the number of classes of particles is 3, but the theory of Wilson can handle an arbitrary number of distinct classes of particles. The probabilities of drawing a particle of type 1, 2 and 3 are denoted as p_1, p_2 *and* p_3 *respectively. In this example, the number of particles sampled is 2. At the end of each branch, the sample composition and probability are given.*

In this model, the samples always contain a fixed number of N particles. It is assumed that the particles in the batch are of uniform volume and can be classified into m distinct classes. In a class, the concentration of the property of interest in a particle and the particle mass is constant. Like Hassialis, Wilson assumes that the probability p_k for any k between 1 and m is equal to the numerical fraction p'_k of particles belonging to the k^{th} class in the batch, i.e. $p_k = p'_k \equiv N_{k,batch}/N_{batch}$ in which $N_{k,batch}$ is the number of particles in the batch belonging to the k^{th} class. An illustration of the sampling process proposed by Wilson is given in Figure 2.2.

The final sampling equation, derived by Wilson, using the notation of Wilson, is:

$$V(a_{sample}) = \frac{1}{2N}\sum_{i=1}^{m}\sum_{j=1}^{m}(\Delta t_i d_i - \Delta t_j d_j)^2\left(\frac{W_i W_j}{d_i d_j}\right) \tag{2.14}$$

where

$V(a_{sample})=$ the variance of the mass concentration in the sample,

N= the number of particles in the sample,

m= the number of particle types in the batch,

$d_k=$ the density of a particle of type k for all k between 1 and m,

$W_k=$ the mass fraction of particles of type k in the batch for all k between 1 and m,

and

$\Delta t_k=$ t_k-t for all k between 1 and m,

where

$t_k=$ the mass concentration in a particle of type k for all k between 1 and m, and

t= the mass concentration in the batch.

A different notation convention will be used in this thesis. The equation of Wilson can be written in this notation, if the following substitutions are made:

m=T

$t_k=a_k$

$t=a_{batch}$

$d_k=m_k/v$

$$W_k = \frac{N_{k,batch} m_k}{M_{batch}} = \frac{m_k p'_k}{\overline{m}}$$

in which

$T=$ the number of different particle types in the batch,
k can represent any integer between 1 and T,
$a_k=$ the mass concentration in a particle of type k,
$a_{batch}=$ the mass concentration in the batch,
$m_k=$ the mass of a particle of type k,
$v=$ the volume of a particle[6] (assumed constant by Wilson),
$M_{batch}=$ the mass of the batch,
$p'_k=$ $N_{k,batch}/N_{batch}$ and
$\bar{m}=$ the average particle mass in the batch M_{batch}/N_{batch}.

When the above variables are substituted and the equation is rewritten, the volume v cancels out of the equation:

$$V\left(a_{sample}\right)=\frac{1}{2N}\sum_{i=1}^{T}\sum_{j=1}^{T}\left(\left(a_i-a_{batch}\right)m_i-\left(a_j-a_{batch}\right)m_j\right)^2\left(\frac{p'_i\,p'_j}{\bar{m}^2}\right) \tag{2.15}$$

The square can be expanded and the terms can be rearranged:

$$V\left(a_{sample}\right)=\frac{\sum_{i=1}^{T}p'_i\left(a_i-a_{batch}\right)^2 m_i^2\sum_{j=1}^{T}p'_j+\left(\sum_{i=1}^{T}p'_i\left(a_i-a_{batch}\right)m_i\right)^2}{\bar{m}^2 N} \tag{2.16}$$

Note that because $p'_i=N_{i,batch}/N_{batch}$ for all i between 1 and T, it follows that

$$\sum_{j=1}^{T}p'_j=1 \qquad \text{and} \qquad \sum_{i=1}^{T}p'_i\,(a_i-a_{batch})m_i=0$$

Hence Equation 2.16 is simplified to:

$$V\left(a_{sample}\right)=\frac{1}{\bar{m}^2 N}\sum_{i=1}^{T}p'_i\,m_i^2(a_i-a_{batch})^2 \tag{2.17}$$

The above equation for the variance is easier to evaluate than the equation originally proposed by Wilson because only a single summation symbol is required.

Gy (Gy, 1979) proposed to model the sample drawing by repeated and independent Bernoulli experiments for every particle in the batch. An illustration of this process is given in Figure 2.3.

[6] Note that Wilson introduced the symbol v (see Wilson, 1964), but this symbol is not part of the adopted notation used throughout Chapter 3 to 9 of this thesis.

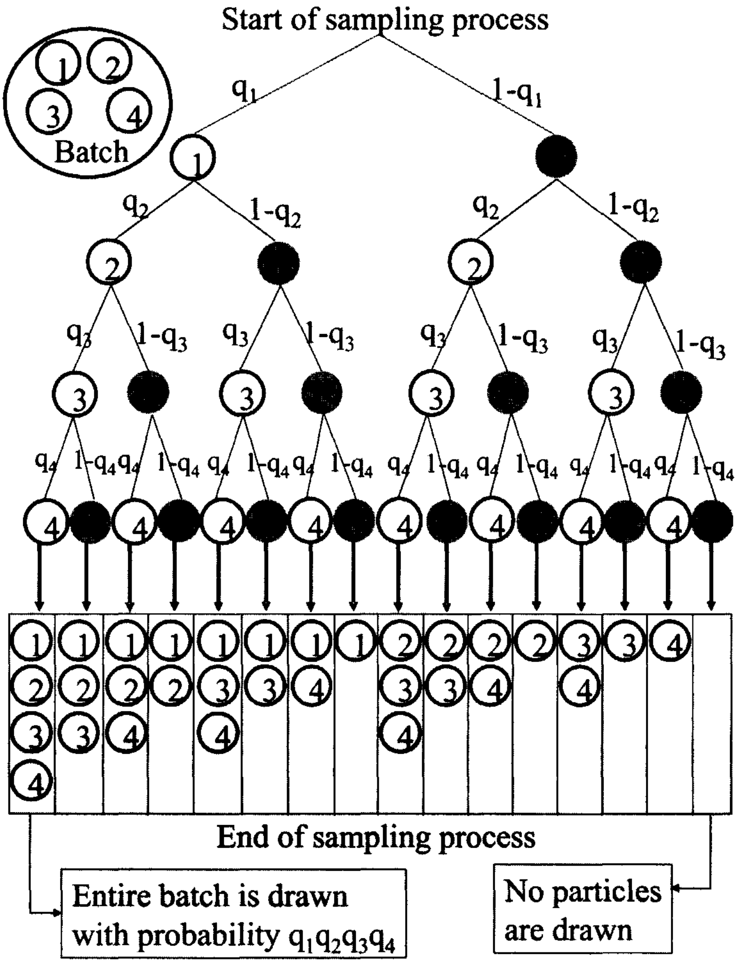


Figure 2.3. *Sampling process according to the model of Gy. In this example, the batch contains four particles. The probabilities of drawing particles 1,2,3 and 4 are indicated by* q_1, q_2, q_3 *and* q_4. *It can be seen that the number of selected particles varies between 4 (the sample on the left) and zero (the 'sample' on the right). At the end of each branch the composition of the sample is given.*

During the kth Bernoulli experiment, the kth particle in the batch has a probability q_k of being selected, while all the other particles have a zero probability of being selected. In standard statistical theory, this process is denoted as Poisson sampling (Särndal *et al*, 1992).

The sampling design is rather artificial because the number of particles in the sample may potentially be any integer between zero and the total number of particles in the batch, N_{batch}. In the special case when all q_i are constant, i.e. $q_i=q$ for all i between 1 and N_{batch}, the total number of particles in the sample is binomially distributed. In Gy's theory, both the mass and volume sampled vary between zero and the mass and volume, respectively, of the entire batch. This is in contradiction with the practice of sampling, since commonly the sampled mass or volume is approximately constant.

In the following, the theoretical consequences of Gy's model are evaluated. Using the indicator I_i, which is one when the i^{th} particle in the batch is selected and zero otherwise, the sample concentration is defined by:

$$a_{sample} = \frac{A_{sample}}{M_{sample}} = \sum_{i=1}^{N_{batch}} I_i a_i m_i \Bigg/ \sum_{i=1}^{N_{batch}} I_i m_i \tag{2.18}$$

where A_{sample} is the total mass of the component of interest in the sample, M_{sample} is the total sample mass, and a_i and m_i are respectively the mass concentration in and the mass of the i^{th} particle in the batch. From the above equation, it follows that the sample concentration will fluctuate both due to statistical fluctuations in the denominator (the sample mass) and in the numerator (the total mass of the component of interest in the sample). Hence, Gy's model does not allow an exact calculation of the expected value and variance of the sample concentration. Therefore, a Taylor linearization is applied (Gy, 1979). In the following, a short and comprehensive derivation will be given, which is basically equivalent to Gy's derivation.

The Taylor linearization implies that a_{sample} is written as a linear function of A_{sample} and M_{sample} plus a rest term:

$$a_{sample} = \frac{A_{sample}}{M_{sample}} = \frac{E(A_{sample})}{E(M_{sample})} + \frac{A_{sample} - E(A_{sample})}{E(M_{sample})} - \frac{E(A_{sample})}{E^2(M_{sample})}\left(M_{sample} - E(M_{sample})\right) + \text{Rest} \tag{2.19}$$

in which $E(A_{sample})$ and $E(M_{sample})$ are the expected value of A_{sample} and M_{sample} respectively. The term Rest is equal to the difference between the left-hand side and the linear terms:

$$\text{Rest} = \frac{A_{sample}}{M_{sample}} - \frac{E(A_{sample})}{E(M_{sample})} - \frac{A_{sample} - E(A_{sample})}{E(M_{sample})} + \frac{E(A_{sample})}{E^2(M_{sample})}\left(M_{sample} - E(M_{sample})\right) \tag{2.20}$$

Rest can be interpreted as the error made during first-order approximation of a_{sample} by linear terms only. The above expression for Rest can be rewritten:

$$\text{Rest} = \left(\frac{A_{sample}}{M_{sample}} - \frac{E(A_{sample})}{E(M_{sample})}\right)\frac{E(M_{sample}) - M_{sample}}{E(M_{sample})} \tag{2.21}$$

From the above equation, it follows that the error made in the linear approximation is proportional to the relative deviation of the sample mass from its expected value, $(E(M_{sample})-M_{sample})/E(M_{sample})$. Hence, the first-order approximation effectively neglects samples in which the difference between M_{sample} and $E(M_{sample})$ is large. On the other hand, if the relative deviation in the sample mass is smaller than one, the error in the first-order approximation is smaller than the error made if A_{sample}/M_{sample} would be replaced simply by $E(A_{sample})/E(M_{sample})$, the zero-th order approximation. The latter observation suggests that the first-order approximation should be preferred instead of the zero-th order approximation.

To the first-order approximation, the expected value of a_{sample} is simply:

$$E(a_{sample}) = \frac{E(A_{sample})}{E(M_{sample})} \tag{2.22}$$

Using Equation 2.19 gives the following equation for the variance of a_{sample}:

$$V(a_{sample}) = V\left(\frac{E(A_{sample})}{E(M_{sample})} + \frac{A_{sample} - E(A_{sample})}{E(M_{sample})} - \frac{E(A_{sample})}{E^2(M_{sample})}(M_{sample} - E(M_{sample})) + \text{Rest}\right) \tag{2.23}$$

where the right-hand side denotes the variance of $E(A_{sample})/E(M_{sample})+(A_{sample}-E(A_{sample}))/E(M_{sample})-(M_{sample}-E(M_{sample}))E(A_{sample})/E^2(M_{sample})$+Rest. Neglecting Rest and using the fact that constant terms do not contribute to the variance yields:

$$V(a_{sample}) = V\left(\frac{A_{sample}}{E(M_{sample})} - \frac{E(A_{sample})M_{sample}}{E^2(M_{sample})}\right) \tag{2.24}$$

where the right-hand side denotes the variance of $A_{sample}/E(M_{sample})-E(A_{sample})M_{sample}/E^2(M_{sample})$. The equation can be rewritten:

$$V(a_{sample}) = \frac{E^2(A_{sample})}{E^2(M_{sample})} V\left(\frac{A_{sample}}{E(A_{sample})} - \frac{M_{sample}}{E(M_{sample})}\right) \tag{2.25}$$

where $V(A_{sample}/E(A_{sample})-M_{sample}/E(M_{sample}))$ is the variance of $A_{sample}/E(A_{sample})-M_{sample}/E(M_{sample})$. In statistical theory, a sample total X_{sample} is a quantity that can be written as a summation over the particles. Hence, when x_i is the value of the i^{th} particle in the batch

$$X_{sample} = \sum_{i=1}^{N_{batch}} I_i x_i \tag{2.26}$$

Under Poisson sampling the following identities hold for the expected value and variance of X_{sample} (Särndal *et al*, 1992):

$$E\left(X_{\text{sample}}\right)=\sum_{i=1}^{N_{\text{batch}}} q_i\, x_i \tag{2.27}$$

$$V\left(X_{\text{sample}}\right)=\sum_{i=1}^{N_{\text{batch}}} q_i\left(1-q_i\right)x_i^2 \tag{2.28}$$

The above relations can be applied to Equations 2.22 and 2.25 by recognizing that A_{sample}, M_{sample}, and $A_{\text{sample}}/E(A_{\text{sample}})-M_{\text{sample}}/E(M_{\text{sample}})$ are the sample totals of $a_i m_i$, m_i and $a_i m_i/E(A_{\text{sample}})-m_i/E(M_{\text{sample}})$ respectively. This results in the following first-order approximations for the expected value and variance:

$$E\left(a_{\text{sample}}\right)=\sum_{i=1}^{N_{\text{batch}}} q_i\, a_i\, m_i \Bigg/ \sum_{i=1}^{N_{\text{batch}}} q_i\, m_i \tag{2.29}$$

$$V\left(a_{\text{sample}}\right)=\left(\sum_{i=1}^{N_{\text{batch}}} q_i\, m_i\right)^{-2}\sum_{i=1}^{N_{\text{batch}}} q_i\left(1-q_i\right)m_i^2\left(a_i-\sum_{j=1}^{N_{\text{batch}}} q_j a_j m_j \Bigg/ \sum_{j=1}^{N_{\text{batch}}} q_j m_j\right)^2 \tag{2.30}$$

Hence, when all q_i are equal, $q_i \equiv q$, the mass concentration in the sample is in the first-order approximation an unbiased estimator for the batch concentration, a_{batch}. This means that $E(a_{\text{sample}})=a_{\text{batch}}$, see also Paragraph 4.2. In this case, the variance of the sample concentration becomes:

$$V\left(a_{\text{sample}}\right)=\frac{(1-q)}{qM_{\text{batch}}^2}\sum_{i=1}^{N_{\text{batch}}} m_i^2\left(a_i-a_{\text{batch}}\right)^2 \tag{2.31}$$

where M_{batch} is the mass of the batch. The variance calculated with the above equation may differ from the actual variance as a result of the artificial sample drawing model assumed by Gy, leading to a binomial distribution of the total number of particles in the sample. However, it will be shown that Equation 2.31 is similar to the equation for the variance that is derived in Chapter 4 for the new sampling theory presented in this thesis. A possible explanation for this fact is that the occurrence of sample masses that deviate largely from the expected value is neglected due to the underlying first-order approximation. Because the equation for the variance derived in Gy's theory is similar to the equation for the variance in the new theory, both equations could provide similar results for specific applications. In the following, two distinct applications will be reviewed.

A derivation of Gy's basic equation (Equation 2.31) was obtained by analysing a mixture of distinct types of materials. Several assumptions were required. Firstly, it was assumed that the particles in the batch can be classified according to volume and type of material and that the concentration in a particle and density of a particle do not vary between particles of a given material type. Secondly, it was assumed that the size distribution in the batch of particles belonging to distinct material types is identical. Thirdly, it was assumed that the volume of each particle in the batch is given by a constant factor f, multiplied by the cube of the particle diameter. Using these assumptions about the composition of the sampled batch and the particle size distribution, Gy obtained the following equation for the factor $\frac{1}{M_{batch}}\sum_{i=1}^{N_{batch}} m_i^2(a_i - a_{batch})^2$ in Equation 2.31:

$$\frac{1}{M_{batch}}\sum_{i=1}^{N_{batch}} m_i^2\left(a_i - a_{batch}\right)^2 = d_{max}^3 fg\ell c \qquad \textbf{(2.32)}$$

where

d_{max}= the typical maximum particle diameter (determined by sieving),

f= the shape factor,

g= the size range factor,

ℓ= the liberation factor, and

c= the mineralogical composition factor

The precise relationship between the above introduced parameters and the masses m_i and concentrations a_i of the particles of the batch can be found in Gy (1979).

Von Blottnitz and Hoberg (von Blottnitz and Hoberg, 1998) arrived at another derivative of Gy's basic equation (Equation 2.31) based on analysing a mixture of plastics. It is assumed that the batch consists of two particle species. The concentrations in the particles of the first and second species are one and zero respectively. It is also assumed that both species have a Rosin-Rammler-Sperling-Bennet (RRSB) particle mass distribution, which is characterized by two parameters. The fraction of particles with mass smaller than m (which can represent any positive number) is parameterized as $1 - e^{-(m/m')^n}$, where m' and n are the two parameters that define the RRSB distribution. Von Blottnitz and Hoberg obtained the following equation for the factor $(1/M_{batch})\sum_{i=1}^{N_{batch}} m_i^2\left(a_i - a_{batch}\right)^2$ in Equation 2.31:

$$\frac{\sum_{i=1}^{N_{batch}} m_i^2\left(a_i - a_{batch}\right)^2}{M_{batch}} = a_{batch}\left(1 - a_{batch}\right)\left[a_{batch}M_A Q(n_A) + \left(1 - a_{batch}\right)M_B Q(n_B)\right] \qquad \textbf{(2.33)}$$

where M_A, M_B, $Q(n_A)$ and $Q(n_B)$ are parameters that define the specific form of the

RRSB particle mass distributions. For a detailed derivation and explanation of these parameters, the reader is referred to Von Blottnitz and Hoberg (1998).

The characteristics of the theories Hassialis, Wilson, Gy and derivatives of Gy's theory can now be projected against the eight criteria:

Criterion 1

- Not met in the theories of Hassialis and Wilson. Equation 2.17 (for T=2 or the general case) does not relate the variance of the sample concentration to the sampled mass or volume. An attempt to overcome this problem by replacing $\overline{m}N$ by the sample mass $M_{sample}(S)$ fails, because generally $\overline{m}N \neq M_{sample}(S)$ when the particles in the batch have a variable mass.
- Not met in the theory of Gy and derivatives of Gy's theory. Equation 2.31 does not depend on the volume or mass sampled. An attempt to overcome this problem by replacing qM_{batch} by $M_{sample}(S)$ fails, because, in Gy's theory, the mass sampled varies between zero and the mass of the entire batch and is therefore generally not equal to qM_{batch}.

Criterion 2

- Met. The theories are based on a model for the drawing of a sample.

Criterion 3

- Not met in the theory of Hassialis.
- Met in the theories of Wilson, Gy and derivatives of Gy's theory.

Criterion 4

- Met in the theories of Hassialis and Wilson. If $\overline{m}N$ is replaced by the sample mass $M_{sample}(S)$ (which is not an accurate substitution), the equation for the variance, Equation 2.17, can be written as $V(a_{sample})=C/M_{sample}(S)$, where C is a material-dependent parameter. In this case, the value of C can be estimated using multiple samples, similar to the manner in which K_s is estimated in the theory of Ingamells/Switzer.
- Met in the theory of Gy and derivatives of this theory. If qM_{batch} is replaced by the sample mass $M_{sample}(S)$ (which is not an accurate substitution), the equation for the variance, Equation 2.31, can be written as $V(a_{sample})=C/M_{sample}(S)$, where C is a material dependent parameter. In this case, C could be estimated, using multiple samples, similar to the way K_s is estimated in the theory of Ingamells/Switzer.

Criterion 5

- Met in the theories of Hassialis and Wilson. The equation for the variance of the sample concentration, Equation 2.17, depends on the variables a_{batch}, a_i, m_i, p_i and $\overline{m}$. These variables depend on all the particles in the batch.
- Met in the theory of Gy and derivatives of this theory. The basic equation (Equation 2.31) requires batch information (the masses of and concentrations in the particles of the batch). Also Equations 2.32 and 2.33, require batch information: for

evaluation of the right-hand side of Equation 2.32, c_{batch}, d_{max}, f, g and ℓ are required. These parameters depend on the particle masses, particle volumes and concentrations in the particles of the batch. For evaluation of the right-hand side of Equation 2.33, the parameters a_{batch}, M_A, M_B, $Q(n_A)$ and $Q(n_B)$ are required. These parameters depend on the particle masses and concentrations in the particles of the batch.

Criterion 6

- Not met. No equations are given for which knowledge of the properties of the particles in a sample is required.

Criterion 7

- Not met. The theories are only applicable to mass concentrations and not to volume concentrations.

Criterion 8

- Not met. The theories of Hassialis and Wilson model the sample drawing as a process leading to a constant number of particles. The theory of Gy (and derivatives of this theory) models the sample drawing as a process leading to a binomially distributed number of particles.

2.4 Results

Current empirical and non-empirical sampling theories were reviewed by considering eight criteria. The first criterion is the ability to provide an accurate relation between the mass or volume sampled and the variance of the sample concentration. The empirical theories of Ingamells/Switzer, Visman and Rasemann/Herbst meet this criterion.

The basis for the derivation of the equation for the variance can be provided by an underlying mathematical model for the drawing of a sample. Therefore, the second criterion is that the theory is based on a model for the drawing of a sample on the level of (groups of) particles. This criterion is met in the non-empirical theories of Hassialis, Wilson and Gy.

Because real batches may contain a wide range of different types of particles, the third criterion is that the theory is applicable to sampling from batches containing any number of distinct types of particles. Only the theory of Hassialis does not meet this criterion.

The fourth criterion is that the theory allows determination of the parameters of the size-variance equation, using the measured sample concentrations of one or more samples of a given size. This criterion is met in the empirical theories of Ingamells/Switzer and Visman and in the non-empirical theories of Hassialis, Wilson and Gy.

Knowledge of the distribution of the particle masses and concentrations of the particles in the batch may be used to calculate the parameters of the size-variance equation if the fifth criterion is met. The fifth criterion is that the theory allows determination of the parameters of the size-variance equation, using prior knowledge of the properties of the particles in the batch. This criterion is met by the empirical theory of Rasemann/Herbst and by the non-empirical theories of Hassialis, Wilson and Gy (see Equations 2.10, 2.17 and 2.31).

When there is no knowledge of the properties of the particles in the batch, analysis of the particles in the sample may provide knowledge of the properties of the particle in the sample. Therefore, the sixth criterion is that the theory allows determination of the parameters of the size-variance equation, using posterior knowledge of the properties of the particles in the sample. Because none of the theories reviewed in this chapter uses this crucial sample information, none of these theories meet this criterion.

The seventh criterion is that the theory applies to mass and volume concentrations. This criterion is not met by any of the reviewed theories, because these theories are only applicable for mass concentrations.

Finally, the eighth criterion is that the theory is able to describe sampling as a process leading to a constant sample mass or volume. This criterion is met by the empirical theories of Ingamells/Switzer, Visman and Rasemann/Herbst.

The results of this review are summarized in Table 2.1.

	Criterion								Total
	1	2	3	4	5	6	7	8	score
Ingamells Switzer	1	0	1	1	0	0	0	1	4
Visman	1	0	1	1	0	0	0	1	4
Rasemann Herbst	1	0	1	0	1	0	0	1	4
Hassialis	0	1	0	1	1	0	0	0	3
Wilson	0	1	1	1	1	0	0	0	4
Gy And derivatives	0	1	1	1	1	0	0	0	4

Table 2.1. *Summary meeting the criteria. A '1' indicates meeting a criterion and a '0' indicates that the criterion is not met. The last column gives the total score of a theory, i.e. the total number of criteria that the theory meets.*

2.5 Conclusions

In Table 2.1, it can be seen that none of the theories meets all criteria. Therefore, further research to develop a theory that meets all eight criteria is required. In order to be able to meet the first, second and third criterion, the new theory must provide the following features:

- The sample drawing must be modelled at the level of (groups of) particles, so that it becomes possible to meet the second criterion.
- Using this model, an equation for the size-variance relationship must be derived, so that the first criterion is met.
- The sample-drawing model must be applicable to batches containing any number of distinct types of particles, so that it becomes possible to meet the third criterion.

Using the new model, the following results must be derived in order to meet the fourth, fifth and sixth criteria:

- The theory must provide an equation for the parameters of the size-variance equation, using the measured sample concentrations of one or more samples of a given size, so that the fourth criterion is met.
- The theory must provide an equation for the parameters of the size-variance equation, using prior knowledge of the properties of the particles in the batch, so that the fifth criterion is met.
- The theory must provide a sample estimator for the variance, whose evaluation requires only the particle masses, particle volumes and concentrations in the particles in the sample, so that the sixth criterion can be met.

Finally, the new theory must provide the following features:

- The theory must be applicable to both mass and volume concentrations, so that the seventh criterion is met.
- The theory must be able to describe sampling as a process that leads to samples of fixed mass or volume, so that it becomes possible to meet the eighth criterion.

In Chapter 3, the underlying sample-drawing model for a new theory is presented.

2.6 References

P.M. Gy (1979) *Sampling of Particulate Materials, Theory and Practice*, 1st edition, Elsevier, Amsterdam, 431 pp.

C.O. Ingamells (1974) New approaches to geochemical analysis and sampling, *Talanta*, **21**, p. 141-155.

C.O. Ingamells and P. Switzer (1973) A proposed sampling constant for use in geochemical analysis, *Talanta*, **20**, p. 547-568.

B. Kratochvil and J.K. Taylor (1981) Sampling for chemical analysis, *Analytical Chemistry*, **53**, p. 924-936.

C. Särndal, B. Swensson and J. Wretman (1992), *Model Assisted Survey Sampling*, Springer, New York, 694 pp.

W. Rasemann and M. Herbst (2000) Ein Probenahmemodell für heterogene Stoffsysteme, *Erzmetall*, **53**, p. 228-237.

R. Smith and G. V. James (1981) *Sampling of bulk materials*, Royal Society of Chemistry, London, 191 pp.

A. F. Taggart (1945) *Handbook of mineral dressing: ores and industrial minerals*, John Wiley, New York, 1915 pp.

J. Visman (1969) *Materials research and standards*, **9**, No. 11, 8 pp.

H. von Blottnitz, H. Hoberg (1998) A new method for the calculation of the relation between sample size and sampling error in the sampling of wastes and recyclables, *Erzmetall*, **51**, p. 595-603.

A.D. Wilson (1964) The sampling of rock powders for chemical analysis, *The analyst*, **89**, p. 18-30.

Chapter 3 Size-based multinomial selections: a model for sampling[7,8]

A mathematical algorithm is presented to serve as a model for ideal sampling from a random arrangement of particles. The concept of ideal sampling is defined and the details of the algorithm are discussed. It is shown that non-ideal sampling and biased sampling are different phenomena, whereas non-ideal sampling can act as a source of biased sampling. The boundary value of the sample size can, with limited effects to the accuracy, be estimated using the sampled mass. Simulations demonstrate the validity of this process.

3.1 Introduction

The second criterion for a sampling theory ("The theory must be based on a model of the drawing of a sample on the level of (groups of) particles.") is addressed in this chapter. In order to develop a realistic model, the physical conditions during sampling have to be taken into account.

Several conditions influencing sampling can lead to a biased selection of particles with respect to size, shape and/or other properties. This biased selection may subsequently result in a biased sample estimate of a batch value or concentration (see Paragraph 4.2 for a mathematical definition of the bias of a sample estimator). Sample drawing is strongly influenced by the properties of the material to be sampled and by the manner of sampling. Material properties that influence sampling are the masses and shapes of the particles, the friction between particles and cohesive forces. Sampling is also influenced by geometry of the sample drawing equipment and operating conditions, for instance the opening time of a shutter when sampling from a stream of moving particles, or the speed with or the depth to which a sampling lance is inserted in a static batch.

During sampling of an arbitrary material, the operating conditions have to be chosen well so that a biased selection of particles is minimized. A discussion on the

[7] The main aspects of this chapter have been published as: B. Geelhoed and H.J. Glass (2001) A new model for sampling of particulate materials and determination of the minimum sample size. *Geostandards Newsletter – The Journal of Geostandards and Geoanalysis*, **25**, p. 325-332.

[8] The definitions of symbols that were adopted in Chapter 2 are voided for Chapter 3 to 8. From now on, a consistent set of symbols and corresponding definitions, to be extracted from the List of Symbols given at the end of this thesis, is used.

minimization of a biased selection of particles can be found in Robinson and Cleary (1999). However, an investigation of this effect is not necessary here because the focus is on ideal sampling, a concept that will be introduced in Paragraph 3.3. To demonstrate the absence of a biased selection of particles during ideal sampling, the sampling error can be split into two parts: the error due to non-ideal sampling, e_{nis}, and the error due to the distribution of non-identical particles, e_{nip}. The precise definition of these errors will be given in Paragraph 3.3. However, it is important to note that only the choice of the location inside the sampled batch and the distribution of the particles in the batch have an influence on e_{nip}. The operating conditions, which may lead to a biased selection of particles, influence only the value of e_{nis}. Because e_{nip} is the sampling error of interest, the operating conditions leading to a biased selection of particles are not further discussed here. Instead, only the distribution of the particles in the batch is of interest. It will be demonstrated that when the particles are in a random arrangement, the size-based multinomial selection model can be applied for the drawing of an ideal sample.

3.2 Basis of new theory

In the non-empirical theories outlined in Chapter 2, the sample drawing model leads to samples in which the total number of particles is constant (Hassialis, Wilson) or binomially distributed (Gy). In the new model, presented in this chapter, these conditions do not exist. The eighth criterion of Paragraph 1.5 states: "The theory must be able to describe the sample drawing as a process leading to an approximately constant sample mass or volume." By assuming that particles are randomly selected from the batch until a fixed parameter Z is reached or exceeded this criterion is met. Z is referred to as the boundary value of the sample size. The aim of introducing Z is purely to model the sample drawing on a theoretical level. In the practice, its precise value is unknown. The boundary value of the sample size Z corresponds to the boundary value of the sample mass M in the mass-based approach or the boundary value of the sample volume V in the volume-based approach. In this model, for any sample S the difference between the sample size obtained $Z_{sample}(S)$ and the boundary value of the sample size is smaller than the largest particle size in the sample, $z_{max}(S)$. The latter symbol represents the largest particle mass in the sample, $m_{max}(S)$, in the mass-based approach and the largest particle volume, $v_{max}(S)$, in the volume-based approach. Hence, the relative difference between the boundary value of the sample size and the sample size obtained tends to approach zero with increasing value of $Z_{sample}(S)$.

In practice, samples often contain a large number of particles. This reduces the relative difference between the boundary value of the sample size and the sample size obtained, and there may be cases in which the contribution of this difference to the sampling error is negligible. Therefore, if the boundary value of the sample size is unknown, the above-described sampling design can be used to model the sample drawing: After a sample S has been drawn, it is assumed that the sample was drawn according to the size-based multinomial selection process, with boundary value of the sample size equal to the actual sample size $Z_{sample}(S)$. This process will be simulated for four distinctly different two-dimensional batches in Paragraph 3.5. It will be demonstrated that the samples obtained are distributed in agreement with the size-based multinomial distribution.

3.3 Ideal sampling

For the definition of ideal sampling, it is necessary to identify two stages during the drawing of a sample. The first stage is the (physical) insertion of the sampling device into the batch and the second stage is the retraction of the sampling device from the batch. All particles that are inside the sampling device after completion of the second stage are selected, and thus form part of the sample S. After the first stage is completed but before the second stage is commenced, the open volume inside the sampling device occupies a certain area of the batch. This area is defined here as the target area.

Sampling is defined as ideal if only the particles whose centres of mass were located inside the target area before the sampling device was inserted will be selected to form part of the sample S after completion of the second stage. The other particles, *i.e.* all the particles that had their centres of mass outside the target area before the first stage was commenced, will not be selected after completion of the second stage. Hence, the drawing of an ideal sample is determined only by the position of the target area in the batch before the first stage is commenced and is therefore insensitive for mechanical factors (*e.g.* friction, cohesion and the speed of the sampling device with which it is inserted) occurring during the first and second stage; these effects may result in biased sampling. In the following the ideal sample is denoted as S'.

As mentioned in Paragraph 3.1, the sampling error is subdivided into a contribution due to non-ideal sampling and a contribution due to the distribution of non-identical particles. Both errors will influence the estimates of batch properties: when estimating the true value of a batch property x_{batch}, using the sample S and estimate $\hat{x}_{batch}(S)$, the value of the sampling error in S is defined as the difference between the sample estimate and the true batch value: $\hat{x}_{batch}(S)-x_{batch}$. The value of the error due to non-ideal sampling, $e_{nis}(S)$, in S is defined as the difference between the actual sample estimate, $\hat{x}_{batch}(S)$, and the estimate that would be derived from the hypothetical ideal sample S', $\hat{x}_{batch}(S')$: $e_{nis}(S)=\hat{x}_{batch}(S)-\hat{x}_{batch}(S')$. In addition, a contribution to the sampling error is introduced by the distribution of non-identical particles. The value of this error in a sample S' is denoted as $e_{nip}(S')$ and is defined as the difference between the estimate derived from the ideal sample and the batch value: $e_{nip}(S')=\hat{x}_{batch}(S')-x_{batch}$. $e_{nip}(S')$. In principle, the errors e_{nis} and e_{nip} fluctuate from sample to sample and can be positive or negative. Using both definitions, the following equation is obtained:

$$\hat{x}_{\mathrm{batch}}(S) = x_{\mathrm{batch}} + \left(\hat{x}_{\mathrm{batch}}(S) - \hat{x}_{\mathrm{batch}}(S')\right) + \left(\hat{x}_{\mathrm{batch}}(S') - x_{\mathrm{batch}}\right) = x_{\mathrm{batch}} + e_{\mathrm{nis}}(S) + e_{\mathrm{nip}}(S') \qquad \textbf{(3.1)}$$

In practice, the average value of e_{nis} is often not equal to zero, leading to a biased sample estimate (see Paragraph 4.2 for a mathematical definition of bias). When non-ideal sampling is controlled, which means that the value of e_{nis} is small, statistical fluctuations of the sample estimate are mainly influenced by e_{nip}. In Paragraph 3.4, a model for the selection of randomly arranged particles by ideal sampling is proposed.

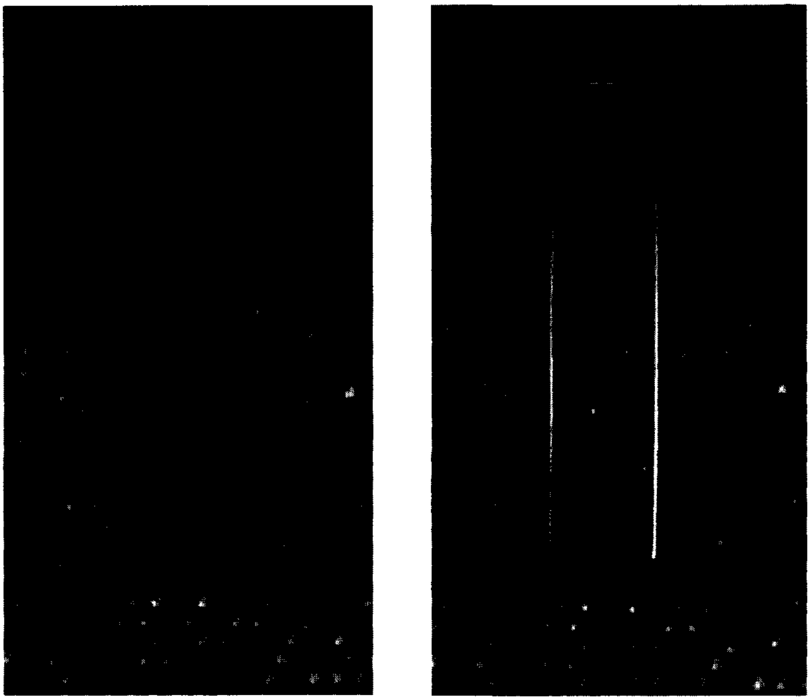


Figure 3.1. *Sampling from a single layer of packed particles, formed under the influence of gravity. The left photograph shows the arrangement of the particles before the first stage is commenced, and the intended position of the sampling lance. Also the target area is indicated in the left photograph. In the left photograph 48 particles are ideally selected. The right photograph shows that disturbance of the arrangement of the particles is evident after completion of the first stage but before the second stage is commenced. Only 46 particles are selected if it is assumed that no particles are lost during the second stage.*

It can be seen from Figure 3.1 that non-ideal sampling can even occur when sampling a single layer of spherical particles with equal diameters. The particles are initially in a static arrangement consisting of only one layer formed under the influence of gravity. Subsequently, an image of the sampling lance is inserted into the particles. The target position of the lance is indicated on the left part of Figure 3.1. Ideally, particles with centres of mass inside the target area are selected. In reality, inserting the lance causes a disturbance of the arrangement. Consequently, it is possible that particles initially expected to be inside the target area fall in practice outside the lance after insertion. Another non-ideal effect is that some particles may be lost when the lance is retracted. It can be derived from Figure 3.1 that 48 particles would be selected in the ideal case. The right-hand part of Figure 3.1 shows that, when assuming that the particles remain in the tube when the sampling lance is pulled up, the actual number of selected particles is 46. This illustrates a deviation from ideal sampling.

The value of e_{nis} depends on the property estimated and its distribution over the particles in the batch and sample. In case of spherical particles with equal dimensions and equal properties, the occurrence of non-ideal sampling generally does not lead to a biased sampling. However, with properties unequally distributed over the particles, the occurrence of non-ideal sampling may lead to a biased sampling.

Finally, some possible numerical values for e_{nis} are derived for the example depicted in Figure 3.1. The situation is considered in which x_{batch} represents the mass concentration of an arbitrary compound, which is estimated using the mass concentration in the sample, here represented by $\hat{x}_{batch}(S)$. It is assumed that all particles have identical masses.

The preceding shows that the ideal sample contains all particles that are finally sampled plus two additional particles. In the following, two numerical examples are given. Firstly, as a worst-case scenario, it is assumed that the concentrations in these two particles are both 1.0, while the concentrations in all other particles are 0.0. In this case, the mass concentration in the ideal sample is:

$$\hat{x}_{batch}(S') = \frac{2}{48} = 0.042 \ \text{g/g}$$

Assuming no particles are lost during the second stage, the mass concentration in the final sample is:

$$\hat{x}_{batch}(S) = 0.000 \ \text{g/g}$$

Hence the value of e_{nis} is:

$$e_{nis}(S) = \hat{x}_{batch}(S) - \hat{x}_{batch}(S') = 0.000 - 0.042 = -0.042 \ \text{g/g}$$

As a second example, the situation is considered in which the concentrations in 24 of the 48 ideally selected particles are 1.0, while the concentrations in the other 24 particles are 0.0. The concentration in the ideal sample is:

$$\hat{x}_{batch}(S') = \frac{24}{48} = 0.500 \ \text{g/g}$$

If it is assumed that the concentrations in the two particles ideally selected, but not forming part of the sample after completion of the first and second stage, are 1.0, the concentration in S is:

$$\hat{x}_{\text{batch}}(S) = \frac{22}{46} = 0.478 \text{ g/g}$$

Hence, the value of e_{nis} is:

$$e_{nis}(S) = \hat{x}_{\text{batch}}(S) - \hat{x}_{\text{batch}}(S') = 0.478 - 0.500 = -0.022 \text{ g/g}$$

This value is smaller than the value of e_{nis} in the previous example.

3.4 Mathematical algorithm

The particles in the batch are classified into T distinct classes differing in one or more parameters. The particle size, particle mass and the concentration of the property of interest are constant within each class. In practice, the choice of a classification may be imperfect due to a remaining variability in the particle sizes, particle masses and concentrations in the particles belonging to a single class. Theoretically, however, classification poses no problems, because every particle in the batch could form a distinct class. In the latter case, T equals the number of particles in the batch.

For sake of simplicity and to proceed on the example presented in the preceding paragraph, it is assumed that a sampling lance is used, but the similar considerations will be possible if using other sampling devices. When the lance is inserted, particles cross the entrance of the lance in succession. It is assumed that the arrangement of particles is not disturbed by insertion of the lance, *i.e.* sampling is ideal. Because the particles are randomly arranged, this corresponds to repeated multinomial selections. With every selection, there are T independent possibilities and the probability of drawing a particle of type i is equal to p_i. The value of p_i varies during the insertion of the sampling lance in the batch. For the first selection, every particle has a selection probability of $1/N_{\text{batch}}$, where N_{batch} is the total number of particles in the batch before sampling. Hence, during the first selection, it is guaranteed that $p_i = p'_i$ in which $p'_i \equiv N_{i,\text{batch}} / N_{\text{batch}}$, where $N_{i,\text{batch}}$ is the total number of particles belonging to the i^{th} class in the batch before sampling and '≡' stands for 'is by definition equal to.' For subsequent selections, the probability of selecting a particle is either zero when a particle is already selected or $1/(N_{\text{batch}}-n_{\text{sample}})$, where n_{sample} is the number of particles selected, when the particle is not selected during any of the previous selections. Therefore, the value of p_i depends on the previously selected particles and the number of particles in the batch. On the condition that n_{sample} particles have been selected of which n_i are of type i (*i.e.* $n_{\text{sample}} = \sum_{i=1}^{T} n_i$), the probability of selecting a particle of type i during the next selection is:

$$p_i = \frac{N_{i,\text{batch}} - n_i}{N_{\text{batch}} - n_{\text{sample}}} \tag{3.2}$$

Dividing the numerator and the denominator on the right-hand side by N_{batch} results in:

$$p_i = \left(p_i' - \frac{n_i}{N_{\text{batch}}} \right) \frac{N_{\text{batch}}}{N_{\text{batch}} - n_{\text{sample}}} \tag{3.3}$$

In this thesis, many results are calculated in the limit of an infinite value of N_{batch}, denoted as $\lim\limits_{N_{\text{batch}} \to \infty}$. By definition, an important property of this limiting process, as used throughout this thesis, is that the taking of the limit $N_{\text{batch}}=\infty$ does not alter the average composition of the batch, *i.e.* the values of p_i' (for all i between 1 and T) remain constant. This is stated as: $\lim\limits_{N_{\text{batch}} \to \infty} p_i' = p_i'$ for all i between 1 and T. In the limit of an infinite value of N_{batch} it follows that:

$$\lim_{N_{\text{batch}} \to \infty} p_i = \lim_{N_{\text{batch}} \to \infty} \left(p_i' - n_i / N_{\text{batch}} \right) N_{\text{batch}} / \left(N_{\text{batch}} - k \right) = p_i' \tag{3.4}$$

This demonstrates that Equation 3.3 can be interpreted as follows: p_i' is the probability of drawing a particle of type i when sampling with replacement or, equivalently, $N_{\text{batch}}=\infty$. To correct for sampling without replacement, p_i' is lowered by n_i/N_{batch}. Subsequently, p_i' has to be multiplied by $N_{\text{batch}}/(N_{\text{batch}}-n_{\text{sample}})$ to assure that the probability of selecting an arbitrary particle remains one.

Using the indicator $I_i(S)$, which is one when the i^{th} particle in the batch is part of S and zero when the i^{th} particle of the batch is not part of S, the sample size of S, $Z_{\text{sample}}(S)$, is defined as:

$$Z_{\text{sample}}(S) = \sum_{i=1}^{N_{\text{batch}}} I_i(S) z_{n(i)} \tag{3.5}$$

in which n(i) is the class of the i^{th} particle in the batch and $z_{n(i)}$ is the size (*i.e.* the mass or volume in the mass-based or volume-based approach respectively and one in the number-based approach) of a particle belonging to the $n(i)^{th}$ class. It is assumed that the multinomial selections are terminated when $Z_{\text{sample}}(S)$ reaches or exceeds the boundary value of the sample size, Z. This is the principle of the size-based multinomial selections. Note that this algorithm only works if Z is smaller than or equal to the size of the batch, Z_{batch}. Due to the discrete nature of the particles, a difference δ(S) may exist between the theoretical boundary value of the sample size and the experimentally obtained sample size. Because δ(S) is smaller than the largest particle size of the sample, the relative difference between the obtained sample size and the boundary value of the sample size $(Z_{\text{sample}}(S)-Z)/Z_{\text{sample}}(S)$ becomes arbitrarily small, if the sample size is sufficiently large. For example, the requirement that the relative difference between the obtained sample size and the boundary value of the sample size is smaller than a factor ε results in a minimum value for the sample size:

$$Z_{\text{sample}}(S) > \frac{Z_{\text{sample}}(S) - Z}{\varepsilon} \tag{3.6}$$

The numerator on the right-hand side cannot be larger than the maximum particle size $z_{\max}(S)$ in the sample S. Therefore, for a sample S, the above condition is certainly met if:

$$Z_{\text{sample}}(S) > \frac{z_{\max}(S)}{\varepsilon} \tag{3.7}$$

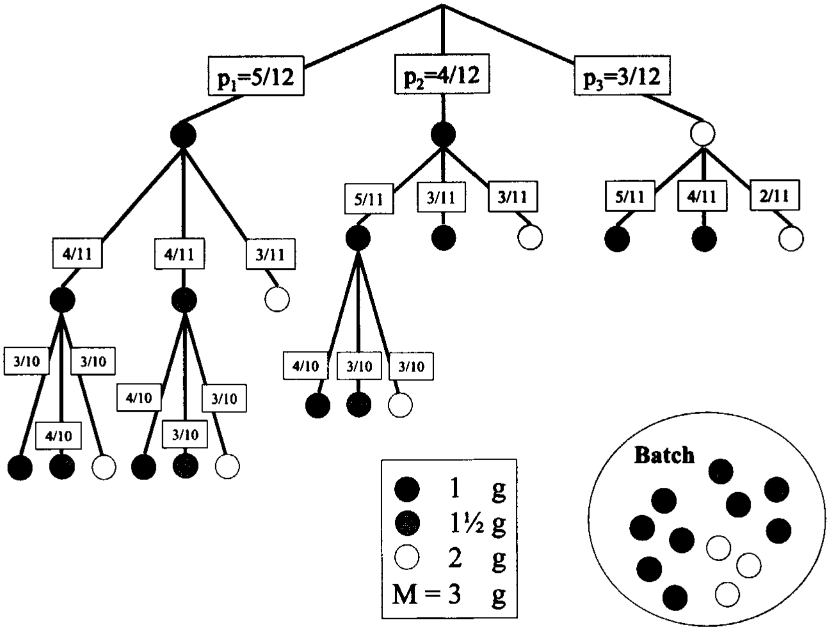


Figure 3.2. *Example of the evolution of* p_i *during a sampling process. Initially, the batch consists of 12 particles belonging to three distinct classes. The particle masses are respectively 1 g, 1.5 g and 2 g. The boundary value of the sample mass* M *is 3 g. In each branch, the probability* p_i *is given.*

In Figure 3.2 the evolution of p_i, calculated using Equation 3.3, is illustrated for sampling from a batch containing 12 particles belonging to three distinct types with particle masses chosen between 1 and 2 g: 1 g, 1.5 g and 2 g respectively. In this example, the boundary value of the sample mass was set to 3 g.

For every possible sample S, the probability P(S) of being drawn is the product of the successive probabilities to draw a particle:

$$P(S) = \prod_{k=0}^{N_{sample}(S)-1} \frac{1}{N_{batch} - k} \qquad \textbf{(3.8)}$$

in which $N_{sample}(S)$ is the total number of particles in S. The probability of going to one of the end-points of the multinomial tree is generally larger than P(S), because the former probability is that of drawing an arbitrary sample from a collection of many samples (namely all samples that are drawn using a sequence equivalent to the sequence leading to the end-point), while the latter probability is only that of drawing a single sample. As the probability of going to one of the end-points of the multinomial tree depends on the obtained sample composition and the batch composition, this probability is denoted as $P(n_1,\dots,n_T,p'_1,\dots p'_T,N_{batch})$. In other words, $P(n_1,\dots,n_T,p'_1,\dots p'_T,N_{batch})$ is the probability of going to an end-point of the multinomial tree leading to a sample with n_i particles of type i for all i between 1 and T. $P(n_1,\dots,n_T,p'_1,\dots p'_T,N_{batch})$ is the product of the successive probabilities p_i. Using Equation 3.3, this can be expressed as:

$$P(n_1,\dots,n_T,p'_1,\dots p'_T,N_{batch}) = \left(\prod_{k=0}^{n_{sample}-1} N_{batch}\big/(N_{batch}-k)\right)\prod_{j=1}^{T}\left[\prod_{i=0}^{n_j-1}\left(p'_j - i/N_{batch}\right)\right] \qquad \textbf{(3.9)}$$

where T is the number of classes and $n_{sample} = \sum_{i=1}^{T} n_i$. It is noted here that $P(n_1,\dots,n_T,p'_1,\dots p'_T,N_{batch})$ is fundamentally different from p_i, the probability to draw a particle of type i during the selection of a particle. Only if Z is reached or exceeded after one particle is selected that belongs to the i^{th} class, $P(n_1,\dots,n_T,p'_1,\dots p'_T,N_{batch}) = p_i$.

When U is the collection of all possible samples, the 'expected value of $\hat{x}_{batch}$' is defined by:

$$E(\hat{x}_{batch}) \equiv \sum_{S \in U} P(S)\hat{x}_{batch}(S) \qquad \textbf{(3.10)}$$

The variance is defined by:

$$V\left(\hat{x}_{batch}\right) \equiv \sum_{S \in U} P(S)\left(\hat{x}_{batch}(S) - E\left(\hat{x}_{batch}\right)\right)^2 \tag{3.11}$$

Instead of a summation over all possible samples, the expected value and variance can also be written as a T-fold summation over all possible values of n_i for all i between 1 and T. Doing this will facilitate the calculation of analytical expressions for expected value and variance in Chapter 4. A general requirement, which is met by all practical estimators, is that, given a sample S and the properties of the T classes, $\hat{x}_{batch}(S)$, depends only on the number of particles in S that belong to the i^{th} class, $N_i(S)$, for all i between 1 and T. Hence $\hat{x}_{batch}(S)$ can be written as a function of the T variables $N_i(S)$:

$$\hat{x}_{batch}(S) = x\left(N_1(S), \ldots, N_T(S)\right) \tag{3.12}$$

When $F(n_1,\ldots,n_T,N_{1,batch},\ldots,N_{T,batch},Z)$ denotes the number of end-points of the multinomial tree resulting in samples with n_i particles of type i for all i between 1 and T, the probability of drawing an arbitrary sample with n_i particles of type i for all i between 1 and T can be written as:

$$P\left(S \middle| N_1(S) = n_1, \ldots, N_T(S) = n_T\right) = F\left(n_1, \ldots, n_T, N_{1,batch}, \ldots, N_{T,batch}, Z\right) P\left(n_1, \ldots, n_T, p_1', \ldots, p_T', N_{batch}\right) \tag{3.13}$$

Using the above equation, the expected value can be written as:

$$E\left(\hat{x}_{batch}\right) \equiv \sum_{S \in U} P(S)\hat{x}_{batch}(S) = \sum_{n_1=0}^{N_{1,batch}} \cdots \sum_{n_T=0}^{N_{T,batch}} \left[P\left(S \middle| N_1(S) = n_1, \ldots, N_T(S) = n_T\right) x\left(n_1, \ldots, n_T\right)\right] =$$

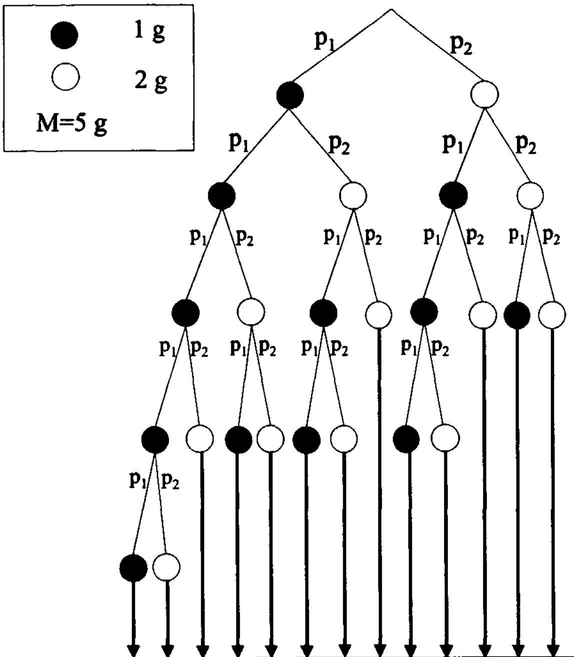
$$\sum_{n_1=0}^{N_{1,batch}} \cdots \sum_{n_T=0}^{N_{T,batch}} \left[F\left(n_1, \ldots, n_T, N_{1,batch}, \ldots N_{T,batch}, Z\right) x\left(n_1, \ldots, n_T\right) P\left(n_1, \ldots, n_T, p_1', \ldots p_T', N_{batch}\right)\right] \tag{3.14}$$

A similar equation can be derived for the variance:

$$\left.\begin{aligned}
&V\left(\hat{x}_{batch}\right) \equiv \sum_{S \in U} P(S)\left(\hat{x}_{batch}(S) - E\left(\hat{x}_{batch}\right)\right)^2 = \\
&\sum_{n_1=0}^{N_{1,batch}} \cdots \sum_{n_T=0}^{N_{T,batch}} \left[P\left(S \middle| N_1(S) = n_1, \ldots, N_T(S) = n_T\right)\left(x\left(n_1, \ldots, n_T\right) - E\left(\hat{x}_{batch}\right)\right)^2 \right] = \\
&\sum_{n_1=0}^{N_{1,batch}} \cdots \sum_{n_T=0}^{N_{T,batch}} \left[F\left(n_1, \ldots, n_T, N_{1,batch}, \ldots, N_{T,batch}, Z\right) \times \right. \\
&\left. \left(x\left(n_1, \ldots, n_T\right) - E\left(\hat{x}_{batch}\right)\right)^2 P\left(n_1, \ldots, n_T, p_1', \ldots p_T', N_{batch}\right) \right]
\end{aligned}\right\} \tag{3.15}$$

Equations 3.14 and 3.15 will be used in the next paragraph. This paragraph concludes with two more examples of the proposed sampling algorithm.

In Figures 3.3 and 3.5, the proposed algorithm is illustrated for the case of two and three distinct particle types respectively in the mass-based approach. The selection of particles is terminated when the sample mass obtained reaches or exceeds the boundary value (5 g and 4 g in Figures 3.3 and 3.5 respectively). Figures 3.4 and 3.6 give the relative frequency of possible routes to a certain sample mass and the probability distribution of the possible sample masses. The relative frequency is defined as the ratio between the number of routes to a certain sample mass and the total number of possible routes. For calculation of the probability distributions in Figure 3.4 and 3.6, it is necessary to assume numerical values for p_1, p_2 and p_3. For the construction of Figures 3.4 and 3.6, it is assumed that $p_1' = p_2' = 1/2$ and $p_1' = p_2' = p_3' = 1/3$ respectively. It is also assumed that the number of particles in the batch is infinite, leading to constant selection probabilities. Therefore, for Figure 3.4, it is assumed that $p_1=p_2=1/2$ and for Figure 3.6, it is assumed that $p_1=p_2=p_3=1/3$. A different choice of values will, of course, lead to different probability distributions. On the other hand, the relative frequency does not depend on the values of the selection probabilities.



M_{sample}	5	6	5	5	6	5	6	5	5	6	5	5	6
●	5	4	3	3	2	3	2	1	3	2	1	1	0
○	0	1	1	1	2	1	2	2	1	2	2	2	3

Figure 3.3. Schematic illustration of the proposed sampling algorithm. At the end of each branch, the actual sample mass and the composition of the sample are given. Selection of particles stops when the boundary value of the sample mass (5 g) is reached or exceeded. It is assumed that a sample never contains all the particles of one type present in the batch, so two choices always remain during each selection.

The relative frequency is zero in the distributions of relative frequencies plotted in Figures 3.4 and 3.6 for sample masses larger than 6 g and 5.5 g respectively. Therefore, both figures demonstrate also that the sample mass can never exceed the boundary value by more than the mass m_{max} of the largest particle batch. In Chapter 7, the proposed sampling process will be simulated and compared with Wilson's multinomial model and Gy's Bernoulli model (see Chapter 2). It will be shown that samples drawn using Wilson's and Gy's procedures have more variation in the sample mass compared to samples drawn by the here described newly developed model.

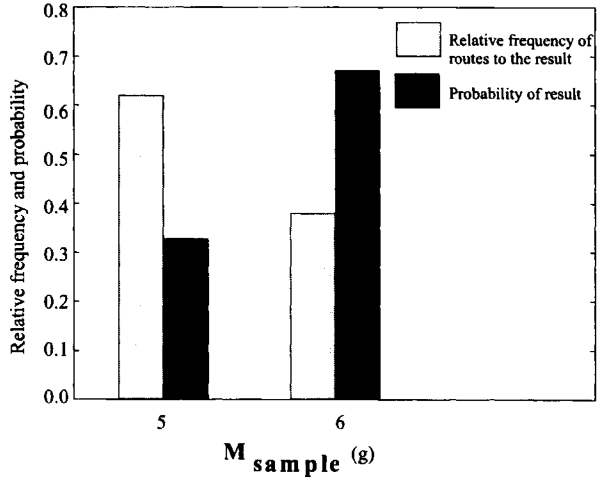


Figure 3.4. Relative frequency of the possible routes to a certain sample mass and the probability distribution of the possible sample masses, for the tree depicted in Figure 3.3. The relative frequency of routes is defined as the number of routes to a certain sample mass divided by all possible routes. The probability of drawing a certain sample mass is the sum of probabilities of all samples with that mass. For calculation of the probabilities, it is assumed that $p'_1 = p'_2 = 1/2$*. It is also assumed that the number of particles in the batch is infinite, leading to constant selection probabilities* $p_1 = p_2 = 1/2$.

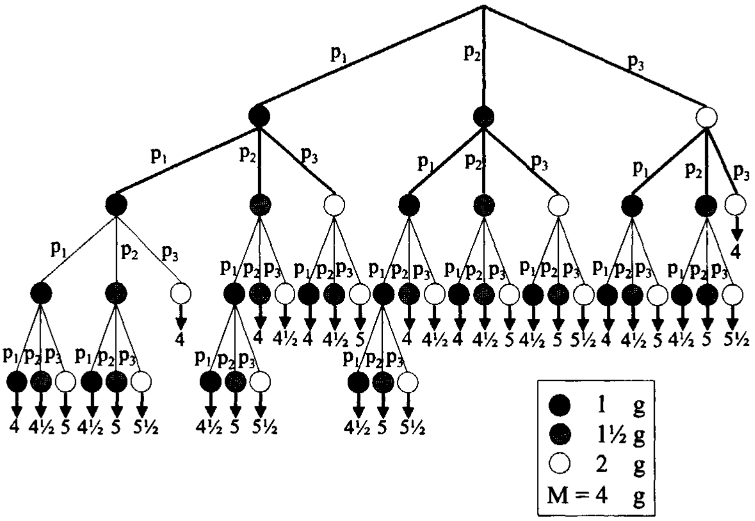


Figure 3.5. Schematic illustration of the proposed sampling algorithm. At the end of each branch, the actual sample mass is given. Selection of particles stops when the boundary value of the sample mass (4 g) is reached or exceeded. It is assumed that a sample never contains all the particles of one type present in the batch, so three choices always remain during each selection.

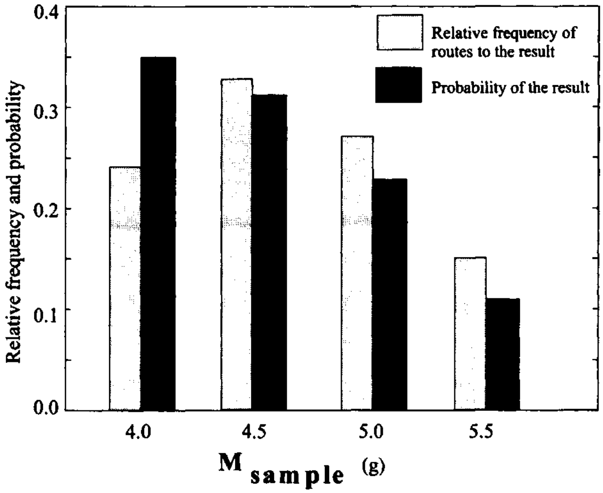


Figure 3.6. *Relative frequency of the possible routes to a certain sample mass and the probability distribution of the possible sample masses, for the tree depicted in Figure 3.5. The relative frequency of routes is defined as the number of routes to a certain sample mass divided by all possible routes. The probability of drawing a certain sample mass is the sum of probabilities of all samples with that mass. For calculation of the probabilities, it is assumed that* $p'_1 = p'_2 = p'_3 = 1/3$. *It is also assumed that the number of particles in the batch is infinite, leading to constant selection probabilities* $p_1=p_2=p_3=1/3$.

The parameters p_i, z_i and Z determine uniquely the complete set of possible sample compositions and their probabilities of being drawn. Therefore, if the mass concentration of the property of interest in a particle of type i is denoted as a_i and the mass of a particle of type i is denoted as m_i, the variance of the mass concentration in a sample, $V(a_{sample})$, is uniquely determined by the parameters p_i, z_i, m_i, a_i and Z. In the next paragraph, the variance is calculated using a calculation of $P(S \mid N_1(S)=n_1,\dots,N_T(S)=n_T)$ for every possible sample composition combined with Equation 3.15.

3.5 Simulations

For the simulations in this paragraph, the mass-based approach is adopted. This implies that the boundary value of the sample size, Z, corresponds to a boundary value of the sample mass, M. Because samples often contain a large number of particles, the difference between the boundary value of the sample mass and the sample mass obtained is relatively small. This indicates that the proposed sampling design can potentially be applied when the boundary value of the sample mass is unknown. In this case, after a sample S has been drawn, it is assumed that the sample was drawn according to the mass-based approach, with boundary value of the sample mass equal to the actual sample mass of S, $M_{sample}(S)$.

In order to test the validity of this assumption, simulations in four distinct two-dimensional batches were performed. The batches contained particles of three distinct types. Each type contained 1,000 particles in the batch, so each batch counted a total of 3,000 particles. The composition of the batches is given in Table 3.1. The particles (diameter 0.01 cm) were positioned at random and non-overlapping positions using the following algorithm: The position of the first particle in the batch is chosen at random. The second particle is then allocated to a preliminary random position in the batch. If the first and second particle overlap, a new random position is chosen for the second particle. This is repeated until there is no overlap between the first and second particle. The allocation of the other particles to a random position is similar: the i^{th} particle is allocated to a random preliminary position in the batch. If there is an overlap with any of the previous i-1 particles, a new random position is chosen for the i^{th} particle. The definitive position of the i^{th} particle is obtained when there is no overlap with any of the previous i-1 particles.

In order to obtain a well-mixed batch, the type of a particle that is allocated to a random position in the batch differs from the previous allocated particle in the following way: after a particle of type 1 is allocated to its final position, the next particle is of type 2. After a particle of type 2 is allocated to its final position the next particle is of type 3. Finally, after a particle of type 3 has been allocated to its final position, the next particle is again of type 1. After all particles have been positioned in the batch according to the above-described procedure, a sample was drawn under ideal conditions (as defined in Paragraph 3.3) with a lance. A graphical illustration of the sampling of a single batch is depicted in Figure 3.7. The procedure to draw a sample was repeated 100 times for each batch: after a single sample was drawn from a batch, all the particles were re-allocated to new positions in the batch by choosing 3,000 new random positions using the algorithm described above.

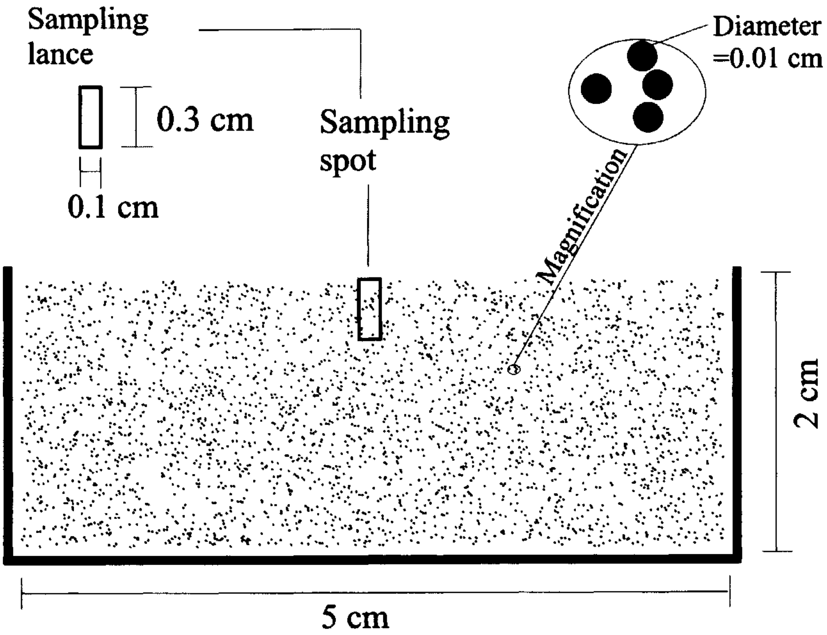


Figure 3.7. Simulation of sampling from a batch consisting of 3,000 particles. Because the target area is 0.3% of the total area of the batch and the batch contains 3,000 particles, on average the samples contain 9 particles.

		Batch 1	Batch 2	Batch 3	Batch 4
Type 1	m_1 (μg)	2.0	1.0	2.0	2.0
10^3 particles	a_1 (g/g)	0.0	0.0	1.0	0.0
Type 2	m_2 (μg)	1.0	1.0	1.0	2.0
10^3 particles	a_2 (g/g)	0.5	0.0	0.0	0.0
Type 3	m_3 (μg)	1.0	1.0	1.0	1.0
10^3 particles	a_3 (g/g)	1.0	1.0	0.0	1.0

Table 3.1. Compositions of the four simulated batches.

Figure 3.8 gives the distributions of sample masses obtained drawn from batch 1, 2, 3 and 4. It can be seen that the sample masses obtained have different values. Hence, for each sample, the sample drawing model is separately modelled by mass-based multinomial selections, using a boundary value of the sample mass equal to the obtained sample mass.

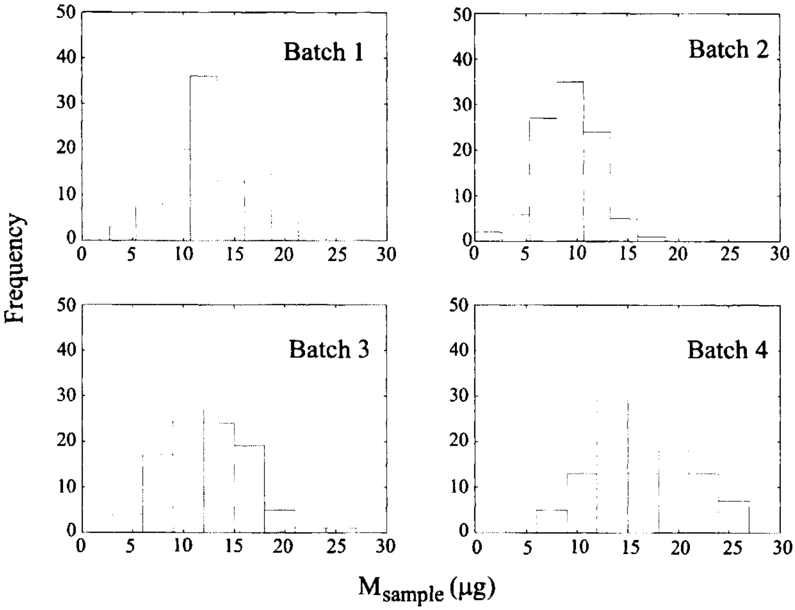


Figure 3.8. Distribution of sample masses drawn from the first, second, third and fourth batch.

For the i^{th} simulated sample S_i, a new variable $z_{score,i}$ was calculated using the following equation:

$$z_{score,i} = \frac{a_{batch} - a_{sample}(S_i)}{\sqrt{V(a_{sample})_{M_{sample}(S_i)}}} \tag{3.16}$$

where

a_{batch}= the mass concentration in the batch (which is constant for all samples drawn from the same batch),

$a_{sample}(S_i)$= the mass concentration in the i^{th} sample and

$V(a_{sample})_{M_{sample}(S_i)}$ = the theoretical variance of the concentration in samples drawn using mass-based multinomial selections with a boundary value of the sample mass equal to $M_{sample}(S_i)$.

The latter variance is calculated using a calculation of $P(S|N_1(S)=n_1,...,N_T(S)=n_T)$ for all possible sample compositions combined with Equation 3.15. The order in which the quantities $P(n_1,...,n_T,p'_1,...,p'_T,N_{batch})$ and $F(n_1,...,n_T,N_{1,batch},...,N_{T,batch},Z)$ were calculated is illustrated in Figure 3.9. Hence, if the sample drawing is successfully described by a mass-based multinomial selection process, the variable $z_{\mathrm{score},i}$ is distributed with a mean equal to zero and a standard deviation equal to one.

The variable χ^2 is defined by

$$\chi^2 = \sum_{i=1}^{N} z_{\mathrm{score},i}^2 / N \tag{3.17}$$

in which N is the number of degrees of freedom, which is equal to 100 here. From statistics (see *e.g.* Barlow, 1989), it is known that χ^2 is approximately normally distributed with mean equal to one and variance equal to 1/(2N). For batch 1, 2, 3 and 4 respectively the following values for χ^2 were obtained respectively: 0.94, 0.84, 0.84 and 0.99. These figures do not differ from one by more that three times the standard deviation ($1/\sqrt{2N} = 0.07$).

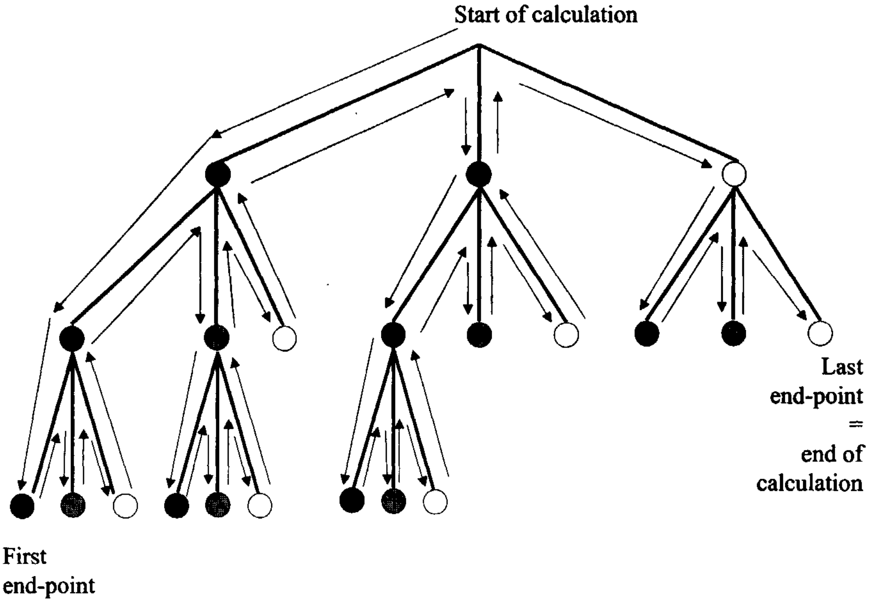


Figure 3.9. *Example of a sequence in which the end-points were calculated, as indicated by the arrows. The depicted multinomial tree is similar to the tree depicted in Figure 3.2.*

Figure 3.10 illustrates that the distributions obtained for the four distinct batches match the theoretical distribution.

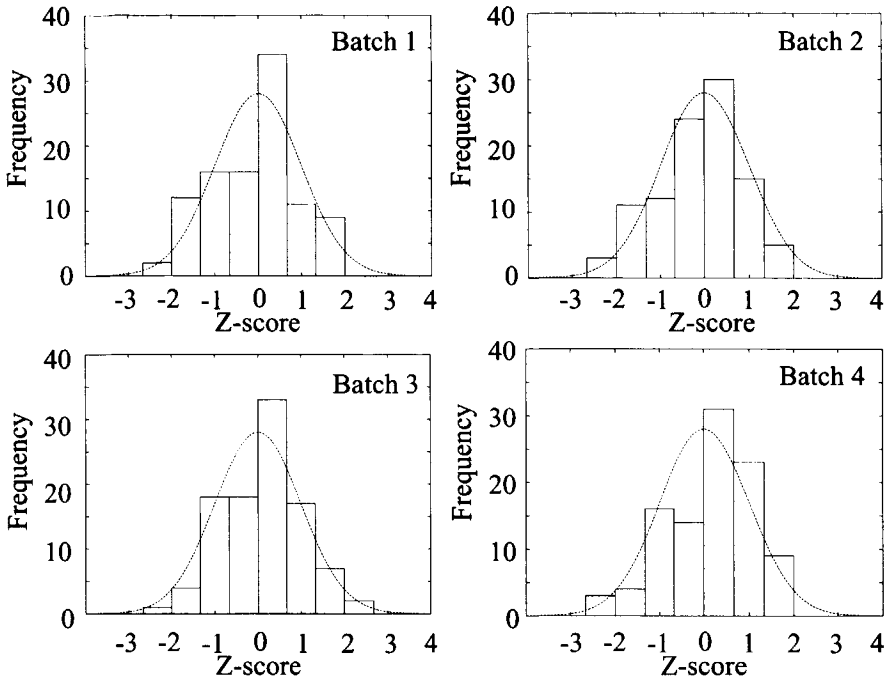


Figure 3.10. Distributions of z-scores for 4 distinct batches. From each batch 100 samples were drawn. After a sample has been drawn, all the particles are replaced in the batch and allocated to new, random positions. The curves represent normal distributions with a mean equal to zero and a variance equal to one.

Hence, the mass-based multinomial selections statistically describe the distribution of mass concentrations in the samples drawn.

3.6 Results

Size-based multinomial selections can model the drawing of an ideal sample from a randomly mixed batch. This process was investigated and verified with simulations using four distinctly different batches. For the simulated batches, the samples were statistically distributed according to the mass-based multinomial distribution. The boundary value of the sample size can, with limited consequences to the accuracy, be estimated using the sampled mass.

3.7 Conclusions

The model presented in this chapter serves as the underlying mathematical algorithm for development of a new sampling theory. Thus the second criteria ("The theory must be based on a model of the drawing of a sample on the level of (groups of) particles.") is met in the new theory. The model leads to samples of either an approximately constant mass or an approximately constant volume. Therefore, the eighth criterion ("The theory must be able to describe the sample drawing as a process leading to an approximately constant sample mass or volume.") is met.

3.8 References

R. J. Barlow (1989), *Statistics. A guide to the use of statistical methods in the physical sciences.* John Wiley & Sons, Chichester, 204 pp.

G. K. Robinson and P. W. Cleary (1999), The conditions for sampling of particulate materials to be unbiased – investigation using granular flow modeling. *Minerals Engineering*, **12**, p. 1101-1118.

Chapter 4 Size-based sampling theory[9]

Because the sample concentration is the ratio of two sample totals, it is important to study the variance of a sample total. It is demonstrated that for calculation of this variance, the covariances between the numbers of particles belonging to the classes in the sample are required. Using a specified method, these covariances are calculated in the size-based approach. As a final result, the variance of the sample concentration, estimated using the properties of the particles in the batch, is calculated.

4.1 Introduction

In view of the first criterion for a sampling theory ("The theory must provide an equation for the variance of the sample concentration, containing the mass or volume sampled and an arbitrary number of additional parameters."), an equation for the size-variance relationship will be derived using the mathematical algorithm, size-based multinomial selections, which was presented in Chapter 3.

Firstly, principles from finite population sampling, relevant to the sampling of particulate materials, are reviewed.

4.2 Principles from finite population sampling

Denoting the actual value of the property simply by x_{batch}, the sample S provides an estimate for x_{batch} denoted as $\hat{x}_{batch}(S)$. This leads to the definition of the estimator $\hat{x}_{batch}$ as a function U→R in which U is the set of all possible samples and R the set of all real values. For any sample S from U, $\hat{x}_{batch}$ provides the corresponding value $\hat{x}_{batch}(S)$, which may vary from sample to sample. The value of the sampling error, $\hat{x}_{batch}(S)-x_{batch}$, will also fluctuate from sample to sample. When it is assumed that sampling is ideal, *i.e.* the ideal sample is equal to the actual sample ($S'=S$), the error due to non-ideal sampling is zero ($e_{nis}(S)=0$)). In this case, it can be derived, using

[9] Part of this chapter has been published in: B. Geelhoed and H. J. Glass (2002) Multinomial distributions applied to random sampling of particulate materials, *Statistica Neerlandica*, **56**, p. 58-76.

Equation 3.1, that:

$$\hat{x}_{batch}(S) - x_{batch} = e_{nip}(S) \tag{4.1}$$

in which $e_{nip}(S)$ is the value of the error due to the distribution of non-identical particles. Because in this work a random variable is defined as a function U→R, $\hat{x}_{batch}$ is by definition a random variable. The sampling errors e_{nip} and e_{nis} are by definition random variables, where the variable e_{nip} is a function of the ideal sample and e_{nis} is a function of both the ideal and the actual (potentially non-ideal) sample. The concept of a random variable is useful for sampling, because it facilitates referring to a sample property without the necessity of referring to its value in a specific sample. This allows definition of statistical properties like the expected value and variance of a random variable x (x can represent e_{nip}, e_{nis}, $\hat{x}_{batch}$ or any other random variable; x(S) represents its value derived from the sample S), which are given below.

The statistical variability of x is caused by the statistical distribution of samples and the fact that x can yield different values for distinct samples. Therefore, in order to define the expected value and variance of x, the probability of drawing a sample S is specified by P(S). P(S) is generally denoted as the sampling design. In Chapter 3, it was demonstrated that size-based multinomial selections can model the drawing of an ideal sampling from a random arrangement of particles. Therefore, in Paragraph 3.4, an expression for P(S) in the size-based approach was derived. However, irrespective of the precise form of P(S), the expected value of x is defined as:

$$E(x) \equiv \sum_{S \in U} P(S)x(S) \tag{4.2}$$

where U is again the collection of all possible samples. For the ideal sampling from a random arrangement of particles, the expression for P(S) derived in Chapter 3, Equation 3.8, may be substituted in the above equation. The bias of an estimator is defined as:

$$B(\hat{x}_{batch}) \equiv E(\hat{x}_{batch}) - x_{batch} \tag{4.3}$$

When the bias is positive, $\hat{x}_{batch}$ will have the tendency to overestimate the true value, while a negative bias indicates the tendency to underestimate the true value. Therefore, it is desirable that the absolute value of the bias is as small as possible. When sampling is ideal, Equation 4.1 may be substituted in the above equation, resulting in:

$$B(\hat{x}_{batch}) = E(e_{nip}) \tag{4.4}$$

An estimator is called unbiased when $B(\hat{x}_{batch})$ is zero (Barnett, 1974). Even an unbiased estimator may result in estimates with large sampling errors. Therefore, the variance of a random variable is defined as a measure of potential variation of around its expected value:

$$V(x) \equiv \sum_{S \in U} P(S)(x(S) - E(x))^2 \tag{4.5}$$

Because a larger variance allows larger deviations from the expected value, it is desirable, for estimators, to have a low variance. When sampling is ideal, Equation 4.1 may be substituted in the above equation. This results in $V(\hat{x}_{batch}) = V(e_{nip})$.

Because P(S) is generally determined by the sampling method, the bias and variance are also determined by the sampling method. Only for the ideal sampling from a random arrangement of particles, Equation 3.8 may be used for substitution of P(S) in Equation 4.4 and 4.5.

The concentration in the sample is an important random variable, because it is generally used as an estimator for the concentration in the batch. Therefore, in Paragraph 4.7, the expected value and variance of the concentration of an arbitrary property in the sample are calculated. The concentration is defined as Y_{batch}/Z_{batch}, where Y_{batch} and Z_{batch} are population totals (see Särndal *et al*, 1992) given by:

$$Y_{batch} = \sum_{i=1}^{N_{batch}} y_{n(i)} \tag{4.6}$$

$$Z_{batch} = \sum_{i=1}^{N_{batch}} z_{n(i)} \tag{4.7}$$

in which N_{batch} is the total number of particles in the batch, n(i) the class of the i^{th} particle in the batch, $y_{n(i)}$ is the value of the property of interest in a particle belonging to the $n(i)^{th}$ class and $z_{n(i)}$ is the size (*i.e.* mass or volume in the mass-based or volume-based approach respectively) of a particle belonging to the $n(i)^{th}$ class. The denominator Z_{batch} is the size of the batch, which can correspond to the mass or volume.

A general expression for the sample concentration is Y_{sample}/Z_{sample}, where

$$Y_{sample} = \sum_{i=1}^{T} N_i y_i \tag{4.8}$$

$$Z_{sample} = \sum_{i=1}^{T} N_i z_i \tag{4.9}$$

where y_i, z_i, N_i and T are the value of the property of a particle of type i, the size of a particle of type i, the number of particles belonging to the ith class in the sample and the number of particle classes respectively. The denominator Z_{sample} is the size of the sample, which can represent the mass of, volume of or number of particles in the sample. Because the denominator and numerator are generally measured without analysing every particle in the sample separately, the sample concentration is extensively used as an estimator for the batch concentration.

4.3 Covariances

The sample concentration is a ratio of two sample totals. Therefore, in this paragraph, the variance of a sample total is considered. Because the particles are classified into T classes, the sample totals Y_{sample} and Z_{sample} were parameterised using the variables y_i, z_i, N_i and T in Equations 4.8 and 4.9. It is noted that N_i is a random variable. For a sample S, the value of N_i is denoted as $N_i(S)$. Consequently, the sample totals Y_{sample} and Z_{sample} are also random variables. Using the definitions of expected value and variance results in:

$$V(Y_{sample}) = \sum_{i=1}^{T}\sum_{j=1}^{T} y_i y_j \left(E(N_i N_j) - E(N_i)E(N_j)\right) \qquad (4.10)$$

A similar expression can be obtained for $V(Z_{sample})$. Hence, an expression for the covariances, $E(N_iN_j)-E(N_i)E(N_j)$, is required. First, in Paragraph 4.4, the special case of sampling a batch containing only two particle classes in the number-based approach is considered. It will be derived that in this special case, N_1 and N_2 are distributed according to the hypergeometric distribution. In Paragraph 4.5, a method will be developed for calculation of the covariances, $E(N_iN_j)-E(N_i)E(N_j)$, in the general case of sampling a batch containing T particle classes in the size-based approach.

4.4 Hypergeometric distribution

In this paragraph, the special case of sampling of a batch containing only two distinct particle classes (T=2) in the number-based approach is considered. The number-based approach can be considered as an alternative to the mass-based or volume-based approach, and is defined as follows: all particle sizes are equal to one and the boundary value of the sample size is an integer N that equals the number of particles in the sample. In the number-based approach with T=2, Equation 3.9 becomes:

$$P(n_1,n_2,p_1',p_2',N_{batch})=\left(\prod_{k=0}^{N-1}N_{batch}/(N_{batch}-k)\right)\left[\prod_{i=0}^{n_1-1}(p_1'-i/N_{batch})\right]\left[\prod_{i=0}^{n_2-1}(p_2'-i/N_{batch})\right] \quad (4.11)$$

As stated in Chapter 3, the probability of drawing an arbitrary sample with n_1 particles of type 1 and n_2 particles of type 2, $P(S\,|\,N_1(S)=n_1,N_2(S)=n_2)$, is the product of $P(n_1,n_2,p_1',p_2',N_{batch})$ and the number of possible end-points of the binomial tree resulting in a sample with n_1 particles of type 1 and n_2 particles of type 2, denoted as $F(n_1,n_2,N_{1,batch},N_{2,batch},N)$:

$$P(S|N_1(S)=n_1,N_2(S)=n_2)=F(n_1,n_2,N_{1,batch},N_{2,batch},N)P(n_1,n_2,p_1',p_2',N_{batch}) \quad (4.12)$$

For the current special situation (number-based approach and T=2), the total number of possible end-points of the binomial tree is 2^N. The total number of possible end-points resulting in a sample with n_1 particles of type 1 and n_2 particles of type 2 is the number of possible permutations of N objects, N!, divided by the number of permutations of n_1 objects, $n_1!$, and the number of permutations of n_2 objects, $n_2!$. This leads to the following expression for $F(n_1,n_2,N_{1,batch},N_{2,batch},N)$:

$$F(n_1,n_2,N_{1,batch},N_{2,batch},N)=\frac{N\,!}{n_1!n_2!} \quad (4.13)$$

provided that $n_1+n_2=N$ and $n_1\leq N_{1,batch}$ and $n_2\leq N_{2,batch}$. If one of these conditions is not met, the value of F is set to zero. Substituting Equation 4.11 and 4.13 into Equation 4.12 gives the following expression:

$$P(S|N_1(S)=n_1,N_2(S)=n_2)= \quad (4.14)$$

$$\frac{N!}{n_1!n_2!}\left(\prod_{k=0}^{N-1}N_{batch}/(N_{batch}-k)\right)\left[\prod_{i=0}^{n_1-1}(p_1'-i/N_{batch})\right]\left[\prod_{i=0}^{n_2-1}(p_2'-i/N_{batch})\right]$$

provided that $n_1+n_2=N$ and $n_1\leq N_{1,batch}$ and $n_2\leq N_{2,batch}$. The following identities can be used for substitution in the above equation:

$$\prod_{k=0}^{N-1}N_{batch}/(N_{batch}-k)=\frac{N_{batch}^N(N_{batch}-N)!}{N_{batch}!} \quad (4.15)$$

$$\prod_{i=0}^{n_1-1}\left(p_1' - i/N_{\text{batch}}\right) = \prod_{i=0}^{n_1-1}\left(\frac{N_{1,\text{batch}}-i}{N_{\text{batch}}}\right) = \frac{N_{1,\text{batch}}!}{N_{\text{batch}}^{n_1}\left(N_{1,\text{batch}}-n_1\right)!} \tag{4.16}$$

$$\prod_{i=0}^{n_2-1}\left(p_2' - i/N_{\text{batch}}\right) = \prod_{i=0}^{n_2-1}\left(\frac{N_{2,\text{batch}}-i}{N_{\text{batch}}}\right) = \frac{N_{2,\text{batch}}!}{N_{\text{batch}}^{n_2}\left(N_{2,\text{batch}}-n_2\right)!} \tag{4.17}$$

Substituting Equations 4.15, 4.16 and 4.17 into Equation 4.14 yields:

$$\left.\begin{aligned}
&P\left(S\middle|N_1(S)=n_1, N_2(S)=n_2\right) = \frac{N!}{n_1!\,n_2!}\,\frac{N_{\text{batch}}^{N}\left(N_{\text{batch}}-N\right)!}{N_{\text{batch}}!}\times \\
&\frac{N_{1,\text{batch}}!}{N_{\text{batch}}^{n_1}\left(N_{1,\text{batch}}-n_1\right)!\,N_{\text{batch}}^{n_2}\left(N_{2,\text{batch}}-n_2\right)!}N_{2,\text{batch}}! = \\
&\frac{N!\left(N_{\text{batch}}-N\right)!}{N_{\text{batch}}!}\,\frac{N_{1,\text{batch}}!}{n_1!\left(N_{1,\text{batch}}-n_1\right)!}\,\frac{N_{2,\text{batch}}!}{n_2!\left(N_{2,\text{batch}}-n_2\right)!}
\end{aligned}\right\} \tag{4.18}$$

provided that $n_1+n_2=N$ and $n_1 \leq N_{1,\text{batch}}$ and $n_2 \leq N_{2,\text{batch}}$. The final obtained expression for $P(S \mid N_1(S)=n_1, N_2(S)=n_2)$:

$$P\left(S\middle|N_1(S)=n_1, N_2(S)=n_2\right) = \frac{N!\left(N_{\text{batch}}-N\right)!}{N_{\text{batch}}!}\,\frac{N_{1,\text{batch}}!}{n_1!\left(N_{1,\text{batch}}-n_1\right)!}\,\frac{N_{2,\text{batch}}!}{n_2!\left(N_{2,\text{batch}}-n_2\right)!} \tag{4.19}$$

is equivalent to the hypergeometric distribution (see Feller, 1968). In the limit of $N_{\text{batch}}=\infty$ (see Paragraph 3.4), Equation 4.14 becomes:

$$\lim_{N_{\text{batch}}\to\infty} P\left(S\middle|N_1(S)=n_1, N_2(S)=n_2\right) = \frac{N!}{n_1!\,n_2!}\,p_1'^{\,n_1}\,p_2'^{\,n_2} \tag{4.20}$$

provided that $n_1+n_2=N$ and $n_1 \leq N_{1,\text{batch}}$ and $n_2 \leq N_{2,\text{batch}}$. The latter two conditions are automatically satisfied in the limit of $N_{\text{batch}}=\infty$. Equation 4.20 is equivalent to the binomial distribution (see Feller, 1968). The cumulative probability to draw a sample with any possible value for N_1 and N_2 (satisfying $N_1+N_2=N$) is one:

$$\left.\begin{aligned} &\lim_{N_{batch}\to\infty}\sum_{n_1=0}^{N}\sum_{n_2=0}^{N}P\left(S|N_1(S)=n_1,N_2(S)=n_2\right)= \\ &\lim_{N_{batch}\to\infty}\sum_{n_1=0}^{N}P\left(S|N_1(S)=n_1,N_2(S)=N-n_1\right)= \\ &\sum_{n_1=0}^{N}\frac{N!}{n_1!\left(N-n_1\right)!}p_1'^{n_1}p_2'^{N-n_1}=\left(p_1'+p_2'\right)^N=1^N=1 \end{aligned}\right\} \qquad (4.21)$$

in which the first equality follows from the fact that P(S | N₁(S)=n₁,N₂(S)=n₂) is zero if $n_1+n_2 \neq N$. For the third equality, the binomial expansion theorem was used.

A method of deriving expected values and covariance of N_1 and N_2 in the limit of $N_{batch}=\infty$ is to differentiate Equation 4.21 with respect to p_1'. In general, it is only correct to differentiate an equation with respect to a variable on the left-hand side and on the right-hand side, if both sides are equal to each other for all possible values of the variable. This is not the case for Equation 4.21, because the cumulative probability is only one if $p_1'+p_2'=1$. Therefore, the relation $p_2'=1-p_1'$ is substituted into Equation 4.21:

$$\sum_{n_1=0}^{N}\frac{N!}{n_1!\left(N-n_1\right)!}p_1'^{n_1}\left(1-p_1'\right)^{N-n_1}=1 \qquad (4.22)$$

The left-hand side and the right-hand side of the above equation are differentiated with respect to p_1':

$$\sum_{n_1=0}^{N}\left(\frac{n_1}{p_1'}-\frac{N-n_1}{1-p_1'}\right)\frac{N!}{n_1!\left(N-n_1\right)!}p_1'^{n_1}\left(1-p_1'\right)^{N-n_1}=0 \qquad (4.23)$$

Substituting $p_2'=1-p_1'$ and Equation 4.20 yields:

$$\lim_{N_{batch}\to\infty}\sum_{n_1=0}^{N}\left(\frac{n_1}{p_1'}-\frac{N-n_1}{p_2'}\right)P\left(S|N_1(S)=n_1,N_2(S)=N-n_1\right)=0 \qquad (4.24)$$

The definition of P(S | N₁(S)=n₁,N₂(S)=n₂) can be substituted back into the above equation:

$$\lim_{N_{batch}\to\infty} \sum_{n_1=0}^{N} \left[\left(\frac{n_1}{p'_1} - \frac{N-n_1}{p'_2} \right) F\left(n_1, N-n_1, N_{1,batch}, N_{2,batch}, N\right) P\left(n_1, N-n_1, p'_1, p'_2, N_{batch}\right) \right] = 0 \quad \textbf{(4.25)}$$

Because $F(n_1,n_2,N_{1,batch},N_{2,batch},N)$ is zero if $n_1+n_2 \neq N$ the summation over n_1 may be replaced by a double summation over n_1 and n_2 if the second argument of F in Equation 4.25, $N-n_1$, is replaced by n_2:

$$\lim_{N_{batch}\to\infty} \sum_{n_1=0}^{N} \sum_{n_2=0}^{N} \left[\left(\frac{n_1}{p'_1} - \frac{N-n_1}{p'_2} \right) F\left(n_1, n_2, N_{1,batch}, N_{2,batch}, N\right) P\left(n_1, n_2, p'_1, p'_2, N_{batch}\right) \right] = 0 \quad \textbf{(4.26)}$$

From Equation 3.14 follows that the above equation is equivalent to the definition of an expected value. Therefore, the above equation results in:

$$\lim_{N_{batch}\to\infty} E\left(\frac{N_1}{p'_1} - \frac{N-N_1}{p'_2} \right) = 0 \quad \textbf{(4.27)}$$

Using $p'_2 = 1 - p'_1$ and rewriting the above equation gives:

$$\lim_{N_{batch}\to\infty} E\left(N_1\right) = Np'_1 \quad \textbf{(4.28)}$$

With a similar derivation, it can be proven that:

$$\lim_{N_{batch}\to\infty} E\left(N_2\right) = Np'_2 \quad \textbf{(4.29)}$$

The same method as above can be used for the calculation of the covariance, $E(N_1N_2)-E(N_1)E(N_2)$. From the definition of expected value and Equation 4.28 follows:

$$\lim_{N_{batch}\to\infty} \sum_{S\in U} N_1(S)P(S) = Np'_1 \quad \textbf{(4.30)}$$

Applying Equation 3.14, the above equation is written as a double summation:

$$\lim_{N_{batch}\to\infty} \sum_{n_1=0}^{N} \sum_{n_2=0}^{N} n_1 P\left(S \middle| N_1(S) = n_1, N_2(S) = n_2\right) = Np'_1 \quad \textbf{(4.31)}$$

Because $P(S \mid N_1(S)=n_1, N_2(S)=n_2)$ is zero if $n_1+n_2 \neq N$ the double summation can be replaced by a single summation if in the summand n_2 is replaced by $N-n_1$. This yields:

$$\lim_{N_{batch}\to\infty} \sum_{n_1=0}^{N} n_1 P\left(S \middle| N_1(S)=n_1, N_2(S)=N-n_1\right) = Np'_1 \tag{4.32}$$

Application of Equation 4.20 results in:

$$\sum_{n_1=0}^{N} n_1 \frac{N!}{n_1!\left(N-n_1\right)!} p_1'^{\,n_1} p_2'^{\,N-n_1} = Np'_1 \tag{4.33}$$

The above equation is valid for all values of p'_1 if $p'_2 = 1-p'_1$. Therefore, in order to differentiate both sides of the equation with respect to p'_1 , the equality $p'_2 = 1-p'_1$ is substituted. This yields after differentiation with respect to p'_1 :

$$\sum_{n_1=0}^{N} n_1 \left(\frac{n_1}{p'_1} - \frac{N-n_1}{1-p'_1}\right) \frac{N!}{n_1!\left(N-n_1\right)!} p_1'^{\,n_1} \left(1-p'_1\right)^{N-n_1} = N \tag{4.34}$$

Substituting $1-p'_1 = p'_2$ and Equation 4.20 yields:

$$\lim_{N_{batch}\to\infty} \sum_{n_1=0}^{N} n_1 \left(\frac{n_1}{p'_1} - \frac{N-n_1}{p'_2}\right) P\left(S \middle| N_1(S)=n_1, N_2(S)=N-n_1\right) = N \tag{4.35}$$

The definition of P(S | N$_1$(S)=n$_1$,N$_2$(S)=n$_2$) can be substituted into the above equation:

$$\lim_{N_{batch}\to\infty} \sum_{n_1=0}^{N} \left[n_1 \left(\frac{n_1}{p'_1} - \frac{N-n_1}{p'_2}\right) F\left(n_1, N-n_1, N_{1,batch}, N_{2,batch}, N\right) P\left(n_1, N-n_1, p'_1, p'_2, N_{batch}\right)\right] = N \tag{4.36}$$

Because F(n$_1$,n$_2$,N$_{1,batch}$,N$_{2,batch}$,N) is zero if n$_1$+n$_2$≠N the summation over n$_1$ may be replaced by a double summation over n$_1$ and n$_2$ if in Equation 4.36 the second argument of F, N−n$_1$, is replaced by n$_2$:

$$\lim_{N_{batch}\to\infty} \sum_{n_1=0}^{N} \sum_{n_2=0}^{N} \left[n_1 \left(\frac{n_1}{p'_1} - \frac{N-n_1}{p'_2}\right) F\left(n_1, n_2, N_{1,batch}, N_{2,batch}, N\right) P\left(n_1, N-n_1, p'_1, p'_2, N_{batch}\right)\right] = N \tag{4.37}$$

Because F(n$_1$,n$_2$,N$_{1,batch}$,N$_{2,batch}$,N) is zero if n$_1$+n$_2$≠N the second argument of P in the above equation, N−n$_1$, may be replaced by n$_2$. Applying equation 3.14, the above equation results in:

$$\lim_{N_{batch}\to\infty} E\left(N_1\left(\frac{N_1}{p_1'}-\frac{N-N_1}{p_2'}\right)\right)=N \qquad (4.38)$$

This is equivalent to:

$$\lim_{N_{batch}\to\infty} E\left(N_1\left(\frac{N-N_2}{p_1'}-\frac{N_2}{p_2'}\right)\right)=N \qquad (4.39)$$

The terms can be rearranged and Equation 4.28 and Equation 4.29 can be substituted into Equation 4.39. This yields:

$$\lim_{N_{batch}\to\infty} \left(E\left(N_1N_2\right)-E\left(N_1\right)E\left(N_2\right)\right)=-Np_1'\,p_2' \qquad (4.40)$$

In the next paragraph, a similar method is used to calculate the expected values and covariances in the size-based approach with T classes.

4.5 Generalised method

As discussed in Chapter 1, a sampling theory is required which models the sample drawing as a process leading to samples of (approximately) fixed mass. In addition, the theory must be applicable to batches containing an arbitrary number of distinct particle classes. The situation considered in the previous paragraph (number-based approach and T=2) does not meet these criteria. Therefore, in this paragraph, a method is applied for calculation of the covariances, $E(N_nN_r)-E(N_n)E(N_r)$, for arbitrary n and r between 1 and T, for the general case of sampling a batch with T classes in the size-based approach in the limit of $N_{batch}=\infty$. The method is basically similar to the method that was applied in Paragraph 4.4 and comprises the following steps:

Step 1

- It is demonstrated that the probability to draw an arbitrary sample, which is one when the evolution of p_i is given by Equation 3.3 with $p_i' = N_{i,batch}/N_{batch}$, remains one for arbitrary values of p_i' (satisfying only $\sum_{i=1}^{T} p_i' = 1$).

Step 2

- An expression for the probability of drawing an arbitrary sample with n_i particles of type i, for i ranging between 1 and T, is derived, using the expression for the probability of going to an end-point of the multinomial tree, $P\left(n_1,\ldots,n_T,p_1',\ldots p_T',N_{batch}\right)$, given by Equation 3.9.

Step 3

- The summation over all possible samples in the definition of the cumulative probability is replaced by a T-fold summation over all possible values of $n_1,\dots,n_T$.

Step 4

- In the limit of $N_{batch}=\infty$, the cumulative probability is differentiated with respect to p'_n, for arbitrary n between 1 and T, to obtain an expression for the expected value of the number of particles belonging to the nth class in the sample.

Step 5

- The expression for the expected value is also differentiated with respect to p'_n, for arbitrary n between 1 and T, to obtain an additional equation.

Step 6

- The equations obtained are rewritten, yielding an equation for the covariance $E(N_nN_r)-E(N_n)E(N_r)$, for arbitrary n and r between 1 and T in the limit of $N_{batch}=\infty$.

Step 1. In the size-based approach, the probability of drawing a sample with arbitrary composition is one when the probability of choosing an arbitrary branch of the T possible branches at any node of the multinomial tree is one. Using Equation 3.3 results in:

$$\sum_{i=1}^{T} p_i = \sum_{i=1}^{T}\left(p'_i - \frac{n_i}{N_{batch}}\right)\frac{N_{batch}}{N_{batch}-n_{sample}} = \left(\frac{N_{batch}\sum\limits_{i=1}^{T} p'_i - n_{sample}}{N_{batch}}\right)\frac{N_{batch}}{N_{batch}-n_{sample}} \qquad \textbf{(4.41)}$$

Because

$$p'_i = N_{i,batch}/N_{batch} \qquad \textbf{(4.42)}$$

for all i between 1 and T, the cumulative of p'_i is one:

$$\sum_{i=1}^{T} p'_i = 1 \qquad \textbf{(4.43)}$$

Substituting this result in Equation 4.41 gives:

$$\sum_{i=1}^{T} p_i = \left(\frac{N_{batch}-n_{sample}}{N_{batch}}\right)\frac{N_{batch}}{N_{batch}-n_{sample}} = 1 \qquad \textbf{(4.44)}$$

Hence, the probability of choosing an arbitrary branch of the T possible branches at any node of the multinomial tree is one. Consequently, the cumulative probability of drawing any arbitrary sample is one. The general situation will be considered where the probability p_i is given by Equation 3.3, with arbitrary p'_i, satisfying only $\sum_{i=1}^{T} p'_i = 1$. Hence, the following derivations are also valid when Equation 4.42 does not hold. Because $\sum_{i=1}^{T} p'_i = 1$, Equation 4.44 remains valid, so the probability of drawing an arbitrary sample remains one. During Step 4, the necessity of this generalisation will be discussed.

Step 2. As defined in Chapter 3, the probability of going to an end-point of the multinomial tree resulting in a sample with n_j particles of class j, for all j between 1 and T, is denoted as $P(n_1,\ldots,n_T,p'_1,\ldots,p'_T,N_{batch})$. Equation 3.9 states that:

$$P(n_1,\ldots,n_T,p'_1,\ldots,p'_T,N_{batch}) = \left(\prod_{k=0}^{n_{sample}-1} N_{batch} \Big/ (N_{batch} - k) \right) \prod_{j=1}^{T} \left[\prod_{i=0}^{n_j - 1} \left(p'_j - i/N_{batch} \right) \right] \quad (4.45)$$

where T is the number of classes and $n_{sample} = \sum_{i=1}^{T} n_i$. In the limit of an infinite value of N_{batch} the above equation becomes:

$$\lim_{N_{batch} \to \infty} P(n_1,\ldots,n_T,p'_1,\ldots,p'_T,N_{batch}) = \prod_{j=1}^{T} \left[\prod_{i=0}^{n_j - 1} p'_j \right] = \prod_{j=1}^{T} {p'_j}^{n_j} \quad (4.46)$$

The probability of drawing an arbitrary sample with n_j particles of class j, for j ranging between 1 and T, is the product of $P(n_1,\ldots,n_T,p'_1,\ldots,p'_T,N_{batch})$ and the number of possible end-points resulting in a sample with n_j particles of class j, for all j between 1 and T. The number of possible end-points resulting in a sample with n_j particles of class j, for all j between 1 and T depends on the number of particles in the different classes in both sample and batch (*i.e.* the variables n_1, ..., n_T, $N_{1,batch}$, ..., $N_{T,batch}$) and the boundary value of the sample size Z (which can represent the boundary value of the sample mass M, volume V or number of particles N). Therefore, the number of possible end-points resulting in a sample with n_j particles of class j, for all j between 1 and T, is denoted as $F(n_1,\ldots,n_T,N_{1,batch},\ldots,N_{T,batch},Z)$. In the number-based approach, F is equal to $N! \big/ \prod_{i=1}^{T} n_i!$ provided that $\sum_{i=1}^{T} n_i = N$ and $n_i \leq N_{i,batch}$ for all i between 1 and T. However, in the mass- or volume-based approach, the relation is not simple. In the following derivations, no explicit knowledge of $F(n_1,\ldots,n_T,N_{1,batch},\ldots,N_{T,batch},Z)$ is needed. The value of $F(n_1,\ldots,n_T,N_{1,batch},\ldots,N_{T,batch},Z)$ is zero for all combinations of n_1, ..., n_T that result in a too low or too high sample size (*i.e.* do not correspond to an end-point of the multinomial tree).

Denoting the probability of drawing an arbitrary sample with n_j particles of class j, for j ranging between 1 and T, as $P(S|N_1(S)=n_1,...,N_T(S)=n_T)$, results in:

$$P(S|N_1(S)=n_1,...,N_T(S)=n_T)=F(n_1,...n_T,N_{1,batch},...,N_{T,batch},Z)P(n_1,...n_T,p'_1,...,p'_T,N_{batch}) \quad \textbf{(4.47)}$$

Step 3. Any end-point resulting in a sample with n_j particles of class j, for j ranging between 1 and T, contributes with $P(n_1,...,n_T,p'_1,...,p'_T,N_{batch})$ to the cumulative probability, $\sum_{S\in U} P(S)$, which is one. Therefore, the expression for the cumulative probability can be written as a T-fold summation[10]:

$$\sum_{S\in U} P(S)=\left(\prod_{s=1}^{T}\sum_{n_s=0}^{N_{s,batch}}\right)\left[F(n_1,...,n_T,N_{1,batch},...,N_{T,batch},Z)P(n_1,...,n_T,p'_1,...,p'_T,N_{batch})\right]=1 \quad \textbf{(4.48)}$$

Taking the limit of $N_{batch}=\infty$ on both sides of the above equation and substituting Equation 4.46 yields:

$$\left(\prod_{s=1}^{T}\sum_{n_s=0}^{N_{s,batch}}\right)\left(F(n_1,...,n_T,N_{1,batch},...,N_{T,batch},Z)\prod_{j=1}^{T}p_j'^{\,n_j}\right)=1 \quad \textbf{(4.49)}$$

Step 4. In the derivation of an expression for the covariances in the limit of $N_{batch}=\infty$, the fourth step is differentiation to p'_n of the above equation. The above result was derived under the constraint $\sum_{i=1}^{T}p'_i=1$. Therefore, the result is not valid for all values of p'_i , where i ranges from 1 to T. Differentiation on both sides of the equation with respect to p'_n , for arbitrary n between 1 and T, is only correct if both sides are equal for all possible values of the variable. This problem is solved by introducing the arbitrary index r, which can represent any integer value between 1 and T, and the index t, ranging from 1 to T, and substituting

$$p'_r=1-\sum_{\substack{t=1\\t\neq r}}^{T}p'_t \quad \textbf{(4.50)}$$

[10] in which the symbolic notation $\sum_{n_1=0}^{N_{1,batch}}\sum_{n_2=0}^{N_{2,batch}}\cdots\sum_{n_T=0}^{N_{T,batch}}=\left(\prod_{s=1}^{T}\sum_{n_s=0}^{N_{s,batch}}\right)$ is used.

This gives:

$$\left(\prod_{s=1}^{T}\sum_{n_s=0}^{N_{s,batch}}\right)\left[F\left(n_1,\dots,n_T,N_{1,batch},\dots,N_{T,batch},Z\right)\left(\prod_{\substack{j=1\\j\neq r}}^{T}p_j'^{\,n_j}\right)\left(1-\sum_{\substack{t=1\\t\neq r}}^{T}p_t'\right)^{n_r}\right]=1 \qquad \textbf{(4.51)}$$

The above equation is valid for all values of p_j', where j can represent any integer between 1 and T.

For arbitrary class n, with n≠r, the expected value of N_n is obtained by taking the derivative with respect to p_n' on both sides of Equation 4.51:

$$\frac{d}{dp_n'}\left(\prod_{s=1}^{T}\sum_{n_s=0}^{N_{s,batch}}\right)\left[F\left(n_1,\dots,n_T,N_{1,batch},\dots,N_{T,batch},Z\right)\left(\prod_{\substack{j=1\\j\neq r}}^{T}p_j'^{\,n_j}\right)\left(1-\sum_{\substack{t=1\\t\neq r}}^{T}p_t'\right)^{n_r}\right]=\frac{d}{dp_n'}1=0 \qquad \textbf{(4.52)}$$

where the arbitrary index r may not be equal to n. Here, the importance of considering the general situation with arbitrary p_i' (satisfying $\sum_{i=1}^{T}p_i'=1$) not necessarily given by Equation 4.42 becomes clear: the zero on the right-hand side of the above equation appears after differentiation, because the cumulative probability is one for arbitrary p_i' (satisfying $\sum_{i=1}^{T}p_i'=1$). If the cumulative probability were only one for p_i' given by Equation 4.42, the above equation would not necessarily be valid. In order to perform the derivation with respect to p_n', summations and differentiation are interchanged. This results in:

$$\left(\prod_{s=1}^{T}\sum_{n_s=0}^{N_{s,batch}}\right)\left[\left(\frac{n_n}{p_n'}-\frac{n_r}{1-\sum_{\substack{t=1\\t\neq r}}^{T}p_t'}\right)F\left(n_1,\dots,n_T,N_{1,batch},\dots,N_{T,batch},Z\right)\left(\prod_{\substack{j=1\\j\neq r}}^{T}p_j'^{\,n_j}\right)\left(1-\sum_{\substack{t=1\\t\neq r}}^{T}p_t'\right)^{n_r}\right]=0 \qquad \textbf{(4.53)}$$

Substitution of $1-\sum_{\substack{t=1\\t\neq r}}^{T}p_t'=p_r'$ results in a simpler equation:

$$\left(\prod_{s=1}^{T}\sum_{n_s=0}^{N_{s,\mathrm{batch}}}\right)\left[\left(\frac{n_n}{p'_n}-\frac{n_r}{p'_r}\right)F\left(n_1,\ldots,n_T,N_{1,\mathrm{batch}},\ldots,N_{T,\mathrm{batch}},Z\right)\left(\prod_{j=1}^{T}p_j'^{\,n_j}\right)\right]=0 \qquad (4.54)$$

In the following developments, the above equation will be transformed into an equation for expected values. First, Equation 4.46 is substituted:

$$\lim_{N_{\mathrm{batch}}\to\infty}\left(\prod_{s=1}^{T}\sum_{n_s=0}^{N_{s,\mathrm{batch}}}\right)\left[\left(\frac{n_n}{p'_n}-\frac{n_r}{p'_r}\right)F\left(n_1,\ldots,n_T,N_{1,\mathrm{batch}},\ldots,N_{T,\mathrm{batch}},Z\right)\right.$$

$$\left.P\left(n_1,\ldots,n_T,p'_1,\ldots,p'_T,N_{\mathrm{batch}}\right)\right]=0 \qquad (4.55)$$

According to Equation 3.14, the above equation is equivalent to the following expected value:

$$\lim_{N_{\mathrm{batch}}\to\infty} E\left(\frac{N_n}{p'_n}-\frac{N_r}{p'_r}\right)=0 \qquad (4.56)$$

Originally the arbitrary indices n and r were chosen unequal. However, the above equation is also valid for n=r, yielding the trivial result 0=0. Equation 4.56 can also be written as:

$$\lim_{N_{\mathrm{batch}}\to\infty}\frac{E(N_r)}{E(N_n)}=\frac{p'_r}{p'_n}=\frac{N_{r,\mathrm{batch}}}{N_{n,\mathrm{batch}}} \qquad (4.57)$$

This implies that in the limit of $N_{\mathrm{batch}}=\infty$ the ratio of the expected values of N_r and N_n is equal to the corresponding ratio of $N_{r,\mathrm{batch}}$ and $N_{n,\mathrm{batch}}$. The above equation can also be written as:

$$\lim_{N_{\mathrm{batch}}\to\infty} E(N_r)=\lim_{N_{\mathrm{batch}}\to\infty} E(N_n)\frac{p'_r}{p'_n} \qquad (4.58)$$

In the derivation of Equation 4.56 no explicit knowledge of the form of $F(n_1,\ldots,n_T,N_{1,\mathrm{batch}},\ldots,N_{T,\mathrm{batch}},Z)$ was used. On the other hand, $F(n_1,\ldots,n_T,N_{1,\mathrm{batch}},\ldots,N_{T,\mathrm{batch}},Z)$ is determined by the structure of the multinomial tree, *i.e.* the location of the end-points. Therefore, it can be concluded that the essential property of the size-based approach, namely that the difference δ between the boundary value of the sample size and the obtained sample size is relatively small, is not a requirement for Equation 4.56 (and hence also for Equation 4.57 and 4.58). However, for the derivation of an expression for the expected value (note: not an equation for a ratio of two expected values) the fact that the parameter δ is relatively small is essential. To demonstrate this, an equation for the precise definition of δ is given:

$$\delta = Z - \sum_{r=1}^{T} z_r N_r \tag{4.59}$$

where z_r is the size, *i.e.* mass or volume of a particle of type r or one. Taking the expected value in the limit of $N_{batch}=\infty$ on both sides of the expression for δ and substituting Equation 4.58 yields:

$$\lim_{N_{batch}\to\infty} E(\delta) = Z - \lim_{N_{batch}\to\infty} \sum_{r=1}^{T} z_r E(N_n)\frac{p'_r}{p'_n} \tag{4.60}$$

This can be written as an equation for the expected value of N_n:

$$\lim_{N_{batch}\to\infty} E(N_n) = \lim_{N_{batch}\to\infty} p'_n\,(Z - E(\delta)) \Big/ \sum_{r=1}^{T} p'_r\, z_r \tag{4.61}$$

The denominator in the above equation is the mean particle size $\bar{z}$ defined as:

$$\bar{z} \equiv \sum_{r=1}^{T} p'_r\, z_r \tag{4.62}$$

Because $-z_{max}<\delta\leq 0$ and thus also $-z_{max}<E(\delta)\leq 0$, E(δ) is often negligible compared with Z. When E(δ) is neglected, the expected value of the number of particles of type n becomes:

$$\lim_{N_{batch}\to\infty} E(N_n) \approx p'_n\, Z/\bar{z} \tag{4.63}$$

Similarly to the above derivation, one can find an expression for the variance $V(N_n) = E(N_n^2) - E^2(N_n)$ and covariance $E(N_nN_r)-E(N_n)E(N_r)$.

Step 5. Substituting $Z-E(\delta)=E(Z_{sample})$ into Equation 4.61 and using Equation 3.14 for evaluation of the expected values in Equation 4.61 results in:

$$\left.\begin{array}{l}\lim\limits_{N_{\mathrm{batch}}\to\infty}\left(\prod\limits_{s=1}^{T}\sum\limits_{n_s=0}^{N_{s,\mathrm{batch}}}\right)\left[\; n_n F\left(n_1,\ldots,n_T,N_{1,\mathrm{batch}},\ldots,N_{T,\mathrm{batch}},Z\right)\times\right.\\ \left.P\left(n_1,\ldots,n_T,p'_1,\ldots,p'_T,N_{\mathrm{batch}}\right)\right]=\\ \lim\limits_{N_{\mathrm{batch}}\to\infty}\left(\prod\limits_{s=1}^{T}\sum\limits_{n_s=0}^{N_{s,\mathrm{batch}}}\right)\left[\left(\sum\limits_{i=1}^{T} z_i\, n_i\right)F\left(n_1,\ldots,n_T,N_{1,\mathrm{batch}},\ldots,N_{T,\mathrm{batch}},Z\right)\times\right.\\ \left.P\left(n_1,\ldots,n_T,p'_1,\ldots,p'_T,N_{\mathrm{batch}}\right)\right]\dfrac{p'_n}{\sum\limits_{r=1}^{T} p'_r\, z_r}\end{array}\right\}\qquad(4.64)$$

From Equation 4.46 and the above equation it follows that:

$$\begin{array}{l}\left(\prod\limits_{s=1}^{T}\sum\limits_{n_s=0}^{N_{s,\mathrm{batch}}}\right)\left[\; n_n F\left(n_1,\ldots,n_T,N_{1,\mathrm{batch}},\ldots,N_{T,\mathrm{batch}},Z\right)\prod\limits_{j=1}^{T} {p'_j}^{n_j}\;\right]=\\ \dfrac{p'_n}{\sum\limits_{r=1}^{T} p'_r\, z_r}\left(\prod\limits_{s=1}^{T}\sum\limits_{n_s=0}^{N_{s,\mathrm{batch}}}\right)\left[\left(\sum\limits_{i=1}^{T} z_i\, n_i\right)F\left(n_1,\ldots,n_T,N_{1,\mathrm{batch}},\ldots,N_{T,\mathrm{batch}},Z\right)\prod\limits_{j=1}^{T} {p'_j}^{n_j}\right]\end{array}\qquad(4.65)$$

In order to differentiate both right-hand side and left-hand side of the above equation with respect to p'_n, for arbitrary r (r≠n), p'_r is replaced by $1-\sum\limits_{\substack{t=1\\ t\neq r}}^{T} p'_t$. This yields after differentiation with respect to p'_n and substituting $1-\sum\limits_{\substack{t=1\\ t\neq r}}^{T} p'_t = p'_r$ back:

$$\left.\begin{aligned}
&\left(\prod_{s=1}^{T}\sum_{n_s=0}^{N_{s,batch}}\right)\left[\left(\frac{n_n^2}{p'_n}-\frac{n_n n_r}{p'_r}\right)F\left(n_1,\ldots,n_T,N_{1,batch},\ldots,N_{T,batch},Z\right)\prod_{j=1}^{T}p_j'^{\,n_j}\right]\\
&=\frac{\sum_{r=1}^{T}p'_r\,z_r+p'_n\left(z_r-z_n\right)}{\left(\sum_{r=1}^{T}p'_r\,z_r\right)^2}\left(\prod_{s=1}^{T}\sum_{n_s=0}^{N_{s,batch}}\right)\left[\left(\sum_{i=1}^{T}z_i\,n_i\right)\times\right.\\
&\left.F\left(n_1,\ldots,n_T,N_{1,batch},\ldots,N_{T,batch},Z\right)\prod_{j=1}^{T}p_j'^{\,n_j}\right]+\frac{p'_n}{\sum_{r=1}^{T}p'_r\,z_r}\left(\prod_{s=1}^{T}\sum_{n_s=0}^{N_{s,batch}}\right)\\
&\left[\left(\sum_{i=1}^{T}z_i\,n_i\right)\left(\frac{n_n}{p'_n}-\frac{n_r}{p'_r}\right)F\left(n_1,\ldots,n_T,N_{1,batch},\ldots,N_{T,batch},Z\right)\prod_{j=1}^{T}p_j'^{\,n_j}\right]
\end{aligned}\right\}\qquad\textbf{(4.66)}$$

When Equation 4.46 is substituted back into the above equation, the definition of mean particle size, $\bar{z}$, is used and Equation 3.14 for the expected value is used to replace the above T-fold summations, the above equation becomes:

$$\begin{aligned}
\lim_{N_{batch}\to\infty} E\left(N_n\left(\frac{N_n}{p'_n}-\frac{N_r}{p'_r}\right)\right)&=\lim_{N_{batch}\to\infty}\frac{\bar{z}+p'_n\left(z_r-z_n\right)}{\bar{z}^2}E\left(\sum_{i=1}^{T}z_i\,N_i\right)\\
&+\lim_{N_{batch}\to\infty}\frac{p'_n}{\bar{z}}E\left(\left(\sum_{i=1}^{T}z_i\,N_i\right)\left(\frac{N_n}{p'_n}-\frac{N_r}{p'_r}\right)\right)
\end{aligned}\qquad\textbf{(4.67)}$$

Substituting $\sum_{i=1}^{T}z_i\,N_i=Z-\delta$, results in the following useful equation for the derivation of variance and covariance:

$$\lim_{N_{batch}\to\infty}E\left(\frac{N_n^2}{p'_n}-\frac{N_nN_r}{p'_r}\right)=\lim_{N_{batch}\to\infty}\left(\frac{Z-E(\delta)}{\bar{z}}-\frac{p'_n(Z-E(\delta))(z_n-z_r)}{\bar{z}^2}-\frac{p'_nE\left(\left[\frac{N_n}{p'_n}-\frac{N_r}{p'_r}\right]\delta\right)}{\bar{z}}\right)\qquad\textbf{(4.68)}$$

As a result of the original assumption that n is unequal to r, the above equation is not valid for n=r. Therefore, the above equation is transformed into an equation that is valid for all values of n and r between 1 and T by introduction of the Kronecker delta, Δ_{nr}. The parameter Δ_{nr} is one when n=r and zero otherwise, where n and r range between 1 and T. The above equation is transformed into the following result that is valid for all values of n and r:

$$\lim_{N_{batch}\to\infty} E\left(\frac{N_n^2}{p'_n}-\frac{N_n N_r}{p'_r}\right)=$$

$$\lim_{N_{batch}\to\infty}\left(\frac{Z-E(\delta)}{\bar{z}}\left(1-\Delta_{nr}\right)-\frac{p'_n\,(Z-E(\delta))(z_n-z_r)}{\bar{z}^2}-\frac{p'_n\,E\left(\left[\frac{N_n}{p'_n}-\frac{N_r}{p'_r}\right]\delta\right)}{\bar{z}}\right) \qquad \textbf{(4.69)}$$

yielding the trivial result 0=0 for n=r. In other words, this gives the correct result for a previously omitted situation.

Step 6. Both the right-hand side and the left-hand side of Equation 4.69 can be multiplied by $p'_r\, z_r$ and summed over all r between 1 and T. For terms that do not depend on r, this operation is equivalent to a multiplication by $\bar{z}$. The result is:

$$\lim_{N_{batch}\to\infty}\left(\bar{z}E\left(\frac{N_n^2}{p'_n}\right)-E\left(N_n\sum_{r=1}^{T} z_r N_r\right)\right)=\lim_{N_{batch}\to\infty}\left((Z-E(\delta))\left(1-\frac{p'_r\, z_r}{\bar{z}}\right)-\frac{p'_n\, z_n (Z-E(\delta))}{\bar{z}}\right)$$

$$+\lim_{N_{batch}\to\infty}\left(\sum_{r=1}^{T} p'_r\, z_r^2\,\frac{p'_n\,(Z-E(\delta))}{\bar{z}^2}-E\left(N_n\delta\right)+\frac{p'_n\,E\left(\sum_{r=1}^{T} z_r N_r \delta\right)}{\bar{z}}\right) \qquad \textbf{(4.70)}$$

The definition of δ is used to eliminate $\sum_{r=1}^{T} z_r N_r$ by substitution of $\sum_{r=1}^{T} z_r N_r = Z-\delta$.

After multiplication on both sides by $\frac{p'_n}{\bar{z}}$ and replacing $\frac{p'_n\,(Z-E(\delta))}{\bar{z}}$ by $E(N_n)$ one obtains:

$$\left.\begin{aligned}
&\lim_{N_{batch}\to\infty}\left(E\left(N_n^2\right)-E^2\left(N_n\right)-\frac{p'_n}{\bar{z}}E(\delta)E\left(N_n\right)+\frac{p'_n}{\bar{z}}E\left(N_n\delta\right)\right)=\\
&\lim_{N_{batch}\to\infty}\left(E\left(N_n\right)\left(1-\frac{p'_r\,z_r}{\bar{z}}\right)-E\left(N_n\right)\frac{p'_n\,z_n}{\bar{z}}+E\left(N_n\right)\sum_{r=1}^{T}p'_r\,z_r^2\,\frac{p'_n}{\bar{z}^2}\right)\\
&-\lim_{N_{batch}\to\infty}\left(\frac{p'_n\,E\left(N_n\delta\right)}{\bar{z}}-\frac{\left(p'_n\right)^2E((Z-\delta)\delta)}{\bar{z}^2}\right)
\end{aligned}\right\}\qquad\textbf{(4.71)}$$

The first two terms on the left-hand side in the above equation are the variance of N_n in the limit of $N_{batch}=\infty$:

$$\left.\begin{aligned}
&\lim_{N_{batch}\to\infty}V\left(N_n\right)\equiv\lim_{N_{batch}\to\infty}\left(E\left(N_n^2\right)-E^2\left(N_n\right)\right)=\\
&\lim_{N_{batch}\to\infty}\left(\frac{p'_n}{\bar{z}}E(\delta)E\left(N_n\right)-\frac{p'_n}{\bar{z}}E\left(N_n\delta\right)\right)+\\
&\lim_{N_{batch}\to\infty}\left(E\left(N_n\right)\left(1-\frac{p'_r\,z_r}{\bar{z}}\right)-E\left(N_n\right)\frac{p'_n\,z_n}{\bar{z}}+E\left(N_n\right)\sum_{r=1}^{T}p'_r\,z_r^2\,\frac{p'_n}{\bar{z}^2}\right)-\\
&\lim_{N_{batch}\to\infty}\left(\frac{p'_n\,E\left(N_n\delta\right)}{\bar{z}}-\frac{\left(p'_n\right)^2E((Z-\delta)\delta)}{\bar{z}^2}\right)
\end{aligned}\right\}\qquad\textbf{(4.72)}$$

A much simpler equation is obtained using the following definition:

$$V\left(N_n+p'_n\,\delta/\bar{z}\right)\equiv E\left(\left(N_n+\frac{p'_n\,\delta}{\bar{z}}\right)^2\right)-E^2\left(N_n+\frac{p'_n\,\delta}{\bar{z}}\right)\qquad\textbf{(4.73)}$$

Rearranging the terms of Equation 4.72 yields:

$$\lim_{N_{batch}\to\infty}V\left(N_n+p'_n\,\delta/\bar{z}\right)=\lim_{N_{batch}\to\infty}E(N_n)\sum_{\substack{r=1\\ r\neq n}}^{T}p'_r\,z_r\left(\bar{z}-p'_n\left(z_n-z_r\right)\right)/\bar{z}^2\qquad\textbf{(4.74)}$$

With large sample masses, *i.e.* N_n is large compared to $p'_n\,\delta/\bar{z}$, the variance of $N_n+p'_n\,\delta/\bar{z}$ will be approximately equal to the variance of N_n. The approximate result in Equation 4.63 will then be valid as well. These approximations yield for the variance of N_n:

$$\lim_{N_{batch}\to\infty} V(N_n) \approx p'_n Z \sum_{\substack{r=1\\ r\neq n}}^{T} p'_r z_r \left(\bar{z} - p'_n \left(z_n - z_r\right)\right)/\bar{z}^3 \tag{4.75}$$

While this result is approximate, an exact solution is obtained by calculating covariances. In order to derive these covariances, the system consisting of Equations 4.61, 4.69 and 4.74 is transformed into a simpler system using the transformed variables N_k^* and Z^*:

$$N_k^* = N_k + \frac{p'_k \delta}{\bar{z}} \tag{4.76}$$

$$Z^* = Z - E(\delta) \tag{4.77}$$

in which k can represent any integer between 1 and T. First, the terms of Equation 4.69 are rearranged providing the following result:

$$\begin{aligned} &\lim_{N_{batch}\to\infty} E\left(\left(N_n + \frac{p'_n \delta}{\bar{z}}\right)\left(\left(N_n + \frac{p'_n \delta}{\bar{z}}\right)\Big/p'_n - \left(N_r + \frac{p'_r \delta}{\bar{z}}\right)\Big/p'_r\right)\right) \\ &= \lim_{N_{batch}\to\infty}\left(\frac{Z-E(\delta)}{\bar{z}}\left(1-\Delta_{nr}\right) - \frac{p'_n\left[Z-E(\delta)\right]}{\bar{z}^2}\left(z_n - z_r\right)\right) \end{aligned} \tag{4.78}$$

where n and r can represent any integer between 1 and T and also the special case in which n=r is allowed. The system of equations (Equation 4.61, 4.74 and 4.78) is transformed into the following system:

$$\lim_{N_{batch}\to\infty} E\left(N_n^*\right) = p'_n Z/\bar{z} \tag{4.79}$$

$$\lim_{N_{batch}\to\infty}\left(E\left(\left(N_n^*\right)^2\right) - E^2\left(N_n^*\right)\right) = \lim_{N_{batch}\to\infty} \frac{p'_n Z^*}{\bar{z}} \sum_{\substack{j=1\\ j\neq n}}^{T} p'_j z_j \frac{\bar{z} - p'_n\left(z_n - z_j\right)}{\bar{z}^2} \tag{4.80}$$

$$\lim_{N_{batch}\to\infty} E\left(N_n^*\left(N_n^*/p'_n - N_r^*/p'_r\right)\right) = \lim_{N_{batch}\to\infty}\left(\frac{Z^*}{\bar{z}}\left(1-\Delta_{nr}\right) - \frac{p'_n Z^*}{\bar{z}^2}\left(z_n - z_r\right)\right) \tag{4.81}$$

The following derivations extract from Equations 4.79 to 4.81 an expression for the covariance between N_n^* and N_r^*. First, Equation 4.80 is rewritten:

$$\lim_{N_{batch}\to\infty} E\left[\left(N_n^*\right)^2\right] = \lim_{N_{batch}\to\infty}\left(\frac{p'_n Z^*}{\bar{z}}\sum_{\substack{j=1\\ j\neq n}}^{T} p'_j\, z_j\,\frac{\bar{z}-p'_n\left(z_n-z_j\right)}{\bar{z}^2}+E^2\left(N_n^*\right)\right) \tag{4.82}$$

The above result is substituted in Equation 4.81:

$$\begin{aligned}&\lim_{N_{batch}\to\infty}\frac{1}{p'_n}\left[\frac{p'_n Z^*}{\bar{z}}\sum_{\substack{j=1\\ j\neq n}}^{T} p'_j\, z_j\,\frac{\bar{z}-p'_n\left(z_n-z_j\right)}{\bar{z}^2}+E^2\left(N_n^*\right)\right]\\ &-\lim_{N_{batch}\to\infty}\frac{1}{p'_r}E\left(N_n^* N_r^*\right)=\lim_{N_{batch}\to\infty}\left(\frac{Z^*}{\bar{z}}\left(1-\Delta_{nr}\right)-\frac{p'_n Z^*}{\bar{z}^2}\left(z_n-z_r\right)\right)\end{aligned} \tag{4.83}$$

This can be rewritten:

$$\begin{aligned}&\lim_{N_{batch}\to\infty}\left(E\left(N_n^* N_r^*\right)-\frac{p'_r}{p'_n}E^2\left(N_n^*\right)\right)=\lim_{N_{batch}\to\infty}\left(-\frac{p'_r Z^*}{\bar{z}}\left(1-\Delta_{nr}\right)+\frac{p'_n p'_r Z^*}{\bar{z}^2}\left(z_n-z_r\right)\right)\\ &+\lim_{N_{batch}\to\infty}\left(\frac{p'_r Z^*}{\bar{z}}\sum_{\substack{j=1\\ j\neq n}}^{T} p'_j\, z_j\,\frac{\bar{z}-p'_n\left(z_n-z_j\right)}{\bar{z}^2}\right)\end{aligned} \tag{4.84}$$

Equation 4.79 implies that

$$\lim_{N_{batch}\to\infty} E\left(N_n^*\right)E\left(N_r^*\right)=\lim_{N_{batch}\to\infty} p'_r\, E^2\left(N_n^*\right)\Big/p'_n \tag{4.85}$$

This can be substituted into Equation 4.84:

$$\begin{aligned}&\lim_{N_{batch}\to\infty}\left(E\left(N_n^* N_r^*\right)-E\left(N_n^*\right)E\left(N_r^*\right)\right)=\lim_{N_{batch}\to\infty}\left(-\frac{p'_r Z^*}{\bar{z}}\left(1-\Delta_{nr}\right)+\frac{p'_n p'_r Z^*}{\bar{z}^2}\left(z_n-z_r\right)\right)\\ &+\lim_{N_{batch}\to\infty}\left(\frac{p'_r Z^*}{\bar{z}}\sum_{\substack{j=1\\ j\neq n}}^{T} p'_j\, z_j\,\frac{\bar{z}-p'_n\left(z_n-z_j\right)}{\bar{z}^2}\right)\end{aligned} \tag{4.86}$$

which is equivalent to:

$$\lim_{N_{batch}\to\infty}\left(E\left(N_n^* N_r^*\right)-E\left(N_n^*\right)E\left(N_r^*\right)\right)=$$

$$\lim_{N_{batch}\to\infty}\frac{p_r' Z^*}{\bar{z}}\left[\sum_{\substack{j=1\\ j\neq n}}^{T} p_j' \, z_j \, \frac{\bar{z}-p_n'\left(z_n-z_j\right)}{\bar{z}^2}-\left(1-\Delta_{nr}\right)+\frac{p_n' z_n - p_n' z_r}{\bar{z}}\right] \tag{4.87}$$

In the special case z_n=1, for all n between 1 and T (the number-based approach), the right-hand side of Equation 4.87 becomes equal to $N_{sample}\left(p_r' \Delta_{nr} - p_n' p_r'\right)$, in which N_{sample} is the total number of particles sampled, which is a well-known result for the standard multinomial distribution (see *e.g.* Tanabe and Sagae (1992) and Pederson, 1973).

If there is a large sample-to-particle size ratio, the variables N_n^* and N_r^* are approximately equal to N_n and N_r. This results in an approximate expression for the covariance between N_n and N_r:

$$\lim_{N_{batch}\to\infty}\left(E\left(N_n N_r\right)-E\left(N_n\right)E\left(N_r\right)\right)\approx$$

$$\lim_{N_{batch}\to\infty}\frac{p_r' Z^*}{\bar{z}}\left[\sum_{\substack{j=1\\ j\neq n}}^{T} p_j' \, z_j \, \frac{\bar{z}-p_n'\left(z_n-z_j\right)}{\bar{z}^2}-\left(1-\Delta_{nr}\right)+\frac{p_n' z_n - p_n' z_r}{\bar{z}}\right] \tag{4.88}$$

The right-hand side of the above equation, which is identical to the right-hand side of Equation 4.87, will be used in the next paragraph for the calculation of the variance of a sample total.

4.6 Expected value and variance of a sample total

A general expression for the expected value and for the variance of a sample total is given, because the sample concentration is the ratio of the sample totals Y_{sample} and Z_{sample}. Parameterisation of the sample total Y_{sample} using T variables y_i, which represent the value for the property of a single particle of type i, results in:

$$Y_{sample}=\sum_{i=1}^{T} y_i N_i \tag{4.89}$$

Hence, a simple expression for the expected value of Y_{sample} is:

$$E\left(Y_{sample}\right)=\sum_{i=1}^{T} y_i E\left(N_i\right) \tag{4.90}$$

The variance of the property can be written as:

$$V\left(Y_{sample}\right)=V\left(\sum_{i=1}^{T} y_i N_i\right)=\sum_{i=1}^{T}\sum_{j=1}^{T} y_i y_j \left(E\left(N_i N_j\right)-E\left(N_i\right)E\left(N_j\right)\right) \qquad \textbf{(4.91)}$$

Hence, an exact calculation of the expected value and variance of a sample total requires expressions for the expected values E(N$_i$) and covariances E(N$_i$N$_j$)−E(N$_i$)E(N$_j$). While this is very complicated, transformed variables N_i^*, N_j^* and Z^* were introduced. In the number-based approach δ is always zero, hence $N_i^* = N_i$. In a mass- or volume-based approach, there may be a difference. However, for large sample sizes, this difference will be relatively small and not influence the final results significantly. This can be demonstrated by using the relation $p_i' = N_{i,batch}/N_{batch}$. Hence, although the variable of interest is Y$_{sample}$, it is easier to calculate the variance and expected value of

$$Y_{sample}+\delta\frac{Y_{batch}}{Z_{batch}}=Y_{sample}+\delta\frac{\sum_{i=1}^{T}N_{i,batch}y_i}{\sum_{i=1}^{T}N_{i,batch}z_i}=Y_{sample}+\delta\frac{\sum_{i=1}^{T}p_i'\ y_i}{\sum_{i=1}^{T}p_i'\ z_i}$$

$$=Y_{sample}+\sum_{i=1}^{T}\frac{p_i'\ \delta}{\bar{z}}y_i=\sum_{i=1}^{T}\left(N_i+\frac{p_i'\ \delta}{\bar{z}}\right)y_i=\sum_{i=1}^{T}N_i^* y_i \qquad \textbf{(4.92)}$$

In a fixed size design with boundary value of the sample size Z, the second term, δY$_{batch}$/Z$_{batch}$, becomes negligible compared to Y$_{sample}$ for large sample sizes. Hence, for large sample sizes, the equations obtained for the variance and expected value can be applied to Y$_{sample}$. The variance is:

$$V\left(Y_{sample}+\delta\frac{Y_{batch}}{Z_{batch}}\right)=V\left(\sum_{i=1}^{T}y_i N_i^*\right)=\sum_{i=1}^{T}\sum_{j=1}^{T}y_i y_j\left(E\left(N_i^* N_j^*\right)-E\left(N_i^*\right)E\left(N_j^*\right)\right) \qquad \textbf{(4.93)}$$

The expression for the covariance, Equation 4.87, is substituted in the above expression for the variance. The result is:

$$\lim_{N_{batch}\to\infty} V\left(Y_{sample}+\delta\frac{Y_{batch}}{Z_{batch}}\right)=$$

$$\lim_{N_{batch}\to\infty}\sum_{i=1}^{T}\sum_{j=1}^{T} y_i\, y_j\,\frac{p'_j\, Z^*}{\bar{z}}\left[\sum_{\substack{k=1\\k\neq i}}^{T} p'_k\, z_k\,\frac{\bar{z}-p'_i\,(z_i-z_k)}{\bar{z}^2}-(1-\Delta_{ij})+\frac{p'_i\, z_i-p'_i\, z_j}{\bar{z}}\right] \qquad \textbf{(4.94)}$$

The factor between square brackets can be simplified:

$$\lim_{N_{batch}\to\infty} V\left(Y_{sample}+\delta\frac{Y_{batch}}{Z_{batch}}\right)=$$

$$\lim_{N_{batch}\to\infty}\sum_{i=1}^{T}\sum_{j=1}^{T} y_i\, y_j\,\frac{p'_j\, Z^*}{\bar{z}}\left[p'_i\,\sum_{k=1}^{T}\frac{p'_k\, z_k^2}{\bar{z}^2}-2p'_i\, z_i\,\frac{\bar{z}}{\bar{z}^2}+\Delta_{ij}+\frac{p'_i\, z_i-p'_i\, z_j}{\bar{z}}\right] \qquad \textbf{(4.95)}$$

Because of symmetry arguments the double summation of the last term between rectangular parentheses becomes zero after summation.

$$\lim_{N_{batch}\to\infty} V\left(Y_{sample}+\delta\frac{Y_{batch}}{Z_{batch}}\right)=\lim_{N_{batch}\to\infty}\sum_{i=1}^{T}\sum_{j=1}^{T} y_i\, y_j\,\frac{p'_j\, Z^*}{\bar{z}}\left[p'_i\,\sum_{k=1}^{T}\frac{p'_k\, z_k^2}{\bar{z}^2}-2p'_i\, z_i\,\frac{\bar{z}}{\bar{z}^2}+\Delta_{ij}\right] \qquad \textbf{(4.96)}$$

This can be rewritten:

$$\lim_{N_{batch}\to\infty} V\left(Y_{sample}+\delta\frac{Y_{batch}}{Z_{batch}}\right)=\lim_{N_{batch}\to\infty}\left(\sum_{i=1}^{T} y_i\, p'_i\,\sum_{j=1}^{T} y_j\,\frac{p'_j\, Z^*}{\bar{z}}\sum_{k=1}^{T}\frac{p'_k\, z_k^2}{\bar{z}^2}\right.$$

$$\left.-2\sum_{i=1}^{T} y_i\, z_i\, p'_i\,\sum_{j=1}^{T} y_j\,\frac{p'_j\, Z^*}{\bar{z}}\,\frac{\bar{z}}{\bar{z}^2}+\sum_{i=1}^{T} y_i^2\,\frac{p'_i\, Z^*}{\bar{z}}\right) \qquad \textbf{(4.97)}$$

Putting the common factor $Z^*/\bar{z}$ out of brackets yields:

$$\lim_{N_{batch}\to\infty} V\left(Y_{sample}+\delta\frac{Y_{batch}}{Z_{batch}}\right)=$$

$$\lim_{N_{batch}\to\infty}\frac{Z^*}{\bar{z}}\left(\sum_{i=1}^{T} y_i\, p'_i\,\sum_{j=1}^{T} y_j p'_j\,\sum_{k=1}^{T}\frac{p'_k\, z_k^2}{\bar{z}^2}-2\sum_{i=1}^{T} y_i\, z_i\, p'_i\,\sum_{j=1}^{T} y_j p'_j\,\frac{\bar{z}}{\bar{z}^2}+\sum_{i=1}^{T} y_i^2 p'_i\right) \qquad \textbf{(4.98)}$$

This can be simplified to:

$$\lim_{N_{\mathrm{batch}}\to\infty} V\left(Y_{\mathrm{sample}} + \delta\,\frac{Y_{\mathrm{batch}}}{Z_{\mathrm{batch}}}\right) = \lim_{N_{\mathrm{batch}}\to\infty} \frac{Z^{*}}{\bar{z}}\sum_{i=1}^{T} p_i' \left(y_i - \frac{z_i}{\bar{z}}\sum_{j=1}^{T} p_j'\, y_j\right)^{2} \tag{4.99}$$

Similarly, the expected value of $Y_{sample}+\delta Y_{batch}/Z_{batch}$ is calculated using Equation 4.79:

$$\lim_{N_{\mathrm{batch}}\to\infty} E\left(Y_{\mathrm{sample}} + \delta\,\frac{Y_{\mathrm{batch}}}{Z_{\mathrm{batch}}}\right) = \lim_{N_{\mathrm{batch}}\to\infty} E\left(\sum_{i=1}^{T} N_i^{*} y_i\right) = \sum_{i=1}^{T} \frac{p_i'\, Z}{\bar{z}}\, y_i \tag{4.100}$$

When the sample is large enough, $\delta Y_{batch}/Z_{batch}$ is negligible compared to Y_{sample} and the relative difference between Z^{*} and Z will be negligible, so that Z^{*} can be replaced by Z. Hence, the following approximate results for the expected value and variance of Y_{sample} are obtained from Equation 4.99 and 4.100:

$$E\left(Y_{\mathrm{sample}}\right) \approx \sum_{i=1}^{T} \frac{p_i'\, Z}{\bar{z}}\, y_i \tag{4.101}$$

$$V\left(Y_{\mathrm{sample}}\right) \approx \frac{Z}{\bar{z}}\sum_{i=1}^{T} p_i' \left(y_i - \frac{z_i}{\bar{z}}\sum_{j=1}^{T} p_j'\, y_j\right)^{2} \tag{4.102}$$

These results are approximations and will therefore not be used in further derivations. Instead, the exact results, Equation 4.99 and 4.100 will be used.

4.7 Sample concentration

In this section, expressions are derived for the expected value and variance of Y_{sample}/Z_{sample}, which is an estimator for Y_{batch}/Z_{batch}. As a first step in the derivation of the expected value, both right-hand side and left-hand side of Equation 4.100 are divided by Z:

$$\lim_{N_{\mathrm{batch}}\to\infty} E\left(\frac{Y_{\mathrm{sample}}}{Z} + \frac{\delta}{Z}\,\frac{Y_{\mathrm{batch}}}{Z_{\mathrm{batch}}}\right) = \sum_{i=1}^{T} \frac{p_i'\, y_i}{\bar{z}} = \frac{Y_{\mathrm{batch}}}{Z_{\mathrm{batch}}} \tag{4.103}$$

When the boundary value of the sample size becomes large, Y_{sample}/Z is approximately equal to the sample concentration Y_{sample}/Z_{sample} and $(\delta/Z)Y_{batch}/Z_{batch}$ can be neglected as δ remains between 0 and $-z_{max}$. Hence, it follows that

$$\lim_{Z\to\infty} E\left(\frac{Y_{\mathrm{sample}}}{Z} + \frac{\delta}{Z}\,\frac{Y_{\mathrm{batch}}}{Z_{\mathrm{batch}}}\right) = \lim_{Z\to\infty} E\left(\frac{Y_{\mathrm{sample}}}{Z_{\mathrm{sample}}}\right) \tag{4.104}$$

Therefore, in order to find an expression for the expected value of the sample concentration, the limit of Z=∞ has to be taken. However, if this is done directly on both sides of Equation 4.103, the following result is obtained:

$$\lim_{Z\to\infty}\ \lim_{N_{\mathrm{batch}}\to\infty} E\left(\frac{Y_{\mathrm{sample}}}{Z}+\frac{\delta}{Z}\frac{Y_{\mathrm{batch}}}{Z_{\mathrm{batch}}}\right)=\frac{Y_{\mathrm{batch}}}{Z_{\mathrm{batch}}} \qquad \textbf{(4.105)}$$

Only if both limits could be interchanged, Equation 4.104 can be substituted into the above equation, providing an equation for the expected value of the sample concentration. However, without further mathematical investigation, interchanging both limits would not be justified. Therefore, in the following, it will be investigated whether both limits may be interchanged.

All previous results containing $\lim\limits_{N_{\mathrm{batch}}\to\infty}$ are based on the following limit (see also Equation 3.4):

$$\lim_{N_{\mathrm{batch}}\to\infty} p_i = \lim_{N_{\mathrm{batch}}\to\infty} \left(p_i' - n_i/N_{\mathrm{batch}}\right) N_{\mathrm{batch}}/\left(N_{\mathrm{batch}}-k\right) = p_i' \qquad \textbf{(4.106)}$$

where p_i is the probability that the next particle sampled is of type i and n_i is the number of particles previously sampled belonging to the i^{th} class. If Z is allowed to go to infinity, it cannot be guaranteed that n_i/N_{batch} remains negligible compared to p_i' in Equation 4.106. Therefore, it would be incorrect to interchange both limits in Equation 4.105. Fortunately, this can be made correct by slightly changing the limiting process. For this, the batch-to-sample size ratio, r_{bs}, is defined as

$$r_{bs} = Z_{\mathrm{batch}}/Z \qquad \textbf{(4.107)}$$

For constant Z, the limit of $N_{\mathrm{batch}}=\infty$ is equivalent to the limit of $r_{bs}=\infty$. Therefore, all previous results derived in the limit of $N_{\mathrm{batch}}=\infty$, can also be read as results valid in the limit of $r_{bs}=\infty$, denoted as $\lim\limits_{r_{bs}\to\infty}$. The advantage, however, of taking the latter limit is that $\lim\limits_{r_{bs}\to\infty} p_i = p_i'$ irrespective of the value of Z, because if $r_{bs}=\infty$, n_i/N_{batch} can always be neglected compared to p_i'. This proves that when in Equation 4.105 $\lim\limits_{N_{\mathrm{batch}}\to\infty}$ is replaced by $\lim\limits_{r_{bs}\to\infty}$, the two limits, $\lim\limits_{Z\to\infty}$ and $\lim\limits_{r_{bs}\to\infty}$, may be interchanged. This results in:

$$\lim_{Z\to\infty}\lim_{r_{bs}\to\infty} E\left(\frac{Y_{sample}}{Z}+\frac{\delta}{Z}\frac{Y_{batch}}{Z_{batch}}\right)=\lim_{r_{bs}\to\infty}\lim_{Z\to\infty} E\left(\frac{Y_{sample}}{Z}+\frac{\delta}{Z}\frac{Y_{batch}}{Z_{batch}}\right)$$
$$=\lim_{r_{bs}\to\infty}\lim_{Z\to\infty} E\left(\frac{Y_{sample}}{Z_{sample}}\right)=\frac{Y_{batch}}{Z_{batch}} \qquad \textbf{(4.108)}$$

As the limit of Z=∞ was effectively taken in order to neglect δ with respect to Z, it is convenient to define the sample-to-particle size ratio, r_{sp}, as:

$$r_{sp}=Z/z_{max} \qquad \textbf{(4.109)}$$

where z_{max} is the largest particle size of the batch. The limiting process, Z=∞, is equivalent to the limit of r_{sp}=∞, denoted as $\lim_{r_{sp}\to\infty}$. Hence, Equation 4.108 becomes:

$$\lim_{r_{bs}\to\infty}\lim_{r_{sp}\to\infty} E\left(Y_{sample}/Z_{sample}\right)=Y_{batch}/Z_{batch} \qquad \textbf{(4.110)}$$

Hence, in the limit of an infinite sample-to-particle size ratio and an infinite batch-to-sample size ratio, the estimator is unbiased. Equation 4.110 is an exact result. However, in practice the batch-to-sample size ratio size and the sample-to-particle size ratio are always finite. Therefore, the expected value of the sample concentration is only approximately equal to the batch concentration, resulting in:

$$E\left(Y_{sample}/Z_{sample}\right)\approx Y_{batch}/Z_{batch} \qquad \textbf{(4.111)}$$

This implies that the sample concentration may contain a (small) bias. Similarly to the above followed method to derive an expression for the expected value, an expression for the variance of the sample concentration is derived. Dividing both left-hand side and right-hand side of Equation 4.99 by Z results in:

$$\lim_{N_{batch}\to\infty} ZV\left(\frac{Y_{sample}}{Z}+\frac{\delta}{Z}\frac{Y_{batch}}{Z_{batch}}\right)=\lim_{N_{batch}\to\infty}\frac{Z^*}{Z\bar{z}}\sum_{i=1}^{T}p'_i\left(y_i-\frac{z_i}{\bar{z}}\sum_{j=1}^{T}p'_j\,y_j\right)^2 \qquad \textbf{(4.112)}$$

Subsequently, the limit of N_{batch}=∞ is replaced by the limit of r_{bs}=∞, the limit of r_{sp}=∞ is taken on both sides of Equation 4.112 and on the left-hand side both limits are interchanged. In addition, the factor Z on the left-hand side is replaced by $Z_{sample}(Z/Z_{sample})$. This results in:

$$\lim_{r_{bs}\to\infty}\lim_{r_{sp}\to\infty}\frac{Z}{Z_{sample}}Z_{sample}V\left(\frac{Y_{sample}}{Z}+\frac{\delta}{Z}\frac{Y_{batch}}{Z_{batch}}\right)=\lim_{r_{bs}\to\infty}\lim_{r_{sp}\to\infty}\frac{Z^{*}}{Z\bar{z}}\sum_{i=1}^{T}p_i'\left(y_i-\frac{z_i}{\bar{z}}\sum_{j=1}^{T}p_j'\,y_j\right)^2 \qquad \textbf{(4.113)}$$

In the limit of $r_{sp}=\infty$, Z/Z_{sample} and Z^{*}/Z are one, Y_{sample}/Z is equal to Y_{sample}/Z_{sample} and $(\delta/Z)Y_{batch}/Z_{batch}$ is zero. Hence, Equation 4.113 becomes:

$$\lim_{r_{bs}\to\infty}\lim_{r_{sp}\to\infty}Z_{sample}V\left(\frac{Y_{sample}}{Z_{sample}}\right)=\frac{1}{\bar{z}}\sum_{i=1}^{T}p_i'\left(y_i-\frac{z_i}{\bar{z}}\sum_{j=1}^{T}p_j'\,y_j\right)^2 \qquad \textbf{(4.114)}$$

The above equation is an exact result. However, in practice the batch-to-sample size ratio and the sample-to-particle size ratio are always finite. Therefore, the variance of the sample concentration multiplied by Z_{sample} is only approximately equal to the right-hand side of Equation 4.114. Therefore, the right-hand side of Equation 4.114 divided by Z_{sample} provides an estimate for the variance, based on the identities of the particles in the batch. This estimate is denoted as $V_{batch}(Y_{sample}/Z_{sample})$, resulting in:

$$V_{batch}\left(\frac{Y_{sample}}{Z_{sample}}\right)=\frac{1}{Z_{sample}\bar{z}}\sum_{i=1}^{T}p_i'\left(y_i-\frac{z_i}{\bar{z}}\sum_{j=1}^{T}p_j'\,y_j\right)^2 \qquad \textbf{(4.115)}$$

The above result provides an estimate for the relation between the variance of the sample concentration and the sample size Z_{sample} (*i.e.* the mass or volume sampled).

4.8 Results

Using a multinomial selection scheme with fixed sample size (mass, volume or number) an expression for the expected value of N_n^{*} and an expression for the covariance between N_n^{*} and N_r^{*} (Equation 4.79 and Equation 4.87 respectively) are now available in the limit of an infinite batch-to-sample size ratio for all values of n and r between 1 and T. These equations were used to derive expressions for the expected value and variance of $Y_{sample}+Y_{batch}(Z-Z_{sample})/Z_{batch}$. Subsequently, these results have been used to derive expressions for the expected value and variance of a ratio of the sample totals Y_{sample} and Z_{sample} in the limit of an infinite sample-to-particle size ratio and an infinite batch-to-sample size ratio (Equation 4.110 and Equation 4.114 respectively).

Using Equation 4.114 an expression (Equation 4.115) for the variance estimated using the identities of the particles in the batch, $V_{batch}(Y_{sample}/Z_{sample})$, was derived

4.9 Discussion

As a consequence of the fact that Equation 4.110 is only valid within the limit of an infinite sample-to-particle size ratio and an infinite batch-to-sample size ratio, the sample concentration may contain a (slight) bias. This bias will be investigated in Chapter 6.

Similarly, Equation 4.114 is also only valid within the limit of an infinite sample-to-particle size ratio and an infinite batch-to-sample size ratio. Therefore, the value of the true variance may differ from the value of $V_{batch}(Y_{sample}/Z_{sample})$ calculated with Equation 4.115. In Chapter 7, this difference will be investigated for the sampling of a batch of wooden chips, two batches of steel slag produced during the production of steel and a batch of recycled plastic chips.

4.10 Conclusions

An equation for the relation between the variance and the sample size was derived (Equation 4.115). Therefore, the first criterion for a sampling theory ("The theory must provide an equation for the variance of the sample concentration, containing the mass or volume sampled and an arbitrary number of additional parameters.") is met. Because the right-hand side of Equation 4.115 depends on the identities of the particles in the batch, the fifth criterion for a sampling theory ("The theory must allow determination of the parameters of the size-variance equation, using prior knowledge of the properties of the particles in the batch.") is also met.

4.11 References

V. Barnett (1974) *Elements of sampling theory*, English University Press, London, 152 pp.

W. Feller (1968) *An introduction to probability theory and its applications, vol. 1*, John Wiley and Sons, 509 pp.

D. G. Pederson (1973) An Approximate Method of Sampling a Multinormal Population, *Biometrics*, vol. **29**, p. 814-821.

C. Särndal, B. Swensson and J. Wretman (1992) *Model Assisted Survey Sampling,* Springer, New York, 694 pp.

K. Tanabe and M. Sagae (1992) An Exact Cholesky Decomposition and the Generalized Inverse of the Variance-Covariance Matrix of the Multinomial Distribution, with Applications, *Journal of the Royal Statistical Society B,* **54**, p. 211-219.

Chapter 5 Estimation of the variance of the sample concentration[11]

The Horvitz-Thompson estimator can provide a general and unbiased estimate for the variance of the π-expanded estimator. It is demonstrated that the sample concentration can be rewritten in the form of a π-expanded estimator, indicating that the Horvitz-Thompson estimator can be applied for estimation of the variance of the sample concentration. Because in this study particles are classified, the behaviour of the π-expanded estimator and Horvitz-Thompson estimator under classification is investigated. Derivations of expressions for the first- and second-order inclusion probabilities, using results from Chapter 4, are performed. These expressions are substituted into the π-expanded estimator and the Horvitz-Thompson estimator. This results in an expression for the variance, estimated using the properties of the particles in the sample. Finally, as an application of the obtained equation for the π-expanded estimator and the variance, the obtained equations are worked out for mass concentrations.

5.1 Introduction

The variance of the sample concentration is a measure of the potential statistical fluctuations of the sample concentration around its expected value. Because it is important to have insight into the magnitude of these variations, it is of practical significance to estimate the numerical relationship between the variance and the sample size. For the drawing of an ideal sample from a random arrangement of particles, the value of the variance of the sample concentration can be calculated using knowledge of the distribution of the particles in the batch and Equation 4.115:

$$V_{\text{batch}}\left(\frac{Y_{\text{sample}}}{Z_{\text{sample}}}\right)=\frac{1}{Z_{\text{sample}}\bar{z}}\sum_{i=1}^{T}p'_i\left(y_i-\frac{z_i}{\bar{z}}\sum_{j=1}^{T}p'_j\,y_j\right)^2 \tag{5.1}$$

[11] The main aspects of this chapter have been published in: B. Geelhoed, H. J. Glass (2004) Estimators for particulate sampling derived from a multinomial distribution, *Statistica Neerlandica*, **58**, p. 57-74.

This can be written as an equation for the variance, containing the sample size and one additional parameter C:

$$V_{batch}\left(Y_{sample}/Z_{sample}\right)=C/Z_{sample} \tag{5.2}$$

where the parameter C is given by:

$$C=\frac{1}{\bar{z}}\sum_{i=1}^{T}p_i'\left(y_i-\frac{z_i}{\bar{z}}\sum_{j=1}^{T}p_j'\,y_j\right)^2 \tag{5.3}$$

Hence, it was demonstrated in Chapter 4 that the fifth criterion for a sampling theory ("The theory must allow determination of the parameters of the size-variance equation, using prior knowledge of the properties of the particles in the batch.") is met.

In this chapter, two alternative methods of estimating the variance will be investigated. The first method is related to the fourth criterion for a sampling theory ("The theory must allow determination of the parameters of the size-variance equation, using the measured sample concentrations of one or more samples of a given size.") and addresses calculation of the parameters of the size-variance equation using analysis results of multiple samples. The second method is related to the sixth criterion for a sampling theory ("The theory must allow determination of the parameters of the size-variance equation, using posterior knowledge of the properties of the particles in the sample.") and addresses calculation of the parameters of the size-variance equation using the results of analyses of the particles in the sample.

5.2 *Estimation of the size-variance relationship using multiple samples: the first method*

In the first method, N_{det} samples are drawn using the size-based multinomial selections, with boundary value of the sample size set to Z, and determination of the concentration in each sample. When the concentration in the i^{th} sample is denoted as $c_{sample,i}$, where i represents any integer between 1 and N_{det}, the variance of a sample can be estimated using the values of $c_{sample,i}$ (see *e.g.* Cohen, 1988):

$$\hat{V}\left(Y_{sample}/Z_{sample}\right)=\frac{1}{N_{det}-1}\sum_{i=1}^{N_{det}}\left(c_{sample,i}-\bar{c}\right)^2 \tag{5.4}$$

where $\bar{c}$ is the arithmetic mean of the N_{det} values of the sample concentration. In order to estimate the parameter C, it is assumed that the value obtained for $\hat{V}\left(Y_{sample}/Z_{sample}\right)$ is equal to the variance calculated with Equation 5.2, where C is replaced by its estimate, $\hat{C}_1$, and Z_{sample} is the average sample mass of the N_{det} samples. This yields:

$$\hat{C}_1 = \frac{1}{N_{det}\left(N_{det}-1\right)}\left(\sum_{j=1}^{N_{det}} Z_{sample,i}\right)\left(\sum_{i=1}^{N_{det}}\left(c_{sample,i}-\bar{c}\right)^2\right) \tag{5.5}$$

where $Z_{sample,i}$ is the size of the i^{th} sample. The above equation provides an estimate for the parameter of the size-variance equation, using the concentrations and masses of an arbitrary number (N_{det}) of samples. Hence, the fourth criterion for a sampling theory is met.

5.3 Estimation of the size-variance relationship using posterior knowledge of the particles in the sample: the second method

The second method is based on the fact that although the distribution of particles in the batch is unknown, the distribution of particles in the sample can in principle be measured. This implies that in the second method, the variance of Y_{sample}/Z_{sample} is estimated using knowledge of the distribution of the particles in the sample. As a first guess, the following ad hoc substitutions are made in Equation 5.1:

- p_i' is replaced by the number fraction of particles of type i in the sample: $N_i \big/ \sum_{j=1}^{T} N_j$
- $\bar{z}$ is replaced by the average particle size in the sample, $\sum_{j=1}^{T} z_j N_j \big/ \sum_{j=1}^{T} N_j$

The result is the following estimator for the variance, denoted as $\hat{V}\left(Y_{sample}/Z_{sample}\right)$:

$$\hat{V}\left(\frac{Y_{sample}}{Z_{sample}}\right) = \frac{1}{Z_{sample}^2}\sum_{i=1}^{T} N_i z_i^2\left(\frac{y_i}{z_i}-\frac{Y_{sample}}{Z_{sample}}\right)^2 \tag{5.6}$$

Because the estimator $\hat{V}\left(Y_{sample}/Z_{sample}\right)$ is a function of the random variables N_i, Y_{sample} and Z_{sample}, the estimator $\hat{V}\left(Y_{sample}/Z_{sample}\right)$ is by definition a random variable. For a sample S, its value is denoted as $\hat{V}\left(Y_{sample}/Z_{sample},S\right)$. Using the definitions in Paragraph 4.2, it is possible to associate a bias to $\hat{V}\left(Y_{sample}/Z_{sample}\right)$. Because the substitutions used to obtain $\hat{V}\left(Y_{sample}/Z_{sample}\right)$ were ad hoc, it is not guaranteed that $\hat{V}\left(Y_{sample}/Z_{sample}\right)$ is unbiased, even in the limit of an infinite batch-to-sample size ratio and an infinite sample-to-particle size ratio. If $\hat{V}\left(Y_{sample}/Z_{sample}\right)$ is biased, the obtained estimate will have the tendency to either underestimate or overestimate the true variance of the sample concentration. Therefore, in the next paragraph, the general and unbiased Horvitz-Thompson estimator is discussed.

5.4 The Horvitz-Thompson estimator

An estimator for the variance of the π-expanded estimator will be determined. The π-expanded estimator, which is a general and unbiased estimator, requires definition of the inclusion probability, π_i, which is the probability that the i^{th} particle of the batch, with i ranging from 1 to N_{batch}, is selected during the drawing of a sample. It is noted that in the theory of finite population sampling, the general terms 'units' and 'population' are common. When dealing with particulate materials the units are the individual particles of material and the population is denoted as the batch. When U_i is the set of samples that contains the i^{th} particle of the batch, the inclusion probability of the i^{th} particle is:

$$\pi_i = \sum_{S\in U_i} P(S) \tag{5.7}$$

It will be demonstrated in this chapter that, if sampling corresponds to size-based multinomial selections, the inclusion probability is approximated by the ratio of the sample size and the batch size.

Because the π-expanded estimator applies for the estimation of a population total, in the following the properties of a population total are defined. Note that because, in this thesis, the population is a batch of particulate material, the population total is often indicated as batch total. In the general sample survey theory (see Särndal *et al*, 1992), the population total Y_{batch} is given by:

$$Y_{batch} = \sum_{i=1}^{N_{batch}} y_{n(i)} \tag{5.8}$$

in which N_{batch} is the total number of particles in the batch, n(i) is the class of the i^{th} particle in the batch and $y_{n(i)}$ is the property of interest in a particle belonging to the $n(i)^{th}$ class.

To investigate the estimation of a batch total, first the sample total is discussed. For definition of the sample total, the indicator I_i is required. The indicator I_i indicates whether the i^{th} particle in the batch is selected or not. When a sample S does not contain the i^{th} particle $I_i(S)=0$; when S contains the i^{th} particle $I_i(S)=1$. The sample total is defined as:

$$Y_{sample} = \sum_{i=1}^{N_{batch}} I_i \, y_{n(i)} \tag{5.9}$$

If $y_{n(i)}>0$, for all i between 1 and N_{batch}, and the sample size is smaller than the batch, for any sample S, $Y_{sample}(S)$ is always smaller than the value of Y_{batch}. When $Y_{sample}(S)<Y_{batch}$ for all possible S, Y_{sample} is a negatively biased estimator for Y_{batch},

which can be demonstrated with the following simple derivation:

$$B\left(Y_{\text{sample}}\right)=E\left(Y_{\text{sample}}\right)-Y_{\text{batch}}=\sum_{S\in U}P(S)Y_{\text{sample}}(S)-Y_{\text{batch}}<\sum_{S\in U}P(S)Y_{\text{batch}}-Y_{\text{batch}}=0 \qquad \textbf{(5.10)}$$

in which the identity $\sum_{S\in U}P(S)=1$ was used. Therefore, the sample total would systematically underestimate the value of the batch total. Fortunately, the sample total of the 'expanded' variable, $y_{n(i)}/\pi_i$, in which π_i is the inclusion probability of the ith particle in the batch, is of the same order of magnitude as the population total. This estimator is by definition the π-expanded estimator, often referred to as the Horvitz-Thompson estimator[12]. It will now be proven that the π-expanded estimator is unbiased.

From the definition of the inclusion probability, the inclusion probability of the ith particle is equal to the expected value of I_i, *i.e.* $\pi_i=E(I_i)$. The π-expanded estimator, denoted $\left\langle Y_{\text{batch}}\right\rangle_\pi$, can be written as:

$$\left\langle Y_{\text{batch}}\right\rangle_\pi=\sum_{i=1}^{N_{\text{batch}}} I_i\frac{y_{n(i)}}{\pi_i}=\sum_{i=1}^{N_{\text{batch}}} I_i\frac{y_{n(i)}}{E(I_i)} \qquad \textbf{(5.11)}$$

This estimator is unbiased because:

$$B\left(\left\langle Y_{\text{batch}}\right\rangle_\pi\right)=E\left(\left\langle Y_{\text{batch}}\right\rangle_\pi\right)-Y_{\text{batch}}=\sum_{i=1}^{N_{\text{batch}}} E\left(I_i\right)\frac{y_{n(i)}}{E\left(I_i\right)}-Y_{\text{batch}}=0 \qquad \textbf{(5.12)}$$

For the variance of the π-expanded estimator, a general Horvitz-Thompson estimator is defined by (Särndal *et al*, 1992):

$$\hat{V}_{HT}\left(\left\langle Y_{\text{batch}}\right\rangle_\pi\right)=\sum_{i=1}^{N_{\text{batch}}}\sum_{\substack{j=1\\ j\neq i}}^{N_{\text{batch}}}\left(\frac{1}{\pi_i\pi_j}-\frac{1}{\pi_{ij}}\right)I_iI_jy_{n(i)}y_{n(j)}+\sum_{i=1}^{N_{\text{batch}}}\left(\frac{1}{\pi_i^2}-\frac{1}{\pi_i}\right)I_iy_{n(i)}^2 \qquad \textbf{(5.13)}$$

in which π_{ij} is the second-order inclusion probability[13], *i.e.* the probability that the particle pair consisting of the ith and jth particle, with $i\neq j$, is selected. In equation:

[12] To avoid confusion in this work the term 'π-expanded estimator' is used and the term 'Horvitz-Thompson estimator' is reserved exclusively for the variance estimator defined in Equation 5.13.

[13] Wherever the symbol π has two indices, as *e.g.* in π_{ij}, this symbol represents a second-order inclusion probability and not the first-order inclusion probability π_k, where $k=i\times j$. When the indices are numbers, *e.g.* i=1 and j=2, a semicolon must be placed between the indices to denote the second order inclusion probability, *e.g.* $\pi_{ij}=\pi_{1;2}$ for i=1 and j=2. The use of a semicolon is essential in this case, because without a semicolon, the symbol becomes π_{12}, which represents the first-order inclusion probability of the twelfth particle in the batch.

$$\pi_{ij} = \sum_{S \in U_{ij}} P(S) \tag{5.14}$$

where U_{ij} is the set of samples that contains the ith and jth particle ($i \neq j$) of the batch[14]. Here, the second-order inclusion probability π_{ij} is not defined for $i=j$, because this combination would not form a pair. If, however, it would be defined that $\pi_{ii}=\pi_i$ for all i between 1 and T, the second term on the right-hand side of Equation 5.13 could be omitted, if the double summation is extended over all values of i and j between 1 and T. This would yield a simpler equation. However, in this thesis, π_{ij} is not defined for $i=j$, avoiding the artificial concept of a 'pair' that consists of one particle. The proof that Equation 5.13 provides an unbiased variance-estimator can be found in literature (Särndal *et al*, 1992). For a sample S, the value of the Horvitz-Thompson estimator is equal to:

$$\hat{V}_{HT}\left(\langle Y_{batch}\rangle_{\pi}, S\right) = \sum_{i=1}^{N_{batch}} \sum_{\substack{j=1 \\ j \neq i}}^{N_{batch}} \left(\frac{1}{\pi_i \pi_j} - \frac{1}{\pi_{ij}}\right) I_i(S) I_j(S) y_{n(i)} y_{n(j)} + \sum_{i=1}^{N_{batch}} \left(\frac{1}{\pi_i^2} - \frac{1}{\pi_i}\right) I_i(S) y_{n(i)}^2 \tag{5.15}$$

In this chapter, Equation 5.13 is the starting point for the derivation of an estimator for the variance of the sample concentration. The derivation contains four essential elements: (*i*) demonstration that if Equation 5.11 and 5.13 are slightly modified, they can be applied to concentrations instead of batch totals, (*ii*) replacement of the summations over all particles by summations over the T particle classes, (*iii*) calculation of the first- and second-order inclusion probabilities in the size-based approach and (*iv*) substitution of the expressions obtained for the first- and second-order inclusion probabilities and calculation of the variance estimator. These four steps are carried out in Paragraphs 5.5, 5.6, 5.7 and 5.8 respectively.

5.5 The π-expanded estimator for the concentration

It may be necessary to estimate the concentration of a property in the batch, expressed as a ratio of batch totals Y_{batch} and Z_{batch}. The denominator Z_{batch} is the size of the batch, which can correspond to the mass or volume. Often the batch concentration is estimated using the corresponding sample ratio Y_{sample}/Z_{sample}, which is denoted as the ratio estimator. Because the denominator and numerator of the ratio estimator are generally measured without analysing every particle in the sample separately, this

[14] Wherever the symbol U has two indices, as *e.g.* in U_{ij}, this symbol represents the set of samples that contains the ith and jth particle of the batch and not the set of samples, U_k, that contain the kth particle of the batch where $k=i \times j$. When the indices are numbers, *e.g.* $i=1$ and $j=2$, a semicolon must be placed between the indices to denote the set of samples that contains the ith and jth particle of the batch, *e.g.* $U_{ij} = U_{1;2}$ for $i=1$ and $j=2$. The use of a semicolon is essential in this case, because without a semicolon, the symbol becomes U_{12}, which represents the set of samples that contain the twelfth particle in the batch.

estimator is extensively used in practice. However, the π-expanded estimator can also be developed by recognizing that Y_{batch}/Z_{batch} is the batch total of $y_{n(i)}/Z_{batch}$. The π-expanded estimator, given by Equation 5.11, becomes:

$$\left\langle Y_{batch}/Z_{batch}\right\rangle_{\pi} = \sum_{i=1}^{N_{batch}} \frac{I_i\, y_{n(i)}}{Z_{batch}\pi_i} \tag{5.16}$$

The value of the π-expanded estimator becomes equal to the value of the ratio estimator for $\pi_i=Z_{sample}/Z_{batch}$. This is of practical significance, because in Paragraph 5.7, it will be demonstrated that, if sampling corresponds to size-based multinomial selections, π_i can be approximated by Z_{sample}/Z_{batch}. In this case, the ratio estimator is unbiased and has a variance estimator given by Equation 5.13, with $y_{n(i)}$ replaced by $y_{n(i)}/Z_{batch}$.

$$\hat{V}_{HT}\left(\left\langle \frac{Y_{batch}}{Z_{batch}}\right\rangle_{\pi}\right) = \sum_{i=1}^{N_{batch}}\sum_{\substack{j=1\\ j\neq i}}^{N_{batch}}\left(\frac{1}{\pi_i\,\pi_j}-\frac{1}{\pi_{ij}}\right) I_i\, I_j\, \frac{y_{n(i)}y_{n(j)}}{Z^2_{batch}} + \sum_{i=1}^{N_{batch}}\left(\frac{1}{\pi_i^2}-\frac{1}{\pi_i}\right) I_i\, \frac{y^2_{n(i)}}{Z^2_{batch}} \tag{5.17}$$

This shows that, in the size-based approach, the positive properties of the Horvitz-Thompson and the ratio estimator can be exploited at the same time.

5.6 Behaviour of the π-expanded and Horvitz-Thompson estimator under classification

In this paragraph, the behaviour of the π-expanded estimator for the batch concentration and the corresponding Horvitz-Thompson estimator for the variance under classification of the particles into T classes is investigated. Equation 5.16 and 5.17 are modified, *i.e.* the summations over all the particles are replaced by summations over the T classes, using the following two general assumptions:

- It is assumed that the inclusion probability is constant for particles of a given type. Denoting the class of the i^{th} particle as n(i) for all i between 1 and N_{batch}, this condition can be mathematically expressed as follows: $\pi_i=\pi_j$ if n(i)=n(j) for all i and j between 1 and N_{batch}.
- For the second-order inclusion probability a similar condition is assumed: $\pi_{ir}=\pi_{js}$ if n(i)=n(j) and n(r)=n(s) for all i, j, r and s between 1 and N_{batch}, i=r and j=s excluded.

The above two conditions can be used to define the new parameters κ_n and κ_{nk} for all n and k between 1 and T, which can be interpreted as the first- and second-order inclusion probability of particles belonging to the nth and kth class[15]. These new parameters are required for the derivation presented in this paragraph of expressions for the π-expanded estimator and Horvitz-Thompson estimator subject to classification.

$$\kappa_n = \pi_i \qquad \text{if} \qquad n(i) = n \tag{5.18}$$

$$\kappa_{nk} = \pi_{ij} \qquad \text{if} \qquad n(i) = n \qquad \text{and} \qquad n(j) = k \qquad \text{and} \qquad i \neq j \tag{5.19}$$

Note that although π_{ij} is not defined for i=j, κ_{nk} is defined for n=k, because different particles may belong to the same class. Using the definition of κ_n, the π-expanded estimator for Y_{batch}/Z_{batch}, given by Equation 5.16, can be written as a summation over the distinct classes.

$$\left\langle Y_{batch}/Z_{batch} \right\rangle_\pi = \sum_{n=1}^{T} \frac{N_n y_n}{Z_{batch}\kappa_n} \tag{5.20}$$

in which now y_n denotes the value of the property of a particle of type n and N_n is the number of particles in the sample belonging to the nth class. Equation 5.20 can be identified as a simple generalization of Equation 5.16. Similarly, the equation for the Horvitz-Thompson estimator for the variance of the sample concentration, Equation 5.17, can be rewritten using summations over the T particle classes. However, this derivation is not a simple generalization.

According to the definitions of κ_n and κ_{nk}, π_{ij} in Equation 5.17 may be replaced by $\kappa_{n(i)n(j)}$ and π_i and π_j may be replaced by $\kappa_{n(i)}$ and $\kappa_{n(j)}$ respectively. Equation 5.17 now becomes:

$$\hat{V}_{HT}\left(\left\langle \frac{Y_{batch}}{Z_{batch}} \right\rangle_\pi\right) = \sum_{i=1}^{N_{batch}} \sum_{\substack{j=1 \\ j\neq i}}^{N_{batch}} \left(\frac{1}{\kappa_{n(i)}\kappa_{n(j)}} - \frac{1}{\kappa_{n(i)n(j)}} \right) I_i I_j \frac{y_{n(i)} y_{n(j)}}{Z^2_{batch}} + \sum_{i=1}^{N_{batch}} \left(\frac{1}{\kappa^2_{n(i)}} - \frac{1}{\kappa_{n(i)}} \right) I_i \frac{y^2_{n(i)}}{Z^2_{batch}} \tag{5.21}$$

The double summation can be extended over all values of i and j when the terms with i=j are subtracted.

[15] Wherever the symbol κ has two indices, as *e.g.* in κ_{nk}, this symbol represents a second-order inclusion probability and not the first-order inclusion probability κ_r, where r=n×k. When the indices are numbers, *e.g.* n=1 and k=2, a semicolon must be placed between the indices to denote the second order inclusion probability, *e.g.* $\kappa_{nk} = \kappa_{1;2}$ for n=1 and k=2. The use of a semicolon is essential in this case, because without a semicolon, the symbol becomes κ_{12}, which represents the first-order inclusion probability of a particle belonging to the twelfth class.

$$\hat{V}_{HT}\left(\left\langle\frac{Y_{batch}}{Z_{batch}}\right\rangle_{\pi}\right)=\sum_{i=1}^{N_{batch}}\sum_{j=1}^{N_{batch}}\left(\frac{1}{\kappa_{n(i)}\kappa_{n(j)}}-\frac{1}{\kappa_{n(i)n(j)}}\right)I_i\,I_j\,\frac{y_{n(i)}y_{n(j)}}{Z_{batch}^2}-$$

$$\sum_{i=1}^{N_{batch}}\left(\frac{1}{\kappa_{n(i)}^2}-\frac{1}{\kappa_{n(i)n(i)}}\right)I_i\,\frac{y_{n(i)}^2}{Z_{batch}^2}+\sum_{i=1}^{N_{batch}}\left(\frac{1}{\kappa_{n(i)}^2}-\frac{1}{\kappa_{n(i)}}\right)I_i\,\frac{y_{n(i)}^2}{Z_{batch}^2} \tag{5.22}$$

The summands with non-zero indicators depend indirectly on the particle numbers i and j via n(i) or n(j) respectively. Therefore, in each summation, terms for which n(i)=n and n(j)=k are equal when n and k are constant. When equal terms are combined, the summations are rewritten as summations over the particle classes n and k:

$$\hat{V}_{HT}\left(\left\langle\frac{Y_{batch}}{Z_{batch}}\right\rangle_{\pi}\right)=\sum_{n=1}^{T}\sum_{k=1}^{T}N_nN_k\left(\frac{1}{\kappa_n\kappa_k}-\frac{1}{\kappa_{nk}}\right)\frac{y_ny_k}{Z_{batch}^2}-\sum_{n=1}^{T}N_n\left(\frac{1}{\kappa_n^2}-\frac{1}{\kappa_{nn}}\right)\frac{y_n^2}{Z_{batch}^2}$$

$$+\sum_{n=1}^{T}N_n\left(\frac{1}{\kappa_n^2}-\frac{1}{\kappa_n}\right)\frac{y_n^2}{Z_{batch}^2} \tag{5.23}$$

Taking the last two terms together results in:

$$\hat{V}_{HT}\left(\left\langle\frac{Y_{batch}}{Z_{batch}}\right\rangle_{\pi}\right)=\sum_{n=1}^{T}\sum_{k=1}^{T}N_nN_k\left(\frac{1}{\kappa_n\kappa_k}-\frac{1}{\kappa_{nk}}\right)\frac{y_ny_k}{Z_{batch}^2}+\sum_{n=1}^{T}N_n\left(\frac{1}{\kappa_{nn}}-\frac{1}{\kappa_n}\right)\frac{y_n^2}{Z_{batch}^2} \tag{5.24}$$

In the following paragraph, in the limit of an infinite batch-to-sample size ratio (*i.e.* $r_{bs}=\infty$) expressions for the first- and second-order inclusion probabilities, κ_n and κ_{nk}, will be derived. In Paragraph 5.8, these expressions will be used to evaluate Equation 5.20 and Equation 5.24.

5.7 First- and second-order inclusion probabilities

The π-expanded estimator for Y_{batch}/Z_{batch} and the Horvitz-Thompson estimator for its variance, given by Equation 5.20 and Equation 5.24 respectively, depend on the first- and second-order inclusion probabilities κ_n and κ_{nk} for all integer values of n and k between 1 and T. In this paragraph, expressions for these inclusion probabilities are derived for the size-based approach. First, the definitions of the inclusion probabilities, Equation 5.7 and Equation 5.14 are rewritten using the indicators I_i and I_j:

$$\pi_i=\sum_{S\in U_i}P(S)=\sum_{S\in U}I_i(S)P(S)=E(I_i) \tag{5.25}$$

$$\pi_{ij} = \sum_{S \in U_{ij}} P(S) = \sum_{S \in U} I_i(S) I_j(S) P(S) = E\left(I_i I_j\right) \tag{5.26}$$

where i may not be equal to j. The results of Chapter 4 can be used for evaluation of the above two expressions. For this purpose, the following expression for N_k is used:

$$N_k = \sum_{r=1}^{N_{batch}} \Delta_{kn(r)} I_r \tag{5.27}$$

for all k between 1 and T. The summation on the right-hand side of the above equation represents a count of particles sampled, which belong to the k^th^ class. While the factor $\Delta_{kn(r)}$ guarantees that only particles belonging to the k^th^ class are counted, the factor I_r guarantees that only particles sampled are counted. To find an expression for the first-order inclusion probability the expected value of Equation 5.27 is calculated. Taking the expected value of both sides of Equation 5.27 gives:

$$E\left(N_k\right) = \sum_{r=1}^{N_{batch}} \Delta_{kn(r)} E\left(I_r\right) \tag{5.28}$$

Substituting Equation 5.25 into the above equation yields the following relation containing the first-order inclusion probability:

$$E\left(N_k\right) = \sum_{r=1}^{N_{batch}} \Delta_{kn(r)} \pi_r \tag{5.29}$$

Using the definition of κ_k, the above equation can be written as:

$$E\left(N_k\right) = \sum_{r=1}^{N_{batch}} \Delta_{kn(r)} \pi_r = \kappa_k \sum_{r=1}^{N_{batch}} \Delta_{kn(r)} = \kappa_k N_{k,batch} \tag{5.30}$$

An expression for the left hand-side of the above equation, the expected value of N_n, was calculated in Chapter 4 in the limit of $N_{batch}=\infty$, or equivalently in the limit of $r_{bs}=\infty$. Taking the limit of $r_{bs}=\infty$ on both sides of the above equation results in:

$$\lim_{r_{bs}\to\infty} E\left(N_k\right) = \lim_{r_{bs}\to\infty} \left(\kappa_k N_{k,batch}\right) \tag{5.31}$$

Substituting Equation 4.61 and the definitions of $\bar{z}$ and Z^* gives:

$$\lim_{r_{bs} \to \infty} \left(\kappa_k N_{k,batch} \right) = \lim_{r_{bs} \to \infty} \left(\frac{p'_k Z^*}{\bar{z}} \right) \tag{5.32}$$

Hence, when the batch-to-sample size ratio is large, but finite, and if Z^* is approximated by Z_{sample} the first-order inclusion probability can be approximated by:

$$\kappa_k \approx \frac{p'_k Z_{sample}}{N_{k,batch} \bar{z}} = \frac{Z_{sample}}{Z_{batch}} \tag{5.33}$$

It is noted that the above equation is only an approximation and will therefore not be used further in theoretical developments. Instead, the correct result, Equation 5.32 is used.

Similarly to the above derivation, an expression for κ_{nk} can be derived. The product of N_n and N_k, for arbitrary n and k between 1 and T is written as:

$$N_n N_k = \left(\sum_{s=1}^{N_{batch}} \Delta_{nn(s)} I_s \right) \left(\sum_{r=1}^{N_{batch}} \Delta_{kn(r)} I_r \right) = \sum_{s=1}^{N_{batch}} \sum_{r=1}^{N_{batch}} \Delta_{nn(s)} \Delta_{kn(r)} I_s I_r \tag{5.34}$$

Taking the expected value of the right-hand side and of the left-hand side yields:

$$E\left(N_n N_k\right) = \sum_{s=1}^{N_{batch}} \sum_{r=1}^{N_{batch}} \Delta_{nn(s)} \Delta_{kn(r)} E\left(I_s I_r\right) \tag{5.35}$$

Only for s≠r the expected value of $I_s I_r$ is equal to π_{sr}. Therefore, the terms for which s=r are excluded from the double summation:

$$E\left(N_n N_k\right) = \sum_{s=1}^{N_{batch}} \sum_{\substack{r=1 \\ r \neq s}}^{N_{batch}} \Delta_{nn(s)} \Delta_{kn(r)} E\left(I_s I_r\right) + \sum_{s=1}^{N_{batch}} \Delta_{nn(s)} \Delta_{kn(s)} E\left(I_s^2\right) \tag{5.36}$$

In the above equation, $E(I_s I_r)$ may be replaced by π_{sr} and $E\left(I_s^2\right) = E\left(I_s\right)$ may be replaced by π_s:

$$E\left(N_n N_k\right) = \sum_{s=1}^{N_{batch}} \sum_{\substack{r=1 \\ r \neq s}}^{N_{batch}} \Delta_{nn(s)} \Delta_{kn(r)} \pi_{sr} + \sum_{s=1}^{N_{batch}} \Delta_{nn(s)} \Delta_{kn(s)} \pi_s \tag{5.37}$$

The equation is slightly simplified by using the identity $\Delta_{nn(s)}\Delta_{kn(s)} = \Delta_{nn(s)}\Delta_{kn}$:

$$E\left(N_n N_k\right) = \sum_{s=1}^{N_{batch}} \sum_{\substack{r=1 \\ r\neq s}}^{N_{batch}} \Delta_{nn(s)}\Delta_{kn(r)}\pi_{sr} + \Delta_{kn} \sum_{s=1}^{N_{batch}} \Delta_{nn(s)}\pi_s \tag{5.38}$$

Using the definitions of κ_n and κ_{nk} in the above equation results in:

$$E\left(N_n N_k\right) = N_{n,batch}\left(N_{k,batch} - \Delta_{nk}\right)\kappa_{nk} + \Delta_{nk} N_{n,batch}\kappa_n \tag{5.39}$$

The terms can be rearranged and $N_{n,batch}\kappa_n = E(N_n)$, from Equation 5.30, can be substituted:

$$N_{n,batch}\left(N_{k,batch} - \Delta_{nk}\right)\kappa_{nk} = E\left(N_n N_k\right) - \Delta_{nk} E\left(N_n\right) \tag{5.40}$$

The right-hand side can be evaluated in the limit of $N_{batch}=\infty$, or equivalently in the limit of $r_{bs}=\infty$, using the results of Chapter 4. Taking the limit of $r_{bs}=\infty$ on both sides of Equation 5.40 gives:

$$\lim_{r_{bs}\to\infty}\left(N_{n,batch} N_{k,batch}\kappa_{nk}\right) = \lim_{r_{bs}\to\infty}\left(E\left(N_n N_k\right) - \Delta_{nk} E\left(N_n\right)\right) \tag{5.41}$$

It will be convenient to introduce a new parameter C_{nk} (interpretation below), defined by:

$$\kappa_{nk} = \kappa_n \kappa_k \left(1 - C_{nk}\right) \tag{5.42}$$

C_{nk} can be interpreted as a small correction necessary because the sample size is fixed. This interpretation is most clearly illustrated when all particles have the same size. The first particle of a pair has a probability of N_{sample}/N_{batch} to be included in the sample. On the condition that the first particle is in the sample, the second particle will have a probability of $(N_{sample}-1)/(N_{batch}-1) \approx (N_{sample}-1)/N_{batch}$ being included in the sample. The product of both probabilities leads to the second-order inclusion probability

$$\kappa_{nk} = \frac{N_{sample}\left(N_{sample} - 1\right)}{N_{batch}^2} = \frac{N_{sample}^2 - N_{sample}}{N_{batch}^2} = \kappa_n \kappa_k \left(1 - \frac{1}{N_{sample}}\right) \tag{5.43}$$

Hence, when all particles have the same size $C_{nk}=1/N_{sample}$. This result supports the above interpretation of C_{nk}. In the following an expression for C_{nk} will be derived which is also valid for particles of varying size. Hereto, Equation 5.42 is substituted into Equation 5.41. This gives:

$$\lim_{r_{bs}\to\infty}\left(N_{n,batch}N_{k,batch}\kappa_n\kappa_k\left(1-C_{nk}\right)\right)=\lim_{r_{bs}\to\infty}\left(E\left(N_nN_k\right)-\Delta_{nk}E\left(N_n\right)\right) \qquad \mathbf{(5.44)}$$

Using $\kappa_nN_{n,batch}=E(N_n)$ and $\kappa_kN_{k,batch}=E(N_k)$ (which follow from Equation 5.30) results in:

$$\lim_{r_{bs}\to\infty}\left(E\left(N_n\right)E\left(N_k\right)\left(1-C_{nk}\right)\right)=\lim_{r_{bs}\to\infty}\left(E\left(N_nN_k\right)-\Delta_{nk}E\left(N_n\right)\right) \qquad \mathbf{(5.45)}$$

This can be transformed into the following equation for C_{nk}:

$$\lim_{r_{bs}\to\infty}C_{nk}=\lim_{r_{bs}\to\infty}\left(\frac{\Delta_{nk}}{E\left(N_k\right)}-\frac{E\left(N_nN_k\right)-E\left(N_n\right)E\left(N_k\right)}{E\left(N_n\right)E\left(N_k\right)}\right) \qquad \mathbf{(5.46)}$$

For evaluation of the above equation, an expression for the covariance in the limit of an infinite value of r_{bs} is required. In Chapter 4, however, an expression for the covariance between the transformed variables N_n^* and N_k^* was derived. In order to use this expression here, first the relation between both covariances is derived:

$$\left.\begin{aligned}
&E\left(N_n^*N_k^*\right)-E\left(N_n^*\right)E\left(N_k^*\right)=E\left(\left(N_n+\frac{\delta p_n'}{\bar{z}}\right)\left(N_k+\frac{\delta p_k'}{\bar{z}}\right)\right)-E\left(N_n+\frac{\delta p_n'}{\bar{z}}\right)E\left(N_k+\frac{\delta p_k'}{\bar{z}}\right)\\
&=E\left(N_nN_k\right)-E\left(N_n\right)E\left(N_k\right)+E\left(\frac{\delta p_n'}{\bar{z}}N_k\right)+E\left(\frac{\delta p_k'}{\bar{z}}N_n\right)+E\left(\frac{\delta p_n'}{\bar{z}}\frac{\delta p_k'}{\bar{z}}\right)\\
&-E\left(\frac{\delta p_n'}{\bar{z}}\right)E\left(N_k\right)-E\left(\frac{\delta p_k'}{\bar{z}}\right)E\left(N_n\right)-E\left(\frac{\delta p_n'}{\bar{z}}\right)E\left(\frac{\delta p_k'}{\bar{z}}\right)
\end{aligned}\right\} \qquad \mathbf{(5.47)}$$

From the above equation follows, after some derivations, that:

$$E\left(N_n^* N_k^*\right) - E\left(N_n^*\right)E\left(N_k^*\right) = E\left(N_n N_k\right) - E\left(N_n\right)E\left(N_k\right) +$$
$$E\left(N_n\right)E\left(N_k\right)E\left(\frac{\delta}{Z^*}\left(\frac{N_k - E\left(N_k\right)}{E\left(N_k\right)} + \frac{N_n - E\left(N_n\right)}{E\left(N_n\right)} + \frac{\delta - E(\delta)}{Z^*}\right)\right) \tag{5.48}$$

This equation can be substituted into Equation 5.46:

$$\lim_{r_{bs}\to\infty} C_{nk} = \lim_{r_{bs}\to\infty}\left(\frac{\Delta_{nk}}{E\left(N_k\right)} - \frac{E\left(N_n^* N_k^*\right) - E\left(N_n^*\right)E\left(N_k^*\right)}{E\left(N_n\right)E\left(N_k\right)}\right.$$
$$\left. + E\left(\frac{\delta}{Z^*}\left(\frac{N_k - E\left(N_k\right)}{E\left(N_k\right)} + \frac{N_n - E\left(N_n\right)}{E\left(N_n\right)} + \frac{\delta - E(\delta)}{Z^*}\right)\right)\right) \tag{5.49}$$

Using the expression for the expected value and the expression for the covariance, Equation 4.87, the expression for the expected value, Equation 4.61, and the definition of $\bar{z}$ results in:

$$\lim_{r_{bs}\to\infty} C_{nk} = \lim_{r_{bs}\to\infty}\left(-\frac{\bar{z}}{p_k' Z^*}\left[\sum_{\substack{r=1\\ r\neq n}}^{T} p_r' z_r \frac{\bar{z} - p_n'\left(z_n - z_r\right)}{\bar{z}^2} - 1 + \frac{p_n' z_n - p_n' z_k}{\bar{z}}\right]\right.$$
$$\left. + E\left(\frac{\delta}{Z^*}\left(\frac{N_k - E\left(N_k\right)}{E\left(N_k\right)} + \frac{N_n - E\left(N_n\right)}{E\left(N_n\right)} + \frac{\delta - E(\delta)}{Z^*}\right)\right)\right) \tag{5.50}$$

The first term on the right-hand side of the above equation,

$$-\frac{\bar{z}}{p_k' Z^*}\left[\sum_{\substack{r=1\\ r\neq n}}^{T} p_r' z_r \frac{\bar{z} - p_n'\left(z_n - z_r\right)}{\bar{z}^2} - 1 + \frac{p_n' z_n - p_n' z_k}{\bar{z}}\right] \tag{5.51}$$

is inversely proportional to the sample size Z^* and hence converges as $1/Z^*$ towards zero as the sample size increases. On the other hand, the second-term on the right-hand side, denoted simply as γ (for reasons explained later),

$$\gamma = E\left(\frac{\delta}{Z^{*}}\left(\frac{N_k - E(N_k)}{E(N_k)} + \frac{N_n - E(N_n)}{E(N_n)} + \frac{\delta - E(\delta)}{Z^{*}}\right)\right) \tag{5.52}$$

will converge more rapidly than $1/Z^{*}$ towards zero, because $(N_k-E(N_k))/E(N_k)$, $N_n-E(N_n))/E(N_n)$ and $(\delta-E(\delta))/Z^{*}$ also approach zero with increasing sample size. Therefore, the first term will be dominant at large sample sizes and the precise structure of the complicated second term will not be relevant in the following derivations and was, therefore, denoted as γ. This results in:

$$\lim_{r_{bs}\to\infty} C_{nk} = \lim_{r_{bs}\to\infty}\left(-\frac{\bar{z}}{p'_k Z^{*}}\left[\sum_{\substack{r=1\\ r\neq n}}^{T} p'_r\, z_r\, \frac{\bar{z} - p'_n\left(z_n - z_r\right)}{\bar{z}^2} - 1 + \frac{p'_n\, z_n - p'_n\, z_k}{\bar{z}} + \gamma\right]\right) \tag{5.53}$$

The above equation can be written as:

$$\lim_{r_{bs}\to\infty} C_{nk} = \lim_{r_{bs}\to\infty}\left(\frac{z_n + z_k}{Z^{*}} - \sum_{r=1}^{T} p'_r z_r^2 \Big/ \left(Z^{*}\bar{z}\right) + \gamma\right) \tag{5.54}$$

A practical problem with application of the obtained expressions for κ_k and κ_{nk} (Equation 5.32 and 5.42 respectively) is that batch information is required: the values of $E(\delta)$ (because $Z^{*}=Z-E(\delta)$), p'_k, and C_{nk}. When the expression for the first-order inclusion probability κ_k is used for calculation of the π-expanded estimator and the Horvitz-Thompson estimator in paragraph 5.8, it will be seen that the associated factor p'_k cancels. This is not the case for $E(\delta)$ and C_{nk}. Fortunately, it will be seen that in the limit of an infinite sample-to-particle size ratio, r_{sp}, $E(\delta)$ will becomes negligibly small and C_{nk} can be approximated by:

$$C_{nk} \approx C_{nk}^{\text{sample}} \tag{5.55}$$

in which C_{nk}^{sample} is defined as:

$$C_{nk}^{\text{sample}} = \frac{z_n + z_k}{Z_{\text{sample}}} - \frac{1}{Z_{\text{sample}}^2}\sum_{r=1}^{T} N_r z_r^2 \tag{5.56}$$

which is based solely on sample information. Therefore, in the next paragraph, the limit of an infinite sample-to-particle size ratio is taken in order to exploit these useful approximations.

5.8 Substitution of first- and second-order inclusion probabilities

Because in the previous paragraph expressions for the first- and second-order inclusion probability were derived in the limit of $r_{bs}=\infty$, in this paragraph, the limit of $r_{bs}=\infty$ is taken of the π-expanded estimator and the Horvitz-Thompson estimator. First, in the expression for the π-expanded estimator, Equation 5.20, the limit of $r_{bs}=\infty$ is taken on both sides:

$$\left.\begin{aligned}
&\lim_{r_{bs}\to\infty}\left\langle Y_{batch}/Z_{batch}\right\rangle_{\pi} = \lim_{r_{bs}\to\infty}\sum_{n=1}^{T}\frac{N_n y_n}{Z_{batch}\kappa_n} = \lim_{r_{bs}\to\infty}\sum_{n=1}^{T}\frac{p'_n N_n y_n}{\bar{z} N_{n,batch}\kappa_n} \\
&= \sum_{n=1}^{T}\frac{p'_n N_n y_n}{\bar{z}\lim\limits_{r_{bs}\to\infty}\left(N_{n,batch}\kappa_n\right)} = \sum_{n=1}^{T}\frac{p'_n N_n y_n}{\bar{z}\lim\limits_{r_{bs}\to\infty}\left(p'_n Z^{*}/\bar{z}\right)} \\
&= \lim_{r_{bs}\to\infty}\sum_{n=1}^{T}\frac{N_n y_n}{Z^{*}} = \lim_{r_{bs}\to\infty}\left(Y_{sample}/Z^{*}\right)
\end{aligned}\right\}\qquad(5.57)$$

in which Equation 5.32 was used. The result obtained

$$\lim_{r_{bs}\to\infty}\left\langle Y_{batch}/Z_{batch}\right\rangle_{\pi} = \lim_{r_{bs}\to\infty}\left(Y_{sample}/Z^{*}\right)\qquad(5.58)$$

indicates that, in the limit of an infinite batch-to-sample size ratio, the π-expanded estimator is equal to the sample concentration if Z^{*} can be replaced by Z_{sample}. Fortunately, this is correct in the limit of an infinite sample-to-particle size ratio. As discussed in Paragraph 4.7, the limit of $r_{sp}=\infty$ and $r_{bs}=\infty$ may be interchanged. Hence, taking the limit of $r_{sp}=\infty$ on both sides of the above equation results in:

$$\begin{aligned}
&\lim_{r_{sp}\to\infty}\lim_{r_{bs}\to\infty}\left\langle Y_{batch}/Z_{batch}\right\rangle_{\pi} = \lim_{r_{sp}\to\infty}\lim_{r_{bs}\to\infty}\left(\frac{Y_{sample}}{Z^{*}}\right) = \lim_{r_{bs}\to\infty}\lim_{r_{sp}\to\infty}\left(\frac{Y_{sample}}{Z^{*}}\right) \\
&= \lim_{r_{bs}\to\infty}\left(\frac{Y_{sample}}{Z_{sample}}\right) = \frac{Y_{sample}}{Z_{sample}}
\end{aligned}\qquad(5.59)$$

Hence, in the limit of both an infinite sample-to-particle size ratio and an infinite batch-to-sample size ratio, the π-expanded estimator is equal to the sample concentration. Similarly to the above derivation, the limit of $r_{bs}=\infty$ is taken on both sides on the equation for the Horvitz-Thompson estimator, Equation 5. 24:

$$\lim_{r_{bs}\to\infty} \hat{V}_{HT}\left(\left\langle \frac{Y_{batch}}{Z_{batch}} \right\rangle_{\pi}\right) = \lim_{r_{bs}\to\infty}\left[\sum_{n=1}^{T}\sum_{k=1}^{T} N_n N_k \left(\frac{1}{\kappa_n \kappa_k} - \frac{1}{\kappa_{nk}}\right)\frac{y_n y_k}{Z_{batch}^2} + \sum_{n=1}^{T} N_n \left(\frac{1}{\kappa_{nn}} - \frac{1}{\kappa_n}\right)\frac{y_n^2}{Z_{batch}^2}\right] \quad \textbf{(5.60)}$$

Substituting the identities $Z_{batch} = \bar{z}N_{n,batch}/p'_n$ and $Z_{batch} = \bar{z}N_{k,batch}/p'_k$ into Equation 5.60 results in:

$$\left.\begin{aligned}
&\lim_{r_{bs}\to\infty} \hat{V}_{HT}\left(\left\langle \frac{Y_{batch}}{Z_{batch}} \right\rangle_{\pi}\right) = \\
&\lim_{r_{bs}\to\infty} \sum_{n=1}^{T}\sum_{k=1}^{T} N_n N_k \left[\frac{p'_n p'_k y_n y_k}{\bar{z}^2}\left(\frac{1}{\left(N_{n,batch}\kappa_n\right)\left(N_{k,batch}\kappa_k\right)} - \frac{1}{N_{n,batch}N_{k,batch}\kappa_{nk}}\right)\right] \\
&+ \lim_{r_{bs}\to\infty} \sum_{n=1}^{T} N_n \left(\frac{1}{N_{n,batch}^2\kappa_{nn}} - \frac{1}{N_{n,batch}\left(N_{n,batch}\kappa_n\right)}\right)\frac{p_n'^2 y_n^2}{\bar{z}^2}
\end{aligned}\right\} \quad \textbf{(5.61)}$$

Equations 5.42 can be substituted into the above equation to eliminate the second-order inclusion probability. The result is:

$$\left.\begin{aligned}
&\lim_{r_{bs}\to\infty} \hat{V}_{HT}\left(\left\langle \frac{Y_{batch}}{Z_{batch}} \right\rangle_{\pi}\right) = \\
&\lim_{r_{bs}\to\infty} \sum_{n=1}^{T}\sum_{k=1}^{T} N_n N_k \left[\frac{p'_n p'_k y_n y_k}{\bar{z}^2}\frac{1}{\left(N_{n,batch}\kappa_n\right)\left(N_{k,batch}\kappa_k\right)}\left(1 - \frac{1}{\left(1 - C_{nk}\right)}\right)\right] \\
&+ \lim_{r_{bs}\to\infty} \sum_{n=1}^{T} N_n \left(\frac{1}{N_{n,batch}^2\kappa_n^2\left(1 - C_{nn}\right)} - \frac{1}{N_{n,batch}\left(N_{n,batch}\kappa_n\right)}\right)\frac{p_n'^2 y_n^2}{\bar{z}^2}
\end{aligned}\right\} \quad \textbf{(5.62)}$$

In order to eliminate the first-order inclusion probabilities, Equation 5.32 can be substituted into the above equation. The result is:

$$\lim_{r_{bs}\to\infty} \hat{V}_{HT}\left(\left\langle \frac{Y_{batch}}{Z_{batch}} \right\rangle_{\pi}\right) = \lim_{r_{bs}\to\infty}\left[\sum_{n=1}^{T}\sum_{k=1}^{T} N_n N_k \frac{y_n y_k}{\left(Z^*\right)^2}\left(1-\frac{1}{\left(1-C_{nk}\right)}\right)+\sum_{n=1}^{T} N_n \frac{y_n^2}{\left(1-C_{nn}\right)\left(Z^*\right)^2}\right] \quad \textbf{(5.63)}$$

From Equation 5.54 it follows that, with increasing sample size, C_{nk} and C_{nn} in the above equation converge towards zero as $1/Z^*$. Note that when all particles have the same size $C_{nk}=1/N_{sample}$ for all n and k between 1 and T (see interpretation below Equation 5.43). Therefore, the Taylor-expansion around $1/N_{sample}$

$$1-\frac{1}{1-C_{nk}} = \frac{-C_{nk}}{1-1/N_{sample}-C_{nk}+1/N_{sample}} = \frac{-C_{nk}}{1-1/N_{sample}}\left(\frac{1}{1-\theta_{nk}}\right) = \frac{-C_{nk}}{1-1/N_{sample}}\left(1+\theta_{nk}+\theta_{nk}^2+\ldots\right) \quad \textbf{(5.64)}$$

where

$$\theta_{nk} = \frac{C_{nk}-1/N_{sample}}{1-1/N_{sample}} \quad \textbf{(5.65)}$$

for all values of n and k between 1 and T, and a similar expansion for C_{nn}:

$$\frac{1}{1-C_{nn}} = \frac{1}{1-1/N_{sample}-C_{nn}+1/N_{sample}} = \frac{1}{1-1/N_{sample}}\left(\frac{1}{1-\theta_{nn}}\right) = \frac{1}{1-1/N_{sample}}\left(1+\theta_{nn}+\theta_{nn}^2+\ldots\right) \quad \textbf{(5.66)}$$

are substituted into Equation 5.63. This yields:

$$\begin{aligned}\lim_{r_{bs}\to\infty} \hat{V}_{HT}\left(\left\langle \frac{Y_{batch}}{Z_{batch}} \right\rangle_{\pi}\right) = & \lim_{r_{bs}\to\infty}\sum_{n=1}^{T}\sum_{k=1}^{T} N_n N_k \frac{y_n y_k}{\left(Z^*\right)^2}\frac{-C_{nk}}{1-1/N_{sample}}\left(1+\theta_{nk}+\theta_{nk}^2+\ldots\right)+ \\ & \lim_{r_{bs}\to\infty}\sum_{n=1}^{T} N_n \frac{1}{1-1/N_{sample}}\left(1+\theta_{nn}+\theta_{nn}^2+\ldots\right)\frac{y_n^2}{\left(Z^*\right)^2}\end{aligned} \quad \textbf{(5.67)}$$

In the limit of $r_{sp}=\infty$, $\theta_{nk}=0$ and $\theta_{nn}=0$. Hence, the series $\left(1+\theta_{nk}+\theta_{nk}^2+\ldots\right)$ and $\left(1+\theta_{nn}+\theta_{nn}^2+\ldots\right)$ in the above equation may be replaced by one in the limit of $r_{sp}=\infty$. Also, Z^* may be replaced by Z_{sample} and the π–expanded estimator may be replaced by the sample concentration Y_{sample}/Z_{sample} in the limit of $r_{sp}=\infty$. Therefore, the advantage of taking this limit would be that unknown values (*i.e.* the values of the π–expanded estimator, θ_{nk}, θ_{nn} and Z^*) are replaced by values known to the sampler (*i.e.* the sample concentration, zero, zero and Z_{sample} respectively). However, in the limit of $r_{sp}=\infty$,

$\hat{V}_{HT}\left(\left\langle Y_{batch}/Z_{batch}\right\rangle_{\pi}\right)=0$, which is not a very useful result. Therefore, the product of the Horvitz-Thompson estimator and Z^* is considered. This results in:

$$\lim_{r_{bs}\to\infty}\lim_{r_{sp}\to\infty}\left(Z_{sample}\hat{V}_{HT}\left(\frac{Y_{sample}}{Z_{sample}}\right)\right)=$$

$$\lim_{r_{bs}\to\infty}\lim_{r_{sp}\to\infty}\left[\sum_{n=1}^{T}\sum_{k=1}^{T}N_nN_k\frac{y_ny_k}{Z_{sample}^2}\frac{-Z_{sample}C_{nk}}{1-1/N_{sample}}+\sum_{n=1}^{T}N_n\frac{1}{1-1/N_{sample}}\frac{y_n^2}{Z_{sample}}\right] \tag{5.68}$$

In the above equation, the only remaining unknown parameter is C_{nk}. Therefore, this parameter is estimated in the limit of $r_{bs}=\infty$ and $r_{sp}=\infty$, using the following result:

$$\lim_{r_{bs}\to\infty}\lim_{r_{sp}\to\infty}Z_{sample}C_{nk}=Z_{sample}C_{nk}^{sample} \tag{5.69}$$

in which C_{nk}^{sample} is defined as:

$$C_{nk}^{sample}=\frac{z_n+z_n}{Z_{sample}}-\frac{1}{Z_{sample}^2}\sum_{r=1}^{T}N_rz_r^2 \tag{5.70}$$

The final result expressed in Equation 5.69 follows from multiplying both sides of Equation 5.54 with Z_{sample}, taking the limit of $r_{sp}=\infty$ on both sides of Equation 5.54, and from the fact that the summation in the right-hand side of Equation 5.54 can be written as a function of a ratio of batch totals $\tilde{Y}_{batch}/Z_{batch}$ if $\tilde{Y}_{batch}$ is defined as the batch total of $\tilde{y}_i=z_i^2$:

$$\lim_{r_{bs}\to\infty}\lim_{r_{sp}\to\infty}Z_{sample}C_{nk}=\lim_{r_{bs}\to\infty}\lim_{r_{sp}\to\infty}\left(\frac{Z_{sample}}{Z^*}\left(z_n+z_k\right)-\frac{Z_{sample}}{Z^*}\frac{\tilde{Y}_{batch}}{Z_{batch}}\right) \tag{5.71}$$

From Equation 4.110, it follows that in the limit of an infinite sample-to-particle size ratio and an infinite batch-to-sample size ratio $\tilde{Y}_{sample}/Z_{sample}$ is unbiased for $\tilde{Y}_{batch}/Z_{batch}$ and its variance is given by Equation 4.115 with y_i replaced by z_i^2. Because the summation in Equation 4.115 applied to $\tilde{Y}_{sample}/Z_{sample}$ is proportional to the particle size only, the variance of $\tilde{Y}_{sample}/Z_{sample}$ is zero in the limit of an infinite sample-to-particle size ratio and an infinite batch-to-sample size ratio. Hence, in the limit of an infinite sample-to-particle size ratio and infinite batch-to-sample size ratio no error is made when $\tilde{Y}_{\text{batch}}/Z_{\text{batch}}$ is replaced by $\tilde{Y}_{\text{sample}}/Z_{\text{sample}}$. In addition, no error is made when Z_{sample}/Z^{*} is replaced by one, leading to Equation 5.69. Substituting this result into Equation 5.68 yields:

$$\lim_{r_{bs}\to\infty}\lim_{r_{sp}\to\infty}\left(Z_{\text{sample}}\hat{V}_{HT}\left(\frac{Y_{\text{sample}}}{Z_{\text{sample}}}\right)\right)=$$

$$\sum_{n=1}^{T}\sum_{k=1}^{T}\left[N_nN_k\frac{y_ny_k}{Z_{\text{sample}}^2}\frac{-\left(z_n+z_n\right)+\frac{1}{Z_{\text{sample}}}\sum_{r=1}^{T}N_rz_r^2}{1-1/N_{\text{sample}}}\right]+\sum_{n=1}^{T}N_n\frac{1}{1-1/N_{\text{sample}}}\frac{y_n^2}{Z_{\text{sample}}} \tag{5.72}$$

The double summation can be written as two separate terms containing one summation each:

$$\lim_{r_{bs}\to\infty}\lim_{r_{sp}\to\infty}\left(Z_{\text{sample}}\hat{V}_{HT}\left(\frac{Y_{\text{sample}}}{Z_{\text{sample}}}\right)\right)=\frac{Y_{\text{sample}}^2}{Z_{\text{sample}}^3\left(1-1/N_{\text{sample}}\right)}\sum_{r=1}^{T}N_rz_r^2-$$

$$2\frac{Y_{\text{sample}}}{Z_{\text{sample}}\left(1-1/N_{\text{sample}}\right)}\sum_{n=1}^{T}N_n\frac{z_ny_n}{Z_{\text{sample}}}+\sum_{n=1}^{T}N_n\frac{y_n^2}{Z_{\text{sample}}\left(1-1/N_{\text{sample}}\right)}\Bigg) \tag{5.73}$$

Using a single summation symbol yields:

$$\lim_{r_{bs}\to\infty}\lim_{r_{sp}\to\infty}\left(Z_{\text{sample}}\hat{V}_{HT}\left(\frac{Y_{\text{sample}}}{Z_{\text{sample}}}\right)\right)=\frac{\sum_{n=1}^{T}\left(\frac{Y_{\text{sample}}^2}{Z_{\text{sample}}^2}N_nz_n^2-\frac{Y_{\text{sample}}}{Z_{\text{sample}}}2N_nz_ny_n+N_ny_n^2\right)}{\left(Z_{\text{sample}}-Z_{\text{sample}}/N_{\text{sample}}\right)}$$

$$=\frac{\sum_{n=1}^{T}N_n\left(y_n-z_n\frac{Y_{\text{sample}}}{Z_{\text{sample}}}\right)^2}{\left(Z_{\text{sample}}-Z_{\text{sample}}/N_{\text{sample}}\right)} \tag{5.74}$$

Application of the above equation results in the following estimator for the variance, based on the identities of the particles in the sample:

$$V_{sample}\left(\frac{Y_{sample}}{Z_{sample}}\right)=\frac{\sum_{n=1}^{T} N_n\left(y_n - z_n\frac{Y_{sample}}{Z_{sample}}\right)^2}{Z_{sample}\left(Z_{sample}-Z_{sample}/N_{sample}\right)} \tag{5.75}$$

For finite values of the sample-to-particle size ratio and batch-to-sample size ratio, $V_{sample}(Y_{sample}/Z_{sample})$ has a possible bias. This bias will be investigated in Chapter 6. For a sample S, the value of $V_{sample}(Y_{sample}/Z_{sample})$ is denoted as $V_{sample}(Y_{sample}/Z_{sample},S)$.

Comparison of Equation 5.75 with the equation for the variance, based on the identities of the particles in the batch, $V_{batch}(Y_{sample}/Z_{sample})=C/Z_{sample}$ (Equation 5.2), results in the following estimate for C:

$$\hat{C}_2=\frac{\sum_{n=1}^{T} N_n\left(y_n - z_n\frac{Y_{sample}}{Z_{sample}}\right)^2}{\left(Z_{sample}-Z_{sample}/N_{sample}\right)} \tag{5.76}$$

The above equation provides an estimate for the parameter of the size-variance equation, using the properties of the particles in the sample. Hence, the sixth criterion for a sampling theory is met.

5.9 Application to mass concentrations

In the mass-based approach, Z_{sample} is the sample mass, M_{sample}, and Z_{batch} is the batch mass, M_{batch}. Applying the estimators (given by Equation 5.59 and Equation 5.75) developed in the previous paragraph to mass concentration in the batch, *i.e.* substituting $y_i=a_im_i$, results in:

$$\lim_{r_{sp}\to\infty}\lim_{r_{bs}\to\infty}\left\langle A_{batch}/M_{batch}\right\rangle_{\pi}=\frac{A_{sample}}{M_{sample}}=a_{sample} \tag{5.77}$$

and

$$V_{sample}\left(a_{sample}\right)=\frac{\sum_{n=1}^{T} N_n m_n^2\left(a_n - a_{sample}\right)^2}{M_{sample}\left(M_{sample}-M_{sample}/N_{sample}\right)} \tag{5.78}$$

in which A_{sample} is the sample total of a_im_i and a_{sample} is the mass concentration in the sample. The right-hand side of Equation 5.78 gives an estimator for the variance of the mass concentration in the sample. For a sample S, the value of $V_{sample}(a_{sample})$ is denoted as $V_{sample}(a_{sample},S)$. In Chapter 6, the ranges of possible biases of both the right-hand side of Equation 5.77 and 5.78 will be calculated for a wide range of distinct batch compositions and values for the batch-to-sample size ratio and sample-to-particle size ratio.

5.10 Results

The size-variance equation, $V_{batch}(Y_{sample}/Z_{sample})=C/Z_{sample}$, contains the parameter C. Two methods were investigated to estimate this parameter. In the first method, C is estimated using an arbitrary number of samples. Equation 5.5 relates the estimate, $\hat{C}_1$, to the sample concentrations and sample sizes of an arbitrary number, N_{det}, of samples.

For the second method, the Horvitz-Thompson estimator was applied. The Horvitz-Thompson estimator can provide a general and unbiased estimate for the variance of the π-expanded estimator. To this end, the sample concentration, Y_{sample}/Z_{sample}, was rewritten in the form of a π-expanded estimator (Equation 5.16). The Horvitz-Thompson estimator was then applied for estimation of the variance of the sample concentration. Because in this study particles are classified, the behaviour of the π-expanded estimator and Horvitz-Thompson estimator under classification was investigated. Expressions for the first- and second-order inclusion probabilities (Equation 5.32 and 5.42 respectively) were derived using results from Chapter 4. These expressions were substituted into the π-expanded estimator and the Horvitz-Thompson estimator. This yielded an expression for the Horvitz-Thompson estimator for the variance of the sample concentration in the limit of an infinite batch-to-sample size ratio and an infinite sample-to-particle size ratio (Equation 5.74). From this expression an estimator, $V_{sample}(Y_{sample}/Z_{sample})$, for the variance was derived (see Equation 5.75). This result was subsequently used to derive an expression (Equation 5.76) for the estimate, $\hat{C}_2$, of the parameter of the size-variance equation

Finally, as an application of the obtained equations for the π-expanded estimator, $\langle Y_{batch}/Z_{batch}\rangle_{\pi}$, and the variance based on the properties of the particles in the sample, $V_{sample}(Y_{sample}/Z_{sample})$, the obtained equations were worked out for mass concentrations (see Equations 5.77 and 5.78).

5.11 Conclusions

Equation 5.5 relates the sample concentrations of one or more samples to the value of the parameter of the size-variance equation. Therefore, the fourth criterion of a sampling theory ("The theory must allow determination of the parameters of the size-variance equation, using the measured sample concentrations of one or more samples of a given size.") is met.

Equation 5.76 relates the properties of the particles in the sample to the value of the parameter of the size-variance equation. Therefore, the sixth criterion for a sampling theory ("The theory must allow determination of the parameters of the size-variance equation, using posterior knowledge of the properties of the particles in the sample.") is also met.

5.12 References

S. S. Cohen (1988) *Practical Statistics*, Arnold, London, 209 pp.

C. Särndal, B. Swensson and J. Wretman (1992) *Model Assisted Survey Sampling,* Springer, New York, 694 pp.

Chapter 6 Evaluation of bias

The sample concentration and the estimator for the variance, based on the properties of the particles in the sample, provide estimators for the batch concentration and the variance of the sample concentration respectively, which are unbiased under certain conditions. To investigate the behaviour, the biases are split into a contribution caused by a finite batch-to-sample size ratio and a contribution caused by a finite sample-to-particle size ratio. Only a theoretical calculation of the range of possible values of the bias in the sample concentration, caused by a finite sample-to-particle size ratio, was presented. For mass concentrations, the remaining biases were investigated using simulations.

6.1 Introduction

In Chapter 4, it was proven that during a size-based selection of particles, the sample concentration, Y_{sample}/Z_{sample}, provides an unbiased estimator for the batch concentration, Y_{batch}/Z_{batch}, in the limit of both an infinite batch-to-sample size ratio and an infinite sample-to-particle size ratio. In practice, however, both ratios are finite and hence, the sample concentration may contain a (small) bias.

The same situation occurs with the estimator for the variance of the sample concentration, $V_{sample}(Y_{sample}/Z_{sample})$, given by Equation 5.75. In Chapter 5, it was proven that in the limit of both an infinite batch-to-sample size ratio and an infinite sample-to-particle size ratio, $V_{sample}(Y_{sample}/Z_{sample})$ is equal to $\hat{V}_{HT}\left(\left\langle Y_{batch} / Z_{batch} \right\rangle_{\pi}\right)$. Because $\hat{V}_{HT}\left(\left\langle Y_{batch} / Z_{batch} \right\rangle_{\pi}\right)$ is unbiased, this implies that $V_{sample}(Y_{sample}/Z_{sample})$ is also unbiased in the limit of both an infinite batch-to-sample size ratio and an infinite sample-to-particle size ratio. In practice, however, both ratios are finite and therefore, $V_{sample}(Y_{sample}/Z_{sample})$ may contain a (small) bias.

Because mass concentrations are important for many practical applications, in this chapter, the magnitudes of the possible biases are investigated for $a_{sample}=A_{sample}/M_{sample}$ and $V_{sample}(a_{sample})$, given by Equation 5.78.

For a_{sample}, a theoretical calculation of the minimum and maximum bias caused by a finite sample-to-particle size ratio is presented in the next paragraph. While a similar calculation for $V_{sample}(a_{sample})$ would be more complicated, the magnitude of the bias in $V_{sample}(a_{sample})$ is studied using simulations with a wide range of distinct batch compositions and values for the batch-to-sample size ratio, r_{bs}, and the sample-to-particle size ratio, r_{sp}. In addition, these simulations were also performed for a_{sample}, validating the theoretical results.

6.2 Theoretical calculation of bias

The bias is divided into two contributions: the bias caused by a finite sample-to-particle size ratio and the bias caused by a finite batch-to-sample size ratio. The advantage of the subdivision of bias is that both contributions can be independently investigated. First, the bias caused by a finite sample-to-particle size ratio, $B_{sp}(\hat{x}_{batch})$, is defined as the bias of $\hat{x}_{batch}$ under sampling with replacement or, equivalently, when the batch-to-sample size ratio, r_{bs}, is infinite:

$$B_{sp}\left(\hat{x}_{batch}\right) \equiv \lim_{r_{bs}\to\infty} B\left(\hat{x}_{batch}\right) = \lim_{r_{bs}\to\infty}\left(E\left(\hat{x}_{batch}\right) - x_{batch}\right) \tag{6.1}$$

The bias caused by a finite batch-to-sample size ratio, $B_{bs}(\hat{x}_{batch})$, is defined as $B(\hat{x}_{batch}) - B_{sp}(\hat{x}_{batch})$, leading to:

$$\begin{aligned} B_{bs}\left(\hat{x}_{batch}\right) &\equiv B\left(\hat{x}_{batch}\right) - B_{sp}\left(\hat{x}_{batch}\right) = E\left(\hat{x}_{batch}\right) - x_{batch} - \lim_{r_{bs}\to\infty}\left(E\left(\hat{x}_{batch}\right) - x_{batch}\right) \\ &= E\left(\hat{x}_{batch}\right) - \lim_{r_{bs}\to\infty} E\left(\hat{x}_{batch}\right) \end{aligned} \tag{6.2}$$

It is noted that, as expected, in the limit of an infinite value of the batch-to-sample size ratio, r_{bs}, the bias caused by a finite batch-to-sample size ratio is zero. Using the above definitions, the total bias of $\hat{x}_{batch}$ can be written as:

$$B\left(\hat{x}_{batch}\right) = B\left(\hat{x}_{batch}\right) - B_{sp}\left(\hat{x}_{batch}\right) + B_{sp}\left(\hat{x}_{batch}\right) = B_{bs}\left(\hat{x}_{batch}\right) + B_{sp}\left(\hat{x}_{batch}\right) \tag{6.3}$$

An analogous equation for relative biases can be obtained (provided that x_{batch} is non-zero) when the left- and right-hand sides of the above equation are divided by x_{batch}:

$$B^{rel}(\hat{x}_{batch}) \equiv \frac{B\left(\hat{x}_{batch}\right)}{x_{batch}} = \frac{B_{bs}\left(\hat{x}_{batch}\right)}{x_{batch}} + \frac{B_{sp}\left(\hat{x}_{batch}\right)}{x_{batch}} = B_{bs}^{rel}\left(\hat{x}_{batch}\right) + B_{sp}^{rel}\left(\hat{x}_{batch}\right) \tag{6.4}$$

in which $B_{bs}^{rel}(\hat{x}_{batch}) \equiv B_{bs}(\hat{x}_{batch})/x_{batch}$ and $B_{sp}^{rel}(\hat{x}_{batch}) \equiv B_{sp}(\hat{x}_{batch})/x_{batch}$.

Theoretical expressions for the maximum and minimum values for the bias due to a finite sample-to-particle size ratio of a_{sample} can be found, using Equation 6.1, applied to a_{sample}:

$$B_{sp}\left(a_{sample}\right) \equiv \lim_{r_{bs}\to\infty} B\left(a_{sample}\right) = \lim_{r_{bs}\to\infty}\left(E\left(a_{sample}\right) - a_{batch}\right) \tag{6.5}$$

For evaluation of the right-hand side of the above equation, results from Chapter 4 can be substituted. For this, first a_{sample} is rewritten as:

$$a_{sample} = A_{sample}/M_{sample} = A_{sample}/(M-\delta) \tag{6.6}$$

Another useful equation is obtained when Equation 4.100 is applied to A_{sample}:

$$\lim_{r_{bs}\to\infty} E\left(A_{sample} + \delta a_{batch}\right) = \sum_{i=1}^{T} \frac{p_i' M}{\overline{m}} a_i m_i = M a_{batch} \tag{6.7}$$

where $\overline{m}$ is the mean particle mass in the batch. This can also be written as:

$$\lim_{r_{bs}\to\infty} E\left(A_{sample}\right) = \lim_{r_{bs}\to\infty} (M - E(\delta)) a_{batch} \tag{6.8}$$

In the following, Equation 6.6 and 6.8 are used to find expressions for the maximum and minimum bias. First, Equations 6.6 is substituted into Equation 6.5:

$$B_{sp}\left(a_{sample}\right) = \lim_{r_{bs}\to\infty} E\left(\frac{A_{sample}}{M-\delta} - a_{batch}\right) \tag{6.9}$$

Because the value of δ varies between zero and $-m_{max}$, a lower limit for the right-hand side is obtained when δ is replaced by $-m_{max}$:

$$B_{sp}\left(a_{sample}\right) \geq \lim_{r_{bs}\to\infty} E\left(\frac{A_{sample}}{M+m_{max}} - a_{batch}\right) \tag{6.10}$$

Substituting Equation 6.8 into the above equation results in:

$$B_{sp}(a_{sample}) \geq \lim_{r_{bs} \to \infty} E\left(\frac{(M - E(\delta))a_{batch}}{M + m_{max}} - a_{batch} \right) \qquad (6.11)$$

A lower limit for the right-hand side of the above inequality is obtained when E(δ) is replaced by zero. This is then also a lower limit for the bias caused by a finite sample-to-particle size ratio:

$$B_{sp}\left(a_{sample}\right) \geq \lim_{r_{bs} \to \infty} E\left(\frac{Ma_{batch}}{M + m_{max}} - a_{batch} \right) = \frac{-m_{max}}{M + m_{max}} a_{batch} \geq \frac{-m_{max}}{M} a_{batch} \qquad (6.12)$$

From the above inequality, it follows that $B_{sp}^{rel}(a_{sample})$ cannot be smaller than minus the inverse of the sample-to-particle size ratio, if size is defined as a mass.

Similarly to the above derivation, an upper limit can be obtained. For this, in Equation 6.9, δ is replaced by zero:

$$B_{sp}\left(a_{sample}\right) \leq \lim_{r_{bs} \to \infty} E\left(\frac{A_{sample}}{M} - a_{batch} \right) \qquad (6.13)$$

Equation 6.8 is substituted into the above inequality:

$$B_{sp}\left(a_{sample}\right) \leq \lim_{r_{bs} \to \infty} E\left(\frac{(M - E(\delta))a_{batch}}{M} - a_{batch} \right) \qquad (6.14)$$

An upper limit for the right-hand side of the above inequality is obtained when δ is replaced by $-m_{max}$. This is then also an upper limit for the bias caused by a finite sample-to-particle size ratio:

$$B_{sp}\left(a_{sample}\right) \leq \lim_{r_{bs} \to \infty} E\left(\frac{(M + m_{max})a_{batch}}{M} - a_{batch} \right) = \frac{m_{max}}{M} a_{batch} \qquad (6.15)$$

From the above equation, it follows that $B_{sp}^{rel}(a_{sample})$ cannot exceed the inverse of the sample-to-particle size ratio, if size is defined as a mass.

From the above-obtained results, Equation 6.12 and 6.15, it follows that the inverse of the sample-to-particle size ratio is an upper limit for the absolute value of the relative bias:

$$\left|B_{sp}^{rel}\left(a_{sample}\right)\right| \leq \frac{m_{max}}{M} \tag{6.16}$$

Hence, for a_{sample}, the relative bias due to a finite sample-to-particle size ratio is between $-m_{max}/M$ and m_{max}/M.

6.3 Simulations

In the previous paragraph, the range of possible values of the bias caused by a finite sample-to-particle size ratio for a_{sample} was calculated. It is, however, more complicated to derive theoretical expressions for the ranges of possible values for the other biases: the bias due to a finite batch-to-sample size ratio of a_{sample} and the biases in the variance estimator, $V_{sample}(a_{sample})$, given by Equation 5.78. To investigate the magnitudes of the biases caused by a finite sample-to-particle size ratio and a finite batch-to-sample size ratio, the probability to draw a sample with n_j particles of class j, for j ranging between 1 and T, denoted as $P(S|N_1(S)=n_1,\ldots,N_T(S)=n_T)$, is calculated for all possible sample compositions and for many distinct batches (20250), covering a broad range of distinct compositions of batch (each batch with given values of p_i' , m_i and a_i). Three classes of particles are defined with the following compositions:

- Particle masses m_1, m_2, and m_3 ranging from 0.9 g, 1.0 g, 1.1 g, 1.2 g, to 1.3 g (5^3 possibilities),
- Numerical fractions of particles p_1' , p_2' , and p_3' chosen from 0.2, 0.4, 0.6 (satisfying $p_1' + p_2' + p_3' = 1$) (3+2+1=6 possibilities).
- Particle concentrations a_1, a_2, and a_3 chosen from 0.0, 0.5 and 1.0 (3^3 possibilities).

Hence, 20250 ($=5^3\times6\times3^3$) different batches were considered. Because of the large number of investigated batches, it is expected that the obtained maximum and minimum value for the relative biases are indicative for batches with similar sample-to-particle size ratio and batch-to-sample size ratio.

First, the bias caused by a finite sample-to-particle size ratio, $B_{sp}\left(\hat{x}_{batch}\right)$, was investigated. Hence, it is assumed that sampling is with replacement. The order in which the probabilities of going to the end-points of the multinomial tree were calculated is illustrated in Figure 6.1.

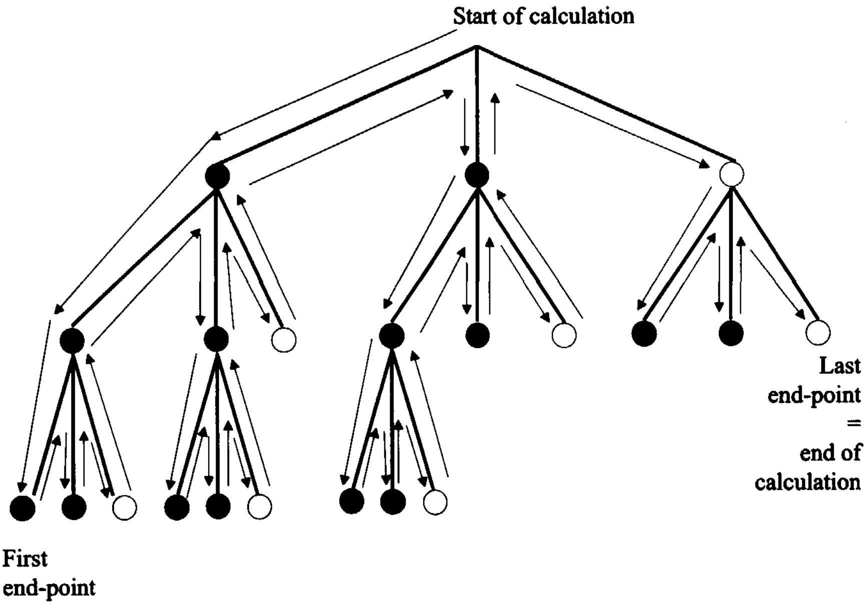


Figure 6.1. *Example of a sequence in which the probabilities of reaching the different end-points of the multinomial tree were calculated, as indicated by the arrows. The depicted multinomial tree is similar to the tree depicted in Figure 3.9.*

The scheme to calculate the probabilities of reaching each possible end-point allowed a calculation of $P(S|N_1(S)=n_1,...,N_T(S)=n_T)$. Hence, the bias of an estimator can be calculated (using $B(\hat{x}_{batch})=E(\hat{x}_{batch})-x_{batch}$ and Equation 3.14). In the following, $\hat{x}_{batch}$ represents either the estimator for the mass concentration in the batch, a_{sample}, or the estimator for the variance of the mass concentration in the sample, $V_{sample}(a_{sample})$, as given by Equation 5.78. For the investigated batches, in Figure 6.2, the maximum and minimum biases caused by a finite sample-to-particle size ratio are plotted as a function of the boundary value of the sample mass.

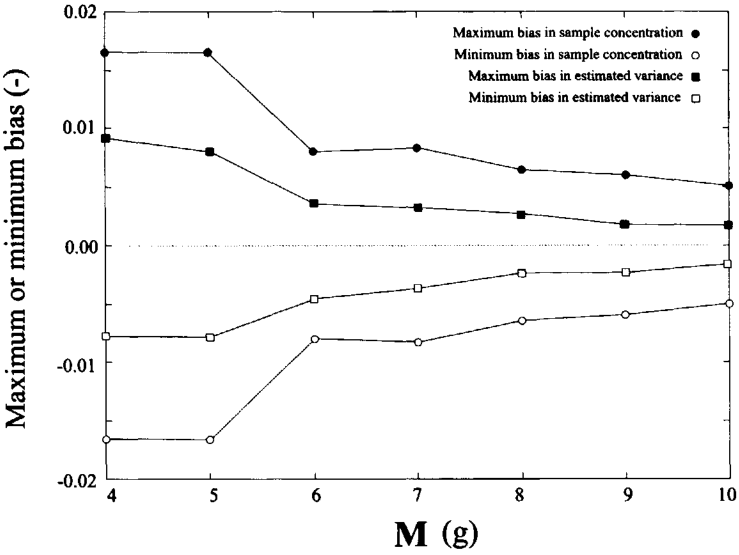


***Figure 6.2.** Maximum and minimum biases caused by a finite sample-to-particle size ratio of the proposed estimators for the mass concentration in the batch and the variance of the mass concentration. Three types of particles were distinguished. For each boundary value of the sample mass the biases were calculated. The maximum and minimum are the maximum and minimum bias respectively that occurs for the 20250 settings indicated in the text.*

From Figure 6.2, it can be seen that the maximum and minimum bias caused by a finite sample-to-particle size ratio tend to become zero if the boundary value of the sample mass becomes larger. Hence, for all simulated compositions the bias caused by a finite sample-to-particle size ratio approaches zero, if the sample mass increases. This must also hold for the relative bias, provided that x_{batch} is non-zero. Relative biases were calculated for the distinct batches. The 20250 distinct batches included 18000 batches with non-zero variance and 19500 with non-zero batch concentration.

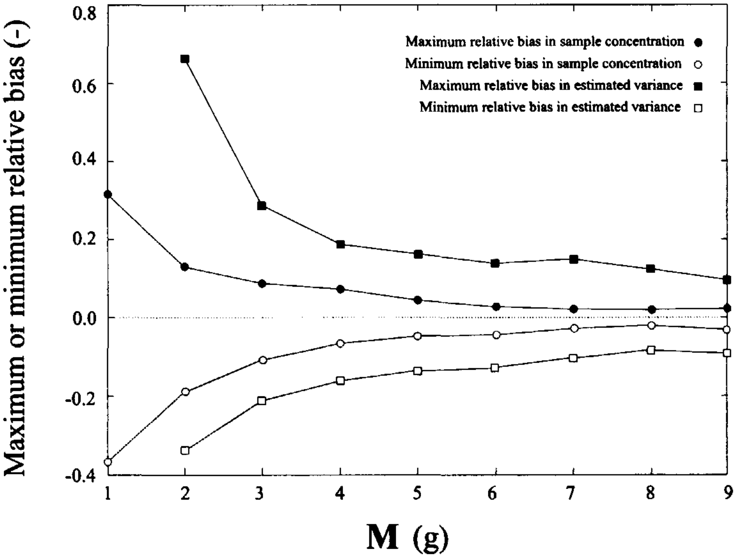


***Figure 6.3.** Maximum and minimum relative biases caused by a finite sample-to-particle size ratio of the proposed estimators for the mass concentration in the batch and the variance of the mass concentration in the sample. The same batches as in Figure 6.2 were used, with the exception of batches resulting in a zero value for the concentration or variance. At* M = *9 g the maximum and minimum biases have reduced to respectively 2.1% and –3.2% for the sample concentration and 9.4% and –9.1% for the variance.*

For all the investigated batches, the sample-to-particle size ratio, r_{sp}, (in the mass-based approach defined as the ratio of M and the maximum particle mass in the batch) cannot be larger than M/0.9. It is therefore assumed that minimum and maximum relative biases can be plotted as a function of the sample-to-particle size ratio by dividing the x-axis (M) of Figure 6.3 by 0.9. The result is plotted in Figure 6.4.

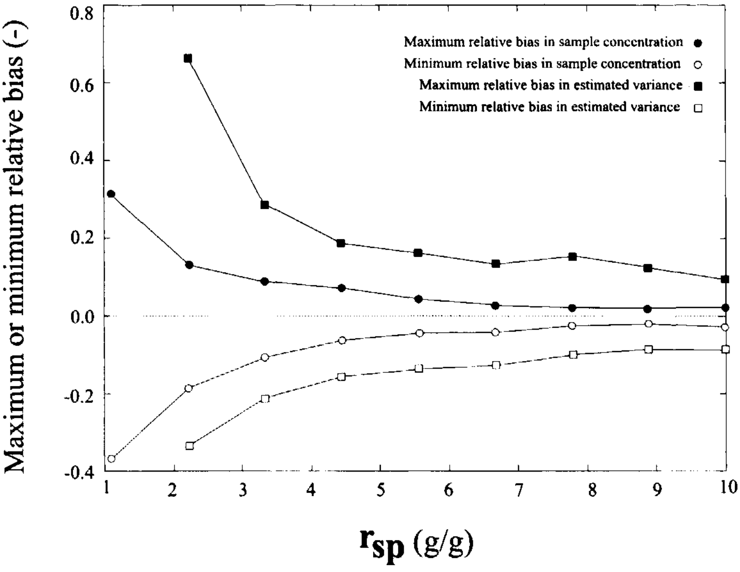


Figure 6.4. Relative bias caused by finite sample-to-particle size ratio as a function of the sample-to-particle size ratio.

It is concluded that when the sample-to-particle size ratio is 10, the biases caused by a finite sample-to-particle size ratio are certainly between −10% and +10%. Hence, it is expected that in general the bias due to a finite sample-to-particle size ratio is between −10% and +10% when the sample-to-particle size ratio is 10 or more. For the relative bias of a_{sample}, the results plotted in Figure 6.4 are consistent with Equation 6.16.

Similar calculations were made to investigate the bias due to a finite batch-to-sample size ratio. A boundary value of the sample mass M=6 g and M=7 g were chosen. The maximum and minimum biases as a function of the number of particles in the batch are depicted in Figure 6.5.

From Figure 6.5, it can be seen that the minimum and maximum biases converge to zero when the number of particles in the batch increases. However, the minimum bias in the variance caused by a finite batch-to-sample size ratio is zero. Thus, for the investigated batches, the bias in the variance is never negative. This outcome is not yet mathematically proven to hold for any arbitrary batch.

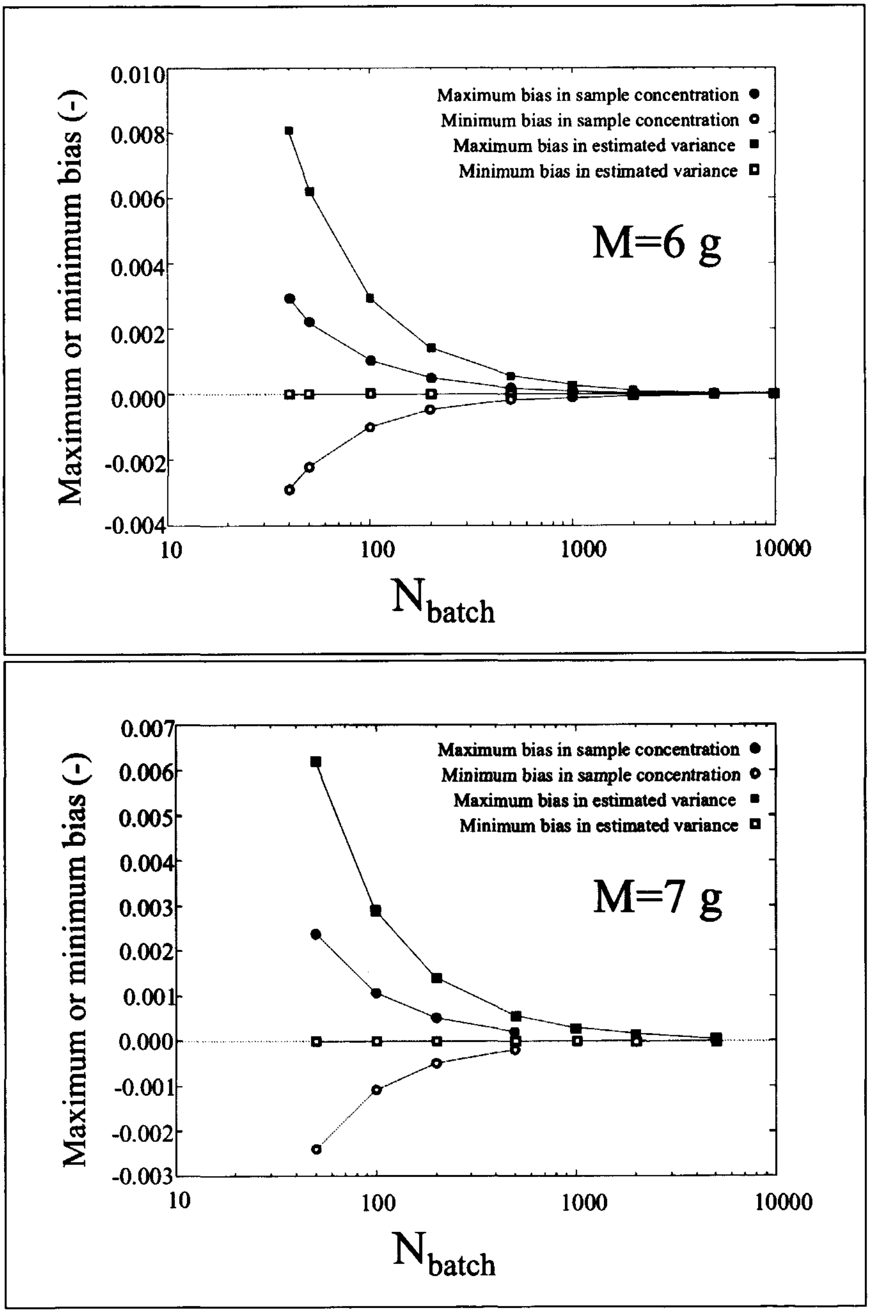


Figure 6.5. Maximum and minimum biases caused by a finite batch-to-sample size ratio of the proposed estimators for the mass concentration in the batch and the variance of the mass concentration. The same batches as in Figure 6.2 were used. The boundary value of the sample mass M ***is 6 g and 7 g.***

The same trend must hold for the relative bias of the estimators: the relative minimum bias in the variance caused by a finite batch-to-sample size ratio cannot be negative and the other relative minimum and maximum biases must converge to zero at increasing batch size. In Figure 6.6, the minimum and maximum relative biases are given as a function of the number of particles in the batch for M=6 g and M=7 g.

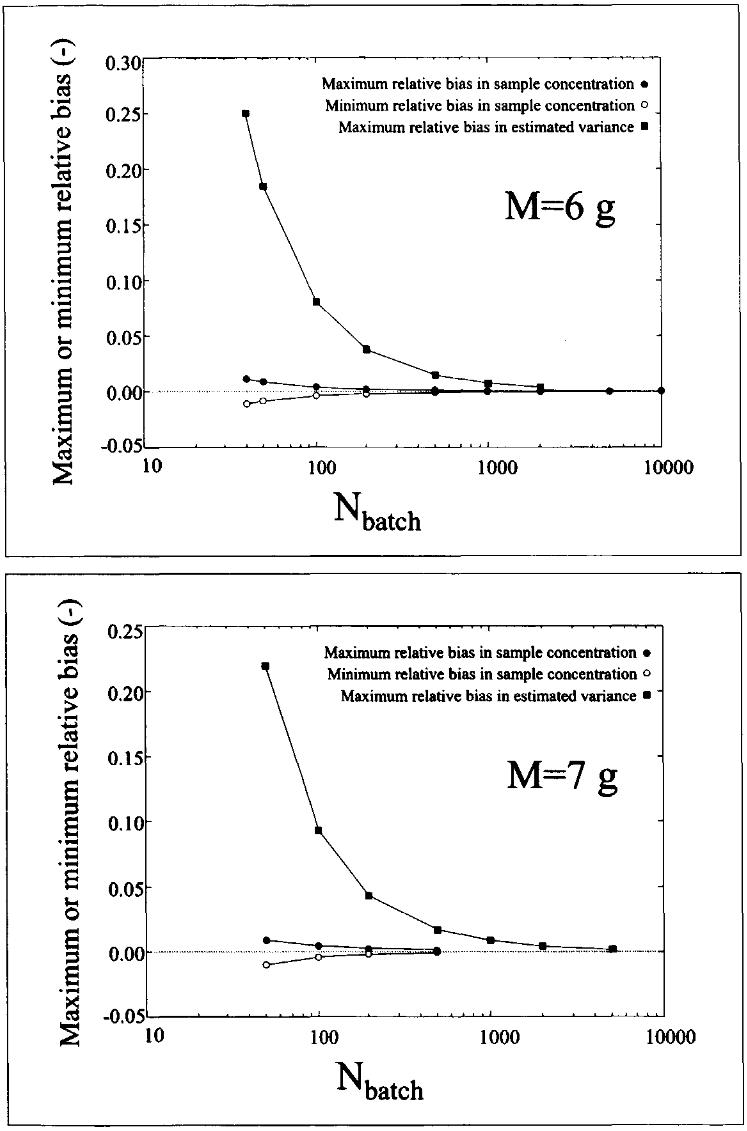


Figure 6.6. *Maximum and minimum relative biases caused by a finite batch-to-sample size ratio of the proposed estimators for the mass concentration in the batch and the variance of the mass concentration. The same batches as in Figure 6.2 were used. The boundary value of the sample mass* M *is 6 g and 7 g.*

From Figure 6.6 it is concluded that the relative bias tends to become zero with increasing batch size. This is also true for increasing batch-to-sample size ratio when this ratio is defined as the mass of the batch divided by the boundary value of the sample mass. For every value of N_{batch} in Figure 6.6 the batch-to-sample size ratio is equal to or smaller than N_{batch}×1.3/M. Therefore, it is assumed that that minimum and maximum relative biases can be plotted as a function of the batch-to-sample size ratio by multiplying the x-axis (N_{batch}) with 1.3 (the mass of the heaviest particle) and dividing by the boundary value of the sample mass (either 6 g or 7 g). Therefore, both graphs of Figure 6.6 are combined by transforming the x-axis to give the batch-to-sample size ratio. Figure 6.7 gives the maximum and minimum biases as a function of the batch-to-sample size ratio.

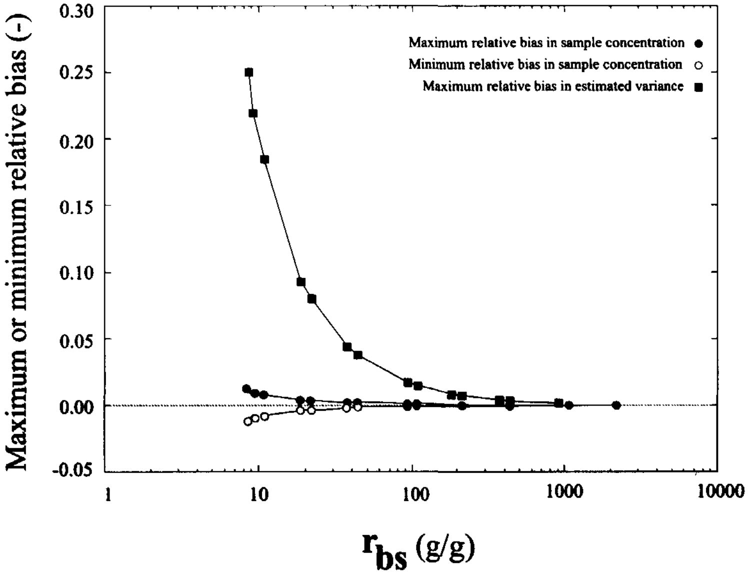


Figure 6.7. Relation between the batch-to-sample size ratio and the maximum and minimum relative biases caused by a finite batch-to-sample size ratio of the proposed estimators for the mass concentration in the batch and the variance of the mass concentration. The same batches as in Figure 6.2 were used. The boundary value of the sample mass M *is 6 g or 7 g.*

Because a broad range of batch compositions was used, the above figure can be used in practice to evaluate the possible relative biases caused by a finite batch-to-sample size ratio.

6.4 Finite population correction

The ratio Y_{sample}/Z_{sample} is a consistent estimator for the batch ratio Y_{batch}/Z_{batch}. This means that when the entire batch is selected as a sample, the estimate is precisely equal to the batch value: $Y_{sample}/Z_{sample}=Y_{batch}/Z_{batch}$. In general, it is a positive property if an estimator is consistent. However, the estimated variance of Y_{sample}/Z_{sample} (Equation 5.75) does not possess this property. Fortunately, it is possible to transform the proposed estimator into a consistent form by subtracting a finite population correction (Barnett, 1974). This correction would be equal to the estimated variance if $Z_{sample}=Z_{batch}$ and therefore does not change the estimator in the limit of an infinite batch size. The (new) estimator becomes:

$$\hat{V}_{fpc}\left(Y_{sample}/Z_{sample}\right)=\left(\frac{1}{Z_{sample}}-\frac{1}{Z_{batch}}\right)\frac{\sum_{i=1}^{T}N_i\left(y_i-z_i\frac{Y_{sample}}{Z_{sample}}\right)^2}{\left(Z_{sample}-Z_{sample}/N_{sample}\right)} \quad \textbf{(6.17)}$$

For estimating mass concentrations in the mass-based approach, the estimator becomes:

$$\hat{V}_{fpc}\left(a_{sample}\right)=\left(\frac{1}{M_{sample}}-\frac{1}{M_{batch}}\right)\frac{\sum_{i=1}^{T}N_i m_i^2\left(a_i-a_{sample}\right)^2}{\left(M_{sample}-M_{sample}/N_{sample}\right)} \quad \textbf{(6.18)}$$

The bias of $\hat{V}_{fpc}\left(a_{sample}\right)$, $B\left(\hat{V}_{fpc}\left(a_{sample}\right)\right)$, is analogous to the previous paragraph split into two contributions: the bias caused by a finite sample-to-particle size ratio, $B_{sp}\left(\hat{V}_{fpc}\left(a_{sample}\right)\right)$, and the bias caused by a finite batch-to-sample size ratio, $B_{bs}\left(\hat{V}_{fpc}\left(a_{sample}\right)\right)$. The former bias is equal to the bias caused by a finite sample-to-particle size ratio of $V_{sample}\left(a_{sample}\right)$:

$$B_{sp}\left(\hat{V}_{fpc}\left(a_{sample}\right)\right)\equiv B_{sp}\left(V_{sample}\left(a_{sample}\right)\right) \quad \textbf{(6.19)}$$

The bias caused by a finite batch-to-sample size ratio is thus defined as the remaining contribution to the total bias:

$$B_{bs}\left(\hat{V}_{fpc}\left(a_{sample}\right)\right)\equiv B\left(\hat{V}_{fpc}\left(a_{sample}\right)\right)-B_{sp}\left(V_{sample}\left(a_{sample}\right)\right) \quad \textbf{(6.20)}$$

This leads to

$$B\left(\hat{V}_{fpc}\left(a_{sample}\right)\right)=B_{bs}\left(\hat{V}_{fpc}\left(a_{sample}\right)\right)+B_{sp}\left(V_{sample}\left(a_{sample}\right)\right) \quad \textbf{(6.21)}$$

Using similar calculations as in the previous paragraph, the maximum and minimum relative biases in $\hat{V}_{fpc}\left(a_{sample}\right)$, caused by a finite batch-to-sample size ratio, were calculated for several values of N_{batch}. Similarly to the procedure in the previous paragraph, each value of N_{batch} is transformed into a value for r_{sp} by multiplying N_{batch} with 1.3 and dividing by M. This allows to construct Figure 6.8, where the maximum and minimum relative biases caused by a finite batch-to-sample size ratio of the variance estimator given by Equation 6.18 (with finite population correction) and the maximum relative bias caused by a finite batch-to-sample size ratio of the variance estimator given by 5.78 (without finite population correction) are plotted versus the batch-to-sample size ratio. The relative bias of the estimator given in Equation 5.78 is potentially larger than the relative bias of the estimator given in Equation 6.18.

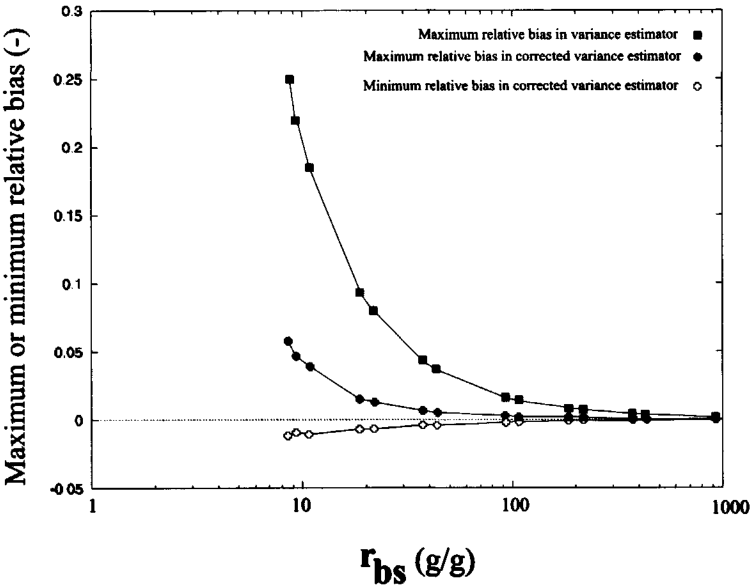


Figure 6.8. Maximum and minimum relative biases of the variances estimated using Equation 5.78 and the corrected estimator Equation 6.19 versus the batch-to-sample size ratio. The same batches as in Figure 6.2 were used. The boundary value of the sample mass M *is 6 g or 7 g.*

Because the finite population correction in Equation 6.18 is positive, the corrected variance estimator has a lower bias than the uncorrected estimator. For small batches, the maximum possible relative bias increases sharply for the uncorrected variance estimator, while for the corrected estimator this bias increases only mildly. For large batches, it can be seen that biases in both estimators become similar. For batch-to-sample size ratio equal to 10, the bias of the corrected variance is between –5% and 5%. For the uncorrected estimator, this range is attained only at a batch-to-sample size ratio of 30.

6.5 Evaluation of total bias

In Paragraph 6.2, the relative bias was split into two contributions: the relative bias due to a finite sample-to-particle size ratio and the relative bias due to a finite batch-to-sample size ratio. For a given sample-to-particle size ratio and a given batch-to-sample size ratio, the ranges of possible values for both biases can be obtained from Figure 6.4 and Figure 6.7 respectively. Choosing a larger sample-to-particle size ratio and a larger batch-to-sample size ratio reduces the biases. However, both ratios can be practically adjusted only by the choice of the sample mass. Choosing a larger sample mass will increase the sample-to-particle size ratio, but decrease the batch-to-sample size ratio. Conversely, choosing a smaller sample mass will increase the batch-to-sample size ratio, but lower the sample-to-particle size ratio. Hence, both ratios cannot be increased simultaneously and a (small) bias cannot be precluded.

In the following, nomograms for the maximum of the absolute value of the relative bias are constructed, which can be used to obtain one of the following three variables: the maximum absolute value of the bias, the sample-to-particle size ratio or the batch-to-sample size ratio, provided the other two are given. To facilitate the construction of these nomograms, the maximum and minimum relative biases are parameterized using simple mathematical functions; see Figure 6.9 and 6.10. These functions are chosen in such a way that the parameterization of a maximum or minimum is respectively larger or smaller than the values obtained with simulations.

From Figures 6.9 and 6.10, it can be concluded that the absolute value of the relative bias in the sample concentration is always smaller than $1/(2.4\times r_{sp})+1/(1.1\times r_{bs})$, which is always smaller than $0.5/r_{sp}+1/r_{bs}$, the expression used for the construction of the nomogram for the relative bias of a_{sample}, see Figure 6.11.

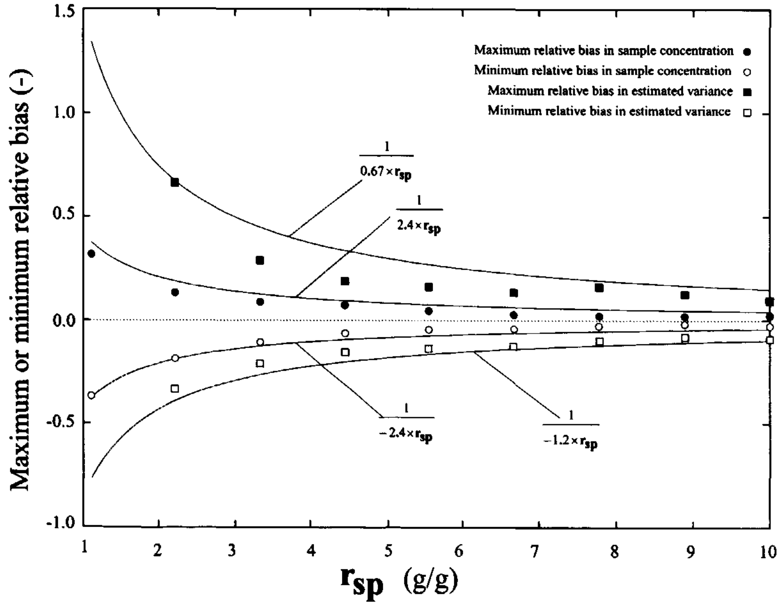


Figure 6.9. *Maximum and minimum relative biases due to a finite sample-to-particle size ratio. Also the parameterizations are shown.*

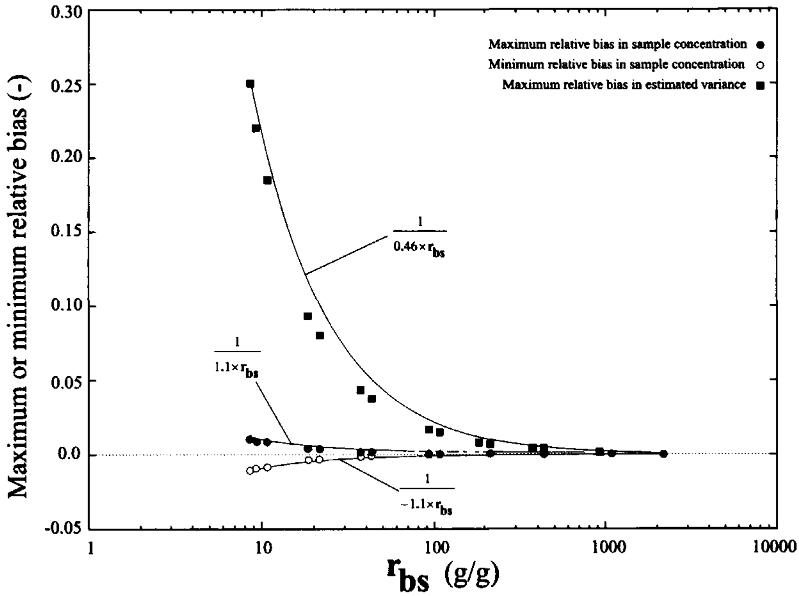


Figure 6.10. Maximum and minimum relative biases due to a finite batch-to-sample size ratio. Also the parameterizations are shown.

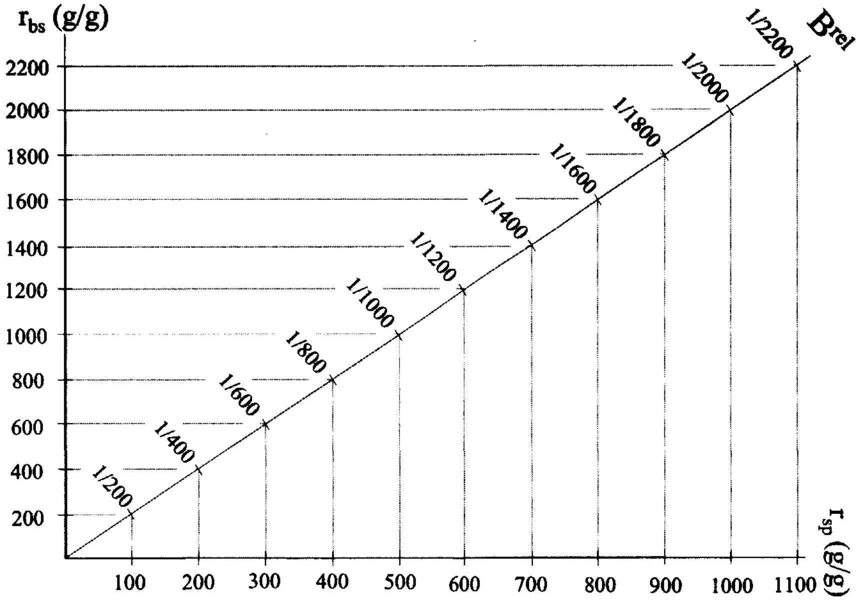


Figure 6.11. Nomogram for the maximum of the absolute value of the relative bias in the sample concentration. The intercept of a straight line through two values on distinct axes with the third axis yields the corresponding value of the third variable.

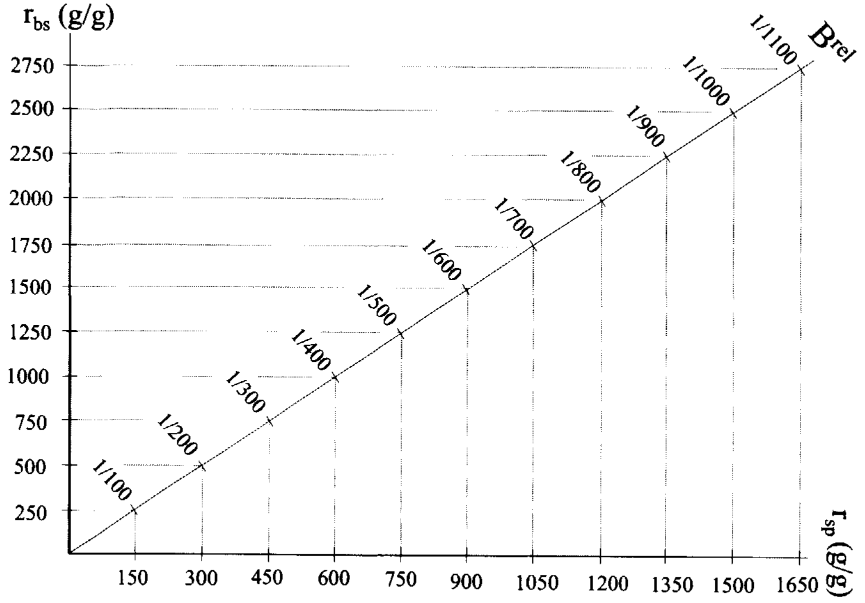


Figure 6.12. *Nomogram for the maximum of the absolute value of the relative bias in the variance of the sample concentration, calculated using Equation 5.78. The intercept of a straight line through two values on distinct axes with the third axis yields the corresponding value of the third variable.*

Analogously, from Figures 6.9 and 6.10 it can be seen that the absolute value of the relative bias in the variance estimator, $V_{sample}(a_{sample})$, is never larger than $1/(0.67\times r_{sp})+1/(0.46\times r_{bs})$, which is smaller than $1.5/r_{sp}+2.5/r_{bs}$, the expression used for construction of the nomogram for the bias of the variance estimator, Figure 6.12.

Figures 6.11 and 6.12 can be used in practice to evaluate the maximum possible bias. However, the above results are only proven valid for sample-to-particle size ratios between 1 and 10 and batch-to-sample size ratio's between 10 and 2×10^3, because for these ratio's the maximum and minimum biases were calculated in Paragraph 6.3. Therefore, further research is proposed to obtain nomograms similar to Figures 6.11 and 6.12 but with extended validity for broader ranges of the batch-to-sample size ratio and sample-to-particle size ratio.

6.6 Results

In Chapter 4, it was proven that the sample concentration, Y_{sample}/Z_{sample}, is an unbiased estimator for the batch concentration, Y_{batch}/Z_{batch}, in the limit of both an infinite sample-to-particle size ratio and an infinite batch-to-sample size ratio. In addition, in Chapter 5 a variance estimator, $V_{sample}(Y_{sample}/Z_{sample})$, was derived that is also unbiased in the limit of both an infinite sample-to-particle size ratio and an infinite batch-to-sample size ratio. In practice, both ratios are finite and hence there are possible biases. There are also possible biases for the above-mentioned estimators applied to mass concentrations in the mass-based approach. Because mass concentrations are important in numerous practical applications, in this chapter the magnitude of the possible biases in a_{sample} and $V_{sample}(a_{sample})$ were investigated as a function of both the sample-to-particle size ratio and the batch-to-sample size ratio.

In order to study the effects of a finite value of r_{sp} and r_{bs} separately, the bias was split into a bias caused by a finite sample-to-particle size ratio, B_{sp} (see Equation 6.1), and a bias caused by a finite batch-to-sample size ratio, B_{bs} (see Equation 6.2). It was demonstrated that the absolute value of the relative bias caused by a finite sample-to-particle size ratio in the mass concentration in the sample, a_{sample}, cannot exceed $1/r_{sp}$ (see Equation 6.16).

While similar calculations for the other biases would be more complicated, simulations with a wide range of distinct batch compositions were performed. Because a wide range of distinct batch compositions was used, it is expected that the obtained minimum and maximum values for the relative biases will be indicative for the sampling of other batches using a similar batch-to-sample size ratio and a similar sample-to-particle size ratio. It was verified that B_{sp}^{rel} and B_{bs}^{rel} converge towards zero as r_{sp} and r_{bs} respectively increase (see Figures 6.4 and 6.7).

For the investigated batches the relative biases caused by a finite sample-to-particle size ratio are between –10% and +10% when the sample-to-particle size ratio is 10 or more. The bias caused by a finite batch-to-sample size ratio is for the sample concentration and for the uncorrected variance estimator between -5% and 5% for a batch-to-sample size ratio of 30 or more.

A finite population correction (see Equation 6.17) was developed and investigated using simulations. Especially for small batch sizes, the proposed correction reduces the possible biases (see Figure 6.10). When it is demanded that the corrected variance estimator has a bias caused by a finite batch-to-sample size ratio between –5% and 5% the batch-to-sample size ratio must be 10 or more. Because for the uncorrected estimator this was achieved at a higher ratio of 30, this demonstrates the potential benefit of using the corrected estimator instead of the uncorrected estimator.

Finally, nomograms were obtained for the maximum of the absolute value of the relative bias in the sample concentration (Figure 6.11) and the variance estimate using the properties of the particles in the sample (Figure 6.12).

6.7 Conclusion

Nomograms are available (see Figures 6.11 and 6.12), which can be used in the practice to evaluate the maximum of the absolute value of the relative bias, given values for the sample-to-particle size ratio and the batch-to-sample size ratio.

6.8 References

V. Barnett (1974), *Elements of sampling theory*, English University Press, London, 152 pp.

Chapter 7 Experimental evaluation of the theory developed[16,17]

The estimators developed in this study and associated analytical uncertainties are evaluated for four samples. Using the experimental results, the new theory is validated by comparing the level of contradiction of the theory with the level of contradiction of the theories of Wilson and Gy. Also the normality of the sample concentration is investigated. It is shown that the new theory is less contradictory than the theories of Gy and Wilson and yields a more Gaussian estimator for the batch concentration.

7.1 Introduction

In Chapter 3, it was demonstrated that the mathematical algorithm indicated as size-based multinomial selections, provides a realistic model for the drawing of a sample from a random arrangement of particles. Based on this algorithm, theoretical results were derived in Chapter 4, 5 and 6. Notable results are expressions for estimators for the concentration in the batch and the variance of the sample concentration, which are of direct practical interest. Therefore, in this chapter, these estimators are evaluated for four samples taken from four distinct batches: a batch of wooden particles, two batches of steel slag and a batch of plastic particles.

Before this is done, the analytical uncertainty is investigated theoretically in Paragraph 7.2. When all particles in the sample are separately analyzed, the estimates for the concentration in the batch and for the variance depend on $N_{\mathrm{sample}}(S)$ determinations. As each determination may have a different uncertainty, the effect of analytical uncertainties in these determinations on the derived estimates must be estimated. Therefore, in the next paragraph, the effect of analytical uncertainty on the derived estimates is investigated theoretically.

[16] Parts of this chapter have been published as: B. Geelhoed, P. Bode, H. J. Glass and M. Stelling (2002) Verification of a New Sampling Theory Using INAA of Recycled Wood, *Transactions of the American Nuclear Society*, **85**, p. 425-426.

[17] Definitions of symbols given in this chapter are not included in the List of Symbols and do not form part of the consistent set of symbols with associated definitions as used for the development of the new theory in Chapter 3 to 6 and Chapter 8.

After these theoretical considerations, the sampling from a batch of wooden particles, two batches of steel slag and a batch of plastic particles are described in Paragraphs 7.3 to 7.5 respectively. For each analyzed material, analysis procedures are described and analysis results are given. These results are used to evaluate the estimators for the batch concentration and its variance and associated uncertainties.

The sampling theories of Wilson, Gy and the present study provide equations for the variance of the sample concentration. In each of these theories, the value for the variance predicted by the model equation (Equations 2.17, 2.31 and 4.115 in the theory of Wilson, Gy and this study respectively) may differ from the actual value, obtained when sampling corresponds exactly to the underlying mathematical algorithm for the drawing of a sample as used in the theory. Therefore, the level of contradiction of a sampling theory is quantitatively defined as the difference between both values for the variance. A sampling theory is internally more consistent if this difference remains small for various types of batches. Therefore, in Paragraph 7.6, the levels of contradiction of the non-empirical theories of Wilson, Gy and the mass-based approach are compared using the analytical results and the general method of bootstrapping.

Another important property of any sampling theory is the degree of normality induced in its estimators. Therefore, in Paragraph 7.7, it is demonstrated theoretically that, for a_{sample}, deviations from the normal (Gaussian) distribution are more likely to occur in the algorithms of Wilson and Gy than in the mass-based approach. To investigate this effect the technique of bootstrapping is used. As expected, it will be seen in Paragraph 7.8 that, in the theories of Wilson and Gy, the estimator for the batch concentration is less normal.

7.2 Evaluation of the analytical uncertainty

According to a guide provided by the International Organisation for Standardization (ISO, 1995) uncertainty in measurement is a parameter, associated with the result of a measurement that characterizes the dispersion of the values that could reasonably be attributed to the measurand. The standard uncertainty is the uncertainty expressed as a standard deviation. If $f(x_1,\dots,x_n)$ is a function that depends on n independent measurands x_i, the combined standard uncertainty $u_c(f)$ is defined as:

$$u_c^2(f)=\sum_{i=1}^{n}u^2(x_i)\left(\frac{\partial f(x_1,\dots,x_n)}{\partial x_i}\right)^2 \qquad \textbf{(7.1)}$$

in which $u(x_i)$ is the standard uncertainty of x_i. When estimating a_{batch}, and the corresponding variance, uncertainty lies in the determinations of the measured concentrations a_i and those of the masses m_i. Because often the masses can be determined much more accurately than the concentrations, in the following derivations, the contribution to the combined standard uncertainty of the standard uncertainties in the masses is neglected. The uncertainty of measurement in the estimated batch concentration, a_{sample}, becomes:

$$u_c^2\left(a_{sample}\right)=\sum_{i=1}^{T}u^2\left(a_i\right)\left(\frac{\partial a_{sample}}{\partial a_i}\right)^2=\sum_{i=1}^{T}u^2\left(a_i\right)\left(\frac{\partial}{\partial a_i}\sum_{j=1}^{T}\frac{N_j a_j m_j}{M_{sample}}\right)^2=\sum_{i=1}^{T}u^2\left(a_i\right)\left(\frac{N_i m_i}{M_{sample}}\right)^2 \quad (7.2)$$

The combined standard uncertainty in the estimated variance, $V_{sample}(a_{sample})$, given by Equation 5.78, results in a much more complicated equation:

$$u_c^2\left(V_{sample}\left(a_{sample}\right)\right)=\sum_{i=1}^{T}u^2\left(a_i\right)\left(\frac{\partial V_{sample}\left(a_{sample}\right)}{\partial a_i}\right)^2=$$

$$\sum_{i=1}^{T}u^2\left(a_i\right)\left(\frac{\partial}{\partial a_i}\sum_{j=1}^{T}\frac{N_j m_j^2\left(a_j-a_{sample}\right)^2}{M_{sample}\left(M_{sample}-M_{sample}/N_{sample}\right)}\right)^2 \quad (7.3)$$

The partial derivatives also act on a_{sample}:

$$\frac{\partial}{\partial a_i}a_{sample}=\frac{\partial}{\partial a_i}\sum_{j=1}^{T}\frac{N_j a_j m_j}{M_{sample}}=\frac{N_i m_i}{M_{sample}} \quad (7.4)$$

This yields the following equation for the uncertainty:

$$u_c^2\left(V_{sample}\left(a_{sample}\right)\right)=\sum_{i=1}^{T}u^2\left(a_i\right)\left(\frac{\partial V_{sample}\left(a_{sample}\right)}{\partial a_i}\right)^2$$

$$=\sum_{i=1}^{T}u^2\left(a_i\right)\left(\sum_{j=1}^{T}\frac{N_j m_j^2\,2\left(a_j-a_{sample}\right)\left(\Delta_{ij}-\frac{N_i m_i}{M_{sample}}\right)}{M_{sample}\left(M_{sample}-M_{sample}/N_{sample}\right)}\right)^2 \quad (7.5)$$

When the standard uncertainties of all a_i are equal and all particle masses are equal, the above expression is simplified. However, when the standard uncertainties and masses are allowed to vary, the above complicated expression must be used. In this chapter, the most general situation in which every particle in the batch belongs to a distinct class, *i.e.* $T=N_{batch}$ and $N_i=I_i$, is assumed for evaluation of Equation 7.2 and 7.5.

The combined standard uncertainty in the estimated standard deviation, $\sqrt{V_{sample}\left(a_{sample}\right)}$, can be related to $u_c^2\left(V_{sample}\left(a_{sample}\right)\right)$:

$$u_c^2\left(\sqrt{V_{sample}\left(a_{sample}\right)}\right)=\sum_{i=1}^{T}u^2\left(a_i\right)\left(\frac{\partial\sqrt{V_{sample}\left(a_{sample}\right)}}{\partial a_i}\right)^2$$

$$=\frac{1}{4V_{sample}\left(a_{sample}\right)}\sum_{i=1}^{T}u^2\left(a_i\right)\left(\frac{\partial V_{sample}\left(a_{sample}\right)}{\partial a_i}\right)^2=\frac{1}{4V_{sample}\left(a_{sample}\right)}u_c^2\left(V_{sample}\left(a_{sample}\right)\right) \tag{7.6}$$

To validate the applicability of the estimators a_{sample} and $V_{sample}(a_{sample})$, actual sample data must be used. Therefore, experimental sampling procedures will be given for four batches. Because mass concentrations are estimated, the mass-based approach is adopted. Equations 7.2, 7.5 and 7.6 are applied for the sampling of a batch of wooden chips, two batches of slag produced during the production of steel and a batch of recycled plastic chips in Paragraphs 7.3 to 7.5 respectively.

7.3 Wooden particles

For technical reasons, recycled wood to be incinerated in a power plant is chopped up into particles with a length up to 20 cm. In theory, when all these particles in the batch are ordered in an arbitrary sequence, a random number generator could be used for random selection of the particles. For every random selection, a random number (integer) would be generated between 1 and $N_{batch}-n_{sample}$, where n_{sample} is the total number of previously selected particles. The generated random number would determine the next particle that is selected in the following way: if the random number is x, the x^{th} particle in the arbitrary sequence would be selected. The mass-based approach could be implemented if these random selections stop if a boundary value M for the sample mass is reached or exceeded.

However, for the batch of wooden chips, with M_{batch}=1.5 kg, considered in this paragraph, the number of chips is practically too large. Therefore, a two-stage sampling procedure was followed. The batch of wooden chips was mixed thoroughly and at a few random locations in the batch subsamples were drawn. This yielded a primary sample containing 38 particles, see Figure 7.1.

Subsequently, all particles in the obtained primary sample are ordered and from this sequence particles can be selected, using random numbers and the above-described sampling design. Mathematically, this is equivalent with selecting with uniform probability particle-after-particle from the batch: The probability P that the i^{th} particle in the batch is selected for the primary sample is given by the ratio of the number of particles in the primary sample to the total number of particles in the batch:

$$P=N_{prim}\big/N_{batch} \tag{7.7}$$

in which N_{prim} is the total number of particles in the primary sample. Subsequently, if during a selection from the primary sample the i^{th} particle is in the primary sample, the probability that the i^{th} particle is selected is given by $1/N_{prim}$. The product of both probabilities yields the probability that during each selection, the i^{th} particle from the batch is selected:

$$\left(1/N_{prim}\right)\left(N_{prim}/N_{batch}\right)=1/N_{batch} \tag{7.8}$$

Hence, the sampling procedure corresponds to selections with uniform probability directly from the batch. The boundary value of the sample mass was set to 20.0 g, resulting in the selection of 18 particles.

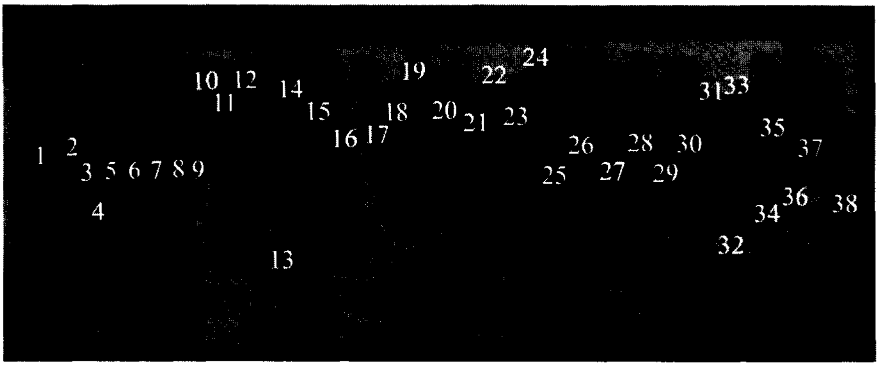


Figure 7.1. *Wooden chips that were randomly selected for the primary sample. In the picture the chips are numbered, starting with chip number 1 on the left to chip number 38 on the right. This ordering is arbitrary. Also a length of 20 cm is indicated. The boundary value of the sample mass was set to 20 g.*

All element determinations were made using Instrumental Neutron Activation Analysis, abbreviated as INAA (see *e.g.* Bode and de Goeij, 1998 and Bode, 2000). Polyethylene capsules with cylindrical dimensions with diameter 9.0 mm and height 15 mm were used for all samples. Each particle was milled and put in one or more capsules, each containing approximately 250 mg of material. After grinding, the chip-material was dried in an oven at 105 °C for 24 hours after which the total dry mass of the chip was determined. A next chip was selected and processed until the dry mass of the total of selected chips reached 20.0 g, resulting in the selection of 18 particles.

The INAA procedure resulted in elemental concentrations and an estimate of the standard uncertainty. The amount of an element in a chip was calculated by adding the absolute amounts in the capsules that belong to that chip. In the Table 7.1, the chip-mass of the selected chips and the concentrations of Zn, As and Cr are given. These elements were selected since they all had a high chip-to-chip variation. Also the combined standard uncertainties in percent provided by the INAA system are given.

Nr	m_i (g)	Zn (mg/kg)	Standard uncertainty Zn (%)	As (mg/kg)	Standard uncertainty As (%)	Cr (mg/kg)	Standard uncertainty Cr (%)
1	0.817	7.60	3.0	3.38E-2	3.6	0.72	15
2	0.335	17.16	2.8	5.79E-2	3.5	5.08	5.4
3	0.271	13.58	3.5	1.67E-2	15	0.67	34
4	0.274	16.31	3.0	2.73E-2	7.9	1.02	22
5	0.107	15.42	6.3	2.13E-2	21	1.18	49
6	1.313	128.94	0.7	10.05E-2	41	0.88	12
7	3.019	32.16	0.6	4.21E-2	2.1	1.91	3.6
8	0.164	14.14	5.6	1.52E-2	21	0.79	47
9	0.349	20.74	3.1	14.79E-2	5.0	1.98	13
10	0.250	3.56	12	2.50E-2	17	0.54	46
11	1.084	58.21	0.9	7.99E-2	2.6	2.52	4.7
12	0.578	36.51	1.5	3.13E-2	8.5	1.02	18
13	0.660	9.34	3.3	2.21E-2	11	1.01	16
14	0.928	36.90	1.0	1.55E-2	8.1	0.59	20
15	6.604	24.21	0.6	2.85E-2	3.2	3.24	1.4
16	0.733	22.92	1.6	4.24E-2	5.5	1.84	7.8
17	2.037	4824.74	0.6	109.97E-2	9.9	7.21	2.4
18	0.444	16.06	2.5	1.83E-2	6.4	1.83	11
$\sum$	20.0						

Table 7.1. *Masses of the selected wooden chips and the associated concentrations of Zn, As and Cr. The standard uncertainties are estimates calculated during the INAA analysis, which include the most significant contributions to the uncertainty of measurement in INAA.*

With the data in Table 7.1 for Zn, As and Cr, the sample concentration and its estimated variance can be calculated, as well as the combined standard uncertainties using Equations 7.2, 7.5 and 7.6. This results in the following estimates and corresponding combined standard uncertainties (Table 7.2):

	Value for Zn (mg/kg)	Value for As (mg/kg)	Value for Cr (mg/kg)
a_{sample}	523	0.15	2.68
$\sqrt{V_{sample}(a_{sample})}$	492	0.11	0.57
$u_c(a_{sample})$	3.0 (= 0.6%)	0.011 (=7.7%)	0.031 (=1.2%)
$u_c\left(\sqrt{V_{sample}(a_{sample})}\right)$	3.0 (= 0.6%)	0.011 (=10.1%)	0.014 (=2.6%)

Table 7.2. *Estimates for the concentrations of Zn, As and Cr, their variances and their combined standard uncertainties. The combined standard uncertainties are based on the standard uncertainties given in Table 7.1.*

7.4 *Steel slag*

In the previous example, the wooden particles had to be milled in order to put a standard amount in each capsule used for analysis. When sampling a batch of steel slag produced during the production of steel, it is technically difficult to mill a single particle with typical diameter of 1 cm to powder. Instead, it is easier to mill larger amounts (several particles) to powder. Therefore, in this example a group of particles constitutes a 'particle'. Although this is not in accordance with the model for sample drawing proposed in Chapter 3, the following sampling procedure assures that sampling corresponds to mass-based multinomial selections of heaps of powder. Two batches of slag produced during the production of steel were milled to powder and dispersed over grids. Batch 1 (with M_{batch}=0.484 kg) and Batch 2 (M_{batch}=0.521 kg), were dispersed over Grid 1 and Grid 2 respectively. Grid 1 and 2 were subdivided in 690 squares and 1380 squares respectively, see Figure 7.2.

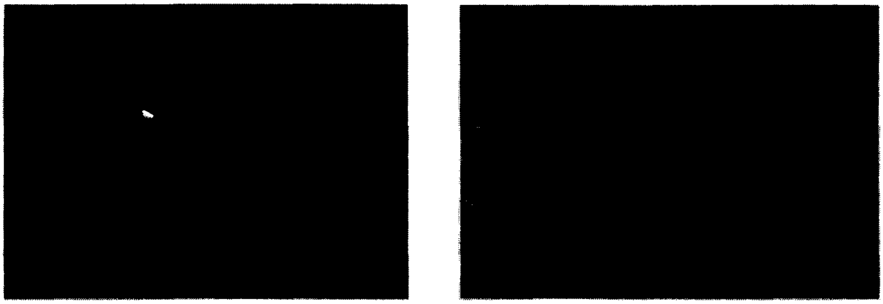


Figure 7.2. *The left graph shows the sampling from Grid 1 containing milled slag (Batch 1) formed during the production of steel. The right graph shows Batch 1 after 30 heaps were selected. Note that for analysis only the first 26 selected heaps were used.*

With the use of random numbers, squares were selected and all the particles in that square were added to the sample until the sample mass reached or exceeded the boundary value. For the sampling of Batch 1 and 2, the boundary value of the sample mass was set to 15 g and 5 g respectively (see Table 7.3 and 7.5), resulting in the selection of 26 and 19 heaps of powder respectively. This sampling protocol, which is not generally used, assures that the heaps of powder are selected according to a mass-based selection.

All element determinations were made using INAA. Polyethylene capsules with cylindrical dimensions with diameter 9.0 mm and height 15 mm were used for all samples. The material from each selected square was weighed and put in a cylindrical capsule. No further sample-handling steps were necessary. The element of interest was Fe, because this element is often heterogeneously distributed in slag, leading to large sampling errors.

The iron concentration and subsample masses drawn from Batch 1 and Batch 2 are given in Table 7.3 and Table 7.5 respectively.

Selection	Mass of selected heap (g)	Fe (g/kg)	Standard uncertainty (%)
1	0.804	11.70	1.6
2	0.427	8.28	1.8
3	0.629	6.13	1.6
4	0.530	9.75	1.8
5	0.465	9.24	2.0
6	0.450	7.24	1.8
7	0.674	10.50	1.5
8	0.572	7.01	1.8
9	0.647	10.80	1.6
10	0.426	10.10	1.7
11	0.704	8.14	1.6
12	0.514	8.45	1.7
13	0.417	8.48	2.8
14	0.631	7.70	1.6
15	0.700	9.48	1.7
16	0.453	9.08	1.7
17	0.472	7.63	1.8
18	0.471	8.91	1.8
19	0.620	8.37	1.7
20	0.795	7.32	1.8
21	0.813	10.00	1.6
22	0.461	7.65	2.1
23	0.372	7.20	2.0
24	0.596	9.56	1.5
25	0.606	9.59	1.5
26	0.772	26.80	1.5
Cumulative	15.021		

Table 7.3. *Masses of the heaps of powder selected from Batch 1 and associated concentrations of Fe. The standard uncertainties are estimates calculated during the INAA analysis, which include the most significant contributions to the uncertainty of measurement in INAA.*

For Batch 1, the statistical evaluation of the sample data yielded the results given in Table 7.4.

	Value for Fe (g/kg)
a_{sample}	9.75
$\sqrt{V_{\text{sample}}(a_{\text{sample}})}$	0.95
$u_c(a_{\text{sample}})$	0.04 (=0.4%)
$u_c\left(\sqrt{V_{\text{sample}}(a_{\text{sample}})}\right)$	0.02 (=2.1%)

Table 7.4. *Estimates for the concentration of Fe, its variance and their combined standard uncertainties for the sampling of Batch 1. The combined standard uncertainties are based on the standard uncertainties given in Table 7.3.*

For Batch 2, the following results were obtained (Table 7.5):

selection	Mass of selected heap (g)	Fe (g/kg)	Standard uncertainty (%)
1	0.174	198.9	1.4
2	0.287	193.0	1.4
3	0.409	199.0	1.4
4	0.410	198.0	1.8
5	0.183	197.8	1.4
6	0.201	195.0	1.4
7	0.214	192.1	1.4
8	0.247	200.0	1.3
9	0.261	198.2	1.4
10	0.221	198.2	1.3
11	0.361	219.1	1.4
12	0.263	198.1	1.4
13	0.297	190.9	1.4
14	0.294	200.0	1.5
15	0.333	202.1	1.4
16	0.254	198.0	1.4
17	0.225	195.1	1.4
18	0.191	206.8	1.4
19	0.231	197.8	1.4
Cumulative	5.056		

Table 7.5. *Masses of the heaps of powder selected from Batch 2 and associated concentrations of Fe. The standard uncertainties are estimates calculated during the INAA analysis, which include the most significant contributions to the uncertainty of measurement in INAA.*

The statistical evaluation of the sample data for batch 2 yielded the following estimates (Table 7.6):

	Value for Fe (g/kg)
a_{sample}	199.2
$\sqrt{V_{sample}(a_{sample})}$	1.7
$u_c(a_{sample})$	0.7 (=0.3%)
$u_c\left(\sqrt{V_{sample}(a_{sample})}\right)$	0.2 (=12.3%)

Table 7.6. *Estimates for the concentration of Fe, its variance and their combined standard uncertainties for the sampling of Batch 2. The combined standard uncertainties are based on the standard uncertainties given in Table 7.5.*

7.5 Shredded plastic

Plastic that is recycled to be used for the manufacturing of new plastic products is for technical reasons chopped up into pieces with typical diameter of several millimetres and typical mass of 20 mg (see Figure 7.3). Because a single particle of this size does not represent a standard amount of material for use of the analysis technique INAA, in this example, a group of plastic flakes constitutes a ‘particle’. Although this is not in accordance with the model for sample drawing proposed in Chapter 3, the following sampling procedure assures that sampling corresponds to mass-based multinomial selections of groups of particles. A batch of plastic flakes with M_{batch}=0.35 kg was equally distributed over a grid. The grid consisted of 1200 squares. Subsequently, using random numbers between 1 and 1200, samples were drawn. This was repeated until the cumulative sample mass reached 3 g, resulting in the selection of 10 groups of particles (see also Table 7.7).

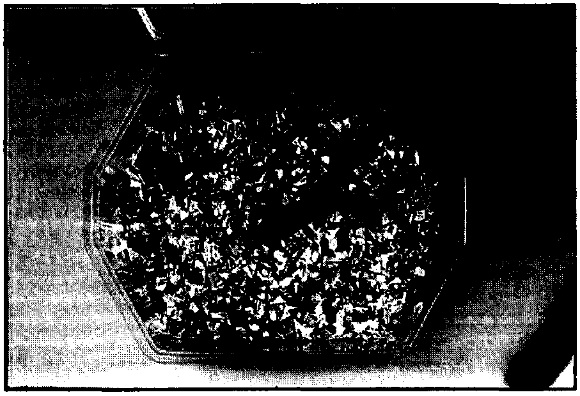


***Figure 7.3.** Top view of a batch of shredded plastic.*

All element determinations were made using INAA. Polyethylene capsules with cylindrical dimensions with diameter 9.0 mm and height 15 mm were used for all samples. After a subsample was drawn, no further sample-handling steps were necessary. The subsamples were weighed into polyethylene capsules. The element of interest was Br, because the variation in concentration between the groups of particles was very large for this element.

Selection	Mass of selected group (g)	Bromine (g/kg)	Standard uncertainty (%)
1	0.323	3.20	2.0
2	0.230	0.88	2.1
3	0.322	2.17	2.0
4	0.289	5.40	1.8
5	0.314	2.29	2.0
6	0.309	4.68	2.0
7	0.321	1.65	1.9
8	0.293	0.09	1.9
9	0.337	0.14	2.2
10	0.249	0.24	2.1
Cumulative	2.987		

***Table 7.7.** Masses of the groups of particles selected from the batch of plastic chips and associated concentrations of Br. The standard uncertainties are estimates calculated during the INAA analysis, which include the most significant contributions to the uncertainty of measurement in INAA.*

The amount of bromine in a group of particles and the associated concentration of Br are given in Table 7.7.

For the batch of shredded plastic, the statistical evaluation of the sample data yielded the following estimates (Table 7.8):

	Value for Br (g/kg)
a_{sample}	2.12
$\sqrt{\hat{V}_{HT}(a_{sample})}$	0.59
$u_c(a_{sample})$	0.02 (=0.8%)
$u_c\left(\sqrt{\hat{V}_{HT}(a_{sample})}\right)$	0.01 (=1.3%)

Table 7.8. *Estimates for the concentration of Br, its variance and their combined standard uncertainties for the sampling of the batch of plastic chips. The combined standard uncertainties are based on the standard uncertainties given in Table 7.7.*

7.6 Validation by using bootstrapping

In this paragraph, the level of contradiction of the new theory is investigated by the general method of bootstrapping using the analytical results of the previous paragraph. During bootstrapping, the sample is "extended" to form a hypothetical batch. The hypothetical batch contains as many classes as particles in the original sample. Each class corresponds to a single particle in the sample and the particles in a class are assumed to be identical to the corresponding particle in the original sample. Then, many samples are drawn from this batch using a specific sampling algorithm. Before the next sample is drawn, the sampled material from the previous sample is put back in the batch. The distribution of samples obtained is finally used as an "estimate" of the distribution of the original sample. Here, the level of contradiction is quantitatively defined as the absolute value of the difference between the variance predicted by the equation provided by a sampling theory and the actual variance if sampling corresponds to the model provided by the theory. Therefore, the method of bootstrapping provides insight into the levels of contradiction of the non-empirical sampling theories of Gy, Wilson and this study.

For each distribution estimate, 10^4 samples were drawn. The mass concentration in the i^{th} sample drawn using a specific sampling algorithm is denoted as $a_{i,sample}$ in which i can represent any integer value between 1 and 10^4. The variance estimate obtained with bootstrapping, denoted here as V^*, is calculated as follows (Särndal *et al*, 1992):

$$V^{*} = \frac{1}{10^4 - 1}\sum_{i=1}^{10^4}\left(a_{i,\mathrm{sample}} - \frac{1}{10^4}\sum_{j=1}^{10^4} a_{j,\mathrm{sample}}\right)^2 \tag{7.9}$$

For each algorithm (Wilson, Gy and this study) the above variance estimate can be compared with the theoretical predictions (Equations 2.17, 2.31 and 4.115 respectively). In the following, bootstrapping is applied for the sample of wooden chips, both slag samples and the sample of shredded plastic.

Wooden chips. The sample consisting of wooden chips is extended to a hypothetical batch containing about 10^4 particles. Because the sample contains 18 particles, it is assumed that the batch contains 18 classes, with each class containing 556 identical particles. The 18 classes contain particles identical to respectively the 18 particles in the sample of wooden chips. Next, for the sampling algorithms of Gy, Wilson and the mass-based approach, a large number of samples (10^4) is drawn from the hypothetical batch, which consists of a population of 10008 (=18×556) particles. For the algorithm of Gy a value for q has to be assigned. Because a larger value of q leads to a larger average sample mass, the value of q is chosen so that the expected value of the sample mass will be 20 g. This value corresponds to q=1/556, because then the expected value of the sample mass using Gy's algorithm is:

$$E\left(M_{\mathrm{sample}}\right) = \sum_{i=1}^{18} N_{i,\mathrm{batch}}\, q\, m_i = \sum_{i=1}^{18} 556 \frac{1}{556} m_i = \sum_{i=1}^{18} m_i = 20.0\,\mathrm{g}$$

in which m_i represents the mass of a particle belonging to the i^{th} class.

Similarly, for the implementation of Wilson's algorithm a value has to be assigned to N. It is assumed that N=18, leading to samples with the same number of particles as the original sample, which was extended to form the hypothetical batch.

For sampling according to the fixed mass design, selections are terminated when the obtained sample mass is larger than or equal to M=20 g.

The theoretical variances of the sampling process according to Wilson, Gy and the size-based approach are respectively given by Equations 2.17, 2.31 and 4.115 with $y_i=a_i m_i$ and $z_i=m_i$. Substituting the parameters of the hypothetical batch into the equations results in numerical values for the variance. For Cr, it is found that in all the three cases the theoretical variance is equal to 0.30.

In table 7.9 the theoretical variances (denoted as V) and variances obtained (denoted as V^*) for Cr are compared. It is observed that in the mass-based approach the agreement between the theoretical and variance obtained is closest.

	V $(mg/kg)^2$	V^* $(mg/kg)^2$	Relative deviation $(V-V^*)/V$
this study	0.30	0.29	0.03
Wilson	0.30	0.36	-0.20
Gy	0.30	0.41	-0.37

Table 7.9. *Comparison between theoretical and obtained variances for Cr for the mass-based approach and Wilson's and Gy's algorithm.*

Slag from production of steel. The sample drawn from Batch 1 is multiplied by 385 to obtain the simulated batch. In this way, the batch contains 26×385=10010 particles. For the mass-based approach, Gy's algorithm and Wilson's algorithm respectively the following parameters are chosen: M=15 g, q=1/385 and N=26. Analogously to the bootstrapping using the sample of wooden chips, the choice of q results in an expected sample masses of 15 g for the algorithm of Gy. The following variances were obtained (Table 7.10):

	V $(g/kg)^2$	V^* $(g/kg)^2$	Relative deviation $(V-V^*)/V$
this study	0.851	0.845	0.0071
Wilson	0.851	0.848	0.0035
Gy	0.851	0.854	-0.0035

Table 7.10. *Comparison of theoretical and obtained variances for Fe in samples from Batch 1 for the mass-based approach and Wilson's and Gy's algorithm.*

Wilson's and Gy's algorithm have relatively less deviation than this study. However, for all algorithms, the relative deviations in the variances are much smaller than in the previous example of sampling a batch of wooden chips.

The sample drawn from Batch 2 is multiplied by 526 to obtain the simulated batch. In this way, the batch contains 19×526=9994 particles. For the mass-based approach, Gy's algorithm and Wilson's algorithm respectively the following parameters are chosen: M=5.0 g, q=1/526 and N=19. Analogously to the bootstrapping using the sample of wooden chips, the choice of q results in an expected sample masses of 5 g for the algorithm of Gy. For Batch 2 the obtained variances are given in Table 7.11. For Wilson's algorithm the agreement between the variance obtained and the theoretical prediction is best.

	V $(g/kg)^2$	V^* $(g/kg)^2$	Relative deviation $(V-V^*)/V$
this study	2.68	2.60	0.030
Wilson	2.68	2.66	0.007
Gy	2.68	2.79	-0.041

Table 7.11. *Comparison of theoretical and obtained variances for Fe in samples from Batch 2 for the mass-based approach and Wilson's and Gy's algorithm.*

Shredded plastic. The sample drawn from the batch of shredded plastic is multiplied by 1000 to obtain the simulated batch. In this way, the batch contains $10\times1000=10^4$ particles. For the mass-based approach, Gy's algorithm and Wilson's algorithm respectively the following parameters are chosen: M=3.0g, q=1/1000 and N=10. Analogously to the bootstrapping using the sample of wooden chips, the choice q results in an expected sample mass equal to the original sample, which was extended to form the hypothetical batch. The following variances for Br were obtained (Table 7.12):

	V $(g/kg)^2$	V^* $(g/kg)^2$	Relative deviation $(V-V^*)/V$
this study	0.31	0.29	0.06
Wilson	0.31	0.32	−0.03
Gy	0.31	0.35	−0.13

Table 7.12. *Comparison of theoretical and obtained variances for Br in samples from the batch of shredded plastic for the mass-based approach and Wilson's and Gy's algorithm.*

Again, for Wilson's algorithm the agreement between the variance obtained and the theoretical prediction is best.

In all four examples studied, the relative deviation between theoretical prediction and variance obtained is for this study positive, with a minimum of 0.0071 (Batch 1 of steel slag) and a maximum of 0.06 (batch of recycled plastic).

For the algorithm of Wilson the minimum and maximum relative deviations observed are −0.20 and 0.007 respectively. This indicates that the range of possible relative deviations is much larger than for this study.

In all four examples studied, the relative deviation between theoretical prediction and variance obtained is for Gy's algorithm negative. For Gy's algorithm the minimum and maximum relative deviations observed are −0.37 and −0.0035 respectively. This also indicates that the range of possible relative deviations is much larger than for this study.

7.7 *Normality*

Bootstrapping can also be used to investigate other properties of the distribution of samples obtained. An important property that can be investigated using bootstrapping is correspondence to a normal or Gaussian distribution. In this paragraph, the degree of normality of the sample concentration, a_{sample}, is investigated.

When samples of constant mass M are considered to be composite samples consisting of N independently drawn increments of mass M/N, the central limit theorem can be used to demonstrate that in the limit of both an infinite sample-to-particle size ratio and an infinite batch-to-sample size ratio, the distribution of a_{sample} is normal. However, a further derivation is not presented here, because in practice both the sample-to-particle size ratio and the batch-to-sample size ratio are finite, resulting in a potentially non-normal distribution of a_{sample}. Below, it is shown that, for a given sample-to-particle size ratio and batch-to-sample size ratio, there are two causes for non-normality: (i) fluctuations in the sample mass and (ii) variation in particle concentrations.

Because a_{sample} is equal to the ratio A_{sample}/M_{sample}, the distribution may become asymmetric, if large fluctuations in M_{sample} may occur. This first source of non-normality can be illustrated numerically as follows: Suppose A_{sample}=0.5 (arbitrary units) and $E(M_{sample})$=1.0 (arbitrary units). If M_{sample} is equal to its expected value, a_{sample}=0.5/1.0=0.5. If M_{sample} is 20% lower than the expected value, but A_{sample} remains equal to 0.5, a_{sample}=0.5/0.8=0.625. On the other hand, if M_{sample} is 20% larger than its expected value, but A_{sample} remains equal to 0.5, a_{sample}=0.5/1.2=0.417. It can be concluded that the deviation in a_{sample} due to a downward fluctuation of the sample mass (0.625−0.5=0.125) is larger than the deviation in a_{sample} due to an upward fluctuation of the sample mass (0.5−0.417=0.083). Because the normal distribution is symmetric, large deviations in the sample mass, which may lead to an asymmetric distribution, are a source of non-normality.

As a low variability in particle concentrations implies that fluctuations in M_{sample} and A_{sample} are more strongly correlated than would be the case with a high variability in particle concentrations, the occurrence of the above mechanism, where it was assumed that A_{sample} remained constant, while M_{sample} varied, becomes less likely. Therefore, the first source of non-normality is counteracted when the variation in particle concentrations is reduced. In other words, reducing the variability in particle concentrations will increase the normality, because the non-normality induced by fluctuations in the sample mass is counteracted. Variability in particle concentrations is a source of non-normality.

For the mass-based approach, the variation in sample masses is maximally m_{max}, while for the algorithm of Gy, the variation in sample masses is equal to M_{batch}, which is generally much larger than m_{max}. The algorithm of Wilson will generally result in a variation in sample masses that lies between the variation for the mass-based approach and the variation for the algorithm of Gy: the variation is $N\times(m_{max}-m_{min})$, where m_{min} is the smallest particle mass in the batch. This shows that, in theory, the algorithm of Wilson may lead to less variation in sample masses than the mass-based approach, if the difference in particle masses, (m_{max}−m_{min}), is smaller than m_{max}/N. However, in practice, there will always be a variation in particle masses and hence, for large samples, the variation in sample masses will be larger for the algorithm of Wilson than for the mass-based approach.

Because fluctuations of the sample mass are generally largest for Gy's algorithm and smallest for the mass-based approach used for the current study, it is expected that the degree of normality will be in the following order: this study≥Wilson>Gy. To test this hypothesis, a quantitative test of normality has to be used. It is chosen here to look at the numbers of samples exceeding the 2σ- and 3σ-levels.

Wooden chips. For the sample of wooden particles, bootstrapping was performed with the same settings as in Paragraph 7.6. In Figure 7.4, the sample concentrations of Cr are plotted against the sample mass obtained, for the mass-based approach, the algorithm of Gy and the algorithm of Wilson.

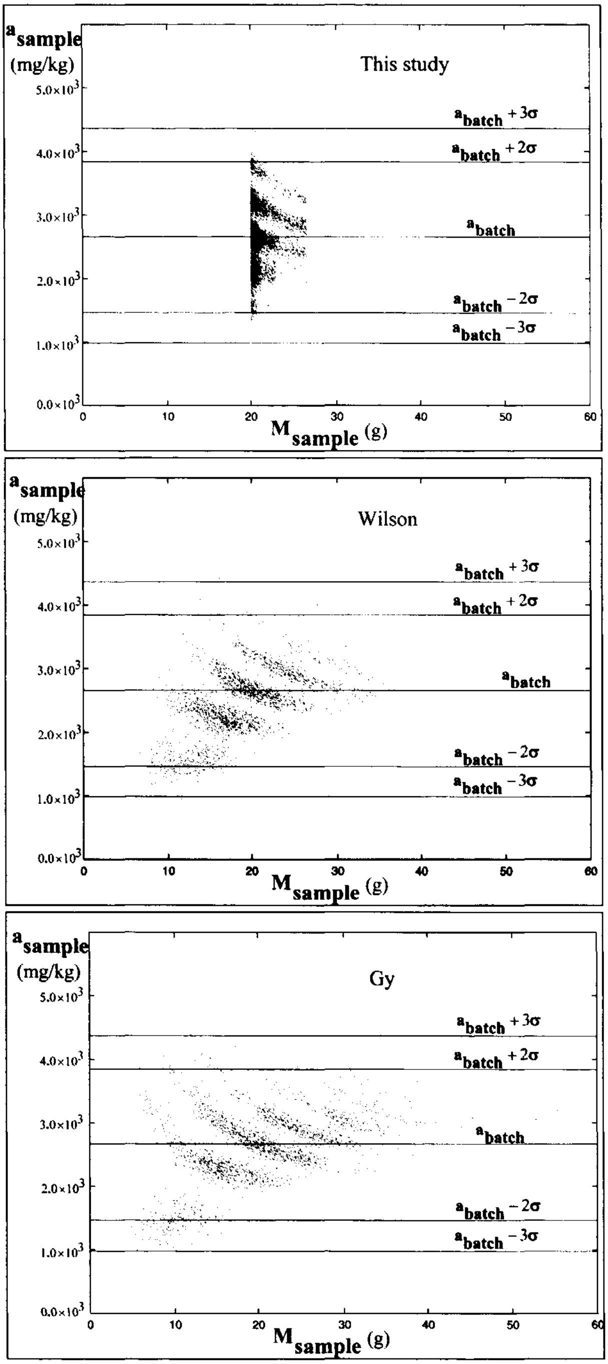


***Figure 7.4.** Sample concentration of Cr versus obtained sample mass for the mass-based approach, Wilson's algorithm and Gy's algorithm. The horizontal lines indicate the batch concentration, the 2σ-levels and the 3σ-levels. Bootstrapping was performed on the sample drawn from the batch of wooden particles.*

Especially samples with a small sample mass have large deviations from the batch concentration. To investigate the degree of normality numerically the total numbers of samples that exceed the 2σ- or 3σ-levels are put in Table 7.13.

	Number of samples exceeding 2σ-levels	Number of samples exceeding 3σ-levels
this study	471	30
Wilson	942	44
Gy	1107	99

Table 7.13. *Number of samples outside the 2σ- and 3σ-levels for the mass-based approach (used in this study) and the sampling algorithms of Wilson and Gy. Bootstrapping was performed on the sample drawn from the batch of wooden particles.*

Steel slag. For Batch 1 of steel slag, bootstrapping was performed with the same settings as in Paragraph 7.6. In Figure 7.5, the concentrations of Fe are plotted versus the obtained sample mass for the mass-based approach, the algorithm of Wilson and the algorithm of Gy.

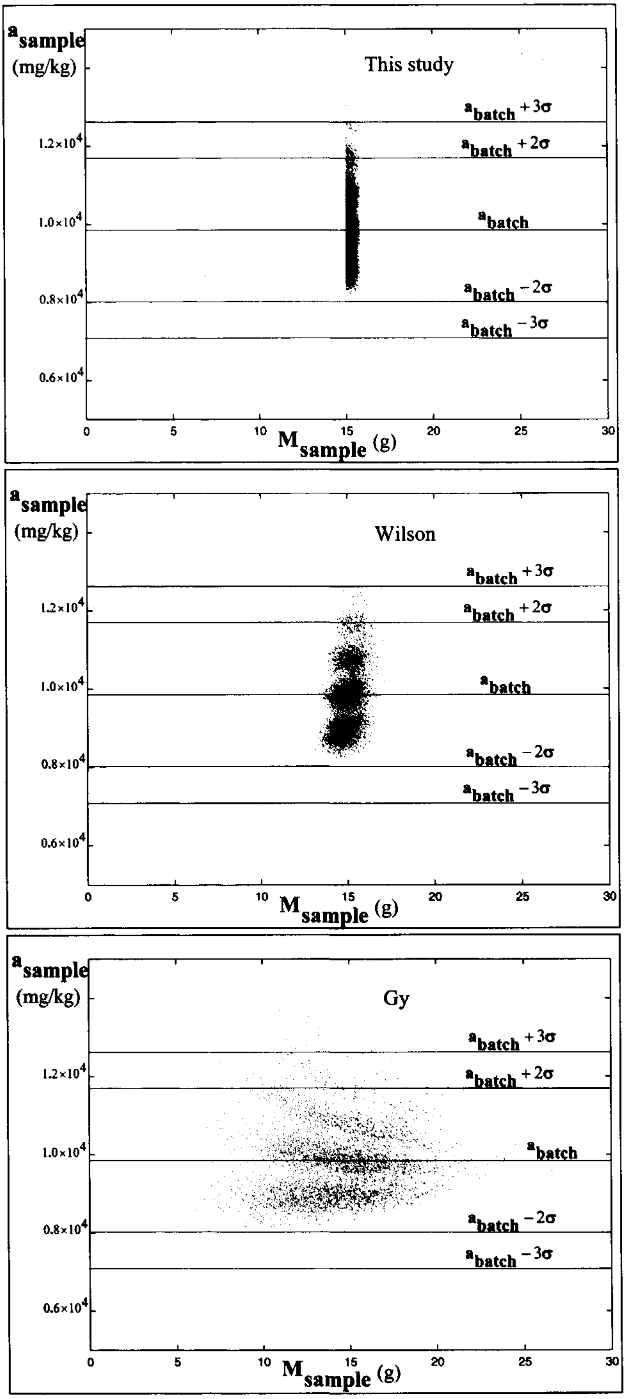


Figure 7.5. *Sample concentration of Fe versus obtained sample mass for the mass-based approach, the algorithm of Wilson and the algorithm of Gy. The horizontal lines indicate the batch concentration, the 2σ-levels and the 3σ-levels. Bootstrapping was performed on the sample drawn from Batch 1 of steel slag.*

	Number of samples exceeding 2σ-levels	Number of samples exceeding 3σ-levels
this study	413	72
Wilson	394	61
Gy	364	73

Table 7.14. *Number of samples outside the 2σ- and 3σ-levels for the mass-based approach (used in this study) and the sampling algorithms of Wilson and Gy. Bootstrapping was performed on the sample drawn from Batch 1 of steel slag.*

For Batch 2 of steel slag, bootstrapping was performed with the same settings as in Paragraph 7.6. In Figure 7.6, the concentrations of Fe are plotted versus the obtained sample mass for the mass-based approach, the algorithm of Wilson and the algorithm of Gy.

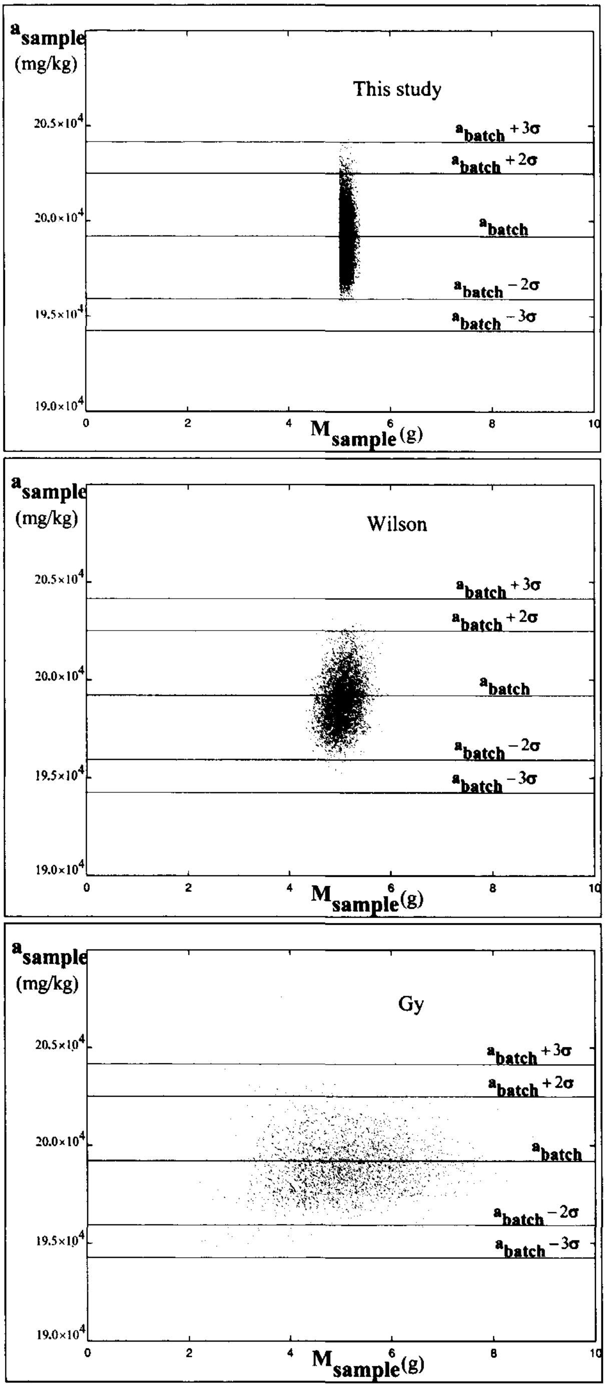


Figure 7.6. *Sample concentration of Fe versus obtained sample mass for the mass-based approach, the algorithm of Wilson and the algorithm of Gy. The horizontal lines indicate the batch concentration, the 2σ-levels and the 3σ-levels. Bootstrapping was performed on the sample drawn from Batch 2 of steel slag.*

	Number of samples exceeding 2σ-levels	Number of samples exceeding 3σ-levels
this study	409	55
Wilson	408	48
Gy	436	62

Table 7.15. *Number of samples outside the 2σ- and 3σ-levels for the mass-based approach (used in this study) and the sampling algorithms of Wilson and Gy. Bootstrapping was performed on the sample drawn from Batch 2 of steel slag.*

Recycled plastic. For the sample of recycled plastic, bootstrapping was performed with the same settings as in Paragraph 7.6. In Figure 7.7, the concentrations of Br are plotted versus the obtained sample mass for the mass-based approach, the algorithm of Wilson and the algorithm of Gy.

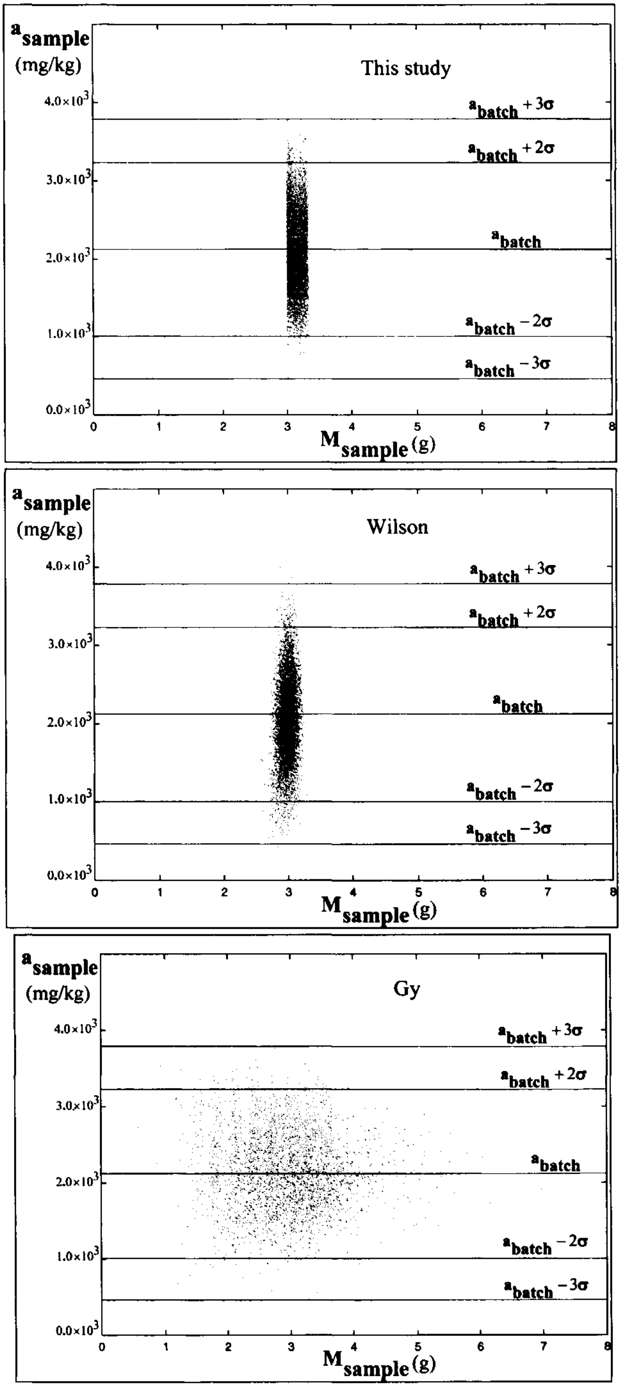


***Figure** 7.7. Sample concentration of Br versus obtained sample mass for the mass-based approach, the algorithm of Wilson and the algorithm of Gy. The horizontal lines indicate the batch concentration, the 2σ-levels and the 3σ-levels. Bootstrapping was performed on the sample drawn from the batch of plastic particles.*

	Number of samples exceeding 2σ-levels	Number of samples exceeding 3σ-levels
this study	369	19
Wilson	479	26
Gy	588	75

Table 7.16. *Number of samples outside the 2σ- and 3σ-levels for the mass-based approach (used in this study) and the sampling algorithms of Wilson and Gy. Bootstrapping was performed on the sample drawn from the batch of plastic particles.*

7.8 Interpretation of results on normality

In the previous paragraph, it was discussed that the degree of normality increases when both sample masses and particle concentrations have less variation. The aim of this paragraph is to interpret the results of the previous paragraph in the context of the above remark. For this, the degree of normality is quantified, using the results of the previous paragraph.

Ideally, when sampling a normal distribution, 0.28% of the samples will be outside the 3σ-levels and 4.56% outside the 2σ-levels (Bendat and Piersol, 1971). Because each time 10^4 sample were drawn, these percentages are respectively exceeded when the actual number of samples outside the 2σ-levels is larger than 456 and the number of sample exceeding the 3-σ level is larger than 28.

The numbers of samples outside the 2σ-levels or 3σ-levels are subject to statistical fluctuations. Assuming that these numbers are distributed following Poisson distributions, the standard deviations of the numbers of samples outside the 2σ-levels or the 3σ-levels are respectively $\sqrt{456}$ and $\sqrt{28}$. Therefore, it is defined here that when the actual number of samples outside the 2σ-levels is larger than $456+3\sqrt{456}=520$ and the actual number of samples outside the 3σ-levels is larger than $28+3\sqrt{28}=44$ the 2σ- and 3σ-levels are significantly exceeded.

Using the above definition the following conclusions can be drawn (see Table 7.17).

		Variation in sample masses (g)	2-σ	3-σ
Wooden chips	this study	6.6		
	Wilson	116.9	*	
	Gy	11,101.7	*	*
Batch 1	this study	0.813		*
	Wilson	12.6		*
	Gy	5,783.1		*
Batch 2	this study	0.410		*
	Wilson	4.5		*
	Gy	2,659.5		*
Shredded plastic	this study	0.3		
	Wilson	1.1		
	Gy	2,987.0	*	*

Table 7.17. *Non-normality of the estimators during different sampling algorithms and associated variation in sample masses. A '*' indicates a significant exceeding of the 2σ- or 3σ-levels (based on 10^4 samples).*

The degree of normality, N, can be defined as follows:

- $N = 2$, if there no significant exceeding of 2σ- and 3σ-levels
- $N = 1$, if there is a significant exceeding of either the 2σ-levels or the 3σ-levels
- $N = 0$, if both the 2σ-levels and the 3σ-levels are significantly exceeded.

From the above definition of N and the results in Table 7.17 follows that for the sampling of wooden chips, the degree of normality is highest for this study (2), lower for Wilson (1) and lowest for Gy (0). This is exactly in accordance with the result of the discussion in the previous paragraph, because it can also be deduced from Table 7.17 that the variation in sample masses varies as mass-based approach<Wilson<Gy.

For the sampling of steel slag (Batch 1 and 2), the degree of normality is equal for all three sampling algorithms ($N = 1$). This result is not in accordance with the results of the discussion in the previous paragraph. An explanation is that the definition of the degree of normality, N, is too coarse to distinguish between the three algorithms. A different definition of the degree of normality, with more than three degrees, may indeed reveal that the nomality varies in the order: this study>Wilson>Gy.

For the sampling of plastic, the degree of normality is lowest for the algorithm of Gy ($N = 0$), while the algorithms of this study and Wilson have the same degree of normality ($N = 2$). This is almost in accordance with the results of the discussion in the previous paragraph, except that the algorithm of this study and Wilson result in the same degree of normality. The latter observation can be explained. It can be deduced from Table 7.17 that the variations in sample masses is not much larger for the algorithm of Wilson than for the mass-based approach. This can also be observed in Figure 7.7. It can therefore be expected that the definition of N is too coarse to distinguish between the algorithms of this study and Wilson. An alternative definition of the degree of normality, allowing for more than three degrees, may reveal that even for the sampling of plastic, the degree of normality is higher for the algorithm of this study than for the algorithm of Wilson.

7.9 Discussion and conclusions

The procedure to estimate the batch concentration, variance and the combined standard uncertainties was illustrated with practical examples (a batch of wooden chips, two batches of slag and a batch of shredded plastic). Despite some large standard uncertainties in analysis results of individual particles or subsamples (up to 49% for the measured Cr concentration in the fifth selected wooden particle) the combined standard uncertainties in the final results were smaller (up to 12.3% for the estimated standard deviation for the sampling from Batch 2 of steel slag).

Next, the validity of the equations for the variance of the theory of Wilson, Gy and this study were compared with bootstrapping using analysis results of the four distinct samples. While the equation for the variance provided by this study (Equation 4.115) is only exact in the limit of an infinite sample-to-particle size ratio and an infinite batch-to-sample size ratio, the approximate nature of Wilson's and Gy's equations are less clear. The results of bootstrapping show that deviations exist in all three algorithms. For the four samples studied, the ranges of relative differences are greater for the algorithms of Wilson and Gy than for this study. Although this shows that the sampling theory presented in this study may have a lower level of contradiction than the theories of Wilson and Gy, no theoretical foundation for this observation is presented in this study. Therefore, further research into the level of contradiction of the three theories is recommended.

In Paragraphs 7.7 and 7.8, the non-normality of the sample concentration was investigated. It was demonstrated theoretically that reducing the variation in concentrations between the particles increases the normality, while increasing the variation in sample masses reduces the normality. The first effect does not depend on the sampling algorithm used, but the second effect does and generally plays a larger role for the algorithm of Wilson and Gy than for this study. It was demonstrated (see Table 7.17) that the algorithm of Gy leads in all four cases to a non-normal ($N<2$) estimator. The algorithm of Wilson leads to a non-normal estimator for Batch 1 and Batch 2 and wooden chips. The estimator obtained with the mass-based approach is only non-normal for Batch 1 and Batch 2. This increased non-normality for the algorithms of Wilson and Gy can therefore be related to the greater variability in sample masses compared with the mass-based approach (see Table 7.17).

7.10 References

J. S. Bendat and A.G. Piersol (1971) *Random Data: analysis and measurement procedures,* John Wiley & Sons, 407 pp.

P. Bode (2000), Automation and Quality Assurance in the Neutron Activation Facilities in Delft, *J.Radioanal.Nucl.Chem.* **245**, p. 127-132.

P.Bode, J.J.M. de Goeij (1998), *Activation Analysis, Encyclopedia of Environmental Analysis and Remediation,* J.Wiley & Sons, New York, p. 68 - 84, ISBN 0-471-11708-0.

ISO (1995) *Guide to the expression of uncertainty in measurement,* 1st ed.

C. Särndal, B. Swensson and J. Wretman (1992) *Model Assisted Survey Sampling,* Springer, New York, 694 pp.

Chapter 8 Minimum sample mass[18]

The minimum sample mass can be estimated using the properties of the particles in the batch or sample. Using the properties of the particles in the sample, a feedback mechanism is proposed to draw additional samples. The mechanism is investigated with simulations. Knowledge of the maximum particle mass and of the minimum and maximum particle concentrations in the sample can be used for estimation of the minimum sample mass.

8.1 Introduction

When the mass concentration in the batch, a_{batch}, is estimated using the mass concentration in the sample, a_{sample}, several factors influence the potential difference between the obtained value for the estimate and the actual batch concentration. The obtained value for the estimate would be exactly equal to the true value if (*i*) the analytical uncertainty is zero, (*ii*) the estimator is unbiased, and (*iii*) the estimator has a zero variance. In practice, the analytical error may be non-zero, the mass concentration in a sample may be slightly biased, and the variance may also be non-zero. Therefore, the three factors influencing the potential difference between the obtained value for the estimate and the actual batch value are analytical uncertainty, bias, and variance. A larger analytical uncertainty, a larger bias or a larger variance leads to more potential difference between the obtained value for the estimate and the actual batch value. Therefore, it is important to have insight in the magnitudes of the analytical uncertainty, bias and variance.

In Chapter 6 and Chapter 7, the magnitudes of bias and analytical uncertainty were investigated. In this chapter, the third aspect, the magnitude of the variance, is addressed. It is investigated how the variance can be reduced by the choice of the sample mass.

[18] Parts of this chapter have previously been published in: B. Geelhoed and H.J. Glass (2001) A new model for sampling of particulate materials and determination of the minimum sample size. *Geostandards Newsletter – The Journal of Geostandards and Geoanalysis*, **25**, p. 325-332.

8.2 *Minimum sample mass*

It is well known that the variance of the sample concentration, which is equal to the variance of the sampling error, reduces with increasing sample size. This condition implies that the relative standard deviation, defined as the square root of the variance divided by the batch concentration,

$$\mathrm{RSD} = \sqrt{V\left(a_{\mathrm{sample}}\right)} \Big/ a_{\mathrm{batch}} \qquad \mathbf{(8.1)}$$

also reduces with increasing sample size. The relation between the sample size and the relative standard deviation can thus be indicated using a function f():

$$\mathrm{RSD} = f\left(M_{\mathrm{sample}}\right) \qquad \mathbf{(8.2)}$$

where $f(M_{\mathrm{sample}})$ is a monotonic decreasing function of M_{sample}. Setting RSD=α, where α is defined as the 'maximum allowable coefficient of variation', for $M_{\mathrm{sample}}=M_{\mathrm{min}}$ and taking the inverse of f, f^{-1}, Equation 8.2 transforms into the following expression for the minimum sample mass:

$$f^{-1}(\alpha) = M_{\mathrm{min}} \qquad \mathbf{(8.3)}$$

For any sample mass larger than the minimum sample mass, it is guaranteed that the relative standard deviation does not exceed the preselected value α. Therefore, M_{min} can be interpreted as the minimum amount of material to be analyzed when it is demanded that the relative standard deviation does not exceed α. In the next paragraph, it is investigated how the minimum sample mass can be calculated using the properties of the particles in the batch.

8.3 *Estimation of the minimum sample mass using the properties of the particles in the batch*

In the size-based approach, the variance can be estimated using the sample size, knowledge of the properties of the particles in the batch and Equation 4.115. In the mass-based approach, this equation becomes:

$$V_{\mathrm{batch}}\left(a_{\mathrm{sample}}\right) = \frac{1}{M_{\mathrm{sample}}\overline{m}} \sum_{i=1}^{T} p_i' \, m_i^2 \left(a_i - a_{\mathrm{batch}}\right)^2 \qquad \mathbf{(8.4)}$$

where a_{sample} is the mass concentration in the sample, a_i is the mass concentration in a particle of type i, and $\overline{m}$ is the mean particle mass in the batch. It is noted that $V_{\mathrm{batch}}(a_{\mathrm{sample}})$ is by definition a random variable, because the sample mass is a random variable. The relative variance estimated with batch information, $V^{rel}_{batch}(a_{sample})$, is defined as:

$$V^{rel}_{batch}(a_{sample}) \equiv \frac{V_{batch}(a_{sample})}{a^2_{batch}} = \frac{1}{a^2_{batch} M_{sample} \overline{m}} \sum_{i=1}^{T} p'_i \, m_i^2 (a_i - a_{batch})^2 \tag{8.5}$$

The relative standard deviation estimated using the properties of the particles in the batch is defined as the square root of $V_{batch}(a_{sample})/a^2_{batch}$. From Equation 8.5, it follows that this standard deviation is inversely proportional to the square root of the sample mass. The condition that the relative standard deviation estimated using the properties of the particles in the batch should not exceed a preselected value α is equivalent to the following inequality:

$$\frac{1}{a^2_{batch} M_{sample} \overline{m}} \sum_{i=1}^{T} p'_i \, m_i^2 (a_i - a_{batch})^2 \le \alpha^2 \tag{8.6}$$

A minimum value is obtained for the sample mass, the minimum sample mass estimated using the properties of the particles in the batch, $M_{min,b}$:

$$M_{min,b} = \frac{1}{\alpha^2 a^2_{batch} \overline{m}} \sum_{i=1}^{T} p'_i \, m_i^2 (a_i - a_{batch})^2 \tag{8.7}$$

When the variance estimated with batch information is equal to the actual variance, *i.e.* when the level of contradiction is zero (see Chapter 7), $M_{min,b}=M_{min}$.

As an example, Equation 8.7 can be applied for $T=2$, $a_1=1$, $a_2=0$ and $m_1=m_2=1$. Because the particle masses are one, $M_{min,b}$ is the minimum number of particles required (denoted as $N_{min,b}$). The batch concentration is equal to p'_1 and p'_2 is equal to $1-p'_1$. This results in the following expression for the minimum number of particles required: $N_{min,b} = (1-p'_1)/(\alpha^2 p'_1)$. This expression corresponds to the equation given in literature for the sampling from a binomial distribution (see *e.g.* Barnett, 1974).

An alternative way of expressing $M_{min,b}$ is:

$$M_{min,b} = \frac{M_{sample} V_{batch}(a_{sample})}{\alpha^2 a^2_{batch}} \tag{8.8}$$

in which Equation 8.4 and 8.7 were used. This shows that the minimum sample mass can be estimated using two batch properties: the batch concentration and the variance estimated using the properties of the particles in the batch, given by Equation 8.4. These quantities are generally unknown. Therefore, in the next paragraph an alternative method to calculate the minimum sample mass will be proposed, where, instead of the above two batch properties, the sample concentration and the estimator $V_{sample}(a_{sample})$ for the variance of the mass concentration in the sample are used.

8.4 Estimation of the minimum sample mass using the properties of the particles in the sample

The minimum sample mass can be estimated by replacing, in Equation 8.8, $V_{batch}(a_{sample})$ and a_{batch} by the approximately unbiased estimators $V_{sample}(a_{sample})$ and a_{sample} respectively. This yields an estimator for the minimum sample mass based on the properties of the particles in the sample:

$$M_{min,s} = \frac{M_{sample} V_{sample}(a_{sample})}{\alpha^2 a_{sample}^2} \tag{8.9}$$

For a sample S, the value of $M_{min,s}$ is denoted as $M_{min,s}(S)$. An alternative estimator for the minimum sample mass is obtained when, in Equation 8.8, a_{batch}^2 is not replaced by a_{sample}^2, but by its approximately unbiased estimator $a_{sample}^2 - V_{sample}(a_{sample})$:

$$M_{min,alt} = \frac{M_{sample} V_{sample}(a_{sample})}{\alpha^2 \left(a_{sample}^2 - V_{sample}(a_{sample})\right)} \tag{8.10}$$

The above equation may of course not be used if $V_{sample}(a_{sample})$ is larger than or equal to a_{sample}^2, because this would lead to a negative or infinite value of $M_{min,alt}$. Using Equation 8.9, the above equation can also be written as:

$$M_{min,alt} = \frac{M_{min,s}}{1 - V_{sample}(a_{sample}) / a_{sample}^2} \tag{8.11}$$

Hence, $M_{min,alt}$ is larger than or equal to M_{min}. Substituting the equation for $V_{sample}(a_{sample})$, Equation 5.78, into Equation 8.9 results in:

$$M_{min,s} = \frac{\sum_{n=1}^{T} N_n m_n^2 (a_n - a_{sample})^2}{a_{sample}^2 \alpha^2 (M_{sample} - M_{sample} / N_{sample})} \tag{8.12}$$

Instead of batch information the above equation uses strictly sample information. The same is true for $M_{min,alt}$ but the result is a much more complicated equation. In the following, a feedback mechanism for the drawing of a (composite) sample is proposed which uses $M_{min,s}$. The scheme is illustrated in Figure 8.1. The aim of the proposed mechanism is to ensure that the relative standard deviation of the mass concentration in the finally obtained composite sample is smaller than or equal to α.

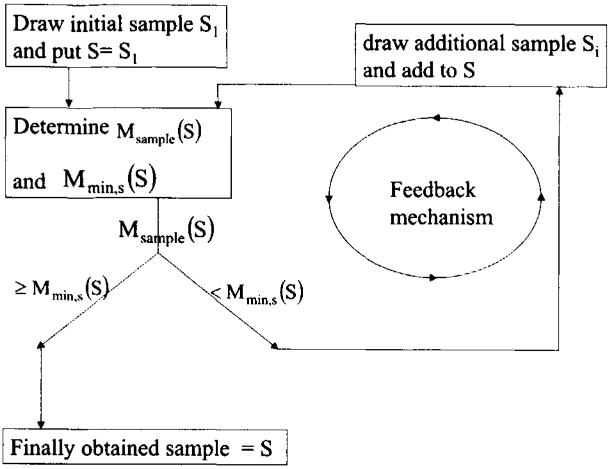

Draw initial sample S_1 and put $S= S_1$
draw additional sample S_i and add to S
Determine $M_{sample}(S)$ and $M_{min,s}(S)$
Feedback mechanism
$M_{sample}(S)$
$\geq M_{min,s}(S)$
$< M_{min,s}(S)$
Finally obtained sample = S

Figure 8.1. Schematic representation of the proposed feedback mechanism to draw a sample. In the depicted scheme $M_{min,s}$ *is used. A similar scheme can be constructed in which* $M_{min,alt}$ *is used.*

Precisely formulated the procedure is:

- An initial sample S_1 is drawn with boundary value of the mass M_1 and from this sample the estimate for the minimum sample mass $M_{min,s}(S_1)$ is derived using Equation 8.12
- If $M_{sample}(S_1)<M_{min,s}(S_1)$ an additional sample, S_2, with boundary value of the sample mass $M_2=M_{min,s}(S_2)-M_{sample}(S_1)$ is drawn and added to the first sample.
- The previous step is repeated as many times as necessary and the boundary value of the sample mass of the ith sample that is added to the composite sample is given by:

$$M_i = M_{min,s}\left(\bigcup_{j=1}^{i-1} S_j\right) - \sum_{j=1}^{i-1} M_{sample}(S_j)$$

  in which $\bigcup_{j=1}^{i-1} S_j$ is the composite sample containing the samples $S_1, S_2,\dots, S_{i-1}$.

- The procedure is terminated if $M_{min,s}\left(\bigcup_{j=1}^{i} S_j\right) \leq \sum_{j=1}^{i} M_{sample}(S_j)$, for any value of i.

In the next paragraph, the relative standard deviation of the composite sample obtained with the proposed procedure is investigated.

8.5 Relative variance

The proposed procedure to draw additional samples from the batch stops if the estimated minimum sample mass is equal to or smaller than the obtained sample mass. This results in:

$$M_{\text{sample}} \geq M_{\text{min,s}} = \frac{M_{\text{sample}} V_{\text{sample}}\left(a_{\text{sample}}\right)}{\alpha^2 a_{\text{sample}}^2} \qquad \textbf{(8.13)}$$

This can be written as:

$$\alpha^2 \geq \frac{V_{\text{sample}}\left(a_{\text{sample}}\right)}{a_{\text{sample}}^2} \qquad \textbf{(8.14)}$$

It follows that the procedure is terminated only if the estimated relative variance is smaller than or equal to α^2. Hence, the procedure assures that the estimated relative variance, $V_{\text{sample}}\left(a_{\text{sample}}\right)/a_{\text{sample}}^2$, is smaller than or equal to α^2. However, this is not necessarily guaranteed for the relative variance, $V_{\text{rel}}\left(a_{\text{sample}}\right) = V\left(a_{\text{sample}}\right)/a_{\text{batch}}^2$. To investigate this effect, two situations are distinguished: (*i*) $M_l > M_{\text{min}}$ and (*ii*) $M_l \leq M_{\text{min}}$.

In the first situation, the initial boundary value of the sample mass is larger than the minimum sample mass. If for every sample S, the estimate for the minimum sample mass is exactly equal to the minimum sample mass, *i.e.* $M_{\text{min,s}}(S) = M_{\text{min}}$, no additional samples would have to be drawn. In this case, sampling corresponds to mass-based multinomial selections with boundary value of the sample mass equal to M_l. Because $M_l > M_{\text{min}}$, the relative standard deviation is smaller than α. In practice, due to possible upward statistical fluctuations in $M_{\text{min,s}}(S)$, it can occur that $M_l < M_{\text{min,s}}(S)$ and thus additional samples have to be drawn. Because variance decreases with increasing sample mass, it is expected that this effect can only lead to a reduction of the variance and not to an increase. Therefore, if M_l is larger than M_{min}, the relative standard deviation is smaller than or equal to α.

In the second situation, the initial boundary value of the sample mass is smaller than or equal to the minimum sample mass. If for every sample S, the estimate for the minimum sample mass would be exactly equal to the minimum sample mass, *i.e.* $M_{\text{min,s}}(S) = M_{\text{min}}$, the finally obtained composite sample will correspond to a sample with boundary value of the sample mass equal to M_{min} and hence will have a relative standard deviation equal to α. Downward statistical fluctuations in $M_{\text{min,s}}(S)$ can lead to a finally obtained composite sample mass smaller than the theoretical minimum sample mass, if $M_{\text{min,s}}(S)$ is smaller than the finally obtained composite sample mass. It is expected that at low values of M_l the estimated minimum sample mass will have much larger statistical fluctuations than at higher values of M_l. Therefore, the occurrence of a finally obtained composite sample mass smaller than the theoretical minimum sample mass will be greater at low values of M_l. Hence, it is expected that at low values of M_l there can be a breakdown of the proposed mechanism: the actual relative standard deviation of the finally obtained composite sample is larger than α.

Hence, although the proposed procedure is objective, it does not guarantee that the relative standard deviation of the concentration in the finally obtained composite sample always stays below the warranted α at low values for M_l. In the following, simulations will be applied to investigate the relation between the initial sample mass and the relative standard deviation of the finally obtained composite sample. Because the estimated minimum sample mass depends on the concentrations and particle masses, simulations were performed with different batches, with a typical particle mass of 1 g. During the simulations, the value of α was fixed at 0.01. For several batches and values of the initial boundary value of the sample mass, M_l, the proposed feedback procedure was repeated 10^4 times. During each feedback procedure, every sample S_i (with i=1,2,3,...) was drawn according to the mass-based multinomial selection scheme. Before a new feedback procedure was started, the particles that were sampled during the previous feedback procedure were put back in the batch, so that all the 10^4 finally obtained composite samples were independent realizations of an identical statistical distribution. In Figure 8.2 the particle masses and concentrations of the studied batches, which represent a range of extreme particle distributions, are graphically defined.

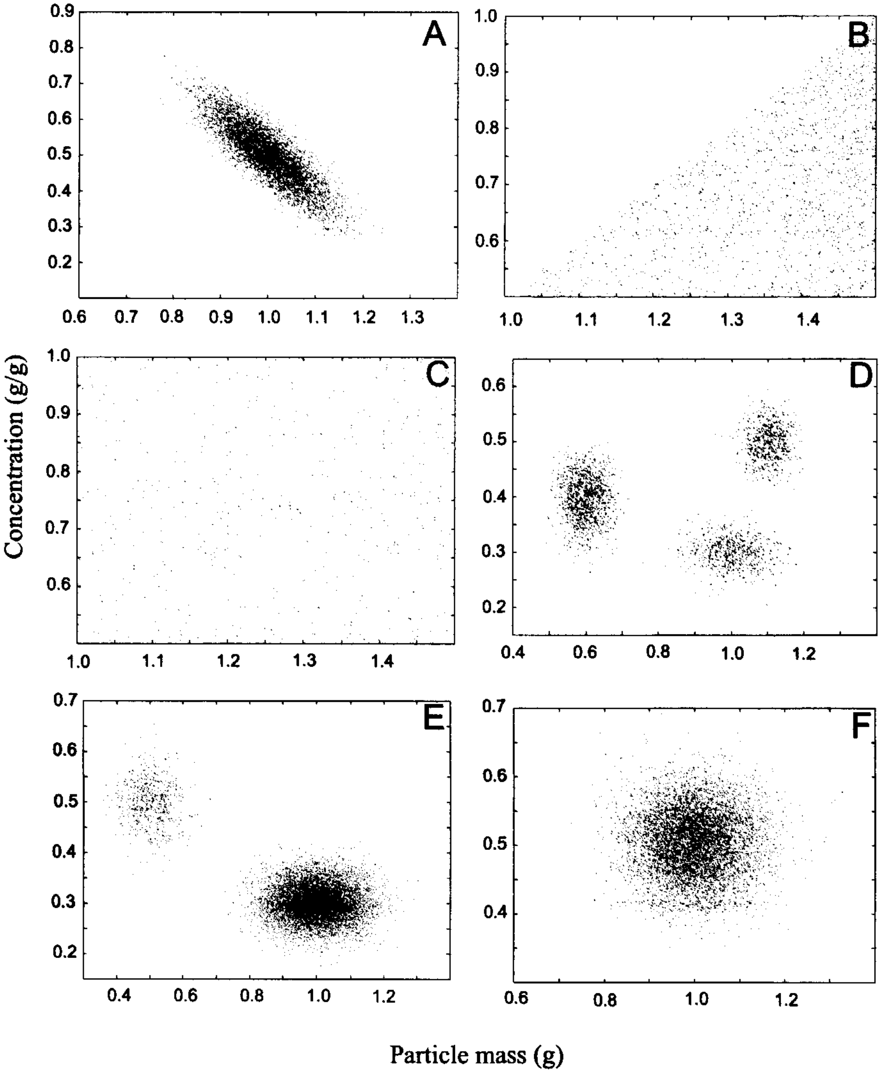


Figure 8.2. *Six particle distributions used for simulations. Every dot represents a particle in the batch. The horizontal axis represents the mass and the vertical axis denotes the concentration in the particle. The total number of particles in each batch is* 10^4.

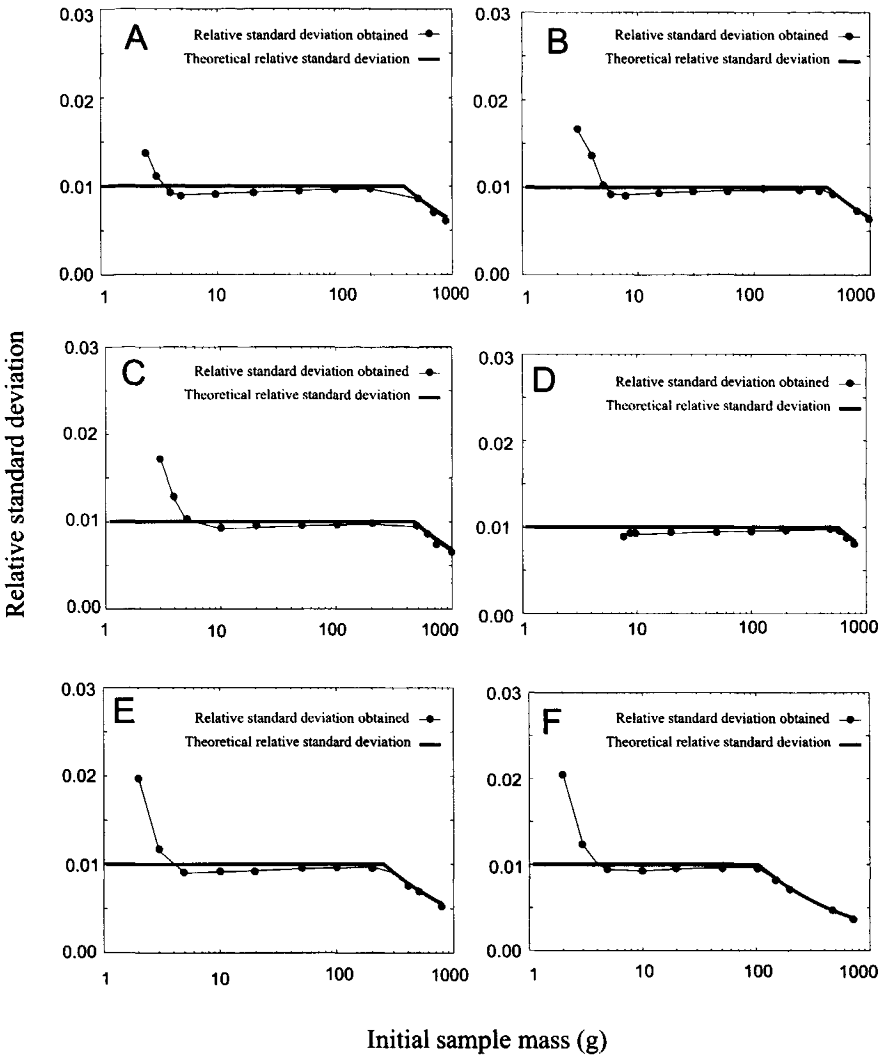


Figure 8.3. *Relative standard deviation calculated with simulated samples as a function of the initial sample mass and the particle distribution from Figure 8.2. For graph A to F in this figure the particle distributions A to F in Figure 8.2 respectively were used. For each point, 10^4 samples were simulated. The solid line represents the theoretical relative standard deviation (for which it is assumed that* $M_{min,s} = M_{min}$ *and* $V(a_{sample})=V_{batch}(a_{sample})$*): 0.01 when the initial sample mass is smaller than* $M_{min,b}$ *and equal to the square root of the relative variance estimated using the properties of the particles in the batch (Equation 8.5) when the initial sample mass is larger than* $M_{min,b}$.

In Figure 8.3, the relative standard deviation of the 10^4 finally obtained composite samples is plotted as a function of the initial sample mass M_I. Graphs A to F represent the results of the simulations for batches A to F in Figure 8.2 respectively. For the sampling of batch D with the initial sample mass smaller than 8 g the finally obtained composite sample masses were comparable to the total mass of batch D. Because in this case it would be better to apply $\hat{V}_{fpc}(a_{sample})$, given by Equation 6.18, instead of $V_{sample}(a_{sample})$, given by Equation 5.78, these data points were omitted in Figure 8.3. In Paragraph 8.9, the effect of omitting a finite population correction is discussed. It will be seen that the results depicted in Figure 8.3 are not significantly influenced.

When assuming that the finally obtained composite sample concentrations are normally distributed, a χ^2-analysis can be applied to prove that the 95% confidence bands corresponding to the obtained relative standard deviations are very narrow. Hence, these confidence bands are not depicted in Figure 8.3.

It is observed that, for a large range of values for the initial sample mass (note the logarithmic scale), the proposed mechanism to draw additional sample ensures that the relative standard deviation is not larger than the warranted value for α (here chosen 0.01). Only for very low values of the initial sample mass, can it be seen that the scheme is indeed inadequate. In Figure 8.4 the results are summarized.

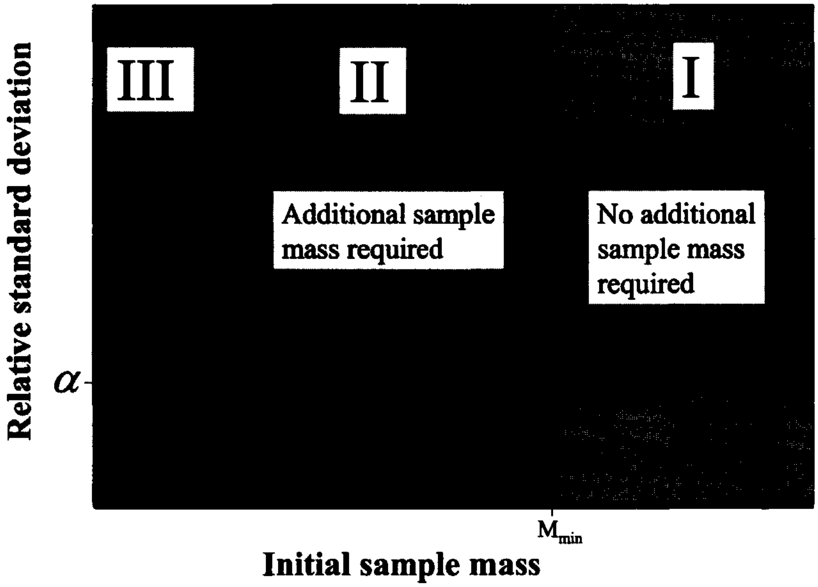

Relative standard deviation
α
III
II
I
Additional sample mass required
No additional sample mass required
M_{min}
Initial sample mass

Figure 8.4. Schematic summary of results obtained with simulations. Depending on the value of the initial sample mass three regions are identified. Only in the third region the relative standard deviation is larger than α.

Three regions can be defined. In region I and II, the sampling scheme proposed is adequate, *i.e.* the relative standard deviation is below the warranted α. In region III this is not guaranteed, due to a too low value of M_1. For batches A to F region III occurs at M_1 below 10 g. Because for these batches the typical particle mass is 1 g, it is expected that in general if a sample contains 10 or more particles, the proposed scheme is adequate, *i.e.* the initial sample mass is in region I or II.

8.6 *Safe value for the variance*

Generally, the mass concentration of a component in the sample can be determined by a single sample analysis instead of analyzing all the particles in the sample separately. Because of this practical convenience, the mass concentration of a component in the sample is extensively used as an estimator for the mass concentration in the batch. An additional advantage, which was demonstrated in Chapter 5, is that the value of this estimator is equal to the value of the unbiased π-expanded estimator, if sampling corresponds to a mass-based multinomial selection of particles in the limit of an infinite sample-to-particle size ratio and an infinite batch-to-sample size ratio. In Chapter 5, a variance estimator $V_{sample}(a_{sample})$, based on the Horvitz-Thompson estimator was derived. However, for evaluation of this estimator, the mass concentrations in all the particles of the sample are required. In practice, these are often unknown.

Therefore, in the next paragraph, a safe value for the estimated variance will be calculated, which is always larger than or equal to the actual variance $V_{sample}(a_{sample},S)$, which would be obtained if all the particles belonging to S were analyzed for their mass concentrations. This will subsequently lead to a larger estimate of the minimum sample mass (see Equation 8.9). Because larger samples lead to smaller variance, it can be expected that if the procedure depicted in Figure 8.1 is modified so that instead of $M_{min,s}$ a larger value is used, the general trend depicted in Figure 8.4, remains valid as an upper bound for the relative standard deviation. Hence, it can still be guaranteed that the relative standard deviation is smaller than α if the initial sample mass is in region I or II. The advantage of application of the safe value of the variance will be that its value can be calculated without analyzing all particles in the sample individually for their mass concentrations.

As an introduction to the calculation of a safe value for the variance, a general technique to obtain extreme values of a function that depends on several variables is discussed. The technique of Lagrange multipliers (see *e.g.* Arfken, 1985) can be applied when searching the extreme values (including maximums, minimums or saddle points) of a function g that depends on multiple variables x_i, denoted as $g \equiv g(x_1,\ldots,x_N)$. When there are no constraints on the variables x_i, the extreme values can be found by equating the partial derivatives of g with respect to the x-variables to zero. This yields N equations for N unknowns. When there is a constraint expressed in the form $f(x_1,\ldots,x_N)=0$ there are N+1 equations with only N unknowns. This may lead to an unsolvable system. Therefore, an extra variable λ, termed as the Lagrange multiplier, is introduced. The partial derivatives with respect to x_i (for all i between 1 and N) and to λ of

$$g(x_1,\ldots,x_N)+\lambda f(x_1,\ldots,x_N) \qquad \textbf{(8.15)}$$

are equated to zero in search for its extreme values. This yields the following system of equations:

$$\frac{\partial}{\partial x_i}\left[g\left(x_1,\ldots,x_N\right)+\lambda f\left(x_1,\ldots,x_N\right)\right]=0 \qquad \mathbf{(8.16)}$$

$$f\left(x_1,\ldots,x_N\right)=0 \qquad \mathbf{(8.17)}$$

Because the final equation is identical to the constraint imposed on the variables x_i, solutions will yield the extreme values of $g(x_1,\ldots,x_N)$ under the constraint $f(x_1,\ldots,x_N)=0$.

In the next paragraph, the above-described technique is applied for the variance estimate $V_{sample}(a_{sample},S)$.

8.7 Calculation of a safe value for the variance

In this paragraph, it is assumed that for the sample S, the value of the sample concentration, denoted as $a_{sample}(S)$, is known and the concentrations a_i are unknown. Therefore, values for the variables a_i will be searched, for which the estimated variance, $V_{sample}(a_{sample},S)$, has a maximum value. It is assumed that every particle in the sample forms a distinct class, hence $T=N_{sample}(S)$ and $N_i(S)=1$ for all i between 1 and T. In a first calculation, an extreme value for $V_{sample}(a_{sample},S)$ will be found by using a Lagrange multiplier and the constraint imposed on the variables a_i $(i=1,\ldots,N_{sample}(S))$:

$$\sum_{j=1}^{N_{sample}(S)} m_j a_j - M_{sample}(S)a_{sample}(S)=0 \qquad \mathbf{(8.18)}$$

When the above equation is satisfied, the mass concentration in a sample in which the concentration in the j^{th} particle is given by a_j is equal to $a_{sample}(S)$. For each possible value of i between 1 and $N_{sample}(S)$, differentiation to a_i of

$$V_{sample}\left(a_{sample},S\right)+\lambda\left(\sum_{j=1}^{N_{sample}(S)} m_j a_j - M_{sample}(S)a_{sample}(S)\right) \qquad \mathbf{(8.19)}$$

yields the following equation:

$$\frac{2m_i^2\left(a_i - a_{sample}(S)\right)+\lambda m_i}{M_{sample}(S)\left(M_{sample}(S)-M_{sample}(S)/N_{sample}(S)\right)}=0 \qquad \mathbf{(8.20)}$$

Equation 8.18 and the $N_{sample}(S)$ results from Equation 8.20 for all possible values of i between 1 and $N_{sample}(S)$ form a system of $N_{sample}(S)+1$ equations. The only solution of this system is that all a_i are given by $a_i=a_{sample}(S)$ for all i between 1 and $N_{sample}(S)$ (and $\lambda=0$). When this result is substituted back in Equation 5.78, it is found that $V_{sample}(a_{sample},S)=0$. Because the estimated sample-to-sample variance is a non-negative quantity it is concluded that a minimum was found. Below, the procedure will be modified in order to obtain a maximum value.

Solutions for the system of equations obtained in a Lagrange procedure contain all possible maximums, minimums or saddle points. However, in the previous calculations only a minimum was found, in spite of the fact that a maximum must exist. The failure can be illustrated in a simple general one-dimensional case.

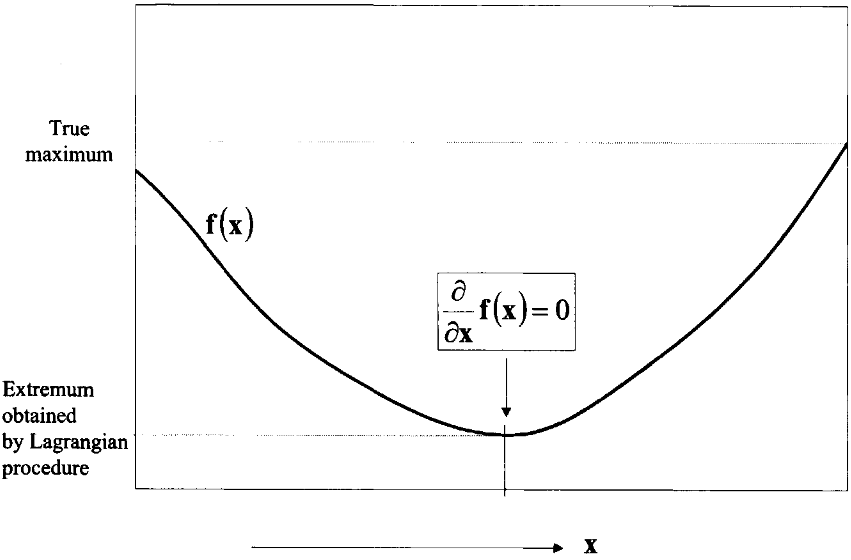

True maximum
f(x)
$\frac{\partial}{\partial x}f(x)=0$
Extremum obtained by Lagrangian procedure
x

Figure 8.5. *The true maximum of a function* f(x) *is not always found by equating the partial derivative to zero. The function* f(x) *has a maximum at the maximum value of* x.

In Figure 8.5, it can be seen that the true maximum is attained at the boundary of the range at which x is defined. A similar situation occurred in the first calculation. Therefore, a transformation of variables is applied. Because each a_i may vary between the minimum and maximum concentration in a particle in the sample, denoted as a_{min} and a_{max} respectively, the following substitution is suggested:

$$a_i = \frac{a_{max} + a_{min}}{2} + \frac{a_{max} - a_{min}}{2}\sin(\phi_i) \qquad \textbf{(8.21)}$$

It follows that a_i cannot be smaller than a_{min} or larger than a_{max}. However the new variables φ_i may take any value between $-\infty$ and $+\infty$. Hence, if a Lagrange procedure is applied using the new variables φ_i, it is expected that all extreme values are found, even the extremes that correspond to one or more values of a_i equal to a_{min} or a_{max}.

Note that instead of the function sin(φ_i) other periodic and differentiable functions that vary between −1 and +1 can be used. This will however not influence the final results, so therefore the well-known sinus function is chosen here. Substitution in Equation 5.78 yields:

$$V_{sample}\left(a_{sample}, S\right) = \frac{\sum_{j=1}^{N_{sample}(S)} m_j^2 \left(\frac{a_{max} + a_{min}}{2} + \frac{a_{max} - a_{min}}{2} \sin\left(\phi_j\right) - a_{sample}(S) \right)^2}{M_{sample}(S)\left(M_{sample}(S) - M_{sample}(S)/N_{sample}(S)\right)} \tag{8.22}$$

The constraint, Equation 8.18, becomes:

$$\sum_{j=1}^{N_{sample}(S)} m_j \left(\frac{a_{max} + a_{min}}{2} + \frac{a_{max} - a_{min}}{2} \sin\left(\phi_j\right) \right) - M_{sample}(S)a_{sample}(S) = 0 \tag{8.23}$$

The first $N_{sample}(S)$ Lagrange equations, obtained by $N_{sample}(S)$ partial derivatives with respect to φ_i, become:

$$\frac{\frac{\partial}{\partial \varphi_i} \sum_{j=1}^{N_{sample}(S)} m_j^2 \left(\frac{a_{max} + a_{min}}{2} + \frac{a_{max} - a_{min}}{2} \sin\left(\varphi_j\right) - a_{sample}(S) \right)^2}{M_{sample}(S)\left(M_{sample}(S) - M_{sample}(S)/N_{sample}(S)\right)}$$

$$+ \lambda \frac{\partial}{\partial \varphi_i} \sum_{j=1}^{N_{sample}(S)} m_j \left(\frac{a_{max} + a_{min}}{2} + \frac{a_{max} - a_{min}}{2} \sin\left(\varphi_j\right) \right) - \lambda \frac{\partial}{\partial \varphi_i} M_{sample}(S)a_{sample}(S) = 0 \tag{8.24}$$

Performing the partial differentiations results in:

$$\frac{2m_i^2 \left(\frac{a_{max} + a_{min}}{2} + \frac{a_{max} - a_{min}}{2} \sin\left(\varphi_i\right) - a_{sample}(S) \right)}{M_{sample}(S)\left(M_{sample}(S) - M_{sample}(S)/N_{sample}(S)\right)} \frac{a_{max} - a_{min}}{2} \cos\left(\varphi_i\right)$$

$$+ \lambda m_i \frac{a_{max} - a_{min}}{2} \cos\left(\varphi_i\right) = 0 \tag{8.25}$$

There is no unique solution of the system of equations (Equation 8.23 and the $N_{sample}(S)$ results provided by Equation 8.25 for all i between 1 and $N_{sample}(S)$). Therefore, the solution that yields the largest value for the variance when it is resubstituted is the true maximum. When for all i (between 1 and $N_{sample}(S)$) $\cos(\varphi_i)$ is non-zero, the solution is again $a_i=a_{sample}(S)$ for all i between 1 and $N_{sample}(S)$). Therefore, in order to find the maximum, for at least one or more i-values, $\cos(\varphi_i)$ must be zero. When this is the case, $\sin(\varphi_i)$ is either +1 or −1 and $a_i=a_{max}$ or $a_i=a_{min}$ respectively.

In view of the above observation, it is convenient to define the indicators $I_{max}(i)$, $I_{min}(i)$ and $I_{other}(i)$ as follows:

- $I_{max}(i)=1$ if $a_i=a_{max}$ and zero otherwise,
- $I_{min}(i)=1$ if $a_i=a_{min}$ and zero otherwise, and
- $I_{other}(i)=1$ if $a_i\neq a_{max}$ and $a_i\neq a_{min}$ and zero otherwise

Using these definitions, the constraint (Equation 8.23) becomes:

$$\sum_{j=1}^{N_{sample}(S)} I_{max}(j)m_j a_{max}+\sum_{j=1}^{N_{sample}(S)} I_{min}(j)m_j a_{min}$$
$$+\sum_{j=1}^{N_{sample}(S)} I_{other}(j)m_j\left(\frac{a_{max}+a_{min}}{2}+\frac{a_{max}-a_{min}}{2}\sin(\varphi_j)\right)-M_{sample}(S)a_{sample}(S)=0 \quad \textbf{(8.26)}$$

and the $N_{sample}(S)$ equations, provided by Equation 8.25, becomes for all values of i satisfying $I_{other}(i)=1$:

$$\frac{2m_i^2\left(\frac{a_{max}+a_{min}}{2}+\frac{a_{max}-a_{min}}{2}\sin(\phi_i)-a_{sample}(S)\right)}{M_{sample}(S)\left(M_{sample}(S)-M_{sample}(S)/N_{sample}(S)\right)}+\lambda m_i=0 \quad \textbf{(8.27)}$$

Resubstituting the definition of a_i into the system of Equation 8.26 and 8.27 yields:

$$\sum_{j=1}^{N_{sample}(S)} I_{other}(j)m_j a_j+\sum_{j=1}^{N_{sample}(S)} I_{max}(j)m_j a_{max}+$$
$$\sum_{j=1}^{N_{sample}(S)} I_{min}(j)m_j a_{min}-M_{sample}(S)a_{sample}(S)=0 \quad \textbf{(8.28)}$$

$$\frac{2m_i^2\left(a_i-a_{sample}(S)\right)}{M_{sample}(S)\left(M_{sample}(S)-M_{sample}(S)/N_{sample}(S)\right)}+\lambda m_i=0 \quad \textbf{(8.29)}$$

The above system of equations can be solved for the variables a_i for all i with $I_{other}(i)=1$. Dividing the left-hand side of Equation 8.29 by m_i yields the following expression:

$$m_i\left(a_i - a_{sample}(S)\right) = -\frac{M_{sample}(S)}{2}\left(M_{sample}(S) - \frac{M_{sample}(S)}{N_{sample}(S)}\right)\lambda \qquad \textbf{(8.30)}$$

Equation 8.28 can be rewritten and the above result can be substituted.

$$-\frac{1}{2}\sum_{j=1}^{N_{sample}(S)} I_{other}(j)\lambda M_{sample}(S)\left(M_{sample}(S) - M_{sample}(S)/N_{sample}(S)\right)$$

$$+\sum_{j=1}^{y} I_{max}(j)m_j\left(a_{max} - a_{sample}(S)\right) + \sum_{j=1}^{z} I_{min}(j)m_j\left(a_{min} - a_{sample}(S)\right) = 0 \qquad \textbf{(8.31)}$$

Defining $M_{max} = \sum_{j=1}^{N_{sample}(S)} I_{max}(j)m_j$ and $M_{min} = \sum_{j=1}^{N_{sample}(S)} I_{min}(j)m_j$ yields an expression for λ:

$$\lambda = \frac{2M_{min}\left(a_{max} - a_{sample}(S)\right) + M_{max}\left(a_{min} - a_{sample}(S)\right)}{M_{sample}(S)\left(M_{sample}(S) - M_{sample}(S)/N_{sample}(S)\right)\sum_{j=1}^{N_{sample}(S)} I_{other}(j)} \qquad \textbf{(8.32)}$$

The above result can be substituted back in Equation 8.30, which results in an equation for a_i (for all values of i with $I_{other}(i)=1$):

$$a_i = a_{sample}(S) - \frac{M_{max}\left(a_{max} - a_{sample}(S)\right) + M_{min}\left(a_{min} - a_{sample}(S)\right)}{m_i \sum_{j=1}^{N_{sample}(S)} I_{other}(j)} \qquad \textbf{(8.33)}$$

Note that it was chosen arbitrarily which particles have $a_i=a_{max}$, $a_i=a_{min}$ or a_i given by Equation 8.33. However, the obtained solution can be substituted in the variance estimator. The extreme value for the variance $\mathrm{Ext}\{V_{sample}(a_{sample},S)\}$ (*i.e.* maximum, minimum or saddle point of $V_{sample}(a_{sample},S)$) consists of three terms:

$$\left.\begin{aligned}
\operatorname{Ext}\left\{V_{\text{sample}}\left(a_{\text{sample}},S\right)\right\} &= \frac{\sum_{j=1}^{N_{\text{sample}}(S)} I_{\max}(j)m_j^2\left(a_{\max}-a_{\text{sample}}(S)\right)^2}{M_{\text{sample}}(S)\left(M_{\text{sample}}(S)-M_{\text{sample}}(S)/N_{\text{sample}}(S)\right)} \\
&+\frac{\sum_{j=1}^{N_{\text{sample}}(S)} I_{\min}(j)m_j^2\left(a_{\min}-a_{\text{sample}}(S)\right)^2}{M_{\text{sample}}(S)\left(M_{\text{sample}}(S)-M_{\text{sample}}(S)/N_{\text{sample}}(S)\right)} \\
&+\frac{\sum_{j=1}^{N_{\text{sample}}(S)} I_{\text{other}}(j)m_j^2\left(a_j-a_{\text{sample}}(S)\right)^2}{M_{\text{sample}}(S)\left(M_{\text{sample}}(S)-M_{\text{sample}}(S)/N_{\text{sample}}(S)\right)}
\end{aligned}\right\} \qquad \mathbf{(8.34)}$$

Substituting the obtained expression for a_j (Equation 8.33) yields after rewriting:

$$\left.\begin{aligned}
\operatorname{Ext}\left\{V_{\text{sample}}\left(a_{\text{sample}},S\right)\right\} &= \frac{\sum_{j=1}^{N_{\text{sample}}(S)} I_{\max}(j)m_j^2\left(a_{\max}-a_{\text{sample}}(S)\right)^2}{M_{\text{sample}}(S)\left(M_{\text{sample}}(S)-M_{\text{sample}}(S)/N_{\text{sample}}(S)\right)} \\
&+\frac{\sum_{j=1}^{N_{\text{sample}}(S)} I_{\min}(j)m_j^2\left(a_{\min}-a_{\text{sample}}(S)\right)^2}{M_{\text{sample}}(S)\left(M_{\text{sample}}(S)-M_{\text{sample}}(S)/N_{\text{sample}}(S)\right)} \\
&+\frac{\left[M_{\max}\left(a_{\max}-a_{\text{sample}}(S)\right)+M_{\min}\left(a_{\min}-a_{\text{sample}}(S)\right)\right]^2}{xM_{\text{sample}}(S)\left(M_{\text{sample}}(S)-M_{\text{sample}}(S)/N_{\text{sample}}(S)\right)\sum_{j=1}^{N_{\text{sample}}(S)} I_{\text{other}}(j)}
\end{aligned}\right\} \qquad \mathbf{(8.35)}$$

To eliminate the summation symbols the following inequalities are used:

$$\sum_{j=1}^{N_{\text{sample}}(S)} I_{\max}(j)m_j^2 \le m_{\max}(S)\sum_{j=1}^{N_{\text{sample}}(S)} I_{\max}(j)m_j = m_{\max}(S)M_{\max} \qquad \mathbf{(8.36)}$$

$$\sum_{j=1}^{N_{\text{sample}}(S)} I_{\min}(j)m_j^2 \le m_{\max}(S)\sum_{j=1}^{N_{\text{sample}}(S)} I_{\min}(j)m_j = m_{\max}(S)M_{\min} \qquad \mathbf{(8.37)}$$

$$\frac{1}{\sum_{j=1}^{N_{sample}(S)} I_{other}(j)} \le \frac{m_{max}(S)}{M_{other}} \tag{8.38}$$

in which $M_{other} = \sum_{j=1}^{N_{sample}(S)} I_{other}(j)m_j$ and $m_{max}(S)$ is the largest particle mass in the sample. Also the following equality is substituted:

$$M_{max}\left(a_{max} - a_{sample}(S)\right) + M_{min}\left(a_{min} - a_{sample}(S)\right) = M_{other}\left(a_{sample}(S) - a\right) \tag{8.39}$$

in which a is the mass concentration in the collection of particles which do not have a maximum or minimum concentration, *i.e.*:

$$a = \sum_{j=1}^{N_{sample}(S)} I_{other}(j)a_j m_j \Big/ \sum_{j=1}^{N_{sample}(S)} I_{other}(j)m_j \tag{8.40}$$

The above substitutions lead to a supreme value for (here defined as a value that is always larger than or equal to) each possible extreme value for the estimated variance:

$$\mathrm{Ext}\left\{V_{sample}\left(a_{sample}, S\right)\right\} \le \frac{m_{max}(S)\left\{M_{max}\left(a_{max} - a_{sample}(S)\right)^2 + M_{min}\left(a_{min} - a_{sample}(S)\right)^2 + M_{other}\left(a - a_{sample}(S)\right)^2\right\}}{M_{sample}(S)\left(M_{sample}(S) - M_{sample}(S)/N_{sample}(S)\right)} \tag{8.41}$$

The values of M_{max}, M_{min} and M_{other} were arbitrary, within the following constraints:

$$M_{sample}(S) = M_{max} + M_{min} + M_{other} \tag{8.42}$$

$$a_{sample}(S)M_{sample}(S) = a_{max}M_{max} + a_{min}M_{min} + aM_{other} \tag{8.43}$$

When Equations 8.42 and 8.43 are used to eliminate a and M_{other} in the supreme value for each possible extreme value for the estimated variance, and the partial derivatives with respect to M_{max} and M_{min} are set to zero (in order to find the combination of M_{max} and M_{min} that yields maximum value), it is found that

$$M_{max} = \frac{a_{sample}(S) - a_{min}}{a_{max} - a_{min}} M_{sample}(S) \tag{8.44}$$

$$M_{min} = \frac{a_{sample}(S) - a_{max}}{a_{min} - a_{max}} M_{sample}(S) \tag{8.45}$$

As a consequence, $M_{other}=0$. Substitution of these results into Equation 8.41 yields a single supreme value for all possible extreme values of the sample-to-sample variance. This is equivalent to an overall supreme value of the sample-to-sample variance:

$$V_{sample}(a_{sample}, S) \leq \frac{m_{max}(S)(a_{max} - a_{sample}(S))(a_{sample}(S) - a_{min})}{M_{sample}(S) - M_{sample}(S)/N_{sample}(S)} \tag{8.46}$$

With respect to the distribution of particle masses in the sample, only the maximum particle mass is needed for evaluation of the right-hand side of the above expression.

8.8 Results

Equations for the minimum sample mass were derived using the properties of the particles in the batch or sample (Equation 8.7 or 8.12 respectively).

A scheme to draw additional sample mass was proposed. The scheme assures that the relative standard deviation estimated using the properties of the particles in the sample does not exceed a preselected value α. Simulations show that for batches A to F in Figure 8.2 the scheme is adequate if the initial sample mass is 10 g or more. Because the typical particle mass for the studied batches was 1 g, the scheme is adequate when a sample contains 10 or more particles.

A maximum value for the estimated variance was derived in case the individual particle concentrations are unknown. A formula (Equation 8.46) is obtained in which only the maximum particle mass and minimum and maximum concentration in the sample are used. When there is no prior knowledge, the two latter quantities are respectively zero and one.

8.9 Discussion

The simulations in Paragraph 8.5 demonstrated that, for batches A to F the scheme is adequate when a sample contains 10 or more particles. Because batches A to F form a wide range of distinct batches, it might be concluded that the scheme is adequate for every arbitrary batch when a sample contains 10 or more particles. More investigation is necessary to strengthen this anticipated conclusion.

Another point of discussion is that, in some cases, the estimated minimum sample mass, $M_{min,s}$ may become comparable to the batch mass ($\approx 0.1 \times M_{batch}$). For calculation of the estimated minimum sample mass, it is then recommended to apply the finite population corrected variance estimator $\hat{V}_{fpc}(a_{sample})$, instead of $V_{sample}(a_{sample})$ in Equation 8.9.

To gauge the influence of omitting a finite population correction in the simulations with batches A to F, the number of times the sample mass was larger than $0.1 \times M_{batch}$ was counted for every point in Figure 8.3. For every point, less than 10% of the simulated samples were heavier than 0.1 times the batch mass. Therefore, it can be concluded that the simulated sampled batches were large enough to justify the absence of a finite population correction.

By comparing Equation 5.78 and 6.18 it follows that $\hat{V}_{fpc}(a_{sample}) = (1 - M_{sample}/M_{batch})V_{sample}(a_{sample})$. Therefore, when the mass of the batch is known, the safe value of $V_{sample}(a_{sample})$, obtained in Paragraph 8.7, can be multiplied by a factor $(1-M_{sample}/M_{batch})$ to find a safe value for $\hat{V}_{fpc}(a_{sample})$. In this way, the results obtained in Paragraph 8.7 can slightly be modified to encompass a finite population correction.

8.10 References

G. Arfken (1985) *Mathematical Methods for Physicists,* Third Edition, Academic Press, Inc, 985 pp.

V. Barnett (1974) *Elements of sampling theory,* English University Press, London, 152 pp.

Chapter 9 Final remarks

This thesis describes the development of a new, non-empirical theory for the sampling of randomly mixed batches of particulate material. The theory can be applied, among others, to calculate the minimum sample mass for use in sampling standards which aim to limit the sampling error to a maximum allowable value. However, limiting the sampling error is not a fundamental requirement of sampling standards in general. Standards may also prescribe a value for the sample mass or volume without taking into account the sampling error associated with the specified sample size. These standards do not require the model-based determination of sample size as outlined in this thesis.

In order to identify the strengths and weaknesses of existing sampling theories, eight general criteria were defined. While other less general criteria may exist, none of the existing sampling theories met all eight criteria. Assuming that none of the reviewed theories are modified, it was concluded that a new theory was required which meets all criteria. It should be noted that the methods used in this study may also be applied to other theories. For example, the Horvitz-Thompson estimator was applied to derive an equation for the sampling variance based on the particle properties in the sample. It is interesting to consider whether the Horvitz-Thompson estimator could be applied in the non-empirical theories of Wilson or Gy.

While the new theory is based on a model for "ideal sampling" from "a random arrangement of particles", only the sampling error due to the distribution of non-identical particles is considered. It does not consider the error due to non-ideal sampling which, for successful application of the theory, must be assumed to be minimized by the practical choice of the operating conditions during sampling. As a result of non-ideal sampling, the actual sample variance may differ from the calculated value. Sampling from "a random arrangement of particles" limits the applicability of the new theory because order or structure may be observed in actual practice. For example, the situation is considered in which a completely segregated batch is sampled containing only two classes of particles: black and white particles. It is assumed that, as a result of segregation, the majority of samples consist entirely of black particles or entirely of white particles. Hence, it would be inappropriate to use the variance estimator provided in this thesis, because its value would be zero for the majority of samples, leading to an underestimation of the true variance. Order or structure could be incorporated into the theory by taking into account the 'variance due to spatial variation'. This has the drawback of introducing an empirical parameter. Furthermore, a model that includes segregation and/or aggregation is likely to be complicated and may yield cumbersome equations.

The new theory models sampling as repeated selections of particles from the batch with the probability of drawing any particle during a specific selection being constant. This mode of sampling is denoted as "equal probability sampling". In practice, deviations from equal probability sampling may occur. For example, when particles are drawn as clusters, the probability of a particle being drawn during a selection is only equal for particles belonging to the same cluster. The current model does not (yet) include these types of sampling processes. For developing a model that can describe sampling processes from batches containing clusters of particles, the basic assumption of equal probability sampling must be abandoned. A promising extension of the current model is to model sampling using unequal selection probabilities which may depend on the properties of the previously selected particles.

The theory provides equations for the expected value and variance of the sample concentration in the limit of an infinite batch-to-sample size ratio and an infinite sample-to-particle size ratio. An expression for the size-variance relationship is obtained using properties of particles in the batch. Under suitable conditions, this equation is exact for finite values of the batch-to-sample size ratio and the sample-to-particle size ratio. However, the actual variance may differ from the calculated variance even if sampling corresponds to ideal sampling from a random arrangement of particles. A more precise equation is likely to be very complicated, which may reduce its attractiveness for practical application. Practical application also suffers if evaluation of the model parameter requires a large number of measurements. Although a method was developed to estimate parameters from a series of samples, the difference between the value estimated and the actual variance remains unknown.

The equation for the variance, estimated using the properties of the particles in the sample, was based on the Horvitz-Thompson estimator. This does not preclude that the estimator may be biased for finite values of the batch-to-sample size ratio and the sample-to-particle size ratio. Even if the estimator is unbiased and sampling corresponds to ideal sampling from a random arrangement of particles, its value will differ from sample to sample and, for a specific sample, may differ from the actual variance. This raises the question of the practical significance of bias. The relative bias, derived from simulations on a wide range of distinct batch compositions, was plotted in a nomogram. It is found that the absolute value of the relative bias in the sample concentration, caused by a finite sample-to-particle size ratio, did not exceed half the inverse of the sample-to-particle size ratio. Another conclusion from the simulated batches is that bias caused by a finite batch-to-sample size ratio in the estimated variance is always positive. It is worthwhile to investigate whether these findings are generally valid.

It is possible that the variance calculated with the theory may differ from the variance which would be obtained if the sample drawing process corresponds exactly to the assumed model. Such a difference could be characterized in terms of the "level of contradiction". Using bootstrapping on four samples, it was shown that the new theory exhibits smaller levels of contradiction than the theories of Wilson and Gy. The new theory also provides a more normal estimator for the batch concentration. However, more insight into the level of contradiction and normality of the estimators of the new theory is required. It would be useful to establish the level of contradiction and normality as a function of the batch-to-sample size ratio, the sample-to-particle size ratio and possibly other factors.

Combining the equation provided by the theory with knowledge of the properties of particles in the batch or sample allows determination of the minimum sample mass. It was demonstrated that the minimum sample mass is accurate if the sample-to-particle size ratio is larger than 10. More study is required to determine the validity of this seemingly arbitrary number. In general, it may be impossible to analyze each particle individually so that unavailable sample information has to be estimated. It should be noted that less sample information should imply a higher calculated minimum sample mass.

Two applications of the determination of the variance associated with the sampling of particulate materials are envisaged: the determination of the minimum sample mass and an application which has not been considered in this thesis: using the variance to assist in decision-making based on sample analyses. For quality and process control, particulate samples are routinely extracted and analysed. With control in general, the objective is to ensure that a limiting value of a property of interest is either exceeded or undercut. Direct comparison of the sample analysis with the limiting value is not possible because the sample analysis does not necessarily reflect the value of the property in the batch. The potential lack of representativity, caused by the sample being smaller than the batch, can be characterized in terms of the variation with respect to the limiting value. By taking into account the variation, a critical value can be defined which is used to make a decision. A decision is made by comparing the critical value with the sample analysis. The permissible variation can be expressed in terms of a number of standard deviations, where the standard deviation is calculated using an appropriate equation provided by the new theory.

Appendix

The Dutch standard NEN 5742 defines scope (sediments and soils), measurands (metals, inorganic compounds, semi-volatile organic compounds and physico-chemical soil properties) and sampling devices to be used. The standard prescribes the way of sampling, which includes a prescribed value for the mass of the sample. Finally, packaging, conservation and transport of the samples drawn and the essential elements of reporting are described.

The prescribed value for the sample mass in the NEN 5742 can be calculated using Gy's theory of particulate materials (Gy, 1979). In the theory of Gy, the sample drawing is modelled using Bernoulli sampling. In a first-order approximation, in which large variations of the sample mass from its expected value are neglected, the variance of the sample concentration becomes:

$$V\left(a_{\text{sample}}\right)=\frac{1-q}{qM_{\text{batch}}^{2}}\sum_{i=1}^{N_{\text{batch}}} m_i^2\left(a_i-a_{\text{batch}}\right)^2 \qquad \text{(A.1)}$$

A derivation of Gy's basic equation (equation A.1) was obtained by analyzing a mixture of distinct types of materials. Several assumptions were required. Firstly, it was assumed that the particles in the batch can be classified according to volume and type of material and that the concentration in a particle does not vary between particles of a given material type. Secondly, it was assumed that the size distribution in the batch of particles belonging to distinct material types is identical. Thirdly, it was assumed that the volume of each particle in the batch is given by a constant factor f, multiplied by the cube of the particle diameter. Using these assumptions about the composition of the sampled batch and the particle size distribution, Gy obtained the following equation for the factor $\left(1/M_{\text{batch}}\right)\sum_{i=1}^{N_{\text{batch}}} m_i^2\left(a_i-a_{\text{batch}}\right)^2$ in Equation A.1:

$$\frac{1}{M_{batch}} \sum_{i=1}^{N_{batch}} m_i^2 \left(a_i - a_{batch}\right)^2 = d_{max}^3 fg\ell c \qquad (A.2)$$

where

d_{max} =	the typical maximum particle diameter (determined by sieving),
f =	the shape factor,
g =	the size range factor,
ℓ=	the liberation factor, and
c =	the mineralogical composition factor of the material.

The precise relationship between the above introduced parameters and the masses m_i and concentrations a_i of the particles of the batch can be found in Gy (1979). It is noted that Gy assumes that ℓ is smaller than or equal to one, although this is not necessarily true for arbitrary batches. Assuming that $qM_{batch}=M_{sample}$ and $1-q$ can be approximated by one, the condition that the relative standard deviation, $\sqrt{V\left(a_{sample}\right)/a_{batch}^2}$, does not exceed α, results in the following value of the minimum sample mass, M_{min}:

$$M_{min} = \frac{1}{\alpha^2 a_{batch}^2} d_{max}^3 fg\ell c \qquad (A.3)$$

To arrive at the value for the sample mass given in the NEN 5742, it is assumed that $c=a_{batch}(1-a_{batch})\rho$, where ρ is the density of the particles. It is further assumed that $\ell=1$, the particles are spheres (i.e. $f=\pi/6$), and a_{batch} can be replaced by p, the numerical fraction of particles in the batch that contain the property of interest. The equation for the minimum sample mass becomes

$$M_{min} = \frac{\pi}{6} \times d_{max}^3 \times \rho \times g \times \frac{1-p}{\alpha^2 p} \qquad (A.4)$$

Equation A.4 corresponds to the equation for the minimum sample mass given in Annex C of the NVN 7302 standard. The NEN 5742 uses Equation A.4 and the following assumptions: the maximum particle size is 10 mm, the density ρ of the particles is 2.6×10^3 kg/m^3, $g=0.25$, $\alpha=0.1$, and $p=0.1$. The result is:

$$M_{\min} = \frac{\pi}{6} \times (0.01)^3 \times 2.6 \times 10^3 \times 0.25 \times \frac{1-0.1}{(0.1)^2 \times 0.1} = 0.3\,\text{kg}$$

An obvious drawback is that the assumptions limit the general applicability of this Dutch standard. Even if the assumptions are correct, the relative standard deviation may still be larger than 10% due to possible flaws in Gy's theory. Moreover, if an alternative prescribed sample mass were to be calculated on the basis of different estimates for the maximum particle size, density and fraction of particles containing the property of interest still using Gy's theory, the relative standard deviation could be larger than 10% due to errors in the assumed batch properties.

References

P.M. Gy (1979) *Sampling of Particulate Materials, Theory and Practice*, 1st edition, Elsevier, Amsterdam, 431 pp.

NEN 5742 (September 2001) Soil - Sampling of soil and sediments for the determination of metals, inorganic compounds, semi-volatile organic compounds and physico-chemical soil characteristics.

NVN 7302 (April 1998) Leaching characteristics of solid earthy and stony building and waste materials - Sampling - Sampling of granular materials from static heaps.

Summary

Standardization of sampling requires that the mass or the volume of the sample is prescribed. In current standards, a prescribed value for the sample mass is derived using empirical relations between assumed properties of the batch and the variance of the sampling error. The potential inaccurate empirical relations and assumed batch properties may lead to an under or overestimation of the potential magnitude of the sampling error. Therefore, the objective of the research described in this thesis is the development of a new, non-empirical theory for the sampling of randomly mixed batches of particulate material, to allow for calculation of the minimum sample mass in sampling standards.

Current empirical and non-empirical sampling theories are reviewed in Chapter 2. None of the empirical and non-empirical theories meet all eight criteria identified in Chapter 1. This justifies the development of a new sampling theory to meet all criteria.

In Chapter 3, a mathematical algorithm is presented to serve as a model for ideal sampling from a random arrangement of particles. The concept of ideal sampling is defined and the details of the algorithm are discussed. It is shown that non-ideal sampling and biased sampling are different phenomena, whereas non-ideal sampling can act as a source of biased sampling. The boundary value of the sample size can, with limited effects to the accuracy, be estimated using the sampled mass. Simulations demonstrate the validity of this process.

Because the sample concentration is the ratio of two sample totals, in Chapter 4, the variance of a sample total is studied. It is demonstrated that for calculation of this variance, the covariances between the numbers of particles belonging to the classes in the sample are required. Using a specified method, these covariances are calculated in the size-based approach. As a final result of Chapter 4, the variance of the sample concentration, estimated using the properties of the particles in the batch, is calculated.

In Chapter 5, the Horvitz-Thompson estimator is used to provide a general and unbiased estimate for the variance of the π-expanded estimator. It is demonstrated that the sample concentration can be rewritten in the form of a π-expanded estimator. This indicates that the Horvitz-Thompson estimator can be applied for estimation of the variance of the sample concentration. Because in this study particles are classified, the behaviour of the π-expanded estimator and Horvitz-Thompson estimator under classification is investigated. Derivations of expressions for the first- and second-order inclusion probabilities, using results from Chapter 4, are performed. These expressions are substituted into the π-expanded estimator and the Horvitz-Thompson estimator. This results in an expression for the variance, estimated using the properties of the

particles in the sample. Finally, as an application of the obtained equations for the π-expanded estimator and the variance, the obtained equations are worked out for mass concentrations.

The sample concentration and the estimator for the variance, based on the particles in the sample, provide estimators for the batch concentration and the variance of the sample concentration respectively, which are unbiased under certain conditions. In Chapter 6, the biases are split into a contribution caused by a finite batch-to-sample size ratio and a contribution caused by a finite sample-to-particle size ratio. Only a theoretical calculation of the range of possible values of the bias in the sample concentration caused by a finite sample-to-particle size ratio is presented. For mass concentrations, the remaining biases are investigated using simulations. Finally, nomograms are obtained for the maximum of the absolute value of the relative bias in the sample concentration and the variance estimate using the properties of the particles in the sample.

In Chapter 7, the estimators developed in this study and associated analytical uncertainties are evaluated for four samples. Using the experimental results, the new theory is validated by comparing the level of contradiction of the theory with the level of contradiction of the theories of Wilson and Gy. Also the normality of the sample concentration is investigated. It is shown that the new theory exhibits lower levels of contradiction than the theories of Gy and Wilson and yields also a more normal estimator for the batch concentration.

In Chapter 8, the minimum sample mass is estimated using the properties of the particles in the batch or sample. Using the properties of the particles in the sample, a feedback mechanism is proposed to draw additional samples. The mechanism is investigated with simulations. Knowledge of the maximum particle mass, the minimum and maximum particle concentrations in the sample can be used for estimation of the minimum sample mass.

Bastiaan Geelhoed

Samenvatting

Standaardisatie van bemonstering vereist dat de massa of het volume van het monster wordt voorgeschreven. In de hedendaagse standaarden wordt een waarde voor de voorgeschreven monstermassa afgeleid aan de hand van empirische relaties tussen veronderstelde eigenschappen van de partij en de variantie van de steekproeffout. De mogelijk onjuiste empirische relaties en veronderstelde eigenschappen van de partij kunnen leiden tot een onder- of overschatting van de mogelijke grootte van de steekproeffout. De doelstelling van het in dit proefschrift beschreven onderzoek is daarom de ontwikkeling van een nieuwe, niet-empirische theorie voor het bemonsteren van willekeurig gemengde partijen korrelvormig materiaal, die het mogelijk maakt de minimale monstermassa in bemonsteringsstandaarden te berekenen.

Hedendaagse empirische en niet-empirische bemonsteringstheorieën worden belicht in Hoofdstuk 2. Geen van de empirische en niet-empirische theorieën voldoen aan de in Hoofdstuk 1 gestelde criteria. Dit rechtvaardigt de ontwikkeling van een nieuwe theorie die wel aan alle criteria voldoet.

In Hoofdstuk 3 wordt een trekkingsschema gepresenteerd dat dient als model voor de ideale monstername uit een stochastiche pakking van deeltjes. Het concept ideaal bemonsteren wordt gedefiniëerd en de details van het algorithme worden besproken. Het wordt aangetoond dat niet-ideaal bemonsteren en onzuiver bemonsteren verschillende verschijnselen zijn, waarbij niet-ideaal bemonsteren een bron van onzuiver bemonsteren kan zijn. De grenswaarde voor de monstergrootte kan, met beperkte gevolgen voor de juistheid, worden afgeschat door gebruik te maken van de monstermassa. Simulaties tonen de geldigheid van dit proces aan.

Omdat de monsterconcentratie de ratio van twee steekproeftotalen is, wordt in Hoofdstuk 4 de variantie van een steekproeftotaal bestudeerd. Het wordt aangetoond dat voor de berekening van deze variantie de covarianties tussen de deeltjesaantallen in het monster van deeltjes behorende tot de klassen vereist zijn. Deze covarianties worden, gebruikmakend van een specifieke methode, berekend in de op grootte gebaseerde aanpak. Als laatste resultaat in Hoofdstuk 4 wordt een vergelijking afgeleid, waarmee de variantie van de monsterconcentratie kan worden berekend met behulp van de eigenschappen van de deeltjes in de partij.

In Hoofdstuk 5 wordt de Horvitz-Thompson schatter voor de variantie gebruikt om een algemene en onbevoordeelde schatting voor de variantie te verschaffen. Het wordt aangetoond dat de monsterconcentratie herschreven kan worden in de vorm van een Horvitz-Thompson schatter. Dit duidt erop dat de Horvitz-Thompson schatter toegepast kan worden op de schatting van de variantie van de monsterconcentratie.

Omdat de deeltjes in deze studie worden geklassificeerd, wordt het gedrag onder klassificatie van de Horvitz-Thompson schatters voor de concentratie en de variantie onderzocht. Gebruikmakend van resultaten uit Hoofdstuk 4 worden afleidingen uitgevoerd van uitdrukkingen voor de eerste en tweede orde insluitverwachtingen. Deze uitdrukkingen worden ingevuld in de Horvitz-Thompson schatters. Dit resulteert in een vergelijking waarmee de variantie kan worden geschat aan de hand van de eigenschappen van de deeltjes in het monster. Tenslotte worden, als toepassing, de verkregen vergelijkingen uitgewerkt voor massa concentraties.

De monsterconcentratie en de schatter voor de variantie, gebaseerd op de eigenschappen van de deeltjes in het monster, verschaffen schatters voor respectievelijk de partijconcentratie en de variantie van de monsterconcentratie, die onder bepaalde voorwaarden zuiver zijn. In Hoofdstuk 6 word de onzuiverheid opgesplitst in een bijdrage ten gevolg van een eindige partij-monstergrootte verhouding en een bijdrage ten gevolg van een eindige monster-deeltjesgrootte verhouding. Alleen de maximale en minimale waarden van de onzuiverheid van de monsterconcentratie ten gevolg van een eindige monster-deeltjesgrootte verhouding worden theoretisch berekend. De overige onzuiverheden worden voor massa concentraties door middel van simulaties onderzocht. Tenslotte worden nomogrammen verkregen voor de maximale absolute waarde van de relatieve onzuiverheid in de monsterconcentratie en in de variantie geschat met behulp van de eigenschappen van de deeltjes in het monster.

In Hoofdstuk 7 worden de schatters die in deze studie ontwikkeld zijn en de bijbehorende analytische onzekerheden geëvalueerd voor vier monsters. Gebruikmakend van de experimentele resultaten wordt de nieuwe theorie gevalideerd door het niveau van tegenstrijdigheid te vergelijken met de theorieen van Gy en Wilson. Er wordt aangetoond dat de nieuwe theorie lagere niveaus van tegenstrijdigheid vertoont dan de theorieen van Gy en Wilson en leidt tot een meer gaussische schatter voor de partijconcentratie.

In Hoofdstuk 8 wordt de minimale monstermassa geschat met behulp van de eigenschappen van de deeltjes in de partij of het monster. Gebruikmakend van de eigenschappen van de deeltjes in het monster wordt een terugkopplingsmechanisme om additionele monsters te nemen voorgesteld. Het mechanisme wordt onderzocht met behulp van simulaties. Kennis van de maximale deeltjesmassa, the minimale en maximale concentraties in het monster kan worden gebruikt voor de schatting van de minimale monstermassa.

Bastiaan Geelhoed

Dankwoord

Hoewel het werken aan het onderzoek dat ten grondslag ligt aan dit proefschrift mij gedurende een periode van enkele jaren vele overuren aan werk heeft bezorgd, kan men natuurlijk nooit helemaal alleen iets dergelijks tot een goed einde brengen. Er zijn velen die ik hier wil bedanken.

Mijn dank gaat allereerst uit naar Hylke Glass en Jeroen de Goeij, die in de functie van promotor vele malen geduldig en kritisch de conceptteksten van dit proefschrift hebben gelezen en waar nodig suggesties tot aanpassing hebben gedaan. Ik dank Hylke ook voor zijn inzet gedurende de eerste jaren van het onderzoek waarin hij als één van mijn begeleiders een bron van inspiratie was.

Ook mijn andere begeleider, Peter Bode, verdient mijn dankbetuiging voor zijn inzet en vooral voor zijn projectmatige aanpak gedurende het onderzoek. Gedurende de eerste jaren van het onderzoek heeft hij mij ook met enthousiasme geholpen bij voorbereidingen van presentaties van onderzoeksresultaten op conferenties.

Mijn collega promovendi Maurice, Jasper, Lennard, Jacco, Astrid, Heleen, Ramon, Rafaella, Marnix en Saskia – en met name mijn kamergenoten Erwin en Heleen - ben ik erkentelijk voor alle leuke momenten die ik samen met hen heb beleefd. Daarnaast wil ik graag alle andere medewerkers en gasten bij de groep FMR bedanken voor de goede en gezellige sfeer en alsook de inzet bij de metingen via neutronenactiveringsanalyse.

Fred Bakker en Marjon Stelling van het NFI ben ik erkentelijk voor hun enthousiaste betrokkenheid bij het onderzoek. De regelmatige werkbesprekingen op het IRI of op het NFI met Fred en later voornamelijk met Marjon heb ik ook als zeer positief ervaren.

From CSM, I would like to thank many people, especially those with whom I worked together on the arsenic project: Chris, Simon, Hylke, Sharon, and Fiona. I also would like to thank Kilian and John for their moral support and Phil Hutchings for carefully reading the draft manuscript of this thesis and his help in improvement of English. Also many thanks to all the other members of CCC and CCA.

Tenslotte dank ik mijn familie, vrienden en kennissen voor hun interesse en vooral de morele ondersteuning gedurende mijn promotieperiode.

7 juli 2004 Bastiaan Geelhoed

Curriculum vitae

Bastiaan Geelhoed

Geboren 20 december 1975 te Amsterdam

1988-1994 Voorbereidend Wetenschappelijk Onderwijs aan het Pieter Nieuwland College te Amsterdam.

1994-1998 Experimentele Natuurkunde aan de Vrije Universiteit Amsterdam.

1998-2002 A.I.O. in dienst van het Interfacultair Reactor Instituut van de Technische Universiteit Delft.

2003-2004 Research Fellow aan de Universiteit van Exeter

List of Tables and Figures

List of Tables

List of Figures

Glossary[19]

π-expanded estimator: An unbiased estimator for a batch total.

Algorithm: Finite set of (simple) instructions used for solving a certain type of problem or achieving a certain result.

Analysis result: The value that is obtained after analysis.

Analyte: The compound of which the concentration or amount is measured.

Batch: The quantity of material whose properties are under study.

Batch size: The mass, volume, or number of particles in the batch.

Batch-to-sample size ratio: The mass or volume of a batch divided by the boundary value of the sample mass or volume respectively.

Batch total: Quantity that can be expressed as a summation over all particles of the batch.

Bias: The difference between the expected value of an estimator and the true value of the quantity that is estimated.

Bootstrapping: A technique for estimation of the variance that uses simulation.

Boundary value of the sample size: Parameters that characterizes the sample size in a model for the drawing of a sample.

Bulk sample: One or more increments of material taken directly from a batch. The bulk sample represents the batch in properties of interest and is, as a consequence, often much larger than the optimum laboratory size.

Classification: A division of the particles of a batch or sample into a finite number of classes, where it is assumed that particles belonging to the same class have identical properties.

Composite sample: Sample that consists of a finite number of increments.

Concentration: Property of a batch, sample or particle that is expressed as a ratio of two quantities, where the denominator is either the mass or volume of the batch, sample or particle respectively.

Contamination: The unwanted addition to the sample of material that influences the final analysis result.

Covariance: Parameter that describes the degree to which two random variables depend on each other statistically.

Empirical theory: A theory for the sampling of particulate materials that is not based on a model for the drawing of a sample on the level of the particles.

Estimate: Value derived from a sample that aims to represent a batch.

Estimation: The process of arriving at an estimate from a sample.

Estimator: Random variable whose value in a sample is an estimate.

[19] The descriptions given are not rigorous definitions.

Expected value: Parameter that describes the average value of a random variable.

Extreme value: Value of a function obtained by equating the partial derivatives to zero. Includes minimum, maximum or saddle point.

Finite population correction: Often a small positive number or random variable that is subtracted from a variance estimate or estimator to account for the finite size of the batch from which the sample was drawn.

Finite population sampling: The drawing of a sample from a batch that contains a finite number of particles.

Gaussian: Corresponding to a bell-shaped probability distribution that is characterized by a mean and standard deviation.

Horvitz-Thompson estimator: Unbiased estimator for the variance of the π-expanded estimator.

Hypergeometric distribution: A special type of probability distribution, used when a sample contains a fixed number of particles that are drawn without replacement, where all possible samples have an equal probability.

Ideal sampling: Hypothetical mode of sampling where there is no influence of unwanted - mostly mechanical - factors that could lead to a biased selection of particles.

INAA: Acronym for the analytical technique instrumental neutron activation analysis.

Increment: An individual portion of material collected by a single operation of a sampling device, from parts of a batch. Increments may be either tested individually or combined and tested as a unit.

Inclusion probability: The probability that a particle is sampled during the drawing of a single sample.

Indicator: Random variable whose value is one if a particle is part of the sample and zero if the particle is not part of the sample

Laboratory sample: A sample as prepared for sending to the laboratory and intended for inspection or testing.

Loss: The unwanted removal from the sample of material that was part of the original sample.

Mass concentration: Property of a batch, sample or particle that is expressed as a ratio of two quantities, where the denominator is the mass of the batch, sample or particle respectively.

Maximum: A value which is larger than or equal to all values in a certain set of values.

Measurand: A quantity subjected to measurement.

Minimum: Value which is smaller than or equal to all values in a certain set of values.

Minimum sample mass: The mass of a sample for which the relative standard deviation is equal to a maximum allowable value.

Minimum sample size: The size of a sample for which the relative standard deviation is equal to a maximum allowable value.

Model: A simplified version of the phenomenon it seeks to describe that focuses on the essentials of the problem.

Multinomial tree: A graphical representation of a sampling process.

Nomogram: A diagram with tree axes which can be used to obtain the value of a third variable when the values of the other two variables are given.

Non-empirical theory: A theory for the sampling of particulate materials that is based on a model for the drawing of a sample on the level of the particles.

Normal: see Gaussian.

Normality: The degree to which a distribution can be characterized by the Gaussian distribution. There is no standard way in which normality is defined.

Parameter of the size-variance relationship: Parameter whose value determines the numerical relation between the sample size and the variance.

Particle size: The mass or volume of a particle.

Particulate material: Material that consists of discrete physical entities of arbitrary size and shape.

Probability: A value that can be assigned to the likelihood of a future possibility. The probability is one when it is certain that the event will happen and zero when it is certain that the event will not happen.

Probability distribution: Variation of the probability as a function of the value of a parameter or a measurand.

Random variable: A function that assigns a value to every possible sample.

Relative standard deviation: The square root of the variance divided by the true value.

Safe value for the variance: Value for the variance based on limited knowledge of the properties of the particles in the sample that is certainly larger than or equal to the value based on all the properties of the particles in the sample.

Sample total: Quantity that can be expressed as a summation over all particles of a sample.

Sample: A portion of material selected from a larger quantity of material.

Sample preparation: The process of extracting a test portion from a laboratory sample.

Sample processing: Sampling, homogenizing, milling, blending, mixing, subsampling, sample preparation, and analysis.

Sample size: The mass of, volume of, or number of particles in a sample.

Sample-to-particle size ratio: The mass or volume of a sample divided by the particle mass or volume respectively.

Sampling: The process of obtaining a sample.

Sampling error: The difference between the estimate derived from a sample and the corresponding true value of the population from which the sample was drawn.

Sampling error due to non-ideal sampling: Part of the sampling error that is caused by non-ideal sampling.

Sampling error due to the distribution of non-identical particles: Part of the sampling error that is caused by the distribution of non-identical particles in the population from which the sample is drawn.

Sampling standard: The recommended process for extracting a sample.

Simulation: Virtual reconstruction of a process.

Size-based multinomial selection: The process of extracting a sample of a certain size from a batch that can contain an arbitrary number of different particles.

Size-variance relationship: Defining how a sample size varies with the variance ascribed to the sample.

Subsample: A portion taken from a sample. A laboratory sample may be a subsample of a bulk sample; similarly, a test portion may be a subsample of a laboratory sample.

Supreme value: A value which is larger than or equal to all values in a certain set of values.

Target area: Part of the volume of the batch whose particles would be selected if sampling were ideal.

Test portion: Quantity of material, of proper size for measurement of the concentration or other property of interest.

Theory: A coherent system of one or more models and corresponding theoretical results in a certain scientific discipline.

Uncertainty: An estimate attached to a measurement result which characterises the range of values within which the true value is asserted to lie.

Variance: Parameter that describes the dispersion of a random variable around its expected value.

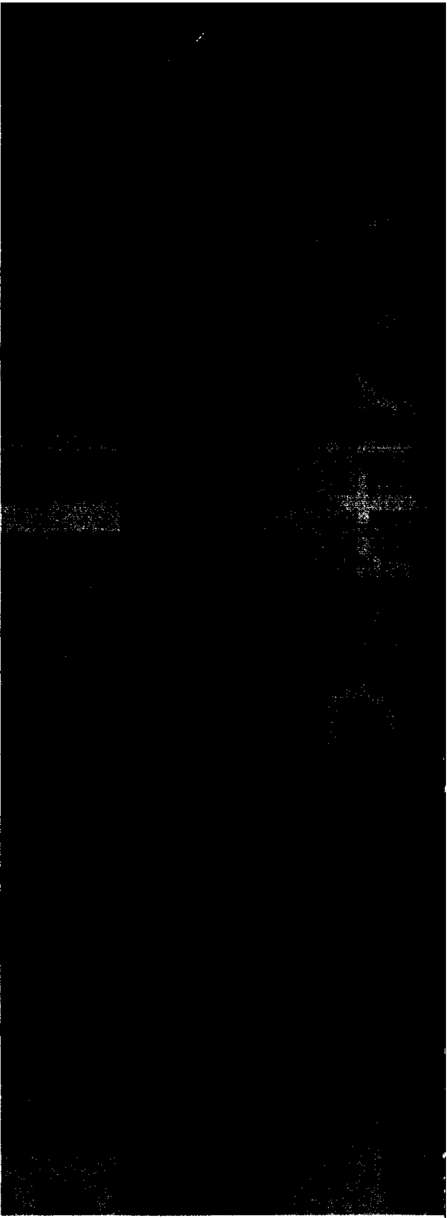
Variance estimator: Random variable that is used to estimate the variance using the properties of the particles in a sample.

Volume concentration: Property of a batch, sample or particle that is expressed as a ratio of two quantities, where the denominator is the volume of the batch, sample or particle respectively.

List of symbols[20,21]

[20] The symbols comprised in this list are used for the development of the new sampling theory, described in Chapter 3 to 6 and Chapter 8. The page number refers to the place(s) where the symbol is defined in this part of the thesis. This definition is consistently used throughout Chaper 3 to 6 and Chapter 8.

[21] When a symbol contains a single index, the variable i is chosen. When a symbol contains two indices, for the first index the variable i is chosen and for the second index the variable j is chosen. The variables i and j can take any arbitrary integer value between 1 and T (the number of particle classes) or between 1 and N_{batch} (the number of particles in the batch). However, some symbols that contain two indices are not defined for i=j.



ISBN 90-407-2517-9

Delft Un

There are many chemical, physical or biological properties of particulate materials that are of crucial importance for economic, agricultural, environmental or health-related reasons. These properties are often estimated using chemical, physical or biological tests on one or more laboratory samples. During the drawing of a laboratory sample, but also during the drawing of the bulk sample from which the laboratory sample is extracted, sampling errors can occur. These sampling errors are caused by the sampling of non-identical particles.

Sampling of Particulate Materials: a New Theoretical Approach presents a new theory for estimation of the variance caused by the sampling of random mixtures of non-identical particles and for determination of the minimum sample size.

TU Delft

Delft University of Technology

Del